Advanced CV methods
Performance evaluation for general object trackers

Matej Kristan

Visual Cognitive Systems Laboratory
Faculty of computer and information science
University of Ljubljana, Slovenia

Emergence of VOT initiative

„Although tracking itself is by and large a solved problem...“,
-- Jianbo Shi & Carlo Tomasi CVPR1994 --

- ~100 tracking papers published annually
- Nonstandard evaluation, source code scarce (before 2013)
- The VOT initiative (February 2013)
- Partners: FRI-UL (SLO), UB (UK), CTU (CZ), AIT (A), LU (S), NICTA (AU), TUT (FI)
- Goal: Establish evaluation standards -> development of trackers
- Problem: Tracking community not tightly integrated

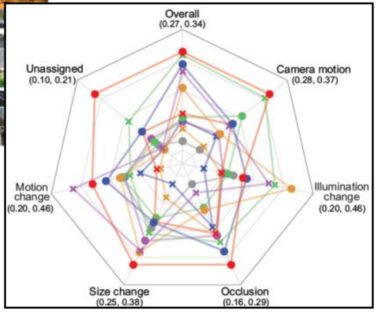
Technical advancements in performance evaluation

Discussion with Tracking community

The four pillars of VOT

- Datasets
- Evaluation methodology
- Evaluation system
- Organization of the VOT challenges

VOT toolkit



VOT2013 benchmark
The first challenge introduced a new evaluation kit plus 16 well-known short videos. 27 single-target trackers submitted by 51 participants participated at the challenge. The results were published in a joint paper presented at an ICCV2013 workshop which was attended by over 70 researchers.

VOT2014 benchmark
The second challenge introduced several improvements in annotations and testing of statistical significance, new set of 25 sequences and an improved evaluation kit. The results were published in a joint paper presented at an ECCV2014 workshop.

VOT2015 benchmark
The third challenge introduced a dataset of 60 challenging sequences, a formalized sequence selection methodology and improvements to evaluation methodology. The results were published in a joint paper presented at an ICCV2015 workshop.

VOT2016 benchmark
The fourth challenge updated the dataset of 60 sequences with new annotations. The results were published in a joint paper presented at a workshop at ECCV2016.

VOT2017 challenge
The VOT2017 challenge will be the 5th visual object tracking challenge. Results will be presented at VOT workshop at ICCV2017. This year the VOT dataset has been refreshed, the winner will be determined on sequestered dataset and a real-time experiment has been introduced.

VOT2018 challenge
The VOT2018 challenge is announced. Stay tuned for more information.

VOT2019 challenge
The VOT2019 challenge will address short-term, long-term, real-time, RGB, RGBT and RGBD trackers. Results will be presented at ICCV2019 VOT workshop.

VOT2020 benchmark
The VOT2020 benchmark addresses short-term, long-term, real-time, RGB, RGBT and RGBD trackers. Results were presented at the ECCV2020 VOT workshop.

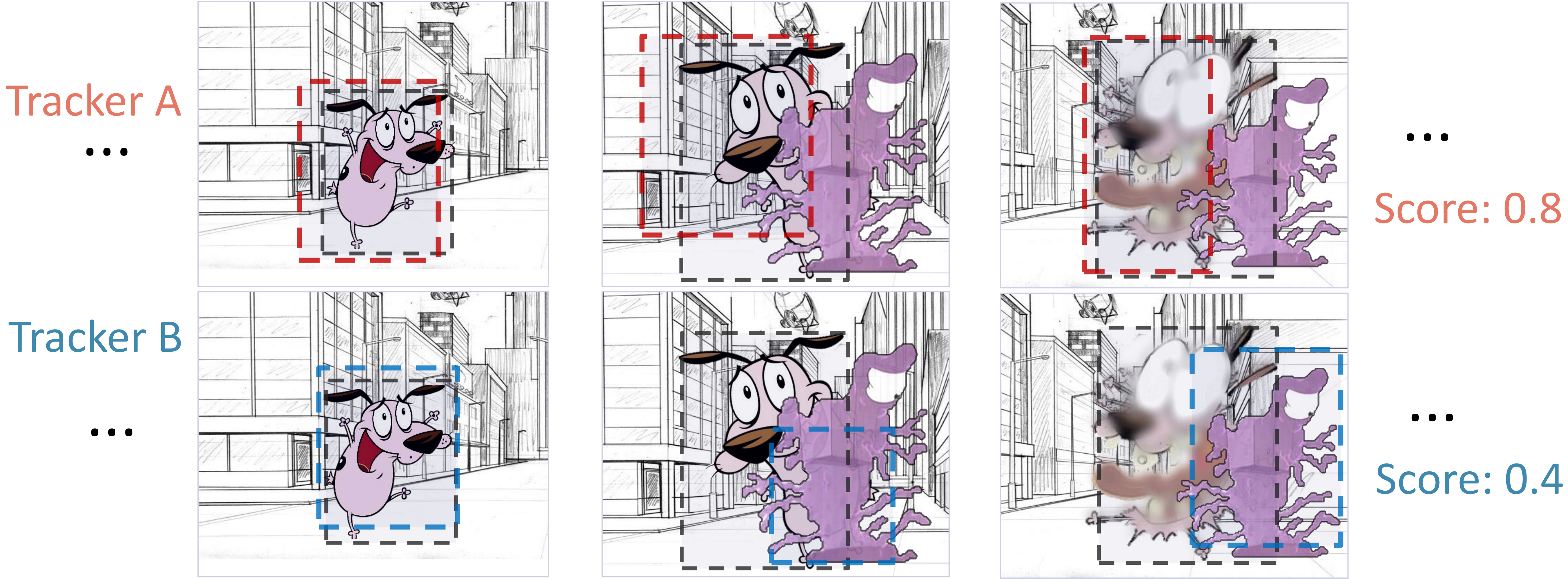
VOT2021 challenge
The VOT2021 challenge addresses short-term, long-term, real-time, RGB and RGBD trackers. Results will be presented at the ICCV2021 VOT workshop.

VOT2022 challenge
The VOT2022 challenge addresses short-term, long-term, real-time, RGB and RGBD trackers.

VOTS2023
The VOTS2023 challenge unifies single/multiple-target, short/long-term tracking and segmentation. Results were presented at the ICCV2023 VOTS workshop.

The purpose of performance evaluation

- Considering two trackers, which one better localizes the object?
- Considering a single tracker, provide a value of tracking success.



EVALUATION METHODOLOGY

(GENERAL OBJECT TRACKERS)

Historical performance measure types: Center error

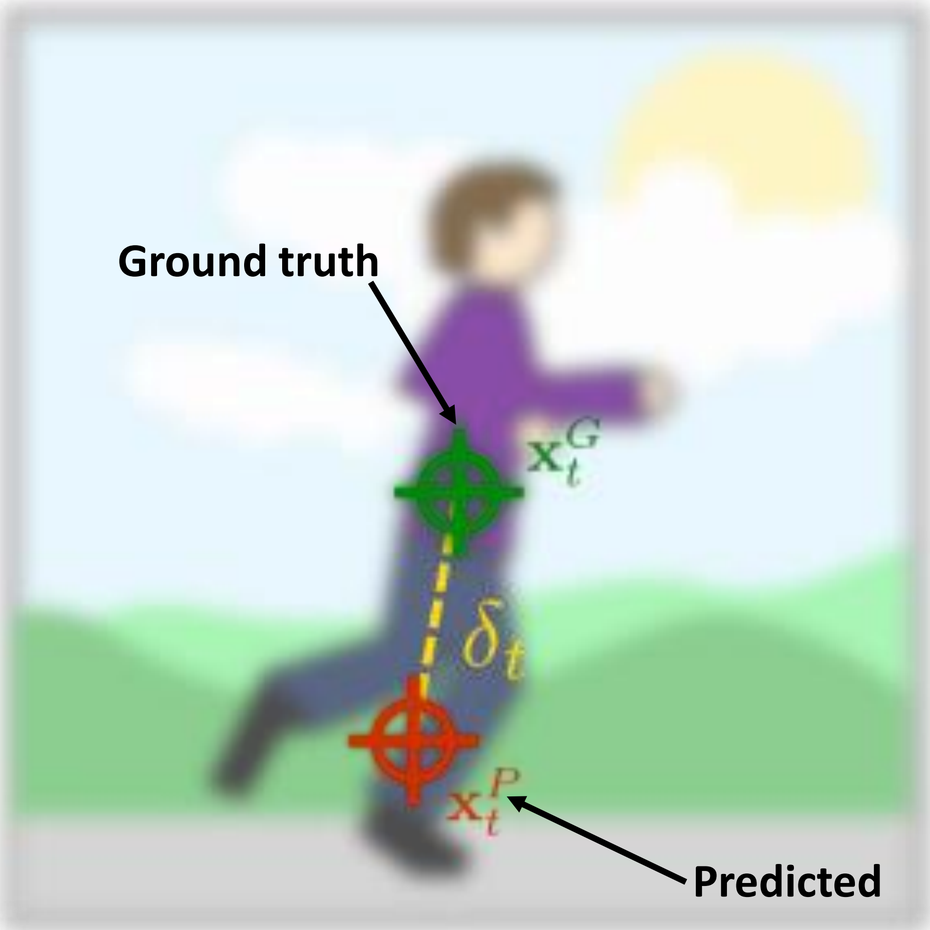
- Distance between ground truth center position and position predicted by the tracker

$$\Delta(\Lambda^G, \Lambda^P) = \{\delta_t\}_{t=1}^N, \quad \delta_t = \|\mathbf{x}_t^G - \mathbf{x}_t^P\|$$

- Summarized as
  - Root-mean-squared error

$$E = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\delta_t^2}$$

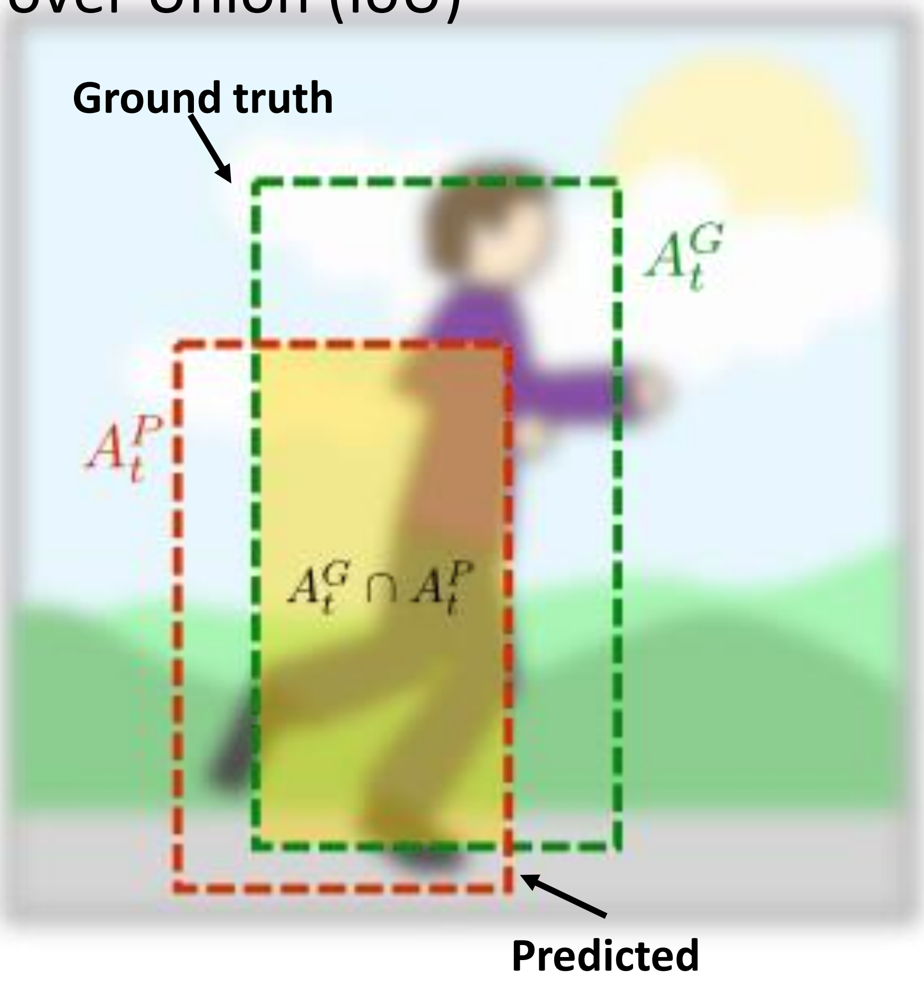
- Drawbacks
  - Does not take into account the size of the object



Measure types: Overlap error

- Overlap between the ground-truth region for the object and the region, predicted by a tracker measured as an Intersection over Union (IoU)

$$\Phi(\Lambda_G, \Lambda_P) = \left\{ \frac{A_t^G \cap A_t^P}{A_t^G \cup A_t^P} \right\}_{t=1}^{N}$$



- Advantages
  - Takes into account the target‘s size
  - Does not compare only estimations of the target center, but the entire bounding box

Measure types: Overlap error

- Overlap between the ground-truth region for the object and the region, predicted by a tracker measured as an Intersection over Union (IoU)

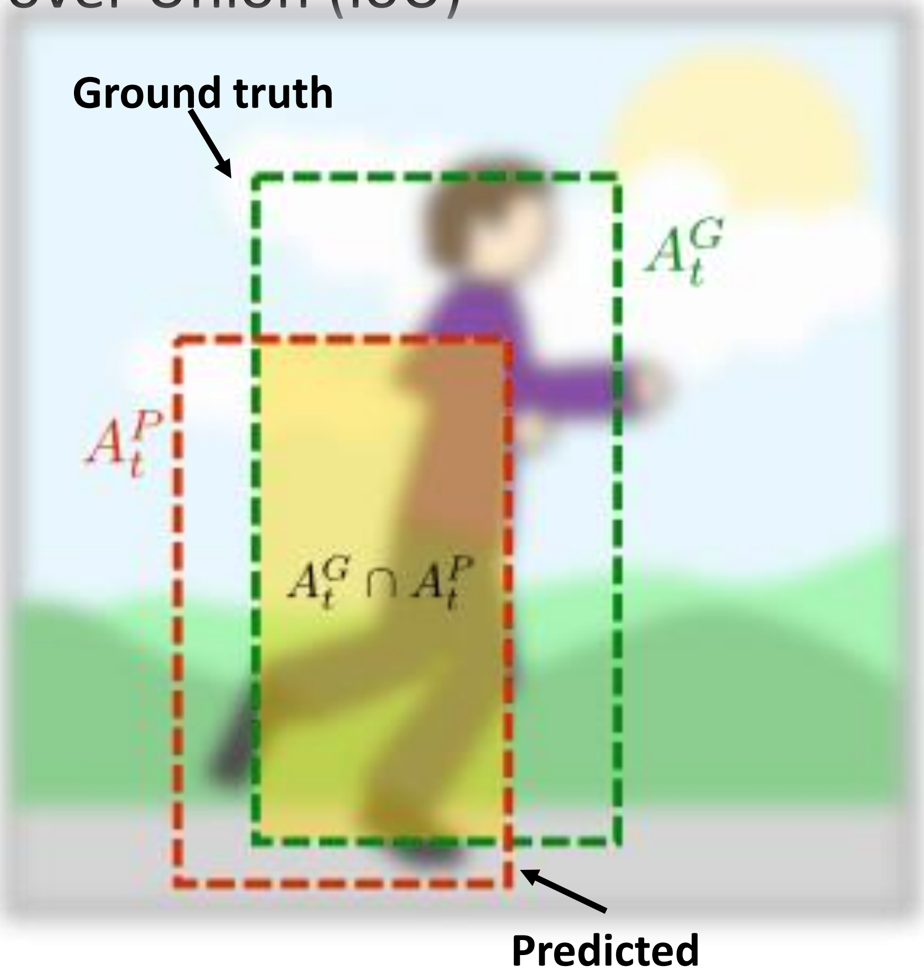
$$\Phi(\Lambda_G, \Lambda_P) = \left\{ \frac{A_t^G \cap A_t^P}{A_t^G \cup A_t^P} \right\}_{t=1}^{N}$$

- Summarized as either
  1. Average overlap

$$E = \frac{1}{N} \sum_{t=1}^{N} \Phi_t$$

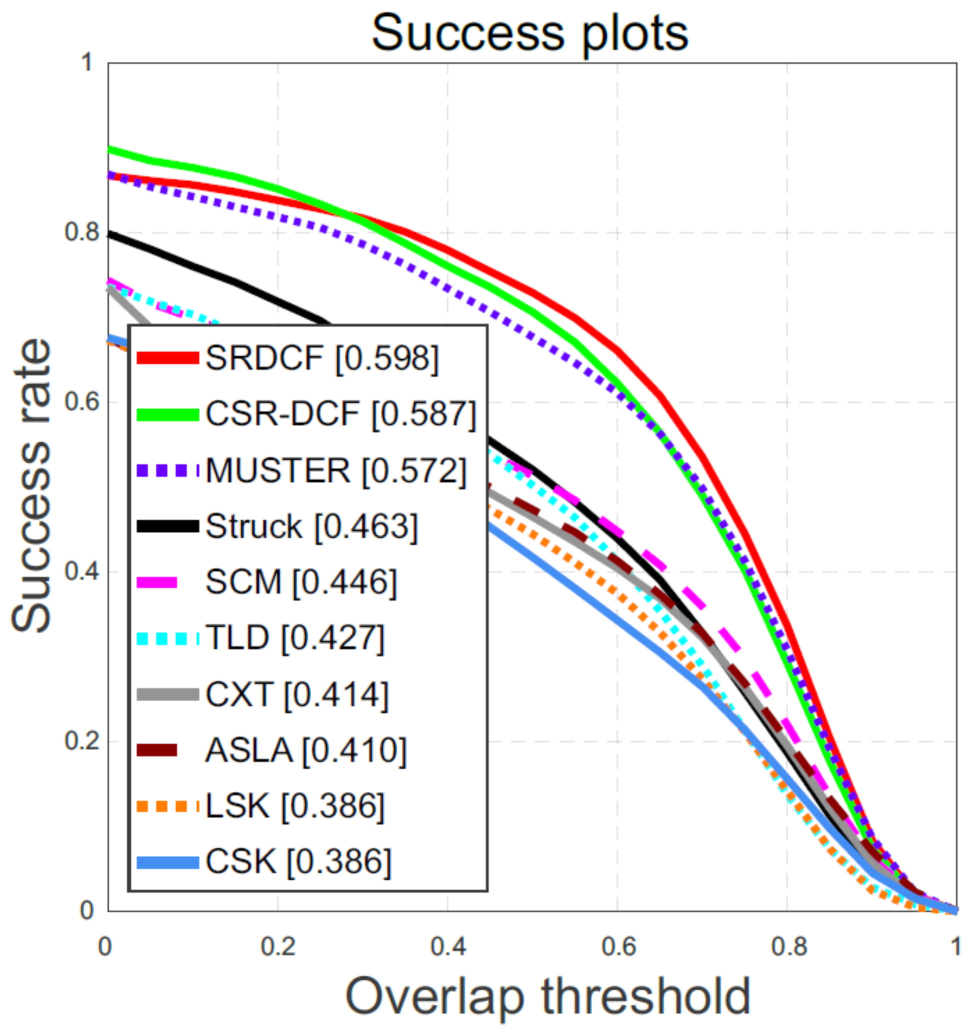
  2. Number of correctly tracked frames

     Number of times when the overlap between the ground truth and the predicted bounding box was sufficiently high, e.g., $\Phi_t > 0.5$.



Measure types: Success plot

- Most popular measure with a simple experimental setup (popularized by [1])
- A tracker is initialized and run until the end of the sequence
- Performance is visualized as portion of frames with overlap > θ_{th}
- The measure: Area under the curve *AUC* (shown[2] to be equal to average overlap)
- Many other measures explored since, by the VOT initiative (see https://www.votchallenge.net/)

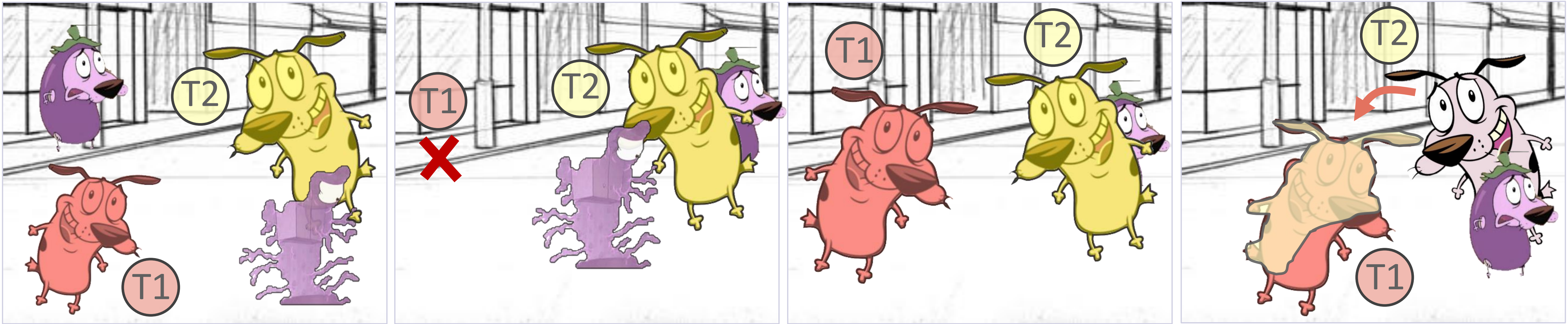


[1]Wu et al. Online Object Tracking: A Benchmark, CVPR 2013

[2]Čehovin Zajc, Leonardis, and Kristan, Visual object tracking performance measures revisited, IEEE TIP 2016

Beyond short-term single-target tracking measures

- General object Short/Long-term, Single/Multi-target trackers
- Initialize on all targets in the first frame and report position in the rest

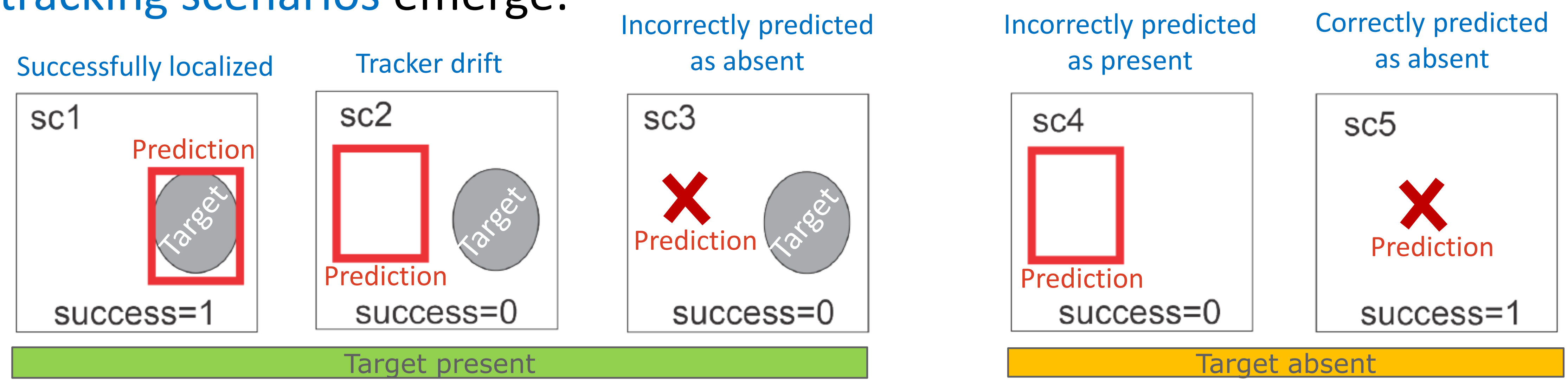


- LT requirement: Determine the target absence and redetect when it reappears
- Drifting off the target to background or another object is considered failure
- A measure introduced in VOTS2023[1]

[1]Kristan et al., The First Visual Object Tracking Segmentation VOTS2023 Challenge Results, ECCVW2023

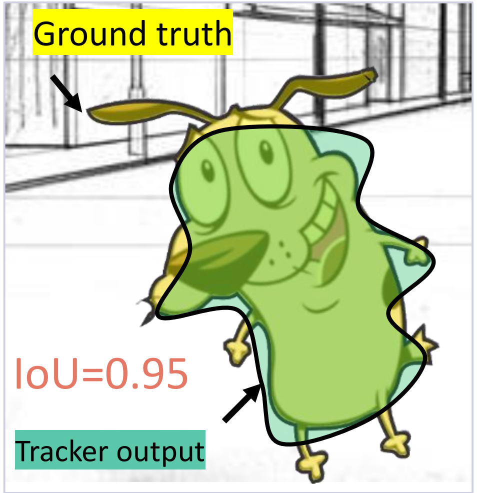
VOTS: Per-target performance measures

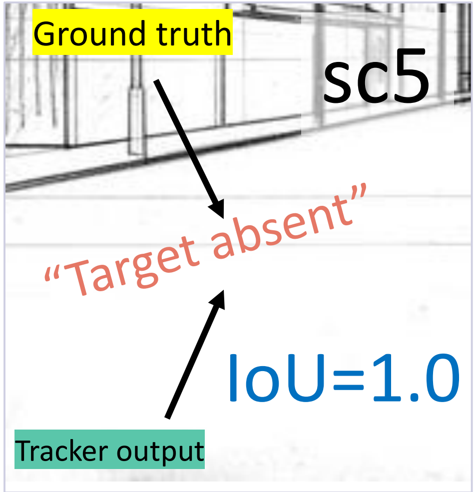
5 tracking scenarios emerge:



- IoU as a standard measure of agreement between prediction and GT
- Require IoU value definition for sc5

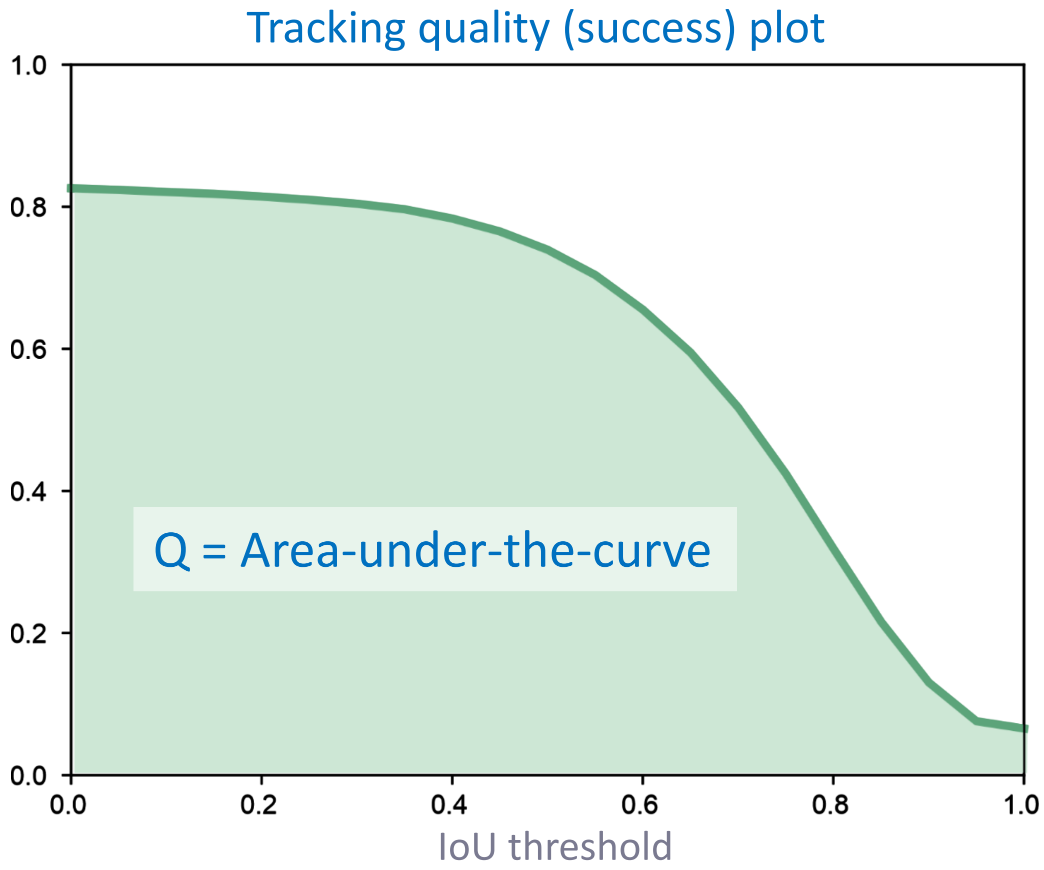
$\text{IoU}_{sc5}=1.0$





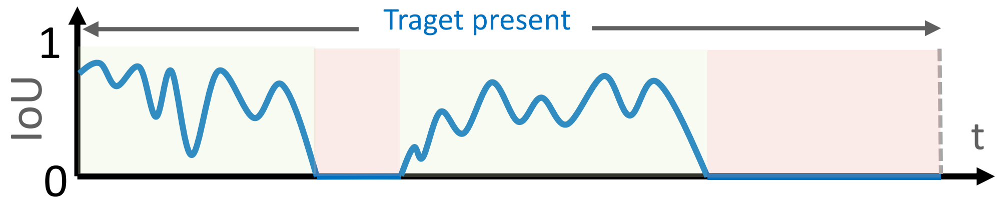
VOTS: Primary performance measure

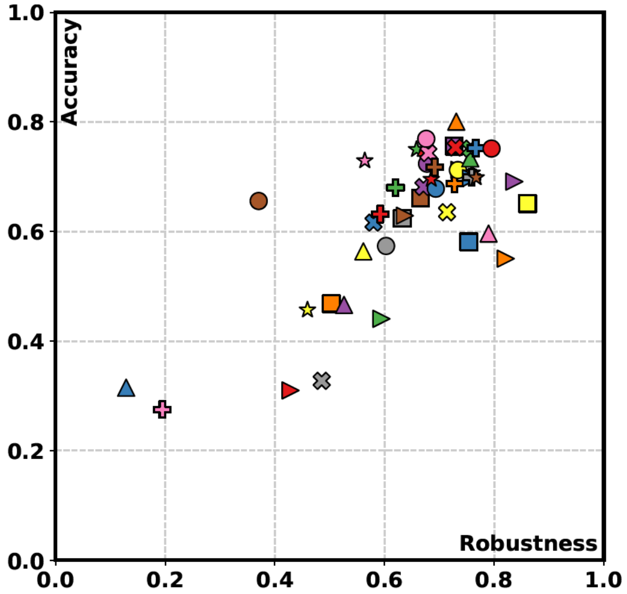
- Performance summarized by the classical success w.r.t IoU plot (i.e., tracking quality plot)
- Success plot calculated individually for each target in each sequence and then averaged
- Primary measure: *Tracking quality Q* (area-under-the-curve)



VOTS: Auxiliary performance measures

- Accuracy/Robustness[1] (@IoU=0.0 when target present)





"Why did the tracker fail while target visible?"

- NotReportedError (NRE): % frames incorrectly predicted target absent
- DriftRateError (DRE): % frames tracker drifted while predicting target present

"How well is target absence determined?"

- AbsenceDetectionQuality (ADQ): % frames target correctly predicted absent

[1] Čehovin Zajc, Leonardis, and Kristan, Visual object tracking performance measures revisited, IEEE TIP 2016

DATASETS

(GENERAL OBJECT TRACKERS)

Currently common tracking benchmarks (modulo VOT)

- Short-term tracking:
  - OTB100[1]: 100 videos, apart from VOT, longest-standing benchmark, outdated now
  - GOT10k[2]: 180 test videos, >10k all videos, highly popular in short-term tracking
  - TrackingNet[3]: 500 videos from YouTube, somewhat skewed content distribution
- Long-term tracking:
  - LaSOT[4]: 280 test videos, average sequence > 2500 frames long
  - UAV123[5]: 123 videos from low-altitude UAVs, average length ~900 frames

[1]Wu et al., Object tracking benchmark. *TPAMI* 2015
[2]Huang et al., Got-10k: A large high-diversity benchmark for generic object tracking in the wild, TPAMI 2021
[3]Muller et al., TrackingNet: A large-scale dataset and benchmark for object tracking in the wild, ECCV2018
[4]Fan et al., Lasot: A high-quality benchmark for large-scale single object tracking, CVPR2019
[5]Muller et al., A benchmark and simulator for UAV tracking, *ECCV*2016

Significant efforts invested by the community

- A common approach

  [Wu et al. CVPR2013, Smeulders et al. PAMI2013, Wang et al. arXiv2015, Wu et al. PAMI2015, …]:

  - Large datasets by collecting many sequences from internet
  - Large dataset $\neq$diverse nor useful

- The VOT approach:
  - Keep it sufficiently small, diverse and well annotated
  - Developed the VOT dataset construction methodology
  - Developed the VOT annotation methodology

The VOT(2015) dataset construction methodology

- Requirements:
  - Diversity in attributes
  - Challenging sequences

ALOV (315 seq.) [Smeulders et al.,2013]
+ OTB (~100 seq.) [Wu et al.,2015]
+ PTR (~50 seq.) [Vojir et al.,2013]
+ >50 new sequences = ~600

~400 sequences

11 dim

11 global attributes (blur, cam motion, etc.)

Clustering: Affinity Propagation [Frey, Dueck 2007]

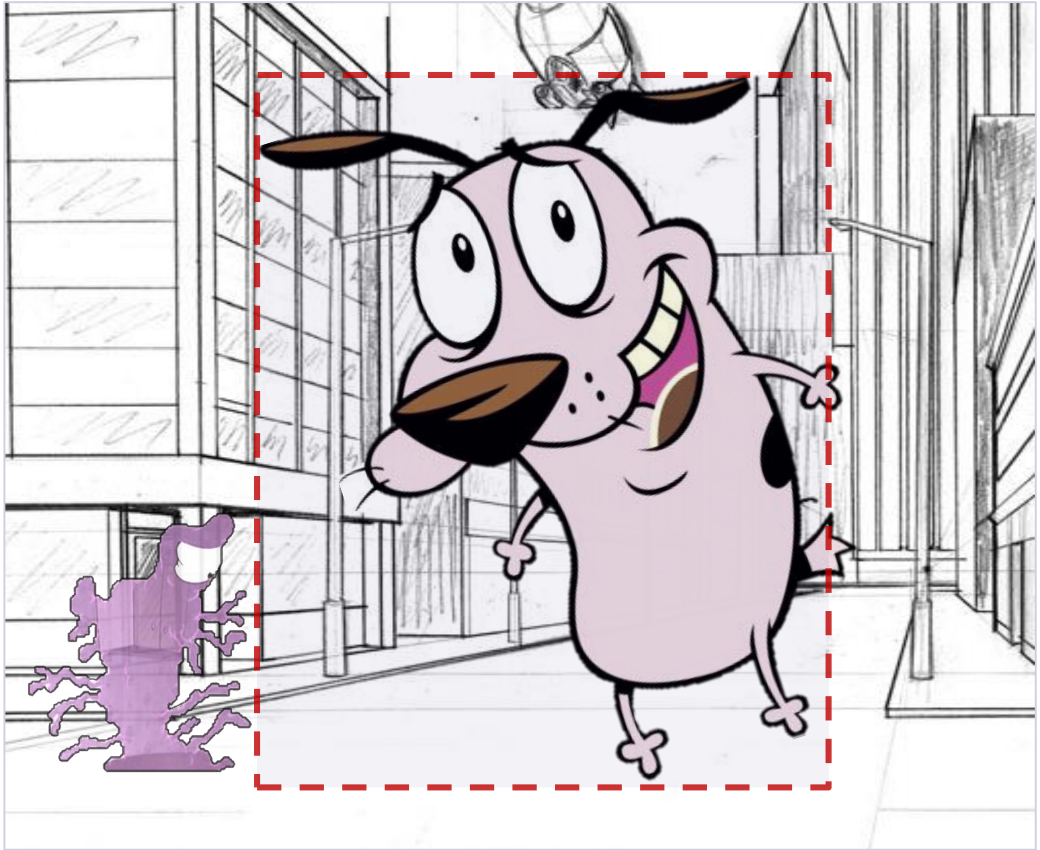
Tracking difficulty estimation of each sequence by standard trackers.

Sampling approach, samples difficult sequences and keeps diversity in attributes

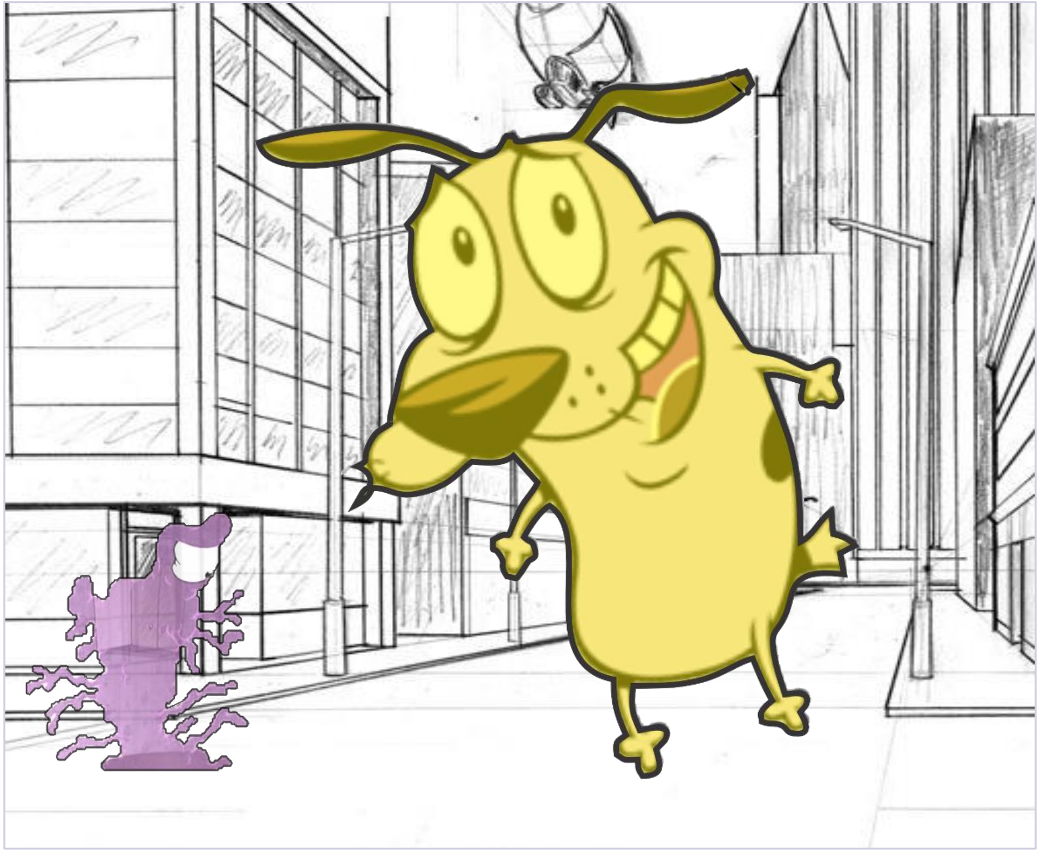
60 sequences

VOT2020 Paradigm shift – revisiting target pose

Bounding box == pose approximation



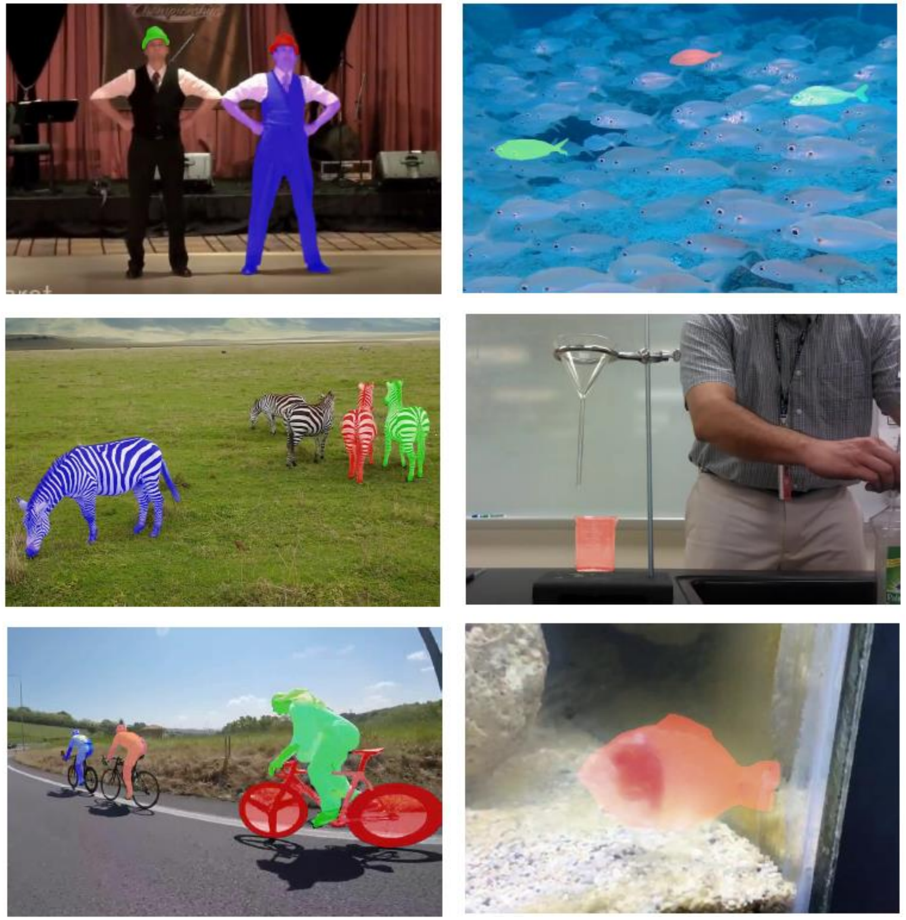
Most accurate pose == segmentation



- Emergence of end-to-end trainable general object segmentation trackers: SiamMask [Wang et al., CVPR2019] & D3S [Lukezic et al., CVPR2020]

VOTS2023 (test) dataset

- Source: LaGOT[1], UTB180[2], TOTB[3], VOT-LT2021,VOT-LT2022, VOT-ST2022
- Selection criteria:
  - Sequences challenging for modern architectures
  - Properties: (i) visually-similar objects, (ii) substantial appearance changes, (iii) cluttered background, (iv) entering-exiting field-of-view
  - Diverse object and scene types (Air, Ground, Underwater)
  - Opaque as well as transparent objects
- Annotation: Segmentation masks
  - Include parts of objects as targets



[1] Mayer et al. ArXiv 2023; [2]Alawode et al. ACCV2022; [3]Fan et al. ICCV2021

VOTS2023 (test) dataset

- Stats: 144 sequences ; 341 targets ; 168 targets leave the FOV at least once
- Sequence properties:
  - min/max = 63/10.7k frames
  - On average 2.37 targets per sequence annotated
  - Median target absence: 18 frames
- To prevent overfitting:
  - Sequences + initialization frames GT publicly available.
  - GT of test frames sequestered, evaluation carried out on a dedicated server.

Importance of training datasets

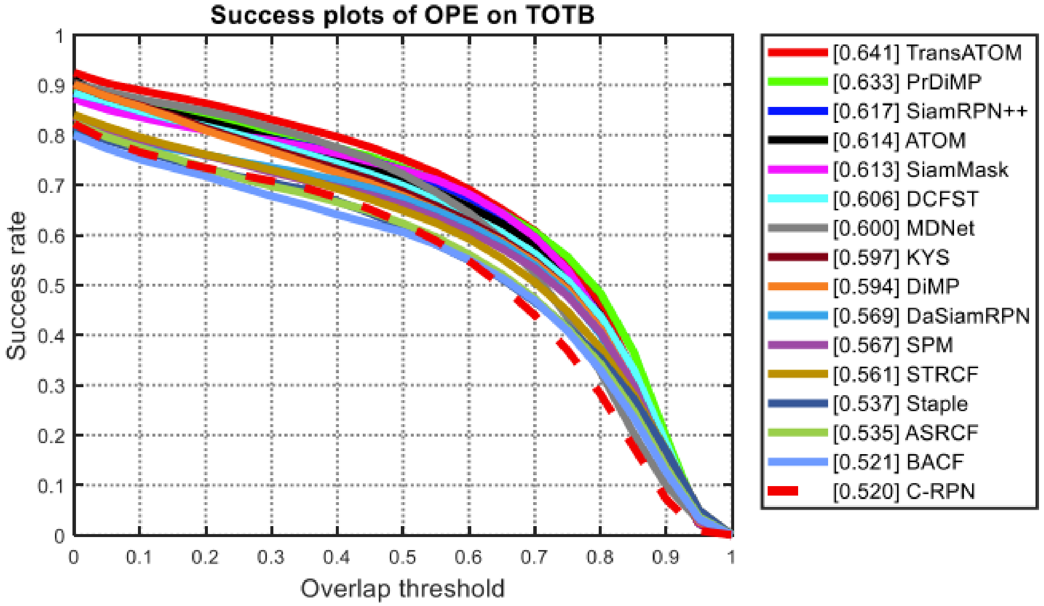
- Currently commonly used single-target training datasets:
  - TrackingNet[1]: 30k training videos from YouTube, box GT
  - GOT10k[2]: ~10k training videos, box GT
  - LaSOT[3]: >1k training videos, box GT
  - COCO[4]: 330k *images*, object detection dataset, augmentation to simulate pairs
  - YoutubeVOS[5]: 3.5k training segmentation videos

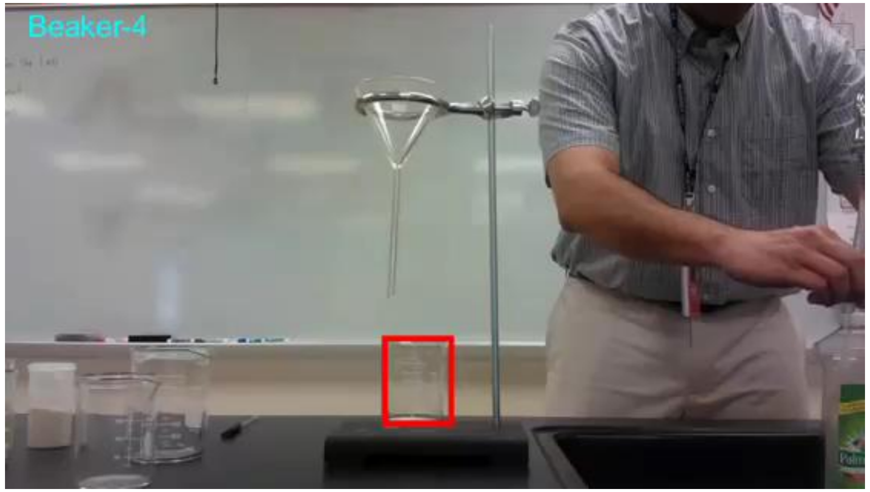
- Evidence emerging that unsupervised pre-training of the tracking architectures leads to improved performance!

[1]Muller et al. ECCV2018 ; [2]Huang et al. TPAMI 2021; [3]Fan et al. CVPR2019 ; [4]Lin et al. ECCV2014; [5]Xu et al., ECCV2018

Importance of training datasets: TOTB example

- Recently a transparent-object tracking benchmark TOTB[1] emerged
- Conjecture of the paper:
  "Classical trackers developed for opaque object tracking significantly underperform!"

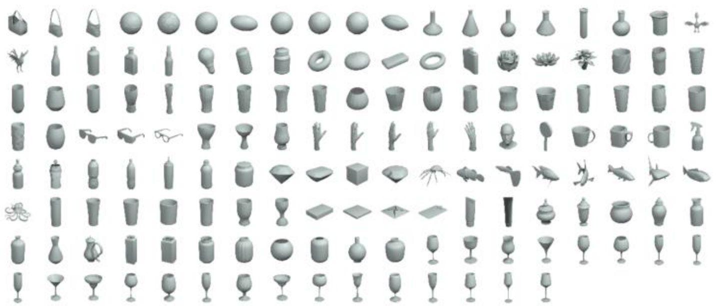


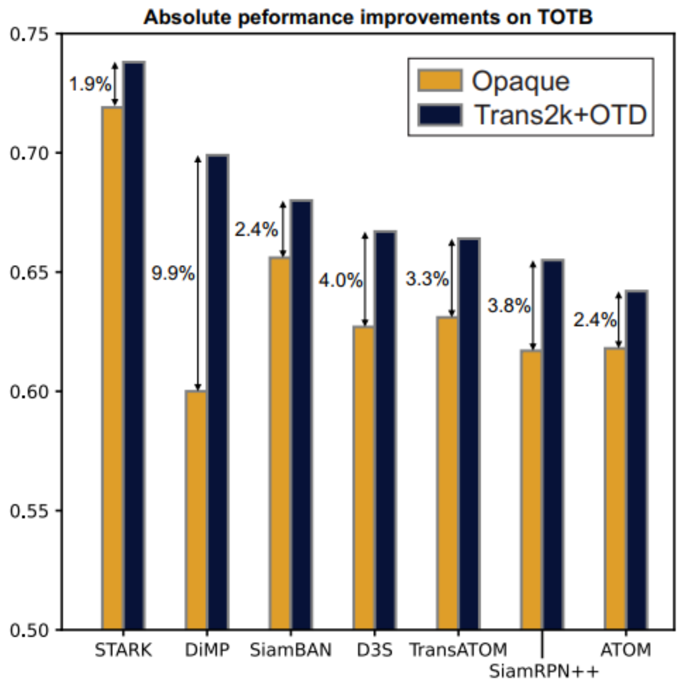


[1] H. Fan, et al., Transparent Object Tracking Benchmark, ICCV 2021

Trans2k: transparent object training dataset

- Trans2k[2] training dataset:
  - Background: videos from GoT-10k
  - Motion: Random periodic trajectory
  - Rendering engine: BlenderProc [1]
  - 2000 sequences (100k frames)
  - Bounding box + segmentation





- Training on Trans2k leads to up to 10 percentage points performance improvements (~16% boost!)

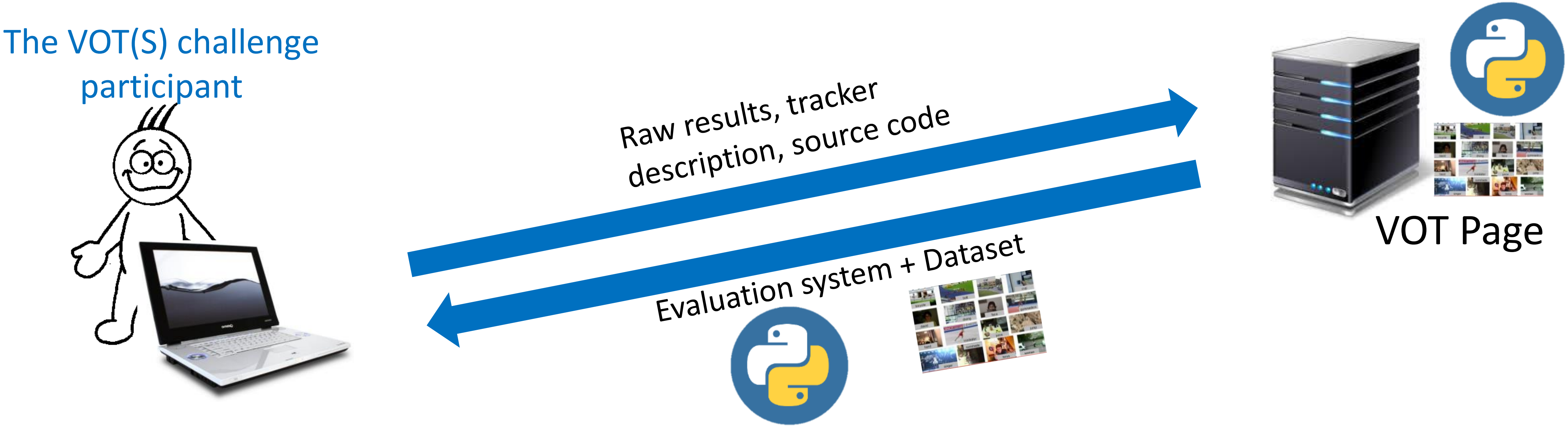
[1]M. Denninger, et al., Reducing the reality gap with photorealistic rendering, ICRSS, 2020

[2]Ž. Trojer, A. Lukežič, J. Matas, M. Kristan, Trans2k: Unlocking the Power of Deep Models for Transparent Object Tracking, BMVC2022, (best paper award), (GIT)

Visual Object Tracking Challenge VOT

VOT CHALLENGES AND BENCHMARKS

The VOT (&VOTS) challenges



- Organization of VOT workshops within ECCV/ICCV
- A paper summarizing the submitted results
  - Participants of sufficiently well performing trackers become coauthors
  - Public release of the submitted tracker code required for the winning position of the competition (since 2017)

A decade of VOT challenges

	Perf. Measures	Dataset size	Target box	Property	Trackers tested
VOT2013	ranks, A, R	16, manual select.	manual	per frame	27
VOT2014	ranks, A, R, EFO	25, manual select.	manual	per frame	38
VOT2015	EAO, A, R, EFO	60, fully auto	manual	per frame	62 VOT, 24 VOT-TIR
VOT2016	EAO, A, R, EFO	60, fully auto	auto	per frame	70 VOT, 24 VOT-TIR
VOT2017	EAO, A, R, EAO_{rt}	60, fully auto + 60 sequestered	auto	per frame	51 VOT / VOT-RT, 10 VOT-TIR
VOT2018	EAO, A, R, EAO_{rt}, LT	60, + sequestered	auto	per frame	72 VOT/VOT-RT ; 15 VOT-LT
VOT2019	EAO, A, R, EAO_{rt}, LT	60, + sequestered	auto	per frame	ST, RT, LT, RGBD-LT, RGBT-ST
VOT2020	*ST Anchor-based*	60, + sequestered		per frame	ST, RT, LT, RGBD-LT, RGBT-ST
VOT2021	*ST Anchor-based*	60, +sequestered		per frame	ST, RT, LT, RGBD-LT
VOT2022	*ST Anchor-based*	60, +sequestered		per frame	STs, STb, RT, LT, RGBD-ST
VOTS2023	*ST/LT, Single/multi target*	144 (gt sequestered)		na	47

- Gradual increase of dataset size and quality
- Gradual refinement of dataset construction
- Gradual refinement of performance measures
- Gradual increase of sub-challenges

Evolution of VOT ST challenge submitted trackers

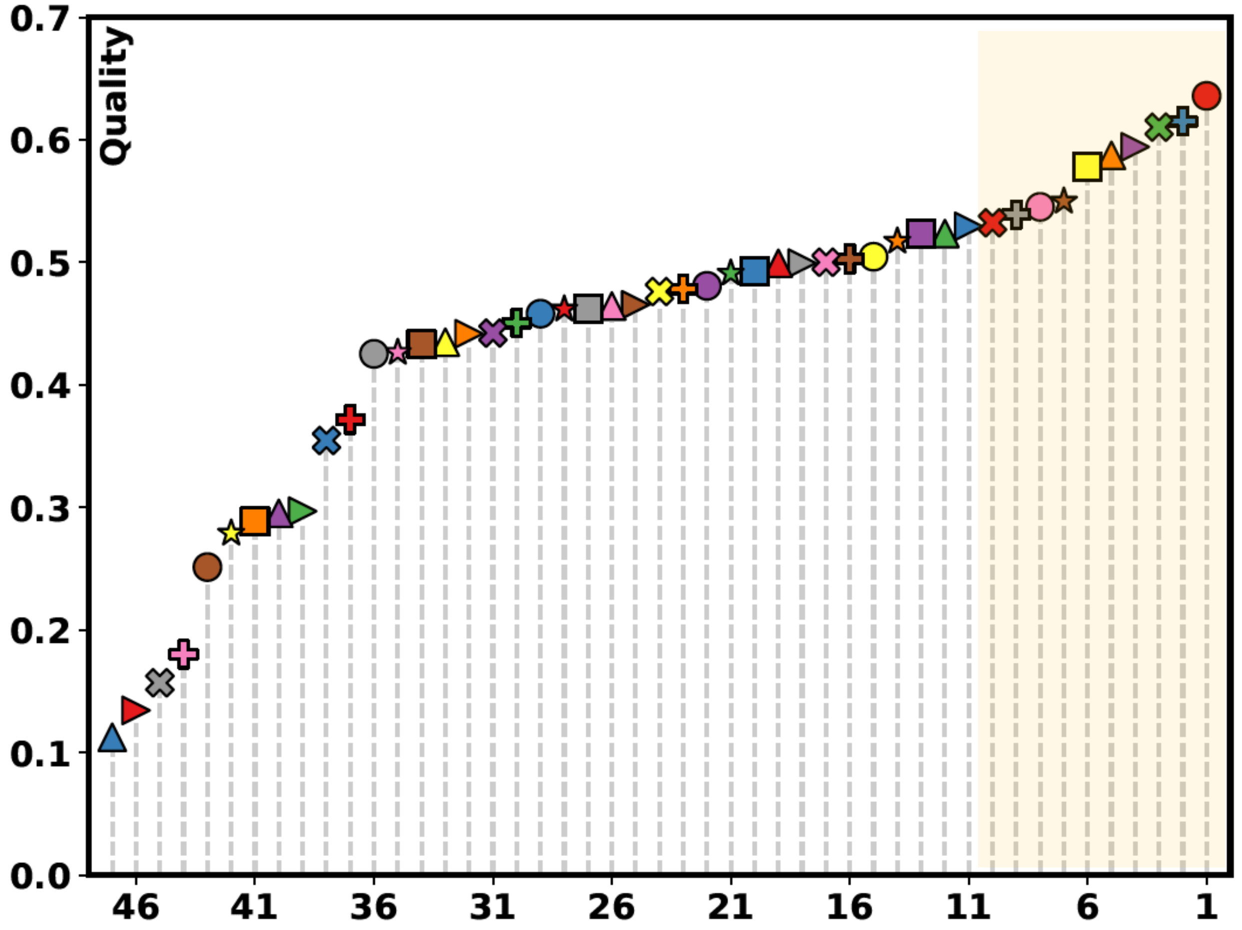
	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Num. trackers	27	38	62	70	51	72	57	37	53	31	
Submitted trackers design types	diverse	8 Dis… tive (sS…) 11 … ive 6 F… d [m… e]	16 DCF [many …	14 CNN 27 DCF …	17 CNN 25 DCF [s… d…	45 CNN 38 DCF	21 Siamese 24 … DCF	68% DCF 46% Siamese	90% CNN 56% DCF 25% … 17% … mer (64% … on)	35% DCF 45% transformers	
Top performers	Diverse	3 DCF	2 CNNs 1 sSVM	DCF+CNN CNN, DCF	DCF+CNN CNN, DCF	DCF+CNN, Siamese, DCF	Deep DCF RPN Siamese	Deep DCF + Segmentation	Transformers	Transformers	

Kristan et al., "The Visual Object Tracking VOT2013 challenge results," ICCV Workshops 2013
Kristan et al., "The Visual Object Tracking VOT2014 challenge results," ECCV Workshops 2014
Kristan et al., "The Visual Object Tracking VOT2015 challenge results", ICCV Workshops 2015
Kristan et al., "The Visual Object Tracking VOT2016 challenge results", ECCV Workshops 2016
Kristan et al., "The Visual Object Tracking VOT2017 challenge results", ICCV Workshops 2017
Kristan et al., "The Visual Object Tracking VOT2018 challenge results", ECCV Workshops 2018
Kristan et al., "The Seventh Visual Object Tracking VOT2019 challenge results", ICCV Workshops 2019
Kristan et al., "The Eighth Visual Object Tracking VOT2020 challenge results", ECCV Workshops 2020
Kristan et al., "The Ninth Visual Object Tracking VOT2021 challenge results", ICCV Workshops 2021
Kristan et al., "The Tenth Visual Object Tracking VOT2022 challenge results", ECCV Workshops 2022
Kristan et al., "The First Visual Object Tracking Segmentation VOTS2023 Challenge Results", ECCVW2023
Kristan et al., "A Novel Performance Evaluation Methodology for Single-Target Trackers", IEEE TPAMI 2016

VOTS2023 challenge results: 47 trackers tested

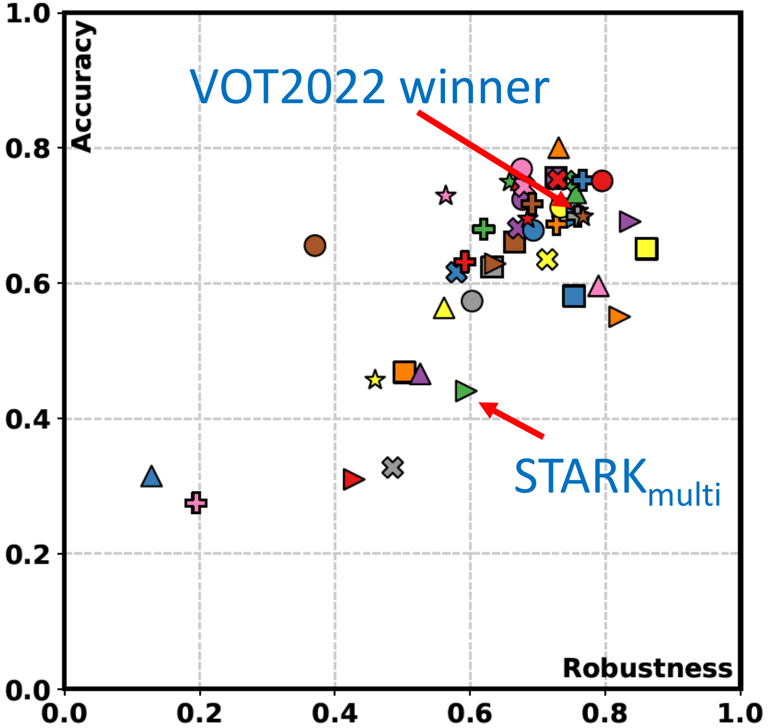
- Top trackers: DMAOT, HQTrack, MVOSTracker, Dynamic$_{DEAOT}$, seqtrack, DMNet, aot, MCMOT, rts_rts50_002, VAPT

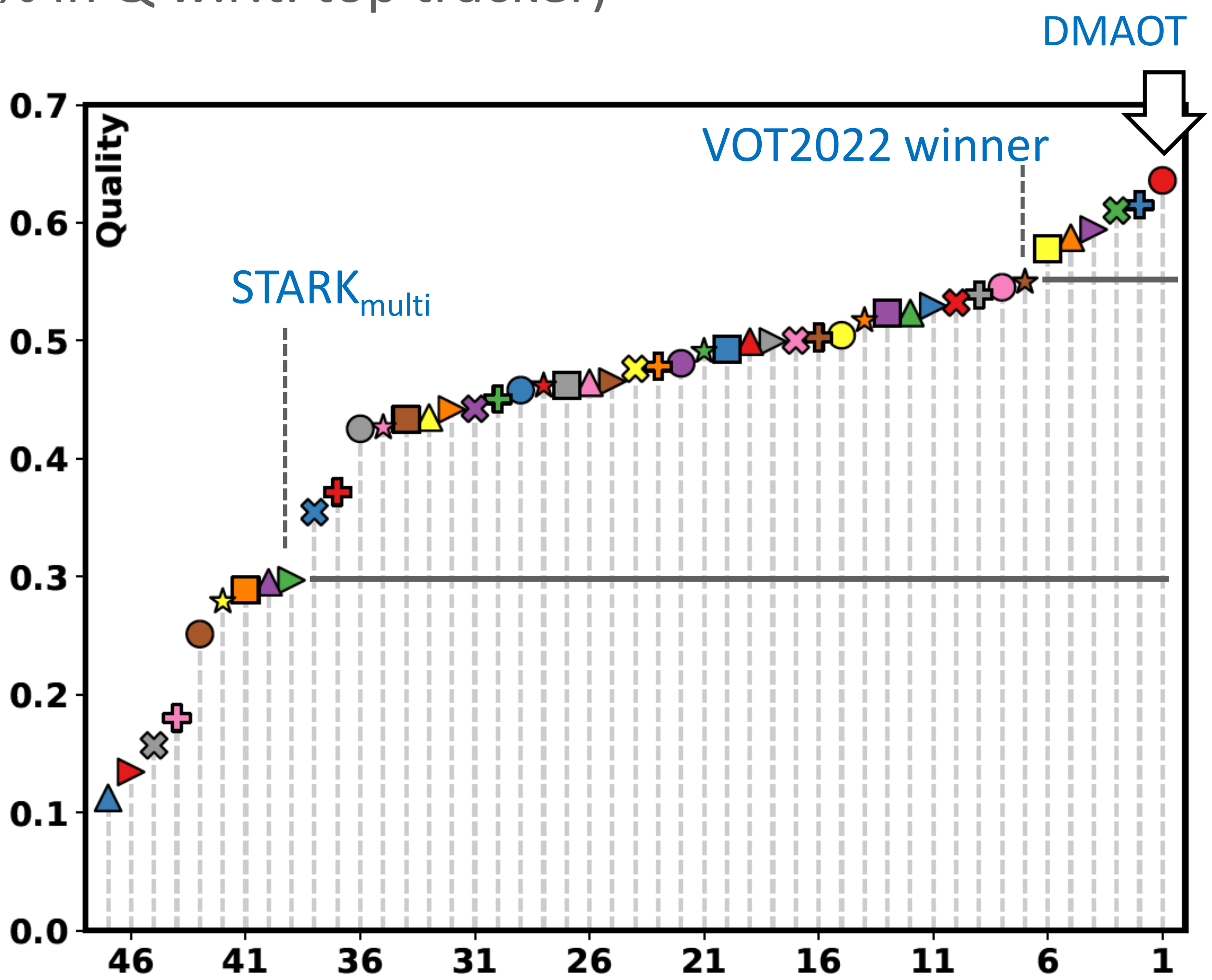
- Dominant design choices:
  - Transformer-based
  - Single-stage ST1/LT0 trackers
  - Same architecture used for frame-to-frame localization and for re-detection



VOTS2023 challenge quality of submissions

- Baseline 1: Independent STARKs[1] (47% in Q w.r.t. top tracker)
  80% of submissions outperform it
- Baseline 2: VOT2022 winner AOT[2]
  13% (top 6 trackers) outperfrom it

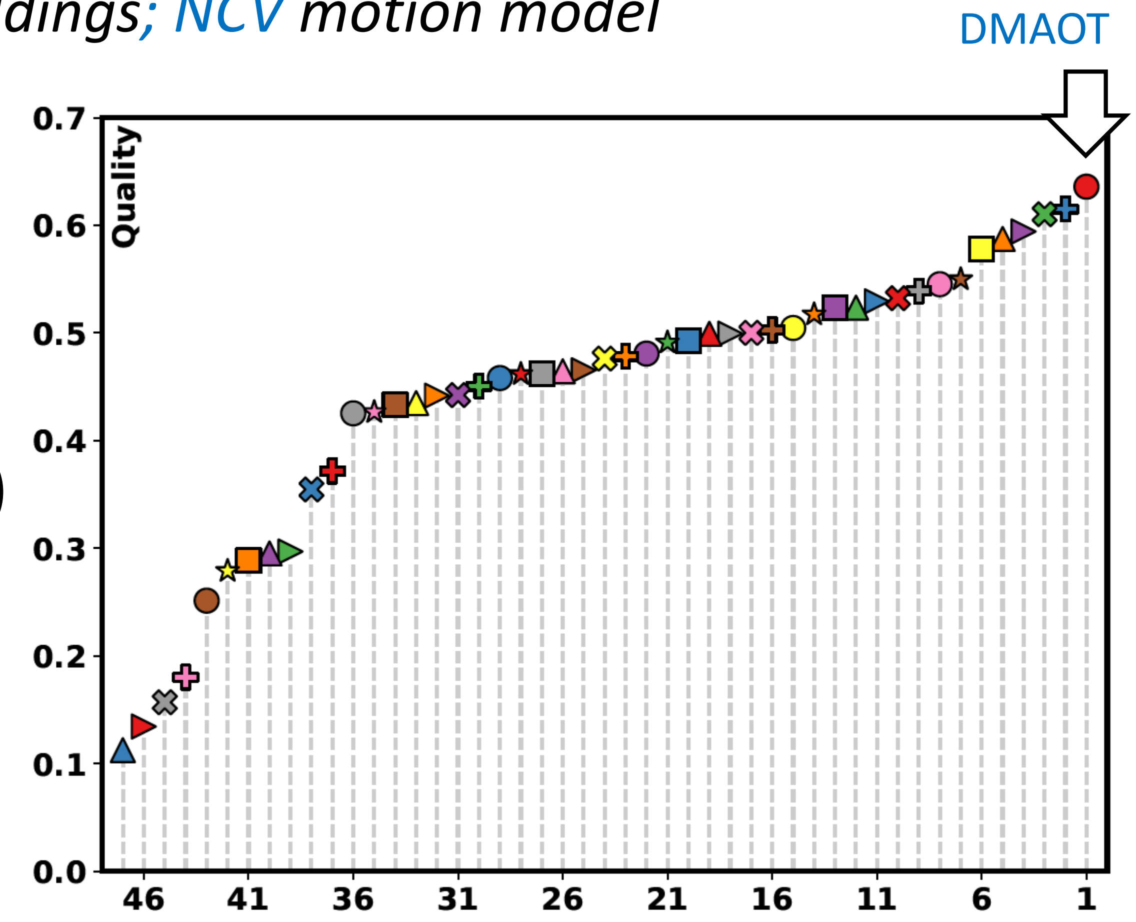




[1] Yan, et al. ICCV2021; [2] Yang, et al. NeurIPS 2021

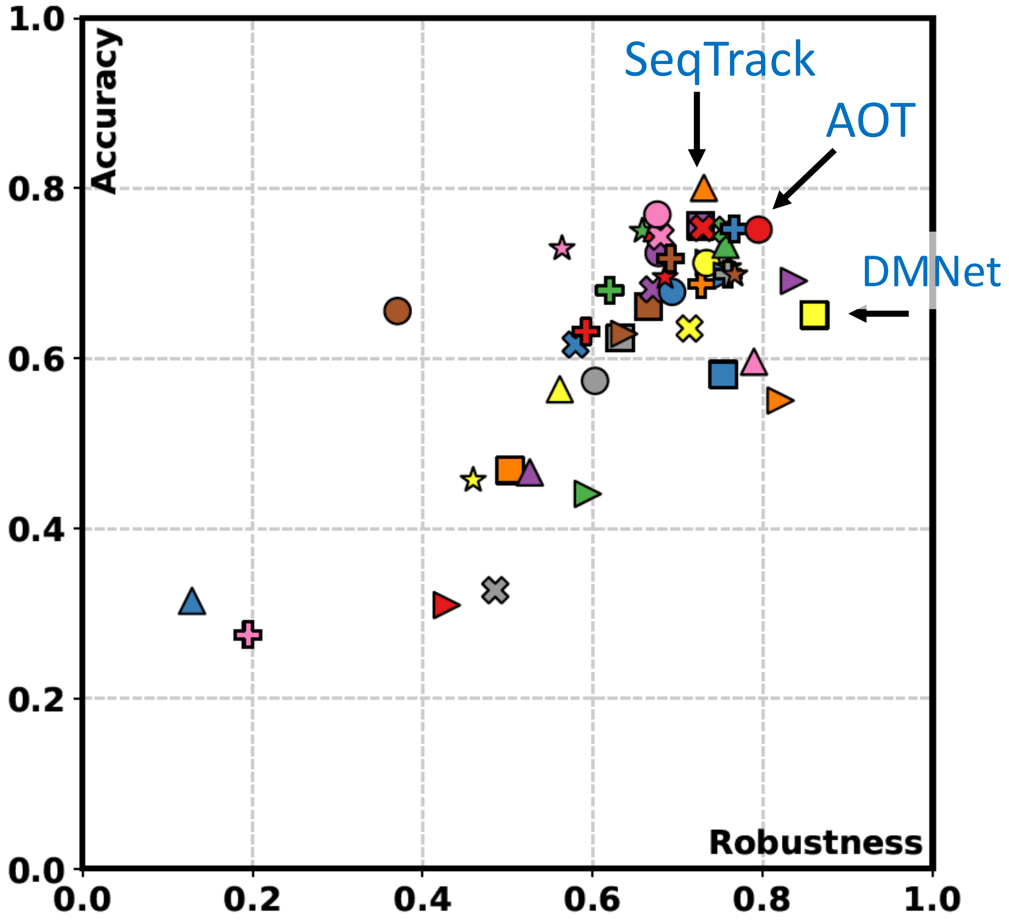
VOTS2023 challenge results

- Top performer DMAOT: Extends the VOT2022 winner AOT
  - *Swin transformer backbone ; Separates long-term and short-term target templates; gated propagation module for visual embeddings; NCV motion model*
- Very good Acc=0.751 & Rob=0.795 (localizes the target 80% of the time)
- Very low drifting (DRE=7%),
- Low false absence prediction (NRE=14%)
- Good target absence prediction: in ADQ=73% cases



VOTS2023 challenge results

- The top-performer in Q (DMAOT) strikes a good balance in Acc/Rob
- Top robustness: DMNet (Rob=0.86) vs (DMAOT Rob = 0.795)
  - Reason might be the use of optimal transport formulation in segmentation/localization
- Top accuracy: SeqTrack
  - Bounding box tracker with SAM[1] segmentation
  - Care taken when to accept the SAM[1] result



[1]Kirillov, et al., Segment Anything, 2023

Summary of tracking performance evaluation

- A number of benchmarks available (VOT, OTB100, GOT10k, LaSOT, TrackingNet)
- Extensive training sets increasingly important (GOT10k, LaSOT, TrackingNet, Trans2k, YoutubeVOS)
- Pretraining and training crucially impacts the performance
- Transformers currently the dominant methodology
- Emergence of pure segmentation-based trackers
- Convergence in tracking (single/multi-target, short/long-term, segmentation)

VOTS results

- Carefully constructed and annotated data sets
- Advanced evaluation protocols
- Advanced and flexible evaluation toolkits

Twitter updates
https://twitter.com/votchallenge