ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 5,Sep. 2023,pp. 1079－1091
DOI:10. 16337/j. 1004-9037. 2023. 05. 007

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

SiamBM:实现更佳匹配的Siamese目标跟踪网络

胡昭华[1,2]，刘浩男[1]，林　潇[1]

(1. 南京信息工程大学电子与信息工程学院,南京 210044;2. 南京信息工程大学江苏省大气环境与装备技术协同创新中心,南京 210044)

摘　要：基于孪生网络的目标跟踪算法通常采用简单的互相关匹配方式,然而这种简单的匹配方式会引入大量无关信息,弱化目标区域的响应。基于无锚框的孪生跟踪网络虽然避免了锚框参数的调整,但由于失去了先验性信息,并不能很好地适应目标物的尺度变化。因此,针对上述所存在的问题,本文提出了一种基于孪生网络的目标跟踪匹配增强算法SiamBM。通过将目标的边界框坐标信息进行编码,为跟踪模型提供有效的指导信息;采用深度可分离互相关级联像素匹配互相关的方式,进一步提高跟踪模型的判别能力;采用多尺度互相关的方式,增强跟踪模型的尺度适应能力。在OTB100数据集上,SiamBM的成功率和精确率分别达到了0.684和0.906,相比基准模型分别提高了5.2%和4.2%。实验结果表明,与目前主流的跟踪器相比,SiamBM取得了相当有竞争力的结果,在各项数据集指标上取得了优越的性能。

关键词：目标跟踪;孪生网络;多方式互相关;多尺度互相关;边界框编码

中图分类号：TP391　　**文献标志码**:A

SiamBM: Siamese Object Tracking Network for Better Matching

HU Zhaohua[1,2], LIU Haonan[1], LIN Xiao[1]

(1. College of Electronics and Information Engineering, Nanjing University of Information Technology, Nanjing 210044, China; 2. Jiangsu Collaborative Innovation Center for Atmospheric Environment and Equipment Technology, Nanjing University of Information Technology, Nanjing 210044, China)

Abstract: Object tracking algorithms based on Siamese networks usually adopt simple cross-correlation matching, but this simple matching method will introduce a lot of irrelevant information and weaken the response of the target region. Although the Siamese tracking network without anchor frame avoids the adjustment of anchor frame parameters, it cannot adapt well to the scale change of the target due to the loss of priori information. Therefore, aiming at the above problems, this paper proposes a object tracking matching enhancement algorithm SiamBM based on Siamese networks. By encoding the boundary frame coordinate information of the target, effective guidance information is provided for the tracking model. The discriminant ability of the tracking model is further improved by means of depth separable cross-correlation and cascade pixel matching cross-correlation. Multi-scale cross-correlation is adopted to enhance the scale adaptability of the tracking model. In the OTB100 dataset, the success rate and accuracy rate of SiamBM reached 0.684 and 0.906, respectively, which increased by 5.2% and 4.2% compared with the benchmark model. The experimental results show that compared with the current mainstream trackers, SiamBM has

收稿日期:2022-05-21;**修订日期**:2022-11-17

achieved quite competitive results and superior performance in various dataset indicators.

Key words: object tracking; Siamese network; multi-modal cross-correlation; multi-scale cross-correlation; bounding-box encoding

引　　言

目标跟踪是计算机视觉领域中一个基础而又具有挑战性的任务，是近几十年来计算机视觉领域最活跃的研究课题之一。目标跟踪的任务定义为：对于一个视频序列，在只给定目标初始帧位置的情况下，跟踪器能够在后续的每一帧中准确地跟踪此目标。目标跟踪在自动驾驶、视频监控、海洋勘探、医学影像等领域都有着广泛的应用，因此备受学术界和工业界的关注。

目前的研究表明，基于孪生网络的目标跟踪算法在跟踪精度和推理速度之间取得了很好的平衡。SINT[1]最早将相似性学习的方式应用到目标跟踪中，其本质是抽取多个候选目标依次送入网络中进行相似度的对比。SiamFC[2]在孪生网络中引入了互相关的结构，解决了SINT速度过慢的问题，真正地实现了速度与精度的平衡。SiamRPN[3]在SiamFC的基础上引入了多通道互相关和目标检测中的RPN区域生成网络[4]，使得回归预测更加精准。SiamRPN++[5]采用了深度可分离互相关，减少了大量参数，稳定了整个训练过程。然而，无论哪一种现有的互相关方式，其本质依然都是两特征图之间固定大小的滑窗卷积操作，因此当物体发生较大形变或者目标区域相对较小时，互相关会引入大量的背景信息，从而影响跟踪精度。

SiamBAN[6]通过直接预测特征图上前景背景的分类得分和4个中心距离偏移量来得到最大响应位置的预测框。这种像素级的无锚框预测方式[7]解决了SiamRPN系列网络存在的问题，减少了参数调整的负担。但由于失去了类似锚框的先验信息以及单一尺度卷积核的采用，跟踪器并不能很好地具备学习和应对目标尺度变化的能力。另外，由于跟踪目标形状的不确定性，固定比例的锚框对跟踪网络并不具备很好的指导性。

因此，根据现有算法研究的不足，本文提出了一种基于孪生网络的互相关匹配增强算法。主要工作如下：

(1)为了能够使得跟踪网络充分利用到有效的先验信息，通过将已给定的目标边界真值框信息编码到网络中，增强了目标区域的前景响应，进一步提升了跟踪精度。

(2)通过分析研究目前互相关结构的特点，采用了一种新的多互相关级联方式，解决了目前互相关匹配的固有问题，减少了无关的背景和干扰信息，提高了跟踪网络的判别能力。

(3)通过引入非常规尺度的卷积核对模板特征和搜索特征进行多方位、多尺度的特征提取并融合，既避免了一系列的锚框参数调整，又能使得跟踪器得到更多的尺度信息。

最终，在能够保证高实时性跟踪的前提下，SiamBM取得了良好的跟踪性能。

1　SiamBM网络框架

1.1　网络整体结构

如图1所示，SiamBM整体的网络结构采用基于无锚框的孪生跟踪网络，主要分为特征提取网络、互相关匹配网络和分类回归网络这3大部分。特征提取网络采用修改后的ResNet50[8]深层神经网络来进行特征提取，ResNet最后两个模块的步长被设置为1，并且扩张卷积的大小被设置为4，增加了感受野的大小。互相关匹配网络由互相关结构和边界框编码模块组成，模板特征和搜索特征经过互相关匹配后与边界框编码后的信息进行融合，作为后续分类回归网络的输入。分类回归网络包含分类和回归

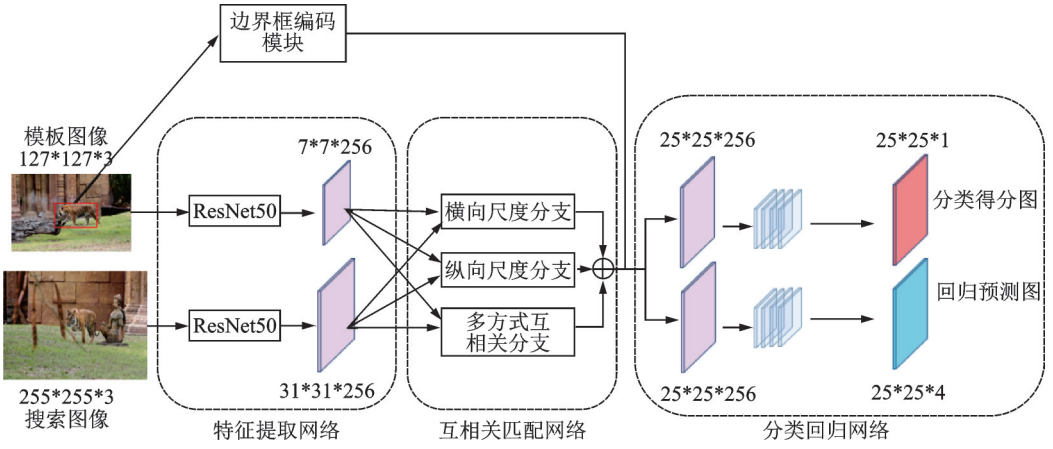


图1 SiamBM网络结构

Fig.1 SiamBM network structure

两个分支,分类分支负责预测目标为前景的得分。回归分支负责预测(l,t,b,r)4个距离,分别代表目标中心位置距回归框4条边的偏移距离。

1.2 边界框编码模块

主流的孪生跟踪网络对于模板帧已给定的边界框信息通常会有两种利用方式,一种是对输入图像进行常规的中心裁剪,另一种是利用边界框的坐标做相关的感兴趣区域(Region of interest,ROI)映射或者提取图像像素掩膜。这两种方式通过利用已有的边界框坐标信息将图像层面的信息进行提取和操作,但是往往忽略了边界框坐标本身这种非结构化的数据信息。因此,本文算法通过将边界框编码后的信息与得到的互相关特征进行融合操作,使得跟踪器的性能得到进一步的提升。

如图2所示,编码模块首先将模板帧的边界框坐标转化成一维的特征向量:$B(x,y,w,h)$,(x,y)代表目标边界框的角点坐标,w代表目标边界框的宽度,h代表目标边界框的高度。特征向量B经过多层全连接层得到

$$B_C=f_C(B) \tag{1}$$

式中:f_C代表全连接层结构,B_C代表特征向量B经全连接层的输出特征。然后互相关网络的输出特征F与B_C进行广播相加操作

$$F_b=F+B_C \tag{2}$$

式中:$F_b\in\mathrm{R}^{(C\times H\times W)}$,$F\in\mathrm{R}^{(C\times H\times W)}$,$B_C\in\mathrm{R}^{(C\times 1\times 1)}$,$C$代表特征图的通道数,$H$和$W$分别代表特征图的高度和宽度。最后$F_b$经过1*1卷积编码得到最终的输出结果

$$F_{BM}=f_g(F_b) \tag{3}$$

式中:f_g代表1*1大小的卷积编码操作,$F_{BM}\in\mathrm{R}^{(C\times H\times W)}$代表最终的边界框编码输出结果。

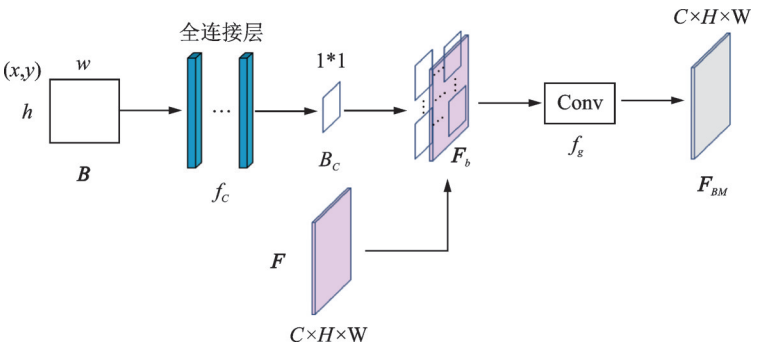


图2 边界框编码模块

Fig.2 Bounding box encoding module

1.3 多方式互相关

作为孪生跟踪网络中最为关键的组成部分，互相关结构的设计对于跟踪器的性能是至关重要的。然而目前已有的互相关方式依然都是两特征图之间进行固定大小的滑窗卷积操作，这种互相关方式是一种针对目标区域的全局信息的匹配，因此当物体发生较大形变或者目标区域相对较小时，互相关会引入大量无关的背景信息，从而干扰对目标的跟踪。

因此，本文对孪生跟踪网络的互相关匹配方式进行了改进，引入了多方式互相关分支。多方式互相关分支通过级联深度可分离互相关和像素匹配互相关两种方式，能够使得网络更好地关注到目标区域，减小干扰物的影响。多方式互相关分支的结构如图3所示，模板分支特征f_z与搜索分支特征f_x首先进行像素匹配互相关操作

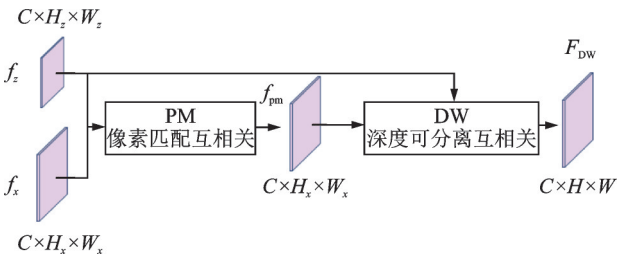


图3 多方式互相关分支结构

Fig.3 Multi-mode cross-correlation branch structure

$$f_{pm}=\mathrm{PM}\left(f_z,f_x\right) \tag{4}$$

式中：PM代表像素匹配互相关，$f_z\in\mathbf{R}^{(C\times H_z\times W_z)}$，$f_x\in\mathbf{R}^{(C\times H_x\times W_x)}$，$C$代表特征图的通道数，$H_z$和$W_z$分别代表模板特征图的高度和宽度，$H_x$和$W_x$分别代表搜索特征图的高度和宽度，$f_{pm}\in\mathbf{R}^{(C\times H_x\times W_x)}$代表输出的像素匹配互相关特征。然后，像素匹配互相关特征f_{pm}与模板分支特征f_z进行深度可分离互相关，最终得到融合两种互相关的特征为

$$F_{DW}=\mathrm{DW}\left(f_{pm},f_z\right) \tag{5}$$

式中：DW代表深度可分离互相关，$F_{DW}\in\mathbf{R}^{(C\times H\times W)}$代表多方式互相关分支的输出特征。像素匹配互相关的方式经常被应用于实时分割领域，其本质是两特征图之间进行像素级别的匹配。

图4详细展示了像素匹配互相关的流程。设定模板分支的特征图为$f_z\in\mathbf{R}^{(C\times H_z\times W_z)}$，搜索分支的特征图为$f_x\in\mathbf{R}^{(C\times H_x\times W_x)}$。首先，$f_z$与$f_x$分别进行经过1*1的卷积层编码得到$f_{z1}$和$f_{x1}$。之后，$f_{z1}$进行维度变换操作分别得到$f_{z11}$和$f_{z12}$，$f_{x1}$进行维度变换操作得到$f_{x2}$，表达式为

$$\begin{cases} f_{z11}=\mathrm{Reshape}\left(\mathrm{Conv}\left(f_z\right)\right) \\ f_{z12}=\mathrm{Reshape}\left(\mathrm{Conv}\left(f_z\right)\right) \\ f_{x2}=\mathrm{Reshape}\left(\mathrm{Conv}\left(f_{x1}\right)\right) \end{cases} \tag{6}$$

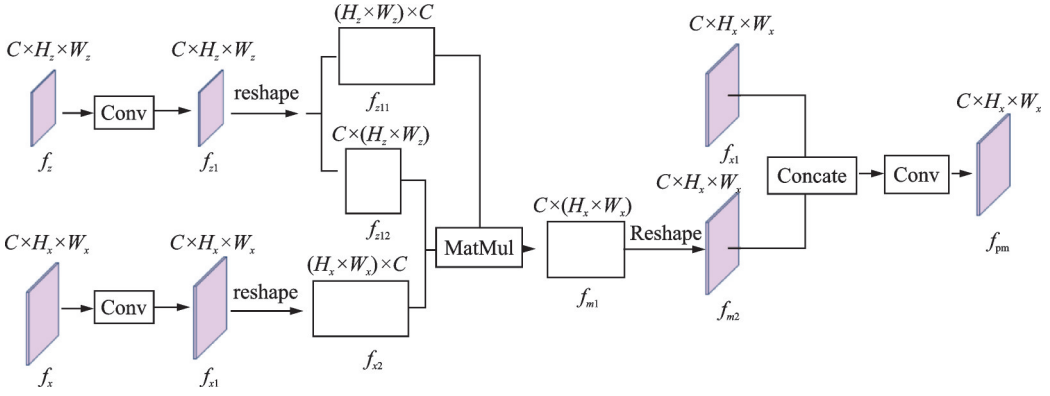


图4 像素匹配互相关结构

Fig.4 Pixel matching cross-correlation structure

式中:$f_{z11}\in\mathbf{R}^{(H_z\times W_z)\times C}$,$f_{z12}\in\mathbf{R}^{C\times(H_z\times W_z)}$,$f_{x2}\in\mathbf{R}^{(H_x\times W_x)\times C}$,Conv代表1*1*$C$维度大小的卷积层,Reshape代表维度变换操作。随后,f_{z12}与f_{x2}进行矩阵相乘操作,将得到的结果再与f_{z11}进行矩阵相乘输出得到f_{m1},并对f_{m1}进行维度变换操作得到f_{m2},表达式为

$$\begin{cases} f_{m1}=\mathrm{MatMul}\left(\mathrm{MatMul}\left(f_{x2},f_{z12}\right),f_{z11}\right) \\ f_{m2}=\mathrm{Reshape}\left(f_{m1}\right) \end{cases} \tag{7}$$

式中:$f_{m1}\in\mathbf{R}^{C\times(H_x\times W_x)}$,$f_{m2}\in\mathbf{R}^{(C\times H_x\times W_x)}$,MatMul代表矩阵相乘操作。在后续阶段,$f_{m2}$与之前编码后的特征$f_{x1}$沿通道进行拼接操作,最终经1*1卷积进行降维后,得到像素匹配互相关的输出f_{pm}

$$f_{\mathrm{pm}}=\mathrm{Conv}\left(\mathrm{Concate}\left(f_{x1},f_{m2}\right)\right) \tag{8}$$

式中:$f_{\mathrm{pm}}\in\mathbf{R}^{(C\times H_x\times W_x)}$,Concate代表沿通道进行拼接操作。

像素匹配的互相关方式通过逐对进行像素匹配,减少了背景信息的干扰,增强了前景信息的提取。并且本文同时级联了两种不同的互相关方式,这是因为像素匹配互相关是一种全局像素匹配的互相关方式,而深度可分离互相关是一种局部匹配的互相关方式,两种方式进行结合后能够在捕捉到全局上下文信息的基础上,再度细化目标局部区域信息的提取,相较于单一的互相关方式,这样能够使得网络具备更好的判别性。

1.4 多尺度互相关

虽然基于无锚框的孪生跟踪网络失去了锚框信息的指导,但对于跟踪器来说,如何去学习并能够适应目标物的尺度变化是更为重要的,而不是局限于使用固定比例的锚框来进行预测框的回归。

在绝大多数场景下,用来滑动提取特征的卷积核一般设置为1*1、3*3、5*5等这种宽高相等的常规卷积核,但这种常规的卷积核在不同场景下并不一定总是最优的。如图5(a)所示,两张图中分别画出了3种不同尺度的卷积核,其中采用虚线框标识的卷积核的选择性要优于其他两种卷积核。可以看到,在不同的场景下,卷积核的尺度设计对于特征的提取也会产生影响。比如在图5(a)的左图中,横向尺度的卷积核能够更好地对此目标物的形状进行建模。相对应地,在图5(a)的右图中,纵向尺度的卷积核能够更好地对目标物的形状进行建模。

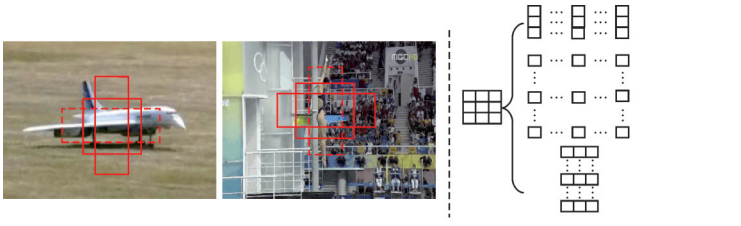


图5 两种场景下的卷积核以及不同方向上的卷积变换

Fig.5 Convolution kernels in two scenarios and convolution transformations in different directions

因此,如图1所示,本文算法在多方式互相关分支的基础之上额外添加了两个多尺度互相关分支,横向尺度分支更加关注于横向区域的特征提取,而纵向尺度分支更加关注于纵向区域的特征提取。图6展示了多尺度互相关分支的结构,可以看到,模板特征和搜索特征经过横向尺度分支分别得到各自的横向尺度特征,两横向特征再进行互相关操作,输出得到横向

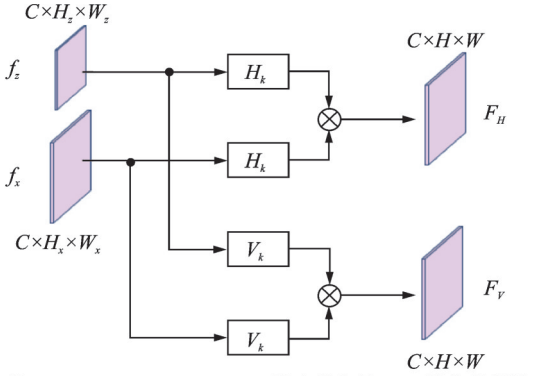


图6 多尺度互相关分支结构

Fig.6 Multi-scale cross-correlation branch structure

尺度分支特征。同样,纵向尺度分支经类似操作,输出得到纵向尺度分支特征

$$F_{\mathrm{H}}=\mathrm{DW}(f_z*H_k,f_x*H_k) \tag{9}$$
$$F_{\mathrm{V}}=\mathrm{DW}(f_z*V_k,f_x*V_k)$$

式中:$f_z\in\mathbf{R}^{(C\times H_z\times W_z)}$、$f_x\in\mathbf{R}^{(C\times H_x\times W_x)}$分别代表互相关匹配网络输入的模板特征与搜索特征,DW代表深度可分离互相关,*代表卷积操作,H_k横向卷积核经3:1的卷积扩张为7*3大小的卷积核,V_k纵向卷积核经1:3的卷积扩张为3*7大小的卷积核。

最终,由图1可见,整个互相关结构的输出可以表示为

$$F=\alpha_1F_H+\alpha_2F_K+\alpha_3F_{\mathrm{DW}} \tag{10}$$

式中:$F_H\in\mathbf{R}^{(C\times H\times W)}$代表横向尺度分支的输出特征,$F_V\in\mathbf{R}^{(C\times H\times W)}$代表纵向尺度分支的输出特征,$F_{\mathrm{DW}}\in\mathbf{R}^{(C\times H\times W)}$代表多方式互相关分支的输出特征,$\alpha_1$、$\alpha_2$、$\alpha_3$分别代表3个特征权重融合系数,随着网络的训练而不断被优化。

1.5 损失函数

对于分类分支而言,分类得分图最终预测目标为前景的概率。对于回归分支而言,回归预测图最终预测4个中心偏移距离。因此联合任务损失函数L为

$$L=\lambda_1L_{\mathrm{cls}}+\lambda_2L_{\mathrm{reg}} \tag{11}$$

式中:L_{cls}代表二值交叉熵损失,L_{reg}代表IOU损失。在实验过程中,设定$\lambda_1=1$,$\lambda_2=1$。

2 网络训练与测试

2.1 网络训练

整个网络在多种大规模数据集上进行端到端的离线训练。网络采用的训练数据集包括ImageNet VID[9]、YouTube-BoundingBoxes[10]、GOT-10k[11]、ImageNet DET[7]、COCO[12]5个大型数据集。输入图像包含模板图像和搜索图像,两图像分别来自同一视频序列的不同图像帧。主干网络的模型参数初始化为在ImageNet[11]上的预训练参数。在训练阶段,预处理后的模板图像和搜索图像作为网络的输入共享同一网络以及网络参数,模板图像和搜索图像同时经过主干特征提取网络、互相关匹配网络和分类回归网络得到最终输出的分类得分图和回归预测图,之后结合联合任务损失函数对整个网络进行端到端的训练优化。

2.2 网络测试

在测试阶段,首先将测试视频序列的第一帧图像进行预处理操作,并作为模板图像送入特征提取网络得到模板特征。此后,模板特征将在网络中被固定,避免后续重复地特征提取,从而加快网络的跟踪速度。同时,将模板帧已给定的边界框信息进行编码,以便后续与互相关特征进行融合。此后,将测试视频序列的后续每一帧以4倍于模板图像的区域进行裁剪操作,裁剪区域的中心为上一帧预测的目标中心点,并将其作为搜索图像送入特征提取网络中,与已经固定的模板特征进行一系列操作,得到分类得分图和回归预测图。最后通过一系列后处理操作,在分类得分最大的位置,对应到回归预测图进行4个偏移量回归,得到此帧中目标的最终预测边界框。

在整个测试的过程中,本文所提出算法的跟踪速度能够稳定保持在58 f/s左右,相较于目前大多数主流的跟踪器,SiamBM在实时性和精度这两个方面的平衡上实现得更加合理。

3 实验结果与分析

3.1 实验环境与参数设置

整个实验在Ubantu 18.04操作系统上进行,编程采用以Python实现的Pytorch框架,硬件配置为AMD Ryzen 7 4800H 2.90 GHz CPU,16 GB内存,Nvidia RTX 2060显卡。实验时,在单个GPU上设置

每次迭代的batch大小为16,并使用带有动量的SGD随机梯度下降法进行梯度回传优化。整个训练的总轮数为50轮,在训练时,前5轮采用0.001逐渐到0.005的学习率进行预热训练,后续的45轮采用0.005逐渐减小到0.000 01的学习率来进行训练,并且在第10轮训练后,主干网络的参数将不再被冻结,跟随整个网络进行端到端的优化训练。权重衰减系数和动量参数分别设定为0.000 1和0.9。

3.2 数据集与评估指标

本文所提出的算法在OTB100[13]、GOT-10k[11]、VOT2019[14]、LASOT[15]4个跟踪基准数据集上进行跟踪器的性能评估。

OTB100是目标跟踪领域中最受广泛使用的基准数据集之一,由100个完全注释过的视频序列组成,平均每个视频序列590帧。OTB数据集中的视频序列包含着各种各样的挑战,比如背景杂乱、遮挡、快速运动、变形等。OTB通过一次评估以精度和成功图的曲线下面积两个指标来评估跟踪器。精度图显示了预测位置与真值框之间的距离在20像素阈值之内所占帧的百分比。成功图显示平均重叠率大于给定阈值的帧的百分比。

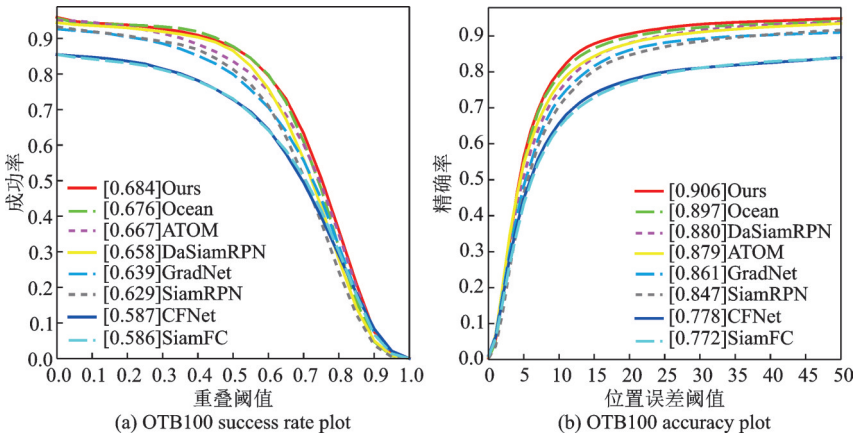
GOT-10k是一个极具挑战性的大规模跟踪数据集,包含超过10 000个视频序列以及180个测试视频序列,其训练集和测试集严格进行类别的区分,避免了对跟踪器的评估结果会对特定类别的视频序列产生偏置。GOT-10k所提供的评价指标包括平均重叠率AO和成功率SR。平均重叠率AO表示所有预测的边界框与真值框之间的平均重叠率。成功率SR具体又分为0.5和0.75两个阈值,SR0.5代表成功跟踪到的帧与真值框重叠率超过0.5的比例,SR0.75代表成功跟踪到的帧与真值框重叠率超过0.75的比例。

VOT2019包含60个具有不同挑战性的视频序列。相较于VOT2018,VOT2019更换了20%的视频序列,并且包含了更多具有挑战性的视频序列。VOT2019提供了3个评估指标:平均重叠期望EAO、准确率A和鲁棒性R。准确率A代表跟踪成功帧的比例,鲁棒性R代表跟踪帧失败的比例,平均重叠期望EAO通过A和R综合计算求得。

LASOT是一个高质量的大规模长期跟踪数据集,包含总共1 400个视频序列,共有70个类别。LASOT测试集包含280个视频序列,平均每个视频序列包含2 500帧左右,其中大量的视频序列都会出现目标物短暂消失的情况。LASOT提供精确率和成功率两个指标,精确率衡量预测框与真值框之间的像素距离,成功率衡量预测框与真值框之间的IOU。

3.3 实验分析与跟踪器对比

图7显示了SiamBM与目前主流的一些跟踪器在OTB基准数据集上的性能对比,对比的算法包括



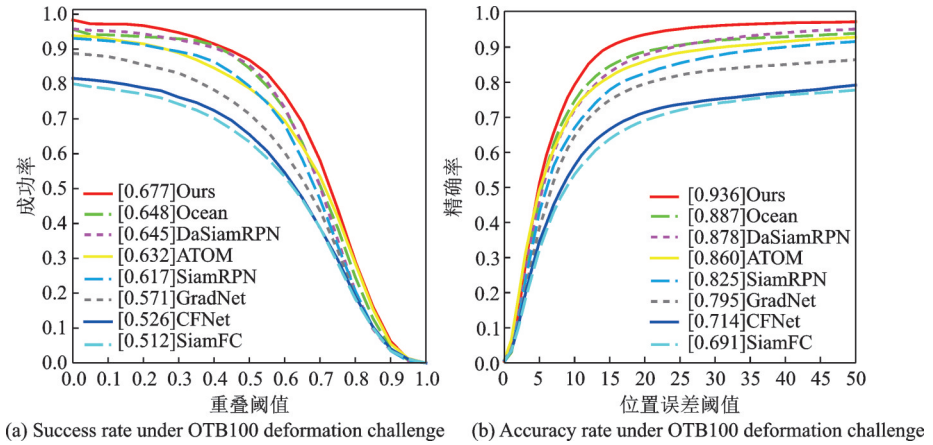
(a) OTB100 success rate plot (b) OTB100 accuracy plot

图7 主流跟踪器在OTB100上的性能评估结果

Fig.7 Performance evaluation results of mainstream trackers on OTB100

Ocean[16]、ATOM[17]、SiamRPN[3]、DaSiamRPN[18]、GradNet[19]、CFNet[20]、SiamFC[2]。由图可见，本文算法在OTB基准上的成功率达到了0.684，精确率达到了0.906，相较于Ocean和ATOM分别提高了0.8%和1.7%。特别地，图8展示了在视频序列的形变挑战下的多种跟踪算法的性能比较。可以看到，在目标形变时，SiamBM展现了强大的适应能力，证实了本文算法解决尺度变化问题的有效性。图9、10还展示了不同跟踪算法在OTB100遮挡、视野消失、运动模糊、背景杂乱、尺度变化挑战下的性能对比，SiamBM在各种挑战下都展现了良好的适应性，取到了良好的跟踪性能。

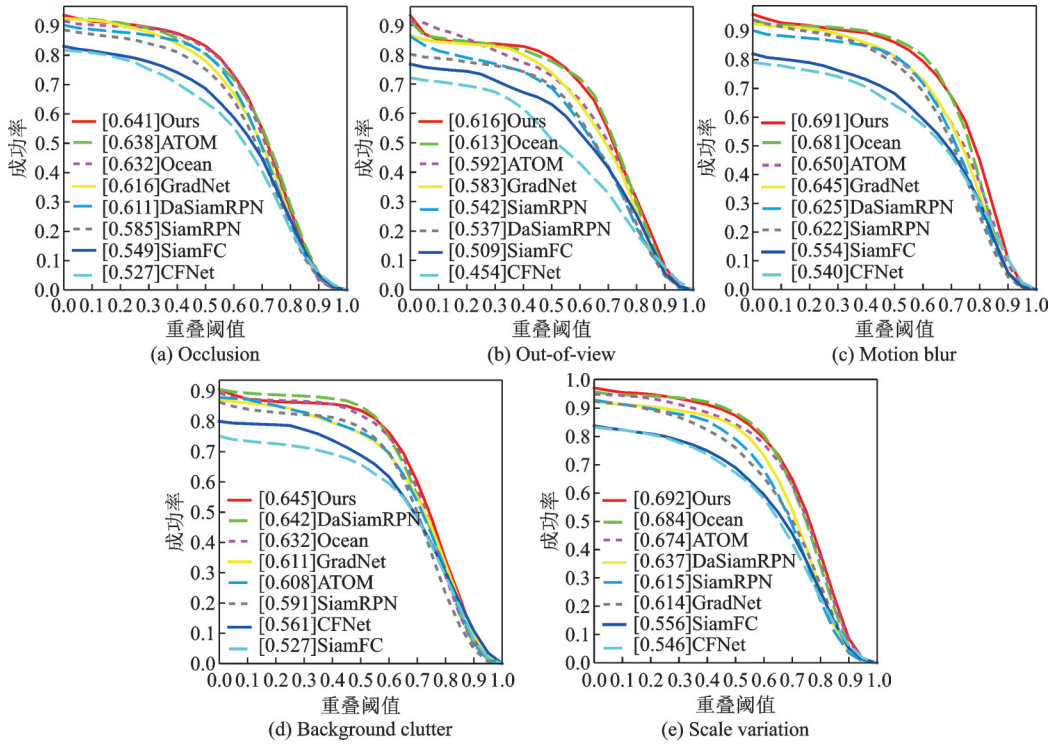
表1展示了SiamBM与主流跟踪器在GOT-10k大规模基准数据集上的性能对比，对比的算法包括



(a) Success rate under OTB100 deformation challenge (b) Accuracy rate under OTB100 deformation challenge

图8 形变挑战下，各种跟踪器的性能对比

Fig.8 Performance comparison of various trackers under the deformation challenge



(a) Occlusion (b) Out-of-view (c) Motion blur

(d) Background clutter (e) Scale variation

图9 各种跟踪算法在OTB100不同挑战下的AUC对比

Fig.9 AUC comparison of various tracking algorithms under different OTB100 challenges

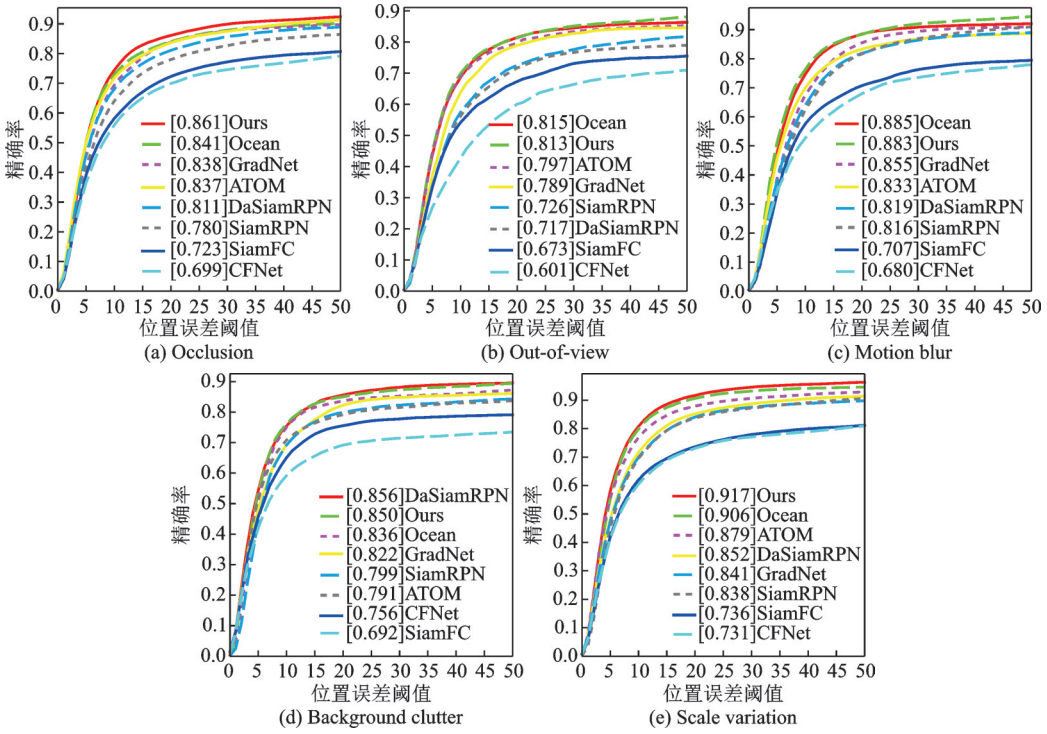


图10 各种跟踪算法在OTB100不同挑战下的精度对比

Fig.10 Accuracy comparison of various tracking algorithms under different OTB100 challenges

表1 GOT-10k上各项跟踪器的性能评估结果

Table 1 Performance evaluation results of various trackers on GOT-10k

跟踪器	平均重叠率	成功率(0.5阈值)	成功率(0.75阈值)
CFNet	0.293	0.265	0.087
MDNet	0.299	0.303	0.099
SiamFC	0.348	0.353	0.098
SiamRPN++	0.517	0.615	0.329
ATOM	0.556	0.634	0.402
SiamFC++	0.595	0.695	0.479
SiamCAR	0.579	0.677	0.437
Ocean	0.592	0.695	0.473
Dimp	**0.611**	**0.717**	**0.492**
Ours	0.604	0.698	0.480

注:加粗字体代表当前性能指标下的最好结果。

DIMP[21]、Ocean[16]、ATOM[17]、SiamRPN++[5]、SiamCAR[22]、SiamFC++[23]、SiamDW[24]、SiamFC[2]、MDNet[25]。具体地,除DIMP之外,SiamBM在平均重叠率AO、成功率SR0.5、成功率SR0.75这3项指标上相较于其他跟踪器均取得了领先。可能的原因是因为DIMP采用的在线更新机制能够及时地跟踪和适应未出现过的类别,从而做出在线调整更新,使得网络具备更好的在线适应能力。整体上可以看出,SiamBM在没有使用在线更新机制的情况下,相较于主流的跟踪器依然取得了强有竞争力的

结果，展现了SiamBM良好的泛化能力。

表2展示了在VOT2019数据集基准上，SiamBM与多种跟踪器在准确率A、鲁棒性R、平均重叠期望EAO三个指标上的性能对比结果。对比的算法包括CSRDCF[14]、MemDTC[7]、SiamCRF_RT[12]、SPM、ROAM++[26]、SiamRPN++[5]、SiamMask[27]、ATOM[17]。其中，准确率A和平均重叠期望EAO越高则表示性能越好，鲁棒性R越低则表示性能越好。在表中可以看到，即使在未采用在线更新机制的情况下，相较于在线跟踪器ATOM，SiamBM在鲁棒性指标上依然取得了有竞争力的结果，充分体现了边界框编码模块和多方式互相关模块的有效性，能够在跟踪过程中抑制干扰物，从而提升模型的辨别性和鲁棒性。在准确率指标A上，SiamBM达到了0.604，在对比中取得了最佳的结果。在平均重叠期望EAO上，SiamBM在性能对比中取得了第二名。得益于互相关匹配的增强，SiamBM在保持良好鲁棒性的同时，在精度上也能够取到良好的性能。

表2 VOT2019上各跟踪器的性能对比

Table 2 Performance comparison of each tracker on VOT2019

跟踪器	准确率↑	鲁棒性↓	平均重叠率
CSRDCF	0.496	0.632	0.201
MemDTC	0.485	0.587	0.228
SiamCRF_RT	0.549	**0.346**	0.262
SPM	0.577	0.507	0.275
ROAM++	0.561	0.438	0.281
SiamRPN++	0.599	0.482	0.285
SiamMask	0.594	0.461	0.287
ATOM	0.603	0.411	**0.292**
Ours	**0.604**	0.416	0.292

注：加粗字体代表当前性能指标下的最好结果。

为了评估跟踪器在长期视频跟踪过程中的鲁棒性，本文算法在大规模长期跟踪数据集LASOT上进行了性能评估。图11展示了SiamBM在LASOT上与主流跟踪器的性能对比，对比的算法包括SiamBAN[6]、SiamCAR[22]、Ocean[16]、SiamRPN++[5]、ATOM[17]、MDNet[25]、VITAL[28]、SiamFC[2]。由图可见，SiamBM在成功率和精确率两个指标上都优于目前主流的跟踪器，在成功率上分别优于Ocean、SiamBAN的0.8%和1%。可以看到，SiamBM在不采用在线更新的机制下，依然能够在LASOT长期跟踪的场景下取得良好的性能，从而验证了跟踪器具备良好的鲁棒性和适应性。

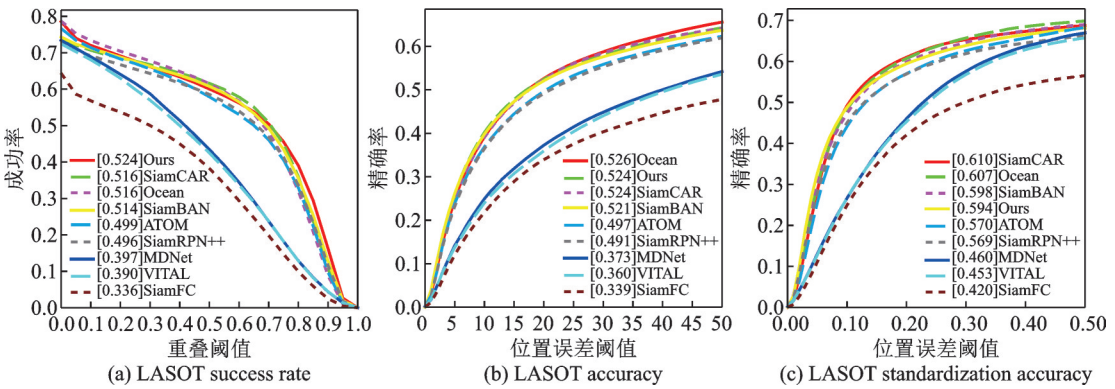


图11 LASOT基准下各项跟踪器的成功率、精度和标准化精度

Fig.11 Success rate, accuracy and standardization accuracy of various trackers under LASOT benchmark

3.4 消融实验

为了验证边界框编码、多尺度互相关、多方式互相关这3个组件结构对于跟踪性能提升的有效性，本文算法在OTB100数据集基准上以成功率为指标进行了相关的消融实验分析，如表3所示。由表3可见，在成功率指标上，3个组件结构分别提升了0.4%、2.2%、0.6%，表明了本文算法的有效性。为了更加具体地验证多方式互相关模块的有效性，如表4所示，本文对两种互相关方式的选择进行了额外的

实验。可以看到,通过合理利用两种互相关方式的特点并进行结合,多方式互相关模块进一步提升了网络的性能。在表5中,本文算法还分析对比了不同尺度的卷积核对于跟踪性能的影响。在图5(b)中可以看到,通过将卷积核的宽高进行不同程度的扩张,能够使得卷积核的形状在横向和纵向上进行拉伸,从而形成横向卷积核和纵向卷积核。同时,考虑到模板特征的实际大小,本文进行了3种不同尺度的扩张实验。由表5可见,实验采用了3组不同尺度的卷积核来进行对比分析,以3*3的卷积核为准进行了3:2、2:1、3:1的横向扩张与2:3、1:2、1:3的纵向扩张,分别得到了3组不同尺度的卷积核。在实验的过程中发现,虽然不同尺度卷积核的组合可能会对跟踪器的性能提升有着微小的差异,但是其都能够很好地提升跟踪器的尺度感知能力,从而带来性能上的提升。

表3 本文算法在OTB100上的消融实验分析

Table 3 Analysis of the ablation experiments of the algorithm in this paper on OTB100

边界框编码	多尺度互相关	多方式互相关	OTB100成功率	跟踪速度/(f•s^{-1})
			0.650	64.7
√			0.654	64.0
	√		0.672	59.8
		√	0.656	62.3
√	√		0.676	59.0
	√	√	0.680	58.7
√	√	√	0.684	58.2

注:"√"代表采用当前模块。

表4 多方式互相关的消融实验分析

Table 4 Analysis of multi-modal cross-correlation ablation experiments

深度可分离互相关(DW)	像素匹配互相关(PM)	OTB100成功率
√		0.650
	√	0.648
√	√	**0.656**

注:"√"代表采用当前模块。

表5 不同尺度的卷积核对性能的影响

Table 5 Effects of convolution kernels of different scales on performance

横向扩张比例/纵向扩张比例	OTB100成功率
3:2 / 2:3	0.683
2:1 / 1:2	0.681
3:1 / 1:3	**0.684**

注:加粗字体代表当前性能指标下的最好结果。

3.5 定性分析

本文算法在OTB100数据集上与3种主流的跟踪器进行了视频序列上的可视化对比与分析。图12展示了在4个具有不同挑战的视频序列下,4种跟踪算法的跟踪可视化结果。在Diving视频序列中,目标在前后视频帧中的姿态发生了较大的变化,ATOM[17]、DaSiamRPN[18]、SiamRPN[3]在目标形变后的预测都产生了较大的偏移,不能很好地将目标框回归到目标物上。而由于多尺度互相关的引入,SiamBM的回归预测更加精准,体现了跟踪器良好的尺度适应能力;在Board视频序列中,在目标物的背景较为杂乱的场景下,其他3种跟踪算法在跟踪过程中都产生了错误的预测,体现了SiamBM良好的判别性;在Girl2视频序列中,在目标物被遮挡之后,SiamBM依然能够正确地保持对目标物的跟踪,体现了跟踪器良好的鲁棒性。在Jump视频序列中,在目标物快速运动的情况下,SiamBM对比其他3种跟踪算法产生了更好的预测结果。

4 结束语

本文在基于无锚框跟踪的孪生网络的基础上,提出了一种匹配增强的目标跟踪算法SiamBM。通过在孪生网络中引入多方式互相关分支,采用两种互相关级联的方式,使其能够对全局上下文特征进行再度细化,可以有效地对目标区域进行监督,增强跟踪器的判别性。同时引入两个多尺度分支,分别在横

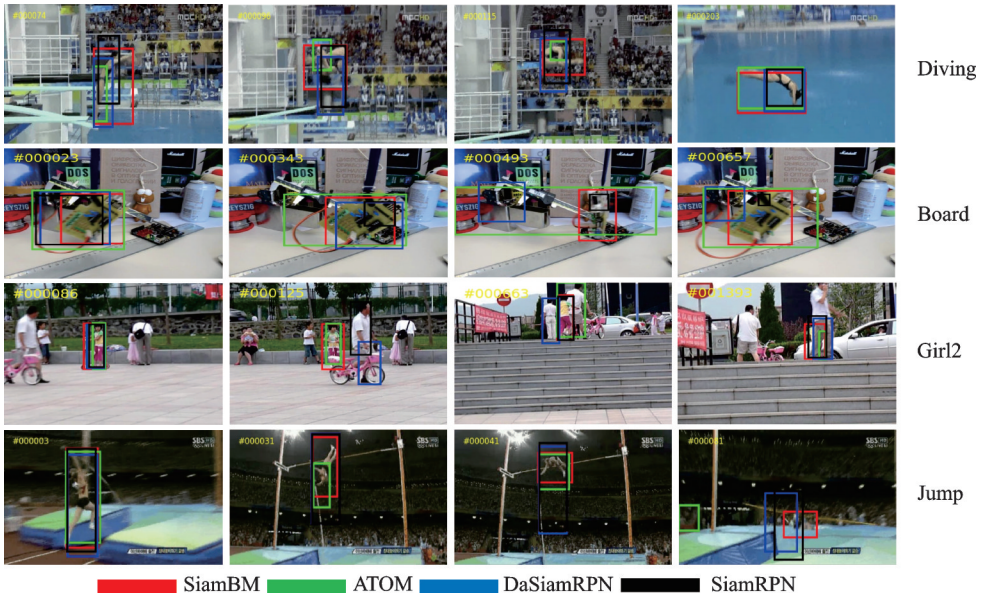


图 12 4种跟踪算法的跟踪结果可视化对比

Fig.12 Visual comparison of tracking results of four tracking algorithms

向、纵向两个方向上进行特征提取,从而能够适应各种目标物的各种形变,提高跟踪器的尺度感知能力。通过将模板帧的边界框信息编码到网络中,能够更好地利用到有效的先验信息,提高跟踪器的性能。通过一系列实验分析可以看到,SiamBM跟踪器在多种流行数据集基准上取到了良好的性能。在未来的工作中,将会试着解决跟踪器出现跟踪丢失的问题,能够在跟踪框漂移后重新关注到目标物本身,使得跟踪器在各种未知的干扰下保持长期稳定的跟踪,这也是整个目标跟踪领域的一个重点和难点。

参考文献:

[1] TAO R, GAVVES E, SMEULDERS A W M. Siamese instance search for tracking[C]//Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition. [S.l.]: IEEE, 2016: 1420-1429.

[2] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2016: 850-865.

[3] LI B, YAN J, WU W, et al. High performance visual tracking with Siamese region proposal network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 8971-8980.

[4] REN S, HE K, GIRSHICK R, et al. Faster r-CNN: Towards real-time object detectionwithregion proposal networks[C]//Proceedings of Advances in Neural Information Processing Systems, [S.l.]: IEEE, 2015.

[5] LI B, WU W, WANG Q, et al. Siamrpn++: Evolution of Siamese visual tracking with very deep networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 4282-4291.

[6] CHEN Z, ZHONG B, LI G, et al. Siamese box adaptive network for visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 6668-6677.

[7] YANG T, CHAN A B. Learning dynamic memory networks for object tracking[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: [s.n.], 2018: 152-167.

[8] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.

[9] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[10] REAL E, SHLENS J, MAZZOCCHI S, et al. YouTube-BoundingBoxes: A large high-precision human-annotated data set for

object detection in video[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 5296-5305.

[11] HUANG L, ZHAO X, HUANG K. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(5): 1562-1577.

[12] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Proceedings of European Conference on Computer Vision. Cham:Springer, 2014: 740-755.

[13] WU Y, LIM J, YANG M H. Online object tracking: A benchmark[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2013: 2411-2418.

[14] KRISTAN M, MATAS J, LEONARDIS A, et al. The seventh visual object tracking vot2019 challenge results[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. [S.l.]: IEEE, 2019.

[15] FAN H, LIN L, YANG F, et al. LASOT: A high-quality benchmark for large-scale single object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 5374-5383.

[16] ZHANG Z, PENG H, FU J, et al. Ocean: Object-aware anchor-free tracking[C]//Proceedings of European Conference on Computer Vision. Cham: Springer, 2020: 771-787.

[17] DANELLJAN M, BHAT G, KHAN F S, et al. ATOM: Accurate tracking by overlap maximization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 4660-4669.

[18] ZHU Z, WANG Q, LI B, et al. Distractor-aware Siamese networks for visual object tracking[C]//Proceedings of the European Conference on Computer Vision (ECCV). [S.l.]: [s.n.], 2018: 101-117.

[19] LI P, CHEN B, OUYANG W, et al. GradNet: Gradient-guided network for visual object tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 6162-6171.

[20] VALMADRE J, BERTINETTO L, HENRIQUES J, et al. End-to-end representation learning for correlation filter based tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2017: 2805-2813.

[21] BHAT G, DANELLJAN M, GOOL L V, et al. Learning discriminative model prediction for tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2019: 6182-6191.

[22] GUO D, WANG J, CUI Y, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 6269-6277.

[23] XU Y, WANG Z, LI Z, et al. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12549-12556.

[24] ZHANG Z, PENG H. Deeper and wider Siamese networks for real-time visual tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 4591-4600.

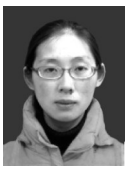
[25] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 4293-4302.

[26] YANG T, XU P, HU R, et al. ROAM: Recurrently optimizing tracking model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 6718-6727.

[27] WANG Q, ZHANG L, BERTINETTO L, et al. Fast online object tracking and segmentation: A unifying approach[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2019: 1328-1338.

[28] SONG Y, MA C, WU X, et al. VITAL: Visual tracking via adversarial learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2018: 8990-8999.

作者简介:



胡昭华(1981-),**通信作者**,女,博士,副教授,硕士生导师,研究方向:视觉跟踪、模式识别,E-mail: zhaohua_hu@163.com。

刘浩男(1998-),男,硕士研究生,研究方向:目标跟踪,E-mail:lhn2014954846@163.com。

林潇(1996-),男,硕士研究生,研究方向:目标跟踪,E-mail:qq1018408006@gmail.com。

(编辑:夏道家)