基于深度学习的目标跟踪方法研究现状与展望

罗海波[1,2,3,4],许凌云[1,2,3,4],惠 斌[1,3,4],常 铮[1,3,4]

(1. 中国科学院沈阳自动化研究所,辽宁 沈阳 110016; 2.中国科学院大学,北京 100049;
3. 中国科学院光电信息处理重点实验室,辽宁 沈阳 110016;
4. 辽宁省图像理解与视觉计算重点实验室,辽宁 沈阳 110016)

摘 要: 目标跟踪是计算机视觉领域的重要研究方向之一,在精确制导、智能视频监控、人机交互、机器人导航、公共安全等领域有着重要的作用。目标跟踪的基本问题是在一个视频或图像序列中选择感兴趣的目标,在接下来的连续帧中,找到该目标的准确位置并形成其运动轨迹。目标跟踪是一个颇具挑战性的问题,目标的非刚性变化往往改变了目标的表观模型,同时复杂的光照变化、目标与场景间的遮挡、背景中相似物体的干扰和摄像机的抖动等使目标跟踪任务变得更加困难。近年来,随着深度学习在目标检测和识别等领域中取得巨大的突破,许多学者开始将深度学习模型引入到目标跟踪中,并在一系列数据评测集上取得了优于传统方法的性能,逐渐开启了目标跟踪领域的新篇章。文中将首先阐述目标跟踪问题的难点和基本解决思路;然后根据利用深度学习算法解决目标跟踪问题的不同思路,对当前出现的此类主流算法进行分析,介绍这些算法各自的优缺点及未来的工作方向。

关键词: 目标跟踪; 深度学习; 计算机视觉; 精确制导

中图分类号: TP391 **文献标志码:** A **DOI:** 10.3788/IRLA201746.0502002

Status and prospect of target tracking based on deep learning

Luo Haibo[1,2,3,4],Xu Lingyun[1,2,3,4], Hui Bin[1,3,4],Chang Zheng[1,3,4]

1. Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China;
3. Key Laboratory of Opto-Electronic Information Processing, Chinese Academy of Sciences, Shenyang 110016, China;
4. The Key Lab of Image Understanding and Computer Vision, Liaoning Province, Shenyang 110016, China)

Abstract: The inverse synthetic aperture lidar (ISAL) have attracted increasing attention for its merits including small visual tracking which is considered as one of the important research topics in the field of computer vision due to its key role in versatile applications, such as precision guidance, intelligent video surveillance, human-computer interaction, robot navigation and public safety. The basic idea for implementing visual tracking is composed of finding the target object in a video or sequence of images, then determining its exact position in the next successive frames and finally generating the corresponding trajectory of this object. Visual tracking, however, is still a challenging problem in practice while taking into account the abrupt appearance changes of the target objects induced by their non-rigid

收稿日期:2016-09-10; **修订日期:**2016-10-20

基金项目:总装预研项目(51301030108)

作者简介:罗海波(1967-),男,研究员,博士生导师,博士,主要从事图像处理、目标跟踪、目标识别等方面的研究。Email: luohb@sia.cn

transformation, the sophisticated lighting variation, the obstruction by the block or similar objects in the background and the camera jitter. Motivated by the successful applications in target detection and recognition in recent years, plenty of deep learning models have been integrated in the visual tracking and better performance over traditional methods was achieved in a series of data evaluations, which opens a new door in the field of visual tracking. In this paper, the overview and progress on visual tracking were summarized. The current challenges and corresponding solving approaches in this field are introduced firstly and in particular, several novel and mainstream visual tracking algorithms based on the deep learning are specially described and analyzed in details, including their basic ideas, advantages and disadvantages and future prospect.

Key words: target tracking; deep learning; computer vision; precision guidance

0 引 言

目标跟踪是计算机视觉领域的重要研究方向之一,也是该领域的一个研究热点[1]。它的基本问题形式是在一个图像序列或视频流中选择一个感兴趣的区域或者物体作为目标,在接下来的连续若干帧中自动找到该目标的位置,得到目标的运动轨迹、具体形态和位置。目标跟踪在军事侦察、精确制导、智能视频监控、人机交互、机器人导航等领域具有广泛的应用,有着重要的实用价值。在这些不同的应用中,人们往往需要通过对图像序列或者视频流中感兴趣的区域或物体进行分析,得到目标的位置、运动轨迹和表观变化,以达到对其进行跟踪、机器人导航避障、进一步进行行为分析和视频高层语意解析等目的。

然而,在实际应用中,目标跟踪面临着诸多挑战。比如,对目标先验知识的不足、复杂的场景变化、摄像机和目标之间的相对不规则运动等,使得设计能适用于所有场景的目标跟踪算法变得非常困难。

一般来说,设计一种目标跟踪算法或者算法框架需重点考虑以下 4 个方面的问题:

(1) 目标或场景的外观模型

外观模型指的是目标在二维图像平面上的外观信息,包括物体及其各部件的几何位置、颜色和上下文信息。不同的跟踪算法往往采用不同的统计模型对目标进行外观模型建模,或者通过提取诸如颜色、纹理、梯度、空间几何等特征以及上述多种不同特征的组合来对目标的外观模型进行表征。例如,L1 算法 [2] 应用稀疏表达来表示目标的外观模型;staple 算法[3]应用了多种特征的组合模型来计算目标在外观模型上的得分,然后根据得分确定目标。

(2) 目标的运动模型

在跟踪问题中,目标的运动往往是不规则的,人们能获知的只是物体运动的连续性和在某一帧中目标出现在搜索区内各个位置的概率。研究目标运动模型的好处是可以提高对模板的采样效率。获知目标的具体运动模型后可以大大减少无效采样空间的相关计算量,从而大幅提升算法的效率并减少漏检率。传统的 NCC 算法[4]将目标模板与全图像进行匹配,增加了许多不必要的计算。粒子滤波算法[5]和光流法[6]可以用于估算目标出现在新一帧中的概率分布,得到目标的运动模型从而避免低概率位置的无效计算。通过计算颜色直方图的偏移概率,Meanshift 算法[7]也同样提升了目标位置搜索的效率。

(3) 目标的内部运动模型

对于目标跟踪的基本问题形式来说,内部运动模型是指物体各个部分的相对运动,即非刚体变换。对于非刚体变换,目前通常的解决方法是建立特殊的模型,例如用于检测算法的可变形模板模型[8]。它通过将一个目标分解为多个相关的子块,通过对各子块的跟踪实现对该目标的跟踪和形态估计。该模型在一定程度上可以解决目标发生形变时的跟踪丢失问题。

(4) 模板的更新策略

不同于一般的目标识别问题,在目标跟踪,尤其是长时间的目标跟踪问题中,恰当的模板更新策

略是必要的。这是因为随着目标的运动和场景的变化，受到场景中干扰物体的遮挡和光照的影响，同时还有快速运动带来的外观模糊等其他变化，目标往往会发生形变。模板更新策略一般需要考虑到计算效率和模板更新速度的平衡、模板可靠性和模板是否新近的平衡。不同的跟踪算法往往根据需要采用给予新近模板和置信度高于一定阈值的模板较大的权重，随机化其他模板权重等方法来提高其跟踪性能。

1 目标跟踪的难点及基本解决思路

1.1 目标跟踪的难点

在对图像序列和视频流的分析中，由于 3D 真实世界投影到 2D 图像平面的信息缺失、图像中噪音的污染、场景的复杂多样性和目标自身的复杂变化，目标跟踪问题面临着诸多挑战。如表 1 所示，根据不同的场景变换和目标变化，一般可以将目标跟踪面临的挑战分为10 类。

表 1 目标跟踪面临的主要挑战

Tab.1 Main challenge in object tracking

No.	Challenge
1	Illumination change
2	Scene complexity
3	Clutter
4	Size change
5	Color change
6	Deformation
7	Aspect-ratio change
8	Blur
9	Scene motion
10	Camera motion

1.2 解决目标跟踪问题的基本解决思路

从数学模型的角度出发，目标跟踪模型一般可以分为判别式模型和生成式模型。生成式跟踪模型通过学习代表目标的外观模型，然后以此为模板在搜索区内进行最小化重构误差的模式匹配，从而完成对目标的跟踪。L1 APG[2]是生成式跟踪模型的代表算法，它假设每一个候选目标均可由字典(即目标模板和碎片模板的组合)稀疏表示，在所有的候选目标中，选择系数最稀疏同时重构误差最小的目标作为跟踪结果。Zhang[9]等人利用循环稀疏结构的目标模板和傅里叶变换增大了样本空间，提高了 L1 跟踪器的计算速度。Tao[10]等人通过引入孪生网络，直接计算目标外观模型的表示误差，以选择出与目标模型最接近的区域作为跟踪结果。

判别式跟踪模型则把跟踪问题看作二元分类问题，着重通过图像序列训练得到区分目标和背景的决策边界。自 TLD 算法[11]提出以来，掀起了一股将判别式模型用于目标跟踪的热潮。其中经典的优秀代表有 Struck[12]和 KCF[13]跟踪算法。Struck 跟踪算法通过为目标区域构建结构化的 SVM 分类器将目标与背景区分开来，KCF 跟踪算法则选用了岭回归模型，并通过巧妙地引入具有循环结构的模板对其进行傅里叶变换，避免了岭回归中的矩阵求逆问题，大幅提升了跟踪的速度和效率。此外，近年来出现了融合多种特征的跟踪算法，例如 staple 跟踪算法高效地将颜色直方图和梯度直方图特征的跟踪结果进行融合，在评测集上取得了较好的结果。此外，目前大部分基于深度学习的目标跟踪方法也可归属于判别式方法。

2 深度学习在目标跟踪中的应用

自 2006 年 Hinton 等[14-15]在深度置信网络方面的重大研究工作发表以来，深度学习作为机器学习的新方向，在人工智能领域的许多重要问题上大显身手。至今已有数种深度学习框架，如深度神经网络、卷积神经网络、深度置信网络和递归神经网络已被成功应用于计算机视觉、语音识别、自然语言处理、音频识别以及生物信息学等领域并获得了极好的效果。

在目标跟踪领域中，自 2013 年以来，基于深度学习的一系列跟踪算法逐渐取得了跟踪精度方面的绝对优势；2016 年，David[16]提出了一种新的深度学习网络框架，第一次将基于深度学习的目标跟踪算法做到了 100 fps 以上，满足了目标跟踪的实时性要求。基于深度学习的目标跟踪算法已经成为目标跟踪领域中不可或缺的重要组成部分。文中将从 3 个方面对当前基于深度学习的目标跟踪算法进行介绍：

(1) 在传统目标跟踪算法框架下，利用深度学习网络提取目标区域的相关特征来提升跟踪算法的准确性；

(2)在传统目标跟踪算法框架下,通过设计针对目标跟踪问题的深度学习网络,实现目标跟踪；

(3) 将其他一些新型的深度学习网络结构应用于目标跟踪中。

2.1 基于深度学习特征的目标跟踪算法

2.1.1 栈式自编码器特征

栈式自编码网络是非监督的深度学习网络之一,由多层稀疏自编码网络组成,分为编码器和解码器两部分。假设观测样本为$\{y_1, y_2, \cdots, y_k\}$稀疏自编码网络致力于优化如下模型：

$$\min_{W,W',b,b'} \sum_{i=1}^{k} \left|\left| y_i - \hat{y}_i \right|\right|^2 + \lambda(||W||_F((||W'||_F) \quad (1)$$

其中,自编码网络通过对神经元的激活或抑制来获取稀疏性结构。让

$$\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^{k} \left[a_j(y^{(i)}) \right] \quad (2)$$

表示隐藏神经元 j 的平均活跃度。为了实现对网络的稀疏性限制,设定参数 ρ 并以和$\hat{\rho}$之间的 KL 距离来表示稀疏性惩罚项：

$$\sum_j \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j} \quad (3)$$

同时，在输入样本中加入随机噪声可以得到更为鲁棒的特征表达，从而使网络具有一定的降噪能力。N. Wang 等人[17]率先将栈式降噪自编码器特征用在了目标跟踪中,删去了自编码器中的解码部分,将其用 softmax 分类器替代,从而实现了对目标和背景的分类。首先,离线训练一个大规模的小样本数据集得到栈式降噪自编码器的基本网络参数；在每一帧图像中挑选出相应的正负样本并根据这些样本的标签对该网络进行在线参数微调。该算法在 VOT 比赛[18]2013 年上取得了第 5 名的好成绩,其优势是利用离线数据库，在很大程度上解决了训练样本不足的问题。然而,作为第一个将深度学习网络运用于目标跟踪的算法,仍然有很大的改进空间,这是因为：(1) 离线预训练时采用的是 32×32 的小样本，这与跟踪时锁定的目标(通常具有不同的尺度)有很大不同，降低了特征的有效性;(2) 提取特征的计算复杂度高,不能很好地满足目标跟踪算法的实时性要求；(3) 将跟踪看作是二元分类问题,容易产生模型错误累积,最终导致跟踪点漂移。

2.1.2 深度卷积特征

卷积神经网络是一种前馈神经网络，由多个卷积层和若干个全联通层组成，与其他深度学习网络相比,采用了共享参数机制,减小了需要估计的参数数量。常见的深度卷积神经网络有 Alexnet、VGG-net、LeNet 等，它们在历届 image net 比赛中都取得了冠亚军的好成绩，成为目标检测和识别领域的标杆性算法。对于目标跟踪中的图像序列而言,不同层的卷积特征往往代表了不同层次的语义信息，将它们有效地组织在一起可大大提升分类器的分类能力。Martin 等人[19]提出将多个不同分辨率下的特征图插值到连续空间域，然后由不同层的卷积特征所对应的置信度图的加权总和求得目标的位置置信度图,实现对目标的跟踪。Y. Qi 等[20]将不同层的卷积特征融合进 KCF 跟踪器中，形成若干个弱分类器,然后通过引进 Hedge 算法把多个弱分类器整合为最后的跟踪器,算法流程如图 1 所示。

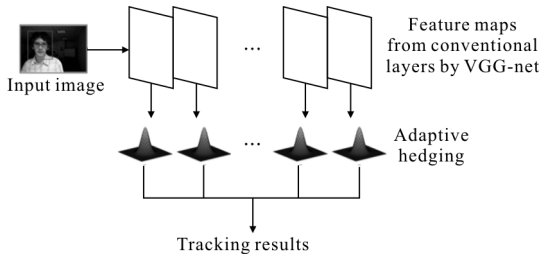

图 1 Hedge 算法流程图

Fig.1 Procedure for Hedge Deep Tracker

L. Wang[21]等将每个不同通道的卷积特征作为一个基学习器,并赋予每个基学习器不同的损失函数进行训练;然后根据各个基分类器之间的相关性来选择尽量不相关的基分类器作为组合,输出最终的跟踪结果。该方法融合了不同通道的特征的优势,这一思想十分值得借鉴。

基于深度卷积特征的相关跟踪算法在近年来的 VOT(Visual Object Tracking)比赛中一直表现优异,这源于深度卷积神经网络强大的表征学习能力和特征泛化能力。与其他基于深度学习的跟踪算法

一样，此类算法需要较大的计算量，难以满足较高的实时性需求；另外，目前的 CNN 特征预训练往往来自 image net 等大规模的图像识别数据库，如何建立一个针对目标跟踪问题且适合作深度学习训练的大规模数据库以获得理想的预训练效果也是一个值得关注的问题。

2.2 面向目标跟踪的深度学习网络

虽然基于深度网络特征的目标跟踪算法取得了十分优异的成果，近年来研究者们仍然希望设计出专门针对目标跟踪问题的深度学习网络，得到计算量小、精准度更佳的目标跟踪算法。

2.2.1 多域卷积神经网络

多域卷积神经网络[22]是 VOT 比赛 2015 年的冠军，该方法设计了一个专门针对跟踪图像序列特性的卷积神经网络，并为不同的视频或图像序列构建了不同的全卷积层，通过多次迭代训练得到跟踪视频图库中的共性深度卷积特征。训练框图如图 2 所示。

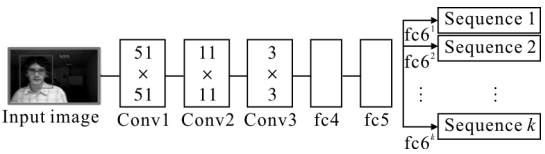

图 2 多域卷积神经网络的流程图

Fig.2 Procedure for MD CNN

和大部分判别式模型一样，多域卷积神经网络将跟踪看作二分类问题，缺乏应对误差累积效应的机制。

2.2.2 树形结构的卷积神经网络

为了缓解判别式模型带来的误差累积效应，参考文献[23]提出了一种针对目标跟踪问题的新型树形深度卷积神经网络算法。

如图 3 所示，树形卷积神经网络考虑了多个卷积神经网络，用它们的线性加权组合来确定目标的位置。该模型每 10 帧增加一个卷积神经网络作为新节点，从之前的卷积神经网络集合中选择可靠性最高的作为自己的父节点，并逐层更新父节点的全链接层权重，同时删去最旧的卷积网络节点。树形卷积神经网络是 VOT2016 的冠军，但由于庞大的计算量，其平均跟踪速度仅为 1.5 fps，远远不能满足实时性要求。

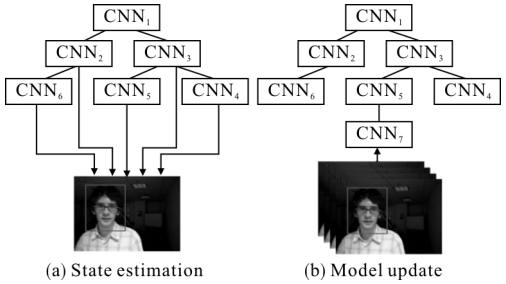

图 3 树形结构的卷积神经网络的状态估计与模板更新过程

Fig.3 State estimation and Model update in TCNN

2.3 深度学习应用于目标跟踪的新思路

近年来，研究者不断探索，不仅仅满足于将目标跟踪问题看作同目标识别问题一样的类别间分类问题，还尝试了细化到个例级别的学习。此外，作为已经在自然语言处理等方向取得巨大突破的深度网络的重要形式，如何将递归神经网络应用于目标跟踪问题也成为目标跟踪算法研究的新思路。

2.3.1 个例学习用于目标跟踪

孪生神经网络是一类由两个或多个具有相同参数和权重的子网络组成的神经网络架构。孪生神经网络在涉及个例之间的相似性度量或两个可比较的事物之间的关系的任务中流行。Tao R.[10]等人将跟踪问题看作模式匹配再认证问题，应用孪生网络从大量的跟踪视频数据集中学习匹配函数，最后选择候选区域中匹配得分最高处作为目标跟踪结果。将跟踪看作单一的模式匹配问题大幅减少了计算量，这类算法很好地满足了实时跟踪的要求，但由于没有充分利用目标的运动轨迹信息，尚有很大的提升空间。

2.3.2 递归神经网络用于目标跟踪

为了缓解跟踪问题中的目标形变和遮挡问题，RTT[24]跟踪器在二维平面上建立了基于多部件的递归神经网络，将候选区域划分为多个大小相等的网格，然后对每一个网格建立 4 个方向的递归神经网络，最后通过融合各个递归神经网络对应的 softmax 分类器的结果得到关于目标位置以及尺度大小的输出。该方法旨在解决跟踪过程中可能产生的误差累积和跟踪点漂移问题，但在目标表征能力和计算速度方面还有很大的提升空间。另一方面，作为在时序分析方面有巨大潜力的深度网络，递归神经网络也被用于寻找帧与帧之间的时序关联。

Guang H.等人[25]通过对单帧视频图像的检测结果建立时序方向的递归神经网络获得跟踪结果,但其采用的是离线检测、在线建立时序方向的思路,尚不能满足目标跟踪的要求。

3 结 论

由于场景和目标变化的复杂性,目标跟踪问题一直是计算机视觉领域最具挑战性的研究方向之一。自 2013 年以来,虽然基于深度学习的目标跟踪算法已经取得了一系列的重大进展,但由于实际场景往往比评测数据复杂,当前的跟踪算法还不能同时满足鲁棒性、实时性和精准度的需要。从跟踪问题的本质出发,目前基于深度学习的跟踪算法在以下 3 个方面仍有较大的提升空间:

(1) 基于深度学习的算法的性能很大程度上依赖于训练数据的数量和好坏,但跟踪问题的难点之一在于样本的缺乏。当前大多数跟踪算法采用用于目标识别的大规模数据库对深度网络进行离线预训练,如何建立面向目标跟踪的有效数据库并将其应用于深度网络的训练,将是未来目标跟踪算法研究的一项重要内容;

(2) 目前大多数基于深度网络的目标跟踪算法只是将问题简单地看作二元分类问题,如能充分利用视频或图像序列中的有效运动信息,将在一定程度上避免跟踪点漂移问题;

(3) 恰当地平衡了深度网络强大的表征能力所需要的计算量和跟踪问题的实时性需求。

参考文献:

[1] Yi lmaz A, Javed O, Shah M. Object tracking [J.] ***ACM Comput Surv***, 2006, 38(4): 13.

[2] Bao C, Wu Y, Ling H, et al. Real time robust L1 tracker using accelerated proximal gradient approach [J]. ***Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit***, 2012, 157 (10): 1830-1837.

[3] Bertinetto L, Valmadre J, Golodetz S, et al. Staple: Complementary Learners for Real -Time Tracking [M]. Oxford: University of Oxford, 2015: 1401-1409.

[4] Luo J, Konofagou E. A fast normalized cross -correlation calculation method for motion estimation [J]. ***IEEE Trans Ultrason Ferroelectr Freq Control***, 2010, 57(6): 1347-1357.

[5] Chang C, Ansari R. Kernel particle filter for visual tracking [J]. ***IEEE Signal Process Lett***, 2005,12(3): 242-245.

[6] Amiaz T, Lubetzky E, Kiryati N. Coarse to over -fine optical flow estimation [J]. ***Pattern Recognit***, 2007, 40(9): 2496-2503.

[7] Comaniciu D, Meer P, Member S. Mean Shift: a robust approach toward feature space analysis [J]. ***IEEE Transaction on Pattern Analasys and Machine Intelligence***, 2002, 24(5): 603-619.

[8] Shi Zelin, Wang Junqing, Huang Shabai. Tracking of deformable objects in complex scene [J]. ***Opto -Electronic Engineering,*** 2005, 32(1): 2-6.

[9] Zhang T, Bibi A, Ghanem B. In defense of sparse tracking: circulant sparse tracker[C]//CVPR, 2016, 3: 1-8.

[10] Tao R, Gavves E, Smeulders A W M. Siamese Instance Search for Tracking[C]//CVPR, 2016: 1420-1429.

[11] Kala l Z, Mikolajczyk K, Matas J. Tracking -Learning -Detection [J]. ***IEEE Transaction on Pattern Analasys and Machine Intelligence***, 2010, 6(1): 1-14.

[12] Hare S, Saffari A, Torr P H S. Struck_ Structured output tracking with kernels [C]//2011 IEEE International Conference on Computer Vision, 2011: 263-270.

[13] Henriques J F, Caseiro R, Martins P, et al. High -speed tracking with kernelized correlation filters [J]. ***EEE Trans Pattern Anal Mach Intell***, 2015, 37(3): 583-596.

[14] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networksr [J]//***Science***, 2006, 313(5786): 504-507.

[15] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. ***Neural Computation***, 2006, 18(7): 1527-1554.

[16] Held D, Thrun S, Savarese S. Learning to track at 100 FPS with deep regression networks[C]//ECCV, 2016, 9905: 749-765.

[17] Wang N, Yeung D Y. Learning a deep compact image representation for visual tracking[C]//Adv Neural Inf Process Syst, 2013: 1-9.

[18] VOT Challenge. VOT2013 Benchmark [EB/OL]. [2017-04-20]: http://www.votchallenge.net/vot2013/.

[19] Danelljan M, Robinson A, Khan F S, et al. Felsberg, Beyond correlation filters: Learning continuous convolution operators for visual tracking [C]//ECCV 2016, 2016, 9909: 472-488.

[20] Qi Y, Zhang S, Qin L, et al. Hedged Deep Tracking [C]//

CVPR, 2016: 4303-4311.

[21] Wang L, Ouyan g W, Wang X, et al. Stct: Sequentially training convolutional networks for visual tracking[C]//***IEEE Conf Comput Vis Pattern Recognit***, 2016: 1373-1381.

[22] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking [J]. ***Computer Science***, 2015: 4293-4302.

[23] Nam H, Baek M, Han B. Modeling and propagating CNNs in a tree structure for visual tracking [C]//European Conference on Computer Vision, 2016: 1-10.

[24] Cui Z, Xiao S, Feng J, et al. Recurrently target-attending tracking [J]. ***IEEE Conf Comput Vis Pattern Recognit***, 2016: 1449-1458.

[25] Ning Guanghan, Zhang Zhi, Huang Chen, et al. Spatially supervised recurrent convolutional neural networks for visual object tracking[C]//CVPR 2016, 2016.