doi:10.12068/j.issn.1005－3026.2025.20240018

结合运动信息与双重注意力机制的两阶段SiamCAR跟踪算法

魏 颖，张家鹏，崔佳琦，黄 通
（东北大学 信息科学与工程学院，辽宁 沈阳 110819）

摘 要：针对单目标跟踪中,因形变、运动模糊、遮挡以及背景干扰导致的跟踪框精度下降问题,特别是在背景干扰下易出现跟踪跳变及漂移问题,提出了一种结合运动信息和双重注意力机制的两阶段跟踪算法.第一阶段,使用带有双重注意力机制的SiamCAR跟踪器对当前帧的目标进行粗定位;第二阶段,利用像素级相似度运算构建边界框精细化模块,在低延迟情况下学习目标的细微特征以提升跟踪精度,并将基于外观特征得到的跟踪框与目标的运动轨迹信息相融合,以改善跟踪漂移及跳变问题.OTB100数据集上的实验结果表明,跟踪框的成功率和精度相比原来分别提高了4.6%和2.8%,在背景干扰下的成功率达到了69.6%.

关 键 词：单目标跟踪;SiamCAR;孪生网络;神经网络;注意力机制
中图分类号：TP 391 **文献标志码**：A **文章编号**：1005-3026(2025)09-0009-08

Two-Stage SiamCAR Tracking Algorithm Combining Motion Information and Dual-attention Mechanism

WEI Ying, *ZHANG Jia-peng*, *CUI Jia-qi*, *HUANG Tong*
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: WEI Ying, E-mail: weiying@ise.neu.edu.cn)

Abstract: In single-object tracking, the accuracy of the tracking bounding box is often compromised by factors such as deformation, motion blur, occlusion, and background interference. In particular, background interference frequently leads to tracking hopping and drift. To mitigate these issues, a two-stage tracking algorithm that integrated motion information with a dual-attention mechanism was proposed. In the first stage, a SiamCAR tracker with a dual-attention mechanism was employed to coarsely locate the target in the current frame. In the second stage, a refinement module of the bounding box was constructed using pixel-level similarity computations to learn the subtle features of the target under low-latency conditions, thereby enhancing the tracking accuracy. Finally, the tracking box obtained based on appearance features was fused with the target's motion trajectory information to mitigate tracking drift and hopping. Experimental results on the OTB100 dataset indicate that the success rate and accuracy of the tracking box have improved by 4.6% and 2.8%, respectively, compared to the original. The success rate in the presence of background interference has reached 69.6%.

Key words: single object tracking; SiamCAR; Siamese network; neural network; attention mechanism

计算机视觉研究领域中,目标跟踪是主要的研究方向之一,分为单目标跟踪和多目标跟踪[1].其中,单目标跟踪是计算机视觉领域的一项重要任务,旨在实时准确地定位视频序列中的特定目标并跟踪其运动轨迹[2].单目标跟踪实际应用广泛,例如自动驾驶、视频监控、面部动作捕捉等领域.

收稿日期：2024-01-17
基金项目：辽宁省重点研发计划项目(2024JH2/102500015);国家自然科学基金资助项目(61871106,62441231);中央高校基本科研业务费专项资金资助项目(N25BSS034).
作者简介：魏 颖(1968—),女,辽宁本溪人,东北大学教授,博士生导师.

近些年,基于孪生网络的跟踪器由于在目标跟踪中的良好表现得到广泛关注.早期的孪生网络在搜索区域的多尺度特征上进行匹配,导致网络跟踪速度较慢;SiamRPN[3]通过在孪生网络中附加一个区域建议提取子网,使得孪生网络的跟踪效率明显提升.目前主流的基于孪生网络的深度学习模型(如SiamRPN等)依赖于一组预定义的锚框,实现了高效的目标跟踪.由于基于锚框的跟踪器的性能对与锚框相关的超参数过于敏感[4],一些基于孪生网络的无锚跟踪器被设计出来,如SiamCAR[5],SiamBAN[6]等.尽管这类基于孪生网络的跟踪方法通过优化特征提取与跟踪模块不断提升了成功率和精度,但在面对复杂相似背景、外观变化、运动模糊及遮挡等情况时,仍难以有效区分目标与相似干扰物.

为了更好地应对上述单目标跟踪问题中的难题,许多学者基于深度学习理论提出了不同措施,以提高算法的性能.通过引入注意力机制[7],模型能够更有效地聚焦于感兴趣区域的细节特征,从而在复杂场景中提升对目标的识别与区分能力,展现出更强的判别性能.早期的跟踪器通常通过多尺度搜索[8]来进行边界框估计,但该策略不准确并且限制了跟踪器的性能.为了获得更鲁棒的跟踪结果,许多先进的跟踪器[9-10]采用多级跟踪策略,通过引入额外的跟踪阶段以获得更精确的边界框估计.其中,Alpha-Refine[11]网络可以在基础跟踪器得到的边界框基础上对目标边界框进行精确定位,从而提升边界框的估计质量.这些跟踪器首先对目标进行粗略定位,然后在附加跟踪阶段对初始结果进行细化,以获得更精确的边界框预测[11].

这些方法通过额外的跟踪阶段实现了更精确的跟踪,但没有将跟踪物体的运动信息引入到算法中,导致在存在相似物体干扰的情况下,跟踪框出现违背运动规律的跟踪跳变和漂移现象.

基于以上观察,本文提出了一个两阶段单目标跟踪框架.第一阶段使用融合双重注意力机制的SiamCAR跟踪算法进行边界框的粗定位,双重注意力机制能够有效提升跟踪器主干网络的特征提取能力,使其生成更具语义信息的特征图,从而增强应对各种跟踪挑战的综合表现;第二阶段设计了引入运动信息的边界框精细化模块,对第一阶段得到的边界框区域进行像素级的相似度计算,学习目标的细微特征,以解决在面对挑战时跟踪框精度不高的问题.接着将基于外观特征得到的跟踪位置与目标的运动轨迹信息相融合,并引入速度信息,利用物体的自然运动规律修正背景干扰下产生的错误跟踪,从而有效解决跟踪漂移及跳变问题.

1 算法框架

本文提出的整体算法框架由两个阶段组成.如图1所示,在第一阶段中,首先对当前帧图像和模板图像进行特征提取操作,获取待跟踪目标以及当前帧图像的特征信息,双重注意力机制的存在可以更有效地提取具有鉴别性的特征.经过分类回归子网后,得到第一阶段输出的较为粗糙的跟踪框,以此为基准在当前帧图像上进行裁剪(两倍大小)并输入第二阶段跟踪模块.使用像素级相似度计算使得第二阶段获得更高质量的特征表示,通过角点检测输出跟踪框.引入前4帧的跟踪结果与角点检测输出的跟踪框一起进行运动轨迹和运动向量计算,判断跟踪状态后进行跟踪修正,以达到精确的跟踪框估计.

1.1 基于双重注意力机制改进的SiamCAR跟踪算法

在第一阶段中,本文提出了一种双重注意力机制并将其引入到SiamCAR的特征提取阶段,主要解决跟踪任务在面对挑战时精度下降的问题.在孪生网络跟踪任务中,图像中的每个加权位置对互相关操作的贡献并不相同,且不同卷积特征通道通常对应不同类型的视觉模式.在复杂跟踪场景下,双重注意力机制使网络更加关注图像的重要区域,并且针对不同语境关注特定的语义属性,从而更加有效提取具有细微特征的判别性能,如图2所示.

如图2左半部分所示,本文在特征提取分支ResNet-50的最后三层引入双重注意力机制模块,该模块用于为特征图中的不同位置和通道分配不同的权重,从而强化关键区域的表达,抑制干扰区域的信息.具体实现如式(1)所示:

$$\left.\begin{aligned}\varphi(X)_i=\mathrm{Att}\left(\varphi(X)_i\right),\\ \varphi(Z)_i=\mathrm{Att}\left(\varphi(Z)_i\right).\end{aligned}\right\} \tag{1}$$

式中:$i=0,1,2$;φ为ResNet-50特征提取网络;Att为双重注意力机制模块;X为搜索区域;Z为模板信息.

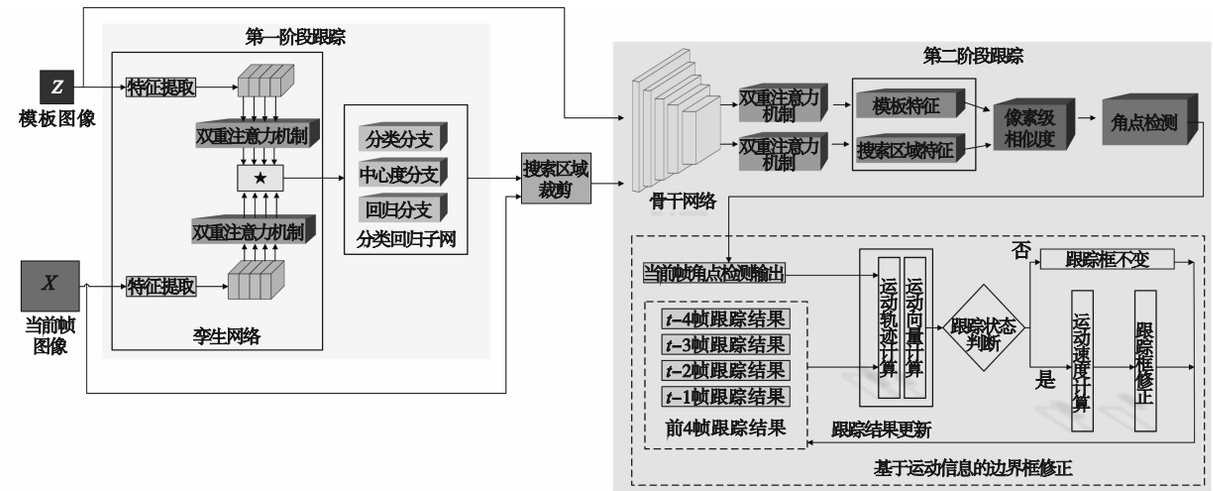

图1　两阶段跟踪算法整体架构

Fig. 1　Overall architecture of two-stage tracking algorithm

如图2右半部分所示,将ResNet-50的最后三层特征提取结果作为双重注意力机制模块的输入特征图,分别通过通道注意力层以及卷积-反卷积层得到输出特征图$F(Z)_i$后,进行互相关操作并进行特征融合.通道注意力层首先通过平均池化层对输入特征在通道维度进行压缩,接着通过卷积层对特征图的通道维度进行压缩,并结合激活函数进行归一化处理;最终生成的空间权重与原特征图逐元素相乘,生成不同权重的特征图,以赋予各通道不同的重要性.卷积-反卷积层提高了目标边界附近的注意力,使模型更加注重目标边界的特征提取.通道注意力机制模块如式(2)所示:

$$\mathrm{Att}_1\left(F(Z)_i\right)=\delta\left(f_c\left(\mathrm{AvgPool}\left(F(Z)_i\right)\right)\right). \quad (2)$$

其中:$F(Z)_i$为输入特征图;δ为Sigmoid激活函数;f_c为全连接层;$\mathrm{AvgPool}\left(F(Z)_i\right)$为平均池化后的特征图;$\mathrm{Att}_1\left(F(Z)_i\right)$为输出的通道注意力参数.

对$\varphi(X)_i,\varphi(Z)_i$进行互相关操作,并进行特征融合,如式(3)所示:

$$\left.\begin{aligned} R_i&=\varphi(X)_i\otimes\varphi(Z)_i,\\ R^*&=\mathrm{Down}\left(\mathrm{Cat}\left(R_1,R_2,R_3\right)\right).\end{aligned}\right\} \quad (3)$$

式中:⊗表示对各通道分别进行互相关运算;Cat表示在通道维度上对特征进行拼接以实现融合;Down表示通过1×1卷积实现特征压缩,以降低维度并提高后续处理效率.压缩后的响应图R^*被用作后续分类和回归子网络的输入.与原始SiamCAR方法一致,分类分支负责判断响应图中每个位置的类别信息,并结合中心度得分共同完成目标位置的判定,最终在回归分支得到相应位置的输出跟踪框.

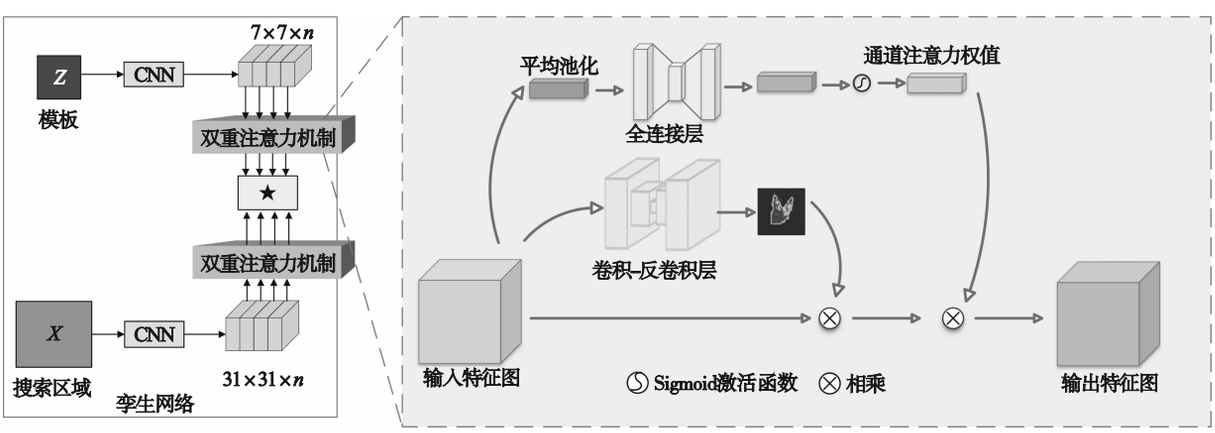

图2　加入双重注意力机制的孪生网络结构示意图

Fig. 2　Schematic diagram of Siamese network structure with dual-attention mechanism

1.2 设计引入运动信息的边界框精细化模块

在第二阶段,受多级跟踪策略的启发[9-11],本文设计了引入运动信息的边界框精细化模块,提升了跟踪器的性能.在该模块中,本文引入了像素级相似度计算[11-12],在小范围搜索区域上进行模板相似度捕捉,相比于一般的相似度提取方式,该方法使网络最大程度保留空间信息以应对复杂跟踪场景.本文设计了基于运动信息的边界框修正模块,将基于外观特征得到的跟踪位置与目标的运动轨迹信息相融合,改善跟踪漂移及跳变问题.此外,该精细化模块可独立训练,并可以直接应用于任何现有的跟踪器,不需要额外的训练.

如图3所示,引入运动信息的边界框精细化模块有两个输入分支:模板分支和搜索区域分支.两个分支采用参数共享的主干网络以及双重注意力机制进行特征提取.所提取的特征图通过相似度计算后,再经过卷积层进一步处理,最后输入到预测头中.作为边界框精细化模块,模板分支的输入为第一帧且带有边界框真值的图像;搜索区域分支的输入则为第一阶段预测结果扩展到两倍大小的同心图像区域.其中,较小的搜索区域(两倍大小)可以减少杂乱背景信息的干扰,有利于精确定位.并且较小的搜索区域降低了计算成本,因此可在增加少量延迟的情况下改进基础跟踪器.

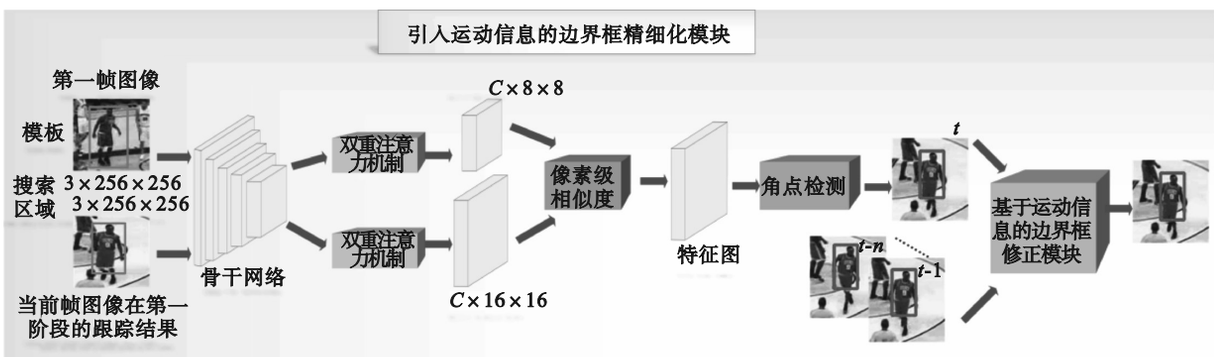

图3 引入运动信息的边界框精细化模块

Fig. 3 Bounding box refinement module with motion information

本文用pixel-wise像素级互相关操作[10]来解决以整个模板特征为内核的互相关操作导致的空间信息模糊问题[11],以获得高质量的特征表示.如图4所示,假设$Z_{\mathrm{f}}\in\mathbf{R}^{C\times H_z\times W_z}$,$X_{\mathrm{f}}\in\mathbf{R}^{C\times H_x\times W_x}$为模板和搜索区域的特征,pixel-wise互相关操作首先将Z_{f}按维度分解为n_z个子核$Z_{\mathrm{fs}}^{i}\in\mathbf{R}^{C\times 1\times 1}$,$Z_{\mathrm{fs}}=\left\{Z_{\mathrm{fs}}^{1},Z_{\mathrm{fs}}^{2},\cdots,Z_{\mathrm{fs}}^{n_z}\right\}$,$n_z=H_z\times W_z$.图中$X_{\mathrm{f}}^{(i,j)}$表示$X_{\mathrm{f}}$第$j$行第$i$列的位置.Pixel-wise互相关计算$X_{\mathrm{f}}$与空间核$Z_{\mathrm{fs}}$的相似度,得到相似度图$S$,即计算模板特征每个特征点与搜索特征的相似度.$S^{(i,j)}$中的第$m$个值表示$X_{\mathrm{f}}^{(i,j)}$与$Z_{\mathrm{f}}$的空间维度中第$m$个位置之间的相似度,如式(4)所示.与普通的互相关操作相比,pixel-wise像素级互相关操作将模板特征的每个部分作为内核,确保每个相似度图编码目标局部区域的信息,同时避免特征模糊的问题.

$$S_m^{(i,j)}=X_{\mathrm{f}}^{(i,j)}\cdot Z_{\mathrm{fs}}^{m},\quad m=1,2,\cdots,n_z. \tag{4}$$

本文在边界框精细化模块中采用角点检测作为检测头来检测边界框的左上角和右下角两个点.检测头采用4个堆叠的Conv-BN-ReLU层处理特征图,并通过一个卷积层来预测两个热力图,分别对应边界框的左上角和右下角.最后,对热力图应用soft-argmax[13],使离散热力图能够精确描述两个角点的位置.

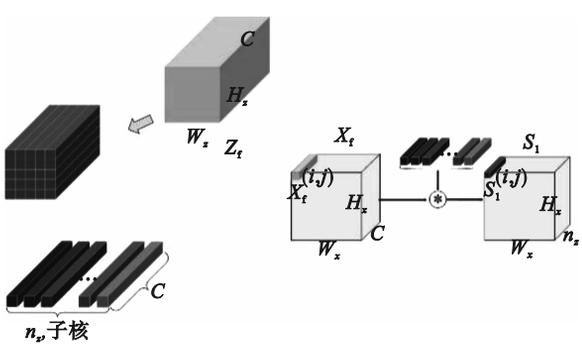

图4 像素级互相关操作

Fig. 4 Pixel-level cross-correlation operation

在基于运动信息的边界框修正模块中,通过提取近几帧的运动信息对当前帧进行修正,以得到当前帧的特征图.当跟踪目标被遮挡时,容易出现跟踪框跳变或漂移的问题,即在遮挡时刻跟踪框跳变(漂移)到邻近物体上,在遮挡物移开后跟踪框才会回到待跟踪物体上,这种跳变往往不符合物体的自然运动规律.为了利用待跟踪物体的运动路径信息,将前4帧的目标跟踪结果保存,并在基于运动信息的边界框修正模块中结合当

前帧结果进行判断,以解决因遮挡或相似物干扰等导致的跟踪失败问题.具体步骤如图5所示,首先得到连续5帧的目标运动轨迹,基于此得到运动向量图.将目标在连续n帧内的平均运动向量定义为

$$v=\frac{1}{n-1}\sum_{i=2}^{n}\left(x_i-x_{i-1},y_i-y_{i-1}\right). \tag{5}$$

从图5中的运动向量图示例可以看出,t时刻运动向量方向与平均运动向量差距过大,即向量夹角超过阈值.而若将$t-2$时刻与t时刻的运动向量相连,该运动向量符合$t-3$帧和$t-2$帧的平均运动向量,依此可以判断,$t-1$时刻发生跟踪框跳变,接下来根据$t-2$,$t-3$,$t-4$三帧的速度以及运动方向对$t-1$帧进行跟踪修正.如式(6)所示,针对跟踪漂移设置了3帧平滑滤波,取$k=t-4,t-3$,可得目标在第$t-1$帧的加权运动速度$\left(v_{t-1}^{x},v_{t-1}^{y}\right)$,其中$\left(x_k,y_k\right)$表示第$k$帧的边界框中心点,$\theta$为学习权重.如式(7),式(8)所示,用目标的运动信息$\left(v_{t-1}^{x},v_{t-1}^{y}\right)$对后续帧运动轨迹进行预测,得到位置信息$\left(x_{t-1},y_{t-1}\right)$,$\left(w_{t-1},h_{t-1}\right)$取前3帧跟踪框长、宽的平均值,即$n$=3.

$$\left.\begin{aligned}v_{k+2}^{x}&=\theta\times\left(x_{k+1}-x_k\right)+(1-\theta)\times v_k^{x},\\ v_{k+2}^{y}&=\theta\times\left(y_{k+1}-y_k\right)+(1-\theta)\times v_k^{y}.\end{aligned}\right\} \tag{6}$$

$$\left.\begin{aligned}x_{t-1}&=x_{t-2}+v_{t-1}^{x},\\ y_{t-1}&=y_{t-2}+v_{t-1}^{y}.\end{aligned}\right\} \tag{7}$$

$$\left.\begin{aligned}w_{t-1}&=\frac{1}{n}\sum_{i=1}^{n}\left(w_{t-2-(n-1)}-w_{t-2-n}\right),\\ h_{t-1}&=\frac{1}{n}\sum_{i=1}^{n}\left(h_{t-2-(n-1)}-h_{t-2-n}\right).\end{aligned}\right\} \tag{8}$$

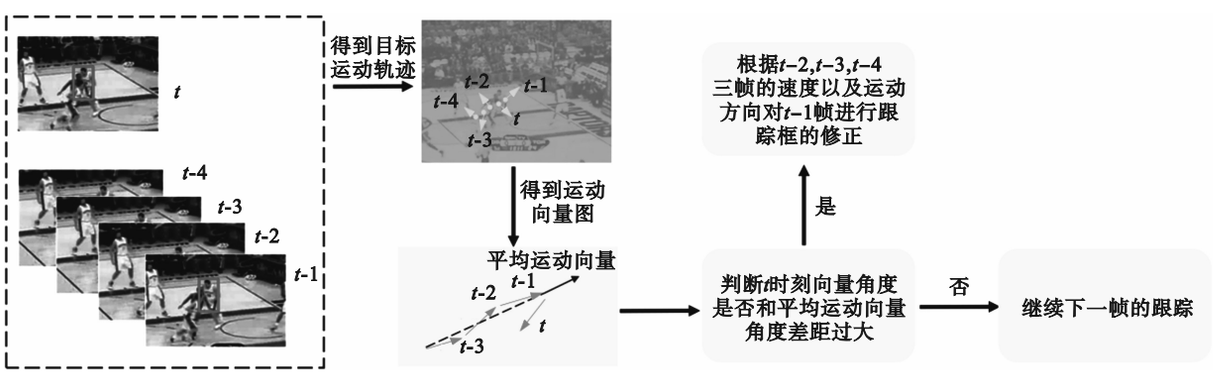

图5　基于运动信息的边界框修正模块

Fig. 5　Bounding box correction module based on motion information

1.3　损失函数

本文整体算法分为两阶段训练.跟踪器的第一阶段的损失函数与基础SiamCAR跟踪器相同,如式(9)所示,由3部分构成,分别为分类损失L_{cls}、中心度损失L_{cen}以及回归损失L_{reg},采用交叉熵损失进行分类,IoU损失进行回归.常数τ_1和τ_2为中心度损失和回归损失的权重.在模型训练时,本文经验性地设置$\tau_1=1$,$\tau_2=3$.

$$L=L_{\mathrm{cls}}+\tau_1 L_{\mathrm{cen}}+\tau_2 L_{\mathrm{reg}}. \tag{9}$$

跟踪器的第二阶段训练采用均方误差(MSE)损失函数,将预测结果转换为[最左,最上,最右,最下]格式的坐标形式,与标签值进行比较,得到MSE值,如式(10)所示,

$$L_{\mathrm{MSE}}=\frac{\sum_{i=1}^{n}\left(F(x)_i-Y_i\right)^2}{n}. \tag{10}$$

其中:$F(x)$为第二阶段预测结果;Y为标签值;n=4.

2　实验与结果分析

2.1　实验设置

本文提出的跟踪方法基于PyTorch框架,使用Python编程语言在PyCharm环境中开发,并在两张RTX 2080 Ti显卡上完成训练与测试.训练阶段选用了两个大规模数据集:GOT-10K和ImageNet.GOT-10K包含约66 GB的图像序列数据,涵盖563类不同目标,总计标注超过150万个真实边界框.特征提取部分采用了文献[14]中调整后的ResNet-50作为主干网络.该网络首先在ImageNet数据集上完成预训练,随后在整体训练过程中加载预训练权重进行初始化.模型共训练50个epochs,采用随机梯度下降(SGD)优化策略,初始学习率设为0.001.对于前10个epochs,骨干网络的参数在训练时被冻结.对于最后40个epochs,ResNet-50的最后3个模块被解冻并一起训练.

本文选用了在目标跟踪研究中具有广泛代表性的数据集OTB100评价算法性能.该数据集涵盖100种不同目标对象,并对其中98条视频序列进行了精确标注.此外,OTB100还包括11类典型挑战因素:快速运动(fast motion,FM)、形变(deformation,DEF)、遮挡(occlusion,OCC)、光照变化(illumination variation,IV)、目标移出视野(out-of-view,OV)、图像低分辨率(low resolution, LR)、尺度变化(scale variation, SV)、外平面旋转(out-of-plane rotation, OPR)、背景干扰(background clutters, BC)、平面内旋转(in-plane rotation, IPR)以及运动模糊(motion blur, MB).每段视频至少包含一种或多种干扰属性.性能评估采用一次性评估(one-pass evaluation, OPE)方法,即将算法预测结果与人工标注的真实轨迹进行对比,以此计算跟踪的成功率和精度.

2.2 消融实验

如表1所示,本文设计了多组消融实验,通过对比验证所提模块为跟踪性能带来的提升.由表1可得,基线算法SiamCAR的成功率为66.3%,精度为88.0%;在特征提取阶段使用双重注意力机制后,算法在成功率和精度上分别提升1.5%和2.1%;在基础跟踪器中单独加入带有运动信息的边界框细化模块(第3行),其成功率和精度与基线相比分别提升了3.8%和0.8%;引入双重注意力机制和边界框精细化模块的最终算法(第4行)与基线算法相比,成功率提升4.6%,精度提升2.8%,证明了所提模块的有效性.虽然本文算法的帧率相较基准算法SiamCAR有所下降,但以较小帧率下降为代价获取更高的跟踪精度是值得的,且改进后仍能满足跟踪的实时性要求.

表1 在OTB100基准上的消融实验

Table 1 Ablation experiment on OTB100 benchmark

SiamCAR	双重注意力机制	引入运动信息的边界框精细化模块	成功率/%	精度/%	帧率/帧·s^{-1}
√			66.3	88.0	51.25
√	√		67.8	90.1	47.6
√		√	70.1	88.8	33.8
√	√	√	70.9	90.8	32.3

2.3 定量分析

为了全面验证本文所提出算法的有效性,将其与多种具有代表性的视觉跟踪方法(包括SiamBAN,SiamRPN++,SiamCAR,SiamFC++和SiamRPN)在OTB100数据集上进行了对比实验,测试结果如图6所示.图中左侧为精度曲线图,右侧为成功率曲线图.从图6可以看出,本文算法在该数据集上的成功率达70.9%,精度为90.8%,在所选的多数对比方法中均取得了更优的表现.与基线算法SiamCAR相比,成功率提高了4.6%,精度提升了2.8%.从定量结果来看,所提出的跟踪器具备良好的跟踪性能.

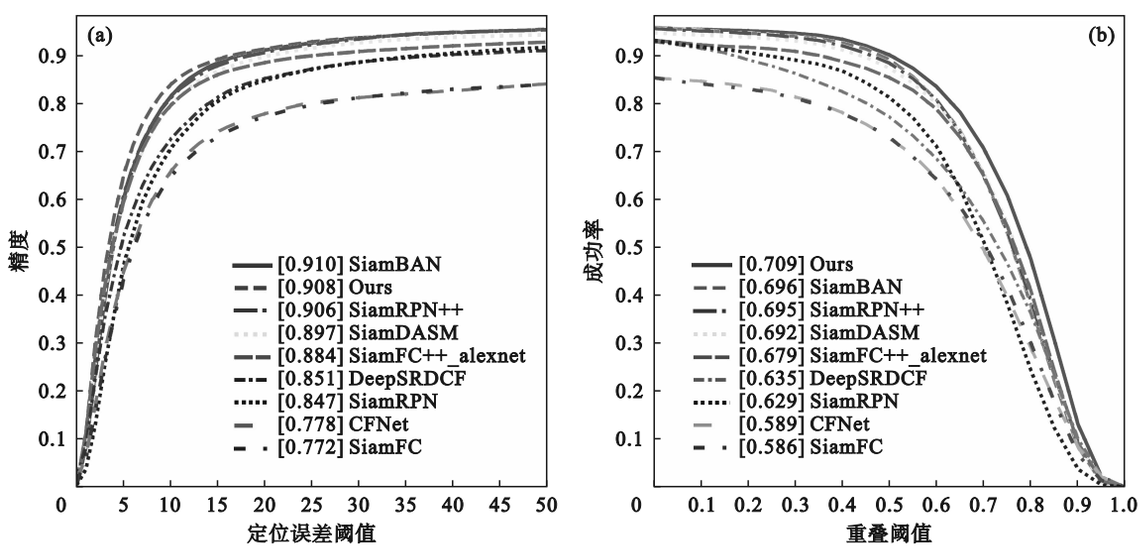

图6 在OTB100上多种算法的精度和成功率对比

Fig. 6 Comparison of precision and success rate of multiple algorithms on OTB100

(a)—精度;(b)—成功率.

此外,为深入评估算法在应对不同挑战因素下的鲁棒性,表2展示了本文方法与其他主流算法在OTB100数据集上多种属性下的成功率对比情况.由表2可见,本文算法在多数挑战条件下均

保持领先,特别是在遮挡、形变、运动模糊、快速运动、出视野、背景干扰和图像低分辨率等关键挑战上表现最优.这一优势主要得益于所引入的具有运动信息感知能力的边界框细化模块,有效提升了对目标相似度的捕捉能力;同时,双重注意力机制也增强了模型对关键区域和通道的响应能力,从而显著提升了模型在复杂场景下的整体跟踪性能.

表2　OTB100数据集不同属性下算法的跟踪成功率对比结果

Table 2　Comparison of tracking success rates of algorithms under different attributes on OTB100 dataset

算法	IV	OPR	OCC	DEF	MB	FM	IPR	OV	BC	LR	SV
本文算法	0. 723	0. 695	**0. 668**	**0. 681**	**0. 738**	**0. 723**	0. 713	**0. 677**	**0. 696**	**0. 721**	**0. 722**
SiamBAN	**0. 724**	0. 687	0. 648	0. 662	0. 698	0. 687	**0. 717**	0. 640	0. 680	0. 719	0. 693
SiamRPN++	0. 714	0. 683	0. 656	0. 667	0. 692	0. 678	0. 700	0. 624	0. 672	0. 720	0. 694
SiamDASM	0. 705	**0. 697**	0. 647	0. 657	0. 716	0. 691	0. 709	0. 625	0. 659	0. 730	0. 704
SiamFC++	0. 695	0. 660	0. 613	0. 646	0. 669	0. 664	0. 695	0. 582	0. 627	0. 699	0. 675
SiamRPN	0. 651	0. 628	0. 588	0. 620	0. 625	0. 602	0. 631	0. 544	0. 594	0. 642	0. 618
SiamFC	0. 572	0. 561	0. 549	0. 512	0. 554	0. 571	0. 559	0. 509	0. 527	0. 618	0. 556
SiamCAR	0. 685	0. 653	0. 619	0. 620	0. 698	0. 684	0. 689	0. 609	0. 617	0. 686	0. 666

2.4　定性分析

为更直观地体现本文算法在实际应用中的跟踪能力,从OTB100数据集中选取了四段典型视频序列进行效果评估.Basketball序列包含光照变化(IV)、遮挡(OCC)、形变(DEF)、外平面旋转(OPR)和背景干扰(BC)等挑战属性;Jump序列涉及尺度变化(SV)、遮挡(OCC)、形变(DEF)、运动模糊(MB)、平面内旋转(IPR)和外平面旋转(OPR);DragonBaby序列面临尺度变化(SV)、遮挡(OCC)、运动模糊(MB)、快速移动(FM)、平面内旋转(IPR)、外平面旋转(OPR)以及出视野(OV);Human3序列具有尺度变化(SV)、遮挡(OCC)、形变(DEF)、外平面旋转(OPR)和背景干扰(BC).图7展示了本文算法与另外6种主流视觉跟踪方法在上述视频序列中的跟踪效果对比,结果显示所提出的方法在多种复杂挑战下依然表现出较强的鲁棒性.

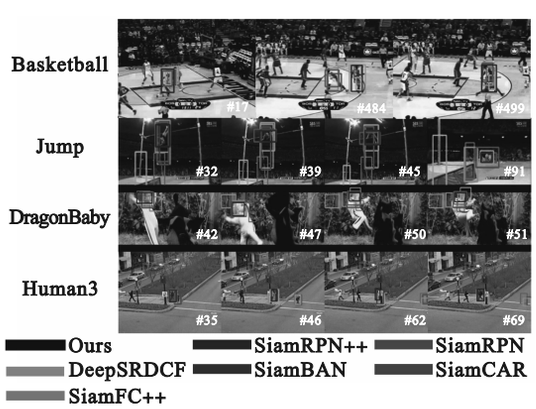

图7　OTB100部分视频序列跟踪结果

Fig. 7　Tracking results of partial video sequences in OTB100

1)背景干扰.在Basketball视频序列中包含多个与目标球员外观极为相似的背景球员,易造成混淆.在该场景下,SiamFC++等跟踪算法在目标识别方面表现不够理想,难以有效区分目标与干扰背景,最终导致跟踪失败.由于相似物体距目标物体非常近,在第17帧中其他框都不同程度地发生了跟踪跳变以及跟踪漂移问题,第484帧中一些跟踪器已经发生了跟踪跳变以及漂移的问题;第499帧的漂移现象更为明显.本文算法不仅未发生跳变和漂移,在能够成功跟踪的跟踪器中其跟踪框也是最精准的.这主要得益于本文引入的双重注意力机制与边界框精细化模块,能够更充分地建模目标特征,即使在背景复杂、干扰较多的场景中,也能准确聚焦于目标区域,从而实现更加稳定可靠的跟踪效果.时序信息的引入成功解决了跟踪跳变以及漂移问题.

2)形变.以视频序列Jump为例,序列中的撑杆跳运动员在从跃起到空中翻转再落地的过程中发生了形变,这要求跟踪算法在目标跟踪任务中具备更强的特征提取能力.具体而言,在跟踪序列的第32帧、第39帧、第45帧以及第91帧中,可以观察到本文算法在目标发生形变时能够实现较为准确的跟踪,并且能够快速响应目标形变.相比之下,其他算法则出现了目标丢失、跟踪响应缓慢或一定程度的跟踪漂移等问题.

3)运动模糊.以视频序列DragonBaby为例,该序列中的小孩在与玩偶互动时经常出现目标模糊的情况,这给跟踪模型的泛化能力提出了较高要求.具体而言,在跟踪序列的第42帧、第47帧、第50帧和第51帧中,当目标出现模糊时,本

文算法能够准确地定位目标,而其他跟踪算法则出现了跟踪丢失、跟踪偏移等问题.

4)遮挡.以视频序列Human3为例,由于目标在运动过程中受到物体遮挡,容易在跟踪过程中丢失跟踪目标.具体而言,在跟踪序列的第35帧、第46帧、第62帧以及第69帧中,由于物体阻挡和其他相似目标的干扰,其他跟踪器已经无法继续准确地跟踪目标,导致目标丢失或跟踪偏移.然而,本文算法在这些情况下仍能保持良好的跟踪能力.

3 结 语

本文针对形变、运动模糊、遮挡以及背景干扰等情况下容易出现的跟踪框精度下降以及跟踪漂移、跳变等问题,提出了一种结合运动信息和双重注意力机制的两阶段单目标跟踪算法.通过引入具有双重注意力机制的SiamCAR跟踪器,在第一阶段对当前帧的目标进行粗定位;在第二阶段利用像素级相似度运算构建边界框精细化模块,学习目标的细微特征以提升跟踪框精度.最终将基于外观特征得到的跟踪位置与目标的运动轨迹信息相融合,改善跟踪漂移问题.实验表明,本文算法与基准算法相比具有更好的跟踪效果,较基准算法其成功率和精度分别提高了4.6%和2.8%;在OTB100不同属性成功率对比结果中,由于运动轨迹信息的引入,本文算法在背景干扰场景下的成功率达到69.6%,增强了对跟踪漂移情况的适应性.此外,本文提出的两阶段跟踪框架适用于其他所有单目标跟踪器,即可以将本文算法的第二阶段应用于其他单目标跟踪算法,无需重新训练即可提高精度.

参考文献:

[1] 魏颖, 徐楚翘, 刁兆富, 等. 基于生成对抗网络的多目标行人跟踪算法[J]. 东北大学学报(自然科学版), 2020, 41(12): 1673-1679, 1720.
(Wei Ying, Xu Chu-qiao, Diao Zhao-fu, et al. A multi-target pedestrian tracking algorithm based on generated adversarial network[J]. *Journal of Northeastern University* (*Natural Science*), 2020, 41(12): 1673-1679, 1720.)

[2] 高文, 朱明, 贺柏根, 等. 目标跟踪技术综述[J]. 中国光学, 2014, 7(3): 365.
(Gao Wen, Zhu Ming, He Bai-gen, et al. Overview of target tracking technology[J]. *Chinese Optics*, 2014, 7(3): 365.)

[3] Li B, Yan J J, Wu W, et al. High performance visual tracking with Siamese region proposal network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 8971-8980.

[4] 郑艳, 赵佳旭, 边杰. 基于SiamBAN跟踪器改进的目标跟踪算法[J]. 东北大学学报(自然科学版), 2023, 44(9): 1227-1233.
(Zheng Yan, Zhao Jia-xu, Bian Jie. Improved object tracking algorithm based on SiamBAN tracker[J]. *Journal of Northeastern University* (*Natural Science*), 2023, 44(9): 1227-1233.)

[5] Guo D Y, Wang J, Cui Y, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, 2020: 6268-6276.

[6] Chen Z D, Zhong B N, Li G R, et al. SiamBAN: target-aware tracking with Siamese box adaptive network[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(4): 5158-5173.

[7] Dong B, Zhuge M, Wang Y, et al. Accurate camouflaged object detection via mixture convolution and interactive fusion[J]. *arXiv Preprint arXiv*, 2101: 05687.

[8] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[C]// Computer Vision-ECCV 2016 Workshops. Cham: Springer International Publishing, 2016: 850-865.

[9] Wang G T, Luo C, Xiong Z W, et al. SPM-tracker: series-parallel matching for real-time visual object tracking[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 3643-3652.

[10] Fan H, Ling H B. Siamese cascaded region proposal networks for real-time visual tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 7944-7953.

[11] Yan B, Zhang X Y, Wang D, et al. Alpha-refine: boosting tracking performance by precise bounding box estimation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, 2021: 5289-5298.

[12] Wang Z Q, Xu J, Liu L, et al. RANet: ranking attention network for fast video object segmentation[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, 2019: 3977-3986.

[13] Luvizon D C, Tabia H, Picard D. Human pose regression by combining indirect part detection and contextual information[J]. *Computers & Graphics*, 2019, 85: 15-22.

[14] Li B, Wu W, Wang Q, et al. SiamRPN++: evolution of Siamese visual tracking with very deep networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 4282-4291.