SiamCross：孪生交叉的目标跟踪对象感知网络

黄旺辉[1),2)] 冯 永[1),2)] 强保华[3),4)] 裴钰璇[1),2)] 罗 越[1),2)]

[1)](重庆大学计算机学院 重庆 400044)

[2)](重庆大学信息物理社会可信服务计算教育部重点实验室 重庆 400044)

[3)](桂林电子科技大学广西可信软件重点实验室 广西 桂林 541004)

[4)](桂林电子科技大学广西光电信息处理重点实验室培育基地 广西 桂林 541004)

摘 要 近来，基于孪生架构的方法因其能在保持良好速度的同时取得较显著的性能，引起了视觉跟踪领域的广泛关注. 然而，孪生网络分支通常是独立的，缺乏信息交互，这限制了模型性能的进一步提升. 为了增强孪生网络分支的协作能力，本文提出基于孪生架构的交叉感知网络模型——SiamCross(Siamese Cross Object-Aware Network). 孪生网络双分支特征提取是提升模型性能的首要核心操作，区分目标和语义背景在很大程度上依赖模型挖掘的特征鲁棒性. 在 SiamCross 中，我们首先基于孪生网络分支的互监督，设计了全新的孪生交叉感知子网络(Siamese Cross-Aware Network，SCAN)用来提取鲁棒特征. SCAN 允许孪生分支彼此全方位高效协同工作，使模板分支可充分利用特征丰富的上下文语义信息，对目标产生更具有区分性的表示；搜索分支结合模板特征，也主动学习到了目标的本质信息. 另一方面，无锚框算法将跟踪任务直接映射为对每个像素的分类和回归，网络分支特征可各自聚焦于目标的局部与全局空间信息. 以上两种特征具有很好的潜在局部-全局互补性. 具体而言，回归特征学习到了更多的目标全局尺寸信息，但同时也引入了周围背景信息，而分类分支专注于学习局部中心定位信息. 二者结合，有利于抑制回归特征的背景信息表达. 同时，回归特征会在目标周边位置进行突出响应，揭示目标所在区域，也为分类分支进行定位提供了有益参考. 为充分利用以上不同的分支空间特征信息来获得更精确的跟踪结果，我们又提出了新型的目标注意力交互网络(Obejct-Attention Interaction Network，OAIN)，并将其融入到 SiamCross 中. OAIN 包含并行交叉注意力模块(Parallel Cross Attention Module，PCA)和自适应可形变交叉对齐模块(Adaptive Deformable Cross Align Module，ADCA). PCA 模块通过对分支中局部与全局信息的巧妙融合，提升了目标状态估计的准确性. 为了进一步使回归特征和目标区域对齐，缓解特征对齐失焦导致的分类分支参考信源可靠性大幅度降低，我们为 ADCA 模块设计了自适应空间转换操作，可以使得回归特征更好反映目标所在区域. 最终，ADCA 模块完善了无锚框网络的高效交互机制. 最后，我们在 OTB2015、VOT2018/2019、GOT-10k 和 LaSOT 五个具有挑战性的公开基准中对 SiamCross 进行了详尽的实验评估. 实验结果显示，SiamCross 与当前先进的跟踪器 SiamRPN++、ATOM 及 DiMP 相比，均取得了更优异的综合表现，并且可实现实时跟踪.

关键词 视觉目标跟踪；孪生网络；信息交互；交叉注意力

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.02151

SiamCross: Siamese Cross Object-Aware Networks for Visual Object Tracking

HUANG Wang-Hui[1),2)] FENG Yong[1),2)] QIANG Bao-Hua[3),4)] PEI Yu-Xuan[1),2)] LUO Yue[1),2)]

[1)](*College of Computer Science, Chongqing University, Chongqing* 400044)

[2)](*Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing* 400030)

[3)](*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi* 541004)

[4)](*Guangxi Key Laboratory of Optoelectronic Information Processing, Guilin University of Electronic Technology, Guilin, Guangxi* 541004)

Abstract Visual object tracking is a fundamental task of computer vision. Siamese based approaches recently attracted extensive attention in the visual tracking community, due to their ability to

收稿日期：2021-09-09；在线发布日期：2022-03-22. 本课题得到之江实验室开放课题(2021KE0AB01)、重庆市技术创新与应用发展专项重点项目(cstc2021jscx-gksbX0058)、国家自然科学基金(61762025)、广西可信软件重点实验室研究课题(kx202006)、广西光电信息处理重点实验室(培育基地)基金(GD18202)、广西自然科学基金重点基金(2019GXNSFDA185007)资助. **黄旺辉**，硕士，主要研究方向为深度学习与目标跟踪. E-mail：huangwh@cqu.edu.cn. **冯 永**(通信作者)，博士，教授，主要研究领域为大数据分析与数据挖掘、人工智能与大数据处理、深度学习与大数据检索. E-mail：fengyong@cqu.edu.cn. **强保华**，博士，教授，主要研究领域为大数据处理与信息检索. **裴钰璇**，学士，主要研究方向为深度学习与视觉目标跟踪. **罗 越**，学士，主要研究方向为深度学习与视觉目标跟踪.

achieve remarkable performance while maintaining good speed. However, Siamese network branches are usually independent and lack information interaction, which limits the further improvement of model performance. In order to enhance the collaboration ability of the Siamese network branch, this paper proposes a cross-awareness network model based on the Siamese architecture, called SiamCross (Siamese Cross Object-Aware Network). The feature extraction of the Siamese network branches is the primary core operation to improve the performance of the model. The distinction between the object and the semantic background largely depends on the feature robustness of model mining. In SiamCross, we first designed a new Siamese Cross-Aware Network (SCAN) based on the mutual supervision of the Siamese network branches. SCAN allows the twin branches to work efficiently with each other in an all-round way, so that the template branch can benefit from the rich contextual semantic information of search features and generate a more differentiated representation of the object; the search branch combines template features and actively learns the essential information of the object. On the other hand, thanks to the anchor-free network clearly formulate tracking tasks like classification and regression for each pixel directly, branch features can focus on local and global spatial information of the object, respectively. The above two features also tend to have good potential local-global complementarity. Specifically, the regression feature learns more information such as the global size of the object, but also inevitably introduces surrounding background information, while the classification branch focuses on learning local center localization information. The combination of the two is beneficial to suppress the expression of background information of regression features. Moreover, the regression feature will respond prominently around the object, revealing the area where the object is located, and also providing a good reference for the classification branch to locate. In order to make full use of the different spatial feature information in the branches to obtain more accurate tracking results, we also propose the Object-Attention Interaction Network (OAIN) and integrate it into SiamCross. OAIN contains two modules, Parallel Cross Attention Module (PCA) and Adaptive Deformable Cross-Align Module (ADCA). The PCA module improves the accuracy of object state estimation by ingenious fusion of local and global in the branch. In particular, in order to better align the regression features with the object region and avoid feature alignment out of focus, its reliability as a reference source for classification branches is greatly reduced. The ADCA module has an adaptive spatial transformation operation, which can make the regression features better reflect the object area. The ADCA module improves the efficient interaction mechanism of the anchor-free network. Finally, we extensively evaluated the proposed tracker on five public challenging benchmarks, including OTB2015, VOT2018/2019, GOT-10k, and LaSOT. The experimental results show that compared with current state-of-the-art trackers (e.g., SiamRPN++, ATOM and DiMP), SiamCross achieved excellent comprehensive performance and could run in real-time.

Keywords visual object tracking; siamese network; information interaction; cross attention

1 引 言

视觉目标跟踪是计算机视觉的一项基本任务.给定了序列初始帧中任意的目标状态后,跟踪器需准确估计后续帧的目标状态.目标跟踪在自动监控、车辆导航、机器人传感、人机交互和增强现实等领域有着广泛的应用.虽然视觉跟踪已经取得了很大的进展,但是由于存在光照变化、遮挡和背景干扰等诸多因素的影响,鲁棒视觉跟踪仍具有很大的挑战性.

基于孪生网络的跟踪算法将视觉目标跟踪问题视为学习一个通用相似得分图.因为其在精度和速度上取得了不错的平衡,近年来受到视觉跟踪领域的广泛关注.跟踪器通常需要对目标进行粗糙的中心定位和精确的状态估计(边界框).目前许多算法[1-2]使用多尺度搜索来估计目标大小,这种方法耗

时且计算负担大[3]. 为了获得精确的目标边界框，SiamRPN[4]引入了包括分类和回归分支的区域候选网络(Region Proposal Network，RPN)[4]. 但是，锚框需要人工参与设计，不仅增加了许多额外的超参数还需要先验知识(如比例分布)，这与通用目标跟踪精神相悖[5]. 与以往的基于锚框的算法相比，无锚框算法[6]具有显著的优势，它不需要预先定义候选框，可以在无任何参考的情况下预测每个像素对应的目标边界框. 近年来，众多高效的算法被提出[4,7]，但如图 1 所示，在诸如背景模糊、光照变化和遮挡等具有复杂背景信息的跟踪问题挑战下，依旧易发生跟踪漂移，亟待采用更好的方法进行优化.

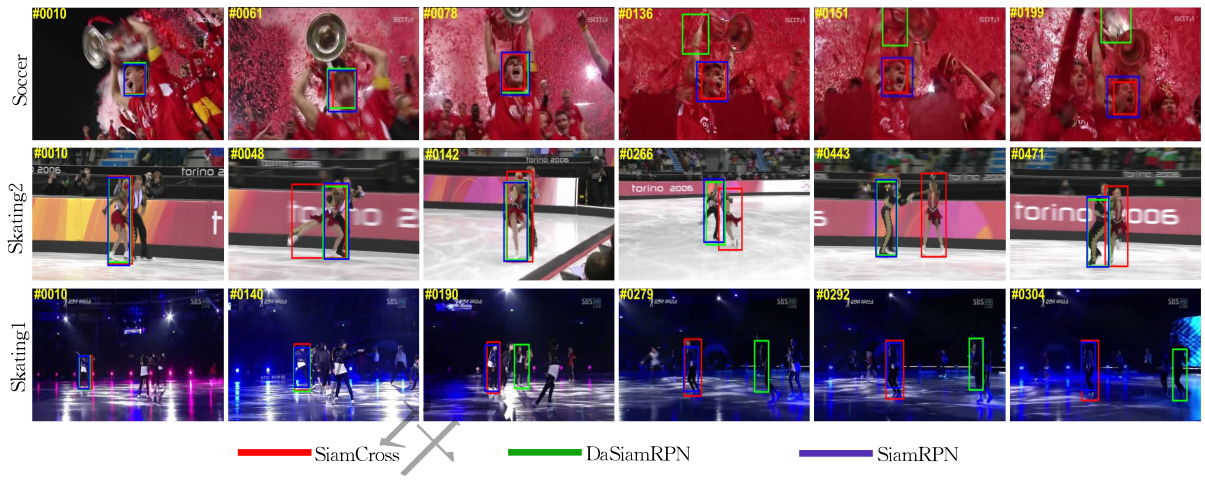


图 1　SiamCross 与 SiamRPN、DaSiamRPN 可视化比较

目前大多数基于孪生区域候选网络(Siamese-RPN)的特征学习方法，没有充分考虑分支之间的特征交互. 因此，导致在部分挑战性的场景中，学习到的特征不具有高区分性和可靠性. 孪生网络分支中提取的特征通常是分开计算的，搜索分支不会主动学习目标信息，目标信息则忽略了搜索图像中丰富的上下文信息，不利于产生更有意义的目标表示. 另一方面，基于无锚框的算法则直接对每个像素进行分类和回归，明确了分类任务和回归任务. 这使得生成的分类特征集中于局部图像特征息，对目标中心位置具有很强的鲁棒性；回归特征则通过学习目标的空间位置全局信息来获得准确的状态估计. 因此以上两种特征也具有很好的互补性. 然而，最近的方法并没有充分探索分类和回归分支之间进行有效交互互补的机制和方案. 这使我们产生了研究动机：可以基于增强分支之间的交互能力去设计更加准确可靠的无锚框跟踪框架. 当前的一些研究[8-9]，试图通过结合分支之间的特征来提高跟踪器的性能，但孪生架构的潜力还亟待进一步挖掘.

本文提出一种高效的基于孪生网络的交叉目标感知网络 SiamCross，探讨了基于孪生网络的双分支模型用于聚焦目标信息的具体策略，以提高模型在复杂场景(如遮挡、光照变化)下的判别能力. 在跟踪过程中，经常同时出现多个对象和遮挡的现象. 为了聚合有意义的上下文相关信息，我们设计了全新的孪生交叉感知子网络 SCAN，利用模板和搜索图像之间的互监督，使生成的特征更好的区分前景和语义背景. 与以往的研究[10]相比，我们明确考虑了初始帧标签信息，增加了前景嵌入位置的影响[11]. 此外，我们还提出了一个新的目标注意力交互网络(OAIN)来进行分类和回归，其中包括并行交叉注意力模块 PCA 和自适应可形变交叉对齐模块 ADCA. PCA 模块由一个平行的通道注意力子模块和空间注意力子模块组成，使得分类特征和回归特征分别专注于学习目标中心位置信息和空间尺寸信息. 这两个特征相结合，进一步提高了目标状态估计的精度. 最后，受益于更精确的边界框预测，我们引入 ADCA 模块将回归特征采样区域与预测的边界框对齐，削弱背景信息影响，来帮助分类分支生成更可靠的分类结果. 本文的主要贡献如下：

(1) 我们提出了孪生交叉目标感知网络 SiamCross，它能够改善模型漂移并减弱目标受相似干扰物的影响，提高跟踪器在复杂场景中的性能.

(2) 我们设计了全新的 SCAN 子网络和 OAIN 子网络，充分挖掘网络分支之间的交互潜力，保持跟踪效率的同时提高了跟踪精度.

(3) SiamCross 在 5 个具有挑战性的基准数据集上达到了当前综合最优的结果并可以实时运行.

2 相关工作

2.1 基于孪生网络的跟踪算法

SINT[12]最早将跟踪任务描述为将目标区域与搜索区域匹配的过程;随后 SiamFC 通过引入互相关操作来计算嵌入空间中样本和候选图像块之间的相似性,奠定了孪生网络跟踪器基础.虽然 SiamFC 取得较高的精度和不错的跟踪速度,但是使用这种多尺度测试来适应目标形变,无法处理目标纵横比改变的情况.SiamRPN 将孪生网络与基于锚框的区域候选网络结合,提供更精确的边界框估计并且保持超高跟踪速度.随后的研究 SiamRPN++[13]和 SiamDW[14],将更深、更强大的主干网络(例如 ResNet[15])引入孪生网络中,极大地提高了跟踪性能.然而,以上基于锚框的方法,需要精心设计锚框参数,不具备很好的泛化能力.最近,SiamBAN[16]等基于无锚框网络的跟踪算法[17],大幅减少了超参数的数量,直接对目标边界框进行分类和回归,提高了模型泛化能力.为了减弱复杂背景下干扰因素的影响,众多策略被提出,其中包括在线更新模板[18-19]、使用更深层的网络[20-21]或注意力机制[22]增强特征表示,以及其他的一些方法[7].

2.2 注意力机制

注意力机制是人脑固有的信号处理机制.计算机视觉中的注意力机制主要分为两类:自注意力和尺度注意力.自注意力解决了卷积神经网络的局部视野问题,它能够为每个位置提供全局视觉.Non-local[10]首先引入了视觉任务中的自注意力机制,来捕捉图像或视频的长程依赖.为了获得纵横路径上完整的图像上下文信息,CCNet[23]则设计了一个循环交叉注意力模块.相比于自注意力机制,尺度注意力基于每个位置的响应值来增强图像中的显著区域,从而抑制其他区域.SENet[24]提出了一个通道注意力模型去显式地学习特征通道之间的相互依赖关系,自动捕获每个特征通道的重要性.CBAM[25]则进一步构建了一个串联的通道注意力模块和空间注意力模块,强调了空间和通道维度上有意义的特征.随后的研究,RASNet[26]和 SiamAttn[22],将注意力机制引入到孪生网络框架中以提高跟踪性能.然而,上述大多数方法均从特征本身生成注意力特征.为了生成更可靠的注意力特征,Ocean+[11]构建了一个具有固定查找表(Look-Up-Table,LUT)的注意力检索网络.本文充分考虑了孪生框架并行分支的优势,从经典注意力结构出发,通过增强分支之间的信息交互,来设计一种全新的孪生交叉目标感知网络框架,进而提高模型的判别能力.

3 SiamCross 框架

如图 2 所示,SiamCross 由提取泛化特征的孪生网络 SCAN(左上部分)和用于分类、回归的 OAIN

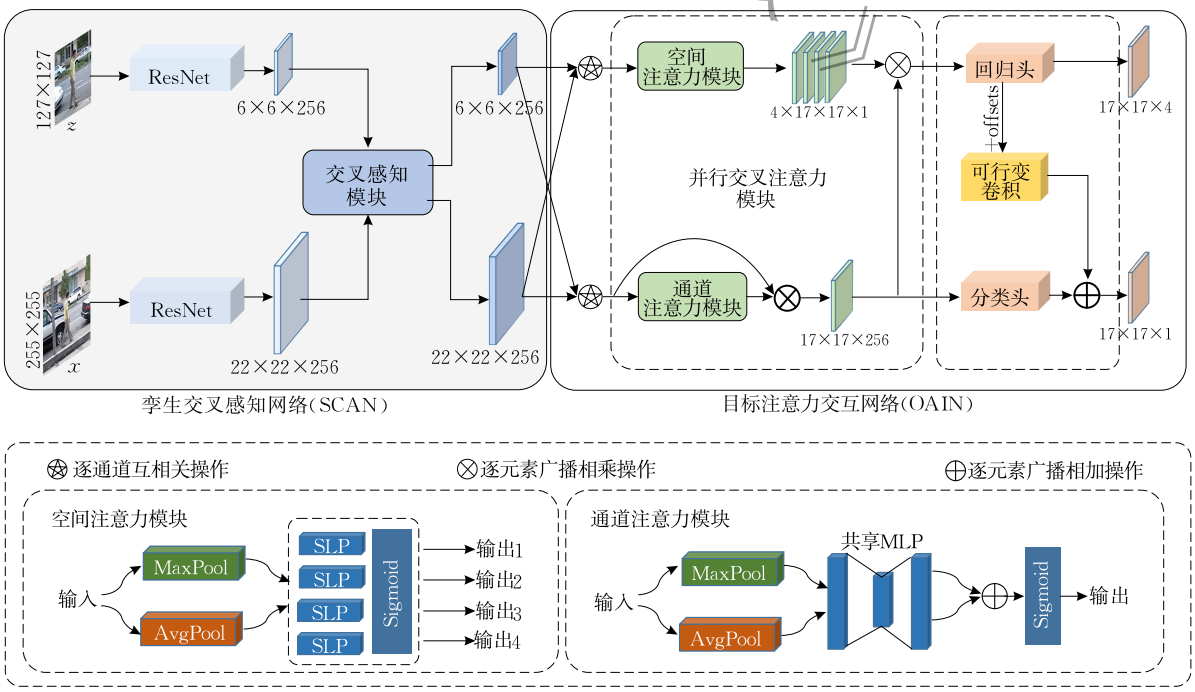


图 2 SiamCross 网络总体结构

网络(右上部分)组成.对于 SCAN,我们首先使用一个共享的深度骨干网络来提取高维度语义特征,然后串联一个交叉感知模块(Cross-Aware Module,CA)来加强生成的特征表示.模板分支特征受益于搜索特征的丰富上下文语义信息,可以对目标产生更有区分度的表示;搜索分支特征则结合模板特征,主动学习到了目标语义信息,背景干扰的影响也因此被削弱.为了更好地对目标进行中心定位和状态估计,我们还设计了一种基于信息交互的 OAIN 网络,充分利用分支之间的交互能力.在具体实现上,通过局部与全局特征的巧妙设计融合,去生成更鲁棒的特征表达.为方便描述,模板帧和搜索帧分别表示为 z 和 x,SiamCross 具体跟踪过程可用算法 1 描述.

算法 1.　SiamCross 跟踪流程.

输入:视频序列 $S=\{S_0,S_1,\cdots,S_n\}$;模板帧状态 B_0

输出:搜索帧目标状态估计 $B=\{B_1,B_2,\cdots,B_n\}$

1. $B\leftarrow B_0$, $n\leftarrow 1$ //初始化目标状态
2. WHILE $i\leq N$ DO
3. $z=127\times127$, $x_i=255\times255$ //裁剪图像
4. //SCAN 网络
5. $f(z)\leftarrow z$, $f(x_i)\leftarrow x_i$ //提取模板和搜索特征
6. $f(z),f(x_i)\leftarrow CA(f(z),f(x_i))$ //送入 CA 模块
7. $F_j\leftarrow f(z)\star f(x)_j$, $j\in\{cls,reg\}$ //互相关计算
8. //OAIN 网络
9. $A^{cls}_{w\times h\times1}$, $A^{align}_{w\times h\times1}$, $A^{reg}_{w\times h\times4}$ //送入 OAIN 网络计算相似度图
10. 分类图$\leftarrow A^{cls}_{w\times h\times1}+A^{align}_{w\times h\times1}$
11. 回归图$\leftarrow A^{reg}_{w\times h\times4}$
12. //更新目标状态
13. $B=B_i$
14. $i\leftarrow i+1$
15. END WHILE

3.1　孪生网络

孪生网络特征提取. 孪生网络由两个并行分支组成,分别以一个模板图像和一个搜索图像作为输入.模板图像在起始帧中以目标对象为中心进行裁剪,而搜索图像则是以上一帧预测的目标位置为中心,在当前帧裁剪出更大的搜索区域.我们采用改进后的 ResNet-50 作为主干网络.具体来说,前三个阶段结构和 ResNet-50 保持大致一致.但在第四个阶段,下采样单元的步长设为 1 来增加特征图尺寸大小,同时将所有的 3×3 空洞卷积的扩张幅度变为 2,以增加其感受野.第五个阶段设计结构类似,但是空洞卷积的扩张幅度为 4.最后,额外增加了一个 1×1 卷积将最后输出特征通道数减少至 256,以降低计算负担.

无锚框网络分类与回归. 基于无锚框的算法将跟踪视为逐像素预测问题[27],即直接对搜索图像上每个位置进行分类,并回归对应的目标边界框.相对于以往基于锚框的算法,响应图 A 的每个位置$(i,j,:)$不再被视为多尺度锚框的中心,而是直接映射到搜索图像中的对应位置(x,y),其对应的值 $A_{cls}(i,j,:)$负责预测相应区域前景-背景概率.回归分支则直接输出一个 4D 向量$\boldsymbol{T}^*(\boldsymbol{l}^*;\boldsymbol{r}^*;\boldsymbol{t}^*;\boldsymbol{b}^*)$,表示从位置$(x,y)$到搜索图像上目标边界四条边的距离,而非通过回归来调整预定义的锚框.整个过程可表述为

$$A^{cls}_{w\times h\times1}=\phi_{cls}(f(z)_{cls}\star f(x)_{cls}),$$
$$A^{reg}_{w\times h\times4}=\phi_{reg}(f(z)_{reg}\star f(x)_{reg}) \tag{1}$$

式(1)中,$f(:)$为用于特征提取的孪生嵌入编码空间,$\star$ 表示深度互相关操作[13].ϕ_{cls}和ϕ_{reg}分别表示分类分支和回归分支的编码空间.$A^{cls}_{w\times h\times1}$表示分类响应图,$A^{reg}_{w\times h\times4}$表示回归响应图.让$(x_0,y_0)$和$(x_1,y_1)$分别代表真值框的左上角和右下角,回归目标$(l^*;r^*;t^*;b^*)$可被计算如下:

$$l^*=x-x_0;\ r^*=x_1-x;$$
$$t^*=y-y_0;\ b^*=y_1-y \tag{2}$$

在线下训练时,我们进行集中采样去学习区域匹配的鲁棒相似度度量[17].只有靠近目标中心 $C_t(x_c,y_c)$半径为 16 像素范围的位置将被回归.

3.2　SCAN 网络

以往大多数基于孪生网络的跟踪器使用从孪生分支提取的特征,来完成不同的跟踪任务.但是在缺少目标模板监督的情况下,从搜索图像中学习到的特征 $f(x)\in R^{C\times(H_X\times W_X)}$ 缺乏被跟踪对象的相关信息.输出的模板特征 $f(z)\in R^{C\times(H_Z\times W_Z)}$ 也无法从搜索图像中丰富的上下文信息中获益.基于以上原因,如图 3 所示,我们提出了一个交叉感知模块,允许孪生分支彼此高效协同工作去学习输入图像对的更有区别性的表示.得益于模板和搜索分支相互监督,模板分支可从搜索分支中获取丰富的上下文信息,从而生成有意义的目标表示;生成的搜索特征具有高区分性,更侧重于跟踪对象本身特性.

对于搜索特征,受文献[11]的启发,我们期望在目标模板监督下为其建立一个空间约束.整个过程被视作检索一个由地址-值对组成的查找表,即 Look-Up-Table.对此,我们首先改变模板特征 $f(z)$

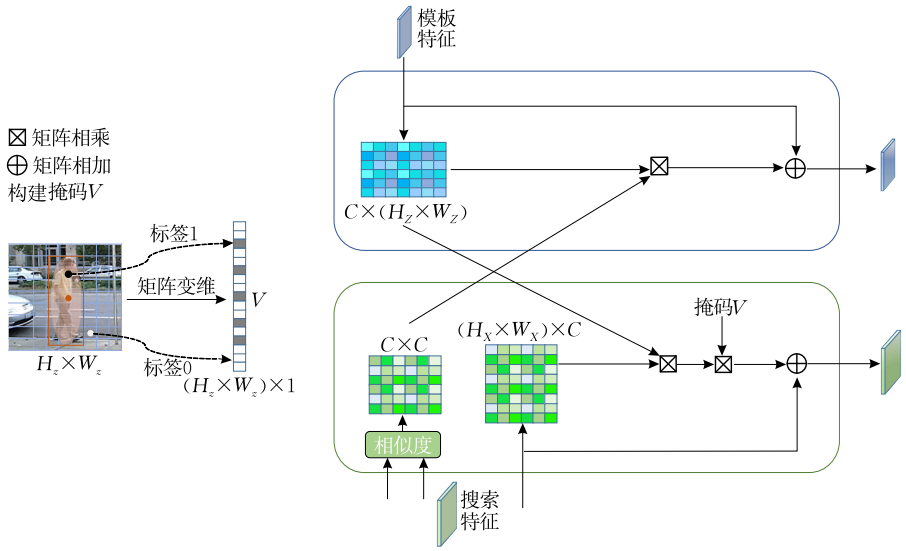


图 3 SiamCross 的交叉感知模块

形状大小为 $\overline{f(z)} \in R^{C \times (H_Z \times W_Z)}$ 并作为地址，同时准备一个对应标记的二进制掩码 $V \in R^{(H_Z \times W_Z) \times 1}$ 作为值，以指示 $f(z)$ 中每个位置是否存在目标. 参考 SiamBAN[16]，我们明确考虑目标尺度和比例的影响，采用椭圆标签进行标签分配. 只有在真值框内的内切椭圆位置被认为是有效目标(记为 1)，其它部分被认为是背景(记为 0). 因此 LUT 中，$f(x)$ 的每个空间位置和目标所有地址相似度矩阵为

$$\boldsymbol{A} = \mathrm{softmax}(\overline{f(x)} \times \overline{f(z)}) \tag{3}$$

其中 $\overline{f(x)} \in R^{(H_X \times W_X) \times C}$ 是 $f(x)$ 改变形状大小后的特征矩阵输出. 利用相似度矩阵 $\boldsymbol{A}$，可以计算出 $f(x)$ 中每个空间位置和目标特征 $f(z)$ 的相似度. 相似度越高，则响应值越大，当前位置属于目标的概率越大. 随后，其注意力矩阵计算为

$$\boldsymbol{M} = \boldsymbol{A} \times V \tag{4}$$

这样，与背景杂波相似的像素被削弱，而属于目标位置的像素被增强. 最后，我们将注意力矩阵编码到搜索特征 $f(x)$ 中：

$$\overline{f(x)} = \alpha \boldsymbol{M} \oplus \overline{f(x)} \tag{5}$$

其中 α 是缩放参数，$\oplus$ 是逐元素广播加法. 最终输出特征和 $f(x)$ 具有相同的形状.

对于目标特征，参考文献[22]，我们还引入了交叉注意力子模块，通过对搜索图像块中的上下文信息进行编码来生成更有意义的目标表示. 即我们从搜索特征而不是目标本身来学习关联映射：

$$\boldsymbol{A}' = \mathrm{softmax}(\widetilde{f(x)} \times \widetilde{f(z)}) \tag{6}$$

其中 $\widetilde{f(x)} \in R^{C \times (H_X \times W_X)}$ 同样是改变形状大小后的特征. 最终的输出特征可通过类似上述过程计算出：

$$\widetilde{f(z)} = \beta \boldsymbol{A}' \times \widetilde{f(z)} \oplus \widetilde{f(z)} \tag{7}$$

其中 β 是缩放参数，输出特征 $\widetilde{f(z)}$ 形状大小为 $R^{C \times H_Z \times W_Z}$. 相较以往基于孪生网络的方法，通常使用经典的视觉注意力机制从特征本身学习权重图，缺乏分支特征之间的相互监督. 跨分支的交互使生成的注意力特征更能感知区分前景和背景.

3.3 OAIN 网络

图 2 显示了互相关操作后连接的具有两个分支的 OAIN 网络. 分类分支对目标中心进行粗略定位，回归分支则对目标空间状态进行精确估计. 以往基于 Siamese-RPN 的算法直接将 RPN 分支用于完成分类和回归任务[15]. 然而，后续实验结果显示，得益于无锚框网络下各个分支的任务清晰明确，分支之间的特征信息互补可进一步提高跟踪性能. 因此，充分利用分支之间的关系是至关重要的. 为了达到这个目标，我们设计了一种新型的 PCA 模块. PCA 模块包括一个通道注意力子模块以生成通道注意力特征 $F_{cls}^{C} \in R^{C \times (H \times W)}$，并视作空间位置权重图与空间注意力子模块输出的 2D 特征进行计算，生成更鲁棒的空间注意力特征 $F_{reg}^{S} \in R^{C \times (H \times W)}$. 最后我们使用一个自适应可形变交叉对齐卷积模块 ADCA 去获得更可靠的定位结果，整体过程如算法 2 所示.

算法 2. OAIN 网络分类与回归.

输入：相似度图 F_{cls}，F_{reg}，分类头 $Head_{cls}$ 与回归头 $Head_{reg}$，权衡参数 ω

输出：分类响应图 $A_{w \times h \times 1}^{cls}$，回归响应图 $A_{w \times h \times 4}^{reg}$

1. $F_{cls}^{C} \leftarrow$ 通道注意力模块(F_{cls})
2. $F_{reg}^{S} \leftarrow$ 空间注意力模块(F_{reg})

3. F_{reg}^{S}←2D 特征 F_{reg}^{S} 与 F_{cls}^{C} 进行计算，参考式(9)

4. $A_{w\times h\times 1}^{cls}$←$Head_{cls}(F_{cls}^{C})$　//规则特征的分类响应图

5. $A_{w\times h\times 1}^{align}$←$ADCA(F_{reg}^{S})$　//对齐特征的分类响应图

6. $A_{w\times h\times 4}^{reg}$←$Head_{reg}(F_{reg}^{S})$　//回归响应图

7. $A_{w\times h\times 1}^{cls}$←$(1-\omega)A_{w\times h\times 1}^{cls}+A_{w\times h\times 1}^{align}$　//分类响应图

PCA 模块. 我们观察到分类特征的特定通道以学习到目标更多的语义信息[13]. 对此，我们通过注意力来进一步强化语义通道信息，并削弱其余通道的影响，使分类特征更具有区分性. 具体来说，对于给定的特征 F_{cls}，我们首先使用一个平均池化层去聚合空间信息和一个最大池化层去获取目标显著特征；然后将具有一个隐藏层的共享多层感知器(MLP)，应用在每个池化后的 1D 特征描述符上[25]. 最后，通过 Sigmoid 函数使得输出特征值映射到区间(0，1)中. 整个注意力特征生成过程可被描述为

$$F_{cls}^{C}=[\sigma(MLP(AvgPool(F_{cls}))+MLP(MaxPool(F_{cls})))]\otimes F_{cls} \tag{8}$$

其中，σ 代表 Sigmoid 函数，$\otimes$ 表示逐元素广播乘法. 隐藏层的输出大小为 $R^{C/r\times 1\times 1}$，为了降低计算负担减少率 r 设为 16.

在回归任务中，期望生成的回归特征更关注于全局空间信息的表达，如目标的尺寸、边界等信息. 为了达到这个目标，我们从特征“自身表达增强”和“分支交互”两个角度去设计了一个全新的网络结构. 对于特征“自身”如何产生更鲁棒的表达，我们很自然地使用空间注意力机制去做到这一点. 因为空间注意力机制会突出感兴趣的特征区域的表达，更加关注于目标的尺寸等全局信息. 另外，很符合直觉的一点就是：在基于无锚框的算法中，回归分支直接输出一个 4D 向量，来表示中心位置到目标边界四条边(上下左右)的距离. 我们由此设计了四个单层感知器 SLP，期望去分别学习不同空间维度的信息，再进行融合互补. 从分支交互角度而言，分类特征对目标中心位置具有更强的鲁棒性和具有全局性信息的回归特征结合可以抑制背景等非语义信息的表达，进一步提高跟踪器的精度.

在具体操作上，我们首先将输入的回归特征 F_{reg} 经过全局平均池化层(AvgPool)和最大池化层(MaxPool)生成两个 2D 空间特征描述符. 与之前的研究[9]不同的是，空间特征描述符进行逐元素相加后被送入四个单层感知器(SLP)去分别学习不同的空间信息(具体消融实验可参见第 4.5 节). 网络输出 4 个不同空间特征，并分别使用 Sigmod 函数进行归一化，然后和分类特征 F_{cls}^{C} 相乘. 最后累加生成回归注意力图：

$$S_i=\sigma_i(SLP_i(Cat(AvgPool(F_{reg})+MaxPool(F_{cls}))))\otimes F_{cls}^{C} \tag{9}$$

$$F_{reg}^{S}=\sum_{i=1}^{4}\gamma_i S_i$$

其中，SLP_i 和 σ_i 分别是第 i 个 SLP 头部和 Sigmoid 函数. SLP 由一个填充为 3 的 7×7 卷积组成以保持输入、输出特征大小相同. 最终的输出 S_i 乘以相应的权重系数 γ_1、γ_2、γ_3 和 γ_4，在我们的实验中简单地将其分别设置为 1、1、1 和 0.05.

自适应可形变交叉对齐模块. 在本文提出的框架中，分类分数反映了搜索图像相应采样区域的目标存在的置信度. 特征采样区域自适应目标的位移和变形，例如聚焦于目标区域，以提高跟踪器在复杂场景中的性能. 如图 4(a)和(b)所示，先前的方法从一个固定的规则区域采样(R_f)或膨胀间隔采样(R_d)，通常不能有效地覆盖目标或包含了过多的背景信息，不利于目标定位. 图 4(c)则展示了带有一个空间转换机制的自适应可形变交叉对齐模块，使

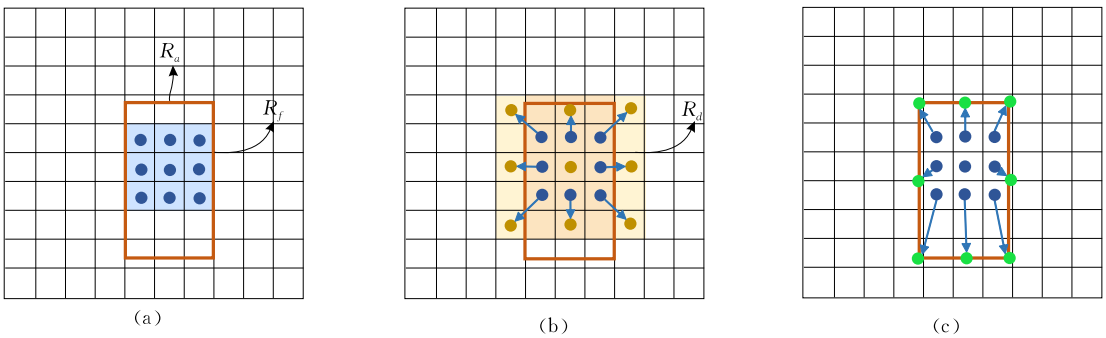


图 4　说明不同的特征采样方法(图中(a)蓝点、(b)黄点、(c)绿点分别表示固定规则卷积、膨胀卷积和自适应可形变卷积的采样位置，卷积核大小都设置为 3×3)

得采样区域从固定区域 R_f 转化到自适应采样区域 R_a(预测的目标边界框).给定输入特征 x,对齐特征 y 上每个位置 $u(dx,dy)$ 的特征对齐变换被定义为

$$y[u]=\sum_{r\in R,o\in O} w(r)*(x[u+g+o]) \quad (10)$$

在上式(10)中,集合 R 表示核大小为 $k\times k$ 的标准卷积,并具有固定规则的采样网格,即 $\left\{\left[-\frac{k}{2},-\frac{k}{2}\right],\left[-\frac{k}{2},0\right],\cdots,\left[0,\frac{k}{2}\right],\left[\frac{k}{2},\frac{k}{2}\right]\right\}$. $o\in O$ 表示从规则的蓝色样本点 $\{(dx,dy)+G\}$ 到自适应绿色采样点 $\{(mx,my)+M\}$(如图 4(c)所示)的 2D 距离偏移向量.位置 (mx,my) 为预测边界框的中心;$M=\left\{\left[-\frac{m_w}{2},-\frac{m_w}{2}\right],\left[-\frac{m_w}{2},0\right],\cdots,\left[0,\frac{m_w}{2}\right],\left[\frac{m_w}{2},\frac{m_w}{2}\right]\right\}$ 表示相对于 (mx,my) 新的自适应采样位置.整个过程与文献[28]相似.

类似的,回归特征保留了更多的目标空间全局尺寸信息,也是分类分支进行定位很好的参考.两种特征的结合又可进一步提高分类精度.如图 2,我们首先将输出的回归特征图和偏移量输入到可形变卷积中,对特征采样区域进行对齐.这样,回归特征可以自适应跟踪过程中的角度变化和形变,聚焦于目标所在区域采样.然后,使用一个 1×1 卷积将输出特征通道减少到 1.我们在实验中发现,将 ADCA 模块作为一个独立的分支进行训练,可以减少训练难度,提高实验效果.在推理过程中,我们结合分类分支和自适应形变交叉对齐模块的输出,以产生更可靠的分类结果.

3.4 损失函数

对于网络不同分支输出,我们遵从最近流行的孪生跟踪器[3,5,16]中设定:分类分支采用交叉熵损失[29]来衡量分类准确性;回归分支则使用具有尺度不变性的 IoU 损失[30],来更好反映预测边界框与真实框重合度.我们提出的多任务损失函数计算如下:

$$\begin{aligned}L(p_{i,j},q_{i,j},\boldsymbol{t}_{i,j})=&\lambda_1\sum_{i,j}L_{cls}(p_{i,j},p^*_{i,j})+\\&\lambda_2\sum_{i,j}L_{align}(q_{i,j},q^*_{i,j})+\\&\lambda_3\sum_{i,j}L_{reg}(\boldsymbol{t}_{i,j}T^*_{i,j})\end{aligned} \quad (11)$$

其中,L_{cls} 和 L_{align} 分别表示分类结果中基于规则区域特征和对齐特征的二值交叉熵损失. $p_{i,j}$ 和 $q_{i,j}$ 是位置 (i,j) 对应输出的分类得分值;$p^*_{i,j}$ 和 $q^*_{i,j}$ 则是对应的真值标签. L_{reg} 为边界框的 IoU 损失,$\boldsymbol{t}_{i,j}$ 为预测的回归距离向量.我们使用权重超参数 $\lambda_1,\lambda_2,\lambda_3$ 来平衡不同的任务,在实验中它们被经验性[5,17]地设置为 $\lambda_1=\lambda_2=1$ 及 $\lambda_3=1.1$.

4 实 验

4.1 实验配置

SiamCross 跟踪器使用改进后的 ResNet-50 作为孪生网络骨干,以及 ImageNet[31]上的预训练参数来初始化模型参数.为了充分学习到真实世界的复杂跟踪场景,我们采用 COCO[32]、ImageNet DET[31]、ImageNet VID[31]、GOT-10k[33]和 YouTube-BB[34]等大型数据集来进行训练.在实验设置中,模板图像输入分辨率为 127×127 像素,搜索图像输入分别率为 255×255 像素.我们在 4 个 GPU 上使用并行 SGD 进行梯度回传,训练的最大回合数为 50,每次迭代批尺寸大小设置为 128.在具体训练时,前 5 个回合进行预热,学习率固定为 0.001;在剩下的回合中,学习率指数从 0.005 逐渐下降到 0.000 01 以避免过拟合.权重衰减系数和动量参数则分别设置为 0.001 和 0.9.

4.2 评价指标

我们在 OTB2015[35]、VOT2018[36]、VOT2019[37]、GOT-10k[33]和 LaSOT[38] 5 个基准平台上进行广泛的评估:(1)OTB2015 包括 100 个视频以及 11 个跟踪挑战,我们使用精确度图(Precision Plots)和成功率图(Success Plots)作为评价指标;(2) VOT2018 与 VOT2019 两个基准都包括 60 个具有不同挑战的视频测试序列,它们的主要评价指标为平均重叠期望(Expected Average Overlap,*EAO*),综合考虑了准确率(Accuracy,*A*)以及鲁棒性(Robustness,*R*);(3) GOT-10k 是一个包含 10 000 个训练子集和 180 个测试子集的大型多类别跟踪基准,我们遵循 GOT-10k 协议,仅使用训练子集来训练 SiamCross,采用平均重叠率(Average Overlap,*AO*)和成功率(Success Rate,*SR*)作为评价指标与其它最先进的或基线的方法进行比较;(4) LaSOT 是最近提出的一个共有 1400 个视频序列的大规模、高质量数据集,我们同样遵从 LaSOT 协议,使用指定训练集来训练模型以及精确度图(Precision Plots)、成功率图(Success Plots)作为主要评价指标.上述指标相关定义可见表 1.

表 1　评价指标及相关定义

评价指标	定义
Overlap Rate	重叠率，定义为预测框与真值框交并比
Precision	精确度，通过比较预测框与真值框中心距离来计算
Success Frame	成功帧，在某帧中预测的边界框和真值框的重叠率超过给定阈值，则该帧称为成功帧
Precision Plots	精确度图，目标预测中心与真值框中心距离小于 20 像素的帧所占比例
Success Plots	成功率图，阈值为[0，1]区间范围内的成功帧曲线下面积(Area Under the Curve，AUC)
Accuracy(A)	准确率，代表跟踪成功帧的比例
Robustness(R)	鲁棒性，代表跟踪失败率
Expected Average Overlap(EAO)	平均重叠期望，综合考虑准确率和鲁棒性
Average Overlap(AO)	平均重叠率，表示所有预测回归框和真值框之间的平均重叠率
Success Rate(SR)	成功率，表示超过指定重叠阈值(0.5 或 0.75)跟踪成功的帧比率

4.3　与目前最先进的跟踪器比较

在本节中，SiamCross 将与 24 个最先进的跟踪器进行比较，包括基于孪生网络和基于相关滤波的算法以及其它的代表性方法.

OTB2015 基准评估. 图 5 显示了 SiamCross 与 DaSiamRPN[7]、ECO_HC[39]、SiamRPN[4]、SiamRPN++[13]、ATOM[21]、GradNet[18]和 Ocean[17]等 9 个顶尖跟踪器的比较情况，SiamCross 都达到最优或次优的跟踪结果. 与 SiamRPN 和 SiamFC 相比，精确度分别提高了 6.4%和 14.4%，成功率提高了 4%和 11.4%. 即使和目前最优的跟踪器 ATOM 以及 Ocean 比较，我们依旧表现出了相当优越的性能. 虽然和 SiamRPN++相比还有些细微差距，但是 SiamRPN++增加多层 RPN 获得更好跟踪结果的同时也较大地增加了模型的复杂度. SiamCross 使用更少运算量(FLOPs，49.2G V.S. 31.4G)却实现了相当接近的跟踪精度.

VOT2018 和 VOT2019 基准评估. 我们在 VOT2018 和 VOT2019 基准上将 SiamCross 和 19 种最先进的方法进行比较. 图 6 根据 EAO 指标对多个跟踪器进行排名，表 2 与表 3 则显示了更多定性比较结果. SiamCross 显著优于其它跟踪器，在 VOT2018 和 VOT2019 基准上 EAO 分别达到 0.471 和 0.339，在 VOT2018 上准确率比 SiamRPN++相对提高 5.7%. 不仅如此，SiamCross 鲁棒性优于所有其他不需要在线更新的孪生跟踪器，即使与在线学习的判别式跟踪器 ATOM 和 DiMP 相比，依旧具有相当的竞争性. 这充分验证分支交叉感知策略的优越性，可以很好地帮助跟踪器抑制干扰物，提高跟踪鲁棒性.

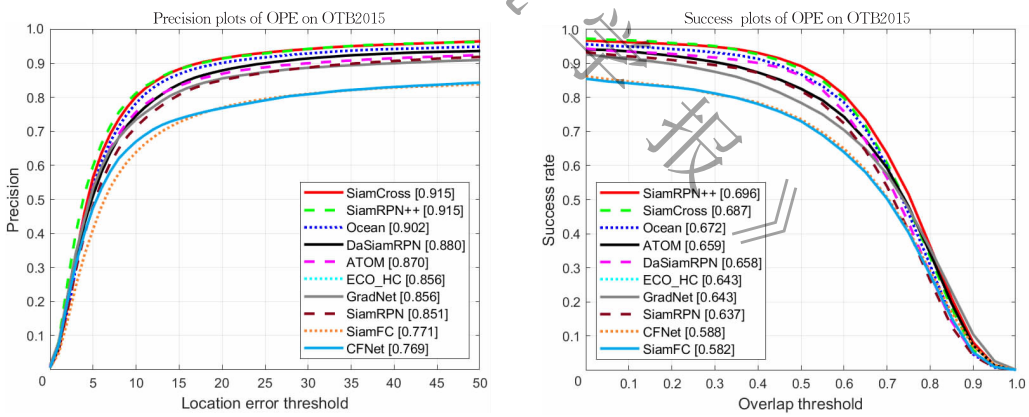
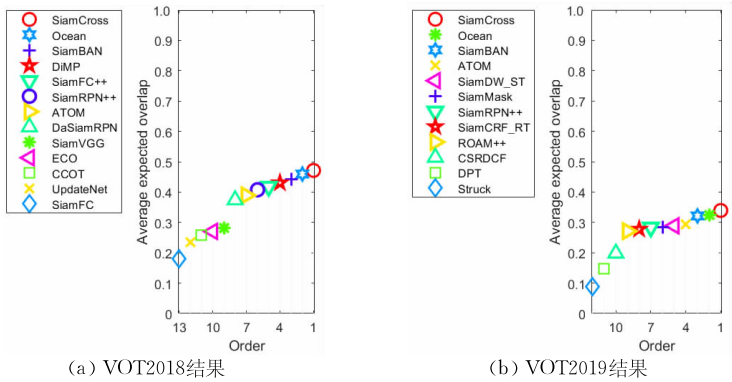


图 5　OTB2015 上精确图和成功率图比较结果(其中“Ocean”特指“Ocean-offline”，下同)



(a) VOT2018结果　　(b) VOT2019结果

图 6　VOT2018 和 VOT2019 上 EAO 得分比较结果

表 2 VOT2018 上详细对比结果

跟踪器	$A\uparrow$	$R\downarrow$	$EAO\uparrow$
SiamFC[1]	0.503	0.585	0.188
SiamVGG[40]	0.531	0.318	0.286
C-COT[41]	0.494	0.318	0.267
ECO[39]	0.484	0.276	0.280
DaSiamRPN[7]	0.590	0.280	0.283
updateNet[19]	0.518	0.454	0.244
ATOM[21]	0.590	0.204	0.401
DiMP[20]	0.597	**0.153**	0.440
SiamRPN++[13]	0.600	0.234	0.414
SiamFC++[5]	0.556	0.183	0.400
SiamBAN[16]	0.597	0.178	0.452
Ocean[17]	0.598	0.169	0.464
SiamCross	**0.608**	0.169	**0.471**

表 3 VOT2019 上详细对比结果

跟踪器	$A\uparrow$	$R\downarrow$	$EAO\uparrow$
Struck[42]	0.417	1.726	0.094
DPT[43]	0.488	1.008	0.153
CSRDCF[37]	0.496	0.632	0.201
ROAM++[44]	0.561	0.438	0.281
SiamCRF_RT[37]	0.550	0.301	0.282
SiamRPN++[13]	0.580	0.446	0.292
SiamMask[45]	0.594	0.461	0.287
SiamDW_ST[14]	0.600	0.467	0.299
ATOM[21]	0.603	**0.301**	0.301
SiamBAN[16]	**0.602**	0.396	0.327
Ocean[17]	0.590	0.376	0.327
SiamCross	0.596	0.329	**0.339**

GOT-10k 基准评估. 为了避免过拟合特定类别,GOT-10k 训练集和测试集中的对象类别没有重复.因此,在该基准进行测试的方法要求具有较强泛化能力来处理未知类别的对象,这极具挑战性.表 4 展现了 GOT-10k 测试数据集的结果.DiMP 在 AO 和 $SR_{0.75}$ 取得了最好的结果.因为在线更新对于跟踪未知的类目标非常重要.然而,SiamCross 依旧在 $SR_{0.5}$ 指标上取得了最好的分数 0.722.与 SiamRPN++相比,SiamCross 在 AO、$SR_{0.5}$ 和 $SR_{0.75}$.指标上显著提高了 9.2%、10.6% 和 15.1%,表明 SiamCross 跟踪时不但能生成精确的目标边界框,同时也具有很好的泛化能力.

表 4 GOT-10k 上详细对比结果

	$SR_{0.5}\uparrow$	$SR_{0.75}\uparrow$	$AO\uparrow$
SiamFC[1]	0.404	0.144	0.374
CFNet[46]	0.265	0.087	0.293
ECO[39]	0.309	0.111	0.316
DaSiamRPN[7]	0.536	0.220	0.444
Ocean[17]	0.695	0.465	0.592
SiamRPN++[13]	0.616	0.325	0.517
SiamFC++[5]	0.695	0.479	0.595
ATOM[21]	0.634	0.402	0.556
DiMP[20]	0.712	**0.492**	**0.611**
SiamCross	**0.722**	0.476	0.609

LaSOT 基准评估. LaSOT 平均视频序列长度超过 2500 帧,每个序列均有来自野外真实世界的各种挑战,目标可能在视野中消失或重新出现.因此跟踪器的鲁棒性对于在 LaSOT 上取得优异表现至关重要.表 5 将 SiamCross 和其它算法在 LaSOT 测试集上的表现进行了对比.SiamCross 在精确度图和成功率图指标上都取得了最好的表现,分别超过了 SiamRPN++ 2.7% 和 2.6%.跟踪器在长时跟踪中易丢失目标,在线更新的跟踪器在这一点具有显著的优势.然而,SiamCross 的性能甚至优于部分最近提出的在线跟踪器 ATOM.从以上比较结果可以看出,SiamCross 不但具有出色的鲁棒性也具有相当大的长时跟踪潜力.

表 5 LaSOT 上详细对比结果

	$Prec.\uparrow$	$Succ.\uparrow$
SiamFC[1]	0.339	0.336
ECO[39]	0.301	0.324
ECO_HC[39]	0.279	0.304
CFNet[46]	0.265	0.296
SiamRPN++[13]	0.491	0.496
ATOM[21]	0.505	0.514
SiamCross	**0.518**	**0.522**

在本文对比实验中,SiamCross 在多个跟踪基准上取得了能与 SOTA 跟踪器相当,甚至更胜一筹的性能.但在 VOT2019 和使用分割掩码标注的 VOT2020 数据集上的表现以及推理速度,我们提出的算法还有进一步提升的空间.

4.4 定性表现

本节将 SiamCross 和当前优秀的跟踪器在 OTB-2015 上不同属性的表现进行定性比较.图 7 和图 8 分别显示了跟踪器在各个属性上的精确度图和成功率图实验结果,图 9 则展示了在具体遮挡、光照、形变等情况下 SiamCross 与 SiamRPN、DaSiamRPN 的可视化对比结果.SiamCross 在所有属性上都实现了最优或次优性能.而大多数算法只在部分挑战中表现良好,这证明 SiamCross 在跟踪现实复杂场景的目标时具有优良的鲁棒性和稳定性.SiamCross 相比 SiamRPN++ 在背景模糊情况下有一定的差距,类似的情况也发生在其它的优秀跟踪器上,如 ATOM 和 Ocean.主要原因是因为 SiamRPN++ 聚合来自不同层次特征的信息有助于区分语义背景干扰,因此模型可以在背景杂乱的情况下获得显著提升.总体上,SiamCross 在快速运动、遮挡、超出视野和其它复杂场景下都具有出色的鲁棒性.

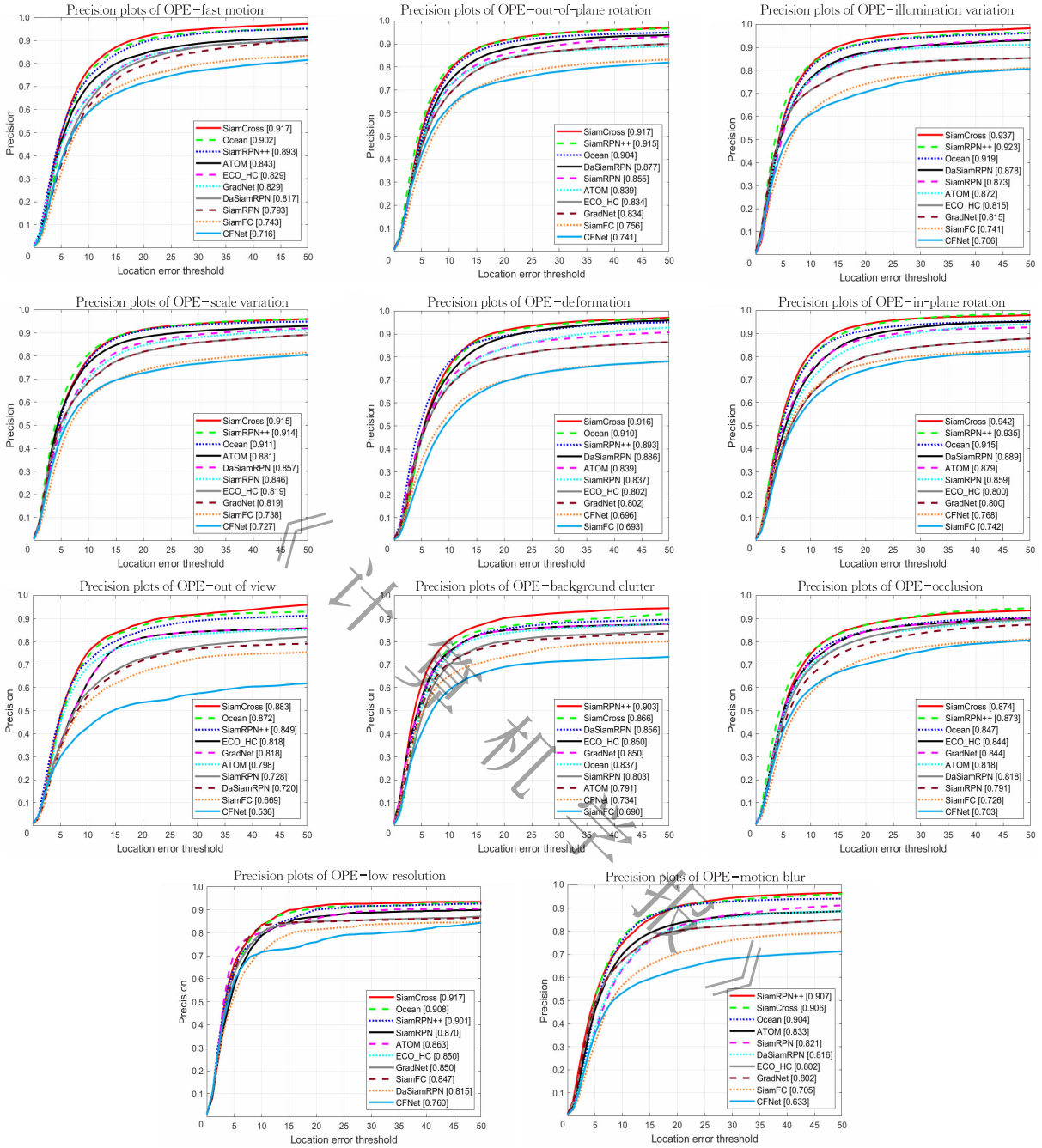


图 7　在 OTB2015 上，SiamCross 与 9 个最先进的追踪器在 11 种不同属性精确度图比较结果

4.5　消融实验

逐组件分析. 为了验证和分析 SiamCross 每个组件的有效性，我们在 VOT2019 上进行消融研究. 从表 6 可知，所有 SiamCross 组件(SCAN 模块、PCA 模块和 ADCA 模块)都可以使得跟踪效果得到提升(*EAO* 分别提高了 3.2%、5.4%和 2.3%). 与 SCAN 和 ADCA 模块相比，本文所提出的 PCA 模块是提高性能更为关键的部分. 从对模型推理速度影响角度分析，PCA 模块对跟踪速度影响较大(下降 18 fps)，而 SCAN 模块影响最小(下降 8 fps). SCAN 模块主要涉及矩阵操作来增强输入特征表示，轻量但可带来 3.2%的较大 *EAO* 值提升. 上述结果也表明无锚框网络分支间交互的巨大潜力. 通过不同模块间的聚合，跟踪器性能可以得到进一步提高. 最终 SiamCross 在 VOT2019 上 *EAO* 达到 0.339，实现了 8.8%的提高.

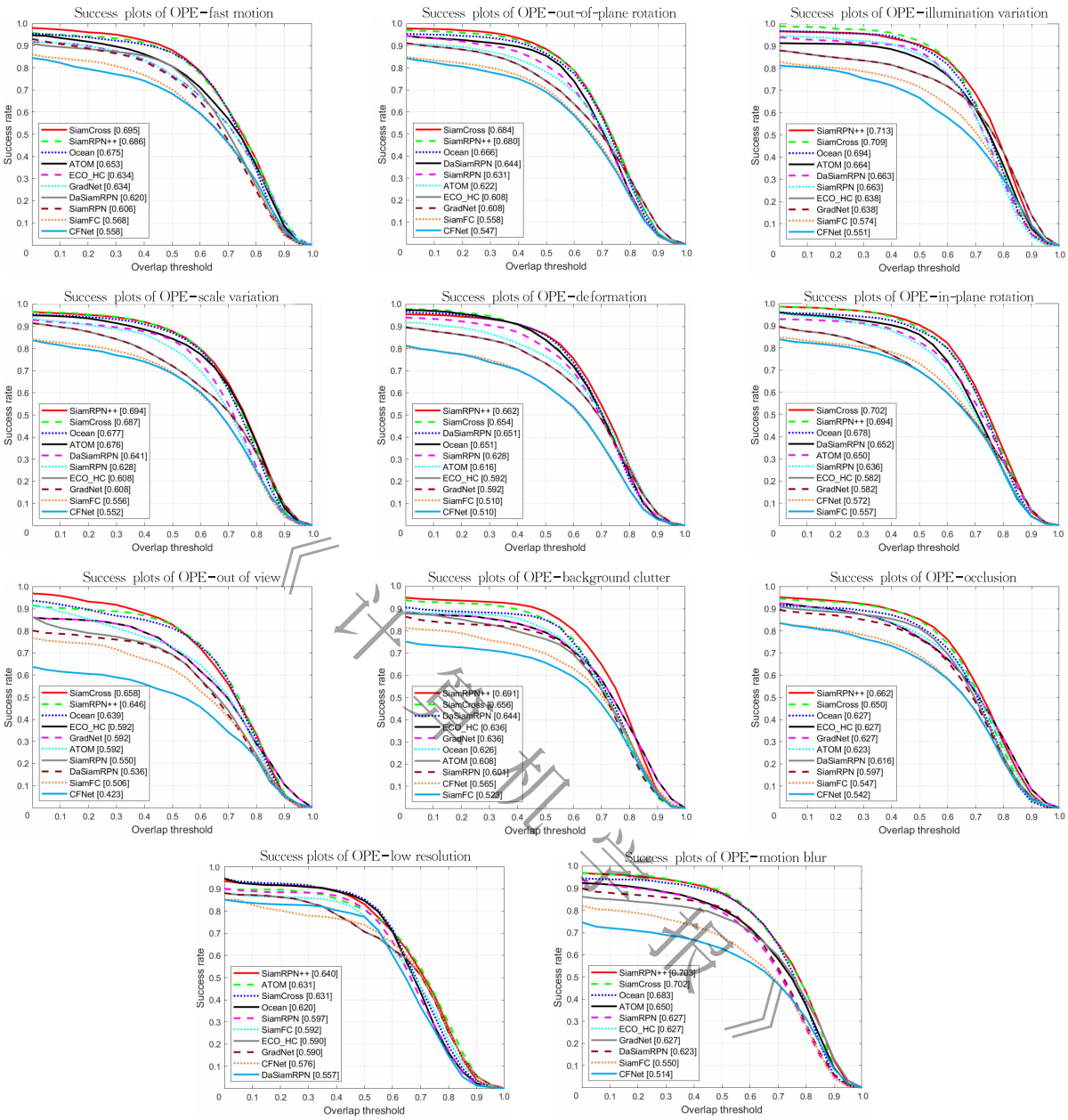


图 8 在 OTB2015 上，SiamCross 与 9 个最先进的追踪器在 11 种不同属性成功率图比较结果

表 6 VOT2019 上逐组件分析

SCAN 模块	PCA 模块	ADCA 模块	*EAO*↑	Δ*EAO*	跟踪速度/fps
			0.251	75	
✓			0.283	↑3.2%	67
	✓		0.305	↑5.4%	55
		✓	0.274	↑2.3%	64
✓	✓		0.322	↑7.1%	48
	✓	✓	0.314	↑6.3%	42
✓	✓	✓	**0.339**	↑8.8%	36

SCAN 模块和 PCA 模块结构分析. 为了更好地细粒度验证我们提出的 SCAN 模块和 PCA 模块的有效性，如表 7 所示，我们还对其进行 5 个额外的对比实验. 对于 SCAN 模块，当增加 LUT 子模块时，*EAO* 提高了 1.1%，这和我们预期的收益稍有出入. 我们认为很可能的一个原因，是在分割任务[11]中采用第一帧的 Mask(掩码)作为 LUT 的 Values(V). 但我们在算法中简单的根据模板帧提供的标签信息，将模板帧中以目标中心的椭圆区域为界限划分，最后进行二值化操作后作为 Values. 这里面尚有很大的探索空间. 对于 PCA 模块，我们着重分析了空间注意力子模块和通道注意力子模块对模型的贡献. 可以看到，我们各个子模块均带来了正向的收益. 特别是空间注意力模块，相对通道注意力模块，提供了 2.8%*EAO* 的更大增幅. 这也侧面验证了我们在局部和全局特征不同维度增强策略上的有效性.

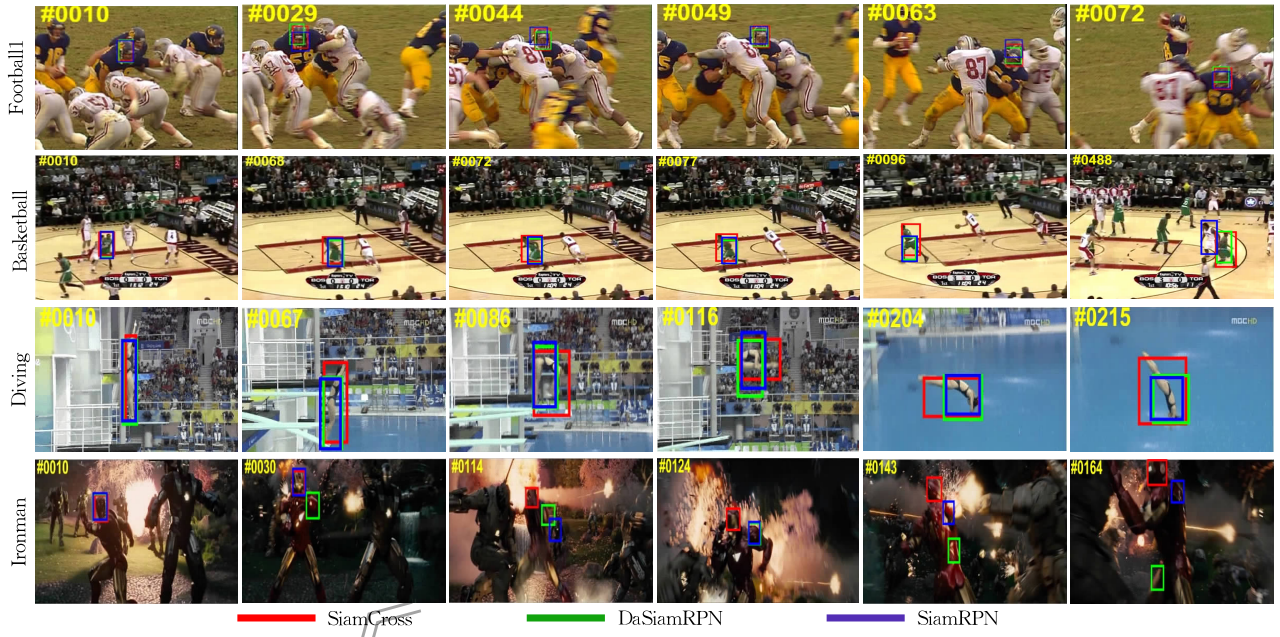


图 9　SiamCross 与 SiamRPN、DaSiamRPN 定性比较结果

表 7　VOT2019 上 PCA 模块和 LUT 模块细粒度贡献分析

LUT 模块	空间注意力模块	通道注意力模块	*EAO*↑
			0.284
✓			0.295
✓	✓		0.323
✓		✓	0.314
✓	✓	✓	**0.339**

空间注意力结构分析. 如表 8 所示，我们在 VOT2019 通过 6 个对比实验研究了不同 SLP 头个数对空间注意力的影响. 在 VOT2019 的测试结果上表明，当空间注意力模块有两个 SLP 输出头时，鲁棒性分数达到最高(0.298). 随着输出头的增加，当使用 4 个输出头时精度和 *EAO* 分别达到了最高值 0.596 和 0.339. 这其中可能的原因是：不同输出头的回归特征关注不同的局部空间目标信息，它们之间的特征融合可以更好地描述全局目标信息. 但当继续增加 SLP 输出头(大于 4)时，一方面带来更多权重超参数调整，不利于更好地进行特征融合；另一方面也增加了模型的复杂度，并未带来实际综合性能的提升.

表 8　VOT2019 上不同数目 SLP 定量比较结果

SLP heads	*A*↑	*R*↓	*EAO*↑
1	0.541	0.334	0.311
2	0.572	**0.298**	0.327
3	0.588	0.342	0.331
4	**0.596**	0.329	**0.339**
5	0.591	0.328	0.334
6	0.573	0.341	0.321

不同卷积特征采样分析. 为了进一步探讨使用不同卷积进行特征采样时对 ADCA 模块的影响，我们在 VOT2019 上进行了相关消融实验. 如表 9 所示，当直接使用规则卷积对回归特征进行采样并(*EAO* 值从 0.322 提高至 0.327)进行交叉融合时，虽然可带来模型性能少量提升，但是远低于使用可形变卷积交叉对齐带来的增幅(*EAO* 值从 0.322 提高至 0.339)，以上同样证明了使用可形变卷积进行特征对齐的必要性. 而当使用膨胀卷积进行特征采样时，实验结果显示，该采样方式并未提高模型性能，甚至出现性能下降的情况(*EAO* 值从 0.322 下降到 0.319). 这表明使用膨胀卷积进行对回归特征进行间隔采样时，并未能很好地对目标区域特征进行采样，导致与分类特征进行交叉融合时带来过多背景语义信息干扰.

表 9　ADCA 模块使用不同卷积特征采样在 VOT2019 上实验比较结果

规则卷积	膨胀卷积	可形变卷积	*EAO*↑
✓			0.327
	✓		0.319
		✓	**0.339**

5　总　结

本文提出了无锚框的孪生网络框架 SiamCross，实现了一个基于孪生网络分支协同交互的高性能跟踪器. 为了使得孪生网络产生的特征更有区分性，从

而增强目标感知能力,我们提出了一个基于分支监督的孪生交叉感知子网络,以提高跟踪器在复杂场景下的表现.为了获得更准确的跟踪结果,我们还提出了一种新型目标注意力交互网络.得益于并行交叉空间-通道注意力机制,分类和回归分支特征更关注局部或全局的空间信息,它们之间的充分交互又进一步提高了跟踪性能.SiamCross 在 5 个流行的基准数据集(OTB2015、VOT2018/2019、GOT-10k 和 LaSOT)上展现了具有相当竞争力的性能,并且达到实时跟踪,验证了其有效性和高效性.另一方面,自 SiamMask[45] 被提出以来,统一跟踪和分割趋势变得愈加普遍.我们未来的工作将研究如何使孪生网络更好地进行协同交互,以及设计更高效的网络来提高算法速度.同时还将考虑植入嵌入式设备中,应用于目标跟踪与分割.

致 谢 感谢罗秋红提出的宝贵意见,对提高本研究的技术内容有很大帮助!

参 考 文 献

[1] Bertinetto L, Valmadre J, Henriques J F, et al. Fully convolutional Siamese networks for object tracking//Proceedings of the European conference on computer vision. Amsterdam, The Netherlands, 2016: 850-865

[2] Lukezic A, Vojir T, Čehovin Zajc L, et al. Discriminative correlation filter with channel and spatial reliability//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 6309-6318

[3] Guo D, Wang J, Cui Y, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6269-6277

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

[5] Xu Y, Wang Z, Li Z, et al. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines//Proceedings of the Association for the Advancement of Artificial Intelligence. New York, USA, 2020: 12549-12556

[6] Tian Z, Shen C, et al. FCOS: Fully convolutional one-stage object detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 9627-9636

[7] 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: 101-117

[8] Du F, Liu P, Zhao W, et al. Correlation-guided attention for corner detection based visual tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6836-6845

[9] Gao P, Yuan R, Wang F, et al. Siamese attentional key point network for high performance visual tracking. Knowledge-Based Systems, 2020, 193: 105448

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

[11] Zhang Z, Liu Y, Li B, et al. Towards accurate pixelwise object tracking via attention retrieval. IEEE Transactions on Image Processing, 2021, 30: 8553-8566

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

[13] Li B, Wu W, Wang Q, et al. SiamRPN++: Evolution of Siamese visual tracking with very deep networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4282-4291

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

[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[16] Chen Z, Zhong B, Li G, et al. Siamese box adaptive network for visual tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6668-6677

[17] Zhang Z, Peng H, et al. Ocean: Object-aware anchor-free tracking//Proceedings of the European Conference on Computer Vision. Cham, 2020: 771-787

[18] Li P, Chen B, Ouyang W, et al. GradNet: Gradient-guided network for visual object tracking//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 6162-6171

[19] Zhang L, Gonzalez-Garcia A, Weijer J V D, et al. Learning the model update for Siamese trackers//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 4010-4019

[20] Bhat G, Danelljan M, Gool L V, et al. Learning discriminative model prediction for tracking//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 6182-6191

[21] Danelljan M, Bhat G, Khan F S, et al. ATOM: Accurate tracking by overlap maximization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 4660-4669

[22] Yu Y, Xiong Y, Huang W, et al. Deformable Siamese attention networks for visual object tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6728-6737

[23] Huang Z, Wang X, Huang L, et al. CCNet: Criss-cross attention for semantic segmentation//Proceedings of the IEEE International Conference on Computer Vision. Seoul, Korea, 2019: 603-612

[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132-7141

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

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

[27] Cui Y, Jiang C, Wang L, et al. Fully convolutional online tracking. arXiv preprint arXiv:2004.07109, 2020

[28] Vu T, Jang H, Pham T X, et al. Cascade-RPN: Delving into high-quality region proposal network with adaptive convolution//Advances in Neural Information Processing Systems. Sydney, Australia, 2019: 1432-1442

[29] de Boer P T, Kroese D P, Mannor S, et al. A tutorial on the cross-entropy method. Annals of Operations Research, 2005, 134(1): 19-67

[30] Yu J, Jiang Y, Wang Z, et al. UnitBox: An advanced object detection network//Proceedings of the 24th ACM International Conference on Multimedia. Amsterdam, The Netherlands, 2016: 516-520

[31] Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211-252

[32] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: Common objects in context//Proceedings of the European Conference on Computer Vision. Zurich, Switzerland, 2014: 740-755

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

[34] Real E, Shlens J, Mazzocchi S, et al. YouTube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 5296-5305

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

[36] 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

[37] Kristan M, Matas J, Leonardis A, et al. The seventh visual object tracking VOT2019 challenge results//Proceedings of the IEEE International Conference on Computer Vision Workshops. Seoul, Korea, 2019: 1-36

[38] Fan H, Lin L, Yang F, et al. LaSOT: A high-quality benchmark for large-scale single object tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 5374-5383

[39] Danelljan M, Bhat G, Shahbaz Khan F, et al. ECO: Efficient convolution operators for tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6638-6646

[40] Li Y, Zhang X. SiamVGG: Visual tracking using deeper Siamese networks. arXiv preprint arXiv:1902.02804, 2019

[41] 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, The Netherlands, 2016: 472-488

[42] Hare S, Golodetz S, Saffari A, et al. Struck: Structured out-put tracking with kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 38(10): 2096-2109

[43] Lukežič A, Zajc L Č, Kristan M. Deformable parts correlation filters for robust visual tracking. IEEE Transactions on Cybernetics, 2017, 48(6): 1849-1861

[44] Yang T, Xu P, Hu R, et al. ROAM: Recurrently optimizing tracking model//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 6718-6727

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

[46] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 2805-2813

HUANG Wang-Hui, M. S. His current research interests include deep learning and visual object tracking.

FENG Yong, Ph. D., professor. His current research interests include big data analysis and data mining, artificial intelligence and big data processing, deep learning and big data retrieval.

QIANG Bao-Hua, Ph. D., professor. His current research interests include big data processing and information retrieval.

PEI Yu-Xuan, Bachelor. Her current research interests include deep learning and visual object tracking.

LUO Yue, Bachelor. Her current research interests include deep learning and visual object tracking.

Background

This work is a part of the "Image and Text Unified Retrieval based on Semantic Deep Understanding", which is mainly supported by the Zhejiang Lab (No. 2021KE0AB01), and the National Natural Science Foundation of China (No. 61762025). In the process of image retrieval, the user's retrieval intention may be ambiguous, or the user lacks understanding of semantic information of input image, which makes the image retrieval results difficult to understand and evaluate, and the retrieval quality is difficult to guarantee. We carry on image and text unified re-between the image primitive semantics information and the hu-man perception information. Aiming at the problem of object recognition and classification accuracy in complex multi-label images, we propose a novel training algorithm, which pre-trains deep neural network using single-label image and fine-tunes the deep neural network using multi-label image. At the same time, we can reduce the candidate boxes combining with abjectness detection technology. Aiming at the isomorphism problem of heterogeneous feature spaces of image and text, we propose heterogeneous space mapping and normalization algorithm to construct unified feature vector model based on deep canonical correlation analysis, and we use the normalized feature to tune deep neural network training process and optimize image and text feature extraction model. In order to improve the user's experience of image and text unified retrieval, a novel image and text semantic automatic summary algorithm is proposed. We consider retrieval time, image and text semantic relevance, user satisfaction and other factors, and a subjective and objective combined sorting and recommendation algorithm is proposed. Finally, we can narrow the semantic gap and achieve efficient and accurate semantics image and text unified retrieval.