动态综述

DOI：10.16818/j.issn1001－5868.2020.06.001

基于深度学习的目标跟踪技术的研究综述

罗　元*，肖　航，欧俊雄

（重庆邮电大学 光电工程学院，重庆 400065）

摘　要： 深度学习技术在计算机视觉领域的应用日趋广泛，基于深度学习的目标跟踪技术是当前计算机视觉领域的一项重要研究课题。文章首先对国内外目标跟踪技术的发展历程进行了回顾和梳理，详细介绍并对比了部分典型的基于深度学习的目标跟踪方法；然后介绍了几种目标跟踪领域的典型数据集；最后对基于深度学习的目标跟踪技术的研究进行了总结，并对未来的发展方向进行了展望。

关键词： 深度学习；计算机视觉；目标跟踪

中图分类号： TP18；TP391.41　**文章编号：** 1001－5868(2020)06－0757－11

Research on Target Tracking Technology Based on Deep Learning

LUO Yuan，XIAO Hang，OU Junxiong

（College of Optoelectronic Engineering，Chongqing University of Posts and Telecommunications，Chongqing 400065，CHN）

Abstract： The application of deep learning technology in the field of computer vision is becoming more and more extensive，and the target tracking technology based on deep learning is an important research topic in the field of computer vision. In this paper，firstly，the development process of target tracking technology at home and abroad is reviewed，and some typical target tracking methods based on deep learning are introduced in detail and compared；then，several typical data sets in the field of target tracking are summarized；finally，the research progress of the target tracking technology are summarized，and the future development directions are prospected.

Key words： deep learning；computer vision；target tracking

0　引言

目标跟踪是计算机视觉领域中具有重要研究意义的课题之一，该课题与目标检测有相似之处，但主要研究内容与研究方法存在一定差异。文献[1]中指出，目标检测是对于给定的任意图像或视频，判定是否存在来自预定义类别的语义对象的任何实例，如果存在，则输出实例的空间位置和所占范围。该过程需要将图像或视频中的目标对象与其他无关的信息进行区分并完成对目标对象的标注，而目标跟踪不仅需要在视频图像序列中标注出目标对象所在的空间位置，还需要将视频连续帧中标注的目标对象连接起来得到目标对象的运动轨迹路径。

近年来，随着计算机性能的不断提高与深度学习技术的不断发展，基于深度学习的目标跟踪技术逐渐成为了研究热点，在取得突破性进展的同时被广泛应用于模式识别[2]、无人驾驶[3]、人机交互[4]、智能机器人[5]和智慧医疗[6]等领域。目前，基于深度学习的目标跟踪技术大量开源，为思想的碰撞创造了更多可能，所以本文旨在对目标跟踪技术的发展历程进行梳理，对基于深度学习的目标跟踪技术中的经典网络进行模型解构、特点剖析并进行性能对比，对目标跟踪领域应用广泛的典型数据集进行简要介绍，最后对基于深度学习的目标跟踪技术的

收稿日期：2020－06－28.

基金项目：国家自然科学基金项目(61801061)；重庆市教委科技项目(KJQN201800607).

* **通信作者：**罗　元　**E-mail：** luoyuan@cqupt.edu.cn

研究进行展望和总结。

1 传统目标跟踪

1.1 均值漂移

均值漂移由 Comaniciu 等[7-8]运用到目标跟踪领域，该方法利用图像特征构造概率密度函数，之后通过迭代沿着概率密度函数的梯度方向搜索函数的局部最大值，并以搜索到的局部最大值作为跟踪目标的位置。

基于均值漂移的目标跟踪方法具有如下优点：1)算法原理简单易实现；2)以灰度统计直方图作为图像特征，对目标旋转、轻微形变的鲁棒性较好；3)对目标中心位置采用分块加权处理，能够有效克服目标被轻微遮挡时错误的位置信息对实时跟踪的影响。但该方法仅使用灰度统计直方图表征目标，图像特征单一，易出现因目标特征不完备而被无关背景信息干扰导致跟踪失败的情况。并且，由于使用大小固定不变的核函数，所以对目标尺度大幅变化的鲁棒性较差。此外，利用概率密度函数搜索局部最大值需要经过多次迭代，导致其难以实现对高速移动目标的跟踪。

1.2 粒子滤波

粒子滤波最早由 Nummiaro 等[9]运用到目标跟踪领域，该方法采用序列蒙特卡洛的滤波方法将贝叶斯滤波方法中的积分运算转化为粒子采样，进而求得样本均值，解决了难以在实际应用中获得贝叶斯滤波概率最优解的问题。粒子滤波算法在应用时首先需要完成状态空间粒子初始化，之后每个迭代周期[10]的主要步骤包括重要性采样、权重更新和重采样，最终根据计算得到的权重估计粒子集合的状态平均值，并根据权重增加的方向实现跟踪，以此作为目标跟踪的输出结果。

基于粒子滤波的目标跟踪方法具有如下优点：1)算法简单，系统易实现；2)采用序列蒙特卡洛的滤波方法，能有效处理非高斯分布的噪声并得到目标最优位置估计；3)具有无参数估计特性，对非线性运动模型表现出优越性，适用于非线性动态系统分析。

但基于粒子滤波的目标跟踪方法的跟踪精度和跟踪速度严重依赖粒子数目且二者难以协同，粒子数越多精度越高但速度越慢，反之亦然。并且在重要性采样过程中使用重要性概率密度函数会导致粒子退化问题的出现。利用重采样解决粒子退化问题时，又易因大量重复的粒子导致粒子多样性的丧失。系统噪声越小，多样性丧失问题越明显。同时，仅以颜色直方图表征目标，目标特征表达单一，易受无关背景或环境信息的影响而导致跟踪失败。

除上述均值漂移和粒子滤波外，传统目标跟踪方法还包括卡尔曼滤波[11-13]等，但传统方法均存在易受背景信息或环境信息干扰而导致跟踪失败，对于形状、尺度和速度等各类变化的鲁棒性亟待提高等不足，所以目前传统目标跟踪方法的使用已经逐渐减少。

2 基于相关滤波的目标跟踪

2.1 MOSSE 与 CSK

最小均方误差滤波(Minimum Output Sum of Squared Error，MOSSE)是由 Bolme 等[14]提出并最先用于目标跟踪领域的基于相关滤波的跟踪方法。该方法在视频第一帧中确定跟踪目标，通过对下一帧图像进行滤波，以滤波后最大响应值所在位置作为目标的新位置，根据新位置完成目标运动轨迹的在线更新。MOSSE 利用傅里叶变换使滤波过程在频域执行，其跟踪速度高达 600～700 f/s，明显优于同时期的 IVT[15]和 MILTrack[16]等跟踪方法。

Henriques 等[17]通过视频帧与帧之间目标移动距离有限的潜在结构信息发现数据矩阵呈环状，并依据循环矩阵理论提出一种循环结构跟踪器(Circulant Structure Tracker，CSK)。该跟踪器利用循环采样使得计算量大大减少，完美解决了密集采样的问题，又利用核函数和傅里叶变换计算相邻两帧图像的相关性，在提高跟踪效果的同时保证了较快的跟踪速度。

无论是 MOSSE 还是 CSK，采用的图像特征均为单通道灰度特征，因此在很大程度上限制了这两种跟踪方法在面对复杂背景或目标与背景颜色相近时的跟踪能力。

2.2 KCF，DCF 与 CN

为了克服利用单通道灰度特征进行目标跟踪时的不足，Henriques 等[18]在 CSK 的基础上进行改善，继而提出了核化相关滤波器(Kernelized Correlation Filter，KCF)和对偶相关滤波器(Dual Correlation Filter，DCF)。KCF 和 DCF 均使用多通道方向梯度直方图(Histogram of Oriented Gradients，HOG)特征替代单通道灰度特征，因为 HOG 特征能够很好地表征目标对象的表面纹理和轮廓形状，所以提升了跟踪器在复杂环境下对运动

模糊和光照变化的鲁棒性。KCF 与 DCF 的区别在于：KCF 使用高斯核函数，DCF 使用线性核函数。文献[18]表明，由于核函数的不同，相比于 DCF，KCF 的平均精度从 72.8%提高到 73.2%，速度从 292 f/s 下降到 172 f/s。

除了 KCF 和 DCF，由 Danelljan 等[19]提出的颜色命名(Color Name，CN)算法也是基于 CSK 进行改进的。CN 采用多通道颜色特征增强了对目标对象的表征能力，提高了跟踪器对光照变化和颜色失真的鲁棒性，同时利用主成分分析方法对特征进行降维，提高了算法的运行速度，此外还引入贝叶斯公式对前景和背景进行区分，降低了复杂背景对目标跟踪的干扰程度。

KCF，DCF 和 CN 通过对图像特征选取的改进提高了跟踪器面对复杂背景时的处理能力，但对于目标本身尺度变化导致的跟踪漂移问题的处理能力仍然有待提高。

2.3　DSST，FDSST 与 SAMF

以进一步提高目标跟踪算法对目标本身尺度变化的鲁棒性为目标，Danelljan 等[20]在 KCF 的基础上引入尺度估计，提出了判辨尺度空间跟踪器(Discriminative Scale Space Track，DSST)。DSST 将目标跟踪视为位置变化和尺度变化两个独立问题，分别训练平移滤波器和尺度滤波器。用平移滤波器确定当前帧的目标位置后，在目标所处位置提取多个尺寸不同的像素块并使用尺度滤波器完成尺度估计，以最大响应值对应的尺度作为当前帧的最优目标尺度。DSST 的分步策略不仅在特征选择上保持高度的灵活性，而且尺度估计模块可以迁移到其他算法，具有良好的推广性，但仍存在计算复杂度高、计算量大的问题。

针对以上问题，Danelljan 等[21]继续提出快速判辨尺度空间跟踪器(Fast Discriminative Scale Space Track，FDSST)。FDSST 使用降维策略和 QR 分解的方法充分减少了算法计算量，文献[21]显示，FDSST 相比于 DSST，平均速度提升超一倍，由 25.4 f/s 提高到 54.3 f/s，并且其他性能也得到一定改善，如：平均距离精度(Mean Distance Precision，Mean DP)由 75.7%提高到 80.2%，平均重叠精度(Mean Overlap Precision，Mean OP)由 67.7%提高到 74.3%。

另一种基于 KCF 改进的解决尺度变化问题的跟踪器是 Li 等[22]提出的多特征集成尺度自适应跟踪器(Scale Adaptive with Multiple Feature Tracker，SAMF)。SAMF 融合了灰度、CN 和 HOG 特征来提高算法性能，同时引入尺度池，对初始目标进行七个不同尺度的缩放，通过相关滤波得到对应的响应值，以最大响应值对应的尺度作为跟踪目标的最优尺度，实现尺度自适应。与 DSST 相比，SAMF 更像是全局寻优的过程，所以响应最高的全局最优在具体到每一帧的时候并不能保证是局部最优，但 SAMF 只需要单个滤波器，无需为尺度估计进行额外的滤波器训练。

除了对图像特征选取的改进和对目标尺度变化的改进，后续还出现了对边界效应问题改进的相关研究，比如 Danelljan 等[23]提出的空间正则相关滤波器(Spatially Regularized Correlation Filter，SRDCF)、Kiani 等[24]提出的有限边界相关滤波器(Correlation Filter with Limited Boundaries，CFLB)和 Galoogahi 等[25]提出的背景感知相关滤波器(Background-Aware Correlation Filter，BACF)等，都对目标跟踪时出现的边界效应问题提供了较好的解决方法。此外，由于近年来深度学习技术的兴起和飞速发展，目标跟踪领域已经出现了相关滤波算法和深度学习相结合，用深度特征替代传统的手工设计的特征的研究，这类研究进一步提高了基于相关滤波的目标跟踪算法的跟踪性能。

3　基于深度网络的目标跟踪

基于深度网络的目标跟踪算法可以分为两类：一类是将深度特征与相关滤波相结合的跟踪方法，如将 HOG 特征替换为利用 VGG-net[26]提取的深度特征的 HCF[27]，HCFT[28]和 HDT[29]，在 SRDCF 基础上改进的 DeepSRDCF[30]和 DMSRDCF[31]，运用传统特征和深度特征相融合的 C-COT[32]和 ECO[33]等；另一类是构造端到端的深度网络的目标跟踪方法，如通过多域学习的 MDnet[34]，融合卷积神经网络(Convolutional Neural Network，CNN)特征和循环神经网络(Recurrent Neural Network，RNN)特征的 SAnet[35]，以及运用多个 CNN 构成树结构的 TCNN[36]等。

3.1　DeepSRDCF 与 DMSRDCF

DeepSRDCF(Deep Features for SRDCF)是由 Danelljan 等[30]在 SRDCF 的基础上引入深度特征提出的。DeepSRDCF 利用 VGG-net 提取深度特征，并探索了不同卷积层的深度特征对跟踪效果的

不同影响，发现使用 VGG-net 的第一层卷积所得特征的效果优于使用深层特征的效果。同时，浅层特征的提取速度更快，保证了算法的运行速度。证明了浅层卷积层比深层卷积层更适用于目标跟踪。

Gladh 等[31]将传统手工特征与深度运动特征融合于 SRDCF 框架，提出了 DMSRDCF（Deep Motion Features for SRDCF）。由于深度运动特征与目标物体的表观特征充分互补，对运动目标的表征更加精确，从而显著提高了算法的跟踪性能。随后，Danelljan 等进一步扩展对 DMSRDCF 的研究工作，并在文献[37]中详细阐述了深度运动特征的适用条件和对目标跟踪的具体影响。

表 1 对比了 SRDCF，DeepSRDCF 和 DMSRDCF 三者在不同数据集中测评的各项性能指标。

表 1 SRDCF，DeepSRDCF 和 DMSRDCF 性能对比

指标	SRDCF	DeepSRDCF	DMSRDCF
Mean OP (OTB-2015)	72.9%	77.3%	84.1%
Mean OP (Temple-color)	62.2%	65.4%	71.2%
鲁棒性 (VOT-2015)	1.24	1.05	0.92
准确度 (VOT-2015)	0.56	0.56	0.58

由表 1 可知，利用深度特征的 DeepSRDCF 和 DMSRDCF 无论是在数据集 OTB-2015 还是在 Temple-color 上，都对 Mean OP 进行了提高。采用深度运动特征的 DMSRDCF 在数据集 VOT-2015 中的鲁棒性和准确度性能表现最佳。

3.2 C-COT 与 ECO

Danelljan 等[32]发现 DCF 方法中仅使用单一分辨率的特征图，基于此改进提出了 C-COT（Continuous Convolution Operators for Visual Tracking）。C-COT 采用多分辨率深度特征图，并通过隐式插值将不同分辨率的特征图插值到连续空间域，实现了传统手工特征与深度特征的融合，能够精准完成亚像素定位，提高跟踪精度，但是由于使用高维特征需要训练大量滤波器，使得算法复杂度高，运算速度较慢。

随后 Danelljan 等[33]从滤波器构造、样本集大小和模型更新策略三个方面对 C-COT 进行改进，继续提出了 ECO（Efficient Convolution Operators for Tracking）。ECO 使用矩阵分解的卷积方法对滤波器进行降维，然后利用混合高斯模型对具有类内相似性和类间差异性的样本组合进行表征，在降低样本数目的同时提高了样本的辨识度。最后利用稀疏型模型更新策略，间隔多帧完成一次更新，通过降低更新频率来提高效率，进一步降低算法复杂度、减少计算量。

C-COT 与 ECO 在数据集 VOT-2016 中的测评结果如表 2 所示。结果显示，ECO 在保证精度的基础上，相比于 C-COT，对 EAO（Expected Average Overlap）和鲁棒性均有所优化。

表 2 C-COT 和 ECO 性能对比

跟踪方式	EAO	鲁棒性	准确度
C-COT	0.331	0.85	0.52
ECO	0.375	0.73	0.53

文献[38]进一步开展了将深度特征运用于目标跟踪的研究，得出了深度特征能够编码高级抽象信息且不受复杂外观变化的影响，但需要大量数据且更深更复杂的网络架构对目标跟踪性能提高无益的结论。基于以上结论，提出了利用数据增强扩充训练集，将深度特征和浅层特征自适应融合的方法，完成对 ECO 的进一步优化，并在数据集 VOT-2017 中进行测评，得到结果如表 3 所示。

表 3 各类算法性能对比

跟踪方式	EAO	鲁棒性	准确度
C-COT	0.267	0.318	0.494
ECO	0.280	0.276	0.483
文献[38]	0.378	0.182	0.532

表 3 显示，利用文献[38]提出的方法对 ECO 优化后，其目标跟踪的各项性能均得到了提高。

3.3 MDnet，SAnet 与 TCNN

MDnet（Multi-Domain Network）是 Nam 等[34]基于迁移学习的思想，采用 VGG-net 的部分结构，提出的一种端到端的多域学习的深度网络，该网络由特征提取通用的共享层和多分支检测的全连接层两部分组成，具体网络结构如图 1 所示。

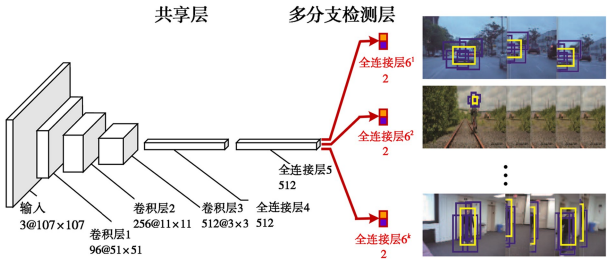


图 1 MDnet 网络结构

在 MDnet 的训练阶段，数据集中每个视频序列

通过训练得到独立的全连接层 6(fc6 层),共享层对所有视频序列通用,以便获得对不同运动目标通用性更强的深度特征;在目标跟踪阶段,保留并固定用于目标特征提取的共享层,根据要进行目标跟踪的新视频序列构建新的全连接层,二者结合组成新的端到端网络。通过新视频序列的第一帧样本更新网络,利用之后每一帧的跟踪结果生成正负样本,并在线微调 fc4～fc6 层,以实现网络的自适应调整,有效区分目标物体和背景的特征表达,保证更好性能的目标跟踪。

由于 MDnet 利用卷积操作将目标跟踪作为目标与背景的二分类问题处理,导致其在跟踪过程中易受目标相似物的干扰,并且对目标遮挡的鲁棒性较差。针对该问题,Fan 等[35] 在 MDnet 的基础上进行改进,提出了 SAnet(Structure-Aware Network),其网络结构如图 2 所示。

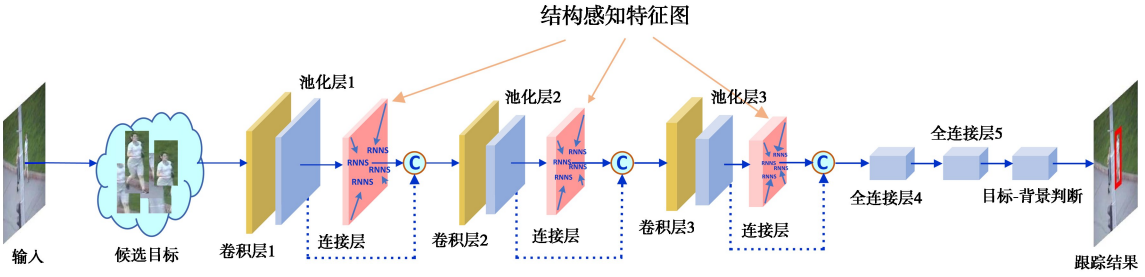


图 2　SAnet 网络结构

SAnet 通过引入 RNN 提取目标物体的结构信息,可以获得视频序列中目标前后帧的长期依赖关系,并将得到的目标显著特征使用跳跃式连接的方式与 CNN 特征融合。融合过程中,CNN 特征主导目标物体与背景之间的判辨,RNN 特征主导目标物体与相似物之间的判辨,以此增强 SAnet 在目标跟踪过程中的抗干扰能力。

除 MDnet 之外,Nam 等[36] 还提出了利用多个 CNN 构成树结构并协作以完成目标跟踪的 TCNN(CNNs in A Tree Structure)。TCNN 针对研究模型可靠性问题,认为以往部分目标跟踪模型随目标物体变化而以统一规则持续更新,但对于部分帧中目标被遮挡致使跟踪失败,导致更新的模型可靠性降低的问题并没有解决。

TCNN 的结构如图 3 所示,图中框线越粗,对应 CNN 模型可靠性越高;框与框的连接线越粗,对应两个 CNN 之间的相关性越高;箭头指示线越粗,对应 CNN 影响目标跟踪的权重越高。

TCNN 对每个 CNN 进行可靠性评估以确定权重,并利用所有 CNN 进行加权检测,充分发挥每个 CNN 的作用。此外,每十帧创建一个新节点,同时删除一个距离创建时刻最久的旧节点,并选择使新节点可靠性最高的节点作为父节点,完成在线微调,避免邻近帧之间的过拟合,以此保证 TCNN 在整个跟踪过程中的可靠性。参与构成 TCNN 的所有 CNN 的结构相同,如图 4 所示。

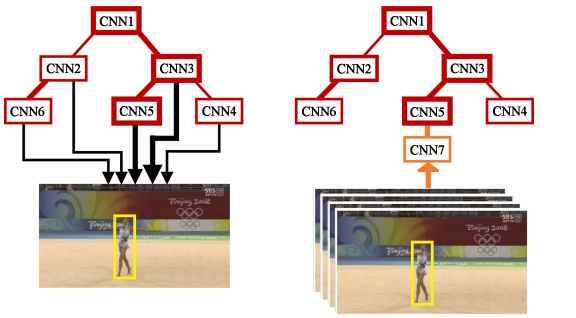


图 3　TCNN 网络结构

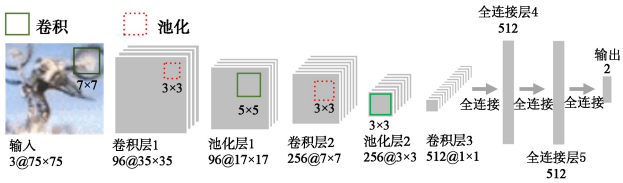


图 4　TCNN 中各 CNN 网络结构

TCNN 中所有 CNN 共享卷积层(Conv 层)的所有参数,通过避免冗余的网络评估减少了非必要的计算,同时节省了存储空间。

MDnet,SAnet 和 TCNN 在 VOT-2015 中的测评结果如表 4 所示。尽管 MDnet,SAnet 和 TCNN 在准确度,鲁棒性和 EAO 方面表现良好,但是 MDnet 和 SAnet 的速度仅为 1 f/s,TCNN 速度较前两者稍快但也只有 1.5 f/s,均难以满足实时跟踪的需求,这与目标跟踪要求精度高且速度快的初衷相悖。随着孪生网络这种特殊网络结构的出现与发展,如今基于孪生网络的目标跟踪已经可以实现精

度与速度的平衡，各项性能相较于以往的深度网络都有了极大改善，已经成为目标跟踪领域的研究热点。

表 4 MDnet，SAnet 和 TCNN 性能对比

跟踪方式	准确度		鲁棒性		EAO
	位列	得分	位列	得分	
MDnet	1.2	0.6	1.62	0.69	0.378 3
SAnet	1.17	0.61	1.58	0.69	0.389 5
TCNN	1.58	0.59	2.83	0.74	0.340 4

4 基于孪生网络的目标跟踪

孪生网络(Siamese Network)由两个及以上数目的子网络构成，子网络结构通常相同，使得孪生网络具有对称性，并且子网络同时接收视频帧输入，实现网络间权值共享，将不同子网络获得的图像特征用于获取相关响应图像，以此判断目标位置并完成目标跟踪。近年来，大量研究在原始孪生网络框架的基础上进行改进，如加入不同功能模块、使用双重孪生结构、优化骨干网络以及利用级联操作等方法，强化了孪生网络用于目标跟踪时的性能。

4.1 SINT 与 SiamFC

SINT(Siamese Instance Search for Tracking)是由 Tao 等[39]提出的基于孪生网络结构并最先用于目标跟踪的方法。SINT 通过生成多个候选目标与示例图像对比，选择差异最小的候选目标作为网络输出结果。虽然 SINT 将目标跟踪过程巧妙转化为目标匹配过程，为目标跟踪的实现提供了全新的思路，但是需要处理大量的候选目标，这导致了算法计算量大、耗时严重的问题。

随后 Bertinetto 等[40]提出了速度更快的 SiamFC(Fully-Convolutional Siamese Networks for Object Tracking)。SiamFC 在孪生网络结构中融入了全卷积网络(Fully Convolutional Network)，其网络结构如图 5 所示。

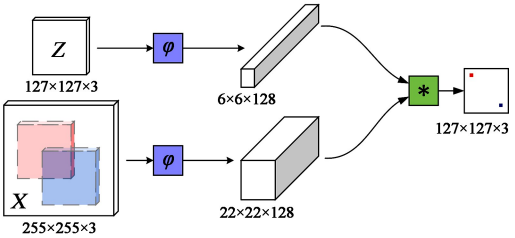


图 5 SiamFC 网络结构

SiamFC 有两条网络分支，其中模板分支用于保存目标模板信息，检测分支用于获取图像特征。通过卷积实现两条分支中图像相关性的计算，得到目标位置的最大响应并以此实现目标定位及完成目标跟踪。尽管 SiamFC 作为早期的基于孪生网络的目标跟踪算法的代表能够满足实时跟踪的需求，但是在精度和鲁棒性上与以往将深度特征和相关滤波相结合的方法仍存在一定差距。

4.2 SiamRPN 与 SiamMask

为了解决 SiamFC 在面对目标尺度变化时鲁棒性较差的问题，Li 等[41]借鉴了目标检测领域的 Faster R-CNN 算法[42]中使用的区域建议网络(Region Proposal Network，RPN)，将 RPN 加入 SiamFC 框架，提出了 SiamRPN。SiamRPN 由两个功能模块组成，Siamese 模块用于特征提取，其结构与 SiamFC 中网络结构相同，RPN 模块用于生成建议，即生成候选目标区域。RPN 模块的两条支路分别用于分类与回归，分类分支实现目标与背景的判辨，回归分支实现对目标的精确定位。SiamRPN 的网络结构如图 6 所示。

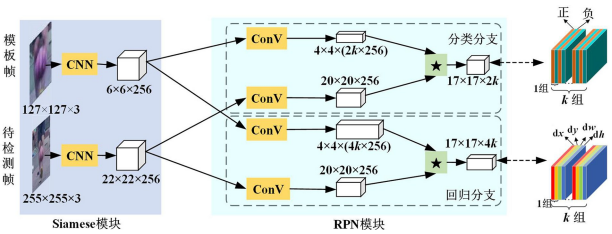
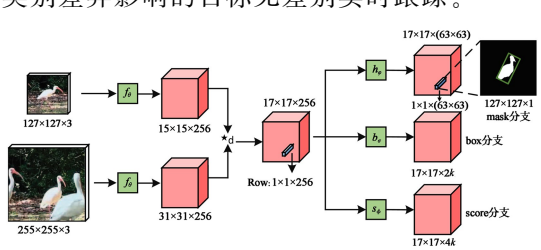


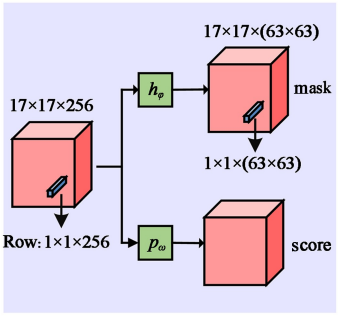
图 6 SiamRPN 网络结构

另一种基于 SiamFC 增加功能模块进行改进的方法是由 Wang 等[43]提出的、可同时实现目标分割与目标跟踪的 SiamMask。SiamMask 有二分支和三分支两种形式，具体结构如图 7 所示，其中三分支结构是在具有 mask 分支和 score 分支的二分支结构的基础上增加了 box 分支。

SiamMask 利用 mask 分支实现目标二值化分割，并将目标分割任务中的细节性特征用于目标跟踪任务。利用 box 分支得到目标候选区域，利用 score 分支完成候选区域的筛选，最终可实现不受目标类别差异影响的目标无差别实时跟踪。



(a) 三分支结构



（b）　二分支结构

图 7　SiamMask 网络结构

4.3　SA-Siam 与 TADT

除了使用功能模块来提升跟踪器在面对目标尺度显著性变化或类别任意时的鲁棒性，还可以通过改变网络结构来达到相同目的。

由 He 等[44]提出的 SA-Siam 使用双重孪生的网络结构，具有语义分支和表观分支，可以充分利用并融合具有互补性质的图像分类任务中学习到的语义特征和相似度来匹配任务中学习到的表观特征，其网络结构如图 8 所示。

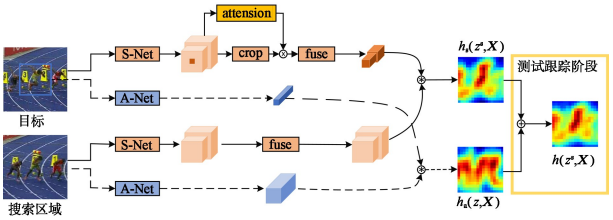


图 8　SA-Siam 网络结构

SA-Siam 的两条分支独立训练以保证两个 Siamese 网络的异质性，直到测试跟踪阶段才完成合并。此外，对于语义分支，加入了通道注意力模块（Channel Attention Module），使对目标跟踪影响更大的通道获得更大的权重，以此增强语义分支面对目标外观变化时的判辨能力，提高网络鲁棒性。

Li 等[45]针对目标性质不可知的跟踪情况，即面对任意尺度的任意目标对象类别，提出了 TADT 网络模型。TADT 使用目标感知模型（Target-Aware Model）指导选择目标跟踪所需要的有效特征。目标感知模型由回归损失（Regression Loss）部分（即岭损失（Ridge Loss）部分）和排序损失（Ranking Loss）部分组成。TADT 的网络结构如图 9 所示。TADT 通过目标感知模型从预先训练的 CNN 中选择具有目标运动特征和尺度敏感特征的目标感知卷积滤波器进行目标识别，之后通过相似度匹配计算相似度得分，以得分最大值所处位置表示目标位置，完成目标跟踪。优化后的 SA-Siam，TADT 和 SiamFC 在不同数据集中的评测结果如表 5 所示。

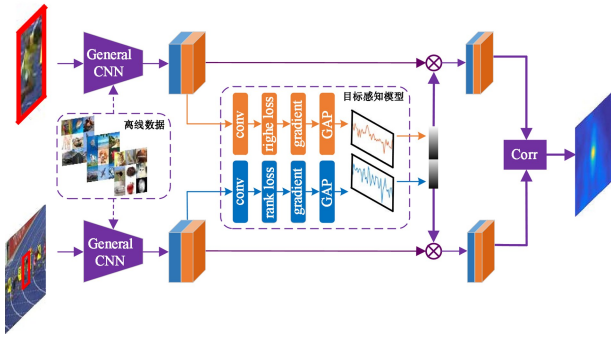


图 9　TADT 网络结构

表 5　各类算法在不同数据集性能对比

数据集	指标	SiamFC	SA-Siam	TADT
VOT-2015	EAO	0.292	0.31	0.327
	准确度	0.54	0.59	0.59
	鲁棒性	1.42	1.26	1.09
VOT-2016	EAO	/	0.291	0.299
	准确度	/	0.54	0.55
	鲁棒性	/	1.08	1.17

4.4　SiamDW，SiamRPN＋＋与 C-RPN

SiamDW 是由 Zhang 等[46]通过对孪生网络的骨干网络进行优化而得到的一种模型。SiamDW 认为网络层数越深，提取到的特征越好，但大量目标跟踪算法仍使用传统的 AlexNet[47]，这限制了算法改进的可能。文献[46]综合分析了 AlexNet，VGG-net，Inception[48]和 ResNet[49]等四种网络在不同网络步长、输出特征尺寸、感受野大小以及有无填充情况下的跟踪效果，发现各影响因素之间并不完全独立，且对于全卷积的孪生网络，填充操作会导致跟踪性能退化。基于以上结论，SiamDW 使用基于残差模块改进的 CIR 单元（Cropping-Inside Residual Units）来削弱由填充操作引起的偏差对目标跟踪的负面影响。

SiamRPN＋＋是由 Li 等[50]在 SiamRPN 的基础上改进骨干网络得到的使用多层特征融合的深层孪生网络模型。SiamRPN＋＋采用空间感知采样策略（Spatial Aware Sampling Strategy），使均匀采样在图像中心附近偏移后进行，克服了深层网络在特征学习过程中对图像中心位置分配更大权重的位置偏见，为使用 ResNet 作为骨干网络创造了条件。SiamRPN＋＋的网络结构如图 10 所示。

由于 ResNet 的不同层级可以获得不同图像特征，因此 SiamRPN＋＋采用了多级级联的方式完成

特征的逐层聚合，有助于目标跟踪性能的提升。SiamRPN++与SiamRPN在VOT-2018中的性能测评结果对比如表6所示。表6显示，对SiamRPN的骨干网络进行改进后得到的SiamRPN++，其各项性能均得到优化。此外，SiamRPN++能以35 f/s的速度保持跟踪的实时性。

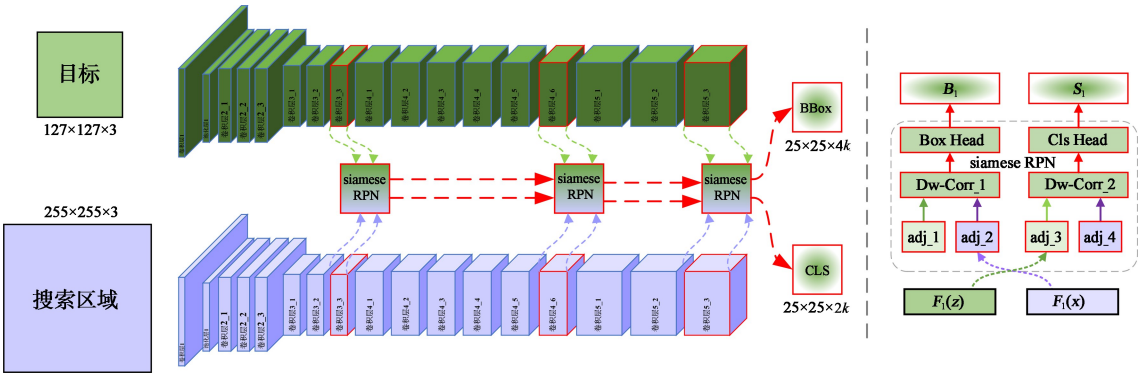


图10 SiamRPN++网络结构

表6 SiamRPN和SiamRPN++性能对比

跟踪方式	EAO	鲁棒性	准确度
SiamRPN	0.383	0.276	0.586
SiamRPN++	0.414	0.234	0.600

另一种以级联的方式对SiamRPN进行改进得到的多层特征融合的网络模型是由Fan等[51]提出的C-RPN，其网络结构如图11所示。

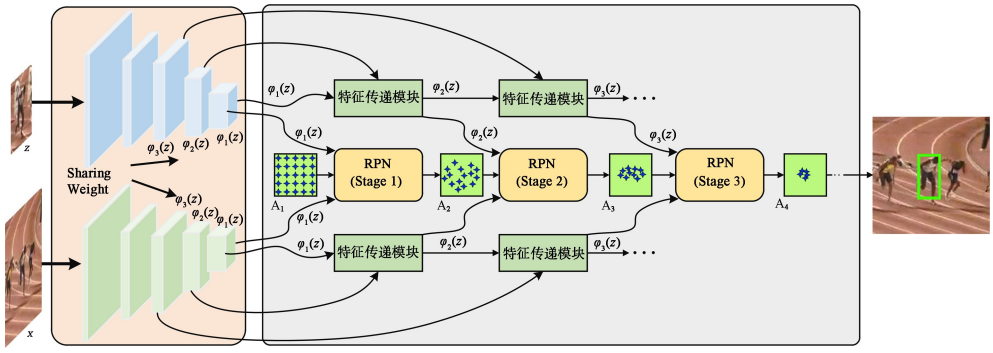


图11 C-RPN网络结构

C-RPN在进行不同层间的特征融合时，不同于SiamRPN++采用的线性加权求和方式，而是运用原创的特征传递模块(Feature Transfer Block)以获得更好的特征表达，有助于目标跟踪精度的提高。C-RPN与SiamRPN在VOT-2016中的性能测评结果对比如表7所示。

表7 SiamRPN和C-RPN性能对比

跟踪方式	EAO	鲁棒性	准确度
SiamRPN	0.344	1.12	0.560
C-RPN	0.363	0.95	0.594

4.5 ATOM

除了本文4.1～4.4小节中所介绍的利用具有结构对称性的孪生网络完成目标跟踪的方法外，还有部分方法虽然借鉴了孪生网络的并行框架，但是结构并不严格对称，如Danelljan等[52]提出的ATOM就可被视为类孪生结构的目标跟踪模型，其网络结构如图12所示。ATOM主要包括目标估计模块和目标分类模块，并借鉴目标检测领域的IoU-Net[53]设计了用于目标估计模块的IoU-Predictor网络。IoU-Predictor网络的结构如图13所示。

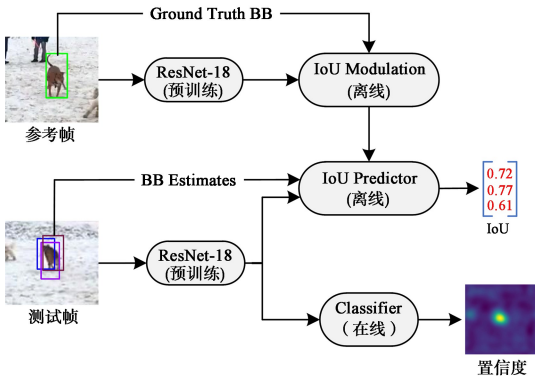


图12 ATOM网络结构

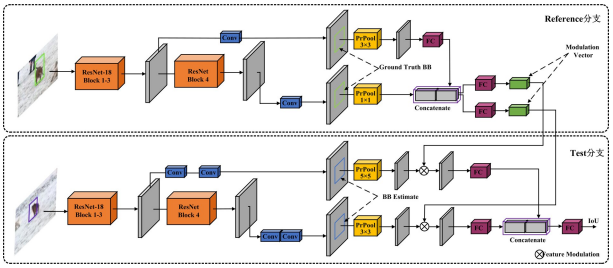


图 13　IoU-Predictor 网络结构

IoU-Predictor 网络主要包含 Reference 分支和 Test 分支。两分支的结构差异主要体现在 PrPool 层尺寸以及 PrPool 层前面放置的 Conv 层数量。

ATOM 在 VOT-2018 中的测评结果如表 8 所示。ATOM 的网络结构虽然并不严格对称，但利用并行框架仍然保证了良好的跟踪性能，并且能够以超过 30 f/s 的速度达到实时跟踪的要求，有效实现速度与精度的平衡。

表 8　ATOM 性能测评结果

跟踪方式	EAO	鲁棒性	准确度
ATOM	0.401	0.204	0.590

5　目标跟踪数据集

5.1　OTB

OTB 包括发布于 2013 年的 OTB-50 以及更新于 2015 年的 OTB-100 数据集，其中 OTB-100 包含 OTB-50 中的所有视频序列。OTB 数据集中的视频序列分为灰度视频序列和彩色视频序列，包括了针对目标形变、目标遮挡、目标尺度变化、光照变化、高速运动、运动模糊、超出观察视野、复杂背景干扰、图像分辨率低以及平面内、外旋转等目标跟踪领域常见问题的样本。此外，OTB 中每个样本序列对应一个用于记录人工标注的跟踪目标对象中心位置和目标对象跟踪框大小的文本文件。

5.2　VOT

VOT 是为目标跟踪领域每年的 VOT 竞赛而准备的数据集，数据集中样本每年持续更新，且难度逐年增大。VOT 数据集中均为彩色视频序列且视频序列的分辨率普遍较高。VOT 涉及的性能测评指标除了 EAO，准确度和鲁棒性，还包括 EFO (Equivalent Filter Operations)。EFO 旨在消除不同硬件平台和不同编程语言等外在因素对不同跟踪器速度的影响，进而以统一标准更客观地评价不同的目标跟踪算法。

5.3　TrackingNet

TrackingNet[54] 是目标跟踪领域的第一个大规模数据集，其包含 3 万多个视频图像序列以及超过 1 400 万个标注框，改变了以往间隔多帧标注一个目标的设计，实现了数据集所含每帧图像中目标的逐一标注，有效解决了基于深度学习的数据需求型跟踪算法缺乏专用的大规模数据集的问题。

5.4　GOT-10k

GOT-10k[55] 是由中国科学院发布的建立在 WordNet 结构主干之上的目标跟踪数据集，同时也是目标跟踪领域第一个使用 WordNet 的语义层次指导类填充的数据集。GOT-10k 的训练集包括 560 多种目标类别和 87 种运动模式，测试集包括 80 多种目标类别和 32 种运动模式，实现了野外常见运动目标对象的广泛覆盖，并且训练集和测试集的样本零重叠。

除上述数据集外，Temple-color，ALOV＋＋，UAV123 以及 VIVID Tracking 等数据集在目标跟踪领域同样得到大量使用。

6　展望与总结

通过对目标跟踪领域不同发展阶段的各类型方法进行特点剖析和性能对比，发现各方法仍存在的问题，结合近年研究趋势，对目标跟踪领域的未来发展方向展望如下：

1)全面或专用数据集的增加。从网络类型出发，针对数据需求型的深度网络，设计涵盖更广泛目标类别、更多种常见动作的超大规模数据集；从应用领域出发，针对不同应用场景，如天上、地下和水中，室内或室外，城市或无人区等，设计样本针对性更强的数据集。

2)硬、软件的升级。提高硬件模块性能，优化软件中的代码效率，减少深度网络因网络结构复杂、计算复杂度高和数据处理量大等特点对跟踪速度产生的负面影响。

3)不同领域经验的借鉴。从目标检测、目标分割、数据增强、迁移学习等领域借鉴成熟的经验，将不同领域处理问题的思想或相关的网络模型改良后运用到目标跟踪领域。

4)网络结构的创新。构造新的深度网络框架，更好地提取或融合不同特征，在保证速度与精度二者平衡的基础上，提高跟踪器对目标形变、目标遮挡、目标尺度变化、光照变化、高速运动、运动模糊、

超出观察视野、复杂背景干扰、图像分辨率低等目标跟踪领域常见问题的鲁棒性。

综上所述，本文首先梳理了目标跟踪领域的发展历程，简单介绍了传统方法和基于相关滤波的方法，详细分析了基于深度网络的目标跟踪方法，并对基于孪生网络这种特殊网络结构的跟踪方法进行了展开，然后介绍了部分典型数据集，最后展望了目标跟踪领域的未来发展方向，为后续研究提供了参考。

参考文献：

[1] 罗 元，王薄宇，陈 旭. 基于深度学习的目标检测技术的研究综述[J]. 半导体光电，2020，41(1)：1-10.
Luo Yuan，Wang Boyu，Chen Xu. Research progresses of target detection technology based on deep learning[J]. Semiconductor Optoelectronics，2020，41(1)：1-10.

[2] Aggarwal J K，Xia L. Human activity recognition from 3D data：A review[J]. Pattern Recognition Lett.，2014，48(10)：70-80.

[3] Brown M，Funke J，Erlien S，et al. Safe driving envelopes for path tracking in autonomous vehicles[J]. Control Engineering Practice，2017，61(4)：307-316.

[4] Severson J. Human-digital media interaction tracking：US，US20100076274 A1[P]. 2017-03-19.

[5] Onate J M B，Chipantasi D J M，Erazo N d R V. Tracking objects using artificial neural networks and wireless connection for robotics [J]. J. of Telecommun.，Electronics and Computer Engin.，2017，9(1/3)：161-164.

[6] Walker S，Sewell C，Park J，et al. Systems and methods for localizing，tracking and/or controlling medical instruments：US，US009629595B2[P]. 2017-4-25.

[7] Comaniciu D，Ramesh V，Meer P. Real-time tracking of non-rigid objects using mean shift[C]//Proc. of IEEE Conf. on Computer Vision and Pattern Recognition CVPR，2000：142-149.

[8] Comaniciu D，Meer P. Mean shift：A robust approach toward feature space analysis[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence，2002，24(5)：603-619.

[9] Nummiaro K，Koller-Meier E，Van Gool L. An adaptive color-based particle filter[J]. Image and Vision Computing，2003，21(1)：99-110.

[10] 李天成，范红旗，孙树栋. 粒子滤波理论、方法及其在多目标跟踪中的应用[J]. 自动化学报，2015，41(12)：1981-2002.
Li Tiancheng，Fan Hongqi，Sun Shudong. Particle filtering：theory，approach and application for multitarget tracking[J]. Acta Automatica Sinica，2015，41(12)：1981-2002.

[11] Comaniciu D，Ramesh V，Meer P. Kernel-based object tracking[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence，2003，25(5)：564-575.

[12] Nguyen H T，Worring M，Boomgaard R V D. Occlusion robust adaptive template tracking [C]// Proc. of Eighth IEEE Inter. Conf. on Computer Vision，2002：678-683.

[13] Peng N S，Yang J，Liu Z. Mean shift blob tracking with kernel histogram filtering and hypothesis testing[J]. Pattern Recognition Lett.，2005，26(5)：605-614.

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

[15] Ross D A，Lim J，Lin R S，et al. Incremental learning for robust visual tracking[J]. Inter. J. of Computer Vision，2008，77(1/3)：125-141.

[16] Babenko B，Yang M H，Belongie S. Visual tracking with online multiple instance learning [C]// Proc. of IEEE Computer Vision and Pattern Recognition，2009：983-990.

[17] Henriques J F，Caseiro R，Martins P，et al. Exploiting the circulant structure of tracking-by-detection with kernels [C]//Proc. of the 12th European Conf. on Computer Vision-Volume Part Ⅳ，2012：1-14.

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

[19] Danelljan M，Khan F S，Felsberg M，et al. Adaptive color attributes for real-time visual tracking[C]//Proc. of 2014 IEEE Conf. on Computer Vision and Pattern Recognition，2014：1090-1097.

[20] Danelljan M，Gustav Hager，Khan F S，et al. Accurate scale estimation for robust visual tracking[C]//Proc. of British Machine Vision Conf.，2014：1-11.

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

[22] Li Y，Zhu J. A scale adaptive kernel correlation filter tracker with feature integration[C]//Proc. of European Conf. on Computer Vision. Springer，Cham，2014：254-265.

[23] Danelljan M，Hager G，Khan F S，et al. Learning spatially regularized correlation filters for visual tracking[C]// Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition，2015：4310-4318.

[24] Kiani H G，Sim T，Lucey S. Correlation filters with limited boundaries[C]// Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition，2015：4630-4638.

[25] Galoogahi H K，Fagg A，Lucey S. Learning background-aware correlation filters for visual tracking[C]// Proc. of 2017 IEEE Inter. Conf. on Computer Vision (ICCV)，2017：1144-1152.

[26] Simonyan K，Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. arXiv e-prints，2014，1409：1-14.

[27] Ma C，Huang J B，Yang X，et al. Hierarchical convolutional features for visual tracking[C]//Proc. of IEEE Inter. Conf. on Computer Vision，2015：3074-3082.

[28] Ma C, Huang J B, Yang X, et al. Robust visual tracking via hierarchical convolutional features[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017: 2709-2723.

[29] Qi Y K, Zhang S P, Qin L, et al. Hedged deep tracking[C]// Proc. of 2016 IEEE Conf. on Computer Vision and Pattern Recognition, 2016: 4311.

[30] Danelljan M, Hager G, Khan F S, et al. Convolutional features for correlation filter based visual tracking[C]//Proc. of 2015 IEEE Inter. Conf. on Computer Vision Workshop (ICCVW), 2015: 621-629.

[31] Gladh S, Danelljan M, Khan F S, et al. Deep motion features for visual tracking[C]//Proc. of 23rd Inter. Conf. on Pattern Recognition, 2016: 1243-1248.

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

[33] Danelljan M, Bhat G, Khan F S, et al. ECO: Efficient convolution operators for tracking[C]// Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2017: 6638-6646.

[34] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]// Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Piscataway, 2016: 4293-4302.

[35] Fan H, Ling H. SANet: Structure-aware network for visual tracking[C]//Proc. of 2017 IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2017: 2217-2224.

[36] Nam H, Baek M, Han B. Modeling and propagating CNNs in a tree structure for visual tracking[J]. arXiv e-prints, 2016, 1608: 07242.

[37] Danelljan M, Bhat G, Gladh S, et al. Deep motion and appearance cues for visual tracking[J]. Pattern Recognition Lett., 2019, 124(6): 74-81.

[38] Bhat G, Johnander J, Danelljan J, et al. Unveiling the power of deep tracking[J]. arXiv e-prints, 2018, 1804: 06833.

[39] Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]//Proc. of 2016 IEEE Conf. on Computer Vision and Pattern Recognition, 2016: 1420-1429.

[40] Bertinetto L, Valmadre J. Fully-convolutional siamese networks for object tracking[C]//Proc. of European Conf. on Computer Vision, 2016: 850-865.

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

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

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

[44] He A F, Luo C, Tian X M, et al. A twofold siamese network for real-time object tracking[C]// Proc. of 2018 IEEE Conf. on Computer Vision and Pattern Recognition, 2018: 4834-4843.

[45] Li X, Ma C, Wu B, et al. Target-aware deep tracking[C]// Proc. of 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019: 1369-1378.

[46] Zhang Z, Peng H. Deeper and wider siamese networks for real-time visual tracking[C]//Proc. of 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019: 4586-4595.

[47] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// Proc. of Conf. on Neural Information Processing Systems, 2012: 1106-1114.

[48] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2015: 1-9.

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

[50] Li B, Wu W, Wang Q, et al. SiamRPN++: Evolution of siamese visual tracking with very deep networks[C]//Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019: 4277-4286.

[51] Fan H, Ling H. Siamese cascaded region proposal networks for real-time visual tracking[C]//Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019: 7944-7953.

[52] Danelljan M, Bhat G. ATOM: Accurate tracking by overlap maximization[C]//Proc. of IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2019: 4655-4664.

[53] Jiang B, Luo R, Mao J, et al. Acquisition of localization confidence for accurate object detection[J]. arXiv e-prints, 2018, 1807: 11590.

[54] Muller M, Bibi A, Giancola S, et al. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild[J]. arXiv e-prints, 2018, 1803: 10794.

[55] Huang L, Zhao X, Huang K. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild[C]// Proc. of IEEE Trans. on Pattern Analysis and Machine Intelligence, 2019: 1-1.

作者简介：

罗　元(1972—)，女，1993 年毕业于浙江大学光学仪器专业，2003 年获重庆大学仪器科学与技术专业博士学位，教授，研究领域为数字图像处理及智能信号处理。