分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2210330130

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　基于深度学习的多目标跟踪算法研究

论文作者　刘聪
校内导师　韩超（教授）
校外导师　郝旭耀

专业学位类别　电子信息
研究方向（领域）　控制工程与自动化技术

论文提交日期：2024 年 5 月 31 日

分类号：TP391　　单位代码：10363
密级：公开　　学　号：2210330130

题　　目　基于深度学习的多目标跟踪算法研究

英文并列

题　　目　Research on Multi-object Tracking Algorithm Based on Deep Learning

学　生　姓　名：刘聪
校　内　导　师：韩超（教授）
校　外　导　师：郝旭耀
专 业 学 位 类 别：电子信息
研究方向（领域）：控制工程与自动化技术

论文答辩日期：2024 年 5 月 23 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：2024 年 5 月 26 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：刘毅　　指导教师签名：韩迅

日期：2024 年 5 月 26 日　　日期：2024 年 5 月 26 日

基于深度学习的多目标跟踪算法研究

摘 要

多目标跟踪是计算机视觉领域中的关键研究方向，旨在对输入视频序列中的多个目标进行定位，同时给予每个跟踪目标唯一的身份标签，并在跟踪过程中保持每个身份标签的稳定性。它在视频监控、人机交互、虚拟现实、自动驾驶等方面有极为广泛的应用。近年来，随着深度学习的发展，基于深度学习的联合式（一步式）多目标跟踪算法因其跟踪效果良好而备受关注。然而，在遮挡等复杂环境下，现有的联合式多目标跟踪器仍然存在跟踪效果差的问题。因此，本文采用联合式多目标跟踪算法框架进行研究，从特征提取、语义信息融合、目标检测等方面提升联合式多目标跟踪算法的性能。本文主要研究工作如下：

（1）针对多目标跟踪过程中目标遮挡的问题，本文提出了一种融合注意力机制的行人多目标跟踪算法。首先在主干网络中采用特征金字塔注意力模块扩大感受野，获得更加丰富的特征信息，以提高对不同尺度目标的检测精度；其次在行人重识别分支网络中融合了一个尺寸感知模块，对来自不同分辨率的语义特征进行融合，提取更加基础的行人特征，从而提高跟踪的精度；最后重新构造检测头，融合小目标检测层，使得所提网络适应不同尺寸的目标。

（2）针对多目标跟踪过程中目标身份标签频繁切换的问题，本文在研究内容（1）的基础上提出了一种基于无锚框检测的行人多目

标跟踪算法。该算法首先在主干网络部分融合深层聚合注意力模块以获得多层次的目标特征，从而扩大感受野，获取全局特征信息；其次采用基于无锚框的 RepPoints 检测头，降低多目标跟踪过程中目标身份标签的切换次数；最后在数据关联部分采用扩展 IOU 匹配，增加检测和轨迹的匹配范围，补偿匹配时的运动估计偏差，从而提高复杂环境下跟踪的鲁棒性。

（3）在 MOT16、MOT17 数据集上检验本文所改进的算法，并与其他先进算法进行比较，实验结果表明，本文所提算法的跟踪性能得到提升。在 MOT16 数据集上，研究内容（1）和研究内容（2）所提算法的多目标跟踪准确率分别达到 75.4%，75.9%。在 MOT17 数据集上，研究内容（1）和研究内容（2）所提算法的多目标跟踪准确率分别达到 74.3%，74.7%。

关键词：多目标跟踪，深度学习，注意力机制，特征信息，行人重识别

Research on Multi-object Tracking Algorithm Based on Deep Learning

ABSTRACT

Multi-object tracking is a key research direction in the field of computer vision. It aims to locate multiple objects in the input video sequence, give each tracking object a unique identification, and maintain the stability of each identification during the tracking process. It has a very wide range of applications in video surveillance, human-computer interaction, virtual reality, autonomous driving and so on. In recent years, with the development of deep learning, one-shot multi-object tracking algorithms based on deep learning have attracted attention due to their good tracking performance. However, in complex environments such as occlusion, existing one-shot multi-object trackers still suffer from poor tracking performance. Therefore, this article adopts a one-shot multi-object tracking algorithm framework for research, aiming to improve the performance of the one-shot multi-object tracking algorithm from aspects such as feature extraction, semantic information fusion, and object detection. The main research work of this article is as follows:

(1) Addressing the issue of object occlusion in multi-object tracking, this article proposes a pedestrian multi-object tracking algorithm that integrates attention mechanism. Firstly, the feature pyramid attention module is used in the backbone network to enlarge the receptive field and obtain more abundant feature information to improve the detection accuracy of different scale objects. Secondly, a size-aware module is integrated into the pedestrian re-identification branch network to fuse

semantic features from different resolutions and extract more basic pedestrian features, thereby improving the tracking accuracy. Finally, the detection head is reconstructed and the small object detection layer is fused to make the proposed network adapt to objects of different sizes.

(2) Addressing the issue of frequent switching of object identification during multi-object tracking, On the basis of research content (1), this article proposes a pedestrian multi-object tracking algorithm based on anchor-free detection. The algorithm first integrates deep layer aggregation attention module in the backbone network to obtain multi-level object features, thereby expanding the receptive field and obtaining global feature information; Secondly, a RepPoints detection head based on anchor-free is adopted to reduce the switching frequency of object identification in the process of multi-object tracking; Finally, in the data association section, extended IOU matching is used to increase the range of detection and trajectory matching, compensate for motion estimation bias during matching, and thus improve the robustness of tracking in complex environments.

(3) The improved algorithm in this paper was tested on the MOT16 and MOT17 datasets and compared with other advanced algorithms. The experimental results showed that the tracking performance of the proposed algorithm was improved. On the MOT16 dataset, the multi-object tracking accuracy of the algorithms proposed in research content (1) and research content (2) reached 75.4% and 75.9%, respectively. On the MOT17 dataset, the multi-object tracking accuracy of the algorithms proposed in research content (1) and research content (2) reached 74.3% and 74.7%, respectively.

KEY WORDS: multi-object tracking, deep learning, attention mechanism, feature information, pedestrian re-identification

目 录

第 1 章 绪 论

1.1 研究背景及意义

近年来，随着人工智能技术[1]的飞速发展，智能设备在各个领域取得了显著的进展，且逐渐融入人们日常生活的方方面面。人工智能的根本目标是通过模拟、仿效或超越人类智能的方式，使计算机系统能够执行需要智力的任务，这包括学习、推理、问题解决和感知等复杂的认知功能。计算机视觉[2]作为人工智能的重要研究方向之一，它涉及到从图像或视频中获取、处理、分析和理解视觉信息，其关键目标是使计算机能够具备对视觉输入信息的感知和理解能力，从而能够模拟人类对视觉信息的解释和推理。

多目标跟踪[3]是计算机视觉领域中至关重要的任务，其主要目的是在图像或视频序列中准确追踪和识别多个目标，并维护它们在时间和空间上的运动轨迹。多目标跟踪包括目标检测[4]、目标识别[5]、数据关联[6]、轨迹更新[7]、运动预测[8]等多个环节，融合了图像处理[9]、模式识别[10]等多个领域的专业知识和技术，是一项复杂性、系统性的任务。

随着社会的发展，多目标跟踪逐渐融入人们生活的方方面面。在视频监控与安防[11]领域中，多目标跟踪算法能够在实时视频流中检测并跟踪多个目标，有助于识别与正常行为不符的异常活动，例如在禁止区域内的人员进入、长时间逗留、物体遗弃等，从而及时发现潜在的安全风险。安防系统可以通过多目标跟踪实现对指定目标的布控，并在目标出现或离开特定区域时触发告警，提高对潜在威胁的识别和应对速度。还可以结合其他智能分析技术，例如人脸识别、行为分析，实现更全面的视频内容理解，提供更精细化的安防决策支持。作为视频监控与安防领域的关键技术，多目标跟踪算法的研究具有巨大的产业价值。在自动驾驶[12]领域中，多目标跟踪可用于用于实时检测和跟踪车辆周围的障碍物，包括其他车辆、行人等。这有助于车辆规避障碍物，避免碰撞，并确保行驶路径的安全。还能够分析其他道路用户的行为，预测其动向和意图。通过多目标跟踪的应用，自动驾驶车辆能够更智能地感知和理解周围环境，从而提高在不同路况下的驾驶安全性和效率。在军事领域中，多目标跟踪主要应用于情报收集与侦察、目标识别与分类、精确火力打击、战场态势感知等方面。在精确火力打击方面，多目标跟踪技术可以帮助士兵和武器快速的锁定目标，实现火力打击的精确性。充分利用

高新技术，可以加强我国国防建设，提高我国综合国力。因此，对于应对现代化局部战争，多目标跟踪技术的研究显得尤为重要。多目标跟踪还在医疗图像分析应用广泛，可用于检测和定位多个病灶，如肿瘤、结节、血管等。通过实时跟踪这些目标，医生能够更准确地确定病变位置，提高诊断的精度。在手术过程中，多目标跟踪技术还可以帮助医生准确定位手术区域，提供实时的导航信息。这对于复杂手术或需要高精度操作的场景尤为重要。通过在医疗图像分析中应用多目标跟踪技术，可以提高医学影像的解读和分析效率，为医生提供更全面、准确的信息，促进精准医疗的发展。

1.2 国内外研究现状

随着科技的不断进步，多目标跟踪领域经历了多年的发展和演变。目前，多目标跟踪算法可分为两大类：传统的多目标跟踪算法和基于深度学习的多目标跟踪算法。

1.2.1 传统的多目标跟踪算法

在跟踪领域的早期，一些常见的早期跟踪算法包括均值滤波[13]（Meanshift Filter）、卡尔曼滤波[14]（Kalman Filter）、相关滤波[15]（Correlation Filtering）等。这些算法具有简洁而直观的数学表达，适用于计算资源较为有限的环境，为多目标跟踪研究奠定了基础。

其中，均值滤波是一种基于统计学的非参数化目标跟踪方法，其核心思想是通过寻找样本分布的峰值来确定目标的位置。均值滤波的优势在于其简单性和计算效率。然而，其对于单一目标、静态场景或近似静态场景表现较好，在处理目标具有快速运动或场景动态变化较大的情况时，均值滤波可能表现效果不佳；卡尔曼滤波是一种递归的状态估计算法。在多目标跟踪中，其被用于分别估计多个目标的运动状态，通过构建状态方程，获取目标的观测数据，从而预测目标下一时刻的状态。但是卡尔曼滤波在多目标跟踪中的应用面临一些困难，特别是当目标之间存在相互遮挡、交叉移动等情况时；相关滤波是一种基于模板匹配的目标跟踪和检测方法，其基本思想是首先定义一个目标模板，并对目标模板进行傅里叶变换，将其转换到频率域。然后将输入图像与在频率域中与傅里叶表示的目标模板进行计算，生成一个相关响应图，并在相关响应图中寻找峰值。最后根据峰值的位置，确定目标在输入图像中的位置，实现目标的定位。2010 年，Bolme

等人[16]首次将相关滤波应用于目标跟踪中，提出了一种新型的最小输出平方误差和相关滤波器，该滤波器提高了目标跟踪的稳定性和速度，能够处理跟踪过程中目标的光照变化、形变、遮挡等情况。2012 年，Henriques 等人[17]针对样本计算量大和跟踪准确率低的问题，提出了一种利用核函数和稠密采样的跟踪方法，有效的提高了目标跟踪的性能。2014 年，Henriques 等人[18]再次提出一种新的核化相关滤波器，与其他核算法不同，该相关滤波器扩展了特征提取的通道数，提高了特征提取的效率，从而提高跟踪的准确率。相关滤波的优势在于其简单性和计算效率，它在处理静态场景或目标纹理明显的情况下表现良好，但在目标遮挡、目标形变、噪声干扰等场景下跟踪效果不佳。

总体而言，这些基于传统算法的多目标跟踪方法在一些特定场景和限制条件下表现良好，但是随着多目标跟踪任务的复杂化和对性能的更高要求，基于传统滤波的跟踪方法已经很少采用。然而，这些传统算法仍然具有重要的作用，它们使研究者们对目标跟踪有了更直观的实践认识。

1.2.2 基于深度学习的多目标跟踪算法

近年来，随着深度学习在计算机视觉领域的普及，基于深度学习的多目标跟踪方法备受关注。这些方法通过利用基于深度学习的技术，结合目标的表征信息、运动信息以及位置信息等特征的相关性来关联同一个目标。相较于传统方法，基于深度学习的多目标跟踪算法无需手动选择特征，而是通过大量数据训练，使模型能够学到更为有效的特征提取能力。这种方法在提高跟踪准确性的同时，也提供了更好的适应性，能够应对复杂场景和目标外观变化的情况。总体而言，可以将基于深度学习的多目标跟踪算法分为基于检测的多目标跟踪算法和联合式多目标跟踪算法。

（1）基于检测的多目标跟踪算法

随着目标检测算法的不断进步，许多多目标跟踪算法采用了基于检测的跟踪方法。在早期的研究中，Bewley 等人[19]首先采用卡尔曼滤波对目标的下一帧检测框进行预测，然后通过测量检测框的重叠度，使用匈牙利算法进行视频帧之间的目标关联。Bochinski 等人[20]通过直接使用 IOU（Intersection over Union）匹配将目标最后一帧的检测框与当前帧中的检测框关联起来，简化了关联过程。Wojke 等人[21]以及 Yu 等人[22]尝试采用行人重识别（Pedestrian Re-identification, ReID）

网络[23]来生成更加稳定的目标运动、表征预测。简而言之，这些方法将每个目标从输入图像中提取出来，然后输入到额外的分支网络中，以提取身份标签（Identification, ID）的外观嵌入。通过计算不同视频帧之间的余弦距离，将目标与现有轨迹段相关联，从而提高多目标跟踪的准确性。

2018 年，Zhou 等人[24]以及 Fang 等人[25]将多目标跟踪任务分为两个独立的步骤：检测和跟踪。首先，通过使用高性能的目标检测算法，如 Faster R-CNN[26]、SDP[27]、DPM[28]等，确定每个目标的检测框。然后，通过将这些检测框与已有的轨迹关联，实现目标的跟踪。由于目标的检测框可以直接由检测算法生成，因此，基于检测的多目标跟踪算法侧重于提高数据关联的性能。2020 年，Braso 等人[29]引入图神经网络来进一步提高目标关联能力，利用图神经网络替代匈牙利算法对目标特征信息进行解析，并将其与边缘分类相关联。这种方法能够更好的获取目标之间的空间和时间关系，从而提高多目标跟踪的鲁棒性和准确性。2021 年，Chen 等人[30]提出了一种基于边缘多通道梯度模型的跟踪算法，该算法使用高精度的算子提取图像的边缘信息，并建立一个融合图像边缘时空光谱信息的多通道梯度模型。在此模型下，引入 ORB 特征来解决目标与目标库匹配的问题，从而解决多目标跟踪过程中的目标遮挡问题。2022 年，Yu 等人[31]针对智能监控视频中光照和遮挡导致行人跟踪精度低的问题，提出了一种基于 Work-Yolo 检测结合 DeepSORT 的行人跟踪算法。该算法采用改进的 YOLOv5 算法，提高了目标检测的准确率，减少了跟踪过程中行人遮挡导致的 ID 切换问题。2023 年，Li 等人[32]提出一种基于 M3-YOLOv5 的轻量级多目标检测跟踪算法，该算法融合注意力机制和 MobileNetV3 算法，降低了整体网络的计算量，提高了跟踪速度。

（2）联合式多目标跟踪算法

近年来，联合式多目标跟踪方法因其跟踪效果良好而备受关注。相比于基于检测的多目标跟踪算法，联合式方法是将目标检测部分和 ID 嵌入提取部分相融合，共享大量目标特征，同时进行目标检测信息的输出和 ID 的嵌入。该方法可以显著的减少计算量，提高跟踪的速度。

2017 年，Xiao 等人[33]在研究中首次对 Faster R-CNN 检测器进行了修改，以同时处理检测任务和 ReID 任务。这一改进包括引入额外的全连接层，用于生成包含目标 ID 嵌入的表示。2019 年，由于 Mask R-CNN 检测网络[34]本身带有一个实例分割分支，能够有效提取目标的像素级特征，从而有效提高多目标跟踪的准

确性。因此，Voigtlaender 等人[35]以 Mask R-CNN 检测网络为基础，并引入了一个特征提取分支，该特征提取分支用于提取每个候选区域的外观特征。2020 年，Wang 等人[36]和 Lu 等人[37]采用了类似的思路，通过重新设计预测头，将基于特征金字塔网络[38]（Feature Pyramid Network, FPN）的检测器转变为联合式跟踪器，提高了跟踪的速度。由于注意力机制能够对重点的候选区域进行关注，并增强这些区域的信息，从而提取目标的更多细微特征信息，并抑制不重要的信息。因此，K. Yoon 等人[39]在联合式学习框架的基础之上，融合注意力机制，将新检测到的目标关联到现有的跟踪轨道上，而不考虑视频帧中出现的目标的数量。该方法提出了一种新的数据关联方式，可以识别检测器输出的假阳性目标。

虽然上述联合式多目标跟踪算法有效的减少了跟踪网络的计算量，提高了多目标跟踪的速度，但其多目标跟踪准确率并未明显优于基于检测的多目标跟踪算法。2021 年，为了提高跟踪的准确率和进一步减少跟踪时间，Zhang 等人[40]在检测部分采用基于无锚框的目标检测方法，并在此基础上添加一个 ReID 网络分支来区分不同目标。该方法解决了检测任务和重新识别任务之间存在的目标特征提取不公平和目标特征维度不匹配的问题。Liu 等人[41]在 FCOS 网络[42]的基础上引入了一种基于可变形卷积的区域转换模块，旨在减少模型对一些无关区域的关注。因为 FCOS 网络可以通过特征金字塔架构融合不同尺度的跟踪目标特征，使得提取的跟踪目标特征更适用于检测任务和 ReID 任务。

此外，2022 年，Mostafa 等人[43]提出了一种实时的轻量级多目标跟踪器，采用一个简化的深层聚合（Deep Layer Aggregation, DLA）网络来提取用于检测的目标特征，利用线性转换器对每帧输入图像的检测热图进行特征提取，生成有效的跟踪特征，从而进一步提高跟踪的速度。Li 等人[44]开发了一种基于卷积神经网络和空间通道注意机制的在线多目标跟踪方法，利用扩展卷积适应目标的变形，使整体网络动态地关注重点任务，忽略不相关的信息。该方法能够使得跟踪过程中的漏检目标数量大幅降低。2023 年，Yan 等人[45]也在 FCOS 网络的基础上融合了一个特征提取分支网络，并提出了一个高效的目标检测框架，在提高跟踪准确率的同时减少了跟踪速度。然而，这些方法在复杂环境下的跟踪效果仍然欠佳，需要进一步进行研究探索。

1.3 多目标跟踪的挑战性分析

多目标跟踪任务是一个综合任务，需要同时解决多个子任务的难点问题。由于该任务在视频数据的基础上进行跟踪，研究者们一直在努力攻克一系列与时空相关性等问题相关的难点。这包括在处理视频数据时面临的挑战，如目标的遮挡、目标外观变化、目标 ID 频繁切换、跟踪实时性，以及在时序数据中保持准确跟踪等问题。部分困难点如下：

（1）目标遮挡问题

目标遮挡是多目标跟踪领域中一个常见而具有挑战性的问题。在目标运动过程中，可能会受到其他目标、场景中的物体或环境的遮挡，导致目标的部分或完全不可见。这种情况使得跟踪系统在准确估计目标位置和轨迹方面面临更大的困难。如何完全解决目标遮挡问题，还需研究者们在前进的道路上继续探索。

（2）目标外观变化问题

在多目标跟踪中，目标外观变化问题是一项具有挑战性的任务。在目标运动的过程中，可能会受到光照、视角、姿态等多方面的变化影响，导致目标的外观在不同时间点呈现出明显的差异。这种外观的不断变化增加了跟踪算法在处理多目标时的复杂性。解决这一问题有待采用更为创新性的方法和技术，以提高系统对目标外观变化的适应性，确保更为准确和鲁棒的跟踪性能。

（3）跟踪实时性问题

多目标跟踪中的实时性问题指的是系统在短时间内需要快速而准确地处理和跟踪多个目标的挑战。实时性在众多应用中都是一个至关重要的要求，尤其在视频监控、自动驾驶、无人机导航等领域更为突出。系统必须以及时、高效的方式处理目标跟踪任务，以满足对即时性能的严格要求。尽管现在的多目标跟踪算法流程结构得到了极大地简化，计算复杂性也大大降低，但在一些高实时性领域依旧有所欠缺。

（4）平衡各子任务问题

多目标跟踪任务涉及到多个子任务，如目标检测、数据关联、运动预测等，如何在有限的计算资源的前提下，平衡这些子任务是确保整个系统高效运行的关键性问题。

1.4 本文主要研究内容和结构安排

本文主要研究基于深度学习的多目标跟踪算法，针对多目标跟踪过程中因目标遮挡、形变等所导致的跟踪准确率不足的问题，提出切实可行的解决方案。本文有五个章节组成，各章节的具体研究内容如下：

第 1 章：绪论。本章节主要描述了多目标跟踪的研究背景及意义，对比了国内外传统的多目标跟踪算法和基于深度学习的多目标跟踪算法的研究发展现状，分析了多目标跟踪的难点问题，并描述本文主要研究内容以及章节结构安排。

第 2 章：多目标跟踪算法相关理论。本章节主要阐述了卷积神经网络的基本原理，从有无锚框的角度描述了目标检测算法，分析了多目标跟踪的基本结构，以及卡尔曼滤波与匈牙利匹配算法原理，最后描述了多目标跟踪的数据集和评价指标。

第 3 章：融合注意力机制的行人多目标跟踪算法。本章节针对多目标跟踪过程中因遮挡引起的跟踪准确率不高的问题，从主干网络和头部网络对联合式多目标跟踪算法进行改进。首先将坐标注意力机制[46]（Coordinate Attention, CA）融入特征金字塔网络架构中，以提高所提网络的检测精度；然后融合空间注意力机制和通道注意力机制，在 ReID 部分提出了一个尺寸感知模块，以适应不同分辨率的目标特征，应对跟踪过程中的遮挡问题；最后由于 YOLOv5s 检测头能够利用基于网格的锚点在不同尺度的特征图上进行目标检测，且具有较高的检测速度和精度，可以实现实时检测，适用于需要快速响应的场景。因此，在检测部分对 YOLOv5s 检测头进行改进，融合小目标检测层，从而获得更好的检测效果。与其他先进方法相比，本章所提网络可以有效提高遮挡情况下的多目标跟踪准确率，跟踪速度也满足实时性要求。

第 4 章：基于无锚框检测的行人多目标跟踪算法。本章节在第三章的基础上进行研究，主要针对多目标跟踪过程中目标 ID 频繁切换的问题。为了解决主干网络提取目标特征不充分的问题，本章节在主干网络部分融合深层聚合注意力模块以获得多层次的目标特征；由于锚框不适合进行目标重识别特征的提取，所以采用基于无锚框的 RepPoints 检测头[47]，从而避免锚框对重识别任务的影响，降低多目标跟踪过程中 ID 切换次数；在数据关联部分采用扩展 IOU 匹配，增加检测和轨迹的匹配范围，从而提高复杂环境下跟踪的稳定性。实验表明，所提算法

具有良好的跟踪性能，且缓解了多目标跟踪过程中目标 ID 的切换频率。

第 5 章：总结与展望。本章节主要对本文的工作进行总结和归纳，分析了本文所提算法的优缺点，并对未来的研究工作进行了展望。

第 2 章 多目标跟踪算法相关理论

深度学习的兴起对计算机视觉领域产生了巨大的影响，深度学习算法如今已经成为解决计算机视觉问题的不可或缺的方法。丰富的实验证明，检测算法的性能直接影响着多目标跟踪算法的整体跟踪性能，所以本章节在描述卷积神经网络的同时，从有无锚框的角度阐述了目标检测算法。然后分析了多目标跟踪算法的基本结构，阐述了卡尔曼滤波与匈牙利匹配算法。最后简要描述了多目标跟踪的数据集和评价指标。

2.1 卷积神经网络概述

卷积神经网络（Convolutional Neural Networks, CNN）是一种深度学习模型，主要用于处理和分析具有网格结构的数据，例如图像和视频。CNN 以生物学上视觉皮层的运作方式为灵感，通过学习图像中的特征来完成各种任务，如分类、检测、分割和识别等。该网络的研究始于二十世纪 80 至 90 年代，早期出现的时间延迟网络和 LeNet-5 是最早的卷积神经网络实例[48]，它们在图像处理和模式识别领域取得了一些突破，并为后来更复杂的卷积神经网络奠定了基础。CNN 的训练通常通过反向传播算法进行，通过优化损失函数来调整网络参数，以使网络的预测结果与实际标签尽可能接近。由于 CNN 能够自动学习图像中的特征，因此在计算机视觉领域取得了巨大的成功，并被广泛应用于图像分类、目标检测、图像分割等任务中。

CNN 的主要结构模块包括：卷积层、池化层、激活函数以及全连接层，其中神经元接收到输入后，通过权重传递给对应的激活函数，进而产生输出结果。激活函数负责引入非线性特性，使网络能够更灵活地学习和捕捉数据中的复杂模式。

（1）卷积层

卷积层的主要作用是利用多个卷积核对输入图像进行特征提取，如图 2-1 所示，在卷积操作中，卷积核以规律的方式在输入特征上滑动，对连接区域内的输入特征进行矩阵元素的乘法求和，并添加偏差量。这种操作有效地捕捉了输入数据中的局部特征，使得卷积层能够学习到空间层次的表示，有助于提取更高级别的特征。

在标准卷积的基础上，部分卷积神经网络采用了更为复杂的卷积操作，包括：反卷积、可分离卷积、可变形卷积等。反卷积也称为转置卷积或上采样卷积，是一种用于将特征图的尺寸扩大的卷积操作。与标准卷积相反，反卷积操作的目标是在输入特征图上进行上采样，从而增加输出的空间分辨率。可分离卷积分为深度可分离卷积和空间可分离卷积，其中深度可分离卷积包含深度卷积和逐点卷积两个部分，广泛应用于深度学习网络，如 MobileNet[49]等。在标准卷积中，是同时考虑图像区域中的所有通道的，而深度可分离卷积通过独立处理通道和空间区域，针对不同的输入通道采用不同的卷积核进行卷积操作。这种分解操作将传统卷积拆分为两个步骤，旨在通过较少的参数来学习更丰富的目标特征表示。空间可分离卷积的核心思想是从图像的空间维度进行卷积操作，其将卷积核分为两个部分，分别对输入的目标特征图进行卷积运算，从而降低计算量。可变形卷积是一种卷积神经网络中的高级卷积操作，旨在提高模型对非规则形状和空间变化的适应性。与传统的固定形状的卷积核不同，可变形卷积允许卷积核的形状在每个位置自适应地变化。可变形卷积通过引入自适应的形状变换机制，对复杂的图像结构和形状变化进行建模，有助于提高卷积神经网络在特定任务上的性能。

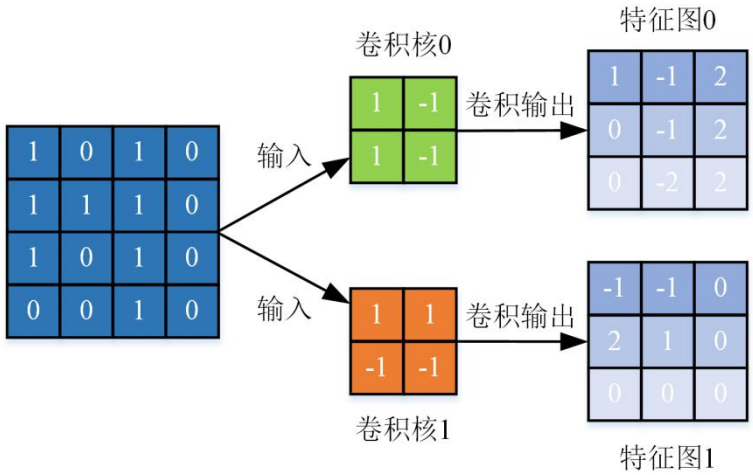

图 2-1 卷积操作示意图

（2）池化层

在卷积层提取特征后，所得的特征图将被输入到池化层，池化层的主要作用是进行特征选择和信息过滤。池化层能够对输入的特征图进行下采样，以减少数据量，同时保留关键的特征信息。这个过程有效地降低了数据量和计算复杂度，有助于提升模型的训练速度和泛化能力。

池化通常可分为最大池化和平均池化。如图 2-2 所示，最大池化操作将输入

特征图划分为多个子特征图，然后从每个子特征图中选取最大像素值作为该子特征图的输出值。通过这种方式，最大池化能够对输入特征图进行目标尺寸和目标特征的下采样，从而减少计算负担。相反，平均池化操作将输入特征图分割成多个子特征图，然后计算每个子特征图中所有像素值的平均值，并将该平均值作为子特征图的输出值，这种方法同样有助于减少计算量。

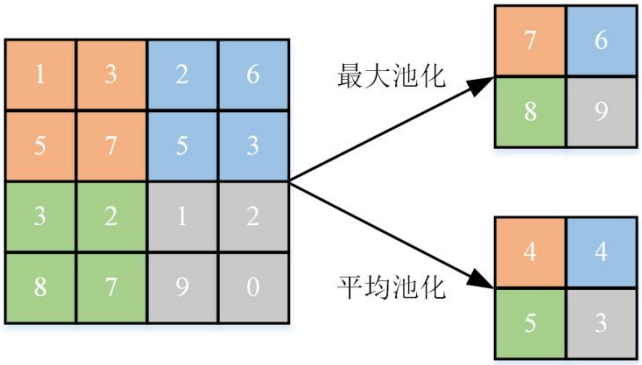

图 2-2 最大池化和平均池化操作示意图

（3）激活函数

激活函数在神经网络中的作用至关重要，它们引入了非线性特性，从而使神经网络能够更好地学习和理解复杂非线性模型。在没有激活函数的情况下，每一层都相当于进行矩阵相乘，每一层的输出都是上一层输入的线性函数。这样，无论神经网络有多少层，整个网络的输出都可以表示为输入的线性组合，无法捕捉到复杂的非线性关系。通过引入激活函数，每一层的输出不再是简单的线性组合，而是经过非线性变换，使得神经网络能够更好地适应和学习非线性模式，提高模型的表达能力和拟合复杂数据的能力。因此，激活函数的引入对于神经网络的深层结构和学习非线性关系起到了至关重要的作用。目前常用的激活函数包括：Sigmoid、Tanh、ReLU 等。

Sigmoid 激活函数是一种常用的激活函数，通常用于二分类问题或者输出层的多分类问题。如图 2-3 所示，Sigmoid 函数的输出在（0,1）之间，可以被解释为概率值，在二分类问题中表示正类别的概率，其函数曲线光滑且可导，这有助于在梯度下降等优化算法中进行参数更新。但是在深层神经网络中，Sigmoid 函数容易导致梯度消失的问题，尤其是在反向传播时，梯度可能变得非常小，从而导致难以更新参数，其方程如式（2-1）所示：

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

其中，x 表示特征输入，$f(x)$ 表示 Sigmoid 函数特征输出。

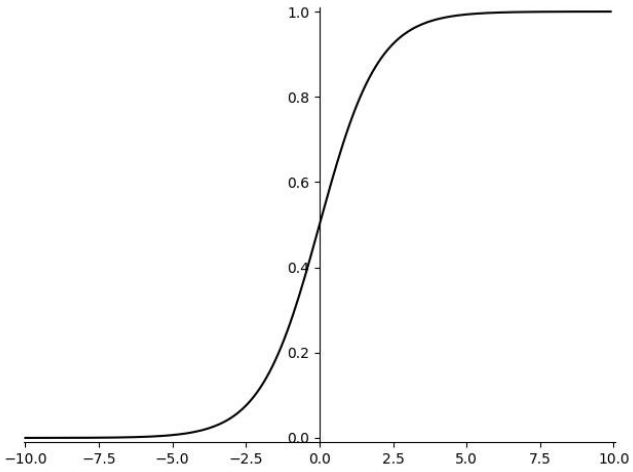

图 2-3 Sigmoid 激活函数曲线图

Tanh 激活函数的输出在（-1,1）之间，这使其对于具有负数输入的神经元而言，可以进行更富有表现力的表示。如图 2-4 所示，与 Sigmoid 激活函数相比，Tanh 激活函数的输出范围更广，包括了负数，因此输出的均值接近于零。这有助于缓解数据分布偏移问题，因为在训练时输入的均值可能在正数或负数方向上变化，其方程如式（2-2）所示：

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

其中，x 表示特征输入，$f(x)$ 表示 Tanh 激活函数特征输出。

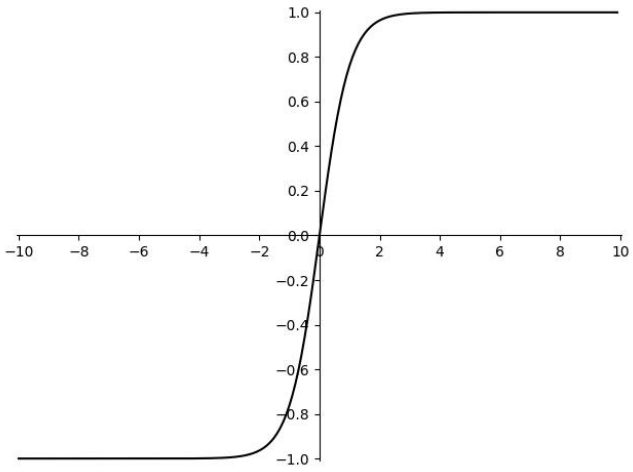

图 2-4 Tanh 激活函数曲线图

ReLU 激活函数又称修正线性单元，是一种非线性激活函数，引入了神经网络的非线性特性，使其能够学习和表示更复杂的函数。相对于 Sigmoid 和 Tanh

激活函数，ReLU 激活函数的计算更加简单和高效。如图 2-5 所示，在正数输入情况下，梯度始终为 1，使得反向传播的计算更为稳定。其导数在正数区域为 1，这有助于缓解梯度消失问题，提高模型的训练效率，其方程如式（2-3）所示：

$$f(x)=\max(0,x) \tag{2-3}$$

其中，x 表示特征输入，$f(x)$ 表示 ReLU 激活函数特征输出。

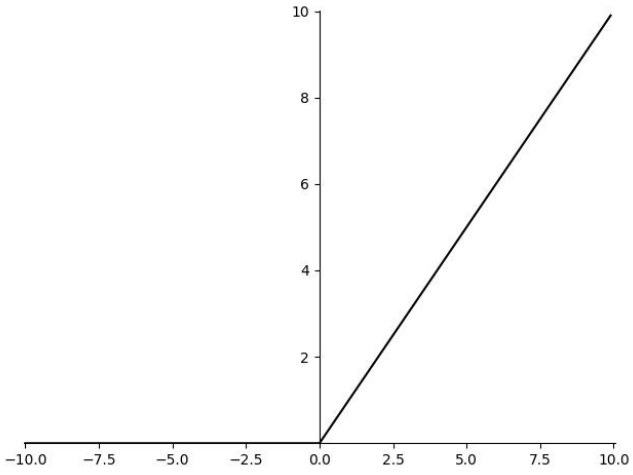

图 2-5 ReLU 激活函数曲线图

（4）全连接层

全连接层将卷积层和池化层输出的特征进行展开，并连接到最终的输出层。如图 2-6 所示，其中每个节点都与上一层的所有节点相连。这种连接方式使得全连接层能够将前一层提取到的特征综合起来。由于每个节点都与上一层的所有节点相连，全连接层的参数数量也是最多的，因此具有较高的灵活性和表达能力。

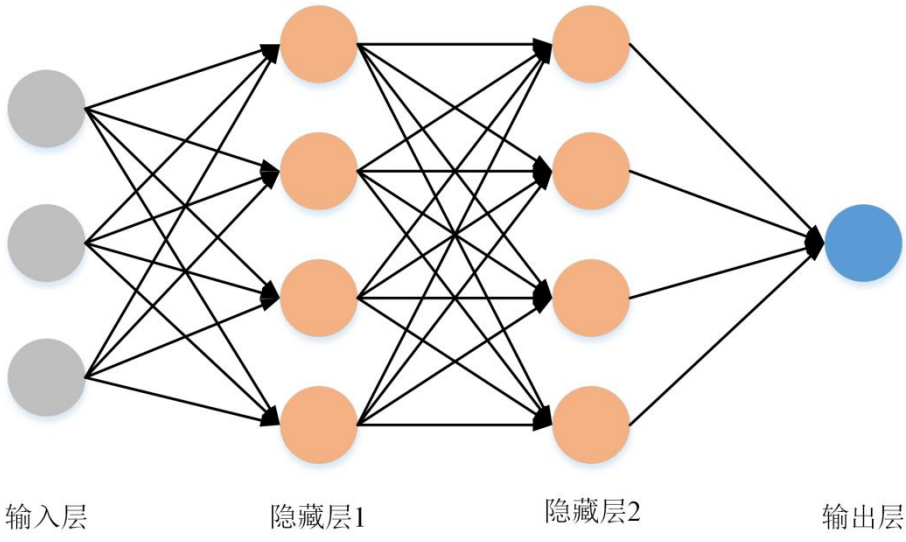

图 2-6 全连接操作示意图

在全连接层中，每个连接都有一个对应的权重，这些权重用于调整输入特征的影响力，从而完成特征的融合和变换。全连接层的输出可以看作是对输入特征的加权组合，通过这种组合可以捕捉到更高层次的抽象特征和关系。全连接层常用于神经网络的最后一层，用来产生最终的输出，或在中间层用于学习更复杂的特征表示。尽管全连接层参数较多，但也因此具备了强大的学习能力，可以适应各种复杂的模式和任务。

2.2 目标检测算法

作为目标跟踪的一部分，目标检测任务不仅需要确定图像中物体的存在，还要为每个检测到的对象分配正确的类别标签，并通过边界框精确标示出它们的位置，其在人脸识别[50]、机器人控制[51]、行人检测[52]等领域应用广泛。随着科技的进步，基于深度学习的目标检测算法已经成为目标检测领域的主流算法，这些算法利用深度神经网络来学习图像中的特征，并通过一系列网络层来实现目标的定位和识别，其可以分为基于锚框的目标检测算法和基于无锚框的目标检测算法。

2.2.1 基于锚框的目标检测算法

基于锚框的目标检测算法通常分为两个主要阶段：生成锚框和检测目标。首先在输入图像上均匀生成一组锚点，这些锚点作为候选框的中心点。锚点的数量和尺寸可以根据实际任务和数据集进行调整。对于每个锚点，生成多个不同尺寸和宽高比例的候选框。这些候选框涵盖了不同形状和大小的目标可能出现的位置；然后，将输入图像和生成的候选框送入深度卷积神经网络进行特征提取，通常使用预训练的卷积神经网络（如 VGG、ResNet）的底层网络作为特征提取器；最后，对每个候选框进行分类计算和回归计算。对于具有相似位置和类别的多个候选框，使用非极大值抑制算法[53]来去除冗余的检测结果，从而保留最相关、最准确的目标框作为输出结果。每个检测结果包括目标的类别和位置等信息。其中，Faster R-CNN、YOLO[54]、SSD[55]、RetinaNet[56]等算法通过基于锚框的方法，实现了对不同尺寸和形状的目标的有效检测，并在实际应用中取得了良好的性能。

然而，由于基于锚框的目标检测算法对锚框的依赖性较强，因此容易产生以下问题：

（1）锚框的设计使得整体网络变得更加复杂，在训练过程中需要消耗大量的计算资源。

（2）为了提高目标匹配的精准性，通常需要预设大量的锚框。然而，这其中大部分锚框只包含背景信息，被视为负样本，而只有少数锚框包含目标信息，被视为正样本。这样的设计可能会导致目标检测中存在的正负样本失衡问题。

（3）锚框设计的质量在很大程度上决定了目标检测算法的性能上限，这使得检测算法受到一定的局限性。

2.2.2 基于无锚框的目标检测算法

近年来，针对基于锚框的目标检测算法存在的一些缺陷，许多研究者致力于探索如何在不使用锚框的情况下实现类似的检测性能。在这一趋势下，涌现出一系列基于无锚框的目标检测算法，避免了设计锚框的步骤，从而提高检测网络对不同类别目标的适应性，并降低计算量。通常可将基于无锚框的检测算法分为基于关键点的无锚框目标检测算法和基于 Transformer 无锚框目标检测算法。

（1）基于关键点的无锚框目标检测算法

基于关键点的无锚框目标检测算法首先利用特征提取网络，对输入视频帧进行特征提取，得到信息丰富的特征图；其次在得到的特征图上采用关键点预测模块，获得目标关键点的坐标信息。关键点预测模块通常由多个子模块组成，每个子模块用于预测一个关键点。这些关键点通常代表目标框的重要位置，如角点、中心等。通过这些坐标信息，确定特征提取后的预测框的关键点位置；然后将定位得到的关键点组合在一起，构成目标框。这个过程通常需要计算边界框的宽度和高度来生成目标的检测框。在关键点组合的过程中，可能会进行关键点的匹配操作，以确定特征关键点的对应关系；最后利用生成的目标框来完成目标检测任务，确定目标的位置和类别。基于关键点的无锚框目标检测算法直接关注目标特征的关键点，避免了锚框的设计，从而提高对多类目标的适应性，并简化模型的结构。

DenseBox 算法[57]是最早尝试基于无锚框检测的算法之一，其创新性在于使用中心点作为对象的初始表征，并在图像的所有位置直接预测对象的四个边界以及类别，提升了算法对不同目标尺寸的适应性。2017 年，Wang 等[58]提出了一种点链接目标检测网络，该网络首先使用全卷积网络对目标边界框的角点、中心点以及其连接进行回归；然后将角点和其连接映射到多个边界框；最后通过融合多个边界框得到目标检测结果。2018 年，Law 等[59]针对角点难以定位目标位置的

问题，提出了 CornerNet 检测算法，以角点作为关键点，通过采用修正后的残差模块、角池化模块和卷积模块，分别预测目标左上角点和右下角点的特征热图、嵌入向量和偏移量。2019 年，Zhou 等人[60]提出了一种名为 CenterNet 的目标检测算法，其创新性在于无需划分输入图像，将多点检测问题转换为单中心点检测问题。与其他方法不同，CenterNet 在后处理阶段无需使用非极大值抑制算法，这降低了整体网络的计算复杂度，在提高目标检测精度的同时保证了目标检测的速度。2020 年，Duan 等[61]在 CornerNet 的基础上提出了一种 CenterNet 算法，旨在解决角点误匹配问题。该算法将每个目标的检测作为一个三元组，而不是一对关键点，并设计了级联角池化模块和中心池化模块，用于提取特征图左上角和右下角的目标信息，以及获得更多中心区域的可识别目标信息，从而提高目标检测的准确率和召回率。这种改进优化了角点的匹配，增强检测网络对目标内部结构的关注，从而在目标检测中取得更为精准和鲁棒的性能。

基于关键点的无锚框目标检测算法简化了模型复杂度，提升了检测性能，为无锚框目标检测提供了一种新的思路。

（2）基于 Transformer 无锚框目标检测算法

近年来，Transformer 模型利用其强大的注意力机制在结构化任务中取得显著成功，如文本翻译和语言建模。在目标检测领域，基于 Transformer 的方法通常由编码器和解码器组成。编码器结合自注意力机制模块和多层前馈神经网络（Feed forward Network, FFN）模块，有效地获得目标的感受野。解码器在编码器的基础上融合了“编码-解码”模块，用于探索编码前后目标特征信息之间的关系，这使得模型能够更好地学习目标之间的联系，提高目标检测的准确性。

基于 Transformer 的目标检测算法通常采用端到端的训练方式，整个模型可以直接在目标检测任务上进行优化，避免了手工设计特征提取器和检测器的繁琐过程，简化了模型的设计和训练流程。它采用自注意力机制的方式，使得模型能够在处理输入视频序列时同时关注到全局信息和局部信息，学习不同目标之间的关联信息。相比之下，传统的卷积神经网络在处理图像时只能通过卷积操作获取局部信息，而往往需要通过多层卷积或者特征金字塔来获取全局信息，增加了模型的复杂度。基于 Transformer 的目标检测方法通过其强大的建模能力，为目标检测任务带来了新的思路和性能提升。

2020 年，Carion 等人[62]提出了一种端到端目标检测器（End-to-end object

detection with Transformers, DETR），将 Transformer 融合到目标检测器中。DETR 的模型结构包括了特征学习模块，可以进一步增强目标特征，从全局目标信息中获取相关的目标特征信息。其可利用前馈神经网络结构，分析每个目标查询中的对象信息，得到最终的预测结果集合。随后，在 DETR 的基础上，Deformable DETR 算法[63]、Dynamic DETR 算法[64]、SAM-DETR 算法[65]、DN-DETR 算法[66]等相继出现，极大地促进了 Transformer 模型在目标检测领域的发展。

2.3 多目标跟踪的基本结构

多目标跟踪的主要任务是在给定图像序列中识别运动的目标，并为这些目标分配唯一的 ID，以实现不同帧之间的目标一一对应，进而输出准确的目标运动轨迹，展示每个目标在时间序列中的移动路径和位置变化。它并非是独立的计算机视觉问题，而是以目标检测为基础进行的更进一步的任务。

随着多目标跟踪算法的发展，其已经形成了一些关键性的结构模块。如图 2-7 所示，多目标跟踪器一般包括目标检测、特征提取、数据关联三个基本模块。

各模块的功能如下：

（1）目标检测：多目标跟踪的第一步通常是目标检测，通过使用目标检测算法来识别图像或视频帧中的目标。流行的目标检测算法包括 Faster R-CNN、YOLO 系列、SSD 等。

（2）特征提取：在多目标跟踪中，特征提取是关键步骤之一，用于捕捉目标的外观信息以及帮助建立目标的表示模型。常用的特征提取方法有：灰度特征、颜色特征、纹理特征、运动特征、形状特征、深度学习特征等。在实际应用中，通常会结合多种特征，形成多维特征向量，以全面捕捉目标的外观信息和运动信息。特征的选择和组合需要根据具体的场景和目标特性进行调整，以取得最佳的跟踪性能。

（3）数据关联：数据关联是多目标跟踪任务中的至关重要的步骤，旨在通过处理和关联目标的观测数据，实现对多个目标的同时跟踪和识别。这一过程涉及从多个传感器或多个时间步骤的数据中准确地建立目标轨迹的问题。在这一阶段，系统模型需要精确地将来自不同源的数据信息相关联，以确保对目标的准确追踪和身份确认。这包括采用各种算法和技术，例如卡尔曼滤波器、粒子滤波器、多假设跟踪以及深度学习方法等，以有效地处理和解决复杂的多目标跟踪挑战。

因此，数据关联在整个多目标跟踪流程中扮演着关键的角色，直接影响模型的准确性和鲁棒性。

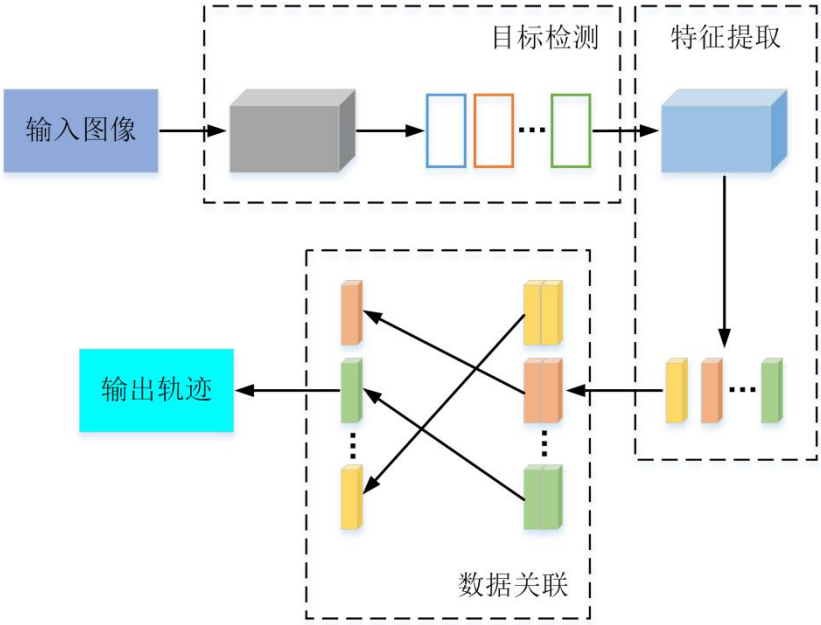

图 2-7 多目标跟踪的基本模块示意图

2.4 卡尔曼滤波与匈牙利匹配算法

卡尔曼滤波（Kalman Filter）是一种用于估计动态系统状态的数学方法，广泛用于机器视觉、雷达定位、航天航空等领域，其基本思想是通过融合系统动态模型和观测数据来对系统状态进行最优估计。通过考虑各个时间点下测量值的联合分布，卡尔曼滤波能够生成更为准确的未知变量估计，相比于仅基于单一测量值的估计方法更为精准。卡尔曼滤波算法一般包括预测和更新两个部分，具体原理如下：

（1）预测部分：卡尔曼滤波算法根据系统的动态模型，利用当前时刻的状态估计和外部输入来预测系统在下一个时刻的状态，并估计预测状态的不确定性，其数学表示如式（2-4）和（2-5）所示：

$$\hat{x}_k^- = F_k \cdot \hat{x}_{k-1} + B_k \cdot u_k \tag{2-4}$$

其中，$\hat{x}_k^-$ 表示 k 时刻预测的系统状态，$\hat{x}_{k-1}$ 表示 $k-1$ 时刻的系统状态，B_k 表示输入矩阵，u_k 表示外部输入。

$$P_k^- = F_k \cdot P_{k-1} \cdot F_k^T + Q_k \tag{2-5}$$

其中，P_k^- 表示 k 时刻预测的状态协方差矩阵，F_k 表示 k 时刻状态转移矩阵，P_{k-1}

表示$k-1$时刻的状态协方差矩阵，F_k^T表示k时刻状态转移矩阵的转置矩阵，Q_k表示过程噪声的协方差矩阵。

（2）更新部分：卡尔曼滤波根据当前时刻的观测值和预测值，计算出当前时刻的状态估计值。由于其已经充分考虑了当前时刻的观测信息，因此这个估计值相较于预测值更加准确。通过在预测部分中计算得到的预测状态协方差矩阵、过程噪声的协方差矩阵以及卡尔曼增益的综合影响下，可以得到状态估计值的误差协方差矩阵。其数学表示分别如式（2-6）、（2-7）和（2-8）所示：

$$K_k = P_k^- \cdot H_k^T \cdot \left(H_k \cdot P_k^- \cdot H_k^T + R_k\right)^{-1} \tag{2-6}$$

$$\hat{x}_k = \hat{x}_k^- + K_k \cdot \left(z_k - H_k \cdot \hat{x}_k^-\right) \tag{2-7}$$

$$P_k = \left(I - K_k \cdot H_k\right) \cdot P_k^- \tag{2-8}$$

其中，K_k表示卡尔曼增益，$\hat{x}_k$表示更新后的系统状态估计，P_k表示更新后的状态协方差矩阵，H_k表示观测矩阵，H_k^T表示观测矩阵的转置矩阵，R_k表示测量噪声的协方差矩阵，z_k表示测量值，I表示单位矩阵。

通过预测和更新两个部分，卡尔曼滤波在每个时刻都通过融合先验信息和测量信息，动态地更新对系统状态的概率分布，从而提高状态估计的精度。在目标跟踪任务中，由于目标运动的不确定性，直接使用目标当前的位置坐标和速度信息计算目标下一时刻的位置是不可靠的。而卡尔曼滤波能够使用滤波算法来消除目标的不确定性噪声，并利用目标的历史位置和速度等状态信息，对目标的下一时刻位置和速度进行合理的预测。此外，卡尔曼滤波还可以根据跟踪模型获得的目标下一时刻位置数据来优化更新其预测结果。这使得卡尔曼滤波算法成为目标跟踪任务中一种强大的运动模型，能够有效地应对目标运动中的不确定性和噪声影响。

匈牙利匹配算法（Hungarian algorithm）是一种用于解决指派问题的优化算法，该算法最初由匈牙利数学家 Dijkstra、Konig 和 Egervary 于 1950 年提出，被广泛应用于解决工作分配和人员分配等实际问题。

匈牙利算法解决的指派问题是指在给定的成本矩阵中寻找一组最优方案分配给执行者，从而最小化总成本。这个问题可以通过构建一个二分图来表示，其中任务和执行者分别形成两个顶点集合，边表示任务分配给执行者的可能性，并

带有相应的成本或权重。匈牙利算法的目标是找到一个最小权重的完美匹配，即每个任务分配给一个执行者，同时每个执行者负责一个任务，以使得总成本最小。具体算法步骤如下：

（1）初始化：对成本矩阵进行初始化。如果成本矩阵不是方阵，可以通过添加零元素将其扩展为方阵。同时，为每一行找到最小值，然后从每个元素中减去相应行的最小值，确保每一行至少有一个零。

（2）零标记：在调整后的成本矩阵中，尽量找到一个零元素，并对其所在的行和列进行标记，以表示该任务已分配给执行者。如果找到的零元素数量等于矩阵的大小，则找到了最优匹配，如果零元素数量不等于矩阵的大小，则进入零划线步骤。

（3）零划线：利用零标记的行和列，在成本矩阵上划线，尽量使每一行和每一列都包含一个零。若有未被划线的零元素，则执行增广路径步骤，否则进行调整元素步骤。

（4）增广路径：增广路径是一条从未被划线的零元素开始到已经被划线的零元素结束的路径。

（5）调整元素：通过增广路径上的零元素，调整未被划线的元素和被划线的元素，使得已划线的元素变成未划线，未划线的元素变成已划线。

重复上述步骤，直到找到最优匹配为止。根据零标记的行和列，找到最优的任务分配方案。

在多目标跟踪中，匈牙利算法通常被用来处理目标数据关联的最优化分配问题。成本矩阵中的元素表示检测到的目标与已知目标之间的关联成本，这些成本可能基于目标之间的距离、运动特征、外观相似度等因素。通过匈牙利算法，系统能够高效地完成目标数据关联，同时考虑到不同目标之间的关联成本。这有助于跟踪系统更准确地维护目标的运动轨迹，识别目标的出现和消失，以及适应动态场景中目标的变化。因此，匈牙利算法在多目标跟踪中的应用为系统提供了一种优化的数据关联策略，提高了整体跟踪性能。

2.5 数据集与评价指标

2.5.1 多目标跟踪数据集

多目标跟踪的目标主要包括车辆和行人，由于行人具有不确定性，其运动轨

迹和形状更为复杂，因此跟踪行人相对更具挑战性，困难程度更大。由此，目前的多目标跟踪算法主要以行人为研究对象，相关的数据集也大多以跟踪行人而创建。其中，常用的数据集包括 MOT15 数据集[67]、MOT16 数据集、MOT17 数据集[68]、MOT20 数据集[69]等。

MOT15 数据集是 MOTChallenge 发布的首个多目标跟踪数据集，由多个分散的视频跟踪数据源整理而成，共包含 22 个视频片段，其中 11 个视频片段用于训练，其余 11 个用作测试，共有 11283 帧图像，包含 1221 个不同的行人目标，提供了 101345 个边界框标签。该数据集的设计目的是为了提供一个标准基准，用于评估和改进多目标跟踪算法，视频中的目标包含了遮挡和背景变化等因素，增加了跟踪问题的复杂性，对算法的鲁棒性提出了挑战。

MOT16 数据集是一个用于多目标跟踪的公开数据集，该数据集由跟踪算法的评估与基准（MOT Challenge）组织提供，并于 2016 年发布。MOT16 数据集包含多个视频序列，每个序列都包含不同数量和类型的目标，例如行人、车辆等。每个序列都配有图像帧以及与之对应的标注文件，标注文件提供了目标的边界框和 ID 信息，用于评估跟踪算法的准确性和鲁棒性。该数据集提供了多种挑战，例如目标遮挡、目标尺度变化、目标外观变化等，从而能够全面评估跟踪算法在复杂环境下的性能表现。MOT16 数据集已成为评估目标跟踪算法性能的标准基准之一，广泛应用于学术研究和实际应用中。

MOT17 数据集是在 MOT16 数据集的基础上提出的。相较于 MOT16 数据集，MOT17 数据集使用了一个全新的、更准确的跟踪场景，每个序列都有 DPM，Faster-RCNN，SDP 这三组检测。

相比于上述多目标跟踪数据集，MOT20 数据集环境更复杂、人群更密集、任务难度更大，其视频最多可达每一帧 246 人。该数据集共包含 8 个视频片段，涵盖了三个不同的场景。其中，4 个视频片段用于训练，而另外 4 个视频片段则用于测试。整个数据集以视频帧的形式提供，总计包含 13410 帧，其中训练视频占了 8931 帧，测试视频为 4479 帧，每个视频帧都配有矩形框标注，但测试数据的标注并未公开。在训练数据中，每帧平均包含 127.04 个行人矩形框标注，而测试数据则平均每帧含有 115.52 个行人矩形框标注。除了行人之外，数据集中还包含其他类型的矩形框标注，例如非移动交通工具等。

由于 MOT16、MOT17 数据集相较于 MOT15 数据集环境场景更丰富，包含

更多的视频序列和目标实例，且现有的优秀算法常用其进行测试比较。而 MOT20 数据集在规模上相对较小，可能限制模型在大规模多目标跟踪任务上的性能表现，尤其是在需要进行长期预测或推断的任务中。其仍然局限于城市街道场景，与 MOT16 和 MOT17 数据集相比，它的场景多样性可能较少，这可能导致模型在更广泛的场景中泛化能力较弱。所以本文选择在 MOT16、MOT17 数据集上训练测试。MOT16、MOT17 数据集部分数据如图 2-8 所示。

图 2-8 MOT16、MOT17 数据集部分数据示意图

2.5.2 多目标跟踪评价指标

为了使模型的评价更加客观准确，本文采用的多目标跟踪评估指标[70]包括：

（1）多目标跟踪准确率(Multi-object Tracking Accuracy, MOTA)

MOTA 用于确定目标的个数，以及目标的相关属性方面的准确率，能够统计跟踪过程中的误差积累情况，包括漏检目标数量、误检目标数量、目标 ID 的切换次数。其方程如式（2-9）所示：

$$MOTA = 1 - \frac{N_{FN} + N_{FP} + N_{IDs}}{N_{GT}} \tag{2-9}$$

其中，N_{FN} 表示整个视频序列的漏检总数，N_{FP} 表示整个视频序列的误检总数，N_{IDs} 表示行人 ID 切换的总次数，N_{GT} 表示真实边界框数量。

（2）多目标跟踪精度（Multi-object Tracking Precision, MOTP）

MOTP 能够确定目标位置上的精确度，用于衡量目标位置确定的精确程度。其方程如式（2-10）所示：

$$MOTP = 1 - \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \tag{2-10}$$

其中，t 表示当前视频帧为第 t 帧，d_t^i 表示第 t 帧中第 i 个预测框与真实框的重叠率，c_t 表示匹配成功的对象个数。

（3）跟踪器 ID 维持能力（Identification F1 Score, IDF1）

IDF1 是通过系统正确识别的检测目标数量与真实目标数和计算检测目标的平均数量之比来表示的，通过这些数量之比平衡了识别精度和召回率。其方程如式（2-11）所示：

$$IDF1 = \frac{2 \times IDTP}{2 \times IDTP + IDFP + IDFN} \tag{2-11}$$

其中，*IDTP* 表示真阳性的 ID 总数，*IDFP* 表示假阳性的 ID 总数，*IDFN* 表示假阴性的 ID 总数。

（4）行人 ID 切换次数（ID Switch, IDs）

IDs 表示多目标跟踪过程中目标 ID 的切换次数。该指标越低，跟踪效果越好。

（5）多数跟踪目标百分比（Mostly Tracked, MT）

MT 表示多目标跟踪过程中高于 80%时间都能成功匹配的轨迹占比。该指标越高，跟踪效果越好。

（6）多数丢失目标百分比（Mostly Lost, ML）

ML 表示多目标跟踪过程中低于20%时间成功匹配的轨迹占比。该指标越低，跟踪效果越好。

除了上述评价指标以外，通常还会采用帧率指标 FPS（算法每秒处理视频的帧数）来反映算法的运行速度。

2.6 本章小节

本章节阐述了多目标跟踪算法相关的理论。首先描述了卷积神经网络的基础理论；其次从有无锚框的角度阐述了目标检测算法；然后分析了多目标跟踪的基本结构，并描述了卡尔曼滤波与匈牙利匹配算法；最后说明了多目标跟踪的数据集和评价指标。本章节的相关理论是多目标跟踪的基础，为后续的实验和分析打下了基础。

第 3 章 融合注意力机制的行人多目标跟踪算法

在遮挡环境下对多个动态目标进行实时跟踪是一个比较困难的问题，设计一个准确率高、实时性好的跟踪器是当下研究的热点和关键。针对多目标跟踪过程中目标遮挡问题，本章节提出了一种融合注意力机制的行人多目标跟踪算法。首先在主干网络中采用特征金字塔注意力模块扩大感受野，获得更加丰富的特征信息，以提高对不同尺度目标的检测精度；其次在行人重识别分支网络中融合了一个尺寸感知模块，对来自不同分辨率的语义特征进行融合，提取更加基础的行人特征，从而提高跟踪的精度；最后重新构造检测头，融合小目标检测层，使得所提网络适应不同尺寸的目标。

3.1 注意力机制

注意力机制[71]是深度学习中一种常用的技术，其模拟了人类在处理信息时的注意力分配方式，使模型能够在处理输入序列时更加关注与当前任务相关的部分，从而提高模型性能。它被广泛地应用于对象检测[72]、图像增强[73]、目标跟踪[74]、图像分割[75]等领域。

在深度学习中，注意力机制的核心理念在于使模型能够在处理输入序列的每个元素时，动态分配不同位置元素的权重。这样一来，模型在学习过程中能够更加专注于与当前任务相关的信息，而不是简单地对所有信息进行均等考虑，这使得模型能够在输入序列中发现并强调更为重要的部分。在图像处理中，也存在一些注意力机制的变种，用于处理图像中不同区域的信息。例如，多头注意力机制是一种常见的注意力机制变体，它通过多个注意力头来学习多个不同的注意力权重。这些权重然后被组合起来，以更全面地捕捉输入序列的信息。

总体而言，注意力机制赋予深度学习模型更灵活的处理序列数据的能力，使得模型能够更好地适应不同长度和不同重要性的输入序列。

3.2 整体算法结构

联合式多目标跟踪算法包含三个部分：主干网络（Backbone）、颈部（Neck）和头部（Head），其中主干网络的作用是对跟踪目标的特征进行初步提取，颈部用于提取更复杂的跟踪目标特征，头部预测跟踪目标的种类和位置，以及进行跟踪目标 ID 的嵌入，总体网络结构如图 3-1 所示。本章节从主干网络和头部对联

合式多目标跟踪算法进行改进，提出了一种融合注意力机制的行人多目标跟踪网络。由于坐标注意机制可以融合位置信息和通道信息，在避免大量计算的同时更准确地定位和识别目标。因此，本章节在主干网络中采用特征金字塔注意力模块扩大感受野，并引入坐标注意力机制。因为尺寸感知模块可以提高 ID 嵌入的关联能力，防止跟踪目标特征语义层的错位，因此将尺寸感知模块集成到头部的 ReID 分支网络中来处理目标遮挡问题。在实际应用场景中，小目标只占几个像素的大小，这很容易导致漏检。为了提高对小目标的检测能力，在头部的检测分支网络部分，对 YOLOv5s 检测头进行改进，增加小目标检测层，使得整体网络对不同尺寸的目标均有更好的效果。

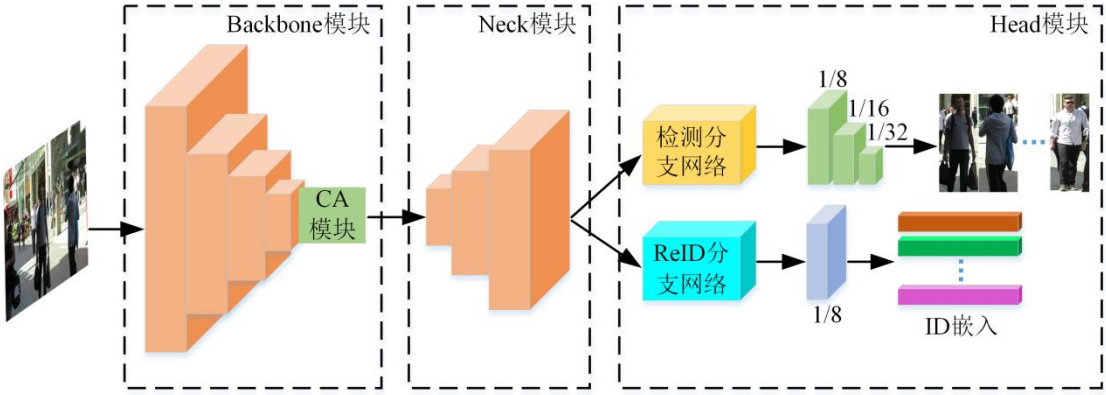

图 3-1 总体网络结构示意图

3.2.1 主干网络

联合式多目标跟踪算法跟踪速度快，但在遮挡的情况下主干网络对跟踪目标的特征提取不充分，故其跟踪精度不高。CA 注意力机制能够同时关注位置特征和通道特征，使得主干网络能够关注更大的区域，大大降低了计算量。为了提高主干网络特征提取的能力，本章节在主干网络上融合 CA 注意力机制和 FPN 网络架构，使其能够提取更多有效的跟踪目标特征，从而提高跟踪精度。

主干网络结构示意图如图 3-2 所示，其工作原理如下：输入的图像由主干网络经过 FPN 网络架构，得到 C×H×W 的特征图 F，在宽度和高度两个方向上分别对其进行全局平均池化，得到在宽度和高度两个方向上的特征图 F_1、F_2。将特征图 F_1 和 F_2 拼接在一起，经过 1×1 的卷积核进行降维，对降维后的特征图进行批量归一化，并用非线性激活函数对批量归一化后的特征图进行优化，得到特征图 F′。将特征图 F′ 在宽度和高度两个方向上分别进行卷积核为 1×1 的卷积运算，

得到与原始特征图F通道数相同的特征图F_h和F_w。将F_h和F_w分别经过Sigmoid激活函数进行优化，获得初始输入特征图F在高度方向上的注意力权重G_h和在宽度方向上的注意力权重G_w。将注意力权重G_h和G_w与初始特征图F进行乘法加权计算，得到带有注意力权重的特征图，以获得更多准确有效的跟踪目标特征。

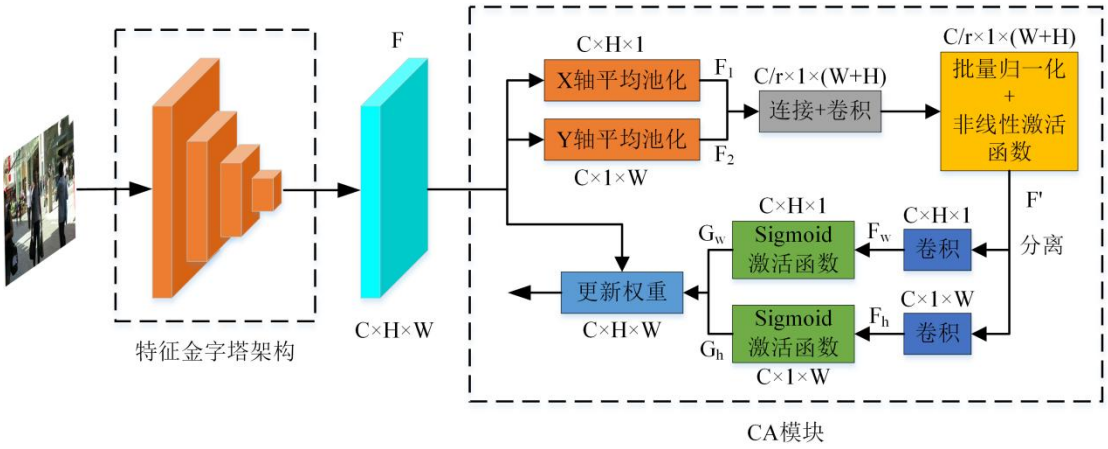

图 3-2 主干网络结构示意图

3.2.2 ReID 分支网络

为了提高ID嵌入的关联能力，解决行人在跟踪过程中间歇性消失的问题，如行人遮挡或短暂消失，本章节在头部的ReID分支网络部分融合了尺寸感知模块。

尺寸感知模块包含空间注意力子模块和通道注意力子模块。空间注意力子模块可以增强跟踪目标相关特征和抑制背景噪声，使得ReID网络能够在不同的分辨率下更好地提取跟踪目标的有效特征。通道注意力子模块对每个通道的权重进行分配，增大有效通道权重，减少无效通道权重，从而使得ReID网络能够更多地关注跟踪目标所在的关键区域。空间注意力子模块的具体实现方式如式（3-1）所示：

$$M_s(F)=\sigma\left(f^{7\times7}\left(\left[AvgPool(F);MaxPool(F)\right]\right)\right) \tag{3-1}$$

其中，F表示输入特征，$M_s(F)$表示空间注意力特征图，$AvgPool$表示平均池化操作，$MaxPool$表示最大池化操作，$f^{7\times7}$表示7×7卷积操作，σ表示Sigmoid激活函数。

通道注意力子模块的具体实现方式如式（3-2）所示：

$$M_c(F)=\sigma\left(MLP\left(AvgPool\left(F\right)\right)+MLP\left(MaxPool\left(F\right)\right)\right) \quad (3\text{-}2)$$

其中，F 表示输入特征，$M_c(F)$ 表示通道注意力特征图，*AvgPool* 表示平均池化操作，*MaxPool* 表示最大池化操作，*MLP* 表示多层感知机，σ 表示 Sigmoid 激活函数。

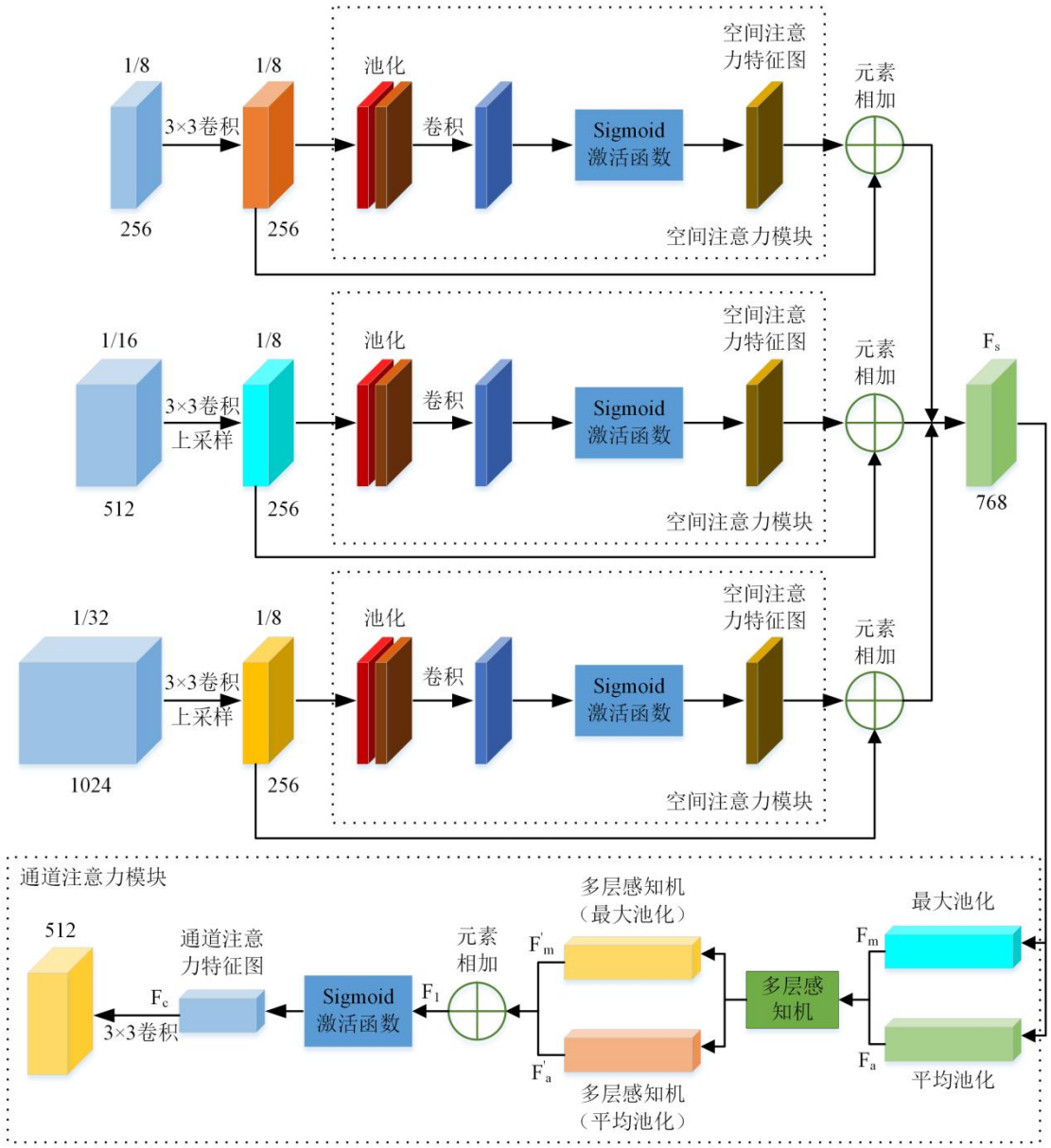

图 3-3 尺寸感知模块结构示意图

尺寸感知模块结构示意图如图 3-3 所示，其工作原理如下：首先将不同尺度的跟踪目标特征图上采样到 1/8，然后通过 3×3 卷积层进行卷积运算，再经过空间注意力子模块的最大池化和平均池化、7×7 卷积运算、Sigmoid 层归一化处理，最后得到空间注意力图。将空间注意力图与最初上采样到 1/8 的特征图进行融合，

以防目标特征信息的丢失。将不同尺度的空间注意力图拼接起来，得到特征图F_s。利用通道注意力子模块对特征图F_s分别进行全局平均池化和最大池化处理，获得处理后的特征图F_a和F_m，然后将特征图F_a和F_m通过一个一维卷积层和一个全连接层进行运算，生成特征图F'_a和F'_m，再将F'_a和F'_m进行相加求和，得到输出特征图F_1，将F_1经过 Sigmoid 层进行归一化处理，生成只有一个通道的特征图F_c。最后，特征图F_c通过一个 3×3 卷积层进行卷积运算，再映射到 512 个信道，为后续的 ReID 任务提供身份信息。

本章节使用 ReID 分支网络来获取目标的外观和结构特征信息。将提取的目标特征表示为特征向量，计算目标之间的相似度或距离来衡量其相似性。ReID 分支网络可以将当前检测到的目标特征与已知的目标特征进行匹配，判断其是否属于同一目标。

3.2.3 检测分支网络

联合式多目标跟踪算法的检测分支网络用于预测跟踪目标的种类和位置，常见的检测分支网络有 YOLO 系列、CenterNet 等，但这些检测分支网络在检测小目标时提取的目标特征信息不足，会出现检测不到小目标的情况。为了提高检测网络对小目标的检测能力，本章节在检测分支网络部分对 YOLOv5s 检测头进行改进。由于 YOLOv5s 检测头最小能检测到分辨率为 8×8 的目标，容易漏检分辨率小于 8×8 的小目标，在其三个检测层的基础上添加了 4 倍下采样的小目标检测层，可以检测到 4×4 分辨率的小目标。改进后的检测分支网络结构如图 3-4 所示：将主干网络提取的特征图进行 4 倍下采样，得到特征图F_2。新增一个上采样模块，用于放大特征图，然后将 4 倍下采样得到的特征图F_2与上采样后的特征图F_1进行拼接，获得特征图F，防止跟踪目标特征信息丢失。将拼接后得到的特征图F经过一个轻量级的卷积模块（Conv3 模块）进一步进行卷积运算，获得特征图FF，用于小目标检测。Conv3 模块由三个 3×3 的深度可分离卷积层组成，在每个卷积层之间引入批量归一化层和 Leaky ReLU 激活函数层，其中第一个卷积层的步幅为 2，可以将特征图的尺寸减半，第二个和第三个卷积层的步幅均为 1，可以获得更丰富的跟踪目标特征信息。最终使用改进后的四个检测层来执行检测任务。

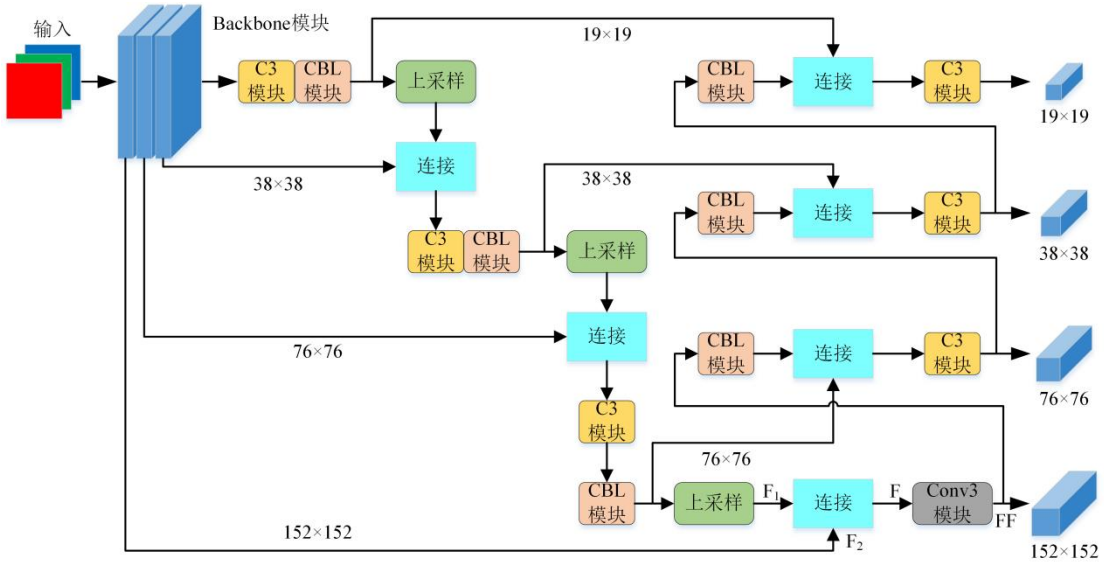

图 3-4 改进后的检测分支网络结构示意图

3.3 实验结果与分析

3.3.1 实验环境

实验环境基于 Win10 操作系统，NVIDIA RTX 2060 显卡，在 Python3.7 版本的服务器条件下，采用 Pytorch1.9.0 版本的深度学习框架实现。在本章模型训练中，对模型进行 30 个 epoch 训练，设置初始学习率为 0.0005，并在 20 个 epoch 后将学习率调整为 0.00005，模型单次训练的数据批次大小为 8。

3.3.2 消融实验

为测试本章节所提算法中不同模块的作用，分别对具有不同模块的网络进行实验：只有原始数据模块的网络（Raw），融合 CA 注意力机制模块的网络（CA），有原始数据模块、CA 注意力机制模块和尺寸感知模块的网络（Scale），本章节所提出的完整网络（Ours）。具有不同模块的网络在 MOT16 数据集上的部分评价指标对比结果如表 3-1 所示，在 MOT17 数据集上的部分评价指标对比结果如表 3-2 所示。

表 3-1 显示在 MOT16 数据集上，融合了多个模块的网络比未添加模块的网络的跟踪性能有了明显的提升，MOTA 提高了 10.7%，IDF1 提高了 15.9%，ID 切换次数明显减少，MT 提高了 10.6%，ML 降低了 6.1%。与只有原始数据模块的网络相比，由于 CA 注意力机制可以使网络更加关注重点区域，所以融合 CA 注意力机制模块的网络 MOTA 和 IDF1 指标分别提高了 3.8%、5.6%。因为尺寸

感知模块可以提高 ID 嵌入的相关能力，防止跟踪目标特征语义层的错位，所以有原始数据模块、CA 注意力机制模块和尺寸感知模块的网络相较于只有原始数据模块的网络，MOTA 和 IDF1 指标分别提高了 6.5%、11.3%。由于本章节所提完整网络在上述网络的基础上增加了小目标检测层，所以 MOTA 和 IDF1 指标相比于只有原始数据模块的网络分别提高了 10.7%、15.9%。

表 3-1 各个模块在 MOT16 数据集上的对比实验

Method	MOTA	IDF1	IDs	MT	ML
Raw	64.7%	56.3%	1855	34.8%	21.2%
CA	68.5%	61.9%	1229	38.3%	19.6%
Scale	71.2%	67.6%	1064	41.9%	17.4%
Ours	75.4%	72.2%	986	45.4%	15.1%

表 3-2 显示在 MOT17 数据集上，添加了多个模块的网络比原始网络的跟踪性能也有了明显的提高。总体来说，MOTA 提高了 11.5%，IDF1 提高了 15.8%，MT 提高了 11.0%，ML 降低了 7.3%，且 ID 切换次数降低。与原始网络相比，融合原始数据模块和 CA 注意力机制模块的准确率提高了 4.3%，融合原始数据模块、CA 注意力机制模块和尺寸感知模块的准确率提高了 7.5%。

表 3-2 各个模块在 MOT17 数据集上的对比实验

Method	MOTA	IDF1	IDs	MT	ML
Raw	62.8%	54.5%	5226	33.1%	23.9%
CA	67.1%	60.2%	4643	37.6%	20.1%
Scale	70.3%	65.7%	3976	40.4%	18.5%
Ours	74.3%	70.3%	3575	44.1%	16.6%

本实验结果表明，CA 注意力机制和所提的尺寸感知模块以及融合的小目标检测层都对提升算法的跟踪效果有较大的影响。在实际应用中，可以提高多目标跟踪的精度，减少 ID 切换次数。无论移除哪个模块，多目标跟踪精度都会降低。然而，由于实际场景的复杂性，它会影响结果的通用性和适用性，需要进一步的研究。

3.3.3 与现有方法对比实验

本章节在 MOT16 数据集和 MOT17 数据集上进行实验，并与其他最先进的多目标跟踪器进行比较。比较结果如表 3-3 和表 3-4 所示，成功率曲线如图 3-5 和图3-6所示。从图中可以看出本章节所提算法在遮挡情况下的成功率表现良好，且明显比其他方法更具有优势。

表 3-3 与其他先进方法在 MOT16 数据集上的对比结果

Method	MOTA	MOTP	IDF1	IDs	MT	ML	FPS
JDE[36]	64.4%	-	55.8%	1544	35.4%	20.0%	18.8
FairMOT[40]	74.9%	-	72.8%	1074	44.7%	15.9%	25.9
MOT_GM[76]	64.5%	76.7%	70.9%	816	36.4%	20.7%	6.5
CRF_RNN[77]	50.3%	74.8%	54.4%	702	18.3%	35.7%	1.5
LMOT[43]	73.2%	-	72.3%	669	44.0%	17.0%	29.6
Ours	75.4%	81.6%	72.2%	986	45.4%	15.1%	25.8

表 3-4 与其他先进方法在 MOT17 数据集上的对比结果

Method	MOTA	MOTP	IDF1	IDs	MT	ML	FPS
JDE[36]	63.0%	-	59.5%	6171	35.7%	17.3%	18.8
FairMOT[40]	73.7%	81.3%	72.3%	3303	43.2%	17.3%	25.9
MOT_ES[78]	73.3%	-	71.8%	3372	41.1%	17.2%	19.0
CRF_RNN[77]	53.1%	76.1%	53.7%	2518	24.2%	30.7%	1.4
LMOT[43]	72.0%	-	70.3%	3071	45.4%	17.3%	29.6
Ours	74.3%	81.7%	70.3%	3575	44.1%	16.6%	25.8

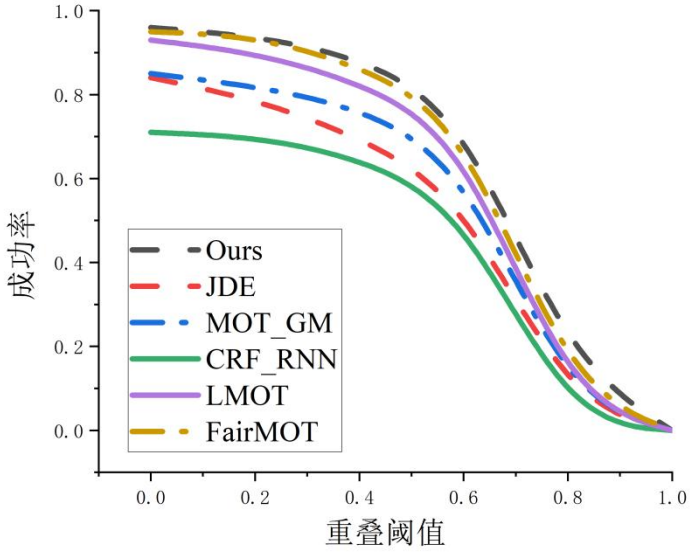

图 3-5 在 MOT16 数据集上的成功率曲线图

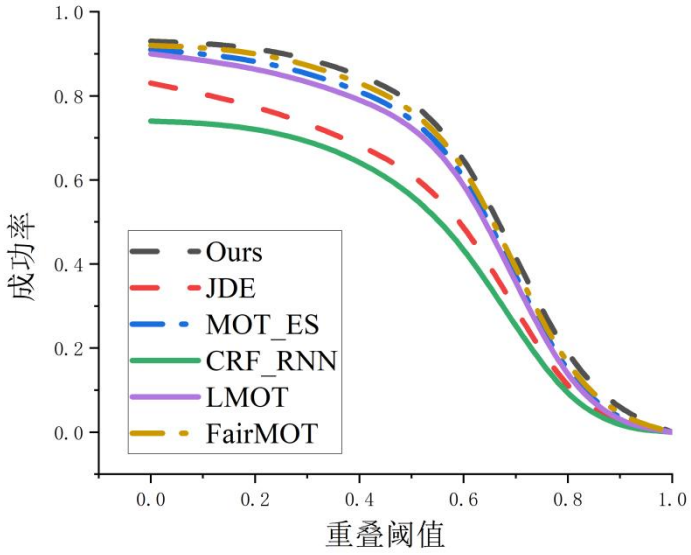

图 3-6 在 MOT17 数据集上的成功率曲线图

基于 2.5.2 节评价指标分析，所提算法与其他先进方法相比具有一定的优势（“-”表示原始论文中未给出数据）。在 MOT16 数据集上，由于本章节算法融合了 CA 注意力机制模块、尺寸感知模块和小目标检测模块，可以融合目标位置信息和通道信息，在避免大量计算的同时更准确地定位和识别目标，并提高 ID 嵌入的关联能力，以及增强网络对小目标的检测能力，所以与 JDE、FairMOT、MOT_GM、CRF_RNN、LMOT 方法相比，多目标跟踪的准确率分别提高了 11.0%、0.5%、10.9%、25.1%、2.2%。与 MOT_GM 和 CRF_RNN 方法相比，IDF1 分别

提高了 1.3%和 17.8%，多目标跟踪精度分别提高了 4.9%和 6.8%。与最优方法 FairMOT 相比，ID 切换次数减少，MT 增加了 0.7%，ML 减少了 0.8%。

在 MOT17 数据集上，与 JDE、FairMOT、MOT_ES、CRF_RNN 和 LMOT 方法相比，多目标跟踪准确率分别提高了 11.3%、0.6%、1.0%、21.2%和 2.3%。与 CRF_RNN 方法相比，IDF1 提高了 16.6%，MOTP 提高了 5.6%。与最佳方法 FairMOT 相比，MOTP 增加了 0.4%，MT 提高了 0.9%，ML 降低了 0.7%。由于网络中添加了 CA 注意机制，并添加了小目标检测头，所以跟踪速度略低，但也满足实时性要求。

综上所述，可以验证出本章节所提算法具有良好的性能，且在实际跟踪场景中具有一定的优势。多目标跟踪效果图如图 3-7 和图 3-8 所示。

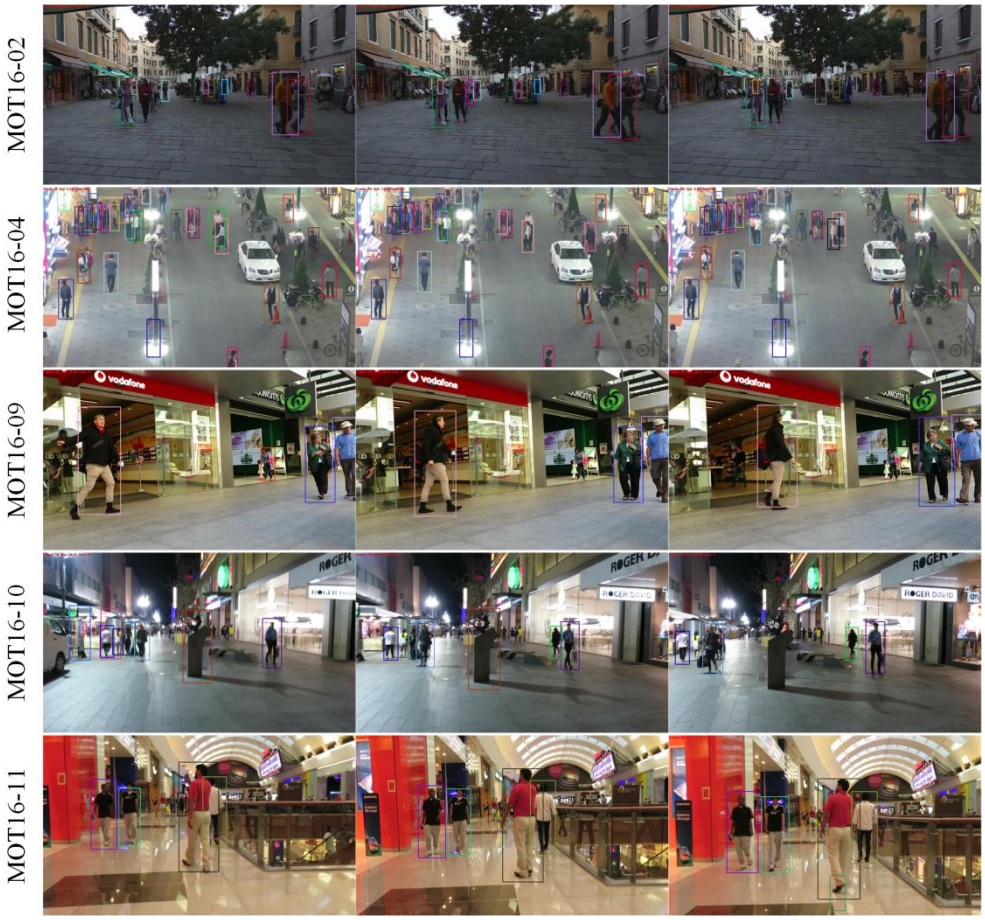

图 3-7 在 MOT16 数据集上的多目标跟踪效果图

图 3-8 在 MOT17 数据集上的多目标跟踪效果图

3.4 本章小节

本章节提出了一种融合注意力机制的行人多目标跟踪算法。所提出的算法包括三个部分的改进：利用坐标注意机制在主干网络中提取充分的跟踪目标特征，有效提高了检测网络在遮挡情况下的检测精度；尺寸感知模块用于融合不同分辨率的目标特征，提高 ID 嵌入的相关性，有效提高了遮挡情况下的跟踪精度；本章节还增加了一个小目标检测层，以增强对小目标的检测能力。实验表明，所提算法在一定程度上解决了行人多目标跟踪中的目标遮挡问题。与其他先进方法相比，该算法可以有效提高多目标跟踪准确率，跟踪速度也满足实时性要求。

第 4 章 基于无锚框检测的行人多目标跟踪算法

基于无锚框的目标检测方法能够更灵活地学习目标的位置和形状，而无需依赖于预设的锚框，这种方法不仅提高了检测的准确性，还降低了设计模型的难度。本章节从无锚框目标检测的角度出发，针对多目标跟踪过程中目标 ID 频繁切换的问题，在第三章研究内容的基础上提出了一种基于无锚框检测的行人多目标跟踪算法。该算法首先在主干网络部分融合深层聚合注意力模块以获得多层次的目标特征，从而扩大感受野，获取全局特征信息；其次采用基于无锚框的 RepPoints 检测头，降低多目标跟踪过程中目标身份标签的切换次数；最后在数据关联部分采用扩展 IOU 匹配，增加检测和轨迹的匹配范围，补偿匹配时的运动估计偏差，从而提高复杂环境下跟踪的鲁棒性。

4.1 无锚框行人检测

无锚框行人检测是指在目标检测任务中，不使用预定义的锚框来生成候选框，而是直接通过网络的学习来预测目标位置。传统的目标检测方法通常使用锚框作为候选框的参考，在图像上放置一组预定义的锚框，然后通过网络预测这些锚框的偏移量和目标分类信息。由于现有的联合式多目标跟踪器大多是直接从基于锚框的目标检测器修改而来的，所以它们是基于锚框的跟踪器。然而，基于锚框的设计并不适合学习目标的 ReID 特征，尽管其检测结果良好，但会导致目标的 ID 切换次数增加，具体体现在以下三点：

（1）忽略 ReID 任务

基于锚框的跟踪器通常首先估计目标的候选框，然后从候选框中提取目标相应的 ReID 特征，所以目标 ReID 特征的质量在很大程度上取决于训练期间候选框的质量。因此，在训练阶段，基于锚框的跟踪器严重偏向于估计目标的候选框，而不是高质量的目标 ReID 特征。这使得大多现有的联合式多目标跟踪器在训练的过程中对 ReID 任务“不公平”。

（2）一个锚框可能对应多个目标

基于锚框的方法通常使用 ROI-Align 方法从目标候选框中提取特征，而 ROI-Align 方法中的大多数采样位置可能属于其他干扰目标或背景，其示意图如图 4-1（a）所示。因此，基于锚框的跟踪器在准确区分目标方面略显不足。

（3）多个锚框对应一个目标

基于锚框的跟踪器在训练过程中，对应于不同目标的多个相邻锚框可能会对应同一个目标，这会导致训练的严重错误，其示意图如图 4-1（b）所示。在目标检测中，为了平衡检测精度和检测速度，特征映射通常被下采样 8、16、32 次。这对于目标检测来说是有利的，但是不利于学习目标 ReID 特征，因为在锚框中提取的特征可能不会与目标的中心对齐。

（a）一个检测框对应多个目标　　（b）多个检测框对应一个目标

图 4-1 锚框问题示意图

无锚框行人检测采用更灵活的方式，通过网络自动学习目标的位置和形状，而不依赖于预定义的锚框。这样的方法通常使用卷积神经网络或其他深度学习架构，通过在整个图像上进行密集的预测，直接输出目标的位置和分类信息。其一般步骤包括：选择适当的深度学习网络架构，如 Faster R-CNN 等；收集和准备用于训练的数据集，包含带有标注的行人图像。标注信息应该包括每个行人的边界框和相应的类别标签；使用准备好的数据集训练深度学习模型。通过最小化目标检测任务的损失函数，使得模型能够准确地预测行人的位置和类别；对模型输出进行一些后处理步骤，例如非极大值抑制，以去除冗余的检测框并提高检测结果的准确性。

4.2 整体算法架构

本章节是在第三章研究内容的基础上进行的，提出了一种基于无锚框检测的行人多目标跟踪算法，其整体算法结构如图 4-2 所示。该算法采用 ResNet-34[79] 作为主干网络，并在主干网络上融合深层聚合注意力模块，以获得多层次的目标特征，从而扩大感受野并获取全局特征信息。由于 RepPoints 检测头能够以一种限制目标空间范围的方式自动排列自己，并且不需要使用锚框来对目标边界框的空间进行采样，所以本章节采用 RepPoints 检测头作为检测分支网络，降低多目标跟踪过程中目标身份标签的切换次数，从而提高多目标跟踪的准确率。在目标

跟踪的过程中，如果目标经历了过快的不规则运动，导致在相邻帧之间位置没有重叠，这种情况会使得目标运动特征的度量失效。因此，本章节在数据关联部分采用扩展 IOU 匹配，增加检测和轨迹的匹配范围，从而提高复杂环境下跟踪的鲁棒性。

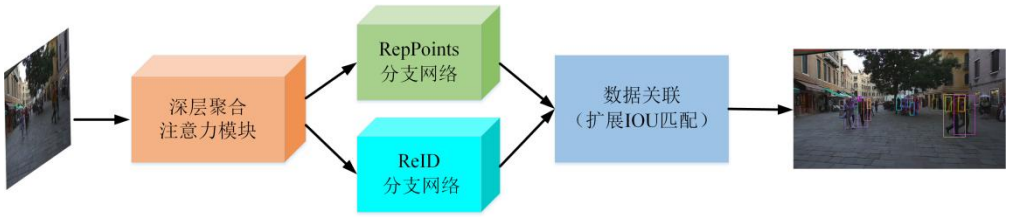

图 4-2 整体算法框架示意图

4.2.1 深层聚合注意力模块

在主干网络部分，为了更加充分的提取跟踪目标的特征，本章节算法在主干网络上融合深层聚合注意力模块，以融合多层特征，扩大感受野，从而获取目标全局特征信息。相比于第三章节的特征金字塔模块，深层聚合注意力模块在低层和高层特征之间有更多的跳越连接，提取的目标特征更为充分，能够更好的传播梯度，使得目标的深层特征和浅层特征更充分的融合。此外，将所有上采样模块中的普通卷积都替换成可变形卷积，从而可以根据目标的尺度和姿态调整感受野的大小。

深层聚合注意力模块的结构示意图如图 4-3 所示，其工作原理如下：将不同尺寸的跟踪目标特征进行多次上采样和下采样操作，充分融合深层和浅层的跟踪目标特征，得到目标特征图 FF 。特征图 FF 分别进行最大池化和平均池化操作，获得最大池化特征图 F_m 和平均池化特征图 F_a 。池化特征图 F_m 和 F_a 分别经过第一个卷积层、ReLU 激活函数层、第二个卷积层，得到特征图 F_{mp} 和 F_{ap} 。将得到的特征图 F_{mp} 和 F_{ap} 相加后输入 Sigmoid 激活函数层，获得特征图 F 。将特征图 F 和原始的特征图 FF 元素相乘，获得特征图 F′，避免跟踪目标特征信息丢失。在通道维度对特征图 F′ 分别进行最大池化和平均池化操作，获得特征图 F_{cm} 和 F_{ca} 。特征图 F_{cm} 和 F_{ca} 经过拼接操作后进行卷积运算和 Sigmoid 激活函数处理，生成特征图 F_s 。最后将特征图 F_s 和特征图 F′ 相乘，获得特征图 FF_1 ，防止目标特征信息丢失，获取充分的跟踪目标特征。

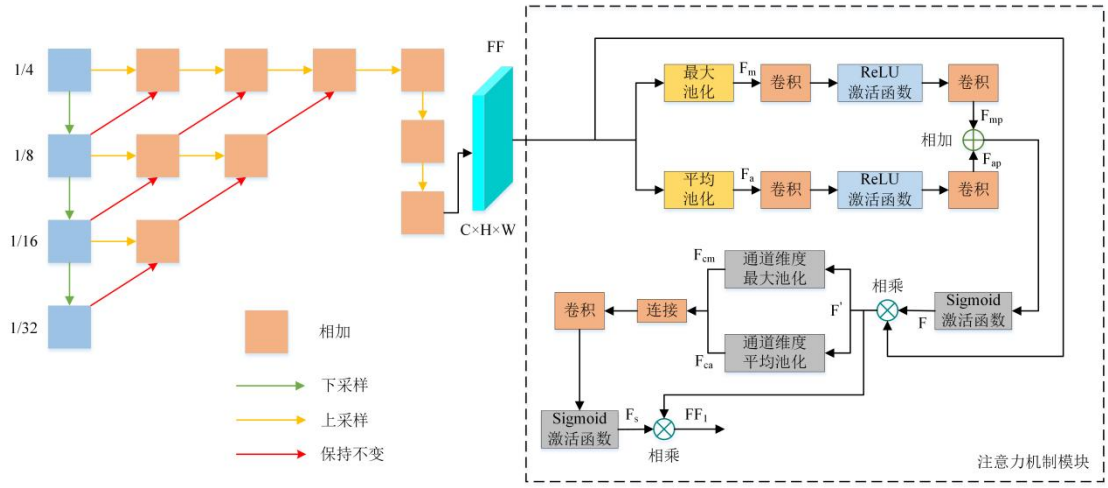

图 4-3 深层聚合注意力模块的结构示意图

4.2.2 RepPoints 检测头

本文第三章所提的一种融合注意力机制的行人多目标跟踪算法，其检测分支网络采用的是基于锚框的 YOLOv5s 检测头。虽然改善了对小目标的检测性能，但基于锚框的方法预设的尺寸和比例，可能不适应目标的多样性，并且目标的不规则形状也难以覆盖。在目标遮挡的场景下，目标处理困难，容易出现误检、漏检的情况，从而导致目标 ID 切换频繁，跟踪准确率降低。

针对以上问题，本章节在检测分支网络部分，采用基于无锚框的 RepPoints 检测头。RepPoints 检测头用一组代表点表示，它具有提供更精确定位和分类目标的能力。与其他目标检测方法不同，RepPoints 采用自下而上的构建方式，通过一组代表点来限制目标空间范围并关注语义上重要的局部区域。这种方法允许 RepPoints 自适应地定位目标，实现准确的目标分类，而无需使用锚框在边界框范围内进行目标特征的采样。与现有的非矩形表示方法不同，RepPoints 通过识别目标的各个点（例如边界框的角点或目标的末端点），实现了更灵活的目标检测方式。其对代表点组建模的方程如式（4-1）所示：

$$R=\left\{\left(x_k,y_k\right)\right\}_{k=1}^{n} \tag{4-1}$$

其中，R 为这组代表点坐标的集合，$\left(x_k,y_k\right)$ 为其中一个代表点坐标，n 为这组代表点总数，n 默认设置为 9。

进一步细化式（4-1）可表示为式（4-2）：

$$R_r=\left\{\left(x_k+\Delta x_k,y_k+\Delta y_k\right)\right\}_{k=1}^{n} \tag{4-2}$$

其中，R_r 表示这组代表点坐标的集合，(x_k, y_k) 表示旧代表点的坐标，$(\Delta x_k, \Delta y_k)$ 表示新代表点对于旧代表点的预测偏移值，n 为这组代表点总数，n 默认设置为 9。

本章节算法的检测分支网络结构示意图如图 4-4 所示，其采用两个非共享的子网络，分别用于生成 RepPoints 代表点（定位）和目标分类。定位子网络首先经过三个 3×3 的卷积层，然后通过两个连续的分支小网络来计算两组 RepPoints 的偏移值。与此同时，分类子网络也应用了三个 3×3 的卷积层，然后经过一个 3×3 可变形卷积层，其输入偏移量与定位子网络中输入第一个可变形卷积层的偏移量相同。在两个子网络的前三个 3×3 卷积层中，还引入了归一化层以提高模型的性能。通过这些设计和策略，可以实现高效而精准的目标检测性能。

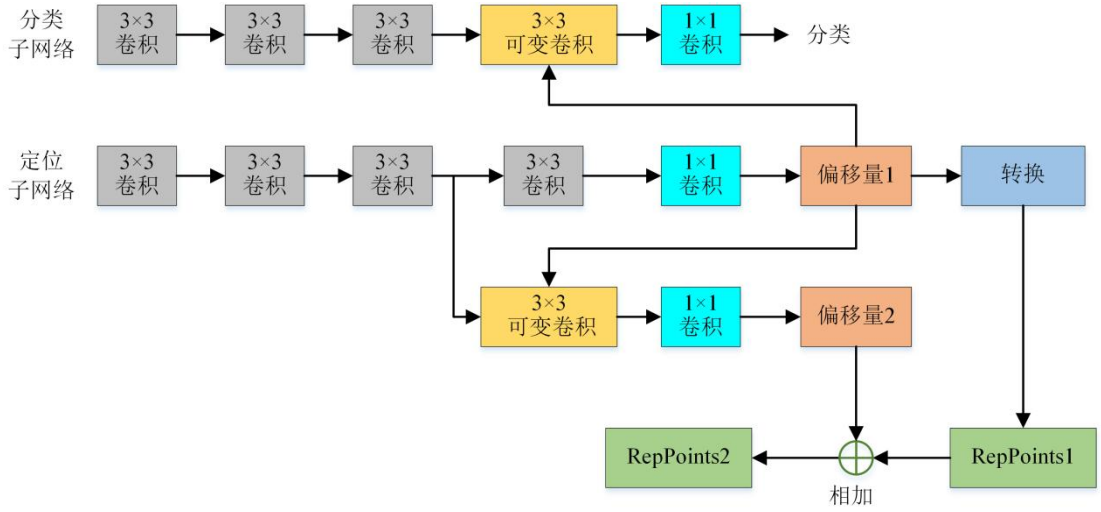

图 4-4 检测分支网络结构示意图

4.2.3 扩展 IOU 匹配

在当前多目标跟踪算法中，数据关联算法主要依赖于目标的表征信息、运动信息和位置信息等特征来建立跟踪目标的轨迹。然而，由于目标遮挡、光照变化、尺度变化以及外观模糊等因素的存在，导致目标的检测效果较为有限。此外，在跟踪过程中，目标可能经历快速且不规则的运动，导致相邻帧之间的目标位置没有重叠，使得多目标跟踪准确率下降。为了解决上述问题，本章节算法采用扩展 IOU 进行数据关联匹配。通过引入扩展 IOU，能够更灵活地考虑目标之间的形状和外观变化，从而提高数据关联的鲁棒性。扩展 IOU 的使用可以降低对目标外观特征的依赖，使得跟踪算法在目标遮挡、光照变化等场景下表现更为稳定。

在复杂环境下，当目标呈现不规则运动时，相邻视频帧中的目标位置未发生

重叠，这种情况下就不适用于采用传统的 IOU 匹配方法，传统的 IOU 匹配原理如图 4-5 所示。为了缓解这一问题，本章节算法采用扩展 IOU 匹配的概念。扩展 IOU 匹配通过扩大 IOU 的匹配面积，拓展检测和轨迹匹配的范围，从而直接匹配相邻视频帧中同一目标不重叠部分，其方程如式（4-3）所示：

$$IOU = \frac{A \cap B}{A \cup B} \tag{4-3}$$

其中，A 表示 A 目标特征图，B 表示 B 目标特征图。

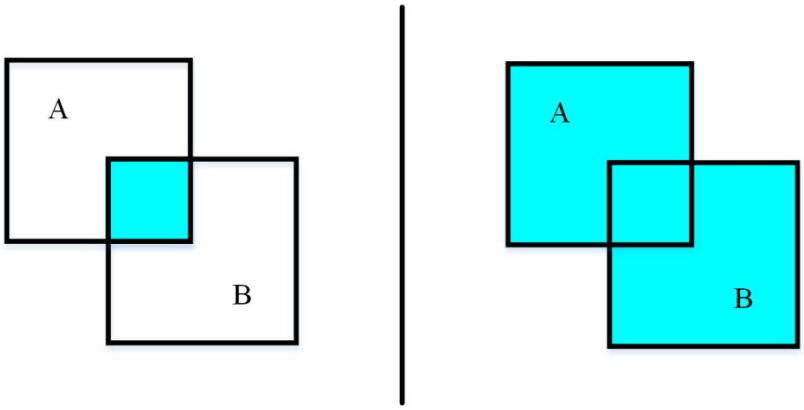

图 4-5 传统 IOU 匹配示意图

扩展 IOU 匹配能够在匹配空间中更灵活地适应目标的不规则运动，同时减少了对目标外观特征的依赖性，从而有效缓解不规则运动对跟踪性能的影响，其原理示意图如图 4-6 所示。

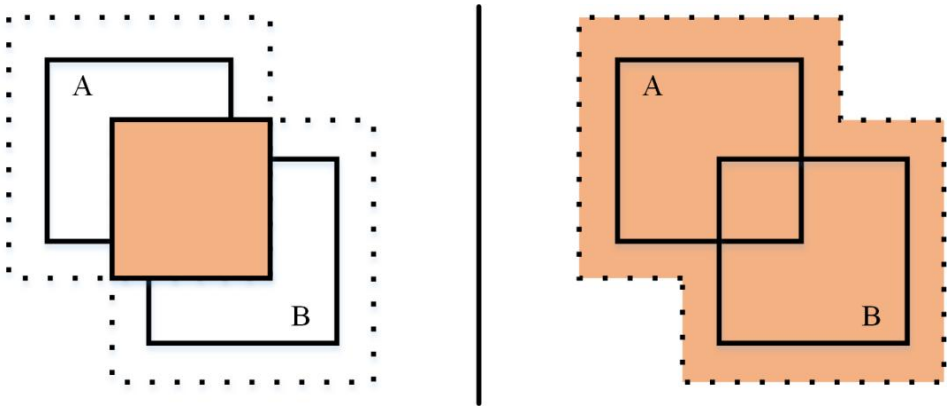

图 4-6 扩展 IOU 匹配示意图

4.3 算法流程

本章节所提的一种基于无锚框检测的行人多目标跟踪算法与本文第三章所提算法基于的跟踪框架一样，都是采用联合式多目标跟踪范式，其具体算法流程图如图 4-7 所示。

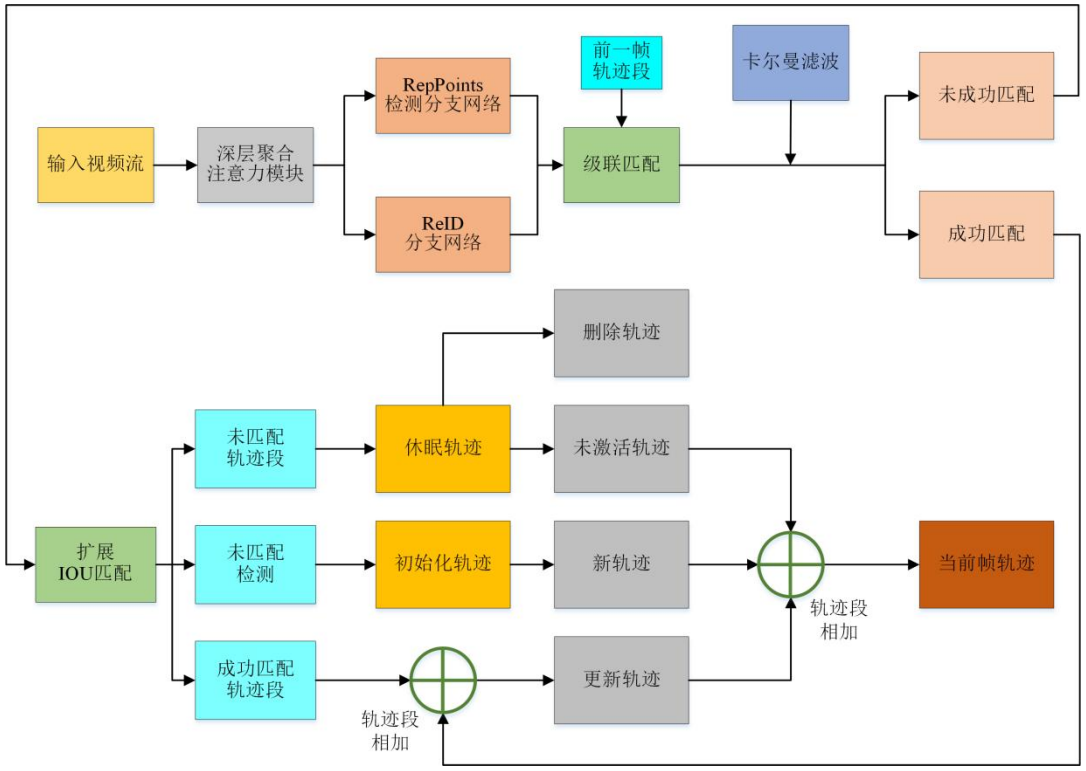

图 4-7 具体算法流程图

具体算法流程如下：

第一步：输入的视频序列经过深层聚合注意力模块，进行主干特征提取。由 ResNet-34 主干网前向传递，分别得到 1/4、1/8、1/16 和 1/32 四个尺寸的特征图。然后对这四个尺寸的特征图进行上采样和下采样操作，并在其之间进行跳跃连接。本章节算法在此处融合了一个注意力机制模块，使得提取的特征更加准确、算法更加优化。

第二步：将深层聚合模块提取出来的特征图经过后续的 RepPoints 检测分支网络和 ReID 检测分支网络做进一步处理。本章节算法采用基于无锚框的 RepPoints 检测分支网络，使算法能够更好的检测目标，降低跟踪过程中目标 ID 的切换次数。ReID 分支网络则采用第三章所提算法的尺寸感知模块，用来提取深度外观特征信息，提高 ID 嵌入的关联能力，增强算法对目标尺度变化的鲁棒性。

第三步：数据关联匹配。首先计算候选框和模板之间 ID 嵌入的相似性，然后利用卡尔曼滤波器进行距离限制，消除了候选框与现有轨迹之间不合理的匹配。采用匈牙利算法寻找最优匹配，未匹配的轨迹段将通过扩展 IOU 匹配再次进行检查。对于匹配成功的轨迹，将更新模板以适应目标外观变化。对于不匹配的候

选框将被初始化为新的轨迹，其中相应的 ID 嵌入同时被初始化为新的轨迹模板。未匹配的轨迹段将被设置为休眠状态。如果一个轨迹段在 30 帧内不能找到它匹配的候选框，就将该轨迹段删除。然后未匹配的轨迹段将被设置为激活状态，当它成功地与一个候选框匹配后终止。成功匹配的轨迹段将被用于在下一帧与候选框连接。

第四步：重复上述步骤直至跟踪结束。

4.4 实验结果与分析

4.4.1 实验环境配置

实验环境基于 Win10 操作系统，NVIDIA RTX 2060 显卡，在 Python3.7 版本的服务器条件下，采用 Pytorch1.9.0 版本的深度学习框架实现。在本章模型训练中，对模型进行 30 个 epoch 训练，设置初始学习率为 0.0005，并在 20 个 epoch 后将学习率调整为 0.00005，模型单次训练的数据批次大小为 8。

4.4.2 消融实验

为测试本章节所提算法中不同模块的作用，分别对具有不同模块的网络进行实验：只有原始数据模块的网络（Original），融合深层聚合注意力模块的网络（DLA-Attention），有原始数据模块、深层聚合注意力模块和 RepPoints 检测头的网络(Rep)，本章节所提出的完整网络（Ours）。具有不同模块的网络在 MOT16 数据集上的部分评价指标对比结果如表 4-1 所示，在 MOT17 数据集上的部分评价指标对比结果如表 4-2 所示。

表 4-1 显示在 MOT16 数据集上，融合了多个模块的网络比未添加模块的网络的跟踪性能有了明显的提升，MOTA 提高了 11.2%，IDF1 提高了 17.1%，ID 切换次数明显减少，MT 提高了 11.0%，ML 降低了 6.9%。与只有原始数据模块的网络相比，由于深层聚合注意力模块能够获得多层次的目标特征，获取全局特征信息，所以融合深层聚合注意力模块的网络 MOTA 和 IDF1 指标分别提高了 5.1%、6.2%。因为 RepPoints 检测头可以自适应地定位目标，实现准确的目标分类，而无需使用锚框在边界框范围内进行目标特征的采样，所以有原始数据模块、深层聚合注意力模块和 RepPoints 检测头的网络相较于只有原始数据模块的网络，MOTA 和 IDF1 指标分别提高了 7.4%、13.2%。由于本章节所提完整网络在上述

网络的基础上采用了扩展 IOU 匹配，提高了数据关联的鲁棒性，所以 MOTA 和 IDF1 指标相比于只有原始数据模块的网络分别提高了 11.2%、17.1%。

表 4-1 各个模块在 MOT16 数据集上的对比实验

Method	MOTA	IDF1	IDs	MT	ML
Original	64.7%	56.3%	1855	34.8%	21.2%
DLA-Attention	69.8%	62.5%	1504	39.5%	18.2%
Rep	72.1%	69.5%	1059	42.1%	16.4%
Ours	75.9%	73.4%	871	45.8%	14.3%

表 4-2 显示在 MOT17 数据集上，添加了多个模块的网络比原始网络的跟踪性能也有了明显的提高。总体来说，MOTA 提高了 11.9%，IDF1 提高了 17.0%，MT 提高了 11.8%，ML 降低了 8.1%，且 ID 切换次数大幅度降低。与原始网络相比，融合原始数据模块和深层聚合注意力模块的准确率提高了 5.4%，IDF1 提高了 6.2%，融合原始数据模块、深层聚合注意力模块和 RepPoints 检测头的准确率提高了 8.5%，IDF1 提高了 12.9%。

表 4-2 各个模块在 MOT17 数据集上的对比实验

Method	MOTA	IDF1	IDs	MT	ML
Original	62.8%	54.5%	5226	33.1%	23.9%
DLA-Attention	68.2%	60.7%	4475	38.3%	19.2%
Rep	71.3%	67.4%	3476	41.2%	17.7%
Ours	74.7%	71.5%	2896	44.9%	15.8%

本实验结果表明，深层聚合注意力模块和 RepPoints 检测头以及数据关联部分的扩展 IOU 匹配都对提升算法的跟踪效果有较大的影响。在实际应用中，本章节所提算法可以提高多目标跟踪的准确率，减少 ID 切换次数。无论移除哪个模块，多目标跟踪准确率都会降低。然而，由于实际场景的复杂性，这会影响结果的通用性和适用性，需要更进一步的探讨研究。

4.4.3 与现有方法对比实验

本章节在 MOT16 数据集和 MOT17 数据集上进行实验，并与其他最先进的

多目标跟踪器进行比较。比较结果如表 4-3 和表 4-4 所示，成功率曲线如图 4-8 和图 4-9 所示。从图中可以看出本章节所提算法的成功率表现良好，且明显比其他方法更具有优势。

表 4-3 与其他先进方法在 MOT16 数据集上的对比结果

Method	MOTA	MOTP	IDF1	IDs	MT	ML	FPS
JDE[36]	64.4%	-	55.8%	1544	35.4%	20.0%	18.8
FairMOT[40]	74.9%	-	72.8%	1074	44.7%	15.9%	25.9
MOT_GM[76]	64.5%	76.7%	70.9%	816	36.4%	20.7%	6.5
CRF_RNN[77]	50.3%	74.8%	54.4%	702	18.3%	35.7%	1.5
LMOT[43]	73.2%	-	72.3%	669	44.0%	17.0%	29.6
Ours1	75.4%	81.6%	72.2%	986	45.4%	15.1%	25.8
Ours2	75.9%	82.5%	73.4%	871	45.8%	14.3%	25.7

表 4-4 与其他先进方法在 MOT17 数据集上的对比结果

Method	MOTA	MOTP	IDF1	IDs	MT	ML	FPS
JDE[36]	63.0%	-	59.5%	6171	35.7%	17.3%	18.8
FairMOT[40]	73.7%	81.3%	72.3%	3303	43.2%	17.3%	25.9
MOT_ES[78]	73.3%	-	71.8%	3372	41.1%	17.2%	19.0
CRF_RNN[77]	53.1%	76.1%	53.7%	2518	24.2%	30.7%	1.4
LMOT[43]	72.0%	-	70.3%	3071	45.4%	17.3%	29.6
Ours1	74.3%	81.7%	70.3%	3575	44.1%	16.6%	25.8
Ours2	74.7%	82.3%	71.5%	2896	44.9%	15.8%	25.7

基于 2.5.2 节评价指标分析，所提算法与其他先进方法相比具有一定的优势（“-”表示原始论文中未给出数据，“Ours1”表示本文第三章节所提算法，“Ours2 表示本章节所提算法”）。在 MOT16 数据集上，由于本章节算法融合了深层聚合注意力模块，采用了 RepPoints 检测头和扩展 IOU 匹配，可以使目标的深层特征和浅层特征更充分的融合，并且不需要使用锚框来对目标边界框的空间进行采

样，以及能够增加检测和轨迹的匹配范围，所以与 JDE、FairMOT、MOT_GM、CRF_RNN、LMOT 方法相比，多目标跟踪的准确率分别提高了 11.5%、1.0%、11.4%、25.6%、2.7%。与 MOT_GM 和 CRF_RNN 方法相比，多目标跟踪精度分别提高了 5.8%和 7.7%，IDF1 分别提高了 2.5%和 19.0%。与 FairMOT 方法相比，ID 切换次数减少，MT 增加了 1.1%，ML 减少了 1.6%。与本文第三章节所提算法相比，MOTA 提高了 0.5%，MOTP 提高了 0.9%，IDF1 提升了 1.2%，IDs 略微降低，MT 增加了 0.4%，ML 降低了 0.8%。

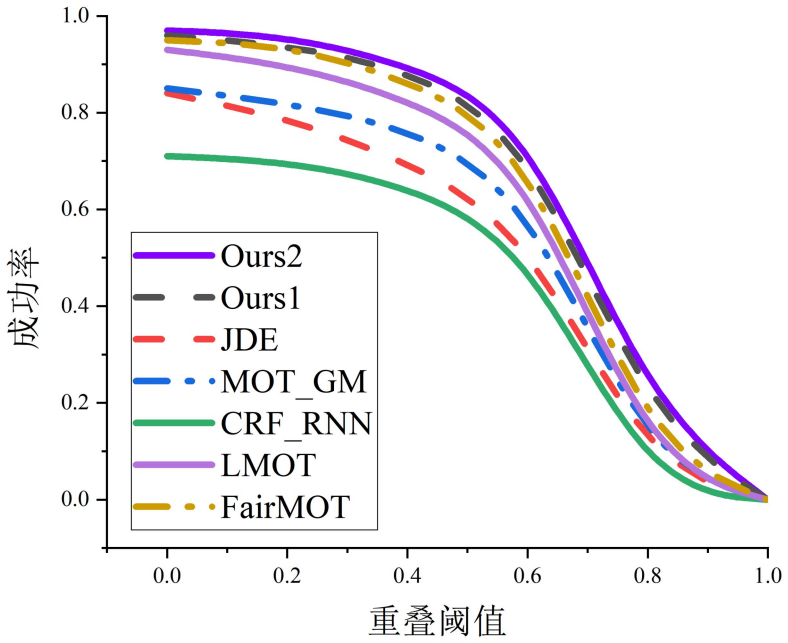

图 4-8 在 MOT16 数据集上的成功率曲线图

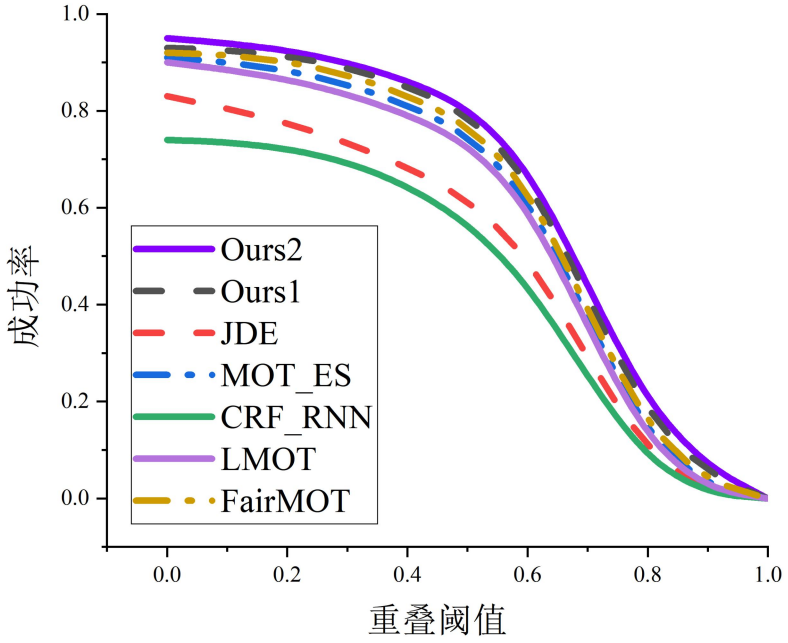

图 4-9 在 MOT17 数据集上的成功率曲线图

在 MOT17 数据集上，与 JDE、FairMOT、MOT_ES、CRF_RNN 和 LMOT 方法相比，多目标跟踪准确率分别提高了 11.7%、1.0%、1.4%、21.6%和 2.7%。与 CRF_RNN 方法相比，IDF1 提高了 12.0%和 17.8%，MOTP 提高了 6.2%。与最佳方法 FairMOT 相比，MOTP 增加了 1.0%，MT 提高了 1.7%，ML 降低了 1.5%。与本文第三章节所提算法相比，MOTA 提高了 0.4%，MOTP 提高了 0.6%，IDF1 提升了 1.2%，IDs 降低，MT 增加了 0.8%，ML 降低了 0.8%。由于本章节算法在主干网络部分融合更多跳跃连接的深层聚合网络，并添加了一个注意力机制，所以跟踪速度略低，但也满足实时性要求。

综上所述，可以验证出本章节所提算法具有良好的性能，且在实际跟踪场景中具有一定的优势。多目标跟踪效果图如图 4-10 和图 4-11 所示。

图 4-10 在 MOT16 数据集上的多目标跟踪效果图

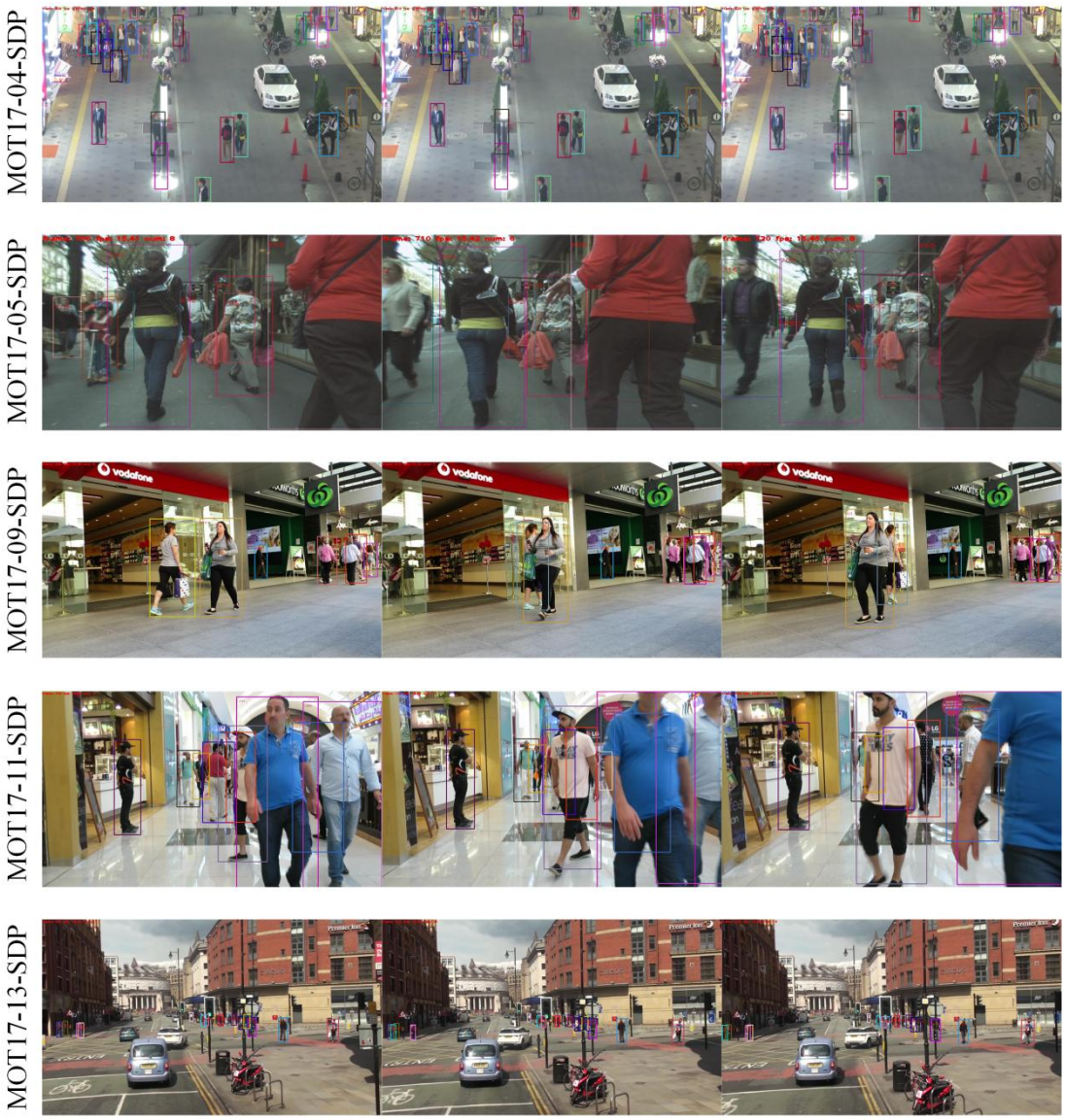

图 4-11 在 MOT17 数据集上的多目标跟踪效果图

4.5 本章小节

本章节在第三章研究内容的基础上提出了一种基于无锚框检测的行人多目标跟踪算法。该算法首先对主干网络进行改进，融合深层聚合注意力模块，获取多层次的目标特征，扩大感受野，从而获得全局的目标特征信息。其次，采用基于无锚框的 RepPoints 检测头，以减少在多目标跟踪中目标身份标签切换的频率。最后，在数据关联方面，采用了扩展 IOU 匹配方法，拓展了检测和轨迹匹配的范围，弥补匹配时可能出现的运动估计偏差，从而提高在复杂环境下的跟踪鲁棒性。实验表明，本章节所提算法具有良好的跟踪性能，在一定程度上解决了多目标跟踪过程中目标 ID 频繁切换的问题，在实际场景应用中具有一定的优势。

第 5 章 总结与展望

5.1 全文总结

本文采用联合式多目标跟踪算法作为基本框架，通过融合注意力机制和无锚框检测跟踪两个方向的研究，探讨影响跟踪器性能的因素。从不同角度对算法进行优化，目的是为了提升多目标跟踪算法的整体性能，同时保持稳定的跟踪速度。

本文的具体研究内容如下：

（1）对目标检测算法和多目标跟踪算法进行描述，分析了多目标跟踪算法的基本结构、跟踪难点，以及相关理论。在联合式多目标跟踪算法的基础上，从不同的角度改善多目标跟踪算法的性能和鲁棒性。

（2）在联合式多目标跟踪算法框架上，首先采用特征金字塔注意力模块融合主干网络，以扩大感受野，获得更加丰富的特征信息，提高对不同尺度目标的检测精度；其次在行人重识别分支网络中融合了一个尺寸感知模块，对来自不同分辨率的语义特征进行融合，提取更加基础的行人特征，从而提高跟踪的准确率；最后重新构造检测头，融合小目标检测层，使得所提网络适应不同尺寸的目标。

（3）在研究内容（2）的基础上，首先引入融合深层聚合注意力模块融合主干网络，以获得多层次的目标特征，从而扩大感受野，获取全局特征信息；其次采用基于无锚框的 RepPoints 检测头，降低多目标跟踪过程中目标身份标签的切换次数；最后在数据关联部分采用扩展 IOU 匹配，增加检测和轨迹的匹配范围，补偿匹配时的运动估计偏差，从而提高复杂环境下跟踪的鲁棒性。

5.2 研究展望

本文叙述了多目标跟踪领域的研究成果和理论，并在联合式多目标跟踪算法的框架上提出了两种改进算法。尽管这两种改进算法取得了一定的效果，但也存在一些局限性。未来的研究方向将在当前基础上展开，进一步深入探讨相关问题。

（1）多目标跟踪任务包含多个子任务，如目标检测、数据关联和运动预测等。在计算资源有限的情况下，如何在这些子任务之间实现更加公平的平衡资源，成为确保整个系统高效运行的关键因素。使模型能够均衡地关注每个子任务的性能表现，是多目标跟踪任务仍然需要解决的问题。

（2）目标在运动过程中可能受到多方面变化的影响，如光照、视角和姿态

等，导致目标在不同时间点呈现出明显的外观差异。这种外观的不断变化增加了跟踪算法在处理多目标时的复杂性。为解决这一问题，需要采用更为创新性的方法和技术，以提高算法对目标外观变化的适应性。

（3）在应对目标长时间的遮挡或目标消失后再次出现在视野中的情况下，如何维持跟踪算法的稳定性仍然是一个重大难题。这可能涉及到新的检测和关联策略，以及对长时间目标缺失时轨迹重建的深入研究，未来需要在此问题上进一步的研究探讨。

参考文献

[1] Ahmed I, Jeon G Piccialli F. From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where[J]. In IEEE Transactions on Industrial Informatics, 2022, 18(8): 5031-5042.

[2] Yu H, Wang Y, Tian Y, et al. Social Vision for Intelligent Vehicles: From Computer Vision to Foundation Vision[J]. In IEEE Transactions on Intelligent Vehicles, 2023, 8(11): 4474-4476.

[3] Luo W, Xing J, Milan, A. Multiple object tracking: A literature review[J]. IEEE Trans. Artif. Intell. 2021, 293: 58-80.

[4] Zou H, Wang J. A Survey of Deep Learning Target Detection Algorithms under Candidate Regions[C]. 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA), 2022: 670-673.

[5] Huang M, Liu F, Yao A, et al. Review of Hybrid Model Used in SAR Target Recognition[C]. 2020 2nd International Conference on Information Technology and Computer Application (ITCA), 2020: 746-750.

[6] Gioele C, Francisco L, Siham T, et al. Deep learning in video multi-object tracking: A survey[J]. Neurocomputing, 2020, 381(2): 61-88.

[7] 伍瀚, 聂佳浩, 张照娓, 等. 基于深度学习的视觉多目标跟踪研究综述[J]. 计算机科学, 2023, 50(04): 77-87.

[8] 刘文强, 裘杭萍, 李航, 等. 深度在线多目标跟踪算法综述[J]. 计算机科学与探索, 2022, 16(12): 2718-2733.

[9] Zhang Y, Zheng X. Development of Image Processing Based on Deep Learning Algorithm[C]. 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2022: 1226-1228.

[10] Zhang X, Liu C, Suen C. Towards Robust Pattern Recognition: A Review[J]. In Proceedings of the IEEE, 2020, 108(6): 894-922.

[11] Mohanapriya D, Mahesh K. Multi object tracking using gradient-based learning model in video-surveillance[J]. In China Communications, 2021, 18(10): 169-180.

[12] Ravindran R, Santora M, Jamali M. Multi-Object Detection and Tracking, Based on DNN, for Autonomous Vehicles: A Review[J]. In IEEE Sensors Journal, 2021, 21(5): 5668-5677.

[13]郑佳慧, 俞晓迪, 赵生妹, 等. 基于均值滤波的关联成像去噪[J]. 光学学报, 2022, 42(22): 49-56.

[14]Kalman R E. A New Approach To Linear Filtering and Prediction Problems[J]. Journal of Basic Engineering, 1960, 82: 35-45.

[15]贺冰, 王法胜, 王星等. 显著性感知三重正则化相关滤波无人机目标跟踪算法[J]. 北京航空航天大学学报, 2023: 1-16.

[16]Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]. 2010 IEEE computer society conference on computer vision and pattern recognition, 2010: 2544-2550.

[17]Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]. European conference on computer vision, 2012: 702- 715.

[18]Henriques J F, Caseiro R, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(3): 583-596.

[19]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.

[20]Bochinski E, Eiselein V, Sikora T. High-Speed tracking-by-detection without using image information[C]. 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017: 1-6.

[21]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.

[22]Yu F, Li W, Li Q, et al. POI: Multiple Object Tracking with High Performance Detection and Appearance Feature[C]. European Conference on Computer Vision (ECCV), 2016: 36-42.

[23]Zhang J, Wang N, Zhang L. Multi-Shot Pedestrian Re-Identification via Sequential Decision Making[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 6781-6789.

[24]Zhou Z, Xing J, Zhang M, et al. Online Multi-Target Tracking with Tensor-Based High-Order Graph Matching[C]. 2018 24th International Conference on Pattern Recognition (ICPR), 2018: 1809-1814.

[25]Fang K, Xiang Y, Li X, et al. Recurrent Autoregressive Networks for Online

Multi-object Tracking[C]. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018: 466-475.

[26] 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, 2017, 39(6): 1137-1149.

[27] Yang F, Choi W, Lin Y. Exploit All the Layers: Fast and Accurate CNN Object Detector With Scale Dependent Pooling and Cascaded Rejection Classifiers[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 2129-2137.

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

[29] Braso G, Leal-Taixe L. Learning a Neural Solver for Multiple Object Tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 6247-6257.

[30] Chen J, Xi Z, Wei C, et al. Multiple Object Tracking Using Edge Multi-Channel Gradient Model With ORB Feature[J]. In IEEE Access, 2021, 9: 2294-2309.

[31] Yu F, Liu Q. Pedestrian Multi-Objective Tracking Based on Work-Yolo[C]. 2022 IEEE 8th International Conference on Cloud Computing and Intelligent Systems (CCIS), 2022: 66-72.

[32] Li X X, Liu Z J, Hu C F, et al. Light-Weight Multi-Target Detection and Tracking Algorithm Based on M3-YOLOv5[C]. 2023 42nd Chinese Control Conference (CCC), 2023: 8159-8164.

[33] Xiao T, Li S, Wang B, et al. Joint Detection and Identification Feature Learning for Person Search[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 3415-3424.

[34] He K, Gkioxari G, Dollar P, et al. Mask R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 2961-2969.

[35] Voigtlaender P, Krause M, Osep A, et al. MOTS: Multi-Object Tracking and Segmentation[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 7934-7943.

[36] Wang Z, Zheng L, Liu Y, et al. Towards Real-Time Multi-Object Tracking[C]. In Proceedings of the European Conference on Computer Vision (ECCV), 2020: 107-122.

[37] Lu Z, Rathod V, Votel R, et al. RetinaTrack: Online Single Stage Joint Detection and Tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020: 14668-14678.

[38] Lin T, Dollar P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 936-944.

[39] 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.

[40] 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, 2021, 129: 3069-3087.

[41] Liu X H, Luo Y C, Yan K D, et al. Part-MOT: A multi-object tracking method with instance part-based embedding[J]. IET Image Processing, 2021, 15(11): 2521-2531.

[42] Tian Z, Shen C H, Chen H, et al. FCOS: Fully Convolutional One-Stage Object Detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 9627-9636.

[43] Mostafa R, Baraka H, Bayoumi A. LMOT: Efficient Light-Weight Detection and Tracking in Crowds[J]. IEEE Access, 2022, 10: 83085-83095.

[44] Li G, Chen X, Li M, et al. One-shot multi-object tracking using CNN-based networks with spatial-channel attention mechanism[J]. Opt. Laser Technol, 2022, 153: 108267.

[45] Yan Y C, Li J P, Qin J, et al. Efficient Person Search: An Anchor-Free Approach[J]. International Journal of Computer Vision, 2023, 131: 1642-1661.

[46] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 13708-13717.

[47] Yang Z, Liu S, Hu H, et al. RepPoints: Point Set Representation for Object Detection[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 9657-9666.

[48] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series[J]. The handbook of brain theory and neural networks, 1995, 3361(10).

[49] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural

Networks for Mobile Vision Applications[J]. arXiv: 1704.04861, 2017.

[50]黄欣研, 刘芳, 鲍骞月, 等. 基于多任务学习和身份约束的生成对抗网络人脸校正识别方法[J]. 电子学报, 2023, 51(10): 2936-2949.

[51]王茂森, 鲍久圣, 鲍周洋, 等. 融合金字塔结构与注意力机制的煤矿井下巡检机器人 PT 目标检测算法[J]. 煤炭科学技术, 2024: 1-11.

[52]Yuan Q, Huang G, Zhong G, et al. Triangular Chain Closed-Loop Detection Network for Dense Pedestrian Detection[J]. In IEEE Transactions on Instrumentation and Measurement, 2024, 73: 1-14.

[53]Sun K, Li Z, Zheng Y, et al. An Area-Efficient Accelerator for Non-Maximum Suppression[J]. In IEEE Transactions on Circuits and Systems II: Express Briefs, 2023, 70(6): 2251-2255.

[54]Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 6517-6525.

[55]Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]. European Conference on Computer Vision (ECCV), 2016: 21-37.

[56]Lin T-Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[J]. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[57]Huang L, Yang Y, Deng Y, et al. DenseBox: Unifying Landmark Localization with End to End Object Detection[J]. arXiv: 1509.04874, 2015.

[58]Wang X G, Chen K B, Huang Z L, et al. Point Linking Network for Object Detection[J]. arXiv: 1706.03646, 2017.

[59]Law H, Deng J. CornerNet: Detecting Objects as Paired Keypoints[C]. European Conference on Computer Vision (ECCV), 2018: 765-781.

[60]Zhou X Y, Wang D Q, Krahenbuhl P. Objects as Points[J]. arXiv:1904.07850, 2019.

[61]Duan K, Bai S, Xie L, et al. CenterNet: Keypoint Triplets for Object Detection[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019: 6568-6577.

[62]Carion N, Massa F, Synnaeve G, et al. End-to-End Object Detection with Transformers[C]. European Conference on Computer Vision (ECCV), 2020: 213-229.

[63] Zhu X Z, Su W J, Lu L W, et al. Deformable DETR: Deformable Transformers for End-to-End Object Detection[J]. arXiv: 2010.04159, 2021.

[64] Dai X Y, Chen Y P, Yang J W, et al. Dynamic DETR: End-to-End Object Detection With Dynamic Attention[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021: 2968-2977.

[65] Zhang G J, Luo Z P, Yu Y C, et al. Accelerating DETR Convergence via Semantic-Aligned Matching[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 939-948.

[66] Li F, Zhang H, Liu S L, et al. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 13609-13617.

[67] Leal-Taixé L, Milan A, Reid I, et al. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking[J]. arXiv: 1504.01942, 2015.

[68] Milan A, Leal-Taixé L, Reid I, et al. MOT16: A Benchmark for Multi-Object Tracking[J]. arXiv:1603.00831, 2016.

[69] Dendorfer P, Rezatofighi H, Milan A, et al. MOT20: A benchmark for multi object tracking in crowded scenes[J]. arXiv: 2003.09003, 2020.

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

[71] Guo M H, Xu T X, Liu J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media, 2022, 8: 331-368.

[72] Aziz L, Sheikh U U, Ayub S, et al. Exploring Deep Learning-Based Architecture, Strategies, Applications and Current Trends in Generic Object Detection: A Comprehensive Review[J]. IEEE Access, 2020, 8: 170461-170495.

[73] Ye W, Yan T, Gao J, et al. LFIENet: Light Field Image Enhancement Network by Fusing Exposures of LF-DSLR Image Pairs[J]. In IEEE Transactions on Computational Imaging, 2023, 9: 620-635.

[74] Ondrašovič M, Tarábek P. Siamese Visual Object Tracking: A Survey[J]. IEEE Access, 2021, 9: 110149-110172.

[75] Minaee S, Boykov Y, Porikli F, et al. Image Segmentation Using Deep Learning: A Survey[J]. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(7): 3523-3542.

[76] Yoo Y S, Lee S H, Bae S H. Effective Multi-Object Tracking via Global Object

Models and Object Constraint Learning[J]. Sensors, 2022, 22(20): 7943.

[77] Xiang J, Xu G H, Ma C, et al. End-to-End Learning Deep CRF Models for Multi-Object Tracking Deep CRF Models[J]. In IEEE Transactions on Circuits and Systems for Video Technology, 2021, 31(1): 275-288.

[78] Xiang X Z, Ren W K, Qiu Y J, et al. Multi-object Tracking Method Based on Efficient Channel Attention and Switchable Atrous Convolution[J]. Neural Processing Letters, 2021, 53: 2747–2763.

[79] He K, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770-778.

攻读硕士期间发表文章

[1] Liu C, Han C. Research on Pedestrian Multi-Object Tracking Network Based on Multi-Order Semantic Fusion[J]. World Electric Vehicle Journal, 2023, 14(10): 272.

致 谢

逝者如斯，不舍昼夜，春去春又来，岁月稍纵即逝。此时，回头想想这段短暂的求学路，时而喜悦，时而惆怅。在这个美丽的校园里，感谢命运的安排，让我有幸结识了许多良师益友，是他们教我如何品味人生，让我懂得如何更好的生活！人生处处是驿站，已是挥手作别之时，在此，向所有帮助过我的人献上我最诚挚的谢意！

首先，谨向我尊敬的导师韩超教授致以诚挚的谢意和崇高的敬意。非常幸运能够成为您的学生，在这短短的三年里，聆听着您孜孜不倦的教诲，感受着您严谨进取的治学精神和乐观向上的生活态度，我不仅体会到知识与研究的魅力，也学会了许多做人的道理。毕业在即，在此谨向您表示我最衷心的感谢，同时祝您工作顺利，合家欢乐，身体健康，一切安好！另外，还要感谢我学生时代所遇到的每一位老师，各位老师道德与学术并重，宽容博大的胸襟、谦逊朴素的为人，令我如沐春风，倍感温馨。永远难忘老师们在个人人生观、世界观上的引领和指导。数载教诲，师恩难报，我在这里对各位老师鞠躬致谢！

其次，请容许我向我最爱的家人表示诚挚的谢意，想到他们，我总是感到温暖而安详。正是因为有你们的鼓励和支持，才有了今天的我。你们的哺育之恩，爱护之情让我永生难忘。感谢所有关心我，爱护我的亲人，祝福你们身体健康，万事如意！

再次，我要感谢同学们对我无私的帮助，同窗之谊，我将终生难忘。虽然相聚在一起的时间不长，但我们的友谊地久天长。在毕业之际，我祝你们有幸福的明天，美好的未来！

同时，还要特别感谢陪伴在我身边的女朋友小王老师。每当遇到挫折和困难时，你都是我坚实的支柱，为我排忧解难，帮我解开心结。感激你一直以来的支持和理解，期许我们能够一起经历未来人生中更多的精彩，共同见证生命中的更多美好瞬间！

最后，我要向从百忙之中抽出时间对本文进行审阅、评议和参与本人论文答辩的各位老师表示诚挚的感谢。

尚德教学 唯实惟新