

分类号: TP391

密 级: 公开

学校代码: 10363

学 号: 2220320168



安徽工程大学

Anhui Polytechnic University

硕士学位论文

题 目 利用先验信息和记忆帧精准
 融合特征的目标跟踪算法

论 文 作 者	<u>曹圣学</u>
校 内 导 师	<u>刘丙友</u>
校 外 导 师	<u>朱标</u>
专业学位类别	<u>电子信息</u>
研究方向(领域)	<u>控制工程</u>

论文提交日期: 2025 年 5 月 16 日

单位代码: 10363
学 号: 2220320168

題目 Target tracking algorithm that precisely futures
using prior information and memory frames

学 生 姓 名：曹圣学

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）：控制工程

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后果由本人承担。

学位论文作者签名：

曹圣学

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

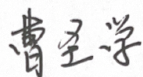
学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 ☐，在 ____ 年解密后适用本版权书。

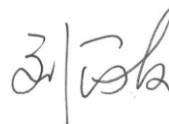
本学位论文属于

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

利用先验信息和记忆帧精准融合特征的目标跟踪算法

摘 要

单目标跟踪任务通过在视频序列的第一帧中指出目标所在的位置，并在后续视频的每一帧中持续捕捉目标位置和状态信息。因此，也可以将单目标跟踪视为一个二分类问题，即准确区分清楚前景目标与杂乱背景。其研究难点在于目标自身变化不定、背景杂乱以及目标与背景的交错融合，使得跟踪器无法分辨目标所在位置，从而影响单目标跟踪的成功率与准确度。目前，基于深度学习的单目标跟踪成为该领域的主流算法，但是在处理复杂场景下的目标消失视野以及背景遮挡问题时仍然会出现跟踪失败问题。针对上述问题，本文提出了以下改进方法：

（1）基于先验信息和多层特征精确融合的孪生网络跟踪模型。首先，为了尽可能保证跟踪目标的简洁性，SiamPIP 深度挖掘初始帧信息，通过二值化掩码操作预处理第一帧图像，从而过滤杂乱背景对跟踪过中的影响。其次，针对 SiamBan 使用修改后的 RestNet50 作为骨干网络并在最后三层均输出特征图，经过简单融合后通过头部网络进行分类和回归问题，SiamPIP 依据分类任务与回归任务所需要的特征图是不完全相同的原理，提出了精准融合特征模块。二分类任务通常需要更深层网络提取的语义信息，以此确保分类的准确性，而回归任务在特征边缘细节上有更高的要求，因此更需要浅层特征。SiamPIP 给予不同权重融合特征后再分别进行分类和回归，提高了跟踪的成功率和准确度。在 OTB 数据集上相比于 SiamBan 跟踪成功率提升 1.0%，准确率提升 1.1%。

（2）基于时间相关性的轻量化跟踪网络模型。先验信息的挖掘

与应用已经十分成熟，关注到 SiamPIP 仅在初始帧做相关研究，未能从跟踪全局的角度上思考，对跟踪器的计算效率也未简化。因此提出了 SiamGPF 算法，该算法挖掘初始帧、当前帧与历史帧之间的联系，初始帧是跟踪过程中最可靠也是唯一的已知信息，但当前帧与初始帧相比可能产生剧烈形变或遮挡，仅依靠初始帧很难准确分辨出目标所在位置，当前帧的前一帧是最能代表目标状态的序列，然而在某一帧跟踪失败时，当前帧的前一帧反而在最大程度上降低跟踪器的解析能力。因此，如何处理好这三者之间的关系是 SiamGPF 其中一个工作，该跟踪器利用残差思想保证初始帧为主要跟踪依据，并且认为初始帧更注重局部信息，使用图注意力机制进行互相关操作，而历史帧侧重关注全局信息，使用全局注意力机制是更有效的，在 SiamGPF 中设计了相关融合器处理两个互相关特征图，最后将结果进行分类与回归任务。SiamGPF 另外一个是为了保证高效的计算，在测试阶段仅保留骨干网络的第二层特征图进行后续计算，这使得跟踪器可以达到 212FPS 跟踪速率。在 OTB 数据集上相比于 SiamBan 跟踪成功率提升 1.7%，准确率提升 1.9%。

关键词：单目标跟踪，孪生神经网络，注意力机制，先验信息，特征融合，历史帧

TARGET TRACKING ALGORITHM THAT PRECISELY FUSES FEATURES USING PRIOR INFORMATION AND MEMORY FRAMES

ABSTRACT

The single-target tracking task works by pointing out where the target is located in the first frame of a video sequence and continuously capturing the target position and state information in each subsequent frame of the video. Therefore, single-target tracking can also be regarded as a binary classification problem, i.e., accurately distinguishing clearly between the foreground target and the cluttered background. The research difficulty lies in the uncertainty of the target itself, the cluttered background, and the interlaced fusion of the target and the background, which makes the tracker unable to distinguish the target location, thus affecting the success rate and accuracy of single-target tracking. At present, single-target tracking based on deep learning has become the mainstream algorithm in this field, but the tracking failure problem still occurs when dealing with the problem of target disappearing from the field of view and background occlusion in complex scenes. Aiming at the above problems, this paper proposes the following improvement methods:

(1) Twin network tracking model based on a priori information and accurate fusion of multi-layer features. Firstly, in order to ensure the simplicity of the tracking target as much as possible, SiamPIP deeply mines the initial frame information and preprocesses the first frame image through the binary mask operation, so as to filter the influence of the cluttered background on the tracking over. Secondly, for the problem that

SiamBan uses the modified RestNet50 as the backbone network and outputs feature maps in the last three layers, and then classifies and regresses them through the head network after simple fusion, SiamPIP proposes an accurate fusion feature module based on the principle that the feature maps required for classification and regression tasks are not identical. The binary classification task usually requires semantic information extracted from deeper networks to ensure classification accuracy, while the regression task has higher requirements on feature edge details and therefore requires shallow features. SiamPIP gives different weights to the fused features and then performs classification and regression respectively, which improves the success rate and accuracy of tracking. The tracking accuracy is improved by 1.0% compared to SiamBan on OTB dataset and the accuracy is improved by 1.1%.

(2) Lightweight tracking network model based on temporal correlation. The mining and application of a priori information has been very mature, and it is concerned that SiamPIP only does relevant research in the initial frame, fails to think from the perspective of tracking global, and does not simplify the computational efficiency of the tracker. Therefore, the SiamGPF algorithm is proposed, which mines the connection between the initial frame, the current frame and the history frame. The initial frame is the most reliable and the only known information in the tracking process, but the current frame may produce drastic deformation or occlusion compared to the initial frame, and it is difficult to accurately distinguish the target's location by relying only on the initial frame, and the previous frame of the current frame is the most representative sequence of the target's state, however, in the case of tracking failure of a frame, the previous frame of the current frame is the most representative of the target's state. However, when the tracking fails in a certain frame, the previous frame of the current frame reduces the tracker's resolution

ability to the greatest extent. Therefore, how to deal with the relationship between these three is one of the work of SiamGPF, the tracker uses the residual idea to ensure that the initial frame is the main tracking basis, and that the initial frame focuses more on local information, and uses the graph attention mechanism to perform the mutual correlation operation, while the history frame focuses on global information, and uses the global attention mechanism to be more effective, and the correlation fusion mechanism is designed to deal with the two mutual correlation feature graphs. In SiamGPF, a correlation fusion machine is designed to process the two inter-correlation feature maps and finally the results are used for classification and regression tasks. Another work in SiamGPF is to ensure efficient computation, only the second layer feature map of the backbone network is retained for subsequent computation during the testing phase, which allows the tracker to achieve a tracking rate of 212 FPS. There is a 1.7% improvement in tracking accuracy and a 1.9% improvement in accuracy compared to SiamBan on the OTB dataset.

KEY WORDS: Single Object Tracking, Siamese Neural Network, Attention Mechanism, Prior Information, Feature Fusion, Historical Frames

目 录

第 1 章 绪论	1
1.1. 研究背景及研究意义.....	1
1.2. 国内外研究现状.....	2
1.2.1. 基于相关滤波的单目标跟踪算法.....	2
1.2.2. 基于深度学习的单目标跟踪算法.....	3
1.3. 本文主要内容.....	5
1.4. 本文组织结构.....	7
第 2 章 单目标跟踪相关理论基础	9
2.1. 卷积神经网络.....	9
2.2. 孪生神经网络.....	13
2.2.1. 孪生网络原理.....	13
2.2.2. 基于孪生网络的单目标跟踪算法.....	14
2.3. 注意力机制.....	17
2.3.1. 通道注意力机制.....	17
2.3.2. 空间注意力机制.....	18
2.3.3. 时间注意力机制.....	19
2.4. 测试数据集及评价标准.....	20
2.5. 本章小结.....	23
第 3 章 基于先验信息和多层特征精确融合的跟踪网络研究	24
3.1. 引言.....	24
3.2. SiamPIP 网络架构.....	25
3.2.1. 先验信息和多层特征精确融合.....	25
3.2.2. 多比例融合网络.....	26
3.2.3. 基于先验信息的特征增强.....	28
3.3. 实验方法与结果分析.....	28
3.3.1. 实验环境介绍.....	29
3.3.2. 实验结果.....	30
3.3.3. 消融实验.....	34
3.4. 本章小结.....	35
第 4 章 基于时间相关性的轻量化跟踪网络研究	36
4.1. 引言.....	36
4.2. SiamGPF 网络架构	37
4.2.1. 基于伪先验信息的网络框架设计.....	37
4.2.2. 基于图注意力机制的局部匹配策略研究.....	39
4.2.3. 基于自注意力的全局匹配策略研究.....	40
4.2.4. 特征匹配结果融合策略研究.....	42
4.2.5. 测试伪先验帧选择策略研究.....	43
4.3. 实验方法与结果分析.....	45
4.3.1. 实验环境介绍.....	45
4.3.2. 实验结果.....	46
4.3.3. 性能与参数量比较.....	51

4.3.4. 消融实验.....	53
4.3.5. 可视化分析.....	56
4.4. 本章小结.....	58
第 5 章 总结与展望	59
5.1. 总结.....	59
5.2. 展望.....	59
参考文献.....	61
攻读学位期间发表的学术成果.....	67
致谢.....	68

第1章 绪论

1.1. 研究背景及研究意义

随着计算机视觉技术和硬件设施的飞速进步，如今的日常生活当中充斥着海量的图像和视频，如何从这些杂乱的图像数据中提取出有效信息并进行分析判断是个极其重要的研究方向。而单目标跟踪作为计算机视觉的重要分支一直备受众多科研人员关注。单目标跟踪技术的含义是：给定一组图像视频序列，指定第一帧中目标所在位置，然后在之后的图像当中持续定位目标，示例如图 1-1 所示。

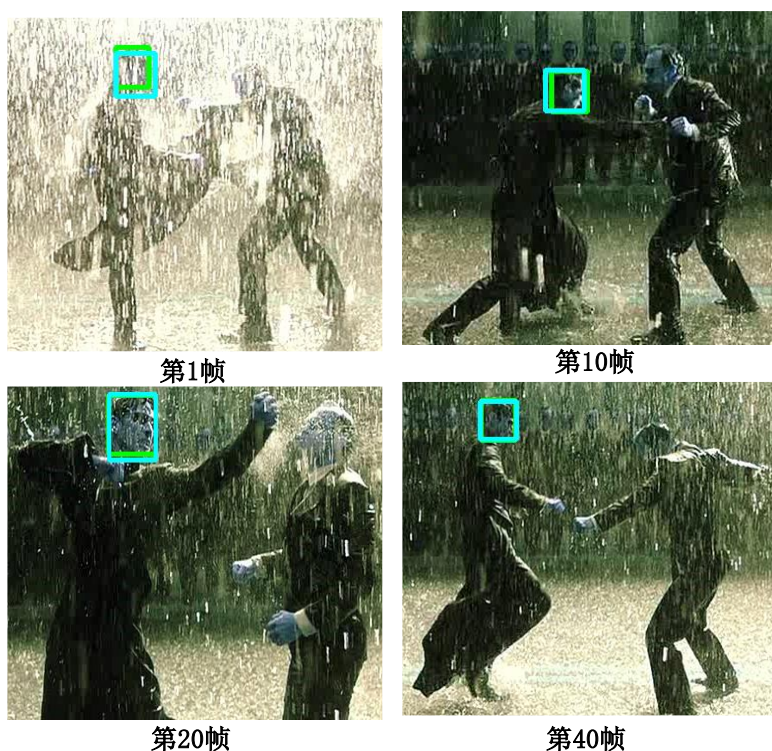


图 1-1 单目标跟踪的矩形框输出图例

目前单目标视频追踪技术已经广泛应用于实际生活领域，并且存在巨大潜力。安防系统^[1]、交通监控^[2]是深度使用目标跟踪技术的领域，将该技术与监控系统相结合，可以对视频中的可疑人员、车辆进行实时检测并跟踪，从而为安全性分析、异常行为检测、交通人流管控等提供重要数据，有效地降低人工计算量。单目标跟踪技术在军事航空^[3]中的应用也十分广泛，该项技术可以迅速准确定位目标从而实现战略性精准打击。无人机在如今的军事行动中已经不可或缺，通过目标跟踪技术进行空中战略侦察，捕捉敌方军事行动以及战略部署，同时无人机可以进行小规模空中打击，对战场起到关键作用，降低了己方人员和物资的损耗。

在高速子弹弹道研究、导弹精准拦截等技术中也着重使用了目标跟踪，文献^[4]通过振镜旋转扩大高速相机视场范围，保证子弹持续出现在视频图像中心位置，这在军事研究领域发挥及其重要的作用。在人机交互领域，目标跟踪技术通过相机等硬件设备采集图像来交由机器处理并获得反馈。文献^[5]创新性的通过捕获人体面部表情细微变化来监测心率变化，其中较为关键的技术就是目标跟踪。无人驾驶^[6]在近几年发展十分迅速，该技术主要使用各种类型的传感器采集车体周边信息，然后经过主控中心计算完成自动避障、路径规划等操作，从而达到安全自动驾驶的目的。目标跟踪技术是无人驾驶软件系统的重要组成部分，当传感器捕捉到周围障碍物并通过该算法分析周围情况，为后续避障等操作铺垫。目标跟踪技术同样可以应用体育赛事、医疗诊断领域，极大提升了各个行业的智能化程度。因此，研究单目标跟踪技术具有重大价值和重要意义。

1.2. 国内外研究现状

1.2.1. 基于相关滤波的单目标跟踪算法

基于相关滤波的单目标跟踪技术在早期以优越的运行速度和准确度引发了大量该领域研究者的关注。该类算法的基本原理是通过不断更新目标模板特征，以确保在视频序列中准确捕捉目标所在位置。该算法首先利用视频序列的首帧信息构建初始目标模板，并通过颜色直方图或梯度特征等方式提取目标与下一帧图像的特征。然后计算目标模板与当前帧图像的相关性，确定目标在当前帧中的位置。目标模板会根据上一帧确定的目标位置进行动态更新，从而实现目标的持续跟踪。

Bolme^[7]等人在 2010 年提出了 MOSSE (Minimum Output Sum of Squared Error) 算法，该算法首次将相关滤波引入单目标跟踪领域。MOSSE 使用快速傅里叶变换将图像数据转换到频域，在数学形式上，通过用乘积运算代替卷积运算，显著提升了跟踪器运算效率。该算法在当时不仅具备较高的跟踪精度，且在 CPU 上运行速度达到了 600-700 FPS (Frames Per Second)。使用卷积核对搜索区域进行卷积运算的公式表达如下：

$$G=F \times H^*, \quad (1-1)$$

其中， G 代表算法中的响应图， F 和 H^* 分别为搜索区域和卷积核的频率域图。

相关研究人员通过优化目标特征类型来增强 MOSSE 跟踪器性能。Henriques

等人提出了 CSK (Circulant Structure with Kernels, CSK)^[8]和 KCF (Kernelized Correlation Filter, KCF)^[9]算法。CSK 在 MOSSE 的基础上引入核技巧来提升图像特征的表达能力,并且在损失函数中添加正则项来避免跟踪器训练过拟合;而 KCF 通过使用循环矩阵计算加快了跟踪速度,同时用多通道 HOG 特征替换了原始的单通道像素特征,使得该跟踪器能够更好地表征物体。这些改进使得相关滤波跟踪算法在精度和速度上超越了以往的相关滤波方法,且仍保持了极高的计算效率,CSK 和 KCF 在单 CPU 上的运行速度分别可以达到 300 FPS 和 200 FPS。

KCF 存在输出跟踪框固定的缺陷,因此在应对尺度变化和遮挡问题时存在局限性。后续研究者为了使跟踪器能够应对更多场景,基于 KCF 进行了大量改进。DSST^[10]进一步探讨了多尺度问题对目标跟踪的影响,在位置计算的基础上添加了缩放尺度回归;Staple^[11]算法融合了 HOG 和 COLOR 特征,能够互补解决各自的不足,从而提升跟踪的稳定性和准确性;SAMF^[12]通过平移滤波器在多尺度缩放的图像块上捕捉目标,取响应值最大的平移量和所在尺度,从而应对目标中心变化和尺度变化的挑战;SRDCF^[13]和 CFLB^[14]的创新点都是扩大搜索区域,同时规定滤波模板的作用范围;LMCF^[15]对平移检测的响应图进行多峰检测,仅在跟踪置信度较高时更新跟踪模型,提升了跟踪器的鲁棒性。

1.2.2. 基于深度学习的单目标跟踪算法

深度学习技术的出现开创了一种新的跟踪范式,通过 Siamese 网络来实现跟踪。SiamFC^[16]通过提取图像对中的特征并使用精确匹配来完成跟踪,该跟踪框架在后续研究中得到了广泛的发展。为了应对多尺度处理挑战,SiamFC 引入了三种不同尺度的搜索区域,可以更好地跟踪变形目标,但是这一方法增加了计算需求,在精度上也存在一定的局限性。SiamRPN^[17]将更快的 R-CNN^[18]概念集成到 Siamese 网络跟踪中,另外一个创新点是将跟踪任务分为分类和回归子任务,相较于 SiamFC,该跟踪器通过采用多个锚框提高了精度,速度得到很大提升。DaSiamRPN^[19]强调训练数据处理的优化,增强了模型对干扰的鲁棒性,从而提高跟踪器的泛化能力。SiamBan^[20]和 SiamCAR^[21]通过利用更深层的网络架构提升对图像数据的特征表达能力。同时,SiamMask^[22]将跟踪和分割任务结合,达到了 55 FPS 的跟踪帧率,跟踪成功率仍然较高。然而有研究者发现随着骨干网络深度的持续增加反而降低了精度,SiamRPN++^[23]通过统一采样、多层融合以

及深度交叉相关技术来解决这一问题,并且引入深度卷积交叉相关性来进行模板特征与搜索图像特征的相关性评估。RASNet^[24]深度使用注意力机制、通道注意力和残差注意力块可以有效应对跟踪目标外观变化。尽管多数 Siamese 网络系列取得了卓越的成果,但它们对超参数的敏感性较高,需要进行大量的参数调优,这违背了深度学习的基本原则。为此,各种无锚点的跟踪策略相继出现,这些策略直接对潜在目标进行前景-背景二值分类,确定边界框与实际目标的接近程度,采用单一或多个头进行目标预测,并在交叉相关后的响应图上进行像素级别的回归。该方法不仅提高了跟踪精度,还有效减少了跟踪器对超参数的敏感性,从而增强模型的泛化能力,为未来的工作奠定坚实基础。SiamGAT^[25]引入了图注意力机制,通过在图像对之间实施局部信息传播进行优化。AutoMatch^[26]则通过神经架构搜索(NAS)^[27]显著增强了跟踪器的效能,旨在找到最优的分类和回归匹配策略。

近年来,Transformer^[28]技术在视觉处理领域的应用显著扩展,TransT 开创性地将该架构应用于目标跟踪领域,打破了传统的交叉相关方法。DropMAE^[29]解决了传统 MAE^[30]方法过度依赖空间线索且忽视帧间时间关系的问题。E.T.Track^[31]则利用单实例级别的注意力层进行实时视觉目标跟踪。HiT^[32]建立了轻量化 Transformer 与跟踪框架之间的联系,将深层特征中的高层信息与浅层高分辨率特征相结合,为跟踪头网络生成更好的特征。同时为了激发学者们对 GOT 研究的兴趣,Mayer^[33]等人提出了一个新的大规模 GOT 基准 LaGOT。

视觉跟踪中的时间信息:在视频单目标跟踪领域,时间信息的利用成为处理目标外观变化和应对复杂场景的重要因素。大多数 Siamese 和 Transformer 跟踪算法通过在线模板更新机制来实现这一目的。UpdateNet^[34]通过动态方式修改模板来应对目标外观的变化,可以更好整合时间信息。DiMP^[35]利用背景数据构建判别式学习框架,并通过在线更新持续优化模型,有效减轻了遮挡和快速运动带来的挑战。DCF 系列^{[36]-[38]}跟踪器中,ATOM 采用先进的优化方法,实时调整模型参数,从而取得了显著的成绩。UpdateNet 认为现有的模板更新方法过于复杂,提出了一种端到端的记忆网络,用于跨帧全局同步数据。然而这种方法存在较大的计算需求,并且降低了跟踪器的帧率,训练过程中需要包含大量的超参数,增加了复杂性。STARK^[39]则利用 Transformer 强大的建模能力,更新动态模板输入并获取时间上下文,将目标跟踪视为令牌序列传播的过程,通过自回归方式将时

间信息纳入跟踪过程，更精准地捕捉目标随着时间演变的信息。尽管这些方法在利用时间信息方面取得了一定成效，但复杂的更新机制和巨大的计算需求也带来了明显局限。因此，本文探讨了一种更简洁的方式来嵌入时间信息。

匹配策略：在 Siamese 跟踪器研究中，模板图像与搜索图像的匹配任务始终是核心问题。SiamFC 率先应用相关性计算方法，将模板特征作为卷积核，以逐通道的方式与搜索图像特征进行卷积运算，最终生成统一样式的单通道特征图。SiamRPN++ 通过融合深层与浅层特征，增强了网络的语义表达能力，并针对交叉相关过程中，由于特征维度增加而导致的参数不平衡问题，引入了深度可分离相关性计算（depth-wise correlation），直接利用逐通道卷积结果进行后续计算。SiamGAT 则批判了传统的全局搜索方法，认为其难以有效应对目标长宽比的变化，因此引入图注意力模块来构建目标与搜索区域之间的局部连接，从而提升对目标形态变化的适应能力。

PG-Net^[40]首先提出了像素到全局相关性的概念，将模板特征划分为尺寸为 1×1 的空间核，并按照通道维度进行计算，旨在缩小匹配范围，从而有效减少背景噪声干扰。SAOT^[41]提出了一种细粒度显著性挖掘模块，用于捕捉目标的局部显著特征，这些特征对于跟踪任务具有较强的判别性。此外，还设计了显著性关联建模模块，用于将捕捉到的显著性特征进行关联学习，从而建立模板与搜索图像之间的有效全局相关性，然而该方法的计算复杂度仍然是一个主要瓶颈。AutoMatch 则利用神经架构搜索技术，提出了一种概念简洁的算法，即通过二值化通道操作自动设计匹配网络，并将学习到的匹配网络集成到基准跟踪器中，从而实现了显著的性能提升。

在上述研究的启发下，本文提出的方法在当前帧采用局部匹配计算，以提高跟踪器目标在运动过程中外观变化的适应能力。同时，在历史帧上采用全局匹配方法，以提供当前场景所需的上下文信息。这种双重策略旨在结合局部匹配的精确性与全局匹配的全局信息，从而实现更准确的跟踪。

1.3. 本文主要内容

本文的工作基于深度学习方法，研究基于孪生神经网络的单目标跟踪模型，针对复杂情况下跟踪器效果不佳甚至跟踪失败的问题，在 SiamBan 的基础上提出了 SiamPIP 和 SiamGPF 两个跟踪器。

本文的主要工作和贡献如下：

1、结合具体领域总结了单目标跟踪任务在国内外发展现状以及存在的研究价值，并阐述了相关基础知识点的概念和原理。

2、提出了基于先验信息和多层特征精确融合的跟踪器（SiamPIP），当面对目标发生剧烈变形，特别是当目标的纵横比发生极端变化时，大多数跟踪器在此情况下容易丢失目标，从而导致跟踪失败。为了解决这一挑战，本文提出了一种基于先验信息和比例融合的无锚点跟踪器。初始帧提供的真实目标框具有较高的纯度，不受噪声干扰，成为整个跟踪任务中的唯一先验信息。因此，通过筛选从骨干网络提取的模板特征，可以有效地利用初始帧的信息。此外，传统的孪生网络模型往往在融合深层和浅层特征时忽视了分类任务与回归任务在特征融合中的作用差异。分类任务通常需要更深层次的语义信息，而浅层网络的特征更适合回归任务，因此在特征融合时需要进行精确区分。在面对严重变形的目标时，SiamPIP 能够准确地区分前景与背景，并保持回归边界框的精确性。通过实验验证，该跟踪器在 OTB100 数据集上获得了 0.707 的 AUC，在 VOT2019 数据集上取得了 0.340 的 EAO，在 UAV123 数据集上获得了 0.643 的 AUC，表现出卓越的性能。

3、提出基于时间相关性的轻量化跟踪网络（SiamGPF）。在 SiamPIP 跟踪器的启发下，本文从如何处理初始帧、历史帧与当前帧的关系入手，利用前一帧（最能准确表达当前帧目标信息的帧）作为历史帧，并与当前跟踪序列进行互相关处理；而初始帧未受杂乱信息污染，与当前跟踪序列进行融合，并利用残差思想，以初始帧为主体结构，保证跟踪器性能不受影响。本文提出了基于图注意力机制的局部匹配模块、基于自注意力的全局匹配模块、特征匹配结果融合模块以及历史帧选择模块。经消融实验和公开测试集评估，结果表明本文提出的方法是有效的。此外，SiamGPF 为了提升跟踪帧率，仅使用修改后的骨干网络的某一层输出结果，使其在 GOT-10k 官方评估器上达到 212FPS 的跟踪帧率。

4、本文充分挖掘初始帧、历史帧和当前帧之间的联系，提出了 SiamPIP 和 SiamGPF 两个离线跟踪器，既保证了跟踪器的鲁棒性，又避免了因模板更新带来的计算负担，为单目标跟踪器在不同状态下所需的目标信息提供了理论依据，并为该领域的发展做出了一定贡献。

1.4. 本文组织结构

本文针对复杂场景下的单目标跟踪问题进行分析和改善，在孪生网络 SiamBan 的基础上进行了多方面改进，包括精准融合、模板滤波、伪先验帧框架等多方面内容。本文共五个章节，图 1-2 详细展示了本文的结构安排。

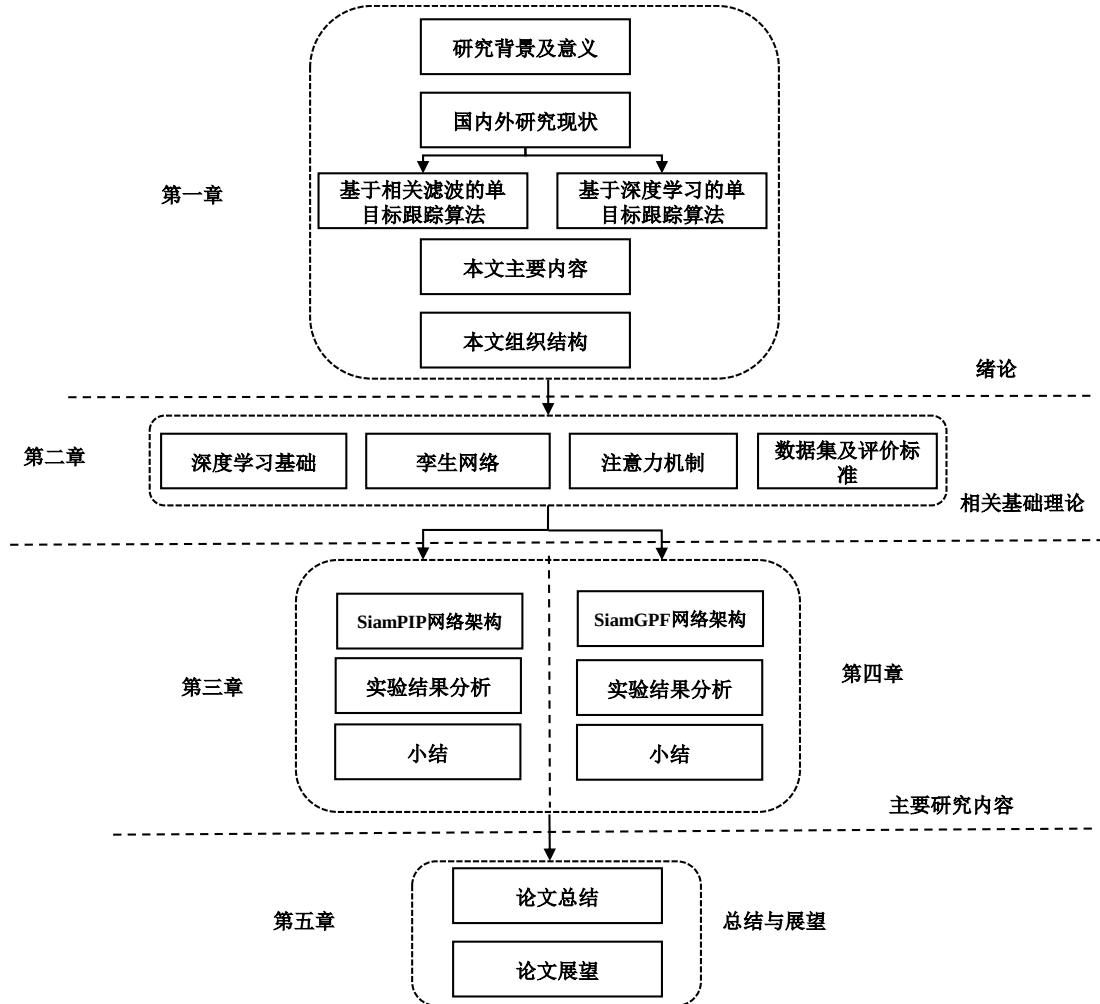


图 1-2 本文章节内容

第一章：绪论。首先结合时代发展背景介绍了目标跟踪技术的应用场景与前景，然后根据发展历程介绍了基于相关滤波的跟踪算法和基于深度学习的跟踪算法，并在第三小节中介绍本文做出的主要工作，最后对文章的主体结构进行总结。

第二章：相关理论基础。本文的主要工作是在孪生网络基础上进行改进，主要涉及网络结构和特征融合方法，因此首先介绍了卷积神经网络的相关概念，其次分析孪生网络在单目标跟踪中的应用，然后结合跟踪器的改进方法，介绍几种注意力机制。最后介绍了测试跟踪器性能的数据集和评价标准。

第三章：本章重点分析跟踪器 SiamPIP，首先介绍该跟踪器的基本结构框架，

其次分析多尺度融合网络和基于先验信息的特征增强方法的原理，最后在测试集上评估跟踪器的综合性能及各部件的贡献度。

第四章：基于时间相关性的轻量化跟踪网络研究。在第三章的思想基础上，本算法将不再局限于挖掘初始帧信息与双输入的孪生架构，而是将历史帧作为伪先验信息融入到孪生网络当中，关注序列之间的联系，并通过修改骨干网络结构减轻跟踪器计算负载。首先介绍了跟踪器 SiamGPF 的基本网络架构，其次分析改进部件的作用和原理，然后通过多个具有不同侧重点的数据集评估跟踪器性能，最后通过详细的消融实验和可视化方法，进一步验证所提出方法的有效性。

第五章：总结与展望。该章节分为两小节，首先总结性介绍了本文做出的贡献，然后分析文中提出算法的一些问题，最后探讨指出未来可以进一步改进该方法的方向。

第2章 单目标跟踪相关理论基础

研究计算机视觉首先需要详细了解卷积神经网络，本章从卷积神经网络基础开始介绍，并详细介绍了该类网络的基本结构。其次，本文的研究内容是基于孪生网络算法，接下来，本章将介绍基于孪生网络的单目标跟踪算法框架以及相关注意力机制。

2.1. 卷积神经网络

卷积神经网络（Convolutional Neural Networks, CNN）^[42]是一种包含卷积运算且具有深度结构的前馈神经网络，是深度学习领域的重要算法之一。它在计算机视觉领域表现出色，能够有效提取特征，在图像分类、目标检测、目标跟踪等任务中取得显著成果。如图 2-1 所示，卷积神经网络通常由卷积层、池化层、全连接层、激活层和输出层等基本模块构成。输入层是卷积神经网络的数据输入，通常是经过裁剪、色彩转换等预处理的图像，并以三维矩阵的形式进行计算。

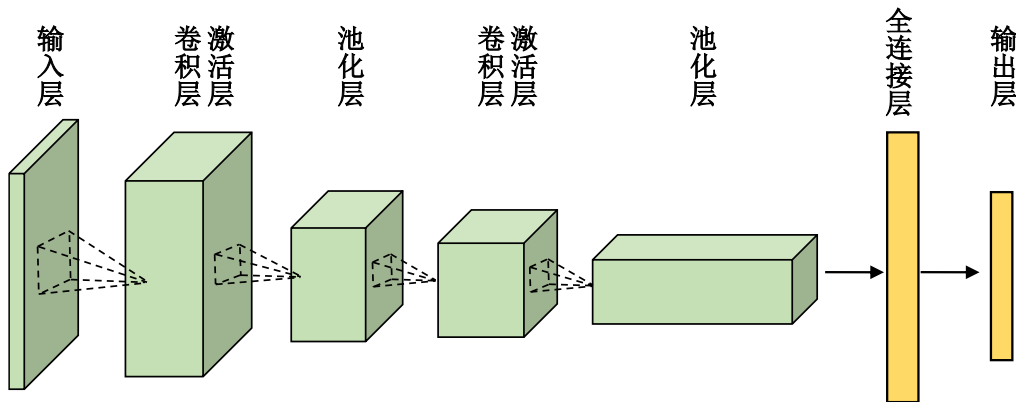


图 2-1 卷积神经网络示意图

卷积层是卷积神经网络的核心部分，由一组可学习的卷积核和偏置参数组成，通过卷积核对输入三维矩阵进行卷积运算，提取特征。在计算机视觉中，卷积操作是常见的基础运算，它使用一个矩阵，被称为卷积核（Kernel），通常是如 1×1 、 3×3 等奇数大小的方形矩阵。卷积核越大，得到的感受野就越大，但计算量也会增大。如图 2-2 所示，卷积运算的过程是将卷积核作为权重，在输入层上大小相同的窗口位置进行加权求和，并将运算结果与偏置参数相加。卷积运算的过程中，引入了步长（Stride）和填充（Padding）的概念^[43]，卷积层由一组

小型的可训练卷积核和偏置参数组成，卷积核的大小和形状可根据具体任务需求进行调整，步长是卷积核逐次滑动的距离，较小的步长会导致输出特征图尺寸较大，较大的步长则会使输出尺寸变小。在卷积过程中，由于周围像素点只参与一次计算，而中心像素点参与多次计算，使用零填充可以避免图像边界信息的丢失。

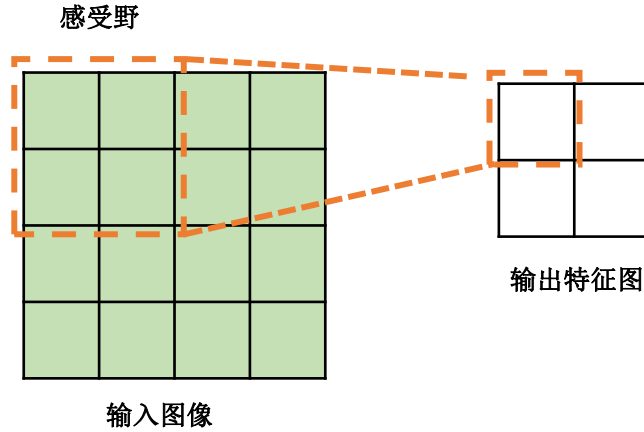


图 2-2 感受野示意图

卷积运算的结果特征图大小与卷积核、步长以及填充层数有关，具体计算公式如（2-1）所示：

$$Size_f = \left[\frac{n+2p-f}{s} + 1 \right] \times \left[\frac{m+2p-f}{s} + 1 \right], \quad (2-1)$$

其中， n 和 m 分别代表输入特征图的宽和高， p 代表填充的大小， f 表示卷积核的大小， s 代表步长。

池化层的主要作用是减少数据的尺寸和计算量，从而提高模型的运算效率，它通过对特征图的每个区域进行聚合操作，压缩卷积特征图的尺寸，去除冗余的特征数据。同时，池化操作没有需要训练的参数，可以增强特征的鲁棒性。常用的池化方式有最大池化和平均池化，见图 2-3。池化操作通过计算特定区域中所有数据的最大值或者平均值来代表该窗口的特征。此外，池化操作还可以帮助减少过拟合，增强模型的泛化能力。

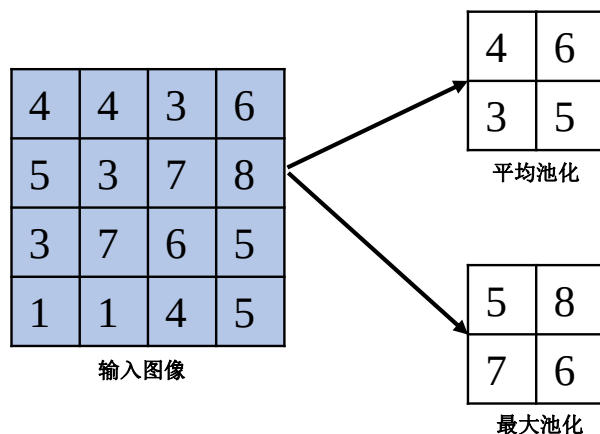


图 2-3 平均池化和最大池化

激活层的主要作用是提供网络的非线性建模能力，没有激活函数的卷积网络仅能表达线性映射，网络深度的增加将变得无效。激活层通过应用非线性函数，使卷积网络能够学习更加复杂的数据表示。常见的激活函数有 ReLU（Rectified Linear Unit）^[44]函数、Sigmoid 函数、tanh 函数等，其定义如下式：

$$f_{\text{sigmoid}}(z) = \frac{1}{1 + e^{-z}}, z \in R \quad (2-2)$$

Sigmoid 函数几何表示见图 2-4：

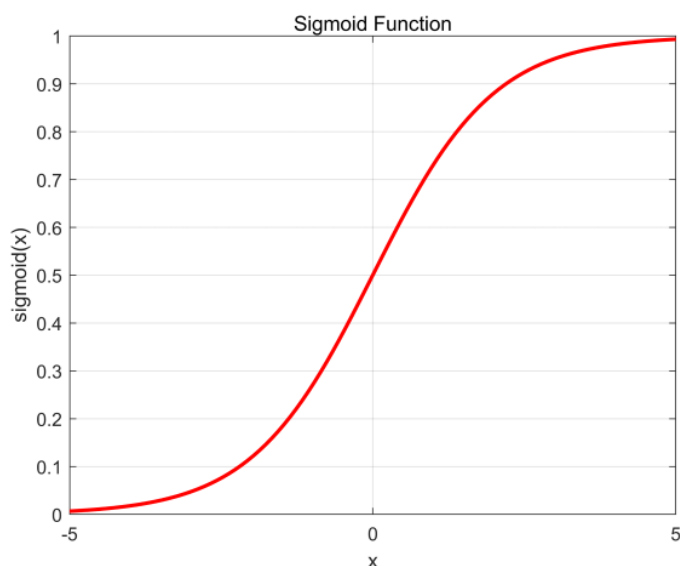


图 2-4 Sigmoid 函数曲线

$$f_{\text{tanh}} = \frac{e^z - e^{-z}}{e^z + e^{-z}}, z \in R \quad (2-3)$$

tanh 函数几何表示见图 2-5：

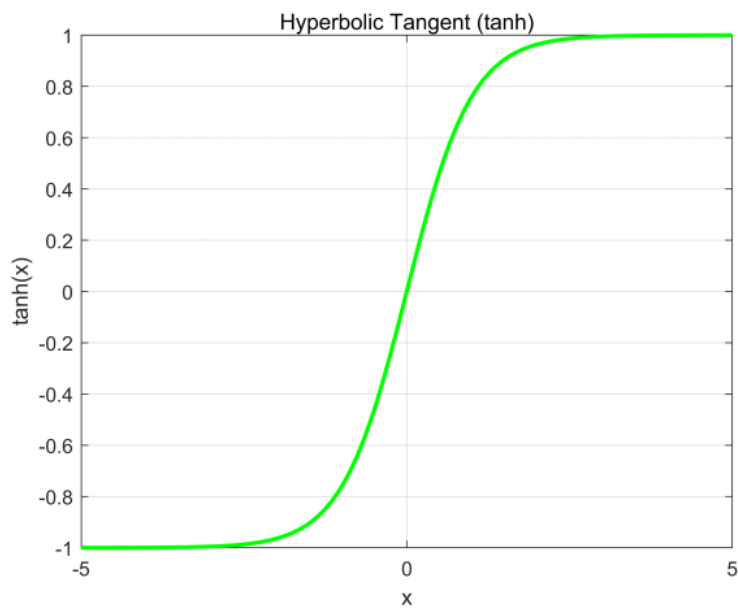


图 2-5 Tanh 函数曲线

$$f_{ReLU} = \begin{cases} 0, & z < 0 \\ z, & z \geq 0 \end{cases} \quad (2-4)$$

ReLU 函数几何表示见图 2-6:

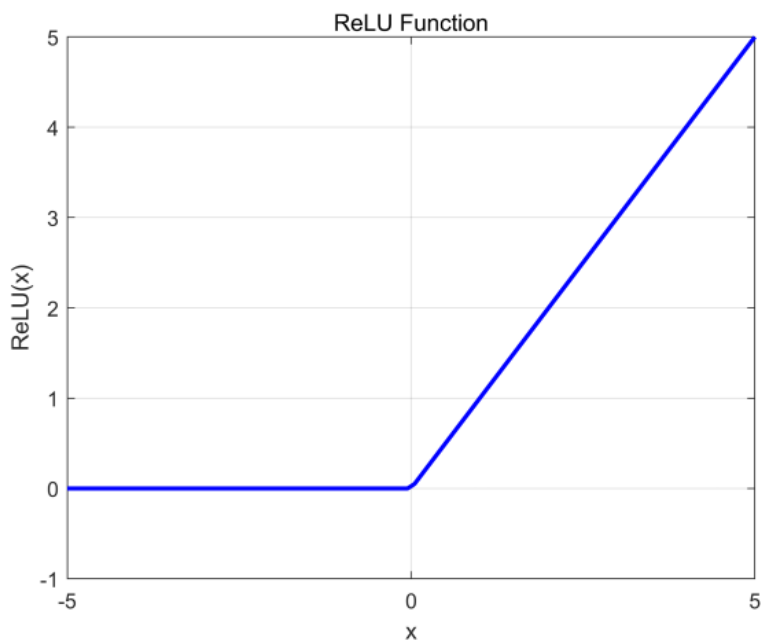


图 2-6 ReLU 函数曲线

在训练中使用最多的是 ReLU 激活函数及其变体，ReLU 函数在正区间内为线性，不存在饱和问题，有效缓解了训练时梯度消失问题，在负区间内为零，不

仅提高了计算效率提高，而且一定程度上缓解了过拟合。Sigmoid 函数的梯度在饱和区间接接近 0，导致权重更新速度变慢，从而影响训练速度。tanh 函数虽然经过原点，能够显著提升效果，但计算复杂，使用较少。随着技术的进步，出现了许多优秀的损失函数，如 EMO^[45]、Huber^[46]等，它们在特定的环境下表现优异，但是 ReLU 仍然是多数场景下表现较好的激活函数。

2.2. 孪生神经网络

孪生网络在单目标跟踪中的应用十分广泛，本文基于孪生网络框架提出了跟踪算法。首先，在第一节介绍了孪生网络与单目标跟踪算法的相关知识；其次，在第二节中介绍了常见的几种注意力机制；最后介绍了单目标跟踪算法在性能测试中使用的数据集和评价标准。

2.2.1. 孪生网络原理

孪生网络（Siamese Network）^[47]由两个或多个相同结构的神经网络组成，形成一个耦合架构，两个网络的结构参数共享。如图 2-7 所示，孪生网络通常用于衡量输入一和输入二的相似度和差异性。早在 20 世纪 90 年代，基于孪生网络的算法就已应用于相似度判别，由于孪生网络参数较少、运行速度快，它逐渐被引入到小样本学习、目标跟踪等领域。

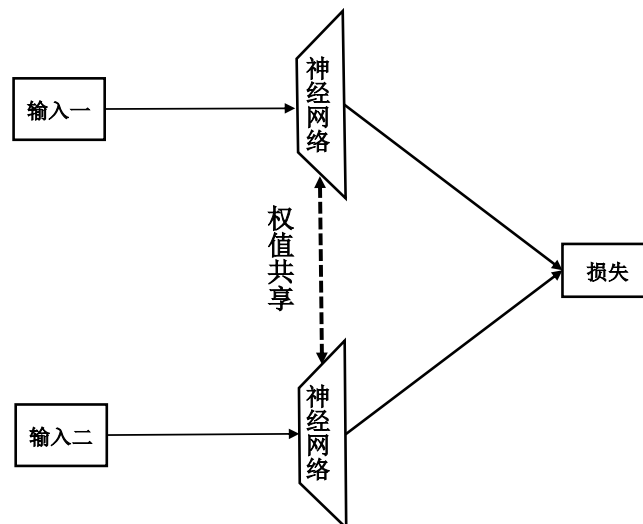


图 2-7 孪生网络原理图

2.2.2. 基于孪生网络的单目标跟踪算法

目标跟踪旨在根据第一帧指定的目标对象，在后续帧中进行持续跟踪。由于无法对图像中的所有对象进行建模，基于孪生网络的单目标跟踪算法将目标跟踪问题转化为相似度匹配问题。这种方法不仅减少了跟踪器的计算量从而提高跟踪速度，还显著提升了准确率。

SiamFC 跟踪算法首次将孪生网络引入单目标跟踪领域，并且以结构简单、跟踪鲁棒性强等优点著称，在随后的几年中迅速成为跟踪领域的主流算法框架，图 2-8 是 SiamFC 跟踪器的主体结构，该跟踪器的骨干网络基于 AlexNet^[48]，结构如图 2-9 所示，该网络为了避免搜索区域周边的候选区域存在无效信息，因此并不使用零填充操作。

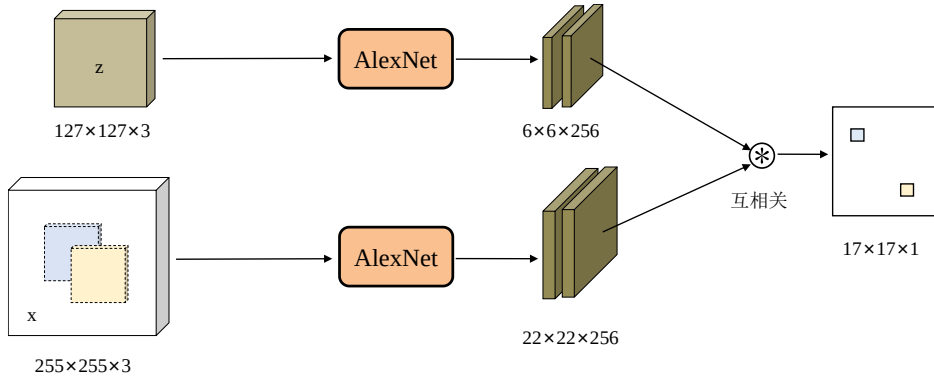


图 2-8 SiamFC 网络结构

SiamFC 的核心思想在于，提取模板图像特征与搜索图像特征时使用的网络结构参数是相同的，在训练过程中无需分别训练骨干网络的超参数。此外，SiamFC 将模板图像特征作为被卷积对象与搜索图像特征完成卷积操作，计算目标与待搜寻区域的匹配度。另外，SiamFC 通过构建搜索区域图像金字塔，解决了相邻帧目标尺度变化过大的问题。

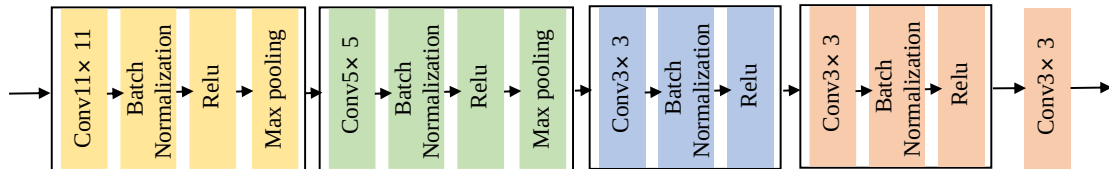


图 2-9 基于 AlexNet 的网络结构

本文的整体算法研究是在 SiamBan 跟踪器的基础上进行的。与 SiamFC 不同，SiamBan 将目标追踪任务分解为分类与回归问题，并采用无锚点边框回归方式，能更精准地预测目标的像素坐标。

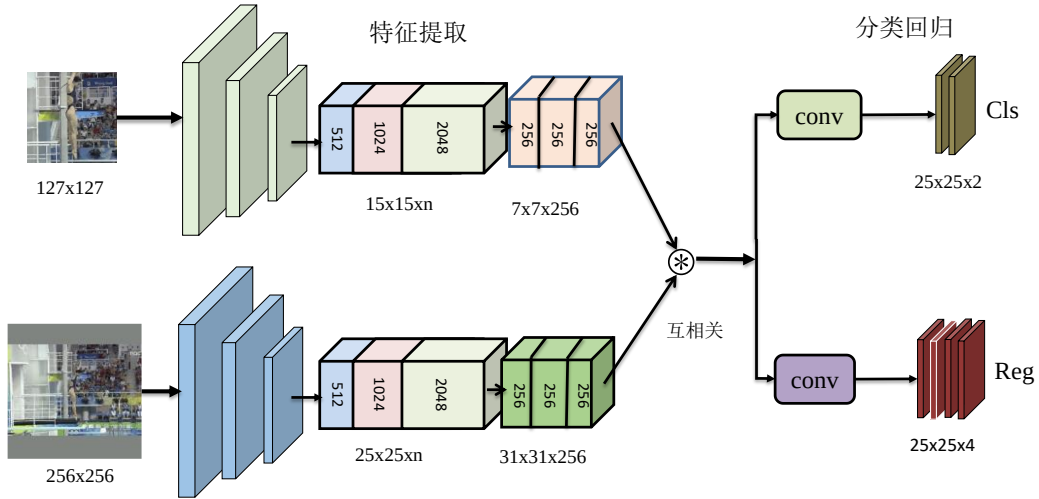


图 2-10 SiamBan 的网络结构

SiamBan 跟踪模型同样基于孪生网络结构进行设计，如图 2-10 所示。它的输入是一对经过预处理的图像，分别为视频序列的第一帧（已给定目标位置）和后续帧图像。SiamBan 的骨干网络采用修改版的 ResNet50。图 2-11 展示了 ResNet 网络瓶颈残差的核心思想。一个标准的残差块可分为两个主要路径：一条是包含多个卷积层和激活函数的残差部分，另一条则是恒等映射。两者输出相加，形成残差瓶颈结构的最终输出。该结构的优点是能够在降低计算量的同时，保持网络的有效性。

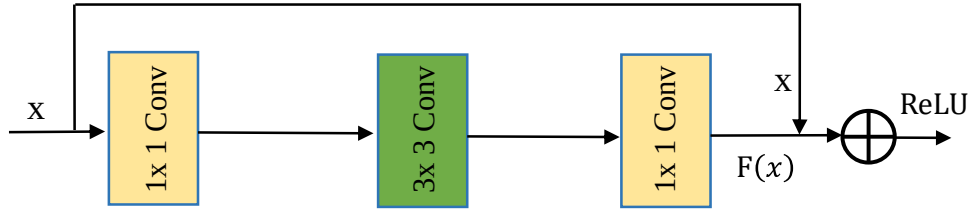


图 2-11 残差结构

修改后的 ResNet50 去掉了最后两个块的步长，并增加了空洞率为 2 和 4 的空洞卷积，从而扩大感受野，确保最后三层的特征图分辨率保持一致。在 neck 网络结构中，模板图像的 7x7 区域被截取，方便后续的互相关运算。SiamBan 的另一个优势是引入了无锚点预测框。SiamRPN 通过交并比来判定正负样本：交并比大于 0.6 的为正样本，交并比小于 0.6 的为负样本。为了降低运算成本，SiamRPN 仅选择 16 个正样本和 48 个负样本。而 SiamBan 在搜索补丁上使用了两个椭圆 E_1 和 E_2 来表示。椭圆定义如下：

$$E_1: \frac{(p_i - g_{x_c})^2}{(\frac{g_w}{2})^2} + \frac{(p_j - g_{y_c})^2}{(\frac{g_h}{2})^2} = 1, \quad (2-5)$$

$$E_2: \frac{(p_i - g_{x_c})^2}{(\frac{g_w}{4})^2} + \frac{(p_j - g_{y_c})^2}{(\frac{g_h}{4})^2} = 1, \quad (2-6)$$

其中, (P_i, P_j) 为搜索补丁上的坐标, g_w, g_c 代表真实框的宽和高, g_{x_c}, g_{y_c} 代表真实框的中心点, 跟踪器以目标对象为中心, 规定在椭圆 E_2 内的部分为正样本, 椭圆 E_1 外的部分为负样本, 二者之间的区域则被忽略, SiamBan 中特征图上的点与搜索补丁上的点的对应关系为:

$$(p_i, p_j) = [\frac{w_{im}}{2} + (i - \frac{w}{2}) \times s, \frac{h_{im}}{2} + (j - \frac{h}{2}) \times s], \quad (2-7)$$

其中, w_{im}, h_{im} 为搜索补丁的宽、高, w, h 为特征图的宽、高, s 为总步长。

此外, SiamBan 在测试过程中需要经过汉明窗处理来减小边缘信息的影响, 从而更能有效应对相似物干扰和尺度变化的挑战。图 2-12 是 SiamBan 在 OTB100 某两个序列中的得分可视化, 第一行和第三行是跟踪框输出图例, 第二行代表未使用余弦窗惩罚的得分热图, 第四行表示未使用惩罚的得分热图。

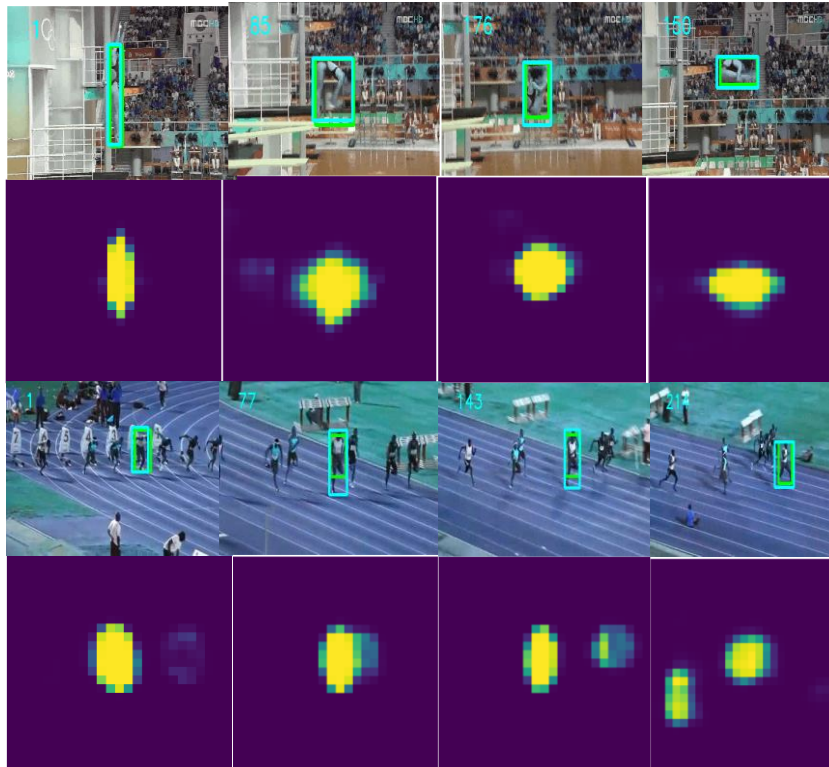


图 2-12 余弦窗惩罚得分热图对比

2.3. 注意力机制

注意力机制源于模仿人类分配注意力的机制，将注意力分配到重要区域，在计算机视觉中表现为对矩阵数据进行加权求和，简单而高效地突出关键信息。这种机制属于动态筛选必要信息的方式，它能够根据输入的重要性，自适应地为各个特征分配权重。注意力机制在许多与卷积神经网络相关的视觉任务中发挥了重要作用，例如目标跟踪、目标检测、语义分割等任务。本文研究的内容同样运用了多种注意力机制，因此将在本节中分别介绍空间注意力机制、通道注意力机制与时间注意力机制等常用注意力机制。

2.3.1. 通道注意力机制

在计算机视觉中，通道注意力机制属于常见的单一注意力机制。特征图通常由多个通道来表达目标的不同信息，但在特定场景中，对这些信息的侧重点存在差异。对所有通道进行相同权重的处理不仅增加了计算量，而且效果不佳。通道注意力机制在训练过程中自适应地为各个通道分配相应的权重，实现对通道的选择性加权，从而解决上述问题。

SE-Net^[49]是 Hu 等人提出的首个应用在计算机视觉中的通道注意力机制，它的核心点在于一个压缩和挤压激励模块，用于收集全局信息、捕获通道关系并提高表征能力。特征图的每个通道对应一种卷积核的卷积结果，即该通道代表了图像中某种特定模式的特征。通过增强或抑制不同通道的特征，可以实现在通道的降维。SE 模块分为两个部分：压缩模块通过全局平均池化来收集全局信息，激励模块通过全连接层和非线性层（ReLU 和 Sigmoid）来捕获通道之间的关系并输出注意力向量。最后，将输入特征与注意力向量相乘，完成输入特征在通道维度上的权重校准，进行注意力的分配。SE 模块的数学表达式如下所示：

$$s = F_{se}(X, \theta) = \sigma(W_2 \delta(W_1 \text{GAP}(X))), \quad (2-8)$$

$$Y = sX, \quad (2-9)$$

其中， σ 为 Sigmoid 函数， W_1, W_2 为可学习权重， δ 为 ReLU 激活函数，GAP 为全局平均池化， X 为输入特征图， Y 为输出特征图。其网络架构如图 2-13 所示：

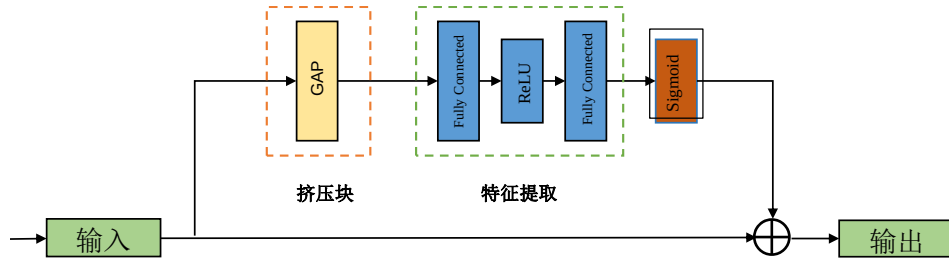


图 2-13 通道注意力机制

SE-Net 结构简单，计算量小，易于嵌入到其他网络中，具有较强的通用性，但挤压模块过于简单，难以有效且充分地提取全局信息，同时激活模块使用全连接层增加了计算负担。

2.3.2. 空间注意力机制

在计算机视觉中，一张图片中包含的空间信息是多元化的，但是并非所有信息都需要关注，例如，在单目标跟踪任务中，仅需关注目标所在区域，其他区域被认定为背景区域。因此，目标跟踪的分类任务本质上是一个二分类问题。空间注意力的作用就是自适应关注图像区域中目标所在的位置，从而提高跟踪器在处理分类任务时的准确性。空间注意力的具体实现以卷积操作为基础，根据输入生成二维空间注意力图，图上每个点代表输入各个部分的重要程度，即权重，然后将权重加权到输入图像上。图 2-14 是 STN^[50]空间注意力机制的结构示意图，它可以处理图像目标在旋转平移以及形变时无法识别的问题，空间注意力机制 STN 与通道注意力机制 SE-Net 存在明显差异。假设输入图像尺寸为 $H \times W \times C$ ，经过 SE-Net 处理后，得到一个大小为 $1 \times 1 \times C$ 的通道权重，并与输入图像相乘。而空间注意力机制得到的空间权重大小为 $H \times W \times 1$ 。两者的主要区别在于，SE-Net 强化通道信息，在处理类间问题时更有效，而 STN 更注重位置关系，能够隐式地确定类内物体定位。

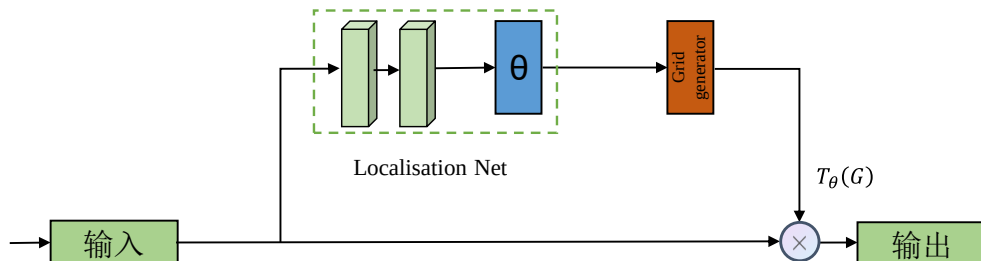


图 2-14 空间注意力机制

SE-Net 和 STN 虽然在各自环境下都表现优异，但是单一的注意力机制可能无法应对复杂情况下的跟踪器建模需求，从而导致跟踪效率低下甚至跟踪失败。为了有效应对这个问题，CBAM^[51]混合注意力机制被提出，如图 2-15 所示，该注意力机制通过结合通道注意力和空间注意力，为模型提供了更加全面和有效的特征提取能力。CBAM 首先关注通道信息再关注空间内哪些部分信息更为重要，这种双重注意力机制能够全面捕捉特征中的关键信息，并且结合了通道和空间注意力机制，能够同时关注通道和空间维度的特征关系，从而增强了网络对不同尺度特征的感知能力。

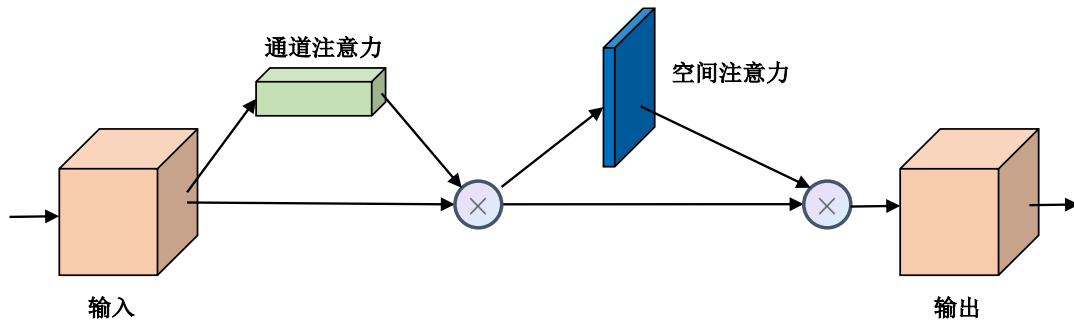


图 2-15 CBAM 注意力机制

2.3.3. 时间注意力机制

时间注意力机制是一种作用于时间维度上的注意力机制，通过大量数据训练出动态权重，一般应用于视频序列分析、自然语言处理、时序数据预测等领域，在跟踪任务中旨在捕捉序列中重要的帧图像，并在序列图像当中自动学习对当前任务贡献最大的时间节点图像，从而提升网络的全局准确性。而 CBAM 注意力机制则更加注重单张图像数据中目标的预测，对任务整体的宏观把控有所欠缺。如图 2-16 所示，BAM^[52]在自注意力机制基础上构建的时空注意力机制。该注意力机制首先通过 1×1 的卷积操作计算输入特征数据来获得 Q、K、V 三个张量，接着对三个张量进行展平操作。输出向量是通过 Value 向量进行加权求和得到的，其中每个 Value 向量的权重是通过 Query 与相应 Key 的关联函数计算得到的。在 BAM 时空注意力机制中，充分考虑了像素之间的时空依赖关系，从而有效应对单目标跟踪中可能出现的挑战。

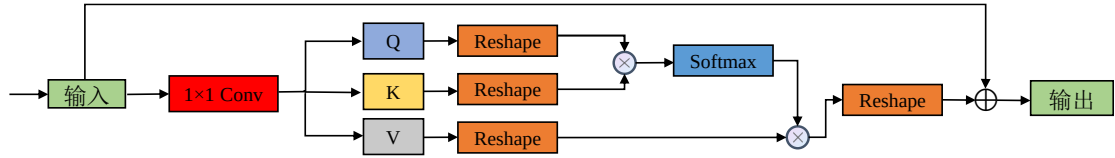


图 2-16 BAM 注意力机制

2.4. 测试数据集及评价标准

评价跟踪器模型的性能通常是通过将训练后的模型应用于测试程序，得到测试结果来进行的，本文将使用 OTB100^[53]、GOT10k^[54]、LaSOT^[55]、LaSO_{Ext}^[56]、TC128^[57]、VOT2016^[58]、NFS^[59]、UAV123^[60]等测试数据集进行测试，这些数据集均采集于现实环境，包含目标剧烈形变、旋转、遮挡、背景杂乱、视角转换等跟踪挑战，各个视频序列是综合评估跟踪器性能的重要数据，且上述测试集有独特的评价标准和特定的拍摄场景，使跟踪器的有效性与落地性高效提升。

OTB100 测试集和评价标准：OTB100 又被称为 OTB2015，是在 2015 年公开发布的提供给单目标跟踪研究者测试跟踪器性能的数据集。该数据集包含 100 个视频序列和 102 个被跟踪目标，其中 25 是灰度视频序列，其余为彩色视频序列。OTB100 中的视频序列多元化，包含但不限于户外场景、室内环境、运动与静止物体。此外，OTB100 数据集为每一个序列视频分类，如快速移动、剧烈形变、旋转、背景杂乱、消失视野、遮挡、运动模糊、相似物干扰等，并且各个视频序列之间可能存在交叉分类，可以更加全面的评估跟踪器的性能。视频序列的每一帧都存在手动标记的目标真实框坐标，用一维数组表示目标左上角横向与纵向像素点坐标和目标框的高和宽，这也是评估跟踪器性能的数据支撑。OTB100 有两个评价标准，分别是跟踪精度和成功率。跟踪精度是指在全部视频序列测试帧中，跟踪器预测的边界框中心点与人工标记真实边界框中心点的距离小于指定阈值（通常设为 20 像素）的序列帧数与所有序列帧数的比值，这一比值可以评估跟踪器跟踪目标的准确性。通过欧式距离计算中心点距离，具体表达如下：

$$\rho = \sqrt{(x_t - x_b)^2 + (y_t - y_b)^2}, \quad (2-10)$$

其中， x_t 、 y_t 代表跟踪器预测边界框的中心坐标点， x_b 、 y_b 表示目标框的真实中心点。

成功率的计算通常是通过跟踪器预测出的边界框与人工标注的目标框的重叠率来衡量的。重叠率大于指定阈值（通常为 50%）的序列帧数与所有序列帧数

的比值，作为跟踪成功的标准。这种指标可以用于评估跟踪器的鲁棒性，通常被称为交并比。具体公式表达如下：

$$\text{success} = \frac{|A^t \cap B^{gt}|}{|A^t \cup B^{gt}|}, \quad (2-11)$$

其中， A^t 为跟踪器预测的边界框， B^{gt} 为人工标记的目标框。当 success 超过 0.5 便定义跟踪成功。

GOT10k 测试集和评价标准：GOT10k 是中国科学院自动化研究所发布的大型单目标跟踪数据集，分为训练集和测试集两部分，并且两者并没有重叠部分。GOT10k 包含了 10000 个实际拍摄的移动物体视频片段和大于 150 万个手工标记的目标框。该数据集中存在丰富的运动类型物体，其中涵盖了 560 多种现实世界运动物体和 80 多种运动的形式。此外，数据集中还提供了物体的可见比例与运动类别。测试集包含 84 个物体类别和 32 个运动类别，共有 180 个视频片段，值得注意的是 GOT10k 测试集为了保证测试的公平性，没有选择公开真实标注的坐标值，而是进行在线高效评估，并且在 GOT10k 工具包中记录每一个序列视频的跟踪时间，用于评估跟踪器的帧率。在评价标准方面，GOT-10k 使用平均重叠率（Accuracy Overlap, AO）和成功率（Success Rate, SR）作为评价指标。AO 的计算如公式下所示：

$$AO = \frac{1}{N} \sum_{i=1}^N IOU_i, \quad (2-12)$$

和公式 2 中的 success 不同，AO 表示的是每一帧交并比的加权平均值。成功率与 OTB100 数据集中成功率图一致，都是评估跟踪成功的帧数在总帧数中的比例。如 $SR_{0.5}$ 表示阈值为 0.5 时的跟踪成功率， $SR_{0.75}$ 代表阈值为 0.75 时的跟踪成功率。

VOT-2016 测试集和评价标准：VOT2016 数据集包含了 60 个各种情景下的视频序列，和 OTB100 数据集中挑战场景相似。这些序列的待追踪目标包括人、车辆、动物等。每个视频序列都详细的给出了注释和标记，例如目标的初始位置和大小，以及序列中每一帧中的真实目标坐标数据。VOT2016 有三个评价标准，分别是 EAO（Expect Average Overlap Rate）、Accuracy 和 Robustness，其中 EAO 用于衡量跟踪器在给定测试序列上的性能，它通过计算一个视频序列的所有帧的平均重叠率来评估跟踪器对目标的定位精度与鲁棒性。Accuracy 是跟踪成功状态

下的平均重叠率。**Robustness** 用来评价跟踪器跟踪目标的稳定性，它和 **VOT** 系列测试集独特的失败重新检测机制相关，**VOT** 认为评测跟踪器时应该在跟踪丢失时检测为错误，并在错误发生的 5 帧之后对跟踪器重新初始化，将该帧图片作为初始帧重新跟踪，这样可以充分利用数据集。之所以是 5 帧而不是立即初始化，是因为错误帧可能是被遮挡以及视野消失的情况，立即初始化会再次跟踪失败，又因为遮挡一般不会超过 5 帧，所以存在这样的设定，**Robustness** 越大，效果越差，具体表达式如下：

$$\rho_R(i) = \frac{1}{N_{rep}} \sum_{k=1}^{N_{rep}} F(i, k), \quad (2-13)$$

其中， $\rho_R(i)$ 是 i 个跟踪器的平均鲁棒性， N_{rep} 代表重复次数， $F(i, k)$ 是第 i 个跟踪器在第 k 次重复中失败的次数。

VOT2019 使 **VOT** 竞赛在 2019 年推出的测试集，相比较 **VOT2016** 更新了部分视频序列，使跟踪场景更具有挑战性。

UAV123 测试集和评价标准：**UAV123** 测试集由低空无人机拍摄的 123 个视频序列组成。所有序列都使用手工标记的四维边框数组表示。该数据集是无人机跟踪系列数据集中的一种，其主要目的是评估基于无人机视角的单目标跟踪算法。和其它测试数据集不同的是由于无人机的拍摄角度较广，目标的视野占比相对较小。因此，该测试集更能有效评估跟踪器对小目标的跟踪性能。在评价指标上，**UAV123** 采用平均精度和成功率指标。

LaSOT 数据集和评价指标：**LaSOT** 是华南理工大学的研究团队提出的一个评估和测试单目标跟踪算法性能的大规模数据集。该数据集分为训练集和测试集且无交叉重叠部分，其目的是提供一个具有高挑战性的长序列和复杂场景下跟踪基准。**LaSOT** 涵盖了 70 个种类，总共 1400 个视频序列，其中 1120 个用于训练，其余用于测试，为了保证不同类型之间的平衡，使每个类别包含了 20 个视频序列，从而减少了测试结果存在类别偏差的可能性。**LaSOT** 数据集和 **GOT-10k** 数据集相比的一个显著特点是其跟踪视频序列的长度，长期的跟踪对跟踪器的鲁棒性要求是更高的，尤其是在目标消失视野并重新出现时。长序列的跟踪场景包含着更多的复杂变化，例如快速移动的目标同时存在遮挡和杂乱背景的场景，这些对于评估跟踪器鲁棒性提出了更加权威的标准。总而言之，**LaSOT** 数据集通过其大规模、长序列和类别均衡的特点，使其无论是在训练数据还是测试性能时都

是不可忽略的数据资源。在评价指标上其成功率与平均精度与 OTB100 数据集一致，但是 LaSOT 提出了归一化精度图，该评价指标是跟踪精度的一个变体，它将位置误差归一化到目标的尺寸上，目的是可以更准确地比较不同尺寸目标下跟踪器的跟踪性能，表示如下：

$$\begin{cases} W = \text{diag}(bb_x^{gt}, bb_y^{gt}) \\ P_{Norm} = \frac{1}{\|W(c^{pr} - c^{gt})\|_2} \end{cases}, \quad (2-14)$$

其中， bb_x^{gt} 是真实目标框的横坐标， bb_y^{gt} 是真实目标框的纵坐标； diag 是矩阵的对角化操作； c^{pr} 代表跟踪器预测的目标框的中心坐标， c^{gt} 是标注目标框的中心坐标； $\|\cdot\|_2$ 是欧式距离符号； P_{Norm} 代表归一化精度。

LaSOT_{Ext} 是 LaSOT 数据集的扩展补充版本，包含 15 个新类别的 150 个具有挑战性的长视频序列子集，其评价标准与 LaSOT 一致，不再做表述。

TC128 数据集和评价指标：TC128 是一个由 128 个彩色视频序列组成的数据集，其中包含了每个视频序列中目标的真实坐标和挑战场景的注释，特别关注在变形和光照变化条件下的追踪挑战，该数据集使颜色信息集成到跟踪当中，更加全面的了解跟踪器的性能。在评价指标上，TC128 采用与 OTB100 相同评价指标，即平均精度和成功率。

NFS 数据集和评价指标：NFS 数据集是专门为评估视频跟踪算法而设计的集合，包含 100 个快速移动的视频序列和九个挑战性属性。这些视频是使用高帧率设备录制的，旨在测试跟踪器处理快速移动目标的性能。NFS 有两个版本，NFS30 和 NFS240，本文使用了 NFS30 版本来测试跟踪器。在评价指标上，NFS30 采用与 OTB100 相同评价指标，即平均精度和成功率。

2.5. 本章小结

本章主要介绍了单目标跟踪相关理论基础，首先阐述了卷积神经网络的基础架构和相关概念，分析卷积、激活、池化、填充等基础模块的作用和推理公式。接着介绍了孪生网络在单目标跟踪中的应用并详细分析了 SiamFC 和 SiamBan 两个经典算法的创新点与逻辑。随后又介绍了多种注意力机制，其中包括通道注意力机制、空间注意力机制、混合注意力机制和时间注意力机制。最后介绍了本文后续实验中常用的多个测试数据集以及相应的评价标准。

第3章 基于先验信息和多层特征精确融合的跟踪网络研究

3.1. 引言

随着深度学习的快速发展,基于孪生网络的跟踪器在性能和速度方面都取得了长足进步。然而,面对物体剧烈变形的任务时,尤其是当目标的长宽比呈现极端比例时,大多数跟踪器容易丢失目标,导致跟踪失败,如图 3-1 所示。

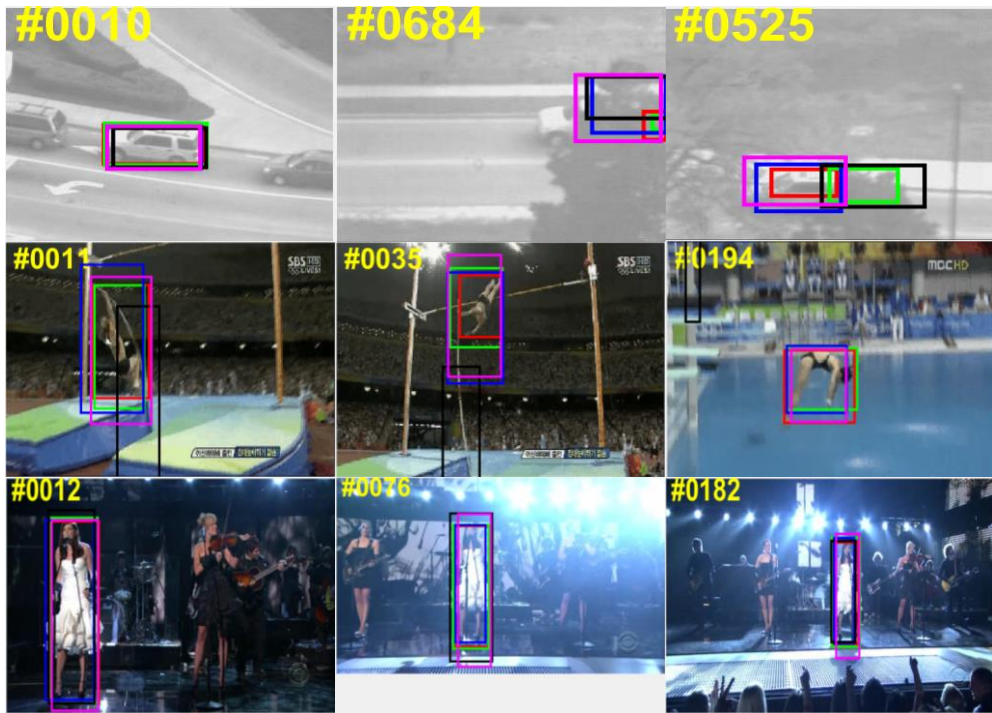


图 3-1 目标产生剧烈形变时跟踪器输出图例

本章节提出的 SiamPIP 可以有效解决这个问题。在跟踪过程中,由初始帧提供的目标信息具有高纯度,不受任何噪声干扰的特性。初始帧是整个跟踪任务中唯一的先验信息,利用初始帧可以过滤从主干网络中提取的模板特征中的杂乱信息。此外,传统的孪生网络模型往往忽略了分类和回归任务之间的差异。分类任务通常需要更深层次的语义信息,而浅层网络的特征更适合回归任务,因此需要精确地融合特征。面对严重变形的物体时,本文提出的模型可以准确地区分前景和背景,同时保持回归边界框的准确性。值得注意的是,该模型在面对运动模糊等挑战时也表现出色。SiamPIP 在 OTB100 上的性能指标达到了 0.707 (AUC),在 VOT2019 上达到了 0.340 (EAO),在 UAV123 上达到了 0.643 (AUC)。

本章节的主要贡献如下:

(1) 跟踪器 SiamPIP 关注目标跟踪中的先验信息并灵活利用掩码特征来区分背景干扰和上下文信息。这使得模板特征更加真实可信，并为后续帧的跟踪提供有效指导。本章节在此思想基础上提出 PI 模块。

(2) 提出了一种准确有效的特征融合方法，根据不同任务，利用不同深度的特征进行不同的比例融合，并提出 PF 模块。这比以往的特征融合方法更精确，不仅提高了跟踪成功率，还显著提升了精度。

3.2. SiamPIP 网络架构

本节首先在 3.2.1 节介绍基于先验信息和多层特征精确融合的跟踪网络主体框架，然后在 3.2.2 节介绍相关理论体系，3.2.3 节给出实验细节。

3.2.1. 先验信息和多层特征精确融合

SiamPIP 结构如图 3-2 所示，包含三个主要部分：特征提取、多级比例融合和预测头。特征提取部分与经典框架 SiamBan 类似，使用 ResNet50 对模板和搜索区域图像进行特征提取，生成相应的特征图。受非局部注意机制（Non-Local）^[61]的启发，对模板特征图 Z_f 和搜索区域特征图 X_f 进行线性映射和重塑操作，以计算特征中的相关性。掩码特征也经过类似过程，再次计算相关性。在共享网络结构参数中，模板特征和搜索区域特征之间缺乏有效的信息交互。因此，PI 模块旨在进行特征增强，以加强 Z_f 对 X_f 的引导。研究中发现，物体分类和边界坐标预测之间的差异性决定了两者所需的特征信息也不同。一些以 SiamCAR 为代表的跟踪器忽略了这一点，它们将多层交叉相关和交叉相关后的特征进行融合，然后将其发送到分类头和回归头，从而获得通道数为 2 和 4 的结果。为了减轻这一问题的影响，采用了 PF 模块，该模块逐渐减少回归任务中使用的多层输入比例，而分类任务中的情况则恰恰相反。这些改进使 SiamPIP 能够更好地处理目标急剧变形的问题。

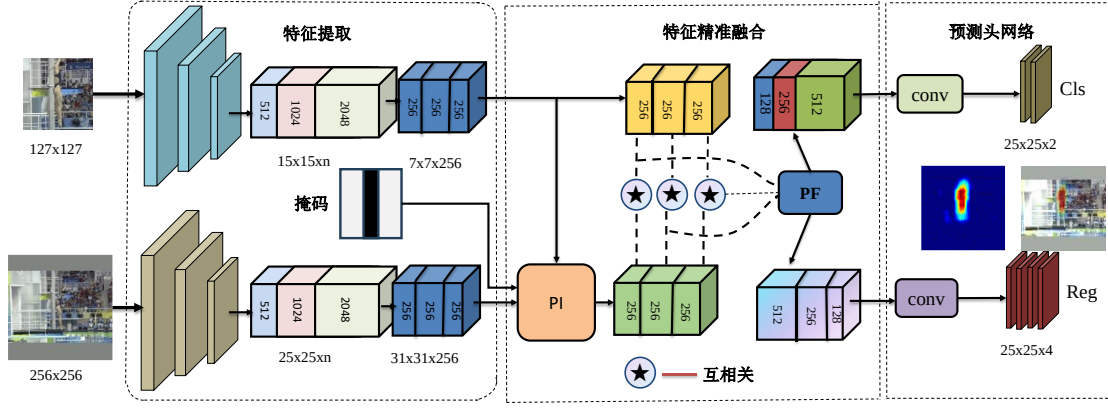


图 3-2 SiamPIP 主体结构

3.2.2. 多比例融合网络

孪生网络系列跟踪算法的独特之处在于网络结构参数的共享，在 SiamRPN 之后，目标跟踪通常使用分类和回归来预测结果。SiamBan 遵循了这种共享思想，其中特征图 P_{cls}^2 和 P_{reg}^4 是输入到头部的单个多层特征相互关联的结果。在目标跟踪中，分类任务只需要区分目标和非目标对，即进行二值分类。深层网络提取的特征包含了更多的语义信息，使得即使在目标发生剧烈变化时，也能有效区分目标，因此更适合二值分类任务。SiamRPN++ 提出了一种将 ResNet50 的 Layer2, Layer3 和 Layer4 的特征统一分辨率后都作为骨干网络输出方法，然而 SiamBan 使用无锚点预测头，它需要将这些多层特征融合在一起，如图 3-3 所示，SiamBan 将三个特征图中的元素求和后除以 3，获得这三个特征图的平均值。在此基础上，本文提出了将多层相互关联的结果进行融合串接的 PF 模块。根据先验信息有不同的融合比例。

$$\begin{aligned} P_{cls}^2 &= \text{Conv0}\left(\text{Cat0}\left([\delta(z)]_{cls} \star [\delta(x)]_{cls}\right)\right), \\ P_{reg}^4 &= \text{Conv1}\left(\text{Cat1}\left([\delta(z)]_{reg} \star [\delta(x)]_{reg}\right)\right), \end{aligned} \quad (3-1)$$

其中， $\star$ 代表 $[\delta(z)]_{cls}$ 和 $[\delta(z)]_{reg}$ 作为卷积核与 $[\delta(x)]_{cls}$ 和 $[\delta(x)]_{reg}$ 做卷积操作， δ 是 1×1 的卷积用于平滑处理特征， z 是模板图像特征， x 是搜索图像特征，由于跟踪的分类只需要对前背景二值分类， P_{cls}^2 是通道为 2 的特征的合成图， P_{reg}^4 表示目标的每个像素点到四条边坐标的距离。

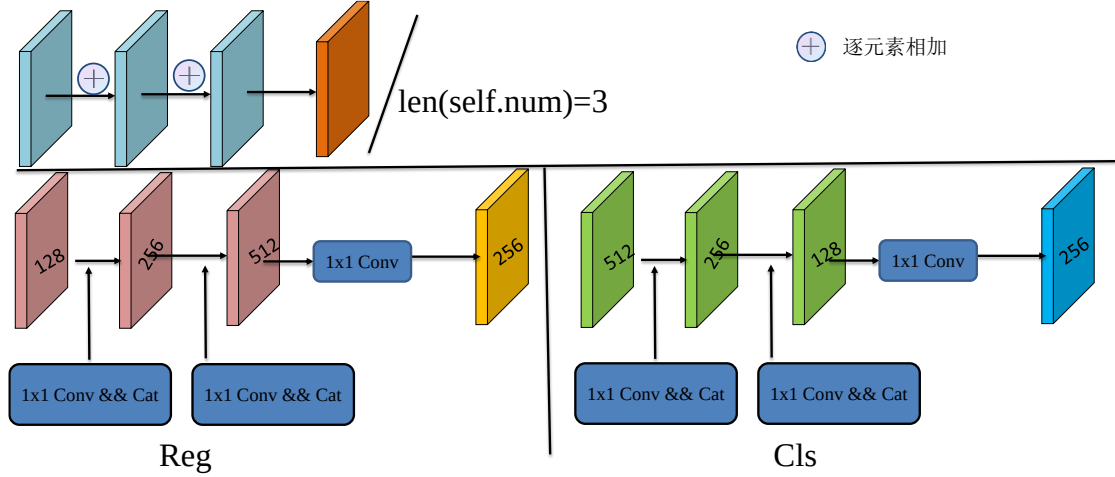


图 3-3 PF 结构

PF 模块的具体做法如图 3-3 所示。SiamPIP 将骨干网的三个输出层的通道数都调整为 256。在回归任务中按照 1: 2: 3 的比例融合三个特征图，在分类任务中按照 3: 2: 1 的比例融合调整通道后的特征图，在互相关运算后得到三个通道数仍为 256 的特征图。接下来，将按如下方式对三个特征图进行通道计数，并将它们融合到头部的输入 Cl_s 中：具体表达如下：

$$Cl_{s_{oc}} = Down\left(Cat\left(\left[\downarrow_0(F_2), \downarrow_1(F_3), \downarrow_2(F_4)\right]\right)\right), \quad (3-2)$$

其中， $Cl_{s_{oc}}$ 为融合的分类分支特征图， F_2 ， F_3 ， F_4 为三层模板特征。 Cat 在第一维（通道）上拼接多个特征图， $Down$ ， $\downarrow_0$ ， $\downarrow_1$ ， $\downarrow_2$ 都是通过卷积运算改变特征图的通道数（oc 为输出通道数）。当然，目标跟踪性能的评价标准离不开边界框的真实坐标，SiamRPN 预设了五个不同纵横比锚框，最后对边界坐标进行了微调，SiamBan 摒弃预设锚框的步骤，直接预测像素点与边界之间的像素距离，得到 Loc ，其中， Loc 为融合回归分支特征映射，与分类操作类似，不同之处在于它将通道数占比的大小进行了反转。本文采用 SiamBan 的无锚回归方法，通过下式得到预测框架 Loc ：

$$\begin{aligned} P_{x_1} &= P_i - d_l^r \\ P_{y_1} &= P_j - d_t^r \\ P_{x_2} &= P_i + d_r^r \\ P_{y_2} &= P_j + d_b^r, \end{aligned} \quad (3-3)$$

其中， P_i ， P_j 为像素点的 x ， y 坐标值， d_l^r ， d_t^r ， d_r^r ， d_b^r 为 Loc 的预测值， (P_{x_1}, P_{y_1}) ， (P_{x_2}, P_{y_2}) 为预测框左上角和右下角的坐标。

3.2.3. 基于先验信息的特征增强

SiamAttn 在非局部注意机制的启发下,利用自注意和交叉注意增强双输入特征,在保证模板图像的纯粹性的同时,增强模板与搜索区域内信息的交互性。Non-Local 运算直接计算两个位置之间的关系,可以快速捕获长距离依赖关系,表示为:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} h(x_i + x_j) g(x_j), \quad (3-4)$$

其中, $h(x_i + x_j)$ 计算 i 与 j 之间的相似度, i 表示特征图中的某个位置, j 表示所有可能位置的索引集合。 x 为输入特征映射, $g(x_j)$ 计算特征映射在 j 位置的表示, $C(x)$ 为归一化过程,用于确保变换前后整体信息的一致性。

SiamPIP 将 Non-Local 扩展到跟踪器模型,如图 3-4 所示,使 Z 、 $Mask$ 和 X 分别为线性映射后相似度计算的输入,并在最后和 X 再次求和,以保证模型的前向性,其表达式为:

$$\begin{aligned} s &= \text{Softmax}(Z_f \otimes X_f) \otimes Mask_f \\ X_{0_f} &= X \oplus (s \odot X), \end{aligned} \quad (3-5)$$

其中, X_{0_f} 为 PI 模块的输出, X 是搜索区域模块线性映射后的特征图的结果, X_f , Z_f 是搜索区域与模板区域特征线性映射后的变形结果, $Mask_f$ 是掩模特征线性映射后的结果。另外, $\otimes$, $\oplus$, $\odot$ 分别代表矩阵乘法、矩阵相加和点乘。

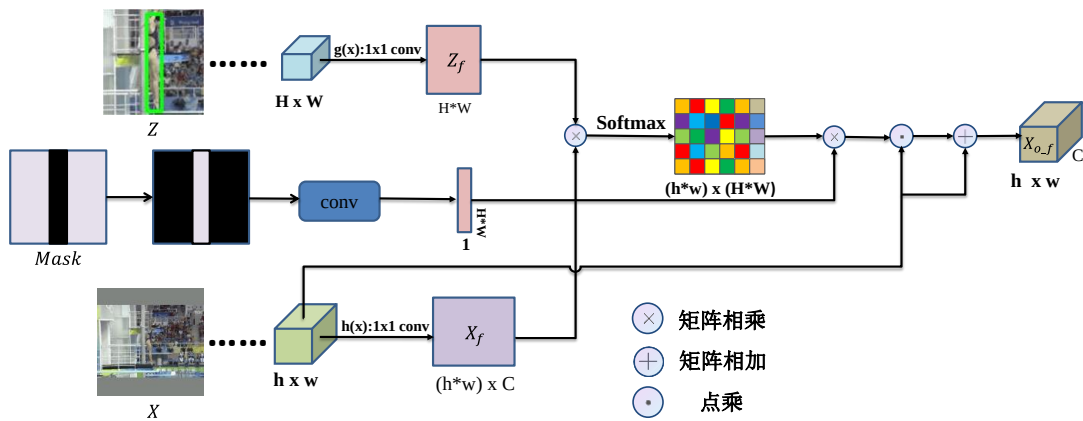


图 3-4 PI 结构

3.3. 实验方法与结果分析

本节将对 SiamPIP 跟踪算法进行训练与测试。实验中所使用的数据集及客观

评价标准已在 2.4 节中详细介绍。算法的训练硬件设备和超参数设置将在 3.3.1 节中说明。SiamPIP 的实验结果与分析分为两部分：在 3.3.2 节，对该算法在 OTB100、VOT2019 和 UAV123 三个公开数据集上整体评估以及各挑战属性的结果进行了分析；随后，3.3.3 节将详细介绍模块消融实验，逐一验证本文提出跟踪器改进点的优越性；最后，3.4 节将对实验结果进行总结。

3.3.1. 实验环境介绍

本章提出的 SiamPIP 算法使用在 ImageNet^[62]上训练的 ResNet50 作为主干网络初始化，并在 ResNet50 的第 3 层和第 4 层分别使用空卷积以扩展感受野，其空率设置为 2 和 4。在训练细节方面，模型共训练了 20 个 epoch，为了减少训练时间并提高训练效率，在前 10 个 epoch 中，冻结了主干网络的参数，仅训练分类回归网络的框架头；在后 10 个 epoch 中，解冻了 ResNet50 的后三层，并与自适应框架头一起训练。SiamPIP 的优化算法采用随机梯度下降方法，权重衰减和动量分别设置为 0.001 和 0.9，在前 5 个 epoch 中使用从 0.001 到 0.005 的热身学习率，并在接下来的训练中使用从 0.005 到 0.00005 的指数衰减。模型批量大小设置为 28，基础学习率为 0.005。训练具体配置见表 3-1，前 10 轮训练耗时 30 小时，后 10 个 epoch 训练耗时 72 小时。使用了六个训练数据集，分别是 GOT-10K、COCO^[63]、LaSOT、ImageNet VID、ImageNet DET 和 Youtube-BB^[64]。每个训练轮次包含 100 万张图像。模板图像的尺寸设置为 127x127，搜索图像的尺寸设置为 255x255。所有图像在训练之前都进行了数据增强。

表 3-1 SiamPIP 跟踪器实验环境配置

软/硬件名称	具体型号
显卡	RTX 2060 12G
处理器	I5-12400KF
内存	16G DDR4
存储硬盘	1TB 固态硬盘
操作系统	Ubuntu 20.04
开发环境	Pycharm
编程语言	Python 3.7
深度学习框架	Pytorch 1.3
统一计算设备架构版本	Cuda10.2

3.3.2. 实验结果

对基于先验信息和多层特征精确融合的跟踪模型进行实验分析,选取网络训练后效果最好的模型,在 OTB100、VOT2019、UAV123 三个常见的测试数据集上与其他跟踪器进行横向对比。接下来,首先展示 SiamPIP 在 OTB100 测试集上的结果,并详细分析在 OTB100 测试集中各种挑战的表现,随后在 VOT2019 和 UAV123 测试集上分析该模型的跟踪结果。

SiamPIP 在数据集 OTB100 上的测试结果:首先利用 OTB-toolkit 工具包测试包括 SiamPIP 在内的各先进跟踪器的跟踪结果,然后在 Matlab 上通过比较预测框与真实框之间的交并比与中心点距离,得出跟踪器的跟踪精确度和准确率。评价标准的具体方法已在 2.4 中说明。如图 3-4 所示, SiamPIP 在 OTB100 上的跟踪精度表现优异,并与 SiamRPN++、SiamFC、Ocean^[65] (离线)、SiamBan、Siamrn^[66]、SiamCAR、SiamDWfc^[67]、DaSiamRPN、ECO^[68]、SiamRPN 等跟踪器进行了对比。SiamPIP 在所有对比跟踪器中表现最好,相比 SiamBan (0.696) 的改进非常显著。此外, SiamPIP 在准确性评估中排名仅次于 Siamrn,并且在多个评估指标上超越了 SiamBan。

通过使用 OTB 评估工具,能够详细评估多种不同的挑战,包括剧烈变形、小目标跟踪等。根据图 3-5, SiamPIP 在处理目标大小变化挑战时取得了 0.920 的精准度,远超 SiamBan (0.910) 和 Siamrn (0.911)。令人惊讶的是,从图 3-6 可以得出 SiamPIP 在处理平面旋转 ($0.721 > 0.717$)、低分辨率 ($0.727 > 0.719$) 和目标形变 ($0.677 > 0.662$) 等挑战时也取得了最高的结果(括号中为 SiamPIP 与 SiamBan 的成功率对比图)。这一结果主要归因于 PI 和 PF 模块的结合。进一步分析表明,大小变换和目标旋转的真实位置在帧与帧之间产生了较大的变化,尤其是在物体宽高比极端的情况下,由于尺寸平滑操作的存在,模型在短时间内无法应对宽度和高度的剧烈变化。然而, PF 使用一个浅层网络来处理回归任务,对位置敏感,使得预测的宽度和高度变化受平滑项的影响。

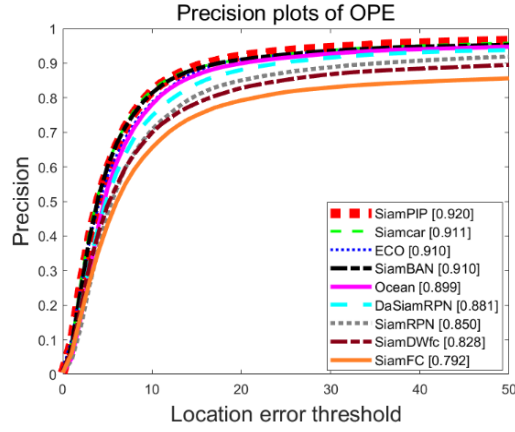


图 3-5 SiamPIP 在 OTB100 上的跟踪结果

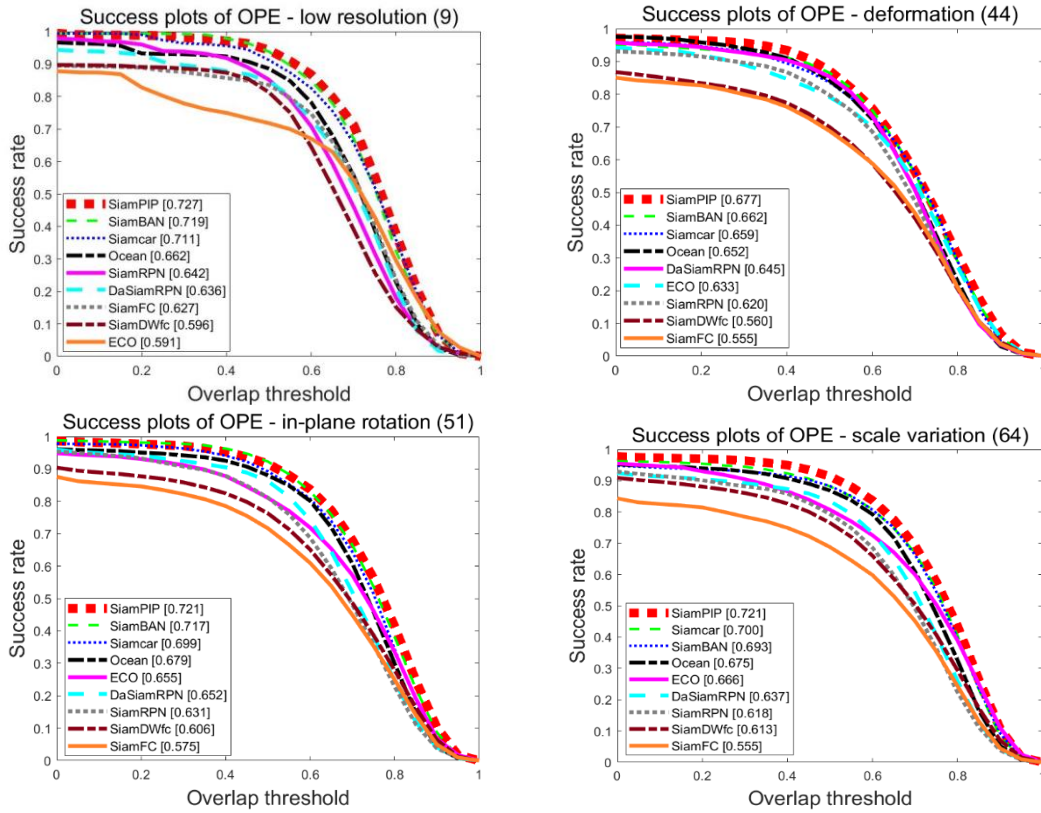


图 3-6 SiamPIP 在 OTB100 上各挑战的跟踪结果

SiamPIP 在数据集 VOT2019 上的测试结果：与 OTB100 相比，VOT2019 包含了更严重和更具挑战性的视频变形。VOT2019 的评估指标更为严格，这类数据集的三个度量分别是 EAO、准确性和鲁棒性，EAO 和准确性得分较高，鲁棒性得分较低，表明跟踪器表现更好。为了详细的展示跟踪器在 VOT2019 上的表现，首先利用 VOT-toolkit 在 Matlab 中将 SiamPIP 和 SiamBan、SiamCAR、ATOM 等跟踪器在 EAO 指标上进行比较，并绘制出气泡图直观展示性能差距。另外还利用 Python 细致的评估跟踪器在精准度和召回率上的结果。通过图 3-7 可以直

观看出 SiamPIP 的 EAO 指标是高于 SiamBan、SiamCAR 等跟踪器。相较于基础跟踪器提升了 1 个点，这得益于提纯初始帧的目标特征与精准融合特征操作。表 3-2 以数值形式显示了比较结果，其中 EAO、准确性和鲁棒性的前三个性能分别用红色、绿色、黄色、绿色和黄色标记标记。从表 3-2 可以看出，模型在 EAO 度量中排名第一，最低的鲁棒性值表示 SiamPIP 在跟踪过程中是稳定的，精度也略低于 ATOM。

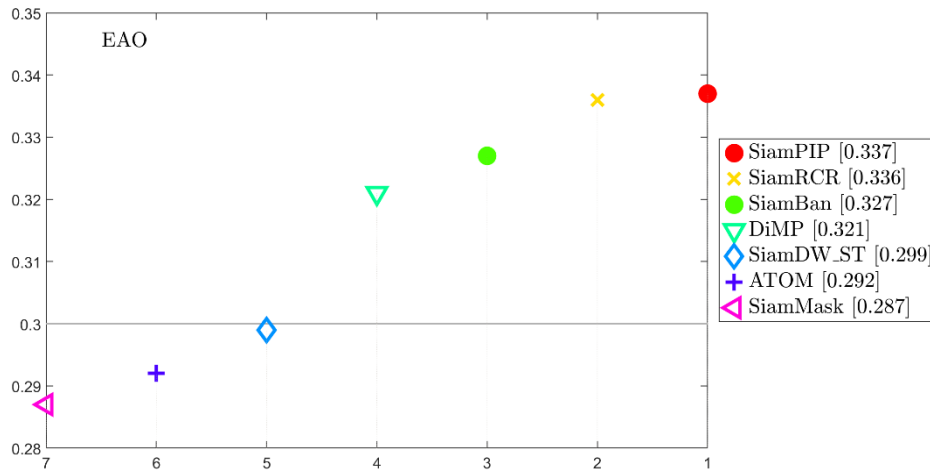


图 3-7 SiamPIP 在 VOT2019 中 EAO 测试结果

表 3-2 SiamPIP 在 VOT2019 中测试结果

Trackers	EAO	Accuracy	Robustness
SA_SIAM_R	0.252	0.563	0.507
SiamMask	0.287	0.594	0.461
ATOM	0.292	0.603	0.411
SiamDW_ST	0.299	0.600	0.467
DCFST	0.317	0.585	0.376
DiMP	0.321	0.582	0.371
SiamBan	0.327	0.602	0.396
SiamRCR	0.336	0.602	0.386
Ours	0.337	0.592	0.321

SiamPIP 在数据集 UAV123 上的测试结果: 图 3-8 是跟踪器在测试集 UAV123 上的运行结果，包含了跟踪成功率与准确性，UAV123 是一个空中视频基准和数据集，所有序列都完全用直立边界框注释，与 OTB100 相比，它存在更多的目标完全消失，包含常见的小目标挑战，长视频鲁棒性挑战。与著名的跟踪器如

SiamBan、SiamCAR、DiMP、Ocean 等相比，SiamPIP 跟踪器获得了先进的分数。本文提出的模型相比与基准跟踪器在成功率和准确率方面分别提高了 1.2%和 1.4%。

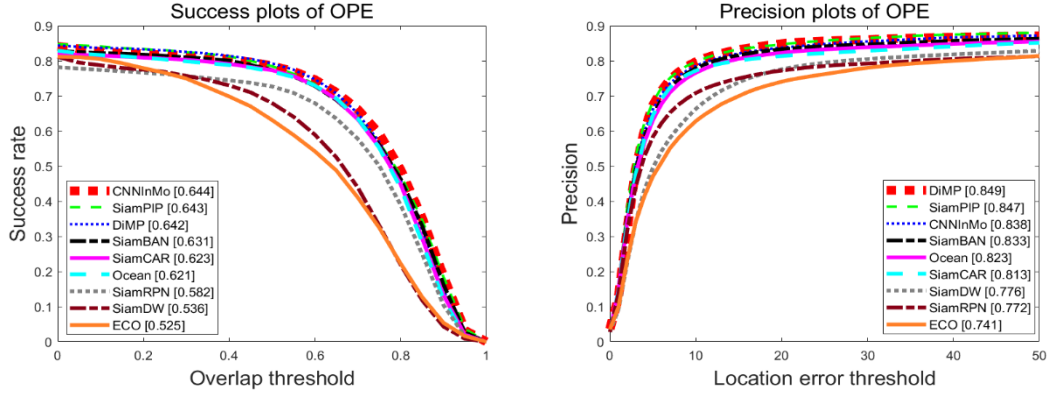


图 3-8 SiamPIP 在 UAV123 中测试结果

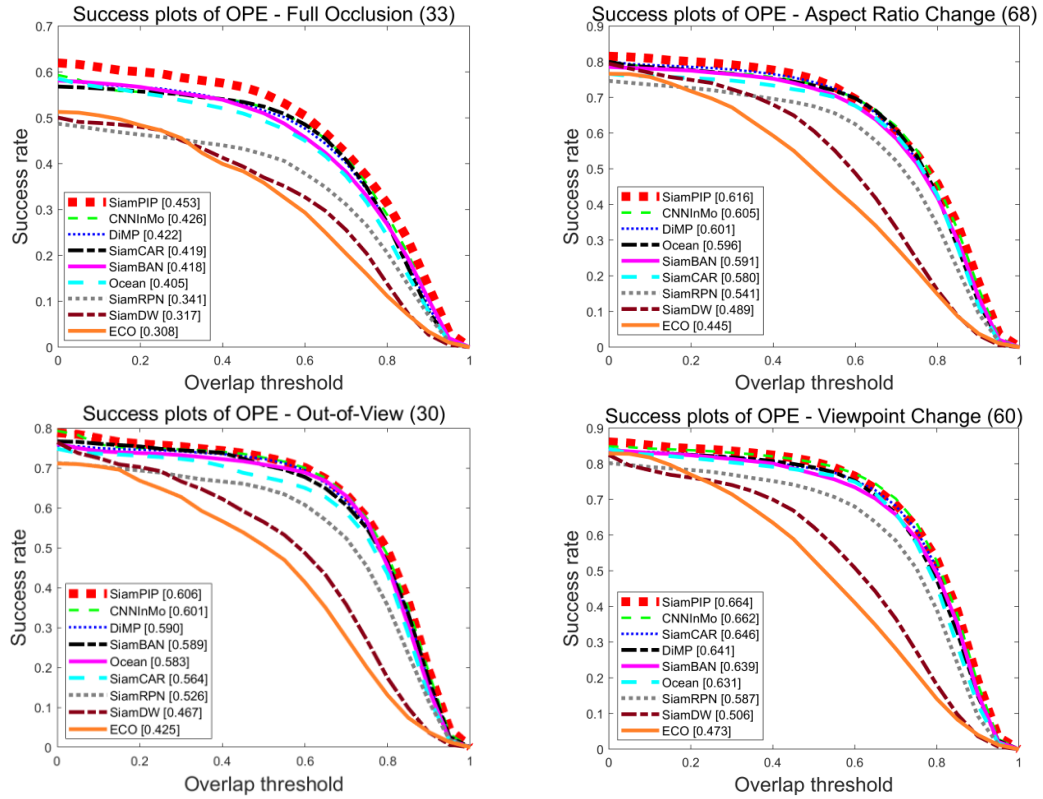


图 3-9 SiamPIP 在 UAV123 中不同挑战测试结果

本文在 UAV123 的各个挑战属性中进行性能测试，从图 3-9 可以看出，SiamPIP 在处理完全遮挡、纵横比变化、超出视野范围、视角转变等挑战时效果最佳，这得益于 PI 与 PF 模块的融合交互，使模型的鲁棒性得到有效提升。

通过在以上三个公开数据集上进行详细的测试实验可以验证跟踪器 SiamPIP 是表现优异的。其次，本文在 3.3 节中展示 SiamPIP 与 Baseline 跟踪器

的详细对比实验

3.3.3. 消融实验

本节所提出的 SiamPIP 是在 SiamBan 的基础上进行的后续工作，遵循了 SiamBan 的训练策略与工作原理。在孪生网络的双输入基础上，加入了先验帧掩码输入，以提升目标特征图的纯度。其次，预测头网络可以分为回归与二分类两个任务，但它们却使用相同的图像特征，这是不合理的。分类通常需要较深层次的特征，而回归任务则可以利用浅层特征来更准确地确定回归坐标值。因此，本文提出了一种根据不同任务采取不同比例精准融合特征的方法，以提升跟踪器的性能。

为了确保所提出的各个组件对 SiamPIP 具有正向增益，基于 SiamBan 跟踪器进行了消融实验。如表 3-3 所示，首先，在 SiamBan 的基础上仅使用 PI 组件，命名为方法 1；然后，仅使用 PF 组件，命名为方法 2；基准跟踪器命名为方法 3；PI 与 PF 联合使用命名为方法 4。如表 2，单独使用 PI 模块，可以将 SiamBan 在 OTB100 中的成功率从 0.696 提高到 0.701，准确率从 0.910 提高到 0.916；使用 PF 模块时，成功率从 0.696 提高到 0.698；PI 和 PF 联合使用时，成功率和准确率分别提高了 1.58%和 1%，从 0.910 提高到 0.920。为了保证实验的严谨性，消融实验分别在 OTB100 和 UAV123 两个测试集上进行。

表 3-3 SiamPIP 各模块消融实验

Tracker	PI	PF	AUC(OTB100)	Pre(OTB100)
1	√	-	0.701	0.916
2	-	√	0.698	0.913
3	Baseline		0.696	0.910
4	√	√	0.707	0.920

如表 3-4 所示，本文的模型在应对目标显著转移的挑战下得到显著提高，将 SiamBan 的成功率提高了 4.04%（从 69.3%提高到 72.1%），准确率提高了 3.10%（从 90.3%提高到 93.1%）。进一步分析表明，PF 模块是提高变形困难处理能力的主要部分，这是因为深度和浅层特性的特定应用保证了跟踪器的鲁棒性。

表 3-4 SiamPIP 在 OTB100 指定挑战下的消融实验

Tracker	PI	PF	AUC_SV(OTB100)	Pre_SV(OTB100)	AUC_SV(UAV123)
1	√	-	0.696	0.906	0.615
2	-	√	0.715	0.921	0.624
3	Baseline		0.693	0.903	0.613
4	√	√	0.721	0.931	0.627

3.4. 本章小结

本章首先详细分析了提纯模板特征和根据具体任务实施多比例精准融合特征的必要性和具体原理，其次在 OTB100、UAV123、VOT2019 三个公开测试集中验证跟踪器的性能，并在 3.3 节中对所提出的 PI 和 PF 部件进行消融实验，实验结果表明去除模板中的杂乱背景信息和多比例融合骨干网络中的三层深浅特征对跟踪器的跟踪效果有明显提升。

第4章 基于时间相关性的轻量化跟踪网络研究

4.1. 引言

尽管许多现代跟踪器表现出令人满意的性能，但只有少数跟踪器深入分析目标的历史状态。STMTrack^[69]尝试通过将多个历史帧与当前帧融合，进行全局匹配在搜索图像踪定位目标，但计算需求过高，限制了其广泛应用。SiamBan 和 SiamCAR 主要解决初始帧与当前帧之间的匹配问题，通常包括三个主要组件：

（1）采用相同的骨干网络从模板图像和搜索图像中提取特征；（2）通过不同的交叉相关技术定位目标，在两个输入之中促进信息交换；（3）利用预测模块确定目标的精确位置。然而，当目标发生显著变形或遭遇严重遮挡时，依赖初始帧中的静态信息可能会导致跟踪失败。图 4-1 对 SiamGPF 在特定挑战下的表现与其他模型进行了比较分析。值得注意的是，SiamR-CNN 通过重新检测策略以应对跟踪漂移，UpdateNet 提出动态更新机制以增强模型的鲁棒性，SiamGAT 利用模板先验数据增强模板和搜索图像之间的局部信息交换。尽管这些方法在某些方面有效，但它们也增加了计算负担，并且面对其他挑战时（如相似物体干扰和背景杂乱）表现较差，说明其泛化性存在局限。

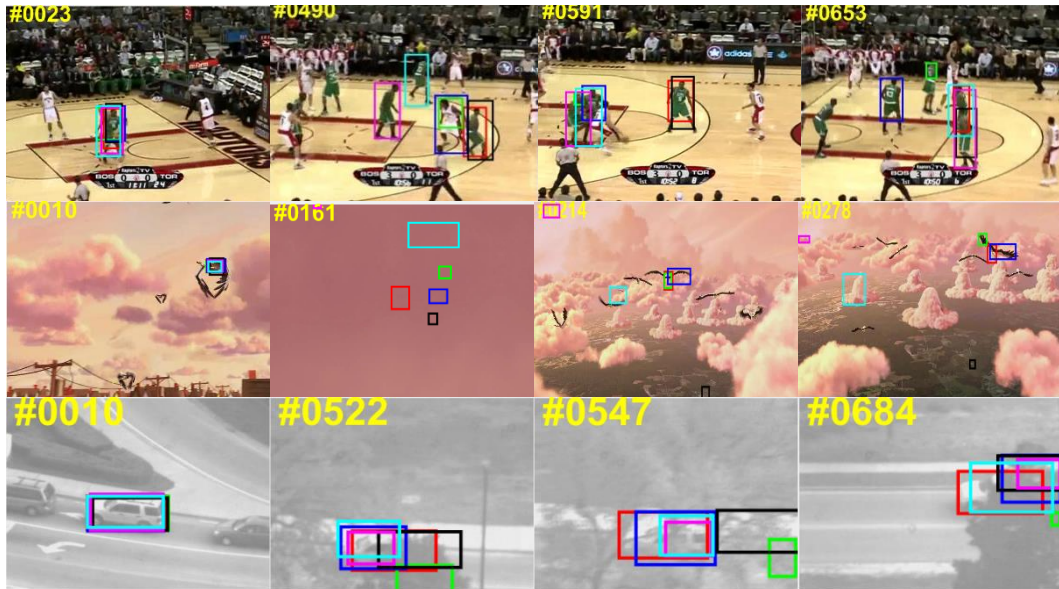


图 4-1 单目标跟踪中的挑战

为了克服上述挑战并保持模型的简洁性，本研究提出了一种结合时间信息的视频跟踪模型。与传统方法仅依赖模板帧不同，SiamGPF 避免了复杂的更新或

重新检测机制，它通过引入历史帧来补充模板信息，并利用图注意力机制实现模板与搜索图像之间的局部信息交互，通过非局部注意力机制的变体在初始帧和历史帧之间实现全局信息交互。提供了两种方案来选择先前的帧序列：使用当前帧的前一帧作为训练和测试的第三个元输入，或者在跟踪完成的图像帧中挑选历史帧。区别在于，训练时选择搜索图像的前一帧作为输入，而测试时可以选择一个在指定范围内具有最高质量的特征图作为输入，也可以选择当前帧的前一帧图像。在训练中，前一帧是从图像对中随机采样的。然而，在测试过程中，在指定范围内选择具有最高质量的特征图作为输入。关于这些策略的详细描述见第 4.3.4 节。

本章节的主要贡献如下：

1、开发了一种创新的先前帧匹配模块，结合基于图注意力机制的局部匹配策略(GM)、基于像素点的全局匹配策略(PM)和融合 GM 与 PM 的模块(FGP)，提出 SiamGPF。通过利用目标的历史状态作为辅助信息，并采用特定的融合策略，该模块能够指导跟踪器对当前帧中目标进行精确定位，从而丰富了时间信息在离线跟踪器中的应用。

2、提出了一种用于推理阶段的先前帧动态选择器(FDS)，可以在指定范围内筛选出一个高质量的单一历史帧作为骨干网络的第三个输入。此方法避免了在线更新策略，同时确保跟踪过程中的鲁棒性。

3、SiamGPF 在多个视觉跟踪基准测试中表现出色，包括 OTB100、LaSOT、LaSOT_{Ext}、UAV123、NFS、TC128、GOT-10k 和 VOT2016。与基准模型相比，本文提出的方法取得了显著的性能提升，并在 GOT-10k 数据集上达到了 212 FPS 的运行速度，证明了其高效性和实用性。

4.2. SiamGPF 网络架构

本节首先在 4.2.1 节中介绍 SiamGPF 跟踪器的总体结构设计，在 4.2.2 节和 4.2.3 节中分别分析 GM 部件与 PM 部件的组成与作用；接着在 4.2.4 节中详细介绍如何融合 PM 与 GM 模块，最后在 4.2.5 节中对测试流程中的先验帧选择策略进行分析。

4.2.1. 基于伪先验信息的网络框架设计

SiamGPF 采用了三元输入机制，隐式替代模板更新策略，避免了在推理过

程中重新计算特征,从而节省了大量的计算资源。这种方法不仅简化了时间数据的融入,还提高了跟踪算法的效率,为实时跟踪应用提供了新思路。模板与搜索图像匹配的任务依然是孪生网络跟踪器研究中的核心主题。**SiamFC** 是应用相关性计算的先驱,采用模板特征作为内核与搜索图像特征逐通道卷积,最终得到统一的单通道特征图。**SiamRPN++**通过集成深浅特征,增强了网络的语义能力,并通过实现深度卷积相关来缓解交叉相关过程中因特征维度增加而导致的参数不平衡问题。图注意力机制批判了传统的全局搜索方法,因为它未能有效处理目标纵横比的变化,因此嵌入图注意力模块,在目标和搜索区域之间建立局部连接,从而改善对这种变化的适应能力。**PG-Net** 首创了像素到全局的相关性概念,通过将模板特征划分为 1×1 的空间内核和通道来缩小匹配范围,有效减少背景噪声的干扰。**SOTA** 提出了细粒度显著性挖掘模块,能够捕捉目标中对跟踪具有判别性的局部显著性,并提出显著性关联建模模块,将捕捉到的显著性区域关联起来,从而学习模板和搜索图像之间的有效全局相关性,但该方法的计算复杂度仍然是一个重要的障碍。**AutoMatch** 利用 NAS 技术提出了一种概念上简单的算法,使用二进制通道操作自动设计匹配网络,并将学习到的匹配网络集成到基线跟踪器中,取得了显著的性能提升。

受到这些前期研究的启发,**SiamGPF** 在当前帧采用局部匹配计算,提高了跟踪器对目标外观变化的响应能力,同时对历史帧应用全局匹配方法,提供了当前场景所需的上下文信息。这种双重策略旨在将局部匹配的精度与全局匹配的泛化性相结合,旨在实现更加可靠的跟踪精度。

图 4-2 描述了框架的三个主要组成部分:特征提取网络、历史帧融合网络和预测头网络。与传统跟踪器不同,**SiamGPF** 在训练阶段省略了 ResNet50 的第 4 层,仅利用第 3 层特征进行后续推理过程。这一战略性修改对于实现 212 FPS 的运行速度至关重要。骨干网络采用共享的特征提取结构和通道调整策略处理三元输入,包括模板图像、搜索图像和记忆帧图像,以促进模型的高效性能。历史帧融合网络集成了三个新颖的组件:GM、PM 和 FGP。本文认为当前帧与历史帧所需的信息存在差异,因此使用了不同的匹配网络,详细的讨论和分析请参见第 4.3.4 节。预测头网络包括分类分支和回归分支,采用无锚点方法来简化回归任务的超参数设置,从而提高跟踪结果的稳定性和准确性。此外,图 4-2 中的虚线表示在测试阶段,搜索区域 b 是由 FDS 处理后的搜索区域得出的,不需要重新

计算,而在训练阶段,搜索区域被识别为模板图像和搜索图像之间的随机帧图像。

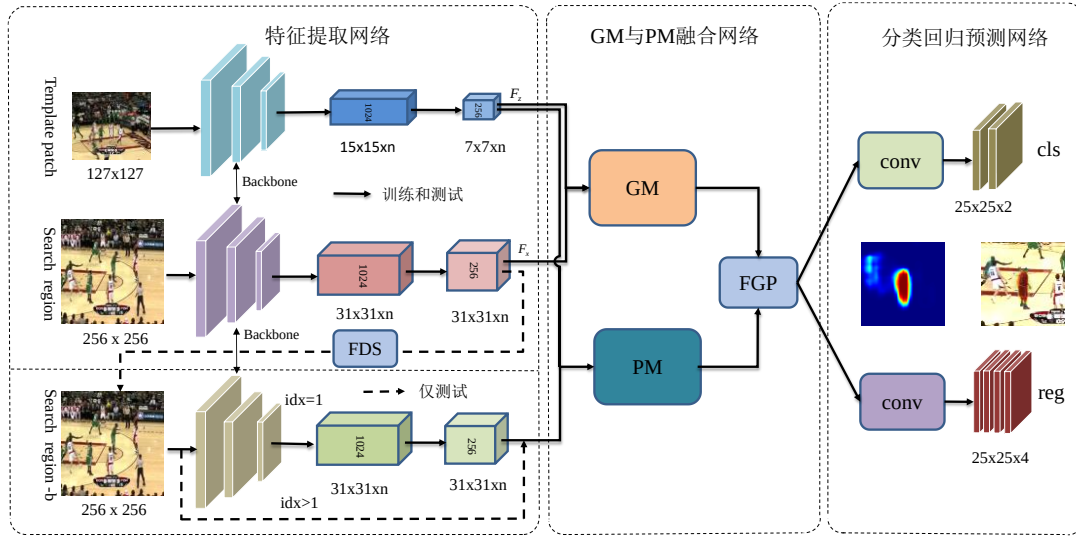


图 4-2 SiamGPF 主体结构

4.2.2. 基于图注意力机制的局部匹配策略研究

匹配网络架构一直是 Siamese 跟踪器研究的核心。由骨干网络生成的特征图 $F_z \in \mathbb{R}^{H_z \times W_z \times C}$ 和 $F_x \in \mathbb{R}^{H_x \times W_x \times C}$ 需经匹配网络处理,生成相似性图 P 。SiamFC 是最早提出相关性计算方法的先驱之一,该方法通过执行通道卷积并随后进行加权平均来完成。具体的公式如下:

$$P = \varphi(F_z, F_x), \quad (3-6)$$

特征的复杂性和全面性表明,单通道相关性计算或原始相关性计算方法不足以准确地描述目标和搜索区域之间的关系。许多知名的跟踪器选择直接将多通道相似度图输入到预测头结构中,以保持得分图的多样性。SiamGAT 认识到目标和搜索区域之间局部对应关系的重要性,通过集成图注意力机制,构建了一个交叉相关模块。AutoMatch 通过应用 NAS 机制自定义匹配网络架构,吸引了众多关注。该方法新颖地使用两个独立的匹配网络分别处理分类和回归任务,从而实现得特征图的拼接。受到这些算法的启发,本文认为初始帧和当前帧在跟踪过程中扮演着不同的角色,因此每个帧都需要不同的匹配网络。对前一帧和当前帧之间相互作用的深入研究表明,记忆帧通常起辅助作用,而当前帧的匹配网络的有效性对于跟踪器的性能至关重要。SAOT 引入了一个显著性挖掘模块,识别出最明显的局部点作为显著点,并利用关联模块将显著区域聚合在一起。这种方法不仅采用局部策略增强模型对目标变形的鲁棒性,还通过集成图注意力模块合并多

个高质量的补丁，从而增强全局信息，提升跟踪精度。与最先进的方法不同，SiamGPF 将局部和全局匹配技术结合在多输入范式中。研究表明，对于当前帧，局部策略能够提高模型的精度；而对于前一帧，全局策略则提高了跟踪器的成功率，具体实验结果详见第 4.3 节。

跟踪器 SiamFC、SiamRPN++、SiamBan、SiamCAR、STMTrack 等采用了整体策略进行跟踪。SiamGPF 对前一帧的功能定位是互补的，利用全局匹配网络区分初始目标定位范围和搜索块中的非目标区域，从而判断后者是背景杂乱还是有效的上下文信息。非局部模块被用来捕捉长距离依赖关系，计算图像特征的时间和空间位置信息，同时计算成本相对较低。

4.2.3. 基于自注意力的全局匹配策略研究

受非局部模块的启发，如图 4-3 所示，通过计算 F_z 和 F_x 内每个像素点的相似度来完成模板和搜索区域之间的匹配过程。为了确保信息的纯性，避免对输入应用线性转换。相似度计算函数采用高斯函数，具体表达式如下：

$$f(x_i, x_j) = \exp(x_i^T, x_j), \quad (3-7)$$

其中， $f(x_i, x_j)$ 是式 3-4 中相似性计算的具体形式，T 代表转置。其中单个元素相似性计算公式如下：

$$\alpha_{ij} = \frac{\exp[(F_i^z \otimes F_j^x) / o]}{\sum_{\forall k} \exp[(F_k^z \otimes F_j^x) / o]}, \quad (3-8)$$

其中， F_i^z 代表模板特征图上像素索引， F_j^x 代表搜索特征图上像素索引， α_{ij} 是相似性矩阵中的一个元素， o 是比例因子，用于在计算时保持数值的稳定性，通常设置为 $\sqrt{c}$ ， c 是特征通道数。 $\otimes$ 是矩阵法乘。

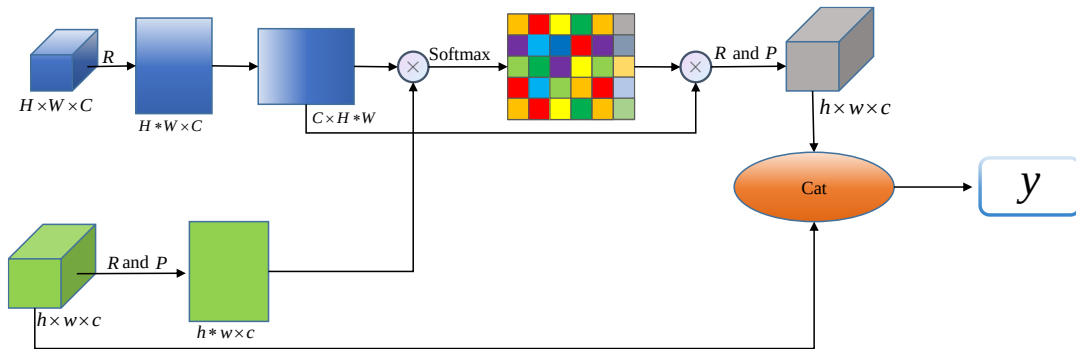


图 4-3 PM 模块结构

由于先前帧的目的是表示目标的历史状态，因此在做相似性比较时应保留特征的原始信息。本文将 α_{ij} 作为权重图与 F_j^x 在通道层次上拼接起来。其方式表达如下：

$$Y = \text{Cat}(F_i^z, (F_x)^T \otimes \alpha), \quad (3-9)$$

其中， Cat 表示相同尺寸的特征图在通道层次上拼接， Y 是最终得到的相似性图，该图随后被送入预测头网络进行最终的位置估计。

在目标跟踪领域，目标变形和遮挡等挑战普遍存在。PG-Net 提出的局部匹配策略将目标划分为一系列小块，分别计算每个小块的相关性，并最终整合这些计算结果以完成匹配操作。传统的跟踪器通常对所有小块赋予相同的重要性，而研究发现，具有显著代表性的小块对于高效跟踪至关重要。当前帧是跟踪器确定目标位置的主要信息来源，其匹配方法对模型的有效性具有决定性影响。与前几帧不同，当前帧需要克服大部分跟踪挑战，并专注于目标的结构本质，从而促进从模板图像特征到搜索图像特征的精细细节传递。最先进的方法利用图注意力网络 GAT 来连接多个显著点，从而完成从模板到搜索区域特征的局部匹配过程，并在多个基准测试中取得了优异的结果。

由于显著点挖掘操作的复杂性，SiamGPF 并不致力于找到这些关键点，一个重要的原因是先前帧的存在提供了目标的全局信息，这可以补充那些缺乏判别能力的 patch，从而提升模型的鲁棒性。图注意力机制允许 patch 根据相邻 patch 的特征及连接的重要性调整其连接权重，从而计算模板图像的目标区域特征和搜索区域特征中的每个像素点之间的相关性，突出更显著的像素点，并且不会过度压缩传递到搜索特征的目标信息。该模型使用二分图的图注意力网络，并定义二分图 $B = (G, A)$ ，将 F_z 和 F_x 的每一个像素点在维持通道数不变的情况下视为图架构的节点，得到 V_z 和 V_x ，对 G 和 A 的定义如下：

$$\begin{aligned} G &= V_z \cup V_x \\ A &= \{(g, a) \mid \forall g \in V_z, \forall a \in V_x\}, \end{aligned} \quad (3-10)$$

其中， G 同时包含模板图像特征节点集 V_z 和搜索区域特征节点集 V_x ，表示 V_z 和 V_x 任意一个节点集之间没有直接的联系，而是两者之间的节点相互关联， A 定义了两个节点集之间所有的连接可能，恰当符合以局部策略计算模板特征和搜索区域

特征之间相关性的理念。对于每个 $i \in V_z$ ，计算 i 对于所有 $j \in V_x$ 的相似性系数，具体表达如下：

$$e_{ij} = \frac{\exp\left[\left(W_z t_z^i\right)^T \left(W_x t_x^j\right)\right]}{\sum_{k \in V_x} \exp\left[\left[\left(W_z t_z^i\right)^T \left(W_x t_x^k\right)\right]\right]}, \quad (3-11)$$

其中， e_{ij} 是节点 i 赋予节点 j 的关注度，首先对 i 和 j 的特征向量 t_z^i 、 t_x^j 进行线性变换，增强特征的泛化性，为了更好的进行二分类，使用矩阵乘法计算相似度，相似度越高表示 j 像素点是前景的概率越大，并给予更多的关注。

为了与全局互相关策略相呼应，且加强特征表示。根据注意力系数，把特征进行加权求和并将注意力系数作为软权重和搜索区域特征节点集 V_x 融合，具体表达如下：

$$\hat{t}_z^i = \sigma \left(\text{Cat} \left(W_v t_z^i \sum_{j \in V_x} e_{ij} W_v t_x^j \right) \right), \quad (3-12)$$

其中， $\hat{t}_z^i$ 表示聚合后的特征， σ 是激活函数，在此使用 ReLU， W_v 同样是线性变换矩阵。

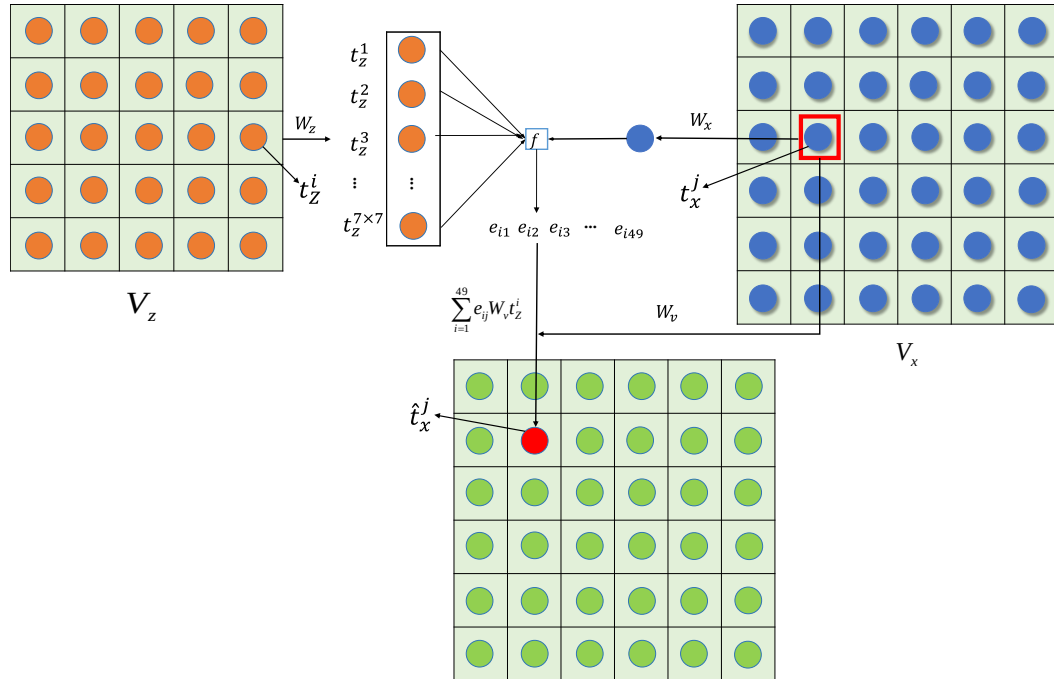


图 4-4 GM 模块结构

4.2.4. 特征匹配结果融合策略研究

本文已经介绍了如何分别利用当前帧和先前帧的特征通过互相关策略进行

相似性计算，那么如何将相似性计算结果融合起来是迫切需要解决的问题。为了保证模型不会存在退化现象，SiamGPF 采用残差结构连接当前帧特征的匹配结果，在特征图融合的方法中，常见的有通道拼接、对应元素相加、相减、点乘以及矩阵相乘等操作。本文提出的基本方法是使用先前帧的信息来补充当前帧。尽管通道拼接方式能够很好地保留参与融合的特征信息，但在实验中发现，仅仅使用特征相加作为融合方式并不足以达到理想的效果。因此，SiamGPF 采用卷积操作来减少通道维度，并引入注意力机制，以便模型能够关注到更重要的通道，从而提高模型的表示能力和泛化性能。如图 4-5 所示，模型采用 Sigmoid 函数的输出作为权重，与输入特征逐元素相乘。这种方法可以增强特征之间的相互联系，并实现更有效的融合效果。为了确保相似性得分图信息融合的可信性，并保持模型的简洁和高效，并没有对融合后的网络进行过多的处理。相反，直接将融合结果送入预测头网络，以评估目标的具体位置。这种直接而高效的方法有助于维持模型的性能，同时简化了整体的网络结构。

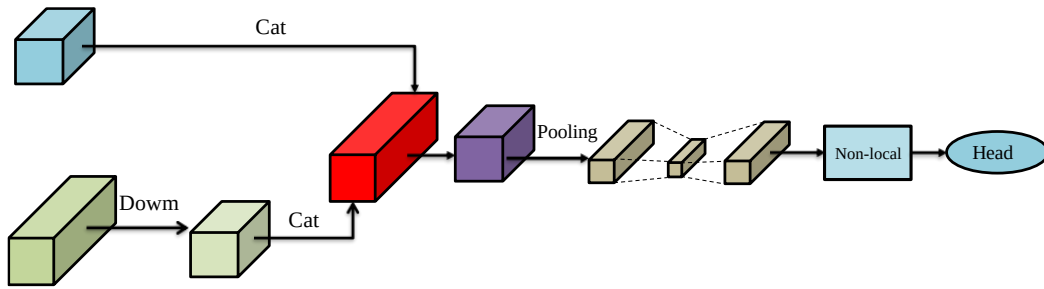


图 4-5 FGP 模块结构

4.2.5. 测试伪先验帧选择策略研究

SiamR-CNN、UpdateNet 等跟踪器使用模板更新策略适应目标的变化，这些方法利用不同的质量评估策略来确定不同帧的可信度，例如，STMtrack 通过融合先前帧、模板帧以及两者之间的采样帧和搜索帧来进行匹配估计。受到这些方法的启发，本文针对动态帧选择的方案根据不同需求进行了调整。模板帧作为单目标跟踪任务的唯一先验信息，提供了最可靠、最纯净的参考。然而，实际的跟踪环境复杂多变，目标的变化可能导致跟踪器在性能不足时难以准确识别。先前帧则最能代表目标相对于初始帧的变化情况。基于此，选择了先前帧 F_{x_t} 作为 SiamGPF 跟踪器的第三元输入。在训练阶段，当抽取训练样本以获取正样本时，从同一序列中选择搜索帧的前一帧作为先前帧的训练数据。然而，在推理过程中，

为了节省计算时间,不再重新提取先前帧的特征,而是采用数据迭代的方式。在每次循环过程中,仅保存上一帧的搜索图像特征,从而显著提高了模型在推理时的速度。具体表达如下:

$$F_{x-t} = \begin{cases} F_x^0 & t=1 \\ F_x^{t-2} & t \geq 2 \end{cases}, \quad (3-13)$$

其中, $t \in \{0,1,2,\dots,N-1\}$, N 代表当前测试序列的图像张数。由于第一帧没有先前帧直接使用第一帧的搜索图像特征 F_x^0 作为先前帧计算。

考虑到目标消失或是中途跟踪失败的案例,仅使用 F_{x-t} 甚至会引入错误信息,本文设计了一个先前帧动态选择器 FDS, 具体表述如下:

$$F_{x-t} = \begin{cases} F_x^0 & t \leq 2 \\ F_x^N & t > 2 \end{cases}, \quad (3-14)$$

其中, F_x^0 代表当前序列的第一帧特征, 因为跟踪第二帧时 F_x^1 是选择器中唯一的候选项, 尽管模型在初始帧就跟踪错误的情况很少存在, 但是却会直接导致跟踪失败, 因此依旧使用 F_x^0 作为先前帧。 N 是本文设计的具体选择方法: 表述如下:

表 4-1 FDS 伪代码

Algorithm 1 Algorithm for FDS

```

1: /*selector*/
2: 输入:  $i, gt\_box, xf, xf_i, auc, outputs['xf'], outputs['pre\_box'], video, \lambda, N$ 
3:  $a = (gt\_box \cap outputs['pre\_box']) / (gt\_box \cup outputs['pre\_box'])$ 
4: if  $i = 1$  then
5:    $xf\_pre[0] = xf$ 
6: else if  $i = 2$  then
7:    $xf\_pre[0] = outputs['xf']$  and  $auc.append(a)$ 
8: else
9:   for 3 to  $len(video)$  do
10:    if  $a \geq \lambda$  then
11:       $xf_i.append(outputs['xf'])$  and  $auc.append(a)$ 
12:      if  $len(xf_i) \geq N$  then
13:         $xf_i.pop(0)$  and  $auc.pop(0)$ 
14:      else
15:        pass
16:       $xf\_pre = xf_i[\max(auc)]$ 
17:   end for
18: 输出:  $xf\_pre$ 

```

4.3. 实验方法与结果分析

在本小节中，将对 SiamGPF 跟踪器的训练测试环境、方法和结果进行详细讲述，并使用可视化方法直观展示本文提出的跟踪器相较于基准跟踪器的效果提升。

4.3.1. 实验环境介绍

所有的训练和推理过程都是在 Ubuntu 22.04 PC 上进行的，具体全硬件配置见表 4-2。整体训练在 60h 内完成。SiamGPF 采用了增强版的 ResNet50 作为骨干网络。该网络使用 ResNet50 的前三个阶段进行训练，并在 layer3 阶段应用空洞卷积以提高特征的分辨率。训练过程包括 30 个 epoch，在前 10 个 epoch 中冻结骨干模型的参数。学习率策略采用线性增长，从 0.001 增加到 0.005，持续 5 个 epoch，随后使用余弦退火调度进行衰减（从 0.005 衰减到 0.00005），在第 15 和第 23 个 epoch 时，衰减率为 0.4。优化方式采用随机梯度下降（SGD），权重衰减为 0.0001，动量为 0.9。模板图像的尺寸为 127×127 ，搜索图像为 255×255 。训练数据集包含五个数据集：LaSOT、GOT-10k、YouTubeBB-BoundingBoxes、ImageNet-DET 和 COCO，经过裁剪变换后，每个 epoch 包含一百万对图像，批次大小为 28。对所有图像应用数据增强技术，如中心裁剪、中心平移和随机模糊，所有图像均以目标为中心。记忆帧处理方法与搜索图像一致，采样范围大于模板帧但小于搜索帧。

表 4-2 Siam GPF 训练实验配置

软/硬件名称	具体型号
显卡	RTX 4070ti 12G
处理器	I5-12490F
内存	32G DDR4
存储硬盘	2TB 固态硬盘
操作系统	Ubuntu 22.04
开发环境	Pycharm
编程语言	Python 3.7
深度学习框架	Pytorch 1.7
统一计算设备架构版本	Cuda11.2

本章开发了两种推理方法：第一种方法使用搜索序列中的前一帧作为初始输入帧，迭代地参与模型的推理过程。与缺乏历史帧功能的模型相比，这种方法只产生最小的额外计算成本。第二种方法利用先前提到的 FDS 来识别最佳历史帧，将图像选择阈值设置为 0.7，并将帧选择限制为最多 12 帧。虽然这种方法相对于第一种方法略微增加了推理时间，但它产生了更好的成功率，这两种策略都表现出明显的优势和局限性。

4.3.2. 实验结果

为了验证本文提出的部件是有效的，SiamGPF 将在七个公开测试集上做广泛验证实验。

SiamGPF 在 LaSOT、NFS、LaSOT_{Ext} 和 GOT-10k 上测试结果：LaSOT 是一个大规模的单目标跟踪数据集，是评估跟踪器性能最具代表性的基准之一。在本研究中，遵循数据集协议 I，对 SiamGPF 和 SiamGPF_P 模型进行了训练（FDS 模块仅在测试中使用）。随后，在测试集上对这些模型进行了评估。本文提出的跟踪器采用了标准化精度和精度指标进行测试，如表 4-3 所示，与知名的跟踪器（如 SiamGAT、SiamBan、CNNInMO^[70]、ATOM、ECO、SiamDW、CGACD^[71]、SiamFC++^[72]和 SiamRPN++）相比，SiamGPF 和 SiamGPF_P 分别达到了 57.4% 和 52.0% 的成功率，标准化精度分别为 65.6% 和 59.2%，精度分别为 58.8% 和 54.3%，均取得优异的结果。此外，如图 4-6 所示，研究详细评估了 LaSOT 中 14 个挑战，包括纵横比变化、低分辨率、视野外、快速运动、完全遮挡、尺度变化、视角变化、背景杂乱、旋转、相机运动、运动模糊、变形、部分遮挡和光照变化（图 4-6 中已缩写）。所有对比结果表明，本文提出的模型在大规模跟踪基准测试中表现出色。

NFS 数据集是专为评估视频跟踪算法而设计的，包含 100 个快速移动的视频序列，并具有九个挑战属性。这些视频使用高帧率设备录制，旨在测试跟踪器处理快速移动目标的能力。NFS 有两个版本，分别是 NFS30 和 NFS240。在本次测试中使用了 NFS30 版本测试跟踪器。如表 4-3 所示，该模型在多种跟踪器中表现优异，成功率（SUC）分别达到了 58.9% 和 57.0%。此外，模型在精度（Pre）指标上也取得了优异的成绩，SiamGPF 达到了 71.2%，SiamGPF_P 达到了 69.0%，明显优于 SiamCAR（52.9%）和基线跟踪器（51.5%）。

LaSOT_{Ext} 是 LaSOT 数据集的扩展版本, 包含 15 个新类别和 150 个具有挑战性的长视频序列。如表 4-3 所示, SiamGPF 在 LaSOT_{Ext} 数据集上的曲线下面积 (AUC) 得分为 39.2%和 36.0%, 精度 (P) 得分分别为 45.5%和 41.4%, 排名仅次于 ATOM。与基线跟踪器相比, AUC 提高了 6.9%和 3.7%, P 分别提高了 8.3%和 4.2%。所有结果与本文研究预期一致, 并和其他评估基准相符。SiamGPF 中提出的前一帧建模确保了更稳定的目标定位性能, 并在大规模长序列数据集上继续表现出色。

GOT-10k 是一个大规模且多样化的视频跟踪数据集, 为跟踪评估提供了新的基准。在训练中每个 epoch 使用了 600 万张图像, 总共训练了 20 个 epoch, 其余训练参数保持不变。如表 4-3 所示, SiamGPF 相较于 Ocean、SiamBan、SiamCAR、ATOM 和 SiamFC++等算法表现出了显著改善。在平均重叠度 (AO) 指标上, 该模型相比 Ocean 提高了 1.2%, 在 SR_{0.75} (重叠阈值为 0.75 时的成功率) 指标上提高了 3.9%。与基线跟踪器相比, AO 提高了 8.3%, SR_{0.5} 提高了 8.3%, SR_{0.75} 提高了 12.0%。这些性能提升得益于本文提出的先前帧模型和动态选择器。这些创新使得模型能够更准确地区分前景和背景, 并通过利用历史帧增强全局信息, 从而更加精确地估计目标的位置。

表 4-3 SiamGPF 在 LaSOT、NFS、LaSOT_{Ext} 和 GOT-10k 上跟踪结果

Method	LaSOT			NFS		LaSOT _{Ext}		GOT-10k		
	SUC	P _{Norm}	P	SUC	P	SUC	P	AO	SR _{0.5}	SR _{0.75}
SiamGPF	57.4	65.4	58.8	58.9	71.2	39.2	45.5	62.8	72.9	52.6
SiamGPF_P	52.0	59.2	54.3	57.0	69.0	36.0	41.4	61.6	71.7	48.7
Baseline	50.9	59.1	51.7	51.5	61.8	32.3	37.2	54.5	64.6	40.6
Ocean	56.0	65.1	56.6	55.3	-	-	-	61.1	72.1	47.3
CNNInMo	53.9	61.6	53.9	56.0	68.5	-	-	-	-	-
CGACD	51.8	62.6	53.5	-	-	-	-	-	-	-
SiamCAR	50.7	60.0	51.0	52.9	65.7	33.9	41.0	56.9	67.0	41.5
SiamBan	51.4	59.8	52.1	59.4	70.0	-	-	57.9	68.4	45.7
SiamFC++	54.4	62.3	54.7	58.1	-	-	-	59.5	69.5	49.7
ATOM	51.5	57.6	50.5	59.0	69.4	37.6	43.0	55.6	63.4	40.2
ECO	32.4	33.8	30.1	46.6	54.7	22.0	24.0	31.6	30.9	11.1
SiamRPN++	49.6	56.9	49.1	50.8	50.9	34.0	39.6	51.7	61.6	32.5
SiamFC	33.6	42.0	33.9	-	-	23.0	26.9	34.8	35.3	9.8

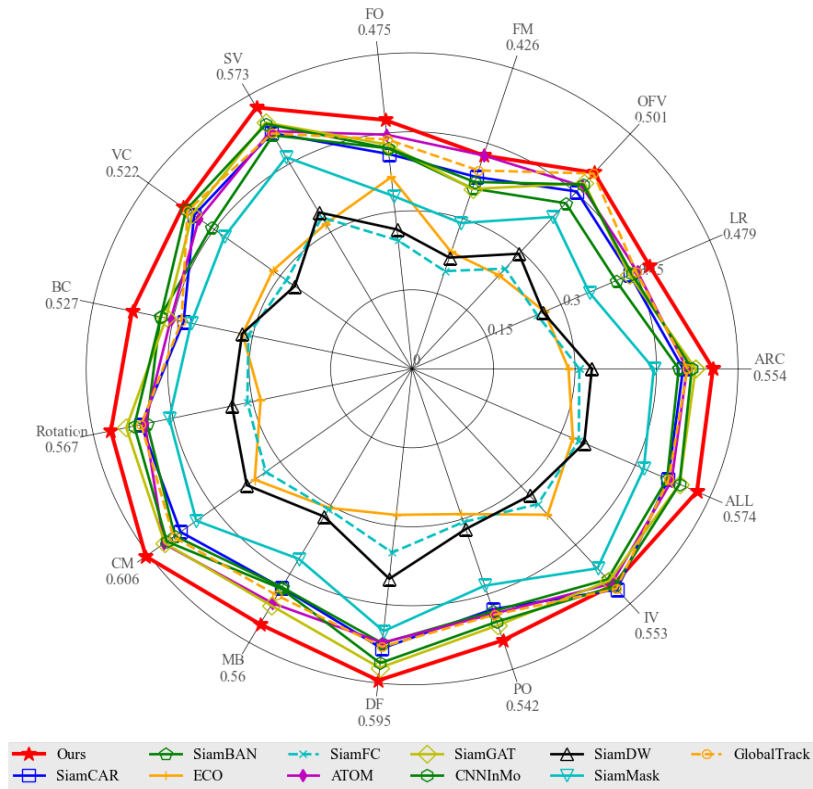


图 4-6 SiamGPF 在 LaSOT 不同属性上挑战成功率

SiamGPF 在 OTB100 上测试结果：该测试集涵盖了多个具有挑战性的场景，如目标变形、运动模糊、相似目标干扰和目标遮挡。在该数据集上，SiamGPF 表现优异。如图 4-7 所示，该模型成功率得分达到 71.3%，精度得分为 92.9%。相比 SiamGAT、CNNInMO 和 SiamRPN 等领先模型表现出更优越的性能。

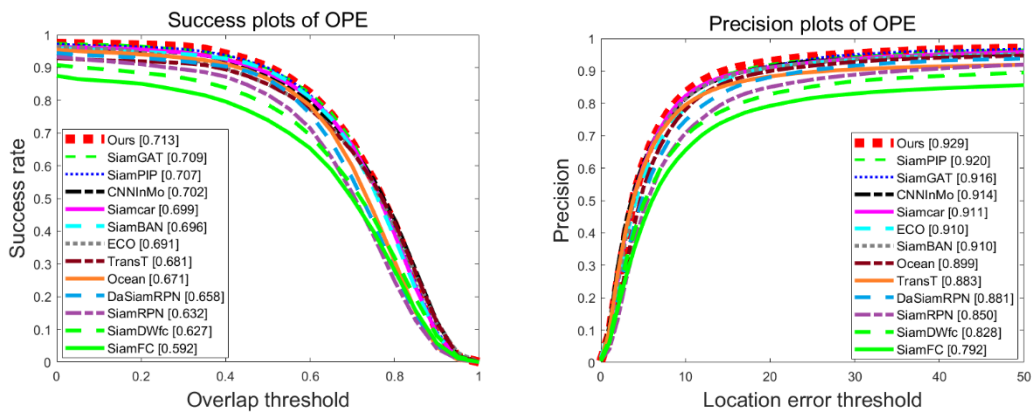


图 4-7 SiamGPF 在 OTB100 上跟踪结果

OTB100 测试集为每个视频序列进行分类标注，如图 4-8 所示，SiamGPF 面对背景模糊成功率得分达到了 0.708，远超 SiamBan，同时，在处理平面外旋转、低分辨率以及光照变化时跟踪成功相比其他著名跟踪器提升很大，通过历史帧信

息辅助当前帧准确捕捉目标信息，并巧妙融合两者互相观结果可以使 SiamGPF 应对复杂环境下的跟踪任务。

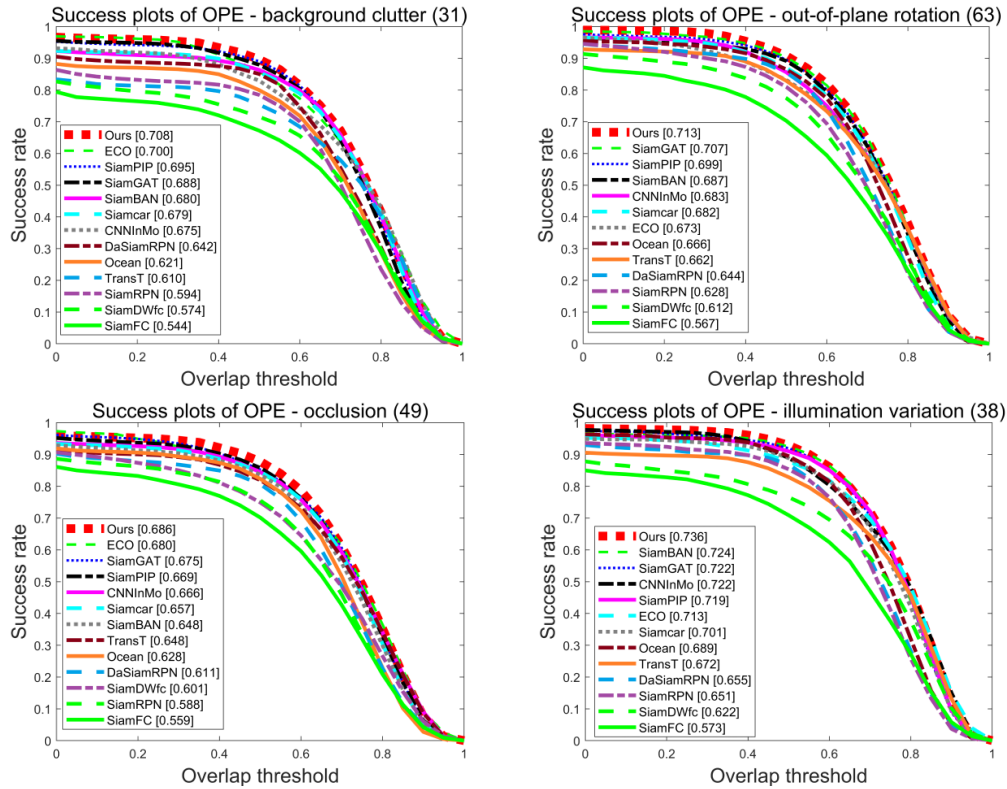


图 4-8 SiamGPF 在 OTB100 中不同挑战属性下跟踪结果

SiamGPF 在 UAV123 上测试结果：如图 4-9 所示，本文提出的 SiamGPF 模型超越了 SiamGAT、SiamBan 和 AutoMatch 等知名算法，成功率达到 65.1%，精度为 84.8%。相比 SiamBan，SiamGPF 在成功率上提升了 2%，在精度上提升了 2.8%。UAV123 标注了不同类型挑战真实数据，从图 4-10 中看以看到 SiamGPF 在相似物干扰、背景杂乱、低分辨率和剧烈形变下表现最佳。值得注意的是，SiamGPF 在保持高跟踪器鲁棒性的同时，实现了高速运行，进一步证明了其在无人机跟踪场景中的高效性和可靠性。

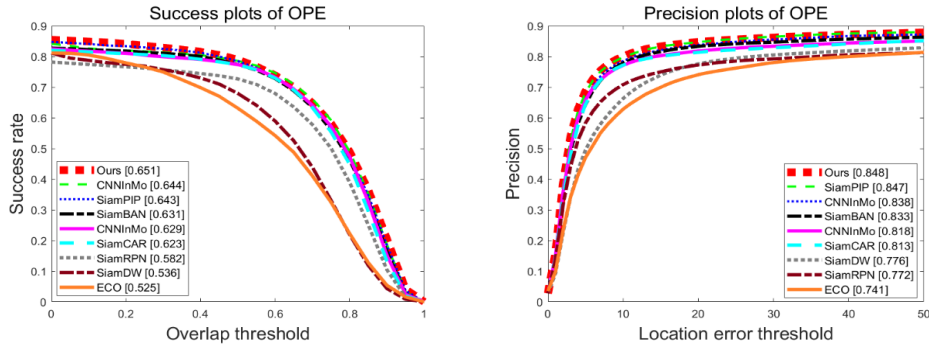


图 4-9 SiamGPF 在 UAV123 上跟踪结果

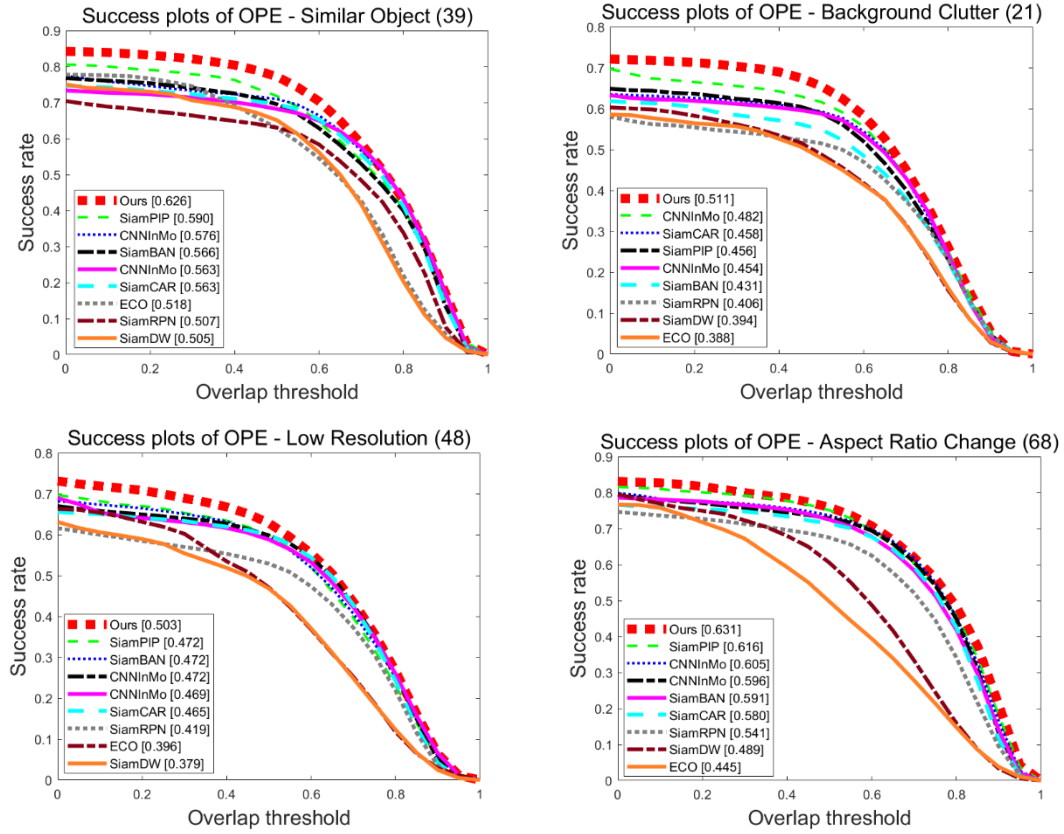


图 4-10 SiamGPF 在 OTB100 上不同属性中的跟踪成功率

SiamGPF 在 VOT2016 上测试结果：VOT2016 通过计算跟踪过程中重叠度低于特定阈值的帧数，并在跳过这些帧后重新初始化跟踪，以评估模型的鲁棒性。如表 4-4 所示，SiamGPF 和 SiamGPF_P 模型在 VOT2016 数据集上的期望平均重叠度（EAO）得分分别为 52.7%和 54.3%。由于 VOT 的独特评估机制，两个模型在准确性上的性能差异较小。然而，在鲁棒性方面，两者的得分均为 11.2，明显优于其他算法。为了更直观地观察模型性能，本文使用了气泡图来展示和排名 EAO 得分，如图 4-11 所示。

表 4-4 SiamGPF 在 VOT2016 上测试结果

Metrics	SiamRPN	SiamDW	UpdateNet	Ocean	SiamBan	SiamGPF_P	SiamGPF
EAO	33.7	37.1	48.1	48.6	49.9	52.7	54.3
A	57.8	58.0	61.0	62.5	63.2	64.3	64.1
R	31.2	24.0	20.6	15.8	14.9	11.2	11.2

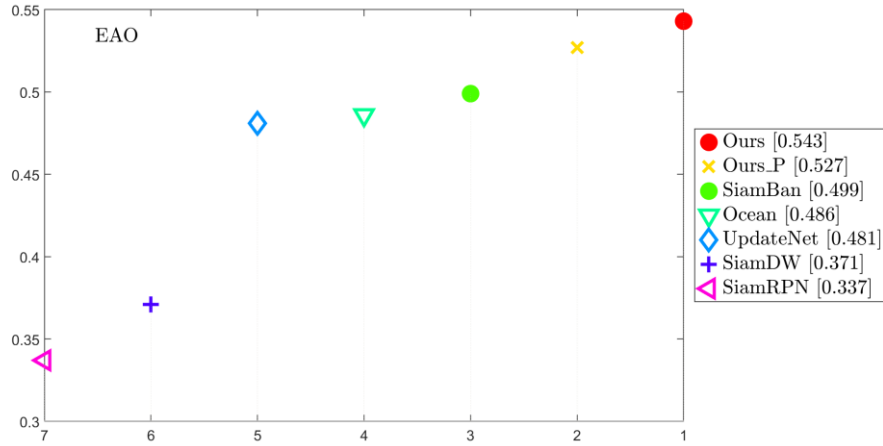


图 4-11 SiamGPF 在 VOT2016 上跟踪结果（EAO）

SiamGPF 在 TC128 上测试结果：为了进一步评估跟踪器的泛化能力，在 TC128 基准数据集上进行了测试。TC128 数据集由 128 个彩色视频序列组成，特别关注在目标变形和光照变化条件下的跟踪挑战。如表 3 所示，SiamGPF 在 SUC 指标上达到 62.6%，相比 SiamBan（57.0%）提升了 5 个百分点，相较于基线提升了 6%。在 Pre 指标上，SiamGPF 达到了 86.1%，比 SiamBan 高出 7 个百分点。这些结果进一步验证了本文的模型在不同跟踪场景下的鲁棒性。

表 4-5 SiamGPF 在 TC128 上跟踪结果

Metrics	SiamRPN++	Siam Ban	Siam CAR	Ocean	SiamDW	CGACD	SiamGAT	SiamGPF	Baseline
SUC	57.3	57.0	56.2	55.1	58.1	60.5	55.9	62.6	56.6
Pre	-	78.6	76.8	78.5	79.9	-	75.3	86.1	78.0

4.3.3. 性能与参数量比较

SiamGPF 不仅在跟踪性能上相较于基准跟踪器有很大的提升，而且跟踪帧率更是达到 221FPS，在本节中将分析 SiamGPF 的参数量与性能之间的联系。

为了确保实验结果的高可信度，将模型在 GOT-10k 数据集上进行了测试，并在其官方服务器上进行了评估。如图 4-12 所示，SiamGPF 模型只有 14.05M 参数，仅是 SiamRPN+的一半（39.95M）。在 GOT-10k 数据集上测试时，SiamGPF 达到了 62.8%的 AO（平均重叠度），超越了 ATOM 模型 7.2 个百分点，并实现

了 212 FPS 的实时操作。该模型通过精简的结构和令人印象深刻的跟踪成功率，展示了其在实际应用中的巨大潜力。同时在表 4-6 中测试了每个算法在 GOT-10k 上的帧率（因为 GOT-10k 网站的评估更具权威性）。

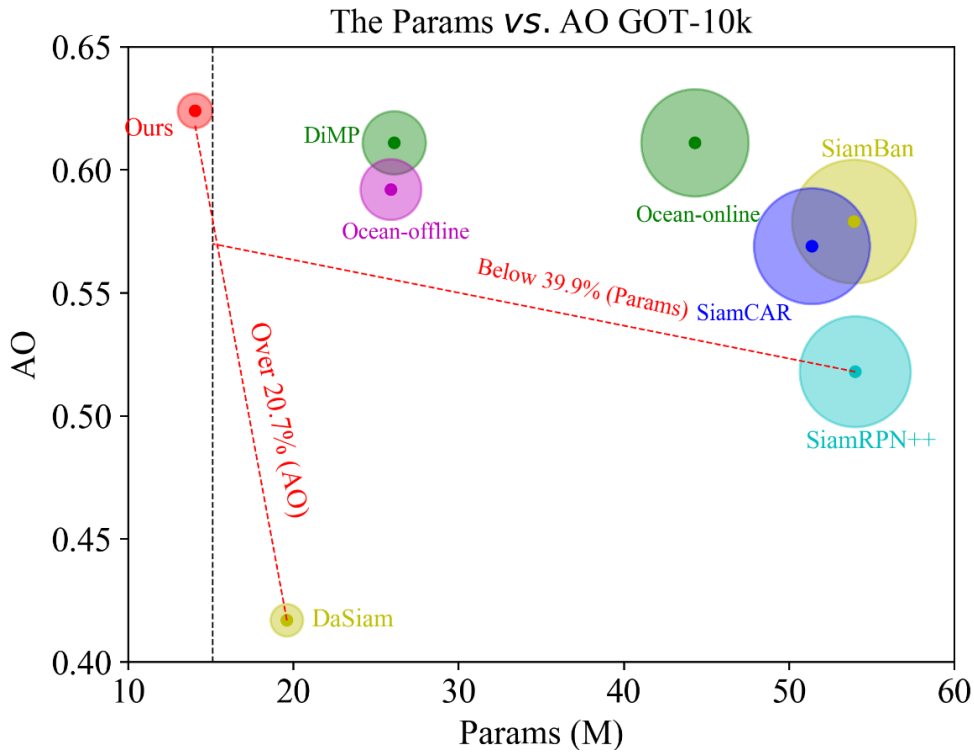


图 4-12 SiamGPF 在 GOT-10k 测试集上跟踪结果以及参数量

表 4-6 SiamGPF 与其它跟踪器跟踪速度比较

Method	FPS	Hardware	Language
SiamGPF	212	RTX4070ti	Python
Ocean-online	98.45	RTX4070ti	Python
CNNInMo	50	RTX4070ti	Python
CGACD	128.34	RTX4070ti	Python
SiamCAR	53	RTX4070ti	Python
SiamBan	93.75	RTX4070ti	Python
SiamFC++ - AlexNet	181.4	RTX4070ti	Python
ATOM	25	CPU	Python
ECO	2.62	CPU	Matlab
SiamRPN++	49.83	RTX2080ti	Python
SiamFC	25.81	Tian X	Matlab

4.3.4. 消融实验

在 4.2 小节中，将 SiamGPF 与其它先进跟踪器在各个公开测试集中测试比较，可以得出将历史帧作为伪先验信息输入到孪生网络中是能够有效提升跟踪器性能的，在本节中，将在两个基准数据集 OTB100 和 UAV123 上对所提出的两个模型进行了深入分析，包括训练阶段的模块消融实验以及推理阶段 FDS 的定性比较。值得说明的是 SiamGPF 的基准跟踪器与 SiamPIP 是不完全相同的，因其特殊的模板匹配机制，使其基准结合 SiamBan 的骨干网络、SiamCAR 的预测头网络，并遵循 SiamBan 的训练策略。

训练消融实验：为了全面了解 SiamGPF 中各个组件的贡献，本文进行了广泛的消融实验。如表 4-7 所示（已应用 FDS），方法 1 采用 GM 作为当前帧的匹配策略，同时保留了原始的深度相关性（DW-Corr）用于前一帧。可以明显看出 GM 提高了基准模型在 OTB100 上的 AUC 得分，从 68.3%提升至 70.5%，这验证了基于局部图注意力匹配策略能够提升信息传递的准确性。此外，部分对部分的匹配关系帮助模型更好地聚焦于目标的位置。为了确保跟踪器的实时性能，该模型仅使用了单层注意力机制。

方法 2 表明，当当前帧使用 DW-Corr，前一帧使用全局 PM 时，OTB100 上的成功率从 68.3%提升到 70.3%，相比方法 1 略有下降。GM 是一种局部互相关策略，擅长解决目标变形和运动模糊问题，但它有明显的缺点：缺乏全局信息，且无法确定捕获的信息是否有效，有时模糊的信息可能会影响跟踪效果。PM 是一种全局匹配策略，可以包含大部分信息，但在识别目标的剧烈变化方面存在局限。因此本文提出了一种替代方案，在两者基础上同时进行匹配，且按照主次关系（GM 为主，PM 为辅），通过将目标历史状态作为 PM 中的伪先验信息替换为搜索区域图像的位置。如表 4-7 中的实验所示，GM 和 PM 都能在一定程度上提升跟踪器性能，而 FGP 和 FDS 的结合进一步提高了性能。

鉴于 FGP 在模型内的作用是信息融合，单独对方法 1 和方法 2 中的 FGP 进行消融实验意义不大。因此，本文仅分析了 GM 和 PM 都使用时的 FGP。在方法 3 中，使用通道拼接来融合当前帧和前一帧的匹配结果。表 4-7 显示，在成功率上并没有提升，反而有所下降，这归因于简单的融合方法导致匹配信息的混乱，过多的全局背景信息错误地影响了模型区分前景和背景的能力。此外，研究中还

尝试了简单的逐元素加法来融合 FGP 中的信息，但结果依旧不理想。虽然这种方法计算简单，并能保留两侧的信息，但在设计时，我并未对 GM 和 PM 赋予相同的重要性。PM 的作用更多是作为信息补充，在 GM 失效时进行错误修正。因此，尽管通道拼接略微增加了计算负担，但它确实代表了最佳的融合方案，并通过后续的瓶颈结构进一步增强了模型的泛化能力。SiamGPF 还通过在 UAV123 数据集上的评估测试进一步验证了本文的分析，结果是互相符合的。

表 4-7 SiamGPF 中各个模块消融实验

Method	GM	PM	FDS	OTB100		UAV123	
				SUC	Pre	SUC	Pre
Baseline	-	-	-	68.3	88.8	61.4	81.4
1	√	-	√	70.5	91.2	63.4	82.5
2	-	√	√	70.3	91.3	62.6	82.4
3	√	√	-	67.7	88.4	61.2	80.9
4	√	√	√	71.0	92.2	65.1	84.8

训练消融实验：本文设计的先前帧动态选择器消融实验如表 4-8 所示，在保证 GM、PM、FDS 正常工作的情况下，方法 1 将先前帧定义为当前帧的前一帧。来自前一帧的目标特征信息对当前帧的定位任务起着重要作用，代表着目标最新的状态。有效利用这一信息可以减轻跟踪器的性能压力。可以看到和 Baseline 相比，方法 1 在 OTB100 数据集上的成功率从 68.3%提升到 70.3%，准确度从 88.8%提升到 92.3%，但是，当进一步分析跟踪器在已知挑战类型（例如目标剧烈形变、局部遮挡、目标从视野中消失等）的表现时，发现方法一的评估结果甚至低于基线模型。这是因为误差累积的原因。当跟踪器在某一帧产生错误定位时，先前帧便会向当前帧引入不正确的特征信息，连续的视频序列会放大这一问题。

因此，本章提出了 FDS，将先前帧定义为目标历史状态中质量得分高于阈值 λ 的搜索图像。为了避免方法一的弊端，该模型提供了 N 个历史帧供选择，具体策略见算法 1。可以看到，方法二在 UAV123 评测标准上相较于方法一从 63.5%提升到 65.1%（AUC），而在 OTB100 上的实验结果表明，FDS 有助于提高跟踪器的鲁棒性，这得益于，模型提供多个高质量的历史帧作为先前帧的候选。此外，本文还对候选帧数量 N 和特征质量得分阈值 λ 在 OTB100 上进行了进一步的分析（以 $T=0.7$ 为例）。研究发现，当 N 过高时，跟踪器性能反而下降，这可

能是因为候选帧保留了低质量的特征图像，FDS 的基本原理导致了错误的放大。因此，适量的 N 和先存后取的迭代方法有助于缓解累积错误。在表 4-9 中也对得分阈值 λ 进行了比较，当 λ 取 0.8 时，跟踪器性能显著下降（从 71.0% 降至 70.0%）。与候选帧数量的原因不同，当在代码中打印候选帧时发现，由于过高的阈值 λ ，当跟踪器开始没有成功定位时，先前帧一直被选择为第一帧搜索图像，这也是 SiamGPF 性能急剧下降的原因。

表 4-8 SiamGPF 在测试中消融实验

Method	Previous frame	FDS	OTB100		UAV123	
			SUC	Pre	SUC	Pre
1	√	-	70.3	92.7	63.5	83.7
2	-	√	71.0	92.2	65.1	84.8

表 4-9 FDS 中超参数消融实验

N	SUC	Pre	P	SUC	Pre
6	70.7	91.6	0.4	70.3	91.3
12	71.0	92.2	0.6	70.7	91.7
18	70.8	92.0	0.7	71.0	92.2
25	70.4	91.1	0.8	70.0	90.6

泛化性消融实验：为了验证模型的泛化性，本文使用 AlexNet 替换了主干网络，并在 OTB100 数据集上进行了测试。结果如表 4-10 所示：与同样使用 AlexNet 的 SiamFC 相比，成功率从 59.2% 提高到 69%，准确率从 79.2% 提高到 89.5%。这足以证明所提出的模型可以提高跟踪器的性能。

表 4-10 SiamGPF 泛化性实验

方法	骨干网络	OTB100	
		Suc	Pre
1	AlexNet(SiamFC)	0.592	0.792
2	AlexNet(SiamGPF)	0.690	0.895

4.3.5. 可视化分析

为了更加直观清晰的说明 SiamGPF 框架是有效的，本小节将在 OTB100 上挑选代表性的视频序列可视化跟踪热图，并和 Baseline 做出比较。

相似物干扰：如图 4-13 所示，前两行分别展示了 SiamGPF 的融合图和热图，相较于基线模型（第三、四行），可以观察到，当存在相似物体的干扰时，SiamGPF 能够更加精准地区分出真实的目标位置，其关注焦点更为集中。这得益于来自先前帧的线索不断为当前帧提供跟踪信息的补充。而基线模型在第三帧时无法明确区分相同类别的干扰物，尽管在第 640 帧有所改善，但其注意力的分散导致目标定位的不准确。

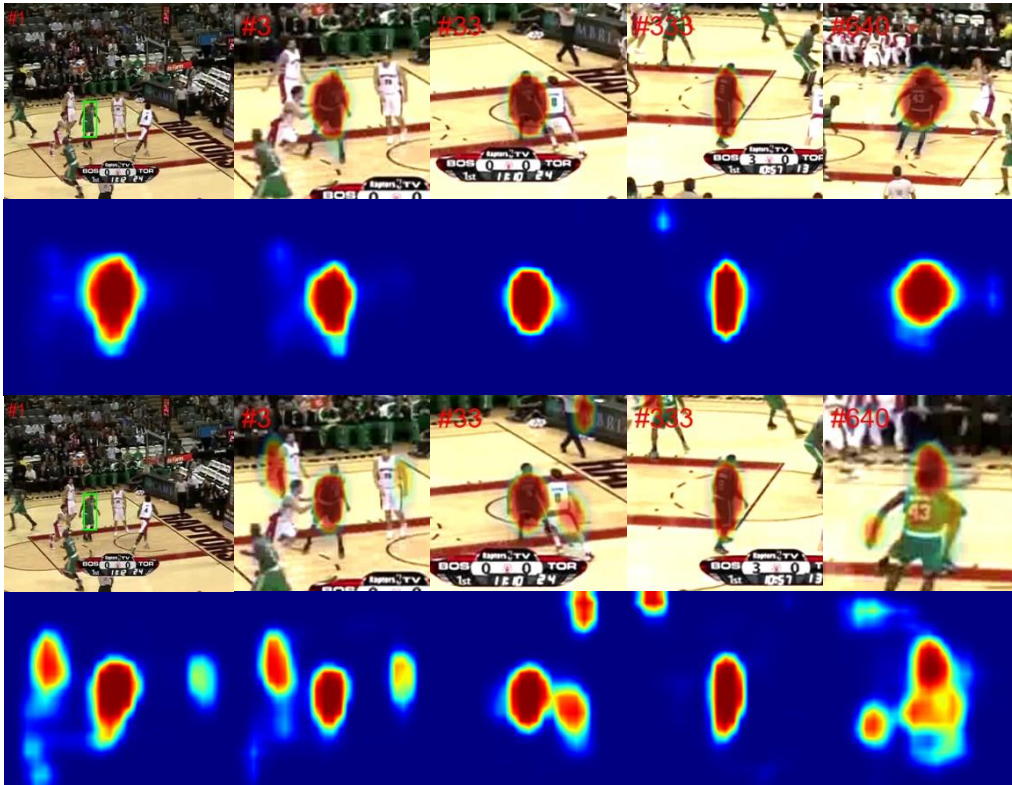


图 4-13 跟踪器在 OTB100 测试集中面对相似物干扰挑战时跟踪可视化

剧烈形变：当前帧的匹配策略依赖于图注意力机制的局部匹配方法，它能够实现目标到搜索区域的局部信息传递，对于应对目标剧烈形变表现出色。如图 4-14 所示，当跳水运动员的身体姿态发生较大变化时，基线模型在第 201 帧就已经跟踪失败，这是因为相较于初始帧，模型无法有效理解物体的变化。而有了先前帧的辅助，模型可以获取目标最近期的变化信息，因而 SiamGPF 模型在跟踪成功率上得到了显著提升。

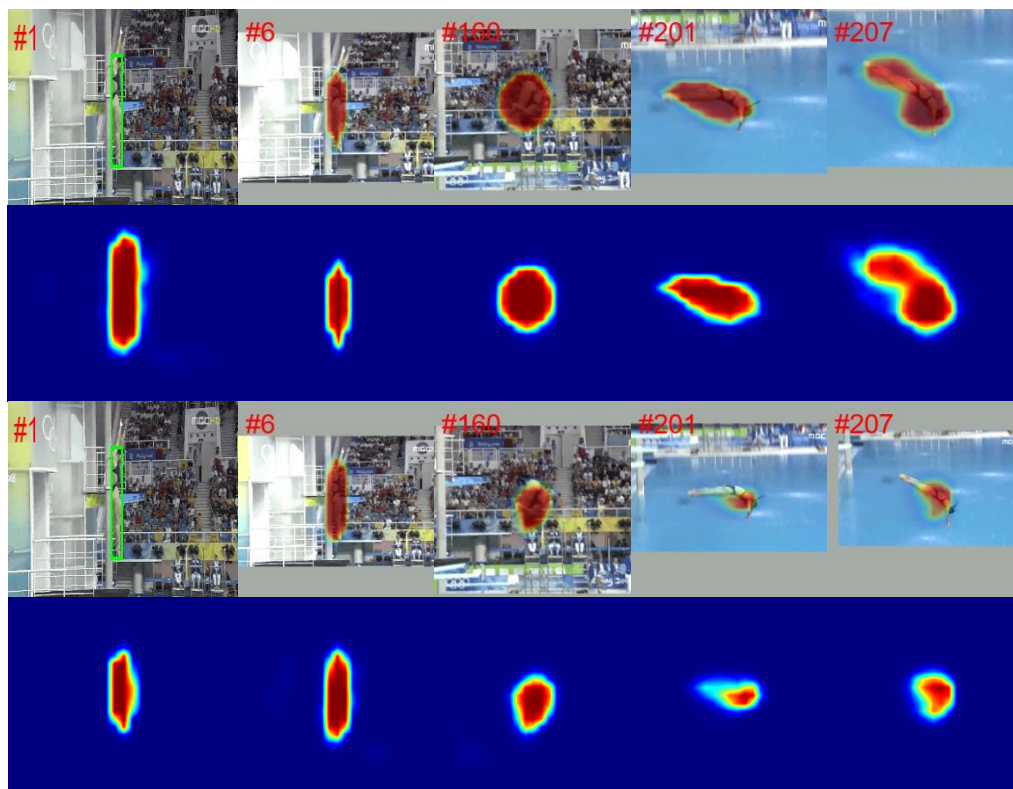
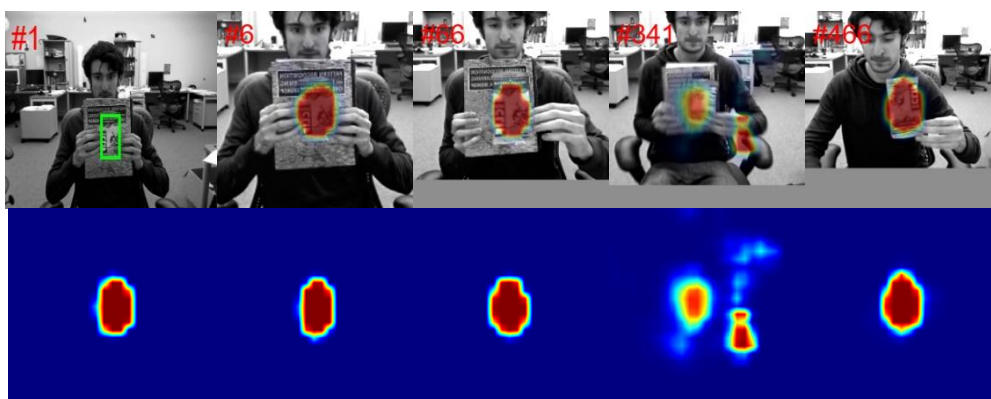


图 4-14 跟踪器在 OTB100 测试集中面对剧烈形变挑战时跟踪可视化

重新定位：SiamGPF 的另一个显著优势在于，在不依赖模板更新策略的情况下，能够对跟踪失败的图像实现重新定位。得益于先前帧的引入，高质量的历史帧得以被储存并作为候选，以供后续选择。先前帧动态选择器向模型提供了可以代表目标最新状态的高质量特征。正如图 4-15 所展示的，尽管 SiamGPF 在第 341 帧的表现并不理想，但得益于先前帧提供的目标关键历史状态，该跟踪器能够及时纠正偏差，重新准确地进行跟踪。相比之下，基线模型在第 196 帧跟踪失败后，其后续序列持续失效。这也从一个侧面印证了 SiamGPF 模型在处理复杂跟踪问题时的出色性能。



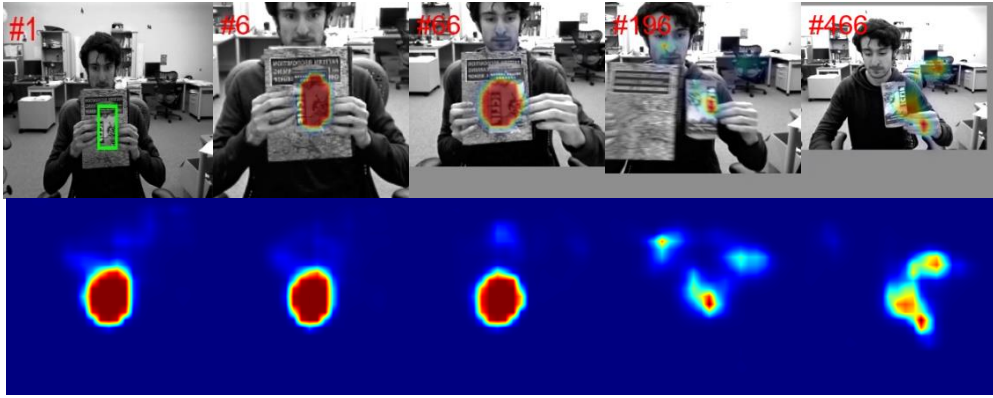


图 4-15 跟踪器在 OTB100 测试集中面对重新定位挑战时跟踪可视化

4.4. 本章小结

在研究 Siam PIP 跟踪器的过程中，发现初始帧信息挖掘工作已经非常成功，但是仅依靠初始帧对跟踪器的性能提升是有限的。此外，三元输入方法对孪生跟踪器框架是一种可行的方案。针对以上问题，本章提出了 SiamGPF 跟踪器，该跟踪器利用跟踪过程中目标的历史帧信息辅助跟踪，能够有效的应对目标产生剧烈变化以及遮挡等困难，并经过量化实验证明，PM、GM、FGP 和 FDS 四个部件是有效的，为了不影响跟踪器运行帧率，仅使用修改后的骨干网络中第二层特征作为后续特征，并且采用 FDS 策略避免先前帧的特征计算，这会大大节省计算资源，这也是该跟踪器帧率高的原因。

第5章 总结与展望

5.1. 总结

由于单目标跟踪在自动驾驶领域、监控安防领域、军事国防领域以及人机交互领域的广泛应用,使其成为计算机视觉领域的重要研究热点之一。复杂场景下的目标遮挡与视野消失一直是多数跟踪器难以应对的难题,本文从先验信息着手,深度挖掘先验信息的内容,研究初始帧、当前帧、历史帧之间的联系,并作出以下工作:

首先对单目标跟踪的研究背景、研究意义以及国内外研究现状进行整理回顾,其次介绍了本文工作所需了解的相关基础知识,如卷积神经网络的结构原理、孪生网络的由来以及在目标跟踪中的应用、注意力机制、相关测试集和评价标准。

以 SiamBan 为基线算法,在分析其在跟踪时面临困难的同时提出了 SiamPIP 跟踪器。该模型包含两个新模块:按任务比例融合相关结果(PF)和利用先验信息将杂乱背景与上下文信息分离(PI)。这两个模块相辅相成,使得目标对象的分类更加准确,同时预测出的目标帧也更接近真实目标。在多个视觉跟踪基准上进行的实验表明, SiamPIP 能够以先进的性能运行,并且达到了每秒 55 帧的速度。

在分析 SiamPIP 的结构时发现其对先验信息的认知具有局限性,因此提出了 SimGPF 跟踪器,该框架跟踪整体的角度上整合了时间相关性,将在跟踪过程中捕捉到的目标变化信息反馈给模型。这种方法不仅赋予先前帧辅助作用,还为当前帧的目标定位提供了全局信息支持。此外,通过 FDS 将先前帧定义为静态或动态帧,以满足不同需求。通过一系列全面的实验发现引入先前帧显著增强了跟踪器的鲁棒性,使其能够适应复杂场景,包括目标变化、遮挡以及视野消失问题。这种新型的 Siamese 跟踪结构不仅表现出强大的通用性,而且由于其简化的网络设计,使得 SiamGPF 能够在 GOT-10k 数据集上以每秒 212 帧的速度实时运行。

5.2. 展望

本文针对跟踪器在复杂场景下面临目标遮挡、视野消失等挑战时难以准确捕捉目标位置的问题,设计了 SiamPIP 和 SimaGPF 两种跟踪器,其中, SiamGPF

在 SiamPIP 思想基础上进行全局性关注先验信息，并取得了理想性能，但是仍存在许多不足的地方。基于本文所提出的研究成果，今后的工作可以从以下几点展开：

本文第四章提出的一种基于时间相关性的跟踪器从全局角度出发，关注跟踪的整体过程，并未深入改进骨干网络结构，但是在进行后续工作时将该思想架构移植到基于 AlexNet 骨干网络的 SiamFC 跟踪器中得到有效提升，在 OTB100 数据集评价标准下，跟踪成功率相比于 SiamFC 提升 9.8%，准确度提升 10.3%，同样将其应用到 OStrack^[73]跟踪器，在 LaSOT 标准下跟踪成功率提升了 1%。因此在跟踪中关注帧与帧之间的联系是行的通的，但是由于计算资源受限，并未过多研究基于 Transformer 的单目标跟踪器，这是本文有待进步之处。

本文提出的 SiamGPF 虽然在跟踪帧率上达到 212FPS，但是高额的计算量还是未能部署到实际任务中，因此继续优化骨干网络，对融合网络结构进行有效剪枝，降低训练成本并部署到无人机领域也是后续值得研究的方向之一。

参考文献

- [1] Chavez-Garcia R O, Aycard O. Multiple sensor fusion and classification for moving object detection and tracking[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 17(2): 525-534.
- [2] Bewley A, Ge Z, Ott L, etc. Simple online and realtime tracking[C]//2016 IEEE international conference on image processing (ICIP). Ieee, 2016: 3464-3468.
- [3] Yang J, Zhao W, Han Y, etc. Aircraft tracking based on fully conventional network and Kalman filter[J]. IET Image Processing, 2019, 13(8): 1259-1265.
- [4] 王义琼,王力超,李子阳,等.偏转速度实时修正的弹道跟踪系统设计[J].兵器装备工程学报,2023,44(03):204-209.
- [5] 任杰.基于深度学习的人脸视频心率检测技术研究[D].南京理工大学,2021.001254.
- [6] 柳曦媛.面向无人驾驶的目标检测与跟踪方法研究[D].中北大学,2024.000949.
- [7] BOLME D S, BEVERIDGE J R, DRAPER B A, etc. Visual Object Tracking Using Adaptive Correlation Filters[C]//Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2010: 2544-2550.
- [8] HENRIQUES J F, CASEIRO R, MARTINS P, etc. Exploiting the Circulant Structure of Tracking-By-Detection with Kernels[C]//Proceedings of the 12th European Conference on Computer Vision (ECCV), 2012: 702-715.
- [9] HENRIQUES J F, CASEIRO R, MARTINS P, etc. High-speed Tracking with Kernelized Correlation Filters[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(3):583-596.
- [10] Danelljan M, Häger G, Khan F S, etc. Discriminative scale space tracking[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(8): 1561-1575.
- [11] Bertinetto L, Valmadre J, Golodetz S, etc. Staple: Complementary learners for real-time tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1401-1409.
- [12] Li Y, Zhu J. A scale adaptive kernel correlation filter tracker with feature integration[C]//Computer Vision-ECCV 2014 Workshops: Zurich, Switzerland, September 6-7 and 12, 2014, Proceedings, Part II 13. Springer International Publishing, 2015: 254-265.
- [13] Danelljan M, Hager G, Shahbaz Khan F, etc. Learning spatially regularized correlation filters for visual tracking[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4310-4318.
- [14] Kiani Galoogahi H, Sim T, Lucey S. Correlation filters with limited boundaries[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4630-4638.

- [15] Wang M, Liu Y, Huang Z. Large margin object tracking with circulant feature maps[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4021-4029.
- [16] Bertinetto L, Valmadre J, Henriques J F, etc. 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.
- [17] Li, J. Yan, W. Wu, etc. High Performance Visual Tracking with Siamese Region Proposal Network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:8971–8980.
- [18] Girshick R, Donahue J, Darrell T, etc. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
- [19] Zhu Z, Wang Q, Li B, etc. Distractor-aware siamese networks for visual object tracking[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 101-117.
- [20] Chen Z, Zhong B, Li G, etc. SiamBan: Target-aware tracking with Siamese box adaptive network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(4): 5158-5173.
- [21] Guo D, Wang J, Cui Y, etc. 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.
- [22] Wang Q, Zhang L, Bertinetto L, etc. 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.
- [23] Li B, Wu W, Wang Q, etc. 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.
- [24] Wang Q, Teng Z, Xing J, etc. Learning attentions: residual attentional siamese network for high performance online visual tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4854-4863.
- [25] Guo D, Shao Y, Cui Y, etc. Graph attention tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 9543-9552.
- [26] Zhang Z, Liu Y, Wang X, etc. Learn to match: Automatic matching network design for visual tracking[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 13339-13348.
- [27] Elsken T, Metzen J H, Hutter F. Neural architecture search: A survey[J]. Journal of Machine Learning Research, 2019, 20(55): 1-21.
- [28] Vaswani A, Shazeer N, Parmar N, etc. Attention is all you need[J]. Advances in neural infor-

- pation processing systems, 2017, 30.
- [29] Wu Q, Yang T, Liu Z, etc. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 14561-14571.
 - [30] He K, Chen X, Xie S, etc. Masked autoencoders are scalable vision learners[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 16000-16009.
 - [31] Blatter P, Kanakis M, Danelljan M, etc. Efficient visual tracking with exemplar transformers[C]//Proceedings of the IEEE/CVF Winter conference on applications of computer vision. 2023: 1571-1581.
 - [32] Liu S, Fan H, Qian S, etc. Hit: Hierarchical transformer with momentum contrast for video-text retrieval[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 11915-11925.
 - [33] Mayer C, Danelljan M, Yang M H, etc. Beyond SOT: tracking multiple generic objects at once[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024: 6826-6836.
 - [34] Zhang L, Gonzalez-Garcia A, Weijer J V D, etc. Learning the model update for siamese trackers[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 4010-4019.
 - [35] Bhat G, Danelljan M, Gool L V, etc. Learning discriminative model prediction for tracking[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6182-6191.
 - [36] Danelljan M, Bhat G, Khan F S, etc. Atom: Accurate tracking by overlap maximization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4660-4669.
 - [37] Danelljan M, Gool L V, Timofte R. Probabilistic regression for visual tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 7183-7192.
 - [38] Bhat G, Danelljan M, Van Gool L, etc. Know your surroundings: Exploiting scene information for object tracking[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 205-221.
 - [39] Yan B, Peng H, Fu J, etc. Learning spatio-temporal transformer for visual tracking[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10448-10457.
 - [40] Liao B, Wang C, Wang Y, etc. Pg-net: Pixel to global matching network for visual tracking[C]//Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16. Springer International Publishing, 2020: 429-444.
 - [41] Zhou Z, Pei W, Li X, etc. Saliency-associated object tracking[C]//Proceedings of the

- IEEE/CVF international conference on computer vision. 2021: 9866-9875.
- [42] LeCun Y, Bottou L, Bengio Y, etc. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [43] 王明华.基于时空注意力的在线目标跟踪算法研究[D].电子科技大学, 2024.004375.
- [44] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//Proceedings of the fourteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2011: 315-323.
- [45] Ren S, Wu Z, Zhu K Q. Emo: Earth mover distance optimization for auto-regressive language modeling[J]. arXiv preprint arXiv:2310.04691, 2023.
- [46] Meyer G P. An alternative probabilistic interpretation of the huber loss[C]//Proceedings of the ieee/cvf conference on computer vision and pattern recognition. 2021: 5261-5269.
- [47] Bromley J, Guyon I, LeCun Y, etc. Signature verification using a siamese time delay neural network[J]. Advances in neural information processing systems, 1993, 6.
- [48] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.
- [49] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
- [50] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[J]. Advances in neural information processing systems, 2015, 28.
- [51] Woo S, Park J, Lee J Y, etc. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision. 2018: 3-19
- [52] Chen H, Shi Z. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection[J]. Remote Sensing, 2020, 12(10): 1662.
- [53] Y. Wu, J. Lim, M.-H. Yang. Online Object Tracking: A Benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013:2411–2418.
- [54] 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, 2019, 43(5): 1562-1577.
- [55] Fan H, Lin L, Yang F, etc. 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.
- [56] Fan H, Bai H, Lin L, etc. Lasot: A high-quality large-scale single object tracking benchmark[J]. International Journal of Computer Vision, 2021, 129: 439-461.
- [57] Liang P, Blasch E, Ling H. Encoding color information for visual tracking: Algorithms and benchmark[J]. IEEE transactions on image processing, 2015, 24(12): 5630-5644.
- [58] Roffo G, Melzi S. The visual object tracking VOT2016 challenge results[C]//Computer Vision--ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II. Springer International Publishing, 2016: 777-823.

- [59] Kiani Galoogahi H, Fagg A, Huang C, etc. Need for speed: A benchmark for higher frame rate object tracking[C]//Proceedings of the IEEE international conference on computer vision. 2017: 1125-1134.
- [60] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 445-461.
- [61] Wang X, Girshick R, Gupta A, etc. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7794-7803.
- [62] Deng J, Dong W, Socher R, etc. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.
- [63] Lin T Y, Maire M, Belongie S, etc. 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.
- [64] Real E, Shlens J, Mazzocchi S, etc. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video[C]//proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5296-5305.
- [65] Zhang Z, Peng H, Fu J, etc. Ocean: Object-aware anchor-free tracking[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 771-787.
- [66] Cheng S, Zhong B, Li G, etc. Learning to filter: Siamese relation network for robust tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 4421-4431.
- [67] 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.
- [68] Danelljan M, Bhat G, Khan F S, etc. Efficient convolution operators for tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 21-26.
- [69] Fu Z, Liu Q, Fu Z, etc. Stmtrack: Template-free visual tracking with space-time memory networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13774-13783.
- [70] Guo M, Zhang Z, Fan H, etc. Learning target-aware representation for visual tracking via informative interactions[J]. arXiv preprint arXiv:2201.02526, 2022.
- [71] Du F, Liu P, Zhao W, etc. Correlation-guided attention for corner detection based visual tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6836-6845.
- [72] Xu Y, Wang Z, Li Z, etc. 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.

- [73]Ye B, Chang H, Ma B, etc. 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.

攻读学位期间发表的学术成果

- [1] **Cao S**, Zhu B, Kong K, etc. SiamGPF: temporal correlation-based visual tracking[J]. Journal of Electronic Imaging, 2024, 33(4): 043041-043041.
- [2] **Cao S**, Zhu B, Li W, etc. A Siamese tracker based on precise fusion of prior information and multi-layer features[C]//2024 36th Chinese Control and Decision Conference (CCDC). IEEE, 2024: 6280-6285.

致谢

感谢三年来的自己，感谢三年来帮助我的老师同学室友。首先感谢导师刘丙友的细心指导，在研究生三年期间对我们无论是在生活中还是学习上帮助甚多，其次感谢王力超老师的指导，感谢万国扬老师的指导、感谢陶亮老师的指导、感谢胡耀聪老师的指导，感谢郭黎老师的指导、感谢耿树巧老师的指导、感谢罗胜平老师的指导，感谢朱文博老师的指导。感谢师兄周星源、章东旭、代浩文、陈邱卓、储冬亭的帮助，感谢同门包俊涛、林腾飞、缪陆、刘俊楠、刘飞瑶的帮助，感谢室友唐斌、温福新、朱家明的帮助，感谢师弟蒋述、陶晓军、陈旺旺、赵宇涵、吴继强、尹冰、汪冉、刘祥玉、黄志成的陪伴。

尚德敏学 唯实惟新

