第 51 卷第 4 期 中国科学技术大学学报 Vol. 51, No. 4
2021 年 4 月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Apr. 2021
收稿日期：2021-02-01；修回日期：2021-04-07 doi:10.52396/JUST-2021-0037

深度视觉目标跟踪进展综述

王宁,席茂,周文罡*,李礼,李厚强*
中国科学技术大学多媒体计算与通信教育部-微软联合实验室,安徽合肥,230027
*通讯作者. E-mail：zwg@ustc.edu.cn；lihq@ustc.edu.cn

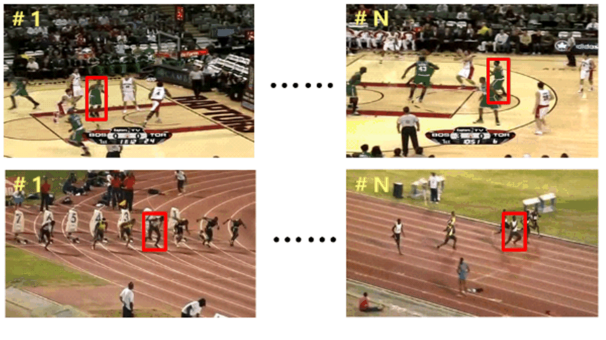
摘要：视频目标跟踪是计算机视觉领域的一个重要研究课题. 近年来,随着深度学习在视觉目标跟踪领域获得了巨大的成功,一系列优秀的深度跟踪算法涌现出来. 本文回顾了近年来深度目标跟踪领域的进展. 首先,我们详细讨论了近十年来跟踪领域数据集的发展趋势,这些数据集不仅全面地评估了算法性能同时为模型训练提供了极大的便利. 其次,我们分类讨论了几大类经典的深度学习跟踪框架,包括深度相关滤波器跟踪、分类式网络跟踪、双路网络跟踪、基于梯度的深度跟踪算法以及基于 Transformer 的跟踪算法. 最后,我们对全文内容进行总结,并指出未来的发展趋势.

关键词：深度目标跟踪;跟踪数据集;相关滤波器;分类式跟踪网络;双路跟踪网络;梯度跟踪网络
中图分类号：P595 **文献标识码**：A

1 引言

视觉目标跟踪是计算机视觉领域的一个基本任务. 目标跟踪旨在基于初始帧中指定的感兴趣目标(一般用矩形框表示),在后续帧中对该目标进行持续的定位,如图 1 所示. 目标跟踪的应用场景非常广泛,包含视频监控、人机交互、机器人、无人驾驶等. 虽然近二十年来,视觉目标跟踪取得了极大的进展,但是一些挑战性因素如目标遮挡、背景杂乱、运动模糊、光照变化等仍是目标跟踪算法面临的主要挑战.

传统的视觉跟踪算法通常采用手工特征来对目标进行表观建模,通过训练鲁棒的辨别式或生成式模型实现目标跟踪,典型的方法包括 MIL[1]、TLD[2]、SCM[3]、STRUCK[4]、KCF[5] 等. 新近比较有挑战性的数据集如 VOT2018[6] 或大规模数据集 TrackingNet[7] 和 LaSOT[8] 上,这些算法的性能远远达不到实际应用的要求.



图中的红色矩形框表示感兴趣的跟踪目标.

图 1 视觉跟踪任务示例

Figure 1. The example of visual tracking

自从 2012 年 AlexNet[9] 在图像分类任务中大放异彩,深度学习受到了广泛关注. 得益于强大的特征提取能力和端到端的训练模式,深度学习技术在计算机视觉、机器学习、自然语言处理等领域都广受关注,并取得了巨大进展. 在过去的五、六年间,基于深度学习的目标跟踪算法获得了巨大突破. 一些经典的深度跟踪算法,如 HCF[10]、MDNet[11]、SiamFC[12]、ECO[13]、SiamRPN[14]、ATOM[15]、DiMP[16] 等不同程度地挖掘了深度学习的潜能并显著提高了跟踪性能. 例如,在经典的 OTB-2015[17] 数据集上,这些深度学习的跟踪算法大幅度超越经典的跟踪器并不断刷新最优性能. 在每年举办的视觉跟踪的挑战赛如 VOT-2018 中,排名前 10 位的算法均不同程度地使用了深度特征. 这些深度学习的跟踪算法采用了各种各样的框架,包含相关滤波器、分类式网络、双路网络等. 在处理跟踪任务的角度,从基于匹配思想的双路网络框架到基于二分类思想的辨别式跟踪器,各种算法框架在性能和效率上各有千秋.

最初的深度跟踪算法主要聚焦于相关滤波器. 通过将传统相关滤波器中的手工特征替换成深度特征,

Citation：王宁,席茂,周文罡,等. 深度视觉目标跟踪进展综述. 中国科学技术大学学报, 2021, 51(4): 335-344.
WANG Ning, XI Mao, ZHOU Wengang, et al. Recent advance in deep visual object tracking. J. Univ. Sci. Tech. China, 2021, 51(4): 335-344.

跟踪性能得到了大幅度提升. 后续研究人员尝试端到端地结合相关滤波器和深度模型,并进一步引出了一系列的基于梯度优化的方案,如 DiMP 算法[16]. 通过将跟踪任务视为模板匹配,基于双路网络跟踪算法(如 SiamFC[12])由于其简洁的框架和高效率而受到了极大关注. 但是该类方法由于忽略了背景信息,因而对相似干扰物的辨别能力较弱,后续工作在双路网络中借鉴相关滤波器来提升模型的辨别能力. 此外,受目标检测领域进展的启发,基于分类式的深度跟踪框架(如 MDNet[11])、双路网络结合区域锚点的多尺度回归[14]等思路同样受到广泛研究. 近期基于 Transformer 的深度跟踪器使用注意力机制进行跟踪模型建模,取得了领先的性能. 我们总结了深度跟踪领域常见的框架及代表性工作,如表 1 所示.

表 1 深度跟踪算法框架概述

Table 1. An overview of deep visual tracking algorithms.

跟踪器类别	主要思想	代表性工作
深度相关滤波器跟踪	将现成的深度特征和相关滤波器算法结合,或将深度网络和相关滤波器端到端联合训练	HCF[10]、C-COT[18]、ECO[13]、CFNet[19]
基于分类式网络的跟踪	将跟踪任务视为前景、背景的二分类任务,在每一帧中分类出最高置信度的候选样本	MDNet[11]、VITAL[20]、RT-MDNet[21]
基于双路网络的跟踪	将跟踪任务视为模板匹配任务,在每一帧中匹配出和初始目标最相似的候选样本	SiamFC[12]、SiamRPN[14]、SiamRPN++[22]
基于梯度优化的跟踪	采用梯度下降(优化)的方法快速求解跟踪中的回归问题来得到具有区分力的跟踪模型	CREST[23]、ATOM[15]、DiMP[16]
基于 Transformer 的跟踪器	借鉴自然语言处理中的 Transformer 结构进行跟踪模型的建模	TrDiMP[24]、TransT[25]、STARK[26]

表 1 按照各种算法出现的先后顺序进行安排. 深度相关滤波器在 2015 年左右提出(如 HCF[10]),并在近年来持续受到关注. 相关滤波器的思想近年来被其他跟踪框架如双路网络和基于梯度的跟踪器所吸纳. 基于分类网络(MDNet)和双路网络(SiamFC)的跟踪算法几乎同期被提出,大致于 2016 年获得了广泛关注. 由于分类网络需要在线的模型微调,导致效率偏低,因而近年来关注度逐渐降低. 双路网络通过汲取相关滤波器的优势(如 CFNet)以及融入区域候选网络(如 SiamRPN)而持续地演变和进化,目前仍是研究的热点. 基于梯度的优化方法在 2019 年左右受到了广泛关注,其代表性工作包括 ATOM 和 DiMP. 该类方法受相关滤波器启发,通过采用快速梯度下降的方法求解具有前景、背景区分能力的滤波器核. 由于利用了背景信息,该类方法相比于双路网络具有更好的干扰物辨别能力. 2021 年,同时期出现了数个基于 Transformer 结构的深度跟踪算法. 该类方法利用注意力机制利用时序信息[24],或对跟踪器建模[25,26],取得了十分突出的性能.

为了总结归纳深度跟踪算法的发展趋势,本文详细梳理了近年来深度跟踪领域的相关工作,并按如下顺序进行阐述:跟踪数据集的发展趋势、结合深度特征的相关滤波器、基于分类网络的跟踪算法、基于双路网络的跟踪算法、基于梯度的深度跟踪算法、基于 Transformer 的深度跟踪等,最后对研究方向进行展望.

2 跟踪数据集发展趋势

数据、算法和算力是人工智能最重要的三个要素. 在计算机视觉任务中,好的数据集往往能够带动相关领域的快速发展. 随着卷积神经网络的快速发展,更多参数量的网络往往需要更多的数据去学习得到一个更好的模型. 因此需要一个良好的标注数据集以促进相关算法的快速发展. 近些年来,视频目标跟踪领域出现了许多不同大小、种类的数据集. 这些数据集引领了目标跟踪算法的进步. 本节将详细介绍目标跟踪领域的常见数据集. 这些数据集的概况和对比如表 2 所示.

OTB: OTB 数据集分为 OTB2013[27]和 OTB2015[17]两个版本. 其中 OTB2013 数据集包含 51 个视频序列,由 Wu 等收集以往目标跟踪领域的常用测试视频构成. 该数据集考虑到很多影响跟踪性能的因素,比如形变、遮挡、光照变化、快速运动、运动模糊等. 同时作者还提出了一系列的评估准则,这些准则与数据集一起为跟踪算法提供了相对统一的测试与评估环境,有利于不同跟踪方法之间的比较,极大地促进了早期目标跟踪任务的发展. OTB2015 是 OTB2013 数据集的扩充,

表 2　目标跟踪任务中常用数据集对比

Table 2. Comparison of benchmark datasets in visual tracking task.

数据集	视频数量	平均帧长	帧率	目标类别	属性	训练/测试集划分	总时长
OTB2013[27]	51	578	30	10	11	无	16.4 分钟
OTB2015[17]	100	590	30	16	11	无	32.8 分钟
TColor-128[28]	128	429	30	27	11	无	30.7 分钟
VOT[6]	60	不定	30	不定	不定	无	不定
NFS[29]	100	3830	240	17	9	无	26.6 分钟
UAV123[30]	123	915	30	9	12	无	62.5 分钟
GOT-10K[31]	10K	150	10	563	6	有	–
TrackingNet[7]	31K	451	30	27	–	有	–
OxUvA[32]	366	–	–	22	–	有	14.3 小时
LaSOT[8]	1400	2,506	30	70	14	有	32.5 小时

通过引入额外的视频,该数据集总共包含 100 个视频. 此外,该数据集还对视频标出了遮挡、形变、快速运动、模糊等 11 个视频属性,便于分析跟踪器应对不同场景的能力.

TColor-128:Liang 等[28]于 2015 年提出此数据集. 针对 OTB 数据集中存在大量的灰度视频,不利于实际场景的评估,TColor-128 数据集收集了 128 个彩色视频,含 27 个类别和 7 个属性. 其中,78 个视频是新收集的,剩余的 50 个视频来自 OTB2015. 其中,新收集的数据集的视频目标跟踪难度高于原有视频.

VOT:VOT 数据集[6]随 2013 年首次举办的单目标视觉跟踪比赛开始发布,每年一期,发展至今. 早期 VOT 数据集主要针对短时目标跟踪,近些年来也会评价算法的实时性以及长时目标跟踪的性能. 该数据集的评价方式与以上的数据集不同,当跟踪器在该数据集上每次跟踪目标失败时,跟踪器都会复位,重新进行初始化. 最终根据失败的次数以及准确度综合成一个统一的指标来评价跟踪器的性能.

NFS:此数据集[29]包含 100 个视频,含 17 个类别和 9 个属性. 不同于常规数据集 30 FPS 的视频采样频率,NFS 视频的帧率达到 240 FPS,较高的帧率往往对跟踪器的跟踪性能有很大的提升. 通过高帧率的摄像设备,很多传统跟踪算法可以与最新的跟踪器相媲美. 和前面数据集一样,该数据集没有划分训练集与测试集.

UAV123:此数据集[30]包含 123 个由无人机拍摄的视频,有 9 个类别 12 个属性,帧率为 30 FPS,不同于以往针对通用单目标视频跟踪数据集,UAV123 数据集针对特定的无人机场景,视频往往是高空俯视角度,特点是背景干净,视角变化较多.

GOT-10K:此数据集[31]含 560 个类别,共有 1 万个视频. 在目标类别、视频数量上均远超以往数据集. 此外该数据集进行训练集和测试集划分,且两者之间没有重叠. 需要说明的是,该数据集的训练和测试视频中的物体类别没有重合,目的在于更加贴近目标跟踪任务的设定,即离线训练阶段跟踪方法没有任何关于待跟踪目标的先验知识,这样可以使跟踪方法更加通用,不依赖于特定物体类别或数据集.

TrackingNet:此数据集[7]包含超过 3 万个视频,共有 27 个目标类别,其视频数量和标注数量比以往的任何跟踪数据集都要大. 同时该数据集也进行训练集和测试集划分,且两者之间没有重叠. 该数据集提供的大规模视频能够有效地缓解当前跟踪领域的训练数据不足问题.

OxUvA:此数据集[32]包含 366 个视频,总时长超过 14 小时. OxUvA 专门针对测试长时跟踪的性能. 长时跟踪,由于目标频繁地被遮挡及超出视野,对跟踪器的鲁棒性有更高的要求. 与此同时,作者还提出了评价长时跟踪算法性能的指标,有助于带动长时跟踪领域的发展.

LaSOT:LaSOT[8]是近年来 Fan 等提出的高标准的数据集,包含 70 个类别 1400 个全部由人工标注的视频,分为训练集与测试集,且两者没有重叠. 该数据集的视频平均长度在 2500 帧左右,算法训练具有挑战性. 同时作者也提出对应的评价标准以并测试了多数当前领先的目标跟踪算法性能.

上述的跟踪数据集各有特色并且近年来提出的新数据集也展现出更大的挑战性. 种类多样的数据集为全面综合评估跟踪算法提供了有力的工具. 此外,近年来的高质量、大规模数据集为深度跟踪算法的训练提供了极大的便利.

3 深度跟踪算法

3.1 深度相关滤波器跟踪

相关滤波器(correlation filter, CF)通过学习一个具有区分力的滤波器来处理待跟踪的图片,其输出结果为一个响应图,表示目标在后续帧中不同位置的置信度. 相关滤波器通过利用循环样本和循环矩阵的性质求解岭回归问题,得到了频域上的高效闭合解,计算效率高. 传统的相关滤波器使用手工特征(如 HOG、ColorName 等)进行学习,较好地兼顾了性能和效率. 由于相关滤波器的学习过程中引入了循环样本,这些样本不可避免地带来了边界效应,因此学者对传统的相关滤波器算法针对如何抑制边界效应开展了大量的研究,典型的工作包括 SRDCF[33]、BACF[34]、ASRCF[35]等. 其余的经典工作包含如何自适应调整学习率(如 SRDCFdecon[36]),如何引入更多的背景信息(如 CACF[37])等.

随着深度学习的日益发展,深度学习与相关滤波器的结合受到了广泛的关注. 在早期的工作中,研究人员探索如何将离线训练好的深度特征(如利用 ImageNet 预训练的 VGG 模型[38])与相关滤波器进行结合. 典型的工作 HCF[10]提出将不同层的深度特征分别训练相关滤波器并进行由粗到精的融合. 高层的语义特征对于目标的抽象表达能力很强,而低层的模型特征擅长刻画目标的纹理、形状等底层信息. 通过将不同尺度特征的滤波响应图进行融合,相关滤波器更好地利用了深度模型. HCF 的框架图如图 2 所示. 在后续的工作中,如何更加充分利用深度特征得到了进一步研究,如 HDT[39]算法研究了如何自适应地改变各尺度特征下滤波器的权重,MCCT[40]跟踪器将滤波器的各种特征进行组合并从中自适应切换等.

深度相关滤波器跟踪代表性的工作 C-COT[18]和 ECO[13]取得了同期十分优异的性能. C-COT 算法重点研究了不同尺度深度特征的分辨率不同而导致的响应图融合问题. 不同于传统方法采用线性插值来调整特征或响应图的尺度,C-COT 采用了连续性插值并和滤波器进行联合优化,取得了良好的效果. ECO 在 C-COT 的基础上,进行了自适应的相关滤波器选取、目标样本的聚类和稀疏的目标更新,获得了效率和存储上的进一步优化并轻微提升了性能. 在 UPDT[41]中,C-COT 和 ECO 系列工作进一步研究了深度特征在相关滤波器中的潜能. 以往的深度滤波器跟踪通常使用较浅的神经网络进行特征提取,使用更深的网络如 ResNet[42]并没有进一步的性能增益. UPDT 算法详细分析了该问题,并提出了适合深度相关滤波器的数据增广、滤波器带宽、融合权重优化等细节,使得深度相关滤波器在使用更深的神经网络后可以得到持续的性能提升. 虽然上述深度相关跟踪在一系列数据集中取得了优异的性能,但是提取特征耗费了大量时间,使其即便使用先进的 GPU 也无法达到实时的速度.

随着技术进一步发展,研究人员发现离线训练的深度特征可能并不是最适合相关滤波器的. 得益于相关滤波器的闭合解,研究人员尝试将滤波器和深度特征提取网络进行联合训练. 这样训练得到的网络不仅能够获得适合滤波器的特征,而且网络的深度也比较浅. 经典的工作包括 CFNet[19]和 DCFNet[43]. CFNet 的流程图如图 3 所示,该方法将相关滤波器嵌入在双路网络中进行端到端的学习,在获得相关滤波器辨别能力的情况下,同时保证了极高的运行效率. 同时期提出的 DCFNet 采用了相似的策略将相关滤波器和主干网络进行联合学习,并在线更新滤波器模板取得了更优的性能. 这类端到端的训练模式效率良好并保持了相关滤波器的频域闭合解. 但是相关滤波器中抑制边界效应等一系列工作(如 SRDCF、BACF)和其他优化(如 C-COT、ECO)等破坏了经典相关滤波器(如 KCF)的闭合解,需要使用交叉迭代的方法进行优化,为端到端训练带来了新挑战. RTINet[44]对此进行了探索并成功将 BACF 和深度学习网络进行联合训练. 在最新的一系列工作中,研究者采用了梯度下降的方法,成功地汲取了上述工作的优势并采用端到端的网络训练优化整个框架.

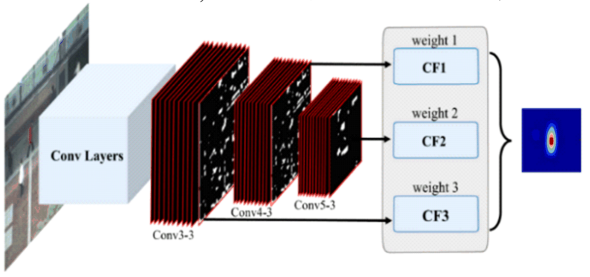


图 2 HCF 算法[10]示意图

Figure 2. Framework illustration of HCF algorithm[10].

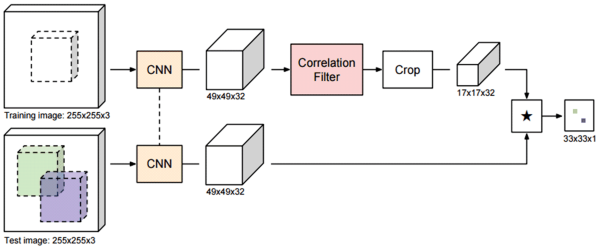


图 3 CFNet[19]框架示意图

Figure 3. Framework illustration of CFNet algorithm[19].

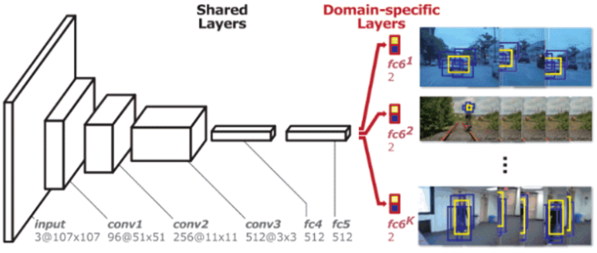


图 4　MDNet 算法[11]框架

Figure 4. Framework illustration of MDNet algorithm[11].

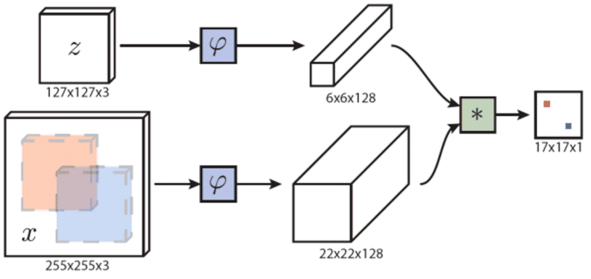


图 5　SiamFC 算法[12]框架

Figure 5. Framework illustration of SiamFC algorithm[12].

3.2　基于分类的深度跟踪器

基于分类的深度跟踪方法受经典的目标检测框架 R-CNN[45]的启发,将目标跟踪任务视为二分类(目标和背景)任务. 该深度跟踪网络包含预训练的卷积层以提取通用的、鲁棒的深度特征,然后利用第一帧的大量正、负样本进行全连接层的训练,使得网络能够适应当前场景下的目标区分. 后续通过适当的网络更新以适应目标的变化,但同时也使得效率降低. MDNet 方法[11]最早将分类式网络引入跟踪领域,并针对跟踪场景中目标和背景在不同视频(目标域)下可能引起歧义(即该视频中的目标可能成为其他视频中的背景物体),引入了多目标域下的训练模式以学习鲁棒的通用目标特征. 在 MDNet 的训练过程中,网络的共享层(包含前三层卷积层和两个全连接层)由训练集中所有视频共同训练. 对每个视频,分别训练独立的分类层(最后一个全连接层)用于区分当前视频域中的目标和干扰物. 经过离线训练阶段,在跟踪时,利用第一帧的标注信息快速微调一个新的全连接层用于辨别当前视频的前、背景. MDNet 的训练图片如图 4 所示.

后续的一系列工作围绕该分类式模型展开. BranchOut[46]算法在 MDNet 的基础上引入了模型集成的思想,提出了多个并行的全连接层,并挑选对当前场景最具区分力的子分类器进行跟踪. VITAL[20]方法在 MDNet 分类模型中引入了生成对抗式网络,通过生成具有遮挡属性的掩膜来干扰分类器,使得分类器学习到的特征更加鲁棒以抑制遮挡掩膜的影响. 分类式跟踪方法的主要弊端是速度慢,在 GPU 中仅能达到 1 FPS,其主要原因在于大量的候选样本都需要进行特征提取. 近期的实时 MDNet 算法(RT-MDNet[21])在分类式网络借鉴 Fast RCNN [47]的思想,对搜索区域进行共享特征的提取再裁剪出样本特征,通过引入 ROI Align 并修改了其他一系列网络细节使得精度下降很少的情况下跟踪速度得到极大提升.

3.3　双路网络跟踪算法

近些年由于深度神经网络快速发展,其强大的特征提取能力使得其对计算机视觉不同领域的表现都有很大的提升. 在视频目标跟踪中,Bertinett 等于 2016 年首次提出双路网络框架(SiamFC[12]),此方法利用卷积网络提取目标模板和搜索区域的特征,然后再进性相关操作生成响应图,其中响应图上的峰值点就是目标所在的位置,如图 5 所示. 不同于传统的相关滤波器算法,双路网络不需要在线参数更新,而且目标模板的特征只需要提取一次即可,因此双路网络的跟踪速度通常很快. 经典的 SiamFC 算法可以在 GPU 下达到 86 FPS 的跟踪速度.

在此之后,考虑到 SiamFC 对目标尺度的回归仍然采用传统缩放形式不能准确地获得目标的尺度信息,Li 等提出 SiamRPN[14],此方法将目标检测中的 RPN 结构引入到 SiamFC 中,通过利用共享的参数提取特征,然后分别通过分类支路获得目标的位置以及回归支路获得目标尺度的精确估计. 如图 6 所示,左边表示共享参数的 CNN 网络;中间部分表示 RPN 模块,这里包含了两支,一支表示分类网络,另一支表示回归网络. 相比于 SiamFC 中采用的传统的图像金字塔的方式来估计目标的大小,SiamRPN 的推理速度更快,可以达到 160 FPS. 此后 Li 等进一步对 SiamRPN 进行拓展提出 DaSiamRPN[48],通过有效地利用负样本对提高了跟踪器的抗干扰能力. 尽管如此,以上方法都是使用 AlexNet 作为特征提取网络,没有用到当前的更深、更强大的特征提取网络. 因此为了能使双路算法充分利用到现有得深层神经网络,Li 等对 ResNet 网络进行更改提出了 SiamRPN++算法[22]. SiamRPN++一方面采取随机平移目标在搜索区域内的位置以解决 CNN 网络边界填充对网络平移不变性的破坏,另一方面也采用了高、中和低层的特征融合的方式获得更好的目标特征表达. 最终通过结合多个数据集进行训练,SiamRPN++在多个目标跟踪的数据集上获得了很好的效果.

研究人员针对 SiamFC 框架,提出了一系列改进,包括集成学习引入互补的双路网络分支[49],引入注意力机制[50,51],无监督学习[52,53],图卷积神经网络[54],采用强化学习来调整模型参数[55]等. Du 等[56]提出了基

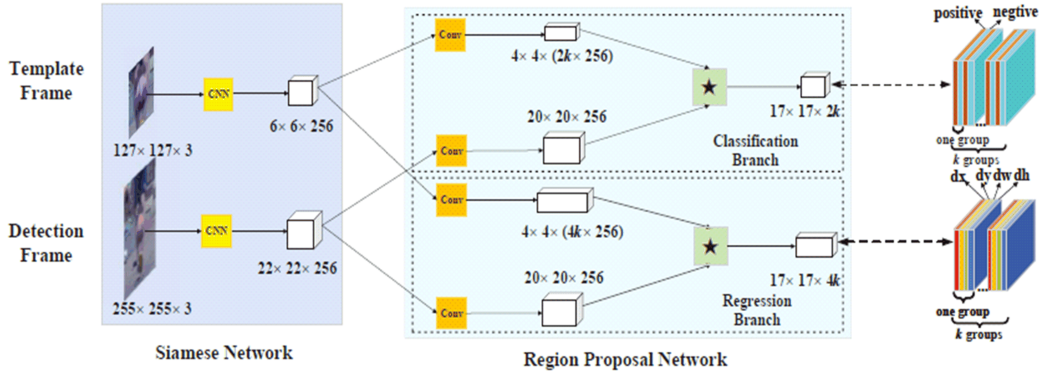


图 6 SiamRPN 算法[14] 框架

Figure 6. Framework illustration of SiamRPN algorithm[14].

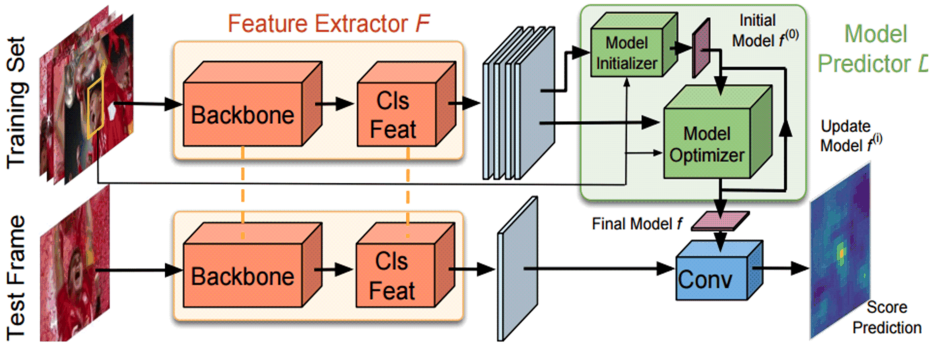


图 7 DiMP 跟踪算法[16] 框架

Figure 7. Framework illustration of DiMP algorithm[16].

于角点预测的双路网络. Siam R-CNN[57] 提出基于重检测的长时双路跟踪框架,取得了十分出色的结果. 针对 SiamRPN 框架,研究人员进一步研究了无锚点的双路网络[58,59]. 基于固定锚点的方法很难约束与目标重合度较小(如 IOU<0.3)的锚点,而无锚点的双路网络能够在较大空间内回归目标区域,因而表现更加出色.

尽管以上的双路网络方法在视频目标跟踪中取得了很大的成功,但是仍然存在缺陷,即其始终使用视频序列的初始帧模板与后续每一帧进行匹配,缺少在线更新过程. 这导致我们的跟踪器不能自动调整参数以适应目标的变化,于是提出了基于双路网络的更新方法. MemTrack[60] 算法通过 LSTM 结构利用历史帧的模板信息去预测当前帧的模板信息. Meta-Tracker[61] 算法额外利用一个元学习网络为双路网络提供新目标的外观信息. 此外,Re2EMA[62] 算法学习一个变换矩阵将过去的模型自适应地变换到新模型. UpdateNet[63] 算法训练一个独立的卷积网络,利用历史模板在下一帧预测一个最优的模板特征. GradNet[64] 算法利用梯度信息去更新模板,这一定程度上可以抑制模板中的背景信息. 以上算法从不同角度利用历史帧信息更新模板,这缓解了固定模板带来的适应目标变化不足的缺陷.

3.4 基于梯度优化的深度跟踪方法

经典的相关滤波器通过闭合解的形式最小化岭回归损失,但是该闭合解主要依赖循环样本. 在使用真实样本的情况下,岭回归损失虽然失去了快速的闭合解,但仍可以通过梯度下降的方法进行求解. 在 CREST[23] 算法中,深度学习中常见的随机梯度下降算法(SGD)被用于优化该岭回归损失,来学习一个类似相关滤波器具备前景、背景区分能力的卷积核. 该卷积核和搜索图片的特征图进行卷积,可以生成响应图用于目标跟踪. 在 CREST 跟踪器中,作者进一步引入残差项来弥补目标外观的快速变化,取得了进一步的性能提升. 在 DLST 算法[65] 中,研究人员发现训练区域中的正负样本不均衡问题导致梯度优化不够精确且效率不高. 该方法基于岭回归中的二范数损失形式,引入了收缩式损失权重,极大地抑制了冗余的、容易分类的负样本的权重,使得学习到的滤波器更加具有区分力且学习的速度更快.

上述的随机梯度下降方案能够学习到鲁棒的滤波器用于目标跟踪,但是随机梯度下降的优化过程通常需要数十次甚至上百次迭代才能较好地收敛,因而一定程度抑制了跟踪器的效率. 在最近的工作中,研究人员转向更加快速的梯度下降方法. ATOM[15] 算法中,

作者采用共轭梯度策略结合深度学习框架进行快速优化. 共轭梯度下降方法估计了每次迭代的学习率和步长,配合深度学习框架的梯度回传求解能力,可以快速地进行优化. 该研究团队在工作 DiMP[16] 中进一步将该思想用在了端到端的学习中,通过神经网络学习到所有的参数,取得了目标领先的性能. DiMP 算法的流程图如图 7 所示. 该算法最核心的设计在于模型预测部分,其包含一个模型初始化和一个模型优化器. 模型初始化主要通过剪裁目标区域特征并通过池化来得到一个良好的初始化滤波器模型,该过程和经典的双路网络类似. 模型优化器部分旨在约束初始化滤波器进行岭回归损失,使得滤波器模型具有背景区分的能力. 由于使用神经网络端到端地估计梯度下降的优化步长,使得模型可以在少数的几次迭代中快速收敛,保持了岭回归损失的区分能力同时保证了跟踪效率. 该算法在数个跟踪数据集上都刷新了当前最好的性能.

D3S 算法[66] 将 ATOM 中基于梯度优化的相关滤波器模型融合在视频目标分割网络中,增强了对干扰物的辨别能力. 在 DiMP 算法的基础上,PrDiMP[67] 采用了基于概率回归的方法对跟踪器和尺度预测模块进行建模,减缓不精确或有歧义的噪声标签对网络的干扰,有效提升了模型精度. KYS 算法[68] 在 DiMP 的基础上考虑了周围环境中背景物体在连续帧间的关联,有效提升了跟踪器对背景区域的感知能力和对干扰物的区分能力.

3.5　基于 Transformer 的深度跟踪方法

Transformer 结构[69] 最初在自然语言处理任务中提出. Transformer 的核心模块是注意力机制,可以将全局信息聚合到需要的位置. 由于该结构可以充分利用 GPU 等硬件设备进行并行计算,因而相比于传统的 RNN 方法更加适合处理长句子. 近年来,注意力机制[70] 在视觉任务中展示了优异的效果. 基于 Transformer 的深度跟踪器在近期取得了优异的性能,获得了广泛的关注.

Wang 等[24] 利用 Transformer 结构探索视频中的时序信息. 传统的跟踪器通常将跟踪任务视为逐帧的目标检测,忽视了丰富的时序关系. 文献[24]提出使用 Transformer 结构将独立的视频帧桥接起来,一方面利用 Transformer 编码器对不同时刻的模板进行相互融合,另一方面利用 Transformer 解码器将不同时刻的视频帧桥接起来,以传递丰富的时序信息,如目标在不同时刻的特征表达以及模板的注意力掩膜等. TransT 算法[25] 借鉴 Transformer 结构改进传统双路网络中的特征融合操作. 传统的双路网络跟踪器利用相关操作将模板特征融合到搜索图片特征中对搜索区域进行分类和回归. 然而这种相关操作没有对模板和搜索区域特征的关系进行充分建模. TransT 算法利用 Transformer 中的注意力机制将模板信息融合到搜索区域中,以便更好地进行目标定位和尺度回归. STARK 算法[26] 借鉴目标检测算法 DETR[71],利用 Transformer 结构进行局部区域中的目标检测实现跟踪任务. 该算法不使用诸如高斯窗等后处理机制,仅使用网络来预测目标的角点进行跟踪. STARK 算法进一步引入时空信息,通过在跟踪过程中将新采集的样本加入 Transformer,跟踪器可以一定程度适应目标的外观变化,取得进一步的性能提升. 值得注意的是,最近提出的 Transformer 跟踪算法[24-26] 在大规模数据集上不断刷新了当前的最优性能,显示了 Transformer 跟踪的巨大潜力.

4　展望

视觉跟踪领域的算法层出不穷,并且各类算法框架都处于不断的发展与完善中. 随着研究的不断深入,深度学习的潜能也进一步被激发. 然而,现有的框架仍存在有待提升的空间.

最近的双路网络方法(如 SiamRPN++)和梯度优化的方法(如 DiMP)为了追求高性能,均采用了很深的 CNN 模型如 ResNet50. 最新的深度模型动辄具有几十甚至上百兆的模型大小,使得这些算法需要极大的存储空间,限制了实际应用. 如何设计适合他们的轻量级模型,例如使用神经网络搜索的方式来获得更优的模型结构,以兼顾低内存消耗和高精度具有重要的研究价值.

此外,随着 CNN 网络越来约深,模型越来约复杂,几大类深度跟踪框架无论双路网络(SiamRPN++)、分类网络(如 RT-MDNet)还是梯度优化的方法(DiMP),都仅能保持 GPU 设备下勉强实时的速度. 视觉跟踪作为很多应用系统中的底层辅助任务,对效率有很高的要求. 期待未来更多的工作能够聚焦于跟踪算法的速度提升. 设计硬件友好的模型运算结构,用于特定场景的高效率视觉跟踪同样具有巨大的应用前景.

基于 Transformer 的视觉跟踪算法刚刚起步,未来有巨大的挖掘空间. 首先,目前的 Transformer 跟踪算法[25,26] 仍没有充分利用背景信息,如何将背景信息引入到 Transformer 结构中提升它的前景、背景区分能力有待探索. 其次,设计可更新的 Transformer 结构,用于适应目标的外观变化亟须探索. 例如,STARK 算法[26] 仅仅粗暴地加入一帧历史样本. 如何更好地利用时序信息以更新 Transformer 模型将有助于达到更优性能. 最后,Transformer 的注意力机制擅长进行多模态信息间的转换以及融合,该框架的兴起为多模态的视觉跟

踪提供了良好的研究契机,如带有红外信息(RGBT 视频中)和深度信息(RGBD 视频中)的视觉跟踪.

目前几大类主流的跟踪算法框架均聚焦于短时视觉跟踪场景. 在实际的复杂场景中,如长时目标跟踪的情况下,如何处理目标的剧烈外观变化、严重目标遮挡、频繁超出视野等问题仍有巨大的提升空间. 应对目标外观的长时间的、剧烈的变化,如何有效对目标外观进行建模以及如何更新跟踪模型充满挑战. 将现有的短时跟踪算法扩展到复杂的实际场景中需要不断探索. Transformer 跟踪器在各种短时跟踪数据集中展示了出色的性能. 而 Transformer 结构本身特别擅长对全局信息进行聚合和分析,能否利用该结构对长时的时序信息进行建模,以判断跟踪失败、超出视野以及目标重检测,需要进一步探索.

5 结论

尽管近十年来视觉目标跟踪技术取得了巨大的进展,但在复杂的实际场景中,计算机跟踪系统和人类的视觉系统仍有巨大差距. 虽然深度学习算法取得了令人瞩目的成绩,但与此同时带来的跟踪效率限制和模型存储消耗等问题仍需进一步完善. 真正意义上的通用、鲁棒、准确且高效率的视觉跟踪研究仍然任重道远. 我们也目睹了近年来的视觉跟踪领域的快速迭代和不断突破,相信在众多研究者的共同努力下,未来的视觉目标跟踪技术会朝着实用、高效、可靠、通用的跟踪技术方向迈进.

致谢

本文工作得到中国科学技术大学青年创新重点基金(YD3490002001)资助.

利益冲突

作者声明本文没有利益冲突.

作者信息

王宁,中国科学技术大学信息学院博士研究生. 研究方向:视觉目标跟踪.

席茂,中国科学技术大学电子工程与信息科学系硕士生.

周文罡,中国科学技术大学信息学院教授,国家“优秀青年基金”获得者. 研究方向:多媒体信息检索和计算机视觉.

李礼,中国科学技术大学信息学院教授,研究方向:图像处理与计算机视觉等.

李厚强,中国科学技术大学电子工程与信息科学系教授,多媒体计算与通信教育部-微软重点实验室主任,国家“杰出青年基金”获得者,“长江学者奖励计划”获得者. 研究方向:图像处理与计算机视觉、强化学习与机器博弈、多媒体信息检索、视频编码与通信等.

参考文献

[1] Babenko B, Yang M H, Belongie S. Robust object tracking with online multiple instance learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(8):1619-1632.

[2] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. IEEE Transactions on Software Engineering, 2011, 34(7):1409-1422.

[3] Zhong W, Lu H, Yang M H. Robust object tracking via sparsity-based collaborative model. Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, 2012.

[4] Hare S, Saffari A, Torr P H S. Struck:Structured output tracking with kernels. Proceedings of the International Conference on Computer Vision, 2011.

[5] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3):583-596.

[6] Kristan M, Leonardis A, Matas J, et al. The sixth visual object tracking VOT2018 challenge results. Proceedings of the European Conference on Computer Vision Workshops, 2018.

[7] Mueller M, Bibi A, Giancola S, et al. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. Proceedings of the European Conference on Computer Vision, 2018.

[8] Fan H, Ling H, Lin L, et al. Lasot: A high-quality benchmark for large-scale single object tracking. Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, 2019.

[9] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks. Proceedings of the Conference on Advances in Neural Information Processing Systems, 2012.

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

[11] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking. Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, 2016.

[12] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking. Proceedings of the European Conference on Computer Vision Workshops, 2016.

[13] Danelljan M, Bhat G, Khan F S, et al. ECO: Efficient convolution operators for tracking. Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, 2017.

[14] Li B, Yan J, Wu W, et al. High performance visual tracking with siamese region proposal network. Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, 2018.

[15] Danelljan M, Bhat G, Khan F S, et al. ATOM: Accurate tracking by overlap maximization. Proceedings ofthe IEEE Conference of Computer Vision and Pattern Recognition, 2019.

[16] Bhat G, Danelljan M, Gool L V, et al. Learning discriminative model prediction for tracking. Proceedings of the International Conference on Computer Vision, 2019.
[17] Wu Y, Lim J, Yang M H. Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9):1834-1848.
[18] Danelljan M, Robinson A, Khan F S, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking. Proceedings of the European Conference on Computer Vision, 2016.
[19] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[20] Song Y, Ma C, Wu X, et al. VITAL: Visual tracking via adversarial learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[21] Jung I, Son J, Baek M, et al. Real-time MDNet. Proceedings of the European Conference on Computer Vision, 2018.
[22] Li B, Wu W, Wang Q, et al. SiamRPN++: Evolution of Siamese visual tracking with very deep networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[23] Song Y, Ma C, Gong L, et al. CREST: Convolutional residual learning for visual tracking. Proceedings of the IEEE International Conference on Computer Vision, 2017.
[24] Wang N, Zhou W, Wang J, et al. Transformer meets tracker: Exploiting temporal context for robust visual tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.
[25] Chen X, Yan B, Zhu J, et al. Transformer Tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021.
[26] Yan B, Peng H, Fu J, et al. Learning spatio-temporal transformer for visual tracking. 2013,arXiv: 17154, 2021.
[27] Wu Y, Lim J, Yang M-H. Online object tracking: A benchmark. Proceedings of the IEEE Conference of Computer Vision and Pattern Recognition, 2013.
[28] Liang P, Blasch E, Ling H. Encoding color information for visual tracking: Algorithms and benchmark. IEEE Transactions on Image Processing, 2015, 24(12):5630-5644.
[29] Kiani Galoogahi H, Fagg A, Huang C, et al. Need for speed: A benchmark for higher frame rate object tracking. Proceedings of the IEEE International Conference on Computer Vision, 2017.
[30] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking. Proceedings of the European Conference on Computer Vision, 2016.
[31] Huang L, Zhao X, HuangK. Got-10k: a large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
[32] Valmadre J, Bertinetto L, Henriques J F, et al. Long-term tracking in the wild: A benchmark. Proceedings of the European Conference on Computer Vision, 2018.
[33] Danelljan M, Hager G, Shahbaz Khan F, et al. Learning spatially regularized correlation filters for visual tracking. Proceedings of the International Conference on Computer Vision, 2015.
[34] Kiani Galoogahi H, Fagg A, Lucey S. Learning background-aware correlation filters for visual tracking. Proceedings of the International Conference on Computer Vision,2017.
[35] Dai K, Wang D, Lu H, et al. Visual tracking via adaptive spatially-regularized correlation filters. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 4670-4679.
[36] Danelljan M, Hager G, Shahbaz Khan F, et al. Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[37] Mueller M, Smith N, Ghanem B. Context-aware correlation filter tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[38] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv: 1409.1556.
[39] Qi Y, Zhang S, Qin L, et al. Hedged deep tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[40] Wang N, Zhou W, Tian Q, et al. Multi-cue correlation filters for robust visual tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[41] Bhat G, Johnander J, Danelljan M, et al. Unveiling the power of deep tracking. Proceedings of the European Conference on Computer Vision, 2018.
[42] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[43] Wang Q, Gao J, Xing J, et al. DCFNet: Discriminant correlation filters network for visual tracking. 2017, arXiv: 1704.04057.
[44] Yao Y, Wu X, Zhang L, et al. Joint representation and truncated inference learning for correlation filter based tracking. Proceedings of the European Conference on Computer Vision, 2018.
[45] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[46] Han B, Sim J, Adam H. BranchOut: Regularization for online ensemble tracking with convolutional neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[47] Girshick R. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 2015.
[48] Zhu Z, Wang Q, Li B, et al. Distractor-aware Siamese networks for visual object tracking. Proceedings of the European Conference on Computer Vision, 2018.
[49] He A, Luo C, Tian X, et al. A twofold Siamese network for real-time object tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[50] Wang Q, Teng Z, Xing J, et al. Learning attentions: Residual attentional Siamese network for high performance online visual tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[51] Yu Y, Xiong Y, Huang W, et al. Deformable Siamese attention networks for visual object tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[52] Wang N, Song Y, Ma C, et al. Unsupervised deep tracking. Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.
[53] Wang N, Zhou W, Song Y, et al. Unsupervised deep representation learning for real-time tracking. International Journal of Computer Vision, 2021, 129(2): 400-418.
[54] Gao J, Zhang T, Xu C. Graph convolutional tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019.
[55] Dong X, Shen J, Wang W, et al. Hyperparameter optimization for tracking with continuous deep *Q*-learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[56] Du F, Liu P, Zhao W, et al. Correlation-guided attention for corner detection based visual tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[57] Voigtlaender P, Luiten J, Torr P H S, et al. Siam *R*-CNN: Visual tracking by re-detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[58] Guo D, Wang J, Cui Y, et al. SiamCar: Siamese fully convolutional classification and regression for visual tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[59] Zhang Z, Peng H, Fu J, et al. Ocean: Object-aware anchor-free tracking. Proceedings of the European Conference on Computer Vision, 2020.
[60] Yang T, Chan A B. Learning dynamic memory networks for object tracking. Proceedings of the European Conference on Computer Vision, 2018.
[61] Park E, Berg A C. Meta-tracker: Fast and robust online adaptation for visual object trackers. Proceedings of the European Conference on Computer Vision, 2018.
[62] Huang J, Zhou W. Re2EMA: Regularized and reinitialized exponential moving average for target model update in object tracking. Proceedings of the AAAI Conference on Artificial Intelligence, 2019.
[63] Zhang L, Gonzalez-Garcia A, Weijer J, et al. Learning the model update for Siamese trackers. Proceedings of the IEEE International Conference on Computer Vision, 2019.
[64] Li P, Chen B, Ouyang W, et al. GradNet: Gradient-guided network for visual object tracking. Proceedings of the IEEE International Conference on Computer Vision, 2019.
[65] Lu X, Ma C, Ni B, et al. Deep regression tracking with shrinkage loss. Proceedings of the European Conference on Computer Vision, 2018.
[66] Lukezic A, Matas J, Kristan M. D3S-A discriminative single shot segmentation tracker. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[67] Danelljan M, Gool L V, Timofte R. Probabilistic regression for visual tracking. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.
[68] Bhat G, Danelljan M, Van Gool L, et al. Know your surroundings: Exploiting scene information for object tracking. Proceedings of the European Conference on Computer Vision, 2020.
[69] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. 2017, arXiv:1706.03762.
[70] Wang X, Girshick R, Gupta A, et al. Non-local neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[71] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. Proceedings of the European Conference on Computer Vision, 2020.

Recent advance in deep visual object tracking

WANG Ning, XI Mao, ZHOU Wengang*, LI Li, LI Houqiang*
MOE-Microsoft Key Laboratory of Multimedia Computing and Communication, University of Science and Technology of China, Hefei 230027, China
* Corresponding author. E-mail: zwg@ustc.edu.cn; lihq@ustc.edu.cn

Abstract: Visual object tracking is an important branch in computer visions. In recent years, with the remarkable success of deep learning techniques, a series of deep tracking algorithms have emerged with impressive performances. In this paper, we review the recent development of deep learning based trackers. First, we revisit the development of tracking benchmarks in the last decade. These tracking datasets not only comprehensively help evaluate the tracking algorithms but also largely support the model training of deep trackers. Next, we discuss several representative tracking frameworks including deep correlation filter tracking, classification-based tracking networks, Siamese tracking networks, gradient-based tracking networks and Transformer based deep trackers. Finally, we conclude the paper and discuss the potential future research directions of the visual tracking.

Keywords: deep visual tracking; benchmark datasets; correlation filter; classification-based tracking networks; Siamese tracking networks; gradient-based tracking networks