**中图法分类号**:TP389.1　　**文献标识码**:A　　**文章编号**:1006-8961(2024)08-2269-34

**论文引用格式**:Hu S Y, Zhao X and Huang K Q. 2024. Visual intelligence evaluation techniques for single object tracking: a survey. Journal of Image and Graphics, 29(08):2269-2302(胡世宇, 赵鑫, 黄凯奇. 2024. 单目标跟踪中的视觉智能评估技术综述. 中国图象图形学报, 29(08):2269-2302)[DOI:10.11834/jig.230498]

# 单目标跟踪中的视觉智能评估技术综述

胡世宇[1,2],赵鑫[1,2],黄凯奇[1,2,3*]

1. 中国科学院大学人工智能学院,北京　100049; 2. 中国科学院自动化研究所智能系统与工程研究中心,北京　100190;
3. 中国科学院脑科学与智能技术卓越创新中心,上海　200031

**摘　要**:单目标跟踪任务旨在对人类动态视觉系统进行建模,让机器在复杂环境中具备类人的运动目标跟踪能力,并已广泛应用于无人驾驶、视频监控、机器人视觉等领域。研究者从算法设计的角度开展了大量工作,并在代表性数据集中表现出良好性能。然而,在面临如目标形变、快速运动、光照变化等挑战因素时,现有算法的跟踪效果和人类预期相比还存在着较大差距,揭示了当前的评测技术发展仍存在滞后性和局限性。综上,区别于以算法设计为核心的传统综述思路,本文依托单目标跟踪任务、从视觉智能评估技术出发,对评测流程中涉及的各个关键性环节(评测任务、评测环境、待测对象和评估机制)进行系统梳理。首先,对单目标跟踪任务的发展历程和挑战因素进行介绍,并详细对比了评估所需的评测环境(数据集、竞赛等)。其次,对单目标跟踪待测对象进行介绍,不仅包含以相关滤波和孪生神经网络为代表的跟踪算法,同时也涉及跨学科领域开展的人类视觉跟踪实验。最后,从"机机对抗"和"人机对抗"两个角度对单目标跟踪评估机制进行回顾,并对当前待测对象的目标跟踪能力进行分析和总结。在此基础上,对单目标跟踪智能评估的发展趋势进行总结和展望,进一步分析未来研究中存在的挑战因素,并探讨了下一步可能的研究方向。

**关键词**:智能评估技术;竞赛和数据集;视觉跟踪能力;单目标跟踪(SOT);目标跟踪算法

# Visual intelligence evaluation techniques for single object tracking: a survey

Hu Shiyu[1,2], Zhao Xin[1,2], Huang Kaiqi[1,2,3*]

*1. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing　100049, China;*
*2. Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences, Beijing　100190, China;*
*3. Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai　200031, China*

**Abstract**: Single object tracking (SOT) task, which aims to model the human dynamic vision system and accomplish human-like object tracking ability in complex environments, has been widely used in various real-world applications like self-driving, video surveillance, and robot vision. Over the past decade, the development in deep learning has encouraged many research groups to work on designing different tracking frameworks like correlation filter (CF) and Siamese neural networks (SNNs), which facilitate the progress of SOT research. However, many factors (e.g., target deformation, fast

**收稿日期**:2023-07-10;**修回日期**:2023-10-07;**预印本日期**:2023-10-14

*通信作者:黄凯奇　kqhuang@nlpr.ia.ac.cn

**基金项目**:科技创新2030——"新一代人工智能"重大项目(2022ZD0116403);国家自然科学基金项目(61721004);中国科学院战略性先导科技专项(XDA27000000)

**Supported by**: Technological Innovation 2030——"New Generation Artificial Intelligence" Major Project (2022ZD0116403); National Natural Science Foundation of China (61721004); Strategic Priority Research Program of Chinese Academy of Sciences (XDA27000000)

motion, and illumination changes) in natural application scenes still challenge the SOT trackers. Thus, algorithms with novel architectures have been proposed for robust tracking and to achieve better performance in representative experimental environments. However, several poor cases in natural application environments reveal a large gap between the performance of state-of-the-art trackers and human expectations, which motivates us to pay close attention to the evaluation aspects. Therefore, instead of the traditional reviews that mainly concentrate on algorithm design, this study systematically reviews the visual intelligence evaluation techniques for SOT, including four key aspects: the task definition, evaluation environments, task executors, and evaluation mechanisms. First, we present the development direction of task definition, which includes the original short-term tracking, long-term tracking, and the recently proposed global instance tracking. With the evolution of the SOT definition, research has shown a progress from perceptual to cognitive intelligence. We also summarize challenging factors in the SOT task to help readers understand the research bottlenecks in actual applications. Second, we compare the representative experimental environments in SOT evaluation. Unlike existing reviews that mainly introduce datasets based on chronological order, this study divides the environments into three categories (i. e. , general datasets, dedicated datasets, and competition datasets) and introduces them separately. Third, we introduce the executors of SOT tasks, which not only include tracking algorithms represented by traditional trackers, CF-based trackers, SNN-based trackers, and Transformer-based trackers but also contain human visual tracking experiments conducted in interdisciplinary fields. To our knowledge, none of the existing SOT reviews have included related works on human dynamic visual ability. Therefore, introducing interdisciplinary works can also support the visual intelligence evaluation by comparing machines with humans and better reveal the intelligence degree of existing algorithm modeling methods. Fourth, we review the evaluation mechanism and metrics, which encompass traditional machine-machine and novel human-machine comparisons, and analyze the target tracking capability of various task executors. We also provide an overview of the human-machine comparison named visual Turing test, including its application in many vision tasks (e. g. , image comprehension, game navigation, image classification, and image recognition). Especially, we hope that this study can help researchers focus on this novel evaluation technique, better understand the capability bottlenecks, further explore the gaps between existing methods and humans, and finally achieve the goal of algorithmic intelligence. Finally, we indicate the evolution trend of visual intelligence evaluation techniques: 1) designing more human-like task definitions, 2) constructing more comprehensive and realistic evaluation environments, 3) including human subjects as task executors, and 4) using human abilities as a baseline to evaluate machine intelligence. In conclusion, this study summarizes the evolution trend of visual intelligence evaluation techniques for SOT task, further analyzes the existing challenge factors, and discusses the possible future research directions.

**Key words**: intelligence evaluation technique; competitions and datasets; visual tracking ability; single object tracking (SOT); object tracking algorithms

## 0 引 言

作为计算机视觉领域的热点研究方向之一,单目标跟踪(single object tracking, SOT)任务旨在对人类动态视觉系统进行建模,让机器在复杂环境中具备类人的运动目标跟踪能力。如图1所示,单目标跟踪的基本定义是:在一段视频序列的首帧指定任意一个目标,并在后续帧持续对该运动目标进行定位(卢湖川 等,2018;李玺 等,2019)。与已有视觉任务相比,单目标跟踪具有时序决策(Yun 等,2017)、类别无关(Wu 等,2013)和实例预测(Huang 等,2019)的特性。其中,时序决策指异于基于图像的检测、识别和分割等静态任务,单目标跟踪需要借助视频相邻帧的时序信息对运动物体进行定位;类别无关即单目标跟踪算法未对目标类别进行任何假设,其可以在开集(open-set)测试环境对任意物体进行定位;实例预测指单目标跟踪需要将目标与视频中的其他物体(包含同类物体)进行区分。

单目标跟踪的任务特点使其具有重要的研究意义和广阔的应用前景。首先,类别无关特性使算法可以在时序上对任意目标的定位结果进行关联,并在检测结果的基础上进行每个目标的身份匹配并生成运动轨迹(Xiang 等,2015)。此外,单目标跟踪可

以在仅提供首帧标注的情况下挖掘视频中的运动物体，并在后续序列中生成大量图像样本，为基于深度学习的视觉类任务提供丰富的数据信息（Wang 等，2017b）。因此，单目标跟踪在智能监控、机器人视觉、无人驾驶等众多应用场景中均发挥着重要作用。单目标跟踪的任务特点也使其面临诸多的挑战因素和技术难点。例如，类别无关的任务特点使得算法难以对目标进行预先建模。此外，目标的表观信息和运动信息在跟踪过程中会因目标形变、快速运动和光照变化等因素的影响而发生改变，对算法的鲁棒性提出较高要求。针对上述问题，众多研究者投入到单目标跟踪任务的算法设计中，并以探寻人类动态视觉系统的感知能力、实现类人鲁棒视觉跟踪为目标开展大量工作。传统方法包含运动模型（Ross 等，2008）、特征表达（Henriques 等，2015）、表观模型（Henriques 等，2012）和算法更新（Wang 等，2016）4 个步骤，并分别进行优化。随着单目标跟踪数据集从小规模（万帧）到大规模（百万帧）的发展，基于数据驱动的深度学习目标跟踪方法（Li 等，2018；Zhu 等，2018；Yan 等，2019；Huang 等，2020）将表观模型与其他模块进行融合，在跟踪准确性和鲁棒性上取得较大提升。

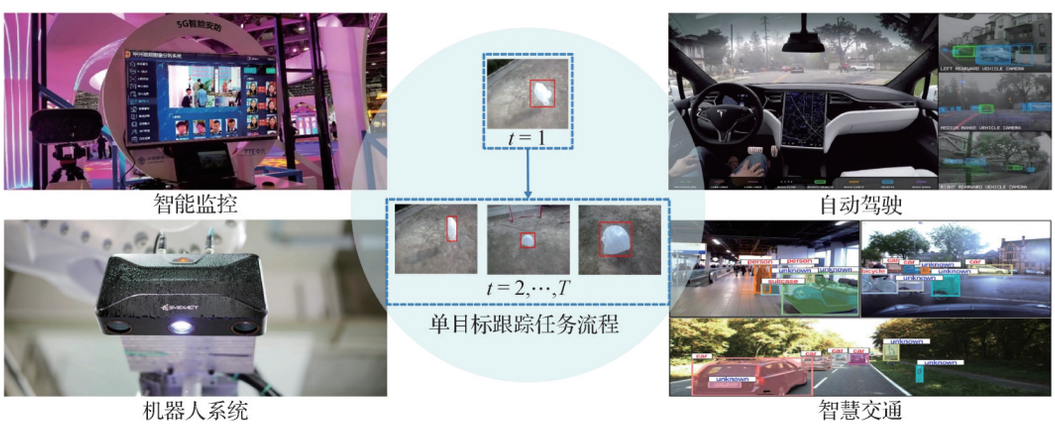



图 1　单目标跟踪执行流程

Fig. 1　The execution process of single object tracking

虽然基于大数据、大算力和大模型的深度学习跟踪方法已经在代表性评测环境中取得良好性能，但其在真实环境中仍然存在鲁棒性不足的现象，与人类的视觉跟踪能力存在较大差距（Hu 等，2023）。例如，当面对快速运动、镜头切换和场景转换等广泛存在于真实应用场景的挑战因素时，人类通常可以在不受干扰的情况下持续定位目标，但算法会因为目标的表观信息和运动信息被破坏而出现跟踪失败。上述性能瓶颈造成大量单目标跟踪算法仍停留在理论研究阶段，距离应用落地仍有较大距离。

究其原因，目前单目标跟踪领域的关注焦点主要集中在算法设计层面，而对于评估技术的研究则相对较少，导致其发展相对滞后，并难以深入挖掘算法的性能瓶颈。这一现象也揭示了评估技术的重要性：算法设计和评估技术就像硬币的两面，开展单目标跟踪的智能评估技术研究，不仅可以为算法的部署与应用提供保证，也可以为算法下一步的发展指明方向。虽然已经有部分学者在代表性综述中提及单目标跟踪评测技术的相关内容（如表 1 所示），但其主要以算法为核心，并从算法设计、模型结构、技术细节等角度开展综述，而通常以较短的篇幅概括评测流程中所涉及的评测环境和评估机制，且仅局限于以“机机对抗”为核心的传统算法评估思路，忽略了对任务评估目的和评估技术发展的思考。

综上，区别于以算法设计为核心的传统综述思路（韩瑞泽 等，2022；Marvasti-Zadeh 等，2022；Javed 等，2023；李成龙 等，2023），本文从单目标跟踪的智能评估技术出发，对一个完整的评测系统需要考虑的 4 个关键环节（评测任务理解、评测环境构建、待测对象建模和评估机制设计）进行详细介绍，并对各个环节之间的关联以及每个环节在完整评测系统中的作用进行阐述，旨在从评估评测的视角为单目标跟踪的下一步研究提供支撑。

1）评测任务理解。对任务定义的深入理解和分析是开展智能评估的基础，只有在深入理解任务特

表1　单目标跟踪领域代表性综述介绍

Table 1　Representative single object tracking surveys

| 综述 | 期刊 | 评测环境 | 待测对象 | 评估机制 |
|---|---|---|---|---|
| 多模态视觉跟踪方法综述(李成龙 等,2023) | 中国图象图形学报 | 介绍5个多模态跟踪数据集(4%) | 介绍以可见光—红外跟踪为主的多模态跟踪算法(71%) | 未涉及 |
| 视频单目标跟踪研究进展综述(韩瑞泽 等,2022) | 计算机学报 | 介绍SOT常见数据集及竞赛(8%) | 介绍SOT代表性方法,主要包含相关滤波和孪生神经网络两类,也涉及针对特定跟踪挑战的方法介绍(50%) | 主要介绍SOT传统的机机对抗指标(4%) |
| Deep learning for visual tracking: a comprehensive survey(Marvasti-Zadeh等,2022) | IEEE Transactions on Intelligent Transportation Systems | 介绍SOT常见数据集及竞赛(8%) | 介绍SOT代表性方法,主要包含对模型结构、执行流程、跟踪范式等各个模块的介绍(55%) | 主要介绍SOT传统的机机对抗指标(5%) |
| Visual object tracking with discriminative filters and siamese networks: a survey and outlook(Javed等,2023) | IEEE Transactions on Pattern Analysis and Machine Intelligence | 介绍SOT常见数据集及竞赛(9%) | 介绍单目标跟踪代表性方法,主要包含相关滤波和孪生神经网络两类(59%) | 主要介绍SOT传统的机机对抗指标(3%) |
| 本文 | 中国图象图形学报 | 从任务设定角度分别介绍SOT通用评测环境、专用评测环境及竞赛评测环境(18%) | 将待测对象分为机器和人类,首先介绍传统方法和代表性深度学习算法(15%),并介绍神经科学家在人类视觉领域开展的相关工作(12%) | 将评估机制分为“机机对抗”和“人机对抗”两类。首先从评估机制和评测指标两方面介绍传统的机机对抗(15%),随后介绍基于人机对抗思想的视觉图灵评测机制(12%) |

注:百分比代表该部分在正文内容的占比,用于对比每篇综述的论述重点。鉴于综述通常还涉及任务介绍及总结展望,因此评测环境、待测对象、评估机制通常仅占正文内容的70%左右。

点的基础上,研究者才能够有针对性地构建出评测实验环境和评测指标,并挖掘出待测对象的性能瓶颈。视觉物体跟踪旨在对运动物体进行持续定位,是人类动态视觉能力的重要组成部分。研究表明,刚出生几周的婴儿就已经初步具备运动感知能力,并可以通过快速学习实现对任意物体的跟踪(Hyvärinen等,2014)。单目标跟踪可以视做对人类视觉跟踪能力的建模,但受限于数据集规模和方法水平,任务定义通常需要添加额外的约束条件以简化难度,因此早期研究主要集中在短时跟踪(short-term tracking)中。随着视觉目标跟踪(visual object tracking, VOT)长时跟踪竞赛和长时跟踪数据集OxUvA(Valmadre等,2018)的发布,目标跟踪的研究重点拓展至长时跟踪(long-term tracking)。长时跟踪拓宽了单目标跟踪的应用场景,但其任务定义中仍隐含“目标连续运动”的假设。针对上述问题,Hu等人(2023)提出全局实例跟踪(global instance tracking, GIT),旨在弥合长时跟踪与人类视觉能力的差异,实现对人类动态视觉能力的全面建模。本文第1节首先从任务发展的角度对短时跟踪、长时跟踪和全局实例跟踪进行介绍,并进一步归纳总结出任务的挑战因素,旨在帮助研究者深入理解单目标跟踪任务,从而为后续评测环境构建、待测对象建模和评估机制设计奠定基础。

2)评测环境构建。良好的评测环境是开展智能评估评测的起点。评测环境需要直观体现评测任务的特点,并通过环境中的挑战因素来体现任务的难点,进而发现待测对象的能力瓶颈。在单目标跟踪领域,研究者早期主要利用零散的视频段对跟踪算法进行评测(Bao等,2012;Henriques等,2012;Wang等,2013;Hare等,2016),缺乏统一的基准(benchmark)对算法性能进行比较。随着人工智能研究热潮的兴起,以Wu等人(2013)提出的OTB(object tracking benchmark)基准和Kristan等人(2013)举办的VOT挑战赛为代表的系列工作为算法研究提供了标准化的评测环境,并通过规范的指标从准确性

和鲁棒性等角度衡量算法的性能，极大地推动了单目标跟踪领域的研究进展。本文第2节将单目标跟踪任务的评测环境分为通用评测环境、专用评测环境和竞赛评测环境3类，并分别进行综述。通用评测环境旨在为目标跟踪任务的评估提供一个全面且综合的评测环境；专用评测环境则针对特定场景（如无人机场景）或特定目标类型（如透明物体）进行设计，旨在研究和评估算法在特定应用场景下的跟踪性能；竞赛评测环境则以竞赛的形式发布，并且依托竞赛目标进行持续维护和更新。综上，高质量的评测环境极大地促进了单目标跟踪任务的研究，并为智能评估发展提供重要的数据支撑。

3）待测对象建模。单目标跟踪任务的待测对象通常为跟踪算法，对各个发展阶段和不同架构的跟踪算法进行评估，可以帮助研究者挖掘算法之间的性能差异，并从细粒度挑战因素下的算法性能瓶颈出发，探寻更加有效的算法优化方案。随着深度学习的发展，诸多研究者从设计高效算法出发开展了大量工作。传统方法将跟踪独立成包含表观模型、特征表达、算法更新和运动模型在内的多个模块，而基于深度学习的方法则将表观模型与其他模块进行融合，并借助大规模数据集和深度神经网络（deep neural network，DNN）来得到具有更快速度和更好性能的模型。在深度学习发展的初期，研究者主要沿用传统框架，仅在特征提取、目标定位等少数模块使用深度学习网络。一部分工作将相关滤波（correlation filter，CF）和深度神经网络进行结合（Ma 等，2015a；Danelljan 等，2016，2017；Bhat 等，2018），在提升模型性能的同时借助相关滤波方法在傅里叶域的计算效率来保障运算速度。另一部分研究者则利用孪生神经网络（siamese neural network，SNN）框架设计目标跟踪算法（Bertinetto 等，2016；Li 等，2018，2019；Zhu 等，2018；Zhang 和 Peng，2019），并取得了显著效果。本文第3节首先对传统跟踪方法进行回顾，然后对基于深度特征的相关滤波方法、基于孪生神经网络的跟踪算法和其他代表性深度学习目标跟踪方法进行介绍。此外，考虑到单目标跟踪算法旨在实现类人的目标跟踪能力，因此研究者也需要同样关注认知神经科学领域的相关工作，并加深对人类动态视觉机理的理解。综上，本文也对跨学科领域开展的人类视觉跟踪实验进行回顾，旨在为基于人类动态视觉能力度量算法智能的评估评测研究提供支撑。

4）评估机制设计。在评测任务理解、评测环境构建、待测对象建模的基础上，评估机制决定着算法研究的具体应用水平，并为下一步算法技术发展指明方向。传统的单目标跟踪任务主要采用以机机对抗为核心的评估机制，评估重点是“向下”的：研究者专注于在评测环境中对比算法的性能，新算法只要超越其他基准方法，即可认为具备优秀的视觉跟踪能力。但是，在传统评估机制中表现优异的算法却在真实任务环境中与人类视觉能力存在较大差距，引发研究者对评估机制进行变革，提出以人机对抗为核心的视觉图灵评估机制。区别于基于大数据、大算力的评估标准，视觉图灵机制将“人”的因素加入到智能评估的回路中，以人类为基准对机器智能开展更有效的评估，评估重点是“向上”的，即只有近似或者超越人类视觉能力的算法，才可以认为具备了视觉智能。本文第4节首先对传统评估机制进行介绍，并对当前代表性跟踪方法在基准数据集上的评估结果开展分析。在此基础上，本文也对视觉图灵机制在视觉任务中的应用范例进行总结，并重点分析单目标跟踪领域的视觉图灵实验，旨在对当前跟踪算法的智能发展水平进行评估。

第5节对单目标跟踪智能评估技术的关键环节进行总结，并归纳出任务定义类人化、评测环境真实化、待测对象多元化和评估方式智能化的发展趋势。在此基础上，对每个模块存在的挑战因素进行分析，并探讨了未来可能的研究方向。

## 1 评测任务

本节从任务定义和任务挑战两方面对单目标跟踪进行介绍，旨在帮助研究者全面理解评测任务，并为开展智能评估奠定基础。其中，任务定义中隐含的约束条件体现出任务特点，约束条件的变化也将导致评估侧重点的改变；任务挑战则代表任务难点，只有在全面理解挑战因素的基础上，研究者才可以针对性地设计评测环境，并通过合理的评估方式准确挖掘出待测对象的能力瓶颈。

### 1.1 任务定义

#### 1.1.1 短时跟踪任务

早期的单目标跟踪主要研究短时跟踪任务，

VOT竞赛(Kristan等,2016)将短时跟踪用5个词表述为:单目标(single-target)、单镜头(single-camera)、无模型(model-free)、短时(short-term)、因果关系(causal)的跟踪。其中,单目标、无模型和因果关系分别对应于实例预测、类别无关和序列决策,用于刻画单目标跟踪区分于其他视觉任务的本质特点;单镜头和短时则用于约束任务边界,将短时跟踪任务建立在相对简单的场景中。如图2(a)所示,短时跟踪将目标运动限制在单一镜头之下,且目标始终存在于画面中。在一段视频序列里,目标可能被局部遮挡,但并未出现明显的表观信息变化。受任务定义强约束的限制,短时跟踪通常难以直接适用于真实环境,与实际应用场景存在较大差异。

1.1.2 长时跟踪任务

取消短时约束、将单目标跟踪拓展至较长的时间序列中,是长时跟踪任务的出发点。然而,研究者对于如何精确定义长时跟踪存在分歧。一方面,目前规模最大的长时跟踪基准LaSOT(large-scale single object tracking)(Fan等,2021a)认为长时跟踪是"时长较长的短时跟踪",因此在数据采集过程中仅延长序列长度,并直接沿用经典的短时跟踪评测方法进行算法评估;另一方面,由VOT(Lukezic等,2021)组织的长时跟踪挑战赛(VOT-LT)认为长时跟踪最重要的特征是允许目标短暂的"消失—再现",因此VOT-LT要求算法具备判定目标消失的能力,并基于这一特点设计评估指标。对比两种定义,允许目标短暂的"消失—再现"更适合作为长时跟踪的决定性因素。通过取消"目标始终存在于画面中"这一隐藏在短时跟踪定义中的约束,更多"长时"视频可以纳入评测环境中,实现从短时到长时的拓展。如图2(b)所示,长时跟踪允许目标因移出画面或者完全遮挡而短暂消失,并要求算法可以对目标再次出现时进行重新定位。

1.1.3 全局实例跟踪任务

人类可以在任意场景中持续定位任意目标,但短时跟踪任务和长时跟踪任务均隐藏"目标连续运动"假设,将任务约束在"单镜头"场景之中,与人类动态视觉能力相距甚远。因此,Hu等人(2023)通过进一步取消"单镜头"约束,提出全局实例跟踪任务来建模人类动态视觉能力。值得注意的是,全局实例跟踪仅包含单目标、无模型和因果关系3项约束,从而实现对单目标跟踪任务原始定义的精准刻画。图2(c)展示了全局实例跟踪的代表性视频序列。与图2(a)(b)相比,随着逐步取消任务约束,单目标跟踪任务场景也从简单场景(短时跟踪)拓展到包含更多挑战性因素的真实场景(全局实例跟踪)中。

## 1.2 任务挑战

如图3所示,单目标跟踪任务面临诸多挑战性,这些挑战因素会改变目标的表观信息和运动信息,导致算法在序列决策的过程中出现累计误差问题,从而无法准确识别和定位目标,最终导致跟踪


图2 单目标跟踪代表性序列

Fig. 2 Representative sequence of single object tracking

((a) short-term tracking; (b) long-term tracking; (c) global instance tracking)


图3　单目标跟踪任务代表性挑战因素

Fig. 3　Representative challenging factors for single object tracking task

失败。本文将目标跟踪任务的常见挑战因素汇总如下：

1）镜头切换（shot-cut）是全局实例跟踪特有的挑战因素，存在淡出—淡入和直接切换两种形式。

2）目标缺失（absent）通常由于目标移出视野或者被完全遮挡，算法需要在目标再次出现在画面中时对其进行重新定位并继续跟踪。

3）目标遮挡（partial occlusion）指目标在运动的过程中由于被障碍物遮挡而丢失部分表观信息。大部分研究者认为超过 10% 的表观信息被遮挡即可视做目标遮挡出现，部分数据集针对遮挡问题提供了细粒度的遮挡等级标注（Li 等，2016；Huang 等，2021）。

4）背景杂乱（background clutter）是背景干扰的一种形式。目标跟踪方法的建模重点通常包括有效提取前景特征和抑制背景信息的干扰，杂乱的背景信息将模糊前景与背景的边界，增大前背景分离的难度，并极易导致算法定位到背景中的干扰物上。此外，背景杂乱对于以透明物体为代表的特殊目标跟踪任务具有极大的挑战性（Fan 等，2021b）。

5）弱光照（dim light）或者强光照（light）是另一种背景干扰，均会导致目标的表观信息发生变化，从而影响算法的跟踪性能。此外，视频序列可能涉及不同光照条件，需要算法在光照变化的过程中依旧保持鲁棒的跟踪能力。

6）运动模糊（motion blur）通常是由于目标运动造成的，会导致连续帧之间的目标表观信息清晰度发生突变，影响算法对目标的特征学习过程。此外，相机运动也会导致视频序列出现画面模糊，对前景信息和背景信息均造成较大影响。

7）目标的快速运动（fast motion）除了会导致运动模糊，也会由于目标位置信息在连续帧之间较大的差异性导致算法跟踪失败。以孪生网络结构为代表的系列跟踪方法（Bertinetto 等，2016；Li 等，2018，2019；Zhu 等，2018）通常基于上一帧的跟踪结果在当前帧设定搜索区域并检索目标位置，快速运动则会造成当前帧的目标位置超出搜索区域，导致算法难以在当前帧正确地对目标进行定位。

8）具有特殊尺寸（special scale）或者特殊形状（special ratio）的目标也会增加跟踪挑战。由于单目标跟踪的评估指标主要建立在中心点距离和矩形框交并比的基础上（详见 4.1 节），因此算法在跟踪微小目标或者长宽比差距悬殊的目标时，通常难以精确拟合目标位置，导致跟踪效果较差。

9）运动目标在视频序列中的尺寸变化（scale variation）和形状变化（ratio variation）是单目标跟踪常见的挑战因素。其中，目标与相机的相对位置发生变化会造成尺寸变化，目标的旋转运动则会造成形状变化，两种变化形式都会导致前景特征发生改变，增加算法跟踪的难度。

在评测环境和实际应用场景中，挑战因素通常组合出现，并通过对目标表观信息和运动信息的破坏，增加单目标视觉物体跟踪的难度。因此，如何在评测环境构建过程中涵盖丰富的挑战因素组合、如何在算法建模流程中针对挑战因素进行优化、如何在评估机制设计过程中有效挖掘挑战因素对算法性能的影响，是值得研究者进一步思考的问题。

## 2 评测环境

本节从通用评测环境、专用评测环境和竞赛评测环境3个方面对单目标跟踪的典型评测环境进行梳理，旨在帮助研究者理解环境特点和评测重点，从而更好地针对评估目标构建评测环境。其中，通用评测环境具备研究起步早、典型工作多和数据覆盖广的特点，旨在构建一个全面的实验平台，并检验待测对象在通用场景下的综合能力。相比之下，专用评测环境则具备“小而精”的特点，主要针对特定任务场景或特殊目标类型进行设计，旨在度量待测对象在特定测评需求下的跟踪效果。竞赛评测环境则依托竞赛进行发布，通常挑选具有高挑战性的视频序列构成测试数据集，旨在快速挖掘算法的能力瓶颈，并从多个维度对参赛算法进行能力排序。

### 2.1 通用评测环境

#### 2.1.1 短时跟踪数据集

如图4所示，在2013年以前，单目标跟踪任务的评估主要基于零散视频序列，缺乏统一的评测环境对算法性能进行分析。作为单目标跟踪领域首个为评估算法效果进行设计的基准数据集，OTB50（http://cvlab. hanyang. ac. kr/tracker\benchmark/index. html）由51段视频构成（Wu等，2013），随后拓展为包含10种运动目标类别、总计100段视频的OTB100数据集（Wu等，2015）。如图5（a）所示，OTB数据集采用水平矩形框为目标位置提供高精度的逐帧标注，为每个视频序列提供多种挑战因素的标注，并对31个代表性算法进行评测。作为规范化单目标跟踪基准的开创性工作，OTB数据集为后续基准的设计提供了良好的范例，并为早期算法研究提供了重要支撑。

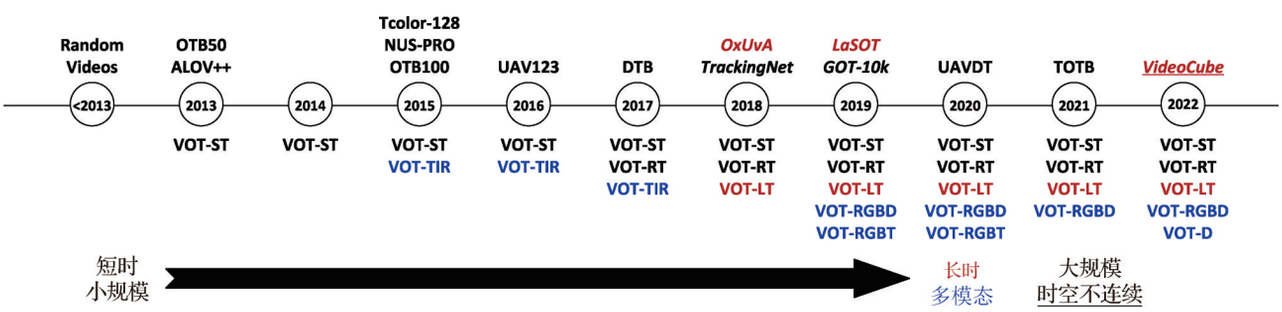



图4 单目标跟踪评测环境发展

Fig. 4 Development of the experimental environments for single object tracking task


(a) OTB100 (Wu等，2015)　(b) LaSOT (Fan等，2021a)　(d) TOTB (Fan等，2021b)

(c) VideoCube (Hu等，2023)　(e) UAV123 (Mueller等，2016)

图5 单目标跟踪代表性数据集

Fig. 5 Representative datasets for single object tracking ((a) OTB100 (Wu et al. , 2015); (b) LaSOT (Fan et al. , 2021a); (c) VideoCube (Hu et al. , 2023); (d) TOTB (Fan et al. , 2021b); (e) UAV123 (Mueller et al. , 2016))

与OTB同期的ALOV++（Amsterdam library of ordinary video）数据集（https://www.crcv.ucf.edu/research/data-sets/alov/）（Smeulders等，2014）由包含不同跟踪难度、目标类别和挑战因素的序列组成，其中包含304段较短的视频序列和10段较长的视频序列，每段序列均携带一个挑战因素标签。TColor-128数据集（https://www3.cs.stonybrook.edu/~hling/data/TColor-128/TColor-128.html）（Liang等，2015）包含78段特有数据和50段来自于OTB100的数据，覆盖27类运动目标和11种跟踪任务常见的挑战因素。区别于包含部分灰度视频序列的OTB数据集，TColor-128中的数据均为全彩序列，旨在评估颜色特征对目标跟踪算法的影响。上述数据集发布时间较早，虽然引领了后续研究的发展，但存在规模较小、实例多样性较弱、视频场景有限等缺陷。

深度神经网络广泛应用于单目标跟踪领域。深度学习方法需要大量标注数据进行驱动，但鉴于上述数据集规模的限制，目标跟踪模型通常先在针对其他视觉任务的大型数据集上进行训练。其中，Russakovsky等人（2015）构建的ImageNet-VID（imagenet video object detection）（https://image-net.org/index.php）和Real等人（2017）构建的YouTube-BB（https://github.com/mbuckler/youtube-bb）是两个代表性的大规模视频物体检测数据集。ImageNet-VID在5 400段视频上对一个或多个运动物体进行标注，YouTube-BB（YouTube bounding boxes）则包含了来自YouTube的38万段视频，并以1 Hz的人工标注频率对540万帧进行注释。虽然二者规模较大，但物体类别较少且包含静态物体，不利于模型训练。因此，如表2所示，若干针对短时跟踪的大型数据集被发布，旨在为基于深度学习的跟踪方法提供充足的高质量训练数据。

Müller等人（2018）构建的TrackingNet数据集（https://tracking-net.org/）包含3万段视频序列和1 400万目标矩形框，覆盖27类运动物体和15种挑战因素。该数据集对YouTube-BB中包含静止物体和噪声的片段进行过滤，并采用DCF（discriminative correlation filter）（Wang等，2017a）算法基于1 Hz的

**表2　单目标跟踪代表性数据集及其属性**

**Table 2　Representative datasets and their attributes of single object tracking task**

| 数据集 | 特点 | 视频 | 帧数 | 挑战因素 | 目标消失 | 镜头切换 | 标签密度/Hz | 目标类别 |
|---|---|---|---|---|---|---|---|---|
| OTB50（Wu等，2013） | 通用 | 51 | 29 k | 11 | N | N | 30 | 10 |
| ALOV++（Smeulders等，2014） | 通用 | 314 | 89 k | 14 | N | N | 30 | 64 |
| OTB100（Wu等，2015） | 通用 | 100 | $59\times10^3$ | 11 | N | N | 30 | 16 |
| Tcolor-128（Liang等，2015） | 通用 | 128 | $55\times10^3$ | 11 | N | N | 30 | 27 |
| UAV123（Mueller等，2016） | 无人机 | 123 | $110\times10^3$ | 12 | N | N | 30 | 9 |
| UAV20L（Mueller等，2016） | 无人机 | 20 | $58.7\times10^3$ | 12 | N | N | 30 | 5 |
| NUS-PRO（Li等，2016） | 行人/刚体 | 365 | $135\times10^3$ | 12 | N | N | 30 | 17 |
| DTB（Li和Yeung，2017） | 无人机 | 70 | $15\times10^3$ | 11 | N | N | 30 | 15 |
| TrackingNet（Müller等，2018） | 通用 | 30k | $14.43\times10^6$ | 15 | N | N | 1（30） | 27 |
| OxUvA（Valmadre等，2018） | 通用 | 366 | $1.55\times10^6$ | 6 | Y | N | 1 | 22 |
| UAVDT（Yu等，2020） | 无人机 | 50 | $80\times10^3$ | 9 | N | N | 30 | 3 |
| GOT-10k（Huang等，2021） | 通用 | 10 k | $1.45\times10^6$ | 6 | Y | N | 10 | 563 |
| LaSOT（Fan等，2021a） | 通用 | 1.55 k | $3.87\times10^6$ | 14 | Y | N | 30 | 70 |
| VideoCube（Hu等，2023） | 通用 | 500 | $7.46\times10^6$ | 12 | Y | Y | 10（30） | 89 |

注：TrackingNet采用人工标注和自动标注相结合的方式，首先以1 Hz的频率进行人工标注，并使用DCF算法（Wang等，2017a）标注剩余帧；OxUvA仅在每秒提供一次人工标注；GOT-10k以10 Hz的采样率从超过40 h的视频中提取145万幅图像，并为每一帧提供人工标注；VideoCube将人工标注和自动标注相结合，其中人工标注频率为10 Hz，并使用PrDiMP算法（Danelljan等，2020）来实现30 Hz的密集标注；Y代表有，N代表无。

人工标注生成中间帧目标矩形框，从而采用人工与算法相结合的方式将标注频率提升到30 Hz，旨在为基于深度学习的跟踪算法提供大规模数据支撑。虽然TrackingNet数据集是目前规模最大、包含序列最多的单目标跟踪数据集，但是其涉及的目标类别有限，且训练集和测试集中的数据分布接近，难以度量跟踪算法对于未知目标类别的泛化能力。

针对上述问题，GOT-10k数据集（http://got-10k.aitestunion.com）（Huang等，2021）从泛化能力评估的角度出发，针对单目标跟踪任务类别无关的特点，构建了一个广泛覆盖563类运动物体、87种运动模式、总计10 000段数据的大型评测环境，并提供150万帧高精度人工标注。GOT-10k采用单词语义关系库WordNet（Miller，1995）来指导物体类别和运动形式的采集，以确保类别选择的全面性，同时规避人为因素带来的偏差。通过训练集—测试集实例类别不重合的开集测试规范，GOT-10k实现对模型在未知运动物体和运动形式上泛化能力的评估，并通过在线评估网站实时更新跟踪算法在测试集上的得分及排名；这一开集测试思路也被后续工作借鉴（Fan等，2021a；Hu等，2023），成为构建大规模单目标跟踪数据集的通用准则。

2.1.2　长时跟踪数据集

无论是以OTB基准为代表的经典目标跟踪数据集还是以GOT-10k和TrackingNet为代表的大规模数据集，均针对单目标跟踪中的短时跟踪任务进行设计（平均时长为10～30 s，且一般假设目标始终出现在画面中）。这种隐含的约束限制了任务的应用场景，因此研究者提出长时跟踪数据集，旨在通过更长的视频序列和更丰富的挑战因素来挖掘任务难点，并为长时跟踪算法提供训练和评估的环境。

OxUvA数据集（https://oxuva.github.io/long-term-tracking-benchmark/）包含366段平均时长为4 320帧的视频，并提供目标消失的标签。然而，其仅提供1 Hz的稀疏人工标注，且未对中间帧进行算法补全。

针对已有目标跟踪数据集规模偏小、序列较短、缺乏高质量密集标注等问题，研究者于2019年发布了LaSOT数据集（https://cis.temple.edu/lasot/），其覆盖70类常见运动目标，总计包含1 400段平均时长为2 502帧的序列。2021年，该数据集引入开集测试规范，并将数据集扩展到1 550段视频和85类运动目标，总计包含387万帧高精度人工标注。如图5（b）所示，LaSOT为每段视频序列提供视觉语义标注，为多模态研究提供支持。

2.1.3　全局实例跟踪数据集

如第1.3节所述，全局实例跟踪进一步拓展视觉目标跟踪的任务边界，并通过取消短时跟踪和长时跟踪任务定义中隐含的连续运动假设，实现对人类动态视觉能力进行建模。作为全局实例跟踪任务的评测环境，Hu等人（2023）构建的VideoCube数据集从传统的单镜头、单场景视频拓展到包含镜头切换和场景转换的视频。镜头和场景作为介于帧和视频之间的两个粒度，可以通过组合的方式丰富视频内容，从而一方面提升单段视频所包含的跟踪挑战因素，另一方面也适配于涉及镜头切换的应用场景。

为了给全局实例跟踪任务提供一个高质量的评测环境，VideoCube数据集通过6D准则对视频叙事内容进行精确刻画，并从多维度向真实世界靠拢。研究者首先将视频叙事内容的定义（发生在时空中的一连串因果关系事件）拆解为6个维度，如图5（c）所示。在此基础上，VideoCube数据集将场景类别、空间连续性、时间连续性、视频时长、运动模式和目标类别作为采集维度，构建一个平均序列长度接近1.5万帧、总计746万帧的大规模单目标跟踪评测环境。此外，区别于其他单目标跟踪数据集的粗粒度挑战因素标注（通常为序列级标注），VideoCube为12种挑战因素提供逐帧的细粒度标注，旨在为复杂场景下算法性能的评估提供高质量的评测环境。

## 2.2　专用评测环境

区别于覆盖多种目标类别和应用场景的通用评测环境，专用评测环境针对特定目标进行设计。

2.2.1　特定对象

行人和刚体是两类常见的运动目标类别。Li等人（2016）构建的NUS-PRO（NUS/BUAA people and rigid objects dataset）数据集包含了365段由移动相机拍摄的视频序列，并为序列中的运动目标提供细粒度的遮挡标注（未遮挡、局部遮挡、完全遮挡），旨在提升算法对行人和刚体的感知能力。

作为一类广泛存在于实际应用场景中的运动目标，透明物体通常被单目标跟踪数据集所忽略。与非透明物体不同，透明物体的表观信息较少，且易受背景信息的影响，因此更具有跟踪挑战性。针对上

述问题，TOTB（transparent object tracking benchmark）数据集（https://hengfan2010.github.io/projects/TOTB/）（Fan等，2021b）采集了225段包含15种常见透明物体的视频序列，并对其进行如图5（d）所示的高精度标注，旨在评估跟踪算法在弱表观信息情况下的运动目标感知能力。

2.2.2 特定场景

与通用场景相比，无人机场景中的目标跟踪任务更具挑战性。其中，目标尺寸较小、运动速度较快、运动模糊情况严重等问题广泛存在于无人机视觉系统中，因而对该场景下跟踪方法提出更高的鲁棒性要求。针对这一场景，Mueller等人（2016）发布UAV123（unmanned aerial vehicle 123）和UAV20L（https://cemse.kaust.edu.sa/ivul/uav123）数据集。如图5（e）所示，该数据集包含大量低分辨率的运动目标，且拍摄视角主要集中在俯视视角，与通用场景存在较大差异。值得注意的是，UAV123和UAV20L中包含部分模拟器生成的虚拟数据。针对无人机场景下真实数据匮乏的问题，研究者发布了DTB（drone tracking benchmark）（https://github. com/flyers/drone-tracking）（Li和Yeung，2017）和UAVDT数据集（https://sites.google. com/site/daviddo0323/projects/uavdt）（Yu等，2020），旨在为无人机视角下的视觉任务研究提供更全面真实的数据支撑。

### 2.3 竞赛评测环境

与基于公开数据集的评测环境不同，部分评测环境以竞赛的形式进行发布。在目标跟踪领域，自2013年起每年固定举行的VOT挑战赛（https://votchallenge.net/index.html）是目前最具影响力的比赛，表3列举了VOT挑战赛的详细信息。

**表3 VOT挑战赛及其子赛事**

**Table 3 VOT challenge and its sub-competitions**

| 年份 | 短时（VOT-ST） | 实时（VOT-RT） | 热成像（VOT-TIR） | 颜色+热成像（VOT-RGBT） | 长时（VOT-LT） | 颜色+深度（VOT-RGBD） | 深度（VOT-D） |
|---|---|---|---|---|---|---|---|
| 2013 | 16（矩形框） | N | N | N | N | N | N |
| 2014 | 25（矩形框） | N | N | N | N | N | N |
| 2015 | 60（矩形框） | N | 20（矩形框） | N | N | N | N |
| 2016 | 60（矩形框） | N | 25（矩形框） | N | N | N | N |
| 2017 | 60（矩形框） | Y | 25（矩形框） | N | N | N | N |
| 2018 | 60（矩形框） | Y | N | N | 35（矩形框） | N | N |
| 2019 | 60（矩形框） | Y | N | 60（矩形框） | 50（矩形框） | 80（矩形框） | N |
| 2020 | 60（分割掩膜） | Y | N | 60（矩形框） | 50（矩形框） | 80（矩形框） | N |
| 2021 | 60（分割掩膜） | Y | N | N | 50（矩形框） | 80（矩形框） | N |
| 2022 | 62（矩形框）/62（分割掩膜） | Y | N | N | 50（矩形框） | 127（矩形框） | 127（矩形框） |

注：数字代表该项赛事的视频序列数量；括号的内容代表该赛事使用的标注方式（旋转矩形框或分割掩膜）；Y代表有，N代表无。

VOT-ST（VOT in short-term）短时跟踪竞赛是一项经典的子赛事，区别于其他数据集使用的水平矩形框，VOT-ST竞赛采用旋转矩形框或分割掩膜为目标位置提供高精度标注（图6（a）（b）），为联合目标分割和目标跟踪任务的研究提供支撑。自2017年开始举办的VOT-RT（VOT in real-time）实时跟踪竞赛可以视做短时跟踪竞赛的拓展，该赛事需要算法在保持短时跟踪鲁棒性的基础上提升运行效率，以满足实时性的要求。2018年起新增的VOT-LT（VOT in long-term）长时跟踪竞赛为长时跟踪任务研究提供了良好的支持，竞赛举办方将“允许目标消失”作为区分短时跟踪和长时跟踪的判别标准，并采集50段长视频序列作为竞赛数据。上述3项子竞赛均基于彩色视频序列，并分别从鲁棒性、实时性和长时跟踪稳定性的角度考察算法。

此外，VOT挑战赛也关注多模态任务。热成像相机和深度相机可以提供额外的环境信息，从而将目标跟踪任务从传统场景拓展至视觉信息较弱的场景。VOT-TIR（VOT with thermal infrared）和VOT-RGBT（VOT with RGB and thermal infrared）基于热成

像图像开展目标跟踪竞赛。热成像信息受光照影响较小，因此在特殊光照下依旧可以为模型提供环境信息，帮助算法根据前景—背景的热成像差异开展目标跟踪（图6(c)）。VOT-D（VOT with depth）和VOT-RGBD（VOT with RGB and depth）则关注深度信息，其可以有效分离前景和背景，同时为目标遮挡问题提供额外的信息支撑（图6(d)）。这些多模态跟踪竞赛不仅为跟踪中的难点问题提供额外的信息源，而且为多模态跟踪算法提供良好的评测环境。


(a) VOT-ST短时跟踪竞赛(基于旋转矩形框)

(b) VOT-ST短时跟踪竞赛(基于分割掩膜)

(c) VOT-RGBT多模态跟踪竞赛(基于热成像图像序列)

(d) VOT-RGBD多模态跟踪竞赛(基于深度图像序列)

图6　单目标跟踪代表性竞赛

Fig. 6　Representative competitions for single object tracking ((a) VOT-ST short-term tracking competition based on rotating bounding-box; (b) VOT-ST short-term tracking competition based on mask; (c) VOT-RGBT multimodal tracking competition based on thermal infrared information; (d) VOT-RGBD multimodal tracking competition based on depth information)

## 3　待测对象

本节从单目标跟踪算法和人类动态视觉跟踪能力两方面对任务的待测对象进行介绍。本节首先以算法的发展历程为切入点，对各个阶段的代表性跟踪算法进行回顾，不仅包含对模型细节的介绍，也阐述了算法为应对评测环境中隐藏的挑战因素而进行模块设计的思路。此外，围绕“智能评估”这一核心目标，本节对认知科学领域与人类动态视觉能力度量有关的典型实验进行回顾，旨在帮助研究者更好地理解算法的建模对象，同时也为基于“人机对抗”思想的视觉图灵评估提供研究基础。

### 3.1　算法

#### 3.1.1　传统跟踪方法

图7所示的传统目标跟踪方法通常包含运动模型、特征表达、表观模型和算法更新这4个步骤。

1）运动模型。单目标跟踪具有时序决策的特点，即模型需要通过对目标的状态进行估计，从而在后续帧中进行目标轨迹预测。预测后续帧中目标位置或对其状态分布进行估计即为单目标跟踪中的运动建模，并以粒子滤波（particle filtering）（Ross等，2008；Huang和Ma，2015）和滑动窗口（sliding window）（Henriques等，2012；Hare等，2016）为代表。

2）特征表达。早期单目标跟踪通常将目标或候选样本作为一个整体进行全局特征提取，并基于此构建表观模型。以灰度特征（Henriques等，2012）、梯度直方图特征（Henriques等，2015）、颜色直方图特征（He等，2013，2017）、哈尔（Haar-like）特征（Hare等，2016）等为代表的方法对运动目标进行整体特征提取和表观建模，但缺乏对背景噪声、遮挡、形变等挑战因素的处理。为了在出现上述影响因素时仍可

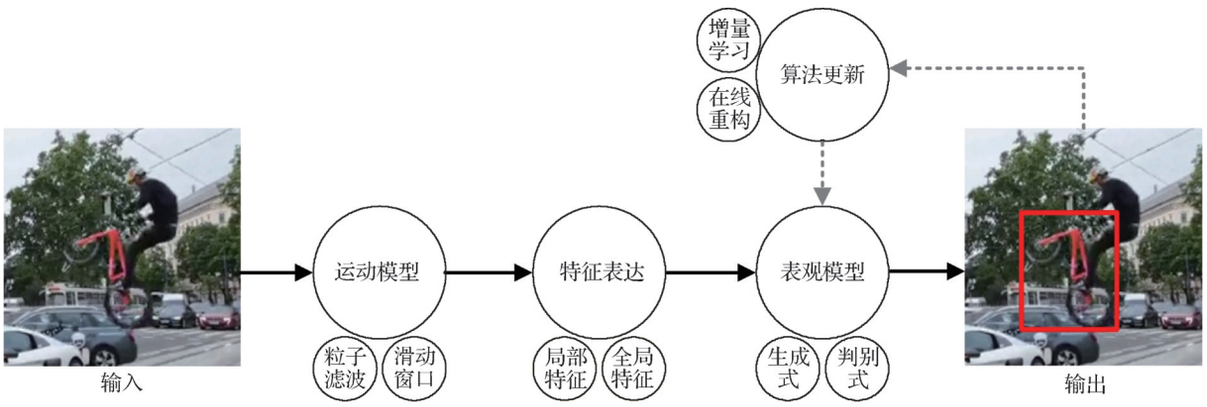



图 7　传统单目标跟踪算法执行流程

Fig. 7　Execution flow of traditional SOT methods

以借助未受干扰的区域特征进行目标跟踪，基于局部特征表达的方法相继提出（Zhang 等，2013）。此类方法将目标或候选样本划分为多个独立区域，通过部件之间的信息融合来实现在遮挡、形变等干扰因素下的目标跟踪。

3）表观建模。根据表观模型的不同，视觉物体跟踪可以划分为生成式方法和判别式方法。生成式方法使用增量子空间（Bao 等，2012；Hare 等，2016）、分块稀疏表达（Henriques 等，2015）等方式维护目标的模板集，并以候选样本与目标模板集的距离进行相似度的衡量；而判别式方法则将跟踪视为分类问题（Henriques 等，2012；Ma 等，2015b），通过更新目标与背景之间的分类器来对运动目标进行建模，并通过分类得分确定目标在当前帧的位置。其中，基于相关滤波的方法（Bolme 等，2010；Ma 等，2015b；Nam 和 Han，2016；Danelljan 等，2017）兼具性能和速度的优势，因而备受关注。

4）算法更新。单目标跟踪任务仅在首帧给定目标的表观信息，但在跟踪过程中表观形态通常会发生改变，所以需要一个合适的更新策略以确保算法在时间序列中可以适应目标的形态变化并实现持续稳定的跟踪。传统跟踪方法通常采用基于增量学习的模型更新（Ross 等，2008；Wang 等，2016）或者在线重构（online refactoring）表观模型（Bao 等，2012；黄凯奇等，2015）。

### 3.1.2　基于深度特征的相关滤波方法

图 8 展示了相关滤波方法的执行流程。相关滤波理论通过循环平移操作实现对训练样本的扩充操作，在早期缺乏大规模数据集的情况下有效弥补了训练数据不足的缺陷；此外，快速傅里叶变换实现了在频域内进行卷积计算，降低了模型的计算量，并显著提升跟踪效率。2010 年，MOSSE（minimum output sum square error）算法（Bolme 等，2010）首次在单目标跟踪领域引入相关滤波，在仅利用灰度特征的情况下实现与其他算法相近的精度，并在单 CPU 的环境中实现 600 帧/s 的速度，引领后续工作开展基于相关滤波的跟踪方法研究。

区别于主要利用手工特征的早期相关滤波方法，随着深度学习技术的发展，部分相关滤波方法开始利用深度特征提升特征表达的能力。其中，VGG（Visual Geometry Group）网络（Simonyan 和 Zisserman，2015）具有结构简洁、性能良好等优点，因此广泛应用在多项视觉任务中。DeepSRDCF（deep spatially regularized discriminative correlation filters）算法（Danelljan 等，2015）和 HCF（hierarchical convolutional features）算法（Ma 等，2015a）均尝试利用 VGG 网络来有效提升算法的跟踪精度。

以 ECO（efficient convolutional operators）（Danelljan 等，2017）为代表的后续工作尝试将更复杂的深度模型与相关滤波方法相结合。然而，研究者发现单纯增加网络深度无法带来更大的收益。针对上述问题，UPDT（unveiling the power of deep tracking）算法（Bhat 等，2018）提出利用深度特征去建模高层语义信息、利用浅层特征去建模纹理信息和颜色信息，最后通过特征融合策略来融合二者的响应图，从而得到最优的结果。

### 3.1.3　基于孪生神经网络的跟踪方法

与利用深度特征替代手工特征的相关滤波方法不同，孪生神经网络算法将单目标跟踪视做局部实例检索任务，能有效规避采样和在线学习带来的计

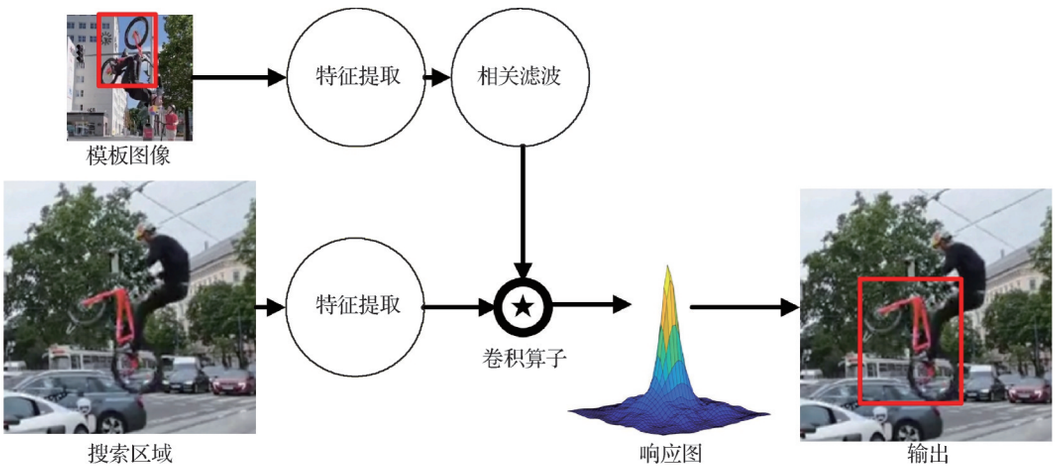



图8　基于相关滤波的单目标跟踪算法执行流程

Fig. 8　Execution flow of SOT methods based on correlation filter

算开销，并成为深度学习时代的主流模型（图9）。

在单目标跟踪的若干步骤中，在线更新过程需要在整个序列决策过程中频繁执行，并且需要在样本池中进行多次迭代来优化表观模型，因此对速度的影响程度最大。对跟踪算法进行更新的核心是确保模型可以在目标表观特征发生改变的情况下与候选样本进行匹配，因此可以将跟踪问题转化为匹配问题。孪生神经网络是针对笔迹匹配问题设计的模型（Bromley 等，1993），将其引入到目标跟踪中并通过在大规模数据集上的训练，可以使模型具备在目标表观变化下的样本匹配能力，从而在舍弃在线更新模块的情况下，依旧可以进行稳定跟踪。

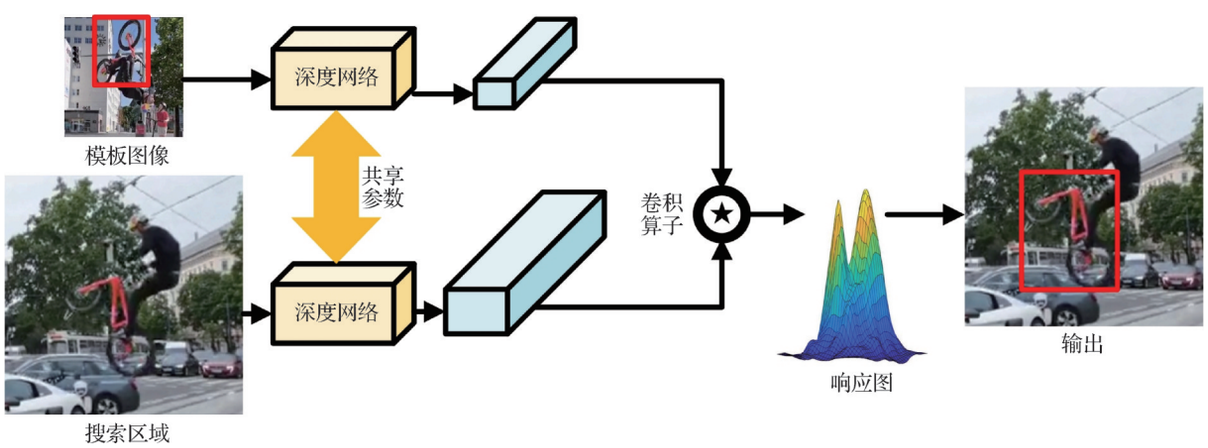



图9　基于孪生神经网络的单目标跟踪算法执行流程

Fig. 9　Execution flow of SOT methods based on siamese neural network

以 SiamFC（siamese full convolution）（Bertinetto 等，2016）为代表的一系列孪生神经网络算法真正将深度学习和目标跟踪领域相结合，并取得了显著效果。如图10（a）所示，SiamFC 将首帧包含目标的区域经过预处理得到模板图像 $z$；其假设目标位置在连续帧中具有相关性，并以上一帧目标所在位置为中心设置当前帧的搜索区域 $x$。SiamFC 的网络结构非常简洁，两个完全一致的全卷积网络 $\varphi$ 负责特征提取，并通过互相关计算将两个分支的特征提取结果生成一张概率分布图，图中概率最大的位置即为当前帧目标的位置。实验表明，SiamFC 在相关数据集上优于传统方法，并保持较高的算法运行速度，为后续基于深度学习的目标跟踪算法研究提供一种可行的思路。然而，SiamFC 仍存在若干问题尚未解决，例如其无法适应目标在运动过程中的形状变化、因缺乏背景信息而存在判别性不足，以及网络层数较浅。针对上述问题，后续工作对 SiamFC 进行了优化，并取得良好效果。

1)应对目标形变问题。研究者提出 SiamRPN(siamese region proposal network)(Li 等,2018)网络,实现对目标矩形框的精确回归。SiamRPN 将跟踪问题抽象成单样本检测问题,并借鉴检测算法的设计思路,利用首帧目标信息来初始化一个局部检测器。如图 10(b)所示,SiamRPN 首先利用孪生神经网络结构,通过一个全卷积网络提取高层特征,并借鉴检测算法 Fast R-CNN(Girshick,2015)的区域候选网络(region proposal network,RPN),实现对目标位置更精确的预测和对矩形框的高精度拟合。

2)增强判别能力。研究者对 SiamRPN 进一步优化,并提出 DaSiamRPN(distractor-aware siamese region proposal network)(Zhu 等,2018)算法。如图 10(c)所示,DaSiamRPN 在目标检测数据集基础上通过数据增广的方式生成图像对,从而扩充训练数据中的正样本。此外,DaSiamRPN 借助检测数据集中来自不同类别或者同一类别不同实例的样本构建难例负样本,从而增强判别能力。针对响应图中得分较高的干扰物对跟踪结果的影响,DaSiamRPN 使用干扰物感知模型,筛选出响应值大于特定阈值的候选框供网络学习。针对孪生网络难以应对长时跟踪中目标消失的问题,DaSiamRPN 提出在目标消失时扩大搜索范围,从而实现稳健的长时跟踪。

3)引入更深的主干网络。SiamRPN++(Li 等,2019)针对孪生神经网络主干网络层数较浅的问题进行改进。如图 10(d)所示,SiamRPN++对已有孪生神经网络跟踪器的采样策略进行优化,让目标在中心点附近进行偏移,缓解深度网络因为破坏了严格平移不变性带来的影响。与 SiamRPN++同期的 SiamDW(deeper and wider siamese network)(Zhang 和 Peng,2019)算法同样关注如何将更深和更宽的卷积神经网络引入孪生神经网络跟踪框架中。SiamDW 借鉴 ResNet(residual network)(He 等,2016)的思想,通过设计内部裁剪残差模块以消除引入填充操作带来的负面影响,并将该模块和基础算法相结合,从而提升速度和精确度。

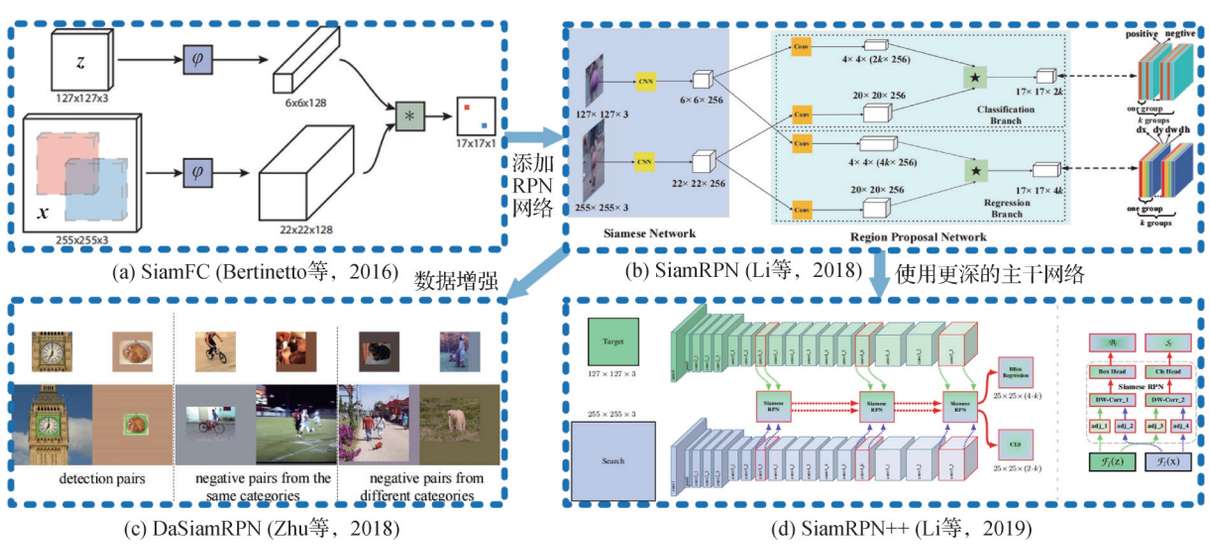



图 10　基于孪生神经网络的代表性跟踪方法框架

Fig. 10　Representative tracking methods based on siamese neural network ((a) SiamFC (Bertinetto et al., 2016); (b) SiamRPN (Li et al., 2018); (c) DaSiamRPN (Zhu et al., 2018); (d) SiamRPN++ (Li et al., 2019))

4)无锚点的跟踪方法。以 SiamRPN 为代表的系列算法均基于锚点(anchor-based)进行目标定位,并严重依赖于锚点的数目、尺寸和形状等参数。受单阶段目标检测算法(Tian 等,2019)的影响,部分算法从无锚点(anchor-free)的角度进行优化。Ocean 算法(Zhang 等,2020)采用基于特征对齐的无锚点网络,以适应跟踪过程中目标形状和尺度的变化问题。以 SiamFC++(Xu 等,2020)和 SiamCAR(siamese fully convolutional classification and regression)(Guo 等,2020)为代表的方法则通过增加一个平行于分类分支的模块来去除对锚点先验的依赖。

5)基于重检测机制的跟踪方法。针对单目标跟踪过程中目标消失的问题,SiamRCNN(siamese region-based convolutional neural network)算法

(Voigtlaender等,2020)通过重检测(re-detection)机制对重新出现在画面中的目标进行精确定位,从而提升长时跟踪的鲁棒性。此外,SiamRCNN借鉴PReMVOS模型(Luiten等,2019)提出的Box2Seg网络生成分割掩膜,从而适配于视频目标分割任务,并在YouTube-VOS(YouTube video object segmentation)(Xu等,2018)基准中取得良好的分割效果。

3.1.4 其他代表性深度学习目标跟踪方法

除相关滤波和孪生神经网络结构之外,也有大量的深度学习方法针对跟踪任务中的特定问题对网络结构进行优化,并提出新的跟踪算法框架。

1)结合相关滤波和孪生神经网络的跟踪框架。ATOM(accurate tracking by overlap maximization)(Danelljan等,2019)设计了一个由离线训练的预测模块和在线训练的分类模块组成的两阶段跟踪框架。该跟踪框架综合了深度学习在大规模数据集上离线训练的优势和相关滤波算法在线更新的优势,取得较好的跟踪效果。DiMP(discriminative model prediction)(Bhat等,2019)算法在此基础上引入判别能力更强的损失函数,以增强模型对于鲁棒特征的学习能力;同时通过对优化器的改进,增加网络的收敛速度。PrDiMP和SuperDiMP(Danelljan等,2020)算法则利用置信度回归的方式提升DiMP对于目标回归的精确度。针对背景干扰问题,KeepTrack(Mayer等,2021)算法首先对训练数据进行重新筛选,并在SuperDiMP的基础上引入候选目标关联模块来关联不同帧之间的背景干扰物,以增加算法的鲁棒跟踪能力。

2)针对长时跟踪任务的跟踪框架。作为一个序列决策任务,长时跟踪会加剧跟踪过程中的累计误差。因此,一系列工作从不同的角度出发,针对长时跟踪任务的难点进行模型设计。GlobalTrack(Huang等,2020)算法提出了一种简洁的全局实例跟踪框架,其将首帧目标作为查询图像(query),并独立对后续每一帧进行目标搜索,以避免累积误差的影响。区别于GlobalTrack的逐帧全局检索思路,另一种长时跟踪框架源自经典的TLD(tracking-learning-detection)(Kalal等,2012)算法,其将长时目标跟踪问题分解为跟踪、学习与检测3个部分,并将局部跟踪算法和全局检测算法相结合,从而应对目标在跟踪过程中发生的形变、部分遮挡等问题。这种建模思路的关键是确定局部搜索和全局搜索的切换时机。针对这一问题,SPLT(skimming-perusal tracking)(Yan等,2019)算法设计了一个离线训练的验证器,通过当前帧的置信度分数决定下一帧的搜索策略。在此基础上,LTMU(long-term tracking with meta-updater)(Dai等,2020)算法引入元学习器来决策跟踪过程中的模型更新问题。通过将几何信息、判别信息、表观信息和时序信息作为元学习器的输入,LTMU实现对是否进行目标更新的自动判别。

3)利用场景信息的跟踪框架。针对传统跟踪算法难以适应相似物体干扰的问题,KYS(know your surroundings)(Bhat等,2020)算法将场景信息引入到跟踪框架中。其通过判定局部区域是否为目标、背景或干扰物,实现对场景信息的编码,并将编码内容在序列中进行传播,最后与目标的表观模型相结合,实现鲁棒跟踪。

4)基于注意力机制的跟踪框架。传统跟踪框架通常包含特征提取、特征融合和目标矩形框估计等多个阶段(multi-stage)。近年来,随着注意力机制(Vaswani等,2017)在诸多视觉任务上的成功应用,部分算法(Chen等,2021;Cui等,2021;Wang等,2021;Yan等,2021;Yu等,2021)将其引入到目标跟踪中,并取得良好的跟踪效果。但是,上述方法仍采用传统跟踪框架,并在卷积神经网络提取到的特征上进行注意力建模。以SwinTrack(Lin等,2022)为代表的部分方法虽然在特征提取和特征融合等阶段均采用Transformer网络,但是仍属于双流(two-stream)多阶段方法,即在特征提取阶段未对模板区域和搜索区域进行交互。因此,MixFormer(Cui等,2022)提出了一个基于注意力机制的跟踪框架,通过混合注意模块完成特征提取和目标信息合并,实现端到端的目标跟踪过程。与多阶段的传统跟踪框架相比,MixFormer充分发挥了注意力机制的优势,并在基准数据集中取得良好的跟踪效果。同期的OSTrack(one-stream track)算法(Ye等,2022)采用相似的思路,将跟踪过程建模为一个单流单阶段(one-stream, one-stage)的网络。与MixFormer相比,OSTrack在速度和精度之间达到较好的平衡。

### 3.2 人类

人类强大的视觉系统是计算机视觉任务的建模目标。因此,将人类作为视觉待测对象并研究其执行计算机视觉任务时的表现,可以为算法建模提供指导。然而,度量人类视觉能力的相关工作主要集

中在认知神经科学领域，因而通常被计算机视觉领域综述所忽略。综上，本文将对涉及人类视觉能力的相关研究进行回顾，旨在为交叉研究提供基础，并为后续基于人类动态视觉能力度量算法跟踪智能的评估评测研究提供基础。

3.2.1 视觉理论

对生物视觉机理的研究可以追溯到半个世纪之前。Hubel和Wiesel(1959，1962)将包含特定模式的幻灯片展示给猫，并记录猫脑神经元在不同模式下的电活动，从而探究生物的视觉信息处理机制。随后，Treisman和Gelade(1980)以包含不同颜色和形状的字符为实验载体分析人类视觉加工能力(图11(a))，并提出特征整合理论。1982年，Marr将生理学、认知神经科学、信息处理等多领域的研究内容进行综合，从信息科学的角度将视觉定义为对于外部图像的有效符号描述，并形成视觉计算理论(Marr，2010)。Chen(1982)基于正方形、圆形等基础几何图形的组合图像，对人类视觉感知系统中的拓扑结构开展研究。Biederman(1987)以包含部件可拆解的日常用品图像作为实验测试数据(图11(b))，通过分析被试者在观测不同图像时的表现，分析人类执行目标识别任务时的能力，并提出成分识别理论。以上述工作为代表的理论研究为计算机视觉发展提供了理论基础，并启发研究者借鉴生物视觉特征开展算法设计(Sudderth等，2005；Li等，2006；Lazebnik等，2006)。

3.2.2 视觉能力

神经科学家将人类视觉能力划分为静态视觉能力和动态视觉能力两大类(Miller和Ludvigh，1962)。作为感知系统的重要组成部分，静态视觉能力和动态视觉能力与人类日常生活关联密切，并吸引诸多神经科学家开展研究，旨在通过观测被试者在执行不同视觉任务时的表现来评估其视觉能力(Land和McLeod，2000；Yu等，2014)。研究发现，动态视觉能力优异的被试者大部分具备良好的静态视觉能力，但静态视觉能力优秀的被试者可能存在运动物体感知缺陷(Miller，1958；Long和Penn，1987)。鉴于良好的静态视觉能力是动态视觉能力的基础，因此本节将对与两种基础视觉能力相关的代表性工作进行综述。

对于静态视觉能力，代表性视觉理论研究工作已经依托目标识别任务从特征提取、信息处理和成分识别等角度开展研究，并形成理论体系(Treisman和Gelade，1980；Chen，1982；Biederman，1987)。此外，临床领域的研究者通常依托4个阈值来设计静态视觉能力度量实验(Ginsburg等，1984)：最小可检测阈值、最小分辨率阈值、最小可感知的对准阈值和最小识别阈值。其中，最小识别阈值通常称为视力，基于此设计的标准视力表通常要求被试者在固定距离(通常为5 m)观测特定字符(通常为大写英文字母)的朝向(如图11(c)所示)，目前已经成为一种广泛应用的静态视觉能力度量范式。

对于动态视觉能力，研究者最初沿用静态视觉能力的度量思路，通过让被试者观测高对比度背景下的运动物体来记录其运动目标感知能力(Kirshner等，1967；Pylyshyn和Storm，1988)。如图11(d)所示，转盘是一种常用的测试仪器，被试者的动态视觉能力取决于其对于旋转字符的感知(Quevedo等，2018)。然而，与广泛应用的基于视力表的静态视觉能力度量范式不同，部分研究者认为基于转盘度量动态视觉能力存在较大局限性(Erickson等，2011)。日常生活中的运动目标感知任务涉及不同的表观信息、背景信息和运动信息，而转盘仅包含白底黑字这一高对比度的前景—背景组合，且仅涉及旋转轨迹，与实际应用场景相距甚远。针对上述问题，Quevedo等人(2012)利用DynVA软件生成不同刺激(stimuli)、背景颜色和运动轨迹的组合，并综合被试者在观看不同组合时的表现作为其动态视觉能力(图11(e))。与传统方式相比，DynVA的评测环境丰富程度已经有较大提升，但仍与真实世界的目标跟踪挑战因素存在较大差异。Hu等人(2023)以全局实例跟踪任务为基础开展人类动态视觉能力的度量实验。研究者从VideoCube数据集中挑选出10段包含不同跟踪难度、视频时长、目标类别、场景类别和运动模式的视频，以3种速度播放给15位被试者，并通过眼动仪记录被试者的动态视觉能力，成为计算机视觉领域对人类动态视觉能力度量的开创性工作(图11(f))。

综上，对于算法性能和人类能力的研究相辅相成、互为引导。一方面，人类动态视觉能力是目标跟踪算法的建模目标，对动态视觉机理的深度研究可以为算法设计提供更科学的理论支持，并为跟踪智能发展指明方向；另一方面，为算法研究所设计的高质量数据集也为人类动态视觉能力研究提供更丰富

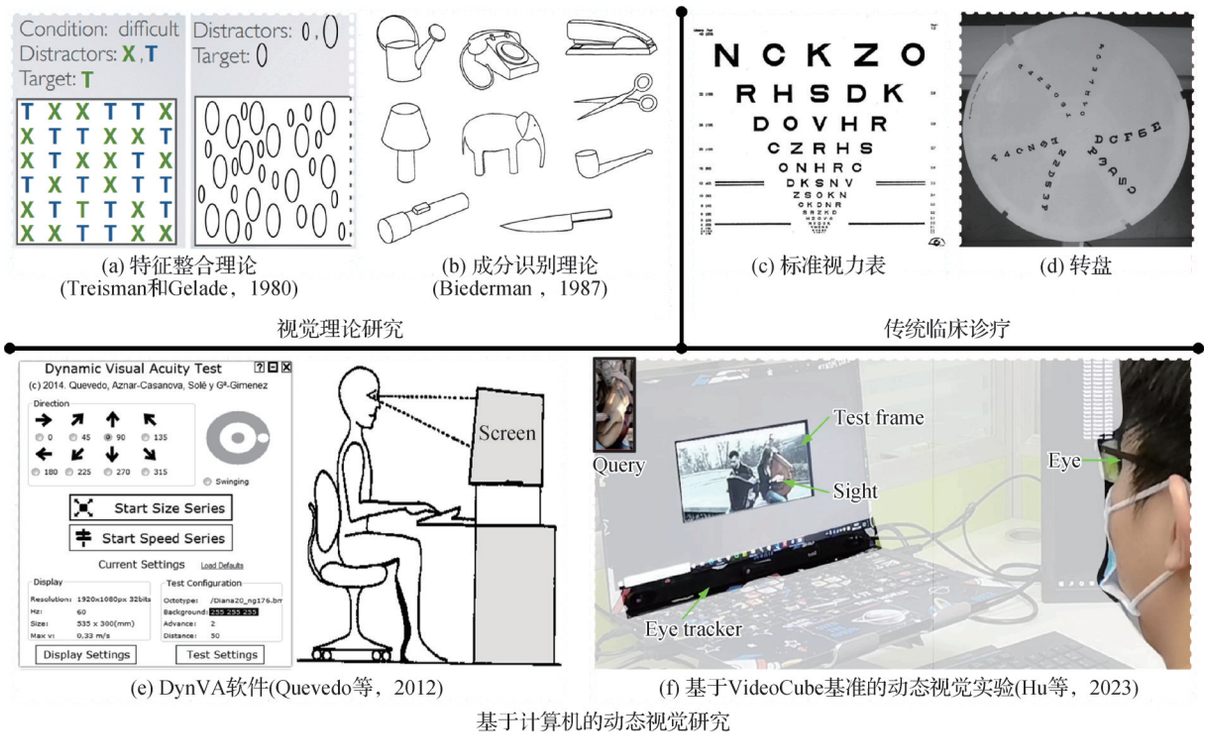



图11　度量人类视觉能力的代表性工作

Fig. 11　Representative works on measuring human visual ability ((a) feature integration theory (Treisman and Gelade, 1980); (b) recognition by component theory (Biederman, 1987); (c) standard optotype; (d) rotator; (e) DynVA software (Quevedo et al., 2012); (f) dynamic visual ability experiment based on VideoCube benchmark (Hu et al., 2023))

的评测环境和更复杂的挑战因素，帮助研究者对人类视觉机理开展细粒度的分析。

## 4　评估方式设计

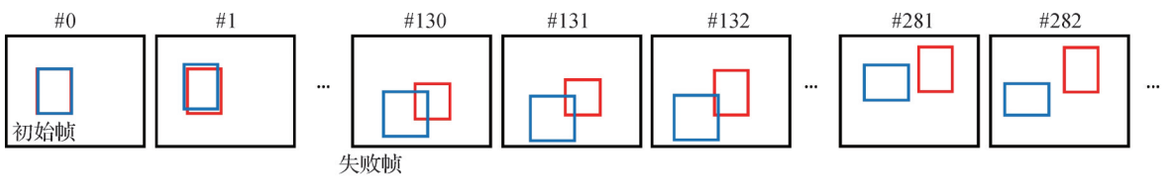
在充分理解评测任务、评测环境和待测对象的基础上，本节从“机机对抗”和“人机对抗”的视角出发，对传统的评估方式和基于视觉图灵思想的评估方式进行介绍。其中，以“机机对抗”为核心的传统评估方式通常专注于测试算法的性能，因此一个良好的算法只需要在评测环境中战胜其他模型（向下对比），这会导致研究者将重点放在模型设计和参数优化中，从而忽略了单目标跟踪任务的建模目标——人类的动态视觉跟踪能力。区别于传统的评测思路，以“人机对抗”为核心的视觉图灵评估则将人类实验者引入到智能评估的回路中，因此算法只有实现近似或者超越人类的视觉跟踪能力，才可以被认为具有智能性（向上对比）。综上，研究者需要在深刻理解两种评测机制特点的基础上，针对评测目标构建完整的智能评估系统，从而实现对待测对象的全面分析。

### 4.1　以“机机对抗”为核心的传统评估方式

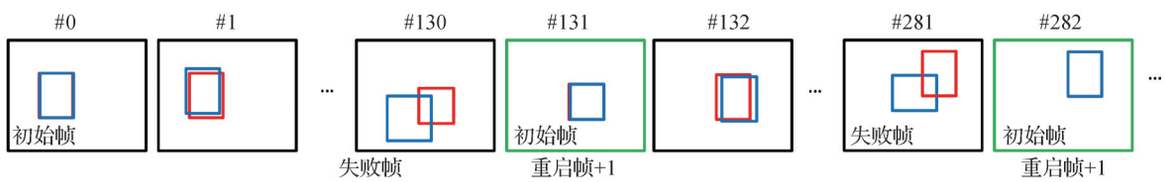
#### 4.1.1　评测机制

评测机制旨在通过合理的实验方式获得原始结果，并为后续的实验分析提供支持。作为规范化单目标跟踪基准的开创性工作，OTB基准（Wu等，2013，2015）共设计了5个评测机制。其中，经典的OPE（one-pass evaluation）机制旨在利用首帧中目标的位置对算法进行初始化，并要求算法在后续帧中以矩形框的形式输出每一帧的目标跟踪结果（如图12（a）所示）。然而，OPE机制包含两个缺点：1）跟踪器受首帧初始化的影响，不同的初始化位置可能会造成较大的结果差异；2）目标跟踪是一个序列决策过程，算法在跟踪失败后通常会在后续帧持续丢失目标，在没有重新初始化机制的情况下，跟踪失败后的结果通常无法提供有意义的评估信息。

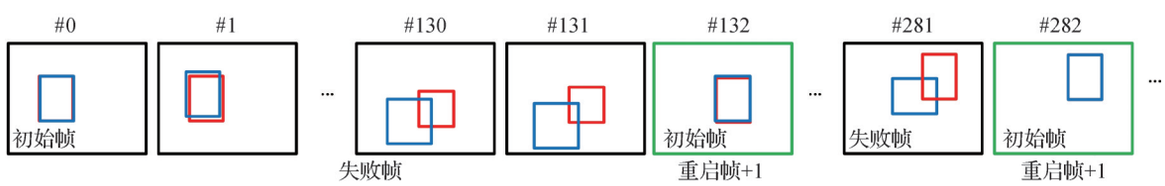
针对缺点1），OTB基准从时间和空间的角度打乱算法的初始化位置，并提出TRE（temporal robustness evaluation）机制和SRE（spatial robustness evaluation）机制。针对缺点2），研究者认为可以通过设




(a) OTB基准提出的OPE机制(Wu等，2015)




(b) OTB基准提出的OPER机制(Wu等，2015)




(c) VideoCube基准提出的R-OPE机制(Hu等，2023)

图 12　单目标跟踪任务评估机制

Fig. 12　Evaluation mechanisms for SOT task ((a) OPE mechanism proposed by OTB benchmark (Wu et al., 2015); (b) OPER mechanism proposed by OTB benchmark (Wu et al., 2015); (c) R-OPE mechanism proposed by VideoCube benchmark (Hu et al., 2023))

计重启机制监控算法的跟踪效果，当发现算法跟踪失败后，用下一帧目标信息对算法进行重新初始化。因此，OTB 基准提出允许算法失败后重启的 OPER（OPE with restart）机制和 SRER（SRE with restart）机制。在 OTB 的基础上，一部分工作仅沿用 OPE 机制进行算法评测，并未包含重启（Huang 等，2021；Fan 等，2021a）。另一部分研究者则认为重启机制对于目标跟踪的评估评测是必要的，并针对重启机制的设计开展研究。

重启帧的选择是重启机制设计的重点。如图 12(b)所示，OTB 基准提出的 OPER 重启机制通常在检测到算法跟踪失败（failure）后，在失败帧的下一帧重新初始化算法，并在最后的评测结果计算中使用全部视频帧参与计算。但 VOT 挑战赛提出的重启机制与 OPER 存在两点差异性（Kristan 等，2016）：1）VOT 认为目标遮挡是导致算法失败的主要原因，失败帧的后续帧有较大概率依旧包含遮挡，因此重启帧与失败帧需要存在一定序列间隔 $N_{\text{skip}}$。鉴于目标遮挡在 VOT 中通常不超过 5 帧，因此 VOT 将 $N_{\text{skip}}$ 设置为 5。2）VOT 认为重启后的算法会在一段时间内获得较高的跟踪得分，因此需要将重启后的部分帧设置为老化期（burn-in period）$N_{\text{burinin}}$，在计算评测结果时去除这部分视频帧，以减少重启对于算法评估的影响。根据视频序列的长度和帧率，VOT 挑战赛将 $N_{\text{burinin}}$ 设置为 10 帧。

针对全局实例跟踪任务设计的 VideoCube 基准提出了一种新的重启机制 R-OPE（restart-based OPE）（Hu 等，2023）。如图 12(c)所示，当检测到在#130 出现跟踪失败时，算法会在距离#130 最近的重启帧（#132）被重新初始化；当#281 再次出现跟踪失败时，算法将于#282 被重新初始化。与其他重启机制相比，R-OPE 机制在重启帧的选择上相对灵活，因此更适合包含丰富挑战因素的全局实例跟踪任务。

### 4.1.2　评测指标

对于视频序列 $\boldsymbol{s}_i = \{\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_t, \cdots\}$，$\boldsymbol{F}_t$ 代表其中第 $t$ 帧图像，单目标跟踪任务的评测指标通常基于当前帧目标矩形框的真值 $\boldsymbol{g}_t$ 和算法预测结果 $\boldsymbol{p}_t$ 之间的位置关系进行计算（目标缺失的视频帧（$\boldsymbol{g}_t = \varnothing$）通常不参与评测计算）。

1）精确度。精确度（precision，PRE）得分主要依赖于真值矩形框中心点 $\boldsymbol{c}_g$ 与预测结果中心点 $\boldsymbol{c}_p$ 之间的距离进行计算（Wu 等，2015），具体为

$$P(\boldsymbol{G}) = \frac{1}{|\boldsymbol{G}|}\sum_{\boldsymbol{s}_i \in \boldsymbol{G}}\frac{1}{|\boldsymbol{s}_i|}\left|\{\boldsymbol{F}_t : d_c \leqslant \theta_d\}\right| \tag{1}$$

$$d_c = \left\| \boldsymbol{c}_p - \boldsymbol{c}_g \right\|_2 \tag{2}$$

式中，$|\cdot|$代表集合的基数(cardinality)，一段视频序列$\boldsymbol{s}_t$的精确度得分定义为中心点距离小于特定阈值$\theta_d$的视频帧所占的比例，计算数据集$\boldsymbol{G}$上所有序列的精确度平均分即可作为算法在整个数据集上的最终结果。基于不同阈值$\theta_d$的统计结果可以绘制成一条精确度曲线，精确度图由不同算法的精确度曲线组成。OTB在对比不同算法的精确度时，一般以$\theta_d = 20$作为排序标准。

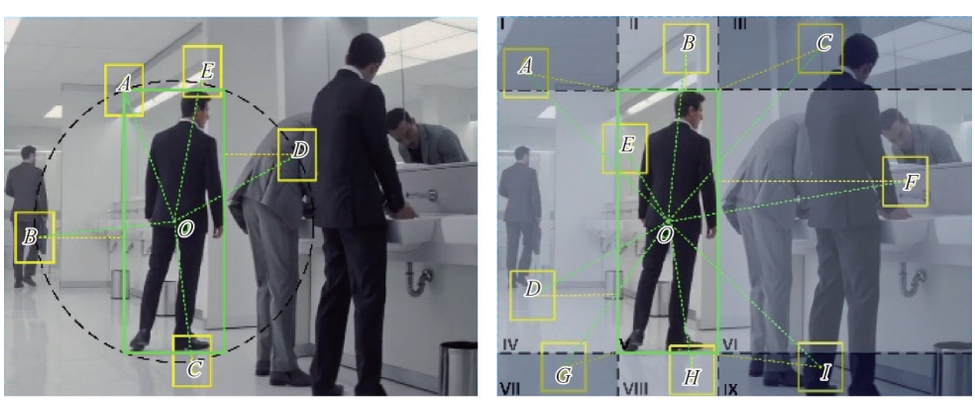
然而，仅依赖中心点距离的传统精确度指标仍存在一定局限性。如图13(a)所示，绿色矩形框表示以$O$为中心点的目标真值矩形框$\boldsymbol{g}_t$，周围的5个黄色矩形框代表5种跟踪器的结果(假设仅存在位置差异)。基于传统精确度指标的计算结果显示，$A$、$B$、$C$、$D$的精度得分相同，$E$的得分最低，这与人类对于算法跟踪效果的直观判定存在差异。分析可知，对于非正方形的目标矩形框，传统精确度计算方式会忽略目标形状和尺度对跟踪精确度的影响。

TrackingNet(Müller等，2018)和LaSOT(Fan等，2021a)采用目标尺度信息对中心点距离$d_c$进行标准化，并将标准化后的数值代入式(1)进行计算。VideoCube(Hu等，2023)则综合考虑目标尺寸、形状和图像分辨率对精确度评估的影响，提出归一化精确度(normalized precision，N-PRE)。如图13(b)所示，跟踪器的精确度得分$d_c'$由两项之和组成：第1项为中心点距离$d_c$(绿色虚线)，第2项为惩罚项$d_c^p$(黄色虚线，表示算法预测结果中心点$\boldsymbol{c}_p$与目标矩形框$\boldsymbol{g}_t$之间的最短距离)。对于算法预测结果中心点$\boldsymbol{c}_p$落入目标矩形框$\boldsymbol{g}_t$中的跟踪器，其精确度得分与传统精确度计算方式相同，即$d_c^p = 0$。随后，精确度得分$d_c'$被归一化到[0, 1]区间，0表示跟踪器中心点为目标中心点$O$，1表示当前帧$\boldsymbol{F}_t$中距离目标中心点$O$最远的点的得分，具体为

$$P'(\boldsymbol{G}) = \frac{1}{|\boldsymbol{G}|}\sum_{\boldsymbol{s}_i \in \boldsymbol{G}}\frac{1}{|\boldsymbol{s}_i|}\left|\{\boldsymbol{F}_t : N'(d_c) \leqslant \theta_d'\}\right| \tag{3}$$

$$N'(d_c) = \frac{d_c'}{\max(\{d_i' | i \in \boldsymbol{F}_t\})} \tag{4}$$

基于不同阈值$\theta_d'$($\theta_d' \in [0, 1]$)可以绘制归一化精确度图。为了避免人工选取阈值引入的差异，VideoCube以预测结果中心$\boldsymbol{c}_p$落入目标矩形框$\boldsymbol{g}_t$的帧在序列中的比例进行排序。




(a) 传统精确度指标(PRE)的局限性　(b) 归一化精确度指标(N-PRE)范例(Hu等，2023)

图13　单目标跟踪任务评价指标

Fig. 13　Evaluation metrics for SOT task

((a) deficiencies of traditional precision metric (PRE); (b) normalized precision metric (N-PRE) (Hu et al., 2023))

2)成功率。成功率(success rate，SR)用于度量真值$\boldsymbol{g}_t$和算法预测结果$\boldsymbol{p}_t$之间的交并比(intersection over union，IoU)，并统计IoU大于特定阈值的比例，具体为

$$S(\boldsymbol{G}) = \frac{1}{|\boldsymbol{G}|}\sum_{\boldsymbol{s}_i \in \boldsymbol{G}}\frac{1}{|\boldsymbol{s}_i|}\left|\{\boldsymbol{F}_t : s_t \geqslant \theta_s\}\right| \tag{5}$$

$$s_t = \Omega(\boldsymbol{p}_t, \boldsymbol{g}_t) = \frac{\boldsymbol{p}_t \cap \boldsymbol{g}_t}{\boldsymbol{p}_t \cup \boldsymbol{g}_t} \tag{6}$$

式中，$\Omega$代表IoU计算。

基于不同阈值$\theta_s$的统计结果可以绘制成一条成功率曲线，成功率图由不同算法的成功率曲线组成。其中，$\theta_s = 0.5$通常用于对目标跟踪算法的成功率进行排序。

3）鲁棒性。假设算法在视频序列$\boldsymbol{s}_i$中失败后一共重启$I_i$次，因为$I_i$没有上界，因此VOT挑战赛（Čehovin等，2016）利用可靠性来进行鲁棒性评估，具体为

$$S \times R_s = \mathrm{e}^{-SM} \tag{7}$$

$$M = I_i / \left| \boldsymbol{s}_i \right| \tag{8}$$

式中，$M$代表平均失败间隔时间，$S$是人工选择的参数（一般选用$S = 30$），可靠性$R_s$是参数$S$下算法跟踪失败次数的减函数。Hu等人（2023）认为鲁棒性不仅与跟踪失败次数有关，也与跟踪序列的挑战因素相关。因此，VideoCube中的鲁棒性定义为

$$R(\boldsymbol{G}) = \frac{1}{|\boldsymbol{G}|} \sum_{\boldsymbol{s}_i \in \boldsymbol{G}} \left[ S(1/\overline{\rho}_i)(1 - I_i/R_i) \right] \tag{9}$$

$$\overline{\rho}_i = \frac{1}{\left| \boldsymbol{s}_i \right|} \sum_{F_t \in \boldsymbol{s}_i} \left[ \frac{COV(F_t, F_{t-1})}{\sigma_{F_t} \sigma_{F_{t-1}}} \right] \tag{10}$$

式中，$\overline{\rho}_i$表示视频序列$\boldsymbol{s}_i$的平均相关系数，$R_i$表示该段视频包含的重启帧总数（注意，$I_i$指代的是算法在序列$\boldsymbol{s}_i$中实际重启的次数，而$R_i$指代的是序列$\boldsymbol{s}_i$中适合作为重启帧的总帧数），$COV(\cdot)$为协方差，$\sigma$为标准差。

4）期望平均覆盖率。为了给算法的性能评估提供一个更加综合的指标，VOT竞赛在结合鲁棒性和成功率的基础上，提出了期望平均覆盖率（expected average overlap，EAO）指标。具体而言，对于一段长度为$N_s$的测评序列，首先基于式（6）逐帧计算算法预测结果与真值之间的交并比，并得到该序列上的平均覆盖率，具体为

$$\Phi_{N_s} = \frac{1}{N_s} \sum_{t=1}^{N_s} \Omega(\boldsymbol{p}_t, \boldsymbol{g}_t) \tag{11}$$

通过对评测环境中所有长度为$N_s$的序列上的平均覆盖率$\Phi_{N_s}$计算均值，可以得到期望平均覆盖率$\hat{\Phi}_{Ns}$。

5）F-score。对于包含检测模块的单目标跟踪算法，Lukezic等人（2021）借鉴检测领域的相关指标，提出F-score来评估算法的性能，该指标的计算过程需要依赖精确率$P_r$和召回率$R_e$的结果，并进行综合计算。鉴于F-score在单目标跟踪任务的评测中并不通用，因此本文不对其展开介绍。

#### 4.1.3 评测结果及分析

表4列举了3.1节介绍的目标跟踪代表性方法在短时跟踪和长时跟踪上的评测结果。表中所列方法均在OPE机制下进行评估。其中，PRE指代经典的精确度得分，采用20像素为阈值从精确度图中计算出表中数据；SR@AUC为成功率图的曲线下面积（area under curve，AUC），其等价于对所有阈值下成功率得分计算均值；$\mathrm{SR}_{0.5}$为以0.5为阈值时从成功率曲线中计算得出的分数；mAO（mean average overlap）为GOT-10k数据集的专用指标，旨在度量不同类别下算法的平均重叠率；NPRE为LaSOT和TrackingNet采用的基于目标尺寸对精确度标准化后的结果。

1）短时跟踪任务。在OTB100基准上，核相关滤波器（kernel correlation filter，KCF）取得了显著优于经典算法TLD的跟踪性能，成为相关滤波跟踪器的代表性工作；SiamFC算法相比于KCF有了7.6%的精确度提升和10.5%的成功率提升，验证了孪生神经网络结构的有效性。由于OTB100发布时间较早、规模较小，且自发布起已公开全部真值，因此2017年之后的大部分算法在该基准上均取得良好的跟踪效果。2019年发布的大规模高质量通用物体跟踪基准GOT-10k旨在对泛化性开展评估，因此与OTB100相比，大部分跟踪算法在GOT-10k上均出现明显的性能下降，凸显出挑战因素对单目标跟踪任务的影响。值得注意的是，KYS算法因对场景信息进行建模，因此对目标突变和相似物体干扰等挑战因素的适应性更强；SiamRCNN算法则通过重检测机制来应对目标跟踪过程中的表观信息和运动信息突变，从而在GOT-10k上获得鲁棒的跟踪效果。

2）长时跟踪任务。长时跟踪取消了“目标始终存在于画面中”的约束，因而对鲁棒性提出更高要求。在2019年之前，因缺乏相关的长时跟踪基准，大部分算法均针对短时跟踪进行建模，例如仅在上一帧目标的周围设置搜索区域，因此难以应对目标消失—再现问题。针对长时跟踪任务的挑战性，利用场景信息的KYS算法和采用重检测机制的SiamRCNN算法可以较好地应对目标消失引起的信息突变。此外，KeepTrack算法从LaSOT基准中挑选出包含大量干扰物的视频用于训练，因而在长时跟踪任务中具有良好的鲁棒性。而基于端到端训练的MixFormer则充分发挥注意力机制的优势，在短时跟踪和长时跟踪中均取得优秀的效果。

3）全局实例跟踪任务。表5列举了3.1节的目标跟踪代表性方法在全局实例跟踪任务上的评测结果，选用的测试环境是大规模多维度全局实例跟踪基准VideoCube，评测机制为OPE机制和R-OPE机

表4　代表性单目标跟踪算法在短时跟踪和长时跟踪上的性能（基于OPE机制）

Table 4　Performance of representative SOT algorithms on short-term and long-term tracking (based on the OPE mechanism)

| 算法 | 短时跟踪 | | | | 长时跟踪 | | |
|---|---|---|---|---|---|---|---|
| | OTB100 | | GOT-10k | | LaSOT | | |
| | PRE | SR@AUC | $SR_{0.5}$ | mAO | PRE | NPRE | SR@AUC |
| TLD（Kalal等，2012） | / | 0.406 | / | / | 0.174 | 0.193 | 0.201 |
| KCF（Henriques等，2015） | 0.695 | 0.477 | 0.263 | 0.279 | 0.166 | 0.190 | 0.178 |
| SiamFC（Bertinetto等，2016） | 0.771 | 0.582 | 0.426 | 0.392 | 0.339 | 0.420 | 0.336 |
| ECO（Danelljan等，2017） | 0.909 | 0.687 | 0.407 | 0.395 | 0.301 | 0.338 | 0.324 |
| SiamRPN（Li等，2018） | 0.851 | 0.637 | / | / | / | / | / |
| DaSiamRPN（Zhu等，2018） | 0.881 | 0.658 | 0.461 | 0.417 | 0.322 | 0.405 | 0.333 |
| ATOM（Danelljan等，2019） | 0.879 | 0.667 | 0.634 | 0.556 | 0.497 | 0.570 | 0.499 |
| SiamRPN++（Li等，2019） | 0.915 | 0.696 | 0.618 | 0.518 | 0.493 | 0.570 | 0.495 |
| SiamDW（Zhang和Peng，2019） | 0.850 | 0.640 | / | / | 0.329 | 0.437 | 0.347 |
| SPLT（Yan等，2019） | / | / | / | / | 0.396 | 0.494 | 0.426 |
| DiMP（Bhat等，2019） | 0.899 | 0.686 | 0.717 | 0.611 | 0.563 | 0.642 | 0.560 |
| GlobalTrack（Huang等，2020） | / | / | 0.681 | 0.579 | 0.528 | 0.597 | 0.517 |
| SiamFC++（Xu等，2020） | / | 0.683 | 0.695 | 0.595 | 0.547 | 0.623 | 0.544 |
| Ocean（Zhang等，2020） | 0.920 | 0.684 | 0.721 | 0.611 | / | / | 0.560 |
| KYS（Bhat等，2020） | / | 0.695 | 0.751 | 0.636 | 0.640 | 0.707 | 0.619 |
| SiamCAR（Guo等，2020） | / | / | 0.670 | 0.569 | / | / | 0.507 |
| PrDiMP（Danelljan等，2020） | / | 0.696 | 0.738 | 0.634 | 0.608 | 0.688 | 0.598 |
| SiamRCNN（Voigtlaender等，2020） | 0.891 | 0.701 | / | 0.649 | 0.722 | / | 0.648 |
| LTMU（Dai等，2020） | / | / | / | / | 0.535 | 0.621 | 0.539 |
| KeepTrack（Mayer等，2021） | / | 0.709 | / | / | 0.702 | 0.772 | 0.671 |
| MixFormer（Cui等，2022） | / | / | 0.857 | 0.756 | 0.763 | 0.799 | 0.701 |

注："/"表示算法的原始论文未在该数据集和指标的组合下进行测试。

制。表中所列方法分别在OPE、R-OPE和视觉图灵这3种机制下进行评估，人类实验者仅在视觉图灵机制下进行评估。其中，PRE指代经典的精确度得分，采用20像素为阈值从精确度图中计算出表中数据；N-PRE为VideoCube提出的携带惩罚项的归一化精确度得分，并以算法成功落入中心区域的比例作为得分；SR@AUC为成功率图的曲线下面积，其等价于对所有阈值下成功率得分计算均值；Robust为VideoCube提出的鲁棒性得分。与短时跟踪或长时跟踪相比，全局实例跟踪取消了连续运动假设，因此评测环境中包含镜头切换和场景转换。对比表4和表5中的OPE机制，所有算法在全局实例跟踪评测环境中性能均出现大幅度下降。一方面体现出全局实例跟踪的难度；另一方面揭示现有跟踪算法在建模过程中依旧依赖于目标连续运动假设，导致算法难以应对真实任务场景中的目标信息突变。不同于OPE机制，R-OPE机制在检测到跟踪失败后，会在距离失败点最近的重启帧对算法进行重新初始化。因此，R-OPE机制下的精确度和成功率得分侧重于体现算法在目标连续运动时的跟踪能力，而通过量化重启次数得到的鲁棒性得分则体现算法在目标信息突变时的重新跟踪能力。

表5　代表性单目标跟踪算法在全局实例跟踪上的性能

Table 5　Performance of representative SOT algorithms on global instance tracking

| 算法 | 全局实例跟踪（VideoCube） | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | OPE | | | R-OPE | | | | 视觉图灵 | |
| | PRE | N-PRE | SR@AUC | PRE | N-PRE | SR@AUC | Robust | PRE | N-PRE |
| TLD（Kalal等，2012） | 0.018 | 0.266 | 0.026 | 0.017 | 0.261 | 0.026 | 0.687 | 0.019 | 0.293 |
| KCF（Henriques等，2015） | 0.010 | 0.026 | 0.079 | 0.223 | 0.621 | 0.391 | 0.722 | 0.005 | 0.141 |
| SiamFC（Bertinetto等，2016） | 0.025 | 0.120 | 0.056 | 0.250 | 0.514 | 0.345 | 0.725 | 0.044 | 0.143 |
| ECO（Danelljan等，2017） | 0.024 | 0.255 | 0.116 | 0.294 | 0.725 | 0.469 | 0.732 | 0.028 | 0.208 |
| SiamRPN（Li等，2018） | 0.119 | 0.456 | 0.283 | 0.316 | 0.712 | 0.496 | 0.734 | 0.132 | 0.371 |
| DaSiamRPN（Zhu等，2018） | 0.115 | 0.453 | 0.281 | 0.317 | 0.710 | 0.495 | 0.734 | 0.136 | 0.390 |
| ATOM（Danelljan等，2019） | 0.115 | 0.425 | 0.251 | 0.338 | 0.737 | 0.517 | 0.736 | 0.151 | 0.408 |
| SiamRPN++（Li等，2019） | 0.198 | 0.538 | 0.351 | 0.375 | 0.734 | 0.525 | 0.737 | 0.262 | 0.521 |
| SiamDW（Zhang和Peng，2019） | 0.075 | 0.463 | 0.146 | 0.272 | 0.714 | 0.458 | 0.731 | 0.106 | 0.431 |
| SPLT（Yan等，2019） | 0.135 | 0.532 | 0.325 | 0.258 | 0.700 | 0.461 | 0.732 | 0.158 | 0.501 |
| DiMP（Bhat等，2019） | 0.176 | 0.520 | 0.356 | 0.364 | 0.753 | 0.550 | 0.738 | 0.260 | 0.487 |
| GlobalTrack（Huang等，2020） | 0.262 | 0.688 | 0.434 | 0.353 | 0.706 | 0.519 | 0.740 | 0.405 | 0.687 |
| SiamFC++（Xu等，2020） | 0.112 | 0.418 | 0.261 | 0.316 | 0.713 | 0.494 | 0.735 | 0.153 | 0.412 |
| Ocean（Zhang等，2020） | 0.179 | 0.523 | 0.328 | 0.379 | 0.730 | 0.505 | 0.735 | 0.256 | 0.476 |
| SiamCAR（Guo等，2020） | 0.095 | 0.321 | 0.151 | 0.340 | 0.701 | 0.476 | 0.733 | 0.142 | 0.400 |
| PrDiMP（Danelljan等，2020） | 0.260 | 0.617 | 0.421 | 0.404 | 0.780 | 0.571 | 0.740 | 0.354 | 0.590 |
| SiamRCNN（Voigtlaender等，2020） | 0.424 | 0.662 | 0.536 | 0.548 | 0.785 | 0.643 | 0.743 | 0.551 | 0.710 |
| LTMU（Dai等，2020） | 0.276 | 0.641 | 0.446 | 0.398 | 0.778 | 0.562 | 0.730 | 0.421 | 0.662 |
| 人类@15 FPS（Hu等，2023） | / | / | / | / | / | / | / | 0.377 | 0.850 |
| 人类@20 FPS（Hu等，2023） | / | / | / | / | / | / | / | 0.243 | 0.805 |
| 人类@30 FPS（Hu等，2023） | / | / | / | / | / | / | / | 0.203 | 0.778 |

注："/"表示算法的原始论文未在该数据集和指标的组合下进行测试。最后3行为人类被试在3种播放速度（15 FPS、20 FPS、30 FPS）下的视觉图灵实验结果。

ATOM系列算法将深度网络在大规模数据集上离线训练的优势和相关滤波在线更新的优势相结合，并通过对损失函数判别性的优化和对目标回归精确度的改进进一步增强算法的鲁棒性，其中代表性的PrDiMP算法在全局实例跟踪任务中展现了良好的跟踪效果。在SPLT算法提出的局部搜索和全局搜索相结合的策略基础上，LTMU算法采用元学习来决策跟踪过程中目标更新问题，因此具有良好的鲁棒性。得益于对序列中所有潜在目标运动信息的存储，SiamRCNN算法可以将目标与画面中的干扰物进行区分，避免重检测时错误定位到干扰物，因而在两种评测机制下均表现良好。GlobalTrack算法在设计思路上更适配全局实例跟踪，并优于同期的其他算法。值得注意的是，在单目标跟踪任务中，利用好连续帧之间的运动物体在表观和运动上的时序相关性有利于实现更稳定、平滑的跟踪。但是，简单的时序依赖性假设会带来严重的累计误差问题。虽然GlobalTrack提供了一个基于全局检索的零累计误差解决方案，但其对于运动模型的处理较为简单，难以充分利用连续视频帧间的时序信息。因此，如何在累计误差与时序依赖关系的有效利用之间取得平衡，是全局实例跟踪算法的建模难点。

### 4.2　以人机对抗为核心的视觉图灵评估

#### 4.2.1　视觉图灵评估范式

以人类视觉系统为建模对象，实现近似或者超越人类的视觉智能，是计算机视觉的发展目标。但

是，以机机对抗为核心的传统评估思路缺乏与人类实验者的视觉能力对比，因此难以综合评估视觉智能的发展程度。针对上述问题，黄凯奇等人(2021)提出视觉图灵评估范式，旨在以人类视觉能力为基准，对算法的视觉智能进行度量。

视觉图灵思想源自图灵测试。图灵首先设计一个模拟游戏，并利用机器代替人类来回答问题。当提问者无法判断对方是人类还是机器时，则可认为机器具备了智能。图灵测试通过一种可操作的方式度量智能，启发研究者基于其核心的人机对抗思想开展工作(Turing，2009)。以AlphaGo(Silver等，2017)和DeepStack(Brown和Sandholm，2018)为代表的智能体在决策问题中战胜了人类精英，成为机器智能发展的里程碑式工作(黄凯奇 等，2020)。

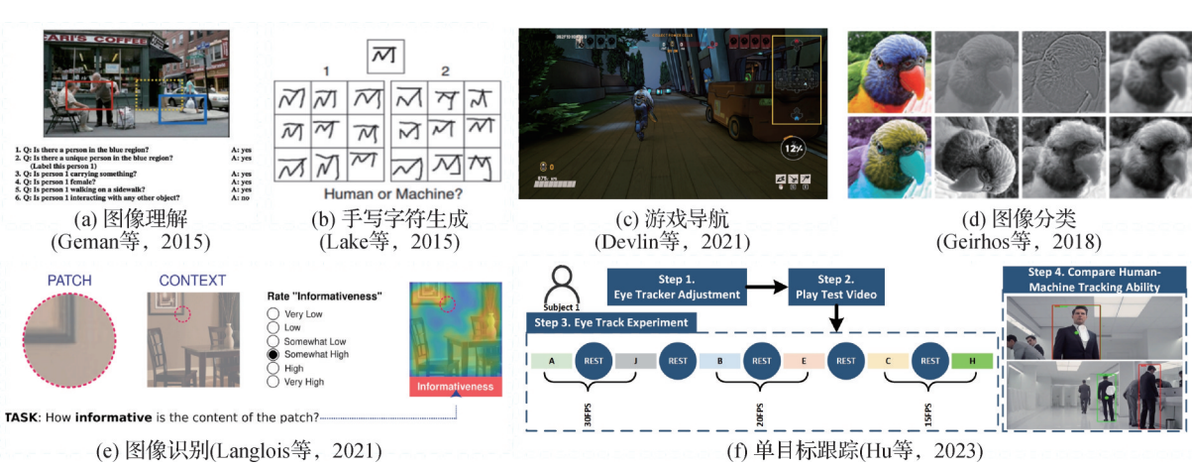
作为图灵测试与计算机视觉领域的交叉点，视觉图灵评估范式逐渐被研究者关注。在图像理解任务中，Geman等人(2015)通过没有歧义的二值问题对机器进行提问，并根据机器的回答判定其是否可以像人类一样对自然图像进行理解(图14(a))。

在生成式视觉任务中，传统评测指标难以对生成效果进行量化评估，因而视觉图灵成为一种可行的评测思路，并应用在手写字符生成(Lake等，2015)、图像染色(Zhang等，2016)等视觉任务中(图14(b))。在游戏导航任务中，Devlin等人(2021)要求智能体和人类在同一游戏中执行导航任务，并将二者的游戏操作视频展示给观测者。如果观测者无法区分人类和智能体，则判定该智能体具备与人类相似的操作风格。

上述工作虽然将图灵测试引入视觉任务中，但这些任务相对宽泛，且评测指标通常基于二值化问题进行设计，难以有效量化人类视觉能力。因此，部分研究者从细粒度任务评估入手开展工作。

1)静态视觉任务。作为一项基础的静态视觉任务，图像分类要求算法根据图像特征对其进行分类，并从已知的类别标签集合中选定一个类别标签。近年来，研究者从构建评测环境、优化评测标准、设计视觉图灵实验等角度开展系列工作(Geirhos等，2018，2020b，2021)，并对人类和算法在图像分类任务中的差距进行度量和分析。研究者首先通过数字图像处理技术对原始图像进行操作(图14(d))，使其与训练数据具有不同分布；然后组织90位被试者执行图像分类任务，并对比人类与52种不同架构的算法在分类任务中的准确率和一致性。研究表明，在图像分类任务中算法与人类的性能差距正在逐渐缩短；其中，大规模预训练模型在部分测试环境中获得了超越人类的性能。Langlois等人(2021)针对图像识别任务设计了视觉图灵实验(图14(e))，通过对比人类与算法在图像识别过程中注意力区域的差异性，为神经网络的可解释性研究提供依据。

2)动态视觉任务。与静态视觉任务相比，动态视觉任务的视觉图灵实验更为复杂，设计难点包括评测环境构建、原始结果获取和评测指标选取等。而神经科学领域相关的人类动态视觉度量



(a) 图像理解 (Geman等，2015)
(b) 手写字符生成 (Lake等，2015)
(c) 游戏导航 (Devlin等，2021)
(d) 图像分类 (Geirhos等，2018)
(e) 图像识别(Langlois等，2021)
(f) 单目标跟踪(Hu等，2023)

图14　视觉图灵评估范式应用示意图

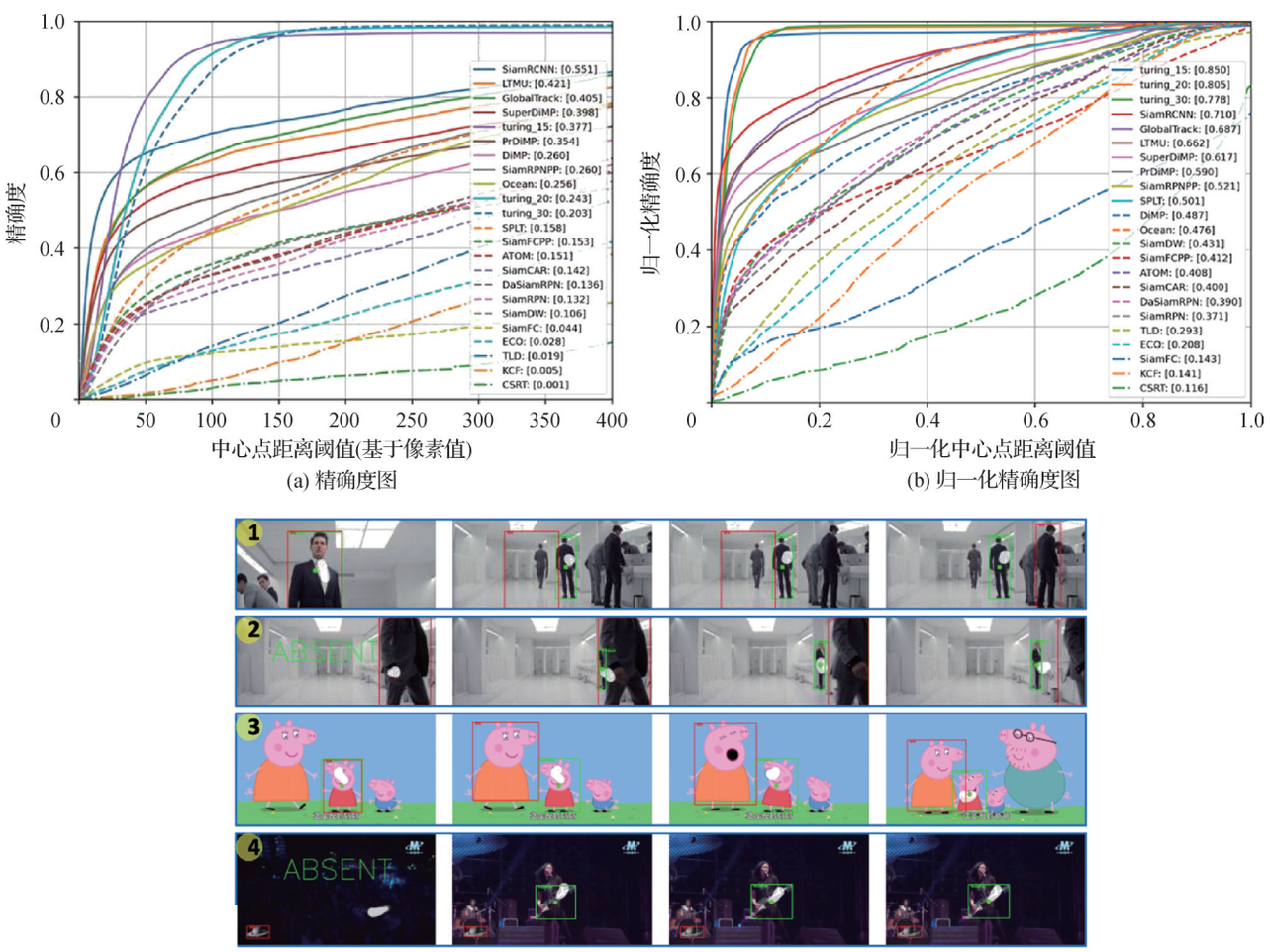
Fig. 14　The applications of the visuhal Turing evaluation mechanism ((a) image comprehension (Geman et al.，2015); (b) handwritten character generation (Lake et al.，2015); (c) game navigation (Devlin et al.，2021); (d) image classification (Geirhos et al.，2018); (e) image recognition (Langlois et al.，2021); (f) single object tracking (Hu et al.，2023))

方案存在评测环境简单、运动轨迹单一、评测指标简略等缺陷，与视觉图灵的初衷相距甚远（详见3.2节）。

针对上述问题，Hu等人（2023）依托VideoCube评测环境开展人类动态视觉能力的度量实验，并在全局实例跟踪任务中对人类和机器的视觉智能进行对比分析（图14(f)）。在视觉图灵实验中，15位被试者需要在3种不同的播放速度下观看测试视频，并通过眼动仪逐帧记录其视线焦点。被试者的视线焦点将被记录为一个二维坐标 $\boldsymbol{c}_p$，并与真值矩形框中心点 $\boldsymbol{c}_g$ 进行对比，最后基于精确度指标进行性能评估（参考4.1.2节的精确度指标计算公式）。实验步骤包含机器校准、操作培训及测试、正式眼动实验和视觉图灵分析，不仅为对比人类和机器的目标跟踪能力提供了一个客观公正的实验方案，也为评估机器跟踪智能提供支持。

### 4.2.2 评测结果及分析

表5和图15展示了基于全局实例跟踪任务的视觉图灵实验结果。考虑到算法在不同硬件环境下执行任务所需的时间不同（即一个低运行速度算法可以通过改善计算资源来提升计算速度），且静态视觉任务中的代表性工作也未对算法的处理速度进行约束，因此研究者主要对比人类实验者与算法的精确度差异，而未对算法的运行速度进行限制。




(a) 精确度图　(b) 归一化精确度图

(c) 人类视觉跟踪能力优于算法的实例

图15　基于全局实例跟踪任务的视觉图灵实验（Hu等，2023）

Fig. 15　Visual Turing experiments based on the global instance tracking task (Hu et al., 2023)

((a) precision plot; (b) normalized precision plot; (c) examples of human visual tracking ability outperforming algorithms)

由图15(a)可知，基于传统精确度指标（详见4.1.2节中式(1)(2)）的评估方式存在一定的局限性，因为其仅度量预测结果 $c_p$ 和目标真值矩形框中心点 $c_g$ 之间的距离，忽略了目标尺寸、画面分辨率和仪器误差的影响，因此直接沿用OTB100基准提出的以中心点距离小于20像素的比例作为视觉图灵评测机制下的排序标准并不合理。

图15(b)展示了基于归一化精确度指标（详见

4.1.2节中式(3)(4))下人类和算法的实验结果。在精确度要求较为严格时(左侧曲线),人类视觉跟踪能力不如算法,其可能受眼动仪误差和被试者注意力偏差的影响。例如,若待跟踪的目标为人类,则被试者倾向于注视目标头部,造成视线焦点与目标中心存在差异。反之,在精确度要求适度放宽的情况下,人类视觉目标跟踪能力会迅速超过算法,并可以在挑战因素下依旧保持稳定跟踪。

图15(c)展示了人类视觉跟踪能力优于算法的部分示例。序列1展示当出现镜头切换时,人类可以迅速在新镜头中定位目标,但算法却漂移到相似物体上。序列2展示当遮挡物被移除时,人类可以快速找到遮挡物后面的目标,但算法却再次定位在相似物体上。序列3展示人类可以区分同类物体并准确找到目标的位置,但算法很难将目标和干扰物分开。序列4展示在镜头切换和弱光照同时出现的情况下,算法无法找寻目标(吉他)的位置,但人类可以根据吉他手的位置在下一镜头对吉他进行快速定位。

通过对比人机的结果,视觉图灵机制揭示了当前算法的局限性,并启发研究者对其进行优化。例如,镜头切换代表连续帧出现目标信息突变,如何有效利用时序依赖关系提升跟踪稳定性,同时应对镜头切换对时序依赖性的破坏,从而实现类人的目标跟踪性能,是算法设计需要进一步思考的要点。此外,在目标信息突变时,如何像人类一样通过高效利用背景信息来对其进行重新定位(序列4),同时规避背景信息中的相似物体干扰(序列1和序列3),也是算法设计需要权衡的关键点。

## 5 结 语

随着大算力、大数据和大模型的普及,诸多单目标跟踪算法在代表性评测环境上已经达到较高的性能,但在真实应用中表现并不如人意,与人类的鲁棒视觉跟踪能力存在较大差异。这一现象揭示了现有评估技术存在较大的局限性,未能真正度量算法在目标跟踪任务上的智能程度。因此,本文首先从跟踪智能效果评估出发,对评测中涉及的评测任务(第1节)、评测环境(第2节)、待测对象(第3节)和评估机制(第4节)进行系统梳理。在此基础上,对单目标跟踪智能评估的发展趋势进行总结和展望,进一步分析存在的挑战因素,并探讨未来可能的研究方向。

### 5.1 评测任务类人化

大部分研究的关注点集中在数据集规模的大小、模型结构的复杂程度、统计指标的计算公式等细节上,忽略了视觉研究的目的,即让机器具备自然视觉的能力。因此,以人类视觉系统为建模对象,全面对比和分析现有任务定义的局限性,并通过去除隐藏的约束条件实现对任务定义的拓展,从而从本质上确保算法研究可以像人类智能靠拢,对于计算机视觉的发展至关重要。

人类可以在任意场景中持续定位任意目标,在复杂场景尤其是对抗环境下依旧保持鲁棒跟踪能力。事实上,从短时跟踪到长时跟踪、再到全局实例跟踪的发展历程,正是单目标跟踪任务去掉隐藏的任务约束、逐步向人类目标跟踪能力靠拢的过程。通过对人类视觉跟踪能力进行精确建模,单目标跟踪任务的应用范围从评测环境拓展至更具挑战性的真实应用场景中。单目标跟踪任务定义的拓展历程,不仅为后续智能跟踪方法的设计奠定基础,其所蕴含的任务定义类人化研究思路也为其他任务提供了一种可以参考的发展路线。

### 5.2 评测环境真实化

评测环境的发展与任务定义的演变息息相关,研究者需要在充分理解任务的基础上,通过合理的构建准则来建模任务真实应用场景,并需要全面覆盖任务挑战因素来为评估评测提供数据支持。

构建评测环境包含数据设计、数据雕刻和数据应用等关键步骤(Liang等,2022)。从2013年之前的零散序列到如今包含746万帧的大规模高质量VideoCube基准,近十年来研究者基于上述流程不断提出高质量的评测环境,促使单目标跟踪的基准逐步从简易的玩具场景向人类所生活的真实应用场景靠拢(图4)。

1)构建设计准则。评测环境的设计重点并非单纯扩大数据规模,而在于如何对任务的真实应用场景进行充分建模。通用评测环境(2.1节)的建模对象通常为真实世界。受限于数据集的规模,以OTB为代表的早期基准仅考虑到目标类别、挑战因素等构建要素,与真实场景仍存在较大的差异。2018年发布的OxUvA和TrackingNet基准代表通用跟踪数据集进入大规模时代。其中,GOT-10k基准从单目

标跟踪任务定义中“类别无关”这一核心特点出发，针对泛化性的评估评测开展研究，并对真实应用场景中存在的563类运动物体和87类运动模式进行采集。VideoCube基准提出6D准则，将视频序列的叙事关系解析为6个采集维度，首次将时空关系纳入到数据采集流程中，从更多元化的环境建模角度实现对真实场景的刻画。

值得注意的是，任务挑战因素（1.2节）始终是基准构建过程中需要充分考虑的要素。2013年提出的OTB50基准对单目标跟踪任务的常见挑战因素进行归纳总结，并为视频序列提供挑战因素的标注。在此基础上，后续代表性基准（Mueller等，2016；Huang等，2021；Fan等，2021a；Hu等，2023）在构建过程中均充分考虑挑战因素的影响，并为目标形变、快速运动和目标遮挡等常见挑战提供标注。其中，VideoCube基准参考人类的视觉系统，将目标跟踪任务的挑战分为静态因素和动态因素，并为挑战因素提供逐帧的细粒度标注。

专用评测环境（2.2节）的构建准则需要充分体现特定对象或者特定场景所引入的新挑战。例如，透明物体跟踪数据集TOTB（Fan等，2021b）中有67.5%的视频序列出现背景杂乱，进一步凸显了透明物体跟踪的挑战性。在基于无人机场景的目标跟踪数据集中（Mueller等，2016；Li和Yeung，2017；Yu等，2020），目标平均相对尺寸（目标矩形框对角线与视频帧对角线的比值）通常低于0.07，远小于通用评测环境（通常在[0.1，0.25]区间），充分体现了无人机场景下的微小目标跟踪挑战。

2）提升数据质量。数据雕刻（data sculpting）对于评测环境的质量至关重要。大量的数据噪声和错误的标签不仅会导致模型学习到错误的信息（Liang和Zou，2022），同时对评测结果的可信度也存在较大影响（Geirhos等，2020a）。诸多基准通过定制精细的任务标注规范来确保数据质量，如LaSOT通过精确调整来提升原始标注的精确度，VideoCube则针对微小区域、透明特征以及大范围摆动的柔性特征提出特殊标注原则。

3）确保评估效果。数据应用包含数据集划分、配套工具设计等步骤，旨在帮助用户高效利用评测环境开展研究，并确保评估结果的公正性。例如，GOT-10k遵循训练、测试类别不重叠的开集测试规范，确保评估结果在未知类别运动物体和运动形式上的泛化能力，这一原则也被LaSOT和VideoCube沿用。此外，数据集以何种形式发布也是需要思考的问题。大部分基准（Wu等，2013，2015；Mueller等，2016；Fan等，2021a）在发布的过程中通常会直接公开所有序列的真值，但泄露测试集的真值将对评估可信度造成较大影响。因此，GOT-10k和VideoCube选择仅公开训练集和验证集的真值，测试集则采用在线评估平台的形式对算法进行排序，为公平评估提供保证。

4）挖掘困难片段。虽然研究者已经从数据设计、数据雕刻和数据应用的角度充分考虑评测环境构建准则，并推动相关研究向真实应用场景靠拢，但是仍有一些问题亟待解决。在单目标跟踪过程中，导致跟踪失败的困难帧一般分布稀疏，而大部分视频帧里的运动目标通常较容易定位。现有基于深度学习的跟踪方法通常无差别地采样视频帧数据用于模型训练，这种做法通常十分低效，难以充分利用关键帧信息来学习如何处理挑战因素。从这一角度出发，如何充分挖掘和利用视频中导致跟踪失败的关键片段，并基于困难片段开展算法训练和评估，值得研究者进行探索。

在算法训练方面，以KeepTrack为代表的部分工作选择从评测环境中手动挑选出包含特定挑战因素的视频序列用于训练；在算法评估方面，Fan等人（2021c）提出TracKlinic基准，首先从评测环境中手动裁剪出2 390段平均长度为115帧的短序列，并确保每个短序列有且只有一个挑战因素，然后通过评估算法在不同短序列上的表现，寻找最易造成算法跟踪失败的挑战因素。但是，上述工作均基于手动筛选的方式对原始评测环境进行挖掘，难以进行大范围推广和应用。

针对这一问题，Hu等人（2024）提出将代表性单目标跟踪评测环境融合为一个包含1 256万帧数据的大型空间SOTVerse，并通过子空间构建准则从中快速抽取包含特定挑战因素的视频序列开展研究。SOTVerse依据单目标跟踪任务特点，分别选取具有代表性的短时跟踪基准、长时跟踪基准及全局实例跟踪基准组成常规空间。随后，SOTVerse基于自动标注算法为10种目标跟踪任务的挑战因素提供逐帧属性标注，并依据用户的研究目标快速从常规空间中抽取相关子序列，形成用户自定义的单目标跟踪任务空间，从而实现基于研究资源高效开展算法

训练和评估。

### 5.3 待测对象多元化

正如5.1节所述，计算机视觉的研究目标是让机器具备类人的视觉智能。机器智能通常划分为计算智能、感知智能和认知智能等若干层次，并在任务定义类人化和评测环境真实化的过程中不断进化。其中，计算智能主要指机器具备科学运算、逻辑处理和统计查询等处理规则化任务的能力；感知智能主要指机器具备通过各种传感器获取信息的能力；认知智能则代表机器具备记忆、理解和推理等更高层的能力。

随着深度学习技术的发展，单目标跟踪算法的智能程度呈现出从感知智能到认知智能的发展趋势。短时跟踪可以视做一种感知智能任务，运动目标始终出现在单一镜头中，因此将模板区域和搜索区域进行匹配的孪生神经网络系列算法（Bertinetto等，2016；Li等，2018，2019；Zhu等，2018）可以凭借简洁的模型设计取得良好的短时跟踪效果。长时跟踪和全局实例跟踪中广泛存在的目标消失及镜头切换等挑战会破坏目标的表观信息和运动信息，当目标再次出现在画面中时，算法需要区分目标和背景中的干扰物，并重新对目标进行定位。在此过程中，算法需要理解目标在运动过程中的表观信息和运动信息变化，并准确辨认出重新出现的目标。因此，跟踪算法在孪生网络的基础上借鉴人类动态视觉能力，通过编码场景信息（Bhat等，2020）、设计重检测机制（Voigtlaender等，2020）、优化目标信息更新机制（Dai等，2020）、利用注意力机制（Cui等，2022）等方法，逐步实现从感知智能向认知智能的进化。

综上，将人类作为待测对象引入目标跟踪中，并以人类视觉能力为标准设计机器智能的进化路线，不仅可以更好地指导算法设计，也为后续通过视觉图灵范式开展算法智能评估奠定了基础。

值得注意的是，现有人类动态视觉能力度量方案仍有巨大的改进空间。正如3.2节所述，对于人类视觉能力的研究和分析主要集中在神经科学领域，但相关研究存在目标结构简单、轨迹单一和背景单调等缺点，与真实应用场景相距甚远，难以综合度量人类的视觉智能。此外，虽然计算机视觉研究者在任务设置、评测环境设计和挑战因素覆盖等方面进行优化，并基于全局实例跟踪任务在VideoCube评测环境中开展动态视觉实验、从而实现对人类动态视觉能力的度量和分析，但仍存在若干问题可以进一步优化。

1）优化度量方式。眼动仪是记录眼动行为最常用的实验仪器（Xia等，2021），其主要采用瞳孔—角膜反射技术，视线方向由瞳孔中心相对于角膜反射的位置确定。然而，由眼动仪获取的眼动数据质量受多种因素影响。例如，被试者在实验过程中可能会出现头部移动，导致仪器不能准确捕获其视线焦点。因此，采用高精度眼动仪并设置严苛的眼动评测环境，或者设计全新的人类视觉跟踪能力度量方案，是可以进一步探究的方向。

2）探究被试特性。研究表明，被试者的生理特征和个人特质等因素对于眼动行为和眼动数据质量均存在一定影响（Burg和Hulbert，1961）。此外，如果采用基于眼动仪的实验方案，则被试者的主视眼、瞳孔大小、睫毛方向和眼睑开合度等生理特征也会影响实验精确度。因此，如何基于被试者的特性进行待测对象选择，从而确保被试者群体具有代表性，值得研究者进一步分析。

### 5.4 评估机制智能化

在任务定义类人化、评测环境真实化、待测对象多元化的基础上，单目标跟踪智能评估从以“机机对抗”为核心的传统评估向以“人机对抗”为核心的视觉图灵评估迈进。区别于基于大数据、大算力的评估标准，视觉图灵机制将“人”的因素加入到智能评估的回路中，以人类为基准对机器智能程度开展更有效的评估，从而打破机器和人类认知的鸿沟，帮助机器更好地对人类的视觉和学习过程进行建模，从而实现真正的人工智能。

正如4.2节所述，视觉图灵源自于经典的图灵测试，并已在图像理解（Geman等，2015）、手写字符生成（Lake等，2015）、游戏导航（Devlin等，2021）、图像分类（Geirhos等，2018）、图像识别（Langlois等，2021）和单目标跟踪（Hu等，2023）等视觉任务中得到初步实践，表明视觉图灵已经逐渐成为下一阶段智能评测发展的新方向。但是，作为一个全新的评估机制，视觉图灵仍处于研究的初期，存在较大的探索和研究空间。

1）解耦视觉能力。视觉图灵的发展是一个任务从粗到细的过程。视觉图灵最初针对图像理解任务（Geman等，2015）以二值化问题的形式进行机器智能研究，需要机器同时具备物体分类、物体定位和关系推理等多项能力，难以针对单项能力进行量化和

评估。研究者在图像分类(Geirhos 等,2018,2020b,2021)和单目标跟踪(Hu 等,2023)任务上开展视觉图灵实验,通过聚焦单一视觉任务使其从宽泛的“智力测试”具象化为可执行的评估范式。然而,即使是单一的目标跟踪任务也涉及多种能力的耦合,因此可以进一步对任务进行拆解,并通过细粒度的评测方案对算法的智能程度进行更全面的分析和评估。具体而言,人类在执行目标跟踪任务的过程中会调用观测、记忆和推理等多种能力,这些能力充分覆盖了感知智能和认知智能的范畴。而目标跟踪算法并非直接建模人类完整的动态视觉能力,而是首先对其进行解耦,并按照智能层级进行不断进化。例如,跟踪算法首先需要在短时跟踪任务上具备良好的感知智能,然后在此基础上进行优化,并逐步向认知智能靠拢。因此,在基于视觉图灵范式开展单目标跟踪算法的智能评估时,可以将人类执行任务所涉及的各项能力进行进一步解耦,并通过分别度量人类和机器在细粒度跟踪任务上的能力找到算法的发展瓶颈。

2)优化评测指标。目前,视觉图灵实验主要沿用“机机对抗”的评估指标对人类和机器的能力进行度量。以单目标跟踪任务为例,人类的结果为视线焦点,而算法的结果为矩形框。如何通过更合理的指标设计,实现对人机跟踪结果更科学的评估和分析,值得进一步关注。此外,5.3 节指出,被试者的动态视觉能力也受其生理特征、认知状态和个人特质等因素的影响。而目前的研究依旧是将人类作为一个整体,并综合被试者的实验表现形成人类基准,缺乏对被试者个体差异的考虑。

一种更合理的思路是通过合理的评测指标,对被试者的能力水平进行分级。例如,借鉴竞技游戏中采用的 Elo(Coulom,2007)评级对被试者进行排位,从而将机器视觉智能发展程度与人类视觉能力进行精确对应。

综上,随着研究的深入,视觉图灵这一全新的评估机制可以有效度量算法发展的科学性,并为实现近似或超越人类视觉能力的算法研究提供支持。

## 参考文献(References)

Bao C L, Wu Y, Ling H B and Ji H. 2012. Real time robust L1 tracker using accelerated proximal gradient approach//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 1830-1837 [DOI: 10.1109/CVPR.2012.6247881]

Bertinetto L, Valmadre J, Henriques J F, Vedaldi A and Torr P H S. 2016. Fully-convolutional siamese networks for object tracking//Proceedings of 2016 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]

Bhat G, Danelljan M, van Gool L and Timofte R. 2020. Know your surroundings: exploiting scene information for object tracking//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 205-221 [DOI: 10.1007/978-3-030-58592-1_13]

Bhat G, Johnander J, Danelljan M, Khan F S and Felsberg M. 2018. Unveiling the power of deep tracking//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 493-509 [DOI: 10.1007/978-3-030-01216-8_30]

Biederman I. 1987. Recognition-by-components: a theory of human image understanding. Psychological Review, 94(2): 115-147 [DOI: 10.1037/0033-295X.94.2.115]

Bolme D S, Beveridge J R, Draper B A and Lui Y M. 2010. Visual object tracking using adaptive correlation filters//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE: 2544-2550 [DOI: 10.1109/CVPR.2010.5539960]

Bromley J, Guyon I, LeCun Y, Säckinger E and Shah R. 1993. Signature verification using a “siamese” time delay neural network//Proceedings of the 6th International Conference on Neural Information Processing Systems. Denver, Colorado, USA: Morgan Kaufmann Publishers Inc.: 737-744

Brown N and Sandholm T. 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. Science, 359(6374): 418-424 [DOI: 10.1126/science.aao1733]

Burg A and Hulbert S. 1961. Dynamic visual acuity as related to age, sex, and static acuity. Journal of Applied Psychology, 45(2): 111-116

Čehovin L, Leonardis A and Kristan M. 2016. Visual object tracking performance measures revisited. IEEE Transactions on Image Processing, 25(3): 1261-1274 [DOI: 10.1109/TIP.2016.2520370]

Chen L. 1982. Topological structure in visual perception. Science, 218(4573): 699-700 [DOI: 10.1126/science.7134969]

Chen X, Yan B, Zhu J W, Wang D, Yang X Y and Lu H C. 2021. Transformer tracking//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8122-8131 [DOI: 10.1109/CVPR46437.2021.00803]

Coulom R. 2007. Computing “Elo ratings” of move patterns in the game

of Go. ICGA Journal, 30(4): 198-208 [DOI: 10.3233/ICG-2007-30403]

Cui Y T, Jiang C, Wang L M and Wu G S. 2021. Target transformed regression for accurate tracking [EB/OL]. [2023-03-14]. https://arxiv.org/pdf/2104.00403.pdf

Cui Y T, Jiang C, Wang L M and Wu G S. 2022. MixFormer: end-to-end tracking with iterative mixed attention//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13598-13608 [DOI: 10.1109/CVPR52688.2022.01324]

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. 2017. ECO: efficient convolution operators for tracking//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6931-6939 [DOI: 10.1109/CVPR.2017.733]

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, Häger G, Khan F S and Felsberg M. 2015. Convolutional features for correlation filter based visual tracking//Proceedings of 2015 IEEE International Conference on Computer Vision Workshop. Santiago, Chile: IEEE: 621-629 [DOI: 10.1109/ICCVW.2015.84]

Danelljan M, Robinson A, Khan F S and Felsberg M. 2016. Beyond correlation filters: learning continuous convolution operators for visual tracking//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 472-488 [DOI: 10.1007/978-3-319-46454-1_29]

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]

Devlin S, Georgescu R, Momennejad I, Rzepecki J, Zuniga E, Costello G, Leroy G, Shaw A and Hofmann K. 2021. Navigation Turing test (NTT): learning to evaluate human-like navigation//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR: 2644-2653

Erickson G B, Citek K, Cove M, Wilczek J, Linster C, Bjarnason B and Langemo N. 2011. Reliability of a computer-based system for measuring visual performance skills. Optometry——Journal of the American Optometric Association, 82(9): 528-542 [DOI: 10.1016/j.optm.2011.01.012]

Fan H, Bai H X, Lin L T, Yang F, Chu P, Deng G, Yu S J, Harshit, Huang M Z, Liu J H, Xu Y, Liao C Y, Yuan L and Ling H B. 2021a. LaSOT: a high-quality large-scale single object tracking benchmark. International Journal of Computer Vision, 129(2): 439-461 [DOI: 10.1007/s11263-020-01387-y]

Fan H, Miththanthaya H A, Harshit H, Rajan S R, Liu X Q, Zou Z L, Lin Y W and Ling H B. 2021b. Transparent object tracking benchmark//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10714-10723 [DOI: 10.1109/ICCV48922.2021.01056]

Fan H, Yang F, Chu P, Lin Y W, Yuan L and Ling H B. 2021c. TracKlinic: diagnosis of challenge factors in visual tracking//Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 969-978 [DOI: 10.1109/WACV48630.2021.00101]

Geirhos R, Jacobsen J H, Michaelis C, Zemel R, Brendel W, Bethge M and Wichmann F A. 2020a. Shortcut learning in deep neural networks. Nature Machine Intelligence, 2(11): 665-673 [DOI: 10.1038/s42256-020-00257-z]

Geirhos R, Meding K and Wichmann F A. 2020b. Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 13890-13902

Geirhos R, Narayanappa K, Mitzkus B, Thieringer T, Bethge M, Wichmann F A and Brendel W. 2021. Partial success in closing the gap between human and machine vision [EB/OL]. [2023-07-10]. http://arxiv.org/pdf/2106.07411.pdf

Geirhos R, Temme C R M, Rauber J, Schütt H H, Bethge M and Wichmann F A. 2018. Generalisation in humans and deep neural networks//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 7549-7561

Geman D, Geman S, Hallonquist N and Younes L. 2015. Visual Turing test for computer vision systems. Proceedings of the National Academy of Sciences of the United States of America, 112(12): 3618-3623 [DOI: 10.1073/pnas.1422953112]

Ginsburg A P. 1984. A new contrast sensitivity vision test chart. Optometry and Vision Science, 61(6): 403-407 [DOI: 10.1097/00006324-198406000-00011]

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

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]

Han R Z, Feng W, Guo Q and Hu Q H. 2022. Single object tracking research: a survey. Chinese Journal of Computers, 45(9): 1877-1907 (韩瑞泽, 冯伟, 郭青, 胡清华. 2022. 视频单目标跟踪研究进展综述. 计算机学报, 45(9): 1877-1907) [DOI: 10.11897/

SP.J.1016.2022.01877］

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］

He K M，Zhang X Y，Ren S Q and Sun J. 2016. Deep residual learning for image recognition//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，USA：IEEE：770-778［DOI：10.1109/CVPR.2016.90］

He S F，Lau R W H，Yang Q X，Wang J and Yang M H. 2017. Robust object tracking via locality sensitive histograms. IEEE Transactions on Circuits and Systems for Video Technology，27(5)：1006-1017［DOI：10.1109/TCSVT.2016.2527300］

He S F，Yang Q X，Lau R W H，Wang J and Yang M H. 2013. Visual tracking via locality sensitive histograms//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland，USA：IEEE：2427-2434［DOI：10.1109/CVPR.2013.314］

Henriques J F，Caseiro R，Martins P and Batista J. 2012. Exploiting the circulant structure of tracking-by-detection with kernels//Proceedings of the 12th European Conference on Computer Vision. Florence，Italy：Springer：702-715［DOI：10.1007/978-3-642-33765-9_50］

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］

Hu S Y，Zhao X and Huang K Q. 2024. SOTVerse：a user-defined task space of single object tracking. International Journal of Computer Vision，132(3)：872-930［DOI：10.1007/s11263-023-01908-5］

Hu S Y，Zhao X，Huang L H and Huang K Q. 2023. Global instance tracking：locating target more like humans. IEEE Transactions on Pattern Analysis and Machine Intelligence，45(1)：576-592［DOI：10.1109/TPAMI.2022.3153312］

Huang K Q，Chen X T，Kang Y F and Tan T N. 2015. Intelligent visual surveillance：a review. Chinese Journal of Computers，38(6)：1093-1118（黄凯奇，陈晓棠，康运锋，谭铁牛. 2015. 智能视频监控技术综述. 计算机学报，38(6)：1093-1118）［DOI：10.11897/SP.J.1016.2015.01093］

Huang K Q，Xing J L，Zhang J G，Ni W C and Xu B. 2020. Intelligent technologies of human-computer gaming. Scientia Sinica Informationis，50(4)：540-550（黄凯奇，兴军亮，张俊格，倪晚成，徐博. 2020. 人机对抗智能技术. 中国科学：信息科学），50(4)：540-550［DOI：10.1360/N112019-00048］

Huang K Q，Zhao X，Li Q Z and Hu S Y. 2021. Visual Turing：the next development of computer vision in the view of human-computer gaming. Journal of Graphics，42(3)：339-348（黄凯奇，赵鑫，李乔哲，胡世宇. 2021. 视觉图灵：从人机对抗看计算机视觉下一步发展. 图学学报，42(3)：339-348）［DOI：10.11996/JG.j.2095-302X.2021030339］

Huang L H and Ma B. 2015. Tensor pooling for online visual tracking//Proceedings of 2015 IEEE International Conference on Multimedia and Expo. Turin，Italy：IEEE：#7177452［DOI：10.1109/ICME.2015.7177452］

Huang L H，Zhao X and Huang K Q. 2019. 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. 2020. GlobalTrack：a simple and strong baseline for long-term tracking//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York，USA：AAAI Press：11037-11044［DOI：10.1609/aaai.v34i07.6758］

Huang L H，Zhao X and Huang K Q. 2021. 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］

Hubel D H and Wiesel T N. 1959. Receptive fields of single neurones in the cat's striate cortex. The Journal of Physiology，148(3)：574-591［DOI：10.1113/jphysiol.1959.sp006308］

Hubel D H and Wiesel T N. 1962. Receptive fields，binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology，160(1)：106-154［DOI：10.1113/jphysiol.1962.sp006837］

Hyvärinen L，Walthes R，Jacob N，Chaplin K N and Leonhardt M. 2014. Current understanding of what infants see. Current Ophthalmology Reports，2(4)：142-149［DOI：10.1007/s40135-014-0056-2］

Javed S，Danelljan M，Khan F S，Khan M H，Felsberg M and Matas J. 2023. Visual object tracking with discriminative filters and siamese networks：a survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence，45(5)：6552-6574［DOI：10.1109/TPAMI.2022.3212594］

Kalal Z，Mikolajczyk K and Matas J. 2012. Tracking-learning-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence，34(7)：1409-1422［DOI：10.1109/TPAMI.2011.239］

Kirshner A. 1967. Dynamic acuity a quantiative measure of eye movements. Journal of the American Optometric Association，38(6)：460-462

Kristan M，Matas J，Leonardis A，Vojir T，Pflugfelder R，Fernández G，Nebehay G，Porikli F and Cehovin L. 2016. A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence，38(11)：2137-2155［DOI：10.1109/TPAMI.2016.2516982］

Kristan M，Pflugfelder R，Leonardis A，Matas J，Porikli F，Cehovin L，Nebehay G，Fernandez G，Vojir T，Gatt A，Khajenezhad A，Salahledin A，Soltani-Farani A，Zarezade A，Petrosino A，Milton A，Bozorgtabar B，Li B，Chan C S，Heng C，Ward D，Kearney D，Monekosso D，Karaimer H C，Rabiee H R，Zhu J K，Gao J，Xiao J J，Zhang J G，Xing J L，Huang K Q，Lebeda K，Cao L J，Maresca M E，Lim M K，El Helw M，Felsberg M，Remagnino P，

Bowden R, Goecke R, Stolkin R, Lim S Y, Maher S, Poullot S, Wong S, Satoh S, Chen W H, Hu W M, Zhang X Q, Li Y and Niu Z H. 2013. The visual object tracking VOT2013 challenge results// Proceedings of 2013 IEEE International Conference on Computer Vision Workshops. Sydney, Australia: IEEE: 98-111 [DOI: 10.1109/ICCVW.2013.20]

Lake B M, Salakhutdinov R and Tenenbaum J B. 2015. Human-level concept learning through probabilistic program induction. Science, 350(6266): 1332-1338 [DOI: 10.1126/science.aab3050]

Land M F and McLeod P. 2000. From eye movements to actions: how batsmen hit the ball. Nature Neuroscience, 3(12): 1340-1345 [DOI: 10.1038/81887]

Langlois T A, Zhao H C, Grant E, Dasgupta I, Griffiths T L and Jacoby N. 2021. Passive attention in artificial neural networks predicts human visual selectivity//Proceedings of the 35th Conference on Neural Information Processing Systems. Virtual: Curran Associates Inc.: 27094-27106

Lazebnik S, Schmid C and Ponce J. 2006. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE: 2169-2178 [DOI: 10.1109/CVPR.2006.68]

Li A N, Lin M, Wu Y, Yang M H and Yan S C. 2016. NUS-PRO: a new visual tracking challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(2): 335-349 [DOI: 10.1109/TPAMI.2015.2417577]

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 C L, Lu A D, Liu L and Tang J. 2023. Multi-modal visual tracking: a survey. Journal of Image and Graphics, 28(1): 37-56 (李成龙, 鹿安东, 刘磊, 汤进. 2023. 多模态视觉跟踪方法综述. 中国图象图形学报, 28(1): 37-56) [DOI: 10.11834/jig.220578]

Li F F, Fergus R and Perona P. 2006. One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4): 594-611 [DOI: 10.1109/TPAMI.2006.79]

Li S Y and Yeung D Y. 2017. Visual object tracking for unmanned aerial vehicles: a benchmark and new motion models//Proceedings of the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA: AAAI Press: 4140-4146 [DOI: 10.1609/aaai.v31i1.11205]

Li X, Zha Y F, Zhang T Z, Cui Z, Zuo W M, Hou Z Q, Lu H C and Wang H Z. 2019. Survey of visual object tracking algorithms based on deep learning. Journal of Image and Graphics, 24(12): 2057-2080 (李玺, 查宇飞, 张天柱, 崔振, 左旺孟, 侯志强, 卢湖川, 王菡子. 2019. 深度学习的目标跟踪算法综述. 中国图象图形学报, 24(12): 2057-2080) [DOI: 0.11834/jig,190372]

Liang P P, Blasch E and Ling H B. 2015. Encoding color information for visual tracking: algorithms and benchmark. IEEE Transactions on Image Processing, 24(12): 5630-5644 [DOI: 10.1109/TIP.2015.2482905]

Liang W X, Tadesse G A, Ho D, Li F F, Zaharia M, Zhang C and Zou J. 2022. Advances, challenges and opportunities in creating data for trustworthy AI. Nature Machine Intelligence, 4(8): 669-677 [DOI: 10.1038/s42256-022-00516-1]

Liang W X and Zou J. 2022. MetaShift: a dataset of datasets for evaluating contextual distribution shifts and training conflicts [EB/OL]. [2023-07-10]. http://arxiv.org/pdf/2202.06523.pdf

Lin L T, Fan H, Zhang Z P, Xu Y and Ling H B. 2022. SwinTrack: a simple and strong baseline for Transformer tracking [EB/OL]. [2023-07-10]. https://arxiv.org/pdf/2112.00995.pdf

Long G M and Penn D L. 1987. Dynamic visual acuity: normative functions and practical implications. Bulletin of the Psychonomic Society, 25(4): 253-256 [DOI: 10.3758/BF03330347]

Lu H C, Li P X and Wang D. 2018. Visual object tracking: a survey. Pattern Recognition and Artificial Intelligence, 31(1): 61-76 (卢湖川, 李佩霞, 王栋. 2018. 目标跟踪算法综述. 模式识别与人工智能, 31(1): 61-76) [DOI: 10.16451/j.cnki.issn1003-6059.201801006]

Luiten J, Voigtlaender P and Leibe B. 2019. PReMVOS: proposal-generation, refinement and merging for video object segmentation// Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: Springer: 565-580 [DOI: 10.1007/978-3-030-20870-7_35]

Lukezic A, Zajc L C, Vojir T, Matas J and Kristan M. 2021. Performance evaluation methodology for long-term single-object tracking. IEEE Transactions on Cybernetics, 51(12): 6305-6318 [DOI: 10.1109/TCYB.2020.2980618]

Ma C, Huang J B, Yang X K and Yang M H. 2015a. Hierarchical convolutional features for visual tracking//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 3074-3082 [DOI: 10.1109/ICCV.2015.352]

Ma C, Yang X K, Zhang C Y and Yang M H. 2015b. Long-term correlation tracking//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 5388-5396 [DOI: 10.1109/CVPR.2015.7299177]

Marr D. 2010. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Massachusetts, USA: The MIT Press

Marvasti-Zadeh S M, Cheng L, Ghanei-Yakhdan H and Kasaei S. 2022. Deep learning for visual tracking: a comprehensive survey. IEEE Transactions on Intelligent Transportation Systems, 23(5): 3943-3968 [DOI: 10.1109/TITS.2020.3046478]

Mayer C, Danelljan M, Pani Paudel D and van Gool L. 2021. Learning

target candidate association to keep track of what not to track//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 13424-13434 [DOI: 10.1109/ICCV48922.2021.01319]

Miller G A. 1995. WordNet: a lexical database for English. Communications of the ACM, 38(11): 39-41 [DOI: 10.1145/219717.219748]

Miller J W. 1958. Study of visual acuity during the ocular pursuit of moving test objects. II. Effects of direction of movement, relative movement, and illumination. Journal of the Optical Society of America, 48(11): 803-808 [DOI: 10.1364/josa.48.000803]

Miller J W and Ludvigh E. 1962. The effect of relative motion on visual acuity. Survey of Ophthalmology, 7: 83-116

Mueller 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. Munich, Germany: Springer: 310-327 [DOI: 10.1007/978-3-030-01246-5_19]

Nam H and Han B. 2016. Learning multi-domain convolutional neural networks for visual tracking//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4293-4302 [DOI: 10.1109/CVPR.2016.465]

Pylyshyn Z W and Storm R W. 1988. Tracking multiple independent targets: evidence for a parallel tracking mechanism. Spatial Vision, 3(3): 179-197 [DOI: 10.1163/156856888x00122]

Quevedo L, Aznar-Casanova J A and Da Silva J A. 2018. Dynamic visual acuity. Trends in Psychology, 26(3): 1283-1297 [DOI: 10.9788/TP2018.3-06En]

Quevedo L, Aznar-Casanova J A, Merindano-Encina D, Cardona G and Solé-Fortó J. 2012. A novel computer software for the evaluation of dynamic visual acuity. Journal of Optometry, 5(3): 131-138 [DOI: 10.1016/j.optom.2012.05.003]

Real E, Shlens J, Mazzocchi S, Pan X and Vanhoucke V. 2017. YouTube-BoundingBoxes: a large high-precision human annotated data set for object detection in video//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 7464-7473 [DOI: 10.1109/CVPR.2017.789]

Ross D A, Lim J, Lin R S and Yang M H. 2008. Incremental learning for robust visual tracking. International Journal of Computer Vision, 77(1): 125-141 [DOI: 10.1007/s11263-007-0075-7]

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S A, Huang Z H, Karpathy A, Khosla A, Bernstein M, Berg A C and Li F F. 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3): 211-252 [DOI: 10.1007/s11263-015-0816-y]

Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y T, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T and Hassabis D. 2017. Mastering the game of Go without human knowledge. Nature, 550(7676): 354-359 [DOI: 10.1038/nature24270]

Simonyan K and Zisserman A. 2015. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2023-07-10]. https://arxiv.org/pdf/1409.1556.pdf

Smeulders A W M, Chu D M, Cucchiara R, Calderara S, Dehghan A and Shah M. 2014. Visual tracking: an experimental survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(7): 1442-1468 [DOI: 10.1109/TPAMI.2013.230]

Sudderth E B, Torralba A, Freeman W T and Willsky A S. 2005. Learning hierarchical models of scenes, objects, and parts//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE: 1331-1338 [DOI: 10.1109/ICCV.2005.137]

Tian Z, Shen C H, Chen H and He T. 2019. FCOS: fully convolutional one-stage object detection//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 9626-9635 [DOI: 10.1109/ICCV.2019.00972]

Treisman A M and Gelade G. 1980. A feature-integration theory of attention. Cognitive Psychology, 12(1): 97-136 [DOI: 10.1016/0010-0285(80)90005-5]

Turing A M. 2009. Computing machinery and intelligence//Epstein R, Roberts G and Beber G, eds. Parsing the Turing Test. Dordrecht: Springer: 23-65 [DOI: 10.1007/978-1-4020-6710-5_3]

Valmadre J, Bertinetto L, Henriques J F, Tao R, Vedaldi A, Smeulders A W M, Torr P H S and Gavves E. 2018. Long-term tracking in the wild: a benchmark//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 692-707 [DOI: 10.1007/978-3-030-01219-9_41]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Voigtlaender P, Luiten J, Torr P H S and Leibe B. 2020. Siam R-CNN: visual tracking by re-detection//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6577-6587 [DOI: 10.1109/CVPR42600.2020.00661]

Wang D, Lu H C and Yang M H. 2013. Online object tracking with sparse prototypes. IEEE Transactions on Image Processing, 22(1): 314-325 [DOI: 10.1109/TIP.2012.2202677]

Wang D, Lu H C and Yang M H. 2016. Robust visual tracking via least soft-threshold squares. IEEE Transactions on Circuits and Systems for Video Technology, 26(9): 1709-1721 [DOI: 10.1109/TCSVT.2015.2462012]

Wang N, Zhou W G, Wang J and Li H Q. 2021. Transformer meets tracker: exploiting temporal context for robust visual tracking//Pro-

ceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1571-1580 [DOI: 10.1109/CVPR46437.2021.00162]

Wang Q, Gao J, Xing J L, Zhang M D and Hu W M. 2017a. DCFNet: discriminant correlation filters network for visual tracking [EB/OL]. [2023-07-10] https://arxiv.org/pdf/1704.04057.pdf

Wang X L, He K M and Gupta A. 2017b. Transitive invariance for self-supervised visual representation learning//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1338-1347 [DOI: 10.1109/ICCV.2017.149]

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]

Wu Y, Lim J and Yang M H. 2015. Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9): 1834-1848 [DOI: 10.1109/TPAMI.2014.2388226]

Xia C, Han J W and Zhang D W. 2021. Evaluation of saccadic scanpath prediction: subjective assessment database and recurrent neural network based metric. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(12): 4378-4395 [DOI: 10.1109/TPAMI.2020.3002168]

Xiang Y, Alahi A and Savarese S. 2015. Learning to track: online multi-object tracking by decision making//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 4705-4713 [DOI: 10.1109/ICCV.2015.534]

Xu N, Yang L J, Fan Y C, Yang J C, Yue D C, Liang Y C, Price B, Cohen S and Huang T. 2018. YouTube-VOS: sequence-to-sequence video object segmentation//Proceedings of the 15th European Conference on Computer Vision. Munich, Germany: Springer: 603-619 [DOI: 10.1007/978-3-030-01228-1_36]

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 34th AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press: 12549-12556 [DOI: 10.1609/aaai.v34i07.6944]

Yan B, Peng H W, Fu J L, Wang D and Lu H C. 2021. Learning spatio-temporal Transformer for visual tracking//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 10428-10437 [DOI: 10.1109/ICCV48922.2021.01028]

Yan B, Zhao H J, Wang D, Lu H C and Yang X Y. 2019. 'Skimming-perusal' tracking: a framework for real-time and robust long-term tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2385-2393 [DOI: 10.1109/ICCV.2019.00247]

Ye B T, Chang H, Ma B P, Shan S G and Chen X L. 2022. Joint feature learning and relation modeling for tracking: a one-stream framework//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 341-357 [DOI: 10.1007/978-3-031-20047-2_20]

Yu B, Tang M, Zheng L Y, Zhu G B, Wang J Q, Feng H, Feng X T and Lu H Q. 2021. High-performance discriminative tracking with Transformers//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 9836-9845 [DOI: 10.1109/ICCV48922.2021.00971]

Yu C S, Wang E M Y, Li W C and Braithwaite G. 2014. Pilots' visual scan patterns and situation awareness in flight operations. Aviation, Space, and Environmental Medicine, 85(7): 708-714 [DOI: 10.3357/asem.3847.2014]

Yu H Y, Li G R, Zhang W G, Huang Q M, Du D W, Tian Q and Sebe N. 2020. The unmanned aerial vehicle benchmark: object detection, tracking and baseline. International Journal of Computer Vision, 128(5): 1141-1159 [DOI: 10.1007/s11263-019-01266-1]

Yun S, Choi J, Yoo Y, Yun K and Choi J Y. 2017. Action-decision networks for visual tracking with deep reinforcement learning//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1349-1358 [DOI: 10.1109/CVPR.2017.148]

Zhang R, Isola P and Efros A A. 2016. Colorful image colorization//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 649-666 [DOI: 10.1007/978-3-319-46487-9_40]

Zhang T Z, Ghanem B, Liu S and Ahuja N. 2013. Robust visual tracking via structured multi-task sparse learning. International Journal of Computer Vision, 101(2): 367-383 [DOI: 10.1007/s11263-012-0582-z]

Zhang Z P and Peng H W. 2019. Deeper and wider siamese networks for real-time visual tracking//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4586-4595 [DOI: 10.1109/CVPR.2019.00472]

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 on Computer Vision. Glasgow, UK: Springer: 771-787 [DOI: 10.1007/978-3-030-58589-1_46]

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. Munich, Germany: Springer: 103-119 [DOI: 10.1007/978-3-030-01240-3_7]

## 作者简介

胡世宇，女，博士研究生，主要研究方向为计算机视觉、视频目标跟踪和认知神经科学。E-mail: hushiyu2019@ia.ac.cn

黄凯奇，通信作者，男，研究员，主要研究方向为计算机视觉、模式识别、视觉监控和认知决策。

E-mail: kqhuang@nlpr.ia.ac.cn

赵鑫，男，副研究员，主要研究方向为计算机视觉和模式识别。E-mail: xzhao@nlpr.ia.ac.cn