计算机科学
COMPUTER SCIENCE

融合注意力特征的无锚框视觉目标跟踪方法

李雪辉，张拥军，史殿习，徐化池，史燕燕

引用本文

李雪辉，张拥军，史殿习，徐化池，史燕燕融合注意力特征的无锚框视觉目标跟踪方法[J]．计算机科学，2023，50(1)：138-146.

LI Xuehui，ZHANG Yongjun，SHI Dianxi，XU Huachi，SHI Yanyan. AFTM:Anchor-free Object Tracking Method with Attention Features [J]. Computer Science，2023，50(1)：138-146.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211000083

融合注意力特征的无锚框视觉目标跟踪方法

李雪辉[1] 张拥军[1] 史殿习[1,2,3] 徐化池[1] 史燕燕[2]
1 国防科技创新研究院 北京 100071
2 国防科技大学计算机学院 长沙 410073
3 天津(滨海)人工智能创新中心 天津 300457
(xhli_niidt@163.com)

摘 要 目标跟踪作为计算机视觉领域的一个重要分支,在智能视频监控、人机交互和自动驾驶等诸多领域具有很高的研究价值。尽管目标跟踪近年来已取得较好的发展,但在复杂跟踪环境下,遮挡、目标形变、光照变化等因素仍会导致跟踪精度下降,跟踪性能不稳定。因此,提出了一种融合注意力特征的无锚框视觉目标跟踪方法(Anchor-Free object Tracking Method,AFTM)。首先,在分类和回归过程中构建自适应生成的注意力权重因子组,实现了一种高效的自适应响应图融合策略,提高了目标定位和边界框尺度计算的准确性;其次,针对数据集中样本类别不均衡的现象,使用可动态缩放的交叉熵损失作为目标定位网络的损失函数,修正模型的优化方向,使跟踪性能更加稳定可靠;最后,设计相应的学习率调整策略,对一定数量的模型进行随机权重平均,增强模型的泛化能力。公开数据集上的实验结果表明,在复杂跟踪环境下,AFTM 具有更高的精度和更稳定的跟踪效果。

关键词:深度学习;目标跟踪;孪生网络;无锚框;注意力机制
中图法分类号 TP391.41

AFTM:Anchor-free Object Tracking Method with Attention Features

LI Xuehui[1],ZHANG Yongjun[1],SHI Dianxi[1,2,3],XU Huachi[1] and SHI Yanyan[2]
1 National Innovation Institute of Defense Technology,Beijing 100071,China
2 College of Computer,National University of Defense Technology,Changsha 410073,China
3 Tianjin Artificial Intelligence Innovation Center,Tianjin 300457,China

Abstract As an important branch in the field of computer vision,object tracking has been widely used in many fields such as intelligent video surveillance,human-computer interaction and autonomous driving. Although object tracking has achieved good development in recent years,tracking in complex environment is still a challenge. Due to problems such as occlusion,object deformation and illumination change,tracking performance will be inaccurate and unstable. In this paper,an effective object tracking method AFTM,is proposed with attention features. Firstly,this paper constructs an adaptively generated attention weight factor group,which implements an efficient adaptive fusion strategy for response map to improve the accuracy of object positioning and bounding box scale calculation in the process of classification and regression. Secondly,aiming at the class imbalance in the data set,the proposed method uses the dynamically scaled cross entropy loss as the loss function of the object positioning network,which can modify the optimization direction of the model and make the tracking performance more stable and reliable. Finally,this paper designs a corresponding learning rate adjustment strategy to stochastically average the weight of a number of models,which can enhance the generalization ability of the model. Experimental results on public data sets show that the proposed method has higher accuracy and more stable tracking performance in complex tracking environment.

Keywords Deep learning,Object tracking,Siamese network,Anchor-free,Attention mechanism

到稿日期:2021-10-14 返修日期:2022-04-15
基金项目:国家重点研发计划(2017YFB1001901);天津市滨海新区合作共建研发平台科技项目(BHXQKJXM-PT-RGZNJMZX-2019001)
This work was supported by the National Key Research and Development Program of China(2017YFB1001901)and Science and Technology Commission of Tianjin Binhai New Area(BHXQKJXM-PT-RGZNJMZX-2019001).
通信作者:张拥军(yjzhang@nudt.edu.cn)

1 引言

目标跟踪作为计算机视觉领域的一个重要课题和研究热点，在智能视频监控、现代化军事、人机交互和自动驾驶等诸多领域得到了广泛应用。目标跟踪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[1]。随着大规模公开标注的图像数据资源的出现以及计算机硬件计算能力的进步和发展，优秀的目标跟踪方法不断涌现，其相关研究也取得了极大的进展。然而，由于真实跟踪场景复杂多变，目标跟踪技术仍面临着遮挡、目标形变、光照变化、背景杂乱等一系列问题的挑战[2]。因此，有效解决上述问题对目标跟踪技术的发展具有重要意义。

近年来，基于孪生网络的目标跟踪算法因具有较高的准确性和实时性，受到了广泛的关注与研究，并取得了显著成果。基于孪生网络的目标跟踪算法通过构建权值共享的孪生网络，将目标模板与候选样本进行匹配，得到图像块之间的相似性程度，从而实现对目标状态的判断。2016 年 Bertinetto 等[3]提出基于端到端全卷积孪生网络的目标跟踪算法 SiamFC，这是孪生网络在视觉目标跟踪领域的首次成功应用。随后，Valmadre 等[4]引入相关滤波的方法，提出了 CFNet 算法，不仅提升了跟踪速度，同时还可以在线学习优化特征表示。为了有效应对目标的多尺度变化，Li 等[5]将目标检测中常用的候选区域生成网络（RPN）应用到了目标跟踪的孪生网络中，提出了 SiamRPN 算法，使模型可以预测任意尺度、形状的目标，算法的准确率和帧率都有了很大的提升。在 SiamRPN 的基础上，DaSiamRPN[6]引入了干扰感知模块，并对训练用的样本集进行有效扩充，实现了长时跟踪，提升了算法在复杂场景下的鲁棒性；SiamRPN＋＋[7]利用深度互相关结构，打破了浅层网络架构的限制，利用多层聚合更大程度地发挥深层网络的作用，在各个权威基准数据集上取得了优秀结果。Zhang 等[8]提出无填充残差单元，设计了一种新的用于孪生跟踪器的更深更宽的网络架构。SiamMask[9]首次将跟踪与分割任务相结合来直接预测物体的掩码（Mask），对跟踪算法整体的精度产生了重大影响。为有效解决模板退化的问题，Zhang 等[10]提出一种孪生网络框架下的模板更新策略，用初始模板、之前累积的模板和当前帧的模板生成下一帧可以用的最优模板，对新的目标模板进行了有效预测。受计算机领域注意力机制成功应用的启发，Yu 等[11]为提升孪生跟踪器的特征学习能力，提出了一种可变形孪生注意力网络，增强了特征对目标外观变化的鲁棒性。为避免预定义锚框及其所带来的相关参数，SiamBAN（Siamese Box Adaptive Network for visual tracking）[12]引入了无锚框的思想，实现了目标尺度和纵横比的准确估计，使跟踪器更加灵活通用。

经过多年的发展，基于孪生网络的目标跟踪算法不断进步，并因其良好的跟踪精度和效率而受到了极高关注，已成为目标跟踪算法的重要分支。此类算法一般采用端到端的框架结构，虽然跟踪速度较快，但降低分类与回归任务的复杂程度，更好地利用具有丰富语义信息的特征表达，以及在复杂多变的真实环境中保证跟踪精度与稳定性等，仍是影响跟踪器性能的关键。针对上述问题，本文提出了一种融合注意力特征的无锚框视觉目标跟踪方法（AFTM）。本文的主要贡献如下：

（1）将孪生网络架构与无锚框的回归方式相结合，实现了简洁方便的回归操作，降低了分类与回归任务的复杂程度，同时无须对锚框超参数进行调优。

（2）在分类和回归模块，构建自适应生成的注意力权重因子组，实现了一种高效的自适应响应图融合策略，得到了更加准确的目标定位和边界框尺度，提高了跟踪精度。

（3）使用可动态缩放的交叉熵损失作为目标定位网络的损失函数，修正模型的优化方向，极大地缓解了样本类别不均衡的问题，使跟踪性能更加稳定可靠。

（4）设计了相应的学习率调整策略，对一定数量的模型进行随机权重平均，以增强模型的泛化能力。

实验结果表明，在 VOT2018 和 VOT2019 数据集上，AFTM 具有更高的精确度和更稳定的跟踪效果。

2 基于孪生网络的目标跟踪算法与目标边界框的回归方法

本节主要对基于孪生网络的目标跟踪算法原理及目标边界框的回归方法进行阐述。

2.1 基于孪生网络的目标跟踪算法原理

孪生网络是一种双分支的网络结构，其架构图如图 1 所示。孪生网络用于接收两分支的输入 X_1 和 X_2，通过一组参数共享的映射函数 $G_W(X)$，将两个输入映射到目标空间，得到 $G_W(X_1)$ 和 $G_W(X_2)$，并在目标空间衡量两个输入的相似程度 E_W。最初，孪生网络主要应用于支票签名对比，随后在计算机视觉、自然语言处理等领域也有了广泛应用。

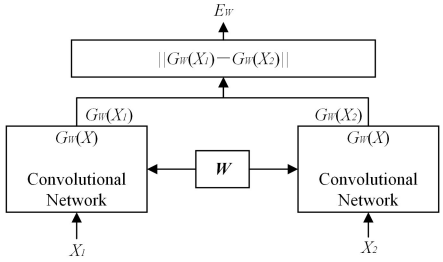


图 1 孪生网络架构图

Fig. 1 Siamese network architecture

基于孪生网络的目标跟踪算法通过构建权值共享的孪生网络，将目标模板区域和待搜索图像区域作为两路并行输入，送到网络中进行特征提取，将得到的目标模板特征与待搜索图像特征进行相关卷积操作，生成相似度响应图，根据响应图来进行前后背景的分类与边界框回归。此类方法在大规模数据集上进行离线训练，当在线跟踪时，网络权重固定，保证了跟踪的速度；其次，由于孪生网络的结构特点，两分支输入能够进行相同语义下的特征映射，有效增强了网络的判别能力；另外，用于匹配的目标模板区域通常不进行更新，避免了目标模板被污染。

2.2 目标边界框的回归方法

在目标跟踪任务中，目标边界框回归指对目标区域生成

精准的位置预测。计算机视觉领域中，目标检测与目标跟踪往往互相启发、共同发展，且两者的回归任务类似，均与类别无关，因此可以将目标检测的回归思想引入到目标跟踪中。当前，常用的边界框回归方法按有无锚框分为以下两种。

(1)基于锚框的回归方法。锚框的概念起源于目标检测算法，其几何化表示即为固定尺寸的矩形边界框。此类方法通过预设多尺度与不同宽高比的锚框作为候选框来实现稠密采样，神经网络对每个候选框进行类别预测后，计算正样本候选框的偏移量，调整正样本候选框的位置以生成预测边界框，在预测边界框和真实目标边界框之间进行损失计算，找到最佳的预测边界框并输出。其中，Faster R-CNN[13]直接利用锚框坐标的偏移量来计算预测边界框的位置；YOLOv2[14]则通过划分网格来计算预测边界框相对于网格位置的偏移量，将其中心点限制在网格内部，使预测边界框能够更快逼近真实边界框。尽管基于锚框的回归方法有着较高的定位精度，但生成锚框需要预设众多参数并进行人工调优，同时生成的锚框数量巨大，带来了更多复杂的计算。

(2)无锚框的回归方法。相比利用锚框回归，此类方法避免了与锚框相关的众多参数与复杂计算，简单方便，并且在定位精度上有着较好的表现。针对去掉锚框后应如何描述候选框的问题，CornerNet[15]提供了一种有效的解决思路，即通过关键点描述候选框。首先，检测目标框的左上角点和右下角点，然后对所有角点进行分类和配对，生成预测的目标边界框。由于仅检测两个角点容易忽略物体内部具有判别性的特征，导致许多误检的产生。Duan 等[16]对目标中心点进行了额外预测，即用 3 个关键点来确定目标，增强了对物体内部信息的感知能力，有效抑制了误检。CenterNet[17]则通过预测目标中心点和目标框的宽高生成预测框。FCOS[18]将输入图像的每个像素都视为样本点，预测样本点到目标框的 4 条边的距离。无锚框的回归方法在设计上直观方便，能够处理待跟踪目标具有较大形变的状况，且泛化能力强，十分适合应用到目标跟踪任务中。

3 融合注意力特征的无锚框视觉目标跟踪方法

针对复杂多变的真实环境中跟踪精度下降、跟踪性能不稳定的问题，本文提出了一种融合注意力特征的无锚框视觉目标跟踪方法——AFTM，其总体框架如图 2 所示。该方法基于孪生网络，将 SiamBAN 作为基本模型，并沿用其无锚框的回归方式，在确保分类与回归任务复杂程度较低的同时，更加关注包含丰富语义信息的特征，使 AFTM 能够在遮挡、目标形变与光照变化等复杂环境下保持较好的跟踪效果。

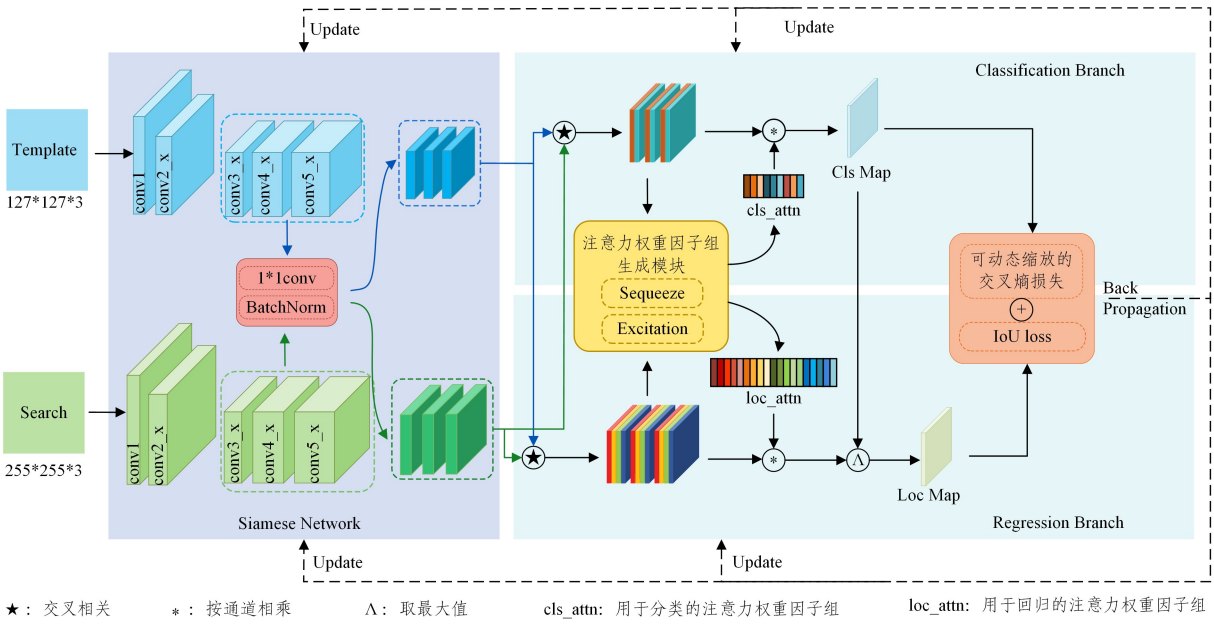


图 2 AFTM 网络框架图

Fig. 2 AFTM network architecture diagram

AFTM 的网络框架由一个用于特征提取的孪生网络和用于分类、回归的两个子网络构成，分类分支负责前后背景的分类，获取目标粗略的位置；回归分支负责预测目标精确的位置，生成预测边界框。在网络的离线训练阶段，接收模板图像与搜索区域图像作为输入，提取特征后分别计算多层的分类与回归响应图；注意力权重因子组生成模块嵌入在两个子网络中，负责生成注意力权重因子组，自适应融合多层响应图，生成最终的分类与回归响应图，在提升响应图质量的同时，使其更加关注对跟踪结果有利的特征，进行准确的目标跟踪；在损失计算部分，使用可动态缩放的交叉熵损失作为分类分支的损失函数，从而提高分类正确率，保证模型的优化方向不被误导，使跟踪性能更加稳定可靠，回归损失则为交并比损失[19]；训练过程采用随机权重平均方法，使模型更加接近最优解，增强模型的泛化能力。通过不断优化损失函数，更新网络参数，得到最终的跟踪模型。

3.1 注意力权重因子组的生成

基于孪生网络的目标跟踪框架中，能否获取到准确的目标定位和预测边界框将直接影响到跟踪结果的好坏，而目标定位和预测边界框均由响应图计算得出。因此，响应图的质量是影响跟踪性能的关键。针对上述问题，我们引入了注意

力机制，构建了注意力权重因子组生成模块，并将其嵌入到分类与回归子网络中，实现了一种高效的自适应响应图融合策略。该策略能够提升响应图质量，得到更加准确的目标定位和边界框尺度，从而提高跟踪精度，得到更好的跟踪效果。

传统的基于孪生网络进行目标跟踪的方法是同等地对待各层、各通道的图像特征，然而浅层与深层特征所包含的语义信息不同，每个通道所蕴含信息的贡献也不等。因此，本文引入注意力机制，按特征的重要性程度赋予其不同的权重，使得网络更加关注有利于目标跟踪的特征，忽略无用信息。注意力权重因子组生成模块的工作原理是接收分类、回归两分支各自计算出的多层响应图，依据其包含目标信息的丰富程度，生成相应的注意力权重因子组，按通道作用于响应图并进行特征融合，得到最终的分类和回归响应图，并据此计算目标定位和预测边界框。

本文方法中，主干网络为 ResNet50，选用 conv3，conv4，conv5 提取到的特征作为分类与回归分支的输入。分类与回归分支接收特征后，使用深度交叉相关[7]（Depth-wise Cross-Correlation）分别计算响应图 P，计算式如式(1)与式(2)所示：

$$P_{\mathrm{cls}}^{W*H*2}=[\varphi(x)]_{\mathrm{cls}}\cdot[\varphi(z)]_{\mathrm{cls}} \tag{1}$$

$$P_{\mathrm{reg}}^{W*H*4}=[\varphi(x)]_{\mathrm{reg}}\cdot[\varphi(z)]_{\mathrm{reg}} \tag{2}$$

其中，x 和 z 分别表示搜索区域图像和模板图像；φ 为孪生网络特征提取操作，$\varphi(x)$ 和 $\varphi(z)$ 为提取到的搜索区域特征与模板特征；cls 与 reg 分别标明计算属于分类或回归分支；$\cdot$ 表示深度交叉相关计算；P_{cls}^{W*H*2} 与 P_{reg}^{W*H*4} 为计算得到的分类响应图和回归响应图。注意力权重因子组生成模块由 3 个子模块构成，分别是 Deform 子模块、Squeeze-and-Excitation（SE）[20] 子模块和 Concatenate 子模块，其运行原理如图 3 所示。

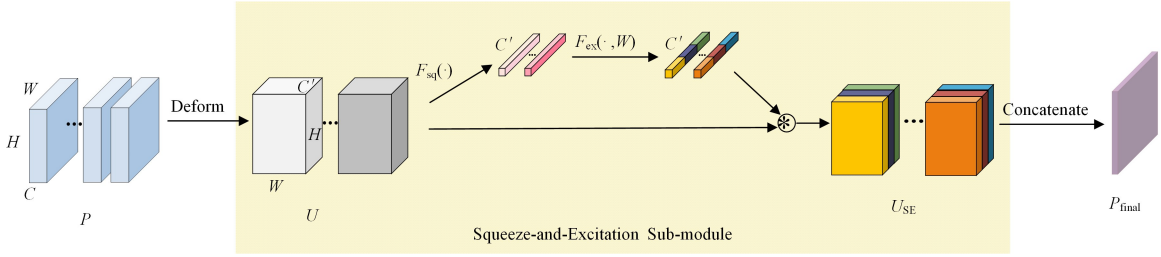


图 3 注意力权重因子组生成模块的运行原理图

Fig. 3 Operation schematic diagram of attention weight factor group generation module

Deform 子模块负责响应图的变形操作，其接收到由分类与回归分支产生的原始多层响应图 P 后，将 P 在不同通道上的特征进行拆分，得到相同通道的特征并进行拼接，输出为变形后的响应特征 U。经过变形操作后，两分支响应图的大小得到了统一。SE 子模块负责生成注意力权重因子组，并对各分支的多层响应图进行加权。该子模块接收变形后的响应特征 U，如式(3)所示，首先通过 Squeeze 操作在空间维度上使用全局平均池化（Global Average Pooling，GAP）压缩每个通道上的空间特征信息，得到每个通道信息的描述子，即全局描述特征 $d_{c'}$；然后，Excitation 操作利用生成的通道信息描述子对通道关系进行建模，如式(4)所示，使用两层全连接层、ReLU 和 sigmoid 激活函数，自适应地学习通道间的非线性关系，得到通道上的激励 s；最后，如式(5)所示，激励 s 作为特征图中通道上的权重，代表着各通道上信息的重要程度，与接收到的变形响应特征 U 按通道相乘后，完成在通道维度上对原特征的重标定，输出加权后的响应特征 U_{SE}。Concatenate 子模块负责通道特征的融合与拼接，接收加权响应特征 U_{SE}，融合相同通道上的特征后，将不同通道的特征进行拼接，输出最终的响应图 P_{final}。

$$d_{c'}=F_{\mathrm{sq}}(u_{c'})=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}u_{c'}(i,j) \tag{3}$$

$$s=F_{\mathrm{ex}}(d,W)=\sigma(W_2\delta(W_1 d)) \tag{4}$$

$$U_{\mathrm{SE}}=s*U \tag{5}$$

其中，$u_{c'}(i,j)$ 为变形操作后每个通道 c' 上的特征图，(i,j) 为特征图上像素点的坐标，$c'\in C'$；$F_{\mathrm{sq}}(\cdot)$ 和 F_{ex}"$(\cdot,W)$ 分别代表 Squeeze 操作与 Excitation 操作；W 是全连接层操作，负责完成通道的降维、升维；$\delta(\cdot)$ 和 $\sigma(\cdot)$ 分别为 ReLU 和 sigmoid 激活函数。

注意力权重因子组生成模块接收分类和回归分支生成的原始多层响应图 P，输出为自适应融合后的最终响应图 P_{final}。此模块引入注意力机制，根据不同特征所包含的不同信息，生成注意力权重因子组，实现了响应特征的自适应加权融合；同时，有侧重地筛选特征信息，对包含跟踪目标的特征赋予较大权重，抑制了其他类别的物体与无用信息，提升了响应图质量，从而更加准确地进行目标状态预测与位置标定。

3.2 可动态缩放的交叉熵损失

目前，基于孪生网络的目标跟踪算法在网络训练中存在样本类别不均衡、分类准确率低的问题，易导致训练所得的跟踪模型在复杂多变的场景中无法准确地跟踪目标，性能发挥不稳定。针对上述问题，本文使用可动态缩放的交叉熵损失[21]（Focal Loss）作为分类分支的损失函数，用于在网络训练时修正模型的优化方向，提高分类正确率与算法精度，减少复杂环境下跟踪误差与跟踪目标偏移情况的发生，保证跟踪的稳定与可靠。

在网络训练中，损失函数负责评估真实值与预测值之间的差异，即衡量模型预测的好坏。尽管相对于利用锚框的跟踪算法而言，无锚框的跟踪算法的负样本数量已大大减少，但仍远多于正样本。因此，为保证模型的优化方向不被大量简单的负样本误导，准确地进行前后背景分类，设计一个有效的分类损失函数十分重要。利用标准的交叉熵损失（Cross Entropy Loss）训练网络时，所得模型容易受到前后背景类别失衡的影响，易分类样本与负样本在总损失函数的输入参数中占较大比重，导致模型训练退化，同时训练效率低下，算法

精度难以提升，判别能力难以增强。AFTM 采用可动态缩放的交叉熵损失作为分类损失函数，使大量简单负样本所带来的损失不再主导梯度下降的方向，更加专注于难分类样本与正样本，极大地缓解了样本类别不均衡的问题。

可动态缩放的交叉熵损失在标准交叉熵损失的基础上，通过引入平衡因子与调制系数实现了损失的动态缩放，自动降低了易分类样本与负样本对损失的贡献，增加了难分类样本与正样本在损失中所占的比重，使模型的预测值能够更好地逼近真实值。

标准交叉熵损失的计算式如式(6)与式(7)所示：

$$p_t=\begin{cases}p, & y=1\\ 1-p, & y=0\end{cases} \tag{6}$$

$$L(p,y)=L(p_t)=-\log(p_t) \tag{7}$$

其中，p 与 $1-p$ 分别为模型对正负样本类别的估计概率，为便于表示，记样本的估计概率为 p_t，$p_t\in[0,1]$。引入平衡因子与调制系数后，可动态缩放的交叉熵损失的计算式如式(8)与式(9)所示：

$$\alpha_t=\begin{cases}\alpha, & y=1\\ 1-\alpha, & y=0\end{cases} \tag{8}$$

$$L_{fl}(p_t)=-\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{9}$$

其中，α 与 $1-\alpha$ 分别为正负样本类别的平衡因子，为便于表示，记平衡因子为 α_t，$\alpha_t\in[0,1]$；$\gamma\in[0,\infty)$ 为聚焦参数，$(1-p_t)^{\gamma}$ 为调制系数。计算损失时，平衡因子 α_t 与调制系数 $(1-p_t)^{\gamma}$ 共同实现对权重的控制，α_t 负责调节比例不均的正负样本，$(1-p_t)^{\gamma}$ 负责控制难易样本所占的比重。在网络的训练过程中，当遇到简单负样本时，p_t 值偏大，调制系数接近于 0，权重降低，减少了此类样本对损失的贡献；当遇到难分类样本时，p_t 值偏小，调制系数接近于 1，损失计算不受影响，使网络倾向于利用此类样本进行参数更新。此外，平衡因子与调制系数对权重的控制是动态变化的，当复杂的样本逐渐变得易于分类时，此类样本对损失的影响也会逐渐下降。

AFTM 的损失函数由分类损失和回归损失两部分组成，如式(10)所示：

$$L=\lambda_1 L_{\text{cls}}+\lambda_2 L_{\text{reg}} \tag{10}$$

其中，分类损失 L_{cls} 采用式(9)中的 L_{fl}，回归损失 L_{reg} 为交并比损失，$\lambda_1=\lambda_2=1$。根据实验结果，分类损失 L_{fl} 中的 α 取 0.25，γ 取 2。改进后的损失函数进一步缓解了样本类别不均衡的问题，使网络训练更加正确有效，同时增强了跟踪模型的判别能力和稳定性，使其能够在复杂环境或相似背景下以较高的精度进行持续可靠的目标跟踪。

3.3 随机权重平均方法

神经网络的训练过程可以看作是最小化损失函数的过程。神经网络训练时，通过误差反向传播不断更新网络的权重参数，当损失函数收敛于某局部最小值时，此时对应的网络权重参数即为模型的一个解。为使得到的模型更加接近最优解，进一步提升模型的整体性能，增强其泛化能力，AFTM 采用随机权重平均方法[22](Stochastic Weight Averaging，SWA)进行训练。

传统的随机梯度下降法(Stochastic Gradient Descent，SGD)寻找的单个局部最优解不一定是全局最优。随机权重平均法 SWA 的原理则是搜寻多个局部解后，对所得解求和取平均作为最终模型的权重参数。实际训练中，数据集的差异性使训练集上训练得到的模型在测试集上的评估结果存在偏差，而通过 SWA 方法训练得到的解位于解空间中较平滑的区域，此时得到的模型泛化能力好，不会导致模型在测试集上的性能出现太大的偏差。因此，AFTM 训练时采用 SWA 方法，在第一阶段通过随机梯度下降法 SGD 进行训练，直至模型对应的损失函数接近收敛；第二阶段，变换学习率策略继续训练，对 SGD 方法搜寻到的较优解求和取平均，并将其作为模型的最终解。

AFTM 设计的学习率策略对应 SWA 方法的两个阶段，在第一阶段通过 SGD 方法收敛损失函数时，采用 Warmup 学习率预热策略与指数衰减策略；进入第二阶段时，将学习率设为恒定值，在训练末尾对得到的较优解求和取平均，得到最终解。在第一阶段，SGD 方法通常采用指数衰减策略收敛损失函数，为避免训练初期直接使用较大学习率引起的模型震荡，故加入 Warmup 学习率预热策略，从较小学习率开始爬升，爬升到指定值后再进行指数衰减，使模型趋于稳定后进行更好的收敛。在第二阶段，学习率已衰减到较小值，此时搜寻的较优解只会在最小值附近小幅度波动，直至模型收敛。因此，在训练末尾对较优解求和取平均时，学习率设为相对较大的恒定值，可避免模型陷入单个局部极小值点，且保证了搜寻到的多个解的差异性。

通过 SWA 方法对神经网络进行训练得到的跟踪模型的整体性能与泛化能力更优，能够在遮挡、目标形变与光照变化等复杂环境下保持较好的跟踪表现。

3.4 算法实现

融合注意力特征的无锚框视觉目标跟踪方法的网络离线训练过程如 3.3 节所述，其具体实现如算法 1 所示。

算法 1 AFTM 算法

输入：第一帧目标模板图像 z_1 与位置 l_1，待跟踪的视频序列及相应参数

输出：第 t 帧目标位置 l_t

Step1 加载训练得到的跟踪模型，利用孪生神经网络 Φ 对模板图像 z_1 提取特征，得到 $\varphi(z_1)$ 并保存；

Step2 进行后续帧的跟踪，在上一帧的目标位置附近裁剪搜索区域 x_t，提取特征得到 $\varphi(x_t)$；

Step3 $\varphi(z_1)$ 与 $\varphi(x_t)$ 进行相关卷积操作，得到多层的分类与回归响应图；

Step4 通过注意力权重因子组生成模块对多层响应图进行自适应融合，生成最终的分类与回归响应图；

Step5 根据最终响应图计算出预测边界框，利用余弦窗和尺度变化惩罚来平滑目标的位置偏移与大小改变；

Step6 选择最佳的预测边界框，用上一帧的目标状态进行线性插值更新其大小，生成预测的目标位置 l_t，重复 Step2— Step6 直到视频序列最后一帧结束。

4 实验及结果分析

4.1 实验设计

AFTM 的实验环境如表 1 所列。主干网络采用在 ImageNet 上预训练过的权值进行初始化，冻结前两层的网络参数。训练数据集采用 ImageNet VID[23]，YouTube-BoundingBoxes[24]，COCO[25]，ImageNet DET[23]，GOT10k[26] 和 La-

SOT[27]，训练数据以图片对的方式输入孪生神经网络，模板分支和搜索分支接收的图片大小分别为 127 * 127 和 255 * 255。网络的离线训练过程如 3.3 节所述，训练周期为 28 个 epoch，batch 大小设为 32，前 5 个 epoch 学习率从 0.001 增加至 0.005 为预热阶段，随后 15 个 epoch 学习率从 0.005 逐渐衰减到 0.00005，然后固定学习率直至训练周期结束。衰减权重和动量设为 0.0001 和 0.9，在前 10 个 epoch 冻结主干网络权值，从第 11 个 epoch 起对整个网络进行端到端的训练。

表 1 实验环境及参数说明

Table 1 Experimental environment and parameters

实验环境		参数说明
硬件配置	CPU	Intel(R) Xeon(R) Gold 6254
	GPU	NVIDIA TITAN V
软件平台	操作系统	Ubuntu 18.04
	代码语言	Python 3.7.9
	所用框架	Pytorch 1.3.1

4.2 定量分析

为验证 AFTM 的有效性，测试数据集采用两个基准数据集 VOT2018[28] 与 VOT2019[29]，并选取当前主流的跟踪算法进行对比。在 VOT2018 上进行对比的算法有 DRT[30]，UPDT[31]，SiamRPN[5]，LADCF[32]，ATOM[33]，SiamRPN＋＋[7]，SiamBAN[12]；在 VOT2019 上进行对比的算法有 SA_SIAM_R[29]，SPM[34]，SiamRPN＋＋[7]，SiamMask[9]，ARTCS[29]，SiamDW_ST[8]，SiamBAN[12]。

跟踪器使用精度(Accuracy)、鲁棒性(Robustness)和期望平均重叠率(Expected Average Overlap，EAO)作为主要评价标准，并在目标丢失时被初始化。其中，Accuracy 计算预测区域与真值区域的平均重叠率，用于评价跟踪的准确度，数值越大，准确度越高；Robustness 计算跟丢时的平均失败帧数，用于评价跟踪情况是否稳定，数值越小，稳定性越好；EAO 计算平均重叠率的期望值，用于描述跟踪器的整体性能，数值越大，跟踪器的效果越好。AFTM 与当前主流的跟踪算法在 VOT2018 数据集上的对比结果如表 2 所列，在 VOT2019 数据集上的对比结果如表 3 所列。图 4、图 5 分别给出了两个数据集上精度与鲁棒性的排序。

表 2 VOT2018 上各算法的实验结果

Table 2 Experimental results of each algorithm on VOT2018

	DRT	UPDT	SiamRPN	LADCF	ATOM	SiamRPN＋＋	SiamBAN	Ours
Accuracy	0.518	0.536	0.588	0.503	0.590	*0.604*	0.597	**0.609**
Robustness	0.201	0.184	0.276	**0.159**	0.203	0.234	*0.178*	0.183
EAO	0.355	0.379	0.384	0.389	0.401	0.417	**0.452**	*0.424*

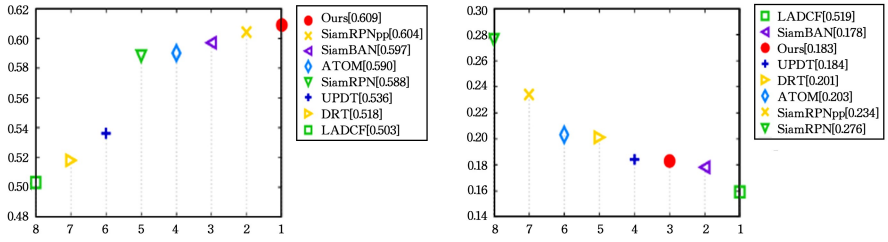
注：加粗数据为最优结果，斜体数据为次优结果

表 3 VOT2019 上各算法的实验结果

Table 3 Experimental results of each algorithm on VOT2019

	SA_SIAM_R	SPM	SiamRPN＋＋	SiamMask	ARTCS	SiamDW_ST	SiamBAN	Ours
Accuracy	0.563	0.577	0.599	0.594	*0.602*	0.600	*0.602*	**0.608**
Robustness	0.507	0.507	0.482	0.461	0.482	0.467	*0.396*	**0.361**
EAO	0.252	0.275	0.285	0.287	0.287	0.299	**0.327**	*0.301*

注：加粗数据为最优结果，斜体数据为次优结果



(a) Accuracy (b) Robustness

图 4 VOT2018 上的精度与鲁棒性排序

Fig. 4 Accuracy and robustness performance on VOT2018

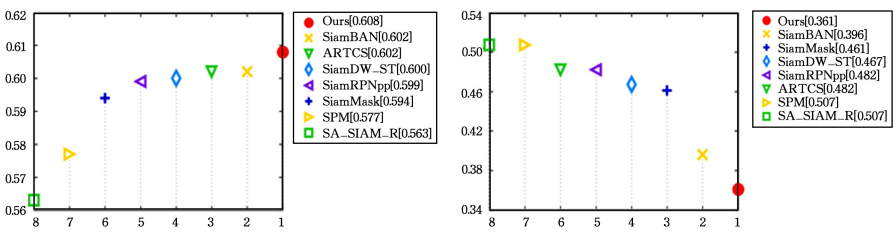


图 5 VOT2019 上的精度与鲁棒性排序

Fig. 5 Accuracy androbustness performance on VOT2019

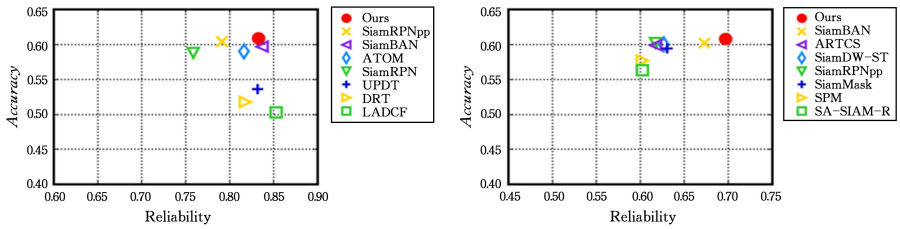
对于精度指标 Accuracy，AFTM 在 VOT2018 数据集与 VOT2019 数据集上分别达到了 0.609 和 0.608，均取得了

最优的表现。在基于孪生网络的跟踪算法中，AFTM 在 VOT2018 数据集上相比 SiamRPN，SiamRPN++与 SiamBAN，精度提升了 0.5%～2.1%不等；在 VOT2019 数据集上相比 SA_SIAM_R，SiamRPN++，SiamMask，SiamDW_ST 与 SiamBAN，其精度提升了 0.6%～4.5%不等。同时，对于非孪生网络的跟踪算法来说，AFTM 在精度上也有明显提升。可以看出，在不同的跟踪场景下，融合注意力特征的无锚框视觉目标跟踪方法能够使预测结果更加逼近待跟踪目标，实现准确的目标跟踪。

对于鲁棒性指标 Robustness，AFTM 在 VOT2018 数据集上的表现优于 UPDT，DRT，ATOM 等 5 种算法，取得了不错的效果；同时，在 VOT2019 数据集上取得了最佳表现 0.361，分别比排名第二位的 SiamBAN 和第三位的 SiamMask 提升了 3.5%和 10%。由此可以看出，对于利用孪生网络或其他神经网络进行跟踪的算法来说，AFTM 的鲁棒性较强，能够尽量避免丢失目标，减少跟踪误差与跟踪目标偏移情况的发生，保证跟踪的稳定性与可靠性。此外，对于 EAO 指标，AFTM 在两个数据集上都能达到次优的效果，说明其在保证高精度与稳定跟踪的情况下，整体性能与其他算法相比也具有较强的竞争力。

图 6 给出了各跟踪算法在 Accuracy 与 Reliability 上的测评图，记做 AR 图。Reliability 为 Robustness 的基底转换值，当跟踪算法的鲁棒性越强，即 Robustness 值越接近于 0 时，Reliability 的值越接近于 1。可得，跟踪算法的 Accuracy 值越高，Robustness 值越低时，该算法的测评点越靠近图的右上方，代表其性能越优。

由图 6 可知，在 VOT2018 数据集上，AFTM 与 SiamBAN 的表现相当，都处于 AR 图的最右上方。相比之下，AFTM 的精度更高，SiamBAN 在 Reliability 上的表现更好。在 VOT2019 数据集上，AFTM 处于 AR 图的最右上方，Accuracy 与 Reliability 上的表现都是最优，相比其他算法更加稳健，跟踪性能更优越。



(a) VOT2018 (b) VOT2019

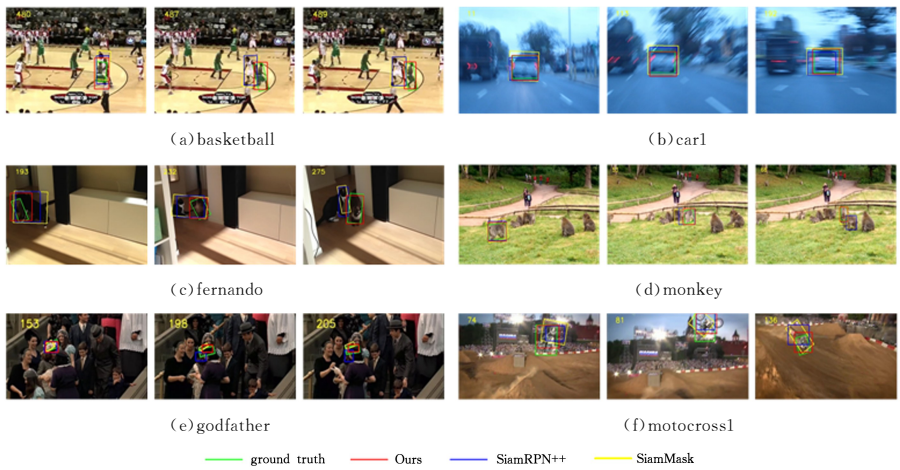
图 6 VOT2018 与 VOT2019 上的 AR 图

Fig. 6 AR map on VOT2018 and VOT2019

4.3 定性分析

为了更加直观地描述不同算法的跟踪性能，我们从测试数据集上选取了 basketball，car1，fernando 等 6 个富有挑战性的视频序列进行定性分析。

图 7 给出了 AFTM，SiamRPN++与 SiamMask 在所选视频序列上的可视化跟踪结果。



(a) basketball (b) car1

(c) fernando (d) monkey

(e) godfather (f) motocross1

图 7 部分视频序列上的跟踪结果

Fig. 7 Partial video sequence tracking results

在 basketball 视频序列中，目标在复杂背景下快速移动，SiamRPN++与 SiamMask 算法均跟踪失败，将相似目标作为了跟踪结果，而 AFTM 通过生成注意力权重因子组，使跟踪器更加关注目标区域的外观特征，减小了复杂背景中多个相似目标带来的不利影响，实现了正确稳定的跟踪。当 car1 视频序列中的目标因相机抖动而产生运动模糊时，各算法都在不同程度上出现了跟踪误差。与其他算法相比，AFTM 通过不断适应场景的抖动变化，抵抗运动模糊因素的干扰，有效减小了跟踪结果与真值之间的误差，使目标跟踪更加准确。在 fernando 视频序列中，目标经历了较大的光照变化并伴有遮挡情况的发生，此时 AFTM 能够及时应对场景的明暗变化，以较高精度持续地跟踪目标，而另外两种算法的跟踪结果

均产生了较大的偏移。当 monkey 视频序列中的目标发生由大到小的尺度变化时，SiamRPN++算法在相似背景下丢失了目标，SiamMask 算法产生了一定的跟踪误差，而 AFTM 能够及时捕获目标外观的大小变化，计算更加准确的跟踪尺度，表现出了较好的跟踪性能。godfather 视频序列中存在严重的遮挡情况，并且目标为不易跟踪的小目标，此时 AFTM 仍能保持正确的跟踪趋势，具有较强的鲁棒性，而其他两种算法的跟踪结果偏移到了目标的边缘区域或附近的背景区域。在 motocross1 视频序列中，目标面临着较大的自身旋转挑战，同时伴有光照变化与运动模糊。在如此复杂的跟踪场景下，3 种算法都产生了一定的跟踪误差，AFTM 由于引入了注意力机制，对获取到的特征信息有侧重地进行筛选，使得跟踪误差相对较小，跟踪结果更加逼近真值，保证了目标跟踪的准确可靠。

结束语 本文针对复杂跟踪环境下，如何持续可靠地进行准确目标跟踪的问题，进行了深入研究，在孪生网络跟踪框架的基础上，提出了一种融合注意力特征的无锚框视觉目标跟踪方法——AFTM。首先，该方法在分类和回归模块中构建自适应生成的注意力权重因子组，实现了一种高效的自适应响应图融合策略，突出有利于目标跟踪的特征，抑制无用信息，使目标跟踪更加准确；其次，采用可动态缩放的交叉熵损失作为目标定位网络的损失函数，进一步缓解了样本类别不均衡的问题，增强了跟踪模型的判别能力和稳定性；最后，在训练过程中对一定数量的模型进行随机权重平均，并设计策略对学习率进行相应的调整，使 AFTM 的整体性能与泛化能力更优。本文对比了 AFTM 与其他主流的目标跟踪算法，实验结果表明，在复杂跟踪环境下，AFTM 具有更高的精度和更加稳定可靠的跟踪效果。跟踪过程中，当目标遇到遮挡、形变或光照变化对目标外观产生影响时，如何对模板进行调整与更新，使得算法能够及时捕捉到目标最新的状态信息，从而进行更好的目标跟踪，是下一阶段的重要工作。

参 考 文 献

[1] LI X，ZHA Y F，ZHANG T Z，et al. Survey of visual object tracking algorithms based on deep learning[J]. Journal of Image and Graphics，2019，24(12)：2057-2080.

[2] LU H C，LI P X，WANG D. Visual Object Tracking：A Survey[J]. Pattern Recognition and Artificial Intelligence，2018，31(1)：61-76.

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

[4] 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. 2017：2805-2813.

[5] 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. 2018：8971-8980.

[6] 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). 2018：101-117.

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

[8] 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. 2019：4591-4600.

[9] 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. 2019：1328-1338.

[10] ZHANG L，GONZALEZ-GARCIA A，WEIJER J，et al. Learning the model update for siamese trackers[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：4010-4019.

[11] YU Y，XIONG Y，HUANG W，et al. Deformable siamese attention networks for visual object tracking[C]// Proceedings of the IEEE/CVF Confe-rence on Computer Vision and Pattern Recognition. 2020：6728-6737.

[12] 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. 2020：6668-6677.

[13] REN S，HE K，GIRSHICK R，et al. Faster r-cnn：Towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems，2015，28：91-99.

[14] REDMON J，FARHADI A. YOLO9000：better，faster，stronger[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：7263-7271.

[15] LAW H，DENG J. Cornernet：Detecting objects as paired keypoints[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018：734-750.

[16] DUAN K，BAI S，XIE L，et al. Centernet：Keypoint triplets for object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：6569-6578.

[17] ZHOU X，WANG D，KRÄHENBÜHL P. Objects as points[J]. arXiv：1904. 07850，2019.

[18] TIAN Z，SHEN C，CHEN H，et al. Fcos：Fully convolutional one-stage object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019：9627-9636.

[19] REZATOFIGHI H，TSOI N，GWAK J Y，et al. Generalized intersection over union：A metric and a loss for bounding box regression[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019：658-666.

[20] HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：7132-7141.

[21] LIN T Y，GOYAL P，GIRSHICK R，et al. Focal loss for dense object detection[C]// Proceedings of the IEEE International

Conference on Computer Vision. 2017: 2980-2988.
[22] IZMAILOV P, PODOPRIKHIN D, GARIPOV T, et al. Averaging weights leads to wider optima and better generalization[J]. arXiv: 1803.05407, 2018.
[23] 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.
[24] REAL E, SHLENS J, MAZZOCCHI S, et al. Youtube-bounding boxes: 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. 2017: 5296-5305.
[25] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// European Conference on Computer Vision. Cham: Springer, 2014: 740-755.
[26] 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.
[27] 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. 2019: 5374-5383.
[28] KRISTAN M, LEONARDIS A, MATAS J, et al. The sixth visual object tracking vot2018 challenge results[C]// Proceedings of the European Conference on Computer Vision(ECCV) Workshops. 2018: 3-53.
[29] 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. 2019: 2206-2241.
[30] SUN C, WANG D, LU H, et al. Correlation tracking via joint discrimination and reliability learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 489-497.
[31] BHAT G, JOHNANDER J, DANELLJAN M, et al. Unveiling the power of deep tracking[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018: 483-498.
[32] XU T, FENG Z H, WU X J, et al. Learning adaptive discriminative correlation filters viatemporal consistency preserving spatial feature selection for robust visual object tracking[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5596-5609.
[33] 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. 2019: 4660-4669.
[34] WANG G, LUO C, XIONG Z, et al. Spm-tracker: Series-parallel matching for real-time visual object tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3643-3652.

LI Xuehui, born in 1997, postgraduate. Her main research interests include computer vision and object tracking.

ZHANG Yongjun, born in 1966, Ph.D, professor. His main research interests include artificial intelligence, multi-agent cooperation, machine learning and feature recognition.

(责任编辑:何杨)