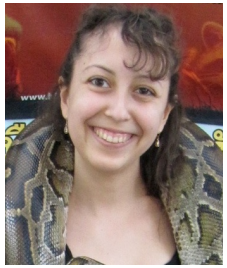
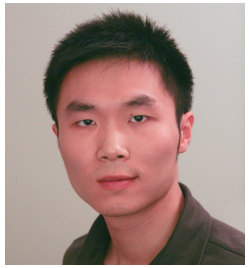
IMAGENET Large Scale Visual Recognition Challenge (ILSVRC)


Olga Russakovsky
(Stanford U.)

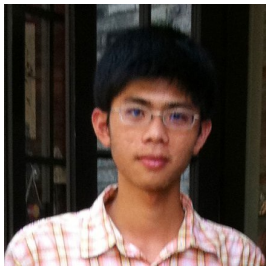

Jia Deng
(U. of Michigan)


Hao Su
(Stanford U.)

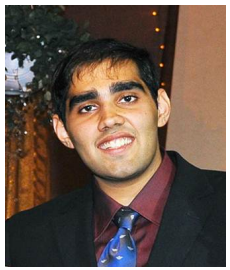

Jonathan Krause
(Stanford U.)


Sanjeev Satheesh
(Stanford U.)


Sean Ma
(Stanford U.)


Zhiheng Huang
(Stanford U.)


Andrej Karpathy
(Stanford U.)


Aditya Khosla
(MIT)


Michael Bernstein
(Stanford U.)


Alexander Berg
(UNC Chapel Hill)


Fei-Fei Li
(Stanford U.)

http://image-net.org/challenges/LSVRC/

Backpack

Flute

Strawberry

Traffic light

Backpack

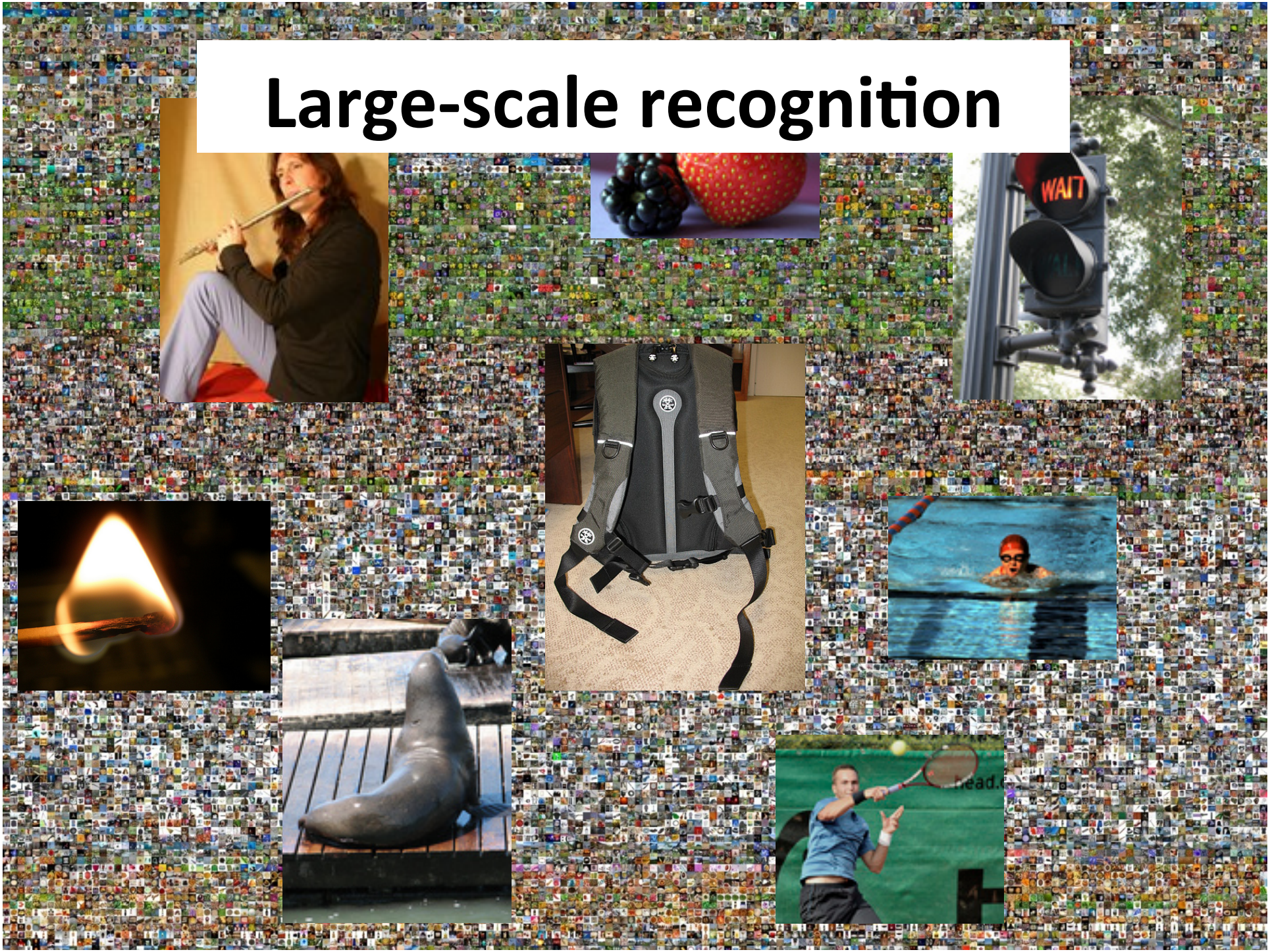
Matchstick

Sea lion

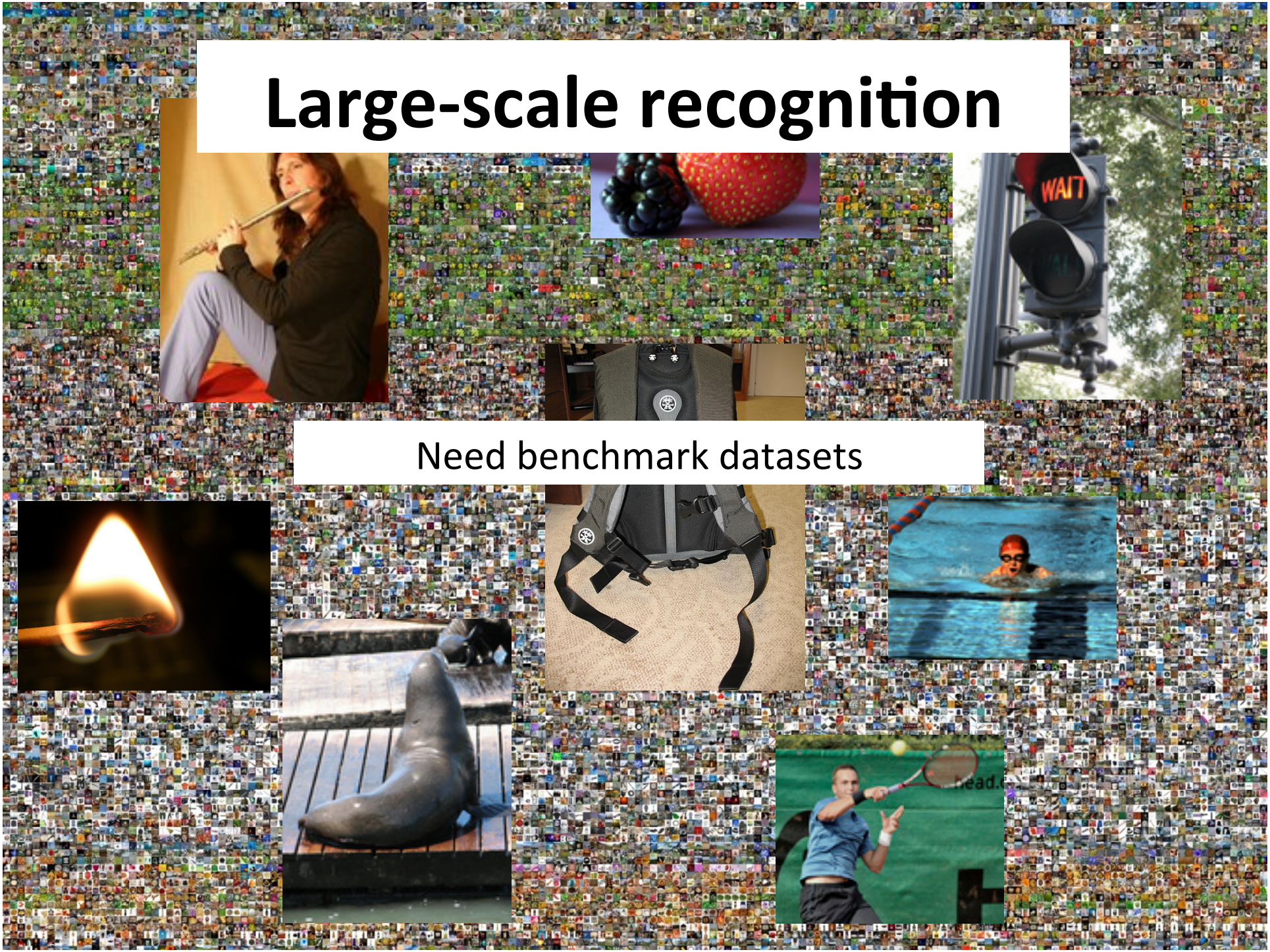
Bathing cap

Racket

Large-scale recognition



Large-scale recognition

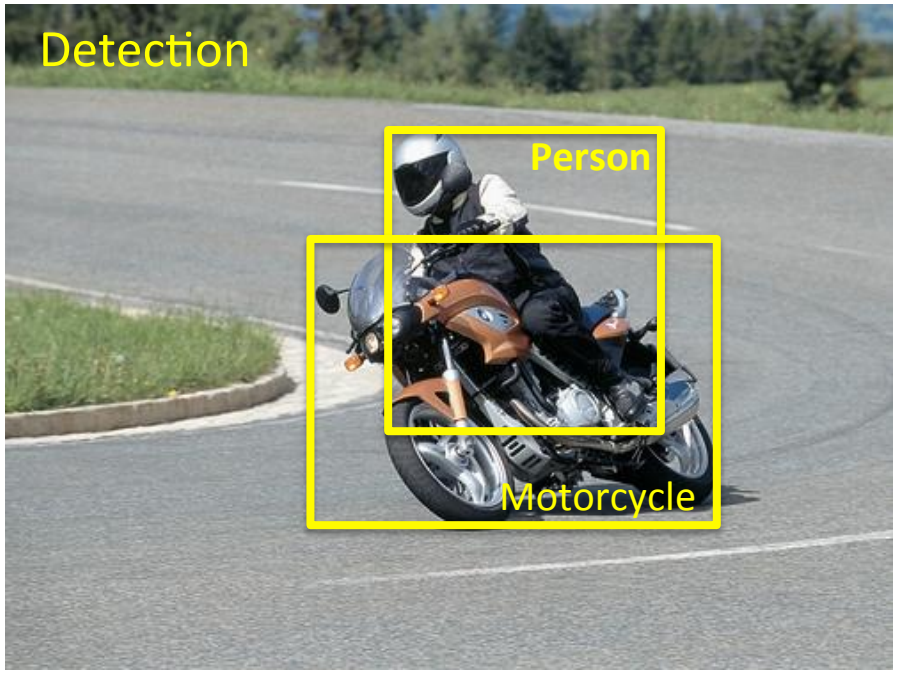


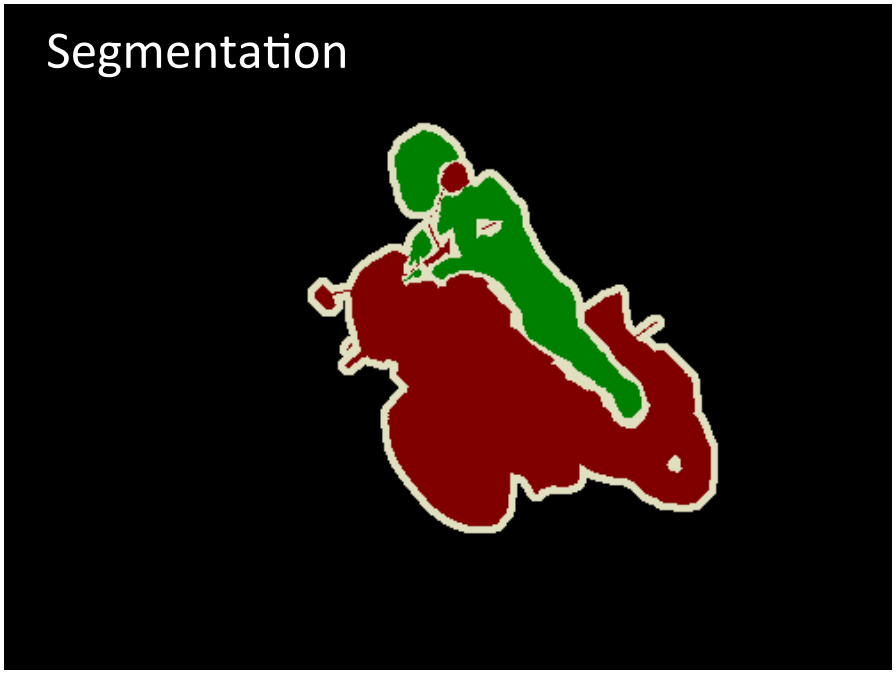
Need benchmark datasets

PASCAL VOC 2005-2012

20 object classes **22,591 images**

Classification: person, motorcycle





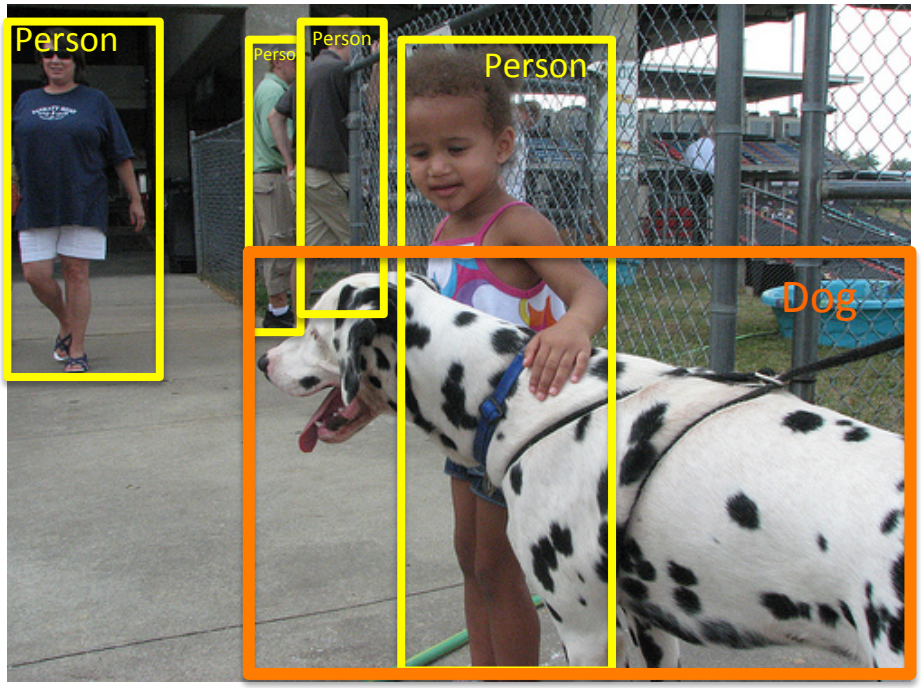
Action: riding bicycle

Everingham, Van Gool, Williams, Winn and Zisserman.
The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2010-2014

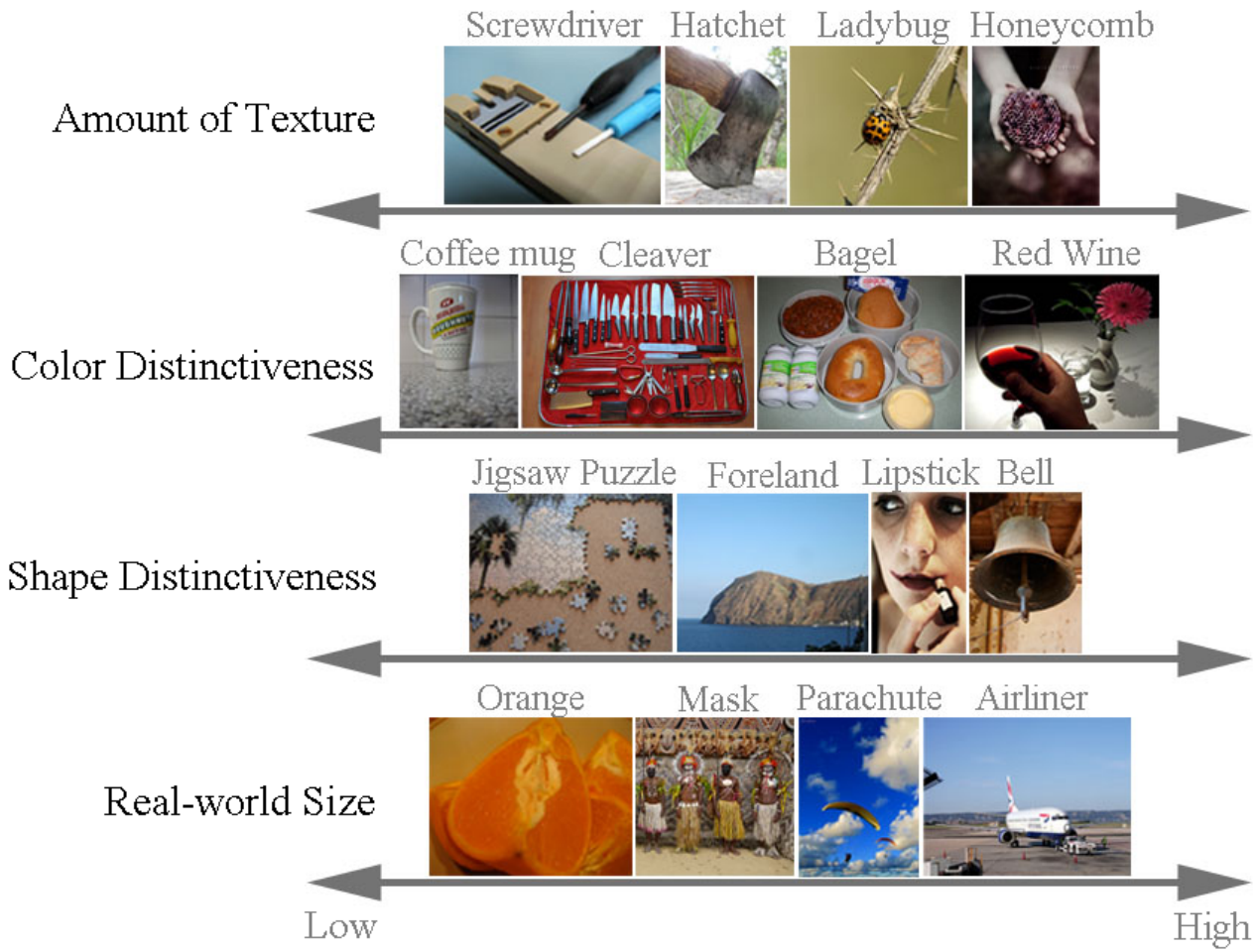
~~20 object classes 22,591 images~~

200 object classes	**517,840 images**	**DET**
1000 object classes	**1,431,167 images**	**CLS-LOC**



http://image-net.org/challenges/LSVRC/

Variety of object classes in ILSVRC



Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going? ICCV 2013
http://image-net.org/challenges/LSVRC/2012/analysis

Variety of object classes in ILSVRC

ILSVRC detection

ILSVRC classification and localization

birds

bird

flamingo cock ruffed grouse quail partridge . . .

bottles

bottle

pill bottle beer bottle wine bottle water bottle pop bottle . . .

cars

car

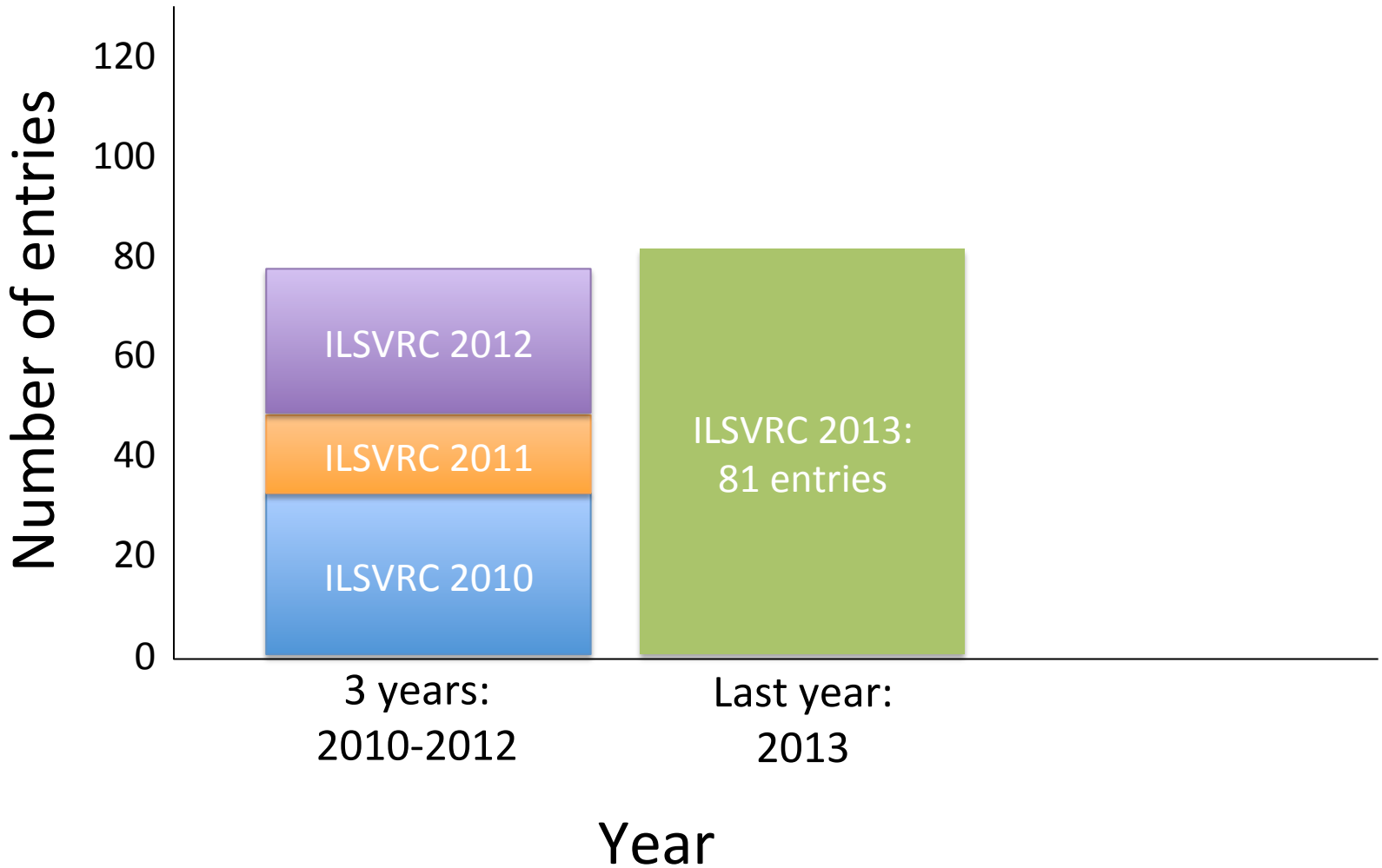
race car wagon minivan jeep cab . . .

Challenge procedure every year

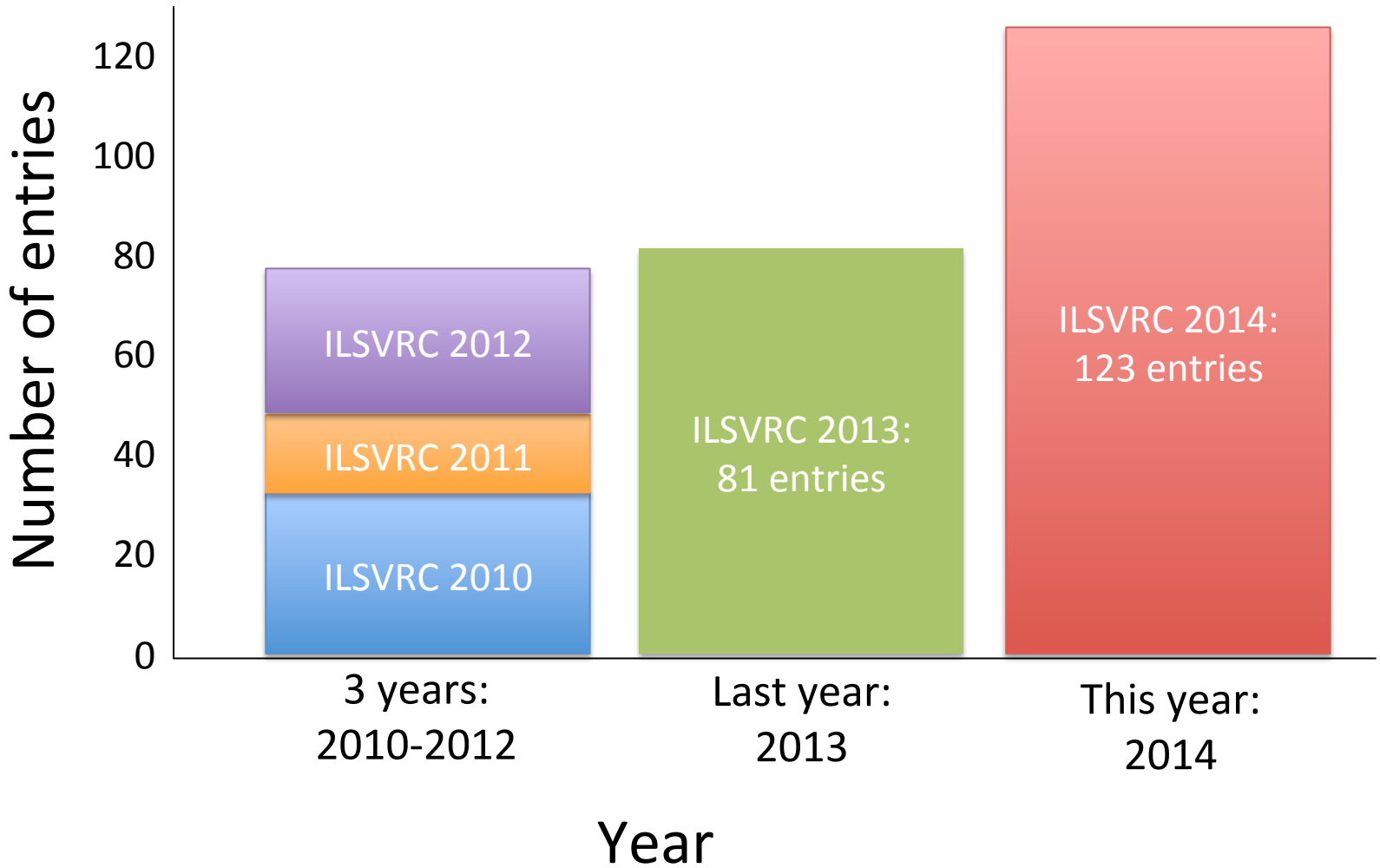
1. Training and validation data released: images and annotations
2. Test data released: images only (annotations hidden)
3. Participants train their models on train & validation data
4. Submit text file with predictions on test images
5. We evaluate and release results, and run a workshop

http://image-net.org/challenges/LSVRC/2014/eccv2014

Participation in ILSVRC over the years



Participation in ILSVRC over the years



Experiment this year: open vs closed submissions

- Offered all teams an option:
  - Open = promise to reveal their method
  - Closed = participate without revealing the method

- Almost all teams chose to be "open" (31/36)
  - And 2 of the "closed" teams still presented spotlights and posters at the workshop

ILSVRC in detail: history and current state-of-the-art

ImageNet Large Scale Visual Recognition Challenge

Olga Russakovsky*, Jia Deng*, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, Li Fei-Fei

http://arxiv.org/abs/1409.0575

- Describes the construction of the ILSVRC datasets
- Highlights the most successful algorithms
- Provides statistical analysis of the results **through ILSVRC2014**
- Compares computer vision accuracy with human-level accuracy

Some questions for today

1. Are computers good at large-scale recognition?

2. Are all objects equally easy for computers?

3. Are we better than computers at recognition?

Some questions for today

1. Are computers good at large-scale recognition?

2. Are all objects equally easy for computers?

3. Are we better than computers at recognition?

ILSVRC image classification task

Steel drum

ILSVRC image classification task

Steel drum

Output:
Scale
T-shirt
<u>Steel drum</u>
Drumstick
Mud turtle

✔

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

ILSVRC image classification task

Steel drum

Output:
Scale
T-shirt
<u>Steel drum</u>
Drumstick
Mud turtle

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

$$\text{Error} = \frac{1}{100{,}000} \sum_{\substack{100{,}000 \\ \text{images}}} 1[\text{incorrect on image } i]$$

ILSVRC2014 classification results

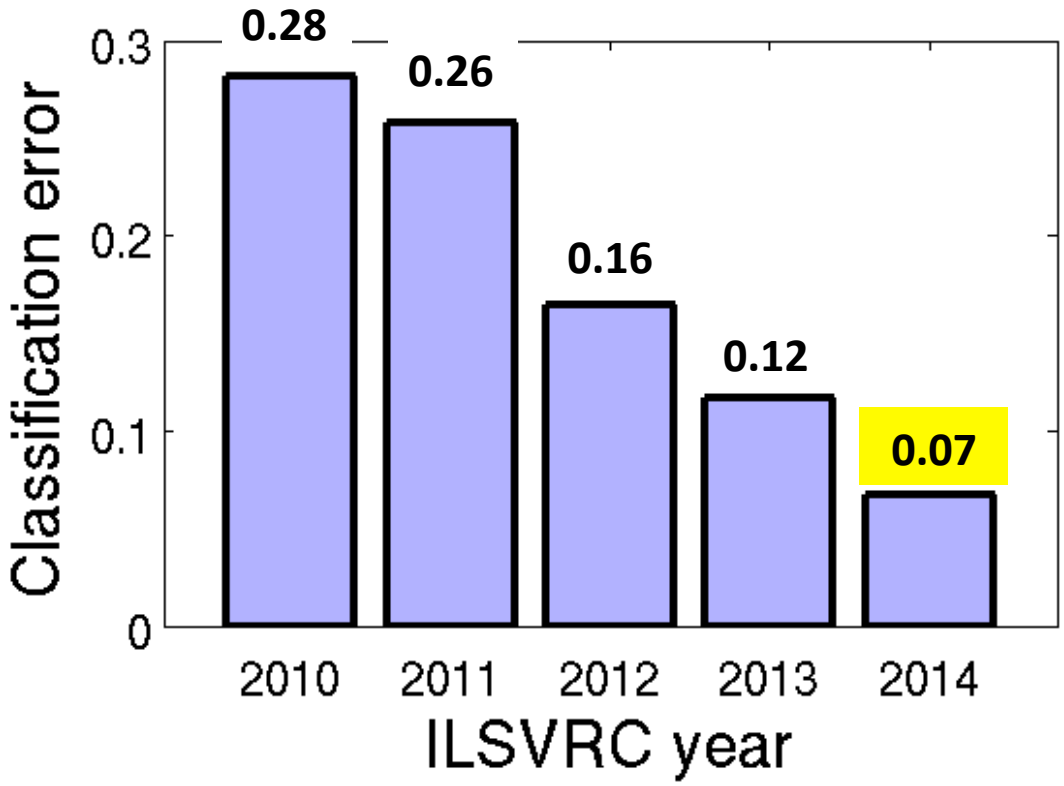
Team Name	Error (%)
GoogLeNet	6.7
VGG	7.3
MSRA Visual computing	8.1
Andrew Howard	8.1
DeeperVision	9.5
NUS-BST	9.8
TTIC_ECP – Epitomic Vision	10.2
XYZ	11.2
Adobe-UIUC, BDC-I2R-UPMC, BREIL_KAIST, Brno University, CASIA_CRIPAC_Weak_Supervision, Cldi-KAIST, DeepCNet, Fengjun Lv, libccv, MIL, Orange-BUPT, PassBy, SCUT_GLH, SYSU_Vision, Trimps-Soushen, UI, UvA-Euvision	

GoogLeNet:
Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Drago Anguelov, Dumitru Erhan, Andrew Rabinovich
Google

VGG:
Karen Simonyan, Andrew Zisserman
University of Oxford

http://image-net.org/challenges/LSVRC/2014/eccv2014

ILSVRC over the years



1.7x reduction in classification error since last year

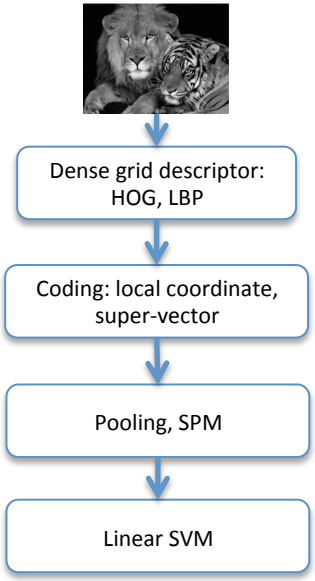
4.2x reduction in classification error since 2010

http://arxiv.org/abs/1409.0575

What changed in ILSVRC classification?

Year 2010

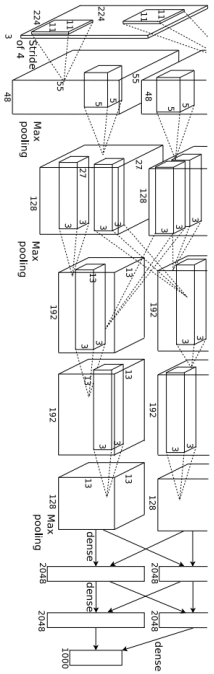
NEC-UIUC



[Lin CVPR 2011]

Year 2012

SuperVision



[Krizhevsky NIPS 2012]

Year 2014

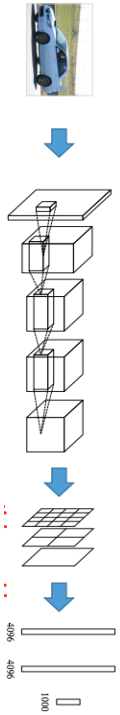
GoogLeNet

[Szegedy arxiv 2014]

VGG

[Simonyan arxiv 2014]

MSRA

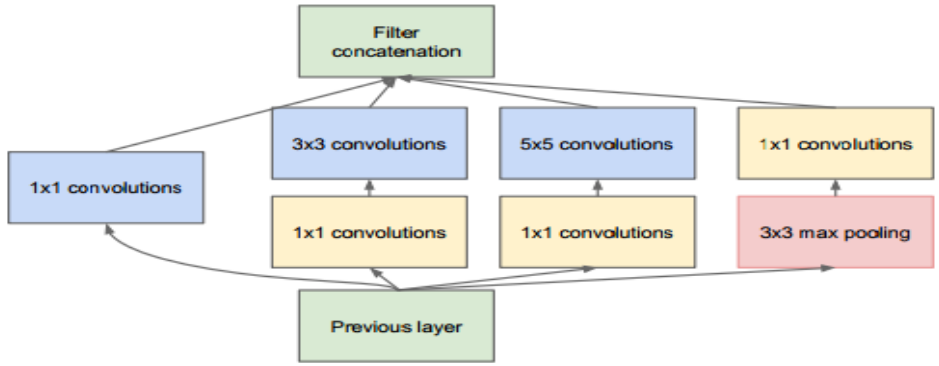


[He arxiv 2014]

What changed in ILSVRC2014 classification?

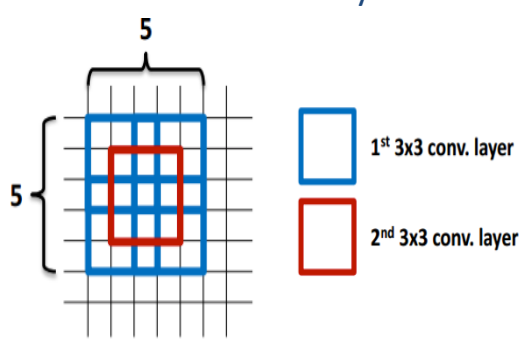
1. Networks became deeper

Inception module [Going deeper with convolutions. Szegedy et al. 2014]



More layers but with smaller kernels (3x3 convolution, 2x2 pooling)

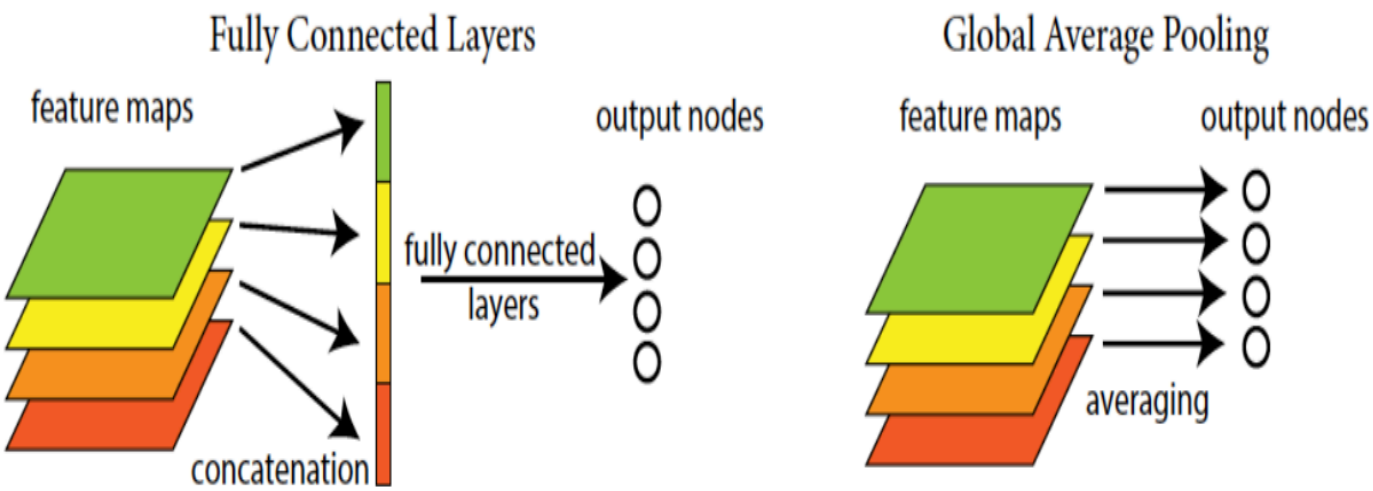
[Very Deep Convolutional Networks. Simonyan and Zisserman. 2014]



http://image-net.org/challenges/LSVRC/2014/eccv2014

What changed in ILSVRC2014 classification?

2. Fully connected layers were (sometimes) removed



Network in Network. Min Lin, Qiang Chen and Shicheng Yan. ICLR 2014
Also used in GoogLeNet and others

http://image-net.org/challenges/LSVRC/2014/eccv2014

3. Almost all successful systems used

- Extensive data augmentation
- Multiscale training across more scales
- Network fusion

And, most importantly, …

http://image-net.org/challenges/LSVRC/2014/eccv2014

3. Almost all successful systems used

- Extensive data augmentation
- Multiscale training across more scales
- Network fusion

And, most importantly, …

Caffe!

http://image-net.org/challenges/LSVRC/2014/eccv2014

Are the winning classification systems really significantly more accurate?

Year	Team name	Error (percent)	99.9% confidence interval
2014	GoogLeNet	6.66	6.40 - 6.92
2014	VGG	7.32	7.05 - 7.60
2014	MSRA	8.06	7.78 - 8.34
2014	AHoward	8.11	7.83 - 8.39
…			
2013	Clarifai	11.20	10.87 - 11.53
…			
2012	SuperVision	15.32	14.94 - 15.69

http://arxiv.org/abs/1409.0575

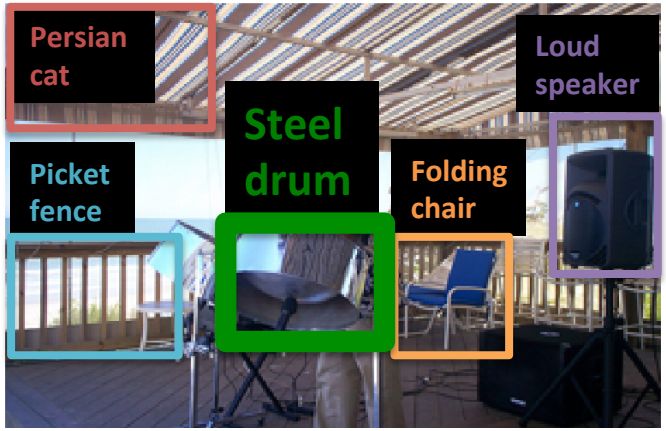
ILSVRC classification + localization task

Steel drum

ILSVRC classification + localization task

Steel drum

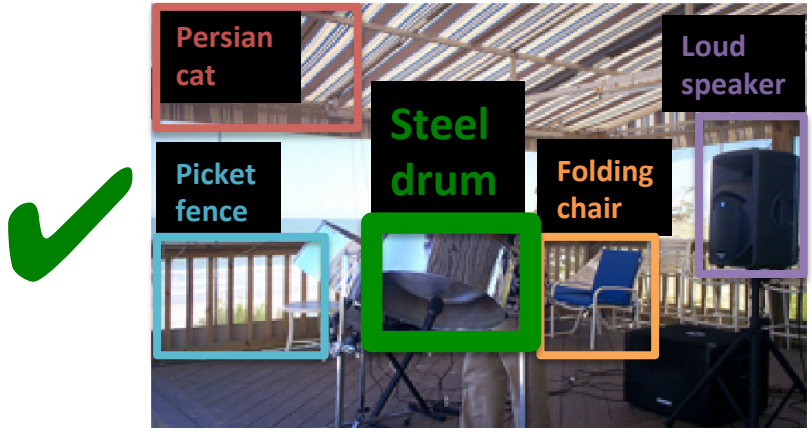
Output



ILSVRC classification + localization task

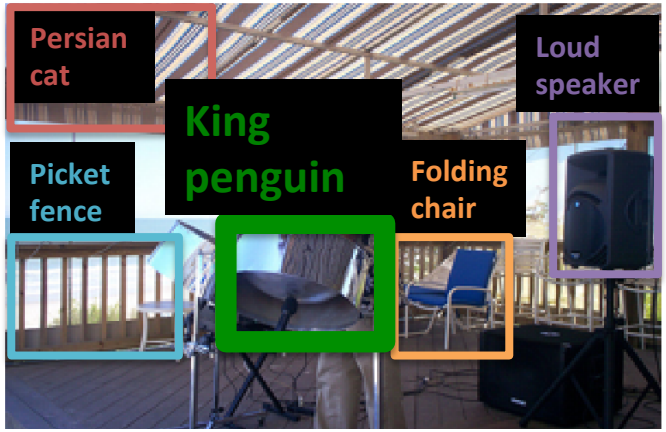
Steel drum

Output



Output (bad localization)

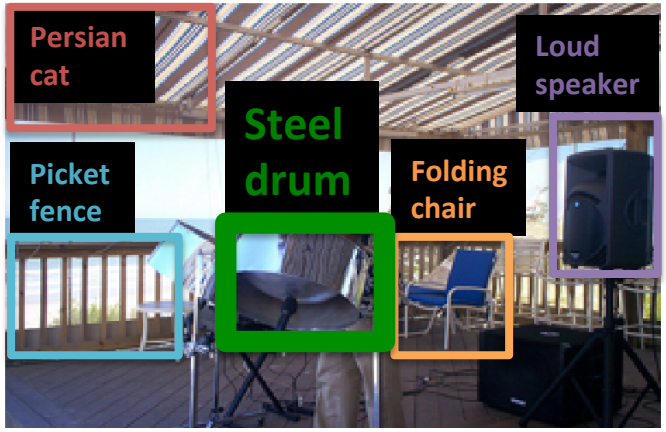
Output (bad classification)



ILSVRC classification + localization task

Steel drum

Output



$$\text{Error} = \frac{1}{100{,}000} \sum_{\substack{100{,}000 \\ \text{images}}} 1[\text{incorrect on image i}]$$

ILSVRC2014 localization results

Team Name	Error (%)
VGG	25.3
GoogLeNet	26.4
Adobe-UIUC	30.1
SYSU_Vision	31.9
MIL	33.7
MSRA Visual computing	35.5
Trimps-Soushen	42.2
ORANGE-BUPT	42.7
Andrew Howard, BDC-I2R-UPMC, BREIL_KAIST, Brno University of Technology, Cldi-KAIST, DeepCNet, DeeperVision, Fengjun Lv, libccv, NUS-BST, PassBy, SCUT_GLH, TTIC_ECP – Epitomic Vision, UI, UvA-Euvision, XYZ	

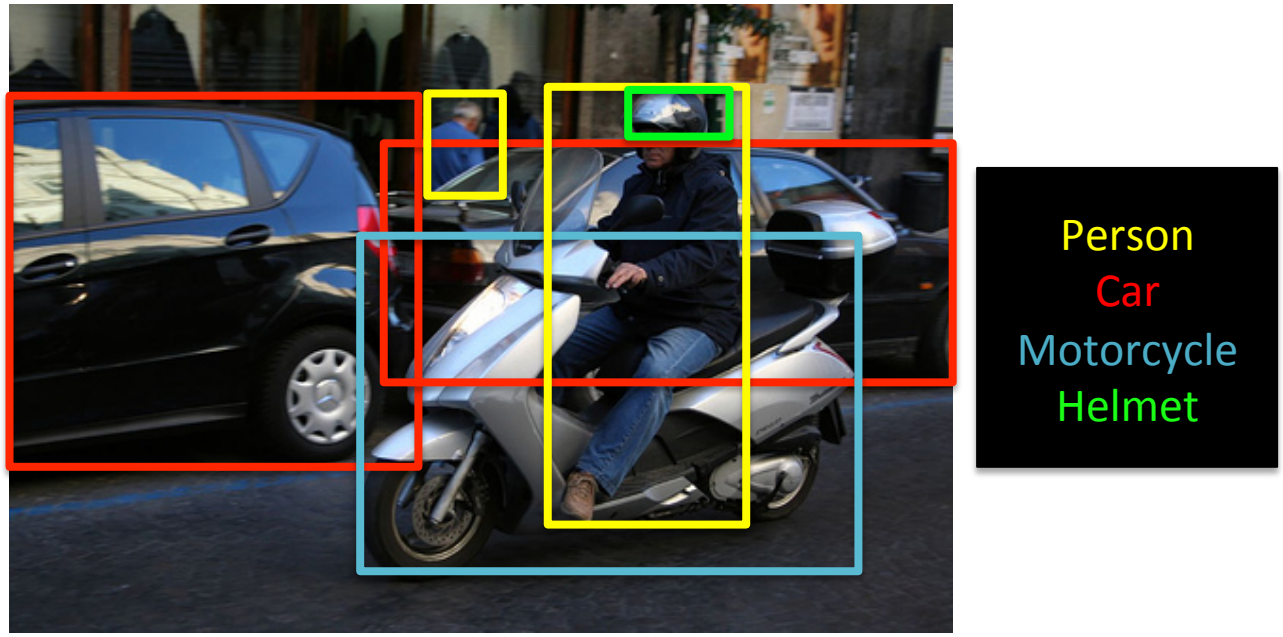
VGG:
Karen Simonyan,
Andrew Zisserman
University of Oxford

GoogLeNet:
Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Drago Anguelov, Dumitru Erhan, Andrew Rabinovich
Google

http://image-net.org/challenges/LSVRC/2014/eccv2014

ILSVRC object detection task

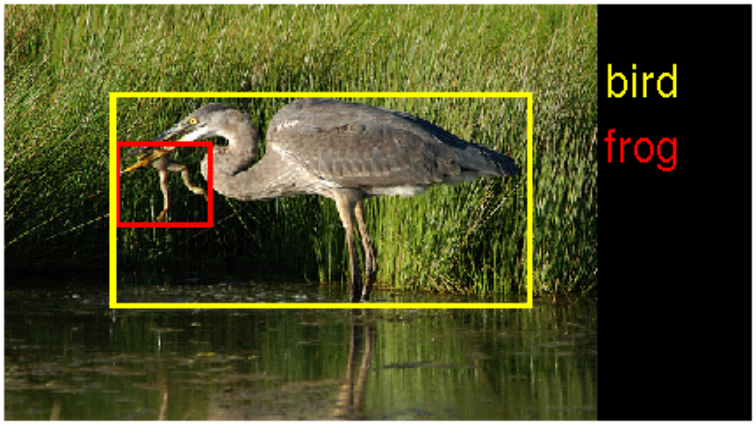
Fully annotated 200 object classes across 120,000 images

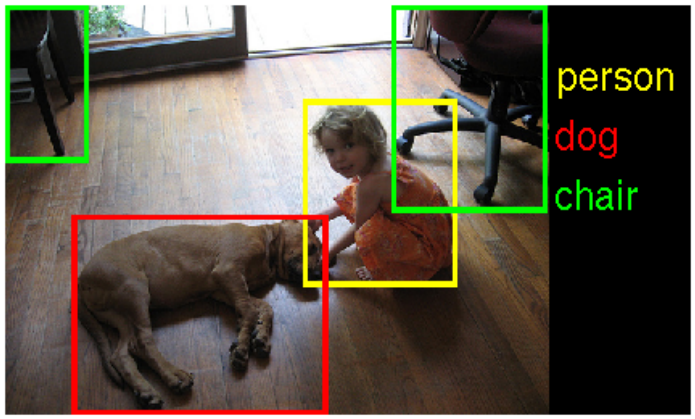


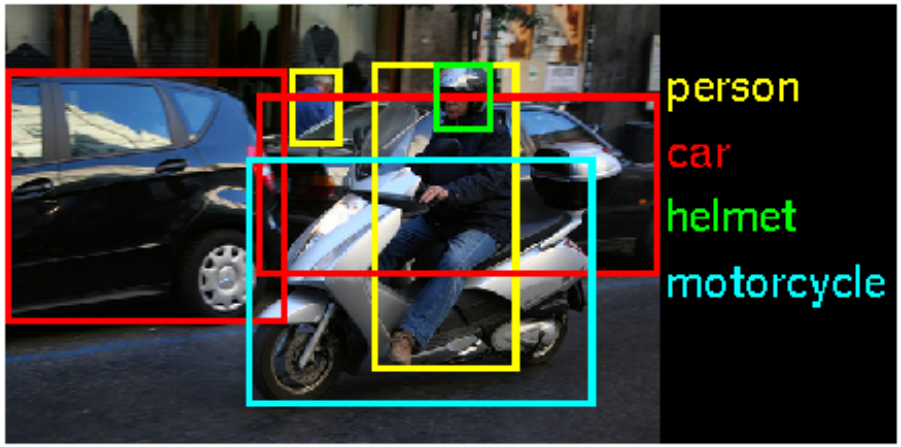
Allows evaluation of generic object detection
in cluttered scenes at scale

Modeled after PASCAL VOC

ILSVRC object detection data

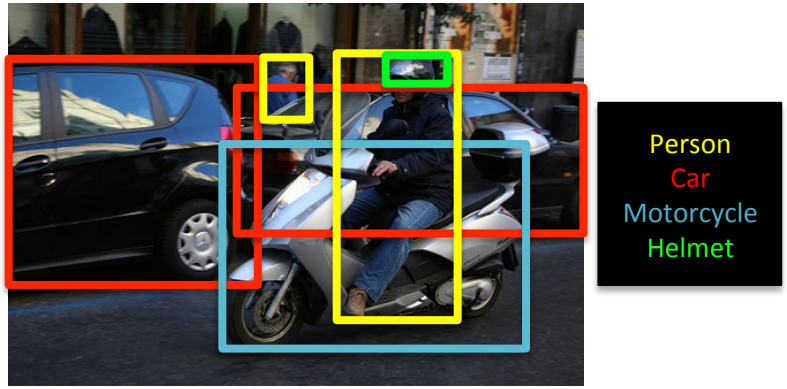






ILSVRC object detection task

All instances of all target object classes expected to be localized on all test images



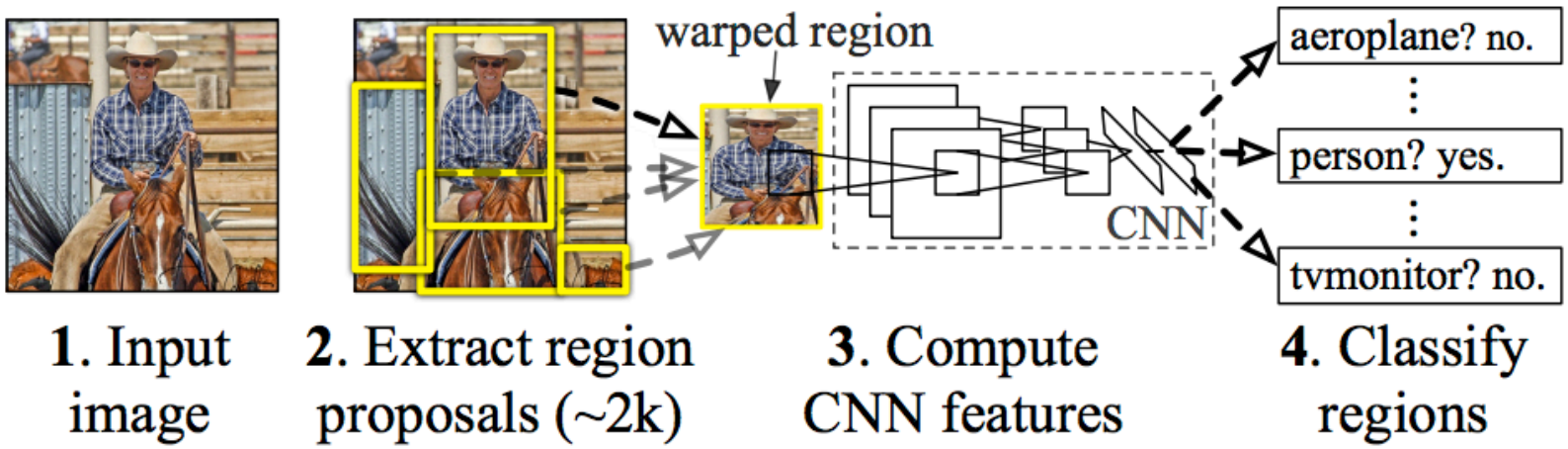
Evaluation modeled after PASCAL VOC:

- Algorithm outputs a list of bounding box detections with confidences
- A detection is considered correct if IOU with ground truth > threshold
- Evaluated by average precision per object class
- Winners of challenge is the team that wins the most object categories

Everingham, Van Gool, Williams, Winn and Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. IJCV 2010.

ILSVRC2014 object detection approach #1

R-CNN: Regions with CNN features

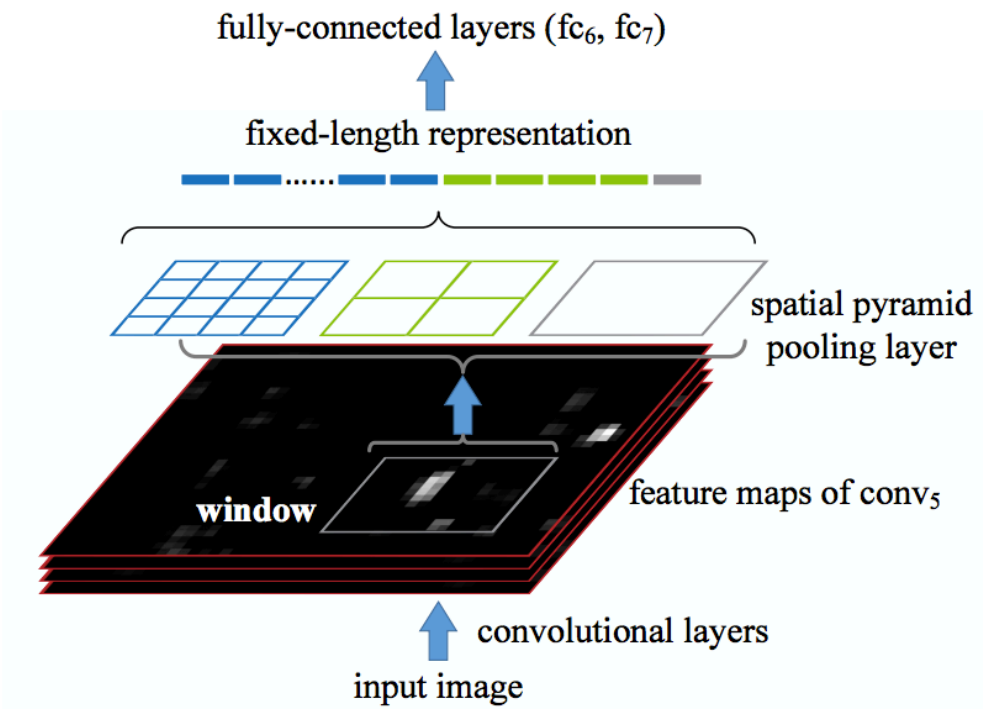


Rich feature hierarchies for accurate object detection
Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik
http://arxiv.org/abs/1311.2524

http://image-net.org/challenges/LSVRC/2014/eccv2014

ILSVRC2014 object detection approach #2

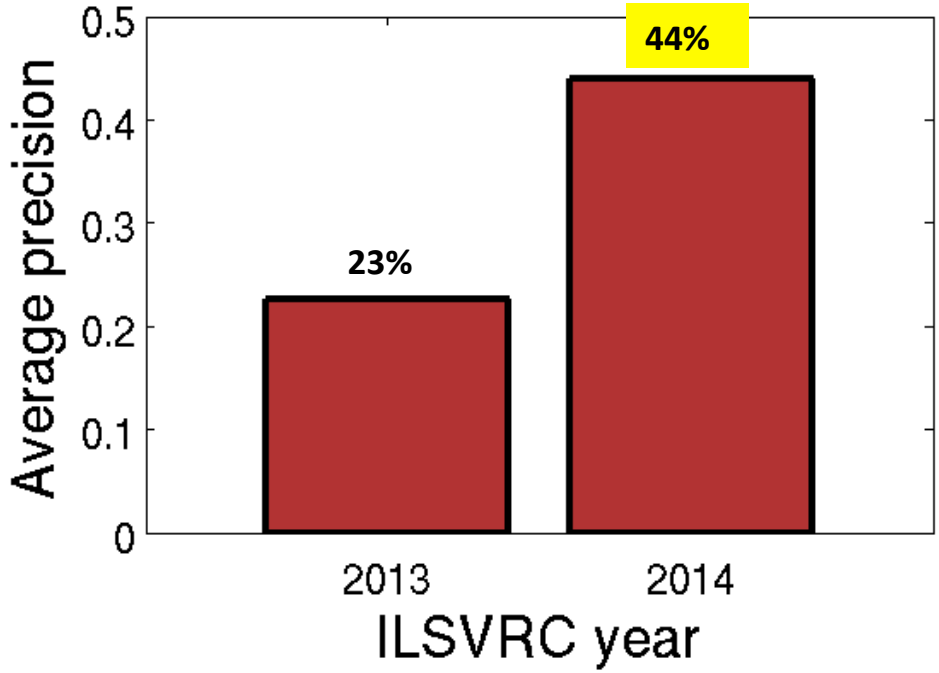
SPP-net: Spatial Pyramid Pooling



Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition
Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun
http://arxiv.org/abs/1406.4729

http://image-net.org/challenges/LSVRC/2014/eccv2014

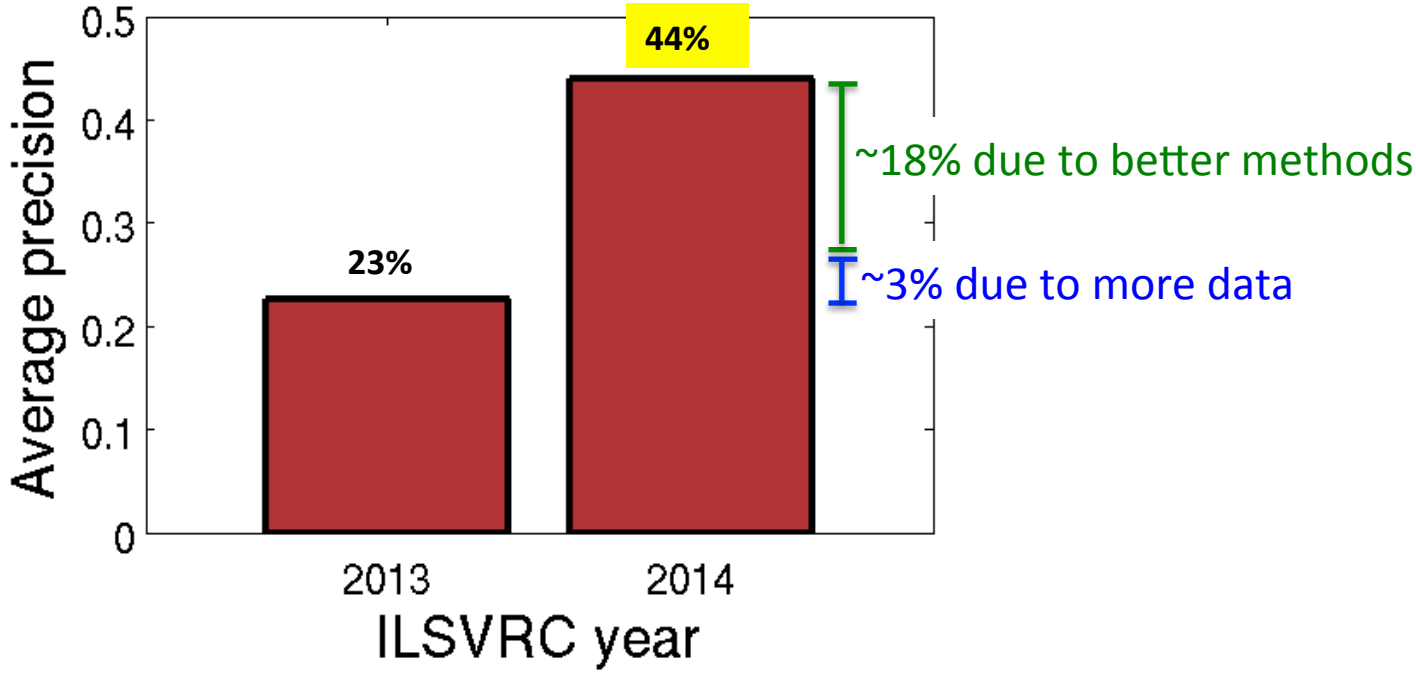
ILSVRC detection since 2013



1.9x increase in object detection average precision in one year

http://arxiv.org/abs/1409.0575

ILSVRC detection since 2013



1.9x increase in object detection average precision in one year

http://arxiv.org/abs/1409.0575

Some questions for today

1. Are computers good at large-scale recognition?

2. Are all objects equally easy for computers?

3. Are we better than computers at recognition?

Some questions for today

1. Are computers good at large-scale recognition?

Yes! *CNNs are getting deeper, accuracy is getting better.*

2. Are all objects equally easy for computers?

3. Are we better than computers at recognition?

Some questions for today

1. Are computers good at large-scale recognition?

   Yes! *CNNs are getting deeper, accuracy is getting better.*

2. Are all objects equally easy for computers?

3. Are we better than computers at recognition?

Easiest classification categories

red fox (100)

hen-of-the-woods (100)

ibex (100)

goldfinch (100)

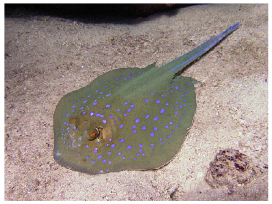
flat-coated retriever (100)

tiger (100)

hamster (100)

porcupine (100)

stingray (100)



Blenheim spaniel (100)

… and 111 more categories with 100% accuracy!

(Highest *accuracy* in percent achieved by any method in ILSVRC2012-ILSVRC2014)

http://arxiv.org/abs/1409.0575

Hardest classification categories

(Highest *accuracy* in percent achieved by any method in ILSVRC2012-ILSVRC2014)

http://arxiv.org/abs/1409.0575

Easiest localization categories

Leonberg (100)

ruffed grouse (100)

ruddy turnstone (100)

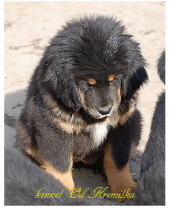
giant schnauzer (99)

tiger (99)

Maltese dog (99)

Japanese spaniel (99)

Tibetan mastiff (99)



hare (99)

African hunting dog (99)

(Highest *accuracy* in percent achieved by any method in ILSVRC2012-ILSVRC2014)

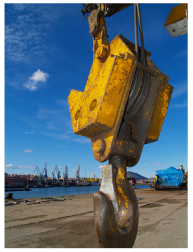
http://arxiv.org/abs/1409.0575

Hardest localization categories

horizontal bar (41)

flagpole (38)

hook (37)



lakeside (36)

letter opener (36)

spotlight (35)

wing (35)

ladle (28)

pole (27)

space bar (23)

(Highest *accuracy* in percent achieved by any method in ILSVRC2012-ILSVRC2014)

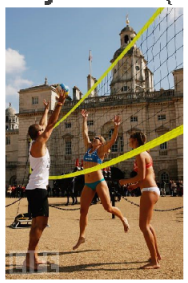
http://arxiv.org/abs/1409.0575

Easiest object detection categories

butterfly (93)

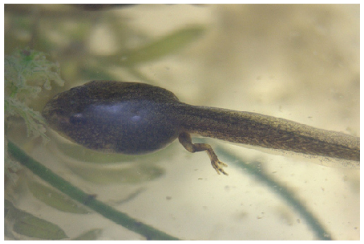
dog (84)

volleyball (83)



rabbit (83)

frog (82)

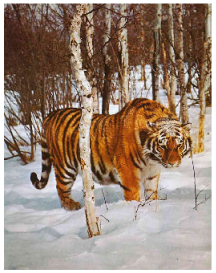


basketball (80)

snowplow (80)

bird (78)

tiger (77)



zebra (77)

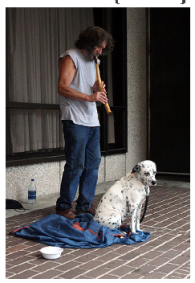
(Highest average precision in percent achieved by any method in ILSVRC2013 and ILSVRC2014)

http://arxiv.org/abs/1409.0575

Hardest object detection categories

lamp (15)

flute (15)



horizontal bar (14)

spatula (13)

nail (13)

ski (12)

microphone (11)

rubber eraser (10)

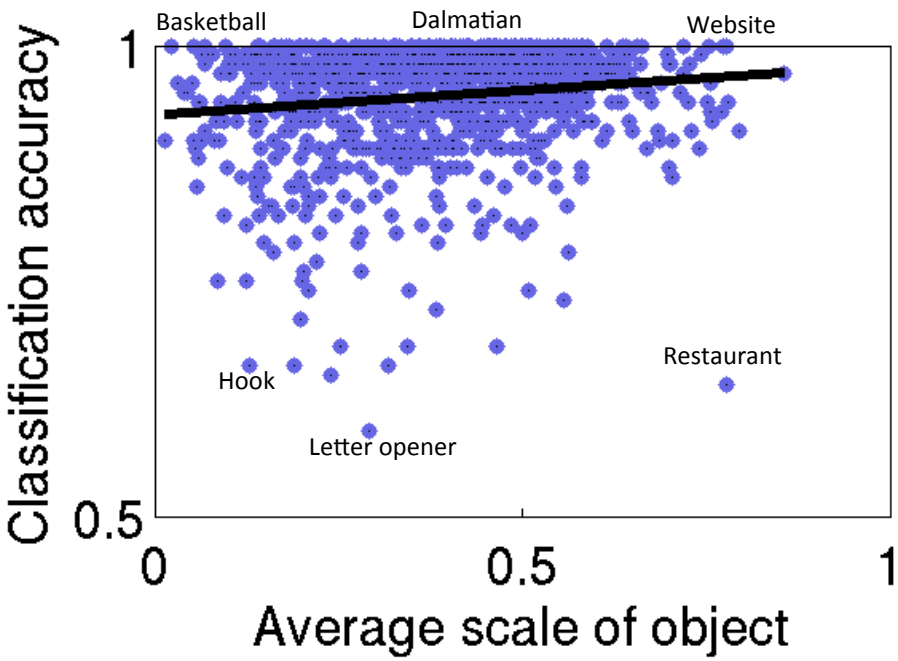
ladle (9)

backpack (8)

(Highest average precision in percent achieved by any method in ILSVRC2013 and ILSVRC2014)

http://arxiv.org/abs/1409.0575

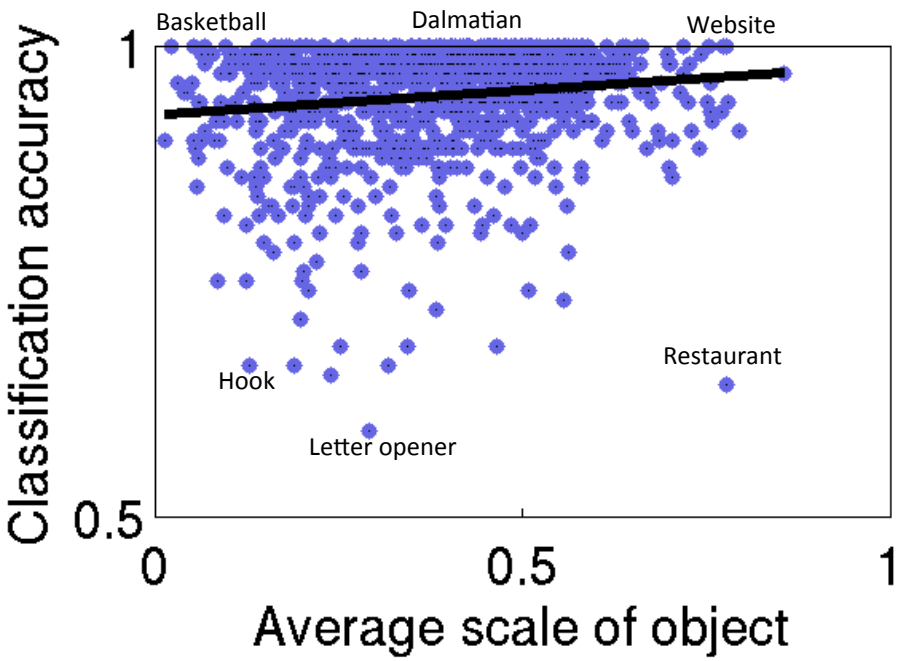
Smaller objects not necessarily harder

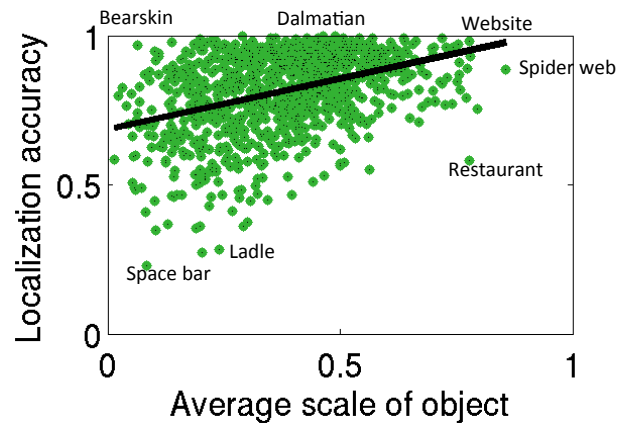


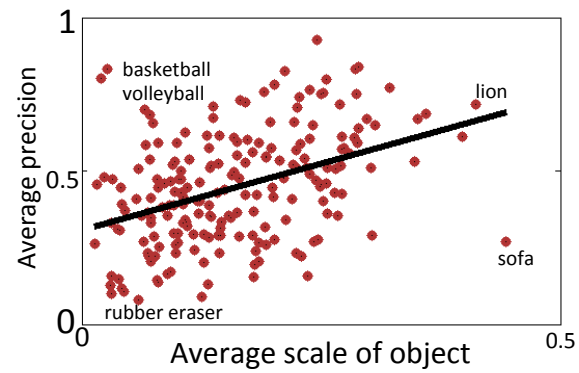
- Each dot is an object class
- X-axis: average fraction of image area occupied by an instance of that class on the validation set
- Y-axis: highest accuracy achieved by any method in ILSVRC2012-ILSVRC2014

http://arxiv.org/abs/1409.0575

Smaller objects not necessarily harder



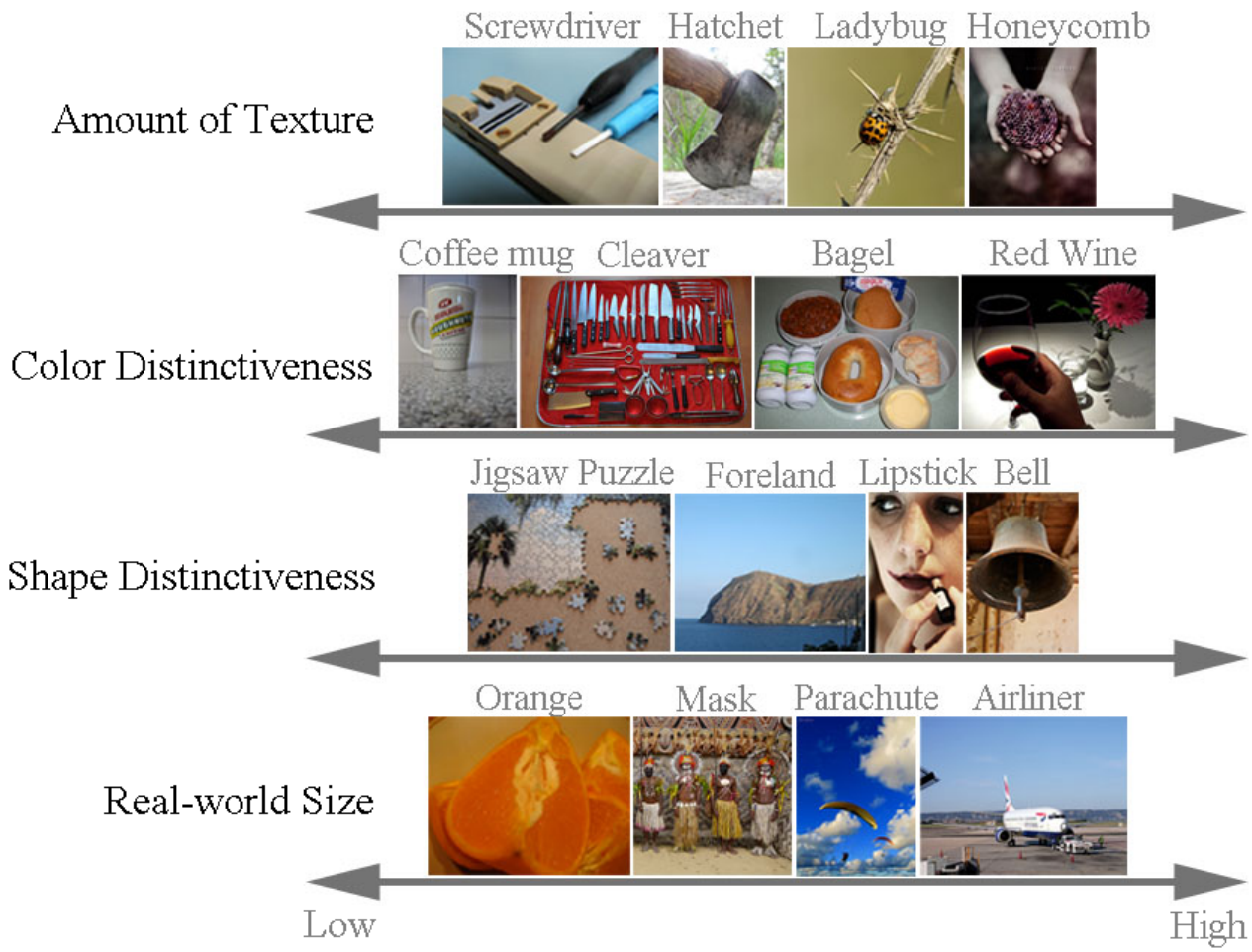




- Each dot is an object class
- X-axis: average fraction of image area occupied by an instance of that class on the validation set
- Y-axis: highest accuracy achieved by any method in ILSVRC2012-ILSVRC2014

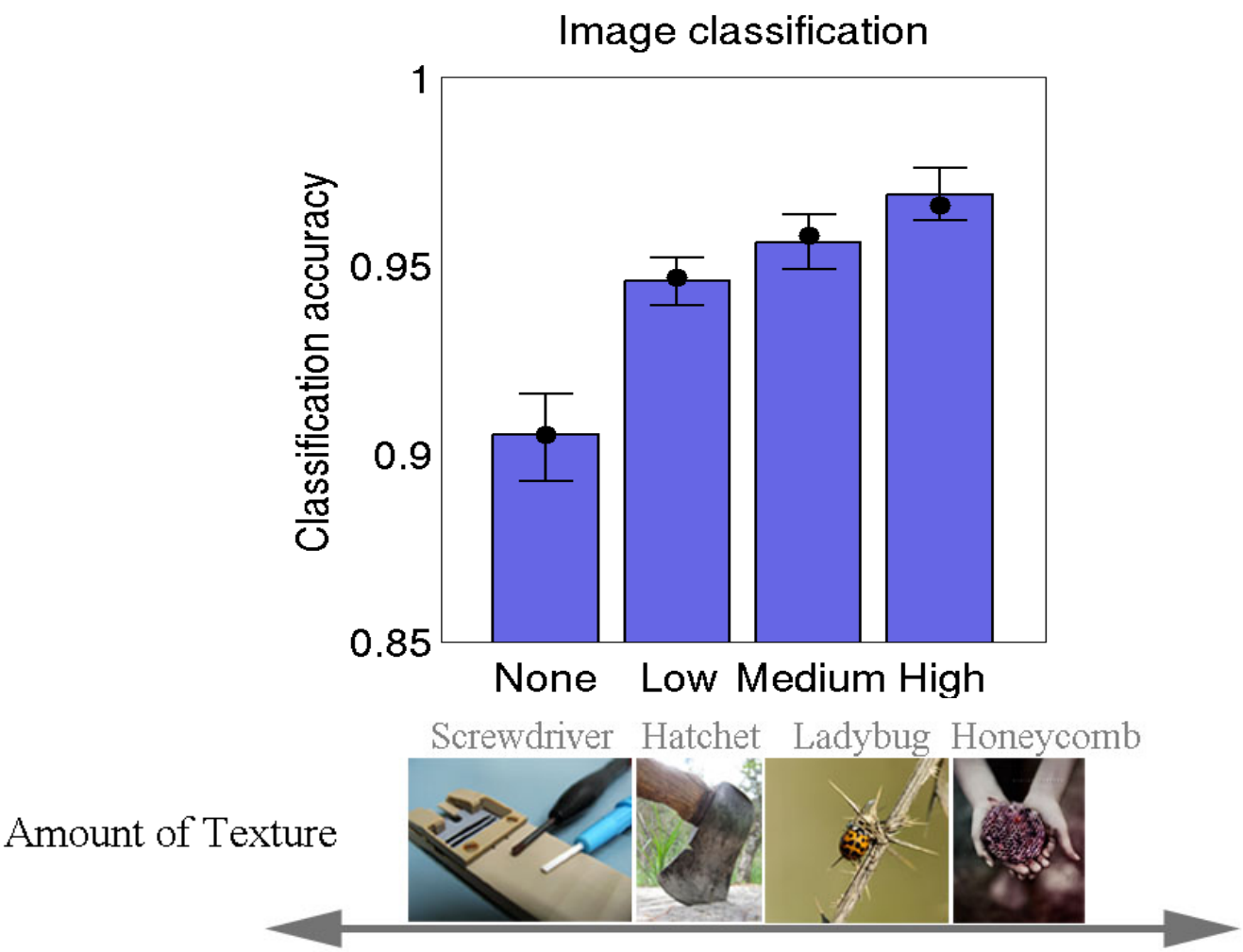
http://arxiv.org/abs/1409.0575

Manually annotated object class properties



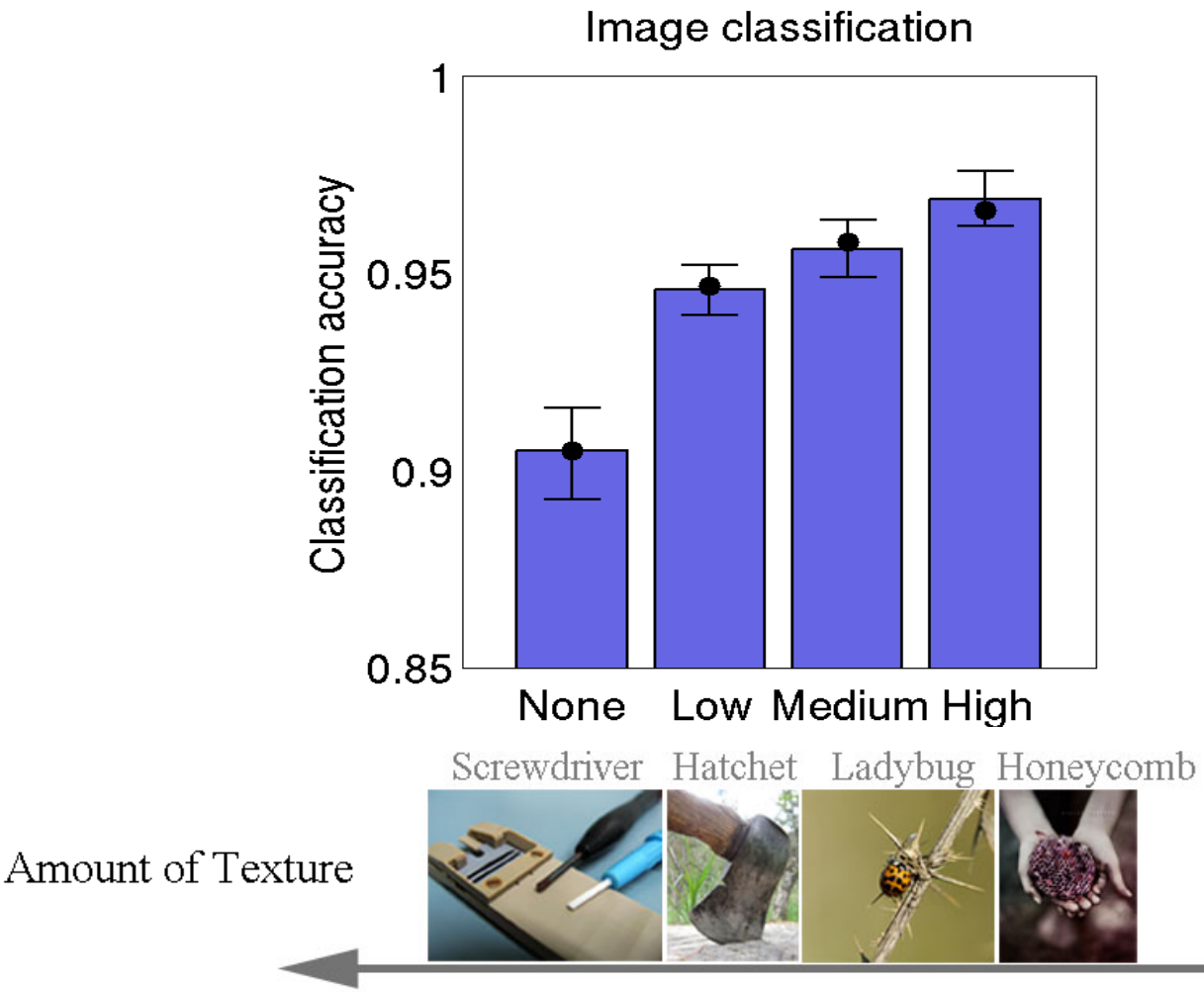
Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going? ICCV 2013

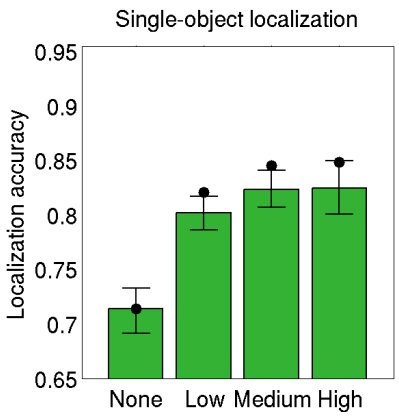
Textured objects are easier

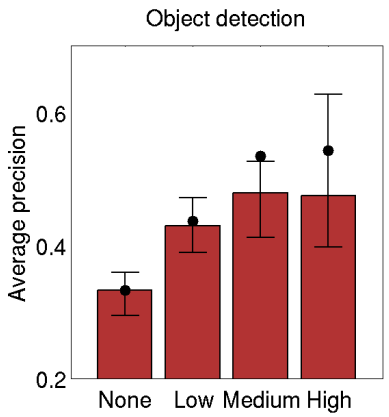


http://arxiv.org/abs/1409.0575

Textured objects are easier







http://arxiv.org/abs/1409.0575

Some questions for today

1. Are computers good at large-scale recognition?

   Yes! *CNNs are getting deeper, accuracy is getting better.*

2. Are all objects equally easy for computers?

3. Are we better than computers at recognition?

Some questions for today

1. Are computers good at large-scale recognition?

   Yes! *CNNs are getting deeper, accuracy is getting better.*

2. Are all objects equally easy for computers?

   No. *Thin and untextured objects are still hard for computers.*

3. Are we better than computers at recognition?

Some questions for today

1. Are computers good at large-scale recognition?

   Yes! *CNNs are getting deeper, accuracy is getting better.*

2. Are all objects equally easy for computers?

   No. *Thin and untextured objects are still hard for computers.*

3. Are we better than computers at recognition?

What is human accuracy on ILSVRC2014 classification?

Andrej Karpathy

But data is manually annotated, isn't human accuracy 100%?

Current crowdsourcing annotation interface

Is this a badger? Yes or No

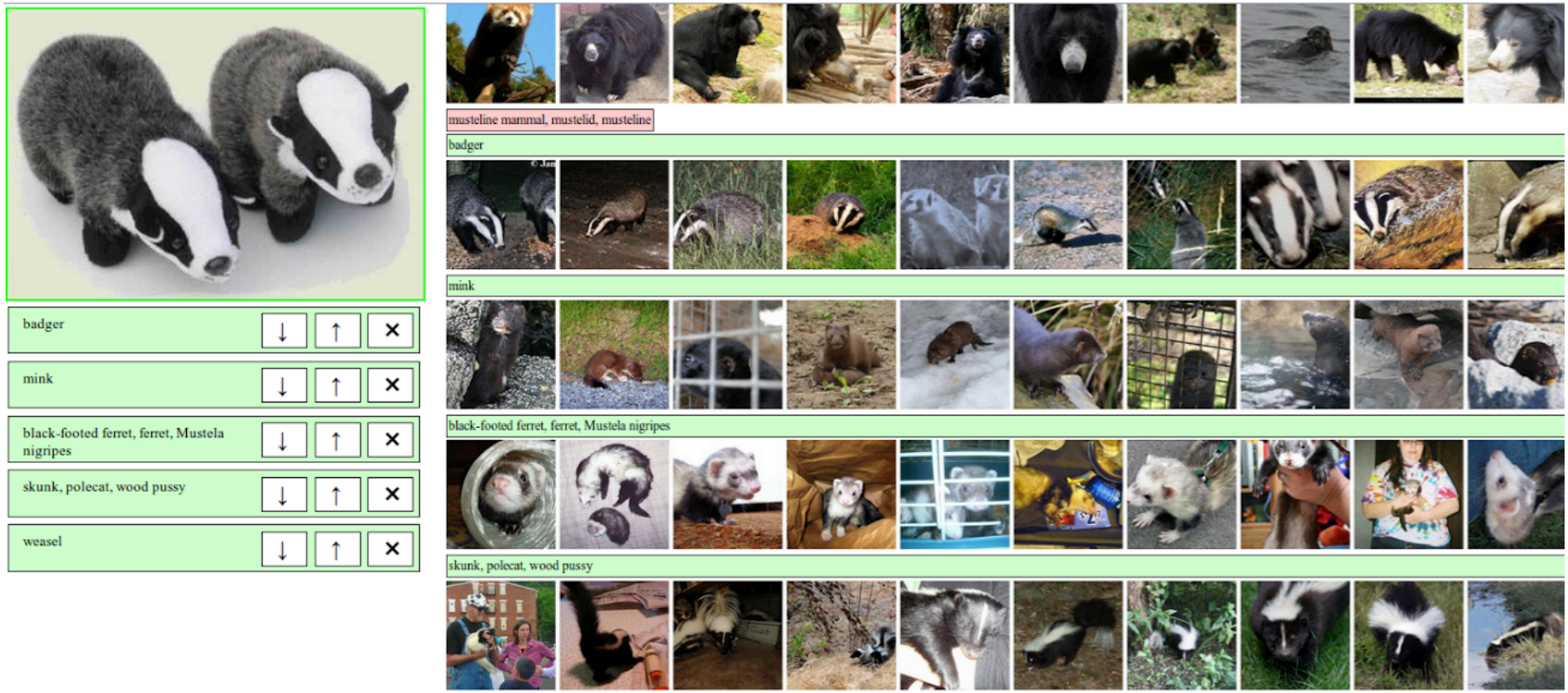
Very different from

Which one of 1000 classes is this?

http://arxiv.org/abs/1409.0575

New web-based annotation interface with 1000 object classes

http://bit.ly/ilsvrclabel



http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

- Compared expert human annotators with winning GoogLeNet entry

	Annotator 1
Total number of images	1500
GoogLeNet classification error	6.8%
Human classification error	5.1%

- Annotator 1 achieved better accuracy than GoogLeNet by 1.7% (p = 0.022)
- Task required *significant* amount of training for humans

http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

- Compared expert human annotators with winning GoogLeNet entry

	Annotator 1	Annotator 2
Total number of images	1500	258
GoogLeNet classification error	6.8%	5.8%
Human classification error	5.1%	12.0%

- Annotator 1 achieved better accuracy than GoogLeNet by 1.7% (p = 0.022)
- Task required *significant* amount of training for humans

http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

	GoogLeNet correct	GoogLeNet wrong
Human correct	1352/1500	72/1500
Human wrong	46/1500	30/1500

http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

	GoogLeNet correct	GoogLeNet wrong
Human correct	1352/1500	72/1500 • Objects very small or thin • Abstract representations • Image filters
Human wrong	46/1500	30/1500

http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

	GoogLeNet correct	GoogLeNet wrong
Human correct	1352/1500	72/1500
Human wrong	46/1500	3

- Objects very small or thin
- Abstract representations
- Image filters

reel	hatchet	sidewinder
stethoscope	vase	maze, labyrinth
whistle	pitcher, ewer	gar, garfish
ice lolly, lolly	coffeepot	valley, vale
hair spray	mask	hammerhead
maypole	cup	sea snake

http://arxiv.org/abs

Human vs computer accuracy on ILSVRC2014 classification

	GoogLeNet correct	GoogLeNet wrong
Human correct	1352/1500	72/1500 • Objects very small or thin • Abstract representations • Image filters
Human wrong	46/1500 • Fine-grained recognition • Class unawareness • Insufficient training data	30/1500

http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

	GoogLeNet correct	GoogLeNet wrong
Human correct	1352/1500	72/1500 • Objects very small or thin • Abstract representations Image filters
Human wrong	46/1500 • Fine-grained recognition • Class unawareness • Insufficient training data	

http://arxiv.org/abs/1409.0575

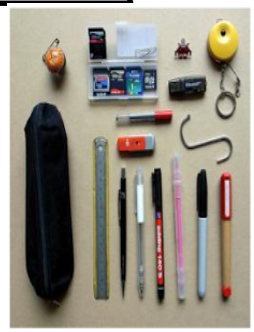
Human vs computer accuracy on ILSVRC2014 classification

	GoogLeNet correct	GoogLeNet wrong
Human correct	1352/1500	72/1500 • Objects very small or thin • Abstract representations • Image filters
Human wrong	46/1500 • Fine-grained recognition • Class unawareness • Insufficient training data	30/1500 • Multiple objects • Incorrect annotations

http://arxiv.org/abs/1409.0575

Human vs computer accuracy on ILSVRC2014 classification

	GoogLeNet correct	GoogLeNet wrong
Human correct	1352/1500	72/1500 • Objects very small or thin • Abstract representations • Image filters
Human wrong	46/1500 • Fine-grained recognition • Class unawareness • Insufficient training data	30/1500 • Multiple objects • Incorrect annotations



rule, ruler
pencil box, pencil case
rubber eraser, rubber
ballpoint, ballpoint pen
pencil sharpener
carpenter's kit, tool kit

http://arxiv.org/abs/1409.0575

Some questions for today

1. Are computers good at large-scale recognition?

   Yes! *CNNs are getting deeper, accuracy is getting better.*

2. Are all objects equally easy for computers?

   No. *Thin and untextured objects are still hard for computers.*

3. Are we better than computers at recognition?

Some questions for today

1. Are computers good at large-scale recognition?

   Yes! *CNNs are getting deeper, accuracy is getting better.*

2. Are all objects equally easy for computers?

   No. *Thin and untextured objects are still hard for computers.*

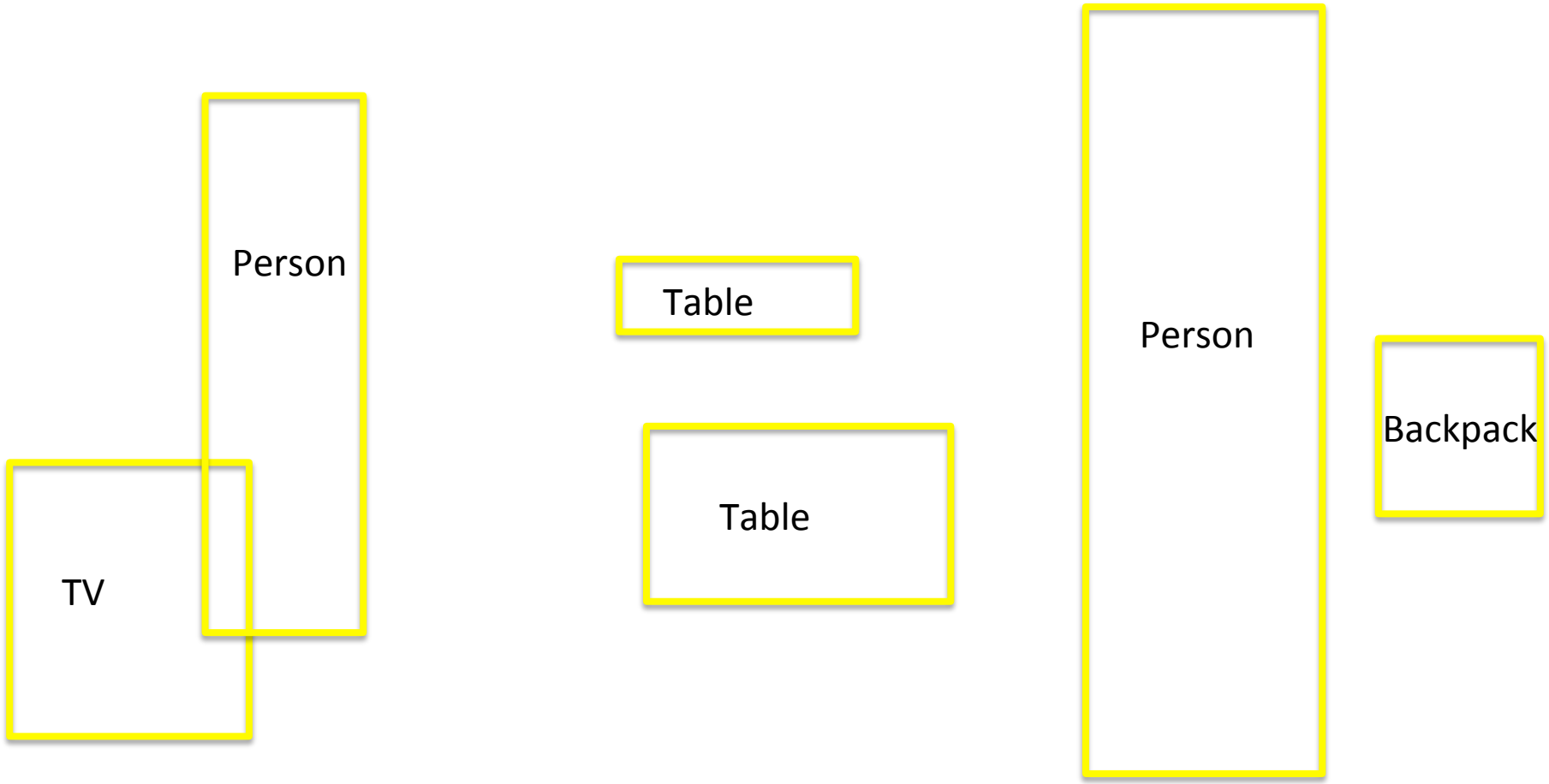
3. Are we better than computers at recognition?

   Not always… *We are worse than computers at large-scale fine-grained classification.*

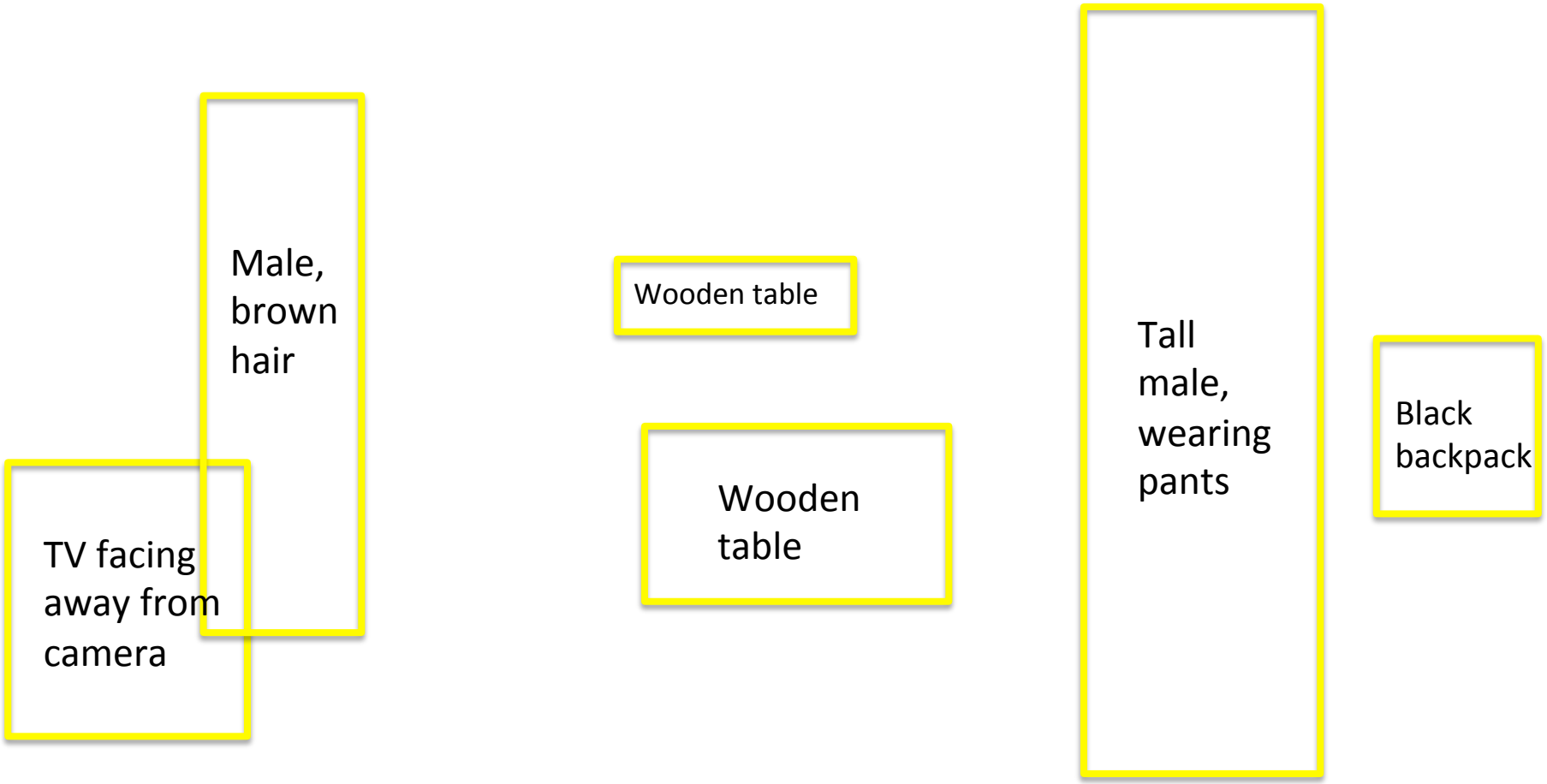
So have we solved computer vision?

So have we solved computer vision?

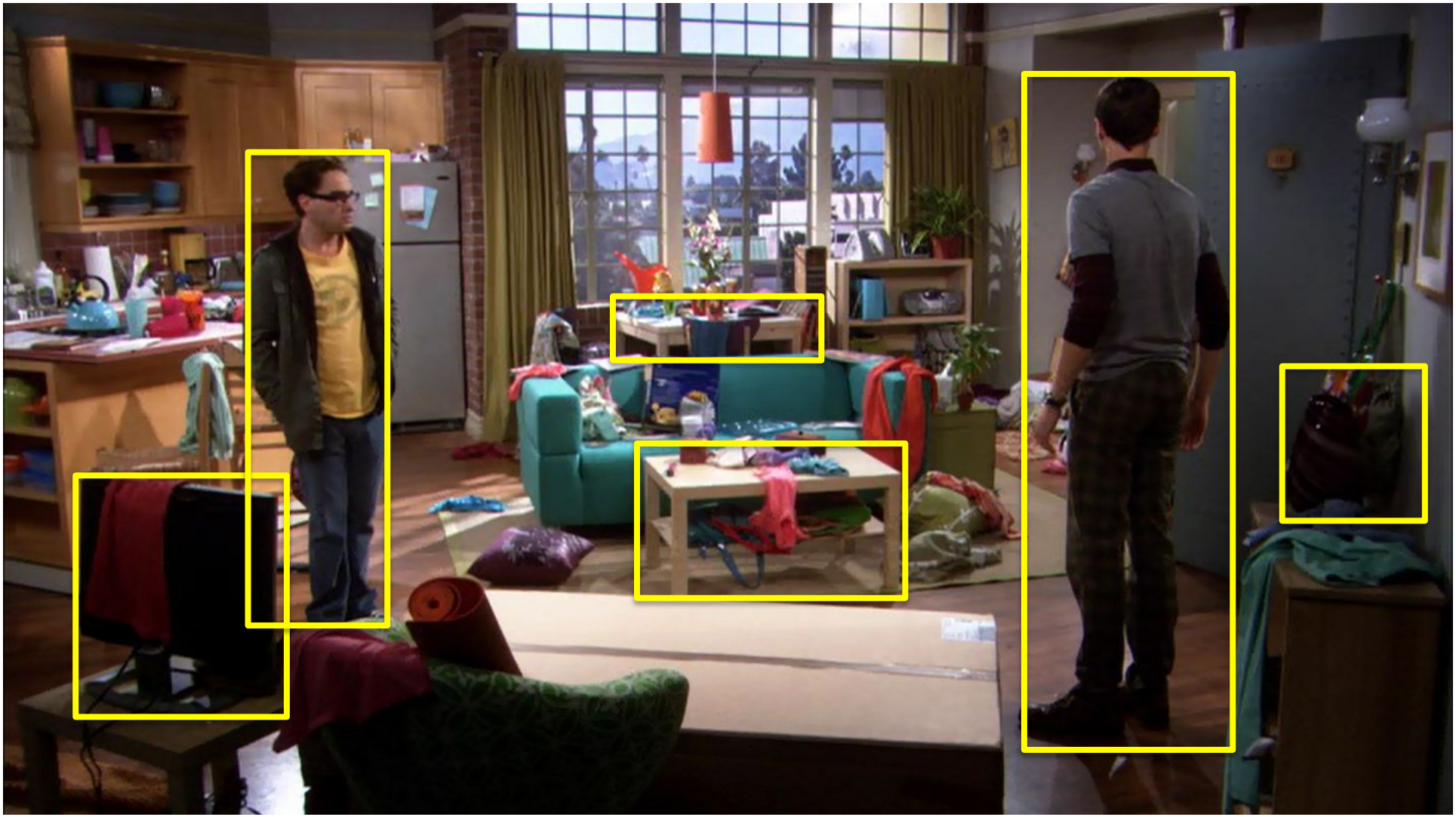
RCNN output:



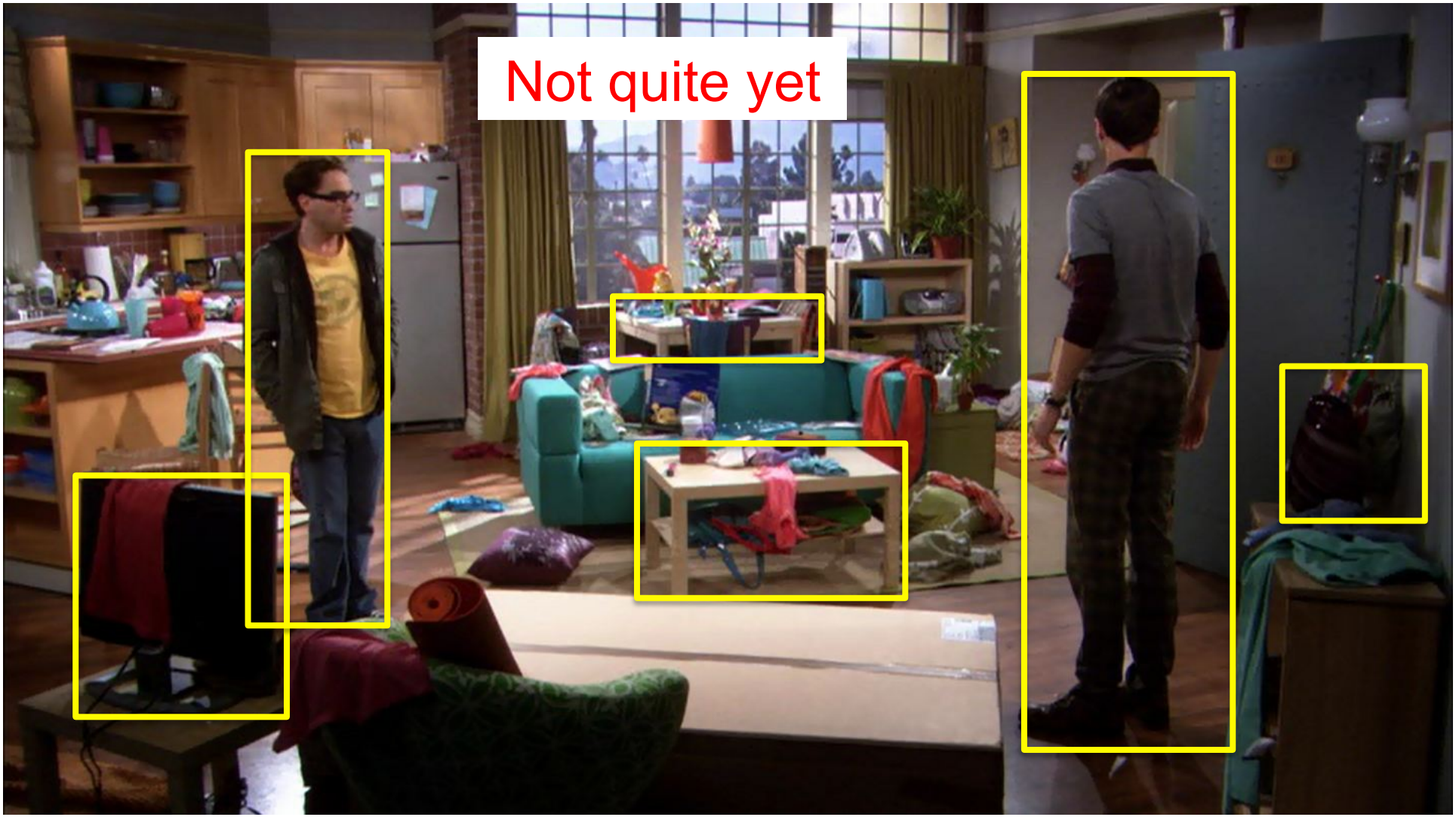
So have we solved computer vision?



So have we solved computer vision?



So have we solved computer vision?



So have we solved computer vision?