分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220320183

安徽工程大学
Anhui Polytechnic University

硕士学位论文

题目　**基于轻量化Transformer模型的高速目标检测与跟踪算法研究**

论 文 作 者　包俊涛
校 内 导 师　刘丙友
校 外 导 师　朱标
专业学位类别　电子信息
研究方向（领域）　新一代电子信息技术

论文提交日期：2025 年 5 月 16 日

分类号：TP391　　单位代码：10363
密　级：公开　　学　　号：2220320183

题　目　基于轻量化 Transformer 模型的高速目标检测与跟踪算法研究

题　目　**Research on High-Speed Object Detection and Tracking Algorithms Based on Lightweight Transformer Models**

学 生 姓 名：包俊涛
校 内 导 师：刘丙友
校 外 导 师：朱标
专业学位类别：电子信息
研究方向（领域）：新一代电子信息技术

论文答辩日期：2025 年 5 月 6 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结果由本人承担。

学位论文作者签名：

日期：2025 年 5 月 16 日

安徽工程大学学位论文版权使用授权书

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

保密 ☐，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：　　指导教师签名：

日期： 2025 年 5 月 16 日　　日期：2025 年 5 月 16 日

基于轻量化 Transformer 模型的高速目标检测与跟踪算法研究

摘 要

在高速运动场景下，目标的实时精准检测与稳定跟踪是计算机视觉领域中的一项重要且具有挑战性的任务。高速运动往往伴随目标的剧烈位移、形变以及环境的快速变化，这使得现有的目标检测和跟踪方法在处理此类问题时面临诸多困难。尤其在自动驾驶、无人机飞行、视频监控等应用中，如何在复杂的动态环境中准确识别并持续跟踪高速运动目标，已成为技术实现的瓶颈。针对这一问题，本文提出并优化了一种基于轻量化变换器（Transformer）的高速目标检测与跟踪框架。具体研究内容如下：

（1）在目标跟踪阶段，提出了一种名为纳米移动视觉变换器（Nano Mobile Vision Transformer，NanoMViT）的轻量级模型，结合了孪生网络与轻量 Transformer，并设计了检测-跟踪一体化协同策略。这一策略能够有效应对目标遮挡及外观剧烈变化，迅速恢复跟踪。实验结果表明，在视觉目标追踪 2018（Visual Object Tracking 2018，VOT2018）、大规模单目标追踪（Large-scale Single Object Tracking，LaSOT）和无人机跟踪基准数据集（DTB70 Drone Tracking Benchmark，DTB70）等数据集上，NanoMViT 与多种主流跟踪算法相比，平均准确率提高了 1~2 个百分点，且在遮挡场景中的鲁棒性显著增强。同时，整体模型的参数量减少了 50%以上。

（2）针对小目标特征易丢失及高帧率下的计算瓶颈问题，本文首先在目标检测阶段引入移动变换器（Mobile vision Transformer，MobileViT）并结合多种注意力机制，通过剪枝、量化和知识蒸馏等技术对大型模型进行高效压缩，提出了基于场景复杂度的自适应移动视觉变换器(adaptive scene complexity 一 based mobile vision

transformer，AC-MViT）。实验结果表明，在上下文常见对象2017(common objects in context 2017，COCO 2017)数据集和模式分析、静态建模和计算学习视觉对象类 2012（pattern analysis，statical modeling and computational learning visual object classes 2012，PASCAL VOC 2012）数据集上，AC-MViT 与原始 MobileViT 相比，在保持检测精度基本不变的同时，浮点运算量降低了约 40%~50%，推理速度提升约 30%。

（3）针对高速摄影下目标检测与跟踪中的运动模糊问题，本文提出了一种新型训练策略。该策略专注于有效应对高速运动所带来的运动模糊现象。通过检测高速运动物体的运动模糊现象来达到快速跟踪的效果。结果表明，所提出的方法在实时性、精度和资源效率等方面表现出显著优势。

关键词：高速目标检测与跟踪，Transformer，MobileViT，自适应剪枝与量化，高速相机

RESEARCH ON HIGH-SPEED OBJECT DETECTION AND TRACKING ALGORITHMS BASED ON A LIGHTWEIGHT TRANSFORMER MODEL

ABSTRACT

In high-speed motion scenarios, real-time and accurate detection and stable tracking of targets is an important and challenging task in the field of computer vision. High-speed motion often involves significant displacement, deformation of the target, and rapid changes in the environment, which makes existing object detection and tracking methods face many difficulties when dealing with such problems. Especially in applications such as autonomous driving, drone flight, and video surveillance, accurately identifying and continuously tracking high-speed moving targets in complex dynamic environments has become a technological bottleneck. To address this issue, this paper proposes and optimizes a high-speed target detection and tracking framework based on a lightweight transformer (Transformer). The specific research content is as follows:

（1）In the target tracking stage, a lightweight model called Nano Mobile Vision Transformer (NanoMViT) is proposed. This model combines a Siamese network with a lightweight transformer and designs a detection-tracking integrated collaborative strategy. This strategy can effectively handle target occlusion and significant appearance changes, rapidly restoring tracking. Experimental results show that, on datasets such as Visual Object Tracking 2018 (VOT2018), Large-scale Single Object Tracking (LaSOT), and A Benchmark and DTB70 Drone Tracking Benchmark (DTB70), NanoMViT improves the average accuracy by 1-2 percentage points compared to several mainstream tracking algorithms and

significantly enhances robustness in occlusion scenarios. At the same time, the overall model's parameter count is reduced by more than 50%.

（2）To address the problem of feature loss of small targets and the computational bottleneck at high frame rates, this paper first introduces a Mobile Vision Transformer (MobileViT) in the object detection stage, and combines multiple attention mechanisms. Through techniques such as pruning, quantization, and knowledge distillation, large models are efficiently compressed. The paper proposes an Adaptive Scene Complexity-based Mobile Vision Transformer (AC-MViT). Experimental results show that, on the Common Objects in Context 2017 (COCO 2017) dataset and the Pattern Analysis, Statistical Modeling, and Computational Learning Visual Object Classes 2012 (PASCAL VOC 2012) dataset, AC-MViT reduces the floating-point operations by about 40%~50% and improves inference speed by about 30%, while maintaining detection accuracy almost unchanged compared to the original MobileViT.

（3）To address the motion blur issue in object detection and tracking under high-speed photography, this paper proposes a novel training strategy specifically designed to effectively handle motion blur caused by high-speed movements. The strategy achieves rapid tracking by detecting motion blur phenomena in high-speed moving objects. Experimental results demonstrate that the proposed method exhibits significant advantages in real-time performance, accuracy, and resource efficiency.

KEY WORDS: High-speed target detection and tracking, Transformer, MobileViT, Adaptive pruning and quantization, High-speed camera

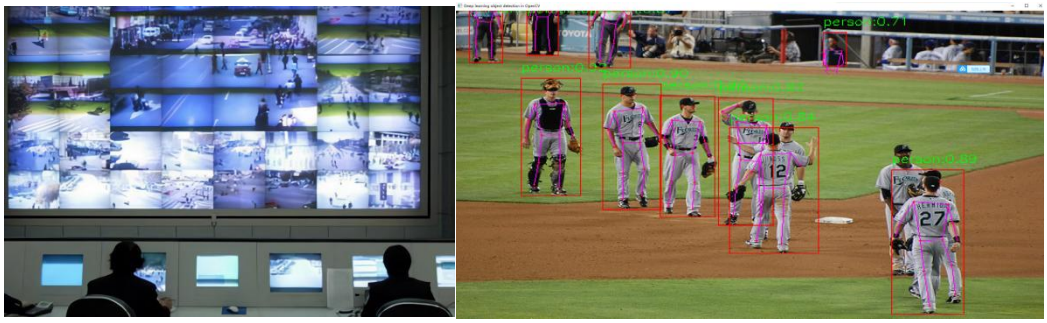
目录

第 1 章 绪论

1.1 课题研究背景和意义

目标检测与跟踪是计算机视觉领域的重要研究方向，旨在从复杂场景中自动识别并定位目标，同时对其进行时序跟踪。这一技术在自动驾驶[1]、智能安防[2]、无人机导航[3]、工业自动化[4]、体育赛事分析[5]等领域具有广泛的应用，如图 1-1 所示。例如，在无人机自主导航中，目标检测与跟踪技术用于实时定位和追踪动态障碍物；在智能安防系统中，通过检测和跟踪可疑人员或车辆的运动轨迹，可以提高事件预警的精准性。然而，高速目标检测与跟踪算法在实际应用中面临多重挑战：一方面，目标的运动速度快、轨迹多变，高速运动时会有残影和拖影；另一方面，实时性要求算法能够在有限的计算资源下完成高精度的检测与跟踪，这对算法的效率和性能提出了更高的要求。

(a)自动驾驶　　(b)无人机导航



(c)智能安防　　(d)体育赛事分析

图 1-1 目标检测跟踪主要研究方向

近年来，深度学习技术的快速发展极大地推动了目标检测与跟踪算法的进步。基于卷积神经网络（Convolutional Neural Networks，CNN）的目标检测方法如更快的区域卷积神经网络（Faster Region-based Convolutional Neural

Network ，Faster R-CNN[6]）和你只看一次（You Only Look Once ，YOLO）系列[7]在精度和速度方面取得了显著成效，而基于孪生网络（Siamese Network，Siamese）的目标跟踪方法如孪生全卷积网络（Siamese Fully Convolutional Network， SiamFC[8]）和孪生区域提议网络（Siamese Region Proposal Network，SiamRPN[9]）也展现了其在高效性上的优势。然而，这些方法在面对高速目标时，仍然存在检测与跟踪精度不足、时延过高等问题，特别是在嵌入式设备或边缘计算平台上。因此，如何在保证检测与跟踪性能的前提下，降低模型复杂度，优化计算效率，成为当前研究的重要方向之一。

相较于传统的 CNN 模型，变换器（Transformer）模型是一种以注意力机制为核心的深度学习架构，Transformer 通过自注意力机制实现了对全局特征的高效建模，能够更好地捕捉目标间的长距离依赖关系。在目标检测领域，检测变换器（Detection Transformer，DETR [10]）模型实现了端到端的检测框架；在目标跟踪领域，跟踪变换器（TrackFormer[11]）模型展现了其强大的建模能力。然而，这些基于 Transformer 的模型普遍存在参数量大、计算复杂度高的问题，限制了其在资源受限场景下的应用。为了弥补这一不足，将量化，剪枝等轻量化[12]技术与 Transformer 模型结合，设计适用于高速目标检测与跟踪任务的高效算法，成为当前研究的一个热点。

基于此，本文以高速目标检测与跟踪的核心，围绕轻量化 Transformer 模型的设计展开探索，旨在解决高速场景中目标检测与跟踪任务的实时性与高效性难题。通过采用剪枝、量化等轻量化技术，引入动态卷积[13]与多注意力机制[14]优化模型性能。本研究为资源受限环境下的高速目标检测与跟踪提供理论支持与技术实现，具有重要的学术价值和工程意义。

1.2 国内外研究现状

1.2.1 目标检测领域研究现状

目标检测技术的发展经历了从传统方法到深度学习方法的不断演进。如图 1-2 所示。

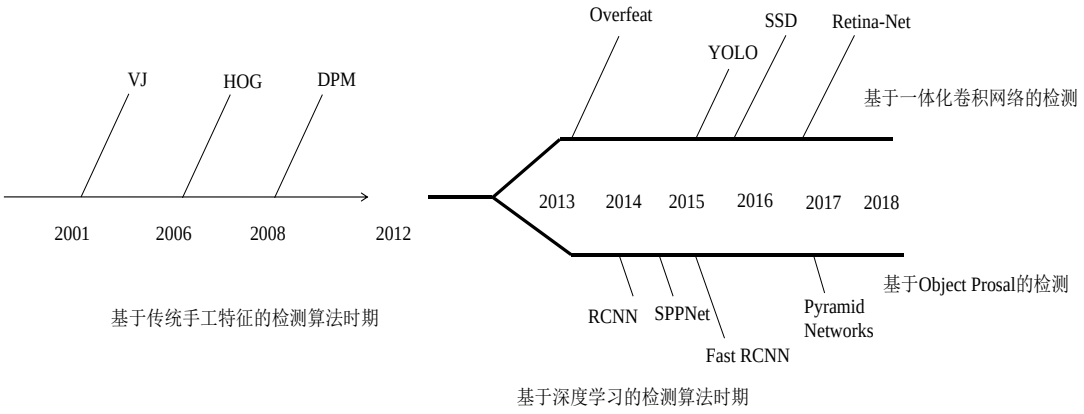


图 1-2 目标检测领域到深度学习方法的时间轴

在早期，基于人工特征的目标检测方法在一定程度上解决了特定场景下的目标检测问题，但由于对复杂场景的适应性差，其性能逐渐被深度学习方法取代。随着深度学习的兴起，基于卷积神经网络的目标检测方法成为研究热点，其中两阶段检测方法和单阶段检测方法单次多框检测器（Single Shot MultiBox Detector ，SSD[15]）取得了广泛应用。两阶段检测方法以较高的检测精度著称，但其计算复杂度较高，实时性较差；单阶段检测方法则以速度快见长，适合实时性要求高的场景。Transformer 模型被引入到目标检测领域，开启了新一轮的研究热潮，DETR 首次采用端到端的目标检测框架，通过自注意力机制对目标进行高效建模，避免了传统检测方法中依赖锚框的设计。然而，DETR 的收敛速度较慢，且在小目标检测和复杂场景下表现有限。因此，一些改进模型 MixFormer[16]通过引入局部注意力机制或多尺度特征融合，进一步提升了检测性能。如图 1-3 所示。

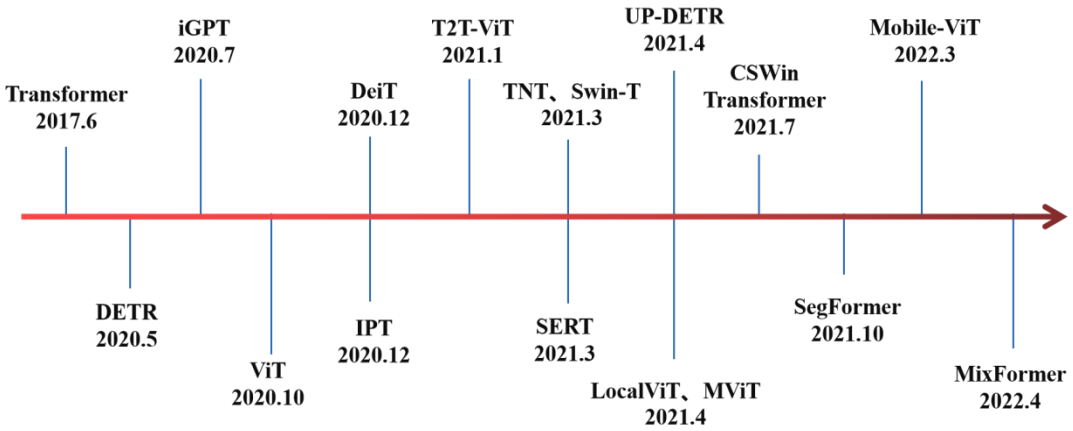


图 1-3 Transformer 模型应用时间轴

1.2.2 目标跟踪领域研究现状

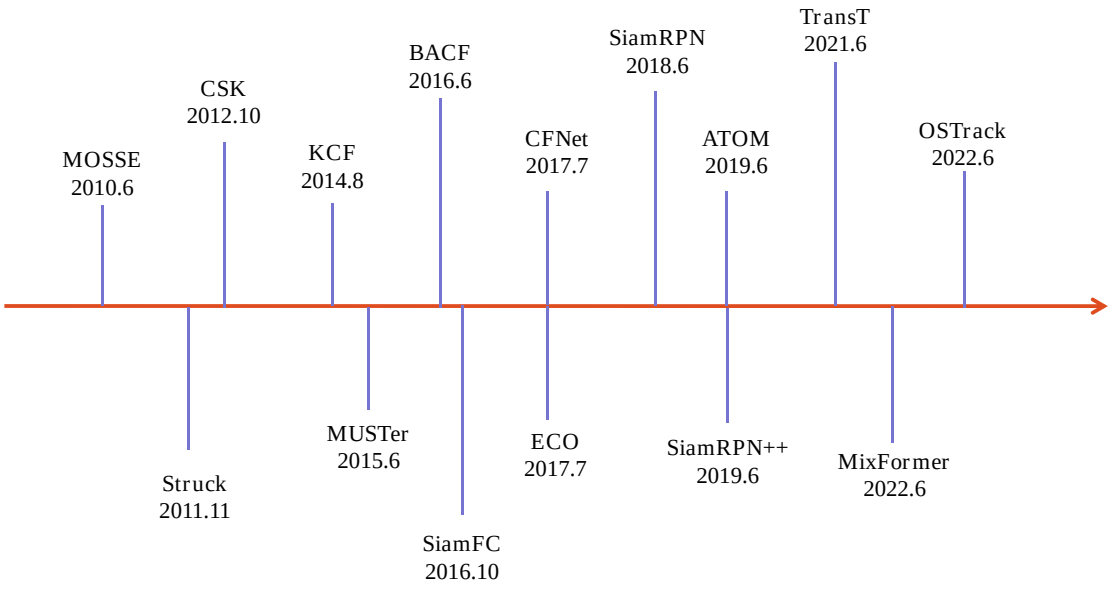


图 1-4 目标跟踪领域研究现状时间轴

目标跟踪作为计算机视觉的重要任务之一，旨在在视频序列中对指定目标进行时序定位。核相关滤波器（Kernelized Correlation Filters ，KCF[18]）和通道与空间可靠性跟踪（Channel and Spatial Reliability Tracking ，CSRT[19]）等传统滤波算法依赖于手工设计的特征和滤波器，对环境变化的鲁棒性较差。近年来，基于深度学习的目标跟踪方法得到了快速发展，其中 SiamFC、SiamRPN 等基于 Siamese 网络的跟踪算法通过构建相似性度量机制，实现了高效的跟踪性能，这些跟踪算法时间如图 1-4 所示。

这些方法在平衡速度与精度方面表现优异，但在复杂场景中，仍可能出现目标漂移或跟踪失败的问题。Transformer 的引入为目标跟踪任务提供了新的思路。TrackFormer 利用自注意力机制实现了对目标的长时间依赖建模，有效提高了跟踪的鲁棒性和精度。此外，基于变换器的运动跟踪器（Transformer Meets Tracker，TMT）[20]等方法通过融合时序特征和空间特征，进一步提升了跟踪性能。然而，与目标检测领域类似，基于 Transformer 的跟踪方法也面临计算复杂度高、模型参数量大的问题，因此难以直接应用于高速场景。

1.2.3 轻量化模型相关研究现状

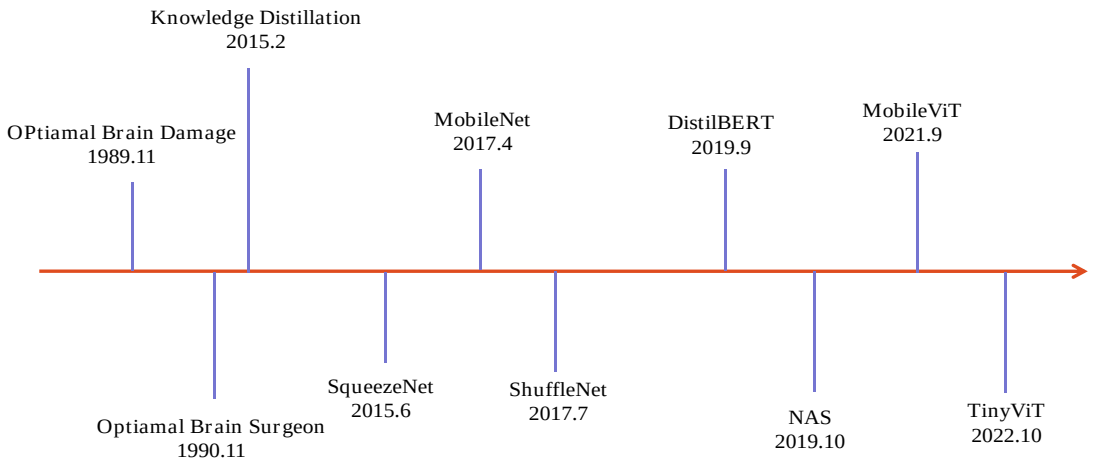


图 1-5 轻量化技术时间轴

轻量化模型的研究旨在通过优化深度学习模型的结构和参数，提高其计算效率和资源利用率。目前，主要的轻量化技术包括模型剪枝、权重量化、知识蒸馏以及架构搜索等，这些轻量化技术具体如图 1-5 所示。

近年来，轻量化 Transformer 模型的研究逐渐兴起。MobileViT 作为一种轻量级 Transformer 模型，通过引入局部注意力机制和高效卷积模块，在保持精度的同时显著降低了模型复杂度。此外，TinyViT[21]等模型进一步探索了 Transformer 的轻量化设计，展现了其在资源受限环境下的应用潜力。当前轻量化 Transformer 在高速目标检测与跟踪任务中的应用研究仍处于起步阶段，相关技术难点还有待进一步突破。

1.3 研究内容与结构安排

1.3.1 主要研究内容

近年来，目标检测与跟踪技术在智能驾驶、无人机导航、视频监控等领域得到了广泛应用。而在高速运动场景下，目标的快速变化、遮挡以及复杂背景带来的干扰使得实时目标检测与跟踪任务面临挑战。传统的目标检测与跟踪算法在硬件资源有限的情况下难以同时保持高精度和实时性。本文以轻量化 Transformer 模型为基础，针对高速目标检测与跟踪任务展开研究，构建高效、高精度的目标检测与跟踪框架。

本文的主要研究内容包括以下几个方面：

（1）针对小目标特征易丢失及高帧率下的计算瓶颈问题，在目标检测阶段引入移动变换器（Mobile vision Transformer，MobileViT）并结合多种注意力机制，通过剪枝、量化和知识蒸馏等技术对大型模型进行高效压缩，提出了基于场景复杂度的自适应移动视觉变换器(adaptive scene complexity－based mobile vision transformer，AC-MViT），并通过实验证明其有效性。

（2）在目标跟踪阶段，提出了一种名为纳米移动视觉变换器（Nano Mobile Vision Transformer，NanoMViT）的轻量级模型，结合了孪生网络与轻量Transformer，并设计了检测-跟踪一体化协同策略。通过实验证明这一策略能够有效应对目标遮挡及外观剧烈变化，并迅速恢复跟踪。

（3）针对高速摄影下目标检测与跟踪中的运动模糊问题，本文提出了一种新型训练策略。该策略专注于有效应对高速运动所带来的运动模糊现象，并在高速相机平台上进行实验评估。

1.3.2 论文结构概述

本文围绕高速目标检测与跟踪展开了相关研究，提出了 NanoMViT 和 AC-MViT 两种算法。通过对比实验和消融实验，验证了这些算法的有效性。此外，还通过实际应用实验进一步验证了两者结合后的效果，如图 1-6 所示。

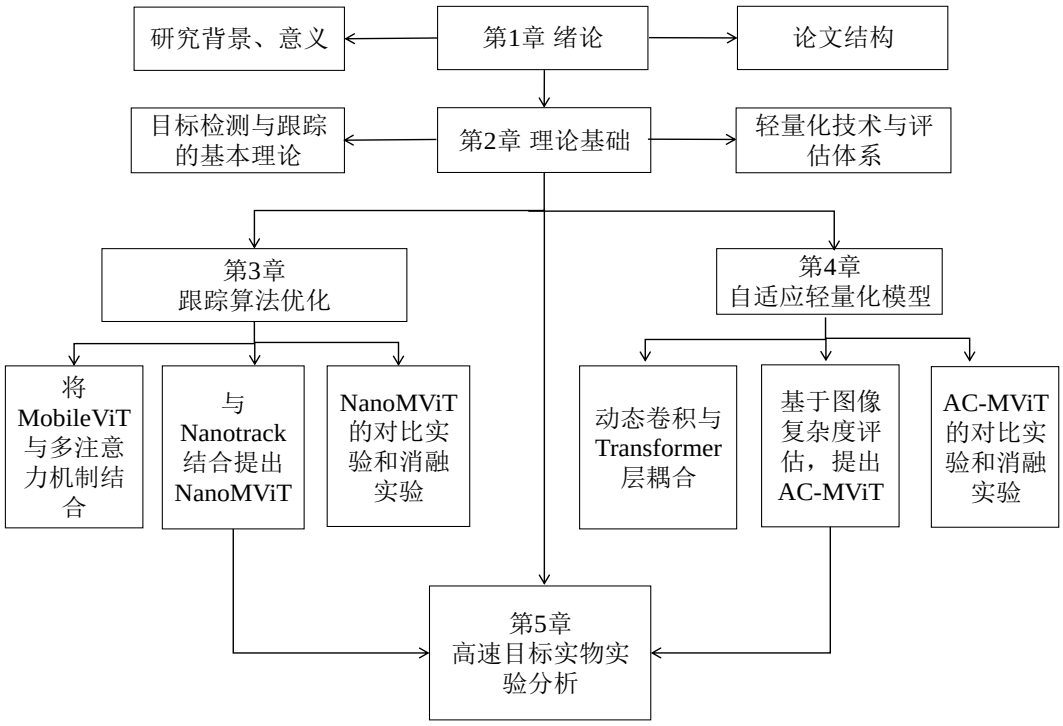


图 1-6 论文的主要框架

本文的具体安排如下：

第 1 章:阐述课题的研究背景与意义，综述国内外研究现状，明确研究目标和内容，并概述论文的整体结构。

第 2 章:介绍目标检测与跟踪的基本理论，Transformer 模型的原理及其在视觉任务中的应用，重点分析轻量化技术（包括剪枝、量化和知识蒸馏）的原理与优势。

第 3 章:结合轻量化 Transformer 模型，提出高速目标检测与跟踪算法框架 NanoMViT，解决高速场景下小目标检测和遮挡跟踪等问题，进行相关性能分析。

第 4 章:设计了 AC-MViT，具体阐述剪枝、量化和知识蒸馏等技术在模型优化中的应用，并通过实验验证其轻量化效果。

第 5 章:对本文提出的算法进行实物实验验证，并结合实际应用场景对算法的综合性能进行讨论。

第 6 章:总结本文的主要研究成果，指出研究中的不足，并提出未来的研究方向。

第 2 章 高速目标检测与跟踪的相关理论

2.1 目标检测理论方法

目标检测在计算机视觉任务中占据核心地位，其任务是对图像中的目标进行定位与分类，典型输出形式通常包括边界框或分割掩膜[23]。随着深度学习的发展，基于 CNN 的目标检测算法在精度与效率上实现了很大的进步，逐渐取代了早期的基于人工特征与传统分类器的检测方法。

QI 最早的传统目标检测方法，如基于方向梯度直方图（Histogram of Oriented Gradients，HOG）特征与支持向量机（Support Vector Machine，SVM）分类器[24]的检测框架，主要依靠人工设计的特征进行目标定位。此类方法在固定场景下或面对特定类型的目标时能够取得一定效果，但对复杂背景、光照变化和尺度多变等问题的适应性较差。另一类较为经典的方法是可变形部件模型（Deformable Parts Model，DPM）[25]，其核心思想是通过可变形部件的组合来描述目标形状，克服了一定程度的目标变形问题。然而，这类方法依然需要较多的先验知识来设计特征，并且计算开销相对较大，当目标运动速度较快或场景复杂度较高时，检测性能会明显下降。

2.2 高速目标检测跟踪流程设计

在高速运动场景中，目标与背景之间的相对位置变化频繁且幅度较大，尤其是在无人机航拍[27]或车辆高速行驶[28]的条件下，视觉数据帧率通常较高，目标在相邻帧间可能产生明显的位移或外观变化。同时，硬件平台往往具有算力受限、存储空间有限等特点，因此需要在算法架构层面进行针对性设计，以满足实时性和高精度的综合要求。

为了实现这一目标，高速目标检测与跟踪系统通常包含几个关键环节：图像获取与预处理、目标检测、目标跟踪以及融合与决策。在图像获取与预处理阶段，根据具体场景选择合适的摄像机或传感器非常重要，同时还需要确保图像数据的清晰度和高帧率。特别是在高速场景下，高帧率能够提供更多的运动信息，帮助减少由于快速运动造成的模糊。然而，高帧率会带来较大的数据量

和实时处理压力，因此，在图像预处理阶段，通常会通过基本的滤波、降噪、分辨率调整等操作来优化图像，为后续的检测和跟踪模块提供清晰、准确的输入数据。

接下来，在目标检测阶段，常见的方法可以分为一阶段检测和两阶段检测。针对高速场景，一阶段检测方法具有更高的推理速度，适用于实时性要求较高的应用，但其在小目标或复杂背景下的精度较低。相反，两阶段检测方法通常能提供更高的精度，但推理速度较慢，可能无法满足高速目标的实时处理需求。因此，选择合适的检测方法取决于对实时性和精度的不同需求。

近年来，Transformer 在视觉领域的应用逐渐深入，基于 Transformer 的目标检测模型展示了端到端训练和全局信息建模的优势。这些模型能够在处理复杂场景时发挥较大优势，但由于其较高的计算量和网络规模，仍然需要进一步进行轻量化和架构优化，以便在高速场景中大规模应用。而在目标跟踪阶段，系统需要根据检测到的目标在连续帧中的位置变化，预测或更新目标的最新位置和运动轨迹，以保持时间上的连续性和身份一致性。对于单目标跟踪而言，基于 Siamese 网络的纳米追踪（Nanotrack）[29]跟踪算法在速率和精度上表现出色，能够在接近实时的条件下完成大部分场景下的跟踪任务。面对高速运动带来的大量交互和遮挡，需要在检测与跟踪模块之间建立有效的信息流，当跟踪失效或出现身份切换时，可以通过检测结果及时进行纠正与重新初始化。具体架构如图 2-1 所示。

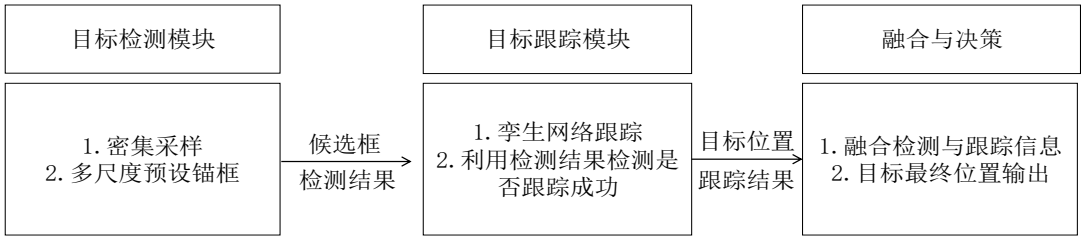


图 2-1 目标检测跟踪流程

2.3 高速目标跟踪理论核心与关键技术

目标跟踪是计算机视觉中继目标检测之后的另一项重要任务，旨在对指定目标在视频序列中的位置进行连续预测并保持身份一致性。与目标检测相比，目标跟踪更加强调利用目标在时间维度上的动态变化信息，并对外观变化、背

景干扰、部分遮挡等情况进行即时处理。在高速场景中，由于目标在相邻帧间的位置变动幅度大，给准确的跟踪带来了额外难度，也对算法的实时性提出了苛刻要求。

在深度学习普及之前，相关滤波类方法凭借在频域计算上的效率优势成为研究热点，如 KCF 和 CSRT 等算法能够在单目标跟踪下实现接近实时的效果。它们的基本思路是学习或在线更新一个与目标相关的滤波器，在搜索区域内对滤波响应进行匹配，从而得到目标的新位置。然而，此类算法往往依赖人工特征，对目标的急剧形变或复杂背景变化的适应性不足。此外，当目标被严重遮挡或发生尺度与外观的多重变化时，滤波器的判定效果会显著下降，导致跟踪失效或漂移。

基于深度学习的目标跟踪在近年来取得突破性进展，尤其是基于Siamese网络的跟踪算法通过将目标跟踪转化为相似性度量问题，在速度与精度之间达到了较好的平衡。SiamFC 开创性地利用端到端的孪生网络结构来学习模板分支与搜索分支之间的相关性，相比传统方法在跟踪精度和对外观变化的适应性上都有明显提升。后续发展出的 SiamRPN 和 SiamRPN++[30]在孪生网络基础上结合了区域提案网络，能进一步准确估计目标的边界框并适应其尺度变化，从而实现更高的跟踪精度。对于高速运动目标，这些方法在理论上可通过提升卷积网络的推理速度和增强滤波器的更新策略来满足实时需求，但在面对遮挡、光照变化和多目标交互时，跟踪仍可能出现明显的漂移。

Transformer在目标跟踪中的应用同样吸引了研究者的注意。TrackFormer 是一种将时序信息和空间信息统一建模的探索性方法，它通过在 Transformer 中对整个视频序列的特征进行自注意力计算，充分挖掘目标在时间和空间维度上的依赖关系，并借助序列中前后帧的目标特征关联来提高跟踪鲁棒性。在高速目标跟踪的核心环节中，时序特征的建模是算法的关键所在。由于目标在帧与帧之间可能存在剧烈的外观变化或位置偏移，如何充分利用历史帧的信息来对目标的运动趋势进行预测，从而在遮挡和缺失的情况下及时恢复跟踪，是待解决的问题之一。许多在线更新策略会在跟踪过程中实时更新目标模型，以便适应最新的外观变化，但过度更新可能导致模型过拟合于暂时的噪声或错误判断，更新不足则会使得模型无法及时学习到目标的真实变化。高速跟踪还需要处理

多目标间的遮挡与身份交换，只有当检测与跟踪模块有效配合，通过检测结果纠正跟踪漂移或重新初始化丢失目标，并利用跟踪维持目标的短时预测与识别，才能在整体系统层面提升效率与鲁棒性。

2.4 轻量化技术理论根源与实现途径

深度学习模型在视觉任务上取得了显著成功，但随之而来的问题是网络规模越来越庞大，参数量与计算量成倍增长，导致推理速度和部署成本上升，尤其在嵌入式或边缘计算平台上难以满足实时需求[31]。为此，各类轻量化技术的研究应运而生，其理论基础在于大量深度神经网络在参数冗余度方面存在共性，如果能够剔除或压缩不重要的权重，就可能在保证精度的同时显著降低模型复杂度。

在剪枝技术中，通过衡量网络中某些权重或滤波器的重要度，可以将那些影响较小的部分剪除，使得网络结构更紧凑。非结构化剪枝[32]主要将小权重直接置零，这种方式自由度高，但剪枝后的网络结构变得稀疏，普通硬件难以充分利用稀疏计算的优势。结构化剪枝则按通道或卷积核维度进行剪枝，网络形状保持规则，更便于直接获得加速效果。对 Transformer 而言，可以对注意力头、多层感知器单元或整个 Encoder/Decoder 层进行结构化缩减，既能降低计算量，也能显著减少内存占用。

量化技术则通过将原本 32 位或 16 位浮点权重转换为低比特宽度的定点数来减少计算和存储需求[33]。常见的量化包括 8 位定点或混合精度方案。合理的量化区间和零点能有效降低量化带来的信息损失。对于注意力机制较为复杂的 Transformer，需要研究是否对注意力权值进行量化处理，以及如何在多头注意力结构中分别设定量化精度，从而在压缩与精度之间取得良好平衡。

知识蒸馏是一种利用教师模型指导学生模型训练的思想，最初在分类任务中得到验证，后来逐渐扩展到检测、分割和跟踪等多种视觉任务[34]。其核心是让学生模型在学习训练集标签的同时，也对教师模型输出的特征或预测分布进行对齐，从而在模型体积更小的情况下取得接近教师模型的表现。对于 Transformer 模型，由于其内部层级结构深且注意力机制复杂，通过特征蒸馏可以让学生模型更好地学习到教师模型在序列编码过程中所捕捉的上下文信息。

值得注意的是，教师模型与学生模型的结构不一定完全相同，可以通过自适应的蒸馏策略使学生模型在保持核心注意力机制的情况下，对不重要的模块或网络层进行裁剪或替换。

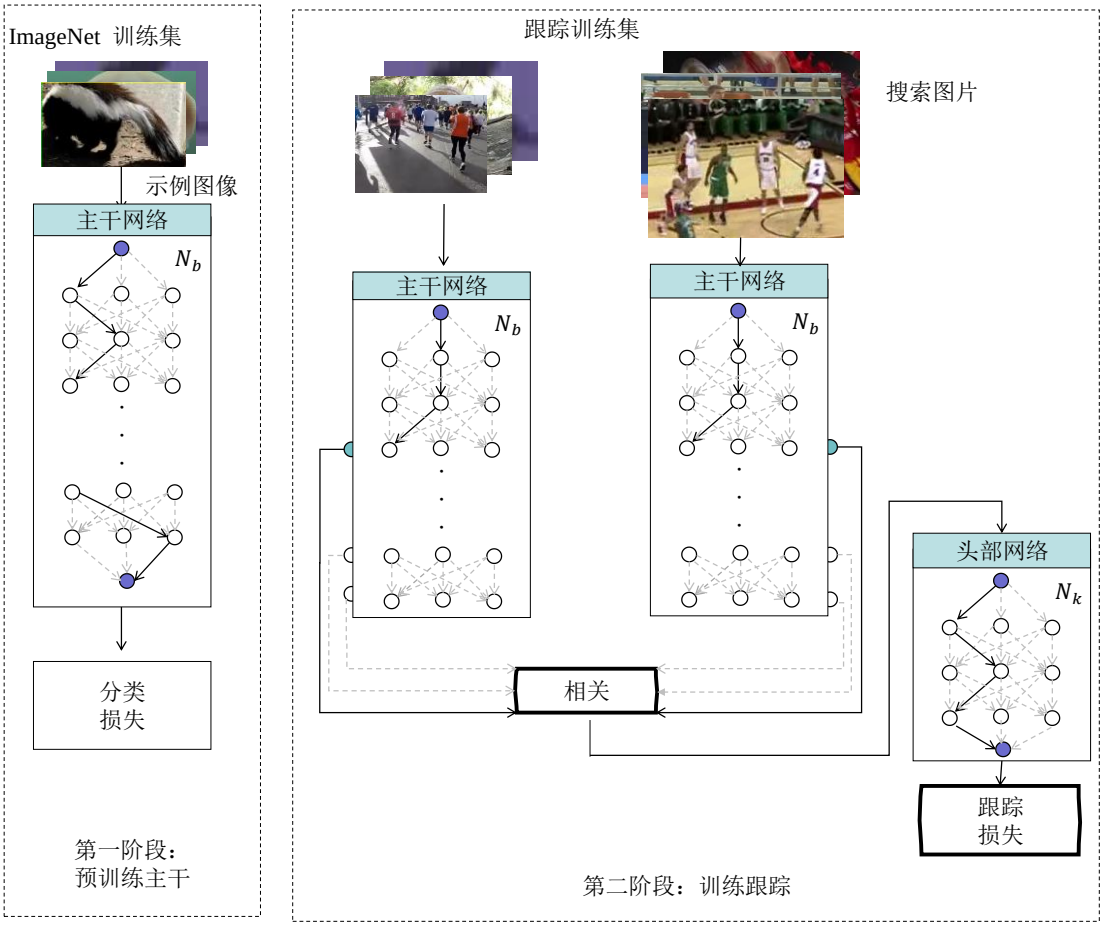


图 2-2 NAS 神经架构

在神经架构搜索（Neural Architecture Search，NAS）[35]中，研究者会预先设定一个包含各种算子与结构组合的巨大搜索空间，通过强化学习或演化算法对模型进行迭代搜索，直到找到最优的结构组合[36]。随着硬件对神经网络计算需求的约束越来越明显，NAS 也开始将延时、能耗和存储占用纳入搜索目标，从而自动生成兼顾准确率与效率的网络结构。目前，针对 Transformer 模型的 NAS 研究还在起步阶段，但已有一些轻量化 Transformer 的探索工作显示出良好的潜力，例如通过控制注意力头数量、缩减深度或引入高效卷积替代部分注意力单元等方式。NAS 神经架构如图 2-2 所示。

2.5 常用数据集及评估体系

高速目标检测与跟踪的研究与应用离不开数据集和评估指标体系的支撑。合适的数据集既能反映算法在实际环境中的表现，也能帮助研究者发现模型在极端情况下的缺陷与改进空间。对于评估体系，需要综合考虑检测与跟踪任务的精度、速度、鲁棒性，以及资源占用等因素。

2.5.1 数据集简介

COCO 2017[37]是一个广泛用于目标检测、实例分割和关键点检测的计算机视觉数据集。它包含超过 33 万张图像，涵盖 80 类物体，并提供边界框、分割掩码、关键点和图像描述等多种标注信息。该数据集的主要挑战包括目标尺度变化大、遮挡严重、场景复杂以及同类目标密集分布，因此在目标检测和跟踪任务中被广泛应用。

PASCAL VOC 2012[38]是一个经典的目标检测、图像分类和语义分割数据集，包含 11,530 张图像和 20 个类别的物体，标注信息包括边界框和分割掩码。该数据集的挑战在于目标种类丰富、姿态多变、存在遮挡以及复杂背景，使其成为深度学习早期目标检测研究的重要基准之一。

VOT2018[39]是一个专门用于视觉目标跟踪任务的数据集，包含 60 个视频序列，总计约 11,500 帧，每帧都提供目标的精确边界框。该数据集设计了专门的评测标准，主要挑战包括目标遮挡、快速运动、尺度变化和运动模糊，使其成为衡量目标跟踪算法鲁棒性的重要基准。

LaSOT[40]是一个大规模的单目标跟踪数据集，包含1,400个长序列视频，总帧数超过 300 万，每帧都提供目标的精确边界框。该数据集的主要特点是长时跟踪，目标往往会经历显著的运动变化、遮挡、背景干扰等复杂场景，旨在推动长期目标跟踪技术的发展，并提升模型在实际应用中的性能。

DTB70[41]数据集由德国航空航天中心于 2012 年发布，专为无人机视觉导航与目标检测研究而设计。它包含 70 个高分辨率视频序列，记录了多种环境条件下的目标跟踪任务。数据集中的视频涵盖了目标的快速移动、遮挡、旋转等复杂运动模式，旨在评估视觉跟踪算法在实际环境中的表现。DTB70 的标注信息详尽，提供了丰富的场景数据，用于训练和验证目标检测及跟踪模型。

2.5.2 评估指标体系的构建

在目标检测任务中，平均精度是最常用的综合指标，通常会在不同的 IoU 阈值下计算平均精度以评判算法的准确性。对于高速检测，则要同时考虑推理速度（如 FPS 或每帧处理时间），因为在高速运动环境中，哪怕精度稍有提升，如果以牺牲大量推理速度为代价，也难以满足实时需求。此外，检测算法的模型参数量和 FLOPs[41]也是评估体系的重要组成部分。

在单目标跟踪任务中，评价指标既包括与检测类似的定位精度，也包括对时间维度和身份管理的考核。对于单目标跟踪，常见指标包括成功率和精度，它们通过将预测框与真实框的重叠率或中心偏差[43]进行统计分析，能直观地反映算法在不同场景下的跟踪效果。

除了精度和跟踪一致性，实时性和资源效率同样是高速场景下的重要维度。在实际工程应用中，如果检测和跟踪模块的处理帧率低于视频输入帧率，系统就无法实时输出目标的最新位置，导致跟踪滞后甚至无法及时应对突发情况。因此，研究者通常会在报告实验结果时同时给出 FPS 或延时数据。对于轻量化后的模型，除了 FPS 以外，参数量和模型大小也会被列入考虑，因为这直接关系到在边缘端设备或无人机载[44]计算平台上的部署可行性。只有在精度和实时性之间达到平衡，同时满足部署硬件的资源约束，目标检测与跟踪算法才能真正具备在高速环境下落地应用的潜力。

通过在不同场景和数据集上的实验对比，可以发现轻量化 Transformer 模型在哪些方面优于传统算法，在哪些问题场景下仍需进一步优化。若发现小目标检测率或对遮挡的适应性仍较弱，则可在多尺度特征融合或时序信息建模[45]方面进行进一步改进。

2.6 本章小结

本章围绕高速目标检测与跟踪的系统架构与理论基础展开了详细论述。首先从系统的整体流程和硬件条件出发，阐明了高速场景下在实时性与精度方面的双重诉求，以及需要在检测、跟踪和融合决策模块之间建立高效协作的必要性。随后介绍了目标检测的理论基础与技术细节，从传统基于人工特征的方法到深度学习的两大流派，再到近年来崭露头角的 Transformer 检测模型，并指出

了高速场景下对多尺度特征提取与小目标检测的特殊需求。接着重点讨论了高速目标跟踪的理论核心，包括传统相关滤波方法和基于深度学习的跟踪框架，以及 Transformer 在跟踪任务中的应用潜力。此后，结合剪枝、量化、知识蒸馏和神经架构搜索等轻量化技术，分析了其内在理论依据与实现途径，并强调了在高速场景中的模型压缩需求。最后，对常用的高速目标数据集及其特性进行了概述，总结了在检测与跟踪中最常使用的评价指标与评估体系，为下一步的模型设计与实验测试提供了重要参考。

第 3 章 基于 Transformer 模型的检测跟踪算法优化

3.1 基于 MobileViT 的检测算法深度优化

3.1.1 MobileViT 的架构

高速运动场景下的目标检测需要综合考虑时空两方面的复杂性。空间层面上，目标往往伴随大尺度变化和运动模糊，且背景干扰与相似物体并存，这使得检测网络必须同时具备强大的局部细节捕捉与多尺度感知能力；时间层面上，高速运动意味着前后帧之间存在显著差异，目标很可能在短时间内完成大范围位移或被遮挡并再现。若检测模型缺少对时间信息的利用，只依赖单帧提取的静态特征，可能难以及时更新对目标的定位，或在类似背景中出现漏检。

在资源受限的移动平台上，高效性是不可或缺的要求。高帧率检测能帮助系统在极短时间内多次刷新目标位置，但与此同时，也对网络计算量提出了严格限制。轻量化网络在此环境中更具适配度；然而，过度压缩模型又易导致对特征表征力的削弱，造成在复杂背景或极度模糊场景下的精度损失。如何在时空特征建模和模型轻量化之间寻找最优平衡，成为面向高速检测的核心问题。

MobileViT 作为融合卷积神经网络与视觉 Transformer 的代表性轻量化架构，初步展示了其在多尺度感知与全局关联建模方面的潜力。通过在前几层保留卷积模块来捕捉局部纹理，并在后续引入自注意力机制用于全局依赖关系的学习，MobileViT 能够兼顾检测速度与特征表达。然而，在特征分辨率、注意力分配策略以及小目标检测等方面，若要应对高速场景中更极端的输入扰动，仍需进一步改进与加强。

MobileViT 的核心思想在于将图像划分为块后先进行卷积处理，再借助精简的 Transformer 模块建模全局上下文信息。轻量化卷积保证了网络在前端能够快速提取特征并适配嵌入式硬件，而后续的 Transformer 部分提供了跨区域、跨通道的长程依赖建模能力，这对于应对快速背景变化、形变剧烈的目标非常重要。具体结构如图 3-1 所示：

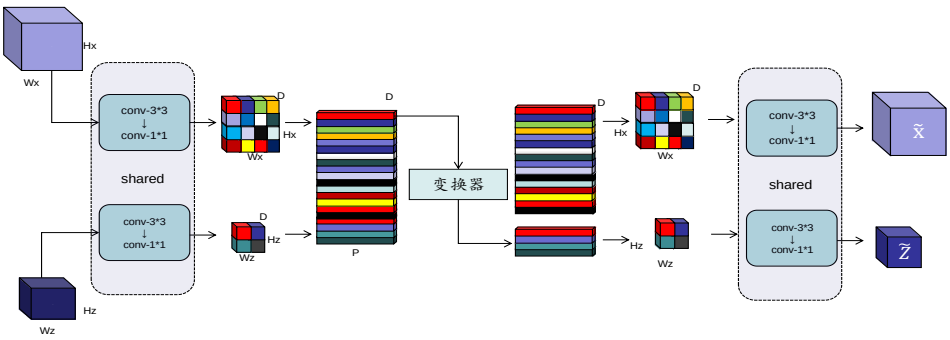


图 3-1 MobileViT 的结构

然而，应用于高速目标检测时，MobileViT 仍存在一些局限。第一，在运动模糊和遮挡较严重的情况下，如何充分保留细节成为难点。若网络的下采样次数较多，小目标或精细边缘信息有可能在浅层阶段过早地被丢弃。第二，Transformer 的多头自注意力计算量随输入分辨率的提升而明显增加，当系统需要保证 30 帧甚至更高帧率进行检测时，模型的推理速度会受到限制。第三，对于大范围尺度变化和类似目标共存的复杂场景，MobileViT 的默认注意力机制虽可一定程度上捕捉全局相关性，但可能缺少对关键区域或关键通道的精细加权，容易产生误检或漏检。第四，在缺乏显式时序融合的情况下，MobileViT 只能处理单帧图像，对连续帧之间的关联利用不足。

3.1.2 mobileViT 的适应性改进

针对这些不足，在保留 MobileViT 轻量化与全局上下文建模优势的基础上，引入多尺度特征融合模块与基于有效通道注意力（Efficient Channel Attention，ECA）[46]和卷积块注意力模块（Convolutional Block Attention Module，CBAM）[47]的通道及空间注意力，提高对高速目标的感知精度。此外，通过量化和剪枝在 Transformer 部分减少计算耗时，并在网络顶层增设时序先验融合模块，弥补其在时间维度上的不足，为后续跟踪环节提供更加稳定的检测候选框输入。

在高速检测场景下，需要兼顾细节纹理与高级语义，因此可以在 MobileViT 的中浅层阶段进行多尺度特征融合：先保留高分辨率的浅层特征以捕捉小目标和局部纹理，再将其与更深层语义特征逐级整合，从而在运动模糊和复杂背景下保留更多有效信息。为了进一步提升模型的辨别能力，可引入多种注意力机制以突出关键区域并抑制冗余干扰。

如图 3-2 所示，ECA 侧重于通道注意力，通过一维卷积近似全连接操作，实现对每个通道重要度的轻量化计算。由于卷积核尺寸通常远小于通道数目，ECA 能以极低的参数开销为所有通道分配自适应权重，特别适合在高速检测中突出那些对目标定位至关重要的特征通道。

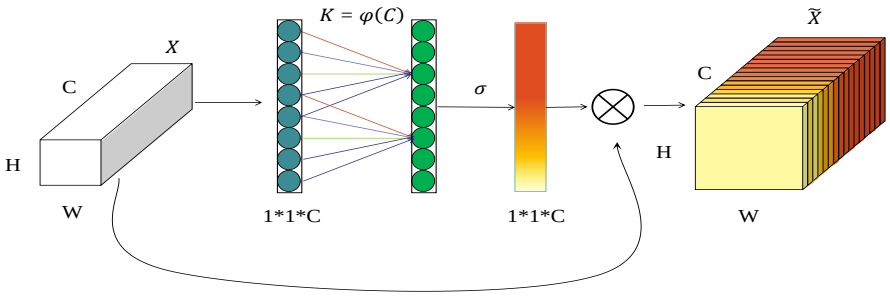


图 3-2 ECA 模块的结构

在需要强调空间维度特征时，可引入 CBAM，如图 3-3 所示。CBAM 既包含通道注意力，也通过卷积获取特征图的空间注意力分布，在原特征图上生成一张空间注意力图，从而在局部范围内突出显著区域并抑制背景干扰。对于运动模糊和复杂背景场景，空间注意力有助于模型精准聚焦于目标所在位置并忽略无关区域。

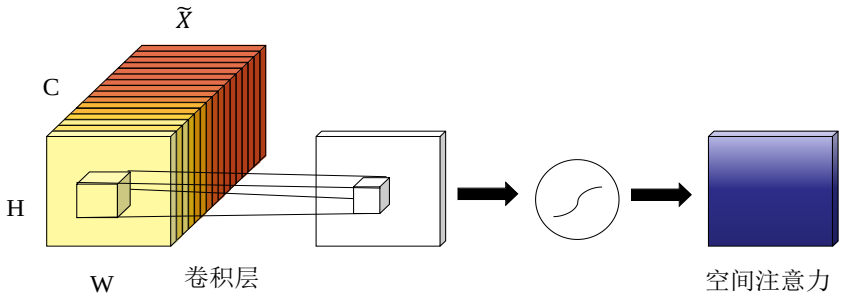


图 3-3 CBAM 模块结构图

在此基础上，将 ECA 与 CBAM 有效结合即可形成 ECSA（Efficient Channel and Spatial Attention），如图 3-4 所示。ECA 以最小的计算代价获得通道级重要度，而 CBAM 的空间注意力机制能进一步突出关键位置，两者协同后使网络同时具备高效的通道筛选与精细的空间感知能力。对于高速检测任务中的快速位移、相似背景和运动模糊等难点，ECAM 可以在轻量化的前提下显著强化模型的判别能力。

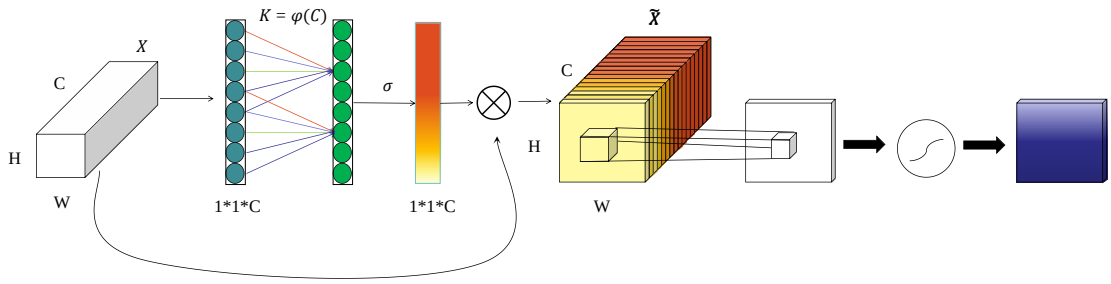


图 3-4 ECSA 模块结构图

在完成多尺度融合与 ECSA 注意力强化的同时，在检测头部或网络输出层融入轻量级时序卷积或小型 Transformer，以利用前后帧的特征差异来预测目标位移与遮挡情况。

3.2 结合 Nanotrack 的跟踪算法改进

3.2.1 Nanotrack 跟踪算法简介

Nanotrack 是一种轻量化单目标跟踪算法，旨在在有限算力和高帧率需求下兼顾速度与精度。Nanotrack 选用 MobileNet 作为骨干，用来提取目标的表观特征，并通过高效的相关操作或卷积方法衡量模板与搜索区域之间的相似度。它会在特征对比之后生成一个响应图，或者借助微型的边界框预测头来回归目标的位置与尺度。与传统的 Siamese 跟踪方法相比，Nanotrack 在网络结构上进行了大量精简，使得参数量更少，推理速度更快。在高速运动或复杂环境中，Nanotrack 仍需面对目标外观发生剧烈变化、被相似物体干扰或在背景杂波中出现遮挡的问题。具体结构如图 3-5 所示。

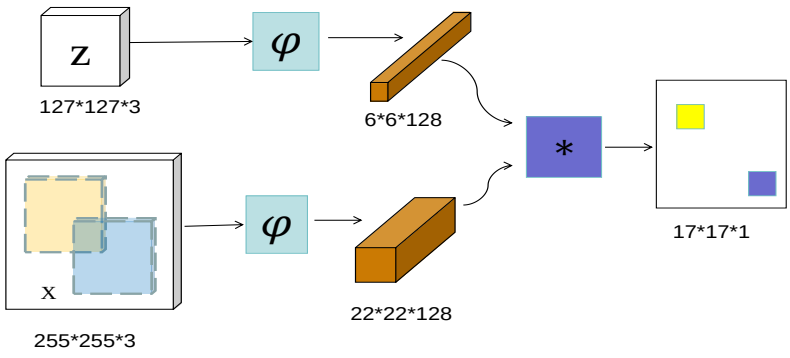


图 3-5 Nanotrack 结构图

与传统的 Siamese 跟踪方法相比，Nanotrack 在网络结构上进行了大量精简，使得参数量更少，推理速度更快。

在高速运动或复杂环境中，Nanotrack 仍需面对目标外观发生剧烈变化、被相似物体干扰或在背景杂波中出现遮挡的问题。

3.2.2 NanoMViT 融合原理与创新架构

为兼顾本地感知与全局建模的需求，提出了 NanoMViT（Nano Mobile Vision Transformer for Object Tracking）来替换原有的 MobileNet 骨干网络，如图 3-6 所示。

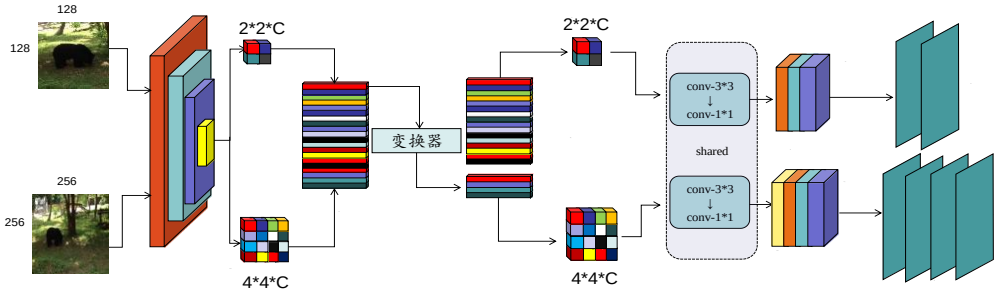


图 3-6 NanoMViT 的结构图

相比传统的轻量级 CNN 模型，MobileViT 能够在保持较低计算量的同时结合 Transformer 的全局特征提取能力，针对目标的外观变化、复杂背景和相似干扰等问题具有更强的适应性。

NanoMViT 通过融合 MobileViT 的骨干结构与 ECSA（Efficient Channel and Spatial Attention）多重注意力模块来提升模型的整体判别能力。MobileViT 背后的原理是将局部卷积特征与 Transformer 的自注意力机制相结合：前者在浅层阶段有效提取目标的局部纹理和边缘信息，后者在深层阶段更好地捕捉全局依赖。这种多层次融合使得网络不仅能在高速运动或复杂环境下有效检测目标，还能在遮挡、快速形变等极端情形下保持较高的鲁棒性。

为了进一步挖掘模型的潜力并减少在移动设备上的计算开销，NanoMViT 还引入了 ECSA 模块。在整体结构中，为了进一步压缩模型规模并提升实时性能，本文在 NanoMViT 的训练阶段结合了量化与剪枝等轻量化技术，使其在高帧率需求下依旧能够快速输出目标位置。

3.3 检测与跟踪一体化融合策略设计

3.3.1 融合策略的信息交互逻辑与同步机制

在实际场景中，检测与跟踪往往并非割裂。单帧级的检测能在较大搜索范围内精准定位目标，而跟踪则以帧间连续的相似度信息进行快速预测，显著减少对每帧进行完整检测的需求。二者结合的一体化融合策略对于高速运动场景至关重要，可以在保证实时性的同时减小漂移风险。

信息交互逻辑通常包含两个方向：当检测分支提供了新的高置信度位置，若与当前跟踪位置相符，则仅在局部做修正；若差距较大，说明跟踪失效或目标严重偏移，此时根据检测结果重新初始化跟踪器。反向地，当跟踪器连续若干帧出现置信度急剧下降，也可触发检测分支在当前帧进行全局搜索，以纠正潜在错误或查找可能的新出现目标。

同步机制则可根据应用实时性与算力状况灵活选择。若系统算力相对充裕，可在每帧同时运行检测与跟踪，通过简单融合策略得到更加稳健的目标位置。若算力有限，则可采用间隔检测模式或置信度自适应模式：仅在若干帧中执行一次检测或在跟踪置信度不足时调用检测，其他帧只使用孪生网络进行跟踪推断。这样能有效降低检测算法的大规模卷积或 Transformer 计算带来的开销，并在目标平滑运动时保持较高的帧率。

3.3.2 联合优化的目标函数实现

为了让检测与跟踪在训练阶段就能够协同提升，可在同一网络或同一训练框架下建立联合目标函数，使检测模块和跟踪模块在共享的特征空间中相互促进。

检测模块一般包含边界框回归损失 L_{reg} 与分类损失 L_{cls}，用于监督网络准确定位目标位置并区分目标/背景，如公式（3-1）所示：

$$L_{det} = \frac{1}{N}\sum_{i=1}^{N}\left(L_{cls}\left(\widehat{P_i}, P_i^*\right) + \lambda L_{reg}(\widehat{t_i}, t_i^*)\right), \tag{3-1}$$

其中，p_i 表示第 i 个候选框的预测类别分别，p_i^*为其真实类别，t_i 为预测的边界框参数，t_i^*为对应的真值，λ 为平衡系数，N 为样本总数。

跟踪模块则有相似度或相关滤波损失，用于衡量搜索帧与模板帧的特征一致性程度，其相似度可定义如公式（3-2）所示：

$$S(z,x) = z * x, \tag{3-2}$$

其中*表示卷积操作，若有标注目标中心位置 u*，则可将偏差最小化作为跟踪损失，如公式（3-3）所示：

$$L_{track} = \sum_j \left\| u_j - u_j^* \right\|^2, \tag{3-3}$$

其中 j 遍历搜索区域上的位置，u_j 为在位置 j 为处所预测的目标响应。

实现二者的端到端学习，可以在同一批训练数据中，对检测分支与跟踪分支的输出分别计算损失，再通过加权求和得到联合损失。若需要保证检测和跟踪各自的性能侧重点，可在损失函数中添加一个一致性正则项，约束检测输出与跟踪预测在目标位置上基本重合，从而减少训练冲突。

在一体化训练中，检测与跟踪同时产生梯度更新，可用加权求和的形式：

$$L_{total} = \alpha L_{det} + \beta L_{track} + \gamma L_{con}, \tag{3-4}$$

其中 α 和 β 是分别控制检测损失与跟踪损失权重的超参数，L_{con} 是可选的一致性正则项，用于约束检测输出与跟踪预测的目标位置在像素坐标或特征空间保持一致，γ 为相应的权重。

3.3.3 检测与跟踪一体化模型的实现与验证

在工程实现层面，可将 MobileViT 与 Nanotrack 统合为一个共享主干网络与分支模块的体系结构：前端的浅层卷积用于提取局部特征并适度下采样，Transformer 与多注意力机制用于进一步全局上下文建模并筛选显著区域。检测分支输出目标边界框候选，跟踪分支通过孪生网络进行相似度匹配，二者在后端进行融合，决定最终目标位置。

为了最大化利用多注意力机制，可以将 ECA 和 CBAM 均嵌入到检测与跟踪共享的 backbone 中，让通道及空间注意力在特征生成阶段就发挥作用，减少后续检测与跟踪分支对噪声的误判。同时，若系统有时序建模需求，也可在检测头或者跟踪器尾部加上轻量级时序注意力模块，把当前帧与前几帧特征进行关联。这种做法可以帮助在目标快速遮挡后出现的场景下更好地恢复跟踪。

验证该一体化模型效果时，需要在检测和跟踪各自的典型数据集上进行测试。对于检测，选用 COCO 数据集，主要观察平均精度、漏检率等指标。对于跟踪，则可在 VOT2018、DTB70、LaSOT 等多个基准数据集上测量准确率、鲁棒值、平均重叠（EAO）和成功率等。同时，还需记录模型在移动端或 GPU 端的实际推理速度，对比纯检测或纯跟踪的性能差距。若整体融合后在高速运动场景下依然能保持较高帧率，同时检测与跟踪精度也显著提升，则说明本章提出的融合策略在工程应用中具有价值。

通过以上几个环节的融合设计与实验分析，可以看出在 MobileViT 与 Nanotrack 等轻量化 Transformer 基础上实现检测-跟踪一体化，不仅可以在高速场景中帮助系统更好地克服大范围位移、遮挡、相似干扰等因素带来的挑战，也能充分发挥检测与跟踪的互补性，既减少了对每帧全局检测的依赖，又避免了仅依赖局部匹配时可能出现的跟踪失效。尤其是在移动设备或无人机平台上，这种小型且高效的检测跟踪一体化模型，对于资源受限环境下的实时目标跟踪应用有着明显优势，也为后续在工业检测、自动驾驶等领域的部署提供了可行思路。后续章节将基于实验结果对上述改进进行验证，并进一步分析不同注意力配置、时序融合策略与结构化剪枝方案所带来的性能影响。

3.4 NanoMViT 的对比实验

基于前文所提出的高速目标检测与跟踪模型展开系统实验，对比分析不同方法在公开数据集与真实场景下的综合性能。主要关注在检测与跟踪的准确率、实时性、轻量化程度以及在高速运动场景中的鲁棒性。为了验证各模块的有效性，还通过消融实验考察多尺度融合策略的性能增益。

3.4.1 设备环境与依赖配置描述

实验主要在配备 Intel Core i9-10900K 处理器和 NVIDIA RTX A6000 显卡。该系统具有有 10752 个 CUDA 核心，并搭配 48GB 内存和 2TB 固态硬盘。在大规模数据集上进行训练与评估时，RTX A6000 的高带宽显存可提供充足算力，保证多线程并行与深度学习加速的顺利进行。

深度学习框架采用 Python 3.9 与 PyTorch 1.10.1，并安装 CUDA 11.3 及 cuDNN 8.2 进行 GPU 加速。模型训练与推理需要依赖 NumPy、OpenCV、TorchVision 和 Matplotlib 等库，用于数据预处理、可视化和结果统计。

3.4.2 跟踪算法性能对比实验

为了评估 NanoMViT 在目标跟踪任务中的优势，选取了 VOT2018、LaSOT 和 DTB70 等主流数据集。比较对象为 SiamVGG[57]，SA_Siam_P，C-COT[58]，SiamFC[56]，SiamTPN[59]，SiamAPN[60]，SiamAPN++[61]，Nanotrack 和 ECO[62]方法。主要指标包括跟踪精度（Precision）、成功率（Success Rate）和平均重叠率（Expected Average Overlap，EAO）。VOT2018 如图 3-7 所示。

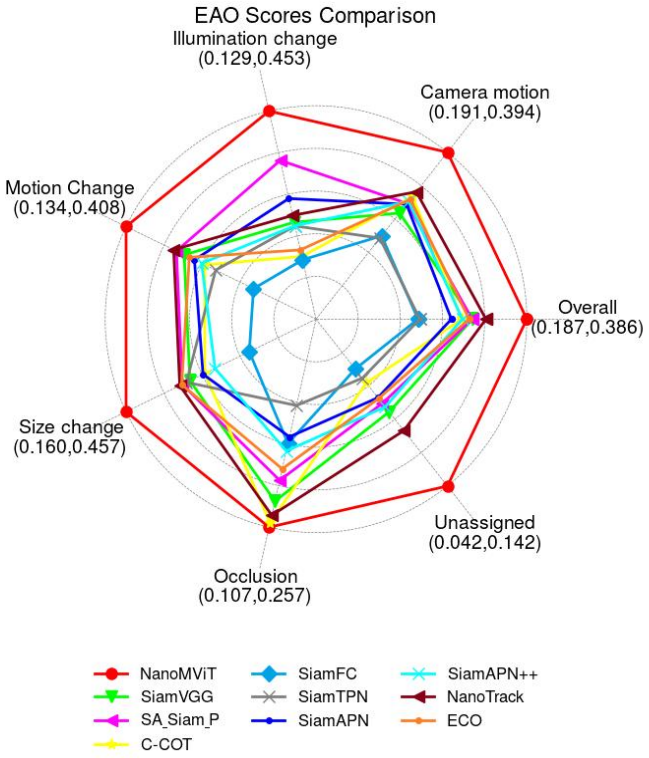


图 3-7 VOT2018 的 ECO 雷达图

从图 3-7 的实验结果可以看出，本文提出的算法在准确率上比基线算法 Nanotrack 提高了 12.9%，在预期平均重叠指标上也实现了 16.3%的提升。雷达图进一步展示了该算法在运动变化、尺寸变化、相机运动和光照变化等多个维度上均取得了显著改善，充分证明了其在应对复杂场景变化时的高适应性和鲁

棒性。然而，在处理目标遮挡问题上，算法的提升相对有限。这主要归因于本文方法中未集成目标检测网络，导致在目标被部分或完全遮挡时，重跟踪的能力受到一定制约，从而影响了整体跟踪性能。未来的研究工作可考虑引入目标检测模块，以进一步增强在遮挡场景下的目标恢复能力，提升算法的整体稳定性和鲁棒性。

在 DTB70 数据集中与 SiamFC，LightTrack[63]，SiamAPN，Ocean[64]，SESiam_FC[65]，SiamDW_FC_Incep22[66]，SiamDW_RPN_Res22，SiamGAT[67]，SiamMask[68]，SiamRPN_alex[69]和 SiamRPN_mobilev2[69]等算法对比，精度准确率图如图 3-8 所示。

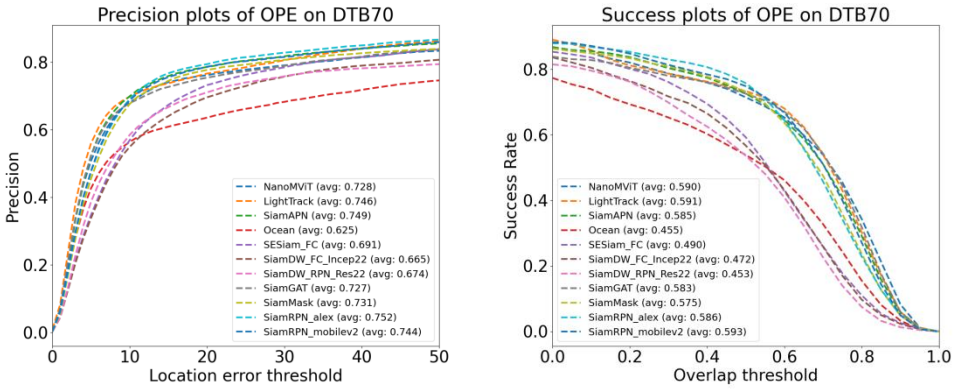


图 3-8 DTB70 对比图

实验结果表明，本文算法在 DTB70 上表现优异，与基线算法 LightTrack 相比，其精度提高了 2.2 个百分点。这一提升充分验证了本文方法在处理动态场景和复杂背景干扰下的高适应性和稳定性，同时也体现了在精度与鲁棒性之间取得的良好平衡。

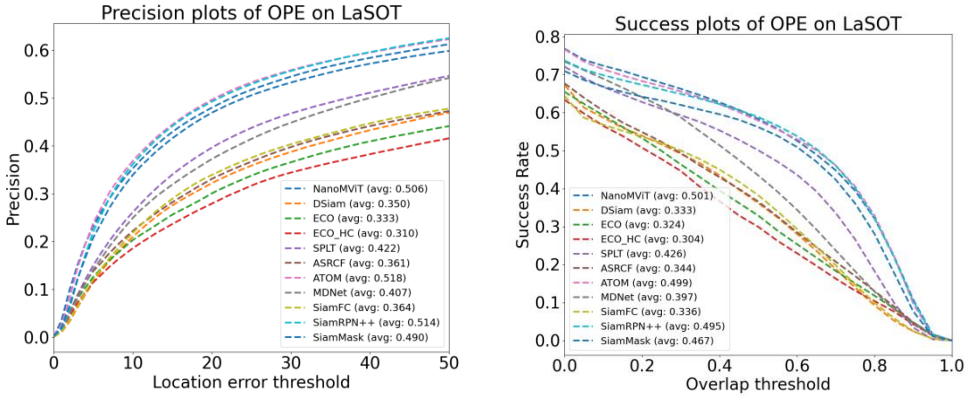


图 3-9 LaSOT 对比图

在 LaSOT 数据集上与多个现有的跟踪器进行比较，包括 DSiam，ECO，SPLT，ASRCF，ATOM，MDNet，SiamFC，SiamRPN++和 SiamMask。如图 3-9 所示。实验结果表明，本文提出的算法在 LaSOT 数据集上也同时具有比较高的性能，与算法 SiamRPN++相比其成功率提高了 0.6 个百分点。这一提升充分验证了 NanoMViT 在复杂场景下也能有较高的成功率和精度。

3.5 消融实验

3.5.1 ECA 消融实验

为了评估 ECA 的有效性，在 VOT2018 数据集上进行了效果消融实验，以 Nanotrack 作为基础算法。实验结果如表 3-2 所示,下划线表示最高分。

表 3-2 ECA 模块在 VOT2018 上的效果消融实验

算法	准确率	鲁棒值	EAO
基准	0.562	0.276	0.311
实验 1	0.563	0.262	0.323
实验 2	0.538	0.220	0.225
实验 3	0.588	0.211	0.342

在第一项实验中，将原有 NanoTrack 模型中的注意力机制模块（Squeeze-and-Excitation Module，SE）[70]模块替换为 ECA 模块。对比实验结果显示，ECA 模块在注意力建模和结果的鲁棒性方面均优于 SE 模块。进一步分析发现，SE 通过全局池化和全连接层进行通道注意力的计算，能够捕捉到全局通道间的依赖关系，但在实际推理中容易忽视部分局部特征的细粒度信息；而 ECA 模块则借助轻量级卷积来在通道间传递信息，更有利于在保留全局信息的同时挖掘局部特征，从而在目标跟踪场景中展现更好的稳定性和适应性。

在第二项实验中，在 NanoMViT 的基础结构上直接添加 SE 模块，对比基准算法后得知，新加入的 SE 并未带来预期的精度收益，反而在一定程度上降低了模型的性能。通过对网络结构的进一步剖析，可以推测原因在于：NanoMViT 本身包含了多层轻量化的卷积和 Transformer 结构，已经在局部与全局特征提取之间达成了较好的平衡；当再叠加 SE 模块后，虽然理论上可强化全局通道注意

力，但全连接操作对局部特征的剥离过于粗糙，不仅无法与 NanoMViT 的轻量化思路很好配合，还可能在反向传播中导致部分关键细节被忽略。

在第三项实验中，将第二项实验中的 SE 模块继续替换为 ECA 模块，结果表明在 NanoMViT 上加入 ECA 后的网络表现优于前述所有对照组。具体而言，ECA 模块采用一维卷积实现通道间注意力的交互，无需全连接操作，不仅保留了全局特征的表征能力，也进一步强化了对局部细节的捕捉。对于 NanoMViT 这样注重多尺度与局部依赖关系的模型而言，ECA 模块的引入恰好能够与主干网络的设计理念相契合：既不大幅增加模型参数量，又能提升在复杂场景下的目标辨识与定位能力。

3.5.2 CBAM 的消融实验

为了评估 ECA 和 CBAM 注意力融合的有效性，在 VOT2018 数据集上进行了效果消融实验，以 Nanotrack 作为基础算法。实验结果如表 3-3 所示,下划线表示最高分。

表 3-3 CBAM 模块参数消融实验。

算法	准确率	鲁棒值	EAO
基准	0.562	<u>0.276</u>	0.311
实验 1	0.588	0.211	0.342
实验 2	0.572	0.232	0.331
实验 3	<u>0.612</u>	0.252	<u>0.362</u>

在将 ECA 模块引入 NanoMViT 的初步实验中，模型在通道维度的注意力分配方面得到了有效增强。ECA 通过轻量级卷积操作来捕捉通道间的相关性，在保持全局信息表征能力的同时，更充分地利用了局部特征，从而对精细纹理和外观变化更为敏感。实验结果表明，与原始基准相比，添加 ECA 后的 NanoMViT 在跟踪精度和稳定性方面均有一定提升，说明在轻量化网络中注重通道间信息交互能切实增强检测与跟踪的效果。

随后又在 NanoMViT 中添加了 CBAM 进行对比，发现仅使用 CBAM 的效果并不如预期。与 SE 模块类似，CBAM 通过通道注意力和空间注意力的联合来突出显著区域，虽然在全局信息提炼上具有一定优势，但在局部细节的保留和对相似目标的精细区分方面并未展现出足够的竞争力。由实验结果可知，单

独使用 CBAM 的效果甚至比在同等条件下添加 SE 模块时更不理想，这可能与 CBAM 对空间注意力的计算方式及其对局部通道特征的处理策略相关。

在对相似物体的跟踪任务中，若两者在外观、颜色或轮廓上极为相近，仅依靠单一注意力机制往往难以实现精准区分。因此，在第三组实验中，针对上述缺陷同时集成了 ECA 与 CBAM 两个模块，试图将二者的优势有机结合：ECA 的轻量级通道卷积机制先对特征进行初步筛选，将无关或冗余的通道响应抑制在较低水平；之后借助 CBAM 的通道与空间关注策略在特征图中进一步突出目标的关键部分，从而形成“先粗略过一遍，再细化强调”的多重注意力调整流程。该方法在相似物体场景下表现出显著的性能增益，特别是在精度要求较高的测评指标上，超过了仅使用 ECA 的基线模型。

结合对消融实验的分析可见，这种分阶段、从粗到细的注意力调控方式在相似物体跟踪情境中不仅有效，而且也是必要的：若仅有 ECA，会在通道选择层面起到积极作用，但对空间维度上的微小差异尚显不足；若单靠 CBAM，则由于其对局部通道特征的处理并不如预期，难以应对高相似度目标在外观或纹理上的微妙差别。唯有将二者配合起来，一方面通过 ECA 提供稳健、高效的通道注意力筛选，另一方面利用 CBAM 在空间注意力中的精细化聚焦，才能在实际跟踪过程中同时兼顾到目标全局轮廓与关键细节，从而有效地减轻目标混淆、遮挡和干扰带来的不利影响。综观全部实验结果，此策略对于需要在不同尺度或外观下保持精确跟踪的应用场景具有明确的实用价值。

3.6 NanoMViT 可视化实验

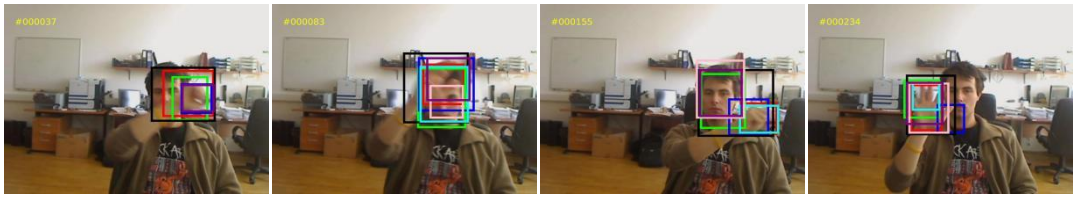
为了更有效地展示 NanoMViT 在复杂环境下的表现，在 VOT2018 基准上进行了定性分析。对 NanoMViT、SiamVGG、SiamFC、SiamTPN、SiamAPN、SiamAPN++和 Nanotrack 的结果进行了直观对比。具体如 3-10 所示。

Girl

Singer1

Singer2

Hand



Leaves

图 3-10 NanoMViT 的可视化

在“Girl”视频序列中，目标人物在运动过程中遭遇了较为严重的遮挡与模糊干扰，尤其在第 112 帧处，由于目标面部被部分阻挡，跟踪器接收到的外观信息不足，导致大多数对照算法出现了不同程度的漂移。更为关键的是，到了第 1250 帧时，场景中的光照、背景和目标外观都已经发生了较大变化，传统方法往往无法成功重定位目标，从而导致跟踪失败。相比之下，NanoMViT 的注意力机制能够在特征提取阶段充分关注到目标的局部区域，并通过 MobileViT 主干网络在模糊和遮挡情况下依然提炼出相对稳定的关键特征，使得在后续帧中能够持续锁定目标位置，避免了跟踪窗口的持续偏移。

“Singer1”序列则主要考验跟踪器应对目标大幅度形变的能力。该序列中，表演者的姿态和服装轮廓在短时间内发生明显变化。当序列运行到第 53 帧时，可以观察到 SiamFC 和 SiamAPN 并未及时更新目标边界框，错失了捕捉目标最

新外观的时机，最终导致了目标跟踪的丢失。而 NanoMViT 借助其对局部细节和全局布局的多重刻画，当目标发生弯腰、转身等快速形变时，依旧能够通过多尺度特征融合实时调整模型对目标的定义范围，因而保证了跟踪过程的连贯性。

“Singer2”序列中，大量背景干扰物和复杂的舞台布景使跟踪器容易把背景区域误判为目标。NanoMViT 架构中的 Transformer 全局注意力机制能够自适应地学习到目标与背景的差异特征，并通过跨通道、跨位置的信息交互，有效削弱复杂背景对模型的干扰，实现更精准的目标定位。

对于“Hand”序列，目标手部在移动过程中常常出现模糊，且场景中还存在与目标外观相近的干扰物体，导致目标与背景或其他手部对象的区分度不足。在第 155 帧时，大多数竞争算法由于无法在短时间内分辨手部与周围相似区域，被模糊误导而跟踪失败。NanoMViT 依靠在轻量化卷积和自注意力层中强化的通道与空间表征能力，即便目标形状模糊、对比度较低，也能依据局部关键纹理与全局上下文维持跟踪窗口的正确位置，实现较为稳定的跟踪序列输出。

“Leaves”序列则更强调对高速运动与相似目标的识别能力：一方面，目标在帧与帧之间的位移大，需要跟踪器具有快速更新能力；另一方面，叶片等相似背景元素的快速抖动也会与目标产生干扰，使得跟踪器较易发生混淆。在这种场景中，若算法缺乏对局部细节及运动趋势的自适应捕捉能力，即便能够短暂锁定目标，也往往会在后续帧中因目标与背景外观相近而出现误判。NanoMViT 在该序列中展现出的优异性能主要得益于其高效的多尺度特征融合与注意力机制。

3.7 本章小结

本章围绕高速目标检测与跟踪任务中实时性与精度之间的矛盾挑战，提出了基于轻量化 Transformer 框架的检测跟踪算法优化方案，并以 NanoMViT 为核心构建模块进行了详细论述。首先，对 MobileViT 在高速检测场景下的时空特征建模进行了深入分析，指出其在捕捉局部细节与全局关联方面的潜力及存在的局限，进而引入多尺度特征融合与多重注意力机制，以强化对小目标、运动模糊和复杂背景的感知能力。其次，在跟踪模块上，通过 Nanotrack 算法的结构

改进，实现了轻量化特征提取与模板匹配，并利用时序融合模块补充单帧检测的不足，从而提高了对目标位移与遮挡变化的适应性。第三，检测与跟踪的一体化融合策略通过信息交互与同步机制实现了两者之间的优势互补，联合优化的目标函数确保了端到端训练中检测定位与跟踪预测的协同提升。最后，通过对比实验、消融分析和可视化定性展示，验证了所提出方法在各项指标上的优势，并证明了其在复杂、高速运动场景中的实际应用价值。

第 4 章 基于 Transformer 模型自适应轻量化模型设计

4.1 基于图像复杂度估计的模型轻量化

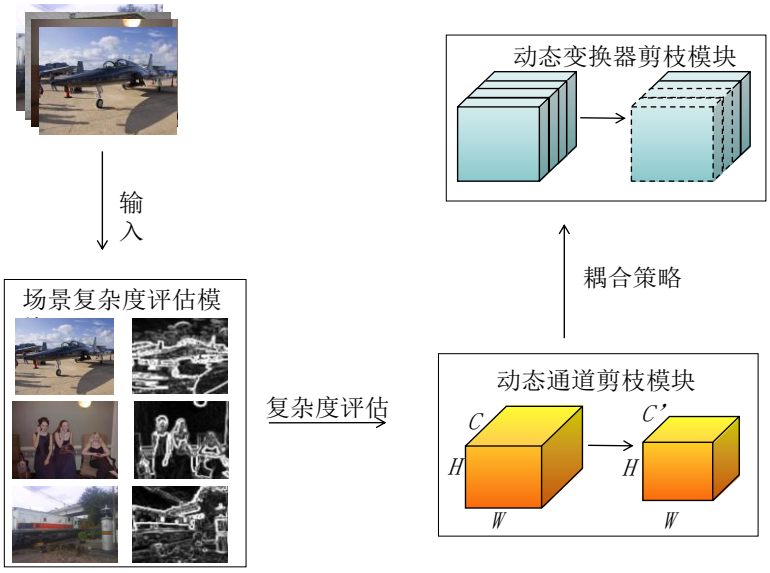


图 4-1 图像复杂度估计的模型流程

基于图像复杂度估计的模型流程如图 4-1 所示。图像复杂度评估是 AC-MViT 剪枝模型得以实现“按需分配”资源的前提。其核心思想是：当输入场景较简单、目标外观变化不大时，无需全面启用所有卷积通道与所有 Transformer 层，即可维持稳定的检测精度；而当场景极其复杂或目标形变显著时，需要保留较完整的网络结构以获取足够的细节表达能力。要想将这种“按需分配”从概念变为现实，需要建立对场景复杂度的可量化评估，并与剪枝、量化等模型压缩技术深度结合，形成动态调度策略，以最小化计算代价为目标，最大化保留关键特征。基于此思路，将从图像复杂度估计、自适应阈值、动态剪枝和 Transformer 剪枝这四个方面讲解。

4.1.1 复杂度评估实现

在对输入图像进行处理时，需要以多种方式量化其视觉与语义复杂度，以便在后续模型优化中依据该复杂度做出自适应调度。首先计算信息熵$H(x_i)$来衡量图像整体的不确定性，如式(4-1)所示：

$$H(x_i) = -\sum_i p_i log_2(p_i), \tag{4-1}$$

其中p_i是像素值为 i 的概率，以此量化图像的信息量和不确定性程度。随后使用 Sobel 算子等边缘检测算法计算图片的边缘密度，通过统计像素值出现的概率分布，可以初步判断画面中是否存在多目标或显著的亮度变化。若信息熵数值偏高，则图像内部的像素分布更为多样，对网络的识别与定位能力提出更高要求。紧接着可以利用 Sobel 等边缘检测算子来获取图像边缘强度分布，如式(4-2)所示:

$$E(x_i) = \frac{\sum_{x,y} \mathbb{I}(A(x,y))}{M \times N}, \tag{4-2}$$

其中$A(x,y)$是边缘检测得到的像素处(x,y)的边缘强度，$\mathbb{I}\big(A(x,y)\big)$是指示函数（当A(x,y)大于阈值时为1，否则为0），$M \times N$是图像的总像素数，边缘密度反映了图像的结构复杂度。若边缘密度较大，往往意味着图像存在大量显著的结构信息，需要模型在卷积层或 Transformer 模块中投入更多计算资源来充分提取目标轮廓或纹理特征。随后基于灰度共生矩阵（GLCM）计算纹理复杂度$T(x_i)$，通过提取对比度、熵、相关性等纹理特征，并进行加权组合，如式(4-3)所示：

$$T(x_i) = w_1 C(x_i) + w_2 E_G(x_i) + w_3 R(x_i), \tag{4-3}$$

其中、$C(x_i)$,$E_G(x_i)$和$R(x_i)$和分别是对比度、GLCM 熵和相关性，w_1、w_2和w_3是相应的权重系数，纹理复杂度能够描述图像的细节丰富程度。通过提取对比度、熵、相关性等特征并进行加权，可以更加全面地反映图像在空间和频域层面上的变化。若这些纹理特征呈现较大数值，通常说明画面蕴含高频细节或重复结构，对检测与分类而言具有更高难度。

同时考虑使用预训练的神经网络对图片进行处理，得到其在特定任务（如分类）下的交叉熵损失$L_{ce}(x_i;W)$，该损失反映了网络对图片的拟合难度，与图像自身的特征相结合，能够更全面地衡量图片的复杂度。

最后通过线性加权的方式将上述各项指标融合为一个综合复杂度指标$C_t(x_i)$，如式(4-4)所示：

$$C_t(x_i) = \alpha_1 L_{ce}(x_i;W) + \alpha_2 H(x_i) + \alpha_3 E(x_i) + \alpha_4 T(x_i), \tag{4-4}$$

其中α_1、α_2、α_3和α_4是根据数据集和任务特点确定的权重参数，通过调整这些参数，可以使综合指标更侧重于不同的复杂度方面。

4.1.2 自适应阈值确定

在得到综合复杂度指标后，需要在训练时对划分复杂与简单图像的阈值进行自适应调节，根据网络性能的反馈信息动态调整阈值C。定义一个关于阈值的损失函数,如式(4-5)所示:

$$L_{th}(C) = \sum_{i=1}^{N} l(x_i, C), \tag{4-5}$$

其中$l(x_i, C)$是根据实例x_i与阈值C的关系以及网络对x_i的预测结果定义的损失项。通过对$L_{th}(C)$关于C求导，并根据导数的正负和大小来调整C的值。通过定义一个关于阈值的损失函数，对阈值的大小进行迭代更新，从而使被判定为复杂或简单的图像各自对应的预测精度和效率都能保持在较佳水平。若网络发现多数样本都被划分为复杂但模型却未能显著提升这部分样本的准确率，则说明阈值设置过高，导致系统在本可轻量化处理的图像上浪费了算力，应适度降低阈值；若在复杂场景中的预测准确率明显提升却造成简单场景精度或速度下降，则需要上调阈值来避免对大批简单样本投入过多资源。通过不断跟踪阈值损失关于阈值的梯度变化，系统能够从训练和验证数据中学到最优的阈值调整规律，使复杂与简单图像的划分更加合理。

一旦阈值确定，网络在推理时就可根据每张输入图片的综合复杂度指标来即时判断其属于复杂还是简单范畴。若判定为复杂，会在后续模型计算中保留更多关键通道或 Transformer 层数；若判定为简单，则可跳过若干冗余模块或采用更低比特量化的方式来降低计算和存储开销。通过这种自适应划分，系统在处理不同图像时可以实现“多用多算，少用少算”，兼顾整体精度与实时性。

4.1.3 动态通道剪枝

对于 mobilevit 中的标准卷积层，可以直接采用动态通道剪枝的卷积层定义。将原始的卷积计算替换为式(4-6)：

$$f_l(x_{l-1}) = \left(\lambda_l(x_{l-1})\left(norm\left(conv_l(x_{l-1}, w_l)\right) + \gamma_l\right)\right)^{+}, \tag{4-6}$$

其中x_{l-1}是第l层卷积层的输入特征图,$\lambda_l(x_{l-1})$是一个基于上一层输出的低开销策略函数，w_l是卷积层的权重张量，$conv_l(x_{l-1}, w_l)$表示使用权重张量对输入特征进行卷积运算，得到的结果是经过卷积操作后的特征图，$norm(Z)$是对特征Z进行归一化操作,γ_l是可训练的参数，$(Z)^{+}$表示激活函数。

在计算时$\lambda_l(x_{l-1})$，利用通道显著性评估策略，如式(4-7)所示：

$$\lambda_l(x_{l-1}) = wta_{dC_l}\big(g_l(x_{l-1})\big), \tag{4-7}$$

其中wta是一个选择函数，$wta_{dC_1}(Z)$返回一个与Z形状相同的张量，除了将Z中绝对值小于dC_l（d是密度参数）个最大元素的项置零。

$g_l(x_{l-1})$是一个用于预测通道显著性的函数，其输出决定了wta函数的输入，进而影响通道的剪枝决策。

之后通过特征下采样公式和基于全连接层的通道显著性预测公式如式(4-8)，式(4-9)所示：

$$ss(x_{l-1}) = \frac{1}{HW}\sum_{x[1]l-1}\sum_{x[2]l-1}\cdots\sum_{x[c]l-1} x_{l-1}\ , \tag{4-8}$$

$$g_l(x_{l-1}) = (ss(x_{l-1})\theta_l + \beta_l)^+, \tag{4-9}$$

其中$ss(x_{l-1})$用于将输入特征（维度）每个通道的空间维度降为标量，θ_l是全连接层的权重张量，β_l是全连接层的偏置向量。如图 4-2 所示。

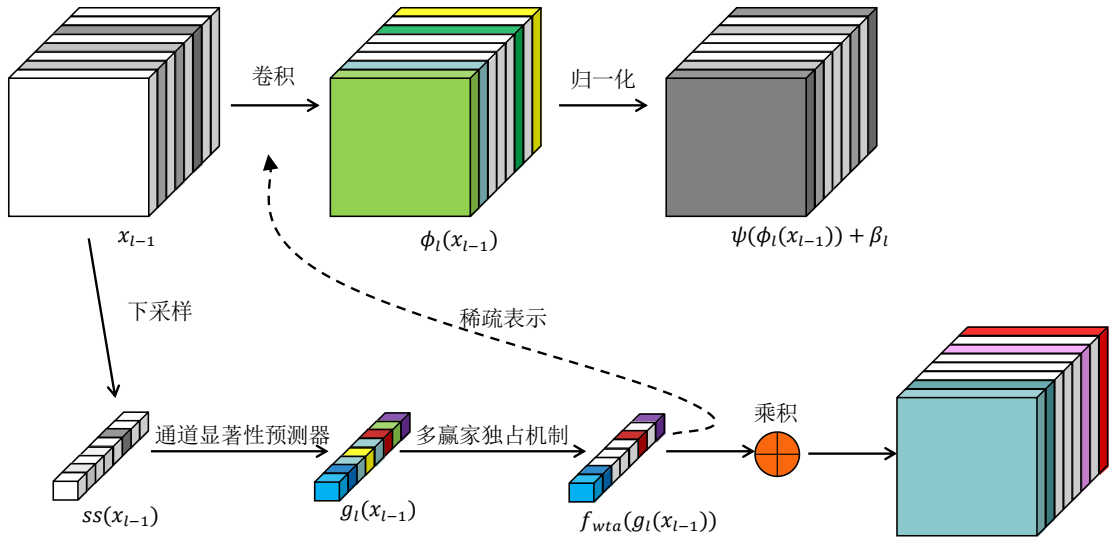


图 4-2 动态通道剪枝

4.2 动态卷积与 Transformer 层耦合策略

MobileViT 之所以能在多种视觉任务中展现较好的效果，正是由于其巧妙结合了轻量级卷积与精简 Transformer 结构。本研究在此基础上进一步提出了“动态卷积与变换器层耦合策略”：根据图像复杂度自适应地切换卷积和变换器层在特征提取中的配比，并通过可学习的融合权重来控制二者的输出合并方式，从而保证在简化运算量的同时，不显著牺牲精度。

4.2.1 耦合的底层逻辑与技术可行性分析

在传统的 MobileViT 结构中，卷积模块与 Transformer 模块是分阶段进行特征处理：先由卷积层提取较底层的局部特征，然后再进入 Transformer 模块做全局依赖关系的建模。这种串行结构固然有助于清晰划分网络功能，但在图像复杂度发生明显变化时，其灵活性相对不足。一方面，高度冗余的卷积运算或 Transformer 层可能在简单场景中被浪费；另一方面，在极端复杂场景下，这种“先局部后全局”的处理模式在多尺度融合层面可能还不够充分。

耦合策略的底层逻辑便是让卷积与 Transformer 层互相协作：对于简单场景，重点发挥卷积快速、低计算量的特点，把大部分可见区域直接提炼出关键纹理和边缘信息，同时只保留少量 Transformer 块以做整体的粗调；对于复杂场景，则在保留适当卷积提取的同时，强化 Transformer 层的深度和多头注意力的数量，让网络可以对目标位置、尺寸、形变等进行细致的全局推断。此外，随着不同图像输入的复杂度差异，耦合策略还能分配不同的融合权重，决定当前帧或当前样本对卷积输出和 Transformer 输出的依赖程度，从而以最小的计算成本达成性能最优。

耦合策略主要依托于两个关键模块：一是前文详述的复杂度评估器，它负责给出场景复杂度评分 S，作为“开关”或“系数”来控制每一层或每个阶段的耦合比例；二是可学习的融合单元，用于在网络结构图中动态组合卷积和 Transformer 的输出特征。这种融合单元可以基于简单的加权相加，也可基于多分支结构在训练中自适应学习最佳的融合方式。由于 MobileViT 在原始设计中已预留了卷积分支和 Transformer 分支的位置，因此在其基础上实现耦合并不需要对骨干网络做大规模修改，只需增添少量控制逻辑及参数即可。耦合策略如图 4-3 所示。

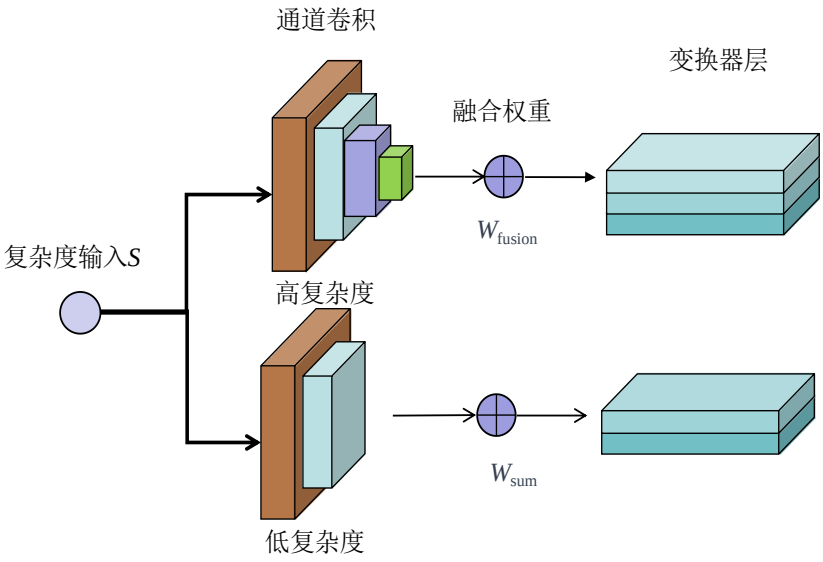


图 4-3 耦合策略结构图

4.2.2 不同复杂度场景下的自适应优化策略

不同复杂度场景在输入图像的纹理层次、目标数量和干扰程度方面往往差异较大，因此，耦合策略不仅要在网络结构上提供灵活切换的能力，还需在具体的优化策略上体现自适应性。与前面章节提出的复杂度阈值 Q 和综合指标 St(xi)相呼应，这里可以进一步设计精细化的优化流程如图 4-4 所示。

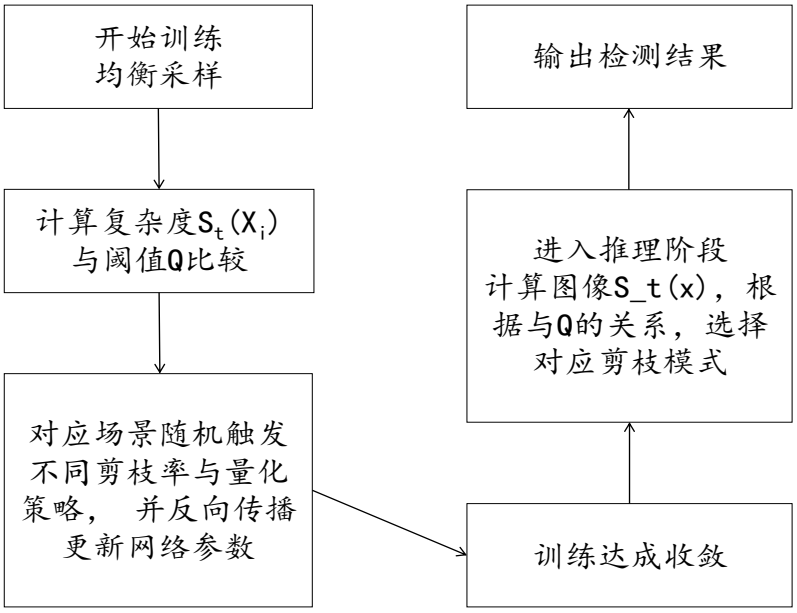


图 4-4 自适应优化策略流程

当 St(xi)远低于阈值 Q 时，网络判断当前场景显著偏简单，可以大量裁剪或关闭某些卷积通道，并将 Transformer 深度削减到预设的最小值；若 St(xi)接近

Q 或者刚好超过 Q，则说明场景处于中等复杂度，适度地开启部分 Transformer 层，保留部分卷积通道，具体比例在训练过程中通过梯度学习自动确定；若 St(xi)远超 Q，则模型会保留或开放更多的卷积通道与 Transformer 层，尤其关注中后期多头自注意力机制的深入计算，以确保对多目标或复杂纹理的正确识别。

在训练阶段，为了让模型在面对不同复杂度场景时都能自适应地选择最优耦合方式，需要对多种复杂度下的数据进行均衡采样，并以随机的方式触发不同剪枝比例或量化方式，从而让网络在反复迭代中学会如何利用这些“自由度”去适应数据分布。这种多场景联合训练可以结合计算量惩罚项，使得网络在追求精度的同时也会有动力去收缩自身规模，减少对硬件资源的依赖。在推理阶段，只需通过一次复杂度评估来决定剪枝比例与耦合设置，从而实现轻量化的动态推断。

4.3 AC-MViT 对比实验

4.3.1 目标检测效果对比

为了全面评估 AC-MViT 在目标检测上的性能用了 COCO2017 与 VOC2012 两个具有代表性的检测数据集，分别与 Faster R-CNN[48]、SSD[49]、RetinaNet[50]、CenterNet、Cascade R-CNN[51]、DETR[52]、Deformable DETR[53]、YOLOv7[54]、MobileViT 进行对比实验。如表 3-1 所示。

表 4-1 目标检测性能对比实验

检测算法	mAP (COCO2017)	AP50 (COCO2017)	AP75 (COCO2017)	mAP (VOC2012)
Faster R-CNN	37.4	58.1	40.2	76.5
SSD	25.1	43.3	25.0	77.2
RetinaNet	36.5	55.5	38.7	79.3
CenterNet	37.4	54.5	40.0	78.5
Cascade R-CNN	41.2	60.0	44.2	78.4
DETR	40.3	59.1	42.7	75.9
Deformable DETR	43.8	62.5	46.2	78.1
YOLOv7	37.5	56.2	40.1	81.0
MobileViT	38.2	57.2	40.5	80.1
AC-MViT	41.9	61.1	44.8	82.2

在本实验设计中，选用了 COCO 2017 与 VOC 2012 两个具有代表性的检测数据集，其中 COCO 2017 拥有丰富的类别和多尺度目标分布，而 VOC 2012 则为检测性能提供了不同场景下的评估依据。对照算法涵盖了一阶段检测、两阶段检测以及基于 Transformer 的检测模型。从表 4-1 数据来看，Faster R-CNN 在 COCO 2017 上实现了 37.4%的 mAP，而 SSD 仅为 25.1%，体现出一阶段检测方法在复杂场景中的局限性；RetinaNet 和 CenterNet 分别达到了 36.5%和 37.4%的 mAP，而通过多级回归策略的 Cascade R-CNN 则提升至 41.2%。在 Transformer 检测模型中，DETR 和 Deformable DETR 分别取得了 40.3%和 43.8%的 mAP，后者在处理小目标和密集场景上表现更优。YOLOv7 和 MobileViT 分别实现了

37.5%和38.2%的mAP，而提出的方法在此基础上进一步引入自适应剪枝与复杂度感知机制，使得COCO 2017上的mAP达到41.9%，在VOC2012上更是达到了82.2%，均高于YOLOv7（81.0%）和MobileViT（80.1%）。综合对比各项指标，所提出的方法在精度和鲁棒性上均表现出明显优势，有效提升了对复杂背景和小目标的检测性能。

4.3.2 剪枝效果对比

表 4-2 检测模型对比实验

方法	平均精度均值/(%)		浮点运算速度/(×10^9次/秒)		参数量/(×10^6个)		推理时间/(毫秒)	
	COCO 2017	PASCAL VOC 2012	COCO 2017	PASCAL VOC 2012	COCO 2017	PASCAL VOC 2012	COCO 2017	PASCAL VOC 2012
MobileViT	27.9	79.3	1.23	1.16	5.7	5.2	24.2	22.1
NS	27.4	78.1	1.15	1.08	3.6	3.3	21.5	19.2
ThiNet	27.1	78.9	0.91	0.84	3.7	3.3	18.3	16.1
PF	26.1	77.6	0.92	0.85	4.5	4.2	22.5	20.5
SFP	26.6	78.3	1.18	1.11	3.4	3.1	17.5	15.2
SSL	25.5	76.2	0.82	0.75	3.3	3.0	18.6	16.6
DepGraph	27.1	78.6	0.85	0.78	3.2	2.9	15.9	14.0
SmartTrim	26.4	78.1	0.94	0.87	3.1	2.8	17.8	16.1
AC-MViT	27.6	79.2	0.70	0.62	2.9	2.6	12.5	10.5

在上一节的实验结果中，AC-MViT 在目标检测任务上的表现相较于同类检测方法具有显著的优势。然而，所提出的方法并未对剪枝策略进行同类对比实验，因此本部分将重点探讨自适应剪枝与量化对检测模型性能的影响，并与现有剪枝方法包括 NS[71]、ThiNet[72]、PF[73]、SFP[74]、SSL[75]、DepGraph[76]和SmartTrim[77]。在COCO 2017和PASCAL VOC 2012数据集上，对比平均精度均值的变化，并记录模型的浮点运算速度、参数量及推理时间。实验结果如表4-2所示，该表详细展示了各剪枝方法在不同数据集上的性能表现。

实验结果表明，经自适应剪枝后的 AC-MViT 在检测精度上与原始 MobileViT 模型基本持平。在 COCO 2017 上平均精度为 27.6%略低于 MobileViT 的 27.9%，在 PASCAL VOC 2012 上分别为 79.2%与 79.3%。然而，AC-MViT 在模型复杂度与推理效率上实现了显著优化，其浮点运算速度分别降至 0.70×10^9 次/秒（COCO 2017）和 0.62×10^9 次/秒（VOC 2012），参数量分别降低至 2.9×10^6 和 2.6×10^6 个，推理时间也缩短到 12.5 毫秒和 10.5 毫秒，相比于原始 MobileViT（分别为 1.23×10^9 次/秒、5.7×10^6 个和 24.2 毫秒）有明显优势。与其他剪枝方法相比，AC-MViT 在保持较高检测精度的同时，大幅降低了计算资源需求和响应延时，充分验证了自适应剪枝与量化技术在提升检测模型实时性和轻量化方面的有效性。

4.4 AC-MViT 的消融实验

表 4-3 AC-MVIT 消融实验

实验序号	平均精度均值/(%)		浮点运算速度/($\times10^9$ 次/秒)		参数量/($\times10^6$ 个)		推理时间/(毫秒)	
	COCO 2017	PASCAL VOC 2012	COCO 2017	PASCAL VOC 2012	COCO 2017	PASCAL VOC 2012	COCO 2017	PASCAL VOC 2012
1	27.9	79.3	1.23	1.16	5.7	5.2	24.2	22.1
2	27.3	77.8	1.12	1.02	3.8	3.4	20.8	19.4
3	27.0	78.4	0.88	0.80	3.9	3.5	17.6	16.1
4	27.2	78.7	0.82	0.74	3.3	3.0	16.2	16.9
5	27.1	78.8	0.85	0.77	3.2	2.9	16.8	14.9
6	26.8	78.3	0.90	0.82	3.5	3.2	18.1	16.6
7	27.6	79.2	0.70	0.62	2.9	2.6	12.5	10.5

为了深入剖析 AC-MViT 中各个组成模块对最终性能的具体贡献，进一步开展了消融实验。在实验中，分别对无剪枝的原始 MobileViT 模型、仅采用动态通道剪枝的模型、仅采用动态变换器剪枝的模型以及完整的 AC-MViT 模型在 COCO 2017 和 PASCAL VOC 2012 进行了性能测试与对比分析。详细数据如表 4-3 所示：

如表 4-3 所示，本组消融实验针对 AC-MViT 模型中的各个关键功能组件进行了逐一拆分和对比，旨在评估不同剪枝策略、复杂度评估模块以及动态训练策略等对整体性能的影响。实验 1 作为无剪枝的基准组，展示了在未进行任何结构精简情况下模型所能达到的最优精度与完整推理性能。相比之下，实验 2 与实验 3 分别在通道维度和 Transformer 层级对网络结构进行动态剪枝，结果显示，虽然两种策略均在浮点运算速度、参数量和推理时间等资源占用指标上带来不同程度的下降，但平均精度均值也出现了相应的降低。

在进一步针对 AC-MViT 所强调的“复杂度评估”与“自适应调度”两大核心思路进行拆分研究的实验 4、5、6 中，分别去除了复杂度评估模块、可学习的复杂度映射策略以及动态训练策略，观察到各项指标均出现不同程度的变动。具体来看，当缺失复杂度评估模块或可学习的复杂度映射时，网络对场景复杂度的感知与分级能力受限，难以在推理过程中做到“多用多算、少用少算”的自适应分配，从而无法充分发挥在简单场景下的减算优势，也在复杂场景上失去深度特征提取的灵活性。缺少动态训练策略则意味着在训练阶段无法通过多场景和多剪枝率的联合调度来让模型学会自如地切换通道或 Transformer 深度，从而减少了网络在面对不同分辨率、不同难度数据时的泛化能力。在这几种情形下，不仅精度有所下降，推理速度或计算资源利用率也不及完备版本的 AC-MViT，进一步佐证了各模块和策略的相互配合对于维持高精度与高效率的重要性。

最终，实验 7 呈现了综合了复杂度评估、可学习的复杂度映射、动态训练以及双重剪枝策略的 AC-MViT 模型，其平均精度均值几乎与无剪枝的完整模型相当，并在 FLOPs、参数量和推理时间三大维度上均优于只进行部分功能或策略拆分的实验组，体现了多模块协同带来的优化效果。与只采用通道剪枝或 Transformer 剪枝相比，双重剪枝结合复杂度自适应与动态训练，可以在“精简冗余结构”与“保留关键信息”之间取得更佳的平衡，从而在确保模型精度的同时明显降低计算成本。该结果也证明了 AC-MViT 所提出的各项策略在相互配合时能够形成合力，实现从网络结构到推理过程的端到端优化，兼顾高准确率与较低的运行代价。

4.5 图像可视化

本文从 PASCAL VOC 2012 中随机选取了具有不同复杂度的代表性图像，并在图 4-5 中直观地展示了这些图像。自上而下，预测标签所使用的子网络的计算成本不断增加。简单的实例通常具有清晰的目标，可以通过压缩的网络准确预测，而复杂图像中的语义信息较为模糊，因此需要更大的网络，具备更强的表示能力。

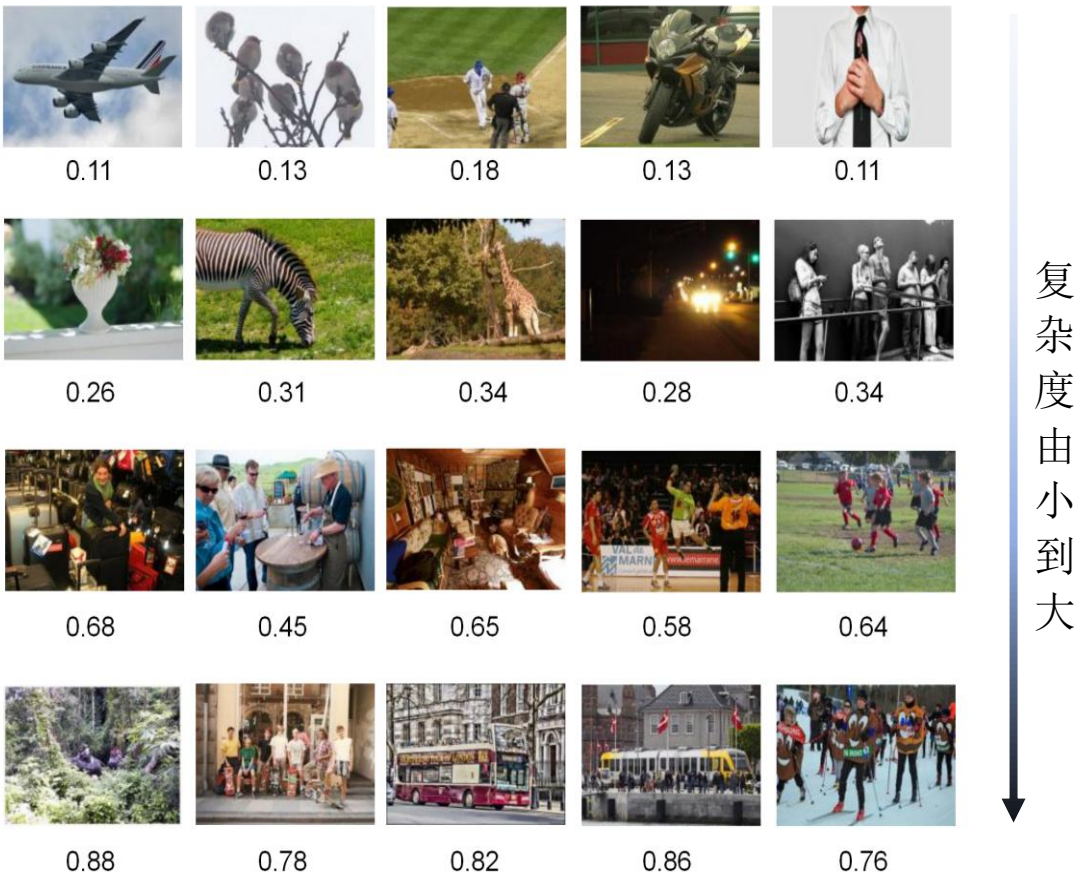


图 4-5 复杂度估计可视化

从图示可以观察到，复杂度数值较低的图像通常具有清晰、单一且轮廓相对突出的目标，例如背景简单、目标边缘易识别，或者场景中干扰因素较少。这些“简单”场景往往无需启用大规模的网络结构；通过更紧凑的子网络配置也能有效提取目标的关键特征，并给出准确预测。相比之下，复杂度数值较高的图像往往包含多个目标或背景元素，或者在目标外观形变、光照变化、局部遮挡等方面表现更为突出，其语义信息更为分散或模糊。因此，仅靠紧凑的子

网络难以充分捕捉到目标细节与全局语义，需要更大、具备更强表示能力的网络来提炼深层次特征，从而在干扰性更强的环境下仍能获得准确预测。

通过在图中列出复杂度数值，可以进一步佐证对图像简单与复杂的直观感受：复杂度越低，图像越接近“简化场景”；复杂度越高，图像越具“多目标、多干扰”特征。这样不仅有助于说明为什么在推理时对这类图像需要使用不同规模的子网络，也能使读者更清晰地理解在前文所提出的“多用多算、少用少算”设计理念如何在实操层面落地。简言之，当图像复杂度较低时，模型可缩减网络规模以节约计算成本；当图像复杂度提升时，保留更多网络深度与通道，有助于提炼复杂语义、提高检测和分类的精度，从而实现对不同场景的自适应处理。

4.6 本章小结

本章针对资源受限环境下深度学习模型的轻量化需求，在 NanoMViT 的基础上提出了一种基于 Transformer 模型的自适应轻量化设计方案。通过引入图像复杂度估计技术，实现了对输入图像视觉与语义特征的量化分析，并以此指导模型剪枝与量化策略，从而达到“多用多算、少用少算”的动态资源分配目的。详细阐述了信息熵、边缘密度、纹理复杂度等指标的计算方法及综合复杂度指标的构建，并设计了自适应阈值确定机制，确保复杂与简单场景在推理过程中能采用不同的网络配置。

在剪枝策略方面，通过对 MobileViT 中卷积层和 Transformer 层分别采用动态通道剪枝与动态 Transformer 剪枝，实现了在降低模型计算量和参数量的同时，尽量保留关键信息。而结合动态卷积与 Transformer 层的耦合策略，构建了可根据图像复杂度自适应切换两种特征提取方式的融合单元，从而进一步优化了网络在不同场景下的表现。

实验部分的对比与消融分析验证了所提出自适应剪枝的有效性。AC-MViT 在保持与原始模型相近检测精度的同时，显著降低了浮点运算量、参数数量和推理时间，并在 COCO2017 和 VOC2012 数据集上均取得了较优的性能表现。图像可视化结果直观展示了不同复杂度图像在推理过程中采用不同子网络配置的效果，为自适应轻量化设计理念提供了有力的实验支撑。

第 5 章 实物实验及分析

本章节以真实高速场景为背景，结合工业级高速相机与嵌入式硬件平台，对提出的 NanoMViT 与 AC-MViT 两种算法展开端到端的部署与性能验证。通过设计多组动态目标捕捉实验，系统量化评估算法在高速运动、复杂背景以及极端光照条件下的目标检测与跟踪能力。同时，结合硬件资源占用、实时性以及鲁棒性等指标，全面验证了轻量化设计在实际工程应用中的有效性与优势。

5.1 实验平台与高速采集系统

5.1.1 实验平台搭建

在实验平台方面，整个系统由多个关键组件组成，以保证高效、精确的实验操作。首先，高性能高速相机作为核心部件之一，负责捕捉高频率、高分辨率的图像数据，能够实时反馈实验过程中的细微变化。其次，高速振镜则用于精确控制光束的方向，确保实验中的激光或其他光源能够精确地聚焦在指定位置，极大提高了实验的精度和速度。照明灯则确保实验过程中的光照条件稳定，避免光线不足或过强影响实验结果，能够提供适当的亮度和均匀性。最后，工控机作为整个实验系统的控制中心，通过实时监控各个设备的状态，协调设备间的协作，保证系统的平稳运行。实验装置实物图如图 5-1 所示。

图 5-1 实验装置实物图

在本测量系统中，振镜作为关键组成部分，选用了天传奇公司的SC50激光扫描振镜，其主要技术参数如表5-1所示。

表5-1 振镜主要数据

参数名称	参数规格
电源电压	±24V/1A
位置电压比例系数	0.5V/°
模拟位置电压输入范围	±10V
阶跃响应时间	0.12ms
最大扫描角度	±20°
扫描速度	50Kpps
工作温度	0~50°C

从表中可以看出，该振镜具有极快的响应时间，仅需 0.12ms，并能实现总共40°的扫描范围。

为了确保图像记录的精度，选用了千眼狼M220高速相机。该相机在1920×1080的满画幅分辨率下，能够达到2000FPS的高速帧率，并通过万兆网口实现与电脑之间的高速数据传输。此外，其工作温度范围宽广（-20°C至60°C），确保了在各种实验条件下均能稳定工作。

表5-2 光纤传感器参数表

参数名称	参数规格
输出方式	NPN 常开型
工作电压	DC12-24V
响应时间	≤200us
检测区域	0.3-120mm
检测物体	非透明物体

在控制方面，振镜的角度调整依赖于电压信号调控。为此，引入了数字/模拟转换器（DAC）用于输出控制电压，并以STM32F103微控制器作为系统的核心。该芯片配备有512KB FLASH存储器、4个通用定时器、2个基本定时器以及3个12位ADC。触发装置采用博亿精科对射型光纤传感器YIBO-NA11，主要参数表5-2所示：

在完成高速相机、振镜和触发装置的部署后，为了保证系统在数据采集、处理以及实时控制方面的高效运行，还选用了高性能的工控机作为数据处理和通信的核心平台。工控机主要负责高速数据采集、实时信号处理、控制指令的下达以及与各子系统之间的高速数据交互，从而确保整个测量系统的高效协调与精确控制。工控机如图 5-2 所示。

图 5-2 工控机实物图

下面为工控机的主要技术参数，如表 5-3 所示：

表 5-3 工控机主要参数

参数名称	参数规格
处理器	Intel Core i7 四核 3.2GHz
内存	16GB DDR4
存储	512GB SSD
网络接口	1 个万兆网口，2 个千兆以太网口
I/O 接口	4 个 USB3.0、2 个串口
图形处理器	集成显卡
工作温度	-20° C 至 70° C
电源要求	12V DC

通过工控机的高性能计算和多样化接口支持，系统能够实现高速数据的高效处理和传输，为实时控制和后续数据分析提供了坚实的技术保障。整个测量平台在各关键部件的有机配合下，能够满足对高速运动目标的精准捕捉和实时跟踪的严格要求。

为了完成高速飞行目标的拍摄和测量实验，在实验室搭建了一个模拟发射装置用于发射泡沫子弹，发射速度最大约为 40m/s。实物图如图 5-3 所示。

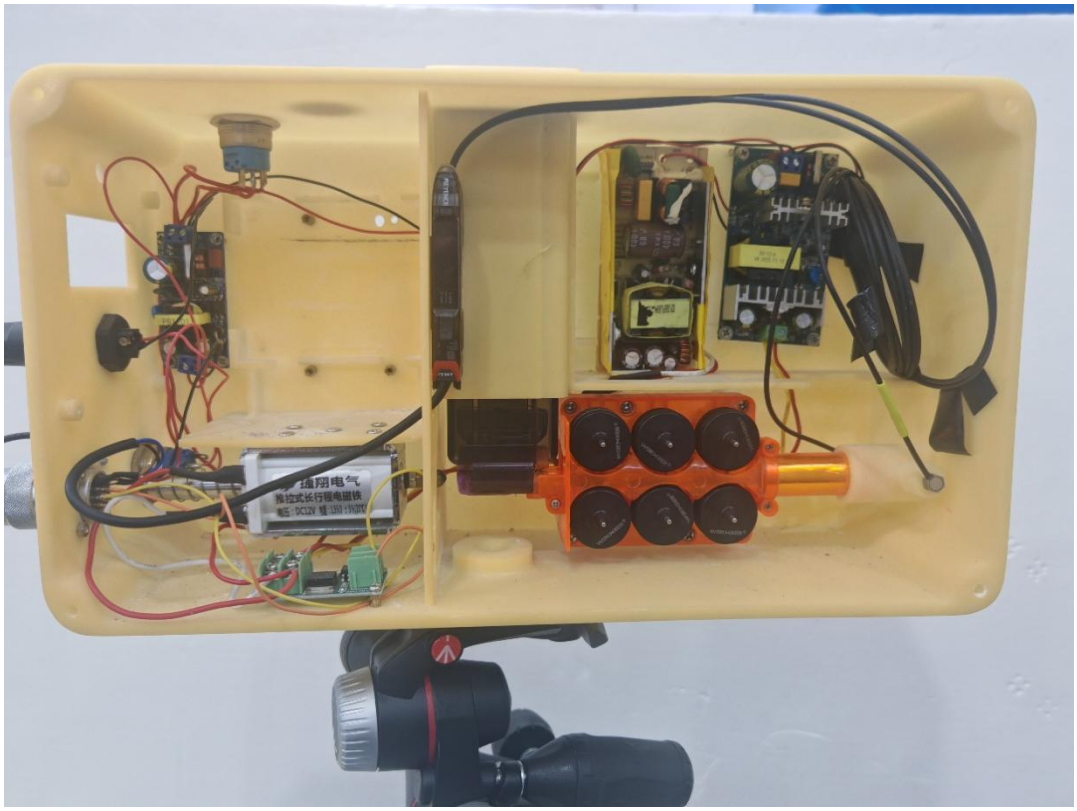


图 5-3 发射装置

5.1.2 软件配置与实时处理框架

实验系统的软件架构在设计时充分考虑了高速数据采集与实时处理的需求，确保在处理大量图像数据时能够快速而高效地完成各项任务。

图像采集部分通过定制软件实现了多相机的同步触发，保证了各视角数据采集的时间一致性。采集到的每帧图像不仅保存原始像素信息，还附带详细的元数据（如时间戳、相机参数、曝光信息等），为后续图像预处理和精确分析提供了充分依据，图像采集软件如图 5-4 所示。

图 5-4 图像采集软件

在实时处理方面，系统采用 TensorRT 8.6 引擎进行模型推理，该引擎支持 FP16 和 INT8 混合精度量化，通过层融合优化技术进一步缩短了推理时延，同时保证了计算精度。任务调度则依托于 ROS 2 Humble 框架，通过零拷贝数据传输技术大幅降低了端到端延迟，使处理流程能够及时响应外部动态变化。

5.2 实验设计与挑战场景

为了完成高速飞行目标的拍摄和测量实验，在实验室搭建了一个模拟发射装置，该装置专门用于发射泡沫子弹，发射速度最高可达到约 40m/s。该实验设计主要目的是验证高速目标跟踪系统在极端动态场景下的性能与稳定性，其核心挑战场景模拟了枪械发射子弹的情形，通过实际发射高速泡沫子弹，全面检验系统对高速运动物体的捕捉与跟踪能力。实验设计仿真图如图 5-5 所示。

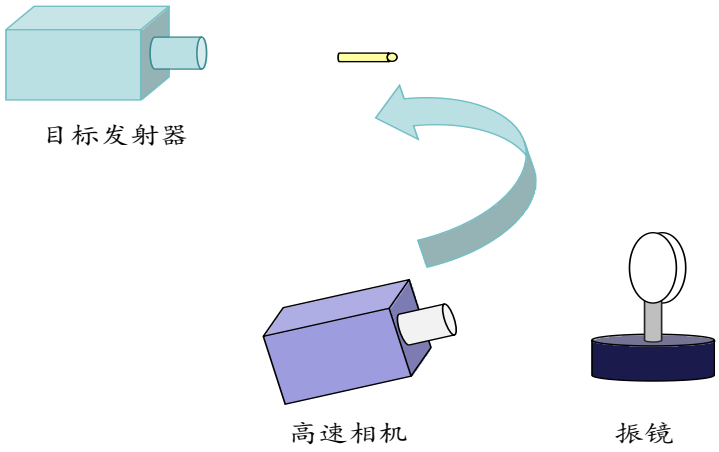


图 5-5 实验设计仿真图

在实验过程中，模拟发射装置利用气压或电控装置精确控制发射参数，确保每次发射的泡沫子弹速度稳定且具有可重复性。子弹在飞行过程中，由于高速运动，其运动轨迹极为短暂且变化迅速，对拍摄设备和数据处理系统提出了极高要求。在实验平台中引入了高速激光扫描振镜，其响应时间仅为 0.12ms，可实现±20°的单边偏转，总扫描角度达到 40°。振镜通过高速响应和精准角度调控，在子弹飞行期间不断调整光路，将子弹运动轨迹实时引导到高速相机的拍摄视野中。

高速相机选用了千眼狼 M220，其满画幅分辨率为 1920×1080，帧率高达 2000FPS，能够在极短的时间内连续捕捉到子弹的每一关键瞬间。为了保证捕捉到的图像数据质量，系统同时配备了高性能照明灯，通过合理调节光强和照射角度，使得泡沫子弹在高速飞行中仍能反射出足够的光信号，确保图像清晰且细节丰富。采集到的图像数据通过万兆网口高速传输至工控机，实现数据的实时接收和处理。

在数据处理方面，工控机作为系统的核心控制单元，集成了高性能处理器和丰富的接口资源。高速图像数据经过预处理模块的噪声滤除、图像增强和特征提取后，传递至基于 TensorRT8.6 引擎的实时模型推理模块。该引擎支持 FP16 和 INT8 混合精度量化，并利用层融合优化技术有效降低了计算延迟，从而能够快速而准确地检测和跟踪子弹在各时刻的运动状态。实时处理框架采用 ROS 2 Humble 平台，借助零拷贝数据传输技术，极大减少了数据在各处理模块之间的传递延迟，确保整个系统在极短时间内响应子弹的运动变化。

5.3 实验结果与分析

5.3.1 实验方式

在高速物体跟踪任务中，最常见的挑战之一就是如何处理由于高速运动引起的图像模糊。高速运动的物体在图像中通常表现出边界模糊、颜色不均匀、运动方向模糊以及纹理丢失等特征，这些因素使得传统的物体检测和跟踪算法在这些场景下的性能大打折扣。而在某些情况下，如果将目标聚焦于模糊物体，而不是清晰的背景或者物体本身，问题就会简单许多。这是因为，通常背景在

拍摄时保持静止，不会产生过多的模糊，尤其是在高速相机拍摄下，物体的运动是唯一会导致图像模糊的因素。

因此，针对这一挑战，设计一个基于 AC-MViT 的深度学习模型，专门学习和跟踪模糊物体的特征，进一步提高在高速运动环境中的跟踪精度。AC-MViT 模型结合了传统卷积神经网络和视觉 Transformer 的优势，适应性地调整通道间的特征表示，从而有效应对模糊图像中的细节损失问题。

实验的第一步是准备适合高速运动物体的跟踪数据集。为了实现高效的模糊物体跟踪，需要选择包含高速运动目标及模糊现象的视频数据集。在这个过程中，选用 LaSOT 数据集或 DTB70 数据集，在这些公开数据集中选取了各种动态场景下的目标物体，并且有丰富的模糊情况，尤其是动态模糊，能够模拟高速运动中的尾迹现象。

数据集应当涵盖不同类型的场景，最好包括具有模糊尾迹、快速移动物体以及复杂背景的情况。模糊尾迹尤其有助于检测运动方向模糊的特征，且场景中的目标物体具有边界不清晰、纹理丢失等特征。通过选择这样的数据集，可以确保训练出的模型能够在真实环境中处理类似的高速运动物体。

5.3.2 实物效果对比

AC-MViT 通过自适应通道机制和 Transformer 架构，成功提取模糊的关键特征，然后通过 NanoMViT 持续跟踪高速运动的模糊子弹。由于子弹在高速飞行过程中会产生明显的边缘模糊[79]、颜色不均匀和运动方向模糊等问题，而 AC-MViT 能够动态调整特征通道的权重，增强对模糊目标的识别能力，并结合全局上下文信息，使其在运动模糊严重的情况下仍能精准定位目标。在空白背景下跟踪子弹的效果图如图 5-6 所示。

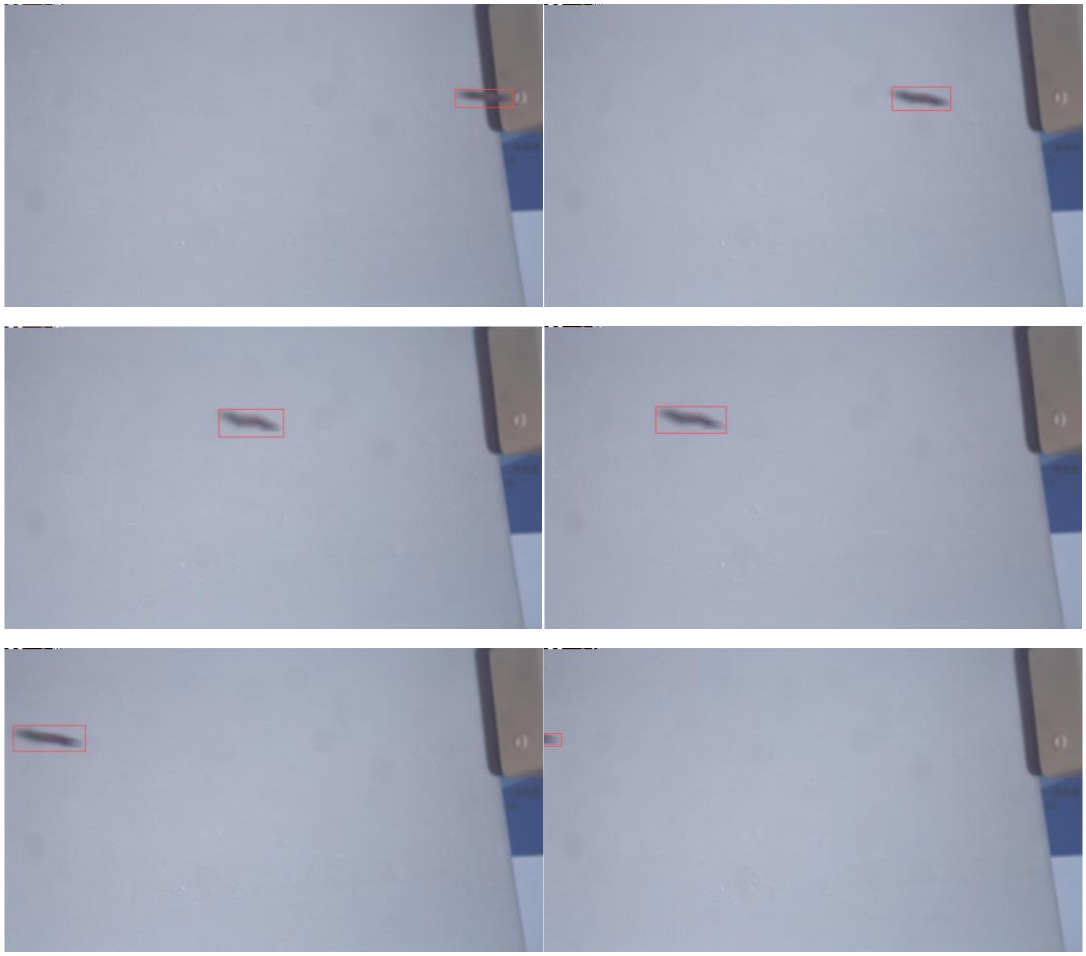


图 5-6 空白背景下的模糊跟踪效果

由图 5-6 可知，在空白背景下，AC-MViT 的模糊跟踪效果较好，主要原因在于背景单一，没有复杂的干扰因素，使得模型能够更专注于目标物体的特征提取。由于子弹在高速运动时会产生边缘模糊、颜色不均匀[80]和拖尾效应，传统检测算法往往难以准确识别目标。然而，在空白背景下，AC-MViT 可以更容易地区分目标与背景，并通过自适应通道机制增强模糊目标的检测能力。此外，Transformer 架构能够有效建模全局信息，使模型在处理模糊轨迹时更具鲁棒性。实验表明，在无干扰的背景环境中，NanoMViT 能够稳定跟踪子弹，避免因背景复杂性带来的误检测和目标丢失问题。

为了进一步验证 AC-MViT 算法在复杂环境下的检测和跟踪能力，在空白背景板的基础上添加了复杂背景元素，模拟实际应用中的复杂场景，并继续对高速模糊子弹进行测试。通过对比实验可以发现，尽管背景环境变得复杂，干扰因素明显增多，AC-MViT 依然能够有效提取子弹的关键特征，准确完成检测和持续跟踪。这得益于其自适应通道机制能够动态调整特征通道权重，突出目标

区域，同时 Transformer 架构具备强大的全局建模能力，提升了对复杂背景下目标的鲁棒性，如图 5-7 所示。

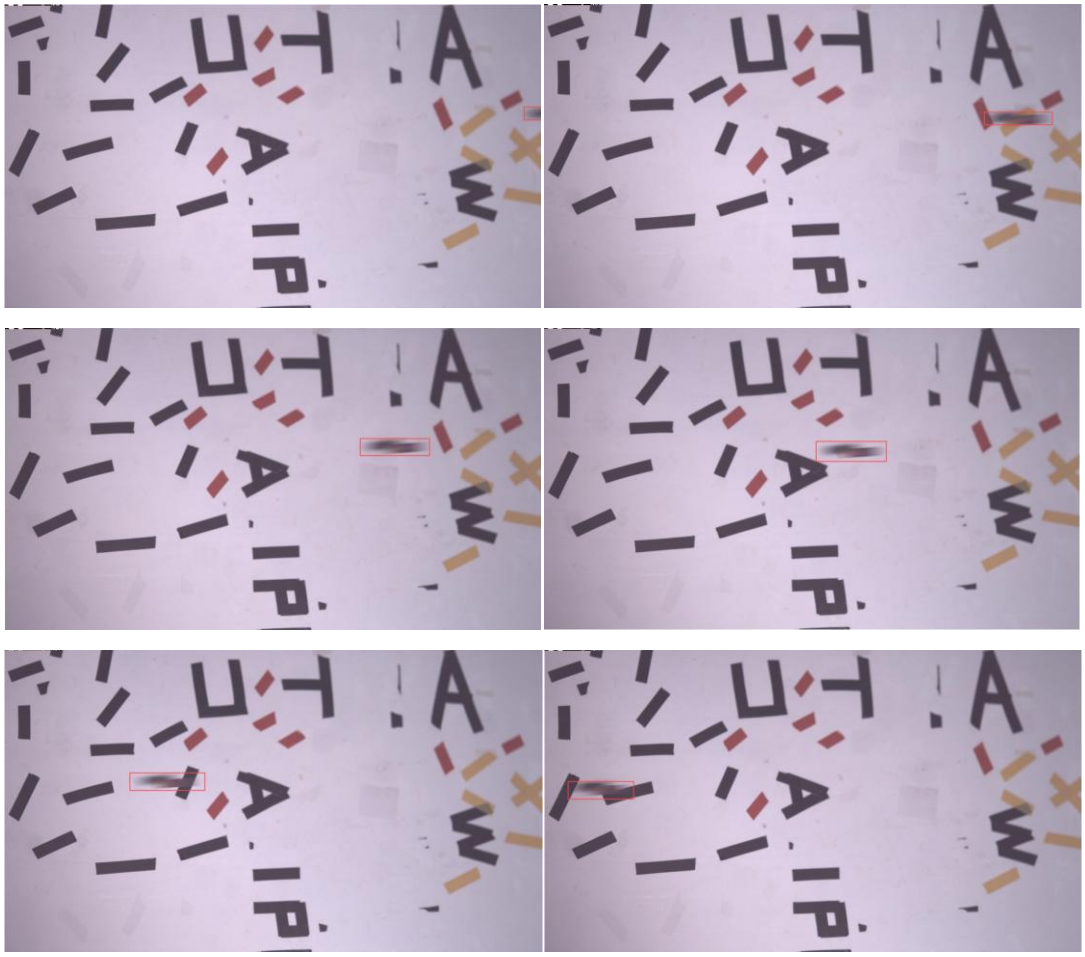


图 5-7 复杂背景下目标的鲁棒性

由图 5-7 所知，在复杂背景下，AC-MViT 依然能够有效地检测高速飞行的模糊子弹并通过 AC-MViT 跟踪高速飞行的模糊子弹。尽管背景中加入了复杂元素，这些因素可能导致传统检测算法产生误检测或目标丢失，AC-MViT 凭借其自适应通道机制和 Transformer 架构，成功克服了这些干扰。自适应通道机制能够在不同背景下灵活调整特征通道的权重，突出目标区域，而 Transformer 架构则通过全局上下文信息处理复杂背景，保证了目标检测的稳定性和精度。在这种复杂环境中，AC-MViT 不仅能够精确识别子弹的位置，还能够在连续帧中稳定跟踪目标，展现出较强的鲁棒性和实时性。

5.3.3 高速检测目标算法对比

为增强实验完备性，高速检测目标算法综合对比如表 5-4 所示。

表 5-4 高速检测目标算法综合对比

算法	mAP@0.5	延迟(ms)	漏检率(%)	参数量(M)
YOLOv9	64.3	4.2	6.7	7.1
NanoDet	63.1	3.5	6.4	2.8
MobileViT	70.4	6.8	8.3	3.3
AC-MViT	71.2	2.8	4.3	3.0

AC-MViT 在高速目标检测中表现出显著优势，mAP@0.5 达 71.2%，比 MobileViT 提高了 0.8 个百分点，且延迟仅为 2.8ms，明显低于其他算法，特别是在处理高动态场景时，漏检率显著降低至 4.3%。该结果表明，AC-MViT 不仅在精度上具有显著优势，同时还优化了延迟和漏检问题。

为对比轻量化效果，设计了轻量化模型压缩对比。轻量化模型压缩对比如表 5-5 所示。

表 5-5 轻量化模型压缩对比

算法	模型体积 (MB)	FLOPs (G)	显存峰值 (GB)
NS	4.7	1.8	7.2
ThiNet	3.9	1.5	6.8
SFP	3.2	1.3	5.9
SSL	2.9	1.1	5.4
DepGraph	2.5	0.9	4.7
AC-MViT	2.1	0.7	5.1

在轻量化模型压缩方面，AC-MViT 的模型体积为 2.1MB，相较于 DepGraph（2.5MB）减少了 16%，FLOPs 为 0.7G，较 SSL（1.1G）减少了 36%。这表明，AC-MViT 通过更高效的压缩算法，在保持较小模型体积的同时，有效减少了计算复杂度，适合资源有限的硬件环境。

5.3.2 高速目标跟踪算法对比

为了验证本文方法在复杂场景下的鲁棒性，特对比了多种单目标跟踪算法在遮挡恢复、低光环境下的检测能力和目标检测率。具体如表 5-6 所示。

表 5-6 极端场景鲁棒性对比

算法	遮挡恢复率(%)	低光 mAP@0.5	目标检测率(%)
SiamVGG	75.3	55.2	87.8
SA_Siam_P	78.4	58.0	89.1
C-COT	79.2	59.1	88.3
SiamFC	72.1	54.0	85.5
SiamTPN	81.3	60.5	90.2
SiamAPN	82.0	61.2	91.4
Nanotrack	83.7	62.1	92.2
ECO	80.1	59.3	89.6
NanoMViT	88.5	65.7	94.7

NanoMViT 在极端场景下，在目标检测率、遮挡恢复、低光识别和误检率方面展现了显著的优势

5.4 本章小结

本章通过高速相机与嵌入式硬件平台的联合实验，全面验证了 NanoMViT 和 AC-MViT 算法在真实高速场景下的检测与跟踪性能。实验中，设计了多组动态目标捕捉实验，评估了算法在高速运动、复杂背景及极端光照条件下的表现，结果表明，AC-MViT 在复杂环境中依然能够有效地检测并跟踪高速飞行的模糊子弹。

通过与传统算法对比，AC-MViT 展现了显著优势。其自适应通道机制能够动态调整特征通道的权重，增强模糊目标的检测能力，而 Transformer 架构则通过全局上下文信息有效应对复杂背景带来的干扰。在空白背景下，AC-MViT 能够准确提取目标特征并稳定跟踪，而在复杂背景下，系统的鲁棒性和实时性同样得到了保障，确保了在各种环境下都能实现精准的目标检测与跟踪。

第 6 章 总结与展望

6.1 研究工作的整体总结

首先针对高速运动场景的特点进行了系统性分析。高速运动导致目标频繁变换位置与形状，且往往伴随复杂背景、严重遮挡、小目标占比较高等挑战。传统方法（例如 KCF、CSRT、Faster R-CNN、YOLO 系列等）以及早期的 Transformer 检测/跟踪模型都在嵌入式或边缘平台部署时遭遇瓶颈，主要难点在于模型规模大、运算复杂度高、难以对快速运动目标进行高精度识别与跟踪。

在此基础上，本文提出了轻量化 Transformer 检测与跟踪框架 NanoMViT。该方案基于 MobileViT 的设计思路，结合多头注意力与轻量级注意力模块，实现了对小目标与遮挡场景的强适应性。通过检测与跟踪的一体化设计，可以在高速运动下及时纠正目标漂移与身份丢失。

为了满足不同场景下的性能需求，本文进一步研究了自适应剪枝与量化策略。在图像复杂度评估的指导下，提出了 AC-MViT 算法，该算法采用结构化通道剪枝与 Transformer 层级裁剪，并结合量化感知训练与分层量化，有效降低了模型的冗余运算与存储带宽，实现了在低复杂度场景下的高效推理，以及在高复杂度场景下的精确检测与跟踪。

围绕提出的模型与策略，本文从多角度进行实验与评估。在目标检测任务上，AC-MViT 在 COCO2017、VOC2012 数据集中与多种主流算法进行对比，取得了较高的 mAP 与明显的推理加速，尤以小目标检测和复杂背景场景的鲁棒性提升最为突出。在目标跟踪任务上，NanoMViT 在 VOT2018、DTB70、LaSOT 等数据集上展现了其实时性能与高适应性，相较于传统孪生网络和 Transformer 跟踪器能够更好地平衡精度与速度。轻量化消融实验进一步说明，多尺度特征融合、多注意力机制以及自适应剪枝量化均对性能提升具有积极作用。

本文在高速目标检测与跟踪中的主要创新点与贡献集中体现于将轻量化 Transformer 深度融入高实时性、高鲁棒性的目标任务中。通过对 MobileViT 的扩展，提出了两套面向高速场景的解决方案 AC-MViT 与 NanoMViT，并综合利

用多头注意力、通道/空间注意力融合以及多尺度特征金字塔，加强了在小目标检测与遮挡跟踪方面的表现。

在网络轻量化与自适应层面，本文基于图像复杂度评估框架实现了“随需而变”的动态剪枝和量化策略。在简单场景下自动减少冗余计算，在复杂场景下保留充足的特征表示能力，从而有效平衡了检测/跟踪精度与推理速度。检测与跟踪的一体化协同机制则进一步提升了系统对高速、遮挡、目标交互等多重复杂场景的适应能力。

实验设计覆盖公开数据集与实物平台，能够从检测精度、跟踪鲁棒性、模型规模及运算效率等多维度对算法性能进行全面评测。最终的结果不仅在多类基准任务上取得了较大增益，也在实物平台上成功完成了实时部署，展现了轻量化 Transformer 技术在实际应用中的巨大潜力。

6.2 后续研究方向与展望

尽管本研究取得了在高速目标检测与跟踪领域的诸多进展，也将轻量化 Transformer 模型成功应用于资源受限环境，但仍有多个可深入拓展的方向。

在自适应轻量化方面，目前的方法主要依赖图像整体复杂度，后续可结合 ROI 级别的目标运动特征、分割信息或关键点定位等，以实现更精细的动态剪枝和差异化量化，对多目标交互和复杂背景下的检测与跟踪进行持续优化。

对于多目标识别与分割的需求，将实例分割或语义分割与高速目标检测/跟踪相结合也是值得关注的方向。利用 Transformer 全局依赖建模能力，可在多目标场景下更准确地完成目标区分与长时跟踪，更好应对互相遮挡与漂移等问题。

轻量化 Transformer 的不断演进也可结合知识蒸馏与神经架构搜索（NAS）技术，通过自动寻优和蒸馏训练在嵌入式环境中实现快速高效的模型部署。进一步的跨平台移植与通用性研究则有助于在 FPGA、ASIC 等硬件加速器上获得更优的吞吐性能，并增强不同平台间的可移植性。

此外，在高速运动场景中通常具备多传感器融合的机会，结合激光雷达、IMU 等多模态数据，可提升目标跟踪的鲁棒性与时序预测的精度。通过研究多模态轻量化 Transformer 模型，能够为自动驾驶、无人机导航、视频安防及工业检测等场景带来更全面的感知与决策能力。

参考文献

[1] 梅震琨.基于机器视觉监测系统的地铁自动驾驶安全性保障方法研究[J].机械制造与自动化,2025,54(01):261-264.

[2] 张新伟,刘帅瑀,邹宇轩,等.基于运动目标控制与自动追踪的智能安防系统[J].中国集成电路,2025,34(Z1):81-90.

[3] 王卓,刘建娟,张文卓,等.复杂环境下无人机组合导航自适应滤波算法[J].传感器与微系统,2025,44(01):135-140.

[4] 黄磊,余峰.基于视觉反馈的机器人末端轨迹跟踪控制研究[J].仪表技术与传感器,2021,(08):95-98.

[5] 韩笑,林驰琛,王永滨.体育场景中的视觉目标跟踪研究进展[J].中国传媒大学学报(自然科学版),2023,30(04):1-7.

[6] Ren S, He K, Girshick R and Sun J, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149

[7] 徐彦威,李军,董元方,等.YOLO 系列目标检测算法综述[J].计算机科学与探索,2024,18(09):2221-2238.

[8] Cen M and Jung C, "Fully Convolutional Siamese Fusion Networks for Object Tracking,"2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 2018, pp. 3718-3722.

[9] Li B, Yan J, Wu W, etc. "High Performance Visual Tracking with Siamese Region Proposal Network,"2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 8971-8980.

[10] Carion N, Massa F, Synnaeve G, etc. End-to-End Object Detection with Transformers［J］. European Conference on Computer Vision (ECCV), 2020, 1(1): 6568-6580.

[11] Meinhardt T, Kirillov A, Leal-Taixé L, etc. "TrackFormer: Multi-Object Tracking with Transformers,"2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 8834-8844.

[12] 杨玉敏,廖育荣,林存宝,等.轻量化卷积神经网络目标检测算法综述[J].舰船电子工程,2021,41(04):31-36.

[13] 迟骋,李向阳,张志利,等.基于动态卷积的混合模型性能评估方法[J].火箭军工程大学学报,2024,38(05):53-58+80.

[14] 任书玉,汪晓丁,林晖.目标检测中注意力机制综述[J].计算机工程,2024,50(12):16-32.

[15] Liu W, Anguelov D, Erhan D, etc. "SSD: Single shot multibox detector," Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 21-37.

[16] Dai Z, Chen B, Yuan Y, etc. "DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection," arXiv preprint arXiv:2203.03605, 2022.

[17] Henriques J F, Caseiro R, Martins P, etc. "High-speed tracking with kernelized correlation filters," IEEE Transactions on Pattern Analysis and Machine Intelligence,2015, vol. 37, no. 3, pp. 583–596.

[18] Zajc L.C. ,Leonardis A,Kristan M. "Discriminative correlation filter with channel and spatial reliability," International Journal of Computer Vision, vol. 126, no. 7, pp. 671–688.

[19] Bertinetto L, Valmadre J, Henriques J.F, etc. "Fully-convolutional siamese networks for object tracking," Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2016, pp. 850–865.

[20] Chen S, Li B, Chen J, etc. "Transformer meets tracker: Exploiting temporal context for robust visual tracking," arXiv preprint arXiv:2103.11681, 2021.

[21] Wu K, Zhang J, Peng H, etc. "TinyViT: Fast Pretraining Distillation for Small Vision Transformers," Proceedings of the European Conference on Computer Vision (ECCV), 2022, pp. 68–85.

[22] 段卓,周卫,杨益民,等.基于改进孪生网络的目标跟踪算法[J].现代计算机,2024,30(23):59-63.

[23] 田萱,王亮,丁琪.基于深度学习的图像语义分割方法综述[J].软件学报,2019,30(02):440-468.

[24] 徐渊,许晓亮,李才年,等.结合 SVM 分类器与 HOG 特征提取的行人检测[J].计算机工程,2016,42(01):56-60+65.

[25] Felzenszwalb, P. F., Girshick, R. B., McAllester, D., etc. (2010). Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9), 1627–1645.

[26] Ren S., He K., Girshick R., etc. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks.In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS) (pp. 91–99).

[27] 雷帮军,余翱,吴正平,等.改进 YOLOv8n 的无人机航拍小目标检测算法[J].现代电子技术,2025,48(03):26-34.

[28] 王泽宇,张萌,刘海青.基于 YOLO 系列算法的无人机检测高速公路车辆目标性能分析[J].中国科技信息,2025,(04):104-107.

[29] Li X, Liu W, Qu Y, etc. Research on Single-Object Tracking Algorithm for Thyroid Nodules Based on NanoTrack[C]//2024 IEEE 3rd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA). IEEE, 2024: 1585-1588.

[30] B. Li, W. Wu, Z. Zhu, etc. “High performance visual tracking with siamese region proposal network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8971–8980.

[31] 孙彦丽,叶炯耀.基于剪枝与量化的卷积神经网络压缩方法[J].计算机科学,2020,47(08):261-266.

[32] Han S., Pool J., Tran J, etc. “Learning both weights and connections for efficient neural networks,” in Advances in Neural Information Processing Systems (NIPS), 2015, pp. 1135–1143.

[33] B. Jacob, S. Kligys, B. Chen, etc. “Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 2704–2713.

[34] X. Zhang, Z. Luo, J. Ma, etc. “Self-Attention Distillation: Towards Efficient Deep Network Solutions,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CJ. Aguinaldo, C. Dhawan, F. Garcia, etc., “Compressing Generative Models via Factorized Adaptation,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6974–6983.VPR), 2021, pp. 13077–13086.

[35] Wang Q., Teng Z., Xing J, etc. (2021). LightTrack: Finding Lightweight Neural Networks for Object Tracking via One-Shot Architecture Search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15180-15189.

[36] X. Guo, K. Yang, W. Zhang, etc. “Single Path One-Shot Neural Architecture Search with Uniform Sampling,” in Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 544–560.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, and Piotr Dollár. "Microsoft COCO: Common Objects in Context." European Conference on Computer Vision (ECCV), 2014.

[38] Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2012). "The PASCAL Visual Object Classes (VOC) Challenge." In *International Journal of Computer Vision (IJCV), 88*(2), 303–338.

[39] Kristan, M., Matas, J., Kaniavsk, I., Vojír, T., & Zhang, L. (2018). "The Visual Object Tracking VOT2018 Challenge Results." Proceedings of the European Conference on Computer Vision (ECCV) Workshops.

[40] Fan, Y., Zhang, S., Xu, F., & Yan, J. (2019). "LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019*, 5370-5379.

[41] Kramer, M., & Wörgo, M. (2017). DTB70: A Benchmark for Drone Tracking. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5503-5509.

[42] M. Mueller, N. Smith, and B. Ghanem, “A benchmark and simulator for UAV tracking,” in Proceedings of the European Conference on Computer Vision (ECCV), 2016, pp. 445–461.

[43] T.-Y. Lin, P. Dollár, R. Girshick, etc. “Feature pyramid networks for object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2117–2125.

[44] 刘通,方璐,高洪皓.边缘计算中任务卸载研究综述[J].计算机科学,2021,48(01):11-15.

[45] 陈科圻,朱志亮,邓小明,等.多尺度目标检测的深度学习研究综述[J].软件学报,2021,32(04):1201-1227.

[46] Wang Q., Xu L., & Wei, Y. (2020). ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11534-11542.

[47] Woo S., Park J., Lee J. Y., etc. (2018). CBAM: Convolutional Block Attention Module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 3-19.

[48] S. Ren, K. He, R. Girshick, etc., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 91–99.

[49] W. Liu, D. Anguelov, D. Erhan, etc., "SSD: Single Shot MultiBox Detector," in European Conference on Computer Vision (ECCV), 2016, pp. 21–37.

[50] T.-Y. Lin, P. Goyal, R. Girshick, etc., "Focal Loss for Dense Object Detection," in IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.

[51] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into High Quality Object Detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 6154–6162.

[52] N. Carion, F. Massa, G. Synnaeve, etc., "End-to-End Object Detection with Transformers," in European Conference on Computer Vision (ECCV), 2020, pp. 213–229.

[53] X. Zhu, W. Su, L. Lu, etc., "Deformable DETR: Deformable Transformers for End-to-End Object Detection," in International Conference on Learning Representations (ICLR), 2021.

[54] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," arXiv preprint arXiv:2207.02696, 2022.

[55] Bertinetto L., Valmadre J., Henriques, J. F., etc. (2016). Fully-Convolutional Siamese Networks for Object Tracking. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 850-865.

[56] Guo Q., Feng W., Zhou C., etc. (2017). Learning Dynamic Siamese Network for Visual Object Tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1781-1789.

[57] He, A., Luo, C., Tian, X., etc. (2018). A Twofold Siamese Network for Real-Time Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4834-4843.

[58] Danelljan M., Robinson A., Khan F. S., etc. (2016). Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), 472-488.

[59] Yang T., Chan A. B., & King I. (2017). Learning to Compare: Siamese Network with Gradient Reversal for Robust Object Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1383-1392.

[60] Liu T., Wang G., & Yang, Q. (2019). Real-Time Visual Tracking via Siamese Weighted ResNet. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 287-295.

[61] Cui Y., Zhang X., Cao Y., etc. (2020). Improved Siamese Attention Pyramidal Network for Real-Time Object Tracking. In arXiv preprint arXiv:2003.12478.

[62] Danelljan M., Bhat G., Khan F. S., etc. (2017). ECO: Efficient Convolution Operators for Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6638-6646.

[63] Yan T., Zhang X., Wang Y., etc. (2021). LightTrack: Finding Lightweight Neural Networks for Object Tracking via One-Shot Architecture Search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15180-15189.

[64] Zhang Z., Peng H., Fu J., etc. (2020). Ocean: Object-aware Anchor-free Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), 771-787.

[65] Li B., Wu W., Wang Q., etc. (2018). High Performance Visual Tracking with Siamese Region Proposal Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8971-8980.

[66] Zhang Z., Peng H., & Fu J. (2019). Deeper and Wider Siamese Networks for Real-Time Visual Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4591-4600.

[67] Guo D., Wang J., Cui Y., etc. (2021). Graph Attention Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 5277-5286.

[68] Wang Q., Zhang L., Bertinetto L., etc. (2019). Fast Online Object Tracking and Segmentation: A Unifying Approach. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1328-1338.

[69] Li B., Yan J., Wu W., etc. (2018). High Performance Visual Tracking with Siamese Region Proposal Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8971-8980.

[70] Hu J., Shen L., & Sun G. (2018). Squeeze-and-Excitation Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7132-7141.

[71] LIU Z,LI J,SHEN Z, etc. Learning Efficient Convolutional Networks through Network Slimming [C]//Proceedings of the IEEE International Conference on Computer Vision.Venice:IEEE,2017:2755-2763.

[72] LUO J,ZHANG H,ZHOU H, etc. ThiNet: Pruning CNN Filters for a Thinner Net[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,41(10):2525-2538.

[73] AHMED W, ANSARI S,HANIF M, etc. PCA driven mixed filter pruning for efficient convNets[J/OL].PLoS One,2022:1-17 (2022-1-24)[2025-2-17].

[74] YAN L,GU S, SHEN C, etc. Soft independence guided filter pruning[J/OL].Pattern Recognition,2024:2423-2475(2024-9-1)[2025-2-17].

[75] WANG L,CHAN R,ZENG T, etc. Probabilistic Semi-Supervised Learning via Sparse Graph Structure Learning[J].In IEEE Transactions on Neural Networks and Learning Systems,2021,32(2):853-867.

[76] PAN Z, CAI J,ZHUANG B.Stitchable Neural Networks(C)//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition,Vancouver:IEEE,2023:16102–16112.

[77] WANG Z,CHEN J,ZHOU W, etc. SmartTrim: Adaptive Tokens and Parameters Pruning for Efficient Vision-Language Models[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation.Torino:ELRA and ICCL,2024:14937–14953

[78] 周星源.基于高速视觉的非接触测量方法研究与应用[D].安徽：安徽工程大学,2024.

[79] 赵俊红.篡改图像边缘模糊操作的被动取证算法[J].华南理工大学学报(自然科学版),2011,39(07):77-82.

[80] 陈卫,邓志良.基于颜色和形状的不均匀光照条件下球体识别[J].电子设计工程,2019,27(21):177-181.DOI:10.14022/j.cnki.dzsjgc.2019.21.039.

攻读学位期间发表的学术成果

论文发表情况：

[1]**J. Bao**, J. Zhao, B. Liu and D. Zhang, "Multi-Attention Fusion Lightweight Single Object Tracking Algorithm," 2024 7th International Conference on Robotics, Control and Automation Engineering (RCAE), Wuhu, China, 2024, pp. 654-658,

[2]**包俊涛**,吴岳,刘丙友.基于场景复杂度的轻量化自适应剪枝模型[J].湖北民族大学学报(自然科学版),2025,43(01):47-52.DOI:10.13501/j.cnki.42-1908/n.2025.03.016.

专利发表情况：

[1]刘丙友、**包俊涛**、齐晶晶、朱标、查放、孔科研、赵继兴，一种基于边缘计算的干线绿波评测系统及方法，CN202410040676

[2]刘丙友、**包俊涛**、齐晶晶、张陆贤、杨潘、赵继兴，一种基于 KCF 算法的干线绿波控制方法及系统，CN202310619873

致谢

在我完成这篇论文的过程中，得到了许多人的帮助与支持。在此，我向所有在学术研究、生活和情感上给予我帮助的人们表达最诚挚的感谢。

首先，我要衷心感谢我的导师刘丙友教授。导师在整个研究过程中给予了我细致的指导与耐心的支持。无论是在选题、研究思路的确立，还是在论文的撰写过程中，刘教授都提供了宝贵的意见和建议，帮助我克服了一个又一个的难题。每当我遇到困惑时，刘教授总是耐心指导，启发我独立思考，培养了我严谨的科研态度和创新精神。刘教授不仅在学术上给予我极大的帮助，也在为人处事方面树立了榜样，使我受益终生。

我还要感谢我的学长学姐们：罗建、朱国武、吴贞犇、陈海峰、王超、俞苏苏、李子阳、王义琼、周星源、代浩文、章东旭、陈邱卓、褚冬亭、程清扬、董家辉、尚永健、徐莎曼。在研究和实验过程中，学长和学姐们都给了我无私的帮助和支持。无论是在数据处理、文献分析，还是在实验方案的讨论中，他们都耐心地指导我，帮助我快速提高自己的科研能力。特别感谢周星源学长，在我遇到实验瓶颈时，他提供了很多宝贵的经验和建议，让我能够顺利地解决问题并推进研究。

在我的同届同学中，林腾飞、蕾圣学、缪陆、刘俊楠、刘飞瑶、李彤彤给了我很多帮助，我们共同讨论问题，分享研究心得，促进了我学术思维的成长。感谢你们的友情和支持，是你们让我的学术之路更加有趣。

此外，我还要感谢我的学弟们：蒋述、陈旺旺、陶晓军、吴纪强、赵宇涵、刘祥玉、黄志成、汪冉、尹冰、黄鑫、李田田、姚浩琳、肖滋乐、袁俊杰、刘康伟。你们在我研究过程中给予了我很多帮助，也让我感受到了团队合作的力量。你们的努力和进取激励了我不断追求更高的学术目标。

同时，我要感谢我的家人，尤其是我的父母。感谢他们始终如一地支持我，给我提供了良好的学习环境和生活条件。在我遇到困难时，父母总是第一时间给予我关心和鼓励，让我能够无后顾之忧地专注于学术研究。他们的无私爱与支持是我能够走到今天的坚强后盾。

最后，我要特别感谢我的女朋友丁萌萌。她在我研究过程中给予了无私的支持与鼓励。无论是在我学术压力大的时候，还是遇到困难的时刻，萌萌总是默默地在背后支持我，给我安慰与动力。在我需要休息时，她陪伴我走出低谷，带给我温暖和力量。

在这篇论文的写作过程中，每一位给予我帮助和支持的人都深深地烙印在我的心中。你们的帮助和鼓励让我在学术的道路上走得更加坚定，也让我深刻感受到团队合作、朋友和家人支持的力量。我将永远铭记并感激你们的帮助与陪伴，未来的日子里，我会更加努力，不辜负所有关心和支持我的人。

再次感谢所有帮助过我的人，正是因为你们的支持，我才能够顺利完成这篇论文。

尚德教学　唯实惟新