融合检测技术的孪生网络跟踪算法综述

张津浦，王岳环

(华中科技大学 人工智能与自动化学院，湖北 武汉 430074)

摘　要：近年来，基于孪生网络的方法在视觉目标跟踪中取得了巨大的进步，但是这类方法在处理跟踪中的目标状态估计以及复杂场景干扰中仍存在较大的提升空间。随着深度学习在目标检测领域取得的成功，越来越多的研究将其成果用于指导目标跟踪技术的发展。对融合检测技术的孪生目标跟踪算法进行了综述。首先介绍检测和跟踪的联系与区别，同时分析检测技术对改进基于孪生网络的跟踪算法的可行性；然后阐述在不同检测框架指导下的孪生目标跟踪算法，以及使用 OTB100、VOT2018、GOT-10k 和 LaSOT 公开数据集对各类算法进行对比和分析；最后对全文进行总结，并对目标跟踪的未来发展方向进行展望。

关键词：目标跟踪；　深度学习；　孪生网络；　目标检测

中图分类号：TP391　　**文献标志码：**A　　**DOI：**10.3788/IRLA20220042

A survey of siamese networks tracking algorithm integrating detection technology

Zhang Jinpu，Wang Yuehuan

(School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China)

Abstract: In recent years, siamese tracking networks have achieved promising performance in visual tracking. However, there is still large room for improvement in the challenge of target state estimation and complex aberrances for siamese trackers. With the success of deep learning in object detection, more and more object detection technologies are used to guide object tracking. This survey reviews the siamese tracking algorithms integrating detection technologies. Firstly, we introduce the relation and difference between detection and tracking, and analyze the feasibility of improving siamese tracking algorithms by detection technologies. Then, we elaborate the existing siamese trackers based on different detection frameworks. Furthermore, we conduct extensive experiments to compare and analyze the representative methods on the popular OTB100, VOT2018, GOT-10k, and LaSOT benchmarks. Finally, we summarize our manuscript and prospect the further trends of visual tracking.

Key words: object tracking;　deep learning;　siamese network;　object detection

收稿日期:2022−01−13；　**修订日期:**2022−03−22

作者简介:张津浦，男，博士生，主要从事深度学习、目标跟踪、目标检测方面的研究。

导师简介:王岳环，男，教授，博士生导师，主要从事精确制导、计算机视觉、实时自动目标识别等方面的研究。

0 引 言

目标跟踪是计算机视觉领域的研究热点之一，其基本流程为：给定视频序列初始帧的目标框，推断后续帧中该目标的位置及形状。目标跟踪技术被广泛应用于自动驾驶、视频监控、人机交互和无人机侦察等领域[1−2]。尽管在上述领域已经取得了巨大的成就，但由于跟踪场景的复杂多样，以及目标运动过程中的形变、遮挡、运动模糊、光照变化、尺度变化、快速移动等情况[3]，当前跟踪算法并不能适应所有跟踪场景，因此研究高性能鲁棒的目标跟踪算法依然是一个富有挑战性的任务。

目前，主流的目标跟踪算法可以分为基于相关滤波的跟踪算法和基于孪生网络的跟踪算法[4]。基于相关滤波的方法属于判别式方法，通过岭回归将跟踪问题转化为前景和背景的二分类问题。传统的相关滤波算法[5−7]采用 HOG[8]、CN[9] 等手工特征表达，速度快，可以在 CPU 上实时运行，但精度一般。结合深度特征的相关滤波算法[10−12]可以获得更强健的特征表达，但会引入较大的计算负担，导致速度大幅下降且难以部署。此外，这类方法使用在 ImageNet[13] 等数据集上离线训练好的模型进行特征提取，无法端到端地根据跟踪任务来优化整个模型。基于孪生网络的跟踪算法[14−16]使用两个共享参数的分支[17]，将跟踪问题转换为模板和搜索区域的相似性度量问题，并且可以端到端进行优化，在精度和速度上取得了较好的平衡。在最近的研究中[18−20]，基于孪生网络的跟踪方法的精度已经超过结合深度特征的相关滤波方法，并且更加容易在边缘设备上部署，因此成为了当下的研究热点。

目标跟踪与目标检测密切关联，Wang[21] 将目标跟踪视为一种特殊的检测任务—实例检测。跟踪与检测的联系和区别如图 1 所示。两者都是在复杂场景中识别目标并进行精确定位。区别在于目标检测包含若干预定义的类别，检测器只检测这些指定类别的对象，并且其输出不区分类内的实例；而目标跟踪是类别不可知的，只在初始帧中给定任意实例，并在后续帧中查找该特定实例。因此，通过适当的初始化，检测器可以从单个图像中学习新的实例来快速转换为跟踪器。近年来，随着深度学习理论在目标检测领域的成熟应用[22]，越来越多的研究者借鉴目标检测方法用于指导跟踪器的设计，弥补现有跟踪方法的缺陷。

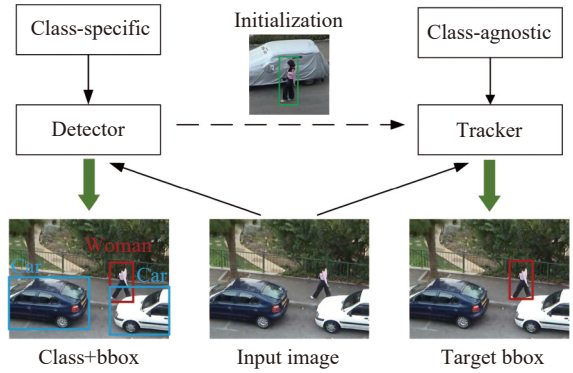

图 1 目标检测与目标跟踪的联系和区别

Fig.1 The relation and difference between object detection and object tracking

文中按照深度学习目标检测框架的分类，包括状态估计方式 (有锚框/无锚框)、阶段数 (单阶段/两阶段) 以及其他类别，对不同检测框架指导下的孪生目标跟踪算法进行完整综述，并根据这类算法在 OTB100、VOT2018、GOT-10 k 和 LaSOT 数据集上的结果进行分析，旨在通过目标检测技术解决跟踪中的关键问题，为目标跟踪的进一步发展提供参考。

1 孪生网络与目标跟踪

孪生网络 (siamese network)[17] 是一种相似性度量方法，近年来，基于孪生网络的方法以其优越的性能和速度引起了广泛的关注[23−24]。SiamFC[14] 是早期最具代表性的基于孪生网络的跟踪方法，网络结构如图 2 所示。孪生网络将目标模板 z 与候选搜索图像 x 送入两个共享权重的 CNN 分支 φ，然后用互相关函数 g 度量二者的相似性，记作 $f(x,z)=g(\varphi(z),\varphi(x))$，相似性最大的位置即为目标所在位置。SiamFC 充分利用离线数据来端到端学习表示物体运动和外观的通用匹配关系，并可以用来定位训练中未曾见过的目标。

SiamFC 提出后受到了广泛的关注，许多跟踪方法都在其基础上进行改进。CFNet[15] 将相关滤波器作为可微分层嵌入到孪生网络框架中进行端到端学习。SA-Siam[25] 设计了融合表观特征和语义特征的双重孪生网络，提升孪生网络的泛化能力。RASNet[26] 将注意力机制引入孪生网络学习更具判别能力的特

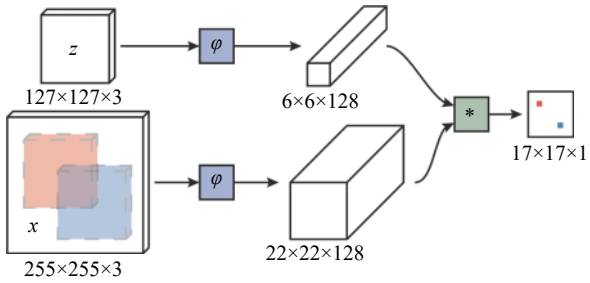

图 2 SiamFC 网络结构

Fig.2 The architecture of SiamFC

征。Dong[27] 利用三元组损失进一步挖掘样本之间的潜在关系，获得更好的训练效果。Cui[28] 设计通道-互联-空间注意力模块，增强模型的适应能力和判别能力。

尽管上述基于孪生网络及其改进的跟踪算法发展迅速，但其存在的一些不足使得精度不如同期的相关滤波方法。主要包括以下问题：(1) 采用多级金字塔的方式进行尺度估计不精确，目标框比例无法改变，且计算量大；(2) 容易受到语义相似物的干扰，如图 3 所示，搜索区域中所有具有语义的物体都有一个较大的响应；(3) 只利用第一帧的模板信息，性能完全依赖孪生网络的泛化能力，较难应对复杂场景下的目标变化。

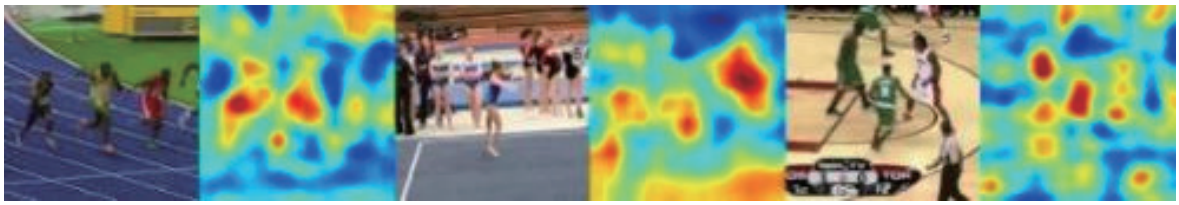

图 3 干扰物对 SiamFC 的影响

Fig.3 The impact of distractors on SiamFC

针对上述问题，研究人员发现，目标检测中的边框回归技术对于目标状态估计具有卓越的性能；两阶段结构能够通过分段处理更细致地捕获样本间的差异；独立的定位质量评价确保了分类和回归的一致性等。这些检测技术均能有效弥补孪生跟踪模型的局限，因此，文中将对这些借鉴检测的模型和思想对孪生跟踪器进行改进的工作进行总结。

2 融合检测技术的孪生跟踪算法

该节详细回顾和分析融合检测技术的孪生跟踪算法，笔者按照状态估计 (state estimation) 和阶段数 (stage number) 对现有跟踪方法进行分类，其他一些特殊的方法 (others) 则单独划为一类进行介绍，整体框架如图 4 所示。其中对于有锚框 (anchor-based) 结构主要包括基于 RPN (RPN-based) 的方法，对于无锚框结构分为基于 FCOS (FCOS-based) 和基于关键点 (key-point-based) 两类方法；在单阶段 (one-stage) 方法中介绍模型更新对孪生跟踪器的优化；其他方法则被划分为基于 IOUNet 的状态估计 (IOUNet-based prediction) 和检测器直接转化跟踪器 (Detector transform tracker)

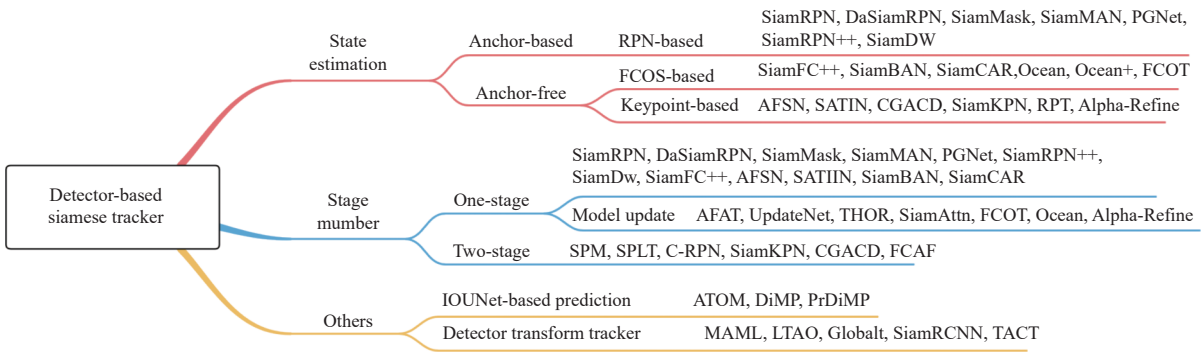

图 4 融合检测技术的孪生跟踪器分类框架

Fig.4 Taxonomy of detector-based siamese tracking methods

两类。

2.1 按状态估计分类

2.1.1 有锚框

SiamRPN[29] 是首个将检测网络融入孪生跟踪器的工作，借鉴了 Faster RCNN[30] 的区域建议网络 (RPN)，在每个滑动窗口位置设置多个锚框，并预测锚框的偏移量进行状态估计。这种方式可以预测具有多种尺度和宽高比的预测框，避免了多尺度搜索。SiamRPN 结构如图 5 所示，由孪生网络和 RPN 两部分组成，二者通过卷积升维操作统一在端到端的框架里面。RPN 输出一个区分目标和背景的分类分支和一个预测目标尺度的回归分支。推理阶段是一个单次检测 (one-shot detection) 的过程，能够超实时运行，在 GTX 1060 中达到 160 FPS。

SiamRPN 的提出使得检测中的锚框机制和 RPN 逐渐成为孪生目标跟踪的新范式，后续有许多研究者对其进行进一步改进[31−35]。DaSiamRPN[35] 引入了难负样本挖掘技术，通过在训练过程中加入具有语义的困难负样本来克服非语义背景和语义干扰之间的数据不平衡问题。SiamMask[31] 增加了掩码预测分支，统一了视频目标跟踪和视频目标分割两个任务，借助分割得到更加准确的旋转目标框。SiamMan[33] 增加了定位分支，可以更好地适应不同目标的运动模式，并减少对预设锚框的依赖。

主流检测方法均使用语义判别能力更强的深层网络如 ResNet[36] 进行特征提取，但是孪生跟踪方法在直接使用深层网络后性能反而下降，原因是深层网络依赖填充操作 (padding) 保证输出特征的分辨率，而填充会破坏孪生架构的空间平移不变性。为了解决这一问题，SiamDW[37] 设计了一个残差裁剪模块去除多余的填充；而 SiamRPN++[38] 设计了一种简单有效的空间感知采样策略，让正样本以均匀分布的采样方式在中心点附近进行偏移，消除填充带来的空间位置偏差。此外，SiamRPN++还提出了类似 FPN[39] 的分层聚合模块学习从浅到深的丰富特征表示以及更轻量级的深度互相关来平衡模板分支和搜索分支的参数。SiamRPN++是数据驱动 (端到端学习) 的深度学习方法在性能上第一次超过相关滤波的方法，并且在 GPU 上能够实时运行。

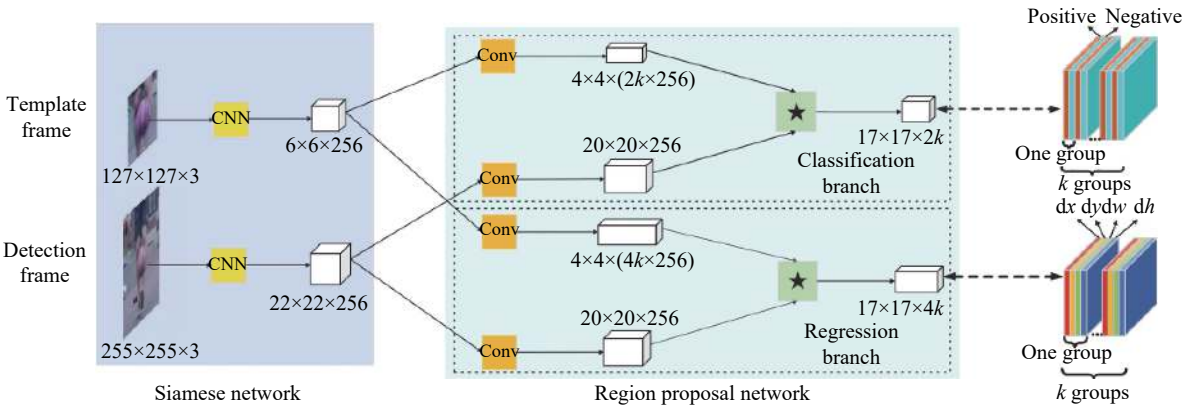

图 5 SiamRPN 结构

Fig.5 The architecture of SiamRPN

2.1.2 无锚框

尽管基于锚框和 RPN 的孪生跟踪器取得了卓越的性能，但其仍然存在一些局限：

(1) 锚框类方法的分类分数衡量的是锚框和目标的相似程度而不是目标本身的置信度。这样容易产生假阳性结果，即可能在非目标区域产生一个不合理的高分；

(2) 锚框的预设依赖大量先验知识 (尺度/长宽比)，需要启发式地调整超参数，影响泛化能力；

(3) 锚框类方法的回归分支训练只针对 IOU 大于一定阈值 (0.6) 的锚框，实际跟踪中当锚框与目标重叠率较小时难以进行调整，缺乏修正弱预测的能力[19]；

(4) 目标的状态估计质量只能使用分类置信度来评价，缺乏独立的质量评估方式[40−41]。

针对上述问题，目标检测中的无锚框回归方式被引入跟踪，它避免了与锚框相关的超参数，更具灵活性和通用性。当前无锚框回归方法主要包含基于 FCOS 和基于关键点两种。

(1) 基于 FCOS 的方法

SiamFC++[41] 首次提出借鉴 FCOS[42] 检测器的无锚框思想来解决上述问题，如图 6 所示。SiamFC++是基于像素级预测的模型，将特征图上每个位置作为训练样本，并预测正样本点相对真实标注框四条边的距离，因此避免了锚框的不匹配问题以及对超参的依赖。此外，无锚框回归在训练过程中考虑了标注框内的所有像素，即使只有一小块区域被识别为前景，也可以预测目标的尺寸。因此，跟踪器能够在一定程度上纠正推理过程中较差的预测。作者还提出了先验空间得分 (Prior Spatial Score) 评估状态估计的质量，抑制距离目标中心较远的位置产生的低质量预测框。SiamFC++凭借更简单灵活的结构和更快的推理速度为后续无锚框孪生跟踪器的发展奠定了基础。

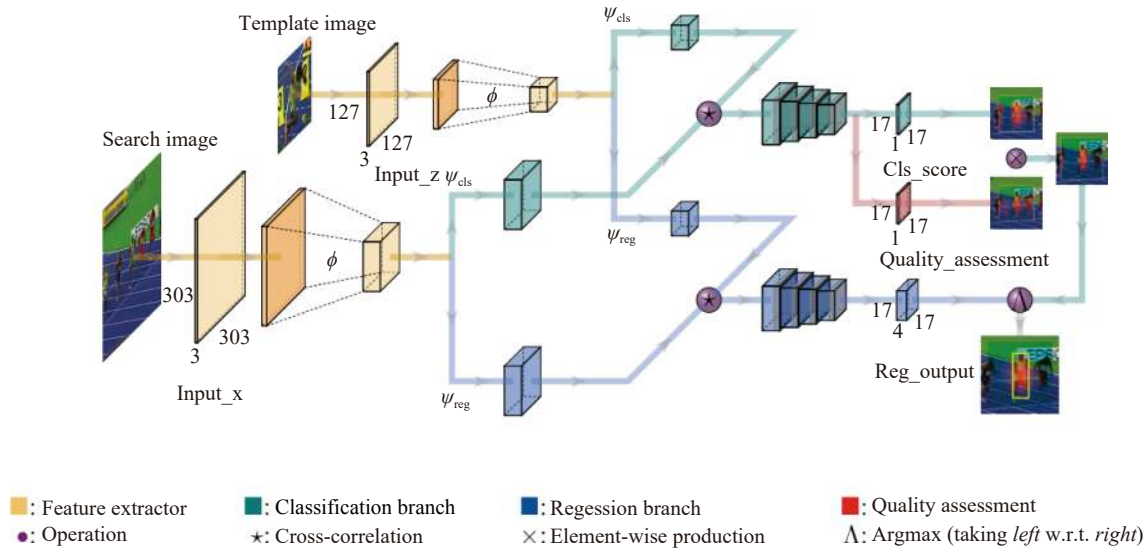

图 6 SiamFC++结构

Fig.6 The architecture of SiamFC++

同时期类似的工作还包括参考文献 [19,40,43−45]。SiamBAN[43] 在训练划分正负样本时设置了椭圆标签，相比传统矩形标签能更准确地标注正负样本。Ocean[19] 额外设计了一个特征对齐模块，将分类分支卷积核的采样位置与边框回归结果对齐，以此来更好地学习对目标敏感的特征并适应尺度的变化。OceanPlus[44] 在 Ocean 的基础上添加注意力检索网络和多分辨率多阶段分割网络进行实例分割。LightTrack[20] 则关注移动端轻量化网络设计，使用神经架构搜索 (NAS) 来设计更轻量级和高效的无锚框跟踪器。

(2) 基于关键点的方法

另一个发展分支则是借鉴 CenterNet[46] 和 CornerNet[47]的关键点 (中心点/角点) 预测模型用于孪生跟踪器中。最早被提出的方法是 SATIN[48]，分别预测了目标的中心点和左上右下两个角点。ASFN[49] 则仿照 CenterNet 预测目标的中心点、中心偏移和尺度。Du[50] 认为传统的互相关操作无法编码角点的空间信息，提出使用逐像素相关 (pixel-wise correlation)，将模板特征的每个位置都与搜索特征计算相似性，类似的工作还包含 Alpha-Refine[51]。

VOT2020[24] 短时赛道的冠军 RPT[52] 受到 Reppoints[53] 的启发，将跟踪目标状态表示为特征点集 (包括语义关键点与边界极值点)，以提升对目标位姿变化、几何结构变化的建模能力。

2.2 按阶段数分类

2.2.1 单阶段

单阶段孪生跟踪器的结构类似单阶段检测器，均在特征图上的所有位置进行密集采样，生成候选框 (锚框) 或直接像素级 (无锚框) 预测分类和回归结果，

主要方法在 2.1 节均有详细介绍。

上述单阶段方法将跟踪问题转化为单帧的独立检测问题，目标模板仅在第一帧初始化并保持不变，跟踪器性能完全依赖模型的泛化能力。当目标发生较大外观变化时，不更新模型往往导致跟踪失败。而得益于单阶段结构的简洁，跟踪社区提出了许多将模型更新融入孪生跟踪算法的解决方案[54–58]。Zhang[57]提出了 UpdateNet 的卷积神经网络更新方式，综合利用第一帧模板、历史累计模板和当前帧模板共同学习，生成最优模板。UpdateNet 解决了线性更新导致的模板信息随时间推移指数衰减的问题，并且针对不同特征维度以不同的程度进行更新，增强了对各种动态变化的适应性。SiamAttn[55]提出一种交叉注意力机制，将搜索分支中丰富的上下文信息编码到模板分支中，提供一种隐式的模板更新。THOR[54]构建了全局动态的目标表示方法，通过提取多个在特征空间中距离尽可能远的模板制作模板集，扩充被跟踪对象特征的多样性。AFAT[56]设计了质量预测网络，通过卷积和 LSTM 从多帧响应映射中提取隐式决策信息，可以从时空角度对潜在的跟踪失败进行可靠和稳健的预测。FCOT[45]则利用无锚框的简洁结构，对回归分支进行在线优化，使跟踪器能更有效处理目标的形变。

2.2.2 两阶段

目标跟踪存在的一大矛盾在于难以平衡跟踪器的鲁棒性 (适应目标外观变化) 和强判别性 (对相似物不漂移)。孪生跟踪器属于模板类方法，正如第 1 节提及的这类方法对于语义干扰物的判别能力较差，而使用两阶段方法进行由粗到细的匹配可以有效缓解这一问题。

SPM[59]受 Faster-RCNN 的两阶段结构启发，将跟踪的鲁棒性和判别性分成两个阶段训练。粗匹配阶段会输出若干个得分最高的候选目标结果送入精匹配阶段。精匹配阶段通过少样本学习[60]区分目标和背景相似物并进行边框回归。最后将两个阶段的输出加权融合。SPLT[61]采用类似的思想，并在此基础上添加了重检测模块用于长时跟踪任务。Zhang[62]等人在两阶段跟踪的第一阶段通过相关滤波调制自适应更新模板，结合时域信息过滤掉易区分的负样本。CGACD[50]则设计了无锚框的两阶段角点检测网络，目的是更好地区分目标和背景物体的角点。

Fan[63]提出了一种多级跟踪框架 C-RPN，通过级联多个 RPN 实现逐层的难负样本采样来解决正负样本不平衡问题，同时充分挖掘各层的特征来实现鲁棒的视觉跟踪。类似的，SiamKPN[64]级联了多个无锚框的关键点预测结构，通过逐渐缩小标签热力图的覆盖范围实现由粗到细的匹配。

2.3 其他

上述介绍的有锚框/无锚框，单阶段/两阶段方法都是基于类似图 2 的具有结构对称性的孪生网络。小节将补充一些使用孪生网络并行架构但是非严格对称的类孪生跟踪器[65–72]，它们同样借鉴了目标检测技术。

(1) 基于 IOUNet 的状态估计

Martin[72,65]认为不能简单地用分类质量来衡量目标状态估计的质量，并借鉴 IOUNet[73]的思想通过 IOU 来评价状态估计结果，整体框架如图 7 所示。与之前的孪生跟踪方法不同，参考分支最后生成的不是特征图，而是经过编码的调制向量。测试分支结合生成的调制向量，预测跟踪框和真值之间的 IOU，并通过梯度上升的方式使 IOU 最大化来得到精细的预测框。PrDiMP[66]在此基础上提出了一种基于概率的回归方法，预测目标状态的条件概率密度，并对标注的噪声和不确定性进行建模。通过最小化二者的 KL 散度来训练网络，使其能够表达目标状态估计中的不确定性。

(2) 检测器转化跟踪器

另外一类方法[67–71]直接用检测器进行跟踪，将目标实例的信息以某种方式编码到待检测图像中，从而将类别感知的检测任务转变成实例感知的跟踪任务。Huang[71]提出一个通用的框架来缩小检测与跟踪之间的差别，整体是 Faster-RCNN 结构，利用元学习 MAML[74]在较少的样本和少量的迭代下学习一个实例分类器区分目标和干扰。同样利用元学习初始化检测器的还有参考文献 [21]，并借鉴 MAML++[75]和 MetaSGD[76]使元学习的训练更加稳定。

GlobalTrack[68]在 Faster-RCNN 的 RPN 部分和预测头部分都加入了目标信息来引导检测网络搜索特定实例，避免了元学习更新带来的不稳定，所以更加适合长时跟踪任务。LTAO[70]端到端地训练了一个线性分类器的权重作为引导信息，具有更强的判别能力。Siam RCNN[69]将第一帧目标标注和上一帧检测结果共同作为引导信息进行重检测，并设计了一种基于轨迹的动态规划算法，能够在长时遮挡后重新检测

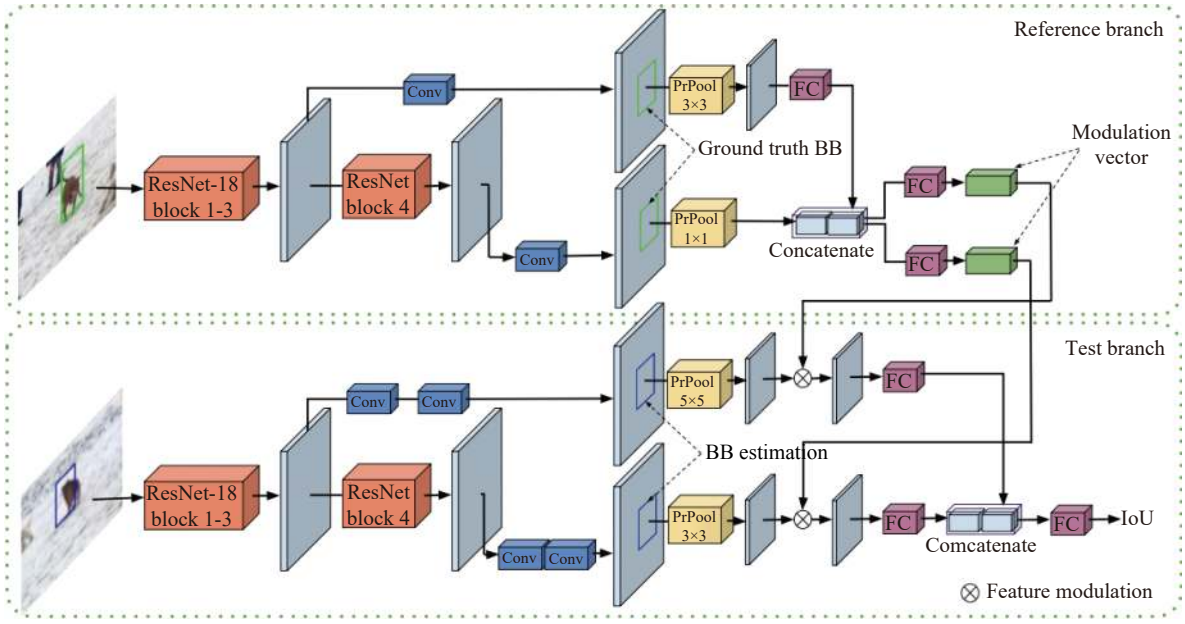

图 7 IOU-Predictor 结构

Fig.7 The architecture of IOU-Predictor

被跟踪对象。TACT[67] 在两阶段结构上增加了三叉戟对齐 (TridentAlign) 模块将目标的特征映射到多个空间维度中，形成特征金字塔来适应尺度变化。上述使用检测器做跟踪的方法都可以在全图进行搜索，更有利于应对一些剧烈运动及长时跟踪的重捕获。

3 实验与分析

该节将在 OTB100[3]、VOT2018[77]、GOT-10 k[78] 和 LaSOT[79] 四个公开数据集上对上述 40 多个跟踪算法进行全面评估。首先对数据集和相应的性能评估方法进行介绍，然后对实验结果进行对比和分析。所有测试结果均来自原论文或官方源码。

3.1 数据集和评估方式

(1) OTB100

Wu[3] 等人 2015 年提出的 OTB100 是目前最为常用的跟踪数据集之一。该数据集包含 100 个完全标注的视频序列，涉及目标跟踪的 11 种属性，包括光照变化、尺度变化、遮挡、形变、运动模糊、快速运动、平面内旋转、平面外旋转、出视野、背景干扰和低分辨率。OTB 的评价指标为距离精度 (Distance Precision) 和重叠成功率 (Overlap Success)，测试时采用一次通过评估 (One-Pass Evaluation, OPE)。

(2) VOT2018

VOT2018[77] 数据集包含 60 个旋转框标注序列，涵盖遮挡、光照变化、运动变化、尺度变化、相机运动和空闲 6 种属性。VOT 具有重启机制，当重叠率为 0 时，跟踪器会被重新初始化。VOT2018 的评价指标为精确性 (Accuracy)、鲁棒性 (Robustness) 和 EAO (Expected average overlap)。

(3) GOT-10 k

GOT-10 k[78] 是一个通用大规模目标跟踪数据集，包含超过 10 K 个视频序列，563 个类别和超过 150 万个标注框，尽可能多地涵盖具有挑战性的现实场景。GOT-10 k 训练集和测试集不存在交集，保证模型的泛化能力。评价指标为平均重叠率 (Average Overlap, AO) 和成功率 (Success Rate, SR)。

(4) LaSOT

LaSOT[79] 包含 1 400 个视频和超过 3.5 M 手工标注图片，是目前最大的密集标注单目标跟踪数据集。该数据集包含 70 个类别，每个类别包含 20 个序列，每个序列平均 2 512 帧，偏重长时跟踪任务且难度相对较大。LaSOT 划分 280 个序列用于测试，评价方式类似 OTB，并增加一个归一化精度 (Normalized Precision) 指标。

3.2 定量结果

表 1 展示了所有跟踪算法的定量比较结果。对于 OTB100 和 LaSOT，按成功率 (AUC) 取 top5，OTB100 上的排名是 RPT, DROL, CGACD, SiamRCNN, Siam-

表 1 跟踪算法在 OTB100, LaSOT, GOT-10 k 和 VOT2018 上的性能对比

Tab.1 Performance comparison of siamese tracking methods on OTB100, LaSOT, GOT-10 k and VOT2018

	TYPE		OTB100		LaSOT		GOT10 k			VOT2018		
	A	S	AUC	PR	AUC.	NPR	AO	SR0.50	SR0.75	A	R	EAO
SiamRPN [29]	T	1	0.637	0.851	0.457	-	-	-	-	-	-	-
DaSiamRPN [35]	T	1	0.658	0.88	0.415	0.496	-	-	-	0.59	0.276	0.383
SiamRPN++ [38]	T	1	0.696	0.915	0.496	0.569	0.518	0.618	0.325	0.6	0.234	0.414
SiamDW [37]	T	1	0.674	0.923	0.384	0.476	0.416	-	-	-	-	0.27
SiamMask [31]	T	1	-	-	-	-	0.514	0.587	0.366	0.61	0.276	0.38
SiamMan [33]	T	1	0.705	0.919	-	-	-	-	-	0.605	0.183	0.462
THOR [54]	T	1	0.648	0.791	-	-	0.447	0.538	0.204	0.582	0.234	0.416
DROL [58]	T	1	**0.715**	**0.934**	0.537	0.624	-	-	-	0.616	-	0.481
SiamAttn [55]	T	1	0.712	0.926	0.56	0.648	-	-	-	**0.636**	0.16	0.47
AFAT [56]	T	1	0.663	0.874	0.492	0.574	-	-	-	0.605	0.239	0.419
UpdateNet [57]	T	1	-	-	0.475	0.56	-	-	-	-	-	0.393
SiamFC++ [41]	F	1	0.683	0.896	0.544	0.623	0.595	0.695	0.479	0.587	0.183	0.426
AFSN [49]	F	1	0.675	0.868	-	-	-	-	-	0.589	0.204	0.398
SATIN [48]	F	1	0.641	0.844	-	-	-	-	-	-	-	-
SiamBAN [43]	F	1	0.696	0.91	0.514	0.598	-	-	-	0.597	0.178	0.452
SiamCAR [40]	F	1	0.697	0.91	-	-	0.569	0.67	0.415	-	-	-
CGACD [50]	F	1	**0.713**	0.922	0.518	0.626	-	-	-	0.615	0.173	0.449
FCAF [80]	F	1	0.649	0.86	-	-	-	-	-	-	-	-
FCOT [45]	F	1	0.693	0.913	0.569	**0.678**	**0.64**	**0.763**	**0.517**	0.6	**0.108**	**0.508**
PGNet [34]	F	1	0.691	0.892	0.531	0.605	-	-	-	0.618	0.192	0.447
Ocean [19]	F	1	0.684	0.92	0.56	-	0.611	0.721	0.473	0.592	**0.117**	**0.489**
Ocean+ [44]	F	1	-	-	-	-	-	-	-	-	-	-
RPT [52]	F		**0.715**	**0.936**	-	-	0.624	**0.73**	0.504	0.629	**0.103**	**0.51**
AlphaRef [51]		1	-	-	**0.589**	**0.649**	-	-	-	**0.633**	0.136	0.476
SiamKPN [64]	F	2	0.712	**0.927**	0.498	-	0.529	0.606	0.362	0.606	0.192	0.44
SPLT [61]	T	2	-	-	0.426	0.494	-	-	-	-	-	-
CRPN [63]	T	2	0.663	-	0.455	0.542	-	-	-	-	-	-
SPM [59]	T	2	0.687	0.889	0.485	-	0.513	0.593	0.359	0.58	0.3	0.338
TACT [67]	T	2	-	-	0.575	0.66	0.578	0.665	0.477	-	-	-
SiamRCNN [69]	T	2	0.701	0.891	**0.648**	**0.722**	**0.649**	0.728	**0.597**	0.609	0.22	0.408
GlobalT [68]	T	2	-	-	0.521	0.599	-	-	-	-	-	-
LTAO [70]	T	2	-	-	-	-	-	-	-	-	-	-
ATOM [72]	*others*		0.667	0.879	0.514	0.576	0.556	0.635	0.402	0.59	0.204	0.401
DiMP [65]	*others*		0.686	0.899	0.569	0.648	0.611	0.717	0.492	0.597	0.153	0.44
PrDiMP [66]	*others*		0.696	0.897	**0.598**	-	**0.634**	**0.738**	**0.543**	0.618	0.165	0.442
SSD-MAML [71]	*others*		0.62	-	-	-	-	-	-	-	-	-
FRCNN-MAML [71]	*others*		0.647	-	-	-	-	-	-	-	-	-
FCOS-MAML [21]	*others*		0.704	0.905	0.523	-	-	-	-	**0.635**	0.22	0.392
Retina-MAML [21]	*others*		0.712	0.926	0.48	-	-	-	-	0.604	0.159	0.452

Note: **Bold** fonts are ranked top-3. '-' means the corresponding results are not given in the original literature. 'TYPE' is the classification basis delineated in this paper, where 'A' indicates the Anchor (Anchor-based 'T '/Anchor-free 'F'), 'S' indicates the Stage number (One-stage '1'/Two-stages '2 '), and 'others' indicates other classes.

CAR; LaSOT 上的排名是 SiamRCNN, PrDiMP, TACT, FCOT, DiMP。按精度 (PR) 排名, OTB100 的前五名是 RPT, DROL, SiamDW, CGACD, Ocean; 而 LaSOT 的前五名是 SiamRCNN, PrDiMP, FCOT, TACT, Ocean。从结果可以发现, 对于 LaSOT 这类较长的视频序列, 排名靠前的算法大多依赖两阶段结构和模型更新。两阶段结构对于鲁棒性和判别性的平衡能有效应对长时跟踪中出现的干扰物以及模型漂移, 而判别式的更新方法也能及时处理目标和场景的各类变化。

对于 VOT2018, 精度 (A) 领先的是 SiamAttn, Alpha-Refine, RPT, PGNet, DROL; 鲁棒性 (R) 领先的是 RPT, FCOT, Ocean, Alpha-Refine, DiMP; EAO 领先的则是 RPT, FCOT, Ocean, DROL, Alpha-Refine。VOT2018 的重启机制使得鲁棒性指标的波动范围很大 (第一名和最后一名的精度差距 0.056, 鲁棒性差距 0.197)。领先的方法大多为灵活的无锚框结构, 它们对 IOU 较小的预测框有更强的矫正能力, 从而避免跟踪失败重启。

对于 GOT-10 k, 平均重叠率 (AO) 领先的是 SiamRCNN, FCOT, PrDiMP, RPT, DiMP; IOU 阈值为 0.5 的成功率 (SR0.50) 排名为 FCOT, PrDiMP, RPT, SiamRCNN, DiMP; IOU 阈值为 0.75 的成功率 (SR0.75) 排名为 SiamRCNN, PrDiMP, FCOT, RPT, DiMP。不难看出, 对边框预测做了特殊处理 (如两阶段预测、不确定性预测、在线优化、关键点表示等) 的方法在 SR0.75 上效果普遍较好。

表 2 不同检测技术用于孪生目标跟踪算法的优缺点对比

Tab.2 Comparison of advantages and disadvantages of siamese trackers with different detection techniques

Taxonomy		Advantage	limitation
State estimation	Anchor-based	First Introducing RPN detection technology;Discarding multi-scale search, and can predict bbox with arbitrary aspect ratio	Relying on prior knowledge;Incapable of rectifying weak prediction
	Anchor-free	Fewer parameters and faster speed;Correcting weak predictions caused by deformation and fast movement	Requiring additional constraints (such as location quality) due to the lack of prior knowledge
Stage number	One-stage	Fast speed;; Easy to add additional modules (e.g. model updates)	Weak discriminability for semantic interference
	Two-stage	Better balance of robustness and discriminability	Complex structure and slow speed
Others	IOUNet-based prediction	More accurate evaluation of location quality	-
	Detector transform tracker	Narrowing the differences between detection and tracking with a common pattern to solve both problems	-

3.3 讨论

综合上述方法描述以及实验分析, 按照文中的分类方式总结了不同检测技术对于孪生目标跟踪算法的优缺点, 如表 2 所示, 并依此归纳出融合检测技术的孪生目标跟踪算法的六条设计经验: (1) 检测网络的预测头部结构可以提升目标状态估计的精度; (2) 无锚框结构相比有锚框结构对于目标形变具有更强的适应性; (3) 两阶段结构面对复杂干扰场景具有更强的判别能力, 而单阶段结构的速度更快; (4) 将时序信息融入检测框架能更好地处理目标和场景的变化; (5) 对状态估计质量单独进行评估可以进一步提升预测目标框的精度; (6) 检测器具有直接转变成跟踪器的潜力。这些经验可以为后续研究者设计跟踪算法提供一定的指导。

4 问题与展望

随着孪生网络和目标检测技术的结合, 目标跟踪领域在尺度估计、抗复杂环境干扰等方面产生了巨大的进步, 但面对复杂环境设计出高精度、高鲁棒性和实时性的跟踪算法仍然有很多困难。根据已有的研究方法、实验结果和最新的研究思路, 笔者对目标跟踪下一步待解决的问题与未来研究方向进行展望。

(1) 目标状态估计的不确定性

在复杂场景中, 边界框的表示具有很强的不确定性, 这会使标注和边界框回归函数的学习变得困难。如图 8 所示, 非刚性形变、遮挡和运动模糊均使得边界框难以划定。

Martin[81] 提出一种基于概率的回归方法, 预测了

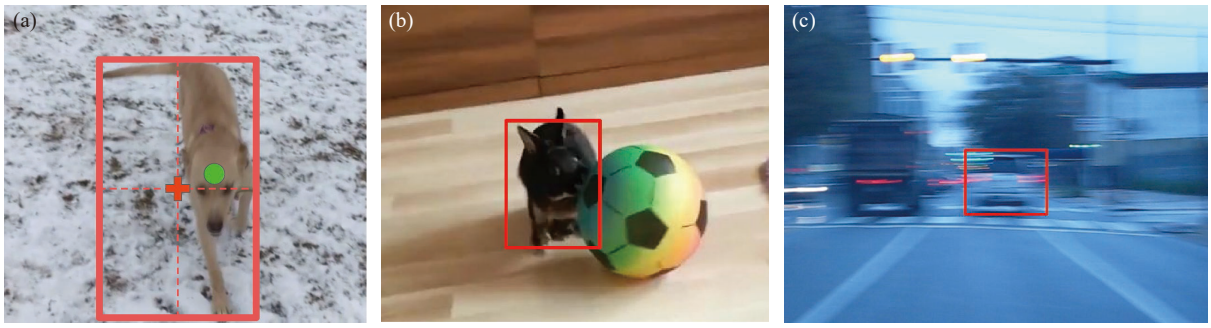

图 8 具有不确定性的边界框。(a) 非刚性形变; (b) 遮挡; (c) 运动模糊

Fig.8 Boundaries with uncertainty. (a) Non-rigid deformation; (b) Occlusion; (c) Motion blur

目标状态的条件概率密度,对来源于不准确标注和任务中模糊情况的标签噪声进行建模。而在目标检测领域中也有大量关于状态不确定性估计的研究[82−85],如何将这种不确定性估计应用于跟踪中,对于进一步提升目标跟踪的状态预测精度有着重要意义。

(2) 训练样本的不平衡

目标跟踪网络在训练时一张图像中仅包含一个正样本,样本不平衡问题相比检测更加严重,网络从大量简单背景或语义干扰中学到的信息的判别能力较弱,直接简单迁移检测模型到跟踪任务中不能完全发挥其优势。Oksuz[86] 总结了目标检测中的各种不平衡问题,包括类别不平衡、尺度不平衡、空间不平衡和多任务损失优化不平衡,并从采样方式、特征、损失函数和生成方法上给出了不同的解决方法。目标跟踪任务同样可以从这些角度出发,研究适合解决跟踪训练中不平衡问题的独特方法,进一步提升数据驱动能力。

(3) 跟踪的域自适应

孪生网络依赖大量离线数据训练相似度度量,对于训练集中未包含的类别,学到的相似度度量不一定可靠,导致泛化能力差。而理想的跟踪器应该具有域自适应能力,在面对类别未知的序列时,能够仅通过少量样本,快速适应特定的目标实例。文中指出了最近一些研究[71,21] 利用元学习初始化检测器,能充分利用初始帧的信息,降低了训练集偏置产生的负面影响。因此,研究元学习等域自适应方法,有助于提高网络模型在目标跟踪任务中的泛化能力,将是未来跟踪领域的重要研究方向。

(4) 其他领域经验的相互借鉴

目标检测的成功经验给目标跟踪带来了许多启发。未来可以持续借鉴包括目标检测、目标分割、少样本学习等领域的思想或相关模型用于目标跟踪领域。同样的,目标跟踪的成果也可以反馈到其他领域,如在视频目标检测或视频目标分割等任务中,利用跟踪对时序关系的建模可以减少漏检误检,进一步提升检测或分割的精度。

最近,NLP 领域中的 Transformer[87] 由于其建立长距离关联和聚合全局信息的优秀能力在多项视觉任务中取得了成功[88]。这项技术同样可以用于目标跟踪。TransT[89] 和 Stark[90] 利用注意力机制取代孪生网络的互相关,解决局部线性的互相关操作缺乏语义和全局信息的瓶颈。TMT[91] 使用 Transformer 进行特征增强,分别将其运用在 siamese 跟踪器和在线的 DiMP 中。SwinTrack[92] 基于 Swin Transformer 设计了一个全部由注意力机制组成的跟踪方法,具有优秀的跟踪性能和速度。可以预见,Transformer 会是未来一段时间的研究热点。

5 结 论

文中回顾了最近热门的融合检测技术的孪生目标跟踪方法,并通过大量实验对其进行评价。这项工作的主要贡献有三个方面。首先,按照状态估计 (有锚框/无锚框),阶段数 (一阶段/两阶段) 和其他几个方面综述了现有的融合检测技术的孪生跟踪器,并从各个角度对这些跟踪器进行了讨论。其次,在主流的 OTB100、VOT2018、GOT-10 k 和 LaSOT 数据集上进行了广泛的实验,比较了具有代表性的方法。这种大规模的评估有助于读者理解检测框架对视觉跟踪的好处。第三,通过对这类方法的发展历史和实验结果进行分析,从目标状态估计的不确定性、训练样本的

不平衡、跟踪的域自适应和其他领域经验的相互借鉴几个方面对目标跟踪存在的问题进行总结，并对未来的发展方向进行展望。

参考文献：

[1] Laurense V A, Goh J Y, Gerdes J C. Path-tracking for autonomous vehicles at the limit of friction[C]//Proceedings of the American Control Conference, 2017: 5586-5591.

[2] Wang Y H, Chai H W, Yang D Y. Improved KCF real-time target tracking algorithm [J]. ***Journal of Huazhong University of Science and Technology***, 2020, 48(1): 5. (in Chinese)

[3] Wu Y, Lim J, Yang M H. Object tracking benchmark [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2015, 37(9): 1834-1848.

[4] Li P, Wang D, Wang L, et al. Deep visual tracking: Review and experimental comparison [J]. ***Pattern Recognition***, 2018, 76: 323-338.

[5] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010: 2544-2550.

[6] Henriques J F, 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.

[7] Danelljan M, Hager G, Khan F S, et al. Discriminative scale space tracking [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2017, 39(8): 1561-1575.

[8] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005: 886-893.

[9] Van De Weijer J, Schmid C, Verbeek J. Learning color names from real-world images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[10] Ma C, Huang J B, Yang X, et al. Hierarchical convolutional features for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 3074-3082.

[11] Danelljan M, Robinson A, Shahbaz Khan F, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking[C]//European Conference on Computer Vision, 2016: 472-488.

[12] Luo H B, Xu L Y, Hui B, et al. Status and prospect of target tracking based on deep learning [J]. ***Infrared and Laser Engineering***, 2017, 46(5): 0502002. (in Chinese)

[13] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge [J]. ***International Journal of Computer Vision***, 2015, 115(3): 211-252.

[14] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking[C]// Proceedings of the European Conference on Computer Vision, 2016: 850-865.

[15] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2805-2813.

[16] Dai K, Wang Y, Yan X. Long-term object tracking based on siamese network[C]//IEEE International Conference on Image Processing (ICIP), 2017: 3640-3644.

[17] Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005: 539-546.

[18] Li B, Wu W, Wang Q, et al. Siamrpn++: Evolution of siamese visual tracking with very deep networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 4282-4291.

[19] Zhang Z, Peng H, Fu J, et al. Ocean: Object-aware anchor-free tracking[C]//Proceedings of the European Conference on Computer Vision, 2020, 12366: 771-787.

[20] Yan B, Peng H, Wu K, et al. LightTrack: Finding lightweight neural networks for object tracking via one-shot architecture search[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 15180-15189.

[21] Wang G, Luo C, Sun X, et al. Tracking by instance detection: A meta-learning approach[C]//Conference on Computer Vision and Pattern Recognition, 2020: 6287-6296.

[22] Zou Z, Shi Z, Guo Y, et al. Object detection in 20 years: A survey[DB/OL]. (2019-05-16)[2022-01-13]. https://doi.org/10.48550/arXiv.1905.05055.

[23] Chen Y F, Wu Y, Zhang W. Survey of target tracking algorithm based on siamese network structure [J]. ***Computer Engineering and Applications***, 2020, 56(6): 10-18. (in Chinese)

[24] Kristan M, Lukeˇ A, Drbohlav O, et al. The Eighth Visual Object Tracking VOT2020 Challenge Results[M]. Switzerland: Springer, 2020.

[25] He A, Luo C, Tian X, et al. A twofold siamese network for real-time object tracking[C]//Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, 2018: 4834-4843.

[26] Wang Q, Teng Z, Xing J, et al. Learning attentions: residual attentional siamese network for high performance online visual tracking[C]//Conference on Computer Vision and Pattern Recognition, 2018: 4854-4863.

[27] Dong X, Shen J. Triplet Loss in Siamese Network for Object Tracking[M]. Switzerland: Springer, 2018: 472-488.

[28] Cui Z J, An J S, Cui T S. Siamese networks tracking algorithm integrating channel-interconnection-spatial attention [J]. ***Infrared and Laser Engineering***, 2021, 50(3): 20200148. (in Chinese)

[29] Li B, Yan J, Wu W, et al. High performance visual tracking with siamese region proposal network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8971-8980.

[30] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2017, 39(6): 1137-1149.

[31] Wang Q, Zhang L, Bertinetto L, et al. Fast online object tracking and segmentation: A unifying approach[C]//Conference on Computer Vision and Pattern Recognition, 2019: 1328-1338.

[32] Chen B X, Tsotsos J K. Fast visual object tracking with rotated bounding boxes[DB/OL]. (2019-09-12)[2022-01-13]. https://doi.org/10.48550/arXiv.1907.03892.

[33] Zhou W, Wen L, Zhang L, et al. SiamMan: Siamese motion-aware network for visual tracking[DB/OL]. (2020-01-18)[2022-01-13]. https://doi.org/10.48550/arXiv.1912.05515.

[34] Liao B, Wang C, Wang Y, et al. Pg-net: Pixel to global matching network for visual tracking[C]//European Conference on Computer Vision, 2020: 429-444.

[35] 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, 2018: 101-117.

[36] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[37] Zhang Z, Peng H. Deeper and wider siamese networks for real-time visual tracking[C]//Conference on Computer Vision and Pattern Recognition, 2019: 4586-4595.

[38] Li B, Wu W, Wang Q, et al. Siamrpn++: Evolution of siamese visual tracking with very deep networks[C]//Conference on Computer Vision and Pattern Recognition, 2019: 4282-4291.

[39] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Conference on Computer Vision and Pattern Recognition, 2017: 936-944.

[40] Guo D, Wang J, Cui Y, et al. SiamCAR: siamese fully convolutional classification and regression for visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 6268-6276.

[41] 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: 12549-12556.

[42] Tian Z, Shen C, Chen H, et al. FCOS: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019: 9627-9636.

[43] Chen Z, Zhong B, Li G, et al. Siamese box adaptive network for visual tracking[C]//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020: 6667-6676.

[44] Zhang Z, Liu Y, Li B, et al. Toward accurate pixelwise object tracking via attention retrieval [J]. ***IEEE Transactions on Image Processing***, 2021, 30: 8553-8566.

[45] Cui Y, Jiang C, Wang L, et al. Fully convolutional online tracking[DB/OL]. (2021-09-26)[2022-01-13]. https://doi.org/10.48550/arXiv.2004.07109.

[46] Zhou X, Wang D, Krähenbühl P. Objects as points[DB/OL]. (2019-04-29)[2022-01-13]. https://doi.org/10.48550/arXiv.1904.07850.

[47] Law H, Deng J. Cornernet: Detecting objects as paired keypoints[C]//Proceedings of the European Conference on Computer Vision, 2018: 765-781.

[48] Gao P, Yuan R, Wang F, et al. Siamese attentional keypoint network for high performance visual tracking [J]. ***Knowledge-based Systems***, 2020, 193: 105448.

[49] Peng S, Wang K, Yu Y, et al. Accurate anchor free tracking[DB/OL]. (2020-06-13)[2022-01-13]. https://doi.org/10.48550/arXiv.2006.07560.

[50] Du F, Liu P, Zhao W, et al. Correlation-guided attention for corner detection based visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 6835-6844.

[51] Yan B, Zhang X, Wang D, et al. Alpha-refine: Boosting tracking performance by precise bounding box estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 5289-5298.

[52] Ma Z, Wang L, Zhang H, et al. Rpt: Learning point set

representation for siamese visual tracking[C]//European Conference on Computer Vision, 2020: 653-665.

[53] Yang Z, Liu S, Hu H, et al. Reppoints: Point set representation for object detection[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019: 9657-9666.

[54] Sauer A, Aljalbout E, Haddadin S. Tracking holistic object representations[DB/OL]. (2019-08-06)[2022-01-13]. https://doi.org/10.48550/arXiv.1907.12920.

[55] Yu Y, Xiong Y, Huang W, et al. Deformable siamese attention networks for visual object tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 6727-6736.

[56] Xu T, Feng Z H, Wu X J, et al. AFAT: Adaptive failure-aware tracker for robust visual object tracking[DB/OL]. (2020-05-27)[2022-01-13]. https://doi.org/10.48550/arXiv.2005.13708.

[57] Zhang L, Gonzalez-Garcia A, van de Weijer J, et al. Learning the model update for siamese trackers[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019: 4009-4018.

[58] Zhou J, Wang P, Sun H. Discriminative and robust online learning for siamese visual tracking[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 13017-13024.

[59] Wang G, Luo C, Xiong Z, et al. Spm-tracker: Series-parallel matching for real-time visual object tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 3643-3652.

[60] Sung F, Yang Y, Zhang L, et al. Learning to compare: Relation network for few-shot learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2018: 1199-1208.

[61] Yan B, Zhao H, Wang D, et al. "Skimming-perusal" tracking: A framework for real-time and robust long-term tracking[C]// Proceedings of the IEEE International Conference on Computer Vision, 2019: 2385-2393.

[62] Zhang H W, Li X X, Zhu B, et al. Two-stage object tracking method based on siamese neural network [J]. ***Infrared and Laser Engineering***, 2021, 50(9): 20200491. (in Chinese)

[63] Fan H, Ling H. Siamese cascaded region proposal networks for real-time visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 7952-7961.

[64] Li Q, Qin Z, Zhang W, et al. Siamese keypoint prediction network for visual object tracking[DB/OL]. (2020-06-07)[2022-01-13]. https://doi.org/10.48550/arXiv.2006.04078.

[65] Bhat G, Danelljan M, Van Gool L, et al. Learning discriminative model prediction for tracking[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019: 6181-6190.

[66] Danelljan M, van Gool L, Timofte R. Probabilistic regression for visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 7181-7190.

[67] Choi J, Kwon J, Lee K M. Visual Tracking by Tridentalign and Context Embedding[M]. Switzerland: Springer, 2020: 504-520.

[68] Huang L, Zhao X, Huang K. Globaltrack: A simple and strong baseline for long-term tracking[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 11037-11044.

[69] Voigtlaender P, Luiten J, Torr P H S, et al. Siam R-CNN: Visual tracking by re-detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 6577-6587.

[70] Dave A, Tokmakov P, Schmid C, et al. Learning to track any object[DB/OL]. (2019-10-25)[2022-01-13]. https://doi.org/10.48550/arXiv.1910.11844.

[71] Huang L, Zhao X, Huang K. Bridging the gap between detection and tracking: A unified approach[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019: 3998-4008.

[72] Danelljan M, Bhat G, Khan F S, et al. ATOM: Accurate tracking by overlap maximization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 4655-4664.

[73] Jiang B, Luo R, Mao J, et al. Acquisition of localization confidence for accurate object detection[C]//Proceedings of the European Conference on Computer Vision, 2018.

[74] Finn C, Abbeel P, Levine S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International Conference on Machine Learning, 2017: 1126-1135.

[75] Antoniou A, Edwards H, Storkey A. How to train your maml[C]//International Conference on Learning Representations, 2019.

[76] Li Z, Zhou F, Chen F, et al. Meta-SGD: Learning to learn quickly for few-shot learning[DB/OL].(2017-09-28)[2022-01-13]. https://doi.org/10.48550/arXiv.1707.09835.

[77] Kristan M, Leonardis A, Matas J, et al. The Sixth Visual Object Tracking VOT2018 Challenge Results[M]. Switzerland: Springer, 2018: 3-53.

[78] Huang L, Zhao X, Huang K. Got-10 k: 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.

[79] Fan H, Lin L, Yang F, et al. Lasot: A high-quality benchmark for large-scale single object tracking[C]//Conference on Computer Vision and Pattern Recognition, 2019: 5374-5383.

[80] Han G, Du H, Liu J, et al. Fully conventional anchor-free siamese networks for object tracking [J]. ***IEEE Access***, 2019, 7: 123934-123943.

[81] Danelljan M, Gool L Van, Timofte R. Probabilistic regression for visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020: 7181-7190.

[82] Choi J, Chun D, Kim H, et al. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019: 502-511.

[83] He Y, Zhu C, Wang J, et al. Bounding box regression with uncertainty for accurate object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 2888-2897.

[84] Zhu B, Wang J, Jiang Z, et al. Autoassign: Differentiable label assignment for dense object detection[DB/OL]. (2020-11-25)[2022-01-13]. https://doi.org/10.48550/arXiv.2007.03496.

[85] Li X, Wang W, Wu L, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[C]//Advances in Neural Information Processing Systems, 2020.

[86] Oksuz K, Cam B C, Kalkan S, et al. Imbalance problems in object detection: A review [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2020, 43(10): 3388-3415.

[87] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 5998-6008.

[88] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 x16 words: Transformers for image recognition at scale[C]// International Conference on Learning Representations, 2021.

[89] Chen X, Yan B, Zhu J, et al. Transformer tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 8126-8135.

[90] Yan B, Peng H, Fu J, et al. Learning spatio-temporal transformer for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision, 2021: 10448-10457.

[91] Wang N, Zhou W, Wang J, et al. Transformer meets tracker: Exploiting temporal context for robust visual tracking[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 1571-1580.

[92] Lin L, Fan H, Xu Y, et al. SwinTrack: A simple and strong baseline for transformer tracking[DB/OL]. (2021-12-08)[2022-01-13]. https://doi.org/10.48550/arXiv.2112.00995.