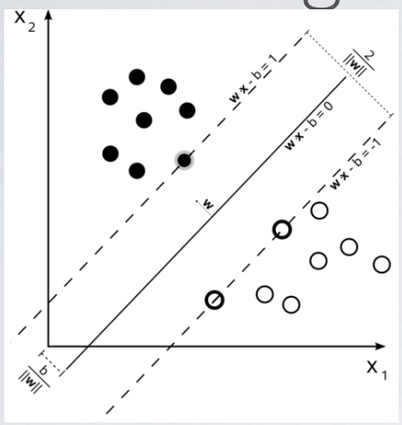
LEARNING TO SEE OBJECTS THROUGH SPACE AND TIME

Philipp Krähenbühl

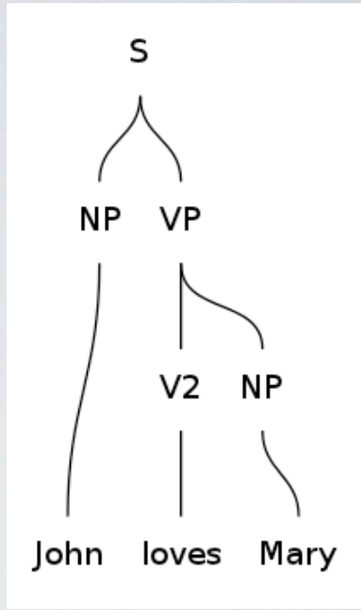
UT AUSTIN AI LAB

Machine learning

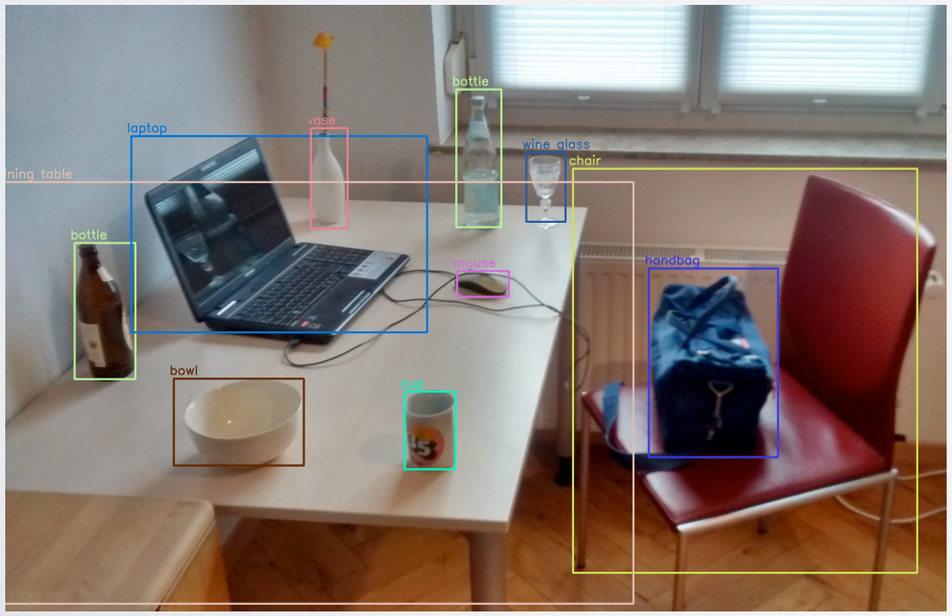


Robotics

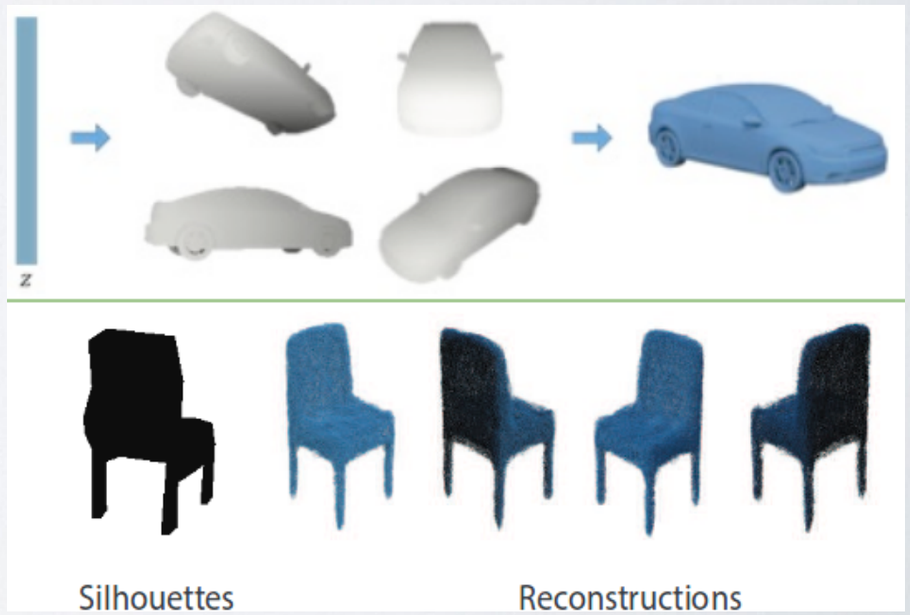
Natural language processing



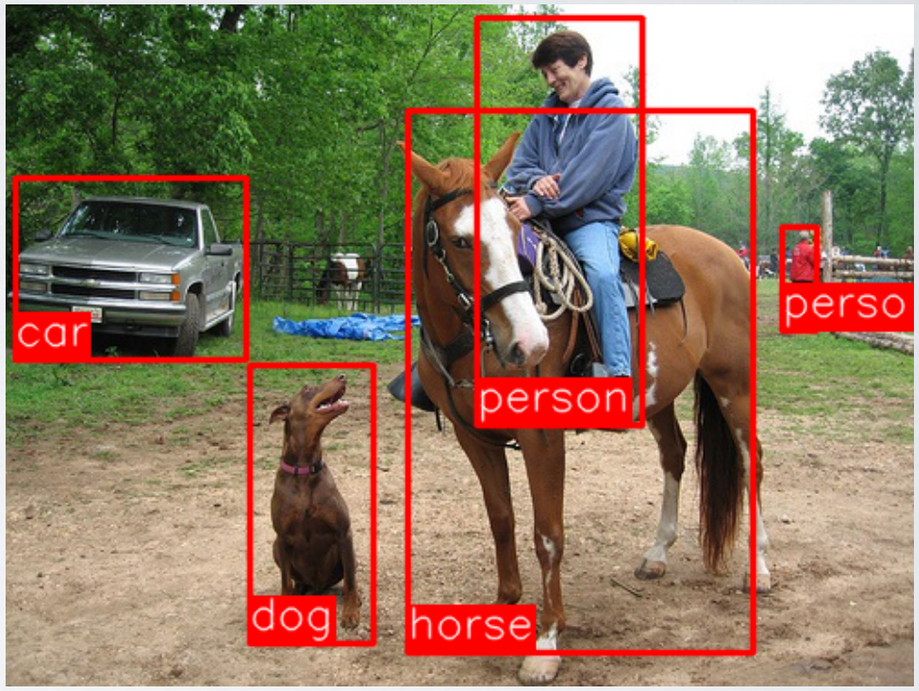
Computer vision

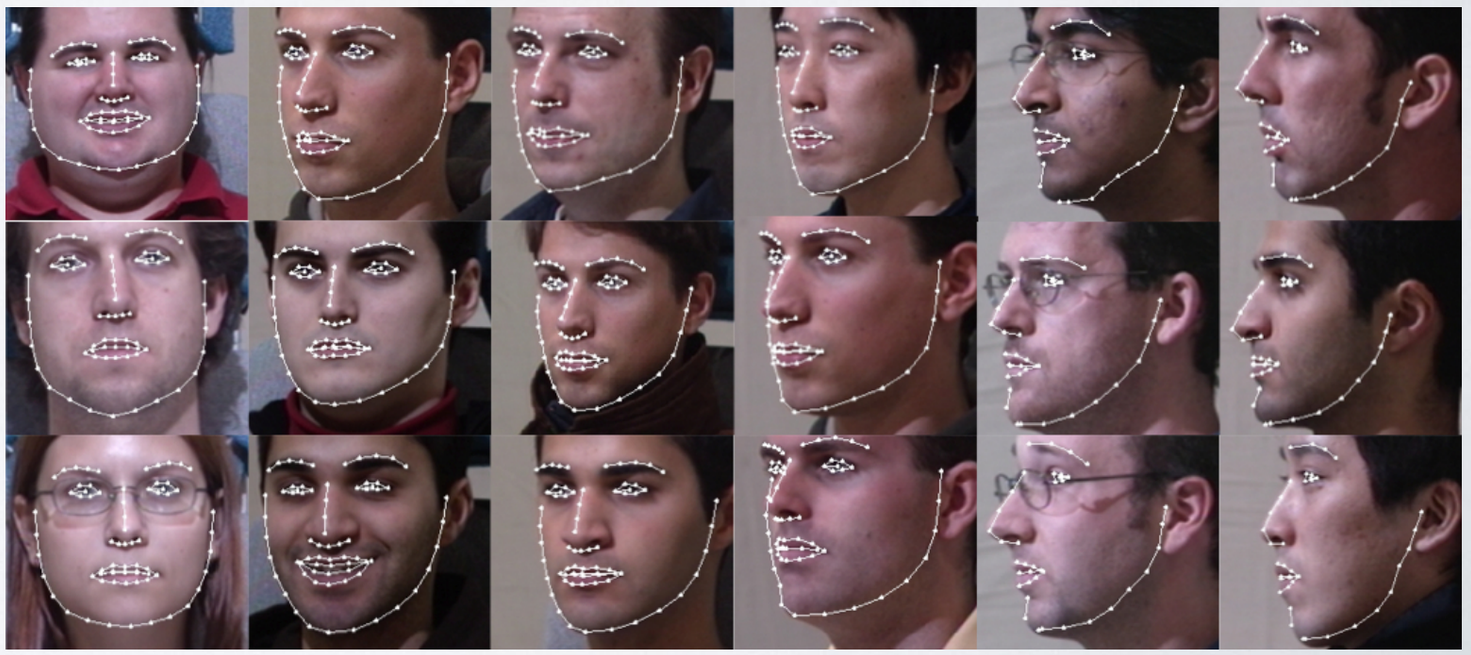


Computer graphics and learning

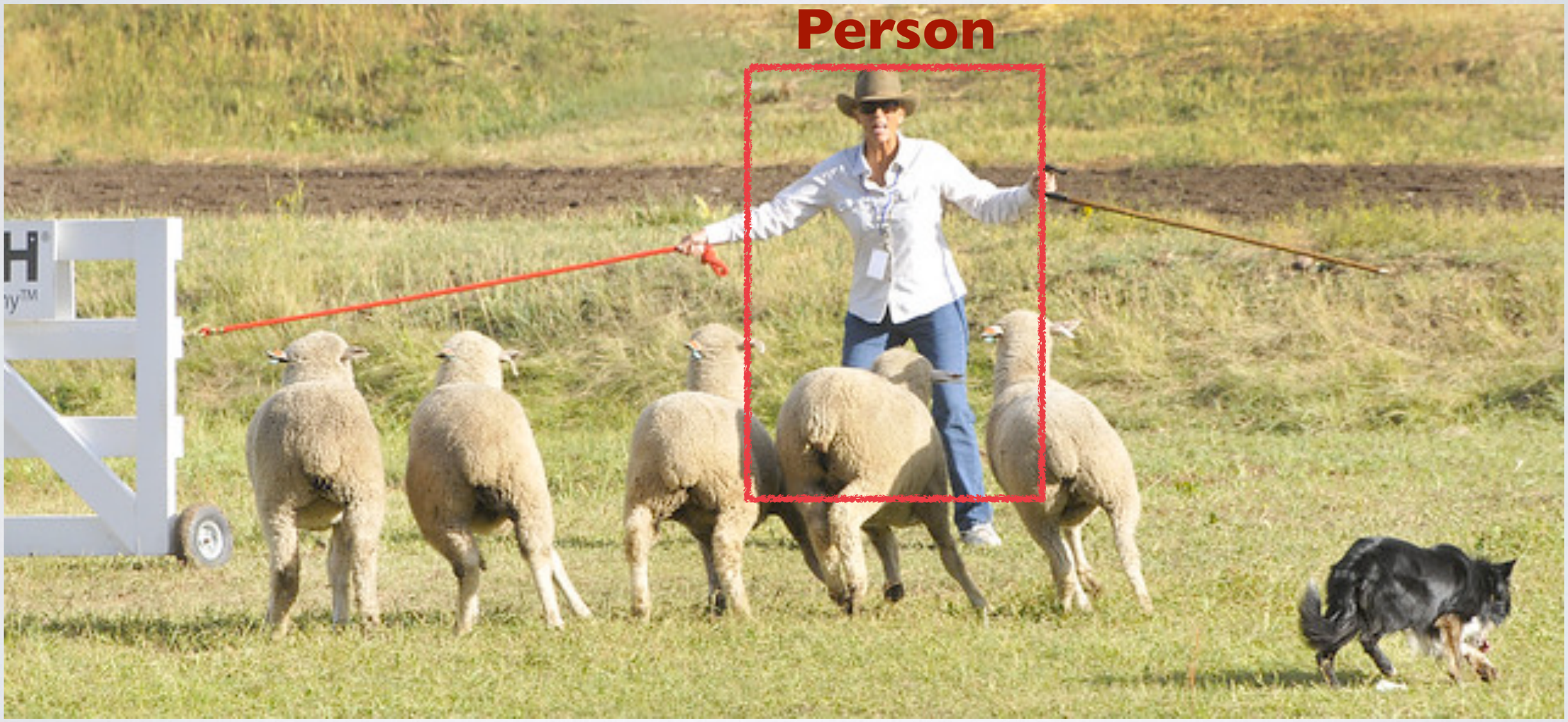


COMPUTER VISION AND DEEP LEARNING

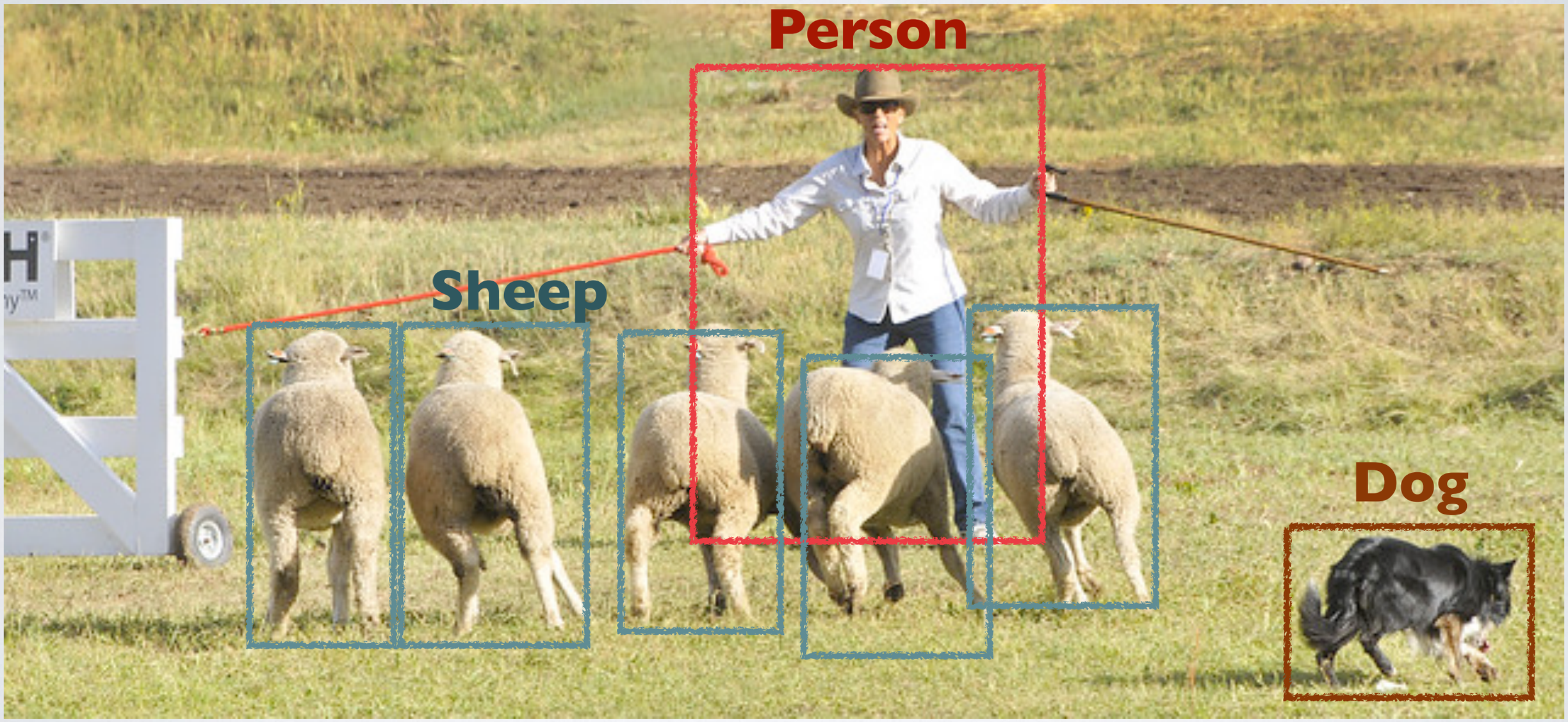




COMPUTER VISION 8 YEARS AGO

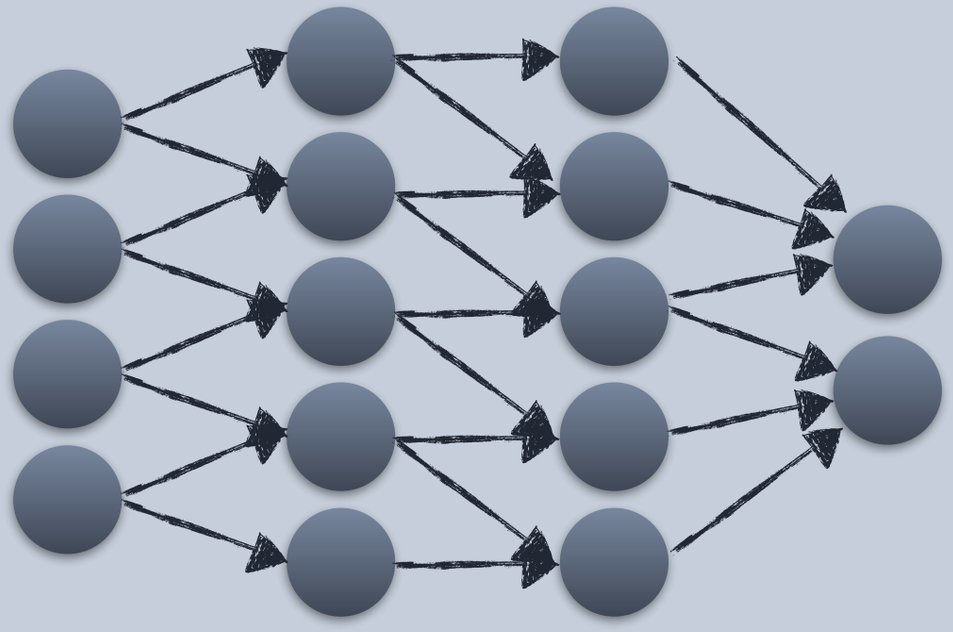


COMPUTER VISION TODAY



WHY DOES COMPUTER VISION WORK TODAY?

Deep Neural Networks



Labeled Data

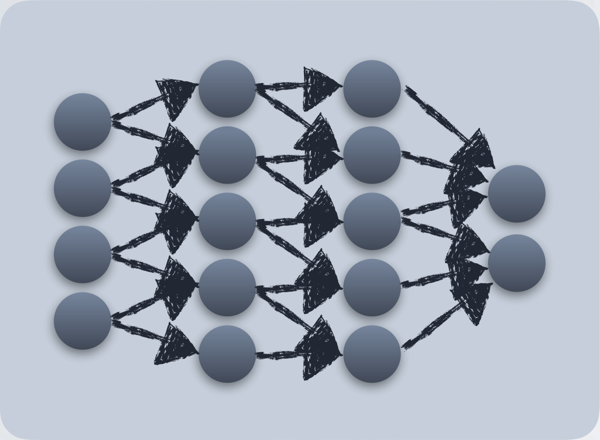
IMAGENET

1.2M images

1k categories

...

DEEP NEURAL NETWORKS



+

48.0% West Highland white terrier
46.5% Sealyham terrier
1.1% wire-haired fox terrier

CONVOLUTIONAL NETWORKS

CONVOLUTIONAL NETWORKS

A RECIPE FOR COMPUTER VISION

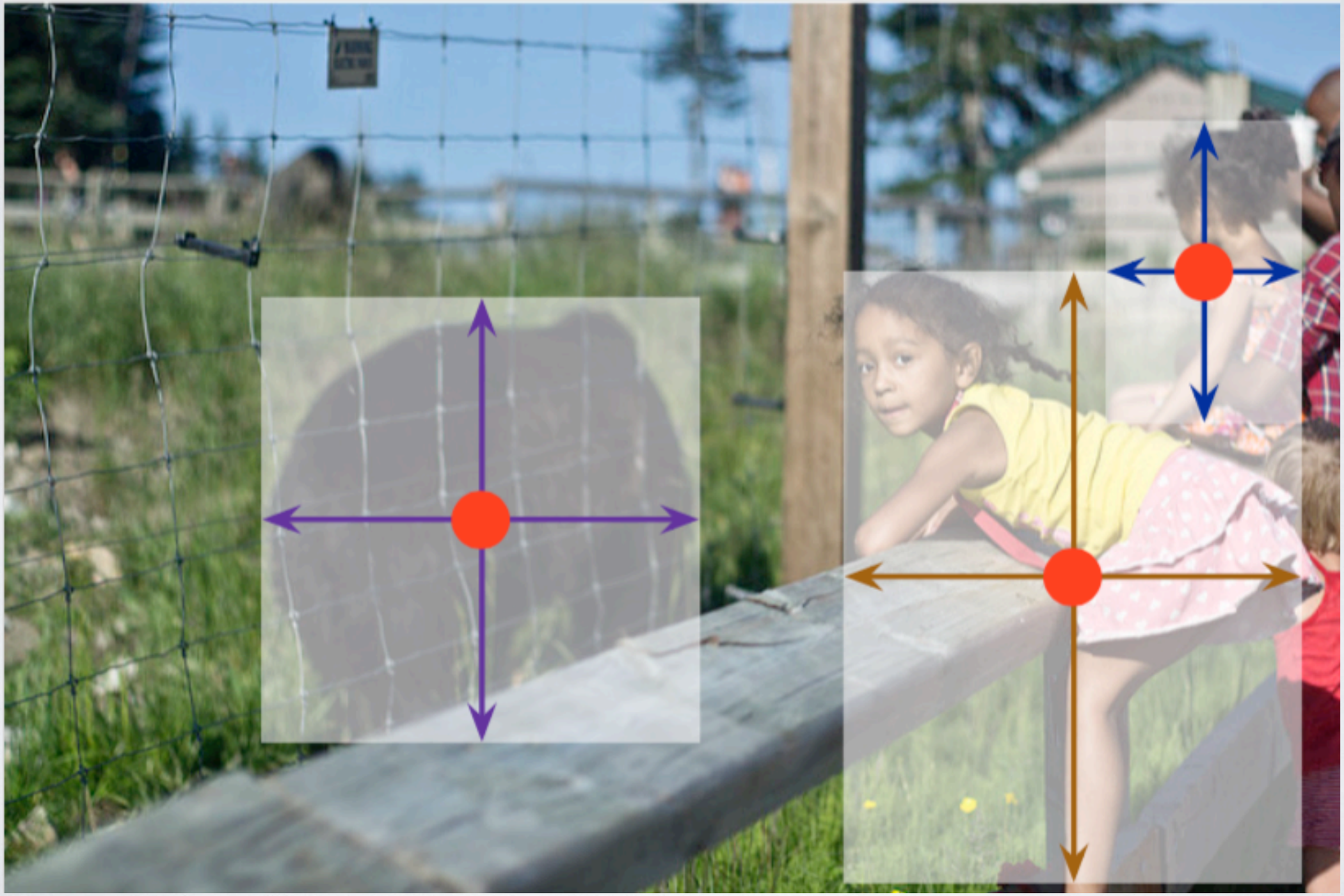
SPARSE AND DENSE CLASSIFICATION

DETECTION

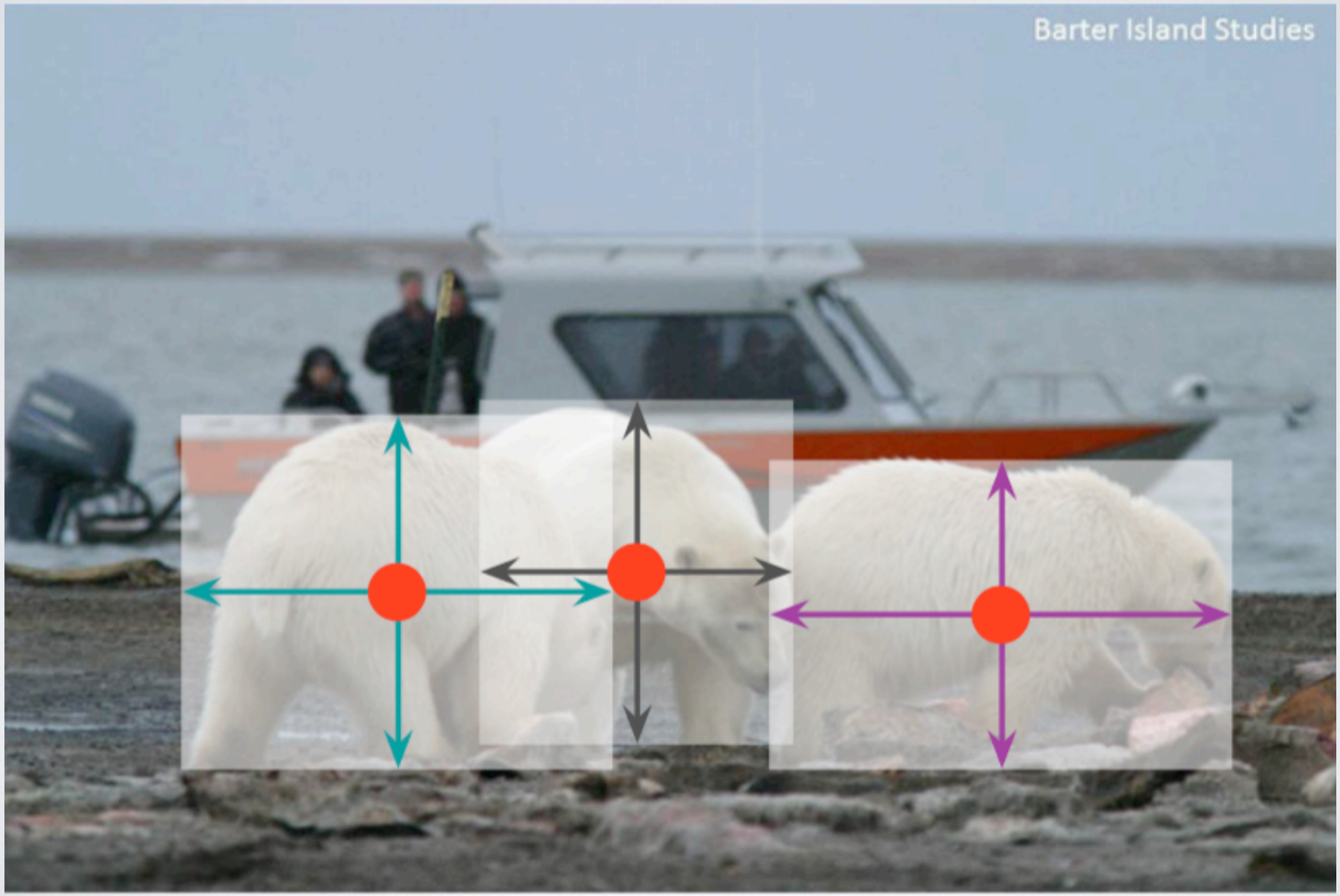
DETECTION

OBJECTS AS POINTS

OBJECTS AS POINTS



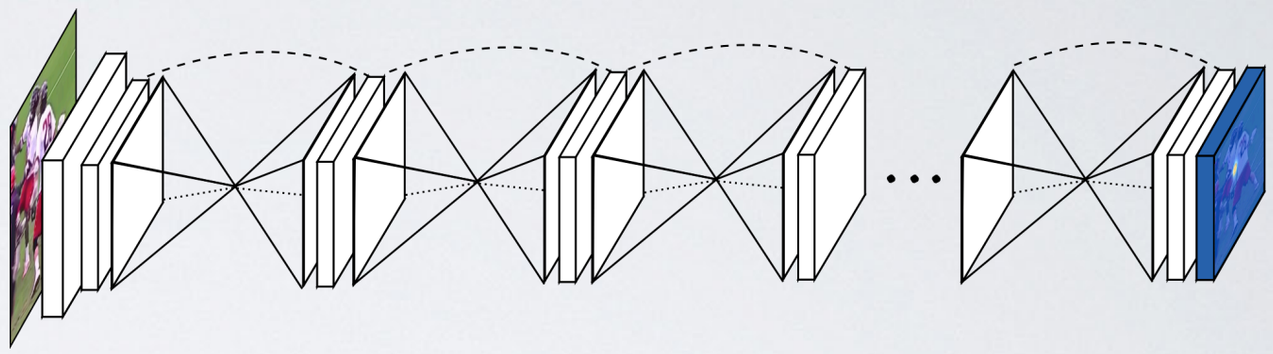
OBJECTS AS POINTS



OBJECTS AS POINTS

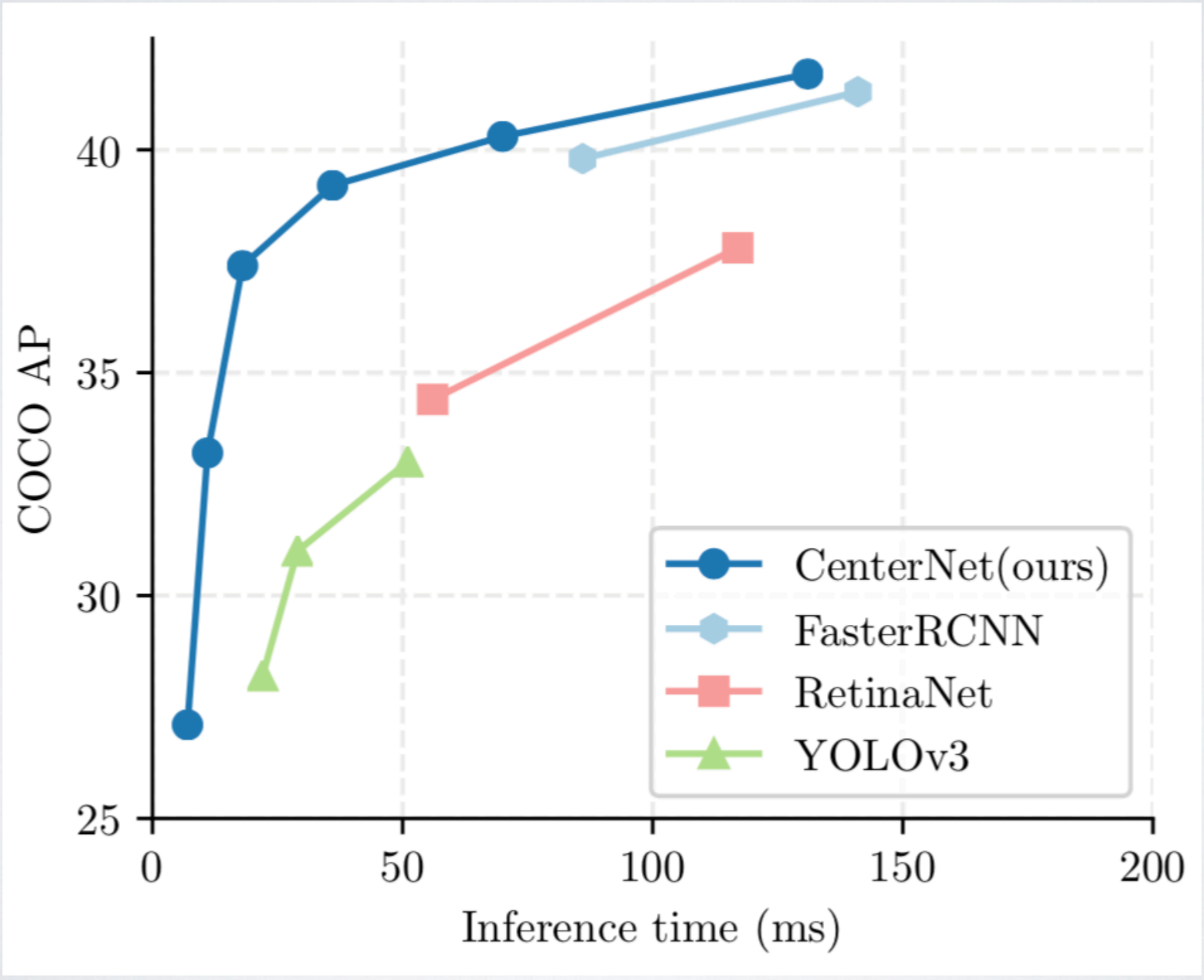
COCO
Common Objects in Context

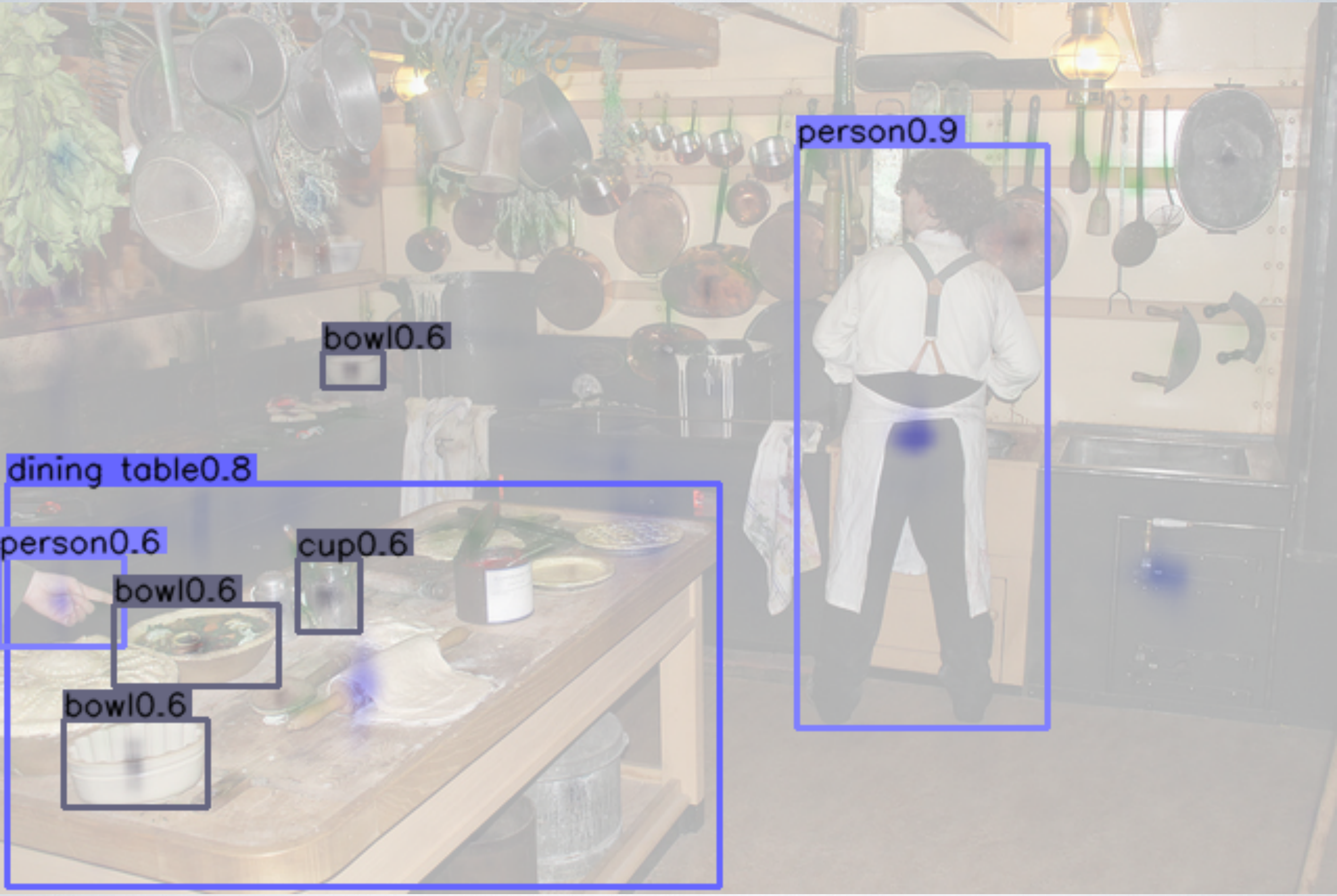
200k images
80 categories
1.5M labels

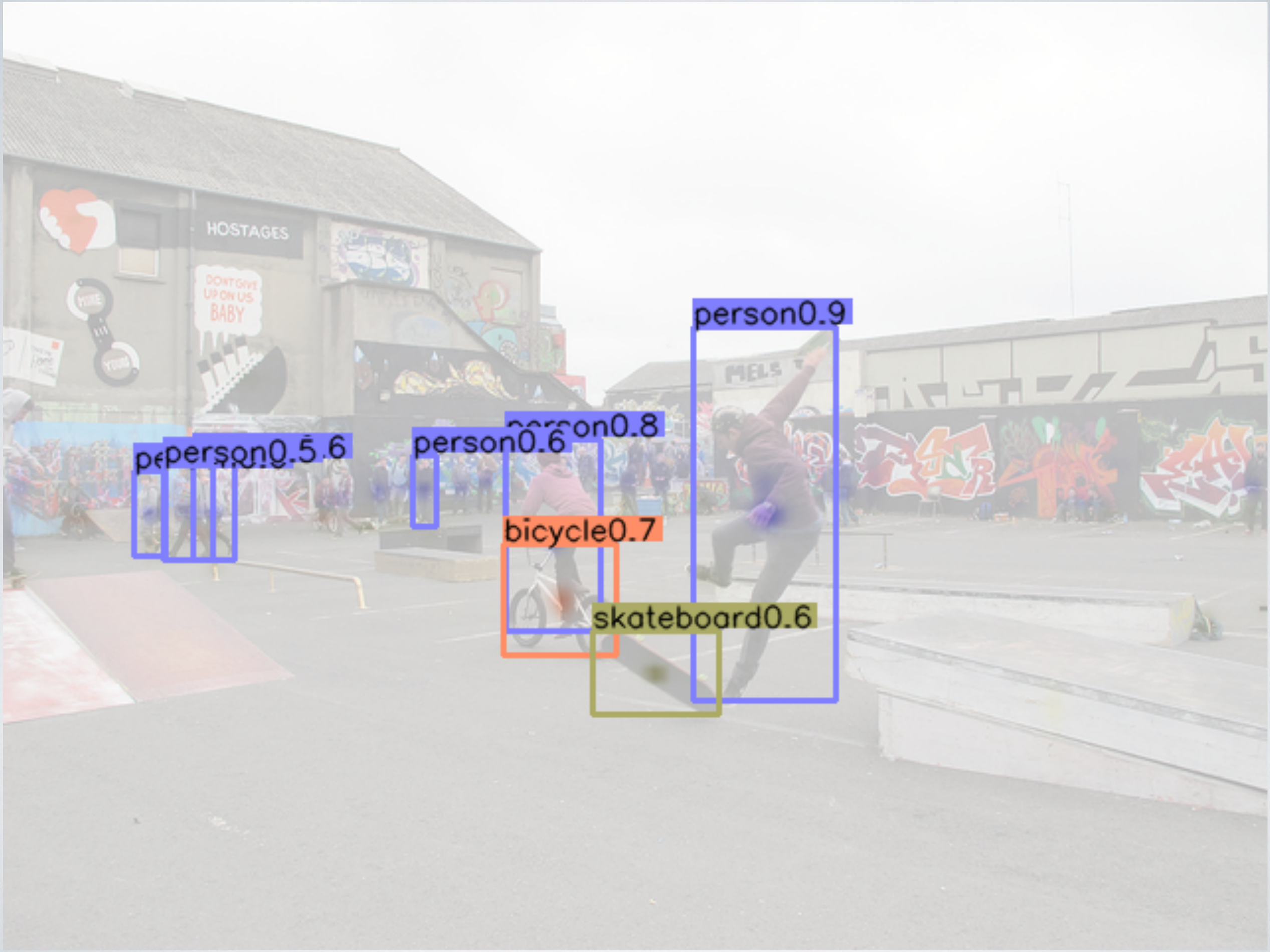


Keypoint prediction network

OBJECTS AS POINTS

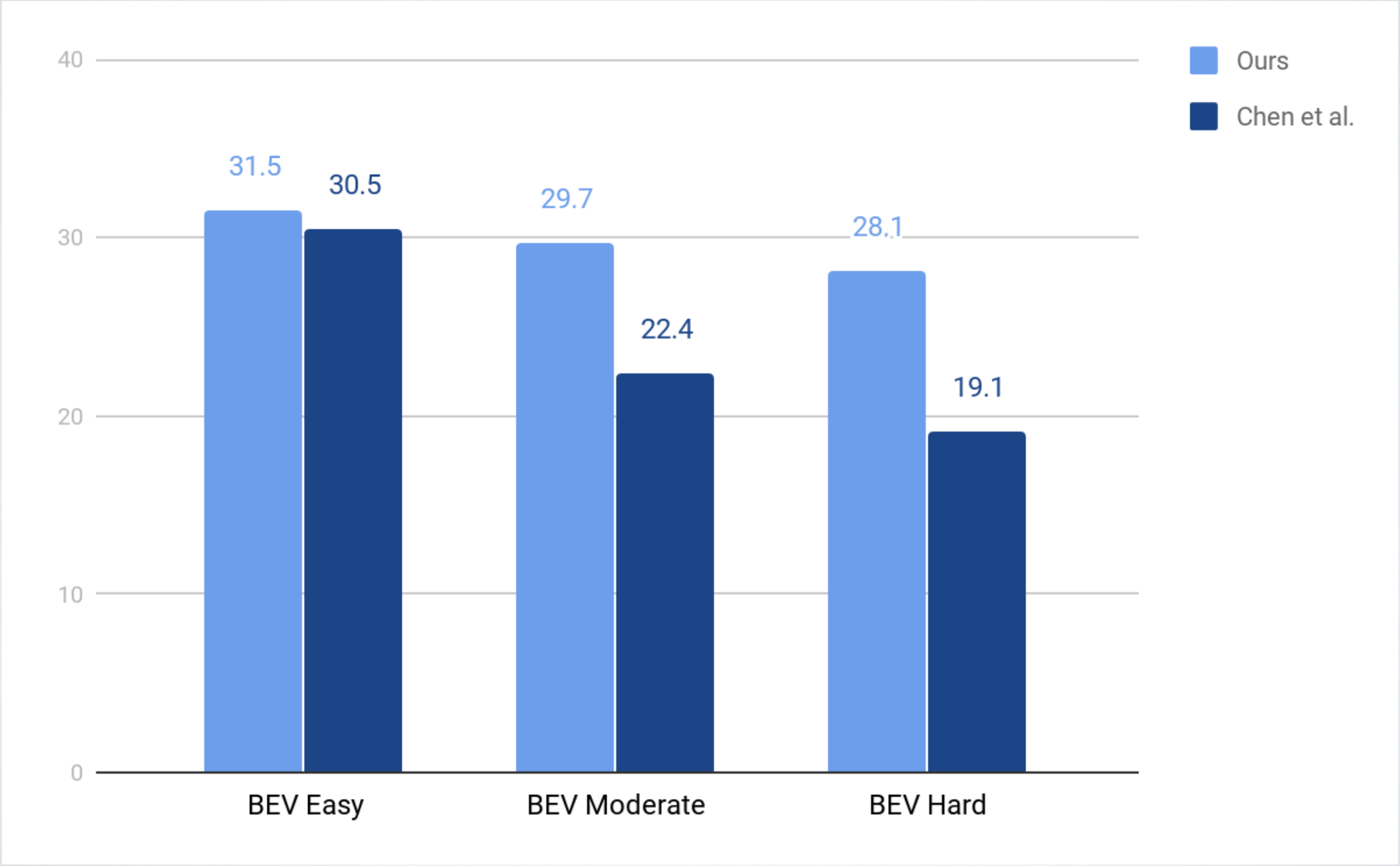



person0.9
bowl0.6
dining table0.8
person0.6
cup0.6
bowl0.6
bowl0.6

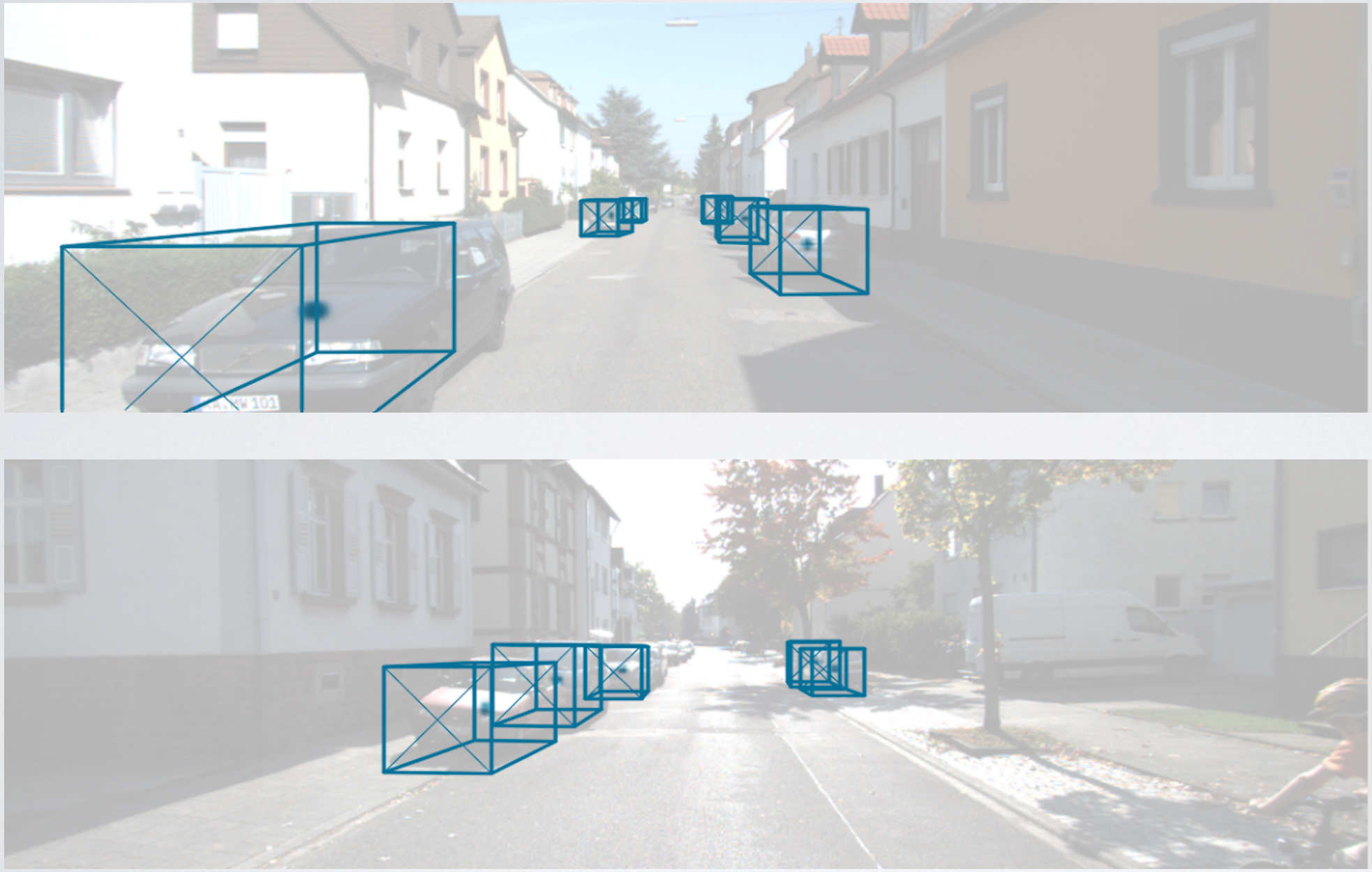

HOSTAGES
DONT GIVE UP ON US BABY
person0.9
person0.8
person0.6
person0.5
bicycle0.7
skateboard0.6

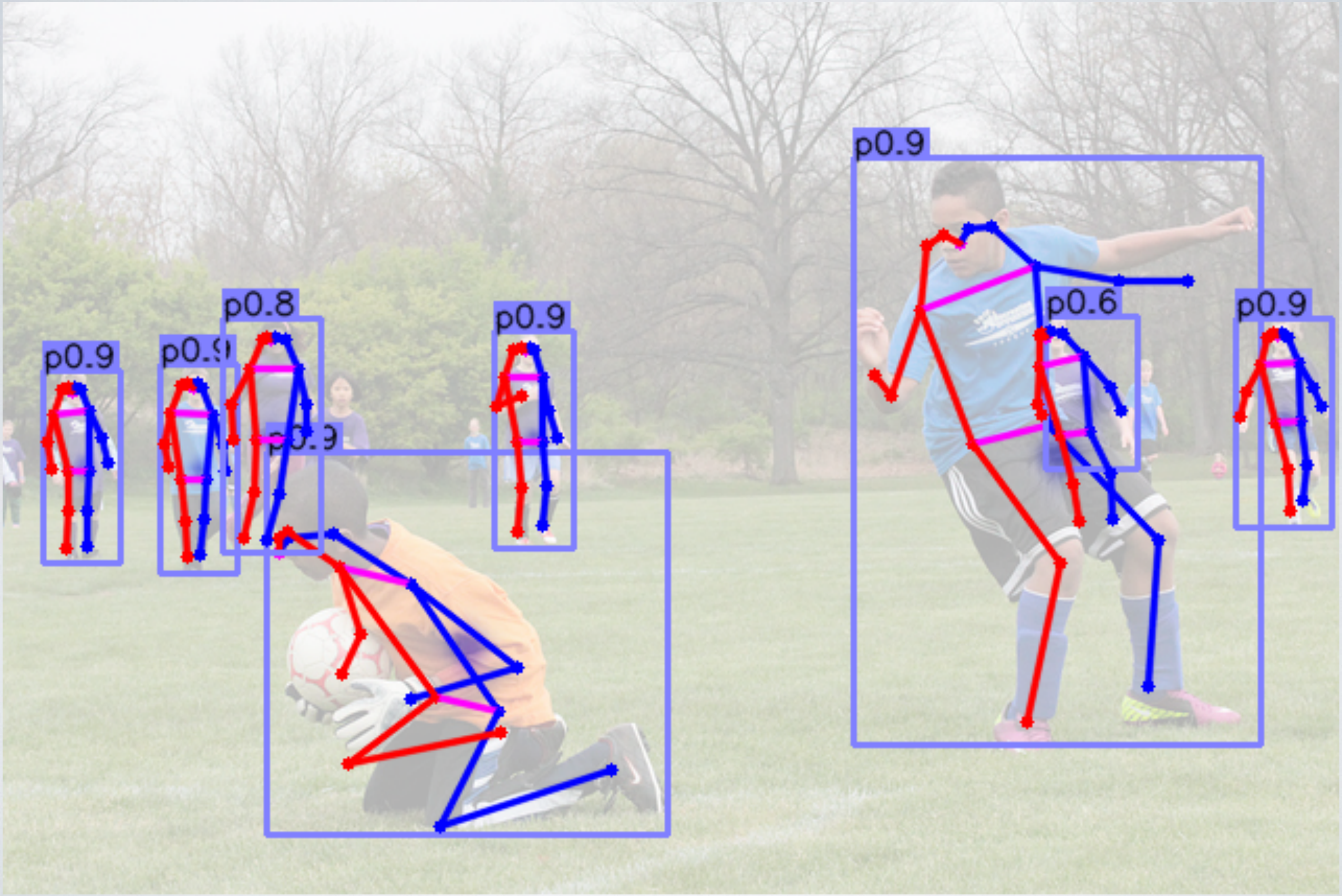

car0.5
person0.5
car0.8
person0.5
car0.8
car0.8
car0.6

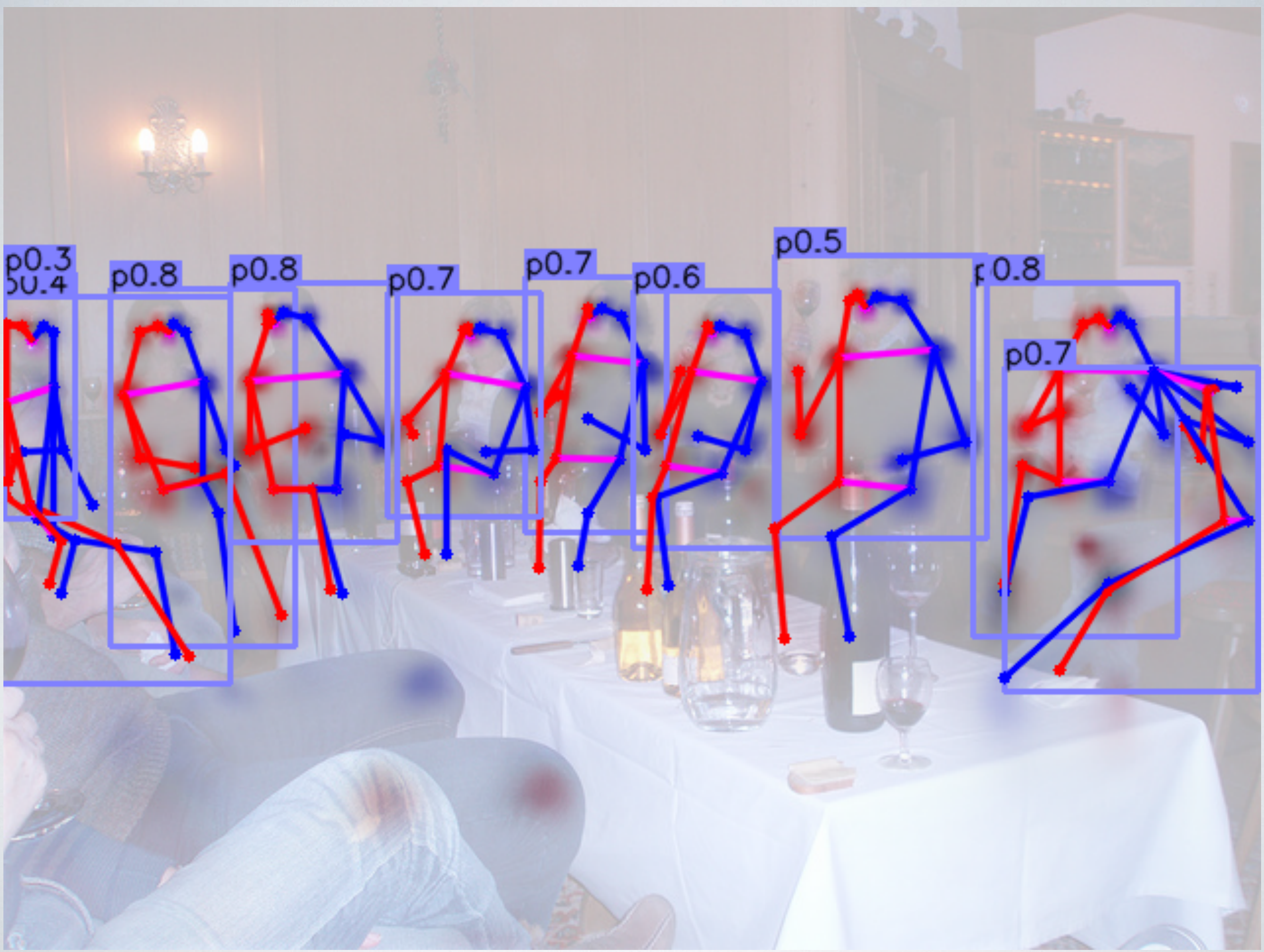
3D LAYOUT ESTIMATION



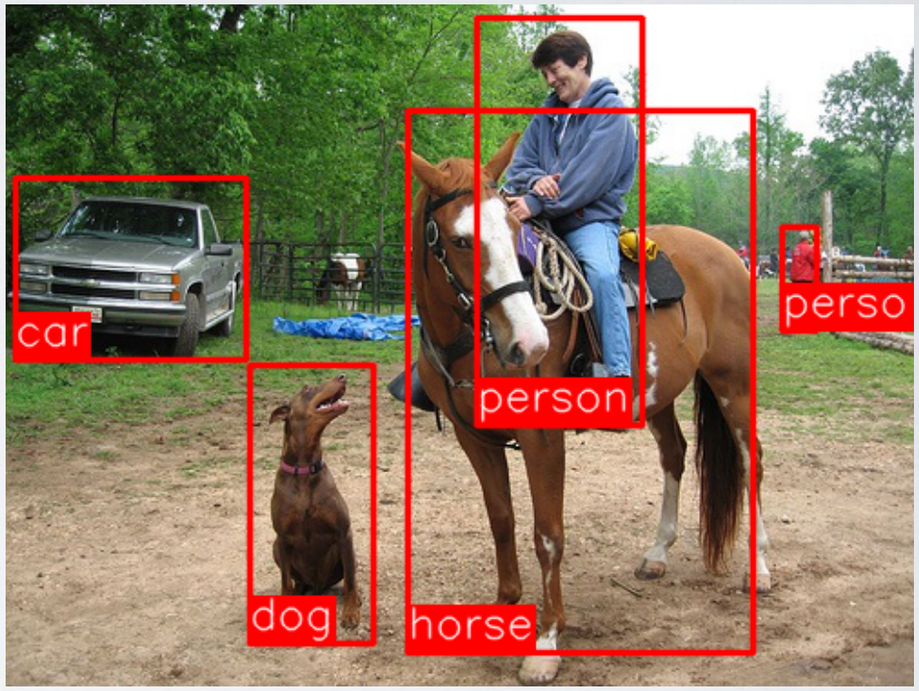
3D LAYOUT ESTIMATION

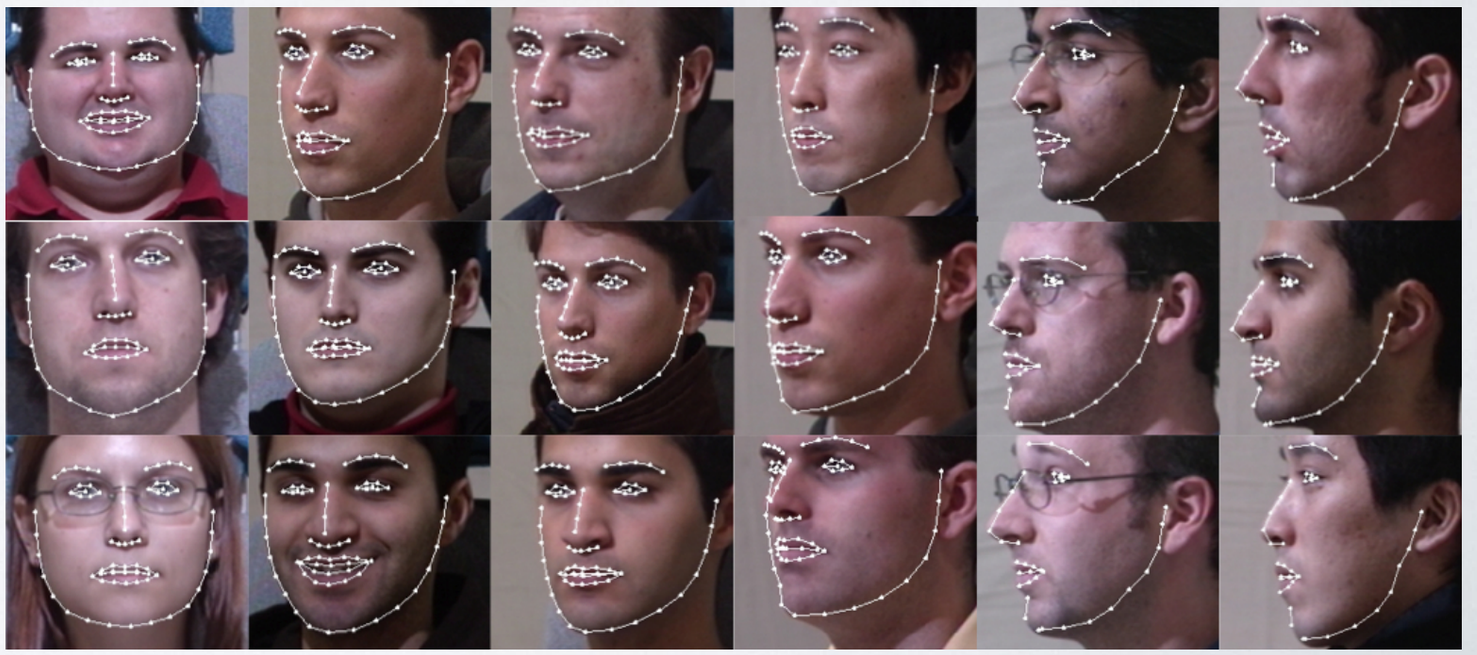



p0.9
p0.9
p0.8
p0.9
p0.9
p0.6
p0.9


p0.3
p0.8
p0.8
p0.7
p0.7
p0.6
p0.5
p0.8
p0.7

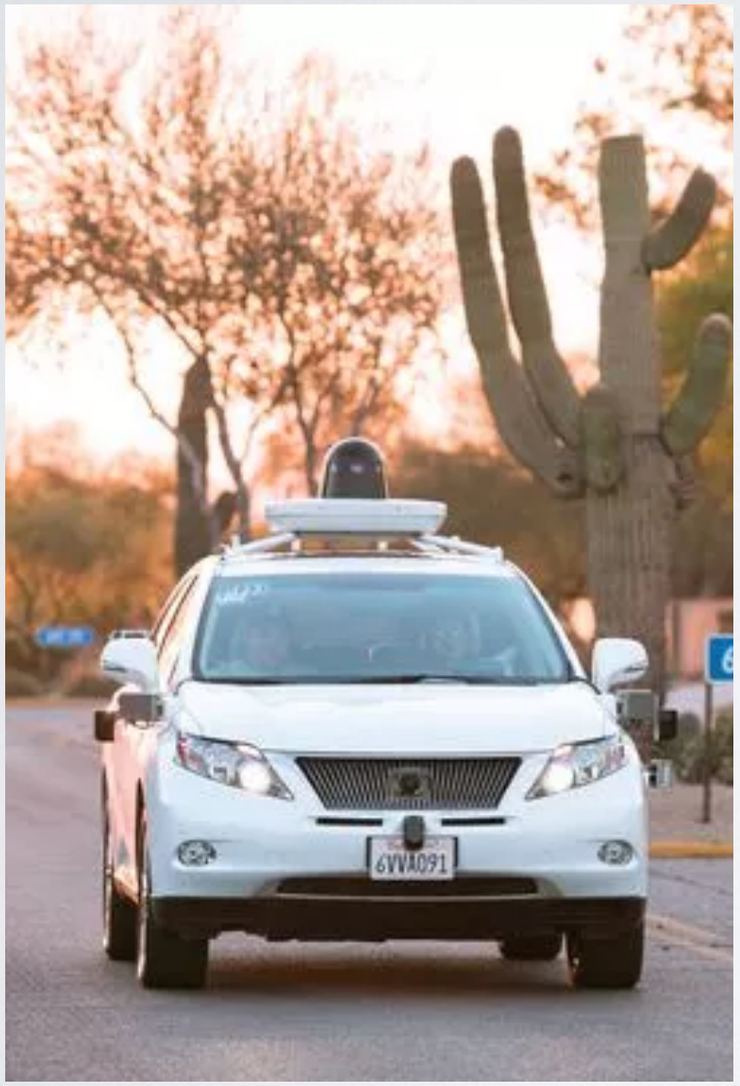
COMPUTER VISION TODAY





OBJECTS THROUGH TIME

MANY DATA SOURCES ARE VIDEO-BASED



IMAGES

- Single snapshots in time
  - Boring
  - Unnatural
  - Incomplete

A FEW IMAGES CAPTURE ARE ENOUGH INFORMATION

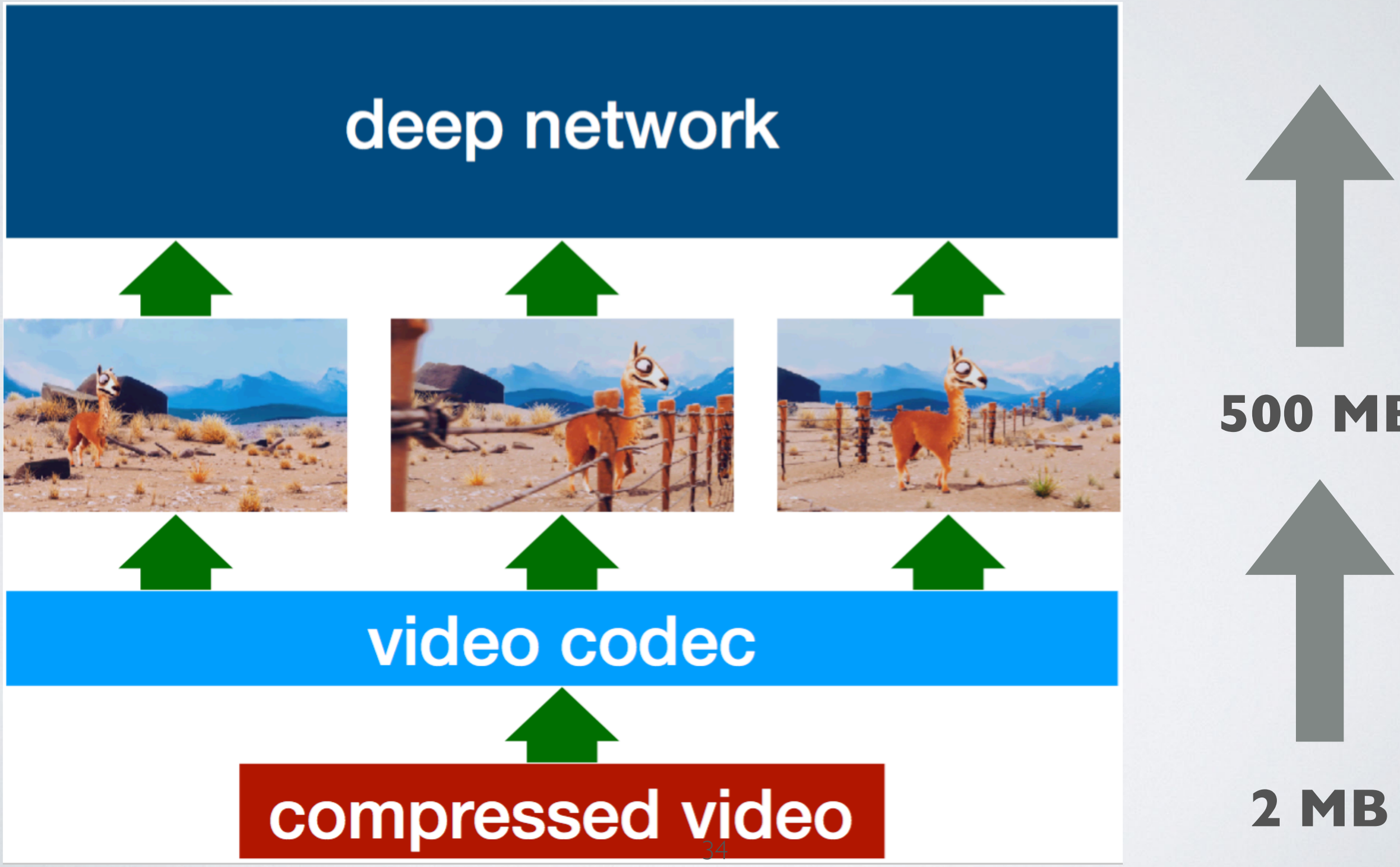
A FEW IMAGES CAPTURE ARE ENOUGH INFORMATION

MOST PHOTO CAMERAS CAPTURE IMAGES

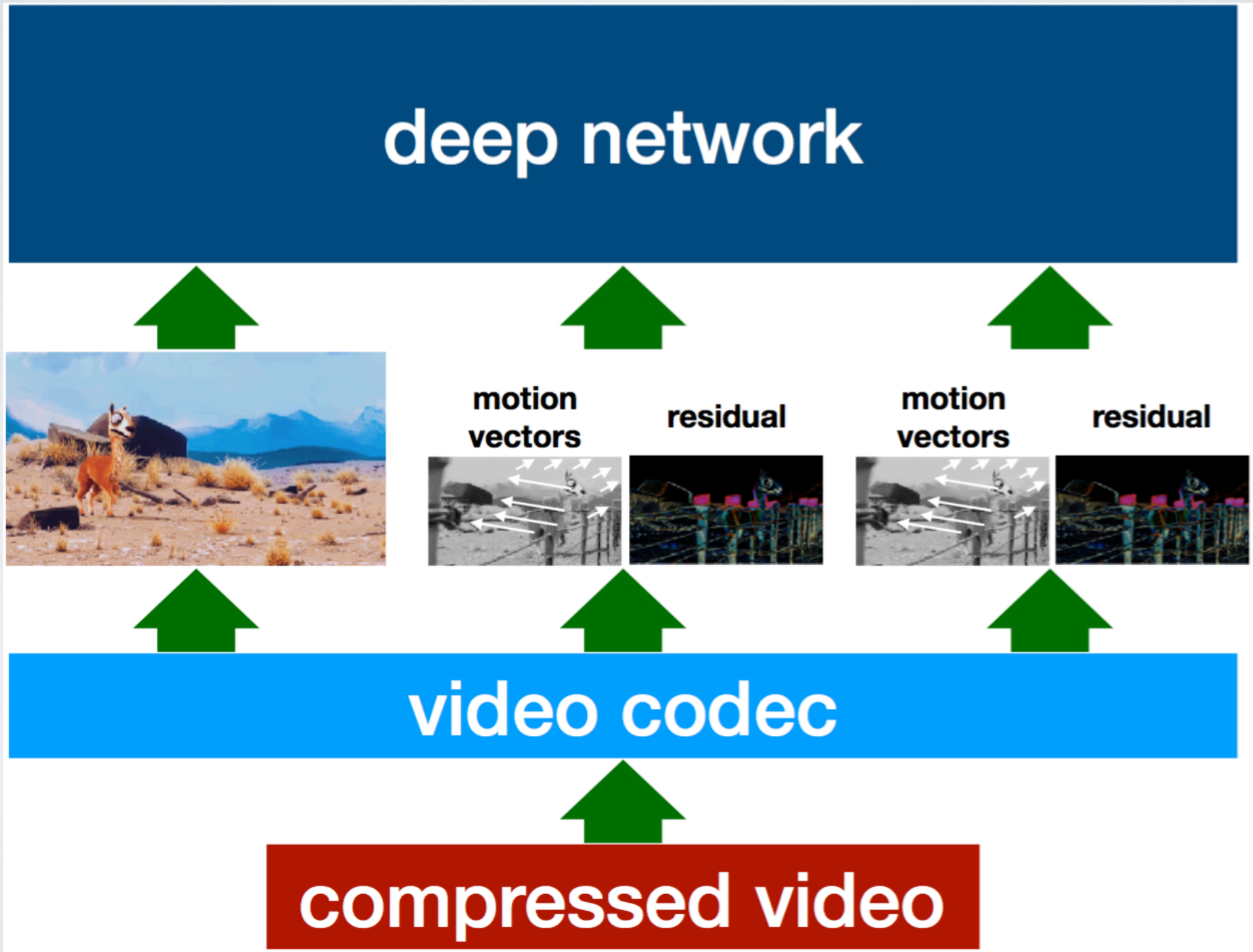
JPEG: **0.6 Mb**

h264: **1.8 Mb**

TRADITIONAL MODELS



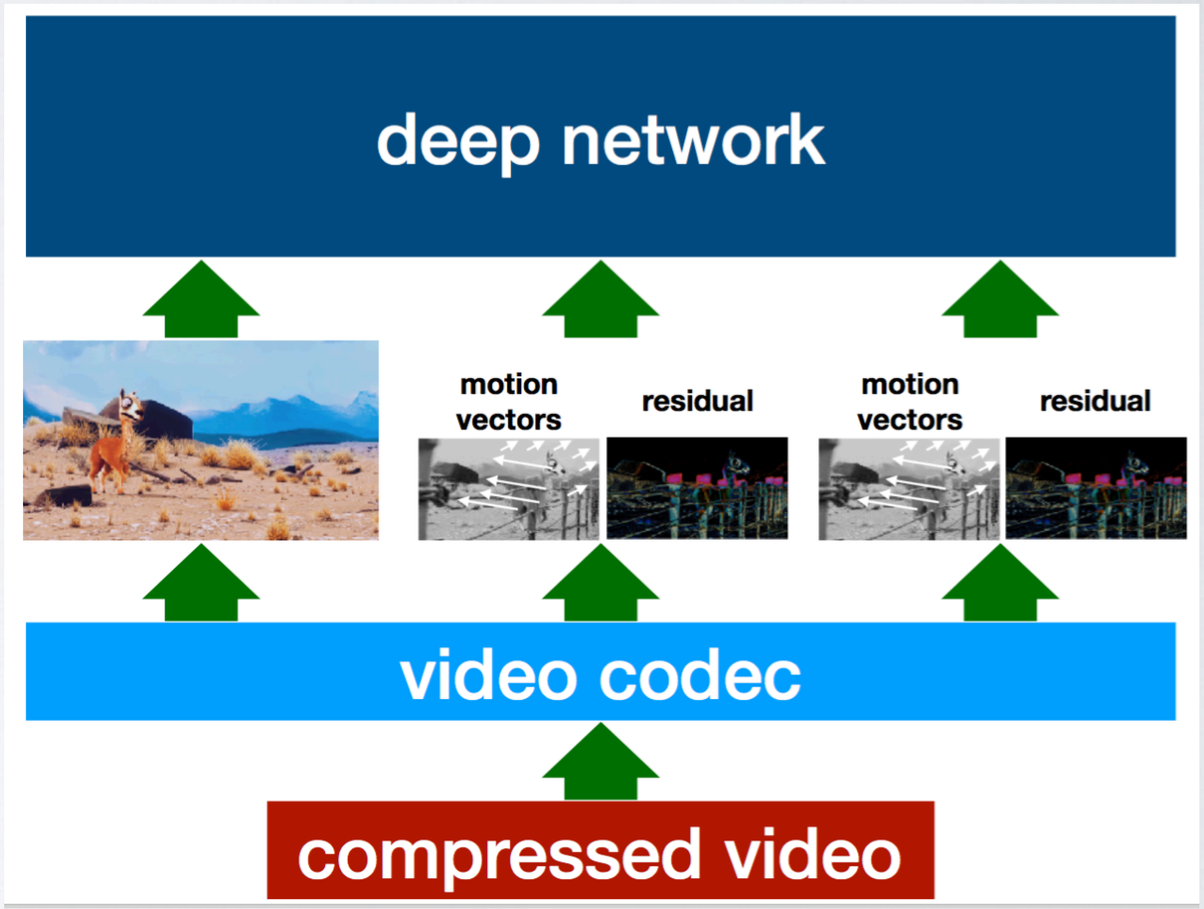
COVIAR



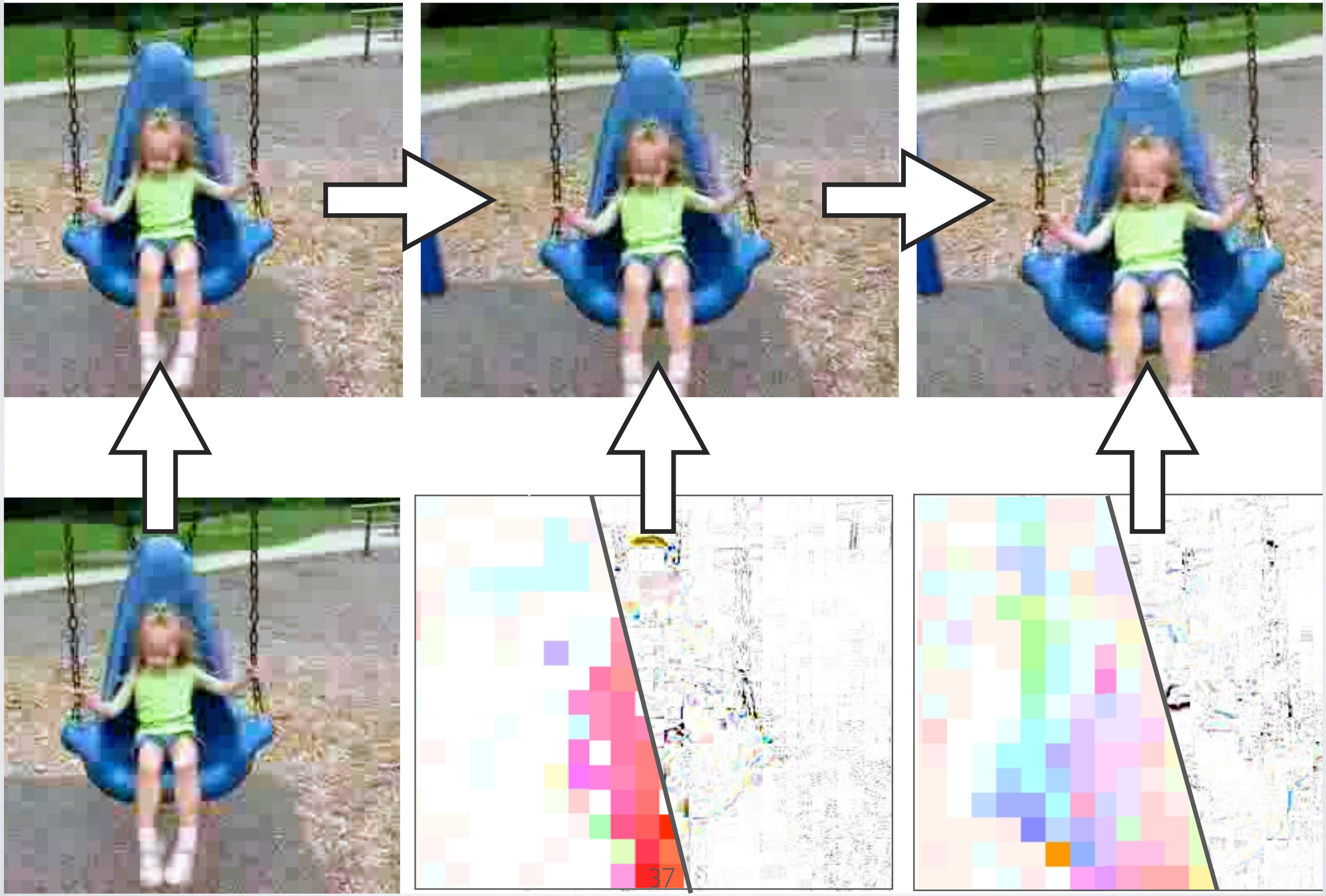
Compressed Video Action Recognition, Chao-Yuan Wu, M. Zaheer, H. Hu, R. Manmatha, A. Smola, P. Krähenbühl, CVPR 2018

COVIAR

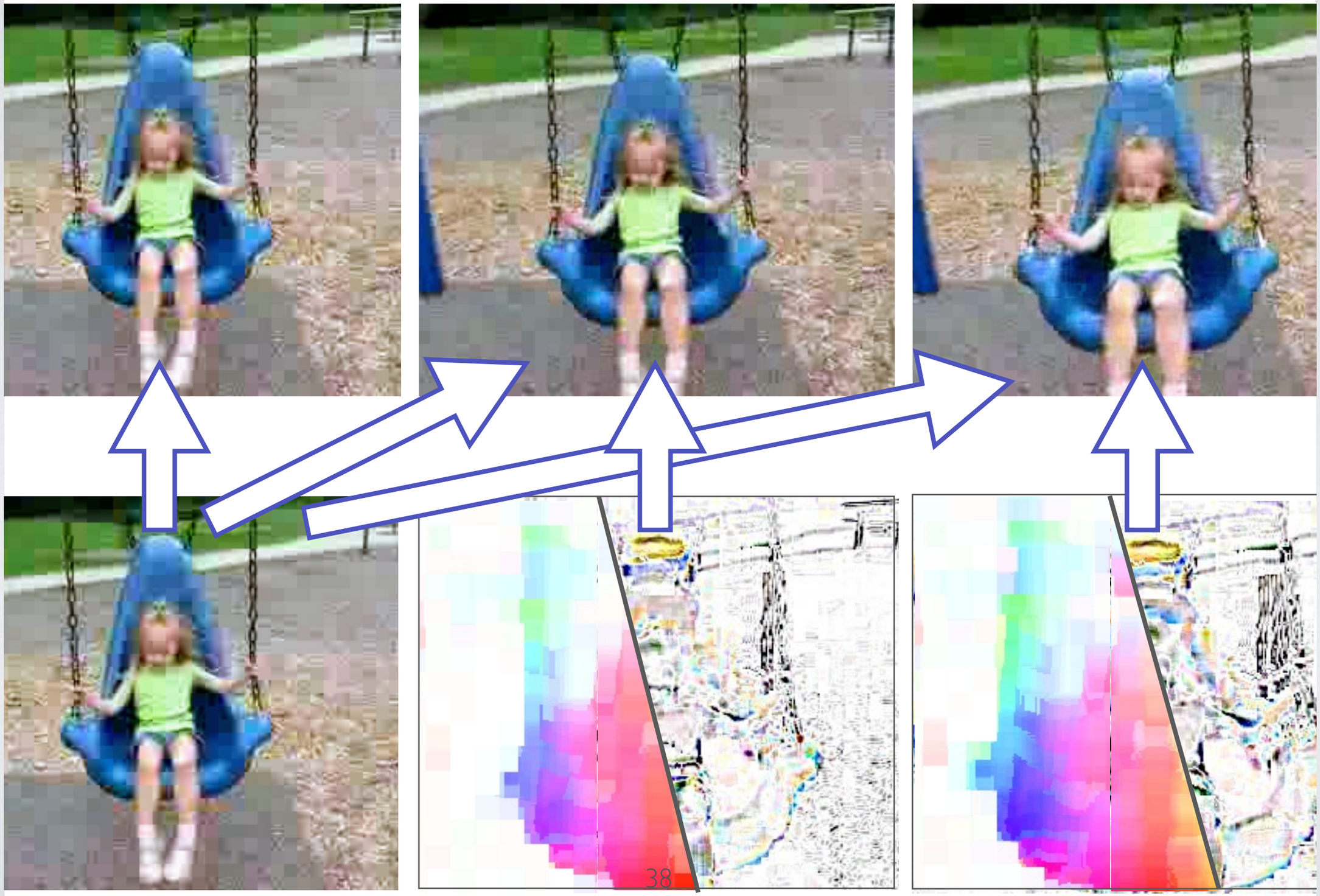
- train network directly on compressed video stream
  - I-frames (images)
  - P-frames
    - motion compensation
    - difference image



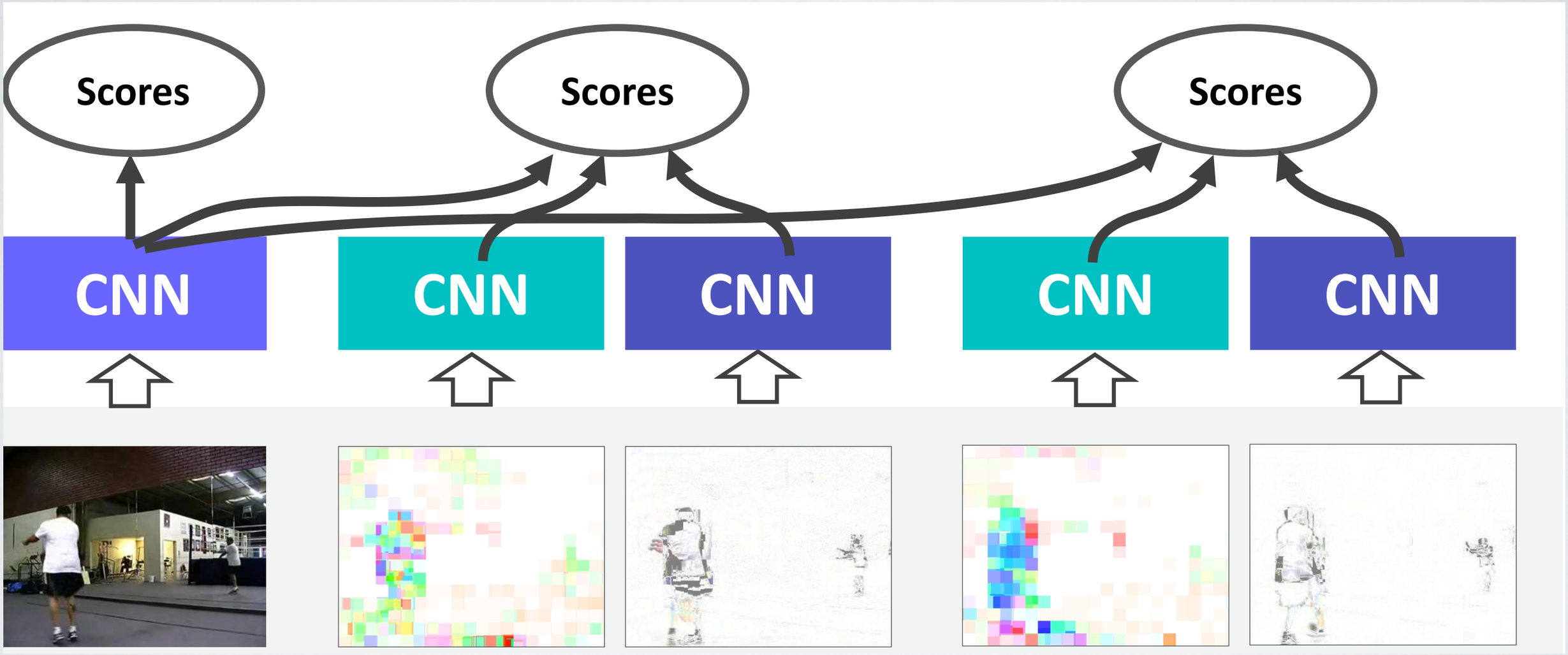
DEPENDENCIES



COVIAR: DEPENDENCIES



COVIAR



RESULTS: SPEED

	GFLOPs	Accuracy (%) UCF-101	HMDB-51
ResNet-50	**3.8**	82.3	48.9
ResNet-152	11.3	83.4	46.7
C3D	38.5	82.3	51.6
Res3D	19.3	<u>85.8</u>	<u>54.9</u>
CoViAR	<u>4.2</u>	**90.4**	**59.1**

RESULTS

	UCF-101	HMDB-51
Without optical flow		
Karpathy *et al.* [14]	65.4	-
ResNet-50 [12] (from ST-Mult [8])	82.3	48.9
ResNet-152 [12] (from ST-Mult [8])	83.4	46.7
C3D [36]	82.3	51.6
Res3D [37]	85.8	<u>54.9</u>
TSN (RGB-only) [41]*	85.7	-
TLE (RGB-only) [5]†	<u>87.9</u>	54.2
I3D (RGB-only) [2]*	84.5	49.8
MV-CNN [46]	86.4	-
Attentional Pooling [10]	-	52.2
CoViAR	**90.4**	**59.1**

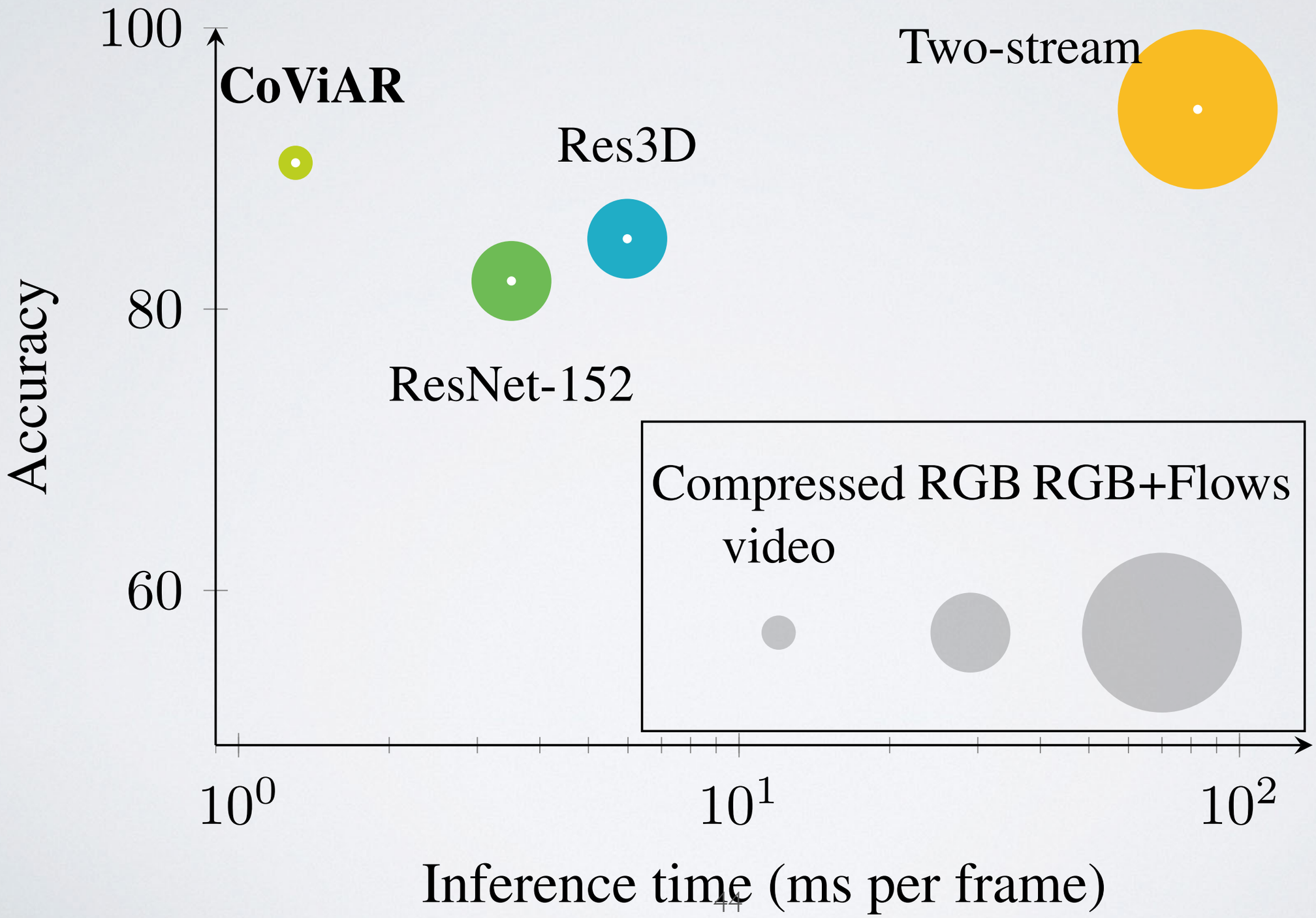
RESULTS

	UCF-101	HMDB-51
With optical flow		
iDT+FV [39]	-	57.2
Two-Stream [30]	88.0	59.4
Two-Stream fusion [9]	92.5	65.4
LRCN [6]	82.7	
Composite LSTM Model [32]	84.3	44.0
ActionVLAD [11]	92.7	66.9
ST-ResNet [7]	93.4	66.4
ST-Mult [8]	94.2	68.9
I3D [2]*	93.4	66.4
TLE [5]†	93.8	68.8
L^2STM [34]	93.6	66.2
ShuttleNet [27]	<u>94.4</u>	66.6
STPN [42]	**94.6**	68.9
TSN [41]	94.2	<u>69.4</u>
CoViAR + optical flow	**94.9**	**70.2**

CHARADES

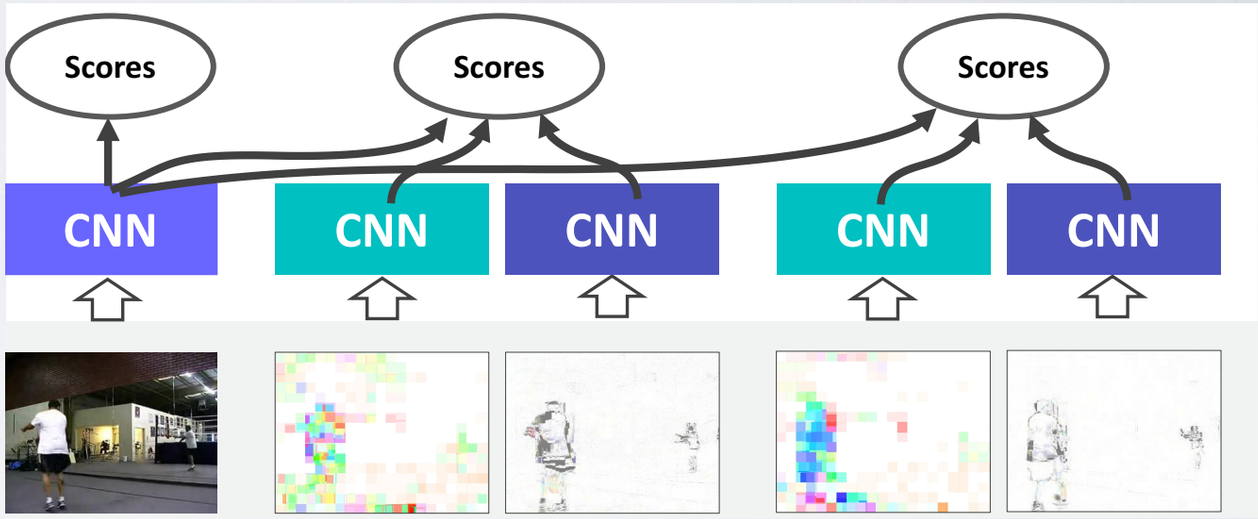
	mAP (%)	wAP (%)
Without optical flow		
ActionVLAD [11] (RGB only)	17.6	25.1
Sigurdsson *et al.* [28] (RGB only)	18.3	-
CoViAR	**21.9**	**29.4**
With optical flow		
Two-stream [30] (from [29])	14.3	-
Two-stream [30] + iDT [39] (from [29])	18.6	-
ActionVLAD [11] (RGB only) + iDT	21.0	29.9
Sigurdsson *et al.* [28]	22.4	-
CoViAR + optical flow	**24.1**	**32.3**

COVIAR


100
80
60
Accuracy
CoViAR
Res3D
Two-stream
ResNet-152
Compressed video
RGB
RGB+Flows
10^0
10^1
10^2
Inference time (ms per frame)

COVIAR

- maintains dependencies
- easy to train
- faster



QUESTIONS


MASTER OF COMPUTER SCIENCE ONLINE
Now accepting applications!
Fall 2019 deadline: June 1
$333 per credit hour
10 Courses
1.5 - 3 years
Top 10 University
On Your Schedule