Object Tracking

Danna Gurari
University of Colorado Boulder
Fall 2024

https://dannagurari.colorado.edu/course/recent-advances-in-computer-vision-fall-2024/

Review

- Last lecture: instance segmentation
  - Motivation
  - Datasets
  - Evaluation metric
  - Mask R-CNN
  - YOLACT
- Assignments (Canvas)
  - Reading assignment was due earlier today
  - Next reading assignments due for next two lectures
  - Project outline due in 1.5 weeks
- Questions?

Object Tracking: Today's Topics

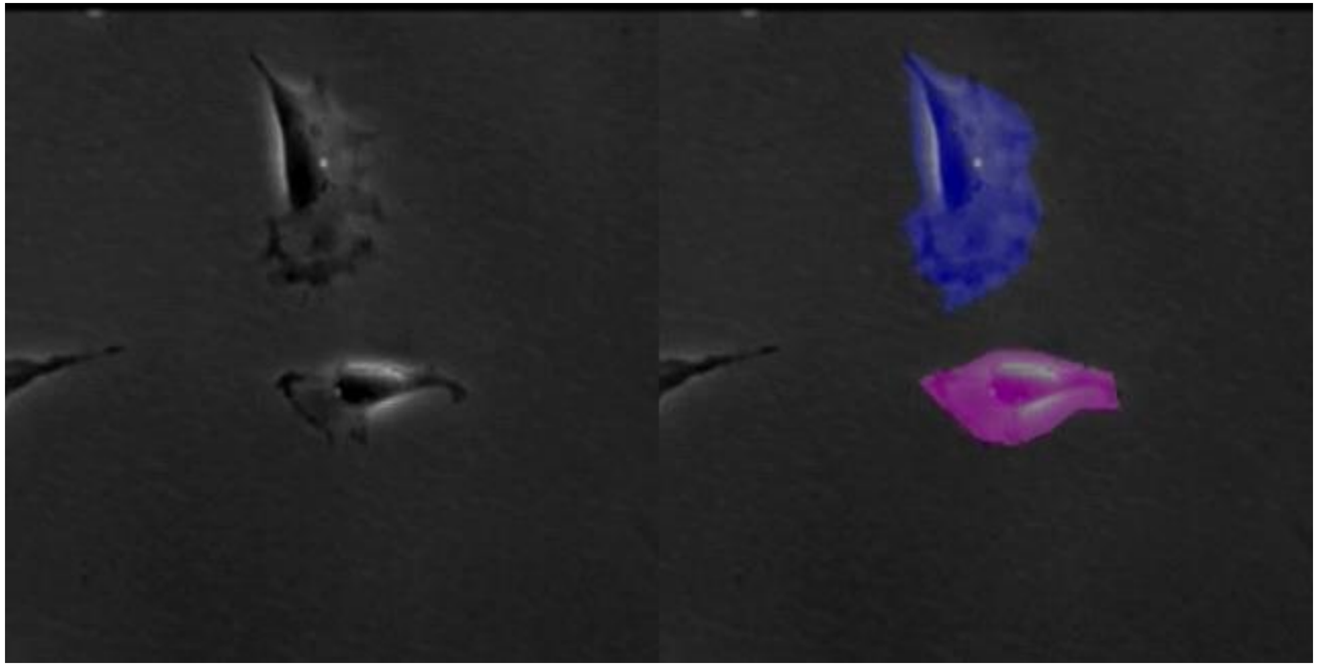
- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models
- Discussion (chosen by YOU ☺)

Object Tracking: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models
- Discussion (chosen by YOU ☺)

Definition

- Identification of the trajectory of an object over time
  - Single object
  - Multiple objects; e.g.,



Definition

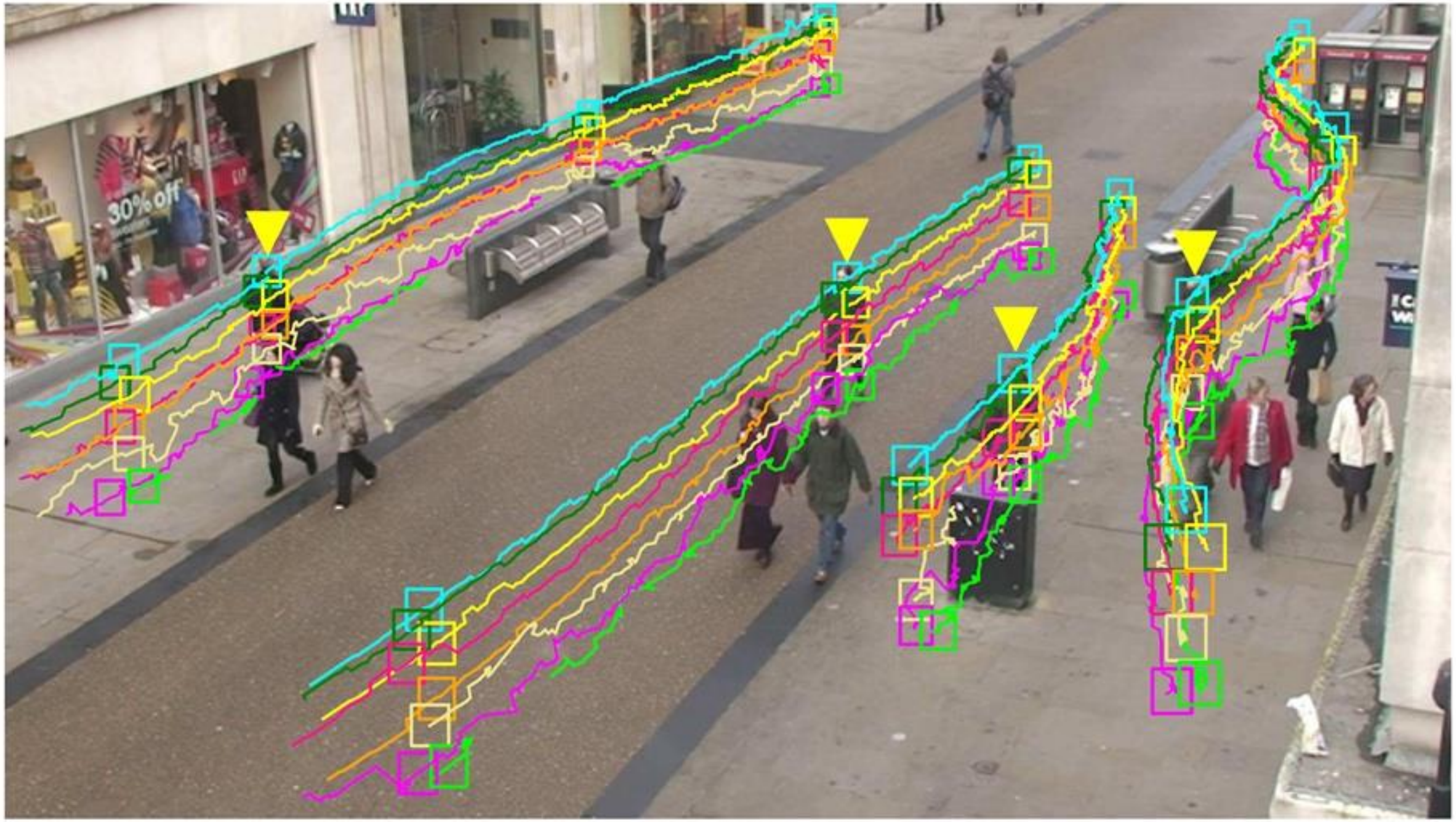
- Identification of the trajectory of an object over time
    - Single object
    - Multiple objects

- How can the trajectory of an object be represented?
    - Bounding box or ellipse
    - Segmentation or coarse outline
    - Position (e.g., object centroid, corner, salient point)

Object Tracking: Today’s Topics

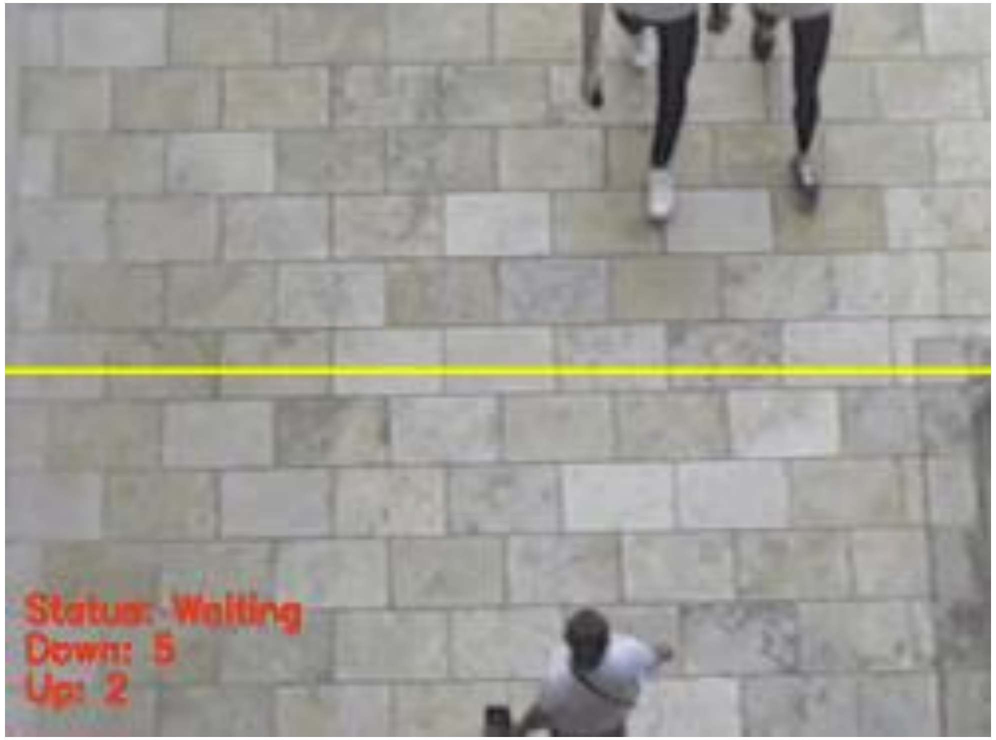
- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models
- Discussion (chosen by YOU ☺)

Surveillance



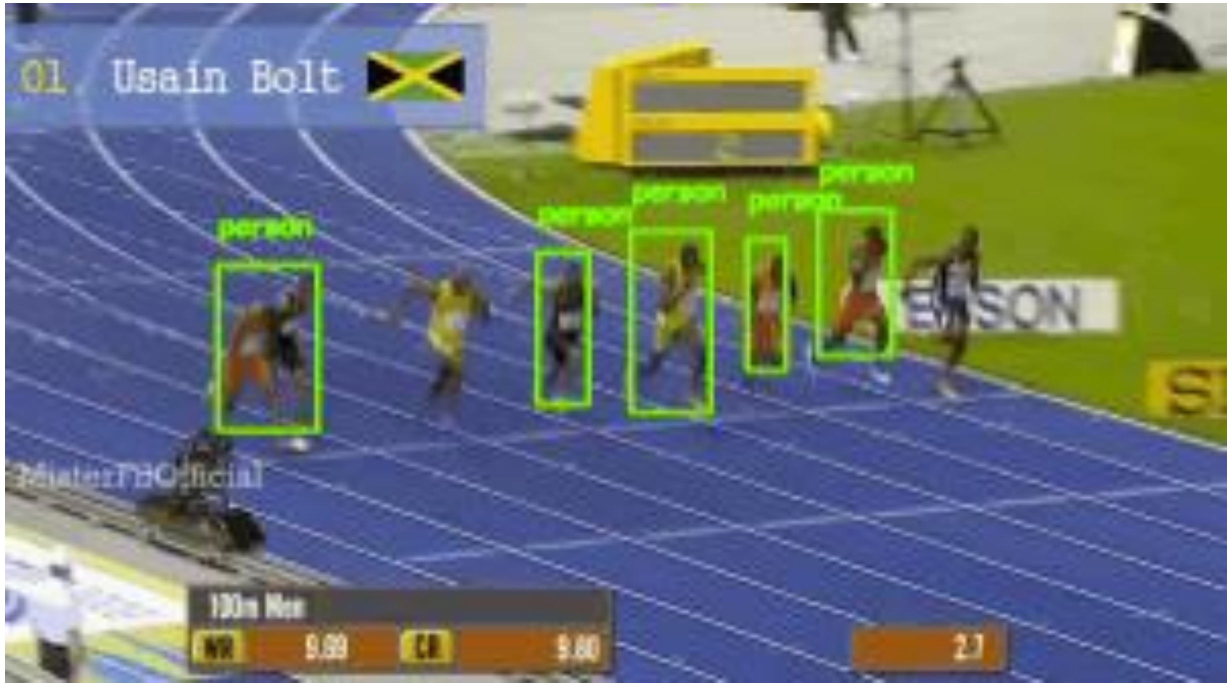
http://www.vision.caltech.edu/koller/MOU-83.html

Business Marketing: People Analytics



https://www.pyimagesearch.com/2018/08/13/opencv-people-counter/
https://www.nytimes.com/2013/07/15/business/attention-shopper-stores-are-tracking-your-cell.html

Sports Analysis



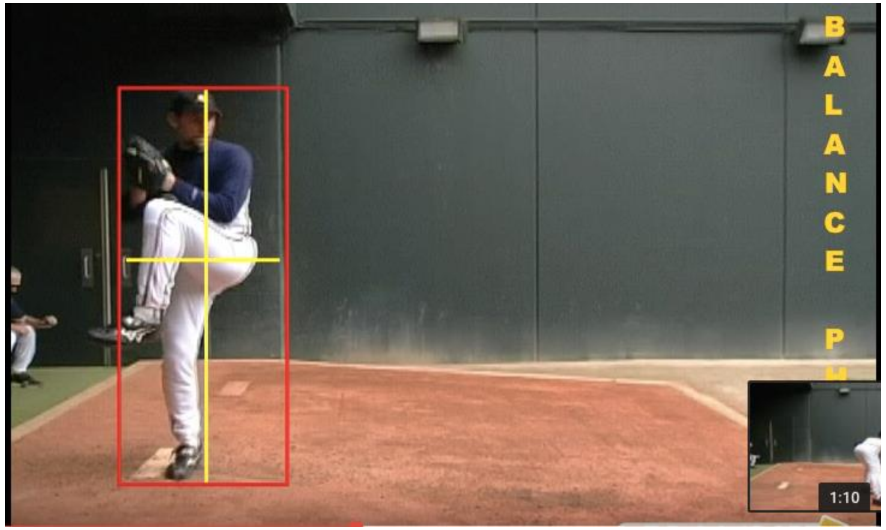
https://www.pyimagesearch.com/2018/10/29/multi-object-tracking-with-dlib/
https://www.pyimagesearch.com/2018/08/06/tracking-multiple-objects-with-opencv/

Sports Performance Analytics

Calculate Bat speed from video!

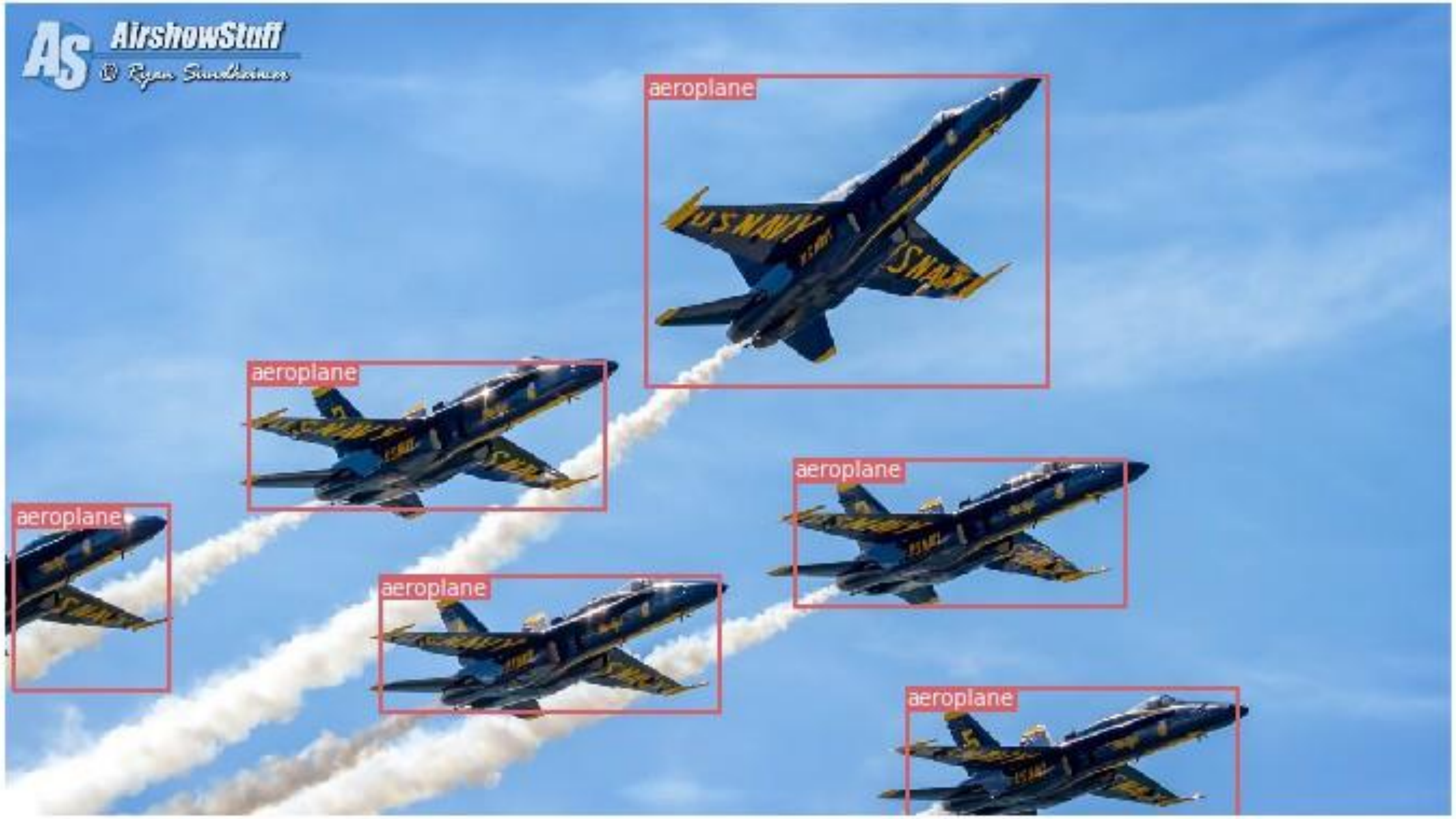
NEW! Track Bowling Ball Path!

Works great for putting!



http://www.motionprosoftware.com/

Military Defense

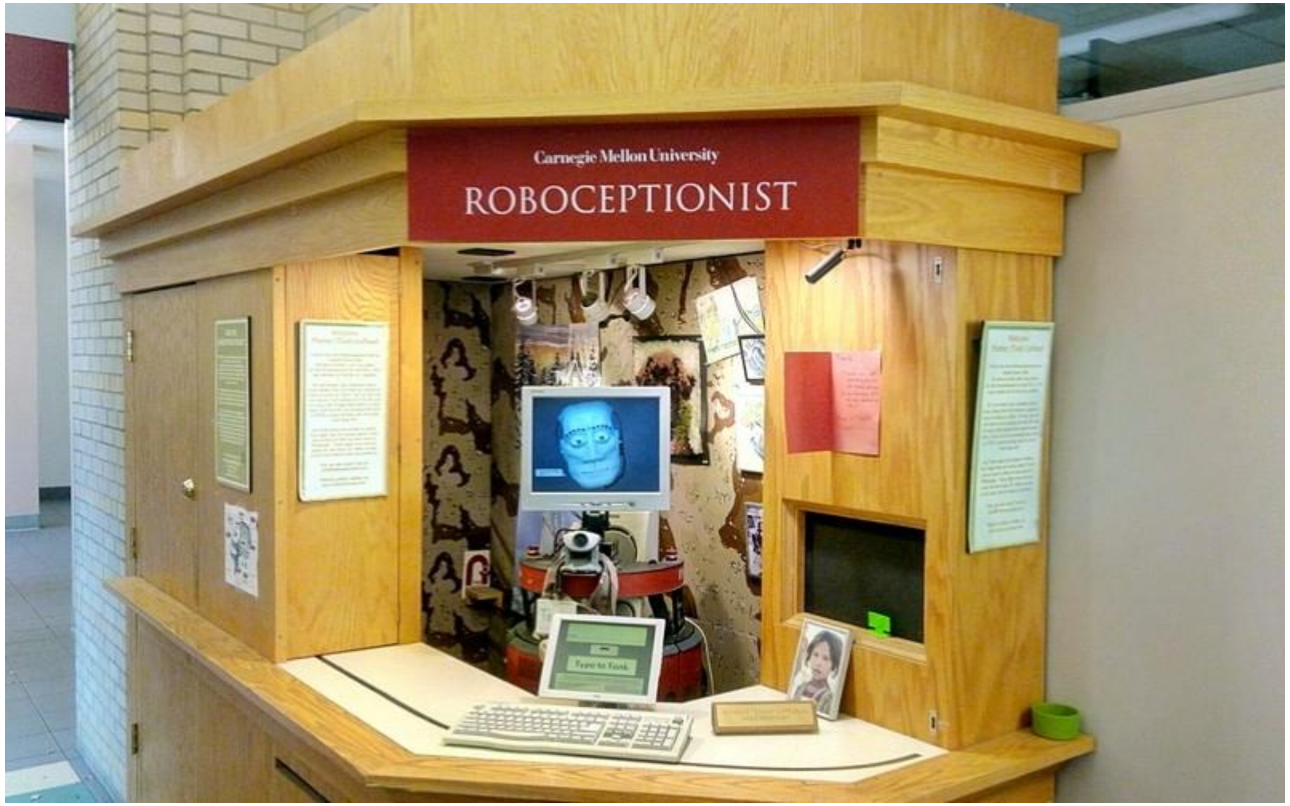


https://towardsdatascience.com/object-detection-and-tracking-in-pytorch-b3cf1a696a98

Self-driving Cars

https://www.pyimagesearch.com/2018/08/06/tracking-multiple-objects-with-opencv/

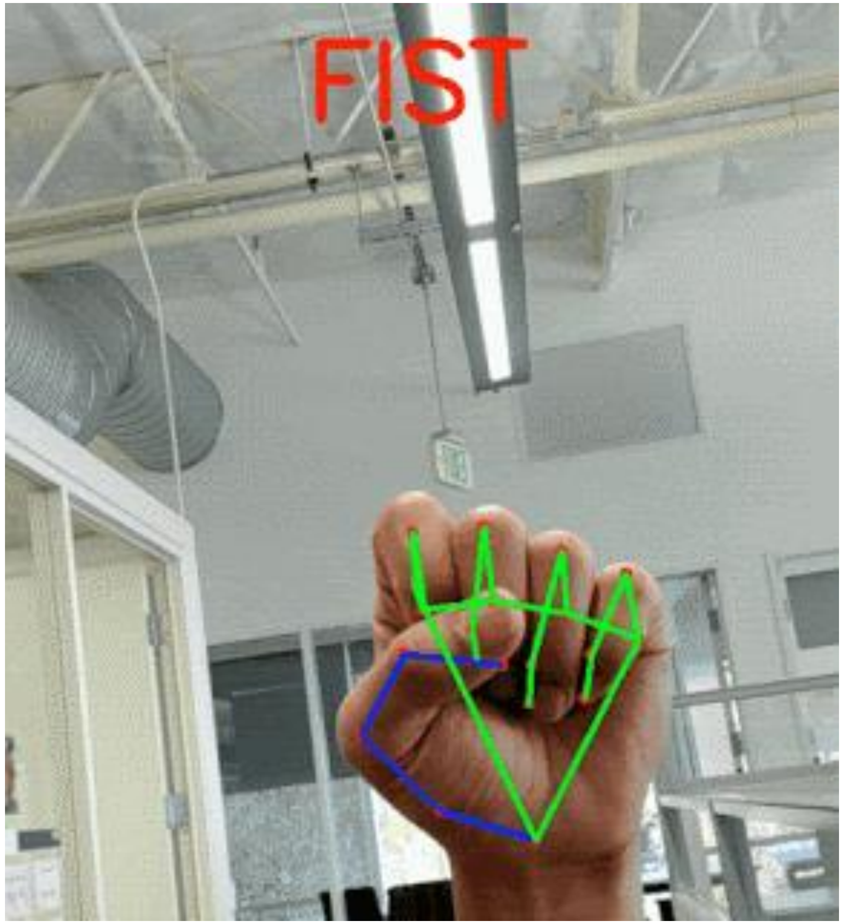
Human Computer Interaction



Roboceptionist

https://alapkshirsagar.github.io/travel/namerica/usa/wv/47.jpg

Sign Language Recognition



https://techcrunch.com/2019/08/19/this-hand-tracking-algorithm-could-lead-to-sign-language-recognition/

Biological Monitoring

Counting bats exiting a cave in Texas:

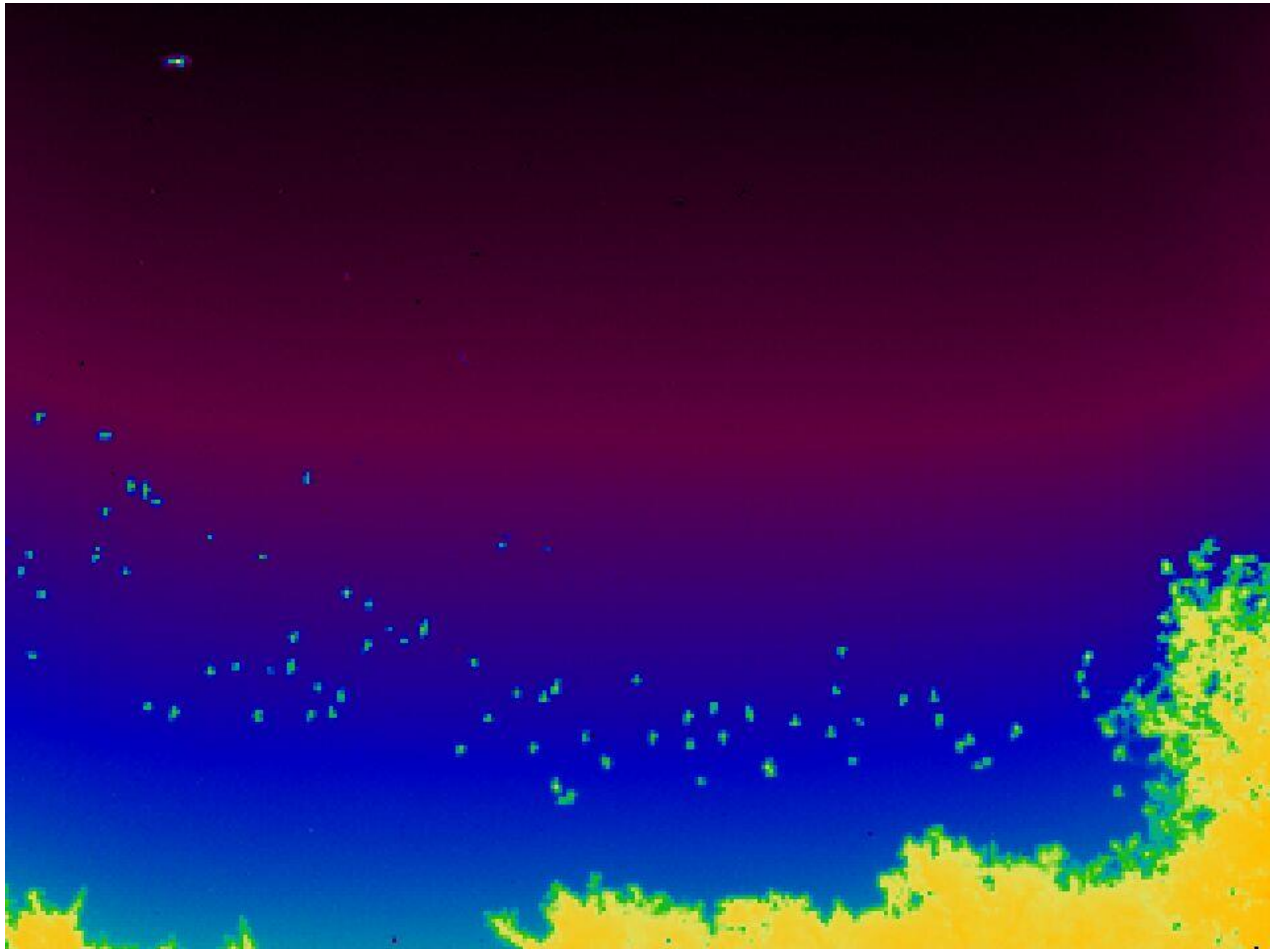


Image source: https://www.cs.bu.edu/fac/betke/research/bats/images2.html

Augmented Reality

https://virtualrealitypop.com/object-recognition-in-augmented-reality-8f7f17127a7a
https://www.geekwire.com/2017/augmented-reality-shopping-phone-patent-hints-amazons-aspirations/

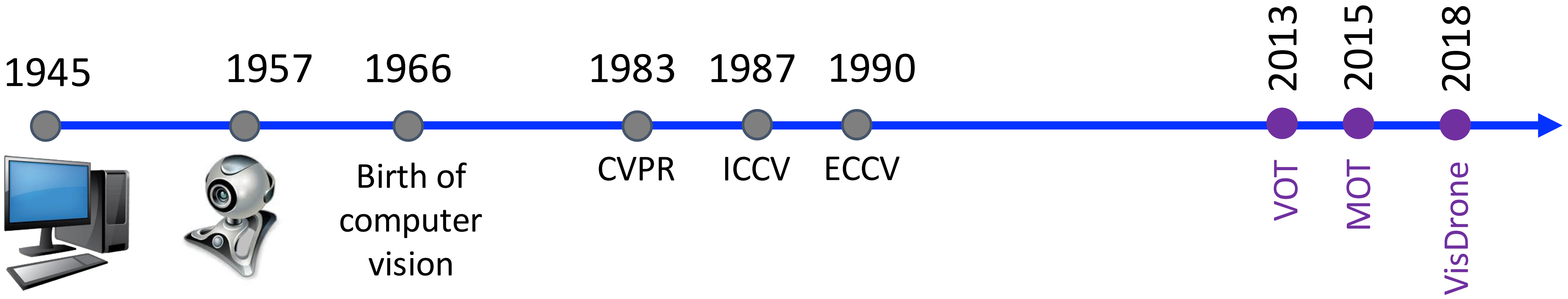
Applications

What other applications can you think of where object tracking could be useful?

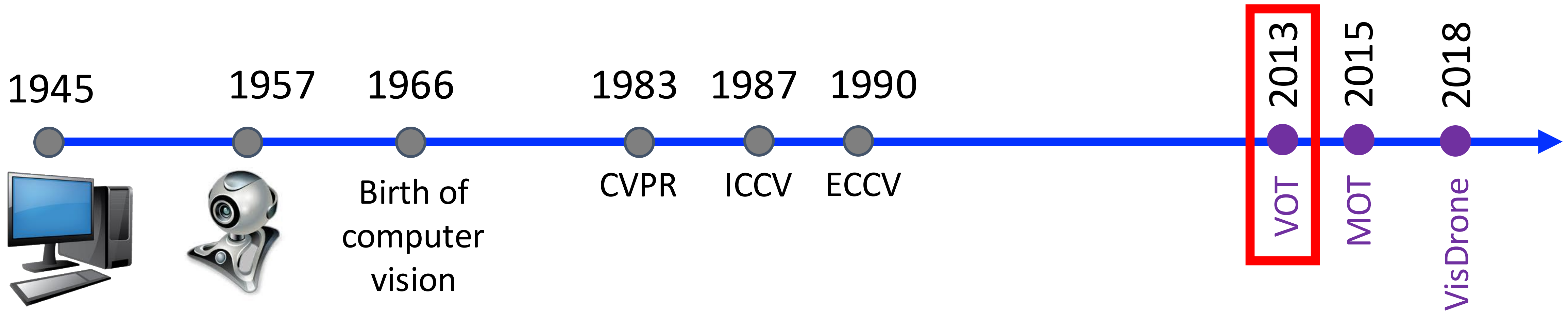
Object Tracking: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models
- Discussion (chosen by YOU ☺)

Object Tracking Datasets



Object Tracking Datasets



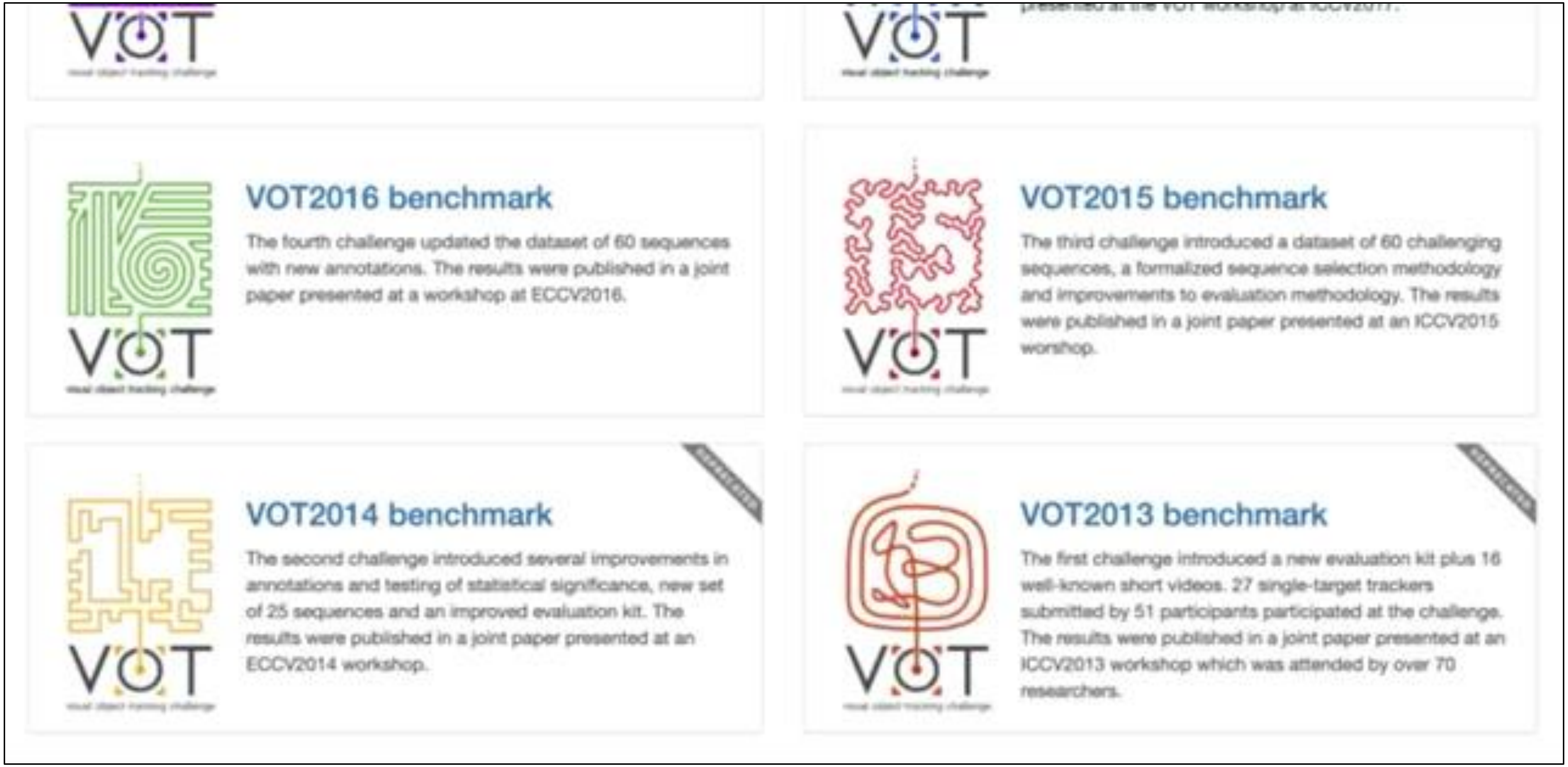
Single Object Tracking Dataset: VOT

- Aggregated 16 videos from existing datasets that used bounding boxes to track a single object in each video
  - Limitation: inconsistent annotation methodologies across videos (e.g., different bounding box criteria)

- Authors re-annotated object tracking for videos they believed had unsuitable annotations

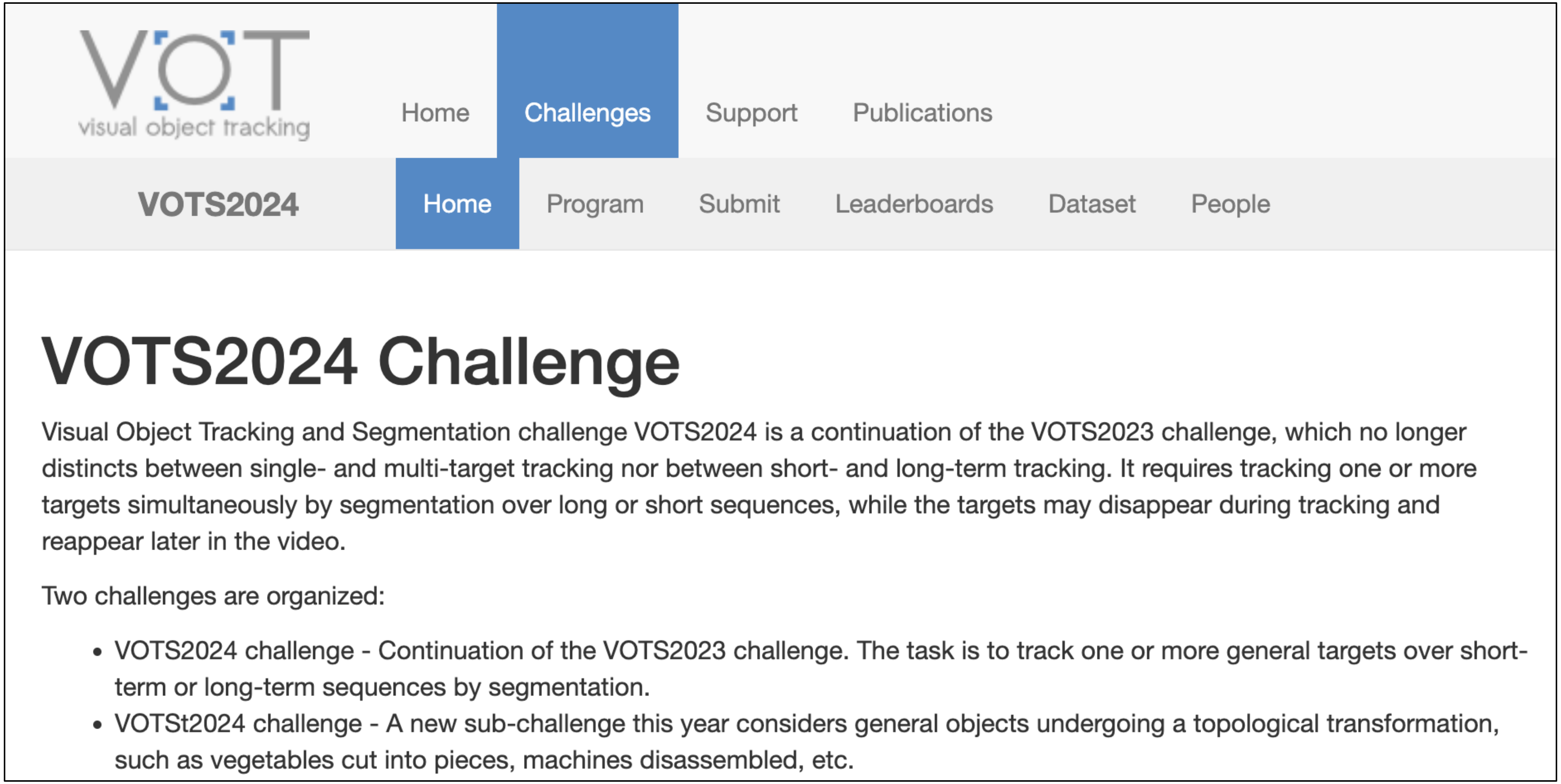
Matej Kristan et al. The Visual Object Tracking VOT2013 Results. ICCV Workshop 2013.

Single Object Tracking Dataset: VOT's Evolution



Matej Kristan et al. The Visual Object Tracking VOT2013 Results. ICCV Workshop 2013.

Single Object Tracking Annual Challenge (12th year in 2024)


visual object tracking

Home | Challenges | Support | Publications

VOTS2024 | Home | Program | Submit | Leaderboards | Dataset | People

VOTS2024 Challenge

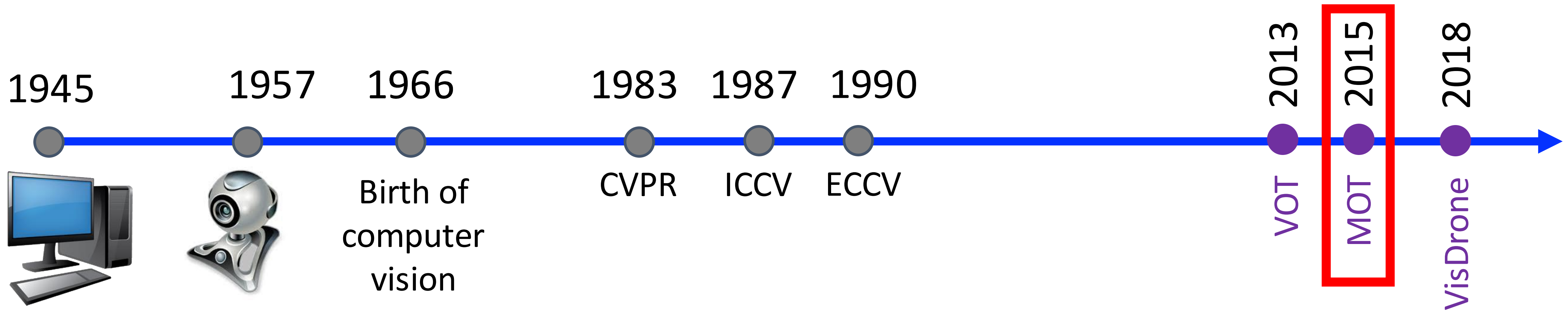
Visual Object Tracking and Segmentation challenge VOTS2024 is a continuation of the VOTS2023 challenge, which no longer distincts between single- and multi-target tracking nor between short- and long-term tracking. It requires tracking one or more targets simultaneously by segmentation over long or short sequences, while the targets may disappear during tracking and reappear later in the video.

Two challenges are organized:

- VOTS2024 challenge - Continuation of the VOTS2023 challenge. The task is to track one or more general targets over short-term or long-term sequences by segmentation.
- VOTSt2024 challenge - A new sub-challenge this year considers general objects undergoing a topological transformation, such as vegetables cut into pieces, machines disassembled, etc.

https://www.votchallenge.net/vot2021/

Object Tracking Datasets



Multiple Object Tracking Dataset: MOT

- Authors aggregated 22 videos that contain a total of 11,286 frames associated with 61,440 annotated bounding boxes
  - **Static and moving camera**; e.g., held by a person, stroller, and car
  - **Multiple viewpoints**; e.g., cameras positioned at a high, medium, and low position (e.g., person's height vs on ground looking up)
  - **Multiple weather conditions**; e.g., sunny vs cloudy vs night time
- 16 videos from existing datasets and other 6 generated by the authors; tracked objects were people and vehicles

Laura Leal-Taixe, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. arXiv 2015.

Multiple Object Tracking Dataset: MOT

- Authors aggregated 22 videos that contain a total of 11,286 frames associated with 61,440 annotated bounding boxes
    - **Static and moving camera**; e.g., held by a person, stroller, and car
    - **Multiple viewpoints**; e.g., cameras positioned at a high, medium, and low position (e.g., person's height vs on the ground looking up)
    - **Multiple weather conditions**; e.g., sunny versus cloudy versus night time
- Annotations:
    - Automatically-generated detections for the dataset provided
    - For existing videos, there GT was used
    - For new videos, the VATIC annotation tool was used to generate tracks

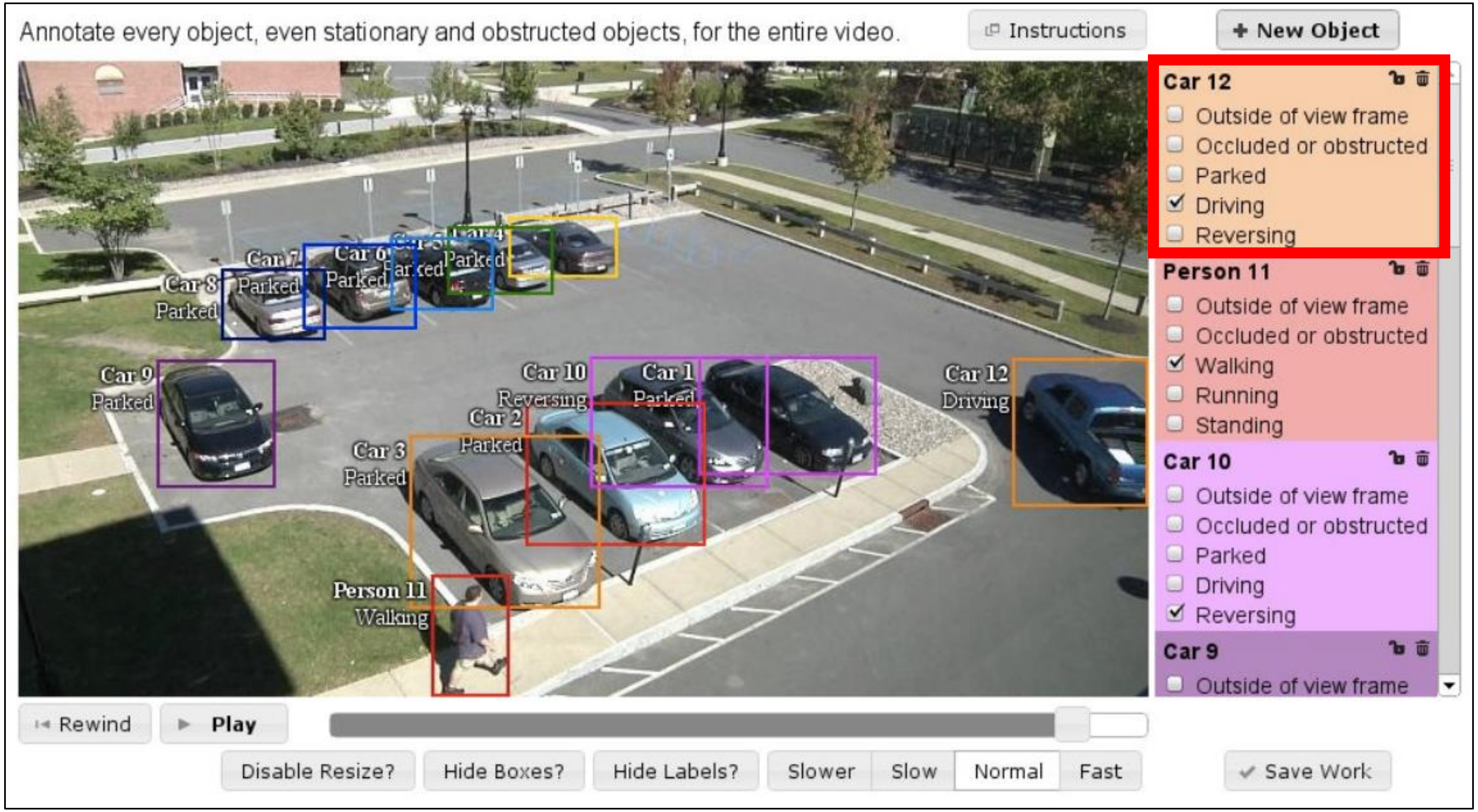
Laura Leal-Taixe, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. arXiv 2015.

Multiple Object Tracking Annotation: VATIC

Demo: https://www.youtube.com/watch?v=ljI5pAowACc

Vondrick et al. IJCV 2012

Multiple Object Tracking Annotation: VATIC



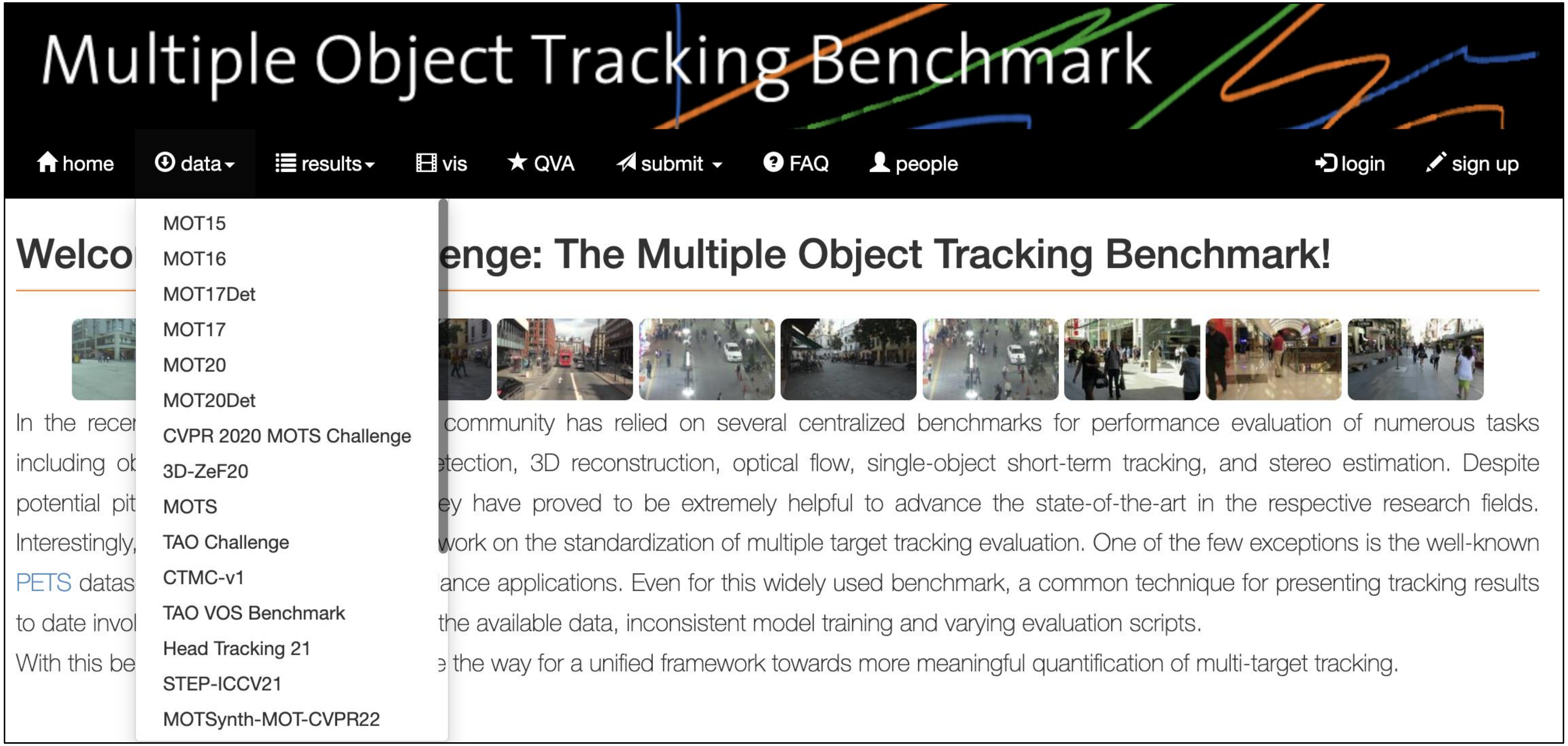
Metadata about each object includes activity and attributes

Vondrick et al. IJCV 2012

Multiple Object Tracking Annotation: VATIC

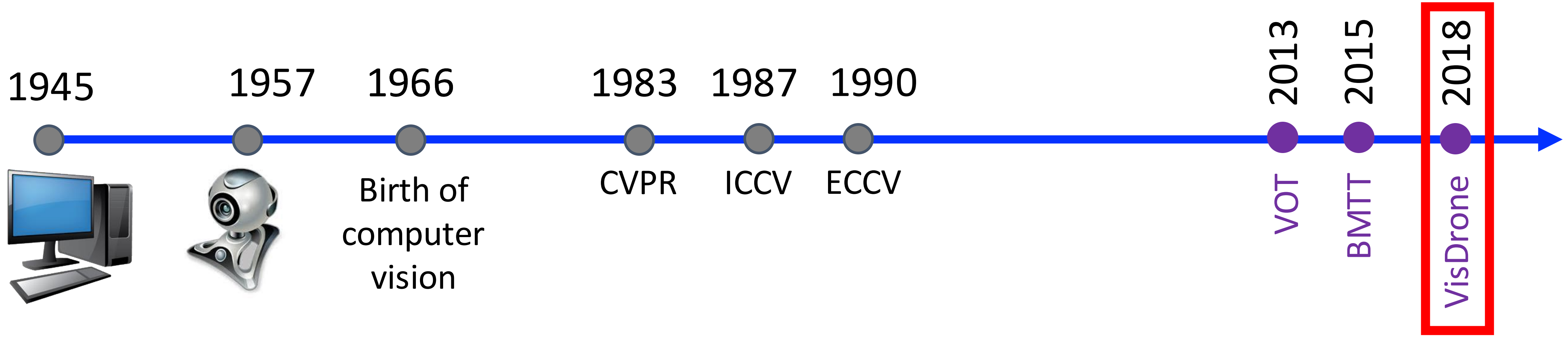
- How to handle occlusions?
  - Instructions: "*Always annotate during occlusions if the position can be determined unambiguously. If the occlusion is very long and it is not possible to determine the path of the object using simple reasoning (e.g. constant velocity assumption), the object will be assigned a new ID once it reappears*"
  - "**visibility**" flag: 0-1 with 1 fully visible and less than 1 indicating occlusions
  - "**confidence**" flag: 1 when box should be considered for evaluation and 0 otherwise (e.g., a pedestrian is too small)
  - Non-tracked categories: "**class**" value is occluder and ignored during evaluation

Multiple Object Tracking Annual Challenge (10^{th} year in 2024)



https://motchallenge.net/

Object Tracking Datasets



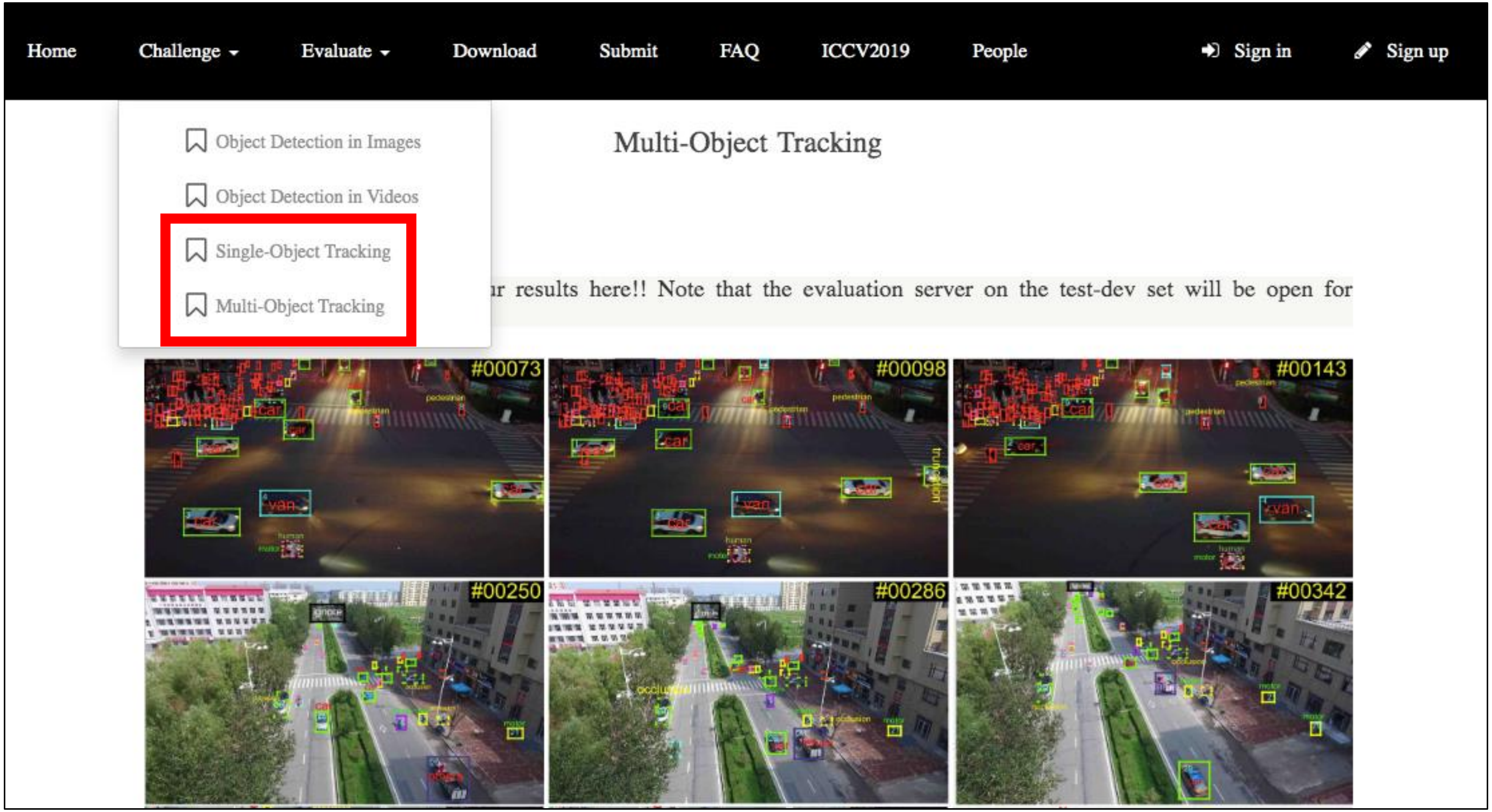
VisDrone

- Authors collected 263 video clips (179,264 frames) from drones in Asia

- Annotations created for over 2.5 million object instances, without description of how annotations were collected

Pengfei Zhu, Longyin Wen, Xiao Bian, Haibing Ling, and Qinghua Hu. Vision Meets Drones: A Challenge. arXiv 2018

VisDrone Challenge



http://www.aiskyeye.com/views/index

Discussion

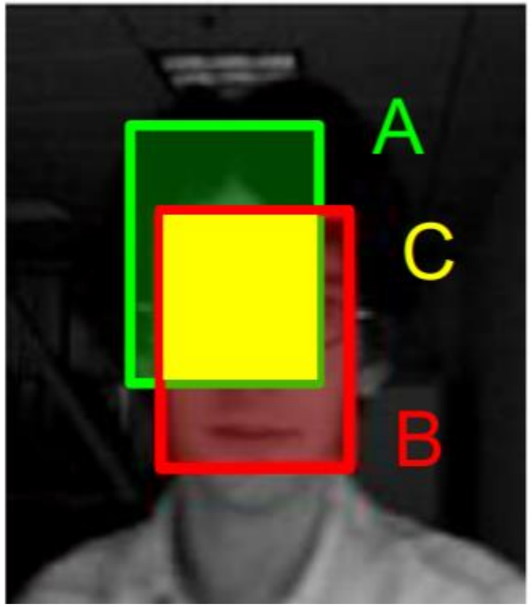
- When designing an annotation protocol, how should these scenarios be handled:
  - Partially visible object
  - Occluded object
  - Object is reflected in reflective surfaces such as mirrors or windows

Object Tracking: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models
- Discussion (chosen by YOU ☺)

Accuracy

Average IoU from a tracker across all video frames

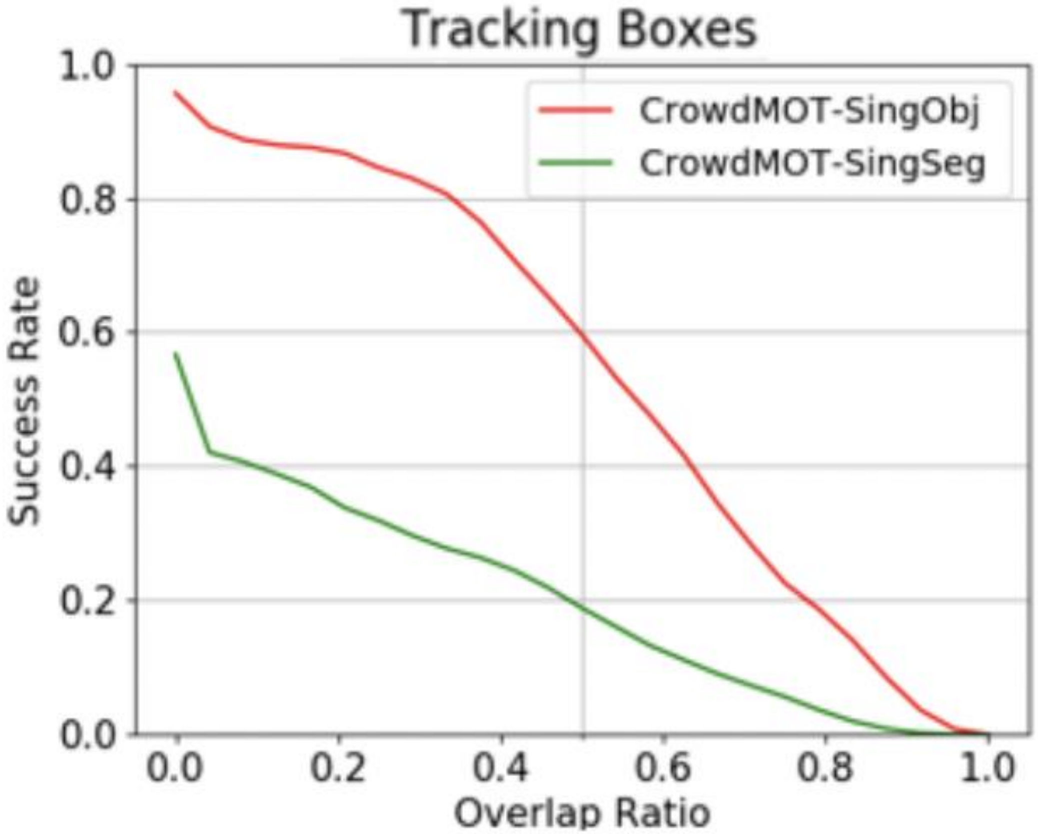


A = Ground Truth
B = Predicted Track
C = Intersection

Figure credit: https://ags.cs.uni-kl.de/fileadmin/inf_ags/opt-ss15/OPT_SS2015_lec11.pdf
Matej Kristan et al. "A Novel Performance Evaluation Methodology for Single-Target Trackers." PAMI 2016

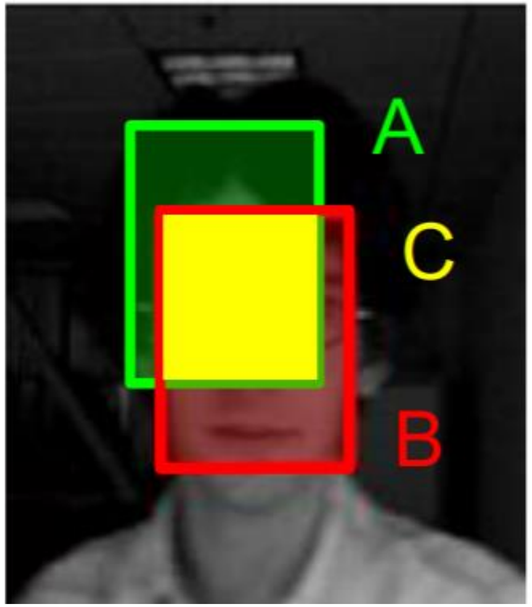
Success Plot

Percentage of frames where the IoU is larger than a given threshold (e.g., 0.5); can create a plot by varying the threshold amount



Robustness

Average number of times a tracker drifts to an IoU value of 0 and so needs to be re-initialized to the ground truth bounding box per video

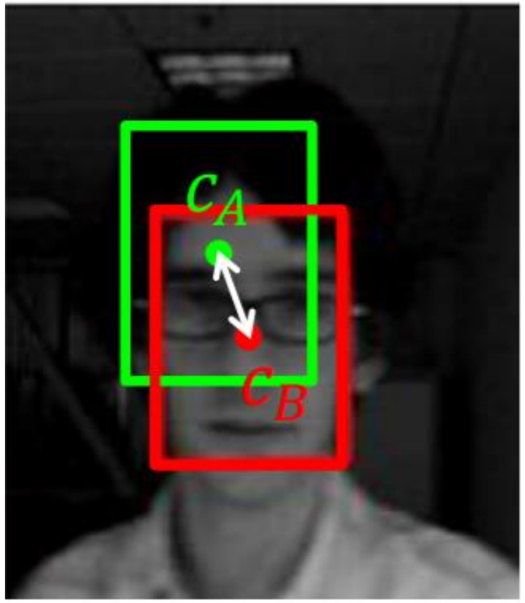


A = Ground Truth
B = Predicted Track
C = Intersection

Figure credit: https://ags.cs.uni-kl.de/fileadmin/inf_ags/opt-ss15/OPT_SS2015_lec11.pdf
Matej Kristan et al. “A Novel Performance Evaluation Methodology for Single-Target Trackers.” PAMI 2016

Precision

Distance between the centers of bounding boxes for each frame



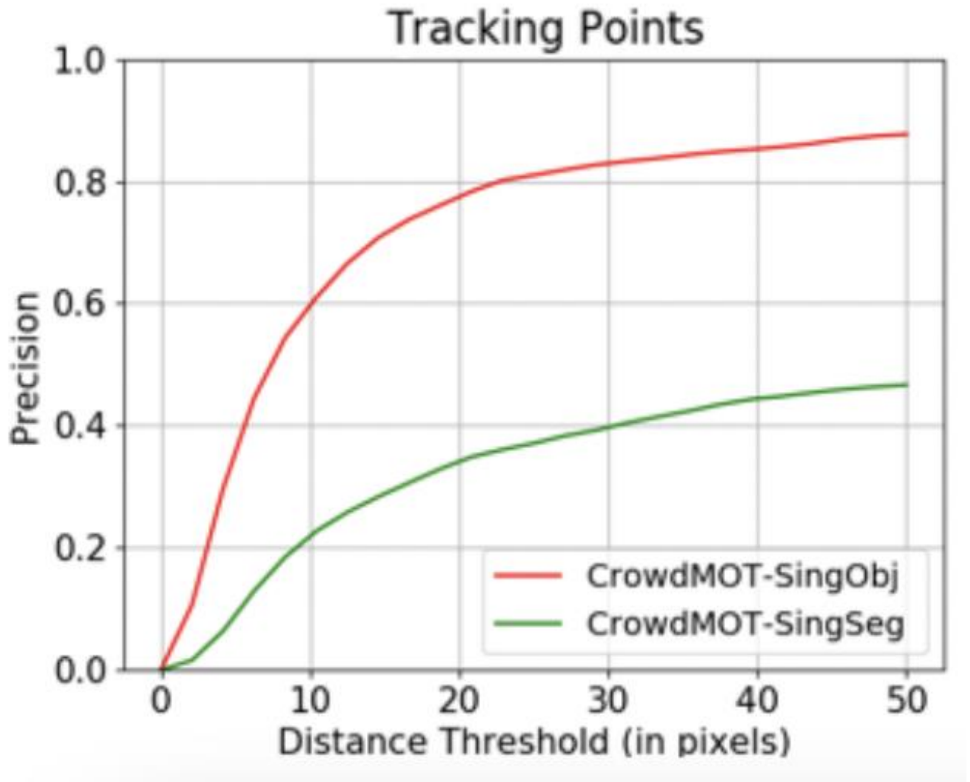
A = Ground Truth
B = Predicted Track

$$p = \|c_A - c_B\|$$

Figure credit: https://ags.cs.uni-kl.de/fileadmin/inf_ags/opt-ss15/OPT_SS2015_lec11.pdf

Precision Plot

Percentage of frames with predicted location within a given threshold distance of ground truth (e.g., 20 pixels); can create a plot by varying the threshold amount



MOTA

Number of missed detections, meaning that IoU is below a given threshold

Number of false positives, meaning that a non-existing detection is introduced

$$1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDSW}_t)}{\sum_t \mathrm{GT}_t}$$

Number of video frames

Number of GT objects

Number of identity switches

Laura Leal-Taixe, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. arXiv 2015.

MOTA

What is the range of possible values?

- (- infinite, 100] (original value usually multiplied by 100)

When is MOTA negative?

- When the number of errors exceed the number of objects in the frames

Number of missed detections, meaning that IoU is below a given threshold

Number of false positives, meaning that a non-existing detection is introduced

$$1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDSW}_t)}{\sum_t \mathrm{GT}_t}$$

Number of video frames

Number of GT objects

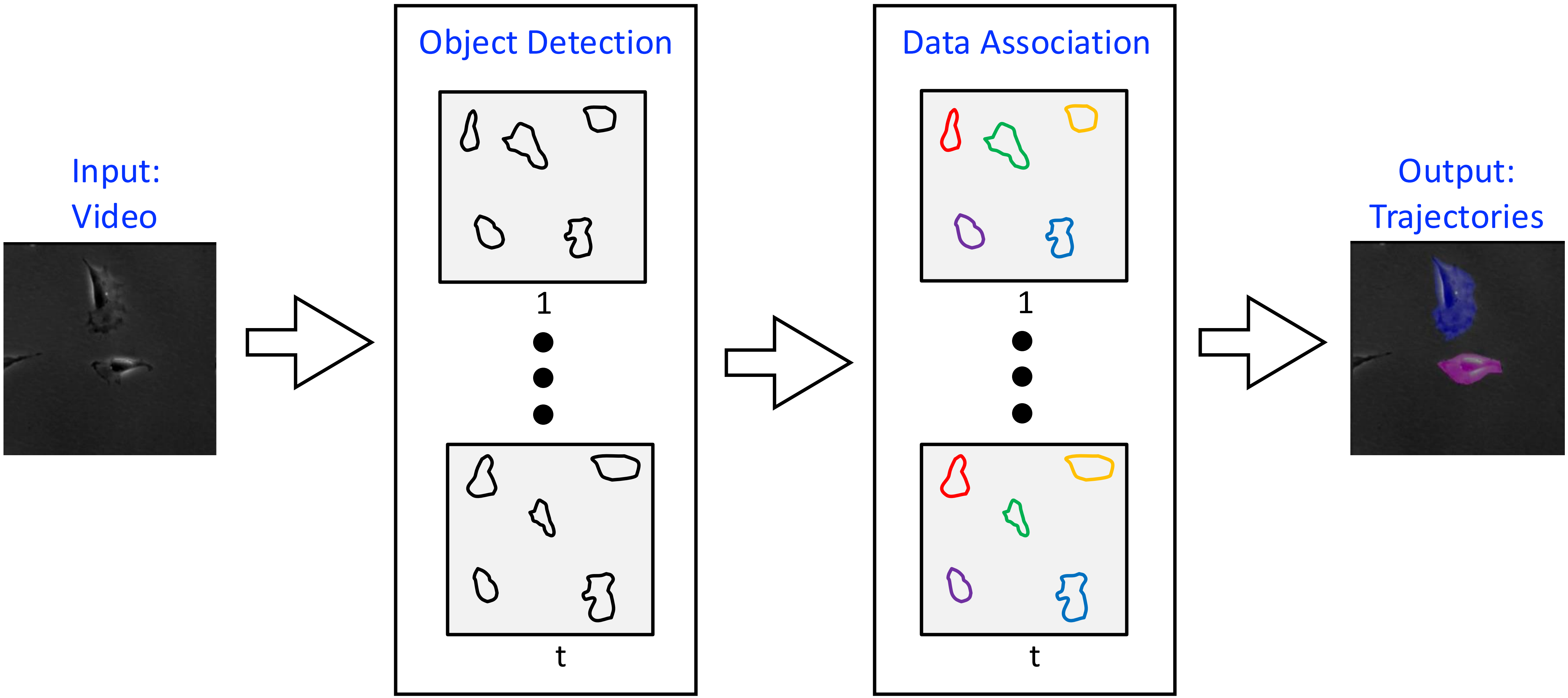
Number of identity switches

Laura Leal-Taixe, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. arXiv 2015.

Object Tracking: Today’s Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models
- Discussion (chosen by YOU ☺)

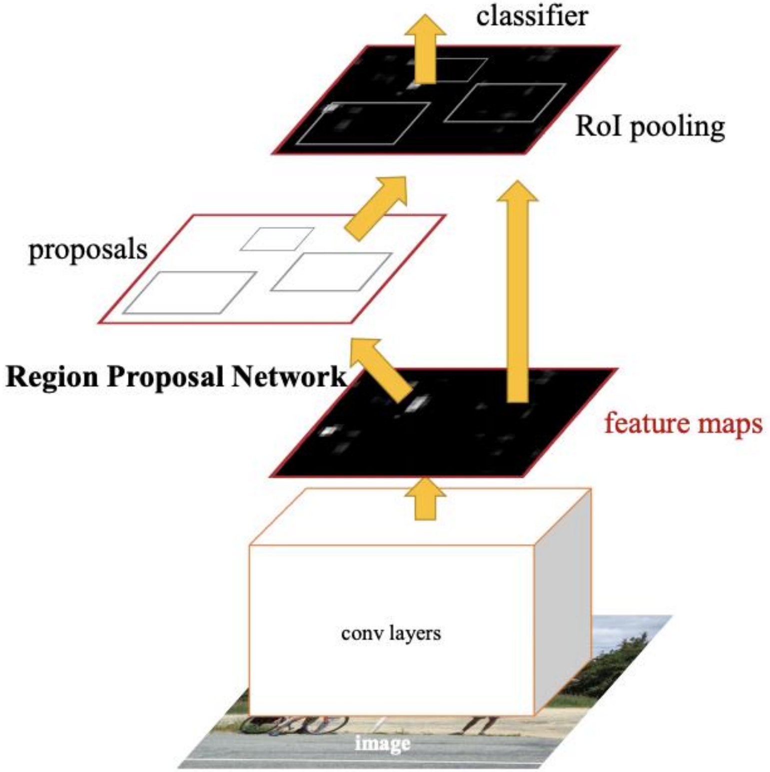
A Common Approach: Tracking-by-Detection



Tracktor – Base Architecture
(FPN Variant of Faster R-CNN)

Input:
Video

Object Detection: Refine Previous Frame Using Regression Head on Faster R-CNN with FPN

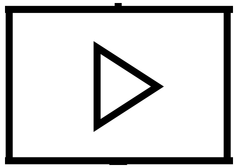


Output:
Trajectories

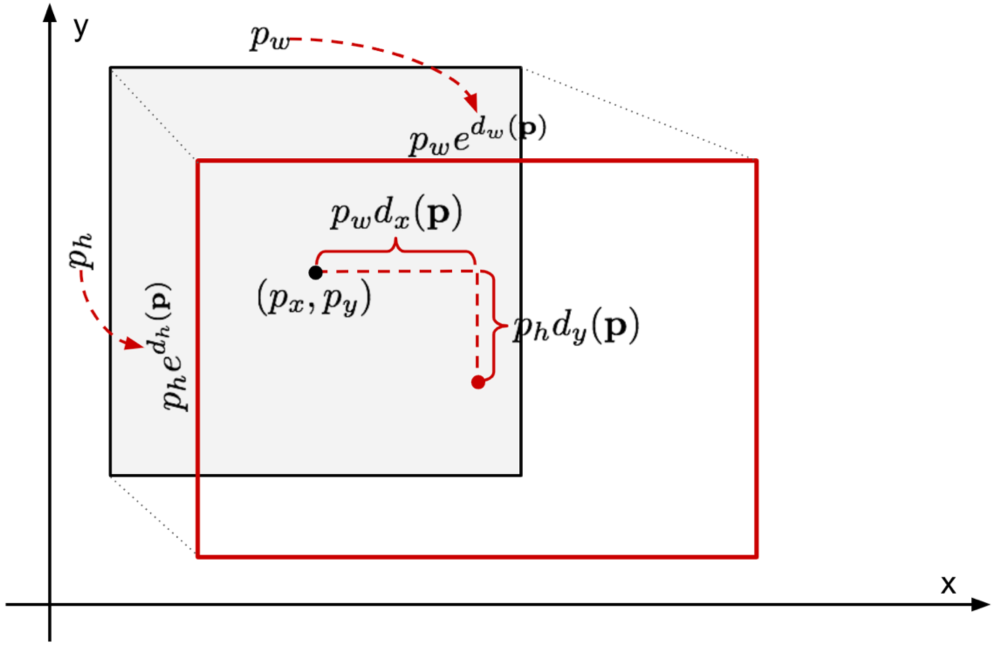
Image Source: Ren Shaoqing Ren et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." Neurips 2015.

Tracktor – Base Architecture (FPN Variant of Faster R-CNN)

Input:
Video



Object Detection: Refine Previous Frame Using Regression Head on Faster R-CNN with FPN



Original region proposal with center (p_x, p_y), width (p_w), and height (p_h) is refined using model parameters (d_x, d_y, d_w, d_y)

Output:
Trajectories

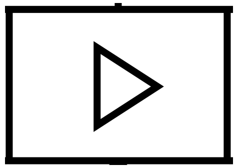
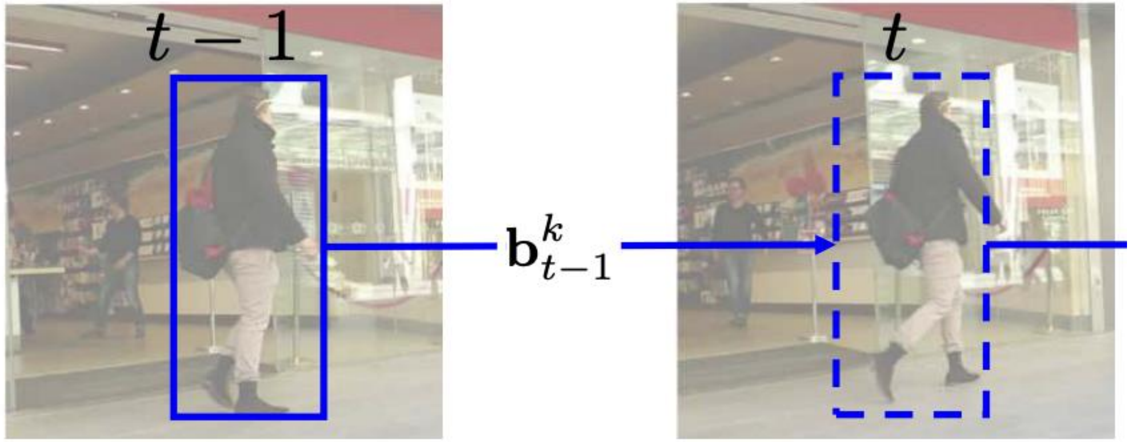


Image Source: https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html#bounding-box-regression

Tracktor – Base Architecture (FPN Variant of Faster R-CNN)

Input:
Video

Object Detection: Refine Previous Frame Using Regression Head on Faster R-CNN with FPN



Original region proposal with center (p_x, p_y), width (p_w), and height (p_h) is refined using model parameters (d_x, d_y, d_w, d_y)

Output:
Trajectories

Image Source: Bergmann, Meinhardt, and Leal-Taixe. "Tracking Without Bells and Whistles. CVPR 2019.

Tracktor – Base Architecture (FPN Variant of Faster R-CNN)

Input:
Video

Object Detection: Refine Previous Frame Using Regression Head on Faster R-CNN with FPN

+

Post-processing: Initialize new objects entering the video for any detections with low IoU with existing active tracks

+

Post-processing: “Kill” tracked object if the predicted classification indicating an object is present falls below a pre-defined threshold

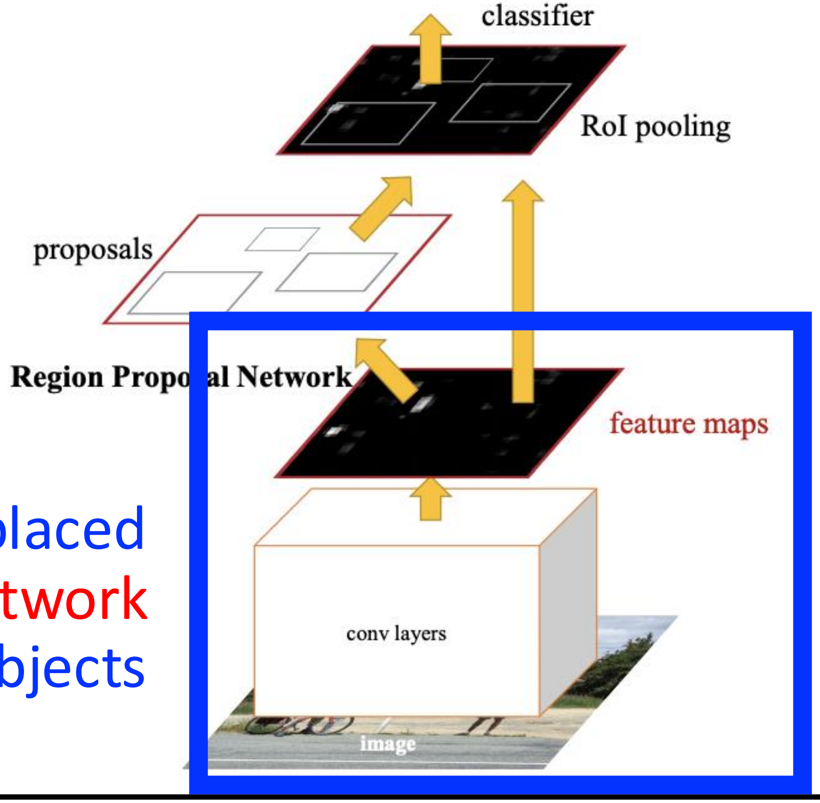
Output:
Trajectories

Image Source: https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html#bounding-box-regression

Tracktor – Base Architecture (FPN Variant of Faster R-CNN)

Object Detection: Refine Previous Frame Using Regression Head on Faster R-CNN with FPN

Input:
Video



Output:
Trajectories

Single-scale feature map replaced with the feature pyramid network towards detecting smaller objects

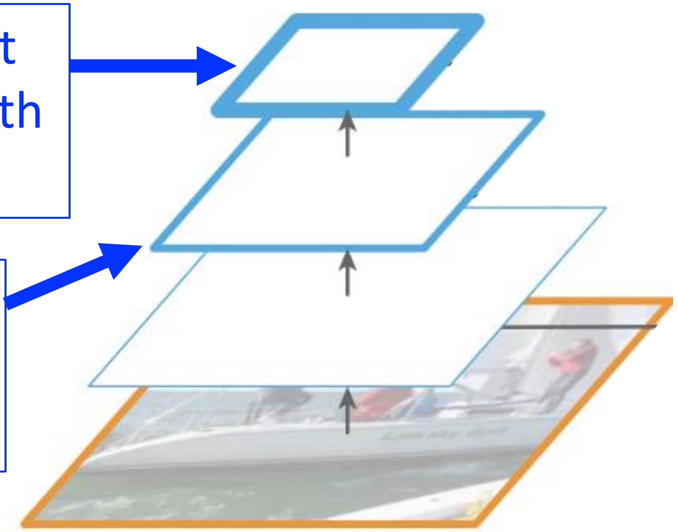
Image Source: Ren Shaoqing Ren et al. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.” Neurips 2015.

Tracktor – Base Architecture (FPN Variant of Faster R-CNN)

Step 1. Compute hierarchy of feature maps at several scales with your favorite backbone architecture (e.g., ResNet)

(Feature Pyramid Network)

Higher layers represent **coarser resolutions** with **stronger semantics**



Lower layers represent **higher resolutions** with **weaker semantics**

Figure source: Tsung-Yi Lin et al. “Feature Pyramid Networks for Object Detection.” CVPR 2017.

Tracktor – Base Architecture (FPN Variant of Faster R-CNN)

Step 1. Compute hierarchy of feature maps at several scales with your favorite backbone architecture (e.g., ResNet)

(Feature Pyramid Network)

Step 2. Fuse semantically stronger, coarser resolution feature maps with higher resolution, semantically weak features maps by upsampling the coarser resolution feature maps

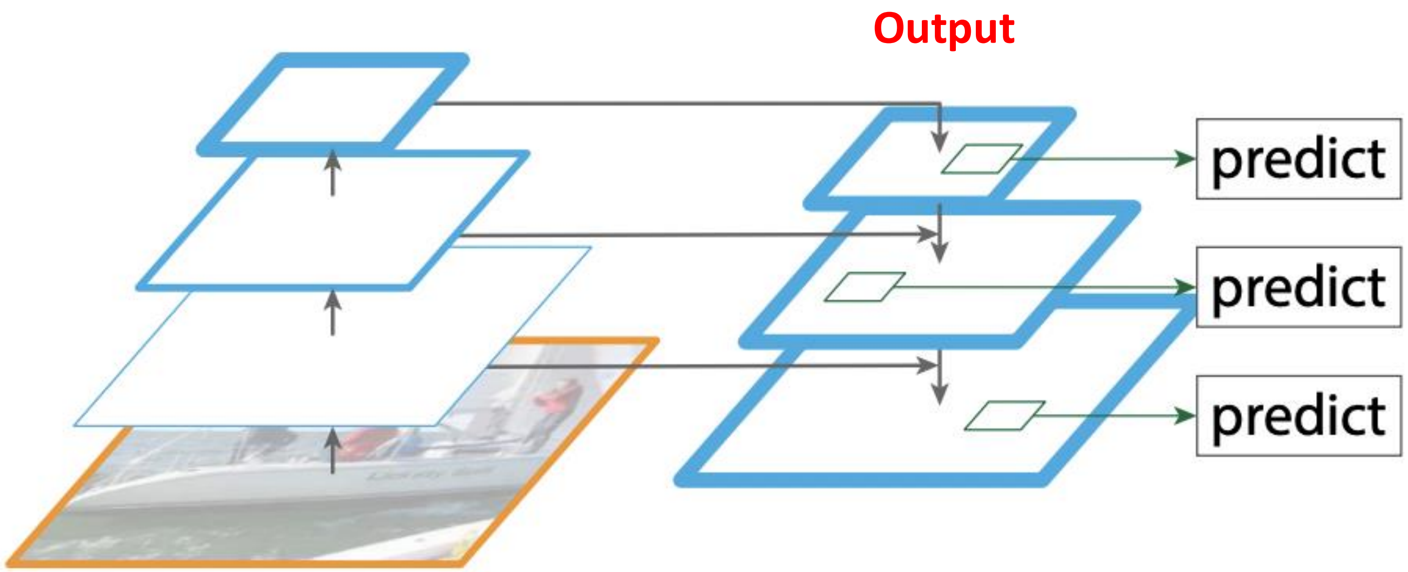


Figure source: Tsung-Yi Lin et al. "Feature Pyramid Networks for Object Detection." CVPR 2017.

Tracktor – Base Architecture (FPN Variant of Faster R-CNN)

- Why not use **image pyramids?** (i.e., convert an image into multiple scales and then extract a semantically strong feature for each scale)
  - Relatively slow at test time (must test an image at every scale)

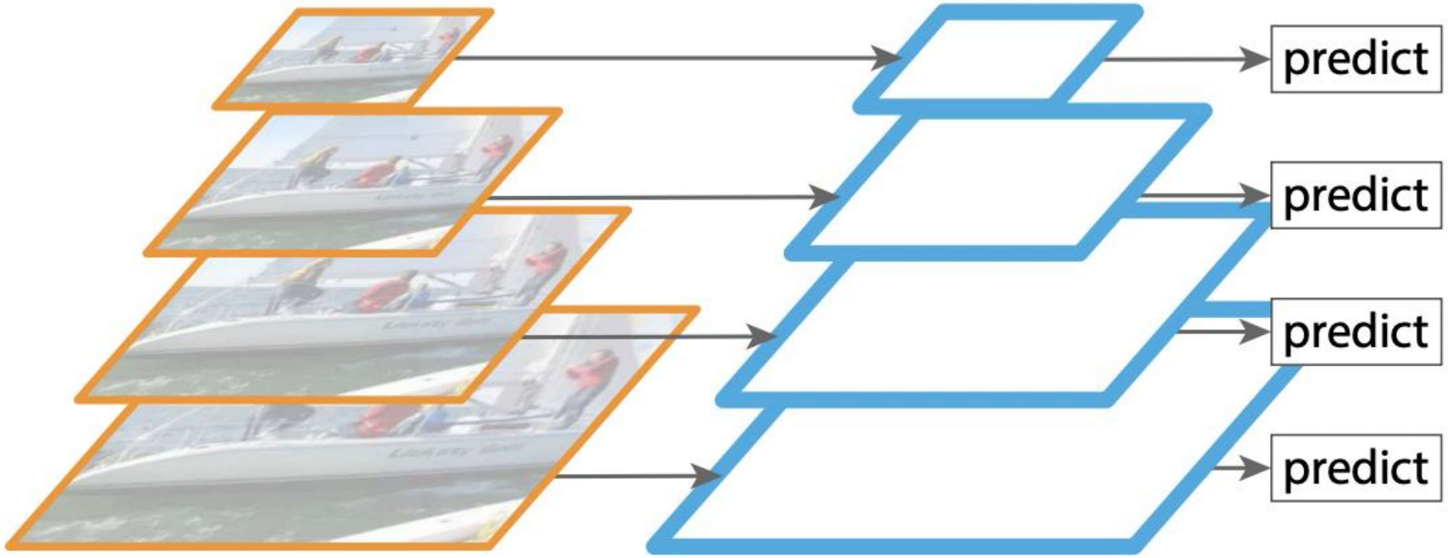


Figure source: Tsung-Yi Lin et al. “Feature Pyramid Networks for Object Detection.” CVPR 2017.

Tracktor++ (i.e., with *More* Post-Processing)

1. **Motion model**: for objects' considerably changing positions between frames

    - Low frame rate: assume constant velocity for all objects

    - Moving camera: apply image registration

2. **Reidentification:** accounts for linking an object that disappears for a short time to itself when it re-appears

    - Compare appearance similarity of killed objects to newly tracked objects

Tracktor++ Performance

State-of-art performance on three datasets with respect to MOTA!

	Method	MOTA ↑
MOT17	Tracktor++	**53.5**
	eHAF [58]	51.8
	FWT [23]	51.3
	jCC [30]	51.2
	MOTDT17 [9]	50.9
	MHT_DAM [32]	50.7
MOT16	Tracktor++	**54.4**
	HCC [44]	49.3
	LMP [59]	48.8
	GCRA [43]	48.2
	FWT [23]	47.8
	MOTDT [9]	47.6
2D MOT 2015	Tracktor++	**44.1**
	AP_HWDPL_p [8]	38.5
	AMIR15 [56]	37.6
	JointMC [30]	35.6
	RAR15pub [17]	35.1

Image Source: Bergmann, Meinhardt, and Leal-Taixe. "Tracking Without Bells and Whistles. CVPR 2019.

Ablation Study of Tracktor++

- Test set: MOT17 which consists of 7 sequences

Method	MOTA ↑
D&T [18]	50.1
Tracktor-no-FPN	57.4
Tracktor	61.5
Tracktor+reID	61.5
Tracktor+CMC	**61.9**
Tracktor++ (reID + CMC)	**61.9**

Greatest boost in performance comes from using a feature pyramid network

Remainder of performance boost stems from the motion model

Image Source: Bergmann, Meinhardt, and Leal-Taixe. "Tracking Without Bells and Whistles. CVPR 2019.

Tracktor++ Weaknesses

- When targets have diminished visibility (i.e., from occlusion)
- When objects are small
- When there is a large gap for a tracked object (i.e., missed detections)

SAM-2:
Semantics-Agnostic, Semi-Automated Tracking

Achieves state-of-the-art performance for video object segmentation, when specifying at the first frame what to track (e.g., click, box, mask)

Demo: https://sam2.metademolab.com/

Today's focus for the programming tutorial!

Ravi et al. SAM 2: Segment Anything in Images and Videos. arXiv 2024

SAM-2:
Semantics-Agnostic, Semi-Automated Tracking

Perspective:
What would it cost to annotate 12,000 1-minute videos (i.e., 200 hours),
with 6 frames sampled per second and 30 seconds to annotate each frame?

Key idea: huge training dataset (all videos and annotations from crowdworkers)!

Ravi et al. SAM 2: Segment Anything in Images and Videos. arXiv 2024

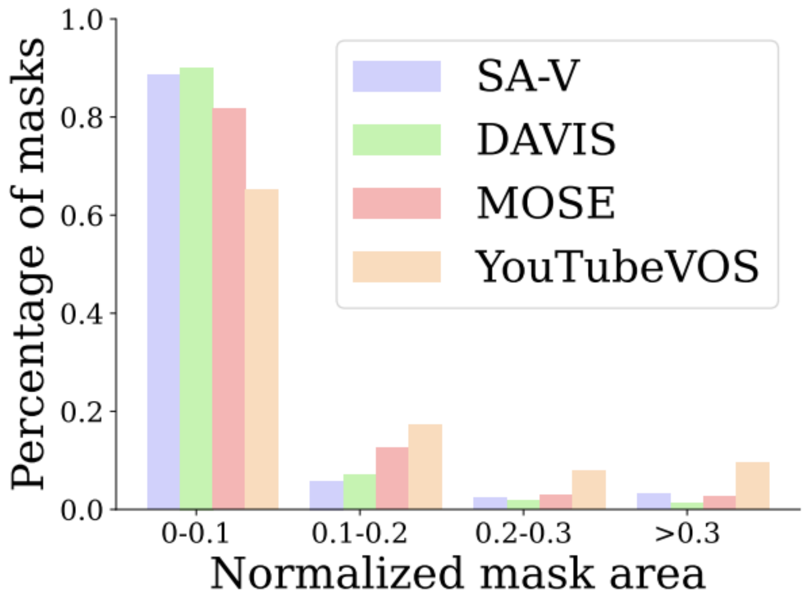
SAM-2:
Semantics-Agnostic, Semi-Automated Tracking

	#Videos	Duration	#Masklets	#Masks	#Frames
DAVIS 2017 (Pont-Tuset et al., 2017)	0.2K	0.1 hr	0.4K	27.1K	10.7K
YouTube-VOS (Xu et al., 2018b)	4.5K	5.6 hr	8.6K	197.3K	123.3K
UVO-dense (Wang et al., 2021b)	1.0K	0.9 hr	10.2K	667.1K	68.3K
VOST (Tokmakov et al., 2022)	0.7K	4.2 hr	1.5K	175.0K	75.5K
BURST (Athar et al., 2022)	2.9K	28.9 hr	16.1K	600.2K	195.7K
MOSE (Ding et al., 2023)	2.1K	7.4 hr	5.2K	431.7K	638.8K
Internal	62.9K	281.8 hr	69.6K	5.4M	6.0M
SA-V Manual	50.9K	196.0 hr	190.9K	10.0M	4.2M
SA-V Manual+Auto	50.9K	196.0 hr	642.6K	35.5M	4.2M

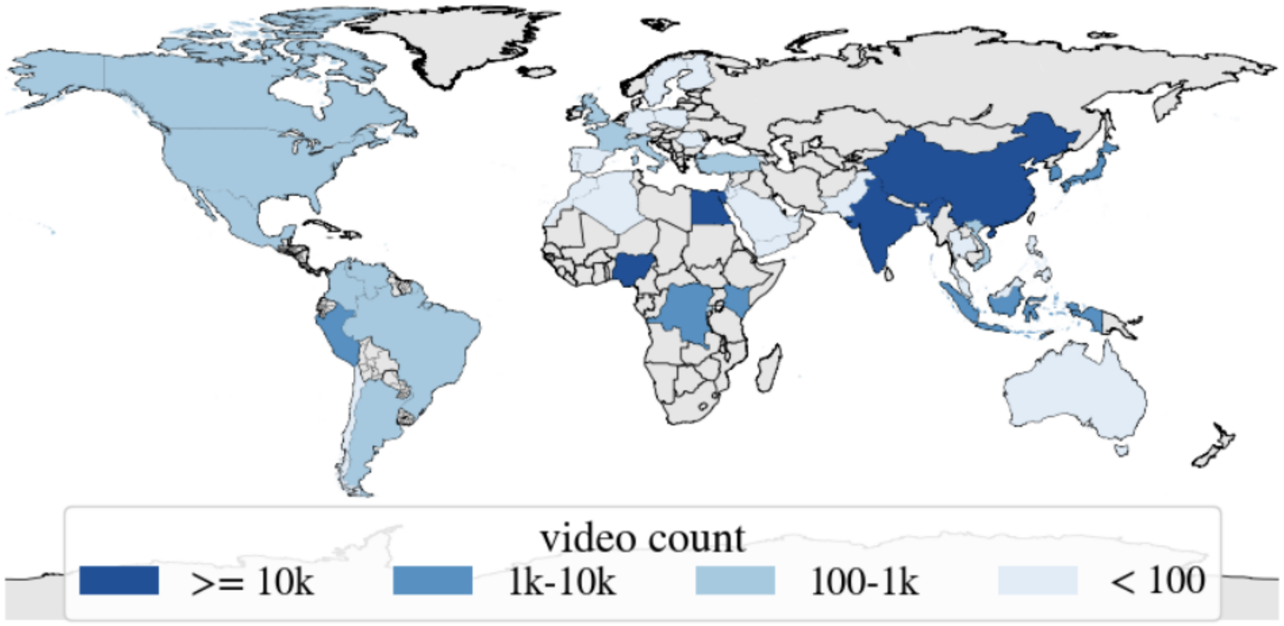
Key idea: huge training dataset (all videos and annotations from crowdworkers)!

Ravi et al. SAM 2: Segment Anything in Images and Videos. arXiv 2024

SAM-2:
Semantics-Agnostic, Semi-Automated Tracking



(a) Size

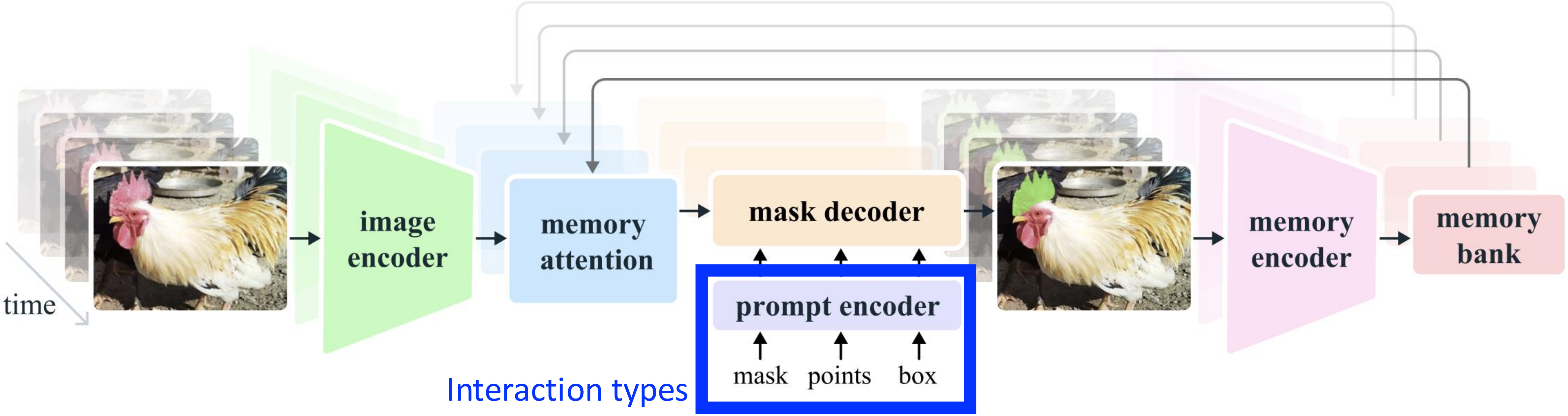


(b) Geography

Masks tend to occupy 10% or less of frames for videos from around the world

Ravi et al. SAM 2: Segment Anything in Images and Videos. arXiv 2024

SAM-2:
Semantics-Agnostic, Semi-Automated Tracking



Interaction types

Architecture extends SAM model with memory to retain tracking information from previous frames

Ravi et al. SAM 2: Segment Anything in Images and Videos. arXiv 2024

Object Tracking: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models
- Discussion (chosen by YOU ☺)

Object Tracking: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models
- Discussion (chosen by YOU ☺)

The End