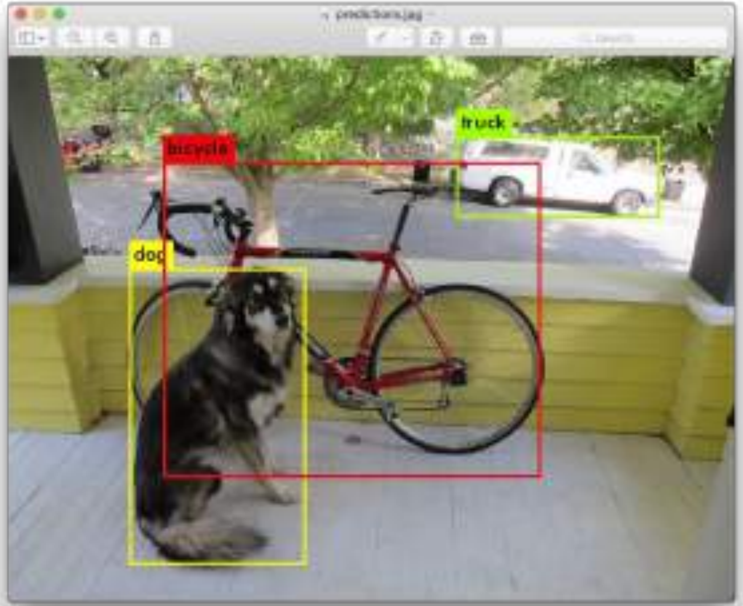
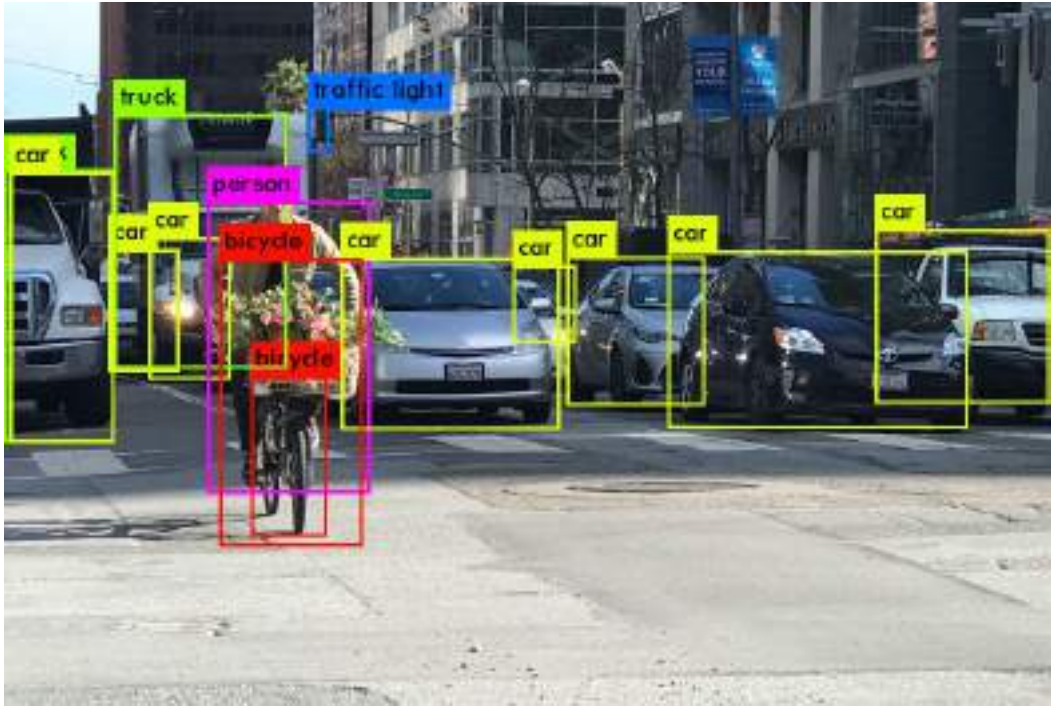
Computer Vision for Embedded Systems

Yung-Hsiang Lu
Purdue University
yunglu@purdue.edu

PURDUE UNIVERSITY

Object Detection

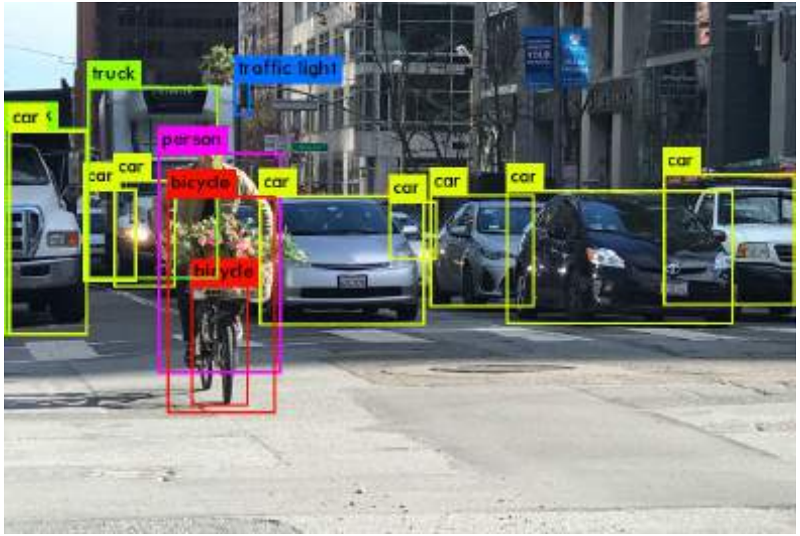




https://pjreddie.com/darknet/yolo/
https://viso.ai/deep-learning/yolov3-overview/

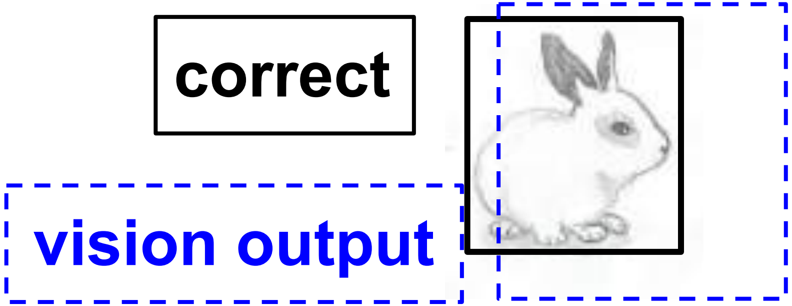
Image Classification vs Object Detection

- Image classification: one dominant object in an image
- Object detection: multiple objects in the same image



Evaluate Object Detection

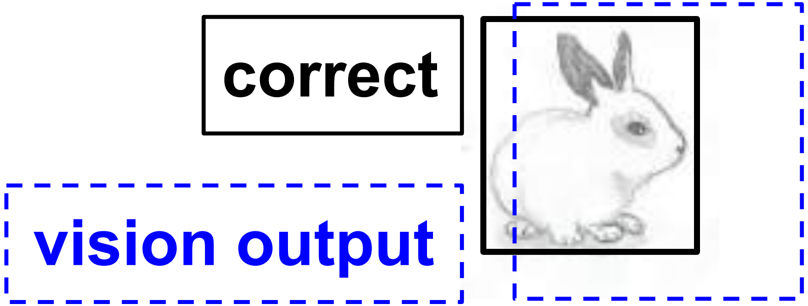
1. correct type of object
2. non maximum suppression
3. correct location (IoU ≥ 0.5)



Intersection over union (IoU)

$$\text{IoU} = \frac{\text{Correct} \cap \text{Vision}}{\text{Correct} \cup \text{Vision}}$$

https://www.deviantart.com/imaginationbutterfly/art/Animal-Drawing-601163034
https://www.template.net/design-templates/drawings/animal-drawings/



$$\text{IoU} = \frac{\text{Correct} \cap \text{Vision}}{\text{Correct} \cup \text{Vision}}$$

Repurpose Image Classifiers

- Apply image classifier at different locations and sizes
- Post-processing: refine bounding boxes, eliminate duplicates, rescore boxes based on other detected objects

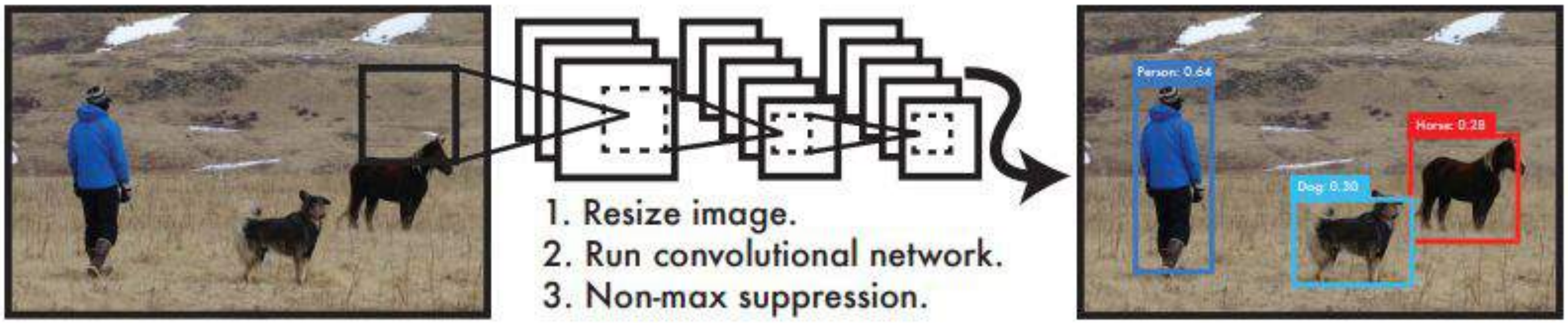
⇒ very slow

https://www.kaggle.com/arunmohan003/yolo-v3-pytorch-tutorial

You Look Only Once (YOLO)

You Only Look Once: Unified, Real-Time Object Detection 2016 (25,000+ citations)

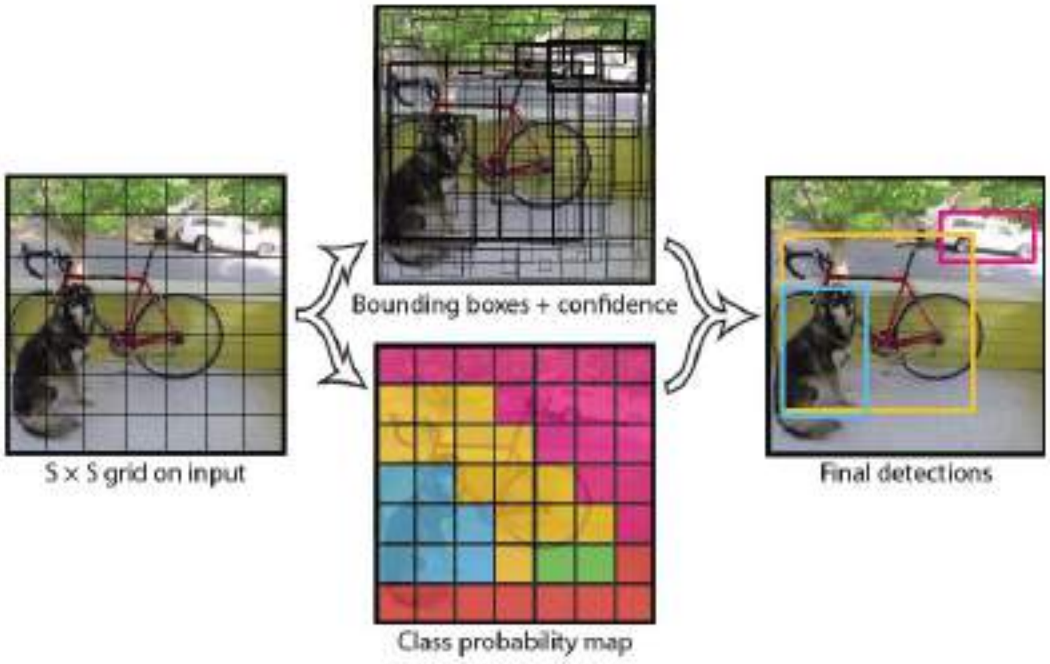
- 45 frames per second (FPS), faster version 155 FPS
- double mAP from earlier fast detectors
- Use 448 x 448 pixels to detect smaller objects
- See the entire images during training ⇒ implicitly include context information
- Testing using natural and artificial images

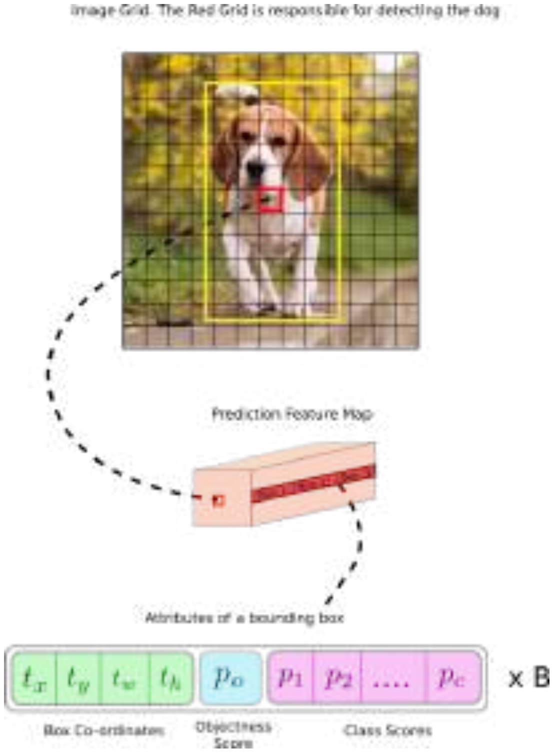


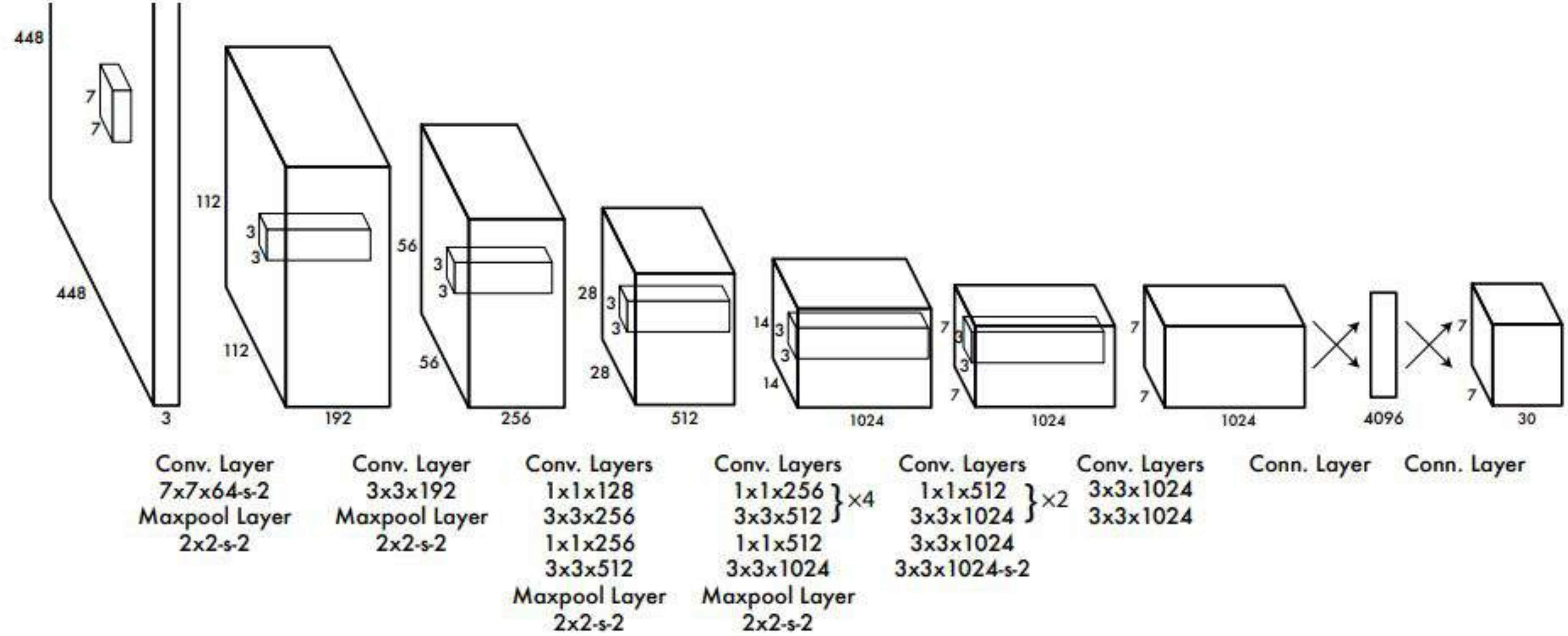
References

1. https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf
2. https://towardsdatascience.com/yolo2-walkthrough-with-examples-e40452ca265f
3. https://www.kaggle.com/arunmohan003/yolo-v3-pytorch-tutorial
4. https://pjreddie.com/darknet/yolo/

Grid and bounding boxes






448
448
7
7
3
112
112
3
3
192
56
56
3
3
256
28
28
3
3
512
14
14
3
3
1024
7
7
3
3
1024
7
7
1024
4096
7
7
30
Conv. Layer
7x7x64-s-2
Maxpool Layer
2x2-s-2
Conv. Layer
3x3x192
Maxpool Layer
2x2-s-2
Conv. Layers
1x1x128
3x3x256
1x1x256
3x3x512
Maxpool Layer
2x2-s-2
Conv. Layers
1x1x256
3x3x512
×4
1x1x512
3x3x1024
Maxpool Layer
2x2-s-2
Conv. Layers
1x1x512
3x3x1024
×2
3x3x1024
3x3x1024-s-2
Conv. Layers
3x3x1024
3x3x1024
Conn. Layer
Conn. Layer

Training

- pretrain first 20 layers for a week
- 88% top-5 accuracy of ImageNet 2012 classification
- convert classification to detection
- add four convolutional and two fully connected layers
- final layer both class probabilities and bounding box
- Leaky ReLU activation
- Learning rate 10^{-2} to 10^{-3} to 10^{-4}
- Dropout 0.5

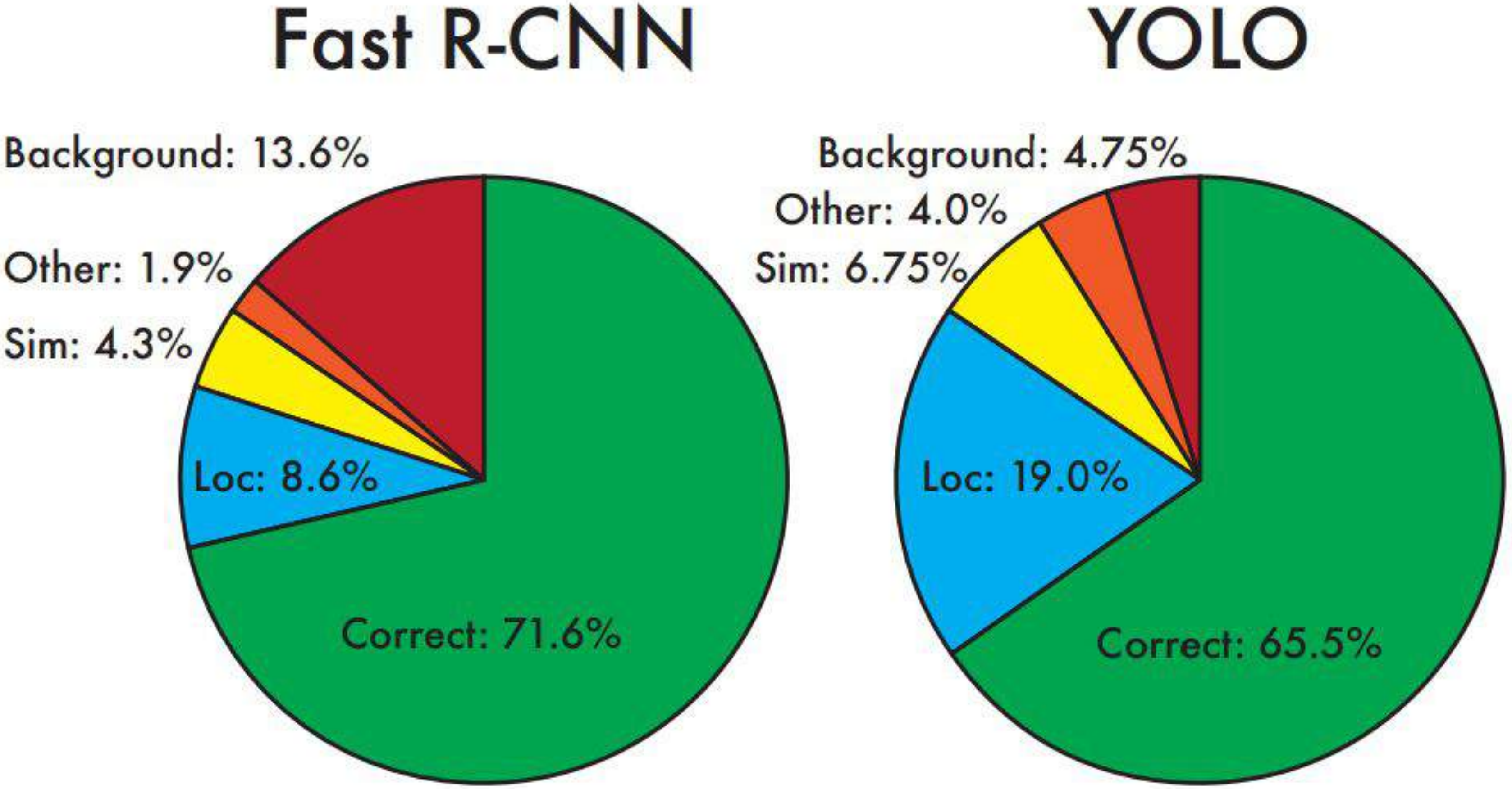
Limitations

- assumption: each grid cell has only one class of object
- unable to detect small objects in groups
- expect aspect ratios
- downsampling
- treat errors in small bounding boxes the same as large boxes

Comparison

- Deformable parts models: disjoint pipeline to extract features, classify regions, predict bounding boxes
- R-CNN: regional proposals, SVM scores bounding boxes, non maximum suppression, 40 seconds / image
- YOLO makes assumption about objects to improve speed, check only 98 bounding boxes / image

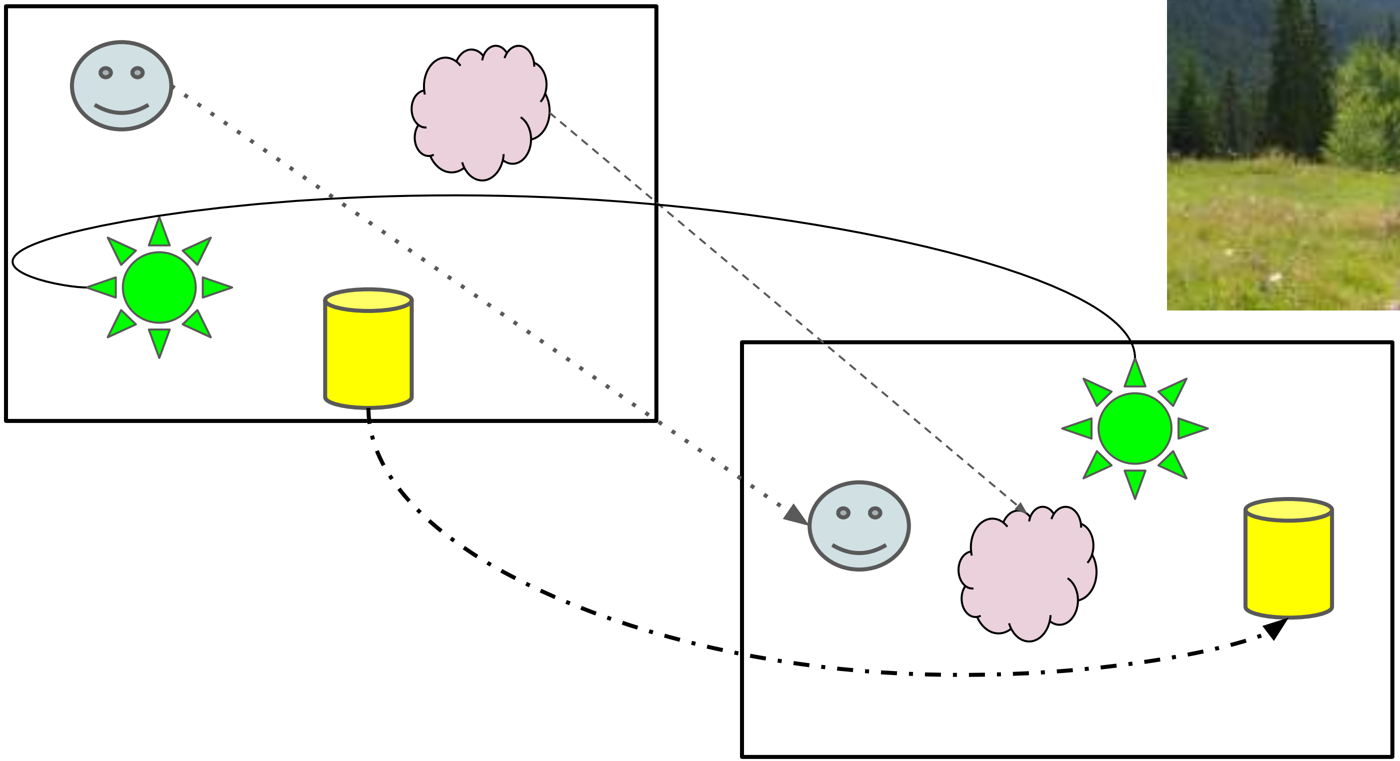
Real-Time Detectors	Train	mAP	FPS
100Hz DPM [30]	2007	16.0	100
30Hz DPM [30]	2007	26.1	30
Fast YOLO	2007+2012	52.7	**155**
YOLO	2007+2012	**63.4**	45
Less Than Real-Time			
Fastest DPM [37]	2007	30.4	15
R-CNN Minus R [20]	2007	53.5	6
Fast R-CNN [14]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[27]	2007+2012	73.2	7
Faster R-CNN ZF [27]	2007+2012	62.1	18
YOLO VGG-16	2007+2012	66.4	21


Fast R-CNN
YOLO
Background: 13.6%
Other: 1.9%
Sim: 4.3%
Loc: 8.6%
Correct: 71.6%
Background: 4.75%
Other: 4.0%
Sim: 6.75%
Loc: 19.0%
Correct: 65.5%

VOC 2012 test	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table
MR_CNN_MORE_DATA [11]	**73.9**	**85.5**	**82.9**	**76.6**	**57.8**	**62.7**	**79.4**	77.2	86.6	**55.0**	**79.1**	**62.2**
HyperNet_VGG	71.4	84.2	78.5	73.6	55.6	53.7	78.7	**79.8**	87.7	49.6	74.9	52.1
HyperNet_SP	71.3	84.1	78.3	73.3	55.5	53.6	78.6	79.6	87.5	49.5	74.9	52.1
Fast R-CNN + YOLO	70.7	83.4	78.5	73.5	55.8	43.4	79.1	73.1	**89.4**	49.4	75.5	57.0
MR_CNN_S_CNN [11]	70.7	85.0	79.6	71.5	55.3	57.7	76.0	73.9	84.6	50.5	74.3	61.7
Faster R-CNN [27]	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3
DEEP_ENS_COCO	70.1	84.0	79.4	71.6	51.9	51.1	74.1	72.1	88.6	48.3	73.4	57.8
NoC [28]	68.8	82.8	79.0	71.6	52.3	53.7	74.1	69.0	84.9	46.9	74.3	53.1
Fast R-CNN [14]	68.4	82.3	78.4	70.8	52.3	38.7	77.8	71.6	89.3	44.2	73.0	55.0
UMICH_FGS_STRUCT	66.4	82.9	76.1	64.1	44.6	49.4	70.3	71.2	84.6	42.7	68.6	55.8
NUS_NIN_C2000 [7]	63.8	80.2	73.8	61.9	43.7	43.0	70.3	67.6	80.7	41.9	69.7	51.7
BabyLearning [7]	63.2	78.0	74.2	61.3	45.7	42.7	68.2	66.8	80.2	40.6	70.0	49.8
NUS_NIN	62.4	77.9	73.1	62.6	39.5	43.3	69.1	66.4	78.9	39.1	68.1	50.0
R-CNN VGG BB [13]	62.4	79.6	72.7	61.9	41.2	41.9	65.9	66.4	84.6	38.5	67.2	46.7
R-CNN VGG [13]	59.2	76.8	70.9	56.6	37.5	36.9	62.9	63.6	81.1	35.7	64.3	43.9
YOLO	57.9	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5
Feature Edit [32]	56.3	74.6	69.1	54.4	39.1	33.1	65.2	62.7	69.7	30.8	56.0	44.6
R-CNN BB [13]	53.3	71.8	65.8	52.0	34.1	32.6	59.6	60.0	69.8	27.6	52.0	41.7
SDS [16]	50.7	69.7	58.4	48.5	28.3	28.8	61.3	57.5	70.8	24.1	50.7	35.9
R-CNN [13]	49.6	68.1	63.8	46.1	29.4	27.9	56.6	57.0	65.9	26.5	48.7	39.5

Tracking Objects (in Video)

Tracking Problem



https://www.dreamstime.com/photos-images/white-horse-run-green-grass.html

Types of tracking problem

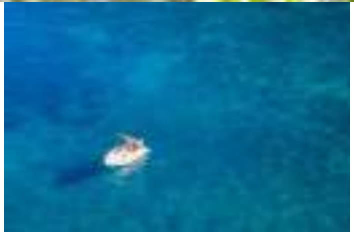
- moving camera?
- single or multiple cameras?
- single or multiple objects?
- major objects or all objects?
- similar or distinct objects?
- occlusion?
- crossing?
- online or offline?
- initial object marking?

https://www.wlfi.com/content/news/Purdue-women-accept-WNIT-bid-will-face-IUPUI-476706723.html

Moving Camera

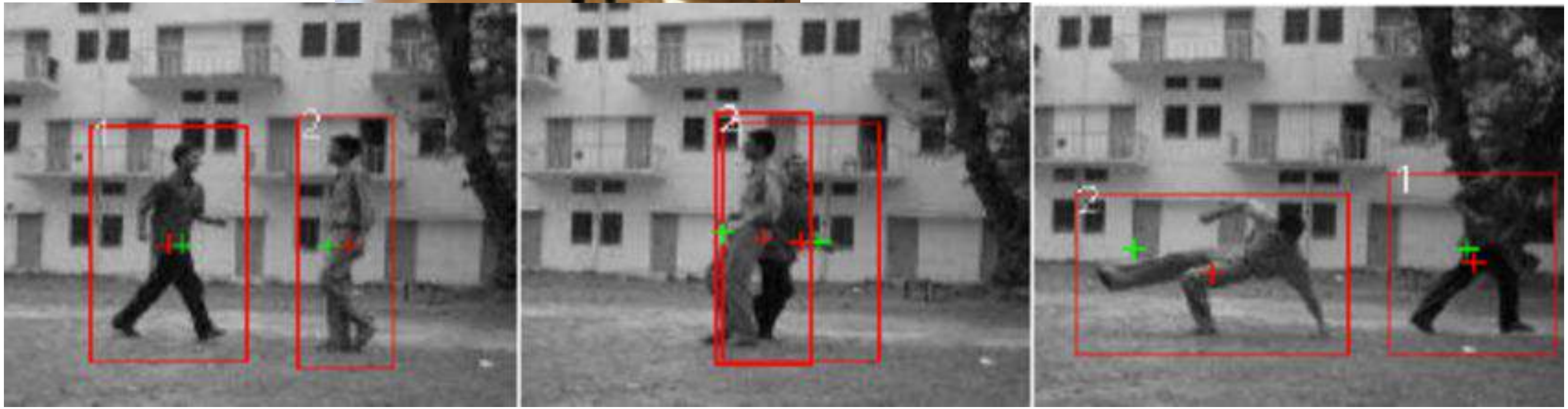
https://www.pexels.com/photo/person-holding-silver-iphone-6-93765/
https://auto.howstuffworks.com/car-driving-safety/safety-regulatory-devices/dashcams.htm
https://www.phase1vision.com/blog/what-is-a-ptz-camera-and-what-is-it-used-for
https://www.adorama.com/alc/what-are-the-best-drones-with-4k-cameras/

Single or Multiple Objects?



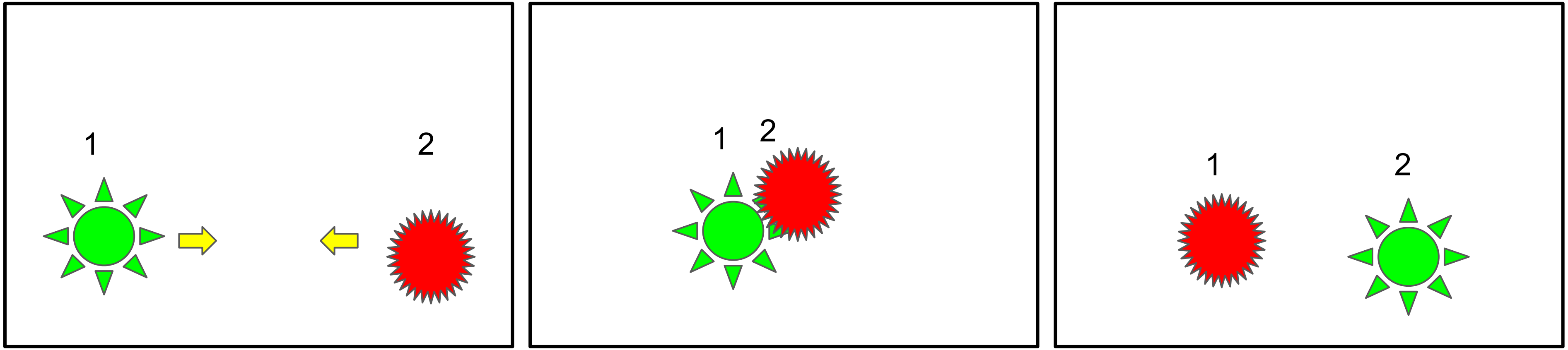
https://www.pexels.com/photo/bird-on-tree-branch-1461867/
https://www.dkfindout.com/us/animals-and-nature/fish/school-fish/
https://bustingbrackets.com/2020/05/27/purdue-basketball-review-2020-21-depth-chart-season-outlook/
https://www.pexels.com/photo/boat-in-the-middle-of-the-ocean-638453/

Occlusion and Crossing



https://www.researchgate.net/figure/Object-Tracking-during-and-after-Occlusion_fig5_220166473
https://kimwilbanks.com/2019/01/12/is-your-name-on-the-column/

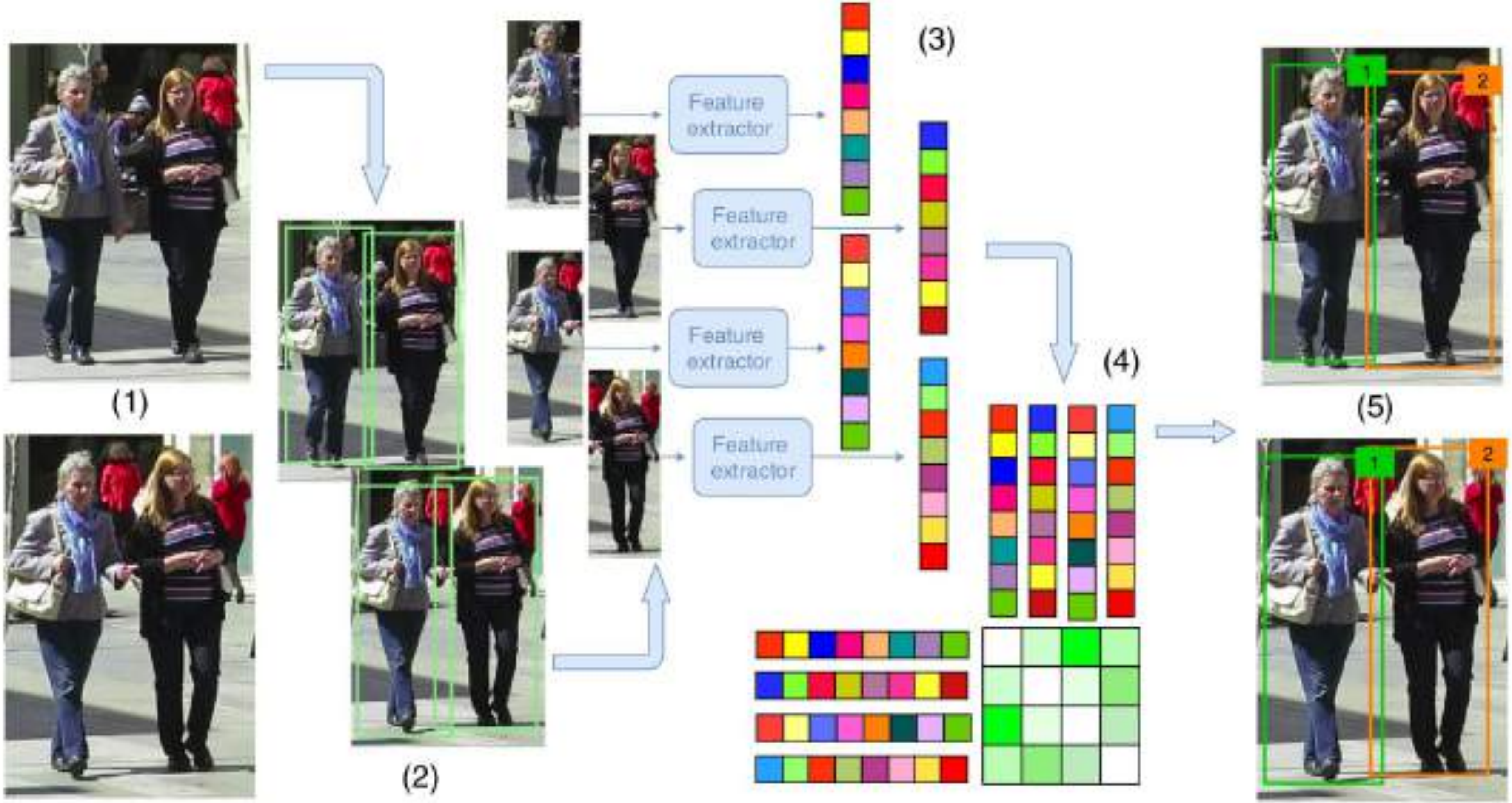
Problem of switched IDs



Deep learning in video multi-object tracking: A survey

Gioele Ciaparrone, Francisco Luque Sánchez, Siham Tabik , Luigi Troiano, Roberto Tagliaferri, Francisco Herrer

Neurocomputing 381 (2020) 61–88


Feature extractor
Feature extractor
Feature extractor
Feature extractor
(1)
(2)
(3)
(4)
(5)
1
2
1
2

Metrics

- object detection: intersection over union (common)
- # frames an object of interest is correctly tracked
- # ID switches
- fragmentation: interruptions in tracking

$$\text{score} = 1 - \frac{FP + FP + IDSW}{GT}$$

Datasets

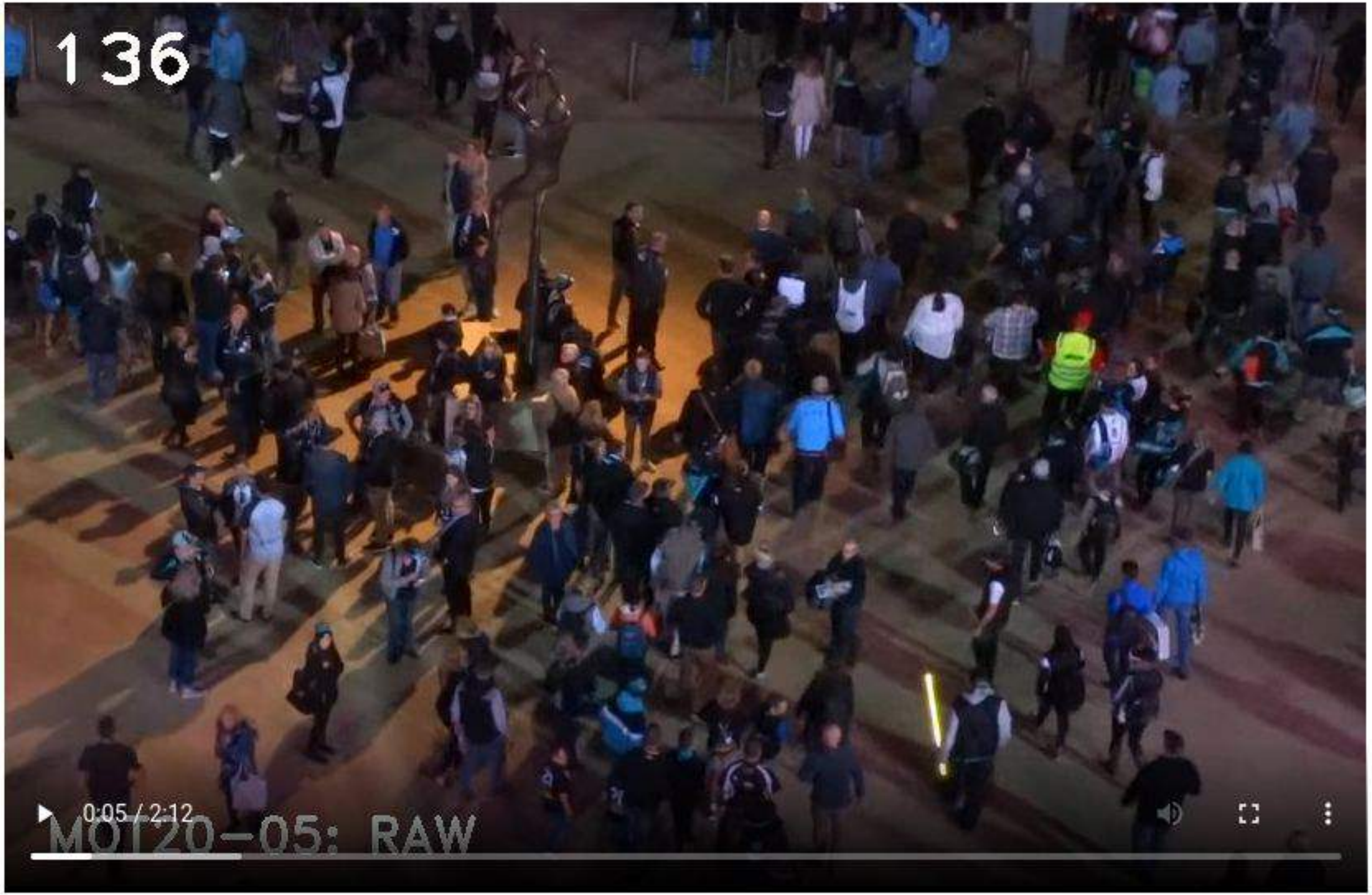

Multiple Object Tracking Benchmark
home
data
results
vis
QVA
submit
FAQ
people
login
sign
MOT15
MOT16
MOT17Det
MOT17
MOT20
MOT20Det
Welco
allenge: The Multiple Object Tracking Benchmark!

MOT 15

Sample	Name	FPS	Resolution	Length	Tracks	Boxes	Density	Description
	Venice-2	30	1920x1080	600 (00:20)	26	7141	11.9	People walking around a large square.
	KITTI-17	10	1224x370	145 (00:15)	9	683	4.7	Walking pedestrians on a sunny day, static camera
	KITTI-13	10	1242x375	340 (00:34)	42	762	2.2	Busy urban environment filmed from a moving car
	ADL-Rundle-8	30	1920x1080	654 (00:22)	28	6783	10.4	A pedestrian scene filmed at night by a moving camera
	ADL-Rundle-6	30	1920x1080	525 (00:18)	24	5009	9.5	A pedestrian street scene filmed from a low angle.

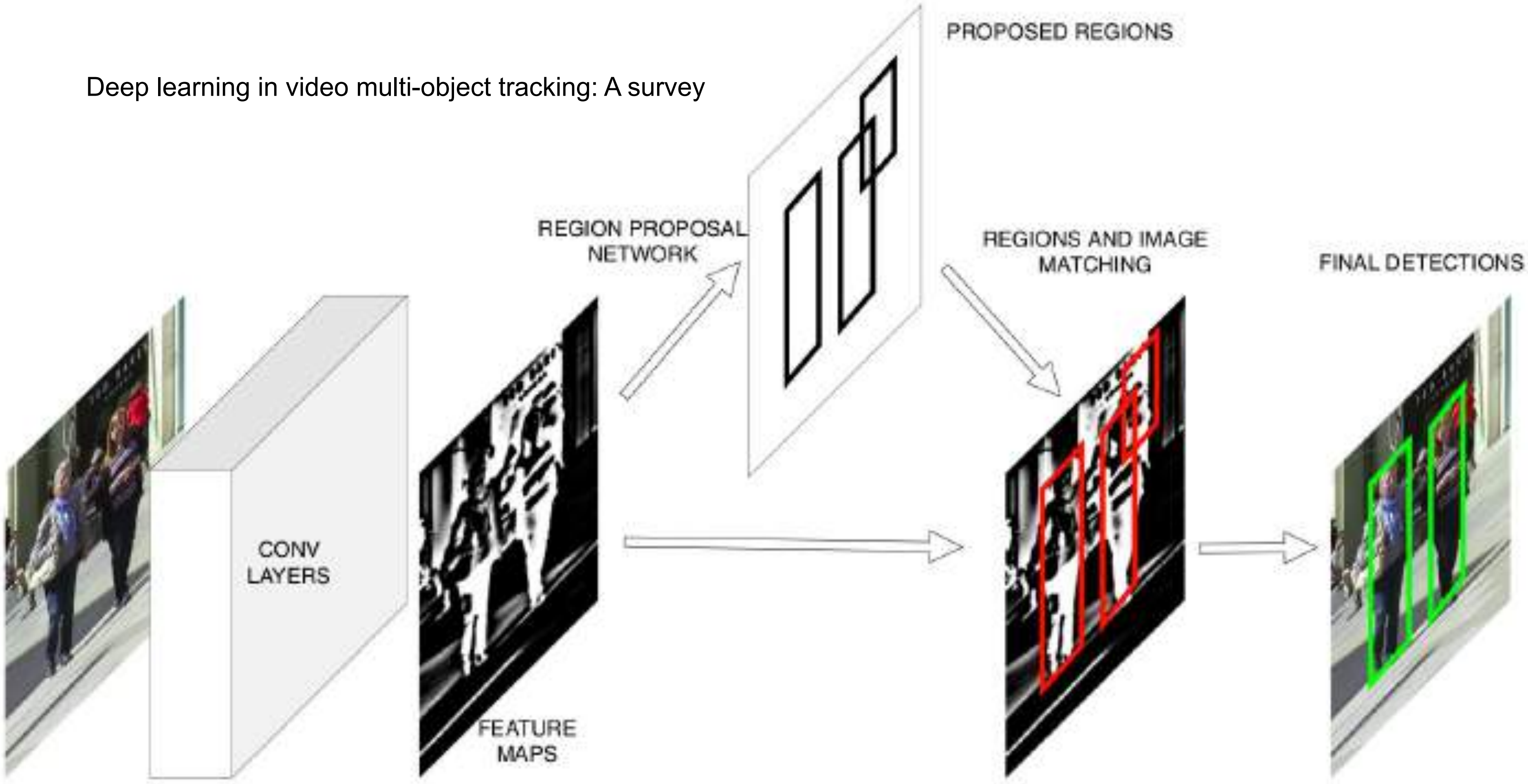
MOT20

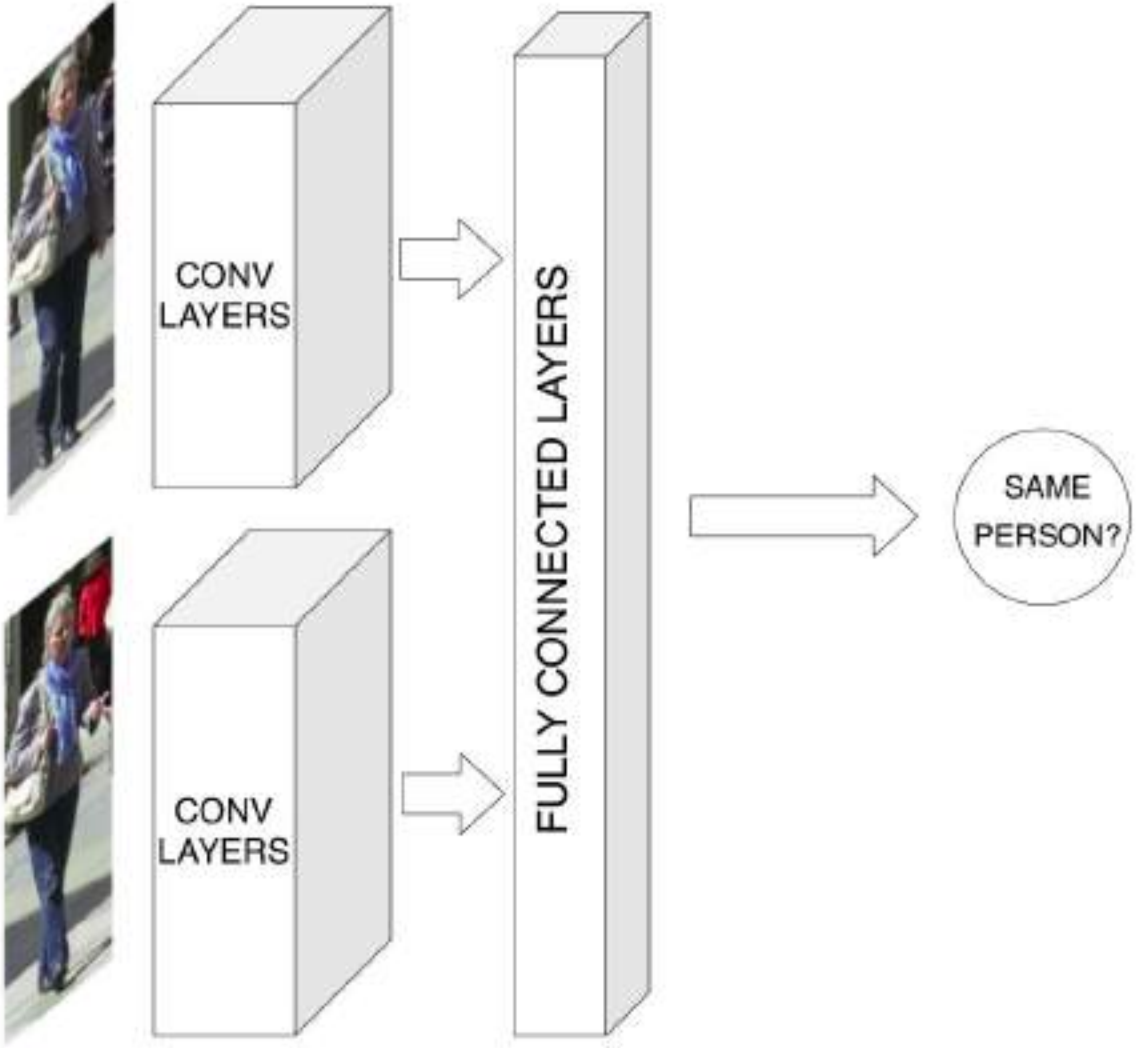
more people each frame


136
0:05 / 2:12
MOT20-05: RAW

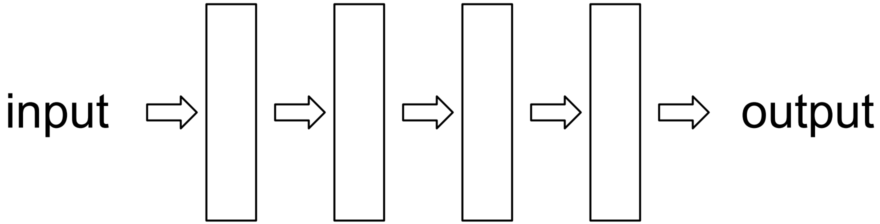
2021 Low-Power Computer Vision Challenge

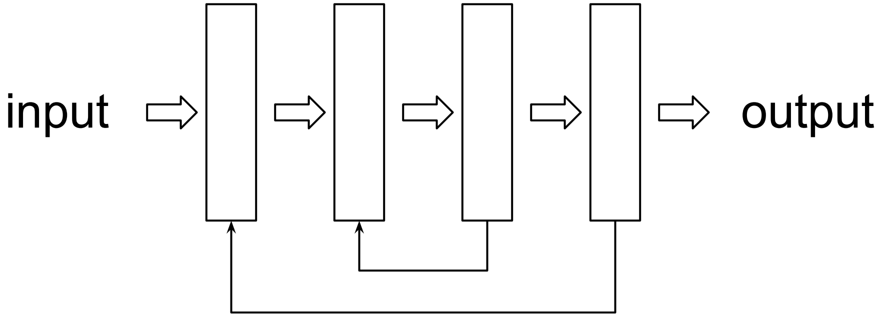
Deep learning in video multi-object tracking: A survey



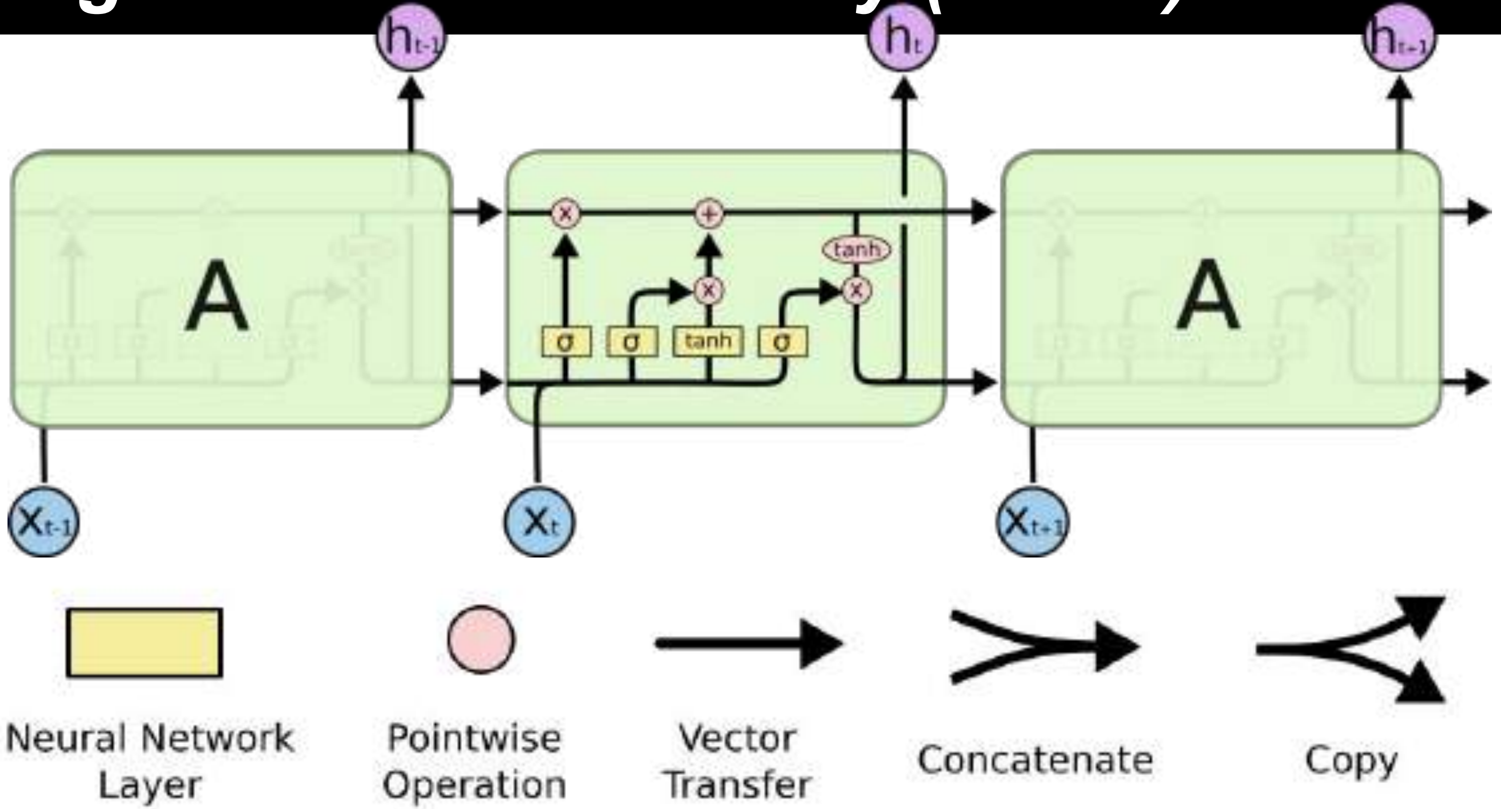

CONV
LAYERS
CONV
LAYERS
FULLY CONNECTED LAYERS
SAME
PERSON?

Feed-Forward vs. Recurrent Networks

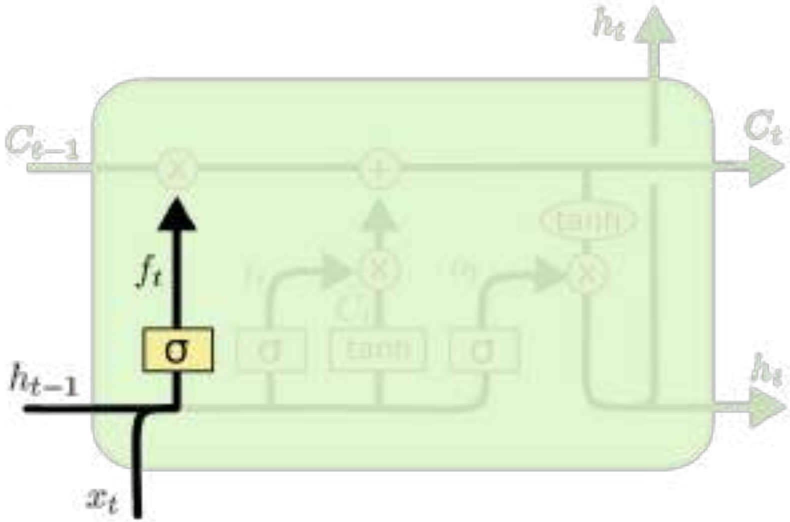




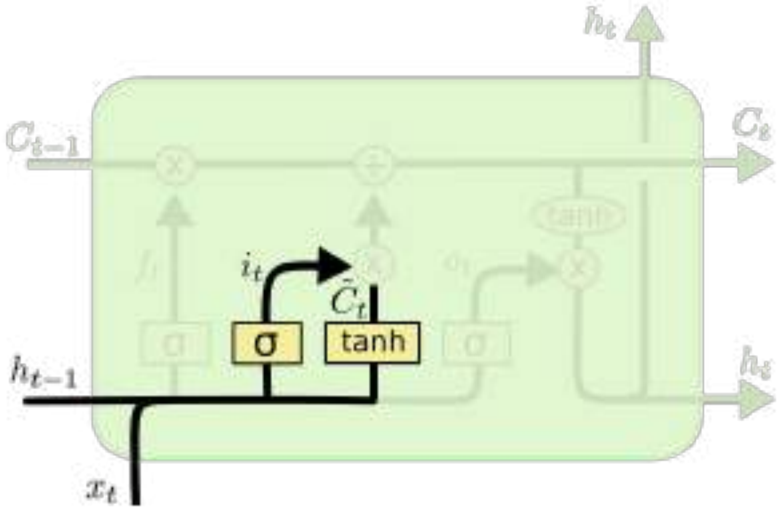
Long Short Term Memory (LSTM)



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

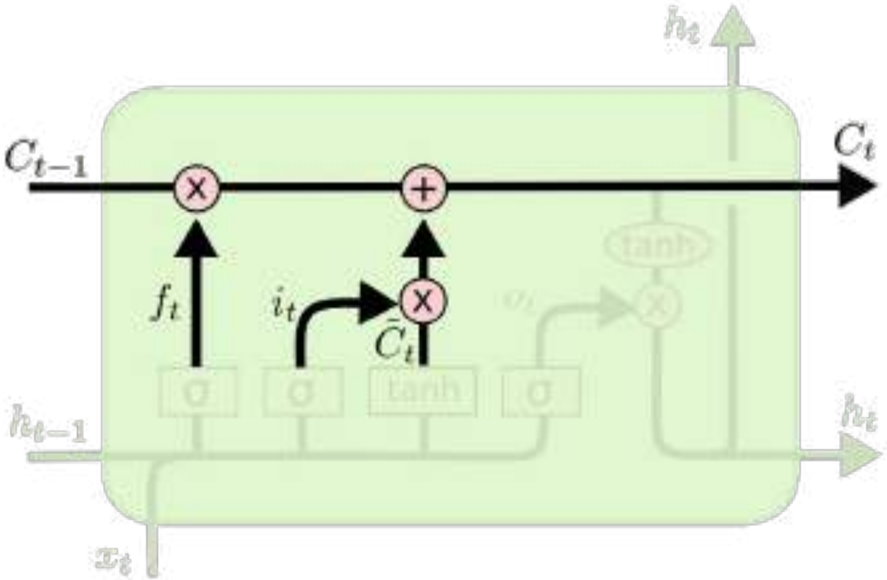


$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

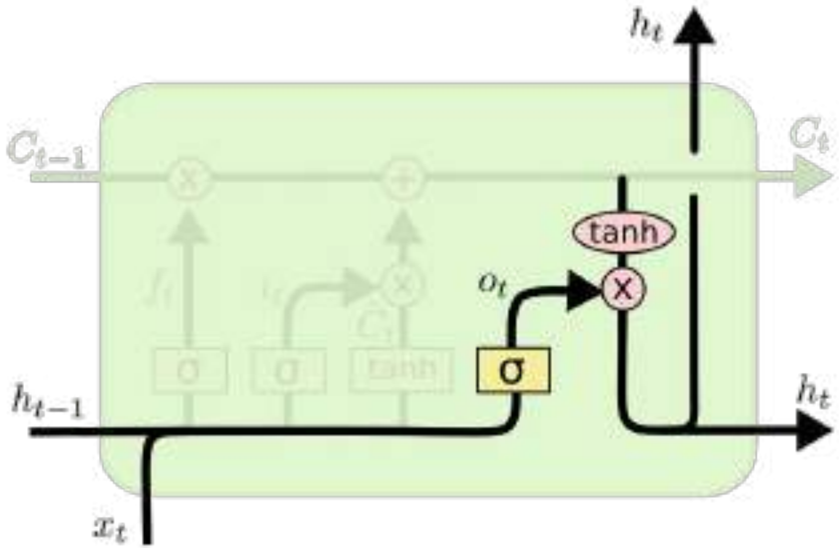


$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

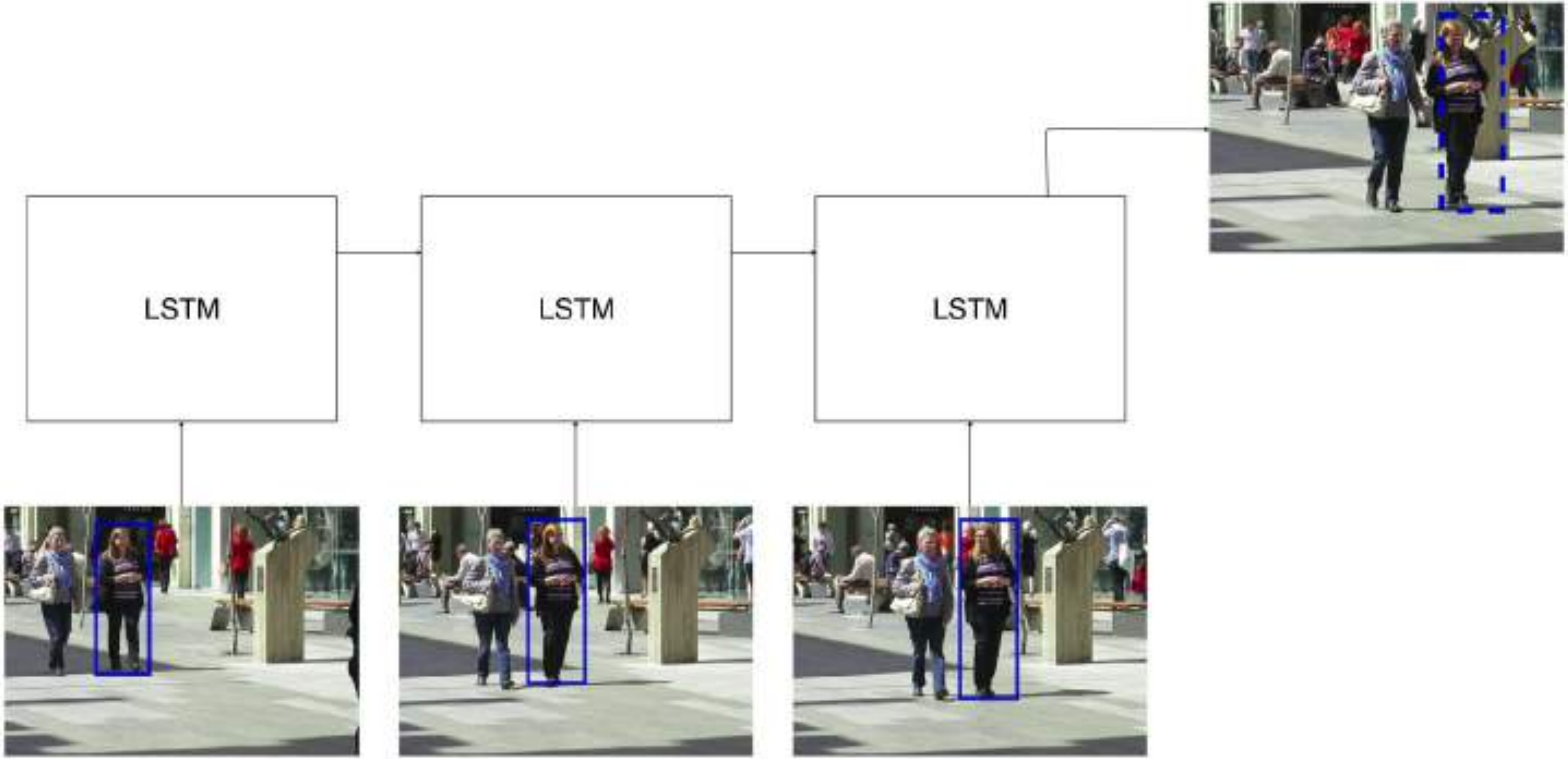


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$


LSTM
LSTM
LSTM

Occlusion and Tracking

Improving Tracking

- Improve detection and neural networks for feature extraction
- Mitigate errors
- Track different types of objects
- Evaluate robustness

Preview: Transformers

“Camera Placement Meeting Restrictions Of Computer Vision”, IEEE International Conference on Image Processing 2020

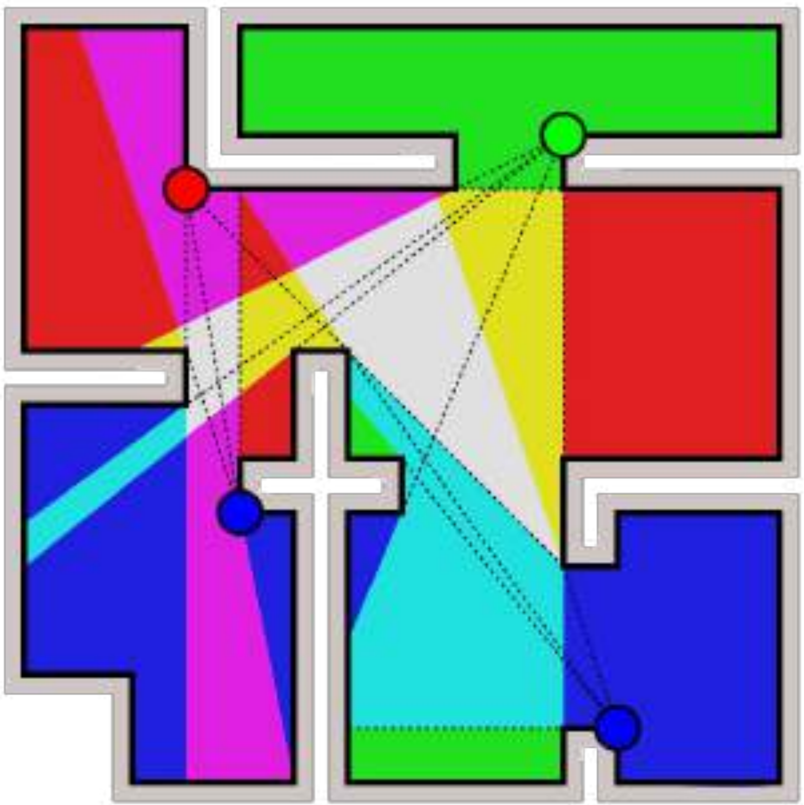
Sara Aghajanzadeh
2020 MSECE Purdue
(2022) doctoral student at U Illinois

Art Gallery Problem

Where to locate guards so that every place in the gallery can be observed by at least one guard.

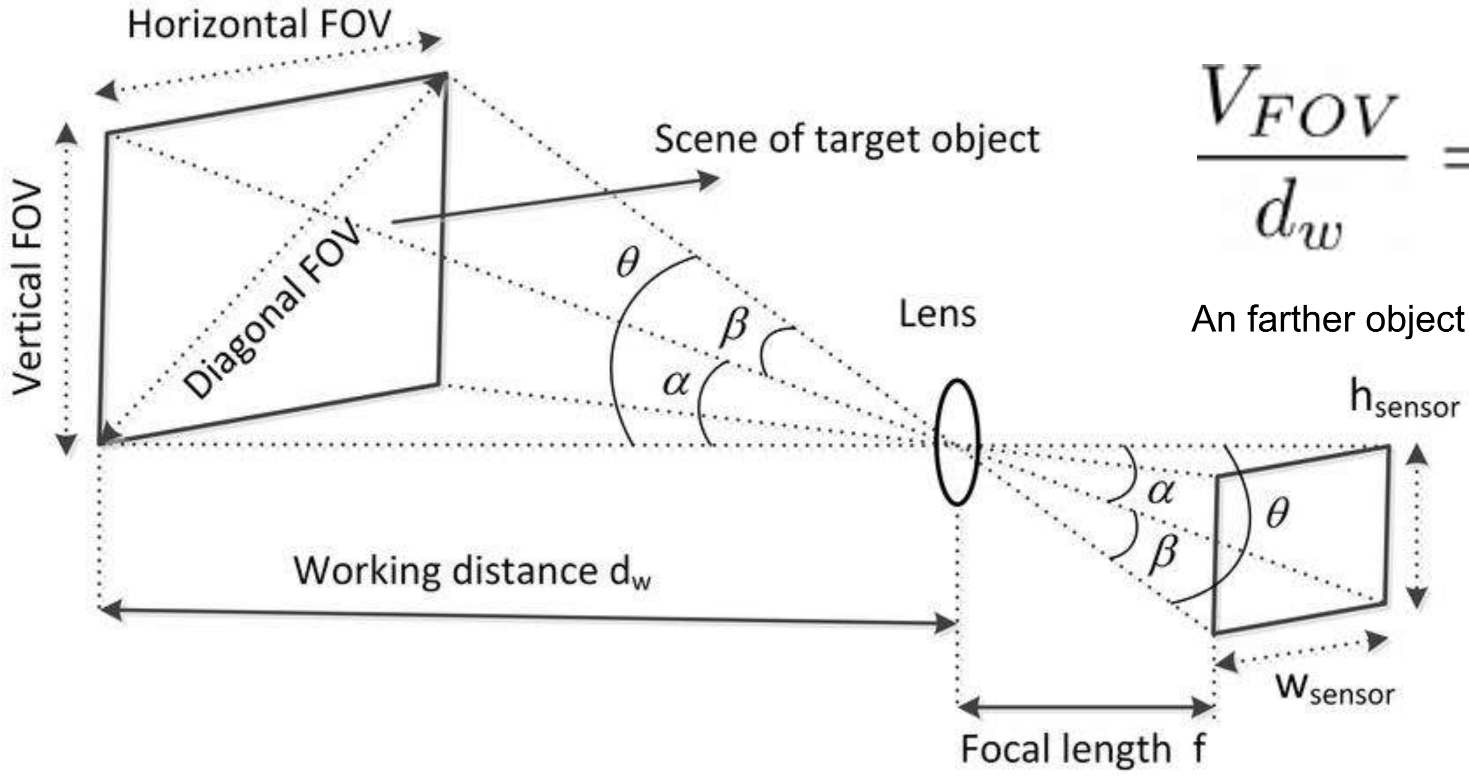
The guards cannot see through walls.

Assumption: each guard can see infinitely far.



https://en.wikipedia.org/wiki/Art_gallery_problem

Camera's Field of View



$$\frac{V_{FOV}}{d_w} = \frac{h_{sensor}}{f}$$

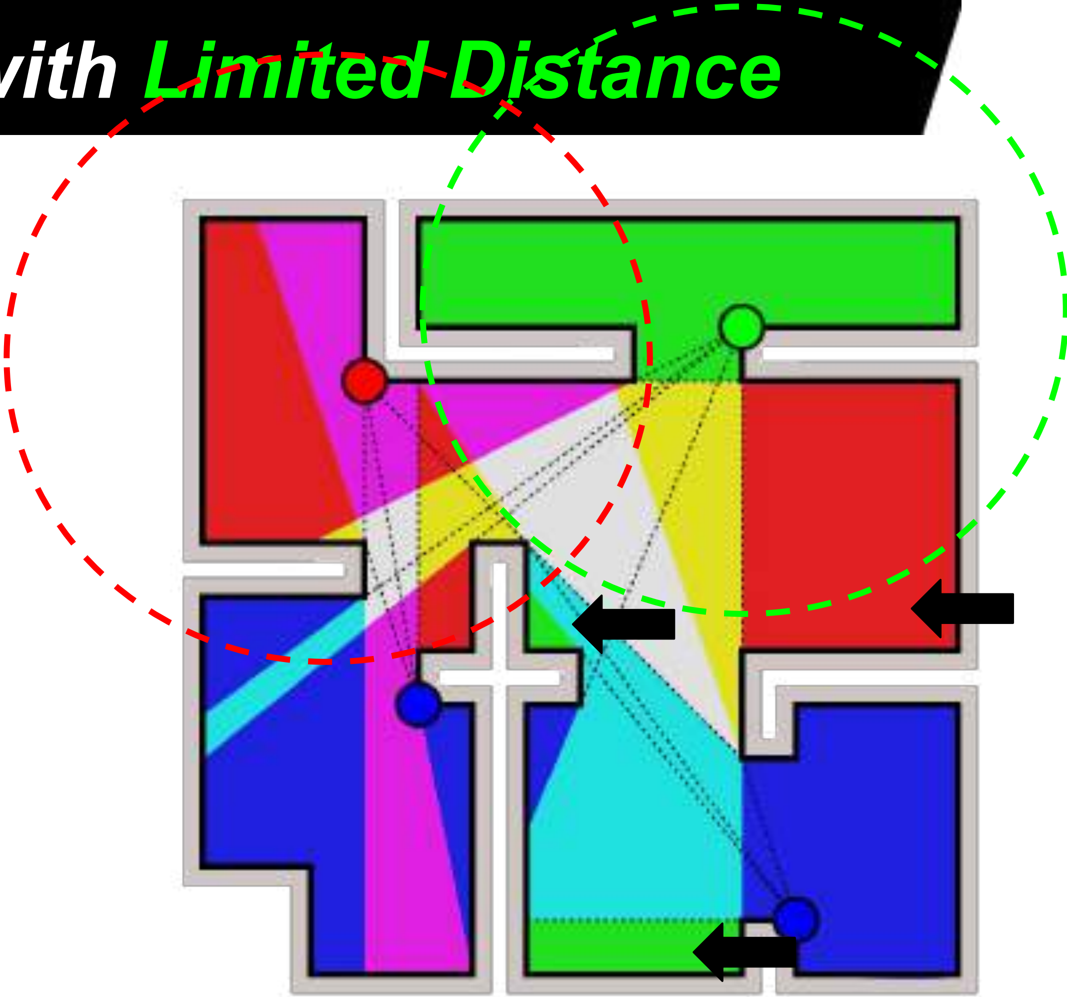
An farther object appears smaller

https://www.researchgate.net/figure/Illustration-of-camera-lenss-field-of-view-FOV_fig4_335011596

Art Gallery Problem with Limited Distance

If a guard has limited viewing distance, the problem is more complex.

The regions marked by black arrows are no longer visible by any guard.



Partition Polygons for Cameras