分类号： TP242.6　　学校代码:10363
密　级： 公开　　学　　号：2220320141

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目 监控视频中的行人检测与跟踪算法研究

论文作者 桑建斌
校内导师 杨会成（教授）
校外导师 毛德平
专业学位类别 电子信息
研究方向（领域） 人工智能

论文提交日期： 2025 年 5 月 12 日

分类号：TP242.6　　单位代码：10363

密　级：公 开　　学　号:2220320141

题　　目 监控视频中的行人检测与跟踪算法研究

英文并列

题　　目 **Research on Pedestrian Detection and Tracking Algorithm in Surveillance Video**

学生姓名：桑建斌

校内导师：杨会成

校外导师：毛德平

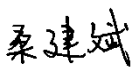
专　　业：电子信息

研究方向：人工智能

论文答辩日期：2025 年 5 月 12 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名： 

日期：　2025 年 5 月 12 日

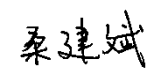
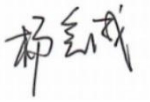
安徽工程大学学位论文版权使用授权书

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

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：  　　指导教师签名： 

日期：　2025 年 5 月 12 日　　日期：　2025 年 5 月 12 日

监控视频中的行人检测与跟踪算法研究

摘 要

视频监控场景中，行人检测与跟踪技术能够对行人目标进行精确识别、进而实现动态行为解析及异常状态预警，广泛应用于公共安全、交通管理等领域。然而，监控视频中，复杂的背景信息干扰和行人特征的多样性，容易导致行人检测算法出现漏检、错检等情况。同时由于行人互相遮挡与外观相似等问题，进一步导致行人跟踪算法效果不佳，给目前的行人检测跟踪算法带来了巨大挑战。针对上述问题，本文提出了一种结合多尺度特征检测和增强局部特征提取的抗遮挡行人检测跟踪框架，通过增强特征提取能力与优化数据关联策略，显著提升了复杂监控场景下的检测精度与跟踪稳定性。本文的主要研究内容与结论如下：

（1）针对行人检测中场景背景干扰、目标尺度多样化、遮挡严重等问题，构建了结合可变形卷积、全局注意力机制和动态任务对齐检测头的多尺度行人检测模型 YOLOv8n-DGT。首先设计了结合可变形卷积的 C2f-DCNv2 模块，增强模型对不同尺度行人特征的动态捕捉能力；然后在颈部网络中融入全局注意力机制抑制背景干扰，强化语义与位置信息融合；最后，设计了动态任务对齐检测头，促进模型分类和定位任务之间的交互，增强对遮挡行人目标的特征提取能力，并采用共享参数策略降低计算复杂度。在 Pascal VOC2012 数据集的实验结果表明，相比基线模型，本文构建的 YOLOv8n-DGT 行人检测模型在计算量未显著增加的情况下，精确率、召回率、平均精度分别提升了 4.8%、2%、2.4%，实现了检测性能与实时性的平衡。

（2）针对行人跟踪中行人互相遮挡、身份切换频繁等问题，采用前文构建的多尺度行人检测模型作为检测器，构建增强局部特征提取的抗遮挡行人跟踪模型。首先，构建多分支协作网络来提取局部行人特征，结合全局对比池化分支与关系分支增强遮挡目标的表征能力。其次，优化数据关联策略，设计动态阈值调整机制并结合 Inner-CIoU 匹配，提升遮挡场景下的匹配精度。最后，在 MOT16 数据集上对改进模型效果进行验证。实验结果表明，相比于基准检测跟踪模型，本文构建的抗遮挡行人检测跟踪模型的 MOTA、HOTA 和 IDF1 分别提高了 3.86%、2.07%和 3.71%，身份切换次数降低了 176 次，对监控场景下的遮挡行人的跟踪效果更加稳定。

综上，本文提出的行人检测与跟踪框架在检测精度、跟踪稳定性与抗遮挡性能上均优于现有主流算法，为智能监控系统提供了高效可靠的技术支撑。

关键词：行人检测，抗遮挡跟踪，可变形卷积，动态任务对齐，多分支协作网络

RESEARCH ON PEDESTRIAN DETECTION AND TRACKING ALGORITHM IN SURVEILLANCE VIDEO

ABSTRACT

In video surveillance scenarios, pedestrian detection and tracking technology can accurately identify pedestrian targets, thus realizing dynamic behavior analysis and abnormal state warning, which is widely used in public safety, traffic management and other fields. However, the complex background information interference and the diversity of pedestrian features in the surveillance video can easily lead to the omission and misdetection of pedestrian detection algorithms. Meanwhile, due to the problems of pedestrians obscuring each other with similar appearance, it further leads to the poor effect of pedestrian tracking algorithms, which brings great challenges to the current pedestrian detection tracking algorithms. To address the above problems, this paper proposes an anti-obscuring pedestrian detection and tracking framework that combines multi-scale feature detection and enhanced local feature extraction, which significantly improves the detection accuracy and tracking stability in complex monitoring scenarios by enhancing the feature extraction capability and optimizing the data association strategy. The main research content and conclusions of this paper are as follows:

(1) A multi-scale pedestrian detection model YOLOv8n-DGT combining deformable convolution, global attention mechanism, and dynamic task-aligned detection head is constructed to address the problems of scene background interference, diverse target scales, and severe occlusion in pedestrian detection.Firstly, the C2f-DCNv2 module

combining deformable convolution is designed to enhance the model's dynamic capture ability of pedestrian features at different scales; Then, a global attention mechanism is incorporated into the neck network to suppress background interference and enhance the fusion of semantic and location information; finally, a dynamic task-aligned detection head is designed to facilitate the interaction between the model classification and localization tasks, to enhance the feature extraction capability of occluded pedestrian targets, and to reduce the computational complexity by adopting a shared-parameter strategy. Experimental results on the Pascal VOC2012 dataset show that compared with the baseline model, the YOLOv8n-DGT pedestrian detection model constructed in this paper improves the precision, recall, and average accuracy by 4.8%, 2%, and 2.4%, respectively, without a significant increase in computation, and achieves a balance between detection performance and real-time performance.

(2) Aiming at the problems of mutual occlusion of pedestrians and frequent identity switching in pedestrian tracking, the multi-scale pedestrian detection model constructed in the previous section is used as a detector to construct an anti-occlusion pedestrian tracking model with enhanced local feature extraction. First, a multi-branch collaborative network is constructed to extract local pedestrian features, combining global contrast pooling and relational branching to enhance the characterization of occluded targets. Second, the data association strategy is optimized, and the dynamic threshold adjustment mechanism is designed and combined with Inner-CIoU matching to improve the matching accuracy in occlusion scenarios. Finally, the effect of the improved model is validated on the MOT16 dataset. The experimental results show that compared with the baseline detection and tracking model, the anti-obscuring pedestrian detection and tracking model constructed in this

paper improves the MOTA, HOTA, and IDF1 by 3.86%, 2.07%, and 3.71%, respectively, and reduces the number of identity switching by 176 times, which makes the tracking effect of occluded pedestrians in surveillance scenes more stable.

In summary, the pedestrian detection and tracking framework proposed in this paper outperforms the existing mainstream algorithms in terms of detection accuracy, tracking stability and anti-obscuration performance, and provides efficient and reliable technical support for intelligent surveillance systems.

KEY WORDS: pedestrian detection, occlusion resistant tracking, deformable convolution, dynamic task alignment, multi-branch collaborative networks

目录

第 1 章 绪论

1.1 研究背景与意义

随着智慧城市建设和社会治理智能化的快速推进，视频监控系统已成为维护公共安全、优化交通管理、提升社会服务的重要支撑手段。近年来，我国政府大力推进“平安城市”“智慧城市”“智慧交通”等国家级战略项目，推动视频监控产业从传统被动式安防向智能化、主动化方向升级。图 1-1 中来自中国报告大厅的数据显示，截至 2023 年，我国视频监控设备市场规模已突破 1271.5 亿元，年增长率达 6.54%。海量监控数据的实时分析与处理需求，对目标检测与跟踪技术的精度、鲁棒性及实时性提出了更高要求[1]。在此背景下，行人检测与跟踪技术作为智能监控系统的核心技术之一[2]，其研究价值与应用潜力日益凸显。因此，越来越多的学者致力于监控视频中的行人检测与跟踪技术研究。

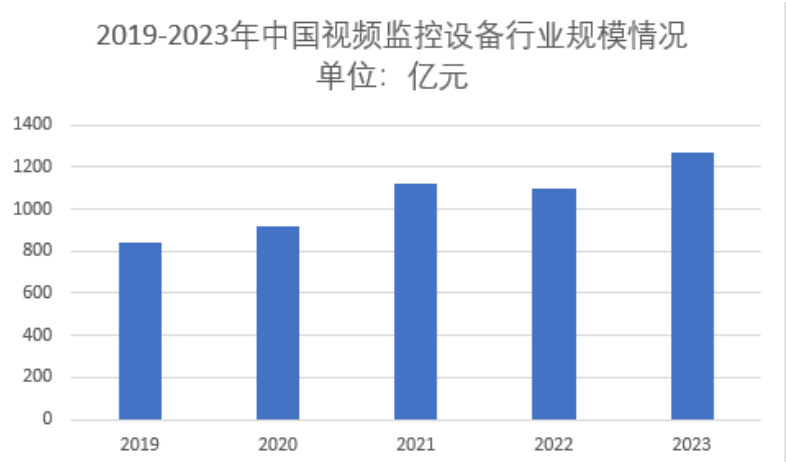


图 1-1 2019-2023 年我国视频监控设备行业规模情况

Fig 1-1 Size of China's Video Surveillance Equipment Industry, 2019-2023

当今社会，行人检测与跟踪技术的应用场景不断拓展，从传统的公共安全、交通管理、金融安全等领域，逐步向智慧城市、工业生产、家庭安防等新兴领域延伸。在智慧城市建设中，行人检测与跟踪技术可以实时统计人流密度、优化交通信号控制、预防公共安全事故，为城市治理提供数据支撑。例如，在交通枢纽与商业中心，该技术可实时监测人流动态，预警踩踏风险；其结合人脸识别技术，还可协助警方快速锁定嫌疑人，提升案件侦破效率。在工业领域，该技术通过追踪工人活动轨迹，规避高危区域风险，优化生产流程布局；在家庭安防场景中，

可区分家庭成员与入侵者并检测人员异常行为，实现智能预警与健康监护。

尽管行人检测与跟踪技术已在当今社会的各个领域得到广泛应用，但在监控视频中实现精准稳定的行人检测与跟踪仍然面临诸多难点和挑战[3]。实际监控场景中行人外观多样、相互遮挡频繁、运动轨迹复杂，加之复杂背景干扰等问题，导致现有算法易出现漏检、错检及身份切换频繁等现象。

在检测领域，YOLO 系列算法凭借“单阶段”设计在速度与精度间实现了较好平衡，但 YOLOv8 等检测算法在密集遮挡场景下仍存在特征提取不充分、分类与定位任务耦合不足等问题。传统卷积的固定感受野限制了其对不同尺寸目标的适应性，而注意力机制的应用多局限于局部特征交互，全局语义信息利用不足。在跟踪领域，基于检测的跟踪框架通过融合运动预测与外观特征提升了稳定性，但 StrongSORT 等算法在长时遮挡场景下仍依赖固定阈值匹配策略，导致身份切换频繁。此外，现有特征提取网络对局部细节的捕捉能力有限，难以应对部分遮挡目标的跟踪需求。针对上述问题，本文从提升行人检测与跟踪的准确性与鲁棒性方面进行研究，以提升对监控视频中遮挡行人的检测与跟踪效果。

1.2 国内外研究现状

1.2.1 行人检测研究现状

目标检测算法分为传统机器学习算法和基于深度学习的目标检测算法[4]。传统目标检测算法采用手工设计特征，并结合机器学习算法完成分类与识别。首先，传统目标检测算法会采用滑动窗口或选择性搜索算法对图像进行处理，以生成候选区域；随后，对这些区域提取特征，并判断其中是否存在目标，若检测到目标，则记录其位置坐标；然后，使用分类器对目标进行分类与识别；最后，通过非极大值抑制（Non-Maximum Suppression, NMS）去除冗余候选框，以提升定位精度。

传统机器学习方法依赖于人工设计特征（如 Haar、HOG、SIFT 等），结合分类完成目标识别和定位。2001 年，P.Viola 等人提出了 Viola-Jones[5]算法，主要用于人脸检测。该算法利用 Haar 特征和 Adaboost 分类器进行级联检测，通过 Adaboost 分类器从大量的 Haar 特征中筛选出少量但关键的特征，最后构建级联分类器逐步筛选候选区域，减少了背景区域的计算开销。但其特征表达能力有限，仅适用于固定视角和简单背景场景。2006 年，Navneet Dalal 等人提出了方向梯

度直方图（Histogram of Oriented Gradients, HOG）+支持向量机（Support Vector Machine, SVM）[6]算法，通过计算局部区域的梯度方向直方图来构建特征向量，并利用支持向量机进行分类，在行人检测任务中取得了进一步突破，然而该算法在复杂场景下的鲁棒性较差，且所采用的手工特征难以捕捉语义信息。2008 年，P. Felzenszwalb 等人提出了可变形部件模型（Deformable Parts Model, DPM）[7]算法，该算法在 HOG 特征的基础上引入了可变形部件模型，将目标拆分为多个可变形部件，通过部件间几何约束建模目标的形变特性，显著提升了形变目标的检测精度。但其多阶段训练流程复杂，且对部件间空间关系的建模能力有限，难以适应行人姿态剧烈变化的场景。

综上，传统机器学习方法过于依赖人工特征设计，无法有效应对复杂背景、光照变化、目标遮挡等问题，且计算效率低，难以满足行人实时检测的需求。随着深度学习技术的高速发展，基于卷积神经网络的检测算法的性能得到了显著提升[8]，并在各个领域得到了大量应用。目前，该检测方法主要分为两阶段目标检测算法和单阶段目标检测算法[9]。

两阶段目标检测算法将目标检测任务分为两个步骤：首先生成一系列候选区域，然后对这些候选区域进行精细的分类和位置回归[10]。2014 年，Girshick 等人提出 R-CNN[11]算法，这是目标检测领域首个结合深度学习的方法。其首先使用选择性搜索算法生成大量候选区域，随后通过卷积神经网络（Convolutional Neural Network, CNN）提取区域特征，并最终借助 SVM 分类器确定目标类别。R-CNN 算法在 VOC2007 行人数据集[12]上取得了显著的性能提升，但存在计算量大、检测速度慢等问题。针对 R-CNN 算法在特征提取过程中存在的重复计算问题，同年，何凯明等人提出了空间金字塔池化网络[13]（Spatial Pyramid Pooling Network, SPPNet）。该算法引入了空间金字塔池化层，使 CNN 能够处理不同尺寸的输入图像并进行特征提取。因此，SPP-Net 只需对整幅图像进行一次特征计算，然后利用 SPP 层对各候选区域执行特征映射，从而显著提升了检测效率。2015 年，Girshick 对 R-CNN 进一步改进，提出了 Fast R-CNN[14]算法。该算法引入了 ROI Pooling（Region of Interest）层来替代 SPP 层，实现了对候选区域的特征提取。此外，Fast R-CNN 算法将分类和边界框回归任务整合到同一个网络中，实现了端到端的训练。然而，其候选区域的生成仍然依赖于传统方法。次年，Ren

等人提出了 Faster R-CNN[15]算法，该算法通过区域建议网络网络[16]（Region Proposal Network, RPN）实现候选区域的生成与特征提取，并通过端到端的训练方式优化检测效果。RPN 网络采用滑动窗口机制，在特征图上生成多个候选区域，并对其进行目标存在性判别与位置回归计算。通过该设计，Faster R-CNN 算法可以一次性完成候选区域生成、特征提取、分类和位置回归四项任务。但其采用多阶段架构，模型参数量较大，难以部署至移动端设备。

虽然两阶段算法在行人检测的精度方面取得了显著提升，但因模型过大，流程复杂，其检测速度始终偏低。而单阶段目标检测算法直接将目标检测转化为分类问题，直接对输入图片或视频进行处理，预测出目标的边界框位置和类别，只需一步即可完成检测。与两阶段算法相比，单阶段算法省去了候选区域提取环节，提升了行人检测速度，并有效降低了模型复杂度，更符合行人检测任务的需求。

YOLO 算法在单阶段目标检测领域具有开创性意义。早在 2016 年，Redmon 等人便提出了 YOLOv1[17]（You Only Look Once）算法，其核心思想是将输入图像划分为 S× S 个网格，将输入图像划分为 S×S（在 YOLOv1 中 S=7）的网格，每个网格负责识别其中的目标，并对该目标的边界框和类别进行预测，大幅提升了检测速度。但由于其输出层采用固定尺寸网格，导致小目标行人漏检率高，定位精度不足。同年，Wei Liu[18]等人提出了 SSD（Single Shot MultiBox Detector）算法。在 SSD 算法中，主干网络通常由 VGG16[19]中的部分卷积层组成，用于对图像进行特征提取。该算法同时融合了 YOLO 中回归策略与 Faster R-CNN 中的锚框机制，并采用多尺度特征图进行预测。这种设计提升了算法对各种尺寸行人的检测能力，从而实现了与两阶段算法相当的精度。但 SSD 算法依赖于较浅层的特征图来检测小目标，导致其在小目标行人检测上效果较差。2017 年，Redmon 等人改进了 YOLOv1 的不足，引入了批归一化（Batch Normalization, BN）、锚框、维度聚类等技术，推出了 YOLOv2[20]算法。该算法采用 Darknet-19 作为新的特征提取网络，简化了网络结构，进一步提升了目标检测速度，但对密集目标的处理仍存在局限性。2018 年，Redmon 等人提出了 YOLOv3[21]算法，在上一代算法基础上，引入多维度检测策略和 Darknet-53 骨干网络，提升了小尺寸目标的检测精度。同时，利用特征金字塔网络（Feature Pyramid Networ, FPN）网络在不同的尺度上进行预测，支持多标签分类任务，但模型参数量较大，边缘设备部署困难。

随后 2020 年，Bochkovskiy 等人提出了 YOLOv4[22]算法，引入了 CSPDarknet53 和多种增强技术（如 Mosaic 数据增强、DropBlock 正则化等），实现了检测性能与速度的平衡，但在复杂背景干扰下仍存在误检问题。同年，Ultralytics 提出 YOLOv5[23]算法，骨干网络由 BottleneckCSP（Cross Stage Partial）和 Focus 模块构成，并采用 FPN 和路径聚合网络（Path Aggregation Network, PAN）相结合的特征融合网络，显著提高了模型的特征提取能力，但其耦合检测头设计导致分类与定位任务的特征交互不足。

近几年 YOLO 系列算法发展迅速，检测速度和精度有了显著的提升，并逐渐成为行人检测的主流算法。2021 年，旷世科技提出了 YOLOX[24]算法，该算法首次引入了解耦头（Decoupled Head）设计，并采用无锚框（Anchor-free）检测策略，使得定位输出仅需四个参数，简化了检测流程，同时进一步提升了定位精度。2022 年，美团公司和 Wang 等人相继提出了 YOLOv6[25]和 YOLOv7[26]算法，其都采用了重参数化卷积的技术，并优化骨干网络，进一步增强了模型的检测性能。但两者复杂的网络结构和多分支设计，导致检测速度有所降低。2023 年，Ultralytics 提出了 YOLOv8，延续了 YOLOv5 中无锚框的检测策略，在骨干网络结构中，将 C3 模块替换为更高效的 C2f 模块，并根据不同尺度模型相应的调整通道数量，检测头也采用了解耦头设计，提升了目标定位精度与检测速度，但其在密集行人场景中仍面临遮挡导致的特征提取不充分问题。

综上，两阶段目标检测算法虽然检测精度较高，但计算复杂、实时性较差；而单阶段目标检测算法虽能实现高效检测，却难以兼顾复杂场景下行人检测的精度和鲁棒性。因此，当前目标检测模型用于行人检测任务的主要挑战是如何在保持实时性能的同时，进一步提升对复杂场景下的不同尺寸行人和遮挡行人的检测精度。针对上述问题，本文将以 YOLOv8 为基准模型，进一步研究多尺度行人检测模型，以解决现有算法在特征提取和多任务交互中的不足，从而为复杂监控场景下的行人检测提供更为精确、鲁棒且高效的解决方案。

1.2.2 行人跟踪研究现状

多目标跟踪（Multiple Object Tracking, MOT）是计算机视觉领域的核心技术，旨在对视频中的多个目标进行持续定位与身份关联。根据发展脉络和技术特点，

MOT 算法可分为经典跟踪算法与基于深度学习的跟踪算法两大类。

经典跟踪算法主要依赖于传统的图像处理技术和机器学习模型，代表方法包括粒子滤波[27]、Meanshift 算法[28]和 Camshift 算法[29]和基于特征点的光流算法[30]。粒子滤波算法通过构造一组随机样本（粒子）来描述目标状态的概率分布，并利用观测值计算每个粒子的权重对目标位置进行估计，但高维状态空间计算复杂度过高，实时性受限。Meanshift 算法则采用核密度估计构建目标颜色直方图模型，并通过迭代搜索概率密度梯度最大方向定位目标。但其缺乏运动模型约束，易受相似背景干扰导致跟踪漂移。Camshift 算法将 Meanshift 算法扩展至视频序列，并引入自适应窗口缩放机制，通过反向投影图动态调整搜索区域尺寸，增强对目标尺度变化的适应性。然而其依赖颜色特征建模，在光照突变场景下鲁棒性显著下降。光流法一般基于亮度恒定假设与空间一致性约束，通过计算相邻帧间像素点的运动矢量场预测目标轨迹。综上，经典跟踪方法在简单背景或运动较为平稳的场景中能够实现基本跟踪，但其对目标形态、遮挡和环境干扰的适应能力较弱。在处理行人多样化和连续遮挡问题时，经典跟踪模型易出现跟踪漂移和身份切换频繁的问题，难以满足实际监控场景中的行人跟踪需求。

基于深度学习的目标跟踪算法依托深度学习模型强大的特征提取与学习能力，在行人跟踪任务中获得了显著的性能提升。相比于传统方法，该类算法能够在高维特征空间中学习更加鲁棒的目标表征，有效降低了由于遮挡、目标形变、光照变化等因素的干扰，从而提升跟踪的准确性与稳定性。根据具体的算法范式，当前基于深度学习的跟踪算法可大致分为两类：基于检测跟踪（Tracking-by-Detection, TBD）和联合检测跟踪（Joint Detection and Tracking, JDT）两类[31]。

TBD 算法首先利用先进的目标检测器在每帧图像中检测目标实例，并结合运动模和数据关联策略进行目标的跨帧关联，从而实现稳定跟踪。相较于传统目标跟踪算法，TBD 算法在外观特征提取方面具有显著优势，能够更有效地区分相似目标，提高匹配精度。JDT 算法则采用端到端框架同时完成目标检测与轨迹更新，即同时预测目标的位置、类别与其在时间序列中的身份信息。2020 年，Zhou 等人基于 CenterNet 结构提出了 CenterTrack 算法[32]，其核心创新在于将目标跟踪任务转化为目标中心点的跟踪任务，利用目标的空间位置进行有效的轨迹预测，避免了传统方法中需要额外重识别特征的繁琐过程。但其缺乏针对目标的

重识别特征，在目标发生长时间遮挡时，易出现身份跳变或跟踪失败。次年，Zhang等人提出了 FairMOT 算法[33]，该算法通过回归目标的中心点以及目标的宽高来进行目标检测，同时引入额外的分支网络进行目标重识别。通过该架构，FairMOT算法能够有效地平衡目标的检测和重识别任务，避免了传统方法中存在的检测与跟踪任务的冲突。但其结构较为复杂，可能在某些硬件资源受限的情况下，计算成本较高。

综上，JDT 算法在应对行人数量变化、行人密集等复杂场景时仍面临较大挑战，尚未大规模应用于行人跟踪任务。目前，TBD 算法是行人跟踪领域研究的主流方向，得益于其模块化设计，该算法可以灵活集成最新的检测、特征提取和数据关联技术，不断推动 MOT 算法在视频行人跟踪任务中性能优化。

2016 年，Bewley 等人提出了 SORT[34]（Simple Online and Realtime Tracking）算法。该算法检测模块集成了 Faster R-CNN 网络，同时引入基于卡尔曼滤波的运动状态估计模型来预测后续帧中的目标位置。然后通过计算检测器输出的目标框与运动模型预测框之间的交并比[35]（Intersection Over Union, IoU）来量化两者的匹配程度，最后采用匈牙利算法构建检测目标与跟踪轨迹的对应关系，有效实现了视频序列中跨帧目标关联。但 SORT 算法忽略了外观特征，导致模型在行人遮挡场景下身份切换（Identity Switches, IDSw）频繁。2017 年，Wojke 等人提出了 DeepSORT[36]算法，在 SORT 算法的基础上增加特征提取网络，进一步提取目标外观特征，同时构建了结合运动信息与外观信息的匹配机制。该算法通过融合目标的外观信息并引入级联匹配策略，有效减少 ID 切换并增强了跟踪稳定性，但其固定阈值匹配策略难以适应跟踪场景的行人数量变化。2022 年，Yunhao Du等人在 DeepSORT 的基础上提出了 StrongSORT[37]算法，其采用指数移动平均特征更新策略[38]（Exponential Moving Average, EMA）更新身份信息，增强相关系数（Enhanced Correlation Coefficient, ECC）[39]运动补偿优化跟踪鲁棒性，并将传统卡尔曼滤波替换为 NSA 卡尔曼滤波[40]，根据目标检测的质量自适应调整噪声尺度，最后用全局线性赋值匹配代替匹配级联，大幅提高了目标跟踪的准确率。但其重识别网络无法充分提取目标局部特征，导致模型在密集场景下的遮挡行人跟踪效果较差。

综上所述，相比于经典跟踪算法，基于深度学习的跟踪算法性能有所提升，

但在面对行人跟踪场景目标密集、遮挡严重等情况时，仍存在匹配不稳定、误检和 ID 频繁切换等问题。针对这些不足，本文拟在 YOLOv8 检测模型和 StrongSORT 跟踪模型的基础上进一步研究，构建增强局部特征提取的抗遮挡行人跟踪算法，以实现对被遮挡行人的稳定跟踪。

1.3 论文的主要研究内容

监控视频中行人检测与跟踪算法在智能安防、智慧交通等领域具有重要应用价值，然而复杂监控场景中存在的行人姿态多样性、行人间相互遮挡以及复杂背景干扰等问题，导致现有行人检测跟踪模型的效果较差。针对上述挑战，本文以 YOLOv8 检测模型与 StrongSORT 跟踪模型为基准，结合特征提取能力增强与数据关联策略优化，提出了融合多尺度特征检测与增强局部特征提取行人检测与跟踪模型，技术路线图见图 1-2，本文主要研究如下：

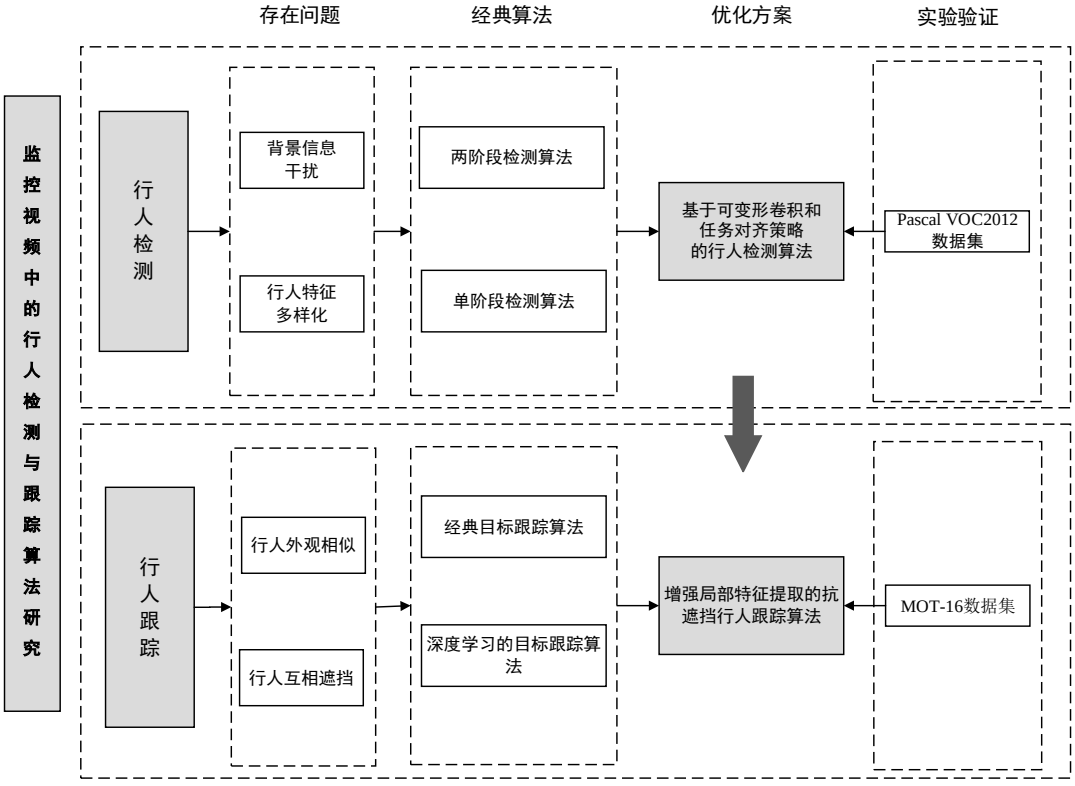


图 1-2 本文技术路线图

Fig 1-2 Technology roadmap for this paper

（1）针对复杂场景下行人检测存在的背景干扰严重、目标遮挡频发、目标特征多样化等问题，提出融合全局注意力机制、可变形卷积与动态任务对齐策略的多尺度行人检测模型。首先，采用可变形卷积重构 C2f 模块，增强模型对不同

尺度行人特征的提取能力。然后，在颈部网络中融入全局注意力机制，降低背景信息干扰，强化语义与位置信息融合；在头部网络部分，设计动态任务对齐检测头从多个卷积层中学习任务交互特征，并采用共享参数策略降低计算复杂度。

（2）在复杂的监控场景中，行人运动不规律、易出现互相遮挡等问题，导致行人跟踪算法的准确度低，身份切换次数频繁。针对上述问题，本文构建了增强局部特征的抗遮挡行人跟踪算法。首先，采用多分支协作网络作为模型特征提取网络，基于多尺度特征提取架构，采用四个协同分支提取行人局部特征，从而增强模型对被遮挡行人的跟踪能力。然后，引入了动态阈值调整机制，根据当前帧检测目标的数量，动态调整最大匹配阈值来适应复杂的跟踪环境，并引入 Inner-CIOU 优化匹配方式，提高跟踪匹配的稳定性。最后，通过实验验证了本文结合多尺度特征检测与增强局部特征提取行人检测跟踪模型的优越性。

1.4 论文的结构安排

本文共五个章节，各章节具体内容如下：

第一章：绪论。本章首先介绍行人检测与跟踪技术的研究背景与应用价值，然后结合国内外研究现状，梳理现有方法在监控场景中的技术瓶颈与面临的挑战，明确了本文研究方向和改进思路。最后，对论文的主要研究内容与组织结构进行概述，引出本文的创新点与研究目标。

第二章：行人检测与跟踪基本原理。本章系统阐述了行人检测与跟踪任务的基本技术原理。首先对卷积神经网络构成及相关原理进行介绍；然后深入分析了 YOLO 系列检测算法的结构与原理，并重点阐述了 YOLOv8 算法的相关原理；最后介绍了跟踪算法的理论基础，包括卡尔曼滤波、匈牙利匹配算法及 SORT 系列跟踪框架，并重点结合 StrongSORT 算法分析其不足之处，为后续算法改进提供了依据。

第三章：结合全局注意力与任务对齐策略的行人检测模型。本章首先分析了 YOLOv8 行人检测算法在复杂场景下存在的不足；然后提出一种结合可变形卷积、全局注意力机制与动态任务对齐检测头的多尺度行人检测模型 YOLOv8n-DGT，并阐述了其构建思路与可行性；最后在 Pascal VOC2012 数据集上进行实验，通过实验与可视化分析，验证了多尺度行人检测模型 YOLOv8n-DGT 的有效

性。

第四章：增强局部特征提取的抗遮挡行人跟踪模型。本章首先对 StrongSORT 不足之处进行分析，然后提出对 StrongSORT 算法的改进思路，阐述构建抗遮挡优化行人跟踪模型的可行性，最后在 MOT-16 数据集上对针对多尺度检测与抗遮挡优化的行人检测跟踪模型进行实验验证，通过消融实验、对比实验与可视化分析，验证了本文构建模型的检测与跟踪性能的优越性。

第五章：总结与展望。总结了全文研究内容与成果，并指出未来模型优化的方向。

第 2 章 行人检测与跟踪基本原理

2.1 卷积神经网络

卷积神经网络是一种基于卷积核的共享权重架构的前馈神经网络，主要由卷积层、激活层、池化层和全连接层所构成[41]，网络结构如图 2-1 所示，其核心思想是通过局部感知和参数共享高效提取数据中的空间或时序特征。CNN 在图像处理领域表现尤为突出，是计算机视觉任务（如图像分类、目标检测）的基础模型，同时也是深度学习中代表性算法之一。在 21 世纪后，以卷积神经网络为基础的深度学习算法得到了快速发展，被广泛运用于人脸识别、物体检测等领域。

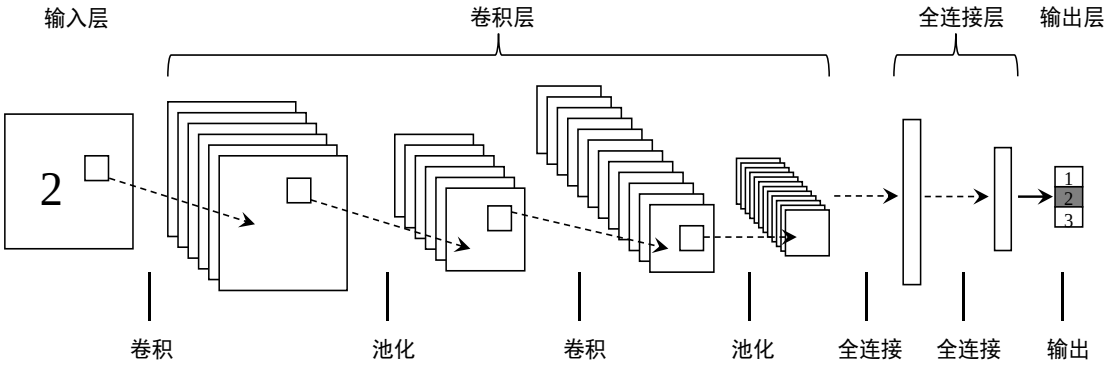


图 2-1 卷积神经网络结构

Fig 2-1 Structure of convolutional neural network

2.1.1 卷积层

卷积层作为卷积神经网络的核心部分，由若干个卷积单元构成，每个单元通过卷积运算提取输入数据的不同特征。通过卷积运算和部分池化操作，卷积层能有效压缩输入数据的维度，进而降低后续网络层的计算量。这有助于模型剔除冗余信息并保持重要特征，从而提升整体运算效率。实际上，卷积运算是一种特殊形式的线性变换，其基本思想是利用卷积核在输入数据上滑动，并对每个局部区域执行加权求和，最终生成特征图。以一个 3×3 的卷积核为例，如图 2-2 所示，卷积操作包含以下步骤：

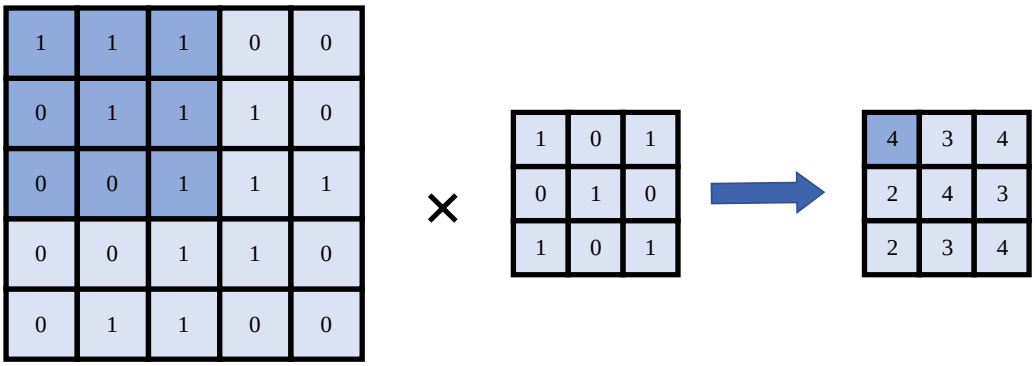


图 2-2 卷积运算流程

Figure 2-2 Convolutional operation flow

（1）确定卷积核大小和步长：卷积核尺寸决定了其感受野的覆盖范围，即可以感知到的输入数据局部区域。步长则表示卷积核在输入数据上每次移动的距离。

（2）进行滑动窗口操作：卷积核在输入数据上按照指定的步长进行滑动，每次滑动到一个新的位置时，就计算该位置局部区域的加权和。

（3）加权求和与偏置：对于每个局部区域，先将卷积核中各个权重与相应位置的输入数据相乘，然后将所有乘积求和得到一个加权总和，再加上偏置项，形成该位置的最终输出。

（4）生成特征图：重复上述步骤，直到卷积核遍历所有输入数据，生成相应的特征图。这个特征图包含了输入数据在卷积核所关注的局部特征上的响应。

2.1.2 池化层

在 CNN 中，池化层具有关键作用。它不仅能降低数据的空间维度和计算复杂度，还能提高模型的泛化能力和训练速度。其原理是对输入特征图进行下采样操作，以压缩数据的空间维度，同时提取输入图像的重要特征。主要流程是通过在输入特征图上滑动一个固定大小的窗口（也称为池化窗口或过滤器）来实现。在每个窗口内，执行特定的下采样操作，最常用的操作方式有两种：最大池化和平均池化，计算方法如图 2-3 所示。最大池化通常挑选池化区域内最大的数值作为输出，这样能够保留特征图中的关键特征，对输入数据的微小变化具有一定的鲁棒性。而平均池化则是计算则通过计算窗口内所有像素的均值来生成输出，以此保留特征图的平均特征，常用于提取图像中的背景信息或全局特征。

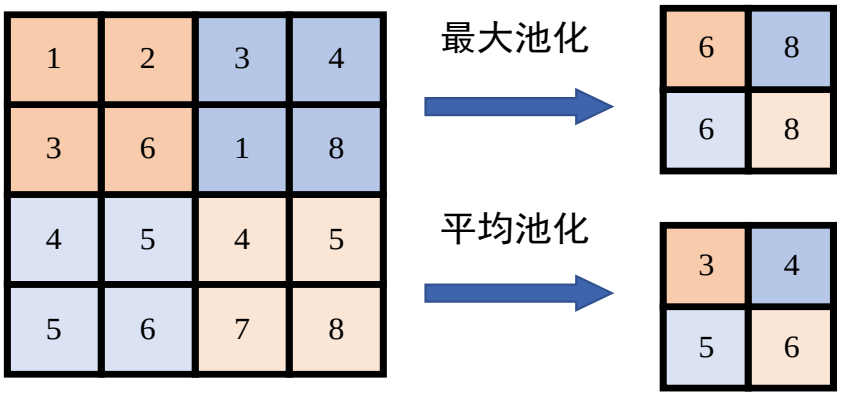


图 2-3 最大池化与平均池化

Figure 2-3 Maximum Pooling and Average Pooling

2.1.3 全连接层

全连接层是 CNN 的重要组成部分。在全连接层中，每个神经元都与前一层的所有神经元相连接，如图 2-4 所示。这种连接方式使得全连接层能够综合前一层提取的特征信息，并与对应的权重进行线性组合，以产生新的特征表示。全连接层执行线性变换操作，主要包括两个参数：权重矩阵和偏置向量，输出结果如下式：

$$z = Wx + b \tag{2-1}$$

其中，x为输入向量，W 为权重矩阵，b为偏置向量，z为输出向量。

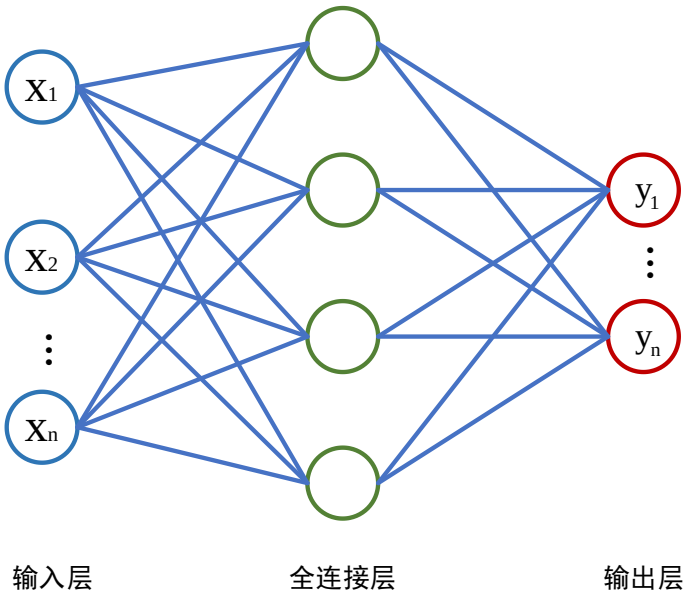


图 2-4 全连接层示意图

Figure 2-4 Schematic diagram of the full connectivity layer

2.1.4 激活函数

激活函数主要用于在 CNN 中引入非线性特性，使模型能够学习和表示复杂的非线性关系。这种非线性特性使得模型能够学习更为复杂的特征和数据分布，从而增强模型对复杂函数的表达能力。在 CNN 中，每个神经元接收来自上一层的输入信号，这些信号经过加权求和后，通过激活函数进行非线性变换，生成输出信号。如果没有激活函数的引入，神经网络的层与层之间仅仅是线性组合，无论网络的深度如何增加，整体上仍是一个线性模型，无法有效地拟合复杂的非线性数据。这个过程可以表示：

$$z = f\left(Wx + b\right) \tag{2-2}$$

其中 x 是输入，W 是权重，b 是偏置，f 是激活函数，z 是输出。常见的激活函数包括：Sigmoid、Tanh 和 ReLU

（1）Sigmoid

Sigmoid 函数是一种应用广泛的激活函数，其数学形式与几何形式如下：

$$f\left(x\right) = \frac{1}{1+e^{-x}} \tag{2-3}$$

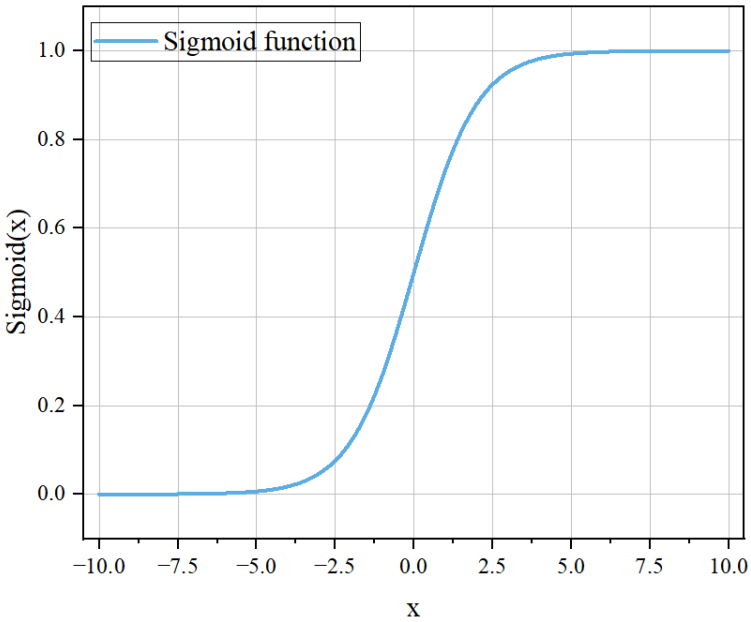


图 2-5 Sigmoid 激活函数

Figure 2-5 Sigmoid activation function

它输出范围在（0，1）之间，主要应用于神经网络的输出层，特别适用于二分类问题中概率值的映射。其优点在于输出范围明确且具有概率解释性，能够直观反映样本归属类别的置信度。然而，Sigmoid 函数存在显著缺陷：当输入较大或较小时，梯度趋近于零，导致梯度消失，影响深层网络的训练；同时，其输出不以零为中心，可能会引发权重更新方向的偏置，降低模型收敛速度。此外，指数运算增加了计算复杂度，在处理大规模数据时影响实时性。随着深度学习发展，ReLU 等激活函数因计算高效的优势逐渐取代了 Sigmoid 在隐藏层的地位，但其在概率建模、门控机制等特定场景仍然适用。如图 2-5 所示为 Sigmoid 函数的几何形式。

（2）Tanh

Tanh 函数（双曲正切函数）是 Sigmoid 函数的平移和缩放版本，其数学形式与几何形式如下：

$$f(x)=\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2-4}$$

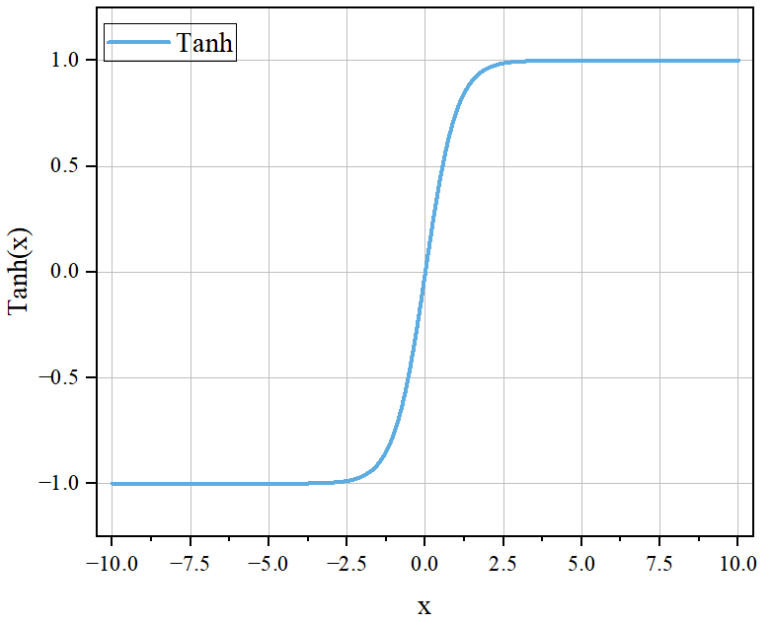


图 2-6 Tanh 激活函数

Figure 2-6 Tanh activation function

其输出范围在（-1，1）之间。与 Sigmoid 函数相比，其零均值特性有效缓解了神经网络训练中梯度更新的偏置问题，从而加速模型收敛。Tanh 函数在输入接近零时梯度较大，有效缓解了浅层网络的梯度消失现象，提高了参数更新效率，

常用于隐藏层的激活函数。但当函数输入绝对值较大时，梯度消失的问题仍然存在。此外，Tanh 函数的计算量过高，对大规模数据处理的效率较差。Tanh 函数因其对称输出特性被广泛使用，但随着 ReLU 等更高效激活函数的出现，其应用逐渐减少，但仍常见于需要输出归一化的场景或特定模型的中间层设计中。如图 2-6 所示为 Tanh 函数的几何形式。

（3）ReLU

ReLU 函数（Rectified Linear Unit）即线性整流函数，数学形式与几何形式如下：

$$f\left(x\right)=max\left(0,x\right) \tag{2-5}$$

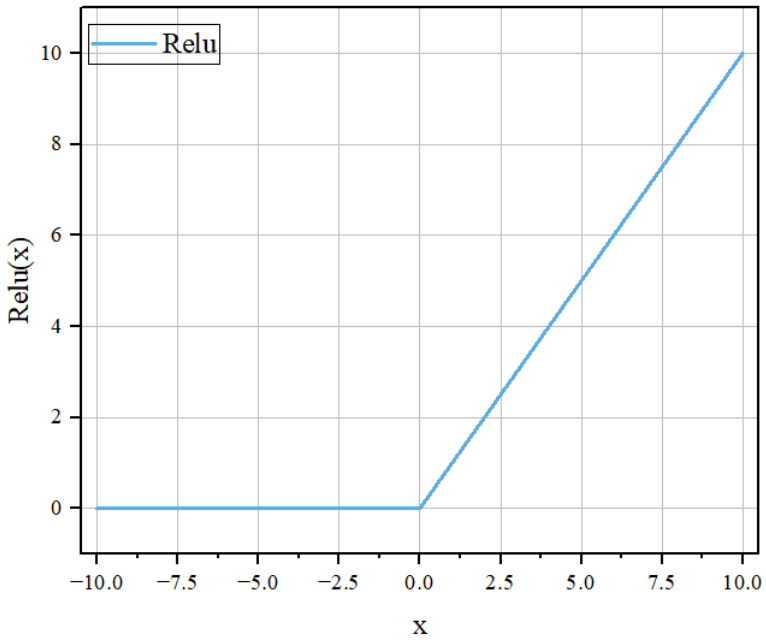


图 2-7 ReLU 激活函数

Figure 2-7 ReLU activation function

ReLU 函数也常用于隐藏层，其核心原理是通过对输入值进行线性整流，保留正数部分并将负数置零，从而为神经网络引入非线性能力。相较于 Sigmoid 等饱和激活函数，ReLU 在输入大于 0 时梯度恒为 1，有效缓解了深层网络中的梯度消失问题。同时其计算仅需简单阈值判断，大幅提升了训练效率。然而，在输入为负值时，其导数为零，可能导致对应神经元永远无法激活。同时 ReLU 函数的输出不以零为中心，可能导致梯度更新不均匀，影响网络性能。尽管存在这些缺陷，ReLU 仍因其高效性和训练稳定性成为神经网络的主流激活函数。如图 2-7 所示为 ReLU 函数的几何形式。

2.2 YOLO 系列算法原理

YOLO 系列算法通常将目标检测任务转化为回归问题，直接从输入图像或视频中预测物体的边界框和类别，无需事先定义候选区域，一般采用卷积神经网络进行特征提取和目标分类定位，只需一次前向传播即可完成检测任务。相比于两阶段算法，YOLO 系列算法省略了候选区域生成的环节，实现了“一步到位”的高效检测。YOLO 检测网络主要由三个部分构成：骨干网络、颈部以及检测头。骨干网络负责从输入数据中提取关键特征；颈部则起到连接骨干网络与检测头的作用，通过融合不同尺度的特征图来增强信息表达；最后，通过检测头对这些多尺度特征对图像中的目标进行类别和位置的预测。

（1）YOLOv1 算法

YOLOv1 算法进行检测时，通常将一张输入为 $448_{\times}448$ 像素的图像，通过 YOLOv1 网络中的基于 Darknet 的全卷积神经网络结构，将图像平均分为 $S_{\times}S$ 个网格，每个网格分别负责预测中心点落在该网格内的目标，这些网格就相当于是两阶段算法的目标的候选区域，通过这种方式 YOLO 算法无需再设计一个 RPN 网络，从而大幅提高了目标检测效率。每个网格预测 N 个边界框和 M 个类别的条件概率。在 YOLOv1 中，N 设置为 2，M 根据使用的数据集确定（如使用 PASCAL VOC 数据集时，M 为 20）。每个边界框包含 5 个预测值：x、y（边界框中心点的坐标）、w、h（边界框的宽和高）以及 Confidence（置信度）。总的来说 $S_{\times}S$ 个网格，网络总输出就是 $S_{\times}S_{\times}$（$5_{\times}N+M$）的张量，然后采用非极大值抑制消除重复检测，最终输出检测结果。在 YOLOv1 中，损失函数主要包括：坐标预测损失、置信度预测损失和类别预测损失，通过最小化这些损失，以优化网络的性能。然而由于 YOLOv1 中的每个网格仅预测了 2 个边界框，并且都属于同一类别，YOLOv1 算法对相互靠近的物体，以及很小的群体检测效果较差。如图 2-8 所示，YOLOv1 的网络结构主要由 24 个卷积层、4 个池化层和 2 个全连接层构成。

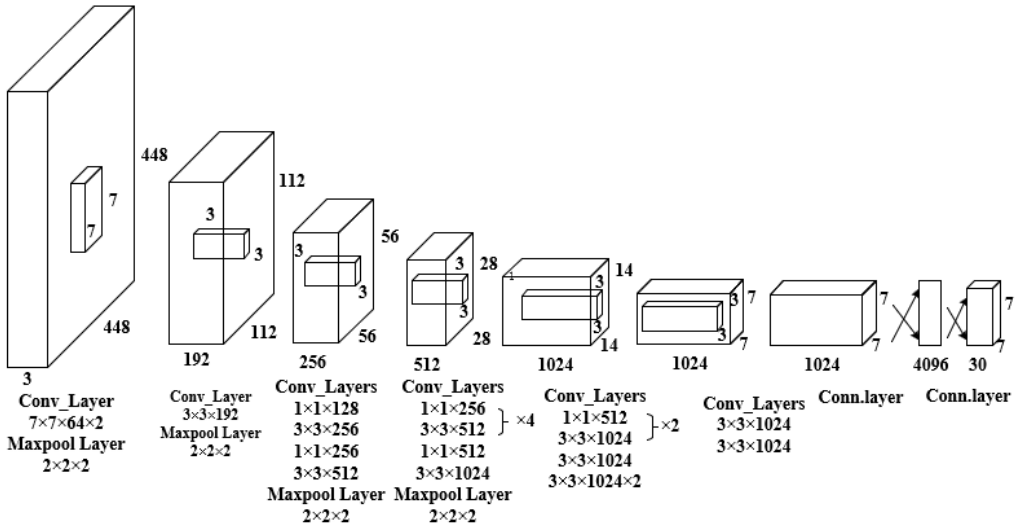


图 2-8 YOLOv1 网络结构

Figure 2-8 YOLOv1 network structure

（2）YOLOv2 和 YOLOv3 算法

为解决 YOLOv1 算法空间信息丢失过多的问题，YOLOv2 算法采用 Darknet-19 作为骨干网络，进一步提升了模型的特征提取能力，并放弃全连接层的设计，在每个卷积层后面都添加了 BN 层，以加快收敛速度。同时 YOLOv2 中引入了锚点框以取代 YOLOv1 中的预测框，提高了网络对不长宽比目标的检测能力。并每隔若干次迭代随机调整输入尺寸（如 320×320 到 608×608，步长 32），增强模型对不同尺度的适应性。

相较于 YOLOv2 中使用的 Darknet-19 网络，YOLOv3 采用了 Darknet-53 作为其骨干网络，Darknet-53 主要由 1×1 和 3×3 的卷积层组成[4]，为了防止过拟合，每个卷积层之后还增加了一个批量归一化层和一个 Leaky ReLU。卷积层、BN 层以及 Leaky ReLU 激活函数共同组成 Darknet-53 的基本 CBL 模块，另外在 YOLOv3 的网络结构中还包含 Res_unit 模块、ResN 模块。YOLOv3 的网络结构如图 2-9 所示。Res_unit 模块是一种常规的残差单元，其将输入特征通过两个 CBL 模块处理后，再与原输入进行拼接。这种设计有效缓解深度网络中的梯度消失问题，并提升特征传递能力。ResX 模块由一个 CBL 层和 X 个 Res_unit 残差单元串联而成，有助于网络提取多尺度的特征信息。YOLOv3 首次引入了多尺度的锚点框检测，分别对应于输入图像的 1/32、1/16 和 1/8 尺度的特征图。例如，在 YOLOv3 网络中输入一张 416×416 像素的彩色图像，首先采用 Darknet53 网络提取目标深层特征，然后在颈部融合不同尺度的特征图，最后检测头从颈部网络

获得多尺度融合后的特征图，并从三个尺度上进行检测，从而生成最终的目标检测结果。

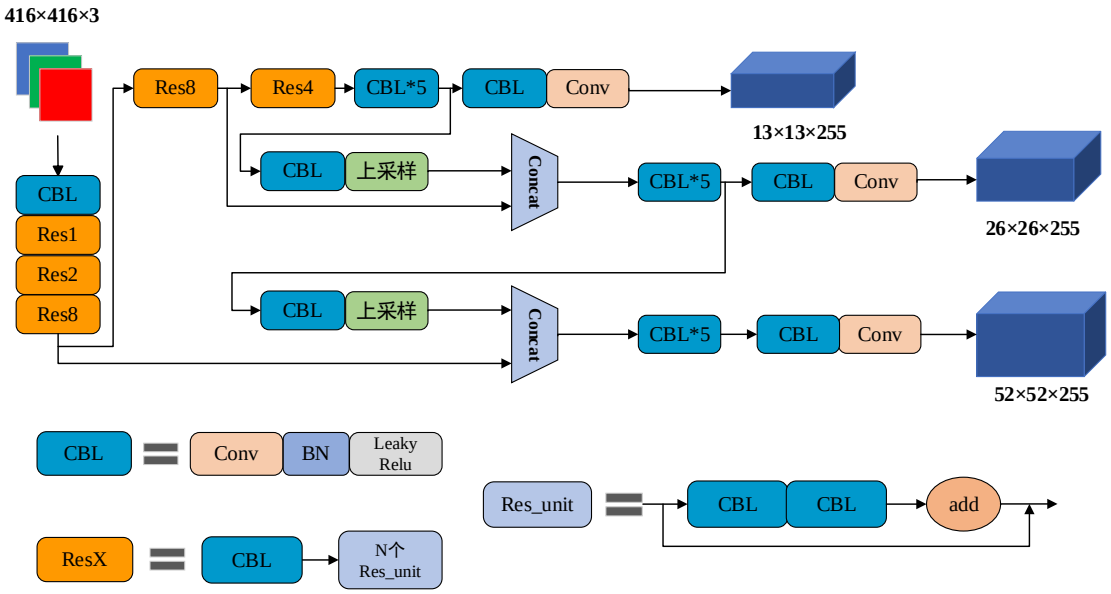


图 2-9 YOLOv3 网络结构

Figure 2-9 YOLOv3 network structure

（3）YOLOv4 和 YOLOv5

在 YOLOv3 算法的基础上，YOLOv4 算法对其网络结构进行了多处优化。在特征提取部分，YOLOv4 引入了 CSPDarknet-53 骨干网络结构，该结构在 Darknet-53 的基础上融入了 CSP 模块，从而增强骨干网络的特征提取能力。并在网络的每个残差模块中都应用了 Mish 激活函数，以在特征转换过程中引入非线性变化，有助于网络更准确地捕获输入数据的复杂特性。在颈部网络部分，YOLOv4 引入了 PANet[42]结构，该结构旨在实现不同尺度特征图之间的信息传递与融合，以获取更好的多尺度特征表示。

YOLOv5 相较于 YOLOv4 在多方面进行了优化和改进，显著提升了模型的目标检测性能。YOLOv5 的网络结构图如图 2-10 所示。YOLOv5 采用 CSPDarknet53 作为骨干网络，并在此基础上融入了 Focus 模块和两种 CSP 模块，以增强特征提取效果。Focus 模块通过切片操作提取特征，有效降低了计算复杂度并减少了原始信息的丢失。其核心思想是将输入特征图进行通道划分和空间划分，在压缩特征图尺寸的同时，保留了其中的重要信息。在网络的末端，采用 SPP 模块来提取不同尺度的上下文信息，这一设计显著提升了模型对于不同尺寸目标的适应性。颈部网络则采用了 FPN 结构和含 CSP 模块的 PAN 结构，通过自上而

下和自下而上的双向特征融合策略，极大增强了特征在不同尺度间的传递和融合。在检测头部分，YOLOv5 则增加了正样本锚框的数量，以加快收敛速度。此外 YOLOv5 采用了 GIOU_Loss 作为边界框损失函数，并使用二进制交叉熵和 Logits 损失函数计算类别概率和目标得分的损失。GIOU_Loss 在 IOU 的基础上引入了预测框和真实框的最小外接矩形，能够更好地反映真实框和预测框的重合程度和远近距离，从而提高了目标检测的准确性。

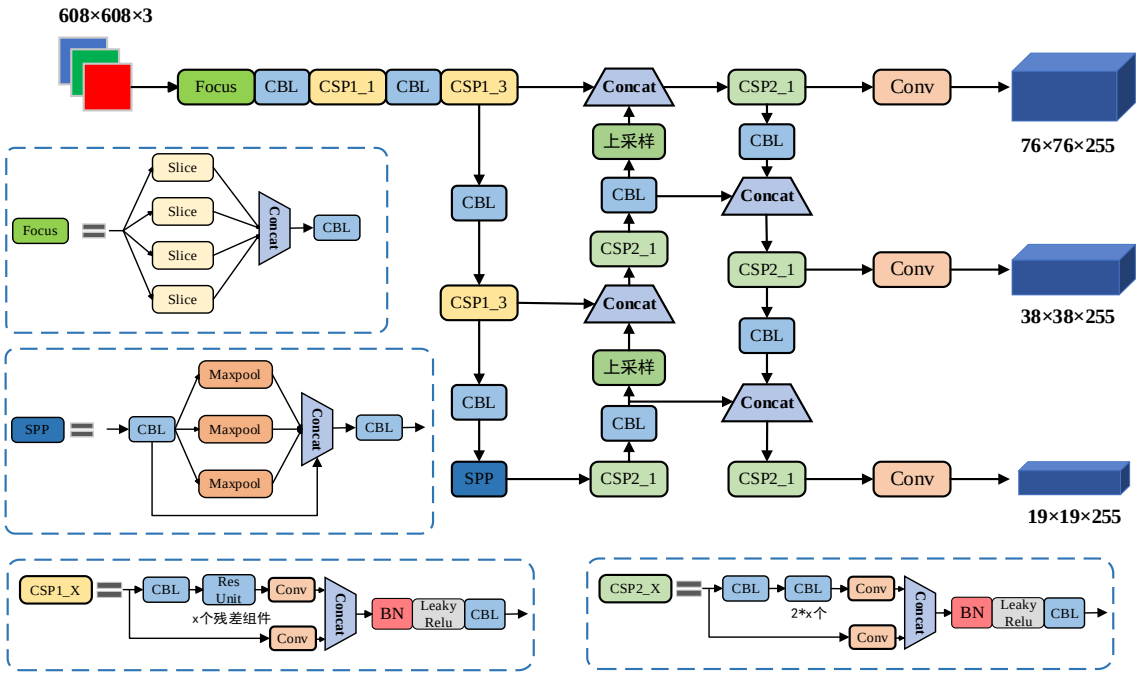


图 2-10 YOLOv5 网络结构

Figure 2-10 YOLOv5 network structure

（4）YOLOv8 行人检测算法原理

YOLOv8 算法根据网络模型大小可以分为 n、m、s、l、x 五种模型，其中 YOLOv8n 的参数量最小、检测速度最快，是最符合实时行人检测需求的模型。YOLOv8n 在继承了 YOLO 系列优点的基础上进行了多项优化，它在目标检测、图像分割、目标跟踪等任务中性能都有显著提升，是当前计算机视觉领域的热门算法之一。其网络结构主要由输入端、骨干网络、颈部网络和检测头四个部分构成。YOLOv8n 的网络结构如图 2-11 所示。

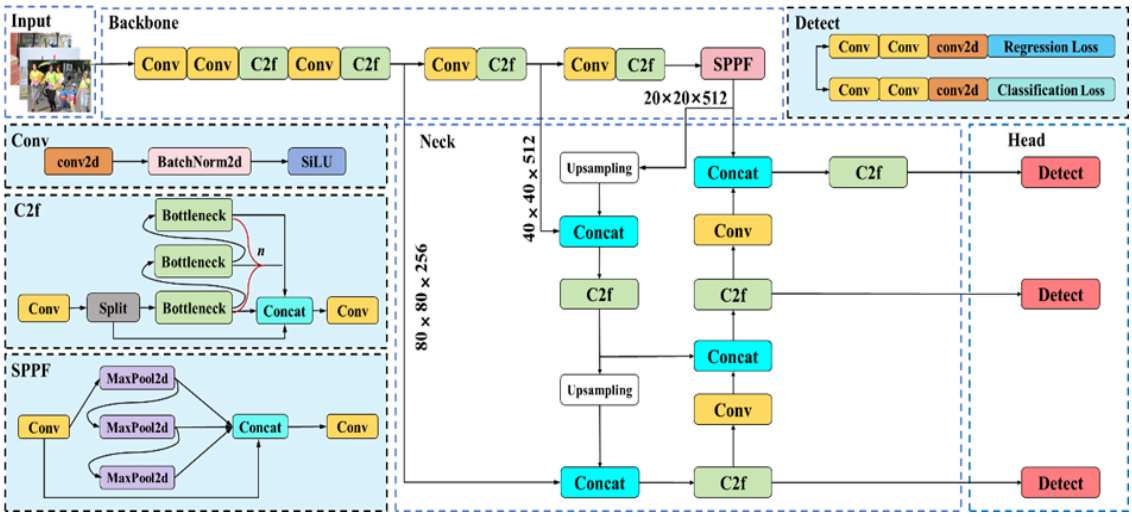


图 2-11 YOLOv8 网络结构图

Figure 2-11 YOLOv8 network structure

输入端：YOLOv8 在输入端采用数据增强策略。YOLOv8 在进行数据预处理时会进行数据增强，主要采用马赛克增强（Mosaic）、混合增强（Mixup）和颜色空间增强（Augment HSV）。马赛克增强是一种将四张图片通过随机缩放、裁剪和排列等操作拼接成一个新的训练样本的方法。与传统方法相比，马赛克增强方法能够同时检测四张图片中的目标，从而增加模型的检测多样性。此外，由于随机缩放和裁剪，马赛克增强能够模拟目标在不同尺度和位置的情况，进一步提升模型的适应性。YOLOv8 算法会在最后 10 个 epoch 关闭马赛克增强，采用 Letter Box Letterbox 技术将图像按原始长宽比缩放，并用灰色边框填充，避免图像失真。

骨干网络：YOLOv8 算法的骨干网络采用了 CSP 思想和快速空间金字塔池化（Spatial Pyramid Pooling – Fast, SPPF）模块等技术，同时使用了残差连接和瓶颈结构来降低模型的体积，提高特征提取能力和计算效率。其骨干网络采用了基于 CSP 结构的网络结构，但进行了多项改进，主要由 CBS 模块、C2f 模块和 SPPF 模块构成。其中最显著的是引入了 C2f 结构，替代了 YOLOv5 中的 C3 结构。C2f 模块的参数量较小且特征提取能力更强。它采用更高效的结构设计，降低了参数量，提高了计算效率。两者的结构差异如图 2-12 所示。C2f 模块由两个 CBS 模块和 n 个 BottleNeck 模块组成，同时引入了更多的跳跃连接结构，并额外增加了分割操作。每个 CBS 模块内部包含一个卷积核大小为 1×1 的卷积层、批归一化层以及 SiLU 激活函数。在特征处理流程上，C2f 模块首先将输入特征图传入第一个 CBS 模块，对特征图的通道数进行调整，使其匹配最终输出所需的通道

维度。然后，采用分割操作将输出结果拆分为两部分，其中一部分输入到 n 个 BottleNeck 模块进行深度特征提取，另一部分则与 BottleNeck 模块处理后的特征图一同拼接，通过多个 Bottleneck 块捕捉更复杂的模式和细节，进一步提炼和增强特征。最后，拼接后的特征图进入第二个 CBS 模块，完成通道降维，使输出特征图达到 C2f 模块预期的通道数，从而确保模型在提取丰富特征的同时，保持较低的计算开销。

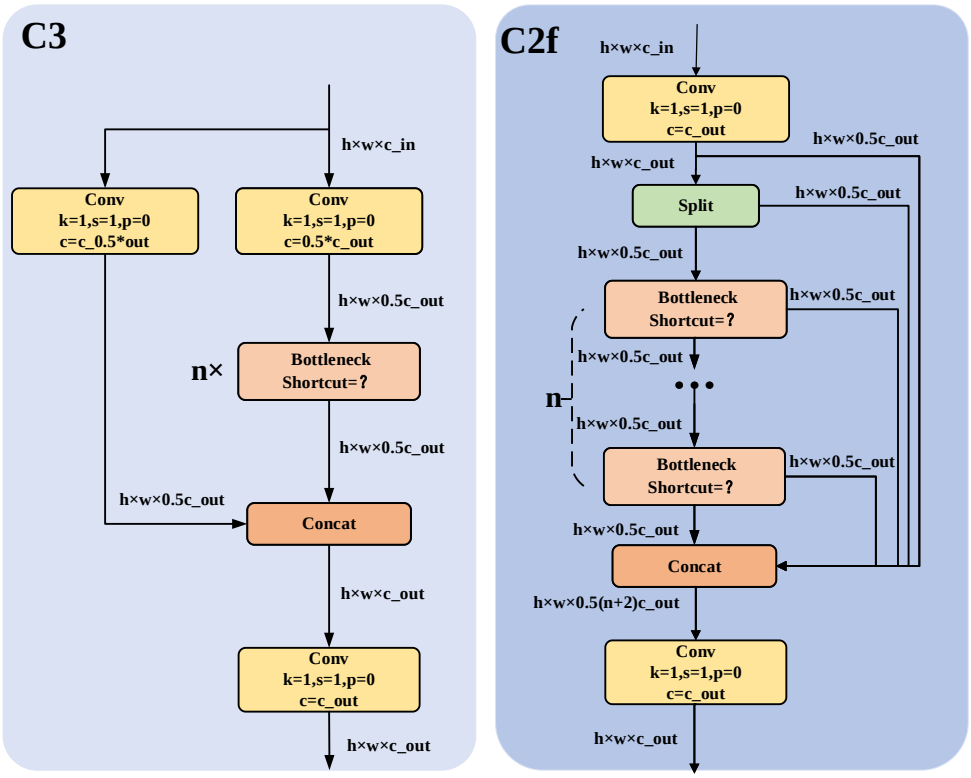


图 2-12 C3 与 C2f 模块结构图

Figure 2-12 Structure of the C3 and C2f modules

在 C2f 模块中，每个 BottleNeck 的输入通道数仅为前一层的一半，显著降低了计算量。同时，梯度流的增加进一步提升了模型的收敛速度和效果。此外，YOLOv8 采用 SPPF 模块作为空间金字塔池化层，通过不同尺度的池化操作来聚合具有不同感受野的特征。这些池化操作的核大小和步长可能不同，例如核大小可能为 5、9、13 等，并且步长通常为 1。最终，这些池化后的特征在通道维度上进行拼接，输出多尺度融合特征，有效扩展了网络的感受野，显著提升了网络对不同尺寸目标的检测性能。

颈部网络：在 YOLOv8 中，Neck 部分沿用了高效的 PAFPN（PAN+FPN，即融合了 FPN 与 PAN 的结构）来构建特征金字塔网络。PAN-FPN 结构通过自顶向

下和自底向上的路径融合不同尺度的特征图。FPN 从最高层的特征图开始，逐层上采样，并与相应的低层特征融合。这种融合通过横向连接，利用 1×1 卷积调整通道数，使上采样特征与低层特征的通道数一致，随后逐层相加实现特征融合。PAN 在 FPN 结构中加入了自底向上的特征传递路径，用于增强底层特征的高层语义信息。它通过自底向上的路径，从底层特征图开始，逐步向上融合更高层次的特征图。同时，在每一层，将自顶向下路径和自底向上路径的特征进行融合，确保每一层的特征都包含不同尺度的信息。这种 PAN-FPN 结构实现了信息的跨尺度传递，使得特征图包含更丰富的上下文和语义信息，提升了网络的表达能力和收敛速度。

检测头：在 Head 部分，YOLOv5 采用耦合头（Coupled Head）结构，这种结构将目标分类和定位任务进行耦合，共享前一层参数。而 YOLOv8 则使用目前主流的解耦头结构，将分类和边界框定位分成两个独立分支的任务，不再共享前一层参数，使得两者可以独立优化，从而提高了检测精度。耦合头和解耦头结构图如图 2-13 所示。与 YOLOv5 使用的 Anchor-Based 方法不同，YOLOv8 采用了 Anchor-Free 的思想。Anchor-Free 方法无需预设锚框，而是通过直接预测边界框的坐标来实现目标检测。由于 Anchor-Free 检测头不用预定义锚框，进一步增强了模型的鲁棒性，使其可以更好地适应各种数据集和目标。

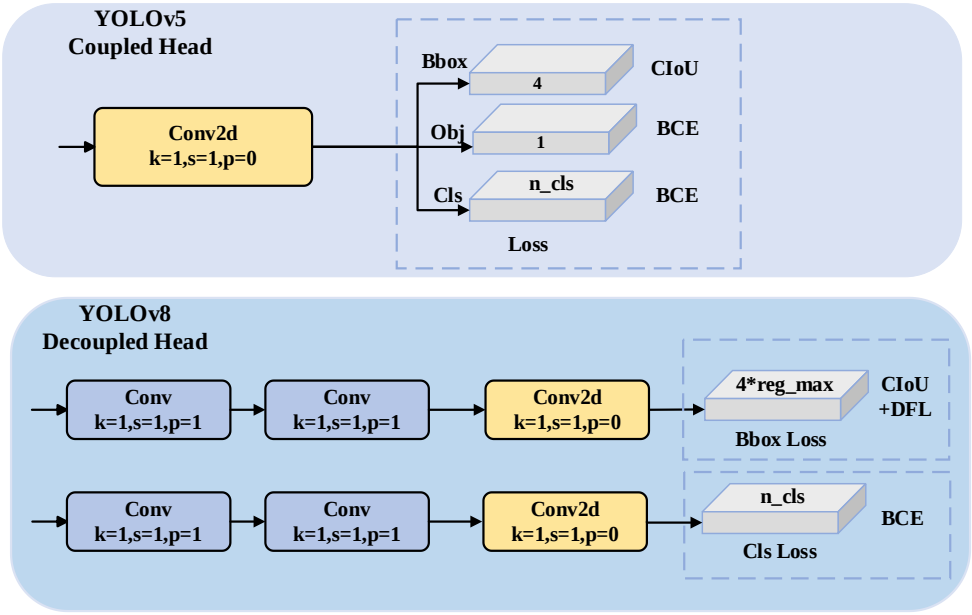


图 2-13 耦合头和解耦头结构图

Figure 2-13.Coupled Head and Decoupled Head structure

在分类损失方面，YOLOv8 使用了 BCE Loss，在回归损失方面 YOLOv8 结合了 DFL Loss 和 CIOU Loss[43]。DFL Loss 用于优化边界框的中心点坐标预测，而 CIOU Loss 根据边界框的宽高比和旋转角度等因素，更精准地评估预测框与实际框之间的差异。YOLOv8 在样本匹配策略上引入了 Task-Aligned Assigner 方法。这种方法可以根据不同任务的需求，自适应地调整正负样本的分配，促使模型在训练过程中能够更关注高质量样本，优化了训练资源的利用率。

2.3 行人跟踪算法理论

基于检测的行人跟踪算法已成为当前行人跟踪领域的主流范式。其基本流程包括三个核心步骤：目标检测、运动预测与数据关联。首先，利用高性能目标检测器（如 Faster R-CNN、YOLO 系列检测器）对输入视频序列的每一帧进行独立的目标检测，生成候选目标的边界框。随后，基于运动建模方法（如卡尔曼滤波）对已跟踪目标的未来状态进行预测，以缓解检测器可能出现的短暂遮挡或丢失问题。最后，通过数据关联策略（如匈牙利匹配算法、级联匹配）在相邻帧之间建立目标的时序关联关系，实现检测框与已有跟踪轨迹的匹配，从而维持目标身份的一致性。此外，为了提升跟踪的稳健性，现代多目标跟踪系统通常结合外观特征与运动特征，并采用多种优化策略（如轨迹管理、重识别模块、贝叶斯滤波）以降低误匹配和目标漂移现象，提高跟踪的准确性和鲁棒性。

2.3.1 卡尔曼滤波

卡尔曼滤波（Kalman Filter, KF）是一种状态估计预测算法，主要应用于目标跟踪和信号处理等领域。该算法基于线性系统的状态空间模型，假设系统噪声和观测噪声均服从高斯分布，并以最小均方误差为最优估计准则，对动态系统的状态进行递归估计。在计算机视觉领域，卡尔曼滤波被广泛用于基于检测的目标跟踪框架，特别是在 SORT 及其改进算法（如 DeepSORT）中发挥关键作用。

卡尔曼滤波的核心思想是利用当前状态的先验估计与观测数据之间的差异进行更新，从而获得更准确的后验状态估计。整个过程由预测和更新两个阶段组成。在预测阶段，卡尔曼滤波利用上一时刻的状态估计值及状态转移模型推测当前时刻的先验状态，并计算预测协方差，以衡量不确定性；在更新阶段，其结合

观测模型和当前测量数据，计算卡尔曼增益，然后利用该增益加权修正预测状态，最终得到后验状态估计。由于其递归计算方式，卡尔曼滤波能够在计算复杂度较低的情况下实时处理系统状态变化，适用于实时目标跟踪任务。卡尔曼滤波预测和更新两个阶段的计算公式如式（2-6）到式（2-10）所示：

（1）预测阶段

基于当前状态和控制输入，预测下一时刻的状态值：

$$\hat{x}_{k|k\text{-}1} = F\hat{x}_{k\text{-}1|k\text{-}1} + Bu_{k\text{-}1} \tag{2-6}$$

其中，$\hat{x}_{k|k\text{-}1}$ 表示时刻 k 的先验状态估计，F 为状态转移矩阵，$\hat{x}_{k\text{-}1|k\text{-}1}$ 为时刻 $k-1$ 的后验状态估计，B 为控制输入模型，u_{k-1} 为控制向量。

基于系统动态模型和过程噪声，预测状态估计的协方差矩阵：

$$P_{k|k\text{-}1} = FP_{k\text{-}1|k\text{-}1}F^T + Q \tag{2-7}$$

其中，$P_{k|k\text{-}1}$ 表示时刻 k 的先验误差协方差矩阵，为时刻 $k-1$ 后验误差协方差矩阵，Q 为过程噪声协方差矩阵。

（2）状态更新阶段

结合观测模型和当前测量数据，计算卡尔曼增益：

$$K_k = P_{k|k\text{-}1}H^T(HP_{k|k\text{-}1}H^T + R)^{-1} \tag{2-8}$$

其中，K_k 为卡尔曼增益，H 为观测矩阵，R 为观测噪声协方差矩阵。

根据实际数据更新状态估计值：

$$\hat{x}_{k|k} = \hat{x}_{k|k\text{-}1} + K_k(z_k - H\hat{x}_{k|k\text{-}1}) \tag{2-9}$$

其中，$\hat{x}_{k|k}$ 为时刻 k 的后验状态估计，Z_k 为时刻 k 的观测值。

更新状态估计的协方差矩阵：

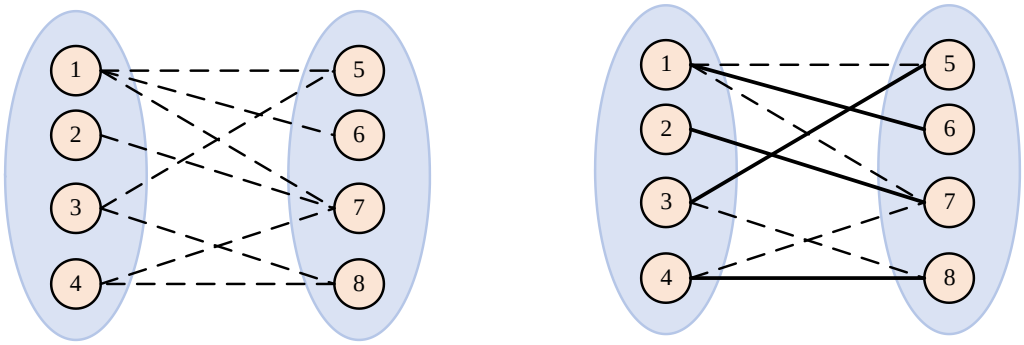
$$P_{k|k} = (I - K_kH)P_{k|k\text{-}1} \tag{2-10}$$

其中，$P_{k|k}$ 为时刻 k 的后验误差协方差矩阵，I 为单位矩阵。通过上述预测和更新步骤，卡尔曼滤波器能够在存在噪声和不确定性的情况下，对目标的运动状态进行精确估计。

2.3.2 匈牙利算法

匈牙利算法（Hungarian algorithm）是在多项式时间内解决指派问题的经典算法，主要用于在给定的成本矩阵中找到最佳的指派方案。在行人跟踪任务中，匈牙利算法被用于将相邻帧中的目标检测框与卡尔曼滤波器预测的跟踪轨迹进行匹配。首先通过构建一个成本矩阵，其中每个元素表示检测框与预测轨迹之间的匹配成本。然后，采用匈牙利算法在该成本矩阵上找到最优匹配方案，以实现检测结果与跟踪轨迹的关联。

匈牙利算法的基本思想是寻找增广路径，并通过增广路径求取二分图的最大匹配，即在两个不相交顶点集之间找到尽可能多的边，使得每个顶点都恰好与一条边相连。这个问题可以转化为一个优化问题，即最大化匹配的权重之和。如图 2-14 所示为典型的二分图，二分图两侧的节点分别代表目标检测框和跟踪轨迹。则匈牙利算法的匹配流程如下：



（a）可匹配的连接图　　（b）最终匹配连接图

图 2-14 匈牙利算法流程

Figure 2-14 Hungarian algorithm flow

（1）首先对左侧的目标节点 1 进行匹配，由图可见节点 1 可以和轨迹节点 5、6、7 匹配，假设节点 1 与节点 5 进行匹配。

（2）由于目标节点 2 仅能和轨迹 7 匹配，接着进行目标 3 的匹配，目标 3 可以和轨迹 5 和 8 匹配，第一步目标 1 已经和轨迹 5 匹配，则将目标 3 与轨迹 8 匹配。

（3）此时，目标节点 4 则没有可以进行匹配的轨迹，为了保证最大化匹配，因此要修改节点 1 的匹配结果，将节点 1 与节点 6 进行匹配。

（4）接着将节点 3 与节点 5 进行匹配，最后节点 4 可以和节点 8 进行匹配，

达到最大化匹配结果。

2.3.3 SORT 系列算法原理

（1）SORT 算法

SORT 算法是一种高效的目标跟踪方法，其核心思想是将目标检测与运动预测相结合，利用卡尔曼滤波器对目标的运动状态进行预测，并通过匈牙利算法将预测结果与当前帧的检测结果进行关联，从而实现目标的持续跟踪。SORT 跟踪算法的流程如图 2-15 所示，Sort 跟踪算法的流程主要分为三个步骤。

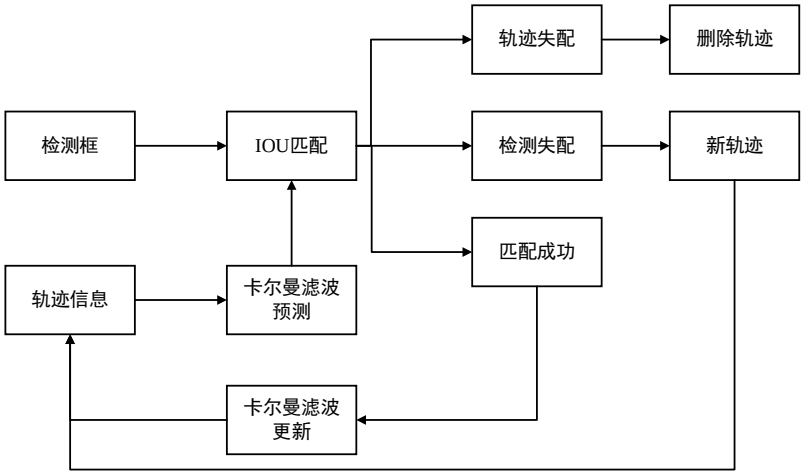


图 2-15 Sort 算法流程

Figure 2-15 Sort algorithm flow

步骤一：采用目标检测模型获取当前帧的检测结果，将第一帧所检测到的目标框初始化为对应的轨迹（Tracks），将所有轨迹储存在 Tracking 列表中，并使用卡尔曼滤波算法预测下一帧目标的位置。

步骤二：在后续的每一帧中，首先使用卡尔曼滤波器预测轨迹的位置；然后，计算当前帧检测结果与预测轨迹之间的 IoU，并基于 IoU 构建代价矩阵；最后，使用匈牙利算法对代价矩阵进行优化求解，实现检测结果与预测轨迹的最佳匹配，并为每个成功跟踪的目标分配一个唯一的 ID。

步骤三：匹配成功后更新目标的轨迹，即将预测框更新为目标框位置的值，然后进行下一帧的预测。若轨迹匹配失败，则存在两种情况。第一种是轨迹失配，说明预测框与真实值偏差较大，将直接删除该轨迹。第二种是检测失配，此时会为其创建新的跟踪轨迹，并进行验证以排除误检。

但是由于 SORT 算法仅使用卡尔曼滤波和匈牙利算法来进行目标跟踪，过于

依赖目标的运动特征来进行轨迹匹配和关联，忽略了目标的外观特征，在遮挡和目标暂时消失的场景下,SORT 算法会出现 ID 频繁切换或目标丢失的问题，导致跟踪效果较差。

（2）DeepSORT 算法

DeepSORT 跟踪算法是在 SORT 算法的基础上进行改进的，该算法的核心思想仍是递归的卡尔曼滤波器和逐帧的匈牙利数据关联。同时为解决 SORT 算法在行人遮挡和暂时消失场景下，跟踪效果差的问题，DeepSORT 算法进行了三处改进。首先引入外观特征，通过在行人重识别数据上预训练的卷积神经网络提取每个检测目标的外观特征向量。这些特征在后续帧中用于计算相似度，有助于在遮挡或相似目标出现时保持一致的 ID 分配；然后在数据关联过程中，DeepSORT 结合了运动信息和外观特征，采用马氏距离度量目标状态预测与实际检测之间的差异，同时计算外观特征之间的余弦相似度。这种多维度的关联策略提高了目标匹配的准确性，减少了 ID 切换的发生；最后 DeepSORT 对卡尔曼滤波器进行了改进，引入了自适应噪声协方差计算和运动一致性约束，以提高在复杂场景下的预测精度和鲁棒性。DeepSORT 算法流程如图 2-16 所示，其流程可分为六个步骤：

步骤 1：使用行人检测模型对当前帧的行人目标进行检测，生成目标的边界框及其置信度分数。并将检测到的每个目标初始化为其相对应的跟踪轨迹，此时轨迹处于未确认状态。通过卡尔曼滤波算法对当前帧的跟踪轨迹进行预测，得到跟踪预测框。

步骤 2：计算这些轨迹预测框与当前帧目标检测框之间的 IoU，构建余弦距离矩阵，然后将余弦距离矩阵作为匈牙利算法的输入，得到线性匹配的结果。匹配结果分为三种情况：第一种情况是未匹配的检测框，将创建新的轨迹以追踪这些新目标。新轨迹初始时处于未确认状态；第二种情况是未匹配的跟踪轨迹，若轨迹处于确认状态，则根据其连续丢失帧数（默认为 30 帧）决定是否删除这些轨迹，以避免追踪无效或已离开视野的目标。若轨迹处于未确认状态，则直接删除；第三种情况是轨迹匹配成功，则使用检测结果更新对应轨迹的卡尔曼滤波器状态，实现轨迹的连续追踪。若轨迹处于未确认状态，则连续匹配一定次数（默认为 3 次）后，将其转换为确认状态。

步骤 3：重复前面两个步骤，直至出现已确认态的跟踪轨迹或视频结束。

步骤 4：通过卡尔曼滤波预测其确认态的轨迹和未确认态的轨迹对应框。然后通过运动信息和外观特征信息将确认态的预测框和检测框进行级联匹配。

步骤 5：经过级联匹配得到三种结果，第一种情况是跟踪轨迹成功匹配。采用卡尔曼滤波更新其对应的跟踪轨迹状态。第二种和第三种情况分别是跟踪轨迹失配和检测框失配，则将之前不确认状态的轨迹和失配的轨迹一起和未匹配的检测结果进行 IoU 匹配，并计算其代价矩阵。然后重复步骤 2 进行线性匹配。

步骤 6：重复上述步骤，直至视频帧结束。

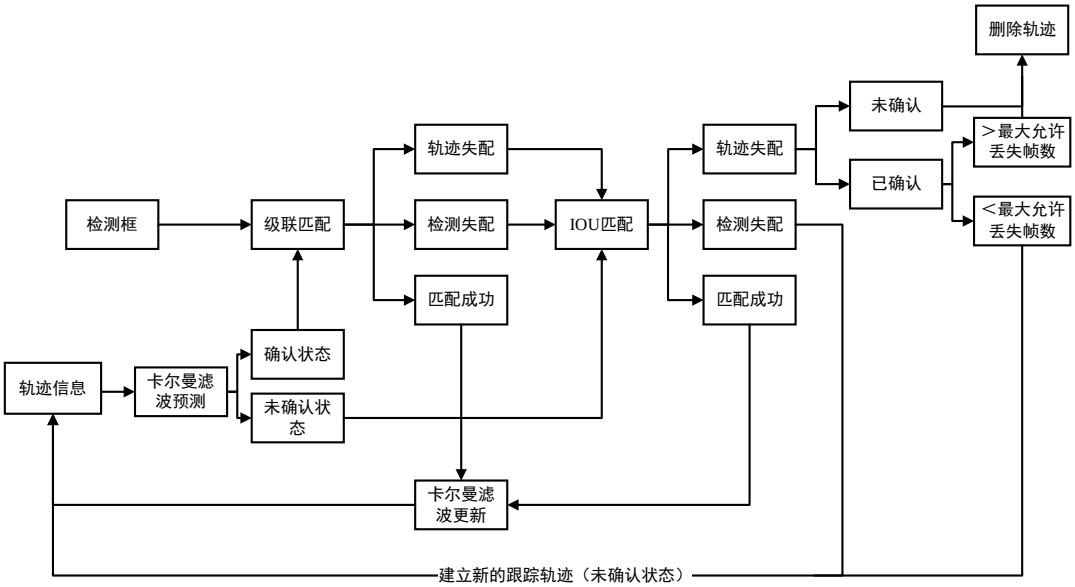


图 2-16 DeepSORT 算法流程

Figure 2-16 DeepSORT algorithm flow

（3）StrongSORT 行人跟踪算法

StrongSORT 跟踪算法由 DeepSORT 算法改进得来，算法流程大致相同，其基本网络结构如图 2-17 所示。StrongSORT 算法相对于 DeepSORT 算法的优化，主要体现在以下三个方面：在外观特征提取方面，StrongSORT 算法采用了更强大的外观特征提取网络，其主要由 ResNet50[44]网络构建的 BoT[45]特征提取器组成，这种设计能够更有效地提取目标的细节和纹理信息，提高跟踪模型特征的表达能力，并引入 EMA 特征更新策略来替换 DeepSORT 算法中的特征信息库；在运动模型方面，StrongSORT 算法首先引入 ECC 算法来进行相机运动的补偿，其次使用 NSA 卡尔曼滤波算法，动态调整噪声协方差，提升运动预测的准确性；在匹配策略方面上，StrongSORT 算法采用全局外观关联补偿局部遮挡或漏检导

致的轨迹中断，减少 ID 切换。

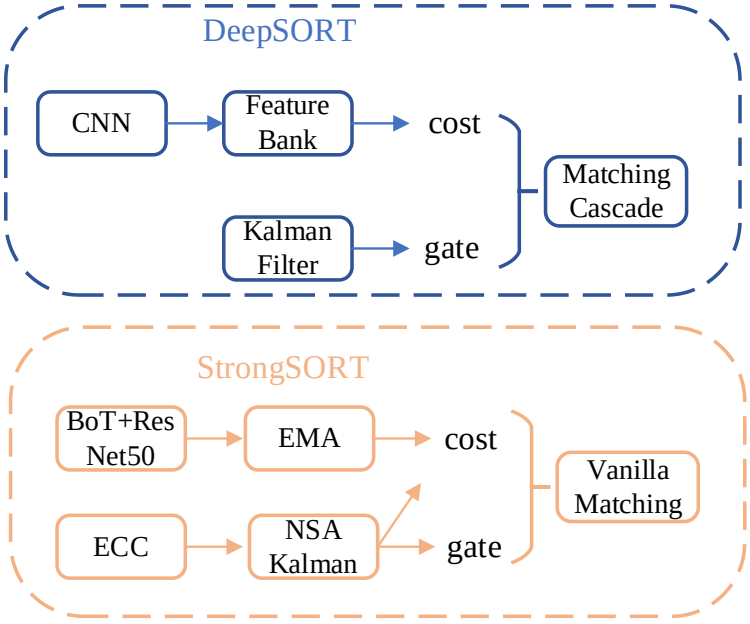


图 2-17 StrongSORT 算法流程图

Figure 2-17 Flowchart of the StrongSORT algorithm

EMA 特征更新策略：虽然 DeepSORT 中的特征库可以保留长期的特征信息，但是保留 100 帧的特征使得跟踪效率低下，对每一帧的检测噪声高度敏感。因此，StrongSORT 采用 EMA 特征更新策略，将过去的特征作为惯性项保留并更新，并以指数移动平均值的方式更新轨迹 i 在 t 帧时刻的表现，通过用 α 加权，有效地更新特征信息，并减少了参数更新的噪音和波动。原理如公式（2-11）所示。

$$e_i^t = ae_i^{t-1} + (1-a) f_i^t \tag{2-11}$$

其中 e_i^t 表示轨迹在 t 帧时刻的外观状态，a 表示动量项，f_i^t 表示轨迹 i 在 t 帧时刻的外观嵌入。

ECC 运动补偿：ECC 算法是一种常用于图像处理和计算机视觉领域的算法，其处理存在光照变化或噪声的图像对齐问题时表现出色。ECC 算法的核心思想是通过找到最佳的几何变换参数，使得两幅图像之间的增强相关系数最大化。这种相关系数对光度失真具有不变性，因此即使在光照条件变化的情况下，也能实现准确的图像对齐。增强相关系数是 ECC 算法的关键部分，主要用于衡量两幅图像之间的相似性。增强相关系数的计算公式如式（2-12）所示，通过这该操作，能够准确量化图像之间的相似度，并基于此实现运动补偿，从而有效提升跟踪算法的精度和稳定性。

$$E_{ECC}(p)=\frac{\bar{i}_r}{\|\bar{i}_r\|}-\frac{\bar{i}_w(p)}{\|\bar{i}_r(p)\|}^2 \tag{2-12}$$

其中，$\|\cdot\|$表示欧几里得范数，p 是形变参数，$\bar{i}_r$ 和$\bar{i}_w(p)$分别是参考（模板）图像i_r和形变图像$i_w(p)$ 的零均值版本。然后，通过最小化增强相关系数函数$E_{ECC}(p)$来解决图像对齐问题。

NSA 卡尔曼滤波：NSA Kalman 算法是基于传统卡尔曼滤波算法的一种优化版本。NSA Kalman 算法的核心思想是在传统卡尔曼滤波的基础上，根据目标检测的质量自适应地调制噪声尺度。在传统卡尔曼滤波中，噪声尺度是一个常数矩阵。但不同的检测方法可能包含不同的噪声矩阵。因此，NSA Kalman 算法通过引入一个自适应计算噪声协方差的公式，能够根据检测的置信度动态地调整噪声尺度，从而提高状态估计的准确性。具体过程如公式（2-13）所示。

$$\tilde{R}_k=(1-c_k)R_k \tag{2-13}$$

其中$\tilde{R}_k$是预设的常数测量噪声协方差，c_k是状态t时刻的检测置信度分数。

全局线性匹配策略：在级联匹配方面，DeepSORT 仍存在不足之处。首先 DeepSORT 算法在匹配优先级上仅根据跟踪器的未匹配时间来进行分配，长时间未匹配的轨迹优先匹配，这种策略容易忽略轨迹的置信度或其他上下文信息。而 StongSORT 算法采用动态优先级调整策略，结合轨迹的置信度、历史匹配质量等动态调整优先级，避免单纯依赖未匹配时间导致的误匹配。同时进行多目标优化，在遮挡场景中，优先匹配高置信度的检测结果，减少低质量检测对匹配的干扰。

其次，DeepSORT 算法仅使用运动（马氏距离）和外观（余弦距离）的加权和作为匹配代价，权重固定，无法适应动态场景。为此，StongSORT 算法中根据场景动态调整运动与外观的权重，同时使用外观和运动信息来解决分配问题。原理如公式（2-14）所示。

$$C=\lambda A_a+(1-\lambda)A_m \tag{2-14}$$

其中C表示总的代价矩阵，λ表示权重因子，A_a表示外观代价矩阵，A_m表示运动代价矩阵。

2.4 本章小结

本章主要介绍了卷积神经网络、行人检测和行人跟踪的基本原理。第一部分主要介绍了卷积神经网络构成及原理；第二部分阐述了 YOLO 系列的算法结构和相关原理，并详细介绍了 YOLOv8 算法的网络结构与优缺点；第三部分主要介绍了行人跟踪算法的原理，对卡尔曼滤波算法，匈牙利匹配算法的基本原理进行介绍与分析，同时对 SORT 系列算法的流程进行阐述。为后续行人检测与跟踪模型的优化研究提供了全面的理论支撑。

第 3 章 基于可变形卷积和任务对齐策略的行人检测模型

3.1 引言

行人检测是计算机视觉领域的研究热点之一，在智能安防、智慧交通、人群分析等多个领域具有广泛应用。但是，由于监控视频场景下的背景信息复杂，行人特征尺度多样化等问题使得现有算法的检测精度较低，漏检率过高。因此，如何在复杂监控场景中精准检测出多尺度的行人目标，已成为行人检测技术研究的重要课题。2019 年，Liu 等[46]采用轻量化网络 MobileNetV2 作为 SSD 模型中的基础网络，并引入空洞卷积，提升了行人检测的精度和速度。同年，Ge 等[47]在 YOLOv3 算法的基础上引入了改进的 K-means 聚类算法并将小目标浅层特征信息进行多尺度融合，提高了小目标行人的检测效果。2022 年，Chen 等[48]提出了一种基于 YOLOv5 的改进行人检测算法，引入了 Res2Block 和坐标注意力[49]（Coordinate Attention, CA），扩大了模型感受野，并使用结构重参数化的方法减少模型的计算量与参数量，但对复杂场景下遮挡行人检测效果较差。Xu 等[50]人采用 Rep-ELAN-W 模块对 YOLOv7 算法进行改进，提取中低维特征图中的小目标特征信息，使模型可以更好的应用于拥挤行人检测场景，但模型体积过于冗余。

针对上述算法的不足，本章选用 YOLOv8 算法作为基准模型，从增强模型特征提取能力与动态任务交互方面进行研究，以提高对监控场景下的多尺度行人目标的检测效果。

3.2 多尺度行人检测模型

在复杂监控场景中，YOLOv8n 算法仍存在以下问题。首先，其网络模型中的标准卷积无法有效的提取不同尺寸的行人特征。其次，复杂场景下的背景特征对行人检测造成了很大的干扰，同时由于行人目标特征信息相似，容易出现错检漏检的情况，导致检测精度下降。

为解决上述问题，本文提出了一种结合可变形卷积、全局注意力机制和动态任务对齐检测头的多尺度行人检测算法 YOLOv8n-DGT。该算法通过在 YOLOv8 骨干网络和颈部网络引入 C2f-DCNv2 模块，增强特征提取网络的学习能力，提

升模型在对复杂检测场景中对不同尺寸目标的特征提取能力。并在骨干网络的 SPPF 模块后增加了全局注意力机制层，通过全局注意力降低背景信息的干扰，减少特征损失同时促进特征交互，改善网络对小尺寸行人的检测效果。最后设计了轻量化动态任务对齐检测头替换了原 YOLOv8 模型的检测头，运用任务对齐和动态特征选择的思想，在压缩模型参数量的同时促进定位特征和分类特征的交互，提高模型检测性能。图 3-1 展示了改进后的 YOLOv8n-DGT 模型的网络结构。

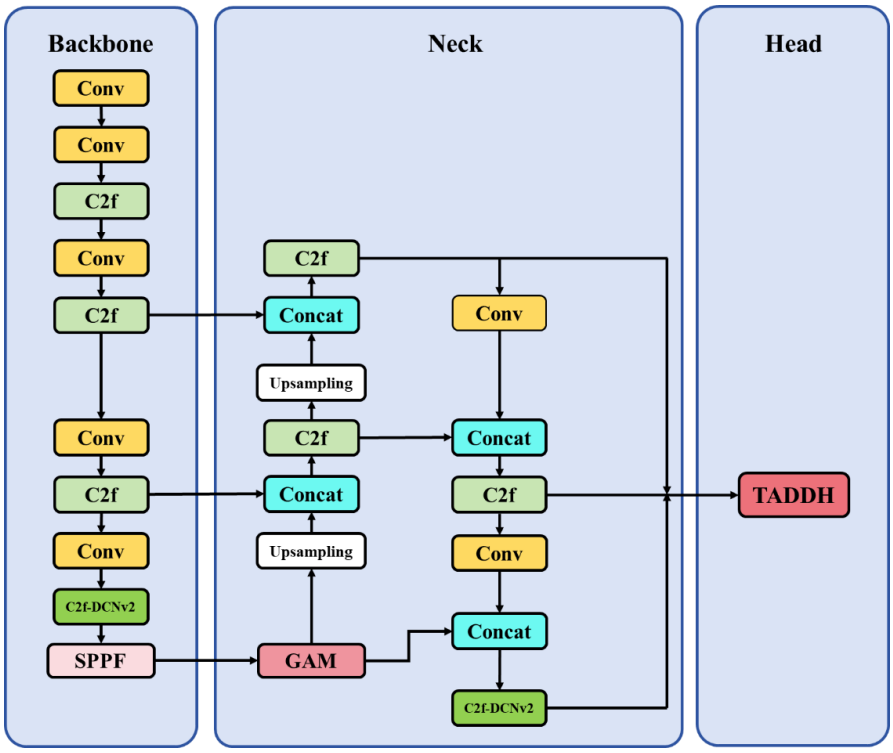


Figure 3-1 YOLOv8n-DGT network structure

3.2.1 结合可变形卷积的 C2f 模块

传统卷积的固定结构在单一场景下具有一定的特征提取能力，然而在复杂场景下，行人目标的尺寸变化较大且位姿多样，导致同一类别的目标存在较大的特征差。因此使用传统的卷积操作无法对不同尺寸的目标都起到很好的效果，而可变形卷积可以根据目标的尺寸和比例自适应并调整，通过不规则的卷积核提取目标特征来适应行人目标动态尺寸的变化。因此本章引入可变形卷网络[51]

（Deformable ConvNets v2, DCNv2）的思想，采用可变形卷积重新构建了 YOLOv8 模型中的 C2f 模块来提升网络的检测能力。

在标准卷积中，每个卷积的采样点无法根据目标大小进行改变，因此不能适应行人目标尺寸的变化。而可变形卷积中则在卷积过程中引入了一个额外的偏移量，根据学习到的偏移量，对卷积核的采样位置进行调整。流程如图 3-2 所示。通过传统卷积对输入图像进行特征提取，生成偏移量和调制量，然后通过偏移量将规则分布的 3×3 像素点进行偏移，并采用偏移后的不规则像素点对特征图进行采样。由于偏移量的存在，卷积核的采样位置可以随输入特征图尺寸的变化而变化，从而使卷积核在提取行人特征时更加贴合目标大小。

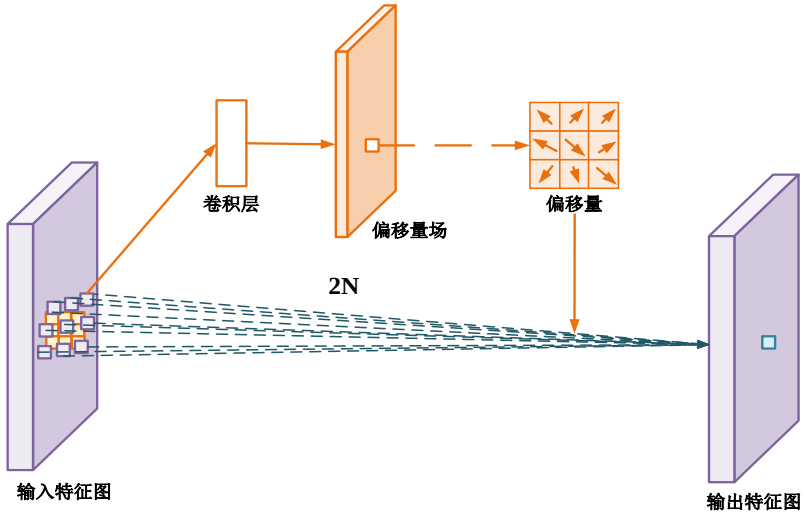


图 3-2 可变形卷积示意图

Figure 3-2 Schematic of deformable convolution

假设输入特征图为 $\mathrm{I} \in \mathrm{R}^{C_{in} \times H \times W}$，则 p_0 位置的标准卷积 $y(p_0)$ 定义如下：

$$y(p_0) = \sum_{p_n \in R} \omega(p_n) \cdot x(p_0 + p_k) \tag{3-1}$$

其中，ω 为采样点权重；p_k 为 p_0 点的偏置项。同时 DCNv2 为了进一步提高对不同尺寸目标的拟合能力，在每个采样点添加权重值，输出特征值 $y(p_0)$ 定义如下:

$$y(p_0) = \sum_{p_n \in R} w(p_k) \cdot x(p_0 + p_k + \Delta p_k) \cdot \Delta m_k \tag{3-2}$$

其中，Δp_k 为输入特征图的偏移量。引入权重系数 Δm_k，对复杂的背景信息

进行过滤，降低模型在提取目标特征时无关背景信息的影响。

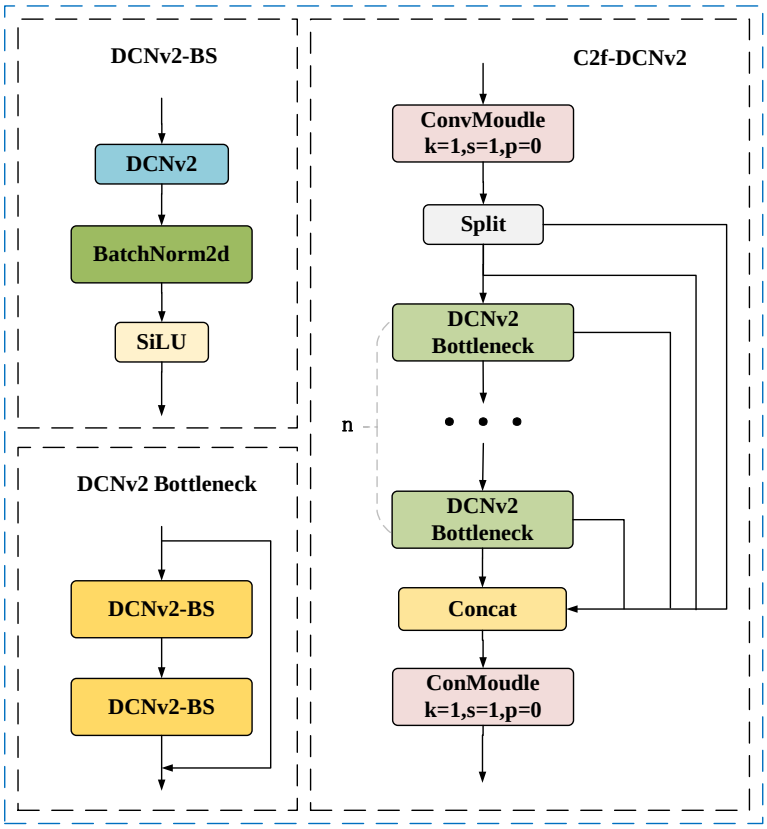


图 3-3 C2f-DCNv2 结构

Figure 3-3 C2f-DCNv2 structure

根据可变形卷积构建的 C2f-DCNv2 模块结构如图 3-3 所示。首先采用 Deformable Conv 替换原本网络中 CBS 模块的标准卷积，构建了 DCNv2-BS 模块。然后采用两个 DCNv2-BS 模块顺序连接并对输入特征进行拼接构建了 DCNv2 Bottleneck 模块，通过该设计使模块可以更好的对输入的不同尺寸特征图进行采样。最后，使用 DCNv2 Bottleneck 替换 C2f 结构中 Bottleneck 结构，构建了 C2f-DCNv2 模块。并使用 C2f-DCNv2 模块替换了原模型骨干网络和颈部网络的第四个 C2f 模块，扩大模感受野，增加采样范围，便于模型提取多尺度特征的上下文信息。使改进后的模型能自适应学习各种尺寸的行人特征，提高模型对目标物体的尺度、背景、形变等信息的学习能力。

3.2.2 全局注意力机制

在行人检测场景中，复杂的背景信息居多，而所检测小目标行人特征过少。为了降低背景信息的干扰，捕捉更全面的行人特征，本章在原 YOLOv8n 网络结

构中增加了一层全局注意力机制[52]（Global Attention Mechanism, GAM），GAM 可以稳定地提高不同网络结构和深度的卷积神经网络的性能，可以有效的降低复杂背景信息的干扰，具有良好的鲁棒性。GAM 在 CBAM[53]（Convolutional Block Attention Module）的基础上进行优化，它沿用了 CBAM 中的顺序通道－空间注意机制的思想，并对空间注意力和通道注意力模块进行了改进，促进了空间注意力和通道注意力的交互，可以减少特征损失并且放大全局维度的交互特征。GAM 模块结构图如图 3-4 所示。

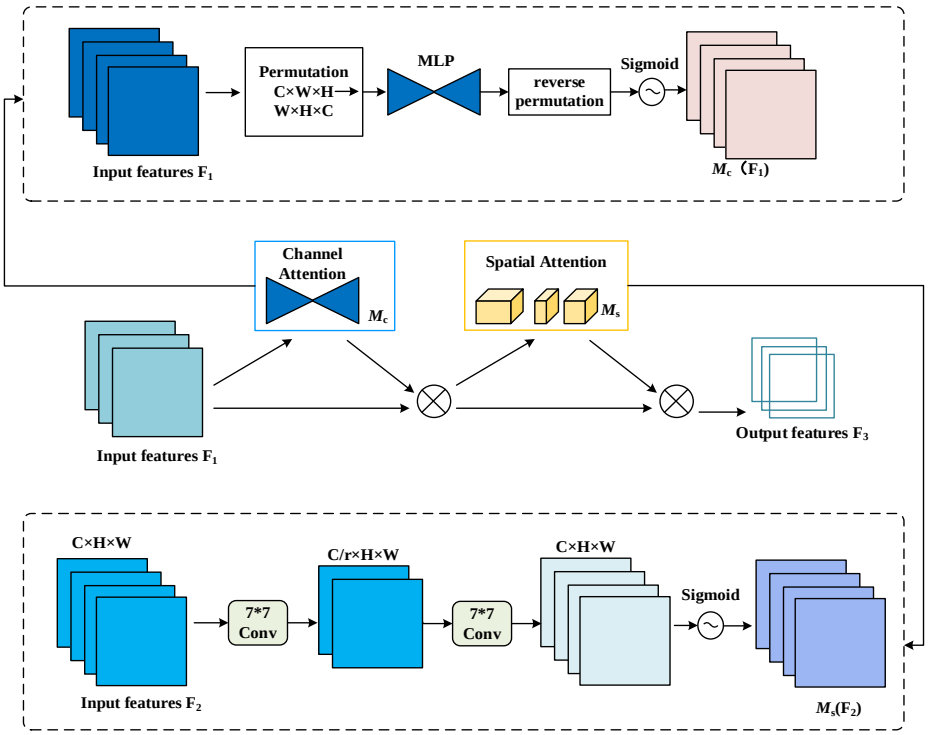


图 3-4 全局注意力机制

Figure 3-4 Global Attention Mechanism

GAM 模块先通过通道注意力模块对原始特征图 F_1 进行校正，得到中间状态 F_2，然后使用空间注意力模块再次校正，得到最终特征图 F_3。以此实现对输入特征图在通道和空间两个维度上的精细处理。

通道注意力子模块首先对输入特征图进行维度转换，然后将转换后的特征图输入多层感知器（Multilayer perceptron, MLP），促进特征向量在不同维度之中进行交互，以此来放大跨维通道-空间依赖性。最后，将 MLP 的输出转换回原始维度，并通过 Sigmoid 函数进行归一化处理，得到通道注意力权重图 M_c。中间状态 F_2 定义如式（3-3）所示：

$$F_2 = M_c(F_1) \otimes F_1 \tag{3-3}$$

为了更好的捕获特征图的空间关系，M_s 空间注意力子模块设计了两个 7×7 的卷积层对空间信息进行融合，并生成空间注意力权重图。并且该模块中通过删去池化层来减少特征信息的损失，增强交互特征的提取能力。中间状态 F_3 定义如式（3-4）所示：

$$F_3 = M_s(F_2) \otimes F_2 \tag{3-4}$$

3.2.3 动态任务对齐检测头

YOLOv8 原有的检测头模型头部使用独立的分类和定位分支,这会导致两个任务之间缺乏交互。同时 YOLOv8 原有头部的缺乏动态学习能力，对多尺度目标检测效果较差。为进一步促进模型分类和定位分支之间的交互，降低模型参数量，提高模型检测的性能，本章设计了一种动态任务对齐检测头（Task Align Dynamic Detection Head, TADDH，结构如图 3-5 所示），并使用该检测头替换原模型的检测头。TADDH 模块借鉴了 Task-aligned head[54]和 Dynamic Head[55]的设计思想，在检测头上设计了任务分解和动态特征选择的结构，利用特征提取器从多个卷积层中学习任务交互特征，促进分类分支和定位分支的交互。通过 Deformable ConvNets v2 自适应学习不同尺寸的目标，从而更好地生成定位分支，同时使用交互特征经过动态特征选择后生成分类分支，提升了模型对交互特征的定位能力和动态学习能力。

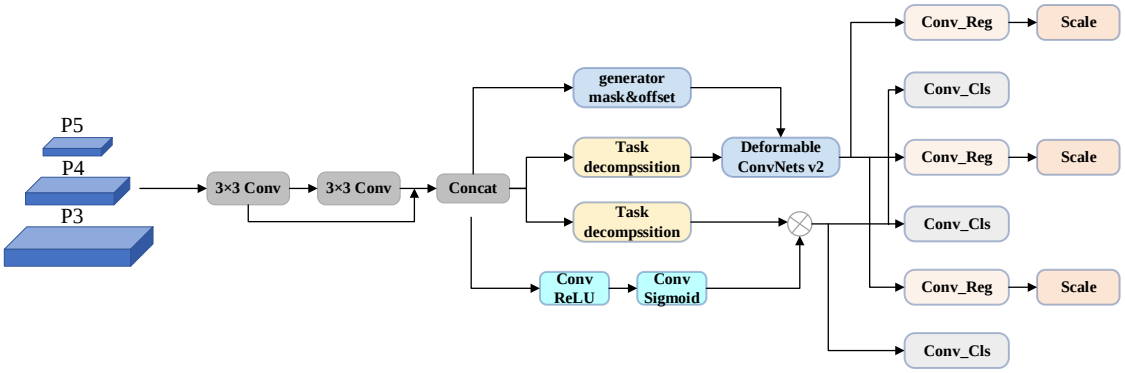


图 3-5 动态任务对齐检测头

Figure 3-5 Task Align Dynamic Detection Head

TADDH 模块首先通过两个卷积核为 3*3 的共享卷积处理 P3、P4 和 P5 的特征图得到拼接后的特征图，通过组归一化的共享卷积，提升检测头定位和分类的

性能，减少模型参数量。然后将通道链接后的特征图输入任务分解模块（结构如图 3-6 所示）中，通过任务分解模块中的特征提取器采用 N 个带激活函数的连续卷积层来计算任务交互特征 X_k^{inter}，计算过程如式（3-5）所示:

$$X_k^{inter}=\begin{cases}\delta\left(conv_k\left(X^{fpn}\right)\right), k=1\\ \delta\left(conv_k\left(X^{inter}\right)\right), k>1\end{cases}, \forall k\in\{1,2,\ldots,N\} \tag{3-5}$$

其中，$X^{fpn}\in R^{H\times W\times C}$ 表示 FPN 特征，$conv_k$ 和 δ 分别代表第 k 个 conv 层和 ReLU 函数。

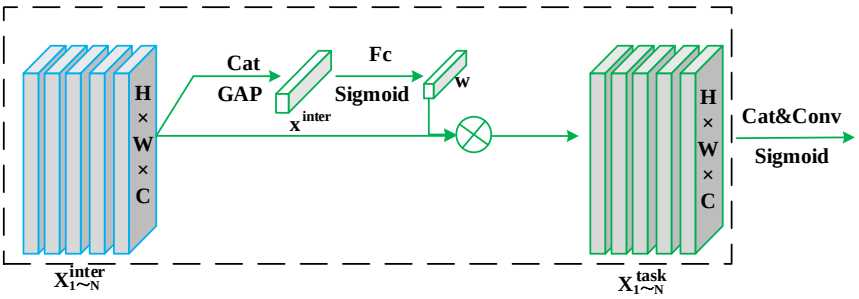


图 3-6 任务分解模块结构

Figure 3-6 Task Decomposition Module Structure

然而，由于原 YOLOv8 模型检测头分支的单一设计，任务交互特征会不可避免的造成两个不同任务之间的特征冲突。因此，任务分解模块中采用层注意力机制（Layer Attention Mechanism, LAM），通过层注意力动态计算这些特定任务特征来促进任务分解。针对分类或定位的每个任务，分别计算特定任务特征，X_k^{task} 的定义如式（3-6）所示：

$$X_k^{task}=\boldsymbol{w}_k\cdot X_k^{inter}, \forall k\in\{1,2,\ldots,N\} \tag{3-6}$$

其中 $\boldsymbol{w}_k$ 是层注意力机制的 $\boldsymbol{w}\in R^n$ 的第 k 个元素，w 是根据跨层任务交互特征计算得出的，能够捕捉到层与层之间的特征依赖关系，公式如式（3-7）所示：

$$w=\sigma\left(fc_2\left(\delta\left(fc_1\left(\boldsymbol{x}^{inter}\right)\right)\right)\right) \tag{3-7}$$

其中 fc_1 和 fc_2 指的是两个完全连接的层。σ 是一个 sigmoid 函数，$\boldsymbol{x}^{inter}$ 是通过对 $\boldsymbol{X}^{inter}$ 应用平均池化得到的，这是 X_k^{inter} 的连接特征。最后，从每个 X^{task} 中预测分类或定位的结果。同时使用拆解前的交互特征生成 DCNv2 所需的 mask 和 offset，然后通过可变形卷积网络生成定位分支，提高了检测头对特征的定位能

力。并且基于 Dynamic Head 尺度感知注意力策略，使用任务拆解前的交互特征通过两个损失函数分别为 ReLU 和 Sigmoid 的卷积层进行动态特征选择，进而生成分类分支。最后为了应对每个检测头所检测的目标尺度不一致的问题，使用 Scale 层对特征图进行缩放。其中共享卷积至生成定位分支和回归分支之间的模块均为共享模块，以此压缩模型体积，降低计算复杂度。

3.3 行人检测实验结果分析

3.3.1 数据集介绍与实验环境

本文的数据集选自开源的 Pascal VOC2012 数据集[56]。该数据集为计算机视觉 PASCAL VOC 挑战赛的公开数据集，其中包含 20 种不同的目标类别。从数据集中选取包含各种复杂场景下行人类别的图片共 9583 张，其中含有大量的小目标行人和不同程度遮挡的行人目标。将数据集随机划分为包含 8633 张图片的训练集和包含 950 张图片的验证集。每张图片均进行了标注，标注包括行人边界框和属性标签等，数据集部分图片如图 3-7 所示。

本章行人检测模型实验的硬件环境为:NVIDIA GeForce RTX3090 GPU 24GB 显存，Intel 酷睿（Core）i9-14900KF @ 3.20GHz CPU 128GB 内存。软件环境：Windows 10 系统，CUDA12.1，Pytorch 2.2.2，编程语言 python 3.9。实验的总 Epochs 设置为 150，Batch_size 设置为 16，实验中选择不采用预训练权重，同时为了防止模型过拟合，本实验设置早停等待轮数为 50。

图 3-7 Pascal VOC2012 数据集部分图片

Figure 3-7 Image of a portion of the Pascal VOC2012 dataset

3.3.2 行人检测评价指标

为了验证本文提出的 YOLOv8n-DGT 算法的优异性，采用精确率（Precision, P）、召回率（Recall, R）、平均精度率（mean Average Precision, mAP）、参数量（Parameters）和每秒浮点运算次数（GFLOPs）作为评价指标。其中 Parameters 和 GFLOPs 是用于衡量模型的体积与计算复杂度的指标。P、R、mAP 则是评估算法检测效果的指标，这三类指标越高，反映该行人检测模型性能更优越。定义如式（3-8）至（3-10）所示：

$$P = \frac{TP}{TP + FP} \tag{3-8}$$

$$R = \frac{TP}{TP + FN} \tag{3-9}$$

$$mAP = \frac{\sum_{i=1}^{k} AP_i}{k} \tag{3-10}$$

其中，TP（True Positive）表示检测正确的目标数量，FP（False Positive）表示检测错误的目标数量，FN（False Negative）表示正确目标中未被检测到的目标数量。

3.3.3 消融实验结果分析

（1）C2f-DCNv2 模块位置分析

C2f-DCNv2 模块可以扩大模型对输入特征的感受野，更好地捕捉目标特征来适应不同行人目标动态尺寸的变化，提升模型检测性能。为了验证 C2f-DCNv2 模块的引入位置对网络性能的影响，设计了如下表 3-1 所示的消融实验。

实验结果表明，单个的 C2f-DCNv2 模块对目标特征的提取效果较差，单独在 YOLOv8n 的骨干网络引入可变形卷积后，提高了骨干网络对不同尺寸特征的提取能力，小幅提升了模型的检测效果。而在颈部网络融入可变形卷积，因为颈部网络本身的特征提取能力足够，模型的检测效果并未有明显变化。在模型骨干网络和颈部网络同时引入 C2f-DCNv2 模块时，大幅提高了网络的特征提取能力，同时平衡了模型的参数量和计算复杂度，P、R、mAP50%和 mAP50-95%分别提升了 1.8%、0.2%、1.3%和 0.7%。综上，引入可变形卷积加强了检测模型对不同

尺寸特征图的提取能力，有利于实现局部特征和全局特征的信息融合，使模型能够更加精确的拟合不同方向和大小的行人特征。因此本章采用在骨干网络和颈部网络都引入 C2f-DCNv2 模块的设计。

表 3-1 C2f-DCNv2 不同位置效果对比

Table 3-1 Comparison of the effect of different positions of C2f-DCNv2

Algorithm	P/%	R/%	mAP50/%	mAP50-95/%	Params/10^6	GFLOPs
YOLOv8n	73.9	80.3	82.1	63.5	3.05	8.1
C2f-DCNv2-backbone	74.7	79.6	82.8	63.9	3.03	8.0
C2f-DCNv2-neck	73.5	79.3	82.0	63.7	3.03	8.0
C2f-DCNv2-all	**75.7**	**80.5**	**83.4**	**64.1**	**3.02**	**7.9**

（2）模块有效性分析

为了进一步分析改进 YOLOv8n 网络的有效性和其中各个模块之间组合对原模型的提升。在参数设置相同的情况下，以 YOLOv8n 为基准设计了如下消融实验。不同改进方法的消融实验结果如表 3-2 所示,并用√表示添加了该改进方法。

表 3-2 消融实验结果

Table 3-2 Results of Ablation Experiments

C2f-DCNv2	GAM	TADDH	P/%	R/%	mAP50/%	mAP50-95/%	GFLOPs
			73.9	80.3	82.1	63.5	8.1
√			75.7	80.5	83.4	64.1	**7.9**
	√		72.4	82.1	82.7	63.3	9.4
		√	73.1	82.9	83.7	64.8	8.6
√	√		73.5	82.4	84.0	64.6	9.1
√		√	77.6	77.4	84.2	64.9	8.5
	√	√	75.5	78.2	84.0	64.8	10.2
√	√	√	**78.7**	**82.3**	**84.5**	**65.2**	9.9

从表中可以看出，分别引入 C2f-DCNv2、GAM 和 TADDH 模块，模型在 P、R 和 mAP50 等方面都有不同程度的提升。C2f-DCNv2 模块对模型准确率和召回率均有所提升，通过引入 C2f-DCNv2 模块提高了模型对不同尺寸行人目标的检测能力，降低了背景信息的干扰提高了 1.6%的准确率和 1.3%的平均精度。GAM

注意力机制通过降低背景信息个干扰，大幅提高了模型的召回率并提高了 0.6%的平均精度。TADDH 模块促进了模型对交互特征的提取，提高了模型召回率 2.6%和平均精度 1.6%，同时采用共享卷积处理交互特征，降低了模型参数量，由此证明本文提出的轻量化检测头的优越性能。将模块进行组合能进一步提高改进算法的平均精度，但召回率会小幅下降。将三处改进结合，由最后一组实验可以看出改进后模型在各项指标上都得到了较大的提升，其中 P、R、mAP50%和 mAP50-90%分别提高了 4.8%、2%、2.4%、1.7%，且模型复杂度并未大幅增加，证明了改进算法的优越性。

3.3.4 对比试验结果分析

为了进一步验证本章改进算法的性能，在 Pascal VOC2012 行人数据集上对于融入了 C2f-DCNv2 模块、GAM 注意力机制和动态任务对齐检测头的行人检测模型 YOLOv8n-DGT 进行验证。对经典的目标检测算法和本文改进算法检测效果进行对比。结果如表 3-3 所示。

表 3-3 不同行人检测算法结果对比

Table 3-3 Comparison of results of different pedestrian detection algorithms

Model	P/%	R/%	mAP50/%	mAP50-95/%	Params/M	GFLOPs
RT-DETR-L[57]	74.0	79.2	82.4	60.6	31.9	103.0
Faster-RCNN	68.7	77.4	77.8	54.4	137.1	370.2
YOLOv5s	69.1	79.3	79.6	60.7	7.0	16.0
YOLOv6n	72.1	78.4	80.3	60.9	4.3	11.1
YOLOv7-tiny	70.6	79.7	81.4	60.8	6.0	13.2
YOLOv8n	73.9	80.3	82.1	63.5	3.0	8.1
Gold-YOLO-N[58]	74.3	81.4	83.4	64.3	5.6	12.1
Mamba YOLO-T[59]	75.1	**83.1**	83.8	64.5	6.1	14.3
YOLOv9t[60]	76.2	81.2	84.3	**65.4**	2.8	12.1
YOLOv10n[61]	74.3	79.1	82.6	62.7	2.6	8.2
YOLOv11n[62]	74.5	80.1	82.3	63.4	**2.6**	**6.6**
YOLOv8n-DGT	**78.7**	82.3	**84.5**	65.2	3.9	9.9

在同样的环境下，将本章改进算法与主流行人检测算法 RT-DETR-L、Faster-RCNN、YOLO 系列算法进行了对比实验。与 YOLOv8n 算法对比，改进模型在参数量仅增加了 0.9M 的情况下，P、R、mAP50%和 mAP50-95%分别提高了 4.8%、2%、2.4%和 1.7%，证明了 YOLOv8n-DGT 算法的优越性。与 YOLOv5s、YOLOv6n、YOLOv7tiny 算法对比，mAP50%分别提高 4.9 个、4.2 个、3.1 个百分点，mAP50-95%分别提高了 4.5 个、4.3 个、4.4 个百分点。并与新推出的 YOLOv9t、YOLOv10n、YOLOv11n 算法相比，在模型参数量相近的情况下，具备了更高的检测精度，验证了本文算法的改进仍具有有效性。与 Gold-YOLO-N 和 Mamba YOLO-T 对比，在检测性能相近的情况下，模型更加轻量化。与经典模型训练过程的指标对比见图 3-8。由下图可以看出改进模型的各项指标均高于目前主流的行人检测算法，且收敛速度更快，曲线更加平稳。

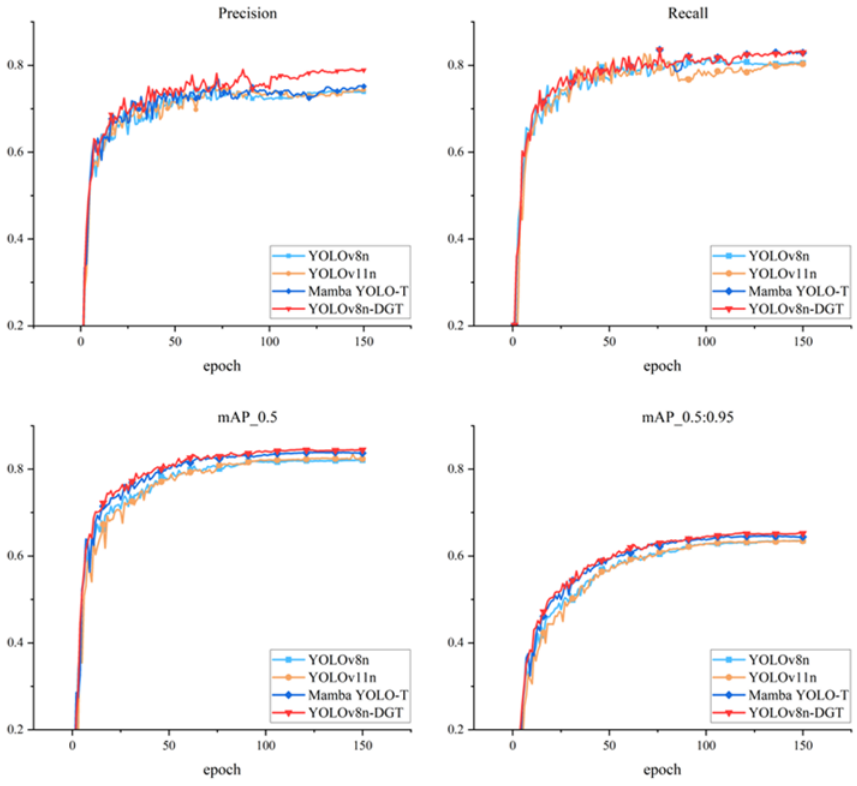
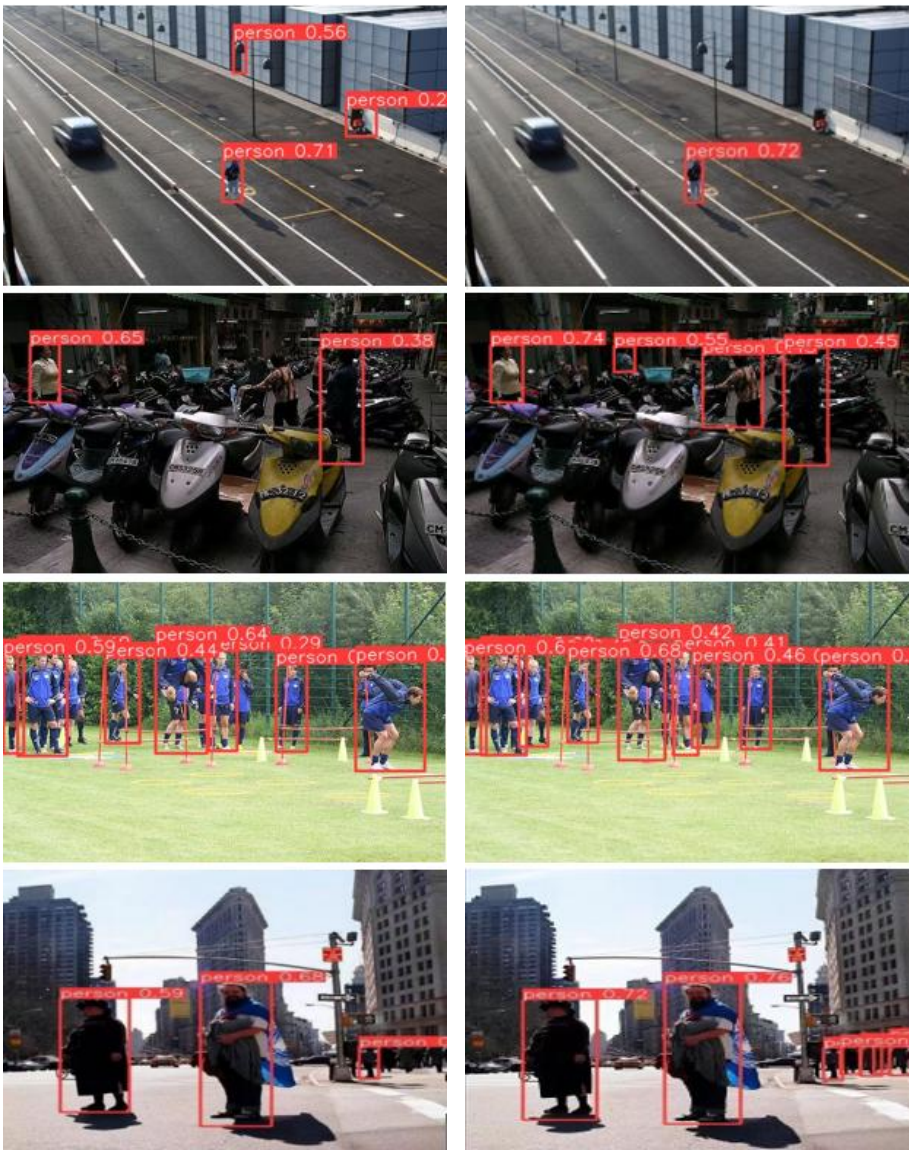


图 3-8 行人检测算法指标对比

Figure 3-8 Comparison of Pedestrian Detection Algorithm Metrics

为了更直观的展示改进模型对复杂场景下的行人目标的检测效果，将本文算法和基准模型 YOLOv8n 在 Pascal VOC2012 验证集上的检测结果进行可视化分

析。对于数据集中的常见场景下原始 YOLOv8n 算法与 YOLOv8n-DGT 行人检测算法对比，如图 3-9 所示。



（a）原算法检测效果　　（b）改进算法检测效果

图 3-9 检测效果可视化对比

Figure 3-9 Comparison of detection effect visualization

对比图 3-9（a）、（b）两组图片可看出，第一张图片中原算法将场景中的电车、路灯误检成了行人，改进算法检测的更准确并未出现误检的情况；在第二张图中原算法在暗光的复杂场景中仅检测出两个目标，而改进算法精准检测出了四个行人目标，降低了漏检率；第三张图在人物互相遮挡的情况下，原算法并未检测出图片左侧和中间部分被严重遮挡的两个目标，而改进算法完整检测出了图中的所有目标，对遮挡目标的检测效果更好；在第四张图中原算法未检测出图片右

下角的小目标行人，而改进算法通过融合可变形卷积，提高了不同尺寸目标的定位能力，精准的检测出了小目标行人。本文改进算法在复杂场景下仍可以有效地检测出被遮挡的目标和小尺寸行人目标，显著提高了网络对不同尺寸行人特征的提取能力和不同特征的融合能力，在复杂场景下具有更好的检测性能。

实验结果表明，与其他的经典行人检测模型比较，YOLOv8n-DGT 检测模型在并未大幅增加参数量的基础上，提升了复杂场景下多尺度行人检测的精度，且满足实时性的检测需求，具有优异性和高效性。

3.4 本章小结

本章提出了一种结合可变形卷积、全局注意力与动态任务对齐检测头的多尺度行人检测算法 YOLOv8n-DGT，解决了 YOLOv8n 行人检测模型在复杂场景下漏检率高，多尺度目标检测效果差的问题。在模型计算复杂度仅增加 1.8GFLOPs 的情况下，大幅提高了检测性能。首先通过引入可变形卷积重构网络中的 C2f 模块，增强模型对多尺度行人特征的动态捕捉能力；然后在颈部网络嵌入全局注意力机制，通过通道与空间注意力交互抑制背景干扰，增强目标区域的语义理解与定位精度；最后设计动态任务对齐检测头，利用任务交互与动态特征选择机制优化检测头分类与定位的协同性能。实验结果证明，和当下主流行人检测模型对比，改进后的 YOLOv8n-DGT 多尺度行人检测模型各项指标均具有优越性，实现了检测性能和计算复杂度的平衡，具有实际的应用价值。

第 4 章 增强局部特征提取的抗遮挡行人跟踪模型

4.1 引言

监控视频中的行人检测跟踪算法主要分为检测与跟踪两部分。行人检测部分，采用前文构建的多尺度行人检测模型 YOLOv8n-DGT 作为检测器，在复杂场景中达到了检测精度与模型复杂度的平衡。行人跟踪部分，本章将采用 StrongSORT 行人跟踪模型作为基线模型，进行相关研究与改进。

StrongSORT 算法是继 DeepSORT 算法之后的又一经典的基于检测的多目标跟踪算法，虽然相比于 DeepSORT 算法取得了不错的跟踪效果，但仍存在一些不足。其中，在外观分支上，StrongSORT 算法使用了由 ResNeSt50 网络构建的 BoT 特征提取网络，该网络对局部特征提取能力较弱，限制了目标跟踪模型对遮挡行人的跟踪效果。在匹配策略上，StrongSORT 算法使用了全局线性匹配策略来代替 DeepSORT 算法中的级联匹配策略，虽然提高了模型匹配的准确度，但在处理非确认态的预测框、未匹配的预测框及未匹配的检测框时，仍采用 IoU 匹配算法对其进行处理。IoU 匹配算法由于阈值的限制且仅考虑检测框与预测框的重叠程度，容易对遮挡后重新出现的目标赋予新的轨迹，进而导致 ID Switch 的发生，从而影响了多目标跟踪算法的稳定性。为解决上述问题，本章基于增强特征提取与抗遮挡匹配优化，构建了增强局部特征提取的抗遮挡行人跟踪模型。

4.2 增强特征提取的行人跟踪模型

本节主要对 StrongSORT 算法中的特征提取网络以及 IoU 匹配策略进行了相关研究和改进，以提高复杂场景下被遮挡行人的跟踪精度与稳定性。首先，由于在视频画面中行人数量较多，易发生被遮挡现象，导致行人部分外观信息丢失，目标信息较少。为此本文采用多分支协作网络 BC-OSNet[63]（Branch-Cooperative OSNet），通过四个协同分支将特征图进行分割后提取局部特征，充分挖掘特征图中丰富的上下文信息与局部特征，以提高跟踪模型的目标局部特征提取能力。同时当目标被长时间遮挡时，其预测位置与实际检测位置的偏离往往超过 IOU 匹配的阈值范围，容易导致检测的结果无法与原轨迹进行匹配，从而产生新的轨迹，

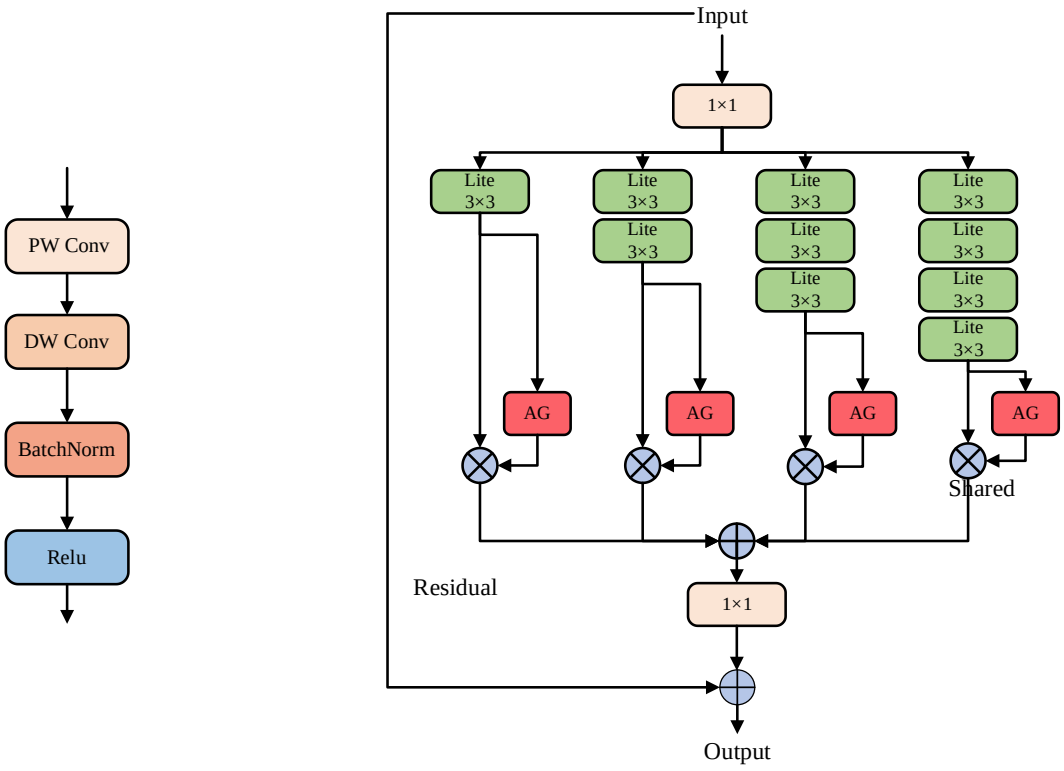
导致身份切换次数偏高。因此针对 IoU 匹配存在的不足，本章在原 StrongSORT 算法 IoU 匹配的基础上，设计了动态阈值调整机制，根据当前帧的检测目标数量，动态调整匹配阈值，同时引入 Inner-CIoU 优化匹配方式，提高行人跟踪模型匹配精度和适应能力。

4.2.1 多分支协作特征提取网络

在外观分支上，虽然 StrongSORT 算法使用了由 ResNeSt50 网络构建的 BoT 特征提取器来提升算法的跟踪性能，但是由于 BoT 特征提取器计算量较高，且无法有效提取行人目标各个区域的细节特征，对被遮挡目标检测效果较差。所以本节对 StrongSORT 算法中的 BoT 特征提取器进行了优化，引入了基于全尺度网络[64]（Omni-Scale Network, OSNet）改进的多分支协作特征提取网络 BC-OSNet，以充分提取行人局部特征和多尺度特征。

OSNet 是一种专为行人重识别（ReID）设计的深度卷积神经网络，与传统的特征提取网络不同，OSNet 网络采用了一种创新的骨干架构，其能够以最低的计算成本提取并融合不同尺度下的行人特征，增强模型的表征能力。OSNet 的核心设计理念主要包含三个方面：

（1）全尺度残差模块。OSNet 使用堆叠不同尺寸的卷积核来提取行人图像在不同感受野中的特征。同时，它设计了全尺度残差模块（Omni-Scale Residual Block, OSBlock），结构图如图 4-1（b）所示，其采用了残差结构和多尺度动态特征融合的思想，内置了多个卷积特征流，通过堆叠不同数量的 Lite 层，每个流都获得了独特的感受野，这种设计使得全尺度残差模块能够在每个残差块内部并行地处理不同尺度的特征提取任务，进一步强化了模型的多尺度特征提取能力。其中的 Lite 模块则如图 4-1（a）所示。Lite3×3 结构包括逐点卷积、卷积核为 3×3 深度卷积、BN（批归一化）层和 ReLU 激活层。由于采用了深度可分离卷积技术，将传统的卷积操作分解为深度卷积和逐点卷积，该结构显著降低了模型的参数量，减少了计算复杂度。



（a）　　（b）

图 4-1 Lite 模块与 OS 模块

Figure 4-1 Lite Module and OS Module

（2）动态特征融合机制，即一个基于注意力机制的统一聚合门。为了更好的提取全尺度特征，OSNet 以动态的方式组合不同流的输出，即根据输入图像给不同的尺度分配不同的权重。该机制能够灵活地调整不同尺度特征的融合权重，使得 OSNet 能够根据不同的输入图像自适应地捕获和融合多尺度特征，从而生成更具辨识度的全尺度特征。具体的计算过程如公式（4-1）所示。

$$\tilde{\boldsymbol{x}}=\sum_{t=1}^{T}G\left(\boldsymbol{x}^{t}\right)\odot\boldsymbol{x}^{t},s.t.\boldsymbol{x}^{t}\triangleq F^{t}\left(\boldsymbol{x}\right) \tag{4-1}$$

其中，$\tilde{\boldsymbol{x}}$ 表示统一聚合门的输出，T 表示 OSBlock 模块中分支的个数，$F^{t}\left(\boldsymbol{x}\right)$ 表示第t个分支该感受野下提取的特征，$\odot$ 代表 Hadamard 积，$G\left(\boldsymbol{x}^{t}\right)$ 由 $\boldsymbol{x}^{t}$ 输入一个微型网络计算后得到。这个微型网络由一个全局平均池化层和一个包含非线性激活函数 ReLU 的 MLP 两部分组成。整个微型网络最终使用非线性激活函数 Sigmoid 进行激活。

（3）深度可分离卷积。OSNet 特征提取网络中采用深度分离卷积代替传统卷积，以此减少模型参数量和运算量。深度可分离卷积（Depthwise Separable

Convolution）将标准卷积分解深度卷积（Depthwise Convolution）和逐点卷积（Pointwise Convolution）。在深度卷积中，每个输入通道分别与一个卷积核进行卷积，提取每个通道的空间特征。随后，逐点卷积使用 1×1 的卷积核，将不同通道的特征进行线性组合，实现通道间的信息融合。这种分解方式显著减少了模型的参数量和计算量。以一个输入为 64×64 大小的 3 通道的特征图为例，若使用标准卷积得到 4 个输出通道，参数量约为 112；而采用深度可分离卷积，参数量仅约为 46，计算参数量减少了近三分之二。

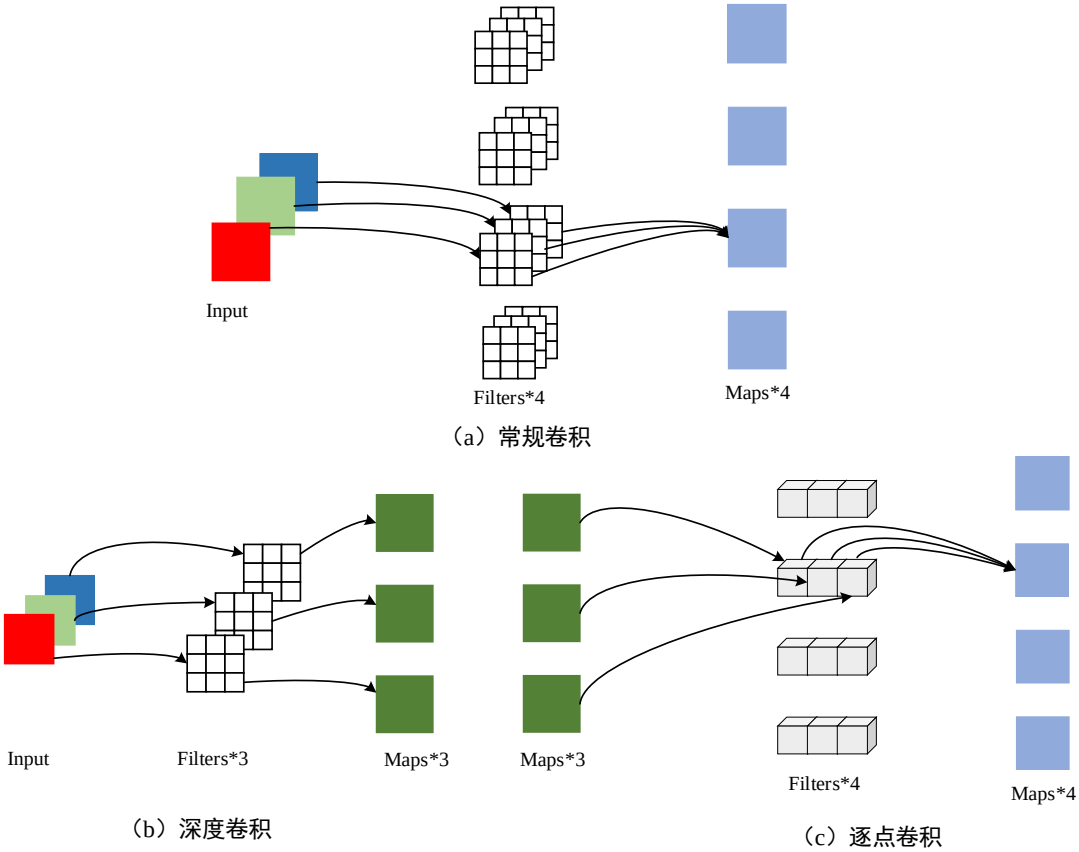


图 4-2 常规卷积与深度可分离卷积操作

Figure 4-2 Conventional Convolution and Depth Separable Convolution Operations

OSNet 的整体结构主要由卷积层、池化层以及 OSBlock 堆叠组成，其网络结构前五层如表 4-1 所示。

BC-OSNet 的整体网络架构如图 4-3 所示，其中输入图片尺寸为 256×128×3。BC-OSNet 网络采用 OSNet 的前 5 层作为共享特征提取层，输出基础特征图。然后，通过局部分支(V1)、全局分支(V2)、全局对比池化分支(V3)和关系分支(V4)四个协同分支更充分的提取行人局部外观特征。

表 4-1 OSNet 网络前五层结构

Table 4-1 The structure of the first five layers in the OSNet network

stage	output	OSNet
conv1	128×64, 64 64×32, 64	7×7 conv, stride 2 3×3 max pool, stride 2
conv2	64×32, 256	OSBlock×2
transition	64×32, 256 32×16, 256	1×1 conv 2×2 average pool, stride 2
conv3	32×16, 384	OSBlock×2
transition	32×6, 384 16×8, 384	1×1 conv 2×2 average pool, stride 2

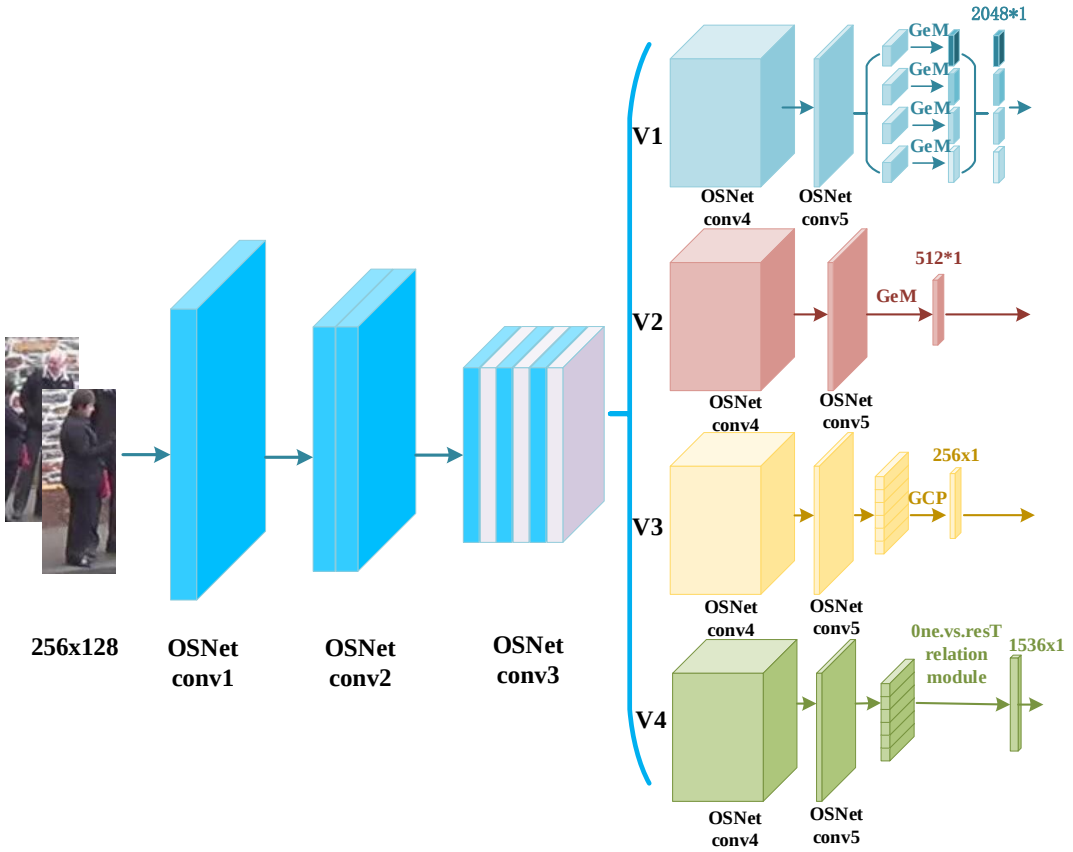


图 4-3 BC-OSNet 结构图

Figure 4-3 Branch-Cooperative OSNet Structure Diagram

（1）局部分支：在局部分支网络中，采用类似于 PCB[66]（Part-based Convolutional Baseline）结构，将特征图水平切分为 4 块，再通过对每块特征图

进行平均池化操作，提取到 4 个 $1\times1\times C$ 局部特征向量。与 PCB 单独计算每一分块的损失不同，BC-OSNet 先将 4 个子区域的特征向量拼接为全局向量 f，计算公式见式（4-2）：

$$f=[f_1^T, f_2^T, \cdots, f_4^T]^T \tag{4-2}$$

f 代表拼接后的特征向量，其中 f_1，f_2，f_3，f_4 表示将特征图水平分割后得到的 4 个列向量。标注数据集记为$\{(x_i,\ y_i), \mathrm{i}=1, 2, \cdots,\ N_s\}$。然后再通过 ID 预测损失优化，如式（4-3）所示：

$$L=-\frac{1}{N_s}\sum_{i=1}^{N_s}\log\left(\frac{\exp\left(\left(W^{y_i}\right)^T f^i+b_{y_i}\right)}{\sum_j\exp\left(\left(W^j\right)^T f^j+b_j\right)}\right) \tag{4-3}$$

其中 W^{y_i} 和 W^j 分别是权重矩阵 W 的第 y_i 列和第 j 列。此处采用广义均值池化（Generalized-Mean, GeM）提取各个分块的特征向量，公式如式（4-4）所示：

$$GeM\left(f_k=[f_0, f_1, \cdots, f_n]\right)=[\frac{1}{n}\sum_{i=1}^{n}(f_i)^{p_k}]^{\frac{1}{p_k}} \tag{4-4}$$

其中 GeM 运算符 f_k 为单个特征映射，当 $p_k\to\infty$时，GeM 等于最大池化，当 $p_k=1$ 时，GeM 等于平均池化，在本地分支中以 $p_k=1$ 初始化 GeM 参数。

（2）全局分支：与 OSNet 的预测分支相似，但 BC-OSNet 使用 GeM 池化来替代 OSNet 中的全局平均池化，以获得预测向量。GeM 池化直接在 conv4 和 conv5 之后执行。在全局分支中，GeM 初始化设置 $p_k=6.5$，由此生成一个 512 维的全局特征向量，用于补充局部分支的细节信息，避免因过度关注局部特征而忽略全局上下文信息。

（3）全局对比池化分支:全局对比池化(Global Contrastive Pooling，GCP)分支主要用于捕捉行人目标与背景的对比差异，抑制背景噪声。GCP 分支进一步在全局平均池化（Global Average Pooling, GAP）和全局最大池化（Global Maximum Pooling, GMP）的基础上提取局部对比特征。在 conv5 卷积层的输出处获得特征映射后，将特征图划分为 6 个水平子区域，分别计算每个区域的 GAP 和 GMP，然后通过全局对比池化得到 256 维的对比特征向量。

（4）关系分支：One-vs.-rest 关系分支主要用于增强局部特征关联。通常，

局部特征往往只包含单个部分的信息，这些信息并不能反映各个局部特征之间的关系。One-vs.-rest 关系分支使得每个局部特征相互关联，增强局部特征与其上下文的关联。与 GCP 分支类似，关系分支网络将特征图划分为 6 个水平特征图，记为($f_1, \cdots, f_6$)，对每个局部特征 f_i，计算其余区域的特征均值 r_i，如式（4-5）所示：

$$\mathbf{r}_i = \frac{1}{5}\sum_{j \neq i}\mathbf{f}_j. \tag{4-5}$$

然后，将特征图 f_i 和 r_i 经过瓶颈层压缩，将通道数从 C 减少到 c，生成 $f_i^{'}$ 和 $r_i^{'}$。再采用关系分支网络融合生成局部关系特征 q_i（q_i 是一个 256 维向量），公式如（4-6）所示，最后将局部关系特征拼接得到 1536 维的最终关系特征。

$$q_i = f_i^{'} + B\left(\mathrm{C}\left(f_i^{'}, r_i^{'}\right)\right), i = 1, \cdots, 6 \tag{4-6}$$

4.2.2 抗遮挡匹配优化

StrongSORT 算法在继承 DeepSORT 算法多目标跟踪框架的基础上，通过优化匹配策略进一步提升跟踪性能。首先通过运动信息构建代价矩阵进行初步关联，然后采用 IoU 匹配算法对非确认态的预测框、未匹配的预测框及未匹配的检测框再次进行匹配；然而在密集人流的长期遮挡场景中，目标可能被其他物体完全覆盖超过数十帧，此时卡尔曼滤波的运动预测误差会随遮挡时长呈指数级累积。当目标再次出现时，其预测位置与实际检测位置的偏离往往超过 IoU 匹配的阈值范围，容易导致检测的结果无法与原轨迹进行匹配，从而产生新的轨迹，进而引起 ID Switch 的问题。

针对该问题，本文设计了动态阈值调整机制，根据当前帧检测目标的数量，动态调整匹配阈值，在目标密度不同的情况下，通过增减 IoU 阈值来适应复杂的跟踪环境。在高密度区域（如人群聚集）下，目标之间的相对位置更加紧密，遮挡严重，需要更高的重叠度才能认为是成功匹配。因此在密集场景下，自动提升阈值，防止因检测框密集造成的 ID 混杂，降低误检率。在目标密度较低时，减小匹配阈值，提高匹配效率。因此本文首先设置一个 IoU 匹配的基础阈值，然后根据检测到的目标数量与某个基准值的比例，来增加一个额外的动态阈值。经过实验推理验证，将基础 IoU 阈值设置为 0.5，基准目标值设为 20，动态系数设置

为 0.1。IoU 匹配阈值逻辑如式（4-7）所示,其中 num 代表场景行人目标数量，$\max\_iou\_dist$ 代表 IoU 最大匹配阈值。

$$\max\_iou\_dist = 0.5 + 0.1 \times \frac{num}{20} \tag{4-7}$$

其次，StrongSORT 算法通过 IoU 匹配对检测框与预测框进行匹配，当两者交并比超过预设阈值时，视为匹配成功；否则，匹配失败。但当检测目标框与跟踪预测框不相交或部分重叠时，IoU 匹配效果较差。在不相交的情况下，IoU 值为 0，无法准确反映检测框与预测框的匹配程度。如图 4-4 所示，虽然这三个框的 IoU 值相同，但是检测框和预测框的相对位置却并不相同。

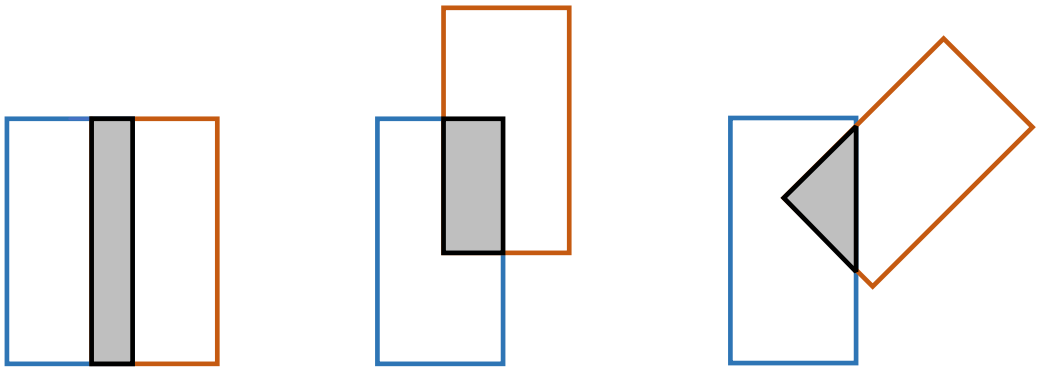


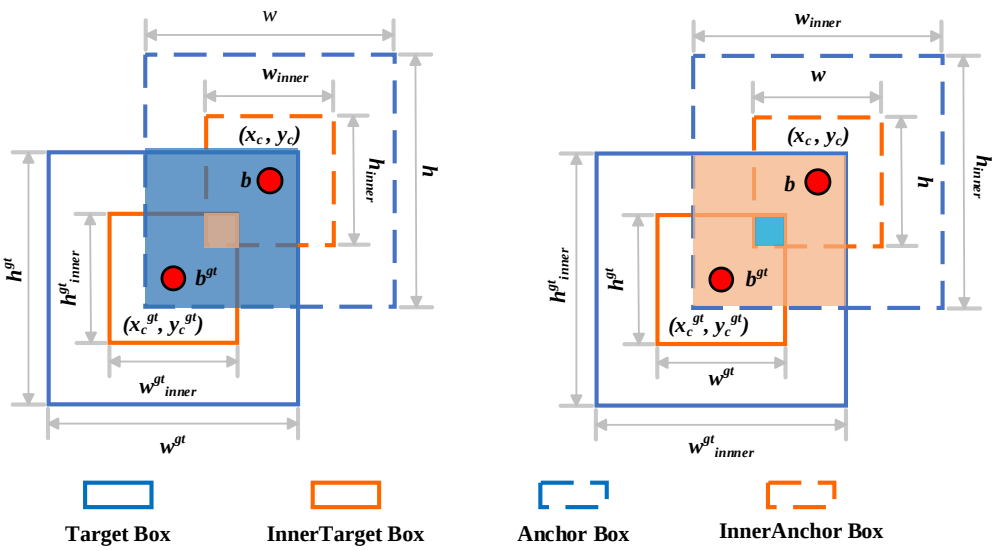
图 4-4 相同 IoU 值的不同重叠情况

Figure 4-4. Different overlapping cases for the same IoU value

针对 IoU 匹配存在的不足，CIoU 在 IoU 的基础上引入了中心点距离和宽高比一致性的惩罚项，综合考量了预测框与真实框之间的重叠面积、中心点距离以及长宽比三个要素，以更全面地衡量预测框与真实框之间的差异，使得目标框的预测更加稳定。公式如式(4-8)所示：

$$CIOU = IOU - \frac{\rho^2\left(b, b^{gt}\right)}{c^2} - \alpha v \tag{4-8}$$

其中 v 代表两个框的长宽比的差距，α 为权重系数。CIoU 在一定程度上优化了 IoU 的不足，但当真实框与预测框的长宽比例相似，实际宽度和高度差异显著时，损失函数中的惩罚项 αv 将会失效，不利检测框与预测框的匹配。



4-5 Inner-IoU 原理图

4-5 Inner-IoU Schematic

针对上述问题，本文采用 Inner-IoU[67]与 CIoU 匹配相结合，构建了 Inner-CIoU。Inner-IoU 原理图如图 4-5 所示。Inner-IoU 引入了辅助边界框，通过调整比例系数 $ratio$，可以更精确地评估目标检测框与预测框重叠程度，有助于对复杂重叠情况下的目标进行匹配。与 IoU 相比，比例系数小于 1 时，辅助边界尺寸小于实际边界，能够加速高 IoU 样本的收敛；比例系数大于 1 时，辅助边界扩展了回归的有效范围，对低 IoU 的回归存在增益[68]。Inner-CIoU 原理如式(4-9)至（4-14）所示，其中 (x_c^{gt}, y_c^{gt}) 表示真值框和内部真值框的中心点坐标，(x_c, y_c) 表示锚框和内部锚框的中心点坐标，w^{gt}，h^{gt} 分别表示真值框的宽度和高度，而 w，h 分别表锚框的宽度和高度，$ratio$ 代表比例因子。

$$\begin{aligned} b_l^{gt} &= x_c^{gt} - \frac{w^{gt} * ratio}{2}, b_r^{gt} = x_c^{gt} + \frac{w^{gt} * ratio}{2} \\ b_t^{gt} &= y_c^{gt} - \frac{h^{gt} * ratio}{2}, b_b^{gt} = y_c^{gt} + \frac{h^{gt} * ratio}{2} \end{aligned} \tag{4-9}$$

$$\begin{aligned} b_l &= x_c - \frac{w * ratio}{2}, b_r = x_c + \frac{w * ratio}{2} \\ b_t &= y_c - \frac{h * ratio}{2}, b_b = y_c + \frac{h * ratio}{2} \end{aligned} \tag{4-10}$$

$$inter = \left(\min\left(b_r^{gt}, b_r\right) - \max\left(b_l^{gt}, b_l\right)\right) * \left(\min\left(b_b^{gt}, b_b\right) - \max\left(b_t^{gt}, b_t\right)\right) \tag{4-11}$$

$$union = \left(w^{gt} * h^{gt}\right) * \left(ratio\right)^2 + \left(w * h\right) * \left(ratio\right)^2 - inter \tag{4-12}$$

$$IoU^{inner} = \frac{inter}{union} \tag{4-13}$$

$$L_{Inner-CIoU} = L_{CIoU} + IoU - IOU^{inner} \tag{4-14}$$

4.3 行人跟踪实验结果分析

4.3.1 数据集介绍与实验环境

本章采用的数据集包含开源的多目标跟踪数据集 MOT16[69]（Multiple Object Tracking 16），数据集部分场景如图 4-6 所示。MOT16 数据集共由 14 个真实场景视频构成，总帧数约 11,000 帧，总目标数超过了 1,300 个，平均每帧包含 10 至 30 个目标，覆盖室内外、拥挤环境、低光照及运动模糊等复杂情况。MOT16 数据集包含训练集和测试集，训练集由 MOT16-02、MOT16-04、MOT16-05、MOT16-09、MOT16-10、MOT16-11、MOT16-13 构成，测试集包含 MOT16-01、MOT16-03、MOT16-06、MOT16-07、MOT16-08、MOT16-12、MOT16-14，训练集的视频序列均包含以 txt 文件储存的标注信息。

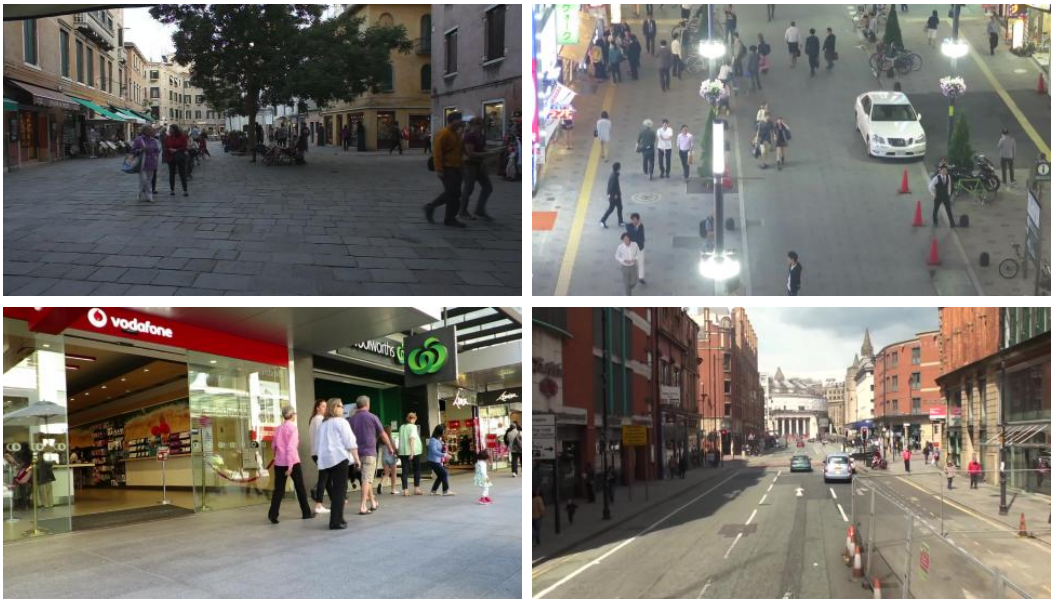


图 4-6 MOT16 数据集部分场景

Figure 4-6 Partial scenario of the MOT16 dataset

本章跟踪模型实验的硬件配置为:NVIDIA GeForce RTX3090 GPU 24GB 显存，Intel 酷睿（Core）i9-14900KF @ 3.20GHz CPU 128GB 内存。软件环境：Windows 10 系统，CUDA12.1，Pytorch 2.2.2，编程语言 python 3.9。

4.3.2 行人跟踪评价指标

本章采用在目标跟踪领域常用的评价指标来量化评价改进跟踪模型的性能，其中包括多目标跟踪精度（Multiple Object Tracking Accuracy, MOTA）、高阶跟踪精度（Higher Order Tracking Accuracy, HOTA）、IDF1 分数（Identification F1 Score, IDF1）和身份切换次数。这些指标从不同维度对跟踪性能进行评估。其中 MOTA 结合了目标检测效果和身份切换次数评估跟踪效果，其值越高，跟踪准确性越好，计算公式如式（4-15）所示，GT 表示真实目标框。

$$MOTA = 1 - \frac{\sum(FN + FP + IDSW)}{\sum(GT)} \tag{4-15}$$

HOTA 相比 MOTA 能够更加全面和均衡的评估跟踪算法的性能，该值更加重视目标检测和数据关联精度的平均性衡量。公式如（4-16）所示，其中 A 表示数据关联分数，TP 表示视频中跟踪正确检测框的数量。

$$HOTA_{\alpha} = \sqrt{\frac{\sum_{C\grave{o}TP}\mathrm{A}(\mathrm{C})}{TP + FN + FP}} \tag{4-16}$$

IDF1 是结合身份识别准确率和召回率的分数，IDSW 则代表身份切换数。如式（4-17）所示，其中 IDTP 表示检测框正确匹配的数量，IDFP 表示检测框错误分配的数量，IDFN 表示检测框被漏分配的数量。

$$IDF1 = \frac{2IDTP}{2IDTP + IDFP + IDFN} \tag{4-17}$$

4.3.3 消融实验结果分析

为了验证本章行人跟踪算法改进的有效性，本节在 MOT16 数据集上采用 YOLOv8n+StrongSORT 算法作为基准对比模型，并通过消融实验的方法，依次引入改进策略，系统的验证各个改进策略的有效性。

从表 4-2 可知，本节共设计了五组消融实验，依次引入 YOLOv8n-DGT、BC-OSNet、抗遮挡匹配优化等改进。第一、二两组实验结果表明，采用第三章设计的 YOLOv8-DGT 模型替换 YOLOv8n 作为前置目标检测器，行人跟踪模型的的跟踪性能采有所提高，其中 MOTA 提升了 2.06%，IDF1 提升了 0.9%，HOTA 提升了 0.47%。从而证明了第三章设计的 YOLOv8-DGT 行人检测模型对于整体跟

踪模型性能提升的有效性。然后在第三组实验中引入 BC-OSNet 特征提取网络，通过四个不同的网络分支提取行人目标的局部特征，增强对局部被遮挡行人的跟踪效果，使行人跟踪模型的 MOTA、IDF1 和 HOTA 进一步提升了 1.68%、1.66%、0.96%。第四组结合了 YOLOv8n-DGT 检测器和抗遮挡匹配优化，通过增加动态阈值调整机制，根据目标密度动态调整匹配阈值并结合 Inner-CIoU 匹配，使得行人跟踪模型具有了更强的鲁棒性，同时降低了行人身份切换次数。第五组实验则是结合所有的改进策略，针对密集场景遮挡情况进行优化，在各项跟踪指标上均取得了较好的效果，MOTA、IDF1 和 HOTA 分别达到了 45.21%，54.59%，42.17%。相比于 YOLOv8n+StrongSORT 基线模型 MOTA、IDF1 和 HOTA 分别提升了 3.86%，3.71%，2.07%，同时行人身份切换次数降低了 176 次。因此充分证明了本章基于 StrongSORT 跟踪模型改进策略的有效性。

表 4-2 消融实验对比结果

Table 4-2 Comparison results of ablation experiments

YOLOv8n-DGT	BC-OSNet	抗遮挡匹配	MOTA%	IDF1%	HOTA%	IDSw
			41.35	50.88	40.12	967
√			43.41	51.78	40.59	886
√	√		45.09	53.44	41.55	895
√		√	45.03	52.49	41.25	843
√	√	√	**45.21**	**54.59**	**42.17**	**791**

4.3.4 对比实验结果分析

为了进一步验证本文改进算法的优越性，将本文改进的算法和目前主流的 DeepSORT、DeepocSORT、ByteTrack、OC-SORT、BoT-SORT、StrongSORT 算法在 MOT16 数据集上进行了对比实验，其中目标检测器均采用第 3 章中改进的 DGT-YOLOv8n。实验结果如表 4-3 所示。

由第一组实验结果可以看出，DeepSORT 算法由于过于依赖目标检测质量，将低分目标检测框直接抛弃，导致对遮挡目标检测效果较差，MOTA 仅有 39.47%，且跟踪过程中的身份切换次数高达 1062 次。ByteTrack 算法相比于 DeepSORT 算法则保留了低置信度目标检测框并进行二次匹配，缓解了遮挡问题，MOTA 提高

了 3.41%，身份切换次数也降低了 314 次。但从第二组实验结果可以看出，ByteTrack 算法舍弃特征提取网络，在长时遮挡情况下跟踪效果仍有待提高。由第三至五组实验结果可知，OC-SORT、BoT-SORT 等 SORT 系列算法的跟踪效果相对 Deepsort 算法均有所提高。StrongSORT 算法在 DeepSORT 算法的基础上引入了 BoT 特征提取网络、EMA 特征更新策略、NSA 卡尔曼滤波等模块，大幅提高了目标跟踪的性能，其 MOTA、IDF1 和 HOTA 相比于 DeepSORT 算法分别提高了 3.94%、10.72%、7.1%。由最后一组实验结果可知，本章增强局部特征提取的抗遮挡 StrongSORT 模型在各项跟踪指标上均取得了不错的效果。MOTA、IDF1 和 HOTA 相比与 StrongSORT 模型分别提高 1.8%、2.81%、1.58%，IDSw 降低了 95 次，验证了本章构建的抗遮挡行人跟踪模型的有效性。同时本章优化算法与 OC-SORT、BoT-SORT、DeepocSORT 三种 SORT 系列算法相比，MOTA、HOTA、IDF1 均有所提升，由此也证明了本章改进算法的优越性。

表 4-3 基于检测的行人跟踪算法性能对比

Table 4-3 Performance comparison of detection-based pedestrian tracking algorithms

Model	MOTA%	IDF1%	HOTA%	IDSw
DeepSORT	39.47	41.06	33.49	1062
ByteTrack[70]	42.88	52.35	40.67	748
OC-SORT[71]	44.64	51.79	40.07	924
BoT-SORT[72]	44.07	53.03	41.21	**433**
DeepocSORT[73]	45.09	52.80	40.62	722
StrongSORT	43.41	51.78	40.59	886
Ours	**45.21**	**54.59**	**42.17**	791

同时，为了更加直观的验证本文改进算法的跟踪效果，在 MOT16 数据集上对 StrongSORT 算法和改进后的算法进行了可视化的对比实验。

（1）街道遮挡场景跟踪效果

如图 4-7 所示，（a）、（b）两组分别为原 StrongSORT 算法和本文改进跟踪算法在 MOT-02 视频序列下的可视化效果，MOT-02 视频序列代表街道遮挡场景，第一、第二和第三行图片分别为视频第 172 帧、第 199 帧和第 231 帧的跟踪效果。

（a）StrongSORT算法跟踪效果　　（b）本文改进算法跟踪效果

图 4-7 MOT-02 视频序列跟踪效果对比

Figure 4-7 MOT-02 Video Sequence Tracking Effect Comparison

从（a）组图片可以看出原 StrongSORT 算法在第 172 帧对蓝色上衣被遮挡行人 ID 编号为 23，该行人在第 198 帧时被 ID-4 行人完全遮挡，直至 231 帧时再次成功跟踪，经过长达 32 帧的遮挡影响行人 ID 被重新分配为 163。原跟踪模型在目标被长时间遮挡的情况下会将其识别为新的目标，造成 ID 频繁切换，影响行人跟踪效果。原跟踪算法在 MOT16-02 视频序列中 IDSw 高达 176 次，同时在跟踪过程中，图片右侧的杂物也被错误识别为行人，进行了无效的跟踪。由（b）组图片可知，在视频第 172 帧，蓝色上衣行人 ID 为 53，由于改进算法针对局部特征提取与匹配策略方面进行了优化，因此该行人经长时间遮挡再次被成功跟踪时，行人 ID 并未发生变化，实现了对遮挡行人的稳定跟踪。本章改进模型在

MOT16-02 视频序列中 IDSw 仅为 69 次，大幅降低了行人 ID 的切换次数，验证了本章抗遮挡行人跟踪模型的跟踪稳定性。并且在整个跟踪流程中，本章改进算法并未出现误检情况，从而验证了前文构建的多尺度行人检测模型 YOLOv8n-DGT 作为前置目标检测器的有效性。

（2）夜晚遮挡场景跟踪效果

（a）StrongSORT算法跟踪效果　　（b）本文改进算法跟踪效果

图 4-8 MOT-04 视频序列跟踪效果对比

Figure 4-8 MOT-04 Video Sequence Tracking Effect Comparison

MOT-04 视频画面代表夜晚遮挡场景，为了验证本章改进算法在夜晚遮挡场景下的跟踪效果，在 MOT-04 视频上进行了可视化分析，跟踪效果如图 4-8 所示。原 StrongSORT 算法在视频第 3 帧时，将道路中间灰色衣服行人 ID 初始化为 11，道路右侧白色上衣行人 ID 为 12。在视频第 71 帧时，ID-11 行人被 ID-12 行人完

全遮挡丢失跟踪目标。经过 10 帧的遮挡，ID-11 于第 81 帧重新出现，但身份切换成了 ID-84，造成了身份切换。原算法对场景左上角密集区域行人跟踪效果较差，行人身份切换频繁。整个跟踪过程中，原跟踪算法对场景中的白色汽车进行了无效跟踪。而由图 4-8（b）可以看出，在截取的视频序列中，同样的两个目标在第三帧初始化 ID 为 13 和 5，在后续的跟踪过程中两位行人的 ID 未发生变化，遮挡情况下，两者的 ID 也并未发生跳变。由于采用多分支协作网络提取行人局部特征，抗遮挡优化模型对场景左上角密集区域行人的跟踪效果更加稳定。采用同时采用多尺度行人检测模型作为检测器，降低了跟踪过程中由误检导致的无效跟踪。

4.4 本章小结

本章针对 StrongSORT 算法在行人遮挡情况下跟踪精度低，身份切换次数较高的问题，构建了基于 YOLOv8＋StrongSORT 抗遮挡优化的行人检测跟踪模型。以前文设计的多尺度行人检测模型 YOLOv8n-DGT 为前置检测器，对 StrongSORT 算法进行改进，构建增强局部特征提取的抗遮挡行人跟踪模型。首先采用多分支协作网络，通过四个协同分支强化对遮挡目标的局部特征提取能力，提升遮挡目标的表征精度。然后优化关联匹配方式，设计动态阈值调整机制，来适应不同目标密度的跟踪场景，同时结合 Inner-IoU 与 CIoU 几何约束策略，优化检测框与预测框的匹配精度，有效减少长时间遮挡导致的身份切换，提升跟踪鲁棒性与稳定性。在 MOT16 数据集上实验结果表明，本章增强局部特征提取的抗遮挡行人跟踪算法对监控场景下的遮挡目标跟踪效果更好。

第 5 章 总结与展望

5.1 研究总结

本研究围绕监控视频场景中的行人检测与跟踪任务，针对复杂背景干扰、行人多样化、遮挡频发及身份切换等问题，提出了一种融合多尺度行人检测与增强局部特征提取的抗遮挡行人跟踪框架。在深入分析现有检测与跟踪算法局限性的基础上，本文从检测模型优化与跟踪策略增强两个层面展开研究。

本文的主要研究内容如下：

（1）由于背景信息干扰严重、行人外观多样化等因素，行人检测算法易出现漏检、错检等问题。针对该问题，本文基于 YOLOv8 检测模型进行优化，提出了一种多尺度行人检测模型。首先，在模型的 C2f 模块中融入可变形卷积，降低模型参数量并提升模型对不同尺度行人特征的提取能力；其次，通过引入全局注意力机制，抑制背景干扰，增强目标区域的语义理解与定位精度；最后，在头部网络采用任务交互与动态特征选择机制，进一步优化检测头分类与定位的协同性能。在 Pascal VOC 2012 数据集上对 YOLOv8-DGT 行人检测模型的性能进行验证，实验结果表明，该模型在维持模型轻量化的前提下，大幅提升了不同尺寸行人目标的检测精度，降低了漏检率，满足了实时行人检测的需求。

（2）针对因行人长时间遮挡导致的跟踪模型精度低、身份切换次数频繁等问题，本文基于 YOLOv8n-DGT 多尺度行人检测器，提出了一种增强局部特征提取的抗遮挡行人跟踪模型。在外观分支方面，构建了多分支协作网络，采用四个分支网络提取遮挡目标特征，增强模型对局部细节的捕捉能力，从而减少因部分遮挡导致的目标信息丢失问题。在关联匹配方面，设计了动态阈值调整机制，根据场景目标密度动态调整最大匹配阈值，结合 Inner-CIoU 匹配策略优化数据关联流程，降低因遮挡导致身份切换次数。最后，在 MOT16 数据集上进行实验验证，实验结果表明，增强局部特征提取的抗遮挡模型显著提高了跟踪精度，降低了行人身份切换次数，能够满足不同监控场景下的行人跟踪需求。

由 PASCAL VOC2012 数据集与 MOT16 数据集上的实验结果表明，相比于基准模型，本文改进的行人检测跟踪模型在检测与跟踪性能上均有显著提升，为

复杂监控场景下的行人检测与跟踪提供了可靠解决方案。

5.2 未来展望

本文通过改进检测模型的特征提取能力与跟踪算法的抗遮挡匹配策略，在行人检测的精度和被遮挡行人的跟踪稳定性均取得了一定的提升，有效解决了复杂监控场景下的行人检测与跟踪难题。然而，实际应用中仍存在不足需进一步研究。

（1）多模态数据融合，本文研究仅基于 RGB 图像数据，难以应对极端暗光场景。可融合红外热成像、毫米波雷达等多源传感器数据，构建跨模态特征提取网络，通过异构信息互补提升全时段、全天候场景下的行人检测与跟踪精度。

（2）模型泛化性增强，本文仅在 MOT16 目标跟踪数据集上进行了训练与测试。但实际监控场景复杂多变，未来需要结合更多数据集训练跟踪模型，以增强模型泛化性适应不同监控场景。

（3）端到端联合优化框架，当前检测与跟踪模块采用分阶段独立训练策略，存在任务特征不兼容问题。未来可设计检测跟踪一体化网络，通过共享特征提取层与联合损失函数，实现端到端优化，降低系统延迟与计算冗余。

参考文献

[1] Hazra S, Mandal S, Saha B, et al. UMTSS: A unifocal motion tracking surveillance system for multi-object tracking in videos[J]. Multimedia Tools and Applications, 2023, 82(8): 12401-12422.

[2] 曹自强, 赛斌, 吕欣. 行人跟踪算法及应用综述[J]. 物理学报, 2020, 69(08): 41-58.

[3] 于明鑫, 王长龙, 张玉华, 等. 复杂环境下视觉目标跟踪研究现状及发展[J]. 航空兵器,2024, 31(03): 40-50.

[4] 郭庆梅, 刘宁波, 王中训, 等. 基于深度学习的目标检测算法综述[J]. 探测与控制学报, 2023, 45(06): 10-20+26.

[5] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2001, 1: I-511-I-518.

[6] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2005, 1: 886-893.

[7] Felzenszwalb P F, Girshick R B, McAllester D. Cascade object detection with deformable part models[C]//Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2010: 2241-2248.

[8] Liu L, Ouyang W, Wang X, et al. Deep learning for generic object detection: A survey[J]. International Journal of Computer Vision (IJCV). 2020, 128: 261-318.

[9] 杨锋, 丁之桐, 邢蒙蒙, 等. 深度学习的目标检测算法改进综述[J]. 计算机工程与应用, 2023, 59(11): 1-15.

[10] 钱承武. 基于深度学习的目标检测算法研究进展[J]. 无线通信技术, 2022, 31(04): 24-29.

[11] 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 (CVPR). 2014: 580-587.

[12] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision (IJCV). 2010, 88: 303-338.

[13] 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 (TPAMI). 2015, 37(9): 1904-1916.

[14] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015: 1440-1448.

[15] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence

(TPAMI). 2016, 39(6): 1137-1149.

[16] Vu T, Jang H, Pham T X, et al. Cascade rpn: Delving into high-quality region proposal network with adaptive convolution[J]. Advances in Neural Information Processing Systems (NIPS). 2019, 32.

[17] 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 (CVPR). 2016: 779-788.

[18] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2016: 21-37.

[19] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[20] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 7263-7271.

[21] Farhadi A, Redmon J. Yolov3: An incremental improvement[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018, 1804: 1-6.

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

[23] Jocher G, Nishimura K, Minerva T, et al. YOLOv5[EB/OL]. (2022-11-22)[2025-3-10]. https://github.com/ultralytics/yolov5.

[24] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[25] Li C, Li L, Jiang H, et al. YOLOv6: A single-stage object detection framework for in dustrial applications[J]. arXiv preprint arXiv:2209.02976, 2022.

[26] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 7464-7475.

[27] Nummiaro K, Koller-Meier E, Van Gool L. An adaptive color-based particle filter[J]. Image and Vision Computing (IVC). 2003, 21(1): 99-110.

[28] Du K, Ju Y, Jin Y, et al. Object tracking based on improved MeanShift and SIFT[C]//2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet). 2012: 2716-2719.

[29] Exner D, Bruns E, Kurz D, et al. Fast and robust CAMShift tracking[C]//Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2010: 9-16.

[30] Sun D, Roth S, Black M J. Secrets of optical flow estimation and their principles[C]// Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). 2010: 2432- 2439.

[31] Cheong B, Zhou J, Waslander S. JDT3D: Addressing the gaps in LiDAR-based tracking-by-

attention[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2024: 161-177.

[32] Zhou X, Koltun V, Krähenbühl P. Tracking objects as points[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2020: 474-490.

[33] Zhang Y, Wang C, Wang X, et al. FairMOT: On the Fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision (IJCV). 2021, 129(11): 3069-3087.

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

[35] Yu J, Jiang Y, Wang Z, et al. Unitbox: An advanced object detection network[C]//Proceedings of the 24th ACM International Conference on Multimedia. 2016: 516-520.

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

[37] Du Y, Zhao Z, Song Y, et al. Strongsort: Make deepsort great again[J]. IEEE Transactions on Multimedia (TMM). 2023: 1763-1773.

[38] Wang Z, Zheng L, Liu Y, et al. Towards real-time multi-object tracking[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2020: 107-122.

[39] Evangelidis G D, Psarakis E Z. Parametric image alignment using enhanced correlation coefficient maximization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI). 2008, 30(10): 1858-1865.

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

[41] 金列俊, 詹建明, 陈俊华, 等. 基于一维卷积神经网络的钻杆故障诊断[J]. 浙江大学学报(工学版), 2020, 54(03): 467-474.

[42] Wang K X, Zou J H, Zhou Y, et al. Few-shot image semantic segmentation with prototype alignment[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 27-28.

[43] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2020, 34(07): 12993-13000.

[44] Wu Z, Shen C, Van Den Hengel A. Wider or deeper: Revisiting the resnet model for visual recognition[J]. Pattern Recognition (PR). 2019, 90: 119-133.

[45] Luo H, Jiang W, Gu Y, et al. A strong baseline and batch normalization neck for deep person reidentification[J]. IEEE Transactions on Multimedia, 2019, 22(10): 2597-2609.

[46] 刘慧, 张礼帅, 沈跃, 等. 基于改进 SSD 的果园行人实时检测方法[J]. 农业机械学报, 2019, 50(04): 29-35-101.

[47] 葛雯, 史正伟. 改进 YOLOV3 算法在行人识别中的应用[J]. 计算机工程与应用, 2019, 55(20): 128-133.

[48] 陈一潇, 阿里甫·库尔班, 林文龙,等. 面向拥挤行人检测的 CA-YOLOv5[J]. 计算机工程与应用, 2022,58(09): 238-245.

[49] Hou Qibin, Zhou Daquan, Feng Jiashi. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 13713-13722.

[50] 徐芳芯, 樊嵘, 马小陆. 面向拥挤行人检测的改进 YOLOv7 算法[J]. 计算机工程, 2024, 50(03): 250-258.

[51] Zhu X, Hu H, Lin S, et al. Deformable convnets v2: More deformable, better results[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 9308-9316.

[52] Zhou K, Tong Y, Li X, et al. Exploring global attention mechanism on fault detection and diagnosis for complex engineering processes[J]. Process Safety and Environmental Protection, 2023, 170: 660-669.

[53] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 3-19.

[54] Feng C, Zhong Y, Gao Y, et al. Tood: Task-aligned one-stage object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 3490-3499.

[55] Dai X, Chen Y, Xiao B, et al. Dynamic head: Unifying object detection heads with attentions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 7373-7382.

[56] 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 (IJCV). 2015, 111: 98-136.

[57] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]// Proceedings of the European Conference on Computer Vision (ECCV). 2020: 213-229.

[58] Wang C, He W, Nie Y, et al. Gold-YOLO: Efficient object detector via gather-and-distribute mechanism[J]. Advances in Neural Information Processing Systems (NIPS). 2023, 36: 51094-51112.

[59] Zhao Z, He P. YOLO-Mamba: object detection method for infrared aerial images[J]. Signal, Image and Video Processing (ISVP). 2024, 18(12): 8793-8803.

[60] Wang C Y, Yeh I H, Mark Liao H Y. Yolov9: Learning what you want to learn using programmable gradient information[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2024: 1-21.

[61] Wang A, Chen H, Liu L, et al. Yolov10: Real-time end-to-end object detection[J]. Advances in Neural Information Processing Systems (NIPS). 2025, 37: 107984-108011.

[62] Khanam R, Hussain M. Yolov11: An overview of the key architectural enhancements[J]. arxiv preprint arxiv:2410.17725, 2024.

[63] Zhang L, Wu X, Zhang S, et al. Branch-cooperative osnet for person re-identification[J]. arxiv preprint arxiv:2006.07206, 2020.

[64] Zhou K, Yang Y, Cavallaro A, et al. Omni-scale feature learning for person re-identification[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2019: 3702-3712.

[65] Park H, Ham B. Relation network for person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2020, 34(07): 11839-11847.

[66] Sun Y, Zheng L, Yang Y, et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 480-496.

[67] Zhang H, Xu C, Zhang S. Inner-iou: more effective intersection over union loss with auxiliary bounding box[J]. arxiv preprint arxiv:2311.02877, 2023.

[68] Wang X, He Z, Liu C, et al. CGA-UNet: Category-guide attention U-Net for dental abnormality detection and segmentation from dental-maxillofacial images[J]. IEEE Transactions on Instrumentation and Measurement (TIM). 2023, 72: 1-11.

[69] Yoon K, Gwak J, Song Y M, et al. Oneshotda: Online multi-object tracker with one-shot-learning-based data association[J]. IEEE Access, 2020, 8: 38060-38072.

[70] Zhang Y, Sun P, Jiang Y, et al. Bytetrack: Multi-object tracking by associating every detection box[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2022: 1-21.

[71] Cao J, Pang J, Weng X, et al. Observation-centric sort: Rethinking sort for robust multi-object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023: 9686-9696.

[72] Li T, Li Z, Mu Y, et al. Pedestrian multi-object tracking based on YOLOv7 and BoT-SORT[C]//Third International Conference on Computer Vision and Pattern Analysis (ICCPA). 2023, 12754: 369-374.

[73] Maggiolino G, Ahmad A, Cao J, et al. Deep oc-sort: Multi-pedestrian tracking by adaptive re-identification[C]//2023 IEEE International Conference on Image Processing (ICIP). 2023: 3025-3029.

附录

附录 1 符号及缩略语的说明

英文缩写	英文全称	中文全称
BC-OSNet	Branch-Cooperative OSNet	—
BN	Batch Normalization	批归一化
CA	Coordinate Attention	坐标注意力
CBAM	Convolutional Block Attention Module	—
CNN	Convolutional Neural Network	卷积神经网络
Conv	Convolution	卷积
CPU	Central Processing Unit	中央处理器
CSP	Cross Stage Partial	—
CUDA	Compute Unified Device Architecture	—
DCN	Deformable ConvNets	可变形卷积网络
DPM	Deformable Parts Model	可变形部件模型
ECC	Enhanced Correlation Coefficient	增强相关系数
EMA	Exponential Moving Average	指数移动平均
Fast R-CNN	Fast Region-based CNN	—
Faster R-CNN	Faster Region-based CNN	—
FN	False Negative	—
FP	False Positive	—
FPN	Feature Pyramid Network	特征金字塔网络
GAM	Global Attention Mechanism	全局注意力机制
GAP	Global Average Pooling	全局平均池化
GCP	Global Contrastive Pooling	全局对比池化
GeM	Generalized-Mean	广义均值池化
GMP	Global Maximum Pooling	全局最大池化

英文缩写	英文全称	中文全称
GPU	Graphics Processing Unit	图形处理器
HOG	Histogram of Oriented Gradients	方向梯度直方图
HOTA	Higher Order Tracking Accuracy	高阶跟踪精度
IDSw	Identity Switches	身份切换次数
IoU	Intersection over Union	交并比
JDT	Joint Detection and Tracking	联合检测跟踪
KF	Kalman Filter	卡尔曼滤波
LAM	Layer Attention Mechanism	层注意力机制
mAP	mean Average Precision	平均精度率
MOT	Multiple Object Tracking	多目标跟踪
MOTA	Multiple Object Tracking Accuracy	多目标跟踪精度
MLP	Multilayer perceptron	多层感知器
NMS	Non-Maximum Suppression	非极大值抑制
OSBlock	Omni-Scale Residual Block	全尺度残差模块
OSNet	Omni-Scale Network	全尺度网络
P	Precision	精确率
PAN	Path Aggregation Network	路径聚合网络
PCB	Part-based Convolutional Baseline	—
R	Recall	召回率
R-CNN	Regions with CNN features	区域卷积神经网络
ROI	Region of Interest	感兴趣区域
RPN	Region Proposal Network	区域建议网络
SE-Net	Squeeze-and-Excitation Networks	—
SORT	Simple Online and Realtime Tracking	—
TBD	Tracking-by-Detection	基于检测跟踪

英文缩写	英文全称	中文全称
SPPF	Spatial Pyramid Pooling - Fast	快速空间金字塔池化
SPPNet	Spatial Pyramid Pooling Network	空间金字塔池化网络
SSD	Single Shot multi-box Detector	—
SVM	Support Vector Machines	支持向量机
TADDH	Task Align Dynamic Detection Head	动态任务对齐检测头
TP	True Positive	—
YOLO	You Only Look Once	—

攻读硕士学位期间的研究成果

1 学术论文

[1] 桑建斌, 张梦雯, 杨思远, 等. 改进 YOLOv8n 的复杂场景行人检测算法[J]. 传感技术学报（已录用）.

[2] Sang J, Zhang M, Yang H, et al. YOLOv8-GDI: A lightweight YOLOv8 for real-time Pedestrian Detection[C]//2024 7th International Conference on Robotics, Control and Automation Engineering (RCAE). IEEE, 2024: 422-428.

致谢

行文至此，落笔为终。回首三载求学路，虽道阻且长，然幸得良师益友相伴、家人倾力支持，方使此篇论文得以成稿。此刻心怀感念，谨以寥寥数语，致谢所有助我前行之人。

首先，衷心感谢我的导师杨会成教授。自大三初聆先生讲授《信号与系统》，严谨治学之风范与妙趣横生之教法便如磁石相引。及至拜入门下读研，五载春秋承蒙教诲。先生学识渊博，治学严谨，从论文选题、实验设计至成果凝练，无不倾注心血。研究困顿之际，先生以“不谋全局者，不足谋一域”之格局点拨迷思；数据偏差之时，又以“见微知著，砥志研思”之严谨导我溯源。从学术思维之塑形到处世之道之启迪，先生皆倾注灼灼心血。更难忘生活迷惘时，先生以"静水流深"之智慧拨云见日。师者如光，虽微致远，此等言传身教必将永镌心碑。

此外，特别感谢实验室的胡耀聪、徐可、朱文博三位老师，他们以亦师亦友之风范润物无声。学术研讨时谆谆善诱，茶歇闲谈间妙语解颐，诸师既授我以专业之能，更示我以治学之乐。实验室昼夜流转的光影里，始终跃动着你们智慧的火花与温暖的守望。

血脉情深，父母之恩山高水长。二十余载求学路，双亲青丝染霜而初心不改。电话那头的殷殷嘱托，归家时灶间的氤氲饭香，低谷时的温言鼓励，皆化作砥砺前行的底气。寸草之心，难报春晖，惟愿此文能为二老鬓间霜雪添染几分欣慰。

同时，感谢我的同门、师兄师姐、师弟师妹以及室友们，愿今后我们都拥有一个美好的未来，在各自的领域里发光发热。

最后，向参与论文评审与答辩的各位专家致以谢意，感谢你们在百忙之中拨冗指导。今当别离，临书惘然。愿将此文化作时光标本，封存师友殷切目光，凝结父母鬓间霜雪，铭记同窗笑泪时光。前路虽远，幸有诸君馈赠之勇气常伴；学术无涯，谨怀此刻感恩之心笃行。

桑 建 斌

2025 年 5 月 7 日于芜湖