基于深度学习的视觉单目标跟踪综述

张长弓，杨海涛†，王晋宇，高宇歌，李高源，冯博迪

(航天工程大学 航天信息学院, 北京 101400)

摘 要：单目标跟踪是一种在视频中利用目标外观和上下文信息对单个目标分析运动状态、提供定位的技术，在智能监控、智能交互、导航制导等方面具有应用前景，但遮挡、背景干扰、目标变化等问题导致实际应用的进展缓慢。随着近年来深度学习的快速发展，研究使用深度学习技术优化单目标跟踪算法已成为计算机视觉领域的热点之一。围绕基于深度学习的单目标跟踪算法，文章在分析了单目标跟踪的基本原理的基础上，从相关滤波、孪生网络、元学习、注意力、循环神经网络和生成对抗网络六个方面，根据核心算法的不同分别进行了概述和分析。此外，文章对研究现状进行了总结，提出了算法的发展趋势和优化思路。

关键词：单目标跟踪；深度学习；孪生网络；相关滤波；元学习；注意力机制

中图分类号：TP18 doi: 10.19734/j.issn.1001-3695.2021.03.0036

Survey on visual single object tracking based on deep learning

Zhang Changgong, Yang Haitao†, Wang Jinyu, Gao Yuge, Li Gaoyuan, Feng Bodi

(*School of Space Information*, *Space Engineering University*, *Beijing* 101400, *China*)

Abstract: Single object tracking(SOT) is a technique to analyze the motion status and provide localization of the single target in video, which uses target appearance and context information. It has promising applications in intelligent surveillance, intelligent interaction, navigation and guidance, etc. However, SOT faces problems such as occlusion, background interference and appearance variation, etc. These problems have led to slow progress in practical applications. With the rapid development of deep learning in recent years, the study of using deep learning techniques to optimize SOT algorithms has become one of the hot spots in computer vision. Around the SOT algorithm based on deep learning, this article respectively provided outline and analysis of each of the six aspects of correlation filter, siamese networks, meta-learning, attention, recurrent neural networks and generative adversarial networks according to the core algorithms, after analyzing the basic principles of SOT. In addition, this article summarized the current state of research and proposed the development trend and optimization ideas of the algorithms.

Key words: single object tracking; deep learning; siamese network; correlation filter; meta learning; attention mechanism

0 引言

单目标跟踪是计算机视觉领域的研究方向之一，其基本任务是获取视频中单个目标在每一帧中的位置信息，为对目标的运动行为及规律的分析和理解提供基础，以便完成更进一步的研究。随着计算机硬件性能的不断提高，越来越多的领域开始出现基于单目标跟踪的具体应用。在智能交通系统中，目标跟踪被用来解决交通问题，如交通监控[1]，行人检测[2]。在军事上，单目标跟踪的应用是当今及未来的热点问题之一，尤其在制导，侦察，飞行器跟踪[3]等方面，其可以用来锁定军事目标，分析战场环境，提供武器导航和导弹预警。基于视频的人机交互领域也非常活跃，手势识别[4]，动作捕捉[5]等应用是计算机获取人体动作高级语义的基础。

单目标跟踪的研究最早可以追溯到 1981 年 Lucas 等对光流法的利用[6]，近年来，随着各种高质量数据集的提出、机器学习和深度学习的发展，单目标跟踪领域涌现出越来越多的优秀算法。目标跟踪的发展大致分为三个阶段：第一阶段集中在 2000 年前后，以 LK-Tracker 为始，这期间主要是经典算法和机器学习在目标跟踪上的应用。这类算法计算复杂度低，运行速度很快，但是鲁棒性和准确性都比较低。第二阶段在 2010 年到 2016 前后，随 MOSSE[9]的提出，相关滤波方法成为研究热点之一，并且凭借速度快，准确性高的特点在各个评估数据集中都有良好的排名。这期间深度学习方法也开始在图像处理领域有所成果，出现了如 MDNet[10]等较为出色的算法。最后一阶段是 2016 年至今，以孪生网络为代表的深度学习方法随着数据集的丰富，不断提高算法的鲁棒性和准确性，在 VOT 等比赛中名列前茅，显示出端到端学习的强大能力。

本文主要围绕深度学习对近几年的优秀算法进行梳理和讨论，旨在为单目标跟踪算法的下一步研究和发展提供参考。

1 基本原理

1.1 基本框架

单目标跟踪是计算机视觉中具有挑战性的研究内容之一，在过去的几十年中取得了长足的发展，并且自提出以来，视觉跟踪的流程就已经确定。对于视频序列，首先根据初始帧中目标的状态初始化跟踪器，然后提取目标特征并建立目标模型，在后续帧中使用跟踪策略如相关滤波、光流、深度学习等，基于目标模型估计目标在当前帧的状态，最后利用当前状态更新目标模型，继续下一帧的跟踪。其基本框架如图

收稿日期：2021-03-17；**修回日期：**2021-04-29

作者简介：张长弓(1996-)，男，山东济宁人，硕士研究生，主要研究方向为目标跟踪；杨海涛(1979-)，男(通信作者)，山东人，副教授，博导，博士，主要研究方向为机器学习(yanghtt@126.com)；王晋宇，男，山西晋中人，硕士研究生，主要研究方向为生成对抗网络；冯博迪，女，陕西渭南人，主要研究方向为图像分类；李高源，男，河南商丘人，硕士研究生，主要研究方向为遥感图像处理；高宇歌，男，安徽人，硕士研究生，主要研究方向为目标识别.

1，主要分为目标初始化，运动模型，提取特征，观测模型和模型更新几个步骤。

a) 初始化目标。跟踪器的工作需要给定一个特定目标，指定目标的过程一般是在视频序列的初始帧中将目标用边界框标出。边界框一般为矩形框，常以左上角坐标及宽和高表示其范围。

b) 运动模型。基于视频和物体运动都是连续的这一特性，可以认为帧之间的目标位置不会相距过远，因此对目标的运动过程建立模型，在前一帧目标位置估计的基础上，利用这一模型可以在目标周围生成生成一组候选区域，这些区域中可能包含目标的当前位置。

c) 提取特征。特征是对目标的抽象的表示，即从目标原始空间映射到某一特征空间。特征提取器使用某些特征表示候选区域中的候选目标。特征提取就是将原始图像数据通过映射变化得到更有利于目标描述的表达方式。

按照表征目标特征的方法，分为基于手工特征的算法和基于深度特征的算法。传统目标算法多使用手工特征，如灰度特征，颜色特征，纹理特征等。深度特征即经过神经网络提取的特征。浅层深度特征包含纹理等外观特征，深层特征含有高度抽象的语义信息如目标的类别。

d) 观测模型。观测模型用于从候选区域提取的特征中判断是否为目标。根据算法不同，一般将其分为两类：生成式方法和判别式方法。

生成式方法主要是利用目标特征学习出代表目标的外观模型，通过它搜索候选区域进行模式匹配，在图像中找到和模型最匹配的区域。生成式模型主要是对后验概率建模，利用统计原理估计联合概率，反映了模板图像和候选图像之间的相似度。其着重描述了目标外观特征的分布，具有较强的特征表征能力。但其忽略了图像背景信息，在遇到遮挡等干扰时容易发生模型漂移，导致跟踪失败。

另一方面，判别式方法则是将目标跟踪作为分类或者回归问题进行解决，其思路是图像特征联合机器学习，以目标信息为正样本，背景信息为负样本，训练一个判别函数来分离目标和背景，寻找最优区域，从而达到跟踪的目的，因此也被称为检测式跟踪。判别式方法在跟踪过程中充分利用了前景目标和背景的特征信息，弥补了生成式方法背景信息利用不充分的缺陷，因而具有较强的鲁棒性。但是判别式方法易受样本数据影响，如正负样本不均，正样本重复等问题。

e) 模型更新。这一步骤主要控制观测模型的更新策略和频率。目标在视频中是动态的，其时空信息和特征持续发生变化，所以需要更新观测模型以适应新的目标特征。如果模型复杂，较高频率的更新会增加计算量，降低跟踪器的运行速度；如果更新频率低于目标的变化速度，易积累误差最终导致模型漂移。

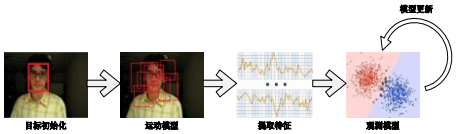

图 1　单目标跟踪的基本框架[24]

Fig. 1　Base framework of single object tracking

1.2　挑战

在视频目标的跟踪过程中，如何快速且鲁棒的跟踪目标一直是一个难题。由于现实生活中物体多样的运动、变化的背景等复杂因素，基于视频的目标跟踪面临着很多挑战，文献[7]总结了以下几点难题：3D 世界在 2D 图像上的投影所带来的信息丢失；图像噪声；物体的复杂运动；物体的非刚性导致的外观变化；部分或完全遮挡；物体的复杂外形；背景光照变化和实时性的要求。目前仍未有一种方法解决以上所有问题，这些挑战仍需要研究人员解决。Wu 等在 OTB100 数据集中进一步将其总结为：形变，光照变化，快速移动，背景干扰，旋转，尺度变化，遮挡[12]。

a) 形变。由于目标是非刚性的，因此在运动过程中会发生一定的形变。依据目标类型不同，可能产生的形变量也不同，较大的形变会导致目标外形轮廓特征剧烈变化(图 2)。

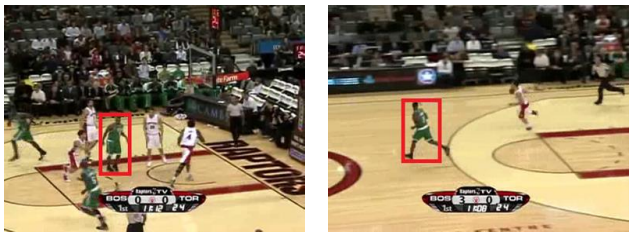

图 2　目标发生形变

Fig. 2　Defomation

b) 光照变化。环境或者目标的亮度会随着太阳、灯光等光源的变化而改变，这种变化会使目标的像素值产生改变，极大的改变目标的颜色相关的特征(图 3)。

图 3　环境光照发生变化

Fig. 3　Illumination variation

c) 运动模糊。当目标相对于图像输入设备的输入速度进行快速运动时，目标周围会出现重影，目标本身会产生一定的模糊效果，不仅视觉上会形变，目标自身的外观特征也会产生损失(图 4)。

图 4　目标产生运动模糊

Fig. 4　Motion blur

d) 背景干扰。在跟踪过程中，可能会在目标周围的背景中出现与目标特征如颜色、外形和纹理相似的物体，使得跟踪器容易误判，跟踪结果漂移到相似目标上(图 5)。

图 5　目标受到背景干扰

Fig. 5　Background clutters

e) 旋转。旋转有两种：平行视频输入设备旋转和非平行旋转。平行旋转使目标像素一定偏移。非平行旋转在损失旧特征的同时增加大量新特征，同时产生形变(图 6)。

(a)原图　　(b)非平行旋转

(c)平行旋转

图 6　目标发生旋转

Fig. 6　Rotation

f) 尺度变化。目标进行相对视频输入设备的运动，导致与设备的距离不断变化，进而使得目标比例发生改变，一般伴随尺度变化产生的还有旋转(图 7)。

图 7　目标发生尺度变化

Fig. 7　Scale variation

g) 遮挡。三维世界中，目标和其他物体之间的相对运动会导致目标被遮挡。遮挡分为部分遮挡和全部遮挡，部分遮挡指目标还保留有部分特征信息，全部遮挡指目标被完全遮挡，失去所有特征。另外，目标超出图像边界也被视为遮挡的一种(图 8)。

图 8　目标发生遮挡

Fig. 8　Occlusion

另外，由于目标跟踪的现实应用意义较强，因此对于实时性也有一定的要求，如果算法和模型过于复杂，会使得跟踪速度降低。例如 Danelljan 等提出的 CCOT[13]算法以其优异的准确度和鲁棒性，在 VOT2016[14]上取得了平均性能排名第一的成绩，但是后来的研究验证 CCOT 的速度只有 0.3 帧/每秒[15]，这种速度下难以进行实际应用。如果要实现视觉上较为平滑的输出视频效果，输出帧速率必须达到至少 15fps[8]。

1.3　数据集

优秀数据集可以很好地推动目标跟踪的发展，它们不仅为各种深度学习算法提供了丰富的训练数据，还为算法的评估提供了有效的测评工具和度量指标。从 PETS，VIVID[16]，CAVIAR 等开始，研究人员提供用以评估和训练的公开数据集，各种具备良好评估标准和视频类别丰富的数据集层出不穷，极大的推动了目标跟踪算法的进步(见表 1)。

a) OTB。2013 年，Wu 等在公开数据集中收集了 50 个视频序列，提出了 OTB2013[17]。OTB2013 包含 51 个视频序列，提供了鲁棒性评价指标 OPE、TRE 和 SRE，同时其含有单目标跟踪的 11 种挑战，对跟踪器的测试十分全面，为跟踪器的检验和比较提供了一个平台，促进了目标跟踪算法的发展。2015 年，Wu 等将数据集扩充到 100 个视频序列，提出了 OTB100[12]。

b) VOT。VOT[14,1824]是目标跟踪领域的大型竞赛数据集，从 2013 年开始承办以来，每年都会进行一次。VOT 为社区提供了一种精确定义且可重复的比较长短期跟踪器的方法，以及一个视觉跟踪领域评估和进步的通用平台。VOT 的挑战内容逐年增多，从最初的短期跟踪到现在的短期、长期、RGB+热成像和 RGB+深度跟踪，丰富性和复杂度都在不断提升。

c) UAV123。Matthias 等于 2016 年构建了 UAV123 数据集[25]，123 个视频序列全部来自于无人机拍摄。UAV123 视频序列平均长度在 1000 帧左右，同时拍摄角度变化大，整体来说目标的变化频繁，对跟踪器的适应能力提出了挑战。

d) TrackingNet。TrackingNet[26]是 Matthias 等于 2018 年设计的大规模数据集，用于户外目标跟踪。该数据集包含 30000 个以上视频序列，具有丰富的目标种类。密集的数据标注使得目标跟踪算法的设计更偏重于挖掘视频中运动目标的时序信息。跟踪器在 TrackingNet 上进行微调后得到的性能平均优于原跟踪算法。

e) GOT-10K。GOT-10K[27]是中科院自动化所 Huang 等于 2018 年提出的通用大规模目标跟踪数据集，为了让训练出的模型具有更强的泛化能力，训练集和测试集之间不存在交集。该数据集包含超过 10000 个视频序列，其中 184 个用于测试，包含丰富的跟踪目标。

f) LaSOT。2019 年，Fan 等提出了大规模的用于目标跟踪的 LaSOT 数据集[28]，该数据集包含 1400 个视频序列，其中测试集包含 280 个序列，每个序列平均 2512 帧，数据集涵盖了包含 70 个类别目标的跟踪任务，每个类别包含 20 个序列。由于该视频序列都超过 1000 帧，偏重于长时跟踪且难度相较于其他数据集较大。

表 1　常见数据集的统计

Tab. 1　Statistics of datasets

数据集	视频数	平均帧数	总帧数
OTB50	51	578	$2.9*10^4$
OTB100	100	590	$5.9*10^4$
VOT2017	60	357	$2.1*10^4$
UAV123	123	915	$1.13*10^5$
GOT-10K	>10000	>1500	$1.5*10^6$
TrackingNet	35242	465	$1.64*10^7$
LaSOT	1400	2506	$3.52*10^6$

2　基于深度学习的单目标跟踪方法

2012 年，以 AlexNet 为代表的深度学习方法在图像处理领域的成功应用，使目标跟踪开始迈入深度学习时代，孪生网络、强化学习、元学习等框架和理论不断地出现在单目标跟踪算法中。近几年的视觉目标跟踪比赛显示，基于深度学习的跟踪器取得了非常好的成绩[2224]。本章按照核心方法的不同类别对基于深度学习的单目标跟踪方法进行介绍。

2.1　相关滤波

相关滤波(CF，correlation filter)是一种判别式机器学习方法，其原理是两个相关信号的卷积响应大于不相关信号的卷积响应。在目标跟踪中，使用目标模板训练的滤波器对视频帧进行滤波处理，在得到的响应图中寻找最大值作为当前帧的目标位置。因此，可以将目标跟踪的过程转换为对搜索区域图像进行相关滤波的过程，寻找目标位置也就是寻找滤波器响应图的最大值位置。其一般框架如图 9。

早期相关滤波算法使用手工特征表示目标，如采用灰度特征的 CSK[29]，采用 HOG 特征的 KCF[30]，这些方法速度很快但是鲁棒性不强，究其原因是手工特征难以适应目标的各

种变化，另一方面，卷积神经网络提取的特征具有良好的抗干扰能力和目标表征能力，因此，目标跟踪领域中出现了使用卷积神经网络提取的特征结合相关滤波方法的研究。

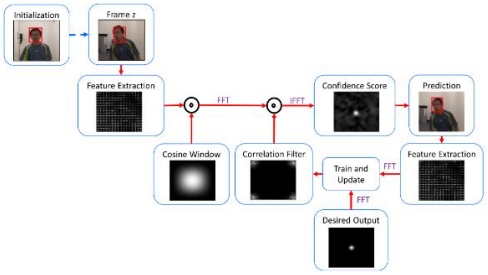

图 9　相关滤波的算法框架

Fig. 9　Algorithm framework of correlation filter

基于上述思想，Ma 等提出了 HCFT[31]算法。HCFT 利用卷积神经网络浅层表征外观，深层表征语义的特点，将不同深度卷积层提取的特征进行滤波并加权合并，由粗到细推断出目标位置。考虑到 HCFT 对尺度变化的鲁棒性较差，作者针对这一问题在其基础上进行改进[32]，使用 EdgeBox[33]算法生成两种区域估计，一是在目标周围以较小步长提供多尺度估计，二是在整个图像中以较大步长提供检测区域估计，获得了较好的鲁棒性。同样基于使用多层卷积神经网络的思路，Danelljan 等在利用空间正则化解决边界效应问题的相关滤波算法 SRDCF[34]的基础上，提出了 CCOT[13]。CCOT 将多层特征图进行三次插值，然后对每个特征通道分别训练一个滤波器，融合滤波响应以估计目标位置。CCOT 具有不错的准确度，但是较高的通道数导致参数爆炸，实时性较差。为解决速度问题，Danelljan 等从滤波器选择、样本集和模型更新策略三个角度对 CCOT 改进，提出了 ECO[15]算法，不仅达到了实时的速度，准确度也有所提升。文献[35]提出了一种筛选特征通道的方法，根据两个连续帧的空间辨识度和时间稳定性计算每个特征通道的友好度，并将较高友好度的通道送入相关滤波器中进行训练和跟踪，可以提高跟踪的准确性。

虽然深度学习朝着更深更复杂的方向发展，但是目标跟踪没有像图像分类领域一样从中受益，如 ECO 算法使用 VGGNet，GoogLeNet 和 ResNet 作为特征提取器得到的性能几乎没有差别。针对这个问题，Bhat 等尝试将 CNN 提取的特征按层次进行不同的处理，提出了 UPDT[36]算法。UPDT 对深层特征使用数据增广扩充训练集，对浅层特征使用不同的高斯标签函数。这种方法增强了算法的准确度和鲁棒性。但数据增广的采用增加了特征提取的负担，减慢了算法的运行速度。

同样受到 SRDCF 启发的还有 19 年的 ASRCF[37]算法。与 SRDCF 类似，ASRCF 加入了自适应空间正则化项以解决边界效应和噪声。另一方面，ASRCF 采用 HOG 特征和卷积特征结合的思路以提升速度。ASRCF 虽然在性能方面稍弱于 UPDT，但实现了 28fps 的实时性速度。

2.2　孪生网络

不同于 VGGNet 等神经网络，孪生网络包含两个平行的输入，并将这两个输入连接起来产生一个输出，以确定两个输入网络间是否含有相同信息，其最早被用来实现签名验证[38]，如图 10。与相关滤波类似，孪生网络也是衡量两个输入的相似性，借由这个概念，孪生网络架构被用来衡量目标模板和待测物体间的相似度以确定目标及其位置。

2016 年，Tao 等将孪生网络应用在目标跟踪上，提出了 SINT[39]算法，开创性的将目标跟踪问题转换为一个图像 patch 块匹配问题。同年，Bertinetto 等提出 SiamFC[40]。两个算法凭借优异的准确性和速度使得孪生网络方法在以相关滤波为主的目标跟踪算法中脱颖而出，成为目标跟踪领域算法的又一热点。

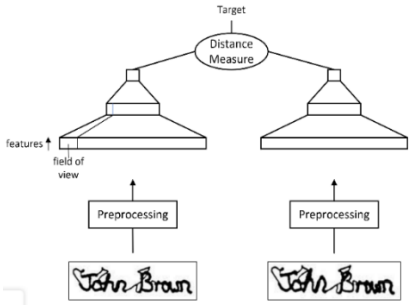

图 10　使用孪生网络结构实现签名验证

Fig. 10　Achieve signature verification byiamese network

如图 11，SiamFC 将模板和当前帧送入同一个嵌入网络 φ(embedding network)， 得到两个 embedding 并进行互相关运算，响应最大的位置即为目标位置。SiamFC 将跟踪任务作为模板匹配问题解决，无须更新模板，高速的同时保证了一定的准确度。

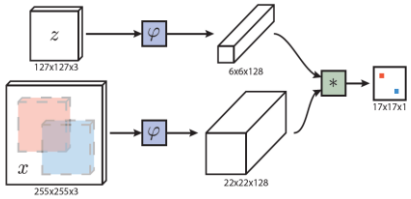

图 11　SiamFC 的网络结构

Fig. 11　Network of siamfc

另一方面，SiamFC 使用五种固定尺度来适应目标的变化，但仍不能提供足够的鲁棒性，同时，目标检测领域有较为优秀的解决尺度问题的方法，因此有学者尝试融合孪生网络和目标检测技术。文献[42]提出 SiamRPN 算法，将目标跟踪作为单样本检测任务进行处理。SiamRPN 借鉴了区域推荐网络 RPN，在孪生网络提取特征后，将特征图送入分类分支和回归分支，使跟踪器可以回归目标位置、形状，在保证实时性的同时，RPN 的 anchor 机制有效的降低了目标尺度变化对跟踪结果的影响。但 SiamRPN 对于模型的泛化性能较弱，在目标丢失的时候仍有较高的响应，于是作者尝试构建难例负样本，并且加入数据增广生成的正样本，提出 DaSiamRPN[43]算法。相比于原算法，DaSiamRPN 增强了分类器的判别能力，提高了模型的泛化性能。为了进一步提升 SiamRPN 的性能，Li 等尝试将骨干网络由 AlexNet 替换为较深的网络 ResNet，发现由于 ResNet 不具备严格平移不变性，会使得孪生网络学习到位置偏见，于是尝试使用均匀分布的采样方式消除了上述影响，提出了 SiamRPN++[44]，同时在特征提取网络后加入了多层融合和深度互相关机制，提升准确度，降低参数量。文献[45]在 SiamRPN 的基础上，设计了一个多 RPN 级联模块，由深到浅融合特征层并分别送入三级 RPN 中，上级 RPN 的分类和回归结果作为下级 RPN 的训练样本，使网络可以逐渐减小相似目标在语义层的干扰，增强了算法的鲁棒性，同时多次回归也有助于提高准确性。

SiamFC 的目标搜索区域较大，因此引入了过多的背景噪声。为了解决这一问题，文献[41]引入了像素全局匹配模块 PGNet，其整体网络结构与级联 RPN 类似，即将不同深度的特征层分别送入模块生成多个分类和回归结果并进行融合，经过测试，PGNet 可以有效的降低背景干扰。

上述算法具有不错的性能和较好的速度，但是目标模板固定容易引起误差积累，导致模型无法适应目标外观的剧烈变化，其鲁棒性存在瓶颈，因此引入在线更新机制是必要的。2017 年，Valmadre 等尝试在孪生网络结构中加入相关滤波层以实现在线更新功能，提出了 CFNet[50]算法。CFNet 推导了

相关滤波的前向和反向传播公式，使相关滤波可以嵌入到预训练网络中实现端到端学习。同样采用端到端学习框架的DCFNet[51]在网络结构上与 CFNet 有所不同，其在 SiamFC 的基础上将互相关运算替换为相关滤波层以实现在线更新，提高了算法的鲁棒性。文献[52]在孪生网络中加入了模型预测器，并引入 Hinge-like 函数和最小均方差函数提高损失函数的判别能力，具有在线更新功能且性能较强。之后作者基于前述算法，提出了一种基于概率的回归方法，即 PrDiMP[53]算法。PrDiMP 可以预测目标状态的条件概率密度，并使用最小化 KL 散度训练回归网络，提高了算法的性能。Voigtlaender 等从难例挖掘和运动轨迹动态规划两个方向，设计了利用初始帧和前一帧为模板的孪生网络检测结构，提出了 SiamRCNN 算法[54]。SiamRCNN 检测所有潜在目标，对检测结果进行跟踪轨迹动态规划。使用基于动态编程的评分算法为所有轨迹进行评估，选取分值最高的轨迹作为目标轨迹，实现了一个较为鲁棒的算法。

基于 anchor 的 RPN 类方法显著提高了跟踪的精度，但 anchor 的设定需要较强的先验知识，且绝大部分 anchor 都作为负样本被学习，因此该类方法在目标尺度和相似目标的处理上存在局限性。文献[46]受目标检测中 anchor-free 机制的启发，与 RPN 类方法同样将跟踪任务分解为分类和状态估计，但摈弃了预先设定的 anchor，提出采用自适应边界框的 SiamBAN 方法，不仅降低参数提升了速度，还具有更高的自由度和通用性。文献[47]与 SiamBAN 的思路相似，其创新在于分类分支增加了中心度(Center-ness)计算，通过惩罚离目标较远的像素可以更好的定位目标中心以提高准确性。基于目标感知的 anchor-free 方法 Ocean[48]针对分类和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。文献[49]将分类分支和回归分支所使用的特征图进行区分，并在分类分支中加入 PSS 评分，可以提高跟踪的准确性。

2.3　元学习

视觉目标跟踪作为一种少样本任务，目标样本不足的问题使得获取正负训练样本难度较大，对于判别式算法来说，会使分类器易于过度拟合并失去泛化能力。许多研究人员尝试基于目标模板扩充训练样本，如 MOSSE 所使用的仿射变换，CSK 的循环采样，UPDT 的数据增广等技术。这些策略或者带来了额外的计算，降低了速度，或者泛化能力不足。基于以上原因，如何在少样本的跟踪器中初始化模型使得模型可以学习到实例的信息而且不会过拟合成为了一个难题。近年来提出的元学习理论可以较好的解决这个问题。元学习可以学习网络参数，自适应的优化目标跟踪模型，其在面对样本不足的问题上有很好的泛化能力。

17 年，Jonghong 等首次将元学习理论引入目标跟踪领域，提出 MLT[55]算法。MLT 采用孪生网络和元学习网络耦合的方式，孪生网络为元学习网络提供训练样本关于损失函数的平均负梯度信息，元学习网络以卷积核和通道注意力反馈目标的特定信息，结合一般特征和目标特定特征，最终得到的响应图可以更加突出目标。文献[56]提出了一种提升跟踪器性能的元学习方法。该方法在初始帧学习模型参数 θ 和梯度下降的动量参数 α，根据两个参数对之后的帧进行评估并更新参数，最后利用 ADAM 算法得到最优的 θ 和 α。文献[57]基于相关滤波跟踪器，通过引入元 Q 学习，实现在不同跟踪场景下网络学习率的动态调整，进而避免在目标发生遮挡时跟踪器学习到错误的目标特征，增强了跟踪器的鲁棒性。

20 年，Wang 等提出了一种基于模型不可知元学习的跟踪器 MAML-Tracker[58]。该方法将跟踪任务作为单实例检测进行处理，给定合适的初始化模型，检测器从单帧中快速学习到实例信息，从而实现跟踪功能。MAML-Tracker 采用双层优化策略，将训练样本分成目标集和支撑集。如图 12，跟踪器尝试找到一组参数使得在固定迭代次数后在目标集上检测器的误差最小，同时生成的参数泛化性能较好。经过测试，算法具有良好的性能和速度，体现了目标检测与少样本学习结合的优秀表现。

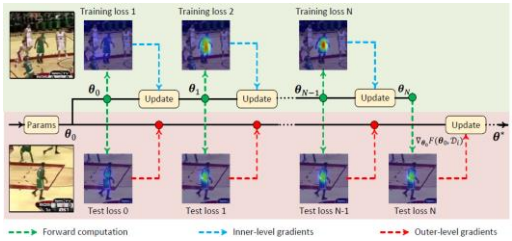

图 12　MAML-Tracker 的参数迭代过程

Fig. 12　Parameters iteration of MAML-Tracker

2.4　注意力

注意力机制的基本思想就是想让跟踪器学会“注意”目标区域，能够忽略无关信息而关注重点信息，从而节省资源，快速获得最有效的信息。其本质是一种通过网络自主学习出的一组权重系数，并以“动态加权”的方式来强调本文所感兴趣的区域同时抑制不相关背景区域的机制。在神经网络中，注意力模块通常是一个额外的神经网络，针对不同的输入学习一个权重。在目标跟踪领域，采用注意力机制可以有效地快速突出目标，减少背景信息的干扰。

17 年，Choi 等将相关滤波与注意力机制相结合，提出了 ACFN[59]算法。ACFN 包含 260 个相关滤波模块，利用注意力网络选择最优相关滤波模块进行跟踪的方法，其可以自适应的选择多种目标特征，减少了计算量，提高了跟踪器的鲁棒性。为了利用目标丰富的运动信息，Zhu 等将光流信息和深度特征与注意力机制相融合，提出了 FlowTrack 方法[60]。FlowTrack 提取历史帧的光流信息和深度特征，送入通道注意力和空间注意力模块，得到目标的运动信息，经过互相关操作计算出目标的位置。FlowTrack 以 12fps 的速度获得了不错的性能。文献[61]对模板和目标分别提取语义和外观特征并进行融合，使两种特征可以互相补充，值得注意的是算法在处理初始帧时加入了通道注意力以增强语义分支的辨别力，最终算法表现良好。Zhou 等在孪生网络外并行的加入了分类分支以提高算法的鲁棒性[62]。该分支以多个帧作为训练集并不断更新，提供更准确的分类结果。同时，作者在特征提取器中引入双重注意机制来提取特定特征，其核心在于特征通道压缩后进行空间和通道注意力特征融合。两个策略使算法对基准跟踪器有较大的性能提升。

注意力机制可以较好的降低目标变化产生的影响，因此有学者尝试以注意力机制代替在线更新。文献[66]将一般注意力，具有残差结构的空间注意力和通道注意力与孪生网络融合，提出 RASNet 方法。RASNet 不仅能够减小深度网络训练中的过拟合问题，还能够提高网络的判别能力和自适应力，多注意力机制代替了在线更新，使得 RASNet 可以很好的适应模型。文献[67]则尝试使用注意力机制提升回归框估计的准确度，其将模板的每个像素与 ROI 计算相似性，然后引入空间注意力强化角点对应区域的特征，另外使用相关性引导的通道注意力强化对特定类别目标相应高的通道，弱化无关通道。Yu 等发现孪生网络独立计算搜索和模板分支的注意力时，会降低算法性能，因此提出了可变孪生注意力网络 SiamAttn[63]。SiamAttn 通过互注意力机制计算模板和搜索区域之间的依赖关系，以隐式的更新模板特征，可变性自注意

力机制则提升了孪生网络跟踪器的特征表达能力，提高了算法的鲁棒性。

最近，由注意力模块组成的 transformer 网络在引入计算机视觉领域后可以实现性能堪比 CNN 的优秀成果，这引起了研究人员的兴趣。文献[64]尝试将 transformer 网络的编码器放入孪生网络的模板分支，解码器放入搜索分支，其中编码器用以加强模板特征，解码器连接两个分支以传播时间上下文，算法以 DiMP 为基准跟踪器，取得了不错的表现。Chen 等基于 transformer 设计了特征融合网络[65]，即孪生网络的两个分支特征进行自注意力增强后使用互注意力融合，这一过程重复 N 次，之后将融合的特征向量送入分类和回归模块得到最终的目标位置，算法表现良好，且速度较快。

2.5　循环神经网络

基于卷积神经网络的目标跟踪方法已经取得了很多成果，但很多方法只关注目标的外观和语义特征，忽略了视频本身的时序信息；只关注目标附近的区域，忽视了对目标周边区域的上下文信息进行建模，时间连续性和空间信息建模方面还有待进一步改善。近年来，循环神经网络 RNN 在时序任务上显示出了突出的性能，因此不少研究者开始探索如何应用循环神经网络来解决现有跟踪任务中存在的问题。

在 16 年，Cui 等将 CF 和 RNN 结合，提出 RTT[68]算法。RTT 对候选区域进行分块，使用多向 RNN 对四个角度的所有分块进行编码，尝试识别有效的部分，从 RNN 产生的置信度用于训练 CF，这提高了 RTT 的鲁棒性，降低了目标被遮挡时对算法的干扰。17 年，Fan 等通过结合 CNN 和 RNN，提出了 SANet[69]算法。SANet 使用 CNN 区别目标和背景，使用 RNN 和有向无环图对目标的自身结构信息建模，以此增强目标和相似物体之间的区别性。由于 CNN 不同层包含的语义不同，所以 SANet 使用多个 RNN 在不同的语义层对目标建模，以达到 CNN 和 RNN 特征融合的目的。19 年，Yun 等提出了 ROLO[70]算法，借鉴目标检测著名的 yolo 算法作为候选样本检测器，同时引入长短期记忆网络 LSTM，将候选样本的图像块放入 LSTM 中输出目标框，并利用 LSTM 保存帧间时序信息。文献[71]利用目标在每帧中的特征及其动态变化设计了一个深度时空网络 ST-LSTM。ST-LSTM 由空间 LSTM 模块和时间 LSTM 模块构成。空间 LSTM 用以捕获目标的空间信息，时间 LSTM 基于空间 LSTM 构造，跨帧获取视频目标的时间变化特性，能够记录目标的运动过程，具有不错的表现。

Liu 等为了解决遮挡问题，将重点放在预测目标在未来帧的位置上，设计了多流卷积 LSTM 网络[72]。算法利用孪生网络比较相邻帧来建立背景运动模型，在去除背景运动干扰后使用轨迹预测模型从在几个历史帧中学习目标的特征，多流卷积 LSTM 网络在这些特征中进行编解码时间演化，并在未来帧中预测目标对象的位置。相比基准跟踪器，算法在鲁棒性特别是遮挡方面具有不错的提升。

2.6　生成对抗网络

目标跟踪一直存在训练样本缺乏、正负样本数量不均衡的问题，由于训练过程每一帧的负样本数量常多于正样本，而且正样本之间的重叠率较高，因此跟踪器面对外观变化时无法保持较高的准确度。利用生成对抗网络 GAN 可以生成不同的正负样本以解决此问题。Song 等利用这一思想，在 MDNet 的基础上加入 GAN，在 18 年提出了 VITAL[73]算法。VITAL 使用生成随机生成一些蒙版，这些蒙版可以自适应的丢弃输入特征来捕获外观变化，然后使用对抗性学习使蒙版可以维持目标最鲁棒的特征，以此解决目标正样本数量少，重叠率高的问题。同年，文献[74]利用 GAN，在 SINT 的基础上，将孪生网络的输出节点链接一个编解码器，以此生成大量增强正样本，解决正样本高重叠率的问题。

2.7　算法对比

本节选取上述每个分类中具有代表性的算法，根据 EAO(expected average overlap)，A(accuracy)，R(robustness)三个指标对算法在 VOT2018 数据集中的实验结果进行对比，对比结果如表 2。

其中，EAO 指跟踪器在一个序列中无重置的重叠率期望值，A 指跟踪器在一个序列中真值与预测边界框的平均重叠率 S(参照式(1))，R 指跟踪器在一个序列中失败的帧数占比。VOT 会根据重叠率设置跟踪器状态，重叠率为 0 即设置当前帧为失败帧，失败 5 帧后会重置跟踪器，因此，EAO 和 A 虽然都是指跟踪器的重叠率，但两者的计算方法有一定的区别，EAO 更能反映跟踪器的整体性能。

$$S=\frac{|r_t\cap r_a|}{|r_t\cup r_a|} \tag{1}$$

其中：r_a 为真值，r_t 为预测边界框。

根据表 2 的对比可以看出，孪生网络方法 Ocean 具有最好的 EAO 评分，其次是增加了注意力机制的 SiamAttn，生成对抗网络 VITAL 则表现最差。基于孪生网络的方法都具有不错的准确度，特别是在孪生网络基础上增加注意力的 SiamAttn，元学习和 RNN 类方法也有较不错的表现。另一方面， UPDT 因为相关滤波器具有在线更新能力，因此获得了不错的鲁棒性，MAML、SiamAttn 等基于孪生网络的跟踪器的鲁棒性相差不大，Ocean 由于添加了在线更新方法因而鲁棒性很强，RNN 方法的鲁棒性则较弱。

表 2　VOT2018 的跟踪器对比结果

Tab. 2　Trackers comparison results of VOT2018

跟踪器	EAO ↑	A ↑	R ↓
相关滤波 UPDT[36]	0.378	0.505	0.140
孪生网络 Ocean[48]	0.489	0.592	0.117
元学习 MAML[58]	0.452	0.604	0.159
注意力 SiamAttn[63]	0.47	0.63	0.16
循环神经网络 STNT[72]	0.398	0.616	0.258
生成对抗网络 VITAL[73]	0.323	无	无

3　结束语

单目标跟踪作为计算机视觉领域的重要研究方向，有着重要的应用价值和研究意义。近年来随着深度学习的发展，出现了很多优秀的单目标跟踪算法。基于深度学习的方法在目标跟踪领域已经取得了很大的突破，其通过卷积神经网络对目标进行卷积操作所提取的深度特征与以往的手工特征相比具有强大的表征能力，其他深度学习理论如对抗生成网络、注意力机制和循环神经网络等在目标跟踪领域的成功应用，也极大的增强了单目标跟踪算法的鲁棒性和准确度。同时近年来，多个目标跟踪大规模数据集的提出，为深度学习的训练和验证提供了辅助。

目前，基于深度学习的方法已经成为单目标跟踪的主流技术，在鲁棒性和准确度方面具有一定的优势。然而，在实际场景的应用中，基于深度学习的方法仍然难以满足跟踪任务对速度和性能的需要，存在着进一步研究和发展的空间。

通过对近几年的深度学习单目标跟踪算法的归纳总结，结合算法存在的问题，本文认为单目标跟踪有如下几个发展趋势：

a) 当前，SiamFC 因其简洁高效的特点，常常作为跟踪算法研究的基础框架，而自 SiamRPN 开始研究人员则将跟踪任务拆解成分类和回归两个分支任务进行处理，分类分支区分目标前背景提高鲁棒性，回归分支提供精确的目标边界框以提高准确性，因此最近的一种趋势是针对两个分支采取

策略进行优化，比如对分类结果进行更细粒度的预测，类似 SiamMask[75]使用目标分割技术以极大提升目标边界框的精度等。当前许多目标检测的优秀设计可以应用于优化工作，目标检测领域技术的进步会给跟踪器发展带来不小的帮助。

b) 另一方面，孪生网络采用模板匹配的思路使得跟踪器面对语义相似的目标时易产生漂移，改进并使用不同的特征提取策略对算法鲁棒性的提升效果不大，目前目标跟踪的另一趋势是设计运动估计模块，采用 LSTM 等网络根据上下文建模不同帧之间像素的时间关系，估计目标运动以较为有效的减少干扰物和背景噪声。另外，深度学习机制可以用来形成运动估计的可解释性仍然未知，从这个角度来看，探索运动估计模块的机制和可解释性来增强现有的跟踪器也是可能的研究方向之一。

c) 深度学习技术的发展总会给目标跟踪带来新的研究思路，如 GAN 解决正样本过少问题，元学习增强跟踪器的泛化性能，图神经网络提升目标特征建模能力[76]等。将具有前景的新颖深度学习技术经过迁移、修改后应用到跟踪器中，可以期待其取得良好的表现，为目标跟踪提供新的研究方向。

总的来说，深度学习在单目标跟踪领域表现出巨大的潜力，但是由于各种影响因素的存在，当前大部分的单目标跟踪算法仍然很难达到理想的跟踪效果，存在很多需要探索完善的地方，因此继续开展基于深度学习的单目标跟踪的研究工作具有十分重要的理论意义和研究价值。

参考文献：

[1] Foresti G L, Snidaro L. Vehicle detection and tracking for traffic monitoring [C]// Image Analysis and Processing. Berlin: Springer Berlin Heidelberg, 2005: 1198-1205.

[2] Ferraz P A P, de Oliveira B A G, Ferreira F M F, *et al.* Three-stage RGBD architecture for vehicle and pedestrian detection using convolutional neural networks and stereo vision [J]. IET Intelligent Transport Systems, 2020, 14 (10): 1319-1327.

[3] Yang Jiachen, Zhao Weirong, Han Yurong, *et al.* Aircraft tracking based on fully conventional network and kalman filter [J]. IET Image Processing, 2019, 13 (8): 1259-1265.

[4] Lim K M, Tan A W C, Lee C P, *et al.* Isolated sign language recognition using convolutional neural network hand modelling and hand energy image [J]. Multimedia Tools and Applications, 2019, 78 (14): 19917-19944.

[5] Ganapathi V, Plagemann C, Koller D, *et al.* Real time motion capture using a single time-of-flight camera [C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010: 755-762.

[6] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision [C]// Proc of the 7th international joint conference on Artificial intelligence. 1981: 674-679.

[7] Yilmaz A, Javed O, Shah M. Object tracking: a survey [J]. ACM Computing Surveys, 2006, 38 (4): 13.

[8] Li Xi, Hu Weiming, Shen Chunhua, *et al.* A survey of appearance models in visual object tracking [J]. ACM Trans on Intelligent System Technology, 2013, 4 (4): 58.

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

[10] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4293-4302.

[11] Wang Naiyan, Shi Jianping, Yeung D, *et al.* Understanding and diagnosing visual tracking systems [C]// IEEE International Conference on Computer Vision. 2015: 3101-3109.

[12] Wu Yi, Lim J, Yang M. Object tracking benchmark [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2015, 37 (9): 1834-1848.

[13] 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, 9914: 472-488.

[14] Kristan M, Leonardis A, Matas J, *et al.* The visual object tracking vot2016 challenge results [J]. Lecture Notes in Computer Science, 2016, 9914: 777-823.

[15] Danelljan M, Bhat G, Khan F S, *et al.* ECO: efficient convolution operators for tracking [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017, 1: 6931-6939.

[16] Collins R, Zhou X, Teh S K. An open source tracking testbed and evaluation web site [C]// IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. 2005

[17] Wu Yi, Lim J, Yang M. Online object tracking: a benchmark [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2013: 2411-2418.

[18] Kristan M, Pflugfelder R, Leonardis A, *et al.* The visual object tracking vot2013 challenge results [C]// IEEE International Conference on Computer Vision Workshops. 2013: 98-111.

[19] Kristan M, Pflugfelder R, Leonardis A, *et al.* The visual object tracking vot2014 challenge results [J]. Lecture Notes in Computer Science, 2015.

[20] Kristan M, Pflugfelder R, Leonardis A, *et al.* The visual object tracking vot2015 challenge results [C]// IEEE International Conference on Computer Vision Workshop. 2015: 564-586.

[21] Kristan M, Leonardis A, Matas J, *et al.* The visual object tracking vot2017 challenge results [C]// IEEE International Conference on Computer Vision Workshops. 2017: 1949-1972.

[22] Kristan M, Leonardis A, Matas J, *et al.* The sixth visual object tracking vot2018 challenge results [J]. Lecture Notes in Computer Science. 2018.

[23] Kristan M, Matas J, Leonardis A, *et al.* The seventh visual object tracking vot2019 challenge results [C]// IEEE/CVF International Conference on Computer Vision Workshop. 2019: 2206-2241.

[24] Kristan M, Leonardis A, Matas J, *et al.* The eighth visual object tracking vot2020 challenge results [J]. Lecture Notes in Computer Science. 2020.

[25] Mueller M, Smith N, Ghanem B. A benchmark and simulator for uav tracking [J]. Lecture Notes in Computer Science, 2016.

[26] Mueller M, Bibi A, Giancola S, *et al.* TrackingNet: a large-scale dataset and benchmark for object tracking in the wild [J]. Lecture Notes in Computer Science, 2018.

[27] Huang Lianghua, Zhao Xin, Huang Kaiqi. GOT-10k: a large high-diversity benchmark for generic object tracking in the wild [C]// IEEE Trans on Pattern Analysis and Machine Intelligence. 2019: 1.

[28] Fan Heng, Lin Liting, Yang Fan, *et al.* LaSOT: a high-quality large-scale single object tracking benchmark [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5369-6378.

[29] Henriques J F, Caseiro R, Martins P, *et al.* Exploiting the circulant structure of tracking-by-detection with kernels [C]// European Conference on Computer Vision. Berlin: Springer Berlin Heidelberg, 2012: 702-715.

[30] Henriques J F, Caseiro R, Martins P, *et al.* High-Speed tracking with kernelized correlation filters [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2014, 37 (3): 583-596.

[31] Ma Chao, Huang Jiabin, Yang Xiaokang, *et al.* Hierarchical convolutional features for visual tracking [C]// IEEE International Conference on Computer Vision. 2015: 3074-3082.

[32] Ma Chao, Huang Jiabin, Yang Xiaokang, *et al.* Robust visual tracking via

hierarchical convolutional features [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2019, 41 (11): 2709-2723.

[33] Zitnick C L, Dollár P. Edge boxes: locating object proposals from edges [J]. Lecture Notes in Computer Science, 2014.

[34] Danelljan M, Häger G, Khan F S, *et al.* Learning spatially regularized correlation filters for visual tracking [C]// IEEE International Conference on Computer Vision. 2015: 4310-4318.

[35] Ge Shiming, Luo ZHao, Zhang Chunhui, *et al.* Distilling Channels for Efficient Deep Tracking [J]. IEEE Trans on Image Processing. 2020, 29: 2610-2621.

[36] Bhat G, Johnander J, Danelljan M, *et al.* Unveiling the power of deep tracking [J]. Lecture Notes in Computer Science, 2018: 493-509.

[37] Dai Kenan, Wang Dong, Lu Huchuan, *et al.* Visual tracking via adaptive spatially-regularized correlation filters [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4665-4674.

[38] Bromley J, Guyon I, Lecun Y, *et al.* Signature verification using a"Siamese"time delay neural network [C]// Proceedings of the 6th International Conference on Neural Information Processing Systems. Denver: Morgan Kaufmann Publishers Inc. , 1993: 737-744.

[39] Tao Ran, Gavves E, Smeulders A. Siamese instance search for tracking [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1420-1429.

[40] Bertinetto L, Valmadre J, Henriques J F, *et al.* Fully-Convolutional siamese networks for object tracking [J]. Lecture Notes in Computer Science, 2016.

[41] Liao Bingyan, Wang Chenye, Wang Yayun, *et al.* PG-Net: Pixel to Global Matching Network for Visual Tracking [C]// Springer International Publishing, 2020: 429-444.

[42] Li Bo, Yan Junjie, Wu Wei, *et al.* High performance visual tracking with siamese region proposal network [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8971-8980.

[43] Zha Yufei, Wu Min, Qiu Zhuling, *et al.* Distractor-Aware visual tracking by online siamese network [J]. IEEE Access. 2019, 7: 89777-89788.

[44] Li Bo, Wang Qiang, Zhang Fangyi, *et al.* SiamRPN+: evolution of siamese visual tracking with very deep networks [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4277-4286.

[45] Fan Heng, Ling Haibin. Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7944-7953.

[46] Chen Zedu, Zhong Bineng, Li Guorong, *et al.* Siamese box adaptive network for visual tracking [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020, 1: 6667-6676.

[47] Guo Dongyan, Wang Jun, Cui Ying, *et al.* SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking [C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6268-6276.

[48] Zhang Zhipeng, Peng Huowen, Fu Jianlong, *et al.* Ocean: Object-Aware Anchor-Free Tracking [C]// Computer Vision – ECCV 2020. 2020: 771-787.

[49] Xu Yinda, Wang Zeyu, Li Zuoxin, *et al.* SiamFC+: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines [J]. Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34 (07): 12549-12556.

[50] Valmadre J, Bertinetto L, Henriques J, *et al.* End-to-End representation learning for correlation filter based tracking [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017, 1: 5000-5008.

[51] Wang Qiang, Gao Jin, Xing Junliang, *et al.* DCFNet: discriminant correlation filters network for visual tracking [J/OL]. Arxiv, 2017. (2017-04-13) [2021-03-16]. https://arxiv. org/abs/1704. 04057.

[52] Bhat G, Danelljan M, Van Gool L, *et al.* Learning discriminative model prediction for tracking [C]// IEEE/CVF International Conference on Computer Vision. 2019, 1: 6181-6190.

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

[54] Voigtlaender P, Jonathon L, Torr H S P, *et al.* Siam R-CNN: visual tracking by re-detection [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6577-6587.

[55] Choi J, Kwon J, Lee M K. Deep meta learning for real-time target-aware visual tracking [C]// IEEE/CVF International Conference on Computer Vision. 2019: 911-920.

[56] Park E, Berg A C. Meta-tracker: Fast and robust online adaptation for visual object trackers [C]// Proc of the European Conference on Computer Vision, 2018: 587-604.

[57] Kubo A, Meshgi K, Ishii S. A Meta-Q-Learning Approach to Discriminative Correlation Filter based Visual Tracking [J]. Journal of Intelligent and Robotic Systems. 2021, 101 (1): Article 11.

[58] Wang Guangting, Luo Chong, Sun Xiaoyan, *et al.* Tracking by instance detection: a meta-learning approach [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 6287-6296.

[59] Choi, J, Chang H J, Yun S, *et al.* Attentional correlation filter network for adaptive visual tracking [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4828-4837.

[60] Zhu Zheng, Wu Wei, Zou Wei, *et al.* End-to-End flow correlation tracking with spatial-temporal attention [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 548-557.

[61] He Anfeng, Luo Chong, Tian Xinmei, *et al.* A Twofold Siamese Network for Real-Time Object Tracking [C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 4834-4843.

[62] Zhou Jinhao, Wang Peng, Sun Haoyang. Discriminative and Robust Online Learning for Siamese Visual Tracking [J]. Proc of the AAAI Conference on Artificial Intelligence. 2020, 34 (07): 13017-13024.

[63] Yu Yuechen, Xiong Yilei, Huang Weilin, *et al.* Deformable siamese attention networks for visual object tracking [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6727-6736.

[64] Wang Ning, Zhou Wengang, Wang Jie, *et al.* Transformer Meets Tracker-Exploiting Temporal Context for Robust Visual Tracking [C]// The IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.

[65] Chen Xin, Yan Bin, Zhu Jiawen. Transformer Tracking [J/OL]. Arxiv, 2021. (2021-03-29) [2021-04-10]. https://arxiv. org/abs/2103. 15436.

[66] Wang Qiang, Teng Zhu, Xin Junliang, *et al.* Learning attentions: residual attentional siamese network for high performance online visual tracking [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 4854-4863.

[67] Du Fei, Liu Peng, Zhao Wei, *et al.* Correlation-Guided attention for corner detection based visual tracking [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6834-6844.

[68] Cui Zhen, Xiao Shengtao, Feng Jiashi, *et al.* Recurrently target-attending tracking [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1449-1458.

[69] Fan Heng, Ling Haibin. SANet: Structure-Aware network for visual tracking [C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 2217-2224.

[70] Ning Guanghan, Zhang Zhi, Huang Chen, *et al.* Spatially supervised

recurrent convolutional neural networks for visual object tracking [C]// 2017 IEEE International Symposium on Circuits and Systems. 2017: 1-4.

[71] Teng Zhu, Xing Junliang, Wang Qiang, *et al.* Deep Spatial and Temporal Network for Robust Visual Object Tracking [J]. IEEE Trans on Image Processing. 2020, 29: 1762-1775.

[72] Liu Yuan, Li Ruoteng, Cheng Yu, *et al.* Object Tracking Using Spatio-Temporal Networks for Future Prediction Location [C]// Computer Vision – ECCV 2020. 2020: 1-17.

[73] Song Yibing, Ma Chao, Wu Xiaohe, *et al.* VITAL: visual tracking via adversarial learning [C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018, 1: 8990-8999.

[74] Wang Xiao, Li Chenglong, Luo Bin, *et al.* SINT+: robust visual tracking via adversarial positive instance generation [C]// Proc of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4864-4873.

[75] Wang Qiang, Zhang Li, Bertinetto L, *et al.* Fast Online Object Tracking and Segmentation: A Unifying Approach [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1328-1338.

[76] Gao Junyu, Zhang Tianzhu, Xu Changsheng. Graph Convolutional Tracking [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4644-4654.