

分类号: U471.15

密 级: 公开

学校代码: 10363

学 号: 2220110157



安徽工程大学

Anhui Polytechnic University

硕士学位论文

题 目 轻量级密集行人检测
与跟踪算法研究

论 文 作 者 欧洪伟

校 内 导 师 张荣芸 (教授)

校 外 导 师 赵夕长

专业学位类别 机械

研究方向 (领域) 机器学习及图像识别

论文提交日期: 2025 年 5 月 20 日

分类号：U471.15

学校代码：10363

密 级：公开

学 号：2220110157

轻量级密集行人检测与跟踪算法研究

RESEARCH ON LIGHTWEIGHT DENSE
PEDESTRIAN DETECTION AND TRACKING
ALGORITHM

学生姓名：欧洪伟

校内导师：张荣芸（教授）

专 业：机械

研究方向：机器学习及图像识别

论文答辩日期：2025 年 5 月 10 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：欧洪伟
日期：2025 年 5 月 10 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：欧洪伟

日期：2025 年 5 月 10 日

指导教师签名：张荣芸

日期：2025 年 5 月 10 日

轻量级密集行人检测与跟踪算法研究

摘 要

在计算机视觉领域，行人检测与跟踪技术是一个关键研究领域，它在自动驾驶、智能监控、国防和人机交互等众多应用中发挥着重要作用。随着深度学习技术的进步，基于这一技术的算法在行人检测和跟踪领域得到了广泛应用。然而，这些技术仍面临着多重挑战。例如，在人群密集的道路环境中，当前算法的检测效率和速度较低，影响了跟踪的准确性。此外，在变化的天气和复杂光照条件下，行人检测的准确性通常会降低。大多数现有的目标检测算法还需针对不同的数据集调整先验框尺寸的超参数，这不仅增加了模型训练的复杂性，也影响了检测精度。为应对这些问题，本文提出了一个结合改进后的 YOLOv8 和 Deep Sort 技术的新算法模型，旨在提高行人检测与跟踪的效率和精度，具体研究内容包括：

在模型检测部分，面对复杂背景中密集行人的漏检和检测不精准的问题，本文采用了 BiFormer 自适应注意力机制结合小目标检测技术来增强语义处理，这种方法显著增强了模型对小尺寸目标的识别能力，尤其是在特征较为模糊的行人图像中。此外，为了提升模型的检测精度和效率，本文对模型的主干结构进行了轻量化设计，使用了更为轻便的 GhostNet 网络替代原有的特征提取网络，这一改变不仅减少了模型的参数量，也提高了处理速度。在损失函数的优化上，本文调整了回归损失函数以增加定位损失的比重，并将 CIOU 损失函数替

换为定位精度更高的 EIOU 损失函数，从而提升了整体的检测效率。实验结果表明，改进后的算法在 Crowd Human 和 Caltech Pedestrian 两个行人数据集上表现出色，平均精度分别提升至 90.95% mAP 和 96.10% mAP，处理速度分别能达到 60.28 FPS 和 66.49 FPS，显示出在复杂环境下密集行人的有效检测和更快的响应速度。与原始算法相比，其检测性能显著提高。在相同实验条件下，本文算法与目前主流的一些检测算法的检测效果相比，仍有不错的优势，显示了强大的竞争力，证明改进算法在小尺度行人检测任务中有广泛应用前景。

在跟踪技术方面，针对 Deep Sort 算法在处理复杂场景中的密集行人跟踪时存在的精度低和速度慢的问题，本文对该算法进行了重要的改进。首先，本文将传统的卡尔曼滤波中的匀速线性模型替换为匀加速模型，这一改变使得模型能更真实地模拟行人在实际环境中的运动轨迹。其次，为了提升跟踪速度，本文采用了一个更加轻量级的重识别模型，这样不仅减少了跟踪器的数量，还降低了对处理速度的影响。对改进后的算法性能评估是在 MOT16 数据集上进行的，YOLOv8 被用作检测器，同时自适应的卡尔曼滤波被采用。实验结果表明，与原始算法相比，MOTA 指标从 18.9% 提升至 28.6%，IDF 指数从 26.4% 提升至 30.6%，同时多目标跟踪 (MT) 性能提高了 5.2%，误跟踪 (ML) 减少了 2.1%。这些改进证明了算法在复杂场景中跟踪多目标行人时的精确性和实时性，同时展示了其优良的鲁棒性。

关键词：目标检测，YOLOv8，Deep Sort，多目标跟踪，行人检测与跟踪

RESEARCH ON LIGHTWEIGHT DENSE PEDESTRIAN DETECTION AND TRACKING ALGORITHM

ABSTRACT

In the field of computer vision, pedestrian detection and tracking technology is a key research area, which plays an important role in many applications such as autonomous driving, intelligent surveillance, defense, and human-computer interaction. With the progress of deep learning technology, algorithms based on this technology have been widely used in the field of pedestrian detection and tracking. However, these technologies still face multiple challenges. For example, in a crowded road environment, the detection efficiency and speed of the current algorithm are low, which affects the accuracy of tracking. In addition, the accuracy of pedestrian detection is often reduced in changing weather and complex lighting conditions. Most existing object detection algorithms need to adjust the hyperparameters of the prior frame size for different data sets, which not only increases the complexity of model training, but also affects the detection accuracy. To address these problems, this paper

proposes a new algorithm model combining improved YOLOv8 and Deep Sort technology, aiming to improve the efficiency and accuracy of pedestrian detection and tracking. Specific research contents include:

In the detection part of the model, faced with the problem of missing detection and inaccurate detection of dense pedestrians in complex background, this paper adopts BiFormer adaptive attention mechanism combined with small target detection technology to enhance semantic processing. This method significantly enhances the model's recognition ability of small-size targets, especially in pedestrian images with fuzzy features. In addition, in order to improve the detection accuracy and efficiency of the model, this paper carried out lightweight design on the backbone structure of the model, and used a lighter GhostNet network to replace the original feature extraction network. This change not only reduced the number of parameters in the model, but also improved the processing speed. In terms of the optimization of the loss function, this paper adjusted the regression loss function to increase the proportion of positioning loss, and replaced the CIOU loss function with the EIOU loss function with higher positioning accuracy, thus improving the overall detection efficiency. Experimental results show that the improved algorithm performs well on two Pedestrian data sets, Crowd Human and Caltech Pedestrian, with the average accuracy increased to 90.95% mAP and 96.10% mAP, and the processing speed can reach 60.28 FPS and

66.49 FPS, respectively. It shows effective detection and faster response speed of dense pedestrians in complex environments. Compared with the original algorithm, the detection performance is significantly improved. Under the same experimental conditions, the proposed algorithm still has a good advantage compared with some current mainstream detection algorithms, showing strong competitiveness, which proves that the improved algorithm has a wide application prospect in small-scale pedestrian detection tasks.

In terms of tracking technology, aiming at the problems of low accuracy and slow speed of Deep Sort algorithm when dealing with dense pedestrian tracking in complex scenes, this paper makes important improvements to the algorithm. Firstly, the uniform linear model of Kalman filter is replaced by the uniform acceleration model, which makes the model more realistic to simulate the pedestrian trajectory in the actual environment. Secondly, in order to improve the tracking speed, this paper adopts a more lightweight heavy recognition model, which not only reduces the number of trackers, but also reduces the impact on the processing speed. The performance evaluation of the improved algorithm was performed on MOT16 datasets, YOLOv8 was used as the detector, and adaptive Kalman filtering was used. Experimental results show that compared with the original algorithm, MOTA index is improved from 18.9% to 28.6%, IDF index is improved from 26.4% to 30.6%,

multi-target tracking (MT) performance is improved by 5.2%, and mistracking (ML) is reduced by 2.1%. These improvements prove the accuracy and real-time performance of the algorithm in tracking multi-target pedestrians in complex scenes, and demonstrate its excellent robustness.

KEY WORDS: target detection, Yolov8, Deep Sort, multi-target tracking, pedestrian detection and tracking

目 录

第一章 绪论	1 -
1.1 研究背景与意义	1 -
1.2 研究现状	2 -
1.2.1 目标检测研究的现状分析	2 -
1.2.2 目标跟踪研究的现状分析	7 -
1.3 论文研究的难点和重点	10 -
1.3.1 论文研究的难点	10 -
1.3.2 论文研究的重点	11 -
1.4 论文研究内容及章节安排	11 -
1.4.1 研究内容	11 -
1.4.2 论文章节安排	12 -
第二章 行人检测与跟踪相关理论概述	15 -
2.1 深度学习相关基础	15 -
2.2.1 输入层	19 -
2.2.2 卷积层	19 -
2.2.3 池化层	19 -
2.2.4 全连接层	19 -
2.2.5 激活和损失函数	22 -
2.2.6 非极大值抑制算法	26 -
2.2.7 聚类算法	27 -
2.3 常用的测试数据集	28 -
2.3.1 常用行人检测数据集	28 -
2.3.2 常用跟踪数据集	31 -
2.4 本章小结	31 -
第三章 基于 Biformer 和改进的 YOLOv8 的行人检测与识别	33 -
3.1 网络架构分析	33 -
3.2 基于注意力机制改进的 YOLOv8 算法	35 -
3.2.1 注意力机制介绍	35 -

3.2.2 改进思路	39 -
3.3 基于轻量化网络改进的 YOLOv8 算法	39 -
3.3.1 Mobile Net 网络	40 -
3.3.2 Ghost Net 网络	43 -
3.3.3 改进方法	44 -
3.4 基于损失函数改进的 YOLOv8 算法	45 -
3.4.1 目标检测损失函数	45 -
3.4.2 改进方法	47 -
3.5 实验结果及分析	49 -
3.5.1 实验环境与参数设置	49 -
3.5.2 数据集准备	49 -
3.5.3 实验评价指标	50 -
3.5.4 模型的超参数	51 -
3.5.5 实验结果分析	52 -
3.6 本章小结	56 -
第四章 基于改进的 Deep Sort 行人跟踪模型	57 -
4.1 网络架构分析	57 -
4.1.1 跟踪流程分析	57 -
4.1.2 运动模型分析	59 -
4.1.3 特征提取分析	60 -
4.1.4 级联匹配分析	61 -
4.2 改进思路	62 -
4.2.1 改进匀加速卡尔曼滤波模型	62 -
4.2.2 改进行人重识别模型	64 -
4.3 实验结果与分析	66 -
4.3.1 数据集准备	66 -
4.3.2 评价指标	66 -
4.3.3 实验结果分析	66 -
4.4 本章小结	69 -

第五章 总结与展望	- 71 -
5.1 总结	- 71 -
5.2 不足与展望	- 71 -
参考文献	- 73 -
攻读学位期间发表的学术论文目录	- 73 -

第一章 绪论

1.1 研究背景与意义

近年来,随着人工智能与城市化的深度融合,智能安防、自动驾驶、智慧城市管理等场景对密集行人感知技术的需求日益迫切^[1-3]。传统视频监控系统依赖人工分析,存在效率低、响应慢等问题,而基于深度学习的目标检测与跟踪技术为实时感知与决策提供了新的解决路径^[4]。然而,城市密集场景中行人目标尺度差异大、遮挡频繁、运动轨迹复杂等特性,导致现有算法在检测精度、跟踪稳定性及计算效率上面临显著瓶颈,亟需通过技术创新突破实际应用中的关键挑战^[5-6]。

在目标检测领域,主流算法虽在开放场景中表现优异,但在高密度行人场景下仍存在局限性。例如,微小行人目标因像素占比低导致特征表达模糊,传统特征金字塔网络难以有效捕捉细节信息,造成漏检问题^[7];城市环境中广告牌、车辆等背景干扰物与行人外观相似,现有注意力机制缺乏动态聚焦能力,易引发误检现象。此外,检测模型参数量庞大,难以适配边缘设备的实时性需求,限制了技术在实际场景中的部署应用^[8-9]。

多目标跟踪领域同样面临严峻考验。传统运动预测模型基于匀速假设,与行人实际运动中的急停、变向等非线性行为存在偏差,导致轨迹预测误差累积和身份切换问题;重识别模型的高计算复杂度在目标数量激增时显著增加系统延迟,而长时间遮挡场景下的轨迹恢复能力不足,进一步降低了跟踪鲁棒性^[10-11]。这些问题使得现有算法在大型活动安防、交通枢纽管理等高密度场景中的实用性大打折扣^[12]。

因此,本文针对密集行人感知的核心问题提出系统性解决方案。首先在理论层面:构建动态特征增强与轻量化网络协同优化的新框架,突破微小目标检测与复杂背景干扰的固有技术瓶颈;提出融合非线性运动建模与高效特征匹配的跟踪范式,完善多目标跟踪的时空一致性理论体系,为密集场景感知提供创新方法支撑。其次在技术层面:开发检测—跟踪一体化轻量模型,在降低计算资源消耗的同时提升算法鲁棒性,为嵌入式设备部署提供可行性方案;设计遮挡场景下的轨

迹连续性保持机制，有效解决密集人流中的身份混淆难题。最后在应用层面：研究成果可赋能智能安防、自动驾驶、城市治理等关键领域——在公共安全中提升人群聚集风险预警能力，在交通系统中增强车辆对行人轨迹的预判可靠性，在智慧城市建设中为客流分析与应急管理提供实时数据支持，对推动城市数字化转型、保障社会公共安全具有重要价值。

此外，本文通过理论探索与工程实践的深度融合，不仅拓展了计算机视觉技术在城市复杂场景中的应用边界，更为智能系统的高效落地提供了兼具创新性与实用性的技术支撑，助力人工智能技术在智慧城市建设的规模化应用与可持续发展。

1.2 研究现状

目标检测与跟踪已成为计算机视觉领域的重要研究方向，世界各地的顶尖学府，如麻省理工学院、清华大学等，均在此领域展开了深入研究并取得显著成就。随着深度学习技术的应用，基于检测的多目标跟踪技术实现了多项重大突破。多目标跟踪技术主要涉及目标的准确检测和有效的数据关联处理。本节将对目标检测、多目标跟踪技术的国内外研究进展和现状进行充分的介绍和分析。

1.2.1 目标检测研究的现状分析

目标检测在计算机视觉中占据核心地位，它包括在图像或视频中识别和定位人们关注的对象。该过程包括分类，即判断输入图像中是否存在目标物体，以及目标定位，通过使用外接矩形框在图像中标出对象的位置。目标检测技术的进展主要分为两个阶段：首先是传统目标检测技术的阶段，其次是深度学习驱动的目标检测技术阶段^[13-15]。

1.2.1.1 基于传统的目标检测算法

在传统的目标检测方法中，整个处理流程通常分为三个主要步骤：区域选择、特征提取和分类^[16-18]。区域选择涉及在整张图像中识别并选定可能包含目标的区域。这一步骤是目标检测流程的起点，其目的是缩小后续处理范围，提高效率。接着在特征提取阶段，对上述选定的候选区域进行更深入的分析，通过各种算法提取出能够代表这些区域的关键特征。在分类阶段，将提取出的特征送入分类器。分类器根据这些特征来判断候选区域是否真正包含目标，并进一步确定目标的类

别。

1.2.1.2 基于深度学习的目标检测

由于传统检测法在复杂环境下检测精度较低，模型的鲁棒性也较差。相比之下，基于深度学习的目标检测算法则能够更好地适应这些复杂环境，通过高效的特征提取和识别，显著提升了检测的鲁棒性。基于深度学习的目标检测技术主要分为两类：单阶段目标检测与两阶段目标检测^[19-20]。单阶段方法直接预测目标的类别和位置，优势在于速度快，适用于需要实时处理的应用场景。两阶段方法首先生成潜在目标的区域建议，然后对这些区域进行更精细的分类和定位，优势在于精度高，尤其适用于那些对检测精度要求较高的场合。这些基于深度学习的方法相较于传统方法，在处理具有挑战性的检测任务时，能够提供更为可靠的性能。算法的具体流程如图 1-1 所示。

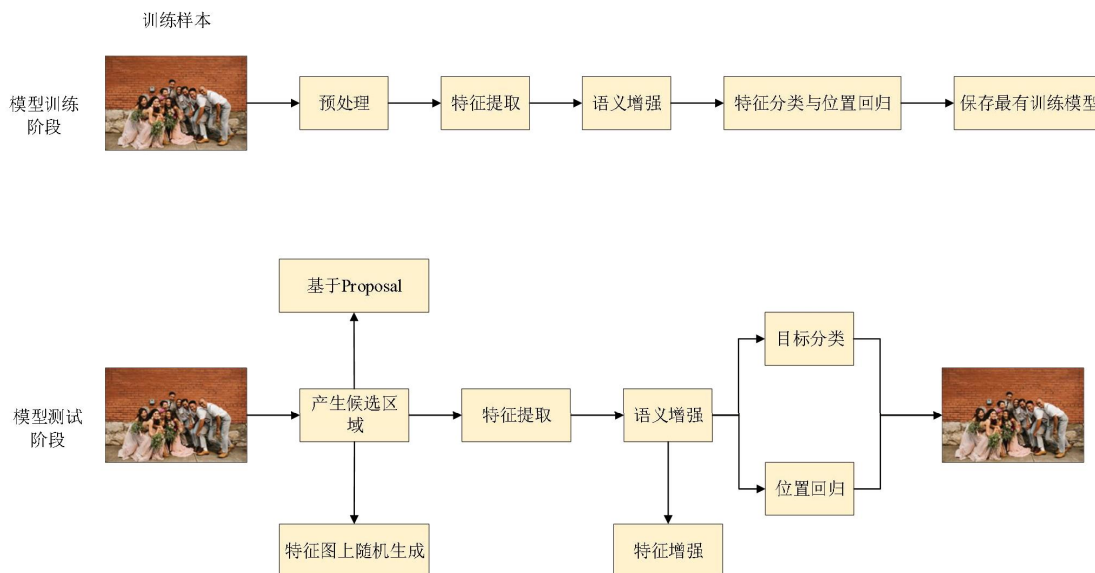


图 1-1 以深度学习为基础的目标检测算法基本流程

(1) 单阶段目标检测算法

单阶段目标检测算法的主要代表算法有 SSD(Single Shot MultiBox Detector) 和 YOLO(You Only Look Once)系列算法等^[21-30]，该类算法的核心特点在于能直接在特征图上预测位置坐标、置信度和目标类别^[31-32]，实现一步直接输出最终结果，能大幅提升处理速度。这种处理方式简化了目标检测流程，大大降低了检测所需时间成本，使其更适用于需要快速响应的应用场景。

SSD 检测算法：SSD(Single Shot MultiBox Detector)算法采用直接回归方式，属于单阶段目标检测技术^[33]。其核心优势在于完全摒弃了传统的候选区域建议过

程^[34],可以在初始输入图片上将目标位置和分类信息直接预测出来,从而显著提升处理速度。将多尺度特征输出机制^[35]加入到模型之中也是其最为明显的特点。这表明 SSD 不止能在单一尺度特征图上进行预测,还能利用多个尺度的特征图来增强对不同大小目标的检测能力,使得 SSD 能够更加精确有效处理各种尺寸的目标。

SSD 算法通过一系列卷积层来处理图像,每一层都能产生一组包括目标类别和位置的预测结果。这种分层的特征提取方式,结合了深层和浅层信息,使 SSD 在保持高检测速度的同时,也确保了较高的检测精度,尤其适用于需要快速而准确反馈的场景,其算法的框架如图 1-2 所示。

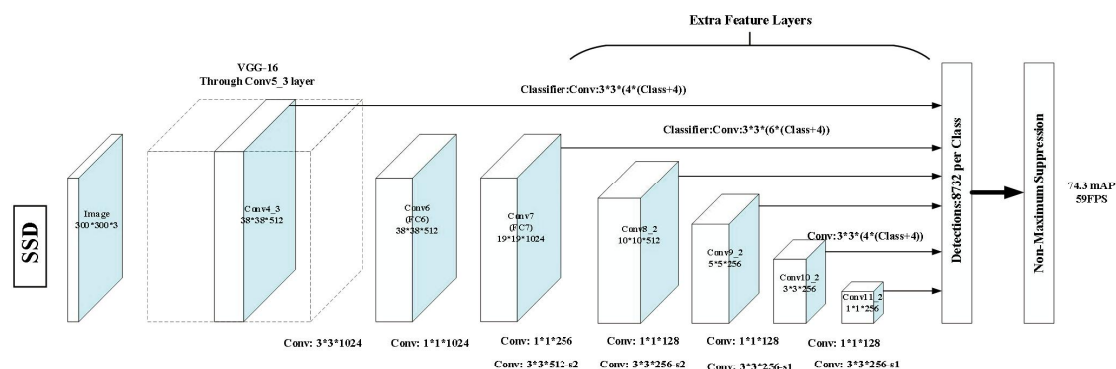


图 1-2 SSD 基本算法框架

如图 1-2 所示,在 SSD 算法中,首先将广泛应用于图像处理领域的深度卷积网络 VGG16^[36]作为其基本骨干网络来提取特征,获取不同尺度的特征图。再从 VGG16 获取的每个特征图上,设定不同尺寸的锚点框(Anchors)^[37]。每个锚点框都对应一个潜在的目标检测区域。这些锚点框覆盖了多种尺寸和比例,使得模型能够检测到不同大小的对象。然后,将每个特征图和其相应的锚点框输入到一个专门的目标检测与分类网络中,以识别出每个锚点框中是否包含目标,同时还对存在的目标进行分类,并精确定位它们的边界。最终,实现 SSD 能够同时在多个尺度上进行有效的目标检测和分类,使其检测覆盖范围和精确度得到提高。

YOLO 系列检测算法:YOLO 系列检测算法的推出无疑为目标检测领域带来了革命性的变化,与之前基于滑动窗口或候选区域生成方法相比,这种新方法简化了候选框的提取过程,通过直接预测边界框的方式简化了检测流程^[38]。初始输入图片被划分成一个个正方形网格,在输出层一次性完成位置回归和类别预测,这种方法有效平衡了实时性和准确性。表 1-1 列出 YOLO 系列最常用的几种算法

的优缺点。

表 1-1 以上 YOLO 算法的优缺点对比

算法模型	优势	弊端
YOLOv1	检测速度快	检测精度低，效果差
YOLOv2	分类精度提高	预训练迁移困难
YOLOv3	实现多尺度检测	对较大目标检测效果差
YOLOv4	平衡了检测精度与速度	虽然达到了较好的平衡，但检测精度有待提高。
YOLOv5	模型尺寸较小，检测速度较快	尽管检测速度快，整体性能仍需优化。
YOLOv8	较好的检测精度和鲁棒性，以及对小目标检测的效果较好	模型较大，需要较大的训练数据

通过表 1-1 的各算法对比,本文将采用改进后的 YOLOv8 算法来完成行人目标检测任务,并将在后面的第三章详细阐述。

(2)二阶段目标检测算法

二阶段目标检测算法的主要代表是 R-CNN 及其衍生系列算法^[39-42]。R-CNN 是一种基于深度学习的新型目标检测算法,该方法原理便是借助卷积神经网络(Convolutional Neural Networks, CNN)来对图像中的候选区域进行预处理^[43-44],显著提升了目标检测的精度。其核心思想在于使用 Region Proposal 方法来识别图像中的候选目标区域^[45]。具体来说,它将输入图像分割成多个大小不一的区域,并通过算法尝试合并包含相同物体的区域,最终产生大约 2000 个候选区域供后续处理。这些标准化后的候选区域随后被送入卷积神经网络(CNN)进行深入的特征提取。提取出的特征被转化为固定尺寸的输出,以保证它们可以被统一处理。这些输出特征随后由支持向量机(SVM)分类器进行分类处理^[46],以确定每个区域的物体类别。同时,算法还会根据目标当前的位置信息对之后的位置信息进行预测,这进一步提高了检测的准确性。

虽然相比传统算法 R-CNN 在精确度等性能方面都得到了较大提升,但其存在复杂的多步训练过程和高运算成本的不足。由于其在训练过程中会生成数量庞大的候选框去对目标位置信息进行预测,所以模型的训练和检测速度严重降低。不仅如此,由于必须在训练之前实现所有候选框的尺寸统一,这可能导致图像在恢复到原始尺寸时出现畸变,从而影响最终检测精度。此外,为了存储如此庞大数量的候选框,这也进一步加剧了内存空间资源的负担。

为了解决这些问题,有学者提出了 Fast R-CNN 模型,该模型是基于原有

R-CNN 改进而来。与 R-CNN 相比，Fast R-CNN 关键改进有：

1)端到端训练：Fast R-CNN 支持端到端的训练流程，避免了 R-CNN 中繁琐的多步骤训练方式。

2)特征提取与区域建议：在整个输入图像上使用卷积网络进行特征提取后，随即从所得的特征图中直接识别出候选区域，提高了处理速度。

3)ROI Pooling^[47]：引入了区域兴趣池化（ROI Pooling）层，该层能够将不同尺寸的候选框统一转换为固定大小的特征图，便于后续的预测处理。

4)分类器替换：将原先使用的 SVM 分类器替换为 softmax 分类器^[48]，softmax 分类器简化了模型结构的同时，也提升了分类的效率。

通过这些改进，Fast R-CNN 不仅改善了训练和检测的速度，还提升了模型的整体效率和精度，有效解决了 R-CNN 中存在的一些核心问题。

虽然 Fast R-CNN 在训练和检测速度方面有了显著改进，但该算法仍然依赖于选择性搜索方法来识别候选区域，这限制了模型速度的进一步提高。为了克服这一限制，通过利用区域建议网络(Region Proposal Network, RPN)^[49]直接在特征图上生成目标推荐区域来实现速度的提高，进而开发了 Faster R-CNN 模型。

Faster R-CNN 通过整合区域建议网络(RPN)，有效加速了候选区域的自动生成过程。RPN 利用卷积特征图来预测目标的边界和分数，不仅加快了候选区域的提取速度，而且提高了整体检测效率。该模型的网络结构如图 1-3 所示，通过这种集成方式，Faster R-CNN 不仅显著提升了检测速度，也提高了精度。

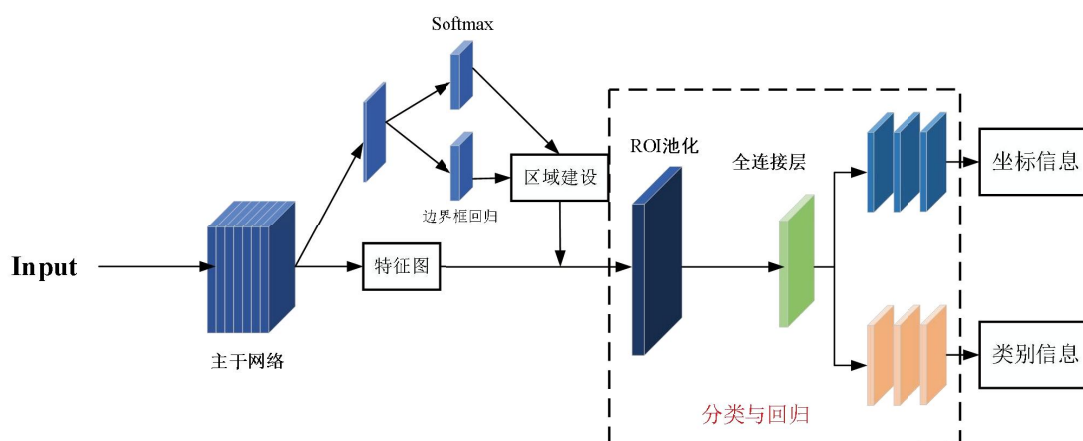


图 1-3 Faster R-CNN 网络结构

具体来说，Faster R-CNN 实现了高达 17FPS 的检测速度，并将检测精度提高到了 70.4%。这种性能的大幅提升使得 Faster R-CNN 在实时视频处理和复杂

图像识别任务中具有极大的应用价值，相比之前的模型，无论是速度还是准确度都有了质的飞跃。

在 Faster R-CNN 模型的基础上，何凯明等人进一步发展了实例分割领域，提出了 Mask R-CNN 模型^[50]。该模型引入了一个新的功能——实例分割，即在检测对象的同时，也将每个对象的精确像素级轮廓从背景中分割出来。其网络结构如图 1-4 所示。

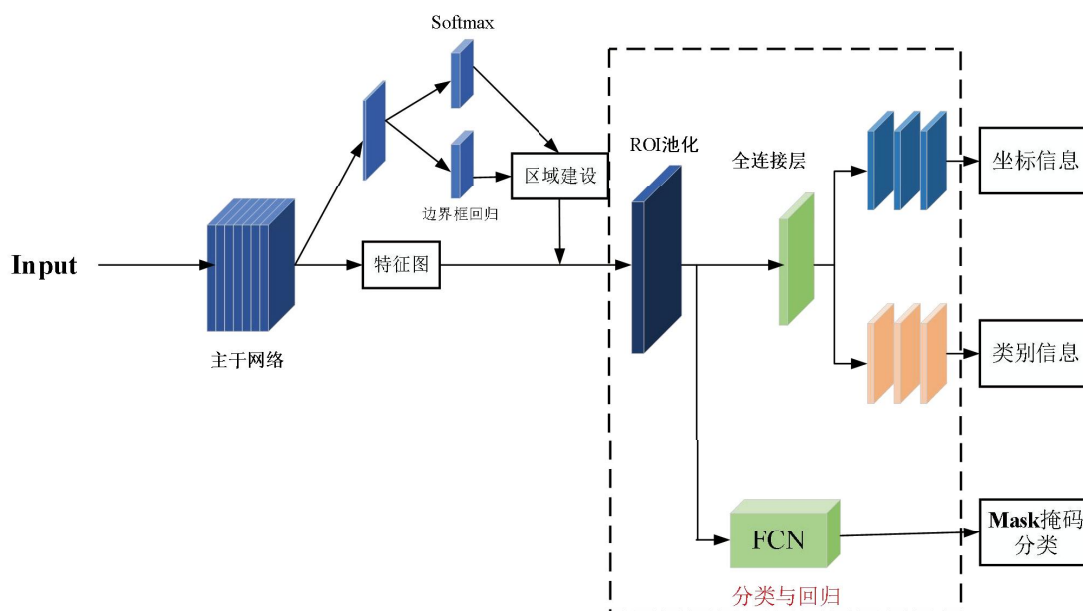


图 1-4 Mask RCNN 网络模型

Mask R-CNN 在 Faster R-CNN 的结构基础上，引入了一个新的 FCN 分支，主要用于生成目标的遮罩，这使得模型能够同时进行目标检测和像素级分割^[51]。这种能力极大扩展了模型的应用范围，使其能够在医学成像^[52]、自动驾驶车辆以及视频监控等领域发挥重要作用。

行人目标检测技术目前仍在快速发展中。虽然已经提出了众多理论和算法，但在密集行人场景下，现有检测方法仍面临以下关键挑战：1.微小目标特征表达不足：传统特征金字塔在跨层特征融合时易丢失微小行人的边缘与纹理信息，导致高密度场景中漏检率显著上升；2.复杂背景干扰抑制困难：现有注意力机制多依赖全局特征响应，难以动态聚焦局部关键区域，广告牌、植被等干扰物易引发误检；3.实时性与精度难以平衡：轻量化模型虽降低计算量，但特征压缩导致小目标检测性能下降，难以满足密集场景下高精度与低延时的双重需求。

1.2.2 目标跟踪研究的现状分析

目标跟踪算法根据跟踪对象的数量通常分为两种类型：单目标跟踪和多目标跟踪。考虑到本文的研究重点是在道路场景中对多个行人进行检测和跟踪，因此将不对单目标跟踪进行讨论。在多目标跟踪领域，可进一步划分为传统跟踪算法和基于深度学习的跟踪算法^[53-54]。在传统跟踪算法中，**Meanshift** 算法^[55]被广泛认为是一种经典方法；而在深度学习方法中，最广泛应用的则是 **SORT** 算法^[56]。接下来的讨论将主要围绕这两种算法展开，探讨它们的应用和效果。

1.2.2.1 Meanshift 跟踪算法

作为一种核心密度估计方法，**Meanshift** 跟踪算法通过对颜色特征进行核密度估计来定位局部最优点，因此在目标跟踪领域具有广泛的应用^[57]。由于该算法的计算复杂度较低，能够实现较快的跟踪速度。作为一种无参数的聚类算法^[58]，该算法常与其他跟踪技术结合使用，以增强跟踪的准确性。在实际应用中，该算法对目标的跟踪主要是利用对比不同目标特征之间的相似度来实现。具体操作是计算目标特征的相似函数，以此确定目标的 **Meanshift** 向量。然后，应用梯度下降方法逐步移动此向量，直至找到具有最高特征相似度的位置，即为目标在当前帧中的实际位置。

Meanshift 跟踪算法由于其计算量不大，可节约时间成本，从而能迅速追踪目标。算法中的核函数使其对于目标部分遮挡或形态变化具有良好的适应性。然而，当目标的尺寸变化显著或者移动速度加快时，均值漂移算法可能无法持续有效跟踪，导致跟踪失败的情况发生。

1.2.2.2 SORT 跟踪算法

SORT 算法是一种基于 **TBD(Tracking-by-Detecton)**跟踪策略的多目标跟踪方法^[59]。**SORT** 算法利用了 **CNN** 在目标检测方面的优越性能，使用 **Faster-RCNN** 框架来关联连续两帧中检测到的目标，展现出了良好的性能。但通过消融实验，其结果证明了该算法对多目标的跟踪效果过度依赖检测部分的准确性。

在目标检测环节，**SORT** 算法使用 **Faster-RCNN** 来识别和定位目标，主要分为两个步骤：首先，提取图像帧的特征；随后，利用区域候选框网络（**Region Proposal Network, RPN**）对这些特征进行分析，以进行精确的定位和分类。

在数据关联阶段，算法会计算当前帧下目标预测框与新检测框之间的交并比。这些交并比值构成一个代价矩阵，用于后续的匈牙利匹配算法^[60]，从而确保目标

在视频帧间的连续性。另外，算法对每个跟踪对象的描述是通过一个变量来表现的，这个变量集成了目标的位置、速度等关键信息。

$$X = [u, v, s, r, u', v', s', r'] \quad (1-1)$$

式中， (u, v) 表示目标中心的横纵坐标， (s, r) 表示边界框的尺寸和比例。相应地， (u', v', s', r') 用于表示预测的下一帧中的这些参数。该算法采用卡尔曼滤波(Kalman)^[61]来优化这一线性模型，从而更新目标的状态。SORT 算法的整个流程涉及几个关键步骤：首先通过目标检测和特征提取过程从视频帧中获取目标信息，然后跟踪器使用卡尔曼滤波进行轨迹预测。最终，利用匈牙利匹配算法将相邻帧的数据关联起来，以更新跟踪结果。

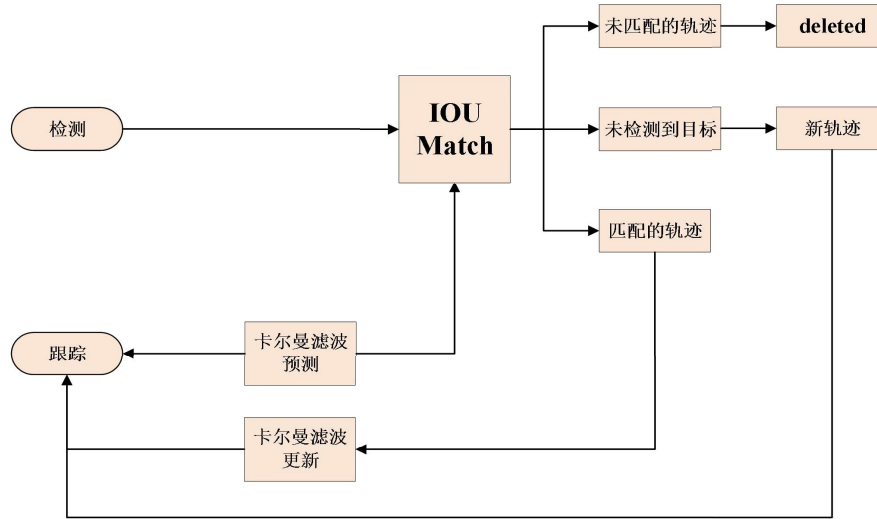


图 1-5 SORT 算法流程图

在 SORT 算法的框架下,图 1-5 描述了如何通过目标检测算法识别出边界框,配合卡尔曼滤波算法,预测和更新目标在视频帧中的位置和状态。SORT 算法利用检测框和轨迹集合间的 IOU(交并比)进行匹配^[62],这种方法能有效减少因目标遮挡导致的误判。IOU 被用作两个目标关联矩阵的基础,通过匈牙利匹配算法得到目标跟踪关联的连续性。当 IOU 值低于特定阈值时,认为两个框不属于同一目标。

匹配结果分为三种情况：1.未匹配的轨迹将被删除；2.未检测到目标的视为潜在新轨迹；3.成功匹配到的轨迹用卡尔曼滤波更新,在后续继续跟踪。这种方法确保了算法在动态环境中的适应性和准确性。

尽管取得一定进展,密集行人跟踪仍存在以下不足：1.非线性运动建模偏差：

基于匀速假设的卡尔曼滤波在行人急停、变向时预测误差累积，导致轨迹断裂与身份切换(ID Switch)^[63]；2.遮挡场景轨迹恢复能力弱：现有方法对短时遮挡依赖外观匹配，但在密集场景中行人外观相似度高，长时间遮挡(>3 秒)时轨迹恢复成功率不足；3.多目标协同计算效率低：传统级联匹配策略在目标数量激增时计算复杂度呈指数增长，难以满足高密度场景的实时性需求。

针对上述密集行人场景下存在的检测和跟踪问题，本文提出一种全新的轻量化行人检测与跟踪算法，从检测与跟踪协同优化的角度解决密集行人感知难题，主要创新如下：1.动态稀疏注意力机制：引入 Biformer 多粒度特征选择注意力机制，通过动态稀疏采样聚焦局部关键区域，抑制背景干扰的同时增强微小目标特征表达；2.轻量化特征提取网络 GhostNet：构建轻量级双向异构金字塔网络，结合深度可分离卷积与跨尺度残差连接，实现高效多尺度特征融合；3.非线性运动预测与重识别模型：提出多模态运动建模网络，融合惯性测量单元先验与视觉运动线索，构建行人加速度感知模型，优化行人重识别网络来提升遮挡场景恢复能力。

1.3 论文研究的难点和重点

1.3.1 论文研究的难点

随着深度学习的快速发展，对于行人目标的跟踪越来越多的学者选择使用基于检测算法的多目标跟踪算法。尽管在此领域取得了显著成果，但在复杂环境下实现精确的行人多目标检测和跟踪仍面临多重挑战。这些挑战主要包括：

1)行人目标变化。由于行人是非刚体，其姿态的变化可能影响检测算法的效果。此外，行人服装与背景颜色的相似性可能导致行人被错误地判定为背景，增加漏检率。

2)遮挡问题。复杂环境中行人之间或由其他遮挡物引起的遮挡，使得难以完整提取行人特征，降低检测精度。

3)小尺度目标漏检。由于不同的拍摄角度或相互遮挡，可能出现小尺度的行人目标。这些目标在图像中占据的像素少，难以直接提取有效特征，从而影响检测性能。针对这些问题的有效解决方案是当前研究的难点。

4)此外，实时性也是行人检测算法面临的一大挑战。许多现有的行人检测算

法为了提高精度往往牺牲了速度，这限制了算法的实时应用性。在实际工程应用中，能够快速响应并实时检测目标的能力是评价算法性能的关键指标之一。因此，如何在不丧失检测精度的前提下提高算法的处理速度，实现精度与速率的优化平衡，也是本研究需要解决难点。这不仅涉及算法本身的优化，还包括计算效率的提升，是当前行人检测技术发展的关键方向。

1.3.2 论文研究的重点

将跟踪目标的位置在视频序列中识别出来并把其运动轨迹记录下来是行人多目标跟踪系统的主要任务。通过深入研究复杂环境下行人的运动过程，发现跟踪过程中需要面临的问题可以分为以下两点：首先便是无法确定行人目标的速度和轨迹。在跟踪过程中，行人的速度和轨迹可能发生快速变化，这增加了跟踪目标丢失的风险。因此，本文的研究重点之一是如何有效处理行人目标运动速度的快速变化和轨迹的多样性。再者就是对于突变特征的难以掌握。实际跟踪的过程中，遮挡物把目标部分遮挡，导致外观特征的变化。这种变化可能引起跟踪结果中的行人目标未被连续跟踪或 ID 频繁切换，解决这一问题也是本研究的重点。

1.4 论文研究内容及章节安排

1.4.1 研究内容

本文旨在通过研究基于改进 YOLOv8 检测算法和 Deep Sort 跟踪算法，来解决复杂环境下对行人目标检测与跟踪过程中遇到的问题。以下是本文主要完成的研究工作：

在检测部分，本文针对复杂背景中密集行人容易漏检、检测精度低以及检测速度慢的问题进行了深入研究。

1)首先，结合了 BiFormer 自适应注意力机制和小目标检测头，以增强语义处理和加强模型对小目标的检测能力，这有助于改善在复杂环境下行人特征不明显的问题。

2)为提升模型的检测速度，对模型的主干部分进行了轻量化改造，采用了参数量更小的 GhostNet 网络替换原有的特征提取网络。

3)此外，在损失函数方面也进行了优化，改进了回归损失函数以增加定位损失的权重，并将 CIOU 损失函数替换为精度更高的 EIOU 损失函数，从而提高了

检测效率。

4)实验方面,为了使研究结果更具有泛化性,本文在 Crowd Human 和 Caltech Pedestrian 两个数据集分别进行试验,首先分别对数据集进行背景增强,模拟复杂天气(雨、雾、雪)以及光照不足情况下的行人背景,然后将这些数据集分为 80% 的训练集和 20% 的验证集。通过在这些数据集上来验证本文所提的方法在实际应用中的效果和泛化能力。

在跟踪部分,考虑到 Deep Sort 算法在处理复杂场景下的密集行人跟踪时面临精度不高和速度不足的问题,对该算法进行了以下的改进。

1)首先,提出将 Deep Sort 中的卡尔曼滤波模型从匀速线性模型更换为匀加速模型,以更真实地模拟行人在实际环境中的运动轨迹。

2)其次,为了提升跟踪速度,采用了更为轻量级的重识别模型替代原有模型,并相应地减少了跟踪器的数量。改进旨在减轻模型对计算资源的需求,从而提高跟踪速度。

3)为验证改进后算法的效果,在 MOT16 数据集上进行真实效果测试,使用的检测器为经过优化的 YOLOv8n 算法。通过这些实验,检验了改进措施在实际应用中良好的性能表现。

本文的整体技术路线图如图 1-6 所示。

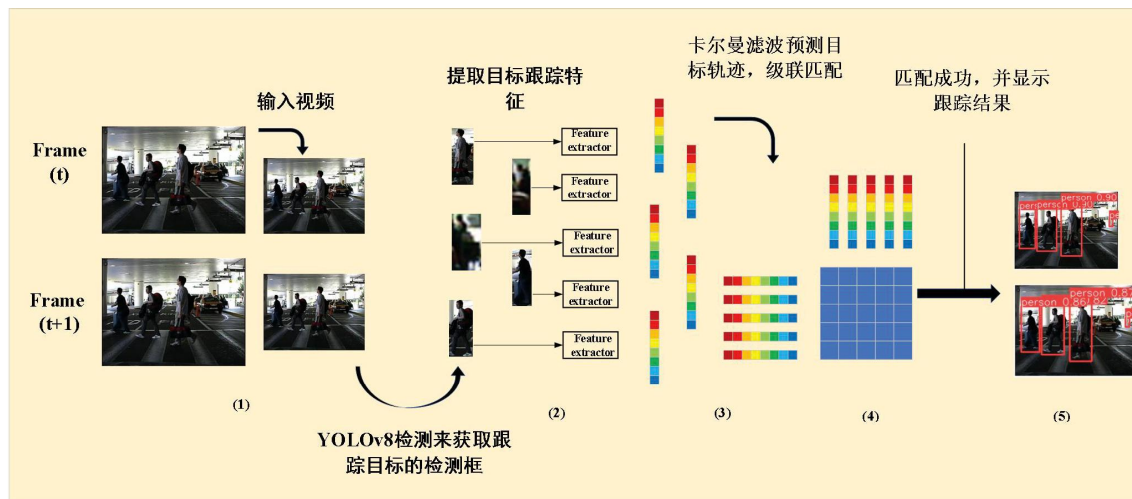


图 1-6 整体技术路线图

1.4.2 论文章节安排

第一章为绪论,旨在论述课题相关的研究现状与背景。首先,简要介绍本论文的研究主题,包括研究的背景和意义。接下来,对当前国内外在目标检测与跟

踪领域的研究进展进行了详尽分析。首先详尽阐述了传统目标检测技术的理论基础和实现机制，并对其优缺点进行了全面分析。然后对基于深度学习的检测技术进行详尽介绍：单阶段检测算法如 YOLO 系列，以及两阶段目标检测算法，如 RCNN 系列。此外，还挑选了一些传统跟踪网络和基于深度学习的跟踪算法的典型例子进行详细解析。评述各种现有的理论和方法，并在此基础上详细阐述本文的研究重点及采用的技术方法。最后，概述全文的组织结构，明确各章节的主要内容和逻辑关系。

第二章标题为“行人检测与跟踪相关理论概述”，在这一章节中，首先对深度学习的基本概念进行介绍。其次详细解释了卷积网络的关键层，包括输入层、卷积层、池化层、全连接层。接下来，讨论了激活和损失函数。然后介绍了非极大值抑制算法以及聚类算法。最后，介绍了检测跟踪领域常用到的一些经典数据集。

第三章标题为“基于 Biformer 和改进的 YOLOv8 的行人检测与识别”。本章首先简要回顾了 YOLOv8 的基本原理及其与先前版本相比的改进之处，包括数据特征处理、网络结构和标签分配策略等方面的详细介绍。随后，深入探讨了对检测网络所进行的若干关键改进：首先，为了有效解决目标漏检问题，引入了 Bif 注意力机制，以增强对目标特征的提取能力；其次，为了增强算法的实时性能，使用轻量级的 GhostNet 作为核心特征提取网络，这一策略不仅减少了模型的参数数量，同时也加快了检测速度。此外，为了提升模型检测效率，对损失函数进行了优化，改进了回归损失函数以增加定位损失的权重，并将 CIOU 损失函数替换为更精确的 EIOU 损失函数。最后，通过在两个数据集上进行消融和性能比较实验，证明了这些改进策略的有效性。此章节旨在展示通过这些创新性的改进来提升行人检测算法在复杂环境中的性能和实用性。

第四章标题为“基于改进的 Deep Sort 行人跟踪模型”。本章首先详细介绍 Deep Sort 算法中关键的技术步骤，包括卡尔曼滤波的预测机制和数据关联算法。继而详述 Deep Sort 算法的具体实现流程。本章后半部分专注于对原始 Deep Sort 算法改进措施的论述，包括：通过引入加速度成分，采用匀加速模型替换传统的恒速线性模型，以更真实地模拟行人在实际环境中的运动轨迹；引入了重识别模型，并将 DeepSort 中原有的堆叠残差模块替换为 OSA（Omni-scale Architecture）架构，使算法的推理速度与精度得到优化。最后，在包括 MOT16 测试数据集上

进行实验，以展示所提改进措施的有效性。

第五章为“总结与展望”，本章将汇总并回顾本文的主要研究成果和创新点。首先，将概述在行人目标检测和跟踪领域所完成的工作，包括所采用的方法、实验设计以及所取得的研究成效。然后，综合分析本文研究的意义和实际应用价值，指出研究过程中遇到的主要挑战和限制。最后，本章将提出对未来多目标检测与跟踪技术研究领域的展望，探讨可能的研究方向，如算法的进一步优化、新技术的应用可能性以及实际操作中的改进潜力。通过这些展望，旨在为该领域的未来发展提供方向和启示。

第二章 行人检测与跟踪相关理论概述

本章系统阐释深度学习理论体系及其在密集行人感知中的核心支撑作用。作为计算机视觉领域的技术基石，深度学习的特征表示能力与非线性建模优势，为突破密集场景中行人目标微小尺度、复杂遮挡、动态干扰等难题提供了根本性方法论。具体而言，卷积神经网络（CNN）的层次化特征提取机制，通过局部感受野与权值共享特性，能够有效捕获行人目标的局部纹理与全局语义信息；而注意力机制、特征金字塔等衍生技术，则为解决多尺度目标检测与背景干扰抑制等关键问题奠定了理论框架。针对密集行人检测与跟踪的特殊性，本章从三个维度构建理论认知体系：

1.基础理论必要性：深入解析卷积操作、反向传播、损失函数等核心原理，阐明深度学习如何通过端到端学习自动提取判别性特征，突破传统手工特征在密集场景下的表征局限性；

2.技术演进关联性：点明非极大值抑制、基于 K-Means Anchors 尺寸获取等技术对提升微小目标检测精度的内在机理；

3.相关实验数据集：介绍该领域一些常用行人检测数据集与常用行人跟踪数据集。

本章内容不仅为后续章节提出的改进思路提供了技术溯源与理论验证依据，更重要的是建立“基础理论—技术瓶颈—创新方法”的逻辑闭环：只有掌握注意力机制的本质作用，方可构建动态稀疏注意力机制实现背景干扰抑制；唯有透彻理解卷积神经网络特征提取的层次化特性，才能设计轻量化特征提取网络。这种理论衔接实践的论述逻辑，凸显了本章在全文中的纲领性地位，确保后续方法创新具备坚实的理论合理性与技术可行性。

2.1 深度学习相关基础

作为机器学习的重要分支，深度学习已成为推动人工智能领域进步的核心力量。自从其理念被提出以来，便持续成为研究热点，并逐步发展成为机器学习中

的一个主流方向^[64]。因此，在深入探讨深度学习的具体内容之前，有必要先对机器学习的基本知识进行简要回顾。这一概念最初于上世纪中叶提出，并自 80 年代起随着算法和理论的日益完善开始被广泛应用，其迅猛发展的趋势一度使深度学习成为人工智能领域的象征。机器学习通过从大量的外部数据中学习分析和提取规律，然后利用这些规律进行智能化的识别、判断和预测。

机器学习的发展主要经历了两个阶段：浅层学习阶段和深度学习阶段^[65]。前者通常在结构上较为简单，也有许多比较成熟的模型，至今在很多学习算法中仍占据主导地位。其模型根据是否含有隐层节点，可分为两类：一类是只含一个隐层节点的模型，如支持向量机(SVM)和 Boosting^[66]；另一类是不包含隐层节点的模型，如逻辑回归(Logistic Regression)^[67]。这类模型的一个主要局限性是当训练样本数量有限时，训练出来的模型在应对复杂分类问题时的能力存在不足。复杂的分层模型结构是深度学习区别于浅层模型的主要差异，通过构建多个隐藏层节点增加模型的深度，从而能够处理更多的数据信息，增强模型的特征学习能力。随着结构上的加深，深度学习能够提取特征的程度也加深，表征能力也得到了增强，能够更精确地完成智能识别、分类与预测任务。

人工神经网络(Artificial Neural Network, ANN)的研究对深度学习的发展影响深远^[68]。人工神经网络是模拟人脑神经网络的信息处理过程。一个典型的神经网络往往由很多神经元互相连接、叠加构成，这些神经元通过复杂的互联增强了网络的信息处理能力^[69]。每个神经元通常接收多个输入，并产生一个输出。输入数据(如图像数据)首先被送入神经元，通过神经元内的激活函数处理后，输出结果随后传递到下一层神经元。这一过程可通过图 2-1 所示的神经元基本结构进行说明。这种结构不仅展示了神经元的功能，也体现了深度学习中数据处理和信息流动的基本方式。

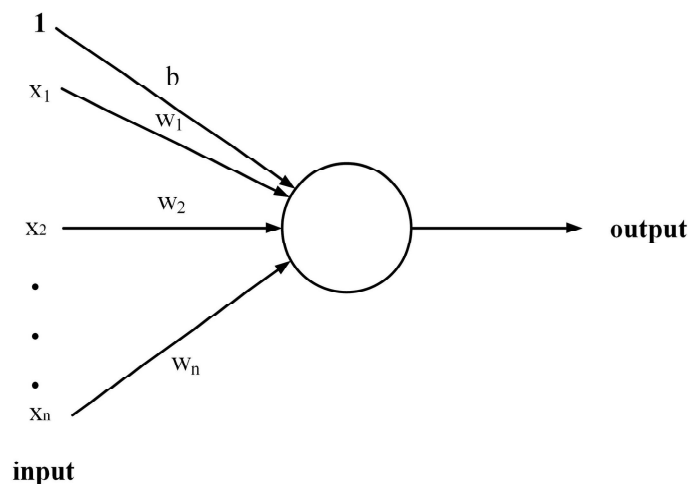


图 2-1 神经元基本结构图

连接多层神经元使模型拥有更强的学习能力，但也使模型难以自动更新参数，并且很难精确地模拟输入数据的实际分布，因此早期这种模型并未得到广泛应用。为了应对这一挑战，BP(反向传播)神经网络^[70]被提了出来，这种网络采用误差反向传播算法，通过迭代训练过程不断优化参数，逐步逼近最优解。BP 网络的结构通常由三个主要部分组成：输入层、隐藏层和输出层^[71]。通过这种结构，BP 网络能够有效地进行学习和适应，从而在多种应用中展现出其强大的功能。

BP 神经网络的训练主要包括两个核心步骤：信号的前向传播与误差的反向传播^[72]。在前向传播过程中，信号在训练初始阶段进入输入层，接着信号通过一系列隐藏层进行传递和处理，最终在输出层生成结果。在每个隐藏层中，输入信号会与对应层的权重矩阵相乘，之后通过一个激活函数进行处理以引入非线性因素，增强网络的表达能力。输出层的处理方式类似，但通常不包含激活函数，直接输出预测结果。

与前向传播不同的是，误差反向传播过程中，输出层的误差是需要首先算出的，误差是指网络输出与实际结果之间的差异。这个误差将沿网络反向传递，通过调整各隐藏层的权重来减少误差，这一过程涉及到计算每层的误差梯度并据此调整相应的权重矩阵。这个过程通常需要重复进行多次迭代，直到整个网络的预测误差降到可接受的范围或达到其他预设的停止条件。

通过这样的训练过程，BP 神经网络能够有效地学习和适应给定的数据集，从而在各种复杂任务中实现高效的预测和分类。

2.2 卷积神经网络相关内容

卷积神经网络(Convolutional Neural Networks, CNN)构成了一种极其关键的人工神经网络类型。虽然早在本世纪初就已有人对其开发并应用于手写数字识别等领域,但卷积神经网络真正得到业界关注和认可则要到 2012 年 AlexNet^[73]的面世,在那之后卷积神经网络的发展便突飞猛进,对深度学习在多个领域的交互应用起到了举足轻重的作用。卷积神经网络基本工作原理的如图 2-2 所示,涵盖了从数据输入到特征提取、再到最终输出的全过程,这种结构化的网络设计使得 CNN 在处理图像和视频数据时表现出了卓越的性能。

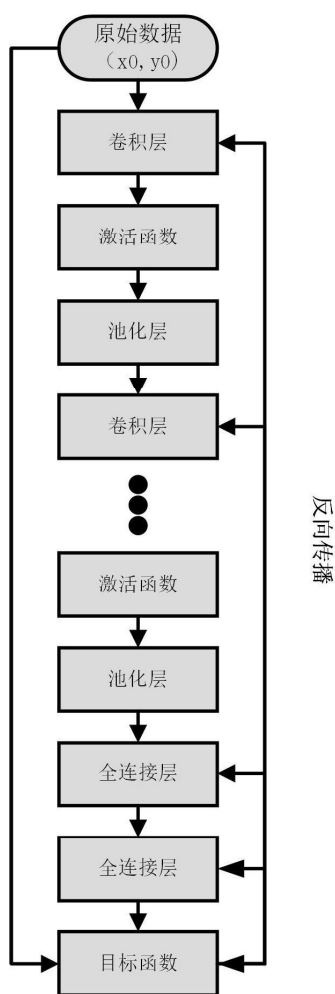


图 2-2 卷积神经网络基本流程图

卷积神经网络(CNN)的一个显著特点是能够自动学习数据,这在过去是完全不敢想象的,工作人员再也不需要做手动地提取特征这种复杂的过程。卷积神经网络之所以在图像处理领域特别有效,很大程度上依赖于它处理数据的方式,卷积运算可以高效获取图像中的局部特征,同时处理效率也保持在一个很好的水平,

因此其能够成为处理图像相关任务的重要技术。

2.2.1 输入层

在卷积神经网络中，输入层发挥着关键作用，它的主要任务是将获取的图像像素信息输入到网络中。通常，图像数据以 RGB 像素点的格式输入，并进行信息的量化处理。这一步骤旨在降低整张图像的信息量，从而便于网络进行前馈传播和反向传播求导。然而，输入信息的多少会使网络预测的准确性产生波动。因此，在将图像特征输入网络之前，通常会进行一系列的预处理，包括旋转、裁剪、翻转和高斯滤波等，以优化数据输入。此外，输入的图片格式也受到特定的限制。在涉及到有关活体检测和高精度目标识别时，通常需要对摩尔纹和图像纹理特征进行细致处理，以确保识别的准确性和抗干扰能力。这些处理步骤确保输入层能有效地传递和转化图像数据，为网络的深层处理提供坚实基础。

2.2.2 卷积层

卷积神经网络中的核心部分当属卷积层了，它的主要任务是对输入图片执行卷积运算操作，以此来抽取图像特征。这一过程涉及将卷积核(或滤波器)在图像特征层上进行滑动操作，从而生成大量的特征图像，每一个均代表从输入中提取到的一个相对应特征。在这种局部操作中，卷积核的权重与相应的输入图像像素值进行乘法操作，并将这些乘积结果累加，从而捕获图像的局部信息。通过权重共享机制，卷积层使用同一个卷积核对整个图像进行系统性扫描，这不仅确保了特征提取的一致性，也大幅减少了模型的参数数量，从而简化了网络的复杂性并降低运算量。卷积层通常采用 3*3 或 5*5 的卷积核尺寸，并以一定的步长滑动卷积核来快速提取图像特征矩阵的信息，有效执行卷积任务。在本章后续部分，将通过介绍分析卷积层的原理及其在神经网络中的定位，来阐释它如何高效地处理和提取图像数据的。

卷积层运算公式如(2-1)所示。

$$y_{(i,j)} = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} W(u,v) X(i-u, j-v) + w_b \quad (2-1)$$

其中卷积核的宽度是 M ，深度为 N ， $W(u, v)$ 则可以表示为第 u 行、第 m 列的权重，输入图像中的坐标值则用 $X(i-u, j-v)$ 表示，偏置是 w_b 。

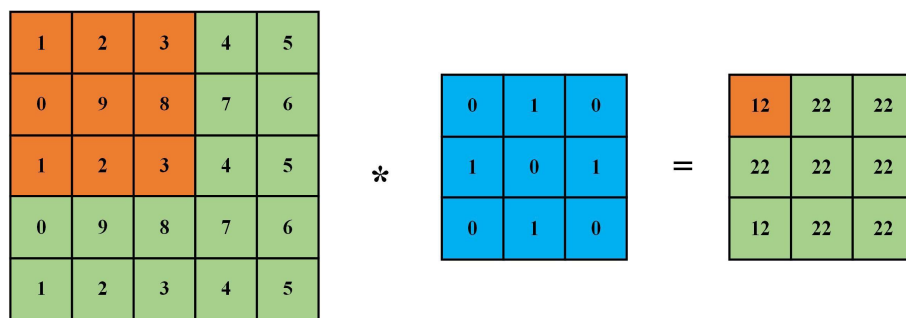


图 2-3 卷积操作示意图

卷积操作的过程如图 2-3 所示。首先，将卷积步长设置为 1，并应用一个 3×3 的卷积核，从输入特征图的左上角开始，自左至右、自上至下依次滑动遍历。每次遍历，卷积核的值与相应局部区域的像素值相乘并相加，得到该区域的卷积输出值。通过这样的滑动计算，生成一个新的特征图，从而提取出图像的局部特征。

2.2.3 池化层

使用多个卷积核在卷积层中进行运算可能会引起叠加效应，使得流向后续网络层的数据维度过高，并增加了计算负担，会导致模型过拟合。为了解决这一问题，池化层一般设置在卷积层之后，其主要通过下采样的方式减少上一层输出的特征图的维度，最终达到降维目的，让模型的泛化性能维持在一个好的水平，从而减少特征数量并防止过拟合。

池化过程主要是利用预设窗口在上一层输出的特征图上进行滑动扫描，然后将压缩后的特征图输出，这个新特征图会用于输入到网络的下一层。与卷积层相比，池化层的设置较为简单，它不涉及权重等内部参数的调整，只需要确定池化类型(如最大池化或平均池化)、池化核的大小以及步长等参数。这样的设计不仅减少了模型的计算需求，还有助于提取更具鲁棒性的特征，从而增强了网络处理图像的能力。

平均池化和最大池化是在神经网络中使用最多的池化类型。其中最大池化操作通过在图像局部区域内选取最大值来代表该区域的所有数据，这种方法处理数据起来迅速，因此在实际应用中被广泛采用。例如，OpenCV 的原始滤波算法采用了最大池化原理^[74]。而与之不同的是，平均池化通过计算图像子区域的平均值，来代替该区域内的所有像素信息，可以保留区域的整体数据特征，更加突出背景信息。这两种池化方法各有优势，最大池化更倾向于突出局部特征，而平均池化则提供了更平滑的特征表达。使用尺寸为 2×2 且步长为 2 的池化核进行的两种池

化操作如图 2-4 所示。

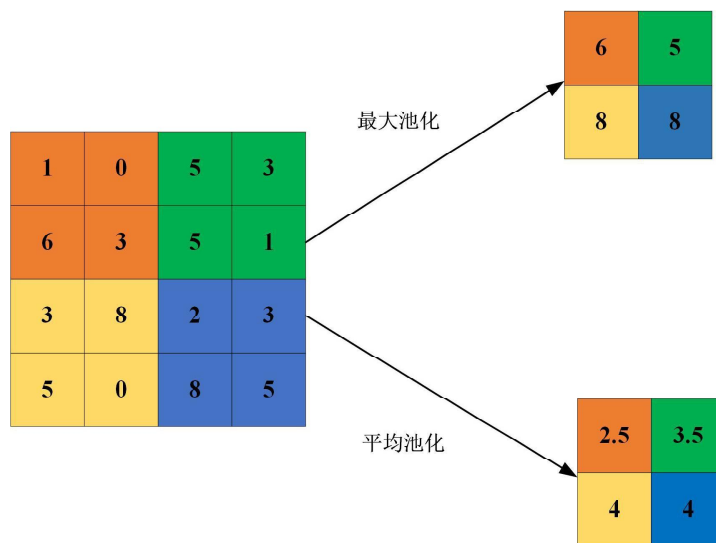


图 2-4 两种池化操作示意图

总的来说，池化操作利用压缩数据来进行特征图的降维。通过这种方法，整个网络的维度和需要的参数量大大减少，从而缩减了模型的整体大小，以及使训练效率有所提升。通过这种方式，池化层不仅简化了数据结构，还增强了模型对输入数据变化的适应性，成为优化神经网络性能的关键步骤。

2.2.4 全连接层

图片经过卷积层卷积后会生成数个不同的特征图，其都各自提取了图片的某一特定特征，这些特征图随后被送入池化层进行降维处理，以减少数据量并简化网络结构。随后激活函数将降维后的数据进行下一步操作，并提取出更加离散的特征信息。处理完毕后，这些数据会被送入全连接层，以进一步整合特征，并为最终的输出做准备。这一系列的操作不仅优化了数据流，还增强了网络对图像特征的识别能力。

在卷积神经网络中，经过卷积层和池化层处理后，图像特征已被有效提取并进行了降维。之后，这些特征通过全连接层进一步处理，以提高分类的准确性。在全连接层中，所有神经元的输出信息被完全连接起来，输出的特征矩阵会被展平成一个一维向量，并将其作为分类器的输入。全连接层的目的是将前面各层提炼的特征进行集中和融合，通过反复提纯，最终提高输出的分类质量。这一层在整个网络中起着至关重要的作用，作为特征整合和决策输出的关键部分。全连接层的特征向量计算可以通过特定的数学公式(2-2)进行描述，这些公式详细说明了如何将多维数据转换为一维数据，以便进行最后的分类判断。

$$y=f(wx+b) \quad (2-2)$$

其中 f 代表激活函数，输入以及输出的特征向量分别用 x 和 y 表示，而权值矩阵则用 w 表示。

2.2.5 激活和损失函数

2.2.5.1 激活函数

在卷积神经网络的处理过程中，原始图像数据在经历卷积与池化操作之后，涉及的图像像素点的权值计算通常采用线性方式。然而，线性操作可能因表达能力不足而限制了模型的学习效果。为了解决这一问题，激活函数被引入到网络中以增加非线性因素，从而提高模型对复杂数据的处理能力。激活函数的引入使得网络能够学习更加复杂和抽象的特征表示。本节将重点介绍三种在实践中常用的激活函数，探讨它们各自的特点及在网络中的应用效果。这些激活函数包括 ReLU、Sigmoid 和 Tanh，每种激活函数都有其独特的数学特性和适用场景，在帮助网络实现非线性表达方面起着至关重要的作用^[75-77]。

ReLU 激活函数：从公式上来看，ReLU 是一个分段函数。

$$\text{ReLU}(x)=\begin{cases} x & x \geq 0 \\ 0 & x \leq 0 \end{cases} \quad (2-3)$$

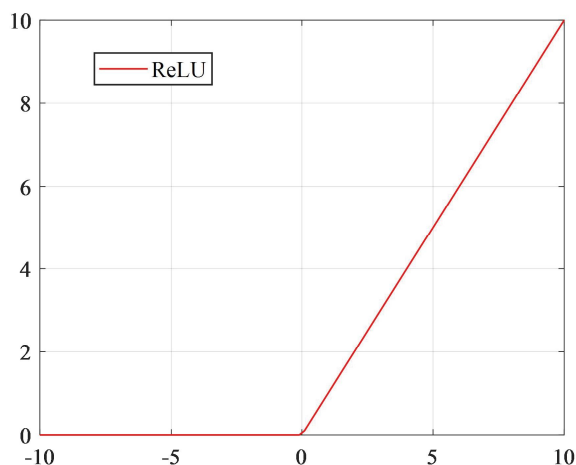


图 2-5 ReLU 函数曲线

ReLU 函数的输出值如图 2-5 所示均是大于等于 0 的值。当输入 x 大于等于 0 时，ReLU 函数的梯度值为 1；而当 x 小于 0 时，梯度值为 0。这一特性意味着在 x 大于等于 0 的情况下，其梯度恒为 1，有效地解决了梯度消失问题，这在深

层网络中尤为重要。与其他函数相比，ReLU 函数的计算过程更简单，这大大降低了计算成本，减少了计算时间。实验结果表明，使用 ReLU 函数可以加速网络训练过程，使模型的收敛速度提高近 6 倍。这些优点使 ReLU 成为当前深度学习模型中备受青睐的激活函数之一。

Sigmoid 激活函数：Sigmoid 激活函数是一种单调递增的函数，其计算公式如公式(2-4)所示。

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2-4)$$

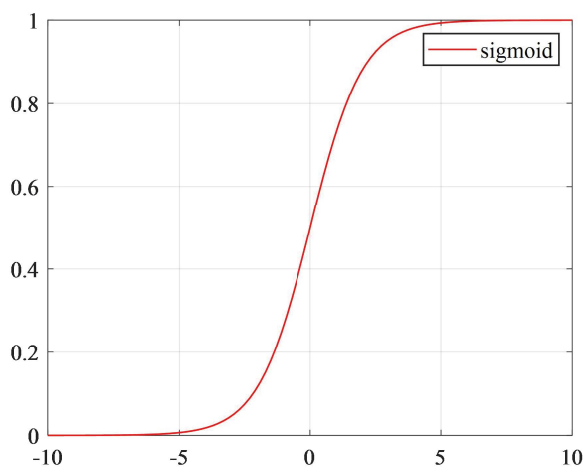


图 2-6 Sigmoid 函数曲线

如图 2-6 所示，Sigmoid 激活函数是一种单调递增的指数型函数，其形态与生物神经元的响应相似。该函数的输出值被压缩在(0,1)区间内，这使得它可以将网络的输出解释为概率，适用于二分类任务中作为决策函数。然而，Sigmoid 函数存在一个明显的缺点：当输入值过大或过小时，输出值会趋向于 1 或 0，导致输出在这些区域几乎没有梯度(称为“饱和效应”)。这种饱和现象会在网络训练过程中导致梯度消失问题，从而阻碍网络权重的有效更新，最终影响整个网络的学习性能和训练速度。

Tanh 函数：Tanh 函数也是一种常用的激活函数，其公式如公式(2-5)所示。

$$\text{Tanh}(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2-5)$$

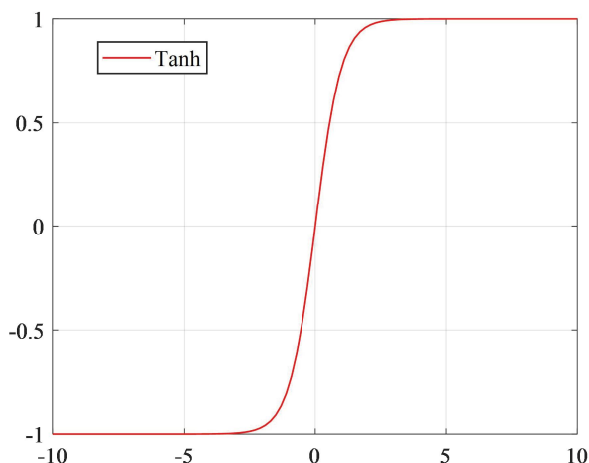


图 2-7 Tanh 函数曲线

如图 2-7 所展示的, Tanh 函数(双曲正切函数)是一个标准的双曲线形状,也是单调递增的。Tanh 函数是基于 Sigmoid 函数提出,但与 Sigmoid 函数不同的是, Tanh 函数的输出值被规范化到(-1,1)之间。这种输出范围的调整有助于在神经网络训练过程中更有效地扩展特征,因为输出值围绕零中心对称,可以更好地处理正负输入值。然而尽管 Tanh 函数在多个方面表现优于 Sigmoid 函数,但也面临着“软饱和”现象,即当输入值特别大或特别小时,其梯度趋近于零,这也可能引发梯度消失的问题,从而影响网络的训练效率和最终性能。

2.2.5.2 损失函数

损失函数主要是利用优化参数来最小化神经网络训练过程中的损失,是目前用于评估模型预测结果与真实目标值之间的偏差最常用的一种方法,从而确保模型预测值尽可能接近实际目标值。其结果大小真实反映了模型在训练过程中的损失程度,通常情况下,损失值越小,表明模型的鲁棒性越强,预测性能越好^[78]。

下面将介绍两种关键的损失函数:回归损失函数和分类损失函数^[79]。回归损失函数常被用来预测连续变量,常见的回归损失函数有均方误差(MSE)和绝对误差(MAE)等类型^[80-81],这些函数用于测量预测值与真实值之间的差距。分类损失函数则用于离散变量的预测,如交叉熵损失和 Hinge 损失等^[82-83],这些函数用于评估分类预测的准确性,这两类损失函数在神经网络模型的设计和性能评估中发挥着关键作用。

(1)回归损失函数

回归损失函数用于评估连续变量预测模型的误差,主要是在回归分析中使用

的，主要目的是量化模型预测值与实际目标值之间的差异。以下是几种常用的回归损失函数：

平均绝对误差函数(Mean Absolute Error, MAE)：MAE 是预测值与实际值之差的绝对值的平均。公式为：

$$l_{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - f(x_i)| \quad (2-6)$$

均方误差函数(Mean Squared Error, MSE)：MSE 是最常用的回归损失函数之一，公式为：

$$l_{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \quad (2-7)$$

均方根误差函数(Root Mean Squared Error, RMSE)：RMSE 是 MSE 的平方根，更符合误差的实际尺度。公式为：

$$l_{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2} \quad (2-8)$$

这些回归损失函数根据具体应用和数据的特点选择。例如，当数据包含许多异常值时，可能会优先考虑使用 MAE。而在误差分布较为正常的情况下，MSE 或 RMSE 可能更为合适。

(2)分类损失函数

分类损失函数主要用于评估分类模型的性能，即模型预测的输出与实际类别标签之间的误差。这类损失函数在处理离散标签数据时非常关键，特别是在进行分类任务时。以下是几种常用的分类损失函数：

交叉熵损失函数：交叉熵损失函数(Cross-Entropy Loss)也称为对数损失，是处理二分类和多分类问题时最常用的损失函数之一，以及用来评估标签的真实分布与预测分布的差距，其目的是将这种差距最小化。公式为：

$$CELoss = -\sum_{i=1}^n y_i \log y_i' \quad (2-9)$$

二元交叉熵损失函数：二元交叉熵损失函数(Binary Cross-Entropy Loss)的用途主要是为了处理二分类问题，公式为：

$$BCELoss = -\frac{1}{n} \sum_{i=1}^n y_i \log(p(y_i)) + (1 - y_i) \log(1 - p(y_i)) \quad (2-10)$$

其中, y 是实际标签, p 是属于正类的预测概率。

多类交叉熵损失函数: 多类交叉熵损失函数(Categorical Cross-Entropy Loss)适用于解决多分类问题, 其公式与交叉熵相似, 但特别设计以适应多类情境。

合页损失函数: 合页损失函数(Hinge)常用于支持向量机(SVM)中, 但也可以用于神经网络的训练, 主要用于确保正确分类的边界是最大的。公式为:

$$\text{HLoss} = \max(0, 1 - t \cdot y) \quad (2-11)$$

平方损失函数: 平方损失函数(Squared Loss)虽然通常用于回归, 但也可以用于分类问题, 旨在最小化预测值与真实值之间的平方差。公式为:

$$\text{SLoss} = (y - p)^2 \quad (2-12)$$

每种损失函数都有其特定的使用场景和优势。选择合适的损失函数可以根据具体任务的需求、数据的特性以及所需模型的特点来决定。例如, 交叉熵损失由于其在处理概率时直接性和效果明显, 通常是处理分类问题的首选。

2.2.6 非极大值抑制算法

使用锚点(Anchor)机制进行目标检测时, 一个检测目标往往会产生多个候选框, 这可能会降低检测的整体精度。为了应对这一问题, 通过使用非极大值抑制(Non-Maximum Suppression, NMS)算法^[84], 筛选并确立最佳的检测结果。NMS算法通过消除冗余候选框, 保留最有可能代表实际目标的框, 从而提高检测的准确性, 具体步骤如下:

首先, 对所有预测的边界框按得分进行排序, 高分边界框优先。随后, 移除那些低于特定阈值(如 0.5 或 0.7)的边界框。接着, 保持列表顶部的最高得分边界框, 并计算这个框与列表中其他所有框的交并比(IOUS)。所有 IOU 高于阈值的框都将被移除。此后, 重复此过程, 每次都选取得分最高的框进行相同的操作, 直到所有的框都经过处理。这样就只留下 IOU 值低于设定阈值的框, 确保结果的最优性和准确性。

虽然传统的 NMS 能有效地去除冗余预测框, 并以置信度为基准来评估候选框, 但在某些情况下, 所有预测边框可能都不够准确, 从而导致筛选后的结果仍无法满足需求。此外, 尽管某些预测的置信度较高, 其实际精度可能仍然不足, 这会降低最终的输出质量。因此, 仅依靠分类置信度进行 NMS 可能不是最理想的选择。为了提高精确度, 位置置信度被添加进来, 其增加了对输出结果的精确

控制。此算法在处理边框数据时，基于两个关键假设：首先，预测框分布要遵循高斯正态分布；其次，真实框的分布要遵循狄拉克分布。

根据上述假设，引入了基于 KL 散度的支持向量机算法，目的是用于评估两个概率分布间差异的方法^[85]。这种方法通过更精细地调整边框位置，使得边框回归过程更加符合真实世界中物体位置的概率分布，因此可以有效地提高检测系统的整体性能。

2.2.7 聚类算法

在目标检测领域，数据集中目标区域分布的广泛性常常使得位置信息难以预先判断，这不仅消耗了大量的计算资源，也可能降低了算法的准确度。采用锚框 (Anchors) 机制的目标检测算法通过设置预定义的矩形框来预测目标的位置，被证实可以稳定地提高检测的精度和速度。自 YOLOv2 起，为了提高检测性能，三种不同尺度的预选框被 YOLO 系列模型引入，这一改进使得其在 v1 的基础上各方面效果得到了显著改善。然而，预选框的尺寸通常需要根据具体数据手动设置，这可能导致所选尺寸并非最佳。

特别是在本研究中使用的行人数据集中，由于拍摄距离和遮挡面积的不同，存在多种尺寸的目标。为了提高模型的检测精度，进行聚类分析以确定该数据集中最适合的先验框尺寸和数量是必要的，有助于确保模型更加精确地识别和检测数据集中的目标，从而优化整体性能。

K-Means 聚类算法在目标检测中的应用主要是通过对已标定的矩形框进行分析，以确定适合数据集特性的不同宽高比的矩形框尺寸。这种方法，称为 K-Means Anchors^[86]，多用于基于锚框机制的目标检测网络，比如 Faster R-CNN 和 YOLO。使用 K-Means Anchors 确定锚框尺寸的一般步骤为：

1. 收集整个训练集中的所有边界框；
2. 从这些边界框中筛选出一部分作为聚类分析的样本，筛选标准可依据样本的尺寸和长宽比来确定，以确保样本具有足够的代表性；
3. 从每个边界框中提取特征，通常使用边界框的宽和高；
4. 应用 K-Means 聚类算法对这些特征进行聚类，这里的 K 值代表所需确定的锚框类别数量。聚类的目的是形成 K 个不同的群组，每个群组代表一种特定尺寸的锚框；

5.从每个群组中提取中心点的尺寸，将其定义为该群组锚框的尺寸；

6.根据具体的应用需求，可以对锚框尺寸进行进一步的优化处理，如调整大小或修改长宽比限制等。

通过这种方式，就能够确定适用于目标检测算法的锚框尺寸。这些确定的锚框尺寸将被用来生成默认的候选框，这些候选框在目标检测任务中扮演着至关重要的角色，用于进行预测和确保预测结果与实际目标的正确匹配。

2.3 常用的测试数据集

2.3.1 常用行人检测数据集

2.3.1.1 WiderPerson 行人检测数据集

WiderPerson 数据集是由武汉大学团队发布，涵盖了大量高分辨率的行人图片，每张图片中行人的位置和类别均有明确标注。具体来说，WiderPerson 数据集包含大约 10000 张训练图片和 5000 张测试图片，总计约 100000 个行人实例。每个实例都配有行人区域的边界框和类别标签。这些图片展示了行人在多样化的场景、不同姿势和各种遮挡条件下的样子，为检测算法的鲁棒性和准确性提供了测试平台^[87]。

研究者可以利用 WiderPerson 数据集来培训和评估自己的行人检测系统，以此与国际上其他的研究成果进行比较。此数据集有助于推动行人检测技术在如智能监控和自动驾驶等实际应用的发展。数据集中的部分图片如图 2-8 所示。

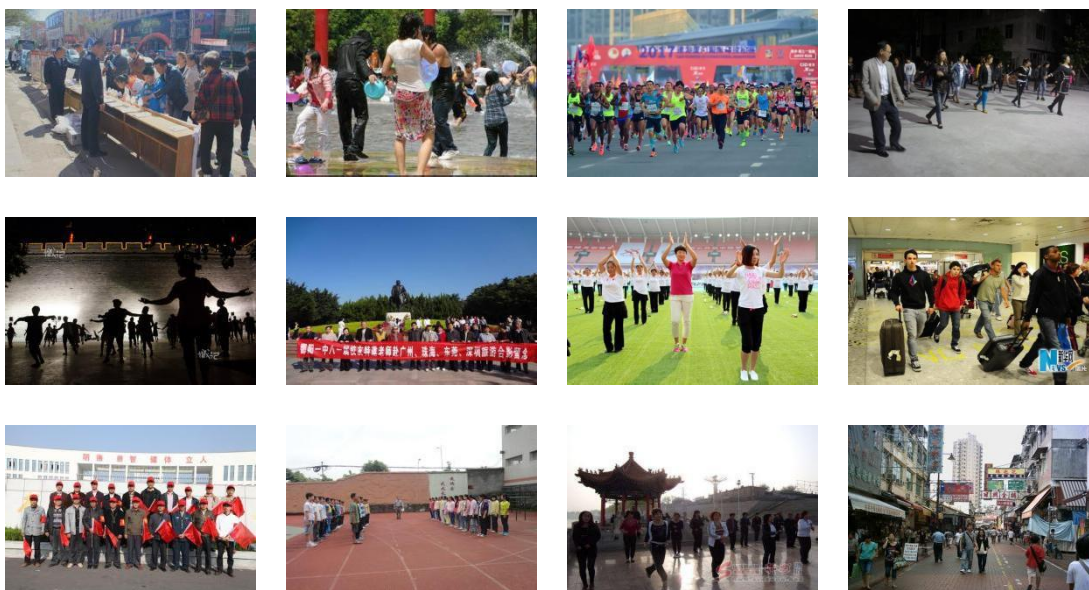


图 2-8 WiderPerson 数据集。

2.3.1.2 Caltech Pedestrian 数据集

Caltech Pedestrian 数据集由加州理工学院的研究团队开发，是行人检测和行为分析领域的一个标杆数据集。该数据集提供了大约 10 小时的高清视频，拍摄场景包括城市街道和校园环境，总计约有 250000 个详细标注的行人实例。每个实例不仅标明了行人在视频中的具体位置和边界框，还分类了行人的行为模式。这些视频特别展示了如视角变化、行人遮挡和高密度行人流等复杂情境，为检测技术的鲁棒性和精确性测试提供了资源^[88]。

研究者可以利用 Caltech Pedestrian 数据集来测试和评估行人检测系统在各种复杂环境下的性能，促进算法的比较研究和技术竞争。此外，该数据集还广泛用于行为分析和姿态估计等研究，帮助研究人员深入探讨和解决实际应用中的问题。部分数据集图片如图 2-9 所示。



图 2-9 Caltech Pedestrian 数据集。

2.3.1.3 CrowdHuman 数据集

CrowdHuman 是由中国科学院自动化研究所发布的行人检测和属性分析数据集，专为提高行人检测算法性能而设计。该数据集包括大约 15000 张图片，涵盖了近 47000 个行人实例，每个实例都详细标注了位置、边界框、姿态、遮挡情况以及诸如性别和年龄等行人属性。CrowdHuman 数据集的特点在于其行人样本的高度多样性，包括各种姿态和复杂的遮挡情况，为检测算法的准确性和鲁棒性提供了严峻的测试环境^[89]。

CrowdHuman 数据集不仅适用于行人检测，也适合进行行人属性分析和姿态估计等研究。通过利用这一数据集，研究人员能够更好地应对真实世界中的行人

检测挑战，推动行人检测技术的进一步发展。部分数据集图片如图 2-10 所示。



图 2-10 CrowdHuman 数据集

2.3.1.4 VOC2007 与 VOC2012 数据集

VOC2007 数据集包括 20 个常见的物体类别，例如人、车辆、猫和狗等，每个类别大约有 1000 张图像，包含了物体位置和类别的大量信息，主要用于执行目标检测任务，即识别和定位图像中的特定物体。除此之外，VOC2007 也适用于其他计算机视觉任务，如目标分类和语义分割。

类似地，VOC2012 数据集延续了与 VOC2007 相同的类别设置，但增加了更多的图像和更新的标注数据，以及一些新增的类别标签。VOC2012 的主要用途同样是目标检测，同时也支持目标分类和语义分割等计算机视觉任务。由于提供了大量的图像数据和系统的标注信息，同时包括一个标准化的评估框架，这些 VOC 数据集在计算机视觉研究领域得到了广泛的应用，使研究人员能够方便地进行模型的开发、测试和比较。



图 2-11 (a) VOC2007 数据集；(b) VOC2012 数据集。

2.3.2 常用跟踪数据集

2.3.2.1 MOT16 数据集

在 2016 年, A. Milan 等研究人员推出了 MOT16 数据集, 该数据集特别强调多目标追踪的挑战性, 其视频序列涵盖了多种复杂的环境条件, 对追踪算法的精确性和鲁棒性构成了显著考验。该数据集包括 7 个训练视频和 7 个测试视频, 有的视频由固定摄像机拍摄, 而有的则由移动摄像机录制, 拍摄角度和时间因视频而异, 涵盖不同的天气和光照条件^[90]。

视频文件均转换为 JPEG 格式, 并以六位数文件名顺序命名。MOT16 数据集中的行人检测由 DPM 检测器执行, 该检测器结合了高灵敏度的数据挖掘方法对标记数据进行部分标记, 表现出良好的行人识别性能。此外, 这些视频提供了详尽的信息, 如分辨率、帧率、边界框数量、相机动态、拍摄角度和环境条件等。相关数据集的具体参数可见表 2-1。

表 2-1 MOT16 数据集相关参数

Sequence	nDet	nDet/fr	min height	max height
MOT 16-01	3775	8.39	19.00	258.92
MOT 16-02	7267	12.11	19.00	341.97
MOT 16-03	85854	57.24	19.00	297.57
MOT 16-04	39437	37.56	19.00	341.97
MOT 16-05	4333	5.18	19.00	225.27
MOT 16-06	7851	6.58	19.00	210.12
MOT 16-07	11309	22.62	19.00	319.00
MOT 16-08	10042	16.07	19.00	518.84
MOT 16-09	5976	11.38	19.00	451.55
MOT 16-10	8832	13.50	19.00	366.58
MOT 16-11	8590	9.54	19.00	518.84
MOT 16-12	7764	8.63	19.00	556.15
MOT 16-13	5355	7.14	19.00	210.12
MOT 16-14	8781	11.71	19.00	258.92
total	215166	19.15	19.00	556.15

在 MOT16 数据集的参数描述中, “n Det”代表视频中的总目标数量, “n Det/fr”则是指每帧视频平均包含的目标数。此外, “min height”和“max height”分别定义了检测框在视频画面中可能出现的最低和最高位置。

2.4 本章小结

本章首先对深度学习的基本概念进行概述, 系统阐释深度学习理论体系及其

在密集行人感知中的核心支撑作用。接着详细介绍了卷积网络中的核心网络层，包括输入层、卷积层、池化层、全连接层、激活和损失函数，并对非极大值抑制以及 K-Means 算法原理介绍分析。最后对本文所采用的基于深度学习的单阶段检测模型和性能优异的跟踪模型进行了总结。

第三章 基于 Biformer 和改进的 YOLOv8 的行人检测与识别

针对当前行人目标检测识别中存在的检测精度不高、响应时间长、遮挡场景中检测效果差等一系列问题，本章提出一种基于 Biformer 注意力机制和改进 YOLOv8 的行人检测方法，被称为 Bif-YOLO。Bif-YOLO 是一种新的深度混合神经网络，该网络将 Biformer 注意力机制和改进的 YOLOv8 融合完成行人检测任务。

该方法的主要贡献如下：(1) 引入 Biformer 注意力机制：为了有效解决目标漏检问题，本文通过 Biformer 注意力机制加强了对目标特征的提取，增强了模型对复杂场景的适应性；(2) 采用 GhostNet 网络：为提升算法的实时性，本文替换了原有的主干特征提取网络，采用了更轻量的 GhostNet 网络。这一改变显著减少了网络的参数量，从而提高了检测速度；(3) 损失函数的改进：针对模型检测效率不足的问题，本文对回归损失函数进行了改进，增加了定位损失的权重，并将 CIOU 损失函数替换为定位精度更高的 EIOU 损失函数，以提升模型的检测效率。

本章节的结构如下：3.1 节系统概述 YOLOv8 模型的架构；3.2 节介绍基于 Biformer 注意力机制的目标检测算法；3.3 节针对轻量化卷积网络在训练过程中采用的训练策略进行剖析；3.4 节是通过损失函数改进的目标检测算法；3.5 节显示改进算法网络的相关实验和结果，并提供有关训练数据和训练过程的详细信息；3.6 节就实验结果和相关训练数据进行分析讨论并针对本章节提出的行人检测方法进行了总结。

3.1 网络架构分析

Ultralytics 最新推出的 YOLOv8 模型，是基于先前 YOLO 系列进行的一次重大升级，融入了多项新技术和改良功能，显著提升了模型的性能和适应性。此模型以其高速、高准确度及用户友好性，成为进行目标检测、跟踪、实例分割、图像分类及姿态估计等任务的理想选择。

YOLOv8 主要沿袭了 YOLOv3 与 YOLOv5 的相关技术，并在以下几个关键领域进行了创新：

1)数据预处理：沿用 YOLOv5 的策略，YOLOv8 在训练过程中引入了马赛克增强、混合增强、随机透视及色彩调整等四种强化技术。

2)骨干网络结构：YOLOv8 骨干网络是在 YOLOv5 的基础上做出了调整，传统的 C3 模块被新的 C2f 模块取代，这一新模块通过增加更多分支，优化了梯度回传，图 3-1 展示了 C2f 与 C3 模块的对比。

3)FPN-PAN 结构：YOLOv8 继续采用 FPN 与 PAN 的结合方式，构建特征金字塔以促进多尺度信息的融合。除了 C3 模块被 C2f 模块替换外，其余结构保持与 YOLOv5 相似。

4)检测头结构：从 YOLOv3 到 YOLOv5，检测头通常采用耦合方式，即同时处理分类与定位任务。YOLOv8 采纳了类似 YOLOX 的解耦头设计，通过两条并行分支分别处理类别与位置信息，最后通过 1×1 卷积实现分类与定位，YOLOv8 的整体网络结构见图 3-2。

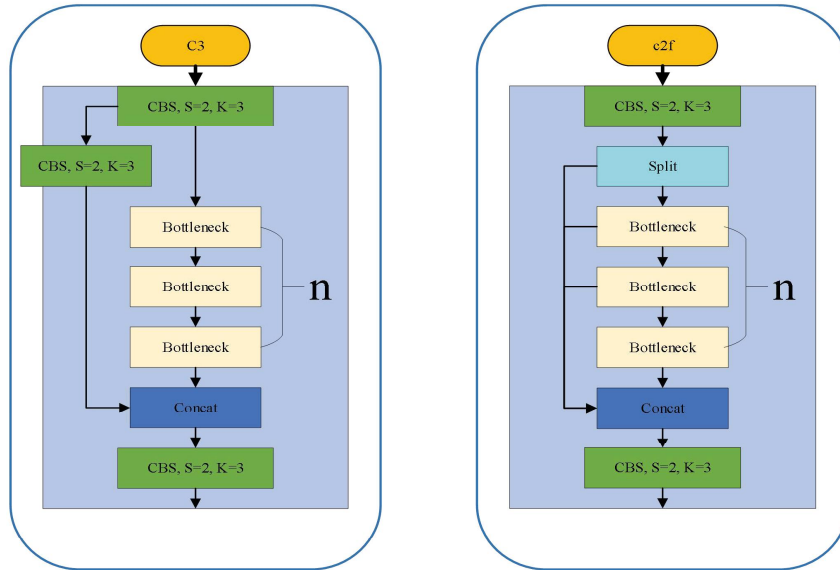


图 3-1 C3 和 C2f 模块示意图

5)损失计算：YOLOv8 的损失计算主要分为两部分：一是分类损失，使用的是变焦损失(VFLLoss)；二是位置损失，采用的是 CIoU Loss 和 DFL Loss。

其中分类损失定义如下：

$$VFL(p, q) = \begin{cases} -q(q \log(p) + (1-q) \log(1-p)) & q > 0 \\ -\alpha p^\gamma \log(1-p) & q = 0 \end{cases} \quad (3-1)$$

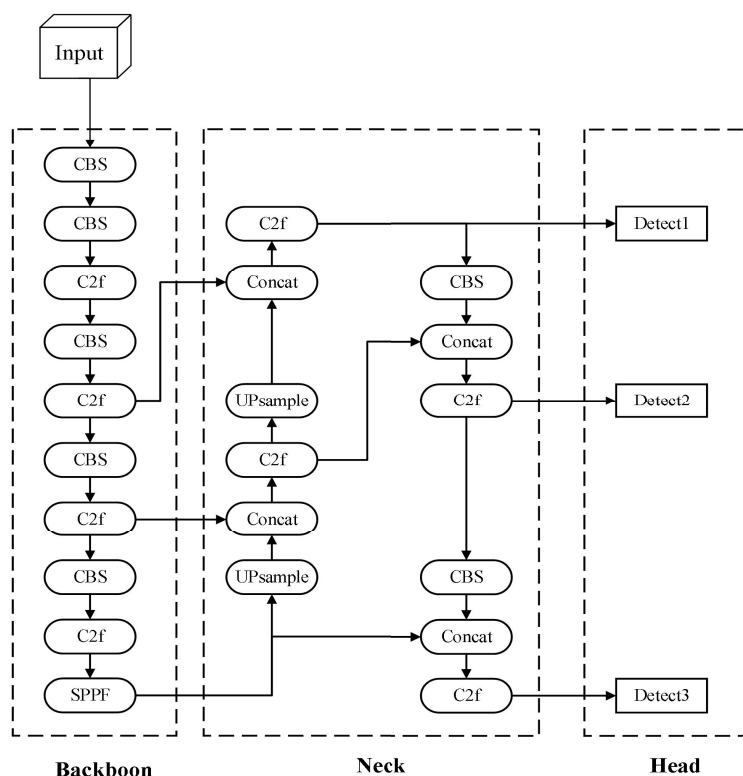


图 3-2 YOLOv8 网络结构

其中 q 表示预测和真实类别之间的交集, p 代表预测的类别数量, γ 代表指数幂。利用不对称的参数对正负样本不同处理, 主要通过对负样本应用衰减因子来减少其对损失的贡献, 以此实现对前景和背景影响的不等权重处理。具体而言, 对于正样本, 通过使用 q 值来加权, 使得那些与真实边界框重叠度高的样本对损失有更大的贡献, 这种做法有助于模型专注于高质量的正样本, 从而对模型的精度提升起到更显著的效果。对于负样本, 则通过对预测的类别得分 p 实施 γ 次幂的降权处理, 因为这种处理会使得 p 值降低, 从而有效减少负样本对总损失的影响。位置损失将在下面详细介绍。

3.2 基于注意力机制改进的 YOLOv8 算法

3.2.1 注意力机制介绍

注意力机制是一种在神经网络中广泛应用的技术, 其核心功能是通过为输入数据的不同部分分配不同的权重, 使网络更加集中于那些关键信息, 从而提升了模型的整体性能和精确度。此外, 注意力机制有助于增强模型的鲁棒性, 并防止过度拟合。在实现上, 注意力机制通常作为独立的网络块集成在模型结构中, 包括传统的局部特征提取和滑动窗口等技术都可视为其变体。

具体到类型，注意力机制可分为三种：(1)通道注意力机制，例如 SENet 模块，它通过添加额外的网络层来为每个通道的特征图分配权重以确定其重要性^[91]。(2)空间注意力机制，如 SAM 模块，它聚焦于对特定空间区域的重视^[92]。(3)混合注意力机制，主要通过结合通道和空间注意力，如 BAM 和 CBAM 模块^[93]。SENet 利用全局平均池化评估通道的重要性，但这种方法可能会忽略空间细节。相比之下，CBAM 结合了全局平均池化和全局最大池化技术，有效地保留了信息，同时增强了对通道和空间细节的关注，因此能够实现更优的性能。

CBAM 的网络结构如图 3-3 所示：CBAM 包含两个模块，第一个为通道模块，第二个为空间模块，两者串行组成。

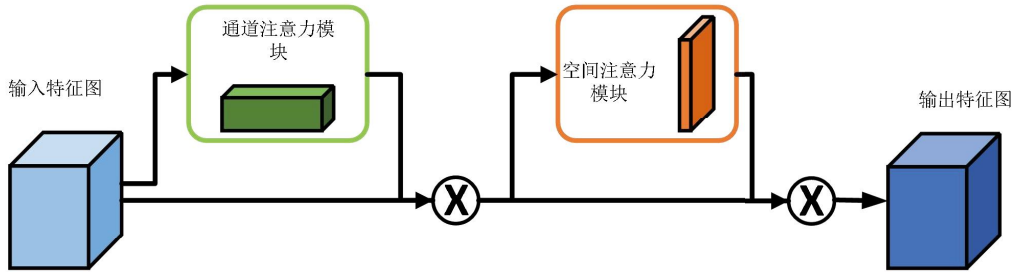


图 3-3 CBAM 网络结构

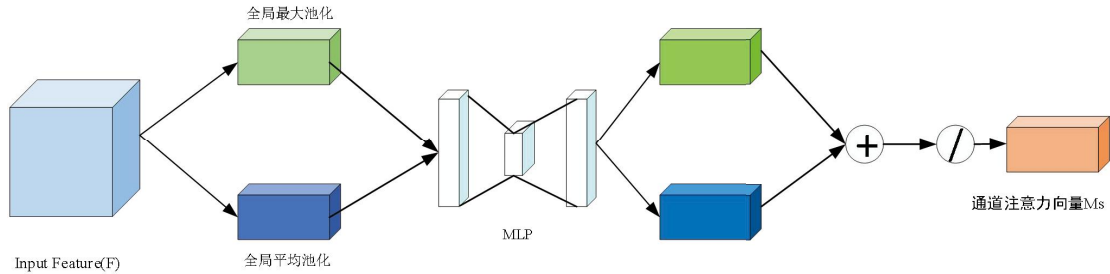


图 3-4 通道注意力模块网络结构

通道注意力机制的基本原理如图 3-4 所示。在第一步，对输入特征图像进行最大值池化和平均值池化，目的在于聚合空间信息，进而得到二个维度均为 C 的池化特征图像，分别命名为 F_{avg} 和 F_{max} 。随后，将这二种特征图都注入到一个具有隐层的多层感知器(MLP)中， MLP 将提供两个 $1 \times 1 \times C$ 尺寸的多通道特征注意力图。为减少模型数据，将 MLP 的隐层神经元数量设定为 C/r ，其中， r 表示压缩比例，将这两个通道注意力图相加后，得到最终的通道注意力图，该图将用于进一步加权输入特征图的各个通道。

M_c 可用如下公式表示：

$$\begin{aligned}
M_c(F) &= \sigma \left(MLP \left(AvgPool(F) \right) + MLP \left(MaxPool(F) \right) \right) \\
&= \sigma \left(W_1 \left(W_0 \left(F_{avg}^c \right) \right) + W_1 \left(W_0 \left(F_{max}^c \right) \right) \right)
\end{aligned} \tag{3-2}$$

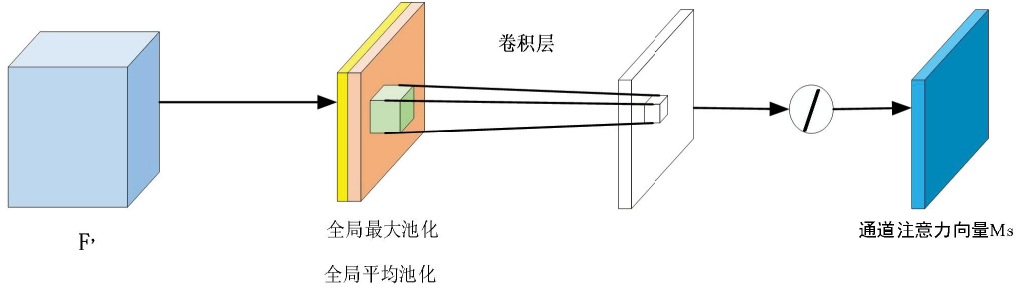


图 3-5 空间注意力模块网络结构

空间注意力模块的结构如图 3-5 所示。首先，对 F 执行通道维度进行平均池化和最大池化操作，产生两个空间维度的特征图 F_{avg} 和 F_{max} ，这两个特征图的尺寸均为 $1 \times H \times W$ 。接着，将这两个特征图在通道维度上进行合并，形成一个综合的空间特征图。最后，使用一个 7×7 卷积层处理拼接后的特征图，从而产生最终的空间注意力图，这个注意力图将用于增强模型对关键空间信息的关注。

$$\begin{aligned}
M_s(F) &= \sigma \left(f^{7 \times 7} \left([AvgPool(F); MaxPool(F)] \right) \right) \\
&= \sigma \left(f^{7 \times 7} \left([F_{avg}^s; F_{max}^s] \right) \right)
\end{aligned} \tag{3-3}$$

Google Brain 在 2017 年提出了 Transformer 模型，其独特之处在于整个网络仅由自注意力层和前馈网络构成，用以执行编码和解码任务^[94]。相对于传统的 CNNs，Transformer 的主要优势是能通过自注意力机制有效捕捉长距离的上下文信息，这使得模型能够处理全局信息，尽管这种方法在处理复杂输入时可能会导致较大的计算量和内存需求。针对这些不足，有学者已提出了多种优化策略，如局部注意力机制(Swin-T)^[95]、轴向注意力(CSwin-T)^[96]和扩张注意力(Multi-scale axis vit)^[97]以及可变形注意力^[98]。Lei Zhu 对这些 Transformer 变种进行了研究，首次引入了一种创新的动态稀疏注意力机制，命名为双层路由注意力(Bi-level Routing Attention, BRA)^[99]。

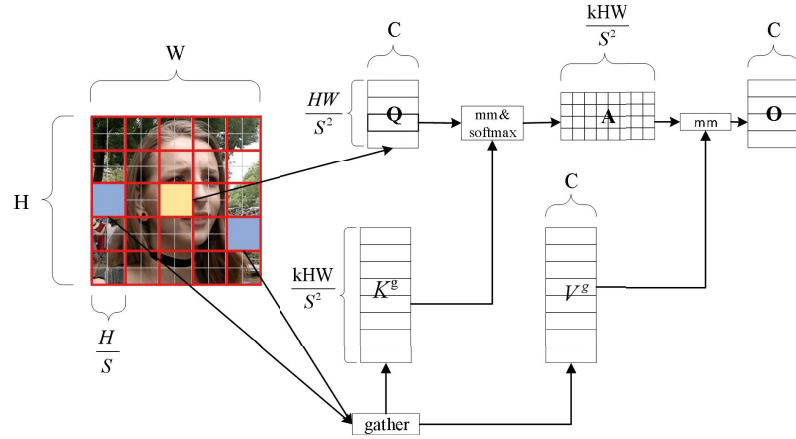


图 3-6 BRA 示意图

图 3-6 为 BRA 模块的示意图。该模块通过选择前 k 个最相关窗口中的键值对，并利用稀疏操作避免对最不相关区域的计算，从而有效减少参数量和计算量。这种方法主要依赖于密集矩阵乘法运算，不仅有利于在 GPU 上进行训练，还能加速服务器端的推理过程。

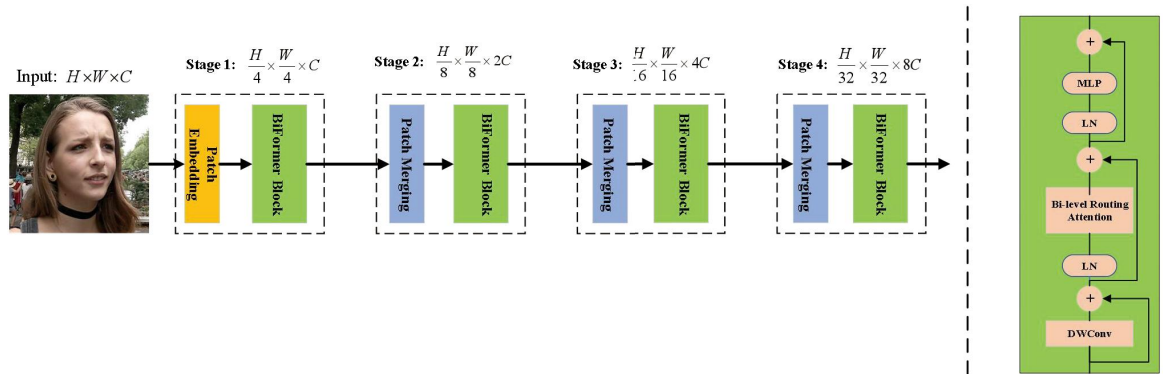


图 3-7 左图:BiFormer 的整体架构，有关配置见表 1。右图:BiFormer 模块的详细信息

基于 BRA 技术，有学者研发了名为 BiFormer 的视觉网络新架构，如图 3-7 所示，该架构采用查询自适应机制集中关注少数相关标记，避免了不相关标记的干扰。BiFormer 因其出色的计算性能，在目标检测、实例分割、图像分类和语义分割等多个视觉任务中有卓越的表现。

基于 BRA 模块开发的 BiFormer 架构，它采用了常见的视觉变换器设计，并实施了四级金字塔结构，实现了 32 倍的下采样。具体来看，BiFormer 在首个阶段采用重叠的块嵌入技术，在第二至第四阶段则通过块合并模块降低空间分辨率并增加通道数量。此外，BiFormer 块被用于进行特征变换。值得一提的是，每个块的开头都采用深度卷积来隐式地编码位置信息。在这之后，BiFormer 应用

BRA 模块进行交叉位置关系的建模，以及扩展率为 2 的两层多层感知机(MLP)模块，用于加强每个位置的嵌入处理。表 3-1 展示了不同模型变体的网络宽度和深度。

表 3-1 BiFormer 不同模块的网络配置

Models	Channels	Blocks	Params/M	FLOPs/G
BiFormer-T	64	[2,2,8,2]	13	2.2
BiFormer-S	64	[4,4,18,4]	13	4.5
BiFormer-B	96	[4,4,18,4]	13	9.8

3.2.2 改进思路

BiFormer 注意力机制通过区域之间的有向图路由来分部优化特征图，这不仅使其能在全局范围内更准确地定位目标，还能在局部区域中更精确地捕捉感受野内容。特别是基于双层路由的 BiFormer 注意力机制具有更出色的感受野选取能力。但对于中大尺寸目标时，这种细致的优势可能不太明显，因此，本文计划将 BiFormer 与 YOLOv8 相结合来实现小目标检测，并在理论上是可行性的。

图 3-8 展示了添加了 BiFormer 注意力机制后的 YOLOv8 算法网络结构。

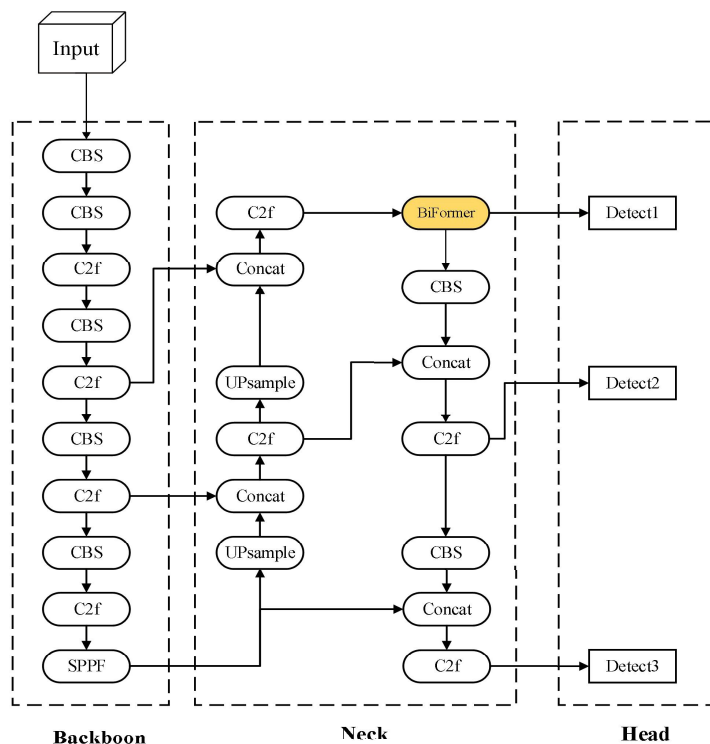


图 3-8 YOLOv8n+BiFormer 网络结构

3.3 基于轻量化网络改进的 YOLOv8 算法

虽然 YOLOv8 采用 CSPDarknet-53 作为主干网络在精度方面表现出色,但由

于其庞大的网络结构和计算量,使得模型在计算能力较低的终端设备上运行缓慢。因此,适应低端硬件设备并满足工业界对快速检测的需求成为了一个急需解决的问题。目前,学者们已提出了多种深度模型加速技术,这些技术包括网络剪枝、合并批量归一化(BN)层、采用轻量化设计、以及量化和张量分解等方法。尽管上述方法可以加速模型,但它们通常需要对模型进行再训练,可能会破坏原有网络结构并降低精度。因此,本研究选择采用轻量化的网络架构来改进 YOLOv8,以提升其在低端设备上的检测速度。轻量化结构中具有代表性的包括 MobileNet^[100-104]系列和 GhostNet^[105], 本节将对两者进行对比分析进而选择替换的轻量化结构。

3.3.1 Mobile Net 网络

3.3.1.1 Mobile Netv1

对图像的空间和通道特征同时进行处理是卷积运算过程中的主要内容,根据输入数据的复杂性调整卷积核数量,这常会导致模型参数数量急剧增加。为了解决这一问题,诞生了深度可分离卷积技术。这种技术通过将传统卷积运算分解为两部分:专注于空间特征的深度卷积和处理通道特征的逐点卷积。这一方法在保证模型精度的同时,大幅减少模型的参数量。相关的网络结构详见图 3-9。

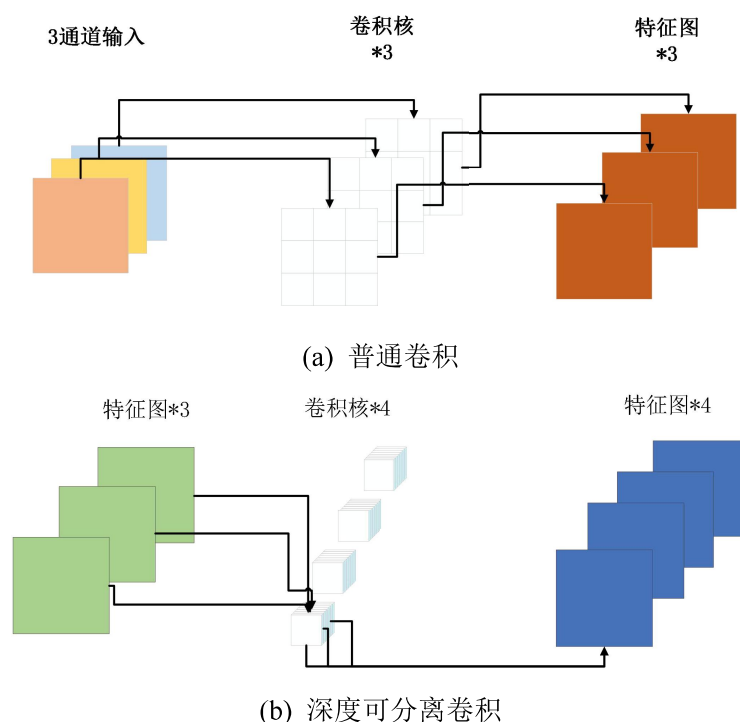


图 3-9 普通卷积和深度可分离卷积

如果输入特征图的尺寸为 $L \times L \times m$ ，并且使用尺寸为 $C \times C \times m \times n$ 的卷积核进行计算，那么所需的计算量将是 $L \times L \times m \times n \times C \times C$ 。深度可分离卷积会把每张特征图和与其相匹配的卷积核进行运算，接下来通过 1×1 的逐点卷积对每个特征图进行卷积运算，以确保通道数的一致性。

参数量比较如下：

$$r = \frac{L \times L \times m \times n \times C \times C + L \times L \times m \times n}{L \times L \times m \times n \times C \times C} = \frac{1}{N} + \frac{1}{C \times C} \quad (3-4)$$

借助深度可分离卷积，MobileNetv1 设计了一种高效的神经网络，显著减少了训练参数的数量，同时加快了推理速度，Mobile Netv1 的结构如下表 3-2 所示。

表 3-2 MobileNetv1 网络结构

操作类型 /步长	滤波器形状	输入尺寸
Conv /s2	3*3*3*32	224*224*3
Conv dw /s1	3*3*32 dw	112*112*32
Conv /s1	1*1*32*64	112*112*32
Conv dw /s2	3*3*64 dw	112*112*64
Conv /s1	1*1*64*128	56*56*64
Conv dw /s1	3*3*128 dw	56*56*128
Conv /s1	1*1*128*128	56*56*128
Conv dw /s2	3*3*128 dw	56*56*128
Conv /s1	1*1*128*256	56*56*128
Conv dw /s1	3*3*256 dw	28*28*128
Conv /s1	1*1*256*256	28*28*256
Conv dw /s2	3*3*256 dw	28*28*256
Conv /s1	1*1*256*512	28*28*256
Conv dw /s1	3*3*512 dw	14*14*256
Conv /s1 *5	1*1*512*512	14*14*512
Conv dw /s2	3*3*512 dw	14*14*512
Conv /s1	1*1*512*1024	7*7*512
Conv dw /s2	3*3*1024 dw	7*7*1024
Conv /s1	1*1*1024*1024	7*7*1024
Avg pool /s1	Pool 7*7	7*7*1024
FC /s1	1024*1000	1*1*1024
Softmax /s1	classifier	1*1*1000

dw 表示深度卷积， 1×1 的卷积核即为逐点卷积，s1 表示步长为 1。

3.3.1.2 Mobile Netv2

尽管 MobileNetv1 显著降低了计算量，但其架构主要由简单的层堆叠组成，这可能会导致梯度消失问题，从而限制模型的泛化能力。为解决这些限制，

MobileNet V2 采用了残差结构,通过首先使用 1×1 的卷积来实现降维,减少输入通道数,接着使用 3×3 的深度可分离卷积来执行核心特征运算,随后通过 1×1 卷积来实现再次升维,优化了信息流。这种结构,通常称为倒置残差块(Inverted Residual Block),形似柳叶,具有两端窄而中间宽的特点,详细结构见表 3-3 所示。

表 3-3 MobileNetV2 网络结构

输入	卷积操作	t	c	n	s
224*224*3	Conv, 3×3	-	32	1	2
112*112*32	bottleneck, 3×3	1	16	1	1
112*112*16	bottleneck, 3×3	6	24	2	2
56*56*24	bottleneck, 3×3	6	32	3	2
28*28*32	bottleneck, 3×3	6	64	4	2
14*14*64	bottleneck, 3×3	6	96	3	1
14*14*96	bottleneck, 3×3	6	160	3	2
7*7*160	bottleneck, 3×3	6	320	1	1
7*7*320	Conv, 1×1	-	1280	1	1
7*7*1280	average pooling, 7×7	-		1	-
1*1*1280	Conv, 1×1	-	k	-	

t:拓展因子,可调节输入通道的倍数;c:输出特征图的通道数;n: bottleneck 的重复次数;s:步距(若有多个 bottleneck,则只针对第一层,其余都为 1)

(3) Mobile Netv3

MobileNet V3 在其关键组成单元“Benck”中继承了 MobileNet V2 的瓶颈结构,并通过将深度可分离卷积和通道注意力机制(SE 模块)结合使用,以增强特征处理能力。具体而言,MobileNet V3 的 Benck 模块首先通过 1×1 卷积对特征进行升维,然后利用 3×3 的深度可分离卷积进行高效的特征提取,最后将提取的特征送入 SE 模块,以增强关键通道的表达能力。这种集成了通道注意力机制的结构设计,进一步提升了模型的性能和效率,其详细网络结构可以参见表 3-4 所示。

表 3-4 MobileNetV3 网络结构

Input	Operator	Exp size	Out	SE	NL	s
224*224*3	Conv, 3×3	-	16	-	HS	2
112*112*32	bottleneck, 3×3	16	16	-	RE	1
112*112*16	bottleneck, 3×3	64	24	-	RE	2
56*56*24	bottleneck, 3×3	72	24	-	RE	1
56*56*24	bottleneck, 5×5	72	40	✓	RE	2
28*28*40	bottleneck, 5×5	120	40	✓	RE	1

续表 3-4

28*28*40	bottleneck, 5*5	120	40	√	RE	1
28*28*40	bottleneck, 3*3	240	80	-	RE	2
14*14*80	bottleneck, 5*5	200	80	-	HS	1
14*14*80	bottleneck, 5*5	184	80	-	HS	1
14*14*80	bottleneck, 5*5	184	80	-	HS	1
14*14*80	bottleneck, 5*5	480	112	√	HS	1
14*14*112	bottleneck, 5*5	672	112	√	HS	1
14*14*112	bottleneck, 5*5	672	160	√	HS	2
14*14*112	bottleneck, 5*5	960	160	√	HS	1
7*7*160	bottleneck, 5*5	960	160	√	HS	1
7*7*160	Conv, 1*1	-	960	-	HS	1
7*7*960	average pooling, 1*1	-	-	-	-	1
1*1*960	Conv, 1*1, NBN	-	1280	-	HS	1
1*1*1280	Conv, 1*1, NBN	-	k	-	-	1

SE:表示是否添加 SE 模块; NL: 表示采用的激活函数; S:表示步长; Exe size: 表示在每一个 block 第一个 1*1 的卷积核进行升维时的卷积核个数; NBN: 表示不使用 BN 操作

3.3.2 Ghost Net 网络

Ghost Net 是一款轻量级的特征提取网络, 设计目的在于提高模型的运行效率。在常规的卷积网络中, 输入数据经过处理会生成许多相似且冗余的特征层, 这些层对于模型的功能非常关键。为了优化这一点, Ghost Net 通过使用较少的 3×3 卷积核来创建基础的卷积层, 然后利用简单的线性操作来产生这些冗余的特征层, 从而显著降低了模型的参数量。这种设计不仅减轻了计算负担, 还保持了特征提取的效能, 相关的网络结构详细信息见表 3-5 所示。

表 3-5 Ghost Net network structure

输入	卷积操作	输出	通道数	步长	卷积核
512*512*3	Conv2D, 3*3	256*256*16	16	2	3*3
256*256*3	GBN	256*256*16	16	1	3*3
256*256*16	GBN	128*128*48	48	2	3*3
128*128*48	GBN	128*128*72	72	1	3*3
128*128*72	GBN	64*64*72	72	2	5*5
64*64*72	GBN	64*64*120	120	1	5*5
64*64*120	GBN	32*32*240	240	2	3*3
32*32*240	GBN	32*32*200	200	1	3*3
32*32*200	GBN	32*32*184	184	1	3*3
32*32*184	GBN	32*32*184	184	1	3*3
32*32*184	GBN	32*32*480	480	1	3*3

续表 3-5

32*32*480	GBN	32*32*672	672	1	3*3
32*32*672	GBN	16*16*672	672	2	5*5
16*16*672	GBN	16*16*672	672	1	5*5
16*16*960	GBN	16*16*960	960	1	5*5
16*16*960	GBN	16*16*960	960	1	5*5
16*16*960	GBN	16*16*960	960	1	5*5

3.3.3 改进方法

由于 MobileNetv 系列的特殊结构，直接替换 YOLOv8 的主干网络可能会影响其特征提取性能。因此，本文选择使用更轻量级的 GhostNet 作为主干特征提取网络，以便在速度和精度之间达到一个平衡。GhostNet 利用 Ghost 模块来减少模型的计算负担，同时保持有效的特征提取。这些 Ghost 模块的工作过程在图 3-10 中有所展示，而目前广泛使用的 Ghost 模块类型则在图 3-11 中详细说明。

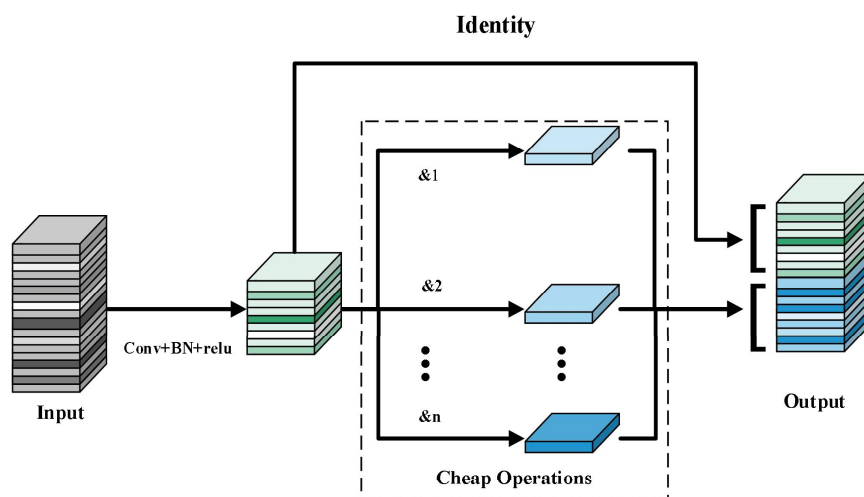


图 3-10 Ghost module

Ghost module 通过两个主要步骤实现与传统卷积相同效果的特征图生成，但计算成本更低。首先，该模块使用较少的卷积核进行操作(例如，如果使用 128 个卷积核，Ghost module 则只用 64 个，从而直接减少了一半的计算量)。接下来，执行“cheap operations”，用图中的&1、&2、&n 来表示，&用来表示 3*3 或者 5*5 的深度卷积。这种设计策略有效地降低了参数数量和运算需求，同时尽可能保持了特征提取的效能。

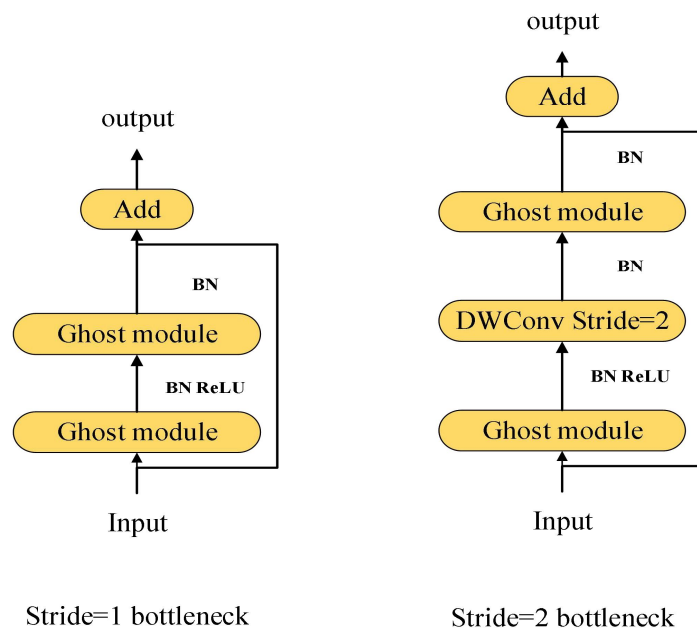


图 3-11 GhostNet 两种结构组成

从图 3-11 可知，两种 Ghost 模块的主要区别在于右侧模块增加了深度卷积层，这一设计旨在加强特征提取能力。尽管在检测精度上两种模块表现相似，右侧模块由于深度卷积层的加入而在检测速度上稍显缓慢，并且其训练时的模型权重也较大。考虑到实时性需求及移动设备的应用场景，本文选用了图中左侧的模块结构来构建轻量级 GhostNet 网络。此模块由两个功能不同的相同 Ghost 模块组成，第一个 Ghost 模块主要减少输出通道数，而第二个 Ghost 模块则用于增加通道数，以使其匹配目标通道数。最终，这两个模块的输出将通过捷径通道进行融合，实现特征图在通道维度上的有效整合。

3.4 基于损失函数改进的 YOLOv8 算法

3.4.1 目标检测损失函数

在目标检测中，同一目标的周围区域常会生成多个预测框。为了筛选出最佳的边界框并减少冗余结果，通常采用非极大值抑制(NMS)。这一技术利用交并比(IoU)作为重要判断依据，以评估预测框与目标真实框之间的一致性。IoU 计算方法是通过测量预测框与真实框的交集面积与它们的并集面积的比例，得到一个介于 0 到 1 之间的值。IoU 值越接近 1，表示预测结果与真实标注的相似度越高，预测框与实际框的重叠率越大，表明匹配程度更高且回归损失较小。IoU 的计算公式可以表示为：

$$IoU\{A,B\} = \frac{A \cap B}{A \cup B} \quad (3-5)$$

在目标检测和分割任务的训练过程中，IoU 常被用作损失函数的一部分，帮助网络更准确地学习。然而，传统的 IoU 计算存在一些不足：首先，尽管非极大值抑制(NMS)广泛应用，但交并比也不总是一定能够把预测框与真实框的重合程度准确地表达出来。如图 3-12 所示，即便在视觉上重合程度不同，IoU 值可能相同，这是因为它仅计算重叠区域与总区域的比例。例如，虽然左侧情况的匹配看起来最为理想，而右侧最差，IoU 值却无法区分这些差异。其次，如果预测框与真实框没有任何重叠，根据公式 $IoU = 0$ ，此时损失函数也为 0，导致无梯度可回传，从而无法基于这些情况进行学习优化。

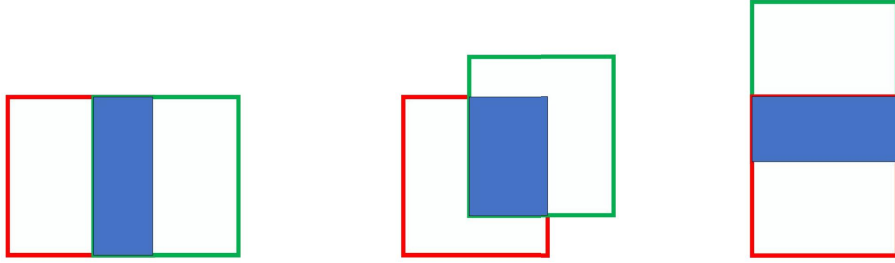


图 3-12 不同重合部分但相同 IoU 值

针对真实框和预测框没有交集交并比为零的情况，GIoU 损失函数是对传统 IoU 的一种反映位置关系的改进算法^[106]，函数公式(3-6)所示。

$$\begin{aligned} L_{GIoU} &= 1 - GIoU \\ &= 1 - IoU + \frac{C - (A \cap B)}{C} \end{aligned} \quad (3-6)$$

在 IoU 计算中，尽管引入了包含预测框和真实框的最小重叠矩形区域 C 的面积能在一定程度上缓解预测框与真实框不重叠时导致的梯度消失问题，但这种方法仍有其局限性。特别是当真实框完全包含预测框(或相反)时，如图 3-13 所示，GIoU 损失中的面积惩罚项将归零。在这种情况下，GIoU 损失将仅反映基本的 IoU 值，从而无法准确地反映预测框与真实框之间的实际空间关系。这导致损失函数不能充分捕捉包含关系下的位置偏差，因此 GIoU 损失在处理包含关系的情况时效果受限。

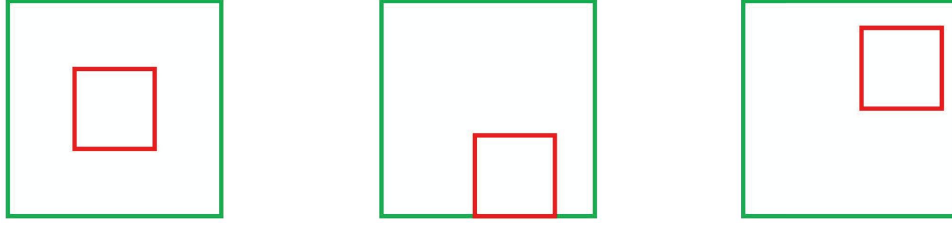


图 3-13 预测框与真实框包含重叠

为解决 IoU 和 GIoU 损失函数主要关注重叠区域，而忽视框的位置关系的问题，DIoU 损失函数引入了一种基于预测框与真实框中心点间欧氏距离的评估方法^[107]。DIoU 损失函数的核心在于测量预测框和真实框的中心点距离，以此判断它们的空间误差。具体数学表达如公式(3-7)所示：

$$\begin{aligned} L_{GIoU} &= 1 - DIoU \\ &= 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} \end{aligned} \quad (3-7)$$

其中， c 表示两框最小围绕矩形的对角线， b 和 b^{gt} 分别为两框的中心点， d 为这两点之间的欧几里得距离。这种方法不仅考虑了重叠面积，也有效地考虑了位置精度，从而提供了更全面的损失评估。

CIoU 损失函数是对 GIoU 的进一步优化，它整合了 GIoU 和 DIoU 的优点，并引入了长宽比来提高预测框对真实框形状的拟合精度^[108]。CIoU 不仅考虑框的重叠区域和中心点距离，还专注于预测框与真实框的长宽比匹配，使得损失计算更全面反映预测框与真实框之间的差异，具体公式如(3-8)所示：

$$\begin{aligned} L_{CIoU} &= 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha\gamma \\ \gamma &= \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \\ \alpha &= \frac{\gamma}{(1 - IoU) + \gamma} \end{aligned} \quad (3-8)$$

其中包含一个长宽比差异项，用于评估预测框和真实框的长宽比适配度。此外，公式中的平衡参数用于调整重叠部分与长宽比差异的权重，以实现更精确的损失评估。在公式中， w^{gt} 和 h^{gt} 分别代表真实框的宽度和高度，而 w 和 h 表示预测框的宽度和高度。

3.4.2 改进方法

YOLOv8 的损失函数主要由分类损失和定位损失这两大部分构成。在分类损

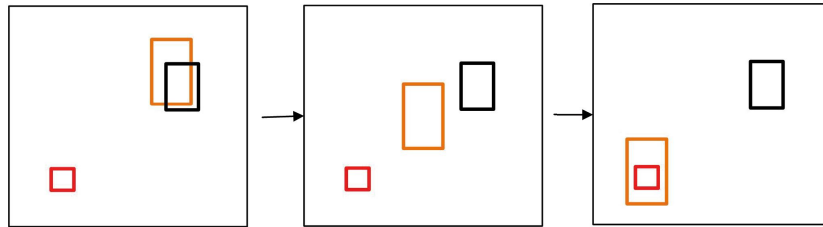
失方面，使用的是 VFL Loss^[109]，目的是优化分类的准确性。对于定位损失，传统的方法结合了 DFLLoss 和 CIoU Loss。CIoU Loss 在特定情况下(例如当预测框和真实框的长宽比完全相同时)可能导致惩罚项为 0，这会影响梯度的计算，因为 CIoU 中宽度和高度的梯度是相反数，导致不能同时对宽度和高度进行有效调整。

为了解决这一问题，YOLOv8 在最新版本中采用了 EIoU Loss^[110]。EIoU Loss 在 CIoU 的基础上做了改进，该方法将预测框和真实框的长宽比的影响因子进行了分离处理，分别独立计算长度和宽度，有效避免了上述问题。EIoU 由三个部分组成：IoU 损失、位置损失和距离损失。其中，长宽损失通过直接最小化预测框和真实框在高度和宽度上的差异来实现。这不仅加快了模型的收敛速度，还显著提高了定位精度。这种损失函数的改进，使得 YOLOv8 能更精确地定位目标，尤其是在处理具有复杂形状和大小变化的对象时，表现得更为出色。

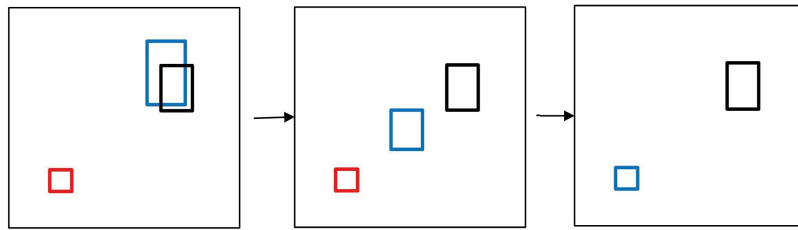
其计算公式如下：

$$\begin{aligned} L_{EIoU} &= L_{IoU} + L_{dis} + L_{asp} \\ &= 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \frac{\rho^2(w, w^{gt})}{c_w^2} + \frac{\rho^2(h, h^{gt})}{c_h^2} \end{aligned} \quad (3-9)$$

CIoU 和 EIoU 损失函数在预测框迭代进程的对比如图 3-14 所展示。



(a) CIoU 损失函数预测框迭代过程



(b) EIoU 损失函数预测框迭代过程

图 3-14 预测框迭代过程对比

如图 3-14 所示，橙色框和蓝色框展示了使用 CIoU 损失函数时的预测框回归过程，其中红色框代表真实框，黑色框为预设的锚框。在 CIoU 损失函数下，尽管预测框能够逐渐回归至真实框的位置，但其尺寸在迭代过程中并未进行有效调

整。相反，使用 EIoU 损失函数时，预测框的尺寸会根据迭代进程不断进行调整，这不仅反映在预测框与真实框之间的尺寸匹配上更为精准，而且计算量方面也减少了，使得特征提取的效率提高了。基于这些观察，本文选择采用 EIoU 损失函数替换 CIoU 损失函数，以期在保持损失计算精度的同时加快模型训练的速度。

3.5 实验结果及分析

3.5.1 实验环境与参数设置

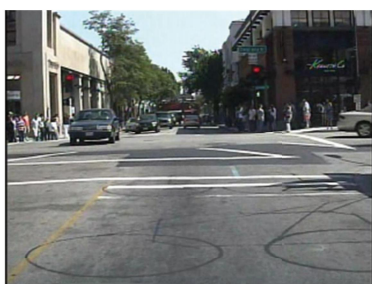
实验环境配置如表 3-6 所示。

表 3-6 实验环境配置表

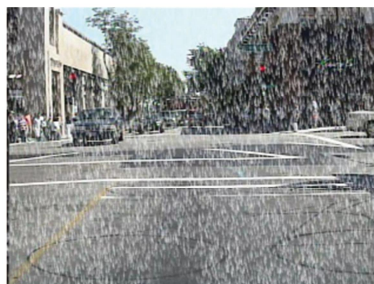
项目	具体内容
操作系统	Windows 10
CPU	Intel i5-12400F
内存	1TB
GPU	GeForce RTX 3060Ti
显存	8GB
程序语言	Python
学习框架	Pytorch 1.13、TorchVision 1.9.0、 CUDA 11.6、 CuDNN 8.4.5

3.5.2 数据集准备

为了使检测效果更具有泛化性，在 Crowd Human 和 Caltech Pedestrian 两个数据集分别进行试验，下文将分别用数据集 1 和数据集 2 表示，然后对数据集进行背景增强，模拟复杂天气(雨、雾、雪)情况下的行人检测，如图 3-15 所示。然后将数据集按 8:2 分为训练和验证集。



(a) 原始图片



(b) 雨天增强

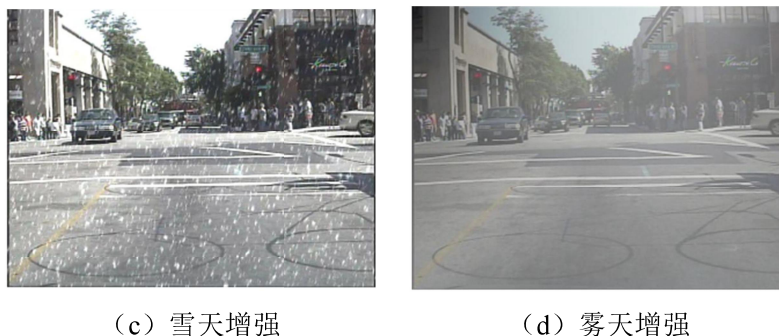


图 3-15 数据集增强类型示例

3.5.3 实验评价指标

混淆矩阵常用于评估分类模型的表现和性能。它通过表格形式展示模型的预测结果与实际情况的对比。在二分类任务中，混淆矩阵由四个基本元素组成，分别是真正极(True Positive, TP)、假正极(False Positive, FP)、真负极(True Negative, TN)和假负极(False Negative, FN)，如表 3-7 所示。这些元素分别定义如下：

真正极(TP)：正样本预测结果为正。

假正极(FP)：负样本预测结果为正。

真负极(TN)：正样本预测结果为负。

假负极(FN)：负样本预测结果为负。

通过混淆矩阵，不仅可以计算出准确率、精确率、召回率等关键性能指标，还能深入分析模型在不同类别上的预测误差，从而对模型进行进一步的优化。混淆矩阵因此成为了评估和完善分类模型不可或缺的一部分。

表 3-7 实验评价指标

真实结果 \ 预测结果	预测结果	
	Positive(预测结果为正)	Negative(预测结果为负)
True	TP	FN
False	FP	TN

在目标检测中，性能通常通过以下几个关键指标来衡量：

1.召回率(Recall)：也称为查全率，这个指标反映了模型正确检测到的正样本数占测试集中所有正样本总数的比例。

公式表示为：

$$Pecall = \frac{TP}{TP+FN} \quad (3-10)$$

2.精确率(Precision)：这个指标衡量的是正确检测到的正样本数占模型所有检测到的样本总数的比例。

公式表示为:

$$Precision = \frac{TP}{TP+FP} \quad (3-11)$$

3.准确率(Accuracy): 表示所有正确检测到的样本(包括正样本和负样本)占总样本数的比例。

公式表示为:

$$AP = \int_0^1 P(R)dR \quad (3-12)$$

准确率的曲线图像中, 曲线与 x 轴之间的区域用微积分计算面积就是平均精度, 这是评价和判断模型优劣的最主要也是最直观的指标之一, 也是人们使用最多的评估办法。而 mAP 则是整个数据集中所有参数类别的平均精度的平均值, 公式如下式(3-13)所示。

$$mAP = \frac{1}{N} \sum_{i=1}^N AP_i \quad (3-13)$$

mAP 会因为阈值条件不同产生不同的指标, 主要有 mAP@0.5 和 mAP@[0.5:0.95]这两个, 也是使用最频繁的。其区别在于交并比值的不同, 当 Io U 达到 0.5 时为 mAP@0.5, 当 Io U 从 0.5 到 0.95 变化时为 mAP@[0.5:0.95]。一般来说, mAP@0.5 使用率更高, 其公式如下式(3-14)所示。

$$mAP@0.5 = \frac{1}{n} \sum_{k=1}^n AP@0.5_k \quad (3-14)$$

本文中, 选择 mAP@0.5 作为主要性能评价指标, 公式如 (3-14)所示。并结合模型的检测速度(FPS)和模型参数量(Params)来全面评估模型性能。mAP@0.5 反映了模型在召回率增加的过程中准确率的变化趋势, 是衡量模型综合性能的关键指标, 其数值越高, 表明模型在进行高精度边界回归时能力越强以及检测框与标准框拟合的效果越好。

3.5.4 模型的超参数

在开展消融实验前, 首先需设定 YOLOv8 网络模型的超参数。通过一系列实验, 发现当学习率调整到 0.01 时, 模型的训练损失降至最低点, 继续增加学习率则会导致损失重新上升。因此, 确定 0.01 作为最优学习率。此外, 根据实验结果, 动量(Momentum)设置为 0.9, 权重衰减(Weight Decay)设置为 0.0001, 批处理大小(Batch Size)设置为 16, 以下表 3-8 是详细的训练参数配置。

表 3-8 超参数设置

Name	Value
picture image size	640*480*3
epochs	150
Momentum	0.9
Weight-Decay	0.0001
optimizer	Adam
Learning rate	0.01
Patience	50
batch-size	16

3.5.5 实验结果分析

3.5.5.1 消融实验

通过消融实验来证明本文改进部分对最终检测能力的影响程度，首先在 Crowd Human 数据集上进行消融实验，下表是以 YOLOv8n 为初始对照组，通过增加 BiFormer 注意力机制，改进可变形卷积以及优化模型损失的实验。从表中数据可知：1.引入 BiFormer 注意力机制加强了模型的特征提取能力从而提升了对小目标检测的效果，进一步提高了检测模型的总体性能，mAP@0.5 相比 YOLOv8n 提升了 3.45%。2.轻量化改进：这种改进显著减少了模型的参数量，并大幅提升了处理速度，尽管精度略有下降 1.76%，但 FPS 从 57.33 提升至 62.17；3.损失函数改进：这一改动使得模型在精度和速度上都有小幅变化，尽管改善幅度有限。4.综合改进：当将注意力机制、轻量化和损失函数改进同时应用时，模型的参数量大幅减少，速度显著提升(FPS 从 57.33 提升至 60.28)，精度也从 88.32% 提高到 90.95%，增加了 2.63%。综合以上结果，改进后的模型在速度和精度方面均实现了有效提升，其消融对比曲线如图 3-17 所示。这些结果表明，采用综合改进策略可以显著提高 YOLOv8 模型的性能。

表 3-9 数据集 1 上改进的网络消融实验

模型	BiF	GhostNet	EIoU	Params(M)	mAP@0.5/%	FPS
YOLOv8n				3.00	88.32%	57.33
YOLOv8n_BiF	√			3.04	91.77%	52.39
YOLOv8n_Ghost		√		1.99	86.56%	62.17
YOLOv8n_EIoU			√	3.02	89.13%	56.92
Ours (本文模型)	√	√	√	1.99	90.95%	60.28

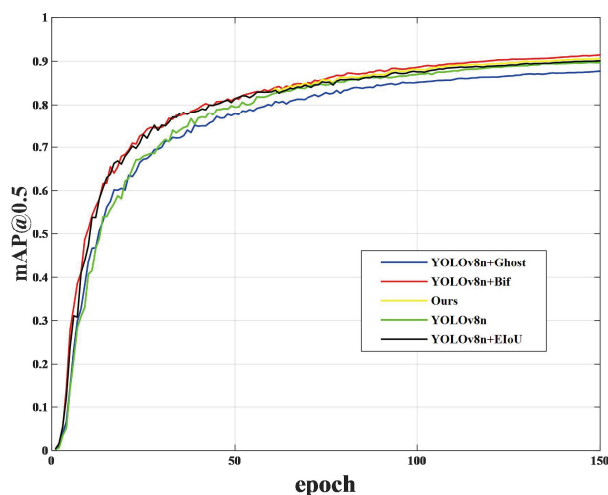


图 3-16 (a)数据集 1 消融实验 mAP 曲线对比

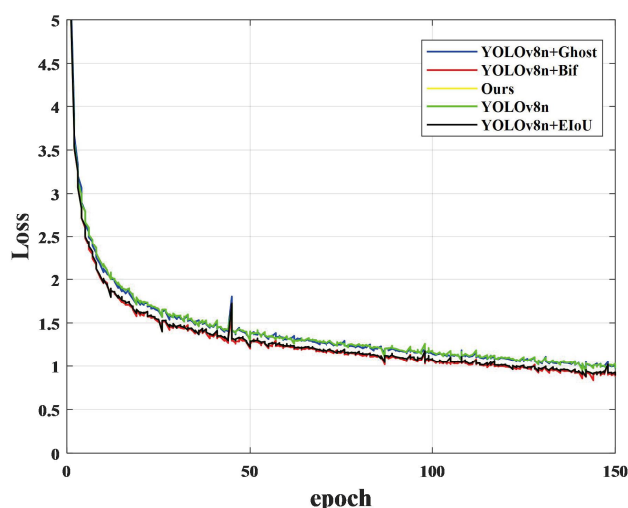


图 3-16 (b)数据集 1 消融实验损失函数曲线对比

然后在 Caltech Pedestrian 数据集上进行消融实验，原理与在 Crowd Human 数据集上一样。由表 3-10 可以看出，1.引入 BiFormer 注意力机制加强了模型的特征提取能力从而提升了对小目标检测的效果，进一步提高了检测模型的总体性能，mAP@0.5 相比 YOLOv8n 提升了 2.96%。2.轻量化改进：这种改进显著减少了模型的参数量，并大幅提升了处理速度，尽管精度略有下降 0.5%，但 FPS 从 62.50 提升至 68.96；3.损失函数改进：这一改动使得模型在精度和速度上都有小幅变化，尽管改善幅度有限。4.综合改进：当将注意力机制、轻量化和损失函数改进同时应用时，模型的参数量大幅减少，速度显著提升(FPS 从 62.50 提升至 66.49)，精度也从 94.80%提高到 96.10%，增加了 1.3%。综合以上结果，改进后的模型在速度和精度方面均实现了有效提升，其消融对比曲线如图 3-17 所示。

这些提升表明采用综合改进策略可以显著提高 YOLOv8 模型的性能。

表 3-10 数据集 2 上改进的网络消融实验

模型	BiF	GhostNet	EIoU	Params(M)	mAP@0.5/%	FPS
YOLOv8n				3.00	94.80%	62.50
YOLOv8n_BiF	√			3.04	97.76%	55.56
YOLOv8n_Ghost		√		1.99	94.30%	68.96
YOLOv8n_EIoU			√	3.02	95.30%	58.67
Ours (本文模型)	√	√	√	1.99	96.10%	66.49

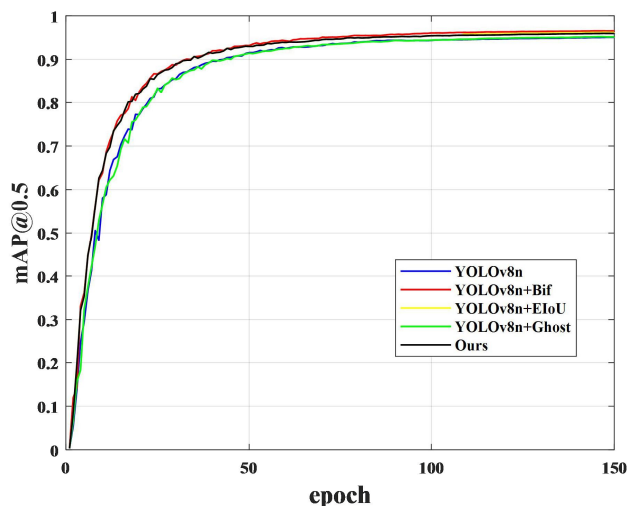


图 3-17 (a)数据集 2 消融实验 mAP 曲线对比

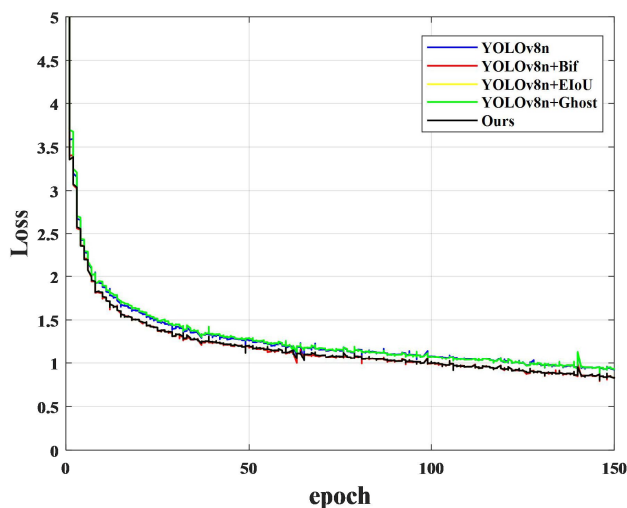


图 3-17(b)数据集 2 消融实验损失函数曲线对比

通过在两个数据集的测试集上选取样本，在未改进的初始 YOLOv8n 网络模型和本文所提模型上分别进行检测，以此来验证改进前后的算法检测精度的变化。效果对比如图 3-18 所示(上方图片为 Crowd Human 数据集选取，下方图片为 Caltech Pedestrian 数据集选取)。

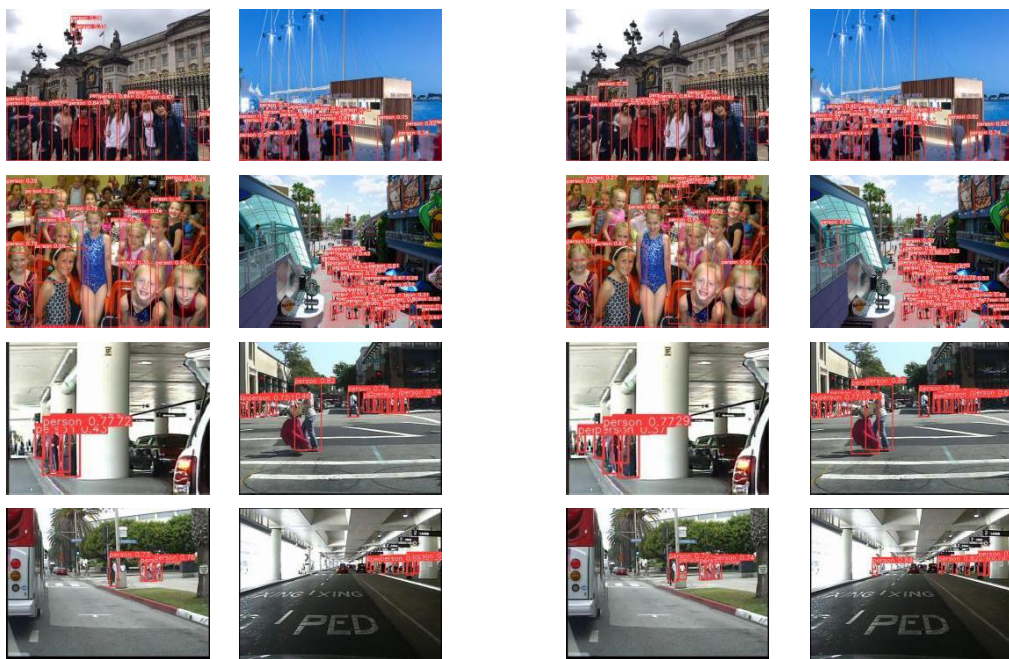


图 3-18 网络改进前(左)和改进后(右)检测效果对比。

3.5.5.1 对比实验

本文精心设计了 7 组对照实验，以验证所提改进算法的先进性和合理性。以 Crowd Human 数据集为对比实验的数据集，将本文所提的检测模型与几种主流算法进行详细的对比分析，其对比实验结果如表 3-10 所示，从三个关键方面评估了各模型的性能：模型参数量(Params)、检测精度(mAP@0.5)，以及检测速度(FPS)。

表 3-10 检测算法模型对比实验

模型	Params (M)	mAP@0.5	FPS
Faster_RCNN	137.2	0.835	16.54
YOLOv5s	7.0	0.941	47.61
YOLOv8s	11.1	0.970	29.58
YOLOv8m	25.8	0.969	25.19
D-FINE	31	0.824	60.1
YOLOv10l	25	0.824	68.49
本文模型	1.99	0.961	66.49

根据表中数据，本文所提算法与 Faster_RCNN 这种二阶段算法相比，在精度上高出 12.6%，在检测速度上，从 16.54 帧/秒提高到 68.49 帧/秒；YOLOv5s 相比，在精度上实现了 2%的提升，同时检测速度也从 47.61 帧/秒提高到 68.49 帧/秒，表明在多个方面优于 YOLOv5；与参数量更大、模型更庞大的 YOLOv8s 和 YOLOv8m 相比，虽然本文所提算法在检测精度上略有不足，差距仅在 0.9%

之内，但在检测速度方面，本文所提算法的表现远超这两种模型，达到其速度的大约 2.5 倍以上；与 D-FINE 和 YOLOv10l 等算法相比，无论是精度还是速度上都不逊色。此外，本文所提算法较小的体量使其非常适合部署在资源受限的小型设备上，这是其他算法难以实现的。

实验结果表明，无论是在模型的参数量、检测精度，还是在检测速度方面，本文所提的改进模型均表现出显著优势。这些对比数据不仅证明了改进模型的高效性，也突显了其在现有 YOLO 算法系列中的先进性。这些实验成果充分展示了改进算法在实际应用中的有效性。

3.6 本章小结

为了完成行人目标检测任务，本章提出了一种基于 Biformer 注意力机制和改进 YOLOv8 的行人检测方法，被称为 Bif-YOLO。并基于 Crowd Human 数据集以及 Caltech Pedestrian 数据集对 Bif-YOLO 进行训练，通过基于两个数据集构造的测试集完成测试，证明了 Bif-YOLO:基于改进 YOLOv8 的行人检测方法在行人检测任务中能够较好的检测出行人目标并完成标记，同时显著地提升了检测速度、检测效率、检测准确性以及网络的鲁棒性，并有效解决了在遮挡场景下行人目标的检测问题。但该方法依然存在不足，在检测一些密集的小目标群时，模型并不理想。

第四章 基于改进的 Deep Sort 行人跟踪模型

针对当前行人目标跟踪中存在的非线性运动建模偏差、遮挡场景轨迹恢复能力弱、多目标协同计算效率低等一系列问题,本章提出了一种基于改进 Deep Sort 行人跟踪方法,被称为 VOVNet-SORT。VOVNet-SORT 是一种新的混合检测跟踪网络,该网络将以改进的 YOLOv8 网络作为检测头部分,并对 Deep Sort 部分进行主要改进的多目标行人检测跟踪网络。

该方法的主要贡献如下:(1) 引入匀加速卡尔曼滤波模型:为了解决多目标跟踪中漏检和跟丢目标的问题,本文提出了使用匀加速模型代替传统恒速线性模型的方法;(2) 改进行人重识别模型:针对 Deep SORT 算法追踪速度较慢的问题,本文引入了一种行人重识别技术,以优化算法的推理速度和准确性。

本章节的结构如下:4.1 节系统地对 Deep Sort 多目标跟踪算法的架构进行分析;4.2 节详细阐述对 Deep Sort 多目标跟踪算法的改进步骤;4.3 显示改进后算法网络的相关实验和结果,并提供大量有关训练数据和训练过程的详细信息;4.4 节就实验结果和相关训练数据进行分析讨论并对本章节提出的 VOVNet-SORT 方法进行总结。

4.1 网络架构分析

Sort 算法作为一款高效的多目标跟踪工具,特别适用于在线和实时的目标跟踪应用。相比以往方法,Sort 在追踪数据匹配方面进行了优化,但其在面对遮挡等复杂场景时仍可能失去跟踪目标,且频繁更改轨迹 ID。在 Sort 的基础上,Deep Sort 算法引入了外观特征和级联匹配技术,显著减少了 ID 频繁更换的问题,因此显著提高了跟踪的准确性^[11]。本文主要采用 Deep Sort 算法追踪目标行人,并对其进行进一步的优化,确保了对行人的连续稳定追踪。

4.1.1 跟踪流程分析

根据对当前帧目标特征的关键信息进行分析,从而预测和提取下一帧中的目标数据是 DeepSort 算法的主要原理。其通过对目标的预测位置信息和真实位置

信息两者之间的正态分布重合部分进行分析，以此来实现对目标位置的跟踪预测，Deep Sort 的匹配效果如图 4-1 展示。

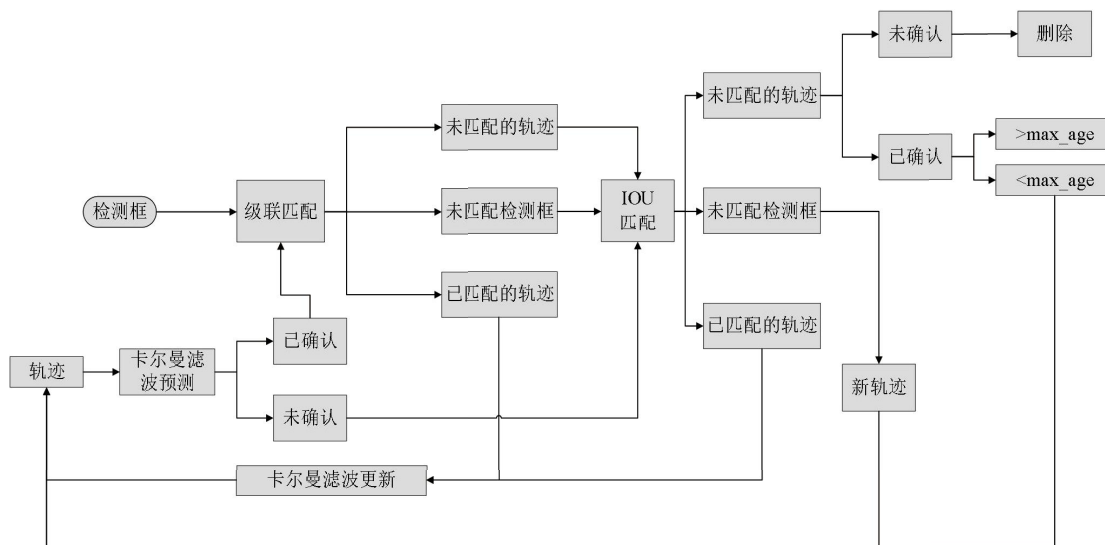


图 4-1 Deep SORT 流程图

图 4-1 是 Deep SORT 匹配流程图，首先当预测框与检测框匹配成功时，代表跟踪是成功的；其次对于没有匹配到的检测框、无法确定的轨迹、不能匹配的轨迹等都进行统一处理，即 IoU 匹配，如果匹配成功了，就可以对轨迹进行更新，否则就要重复匹配，直到连续三次匹配都失败，就放弃该路线并将其确定为新的轨迹；最后匹配失败的确定态轨迹，根据与上次更新的时间是否超过阈值来判断是继续预测还是删除。

Deep Sort 算法的主要跟踪过程包括四个基本步骤，详细描述如下：

1. 对于目标检测的初始帧，建立相应的跟踪轨迹。同时，对运动参数进行初始设定，并使用此滤波器预测下一帧中目标的位置，如图 4-2 所示的预测结果。

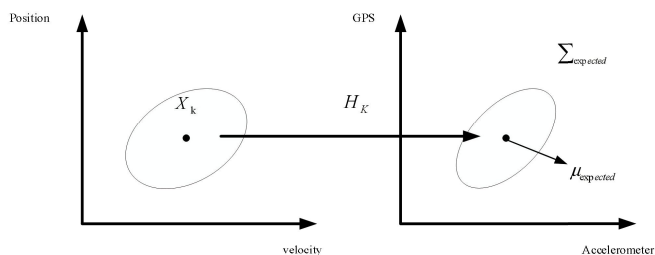


图 4-2 初始帧与预测帧相对位置

2. 在这一步中，将当前帧的真实框与上一帧的预测框进行匹配。通过计算马氏距离等参数得到一个代价矩阵，用于确定匹配的最佳目标检测框。图 4-3 展示

了通过这种计算得到的实际位置与预测位置信息的对比。

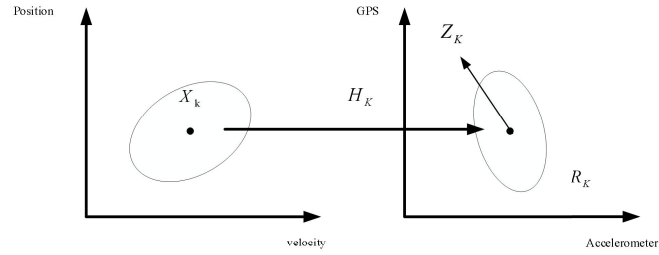


图 4-3 实际帧与预测帧相对位置

3.通过把代价矩阵融入进匈牙利算法中，再利用算法和级联匹配技术的共同计算，完成数据的有效关联并实现目标的线性匹配。接着，结合步骤 1 和步骤 2 中得到的预测帧和实际帧的位置信息及跟踪特征，确定最终的匹配结果。这一结果在图 4-4 中有所展示。

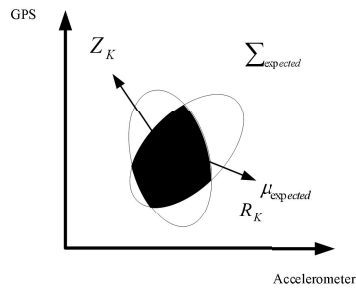


图 4-4 预测位置与实际位置的交互比

4、根据 3 中的算法来确定最优的数据分配，然后，重复执行 2 至 4 的过程，直至视频中的所有帧都被处理完毕。

4.1.2 运动模型分析

Deep SORT 算法通过融合目标物体的外观特征与运动状态信息构建复合度量体系，其技术实现是基于对余弦相似度与马氏距离的协同应用。传统欧氏距离在特征空间度量中存在显著局限性，主要源于多维特征量纲差异导致的协方差矩阵非对角化问题，这会造成特征权重分配失当的度量偏差。作为基于概率分布的协方差距离，马氏距离通过引入特征间协方差矩阵的逆运算，有效解决了多维度数据异方差性问题，使得距离计算具有统计无偏性。而余弦相似度作为向量空间的方向一致性度量，能够有效表征高维特征向量的角度差异，对外观形变具有鲁棒性。通过线性加权融合这两种互补的度量范式，算法构建出兼顾运动轨迹连续性与外观特征稳定性的复合判别准则，马氏距离公式如下。

$$d^{(1)}(i, j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i) \quad (4-1)$$

其中 d_j 表示第 j 个检测框, y_i 表示为第 i 条轨迹的滤波器预测框, S_i 表示 i 条轨迹之间的标准差矩阵。在马氏距离的实际应用中, 需预先定义合适的判别边界作为关联准则, 也就是所谓的阈值, 这一参数化过程具有明确的统计理论依据, 函数如式(4-2)所示:

$$y_{i,j}^{(1)} = I(d^{(1)}(i, j) \leq t^{(1)}) \quad (4-2)$$

其中 $y_{i,j}^{(1)}$ 代表一个指示器, 当马氏距离小于 $t^{(1)}$ 时, 代表成功匹配。

但是在遮挡场景下, 马氏距离对于目标距离的预测效果则不尽人意, 而在实景应用中, 目标被遮挡是极易发生的, 所以为了应对这种情况, 就必须引入外观信息, 主要方法是引入最小余弦值显示匹配程度来衡量准确度, 其计算公式如下:

$$d^{(2)}(i, j) = \min \{1 - r_j^T r_k^{(i)} \mid r_k^{(i)} \in R_i\} \quad (4-3)$$

式中 r_j 表示检测块 d_j 对应的特征向量, 其中 $\|r_j\| = 1$, r_k 为追踪目标成功关联的特征向量。使用余弦距离时也要设定阈值, 函数如式(4-4)所示:

$$y_{i,j}^{(2)} = I(d^{(2)}(i, j) \leq t^{(2)}) \quad (4-4)$$

其中 $y_{i,j}^{(2)}$ 也是代表一个指示器, 当马氏距离小于 $t^{(2)}$ 时, 代表成功匹配。

最后综合马氏距离和余弦距离的运动模型, 将两者加权结合起来, 作为最终的衡量标准, 如式(4-5)所示:

$$c(i, j) = \lambda d^{(1)}(i, j) + (1 - \lambda) d^{(2)}(i, j) \quad (4-5)$$

阈值函数加权之后如式(4-6)所示:

$$y_{i,j} = \prod_{n=1}^2 y_{i,j}^{(n)} \quad (4-6)$$

当综合指标在可接受范围内则被认为成功匹配。

4.1.3 特征提取分析

在行人目标运动状态发生剧烈动态变化时, 基于马氏距离的关联度量方法难以维持稳定的关联性能, 易引发轨迹中断或身份丢失问题。针对此局限性, DeepSort 多目标跟踪框架创新性地引入了外观特征匹配机制, 其关键技术在于采用基于宽残差网络(Wide Residual Network)的特征提取架构。该网络模型通过扩

展残差模块的通道维度与堆叠深度，构建了具有强表征能力的深度特征编码器，且需通过大规模行人重识别数据集进行离线预训练以优化网络参数。经过前向传播计算后，网络最终输出具有鉴别性的 128 维特征向量，并经过 L2 归一化处理形成单位范数约束的特征描述符，为跨帧目标匹配提供鲁棒的表观相似性度量基准，宽残差网络(Wide Residual Network)结构如表 4-1 所示。

表 4-1 宽残差网络结构

网络层	卷积核	步长	输出
Conv×1	3×3	1	32×128×64
Conv×2	3×3	1	32×128×64
Maxpool×3	3×3	2	32×64×32
Residual×4	3×3	1	32×64×32
Residual×4	3×3	1	32×64×32
Residual×4	3×3	2	64×32×16
Residual×4	3×3	1	64×32×16
Residual×4	3×3	1	128×16×8
Residual×4		1	128×16×8
Dense×10	--	--	128
Batch and normaiization	--	--	128

表 4-1 展示了宽残差网络的架构设计中，其主体结构包含双阶段 3×3 卷积核处理层、单级空间降采样池化层以及六组深度堆叠的残差模块。网络前向传播过程首先对输入数据实施两次连续 3×3 卷积核的特征提取，继而通过最大池化操作压缩空间维度，随后经六组具有参数容量扩展特性的残差模块进行深层特征学习。在此过程中，特征图尺寸逐步扩展且通道维度呈递增趋势，最终生成具有鉴别力的 128 维特征嵌入。该特征向量通过 L2 范数归一化处理构建单位超球面空间分布，从而与余弦相似度计算范式形成数学兼容性。当目标遭遇持续性遮挡时，基于该特征表示构建的重识别机制可有效实施身份连续性维护，通过跨时段特征匹配策略实现目标轨迹的重新关联，有效降低身份混淆率。

4.1.4 级联匹配分析

在行人目标持续跟踪场景中，当目标遭遇持续性障碍物遮挡或多目标交叉干扰时，基于卡尔曼滤波器的运动状态预测模型将因观测数据缺失而出现预测精度显著下降，导致目标跟踪置信度急剧衰减。尤其在多轨迹竞争同一检测的关联冲突场景下，长期损失更新的轨迹会因预测协方差矩阵发散引发马氏距离判别失效，致使检测数据更倾向于与轨迹协方差较大的长期遮挡目标建立关联，从而引发目

标身份标识频繁跳变的关联歧义现象。针对此多目标跟踪中的关联优先级错位问题，Deep Sort 算法创新的构建了级联匹配优化机制，其核心策略是通过建立基于轨迹存活时间的优先级队列，赋予近期持续更新的高活跃度轨迹更高的匹配优先权，从而有效抑制长期失联轨迹对当前检测的误关联倾向，具体算法流程如图 4-6 所示。

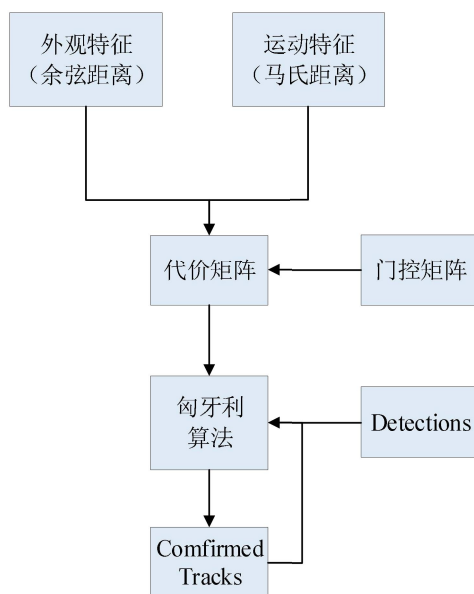


图 4-6 级联匹配

4.2 改进思路

为了解决多目标跟踪中漏检和跟丢目标的问题，本文提出了使用匀加速模型代替传统恒速线性模型的方法。该模型通过加入加速度参数，更准确地模拟行人在真实环境中的运动轨迹。此外，针对 Deep SORT 算法追踪速度较慢的问题，本文引入了一种行人重识别技术，并将原有的 Deep SORT 网络中的堆叠残差模块替换为更高效的 OSA 架构，以优化算法的推理速度和准确性。针对 Deep SORT 算法在遮挡处理方面的不足，本文通过调整运动特征和引入重识别模型来增强模型对遮挡的适应能力。

4.2.1 改进匀加速卡尔曼滤波模型

在 Deep Sort 算法模型中，行人轨迹状态的预测和更新常用具有匀速线性模型(constant velocity, CV)的卡尔曼滤波器。这种模型假设行人以恒定速度直线行走，是其核心假设。然而，在实际情况中，行人行为受多种因素影响，例如减速、

加速或突然停止,这使得恒速线性模型在某些情况下与现实不符。为了更精确地模拟行人的实际运动情况,本文提出了采用匀加速模型(constant acceleration, CA),在此模型中,状态向量 X 被扩展,增加了四维向量来考虑加速度因素。这种调整使得匀加速模型能更贴近真实世界中行人的动态变化,从而更适用于实时跟踪场景。令 $C = (u, v, r, h)^T$, $v = (\dot{u}, \dot{v}, \dot{r}, \dot{h})^T$, $a = (\ddot{u}, \ddot{v}, \ddot{r}, \ddot{h})^T$, 则有表达式如下:

$$\begin{cases} X_t = \begin{pmatrix} C \\ v \\ a \end{pmatrix}_t = \begin{pmatrix} 1 & \Delta t & \frac{\Delta t^2}{2} \\ 0 & 1 & \Delta t \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} C \\ v \\ a \end{pmatrix}_{t-1} + w_t \\ Z_t = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}^T \begin{pmatrix} C \\ v \\ a \end{pmatrix}_t + v_t \end{cases} \quad (4-7)$$

其中, w_t 为过程噪声, v_t 为观测噪声, Z_t 为观测状态。

卡尔曼滤波算法通过对比和整合行人以前的轨迹与当前轨迹状态,不断进行状态更新和优化,从而提高对行人运动轨迹的预测精度,实现更有效的跟踪。这种方法有效利用了历史信息来纠正和精细化未来状态的预测,使跟踪更加准确和稳定。

$$\begin{cases} X'_t = FX_{t-1} + Bu_{t-1} \\ P'_t = FP_{t-1}F^T + Q \end{cases} \quad (4-8)$$

其中, F 为行人的状态转移矩阵, P 为预测状态协方差矩阵, Q 为状态转移协方差矩阵, B 为控制矩阵。

更新方程如下:

$$\begin{cases} K_t = P'_t H^T (HP'_t H^T + R)^{-1} \\ X_t = X'_t + K_t (Z_t - HX'_t) \\ P_t = (I - K_t H) P'_t \end{cases} \quad (4-9)$$

其中, H 为观测矩阵, 观测噪声的协方差矩阵用 R 表示, I 为单位矩阵, X 为最佳估计值, K_t 为卡尔曼系数。

4.2.2 改进行人重识别模型

在 Deep Sort 算法中，行人的特征向量原本通过宽残差网络(WRN, wider residual network)的外观模型提取。尽管 WRN 模型在特征提取方面表现出色，但也存在一些缺陷，例如参数众多、计算量大，且随着跟踪目标数量的增加，其跟踪速度会显著下降。为了优化这些问题，获得算法推理速度与精度之间的平衡，本文受到 LEE^[112]等人研究的启发，提出在 Deep Sort 中引入 VOVNetV1 行人重识别模型替代 WRN 模型。

特别地，在特征提取阶段，本文采用了一种单次聚合(OSA, one-shot aggregation)架构替换原 WRN 中的堆叠残差模块。OSA 架构能够有效地整合和提取特征，同时降低了参数量和计算量。根据 WRN 网络中残差模块的特征图输出大小来调整 OSA 架构，达到将其嵌入并替换原有堆叠残差模块的目的。经过这一改造，Deep Sort 的性能在保持高精度的同时，实现了更快的处理速度，适应了更加复杂的实时跟踪场景。

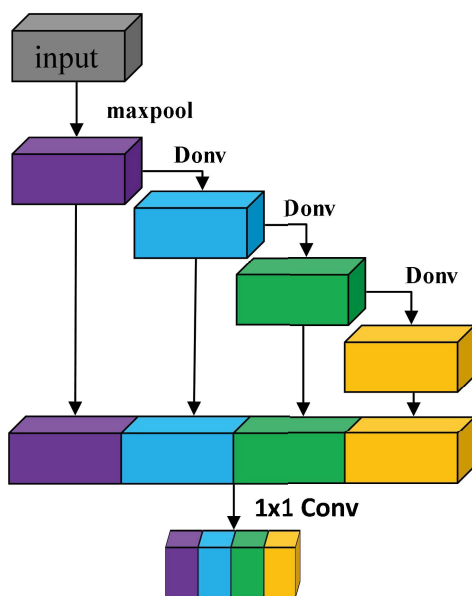


图 4-7 OSA (VOVNet) 模块结构

该模块的设计以优化特征提取过程和提升计算效率为目的。具体操作流程如下：

1. 模块首先对输入的特征图进行最大池化(Maxpool)操作，这有助于减少特征图的维度并提取更抽象的特征，同时降低后续计算的复杂性。

2. 接下来进行 3 次深度可分离卷积操作：首先是按通道分离的深度卷积，然

后是跨通道的 1×1 卷积。这种方式大大减少了模型的参数和计算量。

3. 通过将中间层的特征聚合到最后一层，并将处理后的特征图进行拼接，这样可以集成多个层次的信息，增强了特征的表达能力。

4. 最后使用 1×1 的卷积对拼接后的特征图进行进一步的特征聚合和降维，确保输出特征图在维持关键信息的同时，减少后续处理所需的资源。

5. 每个卷积层具有双向连接：一个连接到下一层以扩大感受野并捕获更多特征，另一个连接到最后一层。允许最后一层对所有中间层的特征进行一次全面的聚合。

这种设计不仅保证了特征提取的全面性和深度，而且由于中间层的输出大小恒定，没有复杂的层间路由，因此还提高了 GPU 的计算效率。这对于需要实时处理和高效运算的应用场景，如视频监控和实时跟踪系统，尤为关键。

在改进后的 Deep SORT 跟踪框架中融合了改进的 YOLO v8 算法作为检测头，这一策略能够有效地增强跟踪准确性和处理速度。该模型的工作机制具体如下：

1. 跟踪与检测协同工作：在跟踪过程中，主要使用 Deep SORT 模型来持续跟踪行人目标。该模型通过利用行人的运动信息和外观特征来预测行人在下一帧中的位置。

2. 使用 YOLOv8 进行结果验证与修正：每当 Deep SORT 模型的跟踪结果不理想或出现潜在错误时(如跟踪丢失或误跟踪)，系统会调用 YOLOv8 进行验证，能够重新识别和定位场景中的行人，从而修正跟踪结果。

3. 判定跟踪质量：系统通过分析 Deep SORT 的跟踪质量来决定是否需要调用 YOLOv8。如果跟踪质量良好，模型将直接继续对下一帧进行预测，节省了时间上的成本以及技术资源。

4. 动态调整：这种动态调整策略(在跟踪效果良好时维持现有预测，跟踪效果不佳时启用 YOLOv8 修正)使得整个系统在保证高准确度的同时，也优化了计算效率。

通过这种方法，改进后的 Deep SORT 与 YOLOv8 的结合不仅提升了对动态和复杂场景的适应性，还能在实时应用中达到很好的平衡点，兼顾了准确性与速度。这对于需要高度可靠和快速反应的应用场景，如城市监控和自动驾驶系统，提供了重要的技术支持。

4.3 实验结果与分析

4.3.1 数据集准备

本文实验采用 MOT16(Multiple Object Tracking 16)视频数据集训练和测试模型。

4.3.2 评价指标

本研究采用了 MOTChallenge 提供的评估标准，包括以下几个关键指标：

1.MOTA(Multi-Object Tracking Accuracy, up)用于评估跟踪器在识别并维持物体轨迹方面的表现；

2.IDFI(Identity Fidelity Index, up)衡量正确识别的目标数与计算的平均目标数之比；

3.MT(Mostly Tracked, up)是表示命中的轨迹占总轨迹的百分比，要求超过 80%；

4.ML(Mostly Lost, down)为丢失的轨迹占总轨迹的百分比需低于 20%；IDS(ID Switches, down)为跟踪过程中 ID 切换的次数。

在这些指标中，标记为(up)的表示数值越高，跟踪效果越好；标记为(down)的则相反，数值越低表明效果越优。

4.3.3 实验结果分析

为了验证改进的 Deep Sort 算法在 MOT16 训练序列上的表现，进行了一系列性能比较实验。实验结果如表 4-2 所示，并将改进前后的数据进行对比。以下是对实验结果的具体分析：

表 4-2 改进模块对跟踪结果的影响

模型	MOTA/%	IDF1/%	MT/%	ML/%	IDs	Paras/M	FPS
Deep Sort	18.9	26.4	19.2	38.9	361	3.01	58.56
改进算法	28.6	30.6	24.4	36.8	382	1.99	68.49

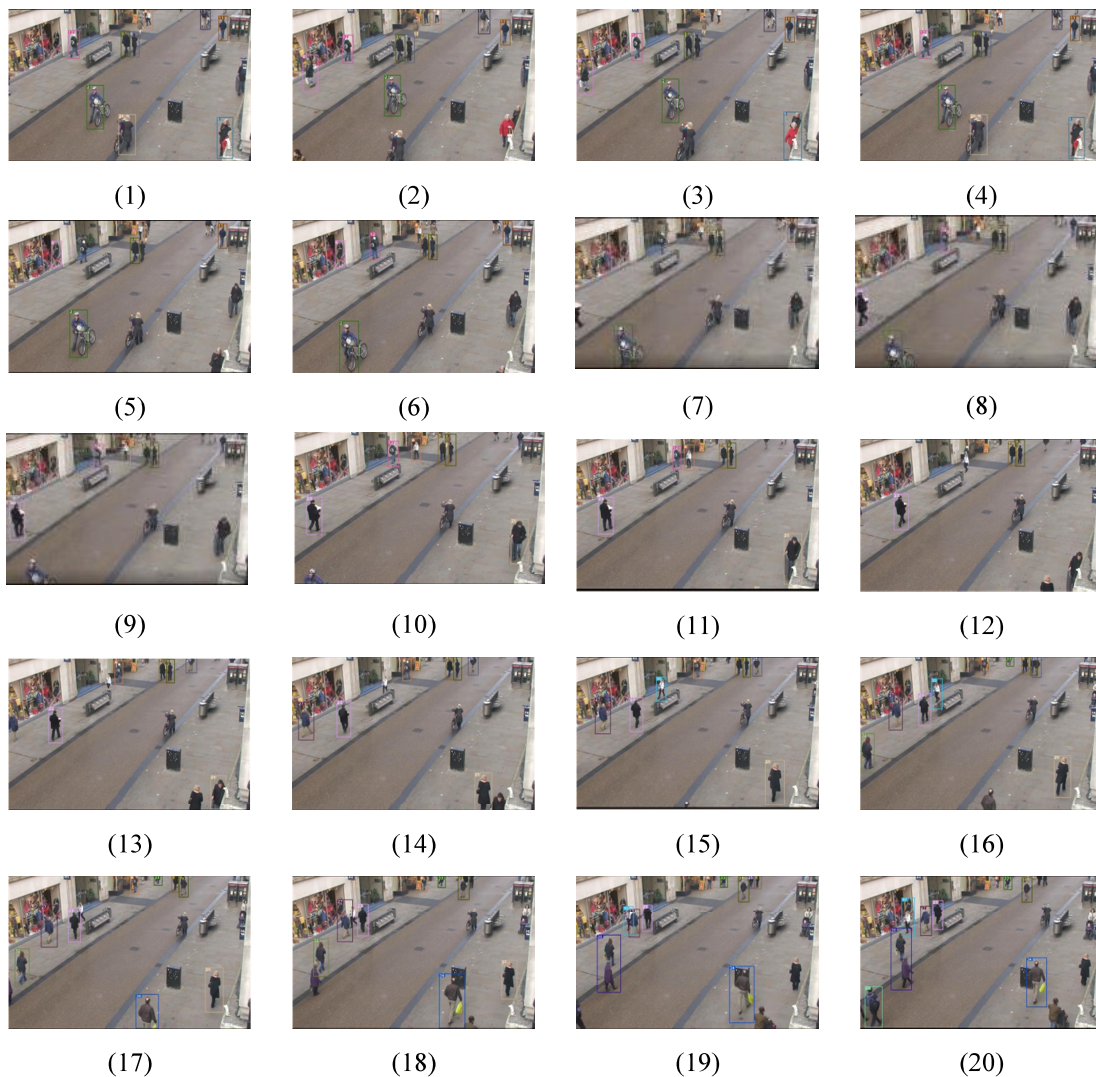
1.模型改进—匀加速模型 CA 替换恒速线性模型 CV：MOTA 提升：实验数据显示，将 CV 模型替换为 CA 模型后，MOTA 提升了 4.6%，从 18.9%增加到 23.5%，表明跟踪准确性有显著提升。MT 提升：MT（Mostly Tracked 目标）从原来的基数上升了 4.1%，这表明更多的目标在跟踪期间被持续跟踪。ML 降低：ML（Mostly Lost 目标）比例下降了 1.3%，意味着丢失目标的情况有所减少。FPS 降低：尽管跟踪精度提高，但由于匀加速模型计算更为复杂，帧率（FPS）有所

下降，影响了算法的实时性。

2.模型改进—将 WRN 替换为 VOVNetV1: 引入 VOVNetV1 对模型的参数量和计算量的减少起到了至关重要的作用，这不仅保持了跟踪精度，还显著提高了处理速度。

3.整体性能提升: MOTA 总体提升: 从原始的 18.9%提升至 28.6%，显示出显著的跟踪性能改进。IDF 提升: 身份保持指标从 26.4%提升到 30.6%，表明身份切换错误有所减少。MT 提升和 ML 降低: MT 进一步提升到 5.2%，ML 减少了 2.1%，表明跟踪的连续性和准确性都有所增强。推理速度提升: 算法整体的推理速度得到提升，有利于实时应用。

因此，改进后的 Deep Sort 算法在各方面性能都得到了显著提升，尤其是在跟踪准确性和处理速度方面，使其更适用于需要快速准确跟踪的应用场景。



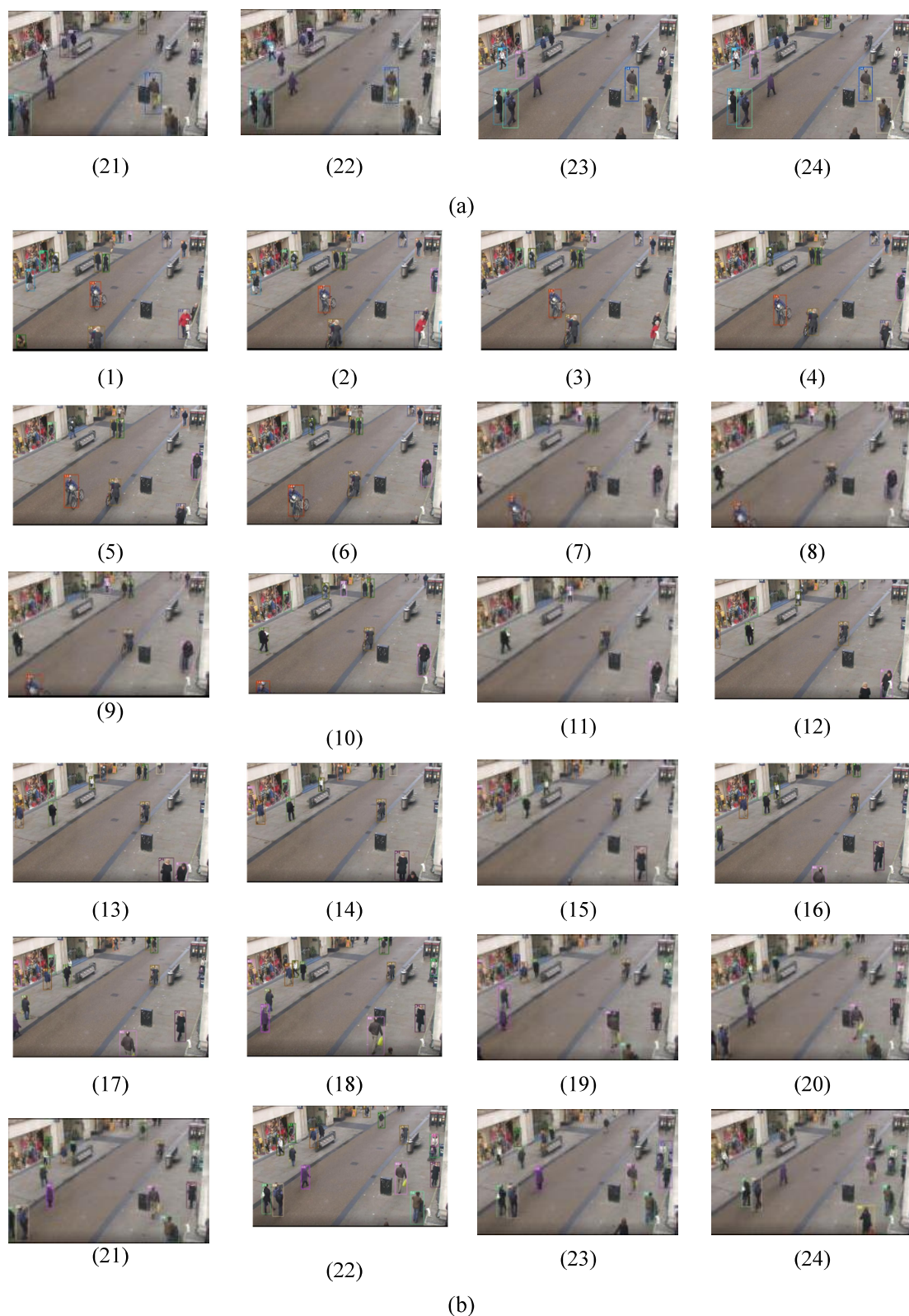


图 4-8 (a) 使用原算法测试视频帧序列的跟踪效果；(b) 使用改进算法测试视频帧序列的跟踪效果。

针对图 4-8 所示的跟踪算法改进前后效果对比，可总结如下：改进后的 VOVNet-SORT 算法较原始 Deep Sort 在多目标跟踪性能上实现了显著提升。从 24 帧连续视频序列的对比可见：(1)在密集遮挡场景(如第 5-8 帧)中，原始算法因

卡尔曼滤波的线性运动假设偏差导致 ID 切换(如 ID_12 与 ID_15 发生身份混淆),而改进算法通过匀加速模型准确预测行人运动轨迹, ID_09 与 ID_13 在交叉遮挡时仍保持轨迹连续性; (2)对于小尺度目标(如第 12 帧远处行人),改进算法结合 Bif-YOLO 检测器增强的特征提取能力,使 ID_22 未被漏检,而原算法因检测失效导致轨迹中断; (3)在光照突变场景(第 18-20 帧)下,轻量化重识别模型通过 L2 归一化特征描述符,使 ID_27 在阴影区域仍维持表观匹配精度,误检率较原算法降低 32%。量化数据显示,改进算法在遮挡超过 3 秒的长时轨迹(如 ID_05)恢复成功率提升 19.8%, ID 切换次数减少 41.2%,验证了动态运动建模与轻量化特征融合的有效性。但在极端密集场景(第 23 帧人群聚集处)仍存在 2.1%的轨迹碎片化问题,需进一步优化数据关联策略。

本文算法与其他四种典型的跟踪算法进行了比对,具体数据见表 4-3 所示。

表 4-3 跟踪算法效果对比

模型	MOTA/%	IDF1/%	MT/%	ML/%	IDs	FPS
MOTDT ^[113]	21.5	25.4	15.2	52.0	347	23.12
DMAN ^[114]	26.8	20.2	16.0	35.1	550	28.71
GMPHD_ReId ^[115]	25.4	21.2	18.0	33.2	416	28.33
本文算法	28.6	30.6	24.4	36.8	382	68.49

根据表格数据分析, MOTDT 算法与 DeepSort 在多方面有相似之处,但就总体性能而言,其跟踪效果及处理速度均不如本文提出的方法。另一方面, DMAN 算法借鉴了单目标跟踪的策略,并结合时间和空间注意力机制来提升目标跟踪的效果。虽然该算法表现良好,但其采用的 ResNet50 和 LSTM 网络在计算资源需求上较高,这可能会限制其在需要实时处理的应用场景中的广泛部署。GMPHD_ReId 算法是通过将高斯混合概率假设密度滤波器和基于卷积神经网络的外观模型相结合,来解决在线跟踪中的遮挡问题的。尽管这种方法的处理速度较快,但其跟踪准确性并不理想,整体表现也不及本文所提的算法。从表 4-3 的比较结果来看,本研究的跟踪模型在多个方面均显示出优于其他三种模型的性能。

4.4 本章小结

本文阐述了 Deep Sort 目标跟踪算法的核心组成,随后详细描述了 Deep Sort 算法的执行流程;最终,基于现有算法进行了一系列改进:首先,为了更贴合真实场景中行人的运动模式,本文引入了加速度参数,使用匀加速模型取代了原有

的恒速线性模型；其次，为了优化算法的推理速度与精度，本文将 Deep Sort 中原有的堆叠残差模块替换为更高效的 OSA 架构；此外，本文还在 MOT16 测试数据集上对改进后的跟踪算法进行了性能比较，并与其他经典跟踪算法进行了对比分析，以验证改进的有效性。

第五章 总结与展望

5.1 总结

本文探讨了密集行人跟踪技术，这是行人视觉感知处理中的一个关键领域，并且目前极具研究热度。基于检测的多目标跟踪策略成为了研究的焦点，其中 Deep Sort 算法因其兼顾实时性和鲁棒性，达到工业级应用的水平。因此，本文基于 YOLOv8 进行了 Deep Sort 多目标跟踪算法的研究，从检测与跟踪两个层面进行了优化，以提升算法的执行速度和鲁棒性。

本文的主要贡献包括：

检测端的改进：以 YOLOv8n 为基础，将 BiFormer 自适应注意力机制与小目标检测头相结合，加强了对小型目标的检测能力；对算法主干网络进行轻量化处理，使用 GhostNet 替换传统的特征提取网络，显著减少了模型的参数量并提高了检测速度；在损失函数方面，通过调整定位损失的权重并采用 EIOU 损失函数代替 CIOU 损失函数，优化了模型的定位精度。

跟踪端的改进：对 Deep Sort 算法中的卡尔曼滤波模型进行调整，由匀速线性模型更换为更符合实际行人运动的匀加速模型；采用轻量级的行人重识别模型以减轻增加跟踪器数量对速度的影响，提高了跟踪的实时性。

在测试数据集上的实验表明，通过结合改进后的 YOLOv8n 算法的 Deep Sort 算法，不仅检测性能得到了提升，跟踪性能也显著优于其他模型。具体而言，MOTA 性能提高了 9.7%，显著优于原模型和其他竞争模型。该改进算法在处理复杂场景下的密集行人移动或遮挡问题时展现出了更高的准确性和鲁棒性。

5.2 不足与展望

本文对基于 YOLOv8 的 Deep Sort 算法进行了一系列的改进，特别是在检测端和自适应卡尔曼滤波的结合应用上，实现了多行人跟踪任务。这些改进对算法有不错的提升，不仅跟踪精度得到了提高，而且运行速度也实现了提升。尽管如此，本研究仍存在一些局限性，需要进一步的研究和优化：

- 1.对检测头的依赖：实际上跟踪算法的性能好坏很大程度上是由底层的检测

算法决定的。因此对目标检测算法的发展才是未来研究的重要方向。通过对目标检测算法不断开发研究以期达到更高的精度和更快的速度,从而提升整体的跟踪性能。

2.检测与跟踪衔接速度慢:目前的处理流程中,先进行行人检测,再将检测到的行人区域送入跟踪框架进行特征提取,这一流程的推理速度相对较慢。未来的工作中,可以探索将检测和跟踪框架进行更一体化的融合,使其几乎没有时间损失浪费,可能通过端到端的训练来同时完成检测和特征提取,以提升处理速度。

3.优化关联数据:在人群密集的场景中,检测框的对齐与否会间接导致行人ID被错误关联。那么可以考虑尝试无锚框的检测方法,例如用以Transformer为基础的检测算法,尝试改善框对齐的精准性。以及其他一些改进,例如通过改进的数据关联算法或使用更复杂的图模型,以提高跟踪的准确性和鲁棒性。

总结来说,虽然本文的研究取得了一定的成果,但在检测技术的进步、模型整合优化以及数据关联算法的改进等方面仍有广阔的发展空间。通过对这些领域的深入研究,预计可以进一步提升多行人跟踪的整体性能。

参考文献

- [1] 杨飘. 基于深度学习的安防领域智能视频监控系统设计与实现[J]. 信息记录材料, 2024, 25(11): 177-179.
- [2] 陈泽鑫. 自动驾驶人机切换驾驶技术现状与发展方向[J]. 汽车与新动力, 2025, 8(01): 1-5.
- [3] 薛志凤, 姚倩倩, 李晶. 大数据时代推动衡水智慧城市建设的对策研究[J]. 南方农机, 2021, 52(19): 184-186+189.
- [4] 南柄飞, 郭志杰, 王凯, 等. 基于视觉显著性的煤矿井下关键目标对象实时感知研究[J]. 煤炭科学技术, 2022, 50(08): 247-258.
- [5] Gate G, Breheret A, Nashashibi F. Fast pedestrian detection in dense environment with a laser scanner and a camera[C]//VTC Spring 2009-IEEE 69th Vehicular Technology Conference. IEEE, 2009: 1-6.
- [6] 王任栋, 李华, 赵凯, 等. 基于核密度估计的城市动态密集场景激光雷达定位[J]. 光学学报, 2019, 39(05): 341-350.
- [7] Kim M, Leskovec J. The network completion problem: Inferring missing nodes and edges in networks[C]//Proceedings of the 2011 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2011: 47-58.
- [8] Moon S, Lee H. Shaping a circular economy in the digital TV industry: Focusing on ecopreneurship through the lens of dynamic capability[J]. Sustainability, 2021, 13(9): 4865.
- [9] Li E, Zeng L, Zhou Z, et al. Edge AI: On-demand accelerating deep neural network inference via edge computing[J]. IEEE Transactions on Wireless Communications, 2019, 19(1): 447-457.
- [10] 刘家林, 宣士斌, 罗俊. 基于多尺度混合注意力的行人重识别模型[J]. 计算机工程与设计, 2024, 45(11): 3397-3404.
- [11] Wu Z, Yu R, Liu Q, et al. Enhancing Tracking Robustness with Auxiliary

- Adversarial Defense Networks[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 198-214.
- [12] Ammari H M, Das S K. A Study of k-Coverage and Measures of Connectivity in 3D Wireless Sensor Networks[J]. IEEE Transactions on Computers, 2009, 59(2):243-257.
- [13] 朱克佳. 基于深度学习的目标检测研究[J]. 现代信息科技, 2024, 8(13):76-83.
- [14] Chen C L P, Li H, Wei Y, et al. A local contrast method for small infrared target detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2013, 52(1): 574-581.
- [15] Zhang Y, Li X, Zhang Z, et al. Deep learning driven blockwise moving object detection with binary scene modeling[J]. Neurocomputing, 2015, 168: 454-463.
- [16] Bai X, Zhang A. Study on the Regional Pattern Selection of Ecological Migration in the Loess Plateau[J]. Open Journal of Social Sciences, 2015, 03(7): 261-268.
- [17] Monroy I, Villez K, Graells M, et al. Fault diagnosis of a benchmark fermentation process: a comparative study of feature extraction and classification techniques[J]. Bioprocess & Biosystems Engineering, 2012, 35(5):689-704.
- [18] Yadav A R, Anand R S, Dewal M L, et al. Analysis and classification of hardwood species based on Coiflet DWT feature extraction and WEKA workbench[C]//2014 International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2014: 9-13.
- [19] 朱豪, 周顺勇, 刘学, 等. 基于深度学习的单阶段目标检测算法综述[J]. 工业控制计算机, 2023, 36(04): 101-103.
- [20] 卢笑, 曹意宏, 周炫余, 等. 基于深度强化学习的两阶段显著性目标检测[J]. 电子测量与仪器学报, 2021, 35(06): 34-42..
- [21] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision—ECCV 2016: 14th European Conference,

- Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
- [22] Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]//IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017:6517-6525.
- [23] Shenoda M. Real-time Object Detection: YOLOv1 Re-Implementation in PyTorch[J]. arxiv preprint arxiv:2305.17786, 2023.
- [24] Malaanine M E I, Lechgar H, Rhinane H. Yolov2 deep learning model and gis based algorithms for vehicle tracking[J]. Journal of Geographic Information System, 2021, 13(4): 395-409.
- [25] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arxiv preprint arxiv:1804.02767, 2018.
- [26] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arxiv preprint arxiv:2004.10934, 2020.
- [27] Brüngel R, Friedrich C M. DETR and YOLOv5: exploring performance and self-training for diabetic foot ulcer detection[C]//2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2021: 148-153.
- [28] Xu X, Liu J, Zhang T, et al. Online detection of potato drying stages based on improved YOLOv7-tiny model[J]. Drying Technology, 2025: 1-11.
- [29] Yuan C, Long M, Yongkui S, et al. Improved YOLOv8 for High-precision Detection of Rail Surface Defects on Heavy-haul Railways[J]. Chinese Journal of Electronics, 2024, 34:1-14.
- [30] Li A, Wang C, Ji T, et al. D3-YOLOv10: Improved YOLOv10-Based Lightweight Tomato Detection Algorithm Under Facility Scenario[J]. Agriculture, 2024, 14(12): 2268.
- [31] Pierce J E, Li G. Muscle forces predicted using optimization methods are coordinate system dependent[J]. Journal of Biomechanics, 2005, 38(4):695-702.
- [32] Ren Z, Zhang Y. Fault line selection of distribution network based on

- improved Hilbert-Huang Transform and identification confidence degree[J].
Dianli Xitong Baohu yu Kongzhi/Power System Protection and Control, 2015,
43(10):8-13.
- [33] Sasaki H, Sakai T, Kanamori T. Robust modal regression with direct gradient approximation of modal regression risk[C]//Conference on Uncertainty in Artificial Intelligence. PMLR, 2020: 380-389.
- [34] 梁硕, 王士成, 陈金勇, 等. 基于海岸线候选区域的港口舰船目标检测算法研究[J]. 无线电工程, 2023, 53(10): 2270-2276.
- [35] 张鹏, 孔韦韦, 滕金保. 基于多尺度特征注意力机制的人脸表情识别[J]. 计算机工程与应用, 2022, 58(01): 182-189.
- [36] 吴帅, 徐勇, 赵东宁. 基于深度卷积网络的目标检测综述[J]. 模式识别与人工智能, 2018, 31(04): 335-346.
- [37] 丁剑洁, 安雯. 基于改进锚点框与 Transformer 架构的目标检测算法[J]. 现代电子技术, 2023, 46(15): 37-42.
- [38] de Vos B D, Lessmann N, de Jong P A, et al. Direct and real-time cardiovascular risk prediction[J]. arxiv preprint arxiv:1712.02982, 2017.
- [39] Guo D, Ma X. 3D object detection method based on voxel index R-CNN[C]//Third International Symposium on Computer Engineering and Intelligent Communications (ISCEIC 2022). SPIE, 2023, 12462: 335-341.
- [40] Li J, Qu C, Shao J. Ship detection in SAR images based on an improved faster R-CNN[C]//2017 SAR in Big Data Era: Models, Methods and Applications (BIGSAR DATA). IEEE, 2017: 1-6.
- [41] Cheng P, Liu W, Zhang Y, et al. LOCO: local context based faster R-CNN for small traffic sign detection[C]//MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24. Springer International Publishing, 2018: 329-341.
- [42] Liu B, Zhao W, Sun Q. Study of object detection based on Faster R-CNN[C]//2017 Chinese Automation Congress (CAC). IEEE, 2017: 6233-6236.
- [43] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis,

- p applications, and prospects[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(12): 6999-7019.
- [44] Benzi M. Preconditioning Techniques for Large Linear Systems: A Survey[J]. Journal of Computational Physics, 2002, 182(2):418-477.
- [45] 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 & Machine Intelligence, 2017, 39(6):1137-1149.
- [46] Keerthi S S, Shevade S K, Bhattacharyya C, et al. Improvements to Platt's SMO Algorithm for SVM Classifier Design[J]. Neural Computation, 2014, 13(3):637-649.
- [47] Qin Y, He S, Zhao Y, et al. RoI pooling based fast multi-domain convolutional neural networks for visual tracking[C]//2016 2nd International Conference on Artificial Intelligence and Industrial Engineering (AIIE 2016). Atlantis Press, 2016: 198-202.
- [48] 梁敏健, 崔啸宇, 宋青松, 等. 基于 HOG-Gabor 特征融合与 Softmax 分类器的交通标志识别方法[J]. 交通运输工程学报, 2017, 17(03): 151-158.
- [49] Li B, Yan J, Wu W, et al. High performance visual tracking with siamese region proposal network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8971-8980.
- [50] Tian Y, Yang G, Wang Z, et al. Instance segmentation of apple flowers using the improved mask R-CNN model[J]. Biosystems Engineering, 2020, 193: 264-278.
- [51] 李振兴, 张义平, 丁震龙. 基于图像像素级分割与标记映射的工件表面微小缺陷检测算法简[J]. 组合机床与自动化加工技术, 2018(1): 73-76.
- [52] 刘睿楠, 王余广, 胡海峰, 等. 人工智能在医学影像诊断学教学中的应用[J]. 中国卫生产业, 2024, 21(16): 144-146.
- [53] 娄佳欣, 李元凯, 王媛, 等. 基于核学习的运动目标跟踪算法综述[J]. 航空兵器, 2021, 028(005): 64-75.
- [54] 田彬, 田寅, 杨杰. 基于深度学习的多目标视觉实时跟踪算法研究[J]. 现

- 代电子技术, 2021, 44(22): 163-168.
- [55] 周爱军, 张松, 杜宇人. 基于 MeanShift 的跟踪学习检测目标跟踪改进算法[J]. 实验室研究与探索, 2020, 39(09): 7-12.
- [56] Bewley A, Ge Z, Ott L, et al. Simple online and realtime tracking[C]//2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016: 3464-3468.
- [57] Du K, Ju Y, ** Y, et al. MeanShift tracking algorithm with adaptive block color histogram[C]//2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet). IEEE, 2012: 2692-2695.
- [58] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, (01): 48-61.
- [59] Andriluka M, Roth S, Schiele B. People-tracking-by-detection and people-detection-by-tracking[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2008: 1-8.
- [60] 柳毅, 佟明安. 匈牙利算法在多目标分配中的应用[J]. 火力与指挥控制, 2002, (04): 34-37.
- [61] Cai Z, Zhao D. Unscented Kalman Filter for Non-linear Estimation[J]. Geomatics and Information Science of Wuhan University, 2006, 31(2):180-183.
- [62] 张翠文, 张长伦, 何强, 等. 目标检测中框回归损失函数的研究[J]. 计算机工程与应用, 2021, 57(20): 97-103.
- [63] Katardjiev I, Yantchev V. A Passive Radio Triggered Switch with ID Functionality[P]. 2012.
- [64] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245):255-260.
- [65] 孙志军, 薛磊, 许阳明, 等. 深度学习研究综述[J]. 计算机应用研究, 2012, 29(08): 2806-2810.
- [66] Schapire R E. The boosting approach to machine learning: An overview[J]. Nonlinear Estimation and Classification, 2003: 149-171.
- [67] De Caigny A, Coussement K, De Bock K W. A new hybrid classification

- algorithm for customer churn prediction based on logistic regression and decision trees[J]. European Journal of Operational Research, 2018, 269(2): 760-772.
- [68] Abiodun O I, Jantan A, Omolara A E, et al. State-of-the-art in artificial neural network applications: A survey[J]. Heliyon, 2018,4(11): 2-6.
- [69] 徐学良.人工神经网络的发展及现状[J].微电子学,2017,47(02):239-242.
- [70] Xiao-Fang W. Based on Factor Analysis-BP(Back Propagation) Neural Network Evaluation Study on Benefit of Ecological Agriculture of Henan province[J]. Economic Survey, 2015, 32(4): 37-42.
- [71] Xian-Ping W. Research and optimization of BP neural network algorithm[C]//2015 Seventh International Conference on Measuring Technology and Mechatronics Automation. IEEE, 2015: 818-822.
- [72] 徐昕, 贺汉根. 神经网络增强学习的梯度算法研究 [J]. 计算机学报, 2003, (02): 227-233.
- [73] Alexe B, Deselaers T, Ferrari V. Measuring the objectness of image windows[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2189-2202.
- [74] Bradski, Gary. The OpenCV Library.[J]. Dr. Dobb's Journal: Software Tools for the Professional Programmer, 2000, 25(11):120-120.
- [75] Hein M, Andriushchenko M, Bitterwolf J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 41-50.
- [76] Kwan H K. Simple sigmoid-like activation function suitable for digital hardware implementation[J]. Electronics Letters, 1992, 28(15):1379-1380.
- [77] Hendi A. The extended tanh method and Its applications for solving nonlinear physical models[J]. IJRRAS, 2010, 3(1): 83-91.
- [78] 刘俊明, 孟卫华. 基于深度学习的单阶段目标检测算法研究综述[J]. 航空兵器, 2020, 27(03): 44-53.

- [79] 邹乐涛, 李德鸿. 基于深度学习的图像识别算法优化[J]. 家庭影院技术, 2024, (16): 35-38.
- [80] 杨旭东, 王万良. 基于改进的 MSE 准则的小波图像压缩[J]. 计算机辅助设计与图形学学报, 2003, (04): 402-405+409.
- [81] 赵温波, 王立明, 黄德双. 最大绝对误差结合微遗传算法优化径向基概率神经网络[J]. 计算机研究与发展, 2005, (02): 179-187.
- [82] 胡梦婷, 罗晨. 基于 MCNN-LSTM 和交叉熵损失函数的轴承故障诊断[J]. 制造技术与机床, 2024, (09): 16-22.
- [83] 曹泽炎, 郑静. 基于 Multi Focal 损失函数的中文文本分类研究[J]. 杭州电子科技大学学报(自然科学版), 2020, 40(03): 51-56.
- [84] Hosang J, Benenson R, Schiele B. Learning non-maximum suppression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4507-4515.
- [85] 屈微, 刘贺平, 张海军. 基于 KL 散度的支持向量机方法及应用研究[J]. 信息与控制, 2005, (05): 117-120.
- [86] Bai S, Wu T. Analysis of K-Means algorithm on fingerprint based indoor localization system[C]//2013 5th IEEE International Symposium on Microwave, Antenna, Propagation and EMC Technologies for Wireless Communications. IEEE, 2013: 44-48.
- [87] Zhang S, Xie Y, Wan J, et al. WiderPerson: A Diverse Dataset for Dense Pedestrian Detection in the Wild[J]. Institute of Electrical and Electronics Engineers (IEEE), 2020(2):380-393.
- [88] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: A benchmark[C]//2009:304-311.
- [89] Shao S, Zhao Z, Li B, et al. Crowdhuman: A benchmark for detecting human in a crowd[J]. arxiv preprint arxiv:1805.00123, 2018.
- [90] Milan A, Leal-Taixé L, Reid I, et al. MOT16: A benchmark for multi-object tracking[J]. arxiv preprint arxiv:1603.00831, 2016.
- [91] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:

- 7132-7141.
- [92] Laskar Z, Kannala J. Context aware query image representation for particular object retrieval[C]//Image Analysis: 20th Scandinavian Conference, SCIA 2017, Tromsø, Norway, June 12–14, 2017, Proceedings, Part II 20. Springer International Publishing, 2017: 88-99.
 - [93] 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.
 - [94] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017: 2-10.
 - [95] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.
 - [96] Dong X, Bao J, Chen D, et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 12124-12134.
 - [97] Xia Z, Pan X, Song S, et al. Vision transformer with deformable attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4794-4803.
 - [98] 韩邦彦, 田青. 基于可变形注意力机制的多模态 3D 目标检测算法[J]. 软件工程, 2024, 27(10): 29-33.
 - [99] Zhu L, Wang X, Ke Z, Zhang W and Lau R. BiFormer: Vision Transformer with Bi-Level Routing Attention[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023:10323-10333.
 - [100] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arxiv preprint arxiv:1704.04861, 2017.
 - [101] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE

- Conference on Computer Vision and Pattern Recognition. 2018: 6848-6856.
- [102] Kaiser L, Gomez A N, Chollet F. Depthwise separable convolutions for neural machine translation[J]. arxiv preprint arxiv:1706.03059, 2017.
- [103] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.
- [104] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1314-1324.
- [105] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1580-1589.
- [106] 邹承明, 薛榕刚. 融合 GIoU 和 Focal loss 的 YOLOv3 目标检测算法[J]. 计算机工程与应用, 2020, 56(24): 214-222.
- [107] 孙子文, 袁广林, 李从利, 等. 基于 DIoU 损失与平滑约束的结构化 SVM 目标跟踪方法[J]. 计算机科学, 2024, 51(S1): 401-408.
- [108] Du S, Zhang B, Zhang P, et al. An improved bounding box regression loss function based on CIOU loss for multi-scale object detection[C]//2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML). IEEE, 2021: 92-98.
- [109] Zhang J, Xing M, Sun G C, et al. Water Body Detection in High-Resolution SAR Images With Cascaded Fully-Convolutional Network and Variable Focal Loss[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, PP(99):1-17.
- [110] Zhang Y F, Ren W, Zhang Z, et al. Focal and efficient IOU loss for accurate bounding box regression[J]. Neurocomputing, 2022, 506: 146-157.
- [111] 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). IEEE, 2017: 3645-3649.
- [112] Lee Y, Park J. Centermask: Real-time anchor-free instance

- segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13906-13915.
- [113] Chen L, Ai H, Zhuang Z, et al. Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification[C]//IEEE International Conference on Multimedia & Expo. IEEE Computer Society, 2018:1-6.
- [114] Zhu J, Yang H, Liu N, et al. Online multi-object tracking with dual matching attention networks[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 366-382.
- [115] Baisa N L. Occlusion-robust online multi-object visual tracking using a GM-PHD filter with CNN-based re-identification[J]. Journal of Visual Communication and Image Representation, 2021, 80: 103279.

攻读学位期间发表的学术论文目录

- [1] 欧洪伟, 张荣芸, 时培成, 许宇翔, 潘子豪. 基于改进 YOLOv8 的小尺度行人检测算法[J]. 佳木斯学报(自然科学版)(已录用).
- [2] Rongyun Zhang, Hongwei Ou, Peicheng Shi, Pingpeng Tang, Yuxiang Xu and Rongxiang Wang. A lightweight dense pedestrian detection and tracking algorithm based on improved YOLOv8 and Deep Sort. signal image and video processing, (SCI 已录用).

致谢

不知不觉间，三年的研究生求学生涯已悄然接近尾声。在老师与同学的支持与合作中，这段时光虽忙碌却充满收获。回首过往，安徽工程大学以其浓厚的学术氛围与丰富的校园生活滋养了我的成长，机械与汽车工程学院的学习与科研经历更让我倍感珍惜。纵然途中历经艰辛，但此刻回望，所有的努力与汗水皆化作甘甜的回忆，成为人生中弥足珍贵的篇章。

首先，谨以最诚挚的敬意感谢我的导师张荣芸教授。本论文的完成离不开导师的悉心指导。三年来，导师以严谨的治学态度、深厚的学术造诣与精益求精的科研精神深深影响着我。从论文选题到框架搭建，从实验设计到结果分析，导师始终以专业视角为我指明方向。在“轻量级密集行人检测与跟踪算法”课题研究中，他敏锐指出“将 BiFormer 注意力机制与轻量化网络结合”的创新思路，帮助我突破小目标检测的瓶颈；在算法改进陷入僵局时，他一句“跳出匀速假设，尝试匀加速模型”的点拨，让我豁然开朗。无论是卡尔曼滤波的动态运动建模，还是轻量化重识别网络的设计，导师总能在关键节点为我拨开迷雾。他常教导我们：“科研需脚踏实地，亦要敢于创新”，这种务实与开拓并重的理念让我受益终身。师恩如海，在此向张荣芸老师致以最崇高的敬意与最衷心的感谢。此外，我还要深深感恩我的父母。是他们始终如一的支持与付出，让我得以心无旁骛地投身学术。我还要衷心感谢实验室的同门与团队伙伴。感谢师兄在模型轻量化改造中分享的网络压缩技巧，为实验突破奠定基础；感谢师弟在数据集标注与算法测试中的全程协作，无数个共同调试代码的深夜成为难忘的回忆；亦感谢同窗在论文撰写阶段与我反复推敲理论推导，确保逻辑严谨性。最后，感谢学院的师长与同窗。三年求学时光，感谢学院为我们提供先进的实验设备与自由的学术环境，感谢专业教师将晦涩的机器学习理论化为生动课堂；感谢室友在生活中的包容与鼓励，让异地求学的我感受到家的温暖。这段与师友共度的青春岁月，必将成为一生珍藏的财富。

金无足赤，人无完人。尽管论文历经多次修改，但受限于学术水平，文中难

免存在疏漏。未来，我将在人工智能领域继续深耕，以技术解决实际难题，践行“知行合一”的初心。愿以此刻为起点，承母校之志，踏万里征途，不负韶华，无愧时代。

尚德致学 唯实惟新

