[['UniSOT: A Unified Framework for Multi-Modality Single Object Tracking',
  'http://ieeexplore.ieee.org/document/11202681',
  'Single object tracking aims to localize target object with specific reference modalities (bounding box, natural language or both) in a sequence of specific video modalities (RGB, RGB+Depth, RGB+Thermal or RGB+Event.). Different reference modalities enable various human-machine interactions, and different video modalities are demanded in complex scenarios to enhance tracking robustness. Existing trackers are designed for single or several video modalities with single or several reference modalities, which leads to separate model designs and limits practical applications. Practically, a unified tracker is needed to handle various requirements. To the best of our knowledge, there is still no tracker that can perform tracking with these above reference modalities across these video modalities simultaneously. Thus, in this paper, we present a unified tracker, UniSOT, for different combinations of three reference modalities and four video modalities with uniform parameters. Extensive experimental results on 18 visual tracking, vision-language tracking and RGB+X tracking benchmarks demonstrate that UniSOT shows superior performance against modality-specific counterparts. Notably, UniSOT outperforms previous counterparts by over 3.0% AUC on TNL2K across all three reference modalities and outperforms Un-Track by over 2.0% main metric across all three RGB+X video modalities.',
  'Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang, Xu Zhou, Feng Wu',
  '3',
  '2026',
  '1']]

[['TOTNet: Occlusion-aware temporal tracking for robust ball detection in sports videos',

  'https://www.sciencedirect.com/science/article/pii/S107731422600024X',
  '',
  'Hao Xu, Arbind Agrahari Baniya, Sam Wells, Mohamed Reda Bouadjenek, Richard Dazeley, Sunil Aryal',
  '1',
  '2026',
  '2'],
 ['Multimodal transformer–diffusion framework for large-scale reconstruction of soccer tracking data',
  'https://www.sciencedirect.com/science/article/pii/S1077314225003492',
  '',
  'Harry Hughes, Patrick Lucey, Michael Horton, Harshala Gammulle, Clinton Fookes, Sridha Sridharan',
  '1',
  '2026',
  '2'],
 ['Distractor suppression Siamese network with task-aware attention for visual tracking',
  'https://www.sciencedirect.com/science/article/pii/S1077314225003303',
  '',
  'Zhigang Liu, Fuyuan Xing, Hao Huang, Kexin Wang, Yuxuan Shao',
  '1',
  '2026',
  '1'],
 ['Transformer tracking with high-low frequency attention',
  'https://www.sciencedirect.com/science/article/pii/S1077314225002863',
  '',
  'Zhi Chen, Zhen Yu',
  '1',
  '2026',
  '1'],
 ['HiViTrack: Hierarchical vision transformer with efficient target-prompt update for visual object tracking',
  'https://www.sciencedirect.com/science/article/pii/S0031320325016553',

```
  '',
  'Yang Fang, Yujie Hu, Bailian Xie, Yujie Wang, Zongyi Xu, Weisheng Li, Xinbo
Gao',
  '2',
  '2026',
  '7'],
 ['LTSTrack: Visual tracking with long-term temporal sequence',
  'https://www.sciencedirect.com/science/article/pii/S0031320326000154',
  '',
  'Zhaochuan Zeng, Shilei Wang, Yidong Song, Zhenhua Wang, Jifeng Ning',
  '2',
  '2026',
  '7'],
 ['A unified spatial-spectral-temporal network for hyperspectral object
tracking',
  'https://www.sciencedirect.com/science/article/pii/S0031320325016681',
  '',
  'Zhuanfeng Li, Jing Wang, Jue Zhang, Dong Zhao, Guanyiman Fu, Jiangtao Wang,
Jianfeng Lu',
  '2',
  '2026',
  '6'],
 ['UniSOT: A Unified Framework for Multi-Modality Single Object Tracking',
  'http://ieeexplore.ieee.org/document/11202681',
  'Single object tracking aims to localize target object with specific reference
modalities (bounding box, natural language or both) in a sequence of specific
video modalities (RGB, RGB+Depth, RGB+Thermal or RGB+Event.). Different
reference modalities enable various human-machine interactions, and different
video modalities are demanded in complex scenarios to enhance tracking
robustness. Existing trackers are designed for single or several video
modalities with single or several reference modalities, which leads to separate
model designs and limits practical applications. Practically, a unified tracker
is needed to handle various requirements. To the best of our knowledge, there is
still no tracker that can perform tracking with these above reference modalities
across these video modalities simultaneously. Thus, in this paper, we present a
unified tracker, UniSOT, for different combinations of three reference
modalities and four video modalities with uniform parameters. Extensive
experimental results on 18 visual tracking, vision-language tracking and RGB+X
tracking benchmarks demonstrate that UniSOT shows superior performance against
modality-specific counterparts. Notably, UniSOT outperforms previous
counterparts by over 3.0% AUC on TNL2K across all three reference modalities and
outperforms Un-Track by over 2.0% main metric across all three RGB+X video
modalities.',
  'Yinchao Ma, Yuyang Tang, Wenfei Yang, Tianzhu Zhang, Xu Zhou, Feng Wu',
  '3',
  '2026',
  '1'],
 ['Toward an Effective Action-Region Tracking Framework for Fine-Grained Video
Action Recognition',
  'http://ieeexplore.ieee.org/document/11173836',
  'Fine-grained action recognition (FGAR) aims to identify subtle and
distinctive differences among fine-grained action categories. However, current
recognition methods often capture coarse-grained motion patterns but struggle to
identify subtle details in local regions evolving over time. In this work, we
introduce the action-region tracking (ART) framework, a novel solution
leveraging a query-response mechanism to discover and track the dynamics of
distinctive local details, enabling distinguishing similar actions effectively.
Specifically, we propose a region-specific semantic activation module that
employs discriminative and text-constrained semantics serve as queries to
capture the most action-related region responses in each video frame,
facilitating interaction among spatial and temporal dimensions with
corresponding video features. The captured region responses are then organized
into action tracklets, which characterize the region-based action dynamics by
linking related responses across different video frames in a coherent sequence.
```

The text-constrained queries are designed to expressly encode nuanced semantic representations derived from the textual descriptions of action labels, as extracted by the language branches within visual language models. To optimize generated action tracklets, we design a multilevel tracklet contrastive constraint among multiple region responses at spatial and temporal levels, which can effectively distinguish individual region responses in each video frame (spatial level) and establish the correlation of similar region responses between adjacent video frames (temporal level). In addition, we implement a task-specific fine-tuning mechanism to refine textual semantics during training. This ensures that the semantic representations encoded by vision language models (VLMs) are not only preserved but also optimized for specific task preferences. Comprehensive experiments on several widely used action recognition benchmarks, i.e., FineGym, Diving48, NTURGB-D, Kinetics, and Something-Something, clearly demonstrate the superiority to previous state-of-the-art baselines.',
  'Baoli Sun, Yihan Wang, Xinzhu Ma, Zhihui Wang, Kun Lu, Zhiyong Wang',
  '5',
  '2026',
  '1'],
 ['Nesterov Accelerated Gradient Tracking With Adam for Distributed Online Optimization',
  'http://ieeexplore.ieee.org/document/11147187',
  'This article presents an accelerated distributed optimization algorithm for online optimization problems over large-scale networks. The proposed algorithm’s iteration only relies on local computation and communication. To effectively adapt to dynamic changes and achieve a fast convergence rate while maintaining good convergence performance, we design a new algorithm called NGTAdam. This algorithm combines the Nesterov acceleration technique with an adaptive moment estimation method. The convergence of NGTAdam is evaluated by evaluating its dynamic regret through the use of linear system inequality. For online convex optimization problems, we provide an upper bound on the dynamic regret of NGTAdam, which depends on the initial conditions and the time-varying nature of the optimization problem. Moreover, we show that if the time-varying part of this upper bound is sublinear with time, the dynamic regret is also sublinear. Through a variety of numerical experiments, we demonstrate that NGTAdam outperforms state-of-the-art distributed online optimization algorithms.',
  'Yanxu Su, Qingyang Sheng, Xiasheng Shi, Chaoxu Mu, Changyin Sun',
  '5',
  '2026',
  '1'],
 ['Virtual Target-Oriented Neural Learning for Robust Optimal Tracking Control of Discrete Strict-Feedback Systems',
  'http://ieeexplore.ieee.org/document/11180916',
  'This article proposes a hierarchical neural learning (HNL) algorithm for optimal tracking control (OTC) of nonlinear strict-feedback systems (SFSs) with unmatched disturbances (uMDs) and unknown dynamics. Leveraging the recursive structure of SFSs, we introduce the virtual target (VT) construction scheme in which each VT is a nonlinear mapping of the current state and desired output, thereby eliminating the noncausal that typically plagues discrete-time SFS control. The VTs serve as auxiliary inputs for low-order subsystems, while a time-varying affine Hamilton–Jacobi–Isaacs (HJI) formulation establishes an explicit relationship between the auxiliary control and the disturbance. The controller is synthesized directly from input–output data, removing the need for an accurate plant model. Within an adaptive dynamic programming (ADP) framework, we further enhance the neural architecture by replacing the conventional action network with a tracking network (T-network) whose energy function merges gradient information with future tracking errors, ensuring that each policy update simultaneously reduces control effort and improves tracking accuracy. Simulations confirm that the proposed HNL scheme achieves outstanding performance in both (optimal) tracking modes, exhibiting strong robustness to uMDs and significant model uncertainties.',
  'Ying Yan, Huaguang Zhang, Jiayue Sun, Zhongyang Ming',
  '5',
  '2026',
  '1'],

['Adaptive Modality Balanced Online Knowledge Distillation for Brain-Eye-Computer-Based Dim Object Detection',
'http://ieeexplore.ieee.org/document/11164371',
'Advanced cognition can be measured from the human brain using brain–computer interfaces (BCIs). Integrating these interfaces with computer vision techniques, which possess efficient feature extraction capabilities, can achieve more robust and accurate detection of dim targets in aerial images. However, existing target detection methods primarily concentrate on homogeneous data, lacking efficient and versatile processing capabilities for heterogeneous multimodal data. In this article, we first build a brain–eye–computer-based object detection system for aerial images under few-shot conditions. This system detects suspicious targets using region proposal networks (RPNs), evokes the event-related potential (ERP) signal in electroencephalogram (EEG) through the eye-tracking-based slow serial visual presentation (ESSVP) paradigm, and constructs the EEG–image data pairs with eye movement data. Then, an adaptive modality balanced online knowledge distillation (AMBOKD) method is proposed to recognize dim objects with the EEG-image data. AMBOKD fuses EEG and image features using a multihead attention module, establishing a new modality with comprehensive features. To enhance the performance and robust capability of the fusion modality, simultaneous training and mutual learning between modalities are enabled by end-to-end online KD (OKD). During the learning process, an adaptive modality balancing module is proposed to ensure multimodal equilibrium by dynamically adjusting the weights of the importance and the training gradients across various modalities. The effectiveness and superiority of our method are demonstrated by comparing it with existing state-of-the-art methods. Additionally, experiments conducted on public datasets and real-world scenarios demonstrate the reliability and practicality of the proposed system and the designed method. The dataset and the source code can be found at: https://github.com/lizixing23/AMBOKD',
'Zixing Li, Chao Yan, Zhen Lan, Xiaojia Xiang, Han Zhou, Jun Lai, Dengqing Tang',
'5',
'2026',
'1']]