The Tenth Visual Object Tracking VOT2022 Challenge Results

Matej Kristan[1], Aleš Leonardis[2], Jiří Matas[3], Michael Felsberg[4], Roman Pflugfelder[5,6,7,8], Joni-Kristian Kämäräinen[9], Hyung Jin Chang[2], Martin Danelljan[10], Luka Čehovin Zajc[1], Alan Lukežič[1], Ondrej Drbohlav[3], Johanna Björklund[11], Yushan Zhang[4], Zhongqun Zhang[2], Song Yan[9], Wenyan Yang[9], Dingding Cai[9], Christoph Mayer[10], Gustavo Fernández[5], Kang Ben[18], Goutam Bhat[10], Hong Chang[24], Guangqi Chen[16], Jiaye Chen[26], Shengyong Chen[43], Xilin Chen[24], Xin Chen[18], Xiuyi Chen[13], Yiwei Chen[35], Yu-Hsi Chen[12], Zhixing Chen[16], Yangming Cheng[55], Angelo Ciaramella[47], Yutao Cui[30], Benjamin Džubur[1], Mohana Murali Dasari[22], Qili Deng[16], Debajyoti Dhar[39], Shangzhe Di[14], Emanuel Di Nardo[46,47], Daniel K. Du[16], Matteo Dunnhofer[51], Heng Fan[48], Zhenhua Feng[50], Zhihong Fu[16], Shang Gao[41], Rama Krishna Gorthi[22], Eric Granger[27], Q.H. Gu[15], Himanshu Gupta[19], Jianfeng He[49], Keji He[13], Yan Huang[13], Deepak Jangid[19], Rongrong Ji[53], Cheng Jiang[30], Yingjie Jiang[26], Felix Järemo Lawin[4], Ze Kang[26], Madhu Kiran[27], Josef Kittler[50], Simiao Lai[18], Xiangyuan Lan[32], Dongwook Lee[34], Hyunjeong Lee[34], Seohyung Lee[34], Hui Li[26], Ming Li[17], Wangkai Li[49], Xi Li[55], Xianxian Li[20], Xiao Li[16], Zhe Li[41], Liting Lin[37], Haibin Ling[40], Bo Liu[25], Chang Liu[18], Si Liu[23], Huchuan Lu[18], Rafael M. O. Cruz[27], Bingpeng Ma[44], Chao Ma[36], Jie Ma[21], Yinchao Ma[49], Niki Martinel[51], Alireza Memarmoghadam[45], Christian Micheloni[51], Payman Moallem[45], Le Thanh Nguyen-Meidine[27], Siyang Pan[35], ChangBeom Park[34], Danda Paudel[10], Matthieu Paul[10], Houwen Peng[28], Andreas Robinson[4], Litu Rout[39], Shiguang Shan[24], Kristian Simonato[51], Tianhui Song[30], Xiaoning Song[26], Chao Sun[55], Jingna Sun[16], Zhangyong Tang[26], Radu Timofte[10,52], Chi-Yi Tsai[42], Luc Van Gool[10], Om Prakash Verma[19], Dong Wang[18], Fei Wang[49], Liang Wang[13], Liangliang Wang[16], Lijun Wang[18], Limin Wang[30], Qiang Wang[35], Gangshan Wu[30], Jinlin Wu[13], Xiaojun Wu[26], Fei Xie[38], Tianyang Xu[26], Wei Xu[16], Yong Xu[37], Yuanyou Xu[55], Wanli Xue[43], Zizheng Xun[14], Bin Yan[18], Dawei Yang[49], Jinyu Yang[41], Wankou Yang[38], Xiaoyun Yang[33], Yi Yang[55], Yichun Yang[30], Zongxin Yang[55], Botao Ye[24], Fisher Yu[10], Hongyuan Yu[13], Jiaqian Yu[35], Qianjin Yu[49], Weichen Yu[13], Kang Ze[26], Jiang Zhai[38], Chengwei Zhang[17], Chunhu Zhang[36], Kaihua Zhang[29], Tianzhu Zhang[49], Wenkang Zhang[38], Zhibin Zhang[43], Zhipeng Zhang[31], Jie Zhao[18], Shaochuan Zhao[26], Feng Zheng[41], Haixia Zheng[54], Min Zheng[16], Bineng Zhong[20], Jiawen Zhu[18], Xuefeng Zhu[26], and Yueting Zhuang[55]

[1] University of Ljubljana, Slovenia
[2] University of Birmingham, United Kingdom
[3] Czech Technical University, Czech Republic
[4] Linköping University, Sweden
[5] Austrian Institute of Technology, Austria
[6] TU Vienna, Austria

[7] TU Munich, Germany
[8] Technion Israel Institute of Technology, Israel
[9] Tampere University, Finland
[10] ETH Zurich, Switzerland
[11] Umeå University, Sweden
[12] Academia Sinica, Taiwan
[13] AI School, China
[14] Beihang University, China
[15] Beijing Jiaotong University, China
[16] ByteDance, China
[17] Dalian Maritime University, China
[18] Dalian University of Technology, China
[19] Dr B R Ambedkar National Institute of Technology Jalandhar, India
[20] Guangxi Normal University, China
[21] Huaqiao University, China
[22] Indian Institute of Technology, India
[23] Institute of Artificial Intelligence, China
[24] Institute of Computing Technology, Chinese Academy of Sciences, China
[25] JD Finance America Corporation, United States of America
[26] Jiangnan University, China
[27] LIVIA-École de technologie supérieure, Canada
[28] Microsoft Research Asia, China
[29] Nanjing University of Information Science and Technology, China
[30] Nanjing University, China
[31] NLP, China
[32] Peng Cheng Laboratory, China
[33] Remark AI, China
[34] Samsung Advanced Institute of Technology (SAIT), Korea
[35] Samsung R&D Institute China Beijing (SRCB), China
[36] Shanghai Jiao Tong University, China
[37] South China University of Technology, China
[38] Southeast University, China
[39] Space Applications Centre, India
[40] Stony Brook University, United States of America
[41] Sustech, China
[42] Tamkang University, Taiwan
[43] Tianjin University of Technology, China
[44] University of Chinese Academy of Sciences, China
[45] University of Isfahan, Iran
[46] University of Milan, Italy
[47] University of Naples Parthenope, Italy
[48] University of North Texas, United States of America
[49] University of Science and Technology of China, China
[50] University of Surrey, United Kingdom
[51] University of Udine, Italy
[52] University of Wurzburg, Germany
[53] Xiamen University, China
[54] Xian Jiaotong University, China
[55] Zhejiang University, China
matej.kristan@fri.uni-lj.si

Abstract. The Visual Object Tracking challenge VOT2022 is the tenth annual tracker benchmarking activity organized by the VOT initiative. Results of 93 entries are presented; many are state-of-the-art trackers published at major computer vision conferences or in journals in recent years. The VOT2022 challenge was composed of seven sub-challenges focusing on different tracking domains: (i) VOT-STs2022 challenge focused on short-term tracking in RGB by segmentation, (ii) VOT-STb2022 challenge focused on short-term tracking in RGB by bounding boxes, (iii) VOT-RTs2022 challenge focused on "real-time" short-term tracking in RGB by segmentation, (iv) VOT-RTb2022 challenge focused on "real-time" short-term tracking in RGB by bounding boxes, (v) VOT-LT2022 focused on long-term tracking, namely coping with target disappearance and reappearance, (vi) VOT-RGBD2022 challenge focused on short-term tracking in RGB and depth imagery, and (vii) VOT-D2022 challenge focused on short-term tracking in depth-only imagery. New datasets were introduced in VOT-LT2022 and VOT-RGBD2022, VOT-ST2022 dataset was refreshed, and a training dataset was introduced for VOT-LT2022. The source code for most of the trackers, the datasets, the evaluation kit and the results are publicly available at the challenge website[56].

Keywords: Visual object tracking challenge, VOT, short-term tracking, long-term tracking, performance evaluation.

1 Introduction

A decade ago, the Visual Object Tracking (VOT) initiative was founded in response to the lack of standardised performance evaluation in visual object tracking. To facilitate the development of this highly active computer vision field, the first VOT2013 challenge [35] was organized in conjunction with ICCV2013. Encouraged by the strong interest of the emerging community, eight VOT challenges have been organized since, with the results presented at the accompanying workshops at major computer vision conferences: ECCV2014 (VOT2014 [36]), ICCV2015 (VOT2015 [34]), ECCV2016 (VOT2016 [32]), ICCV2017 (VOT2017 [31]), ECCV2018 (VOT2018 [30]), ICCV2019 (VOT2019 [28]), ECCV2020 (VOT2020 [29]), ICCV2021 (VOT2021 [33]). The VOT challenge is now the main annual tracking performance evaluation event in computer vision.

The primary mission of the VOT initiative has been the promotion of the development of general trackers for single-camera, single-target, model-free, causal tracking. For nearly a decade the VOT has thus been a community-driven forum for gradual development and in-situ testing of performance evaluation protocols, dataset development and exploration of the tracking challenges landscape. The VOT2013 [35] started with a single short-term tracking challenge; VOT-ST. In VOT2014 [36] the VOT-TIR challenge was added to explore tracking in thermal imagery. In VOT2017 [31] the real-time tracking challenge VOT-RT was established to promote tracking speed and computational efficiency in parallel to robustness. Long-term tracking challenge VOT-LT was introduced in

[56] http://votchallenge.net

VOT2018 [30] and a year later in VOT2019 [28], multi-modal (RGB+thermal and RGB+depth) tracking challenges VOT-RGBT and VOT-RGBD were introduced.

Particular attention has been put on the development of informative performance evaluation measures. Two basic weakly correlated performance measures were introduced in VOT2013 [35] to evaluate the tracking accuracy and robustness of short-term trackers. A ranking-based methodology to identify the top performers was also proposed but was abandoned in VOT2015 [34] in favor of a more principled and interpretable combination of the primary scores in form of the expected average overlap score EAO. For the first seven VOT challenges, the measures were calculated under a reset-based protocol, in which a tracker is reset upon drifting off the target. This protocol was replaced in VOT2020 [29] by the anchor-based evaluation protocol that produces the most stable performance evaluation results compared to related protocols, yet inherits the benefits from the reset-based protocol. Similarly, a performance evaluation protocol and measures tailored for long-term tracking have been developed [41] and applied first in VOT2018 [30]. These measures have consistently shown good evaluation capabilities for long-term trackers.

Several datasets have been developed over the years. A dataset creation and maintenance protocol has been established for the main short-term tracking challenge to produce datasets which are sufficiently small for practical evaluation yet include a variety of challenging tracking situations for in-depth analysis. In VOT2017 [31], a sequestered dataset for identification of the short-term tracking challenge winner was introduced. This dataset has been refreshed along with the public versions over the years. Alongside, datasets specialized for long-term tracking, RGB+thermal and RGBD tracking were constructed and gradually updated.

In most of the VOT challenges, the trackers are required to report the target position as an axis-aligned bounding box. While this is a reasonable target state encoding, the VOT short-term tracking challenge gradually explored more detailed pose encodings to push the bar on tracking accuracy and expand the range of applications. Thus rotated bounding boxes were introduced in VOT2014 [36]. To reduce human annotation bias, VOT2016 [32] introduced fitting rotated bounding boxes to semi-automatically segmented objects in each frame. In VOT2020 [29] bounding boxes were abandoned and the short-term trackers are required to provide full target segmentation (the VOT-ST dataset was accordingly re-annotated to ensure high ground truth accuracy) – with this move, the VOT short-term tracking challenge has started narrowing the gap between visual object tracking and the related field of video object segmentation. The remaining challenges (VOT-LT, VOT-RGBD, VOT-RGBT) maintain axis-aligned target annotation.

This paper presents the tenth edition of the VOT challenges – the VOT2022 challenge. After two years of virtual editions due to the global Covid19 pandemic, the 10th anniversary of VOT was organized in a mixed form with in-person and online attendance, in conjunction with the ECCV2022 Visual Object Tracking

VOT2022 Workshop. In the following, we overview the challenge and participation requirements.

1.1 The VOT2022 challenge

The evaluation toolkit and the datasets are provided by the VOT2022 organizers. The challenges opened in the first week of April and closed on May 3rd. The winners of individual challenges were identified in late June, but not publicly disclosed. The results were presented at the ECCV2022 VOT2022 workshop on 24th October. The VOT2022 challenge contained seven challenges:

1. **VOT-STs2022 challenge** addressed short-term tracking by target segmentation in RGB images.
2. **VOT-STb2022 challenge** addressed short-term tracking by bounding boxes in RGB images.
3. **VOT-RTs2022 challenge** addressed the same class of trackers as VOT-STs2022, except that the trackers had to process the sequences in real-time.
4. **VOT-RTb2022 challenge** addressed the same class of trackers as VOT-STb2022, except that the trackers had to process the sequences in real-time.
5. **VOT-LT2022 challenge** addressed long-term tracking by bounding boxes in RGB images.
6. **VOT-RGBD2022 challenge** addressed short-term tracking by bounding boxes in RGB+depth (RGBD) imagery.
7. **VOT-D2022 challenge** addressed short-term tracking by bounding boxes in depth map images.

The authors participating in the challenge were required to integrate their tracker into the VOT2022 evaluation kit, which automatically performed a set of standardized experiments. The results were analyzed according to the VOT2022 evaluation methodology.

Participants were encouraged to submit their own new or previously published trackers as well as modified versions of third-party trackers. In the latter case, modifications had to be significant enough for acceptance. Participants were expected to submit a single set of results per tracker If a participant coauthored several submissions with a similar design, only the top performer from this *cluster* was considered to compete in the final top-performer ranking and winner identification.

Each submission was accompanied by a short abstract describing the tracker, which was used for the short tracker descriptions in Appendix 5 – the authors were asked to provide a clear description useful to the readers of the VOT2022 results report. In addition, participants filled out a questionnaire on the VOT submission page to categorize their tracker according to various design properties. Authors were encouraged to submit their tracker integrated into a Singularity container provided by VOT, which allows result reproduction and aids potential further evaluation. The participants with sufficiently well-performing submissions who contributed to the text for this paper and agreed to make their tracker code publicly available from the VOT page (or upon request) were offered co-authorship of this results paper. The committee reserved the right to disqualify any tracker that, by their judgement, attempted to cheat the evaluation protocols or failed in the post-hoc evaluation.

Methods considered for prizes in the VOT2022 challenge were not allowed to be trained on certain datasets (OTB, VOT, ALOV, UAV123, NUSPRO, TempleColor and RGBT234), except for VOT-LT2022, where the VOT-LT2021 dataset was allowed. For GOT10k, a list of 1k prohibited sequences was created in VOT2019, while the remaining 9k+ sequences were allowed for learning. The reason was that part of the GOT10k was used in the VOT-ST2022 dataset.

The use of class labels specific to VOT was not allowed (i.e., identifying a target class in each sequence and applying pre-trained class-specific trackers was not allowed). The organizers of VOT2022 were allowed to participate in the challenge but were not eligible to win. Further details are available from the challenge homepage[57].

VOT2022 goes beyond previous challenges by updating the datasets in VOT-ST2022 and VOT-RT2022, introducing a training dataset as well as a sequestered dataset in the VOT-RGBD2022 challenge, introducing a depth-only tracking challenge VOT-D2022 and a new challenging VOT-LT2022 tracking dataset. The Python VOT evaluation toolkit was updated as well.

The remainder of this report is structured as follows. Section 2 describes the performance evaluation protocols, Section 3 describes the individual challenges, Section 4.5 overviews the results and conclusions are drawn in Section 5. Short descriptions of the tested trackers are available in Appendix 5.

2 Performance evaluation protocol

Since VOT2018, the VOT challenges adopt the following definitions from [41] to distinguish between short-term and long-term trackers:

- **Short-term tracker** (ST_0). The target position is reported at each frame. The tracker does not implement target re-detection and does not explicitly detect occlusion.
- **Short-term tracker with conservative updating** (ST_1). The target position is reported at each frame. Target re-detection is not implemented, but tracking robustness is increased by selectively updating the visual model depending on a tracking confidence estimation mechanism.
- **Pseudo long-term tracker** (LT_0). The target position is not reported in frames when the target is predicted not visible. The tracker does not implement explicit target re-detection but uses an internal mechanism to identify and report tracking failure.
- **Re-detecting long-term tracker** (LT_1). The target position is not reported in frames when the target is predicted not visible. The tracker detects tracking failure and implements explicit target re-detection.

[57] http://www.votchallenge.net/vot2022/participation.html

Since the two classes of trackers make distinct assumptions on target presence, separate performance measures and evaluation protocols were designed in VOT to probe the tracking properties.

2.1 The short-term evaluation protocols

The short-term performance evaluation protocol entails initializing the tracker at several frames in the sequence, called the anchor points, which are spaced approximately 50 frames apart. The tracker is run from each anchor - in the first half of the sequences in the forward direction, for anchors in the second half backwards, till the first frame. Performance is evaluated by two basic measures *accuracy* (A) and *robustness* (R).

Accuracy is the average overlap on frames before tracking failure, averaged over all sub-sequences. Robustness is the percentage of successfully tracked subsequence frames, averaged over all sub-sequences. Tracking failure is defined as the frame at which the overlap between the ground truth and predicted target position dropped below 0.1 and did not increase above this during the next 10 frames. This definition allows short-term failure recovery in short-term trackers. The primary performance measure is the expected average overlap EAO, which is a principled combination of tracking accuracy and robustness. Please see [29] for further details on the VOT short-term tracking performance measures.

2.2 The long-term evaluation protocol

The long-term performance evaluation protocol follows the protocol proposed in [41] and entails initializing the tracker in the first frame of the sequence and running it until the end of the sequence. The tracker is required to report the target position in each frame along with a score that reflects the certainty that the target is present at that position. Performance is measured by two basic measures called the tracking precision (Pr) and the tracking recall (Re), while the overall performance is summarized by the tracking F-measure.

The performance measures depend on the target presence certainty threshold, thus the performance can be visualized by the tracking precision-recall and tracking F-measure plots obtained by computing these scores for all thresholds. The final values of Pr, Re and F-measure are obtained by selecting the certainty threshold that maximizes tracker-specific F-measure. This avoids all manually-set thresholds in the primary performance measures.

3 Description of individual challenges

3.1 VOT-ST2022 challenge outline

This challenge addressed RGB tracking in a short-term tracking setup. The initial VOT challenges required target prediction in form of bounding boxes, while a transition to segmentation output requirement has been made in VOT2020.

Nevertheless, to support the very much active community that develops bounding box prediction trackers, the bounding box challenge is re-introduced in VOT2022. Thus the VOT-ST2022 ran two subchallenges: the main segmentation-based short-term tracking challenge VOT-STs2022, and the legacy bounding-box-based short-term tracking challenge VOT-STb2022.

The dataset. Results of the VOT2021 showed that the dataset was not saturated [33], thus the public dataset has been only refreshed by the addition of two sequences which include new challenging scenarios not present in previous VOT datasets: (i) a transparent deforming object and (ii) a flat object with significant out of plane rotations (see Figure 1). The sequestered dataset has been updated with two sequences matching the public dataset extension.

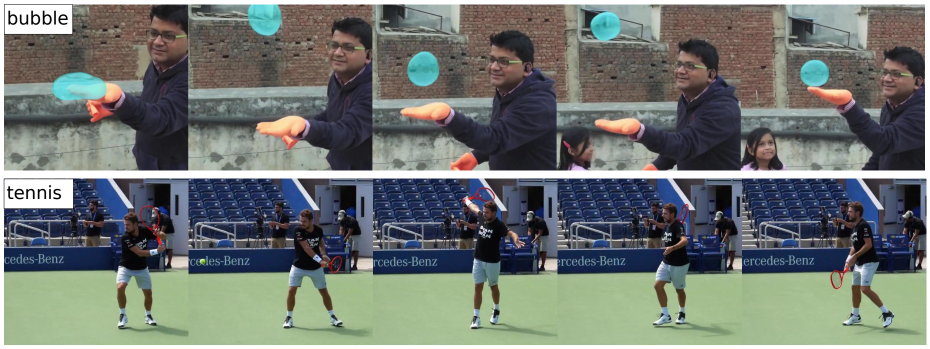

Fig. 1. Two sequences with new challenging scenarios were added to the VOT-ST2022 public dataset. In the sequence 'bubble' the bubble has to be tracked, while in the sequence 'tennis' the racquet is the target object.

The new sequences were frame-by-frame semi-automatically segmented to provide the segmentation ground truth for the main VOT-STs2022 subchallenge. For the legacy VOT-STb2022 subchallenge, the target position was annotated in all sequences by fitting axis-aligned bounding boxes to the target segmentation masks. Per-frame visual attributes were semi-automatically assigned to the new sequences following the VOT attribute annotation protocol. In particular, each frame was annotated by the following visual attributes: (i) occlusion, (ii) illumination change, (iii) motion change, (iv) size change, (v) camera motion.

Winner identification protocol. The VOT-STs2022 winner was identified as follows. Trackers were ranked according to the EAO measure on the public dataset. The top five ranked trackers were then re-run by the VOT2022 committee on the sequestered dataset. The top-ranked tracker on the sequestered dataset not submitted by the VOT2022 committee members is the winner. The same protocol was used to identify the winner of the legacy short-term challenge VOT-STb2022.

3.2 VOT-RT2022 challenge outline

This challenge addressed *real-time* RGB tracking in a short-term tracking setup. The dataset was the same as in the VOT-ST2022 challenge, but the evaluation protocol was modified to emphasize the real-time component in tracking performance. In particular, the VOT-RT2022 challenge requires predisetcting bounding boxes faster or equal to the video frame rate. The toolkit sends images to the tracker via the Trax protocol [54] at 20fps. If the tracker does not respond in time, the last reported bounding box is assumed as the reported tracker output at the available frame (zero-order hold dynamic model). The same performance evaluation protocol as in VOT-ST2022 is then applied. As in VOT-ST2022, two realtime subchallenges were considered: the main segmentation-based realtime subchallenge VOT-RTs2022 and the legacy bounding-box-based realtime subchallenge VOT-RTb2022.

Winner identification protocol. All trackers are ranked on the public RGB short-term tracking dataset with respect to the EAO measure. The winner of the main VOT-RTs2022 subchallenge was identified as the top-ranked tracker not submitted by the VOT2022 committee members. The same methodology was applied to identify the winner of the VOT-RTb2022 challenge.

3.3 VOT-LT2022 challenge outline

This challenge addressed RGB tracking in a long-term tracking setup and is a continuation of the VOT-LT2021 challenge. We adopt the definitions from[41], which are used to position the trackers on the short-term/long-term spectrum. A long-term performance evaluation protocol and measures from Section 2.2 were used to evaluate tracking performance on VOT-LT2022. Compared to VOT-LT2021, a significant change is a new dataset described in the following.

The dataset. The new VOT-LT dataset contains 50 sequences, carefully selected to obtain a dataset with long sequences containing many target disappearances. The LTB50 [41], which was used in VOT-LT2021, is the training set this year. The new VOT-LT dataset contains 50 challenging sequences of diverse objects (persons, cars, motorcycles, bicycles, boats, animals, etc.) with a total length of 168,282 frames. The sequence resolution is 1280 × 720. Each sequence contains on average 10 long-term target disappearances, each lasting on average 52 frames. An overview of the dataset is shown in Figure 2.

The targets are annotated by axis-aligned bounding boxes. Sequences are annotated by the following visual attributes: (i) full occlusion, (ii) out-of-view, (iii) partial occlusion, (iv) camera motion, (v) fast motion, (vi) scale change, (vii) aspect ratio change, (viii) viewpoint change, (ix) similar objects. Note this is per-sequence, not per-frame annotation and a sequence can be annotated by several attributes. Compared with LTB50,the new VOT-LT dataset is more challenging in small objects, similar objects, fast motion, and full/partial occlusions.

Winner identification protocol. The VOT-LT2022 winner was identified as follows. Trackers were ranked according to the tracking F-score on the new LT dataset (no sequestered dataset available). The top-ranked tracker on the dataset

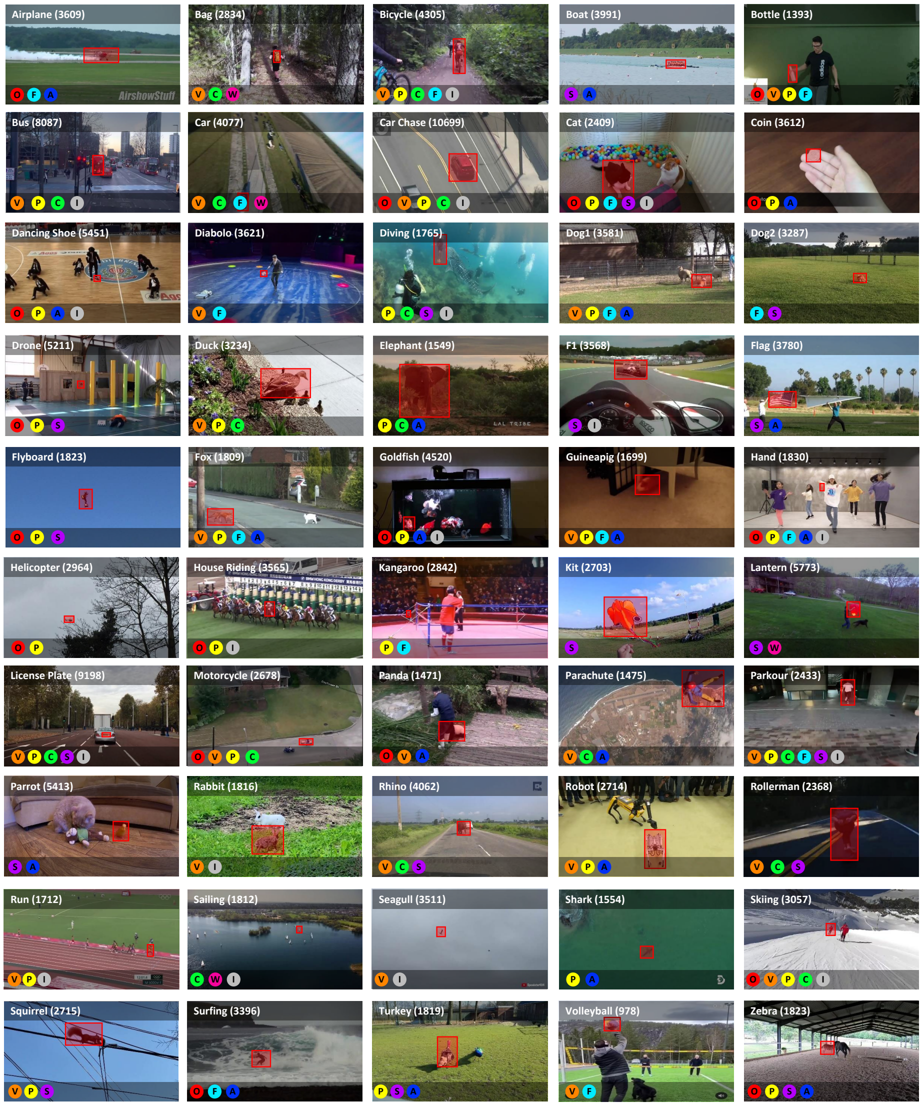

Fig. 2. The new VOT-LT dataset – a frame selected from each sequence. Name and length (top), visual attributes (bottom left): (O) full occlusion, (V) out-of-view, (P) partial occlusion, (C) camera motion, (F) fast motion, (S) scale change, (A) aspect ratio change, (W) viewpoint change, (I) similar objects. The dataset is highly diverse in attributes and target types and contains many target disappearances.

not submitted by the VOT2022 committee members is the winner of the VOT-LT2022 challenge.

3.4 VOT-RGBD2022 challenge outline

The first RGBD (RGB and Depth) challenge was introduced to VOT 2019 and the two first challenges were based on the same public dataset, CDTB [38], which consists of 80 sequences where the target momentarily disappears or is fully occluded. In VOT 2021, the CDTB dataset was replaced with new sequences captured with an Intel RealSense 415 RGBD camera that provides spatially aligned RGB and depth frames. The 2021 dataset contained 80 public and 50 sequestered test sequences. The main motivation for the new dataset was to make it more challenging in the sense that sometimes depth cue is more informative and sometimes RGB. Moreover, separate training and test sequences were provided to allow method fine-tuning with dataset-specific data. More details about the dataset and its properties can be found from [62]. The two major changes as compared to the previous years' RGBD tracks are that 1) the challenge is now a short-term (ST) tracking challenge and 2) the challenge is divided into RGBD and depth-only (D) tracks in order to better understand how much depth contributes to RGBD tracking, i.e. complementarity of the two modalities.

The main motivation to switch from the long-term evaluation to short-term evaluation is that in the long-term setting the target disappearance played an important role and many of the proposed RGBD trackers used the depth channel to assist in occlusion detection, but otherwise the cue was omitted. Now, the two tracks, RGBD and D, provide information about the complementary properties of color texture and depth. It is noteworthy that the RGBD and D challenges use otherwise exactly the same data.

The dataset. Inspired by the recent work on depth-only tracking [63], we converted the long-term sequences from the CDTB dataset used in the first two VOT-RGBD challenges and DepthTrack used in the latest challenge, to short-term sequences. We converted all 80 sequences from CDTB and 50 test sequences of DepthTrack. Since the DepthTrack training sequences were not used they can be used in training learning-based trackers. The short-term sequences were manually checked and sequences with poor depth information or other errors were removed. Finally, 127 sequences were selected and published on the VOT Web site. See Figure 3 for example frames.

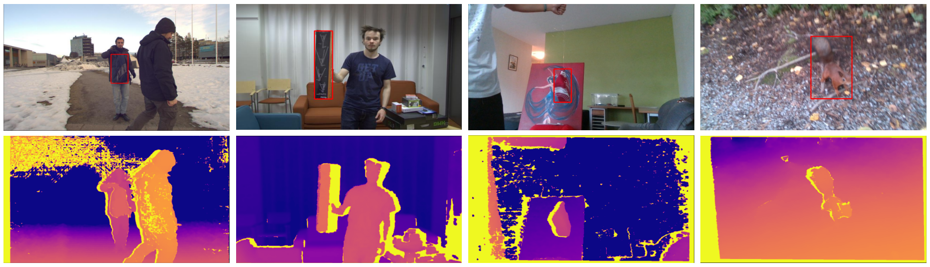

Fig. 3. Samples from the RGBD and D challenge sequences. The first two from the left are from the CDTB sequences and the next two from DepthTrack-test sequences.

VOT-D2022. The data for the VOT-D2022 challenge is exactly the same as for VOT-RGBD except that the RGB frames are removed.
Winner identification protocol. The VOT-RGBD2022 and VOT-D2022 winners were identified as follows. Trackers were ranked according to the EAO measure on the public dataset and the top-ranked tracker on the public dataset not submitted by the VOT2022 committee members is the winner. The same protocol was used to identify the winners of both the VOT-RGBD and VOT-D challenges.

4 The VOT2022 challenge results

This section summarizes the trackers submitted, results analysis and winner identification for each of the VOT2022 challenges.

4.1 The VOT-STs2022 challenge results

The VOT-STs2022 challenge tested 31 trackers, including the baselines contributed by the VOT committee. Each submission included the binaries or source code that allowed verification of the results if required. In the following, we briefly overview the entries and provide the references to original papers in the Appendix A where available.

Of the participating trackers, 13 trackers (42%) were categorized as ST_0, 14 trackers (45%) as ST_1, and 4 (13%) as LT_0. 81% applied discriminative and 19% applied generative models. Most trackers (81%) used a holistic model, while 19% of the participating trackers used part-based models. Most trackers (75%) applied an equally probably displacement within a region centered at the current position[58] or a random walk dynamic model (25%). 42% of trackers localized the target in a single stage, while the rest applied several stages, typically involving approximate target localization and position refinement. Most of the trackers (84%) use deep features. The majority of the submissions (72%) localized the target by segmentation, while the rest reported a bounding box.

The trackers were based on various tracking principles. 11 trackers were based on classical or deep discriminative correlation filters (RTS A.8, ATOM_AR A.18, DiMP_AR A.19, KYS_AR A.20, PrDiMP_AR A.21, CSRDCF A.24, D3Sv2 A.25, SuperDiMP_AR A.22, KCF A.26, LWL A.28, LWL-B2S A.29), 2 trackers were based purely on Siamese correlation (SiamFC A.30, SiamUSCMix A.9), 14 trackers were based on transformers (DAMT A.1, DAMTMask A.2, DGformer A.3, Linker A.4, MixFormerM A.5, MS_AOT A.6, OSTrackSTS A.7, SwinT A.11, SRATransTS A.10, TransLL A.12, TransT A.13, transt_ar A.14, TransT_M A.15, and TRASFUSTm A.16), two were deformable parts trackers (ANT A.17 and LGT A.27), a meanshift tracker (ASMS A.23), and a video-object segmentation method adapted to tracking (STM A.31).

[58] The target was sought in a window centered at its estimated position in the previous frame. This is the simplest dynamic model that assumes all positions within a search region containing the target have an equal prior probability.

In summary, we observe a significant increase in a new class of trackers identified in VOT2021 – the transformers. In fact, 47% of trackers are now from this class, 41% of trackers apply discriminative correlation filters, while 6% apply classical siamese correlation networks.

Results The results are summarized in the AR-raw plots and EAO plots in Figure 4 and in Table 10. The top ten trackers according to the primary EAO measure (Figure 4) are MS_AOT A.6, DAMTMask A.2, MixFormerM A.5, OSTrackSTS A.7, Linker A.4, SRATransTS A.10, TransT_M A.15, DGformer A.3, TransLL A.12 and LWL-B2S A.29. Nine of the top trackers apply transformers as the core tracking methodology and one applied deep DCFs. Seven apply a two-stage target localization, meaning that they first localize the target by a bounding box and the segment the target withing the bounding box with a separate network (two of these apply Alpha-Refine [61] – the winner of VOT-RT2020 challenge). Three of the top 10 trackers are single-stage, meaning that they directly segment the target. Four of the trackers are apply elements (or are extensions) of MixFormers [10], four extend TransT [8] and three apply ViT [16].

The top tracker on the public set according to EAO is MS_AOT A.6, which is based on the recent transformer-based video object segmentation AOT [65]. For normal-sized objects, the tracker acts as a single-stage segmentation method. For tiny objects, the tracker works in a two-stage regime in which the object is first localized by bounding box using MixFormer [10] and then segmented by the AOT.

The second-best tracker is DAMTMask A.2, which is build on top of MixFormer [10] and SuperDiMP [3], and applied a two-stage target localization and segmentation approach. The target location is predicted by RepPoints [64] and a MixFormer-like head is implemented to predict the segmentation mask.

The third-best tracker is MixFormerM A.5, a two-stage tracker which uses a new mixed attention module for simultaneous feature extraction and target information fusion.

The three top performers in EAO are among the top three performers in accuracy (A) and robustness (R) measures as well (Table 10). While these trackers are comparable in target localization accuracy, MS_AOT stands out by its remarkable robustness (Figure 4).

	CM	**IC**	**OC**	**SC**	**MC**
Accuracy	0.62	0.62	0.52	0.64	0.63
Robustness	0.79	0.75	0.68	0.78	0.76

Table 1. VOT-STs2022 tracking difficulty with respect to the following visual attributes: camera motion (CM), illumination change (IC), motion change (MC), occlusion (OC) and size change (SC).

Three of the tested trackers have been published in major computer vision journals and conferences in the last two years (2021/2022). These trackers are

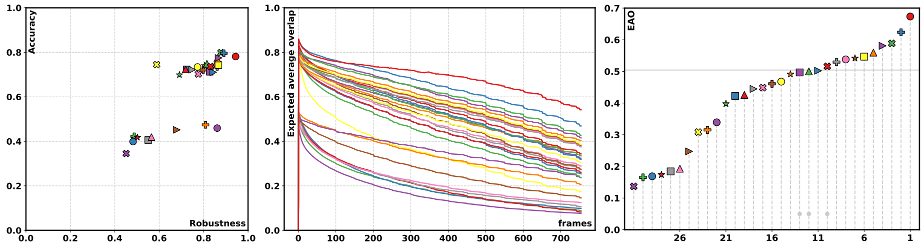

Fig. 4. The VOT-STs2022 AR-raw plots generated by sequence pooling (left) and EAO curves (center) and the VOT-STs2022 expected average overlap graph with trackers ranked from right to left. The right-most tracker is the top-performing according to the VOT-STs2022 expected average overlap values. The dashed horizontal line denotes the average performance of three state-of-the-art trackers published in 2021/2022 at major computer vision venues. These trackers are denoted by gray circle in the bottom part of the graph. See Table 10 for the tracker labels.

indicated in Figure 4, along with their average performance (EAO=0.504), which constitutes the VOT2022 state-of-the-art bound. Approximately 32% of the submissions exceed this bound.

The per-attribute robustness analysis is shown in Figure 5 for individual trackers. The overall top performers remain at the top of per-attribute ranks as well. MS_AOT achieves top robustness in all attributes. According to the median failure over each attribute (Table 1) the most challenging attribute remains occlusion. The drop on this attribute is consistent for all trackers (Figure 5).

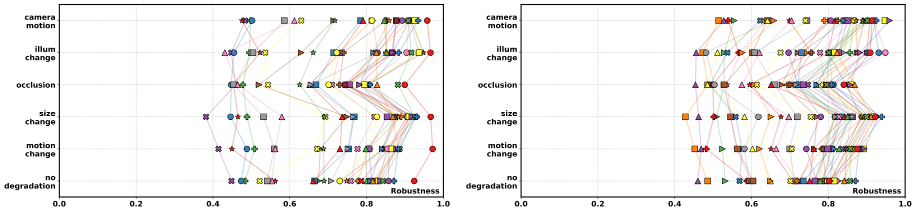

Fig. 5. Robustness with respect to the visual attributes in VOT-STs2022 challenge (left) and in the VOT-STb2022 challenge (right). See Table 10 and Table 12 for VOT-STs2022 and VOT-STb2022 tracker labels, respectively.

The VOT-STs2022 challenge winner The top five trackers from the baseline experiment (Table 10) were re-run on the sequestered dataset. Their scores obtained on the sequestered dataset are shown in Table 2. The top tracker according to the EAO is MS_AOT A.6 and is thus the VOT-STs2022 challenge winner.

	Tracker	EAO	A	R
1.	MS_AOT	0.565	0.823	0.906
2.	DAMTMask	0.513	0.846	0.830
3.	OSTrackSTS	0.500	0.822	0.845
4.	MixFormerM	0.497	0.844	0.819
5.	Linker	0.492	0.829	0.830

Table 2. The top five trackers from Table 10 re-ranked on the VOT-STs2022 sequestered dataset.

4.2 The VOT-STb2022 challenge results

The VOT-STb2022 challenge tested 41 trackers, including the baselines contributed by the VOT committee. Each submission included the binaries or source code that allowed verification of the results if required. In the following, we briefly overview the entries and provide the references to original papers in the Appendix B where available. The trackers were based on various tracking principles. 13 trackers were based on classical or deep discriminative correlation filters (SuperFus B.20, TCLCFcpp B.22, KCF B.36, D3Sv2 B.34, DiMP B.35, ATOM B.32, CSRDCF B.33, SuperDiMP B.41, PrDiMP B.39, FSC2F B.7, oceancycle B.13, DeepTCLCF B.5, KYS B.37), 4 trackers were based purely on Siamese correlation (NfS B.12, SiamUSCMix B.17, SiamVGGpp B.18, SiamFC B.40), 19 trackers were based on transformers (TransT_M B.26, TransT B.25, ADOTstb B.1, GOANET B.8, DAMT B.4, tomp B.23, TransLL B.24, APMT_MR B.2, APMT_RT B.3, DGformer B.6, Linker_B B.9, MixFormer B.10, ViTCRT B.28, MixFormerL B.11, OSTrackSTB B.14, SRATransT B.19, vittrack B.29, SwinTrack B.21, SBT B.16), one ensamble-based (TRASFUST B.27), one was based on meta-learning (ReptileFPN B.15), one was scale-adaptive mean-shift tracker (ASMS B.31), and two were part-based generative trackers (ANT B.30 and LGT B.38).

Results The results are summarized in the AR-raw plots and EAO plots in Figure 6, and in Table 12. The top ten trackers according to the primary EAO measure (Figure 6) are DAMT B.4, MixFormerL B.11, OSTrackSTB B.14, APMT_MR B.2, MixFormer B.10, APMT_RT B.3, ADOTstb B.1, SRATransT B.19, Linker_B B.9, TransT_M B.26. Like in the segmentation tracking challenge VOT-STs2022, all top ten trackers apply transformers. In fact, seven of the top trackers are modifications of segmentation-based counterparts, ranked among the top ten trackers on the VOT-STs2022: MixFormerL, DAMT, OSTrackSTB, MixFormer, SRATransT, Linker, TransT.

All three top-ranked trackers on the public dataset according to EAO, are counterparts of the top-ranked trackers on the main segmentation challenge VOT-STs2022. The two top performers, with equal EAO are MixFormerL B.11 and DAMT B.4. MixFormerL B.11, is a counterpart of the tracker ranked third on VOT-STs2022, while DAMT B.4 is a counterpart of the second-ranked tracker on VOT-STs2022. The two trackers excel in different tracking properties. DAMT is more robust than MixFormerL, while MixformerL is delivers a more accurate target estimation than DAMT. The third-best ranked tracker is OSTrackSTB is a counterpart of the fourth-best ranked tracker on VOT-STs2022.

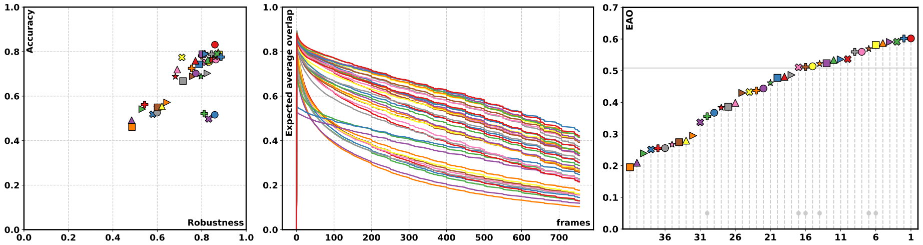

Fig. 6. The VOT-STb2022 AR-raw plots generated by sequence pooling (left) and EAO curves (center) and the VOT-STb2022 expected average overlap graph with trackers ranked from right to left. The right-most tracker is the top-performing according to the VOT-STs2022 expected average overlap values. The dashed horizontal line denotes the average performance of ten state-of-the-art trackers published in 2021/2022 at major computer vision venues. These trackers are denoted by gray circle in the bottom part of the graph. See Table 12 the tracker labels.

Seven of the tested trackers have been published in major computer vision journals and conferences in the last two years (2021/2022). These trackers are indicated in Figure 6, along with their average performance (EAO=0.484), which constitutes the VOT2022 state-of-the-art bound. Approximately 43.9% of the submissions exceed this bound.

The per-attribute robustness analysis is shown in Figure 5 for individual trackers. The overall top performers remain at the top of per-attribute ranks as well, however, none of the trackers consistently outperforms the rest in all attributes. According to the median failure over each attribute (Table 3) the most challenging attribute remains occlusion. The drop on this attribute is consistent for all trackers (Figure 5).

	CM	**IC**	**OC**	**SC**	**MC**
Accuracy	0.68	0.63	0.55	0.66	0.65
Robustness	0.79	0.74	0.69	0.77	0.74

Table 3. VOT-STb2022 tracking difficulty with respect to the following visual attributes: camera motion (CM), illumination change (IC), motion change (MC), occlusion (OC) and size change (SC).

The VOT-STb2022 challenge winner Top trackers from the baseline experiment (Table 12) were re-run on the sequestered dataset. Since some of the top trackers were variations of the same tracker, the VOT committee selected only the top-performing variant as a representative to be run on the sequestered dataset. Note that there are several ways to specify the ground truth against which the predicted bounding boxes from the trackers can be evaluated. The most straight-forward way is to fit bounding boxes to the ground truth masks (as done in the public evaluation). However, the most accurate ground truth target location specification is actually a segmentation mask and the predicted bounding box from the tracker can be considered as its parametric approximation. We thus inspected the tracker performance for winner identification along the bounding box ground truth specification and along the segmentation mask ground truth specification.

The scores using the bounding box ground truth are shown in Table 4, while the scores using the segmentation mask ground truth are shown in Table 5. We observe that the tracker ranks remain the same across the two ground truth specifications, except from the top two, who switch ranks. For this reason, both top-performers are determined as the winners of the VOT-STb2022 challenge, each in its category. The winner of the VOT-STb2022 challenge in the bounding box ground truth category is OSTrackSTB B.14, while the winner in the segmentation mask ground truth category is APMT_MR B.2.

	Tracker	**EAO**	**A**	**R**
1.	OSTrackSTB	0.523	0.800	0.881
2.	APMT_MR	0.508	0.800	0.862
3.	MixFormerL	0.500	0.837	0.825
4.	ADOTstb	0.499	0.812	0.840
5.	DAMT	0.479	0.804	0.826

Table 4. The top five trackers from Table 12 re-ranked on the VOT-STb2022 sequestered dataset using the bounding box ground truth.

	Tracker	**EAO**	**A**	**R**
1.	APMT_MR	0.322	0.528	0.845
2.	OSTrackSTB	0.309	0.517	0.839
3.	MixFormerL	0.306	0.542	0.803
4.	ADOTstb	0.301	0.532	0.806
5.	DAMT	0.289	0.503	0.807

Table 5. The top five trackers from Table 12 re-ranked on the VOT-STb2022 sequestered dataset using the segmentation masks as ground truth.

4.3 The VOT-RTs2022 challenge results

The trackers that entered the VOT-STs2022 challenge were also run on the VOT-RTs2022 challenge. Thus the statistics of submitted trackers were the same as in VOT-ST2022. For details please see Section 4.2 and Appendix A.

Results The EAO scores and AR-raw plots for the trackers participating in the VOT-RTs2022 challenge are shown in Figure 7 and Table 10. The top ten segmentation-based real-time trackers are MS_AOT A.6, OSTrackSTS A.7, SRATransTS A.10, TransT_M A.15, DGformer A.3, MixFormerM A.5, TransLL A.12, TransT A.13 and Linker A.4 and RTS A.8.

Nine of the top ten trackers are based on transformers. Nine trackers are ranked among to top 10 on the VOT-STs2022 challenge: MS_AOT, OSTrackSTS, SRATransTS, TransT_M, DGformer, MixFormerM, TransLL, Linker and rts, while TransT is a variation of TransT_M. The top-ranked tracker on realtime challenge according to EAO is MS_AOT, which is also the top-performer on the VOT-STs2022 public datast, the second-best is OSTrackSTS, which ranks fourth on VOT-STs2022 and the third is SRATransTS, which ranks seventh on VOT-STs2022. This indicates significant advancement in the field of visual object tracking since the inception of the VOT realtime challenges, indicating that the speed limitation of modern robust trackers has been confidently breached by transformers.

Three of the tested trackers have been published in major computer vision journals and conferences in the last two years (2021/2022). These trackers are indicated in Figure 7, along with their average performance (EAO=0.422), which constitutes the VOT2022 state-of-the-art bound. Approximately 45.2% of the submissions exceed this bound.

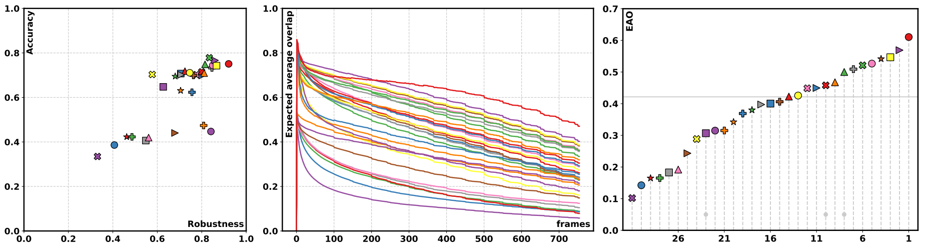

Fig. 7. The VOT-RTs2022 AR plot (left), the EAO curves (center) and the EAO plot (right). The dashed horizontal line denotes the average performance of seven state-of-the-art trackers published in 2021/2022 at major computer vision venues. These trackers are denoted by gray circle in the bottom part of the graph.

The VOT-RTs2022 challenge winner According to the EAO results in Table 10, the top performer and the winner of the segmentation-based real-time tracking challenge VOT-RTs2022 is MS_AOT(A.6).

4.4 The VOT-RTb2022 challenge results

The trackers that entered the VOT-STb2022 challenge were also run on the VOT-RTb2022 challenge. Thus the statistics of submitted trackers were the same as in VOT-STb2022. For details please see Section 4.1 and Appendix B.

Results The EAO scores and AR-raw plots for the trackers participating in the VOT-RTb2022 challenge are shown in Figure 8 and Table 12. The top ten bounding-box-based real-time trackers are OSTrackSTB B.14, APMT_RT B.3, MixFormer B.10, APMT_MR B.2, SRATransT B.19, DAMT B.4, TransT_M B.26, vittrack, SBT B.16, TransT B.25. All of these are based on transformers. Seven are among the top ten performers on the public dataset in VOT-STb2022: OSTrackSTB, APMT_RT, MixFormer,APMT_MR, SRATransT, DAMT and TransT_M. Thus, similarly to VOT-RTs2022, results here show that performance is minimally compromised if at all on account of speed in transformer-based tracking.

The top-performer according to the EAO on the public dataset is OSTrackSTB, which is based on the recent OSTrack [68] and uses a ViT [16] backbone. This tracker is ranked third on VOT-STb2022. The second and the third-best trackers on VOT-RTb2022 are APMT_RT and MixFormer, which are ranked fourth and fifth on VOT-STb2022.

Note that 7 of the tested trackers have been published in major computer vision journals and conferences in the last two years (2021/2022). These trackers are indicated in Figure 8, along with their average performance (EAO=0.421), which constitutes the VOT2022 state-of-the-art bound. Approximately 53.7% of the submissions exceed this bound.

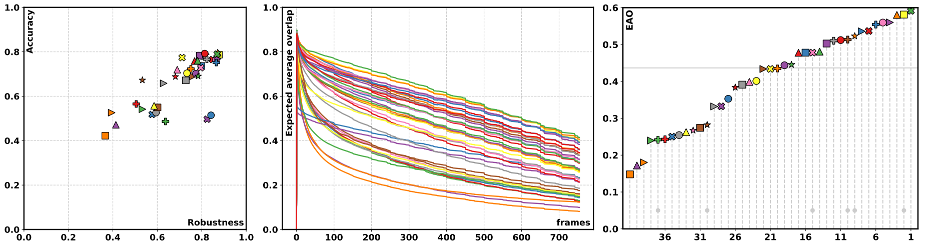

Fig. 8. The VOT-RTb2022 AR plot (left), the EAO curves (center) and the EAO plot (right). The dashed horizontal line denotes the average performance of ten state-of-the-art trackers published in 2021/2022 at major computer vision venues. These trackers are denoted by gray circle in the bottom part of the graph.

The VOT-RTb2022 challenge winner According to the EAO results in Table 12, the top performer and the winner of the bounding-box-based real-time tracking challenge VOT-RTb2022 is OSTrackSTB (B.14).

4.5 The VOT-LT2022 challenge results

Trackers submitted The VOT-LT2022 challenge received 7 valid entries. The VOT2022 committee contributed additional trackers SuperDiMP and KeepTrack as baselines; thus 9 trackers were considered in the challenge. In the following, we briefly overview the entries and provide the references to original papers in Appendix C where available.

All participating trackers were categorized as ST_1 according to the ST-LT taxonomy from Section 2 in that they implemented explicit target re-detection. All trackers were based on convolutional neural networks. Four trackers applied Transformer architecture akin to STARK [60] for target localization (CoCoLoT C.2, mixLT C.5, mlpLT C.6, and VITKT_M C.8). Particularly, VITKT_M C.8 is based purely on a Transformer-backbone [52] for feature extraction. Four trackers applied SuperDiMP structure [3] as their basic tracker (ADiMPLT C.1, mixLT C.5, mlpLT C.6, SuperDiMP C.9). Three trackers selected KeepTrack [46] as their auxiliary tracker due to its robustness to distractors (CoCoLoT C.2, VITKT_M C.8, KeepTrack C.10). One tracker is based on MixFormer [10] to design a long-term tracker that focuses on target recapture (HuntFormer C.4). One tracker extends the D3Sv2 [42] short-term tracker with long-term capabilities (D3SLT C.3). Four trackers combined different tracking methods and switched them based on their tracking scores (CoCoLoT C.2, D3SLT C.3, mixLT C.5, mlpLT C.6, VITKT_M C.8). Among them, two trackers use an online real-time MDNet-based [48] verifier to determine the tracking score (CoCoLoT C.2, D3SLT C.3).

Tracker	**Pr**	**Re**	**F-Score**	**Year**
● VITKT_M	0.629①	0.604②	0.617①	2022
✚ mixLT	0.608②	0.592③	0.600②	2022
✖ HuntFormer	0.586	0.610①	0.598③	2022
▶ CoCoLoT	0.591③	0.577	0.584	2022
▲ mlpLT	0.568	0.562	0.565	2022
□ KeepTrack	0.572	0.550	0.561	2022
★ D3SLT	0.520	0.516	0.518	2022
● Super_DiMP	0.510	0.496	0.503	2022
✚ ADiMPLT	0.489	0.514	0.501	2022

Table 6. List of trackers that participated in the VOT-LT2022 challenge along with their performance scores (Pr, Re, F-score) and ST/LT categorization.

Results The overall performance is summarized in Figure 9 and Table 6. The top-three performers are VITKT_M C.8, mixLT_LT C.5 and HuntFormer C.4. VITKT_M obtains the highest F-score (0.617) in 2022, while last year winner (mlpLT) obtains 0.565. Since the new VOT-LT dataset is more challenging, it should be noted that the average F-Score of these trackers decreased by 11.4% than last year. All the results are based on the submitted numbers, but these were verified by running the codes multiple times. The VITKT_M is composed of a Transformer-based tracker VitTrack, an auxiliary tracker KeepTrack and a motion module. Specifically, the master tracker VitTrack is a Transformer-based tracker composed of a backbone network, a corner prediction head and a classification head. Besides, a simple motion module is used to predict the target current state according to the temporal trajector. When scores of VitTrack and KeepTrack are all lower than a threshold, and the target moves abnormally, this motion module is triggered to predict the current state.

The mixLT architecture is a progressive fusion of multiple trackers, mainly STARK [60] and SuperDiMP. Specifically, it first fuses the results of two trackers, STARK-ST50 and STARK-ST101. The states of two trackers are then corrected based on the fusion resuls. SuperDiMP controlled by meta-updater is introduced for further fusion between dissimilar trackers, in order to improve the robustness of long-term tracking. The final tracking result is determined according to the confidences of the trackers over several frames, and another tracker correction is performed.

Based on MixFormer, the tracker HuntFormer propose an effective motion prediction model that provides a reliable search region for the tracker to recapture the target. Meanwhile, we propose a novel soft-threshold-based dynamic memory update model, which keeps a set of reliable target templates in the memory that can be used to match the target position in the search region. The two modules cooperate with each other, which greatly improves the recapture ability of the tracker.

The VITKT_M achieves an overall best F-score and significantly surpasses mixLT (by 1.7%) and MixFormer (by 1.9%). All of these methods are based on Transformer. Two similar trackers, VITKT_M C.8 and VITKT C.7 were submitted by one team. The only difference is that the VITKT is a more concise version than VITKT_M without the motion module. When ablating the motion module (VITKT C.7), the F-score decreases by 1.2%. Since VITKT C.7 is a minor variant of VITKT_M, we only keep VITKT_M in our ranking.

The VOT-LT2022 challenge winner According to the F-score in Table 6, the top-performing tracker is VITKT_M, closely followed by mixLT and HuntFormer. Thus the winner of the VOT-LT 2021 challenge is VITKT_M C.8.

4.6 The VOT-RGBD2022 challenge results

Eight trackers were submitted to the 2022 RGBD challenge: DMTracker (D.1) keep_track (D.2), MixForRGBD (D.3), OSTrack (D.4), ProMix (D.5), SAMF (D.6), SBT_RGBD (D.7) and SPT (D.8).

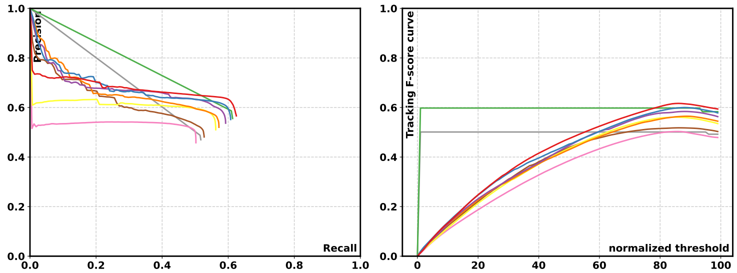

Fig. 9. VOT-LT2022 challenge average tracking precision-recall curves (left) and the corresponding F-score curves (right). Tracker labels are sorted according to maximum of the F-score (see Table 6).

All trackers are based on the popular deep learning-based tracker architectures that have performed well in the previous years VOT RGB challenges. The new deep architecture for this year is MixFormer [10] that is in multiple submissions (MixForRGBD, ProMix and SAMF). The main difference between the submitted trackers is how they fuse the two modalities, depth and RGB, and in their training prodedures. Some teams submitted multiple trackers, but since their architectures are different they were all accepted.

Results The Expected Average Overlap (EAO), Accuracy (A) and Robustness (R) metrics of the submitted and a number of additional trackers are shown in Table 7. The two best performing trackers, MixForRGBD and SAMF, are distinctively better than the next ones. The six best performing trackers are this year submissions while the DepthTrack database baseline, DeT_DiMP50_Max, is the seventh. The two RGB trackers perform the worst as was expected.

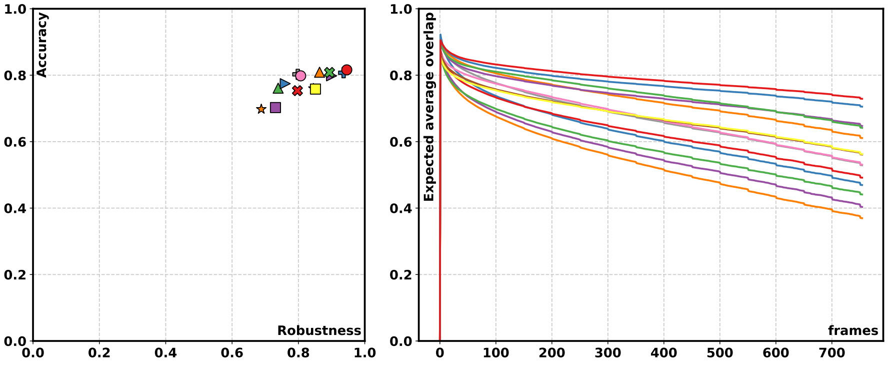

Fig. 10. The VOT-RGBD2022 AR plot (left) and the EAO curves (right).

	Tracker	EAO	A	R
1.	● MixForRGBD	0.779①	0.816	0.946
2.	✚ SAMF	0.762②	0.807	0.936
3.	✖ OSTrack	0.729③	0.808	0.894
4.	▶ ProMix	0.722	0.798	0.900
5.	▲ SBT_RGBD	0.708	0.809	0.864
6.	■ DMTracker	0.658	0.758	0.851
7.	★ DeT_DiMP50_Max	0.657	0.760	0.845
8.	● SPT	0.651	0.798	0.851
9.	✚ STARK_RGBD	0.647	0.803	0.798
10.	✖ keep_track	0.606	0.753	0.797
11.	▶ DRefine	0.592	0.775	0.760
12.	▲ ATCAIS	0.559	0.761	0.739
13.	● DiMP	0.534	0.703	0.731
14.	★ ATOM	0.505	0.698	0.688

Table 7. Results for the eight submitted VOT-RGBD2022 trackers. For comparison, the table also includes the results for the three best performing RGBD trackers from VOT2020 (ATCAIS) and VOT2021 (STARK_RGBD and DRefine), two strong baseline RGB trackers from the previous years (DiMP and ATOM) and the baseline RGBD tracker from the DepthTrack dataset (DeT_DiMP50_Max [62]).

The VOT-RGBD2022 challenge winner The results in Figure 10 show that MixForRGBD and SAMF perform very similarly and are clearly better than the rest. Still, MixForRGBD obtains the best EAO score and is thus the winner of the VOT-RGBD2022 challenge.

4.7 The VOT-D2022 challenge results

The VOT-D2022 challenge uses the same 127 short-time tracking sequences as the above RGBD2022 challenge, but in the D (depth-only) challenge the trackers are provided only the depth map frames. This challenge was added as it was intriguing to study how much RGB adds to the depth cue and what is the complementary power of the two modalities.

The total of six trackers were submitted to the depth-only challenge. The submitted trackers are: CoDeT (E.1), MixFormerD (E.2), OSTrack_D (E.3), RSDiMP (E.4), SBT_Depth (E.5) and UpDoT (E.6).

Not surprisingly the D-only challenge attracted submissions from the same groups that also participated the RGBD challenge. For example, CoDeT is a D-only version of DMTRacker, MixFormerD of MixForRGBD, OSTrack_D of OSTrack, and SBT_Depth of SBT_RGB. RSDiMP is from the same group as the SPT RGBD tracker, but the two architectures are different. The authors of CoDeT also submitted UpDoT which corresponds to standard DiMP trained with two different versions of depth data.

Results The computed performance metrics for the D (depth-only) trackers are in Table 8 and the corresponding graphs in Figure 11. From the results we can see that the depth-only variants of the best performing RGBD trackers also perform well in the D-only challenge (MixFormerRGBD → MixFormerD and OSTrack → OSTrack_D). The only dedicated D-only tracker submitted to the D-only challenge and which does not have an RGBD counterpart, RSDiMP, obtains the second best EAO score. Overall the three best methods, MixFormerD, RSDiMP and OSTrack_D, perform almost on par and are distinctively better than the rest. Therefore, these three trackers are good starting points to understand how to effectively use the depth channel in tracking.

Notably, there is a clear difference between the D-only and RGBD results on the same data (Table 7 vs. Table 8). That confirms that the both modalities, D and RGB, are beneficial for object tracking. For example, the RGB DiMP in Table 7 is clearly better than the depth-only DiMP in Table 8 (EAO 0.534 vs. 0.336), but inferior to the best D-only tracker (MixFormerD 0.600).

	Tracker	EAO	A	R
1.	● MixFormerD	0.600①	0.758	0.806
2.	✚ RSDiMP	0.573②	0.734	0.759
3.	✖ OSTrack_D	0.568③	0.735	0.774
4.	▶ DOT	0.469	0.672	0.673
5.	▲ SBT_Depth	0.462	0.756	0.571
6.	■ UpDoT	0.439	0.652	0.627
7.	★ CoDeT	0.372	0.597	0.594
8.	● DiMP	0.336	0.623	0.496

Table 8. Results for the six submitted VOT-D2022 trackers. For comparison, the table also includes the results for the recent dept-only tracker DOT [63] and RGB DiMP that was trained with RGB but tested with colormap converted depth images.

The VOT-D2022 challenge winner The three best depth-only trackers, MixFormerD, RSDiMP and OSTrack_D, perform on par, but since MixFormerD obtains the best EAO score, it is selected as the winner.

5 Conclusions

Results of the VOT2022 challenge were presented. The challenge is composed of the following challenges focusing on various tracking aspects and domains: (i) the segmentation-based short-term RGB tracking challenge (VOT-STs2022), (ii) the legacy bounding-box-based short-term RGB tracking challenge (VOT-STb2022), (iii) the realtime counterpart of VOT-STs2022 (VOT-RTs2022), (iv) the realtime countrepart of VOT-STb2022 (VOT-RTb2022), (v) the VOT2022 long-term RGB tracking challenge (VOT-LT2021), (vi) the VOT2022 short-term

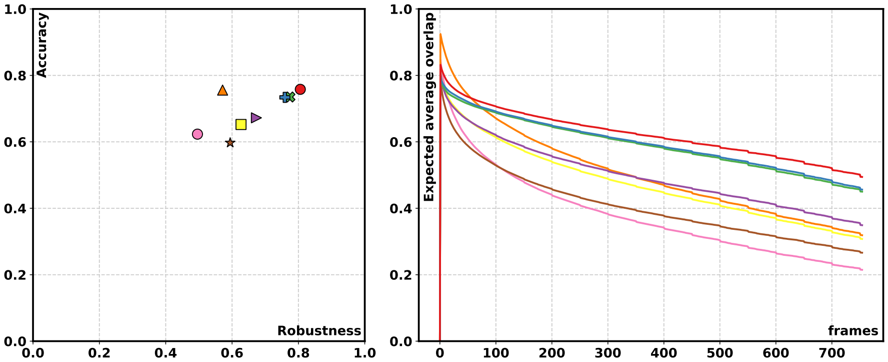

Fig. 11. The VOT-D2022 AR plot (left) and the EAO curves (right).

RGB and depth (D) tracking challenge (VOT-RGBD2022) and its variation (vii) the VOT2022 short-term depth-only tracking challenge (VOT-D2022).

In this VOT edition, new VOT-LT2022, VOT-RGBD2022 and VOT-D2022 datasets were introduced, a legacy bounding-box-based tracking challenge VOT-STb2022 was reintroduced, the VOT-ST2022 public and sequestered datasets were refreshed, and a training dataset has been introduced for VOT-LT2022.

A methodological shift, which was indicated already in the VOT2021 [33], has been made even more aparent this year. Nearly half of the trackers participating in VOT-STs2022 challenge were based on transformers, approximately 40% were using discriminative correlation filters, while only few were based on Siamese correlation trackers (a methodology highly popular in VOT2021). All of the top 9 trackers were based on transformers. Apart from being robust, these trackers are also very fast – 9 of top VOT-STs2022 trackers are among the top trackers on VOT-RTs2022 challenge. Variations of the segmentation trackers were submitted to the legacy bounding-box tracking challenge VOT-STb2022. Seven of the top ten trackers on VOT-STb2022 were modifications of the trackers ranked among the top ten on VOT-STs2022. The winner of the VOT-STs2022 challenge is MS_AOT A.6, while the winner of the VOT-STb2022 challenge in the bounding box ground truth category is OSTrackSTB B.14 and the winner in the segmentation mask ground truth category is APMT_MR B.2. The winner of the VOT-RTs2022 challenge is MS_AOT A.6 and the winner of the VOT-RTb2022 challenge is OSTrackSTB B.14.

The VOT-LT2022 challenge's top-three performers all apply Transformer-based tracker structure for short-term localization and long-term re-detection. Among all submitted trackers, the dominant methodologies are SuperDiMP [3], STARK [60], KeepTrack [46], and MixFormer [10]. The top perfomer and the winner of the VOT-LT2022 is VITKT_M C.8, which ensembles the results of VitTrack and KeepTrack. This tracker obtains a significantly better performance than the second-best tracker.

In the VOT-RGBD2022 and VOT-D2022 challenges, the same tracker architecture obtained the best results in all tracking metrics. There are two interesting points in this this submission that possibly explain its success as compared to others. At first, the tracker is based on the recent Convolutional Visual Transformer (CvT) model and, secondly, the both RGB and depth representations are learned from data. Since there are no depth-only tracking datasets that are sufficiently large for network training, the existing RGB datasets were converted to pseudo depth map datasets using a monocular depth estimation method. These design choices turned out to be the winning ones this year, and therefore the same authors won the VOT-RGBD2022 and VOT-D2022 challenges with their two trackers adopting the same architecture, MixForRGBD and MixFormerD.

For the last decade, the primary objective of VOT has been to establish a platform for discussion of tracking performance evaluation and contributing to the tracking community with verified annotated datasets, performance measures and evaluation toolkits. The VOT2022 was the tenth effort toward this, following the very successful VOT2013, VOT2014, VOT2015, VOT2016, VOT2017, VOT2018, VOT2019, VOT2020 and VOT2021. Since its beginning, the VOT has successfully identified modern milestone tracking methodologies at their inception, spanning discriminative correlation filters, Siamese trackers and most recently the transformer-based architectures. By pushing the boundaries, presenting ever challenging sequences and opening new challenges, the VOT has been successfully fulfilling its service to community. The effort, however, is joint with the tracking community who continually raises to the challenges and is the one generating the fast pace of tracker architecture development. We thank the community for their collaboration and look forward to future developments in this exciting scientific field.

Acknowledgements

This work was supported in part by the following research programs and projects: Slovenian research agency research program P2-0214 and project J2-2506. The challenge was sponsored by the Faculty of Computer Science, University of Ljubljana, Slovenia. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), in particular in terms of the Wallenberg research arena for Media and Language, and the Berzelius cluster at NSC, both funded by the Knut and Alice Wallenberg Foundation, as well as by ELLIIT, a strategic research environment funded by the Swedish government. Besides, this work was partially supported by the Fundamental Research Funds for the Central Universities (No. 226-2022-00051). This work has also received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 899987. Hyung Jin Chang and Aleš Leonardis were supported by the Institute of Information and communications Technology Planning and evaluation (IITP) grant funded by the Korea government (MSIT) (2021-0-00537). Gustavo Fernández was supported by the AIT Strategic Research Program 2022 Visual Surveillance and Insight.

	baseline			realtime			unsupervised
	EAO	AR		EAO	AR		Avg. acc.
Tracker	EAO	A	R	EAO	A	R	AUC
●MS_AOT	0.673①	0.781③	0.944①	0.610①	0.751③	0.921①	0.734②
✚DAMTMask	0.624②	0.796②	0.891②	0.369	0.623	0.756	0.765①
✖MixFormerM	0.589③	0.799①	0.878③	0.521	0.778①	0.834	0.722③
▶OSTrackSTS	0.581	0.775	0.867	0.569②	0.766②	0.860	0.665
▲Linker	0.559	0.772	0.861	0.467	0.709	0.811	0.697
■SRATransTS	0.547	0.743	0.866	0.547③	0.743	0.866②	0.673
★TransT_M	0.542	0.743	0.865	0.542	0.742	0.865③	0.667
●DGformer	0.538	0.744	0.861	0.527	0.744	0.850	0.668
✚TransLL	0.530	0.735	0.861	0.509	0.733	0.846	0.649
✖LWL_B2S	0.516	0.736	0.831	0.458	0.715	0.800	0.616
▶rts	0.502	0.710	0.843	0.450	0.698	0.802	0.592
▲TransT	0.500	0.749	0.815	0.500	0.749	0.815	0.611
■D3Sv2	0.497	0.713	0.827	0.307	0.648	0.627	0.553
★TRASFUSTm	0.491	0.740	0.805	0.342	0.631	0.705	0.616
●SuperDiMP_AR	0.468	0.734	0.773	0.426	0.711	0.746	0.580
✚LWL	0.461	0.721	0.798	0.406	0.699	0.763	0.582
✖SiamUSCMix	0.449	0.702	0.776	0.449	0.702	0.776	0.502
▶KYS_AR	0.445	0.722	0.749	0.397	0.702	0.708	0.574
▲DiMP_AR	0.426	0.723	0.719	0.422	0.718	0.724	0.547
■PrDiMP_AR	0.422	0.723	0.724	0.400	0.707	0.706	0.581
★ATOM_AR	0.398	0.699	0.691	0.380	0.695	0.681	0.505
●DAMT	0.339	0.459	0.861	0.315	0.447	0.841	0.434
✚transt_ar	0.315	0.474	0.809	0.315	0.474	0.809	0.397
✖STM	0.308	0.745	0.589	0.288	0.703	0.577	0.455
▶SwinT	0.247	0.452	0.679	0.243	0.440	0.680	0.369
▲ASMS	0.193	0.419	0.565	0.192	0.419	0.561	0.256
■CSRDCF	0.184	0.406	0.550	0.183	0.406	0.548	0.236
★SiamFC	0.174	0.417	0.502	0.165	0.423	0.462	0.232
●ANT	0.169	0.399	0.482	0.142	0.386	0.407	0.224
✚KCF	0.165	0.423	0.487	0.165	0.423	0.486	0.162
✖LGT	0.137	0.345	0.451	0.102	0.335	0.331	0.162

Table 10. Results for VOT-STs2022 and VOTs-RT2022 challenges. Expected average overlap (EAO), accuracy and robustness are shown. For reference, a no-reset average overlap AO [58] is shown under *Unsupervised*.

	baseline			realtime			unsupervised
	EAO	AR		EAO	AR		Avg. acc.
Tracker	EAO	A	R	EAO	A	R	AUC
●MixFormerL	0.602①	0.831①	0.859	0.512	0.792②	0.814	0.708
✚DAMT	0.602②	0.776	0.887①	0.554	0.752	0.866	0.716③
✖OSTrackSTB	0.591③	0.790	0.869	0.591①	0.790	0.869	0.680
▶APMT_MR	0.591	0.787	0.877③	0.560	0.768	0.871③	0.686
▲MixFormer	0.587	0.797②	0.874	0.580③	0.796①	0.872②	0.696
■APMT_RT	0.581	0.787	0.877②	0.581②	0.787	0.877①	0.721②
★ADOTstb	0.569	0.775	0.862	0.282	0.672	0.533	0.735①
●SRATransT	0.560	0.764	0.864	0.560	0.764	0.864	0.670
✚Linker_B	0.560	0.789	0.844	0.510	0.766	0.823	0.684
✖TransT_M	0.537	0.765	0.849	0.537	0.765	0.849	0.639
▶vittrack	0.536	0.789	0.818	0.536	0.789	0.818	0.679
▲SuperFus	0.534	0.763	0.828	0.481	0.760	0.782	0.629
■SwinTrack	0.524	0.788	0.803	0.503	0.783	0.791	0.626
★SBT	0.522	0.791③	0.813	0.523	0.791③	0.814	0.641
●TRASFUST	0.514	0.754	0.833	0.401	0.704	0.734	0.674
✚TransT	0.512	0.781	0.800	0.513	0.781	0.800	0.641
✖tomp	0.511	0.752	0.818	0.478	0.728	0.796	0.628
▶oceancycle	0.487	0.702	0.825	0.332	0.658	0.629	0.556
▲GOANET	0.481	0.759	0.772	0.478	0.758	0.768	0.654
■SuperDiMP	0.478	0.742	0.788	0.478	0.736	0.798	0.606
★KYS	0.461	0.688	0.797	0.446	0.690	0.784	0.576
●SiamUSCMix	0.444	0.702	0.773	0.444	0.702	0.773	0.556
✚PrDiMP	0.437	0.725	0.755	0.435	0.723	0.751	0.564
✖ViTCRT	0.433	0.774	0.711	0.434	0.774	0.711	0.609
▶DiMP	0.430	0.689	0.760	0.434	0.689	0.761	0.546
▲SiamVGGpp	0.399	0.719	0.690	0.399	0.719	0.690	0.486
■ATOM	0.386	0.668	0.716	0.391	0.672	0.728	0.505
★NfS	0.384	0.688	0.681	0.384	0.688	0.681	0.456
●TransLL	0.367	0.516	0.859	0.353	0.514	0.842	0.473
✚D3Sv2	0.356	0.521	0.811	0.242	0.486	0.637	0.414
✖DGformer	0.337	0.497	0.831	0.332	0.497	0.824	0.462
▶ReptileFPN	0.295	0.572	0.644	0.180	0.526	0.395	0.393
▲FSC2F	0.279	0.554	0.621	0.263	0.557	0.586	0.327
■DeepTCLCF	0.274	0.550	0.601	0.274	0.550	0.601	0.331
★TCLCFcpp	0.267	0.550	0.598	0.267	0.550	0.598	0.329
●ASMS	0.255	0.526	0.599	0.254	0.526	0.594	0.317
✚SiamFC	0.255	0.562	0.543	0.243	0.565	0.505	0.308
✖CSRDCF	0.251	0.519	0.580	0.250	0.518	0.577	0.300
▶KCF	0.239	0.542	0.532	0.240	0.541	0.533	0.234
▲ANT	0.209	0.492	0.484	0.172	0.471	0.414	0.226
■LGT	0.195	0.461	0.486	0.148	0.422	0.366	0.231

Table 12. Results for VOT-STb2022 and VOTb-RT2022 challenges. Expected average overlap (EAO), accuracy and robustness are shown. For reference, a no-reset average overlap AO [58] is shown under *Unsupervised*.

A VOT-STs2022 and VOT-RTs2022 submissions

This appendix provides a short summary of trackers considered in the VOT-ST2022 and VOT-RT2022 tracking by segmentation challenges VOT-STs2022 and VOT-RTs2022.

A.1 DAMT (DAMT)

Z. Fu, L. Wang, J. Sun, X. Li, Q. Deng, D. Du, M. Zheng
{fuzhihong.2022, wangliangliang.makalo, sunjingna, lixiao.dlut, dengqili, dukang.daniel, zhengmin.666}@bytedance.com

We propose a distractor-aware multi-head tracker, called DAMTMask. The tracker is built on top of MixFormer [10], adopting the online classifier of SuperDiMP [3] to coarsely localize targets. Besides, inspired by the excellent object detector RepPoint [64], we design a high-precision head named as *RepHead* for target state estimation. Furthermore, a distractor-aware strategy is proposed to determine whether there are distractors in the current frame. If distractors exist, we use target bounding boxes outputted by *RepHead* to generate training samples and update the classifier.

A.2 DAMTMask (DAMTMask)

L. Wang, Z. Fu, G. Chen, W. Xu, Q. Deng, D. Du, M. Zheng
{wangliangliang.makalo, fuzhihong.2022}@bytedance.com, chenguangqi@bupt.cn, {xuwei.cv, dengqili, dukang.daniel, zhengmin.666}@bytedance.com

This tracker is based on tracker DAMT (A.1). We extend it by using object segmentation to predict target masks.

A.3 Distance Guided Accurate Transformer Tracking (DGformer)

C. Zhang
chunhui.zhang@sjtu.edu.cn

This model is the extension of the TransT tracker [8]. Inspired by recent UAV tracking model [69], we introduce a distance constraint to smooth target localization changes during tracking. We also use the bounding box to refine the object segmentation mask.

A.4 Link Two Frames - Segmentation (Linker)

S. Di, Z. Xun, S. Liu
{dishangzhe, xunzz, liusi}@buaa.edu.cn

We adopt ViT-based [16] as our backbone, which 1) locates the target by associating the current frame with a previous one, and 2) alleviates propagation of uncertainty by reference to the initial template. Besides, we adopt a score branch to measure the tracking accuracy and update the previous frame once its score is above a threshold. We use the Candidate Elimination mechanism from OSTrack to eliminate non-target tokens, we adopt the FCN head from OSTrack, which outputs a classification score map, a local offset map, and a box size map and finally we use AlphaRefine [61] to output the tracking mask. To measure the tracking accuracy we use a score branch similar to MixFormer [10].

A.5 MixFormer-ViT-Base-Seg: End-to-End Tracking and Segmentation with Iterative Mixed Attention (MixFormerM)

C. Jiang, Y. Cui, G. Wu, L. Wang
{mg1933027, cuiyutao}@smail.nju.edu.cn, {gswu, lmwang}@nju.edu.cn

MixFormer-ViT-Base-Seg (MixFormerM) consists of two stages which perform MixFormer-based tracking and segmentation respectively. Our core design is to utilize the flexibility of attention operations, and propose a Mixed Attention Module (MAM) for simultaneous feature extraction and target information integration. Based on MAM, we build our MixFormer tracking framework simply by stacking multiple MAMs with progressive patch embedding and placing a corner head and segmentation head on top. We devise an asymmetric attention scheme in MAM to reduce computational cost, and propose an effective score prediction module to select high-quality templates. MixFormer-ViT-Base-Seg is constructed based on ViT-Base pertained with MAE.

A.6 MS-AOT: Associating Objects with Multi-scale Transformers for Video Object Segmentation (MS_AOT)

Z. Yang, Y. Cheng, Y. Xu, C. Sun, Y. Yang, Y. Zhuang
{yangzongxin, 22151080}@zju.edu.cn, yuanyouxu@outlook.com, {c_sun, yangyics, yzhuang}@zju.edu.cn

The MS-AOT tracker is built based on AOT [66,65,67], a transformer-based video object segmentation method, by applying transformers in multiple feature scales. AOT applies the long short-term transformer (LSTT) [66], which is responsible for propagating the object masks from past frames to the current frame, in the feature scale with a stride of 16. MS-AOT additionally applies LSTT in a finer feature scale with a stride of 8, leading to better performance on small objects. The backbone of MS-AOT is ResNet-50 pre-trained on ImageNet, and we trained MS-AOT on COCO, YouTube-VOS [59], and VIPSeg [47]. MS-AOT has two inference manners. For most of the objects, MS-AOT tracks and segments the objects in an end-to-end way by propagating objects' masks. For a tiny object smaller than 1/900 of the frame size, Mixformer [10] is used to track the object in the current frame and predict a coarse location. Then, a local region surrounding the coarse prediction is cropped and forwarded to MS-AOT for segmenting an accurate segmentation result. Finally, we update the internal template/memory of MixFormer/MS-AOT regarding the segmentation. Besides, the backbone of MixFormer is CvT [57], and we used the official MixFormer checkpoint (`https://github.com/MCG-NJU/MixFormer`).

A.7 One-stream tracker with online template updating for segmentation (OSTrackSTS)

B. Ye, H. Chang, B. Ma, S. Shan, X. Chen
botao.ye@vipl.ict.ac.cn, changhong@ict.ac.cn, bpma@ucas.ac.cn,
{*sgshan, xlchen*}*@ict.ac.cn*

OSTrackSTS is based on the one-stream tracker OSTrack [68], which employs a ViT-Base model for joint feature learning and relation modeling. Compared to the original OSTrackOSTrack [68], we have made several changes. First, we change the input resolution of the search region to 320. Then, a target confidence prediction head is added, which consists of two Transformer layers and an MLP for predicting the target confidence score. We update an online template once the target confidence score is greater than a given threshold β and the main lobe area of the classification graph is smaller than γ. This design tackles the problem of target appearance change. To generate segmentation masks, AlphaRefine [61] is used as a post-processing network. In contrast to the original model, we use a model trained with additional pseudo-mask labels, which was generated by running the original AlphaRefine model on LaSOT and GOT-10K.

A.8 Robust Visual Tracking by Segmentation (RTS)

M. Paul, M. Danelljan, C. Mayer, L. Van Gool
{*paulma, damartin, chmayer, vangool*}*@vision.ee.ethz.ch*

RTS [50] is a unified tracking architecture capable of predicting accurate segmentation masks. To design a *segmentation-centric* approach, we take inspiration from the VOS method LWL [5]. However, to achieve robust and accurate segmentation on tracking datasets, we propose several new components. In particular, we propose an instance localization branch that is trained to predict a target appearance model, which allows the detection of occlusions and to identify the correct target even in cluttered scenes. The output of the instance localization branch is further used to condition the high dimensional mask encoding. This enables the segmentation decoder to focus on the localized target, leading to a more robust mask prediction. Since, our proposed method contains a segmentation and instance memory that needs to be updated with previous tracking results, we design a memory management module. This module first assesses the prediction quality, decides whether the sample should enter into the memory and triggers the tracking model if it should be updated. See [50] for more details.

A.9 Uncertainty-aware Semantic Consistency Siamese Tracker via Mixed Cross Correlation (SiamUSCMix)

J. Ma, B. Zhong, X. Lan, R. Ji, X. Li
majie@stu.hqu.edu.cn, bnzhong@gxnu.edu.cn, xiangyuanlan@life.hkbu.edu.hk,
rrji@xmu.edu.cn, lixx@gxnu.edu.cn

We propose a novel offline tracker (named SiamUSCMix), which consists of a Mixed Cross Correlation module and an Uncertainty-aware Semantic Consistency Siamese Tracker (SiamUSC). Mixed Cross Correlation module aims to explore the perception abilities based on Siamese trackers. As such, we design a mixed feature extractor that fuses template and search features in three diverse approaches. Moreover, D3S is employed to produce a mask prediction as the output.

A.10 Search Region Aware Transformer Tracking for Short-term Segmentation Tracking (SRATransTS)

J. Zhu, X. Chen, S. Lai, D. Wang, H. Lu
{jiawen, chenxin3131, laisimiao}@mail.dlut.edu.cn,
{wdice, lhchuan}@dlut.edn.cn

SRATransTS uses search region aware module (SRA) to obtain a rectified search region for frame-level online tracking. The idea is mainly from SRRT [72], and TransT_M [7] is employed as our basetracker. SRA is a siamese-style matching network, consists of a ResNet-34 backbone network, and a search region classification head. It takes template and X6 search region crop as the inputs and predicts the minimum search region size containing object. The predicted search region size, iou score and location from previous frame are as a comprehensive basis for whether zoom and offset the current frame's search region. The rectified search area can better adapt to the complex motion state of the target to a certain extent. Then the segmentation branch generates the segmentation result of the object. Finally, all the models used are speed up to achieve real-time speed by ONNX (`https://github.com/onnx/onnx`).

A.11 Swin Transformer Tracking (SwinT)

Q. Gu
736446296@qq.com

Based on TransT, we replace the feature extraction and fusion layer with Swin-Transformer-based layers. Furthermore, the feature fusion layer is implemented by a nested Swin Transformer layer called swincross which can combine the feature map of templete and search area in a sliding window way. Specifically, we split the template and the search region into small windows, calculate the attention value of the corresponding window through Swin Transformer, and then shift the pixels of sreach region and repeat the above steps.

A.12 Transformer Tracking with Light-weight Large Receptive Convolution Module (TransLL)

H. Yu, W. Yu, K. He, X. Chen, J. Wu, Y. Huang, L. Wang
{hongyuan.yu, weichen.yu, keji.he}@cripac.ia.ac.cn, chenxiuyi2017@ia.ac.cn,
{jinlin.wu, yhuang, wangliang}@nlpr.ia.ac.cn

In order to ensure a sufficiently large receptive field while reducing computation cost, we introduce a novel light-weight large receptive convolution module (LLconv) by using bilinear sampling on the input feature map to downsample/upsample the input size. To alleviate information loss, we utilize an extra bypassed 1 x 1 convolution module to perform as a high-resolution preserving residue. Thus, the input feature map is connected to the end of an zoomed convolution to preserve the high-resolution of the input. Our LLconv not only guarantees the speed, but also ensures the accuracy. On the basis of TransT [8], we use LLconv to build a segmentation branch. In addition, following ATOM [13], we also add the iou prediction branch to further improve the tracking performance. Our tracker works similarly to the ATOM family methods [13,3,8], please refer to them for more details.

A.13 Transformer Tracking (TransT)

X. Chen, B. Yan, J. Zhu, D. Wang, H. Lu, X. Yang
{chenxin3131, yan_bin, jiawen}@mail.dlut.edu.cn, {wdice, lhchuan}@dlut.edu.cn, xyang@remarkholdings.com

Transformer Tracking presents a transformer-based feature fusion network, which effectively combines the template and the search region features using attention mechanism. TransT [8] consists of three components: the siamese-like feature extraction backbone (ResNet50), the designed feature fusion network, and the prediction head. We extend our transformer tracking framework with a segmentation branch to generate an accurate mask. The segmentation branch fuses the output features of the feature fusion network with the low-level features of the backbone in the FPN style. For more details about TransT, the reader is referred to [8].

A.14 transt_ar (transt_ar)

K. Ben
kierkers@mail.dlut.edu.cn

The tracker takes TransT as the baseline, and adds some prediction heads to improve the accuracy and robustness of the tracker. In addition, it also adds some post-processing and Alpha Refine to further improve the performance of the tracker.

A.15 Multi-Template Transformer Tracking (TransT_M)

X. Chen, J. Zhu, B. Yan, D. Wang, H. Lu, X. Yang
{chenxin3131, jiawen, yan_bin}@mail.dlut.edu.cn,
{wdice, lhchuan}@dlut.edu.cn, xyang@remarkholdings.com

TransT_M [7] is a variant of TransT [8]. We add a segmentation branch, a Multi-Template design, and an IoU prediction head on TransT, forming an end-to-end framework. We concatenate two templates in the spatial dimension and input them into the template branch of TransT. IoU prediction head is a three-layer perceptron to predict the bounding box's IoU and control the updating of the template.

A.16 Tracking by Student FUSing Teachers and AlphaRefine (TRASFUSTm)

M. Dunnhofer, N. Martinel, C. Micheloni
{*matteo.dunnhofer, niki.martinel, christian.micheloni*}*@uniud.it*

The tracker TRASFUSTm is the combination of the TRASFUST bounding-box tracker refsec:tr:vot-ST-B-TRASFUST with the target-dependent segmentation generation method AlphaRefine [61]. After the bounding-box given by TRASFUST, AlphaRefine is run to obtain the segmentation mask of the target.

A.17 ANT (ANT)

Submitted by VOT Committee

The ANT tracker is a conceptual increment to the idea of multi-layer appearance representation that is first described in [53]. The tracker addresses the problem of self-supervised estimation of a large number of parameters by introducing controlled graduation in estimation of the free parameters. The appearance of the object is decomposed into several sub-models, each describing the target at a different level of detail. The sub models interact during target localization and, depending on the visual uncertainty, serve for cross-sub-model supervised updating. The reader is referred to [55] for details.

A.18 Accurate Tracking by Overlap Maximization (ATOM_AR)

Submitted by VOT Committee

This tracker employs the standard ATOM [13] for predicting bounding boxes. The AlphaRefine [61] network is then employed to predict the final mask as a post-processing step.

A.19 Learning Discriminative Model Prediction for Tracking (DiMP_AR)

G. Bhat, M. Danelljan, L. Van Gool, R. Timofte
{*goutam.bhat, martin.danelljan, vangool, timofter*}*@vision.ee.ethz.ch*
This tracker employs the standard DiMP [3] for predicting bounding boxes. The AlphaRefine [61] network is then employed to predict the final mask as a post-processing step.

A.20 Know your surroundings tracker with Alpha refine post-processing step (KYS_AR)

Submitted by VOT Committee

This tracker employs the standard KYS [4] for predicting bounding boxes. The AlphaRefine [61] network is then employed to predict the final mask as a post-processing step.

A.21 PrDiMP50 tracker with Alpha refine post-processing step (PrDiMP-50_AR)

Submitted by VOT Committee

This tracker employs the standard PrDiMP50 [14] for predicting bounding boxes. The AlphaRefine [61] network is then employed to predict the final mask as a post-processing step.

A.22 SuperDiMP50 tracker with Alpha refine post-processing step (SuperDiMP-50_AR)

Submitted by VOT Committee

This tracker employs the standard SuperDiMP50 [3,14,21,22] for predicting bounding boxes. The AlphaRefine [61] network is then employed to predict the final mask as a post-processing step.

A.23 Scale adaptive mean shift (ASMS)

Submitted by VOT Committee

The mean-shift tracker optimizes the Hellinger distance between template histogram and target candidate in the image. This optimization is done by a gradient descend. ASMS [56] addresses the problem of scale adaptation and presents a novel theoretically justified scale estimation mechanism which relies solely on the mean-shift procedure for the Hellinger distance. ASMS also introduces two improvements of the mean-shift tracker that make the scale estimation more robust in the presence of background clutter – a novel histogram colour weighting and a forward-backward consistency check. Code available at `https://github.com/vojirt/asms`.

A.24 Discriminative Correlation Filter with Channel and Spatial Reliability (CSRDCF)

Submitted by VOT Committee

The CSR-DCF [40] improves discriminative correlation filter trackers by introducing the two concepts: spatial reliability and channel reliability. It uses color segmentation as spatial reliability to adjust the filter support to the part of the object suitable for tracking. The channel reliability reflects the discriminative power of each filter channel. The tracker uses only HoG and colornames features. This is the C++ openCv implementation.

A.25 Discriminative Sing-Shot Segmentation Tracker v2 (D3Sv2)

A. Lukežič, J. Matas, M. Kristan
alan.lukezic@fri.uni-lj.si, matas@cmp.felk.cvut.cz, matej.kristan@fri.uni-lj.si

D3Sv2 is an extended version of the D3S [39]. The original method is extended in the following aspects: (i) a better backbone, (ii) channel attention mechanism in the upscaling modules in GIM, (iii) trainable MLP-based similarity computation in GIM, which replaces the 'handcrafted' top-K average operation and (iv) the new scale estimation module used for robust target size estimation.

A.26 Kernelized Correlation Filter (KCF)

Submitted by VOT Committee

This tracker is a C++ implementation of Kernelized Correlation Filter [24] operating on simple HOG features and Colour Names. The KCF tracker is equivalent to a Kernel Ridge Regression trained with thousands of sample patches around the object at different translations. It implements multi-thread multi-scale support, sub-cell peak estimation and replacing the model update by linear interpolation with a more robust update scheme. Code available at `https://github.com/vojirt/kcf`.

A.27 Local-Global Tracking tracker (LGT)

Submitted by VOT Committee

The core element of LGT is a coupled-layer visual model that combines the target global and local appearance by interlacing two layers. By this coupled constraint paradigm between the adaptation of the global and the local layer, a more robust tracking through significant appearance changes is achieved. The reader is referred to [53] for details.

A.28 Learning What to Learn for Video Object Segmentation (LWL)

G. Bhat, F. Järemo Lawin, M. Danelljan, A. Robinson, M. Felsberg, L. Van Gool, R. Timofte
goutam.bhat@vision.ee.ethz.ch, felix.jaremo-lawin@liu.se,
martin.danelljan@vision.ee.ethz.ch, {andreas.robinson, michael.felsberg}@liu.se,
{vangool, timofter}@vision.ee.ethz.ch

LWTL [5] is an end-to-end trainable video object segmentation VOS architecture which captures the current target object information in a compact parametric model. It integrates a differentiable few-shot learner module, which predicts the target model parameters using the first frame annotation. The learner is designed to explicitly optimize an error between target model prediction and a ground truth label, which ensures a powerful model of the target object. Given a new frame, the target model predicts an intermediate representation of the target mask, which is input to the offline trained segmentation decoder to generate the final segmentation mask. In order to guide the learner to focus on the most crucial aspect of the target, a network module is trained to predict spatial importance weights for different elements in the few-shot learning loss. See [5] for more details.

A.29 Learning what to learn tracker with Box2Seg (LWL-B2S)

Submitted by VOT Committee

This is the standard Learning What to Learn (LWL) [5] (A.28) video object segmentation and tracking approach, trained with the annotations generated by the approach [70]. That is, in addition to the YouTubeVOS and DAVIS training datasets, [70] is used to generate masks from bounding box annotated sequences in LaSOT and GOT10k. We then finetune LWL on the combined data. The same inference settings is used as in the standard LWL [5].

A.30 SiameseFC-AlexNet (SiamFc)

Submitted by VOT Committee

SiamFC [2] applies a fully-convolutional Siamese network [9] trained to locate an exemplar image within a larger search image. The architecture is fully convolutional with respect to the search image: dense and efficient sliding-window evaluation is achieved with a bilinear layer that computes the cross-correlation of two inputs. The deep convnet is first trained offline on the large ILSVRC15 [51] video dataset to address a general similarity learning problem, and then this function is evaluated during testing by a simplistic tracker. SiamFc incorporates elementary temporal constraints: the object search is done within a region of approximately four times its previous size, and a cosine window is added to the score map to penalize large displacements. SiamFc also processes several scaled versions of the search image, any change in scale is penalised and damping is applied to the scale factor.

A.31 VOS SOTA method (STM)

Submitted by VOT Committee

STM [49] is a VOS method employing a space-time memory module combined with a dot-product attention layer. Please see the original paper for details [49].

B VOT-STb2022 and VOT-RTb2022 submissions

This appendix provides a short summary of trackers considered in the VOT-ST2022 and VOT-RT2022 tracking by bounding box tracking challenges VOT-STb2022 and VOT-RTb2022.

B.1 Backward Tracking for Robust Template-update (ADOTstb)

D. Lee, S. Lee, Y. Chen, H. Lee, C. Park, S. Pan, J. Yu, Q. Wang
{dw23.lee, seoh.lee, yiwei.chen, hyem.lee, cb0372.park, siyang1.pan, jiaqian.yu, qiang.w}@samsung.com

The proposed object tracker consists of three parts: (i) main tracker, (ii) refinement module, and (iii) template updater. First, we utilized STARK [60] as our main framework for the object-tracking task. The backbone network is replaced by Swin Transformer [37]. Alpha-Refine [61] is applied for each of the bounding box output to improve the box estimation quality of the tracker. We proposed Backward Tracking for the robust template update without any additional network weights/training. We evaluate the quality of the new candidate template by backward tracking. The new candidate template is re-initialized by the current tracking result. Then we track backward to the previous frames with new candidate template. By quantifying the difference of the forward/backward tracking as a confidence value, we could update the template at appropriate timing without any additional network weights/training.

B.2 Adaptive Part Mining Tracker with Multi-Region (APMT_MR)

Y. Ma, D. Yang, Q. Yu, J. He, F. Wang, W. Li, T. Zhang
{*imyc, yangdawei, sa21010105, hejf, wangfei91, lwklwk*}*@mail.ustc.edu.cn, tzzhang@ustc.edu.cn*

Tracker APMT_MR is based on tracker APMT_RT B.3. APMT_RT crops single search region for tracking, which is cropped five times the side length from the center coordinate of the target in the previous frame. Differently, APMT_MR crops three search regions for tracking, including base region, motion region and scale region. The base region is same as the search region used in APMT_RT. The center of the motion region is derive by the displacement of the target between the last two frames. The side length of the scale region is determined by the speed of the target. Furthermore, we design an IoU head to estimate the IoU between the predict bounding box and the ground truth bounding box. If the tracking results of the three search region overlap highly, the box with the highest IoU score is taken as the tracking result, otherwise, the box with the highest confidence score is taken as the tracking result.

B.3 Adaptive Part Mining Tracker (APMT_RT)

Y. Ma, D. Yang, Q. Yu, J. He, F. Wang, W. Li, T. Zhang
{*imyc, yangdawei, sa21010105, hejf, wangfei91, lwklwk*}*@mail.ustc.edu.cn, tzzhang@ustc.edu.cn*

We propose an adaptive part mining tracker called APMT_RT that is composed of a feature extraction backbone, a feature fusion encoder, two decoders (adaptive part mining and object state estimation) and two heads (confidence score and distractor). Swin Transformer-Base [37] is used for feature extraction and a feature fusion encoder is used for spatio-temporal feature fusion. The adaptive part mining decoder is designed for object part division via a cross-attention mechanism. In the object state estimation decoder, each part feature predicts a keypoint, and we derive the bounding box from the mean value and the standard deviation of these keypoints. A confidence score head evaluates the quality of the predicted bounding box to control template update and a distractor head predicts whether distractors exist in the search region. If distractors exist, the drift of the target box is limited to avoid tracking failures. APMT_RT adaptively adjusts the size of the search region based on the speed of the target. We use ONNX (`https://github.com/onnx/onnx`) to speed up APMT_RT for real-time speed.

B.4 DAMT (DAMT)

Z. Fu, L. Wang, J. Sun, X. Li, Q. Deng, D. Du, M. Zheng
{fuzhihong.2022, wangliangliang.makalo, sunjingna, lixiao.dlut, dengqili, dukang.daniel, zhengmin.666}@bytedance.com

We propose a distractor-aware multi-head tracker, called DAMTMask. The tracker is built on top of MixFormer [10], adopting the online classifier of SuperDiMP [3] to coarsely localize targets. Besides, inspired by the excellent object detector RepPoint [64], we design a high-precision head named as *RepHead* for target state estimation. Furthermore, a distractor-aware strategy is proposed to determine whether there are distractors in the current frame. If distractors exist, we use target bounding boxes outputted by *RepHead* to generate training samples and update the classifier.

B.5 TCLCF tracker with deep CNN features (DeepTCLCF)

C. Tsai
chiyi_tsai@gms.tku.edu.tw

DeepTCLCF is the TCLCF tracker combined with a deep convolutional neural network. In the current implementation, we use the first six layers of Darknet19 pretrained model to extract deep features of the tracking target. Moreover, we use three different correlation filters to cooperatively track the same target. DeepTCLCF tracker requires GPU acceleration to achieve real-time performance. Here, we use Nvidia Geforce 1080ti GPU.

B.6 Distance Guided Accurate Transformer Tracking (DGformer)

C. Zhang
chunhui.zhang@sjtu.edu.cn

This model is the extension of the TransT tracker [8]. Inspired by recent UAV tracking model [69], we introduce a distance constraint to smooth target localization changes during tracking. We also use the bounding box to refine the object segmentation mask.

B.7 Fast Saliency-guided Continuous Correlation Filter-based tracker (FSC2F)

A. Memarmoghadam, P. Moallem
{a.memarmoghadam, p_moallem}@eng.ui.ac.ir

The recently proposed ECOhc approach [12] discriminatively tracks the target object via an efficient continuous correlation filter jointly learned over a compact historical sample set of the tracked object described by low-dimensional features. To improve its robustness, we equip the baseline ECOhc with a fast spatio-temporal saliency map constructed by the PQFT approach [20]. The PQFT model utilizes intensity, color, and motion features for quaternion representation of the search image context around the previous pose of the tracked object. Therefore, attentional regions in the coarse saliency map can constrain target confidence peaks. Moreover, to maintain the computational complexity in a reasonable range for real-time tracking, we propose a faster scale estimation algorithm by enhancing the fDSST method [15] via jointly learning the sparsely sampled scale-spaces.

B.8 Generic Occlusion Aware Network for Tracking (GOANET)

M. Dasari, R. Gorthi
{*ee18d001, rkg*}*@iittp.ac.in*

This network formulate occlusion status as a binary classification problem. It learns occlusion status at the frame level during offline training and estimate the same during online tracking. It uses the labels available in LaSOT [19] and VID datasets [51].

B.9 Link Two Frames (Box) (Linker_B)

S. Di, Z. Xun, S. Liu
{*dishangzhe, xunzz, liusi*}*@buaa.edu.cn*

This tracker is similar to the tracker Linker A.4. The difference lies on the refinement process: While Linker uses AlphaRefine [61] to output the tracking mask, Linker_B uses AlphaRefine [61] to output the bounding box.

B.10 MixFormer-ViT-Base: End-to-End Tracking with Iterative Mixed Attention (MixFormer)

Y. Cui, Y. Yang, T. Song, C. Jiang, G. Wu, L. Wang
{*cuiyutao, 181220064, 191098194, mg1933027*}*@smail.nju.edu.cn*,
{*gswu, lmwang*}*@nju.edu.cn*

This tracker is based on tracker MixFormerM (A.5). While MixFormerM is a two stage performing MixFormer-based tracking and segmentation, MixFormer is a one-stage tracker based on ViT-Base/16.

B.11 MixFormer-ViT-Large: End-to-End Tracking with Iterative Mixed Attention (MixFormerL)

Y. Cui, T. Song, Y. Yang, C. Jiang, G. Wu, L. Wang
{*cuiyutao, 191098194, 181220064, mg1933027*}*@smail.nju.edu.cn*,
{*gswu, lmwang*}*@nju.edu.cn*

This tracker is based on tracker MixFormerM (A.5). While MixFormerM is a two stage performing MixFormer-based tracking and segmentation, MixFormerL is a one-stage tracker based on ViT-Large/16.

B.12 Normalization free Siamese Network for Object Tracking (NfS)

H. Gupta, D. Jangid, O. P. Verma, L. Rout, D. Dhar
{*guptah.nitj, ee.deepak.jangid*}*@gmail.com, vermaop@nitj.ac.in,*
{*lr, deb*}*@sac.isro.gov.in*

The NfS Tracker adopts the Normalization Free ResNet (NFNet) architecture as backbone while it utilizes the head of SiamFC++. Additionally, the two networks have been made compatible by introducing a connecting sub-module network. The features have been extracted by NFNet whereas, SiamFC++ performs feature matching by measuring the correlation between the search space and the target. The head structure has classification and regression branches that have been responsible for classification between the one positive or negative patch and refinement of the predicted bounding box respectively. The NfS has been trained on various large scale datasets (GOT10K[59], COCO, ImageNET VID, DET, LaSOT and TrackingNet). The network has been trained for 50 epochs with the learning policy of cosine annealing. The maximum EAO (0.384) has been achieved at the 49th epoch.

B.13 Neighbor ocean with cycle consistency (oceancycle)

Y. Chen
franktpmvu@iis.sinica.edu.tw

We propose a new method to simultaneously track the targets and the objects near the targets called NeighborTrack. If the target is occluded by the neighbors, the chance of wrong matching is reduced because the neighbors will tend to match with the neighbor trajectory rather than the target trajectory. Our method includes two steps. First, we calculate the average IoU between the historical trajectories of neighbors and targets. The neighbor is defined as the target which confidence score is higher than a threshold. The historical trajectory of the neighbor is obtained by tracking the neighbor in the reverse time axis and the historical trajectory of the target is obtained from the tracking result of each historical frame. Second, we apply the Hungarian algorithm to find the best matches of the neighbors and targets. The neighbor trajectories will continue to be used as historical trajectories in the trajectory matching process until there are no new neighbors being matched. The two steps will repeat until the trace is complete.

B.14 One-stream tracker with online template updating (OSTrackSTB)

B. Ye, H. Chang, B. Ma, S. Shan, X. Chen
botao.ye@vipl.ict.ac.cn, changhong@ict.ac.cn, bpma@ucas.ac.cn,
{*sgshan, xlchen*}*@ict.ac.cn*

[59] GOT10k does not use the prohibited 1k sequences and the training dataset follows all the given regulations on the website.

OSTrackT is the same as OSTrackTM A.7, except that the segmentation phase is removed.

B.15 ReptileFPN Meta-Tracker (ReptileFPN)

C. Tsai
chiyi_tsai@gms.tku.edu.tw

ReptileFPN is a tracker based on FPN model and a meta-learning technique called Reptile. Inspired by Reptile Meta-Tracker, we trained a deep learning network offline by repeatedly sampling different tasks. The resulting network can quickly adapt to any domain without the need to train multi-domain branches like MDNet. The original architecture from Reptile Meta-Tracker used VGG like backbone, here we modified it using FPN to further improve the feature extraction ability. During online initialization, the ReptileFPN tracker only requires a few training examples from the first frame and a few steps of optimization to perform well in online tracking. See [25] for more details.

B.16 Correlation-Aware Deep Tracking: Single Branch Transformer Tracking-base (SBT)

F. Xie
jaffe0319@gmail.com

We propose a novel target-dependent feature network inspired by the self-cross-attention scheme. In contrast to the Siamese-like feature extraction, our network deeply embeds cross-image feature correlation in multiple layers of the feature network. By extensively matching the features of the two images through multiple layers, it is able to suppress non-target features, resulting in instance varying feature extraction. The output features of the search image can be directly used for predicting target locations without extra correlation step. Moreover, our model can be flexibly pre-trained on abundant unpaired images, leading to notably faster convergence than the existing methods.

B.17 Uncertainty-aware Semantic Consistency Siamese Tracker via Mixed Cross Correlation (SiamUSCMix)

J. Ma, B. Zhong, X. Lan, R. Ji, X. Li
majie@stu.hqu.edu.cn, bnzhong@gxnu.edu.cn, lanxy@pcl.ac.cn,
rrji@xmu.edu.cn, lixx@gxnu.edu.cn

We propose a novel offline tracker (named SiamUSCMix), which consists of a Mixed Cross Correlation module and an Uncertainty-aware Semantic Consistency Siamese Tracker (SiamUSC). Mixed Cross Correlation module aims to explore the perception abilities based on Siamese trackers. As such, we design a mixed feature extractor that fuses template and search features in three diverse approaches.

B.18 VGG-16 based Siamese Tracker (SiamVGGpp)

H. Gupta, D. Jangid, O. P. Verma, L. Rout, D. Dhar
{*guptah.nitj, ee.deepak.jangid*}*@gmail.com, vermaop@nitj.ac.in,*
{*lr, deb*}*@sac.isro.gov.in*

SiamVGGpp data pipeline consists of two subnetworks: the first one (i.e. backbone) is responsible for feature extraction and a modified VGG-16 has been employed. The second subnetwork (i.e. head) is responsible for feature matching and it utilizes 3 convolution layers with kernel size of 3x3 to perform the correlation operation between the search area and exemplar. The head structure has a classification branch for classification between the one positive or negative patch. It also has a regression branch for the refinement of the predicted bounding box. SiamVGGpp tracker was trained on GOT10k[60], COCO, ImageNET VID, DET, LaSOT and TrackingNet. The network was trained on 40 epochs in total and the adopted learning rate policy was cosine annealing. The maximum EAO (0.399) has been achieved at the 36th epoch.

B.19 Search Region Aware Transformer Tracking for Short-term Bounding-box Tracking (SRATransT)

J. Zhu, X. Chen, S. Lai, D. Wang, H. Lu
{*jiawen, chenxin3131, laisimiao*}*@mail.dlut.edu.cn,*
{*wdice, lhchuan*}*@dlut.edn.cn*

SRATransT uses search region aware module (SRA) to obtain a rectified search region for frame-level online tracking. The idea is mainly from SRRT [72], and TransT_M [7] is employed as our basetracker. SRA is a siamese-style matching network consisting of a ResNet-34 backbone network and a search region classification head. It takes template and X6 search region crop as the inputs and predicts the minimum search region size containing the target. The rectified search area can better adapt to the complex motion state of the target to a certain extent. After obtaining the prediction results of the base tracker, we used Alpha-Refine [61] to improve the accuracy of low scoring boxes. Specifically, we replace the results of boxes whose IoU score is lower than a given threshold γ with the prediction results of Alpha-Refine [61]. Finally, all the used models are speed up using ONNX (`https://github.com/onnx/onnx`) to achieve real-time speed.

B.20 A Decision-level Fusion of Multiple Discriminative Trackers (SuperFus)

S. Zhao, J. Chen, Z. Tang, X. Zhu, T. Xu, X. Wu, J. Kittler, H. Li, X. Li, K. Ze
7201905026@stu.jiangnan.edu.cn, jamescjy98@gmail.com,

[60] GOT10k does not use the prohibited 1k sequences and the training dataset follows all the given regulations on the website.

{zhangyong_tang_jnu, xuefeng_zhu95, tianyang_xu, xiaojun_wu_jnu}@163.com,
j.kittler@surrey.ac.uk, lihui.cv@jiangnan.edu.cn, xilizju@zju.edu.cn,
6213113073@stu.jiangnan.edu.cn

For the outputs of the STARK are top-left and bottom-right response maps, we first transform the top-left and bottom-right response maps to center response map. Then, to align the semantics between SuperDiMP and STARK, responses from two algorithms are interpolated to pixel-level heatmaps. After padding to the same size with the pixel-level heatmap of SuperDiMP, the heatmap of STARK is fused with SuperDiMP with a hyper-parameter weight. Finally, the fused response is obtained by interpolating the pixel-level heatmap to response-level.

B.21 SwinTrack-Base (SwinTrack)

L. Lin, H. Fan, Z. Zhang, Y. Xu, H. Ling
l.lt@mail.scut.edu.cn, heng.fan@unt.edu, zhangzhipeng2017@ia.ac.cn,
yxu@scut.edu.cn, hling@cs.stonybrook.edu

We provide the official results of SwinTrack. SwinTrack is based on the Siamese network architecture. Four main components comprise our fully attentional tracker: the Swin-Transformer backbone, the attentional encoder-decoder network, positional encoding, and the head network. During tracking, the backbone network extracts the features of the template image patch and the search region image patch separately with shared weights, the encoder network (self-attention based) fuse the feature tokens from the template image and the search image by concatenation, and enhances the concatenated tokens layer-by-layer by attention mechanism, positional encoding helps the model to distinguish the tokens from the different source and the different position, the decoder network (cross-attention based) generates the final feature map of the search image and feeds it to the head network to obtain the IoU-Aware classification response map and bounding box estimation map.

B.22 Ensemble correlation filter tracking based on temporal confidence learning (TCLCFcpp)

C. Tsai
chiyi_tsai@gms.tku.edu.tw

TCLCF is a real-time ensemble correlation filter tracker based on temporal confidence learning. In the current implementation, we use two different correlation filters to cooperatively track the same target. TCLCF tracker is a high-speed and robust generic object tracker that does not require GPU acceleration. Therefore, it can be implemented on embedded platforms with limited computing resources.

B.23 Transforming Model Prediction for Tracking (tomp)

C. Mayer, M. Danelljan, G. Bhat, M. Paul, D. Paudel, F. Yu, L. Van Gool
{chmayer, martin.danelljan, goutam.bhat, paulma, paudel}@vision.ee.ethz.ch, fisheryu@ethz.ch, vangool@vision.ee.ethz.ch

We propose a tracker architecture employing a Transformer-based model prediction module. Transformers capture global relations with little inductive bias, allowing it to learn the prediction of more powerful target models. We further extend the model predictor to estimate a second set of weights that are applied for accurate bounding box regression. The resulting tracker tomp relies on training and on test frame information in order to predict all weights transductively. We train the proposed tracker end-to-end and validate its performance by conducting comprehensive experiments on multiple tracking datasets. For more details please see [45].

B.24 Transformer Tracking with Light-weight Large Receptive Convolution Module (TransLL)

H. Yu, W. Yu, K. He, X. Chen, J. Wu, Y. Huang, L. Wang
{hongyuan.yu, weichen.yu, keji.he}@cripac.ia.ac.cn, chenxiuyi2017@ia.ac.cn, jinlin.wu@nlpr.ia.ac.cn, {yhuang, wangliang}@nlpr.ia.ac.cn

In order to ensure a sufficiently large receptive field while reducing computation cost, we introduce a novel light-weight large receptive convolution module (LLconv) by using bilinear sampling on the input feature map to downsample/upsample the input size. To alleviate information loss, we utilize an extra bypassed 1 x 1 convolution module to perform as a high-resolution preserving residue. Thus, the input feature map is connected to the end of an zoomed convolution to preserve the high-resolution of the input. Our LLconv not only guarantees the speed, but also ensures the accuracy. On the basis of TransT [8], we use LLconv to build a segmentation branch. In addition, following ATOM [13], we also add the iou prediction branch to further improve the tracking performance. Our tracker works similarly to the ATOM family methods [13,3,8], please refer to them for more details.

B.25 Transformer Tracking (TransT)

X. Chen, B. Yan, J. Zhu, D. Wang, H. Lu, X. Yang
{chenxin3131, yan_bin, jiawen}@mail.dlut.edu.cn,
{wdice, lhchuan}@dlut.edu.cn, xyang@remarkholdings.com

Transformer Tracking presents a transformer-based feature fusion network, which effectively combines the template and the search region features using attention mechanism. TransT [8] consists of three components: the siamese-like feature extraction backbone (ResNet50), the designed feature fusion network, and the prediction head. For more details about TransT, the reader is referred to [8].

B.26 Multi-Template Transformer Tracking (TransT_M)

X. Chen, J. Zhu, B. Yan, D. Wang, H. Lu, X. Yang
{*chenxin3131, jiawen, yan_bin*}*@mail.dlut.edu.cn,*
{*wdice, lhchuan*}*@dlut.edu.cn, xyang@remarkholdings.com*

TransT_M [7] is a variant of TransT [8]. We add a Multi-Template design, and an IoU prediction head on TransT, forming an end-to-end framework. We concatenate two templates in the spatial dimension and input them into the template branch of TransT. IoU prediction head is a three-layer perceptron to predict the bounding box's IoU and control the updating of the template.

B.27 Tracking by Student FUSing Teachers (TRASFUST)

M. Dunnhofer, N. Martinel, C. Micheloni
{*matteo.dunnhofer, niki.martinel, christian.micheloni*}*@uniud.it*

The TRASFUST tracker [17] consists of two components: (i) a fast processing CNN-based model called the Student; (ii) a pool of off-the-shelf trackers, i.e. the Teachers. The Student, which has the form of a deep regression tracker, is trained offline based on a learning scheme which combines knowledge distillation (KD) and reinforcement learning (RL). Relevant tracking knowledge is acquired through KD from multiple trackers considered as Teachers. After learning, at every frame of a video, the Student is capable of evaluating and selecting the most accurate target localization (as a bounding-box) predicted by the teachers in the pool. In this way, fusion of the underlying teacher trackers is achieved. In this submission, the SuperDiMP [3] and Stark [60] trackers compose the pool of Teachers.

B.28 Tracking Vision Transformer With Class and Regression Tokens (ViTCRT)

E. Di Nardo, A. Ciaramella
{*emanuel.dinardo, angelo.ciaramella*}*@uniparthenope.it*

ViTCRT is a Siamese tracking. The tracker uses a ResNet50 as backbone for feature extraction (pre-trained on ImageNet) extracting the features from the 3rd layer and adjusting the output feature space using sub-sampling. Features are flattened, concatenated and passed to a Vision Transformer (ViT) [16] architecture. Differently from ViT, ViTCRT uses a second token for bounding box regression. At the end a separate MLP is used for each token in order to give an appropriate specialization on a specific task (i.e. classification/regression). Due to the nature of the tokens, that are a unique representation of the inputs, they are augmented following the methodology proposed in STARK [60]. Differently from STARK the importance is given to the tokens. Regression follows the Alpha-Refine [61] method. Foreground/Background classification allows the network to learn a similar representation.

B.29 Visual Object Tracking via Vision Transformer (vittrack)

X. Chen, J. Zhao, H. Peng, D. Wang, H. Lu
{*chenxin3131, zj982853200*}*@mail.dlut.edu.cn, Houwen.Peng@microsoft.com,*
{*wdice, lhchuan*}*@dlut.edu.cn*

Vittrack is composed with a ViT backbone network and a corner prediction head. We concatenate the template and the search region in spatial channel after patch embedding. Then using ViT to extract the features and using the corner prediction head to predict the bounding box. We use the MAE [23] pre-trained parameters of ViT, and fine-tune them on tracking dataset.

B.30 ANT (ANT)

Submitted by VOT Committee

The reader is referred to A.17 for details.

B.31 Scale adaptive mean shift (ASMS)

Submitted by VOT Committee

The reader is referred to A.23 for details.

B.32 Accurate Tracking by Overlap Maximization (ATOM)

Submitted by VOT Committee

The reader is referred to A.18 for details.

B.33 Discriminative Correlation Filter with Channel and Spatial Reliability (CSRDCF)

Submitted by VOT Committee

The reader is referred to A.24 for details.

The CSR-DCF [40] improves discriminative correlation filter trackers by introducing the two concepts: spatial reliability and channel reliability. It uses color segmentation as spatial reliability to adjust the filter support to the part of the object suitable for tracking. The channel reliability reflects the discriminative power of each filter channel. The tracker uses only HoG and colornames features. This is the C++ openCv implementation.

B.34 Discriminative Sing-Shot Segmentation Tracker v2 (D3Sv2)

A. Lukežič, J. Matas, M. Kristan
alan.lukezic@fri.uni-lj.si, matas@cmp.felk.cvut.cz, matej.kristan@fri.uni-lj.si

D3Sv2 [43] is an extended version of the D3S [39]. The original method is extended in the following aspects: (i) a better backbone, (ii) channel attention mechanism in the upscaling modules in GIM, (iii) trainable MLP-based similarity computation in GIM, which replaces the handcrafted' top-K average operation and (iv) the new scale estimation module used for robust target size estimation.

B.35 Learning Discriminative Model Prediction for Tracking (DiMP)

G. Bhat, M. Danelljan, L. Van Gool, R. Timofte
{goutam.bhat, martin.danelljan, vangool, timofter}@vision.ee.ethz.ch

DiMP is an end-to-end tracking architecture, capable of fully exploiting both target and background appearance information for target model prediction. The target model here constitutes the weights of a convolution layer which performs the target-background classification. The weights of this convolution layer are predicted by the target model prediction network, which is derived from a discriminative learning loss by applying an iterative optimization procedure. The model prediction network employs a steepest descent based methodology that computes an optimal step length in each iteration to provide fast convergence. The model predictor also includes an initializer network that efficiently provides an initial estimate of the model weights. The online learned target model is applied in each frame to perform target-background classification. The final bounding box is then estimated using the overlap maximization approach as in [13]. See [3] for more details about the tracker.

B.36 KCF (KCF)

Submitted by VOT Committee

The reader is referred to A.26 for details.

B.37 Know your surroundings tracker with Alpha refine post-processing step (KYS)

Submitted by VOT Committee

The reader is referred to A.20 for details.

B.38 LGT (LGT)

Submitted by VOT Committee

The reader is referred to A.27 for details.

B.39 PrDiMP50 tracker (PrDiMP)

Submitted by VOT Committee

The reader is referred to A.21 for details.

B.40 SiamFC (SiamFC)

Submitted by VOT Committee

The reader is referred to A.30 for details.

B.41 SuperDiMP50 tracker (SuperDiMP)

Submitted by VOT Committee

The reader is referred to A.22 for details.

C VOT-LT2022 submissions

This appendix provides a short summary of trackers considered in the VOT-LT2022 challenge.

C.1 Adaptive DiMP with Dynamic Sample Selection (ADiMPLT)

M. Kiran, L. Nguyen-Meidine, R. Menelau-Cruz, E. Granger
madhu-kiran.madhu-kiran.1@ens.etsmtl.ca, nmlethanh91@gmail.com,
{*rafael.menelau-cruz, eric.granger*}*@etsmtl.ca*

ADiMP [27] treats online learning in tracking as a concept drift problem. It relies on the DiMP tracker model and entropy maximization sampling, along with change detection to adapt to the appearance changes of the target during online tracking. In particular, the new method performs dynamic sample selection and memory replay, to prevent tracking drift and catastrophic forgetting over time and corrupting the tracking model. Our change detection mechanism is proposed to detect gradual changes in object appearance, and select the corresponding samples for online adaption. In addition, abrupt changes allows to manage occlusion, and thereby track objects for a longer time frame. Our entropy maximization sampling strategy allows to maintain a diversified auxiliary buffer for memory replay.

C.2 Combining Complementary Trackers in Long-Term Visual Tracking (CoCoLoT)

M. Dunnhofer, C. Micheloni
{*matteo.dunnhofer, christian.micheloni*}*@uniud.it*

The CoCoLoT tracker generalizes mlpLT [33]. It implements a strategy that combines the complementary behaviors of Stark [60] and KeepTrack [46] trackers. The combination of these trackers is managed by a decision strategy based on an online learned target verifier akin to MDNet [48]. At every frame, the trackers are run in parallel to predict their target localizations. Based on the evaluation of the target localization, the decision strategy selects the output for the current frame and to correct the tracker that performed worse. Additional strategies such as the computation of adaptive search areas and the avoidance of wrong target size estimations, have been implemented to the baseline trackers in order to make their localizations more consistent. The details of CoCoLoT are given in [18].

C.3 A Long Term Discriminative Single Shot Segmentation Tracker (D3SLT)

B. Džubur, M. Kristan, A. Lukežič
bd5830@student.uni-lj.si, {matej.kristan, Alan.Lukezic}@fri.uni-lj.si

The proposed tracker extends the D3Sv2 [42] short-term tracker with long-term capabilities. When the target is lost, the deep DCF module, combined with a motion model uncertainty is used for global re-detection, similarly as in FCLT [44]. Tracking confidence of the patch containing the segmentation output, is determined by the online verifier [26]. The tracker includes additional strategies for robust conflict resolution: (i) First, we perform periodic back-tracking of the re-detected target and the object is briefly tracked backwards in time to verify the validity of the re-detection. (ii) Second, a global search for targets, which are more likely to be the true target, is performed periodically during short-term tracking to revert from false positive tracks.

C.4 HuntFormer: Collaborative Dynamic Memory Update and Motion Prediction for Tracking Target Recapture

Z. Zhang, W. Xue, K. Zhang, C. Zhang, B. Liu, S. Chen
zzbin@stud.tjut.edu.cn, xuewanli@email.tjut.edu.cn, zhkhua@gmail.com, chenvy@tju.edu.cn, kfliubo@gmail.com, sy@ieee.org

We propose the HuntFormer that focuses on effective target recapture. Our HuntFormer is based on MixFormer [10]. An effective motion prediction model provides a reliable search region for the tracker to recapture the target. Meanwhile, we propose a novel soft-threshold-based dynamic memory update model, which keeps a set of reliable target templates in the memory that can be used to match the target position in the search region. The two modules cooperate with each other, which greatly improves the recapture ability of the tracker.

C.5 Progressive Fusion of Similar and Dissimilar Trackers for Long-term Visual Object Tracking (mixLT)

Y. Jiang, T. Xu, Z. Feng, X. Song
1161099088@qq.com, tianyang_xu@163.com, z.feng@surrey.ac.uk, x.song@jiangnan.edu.cn

Tracker mixLT is a progressive fusion of multiple trackers, mainly STARK [60] and SuperDiMP [3]. It first fuses the results of two trackers, STARK-ST50 and STARK-ST101, since they have similar properties and perform well in situations such as occlusion and object disappearance. The states of two trackers are then corrected based on the fusion result. SuperDiMP controlled by meta-updater [11] is introduced for further fusion between dissimilar trackers, in order to improve the robustness of long-term tracking. The final tracking result is determined according to the confidences of the trackers over several frames, and another tracker correction is performed.

C.6 Fusing Complementary Trackers for Long-term Visual Tracking (mlpLT)

M. Dunnhofer, K. Simonato, C. Micheloni
matteo.dunnhofer@uniud.it, simonato.kristian@spes.uniud.it,
christian.micheloni@uniud.it

The mlpLT tracker implements a strategy that combines the Stark [60] and SuperDiMP [3] trackers. Stark was chosen because of its ability in providng accurate bounding-boxes and in re-detecting the target. SuperDiMP was chosen for its robustness. The combination of the two is managed by a decision strategy based on an online learned target verifier. At every frame, the trackers are run to predict their target positions which are then checked by the verifier. Based on such evaluations, the decision strategy selects which localization to give as output. Such an outcome is also employed to correct the worse tracker. mlpLT resulted the winner of the VOT2021-LT challenge and the details of its working are given in [18].

C.7 VITKT

J. Zhao, X. Chen, C. Liu, H. Peng, D. Wang, H. Lu
{*zj982853200, chenxin3131, njx2019*}*@mail.dlut.edu.cn,*
Houwen.Peng@microsoft.com, {*wdice, lhchuan*}*@dlut.edu.cn*

VITKT benefits from the integration of different trackers. Specifically, VITKT consists of three main components including the master tracker VitTrack (B.29), the auxiliary tracker KeepTrack [46], and a metric module MetricNet [71]. VitTrack (B.29) is a Transformer-based tracker composed of a backbone network based on ViT-Base model, a corner prediction head and a classification head. It has a strong ability to handle most challenges. However, we notice that VitTrack (B.29) usually fails when distractors appear. The tracker KeepTrack is employed as an auxiliary tracker. We also employ MetricNet to predict two different distances to evaluate the similarity between the template and the current state. The final state is determined in terms of both the two distances and confidence scores output by trackers.

C.8 VITKT_M

J. Zhao, X. Chen, C. Liu, H. Peng, D. Wang, H. Lu
{*zj982853200, chenxin3131, njx2019*}*@mail.dlut.edu.cn,*
Houwen.Peng@microsoft.com, {*wdice, lhchuan*}*@dlut.edu.cn*

VITKT_M ensembles the results of VitTrack (B.29) and KeepTrack [46]. The ViT-Base model pretrained with MAE is adopted as the backbone, and the two heads (a corner prediction head and a classification head) are implemented by MLP (similar to Stark [60]). The search region and the template patch are concatenated after the patch embedding, and then input to the ViT backbone. The corner prediction and classification heads are performed to output the final state and the corresponding confidence score. We employ KeepTrack as an auxiliary tracker to solve the challenge of distractors. Different from the tracker VITKT C.7, a simple motion module (trained on LaSOT dataset) is implemented to predict the target current state according to the temporal trajectory. The motion module predicts the target state according to the previous temporal trajectory information and it will be triggered when the target is considered moving abnormally.

C.9 (SuperDiMP)

Submitted by VOT Committee

Please see the original paper for details [3].

C.10 KeepTrack (keep_track_lt)

Submitted by VOT Committee

Please see the original paper for details [46].

D VOT-RGBD2022 submissions

This appendix provides a short summary of trackers considered in the VOT-RGBD2022 challenge.

D.1 DMTracker

S. Gao and J. Yang and Z. Li and F. Zheng
12132332@mail.sustech.edu.cn, jinyu.yang96@outlook.com,
liz8@mail.sustech.edu.cn, zhengf@sustech.edu.cn

DMTracker is a Dual-fused Modality-aware (DM) Tracker which aims to learn informative and discriminative representations of the target objects for robust RGBD tracking. The first fusion module focuses on extracting the shared information between modalities based on cross-modal attention. The second fusion aims at integrating the RGB-specific and depth-specific information to enhance the fused features. By fusing both the modality-shared and modality-specific information in a modality-aware scheme, DMTracker can learn discriminative representations in complex tracking scenes.

D.2 keep_track

H. Zheng
zhenghaixia@stu.xjtu.edu.cn

KeepTrack [46] is a tracker that keeps track of distractor objects in order to continue tracking the target. KeepTrack introduces a learned association network which allows to propagate the identities of all target candidates from frame to frame. To tackle the problem of lacking ground-truth correspondences between distractor objects in visual tracking, the tracker uses a training strategy that combines partial annotations with self-supervision. KeepTrack employs super DiMP as the base tracker in order to extract target candidates and propose a target candidate association network that we use to identify the target and distractor across frames. We introduce a novel RGBD tracker (i.e. keep_track) on the top of KeepTrack [46]. We notice that KeepTrack [46] is only a RGB tracker and we design an extra channel to extract features from depth images using ResNet50. The features of depth images and RGB images are merged by element-wise pooling layers to utilise the information in depth as well as we can.

D.3 MixForRGBD

S. Lai and M. Li and J. Zhu and L. Wang and D. Wang and H. Lu
laisimiao@mail.dlut.edu.cn, liming1269521637@dlmu.edu.cn,
jiawen@mail.dlut.edu.cn, ljwang@dlut.edu.cn, wdice@dlut.edu.cn,
lhchuan@dlut.edu.cn

The MixForRGBD is built and based on the tracker MixFormer [10]. Online templates are updated over fixed intervals and when the predicted confidence is larger than a pre-set threshold. The raw depth map is colormap encoded. The colormap encoding is performed by mapping pixels of the normalized depth map to RGB vectors of a predefined color matrix (JET used in the experiments). Two same backbones of MixFormer are constructed to extract and fuse the feature of template and search region between two modalities. An element-wise maximum operation is conducted to merge the two modalities. The whole model and the score branch are fine-tuned on the existing RGB benchmarks LaSOT, COCO, GOT10k and DepthTrack training sets and their generated depth map using the monocular depth estimation algorithm DenseDepth [1]. We also add a simple post-process for penalizing large displacement.

D.4 One-stream tracker with online template updating (OSTrack)

B. Ye and H. Chang and B. Ma and S. Shan and X. Chen
botao.ye@vipl.ict.ac.cn, changhong@ict.ac.cn, bpma@ucas.ac.cn,
sgshan@ict.ac.cn, xlchen@ict.ac.cn

OSTrack is the same as OSTrackSTB B.14, no depth information is used.

D.5 ProMix

Z. Li and J. Yang and S. Gao and F. Zheng
liz8@mail.sustech.edu.cn, jinyu.yang96@outlook.com,
12132332@mail.sustech.edu.cn zhengf@sustech.edu.cn

RGBD prompt tracker (ProMixTrack) transfers the RGB and depth modalities to a single modality by the prompt paradigm. By employing the tracking ability of the pre-trained RGB tracker in the RGB challenge (Mixformer) which is trained with large scale datasets, ProMixTrack can achieve high-performance RGBD tracking even without any extra training on RGBD data.

D.6 SAMF

Z. Fu and J. Sun and L. Wang and Z. Chen and Q. Deng and D.K. Du and M. Zheng
{fuzhihong.2022, sunjingna, wangliangliang.makalo, chenzhixing.omega,dengqili, dukang.daniel, zhengmin.666}@bytedance.com

This is an ensemble method combining MixFormer [10] and SA-Gate [6]. Both the color images and depth maps are used to enhance the power of MixFormer that is built upon transformers. Specifically, two MixFormer backbones are used to extract features of color and depth images, respectively. Then, SA-Gate is adopted to combine the features. It was noticed that this type of combination of color and depth features influences the results significantly. Compared with simple fusion methods such as mean, max, etc., SA-Gate is a better for the fusion of color and depth information in RGBD tracking.

D.7 SBT_RGBD

J. Zhai and W. Zhang and F. Xie and W. Yang and C. Ma
{220211980, wkzhang}@seu.edu.cn, jaffe0319@gmail.com,
wkyang@seu.edu.cn, chaoma@sjtu.edu.cn

This is a method combining SBT-base and DepthTrack. SBT-base is a powerful fully transformer-based tracker. It was noticed that the SBT method is not good at handling the appearance change of the target. To this end, it was combined with the DepthTrack tracker, a powerful online updating tracker. Specifically, when the SBT tracker's confidence is low or the prediction of SBT suddenly strays away, the DepthTrack takes over the tracking process, providing a steady, appearance adaptive result. When the SBT confidence resumes, tracking switches back to SBT. Implementation also includes a refinement module similar to AlphaRefine by modifying the search region of SBT. The refinement module is applied to the final output of the whole tracking system for further boosting the quality of bounding box estimation.

D.8 SPT

X.-F. Zhu and T. Xu and Z. Tang and S. Zhao and H. Li and Z. Kang and X.-J. Wu and X. Li and J. Kittler
xuefeng_zhu95@163.com, tianyang.xu@surrey.ac.uk,
zhangyong_tang_jnu@163.com, 7201905026@stu.jiangnan.edu.cn,
hui_li_jnu@163.com, 6213113073@stu.jiangnan.edu.cn,
wu_xiaojun@jiangnan.edu.cn, xilizju@zju.edu.cn, j.kittler@surrey.ac.uk

In specific, firstly, the search regions and the initial templates of two modalities are input to the ResNet-50 network to extract deep CNN features, respectively. Then, the features of each modality are flattened and concatenated, following a 6-layer stacked transformer encoder to fuse the template-search appearance for the specific modality. Regarding the feature fusion module, firstly, the depth encoder output and the RGB encoder output are concatenated across channels. Then an 1d convolutional layer and a transformer encoder stacking 2 encoder layers are adopted to reduce the channel number of the concatenated features and to further fuse and enhance the features from two modalities, respectively. The rest parts of the framework including the target query, the transformer decoder and the target bounding box prediction head are the same as STARK-S [60].

E VOT-D2022 submissions

This appendix provides a short summary of trackers considered in the depth-only variant of the VOT-RGBD2022 challenge and referred to as VOT-D2022.

E.1 CoDeT

S. Gao and J. Yang and Z. Li and F. Zheng
12132332@mail.sustech.edu.cn, jinyu.yang96@outlook.com,
liz8@mail.sustech.edu.cn, zhengf@sustech.edu.cn

This is a depth-only tracker formed from the Dual-fused Modality-aware Tracker (DMTracker) submitted to the RGBD challenge. The depth frames are converted to depth pseudo colormaps (Co) and depthmaps (De).

E.2 MixFormerD

S. Lai and J. Zhu and L. Wang and D. Wang and H. Lu
{laisimiao, jiawen}@mail.dlut.edu.cn, {ljwang, wdice, lhchuan}@dlut.edu.cn

MixFormerD is a variant of the MixForRGBD that was submitted to the RGBD challenge by the same authors. The variant replaces RGB input by simply replicating the D channel three times.

E.3 OSTrack_D

B. Ye and H. Chang and B. Ma and S. Shan and X. Chen
botao.ye@vipl.ict.ac.cn, changhong@ict.ac.cn, bpma@ucas.ac.cn,
{sgshan, xlchen}@ict.ac.cn

OSTrack_D is the same as (B.14), except that RGB images are replaced by depth maps (converted to pseudo color maps).

E.4 RSDiMP

Z. Tang and X. Zhu and S. Zhao and T. Xu and J. Chen and Z. Kang and H. Li and X. Wu and J. Kittler and X. Li
{zhangyong_tang_jnu, xuefeng_zhu95}@163.com,
7201905026@stu.jiangnan.edu.cn, tianyang_xu@163.com,
jamescjy98@gmail.com, 6213113073@stu.jiangnan.edu.cn,
{hui_li_jnu, xiaojun_wu_jnu}@163.com, j.kittler@surrey.ac.uk, xilizju@zju.edu.cn

RSDiMP is based on SuperDiMP which combines the classifier from DiMP with the bounding box regressor from PrDiMP. In order to better fit the depth data, the third layer of pre-trained ResNet-50 backbone is finetuned and the classifier and IoUNet are trained on several depth datasets, including the synthetic (GOT10K-Depth and LaSOT-Depth from [62]) and real depth datasets (training split of DepthTrack as well as a large-scale dataset collected and annotated by ourselves). Besides, based on the distance statistic that the movement between the adjacent frames is slight, we shrink the scale of search area to a suitable magnitude. The discriminative classifier is updated when the current prediction is thought confident.

E.5 SBT_Depth

W. Zhang and J. Zhai and F. Xie and W. Yang
{wkzhang, 220211980}@seu.edu.cn, jaffe0319@gmail.com,
wkyang@seu.edu.cn

SBT_Depth is a depth-only variant of the SBT_RGBD D.7.

E.6 UpDoT

Z. Li and J. Yang and S. Gao and F. Zheng
liz8@mail.sustech.edu.cn, jinyu.yang96@outlook.com,
12132332@mail.sustech.edu.cn, zhengf@sustech.edu.cn

DepthColormap-DiMP and DepthMap-DiMP are trained from the scratch with generated data and finetuned with the available small RGBD tracking datasets. Specifically, when the depth-only DepthColormap-DiMP tracker's confidence is low or the prediction of it suddenly strays away, the DepthMap-DiMP will update the DepthColormap-DiMP branch for some frames until the confidence resume.

References

1. Alhashim, I., Wonka, P.: High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941 (2018)
2. Bertinetto, L., Valmadre, J., Henriques, J., Torr, P.H.S., Vedaldi, A.: Fully-convolutional siamese networks for object tracking. In: ECCV Workshops. pp. 850–865 (2016)
3. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6182–6191 (2019)
4. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Know your surroundings: Exploiting scene information for object tracking. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII. Lecture Notes in Computer Science, vol. 12368, pp. 205–221. Springer (2020). https://doi.org/10.1007/978-3-030-58592-1_13, https://doi.org/10.1007/978-3-030-58592-1_13
5. Bhat, G., Lawin, F.J., Danelljan, M., Robinson, A., Felsberg, M., Gool, L.V., Timofte, R.: Learning what to learn for video object segmentation. In: European Conference on Computer Vision ECCV (2020)
6. Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G.: Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In: European Conference on Computer Vision. pp. 561–577. Springer (2020)
7. Chen, X., Yan, B., Zhu, J., Wang, D., Lu, H.: High-performance transformer tracking. arXiv preprint arXiv:2203.13533 (2022)
8. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8126–8135 (2021)
9. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR (2005)
10. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13608–13618 (2022)
11. Dai, K., Zhang, Y., Wang, D., Li, J., Lu, H., Yang, X.: High-performance long-term tracking with meta-updater. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6298–6307 (2020)
12. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: ECO: efficient convolution operators for tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 6638–6646. IEEE Computer Society (2017)
13. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: ATOM: Accurate tracking by overlap maximization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4660–4669 (2019)
14. Danelljan, M., Gool, L.V., Timofte, R.: Probabilistic regression for visual tracking. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7181–7190 (2020)
15. Danelljan, M., Häger, G., Khan, F.S., Felsberg, M.: Discriminative scale space tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(8), 1561–1575 (2016)

16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S.e.a.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
17. Dunnhofer, M., Martinel, N., Micheloni, C.: Tracking-by-trackers with a distilled and reinforced model. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (November 2020)
18. Dunnhofer, M., Simonato, K., Micheloni, C.: Combining complementary trackers for enhanced long-term visual object tracking. Image and Vision Computing p. 104448 (2022). https://doi.org/https://doi.org/10.1016/j.imavis.2022.104448, `https://www.sciencedirect.com/science/article/pii/S0262885622000774`
19. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In: IEEE Conf. Comp. Vis. and Patt. Rec. (CVPR) (2019)
20. Guo, C., Zhang, L.: A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Transactions on Image Processing **19**(1), 185–198 (2009)
21. Gustafsson, F.K., Danelljan, M., Bhat, G., Schön, T.B.: Energy-based models for deep probabilistic regression. In: European Conference on Computer Vision ECCV (2020)
22. Gustafsson, F.K., Danelljan, M., Timofte, R., Schön, T.B.: How to train your energy-based model for regression. CoRR **abs/2005.01698** (2020), `https://arxiv.org/abs/2005.01698`
23. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
24. Henriques, J., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on PAMI **37**(3), 583–596 (2015)
25. Jhang, S., Tsai, C.: Reptile Meta-Tracking. In: IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 1–5. IEEE, Taipei, Taiwan (2019)
26. Jung, I., Son, J., Baek, M., Han, B.: Real-time mdnet. In: European Conference on Computer Vision (ECCV) (Sept 2018)
27. Kiran, M., Nguyen-Meidine, L.T., Sahay, R., Cruz, R.M.O.E., Blais-Morin, L.A., Granger, E.: Dynamic template selection through change detection for adaptive siamese tracking. In: International Joint Conference on Neural Networks (IJCNN) (2022)
28. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kamarainen, J.K., Čehovin, L., Drbohlav, O., Lukežič, A., Berg, A., Eldesokey, A., Kapyla, J., Fernández, G., et al.: The seventh visual object tracking vot2019 challenge results. In: ICCV2019 Workshops, Workshop on visual object tracking challenge (2019)
29. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kamarainen, J.K., Čehovin, L., Martin, D., Lukežič, A., Drbohlav, O., He, L., Zhang, Y., Yan, S., Yang, J., Fernández, G., et al.: The eighth visual object tracking vot2020 challenge results. In: ECCV2020 Workshops, Workshop on visual object tracking challenge (2020)
30. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Čehovin, L., Vojíř, T., Bhat, G., Lukežič, A., Eldesokey, A., Fernández, G., et al.: The visual object tracking vot2018 challenge results. In: ECCV2018 Workshops, Workshop on visual object tracking challenge (2018)

31. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Čehovin, L., Vojíř, T., Häger, G., Lukežič, A., Eldesokey, A., Fernández, G., et al.: The visual object tracking vot2017 challenge results. In: ICCV2017 Workshops, Workshop on visual object tracking challenge (2017)
32. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Čehovin, L., Vojíř, T., Häger, G., Lukežič, A., Fernández, G., et al.: The visual object tracking vot2016 challenge results. In: ECCV2016 Workshops, Workshop on visual object tracking challenge (2016)
33. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kämäräinen, J.K., Chang, H., Martin, D., Čehovin, L., Lukežič, A., Drbohlav, O., Kapyla, J., Hager, G., Yan, S., Yang, J., Zhang, Z., Fernandez, G., et. al.: The ninth visual object tracking vot2021 challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision ICCV2021 Workshops, Workshop on visual object tracking challenge. pp. 2711–2738 (2021)
34. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Čehovin, L., Fernández, G., Vojíř, T., Häger, G., Nebehay, G., Pflugfelder, R., et al.: The visual object tracking vot2015 challenge results. In: ICCV2015 Workshops, Workshop on visual object tracking challenge (2015)
35. Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Porikli, F., Čehovin, L., Nebehay, G., Fernández, G., Vojíř, T., et al.: The visual object tracking vot2013 challenge results. In: ICCV2013 Workshops, Workshop on visual object tracking challenge. pp. 98 –111 (2013)
36. Kristan, M., Pflugfelder, R., Leonardis, A., Matas, J., Čehovin, L., Nebehay, G., Vojíř, T., Fernández, G., et al.: The visual object tracking vot2014 challenge results. In: ECCV2014 Workshops, Workshop on visual object tracking challenge (2014)
37. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
38. Lukežič, A., Kart, U., Kämäräinen, J., Matas, J., Kristan, M.: CDTB: A Color and Depth Visual Object Tracking Dataset and Benchmark. In: ICCV (2019)
39. Lukežič, A., Matas, J., Kristan, M.: D3S - A discriminative single shot segmentation tracker. In: Proceedings of the IEEE/CVF CVPR. pp. 7131–7140. IEEE (2020)
40. Lukežič, A., Vojíř, T., Čehovin Zajc, L., Matas, J., Kristan, M.: Discriminative correlation filter with channel and spatial reliability. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6309–6318 (July 2017)
41. Lukežič, A., Čehovin Zajc, L., Vojíř, T., Matas, J., Kristan, M.: Sperformance evaluation methodology for long-term single object tracking. IEEE Transactions on Cybernetics (2020)
42. Lukežič, A., Matas, J., Kristan, M.: A discriminative single-shot segmentation network for visual object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**, 1–1 (12 2021). https://doi.org/10.1109/TPAMI.2021.3137933
43. Lukezic, A., Matas, J., Kristan, M.: A discriminative single-shot segmentation network for visual object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
44. Lukežič, A., Čehovin Zajc, L., Vojiř, T., Matas, J., Kristan, M.: FuCoLoT - A Fully-Correlational Long-Term Tracker. In: ACCV (2018)
45. Mayer, C., Danelljan, M., Bhat, G., Paul, M., Paudel, D.P., Yu, F., Van Gool, L.: Transforming model prediction for tracking. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8731–8740 (June 2022)

46. Mayer, C., Danelljan, M., Paudel, D.P., Van Gool, L.: Learning target candidate association to keep track of what not to track. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13444–13454 (2021)
47. Miao, J., Wang, X., Wu, Y., Li, W., Zhang, X., Wei, Y., Yang, Y.: Large-scale video panoptic segmentation in the wild: A benchmark. In: CVPR (2022)
48. Nam, H., Han, B.: Learning multi-domain convolutional neural networks for visual tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4293–4302 (2016)
49. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9226–9235 (2019)
50. Paul, M., Danelljan, M., Mayer, C., Gool, L.V.: Robust visual tracking by segmentation. In: European Conference on Computer Vision ECCV (2022)
51. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. IJCV **115**(3), 211–252 (2015)
52. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877 (2020)
53. Čehovin, L., Kristan, M., Leonardis, A.: Robust visual tracking using an adaptive coupled-layer visual model. IEEE Trans. Pattern Anal. Mach. Intell. **35**(4), 941–953 (2013). https://doi.org/http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.145
54. Čehovin, L.: TraX: The visual Tracking eXchange Protocol and Library. Neurocomputing (2017). https://doi.org/http://dx.doi.org/10.1016/j.neucom.2017.02.036
55. Čehovin, L., Leonardis, A., Kristan, M.: Robust visual tracking using template anchors. In: WACV. IEEE (Mar 2016)
56. Vojíř, T., Noskova, J., Matas, J.: Robust scale-adaptive mean-shift for tracking. Pattern Recognition Letters **49**, 250–258 (2014)
57. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: ICCV. pp. 22–31 (2021)
58. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: Comp. Vis. Patt. Recognition (2013)
59. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
60. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10448–10457 (2021)
61. Yan, B., Zhang, X., Wang, D., Lu, H., Yang, X.: Alpha-refine: Boosting tracking performance by precise bounding box estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5289–5298 (2021)
62. Yan, S., Yang, J., Käpylä, J., Zheng, F., Leonardis, A., Kämäräinen, J.K.: DepthTrack: Unveiling the power of RGBD tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10725–10733 (2021)
63. Yan, S., Yang, J., Leonardis, A., Kämäräinen, J.K.: Depth-only object tracking. In: British Machine Vision Conference (BMVC) (2021)

64. Yang, Z., Liu, S., Hu, H., Wang, L., Lin, S.: Reppoints: Point set representation for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9657–9666 (Oct 2019)
65. Yang, Z., Miao, J., Wang, X., Wei, Y., Yang, Y.: Associating objects with scalable transformers for video object segmentation. arXiv preprint arXiv:2203.11442 (2022)
66. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. In: NeurIPS. vol. 34 (2021)
67. Yang, Z., Zhang, J., Wang, W., Han, W., Yu, Y., Li, Y., Wang, J., Wei, Y., Sun, Y., Yang, Y.: Towards multi-object association from foreground-background integration. In: CVPR Workshops. vol. 2 (2021)
68. Ye, B., Chang, H., Ma, B., Shan, S.: Joint feature learning and relation modeling for tracking: A one-stream framework. arXiv preprint arXiv:2203.11991 (2022)
69. Zhang, C., Ge, S., Zhang, K., Zeng, D.: Accurate uav tracking with distance-injected overlap maximization. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 565–573 (2020)
70. Zhao, B., Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Generating masks from boxes by mining spatio-temporal consistencies in videos. CoRR **abs/2101.02196** (2021), `https://arxiv.org/abs/2101.02196`
71. Zhao, J., Dai, K., Wang, D., Lu, H., Yang, X.: Online filtering training samples for robust visual tracking. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1488–1496 (2020)
72. Zhu, J., Chen, X., Wang, D., Zhao, W., Lu, H.: Srrt: Search region regulation tracking. arXiv:2207.04438 (2022)