Object Tracking

Danna Gurari
University of Colorado Boulder
Fall 2021

https://home.cs.colorado.edu/~DrG/Courses/RecentAdvancesInComputerVision/AboutCourse.html

Review

- Last lecture:
  - Semantic segmentation problem
  - Semantic segmentation applications
  - Semantic segmentation datasets
  - Semantic segmentation evaluation metrics
  - Computer vision models: fully convolutional networks
- Assignments (Canvas)
  - Reading assignment due earlier today
  - Two reading assignments out that are due next Monday and Wednesday
- Questions?

Object Tracking: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models

Object Tracking: Today's Topics

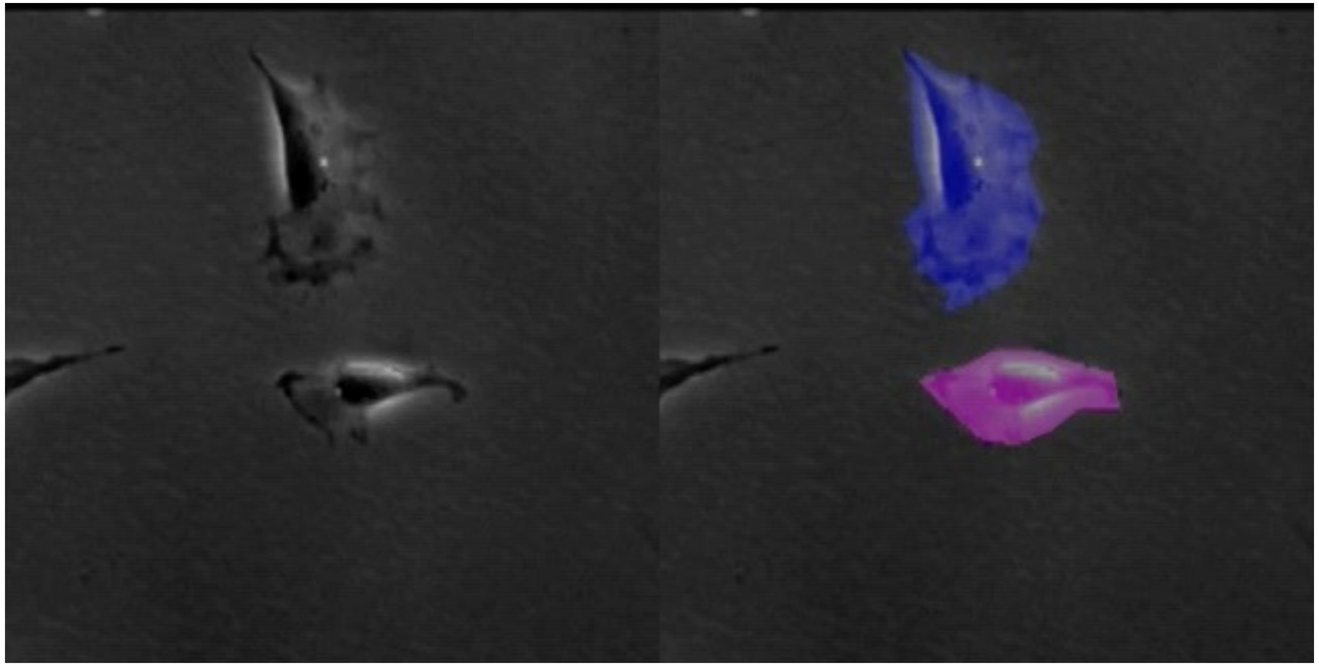
- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models

Definition

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

Input

Output masks overlaid on video



Definition

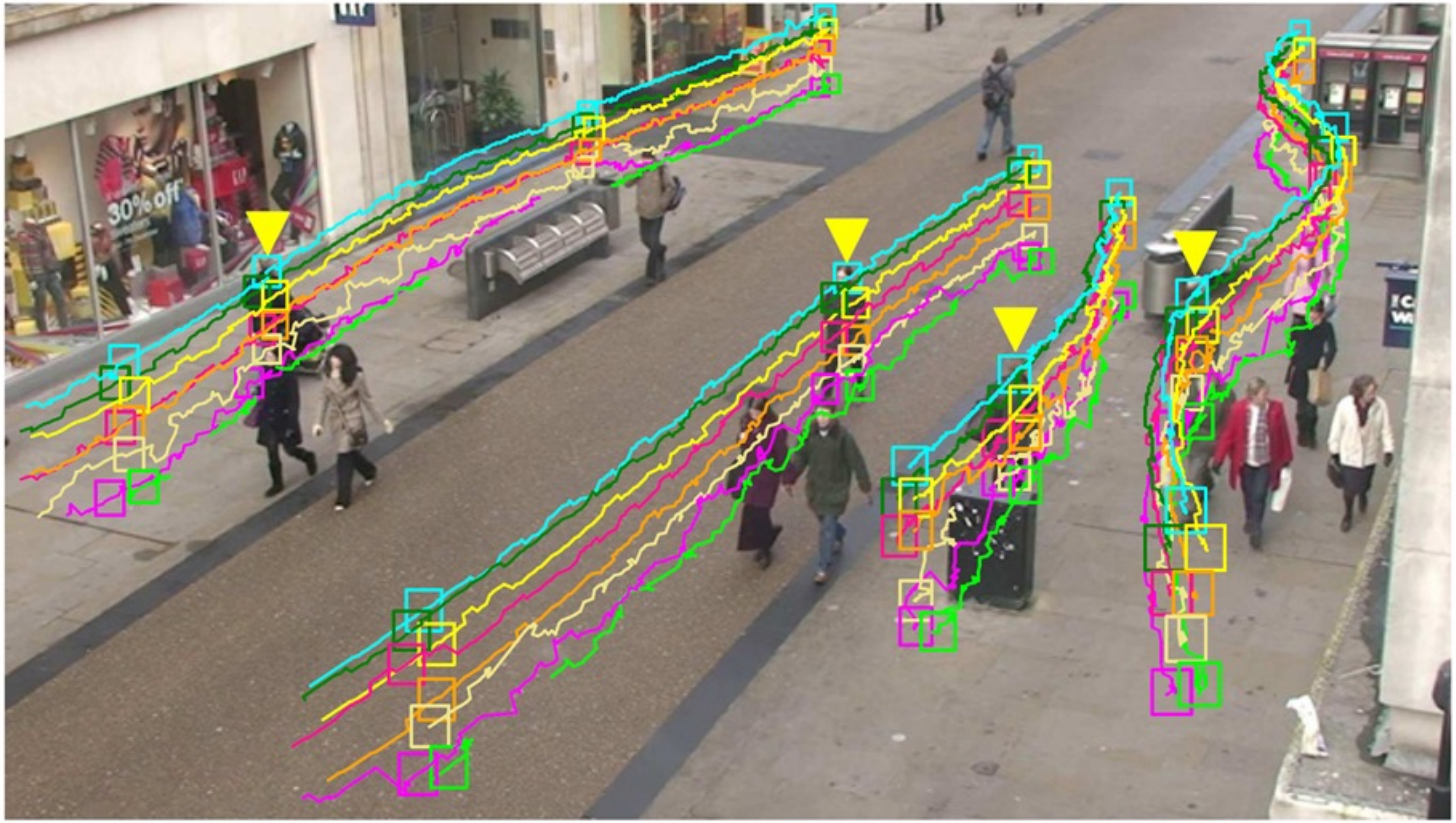
- Identification of the trajectory of an object over time
  - Single object
  - Multiple objects

- 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

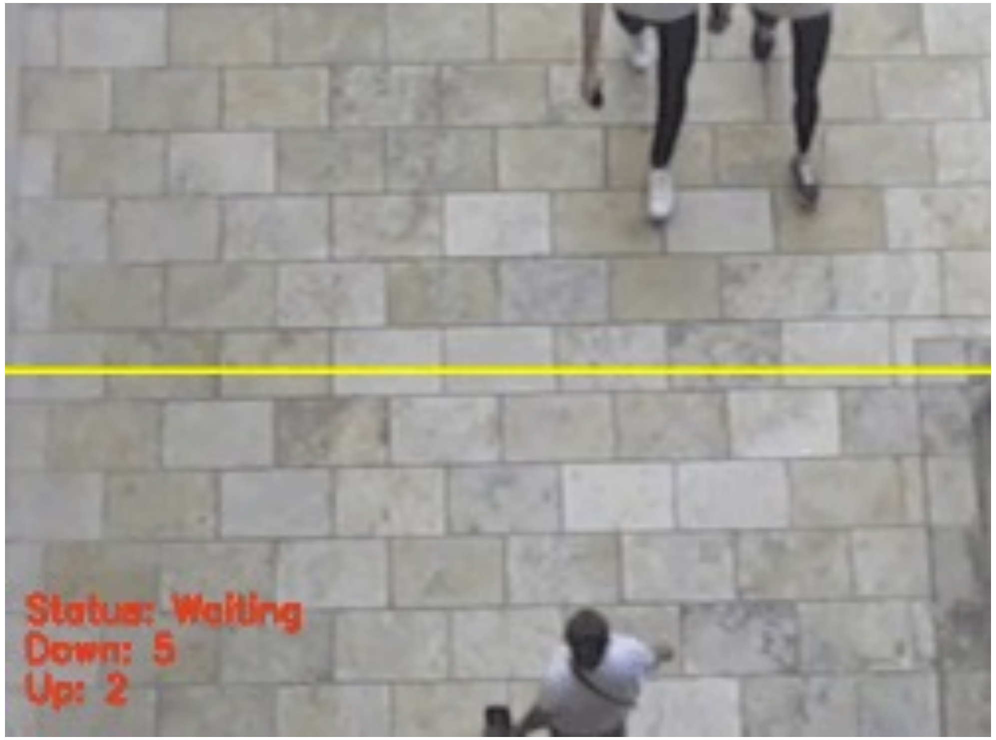
- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models

Surveillance



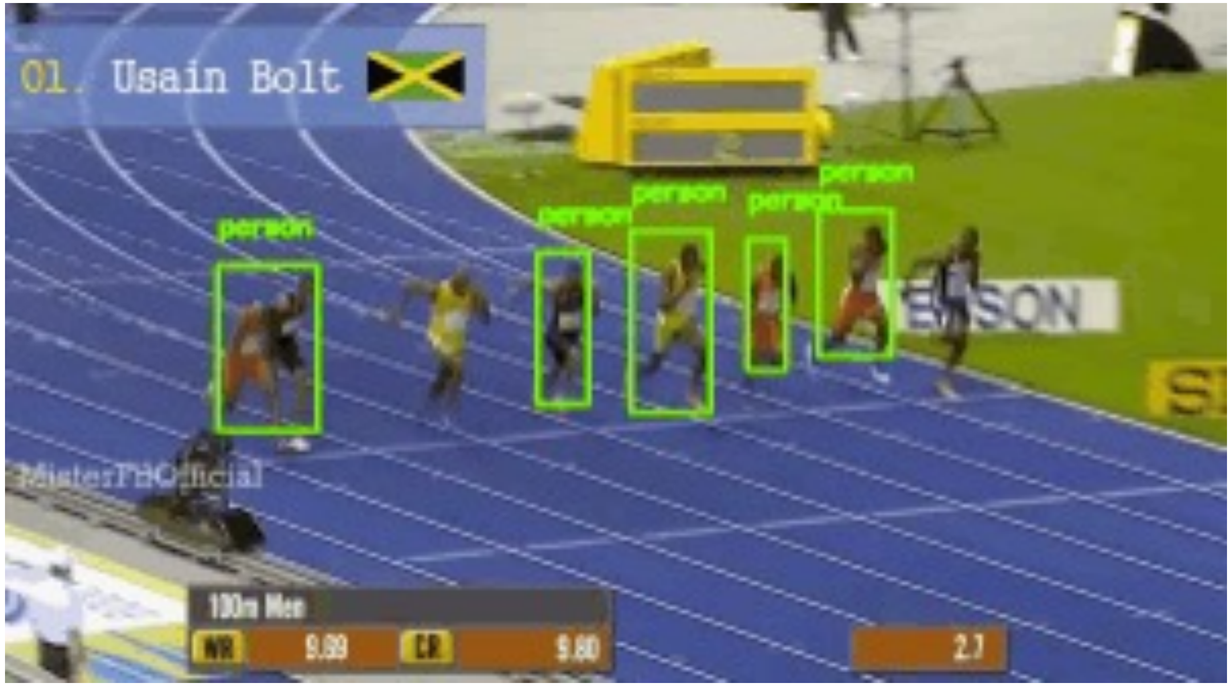
http://www.vision.caltech.edu/koller/MOU-83.html

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/

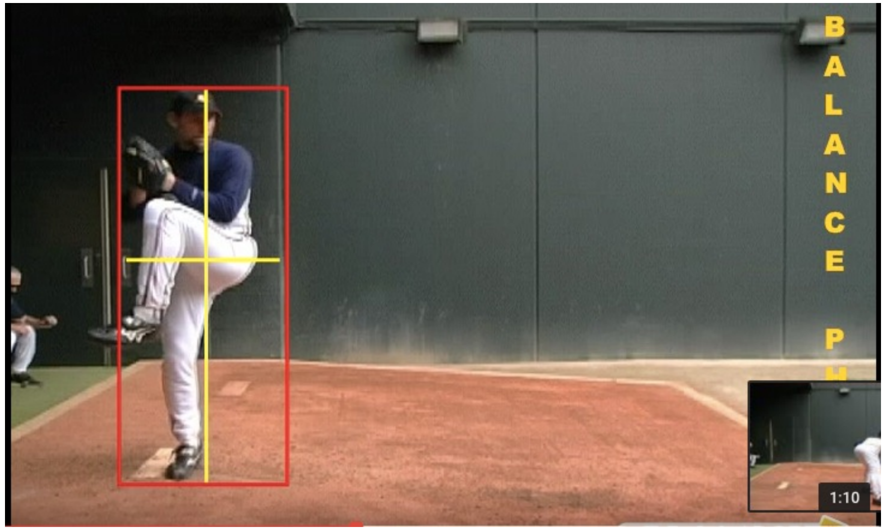
Sports Performance Analytics

Calculate Bat speed from video!

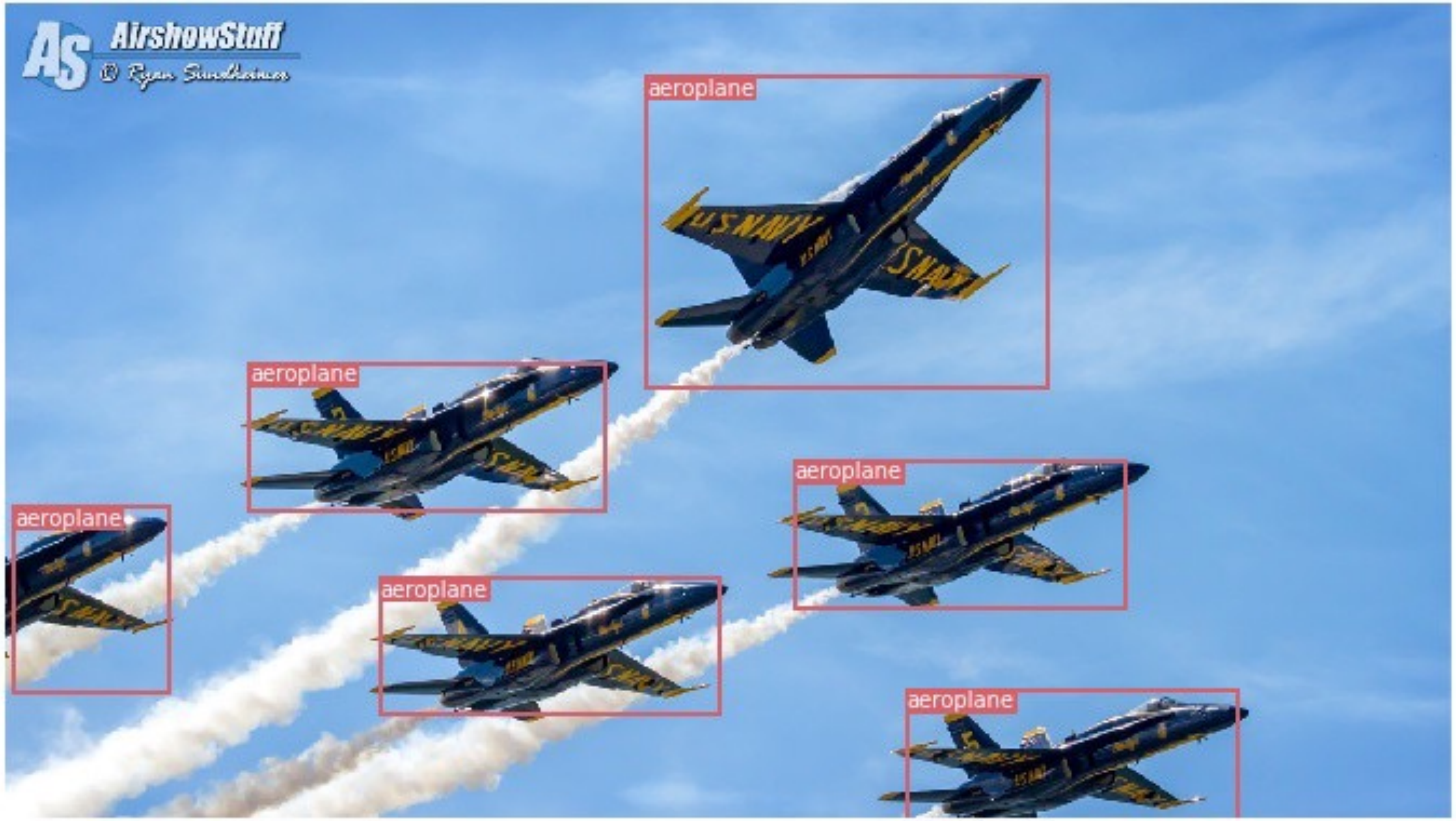
NEW! Track Bowling Ball Path!

Works great for putting!

http://www.motionprosoftware.com/



Military Defense



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

Self-driving Cars

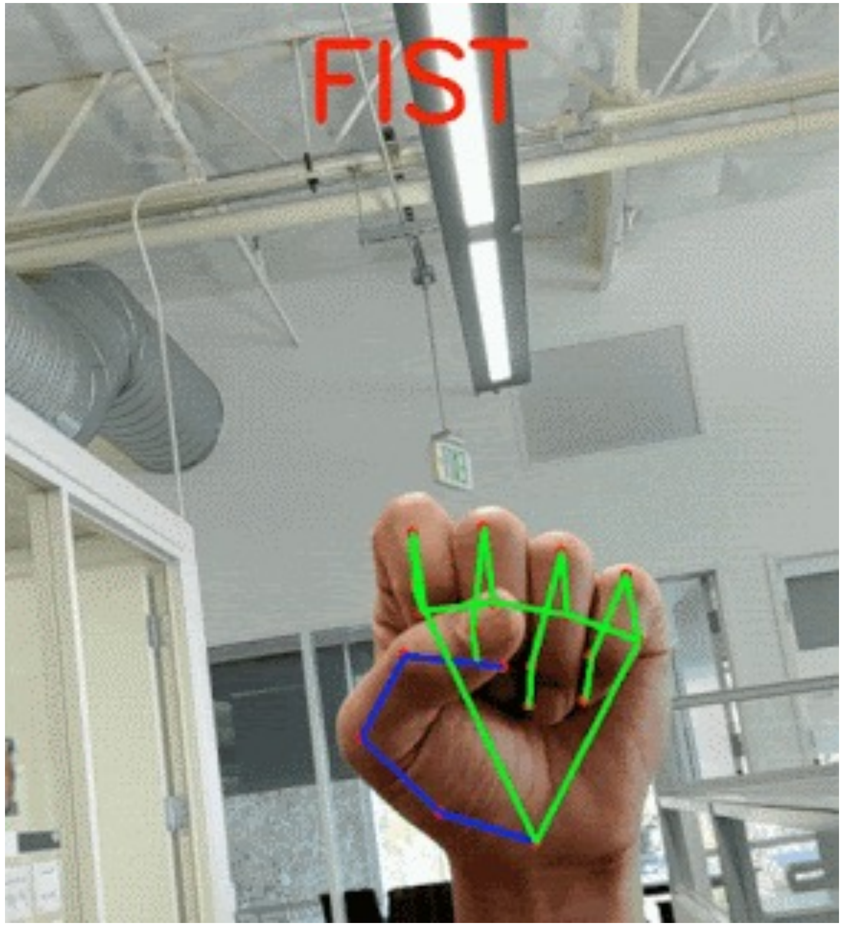
https://www.pyimagesearch.com/2018/08/06/tracking-multiple-objects-with-opencv/

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:

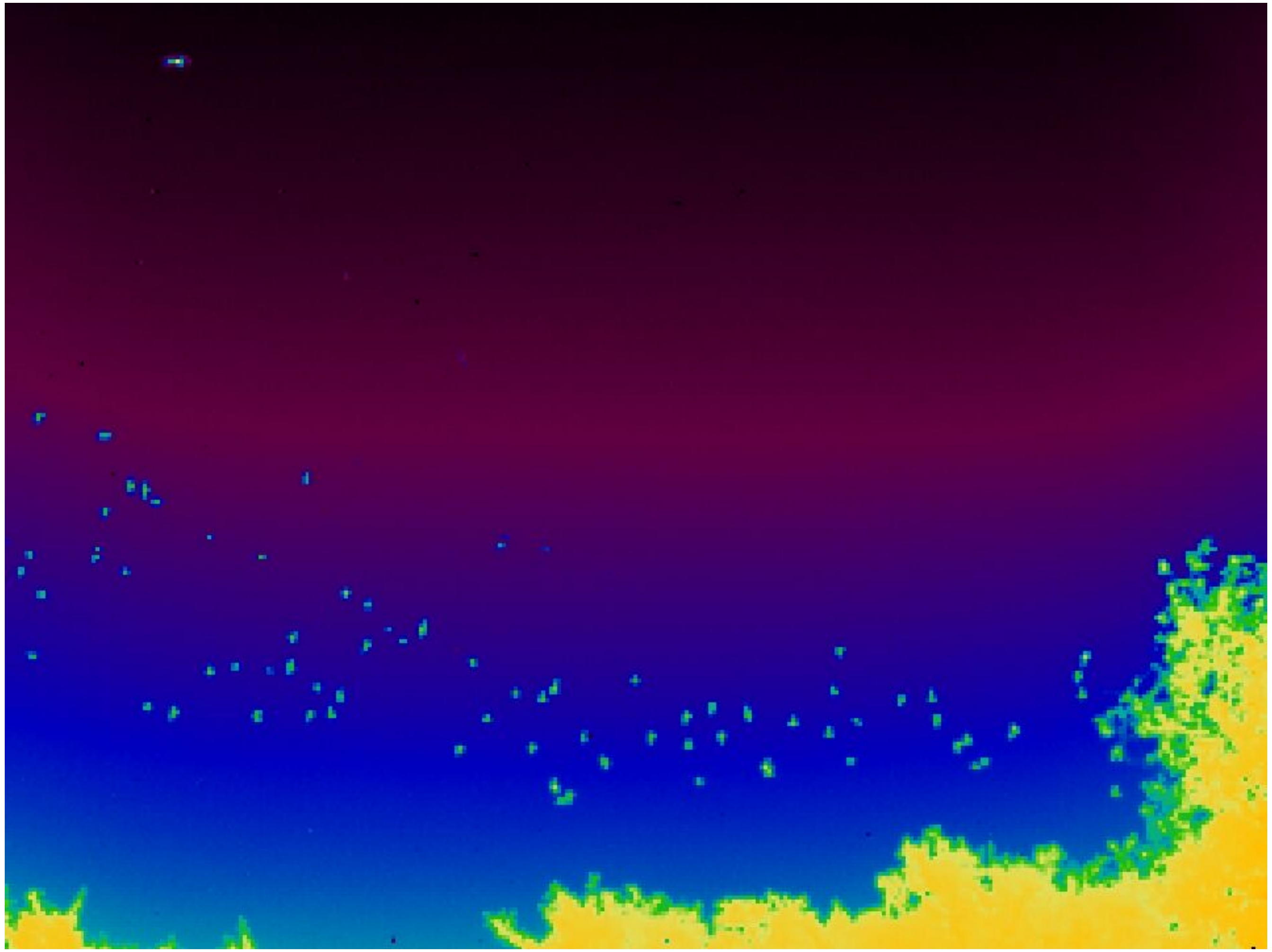


Image source: https://www.cs.bu.edu/fac/betke/research/bats/images2.html

Augmented Reality

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

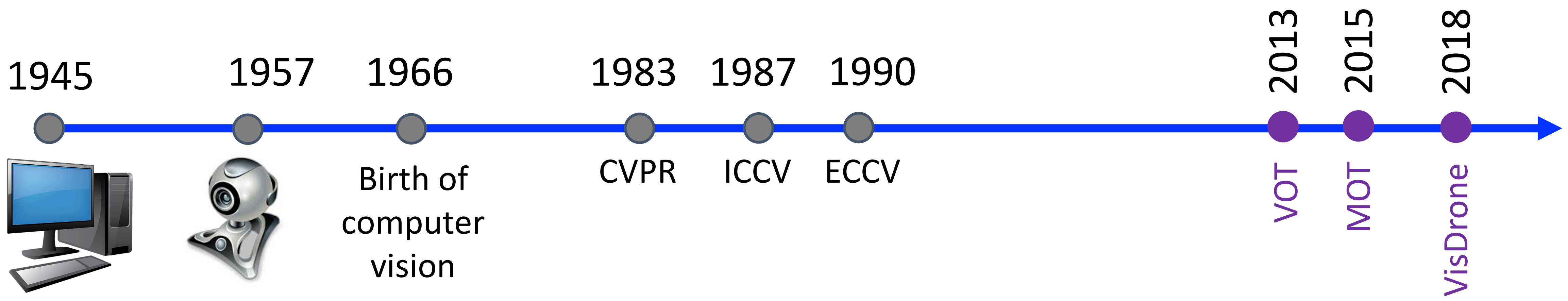
Applications

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

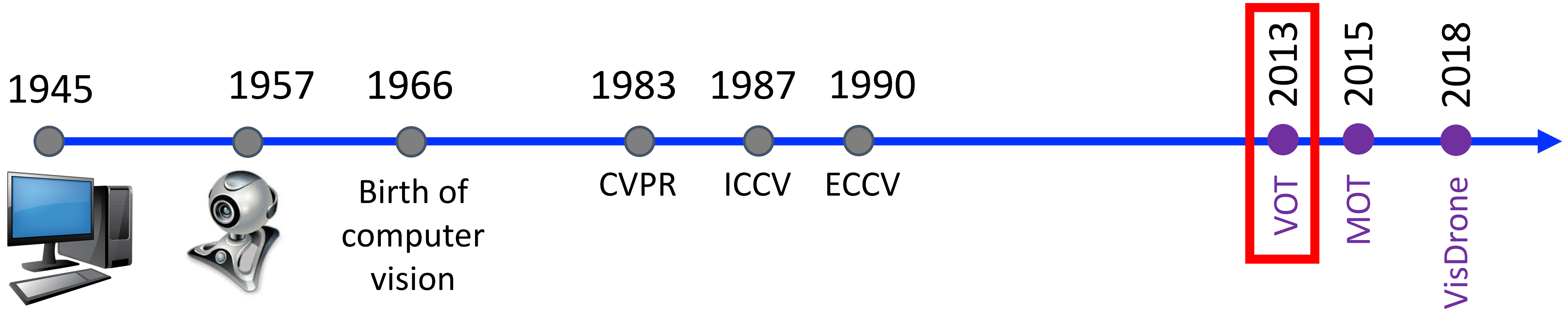
Object Tracking: Today’s Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models

Object Tracking Datasets



Object Tracking Datasets



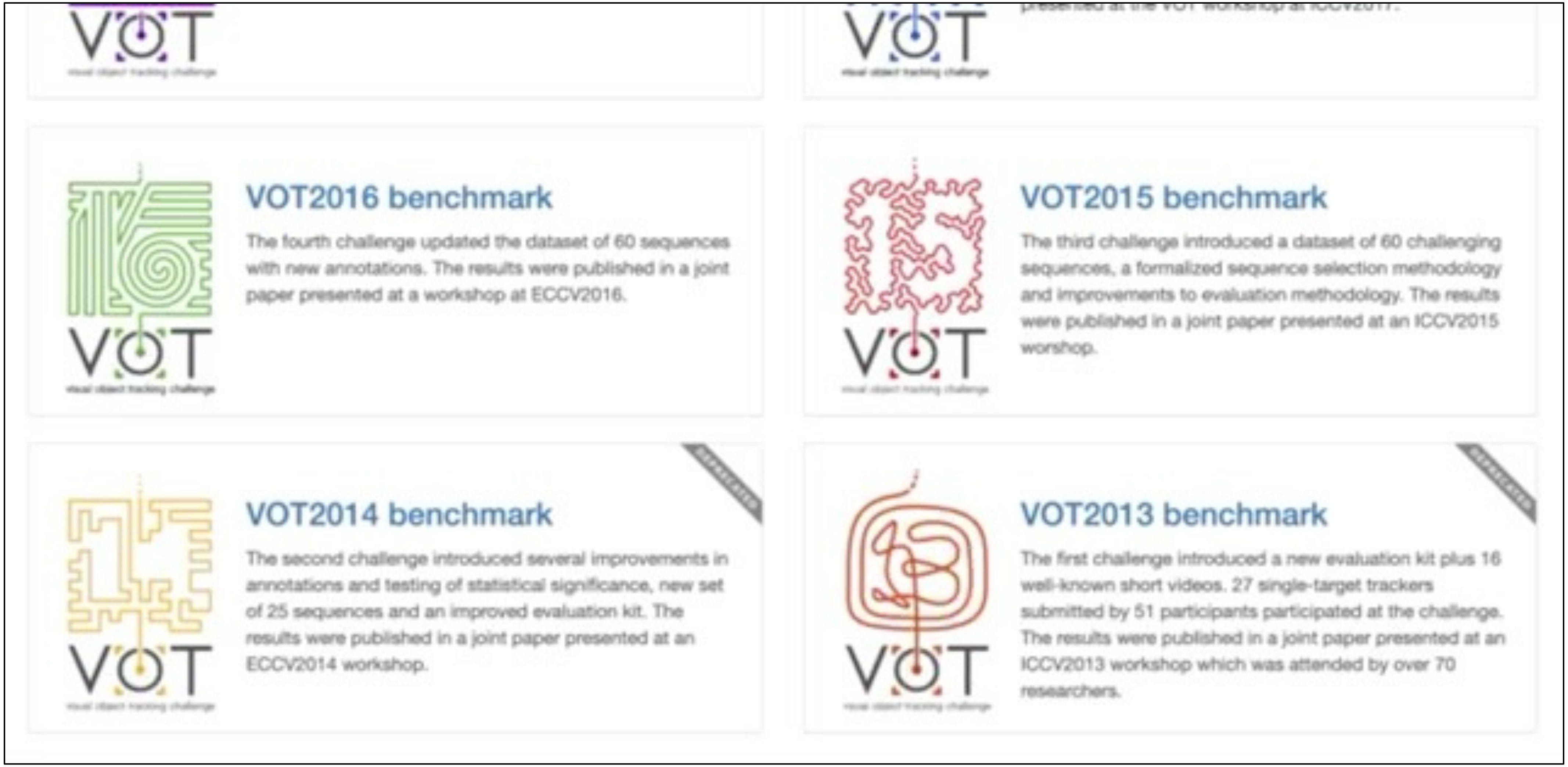
Single Object Tracking Dataset: VOT

- 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 deemed to have unsuitable annotations

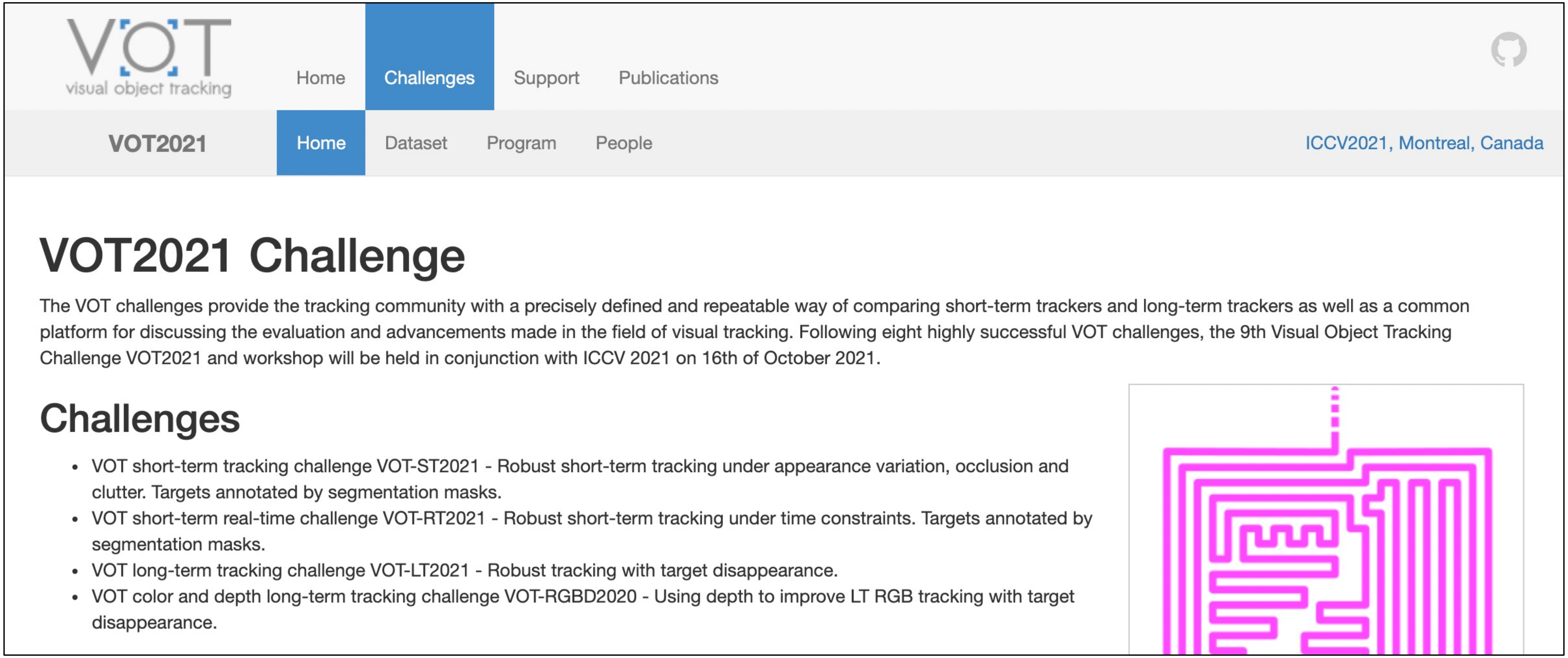
Matej Kristan et al. The Visual Object Tracking VOT2013 Results. ICCV Workshop 2013.

Single Object Tracking Dataset: VOT's Evolution



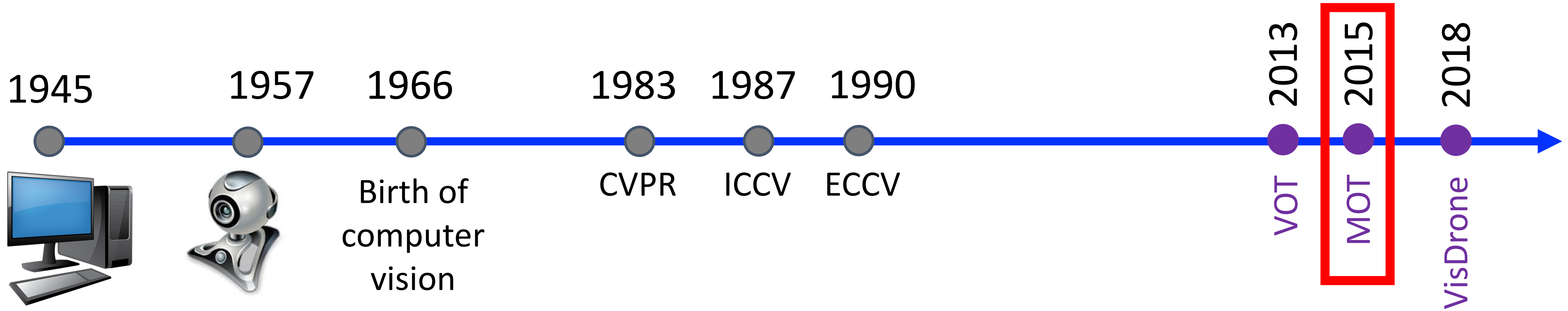
Matej Kristan et al. The Visual Object Tracking VOT2013 Results. ICCV Workshop 2013.

Single Object Tracking Annual Challenge (9th year now)



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 versus on the ground looking up)
  - **Multiple weather conditions**; e.g., sunny versus cloudy versus night time
- 16 of the videos came from existing datasets while the other 6 were generated by the authors; tracked objects were people and vehicles

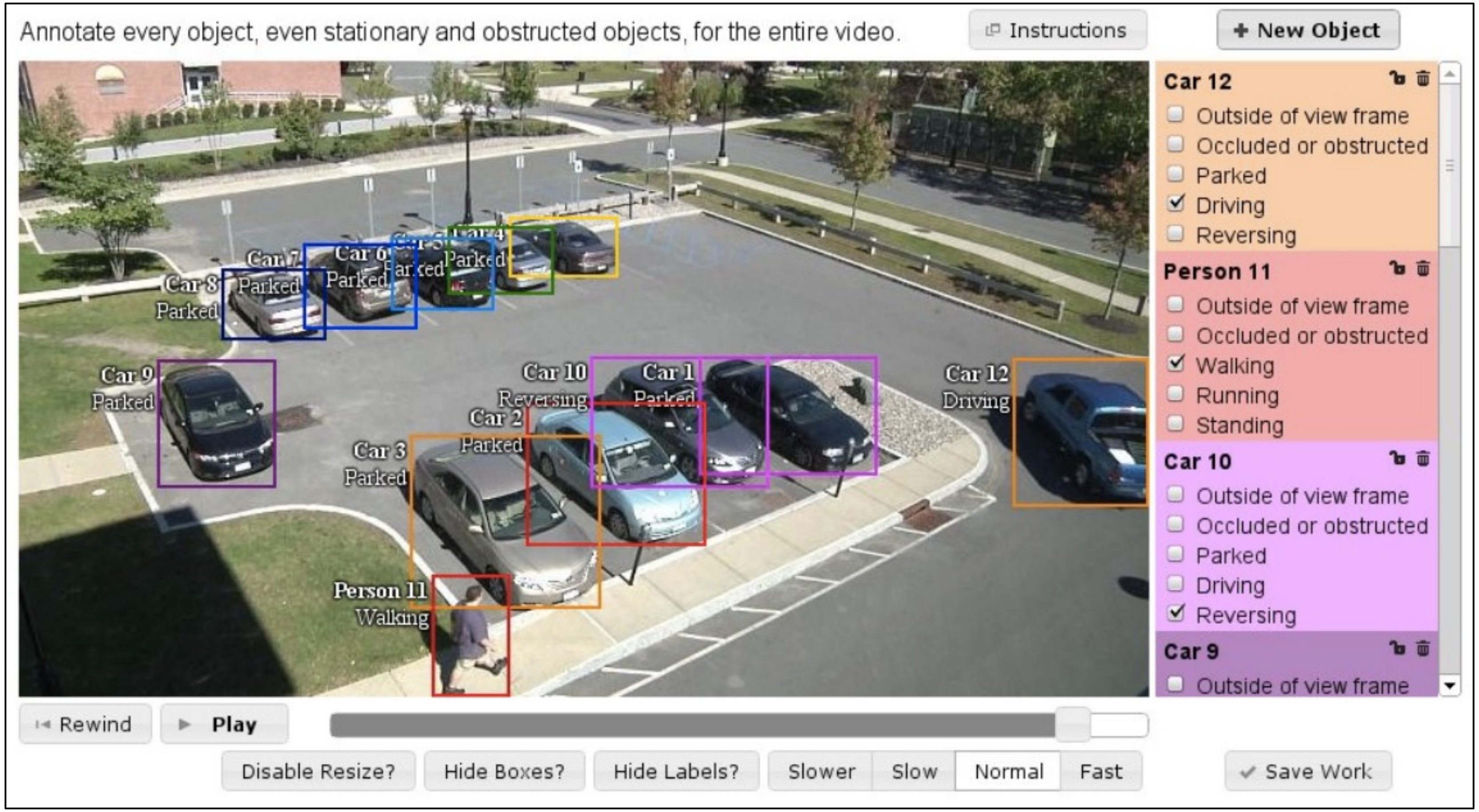
Laura Leal-Taixe, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. arXiv 2015.

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 versus 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.

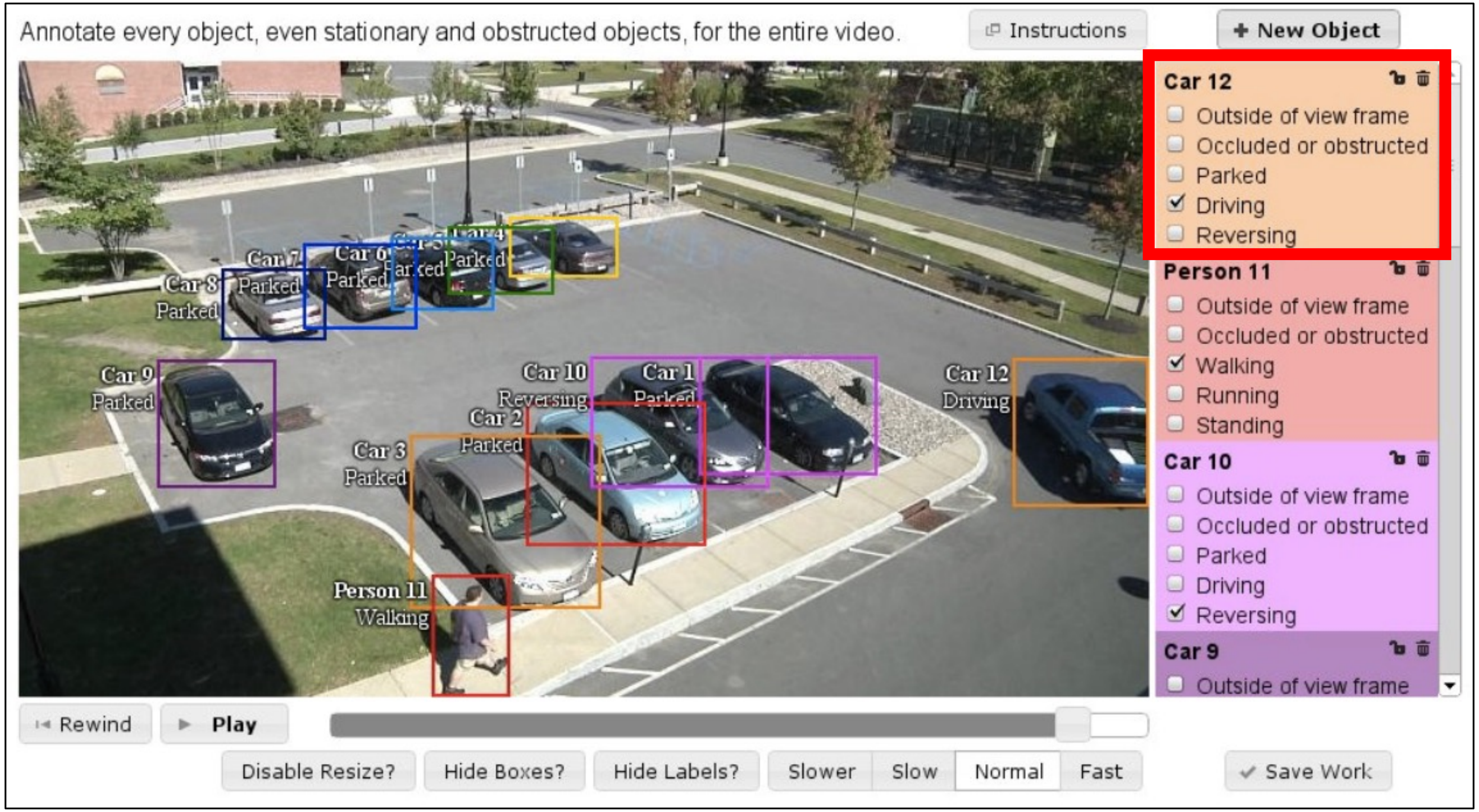
Multiple Object Tracking Annotation: VATIC



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

Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently Scaling Up Crowdsourced Video Annotation: A Set of Best Practices for High Quality, Economical Video Labeling. IJCV 2012.

Multiple Object Tracking Annotation: VATIC



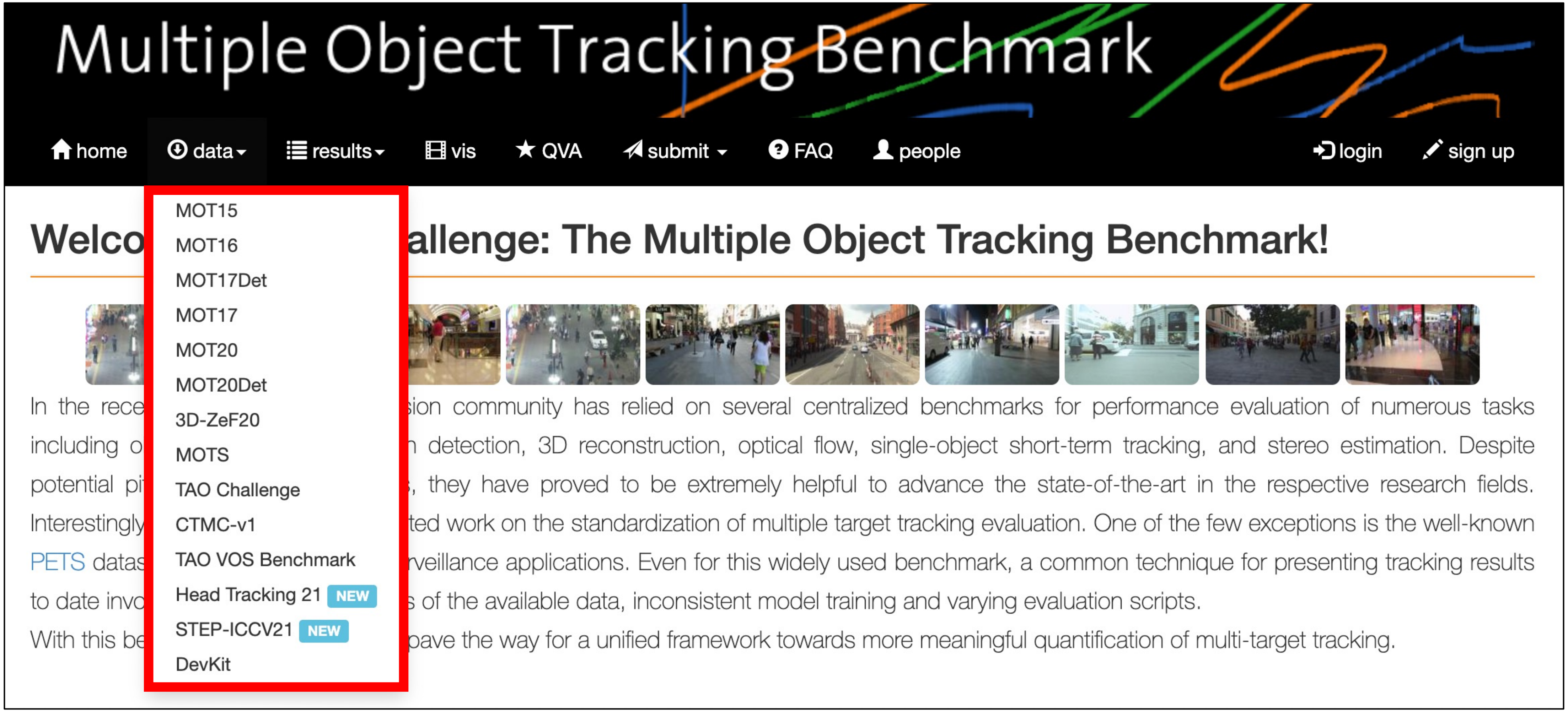
Metadata about each object: e.g., activity, attributes, etc.

Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently Scaling Up Crowdsourced Video Annotation: A Set of Best Practices for High Quality, Economical Video Labeling. IJCV 2012.

Multiple Object Tracking Annotation: VATIC

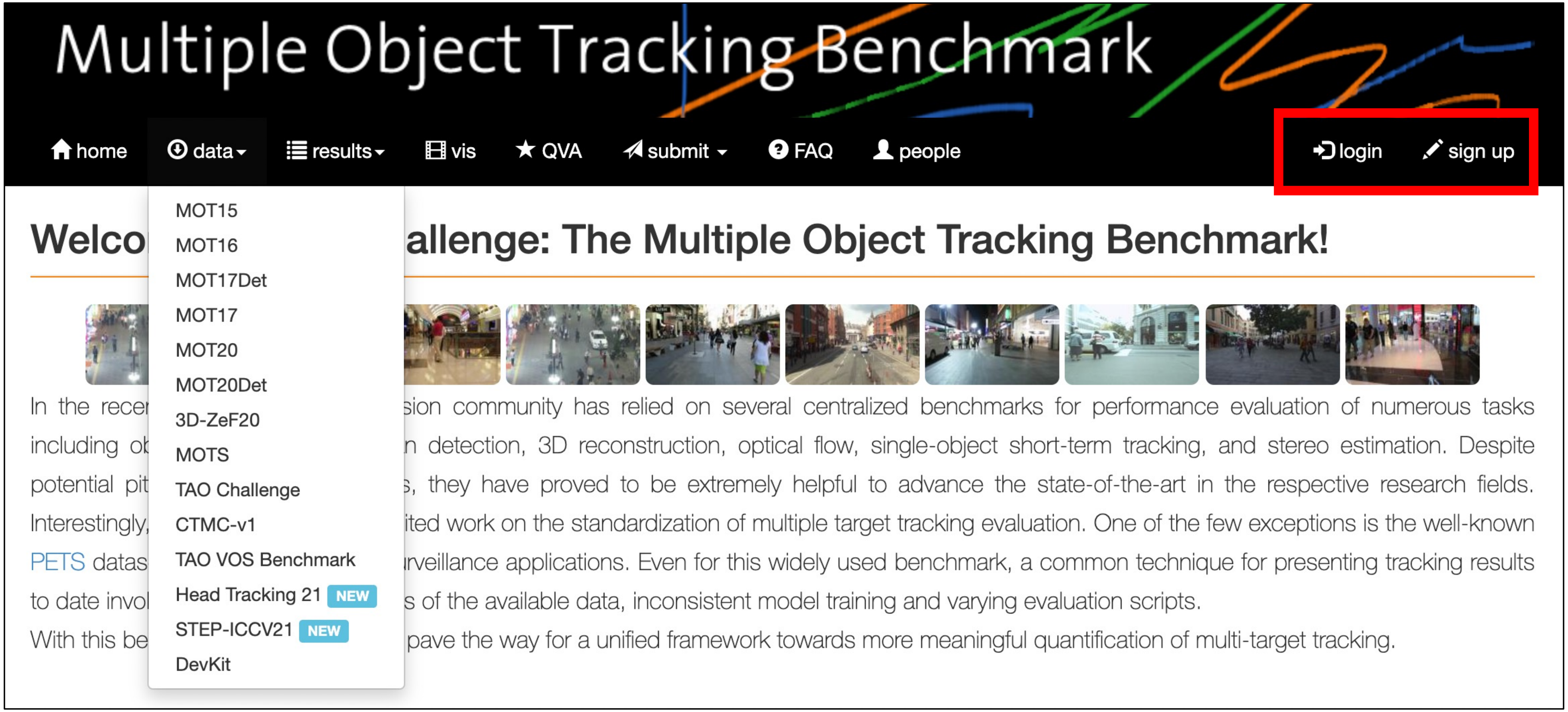
- How to handle occlusions?
  - Annotation instructions: "*Always annotate during occlusions if the position can be determined unambiguously. If the occlusion is very long and it is not possible to determine the path of the object using simple reasoning (e.g. constant velocity assumption), the object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

Single Object Tracking Dataset: MOT's Evolution



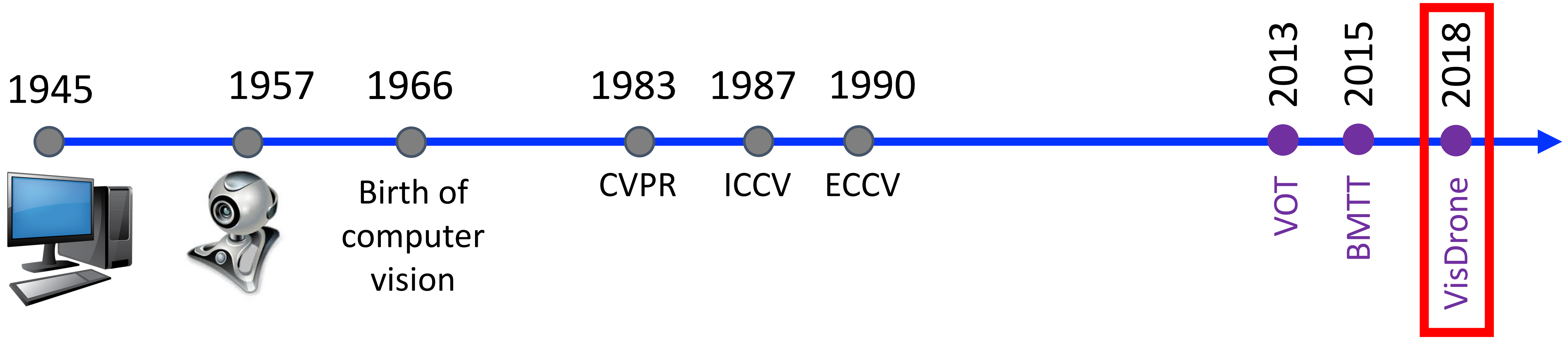
https://motchallenge.net/

Single Object Tracking Annual Challenge (7^{th} year now)



https://motchallenge.net/

Object Tracking Datasets



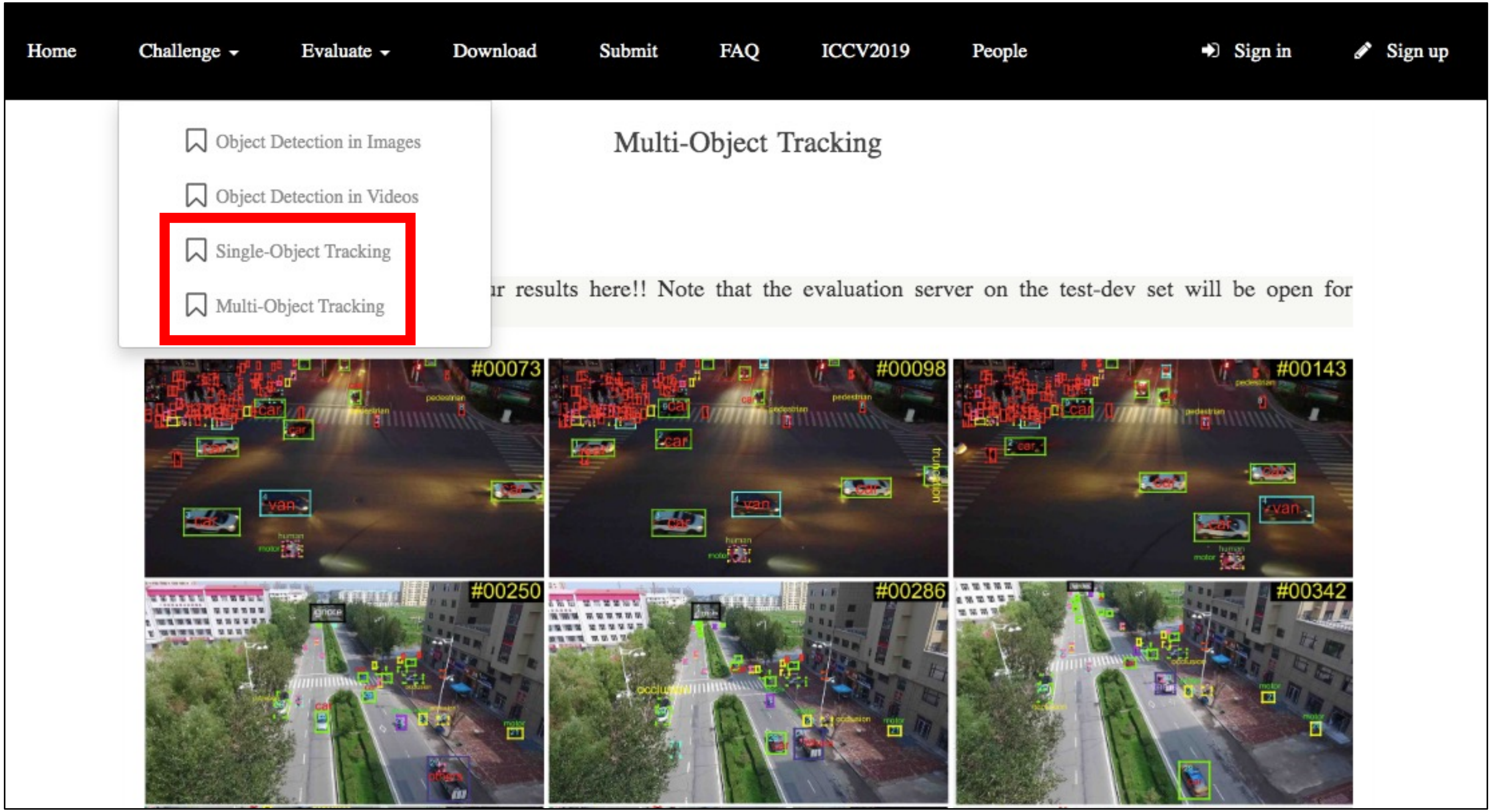
VisDrone

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

- Annotations created for over 2.5 million object instances, however it is unspecified how these 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 to collect **high quality** object tracking annotations, how should these scenarios be handled:
  - Partially visible object
  - Occluded object
  - Object is reflected in reflective surfaces such as mirrors or windows

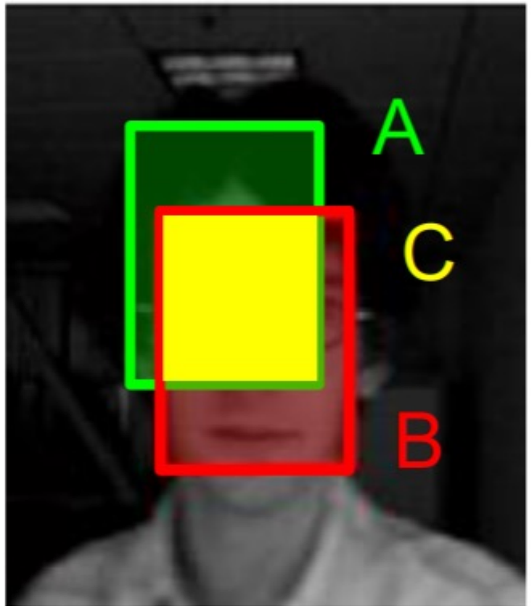
- What will be the total crowdsourcing task cost to annotate 1,000 1-minute videos where you need to track 5 humans/video (assume 30 frames/second)?

Object Tracking: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models

Accuracy

Average IoU from a tracker across all video frames



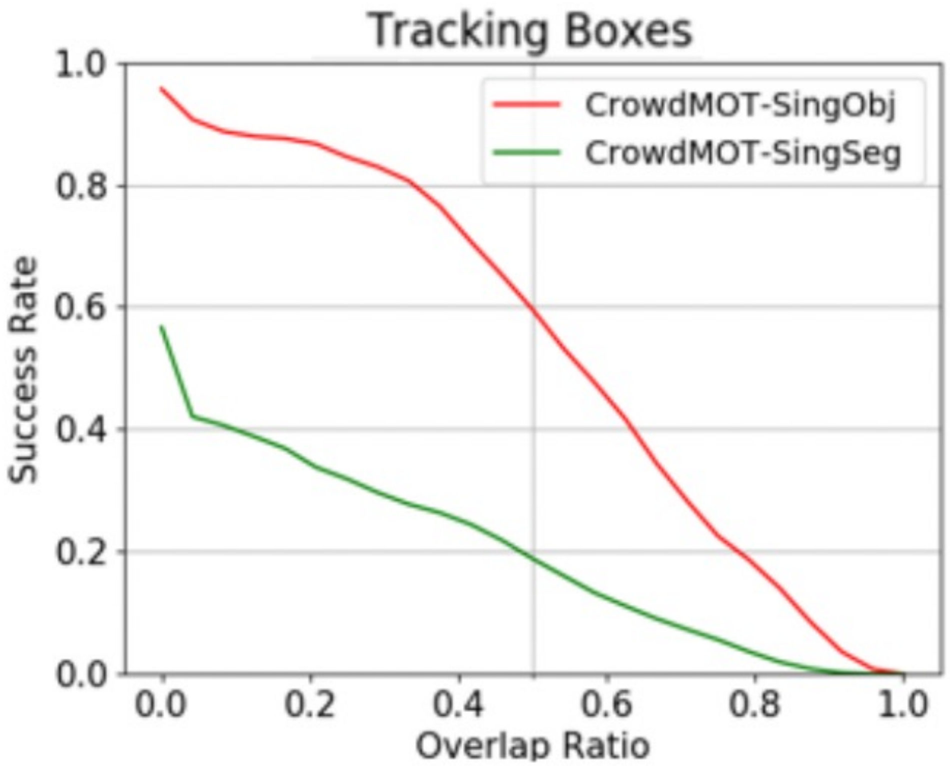
A = Ground Truth
B = Predicted Track
C = Intersection

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

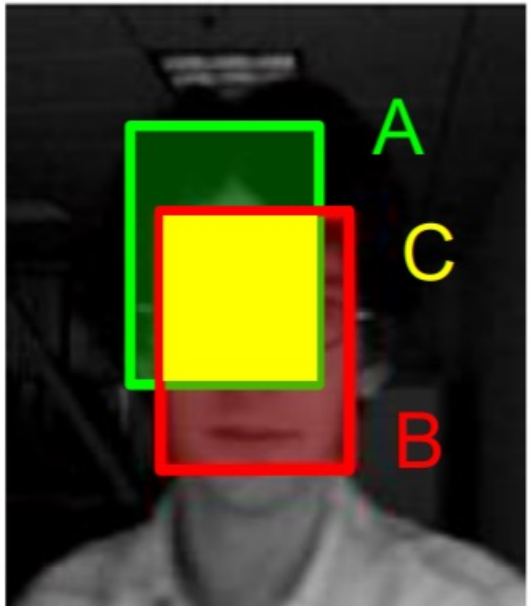
Success Plot

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



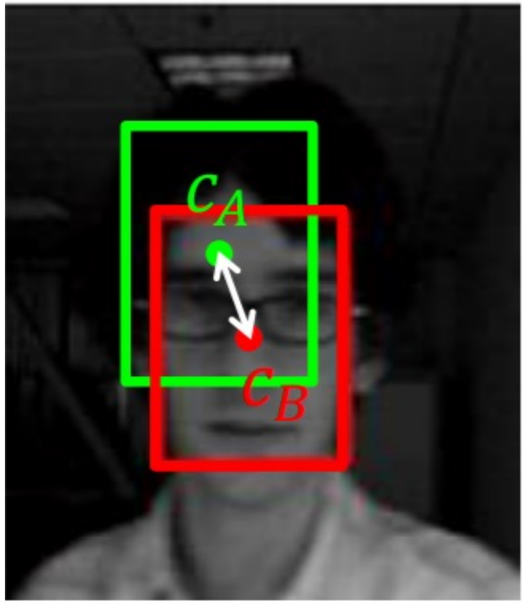
A = Ground Truth
B = Predicted Track
C = Intersection

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



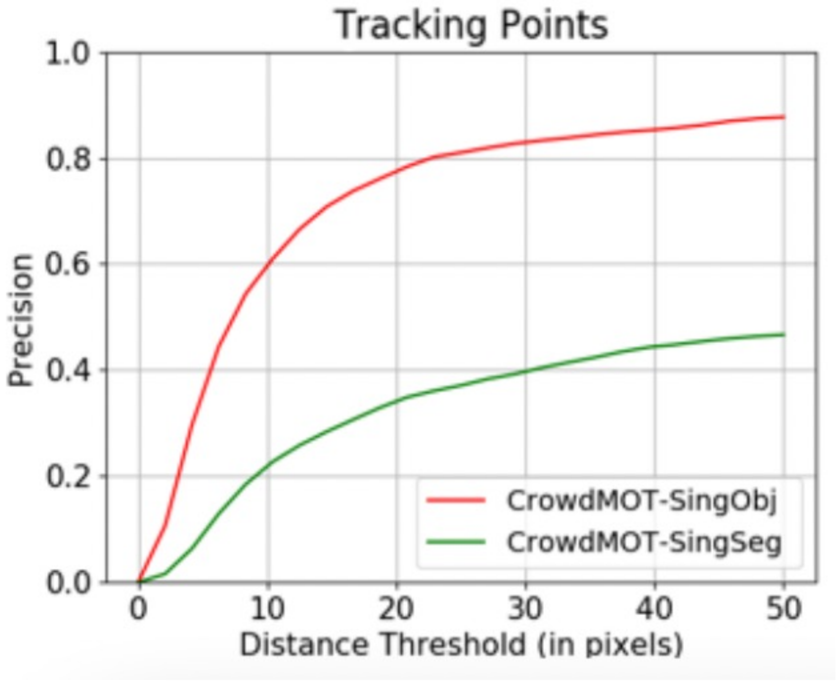
A = Ground Truth
B = Predicted Track

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

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 is 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

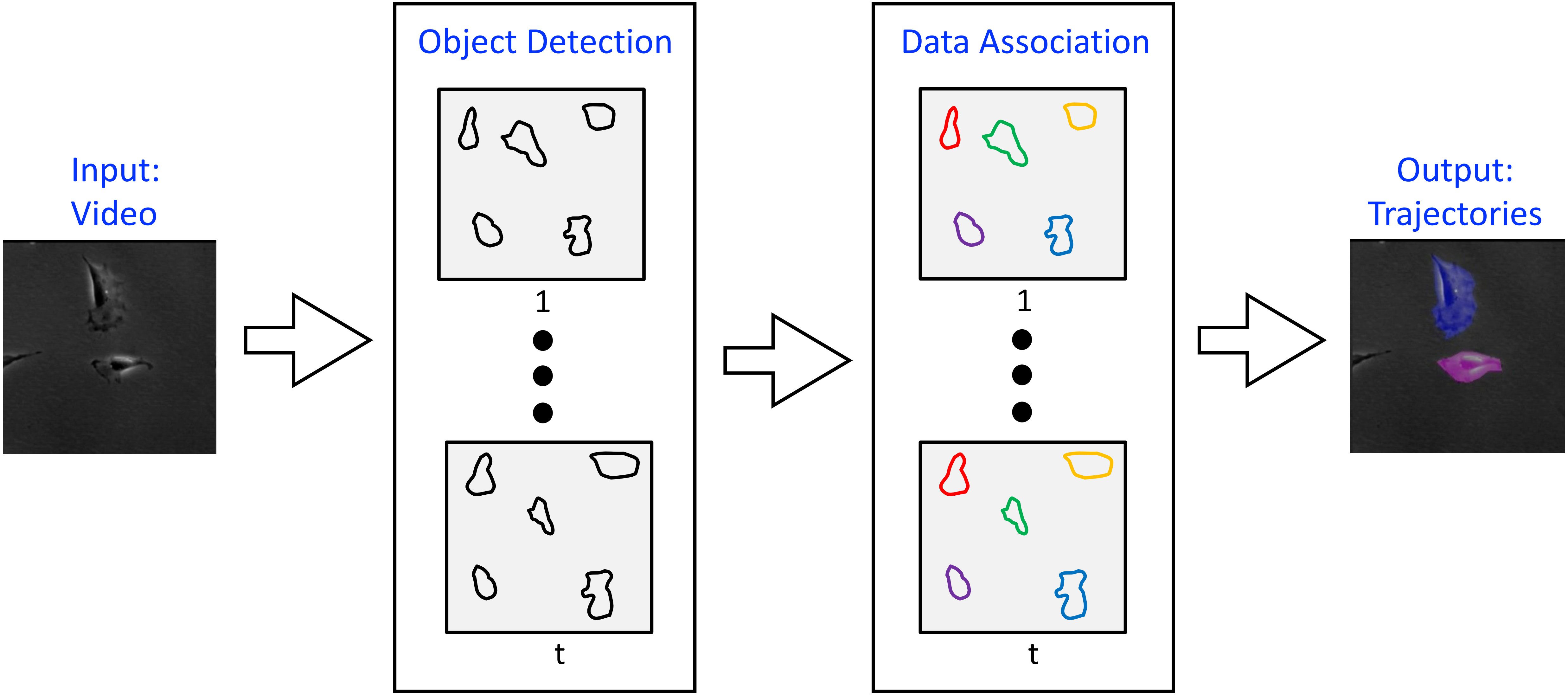
Number of identity switches

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

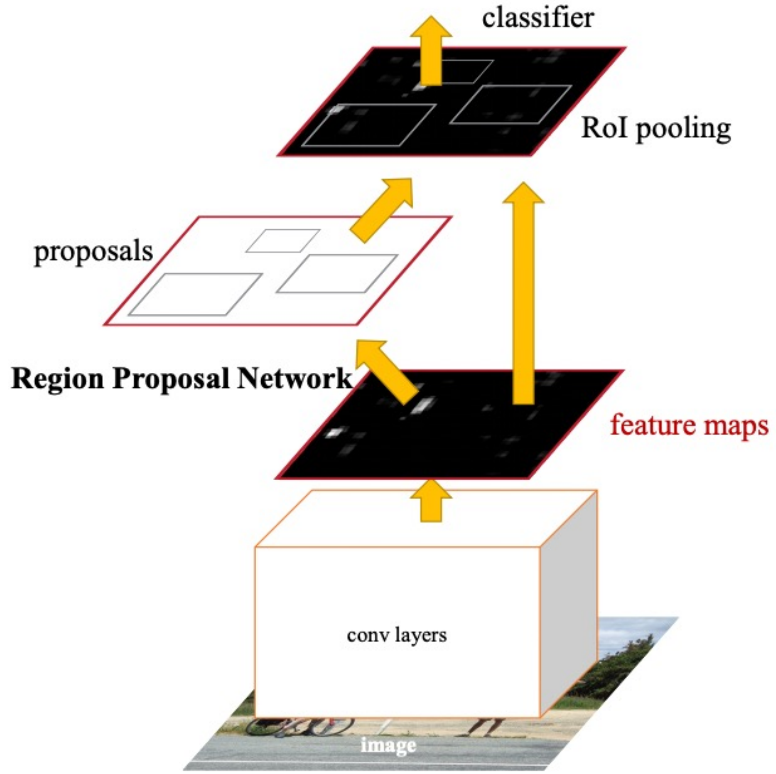
Common Approach: Tracking-by-Detection



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



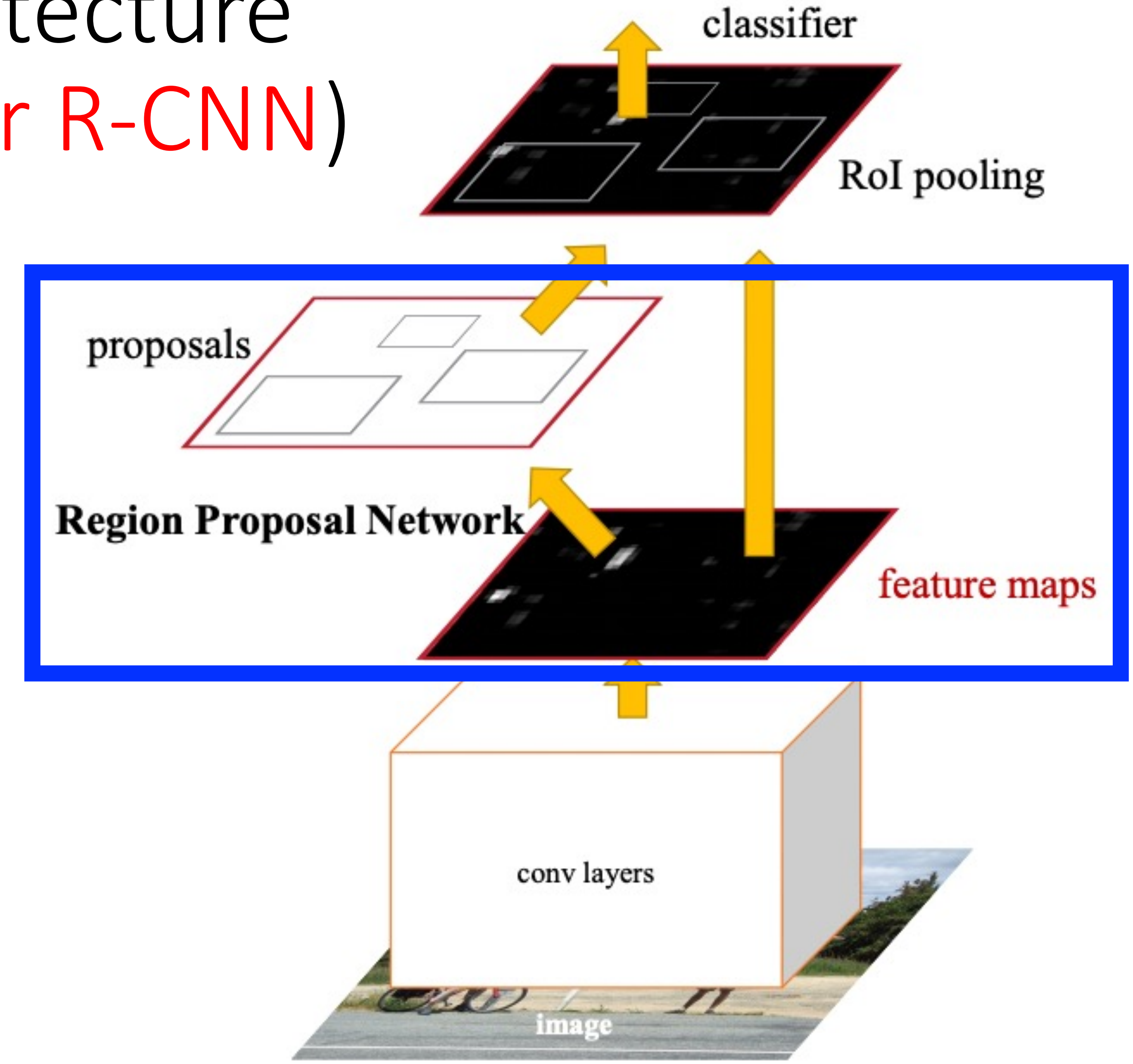
Output:
Trajectories

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)

- Recall:

Embeds a region proposal network in Fast R-CNN using a sliding window approach



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)

- Recall:

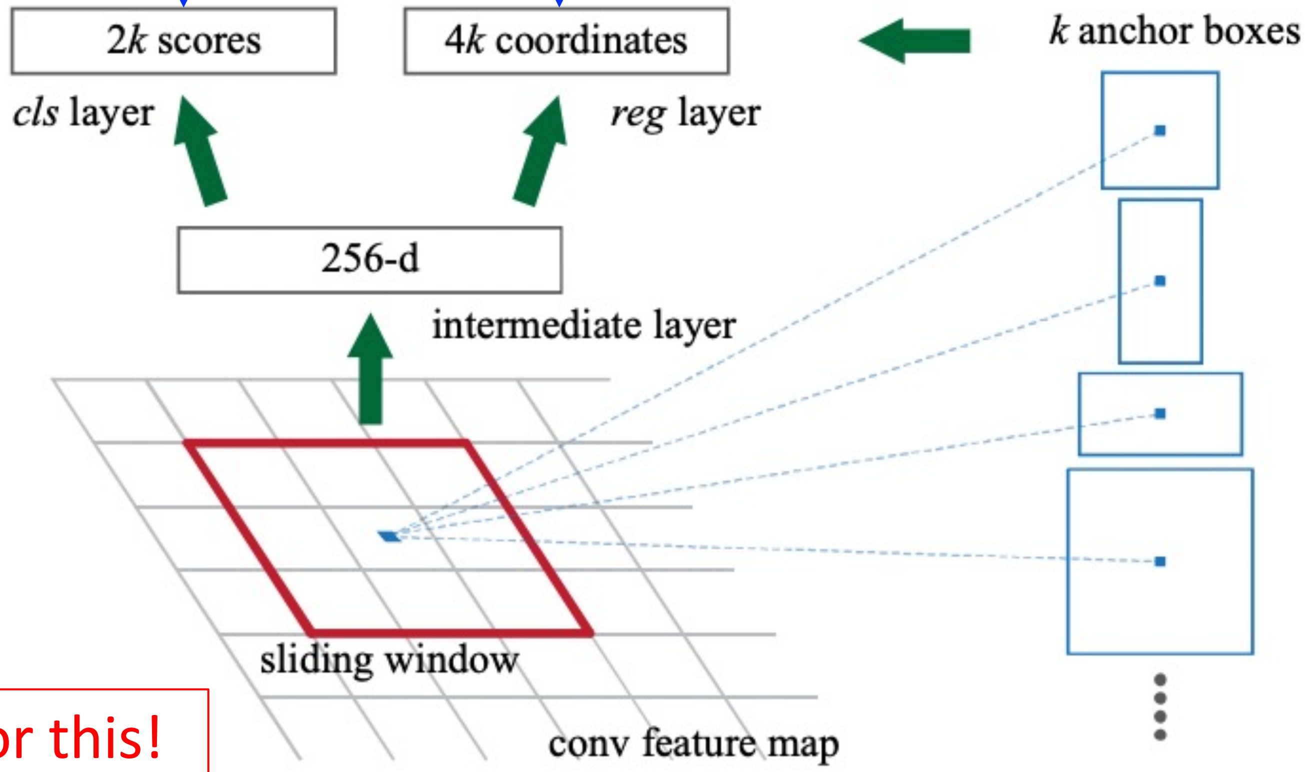
Uses sliding window, since based on convolution

- At each sliding window position, region proposals are predicted with respect to an anchor point (i.e., center of sliding window position)
- At each anchor point, k = 9 anchors are used to represent 3 aspect ratios and 3 scales

Variant addresses and so removes need for this!

Probability of object/not object

Parameters to transform anchor box to GT box (center, width, and height)



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)

- FPN – feature pyramid network
  - Transforms a convolutional network's pyramidal feature hierarchy to have **stronger semantics** at all scales, so that different object sizes are supported

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

Lower layers represent **higher resolutions** with **weaker semantics** (i.e., inadequate representation of high-level concepts since the focus is on low-level features)

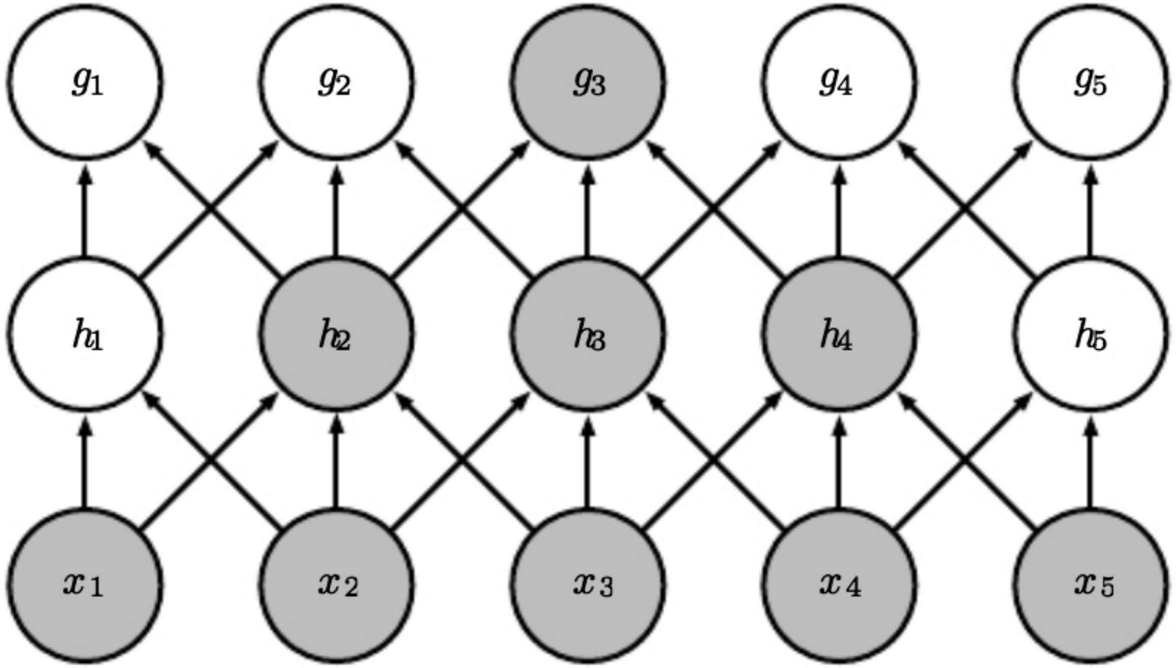


Figure source: https://www.deeplearningbook.org/contents/convnets.html

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

- FPN – feature pyramid network
  - Transforms a convolutional network's pyramidal feature hierarchy to have **stronger semantics** at all scales, so that different object sizes are supported
  - Given a single image scale, its fully convolutional approach generates feature maps with **strong semantics at multiple levels** for use in a downstream task

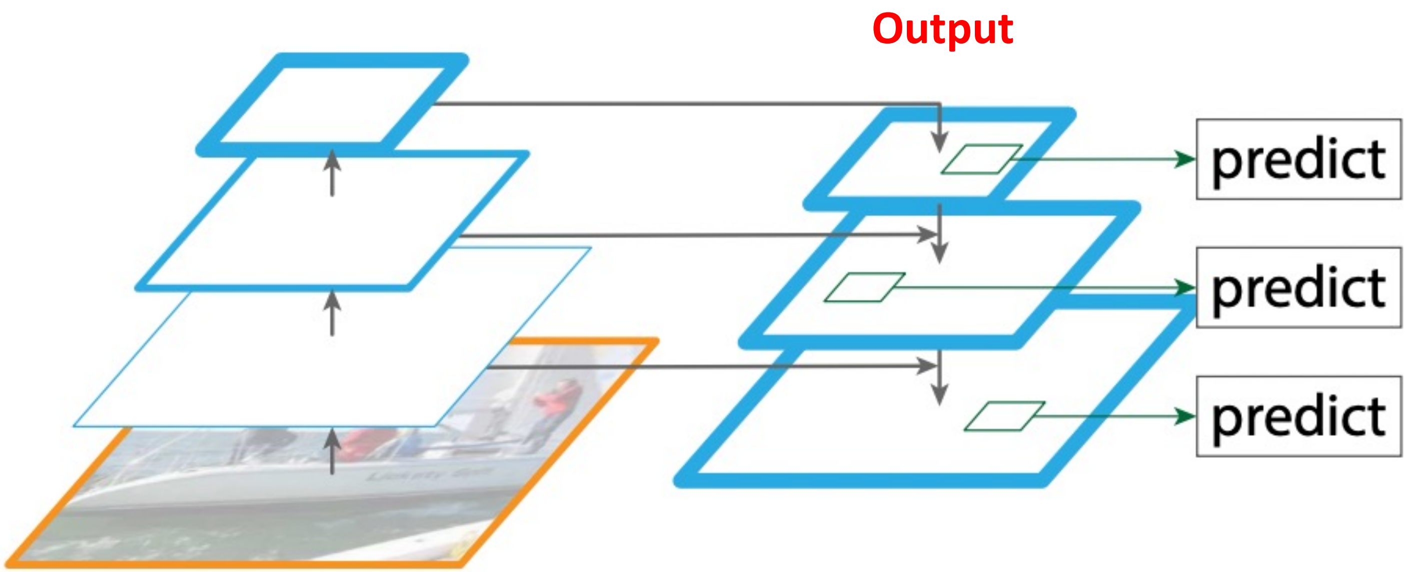


Figure source: Tsung-Yi Lin et al. "Feature Pyramid Networks for Object Detection." CVPR 2017.

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

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

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

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

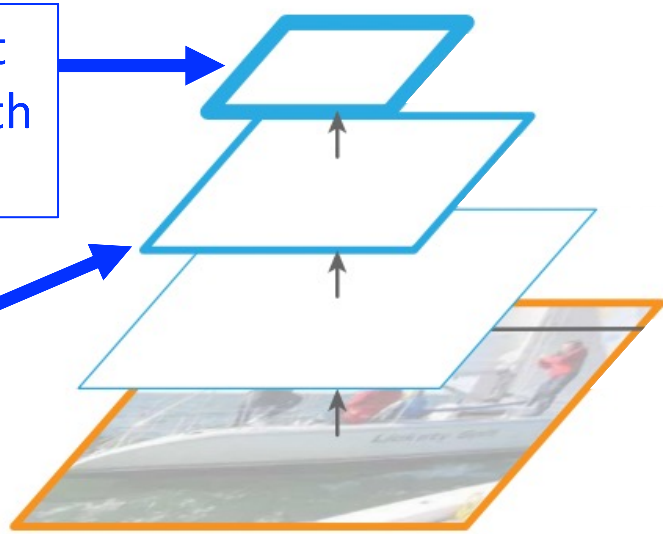


Figure source: Tsung-Yi Lin et al. “Feature Pyramid Networks for Object Detection.” CVPR 2017.

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

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

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

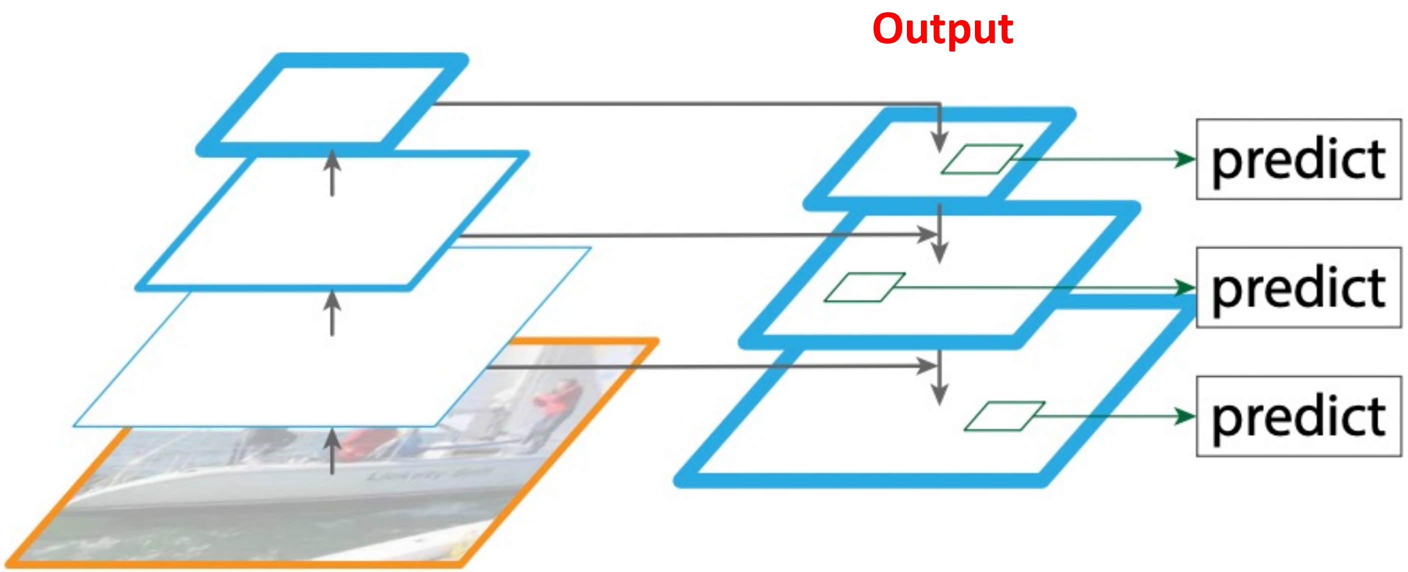
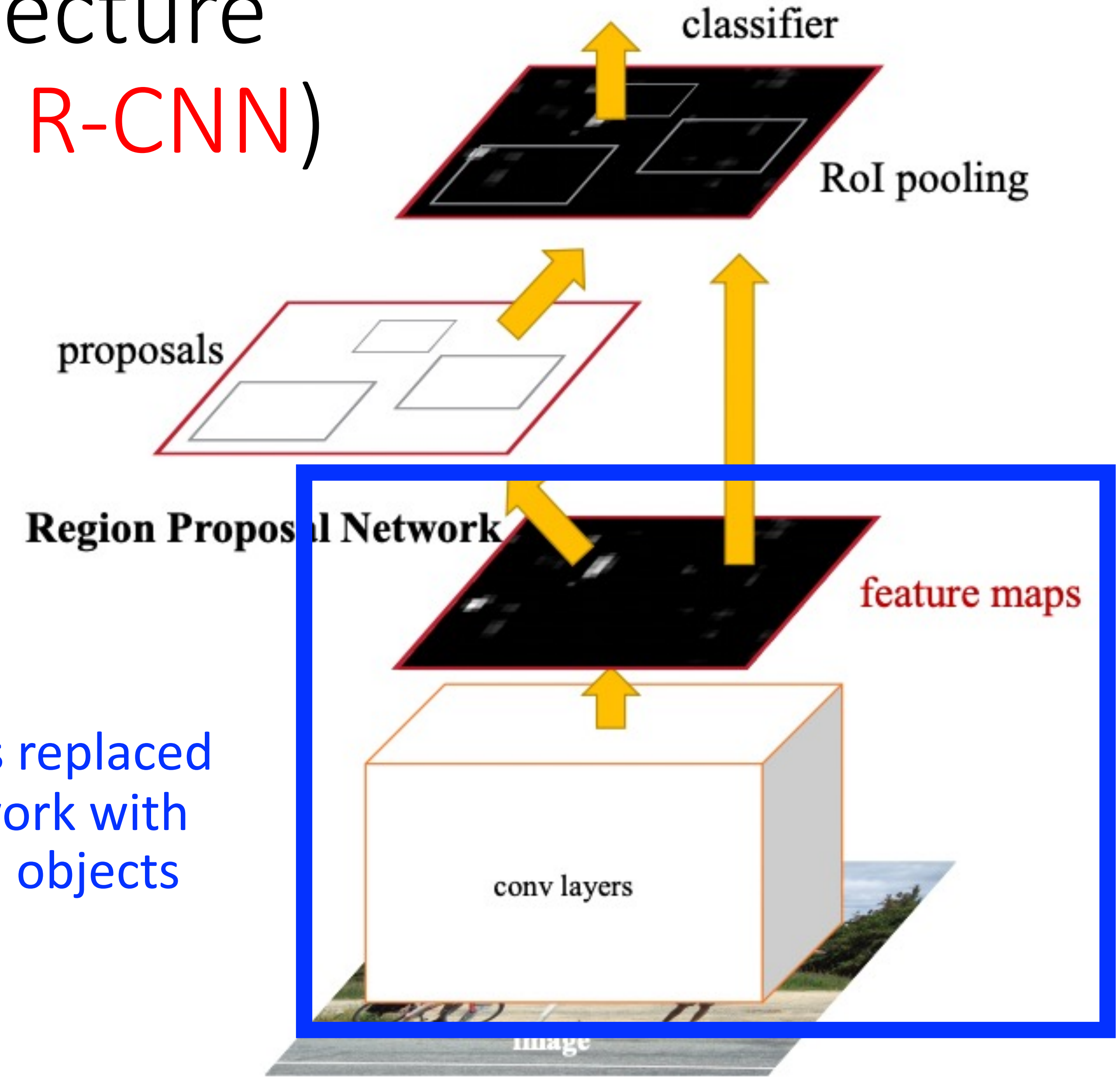


Figure source: Tsung-Yi Lin et al. "Feature Pyramid Networks for Object Detection." CVPR 2017.

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



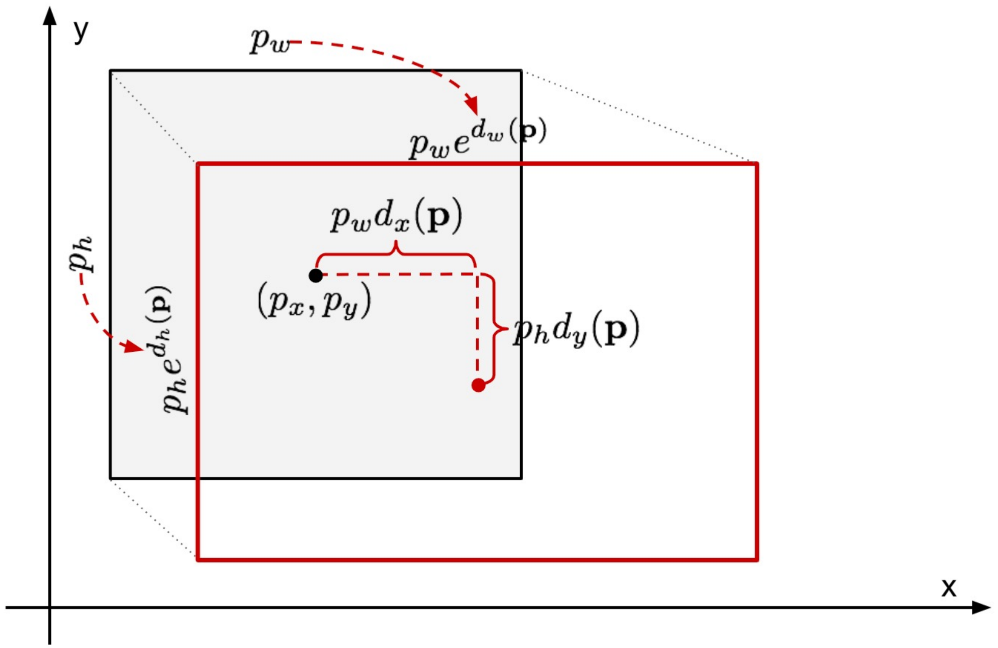
The single-scale feature map is replaced with the feature pyramid network with the aim to detect smaller sized objects

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)

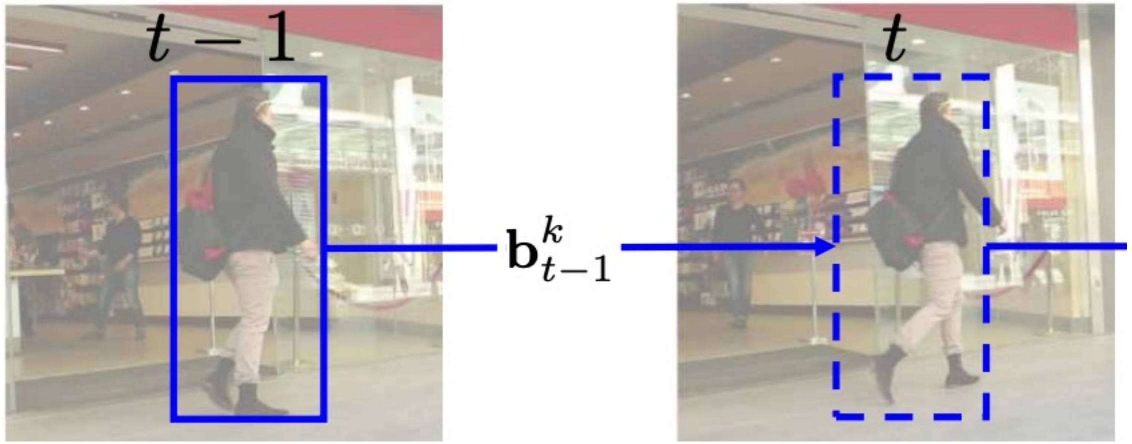
Output: Trajectories

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

1. Motion model to address when an object's position changes a lot between consecutive frames

    - For moving camera, apply image registration

    - For low frame rate, assume constant velocity for all objects

2. Reidentification to enable an object that disappears for a short time to be linked to itself when it re-appears

    - Compare detected object appearance of a "killed" object to newly tracked objects in future video frames using a Siamese network

    - Given two bounding boxes of a "killed" object and candidate detection with high IoU, the Siamese network computes their similarity to determine whether the two tracks should be linked (i.e., if similarity exceeds a pre-defined threshold)

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.

What Makes MOT Difficult?

- 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)

Object Tracking: Today's Topics

- Problem
- Applications
- Datasets
- Evaluation metrics
- Computer vision models


The End