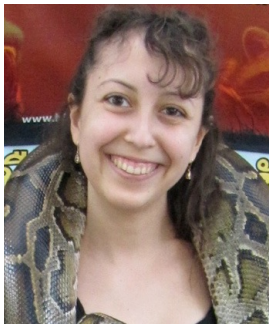
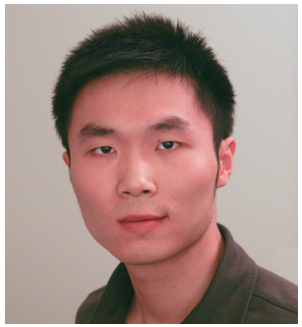
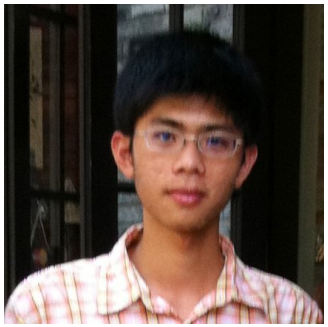
Analysis of Large Scale Visual Recognition

Fei-Fei Li and Olga Russakovsky


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**

Backpack

Flute

Strawberry

Traffic light

Backpack

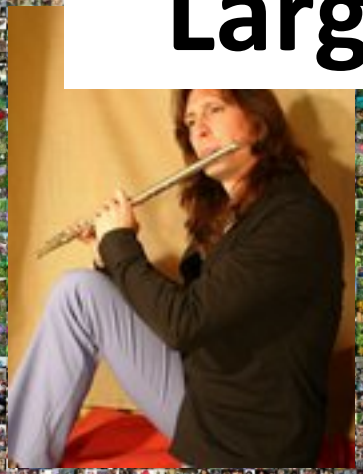
Matchstick

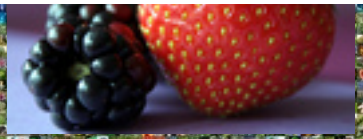
Bathing cap

Sea lion

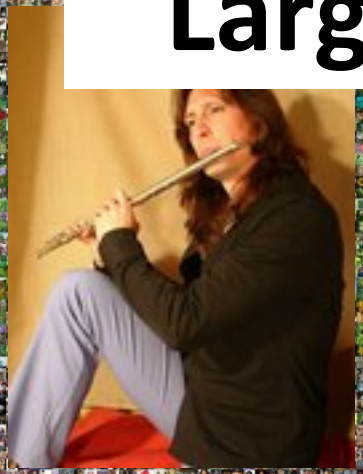
Racket

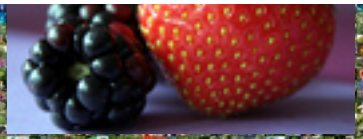
Large-scale recognition





Large-scale recognition



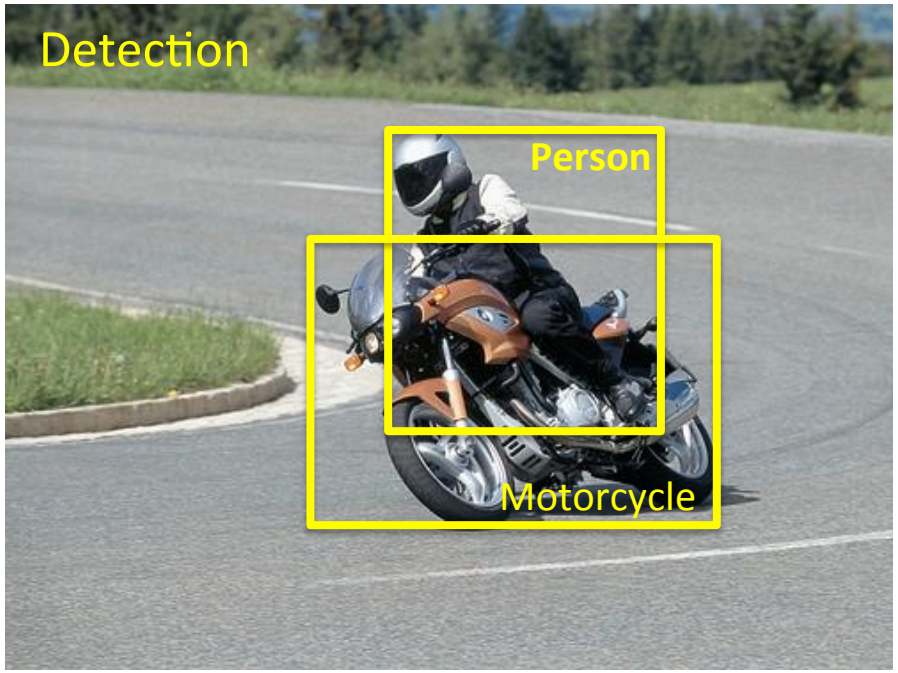


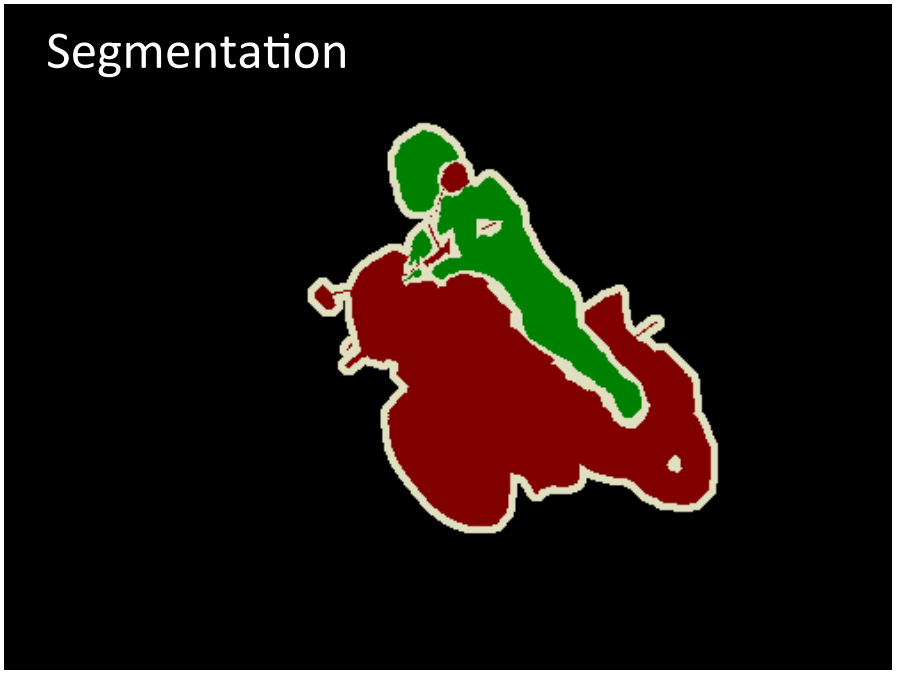
Need benchmark datasets

PASCAL VOC 2005-2012

20 object classes **22,591 images**

Classification: person, motorcycle





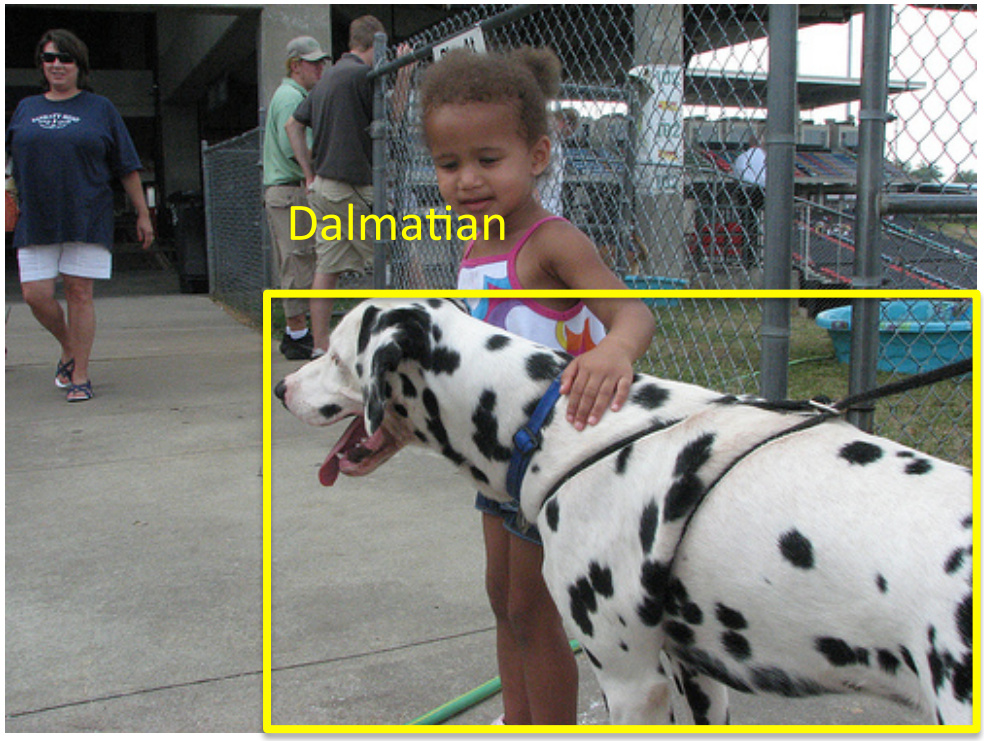
Action: riding bicycle

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

IMAGENET Large Scale Visual Recognition Challenge (ILSVRC) 2010-2012

~~20 object classes~~ ~~22,591 images~~

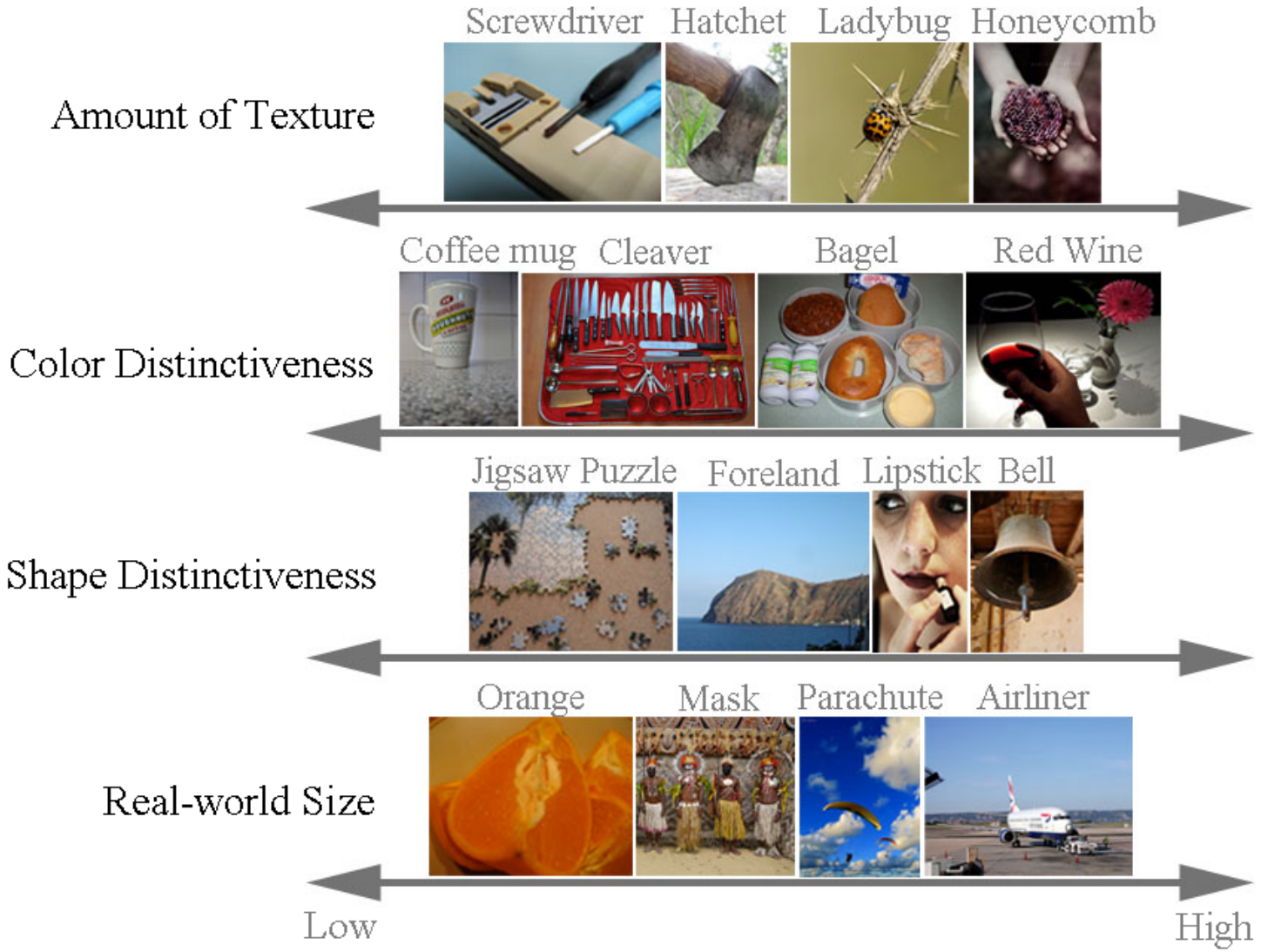
1000 object classes **1,431,167 images**



http://image-net.org/challenges/LSVRC/{2010,2011,2012}

Variety of object classes in ILSVRC

Variety of object classes in ILSVRC



ILSVRC Task 1: Classification

Steel drum

ILSVRC Task 1: Classification

Steel drum

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

ILSVRC Task 1: Classification

Steel drum

Output:
Scale
T-shirt
Steel drum
Drumstick
Mud turtle

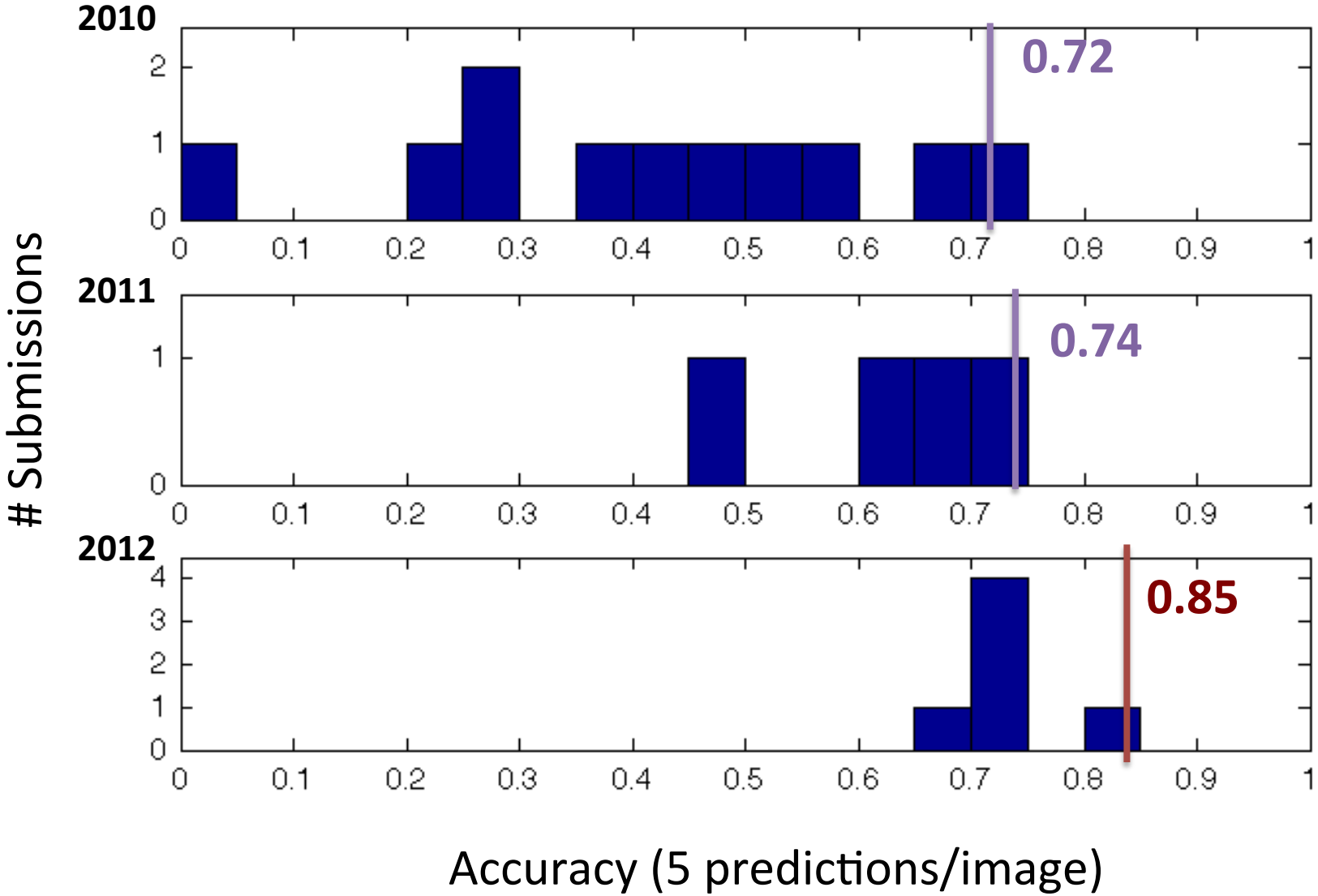
✔

Output:
Scale
T-shirt
Giant panda
Drumstick
Mud turtle

✗

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

ILSVRC Task 1: Classification



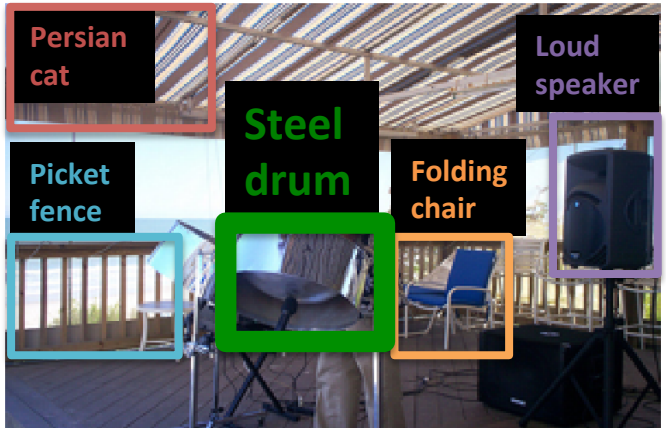
ILSVRC Task 2: Classification + Localization

Steel drum

ILSVRC Task 2: Classification + Localization

Steel drum

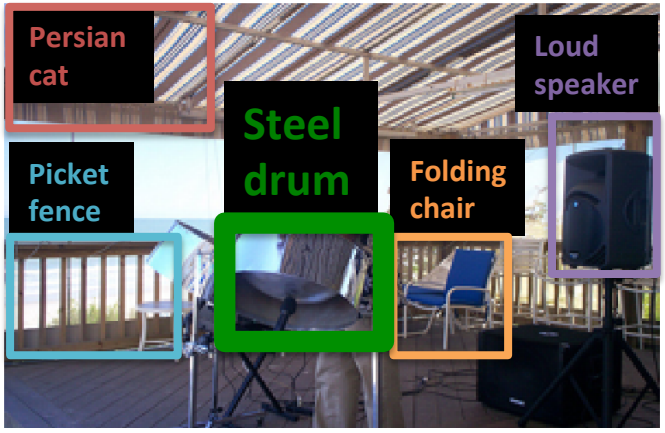
Output



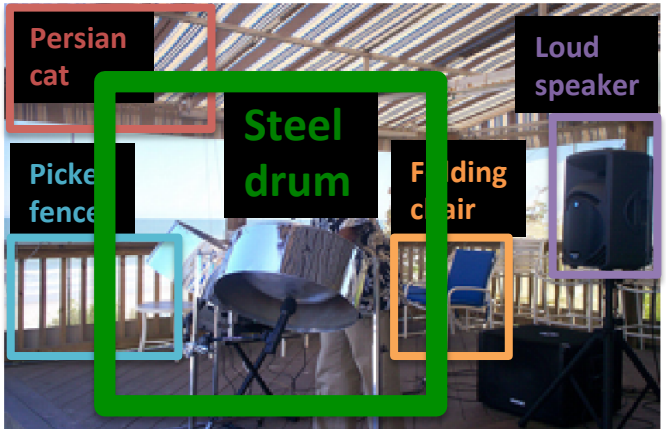
ILSVRC Task 2: Classification + Localization

Steel drum

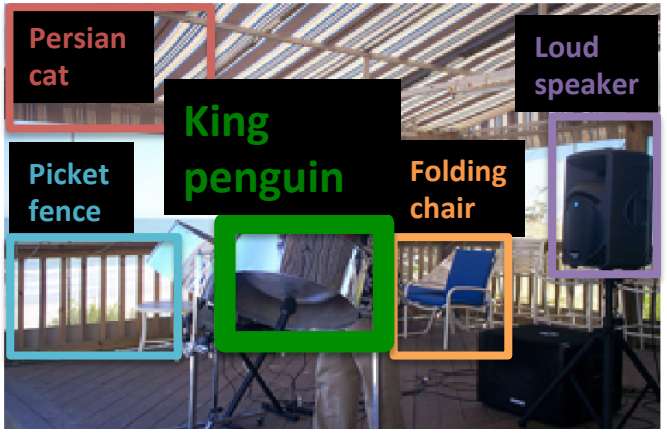
Output



Output (bad localization)



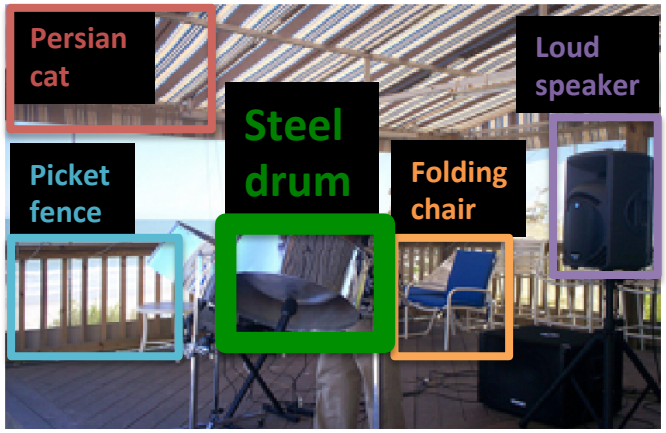
Output (bad classification)



ILSVRC Task 2: Classification + Localization

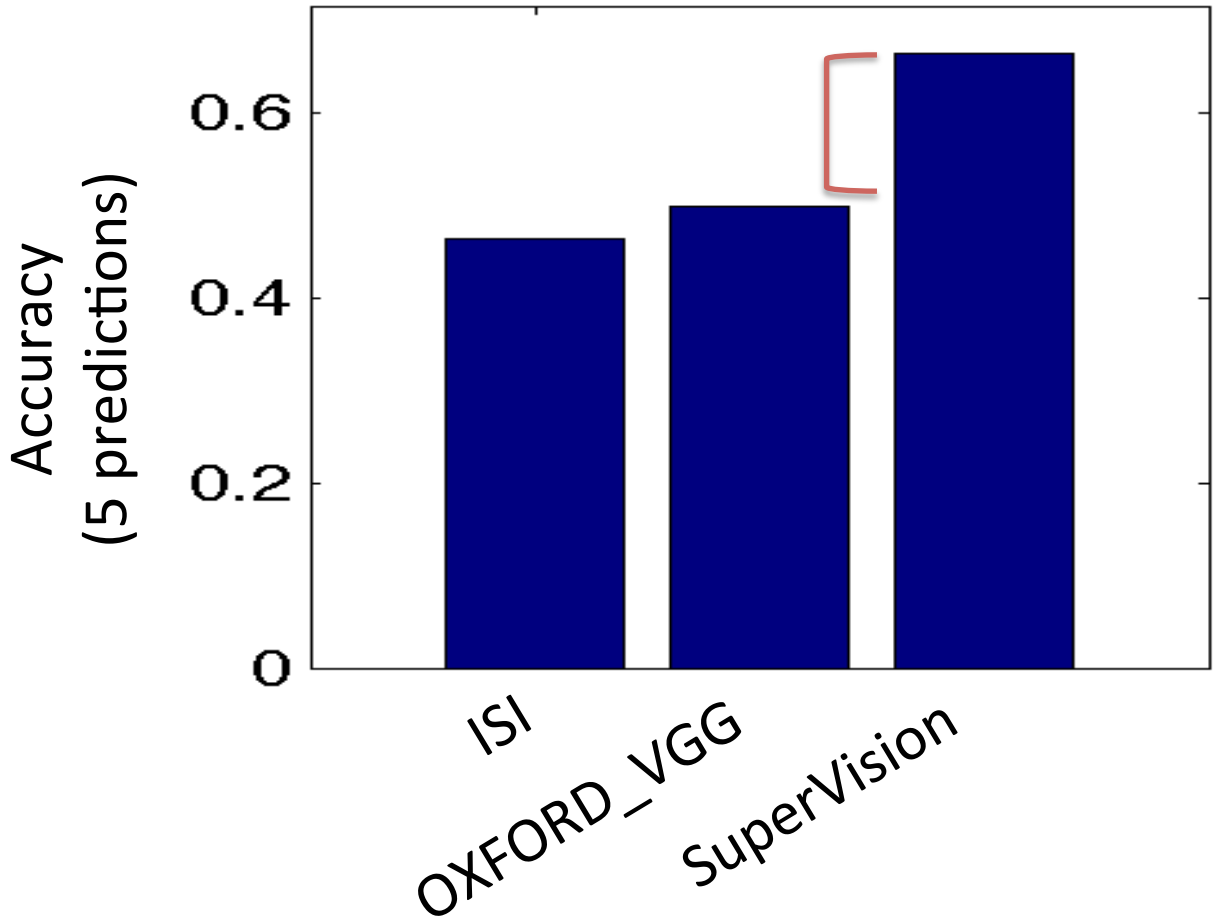
Steel drum

Output



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

ILSVRC Task 2: Classification + Localization



What happens under the hood?

What happens under the hood on classification+localization?

What happens under the hood on classification+localization?

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

Preliminaries:

- ILSVRC-500 (2012) dataset
- Leading algorithms

What happens under the hood on classification+localization?

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

Preliminaries:

- ILSVRC-500 (2012) dataset
- Leading algorithms

What happens under the hood on classification+localization?

- A closer look at small objects
- A closer look at textured objects

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

Preliminaries:

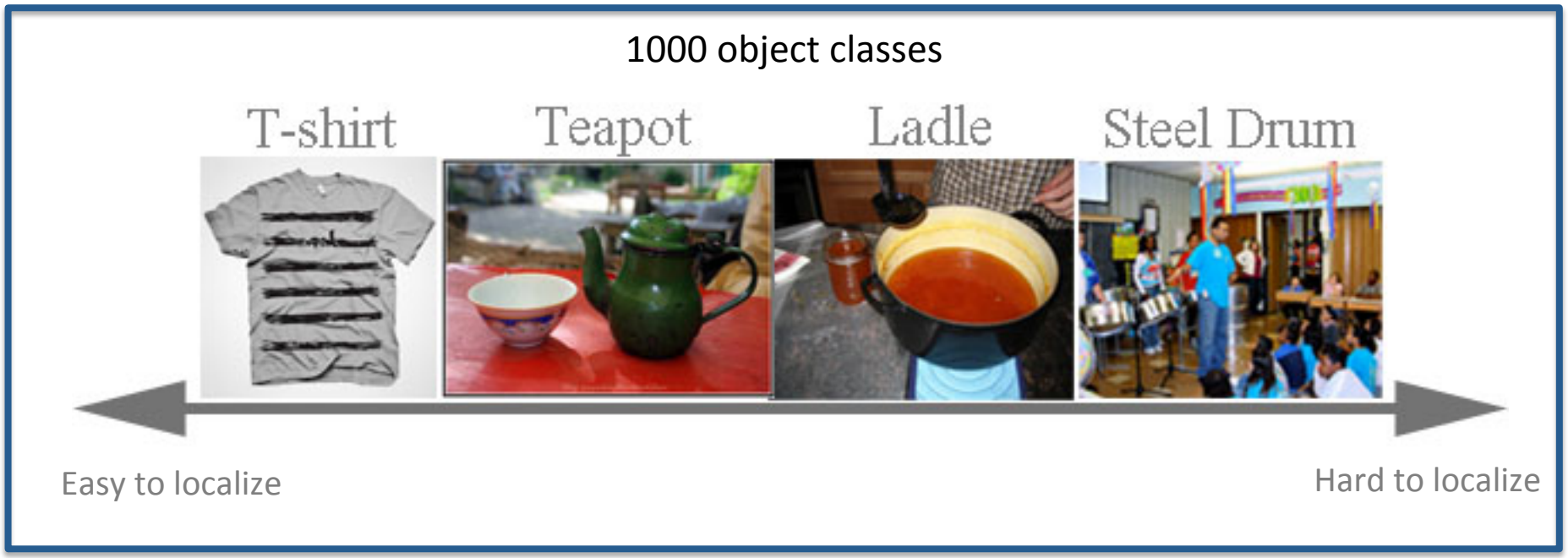
- ILSVRC-500 (2012) dataset
- Leading algorithms

What happens under the hood on classification+localization?

- A closer look at small objects
- A closer look at textured objects

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

ILSVRC (2012)


1000 object classes
T-shirt
Teapot
Ladle
Steel Drum
Easy to localize
Hard to localize

ILSVRC-500 (2012)

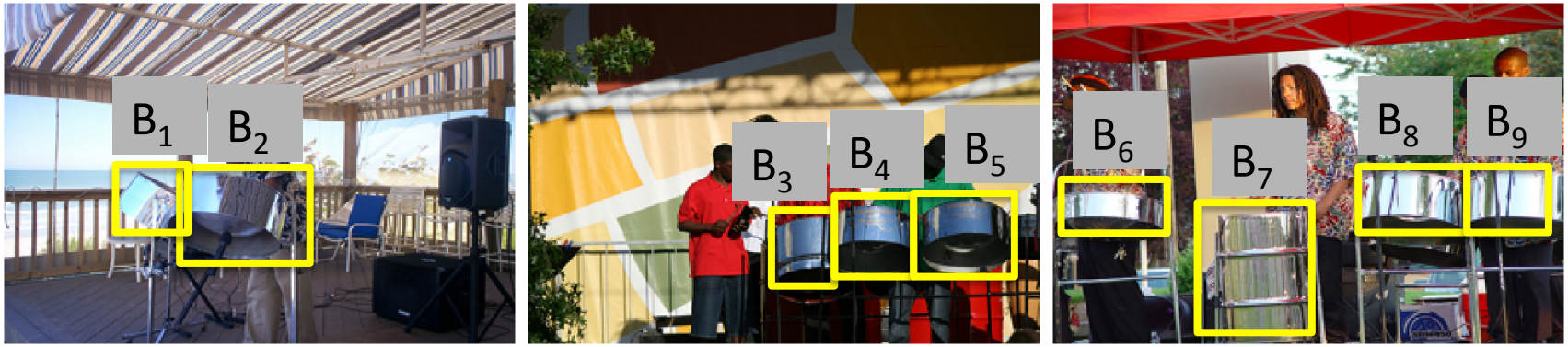
ILSVRC-500 (2012)

Object scale (fraction of image area occupied by target object)

ILSVRC-500 (2012)	500 object categories	25.3%
PASCAL VOC (2012)	20 object categories	25.2%

Chance Performance of Localization

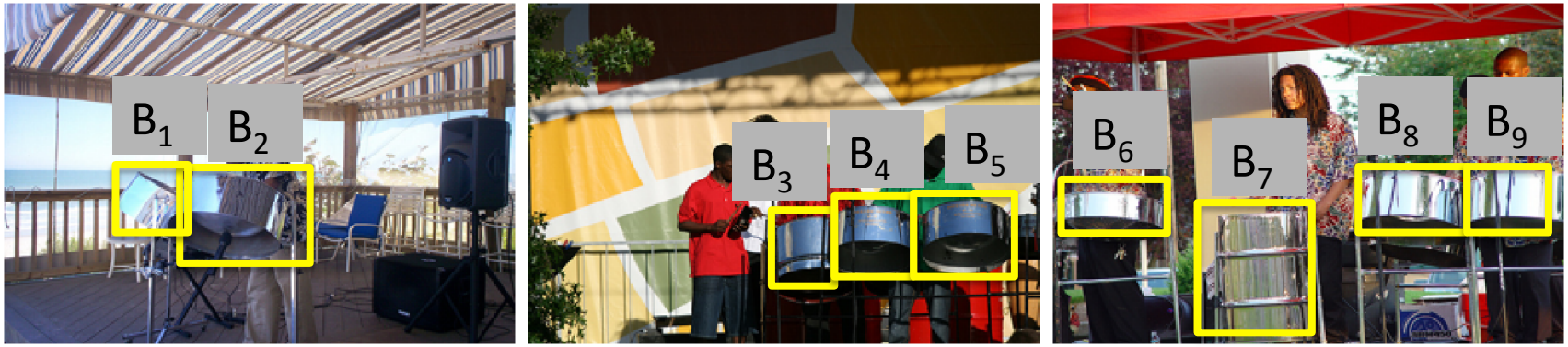
Steel drum



N = 9 here

Chance Performance of Localization

Steel drum

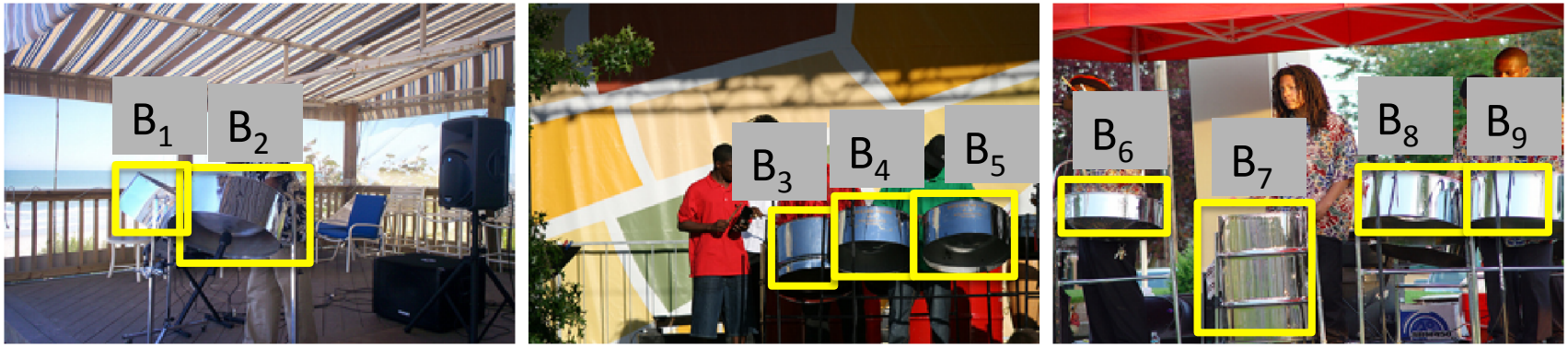


N = 9 here

$$\text{CPL} = \frac{\sum_i \sum_{j \neq i} IOU(B_i, B_j) \geq 0.5}{N(N-1)}$$

Chance Performance of Localization

Steel drum



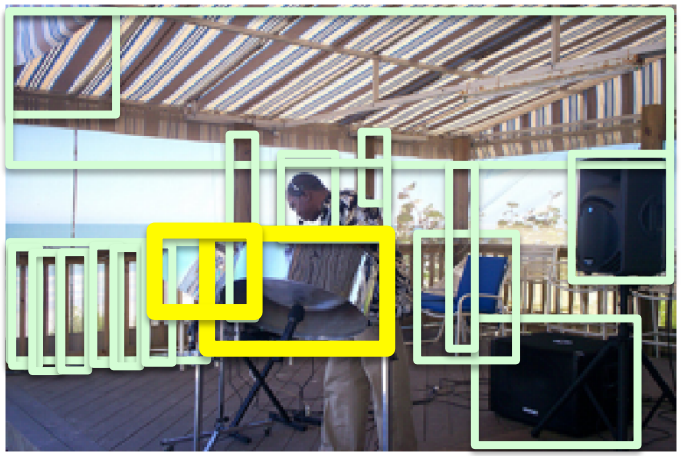
N = 9 here

$$\text{CPL} = \frac{\sum_i \sum_{j \neq i} IOU(B_i, B_j) \geq 0.5}{N(N-1)}$$

ILSVRC-500 (2012)	500 object categories	8.4%
PASCAL VOC (2012)	20 object categories	8.8%

Level of clutter

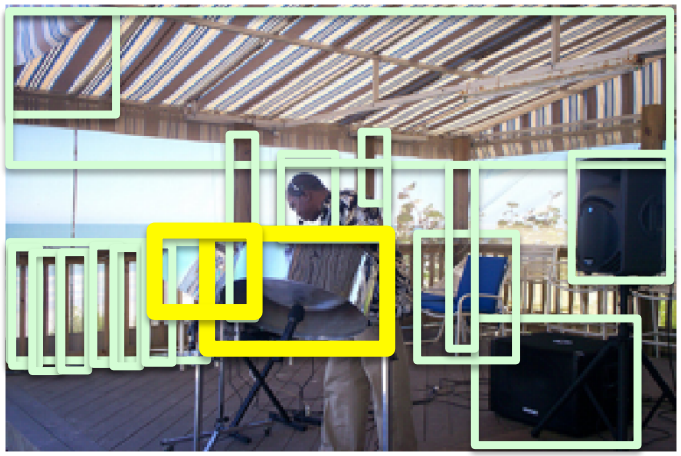
Steel drum



- Generate candidate object regions using method of
  Selective Search for Object Detection
  vanDeSande et al. ICCV 2011
- Filter out regions inside object
- Count regions

Level of clutter

Steel drum



- Generate candidate object regions using method of
  Selective Search for Object Detection
  vanDeSande et al. ICCV 2011
- Filter out regions inside object
- Count regions

ILSVRC-500 (2012)	500 object categories	128 ± 35
PASCAL VOC (2012)	20 object categories	130 ± 29

Preliminaries:

- ILSVRC-500 (2012) dataset – similar to PASCAL
- Leading algorithms

What happens under the hood on classification+localization?

- A closer look at small objects
- A closer look at textured objects

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

SuperVision (SV)

Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton (Krizhevsky NIPS12)

Image classification: Deep convolutional neural networks

- 7 hidden "weight" layers, 650K neurons, 60M parameters, 630M connections
- Rectified Linear Units, max pooling, dropout trick
- Randomly extracted 224x224 patches for more data
- Trained with SGD on two GPUs for a week, fully supervised

Localization: Regression on (x,y,w,h)

http://image-net.org/challenges/LSVRC/2012/supervision.pdf

SuperVision (SV)

Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton (Krizhevsky NIPS12)

Image classification: Deep convolutional neural networks

- 7 hidden "weight" layers, 650K neurons, 60M parameters, 630M connections
- Rectified Linear Units, max pooling, dropout trick
- Randomly extracted 224x224 patches for more data
- Trained with SGD on two GPUs for a week, fully supervised

Localization: Regression on (x,y,w,h)

http://image-net.org/challenges/LSVRC/2012/supervision.pdf

OXFORD_VGG (VGG)

Karen Simonyan, Yusuf Aytar, Andrea Vedaldi, Andrew Zisserman

Image classification: Fisher vector + linear SVM (Sanchez CVPR11)

- Root-SIFT (Arandjelovic CVPR12), color statistics, augmentation with patch location (x,y) (Sanchez PRL12)
- Fisher vectors: 1024 Gaussians, 135K dimensions
- No SPM, product quantization to compress
- Semi-supervised learning to find additional bounding boxes
- 1000 one-vs-rest SVM trained with Pegasos SGD
  - 135M parameters!

Localization: Deformable part-based models (Felzenszwalb PAMI10), without parts (root-only)

http://image-net.org/challenges/LSVRC/2012/oxford_vgg.pdf

Preliminaries:

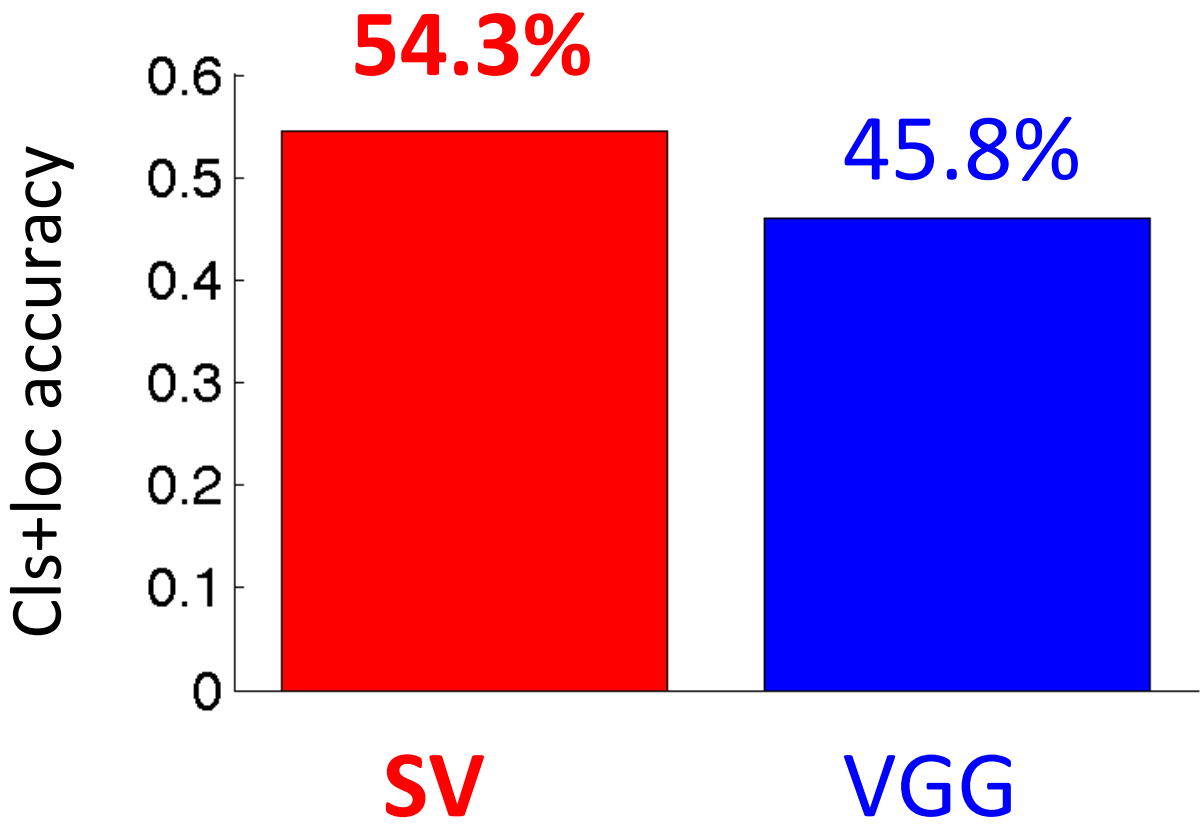
- ILSVRC-500 (2012) dataset – similar to PASCAL
- Leading algorithms: SV and VGG

What happens under the hood on classification+localization?

- A closer look at small objects
- A closer look at textured objects

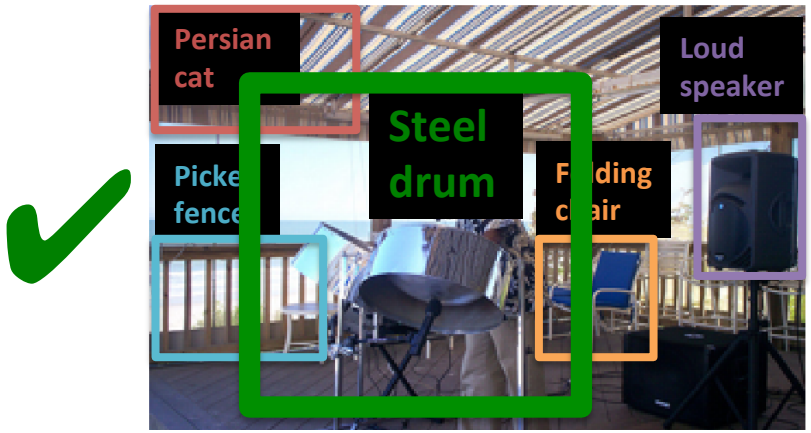
Olga Russakovsky, Jia Deng, Zhiheng Huang, Alex Berg, Li Fei-Fei
Detecting avocados to zucchinis: what have we done, and where are we going?
ICCV 2013 http://image-net.org/challenges/LSVRC/2012/analysis

Results on ILSVRC-500

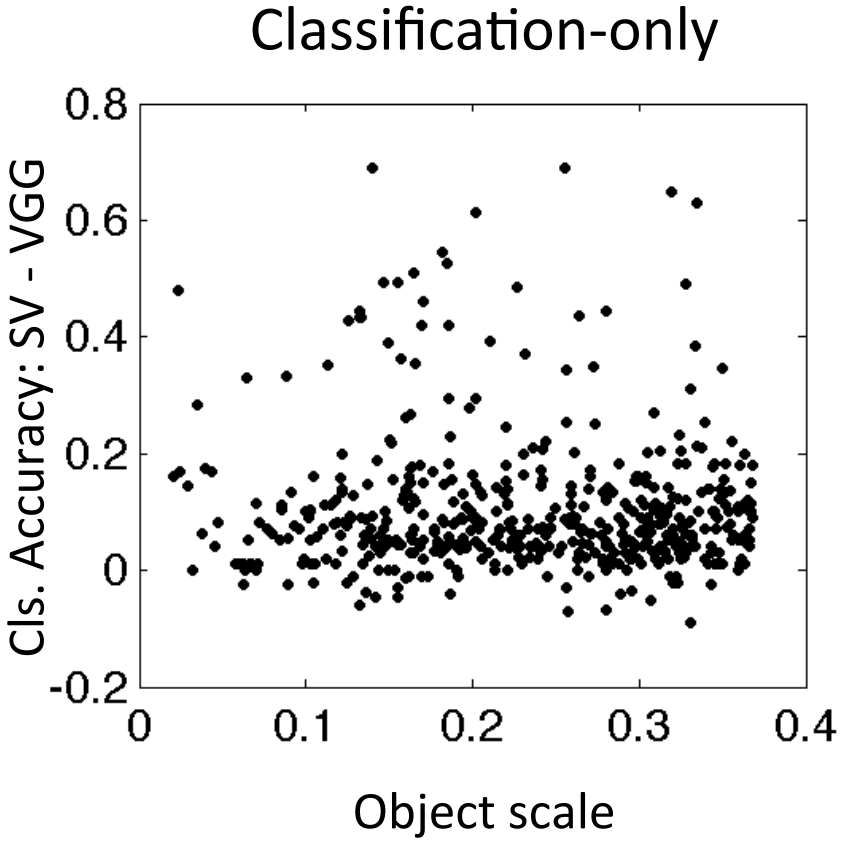


Difference in accuracy: SV versus VGG

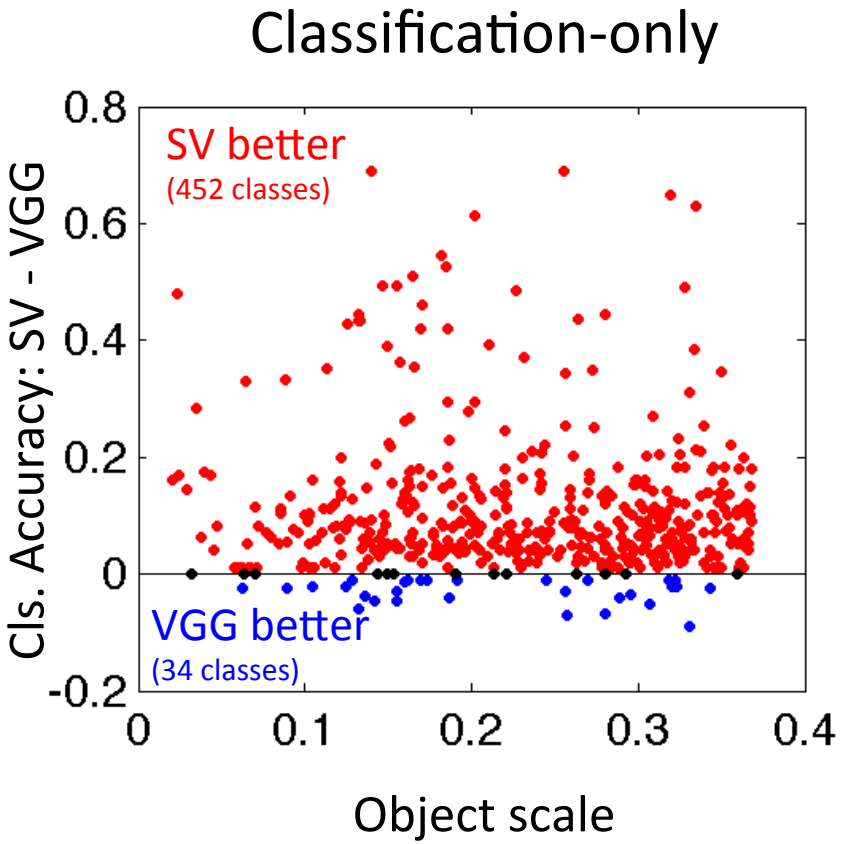
Classification-only



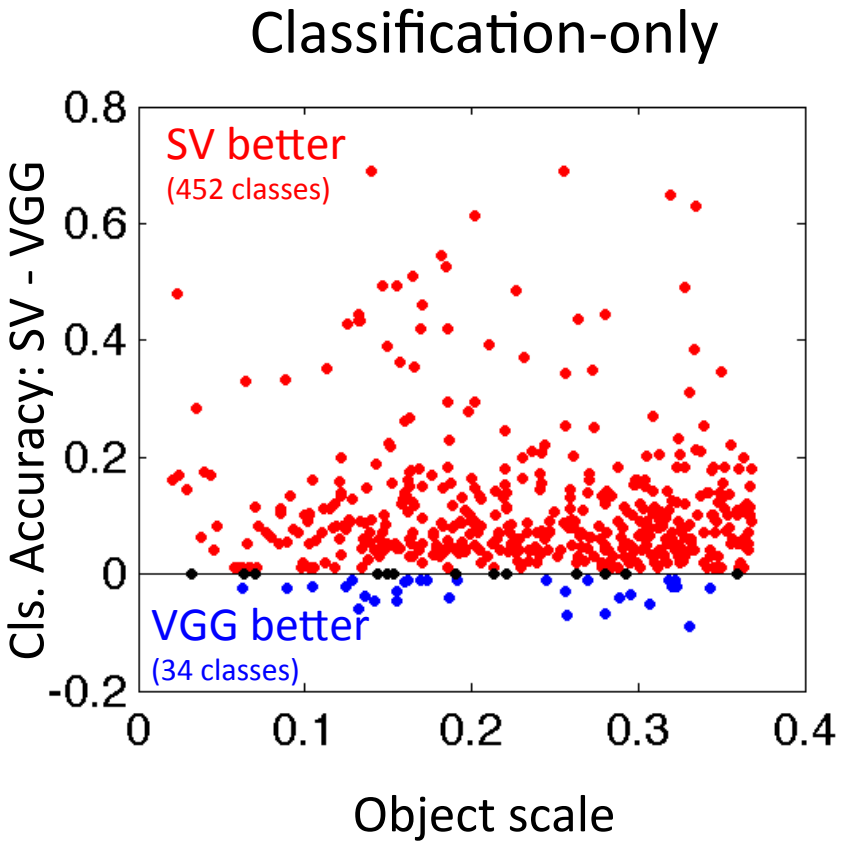
Difference in accuracy: SV versus VGG

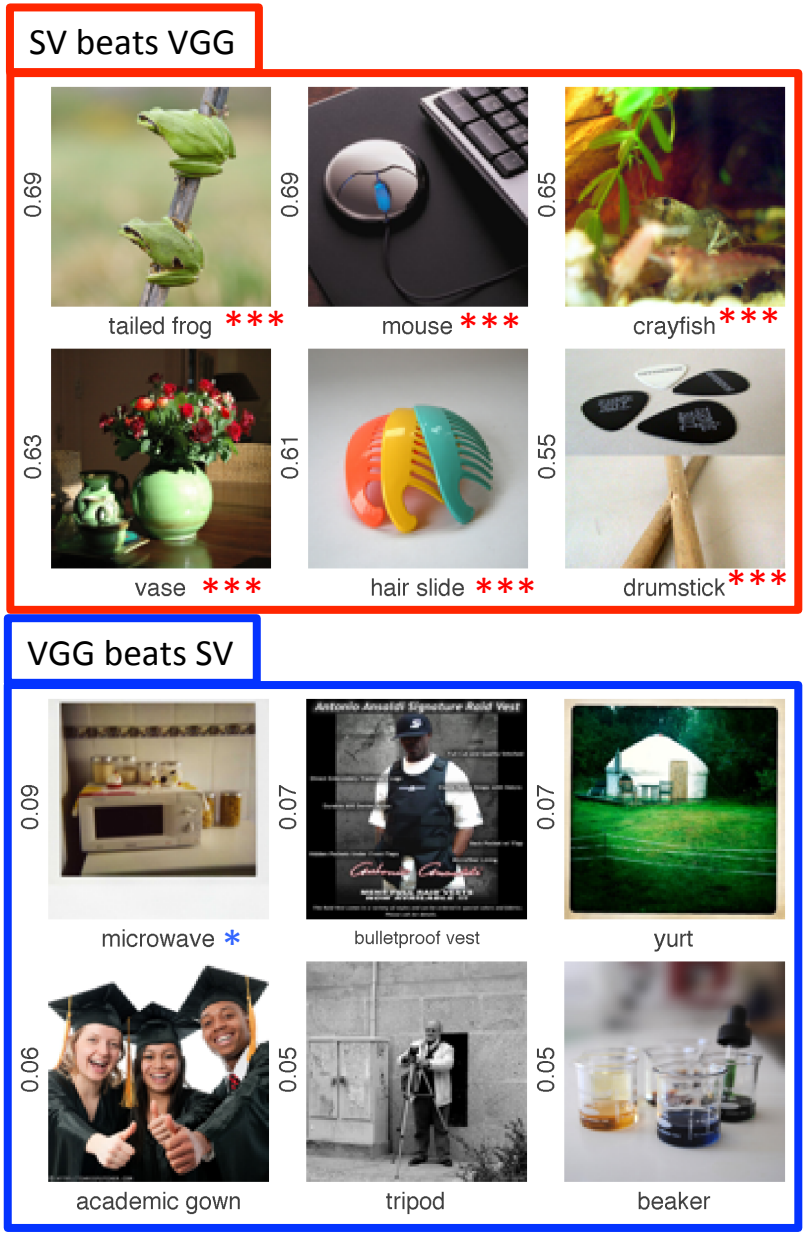


Difference in accuracy: SV versus VGG



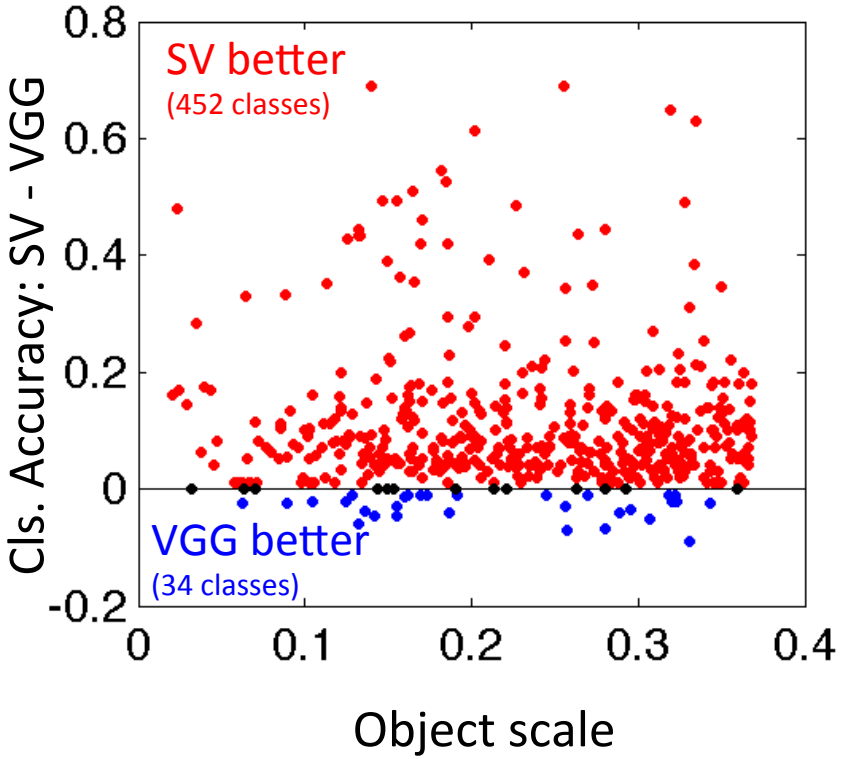
Difference in accuracy: SV versus VGG



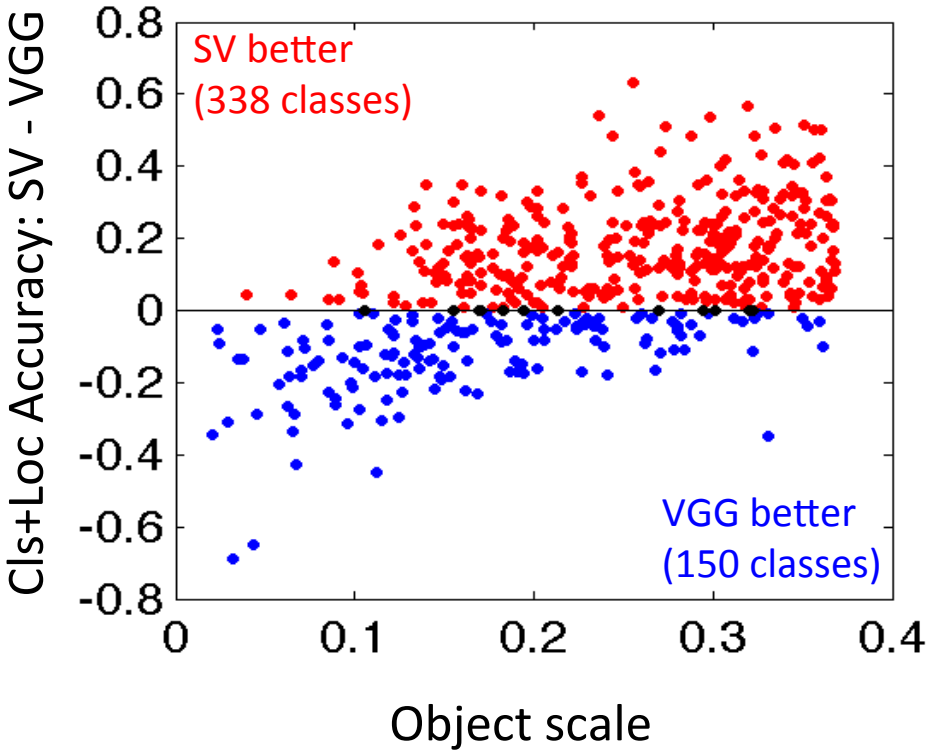


Difference in accuracy: SV versus VGG

Classification-only

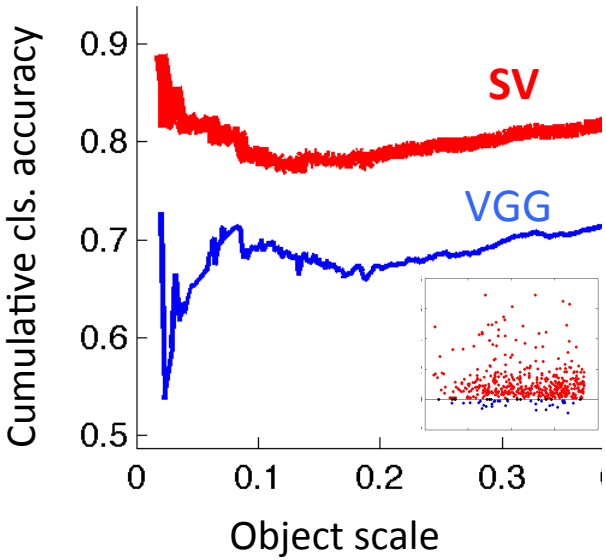


Classification+Localiation

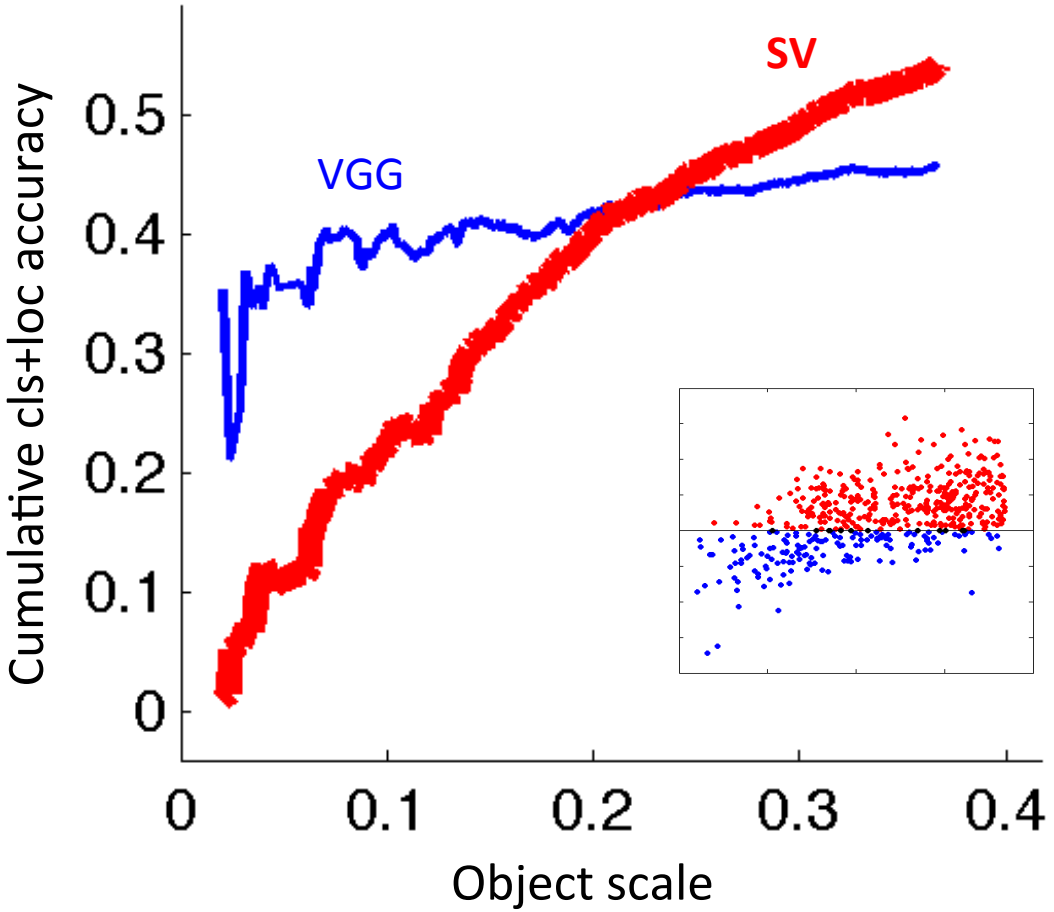


Cumulative accuracy across scales

Classification-only

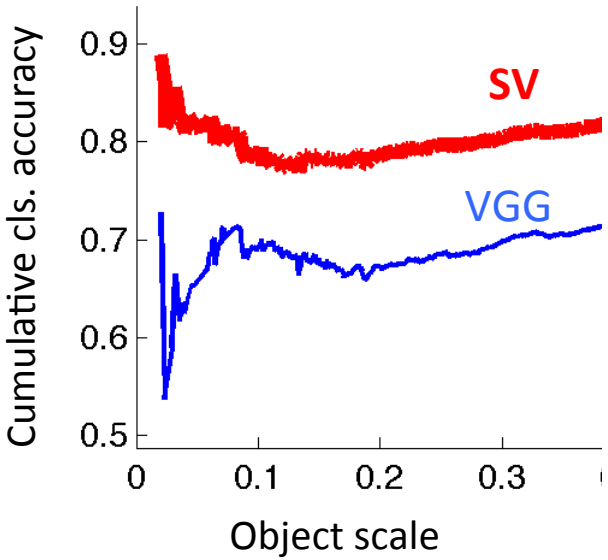


Classification+Localization

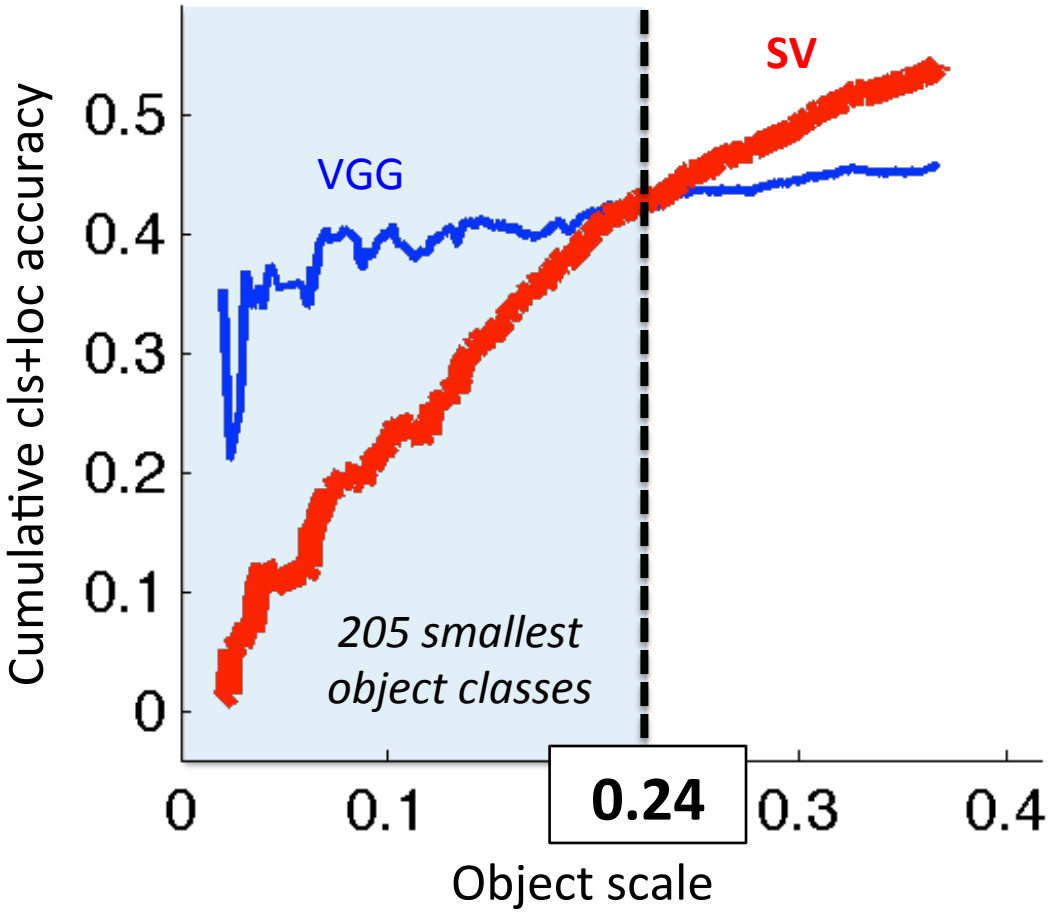


Cumulative accuracy across scales

Classification-only



Classification+Localization



Preliminaries:

- ILSVRC-500 (2012) dataset – similar to PASCAL
- Leading algorithms: SV and VGG

What happens under the hood on classification+localization?

- SV always great at classification, but VGG does better than SV at localizing small objects
- A closer look at textured objects

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

Preliminaries:

- ILSVRC-500 (2012) dataset – similar to PASCAL
- Leading algorithms: SV and VGG

What happens under the hood on classification+localization?

- SV always great at classification, but VGG does better than SV at localizing small objects WHY?
- A closer look at textured objects

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

Preliminaries:

- ILSVRC-500 (2012) dataset – similar to PASCAL
- Leading algorithms: SV and VGG

What happens under the hood on classification+localization?

- SV always great at classification, but VGG does better than SV at localizing small objects
- A closer look at textured objects

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

Textured objects (ILSVRC-500)

Textured objects (ILSVRC-500)

	No texture	Low texture	Medium texture	High texture
# classes	116	189	143	52

Textured objects (ILSVRC-500)

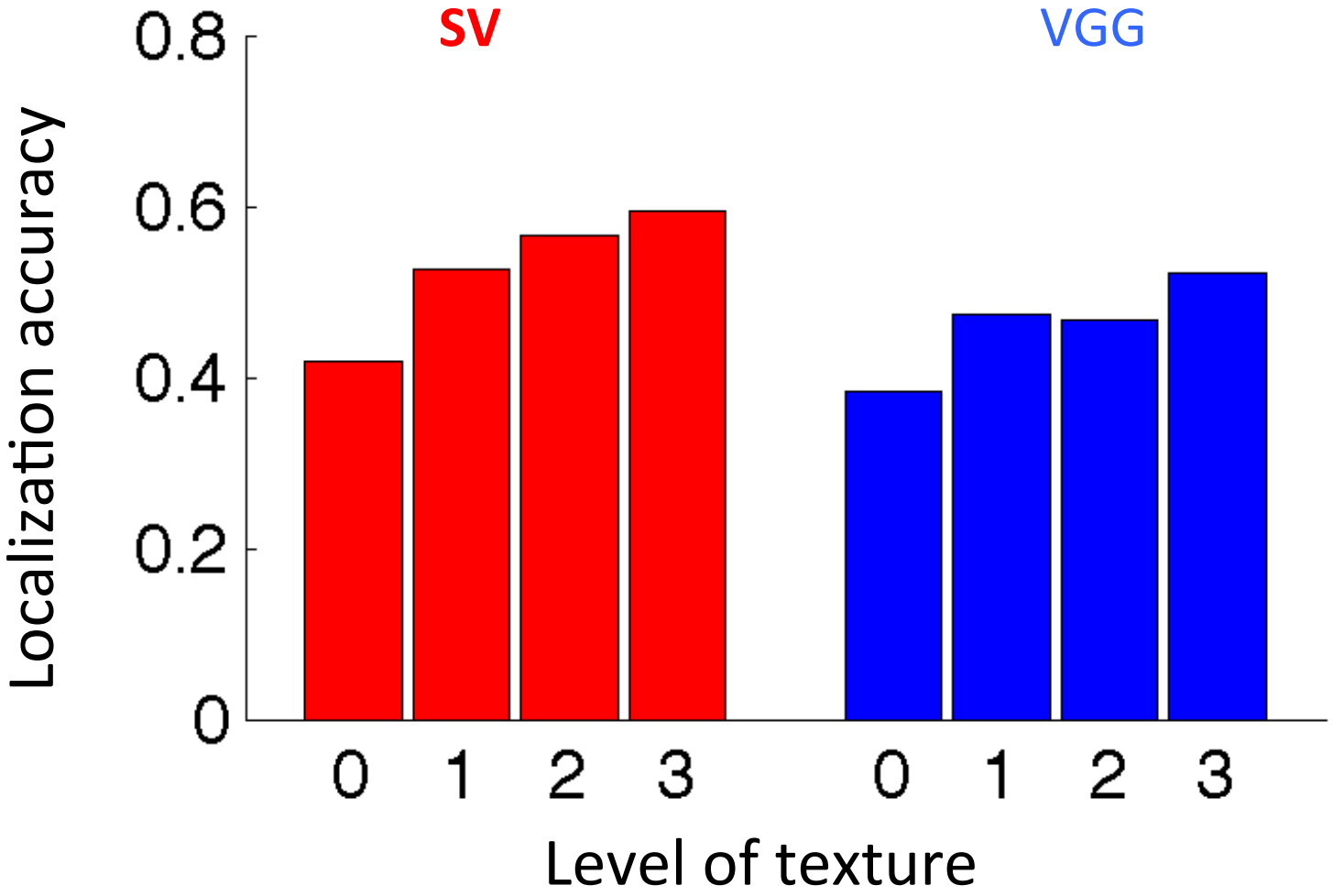
	No texture	Low texture	Medium texture	High texture
# classes	116	189	143	52
Object scale	20.8%	23.7%	23.5%	25.0%

Textured objects (416 classes)

	No texture	Low texture	Medium texture	High texture
# classes	116	~~189~~ 149	~~143~~ 115	~~52~~ 35
Object scale	20.8%	~~23.7%~~ 20.8%	~~23.5%~~ 20.8%	~~25.0%~~ 20.8%

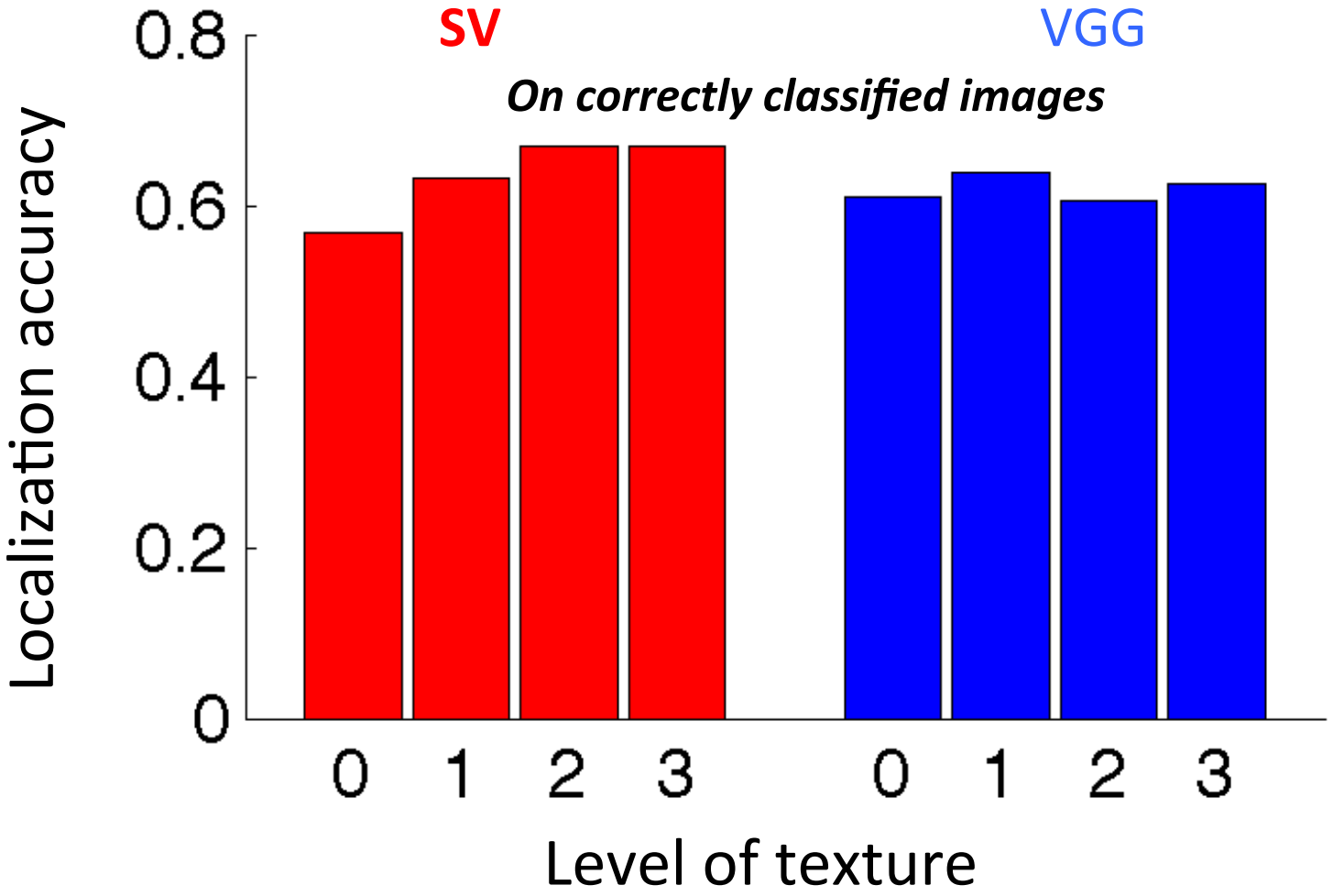
Localizing textured objects

(416 classes, same average object scale at each level of texture)



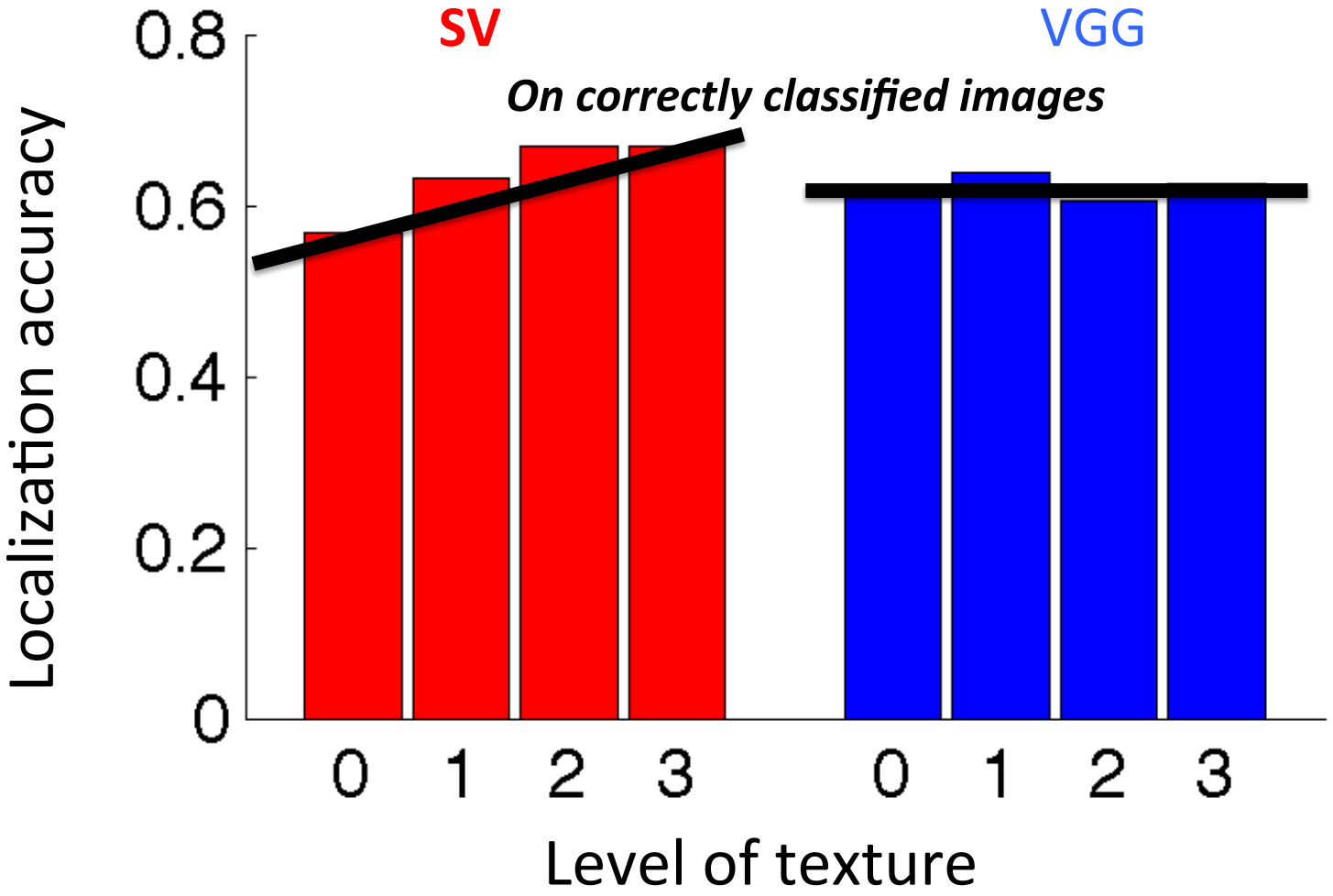
Localizing textured objects

(416 classes, same average object scale at each level of texture)



Localizing textured objects

(416 classes, same average object scale at each level of texture)



Preliminaries:

- ILSVRC-500 (2012) dataset – similar to PASCAL
- Leading algorithms: SV and VGG

What happens under the hood on classification+localization?

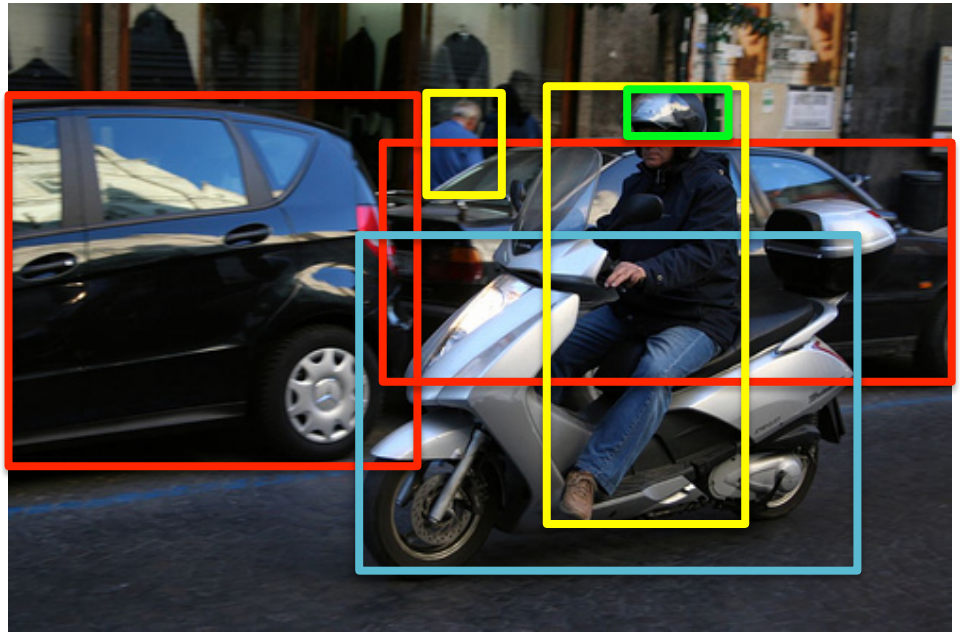
- SV always great at classification, but VGG does better than SV at localizing small objects
- Textured objects easier to localize, especially for SV

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

ILSVRC 2013
with large-scale object detection

NEW

Fully annotated 200 object classes across 60,000 images



Allows evaluation of generic object detection
in cluttered scenes at scale

http://image-net.org/challenges/LSVRC/2013/

ILSVRC 2013
with large-scale object detection

NEW

Statistics		PASCAL VOC 2012		ILSVRC 2013
Object classes		20	10x	**200**
Training	Images	5.7K		**395K**
	Objects	13.6K	25x	**345K**
Validation	Images	5.8K		**20.1K**
	Objects	13.8K	4x	**55.5K**
Testing	Images	11.0K		**40.1K**
	Objects	---		---

More than 50,000 person instances annotated

http://image-net.org/challenges/LSVRC/2013/

ILSVRC 2013
with large-scale object detection

NEW

- 159 downloads so far:

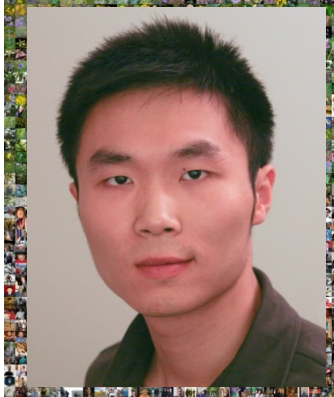
  http://image-net.org/challenges/LSVRC/2013/

- Submission deadline Nov. 15th
- ICCV workshop on December 7th, 2013

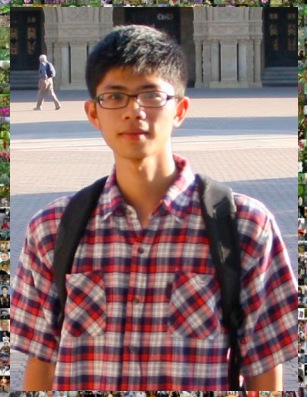
- Fine-Grained Challenge 2013:

  https://sites.google.com/site/fgcomp2013/

Thank you!



Dr. Jia Deng
Stanford U.



Zhiheng Huang
Stanford U.

Jonathan Krause
Stanford U.

Sanjeev Satheesh
Stanford U.

Hao Su
Stanford U.

Prof. Alex Berg
UNC Chapel Hill

Google

Microsoft Research