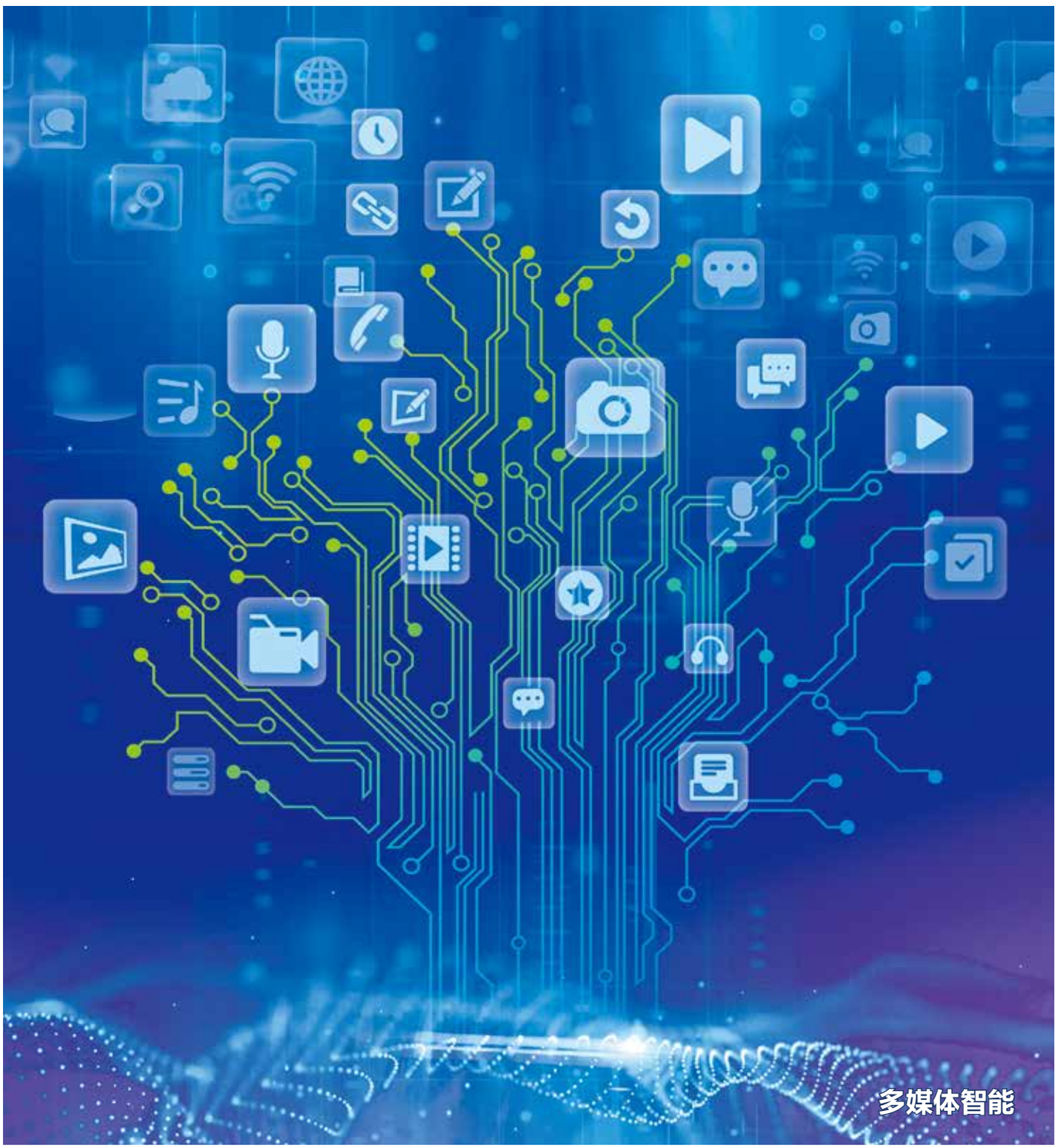
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
09
VOL.27

ISSN1006-8961
CN11-3758/TB



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第27卷第9期（总第317期）

2022年9月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

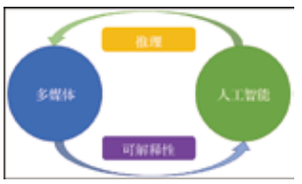
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

2022年9月 第27卷 第9期（总第317期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

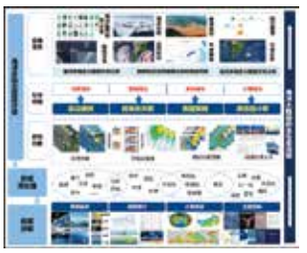
目次



多媒体智能：当多媒体遇到人工智能(第2551页)

学者观点

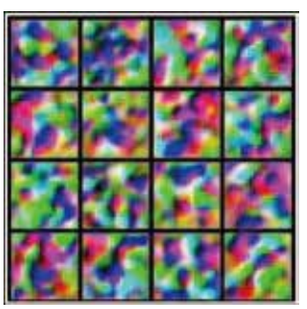
综述



面向海洋的多模态智能计算：挑战、进展和展望(第2589页)

多媒体智能安全

目标智能检测

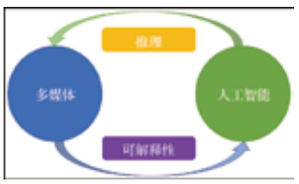


基于真实数据感知的模型功能窃取攻击(第2721页)

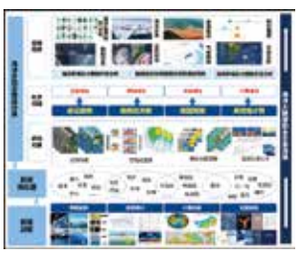
多媒体分析与理解

Volume 27, Number 9
Published September 16, 2022

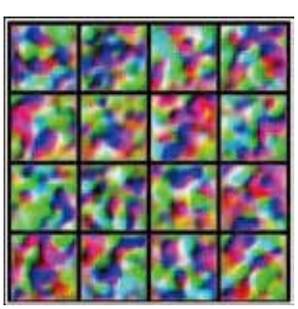
CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Multimedia intelligence: the convergence of multimedia and artificial intelligence(P2551)



Marine oriented multimodal intelligent computing: challenges, progress and prospects(P2589)



Model functionality stealing attacks based on real data awareness(P2721)

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2022)09-2733-16

论文引用格式: Wang M M, Yang X Q and Liu Y. 2022. A spatio-temporal encoded network for single object tracking. Journal of Image and Graphics, 27(09):2733-2748(王蒙蒙,杨小倩,刘勇. 2022. 利用时空特征编码的单目标跟踪网络. 中国图象图形学报,27(09):2733-2748)[DOI: 10.11834/jig.211157]

利用时空特征编码的单目标跟踪网络

王蒙蒙,杨小倩,刘勇*

浙江大学控制科学与工程学院,杭州　310027

摘　要: **目的**　随着深度神经网络的出现,视觉跟踪快速发展,视觉跟踪任务中的视频时空特性,尤其是时序外观一致性(temporal appearance consistency)具有巨大探索空间。本文提出一种新颖简单实用的跟踪算法——时间感知网络(temporal-aware network, TAN),从视频角度出发,对序列的时间特征和空间特征同时编码。**方法**　TAN 内部嵌入了一个新的时间聚合模块(temporal aggregation module, TAM)用来交换和融合多个历史帧的信息,无需任何模型更新策略也能适应目标的外观变化,如形变、旋转等。为了构建简单实用的跟踪算法框架,设计了一种目标估计策略,通过检测目标的 4 个角点,由对角构成两组候选框,结合目标框选择策略确定最终目标位置,能够有效应对遮挡等困难。通过离线训练,在没有任何模型更新的情况下,本文提出的跟踪器 TAN 通过完全前向推理(fully feed-forward)实现跟踪。**结果**　在 OTB(online object tracking: a benchmark)50、OTB100、TrackingNet、LaSOT(a high-quality benchmark for large-scale single object tracking)和 UAV(a benchmark and simulator for UAV tracking)123 公开数据集上的效果达到了小网络模型的领先水平,并且同时保持高速处理速度(70 帧/s)。与多个目前先进的跟踪器对比,TAN 在性能和速度上达到了很好的平衡,即使部分跟踪器使用了复杂的模板更新策略或在线更新机制,TAN 仍表现出优越的性能。消融实验进一步验证了提出的各个模块的有效性。**结论**　本文提出的跟踪器完全离线训练,前向推理不需任何在线模型更新策略,能够适应目标的外观变化,相比其他轻量级的跟踪器,具有更优的性能。

关键词:计算机视觉;目标跟踪;时空特征编码;任意目标跟踪;角点跟踪;时序外观一致性;高速跟踪

A spatio-temporal encoded network for single object tracking

Wang Mengmeng, Yang Xiaoqian, Liu Yong*

College of Control Science and Engineering, Zhejiang University, Hangzhou　310027, China

Abstract: **Objective** Current visual tracking has been developed dramatically based on deep neural networks. The task of single object visual tracking aims at tracking random objects in a sequential video streaming through yielding the bounding box of the object bounding box in the initial frame. It can be an essential task for multiple computer vision applications like surveillance systems, robotics and human-computer interaction. The simplified, small scale, easy-used and feed-forward trackers are preferred due to resources-constrained application scenarios. Most of methods are focused on top performance. Instead, we break this paradox from another perspective through the key temporal cues modeling inside our model and the ignorance of model update process and large-scaled models. The intrinsic qualities are required to be developed in the research community (i. e., video streaming). A video analysis task is beneficial to formulate the basis of the tracking task

收稿日期:2021-12-17;**修回日期**:2022-05-16;**预印本日期**:2022-05-23

*通信作者:刘勇　yongliu@iipc.zju.edu

基金项目:国家自然科学基金项目(61836015)

Supported by: National Natural Science Foundation of China (61836015)

itself. First, the object has a spatial displacement constraint, which means the object locations in adjacent frames will not be widely ranged unless dramatic camera motions happened. Almost visual trackers are followed this path and a new framed search objects is overlapped starting from the location in the last frame. Next, the potential temporal appearance consistency problem, which indicates the target information from preceding frames changes smoothly. This could be regarded as the temporal context, which can provide clear cues for the following predictions. However, the second feature has not been fully explored in the literature. Existing methods is leveraged temporal appearance consistency in two ways as mentioned below: 1) use the target information only in the first frame by modeling visual tracking as a matching problem between the given initial patch and the follow-up frames. Siamese-network-based methods are the most popular and effective methods in this category. They applied a one-shot learning scheme for visual tracking, where the object patch in the first frame is treated as an exemplar, and the patches in the search regions within the consecutive frames are regarded as the candidate instances. The task is transferred to find the most similar instance from each frame. This paradigm ignores other historical frames completely, deals with each frame independently and causes tremendous information loss. 2) Use the given initial patch and the historical target patches both targets at every frame or selected frames to predict the object location in a new frame, including traditional and deep-neural-network-based ones. Traditional trackers based methods like the correlation filter (CF) can learn their models or classifiers from the first frame and update models in the subsequent frames with a small learning rate. Our diverse deep-neural-network-based methods first learn their models offline with vast training data and fine-tune the models online at the initial frame and other frames. However, the solution remains open to balancing the accuracy and latency, especially in deep-neural-network-based methods. Also, network finetuning is forbidden in some practical applications when it is deployed in inference chips, which hinders the wide deployment of these methods. **Method** We facilitate a novel and straightforward tracker to re-formulate the visual tracking problem from the perspective of video analysis. A new temporal-aware network (TAN) is designed to encode target information from multiple frames, which aims at taking advantage of the temporal appearance consistency and the spatial displacement constraint in the forward path without an online model update. To exchange and fuse information from historical frames input, we introduce temporal aggregation modules in TAN and empower our tracker TAN with the ability to learn spatio-temporal features. To balance the computational burden resulting from the multi-frame inputs and tracking accuracy, we employ a shallow network ResNet-18 as our feature extraction backbone and achieve a high speed of over 70 frame/s. Our tracker runs completely feed-forward and can adapt to the target's appearance changes with our offline-trained, temporal-encoded TAN because previous temporal appearance consistency is maintained by the first frame or historical frames, which require expensive online finetuning to be adaptable. To construct a completed simple tracking pipeline further, we design a novel anchor-free and proposal-free target estimation method, which can detect the four corners, including top-left, top-right, bottom-left, and bottom-right, with a corner detection head in TAN. As target locations can be determined by a pair of top-left and bottom-right corners or top-right and bottom-left, we make use of a center score map to indicate the confidence of these two bounding boxes rather than complicated embedding constraints, which can easily locate the target. Thanks to a corner-based target estimation mechanism, our tracker is capable of handling challenging scenarios for significant changes involved. **Result** Without bells and whistles, our method has its potentials on several public datasets, such as online object tracking: a benchmark 50 (OTB50), OTB100, TrackingNet, a high-quality benchmark for large-scale single object tracking (LaSOT), and a benchmark and simulator for UAV tracking (UAV) 123. Our real-time speed optimization and simplified pipeline make TAN more suitable for real applications, especially resource-limited platforms where the large models and online model updates are not supported. **Conclusion** The proposed tracker will provide a new perspective for single-object tracking by mining the video nature, especially the temporal appearance consistency of this task.

Key words: computer vision; object tracking; spatial-temporal feature coding; arbitrary target tracking; corner tracking; temporal appearance consistency; high-speed tracking

0 引 言

单目标跟踪旨在仅给定任意对象在初始帧目标位置的情况下,跟踪后续视频中的对象。目标跟踪是监控系统、机器人和人机交互等大量多媒体认知理解的基础任务。由于大多数实际场景应用中的硬件计算资源有限,且基于在线模型更新的跟踪器网络结构较为复杂和运行速度较慢,难以部署落地,因此探究简单易用的跟踪算法框架十分重要。高性能的跟踪器往往通过使用注意力机制、模板更新等常用策略提升性能。与之不同的是,本文从视频时空特征角度出发,基于序列的时序外观一致性和空间位移约束进行建模,避免使用复杂网络结构和模型更新等策略,较好地平衡了跟踪器的性能和速度。

视觉跟踪任务中的视频时空属性并未充分探究和开发。作为一项视频分析任务,对跟踪来说至少有两个潜在特性和优势。1)目标具有空间位移约束,这意味着除非发生剧烈的物体或相机运动,相邻帧之间的目标位置不会相距太远。现有的视觉跟踪器(任仙怡 等,2002;Henriques 等,2015;Li 等,2018;Zuo 等,2019;宫海洋 等,2018;Danelljan 等,2019;宁纪锋 等,2014;王鑫和唐振民,2010)几乎都遵循空间位移约束,并根据目标在上一帧中的位置在当前帧中开展搜索。2)目标具有时序外观一致性,这表明相邻帧之间的目标外观变化比较微弱,整个序列从时间维度上来看,目标外观是缓慢平滑变化的。时序外观一致性可以提供上下文信息,为后续帧的预测提供有效线索,然而该特性在现有工作中并没有得到充分挖掘和研究。

现有的跟踪器对时序外观一致性的利用主要有两种方式。第 1 种方式是将视觉跟踪建模为初始目标与后续帧的匹配问题,基于孪生网络(siamese network)的跟踪器(Li 等,2018,2019;Bertinetto 等,2016;Xu 等,2020;Zhu 等,2018;Fan 和 Ling,2019;Zhou 等,2020;Chen 等,2020;Guo 等,2020)是其中最经典和有效的方法。它们采用 one-shot 方式进行视觉跟踪,将视频第 1 帧中的目标作为模板,将后续帧中的搜索区域作为候选目标,然后将跟踪任务变为从每一帧中找到与模板最相似的候选目标。这种方式完全忽略了其他历史帧的有效信息而每次独立处理每一帧,造成巨大的信息损失。第 2 种方式,许多传统方法(Henriques 等,2015;丁欢和张文生,2012;陈晨 等,2020;Hare 等,2016)和基于深度学习的方法(Danelljan 等,2019;Bhat 等,2019;宋建锋等,2021;Dai 等,2020;周笑宇 等,2021)同时使用初始目标和历史目标信息(历史目标为每一帧或选定的帧)预测当前帧的目标位置。然而,这种方式对于如何平衡准确率和时间延迟,尤其是基于深度神经网络的方法,仍然是一个较为困难的问题。此外,网络微调在一些实际场景中是无法实现的,如将算法部署到硬件芯片上,严重阻碍了跟踪算法的实际落地应用。

针对上述困难,提出一种新的跟踪器,从视频分析角度重新对跟踪问题建模,本文设计了一种时间感知网络(temporal-aware network,TAN),旨在没有在线模型更新的前提下,利用前向推理中的时序外观一致性,对多个视频帧的信息同时进行编码。在网络中引入了时间聚合模块(temporal aggregation module,TAM),以交换和融合历史帧中的信息,使跟踪器 TAN 能够学习目标的时空特征。为了平衡多帧输入的处理速度和跟踪精度,采用轻量级网络 ResNet-18 作为特征提取骨干网络(backbone network),跟踪器速度达到 70 帧/s。与使用在线模型微调的方法不同,跟踪器 TAN 能够在跟踪过程中进行完全前向推理,在不需要任何网络权重学习或微调的情况下,通过离线训练以学习适应目标的外观变化。此外,为了构建简单通用的跟踪框架,本文设计了一种无锚(anchor-free)、无候选(proposal-free)的目标估计方法,即检测目标的 4 个角点,包括左上角、右上角、左下角和右下角。使用角点检测头(corner detection head)得到 4 个角点后,根据左上—右下角点,右上—左下角点分别得到两组目标候选框,然后通过候选框的中心点置信度分数确定最终目标位置,而不必考虑复杂的约束关系,这能进一步减少目标位置的推理时间。使用这种基于角点的目标估计机制,跟踪器 TAN 能够应对多种具有挑战性的场景,如局部遮挡、外形变化等。

本文提出的跟踪方法简单直观易于实现,达到了领先算法的跟踪性能和很快的推理速度。这种实时有效的跟踪算法框架跟踪器在性能和速度上达到了很好的平衡,更适用于实际场景,尤其对于不支持复杂模型和在线模型更新且计算资源有限的硬件平台。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)提出一种速度与精度平衡的目标跟踪算法,即时间感知网络(TAN),利用视频中的时序外观一致性优势和空间位移约束,对目标时空特征进行建模。在跟踪过程中无需在线模型更新,即可应对目标在时序上的各种变化;2)设计了一种简单有效的目标估计方法,即检测目标边界框的4个角点,该方法能够使跟踪器有效应对多种具有挑战性的场景,如局部遮挡、外形变化等;3)以ResNet-18为主干网络,提出的目标跟踪框架在多个公开数据集上达到领先水平,实现了70帧/s的高速推理。

1 相关工作

关注视频时序特性的研究,尤其是对时序外观一致性,将现有方法分为基于初始模板的跟踪器和基于历史模板的跟踪器两类。此外,针对引入的基于角点检测头的目标估计方法介绍一些相关的跟踪器。

1.1 基于初始模板的跟踪器

基于初始模板跟踪指跟踪器离线训练学习通用特征表示,然后仅使用初始模板对新目标进行跟踪。基于孪生网络的跟踪器(Wang 等,2019;Yu 等,2020;Yang 等,2020)是最经典的 one-shot 学习方法。SINT(siamese instance search for tracking)(Tao 等,2016)和 SiamFC(fully-convolutional siamese networks for object tracking)(Bertinetto 等,2016)作为孪生网络的开创性工作,将视觉跟踪任务表示为初始帧目标与后续帧的成对匹配问题。受目标检测任务中区域候选网络(region proposal network,RPN)和孪生网络跟踪框架的启发,SiamRPN(high performance visual tracking with siamese region proposal network)(Li 等,2018)将跟踪推理视为一个 one-shot 局部检测任务,使用两个权值共享的骨干网络进行特征提取,并使用 RPN 头实现目标分类和边界框回归。SiamRPN ++(Li 等,2019)打破了深度卷积特征中平移不变性的限制,使用 ResNet-50 或更深层的网络作为特征提取主干,跟踪精度在多个公开数据集上提升到更高水平。这些方法通过离线训练且没有模型更新,完全通过初始模板和后续帧的相似性度量进行目标定位跟踪,忽略其他历史帧信息,造成巨大的信息浪费。

1.2 基于历史模板的跟踪器

在考虑使用目标历史模板的信息时,现有方法大多研究模型更新策略(Danelljan 等,2019,2020;Bhat 等,2019)。MDNet(multi-domain convolutional neural network)(Nam 等,2016)由共享层和特定域的多分支层组成,跟踪过程中通过在线微调以适应新目标。UCT(learning unified convolutional networks for real-time visual tracking)(Zhu 等,2017)通过引入峰值降噪比(peak-versus noise ratio,PNR)的方法,避免由于跟踪不准确而引入错误背景信息。ATOM(accurate tracking by overlap maximization)(Danelljan 等,2019)提出一种新的跟踪框架,由目标框回归分支和分类分支组成,其中分类分支可以在线训练以提高跟踪器对背景和目标的区分能力。DiMP(learning discriminative model prediction for tracking)(Bhat 等,2019)和 PrDiMP(probabilistic regression for visual tracking)(Danelljan 等,2020)都是 ATOM 的后续工作,分别改进了分类分支和回归分支。

另一种代表性跟踪器是基于元学习的方法。这类跟踪器首先基于各种检测任务进行训练,然后通过包含目标历史信息的少量训练样本(包括第1帧和后续多帧)微调网络以快速学习适应新目标。Meta-tracker(Park 和 Berg,2018)是第1个在跟踪任务中使用元学习训练模型的方法。Huang 等人(2019a)直接将目标检测器转换到跟踪器中,使用元学习方式学习检测头中的元学习层(meta-layer)。

上述提到的跟踪器试图利用多个历史帧挖掘更多的视频时序特性。然而,本文提出的跟踪器能够实现完全前向推理(fully feed-forward process),无需任何模型更新策略。MemTrack(learning dynamic memory networks for object tracking)(Yang 和 Chan,2018)和 MemDTC(visual tracking via dynamic memory networks)(Yang 和 Chan,2021)是与本文最相似的方法,利用目标历史信息且不在线更新模型权重,但是它们需要使用额外的动态长短期记忆网络(long short-term memory,LSTM)模型适应跟踪过程中目标的外观变化。相比之下,本文方法无需使用任何额外网络来编码或存储历史目标信息,提出的跟踪网络在性能和速度上达到了平衡。

1.3　基于角点的跟踪

一般的多尺度搜索策略无法在涉及目标变化等困难场景下估计密集的边界框，因此如何估计准确的边界框引起了研究人员的兴趣。虽然基于角点的跟踪器可以灵活地应对这些变化，但是这类跟踪器尚未成熟。GOTURN（generic object tracking using regression networks）（Held 等，2016）和 SATIN（siamese attentional keypoint network for high performance visual tracking）（Gao 等，2020）使用全连接网络和互相关操作检测物体角点，但是这些方法并没有表现出很强的性能。CGACD（correlation-guided attention for corner detection based visual tracking）（Du 等，2020）使用基于相关性引导的注意力角点检测来突出角点区域并增强感兴趣区域（region of interest，RoI）特征以实现角点检测，从而实现边界框估计。Ocean（Zhang 等，2020）提出一种基于对象感知的无锚（object-aware anchor-free）网络，以无锚的方式直接预测目标的位置和尺度。SiamKPN（siamese keypoint prediction network for visual object tracking）（Li 等，2020）提出一种从粗到细（coarse-to-fine）的角点检测的级联热图策略。本文设计的角点检测头与以上方法不同，主要体现在两方面。1）直接预测一个目标的 4 个角点，并按对角构成两个边界框，能够应对遮挡、形变等一些具有挑战性的场景；2）采用简单有效的目标框选择机制，通过预测边界框的中心置信度分数而不是复杂的约束来确定目标的最终位置。

2　方　法

2.1　时间感知网络

给定一对输入，包含 K 个历史目标实例 $\{\boldsymbol{I}_j\}_{j=1}^{K}$ 以及当前搜索区域 $\boldsymbol{I}_{\mathrm{S}}$，TAN 的目标是根据这 K 个模板 $\{\boldsymbol{I}_j\}_{j=1}^{K}$，学习预测出在 $\boldsymbol{I}_{\mathrm{S}}$ 中的目标位置。将该问题建模为回归 4 个角点，即左上角、右上角、左下角和右下角，以及目标边界框的中心置信度。

TAN 总体框架如图 1 所示。给定输入模板图像和搜索区域图像，首先使用两个权值共享的骨干网络提取多尺度特征。对于每个尺度，通过相关性模块和两个独立的时间聚合模块进行特征合并，然后分别接角点检测头和置信度分数预测头。在两个头模块中，使用时间融合模块减少时间维度，并使用 Hinge 损失进行监督训练。整体的跟踪框架中有两个权值共享的特征提取主干，提取 K 个模板的特征 $\boldsymbol{\phi}_{\mathrm{T}}^{i} \in \mathbf{R}^{B\times K\times C^i\times H_{\mathrm{T}}^i\times W_{\mathrm{T}}^i}$，以及搜索区域的特征 $\boldsymbol{\phi}_{\mathrm{S}}^{i} \in \mathbf{R}^{B\times 1\times C^i\times H_{\mathrm{S}}^i\times W_{\mathrm{S}}^i}$，其中下标 T 和 S 分别代表模板和搜索区域，上标 $i \in \{0,\cdots,N\}$ 代表主干不同的阶段，B 代表批量大小（batch-size），C 代表通道（channel），H 代表高度（height），W 代表宽度（width）。为了平衡多帧处理时间负担和跟踪精度，使用轻量级网络 ResNet-18 作为特征提取主干，$\boldsymbol{\phi}_{\mathrm{S}}^{i}$ 和 $\boldsymbol{\phi}_{\mathrm{T}}^{i}$ 首先通过权值共享的相关性模块（correlation module），为表述简单，后文省略上标 i。

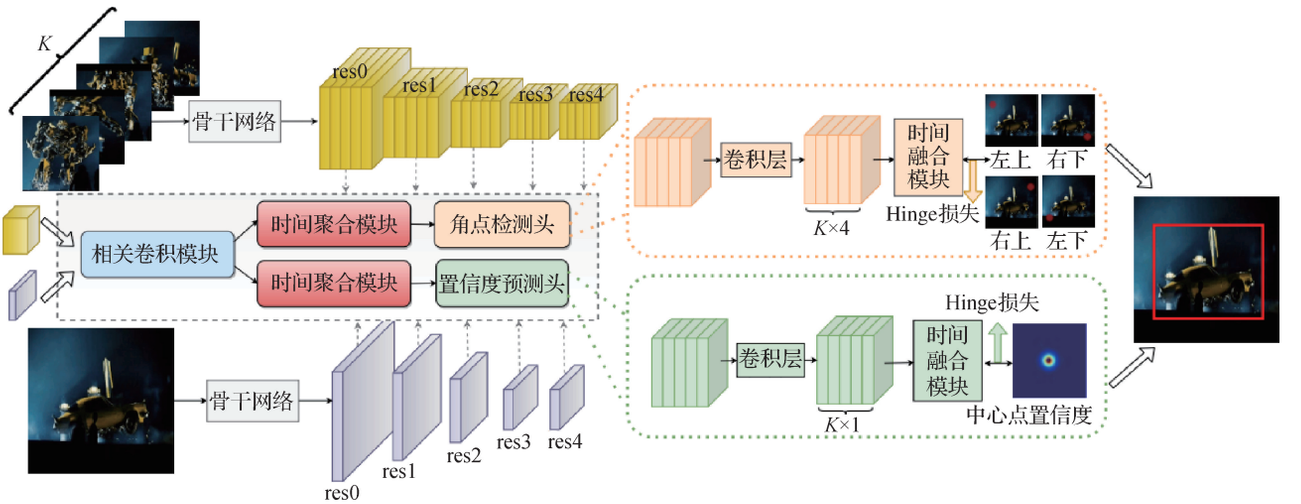


图 1　时间感知网络的总体框架

Fig. 1　Overview of the proposed temporal-aware network

相关性模块结构如图 2 所示，首先是一个 3 × 3 的卷积层，后接批量归一化层（batch normalization，BN）以及 ReLU 激活函数层调整两个输入特征，同时将空间通道减少至 1/2 以减低计算成本。使用一个深度互相关卷积层（depth-wise cross-correlation layer）（Li 等，2019），将 $\boldsymbol{\phi}_{\mathrm{T}}$ 当做卷积核以结合两输入的特征。

$$\boldsymbol{F} = ReLU(BN(\boldsymbol{W}_1 * \boldsymbol{\phi}_{\mathrm{S}})) \bigstar ReLU(BN(\boldsymbol{W}_2 * \boldsymbol{\phi}_{\mathrm{T}})) \tag{1}$$

式中，★ 代表深度互相关卷积层，* 代表普通卷积，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 分别代表 $\boldsymbol{\phi}_{\mathrm{S}}$ 和 $\boldsymbol{\phi}_{\mathrm{T}}$ 的 3 × 3 卷积核权重。

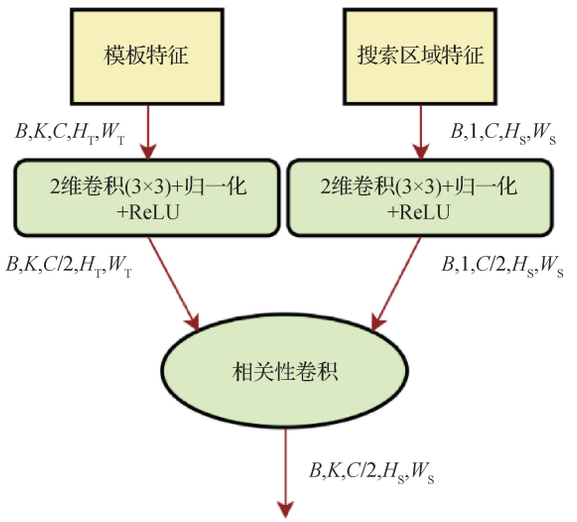


图 2　相关性模块

Fig. 2　Correlation module

由于上述特征在时间维度上是离散的，因此提出一种新的时间建模策略，即时间聚合模块，使网络 TAN 具备学习时空特征的能力。最后，利用角点检测头和置信度检测头得到目标的 4 个角点和中心位置置信度 I_{S} 。

一般来说，多尺度特征融合是提高目标检测和视觉跟踪等任务性能的常用方法。网络深层低分辨率输出产生的定位结果一般粗糙且鲁棒，而浅层高分辨率输出产生的结果趋向于与深层互补。原因是不同层对特征的表示不同，浅层主要表示颜色、纹理等低级特征但缺乏语义信息，而高层获得的高级特征能编码丰富的语义信息。因此，从多个尺度构建层级特征可以鲁棒精确地定位目标。与常用的 ResNet 网络中 4 个阶段特征（res1—res4）的方法对比，本文发现第 1 个池化层（res 0）的特征对跟踪也非常有效，消融实验验证了这一点。

2.2　时间聚合模块

为了将多个目标模板的特征图进行交换和融合以实现时空特征编码，提出时间聚合模块（TAM）如图 3 所示。使用时间卷积融合多个模板特征，首先变换特征 $\boldsymbol{F}$ 维度为 $\hat{\boldsymbol{F}} \in \mathbf{R}^{BH_{\mathrm{S}}W_{\mathrm{S}} \times C/2 \times K}$，然后在最后一个维度上使用 1D 卷积，即在时间维度上对 K 个相关特征进行聚合。考虑到不同通道的语义信息通常是不同的，因此采用深度卷积（depth-wise convolution）方式以确保信息按通道进行融合。数学表达为

$$\hat{\boldsymbol{H}} = \boldsymbol{W}_3 * \hat{\boldsymbol{F}} \tag{2}$$

式中，$\boldsymbol{W}_3$ 是 1D 时间卷积核的权重，$\hat{\boldsymbol{H}}$ 表示相关特征 $\boldsymbol{F}$ 与其他特征融合后的增强特征，将 $\boldsymbol{W}_3$ 的大小设置为 1 × 3。之后，$\hat{\boldsymbol{H}}$ 的维度变换回 $\boldsymbol{H} \in \mathbf{R}^{B \times K \times C/2 \times H_{\mathrm{S}} \times W_{\mathrm{S}}}$，并使用一个普通的 3 × 3 卷积层、BN 层（BN）和 ReLU 层（$ReLU$），以获得在时间融合特征 $\boldsymbol{H}$ 上的时空特征。此外，采用残差连接以保留原始帧的时间信息。上述模块具体表示为

$$\boldsymbol{A} = \boldsymbol{F} \oplus ReLU(BN(\boldsymbol{W}_4 * \boldsymbol{H})) \tag{3}$$

式中，⊕ 代表逐元素相加，$\boldsymbol{W}_4$ 是 3 × 3 空间卷积的权重，$\boldsymbol{A}$ 是时间聚合模块的输出特征。通过提出的时间聚合模块，多个模板的信息可以准确地进行融合增强以更好地定位目标。

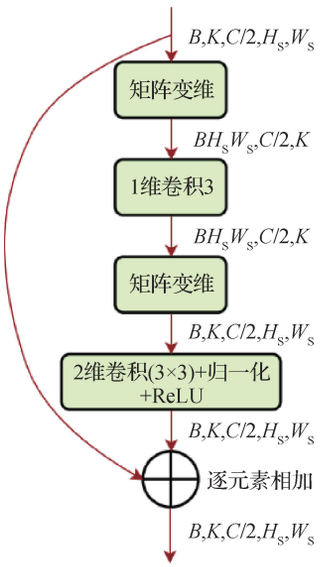


图 3　时间聚合模块

Fig. 3　Temporal aggregation module

2.3　目标定位头

为了构建一个简单实用的跟踪框架，提出一种

新颖的无锚(anchor-free)、无候选(proposal-free)的目标估计方法。将目标定位分解为角点检测和中心置信度预测两个子任务,分别由角点检测头和分数预测头实现。如图 4 所示,目标候选位置由左上—右下角点或右上—左下角点构成的两组候选框组成,然后根据它们在置信度图中的中心置信度分数决定最终位置。

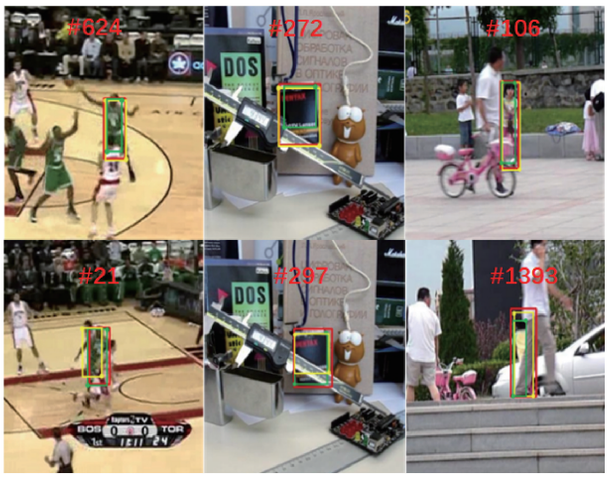


图 4　角点检测结果

Fig. 4　Results of our target localization heads

如图 1 所示,角点检测头学习目标的左上角、右上角、左下角和右下角。这个头部 CNN 结构由一个 3 ×3 空间卷积层、BN 层和 ReLU 激活函数以及这些层的重复结构组成。因为 K 个输入模板对应 K 个输出(每个输出的 4 个通道分别代表 4 个角点),需要进行融合将时间维度从 K 降到 1 以获得 4 个角点的位置置信图,这里使用了池化操作($Pool$),表述为

$$\boldsymbol{P}_c = Pool(ReLU(BN(\boldsymbol{W}_6 * ReLU(BN(\boldsymbol{W}_5 * \boldsymbol{A}))))) \tag{4}$$

式中, $\boldsymbol{W}_5$ 和 $\boldsymbol{W}_6$ 是 3 × 3 空间卷积层的权重, $\boldsymbol{P}_c$ 是 4 个通道的角点预测输出。置信度预测头 $\boldsymbol{P}_s$ 的网络结构除了输出是 1 个通道外,其他部分和角点检测头相同。这两个预测头的权重不共享。

对于每个角点,实际上只有 1 个像素位置是正样本,其他位置都是负样本。然而,一对错误的角点检测,假如它们组成的框与所代表的真实框重叠交并比(intersection-over-union,IoU)足够大,仍然会当做正确的边界框。因此本文通过减少目标中心位置半径 r 内的负样本位置的惩罚以减轻这种错误影响。具体地,使用一个非标准化的 2D 高斯 $\boldsymbol{Y} = e^{-\frac{(x-\hat{x})^2+(y-\hat{y})^2}{2\sigma^2}}$ 作为训练真值,其中心点是正样本点 $(\hat{x},\hat{y})$ 且 $\sigma = r/3$ 。半径 r 由跟踪实例的尺寸大小决定。具体方法是确保半径内一对点生成的边界框与真实边界框的交并比 IoU≥ η ,因此 r 不是一个固定数值,而是根据训练目标样本尺度大小自适应变化。以同样方式生成中心置信度预测真值,中心位置的置信值分数最高,周围位置平滑地逐渐减小到 0,以适应置信度变化。

本文 4 个角点的检测策略不是如 CornerNet(Law 和 Deng, 2019)、CGACD(Du 等, 2020)和 SiamKPN(Li 等,2020)中只检测两个角点,而是发现在某些情况下两个角点会失败,如图 4 中的遮挡场景,4 个角点可以获取更多信息以克服这些困难。在图 4 中,红色框表示目标的真实位置,黄色和绿色框及其对应的角点分别显示了预测的两组候选框“左上—右下”和“右上—左下”。图 4 中,同一列来自同一个视频随机抽取的两帧,从中发现预测多组角点可以克服各种遮挡问题。

2.4　训练和推理

在整个框架的训练和推理方法中,网络框架 TAN 完全离线训练,并且推理过程无需任何模型微调。如图 5 所示,本文算法的损失计算仅存在于训练阶段。由于原始的 res3 和 res4 层的输出分辨率大小相同,所以没有上采样。由于两个头的处理方式相同,因此省去每个原始输出的下标。

2.4.1　离线训练

在离线训练过程中,输入由模板 $\{\boldsymbol{I}_j\}_{j=1}^{K}$ 和搜索区域 $\boldsymbol{I}_S$ 组成。具体地,从训练视频序列中顺序采样 $K+1$ 幅图像,其中前 K 帧是模板 $\{\boldsymbol{I}_j\}_{j=1}^{K}$,最后一帧是搜索区域 $\boldsymbol{I}_S$ 。给定输入,将它们输入特征提取主干网络以获得特征对 $(\boldsymbol{\phi}_T^i,\boldsymbol{\phi}_S^i)$,之后,在每个特征层级中使用相关性模型、时间聚合模块以及两个检测头,得到多尺度预测输出。

通常在输出热度图上,绝大多数的样本值是零,只有少量一部分是非零值。为了解决这个数据不平衡问题,调整预测的热度图($\hat{\boldsymbol{P}}_*^i$)以避免在负样本上过拟合,具体是将背景区域中的置信分数置零,表述为

$$\hat{\boldsymbol{P}}_*^i = 1_{\boldsymbol{P}_*^i<\theta} \odot \max(0,\boldsymbol{P}_*^i) + (1 - 1_{\boldsymbol{P}_*^i<\theta}) \odot \boldsymbol{P}_*^i \tag{5}$$

式中, 1 是其下标的指示函数, ⊙ 代表矩阵点乘, θ 是过滤背景区域的阈值,下标 * 代表角点(cor-

ner)或中心点(center),上标 $i \in \{0,\cdots,N\}$ 表示不同的特征层级。使用 L2 损失函数对真实热度图和预测热度图进行监督,并且以等权重对不同特征尺度的角点和中心点进行加权。训练和推理过程如图 5 所示,对每个尺度的输出都进行监督,最终的训练损失可以表示为

$$L(\{\boldsymbol{I}_j\}_{j=1}^{K},\boldsymbol{I}_{\mathrm{S}}) = \sum_{i=0}^{N}(\|\hat{\boldsymbol{P}}^i_{\mathrm{corner}} - \boldsymbol{Y}^i_{\mathrm{corner}}\|_2^2 + \|\hat{\boldsymbol{P}}^i_{\mathrm{center}} - \boldsymbol{Y}^i_{\mathrm{center}}\|_2^2) \tag{6}$$

式中,$\boldsymbol{Y}^i_*$ 是对应的真实热度图。

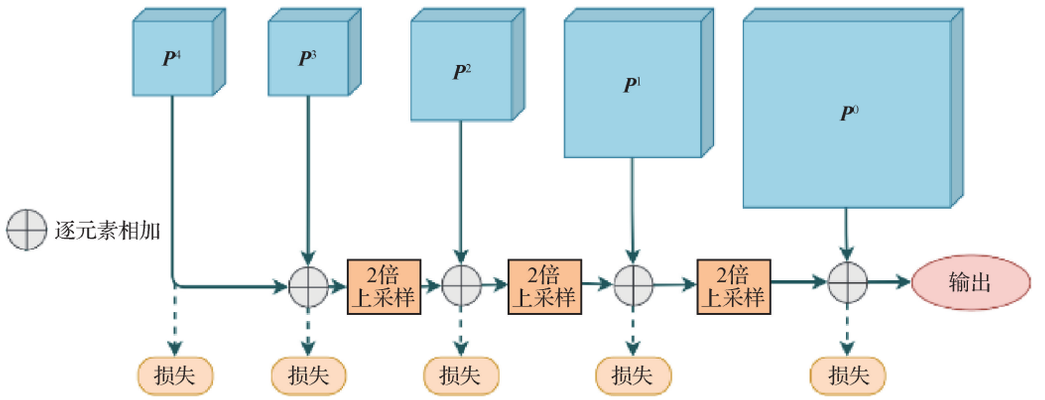


图 5 训练和推理过程

Fig. 5 Training and inference process

2.4.2 推理

在推理阶段需要设计网络的输出解码策略,获得最终的目标位置和构建网络输入。如图 5 所示,当网络输出包含所有尺度的角点和中心置信度的预测热度图时,最简单的融合方式是将低分辨率的输出上采样到更高的分辨率相加并求其均值,并将最高分辨率输出视为最终的热度图。然后,使用 argmax 操作在每个热度图中找到最大响应位置,按角点对(左上—右下,右上—左下)构成两组边界框,最后在中心置信预测热度图中分别找到两个边界框对应中心点的置信度分数 (s_1,s_2),选择较大者作为最终的目标位置。

对于输入的构建,根据前 1 帧的目标位置在当前帧裁剪出搜索区域 $\boldsymbol{I}_{\mathrm{S}}$,并从由高质量历史帧裁剪组成的存储库(memory)中选择模板 $\{\boldsymbol{I}_j\}_{j=1}^{K}$。由于在线跟踪过程没有跟踪质量的反馈信息,本文利用中心置信分数 (s_1,s_2) 作为衡量指标。具体来说,对于初始帧,首先通过添加轻微扰动生成多个初始模板,主要包括位移和尺度变化,并将这些模板放入存储库中。在跟踪过程中,如果跟踪目标区域的中心置信分数 (s_1,s_2) 均高于第 2 帧的中心置信分数阈值 ω,就将其保存进存储库中。选择第 2 帧的原因是其与初始目标最为相似。存储库是一个固定大小的队列,设定其大小为 M,当存储库已满时,将最早的非固定元素移出。从 $\{\boldsymbol{I}_j\}_{j=1}^{K}$ 中选择 K 个模板的方式为

$$[1,rand([2,M-6],K-2),rand([M-5,M],1)] \tag{7}$$

式中,$rand([a,b],n)$ 表示从 $[a,b]$ 中随机选择不重复的 n 个样本。注意,当 $K=1$,则使用初始模板;当 $K=2$ 则使用初始模板和最新的模板;如果 $K>2$,输入的 K 个模板将包含初始模板、最近的模板和中间任意 $K-2$ 个模板。当给定成对输入 $\{\boldsymbol{I}_j\}_{j=1}^{K}$ 和 $\boldsymbol{I}_{\mathrm{S}}$,离线训练模型 TAN 能够完全前向地实现对目标的时空特征建模及目标位置定位。

3 实 验

3.1 实验细节

本文跟踪器 TAN 基于 PyTorch 实现,训练集包括 GOT-10k(a large high-diversity benchmark for generic object tracking in the wild)(Huang 等,2019b)、LaSOT(a high-quality benchmark for large-scale single object tracking)(Fan 和 Ling,2019)、TrackingNet(Müller 等,2018)和 COCO(common objects in context)(Lin 等,2014),每个训练轮次(epoch)从 4 个数据集中采样 200 000 个样本。在目标标注框周围

随机平移和尺度缩放采样得到由模板帧和搜索帧组成的输入。共享的特征提取器主干是在 ImageNet 上预训练的 ResNet-18。整个模型使用 SGD(stochastic gradient descent)优化器训练 50 个轮次,在第 30 个轮次时学习率衰减为原来的 0.1。在训练和测试中,保持输入模板尺寸为 128 像素,搜索区域大小为 256 像素,并且 $\{\boldsymbol{I}_j\}_{j=1}^{K}$ 和 $\boldsymbol{I}_S$ 的裁剪比分别是原始边界框的 1.25 倍和 2.5 倍。训练参数 η 设置为 0.7,推理参数 θ 设置为 0.001,ω 设置为 0.4,M 设置为 50。在单块 1080Ti GPU 上,算法平均跟踪速度达到 70 帧/s。

3.2　对比实验和分析

实验在 5 个大型跟踪数据集基准中进行,包括 OTB50(online object tracking: a benchmark)(Wu 等,2013)、OTB100(Wu 等,2013)、TrackingNet(Müller 等,2018)、LaSOT(Fan 等,2019)和 UAV(a benchmark and simulator for UAV tracking)123(Müller 等,2016),将 TAN 与先进的跟踪器比较和评估。虽然 TAN 实现简单,但仍获得与排名靠前跟踪器相当的结果,并达到 70 帧/s 的推理速度,比大多数跟踪器快。总体来看,跟踪器 TAN 简单实用,达到了精度和速度的平衡。表明了本文方法对视频时序外观一致性利用的有效性。

3.2.1　在 OTB50 数据集上对比实验

OTB50 数据集包含 50 个具有挑战性的视频,视频序列标记为 11 个属性代表不同的困难场景,包括光照变化、尺度变化、遮挡和形变等。OTB 采用 OPE(one pass evaluation)评估方式,对所有序列只测试一次,评测指标是精度值和成功率图下的曲线面积(area under curve,AUC)。精度值代表预测框中心位置与真实框中心位置的距离小于给定阈值 20 个像素的视频帧数所占百分比。成功率图是由预测框和真实框的交并比 IoU 大于给定阈值(0～1)的视频帧数所占百分比绘制的曲线图。成功率通常用曲线下面积(AUC)表示。实验将提出的跟踪器 TAN 与先进的跟踪方法进行对比,结果如图 6 所示。可以看出,TAN 的成功率排列第 1,精度值位居前列。与同样使用 ResNet-18 为骨干网络的 ATOM、DiMP18 和 PrDiMP18 方法相比,这些方法虽然有复杂的在线模型更新策略,但性能仍低于 TAN。而且 TAN 框架由于没有在线模型更新而显示出更快的速度。在成功率和精度值两个指标上,与 ATOM 相比,TAN 表现优秀,分别高出 3.1% 和 3.8%;与 DiMP18 相比,TAN 分别高出 3.9% 和 5.6%。MemDTC 和 MemTrack 通过设计外部动态存储来维护模板以适应目标变化,也能够实现完全前向推理,但其性能与本文方法相差甚远。

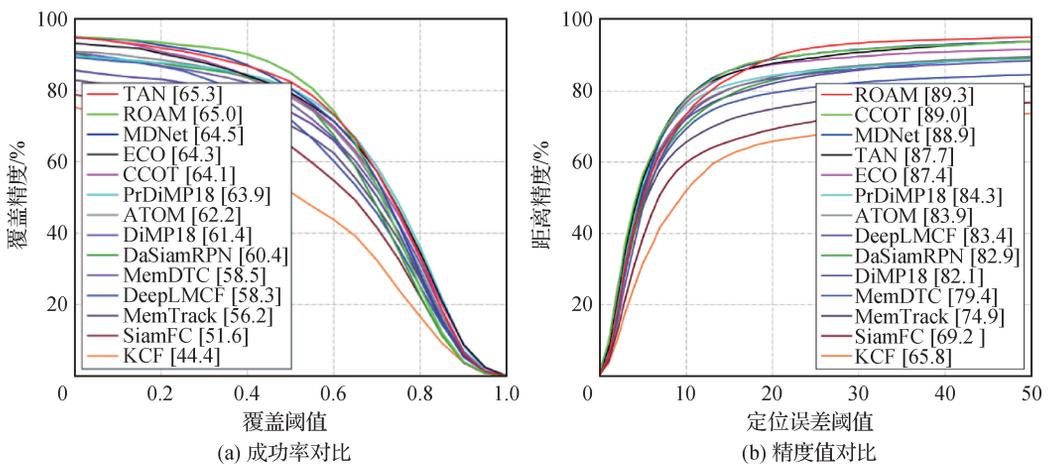


图 6　先进跟踪器在数据集 OTB50 上的对比结果

Fig. 6　Comparison of the quality of state-of-the-art tracking methods on OTB50 dataset

((a) success plots of OPE; (b) precision plots of OPE)

3.2.2　在 OTB100 数据集上对比实验

OTB100 数据集比 OTB50 多 50 个视频序列,评价指标与 OTB50 相同。图 7 展示了提出的框架 TAN 与先进跟踪器的比较结果。可以看出,TAN 成

功率在所有结果中位列第2,稍低于第1的ECO(efficient convolution operators for tracking)算法。但ECO算法使用了模型更新策略提升性能,且运行速度仅为8帧/s,远低于实时性要求。CCOT(continuous convolution operators for visual tracking)的精度值最高,但运行速度低至0.3帧/s,限制其根本不可能应用于实际场景。相比之下,TAN在成功率和精度值两个指标上分别达到68.2%和89.3%。TAN的成功率仅比ECO低0.9%,但速度为70帧/s,远超ECO的8帧/s;TAN的精度值比CCOT低2.2%,但速度是其上百倍。与使用相同骨干网络且需要模型更新的ATOM、DiMP18和PrDiMP18方法相比,TAN取得了更好的成功率(68.2%与66.3%、66.0%、67.9%)和精度值(89.3%与87.4%、85.9%、87.1%),且速度更快(70帧/s与30帧/s、57帧/s、40帧/s)。

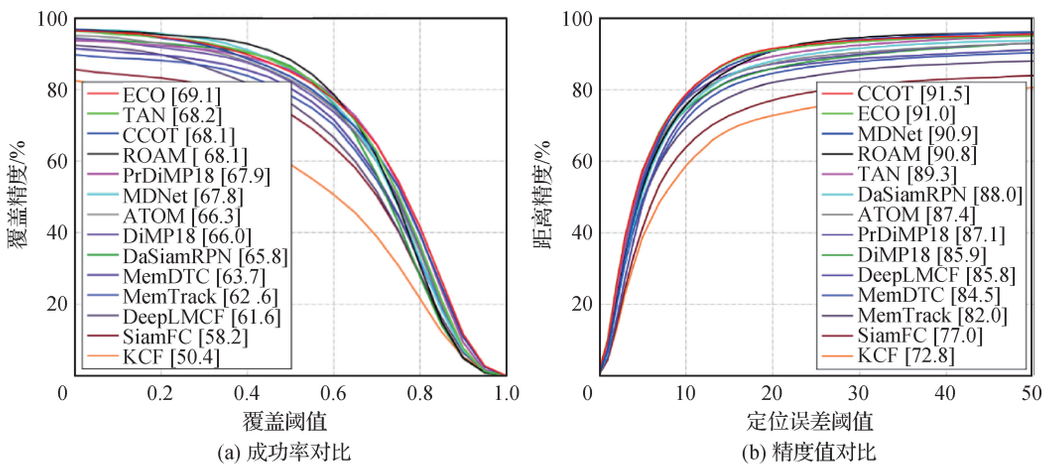


图7 先进跟踪器在数据集OTB100上的对比结果

Fig. 7 Comparison of the quality of state-of-the-art tracking methods on OTB100 dataset

((a) success plots of OPE; (b) precision plots of OPE)

3.2.3 在LaSOT数据集上对比实验

LaSOT数据集提供了大规模、高质量的密集标注,共有1 400段训练视频和280段测试视频,包含70类物体,每类有20段视频序列,视频总体平均长度超过2 500帧,非常适用于评估长时序列跟踪器。LaSOT采用与OTB相同的OPE方式测试成功率和精度。表1展示了TAN与其他先进跟踪器在成功率、精度、归一化精度和速度等方面的比较。跟踪器ATOM和DiMP18性能优于本文的跟踪器TAN,但在速度上逊于TAN。这两个算法在OTB50和OTB100上表现都不如TAN,在该数据集上却表现良好,本文认为是因为它们使用了在线模型更新策略,这对长时序列能更好地表示目标的变化,而本文所用的存储库机制不足以应对长时序列变化的情况。与其他跟踪器相比,跟踪器TAN在性能和速度上都有明显优势。

3.2.4 在TrackingNet数据集上对比实验

TrackingNet大规模数据集由在YouTube上采集的真实视频组成,包含30 000个序列,1 400万个标注及511个测试序列,涵盖了不同的对象类别和场景。实验对511个测试序列采用在线测评的方式,评价指标为成功率、精度和归一化精度,遵循测试规则对所提出的跟踪器TAN进行测试,且与先进跟踪器进行对比,结果如表2所示。可以看出,TAN的表现与在LaSOT数据集中的对比相似,ATOM和DiMP18性能优于TAN,但是跟踪器TAN速度更快,且简单易用,能够部署到一些不支持在线模型更新和硬件计算资源有限的平台上。

3.2.5 在UAV123数据集上对比实验

UAV123数据集包含由无人机捕获的123个视频序列,平均序列长度为915帧,所有视频帧都有边界框标注。数据集包含多种困难场景,如快速运动、尺度变化、光照变化和遮挡等,对跟踪器具有一定的

表 1　先进跟踪器在数据集 LaSOT 上的对比结果

Table 1　Comparison of the quality of state-of-the-art tracking methods on LaSOT dataset

方法	骨干网络	更新策略	成功率	精度	归一化精度	速度/(帧/s)
MDNet	VGG-M	使用	0.397	0.373	0.460	1
ECO	VGG-M	使用	0.324	0.301	0.338	8
ROAM	VGG16	使用	0.390	0.368	–	13
ATOM	Res18	使用	0.515	–	**0.576**	30
DiMP18	Res18	使用	**0.532**	–	–	57
SINT	VGG/AlexNet	不使用	0.314	0.295	–	–
SiamFC	–	不使用	0.336	0.339	0.420	86
DaSiamRPN	AlexNet	不使用	0.415	–	0.496	**160**
TAN	Res18	不使用	0.471	**0.456**	0.538	70

注:加粗字体表示各列最优结果,“–”表示无对应结果。

表 2　先进跟踪器在数据集 TrackingNet test set 上的对比结果

Table 2　Comparison of the quality of state-of-the-art tracking methods on TrackingNet test set

方法	骨干网络	更新策略	成功率	精度	归一化精度	速度/(帧/s)
MDNet	VGG-M	使用	0.606	0.565	0.705	1
ECO	VGG-M	使用	0.554	0.492	0.618	8
ROAM	VGG16	使用	0.620	0.547	0.695	13
ATOM	Res18	使用	0.703	0.648	0.771	30
DiMP18	Res18	使用	**0.723**	**0.666**	**0.785**	57
SiamFC	–	不使用	0.571	0.533	0.663	86
DaSiamRPN	AlexNet	不使用	0.638	0.591	0.733	**160**
TAN	Res18	不使用	0.685	0.642	0.757	70

注:加粗字体表示各列最优结果,“–”表示无对应结果。

挑战性。实验与 OTB 评测方式和指标相同,但 UAV123 数据集的序列长度更长,难度更大。表 3 展示了不同跟踪器在该数据集上的对比结果。可以看出,TAN 的性能(59.2%)优于 MDNet(52.8%)、

表 3　先进跟踪器在数据集 UAV123 上的对比结果

Table 3　Comparison of the quality of state-of-the-art tracking methods on UAV123 dataset

方法	骨干网络	更新策略	成功率	精度	归一化精度	速度/(帧/s)
MDNet	VGG-M	使用	0.528	–	–	1
ECO	VGG-M	使用	0.525	0.741	–	8
CCOT	VGG-M	使用	0.513	–	–	0.3
ATOM	Res18	使用	**0.650**	–	–	30
DiMP18	Res18	使用	0.643	–	–	57
SiamRPN	AlexNet	不使用	0.527	0.748	–	**160**
DaSiamRPN	AlexNet	不使用	0.586	0.796	–	**160**
TAN	Res18	不使用	0.592	**0.807**	**0.761**	70

注:加粗字体表示各列最优结果,“–”表示无对应结果。

ECO(52.5%)和CCOT(51.3%)等使用模型更新的方法。与OTB相比,UAV123中序列长度较长,在线模型更新对性能提升有很大帮助,使用在线更新的跟踪器的ATOM和DiMP18的性能优于未使用的SiamRPN和DaSiamRPN。本文使用的存储库机制也难以应对长序列中一些目标漂移、消失等问题,导致跟踪失败。这个问题在未来会继续探索,寻求更有效的存储库维护和更新策略。

3.3 消融实验

为了分析验证所提出的每个模块的有效性,在数据集OTB100上进行消融实验。

3.3.1 多尺度特征融合的消融实验

如第2.1节所述,通过多尺度构建层级特征可以精确定位目标位置。实验对比了融合浅层特征res0和深层特征res4到不同其他层级特征的效果,这两层级特征在以往跟踪器中很少使用。表4展示了这两层级特征对跟踪器效果的有效性。可以看出,与仅使用res1—res3相比,添加res0能够在成功率和精度上分别带来2.5%和1.3%的增益;与仅使用res1—res4相比,分别提高2.8%和3.7%。同理,与仅使用res1—res3相比,添加res4能够分别带来1.6%和0.9%的提升,与仅使用res0—res3相比,两个指标分别提升了1.9%和3.3%。实验结果表明,使用res0—res4特征融合能得到更好的效果。

表4 多尺度特征融合的消融实验结果

Table 4 The ablation study results of mulit-scale feature fusion

特征融合	成功率	精度
res1—res3	0.638	0.847
res0—res3	0.663	0.860
res1—res4	0.654	0.856
res0—res4	**0.682**	**0.893**

注:加粗字体表示各列最优结果。

3.3.2 时间聚合模块的消融实验

为验证时间聚合模块(TAM)的有效性,对TAM进行消融实验,结果如表5所示。可以看出增加TAM后,成功率和精度分别达到了68.2%和89.3%,提高了3.2%和5.1%。此外,消融实验还测试了网络共享TAM的效果,角点检测头和置信度预测头共享TAM模块时,性能比分开使用TAM分别差了0.6%和1.4%。实验结果表明,时间聚合模块通过融合多个模板的特征,能够明显提升跟踪性能,使用分离TAM能获得更好的效果。

表5 时间聚合模块的消融实验结果

Table 5 The ablation study results of temporal aggregation modules

模型	成功率	精度
不使用TAM	0.650	0.842
角点检测头和置信度预测头共享使用TAM	0.676	0.879
角点检测头和置信度预测头分开使用TAM	**0.682**	**0.893**

注:加粗字体表示各列最优结果。

3.3.3 时间融合模块的消融实验

如第2.3节所述,多个时间特征融合的方式有1×1卷积、最大池化和平均池化3种。1×1卷积是使用有一个输出通道的普通卷积层来减少特征的时间通道维度。最大池化和平均池化是两个普通的池化层。时间融合模块的消融实验结果如表6所示。通过实验发现,使用平均池化能获得比其他两种方式更好的效果。平均池化与1×1卷积核和最大池化相比,成功率和精度分别高出1%、2.7%和0.9%、1.8%。实验结果表明,使用平均池化融合能得到更好的效果。

表6 时间融合模块的消融实验结果

Table 6 The ablation study results of temporal fusion modules

融合方式	成功率	精度
1×1卷积	0.672	0.866
最大池化	0.673	0.875
平均池化	**0.682**	**0.893**

注:加粗字体表示各列最优结果。

3.3.4 模板数量的消融实验

TAN可以在一次前向推理过程中为单帧搜索区域编码多个目标模板。为此,对输入模板数量K进行了试验,对比指标为OTB100上的成功率,验证其对跟踪效果的影响。根据式(7)采样K个目标模板。当$K=1$,仅使用第1帧的模板。图8展示了使用不同K值时的算法跟踪成功率(AUC)。可以看出,K值增加时,跟踪性能得到提升,并在$K=3$处达到峰值,再进一步增加时,性能

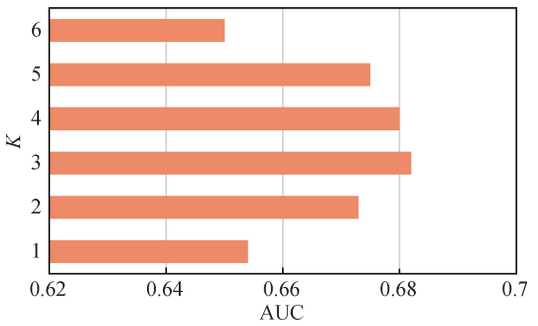


图 8　输入模板数量的消融实验结果

Fig. 8　The results of ablation study of input template number

逐渐下降。不断增加 K 值导致性能变差的原因是额外的模板会带来更多干扰，这与第 2.4 节使用简单模板采样会引入低质量模板有密切关系。该问题将在未来工作中继续改进。因此，在对比实验中为了得到更好效果，设置 $K = 3$。即使用中心点置信度分数能得到更好的效果。

3.3.5　目标框选择机制的消融实验

由于网络框架输出包含中心置信度图和 4 个角点。4 个角点分别由“左上—右下”和“右上—左下”构成两组边界框。对最终目标框的确定方式，本文尝试了 3 种不同方案，即 4 个角点的置信度均值、中心点最大置信度分数值、4 个角点置信度分数和中心点置信度分数的均值。实验结果如表 7 所示。可以看出，通过中心点最大值能够得到最好结果。原因是 4 个角点的分数实际来自 4 个不同通道的角点热度图，这 4 个通道上的热度图峰值能够确定角点位置，但其具体值不在统一的尺度范围内，而中心点置信度在同一个热度图中，能够准确反映对比结果。

表 7　目标框选择机制的消融实验结果

Table 7　The ablation study results of box selection mechanism

类型	成功率	精度
4 个角点的均值	0.669	0.872
中心点最大值	**0.682**	**0.893**
4 个角点和中心点的均值	0.671	0.868

注：加粗字体表示各列最优结果。

3.3.6　速度分析

表 8 对比了与本文方法最相关 ROAM、ATOM 和 DiMP18 等方法在速度—精度上的表现。这 3 种方法与本文方法出发点一致，都试图解决对历史目标跟踪结果的利用，提出了不同的网络结构来提取时序特征。从表 8 可以看出，ROAM 虽然达到了与 TAN 相当的精度，但速度比 TAN 慢了 4.3 倍。ATOM 与 DiMP18 虽然使用与 TAN 一样的基础网络 ResNet-18，但 TAN 在精度和速度上均明显占优。这 3 种方法都依赖在线网络微调改善对目标特征变化的适配性，而 TAN 无需依赖在线模型更新，可以在一次前向推理过程中为目标编码多个目标模板，因此整体跟踪过程是一个纯推理过程，在保证速度的同时可以取得较好的跟踪精度，能够达到速度与精度的平衡。

表 8　不同方法的速度与精度分析结果

Table 8　The results of success and speed for different methods

方法	成功率	速度/(帧/s)
ROAM	0.681	13
ATOM	0.663	30
DiMP18	0.660	57
TAN	**0.683**	**70**

注：加粗字体表示各列最优结果。

4　结　论

从简单实用和轻量级角度出发，通过使用视频时序特性中的时序外观一致性，提出一种新颖有效的跟踪器，设计新的时间感知网络 TAN，通过提出的时间聚合模块提取时空特征，交换和融合来自不同历史帧的信息。同时，设计一个简单有效的目标估计策略检测目标的 4 个角点，并基于中心点置信分数机制确定最终目标框。本文提出的跟踪器完全离线训练，在前向推理中完全不需要任何在线模型更新策略，能够适应目标的外观变化。在实验中，相比其他轻量级的跟踪器，TAN 不包含复杂的性能提升策略，以 70 帧/s 的速度实现了更优或相当的性能。本文工作为单目标跟踪提供了一个新的研究视角，若结合使用跟踪领域常用的模型更新、目标重检测等策略，可进一步增强提出的跟踪框架性能。同时，本文工作也存在提升空间，未来将从更好的时间聚合模块设计、更强的目标估计策略、更好的存储库

维护和更新机制以及有效的模型更新策略等方面进行改进。

参考文献(References)

Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S. 2016. Fully-convolutional siamese networks for object tracking//Proceedings of European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 850-865 [DOI: 10.1007/978-3-319-48881-3_56]

Bhat G, Danelljan M, van Gool L and Timofte R. 2019. Learning discriminative model prediction for tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6181-6190 [DOI: 10.1109/ICCV.2019.00628]

Chen C, Deng Z H, Gao Y L and Wang S T. 2020. Single target tracking algorithm based on multi-fuzzy kernel fusion. Journal of Frontiers of Computer Science and Technology, 14(5): 848-860 (陈晨, 邓赵红, 高艳丽, 王士同. 2020. 多模糊核融合的单目标跟踪算法. 计算机科学与探索, 14(5): 848-860) [DOI: 10.3778/j.issn.1673-9418.1901063]

Chen Z D, Zhong B N, Li G R, Zhang S P and Ji R R. 2020. Siamese box adaptive network for visual tracking//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6667-6676 [DOI: 10.1109/CVPR42600.2020.00670]

Dai K N, Zhang Y H, Wang D, Li J H, Lu H C and Yang X Y. 2020. High-performance long-term tracking with meta-updater//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6297-6306 [DOI: 10.1109/CVPR42600.2020.00633]

Danelljan M, Bhat G, Khan F S and Felsberg M. 2019. ATOM: accurate tracking by overlap maximization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4655-4664 [DOI: 10.1109/CVPR.2019.00479]

Danelljan M, van Gool L and Timofte R. 2020. Probabilistic regression for visual tracking//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7181-7190 [DOI: 10.1109/CVPR42600.2020.00721]

Ding H and Zhang W S. 2012. Multi-target tracking approach combined with SPA occlusion segmentation. Journal of Image and Graphics, 17(1): 90-98 (丁欢, 张文生. 2012. 融合SPA遮挡分割的多目标跟踪方法. 中国图象图形学报, 17(1): 90-98) [DOI: 10.11834/jig.20120113]

Du F, Liu P, Zhao W and Tang X L. 2020. Correlation-guided attention for corner detection based visual tracking//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6835-6844 [DOI: 10.1109/CVPR42600.2020.00687]

Fan H, Lin L T, Yang F, Chu P, Deng G, Yu S J, Bai H X, Xu Y, Liao C Y and Ling H B. 2019. LaSOT: a high-quality benchmark for large-scale single object tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5369-5378 [DOI: 10.1109/CVPR.2019.00552]

Fan H and Ling H B. 2019. Siamese cascaded region proposal networks for real-time visual tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7944-7953 [DOI: 10.1109/CVPR.2019.00814]

Gao P, Yuan R Y, Wang F, Xiao L Y, Fujita H and Zhang Y. 2020. Siamese attentional keypoint network for high performance visual tracking. Knowledge-Based Systems, 193: #105448 [DOI: 10.1016/j.knosys.2019.105448]

Gong H Y, Ren H G, Shi T and Li F J. 2018. Sparse subspace single target tracking algorithm based on improved particle filtering. Modern Electronics Technique, 41(13): 10-13 (宫海洋, 任红格, 史涛, 李福进. 2018. 基于改进粒子滤波的稀疏子空间单目标跟踪算法. 现代电子技术, 41(13): 10-13) [DOI: 10.16652/j.issn.1004-373x.2018.13.003]

Guo D Y, Wang J, Cui Y, Wang Z H and Chen S Y. 2020. SiamCAR: siamese fully convolutional classification and regression for visual tracking//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6268-6276 [DOI: 10.1109/CVPR42600.2020.00630]

Hare S, Golodetz S, Saffari A, Vineet V, Cheng M M, Hicks S L and Torr P H S. 2016. Struck: structured output tracking with kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(10): 2096-2109 [DOI: 10.1109/TPAMI.2015.2509974]

Held D, Thrun S and Savarese S. 2016. Learning to track at 100 fps with deep regression networks//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 749-765 [DOI: 10.1007/978-3-319-46448-0_45]

Henriques J F, Caseiro R, Martins P and Batista J. 2015. High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(3): 583-596 [DOI: 10.1109/TPAMI.2014.2345390]

Huang L H, Zhao X and Huang K Q. 2019a. Bridging the gap between detection and tracking: a unified approach//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3998-4008 [DOI: 10.1109/ICCV.2019.00410]

Huang L H, Zhao X and Huang K Q. 2019b. GOT-10k: a large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(5): 1562-1577 [DOI: 10.1109/TPAMI.2019.2957464]

Law H and Deng J. 2019. CornerNet: detecting objects as paired key-

points[EB/OL]. [2019-05-18]. https://arxiv.org/pdf/1808.01244.pdf

Li B, Wu W, Wang Q, Zhang F Y, Xing J L and Yan J J. 2019. SiamRPN++: evolution of siamese visual tracking with very deep networks//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4277-4286 [DOI: 10.1109/CVPR.2019.00441]

Li B, Yan J J, Wu W, Zhu Z and Hu X L. 2018. High performance visual tracking with siamese region proposal network//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8971-8980 [DOI: 10.1109/CVPR.2018.00935]

Li Q, Qin Z K, Zhang W B and Zheng W. 2020. Siamese keypoint prediction network for visual object tracking[EB/OL]. [2020-06-07]. https://arxiv.org/pdf/2006.04078.pdf

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Müller M, Smith N and Ghanem B. 2016. A benchmark and simulator for UAV tracking//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 445-461 [DOI: 10.1007/978-3-319-46448-0_27]

Müller M, Bibi A, Giancola S, Alsubaihi S and Ghanem B. 2018. TrackingNet: a large-scale dataset and benchmark for object tracking in the wild//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 300-327 [DOI: 10.1007/978-3-030-01246-5_19]

Nam H, Baek M and Han B. 2016. Modeling and propagating CNNs in a tree structure for visual tracking [EB/OL]. [2020-08-25]. https://arxiv.org/pdf/1608.07242.pdf

Ning J F, Zhao Y B and Shi W Z. 2014. Multiple instance learning based object tracking with multi-channel Haar-like feature. Journal of Image and Graphics, 19(7): 1038-1045 (宁纪锋, 赵耀博, 石武祯. 2014. 多通道 Haar-like 特征多示例学习目标跟踪. 中国图象图形学报, 19(7): 1038-1045) [DOI: 10.11834/jig.20140707]

Park E and Berg A C. 2018. Meta-tracker: fast and robust online adaptation for visual object trackers//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 587-604 [DOI: 10.1007/978-3-030-01219-9_35]

Ren X Y, Liao Y T, Zhang G L and Zhang T X. 2002. A new correlation tracking method. Journal of Image and Graphics, 7(6): 553-557 (任仙怡, 廖云涛, 张桂林, 张天序. 2002. 一种新的相关跟踪方法研究. 中国图象图形学报, 7(6): 553-557) [DOI: 10.3969/j.issn.1006-8961.2002.06.006]

Song J F, Miao Q G, Wang C X, Xu H and Yang J. 2021. Multi-scale single object tracking based on the attention mechanism. Journal of Xidian University, 48(5): 110-116 (宋建锋, 苗启广, 王崇晓, 徐浩, 杨瑾. 2021. 注意力机制的多尺度单目标跟踪算法. 西安电子科技大学学报, 48(5): 110-116) [DOI: 10.19665/j.issn1001-2400.2021.05.014]

Tao R, Gavves E and Smeulders A W M. 2016. Siamese instance search for tracking//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1420-1429 [DOI: 10.1109/CVPR.2016.158]

Wang Q, Zhang L, Bertinetto L, Hu W M and Torr P H S. 2019. Fast online object tracking and segmentation: a unifying approach//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 1328-1338 [DOI: 10.1109/CVPR.2019.00142]

Wang X and Tang Z M. 2010. Application of particle filter based on feature fusion in small IR target tracking. Journal of Image and Graphics, 15(1): 91-97 (王鑫, 唐振民. 2010. 基于特征融合的粒子滤波在红外小目标跟踪中的应用. 中国图象图形学报, 15(1): 91-97) [DOI: 10.11834/jig.20100115]

Wu Y, Lim J and Yang M H. 2013. Online object tracking: a benchmark//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 2411-2418 [DOI: 10.1109/CVPR.2013.312]

Xu Y D, Wang Z Y, Li Z X, Yuan Y and Yu G. 2020. SiamFC++: towards robust and accurate visual tracking with target estimation guidelines//Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 12549-12556 [DOI: 10.1609/aaai.v34i07.6944]

Yang T Y and Chan A B. 2018. Learning dynamic memory networks for object tracking//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 153-169 [DOI: 10.1007/978-3-030-01240-3_10]

Yang T Y and Chan A B. 2021. Visual tracking via dynamic memory networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(1): 360-374 [DOI: 10.1109/TPAMI.2019.2929034]

Yang T Y, Xu P F, Hu R B, Chai H and Chan A B. 2020. ROAM: recurrently optimizing tracking model//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6717-6726 [DOI: 10.1109/CVPR42600.2020.00675]

Yu Y C, Xiong Y L, Huang W L and Scott M R. 2020. Deformable siamese attention networks for visual object tracking//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6727-6736 [DOI: 10.1109/CVPR42600.2020.00676]

Zhang Z P, Peng H W, Fu J L, Li B and Hu W M. 2020. Ocean: object-aware anchor-free tracking//Proceedings of the 16th European Conference Computer Vision. Glasgow, UK: Springer: 771-787 [DOI: 10.1007/978-3-030-58589-1_46]

Zhou X Y, Wang L, Ma Y X and Chen P B. 2021. Single object tracking of LiDAR point cloud combined with auxiliary deep neural network. Chinese Journal of Lasers, 48(21): #2110001 (周笑宇, 王玲, 马燕新, 陈沛铂. 2021. 融合附加神经网络的激光雷达点云单目标跟踪. 中国激光, 48(21): #2110001) [DOI: 10.3788/CJL202148.2110001]

Zhou W Z, Wen L Y, Zhang L B, Du D W, Luo T J and Wu Y J. 2020. SiamMan: siamese motion-aware network for visual tracking [EB/OL]. [2020-01-18]. https://arxiv.org/pdf/1912.05515.pdf

Zhu Z, Huang G, Zou W, Du D L and Huang C. 2017. UCT: learning unified convolutional networks for real-time visual tracking//Proceedings of 2017 IEEE International Conference on Computer Vision Workshops. Venice, Italy: IEEE: 1973-1982 [DOI: 10.1109/ICCVW.2017.231]

Zhu Z, Wang Q, Li B, Wu W, Yan J J and Hu W M. 2018. Distractor-aware siamese networks for visual object tracking//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 103-119 [DOI: 10.1007/978-3-030-01240-3_7]

Zuo W M, Wu X H, Lin L, Zhang L and Yang M H. 2019. Learning support correlation filters for visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(5): 1158-1172 [DOI: 10.1109/TPAMI.2018.2829180]

作者简介

王蒙蒙,1994 年生,女,博士研究生,主要研究方向为计算机视觉、深度学习、目标跟踪、行为识别。
E-mail: mengmengwang@zju.edu.cn

刘勇,通信作者,男,教授,主要研究方向为机器人感知和视觉、深度学习、数据挖掘、多传感器融合、机器学习和计算机视觉。
E-mail: yongliu@iipc.zju.edu

杨小倩,女,硕士研究生,主要研究方向为计算机视觉、深度学习、人体姿态跟踪、目标跟踪。
E-mail: yxiaoqian@zju.edu.cn