AAA: Adaptive Aggregation of Arbitrary Online Trackers with a Regret Bound

九州大学　システム情報科学府
修士2年　ソン　ホン

15:20 ~ 15:50

July 14, 2020

KYUSHU
UNIVERSITY

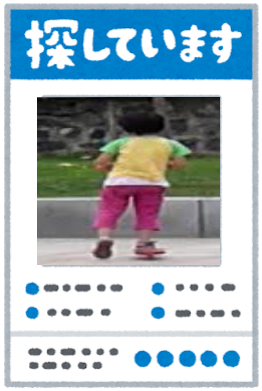
Index

Single object tracking (SOT)

INTRODUCTION

What is online Single Object Tracking (SOT)?

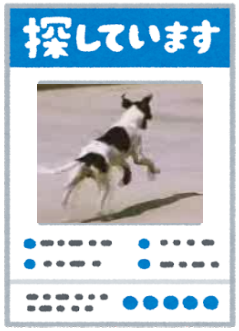
- Given a target location at the first frame,
  we should track the target and predict the location.

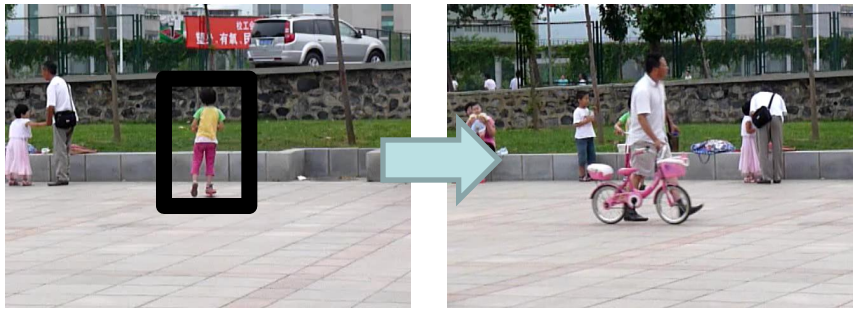


Frame 1 Frame 2 Frame 3

What makes online SOT difficult?

- Depending on videos,
  we may have to track completely different objects.
- There might be a heavy appearance change
  of the target or occlusion even in a video.





What makes online SOT difficult?

No almighty online tracker

Aggregate multiple online tracker for robust tracking!

I track animals well

I track plastics well

I track occluded object well

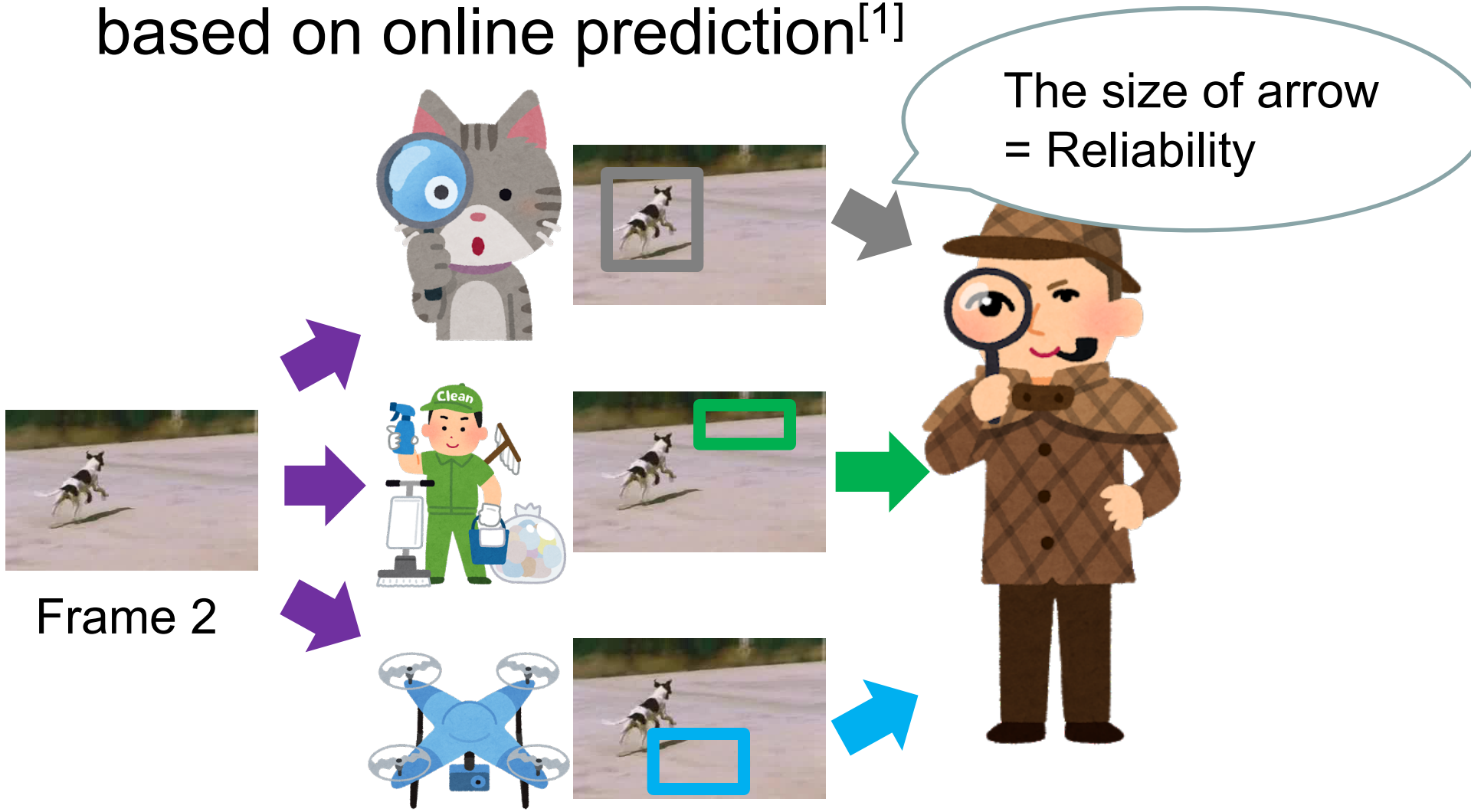
Main contribution

- Depending on videos,
  we may have to track completely different objects.
- There might be a heavy appearance change
  of the target or occlusion even in a video.

Single object tracking (SOT)

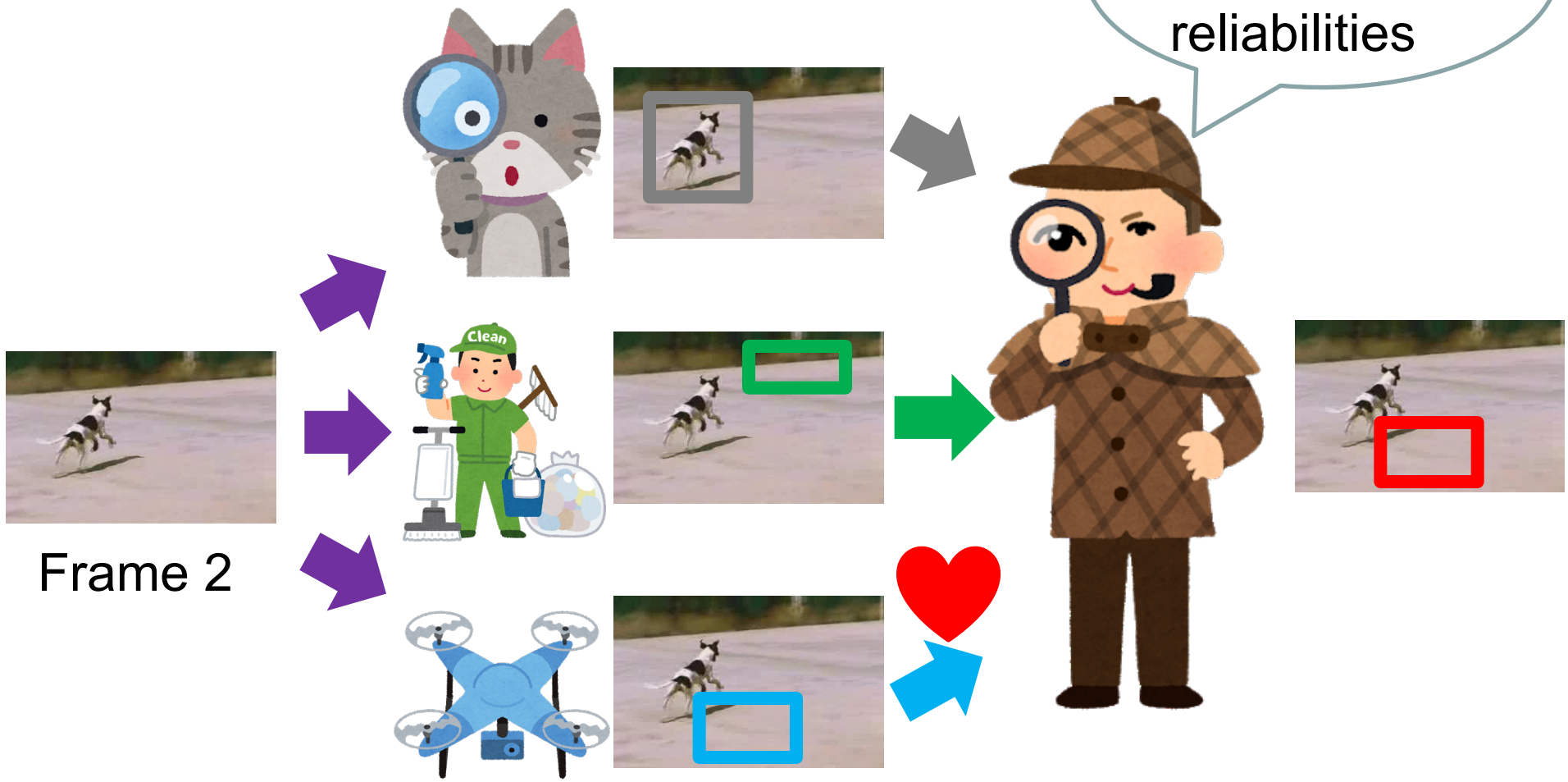
TRACKING WITH MULTIPLE ONLINE TRACKER

Aggregate multiple online trackers based on online prediction[1]



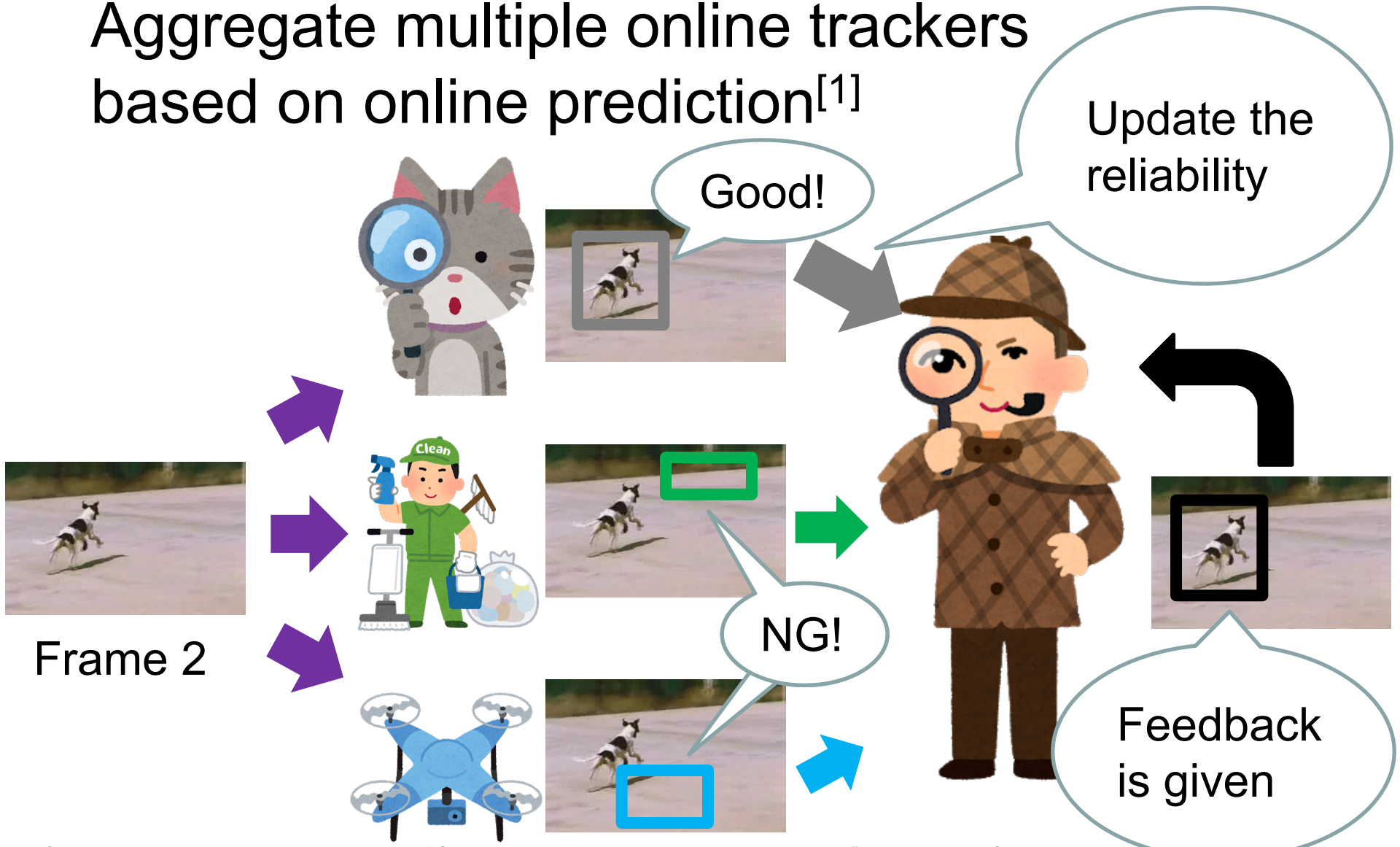
[1] Quanrud, Kent and Daniel Khashabi, “Online Learning with Adversarial Delays,” in Proc. NIPS, 2014.

Aggregate multiple online trackers based on online prediction[1]



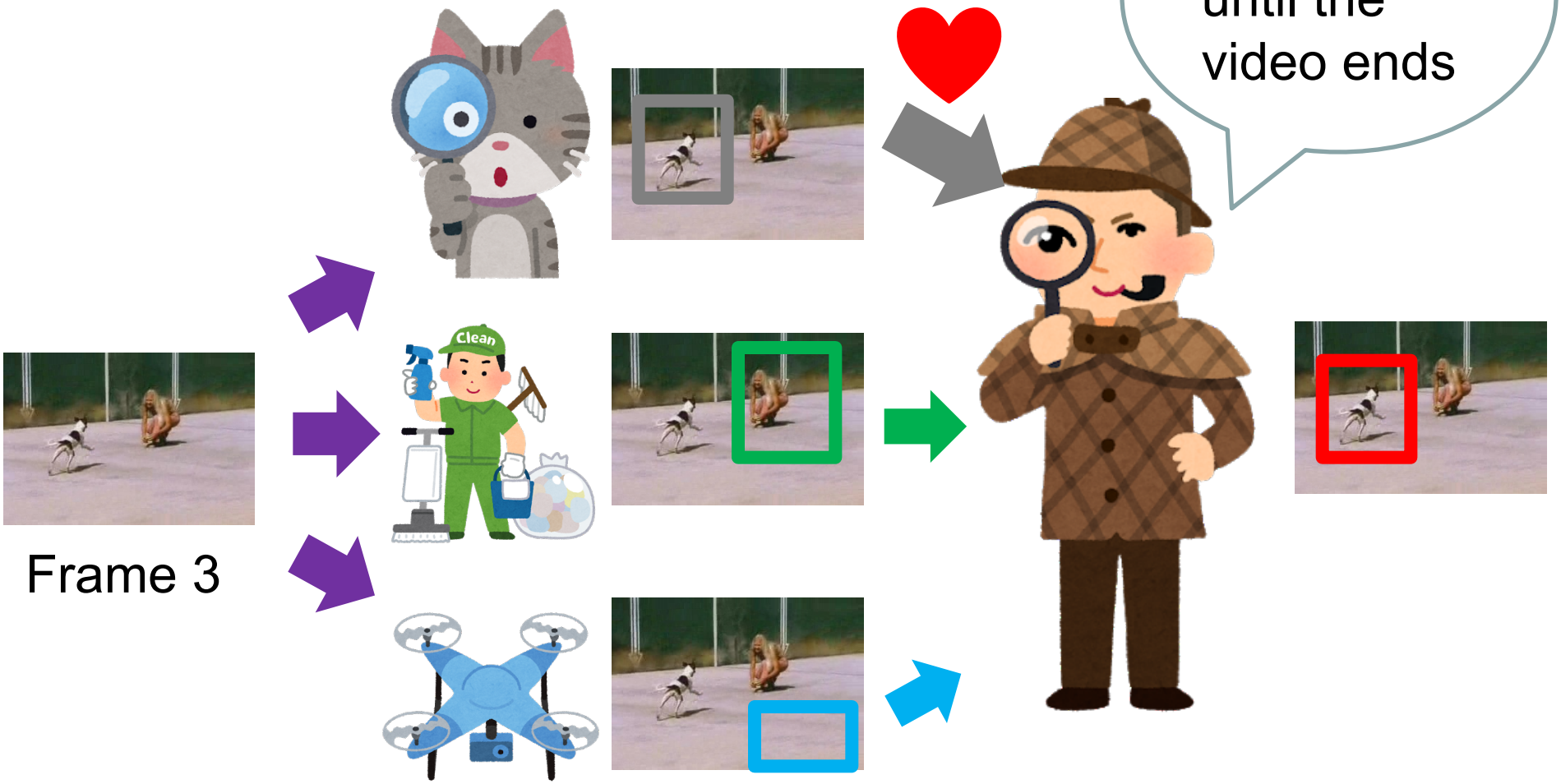
[1] Quanrud, Kent and Daniel Khashabi, “Online Learning with Adversarial Delays,” in Proc. NIPS, 2014.

Aggregate multiple online trackers based on online prediction[1]



[1] Quanrud, Kent and Daniel Khashabi, “Online Learning with Adversarial Delays,” in Proc. NIPS, 2014.

Aggregate multiple online trackers based on online prediction[1]



[1] Quanrud, Kent and Daniel Khashabi, “Online Learning with Adversarial Delays,” in Proc. NIPS, 2014.

How can we get near-true label?















Anchor frame

Frame 1 ⇨ Frame 2 ⇨ Frame 3 ⇨ Frame 4

Anchor frame

How can we get near-true label?

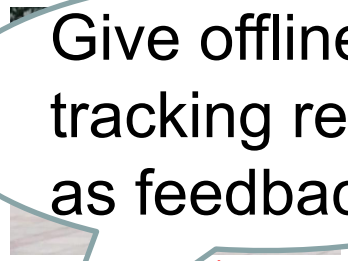
















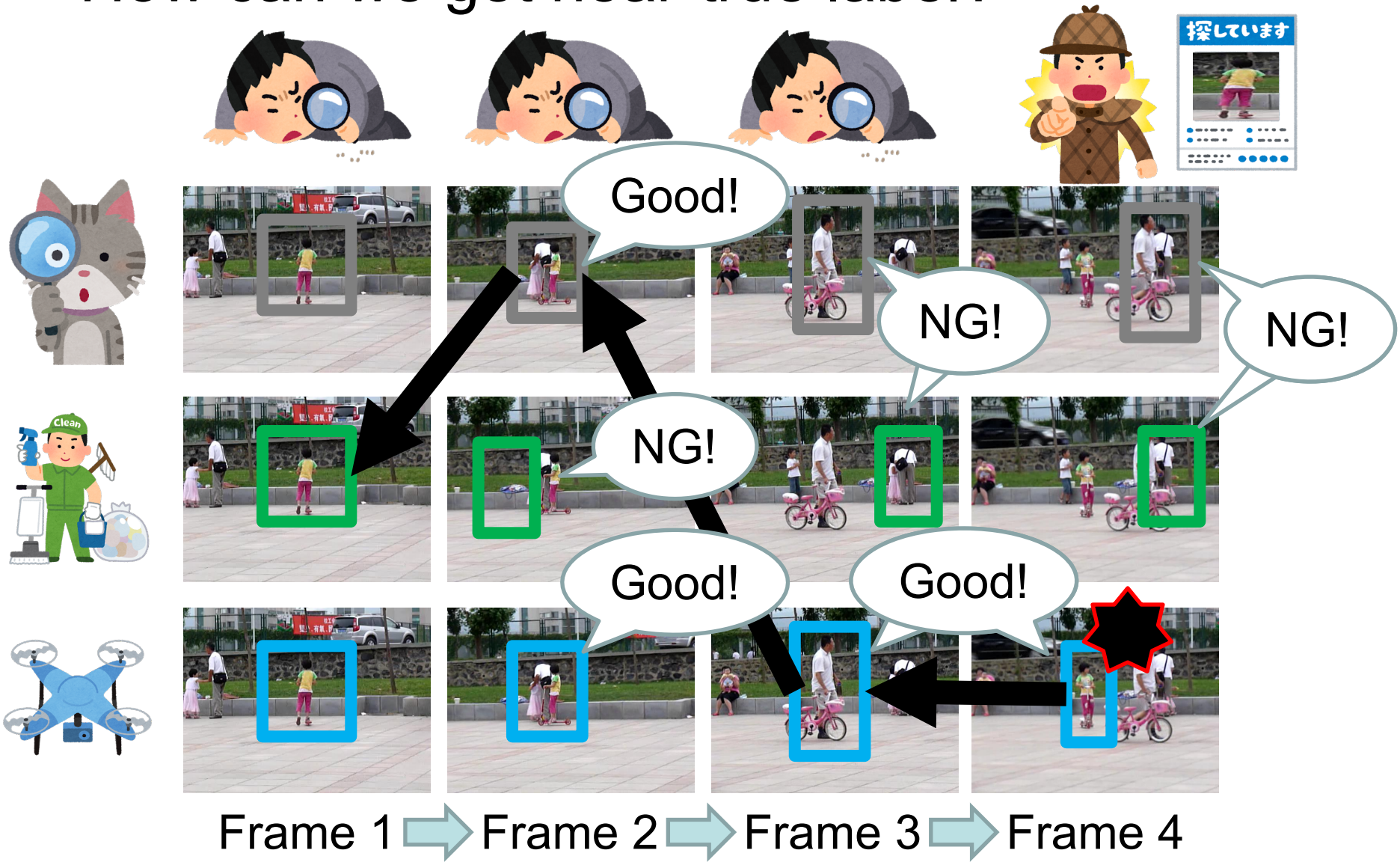
Give offline tracking results as feedback

Anchor frame

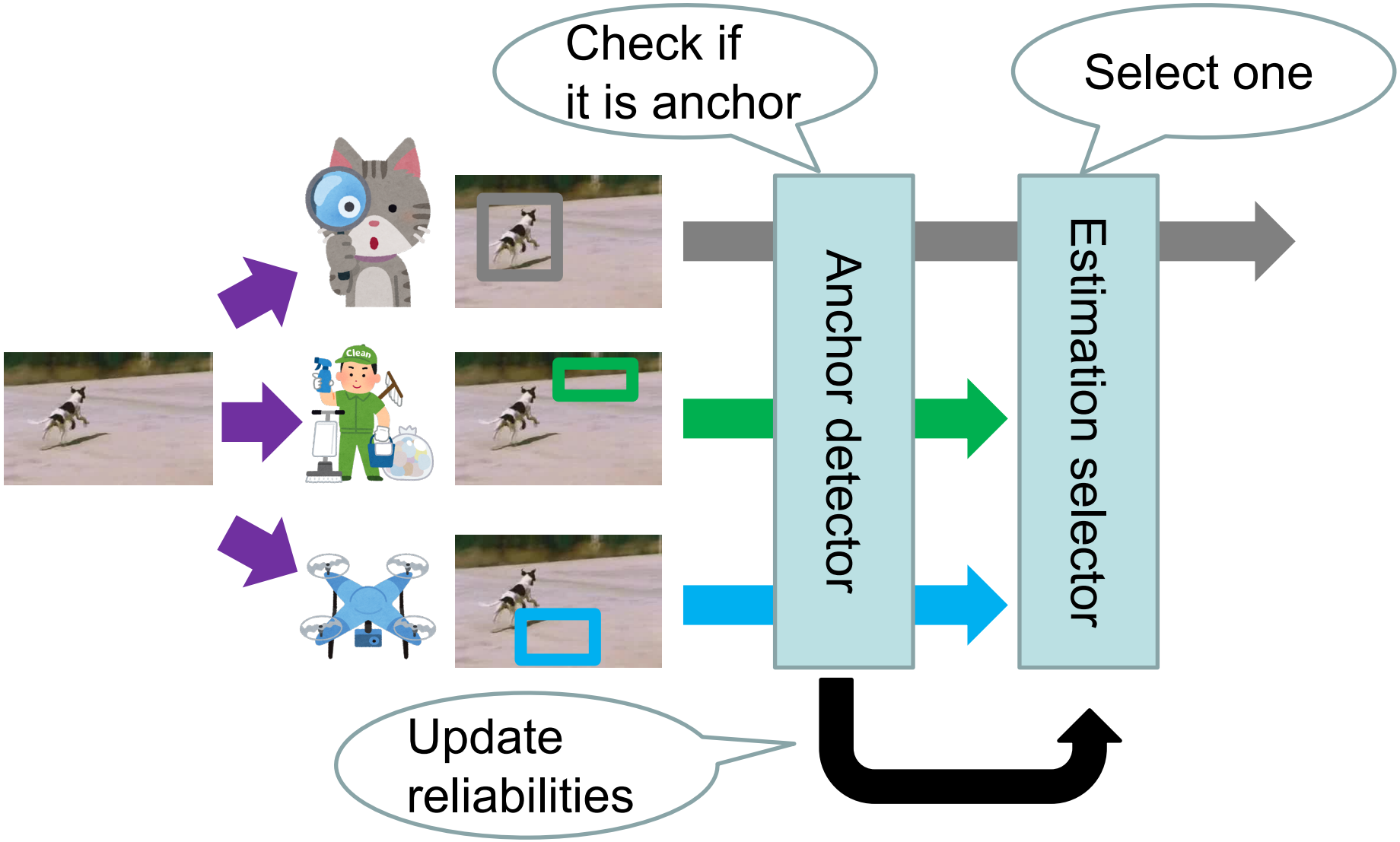
Anchor frame

Frame 1 ⇨ Frame 2 ⇨ Frame 3 ⇨ Frame 4

How can we get near-true label?



Overview of the proposed method



Why should we use online prediction?

Total error L_1^T

Total error L_2^T

Total error L_3^T

minimum

Defined only at the end of the video

$L_2^T (= L_{best}^T)$

Difference
=Regret
$= L_{Ours}^T - L_{best}^T$

Total error L_{Ours}^T

L_{Ours}^T

Why should we use online prediction?

- Surprisingly, the regret of the proposed method is bounded without any assumption.

- We do not need to know who will be the best tracker, which data will be given, and how long the video is.

The length of the video

$$Regret = \boldsymbol{L}^{\boldsymbol{T}}_{\boldsymbol{Ours}} - \boldsymbol{L}^{\boldsymbol{T}}_{\boldsymbol{best}} \leq O\left(\sqrt{T \ln N}\right)$$

The number of the trackers

Single object tracking (SOT)

EXPERIMENTS & RESULTS

Experts and Benchmarks

- We would like to demonstrate that the proposed method is robust for arbitrary videos and trackers.
- For arbitrary video, we evaluate the proposed method on six different benchmarks.
- For arbitrary trackers, we prepare twelve trackers and divide into 2 groups according to their performance.

Comparision with High group

- High group includes six high-performance trackers.
- AAA outperforms the trackers on almost benchmarks

	OTB2015 [14]		TCo128 [15]		UAC123 [16]		NFS [17]		LaSOT [18]		VOT2018 [19]
	AUC	DP	AUC	DP	AUC	DP	AUC	DP	AUC	DP	AO
ATOM [2]	0.67	0.87	0.6	0.81	0.62	0.82	0.58	0.69	0.51	0.51	0.52
DaSiamRPN [3]	0.65	0.88	0.53	0.75	0.57	0.78	0.55	0.67	0.43	0.42	0.43
SiamMCF [4]	0.65	0.85	0.57	0.78	0.54	0.77	0.57	0.7	0.44	0.45	0.45
SiamRPN++ [5]	0.69	0.9	0.58	0.77	0.6	0.8	0.6	0.74	0.49	0.51	0.5
SPM [6]	0.67	0.87	0.58	0.79	0.59	0.77	0.57	0.67	0.47	0.48	0.48
THOR [7]	0.64	0.85	0.52	0.72	0.57	0.77	0.57	0.68	0.4	0.41	0.47
AAA	**0.7**	**0.91**	**0.62**	**0.84**	**0.62**	**0.83**	**0.61**	**0.75**	**0.53**	**0.55**	**0.52**

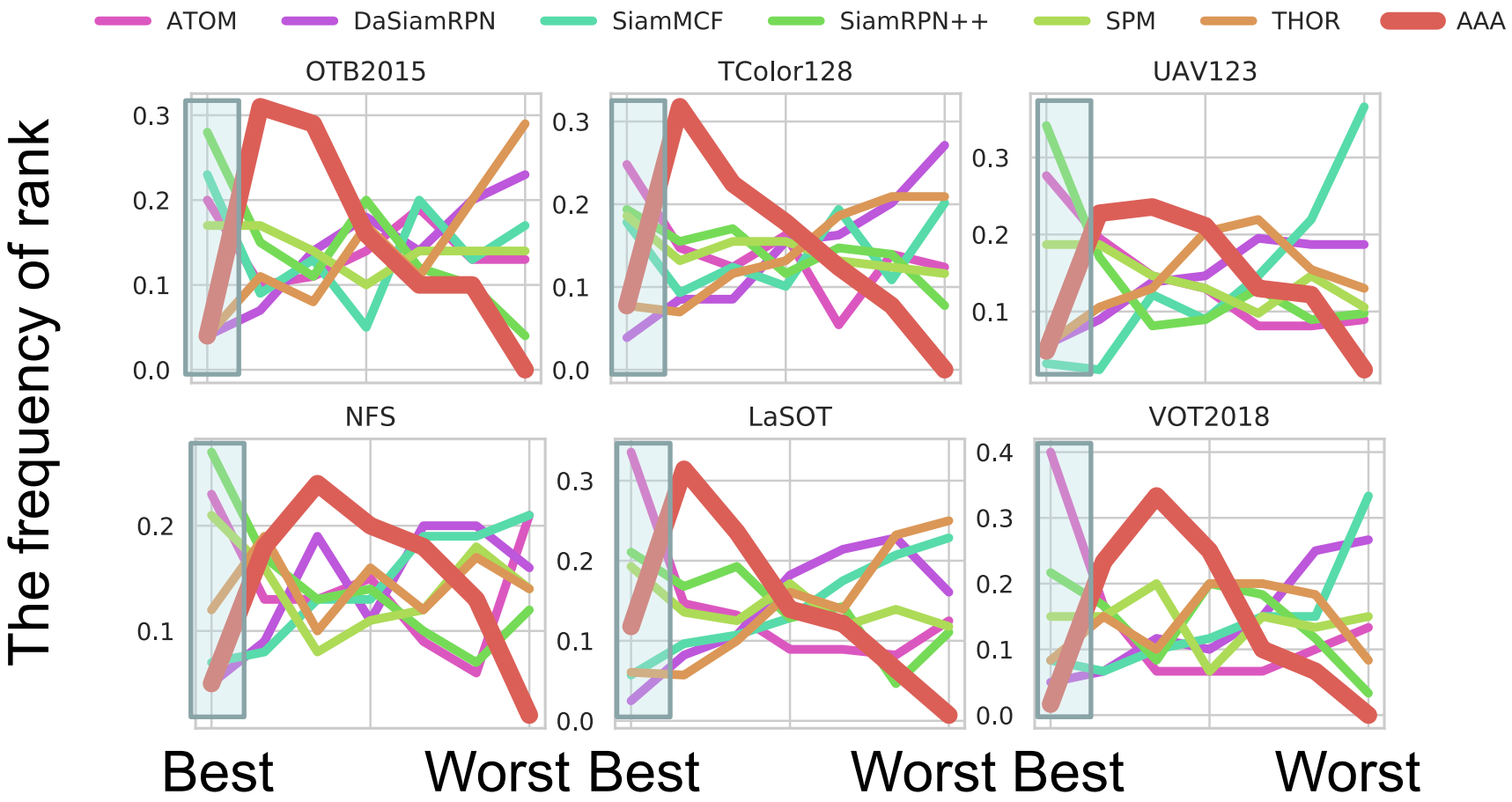
Aggregated tracekrs

2nd

1st

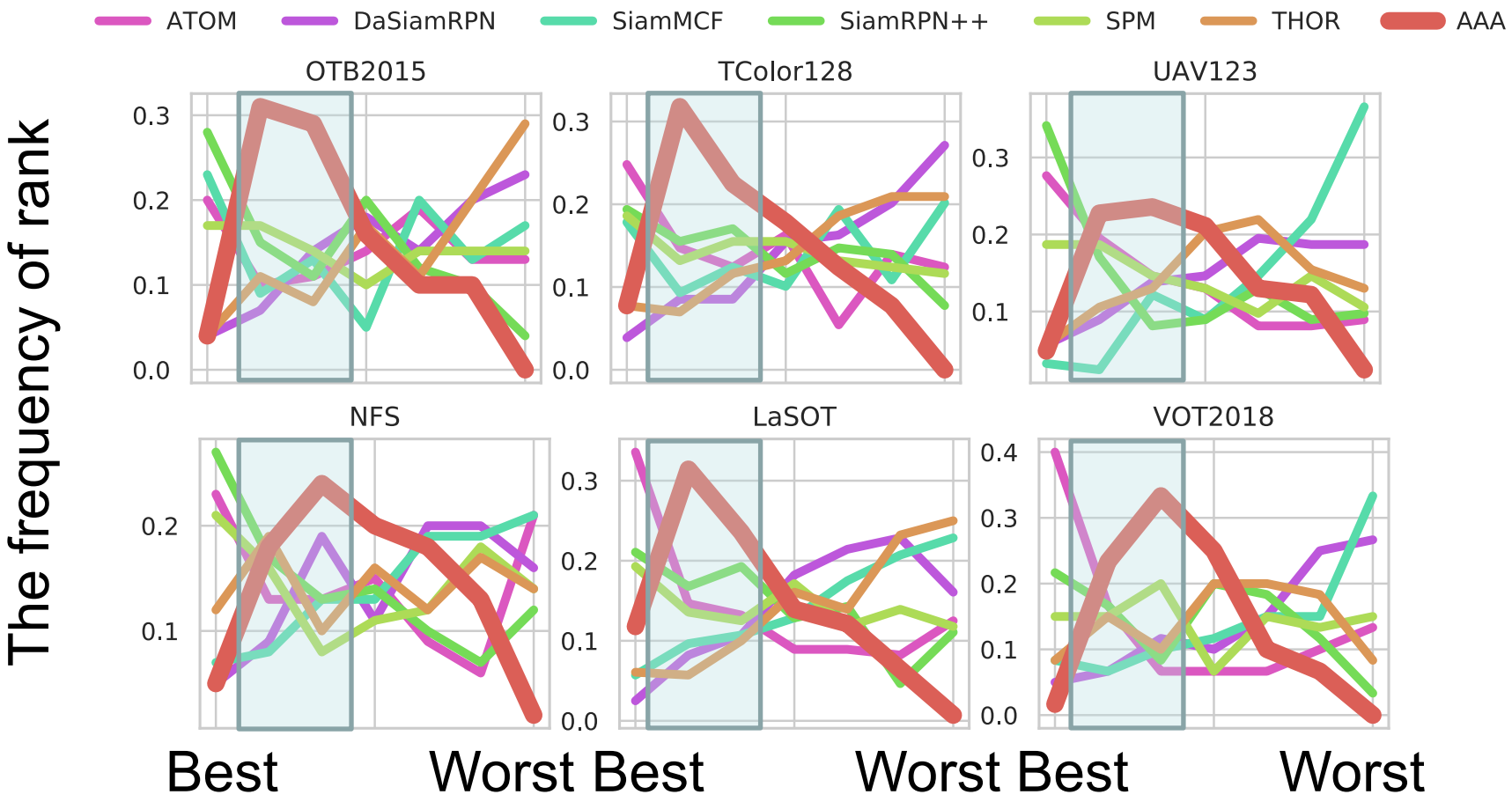
Comparision with High group

- No tracker is always the best, and the best tracker drastically changes over the videos.



Comparision with High group

- AAA achieved the second or third best performance for most videos.



Comparision with Low group

- Low group includes six low-performance trackers.
- AAA achieved a performance like the best tracker.

	OTB2015 [14]		TCo128 [15]		UAC123 [16]		NFS [17]		LaSOT [18]		VOT2018 [19]
	AUC	DP	AUC	DP	AUC	DP	AUC	DP	AUC	DP	AO
GradNet [8]	0.63	0.85	0.56	0.76	0.51	0.74	0.51	0.64	0.36	0.38	0.4
MemTrack [9]	0.63	0.82	0.54	0.74	0.49	0.7	0.5	0.61	0.34	0.35	0.39
SiamDW [10]	0.67	0.91	0.53	0.75	0.46	0.69	0.5	0.63	0.35	0.34	0.37
SiamFC [11]	0.59	0.78	0.52	0.7	0.51	0.74	0.51	0.6	0.35	0.36	0.33
SiamRPN [12]	0.63	0.83	0.52	0.71	0.58	0.77	0.56	0.66	0.45	0.45	0.48
Staple [13]	0.6	0.79	0.51	0.68	0.45	0.64	0.41	0.48	0.24	0.23	0.3
AAA	**0.66**	**0.87**	**0.59**	**0.82**	**0.56**	**0.78**	**0.58**	**0.69**	**0.45**	**0.46**	**0.45**

Aggregated tracekrs

2nd

1st

Multiple object tracking (MOT)

INTRODUCTION

What is the difference from SOT?

- We should track every pedestrian in a video.

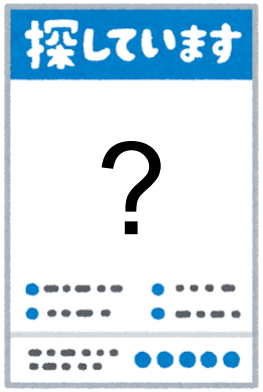
- Some pedestrians enter or leave in the video.

Frame 1 Frame 2 Frame 3

Can the same approach as SOT be used?

- It is difficult to define the anchor frame because there is no template image of the target object we need to track.

- Since the output of multiple object tracker is the location and unique number of the object (i.e., ID), the ID must be matched with each other when the selected tracker changes.



Multiple object tracking (MOT)

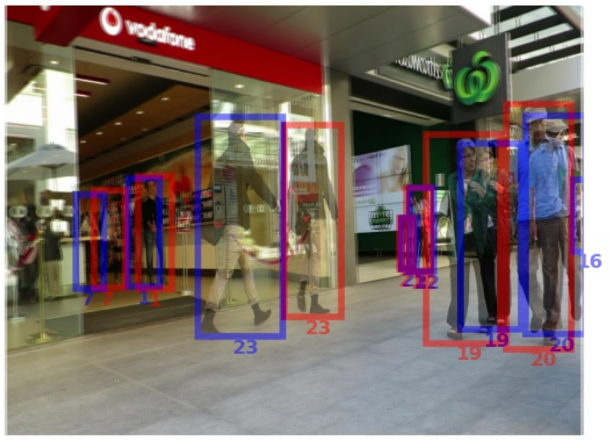
TRACKING WITH MULTIPLE ONLINE TRACKER

For the first step of aggregation

- I assume that anchor frames are at fixed time intervals, for example every 70 frames.

- Furthermore, it is assumed that a bounding box at a position similar to a bounding box in the previous frame is an object of the same ID.

Frame 1 ⋯ Frame 71 ⋯ Frame 141



Frame t-1 and t

Multiple object tracking (MOT)

EXPERIMENTS & RESULTS

Comparision with state-of-the-art trackers

- In a simple experiment, AAA achieved the performance similar to the best tracker.
- There were a lot of cases where IDs were wrongly predicted.

	MOT17[25]			
	FP ↓	FN ↓	IDs ↓	MOTA ↑
DAN[20]	13346	161362	4183	46.9%
DeepSort[21]	11784	161372	1451	48.2%
DeepT[22]	18974	170195	4381	42.5%
MOTDT[23]	8194	155780	1441	50.9%
Sort[24]	15465	167937	2681	44.8%
AAA	**10300**	**149601**	**8638**	**50.0%**

Aggregated tracekrs

1st

2nd

Future work

- As in SOT, I try to define anchor frames that allows offline tracker to achieve good performance in MOT.

- In order to reduce the case of incorrectly predicting ID, re-identification will be applied.

- I will conduct experiments with the large variations of benchmark and aggregated tracker in MOT to demonstrate that the proposed method can achieve state-of-the-art.

Reference

[1] Quanrud, Kent and Daniel Khashabi, “Online Learning with Adversarial Delays,” in Proc. NIPS, 2014.

[2] Danelljan, Martin, et al, "Atom: Accurate tracking by overlap maximization," in Proc. CVPR, 2019.

[3] Zhu, Zheng, et al. "Distractor-aware siamese networks for visual object tracking," in Proc. ECCV, 2018.

[4] Morimitsu, Henrique, "Multiple context features in Siamese networks for visual object tracking,” in Proc. ECCV, 2018.

[5] Li, Bo, et al, "Siamrpn++: Evolution of siamese visual tracking with very deep networks,” in Proc. CVPR, 2019.

[6] Wang, Guangting, et al, "Spm-tracker: Series-parallel matching for real-time visual object tracking," in Proc. CVPR, 2019.

[7] Sauer, Axel, Elie Aljalbout, and Sami Haddadin, "Tracking holistic object representations,” in Proc. BMVC, 2019.

[8] Li, Peixia, et al, "Gradnet: Gradient-guided network for visual object tracking,” in Proc. ICCV, 2019.

[9] Yang, Tianyu, and Antoni B. Chan, "Learning dynamic memory networks for object tracking,” in Proc. ECCV, 2018.

[10] Zhang, Zhipeng, and Houwen Peng, "Deeper and wider siamese networks for real-time visual tracking,” in Proc. CVPR, 2019.

[11] Bertinetto, Luca, et al, "Fully-convolutional siamese networks for object tracking.” in Proc. ECCV, 2016.

Reference

[12] Li, Bo, et al, "High performance visual tracking with siamese region proposal network," in Proc. CVPR, 2018.

[13] Bertinetto, Luca, et al, "Staple: Complementary learners for real-time tracking," in Proc. CVPR, 2016.

[14] Wu, Yi, Jongwoo Lim, and Ming-Hsuan Yang, "Online object tracking: A benchmark," in Proc. CVPR, 2013.

[15] Liang, Pengpeng, Erik Blasch, and Haibin Ling, "Encoding color information for visual tracking: Algorithms and benchmark," TIP 24.12 (2015): 5630-5644.

[16] Mueller, Matthias, Neil Smith, and Bernard Ghanem, "A benchmark and simulator for uav tracking," in Proc. ECCV, 2016.

[17] Kiani Galoogahi, Hamed, et al, "Need for speed: A benchmark for higher frame rate object tracking," in Proc. CVPR, 2017.

[18] Fan, Heng, et al, "Lasot: A high-quality benchmark for large-scale single object tracking," in Proc. CVPR, 2019.

[19] Kristan, Matej, et al, "A novel performance evaluation methodology for single-target trackers," TPAMI, 38.11 (2016): 2137-2155.

[20] Sun, ShiJie, et al, "Deep affinity network for multiple object tracking," TPAMI, (2019).

[21] Wojke, Nicolai, and Alex Bewley, "Deep cosine metric learning for person re-identification," in Proc. WACV, 2018.

[22] Yoon, Young-chul, et al, "Online multi-object tracking with historical appearance matching and scene adaptive detection filtering," in Proc. AVSS, 2018.

Reference

[23] Chen, Long, et al, “Real-Time Multiple People Tracking with Deeply Learned Candidate Selection and Person Re-Identification,” in Proc. ICME, 2018.

[24] Bewley, Alex, et al, "Simple online and realtime tracking,” in Proc. ICIP, 2016.

[25] Milan, Anton, et al, "MOT16: A benchmark for multi-object tracking," *arXiv preprint arXiv:1603.00831* (2016).

[26] Zhang, Li, Yuan Li, and Ramakant Nevatia, "Global data association for multi-object tracking using network flows," in Proc. CVPR, 2008

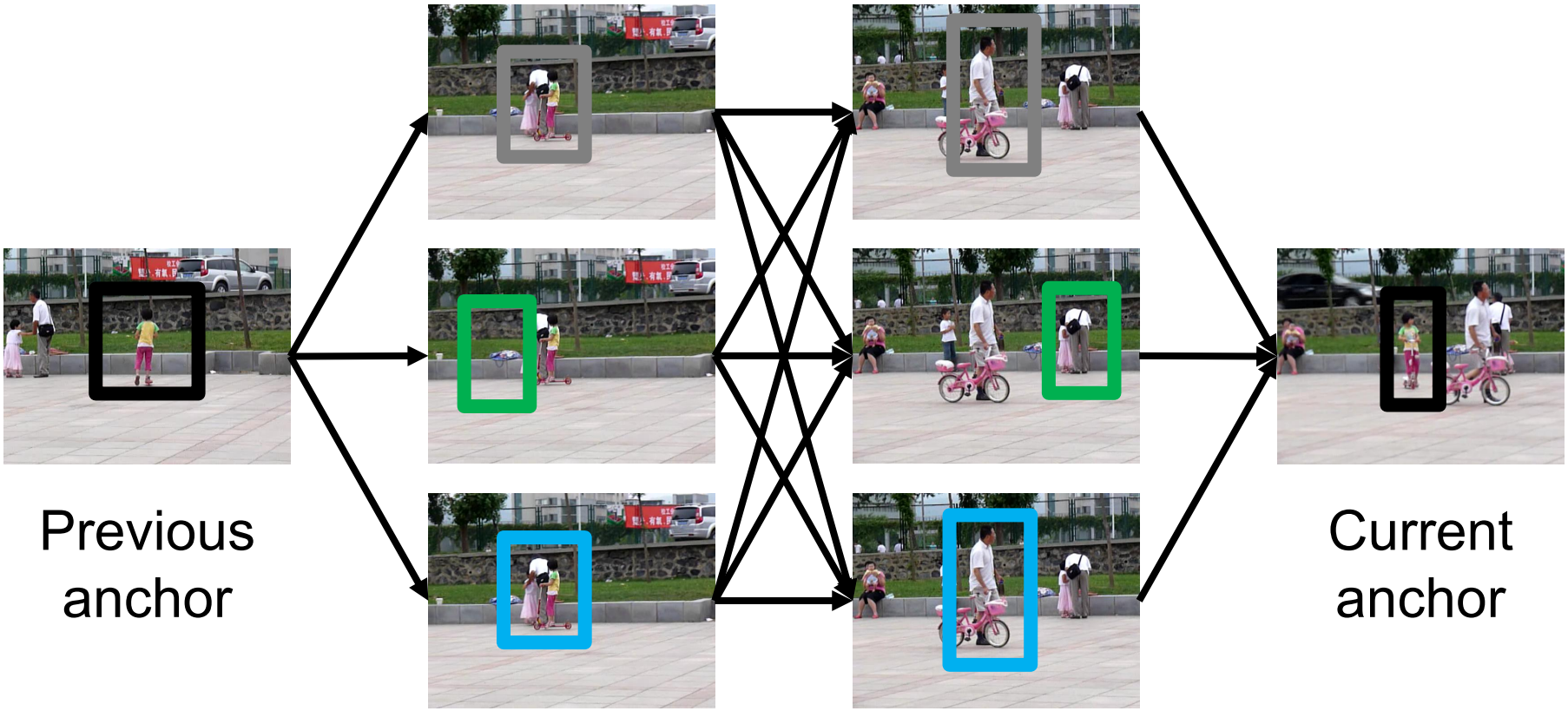
[27] Brasó, Guillem, and Laura Leal-Taixé, "Learning a neural solver for multiple object tracking.”in Proc. CVPR, 2020.

[28] He, Kaiming, et al, "Deep residual learning for image recognition," in Proc. CVPR, 2016.

APPENDIX

How to track the target offline in VOT

- Create a graph with each tracker's estimation as a node as below and find minimum-cost flow[26].



[26] Zhang, Li, Yuan Li, and Ramakant Nevatia, "Global data association for multi-object tracking using network flows," in Proc. CVPR, 2008

How to track the target offline in MOT

- Fortunately, various offline tracking methods have been proposed for MOT.

- In this work, we used state-of-the-art offline multiple object tracker, named MPN-tracker[27].

[27] Brasó, Guillem, and Laura Leal-Taixé, "Learning a neural solver for multiple object tracking.”in Proc. CVPR, 2020.

How to detect anchor frame in VOT

- Extract a target feature vector from the target image which is cropped at the first frame by ResNet[28].
- In the same way, feature vectors are extracted from bounding boxes predicted by trackers every frame.
- If the cosine similarity between the target feature vector and each extracted feature vector by trackers is greater than a threshold, the frame is considered to be an anchor frame.

[28] He, Kaiming, et al, "Deep residual learning for image recognition," in Proc. CVPR, 2016.

How to update the reliability

The offline tracker gives the bounding box of the target object $y^{u+1}, \ldots, y^{t}$ between the previous anchor frame $u+1$ to current anchor frame t. Using the near true label, the loss of tracker i is calculated using the following equation:

$$L_i = \sum_{\tau=u+1}^{t} 1 - \mathrm{IoU}(f_i^{\tau}, y^{\tau}),$$

where f_i^{τ} is the i-th tracker's estimation of the target location at frame τ.

How to update the reliability

Based on this loss, the reliability of tracker i is updated according to following equation:

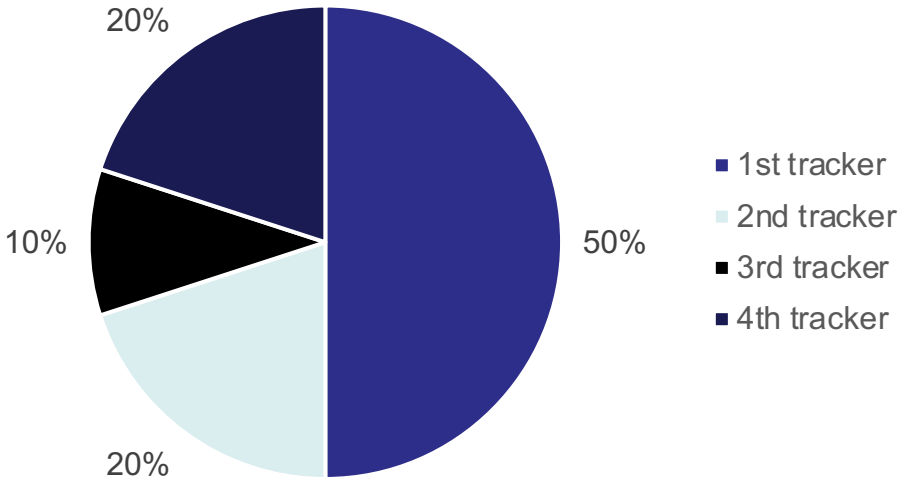
$$w_i^{t+1} = \frac{w_i^t \exp(-\eta L_i)}{\sum_{j=1}^{N} w_j^t \exp\left(-\eta L_j\right)},$$

where η is a learning rate and N is the number of aggregated trackers.

How to select one estimation

- For selecting one of experts' estimation, we use roulette wheel selection, known as fitness proportionate selection, according to their weight.

- Each tracker’s weight represents the probability that the tracker’s estimation will be selected.



Tracking example in VOT

- AAA can follow the best tracker by adaptively aggregating the trackers.