

分类号: TP391.4

学校代码: 10363

密 级: 公开

学 号: 2210330103



安徽工程大学

Anhui Polytechnic University

硕士学位论文

题 目 基于显著性注意力的
目标跟踪方法研究

论 文 作 者 尚永健

校 内 导 师 王力超 副教授

校 外 导 师 郝刚

专业学位类别 电子信息

研究方向 (领域) 检测技术与自动化装置

论文提交日期: 2024 年 5 月 31 日

类号：TP391.4

单位代码：10363

密类级：公开

学号：2210330103

题 目 基于显著性注意力的方法研究

题 目 RESEARCH ON OBJECT TRACKING METHOD
BASED ON SALIENT ATTENTION

学 生 姓 名： 尚永健

校 内 导 师： 王力超

校 外 导 师： 郝刚

专业学位类别： 电子信息

研究方向(领域)： 检测技术与自动化装置

论文答辩日期： 2024 年 5 月 22 日

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：尚永佳

日期：2024年 5月 31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：尚永佳 指导教师签名：王加

日期：2024 年 5 月 31 日

日期：2024 年 5 月 31 日

基于显著性注意力的目标跟踪方法研究

摘要

目标跟踪是计算机视觉中一项至关重要的技术，包含虚拟现实在内的多个尖端应用场景，如智能交通疏导和无人机跟踪等，在各方面都发挥着重要的作用。随着机器学习技术的深度开发，目标跟踪技术取得了显著跃升，但在应对诸如目标物体被遮挡、背景环境复杂多变以及目标短暂脱离视野等实际问题时，实现对复杂场景高效且稳定的目标跟踪依然存在一定的局限性。因此，本论文专注于研究解决这些场景下的跟踪问题，以此为前提优化孪生网络架构的目标跟踪算法。为了增强其鲁棒性和精度定位能力，本文的核心研究内容列举如下：

(1) 针对现有的孪生网络在处理目标形变、平面旋转、遮挡等复杂场景中精度和鲁棒性不足的问题，本论文基于 SiamCAR 算法，设计了一种新颖的跟踪策略 MSSiamCAR，使跟踪器在跟踪过程中，更多地关注目标自身情况，提出局部跟踪策略，跟踪目标的显著区域，这些区域对跟踪器具有明显的判别性。其中显著性捕获可以获得目标特征中对跟踪任务具有判别性的显著点，显著性交互利用图注意力在显著性区域之间建立有效联系，通过学习目标与搜索区域间的关联性来适应目标不同的尺度变化。实验结果表明，相比于 SiamCAR 算法，MSSiamCAR 在 VOT2018 数据集上期望平均重叠率增长了 2.5%，在 GOT-10k 数据集的 AO 、 $SR_{0.5}$ 和 $SR_{0.75}$ 这三个指标中，分别提高了 4.4%、4.3%和 3.1%，跟踪性能优秀，在复杂场景中具有良好的鲁棒性。

(2) 为了进一步提升跟踪器在处理无人机视角下的相似小目标

形变、遮挡等场景的跟踪性能，本论文设计了一种基于局部显著性 Transformer 的无人机目标跟踪方法。设计了一个显著性细化精炼模块，基于对前景和背景的不同关注，并不是所有的 patch 都对跟踪任务具有相同的信息价值。根据不同 patch 得分的高低，区分显著性信息对跟踪任务的贡献。设计了一种新的局部显著性 Transformer 结构，将细化精炼的显著性信息，通过多层次地融合，指导动态词元，在标准的 Transformer 结构上进行了轻量化，通过交叉注意机制来细化显著性信息并最大限度地减少外观信息的损失。提出的跟踪器在四个权威无人机跟踪基准(DTB70、UAV123、UAV10fps、UAV20L)上进行了综合评估，实验结果显示，提出的跟踪器在无人机视角下的挑战中表现出卓越的性能，能够妥善应对并有效解决多种复杂情况下的目标跟踪难题。

(3) 为了在真实世界中对处于复杂环境的目标进行实时跟踪，在论文中搭建了无人机实验平台，将轻量化后的目标跟踪算法部署在搭建的硬件平台中，与同级别轻量化跟踪网络相比，网络模型参数量更少，经过同等硬件条件下的大量对比实验发现，在真实世界跟踪中，本论文提出的跟踪器在速度和精度方面有了显著的提升。由 DTB70 数据集的评估结果可知，本论文提出的跟踪器速度达到了 29 帧，与轻量化的 SaimFC++相比，虽然成功率都达到了 0.6 以上，但是速度却快了一倍。与轻量化的 ECO_HC 相比，虽然速度慢了几帧，但是也达到了实时跟踪的条件并且成功率上升了 1.57%。本论文提出的跟踪器在真实世界跟踪中展现了卓越的性能和极具竞争力的运行速度。

关键词：目标跟踪，显著性，Transformer，注意力机制，孪生网络

RESEARCH ON OBJECT TRACKING METHOD BASED ON SALIENT ATTENTION

ABSTRACT

Object tracking is a crucial technology in computer vision and plays an indispensable role in many cutting-edge application scenarios including virtual reality, such as intelligent traffic management and drone navigation. Although object tracking technology has made significant leaps with the in-depth development of machine learning technology, it is still efficient and stable in complex scenes when dealing with practical problems such as target objects being occluded, complex and changeable background environments, and targets temporarily out of sight. Object tracking still has certain limitations. Therefore, in this article, we focus on studying and solving the tracking problem in these scenarios, and use this as a basis to optimize the object tracking algorithm of the Siamese network architecture to improve its robustness and positioning accuracy. Our main research contents are as follows:

(1) In order to solve the problem of insufficient accuracy and robustness of existing twin networks in dealing with target deformation, plane rotation, occlusion and other complex scenes, we designed a novel tracking strategy MSSiamCAR based on the SiamCAR algorithm so that the tracker can during the tracking process, more attention is paid to the target itself, and local tracking strategies are proposed to track local salient areas of the target. These areas are obviously discriminative for tracking. Saliency mining can find the most eye-catching feature areas in moving targets, allowing the tracker to quickly identify the target.

Inspired by graph attention, we innovatively designed a method whose core goal is to strengthen the deep correlation between salient regional features within moving targets to adapt to and handle diverse and complex transformations of targets. Experimental results show that compared with the SiamCAR algorithm, the expected average overlap rate of MSSiamCAR on the VOT2018 dataset increased by 2.5%, and in the three indicators AO , $SR_{0.5}$ and $SR_{0.75}$ of the GOT-10k dataset, it increased by 4.4%, 4.3% and 3.1%, with excellent tracking performance and good robustness in complex scenes.

(2) In order to further improve the tracking performance of the siamese network in handling similar small target deformations, occlusions and other scenarios from the drone's perspective, we designed a drone object tracking method based on local saliency transformer. A saliency refinement and refinement module is designed, based on different attention to the foreground and background, and not all patches have the same information value for the tracking task. According to the scores in different patches, the contribution of salient information to the tracking task is distinguished. A new local saliency Transformer structure is designed. Based on the refined and refined saliency information, through multi-level integration and guidance of dynamic word elements, the standard transformer structure is lightweight and achieves a novel local saliency transformer that refines saliency information and minimizes the loss of appearance information through a cross-attention mechanism. Our tracker was comprehensively evaluated on four authoritative UAV tracking benchmarks (DTB70, UAV123, UAV10fps, UAV20L). The experimental results show that our tracker performs well in dealing with the problem of target appearance changes from the UAV perspective. Excellent performance, able to effectively adapt to and solve object tracking challenges in a variety of complex scenarios.

(3) In order to track targets in complex environments in real time in the real world, a UAV experimental platform was built in the paper, and the lightweight object tracking algorithm was deployed on the built hardware platform to work with the same level of lightweight tracking network. Compared with the network model, the number of parameters is smaller. After a large number of comparative experiments under the same hardware conditions, it is found that in real-world tracking, the tracker proposed in this article has significantly improved in speed and accuracy. From the evaluation results of the DTB70 data set, it can be seen that the tracker proposed in this article reaches a speed of 29 frames. Compared with the lightweight SaimFC++, the success rate is more than 0.6, but the speed is twice as fast. Compared with the lightweight ECO_HC, although the speed is a bit slower, it still meets the conditions for real-time tracking and the success rate increases by 1.57%. The tracker proposed in this paper demonstrates excellent performance and competitive running speed in real-world tracking.

KEY WORDS: object tracking, saliency, transformer, attention mechanism, siamese network

目录

第 1 章 绪论.....	1
1.1 研究背景及意义.....	1
1.2 国内外研究现状.....	2
1.2.1 基于相关滤波的目标跟踪算法.....	2
1.2.2 基于深度学习的目标跟踪方法.....	3
1.3 目标跟踪面临的挑战.....	4
1.4 本文研究内容与结构安排.....	5
1.4.1 主要研究内容.....	5
1.4.2 本文章节安排.....	6
第 2 章 目标跟踪理论基础.....	8
2.1 引言.....	8
2.2 孪生网络目标跟踪算法.....	8
2.2.1 孪生网络结构.....	8
2.2.2 经典孪生网络 SiamFC.....	9
2.2.3 经典孪生网络 SiamRPN.....	9
2.2.4 无锚框孪生网络 SiamCAR.....	11
2.3 注意力机制.....	12
2.4 数据集及评价指标.....	14
2.4.1 目标跟踪数据集.....	15
2.4.2 相关评价指标.....	15
2.5 本章小结.....	16
第 3 章 基于特征融合和显著性图注意力的目标跟踪方法.....	17
3.1 引言.....	17
3.2 图注意力基本思想.....	18
3.3 特征融合和显著性图注意力网络结构.....	19
3.3.1 特征融合模块.....	19
3.3.2 显著性图注意力模块.....	20
3.4 实验结果分析.....	23
3.4.1 实验细节.....	23
3.4.2 定量分析.....	23
3.4.3 定性分析.....	30
3.4.4 算法实时性分析.....	31
3.4.5 消融实验.....	32
3.5 本章小结.....	33
第 4 章 基于局部显著性 Transformer 的无人机目标跟踪方法.....	34
4.1 引言.....	34
4.2 Transformer 基本思想.....	35
4.3 局部显著性 Transformer 的网络结构.....	37

4.3.1 显著性挖掘模块.....	38
4.3.2 显著性细化精炼模块.....	40
4.3.3 局部显著性 Transformer 模块	42
4.4 实验结果分析.....	46
4.4.1 实验设备及参数设置.....	46
4.4.2 定量分析.....	46
4.4.3 定性分析.....	55
4.4.4 消融实验.....	57
4.5 本章小结.....	58
第 5 章 无人机硬件平台实时跟踪验证	59
5.1 引言	59
5.2 实验平台搭建.....	59
5.3 真实世界实时跟踪分析.....	61
5.4 本章小结.....	64
第 6 章 总结与展望.....	65
6.1 总结.....	65
6.2 展望.....	66
参考文献.....	67
攻读学位期间发表的学术论文目录.....	75
致谢	76

第 1 章 绪论

1.1 研究背景及意义

进入新世纪以来，科技日新月异的快速发展催生了各类视频监控装置的普及，极大地便利了人们的日常生活，同时也积累了前所未有的大规模视频图像数据资源。如此庞大的数据量显然无法单纯依赖人力完成有效处理，于是，计算机视觉技术便在这样的背景下应市场需求而崛起。随着硬件设施性能的不不断提升以及深度学习技术在实践中的成功应用，计算机视觉已经深深植根于现代社会生活之中。诸如目标检测^[1]、目标跟踪^[2]和目标分割^[3]等计算机视觉技术，在现实生活中的多元场景中实现了广泛应用和深度融合。

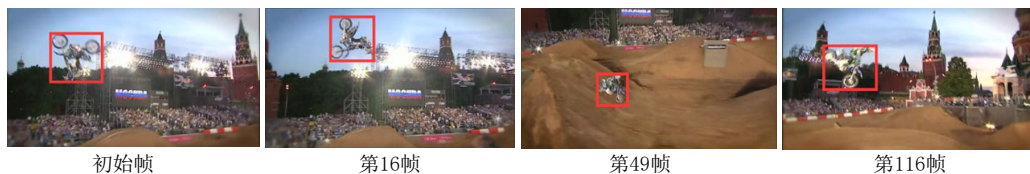


图 1-1 目标跟踪任务示意图

深入研究目标跟踪技术对于计算机视觉领域^[4]来说至关重要，这一基础任务的表现形式可见于图 1-1 所示的目标跟踪示意流程。在视频的首帧选定某一目标后，相应的算法能够在后续连续帧中自主预测该目标的位置和大小变化，进而连续跟踪目标的状态，为后续更深层次的视觉分析做基础。

目标跟踪在视频图像分析中扮演着至关重要的角色，广泛渗透于日常生活的诸多方面。例如，在交通管理中，可用于实时统计车流量、监测道路车辆运行状态；在安全防范系统中，能够进行行人轨迹跟踪^[5]和风险预警；在自动驾驶技术^[6]中，有助于车辆自动保持距离跟随；在智能机器人^[7]研发中，服务于人机互动、人体动作识别乃至移动机器人导航功能；而在军事应用^[8]中，诸如无人机的自主目标锁定与导航引导等，亦离不开目标跟踪技术的支持。因为目标跟踪技术在人工智能各领域具有广泛应用价值，所以其成为当前学术研究的热门话题之一。然而，尽管学者们针对不同情境设计了各式各样的目标跟踪算法，

由于目标运动形态的复杂多样性以及环境条件的千变万化，现有的一些先进的目标跟踪方法往往很难确保在复杂场景下跟踪预测的准确性。因此，如何改进现有目标跟踪算法存在的难题，对推动整个目标跟踪技术的突破和发展具有重要意义。

1.2 国内外研究现状

目标跟踪技术的核心是对一系列图像中目标的动态属性例如位置、尺寸等轨迹要素进行深入剖析，并通过对这些属性参数的运算处理，以展示与目标行为密切相关的联系。生成式跟踪模型聚焦于目标本身，在跟踪过程中并不直接考虑背景干扰，而是基于目标与潜在匹配对象之间的相似性度量进行目标定位，诸如粒子滤波、稀疏表达及均值漂移等经典算法均属此类。相反，判别跟踪模型则重视背景环境对目标识别的影响，通过从初始帧确定的目标区域出发训练高精度的分类模型，旨在从背景纹理中准确分离出目标实体，将跟踪任务当成一个二分类的任务。随着深度学习技术的持续革新与发展，现今的跟踪算法逐渐模糊了生成与判别两种范式的严格界限，越来越多的混合型方法涌现出来。在这个快速演进的过程中，跟踪算法始终保持迭代升级的步伐，涌现出众多创新实现。本文系统梳理并详尽介绍了目标跟踪研究领域的研究现状。

1.2.1 基于相关滤波的目标跟踪算法

该类跟踪模型的关键流程在于运用模板匹配原理，首先将待跟踪目标初始化为滤波模板，接着将此模板与搜索范围内的各个候选区域执行相关性计算，由此输出一张反映候选区域内目标响应程度的图像，其中响应强度最高的位置即指示了当前帧中目标的实际位置。具体而言，操作始于从初始图像序列的第一帧中截取目标区域作为输入特征，继而通过特征提取技术描述目标的本质特征，采用余弦窗口函数以压制目标周边的背景干扰信息。接下来，模型会在已学习的滤波器参数与当前输入特征之间建立频域关联，借助逆傅里叶变换计算得出空间坐标系下的响应分布图，得分数值最大的区域则被认为是目标所在区域。最终，在预测位置提取目标的外观特征，用以进一步的模型训练与更新，为下一帧的目标跟踪做足准备。

若干经典的判别式相关滤波目标跟踪算法包括但不限于 CSK^[9]、KCF^[10]、

DSST^[11]、C-COT^[12]以及 ECO^[13]等。KCF 算法的独特之处在于其利用多通道方向梯度直方图特征，并通过循环平移构造正负样本；相较于 CSK，KCF 不仅实现了高效运行速度，在保持稳定跟踪性能方面表现突出。DSST 算法采用了一种结合三维尺度空间相关滤波器的协同跟踪策略，从而实时估算出最佳匹配尺寸。C-COT 算法由 M.Danelljan 在 2016 年提出，其首先借助预先离线训练好的 VGGNet^[14]抽取原始图像的深层特征，然后在特征图上运用立方插值，并通过海森矩阵确定目标的具体位置。而 ECO^[13]改良了 C-COT 的模型更新机制，利用卷积因子分解特性修整网络结构，并采用隔帧更新策略以减小模型更新频率，同时采用样本分组策略剔除相邻帧中的冗余样本，这不仅提升了跟踪精度，还极大地加快了跟踪速度。

1.2.2 基于深度学习的目标跟踪方法

深度学习技术在目标跟踪领域的推进与深度学习整体进步及大规模数据集的兴起息息相关。2012 年 ImageNet^[15]竞赛的成功举办促使学术界广泛重视并深入探究深度学习技术。随着各类与跟踪任务相关的自然场景数据集不断涌现，许多研究者纷纷尝试将深度学习应用到目标跟踪问题中。得益于当今丰富的数据资源，目标跟踪算法能够通过神经网络训练习得富含深层次语义信息的特征表达，从而增强识别与跟踪能力。尽管当前基于深度学习的跟踪算法在网络框架的复杂性和数据集的利用上取得了显著优势，但在实际应用中仍面临诸如光照变化、长期遮挡、运动模糊、分辨率低下等多种棘手问题，因此，目标跟踪领域的研究者们正在不断优化和完善跟踪算法以应对各种复杂的现实环境挑战。

Tao 等研究者 2016 年在目标跟踪领域引入了创新性的孪生神经网络，并据此提出了一种全新的“孪生网络驱动的目标搜索与跟踪”策略。机智地将跟踪问题与图像配对问题相结合，利用 Siamese^[16]网络学习匹配函数，通过对每帧图像与首帧目标进行相似性比对，以最高得分的候选框作为跟踪结果，这一创新思路为后续目标跟踪提供了新的研究路径。同年，Bertinetto 团队创新提出了名为 SiamFC^[17]的全卷积孪生网络跟踪算法，该算法机智地将复杂的跟踪任务简化为模板匹配的问题。由于 SiamFC 在实现高效成功率和稳健跟踪性能方面的优秀表现，引发了全球研究者的热烈关注。在此基础上，众多改进型跟踪算法应运而生，如 SiamRPN++^[18]和 SiamRPN^[19]等，它们通过在孪生网络后端增设

区域提议网络(RPN)，以提升跟踪精度。SiamRPN 是在 SiamFC 架构中进一步融入了 Faster RCNN^[20]的区域提议网络(RPN)单元，从而增强了对目标位置的精确估计能力，但受限于相对浅层的网络结构，尚无法有效提取深层次特征，针对上述问题，SiamRPN++则率先对 SiamRPN 中采用的 AlexNet^[21]结构进行了升级，将其替换为更深层次的 ResNet^[22]网络，是首次在孪生网络中运用 ResNet-50 网络进行特征提取，明显地改善了模型预测的准确性。SiamDW^[23]则专注于研究特征提取网络的深度与宽度，对 ResNet 和 Inception^[24]等网络架构进行调整，展现出卓越的性能。Guo 等人设计的 SiamCAR^[25]，则独辟蹊径地采取了无锚框的目标跟踪方式，减少了对回归分支锚框超参数的依赖，并加入中心点分支辅助分类，确保精准回归到最佳目标位置。步入 2021 年，TransT^[26]和 Stark^[27]两项研究工作将 Transformer^[28]架构引入目标跟踪，取得了当时最先进的成果。Transformer 架构的优势在于能够有效地捕捉并构建长距离依赖关系，可以并行计算矩阵向量，方便快捷。在之后几年里，跟踪领域出现了基于 Transformer 的结构新潮，比如 OTrack^[29]和 ViPT^[30]等算法。Transformer 在目标跟踪领域的应用相较于传统算法展现出了显著优势：首先，它突破了局部视野限制，能够全面考虑图像整体信息，利于模型理解目标在复杂背景下的全局位置与状态；其次，Transformer 凭借其内在机制能够依据实时的上下文情境动态调整和更新目标模板，有效应对目标因运动、变形等因素引起的外观变化挑战；此外，这种架构还拥有强大的多模态融合能力，可轻松集成诸如光流、深度信息、语义属性等多种类型的数据源，借此增强跟踪结果的鲁棒性。

1.3 目标跟踪面临的挑战

跟踪运动中的目标是一项极为难度的任务，原因在于目标本身的动态变化及其所处环境的复杂性。正如图 1-2 所示，多种典型挑战存在于目标跟踪领域之中：

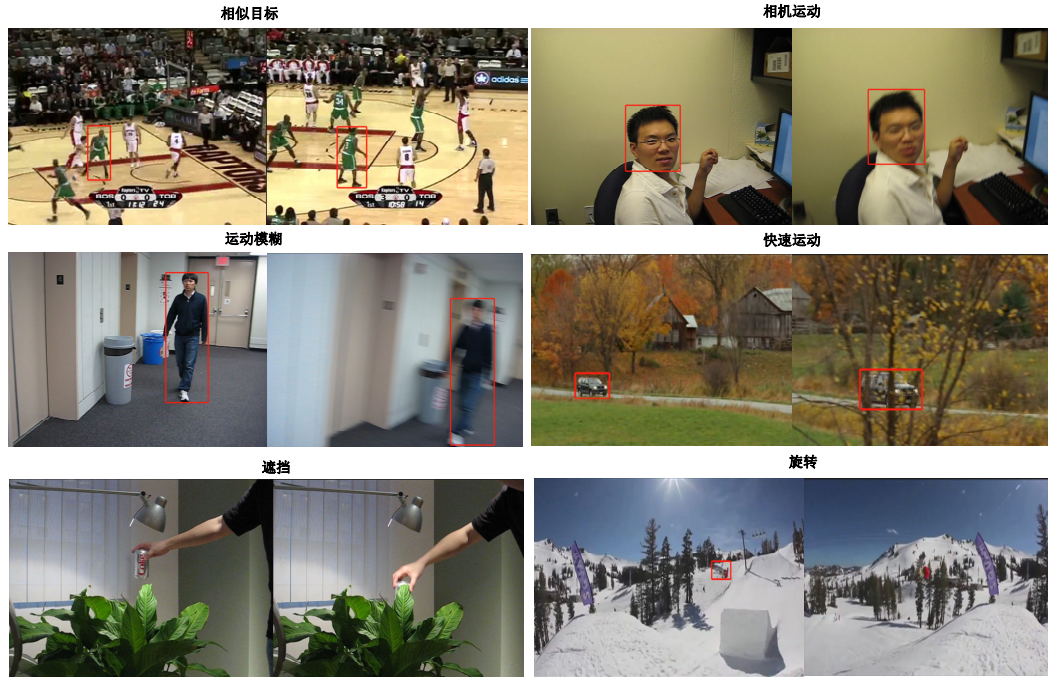


图 1-2 目标跟踪场景

首先，遮挡是此领域尤为突出的挑战，它是导致跟踪失误和失效的关键难点；其次，背景混淆情况下的目标跟踪颇具挑战，即当目标周围存在高度相似的背景元素时，会对目标的识别产生严重干扰；另外，尺度变化也是一个不容忽视的因素，目标因运动距离变化而在视觉空间中呈现出不同大小，如何精确预测并调整目标框的尺寸成为了跟踪中的另一项重要挑战，这对于跟踪器的预测精确度有直接影响。综上所述，无论是目标自身的属性变化还是外部环境条件的影响，都使得精确捕捉运动目标变得极为复杂。除此之外，还有其他一些关键因素亦对跟踪效果产生显著影响，如光照条件的改变、低分辨率的图像、摄像机的移动造成的视角变换、运动引起的图像模糊、目标物体的旋转变换等。同时，考虑到目标跟踪是应用于实时视频分析的场景，因此，设计出既能保证高精度又能满足实时响应速度要求的目标跟踪算法至关重要。

1.4 本文研究内容与结构安排

1.4.1 主要研究内容

针对孪生网络目标跟踪算法在宽高比变化、部分遮挡、相似小目标、快速运动等挑战中跟踪效果不理想甚至跟踪失败的情况，本文以孪生网络为基础构建目标跟踪算法，并在此框架上深入探究如何优化算法以增强其跟踪准确度与

鲁棒性，主要研究步骤如下所述：

(1) 提出一种基于显著性图注意力的孪生网络跟踪器，将跟踪目标的全局信息转换成目标的局部显著性信息。设计了显著性捕获模块来获取目标局部显著性，这些显著性区域对干扰因素具有一定的鲁棒性，设计了显著性交互模块，利用图注意力机制在显著性区域之间建立有效联系，将目标特征信息传播到搜索特征上，使跟踪器对目标特征更加具有判别能力，相比于其他跟踪方法，该方法在精度和成功率方面展现出了显著的进步。

(2) 提出一种使用局部显著性 Transformer 架构的无人机跟踪器。充分利用显著性挖掘和细化模块，将所提取的多层特征进行动态的局部嵌入计算，使跟踪器能够实现卓越的前景与背景区分能力。该方法因为充分利用网络的不同特征，形成了基于局部到全局的注意力机制，因此对于小目标的跟踪能力较强，提高了鲁棒性。

(3) 将局部显著性 Transformer 算法部署到无人机平台，Transformer 架构的一大亮点在于其并行计算能力，针对真实世界的场景中多种无人机挑战，与其他先进跟踪器相比，我们跟踪器的速度、成功率和精度在同一 UAV 硬件平台上有了显著的提高。

1.4.2 本文章节安排

本文的篇章结构划分为六个章节，接下来将逐一阐述各章节的具体组织内容：

第 1 章：阐述了本研究课题的背景及意义，对国内外目标跟踪算法的研究状况进行了综述，并指出了当前目标跟踪领域所面临的若干难点，最后明确了本文研究的主要创新点与贡献。

第 2 章：主要是目标跟踪的基础理论知识。沿时间脉络梳理了基于孪生网络的一系列经典跟踪算法的发展历程；随后对注意力机制的相关概念和技术进行了深入解读。

第 3 章：介绍了一种使用特征融合和显著性图注意力的跟踪器。本章详细介绍了该跟踪算法的结构以及对比实验和消融实验。

第 4 章：介绍了一种使用局部显著性 Transformer 的跟踪器。本章详细介绍了该跟踪算法的结构以及对比实验和消融实验。

第 5 章：介绍了在真实世界中使用第四章提出的跟踪方法，在硬件设备上部署，对运动目标进行实时跟踪，与对比算法进行了速度与准确率的比较。

第 6 章：对本文已完成的各项研究工作进行了全面总结，并在此基础上对后续研究方向及存在的问题进行了前瞻性的展望。

第 2 章 目标跟踪理论基础

2.1 引言

目标跟踪是计算机视觉中一项至关重要的技术，其进步与革新对多元应用场景具有重大价值。本章节系统阐述了几种具有代表性的孪生网络目标跟踪方法，还概览了五个常用的目标跟踪数据集及相应的评估标准。

2.2 孪生网络目标跟踪算法

2.2.1 孪生网络结构

孪生神经网络^[31]架构作为一种深度学习的应用方案，由两个镜像对称且权重完全一致的子网络构成，这两个分支共同分享相同的权重配置与层级结构设计，核心功能在于衡量并区分各自输入数据的相似性和差异性。在这一框架下，原始输入首先通过同一套卷积神经网络进行统一的特征抽取处理，随后运用诸如欧氏距离^[32]、曼哈顿距离^[33]或者余弦相似度等距离^[34]度量准则，比对二者提取后的特征向量间的匹配程度。参照图 2-1 所示，可以直观了解孪生网络的结构。该结构起源于人脸识别^[35]应用领域，并逐渐扩展至目标跟踪等多种应用场景。孪生网络具备以下显著特性：

1. 由于其参数在整个网络中得以共享，孪生网络能够在保证学习效果的同时，显著压缩模型规模，减少训练所需参数数量，有助于避免过拟合现象。
2. 具备良好的非线性关系处理能力，尤其适用于那些难以获取大规模标记数据的任务场景，展现出强大的适应性和有效性。

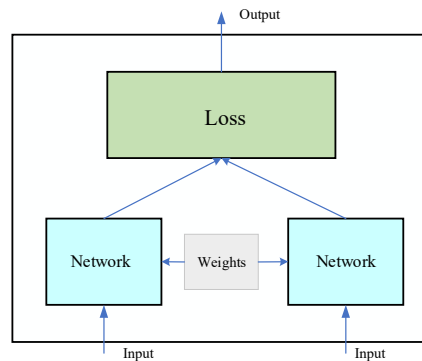


图 2-1 孪生网络结构图

2.2.2 经典孪生网络 SiamFC

SiamFC^[17]是首个将深度学习成功应用于目标跟踪任务的标志性方法，其网络架构如图 2-2 所示。在 SiamFC 中，核心操作是运用全卷积^[36]结构构建特征提取模块，该模块接收一对经过裁剪处理的图像作为输入。具体而言，模板图像源自根据真实边界框裁剪的初始帧，而搜索图像则来自以上一帧预测输出的边界框为基础进行裁剪的新帧。此后在其他目标跟踪算法中，对输入图像进行裁剪以生成模板和搜索图片的做法基本保持一致。SiamFC 的跟踪流程如下：首先，将模板和搜索图像同时送入到 AlexNet 特征提取网络中，根据孪生网络共享权重的特点，我们可以得到维度分别为 $6 \times 6 \times 128$ 和 $22 \times 22 \times 128$ 的模板和搜索特征。之后，模板特征和搜索特征会经过一个互相关操作，二者运算后会产生一个 $17 \times 17 \times 1$ 大小的响应图 heatmap。如图 2-2 所示，图中红色和蓝色区域直观展示了搜索图像中与模板图像相似度较高的位置。SiamFC 为了提升精度，在测试过程中采取了多尺度的方法，这一举措有效地增强了其跟踪性能。然而，SiamFC 的一个主要局限性在于不具备在线学习和自适应能力，无法依据实时获取的数据优化对目标的识别。这意味着虽然预训练模型在初始化阶段能对目标对象形成良好的适配，但对于目标在跟踪过程中可能出现的显著尺寸变化、姿态转变或光照条件变动，模型可能无法做出及时且恰当的响应，从而造成跟踪效率降低。此外，由于搜索区域大小固定及仅依赖单一模板匹配算法，SiamFC 在处理大规模长距离的目标移动时，其跟踪效果可能会大打折扣。

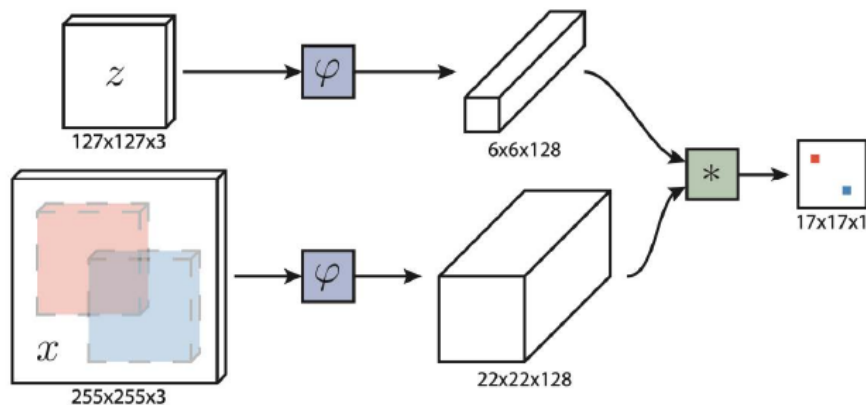


图 2-2 SiamFC 网络结构

2.2.3 经典孪生网络 SiamRPN

在 SiamFC 中采用多尺度策略来获取目标的预测边界框，但这一机制面临

两个挑战。一方面，多尺度状态估计会增加模型的计算负担，进而降低跟踪效率。另一方面，固定的尺度预测方式限制了对目标位置信息的精准捕捉，无法灵活适应目标尺寸的变化。鉴于这些不足，SiamRPN^[19]跟踪算法借鉴了目标检测领域的 Fast-RCNN 思想，将其核心技术 Region Proposal Network(RPN)整合到了目标跟踪框架中。SiamRPN 的具体网络架构在图 2-3 中有详尽展示，其对视频序列的跟踪步骤可概述为：首先将模板和搜索图像一同送入 CNN 以提取特征，所得到的特征维度分别是 $6 \times 6 \times 256$ 和 $22 \times 22 \times 256$ ，随后将这两组特征均馈送到 RPN 结构中。通过互相关操作，这两种特征共同生成了用于分类和回归的向量表达，其数学计算公式参照公式(2-1)给出。

$$\begin{aligned} A_{w \times h \times 2k}^{cls} &= [\varphi(x)]_{cls} \star [\varphi(z)]_{cls} \\ A_{w \times h \times 2k}^{reg} &= [\varphi(x)]_{reg} \star [\varphi(z)]_{reg} \end{aligned} \quad (2-1)$$

在训练阶段，SiamRPN 的总体损失函数是由即回归损失和分类损失^[37]组成的。关于分类损失部分，注意到分类分支仅负责区分目标(前景)与非目标区域(背景)这两种类别，故此，选用适用于二元分类任务的交叉熵损失函数作为分类损失的具体度量标准，该表达式如公式(2-2)：

$$L_{cls} = \frac{1}{N} \sum_i -[y_i \times \log(p_i) + (1 - y_i) \times \log(1 - p_i)] \quad (2-2)$$

y_i 是目标的标签信息，目标设置为 1，背景则设置成为 0， p 是预测为目标的概率。

回归损失如下式所示：

$$smooth_{L1}(x, \sigma) = \begin{cases} 0.5\sigma^2 x^2, & |x| < \frac{1}{\sigma^2} \\ |x| - \frac{1}{2\sigma^2}, & |x| \geq \frac{1}{\sigma^2} \end{cases} \quad (2-3)$$

x 是输入的参数， σ 是控制平滑范围的参数， δ 是坐标偏移量的真实值，最后的回归损失^[38]为：

$$L_{reg} = \sum_{i=0}^3 smooth_{L1}(\delta[i], \sigma) \quad (2-4)$$

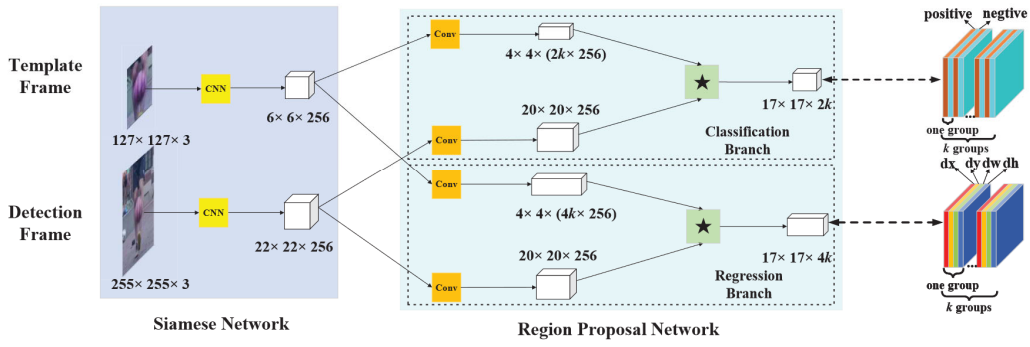


图 2-3 SiamRPN 结构图

2.2.4 无锚框孪生网络 SiamCAR

SiamRPN 跟踪算法在实现目标分类和回归任务时，依赖于 RPN 结构，但该结构中的 Anchor boxes^[39]参数设置繁琐且对超参数调优要求较高。对此，SiamCAR 跟踪算法选择了一种无 Anchor boxes 的途径，设计出一种更为简洁的网络结构，降低了对超参数精细化调节的需求，即便如此，该跟踪器仍然能达到极高的性能标准。SiamCAR 的网络架构如图 2-4 所示，网络结构主要由两大部分构成，即孪生子网络模块和分类回归网络模块。孪生子网络负责特征提取及相似度计算，而分类回归网络则执行具体的分类和回归任务。相较于 SiamRPN，SiamCAR 摒弃了 RPN 结构，转而在 CNN 部分采用更深层次的 ResNet-50 模型。特别是在多级特征利用上，SiamCAR 采取在后三层对应计算互相关后特征拼接的方式计算相似度得分。此外，SiamCAR 还在分类分支中新增了一个中心度分支^[40]，促使同一标签内的目标特征聚集在各自对应的标签类别的中心附近，同时确保不同类别的中心点最大程度分离，这一策略旨在提升对目标中心位置的精确定位能力。

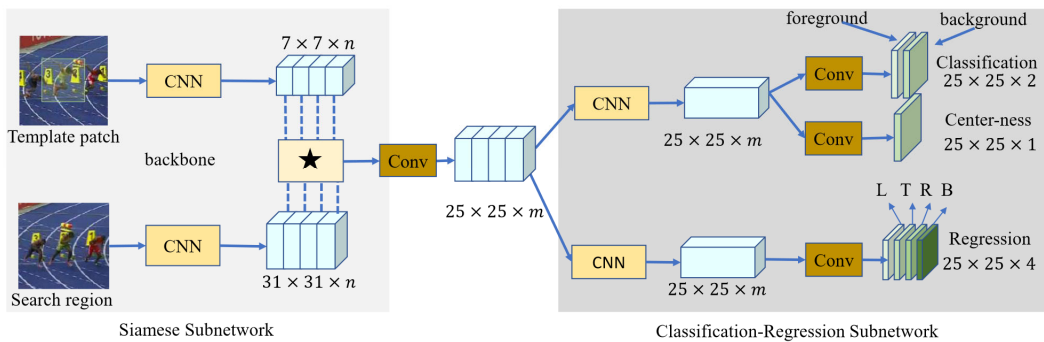


图 2-4 SiamCAR 网络结构

为了提升跟踪性能，通过精心设计分数赋值规则，使得预测框在越接近目

标中心的位置得分更高，远离目标中心的预测框得分则相对较低，从而使预测框更加紧密地围绕在目标周围。响应图上的每一个像素点 (i, j) 都能对应到搜索区域中的一个具体位置 (m, n) 。在分类和回归任务中，生成的特征图谱分为两类通道：一类用于鉴别前景目标与背景差异；另一类则是回归分支的结果，表现为四个通道的特征图，其中，每一个像素点都对应一个包含四个元素的向量信息 $\tilde{t}_{(i,j)} = (\tilde{l}, \tilde{t}, \tilde{r}, \tilde{b})$ ，依据公式(2-5)计算得出，该向量反映出搜索区域中该位置与真实边界框四边的距离信息。采用符号 (m_0, n_0) 和 (m_1, n_1) 分别表示候选框左上角与右下角的真实坐标值。

$$\begin{aligned}\tilde{t}_{(i,j)}^0 &= \tilde{l} = m - m_0, \tilde{t}_{(i,j)}^1 = \tilde{t} = n - n_0 \\ \tilde{t}_{(i,j)}^2 &= \tilde{r} = m_1 - m, \tilde{t}_{(i,j)}^3 = \tilde{b} = n_1 - n\end{aligned}\quad (2-5)$$

特征图中的点 $A_{w \times h \times 1}^{cen}(i, j)$ 的值在搜索图片中得分为 $C(i, j)$ ：

$$C(i, j) = Q(\tilde{t}_{(i,j)}) \times \sqrt{\frac{\min(\tilde{l}, \tilde{r})}{\max(\tilde{l}, \tilde{r})} \times \frac{\min(\tilde{t}, \tilde{b})}{\max(\tilde{t}, \tilde{b})}} \quad (2-6)$$

$$Q(\tilde{t}_{(i,j)}) = \begin{cases} 1 & (\tilde{t}_{(i,j)} > 0), k = 0, 1, 2, 3 \\ 0 & otherwise \end{cases} \quad (2-7)$$

如果点 (i, j) 在搜索区域的 Ground Truth 中，则如公式(2-7)一致。跟踪器中心度分支损失函数 L_{cen} 为：

$$L_{cen} = \frac{-\sum_{Q(\tilde{t}_{(i,j)})=1} C(i, j) * \ln A_{w \times h \times 1}^{cen}(i, j) + (1 - C(i, j)) * \ln(1 - A_{w \times h \times 1}^{cen}(i, j))}{\sum Q(\tilde{t}_{(i,j)})} \quad (2-8)$$

2.3 注意力机制

在人类认知的过程中，面对复杂的现实环境，视觉系统能够有意识地选择筛选出大量的信息，优先聚焦于那些对人类具有重要意义的内容，这一特性在一定程度上归因于人眼视网膜独特的凹面构造，它使得眼球能够在多个视角和全方位范围内有效地捕捉并处理信息。这一生物学原理^[41,42]激发了计算机视觉领域的创新，研究人员尝试模拟类似机制，让计算机视觉算法能够在处理图像时智能地集中资源到图像的关键区域——即通常指代特征响应强度较高的部分，从而提升图像处理系统的效能。

基于此，一种被称为“注意力机制”的技术在计算机视觉中得到广泛应用，并且其概念目前并未有严格统一的界定。一些图像处理手段，如窗口滑动特征提取和显著性检测^[43,44]等，都在某种程度上体现了该思想。注意力机制的核心在于给予不同输入信息不同的关注度，以此来智能识别并强调哪些区域的数据最为关键。参照图 2-5，它通过设计 Query 和 Key 之间的注意力匹配过程来运作，即通过比较 Query 与一系列 Key 之间的关联性来确定权重分配，然后依据这些权重对相应的 Value 信息进行加权整合，最终生成精细化的输出向量。

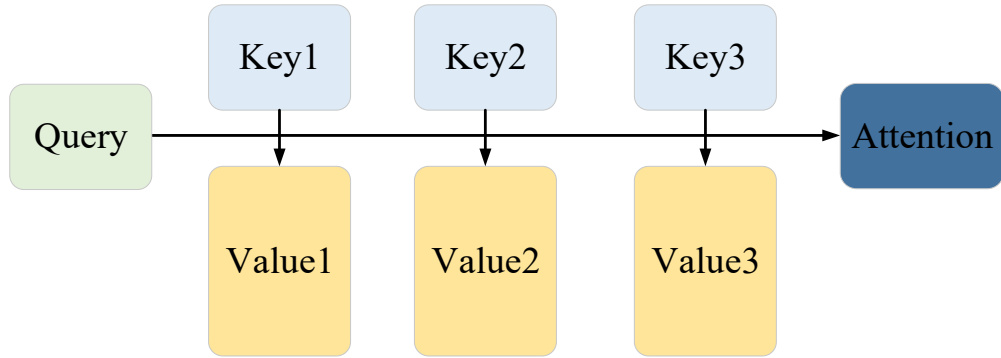


图 2-5 Query、Key、Value 计算

通过衡量 Query 与每一个 Key 之间的匹配程度，我们可以为每个 Key 关联的 Value 赋予一个权重因子，进而将所有 Value 按照各自权重进行累加，从而获得最终的 Attention 得分。换言之，Attention 的本质是在对 Source 中各元素的 Value 进行处理时，采用 Query 与 Key 共同确定 Value 的权重，并据此进行加权求和运算。该机制的基本理念如下所示：

点积：

$$\text{Similarity}(\text{Query}, \text{Key}_i) = \text{Query} \cdot \text{Key}_i \quad (2-9)$$

Cosine 相似性-余弦相似度：

$$\text{Similary}(\text{Query}, \text{Key}_i) = \frac{\text{Query} \cdot \text{Key}_i}{\|\text{Query}\| \cdot \|\text{Key}_i\|} \quad (2-10)$$

分子就是两个向量 Query 与 Key 的点积，分母就是两个特征向量的 L2 范数。

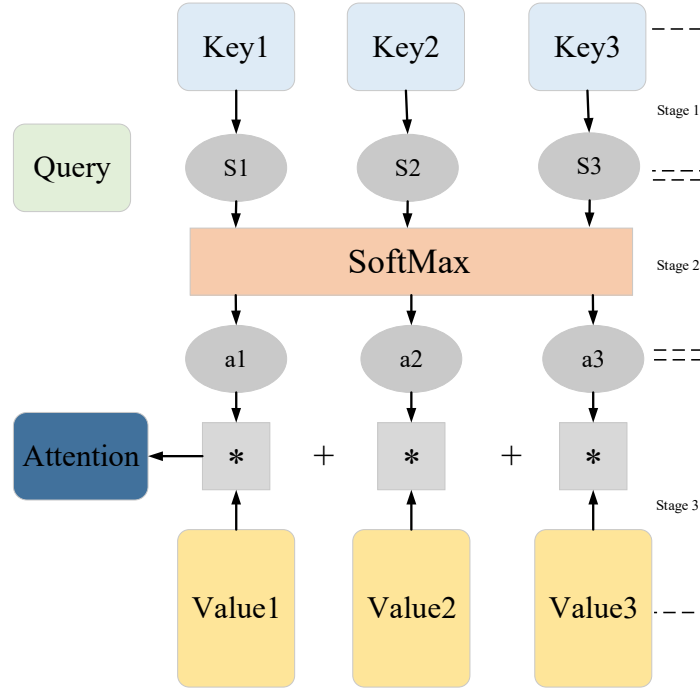


图 2-6 注意力三阶段计算过程

参照图 2-6 所示，在三个计算步骤中，首先计算了 Query 与 Key 之间的相似度，接着在第二阶段采用类似于 Soft max 的计算方法对第一步中得到的相似性得分进行归一化转换。这一转换过程不仅实现了将原始分数归一化为权重总和为 1 的概率分布，而且还能通过 Soft max 内部的特定结构进一步凸显重要元素的权重分量。通常情况下，采用下列方法完成这一计算过程：

$$a_i = \text{Soft max}(Sim_i) = \frac{e^{Sim_i}}{\sum_{j=1}^{L_x} e^{Sim_j}} \quad (2-11)$$

阶段二计算所得到的 a_i 可视为对 Value_{*i*} 元素权重分配的系数，在此基础上进行加权求和运算后，即可得出最终的 Attention 值：

$$\text{Attention}(\text{Query}, \text{Source}) = \sum_{i=1}^{L_x} a_i \cdot \text{Value}_i \quad (2-12)$$

2.4 数据集及评价指标

随着目标跟踪技术的蓬勃发展，层出不穷的新型算法不断涌现，为了更好地满足对这些算法性能进行公正且全面评估的需求，研究者们推出了一系列公共数据集和评价方法。接下来，本文将对所采用的一些通用且公开的目标跟踪数据集及其评价标准作一番简要阐述。

2.4.1 目标跟踪数据集

OTB-100^[45]数据集收录了一百个视频序列，专门记录了包含人物、车辆、动物等各类运动目标在各类复杂环境下的状态。该数据集采用尺度变化、光照变化、相机运动等共计 11 种标签来标识跟踪过程中的难题。在特定情况下，该数据集会对部分具有挑战性的视频序列做细致划分，例如将 Human4、Jogging 以及 Skating2 等视频分割成多个小片段，这些细分的子序列分别代表了不同程度的跟踪难题。

VOT^[46]数据集系列由 VOT 挑战赛主办方发布，其中 VOT2018 版本尤为广泛的应用于各类研究与实践场景之中，包含了 60 个短期跟踪视频序列，涉及行人、车辆、动物等多种对象类型，且在序列中，目标往往伴有大量遮挡和变形等复杂情况。对于 VOT 数据集，其主要的评估标准涵盖了 Accuracy、Robustness 以及 EAO 三个核心维度。

GOT-10k^[47]基准数据集规模庞大且多样化，容纳了百万个众多类别中形态各异的目标物体，每一段视频序列均配备了精细的标注。它的核心目标是通过模拟大量且复杂多变的跟踪场景，助力科研人员深入评估和提升目标跟踪算法在大规模、持久跟踪任务中的性能稳定性和适应能力。它的评价指标为 AO 、 $SR_{0.5}$ 和 $SR_{0.75}$ ，且需要在 GOT-10k 官网上进行测试。

LaSOT^[48]作为一个专为大规模目标跟踪研究打造的权威数据资源库，汇集了 142 类目标和逾万条富含多样变化的视频序列。这些序列中的目标物体位置在每一帧画面中均经过了严谨的标定处理。此数据集被合理划分为训练集和测试集两个独立部分，以便于模型训练与性能验证。其所涵盖的视频场景复杂多变，真实还原了目标跟踪技术在实际应用中可能遭遇的各种难题，为提升跟踪算法的鲁棒性和泛化能力提供了宝贵的实验环境。

UAV123^[49]数据集是专门用于为 UAV 跟踪平台打造的权威数据资源库，涵盖了众多复杂场景，如光照、旋转等，这些都是在空中视角下经常遇到的问题。作为一个广泛应用于目标跟踪和无人机视觉研究领域的重要数据集，UAV123 数据集可以用于视觉跟踪算法性能的评估与比较。

2.4.2 相关评价指标

重叠得分(Overlap Score)，指的是真实的目标位置框与预测的目标位置框的

交并比，公式定义如公式：

$$OS = \frac{|a \cap b|}{|a \cup b|} \quad (2-13)$$

在公式中，我们所指的跟踪算法生成的目标边界框(记为 a)，以及作为参照基准的真实标注边界框(记为 b)。

在目标跟踪评估中，**Success Plot** 是用来反映跟踪性能的一个指标。当某帧画面的 OS 超过预设的临界值时，这帧就被视为有效跟踪(**Success**)。在视频序列的全部帧数中，成功跟踪到目标的帧所占的比例被定义为跟踪成功率。值得注意的是， OS 的取值介于 0 到 1 之间，并以此为基础绘制出 **Success Plot** 曲线；在大多数情况下，这个阈值标准会被设定为 0.5。而 **Precision Plot** 则是通过对单帧精度的考察来呈现跟踪效果的另一个维度。具体来说，当预测框的中心点坐标与实际目标框中心点坐标之差的绝对值小于一个特定的阈值时，则该帧即被认为是精度达标。由于在不同阈值下达到精度要求的帧所占比例各异，故可以通过绘制 **Precision Plot** 曲线来展现这一动态变化。通常，对于判断精度是否达标的阈值设定为 20 像素的距离。此外， AO (**Average Overlap**) 是用来量化真实目标框与预测目标框之间重叠程度的平均值，它的作用与上述的概念相通，都是用来评估跟踪器性能的有效指标。

2.5 本章小结

本章集中探讨了目标跟踪领域的核心理论知识，开始部分详述了孪生网络的基础概念与架构特性，并系统性地剖析了代表性的经典跟踪模型——**SiamFC**、**SiamRPN** 以及无锚框结构的 **SiamCAR**。此外，本章还深入解读了注意力机制的基本原理及其计算流程，并在最后章节全面介绍了目标跟踪研究中广泛采用的数据集和评价体系，为后续章节即将提出的目标跟踪算法的评估工作提供了必要的理论基础。

第3章 基于特征融合和显著性图注意力的目标跟踪方法

3.1 引言

目前, 众多前沿的目标跟踪技术大多依托于孪生网络架构。这种基于孪生网络的目标跟踪方法凭借其卓越的跟踪效能而备受瞩目, 其通过离线训练模型并在在线阶段实施跟踪, 从而实现高效的跟踪性能。一些全局跟踪策略^[18,25], 如 TADT^[50]和 SINT^[51]等, 核心思想是运用统一的距离度量函数或相似性度量手段, 对搜索区的目标特征与预先设定的模板特征进行对照分析, 从而甄选出与模板最为吻合的目标候选区域。然而这些方法在目标出现较大姿态改变、遭受遮挡等情况时, 性能往往有所下滑。此外, 这些算法普遍依赖相同大小的卷积核来计算互相关度量, 这意味着跟踪器在处理尺寸变化或宽高比各异的目标时, 相同大小的卷积核可能存在过拟合现象, 会引入大量冗余甚至误导信息, 从而降低跟踪准确性。另一方面, 基于在线学习^[52]相关滤波器的一些方法能迅速应对目标尺度的变化, 但由于同样采用目标整体进行跟踪, 面对目标外形急剧转变的情形依然力不从心。鉴于以上提到的局限性, 局部跟踪策略^[53-55]得到了众多学者的广泛关注和深入发展。典型做法是将目标分割成多个同等大小的区块, 进而并行跟踪每一个区块, 最后汇总所有区块的信息以估计目标的整体运动状态。不过, 这种方法存在一个潜在缺陷, 即并非所有区块都含有有效目标信息, 在类似目标干扰的情况下可能导致误判。综上所述, 在视觉跟踪领域中, 如何有效地在孪生网络的模板分支与搜索分支之间建立高度相关的响应图, 以便精确地将目标特征映射至搜索空间, 是一个亟待解决的关键问题。因此, 本文创新性地提出了一种名为 MSSiamCAR 的算法, 该算法融合了多层特征并运用显著性注意力机制进行目标跟踪, 采取了像素级精细匹配的方式, 专注于跟踪目标中最具有辨识度和判别能力的部分, 实现更精确的目标定位。能够妥善应对并有效解决多种复杂情况下的目标跟踪难题。

3.2 图注意力基本思想

图注意力网络，顾名思义，是一种基于图形结构并在该结构上运行的神经网络模型，它在图神经网络(GNN)^[54]架构中融入了注意力机制。在图注意力网络的框架下，其对节点特征的表达与普通图神经网络并无本质区别，均通过编码方式将节点特征转化为向量形式。也就是说，在构建图注意力网络时，其原始输入同样是由各个节点特征集合而成，用公式表示：

$$\mathbf{h} = \mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \dots, \mathbf{h}_n, \mathbf{h}_i \in R^F \quad (3-1)$$

其中 n 表示图中节点的数量。 F 表示的是节点特征的数量，使用完全二部图： $G = \{V, E\}$ 表示的是模版和搜索特征之间的关系。其中 $V = V_s \cup V_t$ ， V_s 是模版特征中的每个像素点，可以表示为图中的节点， V_t 表示为搜索区域的节点， $E = \{(u, v) | \forall u \in V_s, \forall v \in V_t\}$ ，图中的每个节点表示为 (i, j) ，其中 $i \in V_s$ ， $j \in V_t$ ，计算节点 i 和 j 之间的互相关得分，记为 $e_{ij} = f(h_s^i, h_t^j)$ ，计算公式如下：

$$f(h_s^i, h_t^j) = (W_s h_s^i)^T (W_t h_t^j) \quad (3-2)$$

h_s^i, h_t^j 表示节点 i, j 的特征向量， W_s, W_t 是线性变换矩阵。在图注意力机制框架中，我们采用内积方法来构建并计算相似性图谱，使用 Softmax 函数，如下式所示对相似度图进行正则化调整。

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in V_t} \exp(e_{ik})} \quad (3-3)$$

计算从 G_t 中的所有节点传递到 G_s 中的 i 个节点的权值。

$$V_i = \sum_{j \in V_t} a_{ij} W_v h_t^j \quad (3-4)$$

其中 W_v 是线性转换矩阵。最后我们将特征权重与节点特征相融合，得到一个更丰富的特征：

$$\hat{h}_s^i = \text{ReLU}(v_i \parallel (W_v h_s^i)) \quad (3-5)$$

其中 $\parallel$ 表示为向量之间的叠加拼接。

3.3 特征融合和显著性图注意力网络结构

在本章节中，我们将详尽阐述 MSSiamCAR 的框架，这项工作最主要的研究内容是：重点关注并跟踪目标物体的局部显著区域，这些区域具备鲜明的辨识度，很容易在搜索区域被定位。利用图注意力来使局部显著性区域之间建立有效联系，将对跟踪起到关键作用，这能够有效的应对跟踪目标的外观变化，例如旋转平移和遮挡。我们的跟踪框架图如 3-1 所示。MSSiamCAR 的网络架构由四部分组成：(1)孪生子网络进行目标特征提取；(2)多层特征融合模块，将 ResNet-50 的 Conv3、Conv4、Conv5 进行特征融合；(3)显著性注意力模块，由显著性捕获和显著性交互组成；(4)正负样本分类、回归和中心度计算。

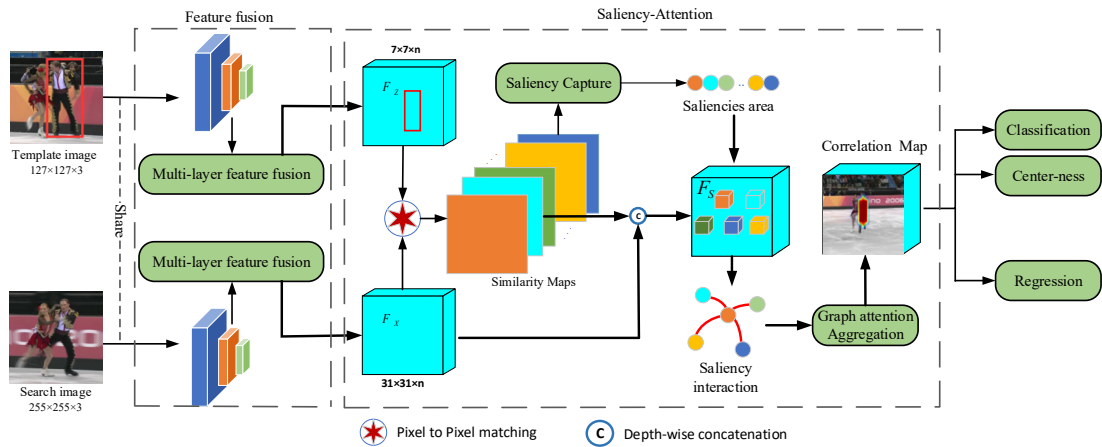


图 3-1 MSSiamCAR 的网络框架

3.3.1 特征融合模块

本章节使用 ResNet-50^[22]作为特征提取网络，根据孪生网络的结构特点，将模版与搜索分支分别送入网络使用相同的权重用于提取特征信息。如图 3-1，两分支输入的参数分别为 $127 \times 127 \times 3$ 和 $255 \times 255 \times 3$ 。在本章节提出的 MSSiamCAR 算法中，为了避免因使用深层的网络结构而导致对关键浅层目标特征的忽视，我们引入了多层特征融合技术。如图 3-2 所示，以模板分支为例，网络含有五个残差块：Conv1、Conv2、Conv3、Conv4、Conv5，我们对后三层进行融合操作，模版分支的输入经过卷积处理后分别得 $15 \times 15 \times 256$ 、 $15 \times 15 \times 1024$ 和 $15 \times 15 \times 2048$ 的特征图。紧接着使用 1×1 的卷积核进行降维，保持他们的维度一致。然后经过 Crop 将大小设置成 $7 \times 7 \times 256$ ，再沿深度方向拼接成 768 维，最后经过逆卷积映射到低维度。搜索分支经过相同的融合步骤可以得到 $31 \times 31 \times 256$ 的搜索特征图 F_X 。

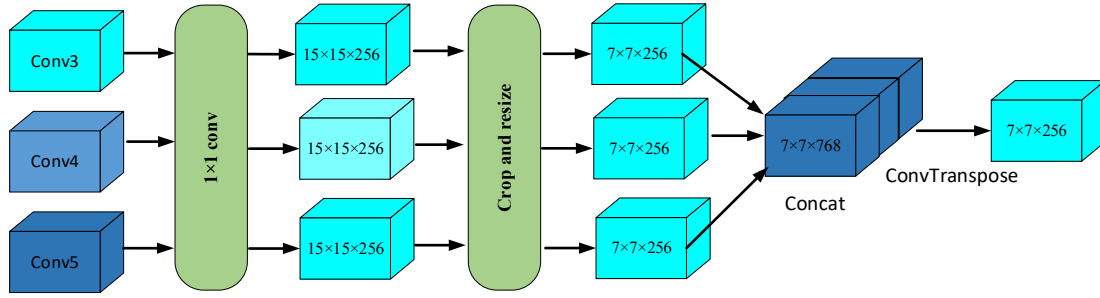


图 3-2 特征融合模块

3.3.2 显著性图注意力模块

在目标物外形发生显著变化的情况下，常规跟踪器的性能往往会下降，为应对这一问题，本文采用了局部跟踪方案，即将目标拆解为多个子区域，并仅关注和跟踪目标的显著区域。本研究设计了一种显著性注意力模块，首先，通过显著性捕获单元从目标模板中提炼出局部显著特征，这些特征具备鲜明的辨识度，将对跟踪效果起到关键作用。随后，在显著性交互单元中将在捕获单元挖掘到的具有辨别能力的显著性区域之间，建立图注意力边的关系，使跟踪器学习在目标和搜索图像之间建立适应目标变化的能力，从而在跟踪过程中提升准确性与鲁棒性。接下来，将分别对显著性捕获模块和显著性交互模块展开详细的阐述。

（1）显著性捕获

将模板特征 $F_z \in \mathbb{R}^{h_z \times w_z \times c}$ 和搜索特征 $F_x \in \mathbb{R}^{h_x \times w_x \times c}$ 划分成 $h \times w$ 个大小为 $1 \times 1 \times c$ 的节点，再令二者做 Pixel-wise Correlation^[18]的操作，计算 F_z 中每个像素点和 F_x 中每个像素点之间的相似度，构建张量大小为 $h_x \times w_x \times h_z w_z$ 相似度图 F_s ，如图 3-3 所示。

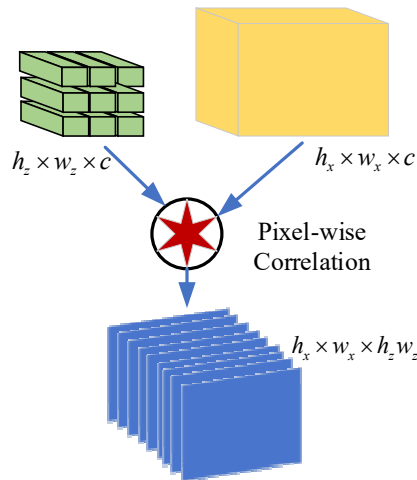


图 3-3 Pixel to Pixel 互相关

像素点 (x,y) 是模板特征 F_z 的一个节点，像素点 (u,v) 是搜索特征 F_x 的一个节点，采用 Cosine 相似性计算匹配程度：

$$S_{((x,y),(u,v))} = \frac{F_{Z(x,y)} \cdot F_{X(u,v)}}{\|F_{Z(x,y)}\| \|F_{X(u,v)}\|} \quad (3-6)$$

式中 $F_{X(u,v)}$ 是 F_x 在沿深度方向 (u,v) 处的向量， $F_{Z(x,y)}$ 同理表示， $\bullet$ 表示点积。如图 3-3 所示，其中蓝色部分为计算相似性后生成的单通道相似图，记为 $S_{(x,y)} \in \mathbb{R}^{h_s \times w_s}$ 。我们认为在相似图中得分最高的点就是该像素在搜索图像中对应的匹配位置。通过对相似图谱中峰值分布的分析，我们能够有效识别并确定图像中的显著区域。采用峰值旁瓣比(Peak-to-Side-lobe Ratio)^[56]这一指标来量化峰值分布的强度特征，参照图 3-4：

$$\text{PSR}(S_{(x,y)}; s) = \frac{F_{\max} S_{(x,y)} - \mu_s(S_{(x,y)})}{\sigma_s(S_{(x,y)})} \quad (3-7)$$

$$\psi \triangleq \text{area}(S_{(x,y)}) | \text{contour}(\text{avg} S_{(x,y)}) \quad (3-8)$$

$$\zeta(S_{(x,y)}) = \text{PSR}(S_{(x,y)}; \text{area} S_{(x,y)} - \psi) \quad (3-9)$$

式(3-7)中 s 描述的是 $S_{(x,y)}$ 中最大值点周围的无关区域， ψ 代表主瓣，它代表紧邻最大值点的区域，其界定范围为峰值点附近的最近边缘， $F_{\max}$ 为 $S_{(x,y)}$ 中的最大值点，在相似图 $S_{(x,y)}$ 中， μ_s, σ_s 分别表示为统计平均值和标准偏差。峰值集中度 ω 与 ψ 占据的面积呈负相关关系：

$$\omega(S_{(x,y)}) = A_{\psi}^{-1} S_{(x,y)} \quad (3-10)$$

综合考虑式(3-9)和(3-10)二者的影响因素，我们引入了正则项作为优化补充来定位目标中心，得到 $s(S_{(x,y)})$ ：

$$s(S_{(x,y)}) = \omega(S_{(x,y)}) [\zeta(S_{(x,y)})]^\beta + \lambda g_{\mu_g, \sigma_g}(x, y) \quad (3-11)$$

β 用于控制公式(3-9)和(3-10)二者对挖掘显著性的影响， $g_{\mu_g, \sigma_g}(x, y)$ 代表与 F_x 中心看齐的高斯分布函数。在搜索图像中，目标的显著性分数 $s(S_{(x,y)})$ 越高，其被准确发现和跟踪的可能性就越大。选取图 3-3 蓝色区域中每个单通道相似图中的显著性点，记为集合 P_z ，同理对应在 F_x 中的结合记为 P_x 。

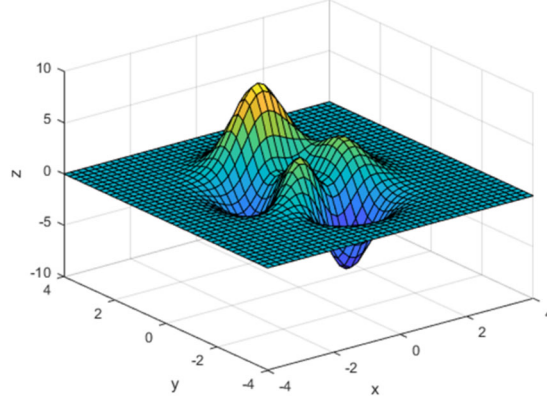


图 3-4 显著性峰值分布示意图

(2) 显著性交互

为了进一步平衡模板特征到搜索区域的数据量，强调之前捕获到的显著性信息，我们使用 Soft max 归一化 $s(S_{(x,y)})$ ：

$$N_{ij} = \frac{\exp(s(S_{(x,y)}))}{\sum_{k \in P_x} \exp(s(S_k))} \quad (3-12)$$

为了结合特征图 F_x 和相似图的有效信息，我们采取了沿通道方向上的拼接操作，生成了一个叠加式的特征图组合 $F_s \in \mathbb{R}^{h_x \times w_x \times (h_z w_z + c)}$ 。我们运用了图注意力来探索处理显著性交互问题，将 P_x 中相应的 k 个显著节点之间进行边的交互连接，借助于图网络中边所携带的信息，对显著节点信息进行整合运算，从而生成各节点的新颖表征。如图 3-5 所示，节点 i 与节点 j 之间的边的交互如下：

$$\alpha_{ij} = \text{Soft max}(\text{ReLU}(\partial^T [W_1 h_i \parallel W_2 h_j])) \quad (3-13)$$

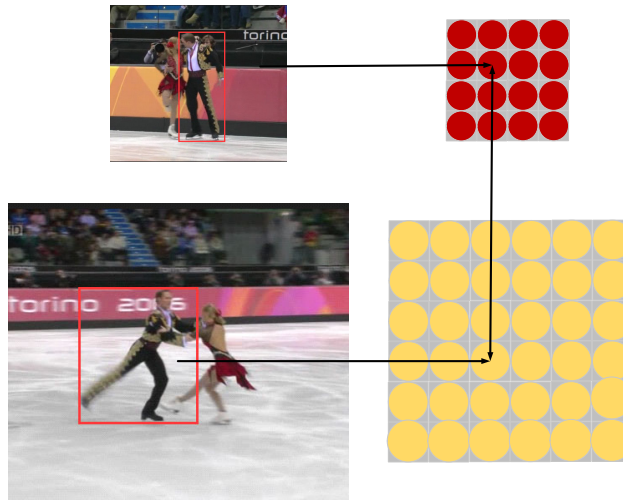


图 3-5 目标显著性特征之间边的交互

∂ 是一个超参数, h_i, h_j 代表节点 i, j 的 $h_z w_z + c$ 维度向量, W_1, W_2 为线性变化矩阵。边缘权重交互后的表征为:

$$\bar{h}_i = \text{ReLU}(\sum_{j \in h_x \times w_x} \alpha_{ij} W h_j) \quad (3-14)$$

对 $\vec{h}_i$ 中的向量执行并行处理, 从而高效地生成用于后续步骤的响应图。

3.4 实验结果分析

3.4.1 实验细节

本研究所采用的算法训练与测试均在一套高端配置的硬件平台上实施, 该平台配备两颗 Intel(R) Gold 6342 系列 CPU, 运行频率为 2.80GHz, 并搭载了四片 NVIDIA GeForce RTX 3090 系列 GPU, 每片 GPU 具备 24GB 独立显存。算法所依托的软件环境基于 Ubuntu 22.04 操作系统, 采用 Pytorch2.0 框架, 编写语言为 Python3.10。为了与 SiamCAR^[25]算法保持一致性, 我们将模板分支和搜索分支的输入图像尺寸均设定为相同的像素值。其中, 模板图像和搜索图像的大小分别为 127×127 和 255×255 。我们选用在 ImageNet 数据集上预训练过的、广泛应用的 ResNet-50 作为基础特征提取网络, 并在此基础上对模型参数进行再训练。训练数据集包含了 COCO、GOT-10k、TrackingNet 和 LaSOT 等多个大型数据集。整个训练过程共计完成了 50 个周期(Epochs), 每个 GPU 批次处理的样本数量设置为 16。在训练过程中, 我们采用了 ADAM 优化器, 并设置了初始学习率为 5×10^{-3} , 随着训练的进行, 学习率逐渐降至 1×10^{-3} , 同时, 权重衰减参数设置为 1×10^{-4} 。整个训练流程遵循了 SiamCAR 算法的基本训练步骤。

3.4.2 定量分析

(1) OTB-100 数据集结果分析

OTB-100^[45]是评估目标跟踪性能的一个基准测试数据集。参照图 3-6 所示的对比结果, 我们把本文提出的 MSSiamCAR 算法与先进的十一种跟踪算法, 包括但不限于 SiamCAR^[25]、SiamRPN++^[18]和 SiamFC^[17]等进行了对照实验。如图 3-6 显示, 其中, 红色实线代表了 MSSiamCAR 算法的性能表现。结果显示, 在成功率和精确度这两项关键指标上, MSSiamCAR 均表现卓越, 分别取得了 0.708 和 0.932 的高分。相较于基准算法 SiamCAR, MSSiamCAR 在成功率上提

高了约 1.1%，精确度上提升了 2.2%。即使与 ECO、SiamBAN 等先进跟踪器相比较，MSSiamCAR 依然展现了明显的优势。

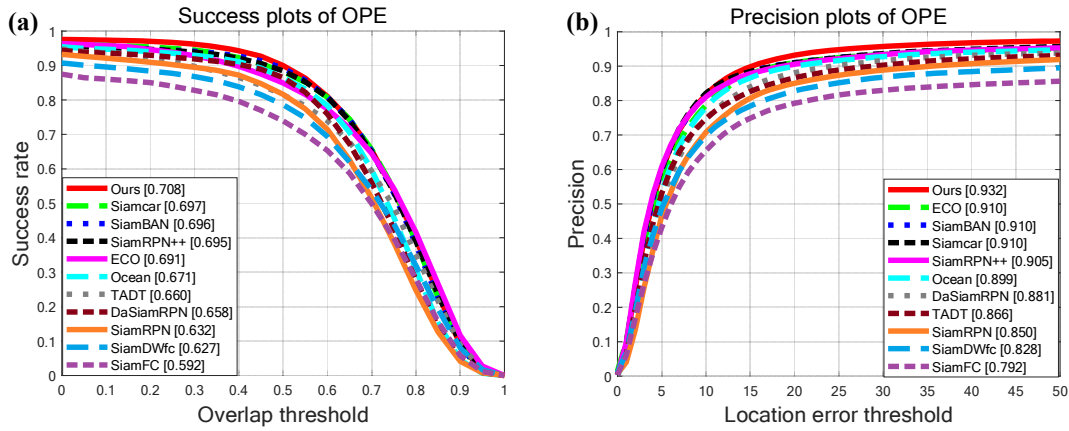


图 3-6 不同方法在 OTB-100 上的成功率 (a) 和精度 (b) 曲线图

(2) VOT2018 数据集结果分析

VOT2018^[46]数据集是 2018 年 VOT 挑战赛发布的专门用于评估目标跟踪性能的标准测试数据集。在 VOT 数据集中，跟踪器性能的优良主要体现在较高的 Accuracy 和 EAO 值，以及较低的 Lost Number 和 Robustness 值。本文提出的 MSSiamCAR 算法在 VOT2018 数据集上与包括 SiamCAR^[25]、DiMP^[57]等在内的多个先进跟踪算法进行了对比实验。从表 3-1 中可以清晰看出，MSSiamCAR 在 EAO 指标上取得了 0.448 的高分，相较于 SiamCAR 提高了 2.5%。同时，该算法在鲁棒性方面表现优秀，目标丢失数目较少。最下方的 MSSiamCAR 在 VOT2018 数据集上的整体性能位居榜首，这主要归功于设计的显著性注意力模块，使得跟踪器在面对目标复杂变化时具有更高的鲁棒性。

表 3-1 对比方法在 VOT2018 上的性能表现

算法	Accuracy	Robustness	Lost Number	EAO
SiamFC	0.492	0.571	122.0	0.207
Siammask	0.609	0.281	60.0	0.381
SiamRPN	0.586	0.276	59.0	0.383
ATOM	0.590	0.203	43.4	0.400
SiamRPNpp	0.601	0.234	50.0	0.415
SiamCAR	0.578	0.197	42.0	0.423
DIMP	0.597	0.152	32.5	0.439
MSSiamCAR	0.574	0.131	28.0	0.448

VOT2018 中的视频序列共包含了五种不同的挑战属性。在图 3-7 中，

“Overall”代表综合性能表现。结果显示，我们设计的跟踪方法在 Camera motion 和 Occlusion 这两种场景中展现出最佳性能，这有力证明了 MSSiamCAR 在这些特定场景下，能够更加稳定且准确地执行目标跟踪任务。

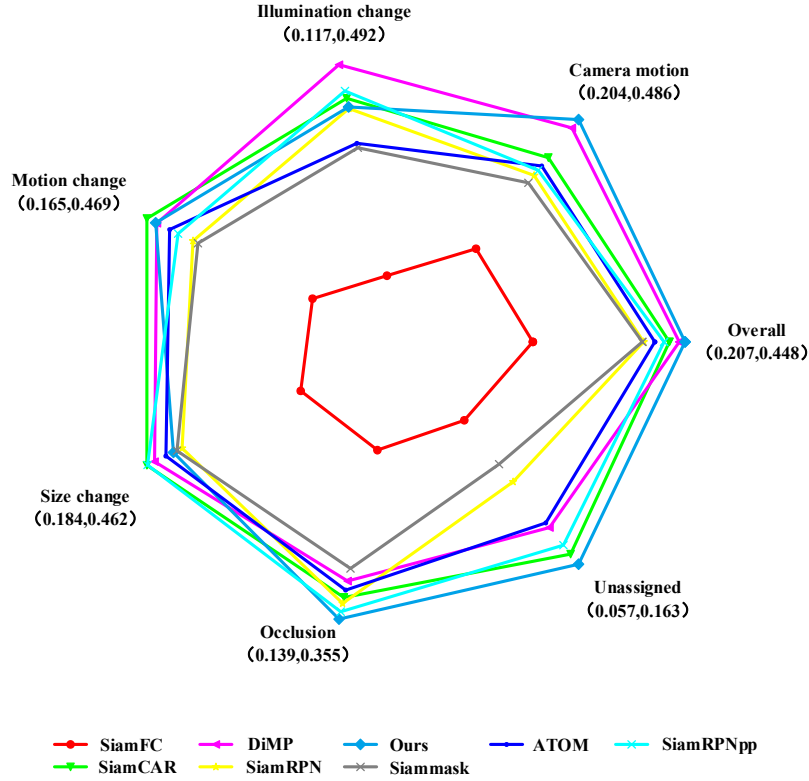


图 3-7 VOT2018 各种属性分析

(3) UAV123 数据集结果分析

UAV123^[49]数据集是一个专为无人机视角目标跟踪而设计的评测数据集，涵盖了多种复杂挑战，可以对跟踪器进行全面且严苛的性能评估。我们在 UAV123 数据集上将所提出的 MSSiamCAR 算法与九种领先的跟踪方法进行了比较，其中包括 SiamCAR^[25]、SiamBAN^[58]、Ocean^[59]、ATOM^[52]、SiamRPN++^[18]、DiMP^[57]、SiamRPN^[19]、SiamDW^[23]以及 ECO^[13]。如图 3-8 所示，我们使用了一阶性能评估(One-Pass Evaluation, OPE)的成功率曲线和精度曲线作为衡量整体性能的指标。在此次对比实验中，MSSiamCAR 算法在成功率和精度上分别取得了 0.639 和 0.844 的好成绩，位列第三，紧随基于判别式学习的顶级跟踪器 ATOM 和 DiMP 之后，且在精度上相较于基准算法 SiamCAR 提升了 1.6%，在成功率上提高了 3%。尽管在总体成功率和精度上略低于 ATOM 和 DiMP，但从图 3-9 所示的结果可以看出，在面对宽高比变化(Aspect

Ratio Change)这一特定挑战场景时, 我们的 MSSiamCAR 算法展现出了显著优势, 相较于 ATOM 算法, 其成功率高出 1.7 个百分点, 精确度提升了 1.6 个百分点。这一结果表明, 在处理无人机视角下目标外形发生显著变化的跟踪任务时, 我们的跟踪器具备较强的竞争实力。

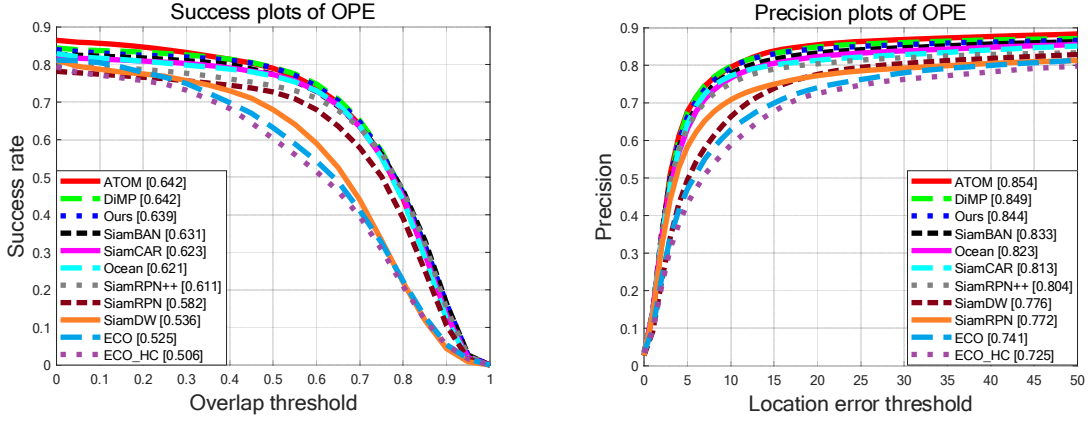


图 3-8 不同方法在 UAV123 上的成功率图和精度图

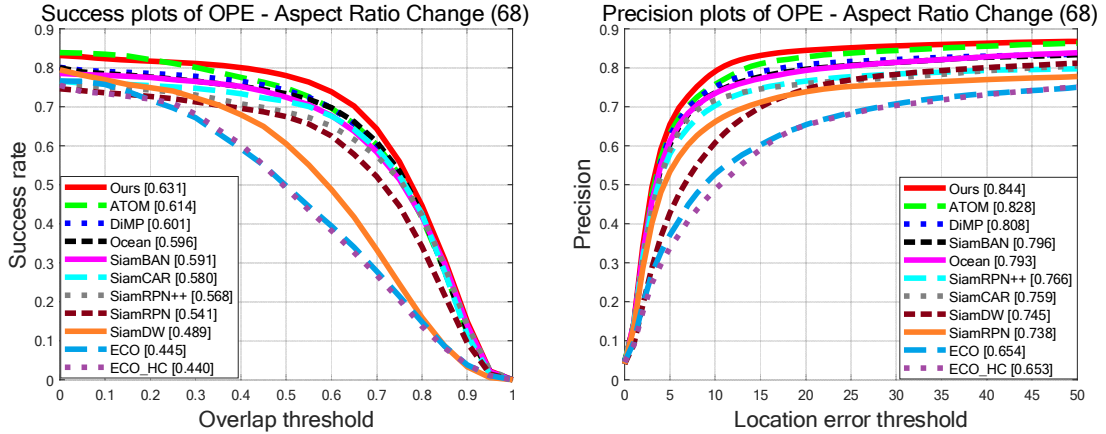


图 3-9 不同方法在 UAV123 上的属性评估

(4) LaSOT 数据集结果分析

LaSOT 数据集是一个规模宏大的目标跟踪评测资源, 如此大规模且具有挑战性的测试数据集, 精度图以及一次通过评估(One-Pass Evaluation, OPE)的成功率图被当作衡量跟踪性能的关键指标。LaSOT 官方还提供了一系列基线算法供比较参考。我们将 MSSiamCAR 算法与包括但不限于 SiamCAR^[25]、SiamBAN^[58]、SiamRPN++^[18]、ECO^[13]、DiMP^[57]等在内的数十种顶尖跟踪方法进行了对比。参照图 3-10 所示的结果, 我们的跟踪器在成功率方面达到了 0.570, 在精度方面达到了 0.650, 根据图中红色实现所示, 我们的方法在这两个评估维度上均展现出最优性能。相较于 SiamCAR, 我们的 MSSiamCAR 在成

功率上提升了 5.4%，在归一化精度上提高了 4%。与最先进的相关滤波类跟踪器 ECO 相比，MSSiamCAR 在这两个指标上分别取得了 2.46%和 3.12%的增长。这一系列领先的结果有力证明了我们的跟踪器在面对视觉对象时具有出色的泛化能力和优越的跟踪性能。

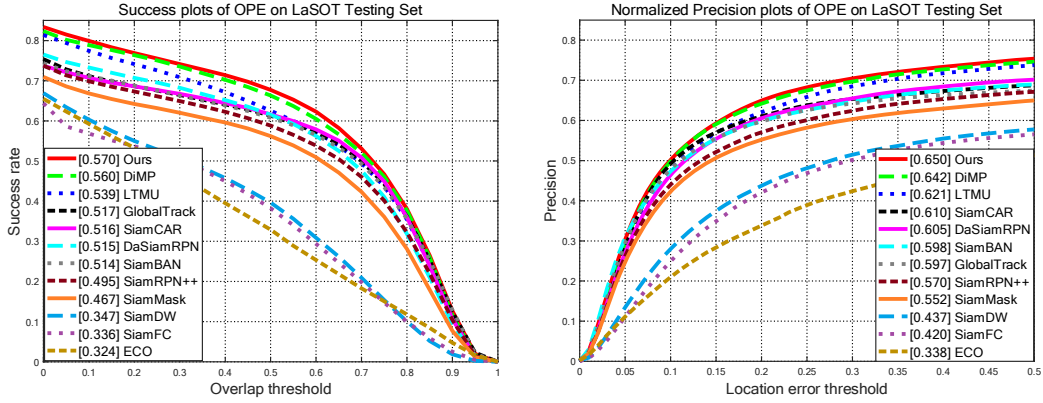


图 3-10 不同方法在 LaSOT 上的成功率和精度图

(5) GOT-10k 数据集结果分析

GOT-10k^[47]是一个大规模数据集，涵盖了超过 10000 段现实世界的物体运动视频。利用这一丰富多样且具有挑战性的数据集来评估我们的方法，可以有效验证跟踪器的泛化性能。我们把我们的方法与其它现有先进的方法进行了比较，这其中包括但不限于 SiamCAR^[25]、ATOM^[52]和 SPM^[60]等。如表 3-2 所示，最后一行高亮部分是我们的结果，我们提出的跟踪器在 AO , $SR_{0.5}$, $SR_{0.75}$ 这三项关键指标上均取得了最优成绩。相较于 SiamCAR，我们的 MSSiamCAR 在 AO , $SR_{0.5}$, $SR_{0.75}$ 上的得分分别提升了 4.4%、4.3%和 3.1%。其中， AO 代表了估计边界框与实际真值框之间的平均重叠度，而 $SR_{0.5}$ 则表示跟踪器在帧与帧之间重叠度超过 50%的成功跟踪比率。图 3-11 展示了不同跟踪方法的成功率曲线对比，从中可以清晰看出，我们的 MSSiamCAR 在各种视觉对象跟踪任务中展现出卓越的泛化能力。

表 3-2 13 种算法在 GOT-10k 中的实验结果

Trackers	AO	$SR_{0.5}$	$SR_{0.75}$	FPS
KCF	0.203	0.177	0.065	94.66
Staple	0.246	0.239	0.089	28.87
DSST	0.247	0.223	0.081	18.25
BACF	0.260	0.262	0.101	14.44
MDNet	0.299	0.303	0.099	1.52
ECO	0.316	0.309	0.111	2.62
CCOT	0.325	0.328	0.107	0.68
SiamFC	0.348	0.353	0.098	44.15
SPM	0.513	0.593	0.359	72.30
SiamRPN++	0.517	0.616	0.325	49.83
ATOM	0.556	0.634	0.402	20.71
SiamCAR	0.569	0.677	0.437	47.79
MSSiamCAR	0.613	0.702	0.468	34.71

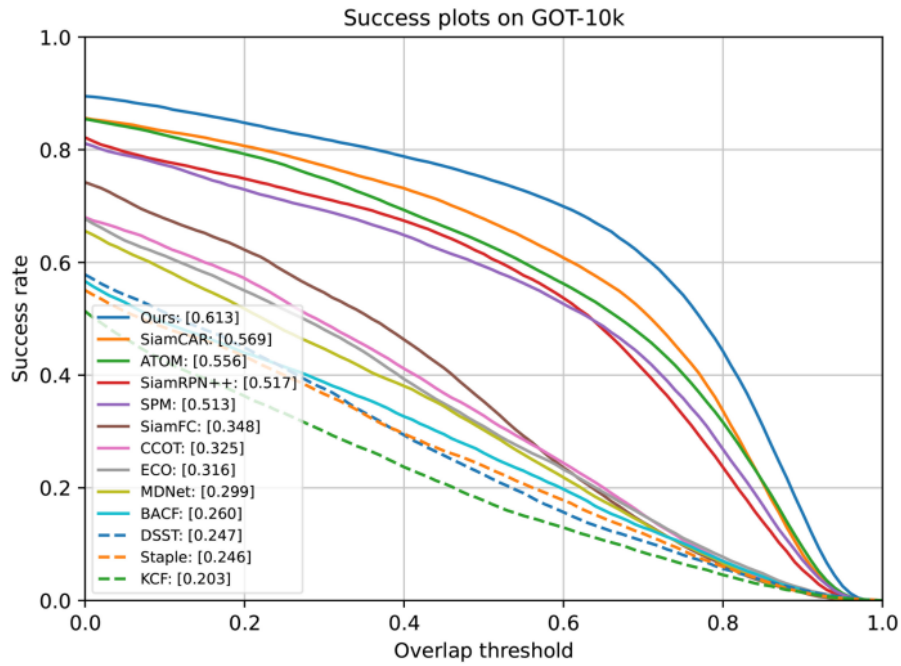


图 3-11 不同跟踪方法在 GOT-10k 上的成功率图

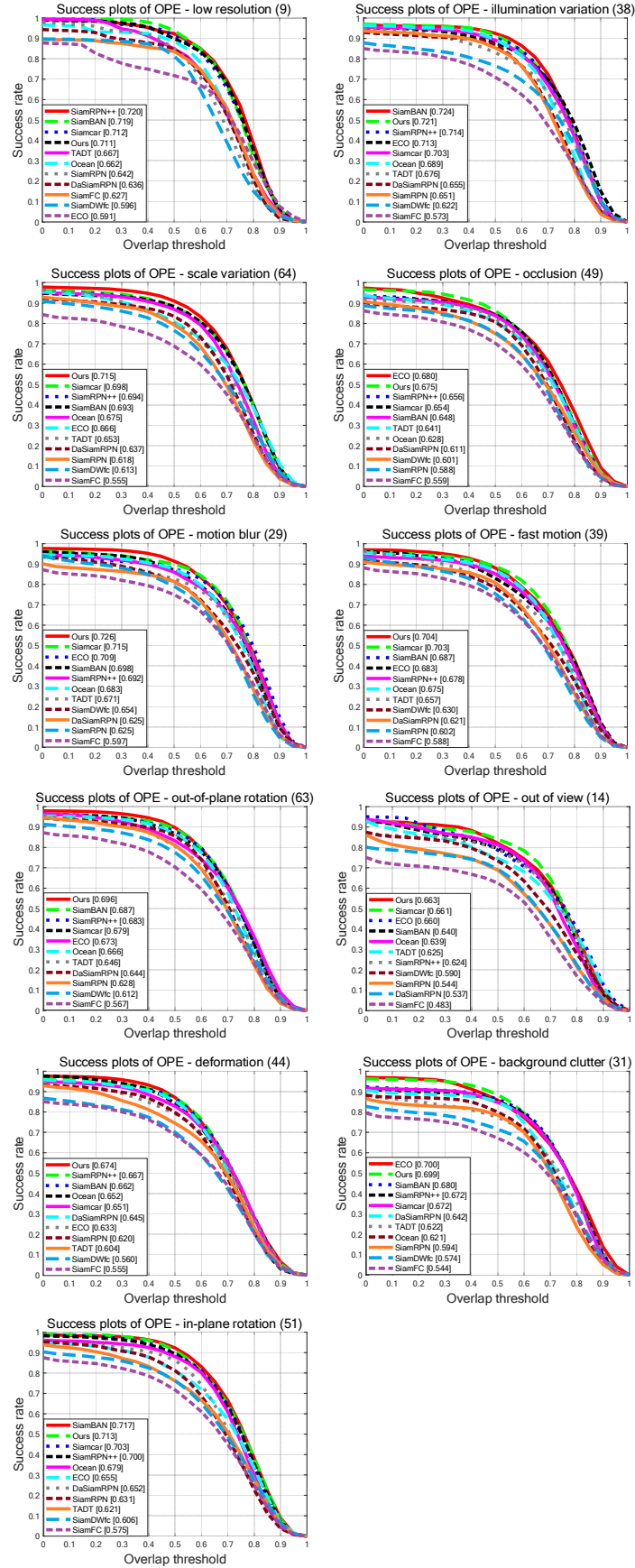


图 3-12 在成功率方面对 OTB-100 不同属性的评估

3.4.3 定性分析

在图 3-12 中展示了对不同属性的跟踪性能评估结果。相较于现有的主流跟踪技术，我们的方法在面对运动模糊、快速运动、形变、尺度变化以及平面外旋转这五大挑战属性时，依旧保持着较高的成功率和精确度。这是因为我们在跟踪过程中聚焦于目标的局部显著区域，在搜索区域中，即使目标出现快速形态变化，也能迅速有效地锁定目标位置。因此，我们提出的跟踪方法对于应对目标外观发生显著改变的场景表现出良好的适应性。

为进一步阐明跟踪器在复杂场景下的表现，图 3-13 展示了在 OTB-100 数据集上的定性分析结果，选取了三个具有挑战性的视频序列进行比较，依次为 (a)CarScale、(b)Diving 和 (c)DragonBaby。观察可知，SiamCAR 在某些情况下无法精确跟踪目标，甚至其他一些跟踪器如 ECO、SiamRPN 在某些帧中会发生误跟踪的现象。在 CarScale 序列中，包含遮挡、快速运动和宽高比尺度变化等多种复杂情况。当车辆在第 165 帧被树木遮挡时，MSSiamCAR 能够精确地锁定目标轮廓，并在车辆快速运动和不断缩放的过程中，继续保持对车辆的有效跟踪，而 ECO 和 SiamRPN 在这一过程中则未能完全跟踪到目标。在 Diving 序列中，目标物体经历了显著的尺度变化。由于运动员动作幅度大，SiamFC 等跟踪器在目标发生形变时出现跟踪失败的问题，而其它算法也不能根据目标形变自适应地调整候选框进行跟踪。相比之下，本文提出的算法仍能保持稳定地跟踪目标。在 DragonBaby 序列中，包含平面内外旋转和运动模糊等挑战。当儿童在跳跃过程中出现部分遮挡时，ECO 和 SiamCAR 等跟踪器出现了跟随错误的情况，但本算法依然能在平面旋转变化的条件下维持良好的跟踪能力。这正是得益于对目标局部显著区域的跟踪，我们的 MSSiamCAR 成功地应对了遮挡、形变和旋转变化的问题，且在红色标记的 MSSiamCAR 轨迹线上，明显提高了跟踪的精确度。

图 3-14 展示了并行计算 $\vec{h}_t$ 后的可视图。可以发现我们的跟踪器能够有效地对目标的显著区域进行精确捕获和交互，确保注意力维持在目标身上，同时渐渐减弱背景信息的干扰，从而使得跟踪结果的准确性得到显著提升。

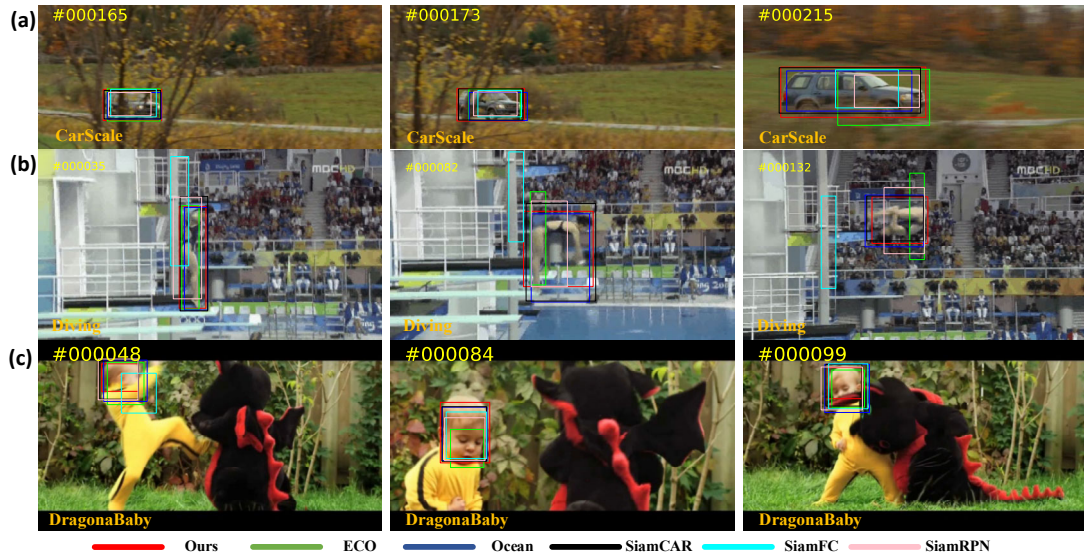


图 3-13 OTB-100 上三个序列可视化分析

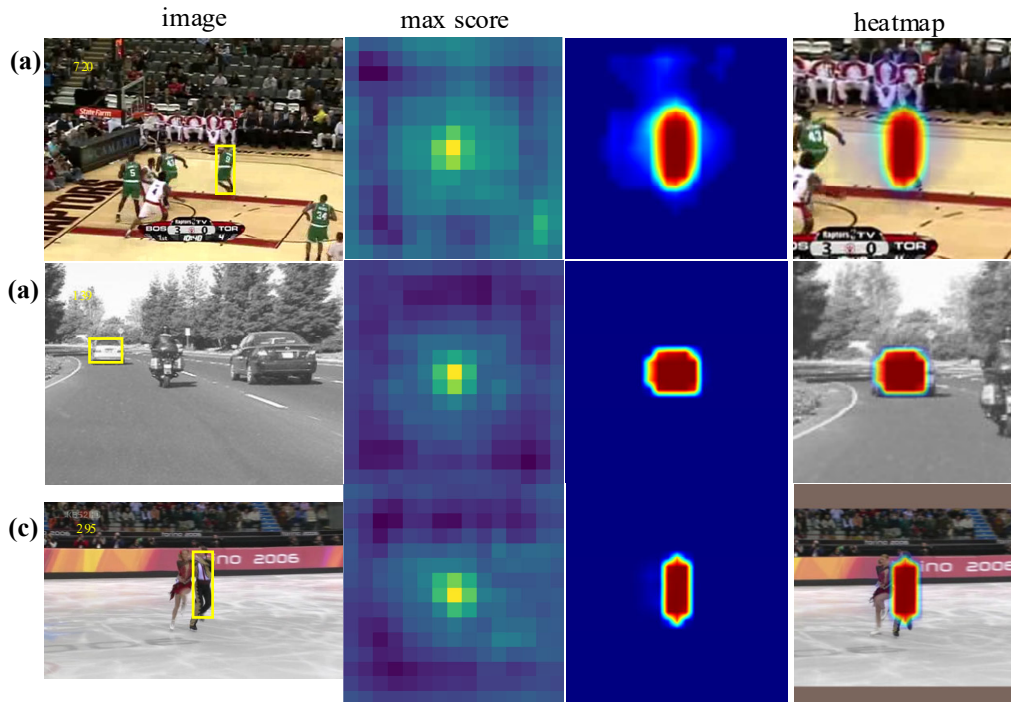


图 3-14 响应图的可视化: (a)Basketball, (b)Car and (c)Skating

3.4.4 算法实时性分析

为了进一步验证 MSSiamCAR 算法的实时性能, 将我们的 MSSiamCAR 与 ECO^[13]、SiamCAR^[25]、SiamRPN++^[18]、Ocean^[59]、ATOM^[52]、DiMP^[57]、SiamFC^[17]、TADT^[50]八个算法在 OTB-100 数据集上进行成功率与速度的测试对比, 如图 3-15 所示, MSSiamCAR 算法成功率在所提对比算法中具有较强优势, 运行速率维持在每秒 35 帧, 满足了普通跟踪器所需的实时性能标准。与

SiamFC 相比，本文算法速度慢了一半，但是跟踪性能却远远高于它，这是因为我们的特征融合网络模块对特征的提取把握的更准确，在采用残差网络作为骨干网络的前提下，相较于 DiMP 跟踪器，我们提出的方案在融入显著性图注意力模块后，虽然运行速度相近，但在性能表现上却取得了显著的进步。

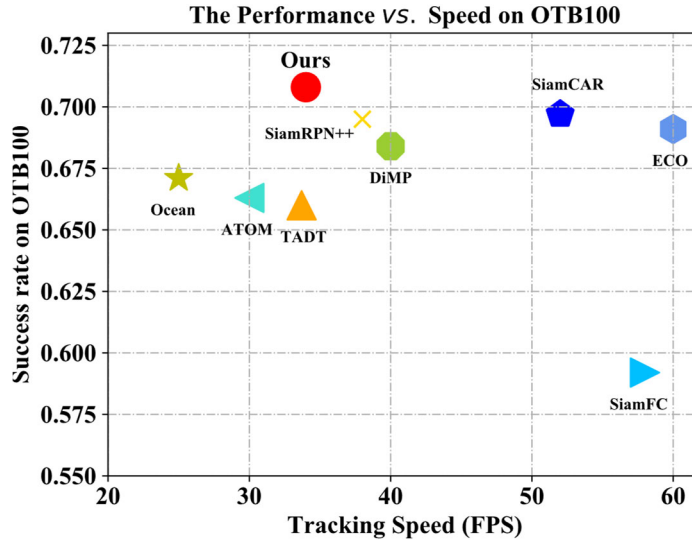


图 3-15 在 OTB-100 上的成功率与速度对比

3.4.5 消融实验

表 3-3 展示了 MSSiamCAR 在 OTB-100 数据集上的成功率(以 AUC 衡量)和精度(以 Pre.表示)。为了验证提出的各项组成单元的有效性，我们构建了五个不同的变体结构，并进行了深入的实验。其中，基础模型(Base model)以残差网络作为骨干网络，并结合了分类和回归分支，成功率和精度分别为 0.640 和 0.861。实验二在 Base model 的基础上加入了 PWS(逐像素互相关和相似度计算)，两项指标分别提升了 3.1%和 1.7%。实验三加入了 S-Interaction(显著性交互)，两项指标分别提升了 0.9%和 2.5%。实验四增加了 S-Capture(显著性捕获)，两项指标分别提升了 1.9%和 2.1%。实验五则加入了 Fusion(特征融合)。观察 1-4 组实验，我们证实了显著性捕获和交互模块提高跟踪性能是有积极作用的。实验五则表明了融合浅层特征的重要性。综上所述，该实验验证了提出的各项组成单元的有效性。

表 3-3 MSSiamCAR 在 OTB-100 上的消融实验

Base model	ResNet50	PWS	Fusion	S-Capture	S-Interaction	AUC	Precision
✓	—	—	—	—	—	0.640	0.861
—	✓	✓	—	—	—	0.671	0.878
—	✓	✓	—	—	✓	0.682	0.903
—	✓	✓	—	✓	✓	0.701	0.924
—	✓	✓	✓	✓	✓	0.708	0.932

3.5 本章小结

本章主要介绍了提出的 MSSiamCAR 算法的网络结构，对提出的特征融合模块、显著性捕获和显著性聚合模块进行了具体介绍分析，介绍了实验硬件设备与实验过程中的一些超参数，在 OTB-100，VOT2018，UAV123 等五个数据集上与其他主流的十几种跟踪方法进行了细致的比较，在各种具有挑战性的场景中进行了定量和定性的分析，将本章中基于基线 SiamCAR 所改进的各部分进行了消融实验，对结果进行分析，证明验证各模块的作用和有效性，并进行了详细的阐述。本章节主要贡献如下：

（1）研究提出一种创新的多层特征融合策略，旨在将高层次的语义理解和底层的特征相结合，以此来最大限度地强化网络在目标特征提取方面的性能。

（2）为了使跟踪器在跟踪过程中更加聚焦于目标对象的特征，我们提出了采用局部跟踪策略，即重点关注并跟踪目标物体的局部显著区域，这些区域具备鲜明的辨识度，对跟踪效果起到关键作用。该策略主要由两个核心组件构成，显著性捕获单元旨在从目标特征中筛选出对跟踪具有辨别能力的区域；而显著性交互单元则是将在捕获单元挖掘到的具有辨别能力的区域之间，建立图注意力边的关系，使跟踪器学习在目标和搜索图像之间建立适应目标变化的能力。

第4章 基于局部显著性 Transformer 的无人机目标跟踪方法

4.1 引言

视觉目标跟踪技术在空中无人机领域的多元化应用中日渐凸显其重要性，如三维空间定位、航拍成像以及特定目标跟踪等任务。无人机跟踪系统的核心任务是在首帧确定的目标模板基础上预测目标在后续帧中的位置。然而，在无人机执行跟踪任务时，频繁变换的飞行姿态与视角引出了一系列严峻挑战，比如纵横比变形、尺寸变化等问题，这些问题极大地削弱了目标模板与其搜索区域在视觉特征上的匹配度，从而严重影响跟踪器的性能表现。

尽管近来已取得一些进步，但在复杂环境下实现稳定高效的无人机跟踪仍是一项艰巨的任务。近年来，基于孪生网络架构的跟踪算法^[61,62]逐渐成为无人机目标跟踪的标准方案。此类方法将跟踪问题简化为模板匹配过程，即将初始帧中选取的目标作为后续帧查找的基准参照物。它们普遍运用互相关运算，通过对模板和搜索区域抽取特征来建立相似性映射图，进而服务于目标的分类与定位。然而，互相关作为一种线性匹配手段，在面对无人机跟踪中目标外观急剧变化的情况时，容易引入背景干扰，导致跟踪稳定性欠佳。为了应对这一问题，许多改进型的孪生网络跟踪器，纷纷引入了特征细化模块以增强特征提取的稳健性。尽管如此，互相关操作在实际应用中依然较易受类似背景信息的误导，在遭遇显著的目标外观变化时，其捕捉感知相似性的能力仍存在局限性。

最近，Transformer^[28]在计算机视觉方面显示出巨大的潜力。特别是，Transformer结构在无人机跟踪的各种场景中表现出优越的性能。为了解决感知相似性下降的问题，一些研究人员尝试用Transformer代替互相关运算。在这项工作中，我们提出了一种用于无人机跟踪的新型显著性指导动态视觉Transformer架构(LSGTT)。在四个无人机基准平台中的跟踪结果显示，提出的LSGTT在无人机视角下的挑战中表现出卓越的性能，能够妥善应对并有效解决多种复杂情况下的目标跟踪难题。

4.2 Transformer 基本思想

Transformer 模型是由谷歌设计的一种革新性框架，旨在有效应对序列到序列^[63](Seq2Seq)问题，它摒弃了传统的循环神经网络如长短时记忆网络(LSTM)，转而采用了全面的注意力架构。在机器翻译等任务上，Transformer 展现出了卓越的表现。该模型创新地引入了多头注意力机制(Multi-head Self-attention)，这是一种特别的神经网络层，其核心功能在于捕捉并量化输入序列内各个向量间的相互依赖关系。通过精心设计的矩阵运算，这种机制能够有效地利用并行计算的优势。Transformer 框架构建在注意力机制的基础之上，包括编码器和解码器两大部分。编码器和解码器均是由多个层次堆叠而成，每一层次包含了自注意力层和位置感知的前馈神经网络(Feed-forward Networks)。整个系统的结构布局清晰，可参阅图 4-1 以获得直观的理解。

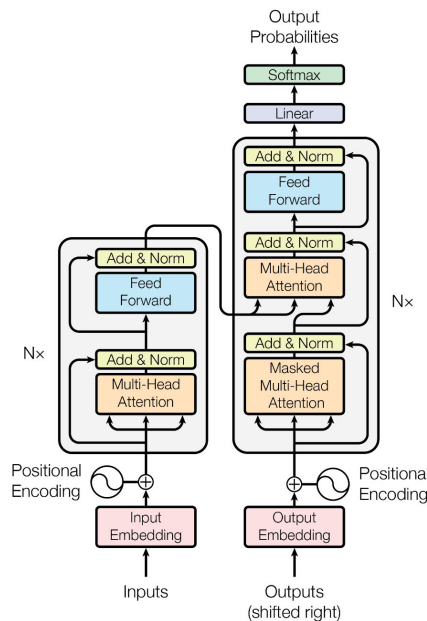


图 4-1 Transformer 框架

编码器部分，如图左边所示，其构造由 N 个堆叠在一起的编码器层(Encoder Layer)构成。每一个编码器层内部包含了按顺序链接的两个主要模块：多头自注意力机制(MHA)和前馈神经网络(FFN)。FFN 模块的设计特色在于其具备并行计算的优势，而 MHA 模块则擅长整合上下文语义信息并精准定位关注点。值得注意的是，每个 MHA 模块和 FFN 模块之后均引入了残差连接机制，这一举措旨在增强信息传递，从而提升整体性能表现。在残差连接之后，信号

会经过一层归一化处理^[64](Layer Normalization)，并将处理后的结果作为下一个编码器层的输入。因此，我们能够将任意一个编码器层的输出表达为以下形式：

$$\text{Output} = \text{LN}(\text{LN}(X + \text{MHA}(X)) + \text{FFN}(\text{LN}(X + \text{MHA}(X)))) \quad (4-1)$$

式中，注意力层的输入表示为 X ，根据前述自注意力原理，若采用自注意力机制，则 Query、Key 和 Value 矩阵均源于通过不同线性变换矩阵 W 所衍生出的结果。参照右方图形所示，解码器架构在很大程度上与编码器保持一致，但存在两个显著差异。首要区别在于，解码器环节中，原先编码器中所采用的多头自注意力被替换为了附带掩码的多头自注意力 (Masked Multi-head Self-attention)，此结构是解码器构建中的核心创新点。另外，原编码器中的第二个多头自注意力层被一种多头交叉注意力机制所取代，换言之，此处的 K、V 矩阵是从编码器输出经过差异化线性转换而得来的结果，而 Q 矩阵的生成则依赖于上一层解码器输出经过相应线性变换后的数据，这三个矩阵共同参与到交叉注意力运算中。至于使用掩模的原因，在于机器翻译等序列到序列任务的性质要求：在翻译当前词时只能参考先前已译出的词以保证语言流畅性，即位置 i 的输出仅能受位置小于 i 的输出影响而不改变它们。为此，Transformer 模型内嵌入了两种不同功能的掩模，即序列掩模和填充掩模。在训练过程中，考虑到 Ground Truth 提供了完整待翻译词汇的准确答案，对输出序列实施掩模操作至关重要，目的是确保当前单词的预测只依赖历史已知信息，不涉及未来信息，通常通过创建下三角矩阵来实现序列掩模效果。同时，将填充掩模中对应位置 i 之前的所有元素赋值为负无穷大，以限制对未出现信息的访问。

关于 Transformer 的组成，编码器部分包含 N 层相同的网络层级堆叠，如图 4-1 左部所示，每一层包括两部分：首先是基于多头注意力机制的自注意力层，并辅以残差模块连接及归一化处理；其次是前馈神经网络层，同样结合了残差模块连接和归一化技术，且在 N 层网络中输入特征维数保持恒定。编码器各层结构复用此模式。至于解码器结构，尽管总体上与编码器相似，但在图 4-1 右部所示的每层网络中，它额外包含了一个带有掩模的自注意力层。解码器的工作流程始于对待解码词进行单层自注意力处理，随后与编码器产生的特征向量通过交叉注意力机制相互作用，最终经由前馈神经网络生成输出。在训练阶段，

解码器虽然接收到整个待预测序列的数据，但需通过掩模策略使其仅能看到历史预测结果而不能窥探未来的预测信息。在实际测试时，则逐步逐词进行预测。值得注意的是，Transformer 网络中的自注意力机制扮演着关键角色，它能够在无视输入序列中元素间距离的情况下捕捉特征向量间的关联。该机制首先将输入向量转化为 Q、K、V 三个向量，各自映射至不同的子空间，接着通过查询向量与各个键值对进行加权计算并运用 Softmax 函数进行规范化，计算公式为：

$$\text{Attention}(Q,K,V)=\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4-2)$$

Q、K、V 分别代表着输入序列经过映射后生成的查询、键和值向量， d 代表键向量的维度大小。原始输入通常是词级别的向量集合，它们会被组织成矩阵形式，并进一步转化为不同的查询与键值对向量集，以便于利用矩阵运算高效并行地处理。

另外，Transformer 模型在自注意力层的核心设计中采用了并发执行多个独立自注意力模块的方式，这种结构被称为“多头注意力”机制。此举旨在使模型能够同时捕获来自序列中不同位置上的多种子空间特征信息。多头自注意力的具体表示为：

$$\begin{aligned} \text{MultiHead}(Q,K,V) &= \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \\ \text{head}_i &= \text{Attention}(QW^Q, KW^K, VW^V) \end{aligned} \quad (4-3)$$

其中，Concat 将多个头的输出拼接到一起，而 W^O 、 W^Q 、 W^K 是对应于不同向量的投影矩阵。在运用 Transformer 架构的跟踪算法中，自注意力机制的独特优势在于其能够促进全局范围内的信息流动与整合，从而为目标跟踪任务引入了一种创新的特征融合手段。这一机制使得模板特征与搜索区域特征之间能够进行更深层的相互作用和信息交流。

4.3 局部显著性 Transformer 的网络结构

在本节中，我们将详细介绍 LSGTT 的框架，如图 4-2 所示，所提出的框架可以分为五个部分，即特征提取网络、特征显著挖掘网络、显著性图精炼细化网络、局部显著性 Transformer 模块以及分类和回归网络，Q，K，V，Q' 代表整个 Transformer 框架的不同输入变量。每个编码器层用于融合不同深度的特征。

通过 Transformer 解码器和 CNN 预测头，最终结果由回归得分选择。动态词元的生成是提高实时跟踪应用 Transformer 架构效率的一个重要方面。对于此操作，一个关键因素是评估每个词元的重要性，利用显著性信息来区分出是有利的运动目标还是无效背景。本文提出的局部显著性 Transformer 网络专门为无人机跟踪任务而设计。通过挖掘出的显著性信息来确定每个词元的重要性，可以区分前景和背景，对无人机视角下的跟踪性能具有明显提升，跟踪速度也有了长足的进步。

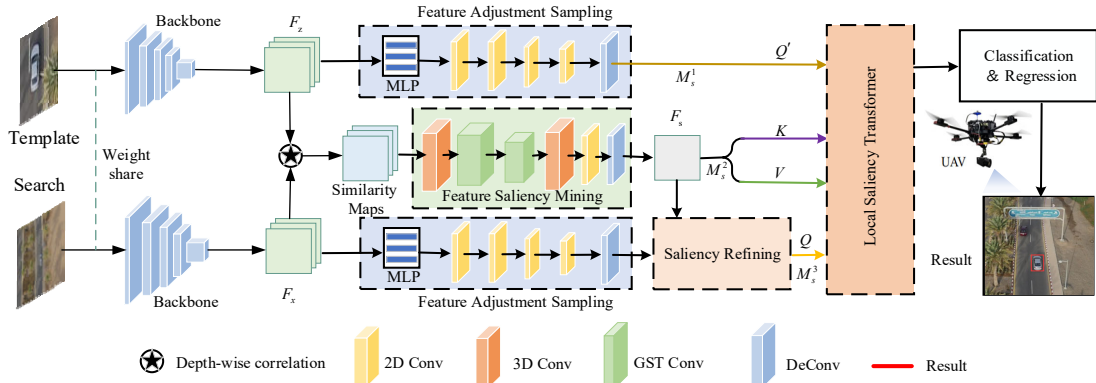


图 4-2 LSGTT 框架结构

4.3.1 显著性挖掘模块

本文选取 AlexNet^[21]作为骨干网络，用于从模板与搜索图像中提取特征信息。模板分支输入张量为 $127 \times 127 \times 3$ 的图片 Z ，搜索分支输入张量为 $287 \times 287 \times 3$ 的图片 X 。为了清楚起见，模板和搜索区域的特征图分别用 F_z 和 F_x 表示，具体来说，特征图 F_z 和 F_x 经过下式计算并调整大小为相似度图 S_1 。

$$S_1 = F_x \star F_z \quad (4-4)$$

其中 $\star$ 表示交叉相关运算符，随后使用空间融合层和通道融合层进一步处理经过 AlexNet 提取特征之后的模板特征图和搜索特征图，空间融合层包含多层感知器 MLP^[65]网络，而通道融合层由卷积-逆卷积网络组成。将两个分支经过特征调整采样后结果，分别表示为 M_s^1 ， M_s^x ，公式如下：

$$M_s^1 = \text{Deconv}(\text{Conv}(\text{MLP}(F_z))) \quad (4-5)$$

$$M_s^x = \text{Deconv}(\text{Conv}(\text{MLP}(F_x))) \quad (4-6)$$

F_z 和 F_x 经过特征调整网络后输出 M_s^1 和 M_s^x ，一个作为局部显著性 Transformer

的指导输入，另一个作为显著性细化精炼模块的输入。将生成的相似度图 S_i 送入显著性挖掘网络模块，其包含 GSTC，3Dconv 和 Deconv 结构。GSTC 是一个我们设计的挖掘显著性信息的网络结构，每个 GSTC 块由两个 Ghost 2D^[66] 空间卷积层、两个 1D 时间卷积层和 ReLU 函数组成。堆叠卷积层能够积累和提取多元化的特征表达，虽然这些信息中包含一定的冗余成分，但这一特点恰恰有助于网络体系对输入数据形成更为综合且深入的认知。但是考虑到无人机硬件跟踪平台的现实可行性，我们可以通过 Ghost 2D 空间卷积层操作提取丰富的特征信息，对于存在冗余性的特征信息部分，我们可以采用更为经济高效的线性变换机制来实现其生成过程，这样不仅能够有效地降低模型所需计算资源，同时设计简单、易于实现。1D 时间卷积层，仅考虑输入序列中每个位置的一小段窗口内的值，这意味着它能捕捉到时间序列中的局部相关性，减轻了模型对全局上下文的过度依赖，同一卷积核在整个时间序列上重复使用，这大大减少了需要训练的参数数量，并使得模型能够更好地泛化到未知的数据。3DConv 增强了在引入时态抽象建模的能力，整合来自不同深度或层面的信息，网络可以学习到连续帧之间的时间依赖性和空间结构，仅考虑每个体素与其邻域内体素的关系，这种局部连接性使得模型能够有效地识别和利用局部特征。因此，这项工作结合了 GhostNet^[66]和 S3D^[67]可实现高效的特征提取。Ghost 2D 空间卷积层的主要流程是要用较少的通道获得特征图，以防止特征冗余，然后生成最终所需的特征。时间信息是通过一维时间卷积层进行积分，该层可以公式化为：

$$\text{STConv} = \text{1DConv}(\text{G2SC}(\text{3DConv}(M_i))) \quad (4-7)$$

其中 M_i 和 M_o 分别表示两个相邻的 Conv 模块的输入和输出。G2SC 和 1DTCConv 对应着 Ghost 2D 空间卷积层和 1D 时间卷积层。需要注意的是，G2SC 中包含了 Ghost 2D 实现廉价的映射操作，3DConv 和 1DConv 分别由两个核大小为 $1 \times 3 \times 3$ 和 $3 \times 3 \times 1$ 的卷积层实现。为了方便起见，STConv 表示单个 Ghost 2D 空间卷积层和 1D 时间卷积层，那么整个 GSTC 块的工作流程可以表示为：

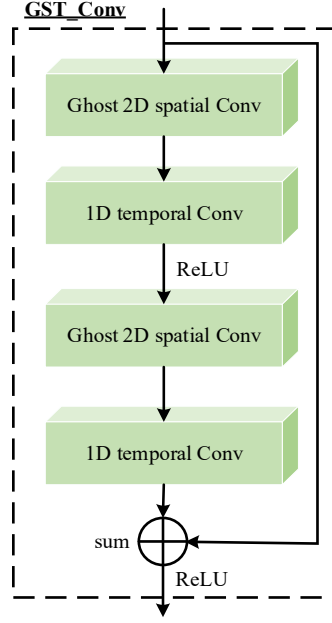


图 4-3 GSTC 结构框架

公式表示为：

$$F_o = F_i + \text{STConv}(\text{ReLU}(\text{STConv}(F_i))) \quad (4-8)$$

其中 F_i 和 F_o 分别代表 GSTC 模块的输入和输出。通过调整卷积核大小和步长，可以实现下采样操作。该分支最终生成显著性特征 M_s^2 ，将结果经过两个不同的分支，一个分支是供局部显著性 Transformer 模块使用，另一个分支是作为显著性细化精炼模块的输入 M_s^2 。

在整个显著性挖掘模块中，MLP 网络结构能够实现空间信息之间的交流，而卷积和反卷积网络结构则促进了跨通道的信息集成。多层感知机(MLP)层在捕获长依赖关系和空间关系方面具有明显的优势，而卷积层在局部特征之间的信息提取和通道信息的融合方面表现出色。这两层的集成增强了表示特征的能力。通过探索相邻帧之间包含的潜在信息，可以提供足够可靠的目标信息，使无人机跟踪更加可靠和鲁棒。

4.3.2 显著性细化精炼模块

为了准备动态词元的生成，特征调整采样网络被设计用来调整特征的大小，如图 4-2 所示。该网络包含多个具有不同感受野的卷积层，以有效解决无人机跟踪挑战，特别是纵横比变化和尺度变化。

基于对前景和背景的不同关注，我们对特征进行分块处理，将特征分割成

等大的 patch，然而并不是所有的 patch 都对跟踪任务具有相同的信息价值。如图 4-4 所示，我们提出了一种显著性细化精炼操作，以根据初始的显著性信息细化后为后续局部 Transformer 生成动态的词元输入。

首先，使用阈值处理显著性挖掘网络生成的 M_s^2 。因为考虑到获得的二进制掩码的采样是不可微分的，所以采用 Gumbel-Softmax 的方法来处理：

$$P = \text{Gumbel-Softmax}(M_s^2) \in \{0,1\} \quad (4-9)$$

输出变量 P 采取了可微分的独热编码形式，这一策略有效地解决了分类器在处理属性类数据时所面临的挑战，并且在某种程度上还增强了特征表达的维度。其取值仅限于 0 和 1 两种状态，使其可用于端到端训练。然后，使用 Swintransformer^[68]中提出的方法使用相同大小的窗口，将整个显著性特征图分成多个块，从而确保掩膜和特征在空间位置上对齐。将搜索分支经过特征调整采样后的特征记为 F_x' ，经过显著性挖掘网络之后的特征记为 F_s ，将他们分成等大的窗口，窗口与窗口之间是一一对应的，在同一深度上，对每个窗口进行求和，生成一幅单通道的特征图，根据划定的阈值进行分类，将显著性信息进一步细化精炼，并根据结果将块分成两部分，分为显著和非显著性区域。对于得分值更高的部分，我们认为相应的区域可能包含跟踪目标。因此，这些块将会被进一步细分成为较小的块，然后被嵌入到词元中。相反，得分值较低的部分，被认为包含更大比例的背景信息，较小的值对应的块被直接嵌入到词元中。由这些非显著性区域产生的词元与从具有较高得分的、进一步细化精炼的显著性区域派生出的词元相结合，形成用于局部显著性 Transformer 的输入 M_s^3 。与直接删除无用的或者目标信息占比小的常见的动态词元生成方法不同，我们所提出的方法将精炼细化后的显著性区域与包含更大比例的背景信息在多个层级上，将整个词元动态的嵌入显著性细化精炼网络。通过显著性信息的得分区分对跟踪任务的贡献，后续的显著性局部 Transformer 模块可以动态地确定每个词元的重要性，为后续跟踪任务提高效率与性能。

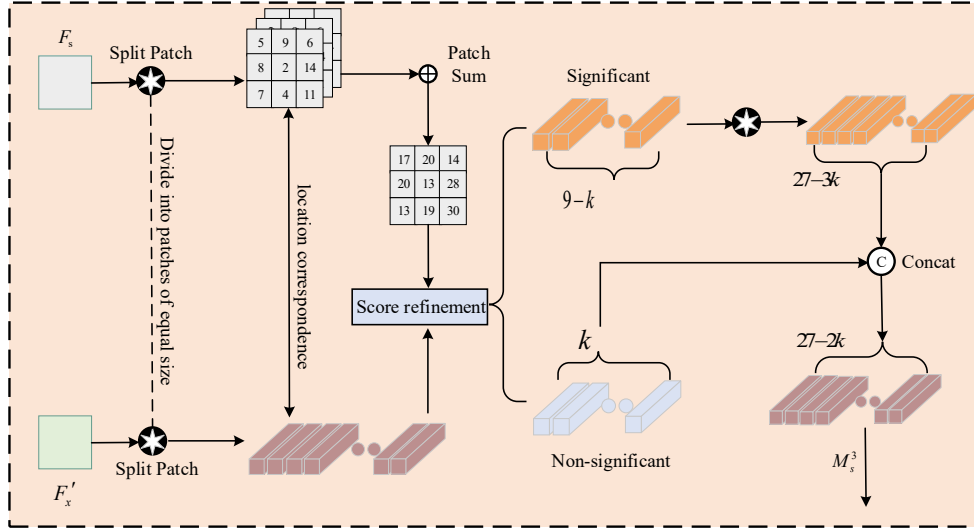


图 4-4 显著性细化精炼结构框架

4.3.3 局部显著性 Transformer 模块

为了细化跟踪目标的显著性信息并消除背景带来的干扰，本章设计了一种新颖的局部显著性Transformer结构LST (local-Saliency Transformer block)。与典型Transformer模块的注意力不同，LST采用局部感受野来消除背景干扰，并将时间和空间复杂度降低到 $O(n)$ 。通过使用卷积层来嵌入Query和Key来调整注意力结构的偏差信息，并使用线性层来保持每个词元的独立性。

参照图 4-5 所示，该模块的输入内容由三个部分整合而成。通过特征调整采样处理的模板特征 M_s^1 ，通过显著性细化精炼模块生成的动态词元 M_s^3 和通过对对象显著性挖掘模块获得的显著性特征 M_s^2 。

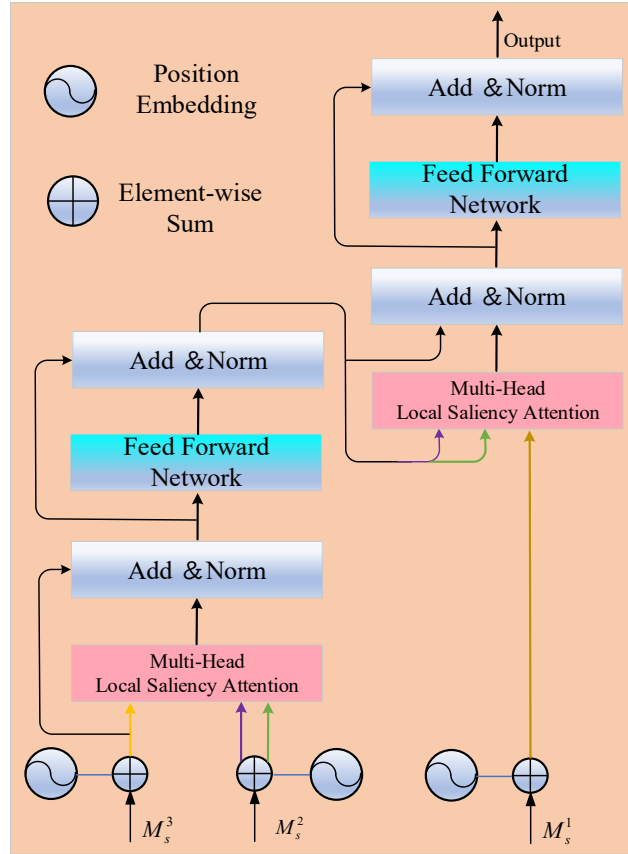


图 4-5 局部显著性 Transformer 框架

对于编码器，不同于标准的 Transformer，考虑到无人机硬件平台的现实可能性，算力有限，所以我们只使用了一个编码器和解码器结构，网络结构中使用动态词元 M_s^1 作为解码器的 Q(Query)，挖掘的显著性特征 M_s^2 作为编码器的 K(Key)和 V(Value)。精炼细化的显著性 M_s^3 作为编码器的 Q。在编码器结构中， M_s^2 提供的 Key 和 Value， M_s^1 提供的 Query，经过位置编码后，进入多头局部显著性注意力模块，如图 4-6 所示，Q 经过进入线性嵌入层，可以保持线性关系，随机降低模型复杂度。

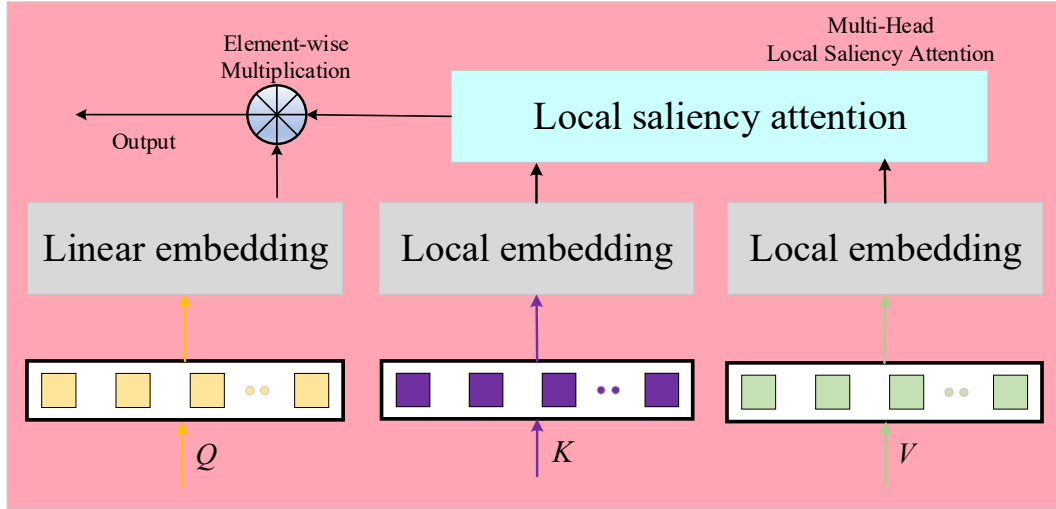


图 4-6 多头注意力结构

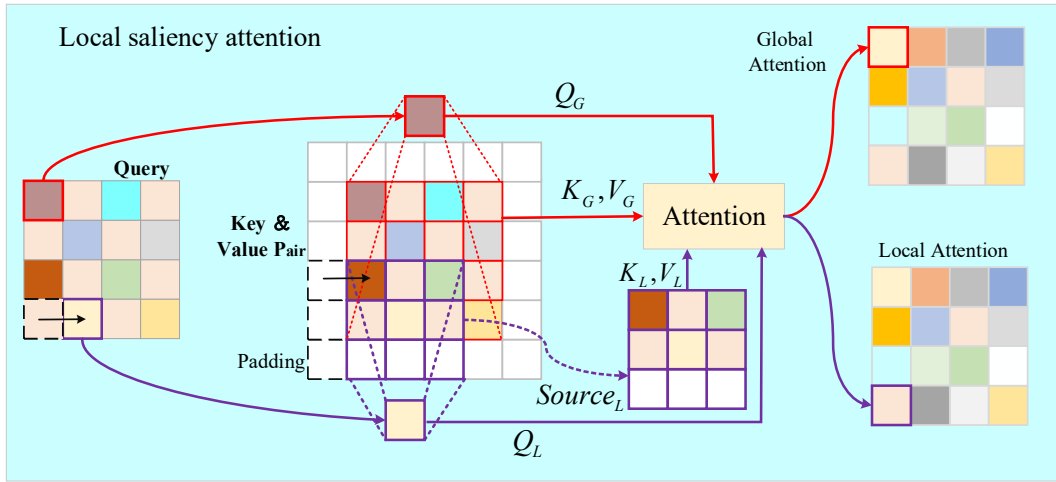


图 4-7 局部显著性注意力结构

K 和 V 进入局部嵌入层，使词元保持局部结构，与全局线性或非线性降维方法相比，局部嵌入更擅长于保留数据的局部流形结构。每个词元可以通过其邻近点的加权线性组合来重构，因此重建的词元之间的相对关系得以尽可能地保持不变，之后进入局部显著性注意力模块，典型的Transformer建立在多头注意力的思想基础上，多头注意力从不同子空间的不同位置构建强大的表达特征。由于需要对两个 $n \times d$ 矩阵进行乘法运算，标准注意力的时间和空间复杂度为 $O(n^2)$ 。实际上，计算完整的注意力图是昂贵的，并且不适合用于局部建模。如图4-7所示，本章通过研究将标准Transformer全部自注意力换成在每个 M_s^3 的 Q 周围设置 $m \times m$ 的局部视野，用局部识别注意力来修改Transformer编码器来聚合特征，代替全局注意力。具体而言， $d = c/h$ ，其中 h 被设置为并行注意力

头的数量。因此每个注意力头被定义成如下公式，通常，点积注意力(Att)计算如下：

$$\text{Att}(Q,K,V)=\text{Softmax}(\frac{QK^T}{\sqrt{d}})V \quad (4-10)$$

其中 $\sqrt{d}$ 是缩放因子，用于避免Softmax函数中的梯度消失。然后，多头注意力模块mAtt的过程表示为：

$$\hat{h}_i = \text{LSA}_i(Q,K,V) = \hat{L}\hat{V}_L = \text{Att}(Q_L W_1^i, K_L W_2^i, V_L W_3^i) \quad (4-11)$$

$\hat{h}_i$ 表示第*i*个多头的计算过程， Q_L, K_L 是如图中所示的局部词元，Att是注意力函数， W_1^i, W_2^i, W_3^i 是对应的权值矩阵， $\hat{L} \in \mathbb{R}^{m \times m}$ 是局部注意力权重图，因此多头局部识别注意力LSA定义为：

$$\text{LSA}(Q,K,V)=\text{Cat}(\hat{h}_1, \hat{h}_2, \dots, \hat{h}_m)W^O \quad (4-12)$$

其中，Cat是在通道维度上连接张量的操作，而 W^O 适用于投影特征的矩阵，权值矩阵都可以看成是全连接层操作，所提出的局部识别注意力在时间和空间复杂度上为 $O(n)$ ，并且基于局部感受野具有更多归纳偏差，以减少冗余信息的干扰，这表明了局部识别注意力在局部建模方面的效率和准确性。采用LRA进行强大的局部信息表示，并利用标准全局注意力构建了局部建模到全局搜索信息处理的机制。之后编码器 M_4 的输出可以通过点积注意力(LSA)获得，计算如下：

$$M_X = \text{LSA}(M_s^3, M_s^2, M_s^2) \quad (4-13)$$

$$M_4 = \text{Norm}(M_s^3 + M_X) \quad (4-14)$$

$$M_5 = \text{Norm}(\text{FFN}(M_4) + M_4) \quad (4-15)$$

其中 M_X 是中间变量，Norm是归一化操作， M_4 和 M_5 是编码器的两个输出。通过利用显著性特征作为先验信息，该模型更加强调通过对象显著性挖掘识别的显著区域。为了在减少外观信息丢失的同时进一步细化显著性信息，解码器被构造为一个标准的Transformer，它将模板中对象的特征作为输入。网络使用模版特征 M_s^1 作为Q，编码器 M_5 的输出作为K和V，形成了基于局部-全局的交叉注意力机制，提高了鲁棒性。整体计算机制与编码器相同。解码器 M_5 的输出可以通过以下方式导出：

$$M_Y = \text{LSA}(M_s, M_s, M_s^1) \quad (4-16)$$

$$M_6 = \text{Norm}(M_s + M_Y) \quad (4-17)$$

其中 M_Y 是中间变量。经过前馈神经网络(FFN)和正则化(Norm)之后, 输出 M_{out} 可以表示为:

$$M_{\text{out}} = \text{Norm}(\text{FFN}(M_6) + M_6) \quad (4-18)$$

通过使用模板特征作为解码过程的指导, 网络能够集中注意物体的显著特征, 并最小化先验信息中背景信息和不利信息。编码器和解码器的组合使得相似性信息能够经历从粗略到细致的逐步细化过程, 从而增强 LSGTT 跟踪器的性能与效率。

4.4 实验结果分析

4.4.1 实验设备及参数设置

本研究所采用的算法训练与测试在一套高端配置的硬件平台上实施, 该平台配备两颗 Intel(R) Gold 6342 系列 CPU, 运行频率为 2.80GHz, 并搭载了四片 NVIDIA GeForce RTX 3090 系列 GPU, 每片 GPU 具备 24GB 独立显存。算法所依托的软件环境基于 Ubuntu 22.04 操作系统, 采用 Pytorch2.0 框架, 编写语言为 Python3.10。其中, 模板图像和搜索图像的大小分别为 127×127 和 287×287 。提出的框架采用在 SiamRPN 中提出的修改后的 AlexNet, 该网络在 ImageNet 上进行了预训练, 然后用超参数作为初始值对预训练模型进行再次训练。我们使用裁剪之后大小为 511×511 像素的 COCO^[69]、YouTube^[70]、ImageNet-VID^[21]、GOT-10k^[47] 的数据集来训练我们的 LSGTT。我们总共训练了 200 个 epochs, 在最后 30 个 epochs 中, 对 AlexNet 的最后三层进行了微调, 每个 GPU 批次处理的样本数量设置为 2048, 训练过程中我们采用 SGD^[71] 优化器, 设置了初始学习率为 5×10^{-1} , 随着训练的进行, 学习率逐渐降至 1×10^{-3} , 并将权重衰减设置为 1×10^{-4} 。

4.4.2 定量分析

(1) DTB70 数据集结果分析

DTB70^[72] 包含了 70 个具有各种极为复杂场景的具有挑战性的无人机序列, 是一个典型的评估跟踪性能的基准。具体而言, 每个序列中存在诸多挑战, 包

括形变、平面旋转、尺度变化等。如图 4-8 所示，将我们的方法与 SiamFC++^[73]、SiamCAR^[25]、SiamRPN++^[18]、SiamBAN^[58]、ECO_HC^[13]、DSST^[11]、KCF^[10]和 SiamFC^[17]八种先进出色的跟踪器进行比较，为了保证提出模块的有效性、公平性和在无人机平台部署的可行性，在所有相关的基于 Siamese 网络的跟踪器中采用了相同的主干网络，即在 ImageNet 上预训练的 AlexNet。所提出的跟踪器在成功率和精确度上都表现优异，分别达到了 0.602 和 0.797。取得了最佳的成功率 (0.602)，超过了排名第三的 SiamBAN(0.557) 和排名第四的 SiamRPN++(0.532) 分别 4.5% 和 7%。在精度上，LSGTT 也表现优于其他方法，取得了最佳的精度 0.797，而排名第二的 SiamFC++ 和排名第三的 SiamBAN 分别获得了 0.790 和 0.770。

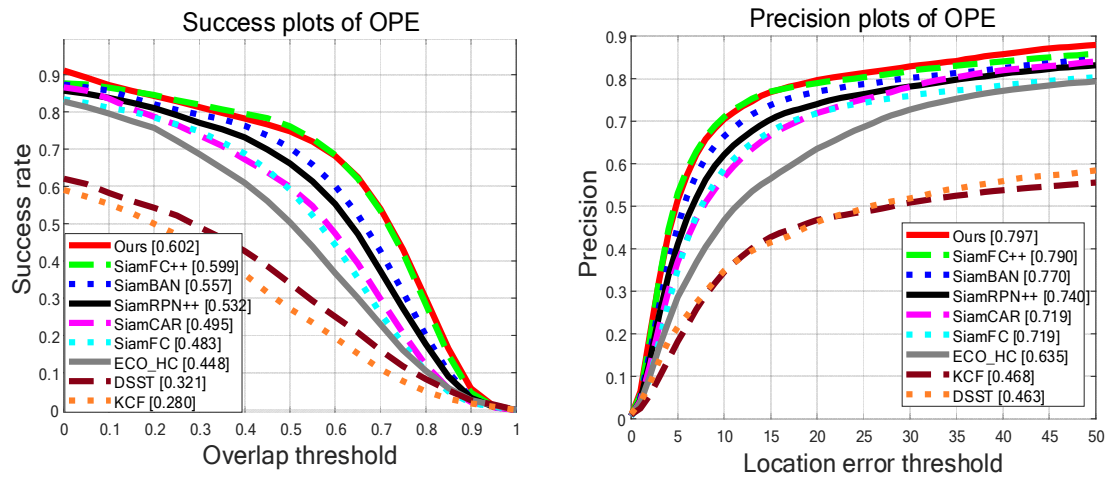


图 4-8 不同方法在 DTB70 上的成功率和精度图

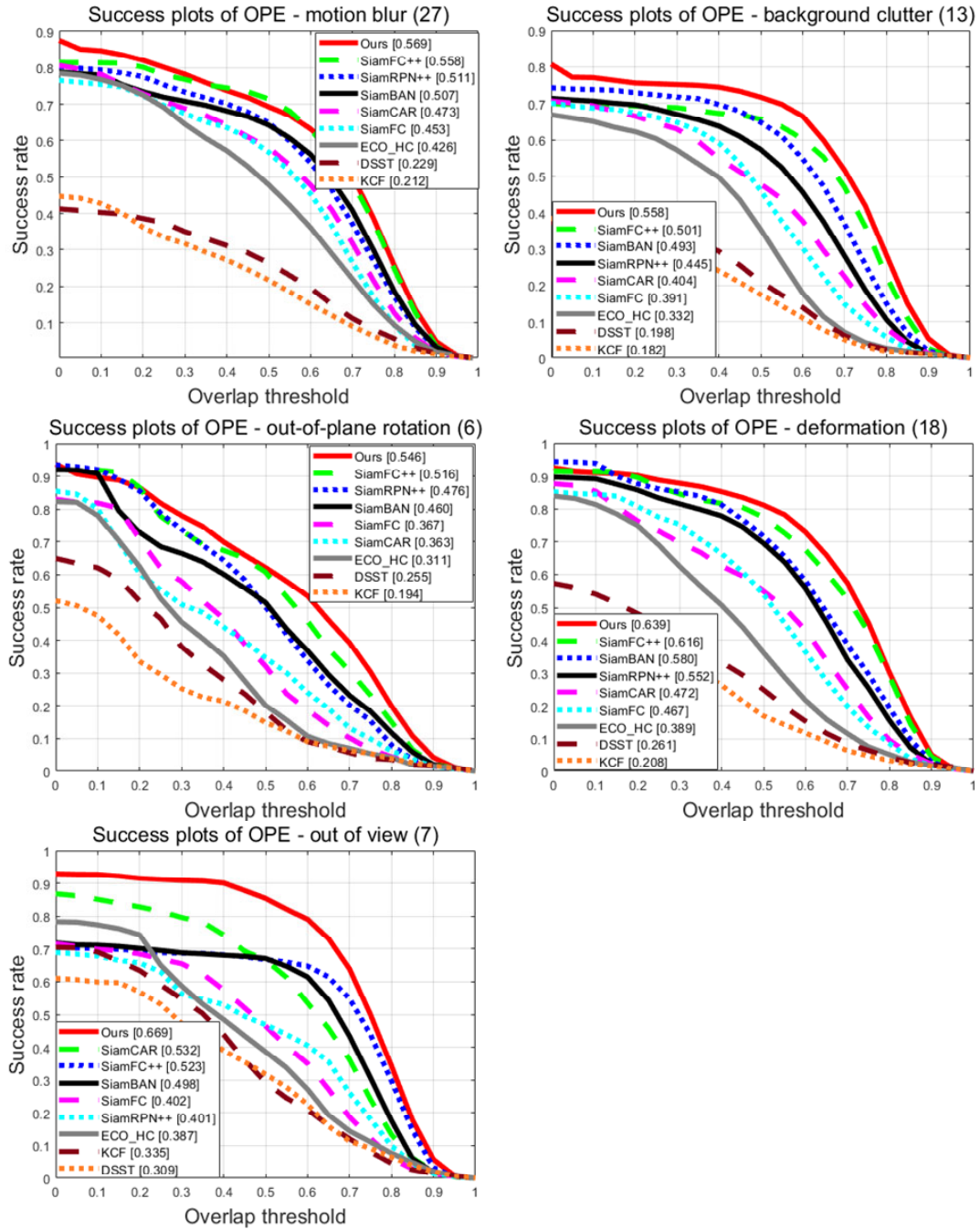


图 4-9 不同方法在 DTB70 上的不同属性成功率图

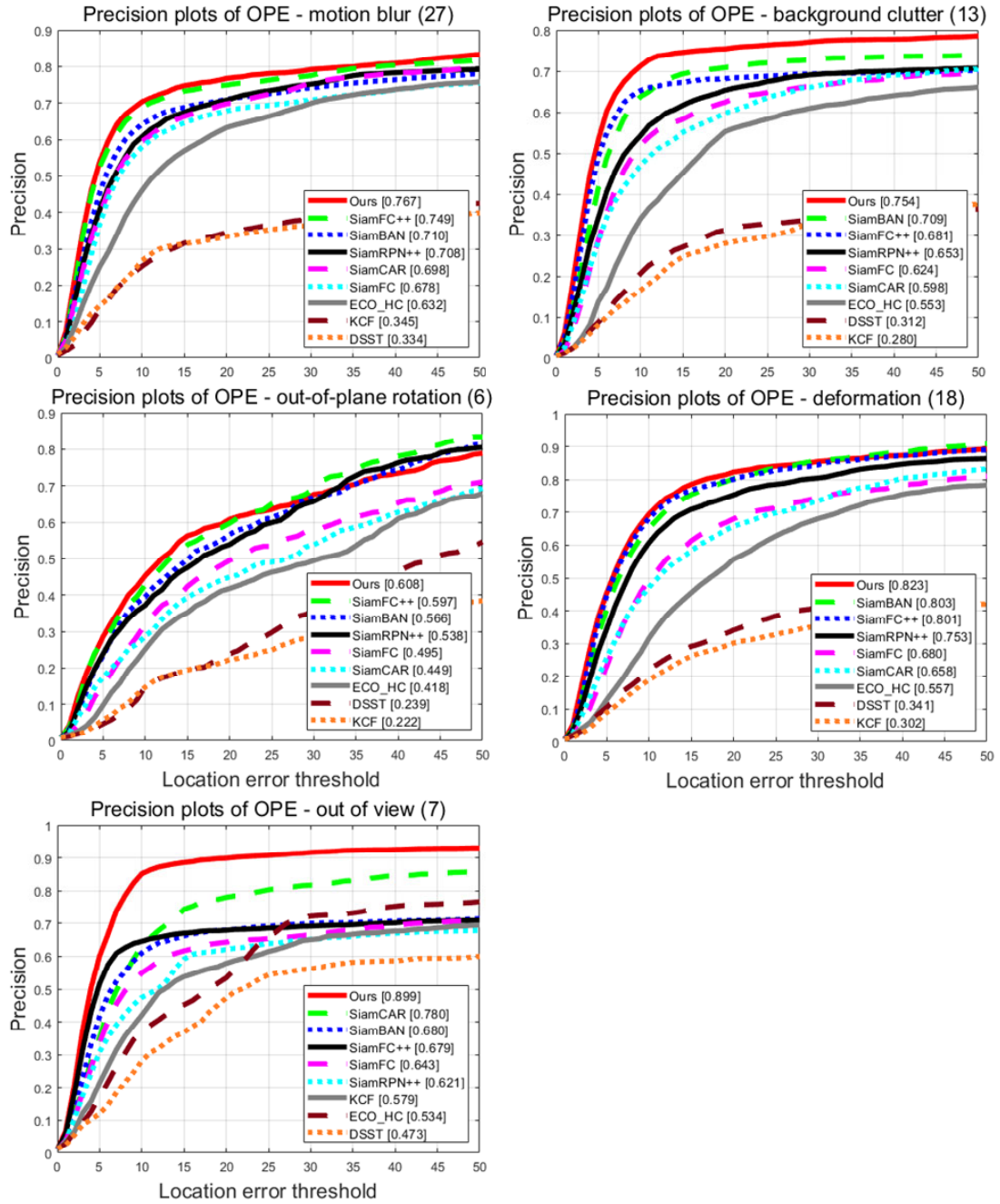


图 4-10 不同方法在 DTB70 上的不同属性精确度图

为了提供更加全面的稳健的分析，我们在 DTB70 数据集上对八名先进的跟踪器进行基于属性的定性比较。如图 4-9 和 4-10 所示，我们的跟踪器在目标丢失的场景中跟踪的精确度达到了 89.9%，成功率达到了 66.9%。相对于其他跟踪器，我们的跟踪器可以很好的处理跟踪无人机视角下的纵横比变化的对象，这是得益于我们的局部显著性 Transformer 模块将目标信息指导融合，强大的局部细节感知框架通过精细的局部建模将全局搜索性能提升，进一步跟踪目标的

显著性区域。

(2) UAV123 数据集结果分析

UAV123@30fps^[49]数据集是一个专为无人机视角目标跟踪而设计的评测数据集，该基准数据集用于评估低空航拍视角下的性能。含有各种无人机视角，在这个基准中，跟踪器面临着许多具有挑战性的航空跟踪场景。我们在 UAV123 数据集上将所提出的 LSGTT 算法与七种领先的跟踪方法进行了比较，其中包括 SiamCAR^[25]、SiamRPN++^[18]、DSST^[11]、KCF^[10]和 ECO_HC^[13]等。为了保证提出模块的有效性、公平性和在无人机平台部署的可行性，在所有相关的基于 Siamese 网络的跟踪器中采用了相同的主干网络，即在 ImageNet 上预训练的 AlexNet。如图 4-12 所示，红色实线是我们的结果，我们的 LSGTT 成绩排名处于领先地位，分别为 0.590 和 0.790，在成功率方面我们的跟踪器比第二名的 SiamBAN 和第三名 SiamRPN++领先了很多，性能分别提升了 4.7%和 6%，在精确度方面，我们的跟踪器也是大幅领先第二名 SiamCAR 和第三名 SiamBAN，性能分别提升了 4.5%和 5.1%。如图 4-11 所示，在无人机视角下的九个挑战场景中，我们的跟踪器与其他先进的跟踪算法相比较，在遮挡、尺度变化、低分辨率等环境中具有压倒性的优势，在快速运动的场景中比第二名 SiamBAN 的成功率高出了 7.3%，该项成功率跟踪结果表示，我们的跟踪器 LSGTT 对无人机视角下的遮挡、目标运动和尺度变化的跟踪任务具有一定的竞争力。如图 4-13 所示，在无人机视角下的十一个挑战场景中，我们的跟踪器在视野丢失、尺度变化和快速运动的场景中表现出色。尤其是在快速运动的场景中，我们的 LSGTT 精确度比第二名 SiamRPN++提升了 7.4%。该项精确度跟踪结果表示，我们的跟踪器 LSGTT 对无人机视角下的跟踪任务具有一定的竞争力，强大的局部细节感知框架通过精细的局部建模将全局搜索性能提高，通过对目标的多种变化进行有效感知，我们能够及时有效地找到目标。显然，我们的跟踪器在面对目标发生显著尺度变化的情况下，依然能够表现出优秀的处理能力。

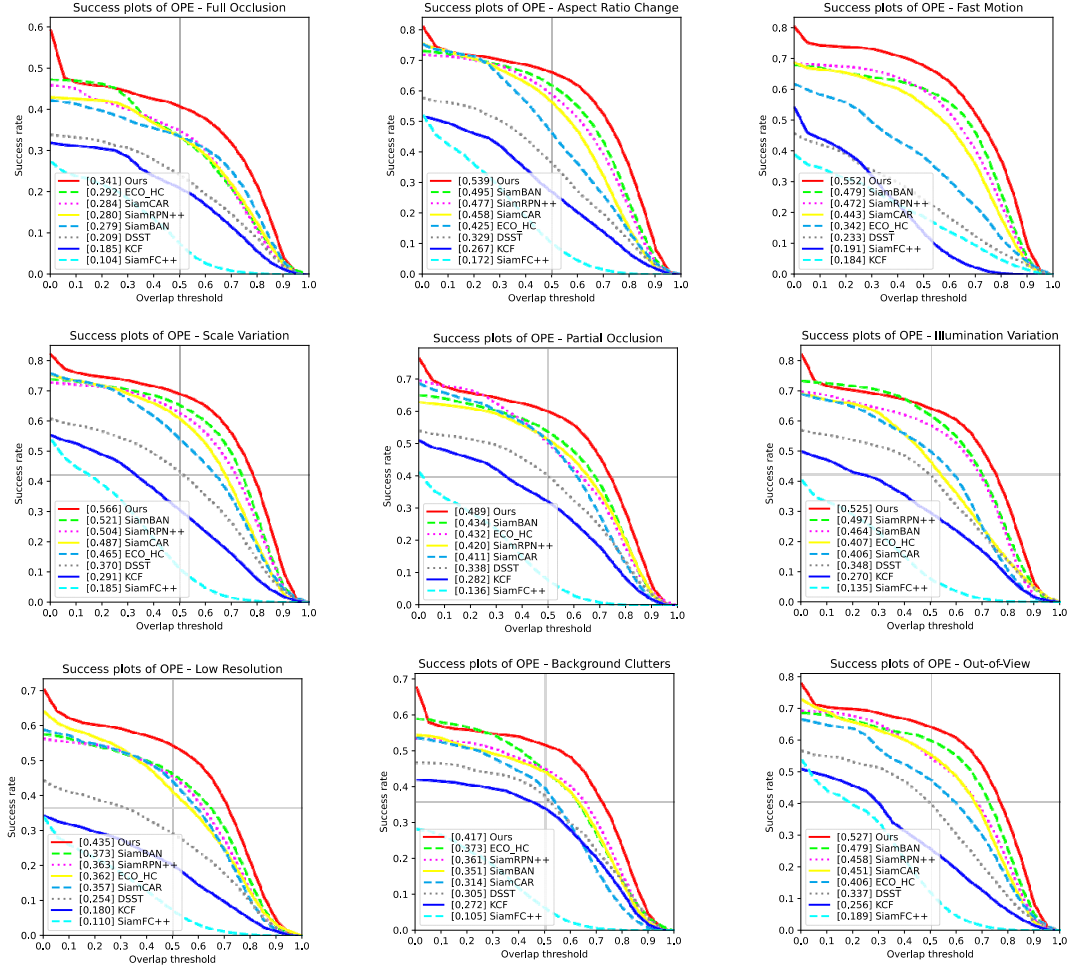


图 4-11 不同跟踪方法在 UAV123 上不同属性的成功率图

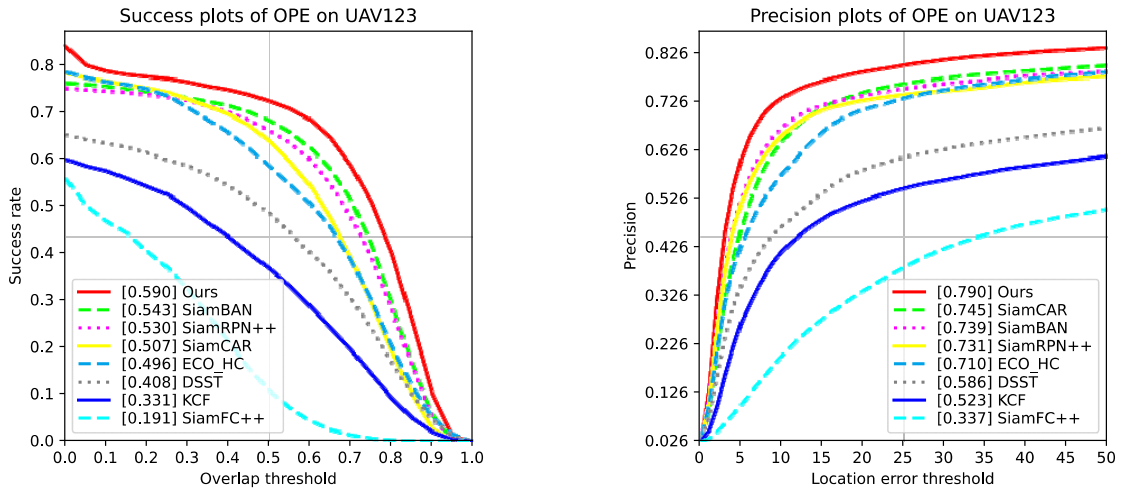


图 4-12 不同跟踪方法在 UAV123 上的成功率图和精度图

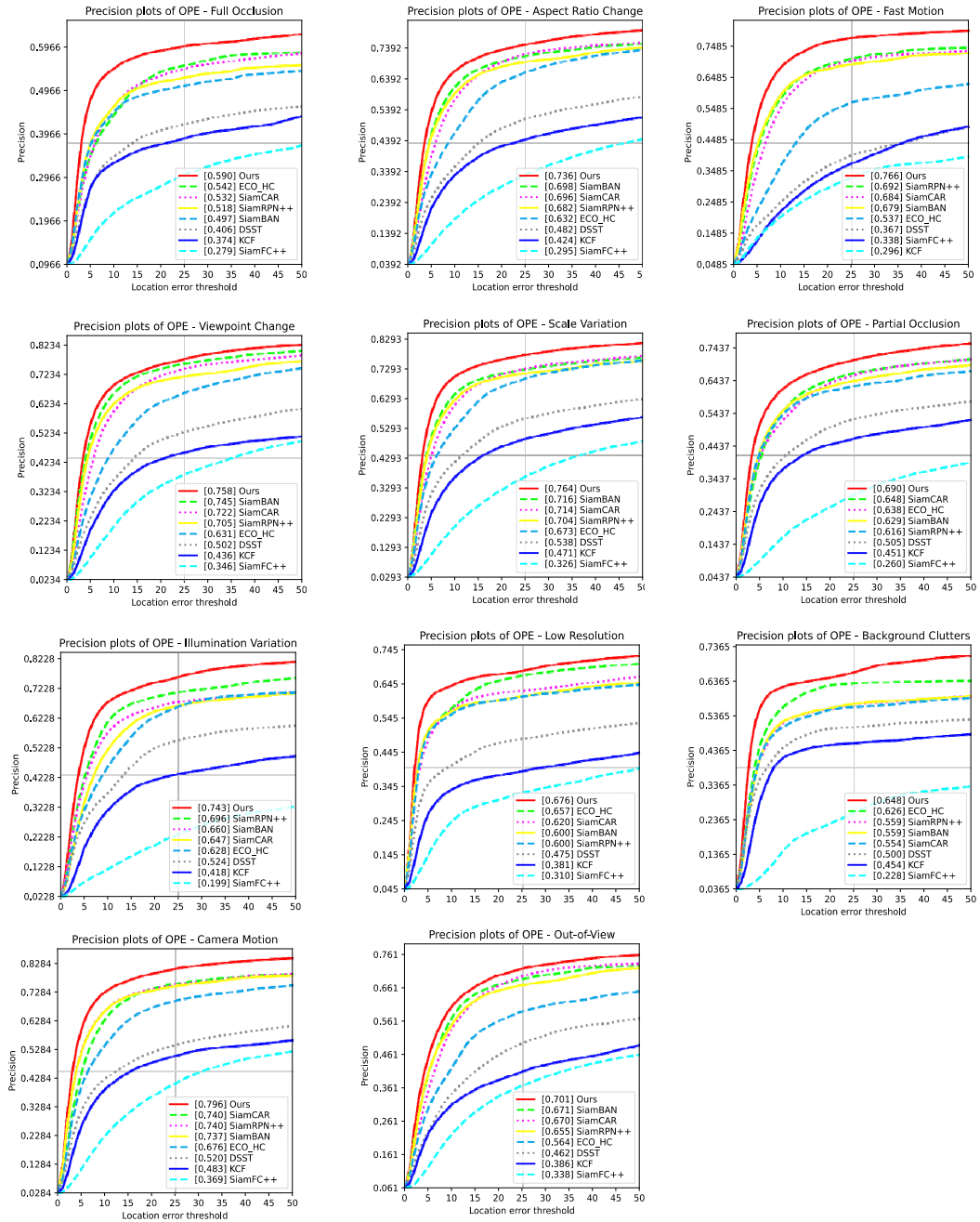


图 4-13 不同跟踪方法在 UAV123 上不同属性的精度图

(3) UAV10fps 数据集结果分析

这个基准中的序列是从UAV@30fps^[49]记录的序列中降采样得到的。因此，UAV123@10fps中物体运动的属性更加严峻和具有挑战性。

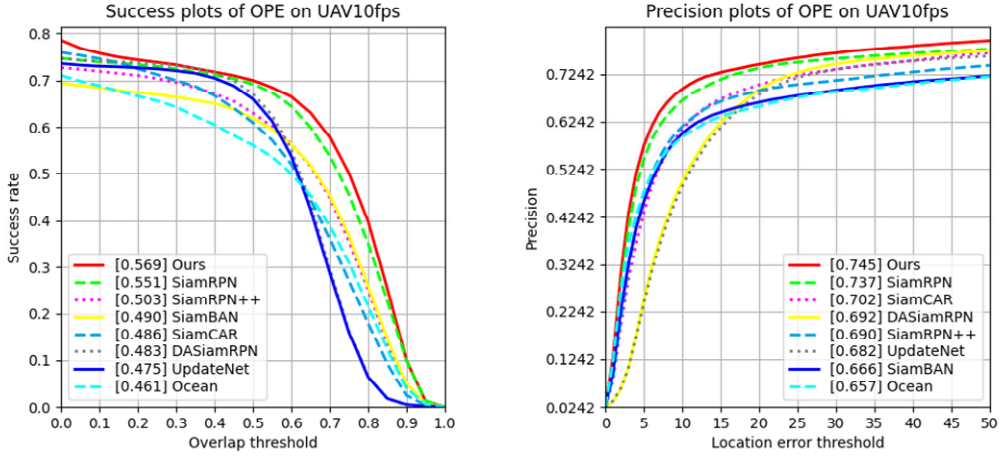


图 4-14 不同方法在 UAV10fps 上的成功率图和精度图

我们将我们的跟踪器 LSGTT 与 7 种最先进的跟踪器进行比较，包括 SiamCAR^[25]、SiamRPN++^[18]、SiamBAN^[58]、Ocean^[59]、SiamRPN^[19]、DASiamRPN^[74]和UpdateNet^[75]。为了保证提出模块的有效性、公平性和在无人机平台部署的可行性，在所有相关的基于Siamese网络的跟踪器中采用了相同的主干网络，即在ImageNet上预训练的AlexNet。如图4-14所示，其中红色实线表明的是我们的跟踪器，我们的 LSGTT成绩排名处于领先地位，分别为0.569和0.745，在成功率方面我们的跟踪器比第二名基于锚框生成式的跟踪器SiamRPN和第三名SiamRPN++领先了很多，性能分别提升了1.8%和5.6%，在精确度方面，我们的跟踪器也是大幅领先SiamCAR和SiamBAN，性能分别提升了4.3%和7.9%。以上基准测试的较高性能表明了LSGTT的准确性和鲁棒性，展示出了强大的局部建模和有效的全局搜索能力。卓越的跟踪性能结果证明了我们的跟踪器在处理各类视觉对象时具有出色的泛化性能。

(4) UAV20L 数据集结果分析

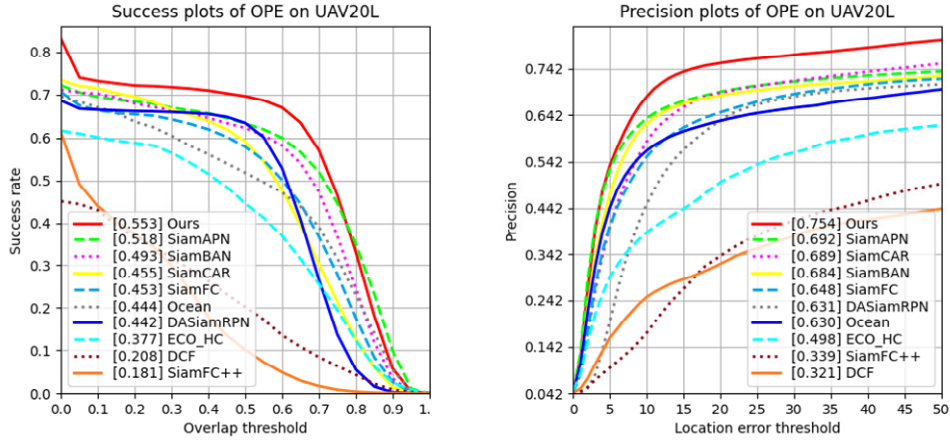


图 4-15 不同跟踪方法在 UAV20L 上的成功率图和精度图

UAV20L^[49]是 UAV123 数据集上的一个用于空中长期跟踪基准，包含 20 个序列，平均序列长度为 2934 帧。在本节中，我们将我们的跟踪器 LSGTT 与多种先进的跟踪器进行比较，包括 SiamFC++^[73]、SiamBAN^[58]、SiamCAR^[25]、SiamAPN^[76]、SiamRPN^[19]、DASiamRPN^[74]、SiamRPN++^[18]、ECO_HC^[13]、DCF^[77]和 Ocean^[59]。为了保证提出模块的有效性、公平性和在无人机平台部署的可行性，在所有相关的基于 Siamese 网络的跟踪器中采用了相同的主干网络，即在 ImageNet 上预训练的 AlexNet。如图 4-15 所示，其中红色实线表明的是我们跟踪器的结果，我们的 LSGTT 成绩排名处于领先地位，分别为 0.553 和 0.754，在成功率方面我们的跟踪器比在嵌入式平台上部署的无人机实时跟踪器 SiamAPN 和第三名 SiamBAN 领先了很多，性能分别提升了 3.5%和 6%，在精确度方面，我们的跟踪器也是大幅领先 SiamAPN 和 SiamCAR，性能分别提升了 6.2%和 6.5%。以上基准测试的较高性能表明了 LSGTT 的准确性和鲁棒性，展示出了强大的局部建模和有效的全局搜索能力。数据结果表明了 LSGTT 有能力并且可以更加鲁棒地跟踪目标。

5. 与深度骨干网络对比

表 4-1 不同骨干网络的跟踪器成功率与精度比较

Trackers	SiamRPN++	SiamDW_RPN	SiamGAT	SiamMask	SiamCAR	Ours
Backbone	MobileNetV2	ResNet-22	GooleNet	ResNet-50	ResNet-50	AlexNet
Success	0.593	0.453	0.583	0.571	0.597	0.602
Precision	0.785	0.709	0.752	0.769	0.801	0.797

为了验证 LSGTT 跟踪器中网络参数对成功率和精确度的影响，将本文提出的方法与其他使用更深层网络的跟踪器进行了对比，例如 SiamRPN++^[18]、SiamMask^[78]、SiamCAR^[25]、SiamDW_RPN^[23]和 SiamGAT^[79]。其中 SiamMask 和 SiamCAR 的骨干网络为 ResNet-50，SiamDW_RPN 为 22 层的残差网络，SiamRPN++使用的是 MobileNetV2^[80]的网络，SiamGAT^[79]的网络为 GooleNet^[81]，它们在 DBT70 上的成功率和精确度结果见表 4-1。可以发现我们的跟踪器仅使用了五层的 AlexNet 网络结构，但是在成功率方面表现最佳，为 0.602。精确度方面也仅仅比使用了 ResNet-50 网络结构的 SiamCAR 低了 0.4%，也处于领先的位置。我们的 LSGTT 跟踪器使用远少于其他对比跟踪器的参数量却达到了优异的跟踪水准，显然，所提出的跟踪器通过强大的局部感知维持了卓越的准确性和鲁棒性。这个对比说明了本文提出方法的出色效率和性能。

4.4.3 定性分析

为了更清晰地比较跟踪器在无人机视角场景下的跟踪性能，图 4-16 显示了在 DTB70 数据集上的定性分析结果，选择了五个具有挑战性的序列，从上到下分别是 (a)Boat3、(b)Car4、(c)Truck1、(d)Truck3、(e)Wakeboard5。可以看出在无人机视角下，在 boat3 序列中，随着轮船不断地移动，SiamCAR 会大范围的框选目标区域，ECO_HC 和 SiamRPN++会小范围的跟踪目标，只有我们的跟踪器 LSGTT 可以精准的跟踪目标轮廓。在 Car4 序列中，有一个广告牌遮挡，第 395 帧中 ECO_HC 会将其误认为是目标，在 507 帧后，过了广告牌，除了我们的跟踪器和 ECO_HC，其余全部跟丢，经过遮挡后无法重新跟踪到目标，我们 LSGTT 可以高效的应对遮挡挑战。在 Truck4 序列中，包含目标外形变化、尺度变化、旋转等挑战。货车在运动过程中，发生旋转，产生了较大的形变，SiamRPN++发生了跟踪错误的情况，其他方法都不能及时调整预测跟踪框，而我们的跟踪器可以鲁棒地跟住目标。在 Truck3 序列中，是一个高无人机视角，

目标在视野内占比很小，SiamBAN 和 SiamRPN++在 246 帧之后直接跟错目标，我们的 LSGTT 方法同样具备出色的小目标物体跟踪能力。在 Wakeboard5 序列中，一个男人在冲浪，伴随着小目标的形变和旋转，我们的 LSGTT 仍能保持超高的跟踪性能，得益于采用 LST 模块进行强大的局部信息表示，并利用标准全局注意力构建了局部建模到全局搜索信息处理的机制，我们的 LSGTT(红色)显著提高了跟踪精度。

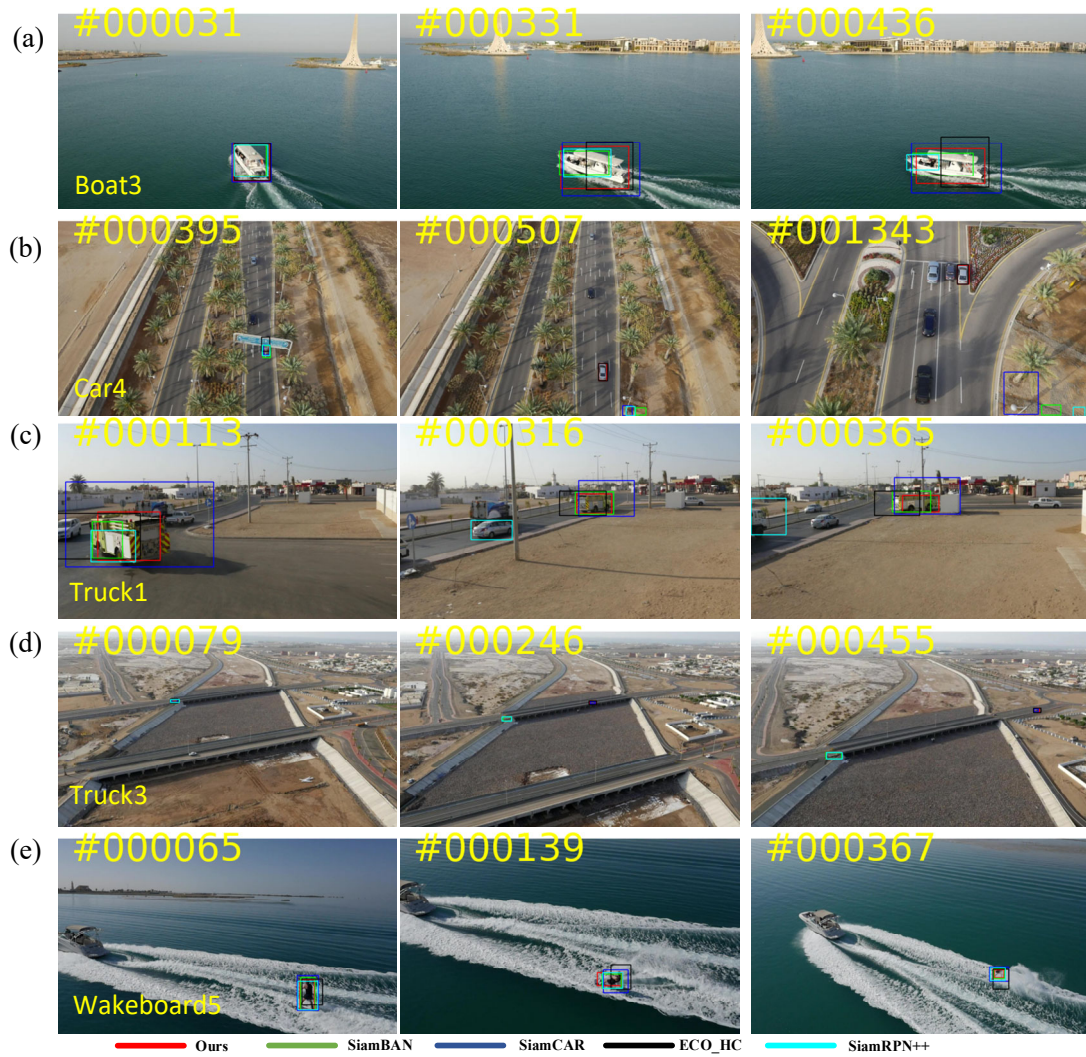


图 4-16 UAV123 上五个序列可视化分析

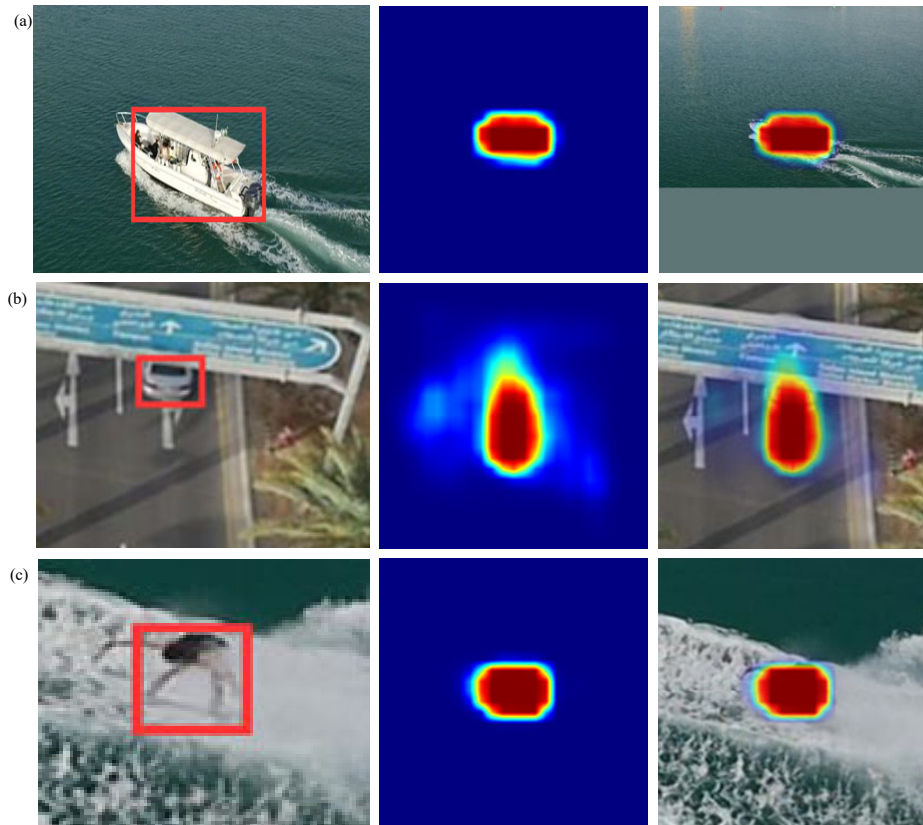


图 4-17 UAV123 上三个序列热力图可视化

图 4-17 显示了所提出的跟踪器的热力图可视化结果。所有的图片来自于基准的数据集 UAV123，我们可以观察到，LSGTT 可以对局部显著性区域进行建模，有效降低背景干扰，专注于跟踪对象的细节方面，应对跟踪目标形变和旋转遮挡等复杂场景，使跟踪结果更加准确。

4.4.4 消融实验

表 4-2 LSGTT 在 DTB70 上的消融实验

Base model	AlexNet	FSM	S-Refine	LST	AUC	Precision
✓	—	—	—	—	0.564	0.754
—	✓	✓	—	—	0.571	0.765
—	✓	✓	✓	—	0.583	0.778
—	✓	✓	✓	✓	0.602	0.797

表 4-2 显示了 LSGTT 在 DTB700 数据集上的成功率(以 AUC 衡量)和精度(以 Pre.表示)。为了验证提出的各项组成单元的有效性，我们构建了四组不同的变体结构，并进行了深入的实验。

Base model 考虑的是只使用 AlexNet 作为特征提取和使用全局注意力的典型单层 Transformers 模块加上分类回归头组成。成功率和精确度分别到达了

0.564 和 0.754。FSM 表示特征显著性提取模块，集成增强了表示目标特征的能力。通过探索相邻帧之间包含的潜在信息，可以提供足够可靠的目标信息。第二组消融实验添加了 FSM 模块成功率和精度分别提升了 0.7%和 1.1%。第三组消融实验添加了显著性精炼细化模块，根据对跟踪任务的贡献，区分出对跟踪有利的 patch，标记出每个词元的重要性，由表可见，Success 和 Precision 分别提升了 1.2%和 1.3%。第四组消融实验是完整版的跟踪流程，采用 LST 进行强大的局部信息表示，并利用标准全局注意力构建了局部建模到全局搜索信息处理的机制，以减少冗余信息的干扰，实验结果表明了局部识别注意力在局部建模方面的效率和准确性，成功率和精确度达到了 0.602 和 0.797。

4.5 本章小结

本章主要介绍了局部显著性Transformer的网络结构，分别介绍了显著性挖掘模块、显著性细化精炼模块、局部显著性Transformer模块。介绍了实验平台与实验过程中的参数设置，在UAV123、UAV10fps、UAV20L、DTB70数据集与主流的八种跟踪方法进行了细致的比较，在各种具有挑战性的场景中进行了定量和定性的分析，将本章中改进的各部分模块进行了消融实验，证明验证各模块的作用和有效性并进行了详细的阐述。通过挖掘出的显著性信息来确定每个词元的重要性，可以区分前景和背景。这项工作的主要贡献如下：

1. 设计了一个目标显著性挖掘模块，通过空间和通道融合的有机结合提取显著性信息来区分前景和背景。
2. 设计了一个显著性细化精炼模块，基于对前景和背景的不同关注，并不是所有的 patch 都对跟踪任务具有相同的信息价值。根据不同 patch 中得分的高低，区分显著性信息对跟踪任务的贡献。
3. 设计了一种新的局部显著性 Transformer 结构，在细化精炼的显著性信息的基础上，通过多层次地融合指导动态词元，在标准的 Transformer 结构上进行了轻量化，实现了一种新颖的显著性融合 Transformer，通过交叉注意机制来细化显著性信息并最大限度地减少外观信息的损失。
4. 对四个权威无人机跟踪基准进行的综合评估证明了LSGTT在解决外观变化方面的强大性能，跟踪结果进一步验证了所提出方法的效率和有效性。

第 5 章 无人机硬件平台实时跟踪验证

5.1 引言

无人机实现实时目标跟踪具有显著的应用价值，为了确保研究成果切实可行，本章采用小巧、低参数量的网络结构，将经过优化的基于 Transformer 框架的跟踪算法移植到我们搭建的无人机硬件系统上。相较于同等轻量级的对比跟踪方法，我们所设计的跟踪器在真实环境中不仅表现出更强的鲁棒性，同时在运行速度上也展现出优越性，有力验证了其卓越的性能优势和实际应用潜能。

5.2 实验平台搭建

我们在真实环境中对第四章所阐述的 LSGTT 目标跟踪方法进行了实时跟踪，该方法搭载于阿木实验室研发的一款名为 P600 旗舰版的无人机硬件平台上，如图 5-1 所示，此平台属准工业级别，特别适应于无人机行业的应用开发及户外环境下各类无人机算法的实时验证工作。为了获取实时图像数据，我们在此无人机上额外安装了专用的图像采集吊舱装置。

鉴于四旋翼无人机兼具优越的机动性能与高度的可扩展性，且结合实验室当前资源情况，选择使用四旋翼无人机来进行图像数据捕获。实验所用无人机配备有卓越的成像系统软硬件，其中包含先进的磁场定向控制系统，保证了电机扭矩输出平稳、噪音低、高效运行，并展现出快速的动态响应能力。无人机整体硬件品质优秀，确保了飞行稳定性，以及长时间续航（空载状态下达 70 分钟），并且搭载 2.4G WiFi 通信模块，提供多种接口便于连接不同外设，增强了整体的拓展性。无人机主体结构采用新型铝合金材料制作而成，内嵌低噪散热风扇，机体尺寸为 94mm×59mm×37mm，重量仅 213 克，具备小巧轻便的特点。操作方面，我们采用 H16 型一体化遥控器，集成了数传、图传和操控功能（工作频段在 2.4-2.483GHz 范围内，数传最远距离达 30 公里，支持 1080P 高清图传），能够在屏幕上实时监控无人机的状态信息、摄像头捕捉的影像画面，并通过摇杆实时操控无人机，实现了无人机飞行的直观可视化管理。

无人机搭载有三轴防抖云台和摄像模组，该模组采用 OVOS12D40 彩色图

像传感器，这是一款高效的 CMOS 图像传感器，能够以最高每秒 60 帧的速度输出一千万像素的高质量图像信号，支持最高一千万像素的高清照片拍摄，以及最高 4K 分辨率 30fps 乃至 60fps 的视频录制。

针对提出的跟踪器技术，我们将其部署在一个通用的无人机平台上，并搭配了典型嵌入式处理器——NVIDIA Jetson AGX Xavier。Jetson AGX Xavier 是由 NVIDIA 公司专门为边缘计算和嵌入式应用场景设计的一款高性能人工智能模块，特别适合对计算强度和能耗效率有极高要求的领域，如机器人技术、无人机、自动驾驶汽车以及其他智能化边缘设备。这款产品内部集成了一系列强劲的硬件加速器，包括用于深度学习加速的 Tensor Cores，能够支撑复杂机器学习、深度神经网络和计算机视觉任务的执行。Jetson AGX Xavier 搭载基于 NVIDIA Volta 架构的 GPU，拥有 512 个 CUDA 核心和 64 个 Tensor 核心，赋予了其强大的计算和深度学习推理能力。

在本章节中，无人机上的 Jetson AGX Xavier 搭载的软件环境基于 Ubuntu 18.04 操作系统，算法开发采用 Python 3.6 语言，并构建于 Pytorch 10.2 框架之上。



图 5-1 无人机硬件平台

5.3 真实世界实时跟踪分析

将第四章的跟踪算法部署在无人机平台上，实现运动目标的实时跟踪。如下图所示，我们测试了在复杂场景中面临的多种挑战，例如快速运动、尺度变化、平面旋转、相机运动、遮挡和视野丢失，如图5-2所示：

对图(a)所展示的场景进行深入分析，我们可以看到，该场景包含了一系列颇具挑战性的目标跟踪问题。首先，目标物体不仅在二维空间内进行了显著的位置移动，它经历了一场从近景向远景的动态过渡，这一过程伴随着深度层面的大幅变化，使得目标物体在三维空间中的相对距离发生了明显变化。随着目标的远近位移，其在平面上的呈现也出现了明显的尺寸调整，即宽高比的变化。当目标从近距离快速奔跑至远距离时，目标物体在图像上的投影面积急剧缩小，从原本在画面中占据较大比例的清晰实体，演变为几乎难以察觉的小型像素块，这无疑增加了识别和跟踪的难度。面对目标物体从大到小、从清晰到模糊的剧烈外观转变，我们的跟踪系统并未失去跟踪精度。相反，它成功地克服了因目标尺度和形态突变带来的困难，始终保持对目标位置的准确感知，并能在每一帧预测中精准锁定目标所在区域，充分验证了该跟踪器在实际应用场景下的高效性能和可靠性。

在深入剖析图(b)所展示的情景时，我们可以观察到一个充满挑战的画面，画面中的目标人物在足球场上正向右侧奔跑并控球，由于其在横向空间上的快速移动与无人机摄像角度的变化，导致目标物体在图像中出现了明显的形状变形。随着目标不断接近无人机摄像范围的边界，最终完全脱离了视线区域，这给跟踪系统带来了显著的难度，即遭遇了遮挡和目标丢失的问题，此刻跟踪器的表现会呈现不确定性，体现在预测的目标框在目标消失前后的区域出现不稳定跳动。不过，在经历了一段时间之后，目标人物携球重新进入足球场的可视区域内，再次被无人机摄像头捕捉到。在这时刻，跟踪系统展现出了优异的快速重定位能力，它能立即精准地确定目标的新位置，并无缝衔接之前的跟踪动作，继续对运动员进行有效跟踪。总结来说，图(b)生动展示了跟踪系统如何应对实际应用场景中的复杂问题，如目标物体因运动造成部分甚至全部遮挡以及短时间内重新出现的情况，它能够在目标失而复得后迅速调整并维持高精度的跟踪效果，这对于包括体育赛事监控在内的各类应用具有重要意义。

在对图(c)所示的具体视觉跟踪场景进行深入解读时，我们可以观察到：首先，在场景的初始阶段，所关注的跟踪目标在画面中处于一个下蹲的姿态，在无人机视野中的画面仅占用了极小的像素范围，这无疑给跟踪技术带来了识别和定位上的挑战。然而，即使在这种不利于识别的情况下，跟踪系统仍在尝试维持对该目标的锁定。紧接着，从画面左侧方向缓缓走入一个与跟踪目标在外观上具有较高相似度的干扰对象。随着这个相似目标不断靠近直至完全遮挡住了原本正在被跟踪的目标，这一阶段对于任何视觉跟踪算法来说都是极端考验，因为它需要在动态变化和遮挡状态下保持对目标的连续感知与识别。然而，我们的跟踪器表现出了一种强大的适应能力和抗干扰特性。在目标被遮挡期间，即便目标本身的直接视觉信息几乎完全消失，跟踪器并没有因此而失去目标。相反，它巧妙地运用了局部显著性分析方法，通过对画面中那些显著的局部特征进行跟踪，成功地识别出了在相似目标背后的原始跟踪目标的真实位置。待相似目标移动离开原地后，真正的跟踪目标得以重新出现在视野范围内。此刻，跟踪器能够在第一时间准确判断出目标重现的信息，并立即恢复对目标的稳定跟踪，整个过程中并未出现因混淆或滞后而导致的跟踪失误现象。

在图(d)所描绘的情景中，我们观察到运动目标在广阔的操场上进行着一系列快速且复杂的动作，从画面左侧开始全速奔跑，继而变为上下起伏的蹲起运动。与此同时，无人机搭载的云台相机积极地同步调整拍摄角度及焦距，始终致力于跟随并捕捉目标的动作轨迹。在此过程中，跟踪系统面临着多重艰巨挑战，包括来自云台相机自身的运动补偿问题、目标物体在高速运动中发生的显著尺度变化，以及目标行为模式的快速变换。尽管存在这些难点，跟踪器依然凭借其卓越的性能和稳定性，有效地应对了所有挑战。它不仅能够实时适应相机运动所带来的影响，还能精确估算并跟踪目标在不同速度和尺度变换条件下的位置，从而确保了无论目标如何快速变换姿势或方位，都能被持续且精准地锁定在镜头之中。简而言之，即使面对如相机机动、目标尺寸变化及瞬时高速运动等多种复杂情况，我们提出的跟踪系统依然能展现出令人称赞的跟踪能力，始终保持对目标的稳定跟踪。

LSGTT在实际应用中的复杂环境里展现出了显著的效果，并在实验测试阶段保持了大约每秒29帧的运行速率。该算法设计了局部-全局相结合的注意力机

制，这一设计不仅有效降低了计算资源的需求，同时在很大程度上抑制了无关干扰信息的影响。具体来说，局部显著性注意力模块增强了Transformer架构对于目标区域局部特征的精细捕获和表达能力，这一系列在复杂场景下取得的良好表现数据有力证明了LSGTT跟踪器具备出色的鲁棒性和竞争优势。

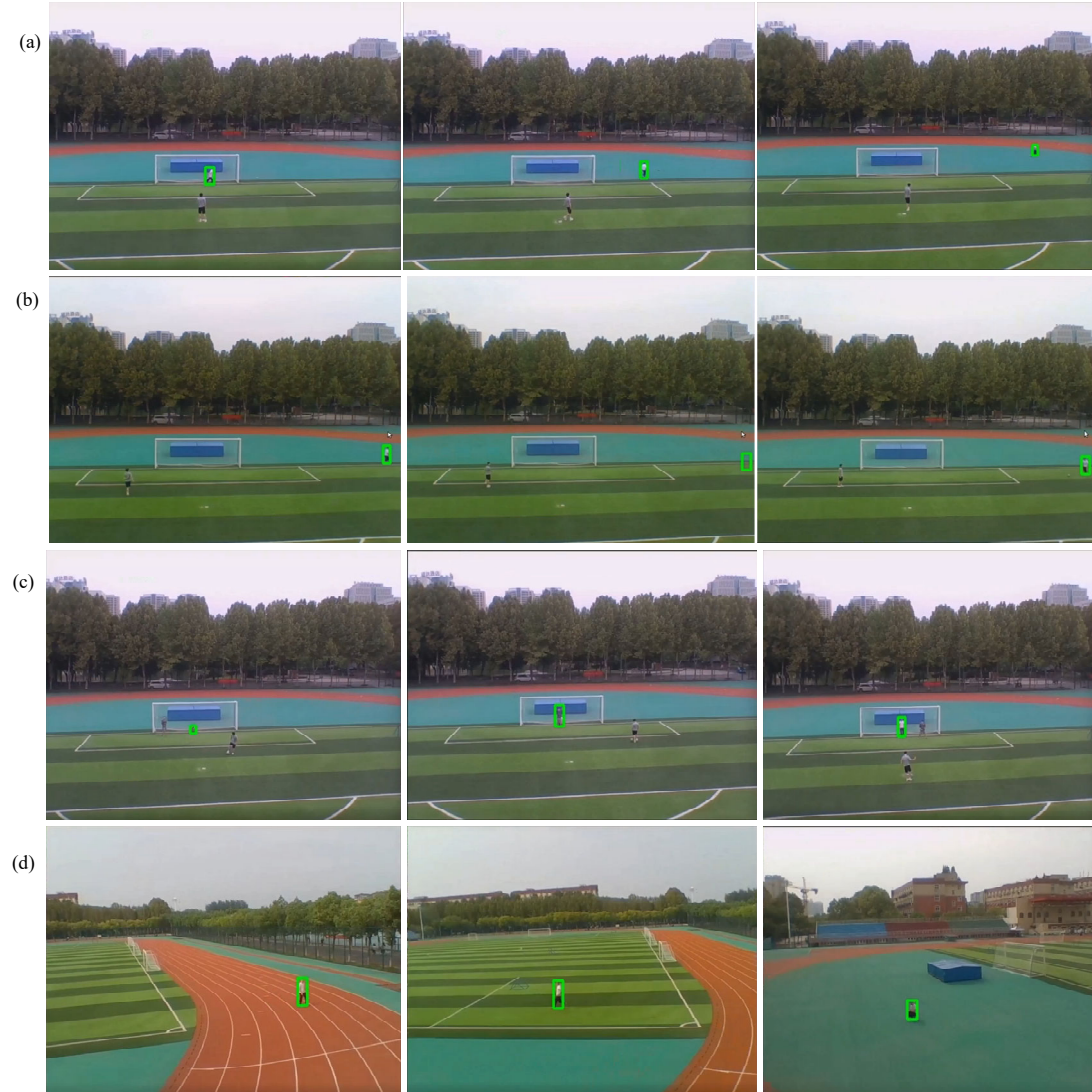


图 5-2 真实世界实时跟踪目标

如图5-3所示，所有的跟踪器结果都是在同一个无人机平台上使用NVIDIA Jetson AGX Xavier获得的，这里显示了一些对比跟踪器的跟踪结果。在DTB70权威无人机跟踪基准下的性能比较。如图所示：LSGTT与轻量化骨干网络的SiamFC++^[73]、SiamBAN^[58]、SiamCAR^[25]、SiamRPN++^[18]等跟踪器相比，我们的跟踪器在性能评估中脱颖而出，不仅实现了最高的成功率，而且还具备最快的运行速度，无论是性能还是效率均位列榜首。相比之下，ECO_HC虽然在速

度上占有优势，但其成功率却垫底。至于SiamFC++，尽管其成功率与我们的跟踪器相当接近，但在速度上却明显落后，运行速度仅为我们的一半，难以满足实时跟踪的需求。得益于高效的局部-全局显著性机制，我们提出的跟踪器在各种基准测试以及实际环境应用中都展现了卓越的性能和极具竞争力的运行速度。

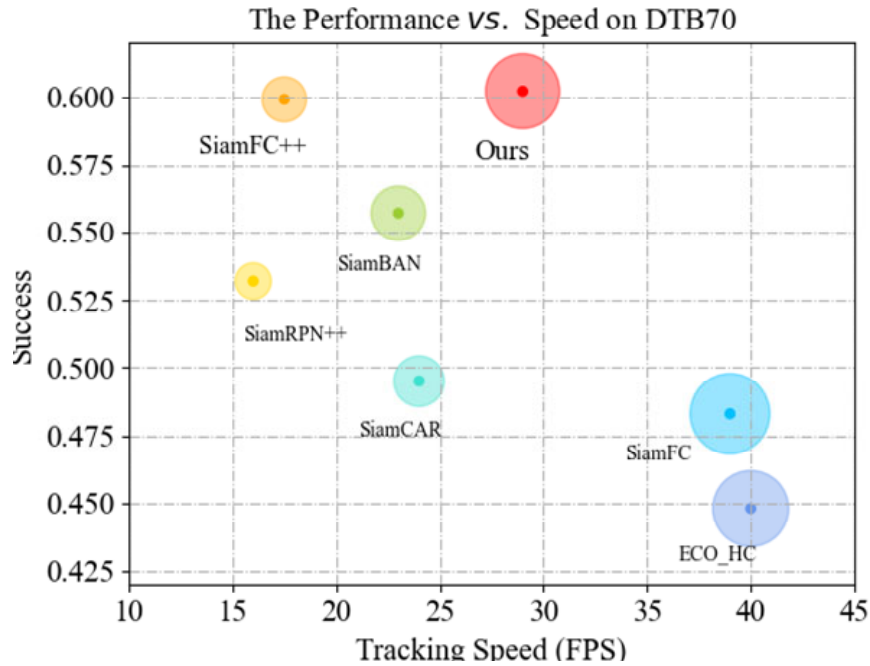


图 5-3 轻量化硬件部署实时跟踪的速度

5.4 本章小结

在本章中，我们将第四章介绍的基于局部显著性Transformer的跟踪算法移植到了无人机硬件设备上，并在现实世界的多元化复杂环境中对运动目标进行了实时追踪。结果显示，无论面对何种复杂场景，该跟踪器均表现出卓越的追踪效果。在同一硬件配置条件下进行的速度与成功率测试中，我们的LSGTT在两项关键性能指标上均占据了首位，证实了其在同类技术中的领先地位。

第 6 章 总结与展望

6.1 总结

目标跟踪是计算机视觉中一项至关重要的技术，在各方面都发挥着至关重要的作用。尽管随着机器学习技术的深度开发，目标跟踪技术已取得显著跃升，但在应对诸如目标物体被遮挡、背景环境复杂多变以及目标短暂脱离视野等实际问题时，实现对复杂场景高效且稳定的目标跟踪依然存在一定的局限性。因此，设计出成功率更高的跟踪方法具有显著的实际价值。本文针对上述问题，主要工作如下：

（1）针对现有的孪生网络在处理目标形变、平面旋转、遮挡等复杂场景中精度和鲁棒性不足的问题，本论文基于 SiamCAR 算法，设计了一种新颖的跟踪策略 MSSiamCAR，使跟踪器在跟踪过程中，更多地关注目标自身情况，提出局部跟踪策略，跟踪目标的显著区域，这些区域对跟踪器具有明显的判别性。其中显著性捕获可以获取目标特征中对跟踪任务具有判别性的显著点，显著性交互利用图注意力在显著性区域之间建立有效联系，通过学习目标与搜索区域间的关联性来适应目标不同的尺度变化。实验结果表明，相比于 SiamCAR 算法，MSSiamCAR 在 VOT2018 数据集上期望平均重叠率增长了 2.5%，在 GOT-10k 数据集的 AO 、 $SR_{0.5}$ 和 $SR_{0.75}$ 这三个指标中，分别提高了 4.4%、4.3%和 3.1%，跟踪性能优秀，在复杂场景中具有良好的鲁棒性。

（2）为了进一步提升跟踪器在处理无人机视角下的相似小目标形变、遮挡等场景的跟踪性能，本论文设计了一种基于局部显著性 Transformer 的无人机目标跟踪方法。设计了一个显著性细化精炼模块，基于对前景和背景的不同关注，并不是所有的 patch 都对跟踪任务具有相同的信息价值。根据不同 patch 得分的高低，区分显著性信息对跟踪任务的贡献。设计了一种新的局部显著性 Transformer 结构，将细化精炼的显著性信息，通过多层次地融合，指导动态词元，在标准的 Transformer 结构上进行了轻量化，实现了一种新颖的局部显著性 Transformer 架构，通过交叉注意机制来细化显著性信息并最大限度地减少外观

信息的损失。提出的跟踪器在四个权威无人机跟踪基准(DTB70、UAV123、UAV10fps、UAV20L)上进行了综合评估,实验结果显示,提出的跟踪器在无人机视角下的挑战中表现出卓越的性能,能够妥善应对并有效解决多种复杂情况下的目标跟踪难题。

(3) 为了在真实世界中对处于复杂环境中的目标进行实时跟踪,在论文中搭建了无人机实验平台,将轻量化后的目标跟踪算法部署在搭建的硬件平台中,与同级别轻量化跟踪网络相比,网络模型参数量更少,经过同等硬件条件下的大量对比实验发现,在真实世界跟踪中,本论文提出的跟踪器在速度和精度方面有了显著的提升。由 DTB70 数据集的评估结果可知,本论文提出的跟踪器速度达到了 29 帧,与轻量化的 SaimFC++相比,成功率都达到了 0.6 以上,但是速度却快了一倍。与轻量化的 ECO_HC 相比,虽然速度慢了几帧,但是也达到了实时跟踪的条件并且成功率上升了 1.57%。本论文提出的跟踪器在真实世界跟踪中展现了卓越的性能和极具竞争力的运行速度。

6.2 展望

本文所提出的跟踪器尽管在效能上有一定的跃升,但依然面临一些待解决的问题:

(1) 虽然 MSSiamCAR 跟踪器中提到的显著性图注意力机制对跟踪效果有所增强,但其在实时跟踪的速度方面仍有提升空间,跟踪器在应对低分辨率以及微小目标时表现欠佳。我们计划在未来的工作中深化研究,以提升该跟踪器在上述复杂场景下的鲁棒性。

(2) 针对在无人机平台上部署的应用场景,跟踪器的参数优化及其运行效率的提升显得尤为重要。这涉及到对跟踪网络架构的精简化改造,包括跟踪网络模型的轻量化设计以及剪枝技术的应用,此方面的深入探索将成为我们后续研究的一项关键议题。

参考文献

- [1] 黄治翔, 张艺骞. 基于深度学习的目标检测综述[J]. 科技资讯, 2023, 21(24): 13–16.
- [2] 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.
- [3] 肖华欣. 图像/视频目标分割技术研究[D]. 2021.
- [4] 李清格, 杨小冈, 卢瑞涛, 等. 计算机视觉中的 Transformer 发展综述[J]. 小型微型计算机系统, 2023, 44(4): 850–861.
- [5] 高帅. 基于行人轨迹预测的局部路径规划与跟踪方法研究[D]. 2022.
- [6] 陈商盈, 倪受东, 童林. 改进 YOLOX 的自动驾驶场景目标检测算法[J]. 计算机工程与应用, : 1–11.
- [7] 高宝林. 基于深度学习的机器人视觉引导抓取策略研究[D]. 2023.
- [8] 张恒. 无人机平台运动目标检测与跟踪及其视觉辅助着陆系统研究[D]. 2010.
- [9] Henriques JF, Caseiro R, Martins P, et al. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels[G]//FITZGIBBON A, LAZEBNIK S, PERONA P, et al. Computer Vision – ECCV 2012. 2012, 7575: 702–715.
- [10] Henriques JF, Caseiro R, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583–596.
- [11] Danelljan M, Häger G, Shahbaz Khan F, et al. Accurate Scale Estimation for Robust Visual Tracking[C]//Proceedings of the British Machine Vision Conference 2014. 2014: 65.1–65.11.
- [12] Danelljan M, Robinson A, Khan FS, et al. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking[G]//2016, 9909: 472–488.
- [13] Danelljan M, Bhat G, Shahbaz Khan F, et al. Eco: Efficient convolution operators for tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 6638–6646.
- [14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

- [15]Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
- [16]Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1420-1429.
- [17]Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking[C]//Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14. Springer International Publishing, 2016: 850-865.
- [18]Li B, Wu W, Wang Q, et al. Siamrpn++: Evolution of siamese visual tracking with very deep networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4282-4291.
- [19]Li B, Yan J, Wu W, et al. High Performance Visual Tracking with Siamese Region Proposal Network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8971–8980.
- [20]Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
- [21]Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84–90.
- [22]He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [23]Zhang Z, Peng H. Deeper and wider siamese networks for real-time visual tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4591-4600.
- [24]Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and

- pattern recognition. 2016: 2818-2826.
- [25]Guo D, Wang J, Cui Y, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6269-6277.
- [26]Chen X, Yan B, Zhu J, et al. Transformer tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8126-8135.
- [27]Yan B, Peng H, Fu J, et al. Learning Spatio-Temporal Transformer for Visual Tracking[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). 2021: 10428–10437.
- [28]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [29]Ye B, Chang H, Ma B, et al. Joint feature learning and relation modeling for tracking: A one-stream framework[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 341-357.
- [30]Zhu J, Lai S, Chen X, et al. Visual prompt multi-modal tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 9516-9526.
- [31]Chopra S, Hadsell R, LeCun Y. Learning a Similarity Metric Discriminatively, with Application to Face Verification[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005, 1: 539–546.
- [32]Danielsson P E. Euclidean distance mapping[J]. Computer Graphics and image processing, 1980, 14(3): 227-248.
- [33]Malkauthekar MD. Analysis of euclidean distance and manhattan distance measure in face recognition[C]//Third International Conference on Computational Intelligence and Information Technology (CIIT 2013). 2013: 503–507.
- [34]Xia P, Zhang L, Li F. Learning similarity with cosine similarity ensemble[J]. Information Sciences, 2015, 307: 39–52.
- [35]Gao Y, Beijbom O, Zhang N, et al. Compact bilinear pooling[C]//Proceedings of

- the IEEE conference on computer vision and pattern recognition. 2016: 317-326.
- [36]LeCun Y, Boser B, Denker JS, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. Neural Computation, 1989, 1(4): 541–551.
- [37]De Boer P-T, Kroese DP, Mannor S, et al. A Tutorial on the Cross-Entropy Method[J]. Annals of Operations Research, 2005, 134(1): 19–67.
- [38]Chai T, Draxler R R. Arguments against avoiding RMSE in the literature[J]. Geoscientific model, 2014.
- [39]Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
- [40]Wen Y, Zhang K, Li Z, et al. A Discriminative Feature Learning Approach for Deep Face Recognition[G]//LEIBE B, MATAS J, SEBE N, et al.Computer Vision – ECCV 2016. 2016, 9911: 499–515.
- [41]Yohanandan S, Song A, Dyer AG, et al. Saliency Preservation in Low-Resolution Grayscale Images[G]//FERRARI V, HEBERT M, SMINCHISESCU C, et al.Computer Vision – ECCV 2018. 2018, 11210: 237–254.
- [42]Yohanandan S, Song A, Dyer A G, et al. Fast efficient object detection using selective attention[J]. arXiv e-prints, 2020: arXiv: 1811.07502.
- [43]Borji A, Cheng M M, Hou Q, et al. Salient object detection: A survey[J]. Computational visual media, 2019, 5: 117-150.
- [44]Wang W, Lai Q, Fu H, et al. Salient Object Detection in the Deep Learning Era: An In-Depth Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 3239–3259.
- [45]Wu Y, Lim J, Yang M-H. Online Object Tracking: A Benchmark[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2411–2418.
- [46]Kristan M, Leonardis A, Matas J, et al. The Sixth Visual Object Tracking VOT2018 Challenge Results[G]//LEAL-TAIXÉ L, ROTH S.Computer Vision – ECCV 2018 Workshops. 2019, 11129: 3–53.
- [47]Huang L, Zhao X, Huang K. GOT-10k: A Large High-Diversity Benchmark for

- Generic Object Tracking in the Wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1562–1577.
- [48]Fan H, Lin L, Yang F, et al. Lasot: A high-quality benchmark for large-scale single object tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5374-5383.
- [49]Mueller M, Smith N, Ghanem B. A Benchmark and Simulator for UAV Tracking[G]//LEIBE B, MATAS J, SEBE N, et al.Computer Vision – ECCV 2016. 2016, 9905: 445–461.
- [50]Li X, Ma C, Wu B, et al. Target-aware deep tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 1369-1378.
- [51]Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1420-1429.
- [52]Danelljan M, Bhat G, Khan F S, et al. Atom: Accurate tracking by overlap maximization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4660-4669.
- [53]Liu S, Zhang T, Cao X, et al. Structural Correlation Filter for Robust Visual Tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 4312–4320.
- [54]Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[J]. arXiv preprint arXiv:1710.10903, 2017.
- [55]Liao B, Wang C, Wang Y, et al. PG-Net: Pixel to Global Matching Network for Visual Tracking[G]//VEDALDI A, BISCHOF H, BROX T, et al.Computer Vision – ECCV 2020. 2020, 12367: 429–444.
- [56]Bolme D, Beveridge JR, Draper BA, et al. Visual object tracking using adaptive correlation filters[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010: 2544–2550.
- [57]Bhat G, Danelljan M, Gool L V, et al. Learning discriminative model prediction for tracking[C]//Proceedings of the IEEE/CVF international conference on

- computer vision. 2019: 6182-6191.
- [58]Chen Z, Zhong B, Li G, et al. Siamese box adaptive network for visual tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6668-6677.
- [59]Zhang Z, Peng H, Fu J, et al. Ocean: Object-aware anchor-free tracking[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. Springer International Publishing, 2020: 771-787.
- [60]Wang G, Luo C, Xiong Z, et al. Spm-tracker: Series-parallel matching for real-time visual object tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3643-3652.
- [61]Fu C, Lu K, Zheng G, et al. Siamese object tracking for unmanned aerial vehicle: A review and comprehensive analysis[J]. Artificial Intelligence Review, 2023, 56(Suppl 1): 1417-1477.
- [62]Cao Z, Fu C, Ye J, et al. Hift: Hierarchical feature transformer for aerial tracking[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 15457-15466.
- [63]Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. Advances in neural information processing systems, 2014, 27.
- [64]Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.
- [65]Tolstikhin I O, Houlsby N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision[J]. Advances in neural information processing systems, 2021, 34: 24261-24272.
- [66]Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.
- [67]Xie S, Sun C, Huang J, et al. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 305-321.

- [68]Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
- [69]Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014: 740-755.
- [70]Xu N, Yang L, Fan Y, et al. Youtube-vos: Sequence-to-sequence video object segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 585-601.
- [71]Large-Scale Machine Learning with Stochastic Gradient Descent Léon Bottou[G]//SUMMA M G, BOTTOU L, GOLDFARB B, et al. 0 ed. Statistical Learning and Data Science. 2011: 33–42.
- [72]Li S, Yeung D-Y. Visual Object Tracking for Unmanned Aerial Vehicles: A Benchmark and New Motion Models[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1).
- [73]Xu Y, Wang Z, Li Z, et al. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12549-12556.
- [74]Zhu Z, Wang Q, Li B, et al. Distractor-aware siamese networks for visual object tracking[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 101-117.
- [75]Zhang L, Gonzalez-Garcia A, Weijer JVD, et al. Learning the Model Update for Siamese Trackers[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 4009–4018.
- [76]Fu C, Cao Z, Li Y, et al. Siamese anchor proposal network for high-speed aerial tracking[C]//2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 510-516.
- [77]Wang Q, Gao J, Xing J, et al. Dcfnet: Discriminant correlation filters network for visual tracking[J]. arXiv preprint arXiv:1704.04057, 2017.

- [78]Wang Q, Zhang L, Bertinetto L, et al. Fast online object tracking and segmentation: A unifying approach[C]//Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition. 2019: 1328-1338.
- [79]Guo D, Shao Y, Cui Y, et al. Graph attention tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 9543-9552.
- [80]Sandler M, Howard A, Zhu M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 4510–4520.
- [81]Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

攻读学位期间发表的学术论文目录

- [1] Wang L, Shang Y, Cheng Q, et al. Multilayer feature fusion and saliency-attention object tracking[J]. Journal of Electronic Imaging, 2023, 32.
- [2] Wang L, Dong J, Cheng Q, etc. Velocity measurement of moving target based on rotating mirror high speed camera[J]. Optical Engineering, 2023, 62(10).
- [3] 王力超, 尚永健, 董家辉, 程清扬, 耿树巧 一种采用显著性捕获交互的目标跟踪方法及系统, 2023.05.
- [4] 王力超, 董家辉, 程清扬, 尚永健, 耿树巧 一种基于转镜式高速相机的运动目标速度测量方法, 2023.05.

致谢

在撰述这篇硕士论文之际，执笔之余，心潮澎湃，感念之情如泉涌，不吐不快。谨以此篇致谢，以表寸草之心，难以言尽。

首先，我真诚地向王力超老师表达我最深沉的敬意与感激。老师之恩，重于泰山，您的博学多识犹如明灯，照亮了我求知之路，您的严谨治学、诲人不倦，使我得以在学术瀚海中砥砺前行。从入学后的论文选题到最后的大论文修改定稿，每一步都离不开王老师的用心指导。每一次组会王老师会认真倾听我的课题汇报，针对课题进展遇到的困难都会给出实质性建议。您的一字一句，一笔一划，无不浸润着智慧的甘霖，滋养我学术之树茁壮成长。其次，我要感谢同窗好友董家辉、程清扬的陪伴与支持。三年寒窗，共研学问，你们的激扬文字、热烈讨论，丰富了我的思考维度，拓宽了我对知识的理解边界。一同度过的日夜，如同璀璨星河，点缀了我人生的旅途，使我在学术探索的道路上不再孤单。再者，衷心感谢我的家人，你们是我背后最坚实的后盾。无论风雨如何变幻，始终给予我无条件的爱与支持，让我在学业的艰辛之中感受到温暖与力量。你们默默付出的岁月和无私的关爱，是我人生中最宝贵的财富。最后，感谢那些在我研究过程中提供帮助的所有人，是你们共同构筑了我学术研究的基础，为我提供了丰富的知识源泉。

在此刻，无数感激之情凝聚成一句深情的话语：感谢你们的陪伴与支持，让我在学术的旅途中从未孤单。再一次，我衷心地向所有人表达感谢，期待我们在未来的日子里，各自在自己专注的领域里放射出璀璨的光芒，成就辉煌。

尚德敏学 唯实惟新

