The Second Visual Object Tracking Segmentation VOTS2024 Challenge Results

Matej Kristan[1], Jiří Matas[2], Pavel Tokmakov[3], Michael Felsberg[4], Luka Čehovin Zajc[1], Alan Lukežič[1], Khanh-Tung Tran[5], Xuan-Son Vu[5,6], Johanna Björklund[5], Hyung Jin Chang[7], Gustavo Fernández[8], Minasadat Attari[29], Antoni Chan[10], Liang Chen[12], Xin Chen[11], Jaired Collins[29], Yutao Cui[17], Ganesh Sai Manas Devarapu[14], Yinglong Du[9], Heng Fan[30], Wan-Cyuan Fan[25], Zhenhua Feng[15], Mingqi Gao[24], Rama Krishna Sai Gorthi[14], Raghav Goyal[25], Jungong Han[31], Bijaya Hatuwal[29], Zhenyu He[13], Xiantao Hu[12], Xingsen Huang[13], Yuqing Huang[13], Dongmei Jiang[20], Ben Kang[11], Palaniappan Kannappan[29], Josef Kittler[32], Simiao Lai[11], Ning Li[12], Xiaohai Li[12], Xin Li[20], Cheng Liang[17], Liting Lin[20], Haibin Ling[30], Ting Liu[18], Ziquan Liu[21], Huchuan Lu[11], Yifei Luo[9], Deshui Miao[13], Juan Mogollon[29], Ziqi Pang[28], Jaswanth Reddy Pochimireddy[14], Viktor Prutyanov, Gani Rahmon[29], Aleksandr Romanov, Liangtao Shi[12], Mennatullah Siam[19,25], Leonid Sigal[25], Arun Kumar Sivapuram[14], Roman Solovyev, Elham Soltani Kazemi[29], Imad Eddine Toubal[29], Jia Wan[13], Limin Wang[17], Xinying Wang[11], Yaowei Wang[20], Yu-Xiong Wang[28], Zhiquan Wang[17], Gangshan Wu[17], Qiangqiang Wu[10], Xiaojun Wu[15], Zihao Xia[17], Jinxia Xie[12], Chenlong Xu[12], Tianyang Xu[15], Yong Xu[23], Chaocan Xue[12], Chao Yang[13], Jinyu Yang[24], Ming-Hsuan Yang[26], Chenyang Yu[11], Ke Yu[27], Chunhui Zhang[22], Jiaming Zhang[17], Zhipeng Zhang[16], Feng Zheng[24], Yaozong Zheng[12], Bineng Zhong[12], Jinglin Zhou[15], Junbao Zhou[28], Yong Zhou[15], Zikun Zhou[20], Guibo Zhu[9], Jiawen Zhu[11], Xuefeng Zhu[15], and Vladimir Zunin

[1] University of Ljubljana, Slovenia
[2] Czech Technical University, Czech Republic
[3] Toyota Research Institute, USA
[4] Linköping University, Sweden
[5] Umeå University, Sweden
[6] DeepTensor AB, Sweden
[7] University of Birmingham, United Kingdom
[8] Austrian Institute of Technology, Austria
[9] CASIA, China
[10] City University of Hong Kong, Hong Kong
[11] Dalian University of Technology, China
[12] Guangxi Normal University, China
[13] Harbin Institute of Technology, Shenzhen, China
[14] IIT Tirupati, India
[15] Jiangnan University, China
[16] KargoBot, China
[17] Nanjing University, China
[18] National University of Defense Technology, China
[19] Ontario Tech University, Canada
[20] Peng Cheng Laboratory, China

[21] Queen Mary University of London, United Kingdom
[22] Shanghai Jiao Tong University, China
[23] South China University, China
[24] Tapall.ai, China
[25] UBC, Canada
[26] University of California at Merced, USA
[27] University of California San Diego, USA
[28] University of Illinois Urbana Champaign, USA
[29] University of Missouri, USA
[30] University of North Texas, USA
[31] University of Sheffield, United Kingdom
[32] University of Surrey, United Kingdom
matej.kristan@fri.uni-lj.si

Abstract. The Visual Object Tracking Segmentation VOTS2024 challenge is the twelfth annual tracker benchmarking activity of the VOT initiative. This challenge consolidates the new tracking setup proposed in VOTS2023, which merges short-term and long-term as well as single-target and multiple-target tracking with segmentation masks as the only target location specification. Two sub-challenges are considered. The VOTS2024 standard challenge, focusing on classical objects and the VOTSt2024, which considers objects undergoing a topological transformation. Both challenges use the same performance evaluation methodology. Results of 28 submissions are presented and analyzed. A leaderboard, with participating trackers details, the source code, the datasets, and the evaluation kit are publicly available on the website[33].

Keywords: VOTS · tracking and segmentation · transformative object tracking · performance evaluation

1 Introduction

Visual object tracking has been continually progressing over several decades, primarily driven by the research efforts of the community and emergence of various benchmarks. Over a decade ago, the VOT[33] initiative was founded to address the lack of performance evaluation consensus in visual object tracking. Through organization of annual challenges in conjunction with the International and European Conferences on Computer Vision (ICCVs and ECCVs), the initiative has explored the tracking landscape and identified major tracking trends, which made the VOT event central to the tracking community.

To facilitate gradual advancements, VOT was initially restricted to single-target tracking and explored short-term and long-term, offline and realtime tracking challenges separately and explored various sensor modalities through

[33] https://www.votchallenge.net/vots2024/

dedicated sub-challenges. Substantial effort was invested in development and revision of performance measures, evaluation protocols and toolkits. To keep raising the bar for ever-improving tracker methodologies, the target location specification has evolved from reporting bounding boxes in the initial challenges [16, 18–20, 22–24] to per-pixel segmentation in the later challenges [15, 17, 21].

A number of related activities emerged, focusing on surveillance scenarios, such as UAVision[34], VisDrone[35] and Anti-UAV[36] and multi-target tracking and segmentation, such as MOTComplex[37], TAO-OW[38], STEP benchmark[39] and LaGOT [31], which consider multiple-object generic tracking by bounding boxes. Video object segmentation in short clips has been explored by DAVIS challenge [5] and continued by the recent endeavours of YouTube-VOS[40]. All these challenges consider objects that may deform, but do not change topology. A new task was introduced recently, which addresses tracking by segmentation of objects that undergo topological transformations [40], such as vegetables cut into pieces and machines being disassembled.

Most of the aforementioned activities are tightly coupled with detectors, and do not directly address the needs of the traditional tracking community interested in general trackers, i.e. trackers whose target is specified at run time and could include arbitrary areas, corresponding to parts of object or regions covered by 'stuff', e. g. a part of road. Moreover, different evaluation protocols are used for different tracking tasks, such as short-term and long-term tracking, and notably between single- and multiple-object tracking.

Driven by substantial increase in maturity of tracking, the VOT initiative proposed in 2023 to no longer make distinction between short-term, long-term, single-target and multiple-target tracking. A unified performance evaluation methodology was proposed and a new series of challenges emerged under the title Visual Object Tracking and Segmentation (VOTS) challenges, with the first challenge organized in conjunction with ICCV2023 [25].

This paper presents the second VOTS2024 challenge, organized in conjunction with the ECCV2024 Visual Object Tracking and Segmentation Workshop, and the results obtained. In the following, we overview the challenge and participation requirements.

[34] CV for UAVs, ICCV 2021, https://sites.google.com/view/uavision2021/
[35] VisDrone challenge, ICCV 2023,http://aiskyeye.com/challenge-2023/
[36] Anti-UAV Challenge, CVPR 2023, https://anti-uav.github.io/
[37] Multiple Object Tracking and Segmentation in Complex Environments, ECCV 2022, https://motcomplex.github.io/
[38] 2nd Workshop on Tracking and Its Many Guises: Tracking Any Object in Open-World, CVPR 2023, https://taodataset.org/workshop/cvpr23/
[39] Segmenting and Tracking Every Point and Pixel: 6th Workshop on Benchmarking Multi-Target Tracking, ICCV 2021 https://motchallenge.net/workshops/bmtt2021/
[40] YouTube-VOS, The 5th Large-scale Video Object Segmentation Challenge, ICCV2023, https://youtube-vos.org/

1.1 The VOTS2024 challenge

The evaluation toolkit and the datasets were provided by the VOTS2024 organizers. The challenge opened on May 13 and closed on June 23 2024. The results, along with the winners were disclosed in early July 2024. The analysis of the results were presented at ECCV2024 VOTS2024 workshop on September 29 2024. The *VOTS2024 Benchmark* opened[41] with a continually updated leaderboard to facilitate tracker development in the post-challenge period. Two challenges were organized in VOTS2024:

- **VOTS2024 challenge** was a continuation of the VOTS2023 challenge, with the task to track one or more general targets over short-term or long-term sequences by segmentation.
- **VOTSt2024 challenge** was a new challenge with the task of tracking objects undergoing topological transformations.

The conceptual difference between the tracked objects considered in the VOTS2024 and VOTSt2024 challenges are visualized in Figure 1.

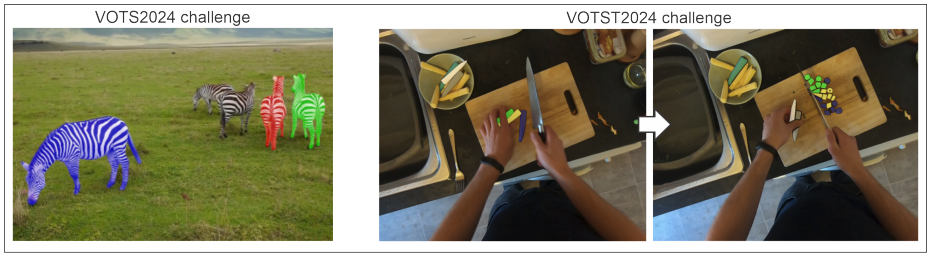

Fig. 1: The VOTS2024 challenge considers tracking objects and/or their parts, whose topology remains constant (left), while VOTSt2024 considers objects whose topology changes during tracking (right).

The participants of the challenges integrated their tracker into the VOTS2024 evaluation kit, which implements the recent evaluation protocols and the datasets, and automatically performed the standardized experiments. Each participant then registered the trackers on the evaluation server and submitted the tracker outputs produced in the experiment. Note that only the initialization frames of ground truth were publicly available, while the ground truth of the remaining frames was sequestered on the server side to prevent overfitting. Furthermore, each registered participant was allowed only 10 attempts to run the evaluation (maximum one per day).

Participants were encouraged to submit their own new or previously published trackers as well as modified versions of third-party trackers. In the latter

[41] `https://eu.aihub.ml/competitions/201`

case, modifications had to be significant enough for acceptance. Each submission was accompanied by a short abstract describing the tracker, which was used for the short tracker descriptions in Appendix A and Appendix B, and a questionnaire to categorize their tracker along various design properties.

Participants with sufficiently well performing submissions (i.e., exceeding the STARK-multi [25] for VOTS2024 and AOTPlus B.1 for VOTSt2024), who contributed with the text for this paper and agreed to make their tracker code publicly available on the VOTS page 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.

The trackers submitted to VOTS2023 could not be re-submitted to VOTS2024 sub-challenge, since the same dataset and the evaluation protocol are used. However, their submission was encouraged to the VOTSt2024 sub-challenge, since it considers a different tracking problem and uses a different dataset. The VOTS committee members could participate in the challenge with their own submissions, but could not compete for the winner title. All co-authors of this paper, including the VOTS2024 committee members and the tracker authors were required to specify the division of work.

Validation and test splits of popular tracking datasets were *not allowed for training* the trackers. These include OTB [42], VOT, ALOV [39], UAV123 [32], NUSPRO [26], TempleColor [29], AVisT [34], LaSOT-val [8], LaGOT [31], GOT10k-val/test [11], TrackingNet-val/test [33], and TOTB [9]. The training split of any dataset is allowed (including LaSOT-train, TrackingNet-train, YouTubeVOS, COCO, etc.). To include transparent objects, the Trans2k [30] dataset[42] was suggested.

The remainder of this report is structured as follows. Section 2.1 overviews the performance evaluation protocol and the two challenges, the results are discussed in Section 3.1, and conclusions are drawn in Section 4. Short descriptions of the tested trackers and divison of work are available in the Appendix A and Appendix B.

2 Description of the individual challenges

The VOTS2024 challenge featured two sub-challenges, (i) the standard VOTS2024 sub-challenge and (ii) the new VOTSt2024 sub-challenge. Both challenges used the unified performance evaluation methodology proposed in VOTS2023 [25]. We thus first review the methodology in Section 2.1 and the individual challenges in Section 2.2 and Section 2.3, respectively.

2.1 The VOTS performance evaluation protocol

The tracker is initialized in the first frame on all specified targets. For each subsequent frame, the tracker is required to report the locations for all visible

[42] `https://github.com/trojerz/Trans2k`

targets in that frame. Specifically, a segmentation mask is required for each visible target, "not present" label is reported for the absent targets. The goal of a tracker is to reliably track each individual target selected in the first frame. Drifting off a target to the background, or another target, is both considered as failed tracking.

In VOTS2023 [25], a notion of tracking success was defined, from which the tracking quality plot can be obtained, replacing the widely-used success plot [42]. Note that, in contrast to the success plot [42], the tracking quality plot also accounts for the long-term tracking properties (see [25] for details). The VOTS primary performance measure is thus the tracking quality Q, which is related to the area under the curve in the tracking quality plot.

In addition, several secondary performance measures were proposed to aid tracking analysis. The first two are localization *accuracy* and *robustness*. The accuracy (Acc) is defined as the sequence-normalized average overlap over successfully tracked frames, while tracking robustness (Rob) is defined as the percentage of successfully tracked frames with visible target, averaged over all sequences. Following our prior work [17], the tracker performance on frames with visible target is summarized by the AR plots [15], with the top-right position indicating the better performance. Three additional secondary measures were proposed to quantify different tracking properties. These are *Not-Reported Error* (NRE), which measures the percentage of frames where the tracker incorrectly reported the target as absent; *Drift-Rate Error* DRE measures the percentage of frames where the tracker drifted off the target; *Absence-Detection Quality* ADQ, gives the percentage of frames with target correctly predicted as absent.

2.2 The VOTS2024 challenge

The VOTS2024 subchallenge was a continuation of the VOTS2023 challenge [25]. This challenge considered tracking by segmentation of one or more targets, whose topology remained constant (i.e., classical objects or their parts). The VOTS2023 dataset was employed, which is composed of 144 sequences, and contains 341 targets in total. The average length of a sequence is ≈2000 frames (min = 63, max = 10700, median = 1810). The number of targets in a sequence ranges from 1 to 8 (median = 2, mean = 2.37). Of the 144 sequences, 93 contain a target which at least once leaves the field of view and then returns. Of the 341 targets, this happens with 168. In cases when the target leaves and returns to the field of view, the median number of absences is 3, with maximum being 23. The median absence length in terms of frame number is 18. All targets are annotated by segmentation masks. The ground truth masks are publicly available only for initialization frames, while the evaluation ground truth is sequestered on the evaluation server and is not publicly available.

Winner identification protocol. The VOTS2024 challenge winner was identified as the tracker that obtained the top tracking quality score Q among all valid submissions. The winners are required to make the tracker code publicly available before the VOTS2024 workshop.

2.3 The VOTSt2024 challenge

The VOTSt2024 subchallenge focusees on complex object transformations, capturing their full temporal extent, and is derived from the work originally presented in [40] at CVPR2023. In particular, the dataset is composed of 713 high-resolution videos, each of which features the tracked objects undergoing major changes of appearance, shape or even aggregate state (e.g. cracking of an egg). In addition, other tracking challenges, such as occlusions and objects leaving the frame for significant time are also common. The videos are 21.2 seconds long on average and are labeled with instance masks at 5 FPS, resulting in 75547 labeled frames in total. Challenge participants were provided with frames extracted at 10 FPS for improved temporal consistency, but only the predictions on labeled frames are used for evaluation. Extra care was taken to account for ambiguous cases when the exact extent of an object mask cannot be determined due to motion blur or the object being semi-transparent, with such regions being annotated with 'Ignore' label. Overall, there are 2.3 masks annotated per frame on average.

The dataset is split into 572 training, 70 validation, and 71 test videos, where annotations for train and validation sets are publicly available, but for the test set, which is used for evaluation in this challenge, only the first frame masks have been released. The standard VOT challenge metrics described in Section 2.1 are used to score the submissions, with the only modification being that, following [40], they are computed over the last 25% for the frames. This modification is proposed to emphasize the robustness of methods to transformations (which are mostly completed by the end of the videos) rather than their generic tracking abilities, which are evaluated by the VOTS subchallenge above.

Winner identification protocol. The VOTSt2024 challenge winner was identified as the tracker that obtained the top tracking quality score Q among all valid submissions. The winners are required to make the tracker code publicly available before the VOTS2024 workshop.

3 The challenge results

This section summarizes the trackers submitted, results analysis and winner identification for each of the two VOTS2024 challenges.

3.1 The VOTS2024 challenge results

A total of 24 trackers was submitted to the VOTS2024 evaluation server. After removing the duplicate, near-duplicate and incomplete submissions, 18 valid entries remained in the VOTS2024 challenge: S3_Track (A.14), dmaot_sam (A.4), HQ-DMAOT (A.9), LY-SAM (A.11), Cutie-SAM (A.2), swinb_dm_deaot_vots (A.16), tapall_ai (A.17), goten (A.8), LoRAT (A.10), ODTrack (A.13), EVPTrack (A.7), DropTrackSaml (A.6), AQATrack (A.1), DPTrack (A.5), VOTS2024_MIEM-HSE_AlphaCHIP (A.18), DITrack (A.3), STARK_SAM_HQ (A.15), and MambaTracking (A.12). Each submission included the link to the source code to allow

verification of the results if required. The authors agreed to make the source codes publicly accessible by the VOTS2024 workshop date. In addition, the VOTS2024 committee included three baselines from VOTS2023 challenge [25]: DMAOT (the winner of VOTS2023 challenge), AOT (foundation of many top trackers since VOTS2023) and the baseline tracker StarkMulti [25].

In the following we summarize the statistics of the submitted trackers (excluding the three baselines) and refer the reader to the Appendix A for the trackers short descriptions. Of the participating trackers, 7 (39%) were categorized by their authors as ST_0, 7 (39%) were categorized as ST_1, 1 (5%) were categorized as LT_0, and 3 (17%) were categorized as LT_1. Most trackers (15; 83%) applied a uniform dynamic model, while (3; 17%) applied a nearly-constant velocity model. The dominant tracking methodology was transformers. In fact, 17 (94%) of the submissions utilized transformers, while one used a Mamba [10]. Most of the trackers localized the targets in multiple stages (10; 56%), while 8 (44%) performed a single-stage localization. Over a two thirds of the submissions utilized the general object segmentation network SAM [14] (12; 67%), one fifth applied object-specific network AlphaRef [46] for target segmentation or for refining the segmentation (4; 22%), while one tracker applied a direct segmentation and one predicted a bounding box as the primary output. Two (22%) trackers applied a fixed template updating mechanism, 5 (28%) updated the template only when confident, 4 (22%) always updated the template, and 5 (28%) never updated the template. All submissions applied the same network for frame-to-frame localization and target re-detection.

The results are summarized in the tracking quality plots and AR plots (Figure 2), and in Table 1. The top 3 trackers according to the primary tracking quality score (Q) are: S3_Track (A.14), dmaot_sam (A.4), and HQ-DMAOT (A.9). The top-performer S3_Track is a single-stage tracker that jointly considers all targets within the same pipeline. The tracker is composed of a semantic-aware feature generation module and a target association module, which correlates the semantic-aware features with the pixels. The resulting correlation features are decoded into the segmentation masks. This tracker extends the XMem [4] model by instance-level information for more robust localization. The next two trackers both two-stage architectures based on the VOTS2023 winner DAMOT [25]. Dmaot_sam applies DMAOT in the first stage to jointly predict the target segmentation masks. In the second stage, a bounding box is fitted to each segmentation mask and used as a prompt in SAM [14], which predicts several masks by default. For each target, the final mask is selected as the one with the largest IoU value with the initial DAMOT mask. HQ-DMAOT follows a similar two-stage strategy.

S3_Track obtains the highest tracking quality (Q=0.722), which is a 10% improvement over the second-best entry. The AR plot (Figure 2) indicates that S3_Track strikes an excellent balance between accurate target segmentation (Acc=0.784) and very good robustness (Rob=0.889) – the latter indicates that this tracker successfully tracked nearly 90% of an average test sequence length, which is a remarkable performance. The tracker drifted off the target in only 4%

of cases (DRE=0.04), and falsely predicted the target as absent in 7% of cases (NRE=0.07). This also means that when the target was present, the tracker failed only in 11% of frames, of this, 36% of failures were due to drifting off the target and 63% of failures were due to incorrectly reporting the target as absent. Overall, the target absence was correctly predicted in 78% of cases (ADQ=0.78).

According to the robustness score, S3_Track is ranked first, while the second-most robust tracker is swinb_dm_deaot_vots, whose robustness is approximately 5% lower than that of S3_Track, but still very high. This tracker is a variant of DMAOT trained for robustness (i.e., segmentation capabilities removed) and combined with HQ-SAM [13] for mask prediction, which, however yields a much lower accuracy score (ranked 8th). We note that all top 10 trackers in robustness apply powerful backbones (ViT and Swin), which likely contribute to their localization robustness. The top tracker in accuracy was dmaot_sam, which outperformed S3_Track by 1.2% in this measure. Note that the segmentation accuracy of dmaot_sam comes from SAM [14] and the novel mask selection procedure.

	Quality	AR		Auxiliary measures		
Tracker	Q ↑	A ↑	R ↑	NRE ↓	DRE ↓	ADQ ↑
S3_Track	0.722①	0.784②	0.889①	0.070	0.041②	0.781③
dmaot_sam	0.653②	0.794①	0.780	0.142	0.078	0.734
HQ-DMAOT	0.639③	0.754	0.790	0.138	0.072	0.750
DMAOT	0.636	0.751	0.795③	0.139	0.066	0.731
LY-SAM	0.631	0.765	0.776	0.140	0.084	0.724
Cutie-SAM	0.607	0.756	0.730	0.210	0.059③	0.851②
swinb_dm_deaot	0.597	0.752	0.845②	0.013	0.141	0.007
tapall_ai	0.589	0.734	0.712	0.249	0.040①	0.894①
goten	0.574	0.782③	0.772	0.013	0.215	0.000
AOT	0.550	0.698	0.767	0.096	0.137	0.470
LoRAT	0.536	0.725	0.784	0.013①	0.203	0.000
ODTrack	0.532	0.723	0.775	0.013	0.212	0.010
EVPTrack	0.531	0.774	0.727	0.013	0.260	0.002
DropTrackSaml	0.529	0.745	0.754	0.013①	0.233	0.000
AQATrack	0.519	0.749	0.741	0.013	0.247	0.000
DPTrack	0.497	0.708	0.741	0.013	0.246	0.004
AlphaCHIP	0.485	0.680	0.727	0.025	0.247	0.131
DITrack	0.479	0.699	0.711	0.022	0.267	0.012
STARK_SAM_HQ	0.457	0.711	0.683	0.013③	0.304	0.001
MambaTracking	0.453	0.699	0.680	0.017	0.302	0.008
StarkMulti	0.297	0.441	0.594	0.220	0.186	0.486

Table 1: Numerical results for VOTS2024 challenge. Tracking quality (Q), accuracy (Acc), robustness (Rob), not-reported error (NRE), drift-rate error (DRE), and absence-detection quality (ADQ).

From Figure 2 we observe that the Q-plot of S3_Track stands out from the rest. It starts higher than for the rest of the trackers and its elbow extends far to the right. This indicates that the tracker is more robust than the other trackers for a range of IoU thresholds (up to over 0.8), further confirming a good balance between the localization robustness and the segmentation accuracy.

The VOTS2024 committee provided an entry-level baseline tracker, Stark-Multi [25] (Q=0.297), for validating the general quality of the submissions. All 18 submissions outperform the this tracker. The second baseline was AOT [47] (Q=0.55), which was the basis of VOTS2022 and VOTS2023 winning trackers and foundation of many submissions this year. Eight submissions (44%) outperform this tracker. The most challenging baseline was the VOTS2023 winner DMAOT [25] (Q=0.636). This year, three trackers (17%) outperform it: S3_Track, dmaot_sam, and HQ-DMAOT. The analysis shows that challenge indeed considers very competitive submissions and that the top trackers have successfully pushed the bar beyond the top performance reached in VOTS2023.

The VOTS2024 challenge winner. The top tracker according to the tracking quality score Q is S3_Track (A.14) and thus the VOTS2024 challenge winner.

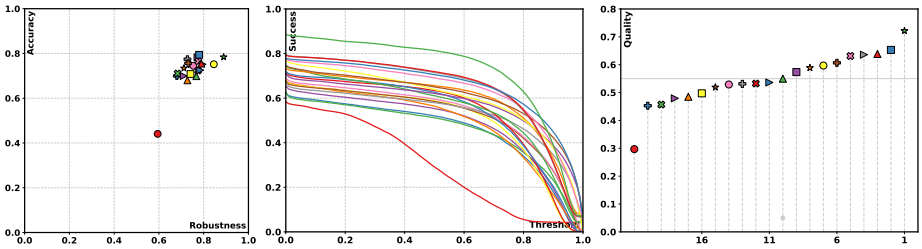

Fig. 2: The VOTS2024 challenge accuracy/robustness AR-plots (left), tracking quality Q-plots (center), and all trackers ranked according to Q score (right).

3.2 The VOTSt2024 challenge results

A total of 8 trackers were submitted to the evaluation server, including the baseline contributed by the VOTS committee. After removing the duplicate, near-duplicate and incomplete submissions, 7 valid entries remained in the VOTSt2024 challenge: AOTPlus (B.1), AQATrack (B.2), S3_Track (B.6), TAM-VT (B.7), EVPTrack (B.3), RMem (B.4), and RMemAOT (B.5).

Each submission included the link to the source code to allow verification of the results if required. The source codes are publicly accessible. In the following we summarize the statistics of the submissions and refer the reader to the Appendix B for the trackers short descriptions.

Of the participating trackers, 1 (14%) was categorized by their authors as ST_0, 1 (14%) as ST_1, 4 (57%) were categorized as LT_0, and 1 (14%) as LT_1. This shows that long-term tracking is essential for the VOTST benchmarks. Most trackers (6; 86%) applied a locally uniform dynamic model, while (1; 14%) applied a nearly-constant velocity model. The dominant network architecture used for tracking was transformers (all the submissions utilized transformers, sometimes in a combinations with a ResNet visual encoder). All of the trackers localized the targets in a single stage, and most of them (5; 72%) directly predicted the object mask as well, while 2 (28%) relied on SAM [14] for segmentation. 3 (43%) trackers applied a fixed template updating mechanism, 3 (43%) always updated the template and 1 (14%) never updated the template. All submissions applied the same network for frame-to-frame localization and target re-detection.

The results are summarized in the tracking quality plots and AR plots (Figure 3), and in Table 2. The top 3 trackers according to the primary tracking quality score (Q) are: RMemAOT (B.5), RMem (B.4), and S3_Track (B.6). All three are single-stage tracker that jointly considers all targets within the same pipeline. The winning submission extends the popular DeAOT [48] tracker by combining it with the ideas from the recent RMem approach, which is also the second best method in this challenge. Specifically, the memory bank in DeAOT is limited to a fixed length to eliminate redundant information and a UCB-inspired [2] update mechanism from RMem is utilized to balance the relevance and freshness of frame cues. The second place tracker, RMem, focuses on the design of the memory mechanism. Their key insight is that expanding memory banks, while seemingly beneficial, actually increases the difficulty for trackers to decode relevant features due to the confusion from redundant information. They instead propose to limit the size of the memory by only including carefully selected 'essential' frames. Finally, the third place tracker is S3_Track, which is the winner of VOTS challenge and is discussed in detail in the previous section. It is notable that all three top-performing methods include sophisticated memory design (with S3_Track being based on XMem [4]), indicating the importance of maintaining an accurate model of the target as it undergoes complex transformations.

The AR plot (Figure 3) indicates that RMemAOT achieves the best combination of accuracy and robustness, closely followed by RMem. In particular, the two methods achieved almost identical robustness, which is not surprising given that they share the same memory design, but the more advanced target segmentation mechanism from DeAOT [48] allows RMemAOT to achieve higher accuracy and thus top quality score. Interestingly, top performance on auxiliary metrics such as drift (DRE) and target absence prediction (AD) is achieved by approaches that show relatively low quality scores (TAMVT and EVPTrack respectively). This indicates that simpler tracker designs can lead to fewer spurious predictions, but more sophisticated approaches are required to achieve top performance on this challenging task. Among the best trackers, RMemAOT demonstrates both lowest drift and most accurate target absence

prediction. This is due to an additional mechanism proposed by the authors where they independently run 3 trackers initialized from the first three frames and fuse their predictions to filter out unstable or anomalous tracking outcomes.

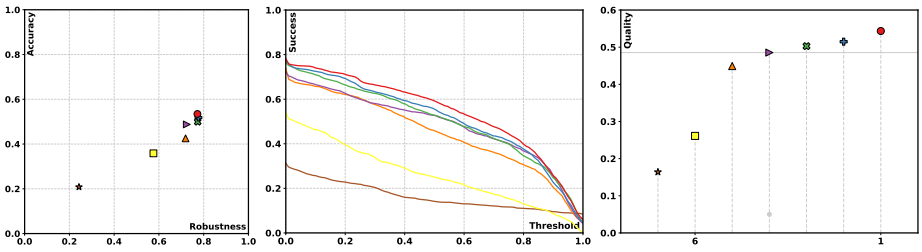

Fig. 3: The VOTSt2024 challenge accuracy/robustness AR-plots (left), tracking quality Q-plots (center) and all trackers ranked according to Q score (right).

	Quality	AR		Auxiliary measures		
Tracker	Q ↑	A ↑	R ↑	NRE ↓	DRE ↓	ADQ ↑
RMemAOT	0.544①	0.534①	0.772③	0.119	0.109②	0.653②
RMem	0.515②	0.518②	0.777①	0.099③	0.125	0.362
ASS_Track	0.503③	0.499③	0.773②	0.083②	0.144	0.287
AOTPlus	0.485	0.488	0.725	0.126	0.150	0.404
TAMVT	0.449	0.425	0.719	0.172	0.109①	0.536③
AQATrack	0.261	0.359	0.575	0.001①	0.423	0.000
EVPTrack	0.164	0.208	0.243	0.644	0.113③	0.751①

Table 2: Numerical results for VOTSt2024 challenge. Tracking quality (Q), accuracy (Acc), robustness (Rob), not-reported error (NRE), drift-rate error (DRE) and absence-detection quality (ADQ).

The VOTSt2024 committee provided a strong baseline tracker for validating that progress is achieved on this novel challenge. The tracker was adopted from the baseline proposed in the VOST paper [40] and extends the very popular AOT approach [47] by increasing the length of the training sequences and replacing the short-term memory module with a recurrent transformer. This tracker called AOTPlus B.1 reached Q=0.485, which is approximately 89% of the Q-score of the top performer. Only 3 submissions outperformed the baseline, indicating that, while non-trivial progress has been achieved, tracking objects under complex transformations remains a challenging problem for existing methods. Notably, while several methods have introduced more sophisticated memory mechanisms, fundamentally all the trackers still heavily rely on appearance-based template matching to identify the target, which is often not adequate when that appearance changes rapidly or is identical between several objects. A new paradigm

which puts more weight on spatio-temporal queues might be required to achieve major breakthroughs.

The VOTSt2024 challenge winner. The top tracker according to the tracking quality score Q is RMemAOT (B.4) and thus the VOTSt2024 challenge winner.

4 Conclusion

The second VOTS2024 challenges and results were presented. The challenges consolidate the merging of the short-term and long-term, single-target and multiple-target tracking with segmentation as the only target location specification, which was proposed in VOTS2023 [25]. In addition to the original VOTS2024 sub-challenge, which considers tracking of classical objects, a new sub-challenge VOTSt2024 was introduced that considers objects whose topology changes during tracking. Both challenges use the same evaluation methodology and remain open with public leaderboards for post-challenge submissions and evaluations.

This paper presents results of 21 trackers evaluated in the VOTS2024 challenge and seven trackers evaluated in the VOTSt2024 challenge. The performance bar in VOTS2024 challenge has been pushed substantially compared to the average performance in VOTS2023 as well as compared to the winner of VOTS2024. The winner of the VOTS2024 challenge is a single-stage tracker, which excels both in segmentation quality as well as in robustness. This indicates that it is worth further exploring pipelines that treat multi-object tracking, segmentation and (re)detection within a single framework.

In the new VOTSt challenge progress with respect to the relatively strong AOTPlus baseline has been achieved as well, primarily by designing more advanced memory mechanisms. This demonstrates that maintaining an up-to-date model of the target is key when its appearance changes rapidly. At the same time, performance on this challenge still remains relatively low, with the top tracker only achieving 75% of the quality score of the VOTS challenge winner. Robustly tracking objects under complex transformations largely remains an open problem with the optimal tracker design yet to be proposed.

The VOT initiative continues to pursue establishment of a platform for discussion and supports the tracking community with challenging general-object tracking problems, performance measures and the toolkits. As observed in several VOT and VOTS challenges so far, segmentation trackers offer remarkable robustness are localization accuracy capabilites. In the past VOT/VOTS experiments indicated that segmentation trackers outperform state-of-the-art bounding box trackers. Nevertheless, we observe that there is still substantial research being invested in the bounding box trackers. While these trackers can be thoroughly analyzed by the previous VOT bounding-box-oriented benchmarks, in particular the VOT-STB2022 [15], we note that also VOTS2023/VOTS2024 can be used by applying general object segmentation algorithms to the bounding box outputs. We thus encourage researcher to use these benchmarks in their evaluation in addition to the classical datasets, since they typically more strongly expose performance differences. They also contain many tracking situations, which are less

well represented in the classical older datasets, thus offering more room for technical advancements. Driven by the remarkable results of segmentation trackers, the VOT initiative plans to continue with the efforts in future VOTS editions, in hope to further facilitate tracking development and breakthroughs.

Statement of the co-authors contributions

The following abbreviations are used for the VOTS2024 organizers: Matej Kristan (MK), Jiri Matas (JM), Pavel Tokmakov (PT), Michael Felsberg (MF), Luka Čehovin Zajc (LCZ), Alan Lukežič (AL), Khanh-Tung Tran (TT), Xuan-Son Vu (SV), Johanna Björklund (JB), Hyung Jin Chang (HJC) and Gustavo Fernandez (GF). The authors of sufficiently well performing trackers contributed in the public paper review and tracker descriptions editing. Their contributions to individual trackers are specified in Appendix A and Appendix B.

Challenge coordination & oversight: MK, JM; Results interpretation: MK, PT, AL; Toolkit development: LČZ, AL; Paper drafting:MK, PT; Paper proofing: MK, JM, GF, MF; Coordination of public review: GF; Camera ready preparation: GF; Evaluation server implementation: KTT, XSV, JB, MF; Evaluation team supervision: LČZ; Website and design: LČZ; Sponsorship acquisition: MK, HJC; Teams coordination: MK.

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 Academic and Research Network of Slovenia (ARNES), the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. Hyung Jin Chang was supported by the Institute of Information and communications Technology Planning and evaluation (IITP) grant funded by the Korea government (MSIT) (No.2024-0-00608). Gustavo Fernández was supported by the AIT Strategic Research Program 2024. The challenge was sponsored by the Faculty of Computer Science, University of Ljubljana, Slovenia and the School of Computer Science at the University of Birmingham, UK.

A VOTS2024 submissions

This appendix summarizes the VOTS2024 challenge trackers and authors' contributions.

A.1 AQATrack for tracking and SAM for segmentation (AQATrack)

J. Xie, L. Shi, C. Xue, X. Hu, N. Li, B. Zhong
xie_jx@stu.gxnu.edu.cn, slt@stu.gxnu.edu.cn, chaocan@stu.gxnu.edu.cn, huxiantao481@gmail.com, ningli65536@mailbox.gxnu.edu.cn, bnzhong@gxnu.edu.cn
Contributions: Conceptualization, JX, LS, CX, XH ,NL; Implementation, JX, LS, CX, XH ,NL; Supervision, JX, BZ

This tracker combines AQATrack [43] and SAM [14]. AQATrack is a robust tracker with spatio-temporal transformers, which adopts simple autoregressive queries to learn spatio-temporal information effectively. Segment Anything Model (SAM) is a strong model for image segmentation. First, we use AQATrack-384 to track the targets. Then, SAM-base is used to generate a mask.

A.2 Putting the Object Back into Video Object Segmentation using high quality masks (Cutie-SAM)

G. Devarapu, J. Pochimireddy, A. Sivapuram, R. Gorthi
ee21b012@iittp.ac.in, ee20b034@iittp.ac.in, ee19D506@iittp.ac.in, rkg@iittp.ac.in
Contributions: Conceptualization, GD, JP, AS, RG; Implementation, GD, JP, AS, RG; Supervision, RG

The Cutie Tracker is an VOS algorithm that utilises deep learning to achieve high accuracy and efficiency in various scenes, addressing challenges like occlusions and fast movements. It employs ResNet-18 and ResNet-50 as backbone models for feature extraction, ensuring detailed feature capture. The architecture includes a query encoder to process the current frame and a mask encoder to process memory frames, providing historical context for consistent tracking. At its core, an object transformer with multiple attention mechanism blocks integrates object queries and object memory with pixel features, enriching them with object-level semantics. The enriched pixel features are then processed by the decoder to generate final object masks. These initial masks are refined using HQ-SAM, where a 0.1 threshold filters out low-confidence predictions. The Intersection over Union (IoU) score, measuring the overlap between predicted masks and ground truth, is calculated, and masks with an IoU score higher than 0.1 are selected.

A.3 Dual Interaction Transformer Tracker (DITrackPython)

Z. Xia, Z. Wang
xzhfirst000430@gmail.com, wangkou120453@gmail.com

Contributions: Conceptualization, ZX, ZW; Implementation, ZX, ZW; Supervision, ZX, ZW

We propose an end-to-end dual interaction transformer tracking architecture (DITrack). It firstly learns target-aware tracking features by mutual guidance, which bridges the template-search image pairs with the bidirectional interaction pathways. This pathway is based on a Vision Transformer backbone pre-trained with context autoencoder (CAE). In the following, we develop a background information fusion module that mitigates the impact of background information and improves inference efficiency. Then, a core feature interaction module is designed to refine the search region response map and filter out multiple noise peaks. The extracted core features of the target interact with the search region to achieve it. The bidirectional and core feature interaction pathways alternately propagate to learn a discriminative single peak representation.

A.4 Associating and Segmenting Objects with Transformers in High Quality (ASOT) (dmaot_sam)

E. Soltani Kazemi, I. Toubal, G. Rahmon, J. Collins, B. Hatuwal, J. Mogollon, P. Kannappan
esdft@missouri.edu, itdfh@missouri.edu, gani.rahmon@mail.missouri.edu, jrcqw5@missouri.edu, bkhcty@missouri.edu, jdmcnw@missouri.edu, pal@missouri.edu
Contributions: Conceptualization, ESK, IET, KP; Implementation, ESK, IET, JC, GR, JM, BH; Supervision, KP

We propose *Associating and Segmenting Objects with Transformers* (ASOT). ASOT is a novel method for enhancing object tracking and segmentation by integrating the strengths of *Decoupled Memory AOT* (DMAOT) [47, 48] and *Segment Anything in High Quality* (HQ-SAM) [13]. Initially, DMAOT, utilizing the Swin Transformer encoder [7], performs object tracking to identify and segment objects. The resultant object masks from DMAOT serve as visual prompts for HQ-SAM to refine the segmentation quality. To mitigate potential object ambiguity in HQ-SAM, we employ a rejection sampling technique on all mask proposals. Each proposal is evaluated based on its Intersection over Union (IoU) with the DMAOT mask, with a rejection threshold set at $\tau_{IoU} = 0.5$. The mask proposal that maximizes the IoU with the DMAOT predicted mask is selected. In scenarios where all proposals are rejected, the original DMAOT mask is retained. To ensure efficient online processing, image features are extracted using the HQ-SAM encoder once per frame, with the decoder queried once per object per frame. This method leverages the high-quality segmentation capabilities of HQ-SAM while maintaining the robust tracking performance of DMAOT. Additionally, high-confidence HQ-SAM outputs can be used as priors to guide DMAOT make better next-frame predictions for objects.

A.5 DPTrack: Dynamically Optimized Prompt for Tracking (DPTrack)

C. Xu, B. Zhong, Y. Zheng
xuchenlong@stu.gxnu.edu.cn, bnzhong@gxnu.edu.cn, 20014083057@stu.hqu.edu.cn
Contributions: Conceptualization, CX; Implementation, CX, YZ; Supervision, BZ

We propose a dynamically optimized prompt-driven tracking framework, called DPTrack, which surpasses traditional template updating strategies by capturing high-quality prompt through video streaming. DPTrack utilizes historical information to generate prompts, while leveraging the attention matrix produced during the tracking procedure to filter out noise prompts (i.e., background and outdated target parts). Finally, AlphaRefine [46] is used as a segmentation network to predict the target mask.

A.6 The short-term DropTrack with SAM-Large model model for mask refinement (DropTrackSaml)

Q. Wu, Z. Liu, J. Wan, A. Chan
qiangqwu2-c@my.cityu.edu.hk, ziquan.liu@qmul.ac.uk, jiawan1998@gmail.com, bchan@cityu.edu.hk
Contributions: Conceptualization, QW, ZL; Implementation, QW, JW; Supervision, AC

DropTrackSaml is an extension of our previous work DropTrackSamb, submitted in VOTS 2023. Our tracker comprises a ViT-based DropTrack motion module initialized with the DropMAE pre-trained weights and a Segment Anything Model (SAM) for mask prediction. Designed as a short-term tracker, it operates without online updating strategies. DropTrackSaml enhances DropTrackSamb by employing SAM-large model, which takes the predicted box from DropTrack as a prompt for high-quality mask prediction. Moreover, the SAM-large model generates three candidate mask predictions, and we select the best one using a combination of the mask prediction score from SAM-large and the IoU between the predicted mask and the predicted box. This method filters out low-quality masks and ensures the final mask adapts well to the tracker.

A.7 Explicit Visual Prompts for Visual Object Tracking (EVPTrack)

L. Shi, J. Xie, X. Hu, C. Xue, N. Li, T. Liu, B. Zhong
slt@stu.gxnu.edu.cn, xie_jx@stu.gxnu.edu.cn, huxiantao481@gmail.com, chaocan@stu.gxnu.edu.cn, ningli65536@mailbox.gxnu.edu.cn, liuting20@nudt.edu.cn, bnzhong@gxnu.edu.cn
Contributions: Conceptualization, LS, XH, NL; Implementation, LS, JX, CX, TL; Supervision, BZ

This tracker combines EVPTrack [38] and SAM [14]. Firstly, we utilize EVPTrack as our primary tracker, which is a novel explicit visual prompts framework.

It uses spatio-temporal information and multi-scale information to achieve robust tracking. Then, we use the SAM model to predict the mask. Segment Anything Model (SAM) is a strong model for image segmentation. In summary, we use EVPTrack to obtain the prediction bounding box and then use SAM to generate the mask based on the prediction bounding box.

A.8 goten_l384_fastitpn (goten_l384_2t_10_80)

B. Kang, X. Chen
kangben@mail.dlut.edu.cn, chenxin3131@mail.dlut.edu.cn
Contributions: Conceptualization, BK, XC; Implementation, BK, XC; Supervision, BK, XC

We use a single target tracker as our basic tracker. Specifically, we follow the one-stream framework of OSTrack [49] to perform feature extraction and feature fusion at the same time. The difference is that we use HiViT [51] as our Backbone and use multiple templates [45] to cope with the challenge of changes in target appearance. After Backbone, we use Center Head [49] to predict the box, and then we input the box result obtained by the single target tracker as a prompt into SAM [14] for mask prediction.

A.9 Improving DMAOT Using HQ-SAM (HQ-DMAOT)

C. Zhang
chunhui.zhang@sjtu.edu.cn
Contributions: Conceptualization, CZ; Implementation, CZ; Supervision, CZ

This model is an extension of DMAOT [47, 48] tracker. In this work, we explore leveraging the off-the-shelf HQ-SAM [13] to refine the segmentation masks predicted by DMAOT. The overall framework is based on the recent DMAOT.

A.10 LoRAT-g-378-SAM (LoRAT)

L. Lin, H. Fan, Z. Zhang, Y. Huang, Y. Wang, Y. Xu, H. Ling
lt.lin@qq.com, heng.fan@unt.edu, zhipeng.zhang.cv@outlook.com,
domaingreen2@gmail.com, wangyw@pcl.ac.cn, yxu@scut.edu.cn,
hling@cs.stonybrook.edu
Contributions: Conceptualization, LL, FH, ZZ; Implementation, LL, YH; Supervision, YW, YX, HL

Motivated by the Parameter-Efficient Fine-Tuning (PEFT) in large language models, we propose **LoRAT**, a method that unveils the power of larger Vision Transformers (ViT) for tracking within laboratory-level resources. The essence of our work lies in adapting LoRA, a technique that fine-tunes a small subset of model parameters without adding inference latency, to the domain of visual tracking. However, unique challenges and potential domain gaps make this transfer not as easy as the first intuition. Firstly, a transformer-based tracker constructs unshared position embedding for template and search image. This

poses a challenge for the transfer of LoRA, usually requiring consistency in the design when applied to the pre-trained backbone, to downstream tasks. Secondly, the inductive bias inherent in convolutional heads diminishes the effectiveness of parameter-efficient fine-tuning in tracking models. To overcome these limitations, we first decouple the position embeddings in transformer-based trackers into shared spatial ones and independent type ones. The shared embeddings, which describe the absolute coordinates of multi-resolution images (namely, the template and search images), are inherited from the pre-trained backbones. In contrast, the independent embeddings indicate the sources of each token and are learned from scratch. Furthermore, we design an anchor-free head solely based on a multilayer perceptron (MLP) to adapt PETR, enabling better performance with less computational overhead. For this challenge, we employ the Segment Anything Model (SAM) to generate mask predictions.

A.11 AOTtracker based on dynamic template updates and SAM correction (LY-SAM)

Y. Zhou, J. Zhou, T. Xu, X. Zhu, Z. Feng, X. Wu, J. Kittler
6233111075@stu.jiangnan.edu.cn, 6233114044@stu.jiangnan.edu.cn, tianyang.xu@jiangnan.edu.cn, 7191905027@stu.jiangnan.edu.cn, fengzhenhua@jiangnan.edu.cn, wu_xiaojun@jiangnan.edu.cn, j.kittler@surrey.ac.uk
Contributions: Conceptualization, YZ, JZ, TX; Implementation, YZ, JZ; Supervision, TX, XZ, ZF, XW, JK

Our tracker is developed based on Decoupled Memory AOT (DMAOT) [47, 48], which exhibits excellent performance in the VOTS2023 challenge. In addition, we employ the Segment Anything in High Quality (HQ-SAM) [13] model to refine the target's boundary segmentation using the mask and bounding box obtained from DMAOT, enhancing the model's segmentation capability. The segmentation ability of the SAM model relies entirely on the localization accuracy of DMAOT. A poorly defined bounding box can degrade the overall performance of the SAM segmentation. To address this, we incorporate the score module of the tracking model and a template updating strategy. By calculating the cosine similarity between the masks obtained from DMAOT and HQ-SAM with the initial template mask separately, we can select the more accurate mask in complex scenes. For long-term sequences, where the initial template may lose benchmarking effectiveness, we replace it with the optimal template to maintain the model's temporal consistency and accuracy.

A.12 MambaTracking: Tracking through sequence modelling of Mamba (MambaTracking)

J. Zhang, C. Liang, Y. Cui, G. Wu, L. Wang
jiamming.zhang@gmail.com, 502023330033@smail.nju.edu.cn, cuiyutao@smail.nju.edu.cn, gswu@nju.edu.cn, lmwang.nju@gmail.com
Contributions: Conceptualization, JZ, CL, GW, LW; Implementation, JZ, CL; Supervision, LW

MambaTracking takes advantage of the sequence modelling capabilities of Mamba to incorporate information from template into search, inside the backbone of VideoMamba [27]. Then, we can easily predict the tracking box through the search features. Besides, we place an Alpha Refine model on top for target segmentation. We first flatten the template and search images to tokens, add their respective position embedding, concatenate them as a whole token sequence, and put the sequence into Mamba Backbone. Finally, we split out the search feature from the token sequence, reshape tokens as 2D shapes, and input it into the corner head to predict box. Besides, we place an Alpha Refine model on top for target segmentation. Due to the bidirectional scanning mechanism of Vision Mamba backbone, we flip the features of each image individually instead of flipping the entire sequence to align the pre-training of the backbone network

A.13 ODTrack: Online Dense Temporal Token Learning for Visual Tracking (ODTrack)

Y. Zheng, B. Zhong, C. Xu, L. Chen, X. Li, S. Lai
20014083057@stu.hqu.edu.cn, bnzhong@gxnu.edu.cn,
xuchenlong@stu.gxnu.edu.cn, liangchen@stu.gxnu.edu.cn,
bruc_0619@stu.gxnu.edu.cn, laisimiao1@gmail.com
Contributions: Conceptualization, YZ; Implementation, YZ, CX, LC, XL, SL; Supervision, BZ

We propose a simple, flexible and effective video-level tracking pipeline, named ODTrack [52], which densely associates the contextual relationships of video frames in an online token propagation manner. ODTrack receives video frames of arbitrary length to capture the spatio-temporal trajectory relationships of an instance, and compresses the discrimination features (localization information) of a target into a token sequence to achieve frame-to-frame association. Finally, AlphaRefine [46] is used as a segmentation network to predict the target mask.

A.14 Associating Semantic Tracker (S3_Track)

D. Miao, X. Huang, X. Li, D. Jiang, M. Yang, Y. Wang
realmiaodeshui@gmail.com, dazseng@gmail.com, lix07@pcl.ac.cn,
jiangdm@pcl.ac.cn, mhyang@ucmerced.edu, wangyw@pcl.ac.cn
Contributions: Conceptualization, XL, DJ, MY, YW; Implementation, DM, XH; Supervision, XL, DJ, MY, YW

Our approach, named S3_Track [28], is developed to provide accurate mask predictions and achieve global search. To better handle long-term sequences, we develop a discriminative query propagation to leverage the modeling of long-term and short-term target appearances. In addition, we design a block that efficiently learns both semantic and detailed information, which can extract rich semantic features from a pre-trained Vision Transformer (ViT) without the need to train all feature extraction parameters. To enhance the performance of our

model, we utilize the MEGA dataset constructed by Cutie, which includes the YouTubeVOS [44], DAVIS [35], OVIS [36], MOSE [6], and BURST [1] datasets.

A.15 Spatio-Temporal Transformer Network for Visual Tracking with Segment Anything in High Quality (STARK_SAM_HQ)

M. Attari
ma8pz@missouri.edu
Contributions: Conceptualization, MA; Implementation, MA; Supervision, MA

This tracker leverages the capabilities of Transformer-based methods to enhance object detection and segmentation performance. Specifically, it employs the STARK (spatio-temporal transformer network for visual tracking) algorithm to accurately detect bounding boxes for individual objects within a scene. Following this, the tracker integrates the Segment Anything Model in High Quality (HQ-SAM) to generate precise segmentation masks using the detected bounding boxes. The combination of STARK and HQ-SAM ensures high accuracy in both localization and segmentation tasks, making this tracker robust and efficient for various applications in computer vision. By harnessing the strengths of Transformer architectures, this approach provides state-of-the-art performance in object tracking and segmentation.

A.16 swinb_dm_deaot_vots (swinb_dm_deaot_vots)

Y. Luo, Y. Du, G. Zhu
luoyifei2023@ia.ac.cn, duyinglong2022@ia.ac.cn, gbzhu@nlpr.ia.ac.cn
Contributions: Conceptualization, YL, YD; Implementation, YL, YD; Supervision, GZ

We came up with a video multi-target tracker, which mainly consists of a video multi-target tracker, and a mask generator. The video multi-target tracker is based on DMAOT, which is a variant of DeAOT that uses intelligent long-term memory and optimizes memory storage through similarity. We uncoupling DeAOT to make it more focused on video multi-target tracking tasks, based on tracking boxes. Use HQ-SAM for mask generation.

A.17 tapall_ai_tracker (tapall_ai)

M. Gao, K. Yu, J. Han, J. Yang, F. Zheng
mingqi.gao@tapall.ai, key022@ucsd.edu, jungonghan77@gmail.com,
jinyu.yang@tapall.ai, f.zheng@ieee.org
Contributions: Conceptualization, JH; Implementation, MG, KY; Supervision, JY, FZ

The tapall_ai_tracker is built based on Cutie [3], a memory-based model considering both pixel-level and object-level information for video object segmentation. We train Cutie with heavy data augmentation for discriminative cross-frame correspondence, such as high-dynamic resolutions. To avoid error

propagation over long-term tracking, we first parse target semantics on the first frame, with a large vision-language model [37]). Then, we filter predictions on remaining frames and only update the tracker with semantically consistent ones. This way, our tapall_ai_tracker always focuses on target-relevant clues rather than drifting to distracting ones. The backbone of Cutie is Wide-Resnet-101 [50].

A.18 VOTS2024_MIEM-HSE_AlphaCHIP: TinySAM+modified pytracking (VOTS2024_MIEM-HSE_AlphaCHIP)

A. Romanov, R. Solovyev, V. Zunin, V. Prutyanov
romeomea@gmail.com, roman.solovyev.zf@gmail.com, volodya12309@gmail.com, vvprutyanov@gmail.com
Contributions: Conceptualization, AR, RS; Implementation, AR, RS, VZ, VP; Supervision, AR

The peculiarity of our tracker lies in its high speed with acceptable accuracy. The architecture of the tracker is two-stage and consists of object localization using a bounding box (ideas from Visual Transformer [7] were used) followed by tracking and segmentation (Mask Transfiner [12]). The class of objects is determined using a visual transformer supplemented with an ANN definition and prediction of the object's position (DenseFusion [41]). Segmentation is performed using Light HQ-SAM[43] to create high-quality target masks based on the predicted bounding box. A Tiny version was chosen as the model for it as it runs much faster than others and predicts with a result comparable to other models. The model considers previous predictions and stores the history of past patterns. If the object goes beyond the boundaries of the image, the model prioritizes searching for it from the side of the last detection. An average tracking frame processing speed of 15 fps was achieved. Light HQ-SAM can be changed to another version of Segment Anything, but this reduces the speed by more than 5 times with an increase in accuracy of no more than 10%.

B VOTSt2024 submissions

This appendix summarizes the VOTSt2024 challenge trackers and authors' contributions.

B.1 AOT+ (AOTPlus)

Submitted by VOTS committee

AOTPlus [40] extends the AOT approach [47] by increasing the length of the training sequences, replacing the short-term memory module with a recurrent transformer and increasing the temporal resolution of the model at test time.

[43] https://github.com/SysCV/sam-hq?tab=readme-ov-file#model-checkpoints

B.2 AQATrack for tracking and SAM for segmentation (AQATrack)

J. Xie, L. Shi, C. Xue, X. Hu, N. Li, B. Zhong
xie_jx@stu.gxnu.edu.cn, slt@stu.gxnu.edu.cn, chaocan@stu.gxnu.edu.cn, huxiantao481@gmail.com, ningli65536@mailbox.gxnu.edu.cn, bnzhong@gxnu.edu.cn
Contributions: Conceptualization, JX, LS, CX, XH, NL; Implementation, JX, LS, CX, XH, NL; Supervision, JX, BZ

This tracker conbines AQATrack [43] and SAM [14]. AQATrack is a robust tracker with spatio-temporal transformers, which adopts simple autoregressive queries to learn spatio-temporal information effectively. Segment Anything Model (SAM) is a strong model for image segmentation. First, we use AQATrack-384 to track the targets. Then, SAM-base is used to generate a mask.

B.3 Explicit Visual Prompts for Visual Object Tracking (EVPTrack)

L. Shi
slt@stu.gxnu.edu.cn
Contributions: Conceptualization, LS; Implementation, LS; Supervision, LS

We propose a novel explicit visual prompts framework for visual tracking, dubbed EVPTrack. Specifically, we utilize spatio-temporal tokens to propagate information between consecutive frames without focusing on updating templates. As a result, we cannot only alleviate the challenge of when-to-update, but also avoid the hyper-parameters associated with updating strategies. Then, we utilize the spatio-temporal tokens to generate explicit visual prompts that facilitate inference in the current frame. The prompts are fed into a transformer encoder together with the image tokens without additional processing. Consequently, the efficiency of our model is improved by avoiding how-to-update. In addition, we consider multiscale information as explicit visual prompts, providing multiscale template features to enhance the EVPTracks ability to handle target scale changes.

B.4 RMem: Restricted Memory Banks Improve Video Object Segmentation (RMem)

J. Zhou, Z. Pang, Y. Wang
junbaoz@illinois.edu, ziqip2@illinois.edu, yxw@illinois.edu
Contributions: Conceptualization, JZ, ZP, YW; Implementation, JZ, ZP, YW; Supervision, YW

RMem rethinks how existing methods organize the memory banks, which is essential for understanding the long-term context for VOS. Specifically, we propose to use a concise restricted memory bank with a bounded size, instead of continuously accumulating frames into the memory bank. In addition, we propose to balance the relevance and freshness of features when updating the memory bank with incoming features. Finally, we demonstrate the potential of RMem's improved training/inference consistency because of the restricted

memory bank: temporal positional embedding. For more information, please refer to our paper [53].

B.5 RMemAOT: Video Object Segmentation with Restricted Memory (RMemAOT)

J. Zhu, C. Yu, X. Wang, H. Lu
jiawen@mail.dlut.edu.cn, yuchenyang@mail.dlut.edu.cn,
wangxinying@mail.dlut.edu.cn, lhchuan@dlut.edu.cn
Contributions: Conceptualization, JZ, CY, XW, HL; Implementation, JZ, CY, XW; Supervision, HL

RMemAOT is developed based on DeAOT [48] and RMem [53], trained on the training split of the VOST [40] dataset. Specifically, the memory bank of DeAOT is restricted to a fixed length to filter out redundant information. Following RMem, the UCB-inspired update mechanism and temporal positional embedding are employed to balance the relevance and freshness of frame cues. To further enhance the robustness, RMemAOT consists of three VOS models that start from the first three frames with minor topological deformation. The tracking results are assembled from the above results.

B.6 Associate semantic Track (S3_Track)

D. Miao, C. Yang, Z. Zhou, X. Li, Z. He, M. Yang
realmiaodeshui@gmail.com, 20b951014@stu.hit.edu.cn, zhouzikunhit@gmail.com,
lix07@pcl.ac.cn, zhenyuhe@hit.edu.cn, mhyang@ucmerced.edu
Contributions: Conceptualization, DM, CY, ZZ; Implementation, DM, CY, ZZ; Supervision, XL, ZH, MY

Our approach, named S3_Track, is developed to provide accurate mask predictions and achieve global search. To better handle heavy transformation sequences, we develop a discriminative query propagation mechanism to leverage the modeling of long-term and short-term target appearances. In addition, we design a block that efficiently learns both semantic and detailed information, which can extract rich semantic features from a pre-trained Vision Transformer (ViT) without the need to train all feature extraction parameters. To verify the performance of our model, we train the model on the VOST train set and test by VOT-Toolkit.

B.7 TAM-VT: Transformation-Aware Multi-scale Video Transformer for Segmentation and Tracking (tamvt)

R. Goyal, W. Fan, M. Siam, L. Sigal
rgoyal14@cs.ubc.ca, wancyuan@cs.ubc.ca, Mennatullah.Siam@ontariotechu.ca,
lsigal@cs.ubc.ca
Contributions: Conceptualization, RG, MS, WF; Implementation, RG, WF; Supervision, RG, WF, MS, LS

The tracker is based on TAM-VT, a novel clip-based DETR-style encoder-decoder architecture that leverages multi-scale matching and decoding to ensure sensitivity and accuracy for small objects. To further enable the capture of objects even during transformations, TAM-VT utilizes a novel transformation-aware loss that focuses learning on video segments where an object undergoes significant deformations. Additionally, a multiplicative time-coded memory is introduced, surpassing vanilla additive positional encoding, to help propagate context across long videos.

References

1. Athar, A., Luiten, J., Voigtlaender, P., Khurana, T., Dave, A., Leibe, B., Ramanan, D.: Burst: A benchmark for unifying object recognition, segmentation and tracking in video. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) pp. 1674–1683 (2023)
2. Auer, P.: Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research **3**(Nov), 397–422 (2002)
3. Cheng, H.K., Oh, S.W., Price, B., Lee, J.Y., Schwing, A.: Putting the object back into video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3151–3161 (2024)
4. Cheng, H.K., Schwing, A.G.: XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: ECCV (2022)
5. Davis, J.W., Gao, H.: Gender recognition from walking movements using adaptive three-mode pca. In: Comp. Vis. Patt. Recognition (2003)
6. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: Mose: A new dataset for video object segmentation in complex scenes. In: Int. Conf. Computer Vision. pp. 20224–20234 (2023)
7. 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)
8. 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)
9. Fan, H., Miththanthaya, H.A., Harshit, Rajan, S.R., Liu, X., Zou, Z., Lin, Y., Ling, H.: Transparent object tracking benchmark. In: Int. Conf. Computer Vision. pp. 10734–10743 (2021)
10. Gu, A., Dao, T.: Linear-time sequence modeling with selective state space. arXiv preprint arXiv:2312.00752 (2023)
11. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. arXiv preprint arXiv:1810.11981 (2018)
12. Ke, L., Danelljan, M., Li, X., Tai, Y.W., Tang, C.K., Yu, F.: Mask transfiner for high-quality instance segmentation. In: CVPR (2022)
13. Ke, L., Ye, M., Danelljan, M., Liu, Y., Tai, Y.W., Tang, C.K., Yu, F.: Segment anything in high quality. In: NeurIPS (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)

15. Kristan, M., Leonardis, A., Matas, J., 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 tenth visual object tracking vot2022 challenge results. In: European Conference on Computer Vision ECCV2022 Workshops (2022)
16. 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)
17. 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)
18. 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)
19. 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)
20. 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)
21. 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)
22. 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)
23. 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)
24. 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)
25. Kristan, M., Matas, J., Danelljan, M., Felsberg, M., Chang, H.J., Zajc, L.v., Lukežič, A., Drbohlav, O., Zhang, Z., Tran, K.T., Vu, X.S., Björklund, J., Mayer, C., Zhang, Y., Ke, L., Zhao, J., Fernández, Gustavo, e.a.: The first visual object tracking segmentation vots2023 challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 1796–1818 (October 2023)
26. Li, A., Li, M., Wu, Y., Yang, M.H., Yan, S.: Nus-pro: A new visual tracking challenge. IEEE-PAMI (2015)

27. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. arXiv preprint arXiv:2403.06977 (2024)
28. Li, X., Miao, D., He, Z., Wang, Y., Lu, H., Yang, M.H.: Learning spatial-semantic features for robust video object segmentation. arXiv preprint arXiv:2407.07760 (2024)
29. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. IEEE Transactions on Image Processing **24**(12), 5630–5644 (2015)
30. Lukežič, A., Trojer, Z., Matas, J., Kristan, M.: A new dataset and a distractor-aware architecture for transparent object tracking. International Journal of Computer Vision (2024). https://doi.org/https://doi.org/10.1007/s11263-024-02010-0
31. Mayer, C., Danelljan, M., Yang, M.H., Ferrari, V., Van Gool, L., Kuznetsova, A.: Beyond sot: Tracking multiple generic objects at once. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6826–6836 (January 2024)
32. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: Proc. European Conf. Computer Vision. pp. 445–461 (2016)
33. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In: ECCV. pp. 300–317 (2018)
34. Noman, M., Ghallabi, W.A., Najiha, D., Mayer, C., Dudhane, A., Danelljan, M., Cholakkal, H., Khan, S., van Gool, L., Khan, F.S.: A benchmark for visual object tracking in adverse visibility. arXiv:2208.06888 (2022)
35. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
36. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H., Bai, S.: Occluded video instance segmentation: A benchmark. IJCV **130**(8), 2022–2039 (2022)
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
38. Shi, L., Zhong, B., Liang, Q., Li, N., Zhang, S., Li, X.: Explicit visual prompts for visual object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4838–4846 (2024)
39. Smeulders, A.W.M., Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual Tracking: An Experimental Survey. IEEE PAMI **36**(7), 1442–1468 (2014)
40. Tokmakov, P., Li, J., Gaidon, A.: Breaking the object in video object segmentation. In: CVPR (2023)
41. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: Computer Vision and Pattern Recognition (CVPR) (2019)
42. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. PAMI **37**(9), 1834–1848 (2015)
43. Xie, J., Zhong, B., Mo, Z., Zhang, S., Shi, L., Song, S., Ji, R.: Autoregressive queries for adaptive tracking with spatio-temporal transformers. In: Proceedings

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19300–19309 (2024)
44. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. In: ECCV (2018)
45. 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)
46. 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)
47. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. In: NeurIPS. vol. 34 (2021)
48. Yang, Z., Yang, Y.: Decoupling features in hierarchical propagation for video object segmentation. Advances in Neural Information Processing Systems **35**, 36324–36336 (2022)
49. 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)
50. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
51. Zhang, X., Tian, Y., Xie, L., Huang, W., Dai, Q., Ye, Q., Tian, Q.: Hivit: A simpler and more efficient design of hierarchical vision transformer. In: The Eleventh International Conference on Learning Representations (2023)
52. Zheng, Y., Zhong, B., Liang, Q., Mo, Z., Zhang, S., Li, X.: Odtrack: Online dense temporal token learning for visual tracking. arXiv preprint arXiv:2401.01686 (2024)
53. Zhou, J., Pang, Z., Wang, Y.X.: Rmem: Restricted memory banks improve video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18602–18611 (2024)