中图法分类号: TP301.6　　**文献标识码:** A　　**文章编号:** 1006－8961(2019)12－2057－24

论文引用格式: Li X, Zha Y F, Zhang T Z, Cui Z, Zuo W M, Hou Z Q, Lu H C and Wang H Z. 2019. Survey of visual object tracking algorithms based on deep learning. Journal of Image and Graphics, 24(12): 2057-2080(李玺, 查宇飞, 张天柱, 崔振, 左旺孟, 侯志强, 卢湖川, 王菡子. 2019. 深度学习的目标跟踪算法综述. 中国图象图形学报, 24(12): 2057-2080) [DOI: 10.11834/jig.190372]

深度学习的目标跟踪算法综述

李玺[1], 查宇飞[2], 张天柱[3], 崔振[4], 左旺孟[5], 侯志强[6], 卢湖川[7], 王菡子[8]

1. 浙江大学计算机科学与技术学院, 杭州　310007; 2. 西北工业大学计算机学院, 西安　710043;
3. 中国科学院自动化研究所, 北京　100190; 4. 南京理工大学计算机科学与工程学院, 南京　210094;
5. 哈尔滨工业大学计算机科学与技术学院, 哈尔滨　150006; 6. 西安邮电大学计算机学院, 西安　710121;
7. 大连理工大学信息与通信工程学院, 大连　116024; 8. 厦门大学信息科学与技术学院, 厦门　361001

摘　要: 目标跟踪是利用一个视频或图像序列的上下文信息,对目标的外观和运动信息进行建模,从而对目标运动状态进行预测并标定目标位置的一种技术,是计算机视觉的一个重要基础问题,具有重要的理论研究意义和应用价值,在智能视频监控系统、智能人机交互、智能交通和视觉导航系统等方面具有广泛应用。大数据时代的到来及深度学习方法的出现,为目标跟踪的研究提供了新的契机。本文首先阐述了目标跟踪的基本研究框架,从观测模型的角度对现有目标跟踪的历史进行回顾,指出深度学习为获得更为鲁棒的观测模型提供了可能;进而从深度判别模型、深度生成式模型等方面介绍了适用于目标跟踪的深度学习方法;从网络结构、功能划分和网络训练等几个角度对目前的深度目标跟踪方法进行分类并深入地阐述和分析了当前的深度目标跟踪方法;然后,补充介绍了其他一些深度目标跟踪方法,包括基于分类与回归融合的深度目标跟踪方法、基于强化学习的深度目标跟踪方法、基于集成学习的深度目标跟踪方法和基于元学习的深度目标跟踪方法等;之后,介绍了目前主要的适用于深度目标跟踪的数据库及其评测方法;接下来从移动端跟踪系统,基于检测与跟踪的系统等方面深入分析与总结了目标跟踪中的最新具体应用情况,最后对深度学习方法在目标跟踪中存在的训练数据不足、实时跟踪和长程跟踪等问题进行分析,并对未来的发展方向进行了展望。

关键词: 视觉目标跟踪; 深度神经网络; 相关滤波器; 深度孪生网络; 强化学习; 生成对抗网络

Survey of visual object tracking algorithms based on deep learning

Li Xi[1], Zha Yufei[2], Zhang Tianzhu[3], Cui Zhen[4], Zuo Wangmeng[5], Hou Zhiqiang[6], Lu Huchuan[7], Wang Hanzi[8]

1. *College of Computer Science and Technology, Zhejiang University, Hangzhou　310007, China;*
2. *School of Computer Science, Northwestern Polytechnical University, Xi'an　710043, China;*
3. *Institute of Automation, Chinese Academy of Sciences, Beijing　100190, China;*
4. *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing　210094, China;*
5. *School of Computer Science and Technology, Harbin Institute of Technology, Harbin　150006, China;*
6. *College of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an　710121, China;*
7. *School of Information and Communication Engineering, Dalian University of Technology, Dalian　116024, China;*
8. *School of Information Science and Engineering, Xiamen University, Xiamen　361001, China*

收稿日期: 2019-07-15; **修回日期:** 2019-08-19; **预印本日期:** 2019-08-26

基金项目: 国家自然科学基金项目(U1605252, 61872307, 61773397)

Supported by: National Natural Science Foundation of China(U1605252, 61872307, 61773397)

Abstract: Object tracking is a fundamental problem in computer vision, which uses context information in a video or image sequence to predict and locate a target (s). It is widely used in smart video monitoring systems, intelligent human interaction, intelligent transportation, visual navigation systems, and many other areas. With the advent of the big data era and the emergence of deep learning methods, tracking performance has substantially improved. In this paper, we introduce the basic research framework of object tracking and review the history of object tracking from the perspective of the observation model. We indicate that deep learning allows for a more robust observation model to be obtained. We review the deep learning methods that are suitable for object tracking from the aspects of deep discriminative model and deep generative model. We also classify and analyze the existing deep object tracking methods from the perspectives of network structure, network function, and network training. In addition, we introduce several other deep object tracking methods, including deep object tracking based on the fusion of classification and regression, on reinforcement learning, on ensemble learning, and on meta-learning. We show the current commonly used databases for object tracking based on deep learning and their evaluation methods. We likewise analyze and summarize the latest specific application scenarios in object tracking from the perspectives of mobile tracking system, detection, and tracking-based system. Finally, we analyze the problems of object tracking, including insufficient training data, real-time tracking, and long-term tracking and specify further research directions for deep object tracking.

Key words: visual object tracking; deep neural network; correlation filter; deep Siamese network; reinforcement learning; generative adversarial network

0 引 言

目标跟踪是计算机视觉的一个重要分支，其利用视频或图像序列的上下文信息，对目标的外观和运动信息进行建模，从而对目标运动状态进行预测并标定目标的位置。目标跟踪融合了图像处理、机器学习、最优化等多个领域的理论和算法，是完成更高层级的图像理解(如目标行为识别)任务的前提和基础(Li 等，2018c; Lu 等，2018)。随着计算机处理能力的飞速提升，各种基于目标跟踪的民用和军用系统纷纷落地，广泛应用于智能视频监控(Huang 等，2015; Collins 等，2000; Haritaoglu 等，2000; Shu 等，2005)、智能人机交互(Bonin-Font 等，2008; Li 等，2003)、智能交通(Lu 等，2010)、视觉导航(Hu 等，2007; Kristan 等，2013)、无人驾驶、无人自主飞行、战场态势侦察(Li 等，2018c; Lu 等，2018)等领域。

国际上，卡内基(梅隆大学、麻省理工学院等多所高校和研究机构最先开展了目标跟踪相关项目的研究(Wang 等，2015; Lowe 等，2004)，并结合多传感器技术，提高了对城市的主动监视和对战场的态势感知能力。IBM 研究院开发的 S3 系统(smart surveillance system)(Haritaoglu 等，2000)能够实现多目标跟踪并完成对目标行为的异常检测。英国的雷丁大学、伦敦大学则致力于民用的视频监控项目(Van 等，2009; 管皓 等，2016)，开发出了能在复杂场景下的行人跟踪和行为理解，以及可用于监测、引导交通流量并实现异常预警的公共交通管理系统。国内也有很多研究所和高校成立了相关课题组，如中国科学院自动化研究所模式识别国家重点实验室、大连理工大学通信与信号处理研究所、西安交通大学图像处理与识别研究所、香港中文大学、清华大学和上海交通大学图像处理与模式识别研究所等，针对目标跟踪在计算机视觉领域中的应用进行研究。除了具体的工程应用，目标跟踪也是各个领域顶级期刊和会议的重要主题，每年都有大量最新最前沿的成果发表在 IEEE TPAMI(IEEE Transactions on Pattern Analysis and Machine Intelligence)、IJCV(International Journal of Computer Vision)、PR(Pattern Recognition)等期刊和 ICCV(IEEE International Conference on Computer Vision)、CVPR(IEEE Conference on Computer Vision and Pattern Recognition)、ECCV(European Conference on Computer Vision)等会议。与此同时，国内举办的各种相关会议和主办的相关期刊也包含目标跟踪主题，如国际图像图形会议(ICIG)、模式识别与计算机视觉会议(PRCV)、视觉与学习青年学者研讨会(VALSE)等会议和《中国科学信息科学》、《计算机学报》、《自动化学报》、《中国图象图形学报》等期刊。这些前沿的研究极大促进了

视觉跟踪领域的理论发展和工程应用,使得相关技术越来越集中应用于国防、企业以及个人生活中。

依据跟踪目标数目的不同,目标跟踪可分为单目标跟踪和多目标跟踪。本文主要关注单目标跟踪。随着深度学习在图像分类和目标检测等计算机视觉任务中的成功应用,深度学习也开始大量应用于目标跟踪算法中,已经取得了很多的研究成果。因此,本文对基于深度学习的目标跟踪算法进行系统的梳理,旨在为目标跟踪的进一步发展提供参考。

1 现有目标跟踪方法简介

目前已有大量的目标跟踪算法,按照其发展脉络,图 1 展示了各时间节点的代表性算法,后文将详细介绍各算法。可以看出,2012 年是一个重要的分界线;以 AlexNet 网络为代表的深度学习方法在图像识别等领域获得了巨大成功,随后迅速被引入到目标跟踪领域中。

视觉跟踪系统的基本框架一般由搜索策略、特征提取和观测模型等模块组成。目前常用的搜索策略包括均值漂移(mean shift)(Comaniciu and Meer,2002)、粒子滤波(particle filter)(Isard and Blake,1998)和循环密集采样(cyclic dense sampling)(Robert 等,2005)等几种方式;通过搜索策略获得候选样本后进行特征提取,主要包括人工特征和学习特征;最后,利用特征判断候选样本是否为跟踪目标的观测模型,通常分为生成式模型和判别式模型。由于观测模型对跟踪结果至关重要,本文主要从观测模型的角度来回顾一下目标跟踪的发展历程(Wang 等,2015)。

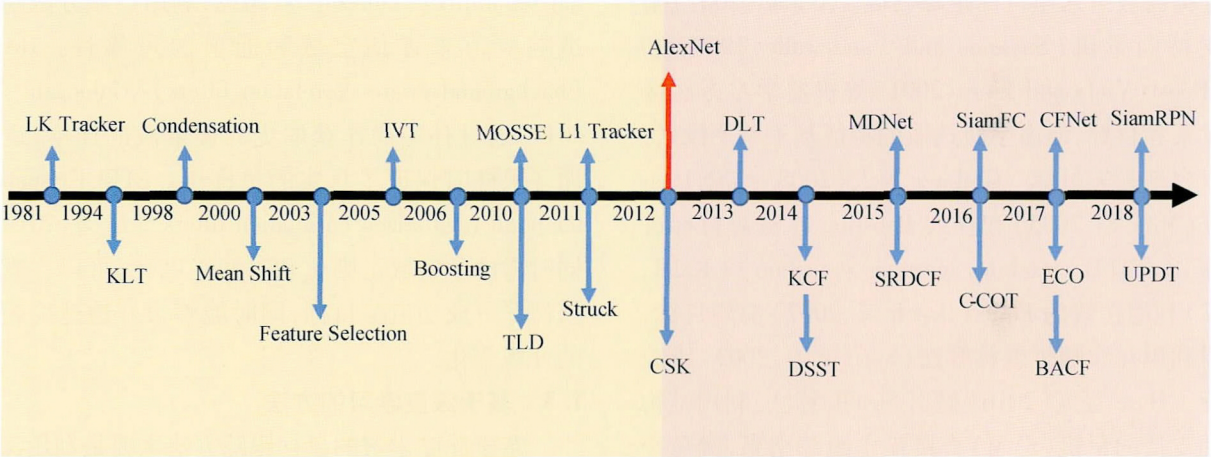


图 1 各时间节点的代表性目标跟踪算法

Fig. 1 Some representative object tracking algorithms

1.1 基于生成式模型的方法

生成式模型提取目标特征构建表观模型,在图像中搜索与模型最匹配的区域作为跟踪结果(Krizhevsky 等,2012)。最早的目标跟踪工作可以追溯到 1981 年提出的 LK 光流法(Horn and Schunck,1981),它假定目标灰度在短时间内保持不变,同时目标邻域内的速度向量场变化缓慢。KLT(Kanade Lucas Tomasi tracking method)(Shi and Tomasi,1994)通过匹配角点实现对目标的跟踪。随后工作则考虑采用原始的外观(Isard and Blake,1998)或者颜色(Comaniciu and Meer,2002)作为主要特征来描述目标,或者采用更为复杂的混合方式描述目标(Jepson 等,2003),这种方法包含三部分,分别描述目标的稳定特征、瞬变特征和噪声过程。然而,上述特征很难描述目标的变化,如当目标角度突然发生剧烈变化时,该外观模型就会失效,甚至丢失目标,导致跟踪任务失败。为了解决这些问题,多角度模型被用来描述目标,它通过计算当前含有目标的图像和用特征向量重建的图像之间的仿射变换差异来跟踪目标。在此基础上,Ross 等人(2008)在线更新特征空间的基,直接将以前检测到的目标作为样本在线学习而无需大量的标注样本。L1 跟踪器(Mei and Ling,2011)把跟踪看做一个稀疏近似问题,通过求解 L1 范数最小化问题,实现对目标的跟踪。同时,SIFT(scale invariant feature transform)(Cruz-Mota 等,2012),SURF(speeded up robust fea-

tures)(Ross 等,2008),最大稳定极值区域(MSER)(Matas 等,2004)等更为鲁棒的局部特征也用来描述目标,以适应目标在局部的各种尺度和旋转的变化。生成式模型不论采用全局特征还是局部特征,其本质是在目标表示的高维空间中,找到与目标模型最相邻的候选目标作为当前估计。但是,此类方法的缺陷在于只关注目标信息,而忽略了背景信息。

1.2 基于判别式模型的方法

与生成式模型不同的是,判别式模型同时考虑了目标和背景信息。判别式模型将跟踪问题看做分类或者回归问题,目的是寻找一个判别函数,将目标从背景中分离出来,从而实现对目标的跟踪。

1)分类判别式模型。早期,Collins 等人(2005)利用线性判别分析自适应地选择对当前背景和目标最具区分力的颜色特征,从而分离出目标。随后,各种分类器被引入至目标跟踪领域。Avidan(2007)采用支持向量机(Suykens and Vandewalle,1999)和 AdaBoost(Viola and Jones,2001)等机器学习方法区分背景和目标,但由于所选取的特征基于单个像素,所以容易丢失目标。Grabner 等人(2008)结合 Haar 特征(Mita 等,2005)和在线 Boosting 算法对目标进行跟踪。TLD(tracking learning detection)(Kalal,2012)利用在线的 Ferns(Bosch 等,2007)检测目标,同时利用在线随机森林算法(Svetnik 等,2003)跟踪目标。Hare 等人(2016)提出 Struck 算法,利用结构化的支持向量机(SVM)直接输出跟踪结果,避免中间分类环节,取得了优异的性能。

2)回归判别式模型。基于回归判别模型的典型方法是相关滤波,其利用循环矩阵,通过快速傅里叶变换实现时域到频域的转换,大大提升了算法的速度。相关滤波因速度优势受到了广泛关注,逐渐成为目标跟踪领域的主流框架。Henriques 等人在 MOSSE 算法(David 等,2010)的基础上,提出 CSK(circulant structure of tracking by detection with kernels)(Henriques 等人,2012)算法,也称为核相关滤波算法,其采用循环移位进行密集采样,并通过核函数将低维线性空间映射到高维空间,提高了相关滤波器的鲁棒性。随后的工作主要从特征选择、尺度估计、正则化等方面对该算法进行改进和提高。特征选择方面,可使用方向梯度直方图(HOG)、CN(color names)等特征更好地表征目标。尺度估计方面,SAMF(scale adaptive multiple feature)(Li 等,2015)同时检测目标位置和尺度的变化,采用图像金字塔进行尺度选择,最佳尺度对应最大响应值;DSST(accurate scale estimation for robustvisual tracking)(Danelljan 等,2014)则将目标跟踪看成位置变化和尺度变化两个独立问题,首先训练位置平移相关滤波器以检测目标中心平移,然后训练尺度相关滤波器来检测目标的尺度变化。由于相关滤波采用循环移位采样,导致除了中心样本以外的其他样本都会存在边界,这称为边界效应。SRDCF(learning spatially regularized correlation filters for visual tracking)(Danelljan 等,2015a)采用了大的检测区域,在滤波器系数上加入权重约束,越靠近边缘权重越大,越靠近中心权重越小,从而使得滤波器系数主要集中在中心区域,有效地缓解了边界效应。CSR-DCF(discriminative correlation filter with channel and spatial reliability)(Lukežic 等,2017)利用空域分割和通道响应值来评估空域和通道的可靠性。BACF(background-aware correlation filters)(Galoogahi 等,2017)通过补零操作获取更大搜索域的样本,进行循环采样时保证了真实的负样本。STRCF(spatial-temporal regularized correlation filters)(Li 等,2018b)同时考虑了空域正则化和时间正则化,可以在有遮挡情况下成功追踪目标,同时能够很好地适应较大的外观变化。

1.3 基于深度学习的方法

基于深度学习的目标跟踪方法主要是利用深度特征强大的表征能力来实现跟踪。按照利用深度特征的方式,可分为基于预训练深度特征的跟踪和基于离线训练特征的跟踪。

1)基于预训练深度特征的跟踪。早期的一些工作(Wang and Yeung,2013)直接利用 ImageNet 数据上的预训练模型提取深度特征。HCF(hierarchical convolutional features for visual tracking)(Ma 等,2015)利用 VGG(visual geometry group)网络的深层特征与浅层特征,融入到相关滤波器获得了很好的跟踪性能。但该算法并没有对尺度进行处理,在整个跟踪序列中都假定目标尺度不变,因此对尺度变化较大的跟踪目标并不鲁棒。HDT(hedged deep tracking)(Qi 等,2016)利用 Hedge 算法将每一层特征训练出来的相关滤波器进行融合提升。C-COT(continuous convolution operators for visual tracking)(Danelljan 等,2016)将浅层表观信息和深层语义信

息结合起来，根据不同空间分辨率的响应，在频域进行插值得到连续空间分辨率的响应图，通过迭代求得最佳位置和尺度。为了解决 C-COT 速度慢的问题，ECO（efficient convolution operators）（Danelljan 等，2017）通过卷积因式分解操作、样本分组和更新策略对其改进，在不影响算法精确度的同时，算法速度提高了一个数量级。UPDT（unveiling the power of deep tracking）（Bhat 等，2018）区别对待深度特征和浅层特征，利用数据增强和差异响应函数提高鲁棒性和准确性，同时利用提出的质量评估方法自适应融合响应图，得到最优的目标跟踪结果。

2）基于离线训练特征的跟踪。基于离线训练特征的跟踪则是通过端到端的方式训练与目标跟踪任务相匹配的特征，从而获得更好的跟踪性能。MDNet（Nam and Han，2016）跟踪算法设计一个轻量级的小型网络学习卷积特征表示目标，利用 SoftMax（Kumarawadu 等，2002）对采样样本分类，其性能表现非常优异，但速度只有 1 帧/s。随后提出的 SiamFC（Bertinetto 等，2016b）算法，则是利用孪生网络（Siamese network），在视频序列 ILSVRC2015 离线训练一个相似性度量函数，在跟踪过程中利用该模型，选择与模板最相似的候选作为跟踪结果。Tao 等（2016）提出 SINT（Siamese instance search network）算法，利用孪生网络直接学习目标模板和候选目标的匹配函数，在线跟踪过程中只用初始帧的目标作为模板来实现跟踪。在孪生网络获得目标位置的基础上，区域提议网络被用来直接估计目标尺度（Li 等，2018d），同时提高了跟踪性能和效率。

如何将相关滤波融入深度学习框架，通过端到端的形式训练最适合相关滤波的深度特征成为许多工作关注的热点问题。Valmadre 等人（2017）提出的 CFNet，首先将相关滤波改写成可微分的神经网络层，和特征提取网络整合到一起以实现端到端优化，训练与相关滤波器相匹配的卷积特征。VOT2017 竞赛冠军算法 CFCF（good features to correlate for visual tracking）（Gundogdu and Alatan，2018）则是通过精调网络模型，学习适用于相关滤波的深度特征，然后将学到的深度特征引入 C-COT 的跟踪框架。最新的一些工作则是将深度学习最新进展，如元学习（Park and Berg，2018）、生成式对抗网络（GAN）（Song 等，2018）等，引入目标跟踪领域，以期获得更好的跟踪性能。

2 适用于目标跟踪的深度学习模型

2.1 深度判别式模型

2.1.1 卷积神经网络

卷积神经网络（CNN）是一种前馈神经网络，其中每个神经元的响应与前一层感受野范围内的神经单元相关。卷积神经网络通常包含卷积层、激活函数、池化层以及全连接层。其中，卷积层由若干卷积单元组成，卷积运算是为了对输入进行特征提取，浅层网络可能是提取一些低级的特征如边缘、轮廓和角点等，深层网络从低级特征中提取更为复杂的特征。池化层即降采样层，通常卷积层之后会得到维度较大的特征，池化后可得到维度较小的特征。全连接层把所有局部特征结合变成全局特征。代表性的 CNN 模型有 AlexNet（Krizhevsky 等，2012）、VGGNet（Simonyan and Eisserman，2015）、GoogLeNet（Szegedy 等，2015）、ResNet（He 等，2016）、DenseNet（Huang 等，2017b）等，网络结构详见对应的参考文献。

2.1.2 循环神经网络

门循环单元（GRU）（Cho 等，2014）是循环神经网络（RNN）的一种，是为了解决长期记忆和反向传播中的梯度问题而提出来的。相比长短期记忆网络（LSTM），使用 GRU 能够达到相当的效果，并且相比之下更易训练，能够很大程度上提高训练效率。GRU 的输入和输出结构与普通的 RNN 是一样的，其原理与 LSTM 非常相似，即用门控机制控制输入、记忆等信息。GRU 有两个门，即重置门和更新门。其中，重置门决定了如何将新的输入信息与前面的记忆相结合，更新门定义了前面记忆保存到当前时刻的比重。如果将重置门设置为 1，更新门设置为 0，即可得到一个标准的 RNN 模型。LSTM 有输入门、遗忘门和输出门三个门，其输入门和遗忘门对应于 GRU 的更新门，GRU 并不会控制并保留内部记忆且没有 LSTM 中的输出门。

传统的 LSTM 使用的是全连接长短期记忆网络，没有考虑空间上的相关性，并且包含了大量冗余的空间数据。针对该问题，研究人员提出了 ConvLSTM 方法（Shi 等，2015），其核心本质与传统 LSTM 相同，都是将上一层的输出作为下一层的输入。不同之处在于，ConvLSTM 加入了卷积结构，使其不仅

具有 LSTM 的时序建模能力，而且还能够像 CNN 一样提取空间特征，并且状态与状态之间的切换替换为卷积计算，从而使其同时具备时空特性。ConvLSTM 可以很直观地扩展到其他具有时空序列的预测问题中，例如跟踪问题。

2.2 深度生成式模型

2.2.1 生成式对抗网络

生成式对抗网络(GAN)主要包括生成器(generator)与判别器(discriminator)。生成器通过学习真实图像分布从而使生成的图像更加真实，以欺骗判别器；判别器则需要对接收的图像进行真假判别。在训练过程中，生成器努力地让生成的图像更加真实，而判别器则努力地去识别出图像的真假。随着训练的迭代，生成器和判别器在不断地进行对抗，期望网络达到一个纳什均衡状态：生成器生成的图像接近于真实的图像分布，而判别器识别不出真假图像。整个系统可以用反向传播进行训练。代表性的生成式对抗网络包括 DCGAN(deep convolution generative adversarial networks)(Radford 等，2016)、WGAN(Wasserstein GAN)(Arjovsky 等，2017)、WGAN-GP(Gulrajani 等，2017)等，详细介绍可参考对应文献。

2.2.2 自编码器

自编码器(AE)(Vincent 等，2010)是一种数据压缩算法，其中数据的压缩和解压缩函数是数据相关的、有损的、从样本中自动学习的。自编码器的结构一般由两个部分组成：编码器和解码器。编码器和解码器可以是任意的模型，通常使用神经网络模型作为编码器和解码器。输入的数据经过神经网络降维得到一个编码，接着又通过另外一个神经网络去解码得到一个与输入原数据尽可能相似的生成数据。通过比较这两个数据，最小化它们之间的差异来训练这个网络中编码器和解码器的参数。当这个过程训练完之后，可以通过编码器生成一个更为紧凑的特征。另外，还可以利用解码器，随机输入一个编码，进而生成一个和原数据尽可能一致的数据。

变分自编码器(VAE)(Kingma and Welling，2013)是 AE 的改进，其结构与 AE 类似，也由编码器和解码器构成。对 AE 而言，输入一幅图像，通过编码器生成一个隐向量，这比随机取一个随机噪声更好，因为这包含着原图片的信息，最后可以将隐向量解码得到与原图片对应的图片。但是 AE 并不能生成任意图片，因为隐向量无法人工构造，需要通过一幅图像输入至编码器才能得到隐向量。为了解决该问题，VAE 应运而生。VAE 在编码过程中会增加限制，迫使其生成的隐向量能够大致遵循一个标准正态分布。因此，在 VAE 训练完成之后，对于生成新数据的任务，只需要给它一个服从标准正态分布的随机隐向量，通过解码器就可得到生成数据。在具体使用中，通常可以利用均值和标准差这两个统计量合成隐向量。这里默认编码之后的隐向量是服从一个正态分布的。通过 VAE 学习到的特征，可以应用于诸如识别、降噪、表示和可视化等任务中。

2.3 其他深度学习模型

其他代表性的深度学习模型有强化学习(Hester 等，2018)和元学习(Al-Shedivat 等，2018)。

强化学习(RL)作为机器学习的一种，已被广泛应用于人工智能领域，例如 Alpha Go、Atari 2600 游戏、自然语言处理等。它主要解决的问题是，对于一个可以感知环境的智能体，通过学习选出能实现目标的最优动作。在强化学习中应用最广泛的为 Q-learning，其通常需要定义一个 Q 函数 Q(s，a)，表示在状态 s 下采取动作 a 能够获得最大回报 R，然后通过迭代的方式不断更新 Q 值。如果 Q 函数足够准确且环境确定，那么只要采取实现最大 Q 值动作的策略即可。传统 Q-learning 将 Q 值存储在一个 Q 表格中，该表格的行表示不同的状态，列表示所有可能的动作。在状态不多的情况下，此方法可以很好地解决一些问题。但在现实中，通常会有近万个不同的状态，因而不可能建立如此大的 Q 表格，这使得 Q-learning 很难用来解决现实问题。因此，提出了 Deep Q-learning，其使用一个深度神经网络来对 Q 函数进行模拟；在 Q 值中使用均方差来衡量当前值与目标值之间的差异，然后将此差异作为目标函数并使用随机梯度下降来优化。为了提升 Deep Q-learning 的性能表现和学习速度，可进一步改进：使用 3 种损失更新网络，即双重 Q 学习损失、监督式大边际分类损失以及在网络的权重和偏置上的 L2 正则化损失，使得网络有更优的能力；通过时间差分更新方式，使得模型速率更高。

虽然强化学习已经在很多方面取得了令人瞩目的成果，但这些成果的一个共同问题是算法的处理受限于环境的稳定性。现实世界中环境通常是不稳定的，会随着时间推移而发生变化，往往会导致通过

强化学习所学到的策略失效，迫使智能体需在训练和运行期间不断调整自我以取得好的成果。为了解决上述问题，可以使用 Meta Learning 的方式对传统强化学习方法进行改进(Al-Shedivat 等，2018)，实现策略的动态自适应。该方法的主要思想是首先训练一个好的初始化网络，在面对新任务时只使用少量数据即可更新出一个适应新任务的网络；主要做法是使用之前的历史经验(如历史的策略和历史的轨迹)创建出新的策略。这样的方式是对于人类思维方式的模仿，即利用历史经验来调整策略，从而快速适应新环境。

3 基于深度学习的目标跟踪方法

基于深度学习的目标跟踪方法可以分别从网络结构、网络功能以及网络训练这 3 个不同的角度进行分类以及阐述。

3.1 按照网络结构分类

3.1.1 基于卷积神经网络的深度目标跟踪方法

卷积神经网络因较强的特征表征能力和泛化能力而广泛应用于目标跟踪中。然而，较之传统的手工特征，深度特征也存在一些不足，如特征维度较高、浅层特征语义信息较弱、深层特征位置信息较弱等。如何把深度特征和手工特征相结合来获取更好的特征描述已成为当前研究者比较关注的问题。Chi 等人(2017)提出 DNT(dual network based tracker)算法，是一个充分利用卷积神经网络的不同层进行特征提取而实现目标跟踪的双重网络。为了突出目标的几何轮廓，首先把卷积神经网络提取的级联特征和拉普拉斯高斯滤波得到的边缘特征整合为粗糙的先验图，再把双重网络的输出和边缘特征整合为混合成分，最后用参考独立成分分析算法得到精确的特征图。

“离线预训练 + 在线微调”是深度学习用于目标跟踪中的常用方法，但从大量离线训练数据中学习通用特征表示，存在耗时、特征针对性不强等问题。针对此问题，Zhang 等人(2016)提出 CNT(convolutional network based tracker)算法，采用一个轻型的两层卷积神经网络；该网络无需大量辅助数据离线训练就能学到较为鲁棒的特征。具体实现为：首先在第 1 帧中使用 k-means 算法从目标区域提取很多归一化的图像块作为固定滤波器，然后结合后续帧目标周围的一系列自适应上下文滤波器来构成特征图集合，最后用自适应阈值的软收缩算法对卷积后的全局表征进行去噪处理，从而得到鲁棒的稀疏表示特征 $\boldsymbol{c}$ 作为目标模板。目标模板的更新策略为

$$\boldsymbol{c}_t = (1-\rho)\,\boldsymbol{c}_{t-1} + \rho\hat{\boldsymbol{c}}_{t-1} \tag{1}$$

式中，$\boldsymbol{c}_t$ 和 $\boldsymbol{c}_{t-1}$ 分别为第 t 帧和第 $t-1$ 帧的目标模板，$\hat{\boldsymbol{c}}_{t-1}$ 为所跟踪的目标在第 $t-1$ 帧时的稀疏表示，ρ 为待学习参数。稀疏表示采用简单的在线更新策略来抑制跟踪器漂移，同时对目标形变更为鲁棒。

3.1.2 基于递归神经网络的深度目标跟踪方法

尽管基于卷积神经网络的目标跟踪方法已经取得了很多成果，但在时间连续性和空间信息建模方面还有待进一步改善，主要原因有：1) 每次只能对当前帧的跟踪目标进行建模，没有考虑当前帧和历史帧之间的关联性；2) 提取出来的深度特征往往随着网络层数的加深变得高度抽象，丢失了目标自身的结构信息；3) 池化操作会降低特征图的分辨率，损失了目标的空间位置和局部结构信息；4) 只关注目标本身的局部空间区域，忽视了对目标周边区域的上下文信息进行建模。近年来，递归神经网络尤其是带有门结构的 GRU(Cho 等，2014)、LSTM(Shi 等，2015)等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递归神经网络来解决现有跟踪任务中存在的问题。

目前大多数基于卷积神经网络的目标跟踪方法都是把目标跟踪作为分类问题处理，导致这些跟踪方法很容易受相似物体的干扰。针对该问题，Fan and Ling(2017)提出 SANet(structure-aware network)算法，引入递归神经网络来提取物体的自身结构信息，结合卷积神经网络来增强模型对相似物体的抗干扰能力。不同于 1 维的时序任务，2 维图像数据中物体的结构信息以无向循环图编码，循环结构使得无法直接应用 RNN 来提取结构信息，因此作者将无向循环图拓扑结构近似为 4 个有向不循环图的组合，如图 2 所示。考虑到卷积神经网络不同层是从不同的角度来提取目标特征，SANet 使用多个 RNN 对目标结构进行建模，提高了物体区分相似物体干扰和背景信息的能力。

MemTrack(Yang and Chan，2018)引入了具有外部存储功能的动态存储网络，通过更新外部存储单元来适应目标形状的变化，不需要高代价的在线网络微调。网络采用具有注意力机制的 LSTM 来控制

存储块的读写过程和模板的各通道门向量,有助于检索外部存储中最相关的模板。为避免模板更新策略所产生的过拟合问题,作者采用初始模板和门控残差模板相结合的方法来适应目标形状的变化,提高了跟踪性能。

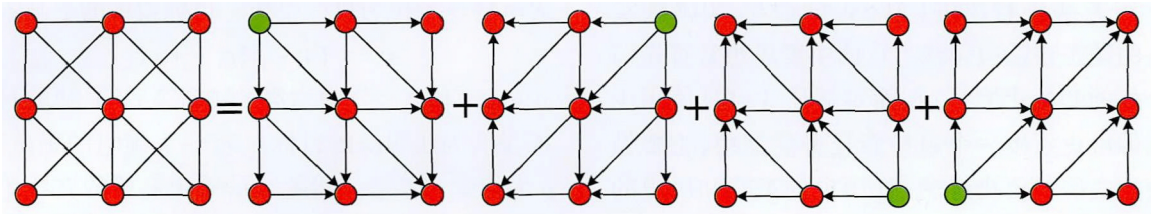


图2　无向循环图的分解

Fig. 2　The decomposition of undirected acyclic graph

3.1.3　基于生成式对抗网络的深度目标跟踪方法

基于深度分类网络的目标跟踪方法存在以下两个方面的问题:1)每一帧中的正样本空间上高度重合,不能获取丰富的表观信息;2)正负样本的比例严重不平衡。VITAL(visual tracking via adversarial learning)(Song 等,2018)通过对抗学习的方法来解决这两个问题。为了增强正样本对形变的鲁棒性,在最后1个卷积层和第1个全连接层之间引入对抗网络随机生成特征的权重掩码,每1个掩码表示一类具体的形变。通过对抗学习能够识别那些长期保留目标形变的掩码。为了解决各类之间的不平衡问题,引入高阶敏感损失函数降低易分负样本对分类网络的影响。SINT++(Wang 等,2018c)假设所有目标样本都位于一个流形空间上,使用变分自编码器生成大量与目标样本相似的正样本,解决了正样本多样性不足的问题。同时,通过深度强化学习用背景图片遮挡样本图片自动生成难区分正样本,解决了难区分正样本少的问题。

现有方法在利用深度神经网络进行目标跟踪时,或作为回归任务,或作为分类任务。受条件对抗网络启发,ADT(adversarial deep tracking)(Zhao 等,2019)使用将二者统一的深度对抗跟踪网络架构,网络由执行回归任务的全卷积孪生神经网络和执行分类任务的分类网络组成,整个网络可以通过对抗学习端到端地进行训练和优化。具体地,用目标块和搜索块组合的数据来训练回归网络,生成能反映目标在每个搜索块中位置和大小的响应图;然后,用目标的模板块、搜索块和响应图组合的数据训练分类网络选出匹配效果最好的响应图和对应的搜索块。跟踪阶段根据上一帧中确定的目标位置从当前帧抽取多个搜索块,回归网络根据搜索块和目标块生成多个响应图,分类网络选出最佳的响应图来确定目标位置。

3.1.4　基于自编码器的深度目标跟踪方法

现有基于深度学习的目标跟踪方法难以满足在线实时跟踪的需求,针对该问题,TRACA(context-aware deep feature compression for high-speed visual tracking)(Choi 等,2018)基于上下文感知的机制选择专家自编码器对深度特征进行压缩,是一种速度快且精度高的基于相关滤波的目标跟踪方法。预训练阶段针对每类目标训练一个自编码器,跟踪阶段根据给定目标选择最佳的专家自编码器进行跟踪。为了在压缩后的特征图上达到更好的跟踪效果,在专家自编码器中引入去除噪声处理和正交损失函数。

基于CNN的目标跟踪方法存在如下局限:1)语义嵌入空间的特征通常分辨率比较低,丢失了实例的细节信息。这些特征主要是针对分类任务学习的,一方面无法明确区分具有相同属性或语义的两个目标,另一方面跟踪器遇到没见过的类别或者目标形状变化较大时容易发生漂移。2)为了提高跟踪速度,网络通常不进行在线更新,不可避免地影响模型的自适应性和跟踪准确性。为了解决上述限制,EDCF(enhanced distributed coordination function)(Wang 等,2018b)使用一种端到端的编解码网络,采用多任务学习策略以相互增强的方式优化相关分析和图像重建,增强跟踪的鲁棒性和自适应性。除了普通的分类损失项之外,解码器在语义嵌入空间引入了重建约束,使得语义表示能够重建成原始图像,缓解了由特征表示分辨率较低所引起的跟踪漂移,保证语义嵌入层能够充分保留最初视觉特征的几何和结构信息,增强了跟踪器的泛化能力和

分辨能力。

3.2 按照网络功能分类

3.2.1 基于相关滤波的深度目标跟踪方法

相关滤波通过构造一个滤波器,与视频帧进行互相关操作,得到一个响应图,其中最高的值指示了目标所在的位置。这一方法充分利用空域卷积可以转换为傅里叶变换域中元素与元素的乘积的理论,极大地降低计算复杂度。给定一帧图像 x 和响应图 g,通过学习一个过滤器 h,使得

$$g = x \odot h^* \tag{2}$$

式中,* 是复共轭操作,$\odot$ 是卷积操作。通过将卷积应用于复共轭的过滤器,实现了互相关操作。为了高效地计算上述方程,分别计算出 x 和 g 的傅里叶变换 X 和 G,然后可以计算得到 H^*,即

$$H^* = X/G \tag{3}$$

式中,除法是元素操作,h 是通过 H^* 计算得到的。

对于目标跟踪任务,相关滤波器从第 1 帧提取的目标区域中训练得到,后续再进行更新。当有新的视频帧时,相关滤波器与之进行互相关操作,得到的最大响应的位置即代表目标的新位置。HCF(Ma 等,2015)通过结合多层 CNN 特征,利用相关滤波来定位被跟踪的目标;其针对每层 CNN 训练一个过滤器,并且按照从深到浅的顺序使用相关滤波,利用深层得到的结果来引导浅层,从而减少搜索空间。CFNet(Valmadre 等,2017)将相关滤波设计成一个可微分的层,采用端到端方式训练网络,提取适用于相关滤波器的特征。该工作基于 SiamFC(Bertinetto 等,2016b)实现,首先设计了两个分别代表当前帧和目标模板的分支,然后通过模板特征和当前帧特征之间的互相关确定目标位置。FlowTrack(Zhu 等,2018b)是一个带有可微分相关滤波层的 Siamese 网络,使用多个之前的视频帧作为模板,结合时空注意力模块计算不同位置不同模板特征的权重,最终确定目标位置。与上述采用离散滤波器的方法不同,C-COT(Danelljan 等,2016)使用连续卷积滤波器进行目标跟踪,ECO(Danelljan 等,2017)针对 C-COT 的过拟合和采样存储问题,使用少量的滤波器参数替代原来大量的过滤器参数,并结合高斯混合模型减少存储的样本数量以及保持样本的差异性。DRT(correlation tracking via joint discrimination and reliability learning)(Sun 等,2018)在 ECO 的基础上引入了稳定性概念,对滤波器的每一部分引入一个权值,由此决定是否使用它进行跟踪。通过构造一个与过滤器大小相同的矩阵,在使用滤波器前与之相乘,最终使得滤波器不可靠部分数值较小,从而提升跟踪精度。

3.2.2 基于分类网络的深度目标跟踪方法

基于分类网络的深度目标跟踪方法通常需要多步完成,首先在目标可能存在的位置产生大量候选框,接着通过分类网络对所有的候选框评估,给出相应的分值,最后所有的候选框都根据得到的分值进行排序,分数最高的候选框就作为目标所在的位置。影响基于分类网络的深度目标跟踪方法性能的主要问题是候选框产生的方式以及数量,产生的过少可能无法包括目标,过多又会影响算法效率。此外,分类网络的质量也直接影响最终的效果,错误的分类将会导致跟踪出现错误。如图 3 所示,基于分类网络的深度目标跟踪方法在当前视频帧中产生若干候选框后分类网络进行二分类,得到前景和背景集合,最后根据分数进行排序,得到最终的目标。

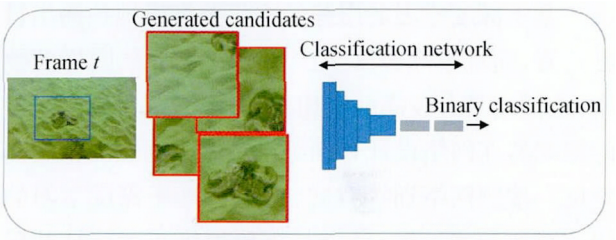


图 3 基于分类网络的深度目标跟踪方法

Fig. 3 Classification network-based deep visual tracking method

MDNet(Nam and Han,2016)提出了一个多域的网络框架,将一个视频序列视为一个域,其中共享的部分用来学习目标的特征表达,独立的全连接层则用于学习针对特定视频序列的二元分类器。VITAL(Song 等,2018)介绍了一个基于生成对抗网络的目标跟踪方法,核心思想是采用 GAN 产生一个权重掩码以选择有判别力的特征,通过掩码与特征图的乘积实现分类。ADNet(Yun 等,2017)是一个与增强学习相结合的目标跟踪算法,增强学习策略网络是通过 CNN 构建的。

3.2.3 基于回归网络的深度目标跟踪方法

基于回归网络的方法通常在之前目标所在的位置基础上,使用一个前向网络,直接回归目标所在的位置。和其他方法相比,基于回归网络的深度目标跟踪方法易于实现,速度较快,且可利用线下和线上

的训练。通常使用卷积神经网络来构建回归网络，将密集的数据特征 X 映射到连续的输出矩阵 Y，通过搜索输出响应图中最大值的位置来估计目标移动到的位置，公式为

$$\arg\min_{W}\|W\cdot X-Y\|^2+\gamma\|W\|^2 \tag{4}$$

式中，W 是卷积核的权值参数，γ 是正则系数。

DSLT(deep regression tracking with shrinkage loss)(Lu 等，2018)使用回归网络将样本映射为一个软标签图，即响应图。然而，前景和背景目标数量上的不平衡会影响回归学习的质量。因此，作者提出了结合收缩损失的回归方法。GOTURN(generic object tracking using regression networks)(Held 等，2016)假定在连续视频帧间目标移动较为缓慢，使用带有两个卷积层分支的神经网络，一个是之前视频帧包含的目标区域，另一个是在当前视频帧中包含的以上一帧目标所在位置为中心一定范围内的区域，二者在全连接层进行融合，回归出目标所在的位置。

3.3 按照网络训练分类

基于深度学习的跟踪器不仅要准确地预测出目标位置，而且跟踪速度至少要达到与视频同样的帧率，这样才具有较高的实用性。因此，基于深度学习的跟踪器在网络设计和训练上需要平衡预测精度和速度。按照网络训练方式，可以将基于深度学习的跟踪算法归为三类：完全基于预训练深度特征的跟踪器；基于离线训练和在线微调结合的跟踪器以及完全基于离线训练网络的跟踪器。

3.3.1 基于预训练网络的深度目标跟踪方法

传统的跟踪方法使用人工特征描述目标，如 HOG、CN 等，而基于预训练特征提取网络的深度目标跟踪方法直接使用预训练网络提取目标特征，从而替代传统人工设计的特征，然后再使用传统的分类(如 SVM)或回归方法预测目标。较之手工特征，深度特征因其表达的丰富性以及平移不变性，可以大大提高跟踪性能。此外，深度特征(特别是浅层特征)具有很强的通用性，在其他视觉任务上学习到的特征可以迁移到目标跟踪中。具体地，该方法一般使用特征通用性较强的图像分类网络(如 AlexNet(Krizhevsky 等，2012)、VGGNet(Simonyan and Zisserman，2015)等)，可以在海量的图像数据集(如 ImageNet)上进行预训练。采用这种方法有很多优点：一方面，完全使用预训练网络，节省了大量的训练时间，而且无需在线更新网络，可以取得较快的跟踪速度；另一方面，预训练网络可以利用大量图像分类的标注数据，解决了目标跟踪中训练样本不足的问题。但是缺点也很明显：一方面，其他任务上训练的网络不能完全适应目标跟踪这个任务；另一方面，目标在跟踪过程中会发生很多形变，网络不更新无法学习到目标的变化信息。

在 HDT(Qi 等，2016)中，使用预训练的 VGGNet 的不同卷积层特征来表征目标，针对每一个卷积层的特征，构建一个基于鉴别相关滤波器(DCF)的弱跟踪器，然后使用集成方法将多个弱跟踪器关联成强跟踪器，从而提高跟踪性能。DeepSRDCF(Danelljan 等，2015b)同样地将深度卷积特征应用到 DCF 跟踪框架中。与 HDT 构建多个跟踪器不同，DeepSRDCF 将预训练 VGGNet 的不同卷积层特征进行线性融合来表征目标。实验表明该方法可以在多个跟踪数据集上取得较好的结果。

3.3.2 基于在线微调网络的深度目标跟踪方法

基于在线微调网络的深度目标跟踪方法结合离线训练和在线微调来更好地表征目标以适应目标变化，网络架构通常特征提取和目标检测两部分。特征提取部分采用预训练的网络进行初始化。在跟踪开始时，首先用第 1 帧的标注样本训练目标检测部分和微调特征提取部分。跟踪过程中，根据预测结果生成一定的正、负样本，然后微调整个网络，进一步提高网络的判别能力，较好地适应目标的变化，显著提高跟踪性能。但是，由于采用在线微调，跟踪速度受到很大影响，一般很难达到实时要求。

在 MDNet(Nam and Han，2016)中，作者采用了特征提取和多分支检测结合的网络结构。在离线训练时，针对每个视频序列构建一个新的检测分支进行训练，而特征提取网络是共享的。这样特征提取网络可以学习到通用性更强的与域无关的特征。在跟踪时，保留并固定特征提取网络，针对跟踪序列构建一个新的分支检测部分，用第 1 帧样本在线训练检测部分。之后再利用跟踪结果生成正负样本来微调检测分支。为了保证跟踪速度，可以每间隔一段时间才微调网络。CREST(convolutional residual learning for visual tracking)(Song 等，2017)是一个端到端的在线学习跟踪网络，其使用 VGG-16 作为目标特征提取网络，然后使用 DCF 来检测目标，其中 DCF 通过一个网络卷积层来实现。此外，为了拟合

响应真实值，采用残差学习来逼近真实值和网络输出相应值之间的残差，从而可以更好地获得目标的表观变化。

3.3.3 基于离线训练网络的深度目标跟踪方法

上述基于在线微调网络的深度目标跟踪方法会使跟踪器的效率大大降低，深度特征的提取和更新很难做到实时。为解决这一问题，提出基于离线端到端训练的全卷积孪生网络的跟踪方法 SiamFC（Bertinetto 等，2016b）；其跟踪速度在 GPU 上可以达到 86 帧/s，而且其性能超过了绝大多数实时跟踪器。SiamFC 主要学习相似度函数，用于目标匹配。孪生网络分别输入初始帧模板以及当前帧的搜索区域，分别使用相同的全卷积网络提取特征，再用相关操作进行模板匹配，生成响应图。响应图中最大值的位置即是目标在搜索区域内的相应位置。网络训练时，SiamFC 采用 ImageNet VID 的视频数据，选取视频中相隔不远的两帧输入网络进行相似度函数学习；跟踪时，训练好的网络无需调整，目标模板也无需更新，从而实现实时跟踪。SiamFC 提出后受到了很多关注，很多跟踪方法都在其基础上进行改进。

由于 SiamFC 网络主要关注外观特征而忽略了高层语义信息，SA-Siam（He 等，2018）采用融合表观特征和语义信息的双重孪生网络跟踪方案，其中一个孪生分支负责表观特征匹配，另一个负责语义信息的匹配。表观分支还是使用原本的 SiamFC，而语义分支则采用 ImageNet 图像分类中训练的网络，并且不进行更新。在训练时，两个分支网络独立训练，以实现互补效果。另外，在语义分支中，还引入了一个空间和通道上的注意力模块，突出跟踪的目标，弱化背景和非目标。SA-Siam 引入语义信息使得跟踪器更加稳定，不易受目标表观变化的影响。

与 SiamFC 和 SA-Siam 采用检测网络的方法不同，GOTURN（Held 等，2016）采用基于孪生网络的回归方法，学习目标表观和运动的变化关系。输入两幅包含目标的图像，GOTURN 首先经过共享参数的孪生网络提取特征，然后回归网络能够比较两幅图像回归出目标的位置，其跟踪速度可以达到 100 帧/s。

4 其他深度目标跟踪算法

4.1 基于分类与回归相融合的深度目标跟踪方法

此类方法将目标跟踪问题转化为目标检测问题，通常采用经典的卷积网络提取目标特征以及区域生成网络（RPN）辅助目标定位。在目标定位的过程中，采用两种类型的子网络，一是分类网络预测前景与背景信息，预测最有可能的目标大致区域，二是位置回归网络对目标区域进行精确的定位预测。代表性的工作有 Siamese-RPN 网络（Li 等，2018a），其网络结构包括特征提取的 Siamese 子网络和产生候选目标区域的 RPN 子网络。Siamese 子网络的输入包括模板帧和检测帧，RPN 子网络则分为分类和回归两个子模块。具体地，首先利用 Siamese 子网络提取跟踪目标的特征作为模板，以待检测位置为中心生成 k 个区域；利用分类模型判断 k 个区域是否为待检测目标，同时用回归模型计算 k 个区域的中心点坐标及目标宽高；最后利用余弦窗和尺度变化惩罚策略对 k 个区域的得分进行排序，通过非极大抑制选出得分最高的框作为跟踪目标的预测框。利用 ILSVRC 和 Youtube-BB 数据集进行离线训练，在 VOT2015、VOT2016 和 VOT2017 数据集上分别进行测试，该方法的速度达到 160 帧/s 的同时获得了先进的跟踪性能。最近提出的 DaSiamRPN 方法（Zhu 等，2018a）对 Siamese-RPN 进行进一步优化和改进，以着重处理训练数据不平衡、自适应的模型增量学习及长程跟踪等问题。在训练阶段采取样本增强策略，利用现有的目标检测数据集（如 ImageNet 检测集和 COCO 检测集）扩充正样本数据，以此提升目标跟踪器的泛化能力，并显式地增加不同视频段同类样本以及不同类样本作为负样本，以此提升目标跟踪器的判别能力。在相似性度量阶段采取非极大抑制选择难以区分的错误样本从而实现更加有效的增量式学习。针对长程跟踪和目标消失问题，该方法提出从局部到全局的搜索策略来检测得分判断目标是否丢失，再进一步实现目标重检测。因此，该方法在目标遮挡和长程跟踪等情况下展示出优越的性能。尽管该类方法在公开数据库上取得了很好的性能，但需依赖于额外的大规模训练数据来保证所训练跟踪模型的鲁棒性。

4.2 基于强化学习的深度目标跟踪方法

该类方法将强化学习的决策策略引入到目标跟踪任务中，以优化深度网络的参数、网络深度、或预测目标移动状态等信息。近期，ADNet（Yun 等，2017）采取马尔可夫决策过程（MDP）的基本策略，将目标移动定义为离散化的动作，特征以及观察的

历史状态形成当前状态，认为目标跟踪是一系列动作预测和状态变化的过程。在学习过程中采用深度神经网络并使用基于矩形框交并比的奖惩机制。在训练阶段，分为监督学习和强化学习两个阶段。在监督学习阶段利用视频序列优化目标位移及尺度变化等动作；在强化学习阶段利用监督学习阶段训练的网络作为初始化，然后采取包含采样状态、动作、激励在内的训练序列进行跟踪仿真。随后，Dong 等人(2018)提出了一种基于超参数优化的深度连续 Q-learning 方法，以解决在线目标跟踪中不同视频的模型超参数适应问题；Huang 等人(2017a)认为跟踪目标的困难程度所依赖的特征复杂度不同，提出了一种自适应的决策过程以学习一个 agent 来决定采取浅层或更深层的特征，有效地提升了目标跟踪的速度；Supančič and Ramanan(2017)针对目前目标跟踪数据的标注困难问题提出了一种弱监督的深度强化学习算法，仅需要在训练过程中标定是否奖励或惩罚而不需要详细的目标框标注，也可以处理部分标注的情况(即形成部分可观察的马尔可夫决策过程)。另外，最近一些研究者基于强化学习中的 Actor-Critic 框架提出了相应的目标跟踪算法(Chen 等，2018；Ren 等，2018)。Actor 网络利用深度网络优化目标位置，Critic 网络计算预测框的得分并反馈至 Actor 网络，从而根据反馈信息更新模型。相比于传统的深度跟踪算法，该类方法不仅可以较好地自适应于新的环境，而且由于模型推理的候选目标框数量少能够提升目标跟踪的速度。

4.3 基于集成学习的深度目标跟踪方法

该类方法的主要策略是先通过一定的规则生成多个分类器，然后采用某种集成策略优化组合，最后综合判断输出最终的目标跟踪结果。早在 Avidan(2007)采用 AdaBoost 加权线性合并多个弱跟踪器来构造一个强跟踪模型。之后，boosting、在线 boosting、多类 boosting 以及多示例 boosting 等技术用来实现集成跟踪器。尽管 boosting 技术能够应用于目标跟踪任务，但其对标签噪声较为敏感。最近，基于深度神经网络的集成跟踪方法引起了许多学者的关注。Nam 等人(2016)提出了一种基于多个卷积神经网络树形结构化的目标表观模型方法。多个卷积网络协同估计目标状态，并通过优化基于树形结构的子模型实现路径的更新。为了节省存储空间和避免冗余的计算，多个卷积网络采取共享底层卷积参数的策略。相似地，Han 等人(2017)采用卷积层共享而全连接层多分枝的集成方法，基于经典的 dropout 方法，在跟踪过程中选取各个分枝模型时采用随机策略，以便增加子跟踪模型的差异性同时避免过拟合问题。从特征表示的角度，Wang 等人(2018a)利用跟踪目标的不同特征来学习相应的判别式相关滤波跟踪专家，然后对专家之间以及专家自身进行评价选择合适的专家进行目标跟踪以及模型更新。Wang 等人(2016)对共享的卷积网络特征谱的每个通道训练一个基学习器。为了降低学习器的相关性以及避免过训练问题，每个基学习器采用不同的损失函数。尽管这些方法尝试利用特征或损失函数增大各子模型的差异性，但每个子模型之间仍然存在过多的冗余信息。对此，部分学者从数据采样的角度缓解此问题，Meshgi 等人(2017)提出了一种基于委员会学习的跟踪方法，每个跟踪器根据训练数据的分布进行采样以使得不同跟踪器之间采用的样本具有差异性。

4.4 基于元学习的深度目标跟踪方法

该类方法利用元学习对目标跟踪模型自适应地优化，使得模型快速地适应于不同视频序列或场景。Learnet 方法(Bertinetto 等，2016a)将目标跟踪模型定义为模板上的动态参数化函数，以便处理在线跟踪时单样本学习的情况。因此，该方法遵循学会学习(learning to learn)的基本思想，让跟踪模型自身根据周围环境定义判别决策。MLT(meta learning for real-time visual tracking)方法(Choi 等，2017)采用梯度预测的策略自适应更新网络参数，采用参数化网络梯度的方法学习网络模型，从而构建了一个元学习网络。此外，也借鉴了经典的 Siamese 匹配网络估计跟踪目标的位置。类似地，Meta-tracker 方法(Park and Berg，2018)也采用基于预测梯度的策略学习方法获得普适性的初始化模型，可以使得跟踪模型自适应于后续帧特征的最佳梯度方向。该方法引入了两个待学习参数：初始化参数 θ_0 和梯度更新参数 α。目标跟踪的元训练过程主要分为两步：1)随机初始化参数 θ_0，将第 1 帧图像输入跟踪模型进行预测，利用预测误差函数以及梯度更新参数 α，反复迭代 T 次作为 θ_1；2)检查参数 θ_1 对后续帧(每次迭代随机取一帧)的鲁棒性，累积损失函数对 θ_1 和 α 的梯度，采用 ADMM 梯度下降算法优化参数 θ_0 和 α。作者将这种思路推广应用于 MDNet 和

CREST 方法,实验表明提出的方法在目标跟踪的速度和精度都有提升。总体上,该类方法尝试研究如何更好地更新模型以使得其自适应于不同的跟踪场景。此外,较之同类的深度目标跟踪方法,该类方法在模型更新环节仅仅需要少量的迭代次数,所以可以有效地提高跟踪速度。

5 数据库与评价标准

在目标跟踪算法日益完善的同时,用于评估各种算法的数据库和评价指标也在日益完善。早期用于算法评估的视频数据集,如 VIVID(Collins 等,2005)、CAVIAR(Fisher,2004)等,专门针对监控场景中的视觉跟踪,目标通常是静止背景下的人或车,数据规模小且大部分目标没有标注。单一的、小规模的且标签缺失的数据集,不仅无法全面衡量算法性能,而且未标注的序列给评估带来了巨大挑战。因此,为进一步推动深度学习在目标跟踪任务中的应用,很多大型数据集陆续建立,评价指标也日趋完善。

5.1 适用于深度学习目标跟踪的视频数据库

2013 年,Wu 等人(2013)建立了一套较为全面的数据库和标准 OTB50 来评估目标跟踪算法。该数据集由 50 个完全标注的视频序列组成,共包含 51 个不同尺寸的目标,总计超过 29 000 帧图像。由于目标在跟踪过程会受到各种干扰因素的影响,为了全面评估跟踪算法在各种因素下的鲁棒性,OTB50 提供了 11 种常见的视频属性标注,即光照变化、尺度变化、遮挡、形变、运动模糊、快速运动、平面内旋转、平面外旋转、移出视野、背景干扰和低分辨率,每一帧图像中至少含有 2 种标注属性。此外,OTB50 数据集整合了 29 个流行的跟踪算法并且统一了输入输出格式以便于大规模的算法性能评估。2015 年,作者进一步将数据集扩展为 100 个视频序列 OTB100(Wu 等,2015),并从中选出 50 个跟踪难度较大的视频构成 TB50。OTB(Wu 等,2013; Wu 等,2015)数据集的出现有助于跟踪算法性能评价的标准化,同时极大地促进目标跟踪方向的研究,至今仍广泛应用于跟踪算法的性能评估。

OTB 数据集的出现也促进了其他视频数据集的发展。为了评估基于颜色信息的目标跟踪算法,2015 年 Liang 等人(2015)建立了 TempleColor128 数据集。该数据集共包含 128 个彩色视频序列,部分序列与 OTB 数据集重合。ALOV + +(Smeulders 等,2014)数据集从 YouTube 网站搜集了 315 个视频序列,共包含 64 种不同类型的跟踪目标,旨在尽可能地囊括现实世界中存在的各种干扰因素,如亮度变化、相似物干扰、遮挡等各种情况。UAV123(Mueller 等,2016)数据集采用无人机以低空的航拍角度拍摄彩色图像,包含 123 个高清序列,共计超过 110K 帧的视频标注,并标注了 12 种视频属性。

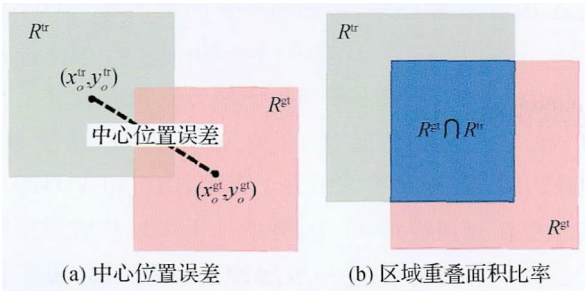
除上述数据集外,一些视频数据库以目标跟踪竞赛的形式也得到了充分发展。NUSPRO(Li 等,2016)竞赛数据集由 YouTube 网站搜集的 365 个视频序列构成并提供一个在线评测系统。该数据集共包含 5 种类别:人脸、行人、运动员、刚性物体和长视频序列,其中每种类别又被细分为 5 ~ 6 种子类,最终形成包含 17 种目标物体的数据集。为分析不同跟踪算法的优缺点,NUSPRO 提供了 12 种干扰目标跟踪的属性标注。近年来,VOT 竞赛数据库受到越来越多的关注。VOT 竞赛从 2013 年开始举办,至今已连续举办 6 届。VOT2013(Kristan 等,2013)仅包含 16 个视频序列,影响力不及同期出现的 OTB50(Wu 等,2013)。VOT2014(Kristan 等,2015b)将视频增至 25 个,并采用多边形区域方式重新标注了样本,较之 OTB 数据集的轴对齐标注更准确。VOT2015(Kristan 等,2015a)、VOT2016 和 VOT2017 进一步扩充视频到 60 个,并增加了 TIR 热成像跟踪子系列。VOT2018 进一步增加了长程目标跟踪任务的新挑战。

ECCV 2018 提出了两个针对深度学习目标跟踪算法的大规模数据集:TrackingNet(Müller 等,2018)和 Long-term Tracking in the Wild(Valmadre 等,2018)。TrackingNet 从 YouTube 视频中进行采样,专门为目标跟踪问题而设计,更接近真实世界中的目标跟踪任务。该数据集包含 30 000 + 个视频,囊括了不同类别的跟踪目标,共计 14 200 000 个标注框。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。此外,TrackingNet 包含各种序列长度的视频,可用来评价短程和长程的目标跟踪算法。Long-term Tracking in the Wild 是专门针对长程目标跟踪算法构建的数据库,共包含 14 h 时长的 366 个视频序列,其中每个视频的平均时长超过 2 min,并带有频繁的目标消

失，增加了跟踪的难度。不同于 TrackingNet，该数据库虽然将数据划分为包含 200 个视频的训练数据集和包含 166 个样本的测试数据集，但是测试数据集的标注是不提供的，作者提供了一个用于评价目标跟踪算法的在线服务器。

5.2 适用于深度学习目标跟踪的评价标准

目标跟踪旨在在给定目标初始位置和大小的前提下，预测目标在后续各帧中的位置和大小。因此，其评价标准通常包含两个基本参数：中心位置误差和区域重叠面积比率，如图 4 所示。在跟踪精确度的评估中，中心位置误差是一个广泛使用的标准，是跟踪目标的中心位置（$x_0^{\mathrm{tr}}, y_0^{\mathrm{tr}}$）和人工标注的准确位置（$x_0^{\mathrm{gt}}, y_0^{\mathrm{gt}}$）之间的平均像素距离。通常，采用一个序列中所有帧的平均中心位置误差来评价跟踪算法对该序列的总体性能。然而，当跟踪器丢失目标时，预测的跟踪位置是随机的，此时平均误差值可能无法准确评估跟踪器的性能。因此，在 OTB 数据集上将其进一步扩展为精确度曲线图，统计在不同阈值距离下的成功跟踪比例，并采用阈值为 20 个像素点所对应的数值作为代表性的精确度评价指标。



(a) 中心位置误差　(b) 区域重叠面积比率

图 4　评价指标

Fig. 4　Evaluation criterion

((a) distance of center point; (b) intersection over union)

由于中心位置误差无法评价目标在跟踪过程中的尺度变化，因此研究者提出基于区域重叠面积比的评价标准，通过统计跟踪算法得到的边界框和人工标注的准确边界框之间的面积重叠比衡量跟踪算法的性能。区域重叠面积比定义为算法预测边界框区域 R^{tr} 和人工标注框区域 R^{gt} 的交集与并集之比为

$$S = |R^{\mathrm{gt}} \cap R^{\mathrm{tr}}| / |R^{\mathrm{gt}} \cup R^{\mathrm{tr}}| \tag{5}$$

OTB 数据集对该参数进行了扩展，在判定重叠率大于给定阈值即跟踪成功的前提下，统计了不同阈值下跟踪成功的帧数占视频总帧数的比例作为成功率，最终绘制了阈值从 0 到 1 变化的成功率曲线图，并使用曲线下面积（AUC）作为成功率评价指标。此外，基于区域重叠面积比评价标准，学者们提出了一系列衡量跟踪算法鲁棒性的评价指标，例如，根据区域重叠面积比，记录跟踪失败次数；在此评价标准上，进一步扩展的评价标准有失败率，其定义为

$$\mathrm{Fr}(F_\tau) = \frac{1}{\log F_\tau} \sum_{f_i \in \mathcal{F}_\tau} -\frac{\Delta f_i}{N} \log \frac{\Delta f_i}{N} \tag{6}$$

$$\Delta f_i = \begin{cases} f_{i+1} - f_i & f_i < \min(\mathcal{F}_\tau) \\ f_1 + N - f_i & f_i = \max(\mathcal{F}_\tau) \end{cases} \tag{7}$$

式中，f_i 为跟踪失败的位置，F_τ 为跟踪失败的次数。当交叠面积低于阈值 τ 时判定为跟踪失败，并重新进行初始化，同时记录跟踪失败的次数 F_τ 和跟踪失败的位置 f_i，每一段的跟踪长度越短，失败率越大。此外，也有一些评价标准将中心位置误差和区域交叠面积比进行结合

$$CoTPS = (1 - \lambda)(1 - \hat{\phi} + \lambda_0^2) \tag{8}$$

式中，$\hat{\phi}$ 为跟踪成功帧的平均重叠率，λ_0 是失败帧所占比例，该指标综合考虑了精确度和鲁棒性，得分越高越好。

中心位置误差和区域重叠面积比是目标跟踪算法评估中两个最基本的度量指标，并在 OTB（Wu 等，2013; Wu 等，2015）数据集中进一步扩展为精确度曲线图和成功率曲线。TempleColor128（Liang 等，2015）、UAV123（Mueller 等，2016）等数据集都沿用了 OTB 提供的评价指标。此外，为进行鲁棒性评估，OTB 提出在时间上（即从不同帧开始跟踪）和空间上（即以不同的边界框开始跟踪）扰乱初始化，以模拟现实世界中由于位置或尺寸方面引入的初始化误差；这两种评估称为时间鲁棒性评估（TRE）和空间鲁棒性评估（SRE）。基于中心位置误差和区域重叠面积比的评价标准是对每个视频序列进行独立的性能评估，为避免单个视频的影响，ALOV + +（Smeulders 等，2014）采用存活曲线来衡量目标跟踪算法在 315 个视频序列上的整体性能，并采用 F-score 进行评估。VOT 竞赛采用了类似的评价标准，提出了准确率与鲁棒性两个基本评价指标，并将其结合为 EAO（expect average overlap）作为整体性能评价指标。准确率即跟踪成功状态下的平均重叠率，而鲁棒性则用跟踪失败总次数来衡量。

现有的深度学习跟踪方法大多采用 OTB 数据集中提出的精确度曲线图和成功率曲线图以及 VOT 竞赛中提出的 EAO 指标。在 ECCV2018 新提

出的大规模视频数据库上,TrackingNet 为了减小目标尺寸对精确度指标的影响,提出了正则化的精确度指标 P_{norm},定义为

$$P_{\text{norm}} = \| W(C^{\text{tr}} - C^{\text{gt}}) \|_2, W = \text{diag}(R_x^{\text{gr}}, R_y^{\text{gr}}) \tag{9}$$

式中,C^{tr} 是预测得到的目标中心位置,C^{gt} 为人工标注的目标中心位置,R_x^{gt} 和 R_y^{gt} 分别为人工标注框的宽度和高度。Long-term Tracking in a Wild 数据集是针对长程视频跟踪任务提出,此类视频面临的最大挑战为目标遮挡和消失视野,因此,该数据集提出一种新的评价指标,来预测每一帧中是否存在目标。类似于二分类问题,作者定义了 TN(true negative)、FP(false positive)、TP(true positive) 和 FN(false negative),并且采用 TPR(true positive rate) 和 TNR(true negative rate) 作为衡量目标跟踪性能的指标。TPR 指在包含目标的帧中被成功预测和正确定位的帧占包含目标的总帧数的比率,TNR 指不包含目标的帧中被成功预测的帧数占不包含目标的总帧数的比例。整合 TPR 和 TNR,还提出了几何均值 $\text{GM} = \sqrt{\text{TPR} \cdot \text{TNR}}$ 作为整体的性能评价指标。

基于上述提出的各种基本评价指标和综合评价指标,为分析不同跟踪算法的特性,研究者进一步提出了针对数据集的目标跟踪方法评价系统(Kristan 等,2016)。它构建了一个包含 25 个视频序列的逐帧标注属性的数据集,并针对该数据集提出了一种简单、易解释的评价方法,即根据性能对目标跟踪算法进行排序。性能主要考虑两方面,一是基于区域重叠面积比的精确度;二是基于失败次数的鲁棒性。该系统通过对数据集的统计分析来解释跟踪算法的性能。

6 应用实例介绍

随着公开的测试图像与视频数据库的不断增多,基于深度学习的目标跟踪技术在实际系统中的应用也取得了较快的进展。下面给出一些应用实例对这一方面进行介绍。

6.1 移动端目标跟踪系统简介

在移动终端中,对人脸实时准确的跟踪是许多应用程序前端输入的前提,只有准确跟踪人脸的位置,才能以更高的精度完成“肤色美白”、“人脸替换”、“虚拟发型”等有趣的功能,如美颜相机、FaceU 激萌、开心魔法、B612 等移动终端的应用程序。

此外,还建立了专门用于移动终端人脸跟踪算法测试的视频数据平台 iBUG。在该平台下,当前主流的目标跟踪算法的跟踪性能均出现不同程度的下降,但基于深度学习的目标跟踪算法依然具有明显的优势,如 DVNet(Schroff 等,2015)、ECO(Danelljan 等,2017) 和 MDNet(Nam and Han,2016) 等方法,其中排名第一的方法 DVNet 是谷歌公司提出的 FaceNet。FaceNet 最先提出时用于人脸识别,但也可以用于目标跟踪领域中,其模型结构如图 5 所示。

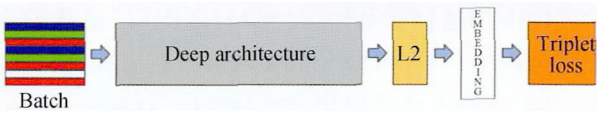


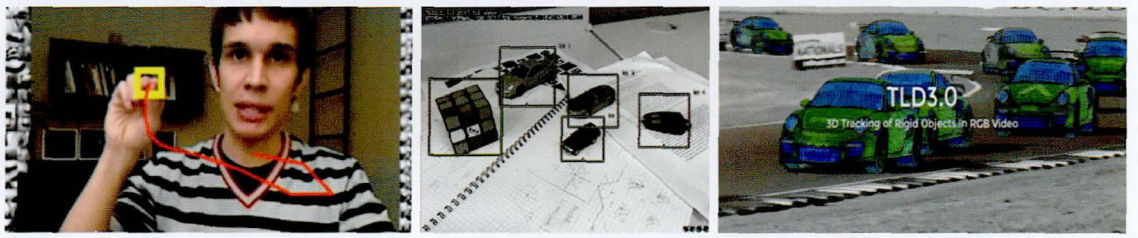
图 5 FaceNet 模型基本结构示意图

Fig. 5 An example of FaceNet model

图中的 Batch 为输入图像,中间的 Deep architecture 对图像提取特征矩阵,然后通过 L2 范数进行归一化,再嵌入成 128 维特征,通过三元组损失(Triplet loss) 进行训练,保证同一身份的目标和不同身份的目标之间的差距足够大。FaceNet 可以直接输入人脸图像来得到特征向量,是一个端到端的系统,不需要人为添加额外的处理(即自身有很强的泛化能力,对于光照、拍摄角度均可实现同一身份目标的高聚类低耦合)。该模型采用两种深度学习网络进行构建,一种是 Z-F Net(Zeiler and Fergus,2014),另一种是 GoogLeNet(Szegedy 等,2015)。在 LFW(labeled faces in the wild) 数据库上用两种方式进行了验证,直接取 LFW 图片的中间部分进行训练,分类准确率达到了 98.87% 左右;若使用额外的人脸对齐工具,分类准确率能达到 99.63% 左右,超过了 DeepID(Sun 等,2015)。

6.2 联合检测与跟踪的长时目标跟踪系统简介

对目标跟踪而言,常用的方法有两种:一种是使用跟踪器根据目标上一帧的位置预测它在下一帧的位置,但这样会积累误差,并且目标一旦在图像中消失,跟踪器就会永久失效,即使物体重新出现也无法完成跟踪;另一种方法是使用检测器对每一帧图像单独检测目标的位置,但这需要提前对检测器进行离线训练,而且只能用来跟踪事先已知的目标。实际的目标跟踪系统期望对视频中的未知目标进行长



(a) TLD1.0跟踪结果　(b) TLD2.0跟踪结果　(c) TLD3.0跟踪结果

图 6　TLD 三个版本的跟踪结果示意图

Fig. 6　Visual tracking results from the three version of TLD method

((a) TLD1.0; (b) TLD2.0; (c) TLD3.0)

时间跟踪，“未知目标”是指任意目标，即在跟踪开始之前不知道目标是什么，“长时间跟踪”意味着在跟踪过程中，目标可能会消失后再重新出现，而且随着光照、背景变化和偶尔的部分遮挡，目标的外观可能会发生很大变化，在这些情况下，依然能够检测到并跟踪目标。因此，单独使用跟踪器或检测器都无法胜任工作，Kalal 博士于 2009 年提出把跟踪器和检测器结合使用，即 TLD 算法。

TLD 算法作为一种长时目标跟踪算法，一经问世，便以其鲁棒的长时跟踪性能获得了人们的关注。2011 年，Kalal 创立了 TLD Vision 公司，以便将这项研究推向实际应用。目前已经从 TLD1.0、TLD2.0 升级到了 TLD3.0，如图 6 所示，分别给出了三个版本跟踪结果的示意图。在 TLD1.0 中，Kalal 给出了 TLD 算法的基本结构(Kalal 等，2012)。TLD 算法主要由 3 个模块构成：跟踪(tracking)模块、检测(detection)模块和学习(learning)模块。TLD 算法的工作流程大致为：首先，检测器通过一系列边界框产生样本，经过级联分类器产生正样本，放入样本集中；然后，使用跟踪器估计出目标的新位置，P 专家根据此位置产生正样本，N 专家再从这些正样本里选出一个最可信的样本，同时将其他正样本标记为负；最后，用正样本更新检测器参数，从而确定下一帧目标边界框的位置。TLD2.0 同样是一个包括跟踪、在线学习和检测的长时目标跟踪系统，但与 TLD1.0 不同的是，其能够进行多目标跟踪。

TLD1.0 和 TLD2.0 版本中尚未使用深度学习技术。在最新提供的 TLD3.0 系统中，实现了对视频中刚性 3D 目标的跟踪，如图 6(c)所示。TLD3.0 系统包含 3 个阶段：第 1 阶段选取多个尺度检测目标位置；第 2 阶段在这些检测结果上定位属于目标的部分；第 3 阶段执行数据关联，并采用 L1 损失函数对齐 3D 目标模型。该系统是基于深度神经网络和 3D 模型的组合，实现了精确性和鲁棒性之间的折中，而所有这些结果的获得都没有进行相机的校准。

6.3　智能监控与安防系统中的应用

传统的监控系统需要依靠人对得到的监控视频进行分析，耗时耗力。智能监控系统可以通过目标跟踪、识别等技术自动实现对目标场景的分析和异常检测。随着深度学习在计算机视觉领域的快速发展，智能视频分析技术已经成为安防企业竞争的关键，相关技术已经达到非常高的精度。传统安防技术更多的是关注事后查证的有效性，但随着高清摄像机的普及，如何利用这些资源使设备“活”起来，已经成为越来越多安防企业发展的重点。有了视频分析，就可以及时发现视频中的异常情况，从而在第一时间做出反应，减少损失。其中，基于深度学习的目标跟踪是其中的热点研究内容。

一个典型的安防系统应具有这些功能：首先在地图图像中标示出粗略的告警点信息；然后根据告警点信息，利用交互式电子地图与定位系统，给出精确的告警点信息；此时，启动视觉跟踪系统对告警点区域进行检测跟踪及异常情况判定；同时，借助其他探测手段进一步与图像信息融合，以提高异常情况判定的准确性。在这一过程中，基于深度学习的目标跟踪技术已经逐渐成为核心技术之一。

6.4　其他方面的应用简介

无人驾驶汽车是计算机视觉技术应用的重要领域。在自动驾驶过程中，通过对车道线、前后方车辆和行人等目标的准确识别，为更高级的行为选择、障碍物规避以及路径规划功能提供了基础，这其中的一项关键技术就是目标跟踪。由于实际路况极为复杂，基于传统目标检测的辅助驾驶技术性能难以得

到大幅提升。随着技术的发展,采用深度学习可以直接学习和感知路面和道路上车辆的特征,经过一段时间的正确驾驶过程,便能学习和感知实际道路情况下的相关驾驶技能,无需再通过感知具体的路况和各种目标,大幅提升了辅助驾驶算法的性能。

基于深度学习的目标跟踪技术也应用于智能机器人中。Agravante 等人(2018)研究了基于深度学习的人体运动跟踪,并将研究成果应用于 UR-5 机器人系统中,实现了对机器手臂书写的实时鲁棒跟踪。该跟踪系统将深度学习技术与预测控制技术相结合,而预测控制技术采用的是动态玻尔兹曼机(DyBM)模型(Osogami and Otsuka,2015),相较于长短时记忆 LSTM 模型(Hochreiter and Schmidhuber,1997)具有更高的跟踪性能。

在无人机应用方面,目标跟踪技术可以作为无人机视觉处理模块,实现对需要拍摄的目标进行持续跟踪,使焦点始终保持在目标上,从而达到更好的拍摄效果。目前,基于深度学习的视觉跟踪技术已经成为无人机视觉跟踪中重要的技术组成部分。

智能交通控制是“智慧城市”的关键内容之一。在城市的主干道,尤其是十字路口,对车辆、行人等目标的自动检测与跟踪是智能交通系统的重要任务,而基于深度学习的目标跟踪技术在其中起着重要作用,借助于云平台,能够及时有效地实现对交通状态的感知,从而提高整个城市的交通效能。

7 问题及展望

7.1 深度学习目标跟踪存在的问题

由于深度学习为构建更加鲁棒的外观模型提供了可能,因此将其应用于目标跟踪任务已成为必然趋势。目前的跟踪算法虽然能很好地应对简单场景,但面对复杂环境,设计出高精度、高鲁棒性和实时性的跟踪算法仍然有很多困难。与检测、识别等其他视觉任务不同,目标跟踪任务因其特殊性,使得在应用深度学习的过程中存在一些问题,如离线训练数据不足、很难实时在线训练和目标遮挡等。

7.1.1 训练数据问题

离线训练网络是深度学习中的重要一环,通过给定训练数据,利用损失函数来实现对网络参数的训练。为了保证训练的网络能够达到目的,实现相应的功能,深度学习需要大量的训练数据。在未采用深度学习方法之前,目标跟踪并没有离线训练的过程,因此只有有限的几个跟踪数据库,如 OTB50,OTB100 和 VOT 等。这些跟踪数据库数据量较少,并且彼此之间存在部分类似的序列,因此并不特别适用于目标跟踪任务的离线训练。

Tao 等人(2016)在 2016 年的 ICCV 上提出 SINT 算法,利用具有 300 个序列的 ALOV300 数据库实现对网络参数的训练。虽然去掉了 ALOV300 中与 OTB 和 VOT 相同的序列,但有些视频序列的背景相似,因此仍存在过拟合问题。同年发表的 SiamFC 算法中,Bertinetto 等人(2016b)将 ILSVRC 数据库应用于目标跟踪任务的离线训练过程。ILSVRC 数据库是检测数据库,包含 4 417 个视频序列,但该数据库建立的目的是目标检测任务,因此在数据库中,目标不会一直存在,并且同一帧中有多个目标,与目标任务不符。除了 ILSVRC 数据库外,SiamRPN(Li 等,2018a)还利用大规模带稀疏标注的视频数据集 Youtube-BB 进行训练,该数据集能够提供 50 倍数量的视频,保证了深度神经网络能够被充分的训练。2018 年 Valmadre 等人建立了 TrackingNet(Müller 等,2018)数据库。TrackingNet 数据库是专为目标跟踪任务设计的数据库,与一般大数据隔几帧标注一个目标不同,其对数据集每一帧中的目标都进行了标注。TrackingNet 数据库包含 3 万多个视频序列和 1 420 万个标注框,数据量较大,在一定程度上可以满足离线训练过程的需求。

7.1.2 实时跟踪问题

深度学习的优势之处在于可以通过大量数据进行学习。但在目标实时跟踪过程中,只有首帧的标注数据是完全准确的,要提取足够的训练数据具有较大困难。深度学习的网络模型较为复杂,网络参数较多,通过大量数据在线训练网络参数来满足跟踪要求,会在很大程度上影响跟踪速度。SiamFC(Bertinetto 等,2016b)只采用了 ILSVRC 数据库实现对网络进行离线训练,而没有进行在线的网络参数训练。这种做法虽然可以使跟踪算法达到实时,但离线训练的网络对当前目标的表达能力有限,很难实现最准确的目标跟踪。在 2017 年 ICCV 上发表的 Dsiam(Guo 等,2017)算法中,Guo 等人在 SiamFC 上进行改进,提出动态孪生网络(dynamic Siamese network),通过在线训练变化矩阵达到离线训练的

目的。但是,该方法并没有从深度学习的角度来完成离线训练过程,即并没有在线训练网络参数,因此还存在较大的问题。

7.1.3 长程目标跟踪中目标严重遮挡和消失问题

目标遮挡是导致跟踪失败的一个重要原因,也是实现长程目标跟踪的关键问题。跟踪任务从始至终都只跟踪一个目标,一旦目标被遮挡,则会极大程度上影响跟踪准确度,甚至导致跟踪失败。因此,当面临遮挡问题时,目标跟踪任务的要求更加严格。目前,目标遮挡可以分为两种情况:部分遮挡和完全遮挡。部分遮挡意味着在图像中还存在部分目标,可以通过对这部分的目标进行判断进而确定目标的位置;完全遮挡则是在图像中找不到目标,可能发生在有大的物体完全遮住了跟踪目标。

随着深度学习在目标跟踪领域的推广,其可以通过训练网络来学习目标的一般变化的特性,为解决目标遮挡问题提供了新的思路。目前,在深度目标跟踪中,解决目标部分遮挡问题的思路主要有两种,一是在离线训练时,在训练样本中增加存在遮挡的目标,通过损失函数,学习当目标被部分遮挡后所产生的变化,接着在线跟踪过程中,利用离线训练好的网络来对测试样本进行判断,进而实现准确跟踪;二是将跟踪目标分为若干部分,分别提取目标的深度特征,当目标被遮挡时,通过匹配测试样本与目标模板的深度特征,若存在较多的部分相似则判断样本为目标。目前解决完全遮挡问题的方法主要是参考目标检测的全图搜索,通过将整张图像的深度特征与目标模板的特征匹配,找到可能是目标的测试样本。但由于整幅图像较大,匹配过程中数据量较大,很难实现对目标的准确判断。

7.2 展望

目前,基于深度学习的目标跟踪方法仍主要停留在基于 ImageNet 预训练的特征应用层面。近年来,ImageNet 的视频数据集也逐渐用来学习更适合目标跟踪的深度特征,并取得了一定进展。随着 TrackingNet 等大规模数据集的出现,使得基于海量跟踪视频端对端地学习深度特征成为可能,有望进一步推进深度学习在视觉目标跟踪中取得突破性进展。相对而言,标注长程跟踪视频和构建大规模数据集的难度更大,如何根据长程跟踪任务的特点及其与短期跟踪任务的联系,结合迁移学习和深度学习构建合适的长期目标跟踪模型,也是未来视觉目标跟踪研究值得关注的一个重要方向。

8 结 论

将深度学习技术引入目标跟踪领域能更好地处理视频目标跟踪问题中存在的各类挑战,从而能有效提升跟踪方法的鲁棒性。具体来说,本文首先简要介绍了适用于目标跟踪的多种深度学习模型。然后,从网络结构、功能划分和网络训练等视角对当前各类基于深度学习的目标跟踪算法进行了详细介绍。接着,总结了基于深度学习的目标跟踪方法的实际应用情况。最后,综合基于深度学习的目标跟踪方法的研究现状,本文分析总结了当前方法存在的 3 个主要问题:1)训练数据问题;2)实时跟踪问题;3)长程目标跟踪中目标严重遮挡和消失视野问题。针对上述三类问题,本文对基于深度学习的跟踪方法的未来发展进行了展望。随着更大规模的数据集的出现,深度学习在视觉跟踪中将有望取得进一步的突破性进展。如何根据任务特点及任务间的联系,结合迁移学习和深度学习构建合适的长期目标跟踪模型,也是未来目标跟踪研究的重要发展方向之一。

参考文献(References)

Agravante D J, De Magistris G, Munawar A, Vinayavekhin P and Tachibana R. 2018. Deep learning with predictive control for human motion tracking [EB/OL]. 2018-08-07 [2019-07-01]. https://arxiv.org/pdf/1808.02200.pdf

Al-Shedivat M, Bansal T, Burda Y, Sutskever I, Mordatch I and Abbeel P. 2018. Continuous adaptation via meta-learning in nonstationary and competitive environments [EB/OL]. 2018-02-23 [2019-07-01]. https://arxiv.org/pdf/1710.03641.pdf

Arjovsky M, Chintala S and Bottou L. 2017. Wasserstein GAN [DB/OL]. [2019-07-02]. https://arxiv.org/pdf/1701.07875.pdf

Avidan S. 2007. Ensemble tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(2): 261-271 [DOI: 10.1109/TPAMI.2007.35]

Bertinetto L, Henriques J F, Valmadre J, Torr P and Vedaldi A. 2016a. Learning feed-forward one-shot learners//Proceedings of International Conference on Neural Information Processing Systems. Barcelona, Spain: NIPS, 523-531

Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S.

2016b, Fully-convolutional siamese networks for object tracking//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 850-865 [DOI: 10.1007/978-3-319-48881-3_56]

Bhat G, Johnander J, Danelljan M, Khan F S and Felsberg M. 2018. Unveiling the power of deep tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 493-509 [DOI: 10.1007/978-3-030-01216-8_30]

Bolme D, Beveridge J R, Draper B A and Lui Y M. 2010. Visual object tracking using adaptive correlation filters//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2544-2550 [DOI: 10.1109/cvpr.2010.5539960]

Bonin-Font F, Ortiz A and Oliver G. 2008. Visual navigation for mobile robots: a survey. Journal of Intelligent and Robotic Systems, 53(3): 263-296 [DOI: 10.1007/s10846-008-9235-4]

Bosch A, Zisserman A and Munoz X. 2007. Image classification using random forests and ferns//Proceedings of 2007 IEEE International Conference on Computer Vision. Rio de Janeiro, Brazil: IEEE, 1-8 [DOI: 10.1109/ICCV.2007.4409066]

Chen B, Wang D, Li P X, Wang S and Lu H. 2018. Real-time 'actor-critic' tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 328-345 [DOI: 10.1007/978-3-030-01234-2_20]

Chi Z Z, Li H Y, Lu H C and Yang M-H. 2017. Dual deep network for visual tracking. IEEE Transactions on Image Processing, 26(4): 2005-2015 [DOI: 10.1109/TIP.2017.2669880]

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: EMNLP, 1724-1734

Choi J, Chang H J, Fischer T, Yun S, Lee K, Jeong J, Demiris Y and Choi J Y. 2018. Context-aware deep feature compression for high-speed visual tracking//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 479-488 [DOI: 10.1109/CVPR.2018.00057]

Choi J, Kwon J and Lee K M. 2017. Deep meta learning for real-time visual tracking based on target-specific feature space [DB/OL] [2019-07-02]. https://arxiv.org/pdf/1712.09153.pdf

Collins R T, Lipton A J and Kanade T. 2000. A system for video surveillance and monitoring [R]. VSAM Final Report, Pittsburgh: Carnegie Mellon University, 329-337

Collins R T and Liu Y X. 2003. On-line selection of discriminative tracking features//Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France. IEEE, 346-352 [DOI: 10.1109/iccv.2003.1238365]

Collins R, Zhou X H and Teh S K. 2005. An open source tracking testbed and evaluation website//Proceedings of 2005 IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. Breckenridge, Colorado: IEEE, #35

Comaniciu D and Meer P. 2002. Mean shift: a robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5): 603-619 [DOI: 10.1109/34.1000236]

Cruz-Mota J, Bogdanova I, Paquier B, Bierlaire M and Thiran J P. 2012. Scale invariant feature transform on the sphere: theory and applications. International Journal of Computer Vision, 98(2): 217-241 [DOI: 10.1007/s11263-011-0505-4]

Danelljan M, Bhat G, Khan F S and Felsberg M. 2017. Eco: efficient convolution operators for tracking//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 6931-6939 [DOI: 10.1109/CVPR.2017.733]

Danelljan M, Häger G, Khan F S and Felsberg M. 2014. Accurate scale estimation for robust visual tracking//Proceedings of the British Machine Vision Conference. Nottingham, UK: BMVA Press [DOI: 10.5244/C.28.65]

Danelljan M, Häger G, Khan F S and Felsberg M. 2015a. Learning spatially regularized correlation filters for visual tracking//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 4310-4318 [DOI: 10.1109/ICCV.2015.490]

Danelljan M, Häger G, Shahbaz Khan F and Felsberg M. 2015b. Convolutional features for correlation filter based visual tracking//Proceedings of 2015 IEEE International Conference on Computer Vision Workshop. Santiago, Chile: IEEE, 621-629 [DOI: 10.1109/ICCVW.2015.84]

Danelljan M, Robinson A, Khan F S and Felsberg M. 2016. Beyond correlation filters: learning continuous convolution operators for visual tracking//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 472-488 [DOI: 10.1007/978-3-319-46454-1_29]

Dong X P, Shen J B, Wang W G, Liu Y, Shao L and Porikli F. 2018. Hyperparameter optimization for tracking with continuous deep Q-learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 518-527 [DOI: 10.1109/CVPR.2018.00061]

Fan H and Ling H B. 2017. SANet: structure-aware network for visual tracking//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, HI, USA: IEEE, 2217-2224 [DOI: 10.1109/CVPRW.2017.275]

Fisher R B. 2004. The PETS04 surveillance ground-truth data sets//Proceedings of 2004 IEEE International Workshop on Performance Evaluation of Tracking and Surveillance. Prague, Czech Republic: IEEE, 1-5

Galoogahi H K, Fagg A and Lucey S. 2017. Learning background-aware correlation filters for visual tracking//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 1144-1152 [DOI: 10.1109/ICCV.2017.129]

Grabner H, Leistner C and Bischof H. 2008. Semi-supervised on-line boosting for robust tracking//Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 234-247 [DOI: 10.1007/978-3-540-88682-2_19]

Guan H, Xue X Y and An Z Y. 2016. Advances on application of deep learning for video object tracking. Acta Automatica Sinica, 42(6): 834-847 (管皓, 薛向阳, 安志勇. 2016. 深度学习在视频目标跟踪中的应用进展与展望. 自动化学报, 42(6): 834-847) [DOI: 10.16383/j.aas.2016.c150705]

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V and Courville A C. 2017. Improved training of wasserstein GANs//Proceedings of Advances in Neural Information Processing Systems. Long Beach, CA, USA: NIPS, 5767-5777

Gundogdu E and Alatan A A. 2018. Good features to correlate for visual tracking. IEEE Transactions on Image Processing, 27(5): 2526-2540 [DOI: 10.1109/TIP.2018.2806280]

Guo Q, Feng W, Zhou C, Huang R, Wan L and Wang S. 2017. Learning dynamic Siamese network for visual object tracking//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 1781-1789 [DOI: 10.1109/ICCV.2017.196]

Han B, Sim J and Adam H. 2017. BranchOut: regularization for online ensemble tracking with convolutional neural networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 521-530 [DOI: 10.1109/CVPR.2017.63]

Hare S, Golodetz S, Saffari A, Vineet V, Cheng M M, Hicks S L and Torr P H S. 2016. Struck: structured output tracking with kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10): 2096-2109 [DOI: 10.1109/TPAMI.2015.2509974]

Haritaoglu I, Harwood D and Davis L S. 2000. W^4: real-time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8): 809-830 [DOI: 10.1109/34.868683]

He A F, Luo C, Tian X M and Zeng W. 2018. A twofold siamese network for real-time object tracking//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 4834-4843 [DOI: 10.1109/CVPR.2018.00508]

He K M, Zhang X Y, Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 770-778 [DOI: 10.1109/CVPR.2016.90]

Held D, Thrun S and Savarese S. 2016. Learning to track at 100 FPS with deep regression networks//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 749-765 [DOI: 10.1007/978-3-319-46448-0_45]

Henriques J F, Rui C, Martins P Vineet V, Cheng M, Hicks S L and Torr P H S. 2012. Exploiting the circulant structure of tracking-by-detection with kernels//Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 702-715 [DOI: 10.1007/978-3-642-33765-9_50]

Henriques J F, Caseiro R, Martins P and Batista J. 2015. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3): 583-596 [DOI: 10.1109/tpami.2014.2345390]

Hester T, Vecerik M, Pietquin O, Lanctot M, Schaul T, Piot B, Dan H, Quan J, Sendonaris A and Dulacarnold G. 2018. Deep Qlearning from demonstrations//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, LA, USA: AAAI

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

Horn B K P and Schunck B G. 1981. Determining optical flow. Artificial Intelligence, 17(1-3): 185-203 [DOI: 10.1016/0004-3702(81)90024-2]

Hu W M, Xie D, Fu Z Y, Zeng W and Maybank S. 2007. Semantic-based surveillance video retrieval. IEEE Transactions on Image Processing, 16(4): 1168-1181 [DOI: 10.1109/TIP.2006.891352]

Huang C, Lucey S and Ramanan D. 2017a. Learning policies for adaptive tracking with deep feature cascades//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 105-114 [DOI: 10.1109/ICCV.2017.21]

Huang G, Liu Z, Van Der Maaten L and Weinberger K Q. 2017b. Densely connected convolutional networks//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2261-2269 [DOI: 10.1109/CVPR.2017.243]

Huang K Q, Chen X T, Kang Y F and Tan T N. 2015. Intelligent visual surveillance: a review. Chinese Journal of Computers, 38(6): 1093-1118 (黄凯奇, 陈晓棠, 康运锋, 谭铁军. 2015. 智能视频监控技术综述. 计算机学报, 38(6): 1093-1118) [DOI: 10.11897/SP.J.1016.2015.01093]

Isard M and Blake A. 1998. CONDENSATION—conditional density propagation for visual tracking. International Journal of Computer Vision, 29(1): 5-28 [DOI: 10.1023/A:1008078328650]

Jepson A D, Fleet D J and El-Maraghi T F. 2003. Robust online appearance models for visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(10): 1296-1311 [DOI: 10.1109/TPAMI.2003.1233903]

Kalal Z, Mikolajczyk K and Matas J. 2012. Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(7): 1409-1422 [DOI: 10.1109/TPAMI.2011.239]

Kingma D P and Welling M. 2013. Auto-encoding variational bayes [DB/OL] [2019-07-02]. https://arxiv.org/pdf/1312.6114.pdf

Kristan M, Matas J, Leonardis A, Felsberg M, Cehovin L, Fernandez G, Vojir T, Hager G, Nebehay G and Pflugfelder R. 2015a. The visual object tracking VOT2015 challenge results//Proceedings of 2015 IEEE International Conference on Computer vision Workshop.

Santiago, Chile: IEEE, 564-586 [DOI: 10.1109/ICCVW.2015.79]

Kristan M, Matas J, Leonardis A, Vojíř T, Pflugfelder R, Fernández G, Nebehay G, Porikli F and Čehovin L. 2016. A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(11): 2137-2155 [DOI: 10.1109/TPAMI.2016.2516982]

Kristan M, Pflugfelder R, Leonardis A, Matas J, Čehovin L, Nebehay G, Vojíř T, Fernández G, Lukežič A, Dimitriev A, Petrosino A, Saffari A, Li B, Han B, Heng C, Garcia C, Pangeršič D, Häger G, Khan F S, Oven F, Possegger H, Bischof H, Nam H, Zhu J, Li J, Choi J Y, Choi J-W, Henriques J F, van de Weijer J, Batista J, Lebeda K, Öfjäll K, Yi K M, Qin L, Wen L, Maresca M E, Danelljan M, Felsberg M, Cheng M-M, Torr P, Huang Q, Bowden R, Hare S, Lim S Y, Hong S, Liao S, Hadfield S, Li S Z, Duffner S, Golodetz S, Mauthner T, Vineet V, Lin W, Li Y, Qi Y, Lei Z and Niu Z H. 2015b. The visual object tracking VOT2014 challenge results//Proceedings of 2014 European Conference on Computer vision. Zurich, Switzerland: Springer, 191-217 [DOI: 10.1007/978-3-319-16181-5_14]

Kristan M, Pflugfelder R, Leonardis A, Matas J, Porikli F, Cehovin L, Nebehay G, Fernandez G, Vojir T, Gatt A, Khajenezhad A, Salahledin A, Soltani-Farani A, Zarezade A, Petrosino A, Milton A, Bozorgtabar B, Li B, Chan C S, Heng C, Ward D, Kearney D, Monekosso D, Karaimer H C, Rabiee H R, Zhu J, Gao J, Xiao J, Zhang J, Xing J, Huang K, Lebeda K, Cao L, Maresca M E, Lim M K, Helw M E, Felsberg M, Remagnino P, Bowden R, Goecke R, Stolkin R, Lim S Y, Maher S, Poullot S, Wong S, Satoh S, Chen W, Hu W, Zhang X, Li Y and Niu Z. 2013. The visual object tracking vot2013 challenge results//Proceedings of 2013 IEEE International Conference on Computer Vision Workshops. Sydney, NSW, Australia: IEEE, 98-111 [DOI: 10.1109/ICCVW.2013.20]

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: Curran Associates Inc, 1097-1105

Kumarawadu S, Watanabe K, Kiguchi K and Izumi K. 2002. Adaptive output tracking of partly known robotic systems using softmax function networks//Proceedings of 2002 International Joint Conference on Neural Networks. Honolulu, HI, USA, USA: IEEE, 483-488 [DOI: 10.1109/IJCNN.2002.1005520]

Li A N, Lin M, Wu Y, Yang M and Yan S. 2016. NUS-PRO: a new visual tracking challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2): 335-349 [DOI: 10.1109/TPAMI.2015.2417577]

Li B, Yan J J, Wu W, Zhu Z and Hu X. 2018a. High performance visual tracking with siamese region proposal network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 8971-8980 [DOI: 10.1109/CVPR.2018.00935]

Li F, Tian C, Zuo W M, Zhang L and Yang M H. 2018b. Learning spatial-temporal regularized correlation filters for visual tracking//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 4904-4913 [DOI: 10.1109/CVPR.2018.00515]

Li P X, Wang D, Wang L J and Lu H. 2018c. Deep visual tracking: Review and experimental comparison. Pattern Recognition, 76: 323-338 [DOI: 10.1016/j.patcog.2017.11.007]

Li B, Yan J J, Wu W, Zhu Z and Hu X L. 2018d. High performance visual tracking with Siamese region proposal network//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 8971-8980 [DOI: 10.1109/cvpr.2018.00935]

Li T H S and Chang S J. 2005. Autonomous fuzzy parking control of a carlike mobile robot. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 33(4): 451-465 [DOI: 10.1109/TSMCA.2003.811766]

Li Y and Zhu J K. 2015. A scale adaptive kernel correlation filter tracker with feature integration//Proceedings of 2014 European Conference on Computer Vision. Zurich, Switzerland: Springer, 254-265 [DOI: 10.1007/978-3-319-16181-5_18]

Liang P P, Blasch E and Ling H B. 2015. Encoding color information for visual tracking: algorithms and benchmark. IEEE Transactions on Image Processing, 24(12): 5630-5644 [DOI: 10.1109/TIP.2015.2482905]

Lin Y M, Shen J, Cheng S Y and Pantic M. Mobile face tracking: a survey and benchmar [DB/OL]. [2019-07-02]. https://arxiv.org/pdf/1805.09749v1.pdf

Lowe D G. 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2): 91-110 [DOI: 10.1023/B:VISI.0000029664.99615.94]

Lu H C, Fang G L, Wang C and Chen Y W. 2010. A novel method for gaze tracking by local pattern model and support vector regressor. Signal Processing, 90(4): 1290-1299 [DOI: 10.1016/j.sigpro.2009.10.014]

Lu H C, Li P X and Wang D. 2018. Visual object tracking: a survey. Pattern Recognition and Artificial Intelligence, 31(1): 61-76 (**卢湖川, 李佩霞, 王栋**. 2018. **目标跟踪算法综述. 模式识别与人工智能**, 31(1): 61-76) [DOI: 10.16451/j.cnki.issn1003-6059.201801006]

Lu X K, Ma C, Ni B B, Yang X, Reid I and Yang M-H. 2018. Deep regression tracking with shrinkage loss//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 369-386 [DOI: 10.1007/978-3-030-01264-9_22]

Lukežic A, Vojír T, Zajc L C, Matas J and Kristan M. 2017. Discriminative correlation filter with channel and spatial reliability//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Rec-

ognition. Honolulu, HI, USA: IEEE, 4847-4856 [DOI: 10.1109/CVPR.2017.515]

Ma C, Huang J B, Yang X K and Yang M H. 2015. Hierarchical convolutional features for visual tracking//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 3074-3082 [DOI: 10.1109/ICCV.2015.352]

Matas J, Chum O, Urban M and Pajdla T. 2004. Robust wide-baseline stereo from maximally stable extremal regions. Image and Vision Computing, 22(10): 761-767 [DOI: 10.1016/j.imavis.2004.02.006]

Meshgi K, Oba S and Ishii S. 2017. Efficient diverse ensemble for discriminative co-tracking [DB/OL]. [2019-07-02]. https://arxiv.org/pdf/1711.06564.pdf

Mita T, Kaneko T and Hori O. 2005. Joint haar-like features for face detection//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing: IEEE, 1619-1626 [DOI: 10.1109/ICCV.2005.129]

Mueller M, Smith N and Ghanem B. 2016. A benchmark and simulator for UAV tracking//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 445-461 [DOI: 10.1007/978-3-319-46448-0_27]

Müller M, Bibi A, Giancola S, Al-Subaihi S and Ghanem B. 2018. TrackingNet: a large-scale dataset and benchmark for object tracking in the wild//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 310-327 [DOI: 10.1007/978-3-030-01246-5_19]

Mei X and Ling H. 2011. Robust visual tracking and vehicle classification via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 33(11): 2259-2272

Nam H, Baek M and Han B. 2016. Modeling and propagating CNNs in a tree structure for visual tracking [DB/OL]. [2019-07-02]. https://arxiv.org/pdf/1608.07242.pdf

Nam H and Han B. 2016. Learning multi-domain convolutional neural networks for visual tracking//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 4293-4302 [DOI: 10.1109/CVPR.2016.465]

Osogami T and Otsuka M. 2015. Seven neurons memorizing sequences of alphabetical images via spike-timing dependent plasticity. Scientific Reports, 5: #14149 [DOI: 10.1038/srep14149]

Park E and Berg A C. 2018. Meta-tracker: fast and robust online adaptation for visual object trackers//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 587-604 [DOI: 10.1007/978-3-030-01219-9_35]

Qi Y K, Zhang S P, Qin L, Yao H, Huang Q, Lim J and Yang M H. 2016. Hedged deep tracking//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 4303-4311 [DOI: 10.1109/CVPR.2016.466]

Radford A, Metz L and Chintala S. 2016. Unsupervised representation learning with deep convolutional generative adversarial network [DB/OL]. [2019-07-02]. https://arxiv.org/pdf/1511.06434.pdf

Ren L L, Yuan X, Lu J W, Yang M and Zhou J. 2018. Deep reinforcement learning with iterative shift for visual tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 697-713 [DOI: 10.1007/978-3-030-01240-3_42]

Ross D A, Lim J, Lin R S and Yang M H. 2008. Incremental learning for robust visual tracking. International Journal of Computer Vision, 77(1-3): 125-141 [DOI: 10.1007/s11263-007-0075-7]

Schroff F, Kalenichenko D and Philbin J. 2015. FaceNet: a unified embedding for face recognition and clustering//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 815-823 [DOI: 10.1109/CVPR.2015.7298682]

Shi J B and Tomasi C. 1994. Good features to track//Proceedings of 1994 IEEE Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 593-600 [DOI: 10.1109/CVPR.1994.323794]

Shi X J, Chen Z R, Wang H, Yeung D Y, Wong W K and Woo W-C. 2015. Convolutional LSTM network: a machine learning approach for precipitation nowcasting//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press, 802-810

Shu C F, Hampapur A, Lu M, Brown L, Connell J, Senior A and Tian Y. 2005. IBM smart surveillance system (S3): a open and extensible framework for event based surveillance//Proceedings of 2005 IEEE Conference on Advanced Video and Signal Based Surveillance. Como, Italy: IEEE, 318-323 [DOI: 10.1109/AVSS.2005.1577288]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition//Proceedings of International Conference on Learning Representations. San Diego, CA: ICLR

Smeulders A W M, Chu D M, Cucchiara R, Calderara S, Dehghan A and Shah M. 2014. Visual tracking: an experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7): 1442-1468 [DOI: 10.1109/TPAMI.2013.230]

Song Y B, Ma C, Gong L J, Zhang J, Lau R W and Yang M H. 2017. CREST: convolutional residual learning for visual tracking//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2574-2583 [DOI: 10.1109/ICCV.2017.279]

Song Y B, Ma C, Wu X H, Gong L, Bao L, Zuo W, Shen C, Lau R and Yang M H. 2018. Vital: visual tracking via adversarial learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 8990-8999 [DOI: 10.1109/CVPR.2018.00937]

Sun C, Wang D, Lu H C and Yang M-H. 2018. Correlation tracking via joint discrimination and reliability learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recogni-

tion. Salt Lake City, UT, USA: IEEE, 489-497 [DOI: 10.1109/CVPR.2018.00058]

Sun Y, Wang X G and Tang X O. 2015. Deeply learned face representations are sparse, selective, and robust//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2892-2900 [DOI: 10.1109/CVPR.2015.7298907]

Supančič III J and Ramanan D. 2017. Tracking as online decision-making: Learning a policy from streaming videos with reinforcement learning//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 322-331 [DOI: 10.1109/ICCV.2017.43]

Suykens J A K and Vandewalle J. 1999. Least squares support vector machine classifiers. Neural Processing Letters, 9(3): 293-300 [DOI: 10.1023/A:1018628609742]

Svetnik V, Liaw A, Tong C, Culberson J C, Sheridan R P and Feuston B P. 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of Chemical Information and Computer Sciences, 43(6): 1947-1958 [DOI: 10.1021/ci034160g]

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Tao R, Gavves E and Smeulders A W M. 2016. Siamese instance search for tracking//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 1420-1429 [DOI: 10.1109/CVPR.2016.158]

Valmadre J, Bertinetto L, Henriques J F, Tao R, Vedaldi A, Smeulders A, Torr P and Gavves E. 2018. Long-term tracking in the wild: a benchmark//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 692-707 [DOI: 10.1007/978-3-030-01219-9_41]

Valmadre J, Bertinetto L, Henriques J, Vedaldi A and Torr P H. 2017. End-to-end representation learning for correlation filter based tracking//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 5000-5008 [DOI: 10.1109/CVPR.2017.531]

Van De Weijer J, Schmid C, Verbeek J and Larlus D. 2009. Learning color names for real-world applications. IEEE Transactions on Image Processing, 18(7): 1512-1523 [DOI: 10.1109/TIP.2009.2019809]

Vincent P, Larochelle H, Lajoie I, Bengio Y and Manzagol P A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11: 3371-3408

Viola P and Jones M. 2001. Fast and robust classification using asymmetric adaboost and a detector cascade//Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver, British Columbia, Canada: MIT Press, 1311-1318

Wang L J, Ouyang W L, Wang X G and Lu H. 2016. STCT: sequentially training convolutional networks for visual tracking//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 1373-1381 [DOI: 10.1109/CVPR.2016.153]

Wang N Y and Yeung D Y. 2013. Learning a deep compact image representation for visual tracking//Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: ACM, 809-817

Wang N Y, Shi J P, Yeung D Y and Jia J. 2015. Understanding and diagnosing visual tracking systems//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 3101-3109 [DOI: 10.1109/ICCV.2015.355]

Wang N, Zhou W G, Tian Q, Hong R, Wang M and Li H. 2018a. Multi-cue correlation filters for robust visual tracking//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 4844-4853 [DOI: 10.1109/CVPR.2018.00509]

Wang Q, Zhang M D, Xing J L, Gao J, Hu W and Maybank S. 2018b. Do not lose the details: reinforced representation learning for high performance visual tracking//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: IJCAI, 985-997 [DOI: 10.24963/ijcai.2018/137]

Wang X, Li C L, Luo B and Tang J. 2018c. SINT++: robust visual tracking via adversarial positive instance generation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 4864-4873 [DOI: 10.1109/CVPR.2018.00511]

Wu Y, Lim J and Yang M H. 2015. Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1834-1848 [DOI: 10.1109/TPAMI.2014.2388226]

Wu Y, Lim J and Yang M H. 2013. Online object tracking: a benchmark//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2411-2418 [DOI: 10.1109/CVPR.2013.312]

Yang T Y and Chan A B. 2018. Learning dynamic memory networks for object tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 153-169 [DOI: 10.1007/978-3-030-01240-3_10]

Yun S, Choi J and Yun Y. 2017. Action-decision networks for visual tracking with deep reinforcement learning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 1349-1358 [DOI: 10.1109/CVPR.2017.148]

Zeiler M D and Fergus R. 2014. Visualizing and understanding convolutional networks//Proceedings of the 13th European Conference on

Computer Vision. Zurich, Switzerland: Springer, 818-833 [DOI: 10.1007/978-3-319-10590-1_53]

Zhang K H, Liu Q S, Wu Y and Yang M H. 2016. Robust visual tracking via convolutional networks without training. IEEE Transactions on Image Processing, 25(4): 1779-1792 [DOI: 10.1109/TIP.2016.2531283]

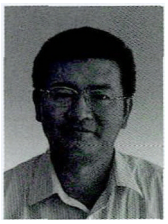
Zhao F, Wang J Q, Wu Y and Tang M J I T. 2019. Adversarial deep tracking. IEEE Transactions on Circuits and Systems for Video Technology, 29(7): 1998-2011 [DOI: 10.1109/TCSVT.2018.2856540]

Zhu Z, Wang Q, Li B, Wu W, Yan J and Hu W. 2018a. Distractor-aware Siamese networks for visual object tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer, 103-119 [DOI: 10.1007/978-3-030-01240-3_7]

Zhu Z, Wu W, Zou W and Yan J. 2018b. End-to-end flow correlation tracking with spatial-temporal attention//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 548-557 [DOI: 10.1109/CVPR.2018.00064]

作者简介

李玺，1981 年生，男，教授，主要研究方向为计算机视觉、模式识别以及深度学习。
E-mail: xilizju@zju.edu.cn



王菡子，通信作者，男，教授，主要研究方向为计算机视觉和模式识别。
E-mail: Hanzi.Wang@xmu.edu.cn

查宇飞，男，副教授，主要研究方向为视频目标跟踪。
E-mail: zhayufei@126.com

张天柱，男，教授，主要研究方向为模式识别与智能系统。
E-mail: tzzhang@ustc.edu.cn

崔振，男，教授，主要研究方向为计算机视觉与模式识别。
E-mail: zhen.cui@njust.edu.cn

左旺孟，男，教授，主要研究方向为计算机视觉和深度学习。
E-mail: cswmzuo@gmail.com

侯志强，男，教授，主要研究方向为图像处理、计算机视觉和信息融合。E-mail: hzq@xupt.edu.cn

卢湖川，男，教授，主要研究方向为计算机视觉、模式识别、图像处理。E-mail: lhchuan@dlut.edu.cn