

文章编号:1007-2780(2024)08-1079-11

基于双分支在线优化和特征融合的视频目标跟踪算法

李新鹏¹, 王 鹏^{1*}, 李晓艳¹, 孙梦宇¹, 陈遵田², 郜 辉¹

(1. 西安工业大学电子信息工程学院, 陕西 西安 710021;

2. 西安机电信息技术研究所, 陕西 西安 710065)

摘要:针对D3S算法对跟踪目标的判别能力不足的问题,提出了一种基于双分支在线优化和特征融合的视频目标跟踪算法。首先,构建双分支的在线优化分类器,实现对目标的二次定位,得到更准确的目标位置响应图;然后,在特征层上实现响应图与搜索特征的融合,并通过encoder模块促进融合过程,进一步突出跟踪目标的特征;最后,通过encoder模块实现模板特征的更新,拟合特征间的差异,提高分割模块的判别能力。在VOT2018和UAV123数据集进行实验,改进后算法与原算法相比,在VOT2018数据集上EAO提高了2.9%,在UAV123数据集上成功率提高了2.4%,准确率提高了2.9%。实验结果表明,本文方法提高了算法的判别能力,并且进一步提升了精度和鲁棒性。

关键词:视频目标跟踪;目标分割;在线优化;特征融合;注意力机制

中图分类号:TP391 文献标识码:A doi:10.37188/CJLCD.2023-0256

Video object tracking algorithm based on dual-branch online optimization and feature fusion

LI Xinpeng¹, WANG Peng^{1*}, LI Xiaoyan¹, SUN Mengyu¹, CHEN Zuntian², GAO Hui¹

(1. School of Electronics and Information Engineering, Xi'an Technological University, Xi'an 710021, China;

2. Xi'an Institute of Electromechanical Information Technology, Xi'an 710065, China)

Abstract: In response to the issue of inadequate discrimination capability in the D3S algorithm for tracking

收稿日期:2023-07-25;修订日期:2023-08-08.

基金项目:国家自然科学基金(No.62171360);陕西省科技厅重点研发计划(No.2024GX-YBXM-162);陕西省电子设备智能测试与可靠性评估工程技术研究中心项目(2023-ZC-GCZX-0047);2022年度陕西高校青年创新团队项目;山东省智慧交通重点实验室(筹)项目;2023年陕西省高校工程研究中心项目;西安市军民两用智能测评技术重点实验室项目

Supported by National Natural Science Foundation of China (No.62171360); Key Research and Development Program of Shaanxi Provincial Science and Technology Department (No.2024GX-YBXM-162); Project of Engineering Technology Research Center of Shaanxi Province for Intelligent Testing and Reliability Evaluation of Electronic Equipments (No.2023-ZC-GCZX-0047); Project of 2022 Shaanxi University Youth Innovation Team; Project of Shandong Key Laboratory of Smart Transportation (Preparation); Project of 2023 Shaanxi University Engineering Research Center; Project of Xi'an Key Laboratory of Military Civilian Dual-Use Intelligent Evaluation Technology

*通信联系人, E-mail: wp_xatu@163.com

target, a video object tracking algorithm based on dual-branch online optimization and feature fusion is proposed. Firstly, a dual-branch online optimization classifier is constructed, which achieves secondary location of the target, resulting in a more accurate target position response map. Secondly, the fusion of the response map and search features is realized on the feature layer, and the encoder module promotes the fusion process, further highlighting the features of tracking target. Finally, by updating the template features with the encoder module, the differences between features are fitted, thereby enhancing the discriminative capability of the segmentation module. Experimental evaluations are conducted on the VOT2018 and UAV123 datasets. Compared with the original algorithm, the improved algorithm improves EAO by 2.9% on the VOT2018 dataset, increases success rate by 2.4% and accuracy by 2.9% on the UAV123 dataset. The experimental results demonstrate that the method in this paper improves the algorithm's discriminative ability and further improves accuracy and robustness.

Key words: object tracking; object segmentation; online optimization; feature fusion; mechanism of attention

1 引 言

视频目标跟踪是计算机视觉领域的基础任务之一。在视频目标跟踪的任务中,只需在视频序列的第一帧中用边界框给定跟踪目标的状态,就可以在后续帧中完成对跟踪目标的持续预测。视频目标跟踪目前广泛应用于多种领域,如视频监控、无人驾驶、人机交互、机器人视觉等^[1]。然而,由于视频序列中存在光照变化、运动模糊、遮挡等干扰因素,视频目标跟踪算法对视频序列中目标的跟踪精度有很大影响。因此,开发一个高精度的视频目标跟踪算法十分具有挑战性。

随着深度学习技术的不断发展,深层卷积网络因其强大的特征提取能力和泛化能力在视频目标跟踪领域得到了广泛的应用^[2-4]。SiamFC^[5]算法将跟踪问题转化为相似性学习问题,通过搜索区域与模板区域之间的匹配来实现跟踪,取得了精度与速度的平衡。SiamRPN^[6]算法中引入了区域候选网络 RPN,通过分类分支实现对目标和背景的分类,并通过回归分支对边界框的尺寸和位置进行回归。SiamRPN++^[7]算法在训练过程中采用位置均衡的采样策略解决了平移不变性问题,并将深层网络 Resnet50 引入跟踪算法。ATOM^[8]算法设计了在线优化的分类器来估计目标位置,通过实时更新的分类器来适应跟踪过程中目标的外观变化。DiMP^[9]算法改进了在线优化的分类器,将分类器的超参数从人工预设改为网络自适应学习,使分类器的泛化能力得到提高。TransT^[10]算法设计了结合上下文增强和交叉特

征增强模块的 transformer^[11-12]特征融合网络,有效地融合了模板特征和搜索区域中的感兴趣区域特征。ToMP^[13]算法提出基于 transformer 的模型预测,能够捕获全局关系,并通过模型预测来实现边框回归。UTT^[14]算法提出一种基于 transformer 的跟踪器来实现单目标跟踪与多目标跟踪的统一,通过目标特征和跟踪帧特征之间的相关性定位目标。

上述算法虽然能生成高质量的边界框,但是边界框只能简单表示目标的外观,还会引入背景信息,阻碍了算法精度和鲁棒性的提升。因此,近几年开始出现将分割方法引入跟踪领域的趋势。SiamMask^[15]算法在 RPN 结构的基础上额外引入了分割分支,并使用分割数据集 YouTube-VOS^[16]进行训练以生成目标的二进制分割掩码。AR^[17]算法采用角点预测的方式获取预测的边界框并用辅助的掩码头部来辅助训练,整个模块独立训练,能直接应用于现有的跟踪器中。RTS^[18]算法通过包含历史信息的样本池生成分割掩码,并将实例定位分支获取的位置信息引入编码过程来改进分割掩码。D3S^[19]算法通过两个互补的网络分支来生成分割掩码。GIM 分支中的分割网络能将搜索区域中所有的相似目标通过掩码分割出来,GEM 分支的在线学习分类器能预测出准确的跟踪目标位置,并作为多个掩码的选择器。

尽管 D3S 算法在跟踪性能上取得了很大进展,但是 D3S 算法也存在部分问题:该算法的分割网络依赖于分类器模块提供的目标位置信息,

当分类器预测的位置信息出错时会导致错误的分割结果;该算法仅在掩码层上进行了响应图与掩码的拼接,没有充分利用获取的位置信息,使算法的判别能力不足;该算法中的模板特征在第一帧获得后不再进行更新,未更新的模板特征会导致分割网络的判别能力不足。

为了解决上述问题,本文在D3S算法的基础上提出了基于双分支在线优化和特征融合的视频目标跟踪算法。首先,构建了双分支的在线优化分类器,获取两个阶段的响应图并加权融合,输出更精准的位置信息;然后,在特征层融合了搜索特征与响应图,并通过transformer模型中的encoder模块来促进特征融合过程,突出跟踪目标特征并抑制相似干扰目标的特征;最后,通过encoder模块更新模板特征,拟合模板特征与搜索特征的空间信息差异,进一步提高分割模块的判别能力。

2 D3S 算法概述

D3S算法借鉴了ATOM、VideoMatch^[20]和U-net^[21]等算法,提出了一种判别式单样本分割跟踪器(Discriminative Single Shot Segmentation Tracker)。D3S算法结构如图1所示。该算法将Resnet50作为主干网络,经主干网络提取后的特征会输入到GIM分支与GEM分支。GEM分支的在线学习分类器能预测出目标在搜索区域中的位置响应图;GIM分支的分割网络则通过相似性分割的方法从搜索特征与模板特征中获得低分辨率的分割掩码。响应图与分割掩码经过通道拼接后输入到U-Net结构的上采样网络中,通过连续的上采样以及与Resnet50网络浅层特征的特征融合,输出与搜索区域相同尺寸的高精度掩码,再经过椭圆拟合法将掩码转换成对应的旋转边界框,作为跟踪算法的输出结果。

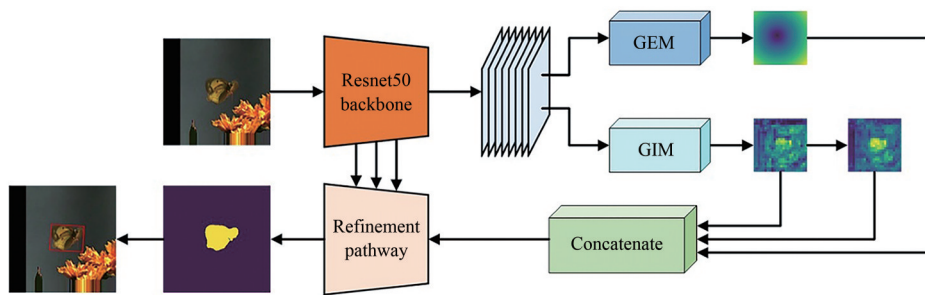


图1 D3S算法网络结构

Fig. 1 Network architecture of D3S algorithm

3 基于双分支在线优化和特征融合的视频目标跟踪算法

针对D3S算法中特征对跟踪目标的判别能力不足的问题,本文提出了基于双分支在线优化和特征融合的视频目标跟踪算法,算法结构如图2所示。图2中下方的矩形框部分为本文中双分支在线优化部分,该模块在ATOM算法提出的分类器的基础上,构建了本文所提出的双分支结构。该模块通过两组独立学习的分类器从搜索区域中获取二阶段的目标位置响应图,并通过加权融合得到高斯响应图,输出更准确的目标位置给后续的分割网络。图2中上方的矩形框表示本文

中构建的基于transformer的特征融合与模板更新部分,其中引入了transformer中的主流encoder结构。transformer通常通过自注意力机制来更新输入特征,而本文以互注意力机制对输入的成对特征进行权值更新,实现了本文所提出的搜索特征与响应图的融合以及模板特征的更新。该模块在搜索特征与模板特征输入相似性分割模块之前,对搜索特征与响应图进行特征融合,再通过encoder模块来促进搜索特征与响应图的融合,使搜索特征中的跟踪目标特征更加突出,再将更新后的搜索特征与模板特征输入encoder模块进行模板特征的更新,使模板特征适应更新后的跟踪目标特征,算法的判别能力得到提高。

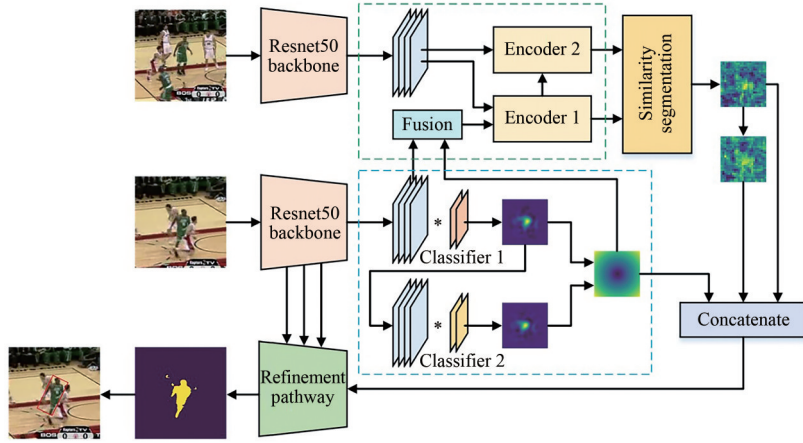


图2 本文提出算法的网络结构

Fig. 2 Network architecture of the algorithm proposed in this article

3.1 双分支在线优化

针对D3S算法中的分类器提供位置信息不够准确的问题,本文构建了双分支的在线优化分类器,通过两组独立训练的分类器来实现对跟踪目标的二次定位,并通过加权融合响应图来提供更精确的目标位置信息。

本文所构建的双分支分类器均为双层的全卷积网络构成,可以定义为:

$$f(x; w) = \phi_2(w_2 * \phi_1(w_1 * x)), \quad (1)$$

式中: x 为输入分类器的主干特征, $w = \{w_1, w_2\}$ 为分类器的网络参数, $*$ 表示卷积操作, ϕ_1 、 ϕ_2 表示激活函数。

分类器的训练方法源于相关滤波器DCF,基于 L_2 损失完成了对目标函数的构建,并在跟踪过程中通过反向传播公式实现对分类器的训练更新,使分类器能够实现快速的二分类。具体函数如式(2)所示:

$$L(w) = \sum_{j=1}^m \gamma_j \|f(x_j; w) - y_j\|^2 + \sum_k \lambda_k \|w_k\|, \quad (2)$$

式中: x_j 为主干特征, γ_j 为权重, w_1 为 1×1 、64通道的卷积层, w_2 为 4×4 、64通道的卷积层, y_j 为标准边界框, λ_k 为权重参数 w_k 的正则化项。

在算法的初始化阶段,模板图片经过数据扩充以及主干网络提取后得到30组的模板特征。这些模板特征与对应的目标边界框将用于分类器 f_1 的初始化。分类器 f_1 完成初始化后将参数传递给分类器 f_2 ,减少算法初始化的时间。

在算法的推理阶段,以上一帧的预测位置为

初始位置进行裁剪获取搜索特征 x_1 ,搜索特征 x_1 经过分类器 f_1 的二分类后得到一阶段的位置响应图。该响应图最大值的位置将作为偏移参数,对初始位置进行偏移。后续的跟踪步骤都将以偏移后的位置为中心进行,保证跟踪目标处于搜索区域的中心附近。

然后以偏移位置为中心重新裁剪获得搜索特征 x_2 ,输入分类器 f_2 预测出二阶段的位置响应图。获得两组响应图后,需要通过Argmax方法转换成高斯响应图,高斯响应图能更直观地表示距离度量信息。

经过实验发现,两个高斯响应图经过加权融合后输入后续的分割模块能取得更好的跟踪效果,其中效果最佳的加权融合方法的公式如式(3)所示:

$$\text{map}_x = \frac{\text{map}_1 + \max \cdot \text{map}_2}{1 + \max}, \quad (3)$$

式中: map_1 是一阶段的高斯响应图,由于位置偏移, map_1 将设为以中心为最大值的高斯响应图; map_2 为二阶段的高斯响应图; $\max$ 为二阶段响应图中置信度的最大值; $\cdot$ 表示矩阵数乘。

加权融合后的高斯响应图能进一步提高跟踪精度和鲁棒性,在目标受到相似干扰物遮挡时也能更好地分辨出目标与相似干扰物。

在线优化阶段,初始区域裁剪的搜索特征和目标边界框作为样本在线训练分类器 f_1 ,而分类器 f_2 使用偏移后区域的搜索特征和边界框进行在线训练。双分类器的在线优化均为每10帧进行一次。

3.2 基于 transformer 的特征融合

D3S 算法仅在掩码层上通过响应图与分割掩码之间的通道拼接来给后续分割模块提供目标位置信息,没有充分利用响应图提供的定位信息。因此本文在特征层上构建了基于 transformer 的特征融合模块,更好地利用 3.1 节中获取的加权融合响应图。

为了对搜索特征与高斯响应图进行特征融合。首先将分类器中获得的高斯响应图通过 softmax 层进行响应图的数值归一化。然后将响应图与搜索特征进行矩阵点乘,通过响应图包含的模板位置信息对搜索特征的特征信息进行空间上的筛选。

经过与响应图的特征融合后,搜索特征中包含的特征会根据其离目标中心位置的距离受到不同程度的抑制,导致搜索特征丢失了一部分的特征信息,因此要将融合后的搜索特征与模板特征输入到 encoder 模块中进行空间信息重构。

视频目标跟踪领域中的 transformer 模型通常为 N 层 encoder 和 decoder 堆叠而成。本文中使用了由多头注意力模块和前馈网络构成的单层 encoder,能学习输入特征间的时空上下文信息来更新输入的 Query 值。encoder 的具体结构如图 3 所示。

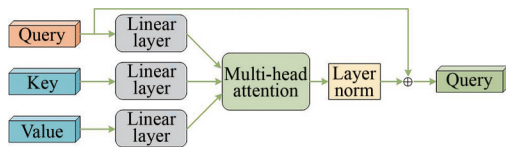


图3 encoder模块结构

Fig. 3 Structure of encoder module

特征输入 encoder 前需要降维后两维,表示为 $F_z, F_x \in R^{n \times d}$ 。 F_z 为降维后的模板特征, F_x 为降维后的搜索特征, n 为特征的总像素数, d 为通道数。encoder 模块的输入如式(4)所示:

$$Q = F_z W_{q1}, K = F_z W_{k1}, V = F_z W_{v1}, \quad (4)$$

其中: Q, K, V 分别对应图 3 中的 Query、Key、Value 值。 $W_{q1}, W_{k1}, W_{v1} \in R^{d \times d_a}$ 由一层可学习的全连接层构成,能将输入特征投影到 $d_a - d$ 的空间上。

encoder 模块的输入特征会进行如式(5)所示的操作:

$$F'_x = F_x + \psi \left(\text{softmax} \left(\frac{Q(K)^T}{\sqrt{d_a}} \right) (V) \right), \quad (5)$$

其中: softmax 用于对注意力机制进行归一化, ψ 为输入维度 d_a 和输出维度 d 的线性层, F'_x 为权值更新后的模板特征。式(5)中采用了残差结构使模型的收敛更稳定。

通过 encoder 的互注意力机制可以实现特征间的信息交互,帮助搜索特征补全在与响应图融合过程丢失的边缘特征,使搜索特征在突出跟踪目标特征的同时能够实现对跟踪目标的高精度分割。

3.3 基于 transformer 的模板特征更新

D3S 算法中没有采用模板更新策略,模板特征在第一帧获取后不再更新,因此本文构建了基于 transformer 的模板特征更新模块,使模板特征在跟踪过程中更新,更好地适应特征融合后的搜索特征,提高分割网络对跟踪目标的判别能力。

首先将 3.2 节中模板特征与特征融合后的搜索特征进行降维,表示为 $F_z, F'_x \in R^{n \times d}$ 。降维后的特征输入到 encoder 中, encoder 与 3.2 节中结构相同,仅改变了 encoder 的输入参数:

$$Q' = F_z W_{q2}, K' = F'_x W_{k2}, V' = F'_x W_{v2}, \quad (6)$$

其中: Q', K', V' 为 3.3 节中 encoder 模块中的 Query、Key、Value 输入。 $W_{q2}, W_{k2}, W_{v2} \in R^{d \times d_a}$ 为一层可学习的全连接层。

encoder 模块的输入特征会进行如式(7)所示的操作:

$$F'_z = F_z + \psi \left(\text{softmax} \left(\frac{Q'(K')^T}{\sqrt{d_a}} \right) (V') \right). \quad (7)$$

在后续相似性分割模块中,模板特征将在搜索特征上进行相似性搜索,而 encoder 关注了模板特征和搜索特征之间的相似性分数,并通过全局空间信息的加权融合进一步学习。模板特征经过 encoder 的权重更新后可以在全局空间信息上与搜索特征更加相似,可以帮助分割网络更好地生成跟踪目标对应的分割掩码,使分割网络的判别能力得到提高。

4 实验与分析

4.1 实验环境和训练细节

本文所提出的双分支在线优化部分为在线

学习模块,在推理过程中完成初始化和优化过程,因此不参与训练过程。基于 transformer 的特征融合以及模块更新模块在分割网络上进行了改进,需要进行离线训练,训练的实验环境如下:实验操作系统为 Liunx18.04, GPU 为 NVIDIA Quadro P6000 24 GB, CPU 为 Inter(R) Xeon(R) Gold 5118 CPU @ 2.30 GHz, 内存 64 GB。使用 Pytorch 深度学习框架,编程语言为 Python3.6。

本文算法的训练数据集沿用了 D3S 算法所使用的 Youtube-VOS 视频分割数据集,其中训练集以及验证集采用了该分割数据集中的默认配置。在训练过程中,优化器为 Adam;损失函数为交叉熵损失;训练共 40 个 Epoch,每个 Epoch 进行 1 000 次迭代;BatchSize 设为 64;初始学习率为 10^{-3} ,学习率每经过 15 个 Epoch 衰减为 0.2;验证集每 5 个 Epoch 进行执行一次,每个 Epoch 迭代 50 次。图 4 为本文算法在训练过程中获得的训练集以及验证集损失曲线图。

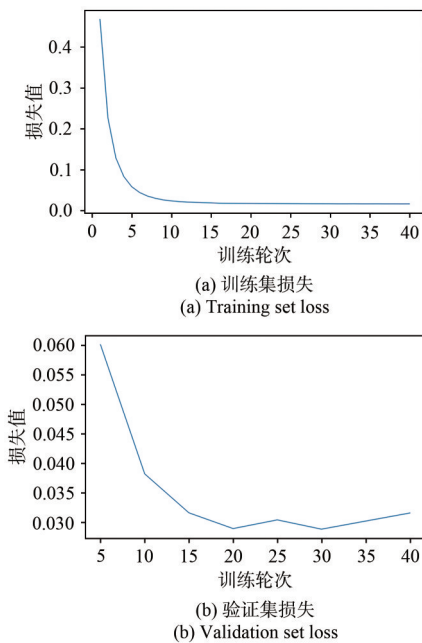


图 4 本文算法的训练损失

Fig. 4 Training loss of the algorithm proposed in this article

4.2 定量分析

4.2.1 VOT 数据集

VOT2018^[22]是2018年视觉跟踪挑战赛使用的测试数据集,其中包含了旋转边界框标注的60个视频序列和6种视觉属性,并加入了跟踪失

败后5帧重启的机制。该数据集可以公正地评价跟踪器的性能。

VOT 数据集使用 EAO、Accuracy 和 Robustness 这3项指标来评估跟踪算法。其中 Accuracy 和 Robustness 表示跟踪结果的平均重叠率和跟踪失败次数,如式(8)和(9)所示。平均重叠期望(EAO)是 VOT 数据集的核心评价指标,表示跟踪结果的非重置重叠的期望值,如式(10)所示。EAO 和 Accuracy 越高,算法的精度越高;Robustness 越低,算法的鲁棒性越好。

$$\phi_t = \frac{A_t \cap A_g}{A_t \cup A_g}, \quad (8)$$

$$\rho_i(i) = \frac{1}{N_{re}} \sum_{k=1}^{N_{re}} F(i), \quad (9)$$

$$\hat{\phi} = \frac{1}{N_h - N_l} \sum_{N_l=N_i}^{N_h} \hat{\phi}_{N_i}, \quad (10)$$

其中: A_t 为第 t 帧的预测框, A_g 为第 t 帧的真实框, ϕ_t 为跟踪器在第 t 帧的重叠率, N_{re} 为测试次数, $F(i)$ 为第 i 个跟踪器的跟踪失败次数, $\rho_i(i)$ 为第 i 个跟踪器的 Robustness, N_h 、 N_l 为测试序列选取的片段长度, $\hat{\phi}_{N_i}$ 为测试序列选取片段的精度, $\hat{\phi}$ 为跟踪器在该测试序列的 EAO。

为了验证本文提出算法的跟踪性能,在 VOT2018 数据集上将本文算法与基准算法 D3S 以及 SiamRPN++、SiamRPN、ATOM、SiamMask 等算法进行对比实验,结果如表 1 所示,最优结果加粗表示,次优结果用下划线表示。其中本文提出的算法相较于基准算法 D3S 在 EAO 指标上提高了 2.9%,在 Accuracy 和 Robustness 指标上分别提高了 2.0% 和 1.0%,整体指标相较于其他算法有很大的领先。

为了展示对比算法在各种视觉属性下的跟踪性能,在 VOT2018 数据集上进行了各种视觉

表 1 VOT2018 数据集的实验结果

Tab. 1 Experimental results on VOT2018 dataset

Tracker	EAO	Accuracy	Robustness
Ours	0.518	0.660	0.140
D3S	<u>0.489</u>	<u>0.640</u>	<u>0.150</u>
SiamRPN++	0.414	0.600	0.234
SiamRPN	0.383	0.588	0.276
ATOM	0.401	0.590	0.204
SiamMask	0.380	0.609	0.276

属性下的算法对比实验,实验结果如表2、表3所示,其中最优结果加粗表示,次优结果用下划线表示。表3的数值为算法的失败次数,数值越小鲁棒性越好。从表2可以看出,在遮挡场景下,D3S算法的精度指标在所有对比算法中是最低的,说明D3S算法区分复杂背景和跟踪目标的能力较

差,对跟踪目标的判别能力不足。而改进后的本文算法精度在所有视觉属性挑战下,除了光照变换场景精度有一定下降,其他场景的精度结果均为第一,在鲁棒性结果中本文算法均为第一。实验结果表明,本文提出的算法大幅提升了遮挡场景下的跟踪效果,提高了对跟踪目标的判别能力。

表2 VOT2018数据集视觉属性实验的精度结果

Tab. 2 Precision results of visual attribute experiment on VOT2018 dataset

Tracker	运动模糊	光照变化	快速运动	遮挡	尺度变化	无标签
Ours	0.655 0	<u>0.621 4</u>	0.642 0	0.508 1	0.662 3	0.679 9
D3S	<u>0.641 2</u>	0.649 0	<u>0.618 3</u>	0.444 1	<u>0.660 5</u>	<u>0.673 5</u>
ATOM	0.598 4	0.590 4	0.599 7	0.503 7	0.604 5	0.581 6
SiamMask	0.620 3	0.615 8	0.620 0	0.467 8	0.628 0	0.607 9
SiamRPN	0.591 4	0.564 8	0.588 8	0.501 6	0.574 4	0.589 3
SiamRPN++	0.620 1	0.604 2	0.602 8	<u>0.507 5</u>	0.575 8	0.603 5

表3 VOT2018数据集视觉属性实验的鲁棒性结果

Tab. 3 Robustness results of visual attribute experiment on VOT2018 dataset

Tracker	运动模糊	光照变化	快速运动	遮挡	尺度变化	无标签
Ours	15.0	0.0	5.0	15.0	3.0	6.0
D3S	<u>16.0</u>	0.0	<u>6.0</u>	16.0	3.0	11.0
ATOM	18.8	0.8	11.4	<u>15.6</u>	5.6	9.4
SiamMask	29.0	0.0	17.0	18.0	6.0	<u>9.0</u>
SiamRPN	29.0	2.0	19.0	17.0	10.0	11.0
SiamRPN++	24.0	1.0	14.0	17.0	8.0	10.0

4.2.2 UAV数据集

UAV123^[23]是一个无人机低空拍摄的短时跟踪数据集,包含矩形边界框标注的123个视频序列,并根据OTB^[24]数据集所包含的视觉属性进行了注释。该数据集可以评价算法在无人机拍摄场景下的跟踪器性能。

UAV123数据集使用成功率与准确率两种评价指标来对算法的精度和鲁棒性进行评价。成功率表示跟踪结果的平均重叠率,计算方法与4.2.1节中的Accuracy指标相同。准确率表示跟踪器在测试序列下的中心位置误差,可以表示预测生成的边界框与标准边界框的位置误差,其计算如式(11)所示:

$$d = \sqrt{\|c - c_t\|^2}, \quad (11)$$

其中: c 表示预测的目标位置, c_t 是人工标定的目标位置, d 表示中心位置误差值。

在UAV123数据集上将本文算法与基准

算法D3S以及SiamRPN++、DaSiamRPN^[25]、UPDT^[26]、ECO^[27]等算法进行对比实验,结果如图5所示。本文算法相较于基准算法D3S在成功率上提高了2.4%,在准确率上提高了2.9%,且相较于次优的SiamRPN++算法实现了较小的领先,这主要是因为UAV数据集依然使用常规的矩形边界框进行标注,且存在一定量的小目标跟踪场景,基于分割网络的跟踪算法在其中无法发挥较好的效果。

为了展示对比算法在遮挡以及相似目标等视觉属性下的跟踪性能,在UAV123数据集上进行了对应视觉属性下的算法对比实验,实验结果如图6所示。从图6可以看出,D3S算法整体成功率与准确率均处于对比算法的下游水平,说明D3S算法在区分目标与复杂背景以及相似目标的能力较差,对跟踪目标的判别能力不足。而本文算法在遮挡场景下,成功率与准确率相较于原

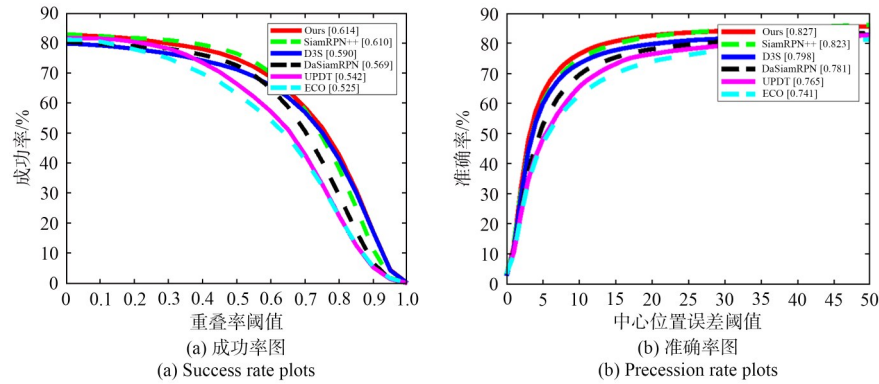


图5 UAV123数据集的实验结果

Fig. 5 Experimental results of UAV123 dataset

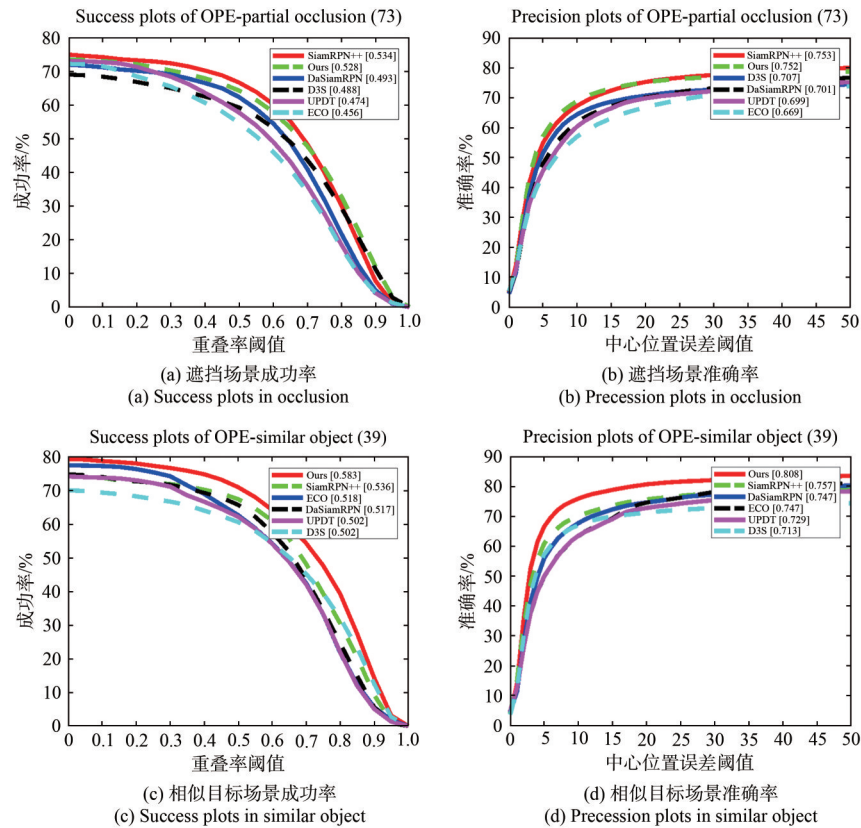


图6 UAV123数据集下遮挡、相似目标属性的实验结果。

Fig. 6 Experimental results of occlusion and similar object attribute in UAV123 dataset

算法提高了4.0%和4.5%,但是稍微落后于最优的SiamRPN++算法。而在相似目标属性下,本文算法的成功率与准确率比原算法提高了9.5%与8.1%,并在所有对比算法中实现了最优。从图6可以看出,本文算法在遮挡以及相似目标场景下的算法精度得到了很大的提高,其中相似目标场景下的提升相对更高,说明本文提出的方法能有效提高算法的判别能力,使算法能更

好地区分跟踪目标与相似干扰目标。

4.3 消融实验

为了验证本文提出方法的有效性,将本文中提出的3个模块在VOT2018数据集上进行了消融对比实验,实验结果如表4所示,最优结果加粗表示,次优结果用下划线表示。其中BF表示双分支在线优化,FF表示基于transformer的特征融合,TU表示基于transformer的模板更新。

表 4 VOT2018数据集上的消融实验结果

Tab. 4 Ablation experiments results on VOT2018 dataset

消融实验	Accuracy	Robustness	EAO
baseline	0.640	0.150	0.489
baseline+BF	0.655	0.164	0.498
baseline+FF	0.649	0.155	0.504
baseline+BF+FF	<u>0.657</u>	0.160	0.506
baseline+FF+TU	0.651	<u>0.145</u>	<u>0.509</u>
baseline+BF+FF+TU	0.660	0.140	0.518

由表 4 可以看出,引入双分支在线优化后,EAO 与 Accuracy 提高到 0.489 与 0.655,但是鲁棒性降到了 0.164,说明引入双分支分类器得到更精确的位置信息后能提高算法的精度,但是更大的位置偏移也会带来额外的跟踪失败。加入特征融合模块后,EAO 与 Accuracy 提升到 0.504 与 0.649,说明促进响应图与特征的融合能一定程度上提高算法的精度,但是只引入特征融合模块会导致搜索特征的变化过大,导致算法的鲁棒性下降。而同时引入两个模块后,算法精度提高有限,且鲁棒性相较于原算法有一定的下降,说明两个模块之间存在耦合问题。模板更新是在特征融合基础上的改进,做了特征融合与模板更新两个模块的组合实验。两个模块同时加入后,算法的精度和鲁棒性都有了一定的提高,说明特征融合与模板更新同时加入后,跟踪目标特征突出后算法的精度有了进一步的提高,而对模板特征的更新也使算法稳定性更好,减少了跟踪失败的次数。最后,3 个模块同时引入后,3 项指标都达到了最优,说明 3 个模块实现了互补,使算法的跟踪精度和鲁棒性得到了最大的提升。

4.4 实时性实验

本文在 UAV123 数据集上进行了实时性实验,对本文算法与原算法的推理速度进行了对比,具体实验结果如表 5 所示。可以看出,本文所提出的算法相较于 D3S, FPS 数值有一定的下降,因为本文算法所使用的双分支分类器以及基于 transformer 的特征融合和模板更新模块均引

入了额外的网络结构,增加了网络的参数量,对算法的速度造成了一定的影响。

4.5 定性分析

本文在 VOT2018 数据集中选取了 4 个视频序列进行跟踪结果可视化实验,结果如图 7 所示,其中图 7(a)和(b)为遮挡属性下的序列,图 7(c)和(d)为相似目标属性下的序列。从跟踪结果可以看出,本文算法整体生成的边界框准确性相较于原算法有一定的提高。其中 basketball, fish3 序列下的跟踪结果表明,本文算法所生成的边界框在遮挡场景下能更好地划分出跟踪目标的轮廓;而 bolt2 和 girl 序列的跟踪结果图表明,本文算法在存在相似目标的场景下能更好地维持对正确目标的跟踪,并在跟踪到错误目标后也能够重新实现对原目标的跟踪。通过整体的实验结果可以看出,本文算法提高了对跟踪目标的判别能力,能更好地从复杂背景和相似目标中区分跟踪目标,也使得算法生成的边界框准确性更好。

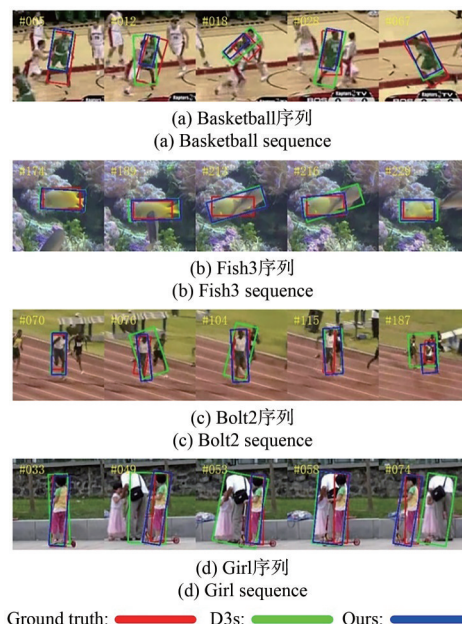


图 7 VOT2018数据集 4 个序列的跟踪结果可视化

Fig. 7 Visualized results of four sequences on VOT2018 dataset

表 5 UAV123数据集上的实时性实验结果

Tab. 5 Real time experiments results on VOT2018 dataset

实时性实验	FPS
D3S	22.9
Ours	17.6

5 结 论

针对 D3S 算法对跟踪目标的判别能力不足的问题,本文提出了一种基于双分支在线优化和特征融合的视频目标跟踪算法。首先构建了

双分支的在线优化分类器,通过双分类器获得加权融合的响应图,为后续的分割模块提供更高精度的位置信息;然后通过 encoder 中的互注意力机制,促进位置信息与搜索特征的融合,进一步突出跟踪目标的特征;最后通过 encoder 对模板特征进行权重更新,拟合了特征间的空间信息差异,使分割掩码判别效果更好。在公开数

据集 VOT2018 和 UAV123 上进行了实验分析, VOT2018 数据集上本文算法的 EAO、Accuracy 和 Robustness 指标分别达到了 0.518、0.660 及 0.140,相较于原算法分别提高了 2.9%、2.4%、1.0%,取得了优异的表现。实验结果表明,本文算法有效地提升了算法对目标的判别能力,并使算法的整体精度和鲁棒性有了较大的提升。

参 考 文 献:

- [1] 苗宗成,高世严,贺泽民,等. 基于孪生网络的目标跟踪算法[J]. 液晶与显示,2023,38(2):256-266.
MIAO Z C, GAO S Y, HE Z M, *et al.* Single-objective tracking algorithm based on siamese networks [J]. *Chinese Journal of Liquid Crystals and Displays*, 2023, 38(2): 256-266. (in Chinese)
- [2] 李玺,查宇飞,张天柱,等. 深度学习的目标跟踪算法综述[J]. 中国图象图形学报,2019,24(12):2057-2080.
LI X, ZHA Y F, ZHANG T Z, *et al.* Survey of visual object tracking algorithms based on deep learning [J]. *Journal of Image and Graphics*, 2019, 24(12): 2057-2080. (in Chinese)
- [3] 韩瑞泽,冯伟,郭青,等. 视频单目标跟踪研究进展综述[J]. 计算机学报,2022,45(9):1877-1907.
HAN R Z, FENG W, GUO Q, *et al.* Single object tracking research: A survey [J]. *Chinese Journal of Computers*, 2022, 45(9): 1877-1907. (in Chinese)
- [4] 刘艺,李蒙蒙,郑奇斌,等. 视频目标跟踪算法综述[J]. 计算机科学与探索,2022,16(7):1504-1515.
LIU Y, LI M M, ZHENG Q B, *et al.* Survey on video object tracking algorithms [J]. *Journal of Frontiers of Computer Science and Technology*, 2022, 16(7): 1504-1515. (in Chinese)
- [5] BERTINETTO L, VALMADRE J, HENRIQUES J F, *et al.* Fully-convolutional Siamese networks for object tracking [C]//*Proceedings of the European Conference on Computer Vision*. Amsterdam: Springer, 2016: 850-865.
- [6] LI B, YAN J J, WU W, *et al.* High performance visual tracking with Siamese region proposal network [C]//*Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE, 2018: 8971-8980.
- [7] LI B, WU W, WANG Q, *et al.* Siamrpn++: Evolution of Siamese visual tracking with very deep networks [C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 4277-4286.
- [8] DANELLJAN M, BHAT G, KHAN F S, *et al.* ATOM: Accurate tracking by overlap maximization [C]//*Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 4655-4664.
- [9] BHAT G, DANELLJAN M, VAN GOOL L, *et al.* Learning discriminative model prediction for tracking [C]//*Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision*. Seoul: IEEE, 2019: 6181-6190.
- [10] CHEN X, YAN B, ZHU J W, *et al.* Transformer tracking [C]//*Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021: 8122-8131.
- [11] VASWANI A, SHAZEER N, PARMAR N, *et al.* Attention is all you need [C]//*Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach: Curran Associates Inc., 2017: 6000-6010.
- [12] 张印辉,庄宏,何自芬,等. 氨气泄漏混洗自注意力轻量化红外检测[J]. 中国光学(中英文),2023,16(3):607-619.
ZHANG Y H, ZHUANG H, HE Z F, *et al.* Lightweight infrared detection of ammonia leakage using shuffle and self-attention [J]. *Chinese Optics*, 2023, 16(3): 607-619. (in Chinese)
- [13] MAYER C, DANELLJAN M, BHAT G, *et al.* Transforming model prediction for tracking [C]//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 2022: 8721-8730.
- [14] MA F, SHOU M Z, ZHU L C, *et al.* Unified transformer tracker for object tracking [C]//*Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE, 2022: 8771-8780.

- [15] WANG Q, ZHANG L, BERTINETTO L, *et al.* Fast online object tracking and segmentation: A unifying approach [C]//*Proceedings of the 2019 IEEE/CVF conference on Computer Vision and Pattern Recognition*. Long Beach: IEEE, 2019: 1328-1338.
- [16] XU N, YANG L J, FAN Y C, *et al.* YouTube-VOS: Sequence-to-sequence video object segmentation [C]//*Proceedings of the 15th European Conference on Computer Vision (ECCV)*. Munich: Springer, 2018: 603-619.
- [17] YAN B, ZHANG X Y, WANG D, *et al.* Alpha-refine: Boosting tracking performance by precise bounding box estimation [C]//*Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville: IEEE, 2021: 5285-5294.
- [18] PAUL M, DANELLJAN M, MAYER C, *et al.* Robust visual tracking by segmentation [C]//*Proceedings of the 17th European Conference on Computer Vision*. Tel Aviv: Springer, 2022: 571-588.
- [19] LUKEŽIČ A, MATAS J, KRISTAN M. D3S-A discriminative single shot segmentation tracker [C]//*Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE, 2020: 7131-7140.
- [20] HU Y T, HUANG J B, SCHWING A G. VideoMatch: Matching based video object segmentation [C]//*Proceedings of the 15th European Conference on Computer Vision (ECCV)*. Munich: Springer, 2018: 56-73.
- [21] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation [C]//*Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*. Munich: Springer, 2015: 234-241.
- [22] KRISTAN M, LEONARDIS A, MATAS J, *et al.* The sixth visual object tracking VOT2018 challenge results [C]//*Proceedings of the European Conference on Computer Vision*. Munich: Springer, 2018: 3-53.
- [23] DU D W, QI Y K, YU H Y, *et al.* The unmanned aerial vehicle benchmark: Object detection and tracking [C]//*Proceedings of the 15th European Conference on Computer Vision*. Munich: Springer, 2018: 375-391.
- [24] WU Y, LIM J, YANG M H. Online object tracking: A benchmark [C]//*Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*. Portland: IEEE, 2013: 2411-2418.
- [25] ZHU Z, WANG Q, LI B, *et al.* Distractor-aware Siamese networks for visual object tracking [C]//*Proceedings of the 15th European Conference on Computer Vision*. Munich: Springer, 2018: 103-119.
- [26] BHAT G, JOHNANDER J, DANELLJAN M, *et al.* Unveiling the power of deep tracking [C]//*Proceedings of the 15th European Conference on Computer Vision*. Munich: Springer, 2018: 493-509.
- [27] DANELLJAN M, BHAT G, KHAN F S, *et al.* ECO: Efficient convolution operators for tracking [C]//*Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE, 2017: 6931-6939.

作者简介:



李新鹏,男,硕士研究生,2021年于西安工业大学获得学士学位,主要从事计算机视觉和目标跟踪方面的研究。
E-mail:602219008@qq.com



王 鹏,男,博士,教授,2022年于西北工业大学获得博士学位,主要从事智能信息感知与模式识别、武器测试技术及嵌入式系统等方面的研究。
E-mail:wp_xatu@163.com