计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.009487]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于孪生网络的串联互相关目标跟踪①

陈凤姣[1], 程　旭[2]

[1](南京信息工程大学 软件学院, 南京 210044)
[2](南京信息工程大学 计算机学院、网络空间安全学院, 南京 210044)
通信作者: 程　旭, E-mail: xcheng@nuist.edu.cn

摘　要: 针对现有孪生网络目标跟踪技术只对模板特征和搜索特征进行一次融合操作, 使得融合特征图上的目标特征相对粗糙, 不利于跟踪器精确跟踪定位的问题, 本文设计了一个串联互相关模块, 旨在利用现有的互相关方法, 对模板特征和搜索特征做多次的互相关操作增强融合特征图上的目标特征, 提升后续分类和回归结果的准确性, 以更少的参数实现速度和精度之间的平衡. 实验结果表明, 所提出的方法在 4 个主流跟踪数据集上都取得了很好的结果.

关键词: 深度学习; 目标跟踪; 视频监控; 孪生网络

引用格式: 陈凤姣,程旭.基于孪生网络的串联互相关目标跟踪.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9487.html

Sequential Cross-correlation Object Tracking Based on Siamese Network

CHEN Feng-Jiao[1], CHENG Xu[2]

[1](School of Software, Nanjing University of Information Science and Technology, Nanjing 210044, China)
[2](School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Existing Siamese network object tracking techniques perform only one fusion operation of template features and search features, which makes the object features on the fused feature map relatively coarse and unfavorable to the tracker's precise positioning. In this study, a serial mutual correlation module is designed. It aims to use the existing mutual correlation method to enhance the object features on the fused feature map by performing multiple mutual correlation operations on the template features and the search features, so as to improve the accuracy of the subsequent classification and regression results and strike a balance between speed and accuracy with fewer parameters. The experimental results show that the proposed method achieves good results on four mainstream tracking datasets.

Key words: deep learning; object tracking; video surveillance; Siamese network

1　引言

目标跟踪是计算机视觉领域热点研究问题. 它被广泛应用于体育赛事转播、视频监控[1]和无人机、无人车、机器车等领域. 近年来, 凭借着强大的表征学习能力, 深度学习发展迅速, 基于卷积神经网络 (CNN) 的深度跟踪器在速度和精度方面取得了很好的成绩. 其中, 基于孪生网络的跟踪模型是所有深度学习方法中最流行的跟踪器. 孪生跟踪器主要遵循检测跟踪的思想, 首先, 孪生跟踪器学习一个模板特征映射, 然后计算与搜索区域特征的相似度, 最后选择得分最高的位置作为目标的预测. 如何获得准确可靠的特征一直是跟踪技术努力的方向, SiamFC[2]、SiamRPN[3]、SiamRPN++[4]、SiamAttn[5]、SiamCAR[6]、Ocean[7]等许多跟踪器在这方面取得很大的突破.

对于从孪生网络中获取的模板特征和搜索特征, 通常使用互相关运算来确定目标是否出现在搜索分支

① 收稿时间: 2023-10-12; 修改时间: 2023-11-27, 2023-12-06; 采用时间: 2023-12-27; csa 在线出版时间: 2024-04-01

特征上. 互相关运算是跟踪网络中最重要的部分, 模板特征和搜索特征互相关运算输出的融合特征图对后续分类和回归目标至关重要. 在之前的文章中, 有 3 种经典的互相关操作: SiamFC[2]中的传统互相关、SiamRPN[3]中的 Up-channel 互相关和 SiamRPN++[4]的深度可分离互相关 (depth-wise cross-correlation). 然而, 这 3 种互相关方法只对模板特征和搜索特征进行一次交互操作, 使得融合特征图上的目标特征相对粗糙, 不利于跟踪器精确定位. 经过实验分析, 模板特征和搜索特征进行多次互相关运算可以逐步细化特征, 引导跟踪网络学习到更多的特征, 从而提高跟踪模型的性能. 为此, 本文设计了一个串联互相关模块, 旨在利用现有的深度可分离互相关方法, 对模板特征和搜索特征做多次的互相关操作增强融合特征图上的目标特征, 提升后续分类和回归结果的准确性, 以更少的参数实现速度和精度之间的平衡. 此外, 为了构建简单通用的跟踪模型, 本文还采用一种无锚、无候选框的目标预测方法. 融合特征图输入预测头部分, 输出 3 个分支: 分类分支、筛选分支、回归分支. 使用串联互相关模块和无锚框的目标预测方法, 跟踪模型能够应对多种具有挑战性的场景, 如目标消失再出现、背景昏暗等.

本文提出的框架为视觉跟踪领域提供新的思考视角, 主要贡献在于: 1) 提出一种新的融合模板特征和搜索特征的方法, 细化了融合特征图上的目标特征, 提高了跟踪器的跟踪性能; 2) 设计一种简单有效的目标预测方法, 该方法能够使跟踪器有效应对多种具有挑战性的场景, 目标消失再出现、背景昏暗等; 3) 以 ResNet-50 为特征提取网络, 提出的目标跟踪方法在多个公开数据集上领先其他算法.

2 相关工作

近些年, 基于孪生网络的跟踪方法不断的融合其他领域优秀的方法, 孪生网络的跟踪精度和成功率都有了很大的提升. 孪生跟踪器 SiamFC[2]使用全卷积网络提取特征, 对提取到的模板图像特征和搜索区域特征进行相似性匹配, 输出响应得分图, 得分最高的位置即为目标所在. 由于 SiamFC[2]在跟踪速度与跟踪精度之间的良好平衡, 为跟踪领域带来了新的方向, 成为当前最经典的孪生网络目标跟踪算法. 孪生跟踪器 SiamRPN[3]通过引入目标检测中的区域建议网络 (RPN) 用于跟踪任务中的目标预测, 得到更精确的目标边界框, 避免了多尺度测试带来的高计算消耗, 提高了跟踪速度. Wang 等人[8]提出可同时实现视频目标跟踪和目标分割这两个任务的 SiamMask 算法, 在用于目标跟踪的孪生网络上增加一个掩码分支来实现目标分割, 同时优化网络. SiamMask[8]模型简单, 功能多样, 速度快, 其效果也超越了其他跟踪算法. 孪生跟踪器 SiamRPN++[4]使用一个简单但是有效的空间采样方式来打破孪生网络用于跟踪不能有效利用深度网络的限制, 同时, 将传统的互相关操作用深度可分离互相关代替, 得到多通道具有不同语义特征的得分图, 使得跟踪目标更准确. 孪生跟踪器 U-AST[9]解决了现有孪生网络仅依靠预测到的 4 条边来实现有效的回归, 无法解决复杂场景中目标边界框准确回归的问题. 孪生跟踪器 GSiamMS[10]通过集成 Res2Net 模块和 Transformer 模块来应对现有孪生跟踪器很难处理目标受到大规模变化、类似物体干扰的场景, 以及孪生网络无法建立模板分支和搜索分支之间连接的问题.

3 方法

如图 1 所示, 论文整体框架有特征提取网络, 串联互相关模块, 预测头 3 个部分构成. 首先将经过剪裁的图像对输入到特征提取网络输出模板特征和搜索特征, 分别记为$\varphi(Z)$和$\varphi(X)$. 其次, 将模板特征和搜索特征输入串联互相关模块输出融合特征图. 最后, 融合特征图被送入预测模块进行目标定位.

3.1 特征提取网络

论文使用 ResNet-50 作为特征提取网络. 将视频帧尺寸为 127×127 和 255×255 的图像对输入特征提取网络进行目标对象的特征提取, 得到不同阶段不同大小的特征图. 经过实验对比分析, 选取特征提取网络最后 3 个阶段的特征输入到串联互相关模块, 跟踪器的速度和准确度最好. 最后将 3 个阶段的细化特征利用一个 *Cat* 函数把 3 张特征图进行一个拼接操作, 得到一张整体特征融合图, 记为S.

3.2 串联互相关模块

串联互相关模块的详细结构如图 2 所示. 以前的孪生网络互相关模块都是对模板特征和搜索特征进行一次互相关操作, 得到的是目标对象在搜索区域上的基本匹配信息, 使得融合特征图上的目标特征相对粗糙, 不利于跟踪器后续对目标精确定位. 为了解决上述问题, 对模板特征和搜索特征进行二次互相关运算, 不

仅可以引导跟踪网络学习到更多明显的特征, 而且还可以提高跟踪模型的性能. 第 1 个互相关用于初始融合模板特征和搜索特征, 第 2 个互相关用于细化特征, 提高后续分类和回归任务的准确性. 串联互相关模块的输入是特征提取网络输出的模板特征$\varphi(Z)$和搜索特征$\varphi(X)$. 首先, 对模板特征和搜索特征做一个深度互相关操作得到一个融合特征φ_{fusion}. 上述操作具体表达式为:

$$\varphi_{\text{fusion}} = \varphi(Z) * \varphi(X) \tag{1}$$

其中, $*$表示深度可分离互相关, 融合特征φ_{fusion}得到的是目标对象在搜索区域上的基本匹配信息. 为了细化特征提高召回率, 需要对搜索特征和融合特征进行加权求和. 由于搜索特征图的长宽为 31×31, 融合特征φ_{fusion}长宽为 25×25, 使用一个 Upsample 改变融合特征图的长宽, 以便进行后续的加权融合操作. 这个 Upsample 操作是由两个连续的$Conv+BN+ReLU$组成. 第 1 个$Conv+BN+ReLU$块的卷积核大小被设置为 3, Padding 设置为 3, 第 2 个块$Conv+BN+ReLU$的卷积核大小也设置为 3, Padding 设置为 2. 使用两个不同大小的$Conv+BN+ReLU$将融合特征分两次进行上采样, 这样做不仅能学习到不同尺度的细粒度语义信息, 还能减少信息丢失, 有利于准确跟踪目标, 提高跟踪网络的鲁棒性, 从而使得后续预测网络能够获得更多的目标特征信息. 然后, 设置一个加权融合策略来聚合融合特征φ_{fusion}和搜索特征$\varphi(X)$, 具体表达式如下所示:

$$\varphi_{\text{sum}} = \lambda_1 Upsample(\varphi_{\text{fusion}}) + \lambda_2 \varphi(X) \tag{2}$$

其中, 经过反复实验λ_1和λ_2的值分别被设置为 0.8 和 0.2, 才能在聚合互相关特征φ_{fusion}的同时不过度隐藏原始搜索特征$\varphi(X)$.

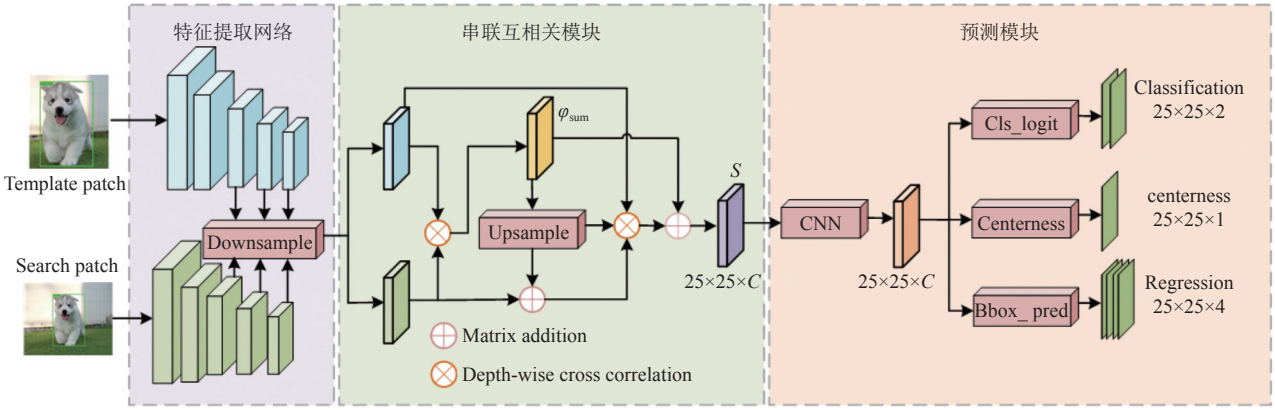


图 1 跟踪框架

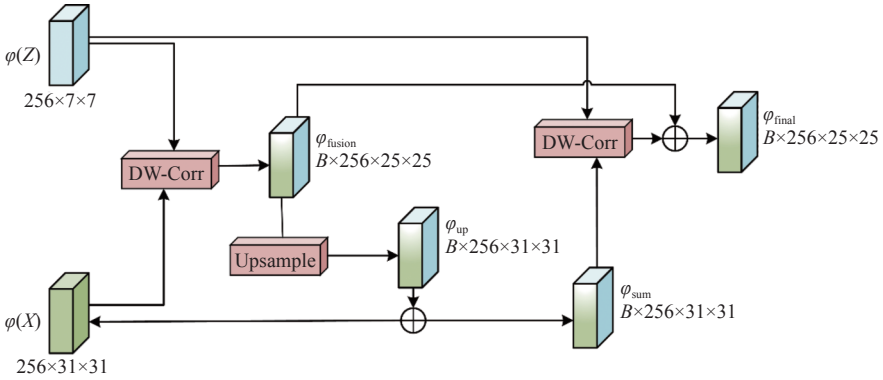


图 2 串联互相关模块详细图

最后, 对φ_{sum}和模板特征$\varphi(Z)$执行第 2 个深度互相关操作得到的结果与第 1 个深度互相关结果相加得到最后的融合特征φ_{final}. 如式 (3) 所示:

$$\varphi_{\text{final}} = (\varphi(Z) * \varphi_{\text{sum}}) + \varphi_{\text{fusion}} \tag{3}$$

本文把特征提取网络后 3 个阶段的模板特征和搜索特征分别输入串联互相关模块得到 3 张融合特征图, 然后, 利用一个 *Cat* 函数把 3 张特征图进行一个拼接操作, 得到一张整体特征融合图S. 具体定义如下所示:

$$S = Cat(\varphi_{\text{final}}, \varphi_{\text{final}}, \varphi_{\text{final}}) \tag{4}$$

其中, 3 个φ_{final}分别代表 3 个阶段的模板特征和搜索特征输入到串联互相关模块的输出.

3.3 预测头

如图 1 所示, 串联模块输出的整体特征图S输入预测模块, 输出分类特征图$M^{\text{cls}}_{w\times h\times 2}$、回归特征图$M^{\text{reg}}_{w\times h\times 2}$和筛选特征图$M^{\text{cho}}_{w\times h\times 1}$这 3 张特征图. 分类特征图上的每一个点$(i,j,:)$表示一个 2D 向量, 其意义是表示输入搜索区域中对应位置的前景和背景分数. 相似的, 回归特征图上的每一个点$(i,j,:)$表示一个 4D 向量$reg(i,j)=(l,t,r,b)$, 输出检测目标的 4 个角点. 筛选特征图上的每个点值给出相应位置的得分, 根据这些得分排除异常值, 筛选最优值.

本文采用交叉熵损失进行分类, *IOU* 损失进行回归预测. (x_0,y_0)和(x_1,y_1)分别表示 ground truth bounding box 的左上角和右上角. (x,y)表示(i,j)的对应位置, 回归特征图上的回归目标可以通过计算得出:

$$\begin{cases} l = x - x_0,\ t = y - y_0 \\ r = x_1 - x,\ b = y_1 - y \end{cases} \tag{5}$$

利用回归目标, 可以计算出 ground truth bounding box 和这个预测的 bounding box 的 *IOU* 损失. 通过以下列公式计算回归损失:

$$L_{\text{reg}} = \frac{1}{\sum\prod(reg(i,j))}\sum_{i,j} reg(i,j)$$
$$L_{IOU}(M^{\text{reg}}(i,j:), cls(x,y)) \tag{6}$$

其中, L_{IOU}代表 *IOU* 损失, $\prod(\cdot)$表示一个指标函数, 定义为:

$$\prod(reg(i,j)) = \begin{cases} 1, & \text{if } reg^{k}_{(i,j)} > 0,\ k = 0,1,2,3 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

实验观察到, 远离目标中心的位置会产生低质量的预测框, 这降低了跟踪的性能. 因此在预测头部分添加了一个筛选分支, 筛选特征图上的每个点值给出相应位置的得分, 根据这些得分排除异常值筛选最优值. 筛选分支上每个点的得分$C(i,j)$可以定义为:

$$C(i,j) = \prod(reg(i,j)) * \sqrt{\frac{\min(l,r)}{\max(l,r)} \times \frac{\min(t,b)}{\max(t,b)}} \tag{8}$$

如果(x,y)是背景, 则$C(i,j)$值设为 0, 筛选分支的损失可以定义为:

$$L_{\text{cho}} = \frac{-1}{\sum\prod(reg(i,j))} \sum_{\prod(reg(i,j))==1} C(i,j)\times\log M^{\text{cho}}_{w\times h\times 1}(i,j) + (1-C(i,j))\times\log\left(1 - M^{\text{cho}}_{w\times h\times 1}(i,j)\right) \tag{9}$$

本文中的总损失函数定义如下:

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

其中, L_{cls}表示分类的交叉熵损失, L_{reg}和L_{cho}用的是 *IOU* 损失, 常数λ_1和λ_2分为L_{reg}和L_{cho}的权重参数. 在模型训练时, λ_1和λ_2分别被设置为λ_1=1, λ_2=3.

4 实验分析

4.1 实验细节

实验配置: 本文遵循孪生网络的目标跟踪实验, 将数据集图片尺寸预处理为 127×127×3 (宽, 高, 通道数) 的模板图像, 255×255×3 (宽, 高, 通道数) 的搜索图像. 本文硬件平台配置环境为 RTX-2080 Ti GPU (11 GB memory)、Intel Core i9-9900 CPU 以及内存容量为 64 GB.

参数配置: 本文使用在 ImageNet 上预训练的权重初始化骨干网络, 并冻结前两层的参数. 本文一共训练了 20 个 epoch, 前 5 个 epoch 的学习率为 0.001 到 0.005, 后 15 个 epoch 的学习率从 0.005 指数衰减到 0.000 05. 训练过程中, 本文设置批量大小为 32, 随机梯度下降算法 (stochastic gradient descent, SGD) 的动量为 0.9, 权重衰减为 0.005.

4.2 实验数据集

训练数据集: 为了增强网络的泛化性, 本文整个网络在 ImageNet VID[11]、COCO[12]、ImageNet DET[11]和 LaSOT[13]这 4 个数据集上进行端到端的训练. 具体地, 对于每个视频序列, 在视频的第 1 帧裁剪目标模板, 在后续的每帧中采样一次, 并裁剪搜索区域, 其中模板区域大小裁剪为 127×127, 搜索区域大小裁剪为 255×255.

测试数据集: 将本文跟踪器算法与最先进的跟踪器在 OTB100[14]、UAV123[15]、GOT-10k[13]和 LaSOT[16]这 4 个跟踪基准上进行比较. 本文的跟踪算法达到了最先进的结果, 并以每秒 40 帧的速度运行.

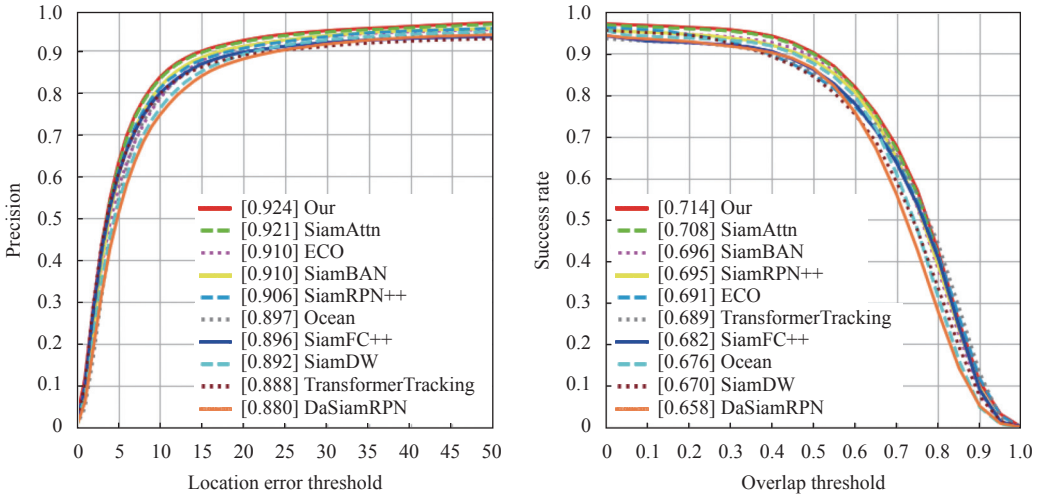
4.3 与现有算法对比分析

本文实验在 4 个大型跟踪数据集上进行, 包括 OTB100[14]、UAV123[15]、LaSOT[16]和 GOT-10k[13], 将本文算法与先进的跟踪器比较和评估. 本文跟踪器获得

了与排名靠前跟踪器相当的结果. 总体来看, 本文算法简单实用, 达到了精度和速度的平衡.

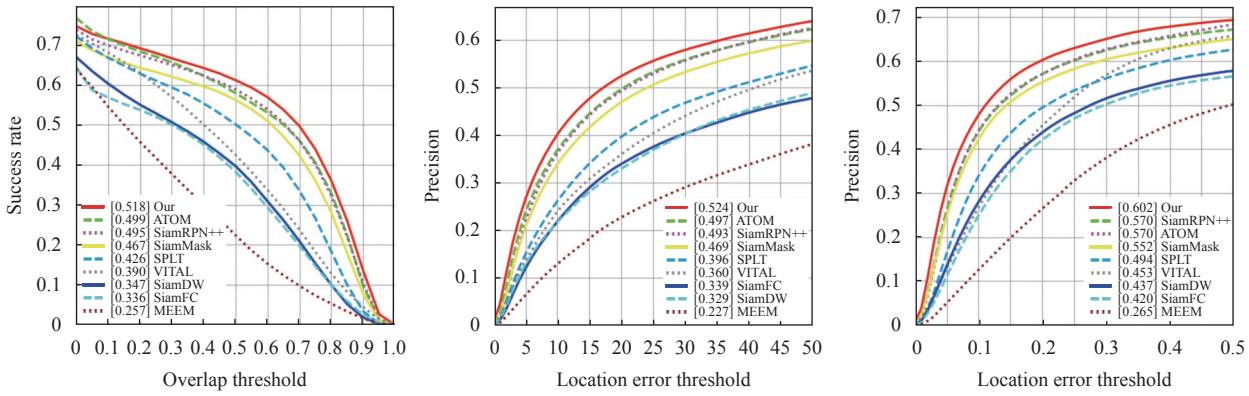
在 OTB100 数据集上对比实验: 图 3 为本文算法与其他主流目标跟踪模型在数据集 OTB100 上的跟踪成功率和精度曲线图. 主流跟踪模型包括 SiamAttn[5]、SiamBAN[17]、SiamRPN++[3]、ECO[18]、SiamFC++[19]、Ocean[7]、SiamDW[20]和 DaSiamRPN[21]. 由图 3 所知, 本文跟踪器达到了最好的性能. 与最近的 SiamBAN 相比, 本文跟踪器的成功率提高了 1.8%, 精度提高了 1.4%.

在 LaSOT 数据集上对比实验: 图 4 为本文算法与现有主流跟踪器在数据集 LaSOT 上比较的结果, 包括、ATOM[22]、SiamRPN++[3]、SiamMask[8]、SiamDW[20]等. ATOM 的研究结果在其作者的网站上提供, 而其他的研究结果则由 LaSOT 的官方网站提供. 与 SiamDW[20]相比, 本文跟踪器的成功率、精度和归一化精度分别提高了 17.1%、19.5% 和 12.9%.



(a) 跟踪精度曲线 (b) 跟踪成功率曲线

图 3 主流跟踪算法在 OTB100 数据集上的成功率图和精度图



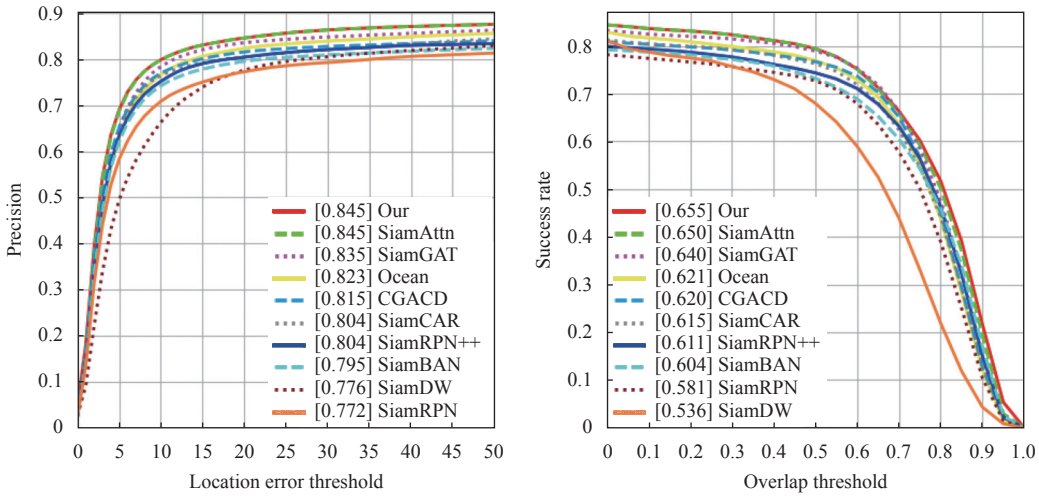
(a) 跟踪成功率曲线 (b) 跟踪精度曲线 (c) 归一化精度曲线

图 4 主流跟踪算法在 LaSOT 数据集上的成功率图、精度图以及归一化精度图

在 UAV123 数据集上对比实验: 图 5 为本文算法与其他主流目标跟踪模型在数据集 UAV123 上的跟踪成功率和精度曲线图. 由图 5 可知本文跟踪模型在成功率和精度指标上分别达到 65.5% 和 84.5%, 均优于其他对比主流跟踪器, 甚至比 SiamAttn[5]跟踪器高出 0.5% 的成功率.

在 GOT-10 数据集对比实验: 表 1 给出了本文跟踪模型在 GOT-10k 数据集上的实验结果, 其中, 粗体代表跟踪器最好性能. 由表 1 可知, 本文方法在指标 AO、$SR_{0.5}$和$SR_{0.75}$上的性能分别为 61.3%, 3.8%, 50.3%.

与 8 种其他的主流跟踪方法相比, 本文算法在指标AO、$SR_{0.5}$和$SR_{0.75}$都有很明显的提高. 尤其是与 SiamCAR[6] 跟踪器相比, 本文跟踪模型在指标AO、$SR_{0.5}$和$SR_{0.75}$分别提高 4.4%、6.8% 和 8.8%. 这得益于本文提出串联互相关模块可以使模板特征和搜索特征深度融合交互, 从而在面对复杂跟踪场景时, 展现出较优的鲁棒性.



(a) 跟踪精度曲线　　(b) 跟踪成功率曲线

图 5　主流跟踪算法在 UAV123 数据集上的成功率图和精度图

表 1　跟踪算法在 GOT-10k 数据集上的结果 (%)

指标	ECO[18]	ATOM[22]	SiamRPN++[4]	SiamFC++[19]	SiamCAR[6]	D3S[23]	DCFST[24]	Ocean[7]	本文算法
AO	31.6	55.6	51.7	59.5	56.9	59.7	**63.8**	61.1	61.3
$SR_{0.5}$	30.9	63.4	61.6	69.5	67	67.6	72.3	72.1	**73.8**
$SR_{0.75}$	11.1	40.2	32.5	47.9	41.5	46.3	49.8	47.3	**50.3**

5　消融实验

为了验证所提出的每个模块的有效性, 在数据集 OTB100[14] 上进行消融实验.

多阶段特征融合消融实验: 为了验证特征提取网络的各阶段特征怎么组合输入串联互相关模块能够实现最好的跟踪性能, 实验对比了 res2, res3, res4 不同阶段特征的组合方式的效果. 如表 2 所示, 其中, 粗体代表最好性能. 把骨干网络最后 3 层特征组合送入串联互相关模块中, 跟踪器实现了最好的跟踪性能.

表 2　本文跟踪器在 OTB100 数据集上特征组合分析

组合方式	成功率	精度
res2	0.580	0.650
res3	0.530	0.590
res4	0.500	0.551
res2+res3	0.583	0.721
res2+res4	0.601	0.737
res3+res4	0.630	0.715
res2+res3+res4	**0.714**	**0.924**

串联互相关模块的消融实验: 表 3 对串联互相关模块进行消融实验来验证其有效性, 其中, 粗体代表最好性能. 可以看出用串联互相关模块替代经典的互相关操作, 成功率和精度分别达到了 71.4% 和 92.4%.

表 3　本文方法在 OT100 数据集上不同互相关操作结果

互相关方法	成功率	精度
传统互相关	0.520	0.657
Up-channel	0.537	0.690
Depth-wise	0.647	0.840
串联互相关模块	**0.714**	**0.924**

6　结论与展望

针对现有互相关方法只对模板特征和搜索特征进行一次交互操作, 使得融合特征图上的目标特征相对粗糙, 不利于跟踪器精确定位的问题. 为此, 本文设计了一个串联互相关模块, 旨在利用现有的深度互相关方法, 对模板特征和搜索特征做多次的互相关操作增强融合特征图上的目标特征, 提升后续分类和回归结果的准确性, 以更少的参数实现了速度和精度之间的平衡. 此外, 为了构建简单通用的跟踪模型, 本文还采用一种无锚、无候选框的目标预测方法, 引导跟踪网络学习到更多明显的特征, 从而提高跟踪模型的性能.

参考文献

1 Xing JL, Ai HZ, Lao SH. Multiple human tracking based on multi-view upper-body detection and discriminative learning. Proceedings of the 20th International Conference on Pattern Recognition. Istanbul: IEEE, 2010. 1698–1701.

2 Bertinetto L, Valmadre J, Henriques JF, *et al*. Fully-convolutional Siamese networks for object tracking. Proceedings of the 2016 European Conference on Computer Vision. Amsterdam: Springer, 2016. 850–865.

3 Li B, Yan JJ, Wu W, *et al*. High performance visual tracking with Siamese region proposal network. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8971–8980.

4 Li B, Wu W, Wang Q, *et al*. SiamRPN++: Evolution of Siamese visual tracking with very deep networks. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4277–4286.

5 Yu YC, Xiong YL, Huang WL, *et al*. Deformable Siamese attention networks for visual object tracking. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6727–6736.

6 Guo DY, Wang J, Cui Y, *et al*. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6268–6276.

7 Zhang ZP, Peng HW, Fu JL, *et al*. Ocean: Object-aware anchor-free tracking. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 771–787.

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

9 Zhang DW, Fu YW, Zheng ZL. UAST: Uncertainty-aware Siamese tracking. Proceedings of the 39th International Conference on Machine Learning. Baltimore: PMLR, 2022. 26161–26175.

10 Liang W, Ding DR, Wei GL. Siamese visual tracking combining granular level multi-scale features and global information. Knowledge-based Systems, 2022, 252: 109435. [doi: 10.1016/j.knosys.2022.109435]

11 Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. International Journal of Computer Vision, 2015, 115(3): 211–252. [doi: 10.1007/s11263-015-0816-y]

12 Lin TY, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. Proceedings of the 13th European Conference on Computer Vision. Zurich: Springer, 2014. 740–755.

13 Huang LH, Zhao X, Huang KQ. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1562–1577. [doi: 10.1109/TPAMI.2019.2957464]

14 Wu Y, Lim J, Yang MH. Online object tracking: A benchmark. Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013. 2411–2418.

15 Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 445–461.

16 Fan H, Lin LT, Yang F, *et al*. LaSOT: A high-quality benchmark for large-scale single object tracking. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 5369–5378.

17 Chen ZD, Zhong BN, Li GR, *et al*. Siamese box adaptive network for visual tracking. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 6667–6676.

18 Danelljan M, Bhat G, Khan FS, *et al*. ECO: Efficient convolution operators for tracking. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6931–6939.

19 Xu YD, Wang ZY, Li ZX, *et al*. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020. 12549–12556.

20 Zhang ZP, Peng HW. Deeper and wider Siamese networks for real-time visual tracking. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4586–4595.

21 Zhu Z, Wang Q, Li B, *et al*. Distractor-aware siamese networks for visual object tracking. Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich: Springer, 2018. 103–119.

22 Danelljan M, Bhat G, Khan FS, *et al*. ATOM: Accurate tracking by overlap maximization. Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 4655–4664.

23 Lukežič A, Matas J, Kristan M. D3S —A discriminative single shot segmentation tracker. Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 7131–7140.

24 Zheng LY, Tang M, Chen YY, *et al*. Learning feature embeddings for discriminant model based tracking. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 759–775.

(校对责编: 孙君艳)