计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(5):40−51 [doi: 10.15888/j.cnki.csa.008476]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于深度学习的单目标跟踪算法综述①

王红涛, 邓淼磊, 赵文君, 张德贤

(河南工业大学 信息科学与工程学院, 郑州 450001)
通信作者: 邓淼磊, E-mail: dmlei2003@163.com

摘　要: 单目标跟踪是计算机视觉领域中的研究热点. 传统算法如相关滤波的跟踪速度较快, 但由于提取到的颜色、灰度等手工特征较为粗糙, 跟踪精度往往不高. 近年来随着深度学习理论的发展, 使用深度特征的跟踪方法能够在跟踪的精度和速度方面达到很好的平衡. 本文首先介绍单目标跟踪的相关背景, 接着从相关滤波单目标跟踪、深度学习单目标跟踪两个阶段对单目标跟踪领域发展过程中涌现出的多个算法进行梳理, 并详细介绍目前主流的孪生网络算法. 最后通过大型数据集对近年来优秀算法进行对比分析, 针对其缺点与不足, 对该领域未来的发展前景做出展望.

关键词: 计算机视觉; 单目标跟踪; 相关滤波; 深度学习; 孪生网络; 注意力机制

引用格式: 王红涛,邓淼磊,赵文君,张德贤.基于深度学习的单目标跟踪算法综述.计算机系统应用,2022,31(5):40–51. http://www.c-s-a.org.cn/1003-3254/8476.html

Survey on Single Object Tracking Algorithms Based on Deep Learning

WANG Hong-Tao, DENG Miao-Lei, ZHAO Wen-Jun, ZHANG De-Xian

(College of Information Science and Engineering, Henan University of Technology, Zhengzhou 450001, China)

Abstract: Single object tracking is a research focus in the field of computer vision. Traditional algorithms including correlation filtering have fast tracking speed but generally low tracking accuracy due to the roughness of extracted manual features such as color and gray levels. With the development of deep learning theory in recent years, tracking methods using deep features can achieve a good balance between tracking accuracy and speed. This study first introduces the relevant background of single object tracking and then sorts out multiple algorithms that have emerged in the development of single object tracking from the two stages of single object tracking based on correlation filters and deep learning. The current mainstream Siamese network algorithms are also introduced in detail. Finally, a large data set is used to compare and analyze the excellent algorithms that have emerged in recent years. In view of the shortcomings and deficiencies of these algorithms, the development prospects of this field are provided in this study.

Key words: computer vision; single object tracking; correlation filter; deep learning; Siamese network (SiameseNet); attention mechanism

单目标跟踪是计算机视觉领域基础且具有广泛实用性的任务之一, 其跟踪方法就是在视频第一帧中获取目标区域的特征信息, 以此为依据在后续帧中对目标状态进行估计并进行准确定位.目前单目标跟踪技术在许多领域均有应用, 例如在智能视频监控领域[1], 可以根据监控目标的移动轨迹自动跟踪, 并自动适应监控场景中的其他变化, 而不需要人为的干预; 在自动驾驶领域[2,3], 通过采用视觉目标跟踪技术, 对车辆周围的

① 收稿时间: 2021-08-02; 修改时间: 2021-08-31; 采用时间: 2021-09-09; csa 在线出版时间: 2022-02-25

汽车、行人、道路进行区分, 进而合理避障, 安全行驶; 在智能人机交互领域[4,5], 通过对人体的动作捕捉或眼球跟踪, 以此产生对应的交互操作. 但由于现实场景比较复杂, 在应对遮挡、运动模糊、背景干扰等问题时, 跟踪效果并不是十分理想.

本文旨在通过对单目标跟踪领域算法的分类梳理与分析, 为研究人员的进一步研究提供一份较高质量的研究综述, 主要结构如下:

文中第 1 节主要介绍单目标跟踪相关背景, 包括跟踪框架和跟踪挑战以及深度学习中的主流网络; 第 2 节对相关滤波类算法的发展进行简单回顾, 并介绍其基于深度特征的改进; 第 3 节对深度学习类单目标跟踪算法进行系统性的梳理, 并通过数据分析总结其优缺点; 第 4 节对本文内容进行总结并针对算法的缺点与不足对未来的发展趋势做出展望.

1 相关背景

1.1 跟踪框架

单目标跟踪的跟踪框架一般可以分为 5 部分, 分别是运动模型、特征提取、观测模型、模型更新和集成处理[6], 其中观测模型是目标能否成功跟踪的关键, 一般可以分为生成式模型和判别式模型.

生成式方法首先通过特征学习得到目标的外观模型, 接着在后续帧中进行模板匹配, 寻找最匹配区域, 以此作为目标位置. 比较著名的生成式方法[7] 有卡尔曼滤波[8]、粒子滤波[9]、均值漂移[10] 等. 但生成式方法没有考虑帧中的背景信息, 所以当面对光照变化、运动模糊等挑战时跟踪准确率往往很低.而判别式方法将跟踪问题转换为一个二分类问题, 通过提取到的目标区域和背景区域的特征, 训练一个分类器在后续帧中对目标与背景进行区分. 由于判别类方法很好的利用了背景中的特征信息, 因此跟踪往往更为准确.

1.2 跟踪挑战

尽管国内外学者已经对视觉目标跟踪技术研究多年, 但在实际应用场景中, 想要对目标实现实时、准确、稳定的跟踪仍是一个很大的难题, 主要面临的挑战[11] 如下:

1) 遮挡 (occlusion). 在现实场景中, 正在运动中的目标很容易发生相互遮挡, 进而丢失部分或全部信息.部分信息丢失时, 可以通过对目标分块或者及时更新模板的方法进行解决, 而全部信息丢失目前并没有好的办法完全解决.

2) 图像模糊 (image blur). 运动模糊、光照变化、图像分辨率较低等情况, 都会导致目标出现模糊效果, 使目标外观特征信息受损, 进而影响到后续的特征提取与匹配.

3) 形变 (deformation). 目标在运动过程中很容易发生形态变化, 如果形变过大则会导致跟踪发生漂移. 应对这个挑战的关键在于能够及时的更新目标的表观模型, 使其很快适应表观的变化.

4) 尺度变化 (scale variation). 一般由物体在运动过程中距离镜头的远近产生了较大的变化或者非刚性的物体在运动过程中发生旋转导致较大的尺度变化, 目前的尺度自适应算法已经较好的解决了这类问题.

5) 背景干扰 (background clutters). 背景干扰主要是图像背景与目标特征相似, 这会导致跟踪器在提取目标特征信息时无法较好的区分目标和背景, 导致信息提取错误, 最终跟踪错误. 因此选择有效的特征对目标和背景进行区分非常必要.

1.3 深度学习中典型的神经网络

深度学习在单目标跟踪领域的成功, 很大程度上得益于神经网络能提取到更优更精细的深度特征, 进而更好的用于后续的识别、跟踪. 本节简要介绍深度学习网络中主流的卷积神经网络、循环神经网络和生成式对抗网络.

1.3.1 卷积神经网络

卷积神经网络 (convolutional neural networks, CNN) 是一类具有强大表征学习能力的前馈神经网络, 一般由卷积层、池化层、全连接层 3 部分组成. 经典的卷积神经网络 LeNet-5[12] 的网络结构如图 1 所示.

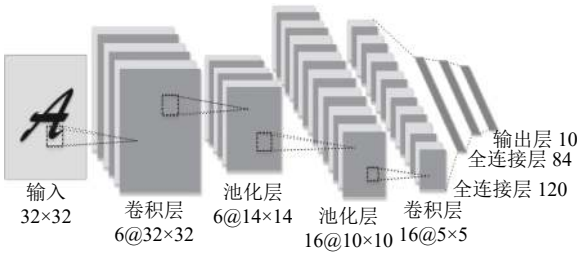

图 1 LeNet-5 网络结构

2012 年 Krizhevsky 等人提出了更深且性能更为优越的 AlexNet 模型[13]. AlexNet 一共 8 层, 通过引入 dropout 方法使全连接层中部分神经元失去作用来缓

解过拟合, 且使用双 GPU 加速进行训练, 在该年的 ImageNet 图像识别大赛中取得了很好的成绩. 2014 年, 牛津大学团队基于如何设计一个更深的网络的设想提出了 VGGNet[14], 将神经网络的深度提升为 11/13/16/19 层. 同年, Szegedy 提出的 GoogleNet[15] 将网络的深度增加到了 22 层, 并通过在网络中引入 Inception 单元进一步提升了模型的整体性能. 2015 年 He 等人通过残差学习的构想成功训练出超深的神经网络 ResNet[16], 且性能并没有下降. 2017 年, Howard 等人提出了一种简单而有效的深度可分离卷积网络 MobileNet[17], 为以后的模型发展提供了另一种可能性, 各模型主要参数对比如表 1 所示.

表 1 深度模型参数对比

模型	提出时间 (年)	ILSVRC成绩	层数	数据增强	卷积核大小	参数量 (Million)
AlexNet	2012	分类、检测、定位冠军	8	Y	3, 5, 11	60
VGGNet	2014	定位比赛冠军	11/13/16/19	Y	3	138
GoogLeNet	2014	分类比赛冠军	22	Y	1, 3, 5, 7	6.8
ResNet	2015	分类、检测、定位冠军	50/101/152	Y	1, 3, 5, 7	—
MobileNet	2017	—	28	Y (少量)	1, 3	4.2

1.3.2 循环神经网络

循环神经网络 (recurrent neural networks, RNN)[18] 适用于处理具有时间序列特性的数据, 其基础结构如图 2 所示. 其中, x_t 为 t 时刻输入层的值, o_t 为 t 时刻输出层的值, u、v、w 为网络中的共享参数. 从图中可以看出, 在 t 时刻的隐藏层的值 h_t 由前一时刻隐藏层的值 h_{t-1} 和输入 x_t 共同得到, 且 h_t 的值不仅用于产生 o_t 的值, 也作用于下一时刻 h_{t+1} 的值, 正是由于这种“记忆”功能, 使得循环神经网络能够挖掘出数据中的时序信息以及语义信息.

1.3.3 生成式对抗网络

受博弈论二元零和博弈的启发, 2014 年, Goodfellow 等人提出生成式对抗网络 (generative adversarial network, GAN)[19], 相比于传统的卷积神经网络和循环神经网络等, GAN 是一种全新的无监督网络架构, 其中包含生成模块和判别模块, 二者互为对抗学习的目标, 其网络结构如图 3 所示.

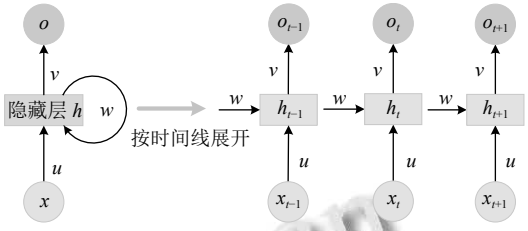

图 2 RNN 模型按时间线展开

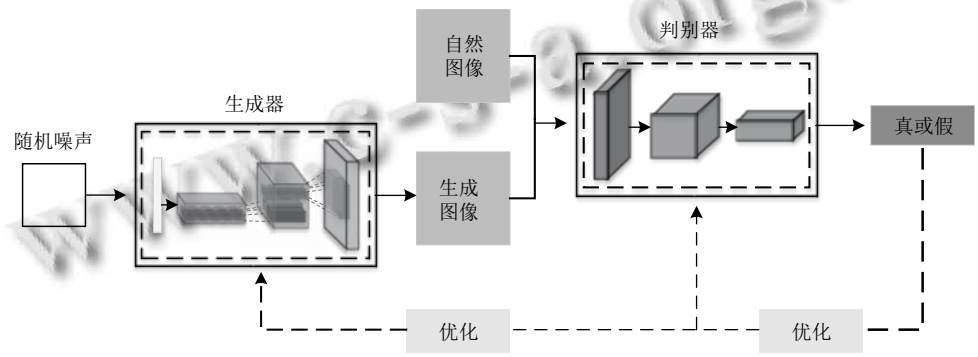

图 3 GAN 网络结构

在 GAN 的训练过程中, 需要固定其中一个模块, 然后优化另一个模块, 如此交替进行, 直至两个模块达到平衡状态, 此时判别网络已经无法判别出数据的真伪. GAN 发展至今已经有了大量的实际用例, 除了应用于图像生成领域之外, 还可以应用于视频生成[20]、艺术品生成[21]、3D 建模[22] 等领域.

2 基于相关滤波跟踪算法

相关滤波算法跟踪过程是首先利用初始帧中目标区域特征来训练一个滤波器, 接着在后续帧中进行相

同区域的特征提取, 然后在频域中对提取到的特征进行相关滤波操作, 最后将响应图中得分最大的区域作为目标区域. 在算法的发展初期提取的目标特征一般为颜色特征、灰度特征等手工特征.

2010 年, Bolme 等人提出了使用单帧即可训练出稳定滤波器的 MOSSE 算法[23]. MOSSE 算法结构简单, 对外观、光照等变化鲁棒性很强, 同时可以高效处理视频帧. 后续 Henriques 等人基于 MOSSE 提出了带有核循环结构的 CSK[24] 和扩展了方向梯度直方图 (histogram of oriented gradients, HOG) 特征的 KCF[25], 进一步提升了跟踪性能. DSST[26] 在 KCF 的基础上引入多特征融合机制, 通过训练尺度滤波器和位置滤波器分别进行目标尺度估计和目标定位, 能够较好地应对跟踪过程中出现的尺度变化, 但跟踪效率有所下降.

总的来说, 虽然相关滤波类算法的速度较快, 但无法在速度与精度之间保持较好的平衡. 随着深度学习的发展, 研究人员开始考虑利用深度学习所能提取的更为鲁棒的深度特征来代替传统相关滤波方法中使用的 CN、HOG 等手工特征, 以提升模型的性能.

2015 年, HCF[27] 以 KCF 为基本的跟踪框架, 利用 VGG-19 中的 Conv3-4、Conv4-4、Conv5-4 层的输出自适应的学习 3 个相关滤波器以对目标外观进行编码, 最后对得到的响应图进行逐层推断以得到最终的预测结果, 由于采用深层网络的输出包含更多的语义信息, HCF 对目标外观变化具有很好的鲁棒性. C-COT[28] 使用隐式插值模型来学习一组连续卷积滤波器, 并通过滤波器与多分辨率特征图进行运算产生目标的连续域置信度图, 进而对目标进行定位, 取得了很好的跟踪效果. 但由于 C-COT 需要处理高维特征, 计算量很大, 所以跟踪效率降低.

2017 年, Danelljan 等人以提高效率的角度出发, 从模型大小、样本集大小、更新策略 3 个方面对 C-COT 进行了改进, 提出了 ECO 算法[29]. ECO 首先采用分解卷积的方法在特征提取上做了简化, 以选取贡献度较高的维度特征. 其次采用高斯混合模型合并相似样本, 并构建差异性较大的样本分组, 使样本集同时具有多样性和代表性. 最后每隔 6 帧进行一次模型更新, 不但提高了算法速度, 而且提高了应对遮挡等挑战的稳定性. 2018 年, Bhat 等人提出了 UPDT 算法[30], 通过对网络中深层和浅层特征的系统性分析, 针对不同的层次提出了不同的处理策略, 最后将处理后的特征相结合进而提升网络的鲁棒性和准确度. 2019 年, Danelljan 等人认为采用多尺度搜索的方式来估计目标状态的处理方式较为简单, 提出了 ATOM[31]. ATOM通过在线训练的目标分类模块对目标进行粗略的定位, 接着通过离线训练的目标估计模块进行精确定位, 在速度达到实时的情况下取得了较好的精度.

3 基于深度学习的跟踪算法

近年来基于深度学习的单目标跟踪算法在很多视觉跟踪挑战赛中取得了很好的成绩, 目前可以主要分为基于孪生网络、基于循环神经网络、基于生成对抗网络单目标跟踪, 其中基于孪生网络的单目标跟踪算法目前已经成为了单目标跟踪领域主流的算法.

3.1 孪生网络跟踪算法

孪生网络以两个样本为输入, 输出其嵌入高维度空间的表征, 以比较两个样本的相似程度, 最后将相似度得分图上得分最高的区域当做目标区域. 2016 年, Tao 等人最先提出了孪生网络跟踪算法 SINT[32], SINT 通过离线训练学习到一个匹配函数, 然后将带有目标框的第一帧数据和一系列候选框送入网络当中, 利用匹配函数在后续帧中进行目标定位. 由于在跟踪过程中需要处理多个候选框, 所以比较耗时.

同年, Bertinetto 等人提出了 SiamFC[33] , SiamFC 采用离线训练并在线微调的跟踪方式, 在保持跟踪精度的同时, 提高了跟踪效率, 成为了未来 Siamese 网络改进的基础. 如图 4 所示, SiamFC 将目标模板 z 和 x 分别作为模板分支和搜索分支的输入, 两个分支权值共享, 并通过特征变换φ提取模板和搜索区域的特征$\varphi(z)$、$\varphi(x)$. 然后, 将$\varphi(z)$当做卷积核在$\varphi(x)$上做滑动卷积, 最后通过得到的相似度得分来确定目标位置.

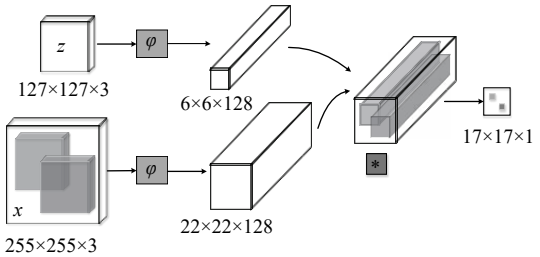

图 4 SiamFC 跟踪框架

虽然 SiamFC 在速度上远超实时, 但其精度却低于一些相关滤波类算法. 表 2 展示了 VOT2015[34] 挑战赛中的前 3 跟踪器与 SiamFC 和 SiamFC-3s 的性能对比,

其中, SiamFC-3s 代表 3 个搜索尺度的 SiamFC. VOT 系列评价指标相较于 OTB 系列拥有独特的“重启”机制, 即在跟踪失败 5 帧后会将跟踪器重置, 更加充分地利用了数据集. 其评价指标主要为精确率 (A) 和期望平均重叠率 (EAO), 其中, *为消除不同计算机性能影响的 EFO (equivalent filter operations) 速度, ↑代表数据越大性能越好, ↓代表数据越小性能越好.

表 2 SiamFC 与 VOT2015 挑战赛 top-3 算法对比

Tracker	A (↑)	# Failure frame (↓)	EAO (↑)	Speed (fps)
MDNet	0.562	46	0.357 5	1
EBT	0.448	49	0.304 2	5
DeepSRDCF	0.535	60	0.303 3	<1*
SiamFC-3s	0.534	84	0.288 9	86
SiamFC	0.524	87	0.274 3	58

3.1.1 基于注意力机制的改进

注意力机制的作用是基于原有的数据找到其之间的关联性并突出某些重要的特征, 同时对不相关的特征进行抑制. 在目标跟踪领域可以高效地获取目标的特征表达, 进一步提升算法的鲁棒性.

2018 年 He 等人提出了 SA-Siam[35], 用于提升 SiamFC 网络的泛化性能. 其跟踪架构如图 5 所示, SA-Siam 的外观分支的基本结构与 SiamFC 基本相同, 主要用于获取目标的表观特征; 语义分支使用 AlexNet网络, 并且不进行更新, 主要用于高层语义信息的提取, 且为了增强语义分支的分辨能力, 添加了一个通道注意力模块, 使语义分支也能够关注到更具有表观信息的通道. 两个分支独立训练, 并在最后进行融合, 进一步提升了模型的泛化能力.

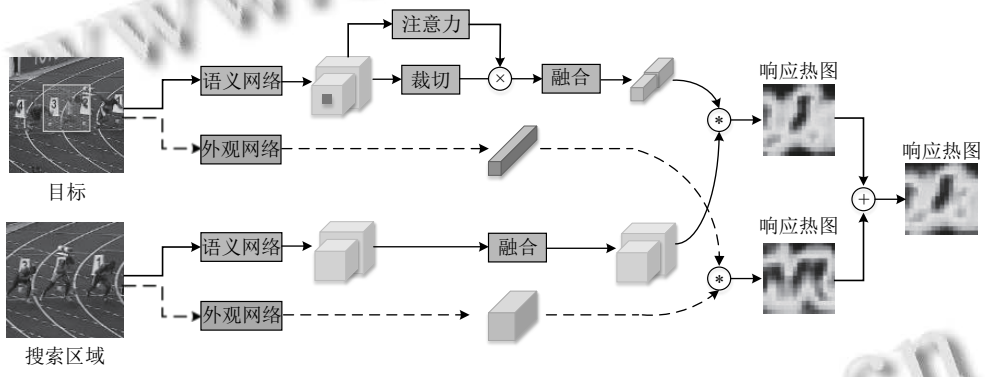

图 5 SA-Siam 跟踪框架

同年, Wang 等人基于 Siamese 网络离线训练无法很好的适应目标变化的问题, 提出了 RASNet[36]. RASNet 通过多注意力的结合代替在线更新, 其中残差注意力模块负责表示目标的全局信息, 生成注意力模块负责表示目标的空间信息, 通道注意力负责表示特征的通道信息, 并通过注意力机制提供的加权互相关将判别学习从特征表示学习中解耦出来, 提升了网络的判别能力和泛化能力, 进一步缓解了过拟合. 2020 年, Yu 等人提出可变形孪生注意网络 SiamAttn[37], 从另一个角度解决离线训练导致的特征学习不足以及目标与搜索分支独立计算导致性能无法提升的问题. SiamAttn 通过自注意力机制学习丰富的上下文信息提升了网络的特征表达能力, 并通过互注意力机制在互相关操作之前交互模板和搜索区域之间的信息, 以隐式的更新模板特征, 提升了算法的鲁棒性.

2021 年, Chen 等人受 Transformer[38] 的启发提出了一种新的基于注意力机制的特征融合网络 TransT[39], 用于解决基于 Siamese 跟踪器中进行特征融合所采用的互相关操作会丢失语义信息, 并容易陷入局部最优的问题. 该网络以模板分支和搜索分支的特征为输入, 通过两个自注意力模块来增强特征表达, 然后通过两个交叉注意特征增强模块来进行特征融合, 将这一步骤重复 N 次, 最后通过附加的交叉特征增强模块将特征有效地融合在一起, 进而获取更为精确的跟踪结果. TransT 的跟踪速度为 50 fps, 满足了实时, 且在大规模数据集 LaSOT[40]、TrackingNet[41]、GOT-10k[42] 均取得了很优的性能.

3.1.2 基于锚 (anchor-based) 的改进

锚的概念最初来源于 Faster-RCNN[43] 中的区域生成网络 (region proposal network, RPN), 用于解决单窗

口无法检测多个目标以及目标多尺度变化的问题. 2018 年, Li 等人提出了 SiamRPN[44], 将用于目标检测的 RPN 模块应用到了目标跟踪当中, 其模型框架如图 6 所示, 其中, k 表示锚框的个数. SiamRPN 的 RPN 模块包含分类和回归两个分支, 在跟踪阶段将目标跟踪定义为局部一次性检测任务, 利用回归分支可以让定位框的定位更加准确, 避免了耗时的多尺度测试, 提升了跟踪精度的同时跟踪速度也达到了 160 fps, 但其在处理相似物体干扰方面效果较为一般, 且泛化能力较弱.

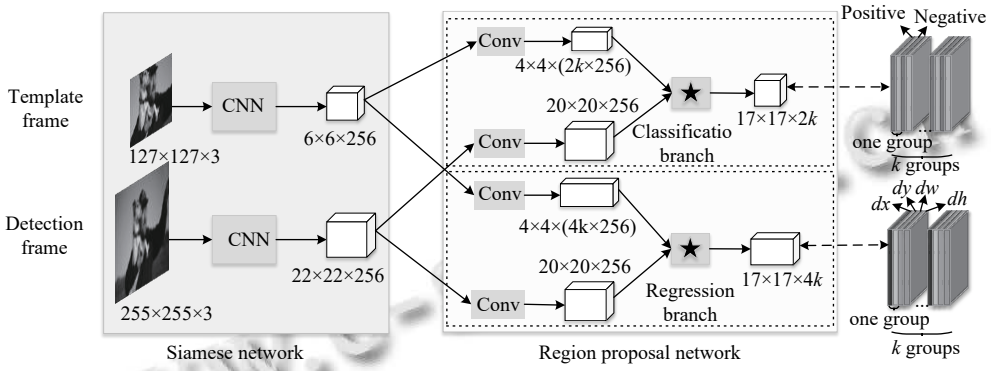

图 6 SiamRPN 跟踪框架

后续仍基于锚的研究主要分为 3 大类, 分别是通过增加训练集的方法训练出更鲁棒的网络, 设计更为强大的骨干网络和更为有效的利用 RPN 模块. 2018 年, Zhu 等人基于 SiamRPN 提出了 DaSiamRPN[45], 在使用 SiamRPN 训练集 ILSVRC2015-VID[46] 和 YouTube-BB[47] 的基础上, 引入用于目标检测的 ILSVRC2015[46] 和 COCO[48] 静态图片数据集来构造训练样本对, 扩充正样本数据, 并使用来自相同类别的负样本对来增加困难负样本数据, 通过正负训练样本的扩充提高了模型的泛化能力和判别能力, 获得了更加鲁棒的跟踪结果.

对于目标跟踪而言, 进行特征匹配、目标定位的关键在于是否能够提取到更为鲁棒的特征, 但图像填充的负面影响使得之前的工作均只能使用 AlexNet 等浅层的网络, 提取的目标特征往往不够具体、全面. 2019 年, Li 等人尝试将深层网络 ResNet 作为孪生网络跟踪器的骨干网络提出了 SiamRPN++[49], 阐述了 Siamese 网络不能使用深度网络进行提取的原因: 缺乏严格的平移不变性, 这种不变性的缺乏会使得网络学习到位置偏见, 使网络更关注图像正中心位置的特征. Li 等人[49] 改变了原有的采样策略, 使得跟踪器在图像中心一定范围内进行均匀采样, 且通过多层特征融合模块解决了模板分支和搜索分支参数量极不平衡的问题, 使得模型更容易被训练, 他们认为 SiamRPN++的出现使得孪生网络类方法第一次在精度上超过了相关滤波类方法.

3.1.3 基于无锚 (anchor-free) 的改进

通过引入 RPN 模块来进行跟踪的方法已经在精度方面达到了很高的水平, 但由于 RPN 模块取消了多尺度搜索, 所以需要在 RPN 模块中精心设计锚框, 小心调整锚框数量、大小和高宽比等超参数, 复杂度提升, 比较耗费时间. 且这些跟踪器在处理大尺度变化和姿态变化等问题时仍存在困难. 后续的研究者们开始研究新的方法来继续改进目标跟踪的精度和速度. 而基于无锚的方法由于其结构简单, 性能优越, 近年来成为单目标跟踪任务中的热门方法. 与基于锚的方法不同, 无锚的方法可以直接预测物体的位置.

2020 年, Chen 等人提出了 SiamBAN[50], 利用全卷积网络的表达能力将跟踪问题看作是分类回归问题, 不需要多尺度搜索模式和预先定义的候选框. 如图 7 所示, SiamBAN 由一个 Siamese 主干网络和多个自适应头模块组成, 其中自适应头模块将主干网络从 Conv3、Conv4、Conv5 层中提取到的特征分别通过深度互相关操作进行融合, 并通过分类分支对目标和背景进行区分, 通过回归分支输出预测框与真实框的偏移量, 并根据偏移量来更新预测框, 最终实现目标跟踪定位. SiamBAN 在多个大型基准数据集上均取得了很优的结果.

SiamCAR[51] 同样基于无锚的设计, 在网络中增加了中心度分支, 通过中心度分支对远离中心点的位置进行约束, 即分配更低的权重. 然后一方面通过分类得分图与中心度得分图得到最佳中心点, 另一方面通过偏差回归得到最佳中心点所对应的偏差坐标, 最后组合起来实现目标跟踪. Ocean[52] 针对分类和回归分支采用不同的采样策略, 对分类分支通过引入特征对齐模块, 将卷积核的固定采样位置转换为与预测的边界框对齐, 增强了分类结果. 同时在孪生结构外引入了在线更新分支, 并与分类分支进行特征融合以增强算法鲁棒性.

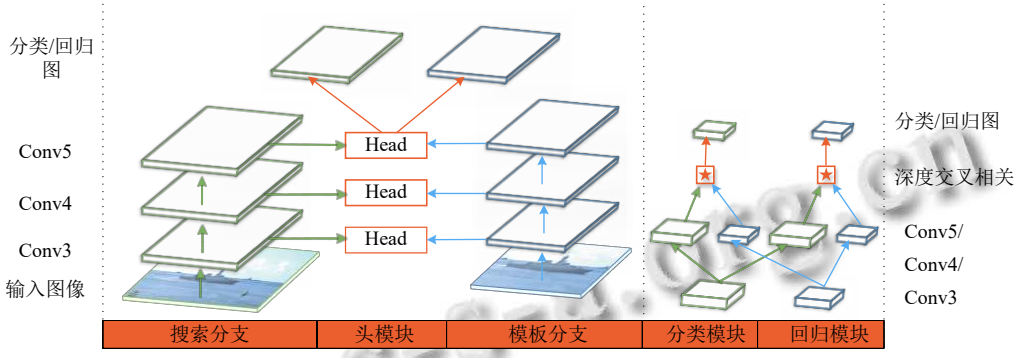

图 7 SiamBAN 跟踪框架

3.2 循环神经网络跟踪算法

虽然卷积神经网络能够提取出更鲁棒的特征表示且泛化能力较强[53], 但基于 CNN 的跟踪算法在面对相似目标时往往不能进行有效的区分, 而循环神经网络可以对具有时序特性的数据进行进一步的信息挖掘, 放大相似目标之间的区别性. 于是研究人员开始探索怎样将循环神经网络用于单目标跟踪领域来解决现有的一些问题.

2016 年, Cui 等人提出了基于循环神经网络的算法 RTT[54], RTT 首先根据上一帧中的目标框确定一个新的候选区域, 接着对这个候选区域进行分块, 并使用多方向 RNN 从 4 个方向遍历每个分块区域, 最后得到目标置信图, 进而完成对目标位置的预测. 其中分块提取特征并使用 RNN 进行处理的策略使得 RTT 能更好的处理遮挡以及目标外观变化的问题. 文献 [55] 提出了基于循环神经网络的结构感知视觉跟踪网络 SANet, SANet 首先利用 CNN 对每帧中的目标和背景进行有效区分, 然后使用 RNN 在不同深度的层中利用目标的自身结构信息对目标进行建模, 最后通过跳跃式连接, 将 CNN 特征与 RNN 特征进行融合, 有效地对目标和相似物体进行区分.

由于 RNN 无法解决长输入序列信息传递时网络容易产生的一系列梯度问题, Hochreiter 等人在 RNN 的基础上提出了长短期记忆网络 LSTM[56], LSTM 允许网络保留或遗忘传递过程中的信息, 使得网络可以处理任意长度的序列信息. 2018 年, Yang 等人提出了一个动态的记忆网络 MemTrack[57], 通过 LSTM 控制模块对记忆模块的读取与写入动态的控制最后的模板生成, 很好的解决了孪生网络等模板匹配类算法无法很好适应目标外观变化的问题, 且速度超过了实时.

3.3 生成式对抗网络跟踪算法

基于 GAN 可以生成不同的新数据的启发, 2018 年, Song 等人将对抗学习的思想应用在视觉跟踪领域, 以解决训练样本中正负样本数极度不平衡以及正样本之间差异性较小的问题, 提出了 Vital[58]. Vital 通过一个生成对抗网络来产生一个随机遮挡掩膜, 在对抗学习的作用下, 遮挡掩膜保留了目标特征中最鲁棒部分, 使得跟踪器能够在长时间内获得大量的外观变化, 从而获得更加鲁棒的跟踪效果. 同年, Wang 等人提出了 SINT++[59], 该算法假设所有的目标样本位于同一流形空间内, 通过使用对抗学习来对正样本生成遮挡块以此来构造困难正样本, 解决了正样本之间差异性较小的问题, 对遮挡和尺度变化的目标具有很强的鲁棒性.

2020 年, Yan 等人着重研究对抗学习中的对抗攻击, 针对 SiamRPN++[49] 提出了一种有效且高效的冷却收缩攻击策略[60]. SiamRPN++将得到的分类热图中响应最高的区域作为目标存在的区域, 冷却收缩策略通过攻击 SiamRPN++中的搜索区域使得目标区域的响应值降低, 导致跟踪器无法准确定位目标, 并使用收缩损失最小二乘法迫使 SiamRPN++的回归预测边界框变小, 进而导致跟踪失败, 成功欺骗了 SiamRPN++. 且此方法具有良好的迁移性, 能够成功应用到 DaSiamRPN

和 DiMP[61] 等高性能跟踪器, 为以后研究更为高效的对抗生成学习提供了可能.

3.4 算法对比分析

目前单目标跟踪技术的研究重点正从短时跟踪向长时跟踪发展, 因为长时跟踪更能反映出评测算法的实用性. 其中 LaSOT[40] 为 2019 年 Fan 等人提出的偏向于长期跟踪的数据集, 包括 1 400 个序列视频, 每个视频平均 2 512 帧, 每帧都具有高质量的手工密集注释, 并且为每个视频提供了额外的语言描述, 很大程度上满足了复杂算法的特征获取.

在此我们选取包括上述介绍的近年的优秀跟踪算法, 根据 AUC、P_{Norm}、P 三个评价指标在 LaSOT 数据集上的实验结果进行对比, 其中, AUC 是以纵轴为真正率 (true positive rate, TPR), 横轴为伪正率 (false positive rate, FPR) 的坐标轴上形成的 ROC 曲线下的面积 (参照式 (1)、式 (2)), 能够直观反映出跟踪器分类效果的好坏. P_{Norm} 为归一化方法去除图像和回归框大小对精度影响后的精确度值, P 为精确度值.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

$$P = \frac{TP}{TP + FP} \tag{3}$$

其中, TP (true positive) 代表真正类, TN (true negative) 代表真负类, FP (false positive) 代表假正类, FN (false negative) 代表假负类.

结果如表 3 所示, 其中, ↑表示数据越大性能越好. 首先, 相关滤波类算法 ATOM 由于综合了相关滤波类算法在线更新和孪生网络类算法离线训练的优点, AUC 评分优于 SiamRPN++和 SiamBAN, 但整体性能相较于深度学习类方法相比仍处于下游. 其次, 在深度学习类方法中, 基于生成对抗网络的 Vital主要关注目标外观变化问题, 且没有添加回归机制, 整体评分最低. 循环神经网络类算法 ROAM[62] 表现也相对较差, 孪生网络类算法则整体表现很好. 最后, 在孪生网络算法中, 添加注意力机制的 TransT 能够更为有效的关注有用的特征表达, 取得了最好的性能评分, 同样基于注意力的 SiamAttn 也取得了较优的性能. 各类别主流算法的分析和总结如表 4 所示.

表 3 LaSOT 数据集上算法的性能比较 (★ 为 LaSOT[39] 评测结果)(%)

性能指标	Vital★	SiamRPN++	ATOM	ROAM++	SiamBAN	SiamAttn	Ocean	SiamR-CNN	TransT
AUC得分 (↑)	41.2	49.6	51.5	44.7	51.4	56.0	56.0	64.8	64.9
归一化精确率 (↑)	48.4	57.0	57.6	54.3	59.8	64.8	65.1	72.2	73.8
精确率 (↑)	37.2	—	50.5	44.5	52.1	—	56.6	68.4	69.0

表 4 主流算法的分析与总结

方法类别	主流算法	算法特点	方法的优点	方法的缺点
相关滤波跟踪算法	ATOM	在线训练分类模块对目标粗略定位, 离线训练状态估计模块进行精确定位	具有在线更新的设计, 目标定位更为准确	模型设计复杂度高, 且在线更新导致速度降低
孪生网络跟踪算法	SiamFC	将模板特征与搜索区域特征进行互相关操作, 获取响应得分图, 本质是模板匹配寻找最优结果	权值共享, 降低了网络的复杂度, 并充分利用了CNN强大的特征表达能力, 能够在算法的性能和速度之间保持较好的平衡	为了跟踪速度的提升, 很多算法没有设计进行模板更新, 无法很好的适应目标和背景变化; 且在跟踪中容易造成误差累积, 导致算法的鲁棒性相对较差
	基于注意力机制 TransT	利用Transformer结构来替代Siamese网络的互相关操作, 获取到了更多的语义信息		
	基于锚框 SiamRPN++	改变了Siamese网络原有的采样策略, 成功将深层网络ResNet作为骨干网络进行训练; 提出了深度互相关操作, 大大减少了网络中的参数量		
	基于无锚框 SiamBAN	取消了多尺度搜索和预先定义的候选框, 降低了网络设计的复杂度; 改变了正负样本的定位策略		
循环神经网络跟踪算法	MemTrack	通过引入LSTM来动态的控制模板生成, 进而适应目标形状的变化	较为充分的利用了网络间的时序信息, 能够较好的应对跟踪过程中目标的外观变化以及相似目标的区分	网络参数量大, 且对处理网格化的数据(如图像)能力不如CNN, 整体性能相对较差
	ROAM++	使用LSTM循环生成用于跟踪模型优化的自适应学习速率, 进而循环优化模型		
生成式对抗网络跟踪算法	Vital	通过GAN生成随机遮挡掩膜, 用来捕获目标的一系列外观变化; 提出高阶代价敏感损失函数, 降低易被分类为正样本的负样本带来的分类影响	利用GAN可以缓解训练样本分布的不平衡问题, 使网络更好的适应遮挡、形变等挑战	GAN本身在训练时不容易收敛, 且容易受到模式坍塌的影响

4 结论与展望

单目标跟踪目前仍是计算视觉领域具有高实用性的热门课题之一, 虽然深度学习与单目标跟踪结合时间较短, 但大量的优秀算法已经被提出. 与相关滤波类算法相比, 基于深度学习类算法, 特别是孪生网络跟踪算法, 由于其基于 CNN 的强大特征表达能力和离线训练的策略, 使得其在跟踪精度与速度方面能够达到很好的平衡. 而 RNN 和 GAN 在单目标跟踪领域的成功应用也使得跟踪器能更好的挖掘时序序列的信息并更好的处理正负样本失衡的情况. 但在复杂的实际应用场景中, 这些算法的整体性能仍有待于优化. 本文对基于深度学习的单目标跟踪算法进行了综述, 并对未来的发展趋势做了展望.

(1) 注意力机制的进一步探索

在单目标跟踪领域, 注意力机制可以高效地帮助跟踪器获取目标的特征表达, 进而提升网络的分辨能力, 因此一直被广泛使用. 但注意力机制的添加并不是简单的模块堆叠, 怎样合理使用至关重要. 最近 Transformer 网络因为合理嵌入多种注意力机制被广泛应用于多个领域. 2021 年, Yan 等人基于 Transformer 提出 STARK[63], 通过学习一种强大的时空联合表示, 可以捕获丰富的全局信息, 进一步刷新了 LaSOT[40] 等大型数据集上的最佳性能, 也证明了未来注意力机制仍有很大的潜力, 需要研究人员进一步探索.

(2) 相关领域中的技术迁移

2018 年, RPN 模块的引进使得 Siamese 系列跟踪器的性能得到了大幅度提升, 特别是 SiamRPN++[49] 的出现使孪生网络类算法精度得到了进一步提升. 2020 年, Voigtlaender 等人借鉴目标检测领域中的 Faster-RCNN 提出再检测孪生网络架构 SiamR-CNN[64], 在多个数据集上均取得了很优的性能, 未来通过引入其他领域的相关技术仍是目标跟踪发展的一种趋势.

(3) 在线更新机制的增加

虽然目前基于深度学习技术的单目标跟踪已经取得了很好的发展, 但在工业界中的应用却仍有不足. 主要原因在于大多数算法为了保持高速的跟踪状态往往使用预训练数据, 以提供通用的目标表示, 并降低由于训练数据不平衡导致的过拟合风险, 在跟踪过程中并不进行模板更新, 十分依赖离线训练过程中所学习到的模板. 这样一方面当跟踪出现错误时, 容易造成误差累计, 出现跟踪漂移等情况; 另一方面, 当跟踪器遇到“没有见过”的目标时, 跟踪效果往往较差. 为适应在现实场景中的应用, 在线更新机制应是未来单目标跟踪发展的重点之一.

(4) 与 GAN 进一步结合

在现实应用场景中, 遮挡是常见的挑战之一, 由于遮挡时会丢失相当部分的特征信息, 往往会使跟踪发生漂移. 而 GAN 可以通过构造网络中的困难负样本, 使网络获取更具判别力的特征, 进而增加跟踪器的判别能力, 更好的应对跟踪中遮挡情况的出现. 与 GAN 的进一步结合, 会使跟踪器进一步应用于现实场景中.

(5) 设计适用于长时跟踪的跟踪器

以往的跟踪器大都针对短时跟踪任务, 而长时跟踪过程中会遇到更多的跟踪挑战, 更贴近于现实应用场景. 随着 LaSOT[40]、GOT-10k[42] 等长时跟踪数据集的出现, 对现有的跟踪器提出了更高的要求. 而现有的解决方法往往是在短时跟踪算法的基础上添加重检测模块, 并不能很好的应对长时跟踪任务. 在以后的发展过程中, 设计针对性的长时跟踪算法也是研究重点之一.

参考文献

1 Lee KH, Hwang JN, Okopal G, *et al*. Ground-moving-platform-based human tracking using visual SLAM and constrained multiple kernels. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(12): 3602–3612. [doi: 10.1109/TITS.2016.2557763]

2 Gao M, Jin LS, Jiang YY, *et al*. Manifold Siamese network: A novel visual tracking ConvNet for autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 2020, 21(4): 1612–1623. [doi: 10.1109/TITS.2019.2930337]

3 Brown M, Funke J, Erlien S, *et al*. Safe driving envelopes for path tracking in autonomous vehicles. Control Engineering Practice, 2017, 61: 307–316. [doi: 10.1016/j.conengprac.2016.04.013]

4 Li K, Cheng J, Zhang QS, *et al*. Hand gesture tracking and recognition based human-computer interaction system and its applications. Proceedings of 2018 IEEE International Conference on Information and Automation. Wuyishan: IEEE, 2018. 667–672.

5 Lim KM, Tan AWC, Lee CP, *et al*. Isolated sign language recognition using convolutional neural network hand modelling and hand energy image. Multimedia Tools and

Applications, 2019, 78(14): 19917–19944. [doi: 10.1007/s11042-019-7263-7]

6 Wang NY, Shi JP, Yeung DY, *et al*. Understanding and diagnosing visual tracking systems. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3101–3109.

7 孟琭, 杨旭. 目标跟踪算法综述. 自动化学报, 2019, 45(7): 1244–1260.

8 Zhang ZT, Zhang JS. A new real-time eye tracking based on nonlinear unscented Kalman filter for monitoring driver fatigue. Journal of Control Theory and Applications, 2010, 8(2): 181–188. [doi: 10.1007/s11768-010-8043-0]

9 Chang C, Ansari R. Kernel particle filter for visual tracking. IEEE Signal Processing Letters, 2005, 12(3): 242–245. [doi: 10.1109/LSP.2004.842254]

10 Du K, Ju YF, Jin YL, *et al*. Object tracking based on improved MeanShift and SIFT. Proceedings of the 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet). Yichang: IEEE, 2012. 2716–2719.

11 Wu Y, Lim J, Yang MH. object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834–1848. [doi: 10.1109/TPAMI.2014.2388226]

12 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]

13 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 26th Annual Conference on Neural Information Processing Systems. Lake Tahoe: NIPS, 2012. 1106–1114.

14 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556, 2014.

15 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 1–9.

16 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

17 Howard AG, Zhu ML, Chen B, *et al*. MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv: 1704.04861, 2017.

18 Lipton ZC, Berkowitz J, Elkan C. A critical review of recurrent neural networks for sequence learning. arXiv: 1506.00019, 2015.

19 Goodfellow I, Pouget-Abadie J, Mirza M, *et al*. Generative adversarial nets. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: ACM, 2014. 2627–2680.

20 Vondrick C, Pirsiavash H, Torralba A. Generating videos with scene dynamics. Advances in Neural Information Processing Systems, 2016, 29: 613–621.

21 Wu JJ, Zhang CK, Xue TF, *et al*. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. Proceedings of Conference on Neural Information Processing Systems. Barcelona: NIPS, 2016. 82–90.

22 Ghosh P, Gupta PS, Uziel R, *et al*. GIF: Generative interpretable faces. Proceedings of 2020 International Conference on 3D Vision (3DV). Fukuoka: IEEE, 2020. 868–878.

23 Bolme DS, Beveridge JR, Draper BA, *et al*. Visual object tracking using adaptive correlation filters. Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco: IEEE, 2010. 2544–2550.

24 Henriques JF, Caseiro R, Martins P, *et al*. Exploiting the circulant structure of tracking-by-detection with kernels. Proceedings of the 12th European Conference on Computer Vision. Florence: Springer, 2012. 702–715.

25 Henriques JF, 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. [doi: 10.1109/TPAMI.2014.2345390]

26 Danelljan M, Häger G, Khan FS, *et al*. Accurate scale estimation for robust visual tracking. Proceedings of British Machine Vision Conference 2014. Nottingham: BMVA Press, 2014. 1–11.

27 Ma C, Huang JB, Yang XK, *et al*. Hierarchical convolutional features for visual tracking. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3074–3082.

28 Danelljan M, Robinson A, Khan FS, *et al*. Beyond correlation filters: Learning continuous convolution operators for visual tracking. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 472–488.

29 Danelljan M, Bhat G, Khan FS, *et al*. ECO: Efficient convolution operators for tracking. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern

Recognition. Honolulu: IEEE, 2017. 6931–6939.

30 Bhat G, Johnander J, Danelljan M, *et al*. Unveiling the power of deep tracking. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 483–498.

31 Danelljan M, Bhat G, Khan FS, *et al*. ATOM: Accurate tracking by overlap maximization. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4660–4669.

32 Tao R, Gavves E, Smeulders AWM. Siamese instance search for tracking. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1420–1429.

33 Bertinetto L, Valmadre J, Henriques JF, *et al*. Fully-convolutional siamese networks for object tracking. Proceedings of European Conference on Computer Vision. Amsterdam: Springer, 2016. 850–865.

34 Kristan M, Matas J, Leonardis A, *et al*. The visual object tracking VOT2015 challenge results. Proceedings of 2015 IEEE International Conference on Computer Vision Workshop. Santiago: IEEE, 2015. 564–586.

35 He AF, Luo C, Tian XM, *et al*. A twofold siamese network for real-time object tracking. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4834–4843.

36 Wang Q, Teng Z, Xing JL, *et al*. Learning attentions: Residual attentional siamese network for high performance online visual tracking. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4854–4863.

37 Yu YC, Xiong YL, Huang WL, *et al*. Deformable siamese attention networks for visual object tracking. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6727–6736.

38 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of Annual Conference on Neural Information Processing Systems. Long Beach: NIPS, 2017. 5998–6008.

39 Chen X, Yan B, Zhu JW, *et al*. Transformer tracking. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 8126–8135.

40 Fan H, Lin LT, Yang F, *et al*. LaSOT: A high-quality benchmark for large-scale single object tracking. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5369–5378.

41 Muller M, Bibi A, Giancola S, *et al*. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 300–327.

42 Huang LH, Zhao X, Huang KQ. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1562–1577. [doi: 10.1109/TPAMI.2019.2957464]

43 Ren SQ, He KM, Girshick RB, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of Annual Conference on Neural Information Processing Systems. Montreal: NIPS, 2015. 91–99.

44 Li B, Yan JJ, Wu W, *et al*. High performance visual tracking with siamese region proposal network. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8971–8980.

45 Zhu Z, Wang Q, Li B, *et al*. Distractor-aware siamese networks for visual object tracking. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 103–119.

46 Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

47 Real E, Shlens J, Mazzocchi S, *et al*. YouTube-boundingboxes: A large high-precision human-annotated data set for object detection in video. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 7464–7473.

48 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755.

49 Li B, Wu W, Wang Q, *et al*. SiamRPN++: Evolution of siamese visual tracking with very deep networks. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4277–4286.

50 Chen ZD, Zhong BN, Li GR, *et al*. Siamese box adaptive network for visual tracking. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6667–6676.

51 Guo DY, Wang J, Cui Y, *et al*. SiamCAR: Siamese fully convolutional classification and regression for visual

tracking. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6268–6276.

52 Zhang ZP, Peng HW, Fu JL, *et al*. Ocean: Object-aware anchor-free tracking. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 771–787.

53 陆峰, 刘华海, 黄长缨, 等. 基于深度学习的目标检测技术综述. 计算机系统应用, 2021, 30(3): 1–13. [doi: 10.15888/j.cnki.csa.007839]

54 Cui Z, Xiao ST, Feng JS, *et al*. Recurrently target-attending tracking. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 1449–1458.

55 Fan H, Ling HB. SANet: Structure-aware network for visual tracking. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE, 2017. 2217–2224.

56 Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

57 Yang TY, Chan AB. Learning dynamic memory networks for object tracking. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 153–169.

58 Song YB, Ma C, Wu XH, *et al*. VITAL: Visual tracking via adversarial learning. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8990–8999.

59 Wang X, Li CL, Luo B, *et al*. SINT++: Robust visual tracking via adversarial positive instance generation. Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4864–4873.

60 Yan B, Wang D, Lu HC, *et al*. Cooling-shrinking attack: Blinding the tracker with imperceptible noises. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 987–996.

61 Bhat G, Danelljan M, Van Gool L, *et al*. Learning discriminative model prediction for tracking. Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6181–6190.

62 Yang TY, Xu PF, Hu RB, *et al*. ROAM: Recurrently optimizing tracking model. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6717–6726.

63 Yan B, Peng HW, Fu JL, *et al*. Learning spatio-temporal transformer for visual tracking. arXiv: 2103.17154, 2021.

64 Voigtlaender P, Luiten J, Torr PHS, *et al*. Siam R-CNN: Visual tracking by re-detection. Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6577–6587.