分类号：TP391.41
密　级：公开

学校代码：10363
学　　号：2220910103

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目 复杂环境下动态小目标检测与跟踪关键技术研究

论 文 作 者　张紫阳
校 内 导 师　曹雏清（高级工程师）
校 外 导 师　陈智君
专业学位类别　计算机技术
研究方向（领域）　智能系统与应用

论文提交日期：2025 年 5 月 9 日

分类号：TP391.41　　单位代码：　10363
密　级：公开　　学　　号：2220910103

题　目　复杂环境下动态小目标检测与跟踪关键技术研究

英文并列

题　目　Research on Key Technologies of Dynamic Small Object Detection and Tracking in Complex Environments

学 生 姓 名：张紫阳
校 内 导 师：曹雏清（高级工程师）
校 外 导 师：陈智君
专业学位类别：计算机技术
研 究 方 向：智能系统与应用

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：张紫阳

日期：2025 年 5月 9 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：张崇阳　　指导教师签名：

日期：2025年 5月 9日　　日期：2025年 5月 9日

复杂环境下动态小目标检测与跟踪关键技术研究

摘 要

目标检测与跟踪技术在自动驾驶、安防监控和无人机巡检等领域取得了广泛应用。然而，复杂环境下动态小目标检测与跟踪仍然存在较大的挑战。小目标由于在图像中占比低，特征信息稀疏，易受背景干扰、运动模糊与部分遮挡影响，传统手工设计的检测模型存在特征提取能力不足、计算冗余度高等问题的同时，而且现有跟踪方法在遮挡等复杂场景下常出现身份漂移和轨迹断裂等问题。因此设计出具有强大特征提取能力和高效跟踪性能的目标检测与跟踪算法具有重要的现实应用意义。为此本文对复杂环境下动态小目标检测与跟踪技术展开研究，具体内容如下：

首先，针对传统目标检测方法中难以对小目标细微特征进行有效捕获的问题，本文提出一种基于神经架构搜索的小目标检测算法，采用了双边约束采样策略的神经架构搜索技术来自动寻找最优的主干架构，并确保了超网训练过程中各子网均能得到公平训练，并聚合各子网梯度更新超网权重，以提升主干网络的特征提取能力，最终在性能与复杂度之间达到理想平衡。在此基础上，将搜索到的主干网络替换 Faster R-CNN 原有的特征提取器，并与特征金字塔网络结合，以增强对不同尺度目标的检测效果。实验结果表明，该方案在显著降低模型参数量和计算量的同时，有效提升了小目标检测精度。

其次，针对现有跟踪方法在遮挡等复杂场景下常出现身份漂移和轨迹断裂等问题，本文提出基于线性注意力机制的多尺度特征融合跟踪算法。该方法利用检测器输出的位置信息和特征，通过具有引入线性注意力机制的 Cross-Transformer 将低层的细节信息与高层的语义

信息融合，从而大幅增强特征表达能力。结合相似性学习策略，算法能够更高效地进行目标匹配和轨迹关联，在处理遮挡、目标重叠及身份混淆等复杂场景时表现出更强的鲁棒性。实验在 MOT 和 VisDrone 数据集上验证了该方法在提高目标跟踪准确性和稳定性方面的显著优势。

总之，本文工作为复杂环境下动态小目标的检测与跟踪提供了一种高效、实用的解决方案，并为后续研究指明了新的发展方向。

关键词：小目标检测与目标跟踪，神经架构搜索，Faster R-CNN，特征融合，线性注意力

RESEARCH ON KEY TECHNOLOGIES OF DYNAMIC SMALL OBJECT DETECTION AND TRACKING IN COMPLEX ENVIRONMENTS

ABSTRACT

Object detection and tracking technology has been widely used in the fields of autonomous driving, security monitoring, and drone inspection. However, there are still great challenges in the detection and tracking of dynamic small objects in complex environments. Small objects account for a low proportion in the image, have sparse feature information, and are easily affected by background interference, motion blur, and partial occlusion. Traditional manually designed detection models have problems such as insufficient feature extraction capabilities and high computational redundancy. In addition, existing tracking methods often have problems such as identity drift and trajectory breakage in complex scenarios such as occlusion. Therefore, it is of great practical application significance to design a object detection and tracking algorithm with powerful feature extraction capabilities and efficient tracking performance. To this end, this paper studies the detection and tracking technology of dynamic small objects in complex environments. The specific contents are as follows:

First, in order to solve the problem that it is difficult to effectively capture the subtle features of small objects in traditional object

detection methods, this thesis proposes a small object detection algorithm based on neural architecture search. The neural architecture search technology with bilateral constrained sampling strategy is used to automatically find the optimal backbone architecture, ensure that each subnet can be trained fairly during the supernet training process, and aggregates the gradients of each subnet to update the supernet weights to improve the feature extraction ability of the backbone network, and finally achieve an ideal balance between performance and complexity. On this basis, the searched backbone network replaces the original feature extractor of Faster R-CNN and is combined with the feature pyramid network to enhance the detection effect of objects of different scales. Experimental results show that this scheme significantly reduces the number of model parameters and the amount of calculation while effectively improving the detection accuracy of small objects.

Secondly, in response to the problems of identity drift and trajectory breakage that often occur in existing tracking methods in complex scenes such as occlusion, this thesis proposes a multi-scale feature fusion tracking algorithm based on linear attention mechanism. This method uses the position information and features output by the detector, and fuses the low-level detail information with the high-level semantic information through the Cross-Transformer with the introduction of linear attention mechanism, thereby greatly enhancing the feature expression ability. Combined with the similarity learning strategy, the algorithm can more efficiently perform object matching and trajectory association, and show stronger robustness when dealing with complex scenes such as occlusion, object overlap and identity

confusion. Experiments on MOT and VisDrone datasets verify the significant advantages of this method in improving object tracking accuracy and stability.

In summary, this work provides an efficient and practical solution for the detection and tracking of dynamic small objects in complex environments, and points out a new development direction for subsequent research.

KEY WORDS: Small object detection and object tracking, Neural architecture search, Faster R-CNN , Feature fusion, Linear attention

目录

图录

表录

第1章 绪论

1.1 研究背景及意义

随着社会与科技的快速发展，深度学习在计算机视觉领域的应用愈发广泛，并不断推动目标检测和跟踪技术的革新与完善。这些技术已逐步渗透到我们生活的各个方面。例如，在医学影像中，高效而准确的目标检测能够迅速定位病灶区域[1-4]，从而降低漏诊和误诊风险，提高诊断质量；在自动驾驶领域，能够实时准确地检测与跟踪远距离或小尺寸处于运动中的行人和车辆[5-7]对于保障行车安全和优化驾驶决策至关重要；而在军事应用中，有效跟踪高速运动目标则是实现精确打击和快速反应的核心技术[8,9]。可见，目标检测与跟踪技术不仅仅是一项热门的研究方向，也关乎社会生活与国防安全，具有重要的现实意义。

尽管目标检测与跟踪技术已有了长足的发展与应用[10]，在众多实际应用场景中，尤其是复杂环境下的动态小目标检测与跟踪仍存在明显瓶颈和技术挑战。小目标通常在图像中所占的像素面积十分有限，加之可能处于高空视角下或者与背景、杂物混杂在一起，导致其外观特征模糊不清，目标检测算法主干网络对于边缘信息以及图像特征难以准确提取。而复杂环境通常包含复杂多变的背景信息，如光照变化、阴影干扰、物体遮挡及背景相似物的存在等因素。这些因素使得小目标的视觉特征进一步减弱或模糊，加剧了特征的识别难度，严重影响小目标检测与跟踪的准确性与鲁棒性。此外，随着深度网络的层数和宽度不断增加，并且为了提升小目标检测的精度，一些专门面向小目标检测的手工设计网络往往引入了更为复杂的结构模块[11-13]，试图通过扩大网络容量与表达能力来提升检测精度。然而，这使得参数量与计算复杂度的显著提高的，这不仅加剧了网络训练的难度和对硬件资源的依赖，也在一定程度上限制了算法在实时性或嵌入式部署方面的应用前景。同时在复杂环境往往伴随着小目标的高速运动或轨迹的不确定性，在拥挤的城市道路环境中，小目标的随机运动轨迹可能相互交织，增加了目标关联与身份确认的困难，其运动轨迹表现出很强的随机性和不确定性，极大地增加

了检测与跟踪算法对目标的持续感知难度，容易导致目标漏检、误检或跟踪失败。

因此，如何在保证网络轻量化与高效性的前提下，提高对动态小目标的检测与跟踪精度，是当前学术界和工业界共同关注的关键问题。针对复杂环境中的小目标检测与跟踪问题，开展关键技术研究不仅能推动计算机视觉领域的进步，还具有广泛的实际应用意义。

1.2 国内外研究现状

目标检测与跟踪在计算机视觉领域的应用十分广泛，尤其在自动驾驶、安防监控等场景中发挥着至关重要的作用。尽管深度学习方法已大幅提高了检测与跟踪的性能，但在复杂环境下，特别是对小目标的检测与跟踪依然面临诸多难题。

1.2.1 目标检测技术研究现状

近年来，随着深度学习技术的迅猛发展，目标检测技术在计算机视觉领域取得了突破性进展，被广泛应用与各种场景。总体来看，目标检测主要可以分为两大类[14-16]：基于传统机器学习的方法和基于深度学习的方法。

早期的传统目标检测方法主要依赖于手工设计的特征，由于当时缺乏有效的图像表示，人们只能设计复杂的特征表示及各种加速技术对有限的计算资源物尽其用。早期的传统方法依赖人工设计的特征，如 HOG[17]、SIFT[18]等，然后结合滑动窗口或级联分类器对图像进行遍历搜索。这类方法在简单背景下尚能取得一定效果，但一旦面对光照变化、尺度多变或背景复杂的场景，性能往往大打折扣。深度学习的兴起，为目标检测提供了更加丰富的特征表达能力[19]，使检测精度和速度得到显著提升，逐渐成为主流。

常见的深度学习检测框架包括两阶段[20]和一阶段[21]算法。以两阶段算法为例，R-CNN[22] 架构如图 1-1 所示，标志着深度学习在目标检测领域的突破性进展，但它仍然存在一些局限性，大量重叠提议上的冗余特征计算导致检测速度极慢。

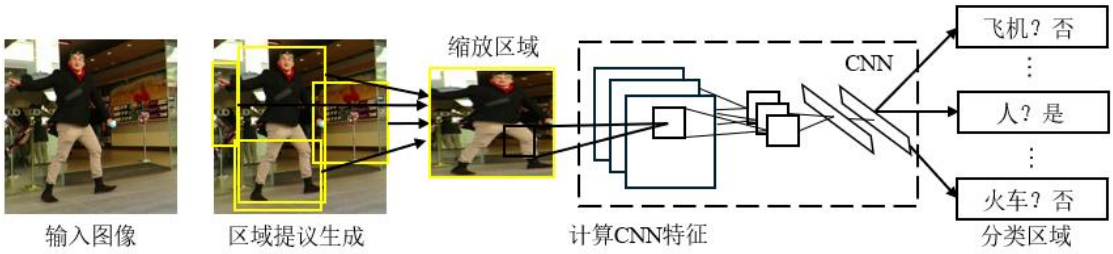

图 1-1 R-CNN 框架

Fig.1-1 R-CNN frame

而 Faster R-CNN[23] 等，通常先通过区域提议网络（Region Proposal Network, RPN）生成潜在候选框，再对候选区域进行精细分类和边界回归，如图 1-2 所示；一阶段算法则以 YOLO[24]和 SSD[25]为典型，通过端到端的方式直接对图像进行预测，具有更快的推理速度。

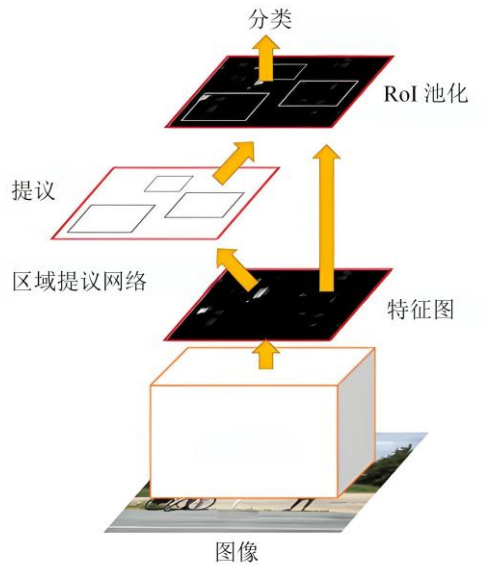

图 1-2 区域提议网络

Fig.1-2 Region proposal network

无论是一阶段还是两阶段算法，在面对常规尺寸的目标时已取得优异表现，但对于小目标的检测仍存在一定的瓶颈。小目标一般指像素面积小于 32×32 或者占原图像比例小于百分之一的物体[26]，由于在图像中所占像素点极少，而当前流行的特征提取网络通常对特征图进行下采样，以减少空间冗余并学习高维特征，这不可避免地会消除微小物体的表示，导致漏检与误检频发。

为进一步提升对多尺度目标，特别是小目标的检测性能，研究者们提出了多

种改进策略。其中，多尺度特征融合方法十分普遍。例如，特征金字塔网络[27]（Feature Pyramid Networks，FPN）采用自上而下与横向连接的方式，将深层语义信息与浅层空间细节融合，从而增强了网络对不同尺度目标的感知能力。此外，Anchor-free 检测器直接在特征图上回归目标的中心及尺寸，减少了对锚框超参数的依赖，在小目标检测任务中展现出一定优势[28,29]。然而，当目标密集或背景干扰严重的情况下，现有方法在定位精度和稳定性上仍难满足要求。为此，部分研究使用反卷积层来提高特征图的分辨率[30]，以进一步优化小目标的定位准确性。

近年来，Transformer 架构逐渐被引入到目标检测任务[31]中，利用注意力机制强调重要区域[32]，同时抑制背景和噪声信息，从而更好地强调那些易受干扰的小目标。如 DETR[33]、Deformable DETR[34]等方法利用自注意力机制进行全局信息建模，简化了传统检测流程，并显著增强了不同特征之间的交互。然而，这类模型通常参数量庞大、并且训练过程收敛较慢，同时复杂的模块设计也会大幅增加模型的整体复杂度和计算成本。

除了网络架构的不断创新，模型轻量化与压缩技术也受到越来越多的关注。研究者利用剪枝、量化以及神经架构搜索[35]（Neural Architecture Search, NAS）等方法，有效减少了模型的参数数量和计算复杂度，从而更适合在资源受限的设备上部署。然而，对于小目标检测任务来说，如何在减小模型体积的同时保持高精度仍是一大挑战。过度精简模型可能会削弱其捕捉细微特征的能力，导致检测效果下降；而为提高准确率而不断扩展网络规模，则会显著增加计算成本，难以满足实时性要求。因此，如何通过合理的结构优化与高效的特征融合，在确保小目标检测精度的同时兼顾模型的实时性能，依然是当前学术界和工业界亟待解决的关键问题。

1.2.2 神经架构搜索技术研究现状

神经架构搜索[38,39]旨在利用自动化方法在庞大的网络设计空间中寻找最优架构，以替代传统依赖专家经验和大量试验确定网络结构的方式。传统的网络设

计过程不仅耗时费力，而且往往难以探索所有可能的结构组合，容易错过那些具有潜在优越性能的架构。NAS 的提出正是为了解决这一问题，通过自动化搜索机制，使得网络能够在精度、计算效率和模型大小之间取得更好的平衡。

早期的 NAS 方法主要基于强化学习[40,41]和进化算法[42,43]。在强化学习方法中，一个控制器网络会通过策略梯度等方式生成候选网络架构，然后根据这些架构在特定任务（例如图像分类或目标检测）上的表现给予奖励信号进行反馈调整。尽管这种方法能够在一定程度上实现自动化网络设计，但其搜索过程通常计算资源消耗巨大且速度较慢，限制了其在大规模问题中的推广应用。

为了降低搜索成本，研究者们又探索了基于进化算法的 NAS 方法。进化算法通过模拟自然选择和遗传进化过程，在一个网络结构种群中不断进行交叉、变异和选择操作，从而逐步优化候选架构的性能。与强化学习相比，进化算法在某些任务上表现出较好的搜索稳定性和并行计算能力，但同样面临搜索效率低和计算资源需求较高的问题。这两类方法的提出，为自动化设计网络奠定了基础，同时也暴露出高计算开销和低效性等瓶颈问题。

近几年来，基于梯度的 NAS[44,45]方法成为研究热点。这类方法的核心思想是将离散的网络架构搜索问题转化为连续优化问题，通过构建一个可微分的搜索空间，使得网络结构参数和权重可以同时进行优化。以 DARTS[46]（Differentiable Architecture Search）为代表的方法，通过引入架构参数，使得每个候选操作在网络中的重要性可以用连续变量来表示，然后利用标准的反向传播算法进行梯度下降优化，从而大幅降低了搜索时间和计算成本。基于梯度的 NAS 方法在搜索效率和稳定性上都取得了显著提升，但也存在易陷入局部最优、对搜索空间设计要求较高等问题。

除了搜索算法本身，搜索空间的设计也是 NAS 研究中的关键环节。搜索空间决定了算法可以探索的网络架构范围，其设计直接影响搜索效率和最终模型的性能。早期的搜索空间往往较为宽泛，包含大量冗余和无效的架构，导致搜索过程漫长且效果不稳定。为此，逐渐引入领域知识和任务约束[47]，对搜索空间进

行合理裁剪。例如，在图像处理任务中，可以根据卷积网络的特性，预先设定一组常用操作（如卷积、池化、跳跃连接等）和连接模式，从而使搜索空间既具有多样性，又符合实际应用需求。这种针对性设计大大提高了 NAS 的搜索效率，同时也保证了搜索到的架构更适合特定任务。

在具体应用方面，NAS 技术已逐步渗透到各类深度学习任务[49,50]中，尤其在目标检测领域展现出巨大的应用潜力。传统目标检测模型往往依赖于由专家设计的骨干网络，再配合检测头进行目标定位与分类。然而，手工设计的网络往往难以同时满足对不同尺度目标的有效的特征提取。利用 NAS 技术，可以自动搜索出更适合视觉任务的网络结构[51,52]。例如，通过在搜索过程中考虑多尺度特征融合和高分辨率特征提取的需求，因此能够设计出在小目标检测上具有更高敏感度和准确率的架构，同时还能兼顾整体网络的计算效率和资源占用。这样的自动化网络设计方法，既减轻了人工设计的负担，又为检测任务提供了新的思路和优化空间。

此外，NAS 技术与模型压缩技术的结合，也成为当前研究的重要方向[53,54]。对于实际应用场景来说，模型的轻量化和高效性尤为关键。通过 NAS 搜索得到的最优架构，往往可以在性能上有所保证，但其参数量和计算复杂度仍可能较高。此时，结合剪枝和量化等技术，可以在保证模型检测精度的同时，进一步降低模型的运算开销，使其更适合在移动设备或嵌入式平台上部署。这种多技术协同的策略，为解决实际应用中轻量化与高精度难以兼顾的问题提供了有效途径。

尽管 NAS 技术在理论和实践中均取得了显著进展，但其应用仍面临不少挑战。例如，在训练过程中，各个通道训练存在不公平性问题，庞大的搜索空间和随机性因素常常导致最终生成的网络结构存在较大差异，生成网络并非最优架构问题尚未得到根本解决。

1.2.3 目标跟踪技术研究现状

目标跟踪[55-57]作为计算机视觉中的一项核心技术，如图 1-3 所示，旨在从连续的视频序列中实时、准确地捕捉目标的运动轨迹。早期的目标跟踪[58,59]主要

依赖人工设计的低级特征，如颜色、纹理和梯度等，再结合模板匹配、粒子滤波或卡尔曼滤波等算法实现对目标位置的估计。这些传统方法实现简单、对硬件资源需求较低，但在应对目标外观快速变化、光照干扰等情况时，往往缺乏足够的鲁棒性，导致跟踪容易出现漂移或丢失。随着应用场景的不断拓展，尤其是在视频监控、无人驾驶和智能安防等领域，对算法的实时性与精准度提出了更高要求。

图 1-3 目标追踪

Fig.1-3 Object tracking

深度学习的崛起为目标跟踪提供了新的技术路径。通过 CNN 提取更具判别力的特征，学术界和工业界先后发展出多种深度跟踪算法[60]，大致可分为单目标跟踪和多目标跟踪两大方向。单目标跟踪[61]（SOT）聚焦于在连续帧中跟随一个给定目标，深度学习在该领域的代表性工作包括基于孪生网络的跟踪器[62-64]其核心思想是利用双分支网络计算模板与搜索区域间的相似度，从而实现端到端的目标匹配。在面对快速运动或外观变化时，部分方法会在跟踪过程中进行在线更新，以适应新的目标特征，但也可能带来模型漂移的风险。此外，近年兴起的注意力机制与 Transformer 架构也逐渐被引入单目标跟踪领域[65]，通过全局信息建模改善了在遮挡和复杂场景下的表现。

多目标跟踪[66-68]（MOT）则需要在视频序列中同时跟踪多个目标，因而必须兼顾目标检测与身份关联等多重需求。传统多目标跟踪大多采用“检测-关联”[69]的分步策略，即先使用检测器在每一帧中定位潜在目标，再通过数据关联算法将跨帧出现的目标加以匹配。典型的关联算法包括匈牙利算法、网络流模型等，它们通常基于目标之间的空间重叠度、运动轨迹以及外观相似度完成匹配。然而，在目标数量密集或出现高度重叠、相似目标的场景下，这类方法常会产生身份切换（ID Switch）或漏匹配。为应对这些挑战，一些工作开始在数据关联阶段引

入更加深度的外观特征[70]和 Re-ID[71]技术，以提高跨帧匹配的精度。

在这一背景下，SORT[72]成为早期极具影响力的工作之一。SORT 采用一个轻量的卡尔曼滤波器来预测目标的运动状态，再结合匈牙利算法进行数据关联，具有实现简单、实时性高的优势。然而，由于缺乏外观特征支撑，SORT 在目标相似或拥挤环境中容易发生 ID 切换。为此，Deep SORT[73]在 SORT 的基础上进一步引入了深度外观特征，通过在关联阶段计算检测框的深度嵌入向量相似度，显著提高了对相似目标的区分能力，因而在现实监控和赛事分析等场景中获得广泛应用。

除了检测-关联策略，近年也出现了将检测与跟踪合二为一的端到端多目标跟踪框架[74]。例如，CenterTrack[75]将 CenterNet 的目标检测思想与跟踪相关信息相结合，在每一帧中同时预测目标中心点与运动偏移量，从而实现检测和关联的统一建模。该方法通过在多尺度特征图上直接回归目标中心并匹配前后帧目标运动位移，避免了显式的外观特征提取过程，因而在保持较高检测与跟踪精度的同时，实现了更快的推理速度。类似的思路也可见于 FairMOT[76]、JDE[77]等工作，尽管具体实现各不相同，但共同点在于将检测器与跟踪器深度融合，一定程度上减少了检测误差对后续关联的影响，并提高了整体的跟踪效率。

然而，不论是“检测-关联”还是“端到端”策略，多目标跟踪依然面临多种复杂环境的挑战。复杂环境通常包含复杂多变的背景信息，如光照变化、阴影干扰、物体遮挡及背景相似物的存在等因素。首先，目标可能频繁的出现部分或完全遮挡，尤其在拥挤场景中，如何在目标可见区域极小的情况下保持对身份的持续跟踪并正确恢复，依旧是难题。其次，不同目标之间经常存在高度外观相似，若仅依赖运动信息进行关联，很容易导致身份切换；但若单纯依赖外观特征，若该特征对环境光照、角度变化敏感，也可能造成关联失败。此外，实时性与计算成本同样考验算法设计，如何兼顾高精度与高效率始终是核心需求。

虽然目标跟踪技术已从传统方法向深度学习驱动的综合性框架转变，并在单目标和多目标领域都获得了显著进步。单目标跟踪依托孪生网络、在线更新以及

Transformer 结构，不断突破在遮挡与外观变化中的瓶颈；多目标跟踪则从 SORT、Deep SORT 等检测-关联方法逐步演进到 CenterTrack 等端到端框架，初步实现了检测与关联的深度融合。然而，物体遮挡、高相似目标、光照变换等复杂环境以及实时跟踪性能等仍是现阶段研究和应用中的难点。随着深度学习和硬件加速器的发展，目标跟踪在精准度、鲁棒性和实时性方面有望迎来进一步的重大提升。

1.3 主要研究内容

本课题主要进行了关于复杂情况下的动态小目标检测和跟踪关键技术研究，利用深度学习基本理论和关键技术，分析常见的目标检测和跟踪算法在小目标任务中的有限性和难点，寻找出适用于小目标检测的网络模型，旨在不显著增加计算成本和参数量的前提下，提高模型对小目标的检测能力，并提高目标跟踪算法在跟踪过程中应对复杂的背景、光照变化、遮挡以及目标运动的不确定性等因素的鲁棒性。因此，本研究的主要内容如下：

(1)基于神经架构搜索的小目标检测算法研究。由于小目标在图像中占据的像素较少，为了获取更多有效信息，通常需要引入多尺度表征或注意力机制，但这不可避免地增加了检测模型的复杂度。针对手工设计的小目标检测模型存在检测能力较弱、模型复杂度高、计算成本和参数量昂贵以及对硬件部署要求高等问题，本研究提出利用神经架构搜索技术来搜索最优且高效的特征提取网络，作为小目标检测的骨干网络，以实现精细的特征提取。在此过程中，解决了为避免引入多余参数而使用权重纠缠所导致的子网训练不公平问题。通过设计专门的搜索空间和评价指标，利用 NAS 技术找到在性能（检测精度）和效率（计算成本、参数量）之间达到最佳平衡的网络架构。将 NAS 搜索的主干网络集成到 Faster R-CNN 目标检测框架中，增强模型对目标的精细特征提取能力，而无需显著增加模型的复杂度。这使得模型在保持检测精度的同时，降低了对计算资源和硬件部署的要求，从而有效地完成小目标检测任务。

(2) 基于线性注意力机制的多尺度特征融合的目标跟踪算法。针对复杂场景中动态小目标易受背景干扰、遮挡及运动不确定性影响导致的跟踪精度下降问题，本研究提出一种基于线性注意力机制的多尺度特征融合的目标跟踪算法。首先，基于神经架构搜索的小目标检测模型，获取高精度的目标初始定位信息；其次，引入具有线性注意力机制的 Cross-Transformer 模块，在降低传统 Transformer 计算复杂度的同时，通过双向特征交互机制（Bottom-to-Up 与 Up-to-Bottom）融合低层结构细节与高层语义信息，增强小目标的细粒度表征能力，解决因目标尺度小、纹理模糊导致的特征提取不足问题。此外，结合对比学习策略，通过密集采样候选区域与双向 Softmax 匹配，优化目标相似性度量，有效应对遮挡与相似目标干扰，并结合置信度阈值调整策略，减少因目标短暂消失或剧烈形变引起的身份切换与轨迹断裂，以提升跟踪的稳定性与鲁棒性。

1.4 论文组织结构

第一章为绪论。本章首先阐述了复杂环境中小目标检测与跟踪所面临的主要挑战，详细讨论了各种因素带来的检测与跟踪难题，探讨了该课题的研究背景与意义。随后，对国内外在小目标检测、目标跟踪以及神经架构搜索等相关领域的研究现状进行了初步分析。最后，本章对本文的主要研究内容、技术路线和文章结构安排进行了详细介绍。

第二章为相关技术概述。本章全面回顾了小目标检测与跟踪领域的国内外研究进展。首先，梳理了传统检测方法及基于深度学习的检测技术，并分析了它们在小目标检测中的优势与不足。接着，深入探讨了 NAS 技术在网络优化中的应用，介绍了其从强化学习、进化算法到基于梯度优化方法的发展历程，以及一些性能评估策略的应用。最后，综述了目标跟踪领域的关键技术，为本文后续工作提供了充分的理论依据和技术参考。

第三章为基于神经架构搜索的小目标检测算法。本章在公平训练网络的基础上，在不同资源约束下自动搜索出最优的网络主干架构，实现对图像中细微特征

信息的高效提取，并将 Faster R-CNN 的主干网络更换为搜索出的网络。最终通过结果证明所提出的方法不仅提升了小目标检测的精度，同时也降低了需要的计算量和参数。

第四章为基于线性注意力机制的多尺度特征融合跟踪算法。该框架引入线性注意力机制，通过多尺度特征的融合，使模型能够更好地捕捉高层语义信息和低层结构细节之间的关系，以显著增强特征表示能力，从而增强了其定位能力和对复杂场景的理解，在处理分类错误和遮挡方面表现出更强的鲁棒性。最后在目标跟踪数据集上进行评估和可视化，证明了算法在复杂环境能够显著提高了目标跟踪的准确性与稳定性。

第五章，总结与展望。总结本文的主要内容及解决的问题，以及有哪些未解决的问题，之后将针对未解决问题如何展开研究等。

第2章 相关技术概述

2.1 目标检测技术

在深度学习方法普及之前，目标检测主要依赖于人工设计的特征和统计学习方法。早期的目标检测研究者们通过手工设计的特征提取图像中的局部信息，再结合滑动窗口或特征金字塔技术，在不同尺度上搜索候选区域。为了减少搜索区域的不确定性，级联分类器被广泛应用于人脸检测领域，并在一定程度上推动了目标检测技术的发展。

然而，传统方法存在较大局限性。首先，人工设计的特征往往难以表达物体的高层语义信息，对于目标的复杂外观和姿态变化缺乏足够的鲁棒性；其次，滑动窗口和金字塔搜索方式计算量庞大，难以满足实时性要求；再者，在处理背景杂乱、光照变化和部分遮挡等问题时，这些方法常常出现漏检或误检。尤其对于小目标，由于在图像中占比微小，传统特征很难捕捉到有效信息，导致检测性能大幅下降。尽管通过引入多尺度特征、多通道融合以及启发式筛选策略能够在一定程度上缓解上述问题，但整体检测精度和效率仍难以令人满意。

随着深度学习的发展，卷积神经网络（Convolutional Neural Network，CNN）在目标检测中的应用逐渐取代了传统的手工特征提取方法。CNN 能够自动从数据中学习多层次的特征表示，从低级的边缘和纹理到高级的语义信息，具有很强的特征表示能力，因此它们被用于目标检测架构。此后，目标检测技术主要分为两类检测器：二阶段检测器与一阶段检测器。两阶段目标检测器是一类在目标检测任务中广泛应用的深度学习模型，其主要特点是将检测过程划分为两个阶段：首先，通过算法（如选择性搜索）生成大量可能包含目标的候选区域；然后，对这些候选区域进行分类，并精确调整其边界框位置。

2.1.1 Fast R-CNN

为了解决 R-CNN 的不足，Girshick 等人在其基础上提出了 Fast R-CNN。Fast R-CNN 可以让检测器和边界框回归器可以在相同的网络配置下同时训练，

采用多任务损失函数，而不是像 R-CNN 那样单独训练它们。在 Fast R-CNN 中，不是每幅图像执行 CNN 2000 次，而是只运行一次并获得所有感兴趣的区域。然后，在最后的卷积层和初始全连接层之间加入 RoI 池化层，这样就可以为所有区域的提议提取一个固定长度向量的特征。

首先，将整张输入图像传入卷积神经网络，生成共享的卷积特征图。然后，使用选择性搜索等方法生成候选区域（Region of Interest，RoI）。接着，利用 RoI 池化层从共享的卷积特征图中提取每个候选区域的固定大小特征。最后，将提取的特征输入全连接层，分别进行分类和边界框回归，输出每个 RoI 的类别和精确的边界框位置。虽然 Fast R-CNN 消耗更少的计算时间，并提高了检测精度。然而，它是基于传统的区域建议方法，它使用选择性搜索方法，使其耗时。

2.1.2 Faster R-CNN

尽管 Fast R-CNN 在目标检测的速度和准确性方面取得了显著进步，但它仍然依赖于诸如选择性搜索这样的传统区域提议算法来生成候选区域。这种依赖导致了候选区域生成过程的效率瓶颈。为了解决这个问题，Faster R-CNN 应运而生，它是首个真正意义上的端到端深度学习目标检测模型。Faster R-CNN 通过引入一个名为区域提议网络的新模块，如图 2-1 所示，其取代了传统的区域提议方法，从而大幅提升了检测速度。

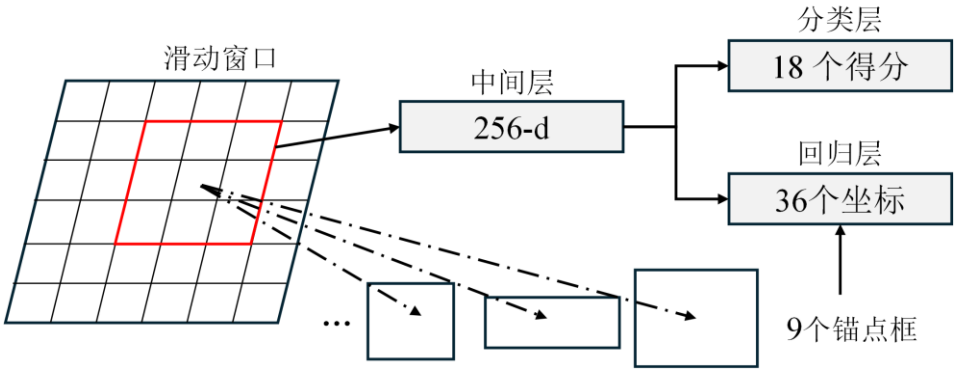

图 2-1 区域提议网络流程

Fig.2-1 Region proposal network flow

首先，输入图像通过一个预训练的卷积神经网络（例如 VGG16 或 ResNet），这个网络被称为主干网络。该网络会输出一张或多张特征图，这些图包含了图像

的空间信息和高层次的语义信息。假设输入图像尺寸为$W \times H$,经过若干次下采样后，得到的特征图尺寸为$\frac{W}{S} \times \frac{H}{S}$,其中$S$是下采样的比例因子。

在特征图上的每个位置，RPN 会生成一组锚点。每个锚点是一个矩形框，具有特定的中心位置、宽度和高度。对于每个位置，会生成多个不同尺度和宽高比的锚点。在 Faster R-CNN 中每个位置会有 9 个锚点，分别是三种不同的尺寸（128×128，256×256，512×512）和三种宽高比（1:1, 1:2, 2:1）。特征图大小定义为$W_f \times H_f$，那么总会有$W_f \times H_f \times 9$个锚点。

接下来，RPN 会在特征图上使用滑动窗口机制进行扫描。对于特征图上的每一个位置，应用一个 3×3 卷积核来提取特征，并将这些特征映射到一个更高维度的空间（如 256 维）。具体来说，特征图的通道数为 C，那么 3×3 卷积层的参数可以表示为：

$$W_{conv} \in \mathbb{R}^{C \times 256 \times 3 \times 3}, b_{conv} \in \mathbb{R}^{256} \tag{2-1}$$

特征向量f可以表示为：

$$f = ReLU(W_{conv} * F + b_{conv}) \tag{2-2}$$

其中，F是特征图的一个局部区域，*表示卷积操作，ReLU 是非线性激活函数。

后续，分类分支用于判断每个锚点是否包含前景（即可能是物体）。这通常是一个二分类问题，使用 softmax 函数来计算概率：

$$p_{cls} = softmax(W_{cls} f + b_{cls})S \tag{2-3}$$

其中$W_{cls} \in \mathbb{R}^{2 \times 256}$和$b_{cls} \in \mathbb{R}^2$是分类分支的权重和偏置。$p_{cls}$是一个二维向量，分别表示前景和背景的概率。

回归分支用于调整锚点的位置和大小，使其更接近真实的边界框。回归分支输出四个值：$t_x, t_y, t_w, t_h,$，他分别表示相对于锚点的中心坐标偏移和宽高变化。具体公式如下：

$$\begin{aligned} t_x &= (x - x_a)/w_a \\ t_y &= (y - y_a)/h_a \\ t_w &= \log(w/w_a) \\ t_h &= log(h/h_a) \end{aligned} \tag{2-4}$$

其中，(x_a, y_a, w_a, h_a)是锚点的位置和大小，而(x, y, w, h)是预测的真实边界框的位置和大小。回归分支的计算公式为：

$$t = W_{reg} f + b_{reg} \tag{2-5}$$

其中，$W_{reg} \in \mathbb{R}^{4\times 256}$和$b_{reg} \in \mathbb{R}^4$是回归分支的权重和偏置。

RPN 的损失函数由两部分组成：分类损失和回归损失。分类损失通常采用交叉熵损失函数，计算公式为：

$$L_{cls} = -\sum_i y_i \log\ p_{i,cls}\ + (1 - y_i) \log\ 1 - p_{i,cls} \tag{2-6}$$

其中y_i是真实标签（0 或 1），$p_{i,cls}$是预测的概率。

回归损失通常采用平滑 L1 损失函数，定义为：

$$L_{reg} = \sum_i \rho(t_i + t_i^*) \tag{2-7}$$

其中ρ指的是 Smooth L1 损失函数，计算每个正样本锚点的预测偏移量t_i与其对应的真值偏移量t_i^*之间的差异。

Faster R-CNN 的总的损失函数可以写作：

$$L(p,t) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i L_{reg}(t_i, t_i^*) \tag{2-8}$$

其中N_{cls}和N_{reg}分别用于计算分类损失和回归损失的样本数，λ是一个平衡因子，用来调整分类损失和回归损失之间的权重，p_i和p_i^*指的是预测的概率分布以及对应的真实标签。

总的来说，Faster R-CNN 不仅继承了 Fast R-CNN 的优点，还通过集成 RPN 实现了对整个目标检测流程的优化，使得该模型既能够快速生成高质量的区域建议，又能精确地定位和分类图像中的物体。这一改进标志着目标检测技术的重大突破，为实时应用提供了强有力的支持。

2.1.3 Feature Pyramid Networks

虽然之前的方法加快了目标检测的实时率，但是其在应对小目标或者一些分辨率较低的目标时，检测率还是较低，因为虽然 CNN 的更深层的特征有利于类别识别，但不利于定位对象。特征金字塔网络被提出用来解决这一问题，如图 2-2 所示，这是一种利用深度卷积神经网络固有的多尺度金字塔层次结构，以低成本构建特征金字塔的方法。FPN 能够接受任意尺寸的图像作为输入，并在多个层次上输出相同尺寸的特征图。该方法在许多应用中表现出显著的提升。

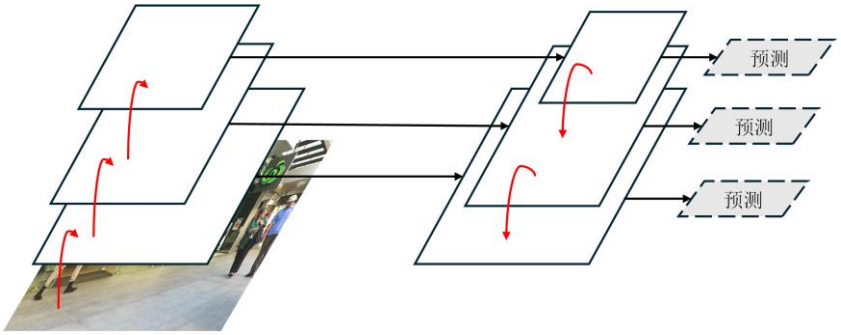

图 2-2 特征金字塔网络

Fig.2-2 Feature Pyramid Networks

需要注意的是，FPN 本身并不是一个目标检测器，而是一个特征提取器，通常与目标检测器结合使用。FPN 的架构通过自顶向下的路径和横向连接，将语义丰富的低分辨率特征与语义较弱的高分辨率特征相结合。具体而言，FPN 利用卷积神经网络的序列结构，构建了自底向上的路径和自顶向下的路径，并通过横向连接进行融合。在自底向上的路径中，图像作为输入传递给 CNN，经过池化层处理后，特征图被调整为相同的尺寸。对于 FPN 的每个阶段（即每个分辨率层级），定义一个金字塔层级。在自顶向下的路径中，通过上采样将高分辨率的特征图恢复到与自底向下路径相同的尺寸，然后通过横向连接，将这些特征与自底向下路径的特征进行融合。每个横向连接将自底向下和自顶向下路径中相同尺寸的特征图进行组合。

FPN 的过程提供了一种广泛的解决方案，用于生成具有丰富语义内容的多尺度特征图。FPN 不依赖于 CNN 的特定架构，可以与不同阶段的目标检测方法结合使用，如区域提议网络、Fast R-CNN 等。通过这些改进，FPN 在多尺度目标

检测中表现出色，尤其在小目标检测方面，显著提高了检测精度和效率。然而，在一些方法中，为了提升对小目标的检测能力，往往在已经庞大的网络上叠加更多冗余模块，这虽然在一定程度上增强了检测能力，但也不可避免地增加了网络的体量。因此，如何在不增加额外模块的前提下改善网络对小目标的检测性能仍然是一大挑战。

2.2 神经架构搜索

长期以来，神经网络的优良架构主要依赖于领域专家的手工设计，这一点可以从许多先进技术中得到验证，如 DenseNet、ResNet 和 VGG 等。这些有前景的卷积神经网络模型都是由在神经网络和图像处理方面具有丰富经验的研究人员设计的。人工设计的 CNN 架构需要深入的领域知识，并且在试错过程中需要大量的计算资源来评估不同架构的性能。然而，在实际应用中，大多数最终用户并不具备这样的专业知识。此外，神经网络架构通常与特定问题密切相关。如果数据的分布发生变化，架构需要相应地重新设计。神经架构搜索旨在通过自动化搜索，从预定义的搜索空间中选择最合适的架构，并将其传递给性能估计。通过评估架构的性能，并将估计结果反馈给搜索策略，经过多轮优化，最终找到最优架构，甚至可以超越人工设计的模型，其过程如图 2-3。

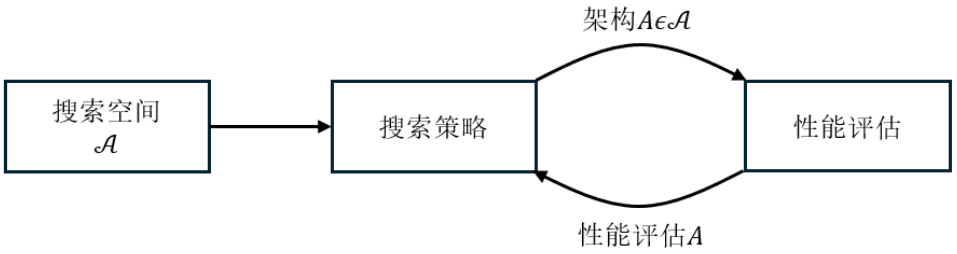

图 2-3 神经架构搜索方法

Fig.2-3 Neural architecture search methods

NAS 可以通过以下公式化的优化问题来建模：

$$\begin{cases} \arg\min_{A} = \mathcal{L}(A, D_{train}, D_{val}) \\ s.t. A \in \mathcal{A} \end{cases} \tag{2-9}$$

其中，$\mathcal{A}$表示潜在神经架构的搜索空间，$\mathcal{L}\cdot$测量架构A在训练数据集D_{train}上训练之后在验证数据集D_{val}的性能，目标是找到一个最佳的网络架构，使其在验证数据集原则上表现最佳。

2.2.1 搜索空间

一个相对简单的搜索空间如图 2-4 所示是链式结构神经网络的空间。链式结构的神经网络架构A可以被写为n层的序列，其中第n层l_n从$n-1$层接收其输入，简而言之，$A = l_n \circ \ldots l_i \circ \ldots l_2 \circ l_1$。搜索空间然后可以被参数化为：（1）最大层数$n$；（2）每层执行的操作类型，如，池化、卷积或更高级的操作，如可分离卷积或者扩张卷积。（3）与操作相关联的超参数，例如，卷积层的卷积核数目、核大小和跨步。其中（3）的参数是以（2）为条件的，因此搜索空间的参数化不是固定长度的，而是条件空间，这样的搜索空间是全局搜索空间，其搜索空间更全面也更灵活，但其也带来了十分昂贵的计算代价。

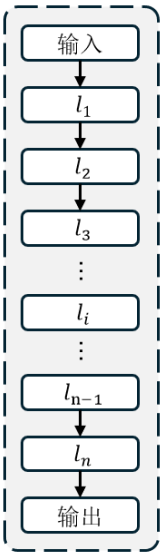

图 2-4 链式神经网络架构

Fig.2-4 Chain neural network architecture

为了提高搜索效率，某些实现可能会对搜索空间进行限定或简化。具体来说，

一些 NAS 方法会将网络划分为基本单元（cell），基于单元的结构通过组合预定义的网络单元来构建网络，每个单元作为一个可重复使用的子网络，具有自己的输入和输出，并能够根据一定规则进行连接与堆叠。构建这种结构通常包括两个步骤：首先，搜索出性能优秀的不同单元结构，如图 2-5 的 a 和 b 结构；其次，将这些不同的单元堆叠，形成最优的网络架构，如图 2-5 的 c 架构。在搜索过程中，通常会对搜索空间进行一定的限定或简化，以提高效率并减少计算复杂度。由于这种方法引入了从局部到全局的设计思路，搜索到的网络模型性能通常较为优秀，因此已成为主流方法之一。但这些 NAS 方法通常需要为每一个候选架构单独训练网络，这意味着每次搜索一个新的架构时，都要进行多次训练，这对于大规模的网络搜索而言是非常低效且计算资源消耗巨大的。

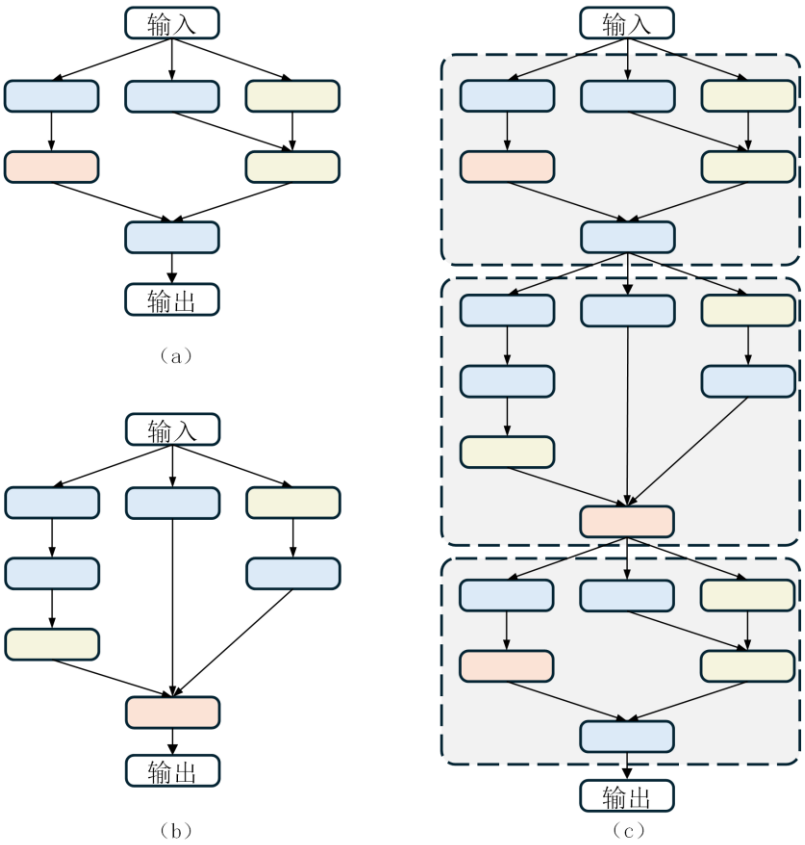

图 2-5 单元搜索空间

Fig.2-5 Cell search space

2.2.2 搜索策略

搜索策略指的是如何有效地探索定义好的搜索空间，以最小的时间成本和计

算成本找到最优的神经网络架构。由于神经架构空间通常非常庞大，搜索策略的设计对于提高效率、减少计算开销至关重要。常见的搜索策略包括进化方法，强化学习（RL）和基于梯度的方法。

（1）基于强化学习的方法

在 NAS 任务中，架构生成过程可以被视为一个强化学习问题，智能体选择动作（如选择层类型、超参数等），并通过测试集上的效果预测函数获得回报，这个函数类似于工程优化中的代理模型。由于在 NAS 任务中，智能体与环境之间没有实际交互过程，搜索问题可以简化为一个无状态的多臂老虎机（MAB）问题。强化学习在 NAS 中的应用始于 2016 年， MIT 提出了 MetaQNN[78]，将架构搜索建模为一个马尔可夫决策过程，使用 Q-learning 算法来生成 CNN 架构。在每一层的选择中，MetaQNN 通过学习智能体选择层的类型和相应的超参数。生成的网络架构经过训练和评估后，得到的精度作为回报，反馈给 Q-learning 算法用于更新策略，使用 10 个 GPU 并花费 8 到 10 天的时间，成功地击败了同类人工设计的网络架构。Google 的工作则采用了 RNN 作为控制器，生成描述网络结构的字符串。这些字符串用于定义网络架构，并通过训练网络获得准确率评估，利用 REINFORCE 算法优化控制器的参数，使得它能够生成更高准确率的网络架构，性能超过了手动设计的网络架构，发现了比传统 LSTM 更优的结构。

（2）基于进化算法的方法

进化算法（Evolutionary Algorithm, EA）是一类受自然选择和遗传学启发的随机优化方法，在 NAS 中通过模拟生物进化过程优化神经网络架构，其核心思想是从一组候选架构中通过选择、交叉和变异操作逐步“进化”出性能更优的解决方案，如图 2-6 所示。进化算法以其强大的全局搜索能力著称，能够高效探索大规模、高维且离散的搜索空间，避免陷入局部最优解，特别适用于 NAS 中网络层类型、连接方式和超参数组合的优化问题。

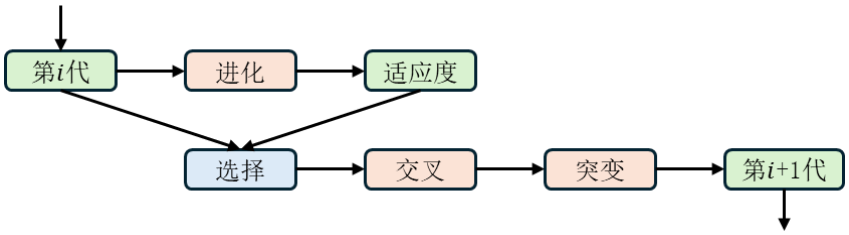

图 2-6 基于进化算法的方法

Fig.2-6 Methods based on evolutionary algorithms

进化算法通常从一个初始种群开始，每个个体代表一个神经网络架构，初始种群可随机生成或基于先验知识设计，种群规模需在多样性与计算开销间权衡，通常设置为几十到几百个个体。接下来，进化算法通过训练每个架构并在验证集上评估适应度（如准确率、计算复杂度 FLOPs 或推理延迟）量化性能，这一过程因涉及大量训练而成为计算开销最大的环节。为挑选优质个体，选择操作从种群中筛选适应度较高的父代，常用方法包括轮盘赌选择（依适应度比例分配概率）、锦标赛选择（随机比较若干个体选优）和排名选择（基于排名而非绝对值选择），旨在保留优秀基因同时维持种群多样性。父代随后通过交叉操作（交换架构的部分结构，如层或模块）和变异操作（随机修改层类型、卷积核大小等超参数）生成子代，交叉继承父代优势，变异增强探索能力，避免搜索停滞。新一代架构生成后，重复适应度评估与遗传操作的迭代过程，并常采用精英策略直接保留前代最优个体，确保优秀架构不因随机操作丢失，推动搜索逐步逼近全局最优。然而，由于每个架构需要训练评估，适应度评估过程计算开销较大。

（3）基于梯度的方法

基于梯度的方法凭借其高效的搜索能力，始终在机器学习领域占据核心地位。与黑箱优化方法相比，梯度法通过显式利用目标函数的微分信息，显著提升了搜索效率。这一优势使得神经架构搜索领域近年重新回归梯度法的研究主线，其中里程碑式的工作当属 DARTS，DARTS 将离散的神经网络架构选择（如卷积核类型、连接方式）转化为连续可微的概率分布，构建基于 cell 的层级化搜索空间。通过对架构参数和网络权重同时进行松弛化，实现了搜索过程的端到端可微分。首次引入解耦式优化策略，将结构搜索问题建模为双层优化:

$$\begin{cases} \alpha^* = \arg\min_{\alpha} ACC_{val}(w^*(\alpha), \alpha) \\ s.t. w^*(\alpha) = \arg\min_{w} \mathcal{L}_{train}(w, \alpha) \end{cases} \tag{2-10}$$

其中，内层优化网络权重 w，外层优化架构参数 α，通过交替梯度下降实现联合优化从而寻找出最佳架构α^*。相比传统 NAS 方法（如基于强化学习的 NAS 需数千 GPU 小时），DARTS 效率提升超百倍。

尽管 DARTS 开创了可微分架构搜索的先河, 在超大规模数据集 ImageNet 或其他数据集上直接搜索成本仍较高，且跨任务迁移能力有限。近年来，梯度法在 NAS 领域的演化呈现轻量化与通用化，通过权重共享、超网（Supernet）等技术，进一步将搜索耗时压缩至数小时级别。

2.2.3 性能评估

性能评估旨在量化候选网络结构的潜在性能如分类准确率、推理速度等，以引导搜索方向。候选架构的性能评估通常需要经历完整的训练验证流程：首先基于训练集优化模型参数，然后在验证集进行精度验证。其核心矛盾在于评估效率与评估可靠性的权衡：理想情况下需对每个架构独立训练至收敛，但从头训练这么多结构计算成本难以承受。因此采用近似评估策略来降低计算资源消耗。然而，这些效率优化策略可能引入系统性偏差，导致性能评估失真。

一些方法[79]使用权重共享来代替重新初始化并且将离散的搜索空间转化为连续的搜索空间，使用梯度优化的方式来建模神经架构搜索问题，大大加速了搜索过程。将之前训练过的子网的权重作为当前要评估的子网的初始权重可以有效的提高训练速度，加速收敛，避免从头开始训练。然而权重共享超网中各个子网耦合度高，不能保证从超网继承权重的方式是有效的，并且同时优化网络权重参数和架构参数会不可避免对架构引入某些偏好，这样在优化过程中会偏向于训练某些权重，造成不公平训练。本质是对架构的相对排序而非绝对性能评估，因此超网训练阶段可能陷入局部最优，使得排名靠前的架构在独立训练时表现平庸。

SPOS[80]提出了一个单路径的超网结构, 减少权重之间的耦合度，该方法将传统网络中的每个层级重构为包含多尺度候选操作的可搜索模块（如不同卷积核

尺寸），并通过路径采样机制实现结构搜索。在保证搜索空间连续性的同时显著降低了内存占用。首先将其将超网权重优化为

$$W_{\mathcal{A}} = \underset{W}{\operatorname{argmin}} \mathcal{L}_{\text{train}}(\mathcal{N}(\mathcal{A}, W)). \tag{2-11}$$

W 代表超网权重，$\mathcal{N}(\mathcal{A}, W)$ 代表超网中编码的搜索空间。

其次，架构搜索过程可以表示为：

$$\alpha^{*} = \underset{\alpha \in \mathcal{A}}{\operatorname{argmax}} \ \mathrm{ACC}_{\text{val}} \ \ \mathcal{N}(\alpha, W_{\mathcal{A}}(\alpha)) \ . \tag{2-12}$$

α 代表被采样的子网架构，它会继承超网的权重 $W_{\mathcal{A}}(\alpha)$，然后在这个过程中挑选验证集上准确率最高的子网结构。在权重共享机制下，采样生成的子网架构通过继承超级网络的参数进行性能评估。这种策略虽然能够快速筛选验证集精度最优的候选结构，但共享参数架构的性能与其独立训练的真实表现存在显著偏差。这种系统性误差严重削弱了超网对子网潜力的评估能力。更关键的是，由于搜索阶段参数未被充分优化，最终选定的最优架构仍需进行解耦式再训练以获得实际可用的模型参数。这种训练流程的割裂不仅增加了整体计算开销，更揭示了权重共享范式在搜索效率与评估可靠性之间的根本性矛盾。

因此 AutoFormer[81]提出权重纠缠策略在每一层中共享其公共部分的权重，如图 2-7，同一层中权重纠缠可以在训练子网中对同一个通道多次训练，而不像权重共享，不同子网架构训练不同的通道，因此其搜索和评估效率更高，权重纠缠训练的子网可以达到与重新开始训练的子网相当的性能，并且降低了内存成本。

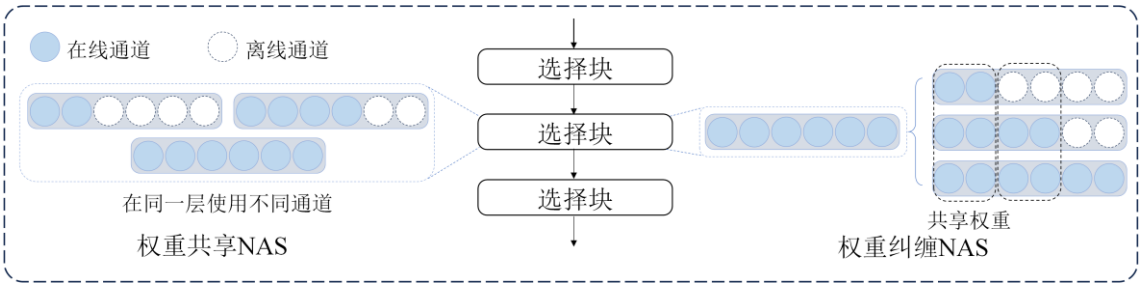

图 2-7 权重纠缠

Fig.2-7 Weight entanglement

权重纠缠虽然显著减少了参数数量和搜索成本，但也引入了超网中子网训练

不足的问题。具体来说，现有的方法使用单边增强原理创建与超网不同宽度的子网，通常选择超网前面的通道来构建采样子网，而不是后面的通道。结果是，前面的通道在各种子网中都得到了训练，因而拥有更多的训练时间。如图 2-8 所示，在构建宽度为 1 到 6 的子网时，第一个通道总是会被采样和训练，而只有在构建宽度为 6 的子网时，才会使用最后一个通道。这种训练中的偏差限制了超网的排序能力，导致难以评估搜索的有效性。

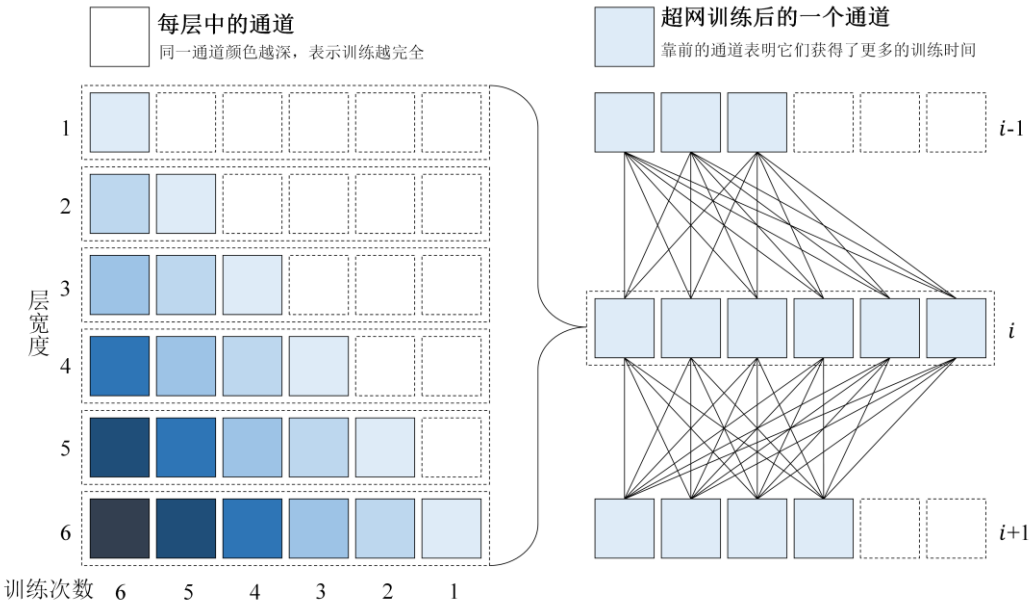

图 2-8 单边增强原理

Fig.2-8 Unilaterally augmented principle

2.3 目标跟踪

目标跟踪是计算机视觉领域的重要任务之一，其核心是从视频序列中实时、稳定地跟踪目标的运动轨迹，构建有效的目标表示、预测其运动轨迹，并在连续帧中实现目标匹配。

2.3.1 卡尔曼滤波

卡尔曼滤波是一种递归最优估计算法，主要用于在噪声存在的情况下估计线性动态系统的状态。其基本过程分为两个阶段：预测和更新。

在预测阶段 k-1 时，假设系统状态的估计为 $x_{k-1|k-1}$，通过状态转移模型对时刻 k-1 的状态进行预测：

$$x_{k|k-1} = F_k X_{k-1|k-1} + B_k u_k \tag{2-13}$$

其中，F_k是状态转移矩阵，B_k是控制输入矩阵，u_k为控制输入。同时，预测误差协方差更新为：

$$P_{k|k-1} = F_k P_{k-1|k-1} F_k^T + Q_k \tag{2-14}$$

这里，$P_{k-1|k-1}$是上一步状态估计的误差协方差，Q_k而是过程噪声的协方差矩阵。在更新阶段，当时刻 k 得到观测z_k，首先计算预测观测与实际观测之间的差异（称为创新或残差）：

$$y_k = z_k - H_k x_{k|k-1} \tag{2-15}$$

其中，H_k是观测矩阵。接着，计算创新协方差：

$$S_k = H_k P_{k|k-1} H_k^T + R_k \tag{2-16}$$

R_k表示观测噪声的协方差。然后，求解卡尔曼增益：

$$K_k = P_{k|k-1} H_k^T S_k^{-1} \tag{2-17}$$

利用卡尔曼增益更新状态估计：

$$x_{k|k} = x_{k|k-1} + K_k y_k \tag{2-18}$$

并更新误差协方差：

$$P_{k|k} = (I - K_k H_k) P_{k|k-1} \tag{2-19}$$

以上预测和更新步骤不断交替进行，使得卡尔曼滤波器能够根据观测数据动态校正系统状态的估计，从而实现对系统状态的最优估计。这些使得卡尔曼滤波在处理线性动态系统和高斯噪声问题时表现优异。

2.3.2 双向 Softmax 匹配

跨帧目标关联是目标跟踪核心挑战之一，为了实现更鲁棒的目标关联，QDTrack[82]提出双向 Softmax 用于对象关联的策略，该机制在匹配过程中确保了匹配双方在特征空间内的双向一致性，从而有效降低了误匹配的风险。该方法的核心思想是，对于当前帧中每个检测到的目标以及候选匹配目标，均采用 softmax 归一化对它们之间的相似度进行处理，这样只有当两个对象在彼此的候

选集合中均为最佳匹配时，其匹配得分才会较高。

具体来说，假设当前帧中有N个检测对象，其特征向量分别为n_i（其中i=1,2,...,N），而在候选匹配集合中有M个对象，其特征向量为m_j（其中j=1,2,⋯,M）。对于任意一对(i,j)，首先计算它们的内积$n_i \cdot m_j$作为相似度的度量。接下来，双向 softmax 的计算分为两个方向：一方面，将检测对象i与所有候选对象之间的相似度进行归一化，得到第一项

$$P_1(i,j)=\frac{\exp(n_i \cdot m_j)}{\sum_{k=1}^{M}\exp(n_i \cdot m_k)} \tag{2-20}$$

另一方面，将候选对象j与所有检测对象之间的相似度进行归一化，得到第二项:

$$P_2(i,j)=\frac{\exp(n_i \cdot m_j)}{\sum_{k=1}^{N}\exp(n_k \cdot m_j)} \tag{2-21}$$

最后，将这两项取平均，得到最终的匹配得分：

$$f(i,j)=\frac{1}{2}\left[P_1(i,j)+P_2(i,j)\right]=\frac{1}{2}\left[\frac{\exp(n_i \cdot m_j)}{\sum_{k=1}^{M}\exp(n_i \cdot m_k)}+\frac{\exp(n_i \cdot m_j)}{\sum_{k=1}^{N}\exp(n_k \cdot m_j)}\right] \tag{2-22}$$

这种双向归一化策略确保了一个匹配对只有在两侧都表现为最近邻时，才能获得较高的得分；如果在任一方向上未能形成一致的匹配，则对应的得分将会降低。这样可以在实际的目标关联过程中，通过简单的最近邻搜索来选择最佳匹配，从而有效地解决由于遮挡、目标重叠或特征模糊引起的误匹配问题。双向 Softmax 不仅简化了传统数据关联算法的复杂性，而且通过增强匹配的一致性，有助于在复杂场景下提升多目标跟踪的鲁棒性和整体性能。

2.4 本章小结

本章系统回顾了目标检测与跟踪领域的相关技术。从目标检测角度出发，介绍了传统手工特征方法和基于深度学习的两阶段检测器，并探讨了多尺度特征融合等方法对小目标检测性能的提升。接着，详细分析了神经架构搜索的基本原理

和关键策略，包括强化学习、进化算法与基于梯度的方法，以及搜索空间设计对架构优化的重要作用。最后，综述了目标跟踪的关键技术，从传统的卡尔曼滤波到双向 Softmax 在解决遮挡、相似目标干扰中的关键技术原理。本章为后续章节中提出的基于神经网络搜索的小目标检测与基于线性注意力机制的多尺度特征融合的目标跟踪算法提供了理论依据和技术支撑，同时也指出了现有方法的不足，为进一步研究指明了方向。

第3章 基于神经架构搜索的小目标检测算法

3.1 引言

在复杂环境的目标检测任务中，目标检测算法在面对图像中占比较小的小目标时，基于人工设计的卷积神经网络往往表现出特征提取能力不足的问题。小目标的细微特征信息被严重压缩，从而导致检测精度下降。为提升小目标检测的精度，通常通过增加网络宽度或深度提升特征表达能力，或者引入注意力机制、超分辨率重建模块等模块以增强对细粒度信息的捕捉。然而，手工设计的复杂模块虽能提升检测性能，却不可避免地导致模型结构变得更加复杂，参数量和计算量激增，进而影响实时性和硬件部署效率。神经架构搜索为解决这一矛盾提供了新思路，其能在不增加额外复杂模块的前提下，通过自动化网络结构优化，在给定计算约束下挖掘最优特征提取网络架构。然而，NAS 通过权重纠缠虽然显著减少了参数数量和搜索成本，但也引入了超网中子网训练不公平的问题，导致使用One-shot权重进行评估的性能与实际架构的性能之间相关性较弱，从而可能导致次优的搜索结果。

基于此，本章提出了一种新颖的神经架构搜索方法，该方法在公平训练网络的基础上，进行梯度聚合更新超网权重，能在不同资源约束下自动搜索出最优的网络主干架构，实现对图像中细微特征信息的高效提取。该方法不仅能够显著提升小目标检测的精度，同时保证对常规尺寸目标的检测能力不受影响。通过合理定义搜索空间以及对超网公平的训练策略，期望在降低模型复杂度和计算资源消耗的同时，实现检测精度与实时性能之间的最佳平衡，从而为小目标检测问题提供一种高效而实用的解决方案。

3.2 基于神经架构搜索的小目标检测算法框架

基于神经架构搜索的小目标检测算法整体框架如图 3-1 所示，整体流程由神经架构搜索与 Faster R-CNN 检测部分组成。

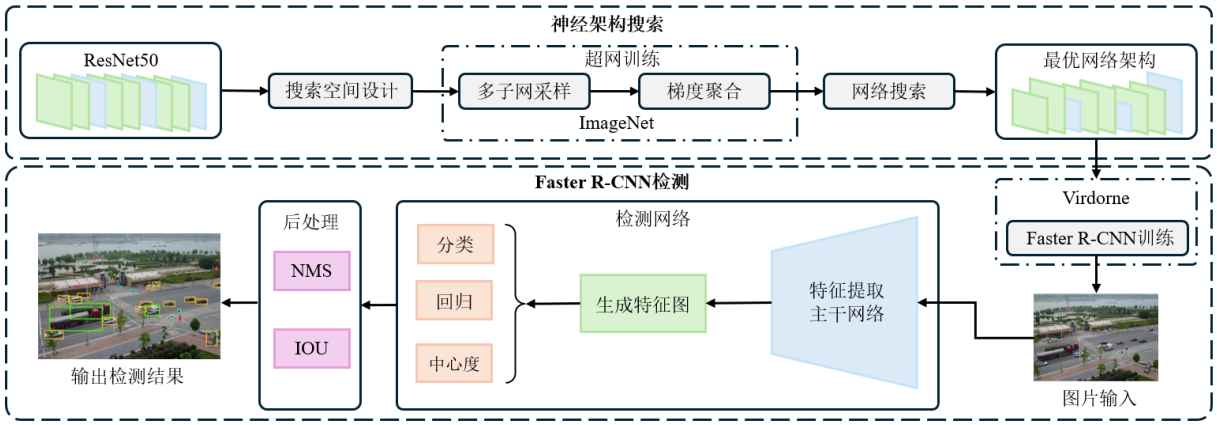

图 3-1 基于神经架构搜索的小目标检测算法框架

Fig.3-1 Small object detection algorithm framework based on neural architecture search

在神经架构搜索阶段，利用神经架构搜索来识别宽度可调的网络 ResNet50，并在预设的搜索空间下提取有代表性的子网。通过约束采样的上下边界，分别对最小、最大及若干中等大小的子网进行采样。在每次训练迭代中，对这些模型执行前向和后向传播，聚合其梯度更新超网的权重，从而提高未采样子网的准确性，拓宽对网络整体性能的理解。此外，通过在优化后的超级网络上进行进化搜索，有效地分配有限的资源，在不同的资源约束下找到最优架构。在检测阶段 Faster R-CNN 目标检测算法获得完成进化搜索后所选取的新的 ResNet50 主干网络，并在小目标检测数据集上进行端到端训练。训练阶段首先通过候选区域生成网络获取潜在目标区域，再利用改进后的主干网络提取深层特征，结合区域特征映射进行分类与边界框回归。由于网络架构在搜索阶段已充分兼顾了不同宽度与深度的子网训练，因而在面对小目标时能保留更多细节特征，有效提升检出率和定位精度。最终，经过后处理后输出检测结果。整个流程在保证实时性与模型轻量化的同时，显著提升了对小目标的检测性能。

3.2.1 神经架构搜索

为了保证搜索效率，我们仅对网络的宽度进行搜索。在一个给定的神经网络 N 中，该结构由 M 层组成，包括归一化层和卷积层，其架构表示为 $A=\{R_1,R_2,...,R_m,...,R_M\}$，$R_m$ 表示第 m 层的输出通道数。通道搜索的目标是逐层识别需要保留的通道。具体来说，对于第 m 层，需要去除的冗余通道集为 U_m^p，因

此有$U_m^p \subset [1, R_m]$，其中$[1, R_m]$表示在该层中从 1 到R_m所有通道的完整集合。每一层的最终宽度$R_m^{'}$可以表示为：

$$R_m^{'} = R_m - \left| U_m^{pruned} \right| \tag{3-1}$$

基于 AutoML 将其概念化为在搜索空间$\mathcal{A}$的自动通道搜索，量化为:

$$|A| = \prod_{m=1}^{M} R_m \tag{3-2}$$

例如，网络由$M = 50$层组成，每层有$R_m = 40$通道，总数则为50^{60}。因此，如果对每一层的每个通道都进行独立搜索，搜索空间将非常庞大。为了减轻搜索的难度，引入分组策略。具体而言，同一层中的所有通道被均匀地划分为K组，从而将每一层选择的数量减少到K，则第m层的宽度选择为$(R_m / K) \times [1:K]$。该策略有效地将搜索空间大小缩小到K^M。

3.2.2 双边约束采样

为在超网训练过程中同时反映网络层面的训练动态，确保训练的公平性，提出了一种双边约束采样方法。该方法对大型、小型和中型模型进行迭代采样，执行前向和反向传播，并在验证期间提供可用的搜索空间边界，从而增强了对架构进行准确排序的能力，保证了网络对于特征提取能力，确保了搜索过程的有效性。

在神经网络中，线性层起着至关重要的作用。第m线性层的输出可以表示为:

$$y_m^r = \sum_{j=1}^{r} \omega_m^j x_m^j \tag{3-3}$$

r表示为通道数，$\omega_m = \{\omega_m^1, \omega_m^2, ..., \omega_m^j, ..., \omega_m^r\}$表示第$m$层相关的权重，$x_m = \{x_m^1, x_m^2, ..., x_m^j, ..., x_m^r\}$是第$m$层的输入特征，$y_m^r$表示聚合的预测输出。从理论上讲，较宽的网络被认为比较窄的网络更有效，主要是因为在较宽的网络中，可以通过将某些权重设为零来模拟较窄网络的架构。这个特性表明，最小化第m层中最大和最小宽度配置之间的输出差异，可以有效地约束该层内的误差，如下所示：

$$\left| y_m^r - y_m^{t+1} \right| \le \left| y_m^r - y_m^t \right| \le \left| y_m^r - y_m^{t_0} \right| \tag{3-4}$$

其中，y_m^t表示第m层中前t个通道的累积输出，t_0是一个预定义的超参数，通常采$t_0=\lceil 0.5r \rceil$。这表明，表示完全聚合的预测输出y_m^r和部分聚合的预测输出y_m^t之间差异的误差τ逐渐减小，并遵循以下约束:

$$0 \le \tau_m^{t+1} \le \tau_m^t \le \tau_m^{t_0}, \tau_m^t = \left| y_m^r - y_m^t \right| \tag{3-5}$$

因此优化超级网络的最窄宽度(下边界)和最宽宽度(上边界)可以间接优化包含不同宽度的子网。这样在层级的方面可以均匀训练每个通道，避免通道级不平衡训练的方法。这种整体方法确保了网络架构的更统一的优化。因此双边约束采样方法，在每次训练迭代中选择最小、最大和一组中等大小的子网进行统一采样，该方法结合子网梯度来全面更新超网，能在验证过程中提供搜索空间的性能边界，还使得网络能够准确地推断未采样子网的性能，从而增强了对网络整体的理解和优化能力。

有M层堆栈的子网$\alpha \in A$，我们将其架构和权重表示为:

$$\begin{cases} \alpha = (\alpha_1, \alpha_2, ..., \alpha_m, ..., \alpha_M) \\ \upsilon = (\upsilon_1, \upsilon_2, ..., \upsilon_m, ..., \upsilon_M) \end{cases} \tag{3-6}$$

其中α_m是第m层的采样块，υ_m表示相应块的权重。

对于体系结构的每一层，都有 n+2 种可能的块配置。 α_m和υ_m都是从潜在搜索空间中预定义的 n 候选块集合中采样的，其公式为:

$$\begin{cases} \alpha_m = (\alpha_m^1, \alpha_m^2, ..., \alpha_m^j, ..., \alpha_m^n) \\ \upsilon_m = (\upsilon_m^1, \upsilon_m^2, ..., \upsilon_m^j, ..., \upsilon_m^n) \end{cases} \tag{3-7}$$

其中α_m^j表示为搜索空间内的候选块，υ_m^j是其相应的权重。

因为遵循权重纠缠策略，在该策略中，允许同一层上的候选节点共享尽可能多的权重。为此，对于同一层中的任意两个α_m^j和α_m^k，表示为:

$$\upsilon_m^j \subseteq \upsilon_m^k \; or \; \upsilon_m^k \subseteq \upsilon_m^j \tag{3-8}$$

这种策略使得权重更新中的υ_m^j和υ_m^k相互纠缠。为了实现优化所有宽度的子网，在设计中定义了最小和最大子网之间的目标边界，利用权重纠缠超网的概念。具体来说，迭代地处理边界和 n 个中间随机子网，合并它们的梯度信息来优化超

网。因此，采样的网络配置定义为:

$$\begin{cases} S_\alpha = \{\alpha^{\min}, \alpha^{r_1}, \alpha^{r_2}, ..., \alpha^{r_N}, \alpha^{\max}\} \\ S_\upsilon = \{\upsilon^{\min}, \upsilon^{r_1}, \upsilon^{r_2}, ..., \upsilon^{r_N}, \upsilon^{\max}\} \end{cases} \tag{3-9}$$

其中 $\alpha^{\min}$ 与 $\alpha^{\max}$ 表示最窄的子网(下边界)和最宽的子网(上边界)，$\alpha^{r_1} = \{\alpha_1^{r_1}, \alpha_2^{r_1}, ..., \alpha_m^{r_1}, ..., \alpha_M^{r_1}\}$ 和 $\alpha^{r_2} = \{\alpha_1^{r_2}, \alpha_2^{r_2}, ..., \alpha_m^{r_2}, ..., \alpha_M^{r_2}\}$ 表示随机选择的两个子网。在图 3-2 中，我们通过将 m 层中的中间子网数量 n 设置为 2 来说明双边约束采样方法。通过双边约束采样方法，保证了每个子网得到公平的训练，同时避免了在确定最优网络后的再训练。

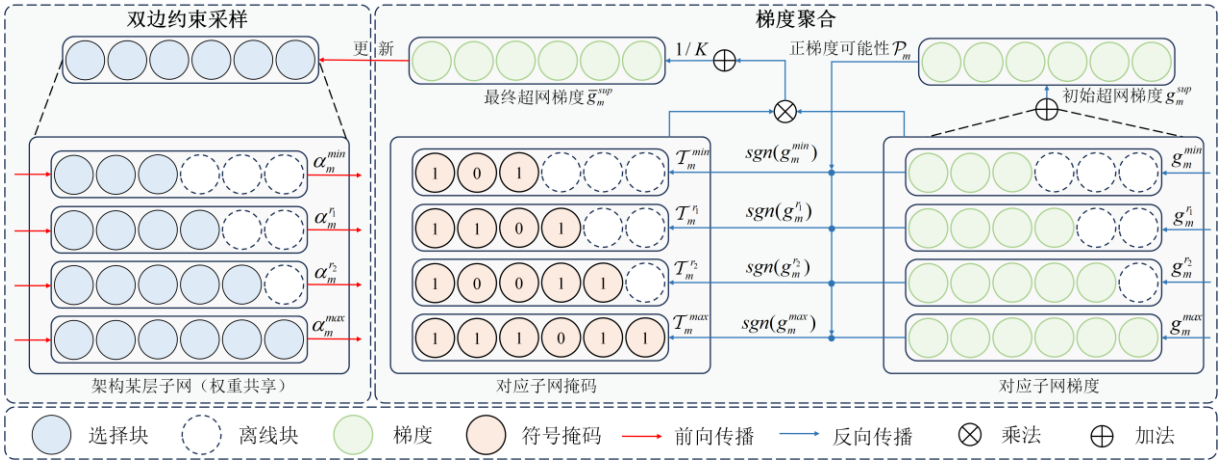

图 3-2 双边约束采样训练主干网络

Fig.3-2 Bilateral constraint sampling trains backbone network

3.2.3 梯度聚合

得到采样子网后，需要利用其梯度来优化网络。因此需要执行反向传播步骤，得出每个采样子网的梯度，其公式为:

$$g^k = \nabla L^k(h^k(f^k(x;\upsilon^k);\varphi^k);p_{gt}) \tag{3-10}$$

其中变量 x 为输入结构 $f^k(\cdot)$ 为第 k 个子网的主干网络，其对应权重 υ^k，$h^k(\cdot)$ 为第 k 个子网的任务头网络，其对应 φ^k，p_{gt} 为 x 对应的真值结果。另外 g^k 和 L^k 分别表示第 k 个采样子网的梯度和损失函数。

一般情况下，可以将所有子网的梯度相加，得到超网的梯度，如下所示:

$$g^{sup} = g^{min} + g^{r_1} + g^{r_2} + ... + g^{r_N} + g^{max} \tag{3-11}$$

公式中 g^{sup} 表示超网的梯度。优化超网的一种方法是通过公式 3.11 中的梯度直接更新其参数。然而，在超网训练过程中，由于超网与其子网之间存在权重共享的区域，而不同子网的梯度方向不同会出现梯度冲突。为此需要对梯度方向进行聚合，通过符号掩码调整和对齐多个子网在共享区域中的梯度方向，从而实现对超网的有效优化。因此定义正梯度可能性的概念，并将其公式化为：

$$\mathcal{P}[i] = \frac{1}{2} + \frac{1}{2} \cdot \frac{\sum_k g^k[i]}{\sum_k \left| g^k[i] \right|}, 0 \le \mathcal{P}[i] \le 1 \tag{3-12}$$

其中，$\mathcal{P}[i]$表示在 n+2 个子网上计算梯度的第 i 个分量为正的可能性。然后，根据 $\mathcal{P}[i]$计算第 k 个子网的第 i 个分量的符号掩码，公式为：

$$\mathcal{T}^k[i] = \begin{cases} 1 & sgn(g^k[i]) > 0 \quad and \quad \mathcal{P}[i] > \bar{\mathcal{E}} \\ 1 & sgn(g^k[i]) < 0 \quad and \quad 1 - \mathcal{P}[i] > \bar{\mathcal{E}} \\ 0 & otherwise \end{cases} \tag{3-13}$$

其中，$\mathcal{T}^k[i]$表示第 k 个子网中第 i 个分量的符号掩码，sgn(·)表示 sign 函数，而阈值 $\bar{\mathcal{E}}$ 设置为 0.5。后续阶段为每个子网 k 定义共识梯度 $\bar{g}^k$，其第 i 个分量如下所示：

$$\bar{g}^k[i] = g^k[i] \cdot \mathcal{T}^k[i], i = 1, 2, \ldots, I \tag{3-14}$$

其中，I 表示梯度向量中的分量的总数。最后，超网的最终梯度向量计算如下：

$$\bar{g}^{sup} = \frac{1}{K}(\bar{g}^{min} + \bar{g}^1 + \cdots + \bar{g}^2 + \cdots + \bar{g}^{max}) \tag{3-15}$$

将 $\bar{g}^{sup}$ 用于更新超网的参数。梯度聚合使所有子网的梯度方向与超网的优化方向保持一致，从而使网络能够更有效地训练。

3.2.4 Faster R-CNN 检测网络模型构建

Faster R-CNN 是一种两阶段目标检测算法，其结构主要由输入端、主干网络、颈部网络和检测头四个部分构成。图 3-3 展示了主干与颈部网络之间的连接

框架。图中展示的 ResNet50 主干网络为搜索得到的 3G 版本，其中卷积核上标记的红色数字表示输入输出通道已发生改变。可以发现在浅层网络时候，ResNet50 的 BottleNeck 发生较大变动，深层变化较小。经过更换主干网络后的 Faster R-CNN，不仅实现了网络的轻量化，同时也大幅减少了参数量。在实际使用中，需要对 3G 版本各层输出与 FPN 的输入进行通道对齐和适配。相较于原始网络，ResNet50-3G 不仅能够在相同硬件环境下处理更高分辨率的图像或实现更快的推理速度，而且通过 FPN 将多层级特征进行融合，最大化地利用了 ResNet50 的多阶段特征，最终在保持较高精度的同时进一步提升了网络效率。

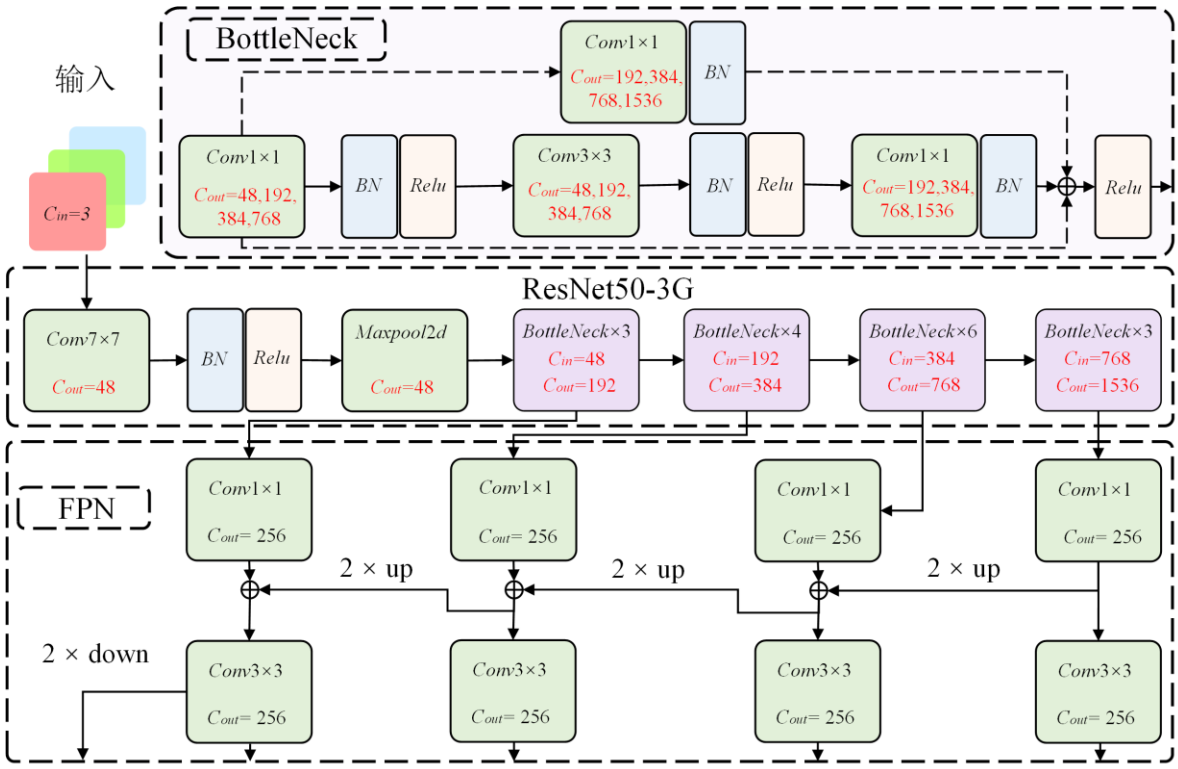

图 3-3 Faster R-CNN 主干与颈部网络

Fig.3-3 Faster R-CNN backbone and neck networks

3.3 神经网络搜索实验结果与分析

在本节中，先在 ImageNet 数据集上评估 ResNet50 架构来进行全面的实验，以证明双边约束采样的神经架构搜索方法得到的网络架构在大规模图像分类任务中的有效性，并展示该网络在特征提取能力与体系结构设置方面的优势。

3.3.1 ImageNet 数据集

ImageNet 是一个规模大、标注精细的图像数据集，并且具有足够的多样性，以便对机器学习模型进行训练和评测，从而推进机器在图像理解任务上的水平。为图像分类、目标检测、图像检索等核心任务提供高质量、精细化标注的海量图像数据。

ImageNet 常用的 ILSVRC 竞赛版本一般会从其中选取 1000 个类别，每个类别的训练图像数量通常在数百到上千张不等，验证集和测试集也各包含 5 万张图像。这些类别类型丰富，包括动物、交通工具、日常用品、植物、器具等，大多都是易于人类识别且相互之间可区分的对象。所有图像都有对应的标签，有些版本还提供了边界框，用于目标分类和检测任务。因此在 ImageNet 数据集上评估搜索出的 ResNet50 网络，更能体现其特征提取能力的有效性。

3.3.2 实验细节

本文实验平台环境设置具体如下表 3.1 所示。

表 3-1 实验环境设置

Table.3-1 Experimental environment setting

名称	属性
CPU	英特尔 至强 Gold 6226R
GPU	Nvidia GeForce GTX 3090Ti×10
System	Ubuntu 20.04.6 LTS
Python	3.7
PyTorch	1.11.0
CUDA	10.2

在搜索空间的设计上，首先以原始的 ResNet50 作为基线配置，并将其通道数划分为 K=10 或 K=20，以方便对不同层的宽度进行灵活调整。在此过程中，为了保证网络的整体结构完整性，允许对每层通道数的最大裁剪比例不超过 80%。也就是说，每一层至少保留 20% 的通道数，以维持其基本的特征提取能力。通过这种分组及裁剪的方式，能够在保留网络主干结构的前提下，实现对网络宽度的精细化搜索。在 ResNet50 的超网训练过程中，将采用 LAMB 优化器，初始学习率设定为 0.008，权重衰减设置为 K=20，训练总周期为 120 个。子网

络的进化搜索实现主要遵循 SPOS 中概述的准则，其中种子数量P设定为 128，k 设定为 64，变异概率设定为 0.2。

3.3.3 实验评价指标

为了更准确地分析搜索得到的主干网络在图像分类任务中的性能和算法在目标检测任务中的表现，本节实验采用了一系列量化指标进行评估。首先，在 ImageNet 数据集上，我们使用 Top-1 精度作为主要指标，只看模型给出的最高分的那个类别是否等于真实标签。通过公式：

$$\text{Top-1精度}=\frac{\text{正确预测样本数}}{\text{总样本数}}\times 100\% \tag{3-16}$$

而 Top-5 精度则是看模型最高分到第五高分这五个类别是否包含真实标签来衡量网络的分类能力。其公式如：

$$\text{Top-5精度}=\frac{\text{正确预测样本数}}{\text{总样本数}}\times 100\% \tag{3-17}$$

此外，本章实验还统计模型的浮点运算次数（FLOPs）和参数量，分别使用卷积层 FLOPs 的估算公式

$$FLOPs=\text{输出宽度}\times\text{输出高度}\times\text{输入通道数}\times(\text{卷积核宽}\times\text{卷积核高})\times\text{输出通道} \tag{3-18}$$

以及参数总数来衡量模型的计算效率和复杂度。通过上述指标的综合分析，我们能够深入评估和对比不同网络架构在大规模图像分类的性能优势与不足，为进一步优化网络结构提供理论依据和实践指导。

3.3.4 搜索网络性能评估

首先，在 ImageNet 数据集上，我们验证通过双边约束采样搜索到的 ResNet50 架构在大规模图像分类任务中的效果，以体现其特征提取能力，并于 DMCP[83]、AutoSlim[84]、AutoPruner[85]、CATRO[86]、BCNet 和 PEEL[87]等方法进行了对比实验。在基准测试中，基线 ResNet50 模型拥有 25.5M 参数和 4.1G FLOPs，能够实现 77.6% 的 Top-1 精度。

以 224×224 作为图片输入分辨率，实验结果如表 3-2 所示，通过双边约束采样对原始 ResNet50 架构进行搜索后，在不同 FLOPs 约束下均取得了最佳

Top-1 准确率。与基线模型相比，ResNet-3G 在参数量减少 2.6M 且 FLOPs 减少四分之一的情况下，Top-1 精度反而更高；同时，与当前先进的 BCNet 方法相比，本文采用双边约束采样获得的 3G FLOPs 和 2G FLOPs 新版 ResNet50 均在 Top-1 准确率上超出 BCNet 0.6 个百分点。其中，2G FLOPs 的 ResNet50 将原始模型的计算量减半，但 Top-1 准确率仅降低了 0.1%。此外，在相似的参数下，本文方法的 1G FLOPs 的 ResNet50 的 Top-1 准确率比 BCNet 提高了 0.5%。这表明，与 BCNet 相比，双边约束采样在训练超网方面更加高效。而且在不同 FLOPs 约束下，新搜索到的 ResNet50 架构在分类性能和特征提取能力上均优于经过蒸馏处理的 AutoSlim，充分证明了在预设上下界内通过插值采样多个子网络的方法的有效性和优势。

表 3-2 ResNet50 在 ImageNet 数据集上不同技术的性能比较

Table.3-2 ResNet50 performance comparison of different technologies on the ImageNet dataset

ResNet50 性能比较				
方法	FLOPs(G)	参数量(M)	测试精度(%)	
			Top-1	Top-5
基线	4.1	25.5	77.6	—
AutoSlim*	3.0	23.1	76.0	—
BCNet	3.0	22.6	77.3	93.7
PEEL	3.0	—	77.6	—
ResNet-3G	3.0	22.9	77.9	93.9
DMCP	2.2	—	76.2	—
AutoSlim*	2.0	20.6	75.6	—
CATRO	2.2	—	76.0	92.8
BCNet	2.0	18.4	76.9	93.3
PEEL	2.2	—	76.5	—
ResNet-2G	2.0	19.1	77.5	93.5
AutoPruner	1.4	—	73.1	91.3
AutoSlim*	1.0	13.3	74.0	—
BCNet	1.0	12	75.2	92.6
PEEL	1.1	—	74.7	—
ResNet50-1G	1.0	12.4	75.7	93.1

同时也搜索出一份 MoblieNetV2 网络，以验证搜索后的网络对于检测任务的提升作用，如表 3-3，方便后续的性能对比。

表 3-3 MoblieNetV2 在 ImageNet 数据集上不同技术的性能比较

Table.3-3 Performance comparison of different technologies for 3MoblieNetV2 on the ImageNet dataset

MobileNetV2 性能比较				
方法	FLOPs(M)	参数量(M)	测试精度(%)	
			Top-1	Top-5
基线	300	3.5	72.6	—
AutoSlim*	305	5.7	71.8	—
BCNet	305	4.8	73.9	91.5
MobileNet-300M	305	4.7	74.3	91.8
MobileNet-200M	217	2.9	73.1	91.6

3.4 目标检测实验结果与分析

将通过进化搜索获得的新的 ResNet50 主干网络替换原有 Faster R-CNN 检测器的特征提取网络，对齐通道，并在 VisDrone2019 小目标检测数据集上进行端到端训练。通过详细的实验征明该改进主干网络在对小目标细节信息的保留与捕捉方面的增强，在减少模型参数量与计算开销的同时，进一步提升了小目标检测精度。

3.4.1 VisDrone 数据集

VisDrone 数据集是近年来无人机视角下视觉研究中广受关注的大规模数据集之一。它涵盖了更广阔、动态变化更频繁的拍摄视角和背景环境，包含晴天、阴天、傍晚、夜晚等多种光照条件，环境复杂度极高。在同一张图像或同一帧视频中，目标可能从远处极小的行人到近距离的大型车辆不等，大小跨度巨大；同时，场景中的目标往往数量众多、分布密集。

由于上述特点，其小目标检测极具挑战性，对算法的泛化能力要求也更高。VisDrone-Detection 子集主要面向小目标检测任务，在该数据集中，提供了数千张无人机拍摄图像，训练集为 6471 张图片，验证集是 548 图片。标注了包括行

人、车辆在内的多种目标的边界框和类别标签，共计 11 个类别。其数据分布如图 3-4 所示，大部分由小目标构成我们在此选取其中的 10 个类别用于训练。

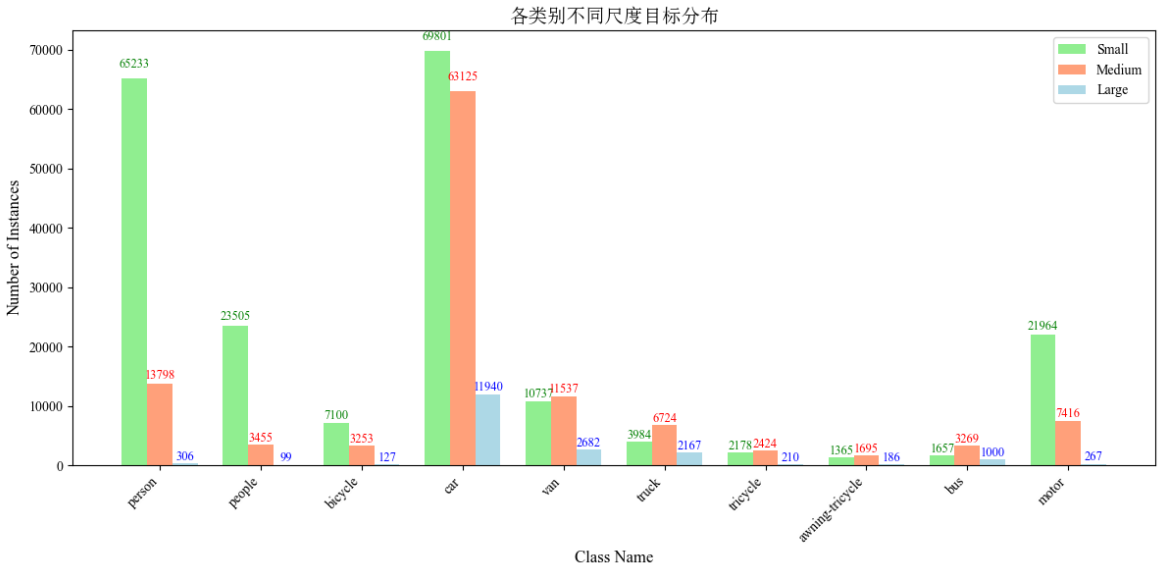

图 3-4 VirDrone 数据集标签分布

Fig.3-4 Label distribution of the VirDrone dataset

3.4.2 实验细节

采用 Faster R-CNN 作为目标检测框架，使其核心结构由 ResNet50 和 FPN 组成。对于 R-CNN 部分，采用随机采样策略，实验图片输入分辨率为 1344×768，Resize 为 640×640 分辨率。每张图像中从候选区域中采样 512 个 RoIs，其中正样本比例设置为 0.25；而在 RPN 部分，每张图像采样 256 个候选锚框，正样本比例为 0.5。训练总共进行 20 个 epoch，优化器采用 SGD，初始学习率设定为 0.02，动量为 0.9，权重衰减为 0.0001。学习率调度采用了两阶段策略：首先，在前 500 次迭代内使用 LinearLR 调度器，将学习率从初始值按照起始因子 0.001 线性升至目标值；随后，在第 12 个和第 17 个 epoch 时，利用 MultiStepLR 将学习率分别衰减 0.1 倍，从而使训练在后期更为细致，并有助于模型收敛。

3.4.3 实验评价指标

在目标检测任务中，采用平均精度均值（mAP）作为评价标准，对于二分类情形下，每个类别都会基于检测结果统计真阳性（TP）、假阳性（FP）和假阴性（FN）。其中，准确率（Precision）可表示为：

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{3-19}$$

而召回率（Recall）为：

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{3-20}$$

随后，通过在不同阈值下计算并绘制准确率-召回率曲线，可得到每个类别的平均精度，再对各类别 AP 求平均，即可得到平均精度均值 mAP。公式可写为

$$\text{AP} = \int_0^1 p(r)dr, \quad \text{mAP} = \frac{1}{C}\sum_{i=1}^{N}\text{AP}_i, \tag{3-21}$$

其中 $p(r)$ 表示在给定召回率 r 下的准确率，C 为类别总数。

3.4.4 实验结果分析

在训练后,更换主干 Faster R-CNN 后以及原始 Faster R-CNN 的平均精度均值、浮点运算量 GFLOPs 以及参数量如表 3-4 所示，其中 mAP_{50} 是在 IoU 阈值为 0.5 时的平均精度均值是作为快速对比模型效果的“基准分数”。

表 3-4 不同主干网络下 Faster R-CNN 实验对比结果。

Table.3-4 Comparison results of Faster R-CNN experiments under different backbone networks

算法模型	主干网络	mAP (%)	mAP_{50} (%)	GFLOPs	Params (M)
Faster R-CNN	ResNet50	20.1	35.6	71.5	41.39
Faster R-CNN	ResNet50-3G	20.7	36.8	68.3	38.42
Faster R-CNN	ResNet50-2G	16.6	32.6	65.2	35.76

可以看出在更换主干网络后，新的 Faster R-CNN 各种指标都优于原始 Fast R-CNN 的方法，mAP 与 mAP50 都有明显上涨，表 3-4 显示检测网络 FLOPs 与参数都有下降，这表明更换主干后的检测算法计算复杂度更低，验证新的网络对特征表达能力的增强效果，也能更好满足复杂环境下小目标检测实时性要求。需要注意的是，当将主干网络进一步压缩到 2G 版本时，虽然 FLOPs 与参数量继续下降，但检测精度也出现明显下滑。

从图 3-5 可以看出在经过 1200Steps 的训练后，检测器的 mAP_{50} 精度变化都趋于稳定，同时可以能看出 Faster R-CNN-3G 的最终 mAP_{50} 精度要高于 Faster

R-CNN。而 Faster R-CNN-2G 由于主干网络的通道减小过多，性能表现不如其他两个检测模型。为兼顾精度与效率，后续实验选用 3G 版本的 ResNet50 作为主干网络。

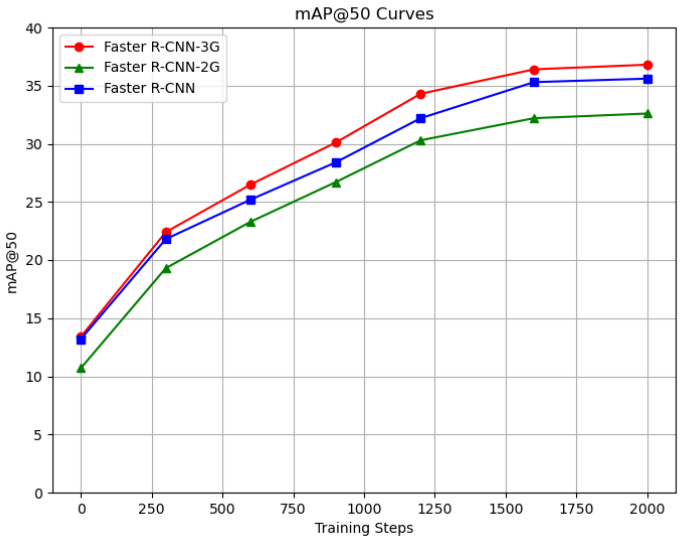

图 3-5 训练中 mAP_{50} 变化情况

Fig.3-5 Changes in mAP_{50} during training

为进一步验证所训练的 Faster R-CNN 小目标检测模型的性能，以及双边约束采样的神经架构搜索方法的有效性，通过与 YOLOX-S 及以更改 MobileNetV2 为 Faster R-CNN 主干网络的算法进行了实验对比，并统一在 VisDrone 数据集上对各模型进行了训练。对比结果如表 3-5 所示。与单阶段的目标检测器 YOLOX-S 进行对比可以发现，使用新 ResNet50 主干网络 mAP_{50} 超过了其 10%，后续再将 Faster R-CNN 的主干网络更换为 MobileNetV2，使用搜索得到的 MobileNetV2-300M 不仅参数量与 FLOPs 降低，其平均精度均值相对于原始 MoibleNetV2 也有大幅提升，这些再次证明我们提出的使用双边约束采样的神经架构搜索方法搜索得到的主干网络具备跨检测框架迁移能力，能保持甚至提升性能，以及其对特征提取能力的有效提升。但由于 MoibleNetV2 过于轻量化，并且精度下降过大，将使用 ResNet50 作为后续实验的主干网络。

表 3-5 不同检测算法不同主干网络对比结果

Table.3-5 Comparison results of different detection algorithms and different backbone networks

算法模型	主干网络	mAP_{50} (%)	GFLOPs	Params（M）
Faster R-CNN	ResNet50	35.6	71.5	41.39
RetinaNet	ResNet50	21.5	—	—
YOLOX-S	CSPDarknet	26.7	26.8	9.0
Faster R-CNN	ResNet50-3G	36.8	68.3	38.42
Faster R-CNN	MobileNetV2	10.0	46.2	19.49
Faster R-CNN	MobileNetV2-300M	13.3	46.5	20.26

图 3-6 显示不同检测算法性能对比示意图也能直观看出，改进后的 Faster R-CNN 的检测效果优于其他模型，在搜索的 ResNet50 使得精度取得最好成绩的同时 FLOPs 以及 Params 参数量也有大幅下降，同时在 MobileNetV2 上在计算量和参数大小相似情况下，其精度也有较大提升，能更好的对复杂视觉图像中的小目标进行检测。

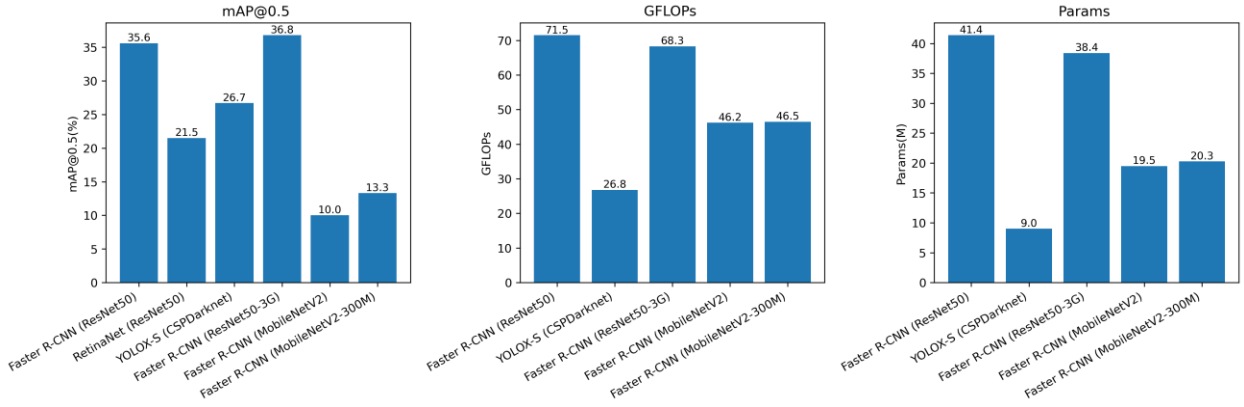

图 3-6 不同检测算法性能对比示意图

Fig.3-6 Performance comparison diagram of different detection algorithms

在图 3-7 中展示了不同检测算法在 VisDrone 测试集上的检测效果图。上侧为 Faster R-CNN 的结果，下侧为更换主干网络后的检测结果。整体对比表明，在针对小目标的场景时，更换主干网络后的算法在检测距离拍摄位置较远的小目标时更加完整，能够检测到更多在马路上的行人，同时对被树木遮挡住大部分的目标也能精确检测，而原始算法则出现较为严重的漏检问题，甚至出现了错误类别的检测结果或者是误检测。由此可见，本文提出的基于神经网络搜索的小目标

检测算法，能够在背景复杂且包含大量小目标的场景下完成更为出色的目标检测任务。

图 3-7 检测效果图

Fig.3-7 Detection effect diagram

在更高空的俯视视角或光线多变、环境昏暗等复杂场景下，该方法依然展现出良好的鲁棒性与适应能力，能够在多变的拍摄角度与光照条件中精准识别绝大部分目标。如图 3-8 所示，在高空拍摄的画面中，大部分汽车都能被准确识别；而在夜间拍摄时，即使部分小目标位于阴影区域，也能成功检测到，进一步验证了该算法的可靠性。

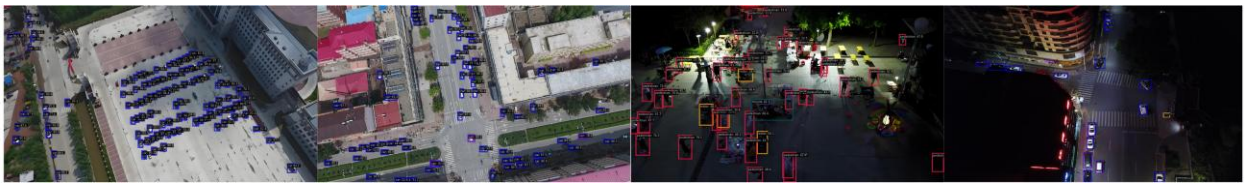

图 3-8 复杂环境下检测效果图

Fig.3-8 Detection effect diagram in complex environment

3.5 本章小结

本章提出了一种基于神经架构搜索的小目标检测方法，以解决传统目标检测网络在处理小目标时特征表达不足的问题。针对手动设计复杂模块所带来的网络冗余和计算成本增加的问题，本章利用神经架构搜索自动优化网络结构，实现在不额外增加复杂模块的、减小计算量的前提下能够有效增强网络对细微特征信息

的捕获能力。实验结果显示，改进后的主干网络在提高检测精度的同时，有效降低了模型参数量与计算成本，在对比传统 Faster R-CNN、RetinaNet、YOLOX 等主流检测模型时表现出优势。

第4章 基于线性注意力机制的多尺度特征融合跟踪算法

4.1 引言

目标跟踪作为计算机视觉领域的核心任务之一，在复杂场景下的鲁棒性仍面临严峻挑战。当前主流的"检测-跟踪"范式通常依赖交并比（IoU）或中心距离进行目标关联，但这类方法在应对非线性运动、动态相机视角及非均匀帧率等复杂情况时存在显著局限。现有研究多将视觉特征作为辅助模块，用来重新识别或增强关联等任务。QDTrack 通过充分挖掘图像和对象信息，其可以更好地学习稀疏边界框和边界框之外的相似性，图像中的许多潜在目标区域特征能够提供价值的训练监督，最终有助于跨帧匹配和跟踪对象。然而，当目标在复杂环境下被部分或完全遮挡时，基于相似外观的判别方法往往无法可靠地区分目标，出现分类错误；尤其当目标尺寸较小，特征难以有效提取，导致跟踪器难以保持准确关联并最终出现跟丢或漂移。

针对上述问题，本章提出一种基于线性注意力机制的多尺度特征融合跟踪算法 CTTrack。该框架引入线性注意力机制，通过多尺度特征的融合，使模型能够更好地捕捉高层语义信息和低层结构细节之间的关系，以显著增强特征表示能力。这种融合增强了定位能力和对复杂场景的理解，在处理分类错误和遮挡方面表现出更强的鲁棒性，从而实现了更稳健的跟踪性能。

4.2 基于线性注意力机制的多尺度特征融合跟踪框架

基于线性注意力机制的多尺度特征融合跟踪，其整体框架流程如图 4-1，所提出的跟踪算法由多尺度特征融合、相似性学习和轨迹关联与处理部分组成。

该算法针对输入图像及参考帧，通过边界框预测头生成密集的候选边界框，并经过非极大值抑制筛选后，获得准密集采样样本。同时引入基于线性注意力机制的 Cross-Transformer，以融合底层特征中的空间结构信息与高层特征中的语义信息，实现跨层的多尺度特征融合。随后利用嵌入特征对候选区域进行正负样

本识别，通过相似度学习优化特征空间，使相同实例的特征更加接近，不同实例的特征更加分离。接下来采用双向 Softmax 计算嵌入特征间的相似性，以增强匹配过程中的双向一致性，得到可靠的匹配分数。在轨迹关联与处理阶段，首先通过跨类别的 NMS 去除检测结果中重复的候选框，避免轨迹的冗余创建。然后根据计算出的匹配分数，若当前检测目标与已有轨迹的相似度超过阈值，则关联并更新相应轨迹；若目标无法匹配现有轨迹但置信度较高，则建立新的轨迹；若目标置信度较低，则暂作为背景对象在一定帧数内保留，以减少频繁出现的误检对跟踪准确性的影响。

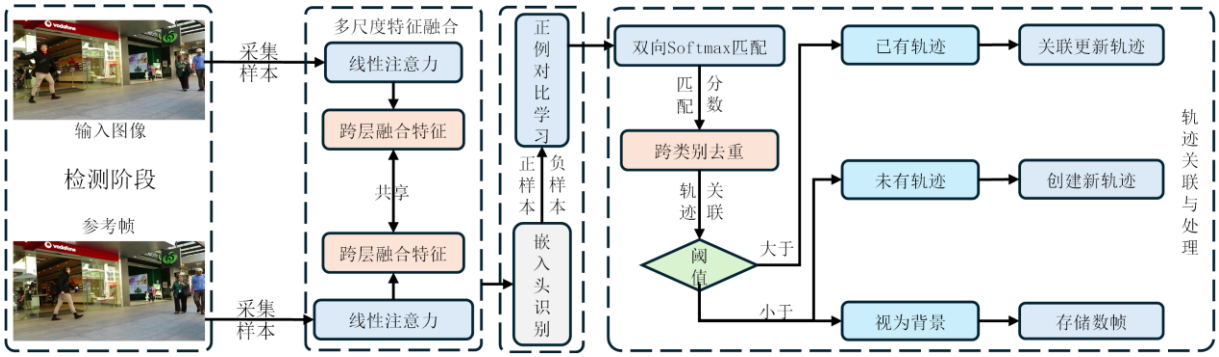

图 4-1 基于线性注意力机制的多尺度特征融合跟踪算法流程

Fig.4-1 Multi-scale feature fusion tracking algorithm based on a cross-attention mechanism flow

4.2.1 特征图采样

特征提取器通常包括一个用于提取特征的模型，沿着有一个特征金字塔网络来处理多尺度特征提取。目标位置和类别由边界框预测模块使用由提取器生成的特征图来识别。为了细化候选边界框，非最大值抑用于过滤和采样准密集样本，从而产生可能包含对象的区域，这些区域通常包括多个重叠的边界框。为了实现更准确的定位，利用图像中的目标区域，并使用 RoI Align 来获取相应的特征图。

对于每个候选区域，首先将其从原图坐标投影到对应的特征图坐标：

$$x_{feat} = \frac{x_{input}}{R}, \quad y_{feat} = \frac{y_{input}}{R} \tag{4-1}$$

其中，R 表示特征图与原始输入图像之间的下采样比。随后将映射到特征图

的候选框区域均匀地划分成固定大小的网格，候选框在特征图上的坐标区域为(x_1, y_1, x_2, y_2),网格的宽度、高度计算为：

$$grid_w = \frac{x_2 - x_1}{W}, \quad grid_h = \frac{y_2 - y_1}{H} \tag{4-2}$$

其中，W，H为预设的固定输出特征尺寸。

随后对于划分好的每个网格点，进行双线性插值。首先确定网格点的坐标，找到与其相邻的 4 个特征图像素点。假设待插值的坐标为(x, y),相邻四个点为：$(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)$,双线性插值公式为：

$$\begin{aligned} f(x,y) = f(x_1,y_1)\cdot(1-\Delta x)\cdot(1-\Delta y) + f(x_2,y_1)\cdot\Delta x\cdot(1-\Delta y) + \\ f(x_1,y_2)\cdot(1-\Delta x)\cdot\Delta y + f(x_2,y_2)\cdot\Delta x\cdot\Delta y \end{aligned} \tag{4-3}$$

其中$\Delta x = x - x_1, \Delta y = y - y_1$,通过这种插值方法，从特征图中准确地获取非整数坐标位置的特征值，避免了直接整数量化导致的精度损失。最终将插值后的特征值，组成一个固定尺寸的输出特征图。

4.2.2 多尺度特征融合

为了解决传统注意力机制的二次复杂性所带来的计算挑战，本文引入了带有线性注意力机制的 Cross-Transformer。线性注意力将计算成本降低到线性时间，使其显着更有效，适用于实时多目标跟踪任务。这种效率的提高使模型能够更有效地处理更长的序列，并支持实时应用程序，这对于在复杂的动态环境中进行跟踪至关重要。此外，线性注意力集中在基本特征的相互作用上，提高了泛化能力，使模型在处理稀疏或噪声数据时更有效。

线性注意力包括多头注意力机制、前馈网络和归一化层。将 Transformer 块的输入特征表示为 X 和 S。给定 X 和 S，首先通过线性投影生成查询、键和值，表示为 Q、K 和 V。该过程表示如下：

$$Q = XW_Q, K = SW_K, V = SW_V \tag{4-4}$$

其中 W_Q、W_K 和 W_V 是使用线性层实现的可学习权重矩阵。之后，使用特征映射函数将查询和键映射到新的特征空间，以获得非线性表示：

$$Q' = \phi(Q), K' = \phi(K) \tag{4-5}$$

其中 $\phi(\cdot)=elu(\cdot)+1$，然后，为了确保数值稳定性并防止可能的溢出，通过将值向量 V 除以序列长度 S_v 来对其进行归一化：

$$V'=\frac{V}{S_v} \tag{4-6}$$

在此基础上，使用线性注意力来全局地建模特征之间的关系，其由下式给出：

$$G=\frac{Q'\cdot(\sum_{s=1}^{S}K_s'V_s')}{\sum_{s=1}^{S}Q_s'K_s'+\epsilon}S \tag{4-7}$$

其中，ϵ 是为了避免被零除而引入的正则化常数。S 表示键和值序列的长度。G 表示检索到的消息，通过线性注意力捕获聚合的和上下文感知的特征表示。此外，MLP 进一步对通过线性注意力机制检索出的特征消息 G 进行特征变换和增强：$G^*=GW_O$。

最后，通过前馈网络（Feed-forward network，FFN）对经过多头注意力层融合后的特征进行进一步的提炼。该前馈网络由两个线性变换层以及它们之间的非线性激活函数构成，表示为：

$$G'=\varrho([G^*\,\|\,X]W_1)W_2 \tag{4-8}$$

其中，W_1 和 W_2 是线性层对应的参数矩阵，$[\cdot\|\cdot]$ 表示沿着通道维度进行特征拼接操作，ϱ 表示非线性激活函数 ReLU。

随后使用 Cross-Transformer 结构用来在特征提取阶段融合来自不同层次的特征信息，以提高模型的整体特征表达能力，如图 4-2 所示。

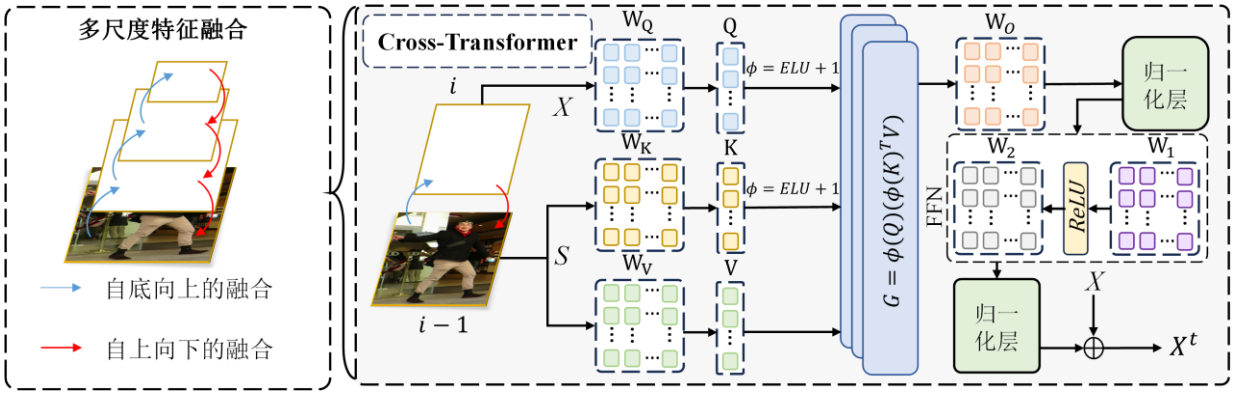

图 4-2 多尺度特征融合

Fig.4-2 Multi-scale feature fusion

Cross-Transformer 模块通过线性注意力机制结合相邻层级的特征表示，从而实现不同特征层间信息的充分交互，其输入特征X与S分别来自不同的特征层。其中 $X=F_i\in\mathbb{R}^{h_i\times w_i\times d}$ 表示第i层的特征， $S=F_{i-1}\in\mathbb{R}^{h_{i-1}\times w_{i-1}\times d}$ 表示第 $i-1$ 层的特征，h，w 分别表示特征图的高度和宽度，d 为特征维度。

图 4-2 中，Cross-Transformer 通过一种双向的信息融合策略实现不同层级特征的有效交互：

首先，模型逐步地将底层特征中精细的结构信息传递至高层，以此实现一种自下而上的融合机制。这种融合方式使得高层特征在保留抽象语义的同时，还能保留底层特征的丰富空间细节，从而增强模型捕捉复杂空间结构的能力：

$$X^t=CT(X,S;\varphi)+X \tag{4-9}$$

接着，模型再逐渐将高层特征所蕴含的语义信息回传到底层，以实现自上而下的融合策略。这种策略使底层特征在保持空间位置精度的同时，获取更丰富的全局上下文信息。因此，底层特征不仅能够保留原有的局部细节，还能融入更多的高层语义信息，从而获得更精确、更具上下文感知的目标特征表示。

$$S^t=CT(S,X;\varphi)+S \tag{4-10}$$

其中，X^t 和 S^t 是更新后的特征，函数 $CT(\cdot)$ 表示特征聚合操作，符号φ为包含跨层 Transformer 结构的 Transformer 模块的参数集合，可以具体表示为：$\varphi=\{W_Q,W_K,W_V,W_O,W_1,W_2\}$。

4.2.3 相似性学习

在训练过程中，首先将关键帧记作 P1，并从其时间相邻范围内随机选取一帧作为参考帧 P2。接下来，分别在关键帧与参考帧中采样大量候选区域，通过之前描述的特征提取步骤获得区域对应的特征表示。随后，通过非极大值抑制在候选区域内进一步筛选获得高质量的目标样本用于相似性学习，具体流程如图 4-3。这种策略能捕获更多图像中的信息区域，极大地提高用于实例相似性学习的训练数据多样性。

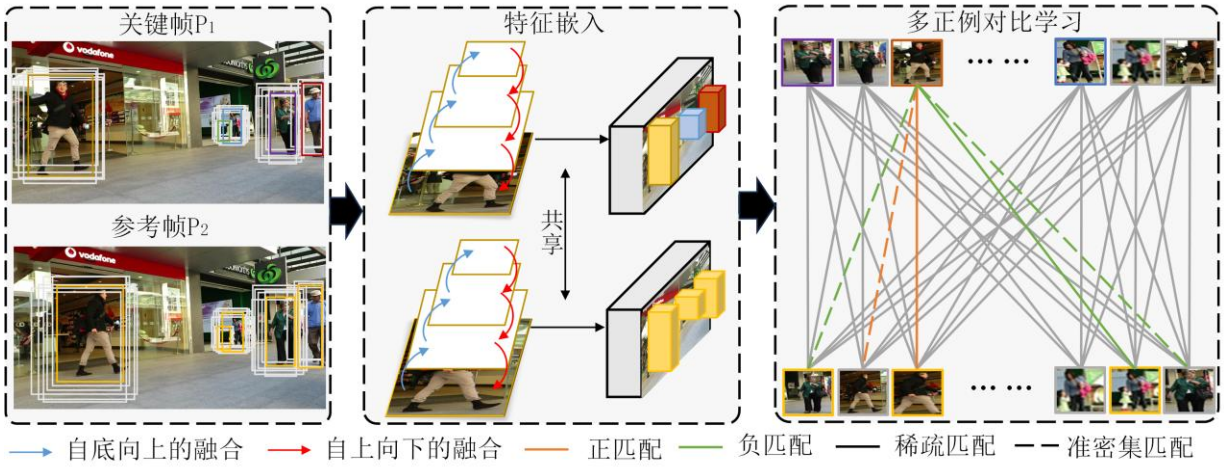

图 4-3 相似性学习流程

Fig.4-3 Similarity learning process

定义训练样本与参考样本之间的正负关系方式为：将并集交集（IoU）大于 γ_1 的区域定义为真实对象的正样本，将 IoU 小于 γ_1 的区域定义为负样本。如果两个帧中的区域对应于相同的实况对象，则我们认为匹配是正的，否则，它是负的。随后，再将包含在关键帧中的样本定义为 M，将包含在参考帧中的对比样本定义为 N。对于每个训练样本，使用对比学习框架的特征嵌入模块的优化过程，其损失函数可以表示为：

$$\ell_{embed} = \log\left[1+\sum_{n^+}\sum_{n^-}\exp(m\cdot n^- - m\cdot n^+)\right] \tag{4-11}$$

其中 m 表示训练样本的特征嵌入，而 n^+和 n^-分别表示正负样本的特征表示。然后，使用 L2 损失作为辅助损失：

$$\ell_{aux} = \left(\frac{m\cdot n}{\|m\|\|n\|} - c\right)^2 \tag{4-12}$$

在两个样本之间的匹配被认为是肯定的情况下，设置 c=1；否则，c=0。损失函数通过计算余弦相似度和标签 c 之间的平方差来惩罚与期望相似度的偏差。整体训练目标结合上述两个损失函数，以端到端的方式优化：

$$\ell = \ell_{det} + \gamma_1\ell_{embed} + \gamma_2\ell_{aux} \tag{4-13}$$

其中，γ_1 和 γ_2 默认设置为 0.25 和 1.0。

4.2.4 目标关联与轨迹管理

为防止类别内部的非极大值抑制导致同一位置存在多个类别不同的检测框。在目标跟踪任务中，这种现象会带来冗余的目标特征嵌入，导致产生多余的轨迹。因此，使用跨类别非极大值抑制来去除这种重复检测，避免后续的重复轨迹创建问题。

然后，采用双向 Softmax 匹配策略计算检测目标与匹配候选目标之间的相似度。假设在 t 帧处，共有 Q 个检测目标，其对应的特征嵌入记为 q；同时，从之前若干帧中，我们已收集了 P 个候选轨迹，其特征嵌入 p。此时，定义检测目标与候选轨迹之间的相似度得分 $Z(i,j)$ 如下：

$$Z(i,j)=\frac{1}{2}\left(\frac{\exp(q_i\cdot p_j)}{\sum_{l=0}^{P-1}\exp(q_i\cdot p_l)}+\frac{\exp(q_i\cdot p_j)}{\sum_{l=0}^{P-1}\exp(q_l\cdot p_i)}\right) \tag{4-14}$$

该公式将每个检测到的对象与所有匹配候选对象的相似性和每个匹配候选对象与所有检测到的对象的相似性相结合，确保匹配中的双向一致性。高分指示检测到的对象和匹配候选者是特征空间中彼此最相似的对应物。

在匹配过程期间，将检测到的对象与匹配候选相关联的决定是基于相似性得分 $Z(i,j)$ 和预定义阈值 α_m。如果分数超过相似度阈值并且检测置信度高于 α_o，则认为匹配成功，并更新相应的轨迹。否则，检测置信度高于 α_n，则初始化新的轨迹；如果置信度低，则将检测结果作为背景对象保留至后续若干帧内再进行匹配判断，以降低误检测对轨迹关联造成的不良影响。

4.3 实验结果与分析

4.3.1 实验指标评价

多目标跟踪算法生成的目标跟踪框应清晰、准确地匹配真实目标的位置，对同一个目标，在整个跟踪过程中应保持稳定的 ID，不应发生错误的身份切换；当目标出现时算法能够迅速创建轨迹，当目标消失时能及时结束轨迹，且避免漏检和误检；因此，需要利用一系列公认的评价指标，来量化评估算法的表现。本

研究将采用以下几种标准指标，对所提出的跟踪算法（CTTrack）进行全面的性能评估。

1）FN：假阴性，漏检数量，表示真实存在的目标被算法漏掉未跟踪的数量。

2）FP：假阳性，误检数量，表示算法错误地跟踪了不存在的目标的数量。

3）IDF1：身份一致性指标，其专注于目标的身份保持能力，用来评估长期跟踪的稳定性。

$$IDF1 = \frac{2 \cdot IDTP}{2 \cdot IDTP + IDFP + IDFN} \tag{4-15}$$

其中，*IDTP* 表示目标 ID 正确匹配的数量；*IDFP* 表示算法生成了多余或错误的 ID；*IDFN* 表示目标真实存在但算法未能成功分配 ID。*IDF*1 其数值越高，表示算法在跟踪过程中，目标身份维持更加稳定一致。

4）HOTA：高阶跟踪精度，能够同时兼顾目标的检测精度和轨迹关联的精度，与传统指标不同的是，HOTA 更强调检测与轨迹关联之间的平衡。

$$HOTA = \sqrt{DetA \times AssA} \tag{4-16}$$

其中，*DetA* 为检测精度，考察算法对目标位置检测的准确性；*AssA* 为关联精度代表算法在各帧之间对同一目标进行正确关联的能力。HOTA 能够更全面地评价目标跟踪算法的综合表现。

5）MOTA：多目标跟踪准确度 ，MOTA 衡量了算法综合的跟踪准确性，包括对漏检、误检和身份切换的整体评价：

$$MOTA = 1 - \frac{\sum(FP + FN + IDS)}{\sum GT} \tag{4-17}$$

其中，IDS 代表跟踪过程中身份切换的总次数；GT 代表所有帧中真实目标的总数量。MOTA 值越高说明算法综合性能越好，意味着算法对目标轨迹的整体保持能力越强。

4.3.2 实验细节

CTTrack 先采用在 COCO-person 数据集上训练的 Faster R-CNN 加特征金字塔结构的检测器对预训练权重进行初始化，实验环境基于 PyTorch 1.11 版本，并

在单张 NVIDIA GeForce RTX 3090 GPU 上进行训练和测试。

本节将实验将在 MOT17 等多个数据集上验证基于线性注意力机制的多尺度特征融合跟踪算法 CTTrack 的性能，将 QDTrack 作为基线，并与其他算法进行了广泛的比较。其中 MOT17 是 2017 年发布的多目标跟踪竞赛标准数据集，共包含 14 个视频序列。这些视频均采自不同场景，背景复杂且包含大量动态元素。在这些视频中，行人经常密集出现，导致频繁互相遮挡和复杂交互的同时也有小目标的存在，从而能够对 CTTrack 的性能进行综合的评估。

在 MOT17 数据集上训练时，我们使用随机梯度下降（SGD）优化器，初始学习率设置为 0.02，动量因子设为 0.9，权重衰减设为 0.001。训练过程中，学习率采用阶梯衰减策略，具体而言，在第 3 个 epoch 时将学习率降低为原来的十分之一，整个训练过程总共持续 4 个 epoch。

为了进一步展示 CTTrack 模型在不同场景下泛化能力方面的优势，将 LVIS 训练集的前半部分数据与完整的 COCO 训练集进行合并构成联合数据集，以此作为训练数据；同时，使用 TAO 验证集评估模型泛化性能。在训练该联合数据集时，仍然沿用了在 MOT17 数据集中采用的相同优化器配置。因为联合数据集规模更大，为此相应地延长了训练轮次，并将学习率衰减的 epoch 调整为第 16 个和第 22 个 epoch，训练过程总共进行了 24 个 epoch。

4.3.3 实验结果分析

为证明本章提出的算法针对 QDTrack 的不足在性能上实现了优化，表 4-1 为我们使用 MOT17 数据库中的视频序列对 QDTrack 与本章提出的算法进行跟踪性能测试和评估结果。

首先在从数据可看出，在衡量身份一致性方面的 IDF1 指标上，我们的方法显著优于 QDTrack，IDF1 提高了约 4.4，表明本章提出的算法在身份保持和长期跟踪稳定性方面更加出色。FN 与 FP 的双双大幅降低，表明本章方法在避免漏跟踪以及提高目标识别精度、减少误识别方面均有明显进步。此外，本章算法在 MOTA 和 HOTA 这两项综合性能指标上也明显优于 QDTrack，进一步说明其在

降低漏检、误检及身份切换方面的整体能力更强。

表 4-1 两种算法性能对比

Table.4-1 Performance comparison between the two algorithms

算法	IDF1(%)	FN	FP	MOTA(%)	HOTA(%)
QDTrack	68.5	42939	8373	68.2	57.1
Ours	72.9	36774	4614	73.7	61.3

为了更直观地展示实验结果，在图 4-4 中可视化了 QDTrack 与本文跟踪算法的过程。前两行的四列图片依次展示了 MOT17 视频中的第 1 帧、第 15 帧、第 75 帧以及第 105 帧，第三行为放大后的细节对比。从第二列可以看到，QDTrack 在右上角坐着的男性小目标上无法实现有效跟踪，而本文提出的方法却能稳定地跟踪该目标。在第 105 帧时，当两人进入商店门口，男生几乎完全被遮挡仅暴露极少量特征，而本文算法依旧能够准确跟踪。以上结果表明，当面对特征表达有限的小目标或严重遮挡场景时，本文算法依然能提取到有效的特征信息并保持稳健的跟踪性能。

图 4-4 跟踪结果可视化

Fig.4-4 Tracking result visualization

同时为了证明本章方法性能的优秀，继续通过在 MOT 数据集与 SORT、DeepSORT、Tracktor、QDTrack 等算法的对比实验，结果如表 4-2 所示。本章节算法的 IDF1 取得了 72.9%的优异成绩，优于 DeepSORT（69.6%）、Tracktor（66.9%）和 QDTrack（68.5%），表明其在目标遮挡与长时间跟踪等复杂场景

中拥有更稳定的身份保持能力。同时，该方法在 FN 与 FP 上大幅下降，展示出对小目标与复杂环境的更强鲁棒性。在MOTA指标上，本章算法取得了73.7%，凸显了其在漏检、误检以及身份切换等方面的综合优势；HOTA 指标则同样达到最优，说明其在检测精度与轨迹关联间实现了更佳平衡。图 4-5 给出了性能可视化对比结果。本章方法通过多尺度特征融合与 Cross-transformer 的有效结合，不仅能更好地应对 MOT17 中小目标的跟踪需求，还在精度与鲁棒性方面全面超越了传统方法。

表 4-2 MOT 数据集的性能比较

Table.4-2 Performance comparison of MOT datasets

算法	IDF1(%) ↑	FN ↓	FP ↓	MOTA(%) ↑	HOTA(%) ↑
SORT	57.8	82451	15171	62.0	-
DeepSORT	69.6	40326	15060	63.8	-
Tracktor	66.9	45762	11088	64.1	-
QDTrack	68.5	42939	8373	68.2	57.1
Ours	72.9	36774	4614	73.7	61.3

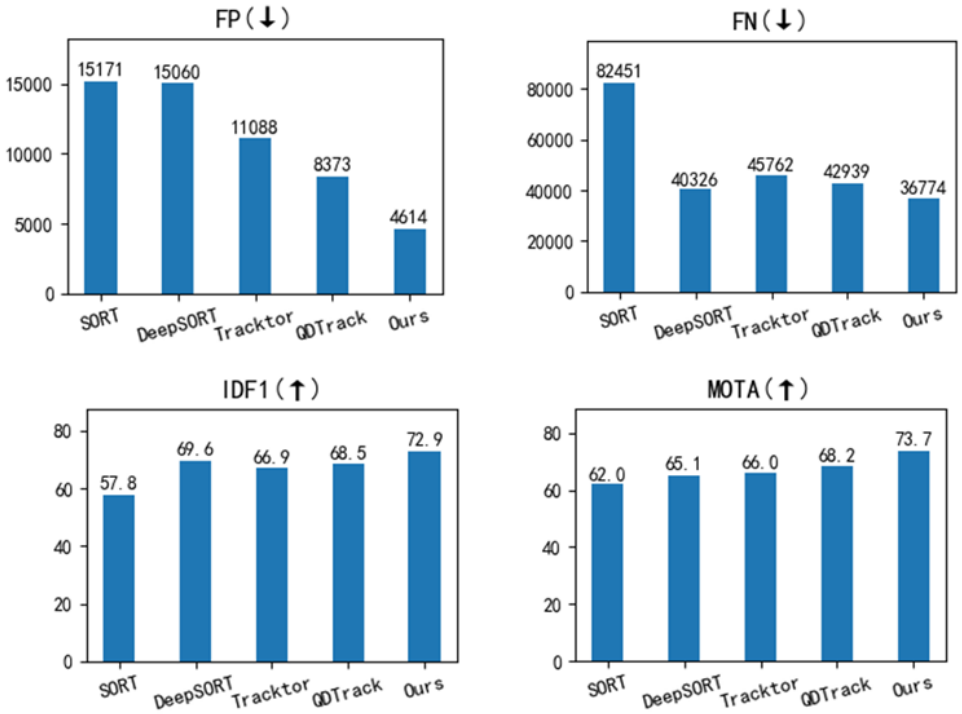

图 4-5 跟踪性能比较

Fig.4-5 Tracking performance comparison

为了进一步展示 CTTrack 在不同场景下泛化能力方面的优势，表 4-3 显示了

本章方法和基线在联合数据集上的性能比较。其中，CTTrack 在所有指标上都实现了更高的 AP，其 AP 为 19.8%，优于 QDTrack 的 17.2%。对于 AP_{50} 和 AP_{75}，CTTrack 分别达到 30.9%和 20.1%。更重要的是，CTTrack 在检测各种大小的物体方面表现出了显着的改进，AP_S（小物体）从 5.3%增加到 7.6%，AP_M（中等物体）达到 15.1%，AP_L（大物体）也有所改进。这些结果表明，我们的基于线性注意力机制的多尺度特征融合跟踪算法有效地处理不同尺度的对象，提高了在复杂环境中的整体跟踪性能，进一步证实了 CTTrack 在不同跟踪条件下的鲁棒性和适应性。

表 4-3 联合数据集性能对比结果

Table.4-3 Performance comparison results of the joint dataset

算法	AP ↑	AP_{50} ↑	AP_{75} ↑	AP_S ↑	AP_M ↑	AP_L ↑
QDTrack	17.2	28.6	17.7	5.3	13.0	22.1
Ours	19.8	30.9	20.1	7.6	15.1	24.0

为了证明 CTTrack 模型中基于线性注意力机制的多尺度特征融合模块的有效性，设计了一项由三个实验组成的消融实验，如表 4-4。

表 4-4 消融研究结果

Table.4-4 Ablation study results

	基线	B2U	U2B	IDF1	MOTA	HOTA
实验 1	√			68.5	68.2	57.1
实验 2	√	√		70.2	69.9	59.1
实验 3	√		√	70.6	70.2	59.7
实验 4	√	√	√	72.9	73.7	61.3

在实验 1 中，我们以基本的 QDTrack 模型作为基线；实验 2 中，采用自底向上（B2U）Transformer 编码器，以提升特征表达质量；实验 3 则整合了自上而下（U2B）Transformer 编码器。最终，在实验 4 中，我们同时引入了 B2U 和 U2B Transformer 模块，实现不同特征层之间的功能交互。MOT17 数据集上的实验结果表明，单独添加自底向上或自上而下的 Transformer 编码器均能提高模型的检测精度和多目标跟踪性能，而在实验 4 中结合两种方向的模块后，模型在 IDF1 和 HOTA 指标上均有显著提升，同时身份切换和碎片化现象大幅减少。该

消融研究验证了 Transformer 在特征提取及多尺度特征交互方面的有效性，表明将两种方向的特征表达相结合，能够显著提升跟踪精度和稳定性

4.4 更改检测器的 CTTrack 评估结果

为进一步提升CTTrack在目标跟踪中的表现以及增强其在复杂环境下的鲁棒性，本节将在第3章提出的基于神经架构搜索的小目标检测算法作为CTTrack的检测器进行深入实验。通过在 MOT17 数据集上进行评估，并在 VisDrone 数据集上跟踪实验并可视化，来检验在多种复杂环境下的检测有效性以及跟踪算法的泛化性能。

在 MOT17 数据集上的性能评估结果如表 4-5 所示。可以发现，引入线性注意力机制来融合多尺度特征后，CTTrack 的性能虽然得到提升，但由于增加了额外的计算量，FPS 略有下降。相较之下，更改检测器的 CTTrack 在检测阶段中采用了前章基于神经架构搜索的小目标检测算法，使主干结构更为轻量化且参数量更少的网络，从而在保持甚至提升检测精度的同时，提高了推理帧率，保证了跟踪过程的连贯性。漏检、误检以及身份切换等综合指标的改善，进一步证明了该算法的强大特征提取能力与有效性。

表 4-5 改进 CTTrack 性能对比结果

Table.4-5 Improved CTTrack performance comparison results

算法	IDF1(%)	FN	FP	MOTA(%)	HOTA(%)	FPS
QDTrack	68.5	42939	8373	68.2	57.1	10.5
CTTrack	72.9	36774	4614	73.7	61.3	7.9
改进 CTTrack	74.5	35434	4320	74.9	62.7	8.7

图 4-6 展示了在 MOT17 数据集上的跟踪可视化结果。该场景行人密集、目标容易发生重叠，同时由于拍摄视角较高，目标尺寸相对较小。在第 1 帧与第 15 帧的下方位置，基线算法和原始 CTTrack 都未能区分紧挨在一起的两位身穿黑色西装的男性，而改进 CTTrack 由于采用了特征提取能力更强的检测算法，能够将这两名男性识别为不同个体。此外，在第 15 、 55 以及 70 帧中，最右侧的一名女性被广告牌部分遮挡，导致基线算法丢失了该目标；然而，本章方法均能在遮挡情形下持续跟踪该女性，尤其是改进 CTTrack 在目标仅剩极少可见特

征时，仍能准确定位并保持稳定关联，展现出更优异的跟踪鲁棒性。

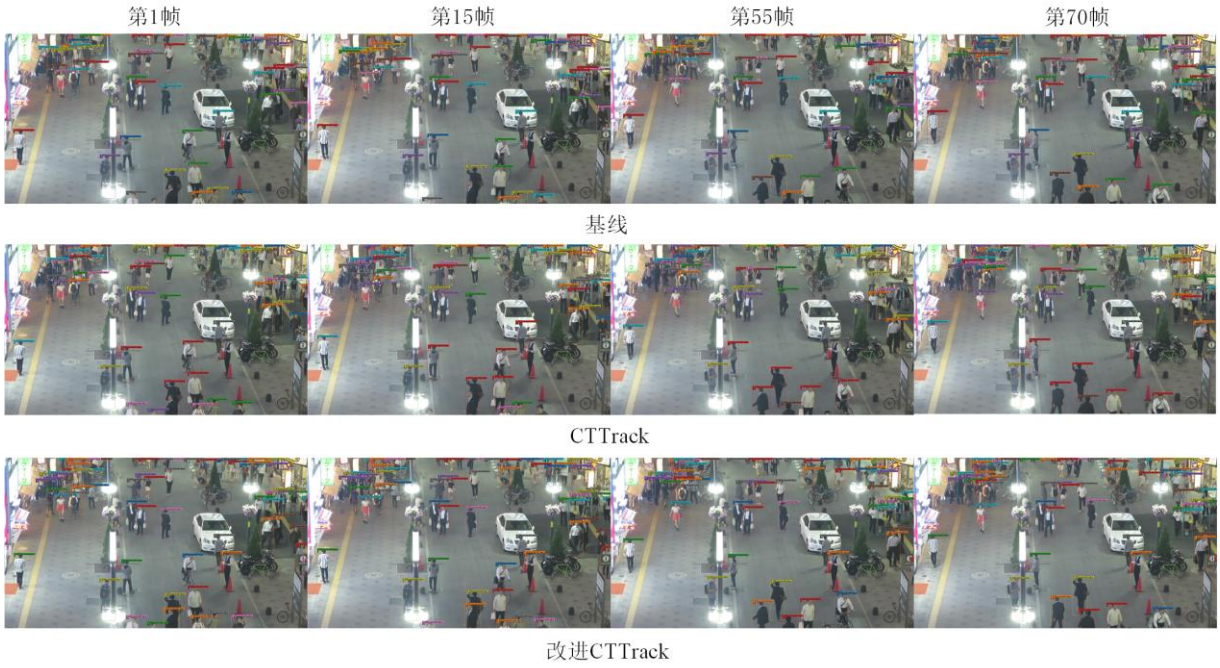

图 4-6 MOT17 数据集跟踪可视化

Fig.4-6 MOT17 dataset tracking visualization

为验证本章方法的泛化性能，我们在 VisDrone-MOT 数据集上进行推理并对结果进行可视化，对比结果显示如图 4-7。

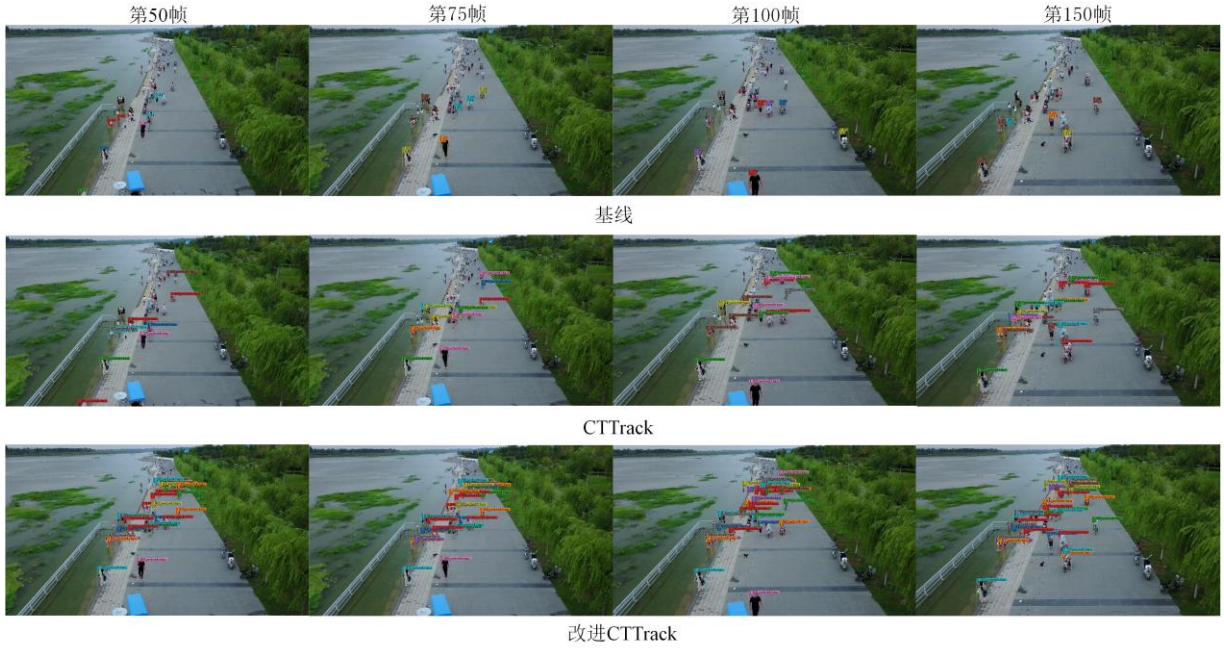

图 4-7 VisDrone-MOT 数据集跟踪可视化

Fig.4-7 VisDrone-MOT dataset tracking visualization

基线方法在面对较多小目标时存在非常多的漏检，且对同一目标的轨迹关联并不稳定。相比之下，本章提出的跟踪算法利用多尺度特征融合，特征表达能力更强，能够捕捉更多小目标，有效减少了漏检情况，轨迹关联也较为稳定；但当无人机向前飞行导致相机运动时，仍有部分轨迹未能成功关联。进一步替换检测器后，改进的 CTTrack 有效降低了此类关联失败的数量，展现出更好的适应性与鲁棒性。

在光线多变、人流密集、遮挡较多及相机运动的复杂环境下，结合小目标检测器后的 CTTrack 仍展现出优秀的跟踪能力。如图 4-8 所示，跟踪器依旧能稳定跟踪到背景中一些较小目标的行人；同时，在行人被广告牌大部分遮挡的情况下，跟踪器依然能够实现稳定追踪。进一步展示了该方法跟踪的稳定性与准确性。

图 4-8 复杂环境下改进 CTTrack 跟踪可视化

Fig.4-8 Improving CTTrack tracking visualization in complex environments

4.5 本章小结

本章提出了一种基于线性注意力机制的多尺度特征融合目标跟踪算法，通过引入具有线性注意力机制的 Cross-Transformer 来进行多尺度特征交互，以提升对小目标及复杂场景的检测与关联能力，实验结果表明，该方法在人群密集遮挡严重以及具有多数小目标的数据集上有效降低了漏检、误检及身份切换等问题，而进一步将基于神经架构搜索的小目标检测算法整合至 CTTrack 后，检测精度和推理帧率均得到提升，展现出更强的泛化性能和鲁棒性。

第 5 章 总结与展望

5.1 工作总结

本文围绕复杂环境下动态小目标检测与跟踪关键技术展开深入研究，以解决实际场景中存在的小目标检测精度低、跟踪稳定性差等难题。本文在对现有目标检测与跟踪方法深入分析的基础上，结合当前计算机视觉领域中的前沿技术，提出了一种基于神经架构搜索的小目标检测算法，并进一步设计了一种基于线性注意力机制的多尺度特征融合的目标跟踪算法。这些方法的引入有效地解决了小目标特征难以捕捉、目标跟踪易受遮挡及背景干扰的问题，显著提升了检测与跟踪任务的性能。

在目标检测方面，本文提出了一种新颖的神经架构搜索方法。通过精心设计的搜索空间与双边约束采样策略，本文构建了一种公平且高效的网络搜索框架，能够自动寻找出适合高效的最优主干网络结构，在每个训练迭代中同时采样并训练最大、最小及若干中间子网，以实现对子网络的公平训练，从而避免了传统权重共享模式下出现的训练不均衡问题。同时，通过梯度聚合的方法，融合其梯度以更新超网充分挖掘网络潜在的性能，显著降低了模型的参数量和计算开销。在复杂环境的小目标检测任务中，更换为经过 NAS 搜索的主干网络后，Faster R-CNN 在小目标检测任务上的性能明显优于传统方法，同时模型复杂度显著降低，兼顾了检测的精度和实时性。

在目标跟踪方面，本文提出了一种基于线性注意力机制的多尺度特征融合的目标跟踪算法，以解决复杂环境中动态小目标易受背景干扰、遮挡及运动不确定性影响而导致的跟踪精度下降问题。本文通过带有线性注意力机制的 Cross-Transformer，有效降低了传统 Transformer 模型高昂的计算复杂度，实现了更高效的全局信息建模能力，并将低层次的空间细节与高层次的语义信息进行融合，解决了传统方法中特征提取不充分的问题。此外，结合对比学习与双向 Softmax 匹配机制，算法进一步提升了目标跟踪的稳定性，降低了遮挡、目标重叠或运动

不确定性带来的影响。在实验验证阶段，证明了本文提出的算法在多目标跟踪精度与稳定性等指标上均优于现有主流方法。这进一步说明了本文提出的基于线性注意力的多尺度特征融合目标跟踪方法能够更有效地应对复杂环境下小目标的动态跟踪需求。

5.2 工作展望

尽管本文的研究工作在理论和实践应用方面均取得了积极的成果，但由于当前复杂场景下的小目标检测与跟踪任务仍然十分艰巨，因此仍存在进一步研究和完善的空间：

当前的 NAS 方法主要聚焦于对网络宽度的搜索，未来可以扩展到更为灵活的搜索空间，如网络深度、结构类型和空间连接方式等，以进一步提高网络结构的通用性与鲁棒性。同时，由于 NAS 本身搜索过程仍然存在计算开销较高的问题，如何更高效地进行超网训练和快速搜索仍值得进一步深入探索。此外，通过构建出一个大型的小目标检测数据集，然后在此数据集上直接使用 NAS 对目标检测网络进行搜索，也能提高小目标检测网络的性能，提高小目标检测的精度。

在目标跟踪算法方面，使用 Cross-Transformer 进行多尺度特征融合虽然提升了小目标的跟踪鲁棒性，但仍未完全解决真实情况下在极端密集场景、剧烈光照变化或快速运动场景下目标身份切换的问题。后续可尝试将跨模态信息，如激光雷达数据，融合到跟踪网络中，以提高复杂环境下目标跟踪的可靠性。

参考文献

[1] 尹艺晓, 马金刚, 张文凯, 等. 从 U-Net 到 Transformer: 混合模型在医学图像分割中的应用进展[J]. Laser & Optoelectronics Progress, 2025, 62(2): 0200001-0200001-23.

[2] 赵阳, 李俊诚, 成博栋, 等. 深度学习在口腔医学影像中的应用与挑战[J]. 中国图象图形学报, 2024, 29(3): 586-607.

[3] 黄晓鸣, 何富运, 唐晓虎, 等. U-Net 及其变体在医学图像分割中的应用研究综述[J]. 中国生物医学工程学报, 2022, 41(5): 567-576.

[4] 扈拯宁, 黄强, 谢尧, 等. 基于卷积神经网络的卵巢囊腺瘤 CT 图像病灶分割[J]. 计算机应用与软件, 2025, 42(1): 189-196.

[5] 陈志宏, 王明晓. 计算机视觉在智慧安防中的应用[J]. 电信科学, 2021, 37(8): 142-147.

[6] 曹家乐, 李亚利, 孙汉卿, 等. 基于深度学习的视觉目标检测技术综述[J]. 中国图象图形学报, 2022, 27(6): 1697-1722.

[7] Boukerche A, Ma X. Vision-based autonomous vehicle recognition: A new challenge for deep learning-based systems[J]. ACM Computing Surveys (CSUR), 2021, 54(4): 1-37.

[8] 熊守丽. 无人船红外图像单目视觉检测与跟踪研究[J]. 舰船科学技术, 2024, 46(7): 159-162.

[9] 陈琳, 刘允刚. 面向无人机的视觉目标跟踪算法: 综述与展望[J]. 信息与控制, 2022, 51(1): 23-40.

[10] Cheng G, Yuan X, Yao X, et al. Towards large-scale small object detection: Survey and benchmarks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13467-13488.

[11] Zhu X, Lyu S, Wang X, et al. TPH-YOLOv5: Improved YOLOv5 based on

transformer prediction head for object detection on drone-captured scenarios [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 2778-2788.

[12] Zhong R, Peng E, Li Z, et al. SPD-YOLOv8: a small-size object detection model of UAV imagery in complex scene[J]. The Journal of Supercomputing, 2024, 80(12): 17021-17041.

[13] Xiao J, Guo H, Zhou J, et al. Tiny object detection with context enhancement and feature purification[J]. Expert Systems with Applications, 2023, 211: 118665.

[14] 刘洪江, 王懋, 刘丽华, 等. 基于深度学习的小目标检测综述[J]. 计算机工程与科学, 2021, 43(08): 1429.

[15] 张阳婷, 黄德启, 王东伟, 等. 基于深度学习的目标检测算法研究与应用综述[J]. Journal of Computer Engineering & Applications, 2023, 59(18).

[16] Xu S, Zhang M, Song W, et al. A systematic review and analysis of deep learning-based underwater object detection[J]. Neurocomputing, 2023, 527: 204-232.

[17] Mizuno K, Terachi Y, Takagi K, et al. Architectural study of HOG feature extraction processor for real-time object detection[C]//2012 IEEE Workshop on Signal Processing Systems. IEEE, 2012: 197-202.

[18] Piccinini P, Prati A, Cucchiara R. Real-time object detection and localization with SIFT-based clustering[J]. Image and Vision Computing, 2012, 30(8): 573-587.

[19] 李寰宇, 毕笃彦, 杨源, 等. 基于深度特征表达与学习的视觉跟踪算法研究[J]. 电子与信息学报, 2015, 37(9): 2033-2039.

[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(6): 34-42.

[21] 陈旭, 彭冬亮, 谷雨. 基于改进 YOLOv5s 的无人机图像实时目标检测[J]. 光电工程, 2022, 49(3): 210372-1-210372-13.

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

[23] 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, 2016, 39(6): 1137-1149.

[24] Jiang P, Ergu D, Liu F, et al. A Review of Yolo algorithm developments[J]. Procedia computer science, 2022, 199: 1066-1073.

[25] Tripathy S, Satpathy M. SSD internal cache management policies: A survey[J]. Journal of Systems Architecture, 2022, 122: 102334.

[26] Tong K, Wu Y. Deep learning-based detection from the perspective of small or tiny objects: A survey[J]. Image and Vision Computing, 2022, 123: 104471.

[27] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[28] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9627-9636.

[29] Duan K, Bai S, Xie L, et al. Centernet: Keypoint triplets for object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6569-6578.

[30] Cai Z, Fan Q, Feris R S, et al. A unified multi-scale deep convolutional neural network for fast object detection[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016,

Proceedings, Part IV 14. Springer International Publishing, 2016: 354-370.

[31] Li Y, Miao N, Ma L, et al. Transformer for object detection: Review and benchmark[J]. Engineering Applications of Artificial Intelligence, 2023, 126: 107021.

[32] Shi H, Zhou Q, Ni Y, et al. DPNET: dual-path network for efficient object detection with lightweight self-attention[C]//2022 IEEE international conference on image processing (ICIP). IEEE, 2022: 771-775.

[33] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229.

[34] Zhu X, Su W, Lu L, et al. Deformable detr: Deformable transformers for end-to-end object detection[J]. arXiv preprint arXiv:2010.04159, 2020.

[35] Ghiasi G, Lin T Y, Le Q V. Nas-fpn: Learning scalable feature pyramid architecture for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7036-7045.

[36] Wang N, Gao Y, Chen H, et al. NAS-FCOS: efficient search for object detection architectures[J]. International Journal of Computer Vision, 2021, 129: 3299-3312.

[37] Wang S. An augmentation small object detection method based on NAS-FPN[C]//2020 7th International Conference on Information Science and Control Engineering (ICISCE). IEEE, 2020: 213-218.

[38] 丁丁, 刘文哲, 盛常冲, 等. 神经网络架构搜索研究进展与展望[J]. 国防科技大学学报, 2024, 45(6): 100-131.

[39] Gao Y, Yang H, Zhang P, et al. Graph neural architecture search[C]//International joint conference on artificial intelligence. International Joint Conference on Artificial Intelligence, 2021:1403-1409.

[40] Jaafra Y, Laurent J L, Deruyver A, et al. Reinforcement learning for neural architecture search: A review[J]. Image and Vision Computing, 2019, 89: 57-66.

[41] Balaprakash P, Egele R, Salim M, et al. Scalable reinforcement-learning-based neural architecture search for cancer deep learning research[C]//Proceedings of the international conference for high performance computing, networking, storage and analysis. 2019: 1-33.

[42] Liu Y, Sun Y, Xue B, et al. A survey on evolutionary neural architecture search[J]. IEEE transactions on neural networks and learning systems, 2021, 34(2): 550-570.

[43] Yang Z, Wang Y, Chen X, et al. Cars: Continuous evolution for efficient neural architecture search[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1829-1838.

[44] Santra S, Hsieh J W, Lin C F. Gradient descent effects on differential neural architecture search: A survey[J]. IEEE Access, 2021, 9: 89602-89618.

[45] Akimoto Y, Shirakawa S, Yoshinari N, et al. Adaptive stochastic natural gradient method for one-shot neural architecture search[C]//International Conference on Machine Learning. PMLR, 2019: 171-180.

[46] Liu H, Simonyan K, Yang Y. Darts: Differentiable architecture search[J]. arXiv preprint arXiv:1806.09055, 2018.

[47] Pham H, Guan M, Zoph B, et al. Efficient neural architecture search via parameters sharing[C]//International conference on machine learning. PMLR, 2018: 4095-4104.

[48] Baymurzina D, Golikov E, Burtsev M. A review of neural architecture search[J]. Neurocomputing, 2022, 474: 82-93.

[49] Xue Y, Chen C, Słowik A. Neural architecture search based on a multi-objective evolutionary algorithm with probability stack[J]. IEEE Transactions

on Evolutionary Computation, 2023, 27(4): 778-786.

[50] Xie T, Dai K, Sun Q, et al. CO-Net++: A Cohesive Network for Multiple Point Cloud Tasks at Once With Two-Stage Feature Rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46: 10911-10928.

[51] Xiong Y, Liu H, Gupta S, et al. Mobiledets: Searching for object detection architectures for mobile accelerators[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 3825-3834.

[52] Guo J, Han K, Wang Y, et al. Hit-detector: Hierarchical trinity architecture search for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11405-11414.

[53] Su X, You S, Wang F, et al. Bcnet: Searching for network width with bilaterally coupled network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2175-2184.

[54] Xu X, Chen J, Li Z, et al. Towards efficient filter pruning via adaptive automatic structure search[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108398.

[55] Luo W, Xing J, Milan A, et al. Multiple object tracking: A literature review[J]. Artificial intelligence, 2021, 293: 103448.

[56] 李玺, 查宇飞, 张天柱, 等. 深度学习的目标跟踪算法综述[J]. 中国图象图形学报, 2019, 24(12): 2057-2080.

[57] 韩瑞泽, 冯伟, 郭青, 等. 视频单目标跟踪研究进展综述[J]. 计算机学报, 2022, 45(9): 1877-1907.

[58] 孟琭, 杨旭. 目标跟踪算法综述[J]. 自动化学报, 2019, 45(7): 1244-1260.

[59] 卢湖川, 李佩霞, 王栋. 目标跟踪算法综述[J]. 模式识别与人工智能, 2018, 31(1): 61-76.

[60] Hassan S, Mujtaba G, Rajput A, et al. Multi-object tracking: a systematic literature review[J]. Multimedia Tools and Applications, 2024, 83(14): 43439-43492.

[61] Zheng L, Tang M, Chen Y, et al. Improving multiple object tracking with single object tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 2453-2462.

[62] Yan Y, Huo W, Ou J, et al. Improved SiamFC Object Tracking Algorithm Based on Anti-Interference Module[J]. Journal of Sensors, 2022, 2022(1): 2804114.

[63] 金久才, 张家林, 刘德庆, 等. 基于改进 SiamRPN 的高速无人船海上目标视觉跟踪研究[J]. 海洋科学进展, 2024, 42(3): 545-554.

[64] Huang Z, Zhao H, Zhan J, et al. A multivariate intersection over union of SiamRPN network for visual tracking[J]. The Visual Computer, 2022, 38(8): 2739-2750.

[65] Kugarajeevan J, Kokul T, Ramanan A, et al. Transformers in single object tracking: An experimental survey[J]. IEEE Access, 2023, 11: 80297-80326.

[66] Meinhardt T, Kirillov A, Leal-Taixe L, et al. Trackformer: Multi-object tracking with transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 8844-8854.

[67] Zeng F, Dong B, Zhang Y, et al. Motr: End-to-end multiple-object tracking with transformer[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 659-675.

[68] Zhang Y, Sun P, Jiang Y, et al. Bytetrack: Multi-object tracking by associating every detection box[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 1-21.

[69] Sun Z, Chen J, Chao L, et al. A survey of multiple pedestrian tracking based

on tracking-by-detection framework[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(5): 1819-1833.

[70] Zheng C, Yan X, Gao J, et al. Box-aware feature enhancement for single object tracking on point clouds[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13199-13208.

[71] 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.

[72] 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.

[73] 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.

[74] Fischer T, Huang T E, Pang J, et al. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15380-15393.

[75] Zhou X, Koltun V, Krähenbühl P. Tracking objects as points[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 474-490.

[76] 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.

[77] Tsai C Y, Su Y K. MobileNet-JDE: a lightweight multi-object tracking model for embedded systems[J]. Multimedia Tools and Applications, 2022, 81(7): 9915-9937.

[78] Baker B, Gupta O, Naik N, et al. Designing neural network architectures using reinforcement learning[J]. arXiv preprint arXiv:1611.02167, 2016.

[79] Kim Y, Li Y, Park H, et al. Neural architecture search for spiking neural networks[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 36-56.

[80] Guo Z, Zhang X, Mu H, et al. Single path one-shot neural architecture search with uniform sampling[C]//Computer vision–ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XVI 16. Springer International Publishing, 2020: 544-560.

[81] Chen M, Peng H, Fu J, et al. Autoformer: Searching transformers for visual recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12270-12280.

[82] Fischer T, Huang T E, Pang J, et al. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15380-15393.

[83] Guo S, Wang Y, Li Q, et al. Dmcp: Differentiable markov channel pruning for neural networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1539-1547.

[84] Yu J, Huang T. Autoslim: Towards one-shot architecture search for channel numbers[J]. arXiv preprint arXiv:1903.11728, 2019.

[85] Luo J H, Wu J. Autopruner: An end-to-end trainable filter pruning method for efficient deep model inference[J]. Pattern Recognition, 2020, 107: 107461.

[86] Hu W, Che Z, Liu N, et al. CATRO: Channel Pruning via Class-Aware Trace Ratio Optimization[J]. IEEE transactions on neural networks and learning systems, 2024, 35(8): 11595-11607.

[87] Hou Y, Ma Z, Liu C, et al. Network pruning via resource reallocation[J].

Pattern Recognition, 2024, 145: 109886.

攻读硕士学位期间取得的成果

研究生期间发表的论文：

[1] Zhang Z, Cao C, Zheng F, et al. GCPNet: Gradient-aware channel pruning network with bilateral coupled sampling strategy[J]. Expert Systems with Applications, 2025, 266: 126104. （SCI Q1, 中科院一区, 第一作者）

[2] Zhang Z, Cao C, Zheng F. CTTrack: Leveraging Cross-Transformer for Multi-Scale Feature Fusion in Multi-Object Tracking[C]//2024 5th International Conference on Computers and Artificial Intelligence Technology (CAIT). IEEE, 2024: 158-164. （EI 会议, 第一作者）

研究生期间发布的发明专利：

1、一种加快网络收敛的基于符号的梯度归一化方法（实质审查、排名第一）

致谢

时光荏苒，白驹过隙。三年的硕士研究生生活即将画上句号，回首求学与科研的这段时光，既有追求学术道路上的艰辛与挑战，也有收获知识与成长的喜悦与满足。论文的完成不仅是我个人努力的结果，更离不开众多师长、同学、朋友以及家人的支持与帮助。在此，我怀着无比感激的心情，向所有在我学习与生活中给予帮助的人致以最诚挚的谢意。

首先，我要衷心感谢我的导师曹雏清老师。曹雏清老师渊博的学识、严谨的治学态度和一丝不苟的科研作风一直深深感染着我。从课题的选题到实验设计，从论文的撰写到最终的定稿，曹老师始终给予我悉心的指导与耐心的点拨。在遇到科研困惑与学术瓶颈时，导师总是耐心倾听，及时给予鼓励与建议，使我在科研道路上不断前行。导师不仅是我学术上的引路人，更是我人生道路上的榜样，曹老师以实际行动诠释了作为一名学者的责任与担当，让我深深敬佩与钦佩。

其次，我要衷心感谢课题组的郑方军同学，在实验方案设计、数据分析等方面给予了我无私的支持与耐心的帮助。与同门师兄弟们一起在实验室讨论问题、分享心得的日子，成为我研究生生活中最珍贵、最难忘的回忆。同时，我要特别感谢哈尔滨工业大学的谢涛博士和戴崑博士，感谢他们在科研道路上对我的悉心指导与宝贵帮助。他们严谨的学术态度、扎实的专业素养和耐心的教诲，不仅拓展了我的学术视野，更激励我在科研探索中不断前行、勇攀高峰。

我要感谢我的家人和女朋友。感谢父母无私的爱与默默的支持，是您们用无尽的关怀、包容和鼓励为我撑起一片温暖的天空。无论我身处何方，您们永远是我最坚强的后盾。感谢毛毛在我学业与人生道路上的理解、鼓励和支持，让我有勇气不断追求自己的梦想，在遇到困难与挫折时，及时给予我鼓励让我重拾信心。

我还要感谢论文评审和答辩委员会的各位专家老师，感谢您们在百忙之中审阅我的论文，并提出宝贵的修改意见。老师们严谨细致的学术态度让我受益匪浅，您的建议和意见不仅完善了我的论文，更提升了我的学术视野和研究能力。

回顾整个研究生阶段，有过欢笑，有过泪水，有过迷茫，也有过收获。感谢这段宝贵的时光，感谢所有遇见的人。论文的完成并不代表学习的结束，而是新的起点。未来无论走向何方，我都将以更加严谨、勤奋、踏实的态度，继续努力，不负青春，不负韶华。