控制与决策
Control and Decision

无锚双注意力孪生网络的视觉跟踪

郭文, 梁卜文, 丁昕苗

引用本文:
郭文,梁卜文,丁昕苗. 无锚双注意力孪生网络的视觉跟踪[J]. 控制与决策, 2024, 39(2): 633−640.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1303

无锚双注意力孪生网络的视觉跟踪

郭 文[1,2], 梁卜文[1], 丁昕苗[2†]

(1. 山东工商学院 计算机科学与技术学院，山东 烟台 264005；
2. 山东工商学院 信息与电子工程学院，山东 烟台 264005)

摘 要: 针对跟踪过程中因光照变化、快速运动及尺度变化等造成的角点定位精准度下降问题,受SiamCAR的跟踪框架启发提出一种无锚双注意力孪生网络的视觉跟踪算法. 首先,算法的主干网络采用ResNet-50并结合增强多层融合特征图进行特征提取,充分利用网络浅层特征的定位信息和深层次的语义信息,提高算法对目标特征的语义理解能力;然后,构建混合注意力模块缓解无锚跟踪器角点定位不准确问题,提高算法的跟踪准确性和定位精度;最后,在GOT10K、UAV123、LaSOT等数据集上进行广泛实验,并与当前的先进跟踪器进行比较,该算法可以较好地抵抗光照变化、快速运动及尺度变化等多种复杂因素带来的影响,同时,在多项评测指标上获得了良好的跟踪性能.

关键词: 无锚；注意力特征图；双注意力；孪生网络；视觉跟踪

中图分类号: TP391 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1303

引用格式: 郭文,梁卜文,丁昕苗. 无锚双注意力孪生网络的视觉跟踪[J]. 控制与决策, 2024, 39(2): 633-640.

Dual attention Siamese network with anchor free for visual tracking

GUO Wen[1,2], *LIANG Bo-wen*[1], *DING Xin-miao*[2†]

(1. School of Computer Science and Technology，Shandong Technology and Business University，Yantai 264005，China；
2. School of Information and Electronic Engineering，Shandong Technology and Business University，Yantai 264005，China)

Abstract: Aiming at the problem of decreased corner positioning accuracy caused by light changes, fast motion and scale changes in the tracking process, we propose a visual tracking algorithm motivated by the framework of SiamCAR. Firstly, the research method uses improved ResNet-50 as a feature extraction backbone network and combines with enhanced multi-layer fusion feature map to extract feature, which makes full use of the location information of shallow features and deep semantic information of the network, and improves the semantic understanding ability of the algorithm to target features. Secondly, a hybrid attention block is constructed to alleviate the inaccurate corner location of the anchor tracker, which improves the tracking accuracy and positioning accuracy of the algorithm. Finally, extensive experiments are carried out on GOT10K, UAV123, LaSOT and other datasets. Besides, compared with current advanced trackers, the proposed algorithm can better resist the influence of various complex factors such as illumination variation, rapid motion and scale variation, at the same time, obtain good tracking performance on a number of evaluation indicators.

Keywords: anchor free；attention feature map；dual attention；Siamese network；visual tracking

0 引 言

视觉跟踪技术是一项计算机视觉领域极具实用价值的任务. 随着人机交互、无人驾驶、智能安防等热度[1]的提升,视觉跟踪已成为该领域内的研究热点. 视觉跟踪技术取得了较大突破,但是现实场景中的光照变化、背景杂波、遮挡、低分辨力、尺度变化、外观变化和快速运动[2]仍然是视觉跟踪面临的挑战.

在视觉跟踪领域,当前的两种主流方法分别是基于相关滤波[3]和基于深度学习[4]的跟踪方法. 相关滤波是一种判别式机器学习算法[5],其主要原理是判别两个相关信号的输出响应是否大于不相关信号的输出响应. 在视觉跟踪过程中,使用目标模板训练的滤波器对后续视频帧进行滤波处理,寻找响应图中最大输出响应,作为当前帧的目标位置. 因此,基

收稿日期: 2022-07-21；**录用日期:** 2022-10-21.

基金项目: 国家自然科学基金项目 (62072286, 61876100, 61572296)；山东省研究生教育创新计划项目 (SDYAL21211).

†通讯作者. E-mail: dingxinmiao@126.com.

*本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

于相关滤波的视觉跟踪即转化为对后续搜索帧进行相关滤波的过程. 在初始的相关滤波算法中主要通过设计底层的手工特征去表示目标,如采用单一灰度特征的MOSSE (minimum output sum of squared error filter)[6]、CSK (exploiting the circulant structure of tracking-by-detection with kernels)[7]和采用方向梯度直方图(HOG)特征的KCF (kernel correlation filter)[8],但在实际的应用场景中,单一的手工特征很难应用于实时的复杂场景,而深度神经网络[9]以其良好的目标表征能力和抗干扰能力在计算机视觉领域引起了广泛的关注.

深度学习算法在视觉跟踪上的应用很多是围绕孪生网络[10]展开的. 研究学者从该网络得到应用起一直致力于在特征提取、分类设计、数据增强、边界框回归等方面设计更鲁棒的视觉跟踪算法. 孪生网络不同于VGGNet (visual geometry group net)[11]等神经网络,它将视觉跟踪任务定义为目标匹配问题,主体是由两个平行分支组成,通过判定目标模板帧与后续搜索帧之间的相似度确定其位置. 在视觉跟踪领域首次使用孪生网络的算法是SINT (Siamese instance search for tracking)[12],为了提高跟踪的鲁棒性,该算法使用特殊的样本采样方法,即半径采样方法,使得跟踪精度达到了当时先进的水平. SiamFC (fully-convolutional Siamese networks for object tracking)[13]将孪生网络跟踪方法推向了前沿,该算法是一种端到端训练的全卷积孪生网络算法,以其良好的速度和准确性在传统相关滤波算法中脱颖而出,它的出现使得利用卷积神经网络的深度算法可以与传统的相关滤波算法抗衡. 但另一方面,SiamFC使用浅层网络,搜索区域太广,并且固定5种不同大小尺度的锚框,在面对复杂环境(快速运动、光照变化、遮挡)时,容易出现跟踪漂移和目标丢失问题. 随后,受目标检测中预定义锚框的启发,跟踪领域学者尝试在跟踪领域融合目标检测技术. 引入经典的目标检测算法Faster R-CNN[14]中区域建议网络(RPN),提出了基于锚框的SiamRPN (high performance visual tracking with Siamese region proposal network)[15]跟踪算法,将视觉跟踪作为单样本检测问题. 通过滑动窗口的方式,设定不同比例尺寸的锚框,根据生成的锚框以及网络特征,直接预测目标中心的位置,优化了目标尺度变化问题. SiamMASK (fast online object tracking and segmentation)[16]为了进一步提升RPN对目标包围框的精准度,增加了分割分支,从而获得更贴合的目标包围框. DaSiamRPN (distractor-aware Siamese networks for visual object tracking)[17]针对RPN正负样本不均衡问题,通过数据增广,增加正样本对,扩大训练集的种类,在一定程度上提高了网络的泛化能力.

无论是添加检测头还是增加正样本采样,目的是保证参数量的同时,提升视觉跟踪的鲁棒性. 受图像分类、目标检测领域注意力的影响,文献[18]以ResNet网络为主干,在主干网络之后增加了通道注意力,通过跨通道学习,提升了网络模型对首帧的利用率; SiamCAR (Siamese fully convolutional classification and regression for visual tracking)[19]则在通道降维的基础上使用Anchor-free机制减少了整体的参数量,并取得了良好的结果. 上述方法虽然可以较好地解决深度视觉跟踪的问题,但是仍然存在以下问题限制了其性能的提升:

1) 当前算法对模板特征进行了多类型的提取,但是随着网络深度的增加,基于CNN网络架构中的参数量也越来越大,增加了算法的复杂度,从而给网络的训练和后期的推理带来负面影响.

2) 无论是传统卷积神经网络提取特征,还是通过现有注意力感知的特征,通常使用最大池化操作或平均池化[20]操作建立像素点与全局的关系,在处理过程中容易将区分度不明显的区域舍去,这减少了跟踪过程中有用信息的使用,从而损失鲁棒性.

3) 当前孪生网络在进行特征融合后直接预测锚点位置,缺乏特征语义信息的准确表达、对锚点的精准定位以及对锚框的准确回归,这使得跟踪过程中容易出现跟踪漂移现象.

为了解决上述问题,本文在SiamCAR的基础上提出一种创新的无锚双注意力孪生网络跟踪框架,如图1所示. 该框架为了避免繁杂的超参数设计,在沿用孪生网络的基础上放弃锚框,整个网络架构主要由3个子网络组成:用于逐像素特征提取的主干网络、特征增强的混合注意力网络和分类头回归网络.

本文工作的主要贡献如下:

1) 设计一种主干网络语义增强模块,在没有增加主干网络参数的同时,提升主干网络的表征能力,可有效地提升网络对弱目标跟踪的鲁棒性,增强网络跟踪的抗干扰能力.

2) 设计一个混合特征增强网络模块,有效地强化特定类别响应度高的特征区域,弱化无关特征区域,提高角点回归框估计的准确度,提升算法的鲁棒性.

3) 改进多层融合方式,使得融合特征更有利于目标描述的层次表达,提升网络对跟踪过程中尺度变化的适应能力,所提出的方法在LaSOT[21]、UAV123[22]、GOT-10K[23]等数据集的跟踪评测中在时间和准确度上获得了优良的性能.

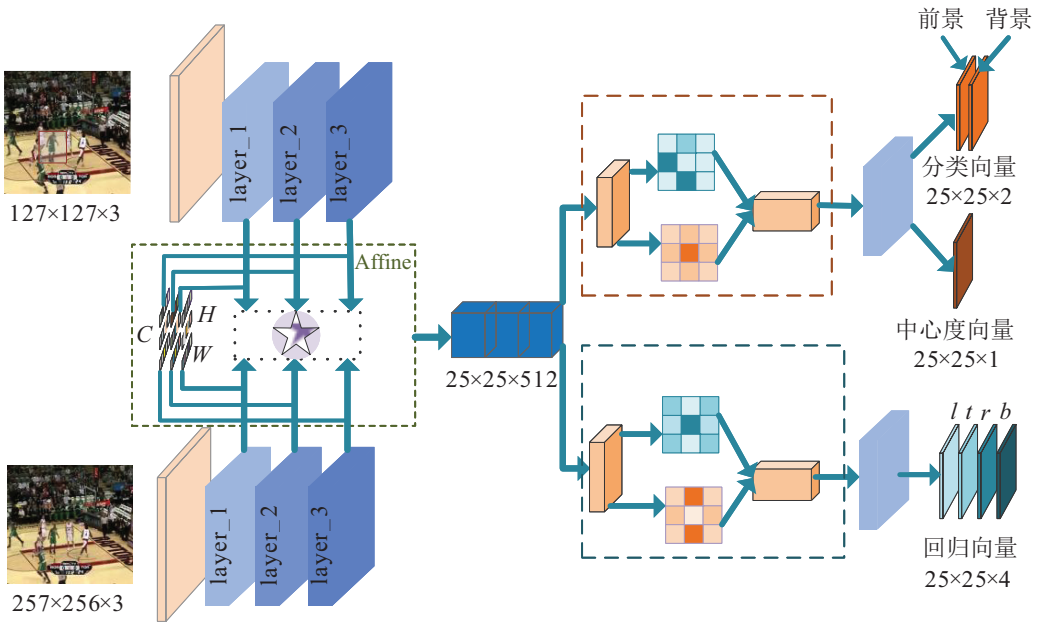

图 1　无锚双注意力孪生网络框架

1　无锚双注意力孪生网络视觉跟踪框架

1.1　特征提取网络

特征信息语义表达是否充分,决定着跟踪结果的好坏. 随着网络层数的加深,网络模型的表达能力会更强,但是随着网络的加深会引起梯度的消失或爆炸. ResNet50网络利用1×1卷积残差连接结构,在增加网络深度的同时有效缓解了神经网络梯度消失和模型退化等问题,为此,使用经典的深度卷积网络ResNet50作为模板分支和搜索分支特征提取的主干网络. 但是,随着每层步长的叠加,特征图的大小成反比例增长. 为了保留更多的目标信息,感知目标图像立体信息,决定将倒数第3层、第4层的步长缩减,另外,为了在相同参数量的情况下获得更多的全局信息,通过使用空洞间隔为1的空洞卷积替换原有3×3卷积. 在以往的实验中发现,直接进行互相关操作,对于弱目标特征会造成大部分特征的损失. 为了更好地利用特征图的有效信息,采用空间感知模块对主干网络提取的特征图逐像素分配唯一权重. 考虑主干网络的参数量,受文献[24]启发,抛弃空间注意力与通道注意力的并行或串行组合,而是将注意力表达方式从二维扩展到三维,通过测量一个神经元与其他神经元之间的线性可分,使得特征向量与背景向量分离. 为此为每个神经元定义一个能量函数,保证每个神经元权重的唯一性,即

$$e_t(w_t,b_t,y,)=(y_t-\hat{t})^2+\frac{1}{M-1}\sum_{i=1}^{M-1}(y_0-\hat{x}_i)^2. \tag{1}$$

其中:$\hat{t}=w_tt+b_t$和$\hat{x}_i=w_tx_i+b_t$是t和x_i的线性变换,t和x_i是输入特征的单个通道中定位的目标神经元和其他神经元,i是空间维度上的索引,$M=H\times W$是该通道上神经元的数量,w_t和b_t是变换权重和偏移量,并且式(1)中全部是标量. 通过最小化式(1)找到目标神经元与同一通道中其他神经元之间的线性可分性. 随后按照注意力机制的定义需要对特征进行增强处理. 为了适应视觉跟踪任务中对弱目标跟踪的鲁棒性,在通道维度使用最大池化后的特征与增强特征自适应融合,通过给不同神经元一个明确的权重,大幅度提高目标分类和检测精度.

1.2　多层特征融合的相似度级联方法

不同深度的特征通道往往具有不同的语义信息,浅层的特征通道往往聚焦于目标物体的外观形状、颜色、纹理等手工特征上,这些信息的有效利用对于目标的定位有着至关重要的作用,但是浅层特征也有弊端,不能包含丰富的语义信息,对于各个节点之间的内在联系不能充分表达,而深层特征却可以弥补这一不足. 同时,随着网络层数的加深、原有主干网络中步长的叠加,特征图分辨率越低,目标物体的语义信息得到越充分的挖掘,这对目标的运动状态、模糊场景中的鲁棒跟踪是有增益的. 为此,抛弃只使用网络的最后一层特征,改用主干网络中最后3个残差模块的输出特征,获取不同层次的语义信息,其次每个通道可以看作是目标信息的一种特征. 为此,在通道维度执行互相关操作. 具体而言,数据经过预处理操作,将模板帧和搜索帧分别输送到主干网络,为了方便理解,将输入模板帧记为Z,搜索图像表示为X,得

到相应网络输出的特征图$\varphi(z),\varphi(x)$,为了找到嵌入在语义空间中目标对象位置信息和尺度信息,对两个特征图在通道维度执行互相关操作,得到目标深度响应图,其计算方式为

$$F(\text{out})=\varphi_k(z)*\varphi_k(x). \tag{2}$$

其中:$*$是特征图在通道上的互相关操作,k是主干网络中使用的层数,out是输出的互相关特征图,F是输出后的特征图通过1×1卷积降维操作. 模板帧和搜索帧在经过神经网络处理后,原有的语义信息会有浮点级别的损失,此时,若进行简单的拼接或者线性的相加会对有用信息造成浪费. 对此,提出对信息利用率更高的Affine模块. 本文认为每一层分支输出特征图都是带有目标不同语义信息的图片,且大胆假设是经过弱化后的目标信息. Affine模块是独立地应用于特征矩阵X_k每一列的仿射运算,应用于每个注意力模块与原始特征图融合的过程中,类似于“层归一化”,该算子不仅使训练更加稳定,而且通过算子的调节,模块的自适应融合避免了一些列超参设计. 其次通过高层语义特征和低层语义特征的增强,更多有效语义信息得到保存,尤其是在解决复杂场景中的光照、快速移动上有着实质性收益,其计算方式为

$$\text{Affine}_{\alpha,\beta}(X_k)=\alpha\text{Diag}(X)+\beta. \tag{3}$$

其中:α, β是可训练参数,$\text{Diag}(X)$是经过逐像素特征增强特征图. 此操作按照元素重新分配实现逐像素点的运算,用该方法代替层归一化,且与层归一化相比在进行推理运算时基本不增加时间成本. $\{R^q, R^k, K^v\}\in\boldsymbol{R}^{C\times H\times W}$是卷积神经网络运算得到的特征图.

1.3 双重注意力增强模块

相比于SiamFC、SiamRPN的主干网络,ResNet网络具有更深层次语义信息,参数量也得到了大幅增长. 为轻量化网络架构,将融合后的特征图通道统一降维到255. 通过降维可以显著降低参数量,保持架构的轻量化. 在SiamCAR跟踪算法中,降维后的特征图经过卷积模块,送入分类分支和回归分支,进行目标中心点的确定以及包围盒的回归. 由于卷积运算将导致局部感受野的偏见,对应于具有相同标签像素的特征会有一些差异,这些差异导致了类内不一致,从而影响了识别的准确性. 另外,降维后的特征图减少计算量的同时也会将更多的特征进行富集,这对后续的目标确定和包围框的回归是不友好的. 受文献[18]启发,本文通过引入双重混合注意力模块逐像素捕获更加精细的语义信息,以更好的方式表示特征.

1.3.1 空间注意力模块

对于场景理解,具有判别力的特征表示是关键的,可以通过获取全局上下文信息进行判定. 但是,传统的全卷积网络产生的局部特征可能会导致锚点的错误分类. 为了能够捕获特征图上任意两点之间的位置关系,使得目标回归更加精准,使用位置注意力模块通过编码更广范围的语义信息到局部感受野中以增强特征图的表达能力. 首先,将融合特征通过3个不同的卷积层生成3个新的特征图表示为$\{R^q, R^k, K^v\}\in\boldsymbol{R}^{C\times H\times W}$,通过$R^q$、$R^k$的相似度图的矩阵运算生成空间注意矩阵,使得特征图中任意两个像素之间的空间关系得到映射,将特征图大小转化为$\{R^q, R^k\}\in\boldsymbol{R}^{C\times(H\times W)}$,随后通过使用softmax函数对这些权重进行归一化,这里的权重对应于两个位置的相似度,从而达到获取全局位置信息的作用,此时需要将权重和相应的R^v进行矩阵乘法运算,最后将原特征图与注意力权重矩阵自适应融合,具体过程如图2所示.

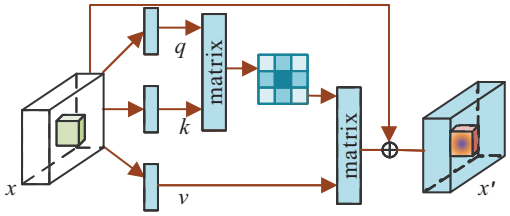

图 2 空间注意力

1.3.2 通道注意力模块

高层特征通过通道映射被认为携带不同特征的语义信息,并且不同的特征语义信息彼此呼应、彼此关联. 为此利用通道映射之间的相互依赖关系,建模特征映射之间的关系,并改进特定语义信息的特征表示,具体过程如图3所示.

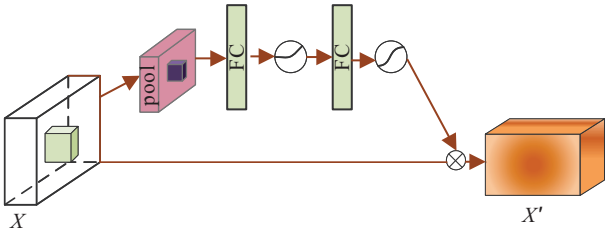

图 3 通道注意力

首先将融合特征图$H\times W\times C$,通过自适应池化压缩到$1\times1\times C$大小的通道向量,即让全局空间信息压缩到通道描述符中,通过全连接层降低维度,随后再使用全连接层恢复维度,最后再进行SIG (sigmoid)激活将特征向量重新映射,得到想要的通道维度权重矩阵,再与原特征图相乘得到最后的通道注意力模块.

1.4 分类和回归网络

深度互相关目标响应图在搜索区域图中都有唯一的位置进行一一映射. 基于RPN网络算法通过滑动窗口的方式依次寻找响应图中对应搜索区域的最大位置,并以该锚点作为手动设计锚框的中心. 本文算法与此不同的是采用无锚框机制的分类回归网络,直接通过角点和中心点的计算来确定目标信息的分类和回归,提高检测速度的同时避免了一系列超参数设计.

如图1所示,传统的视觉跟踪任务被分解为目标响应的分类与回归任务. 分类任务是通过前、后背景的分离预测该目标位置的类别;回归任务是通过优化边界区域来获得精准的目标边界框,从而获得后续帧中目标的详细信息. 前、后背景的分离是通过判定经过逐像素特征强化之后特征图上的位置(i,j)映射回搜索区域的具体位置(x,y)是否落在目标真实的边界框内来实现的,如果在真实边界框内则判定为前景,否则判定为背景,因此分类分支特征图中用一个二维向量代表前、后背景的分数. 回归分支是通过判定前景像素点到真实边界框4条边的距离,得到后续帧中特征图的大小,因此回归特征图上每个点输出一个四维向量,具体计算方式为

$$\begin{aligned} d_l &= x - x_0,\\ d_t &= y - y_0,\\ d_r &= x_1 - x,\\ d_b &= y_1 - y. \end{aligned} \tag{4}$$

其中:(d_l,d_t,d_r,d_b)是像素点到边界框4条边的距离,(x_0,y_0)和(x_1,y_1)是真实边界框的左上角和右下角坐标. 通过实验发现,使用上述判别方法与基于锚框的方法在性能上还是存在一些差距的,主要原因是远离目标中心点位置的锚点会产生低质量的预测边界框,从而降低跟踪鲁棒性. 为了描述前景样本点到目标标注中心点的距离,本文通过添加平行分支-中心度分支降低异常值. 该分支会根据距离中心点远近计算对应位置的中心度得分$C(i,j)$,从而给每一个位置点赋予相应的权重. 具体运算方式为

$$C(i,j) = I * \sqrt{\frac{\min(d_l,d_t)}{\max(d_l,d_t)} \times \frac{\min(d_t,d_b)}{\max(d_t,d_b)}}. \tag{5}$$

其中:I表示是否落在真实边界框内,若属于首帧标注的真实框内则赋值为0,否则标注为1. 另外考虑到数据集中搜索区域与目标区域所占比率不大,不存在样本不均衡问题,为此对于分类损失函数选用二次交叉熵函数,而对于搜索区域回归任务则采用IoU损失函数,最终将多任务总损失函数定义为

$$L = L_{\rm cls} + \lambda_1 L_{\rm cen} + \lambda_2 L_{\rm reg}. \tag{6}$$

其中:$L_{\rm cls}$是交叉熵损失函数;$L_{\rm cen}$是中心度损失函数;$L_{\rm reg}$是IoU损失函数;λ_1,λ_2是相应的超参数权重,在实验过程中分别设置为1和3.

2 实验结果与分析

2.1 实验细节

本文实验是在PyTorch框架下用Python语言编写,GPU为Tesla V100,主干网络采用ImageNet预训练过后的ResNet50,以往的实验表明,该预训练参数对跟踪任务是一个很好的初始化. 在训练过程中遵循端到端的训练,批次大小设置为80,训练20个周期,前5个周期使用初始化学习率0.001修改到0.005,在后15个周期使用指数衰减方法学习率从0.005到0.000 05,优化器选用梯度随机下降法(SGD). 在整个训练过程中分为两个阶段,前10个周期冻结主干参数,使用主干网络的预训练参数,训练除主干以外的模型参数. 在第11个周期以及之后解冻主干参数,训练模型中所有的参数. 本文的训练数据集为COCO[25]、ImageNet Det、ImageNet VID[26]、YouTube-BoundingBoxes[27].

2.2 定量实验分析

本文在UAV123、GOT-10 k数据集上进行了短时单目标跟踪实验,在LaSOT数据集上评测长时跟踪的效果. 长时跟踪的目标可能会离开视野或长时间处于完全遮挡状态,这比短时跟踪更具挑战性, 并且这个评测基准是目前测试较大的单目标跟踪基准.

1) UAV123对比实验分析.

UAV123数据集总共包含123个视频序列,超过110 K帧. 所有的序列都使用竖直边框进行了完整注释. 该数据集对于跟踪而言主要的挑战包括快速运动、大尺度变化、光照变化大和遮挡等问题,相比于OTB 100[26],该数据集存在不少目标完全消失的情况,增加了数据集测验的难度. 在此数据集上,将本文算法与包括SiamRPN++ (evolution of Siamese visual tracking with very deep networks)、SiamDW (deeper and wider Siamese networks for real-time visual tracking)、SiamBAN (Siamese box adaptive network for visual tracking)、Ocean (object-aware anchor-free tracking)和SiamCAR在内的5种先进方法进行比较. 使用OPE的成功图和精度图作为指标评价整体性能,该生成图是按官方统一标准绘制. 如图4所示,相比于基准网络框架SiamCAR,本文算法在准确度上提高了2.1 %,

与其他先进的RPN算法相比,本文算法通过更简单的网络和启发式调整参数来获得具有竞争力的结果.

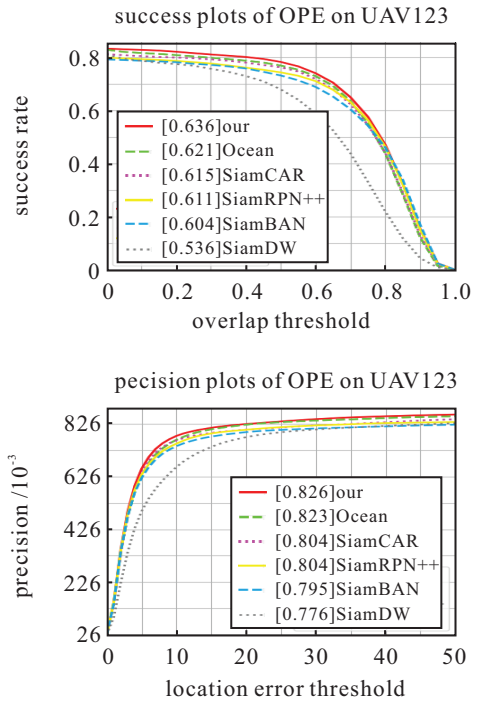

图 4 UAV123数据集的跟踪成功率与精确度对比

2) GOT-10k对比实验分析.

表 1 在GOT-10 K的细节对比

tracker	AO	SR0.5	SR0.75	FPS
KCF	0.203	0.177	0.065	94.6
fDSST	0.206	0.187	0.075	67.0
SRDCF	0.236	0.227	0.094	68.5
Staple	0.246	0.239	0.089	68.3
SAMF	0.246	0.241	0.084	69.3
CFnet	0.293	0.265	0.087	35.62
MDNet	0.299	0.303	0.099	1.52
ECO	0.316	0.309	0.111	2.62
SiamRPN	0.483	0.581	0.270	**97.55**
SiamRPN++	0.517	0.616	0.325	72.30
SiamCAR	0.569	0.670	0.415	49.83
ours	**0.594**	**0.694**	**0.464**	30.37

GOT-10 K是国际权威的通用单目标跟踪算法评测大型数据集. 它包含87种运动模式,560种运动物体,150万以上的手动标注的目标框,并且官方为了确保公平性,使该数据集中的训练集与测试集中数据类型是零重叠,具备场景丰富、算法挑战难度高等特点. 算法的评测需要在给定的训练数据集上训练模型,并在给定的测试数据集上对他们进行测试. 最后,将测试结果上传到官网,官网自动对结果进行分析评测,因此GOT-10 K的跟踪结果比其他基准更具有说服力. 如表1所示,本文各项指标比基准网络框架SiamCAR对AO、SR0.5和SR0.75的评分分别提高了2.5 %、2.4 %和4.9 %,评价指标主要参考的是平均重叠率(AO)和成功率(SR).

3) LaSOT对比实验分析.

LaSOT是一个高质量、类别丰富的大规模数据集,包含70种类别,视频序列达到1 400个. 与之前的数据集相比,LaSOT中平均序列长度超过2 500帧. 每个序列都有来自野外的各种挑战,目标可能会消失并在视野中重新出现,这考验了跟踪器在长时跟踪中重新跟踪目标的能力. 以LaSOT官网发布的测试集对本文跟踪器进行评估,该测试集含有高达280个的视频序列,丰富的场景很好地验证了跟踪器的泛化能力. 将本文算法与现有的先进跟踪算法进行比较,算法主要包括SiamCAR、 DiMP-50 (learning discriminative model prediction for tracking)[28]、ATOM (accurate tracking by overlap maximization)[29]、SiamRPN++等.

实验结果如表2所示. 表2结果表明,本文算法与基准算法SiamCAR相比,在归一化精度、精度、准确度3个指标上分别提升了0.1 %, 1.1 %, 2.7 %.

表 2 在LaSOT的细节对比实验

tracker	归一化精度	精度	准确度
MDNet	0.460	0.373	0.397
SiamFC	0.420	0.339	0.336
DiMP-50	**0.648**	–	0.569
StructSiam	0.418	0.333	0.335
ATOM	0.515	0.303	**0.576**
SiamRPN++	0.496	0.491	0.569
SiamCAR	0.600	0.510	0.507
ours	0.601	**0.521**	0.534

2.3 消融实验

为了详细研究本文算法各个模块带来的性能增益,对所使用的具体模块进行了消融实验,以评估它们对算法的特定贡献. 此处以SiamCAR模型为基础网络,逐步添加空洞卷积(Atrous)、 特征增强模块(Affine)和混合注意力模块(TAta),算法的评估在OTB 100数据集上进行,本研究采用本地机器上的测试结果进行实验评估和消融分析,结果如表3所示.

表3 每种改进策略产生的性能增益对比 %

SiamCAR	Atrous	Affine	TAta	Succ	Prec
√	–	–	–	66.3	87.6
√	√	–	–	67.0	85.3
√	–	√	–	67.2	86.4
√	–	–	√	67.3	87.5
√	–	√	√	68.5	87.6
√	√	√	–	68.3	86.3
√	√	√	√	**69.3**	**89.6**

由表3可知,每个模块对模型性能的提升均有不同程度的贡献,其中空洞卷积为原本的框架提升了感受野,使得成功率有了部分提高,扩大感受野的同时

也为锚点的选择造成了困难,导致准确度降低,当选择Affine模块和TAta模块时有大幅增益,这主要得益于Affine模块增强了主干提取的特征,丰富了语义信息,并且有TAta对锚点区域的特征增强,并且该模块有全局感知的作用,故在两项指标中均有提升. 之后通过添加空洞卷积获得更大的感受野,使得语义信息更加丰富,模型性能可以得到进一步提高,最终结合所有的模块,本文方法达到了较好的性能,Succ和Prec分别为69.3 %和89.6 %.

为了更直观地展示各个模块的性能,对于相关模块效果,在OTB100测试集中随机采样通过热度图来展示,具体效果如图5所示,在单独使用Affine模块和TAta模块时相对基准算法都能有不同程度的优化,像在Basketball图与Soccer图中,这两种背景颜色极其相似的情况下,Affine模块的感知能力就能很好地体现出来,在Dog图中,目标相对清晰,TAta良好的定位回归能力发挥了作用,引导网络精准定位. 同时也验证了好的目标跟踪不仅需要良好的特征,更需要对特征的判断,从而实现跟踪性能的鲁棒性. 但本文的实验也有不足,对于大尺度旋转很难实现对边界框的准确覆盖,分析原因是因为本文框架是基于无锚的跟踪,后处理中窗口的惩罚判定不够精准.

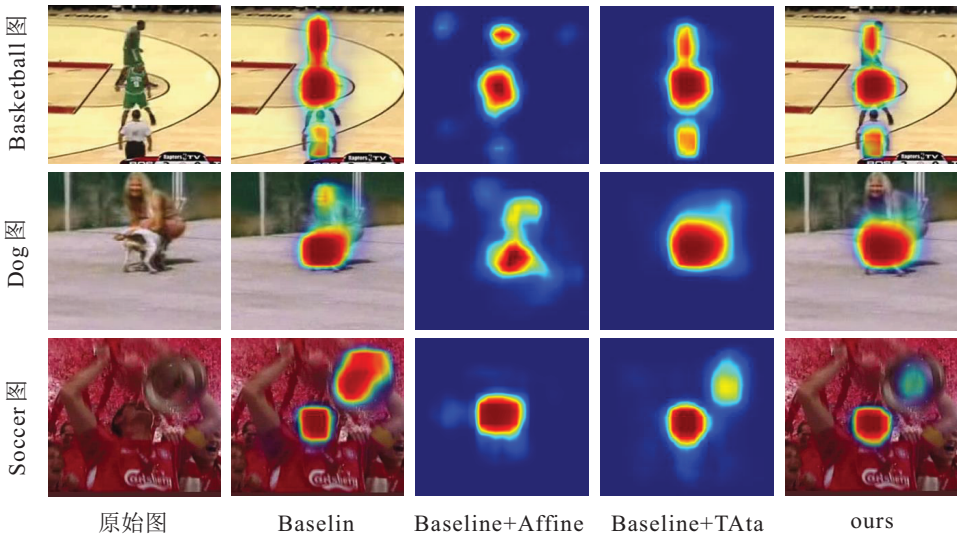

图5　可视化注意力模块热度图

3　结　论

本文提出了一种基于无锚双注意力孪生网络视觉跟踪算法. 考虑到跟踪网络的实际应用场景,所提出的跟踪算法利用逐像素增强机制提升主干网络提取的语义特征,提高了物体语义的判别性. 使用双路注意力增强锚点区域信息,实现了目标定位的精确表达. 实验结果通过与具有挑战性的视频序列上的其他视觉跟踪器进行比较,验证了该算法的有效性和鲁棒性.

本文实验发现,可以使用窗函数加强算法后处理,窗函数的辅助作用可以提高其整体泛化性能. 计划在未来的工作中加强算法后处理操作与目前工作的融合.

参考文献(References)

[1] Boudoukh G, Leichter I, Rivlin E. Visual tracking of object silhouettes[C]. The 16th IEEE International Conference on Image Processing. Cairo, 2010: 3625-3628.

[2] 刘如浩, 张家想, 金辰曦, 等. 基于可变形卷积的孪生网络目标跟踪算法[J]. 控制与决策, 2022, 37(8): 2049-2055.
(Liu R H, Zhang J X, Jin C X, et al. Target tracking based on deformable convolution Siamese network[J]. Control and Decision, 2022, 37(8): 2049-2055.)

[3] Danelljan M, Robinson A, Shahbaz Khan F, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking[C]. European Conference on Computer Vision. Amsterdam, 2016: 472-488.

[4] Kristan M, Leonardis A, Matas J, et al. The eighth visual object tracking VOT2020 challenge results[C]. Proceedings of European Conference on Computer Vision. Cham, 2020: 547-601.

[5] Danelljan M, Häger G, Khan F S, et al. Learning spatially regularized correlation filters for visual tracking[C]. 2015 IEEE International Conference on Computer Vision. Santiago, 2016: 4310-4318.

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

[7] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with

kernels[C]. European Conference on Computer Vision. Berlin, 2012: 702-715.

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

[9] Xu Y L, Wang J B, Li H, et al. Patch-based scale calculation for real-time visual tracking[J]. IEEE Signal Processing Letters, 2016, 23(1): 40-44.

[10] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 5000-5008.

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J/OL]. 2014, arXiv: 1409. 1556.

[12] Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, 2016: 1420-1429.

[13] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[C]. European Conference on Computer Vision. Cham, 2016: 850-865.

[14] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137−1149.

[15] 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.

[16] Wang Q, Zhang L, Bertinetto L, et al. Fast online object tracking and segmentation: A unifying approach[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 1328-1338.

[17] Zhu Z, Wang Q, Li B, et al. Distractor-aware Siamese networks for visual object tracking[C]. European Conference on Computer Vision. Cham, 2018: 103-119.

[18] Wang Q, Teng Z, Xing J L, et al. Learning attentions: Residual attentional Siamese network for high performance online visual tracking[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 4854-4863.

[19] 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. Seattle, 2020: 6268-6276.

[20] Du F, Liu P, Zhao W, et al. Correlation-guided attention for corner detection based visual tracking[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 6835-6844.

[21] Fan H, Bai H X, Lin L T, et al. LaSOT: A high-quality large-scale single object tracking benchmark[J]. International Journal of Computer Vision, 2021, 129(2): 439-461.

[22] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[C]. European Conference on Computer Vision. Cham, 2016: 445-461.

[23] Huang L H, Zhao X, Huang K Q. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1562-1577.

[24] Yang L, Zhang R Y, Li L, et al. SimAM: A simple, parameter-free attention module for convolutional neural networks[C]. International Conference on Machine Learning. Vienna, 2021: 11863-11874.

[25] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[C]. European Conference on Computer Vision. Cham, 2014: 740-755.

[26] 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.

[27] Real E, Shlens J, Mazzocchi S, et al. YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 7464-7473.

[28] Bhat G, Danelljan M, Van Gool L, et al. Learning discriminative model prediction for tracking[C]. 2019 IEEE/CVF International Conference on Computer Vision. Seoul, 2020: 6181-6190.

[29] Danelljan M, Bhat G, Khan F S, et al. ATOM: Accurate tracking by overlap maximization[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 4655-4664.

作者简介

郭文 (1978−), 男, 教授, 博士, 从事计算机视觉、多媒体计算等研究, E-mail: wguo@sdtbu.edu.cn;

梁卜文 (1994−), 男, 硕士生, 从事计算机视觉的研究, E-mail: 2020410017@sdtbu.edu.cn;

丁昕苗 (1979−), 女, 教授, 博士, 从事计算机视觉、视频理解等研究, E-mail: dingxinmiao@126.com.