Previously at ACVM

- Long-term tracking:
  - Identify target disappearance
  - Detect the target when it reappears
- Three architectures:
  - TLD (NCC gray-scale patch + flow)
  - ALIEN (Keypoints)
  - FCLT (DCF)
- SoTA deep tracker
  - MBDMD

Advanced CV methods
Performance evaluation for single-target 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 --

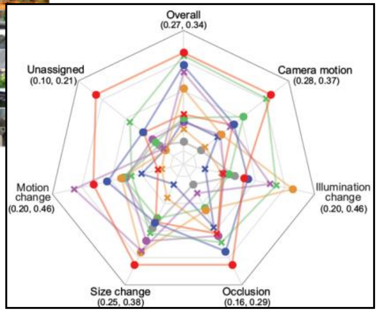
- ~100 tracking papers published annually
- Nonstandard evaluation, source code scarce (before 2013)
- The VOT initiative (February 2013)
- Partners: FRI-UL (SLO), UB (UK), CTU (CZ), AIT (A), LU (S), NICTA (AU), TUT (FI)
- Goal: Establish evaluation standards -> development of trackers
- Problem: Tracking 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 worshop.

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.

Visual Object Tracking Challenge VOT

DATASET (SHORT-TERM TRACKERS)

Related datasets

- 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

- 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

The VOT dataset annotation protocol

- Each image annotated by 6 attributes:

  Occlusion, Illumination change , Object motion, Object size change, Camera motion, Unassigned

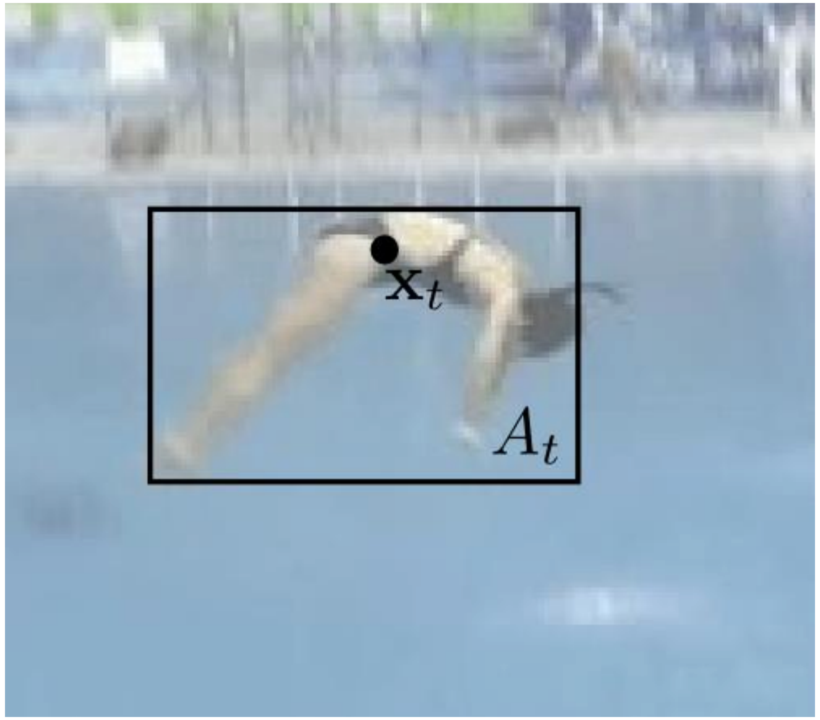
(i)	0	1	1	0
(ii)	0	0	0	0
(iii)	0	0	0	0
(iv)	1	1	1	0
(v)	0	0	0	0
(vi)	0	0	0	1

Target ground truth position annotation

- Comparing tracking result against a ground-truth
  - Sequence manually annotated by an expert annotator

- Different kinds of annotations historically used
  - Object center point
  - Bounding box (more informative)

$$\Lambda = \{(A_t, \mathbf{x}_t)\}_{t=1}^{N}$$

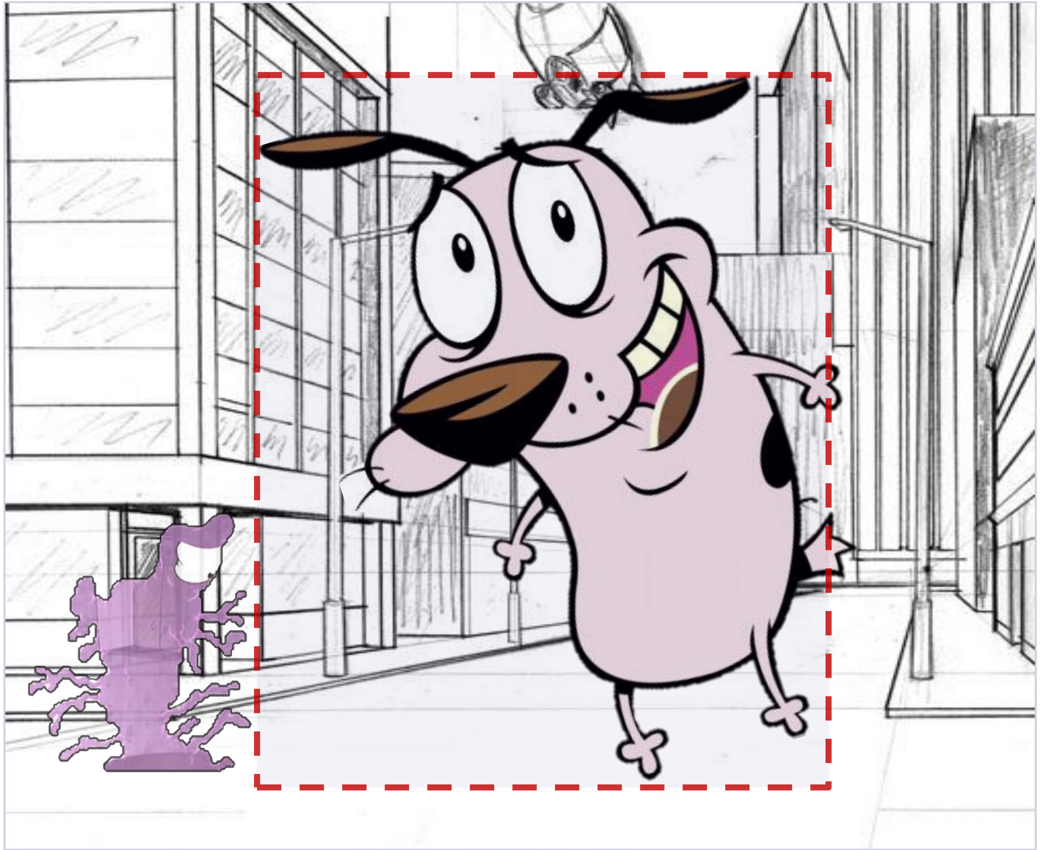


The VOT (2016) dataset annotation protocol

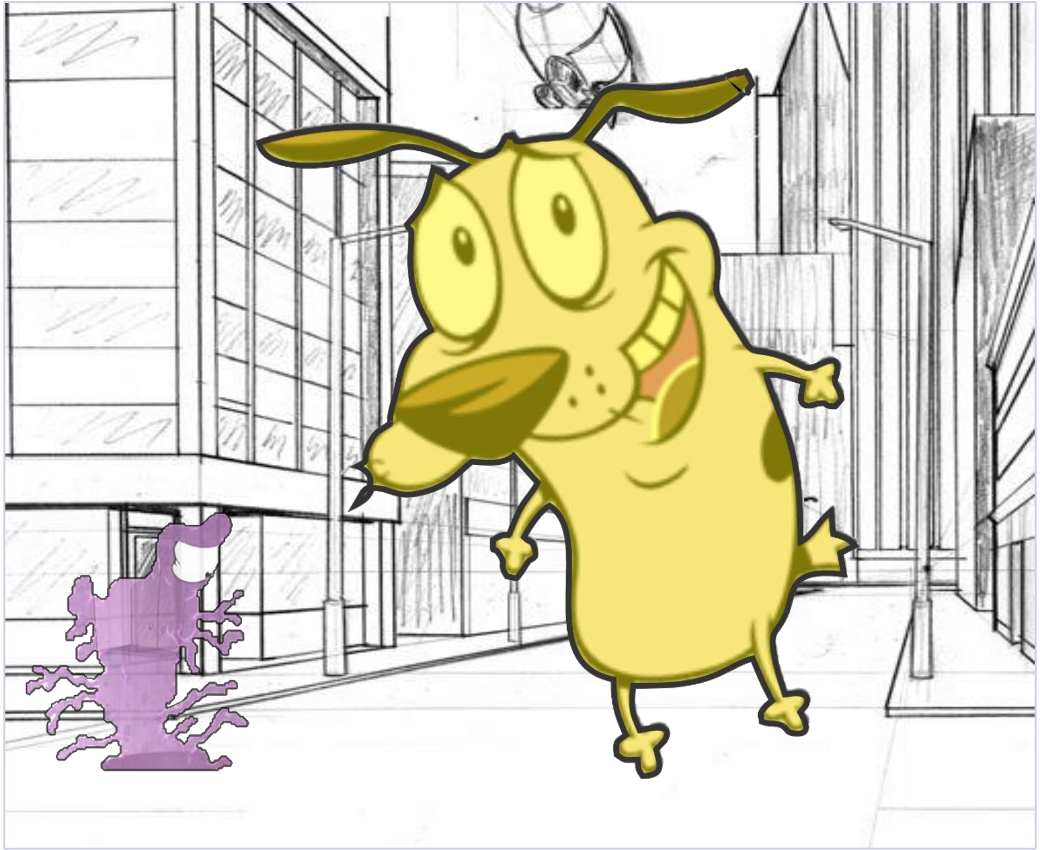
- Each image semi-automatically segmented
- A bounding box fitted automatically to segmentation mask

VOT2020 Paradigm shift – revisiting target pose

Bounding box == pose approximation



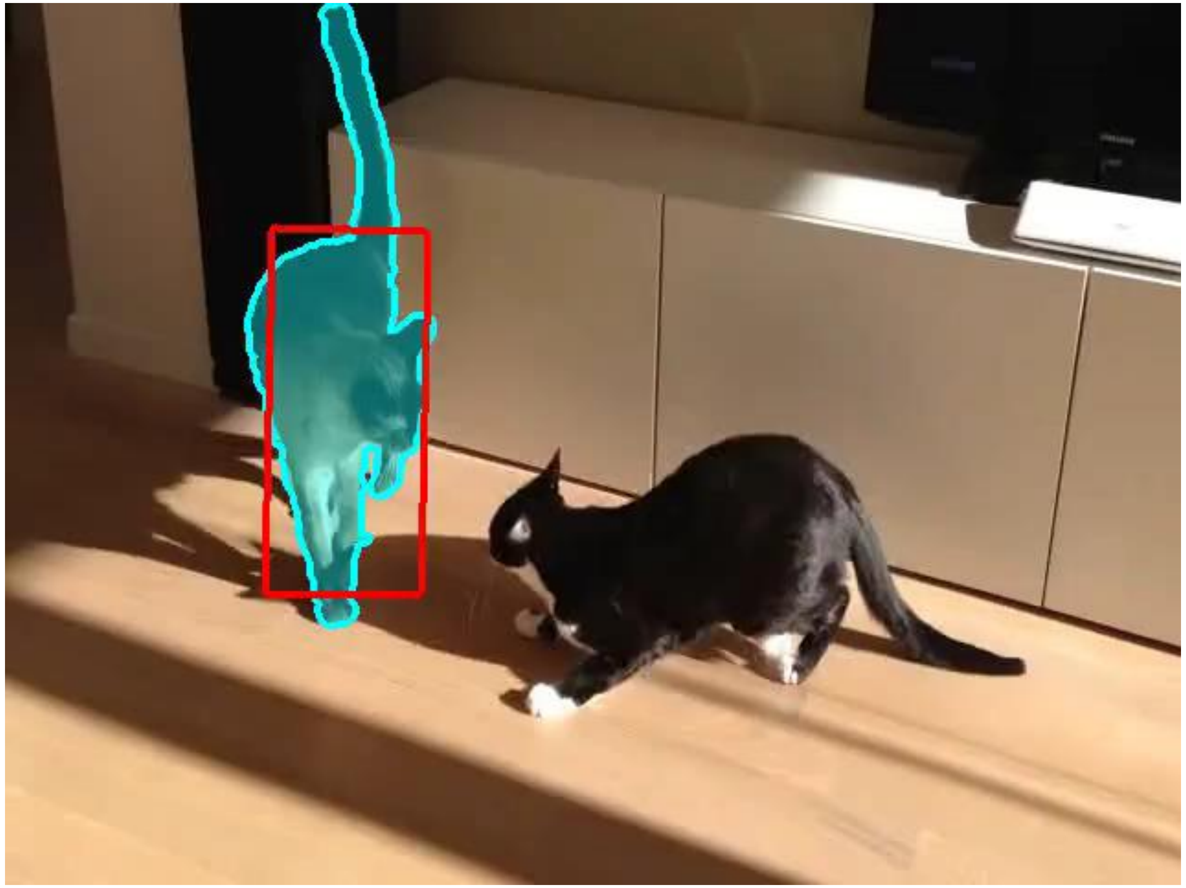
Most accurate pose == segmentation



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

The VOT-ST2020 (onward) dataset

- Public dataset (60 sequences) +
  Sequestered dataset (60 sequences)
  Winner identified on *sequestered dataset*
- Both datasets refreshed
  - A challenging sequence added to each
- All frames manually segmented!
- Bounding boxes not provided (obsolete)
  - Reintroduced in 2022 ☺
- Each frame annotated by 6 attributes:
  Occlusion, Illumination change , Object motion, Object size change, Camera motion, Unassigned



Red – VOT2019 annotation by a bounding box
Blue – VOT2020 annotation by a segmentation mask

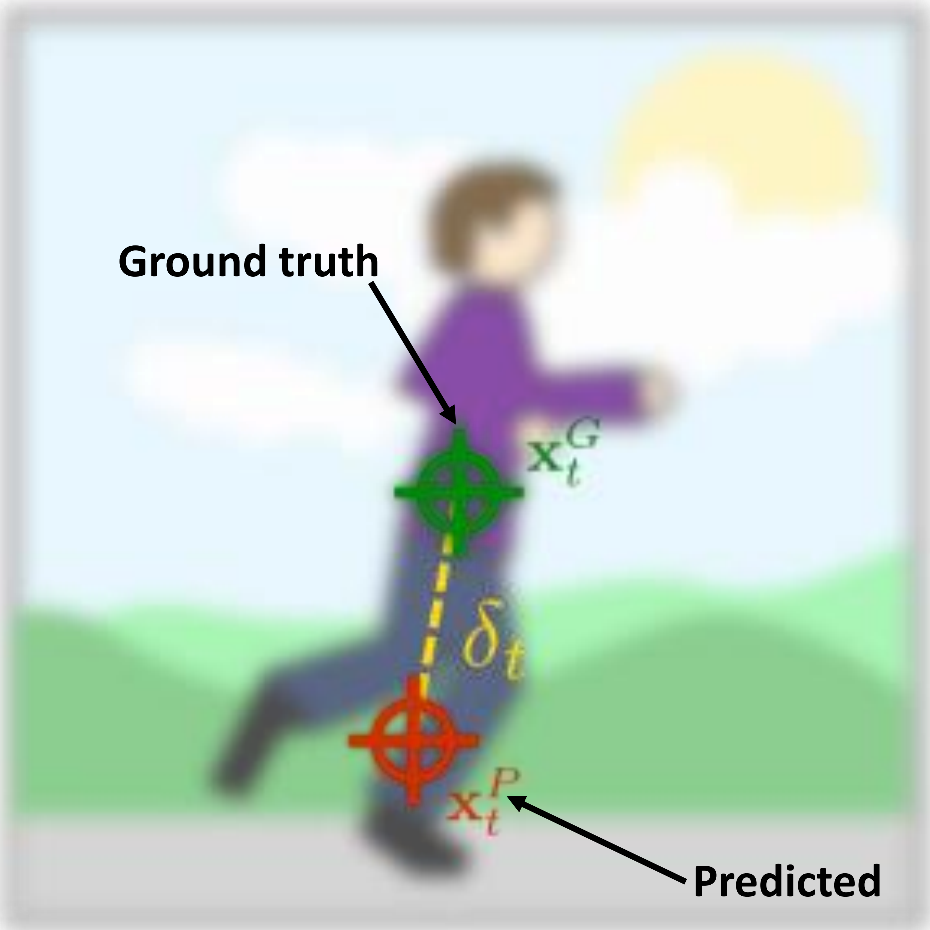
Visual Object Tracking Challenge VOT

EVALUATION METHODOLOGY (SHORT-TERM 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: Center error

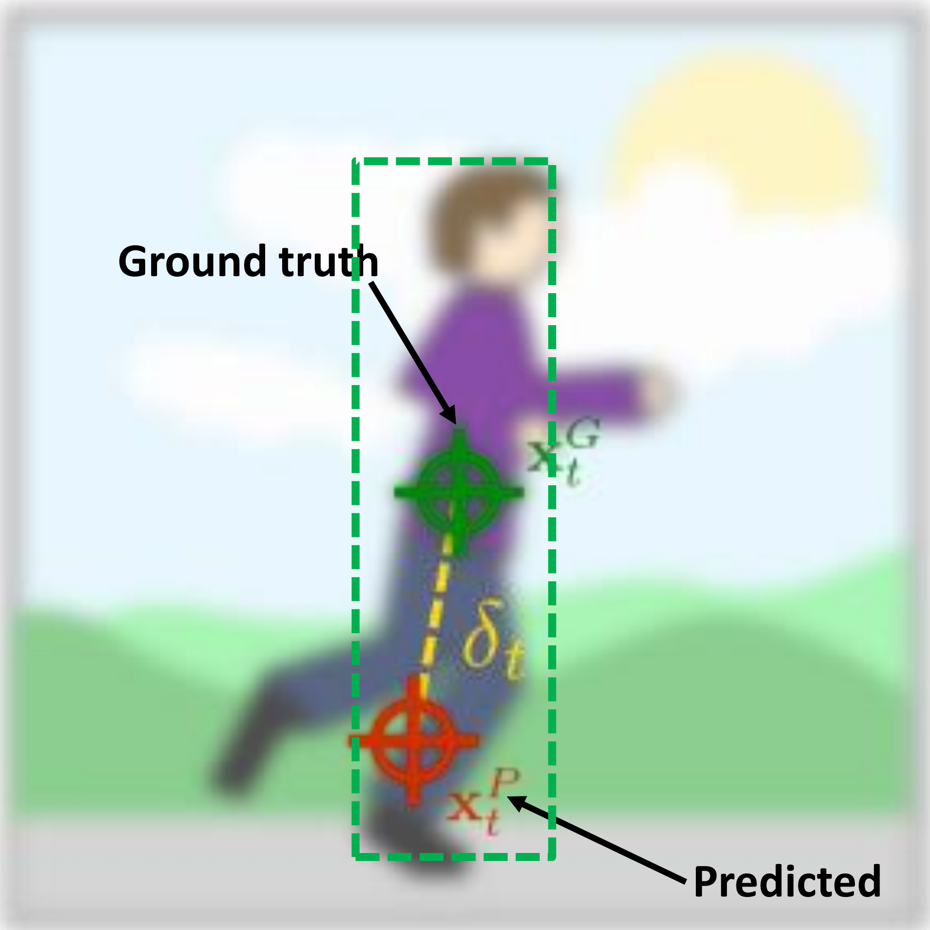
- Distance between center position of ground truth and position predicted by the tracker

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

- Take into account the size as well by normalizing with the size of the GT bounding box (A_t^G):

$$\widehat{\Delta}(\Lambda^G, \Lambda^P) = \left\{\widehat{\delta}_t\right\}_{t=1}^{N}, \quad \widehat{\delta}_t = \|\frac{\mathbf{x}_t^G - \mathbf{x}_t^P}{size(A_t^G)}\|$$

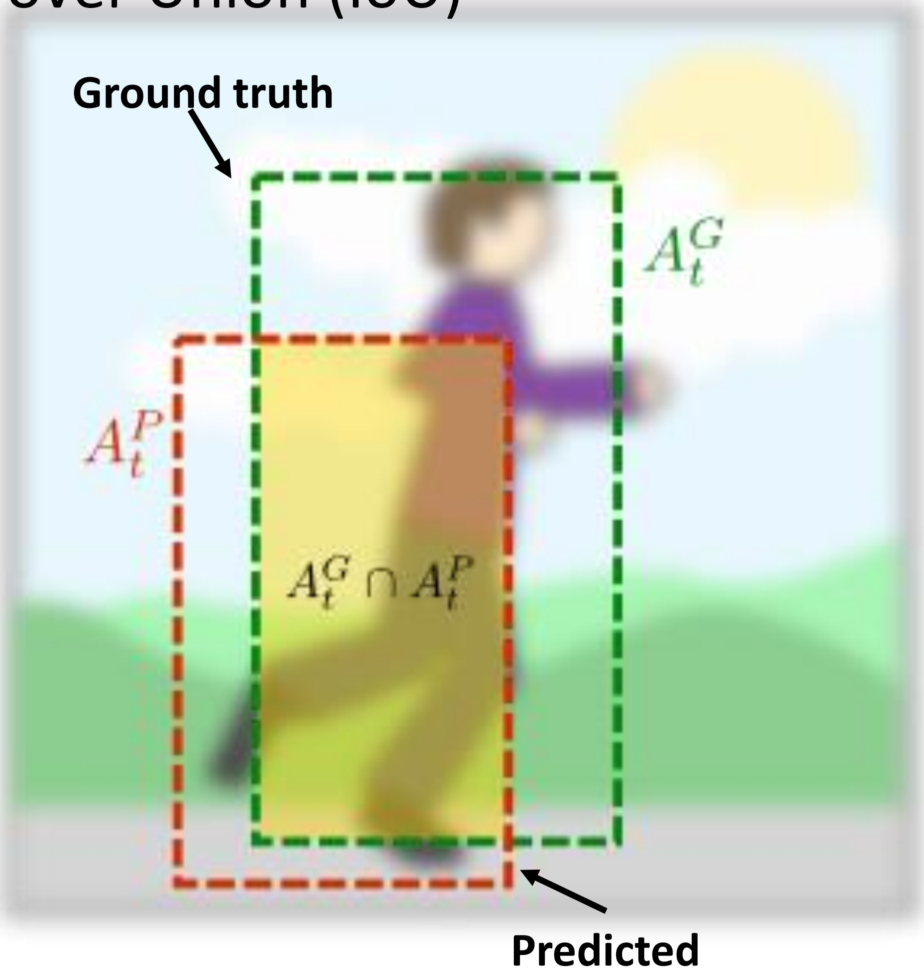
- Drawback: the error is unbounded, and does not take into account the estimated size of the target



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)

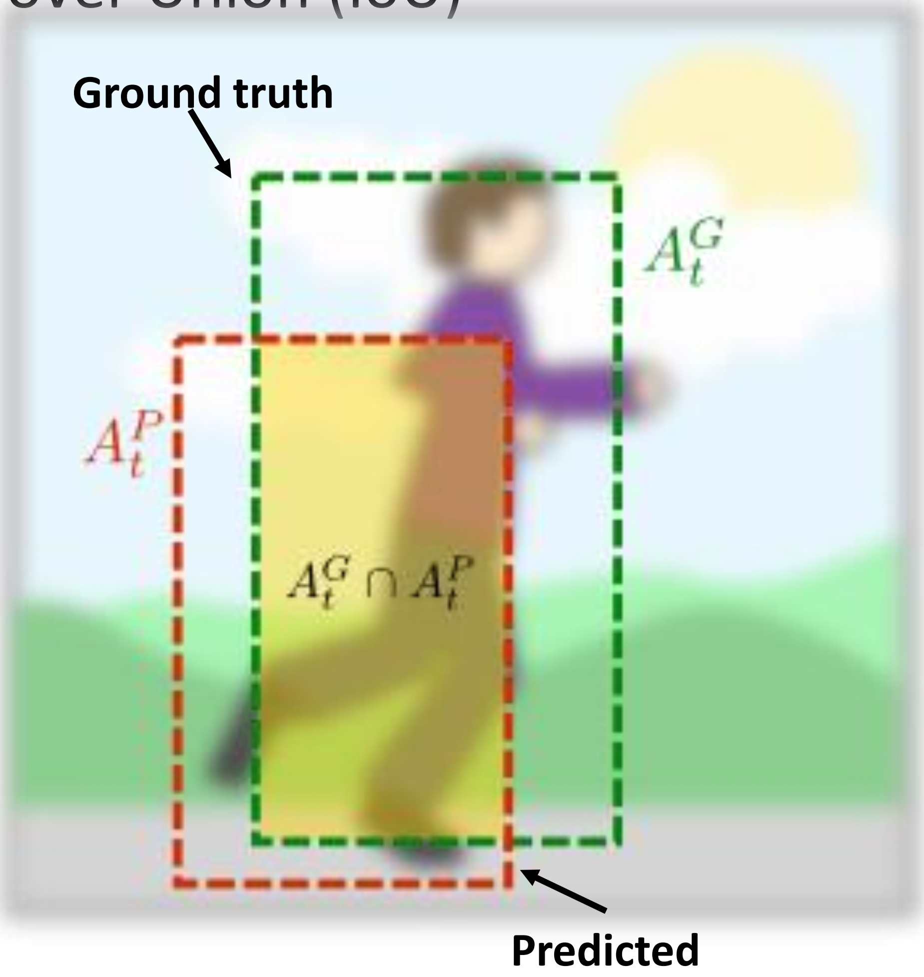
$$\Phi(\Lambda_G, \Lambda_P) = \left\{ \frac{A_t^G \cap A_t^P}{A_t^G \cup A_t^P} \right\}_{t=1}^{N}$$

- 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., $\Phi_t > 0.5$.



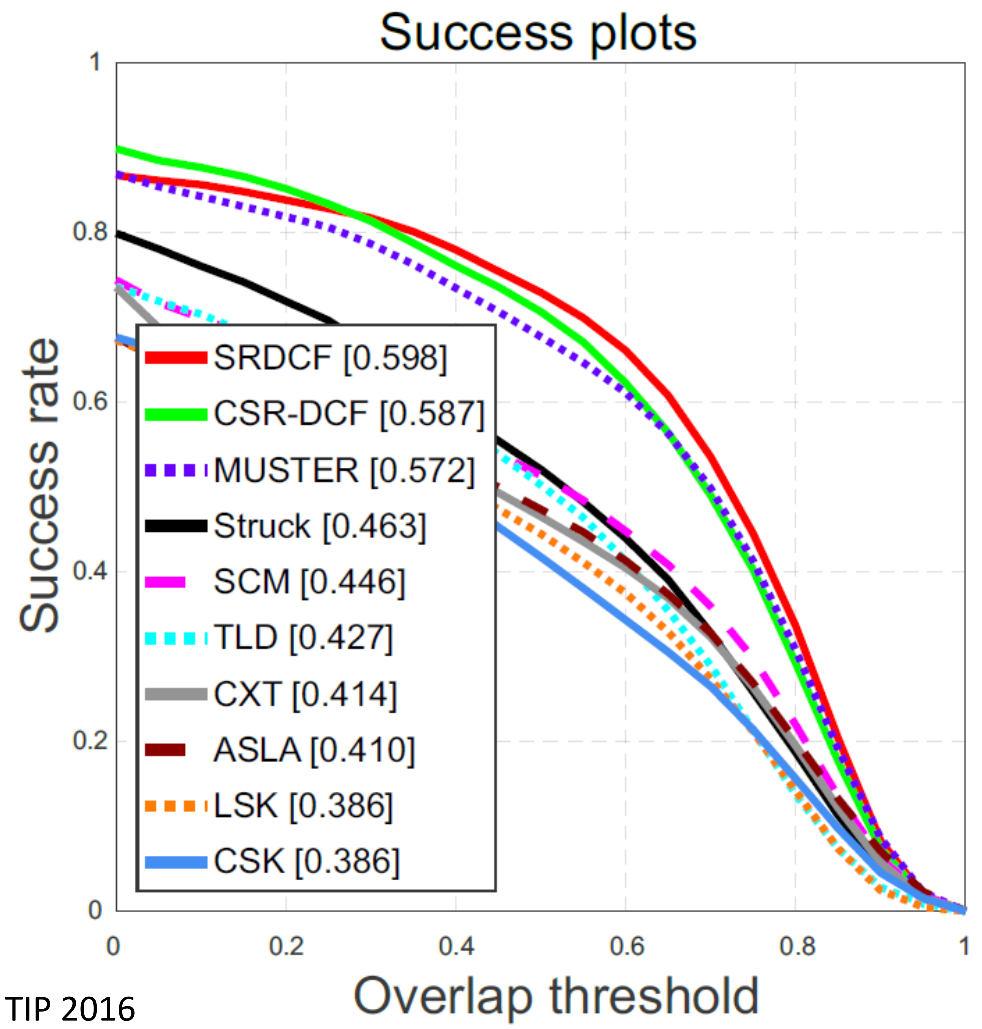
Measure types: Success plot

- A 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}

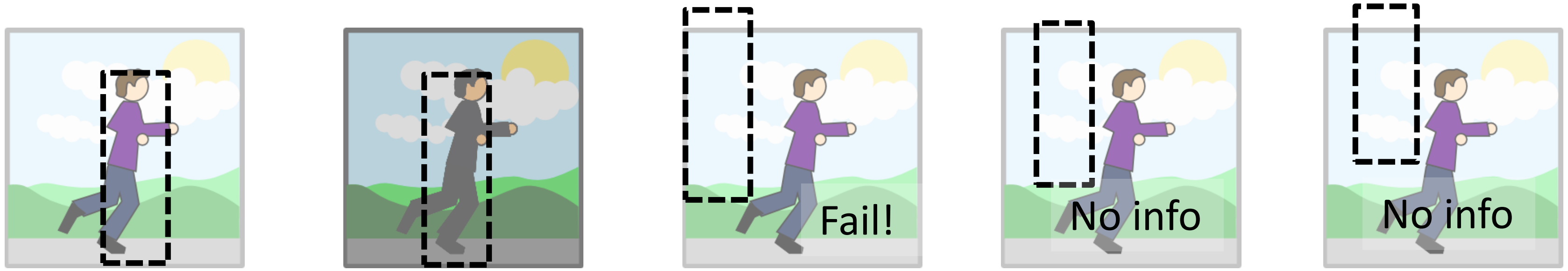
- The measure: Area under the curve *AUC*
  (shown[2] to be equal to average overlap)



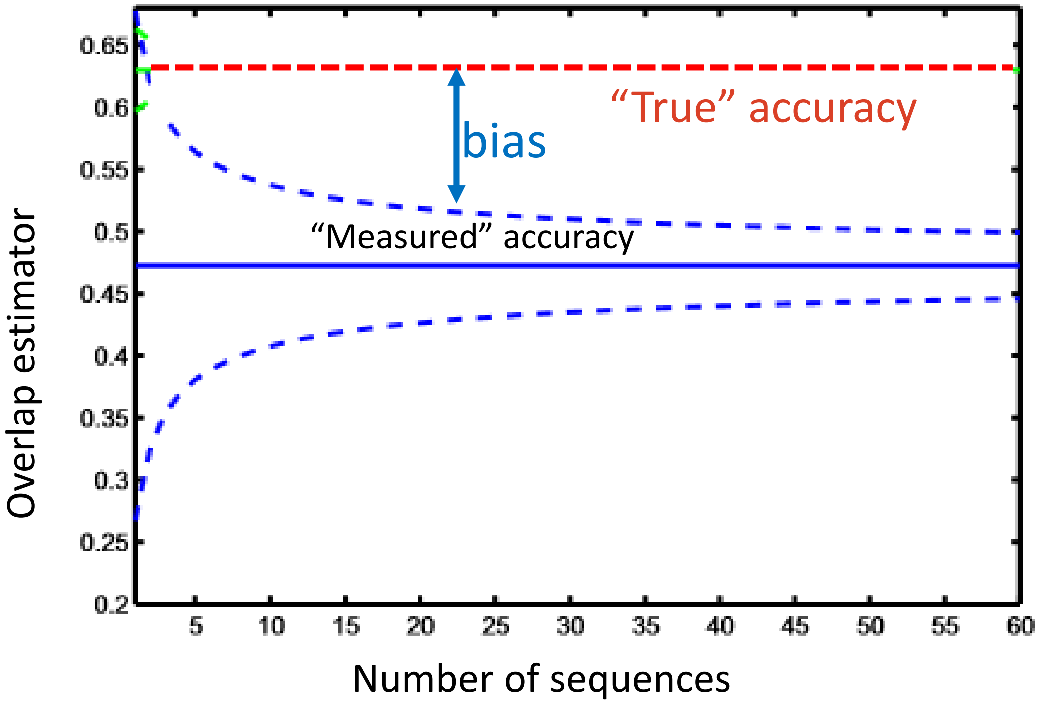
[1]Wu et al. Online Object Tracking: A Benchmark, CVPR 2013
[2]Čehovin Zajc, Leonardis, and Kristan, Visual object tracking performance measures revisited, IEEE TIP 2016

Measure types: Success plot

- But the tracker may fail at a random position



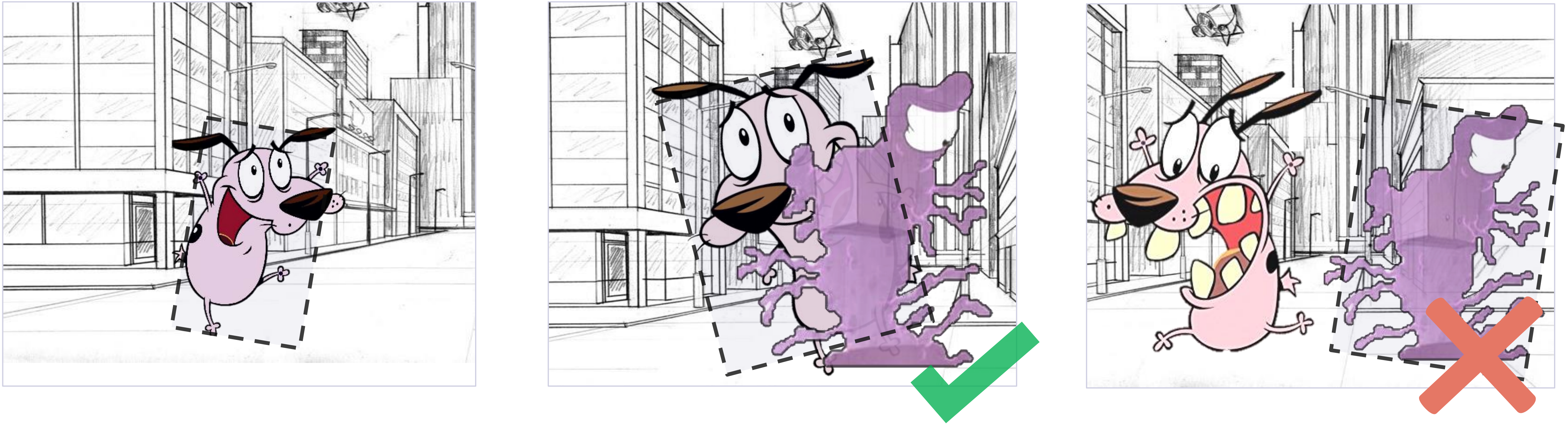
- The overlap drops to 0 after the failure
- Benefits: Simple experiment
- Drawback: Affected by point of failure and sequence length



Measure types: Failure rate

So, which measure should we use??

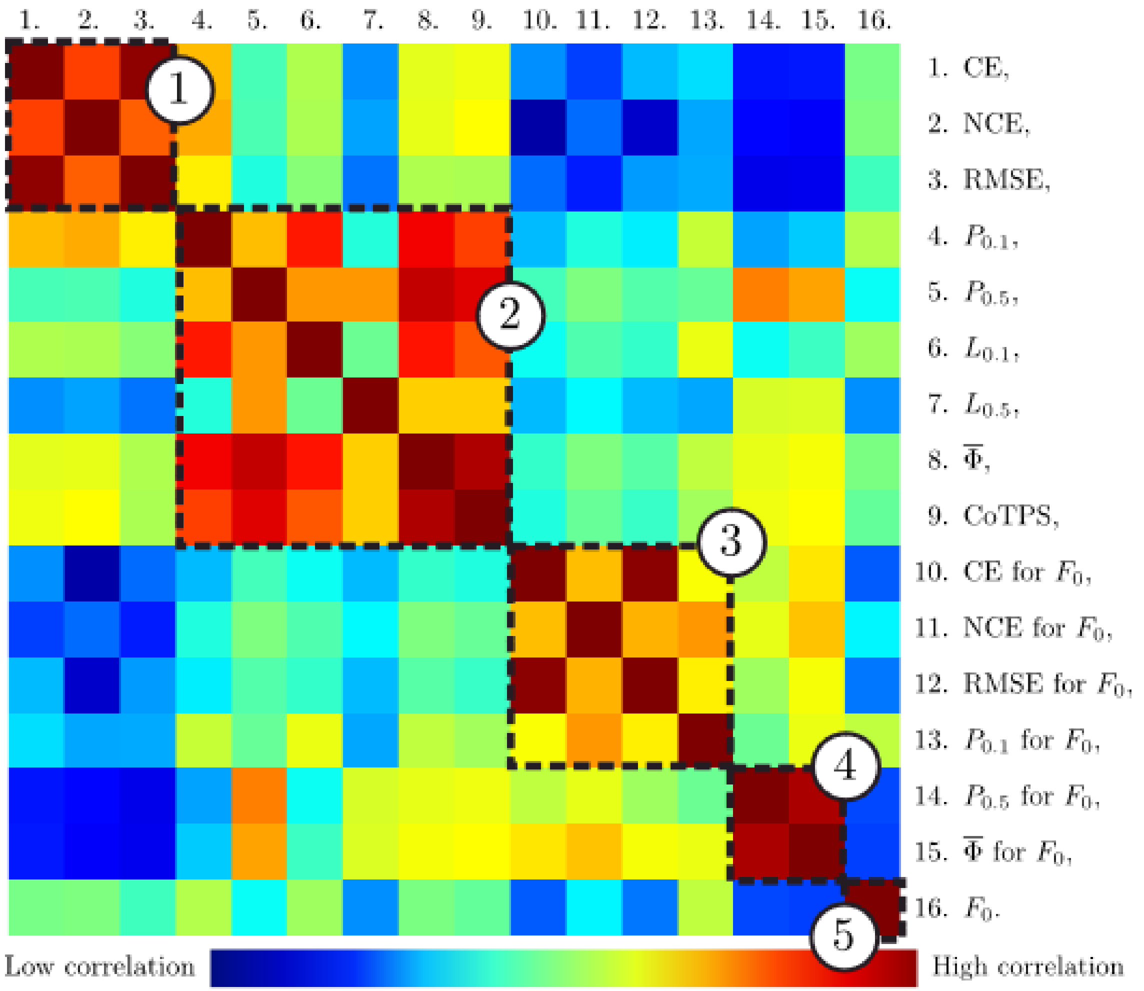
- In a *short-term setup*, a drift is considered a failure and tracker is reset



- Counts the number of times the tracker failed and had to be reinitialized
- Benefits: Entire sequence is used for evaluation
- Drawback: Requires interactive experiment

The VOT (2013) performance measure selection

- Run 13 trackers on 25 sequences
- Tested the equivalence between measures by calculating correlations among all measure pairs
- Several correlated clusters of measures automatically detected by running Affinity Propagation

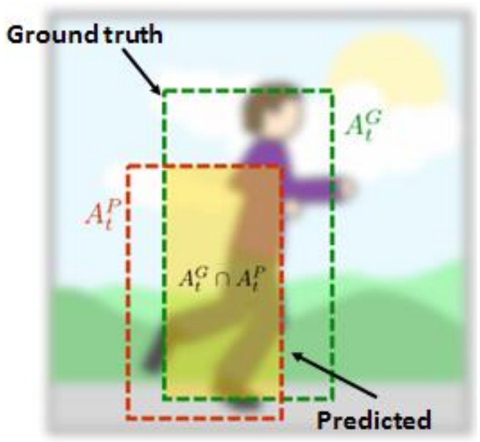


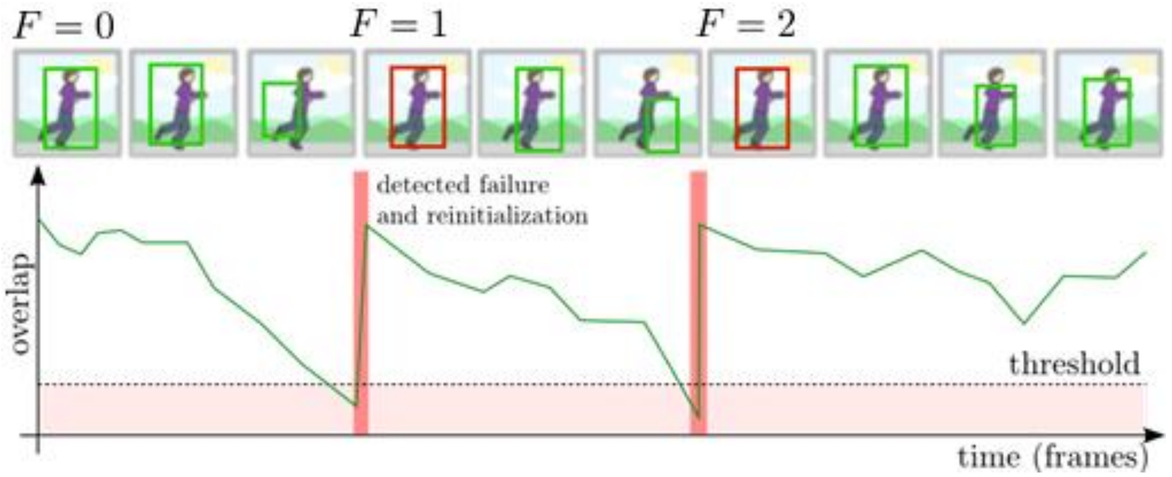
Čehovin, Leonardis, Kristan. *Visual object tracking performance measures revisited*, IEEETIP 2016

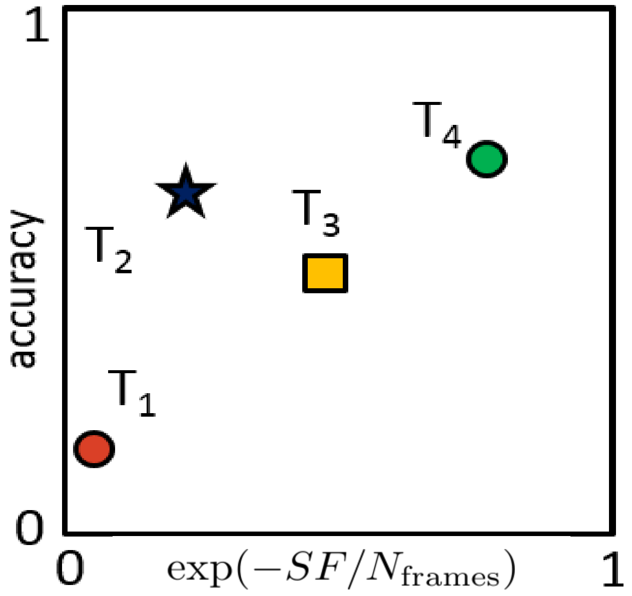
Evaluation methodology

- Two weakly correlated measures[2] chosen according to[1]:
  - Robustness (number of times a is reinitialized)
  - Accuracy (average overlap while tracking)
- Expected average overlap EAO: principally combines A & R

  expected overlap the tracker obtains on a short-term sequence of an average length





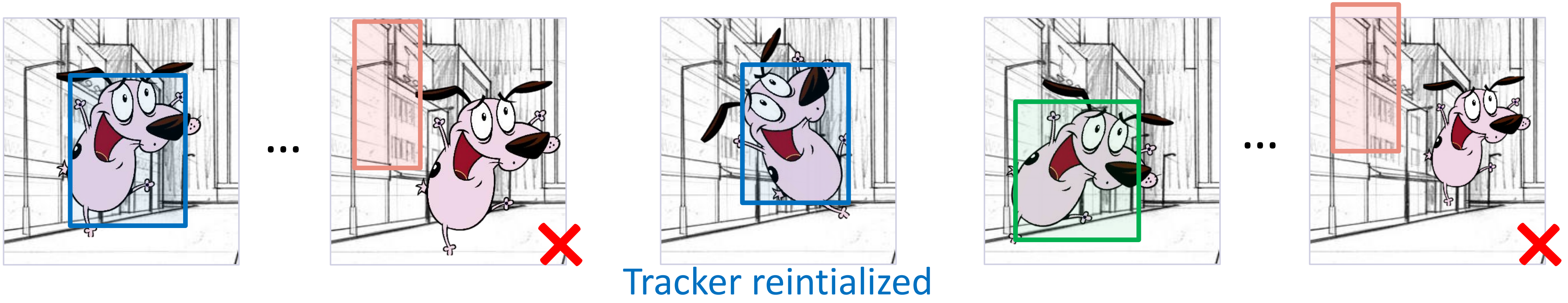


[1]Čehovin, Leonardis, Kristan. *Visual object tracking performance measures revisited*, IEEETIP 2016

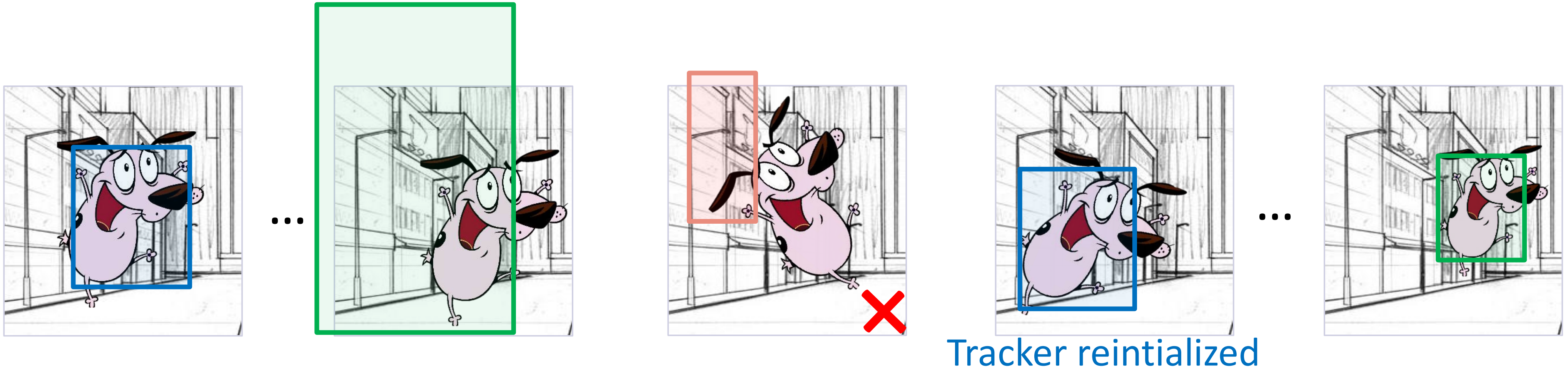
[2]Kristan et al., A Novel Performance Evaluation Methodology for Single-Target Trackers, IEEETPAMI 2016

… but trackers were getting better

- A failure at some frame affects the next failure (a tuning opportunity)

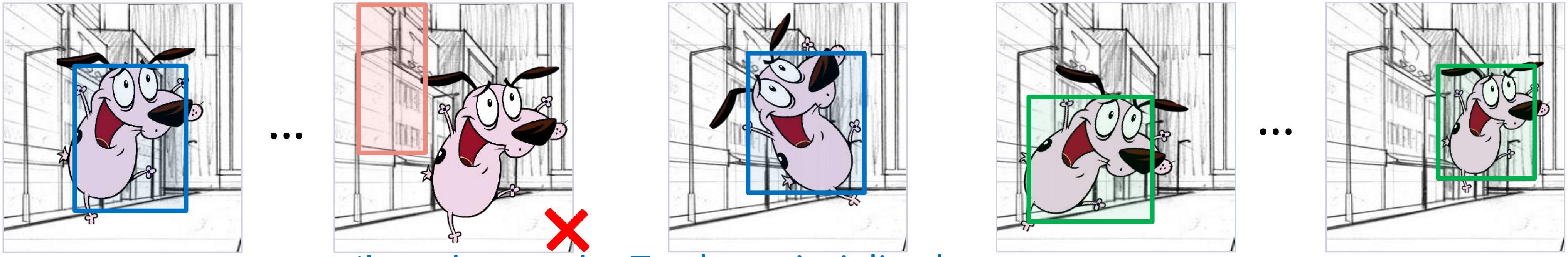


- Intentional bounding box over-inflation



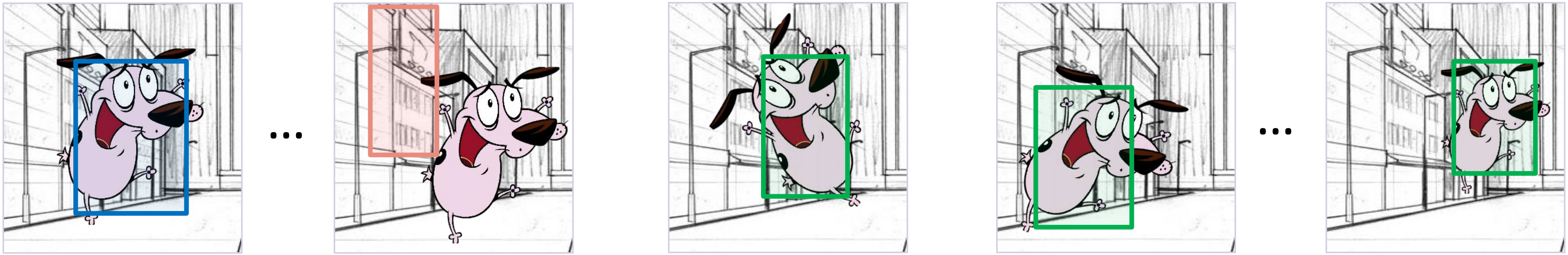
… but trackers were getting better

- Failure definition (0 overlap) penalizes even short-term failures



Failure detected Tracker reintialized

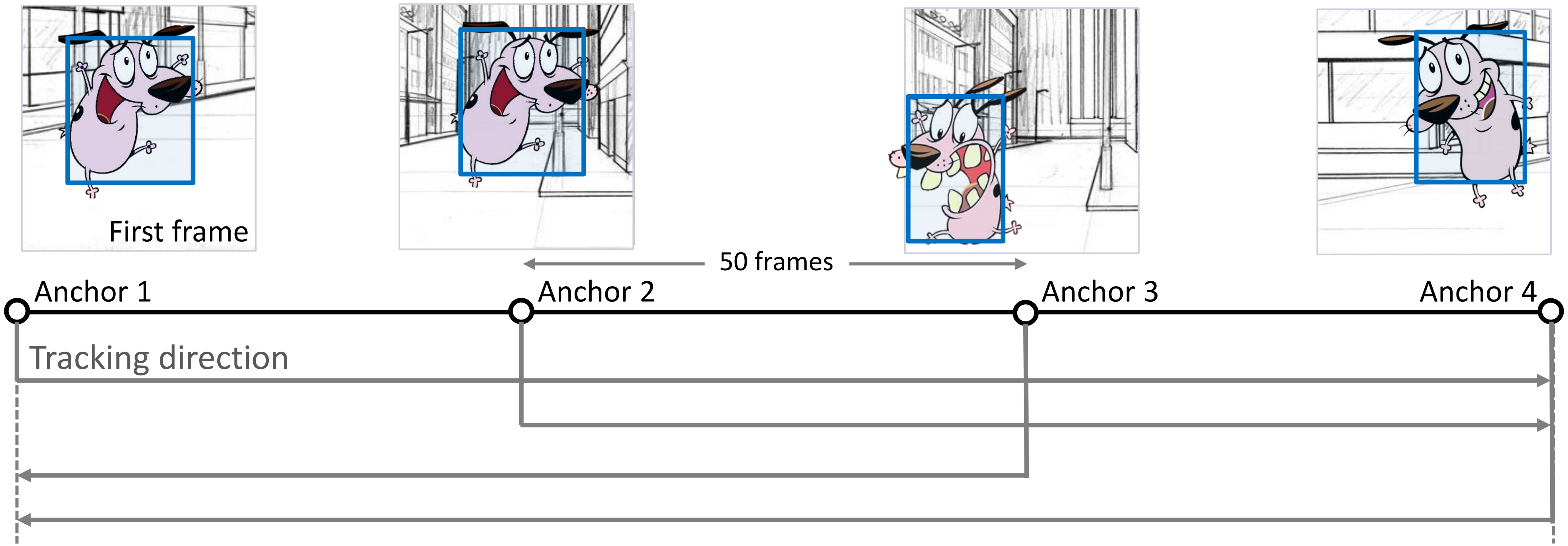
- A tracker might have recovered from a *short-term* failure



If not reset

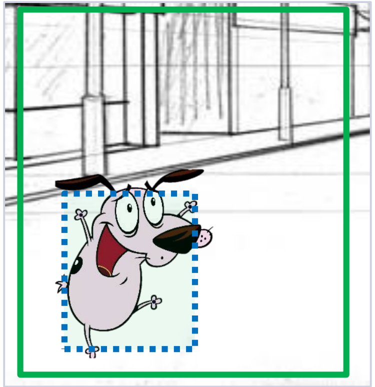
VOT2020 Anchor-based protocol

- Introduce initialization points (anchors) equal for all trackers
- Track in the direction of the largest number of tracking frames
- Each anchor produces one subsequence



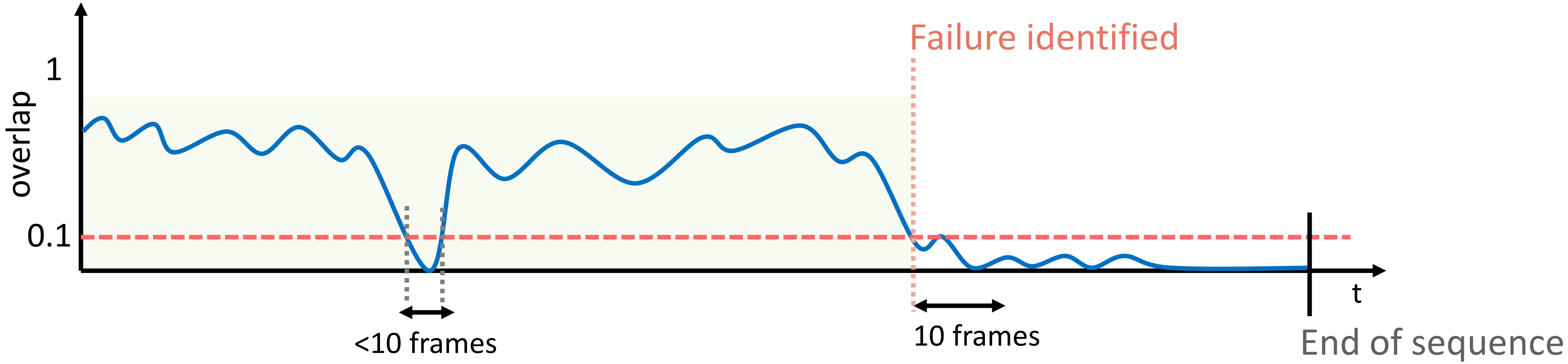
Accounting for short-term failure recovery

- Potential failure: overlap < $\theta_\Phi = 0.1$



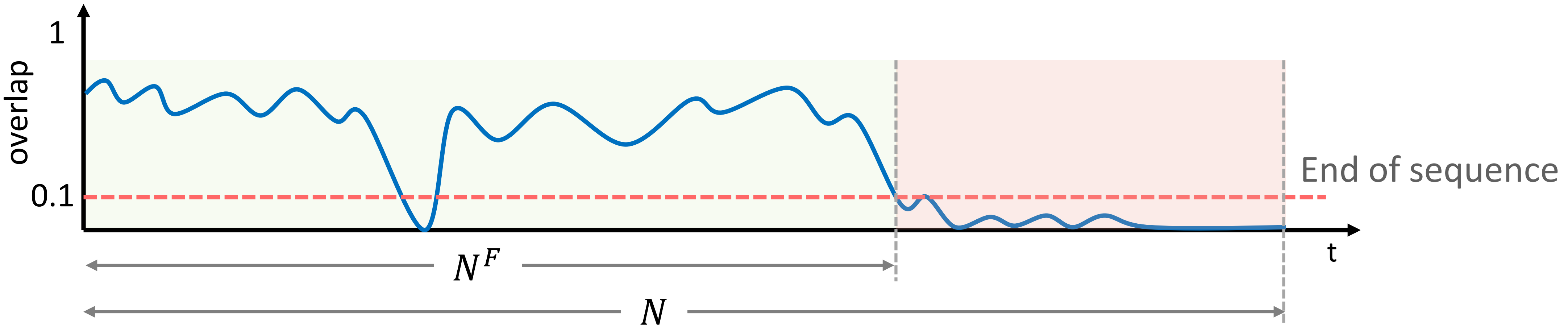
Prevent *“gaming”* where a tracker would predict the *“entire image”* as a bounding box to prevent reset identification

- Failure if the tracker does not recover within θ_N=10 frames



VOT performance measures (since 2022)

- Accuracy (A): average overlap on the successfully tracked period
- Robustness (R): Percentage of the tracked sub-sequence (N^F/N)



- Overall A/R: weighted average over all sequences
- EAO measure – combines the per-subsequence results

Kristan et al., [The new VOT2020 short-term tracking performance evaluation protocol and measures](), VOT 2020 webpage

Visual Object Tracking Challenge VOT

DATASETS & PERFORMANCE MEASURES

(LONG-TERM TRACKERS)

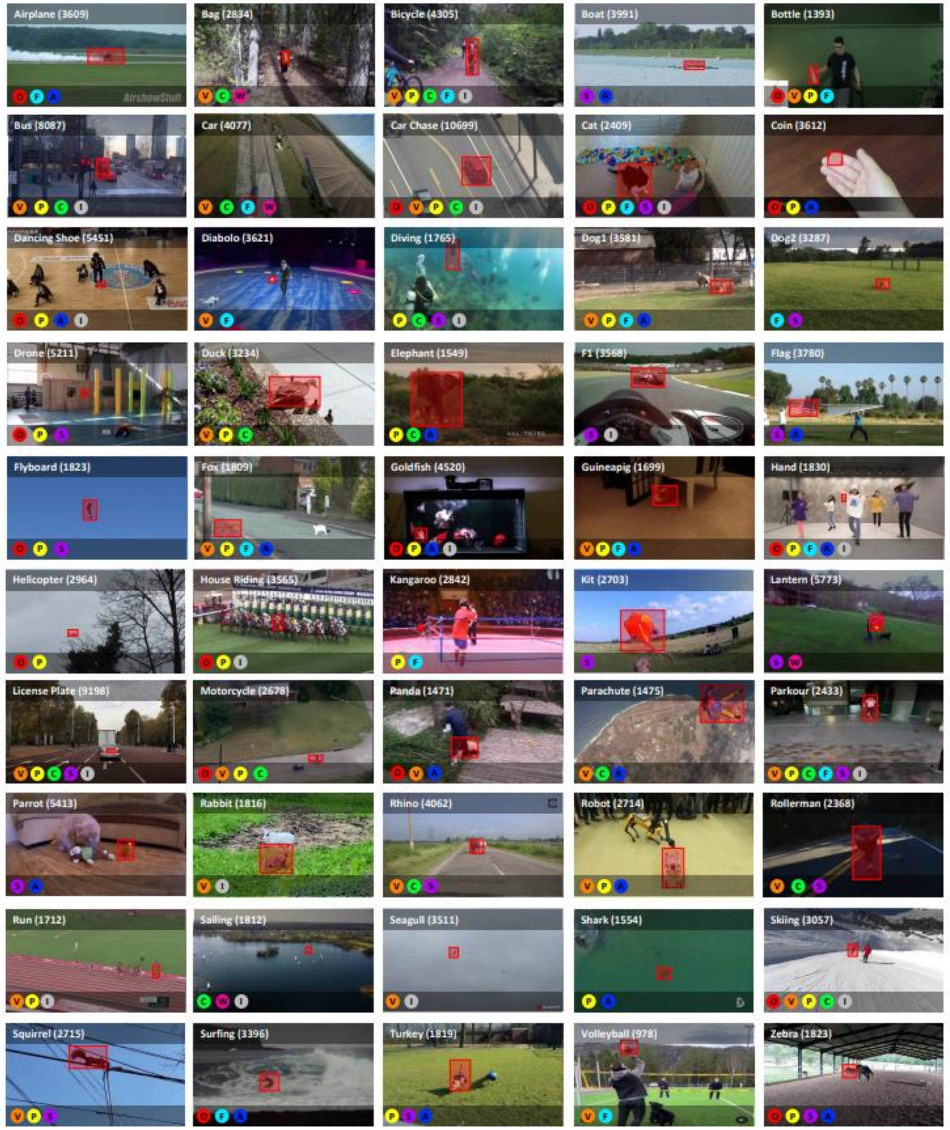
Long-term tracking evaluation



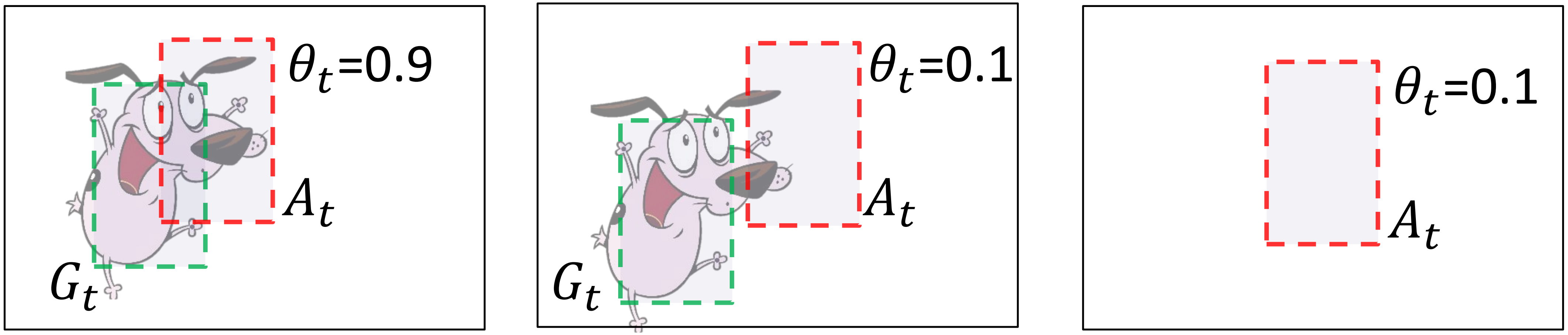
- Required long-term tracker properties:
  - Determine whether the target has been lost (or disappeared)
  - Re-detect the target when it reappears
- Tracker output at each frame: bounding box + certainty score

VOT2022 LT tracking dataset

- 50 sequences (168,282 frames)
  (average sequence length >4k frames)
- Axis-aligned bounding box annotations
  (persons, car, motorcycle, bicycle, boat, animals, etc.)
- Resolution: 1280x720
- Average per sequence disappearance: 10
- Average target absence period: 52 frames
- Nine per-sequence attributes:
  (1) full occlusion, (2) out-of-view motion, (3) partial occlusion, (4) camera motion, (5) fast motion, (6) scale change, (7) aspect ratio change, (8) viewpoint change, (9) similar objects



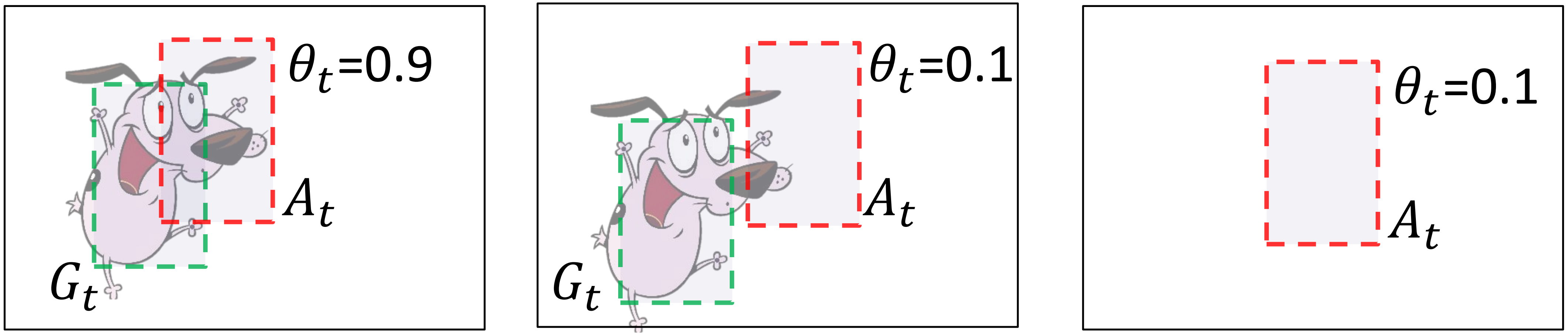
LT performance measure design



- Requirements: (i) localization accuracy, (ii) target absence prediction accuracy, (iii) re-detection accuracy
- Precision (Pr) … % of all predictions A_t that agree with GT G_t
- Recall (Re) … % of all GT boxes that that agree with predictions A_t
- F-measure … a standard Pr/Re tradeoff $F = 2PrRe/(Pr + Re)$

[1]Lukežič, Čehovin Zajc, Vojíř, Matas, Kristan, Performance evaluation methodology for long-term single-object tracking, TCyb2020

LT performance measure design



- Agreement = sufficient overlap:

Detection "uncertainty" threshold

$$\Omega(A_t, G_t) \geq \tau_\Omega \longrightarrow \Omega(A_t(\tau_\theta), G_t) \geq \tau_\Omega$$

- Precision and Recall depend on two thresholds: $\mathrm{Pr}(\tau_\theta, \tau_\Omega)$, $\mathrm{Re}(\tau_\theta, \tau_\Omega)$

- The overlap threshold is avoided by integrating it out

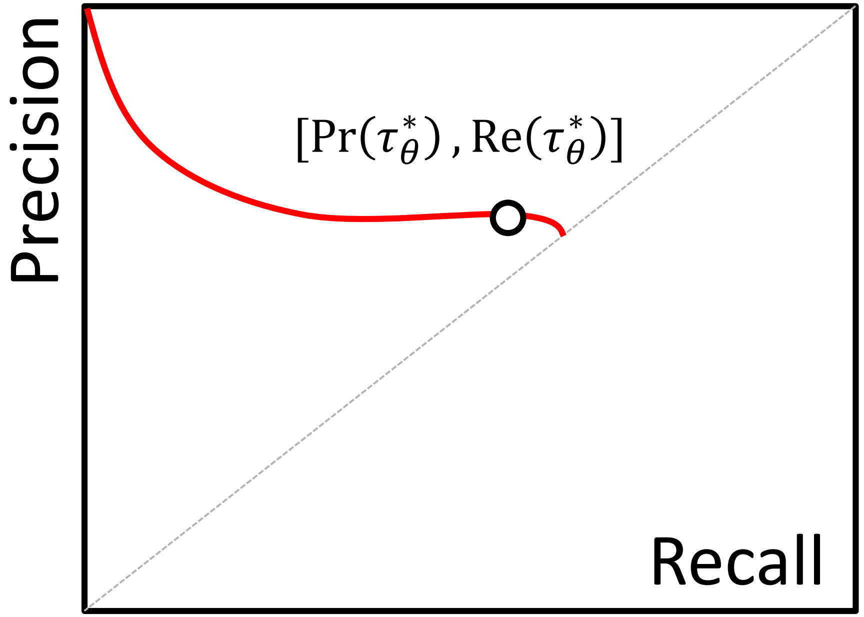
$$Pr(\tau_\theta) = \int_0^1 Pr(\tau_\theta, \tau_\Omega) d\tau_\Omega = \frac{1}{N_p} \sum_{t \in \{t : A_t(\tau_\theta) \neq \emptyset\}} \Omega(A_t(\tau_\theta), G_t),$$

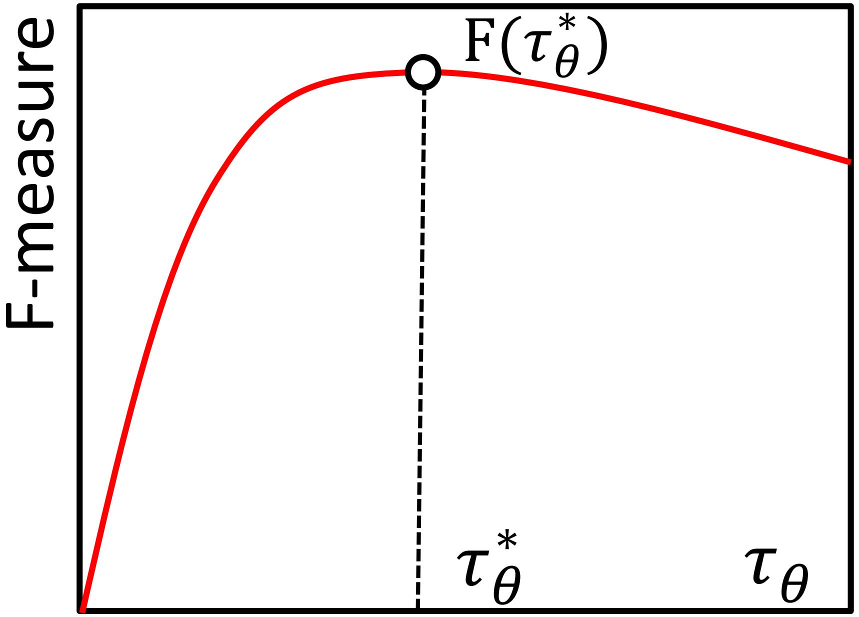
$$Re(\tau_\theta) = \int_0^1 Re(\tau_\theta, \tau_\Omega) d\tau_\Omega = \frac{1}{N_g} \sum_{t \in \{t : G_t \neq \emptyset\}} \Omega(A_t(\tau_\theta), G_t)$$

[1]Lukežič, Čehovin Zajc, Vojíř, Matas, Kristan, Performance evaluation methodology for long-term single-object tracking, IEEE TCyb 2020

Primary LT performance measures

- Primary measures are $\mathrm{Pr}(\tau_\theta^*)$, $\mathrm{Re}(\tau_\theta^*)$ and $\mathrm{F}(\tau_\theta^*)$ evaluated at detection certainty threshold that maximizes the tracker F-measure





- Primary scores thus fully avoid manually setting the thresholds
- In short-term setup, $\mathrm{F}(\tau_\theta^*)$ reduces to a standard ST measure!

Visual Object Tracking Challenge VOT

EVALUATION SYSTEM

The VOT evaluation system

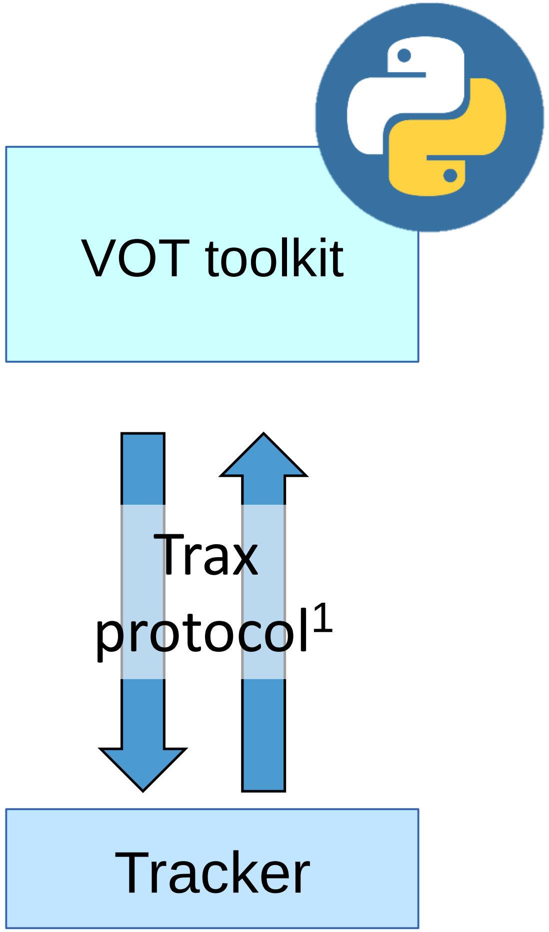
- A toolkit automatically performs a battery of standard experiments

  Currently the most advanced toolkit in visual tracking. Early Matlab toolkits[1] now obsolete, the most recent toolkit in Python.

- Download from the VOT homepage

  https://www.votchallenge.net/howto/tutorial_python.html

- Plug and play!
  - Supports major programming languages and operating systems



[1]Luka Čehovin, TraX: The visual Tracking eXchange Protocol and Library, Neurocomputing, 2017

Short-Long-term tracking

OTHER POPULAR BENCHMARKS & THE ROLE OF TRAINING

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

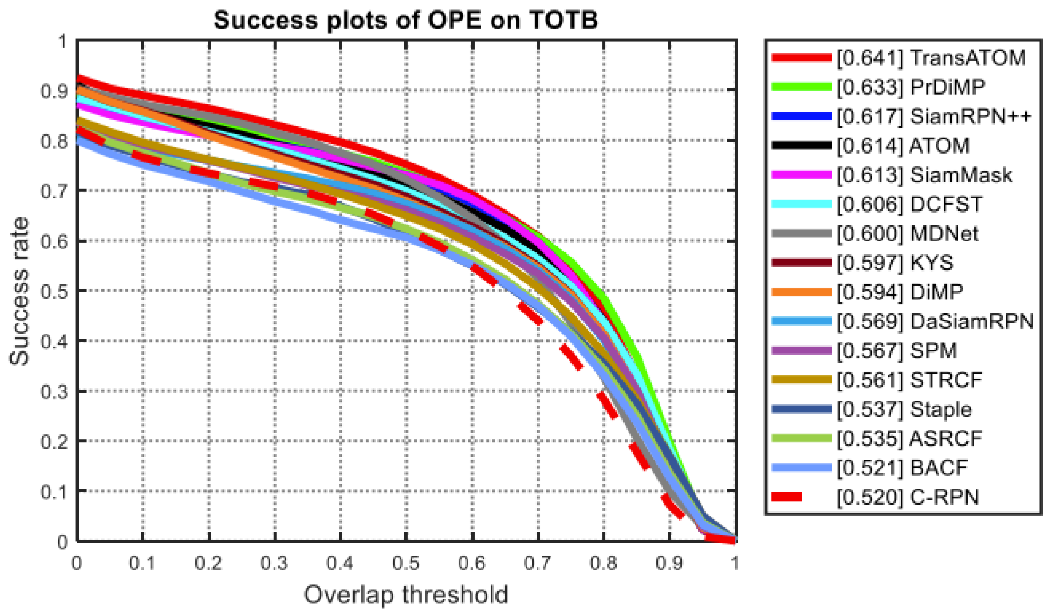
Importance of training sets

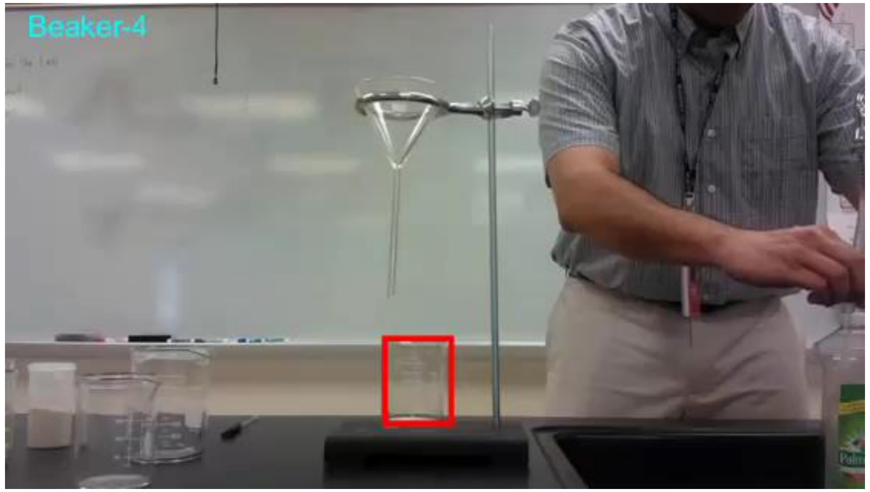
- Currently commonly used single-target training datasets:
  - TrackingNet[1]: 30k training videos from YouTube, box GT
  - GOT10k[2]: ~10k training videos, box GT
  - LaSOT[3]: >1k training videos, box GT
  - COCO[4]: 330k *images*, object detection dataset, augmentation to simulate pairs
  - YoutubeVOS[5]: 3.5k training segmentation videos
- Evidence emerging that unsupervised pre-training of the tracking architectures highly important for obtaining top 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!"

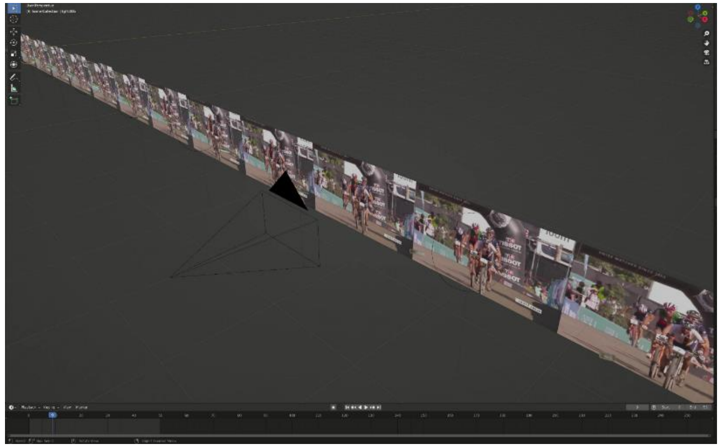


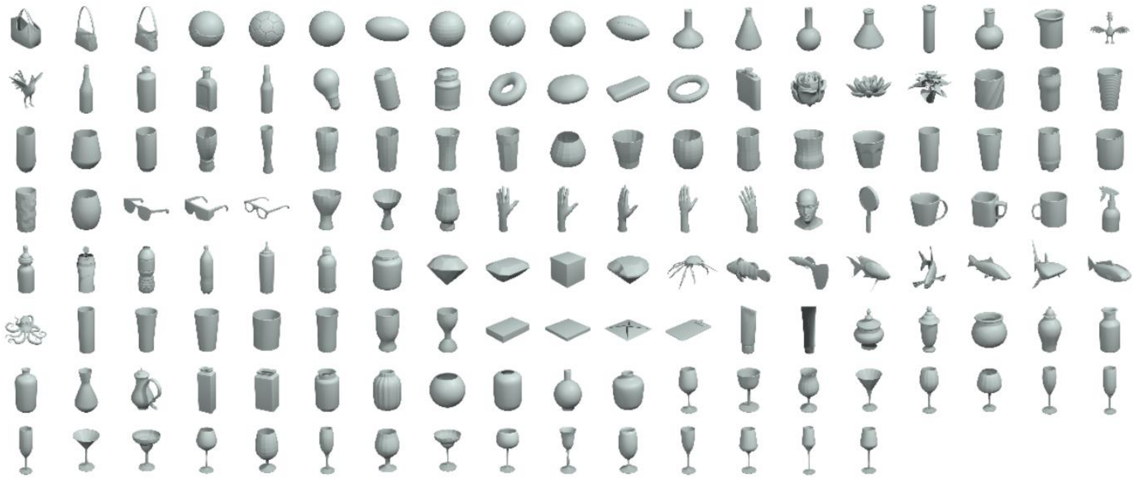


[1]H. Fan, et al., Transparent Object Tracking Benchmark, ICCV 2021

Importance of training datasets: TOTB example

- Transparent objects (glass/plastic) well rendered by modern renderers
- Benefits: Potentially unlimited training sequences, automatic annotation
- Trans2k[2] training dataset:
  - Background: existing video from GoT-10k
  - Motion: Random periodic trajectory
  - Rendering engine: BlenderProc [1]





[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]())

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

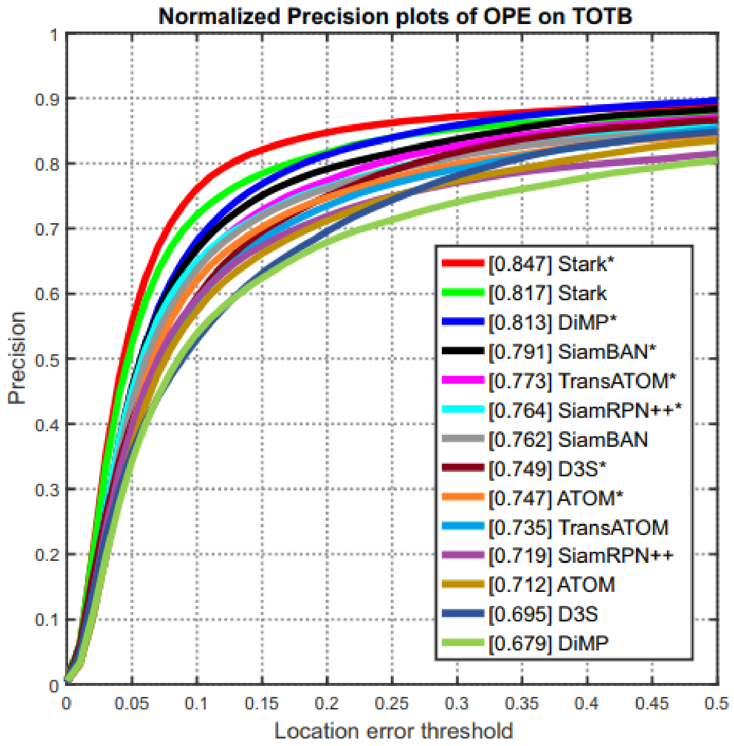
Trans2k: transparent object training dataset

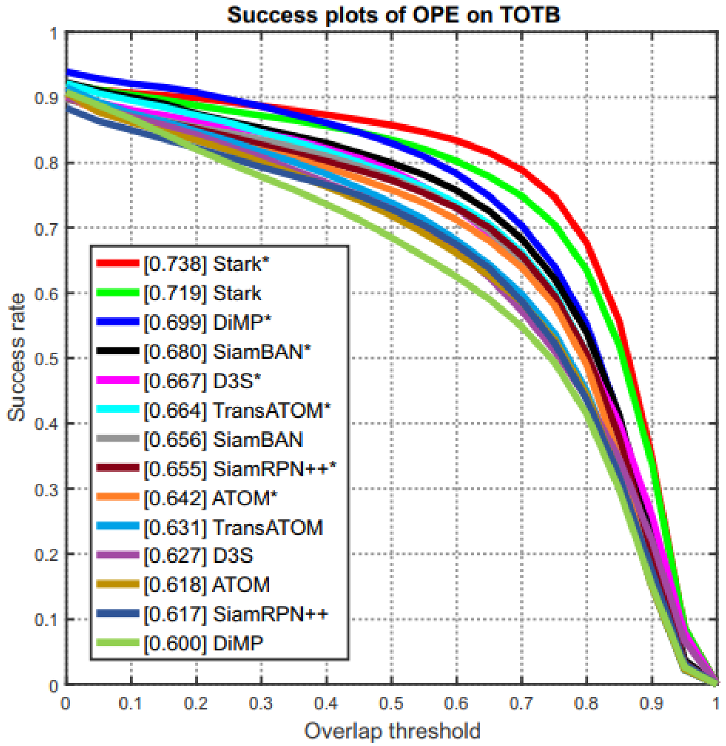
- 2000 training sequences
- 104,343 frames
- Target position annotation: Bounding box + segmentation

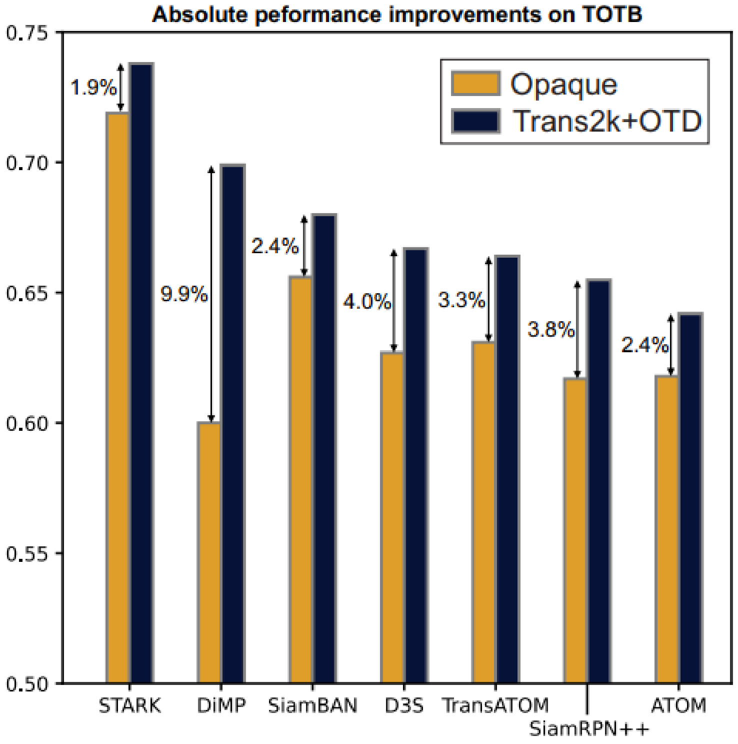
[1]Trojer, Lukežič, Matas, Kristan, [Trans2k: Unlocking the Power of Deep Models for Transparent Object Tracking](), BMVC2022 ([GIT]())

Trackers trained on Trans2k

- Standard trackers re-trained on Trans2k+GOT10k
- Evaluated on TOTB[1] [1]H. Fan, et al., Transparent Object Tracking Benchmark, ICCV 2021





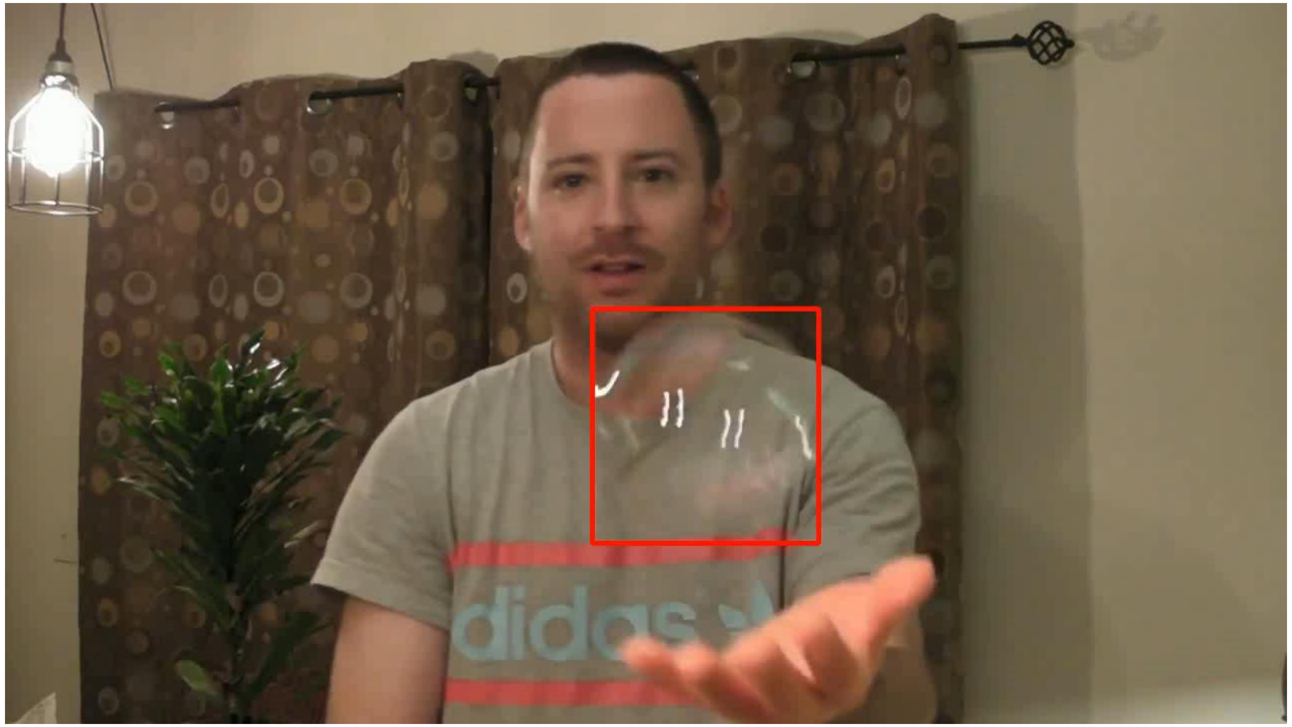


- Up to 10 percentage points performance improvements (~16% boost!)

[1]Trojer, Lukežič, Matas, Kristan, Trans2k: Unlocking the Power of Deep Models for Transparent Object Tracking, BMVC2022 (GIT)

Trackers trained on Trans2k

DiMP [1]



STARK [2]

Original tracker
Re-trained on Trans2k
Ground-truth

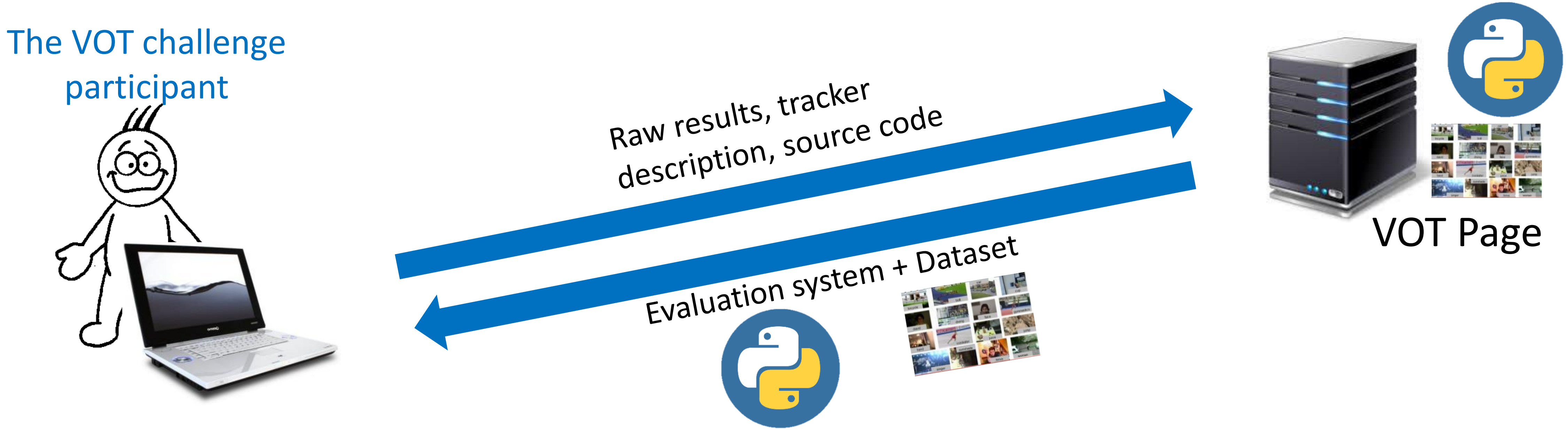
[1] M. Danelljan, et al., Learning discriminative model prediction for tracking, ICCV 2019
[2] B. Yan, et al., Learning spatio-temporal transformer for visual tracking, ICCV 2021

[1]Trojer, Lukežič, Matas, Kristan, Trans2k: Unlocking the Power of Deep Models for Transparent Object Tracking, BMVC2022 (GIT)

Visual Object Tracking Challenge VOT

THE CHALLENGES AND WORKSHOPS

Building the community: The VOT challenge



- 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)

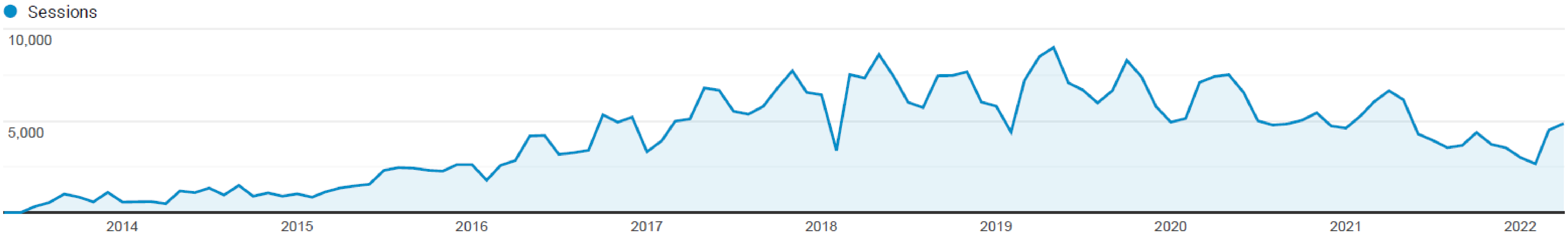
The VOT challenge evolution

	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

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

The VOT community evolution

VOT2013 VOT2014 VOT2015 VOT2016 VOT2017 VOT2018 VOT2019 VOT2020 VOT2021 VOT2022



- Results paper @ major CV conference (ECCV/ICCV) workshops
- Annually ~100 coauthors on the results papers
- On average >60 trackers evaluated annually

Evolution of VOT ST challenge submitted trackers

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
# track.	27	38	62	70	51	72	57	37	53	31
Submitted trackers design types										
Top performing										

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., "A Novel Performance Evaluation Methodology for Single-Target Trackers", IEEE TPAMI 2016

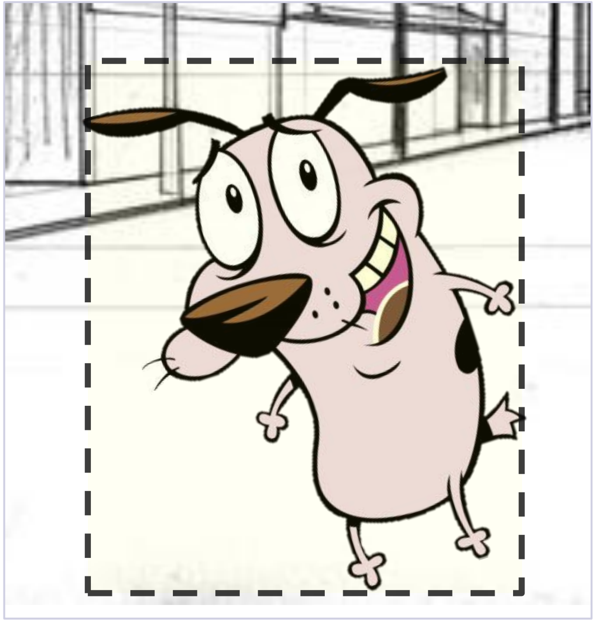
VOT-ST2022 challenge variations

- Bounding boxes abandoned in VOT2020, but reintroduced in 2022 due to pertaining significant research interest in the community
- Standard VOT anchor-based evaluation used (A, R, EAO)

VOT-STs2022

Segmentation mask

VOT-STb2022



Bounding box

Realtime constraint:

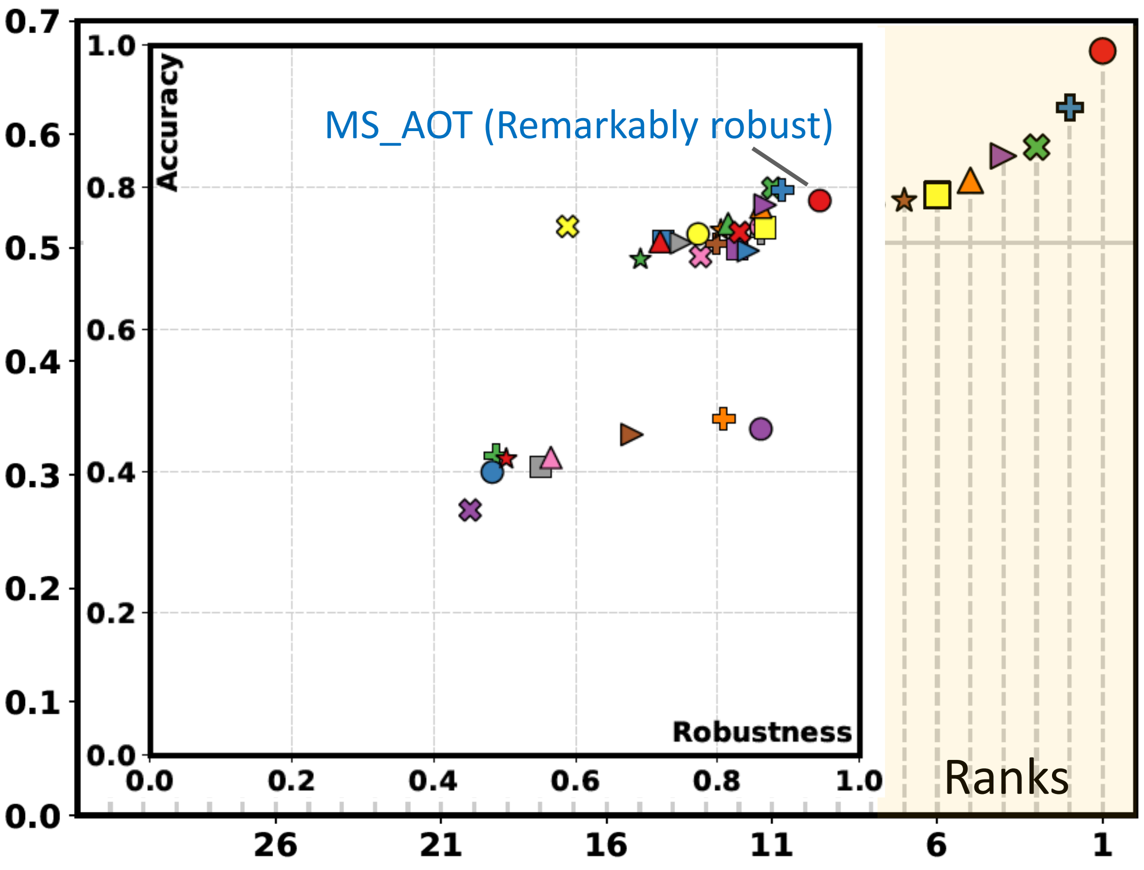
- *Process @20fps*
- *Winners identified on the public dataset*

Variants:

- VOT-RTs2022
- VOT-RTb2022

VOT-STs2022 results on public dataset (31 trackers)

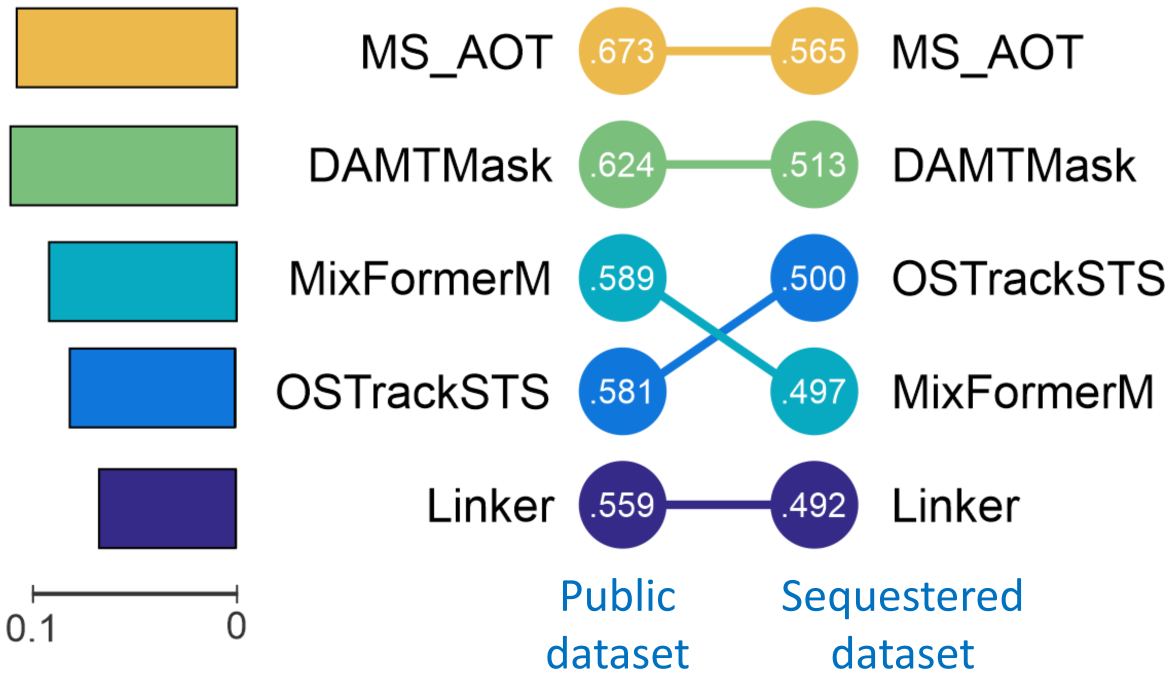
- Top trackers: (1) MS_AOT, (2) DAMTMask (3) MixFormerM, (4) OSTrackSTS, (5) Linker, (6) SRATransTS, (7) TransT_M, (8) DGformer, (9) TransLL, (10) LWL-B2S
- Core methodology:
  - 9 transformers, 1 deep DCF
  - Most use: Mixformer[1], TransT[2]
  - 7 two-stage:
    (i) box localization + (ii) segmentation
- Top performer (MS_AOT) stands out:
  - Single-stage, based on pure
    video object segmentation method[1]

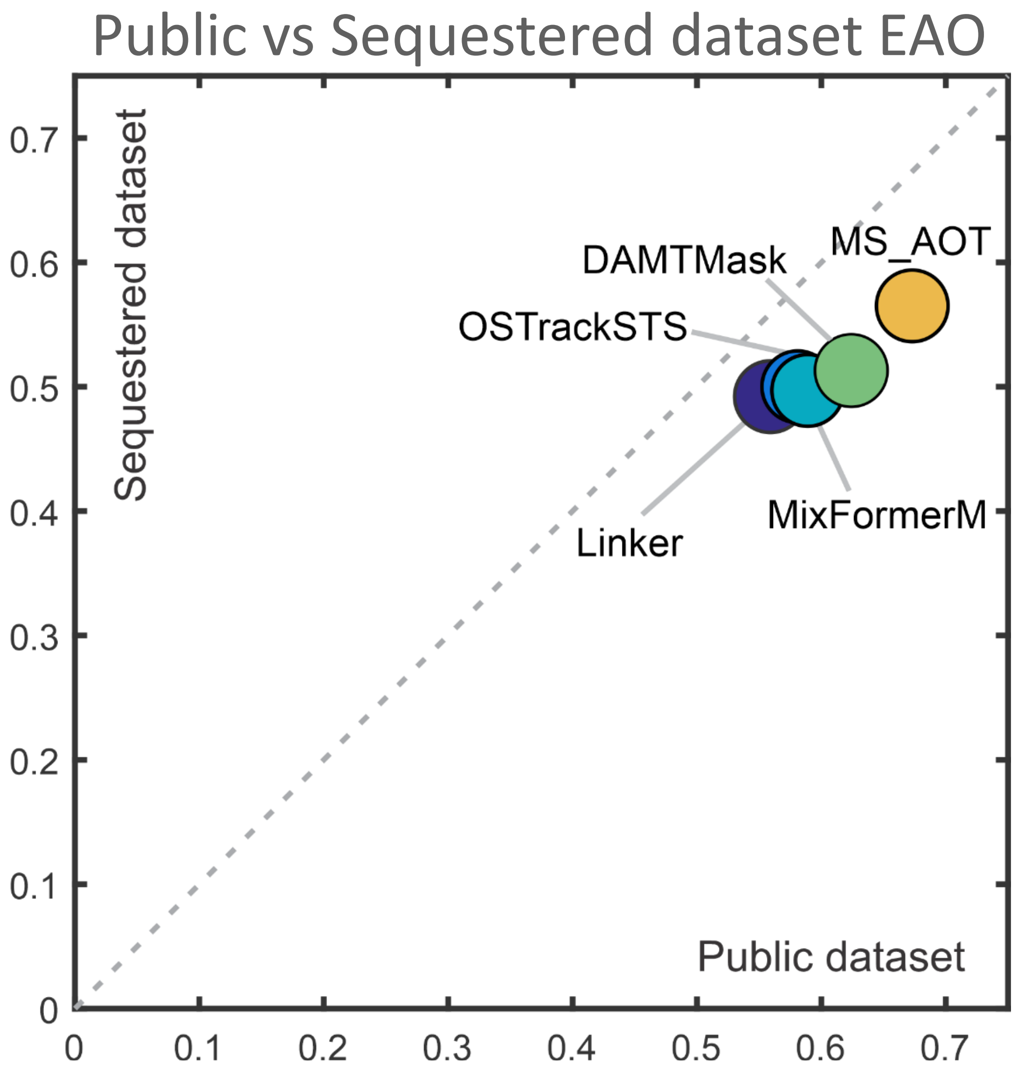


[1]Cui et al. CVPR2022, [2]Chen et al. CVPR2021, [3]Yang et al. Neurips 2021

VOT-STs2022 results on sequestered dataset

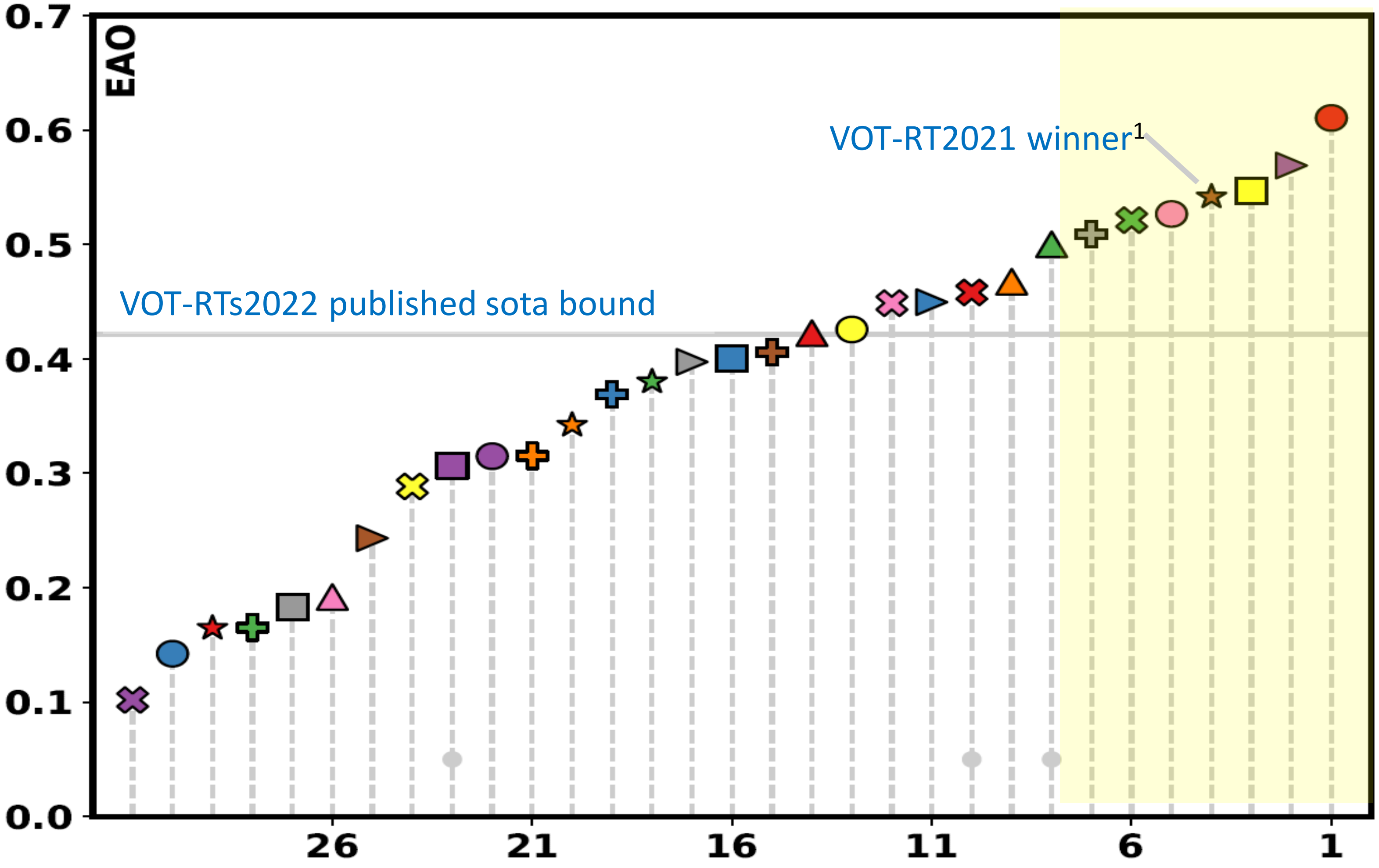
- Comparable results between public and sequestered set
  - Slight relative performance differences
  - Clearly stands out: MS_AOT





VOT-RTs2022 realtime challenge results

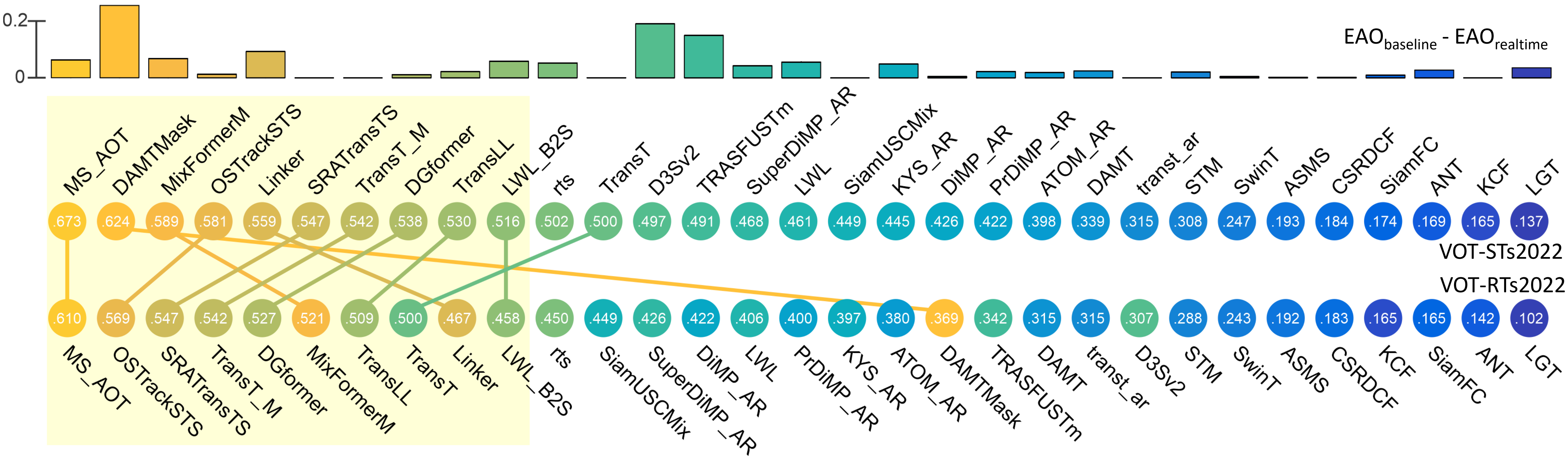
- Top 10: (1) MS_AOT, (2) OSTrackSTS, (3) SRATransTS, (4) TransT_M, (5) DGformer, (6) MixFormerM, (7) TransLL, (8) TransT, (9) Linker, (10) RTS



- 9 are transformers
- 3 outperform the VOT-RT2021 winner[1]
- Top: MS_AOT
- 45% of submissions outperform VOT-RT2022 sota bound

[1]TransT_M [Chen et al., Arxiv2022]

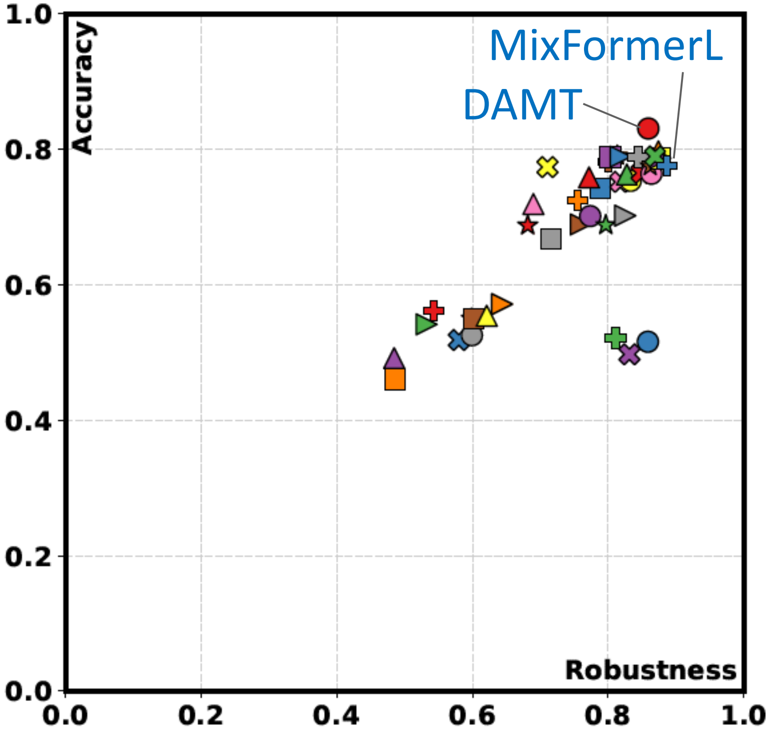
VOTs2022 Realtime vs Baseline results



- 9 top VOT-RTs2022 trackers among top 10 on VOT-STs2022 challenge!
- The top RT tracker MS_AOT is top in VOT-STs2022

VOT-STb2022 results on public dataset (41 trackers)

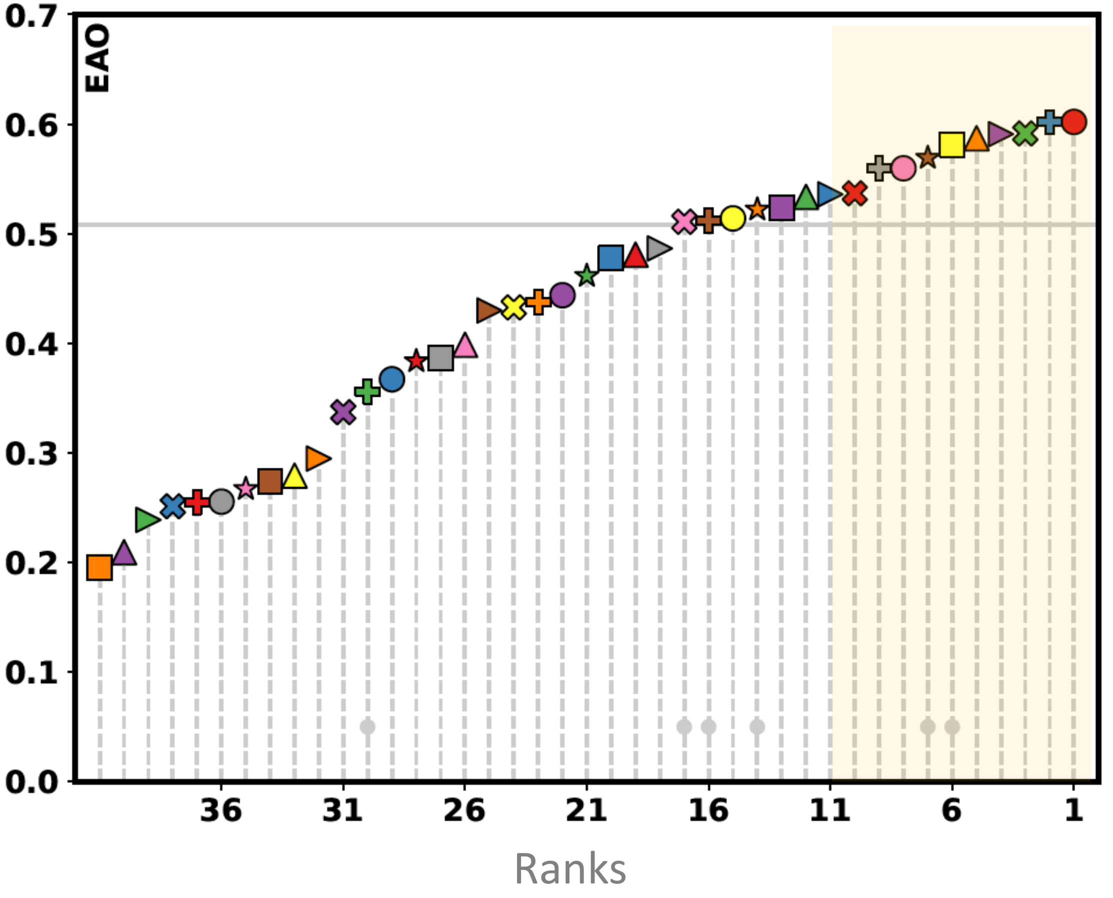
- Top trackers: (1) DAMT, (2) MixFormerL, (3) OSTrackSTB, (4) APMT_MR, (5) Mixformer, (6) APMT_RT, (7) ADOTstb, (8) SRATransT, (9) Linker_B, (10) TransT_M
- All top trackers are transformers



Same (top) EAO:
DAMT & MixFormerL

More accurate

More robust



Box trackers vs Segmentation trackers

- VOT-STb2022 top 10 performers:
  - 7 perform well in VOT-STs2022
  - 3 of top 4 VOT-STb2022 are among top VOT-STs2022 (3rd, 2nd, 4th)

VOT-STb2022

Tracker	EAO	A	R
MixFormerL	0.602①	0.831①	0.859
DAMT	0.602②	0.776	0.887①
OSTrackSTB	0.591③	0.790	0.869
APMT_MR	0.591	0.787	0.877③
MixFormer	0.587	0.797②	0.874
APMT_RT	0.581	0.787	0.877②
ADOTstb	0.569	0.775	0.862
SRATransT	0.560	0.764	0.864
Linker_B	0.560	0.789	0.844
TransT_M	0.537	0.765	0.849
vittrack	0.536	0.789	0.818
SuperFus	0.534	0.763	0.828
SwinTrack	0.524	0.788	0.803

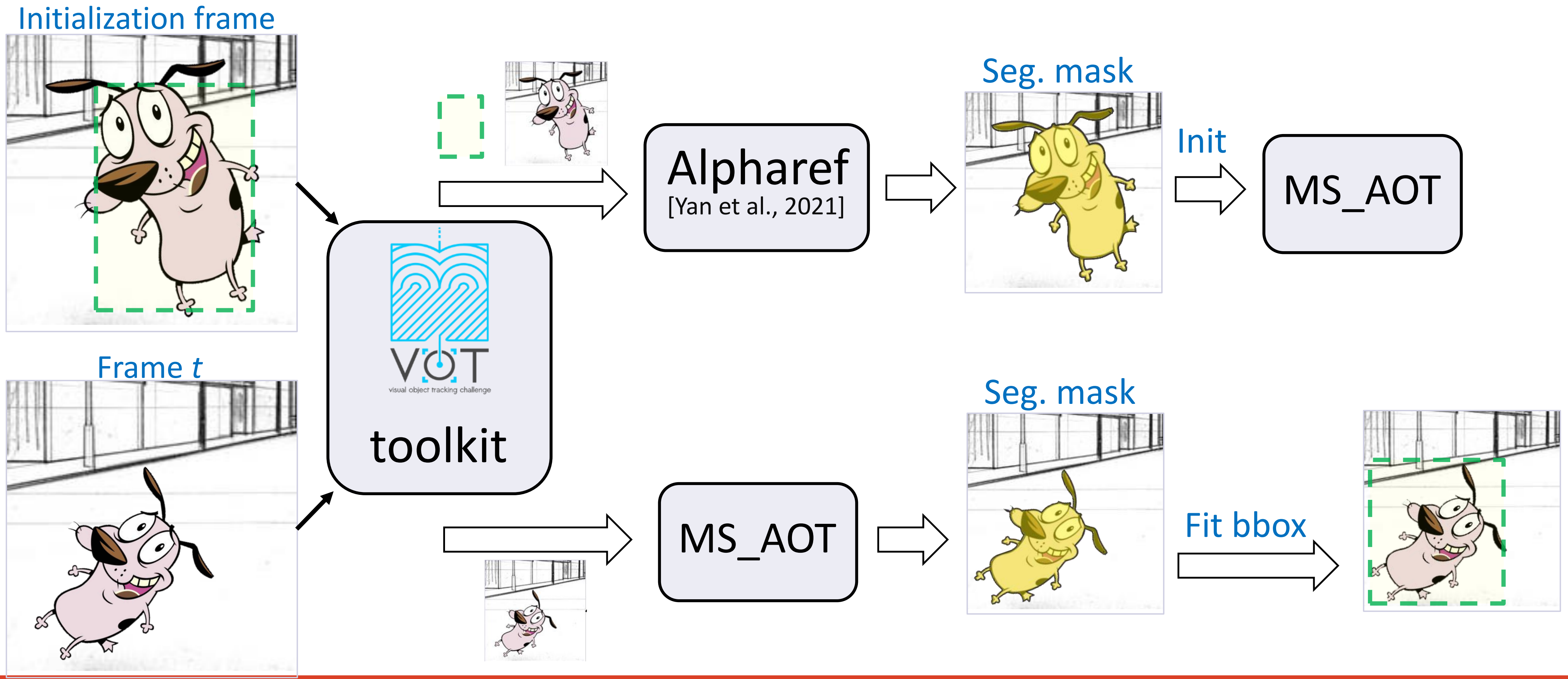
VOT-STs2022

Tracker	EAO	A	R
MS_AOT	0.673①	0.781③	0.944①
DAMTMask	0.624②	0.796②	0.891②
MixFormerM	0.589③	0.799①	0.878③
OSTrackSTS	0.581	0.775	0.867
Linker	0.559	0.772	0.861
SRATransTS	0.547	0.743	0.866
TransT_M	0.542	0.743	0.865
DGformer	0.538	0.744	0.861
TransLL	0.530	0.735	0.861
LWL_B2S	0.516	0.736	0.831
rts	0.502	0.710	0.843
TransT	0.500	0.749	0.815
D3Sv2	0.497	0.713	0.827

Box trackers vs Segmentation trackers

- VOT-STs2022 winner MS_AOT run on public STb2022 dataset
  - Initialize by AlphaRef[1] ; Output is bounding box fited to mask prediction

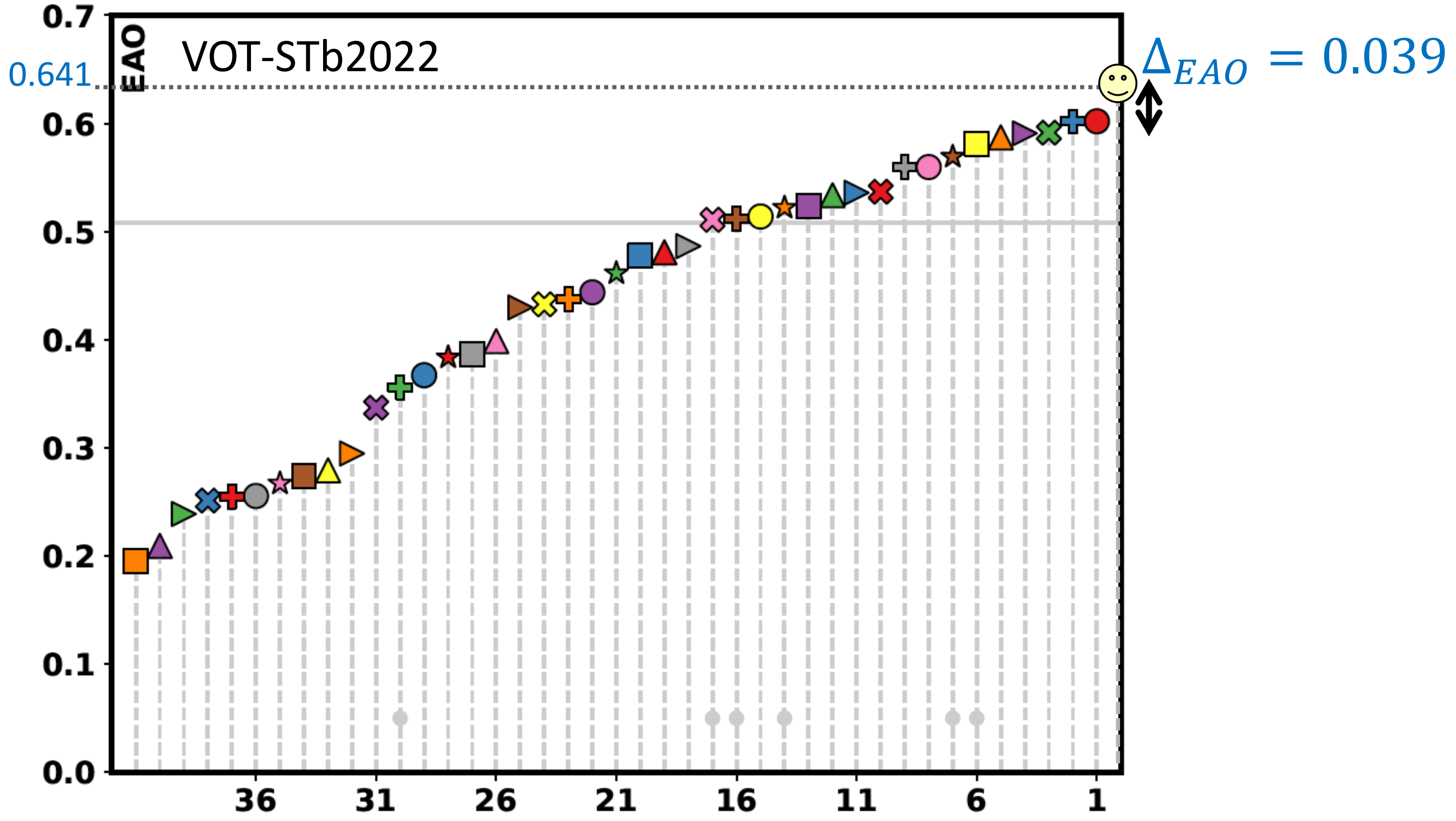
[1][Yan et al., CVPR2021]

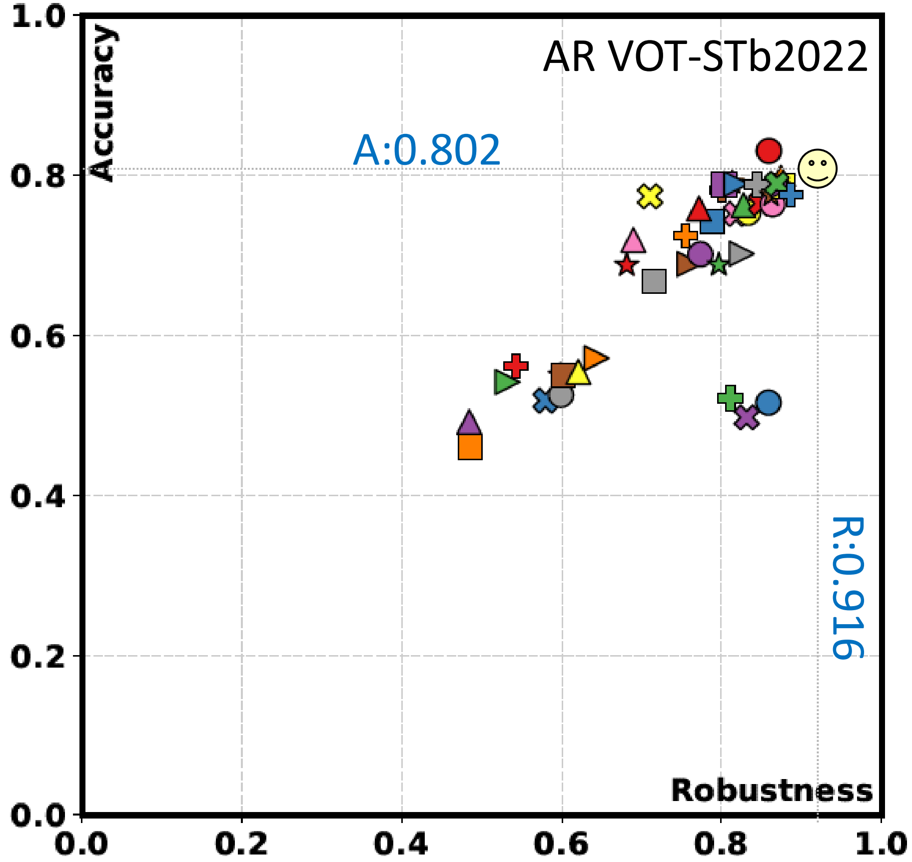


Box trackers vs Segmentation trackers

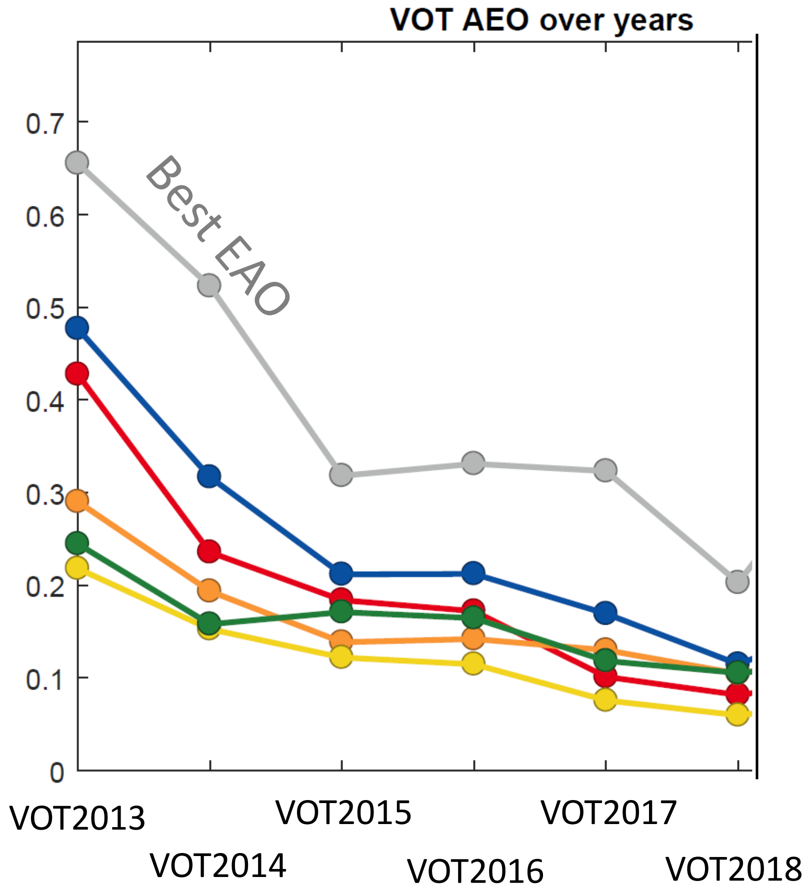
- VOT-STs2022 winner MS_AOT run on public STb2022 dataset
  - Initialize by AlphaRef[1] ; Output is bounding box fited to mask prediction [1][Yan et al., CVPR2021]

    (MS_AOT) EAO: 0.641, A:0.802, R:0.916





The VOT ST datasets tracking difficulty



Dataset increasingly more challenging

VOT-LT2022 results

- Top 3 trackers: *VITKT_M, mixLT,* and *HuntFormer*
  - *Fusion of multiple trackers and motion prediction model*
- Top performance: VITKT_M
  - Trackers: STARK[1] + KeepTrack[2]
  - A simple motion module (~1.2% improved)
- Second-best (~1.7% Worse) : mixLT
  - STARK + SuperDiMP[3]
- Baseline: mlpLT (winner of VOT-LT2021)
  - 4 trackers outperformed the VOT-LT2021 winner

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

[1] Yan et al. ICCV2021
[2] Mayer et al. ICCV2021
[3] Bhat et al. ECCV2020

VOT-LT: 2018 vs 2020

FCLT [Lukežič et al. ACCV2018]

A state-of-the-art LT tracker in 2017

LTMUB [Dai et al. VOT2020]

Top LT tracker in VOT2020

VOT2022 ST/RT/LT challenges summary

- VOT-ST2022:
  - Transformers became the dominant methodology of top trackers
  - Observed emergence of remarkably robust segmentation trackers
  - The same segmentation tracker won STs & RT challenge, and would have won STb (!)
  - Invest more research into purely segmentation trackers

- VOT-LT2022:
  - Top tracker: Transformer-based & mixed with distractor tracking + motion model
  - Significant advancements made since 2018

Beyond the VOT challenges

- VOT has focused on (short-term, long-term) single-target tracking
- In parallel, substantial advances made in:
  - Video object segmentation (but focused on video editing of short videos)
    YouTubeVOS
  - Multiple target trackers (but focused on pre-trained categories, e.g., people)
    MotComplex, TAO-OW, STEP (with segmentation)
- A new chapter: Visual Object Tracking Segmentation VOTS2023
  - Short and Long-term tracking converged
  - Primary output: segmentation
  - Tracking of multiple general targets
  - Challenge opened last week, results presented @ICCV2023

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)

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

Twitter updates
https://twitter.com/votchallenge

Thanks

- The VOT2022 committee


M. Kristan


J. Matas


A. Leonardis


J. K. Kamarainen


H. J. Chang


R. Pflugfelder


G. Fernandez


L. Čehovin


A. Lukežič


M. Felsberg

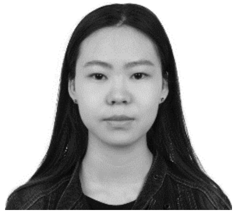

M. Danelljan


O. Drbohlav


Z. Zhang


Y. Song


Wenyan Yang


Z. Yushan


C. Mayer


Dingding Cai


Johanna Björklund

- Everyone who participated or contributed

Kang Ben18, Goutam Bhat10, Hong Chang24, Guangqi Chen16, Jiaye Chen26, Shengyong Chen43, Xilin Chen24, Xin Chen18, Xiuyi Chen13, Yiwei Chen35, Yu-Hsi Chen12, Zhixing Chen16, Yangming Cheng55, Angelo Ciaramella47, Yutao Cui30, Benjamin Dˇzubur1, Mohana Murali Dasari22, Qili Deng16, Debajyoti Dhar39, Shangzhe Di14, Emanuel Di Nardo46,47, Daniel K. Du16, Matteo Dunnhofer51, Heng Fan48, Zhenhua Feng50, Zhihong Fu16, Shang Gao41, Rama Krishna Gorthi22, Eric Granger27, Q.H. Gu15, Himanshu Gupta19, Jianfeng He49, Keji He13, Yan Huang13, Deepak Jangid19, Rongrong Ji53, Cheng Jiang30, Yingjie Jiang26, Felix J¨aremo Lawin4, Ze Kang26, Madhu Kiran27, Josef Kittler50, Simiao Lai18, Xiangyuan Lan32, Dongwook Lee34, Hyunjeong Lee34, Seohyung Lee34, Hui Li26, Ming Li17, Wangkai Li49, Xi Li55, Xianxian Li20, Xiao Li16, Zhe Li41, Liting Lin37, Haibin Ling40, Bo Liu25, Chang Liu18, Si Liu23, Huchuan Lu18, Rafael M. O. Cruz27, Bingpeng Ma44, Chao Ma36, Jie Ma21, Yinchao Ma49, Niki Martinel51, Alireza Memarmoghadam45, Christian Micheloni51, Payman Moallem45, Le Thanh Nguyen-Meidine27, Siyang Pan35, ChangBeom Park34, Danda Paudel10, Matthieu Paul10, Houwen Peng28, Andreas Robinson4, Litu Rout39, Shiguang Shan24, Kristian Simonato51, Tianhui Song30, Xiaoning Song26, Chao Sun55, Jingna Sun16, Zhangyong Tang26, Radu Timofte10,52, Chi-Yi Tsai42, Luc Van Gool10, Om Prakash Verma19, Dong Wang18, Fei Wang49, Liang Wang13, Liangliang Wang16, Lijun Wang18, Limin Wang30, Qiang Wang35, Gangshan Wu30, Jinlin Wu13, Xiaojun Wu26, Fei Xie38, Tianyang Xu26, Wei Xu16, Yong Xu37, Yuanyou Xu55, Wanli Xue43, Zizheng Xun14, Bin Yan18, Dawei Yang49, Jinyu Yang41, Wankou Yang38, Xiaoyun Yang33, Yi Yang55, Yichun Yang30, Zongxin Yang55, Botao Ye24, Fisher Yu10, Hongyuan Yu13, Jiaqian Yu35, Qianjin Yu49, Weichen Yu13, Kang Ze26, Jiang Zhai38, Chengwei Zhang17, Chunhu Zhang36, Kaihua Zhang29, Tianzhu Zhang49, Wenkang Zhang38, Zhibin Zhang43, Zhipeng Zhang31, Jie Zhao18, Shaochuan Zhao26, Feng Zheng41, Haixia Zheng54, Min Zheng16, Bineng Zhong20, Jiawen Zhu18, Xuefeng Zhu26, and Yueting Zhuang55

- VOT2022 sponsor:

University *of Ljubljana*
Faculty *of Computer and Information Science*