长时视觉目标跟踪前沿简介

刘畅，陈晓凡，薄纯娟，王栋

大连理工大学

一、 引言

视觉目标跟踪技术是计算机视觉任务中的一个重要研究课题，旨在给定第一帧的目标初始状态后，对后续帧中目标的位置、尺度等信息进行预测。其在日常生活中有着广泛的应用场景，是智能监控、智慧交通、自动驾驶等任务的基础。近年来，目标跟踪研究成果及公开评测数据集大多强调秒级到分钟级短时跟踪场景中算法的精确性和鲁棒性，而现实场景中分钟级甚至小时级长时视觉目标跟踪更贴近实际应用需求，是未来跟踪的重点研究方向。相对短时目标跟踪，长时跟踪中不仅存在更为复杂的姿态、形变、模糊、遮挡、光照、背景杂乱等困难，而且包含更多的视频帧数、目标消失再出现、显著的目标尺度及外观差异变化等更加严峻的挑战，需要其兼备精准快速跟踪和丢失判断，以及长时全图重找回的能力。部分研究表明[1][2][3]，短时跟踪器在较长的视频序列上的性能较差。究其原因是短时跟踪器在长时场景中，更容易因模板污染、误差累积造成跟踪漂移、甚至失败，同时在处理更为复杂的挑战因素和频繁的目标丢失重找回情况时，短时目标跟踪算法也更容易受到外观模型判别力不足和搜索区域的限制。2018年来，随着VOT竞赛增设长时目标跟踪赛道，以及相关长时跟踪数据集的提出[2][3][4]，长时目标跟踪领域备受关注，但长时目标跟踪的发展仍处于起步阶段。为此，本文从传统特征和深度特征两个角度，归纳总结了现有的长时目标跟踪算法，并在最后尝试对长时目标跟踪未来的发展方向进行探讨。更多资源请参考：https://github.com/wangdongdut/Long-term-Visual-Tracking

图 1 长时跟踪视频采样帧示意图，绿色框为目标真值，无框帧为目标不在视野

二、 长时视觉目标跟踪研究现状

1. 基于传统特征的长时目标跟踪算法

Kalal 等人的 TLD 算法[5]提出“局部跟踪-全局检测”的长时跟踪范式，该算法利用前向-反向光流匹配作为局部跟踪器，采用基于集成学习的在线检测器来在全图范围内进行目标重检测，并使用 P-N 学习模块辅助跟踪和检测模块的在线更新。Ma 等人[6]使用基于 HOG 特征的判别式相关滤波器作为局部跟踪，并使用在线随机蕨分类器实现全局检测。Nebehay 等人[7]使用局部和全局的关键点匹配实现长时目标跟踪。Hong 等人[8]使用基于 HOG 和 SIFT 特征的集成相关滤波跟踪器作为局部跟踪模块，并利用长短时分离的目标历史样本存储结合关键点匹配实现全图检测。FuCoLoT 算法[9]尝试将搜索区域尺度与目标尺度估计解耦，设计使用 ADMM 优化的基于 HOG 和 Color Names 特征的相关滤波跟踪器作为局部和全图的跟踪和检测，避免了采用多滑动窗搜索的时间消耗，在速度上具有显著优越性。

另一类工作尝试每一帧在全图中检测最佳目标正样本，将局部跟踪简化为辅助时序约束项。Zhu 等人[10]使用基于边缘的检测算法在全图中获取目标候选框，再使用在线学习 SVM 分类器为候选框打分，结合时序约束获得连续的目标跟踪框输出。文献[11]使用多个基于关键点的方法在全图中进行检测以获得待跟踪目标位置，并采用基于“生长-剪枝”的更新样本选择方法以实现更

好的检测效果。文献[12]使用基于 HOG 特征的 SVM 分类器作为检测手段获得目标候选区域，并使用动态规划求解最短路径问题从而实现跟踪，同时使用基于自步学习的数据选择策略对检测器进行重新学习。总体说来，以上算法尽管能够处理简单目标和场景的长时跟踪问题，但由于传统手工特征表达能力的局限性，这些算法处理复杂真实环境中长时跟踪问题的能力有限。

2．基于深度特征的长时目标跟踪算法

深度学习的迅猛发展为计算机视觉的各个领域带来了新的发展动力。Andrew G 等人提出的 MobileNet 网络以及 He 等人提出的 ResNet 网络等基础分类网络展现了强大的特征提取能力，被应用于各类视觉任务中。Martin 等人在相关滤波算法中使用深度卷积网络作为特征提取器，获得了明显的性能提升，但深度特征提取手段和传统外观建模手段间的耦合程度不足，限制了性能的提升空间。Bertinetto 等人提出了结构简单的端到端深度孪生卷积网络 SiamFC[13]，该方法不仅跟踪性能优秀，同时具有更快的处理速度，成为短时目标跟踪领域的主流框架之一。优秀的深度特征跟踪器被部署在长时跟踪算法中补充或取代传统特征跟踪器。Fan 等人[14]提出基于 fDSST[15]局部跟踪器和一个辅助更新的验证网络的方法，尽管局部跟踪器延续了传统特征，其验证网络采用的是基于深度特征的孪生卷积结构。Valmadre 等人[4]和 Lee[16]等人均使用深度孪生网络 SiamFC[13]作为局部跟踪器，且后者通过离线训练的网络生成稀疏的候选区域并计算相似度。Wu 等人[17]有效结合离线深度孪生网络 SiamRPN 和在线更新深度分类网络 MDNet 进行局部跟踪和全局检测的候选目标判别。Zhang 等人提出的 MBMD[18]算法利用一个离线训练的回归网络在局部区域内直接回归出待跟踪目标的包围框，利用滑动窗搜索的方式重复利用局部回归网络来实现目标消失后的重找回，并利用一个在线学习的验证器来实现跟踪器在局部跟踪与全局检测之间的切换，该算法获得了 VOT2018 首届长时比赛冠军。Yan 等人[19]采用深度孪生网络 SiamRPN 作为局部跟踪器，同时设计了“略读-精读”模型的长时跟踪算法框架（如图 2 所示），该框架提出一个轻量级的“略读”模块来减少滑窗数量，从而显著提升了全局检测模块的速度。Zhu 等人[20]则部署拓展短时跟踪器并使用其鲁棒性得分作为全图范围与局部范围搜索的标志。Dai 等人[21]使用能够在线学习的深度跟踪算法作为局部跟踪模块，并提出元更新模块从时空多线索信息中预测更新可靠性（如图 3 所示），有效降低累积噪声，显著提升了算法的精度，并获得了 VOT2019 和 VOT2020 两届长时比赛冠军。RLTDiMP[22]算法中同样采用在线学习的深度跟踪算法作为局部跟踪器，并使用背景增强促使在线更新过程中进行更具判别性的特征学习，并使用随机滑窗减少时间消耗。

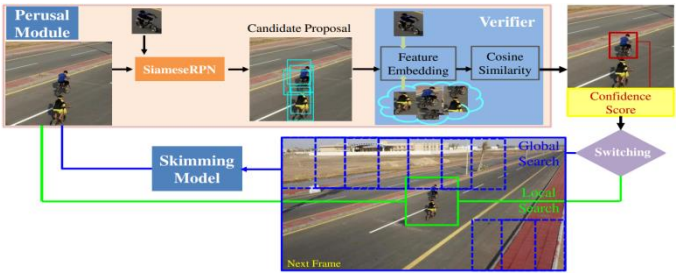

图 2 基于“略读-精读”框架的 SPLT 长时跟踪算法

（本图来自文献[19]）

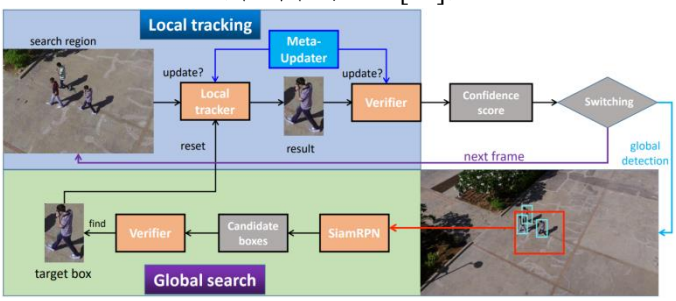

图 3 基于元更新器模型的 LTMU 长时跟踪算法

（本图来自文献[21]）

此外，一些工作将长时目标跟踪视为对于特定目标的检测任务，以基于深度卷积网络的目标检测器为基础，结合孪生卷积网络结构进行改进，充分吸收了深度学习在目标检测领域的发展。Paul 等人[23]基于 Faster RCNN[24]设计了双阶段的孪生卷积网络，改造的 Re-detection 模块以模板和搜索区域中候选框的特征作为输入，进行相似度估计和包围框回归，并基于此设计了精巧的轨迹动态规划来有效抑制干扰，从而提升跟踪的准确率。Dave 等人[25]基于 Mask RCNN 结构将特定类别的目标检测器转换为特定对象的目标检测器，并提出了一个计算具有判别力目标模板的端到端轻量策略，能够有效处理跟踪中的干扰。Huang 等人[26]也采取了相似的方法，基于 Faster RCNN 构造了一个针对特定对象的目标检测器（如图 4 所示），目标模板特征使用 ROI Align 进行提取，同时设计网络使得其学习如何使用模板特征

编码搜索区域的特征来获得更好的特定目标检测效果。Janghoon 等人[27]同样采用了与前述工作相似的思路，但借鉴了目标检测中的 Anchor-free 结构来替代 Anchor-based 结构，完成搜索区域候选框与模板的粗略匹配，并添加了使用上下文嵌入信息学习的精细匹配阶段提升分辨干扰物与跟踪目标的能力，同时其减小全图输入的分辨率达到了速度和准确率的较好平衡。Li 等人[28]同样通过 Faster RCNN[24]构造针对特定对象的目标检测器来获取候选目标区域，但使用运动模型预测定位分布感知时序信息，将潜在干扰物剔除以达到更准确的跟踪效果。

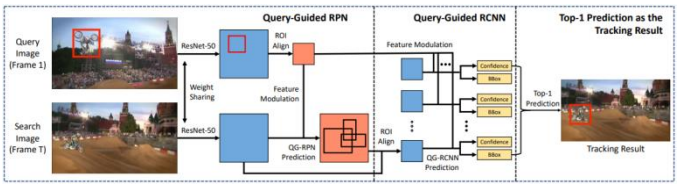

图4 基于特定目标检测的GlobalTrack长时跟踪算法（本图来自文献[26]）

3．长时目标跟踪数据集

长时目标跟踪数据集相较短时数据集，不仅具有更多的视频平均帧数，同时具有较多的目标丢失和重新出现场景，跟踪挑战也更加复杂多样。Muller 等人[1]在 2016 年提出了无人机视角下的 20 个长时目标跟踪视频序列。2018 年，OxUvA[4]、TLP[2]、LTB[3]等长时数据集及相对应的评价标准相继推出，极大丰富了长时目标跟踪的评测体系。OxUvA 数据集[4]包含 366 段轨迹，包括 dev 部分 200 段和 test 部分 160 段，平均时长 2.4 分钟采用 TPR、TNR、MaxGM 作为评价标准。TLP[2]包含真实场景下的 50 段视频，平均时长超过 8 分钟，使用 Success、Precision 作为评价标准。VOT 竞赛的长时目标跟踪赛道继承了[3]工作的数据集及评价标准，进一步推动了长时跟踪领域的发展。VOTLT2018 数据集包含 35 段长时视频，VOTLT2020 数据集与 VOTLT2019 数据库相同，包含 50 段长时视频，使用 Precision、Recall、F-score 作为评价标准。更多介绍和对比请参考：https://github.com/wangdongdut/Long-term-Visual-Tracking

4．长时目标跟踪实验结果

表格 1 和 2 简要汇总了当前长时跟踪算法的性能，排名前列的跟踪器均为使用深度特征的跟踪器。从表中看出，LTMU 和 Siam R-CNN 算法具有显著的精度有效，但即使在 GPU 环境下，前者的运行速度不到 15 帧/秒，后者不到 5 帧/秒。

表 1 VOTLT2019(2020) 及 OxUvA 数据集实验结果

Dataset	VOTLT2019(2020)			Dataset	OxUvA		
Tracker	F-score (↑)	Precision (↑)	Recall (↑)	Tracker	MaxGM (↑)	TPR (↑)	TNR (↑)
LTMU [21]	0.697	0.721	0.674	LTMU [21]	0.751	0.749	0.754
LT_DSE	0.695	0.715	0.677	Siam R-CNN [23]	0.723	0.701	0.745
Megtrack	0.687	0.703	0.671	LTA [25]	0.716	0.655	0.782
CLGS	0.674	0.739	0.619	TACT [27]	0.709	0.809	0.622
RLTDiMP [22]	0.670	0.657	0.684	SPLT [19]	0.622	0.498	0.776
Siam R-CNN [23]	0.670	0.658	0.676	GlobalTrack [26]	0.603	0.574	0.633
SiamDW-LT	0.656	0.678	0.635	MBMD [18]	0.544	0.609	0.485
TACT [27]	0.569	0.578	0.561	SiamFC+R [4]	0.454	0.427	0.481
MBMD [18]	0.575	0.623	0.534	TLD [5]	0.431	0.208	0.895
SPLT [19]	0.565	0.587	0.544	LCT [6]	0.396	0.292	0.537
GlobalTrack [26]	0.536	0.565	0.510	EBT [10]	0.283	0.321	0
FuCoLoT [9]	0.411	0.507	0.346				

三、 结束语

长时视觉目标跟踪仍然是一个充满挑战性的研究方向，其发展仍处于起步阶段。本文从传统特征和深度特征两个方向梳理了现有的长时目标跟踪算法，并整理了长时目标跟踪数据集，可以发现该领域虽然已经显现了许多优秀的算法，但仍存在一些亟待解决的问题。首先，在长时目标跟踪中，如何减弱误差累积及模板污染仍然是一个重要研究方向。其次，对于目标丢失及重检测的处理仍然是制约长时目标跟踪性能的重要因素。再次，现有的前沿长时跟踪算法速度较慢，许多算法都无法达到实时性能，不适于今后的实际应用需求，距离应用部署仍有较广的发展空间。

表 2　TLP 数据集实验结果

Tracker	Success (↑)	Precision (↑)
Siam R-CNN [23]	0.601	0.630
LTMU [21]	0.571	0.608
SuperDiMP	0.562	0.580
RLTDiMP [22]	0.528	0.533
TACT [27]	0.523	0.545
GlobalTrack [26]	0.520	0.556
MBMD [18]	0.492	0.502
SPLT [19]	0.416	0.403
TLD [5]	0.122	0.116
LCT [6]	0.101	0.071

【参考文献】

[1] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. *Far East Journal of Mathematical Sciences*, 2(2):445-461, 2016.

[2] Abhinav Moudgil and Vineet Gandhi. Long-term visual object tracking benchmark. *Asian Conference on Computer Vision*, pages 629-645, 2018.

[3] A Lukežič and LČ Zajc and T Vojíř and Matas, J. and Kristan, M. Now you see me: evaluating performance in long-term visual tracking. *arXiv preprint arXiv:1804.07056*, 2018.

[4] Jack Valmadre, Luca Bertinetto, Joao F. Henriques, Ran Tao, Andrea Vedaldi, Arnold W.M. Smeulders, Philip H.S. Torr, and Efstratios Gavves. Long-term tracking in the wild: a benchmark. *European Conference on Computer Vision*, pages 692-707, 2018.

[5] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(7):1409-1422, 2011.

[6] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming-Hsuan Yang. Long-term correlation tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5388-5396, 2015.

[7] Georg Nebehay and Roman Pflugfelder. Clustering of static-adaptive correspondences for deformable object tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2784-2791, 2015.

[8] Zhibin Hong, Zhe Chen, Chaohui Wang, Xue Mei, Danil Prokhorov, and Dacheng Tao. Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 749-758, 2015.

[9] A Lukežič and LČ Zajc and T Vojíř and Matas, J. and Kristan, M. Fucolot - a fully-correlational long-term tracker. *Asian Conference on Computer Vision*, pages 595-611, 2018.

[10] Gao Zhu, Fatih Porikli, and Hongdong Li. Beyond local search: Tracking objects everywhere with instance-specific proposals. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 943-951, 2016.

[11] Mario Edoardo Maresca and Alfredo Petrosino. Matrioska: A multi-level approach to fast tracking by learning. *International Conference on Image Analysis and Processing*, pages 419-428, 2013.

[12] James Steven Supancic and Deva Ramanan. Self-paced learning for long-term tracking. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2379-2386, 2013.

[13] Luca Bertinetto, Jack Valmadre, Jo~ao F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. *European Conference on Computer Vision*, pages 850-865, 2016.

[14] Heng Fan and Haibin Ling. Parallel tracking and verifying. *IEEE Transactions on Image Processing*, 28(8):4130-4144, 2019.

[15] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative scale space tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1561-1575, 2017.

[16] Hankyeol Lee, Seokeon Choi, and Changick Kim. A memory model based on the siamese network for long-term tracking. *European Conference on Computer Vision Workshops*, pages 100-115, 2018.

[17] Han Wu, Xueyuan Yang, Yong Yang, and Guizhong Liu. Flow guided short-term trackers with cascade detection for long-term tracking.

IEEE/CVF International Conference on Computer Vision Workshops, pages 170-178, 2019.

[18] Yunhua Zhang, Dong Wang, Lijun Wang, Jinqing Qi, and Huchuan Lu. Learning regression and verification networks for long-term visual tracking. *arXiv preprint arXiv:1809.04320*, 2018.

[19] Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'Skimming-Perusal' Tracking: A framework for real-time and robust longterm tracking. *IEEE/CVF International Conference on Computer Vision*, pages 2385-2393, 2019.

[20] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. *European Conference on Computer Vision*, pages 101-117, 2018.

[21] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6298-6307, 2020.

[22] Seokeon Choi, Junhyun Lee, Yunsung Lee, and Alexander Hauptmann. Robust long-term object tracking via improved discriminative model prediction. *European Conference on Computer Vision Workshops*, pages 602-617, 2020.

[23] Paul Voigtlaender, Jonathon Luiten, Philip H.S. Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6578-6588, 2020.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137-1149, 2017.

[25] Achal Dave, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Learning to track any object. *arXiv preprint arXiv:1910.11844*, 2019.

[26] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. *AAAI Conference on Artificial Intelligence*, pages 11037-11044, 2020.

[27] Janghoon Choi, Junseok Kwon, and Kyoung Mu Lee. Visual tracking by tridentalign and context embedding. *Asian Conference on Computer Vision*, 2020.

[28] Zhenbang Li, Qiang Wang, Jin Gao, Bing Li, and Weiming Hu. Globally spatial-temporal perception: a long-term tracking system. *IEEE International Conference on Image Processing*, pages 2066-2070, 2020.