# 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

Download the presentation file here!

1. Single object tracking (SOT)
    1. Introduction
    2. Tracking with multiple online tracker
    3. Experiments & results

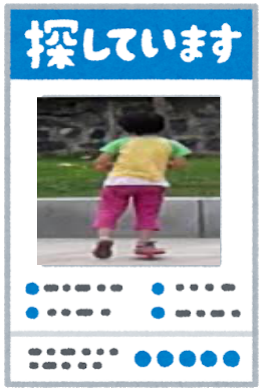
2. Multiple object tracking (MOT)
    1. Introduction
    2. Tracking with multiple online tracker for MOT
    3. Experiments & results

Single object tracking (SOT)

# INTRODUCTION

# What is online Single Object Tracking (SOT)?

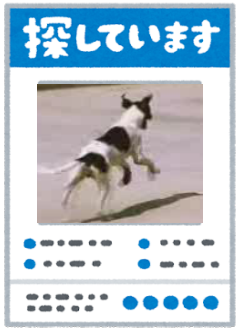
- Given a target location at the first frame,
  we should track the target and predict the location.




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

# Use multiple online trackers for robust tracking

I track animals well

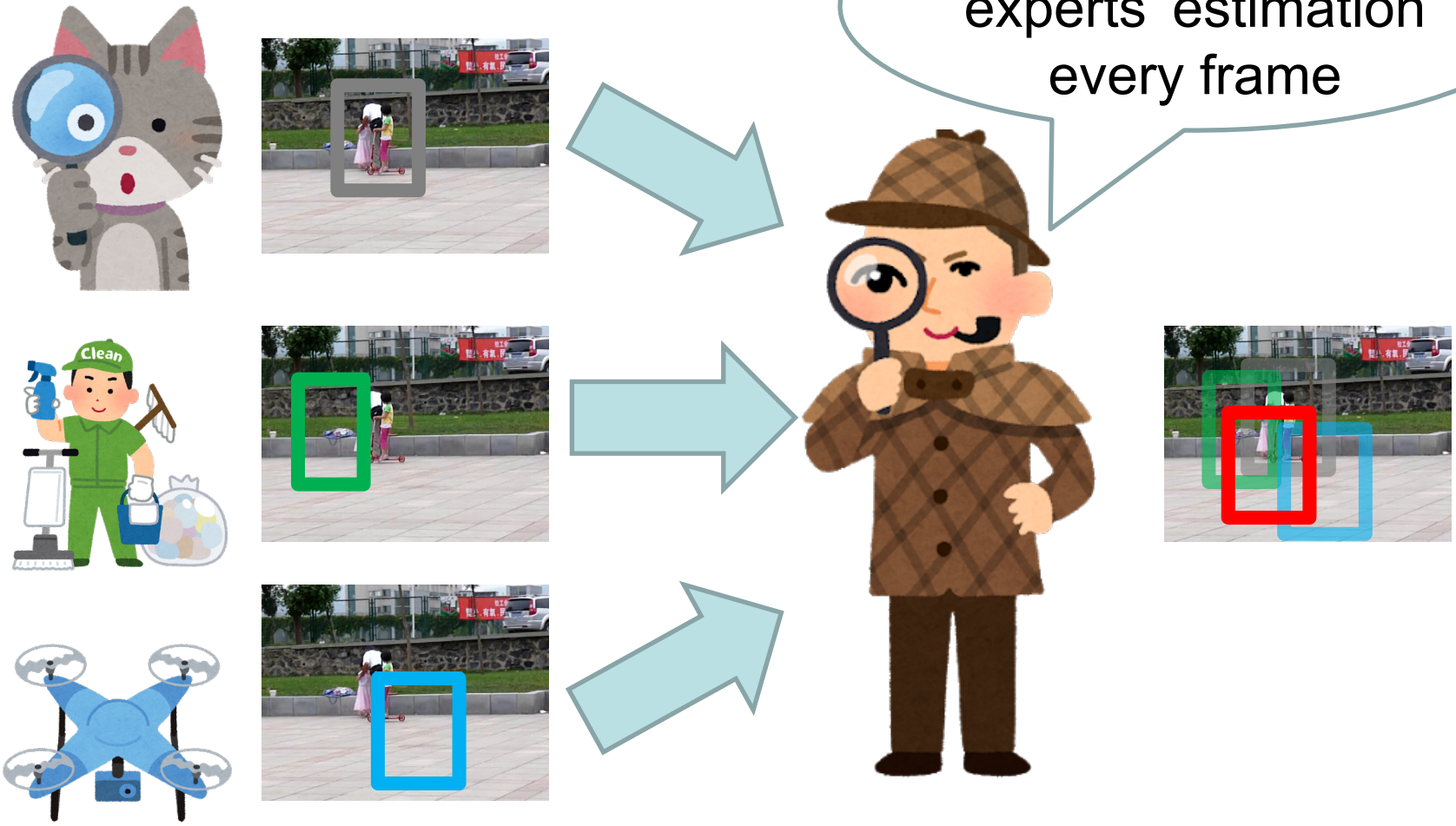

I track plastics well

I track occluded object well

We call them “experts”

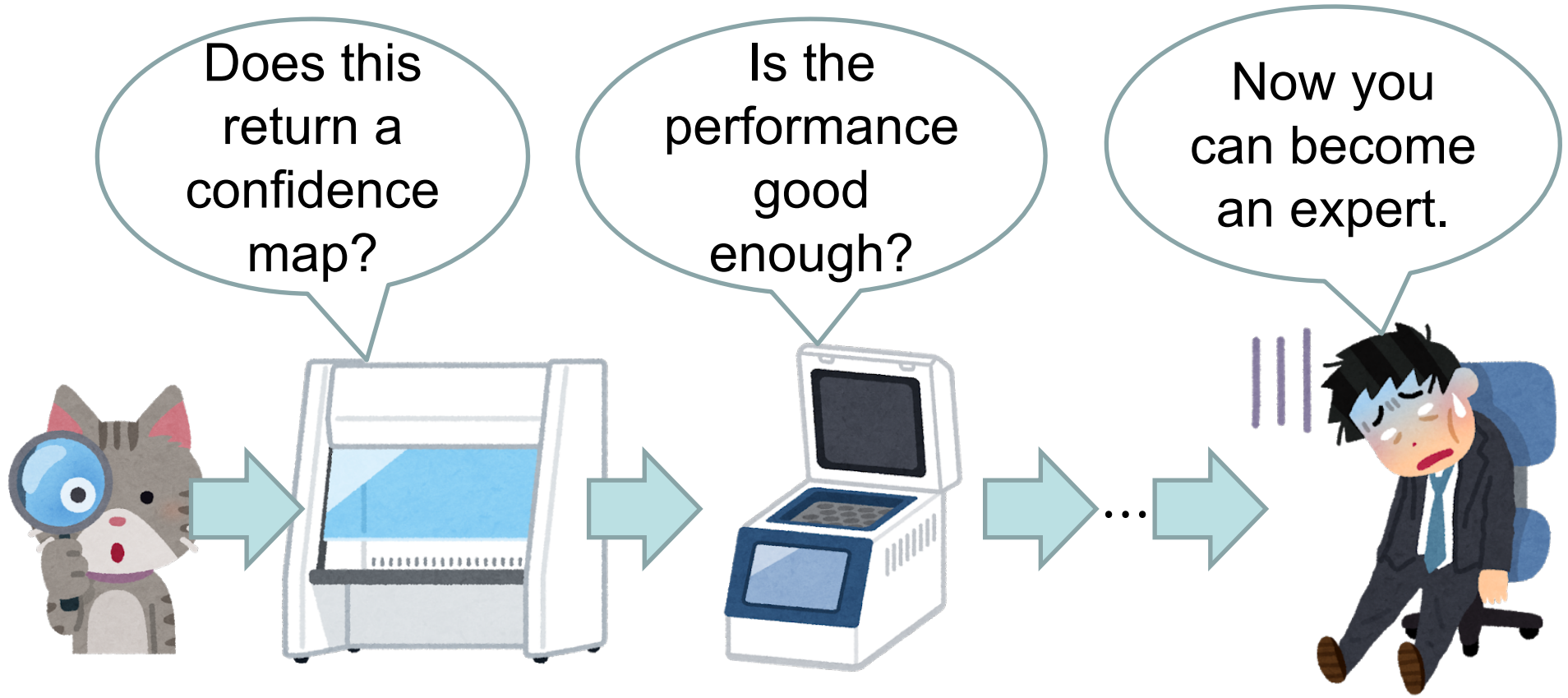
# Use multiple online trackers for robust tracking

For example…



# Problems with using experts

- Most algorithms using experts require a lot of restrictions to use online trackers as experts.



# In our work

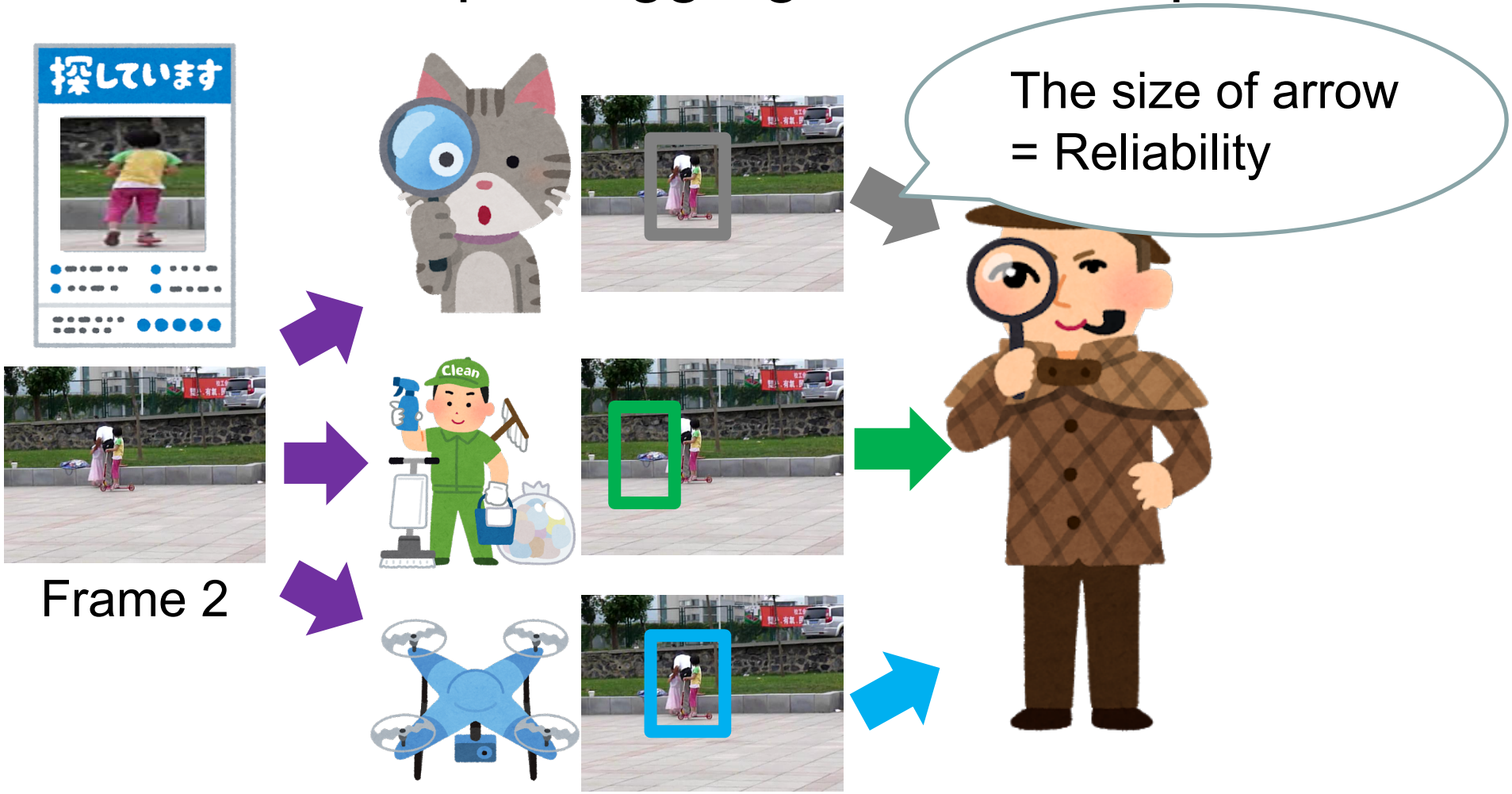
- The proposed method can aggregate arbitrary trackers.

- Moreover, The performance of the proposed method is theoretically guaranteed.

Single object tracking (SOT)

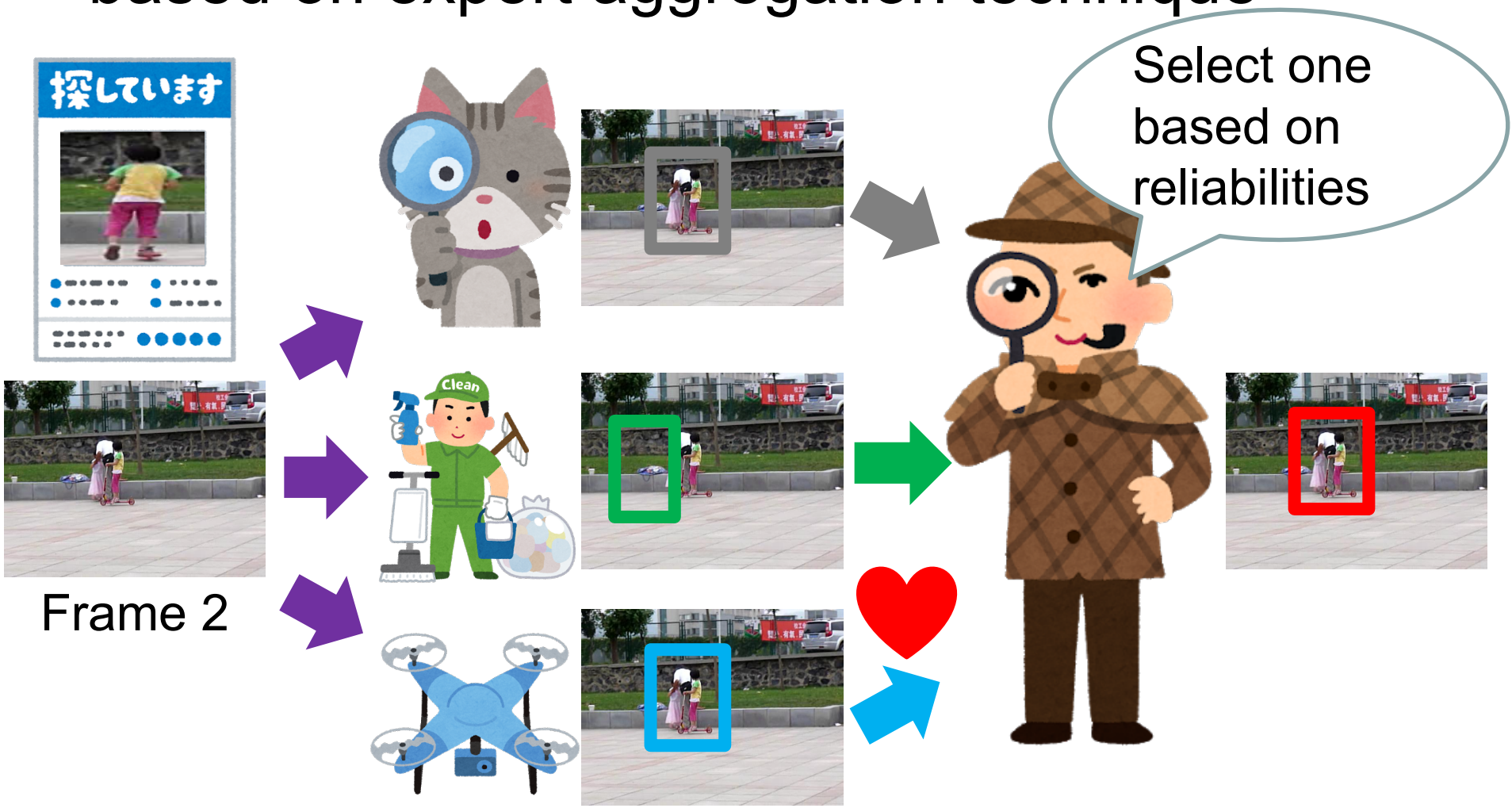
# TRACKING WITH MULTIPLE ONLINE TRACKER

# Aggregate multiple online trackers based on expert aggregation technique[1]



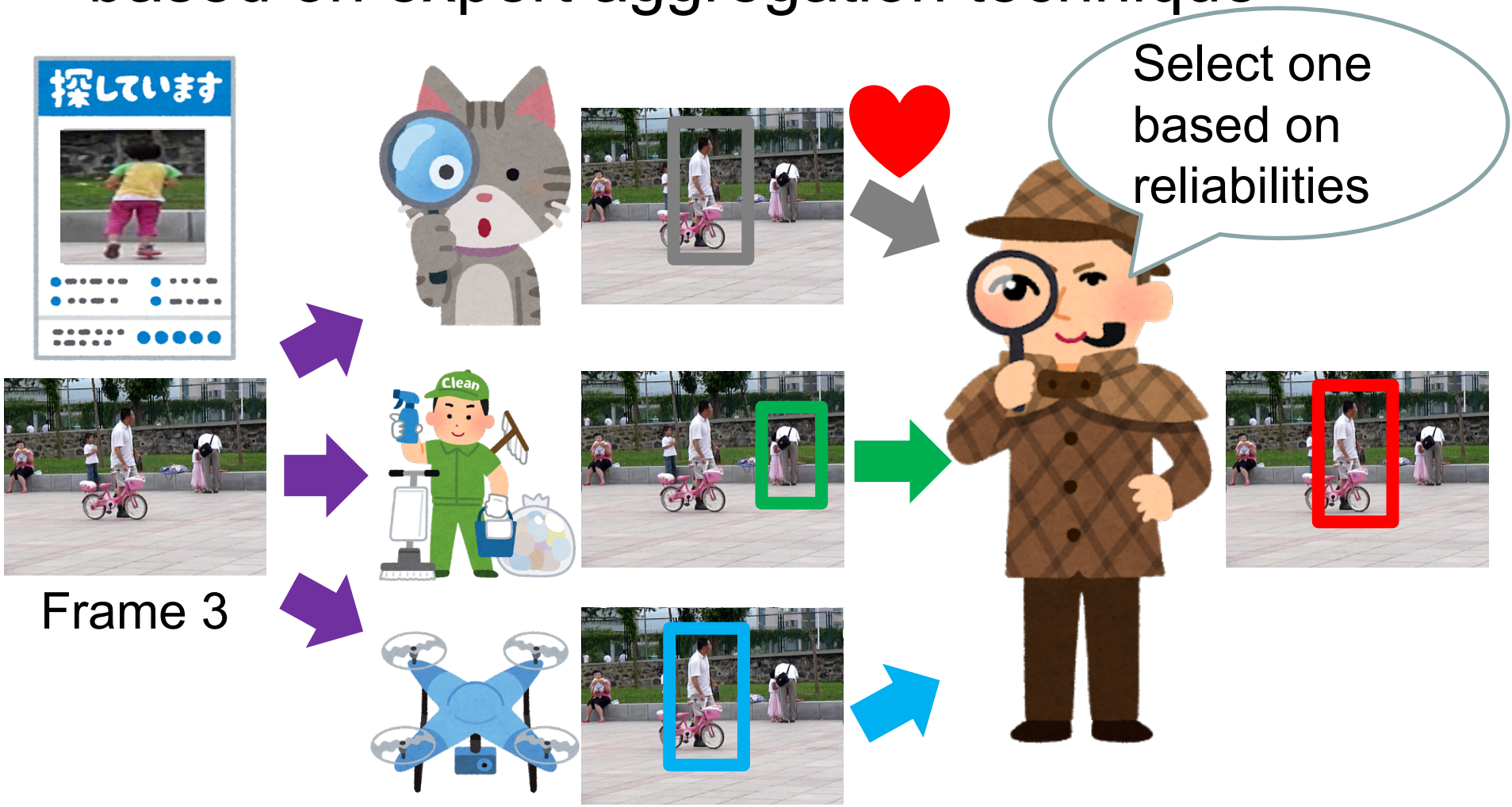

[1] Quanrud, Kent and Daniel Khashabi, “Online Learning with Adversarial Delays,” in Proc. NIPS, 2014.

# Aggregate multiple online trackers based on expert aggregation technique[1]



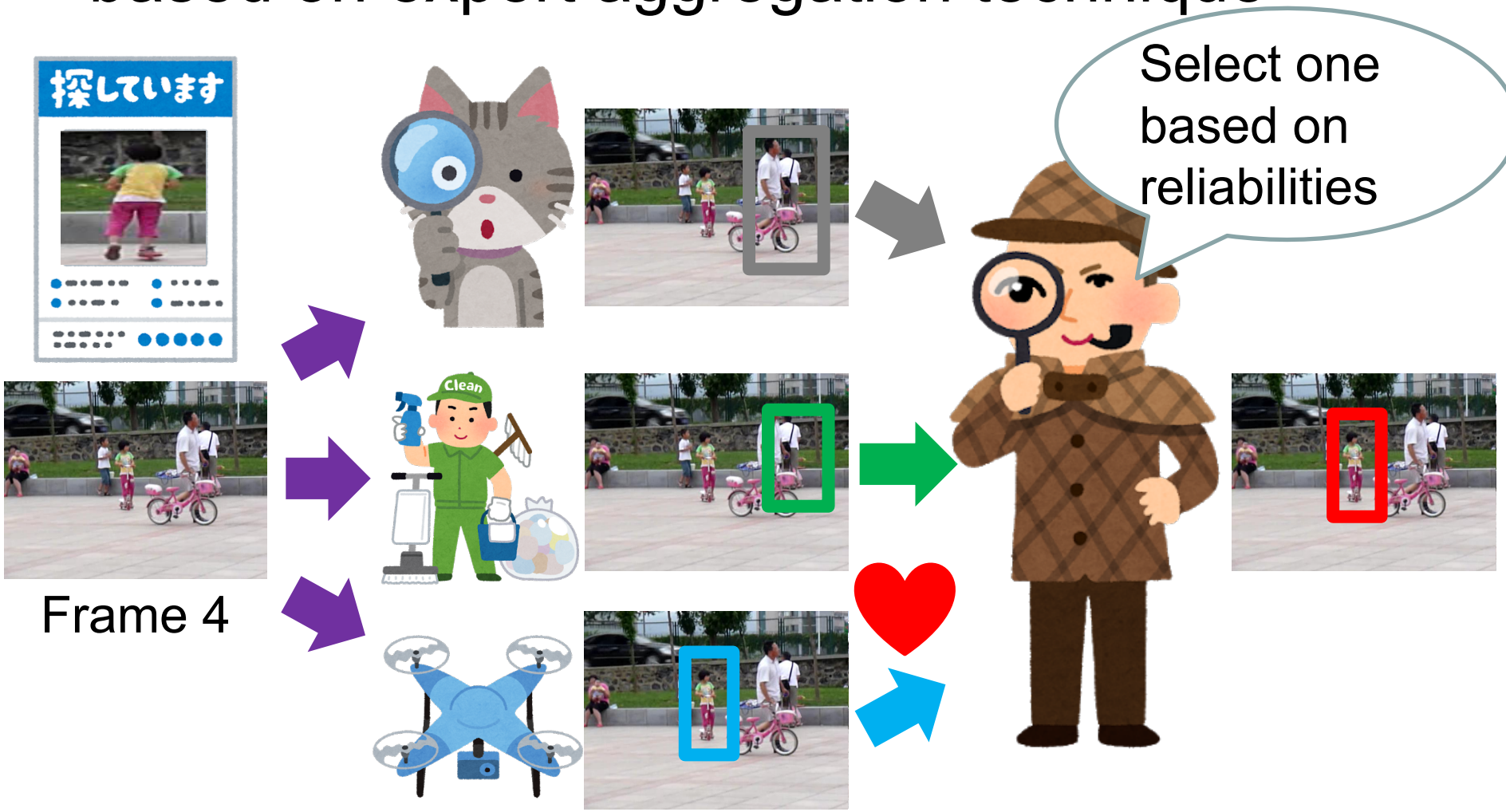

[1] Quanrud, Kent and Daniel Khashabi, “Online Learning with Adversarial Delays,” in Proc. NIPS, 2014.

# Aggregate multiple online trackers based on expert aggregation technique[1]




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

# Aggregate multiple online trackers based on expert aggregation technique[1]




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

# Aggregate multiple online trackers based on expert aggregation technique[1]


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

# Aggregate multiple online trackers based on expert aggregation technique[1]


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

# How can we get feedback?


Where is she?


Frame 1 ⇨ Frame 2 ⇨ Frame 3 ⇨

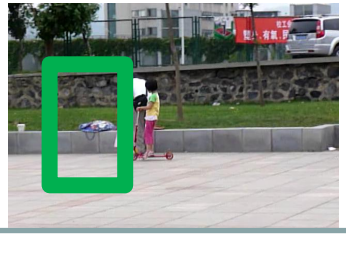
# How can we get feedback?










Here she is!

Anchor frame =
When the target location can be predicted with high confidence

Frame 1 ⇨ Frame 2 ⇨ Frame 3 ⇨ Frame 4

# How can we get feedback?

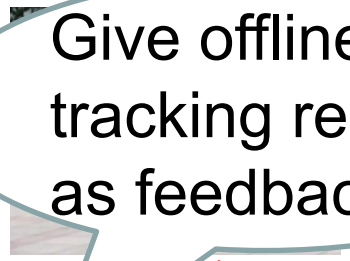




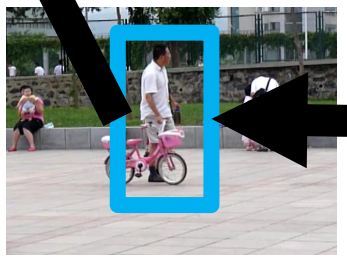








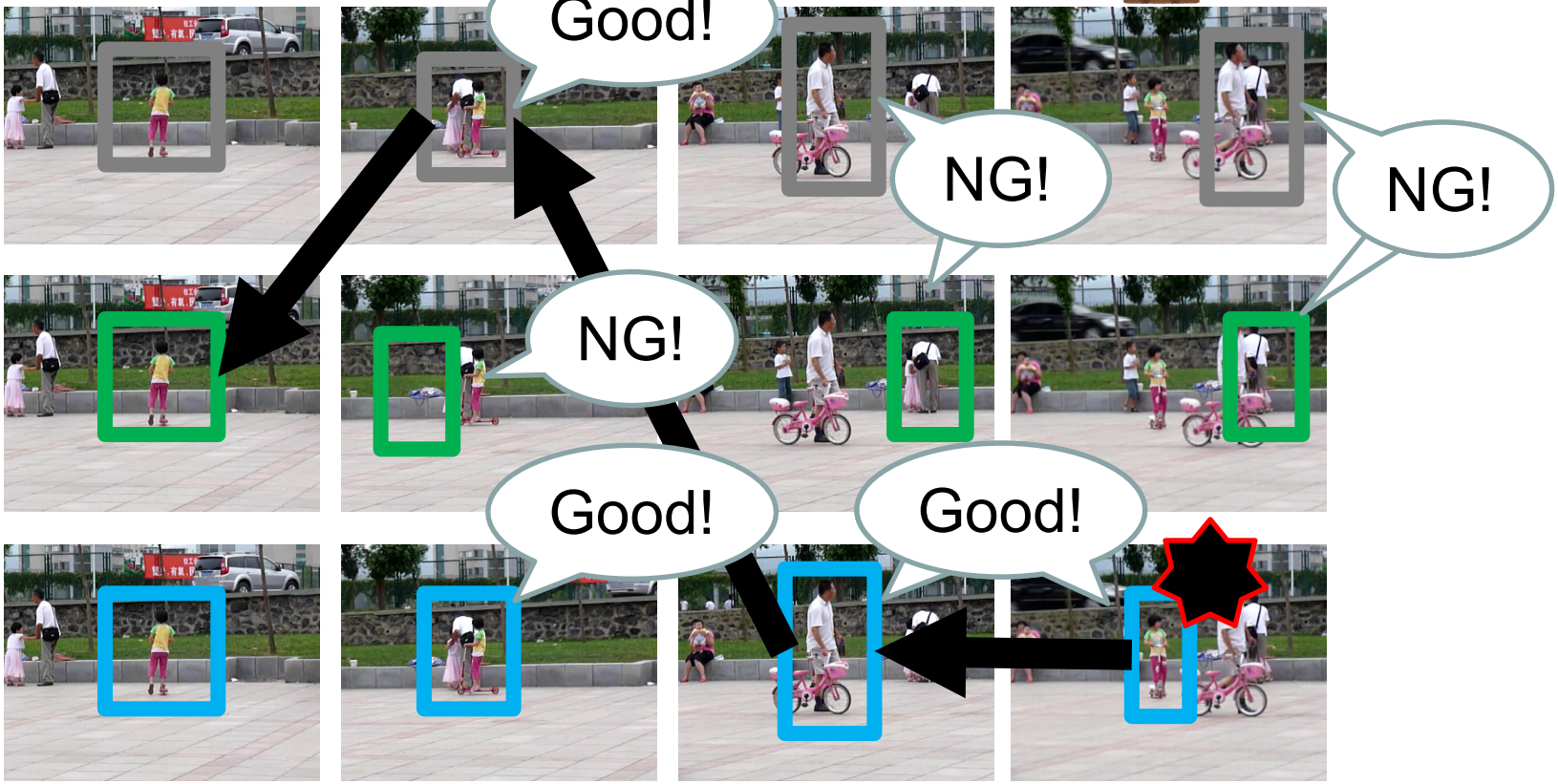

Give offline tracking results as feedback

Anchor frame

Anchor frame

Frame 1 ⇨ Frame 2 ⇨ Frame 3 ⇨ Frame 4

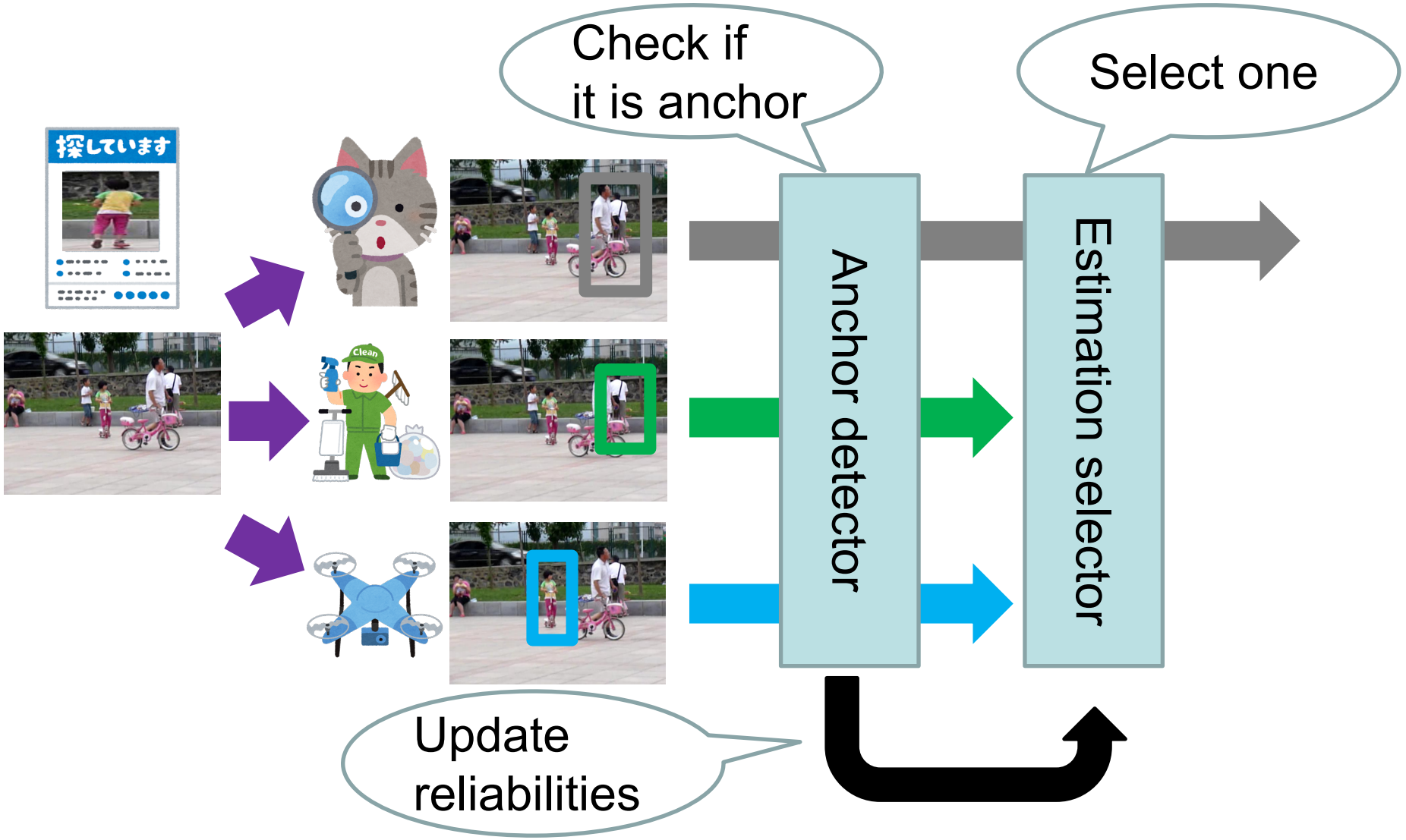
# How can we get feedback?



Frame 1 ⇨ Frame 2 ⇨ Frame 3 ⇨ Frame 4

# 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 expert, 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 experts

Single object tracking (SOT)

# EXPERIMENTS & RESULTS

# Experts and Benchmarks

- I would like to demonstrate that the proposed method is robust for arbitrary videos and experts.
- For arbitrary video, I evaluated the proposed method on six different benchmarks.
- For arbitrary experts, I prepared twelve experts and divide into 2 groups according to their performance.

# Comparision with High group

- High group included six high-performance experts.
- Ours outperformed the experts 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 |
| **Ours** | **0.7** | **0.91** | **0.62** | **0.84** | **0.62** | **0.83** | **0.61** | **0.75** | **0.53** | **0.55** | **0.52** |

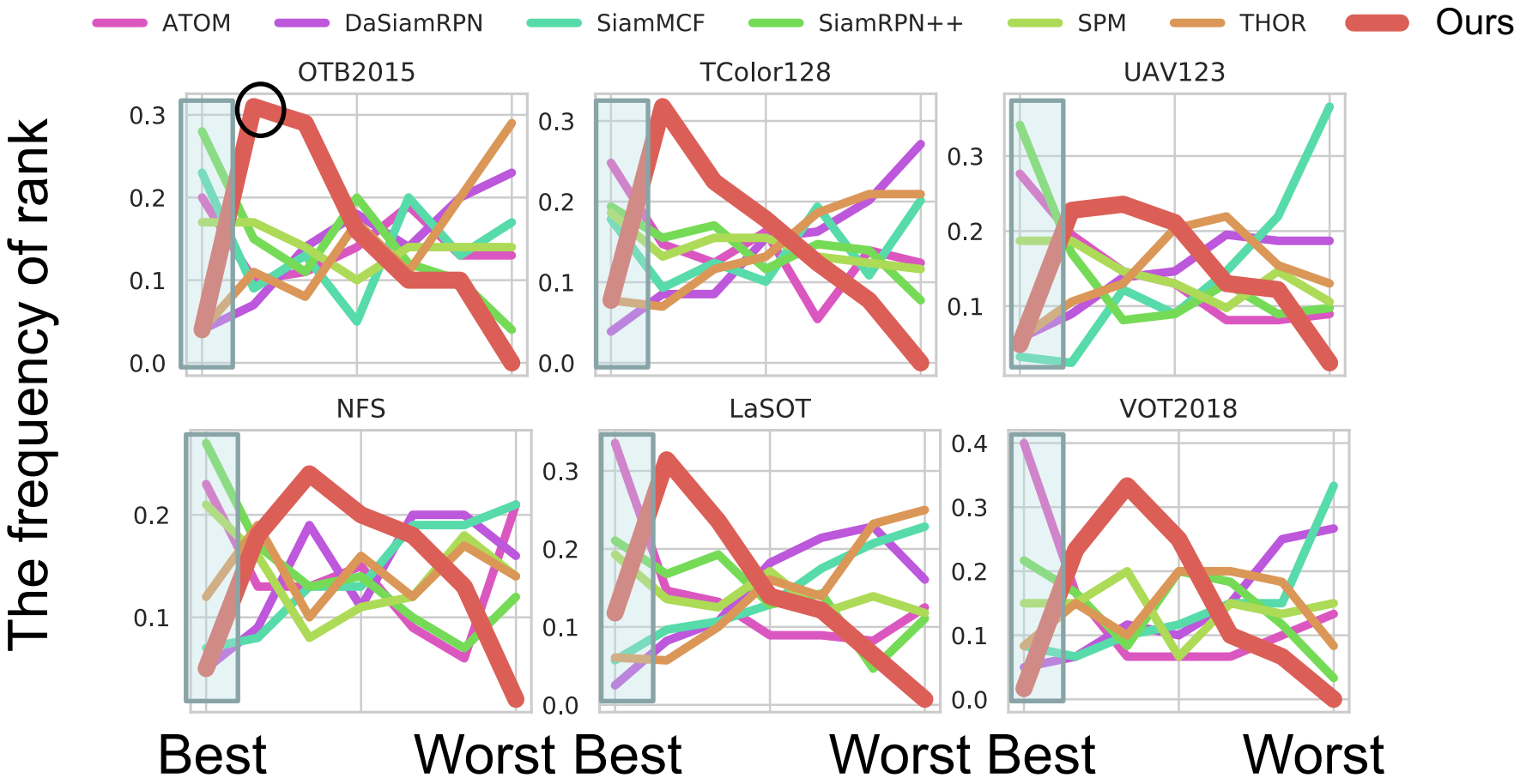
Experts

2nd

1st

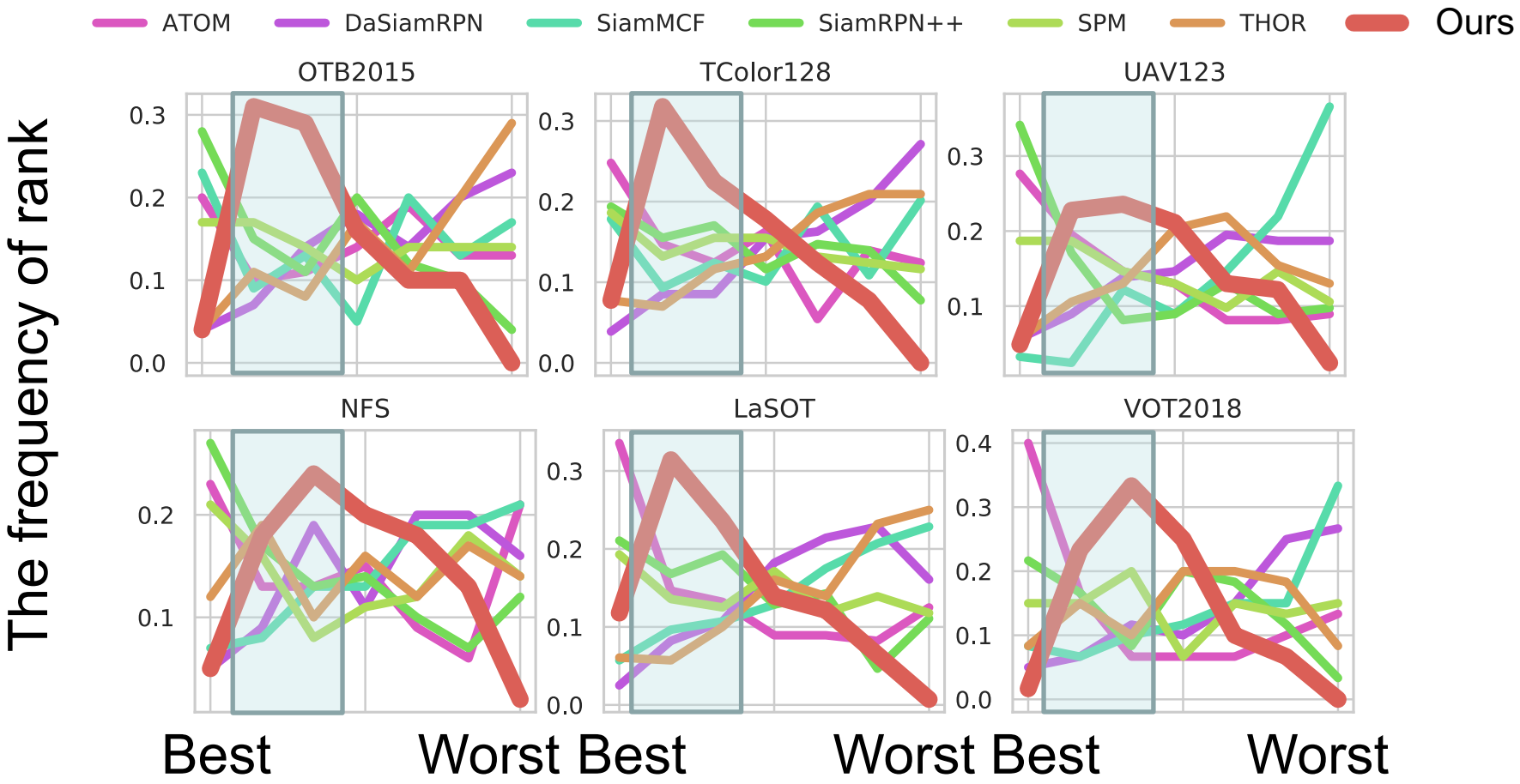
# Comparision with High group

- No tracker was always the best, and the best expert drastically changed over the videos.



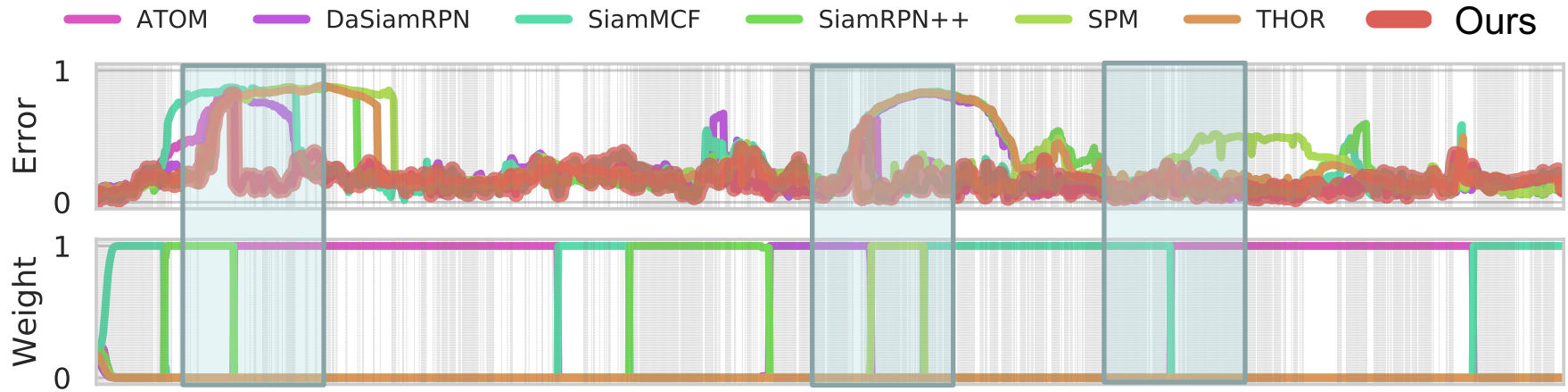
# Comparision with High group

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

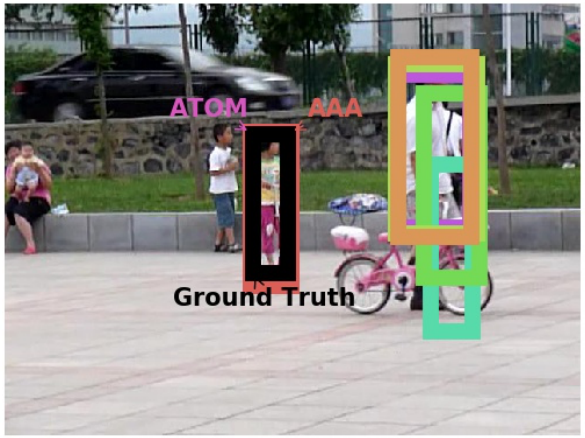


# Tracking example in SOT

- Ours(AAA) can follow the best expert by adaptively aggregating the trackers.








# Comparision with Low group

- Low group included six low-performance experts.
- Ours achieved at least the second best performance for all benchmarks.

| | 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 |
| **Ours** | **0.66** | **0.87** | **0.59** | **0.82** | **0.56** | **0.78** | **0.58** | **0.69** | **0.45** | **0.46** | **0.45** |

Experts

1st

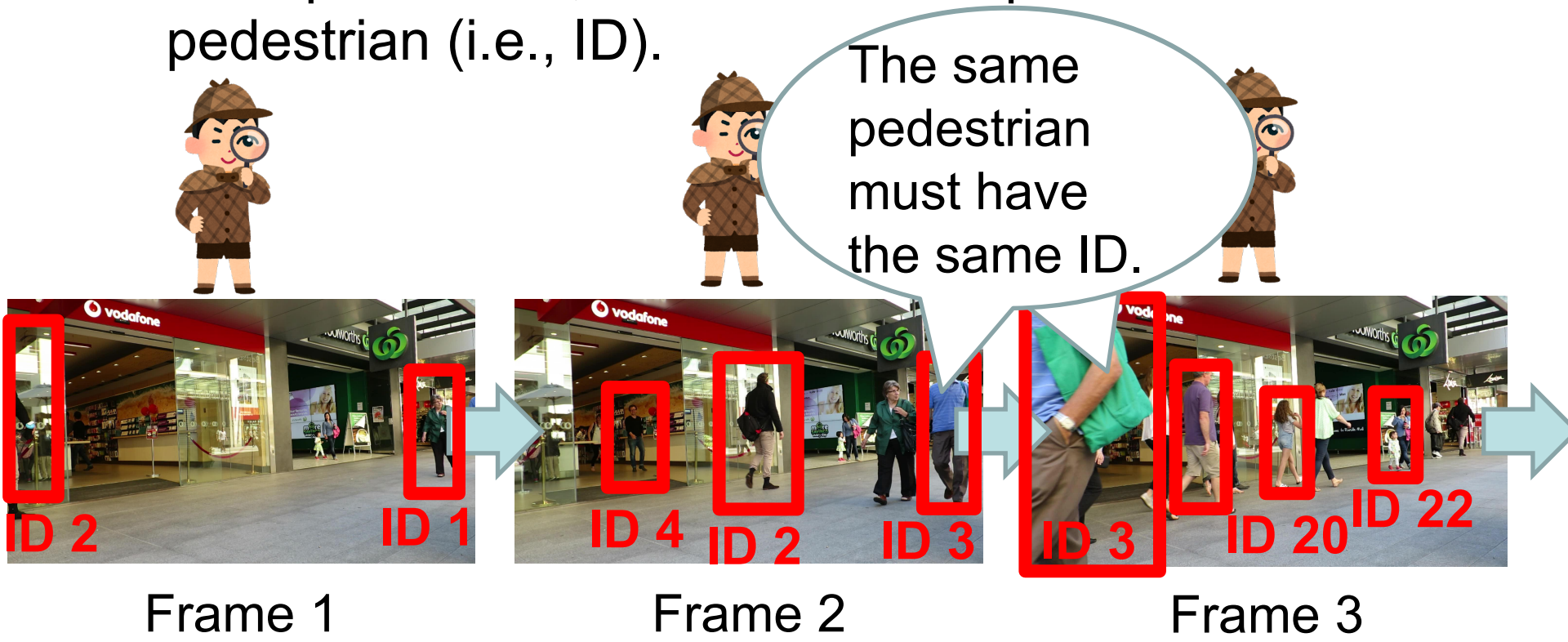
2nd

Multiple object tracking (MOT)

# INTRODUCTION

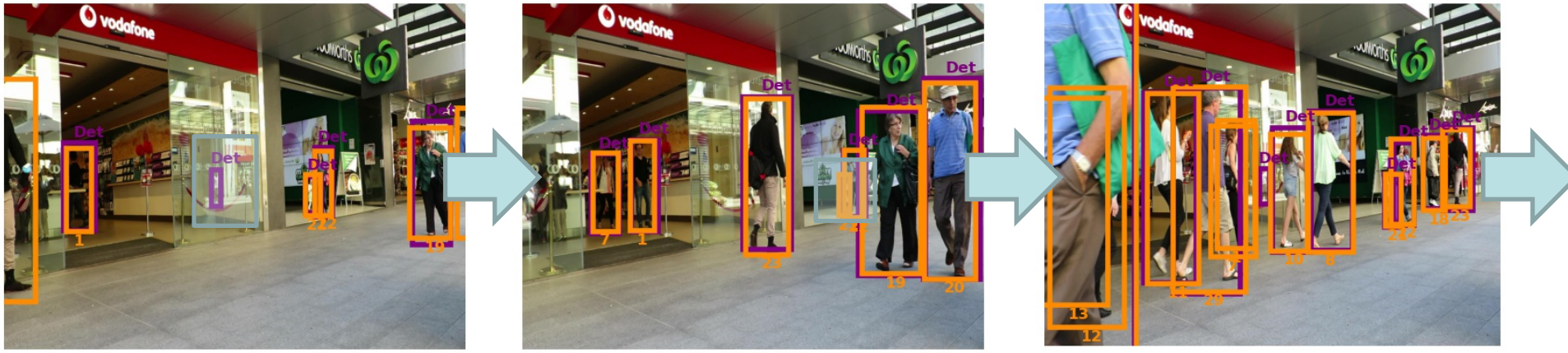
# What is online Multiple Object Tracking (MOT)?

- We should track every pedestrian in a video from the first frame.
- It is necessary to predict not only the location of each pedestrian, but also the unique number of the pedestrian (i.e., ID).



# What is the difference from SOT?

- The initial location of pedestrians is not given. Instead, human detection results using some human detectors are given.
- Some pedestrians enter or leave the scene.

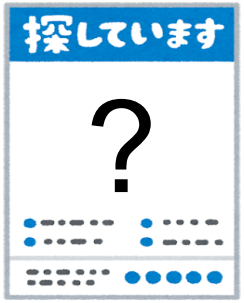



Frame 1 Frame 2 Frame 3

Orange: ground truth, purple: human detection results

# Can the same approach as SOT be used?

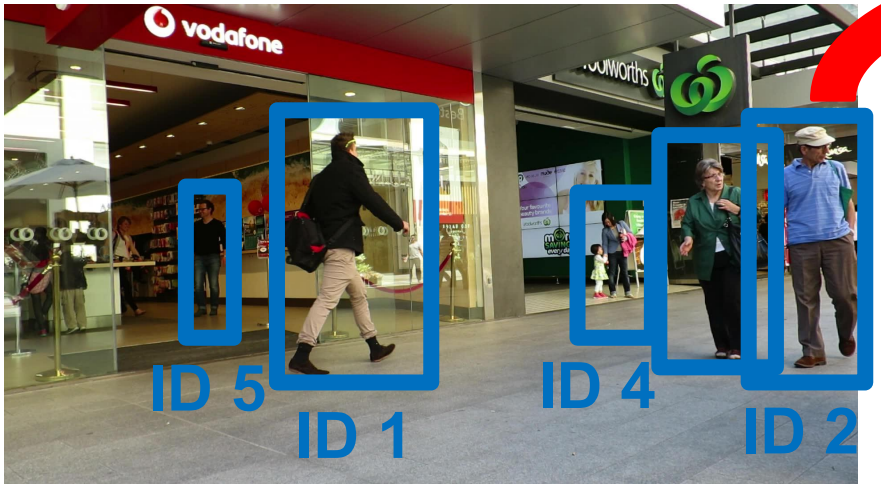
- The definition of the anchor frame should be changed because it is difficult to predict the location of “all pedestrians” with high confidence.





# Can the same approach as SOT be used?

- Since the output of experts is the location and ID, the ID must be matched with each other when the selected tracker changes.




Frame $t-1$
Tracker 1 is selected


Frame $t$
Tracker 2 is selected

Multiple object tracking (MOT)

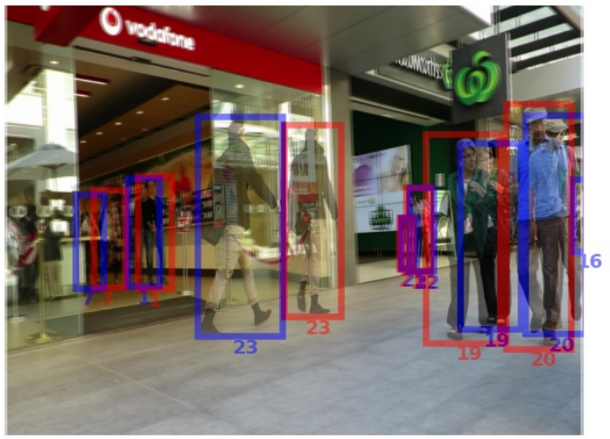
# TRACKING WITH MULTIPLE ONLINE TRACKER

# For the first step step to expansion for MOT

- I assume that every 70 frames are anchor 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

# What I try to reduce

- The main evaluation criterion is MOTA, which consists of FP, FN, and IDs.

- Since FN is very large in most data, FN is evaluated as the loss of expert.

| Sequence | MOTA | IDF1 | MOTP | MT | ML | FP | FN | Recall | Precision | FAF | ID Sw. | Frag |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MOT17-01-DPM | 41.1 | 47.9 | 76.8 | 7 | 9 | 529 | 3,245 | 49.7 | 85.8 | 1.2 | 24 | 43 |
| MOT17-01-FRCNN | 40.8 | 49.2 | 76.7 | 8 | 9 | 625 | 3,166 | 50.9 | 84.0 | 1.4 | 27 | 47 |
| MOT17-01-SDP | 41.5 | 47.8 | 76.3 | 9 | 8 | 668 | 3,075 | 52.3 | 83.5 | 1.5 | 31 | 55 |
| MOT17-03-DPM | 79.3 | 69.0 | 78.6 | 92 | 11 | 1,193 | 20,320 | 80.6 | 98.6 | 0.8 | 178 | 471 |
| MOT17-03-FRCNN | 79.9 | 69.6 | 78.6 | 90 | 9 | 1,278 | 19,555 | 81.3 | 98.5 | 0.9 | 179 | 465 |
| MOT17-03-SDP | 80.4 | 69.8 | 78.5 | 92 | 10 | 1,351 | 18,959 | 81.9 | 98.4 | 0.9 | 181 | 473 |
| MOT17-06-DPM | 53.0 | 40.8 | 77.4 | 58 | 72 | 632 | 4,811 | 59.2 | 91.7 | 0.5 | 97 | 203 |

**MOTChallenge** https://motchallenge.net/method/MOT=3246&chl=10

# Comparision with state-of-the-art trackers

- In a simple experiment, Ours 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% |
| **Ours** | **10300** | **149601** | **8638** | **50.0%** |

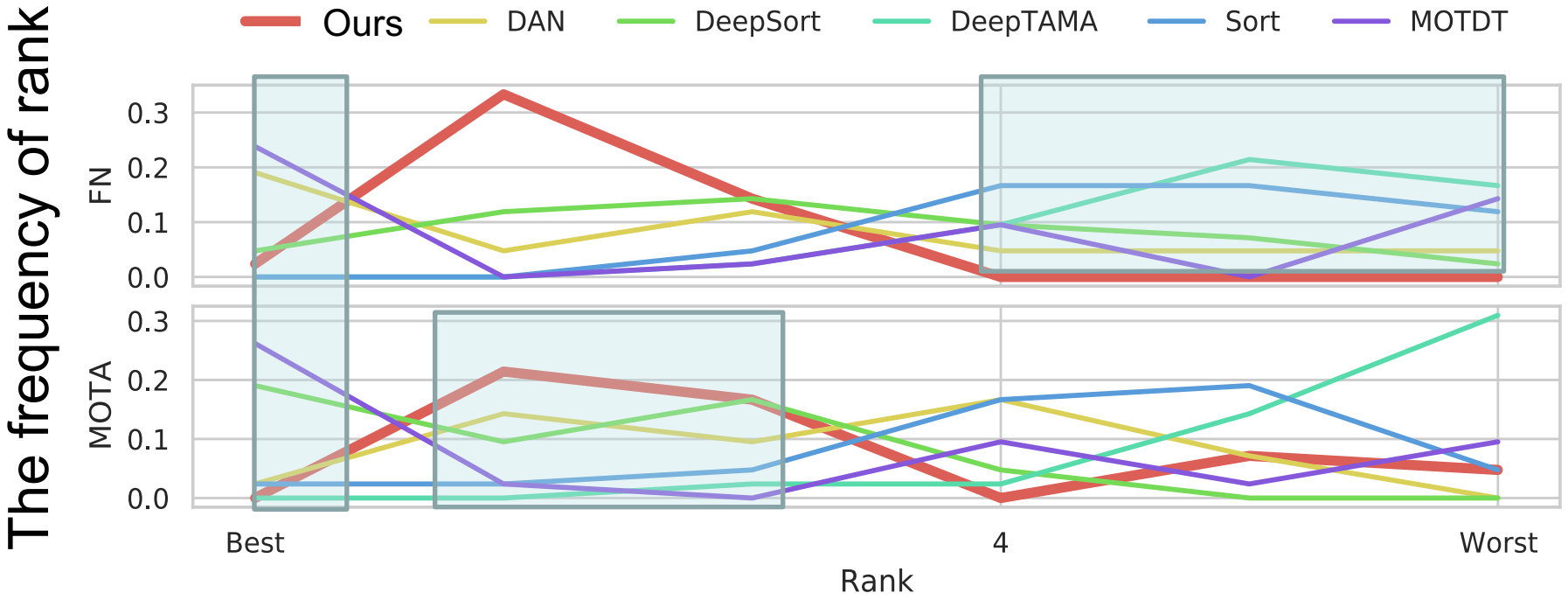
Experts

1st

2nd

# Comparision with state-of-the-art trackers

- Similar to the SOT tasks, no tracker was the best.
- Ours achieved at least third lowest FN for all videos.
- Good performance was also obtained for MOTA.



# Future work

- I try to define anchor frames that allows the proposed method to achieve good performance for MOT.

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

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

# 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.

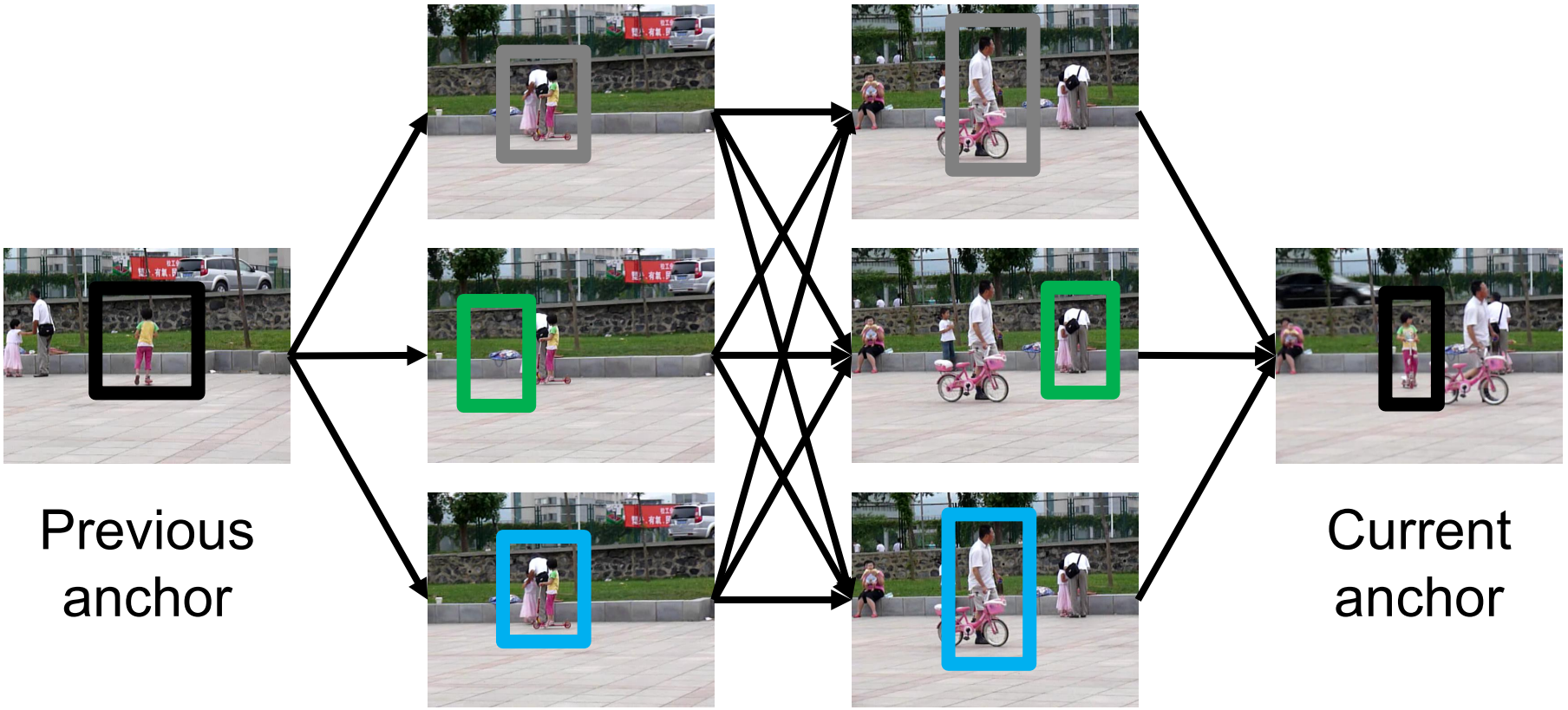
[29] Bernardin, Keni, and Rainer Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing* 2008 (2008): 1-10.

[30] Wang, Ning, et al, "Multi-cue correlation filters for robust visual tracking,” in Proc. CVPR, 2018.

# APPENDIX

# How to track the target offline in SOT

- 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 SOT

- 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 calculate the loss of experts in SOT

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 offline tracking results, the loss of tracker $i$ is calculated using the following equation:

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

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

# How to calculate the loss of experts in MOT

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 $t$. Based on the offline tracking results, the loss of tracker $i$ is calculated using the following equation:

$$L_i = \sum_{\tau=u+1}^{t} \text{false negative of expert } i \text{ at frame } \tau,$$

where the false negative of the expert is calculated by CLEAR-MOT[29].

[29] Bernardin, Keni, and Rainer Stiefelhagen, "Evaluating multiple object tracking performance: the CLEAR MOT metrics," *EURASIP Journal on Image and Video Processing* 2008 (2008): 1-10.

# Why we use the FN as the loss of tracker

- MOTA can be calculated as follows:

$$MOTA = 1 - \frac{\sum_t FP + FN + IDs}{\sum_t GT}$$

- However, for the MOT benchmark, the FN of all experts is far greater than the FP and IDs.

- In addition, we have to make sure that experts with small cumulative losses have a high MOTA.

- Therefore, we used the cumulative sum of the expert's FNs as the expert's loss.

# How to update the reliability

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

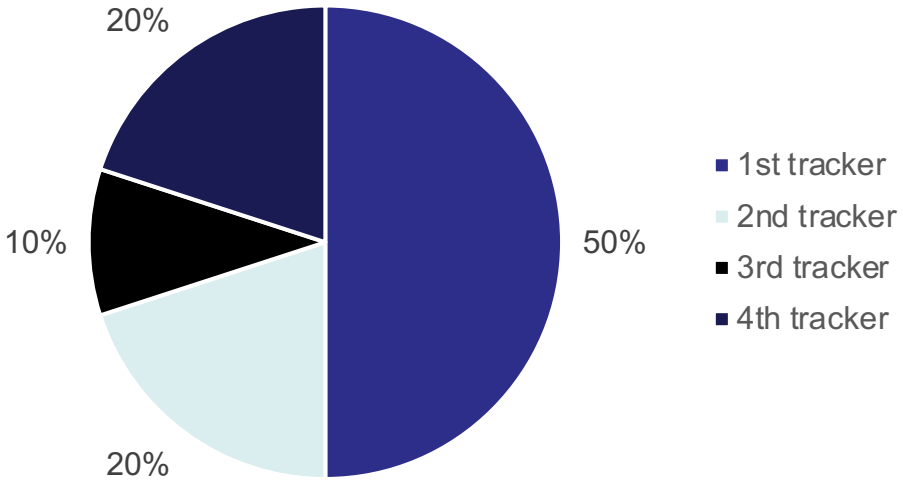
$$w_i^{t+1} = \frac{w_i^t \exp(-\eta L_i)}{\sum_{j=1}^{N} w_j^t \exp\left(-\eta L_j\right)},$$

where $\eta$ 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.



# Comparision with High group

- I compared the proposed method with other aggregation-based trackers.
- In other trackers, performance was lower than most experts.

| Tracker | OTB2015 | | TColor128 | | UAV123 | | NFS | | LaSOT | | VOT2018 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AO |
| ATOM | 0.67 | 0.87 | *0.60* | *0.81* | **0.62** | *0.82* | 0.58 | 0.69 | *0.51* | *0.51* | *0.52* |
| DaSiamRPN | 0.65 | 0.88 | 0.53 | 0.75 | 0.57 | 0.78 | 0.55 | 0.67 | 0.43 | 0.42 | 0.43 |
| SiamMCF | 0.65 | 0.85 | 0.57 | 0.78 | 0.54 | 0.77 | 0.57 | 0.70 | 0.44 | 0.45 | 0.45 |
| SiamRPN++ | *0.69* | *0.90* | 0.58 | 0.77 | 0.60 | 0.80 | 0.60 | *0.74* | 0.49 | 0.51 | 0.50 |
| SPM | 0.67 | 0.87 | 0.58 | 0.79 | 0.59 | 0.77 | 0.57 | 0.67 | 0.47 | 0.48 | 0.48 |
| THOR | 0.64 | 0.85 | 0.52 | 0.72 | 0.57 | 0.77 | 0.57 | 0.68 | 0.40 | 0.41 | 0.47 |
| MCCT[30] | 0.64 | 0.83 | 0.53 | 0.72 | 0.58 | 0.76 | 0.57 | 0.69 | 0.42 | 0.44 | 0.40 |
| Random | 0.66 | 0.87 | 0.56 | 0.77 | 0.58 | 0.79 | 0.57 | 0.69 | 0.46 | 0.46 | 0.48 |
| Max | 0.68 | 0.89 | 0.56 | 0.77 | 0.58 | 0.78 | *0.61* | 0.74 | 0.46 | 0.46 | 0.46 |
| AAA | **0.70** | **0.91** | **0.62** | **0.84** | *0.62* | **0.83** | **0.61** | **0.75** | **0.53** | **0.55** | **0.52** |

[30] Wang, Ning, et al, "Multi-cue correlation filters for robust visual tracking," in Proc. CVPR, 2018.

# Comparision with Low group

- Even when we aggregate experts in Low group, no other trackers can outperform the experts.

| Tracker | OTB2015 | | TColor128 | | UAV123 | | NFS | | LaSOT | | VOT2018 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AO |
| GradNet | 0.63 | 0.85 | *0.56* | *0.76* | 0.51 | 0.74 | 0.51 | 0.64 | 0.36 | 0.38 | 0.40 |
| MemTrack | 0.63 | 0.82 | 0.54 | 0.74 | 0.49 | 0.70 | 0.50 | 0.61 | 0.34 | 0.35 | 0.39 |
| SiamDW | **0.67** | **0.91** | 0.53 | 0.75 | 0.46 | 0.69 | 0.50 | 0.63 | 0.35 | 0.34 | 0.37 |
| SiamFC | 0.59 | 0.78 | 0.52 | 0.70 | 0.51 | 0.74 | 0.51 | 0.60 | 0.35 | 0.36 | 0.33 |
| SiamRPN | 0.63 | 0.83 | 0.52 | 0.71 | **0.58** | *0.77* | *0.56* | *0.66* | **0.45** | *0.45* | **0.48** |
| Staple | 0.60 | 0.79 | 0.51 | 0.68 | 0.45 | 0.64 | 0.41 | 0.48 | 0.24 | 0.23 | 0.30 |
| MCCT[30] | 0.59 | 0.79 | 0.49 | 0.66 | 0.50 | 0.70 | 0.51 | 0.63 | 0.32 | 0.34 | 0.32 |
| Random | 0.62 | 0.83 | 0.53 | 0.72 | 0.50 | 0.71 | 0.50 | 0.60 | 0.35 | 0.35 | 0.38 |
| Max | 0.63 | 0.83 | 0.52 | 0.71 | 0.51 | 0.73 | 0.53 | 0.64 | 0.35 | 0.35 | 0.38 |
| AAA | *0.66* | *0.87* | **0.59** | **0.82** | *0.56* | **0.78** | **0.58** | **0.69** | *0.45* | **0.46** | *0.45* |

[30] Wang, Ning, et al, "Multi-cue correlation filters for robust visual tracking," in Proc. CVPR, 2018.

# Comparision with Mix group

- Mix group included three high-performance experts and three low-performance experts.

| Tracker | OTB2015 | | TColor128 | | UAV123 | | NFS | | LaSOT | | VOT2018 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AO |
| ATOM | 0.67 | 0.87 | *0.60* | *0.81* | **0.62** | **0.82** | 0.58 | 0.69 | **0.51** | *0.51* | **0.52** |
| SiamRPN++ | **0.69** | **0.90** | 0.58 | 0.77 | *0.60* | 0.80 | **0.60** | **0.74** | 0.49 | 0.51 | *0.50* |
| SPM | 0.67 | 0.87 | 0.58 | 0.79 | 0.59 | 0.77 | 0.57 | 0.67 | 0.47 | 0.48 | 0.48 |
| MemTrack | 0.63 | 0.82 | 0.54 | 0.74 | 0.49 | 0.70 | 0.50 | 0.61 | 0.34 | 0.35 | 0.39 |
| SiamFC | 0.59 | 0.78 | 0.52 | 0.70 | 0.51 | 0.74 | 0.51 | 0.60 | 0.35 | 0.36 | 0.33 |
| Staple | 0.60 | 0.79 | 0.51 | 0.68 | 0.45 | 0.64 | 0.41 | 0.48 | 0.24 | 0.23 | 0.30 |
| MCCT[30] | 0.60 | 0.78 | 0.50 | 0.66 | 0.53 | 0.71 | 0.53 | 0.65 | 0.36 | 0.38 | 0.33 |
| Random | 0.64 | 0.84 | 0.55 | 0.75 | 0.54 | 0.74 | 0.53 | 0.63 | 0.40 | 0.41 | 0.42 |
| Max | 0.65 | 0.84 | 0.56 | 0.76 | 0.55 | 0.76 | 0.57 | 0.68 | 0.40 | 0.42 | 0.43 |
| AAA | *0.68* | *0.89* | **0.62** | **0.83** | 0.60 | *0.81* | *0.59* | *0.71* | *0.51* | **0.53** | 0.49 |

[30] Wang, Ning, et al, "Multi-cue correlation filters for robust visual tracking," in Proc. CVPR, 2018.

# Comparision with SiamDW[10]

- we implemented experts using several parameter sets (e.g., backbone networks, the weight of the network, and hyper-parameters) in SiamDW[10]

| Tracker | OTB2015 | | TColor128 | | UAV123 | | NFS | | LaSOT | | VOT2018 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AO |
| SiamDW_SiamFCRes22 | 0.64 | 0.84 | *0.58* | *0.79* | 0.51 | 0.73 | 0.52 | 0.64 | *0.38* | *0.39* | 0.38 |
| SiamDW_SiamFCIncep22 | 0.61 | 0.81 | 0.55 | 0.76 | 0.50 | 0.72 | 0.51 | 0.64 | 0.36 | 0.38 | 0.35 |
| SiamDW_SiamFCNext22 | 0.62 | 0.82 | 0.57 | 0.76 | 0.49 | 0.71 | 0.51 | 0.63 | 0.37 | 0.38 | 0.32 |
| SiamDW_SiamRPNRes22 | **0.67** | **0.91** | 0.53 | 0.75 | 0.46 | 0.69 | 0.50 | 0.63 | 0.35 | 0.34 | 0.37 |
| SiamDW_SiamFCRes22_VOT | 0.63 | 0.84 | 0.56 | 0.77 | 0.51 | 0.73 | 0.52 | 0.65 | 0.36 | 0.37 | 0.37 |
| SiamDW_SiamFCIncep22_VOT | 0.60 | 0.80 | 0.54 | 0.75 | 0.50 | 0.73 | 0.49 | 0.61 | 0.35 | 0.36 | 0.35 |
| SiamDW_SiamFCNext22_VOT | 0.61 | 0.81 | 0.54 | 0.74 | 0.49 | 0.72 | 0.51 | 0.63 | 0.35 | 0.38 | 0.34 |
| SiamDW_SiamRPNRes22_VOT | 0.66 | *0.90* | 0.53 | 0.74 | 0.46 | 0.69 | 0.51 | *0.66* | 0.35 | 0.35 | **0.43** |
| MCCT[30] | 0.63 | 0.83 | 0.54 | 0.74 | 0.50 | 0.71 | 0.51 | 0.65 | 0.34 | 0.36 | 0.34 |
| Random | 0.63 | 0.84 | 0.55 | 0.76 | 0.49 | 0.71 | 0.51 | 0.64 | 0.36 | 0.37 | 0.37 |
| Max | 0.64 | 0.86 | 0.55 | 0.76 | *0.51* | *0.74* | *0.53* | 0.66 | 0.36 | 0.36 | 0.37 |
| AAA | *0.66* | 0.88 | **0.60** | **0.82** | **0.52** | **0.75** | **0.55** | **0.68** | **0.42** | **0.43** | *0.42* |

[10] Zhang, Zhipeng, and Houwen Peng, "Deeper and wider siamese networks for real-time visual tracking," in Proc. CVPR, 2019.
[30] Wang, Ning, et al, "Multi-cue correlation filters for robust visual tracking," in Proc. CVPR, 2018.

# Comparision with SiamRPN++[5]

- we implemented experts using several parameter sets (e.g., backbone networks, the weight of the network, and hyper-parameters) in SiamRPN++[5]

| Tracker | OTB2015 | | TColor128 | | UAV123 | | NFS | | LaSOT | | VOT2018 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AO |
| SiamRPN++_AlexNet | 0.66 | 0.87 | 0.57 | 0.77 | 0.58 | 0.77 | 0.54 | 0.65 | 0.45 | 0.45 | 0.47 |
| SiamRPN++_AlexNet_OTB | 0.66 | 0.86 | 0.55 | 0.75 | 0.58 | 0.78 | 0.54 | 0.66 | 0.43 | 0.43 | 0.45 |
| SiamRPN++_ResNet-50 | 0.65 | 0.86 | 0.56 | 0.75 | 0.61 | 0.80 | 0.58 | 0.71 | 0.50 | 0.51 | 0.51 |
| SiamRPN++_ResNet-50_OTB | **0.69** | **0.90** | *0.58* | 0.77 | 0.60 | 0.80 | 0.60 | *0.74* | 0.49 | 0.51 | 0.50 |
| SiamRPN++_ResNet-50_LT | 0.63 | 0.84 | 0.58 | *0.79* | 0.61 | *0.81* | 0.56 | 0.68 | *0.52* | *0.54* | *0.51* |
| SiamRPN++_MobileNetV2 | 0.65 | 0.86 | 0.56 | 0.76 | 0.60 | 0.79 | 0.57 | 0.70 | 0.45 | 0.46 | 0.50 |
| SiamRPN++_SiamMask | 0.65 | 0.85 | 0.54 | 0.73 | 0.60 | 0.80 | 0.58 | 0.72 | 0.47 | 0.48 | 0.48 |
| MCCT [30] | 0.64 | 0.84 | 0.55 | 0.75 | *0.61* | 0.81 | 0.59 | 0.72 | 0.48 | 0.50 | 0.45 |
| Random | 0.66 | 0.86 | 0.56 | 0.76 | 0.60 | 0.79 | 0.57 | 0.69 | 0.47 | 0.48 | 0.49 |
| Max | 0.66 | 0.87 | 0.56 | 0.76 | 0.60 | 0.80 | *0.60* | 0.73 | 0.47 | 0.48 | 0.50 |
| AAA | *0.68* | *0.89* | **0.61** | **0.83** | **0.64** | **0.85** | **0.61** | **0.74** | **0.54** | **0.56** | **0.52** |

[5] Li, Bo, et al, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in Proc. CVPR, 2019.
[30] Wang, Ning, et al, "Multi-cue correlation filters for robust visual tracking," in Proc. CVPR, 2018.

# Accuracy in anchor frames

- We also evaluate whether the proposed method can actually detect the target location with high confidence in anchor frames

| Tracker | OTB2015 | | TColor128 | | UAV123 | | NFS | | LaSOT | | VOT2018 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AUC | DP | AO |
| ATOM | 0.72 | 0.93 | *0.66* | 0.88 | *0.71* | *0.91* | 0.66 | 0.80 | *0.66* | *0.70* | *0.61* |
| DaSiamRPN | 0.70 | 0.92 | 0.61 | 0.84 | 0.68 | 0.88 | 0.64 | 0.79 | 0.59 | 0.62 | 0.57 |
| SiamMCF | 0.71 | 0.91 | 0.66 | 0.88 | 0.66 | 0.88 | 0.67 | 0.83 | 0.60 | 0.64 | 0.57 |
| SiamRPN++ | *0.73* | *0.94* | 0.64 | 0.85 | 0.70 | 0.89 | *0.68* | *0.84* | 0.64 | 0.69 | 0.60 |
| SPM | 0.72 | 0.92 | 0.65 | *0.89* | 0.69 | 0.87 | 0.65 | 0.78 | 0.63 | 0.67 | 0.60 |
| THOR | 0.68 | 0.90 | 0.59 | 0.79 | 0.67 | 0.87 | 0.64 | 0.79 | 0.55 | 0.58 | 0.59 |
| AAA | **0.74** | **0.94** | **0.70** | **0.92** | **0.72** | **0.93** | **0.70** | **0.87** | **0.70** | **0.76** | **0.64** |

# Tracking example in MOT