基于多帧一致性修正的自监督孪生网络目标跟踪方法

程 旭[1),2),3)] 刘丽华[1),2)] 王莹莹[1),2)] 余梓彤[3)] 赵国英[3)]

[1)](南京信息工程大学计算机学院 南京 210044)

[2)](南京信息工程大学数字取证教育部工程研究中心 南京 210044)

[3)](奥卢大学机器视觉与信号分析研究中心 芬兰 奥卢 FI-90014)

摘 要 深度学习技术促使目标跟踪领域得到了飞速发展,但有限的标注数据限制了深度模型的高效训练.因此,自监督学习应用于目标跟踪领域来解决模型训练需要大量标注数据的问题.然而,现有基于自监督学习的跟踪器大多提取目标浅层信息,缺乏对目标关键特征的高效表达,且忽视了因目标遮挡等挑战导致的反向验证难度大的问题,致使跟踪精度下降.为解决上述问题,本文提出一种基于多帧一致性修正的自监督孪生网络跟踪方法,由前向多帧反序验证策略、混序修正模块和视觉特征增强模块三部分共同构成.首先,前向多帧反序验证策略从多条路径中自适应选择最优目标轨迹来构造循环一致性损失优化函数,面对目标遮挡、背景干扰、形变等挑战时能够合理规划路径.其次,针对多条路径对同一帧目标预测位置的不一致问题,提出混序修正模块来修正跟踪偏移,增强了前向跟踪时特征提取网络的鲁棒性.此外,视觉特征增强模块通过自适应加权融合目标的全局上下文信息与局部语义特征信息,增强了模型对目标自身特征的表达能力.最后,本文方法在 OTB2013、OTB2015、TColor-128 和 VOT-2018 四个公开数据集上进行了验证.实验结果表明:在光照、形变、背景干扰等复杂场景下,相比于现有 21 种主流跟踪算法,本文方法在四个数据集上的精确度平均提高了 4.6%,比基于自/无监督学习的跟踪器平均提高了 5.8%的精确度.

关键词 视频监控;目标跟踪;自监督学习;循环一致性损失;视觉注意力机制

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02544

A Multi-Frame Consistency Correction Based Self-Supervised Siamese Network Method for Object Tracking

CHENG Xu[1),2),3)] LIU Li-Hua[1),2)] WANG Ying-Ying[1),2)] YU Zi-Tong[3)] ZHAO Guo-Ying[3)]

[1)](*School of Computer Science, Nanjing University of Information Science and Technology, Nanjing* 210044)

[2)](*Engineering Research Center of Digital Forensics, Nanjing University of Information Science and Technology, Nanjing* 210044)

[3)](*Center for Machine Vision and Signal Analysis, University of Oulu, Oulu FI-90014, Finland*)

Abstract Visual object tracking is an important yet challenging task in computer vision with a wide range of applications, such as video surveillance, robotics, action recognition, scene understanding, intelligent transportation, visual navigation, and human-machine interaction, etc. It aims to estimate the state of an arbitrary object in video frames, given the object bounding box in an initial frame. In recent years, deep learning technology has promoted the rapid development in the object tracking field, numerous visual tracking methods based on deep learning have made great progress, especially for Siamese trackers which aim to learn a decision making-based similarity evaluation. Nevertheless, the insufficient labeled data limits the efficient training of deep network model. Therefore, self-supervised learning strategy is applied to the object tracking to

收稿日期:2022-01-04;在线发布日期:2022-09-27.本课题得到国家自然科学基金(61802058,61911530397)、国家留学基金资助项目(201908320175)、中国博士后科学基金资助项目(2019M651650)和江苏省研究生科研与实践创新计划项目(KYCX22_1220)资助.

程 旭(通信作者),博士,副教授,中国计算机学会(CCF)会员,主要研究领域为计算机视觉、模式识别.E-mail:xcheng@nuist.edu.cn.**刘丽华**,硕士研究生,主要研究领域为目标检测与跟踪.**王莹莹**,硕士研究生,主要研究领域为智能系统的对抗攻击.**余梓彤**,博士研究生,主要研究领域为计算机视觉与生物识别安全.**赵国英**,博士,教授,主要研究领域为计算机视觉、视频图像处理、智能人机交互.

solve the problem of model training that requires a large number of labeled data. However, the existing self-supervised trackers mostly extract shallow information of the object and lack the efficient representation of key features of the object. In addition, they also ignore the difficulty of reverse verification caused by the challenges such as object occlusion, resulting in a decrease in tracking accuracy. In order to solve the above problems, a multi-frame consistency correction based self-supervised Siamese network tracking method (MCCSST) is proposed in this paper, which consists of a forward multi-frame reverse order verification strategy, a mixed order correction module and visual feature enhancement module. Firstly, the forward multi-frame reverse order verification strategy can adaptively select the optimal tracking trajectory from multiple paths to construct the cycle-consistency loss optimization function, so as to reasonably avoid the challenges of object occlusion, background clutter, deformation and so on. Secondly, for the problem of inconsistent object localization by multiple paths in the same frame, a mixed order correction module is proposed to correct the tracking drift and enhance the robustness of the object feature extraction, which utilizes temporal information of a video to better focus on the object's own features during the forward tracking. In addition, the visual feature enhancement module, consisting of channel correlation branch, convolution block branch and spatial correlation branch, is utilized to enhance the object features representation ability by adaptively weighted fusing the global context information and local semantic feature information of the object. In order to improve channel category and spatial position information of the object, while suppressing irrelevant background information, we further develop an adaptive feature fusion scheme to fuse multi-dimensional feature maps of three branches. Based on Siamese network architecture, the Discriminant Correlation Filters Network with Vital Feature Enhancement (DCFNet-VFE) is designed as our baseline, and then the object location is achieved through the filter layer. Finally, the proposed method is verified on four public object tracking benchmark datasets: OTB2013, OTB2015, TColor-128 and VOT-2018. The experimental results show that, under the complex scenes (e.g., illumination, deformation, background interference), the accuracy of the proposed method on the four benchmarks is improved by 4.6% on average over the compared twenty-one state-of-the-art trackers, which is an average of 5.8% higher than that of the self/unsupervised learning-based trackers.

Keywords　surveillance; object tracking; self-supervised learning; cycle-consistency loss; visual attention mechanism

1　引　言

目标跟踪技术是计算机视觉领域近十年来活跃且重要的研究热点之一，现已广泛应用于视频监控[1]、智能人机交互[2]、自动驾驶[3]等领域. 目标跟踪定义为给定视频初始帧的目标位置，预测后续视频序列的目标状态. 近年来，研究人员提出了大量行之有效的方法[4-12]，但仍然面临遮挡、形变、运动模糊和光照变化等诸多挑战.

最近，基于深度学习的目标跟踪方法[4-9]在跟踪精度方面大幅优于传统方法，这归功于深度模型强大的特征表达能力. 其先利用预训练的卷积神经网络(Convolutional Neural Network，CNN)提取目标特征，再将目标模板与当前帧搜索区域进行匹配，得到的相似度最大的搜索区域即为目标预测位置. 这些方法的性能在一定程度上得到了提升，但在线跟踪时未引入更新机制，导致模型泛化性差. 牛津大学视觉几何组(Visual Geometry Group，VGG)发布了一系列以 VGG 开头的卷积网络模型. Song 等人[9]提出了采用 VGG[13]作为特征提取网络实现对目标的精确跟踪，利用随机梯度下降(Stochastic Gradient Descent，SGD)策略在线更新模型，但跟踪速度较慢. 类似的方法还包括 ECO (Efficient Convolution Operators)[5]，MDNet (Multi-Domain Network)[14]，C-COT (Continuous Convolution Opera-

tor Tracker)[15]等.上述模型属于监督学习的目标跟踪方法,存在两个明显的缺点:(1)模型训练需要大量标注的数据集,而在有限标注数据下训练的特征提取网络不能够充分提取目标特征,且模型容易过拟合;(2)基于预训练的深度模型[5,9,14-16]需要深层 CNN 来提取目标特征,但在线微调网络导致计算复杂度较高,无法满足实时性需求.

为了解决上述问题,自监督学习技术应用于目标跟踪领域,仅使用视频序列初始帧的目标信息,以基于视频时序的循环一致性作为自监督信号来实现目标跟踪.最初,Wang 等人[17]利用循环一致性损失函数优化目标跟踪模型,直接计算目标初始标签与反向跟踪生成的伪标签之间的差异,但忽略了中间训练样本对的轨迹状态.为此,Yuan 等人[18]提出了多循环一致性目标损失函数,进一步考虑了所有训练样本的前向和反向跟踪生成的伪标签之间的差异,增强了跟踪模型的鲁棒性.上述方法[17-18]通过最小化前向与反向轨迹之间的整体误差来训练跟踪模型,但是前向跟踪轨迹对反向轨迹有密切的影响,当视频序列中间帧存在目标遮挡、形变等挑战时会直接影响前向跟踪的结果,进而造成反向跟踪的初始伪标签偏离目标对象,影响整体模型的全局优化.

针对这一问题,本文提出一种基于多帧一致性修正的自监督孪生网络跟踪方法,由前向多帧反序验证策略、混序修正模块和视觉特征增强模块三部分组成.前向多帧反序验证策略通过交换一组训练样本对的中间相邻帧次序构造不同路径,自主选择最优循环一致性损失函数优化跟踪模型.同时本文还提出混序修正模块和视觉特征增强模块,前者利用相邻帧的时序信息来提高跟踪的鲁棒性,后者增强了特征提取模块对跟踪对象的关键特征提取.

本文主要贡献如下:

(1)提出了一种基于多帧一致性修正的自监督孪生网络跟踪方法(Multi-frame Consistency Correction Based Self-supervised Siamese Network Tracker,MCCSST),由前向多帧反序验证策略、混序修正模块和视觉特征增强模块共同构成.前向多帧反序验证策略通过改变视频帧位置以多条路径进行前向与反向跟踪,自适应选择最优轨迹来优化跟踪模型,解决了目标遮挡、形变等挑战所带来的反向验证难度大的问题,提高了跟踪模型的鲁棒性与泛化性.

(2)针对现有网络难以在复杂场景下充分挖掘目标深层语义特征的问题,本文设计了混序修正模块和视觉特征增强模块来联合增强网络模型对目标的表达能力.混序修正模块关注于视频时序信息,通过多条路径在同一帧生成的不同伪标签来修正跟踪偏移.视觉特征增强模块关联目标特征响应图的空间信息和通道信息,自适应加权融合目标全局上下文信息与局部语义特征信息,使模型关注于目标自身特征,以提高跟踪器的判别能力,有效地降低了跟踪过程中发生跟踪漂移的风险.

(3)本文方法在光照、形变、背景干扰等复杂场景下进行了验证.实验结果表明:相比于现有的主流跟踪方法,本文方法在四个数据集上的精确度平均提高了 4.6%,比基于自/无监督学习的跟踪器平均提高了 5.8%的精确度.

2 相关工作

本节从基于监督学习的目标跟踪方法、基于自/无监督学习的目标跟踪方法以及视觉注意力机制三个方面介绍本文的相关工作.

2.1 基于监督学习的目标跟踪方法

近年来,基于监督学习的目标跟踪模型依赖于大规模标注数据的训练[4-5,7-8,12,15,19-20],性能得到了显著提升,特别是基于孪生网络跟踪方法[4,7-8,12,19-20]的发展,其将跟踪任务阐述为相似度匹配问题,架构如图 1(a)所示.Bertinetto 等人[4]提出 SiamFC(Fully-Convolutional Siamese Networks)跟踪器,其利用共享参数的全卷积网络提取目标模板与视频目标搜索帧的特征,再对其做互相关运算,生成最大响应值作为目标预测位置.但该模型未采取在线微调策略导致泛化能力差,同时跟踪过程中易受到背景中相似目标的干扰.后续,Li 等人[7]提出了基于区域提案网络(Region Proposal Network,RPN)的 SiamRPN 跟踪器,抛弃了传统的多尺度测试与在线微调,实现了对目标的精细定位.该类代表性方法还包括 DaSiam(Distractor-aware Siamese Region Proposal Networks)[21],SA-Siam(A Twofold Siamese Network)[22],SiamCRPN(Siamese Cascaded Region Proposal Networks)[23]等.

基于相关滤波(Correlation Filter,CF)的目标跟踪方法在速度方面有着极大优势,其利用循环矩阵,通过快速傅里叶变换将空间域的训练样本转换到频域计算.经典方法有 KCF(Kernel Correlation Filter)[10],MOSSE(Minimum Output Sum of Squared Error)[11]等.后续,该方法从特征选择[24]、尺度估

计[25-26]、目标分块[27]、边界效应[28-29]和响应自适应[30]五个方面进行了改进，以提升跟踪器的性能. 最近，深度学习和相关滤波相结合的目标跟踪方法表现出更强的鲁棒性，代表性方法包括 ECO[5]，C-COT[15]，HDT (Hedged Deep Tracking)[16]，MCCT (Multi-cue Correlation Filter Tracker)[31]等. Valmadre 等人[20]提出端到端训练的 CFNet (Cascade and Fused Network)跟踪器，将相关滤波器建模为一个附加层融入孪生跟踪网络架构中，实现了利用较少卷积层丰富目标特征的目的. 但是由于迭代时过多的网络参数折戟了 C-COT [15]的跟踪速度，Danelljan 等人[5]提出了 ECO 跟踪器，从简化训练集、模型稀疏更新和高效的卷积操作三个方面来降低 C-COT 模型的复杂度.

虽然上述方法取得了良好性能，但模型训练需要依赖于大规模标注数据集. 在现实生活中，数据标注成本很高，同时若利用过少的标注数据训练会导致模型过拟合且难以高效地提取目标关键信息. 为此，引入自/无监督学习方式可突破有限标注数据集的局限，为真实场景下的目标跟踪理论发展提供了可能.

2.2　基于自/无监督学习的目标跟踪方法

如何利用大量未标记的视频序列训练跟踪模型是自/无监督学习重点解决的问题. 通常，基于自/无监督学习目标跟踪方法利用未标记的视频数据离线训练跟踪模型，再微调模型使其适用于目标跟踪任务. 该方法可分为基于视频帧间相关性目标跟踪方法[32-35]和基于视频时序的循环一致性目标跟踪方法[17-18,36].

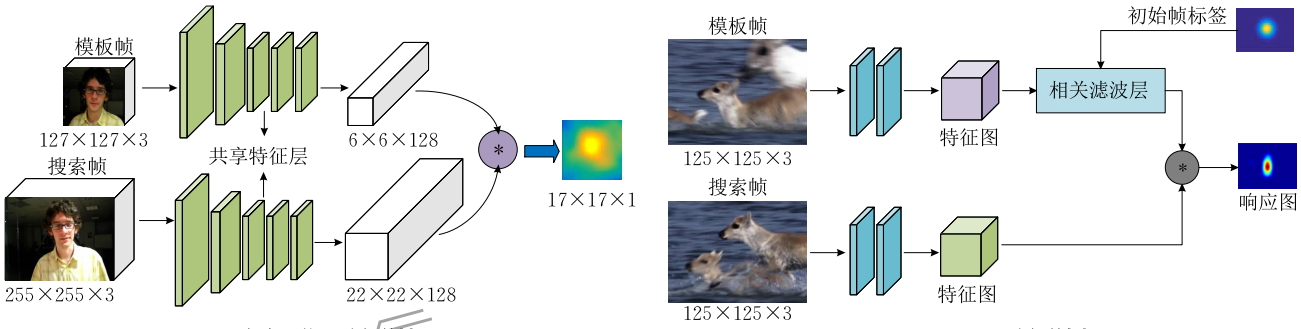
视频帧间相关性指视频本身具有连续两帧图像差异不大的特性，利用这一特性作为自监督信号来学习目标的表观特征. Vondrick 等人[32]采用学习嵌入的方法，以视频着色形式复制模板区域颜色给灰度搜索区域来学习跟踪对象. 在此基础上，文献[33]通过删除颜色通道迫使网络降低对低级色彩信息的依赖，结合预先采样和像素循环一致性策略联合训练递归的跟踪模型，从而大幅提升跟踪性能. Li 等人[34]巧妙地设计了一种共享的帧间亲和矩阵，以区域级和像素级方式分别对视频帧建模，通过在不同层次找到准确的特征对应来提升跟踪性能. 最近，Lai 等人[35]提出了密集跟踪模型(Memory-Augmented Self-Supervised Tracker，MAST)，采用内存组件扩充自监督跟踪架构来关联多个参考帧，有效解决了由误差累积导致的跟踪器漂移问题.

视频时序的循环一致性源自文献[37]，为了评估跟踪结果的可靠性而提出了跟踪-学习-检测(Tracking Learning Detection，TLD)算法，通过衡量跟踪的前向轨迹与反向轨迹之间的误差来评判. 基于此，Wang 等人[17]提出了无监督深度跟踪器(Unsupervised Deep Tracking，UDT)，网络架构如图 1(b)所示，将模板帧与搜索帧经过两层卷积得到特征图，再在傅里叶域实现搜索帧中目标的快速定位. UDT 将目标运动轨迹的循环一致性作为自监督信号训练模型，通过归一化运动信息获得权重系数计算代价损失函数来减少噪声样本，提升了跟踪器的稳定性，奠定了无监督学习的目标跟踪框架. Yuan 等人[18]进一步提出了自监督深度相关跟踪器(Self-supervised Deep Correlation Tracker，Self-SDCT)，采用循环轨迹一致性学习目标特征并引入多循环一致性损失来优化跟踪模型，达到了与监督学习目标跟踪算法[4, 20, 38]相媲美的效果. 针对跟踪过程中目标遮挡、消失后又重现等问题，Wang 等人[36]提出了跨帧识别目标对象的方法，模型自主选择最优长度循环来适应目标对象的局部变换. 上述跟踪器[17-18,36]利用前向与反向跟踪差异构造一致性，然而，前向与反向的轨迹信息密切关联，前向跟踪的失败会导致反向验证难度增大. 因此，本文提出一种高效的基于多帧一致性修正的自监督孪生网络跟踪方法，通过多条路径的自适应选择来保证前向跟踪的有效性，使用混序修正和视觉特征增强来充分挖掘大量未标注视频中的丰富信息，以提升跟踪器的判别能力.

2.3　视觉注意力机制

注意力在人类的视觉信息感知中起着至关重要的作用. 在计算机视觉领域，视觉注意力机制通常用于突出描述图像边缘、纹理、视觉变化等信息. Hu 等人[39]提出 SENet (Squeeze-and-Excitation Networks)来学习响应图通道间的关系，再利用通道间的关联性加权原始特征图. 在此基础上，Woo 等人[40]和 Park 等人[41]提出 CBAM (Convolutional Block Attention Module)和 BAM (Bottlenet Attention Module)，分别以串联和并联的方式关联通道和空间注意力模块，前者采取 Softmax 函数加权响应图通道权重，后者以自注意力形式关注目标特征内部信息，增强了特征间的信息流. 文献[42]借鉴通道注意力模块和多分支卷积感受野的思想，动态计算各卷积核得到相应通道的权重，再自适应地将各个卷积核的结果进行融合，该工作验证了采用多

个不同卷积核学习可更好地提升模型对目标的特征表达能力. Zhang 等人[43]提出了即插即用的 EPSA-Net (Efficient Pyramid Split Attention),其利用多尺度特征图提取空间信息,同时重新修正通道间的注意力权重,在图像分类和目标检测任务上表现出优异的性能. 此外,自注意力模块[44-45]也被应用于各类视觉任务以提取特征图内部空间信息. 受到非局部均值(Non-Local Means)操作的启发,文献[44]提出了一种简单的非局部算子,用于捕获目标的长距离依赖关系,提高了模型的泛化能力. 最近,注意力机制采用自注意力特征融合[46]、多级特征融合[47]、多层感知器(Multilayer Perceptron, MLP)[48]、Transformer[49]等技术来增强特征图之间的信息流.



(a) 孪生网络跟踪架构[4] (b) UDT跟踪架构[17]

图 1 基于监督学习和自/无监督学习的目标跟踪模型架构

3 模型框架

针对现有基于自监督学习的跟踪模型[17,32]在面对复杂场景下前向跟踪定位不准确导致反向验证难度大的问题,本文提出一种基于多帧一致性修正的自监督孪生网络跟踪方法(MCCSST),利用孪生网络跟踪模型进行跟踪以获得目标相应的伪标签. MCCSST 跟踪模型由前向多帧反序验证策略、混序修正模块和视觉特征增强模块三部分构成. 在前向多帧反序验证策略中,一组训练样本对的末尾帧由两条不同路径分别生成两个不同的伪标签,再利用这两个伪标签反向跟踪至初始帧,即在初始帧中存在两个预测值与一个真实值,通过三者之间的循环一致性损失优化跟踪模型. 混序修正模块利用不同路径在同一中间相邻帧上生成的不同跟踪结果来修正跟踪器偏移,以提高跟踪器的鲁棒性. 视觉特征增强模块充分挖掘目标的关键特征,从而提高跟踪模型的辨别能力. 基于多帧一致性修正的自监督孪生网络跟踪方法结构如图 2 所示.

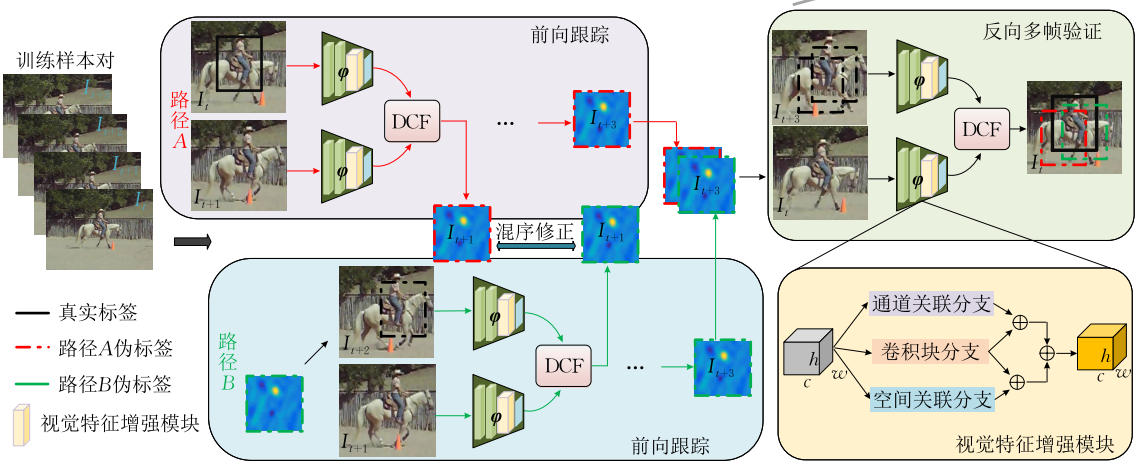


图 2 基于多帧一致性修正的自监督孪生网络跟踪方法结构框图

3.1 问题与动机

最近,自监督学习方法应用于目标跟踪领域,以解决基于有限数据集训练下的模型泛化性差的问题. Wang 等人[17]提出的 UDT,将训练过程分为"前向跟踪"和"反向验证"两部分. 前向跟踪时,使用视频初始帧对后续帧进行预测并生成所有样本对的目

标伪标签；反向验证时，利用“伪标签”对初始帧进行反向跟踪，通过在初始帧构造视频时序的循环一致性作为自监督信号来优化跟踪模型. 具体过程如下.

首先，分别将目标模板 $\boldsymbol{X}$ 和搜索补丁 $\boldsymbol{Z}$ 送入 Siamese Network，得到特征图 $\boldsymbol{\varphi}(\boldsymbol{X})$ 和 $\boldsymbol{\varphi}(\boldsymbol{Z})$；再将判别相关滤波器(Discriminative Correlation Filter，DCF)作为附加层融入到孪生跟踪网络框架中，使用带有真实标签 $\boldsymbol{Y}$ 的模板特征 $\boldsymbol{\varphi}(\boldsymbol{X})$ 训练滤波器 $\boldsymbol{W}$，如式(1)所示：

$$\boldsymbol{W}=\arg\min_{\boldsymbol{W}}\|\boldsymbol{W}\times\boldsymbol{\varphi}(\boldsymbol{X})-\boldsymbol{Y}\|_2^2+\lambda\|\boldsymbol{W}\|_2^2 \tag{1}$$

式中，$\times$ 表示傅里叶频域的卷积运算；λ 为正则化参数. 将式(1)转换为傅里叶变换如下：

$$\boldsymbol{W}=\vartheta^{-1}\left(\frac{\vartheta(\boldsymbol{\varphi}(\boldsymbol{X}))\odot\vartheta'(\boldsymbol{Y})}{\vartheta'(\boldsymbol{\varphi}(\boldsymbol{X}))\odot\vartheta(\varphi(\boldsymbol{X}))+\lambda}\right) \tag{2}$$

式中，ϑ，ϑ^{-1}，$\odot$，ϑ' 分别表示离散傅里叶变换，离散傅里叶逆变换，矩阵元素点乘与复共轭运算.

然后，将训练好的滤波器 $\boldsymbol{W}$ 与搜索补丁特征图 $\boldsymbol{\varphi}(\boldsymbol{Z})$ 进行卷积运算操作，得到响应图 $\boldsymbol{R}$：

$$\boldsymbol{R}=\boldsymbol{W}\times\boldsymbol{\varphi}(\boldsymbol{Z})=\vartheta^{-1}(\vartheta(\boldsymbol{\varphi}(\boldsymbol{Z}))\odot\vartheta'(\boldsymbol{W})) \tag{3}$$

最后以最大响应值 $\boldsymbol{R}$ 作为目标中心，生成响应位置的高斯伪标签. 利用该伪标签和 $\boldsymbol{\varphi}(\boldsymbol{Z})$ 更新滤波器 $\boldsymbol{W}$. 循环往复以上步骤.

然而，UDT 仅利用两帧之间的轨迹信息优化跟踪模型，其跟踪性能受到了严重限制. 主要表现在两个方面：(1) 当视频序列出现目标遮挡、形变等挑战时，模型前向预测得到不正确的伪标签后，再通过反向跟踪至初始模板，造成反向验证难度增大，极大地影响了跟踪模型的全局优化；(2) 面对复杂场景时，其难以在背景相似物干扰下准确提取目标特征，导致跟踪偏移.

为解决上述问题，本文提出基于多帧一致性修正的自监督孪生网络跟踪方法，该方法提出对多条路径进行前向与反向跟踪，再自主选择最优路径解决中间帧存在目标遮挡、形变等带来的反向验证难度大的问题，稳定自监督跟踪模型的训练. 此外，本文还提出混序修正模块和视觉特征增强模块来提高模型对目标关键特征的表达能力.

3.2　前向多帧反序验证策略

本节联合目标前向与反向跟踪过程阐述前向多帧反序验证策略，通过设计多路径前后帧目标轨迹循环一致性损失函数，选择全局最优反序路径来提升面对目标遮挡、形变等复杂场景下的跟踪精度. 前向多帧反序验证策略过程如图 3 所示.

前向跟踪：假定在视频帧 I_t 中给定目标真实标签 $\boldsymbol{Y}_t$，采用孪生相关滤波跟踪器[38]对下一帧 I_{t+1} 进行前向跟踪，预测得到目标的位置和大小：

$$\boldsymbol{R}_{t+1}=\boldsymbol{\varphi}(I_t,I_{t+1},\boldsymbol{Y}_t;\boldsymbol{\theta}) \tag{4}$$

式中，$\boldsymbol{\varphi}$ 为孪生相关滤波跟踪网络；$\boldsymbol{\theta}$ 为网络参数；$\boldsymbol{R}_{t+1}$ 为在 I_{t+1} 帧上跟踪结果的高斯伪标签. 式(4)表示从 I_t 帧到 I_{t+1} 帧完成一次前向传播过程.

本文方法将前向跟踪拓展至多帧，利用视频帧间固有的时序信息，为模型训练提供更多的监督信号. 任取一段视频，每十帧随机选择四帧图像作为一组训练样本对$\{I_t,\ I_{t+1},\ I_{t+2},\ I_{t+3}\}$，跟踪器对每组训练样本中的四帧图像依次进行前向跟踪生成相应的高斯伪标签. 末尾帧生成的伪标签可由式(4)进一步表示为

$$\boldsymbol{R}_{t+3}=\boldsymbol{\varphi}(I_{t+2},I_{t+3},\boldsymbol{\varphi}(I_{t+1},I_{t+2},\boldsymbol{R}_{t+1});\boldsymbol{\theta}) \tag{5}$$

若训练样本对的中间帧出现目标遮挡或消失时，前向跟踪往往会定位失败且生成错误的高斯伪标签. 这一误差将在后续训练中逐渐累积，直接影响训练样本最后一帧的跟踪结果. 如果将这一误差伪标签作为反向跟踪初始值，则会导致反向验证失败.

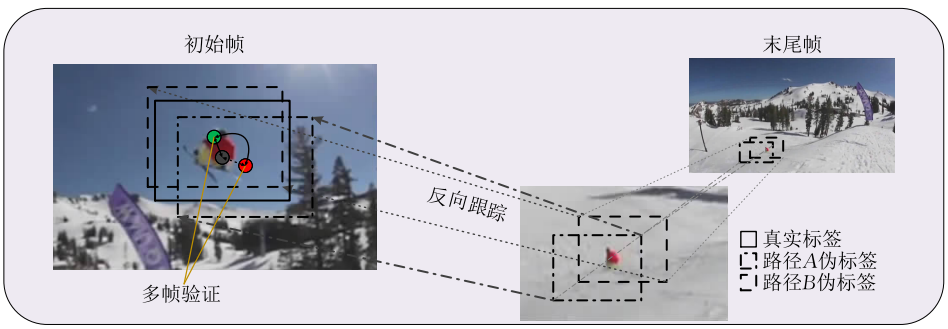


图 3　前向多帧反序验证策略

为了克服这一缺点，本文把一组训练样本对的首尾帧作为参考帧，中间相邻两帧视作时序帧. 交换时序帧 I_{t+1} 和 I_{t+2} 次序，让模型学会利用视频时序信息推理跟踪，即以 $I_t\rightarrow I_{t+2}\rightarrow I_{t+1}\rightarrow I_{t+3}$ 的顺序再

次执行前向跟踪，在末尾帧生成该路径下相应的高斯伪标签，表达如下：

$$\boldsymbol{R}'_{t+3}=\boldsymbol{\varphi}(I_{t+1},I_{t+3},\boldsymbol{\varphi}(I_{t+2},I_{t+1},\boldsymbol{R}'_{t+2});\boldsymbol{\theta}) \tag{6}$$

根据式(5)和式(6)，在一组训练样本对的最后一帧 I_{t+3} 上分别生成两个高斯伪标签 $\boldsymbol{R}_{t+3}$ 和 $\boldsymbol{R}'_{t+3}$. 将生成伪标签 $\boldsymbol{R}_{t+3}$ 的视频帧输入路径记作 T_A；生成伪标签 $\boldsymbol{R}'_{t+3}$ 的输入路径记作 T_B. 本文方法在拓展多帧跟踪的同时，通过改变相邻帧序列来处理目标遮挡、消失、复现等复杂情形，模型学习多条路径下的最优解完成鲁棒的目标跟踪.

反向跟踪：在一组训练样本对的末尾帧 I_{t+3} 上由 T_A 和 T_B 经过前向跟踪分别生成高斯伪标签 $\boldsymbol{R}_{t+3}$ 和 $\boldsymbol{R}'_{t+3}$，再分别以它们作为初始值进行反向跟踪至第一帧 I_t，表达为

$$\begin{aligned}\boldsymbol{R}_{t1}&=\boldsymbol{\varphi}(I_{t+3},I_t,R_{t+3};\boldsymbol{\theta})\\ \boldsymbol{R}_{t2}&=\boldsymbol{\varphi}(I_{t+3},I_t,\boldsymbol{R}'_{t+3};\boldsymbol{\theta})\end{aligned} \tag{7}$$

式中，$\boldsymbol{R}_{t1}$ 和 $\boldsymbol{R}_{t2}$ 表示分别由路径 T_A 和路径 T_B 循环反向跟踪回到第一帧 I_t 生成的高斯伪标签.

前向多帧反序验证：为了解决 UDT 跟踪器中反序验证失效的问题，本文提出通过改变前向跟踪的视频帧输入次序，将反向跟踪在训练样本对的初始帧 I_t 上生成的 $\boldsymbol{R}_{t1}$ 和 $\boldsymbol{R}_{t2}$ 分别与真实标签 $\boldsymbol{Y}_t$ 构造循环一致性损失函数，完成跟踪模型的自监督训练. 同时，$\boldsymbol{R}_{t1}$ 与 $\boldsymbol{R}_{t2}$ 两者本身也应互相影响. 定义损失函数如下：

$$\begin{aligned}\ell_1&=\|\boldsymbol{R}_{t1}-\boldsymbol{Y}_t\|_2^2\\ \ell_2&=\|\boldsymbol{R}_{t2}-\boldsymbol{Y}_t\|_2^2\\ \ell_t&=\|\boldsymbol{R}_{t1}-\boldsymbol{R}_{t_2}\|_2^2\end{aligned} \tag{8}$$

式中，ℓ_1 表示路径 T_A 的反序验证循环一致性损失函数；ℓ_2 为路径 T_B 的反序验证循环一致性损失函数；ℓ_t 表示路径 T_A 与路径 T_B 不同运动轨迹之间的一致性损失函数.

综上所述，前向多帧反序验证的总损失函数定义为

$$\ell=\min(\ell_1,\ell_2)+\alpha\ell_t \tag{9}$$

式中，$\min(\ell_1,\ell_2)$ 表示选择最佳路径的反序验证循环一致性损失函数来优化网络模型. 当两条路径遭受定位不准确时，由于 ℓ_t 中缺少真实标签 $\boldsymbol{Y}_t$ 的信息，无法判定其可靠性，故通过设置权重参数 α 来减小模型训练过程的不稳定性(本文经验设置 $\alpha=0.1$).

3.3 混序修正模块

针对多条路径在同一帧中预测目标位置不一致的问题，本文进一步提出混序修正模块来纠正跟踪偏移，利用视频的时序信息，更好地关注目标自身特征，增强前向跟踪时特征提取网络的鲁棒性. 混序修正模块结构如图 4 所示.

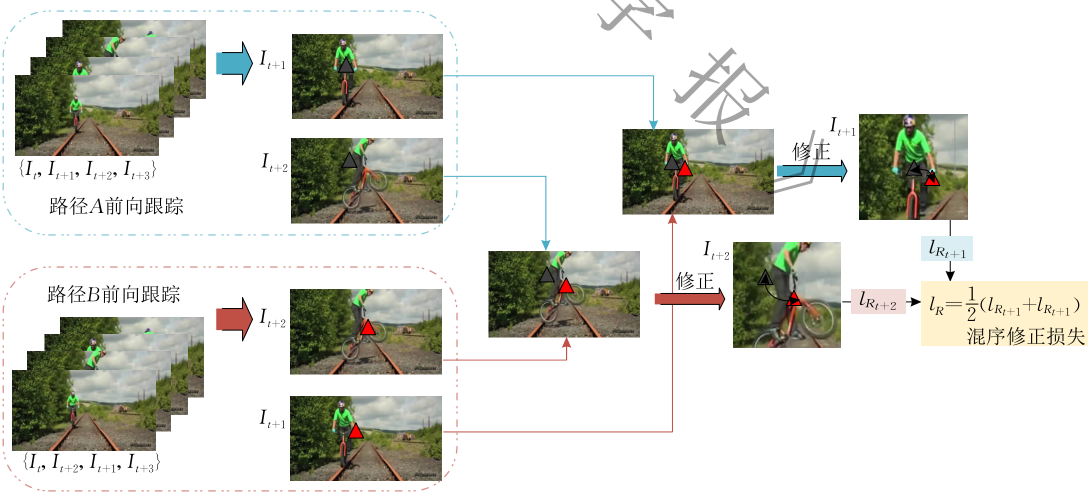


图 4 混序修正模块结构

理想状态下，无论前向跟踪时视频帧的输入路径顺序如何变化，同一帧理应得到相同的跟踪结果. 但实验发现，视频帧的输入顺序对模型训练优化影响极大. 针对这一问题，本文对前向跟踪中由两种路径在同一帧上分别生成的响应图进行比较学习，提取更细致的目标语义特征，及时修正由背景干扰引起的跟踪漂移.

图 4 中，训练样本对依据路径 T_A 进行前向传播，由第一帧 I_t 信息推理获得帧 I_{t+1} 的目标跟踪响应图 $\boldsymbol{R}_{t+1}$；在路径 T_B 中，通过推理对训练样本中第一帧 I_t 与第二帧 I_{t+2} 联合预测得到帧 I_{t+1} 的响应图 $\boldsymbol{R}'_{t+1}$. 如果跟踪器足够鲁棒，$\boldsymbol{R}_{t+1}$ 与 $\boldsymbol{R}'_{t+1}$ 的中心位置(即三角形)应完全一致. 但在复杂场景下，当路径 T_A 中帧 I_{t+1} 的目标发生形变或遮挡时，$\boldsymbol{R}_{t+1}$ 会被极

大地影响，反观 T_B，$\boldsymbol{R}'_{t+1}$ 利用前两帧的时序信息因而更加精确. 因此，本文方法通过约束在帧 I_{t+1} 上得到的两种不同结果，纠正误差较大的响应图，从而降低发生跟踪偏移的风险. 同理，上述过程可推广至由 T_A 和 T_B 在帧 I_{t+2} 上生成的结果响应图 $\boldsymbol{R}_{t+2}$ 与 $\boldsymbol{R}'_{t+2}$. 为此，将同一帧不同路径下的目标定位一致性损失定义为

$$\begin{aligned}\ell_{R_{t+1}} &= (\boldsymbol{R}'_{t+1}-\boldsymbol{R}_{t+1})_2\\ \ell_{R_{t+2}} &= (\boldsymbol{R}'_{t+2}-\boldsymbol{R}_{t+2})_2\end{aligned} \tag{10}$$

进一步，本文将混序修正模块的损失函数定义为

$$\ell_R = \frac{1}{2}(\ell_{R_{t+1}}+\ell_{R_{t+2}}) \tag{11}$$

式中，ℓ_R 表示一组训练样本对中的两帧联合训练的总损失函数.

综上所述，本文将 MCCSST 跟踪模型的训练总损失函数定义为

$$\ell_{总} = \ell_R + \ell \tag{12}$$

3.4　视觉特征增强模块

UDT 跟踪器骨干网络仅依赖两层卷积提取的目标特征是浅层信息，难以在复杂场景下充分挖掘目标的深层语义特征与时空信息. 为解决这一问题，本文提出视觉特征增强模块来丰富目标的关键特征，从响应图的空间关联性和通道关联性提取目标上下文信息与语义特征，完成全局信息与局部信息的目标特征自适应加权融合，图 5 为视觉特征增强模块结构，上中下三个分支分别代表通道关联分支、卷积块分支和空间关联分支.

图 5 中，给定视频帧特征图 $\boldsymbol{U}\in\mathbf{R}^{H\times W\times C}$（$H$，$W$ 和 C 分别表示特征图的高、宽和通道数），将其分别送入通道关联分支、卷积块分支和空间关联分支得到增强后的特征图. 其中，通道关联分支利用全局平均池化（Global Average Pooling，GAP）和 Sigmoid 函数得到不同权重的结果特征图 $\boldsymbol{U}_1\in\mathbf{R}^{H\times W\times C}$. 线性全连接层位于通道关联分支中，为了完整地保留中间层特征信息，本文将中间线性全连接层设置为浅层特征图通道数的八分之一，这在一定程度上减少了计算量. 在卷积块分支中，将 1×1 卷积、最大池化（Max Pooling）和修正线性单元（Rectified Linear Unit，ReLU）组成卷积块来提取视频帧中目标的关键特征得到 $\boldsymbol{U}_2\in\mathbf{R}^{H\times W\times C/8}$；空间关联分支采取自注意力方式提取响应图内部空间关联信息，考虑到响应图中目标与周围上下文信息的关联性与区分性，采用最大池化和上采样操作关联特征图内部空间信息得到 $\boldsymbol{U}_3\in\mathbf{R}^{H\times W\times C/8}$.

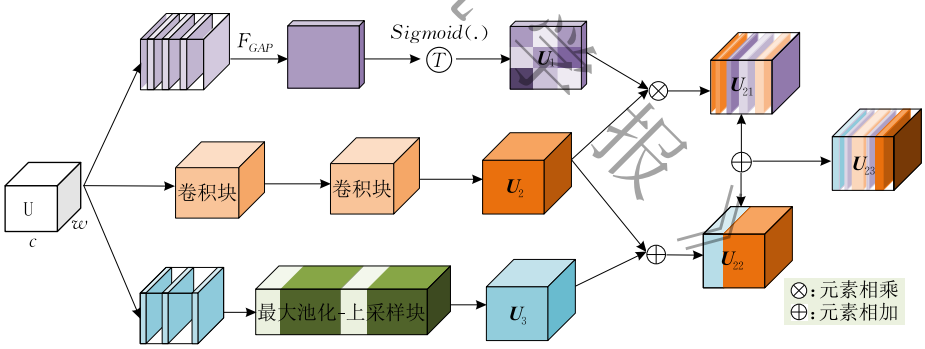


图 5　视觉特征增强模块结构图

为了有效增强跟踪目标的通道类别与空间位置信息，同时抑制无关背景信息，本文提出一种自适应特征融合方式来融合上述三分支的多维度特征图. 首先，通道关联特征图 $\boldsymbol{U}_1$ 和卷积块特征图 $\boldsymbol{U}_2$ 通过逐元素相乘得到通道自适应的响应图 $\boldsymbol{U}_{21}$，该响应图中包含通道信息和全局语义特征；其次，空间关联特征图 $\boldsymbol{U}_3$ 和卷积块特征图 $\boldsymbol{U}_2$ 通过逐元素相加得到响应结果图 $\boldsymbol{U}_{22}$，用以丰富空间维度上特征图的全局语义与局部特征信息；最后，为了匹配通道信息与空间位置信息，采取阈值权重相加的方式得到最终特征响应图 $\boldsymbol{U}_{23}$，如式(13)所示.

$$\boldsymbol{U}_{23} = \gamma(\boldsymbol{U}_1\otimes\boldsymbol{U}_2)\oplus(1-\gamma)(\boldsymbol{U}_3\oplus\boldsymbol{U}_2) \tag{13}$$

式中，⊗和⊕分别表示逐元素相乘和逐元素相加；γ 为预定义的权重参数（默认设置为 0.5）.

3.5　MCCSST 跟踪模型的训练

网络结构：本文提出了关键特征增强判别相关滤波器网络模型（Discriminant Correlation Filters Network with Vital Feature Enhancement，DCFNet-VFE），用于提取目标的关键特征. DCFNet-VFE 模型为孪生网络架构，该架构有两个分支，分别输入模板帧和搜索帧，经过参数共享的 CNN 处理得到相应特征图，后续再经过滤波层实现目标的定位.

在 DCFNet-VFE 网络模型中，前两层卷积的滤

波器尺寸沿用 DCFNet[10] 模型，分别为 3×3×3×32 和 3×3×32×32，使用 ReLU 激活函数来缓解过拟合问题；然后，添加视觉特征增强模块来增强响应图中目标特征的提取与多维度特征图信息的融合；再在卷积层的末端加入局部响应归一化(Local Response Normalization，LRN)层.

模型训练：将一组训练样本对按不同次序分别输入 DCFNet-VFE 网络中进行前向跟踪，在训练样本对的末尾帧生成两条路径的不同预测结果，再分别将这两条路径得到的结果作为伪标签反向跟踪到初始模板帧. 在训练过程中，对不同路径下一组训练样本对的中间两个相邻时序帧联合优化. 通过联合前向多帧反序验证模块与混序修正模块之间的损失训练优化跟踪模型，直至模型收敛. 算法 1 总结了 MCCSST 跟踪模型的训练过程.

算法 1. MCCSST 跟踪模型的训练.

输入：未标注视频序列，N 组训练样本对，T 次迭代

输出：跟踪网络 $\boldsymbol{\varphi}_{\theta}(\cdot)$

1. 从原始未标记的视频帧 I_i 中裁剪中心补丁 P_i；
2. 采用随机权重 $\boldsymbol{\theta}$ 初始化 CNN 模型 $\boldsymbol{\varphi}_{\theta}(\cdot)$；
3. FOR $t=1$：T DO
4. FOR $i=1$：N DO
5. 将 P_i 输入特征提取网络； //视觉特征增强模块
6. 前向跟踪：依据路径 T_A，依次获得 $\boldsymbol{R}_{t+1}$、$\boldsymbol{R}_{t+2}$、$\boldsymbol{R}_{t+3}$；
依据路径 T_B，获得 $\boldsymbol{R}'_{t+2}$、$\boldsymbol{R}'_{t+1}$、$\boldsymbol{R}'_{t+3}$；
7. 反向跟踪：利用式(7)在 P_i 上获得响应图 $\boldsymbol{R}_{t1}$ 与 $\boldsymbol{R}_{t2}$；
8. 利用式(8)计算 $\boldsymbol{R}_{t1}$、$\boldsymbol{R}_{t2}$ 与真实标签 $\boldsymbol{Y}_t$ 之间的循环一致性损失，以及 $\boldsymbol{R}_{t1}$ 与 $\boldsymbol{R}_{t2}$ 的损失；
9. 由式(9)选择最佳路径获得前向多帧反序验证模块的总损失函数 l；//前向多帧反序验证策略
10. 利用式(10)纠正 $\boldsymbol{R}_{t+1}$ 与 $\boldsymbol{R}'_{t+1}$ 和 $\boldsymbol{R}_{t+2}$ 与 $\boldsymbol{R}'_{t+2}$ 在不同路径下生成的跟踪结果差异； //混序修正模块
11. 由式(12)优化跟踪模型 $\boldsymbol{\varphi}_{\theta}(\cdot)$；
12. END
13. END

3.6 在线跟踪

跟踪时，MCCSST 模型以前向跟踪的方式在线跟踪. 首先，将目标模板帧与搜索帧分别输入孪生网络 DCFNet-VFE 的两个分支，经过参数共享的特征层进行特征提取得到相应的特征图，再将相关滤波器与对应的搜索帧特征图进行卷积运算操作，得到预测响应图，并以最大响应值为目标中心，生成响应位置的高斯伪标签，作为目标跟踪的结果.

同时，为了使跟踪模型能够自适应外界复杂环境对目标表观的影响，采取线性加权策略来逐帧更新相关滤波层的模板参数，避免发生跟踪漂移，实现对目标对象的快速定位. 参数更新如下：

$$\boldsymbol{W}_t = (1-\alpha_t)\boldsymbol{W}_{t-1} + \alpha_t \boldsymbol{W} \tag{14}$$

式中，$\alpha_t \in [0,1]$ 为线性插值系数，$\boldsymbol{W}$ 表示当前相关滤波器. 同时本文采用尺度因子 α^s 为 $\{a^s \mid a=1.015, s=\{-1,0,1\}\}$ 的金字塔策略估计目标对象的尺度.

本文 MCCSST 算法流程如图 6 所示.

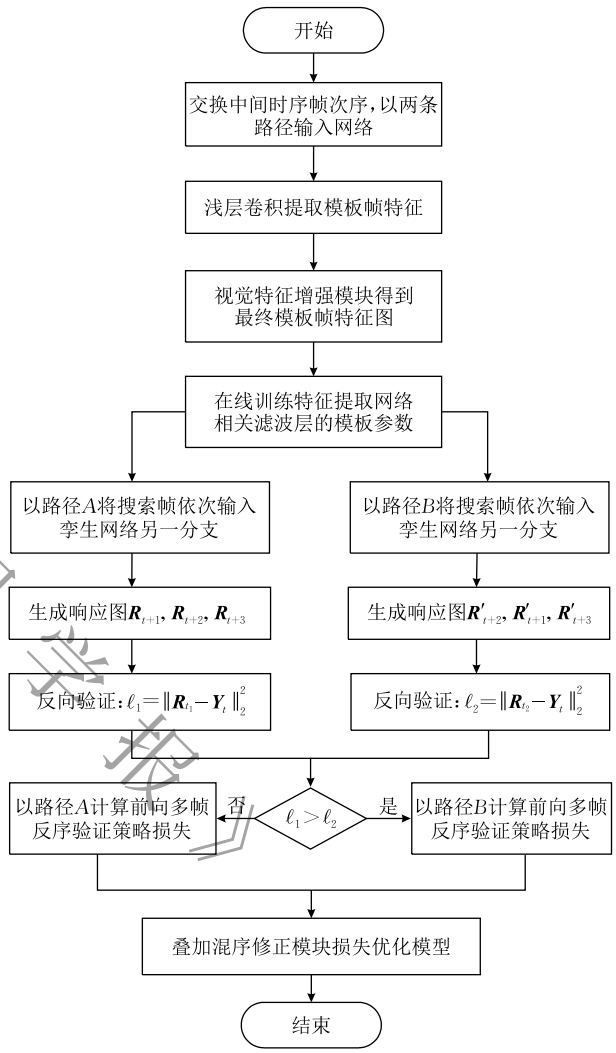


图 6 基于多帧一致性修正的自监督孪生网络跟踪方法流程图

4 实验结果及分析

4.1 实验细节

参数设置：本文方法在 Pytorch 深度学习架构下开展实验验证. 硬件平台配置环境为 RTX-2080Ti GPU (11GB memory) 和 Intel-i9 CPU(64GB memory). 为了与监督学习跟踪器[4,7,38] 和自/无监督学习跟踪器[17-18] 进行公平比较，本文选取 ILSVRC2015[50] 作为训练数据集，该数据集包含 30 种基本类别，充分考虑了

运动类型，背景干扰等多种因素，训练视频序列共计3862 段，验证视频为 555 段.

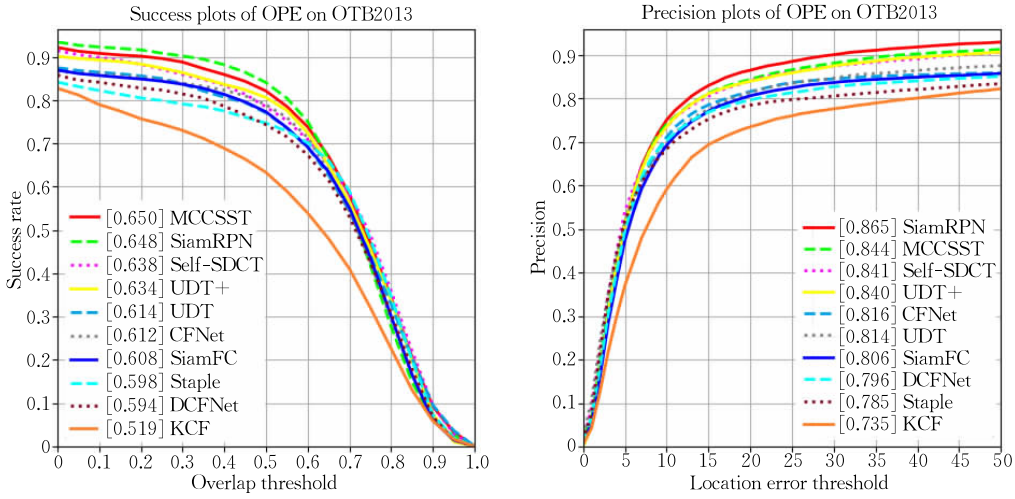
首先确定视频序列每一帧图像尺寸为 1280×720×3 像素的中心位置，再以该位置为中心裁剪得到大小为 640×360×3 的中心补丁. 然后，依据跟踪模型输入所需要的图像尺寸(125×125×3)和中心补丁尺寸确定宽高缩放比例，并利用其计算图像像素的平移距离，进而生成相应的映射矩阵，最后利用仿射变换将图像尺寸调整为 125×125×3，将其作为跟踪模型的输入. 预处理后，视频每 10 帧随机采样 4 帧裁剪图像作为一组训练样本对，将初始帧设为模板图像，其余为搜索图像. 以模板图像的中心作为跟踪目标，并给出真实标签. 训练过程只需迭代 25 次即可收敛；批量大小设置为 32；学习率设置为 0.0001；SGD 优化算法的动量为 0.9；权重衰减设置为 0.005 来训练优化跟踪模型.

评估标准：选取 OTB2013[51]、OTB2015[52]、TColor-128[53]和 VOT-2018[54]四个数据集来评价跟踪器的性能. 这些数据集中包含各种复杂性挑战，如遮挡、形变、光照变化以及快速运动等，且训练数据集与测试数据集无交叉. OTB 数据集采用精确度(Precision，P)和成功率(Success Rate，SR)作为评估准则. P 表示跟踪算法预测的目标位置中心点与人工标注目标中心点的中心误差；SR 代表跟踪算法得到的预测目标框与目标原始边界框重合率大于 0.5 的百分比. VOT-2018 数据集同时衡量算法的精确度(Accuracy，A)、鲁棒性(Robustness，R)以及平均重叠期望(Expect Average Overlap，EAO). A 表示跟踪算法的准确度；R 为评价跟踪算法稳定性的指标；EAO 是衡量算法精确性的指标. TColor-128 数据集选择指标 P 和成功率图的曲线下面积(Area Under Curve，AUC)来衡量算法性能.

4.2 定量评估

本文 MCCSST 方法在 OTB2013、OTB2015、TColor-128 和 VOT-2018 共 4 个基准数据集上进行了定量评估，参与评估的跟踪方法包括：监督学习跟踪器 SiamFC[4]、ECO[5]、SiamRPN[7]、HDT[16]、CFNet[20]、DCFNet[38]、SiamATL[55]、TADT[56]、SiamDW[57]和自/无监督跟踪器 KCF[10]、UDT[17]、UDT+[17]、Self-SDCT[18]、Staple[58]、ARCF[59]、SITUP[60]、DSST[61]、S^2 SiamFC[62]、CycleSiam[63]、AlexPUL[64]、CCCUT[65].

OTB2013 数据集：该数据集由 50 段完整注释的视频序列组成，全部视频序列涉及目标跟踪的 11 种属性，包括：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、出视野、背景干扰和低像素，且每段视频序列至少包含两个或多个属性. 图 7 为 MCCSST 在 OTB2013 数据集上的实验结果. 由图 7 可知，本文 MCCSST 方法比 9 种主流跟踪算法在成功率指标上平均提高 4.3%，精度平均高出 3.3%. 同时，MCCSST 跟踪性能大幅超越了主流的自监督学习跟踪器，特别是与当前性能较好的自监督 UDT 跟踪器相比，本文方法在成功率和精度方面分别提高了 3.6%与 3.0%，这得益于本文设计的多路径自适应选择策略，该策略在面对复杂环境挑战时能够合理规划路径，以避免反向验证的难度增大.

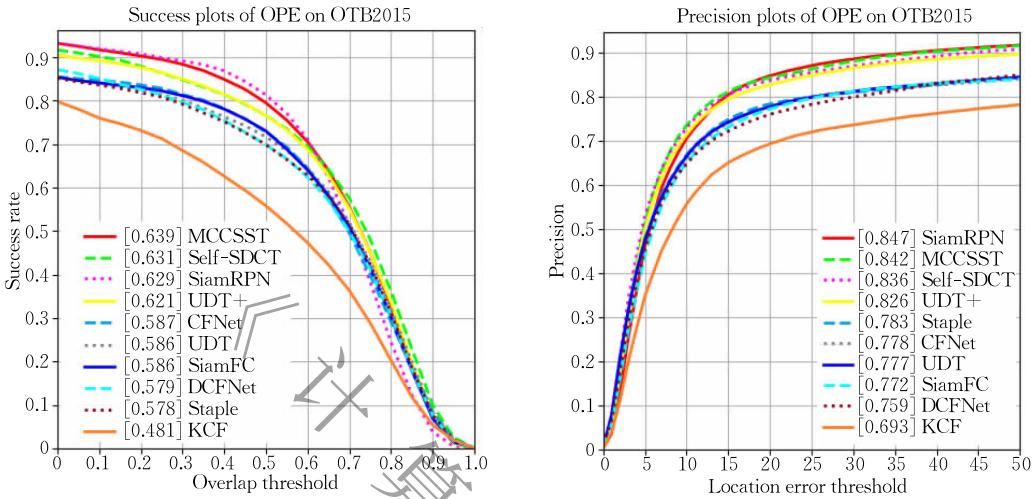


(a) 跟踪成功率曲线　　(b) 跟踪精度曲线

图 7　主流跟踪算法在 OTB2013 数据集上的精度和成功率

OTB2015 数据集：该数据集是 OTB2013 数据集的进一步扩充，拥有 98 段视频，共计 100 个测试场景，涉及到灰度图像和彩色图像. 图 8 为 MCCSST 在 OTB2015 数据集上的实验结果.

由图 8 可知，本文 MCCSST 方法在精度和成功率指标上均优于其他主流跟踪器，精度和成功率分别达到了 0.842 和 0.639，与基于自/无监督学习的跟踪器相比，本文方法平均提高了 5.9%的精度和 6.0%的成功率. 这是因为 MCCSST 通过纠正同一帧中多条路径生成的伪标签偏差来降低跟踪过程中发生跟踪偏移的风险，因而，MCCSST 跟踪性能达到了最佳，比基于监督学习的 SiamRPN 跟踪器高出 1%的成功率.



(a) 跟踪成功率曲线 (b) 跟踪精度曲线

图 8 主流跟踪算法在 OTB2015 数据集上的精度和成功率

TColor-128 数据集：该数据集认为对于视觉推理颜色信息可以提供更丰富的判别线索. 因此，为了评估基于颜色信息的目标跟踪算法性能，其设计包含了 128 段彩色视频序列，部分序列与 OTB 数据集重合，但更具挑战性. 表 1 为主流目标跟踪器在 TColor-128 数据集上的实验结果. 由表 1 可知，MCCSST 跟踪性能比所有参与比较的跟踪器在精度和 AUC 上平均提高了 6.2%和 4.8%. 而与 SITUP 跟踪器相比，本文 MCCSST 方法在精度和 AUC 上分别高出 10.4%和 8.1%，这是由于 SITUP 跟踪器使用相关滤波特征提取方法，对目标关键特征的提取能力较弱，而本文提出的视觉特征增强模块，从响应图的空间关联性和通道关联性提取目标上下文信息与语义特征，完成全局信息与局部信息的目标特征自适应加权融合，极大的丰富了目标的关键特征. 实现了具有竞争力的跟踪性能.

VOT-2018 数据集：该数据集包含 60 段视频序列，以短视频序列为主且视频均为彩色序列，难于 OTB 数据集. 在目标丢失时，该数据集有重新初始化机制. 表 2 给出了 MCCSST 在 VOT-2018 数据集上的实验结果. 由表 2 可知，本文方法在精度、鲁棒性和 EAO 三个指标上性能均优于所有的自/无监督学习跟踪器，MCCSST 在指标 A 和 EAO 上平均高出 5.9%和 3.4%，在 R 指标方面平均优于 5.6%. 特别地，与 CycleSiam 跟踪器相比，MCCSST 方法在准确度方面提高了 14.2%. 原因在于 CycleSiam 跟踪器仅通过前向与反向构造一致性损失来优化网络模型，而本文提出的混序修正模块利用视频间的时序信息，极大提升了特征提取网络的辨别能力. 同时，本文 MCCSST 方法的性能也优于最新的无监督跟踪器 AlexPUL，利用了混序修正模块和视觉特征增强模块联合增强网络模型对目标的表达能力和背景辨别能力，可更好的适应于复杂场景.

表 1 跟踪算法在 TColor-128 数据集上的结果(粗体为监督跟踪器最好性能，粗斜体为自/无监督跟踪器最好性能)

跟踪算法	自/无监督学习跟踪算法						监督学习跟踪算法				
	MCCSST	UDT	Self-SDCT	DSST	ARCF	SITUP	SiamFC	HDT	CFNet	TADT	SiamATL
P	***0.743***	0.658	0.729	0.535	0.709	0.639	0.694	0.686	0.607	0.759	**0.794**
AUC	***0.551***	0.507	0.540	0.405	0.525	0.470	0.505	0.480	0.456	0.563	**0.577**

表 2　跟踪算法在 VOT-2018 数据集上的结果(粗体为监督跟踪器最好性能,粗斜体为自/无监督跟踪器最好性能)

目标跟踪算法	监督学习	自/无监督学习	A(↑)	R(↓)	EAO(↑)
DCFNet[38]	√		0.470	0.543	0.180
ECO[5]	√		0.484	**0.276**	**0.280**
SiamFC[4]	√		0.503	0.585	0.188
SiamDW[61]	√		**0.525**	0.412	0.269
SiamATL[59]	√		0.514	0.549	0.247
KCF[10]		√	0.447	0.773	0.135
S^2SiamFC[62]		√	0.463	0.782	0.180
CycleSiam[63]		√	0.377	0.750	0.131
AlexPUL[64]		√	0.515	0.693	0.182
CCCUT[65]		√	0.499	0.716	0.161
MCCSST		√	***0.519***	***0.687***	***0.192***

4.3　定性分析

为了更直观地展示本文 MCCSST 跟踪算法的实际效果,从 OTB2015 数据集中选取五个具有代表性的视频序列进行可视化效果展示,分别为 Tiger、Lemming、Car、Football 和 Skating. 视频帧中包含了大多数跟踪场景中遇到的挑战,如复杂背景、目标遮挡、快速移动和尺度变化等. 图 9 给出了 MCCSST 方法与其他主流跟踪算法的跟踪结果.

由图 9 可知,在视频序列 Tiger、Car 和 Football 中,面对目标快速移动、尺度变化和复杂背景挑战时,本文 MCCSST 跟踪方法明显优于 SiamDW[57],

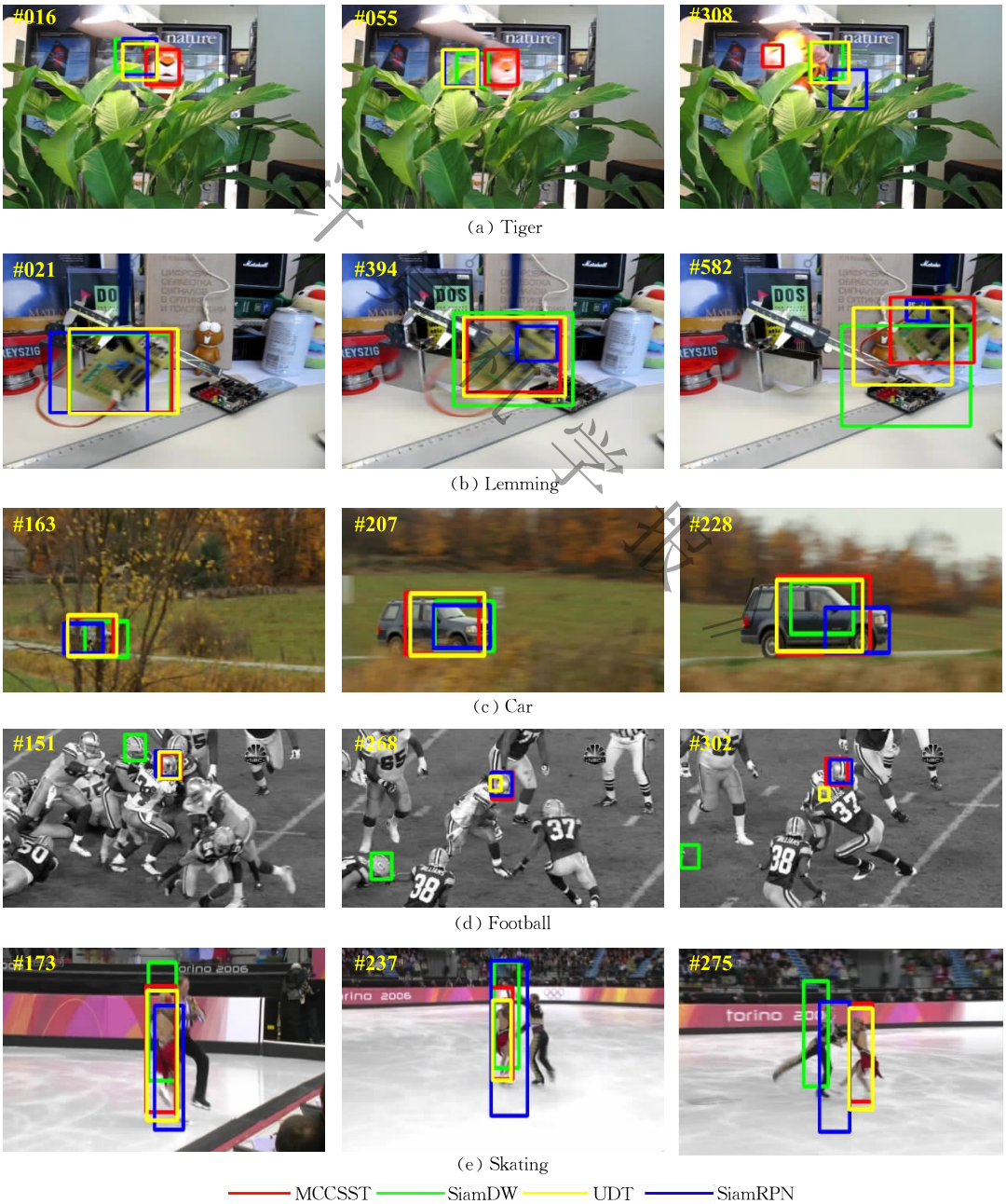


图 9　四种代表性跟踪算法的可视化跟踪结果

这得益于本文提出的混序修正模块和视觉特征增强模块.与 SiamDW 相比,混序修正模块更加关注视频时序信息,同时视觉特征增强模块也提升了模型的辨别能力.因此,本文 MCCSST 跟踪方法可更好的自适应于复杂场景,跟踪性能明显优于其他跟踪算法.即使在面对目标遮挡、快速运动和背景干扰的复杂情况下,本文方法依然能够准确定位目标.

4.4 消融实验

4.4.1 各模块的性能影响

本文在数据集 OTB2013 和 OTB2015 上通过 3 组对比实验详细分析了训练过程中前向多帧反序验证策略、混序修正模块和视觉特征增强模块这 3 个模块对跟踪模型性能的影响,从而验证本文所提出的 MCCSST 跟踪算法的有效性.其中,FMR 表示前向多帧反序验证策略;MOC 为混序修正模块;VFE 为视觉特征增强模块.

由表 3 可知,当三个模块同时作用于跟踪模型的训练时,跟踪效果最佳,AUC 和精确度在 OTB2013 数据集上达到 0.650 和 0.844;在 OTB2015 数据集上达到了 0.639 和 0.842.当三个模块均不参与训练时,模型退化为 UDT 跟踪器,实验结果表明各方面跟踪性能都降为最低.进一步,由表分析知,每个模块对跟踪性能均有积极影响,删除三个模块中的任何一个,跟踪性能均会降低.当 VFE 单独存在时,其在 OTB2015 数据集上的 AUC 为 0.627;当 FMR 和 VFE 同时存在时,AUC 在 OTB2015 数据集上仅次于三个模块同时存在,可达 0.633,这充分验证了 FMR 和 VFE 的有效性;而当 FMR 和 MOC 同时存在时,AUC 在 OTB2015 数据集上可达到 0.630.该实验也直接反映了 MCCSST 跟踪算法中前向多帧反序验证策略、混序修正模块和视觉特征增强模块三个部分的重要性.

表 3 MCCSST 不同模块性能比较(粗体为最好性能)

FMR	MOC	VFE	OTB2013		OTB2015		
			P	AUC	FPS	P	AUC
√	√	√	**0.844**	**0.650**	**0.842**	**0.639**	52
—	—	—	0.829	0.627	0.823	0.618	**65**
√	—	√	0.840	0.642	0.837	0.633	58
√	√	—	0.837	0.638	0.831	0.630	54
—	—	√	0.834	0.633	0.826	0.627	61

4.4.2 视觉特征增强模块

为了验证视觉特征增强模块中通道关联分支、卷积块分支和空间关联分支的有效性,本节在 OTB2015、TColor-128 和 VOT-2018 三个基准数据集上做了 3 组对比实验,评价指标均为精确度.如表 4 所示.

表 4 视觉特征增强模块不同分支性能比较

通道关联分支	卷积块分支	空间关联分支	OTB2015	TColor-128	VOT-2018
√	√	√	**0.842**	**0.743**	**0.519**
√	√	—	0.830	0.729	0.507
—	√	√	0.836	0.738	0.511

实验分析了三个分支不同搭配下本文方法的跟踪性能:(1)三个分支同时存在,即视觉特征增强模块;(2)仅叠加通道关联分支和卷积块分支;(3)仅叠加空间关联分支和卷积块分支.由表 4 可知,完整的采用三个分支自适应融合,在三个数据集上精确度都达到最高.当删除任意分支时,跟踪性能均有所下降,如删除空间关联分支,在数据集 TColor-128 上精确度下降 1.4%;删除通道关联分支,在 VOT-2018 上精确度下降 0.8%.从而验证了视觉特征增强模块各个分支对 MCCSST 跟踪模型性能的有效性.

4.4.3 帧选取范围的影响

本文在数据集 OTB2015 上对一组训练样本对的总帧数选取范围对模型性能的影响进行了四组对比实验,分别将总帧数拓展为 8 帧、10 帧、12 帧和 15 帧,通过指标 P 和 AUC 评估,结果如表 5 所示.

表 5 不同总帧数选取范围模型的性能比较(粗体表示最优结果)

选取帧数/总帧数	4/8	4/10	4/12	4/15
P	0.833	**0.842**	0.837	0.824
AUC	0.634	**0.639**	0.635	0.618

由表 5 可知,总帧数由 8 帧扩展至 10 帧有利于本文 MCCSST 跟踪模型的前向多帧反序验证和混序修正的优化;当总帧数选取为 12 时,性能较 10 帧略有降低;若将总帧数设置为 15 帧时,性能会大幅降低,原因在于帧间隔过大不利于网络学习连续目标特征,同时也难以提取目标关键特征导致模型无法有效辨别目标表观信息.而每 10 帧随机选择 4 帧图像作为一组训练样本对的跟踪性能最好,精度和 AUC 分别达到 0.842 和 0.639,远远超过每 15 帧随机选择 4 帧图像作为一组训练样本对时的跟踪性能.

进一步,本文选取帧数为 4 作为一组训练样本对是充分考虑模型的各模块功能.若将选取帧数设置为 3,将导致 MCCSST 跟踪模型的前向多帧反序验证模块与混序修正模块同时失效;若扩大选取帧数为 5 或 6,将极大增加模型的复杂度,同时导致模型训练的可靠性降低.因此,综合衡量模型的优化难易与性能优劣,将一组训练样本对的选取帧数设置为 4.

4.4.4　MCCSST 模型的复杂度分析

MCCSST 模型的计算量主要集中在特征提取与增强过程.模型前向跟踪时仅通过模板匹配获取下一帧的目标位置,该过程不参与反向传播和更新网络参数.此外,训练过程中采用批处理与并行处理策略,使其在原先基准模型上仅增加了少许参数量,该参数量远远小于特征提取网络,在几乎不增加模型训练代价的同时最大化提升跟踪性能.进一步,在通道关联分支中,本文将中间线性全连接层设置为浅层特征图通道数的八分之一,这在一定程度上也减少了计算量.

实时性对于衡量跟踪器的性能至关重要,因此在 OTB2015 数据集上对提出的 MCCSST 速度进行了分析,并与 Self-SDCT[18]、UDT＋[17]、SiamFC[4]、SiamATL[59]等 8 种代表性跟踪器进行了比较,结果如表 6 所示.由表 6 可知,MCCSST 跟踪算法速度达到 52FPS,比无监督跟踪器 Self-SDCT 高出 4FPS,几乎与 UDT 升级版 UDT＋持平.相对于 AUC 而言,虽然本文 MCCSST 跟踪器性能略低于 SiamATL 跟踪器,但速度却是其近 2.5 倍.

表 6　主流跟踪算法在 OTB2015 数据集的速度(粗体和粗斜体分别为监督学习跟踪器和自/无监督学习跟踪器的最好性能)

跟踪器	自/无监督学习跟踪器					监督学习跟踪器			
	MCCSST	UDT＋	Self-SDCT	CycleSiam	CCCUT	DaSiam	SiamFC	DCFNet	SiamATL
FPS	52	55	48	59	***68***	25	**86**	70	21

同时,为了更直观地验证本文提出的 FMR、MOC 和 VFE 三个模块对模型速度的影响,表 3 给出了相应模块的跟踪速度.由表 3 可知,当三个模块均不存在时,跟踪速度略有提升但跟踪性能最低.三个模块相继加入后,跟踪性能均得到了提升,跟踪速度略微降低.当三个模块全部存在时,其跟踪性能达到最高,且跟踪速度达到 52FPS,满足实时性要求.

此外,从计算复杂度考虑,本文采用浮点运算次数(FLoating-point Operations,FLOPs)和模型参数量(Parameters)分析了提出的 DCFNet-VFE 网络的复杂度,如表 7 所示.

表 7　模型的浮点运算次数与参数量指标

模型	FLOPs/GB	Parameters/MB
AlexPUL	0.647	2.929
ResPUL	2.650	1.445
DCFNet-VFE	1.247	0.858

由表 7 可知:DCFNet-VFE 网络的 FLOPs 为 1.247 GB,介于 AlexPUL[64]和 ResPUL[64]之间.且 DCFNet-VFE 模型参数量仅为 0.858 MB,是 3 个模型中最低的,这得益于轻量化网络模型 DCFNet 和视觉特征增强模块.从而也验证了本文 DCFNet-VFE 模型的复杂度不高,为进一步的落地应用提供了可能.

5　总　结

本文提出一种基于多帧一致性修正的自监督孪生网络跟踪方法,由前向多帧反序验证策略、混序修正模块和视觉特征增强模块三个模块构成.前向多帧反序验证策略通过自适应选取多条路径中的最优目标轨迹来优化跟踪模型.混序修正模块通过限制多条不同路径在同一帧生成的不同响应图来修正跟踪偏移,从而提高前向跟踪的鲁棒性.视觉特征增强模块利用自适应特征加权方法融合目标的全局上下文信息与局部语义特征信息,克服了原有自监督目标跟踪方法由于浅层网络无法完整表征目标特征的不足.最后,本文方法在四个公开数据集上进行了实验,结果验证了本文方法的有效性.

致　谢　本论文的数值计算得到了南京信息工程大学高性能计算中心的计算支持和帮助.同时感谢编辑部和审稿专家给予本文的宝贵意见.

参　考　文　献

[1] Huang Kai-Qi, Chen Xiao-Tang, Kang Yun-Feng, Tan Tie-Niu. Intelligent visual surveillance:a review. Chinese Journal of Computers, 2015, 38(6): 1093-1118(in Chinese)
(黄凯奇, 陈晓棠, 康运锋, 谭铁牛. 智能视频监控技术综述. 计算机学报, 2015, 38(6): 1093-1118)

[2] Choi C, Christensen H I. Real-time 3D model-based tracking using edge and keypoint features for robotic manipulation//Proceedings of the IEEE International Conference on Robotics & Automation. Anchorage, USA, 2010: 1050-4729

[3] Choi W. Near online multi-target tracking with aggregated local flow descriptor//Proceedings of the International Conference on Computer Vision. Santiago, Chile, 2015: 3029-3037

[4] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-Convolutional siamese networks for object tracking//Proceedings of the European Conference on Computer Vision Workshops.

Amsterdam, Netherlands, 2016: 850-865

[5] Danelljan M, Bhat G, Khan F S, et al. ECO: Efficient convolution operators for tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2016: 1063-6919

[6] Li B, Wu W, Wang Q, et al. SiamRPN++: Evolution of siamese visual tracking with very deep networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4277-4286

[7] Li B, Yan J, Wu W, et al. High performance visual tracking with siamese region proposal network//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8971-8980

[8] Zhang Y, Wang L, et al. Structured siamese network for real-time visual tracking//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 351-366

[9] Song Y, Chao M, Wu X, et al. VITAL: Visual tracking via adversarial learning//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8990-8999

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

[11] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA, 2010: 2544-2550

[12] Liang Z, Shen J. Local semantic siamese networks for fast tracking. IEEE Transactions on Image Processing, 2019, 29(99): 3351-3364

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

[14] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1063-6919

[15] Danelljan M, Robinson A, Khan F S, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking//Proceedings of the European Conference on Computer Vision. Amsterdam, Holland, 2016: 472-488

[16] Qi Y, Zhang S, Qin L, et al. Hedged deep tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4303-4311

[17] Wang N, Song Y, Ma C, et al. Unsupervised deep tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1308-1317

[18] Yuan D, Chang X, Huang P Y, et al. Self-supervised deep correlation tracking. IEEE Transactions on Image Processing, 2020, 30: 976-985

[19] Wang Q, Zhang L, Bertinetto L, et al. Fast online object tracking and segmentation: a unifying approach//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1328-1338

[20] Valmadre J, Bertinetto L, Henriques J F, et al. End-to-end representation learning for correlation filter based tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 1063-6919

[21] Zhu Z, Wang Q, Li B, et al. Distractor-aware siamese networks for visual object tracking//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 103-119

[22] He A, Luo C, Tian X, et al. A twofold siamese network for real-time object tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4834-4843

[23] Fan H, Ling H. Siamese cascaded region proposal networks for real-time visual tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 7944-7953

[24] Danelljan M, Khan F S, Felsberg M, et al. Adaptive color attributes for real-time visual tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1090-1097

[25] Ding G, Chen W, Zhao S, et al. Real-time scalable visual tracking via quadrangle kernelized correlation filters. IEEE Transactions on Intelligent Transportation Systems, 2017, 19(99): 140-150

[26] Yang L, Zhu J. A scale adaptive kernel correlation filter tracker with feature integration//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 254-265

[27] Fan H, Xiang J. Robust visual tracking via local-global correlation filter//Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017: 4025-4031

[28] Feng W, Han R, Guo Q, et al. Dynamic saliency-aware regularization for correlation filter-based object tracking. IEEE Transactions on Image Processing, 2019, 28(7): 3232-3245

[29] Li F, Tian C, Zuo W, et al. Learning spatial-temporal regularized correlation filters for visual tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4904-4913

[30] Sui Y, Wang G, et al. Correlation filter learning toward peak strength for visual tracking. IEEE Transactions on Cybernetics, 2017, 48(4): 1290-1303

[31] Wang N, Zhou W, Tian Q, et al. Multi-cue correlation filters for robust visual tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4844-4853

[32] Vondrick C, Shrivastava A, Fathi A, et al. Tracking emerges by colorizing videos//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 391-408

[33] Lai Z, Xie W. Self-supervised learning for video correspondence flow. arXiv preprint arXiv:1905.00875, 2019

[34] Li X, Liu S, Mello S D, et al. Joint-task self-supervised learning for temporal correspondence. arXiv preprint arXiv: 1909.11895, 2019

[35] Lai Z, Lu E, Xie W. MAST: A memory-augmented self-supervised tracker//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6478-6487

[36] Wang X, Jabri A, Efros A A. Learning correspondence from the cycle-consistency of time//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 2566-2576

[37] Kalal Z, Mikolajczyk K, Matas J. Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(7): 1409-1422

[38] Wang Q, Gao J, Xing J, et al. DCFNet: Discriminant correlation filters network for visual tracking. arXiv preprint arXiv:1704.04057, 2017

[39] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023

[40] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018: 3-19

[41] Park J, Woo S, Lee J Y, et al. BAM: Bottleneck attention module. arXiv preprint arXiv:1807.06514, 2018

[42] Li X, Wang W, Hu X, et al. Selective kernel networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 510-519

[43] Zhang H, Zu K, Lu J, et al. EPSANet: An efficient pyramid split attention block on convolutional neural network. arXiv preprint arXiv:2105.14447, 2021

[44] Wang X, Girshick R, Gupta A, et al. Non-local neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7794-7803

[45] Zhang Q, Yang Y. ResT: An efficient transformer for visual recognition. arXiv preprint arXiv:2105.13677, 2021

[46] Liu H, Liu F, Fan X, et al. Polarized self-attention: Towards high-quality pixel-wise regression. arXiv preprint arXiv:2107.00782, 2021

[47] Chen Y, Kalantidis Y, Li J, et al. A2-Nets: Double attention networks. arXiv preprint arXiv:1810.11579, 2018

[48] Ding X, Xia C, Zhang X, et al. Repmlp: Re-parameterizing convolutions into fully-connected layers for image recognition. arXiv preprint arXiv:2105.01883, 2021

[49] Hou Q, Jiang Z, Yuan L, et al. Vision permutator: A permutable mlp-like architecture for visual recognition. arXiv preprint arXiv:2106.12368, 2021

[50] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2014, 115(3): 1-42

[51] Wu Y, Lim J, Yang M H. Online object tracking: A benchmark//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA, 2013: 2411-2418

[52] Wu Y, Lim J, Yang M H. Object tracking benchmark. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9): 1834-1848

[53] Liang P, Blasch E, Ling H. Encoding color information for visual tracking: Algorithms and benchmark. IEEE Transactions on Image Processing, 2015, 24(12): 5630-5644

[54] Kristan M, Leonardis A, Matas J, et al. The sixth visual object tracking vot2018 challenge results//Proceedings of the European Conference on Computer Vision Workshops. Munich, Germany, 2018: 3-53

[55] Huang B, Xu T, Shen Z, et al. SiamATL: Online update of Siamese tracking network via attentional transfer learning. IEEE Transactions on Cybernetics, 2021, 99: 1-14

[56] Li X, Ma C, Wu B, et al. Target-aware deep tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 1369-1378

[57] Zhang Z, Peng H. Deeper and wider siamese networks for real-time visual tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4591-4600

[58] Bertinetto L, Valmadre J, Golodetz S, et al. Staple: Complementary learners for real-time tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1401-1409

[59] Huang Z, Fu C, Li Y, et al. Learning aberrance repressed correlation filters for real-time UAV tracking//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 2891-2900

[60] Ma H, Acton S T, Lin Z. SITUP: Scale invariant tracking using average peak-to-correlation energy. IEEE Transactions on Image Processing, 2020, 29: 3546-3557

[61] Danelljan M, Gustav H, et al. Discriminative scale space tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1561-1575

[62] Sio C H, Ma Y J, Shuai H H, et al. S2SiamFC: Self-supervised fully convolutional siamese network for visual tracking//Proceedings of the 28th ACM International Conference on Multimedia. Westminster, USA, 2020: 1948-1957

[63] Yuan W, Wang M Y, Chen Q. Self-supervised object tracking with cycle-consistent siamese networks. arXiv preprint arXiv:2008.00637, 2020

[64] Wu Q, Wan J, Chan A B. Progressive unsupervised learning for visual object tracking//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 2993-3002

[65] Zhu J, Ma C, Jia S, Xu S. Contrastive cycle consistency learning for unsupervised visual tracking//Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision. Beijing, China, 2021: 564-576

CHENG Xu, Ph. D., associate professor, master supervisor. His research interests include computer vision, pattern recognition.

LIU Li-Hua, master student. Her research interest includes object detection and tracking.

WANG Ying-Ying, master student. Her research interest includes adversarial attacks on intelligent systems.

YU Zi-Tong, Ph. D. candidate. His research interest includes computer vision and biometric security.

ZHAO Guo-Ying, Ph. D., professor, Ph. D. supervisor. Her research interests include computer vision, video image processing, and intelligent human-computer interaction.

Background

Visual object tracking is one of the fundamental tasks in computer vision, which has a wide range of applications including video surveillance, action recognition, scene understanding, intelligent transportation, visual navigation, and human-machine interaction, etc. It aims to estimate the object location in the following frames, given its initial state in the first frame. In recent years, the rapid development of the internet and intelligent device terminals has led to an exponential increase in video data. In order to effectively understand and analyze video big data, there is an urgent need to design robust visual trackers which can automatically locate target objects in video sequences.

The currently existing deep learning-based trackers were proposed to significantly improve tracking performance, but the limited labeled data limits the efficient training of deep network model. Although self-supervised learning-based trackers were attempted to handle this issue, they lack the efficient representation of object features, and ignores the difficulty of reverse verification caused by the challenges such as object occlusion, resulting in a decrease in tracking accuracy. Thus, the existing approaches cannot well address those significant challenges.

In this paper, we propose an effective multi-frame consistency correction based self-supervised Siamese network tracking method (MCCSST) for visual tracking. In our design, we present an effective forward multi-frame reverse order verification strategy that adaptively selects the optimal object path to construct the cyclic-consistency loss optimization function. To address the problem of inconsistent object location from the multi-path at the same frame, we introduce a simple yet effective mixed order correction module to correct the tracking drift. In addition, we further develop the visual feature enhancement module to enhance the object features representation ability by adaptively weighted fusion the global context information and local semantic feature information of the object. Our method is verified on four public datasets: OTB2013, OTB2015, TColor-128 and VOT-2018. The accuracy of the proposed method is improved by 4.6% on average over the compared twenty-one state-of-the-art trackers, which is an average of 5.8% higher than that of the self/unsupervised learning-based trackers.

This research is supported by the National Natural Science Foundation of China (Nos. 61802058 and 61911530397), the China Scholarship Council (CSC) (No. 201908320175), the China Postdoctoral Science Foundation (No. 2019M651650), and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund.

Before this work, our research group has already published more than 40 high-quality papers in the field of object tracking, object detection and activity recognition.