Object Detection with Deep Learning

CVPR 2014 Tutorial

Pierre Sermanet, Google Research

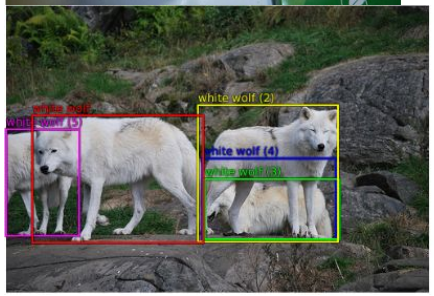
What is object detection?

- classification
- localization
- detection
- segmentation

Top 5:
pencil sharpener
pool table
hand blower
oil filter
packet

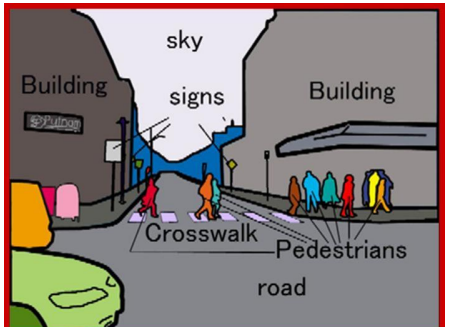
Groundtruth:
pencil sharpener

ILSVRC2012_val_00010000.JPEG



Groundtruth:
white wolf
white wolf (2)
white wolf (3)
white wolf (4)
white wolf (5)

Groundtruth:
tv or monitor
tv or monitor (2)
tv or monitor (3)
person
remote control
remote control (2)



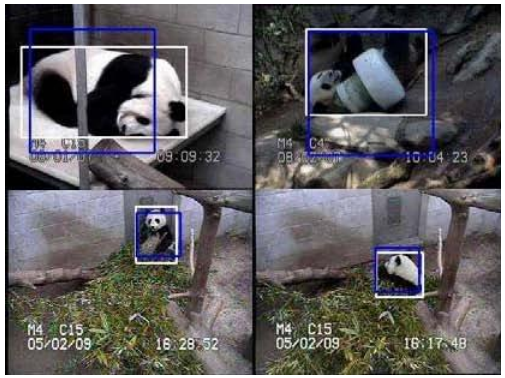
difficulty

Why is object detection important?

- Perception is one of the biggest bottlenecks of
  - Robotics

  - Self-driving cars

  - Surveillance



Is it deployed?

- **classification**
  - personal image search (Google, Baidu, Bing)
- **detection**
  - face detection
    - cameras
    - election duplicate votes
    - CCTV
    - border control
    - casinos
    - visa processing
    - crime solving
    - prosopagnosia (face blindness)
  - objects
    - **license plates**
    - **pedestrian detection** (Daimler, MobileEye):
      - e.g. 2013 Mercedes-Benz E-Class and S-Class: warning and automatic braking reducing accidents and severity
    - **vehicle detection** for forward collision warning (MobileEye)
    - **traffic sign detection** (MobileEye)

What datasets for detection?

- **PASCAL** pascallin.ecs.soton.ac.uk/challenges/VOC
- **ImageNet** www.image-net.org/challenges/LSVRC/2014
- **Sun** sundatabase.mit.edu
- **Microsoft COCO** mscoco.org

	# classes	average # categories per image	average # instances per image	average object scale	average resolution	# images				# objects			
						total	train	val	test	total	train	val	test
PASCAL	20	1.521	2.711	0.207	469x387	22k	6k	6k	10k	42k?	14k	14k	-
ImageNet13	200	1.534	2.758	0.170	482x415	516k	456k	20k	40k	648k?	480k	56k	-
Sun	4919	9.8	16.9	0.1040	732x547	16873	-			285k	-		
COCO	91	3.5	7.6	0.117	578x483	328k	164k	82k	82k	2500k	~1250k	~625k	~625k

The pascal visual object classes (voc) challenge. Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. *International journal of computer vision* 88, no. 2 (2010): 303-338.
Imagenet: A large-scale hierarchical image database. Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248-255. IEEE, 2009.
SUN Database: Large-scale Scene Recognition from Abbey to Zoo, Jianxiong Xiao, James Hays, Krista Ehinger, Aude Oliva, and Antonio Torralba. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010.
Microsoft COCO: Common Objects in Context, Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick, http://arxiv.org/abs/1405.0312, May 2014

What datasets for detection?

- **Microsoft COCO**
  - release: summer 2014
  - segmented instances
  - non-iconic images
  - ~80% of images have >1 categories or instances

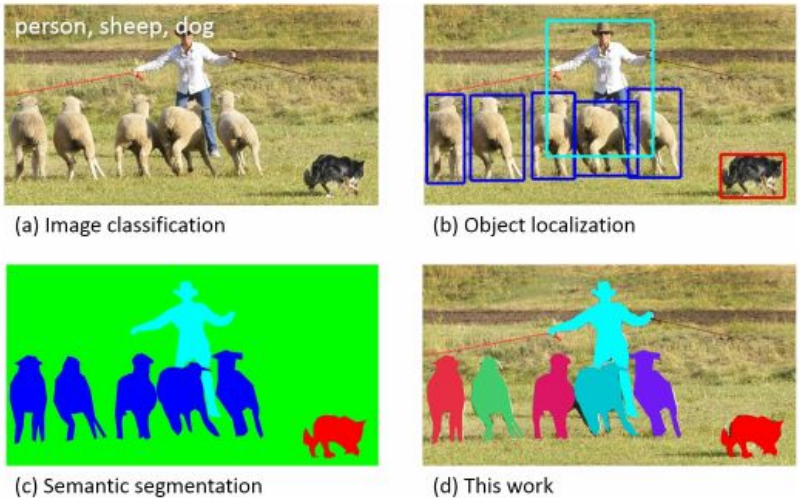
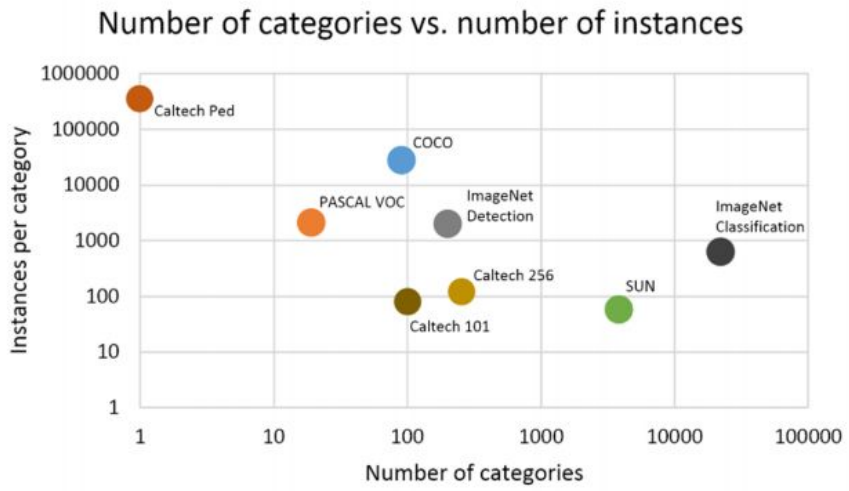


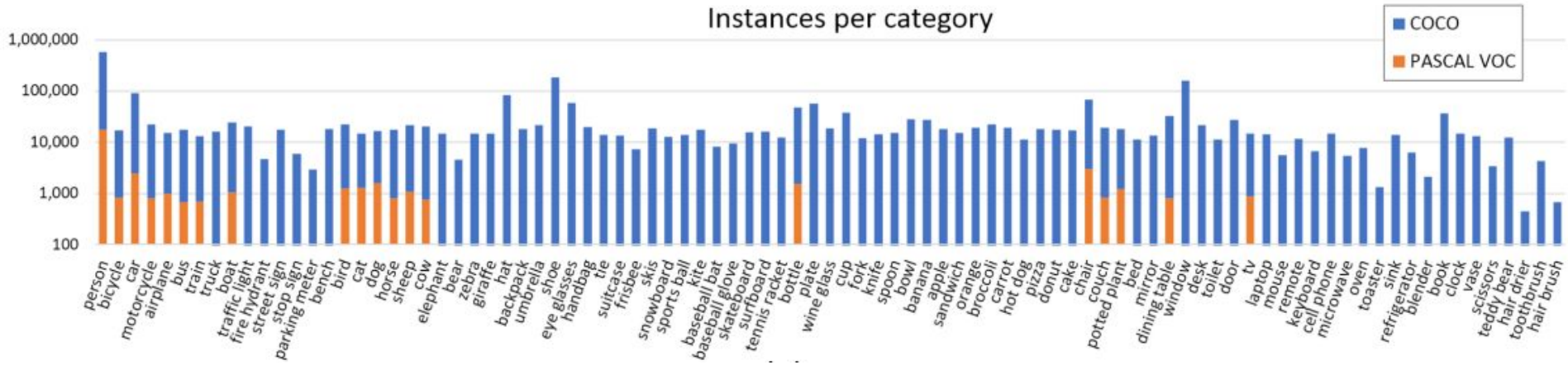
Fig. 2: Example of (a) iconic object images, (b) iconic scene images, and (c) non-iconic images.

Microsoft COCO: Common Objects in Context, Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick, http://arxiv.org/abs/1405.0312, May 2014

What datasets for detection?

- **Microsoft COCO**
  - useful/common categories
  - even distribution (as opposed to long-tail for SUN)

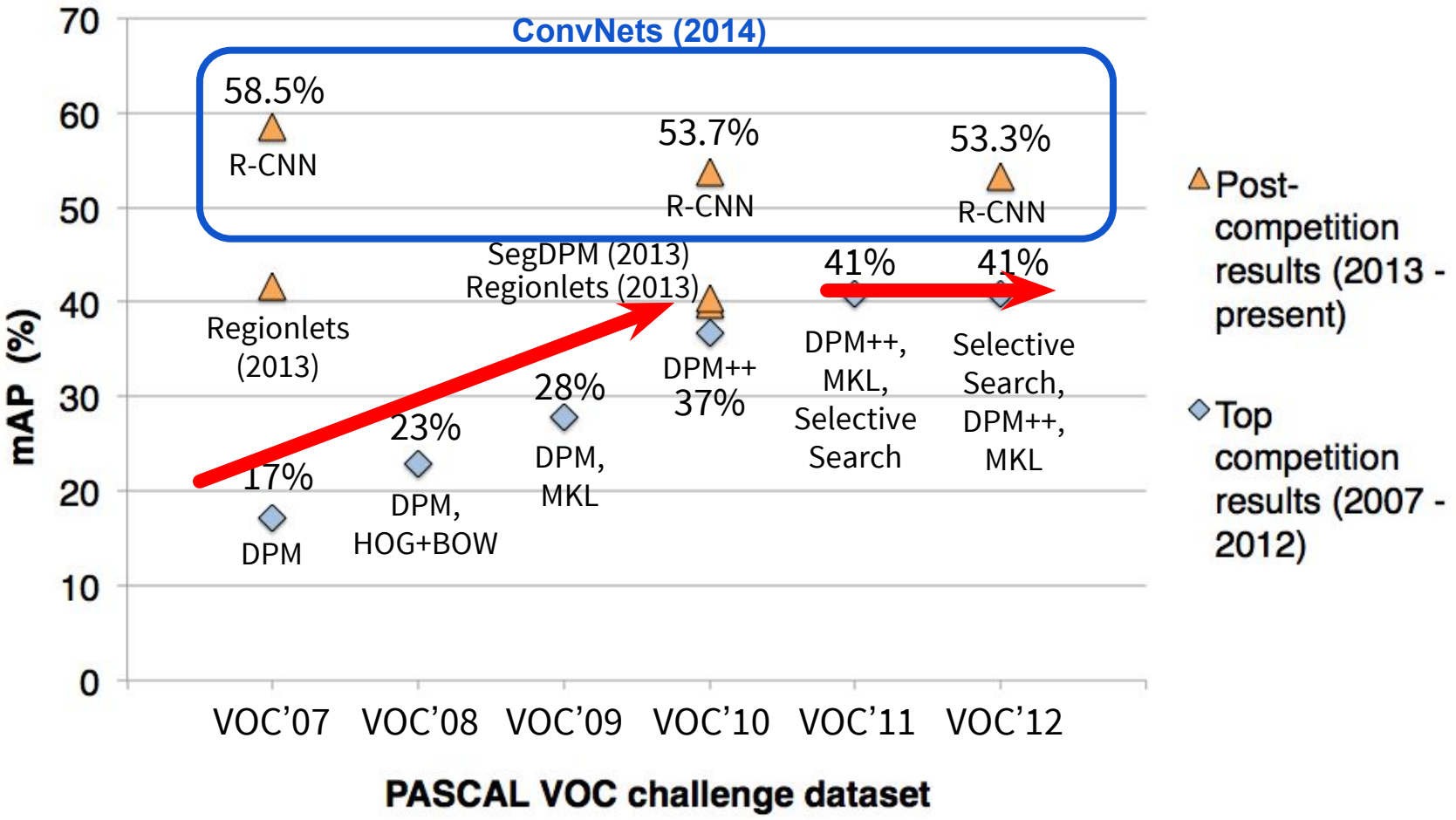




Microsoft COCO: Common Objects in Context, Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick, http://arxiv.org/abs/1405.0312, May 2014

Recent history of object detection

- **Large improvements** using Deep Learning [Girshick'13/14]



Rich feature hierarchies for accurate object detection and semantic segmentation. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *arXiv preprint arXiv:1311.2524* (2013).

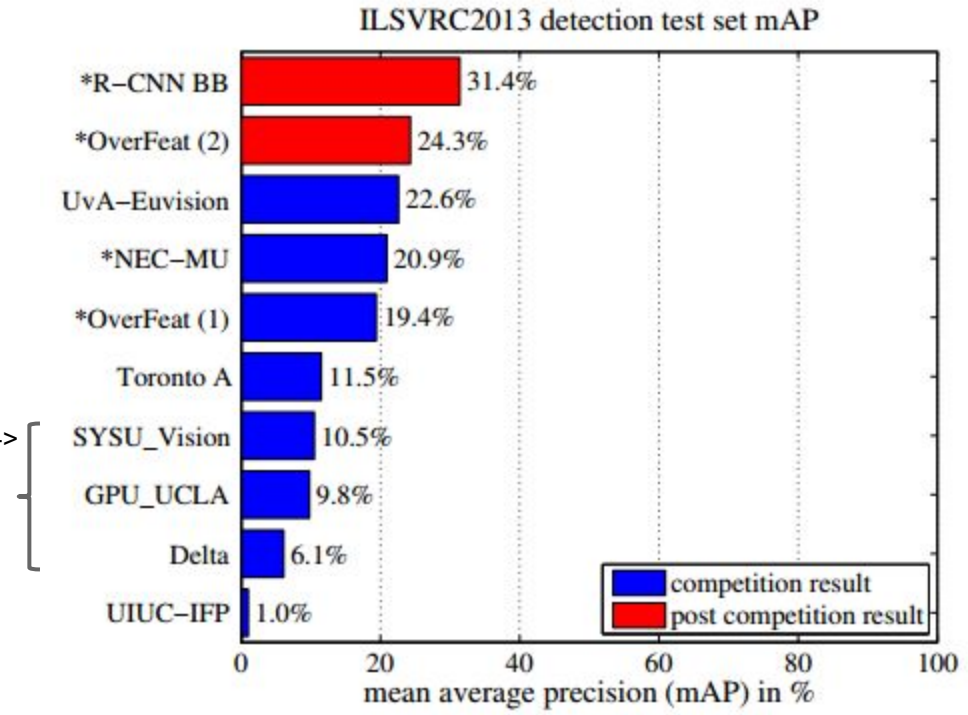
The PASCAL Visual Object Classes Challenge - a Retrospective, Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. Accepted for International Journal of Computer Vision, 2014

Recent history of object detection

- **Large improvements** using Deep Learning
- ImageNet 2013 detection (new challenge)
  - top entries all use **Convolutional Networks (ConvNets)**

No ConvNets

Deformable Parts Models (DPM) ->



Rich feature hierarchies for accurate object detection and semantic segmentation. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *arXiv preprint arXiv:1311.2524* (2013).

Overfeat: Integrated recognition, localization and detection using convolutional networks. Sermanet, Pierre, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. *arXiv preprint arXiv:1312.6229* (2013), International Conference on Learning Representations (ICLR)` 2014.

ConvNets breakthroughs for visual tasks

	Dataset	Performance	Score
[Sermanet et al 2014]: OverFeat (fine-tuned features for each task) (tasks are ordered by increasing difficulty)			
• image classification	ImageNet LSVRC 2013	competitive	13.6 % error
	Dogs vs Cats Kaggle challenge 2014	**state of the art**	98.9%
• object localization	ImageNet LSVRC 2013	**state of the art**	29.9% error
• object detection	ImageNet LSVRC 2013	competitive	24.3% mAP
[Razavian et al, 2014]: public OverFeat library (no retraining) + SVM **(simplest approach possible on purpose, no attempt at more complex classifiers)** (tasks are ordered by "distance" from classification task on which OverFeat was trained)			
• image classification	Pascal VOC 2007	competitive	77.2% mAP
• scene recognition	MIT-67	**state of the art**	69% mAP
• fine grained recognition	Caltech-UCSD Birds 200-2011	competitive	61.8% mAP
	Oxford 102 Flowers	**state of the art**	86.8% mAP
• attribute detection	UIUC 64 object attributes	**state of the art**	91.4% mAUC
	H3D Human Attributes	competitive	73% mAP
• image retrieval (search by image similarity)	Oxford 5k buildings	**state of the art**	68% mAP?
	Paris 6k buildings	**state of the art**	79.5% mAP?
	Sculp6k	competitive	42.3% mAP?
	Holidays	**state of the art**	84.3% mAP?
	UKBench	**state of the art**	91.1% mAP?

Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, Yann LeCun, **OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks**, http://arxiv.org/abs/1312.6229, ICLR 2014
Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson, **CNN Features off-the-shelf: an Astounding Baseline for Recognition**, http://arxiv.org/abs/1403.6382, DeepVision CVPR 2014 workshop

ConvNets breakthroughs for visual tasks

	Dataset	Performance	Score
[Zeiler et al 2013]			
• image classification	ImageNet LSVRC 2013	**state of the art**	11.2% error
	Caltech-101 (15, 30 samples per class)	competitive	83.8%, 86.5%
	Caltech-256 (15, 60 samples per class)	**state of the art**	65.7%, 74.2%
	Pascal VOC 2012	competitive	79% mAP
[Donahue et al, 2014]: DeCAF+SVM			
• image classification	Caltech-101 (30 classes)	**state of the art**	86.91%
• domain adaptation	Amazon -> Webcam, DSLR -> Webcam	**state of the art**	82.1%, 94.8%
• fine grained recognition	Caltech-UCSD Birds 200-2011	**state of the art**	65.0%
• scene recognition	SUN-397	competitive	40.9%
[Girshick et al, 2013]			
• image detection	Pascal VOC 2007	**state of the art**	48.0% mAP
	Pascal VOC 2010 (comp4)	**state of the art**	43.5% mAP
	ImageNet LSVRC 2013	**state of the art**	31.4% mAP
• image segmentation	Pascal VOC 2011 (comp6)	**state of the art**	47.9% mAP
[Oquab et al, 2013]			
• image classification	Pascal VOC 2007	**state of the art**	77.7% mAP
	Pascal VOC 2012	**state of the art**	82.8% mAP
	Pascal VOC 2012 (action classification)	**state of the art**	70.2% mAP

M.D. Zeiler, R. Fergus, **Visualizing and Understanding Convolutional Networks,** Arxiv 1311.2901 http://arxiv.org/abs/1311.2901

J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. **Decaf: A deep convolutional activation feature for generic visual recognition**. In ICML, 2014, http://arxiv.org/abs/1310.1531

R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. **Rich feature hierarchies for accurate object detection and semantic segmentation**. arxiv:1311.2524 [cs.CV], 2013, http://arxiv.org/abs/1311.2524

M. Oquab, L. Bottou, I. Laptev, and J. Sivic. **Learning and transferring mid-level image representations using convolutional neural networks.** Technical Report HAL-00911179, INRIA, 2013. http://hal.inria.fr/hal-00911179

ConvNets breakthroughs for visual tasks

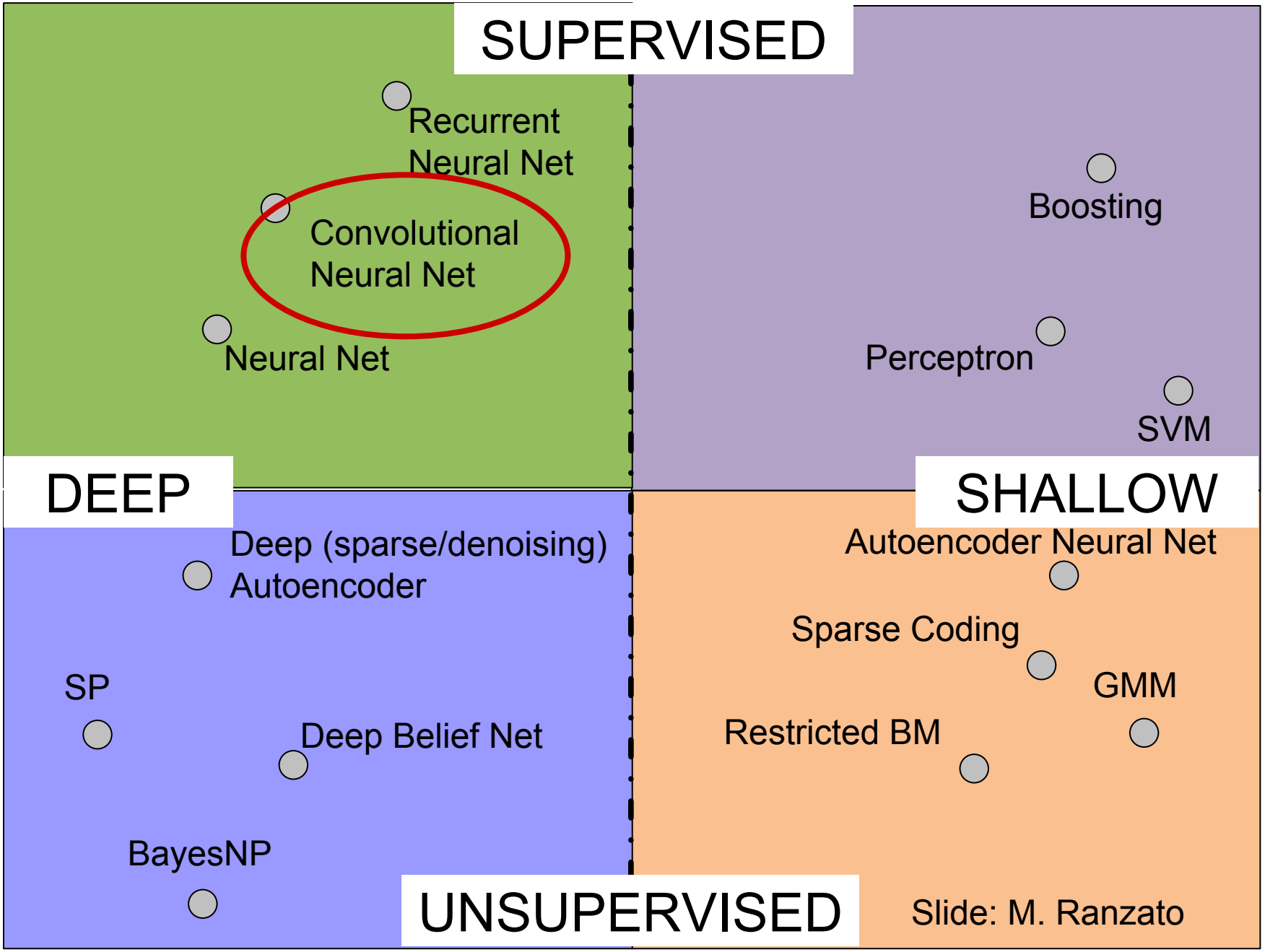
	Dataset	Performance	Score
[Khan et al 2014] • shadow detection	UCF CMU UIUC	**state of the art** **state of the art** **state of the art**	90.56% 88.79% 93.16%
[Sander Dieleman, 2014] • image attributes	Kaggle Galaxy Zoo challenge	**state of the art**	0.07492

S. H. Khan, M. Bennamoun, F. Sohel, R. Togneri. **Automatic Feature Learning for Robust Shadow Detection,** CVPR 2014
Sander Dieleman, Kaggle Galaxy Zoo challenge 2014 http://benanne.github.io/2014/04/05/galaxy-zoo.html

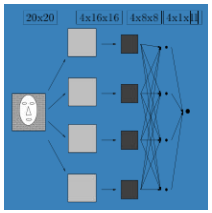
ConvNets breakthroughs for visual tasks

[Razavian et al, 2014]:
"It can be concluded that **from now on, deep learning with CNN has to be considered as the primary candidate in essentially any visual recognition task**."

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson, **CNN Features off-the-shelf: an Astounding Baseline for Recognition**, http://arxiv.org/abs/1403.6382, DeepVision CVPR 2014 workshop

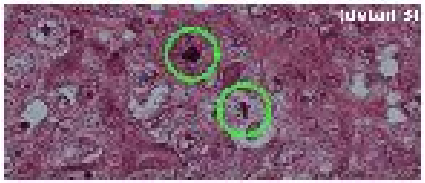

SUPERVISED
Recurrent
Neural Net
Convolutional
Neural Net
Neural Net
Boosting
Perceptron
SVM
DEEP
SHALLOW
Deep (sparse/denoising)
Autoencoder
Autoencoder Neural Net
Sparse Coding
GMM
SP
Deep Belief Net
Restricted BM
BayesNP
UNSUPERVISED
Slide: M. Ranzato

History of detection with ConvNets

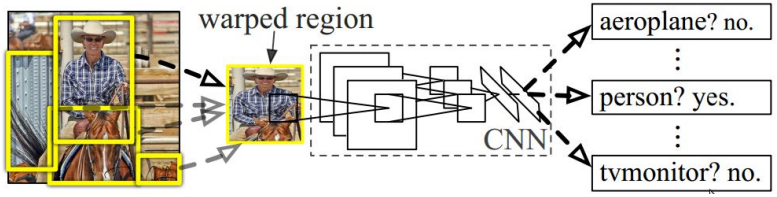
Vaillant, Monrocq, LeCun 1994
Osadchy, LeCun, Miller 2004
Face detection with pose estimation!

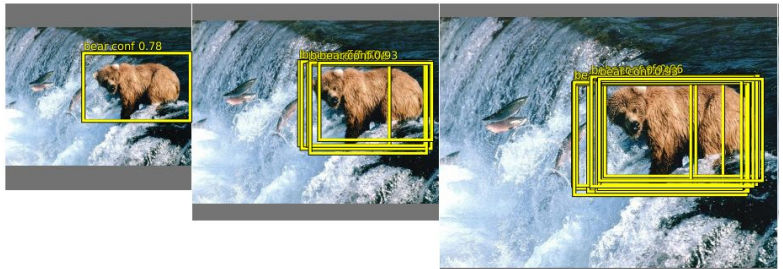
LeCun, Huang, Bottou 2004
NORB dataset



Cireşan et al. 2013
Mitosis detection

Sermanet et al. 2013
Pedestrian detection

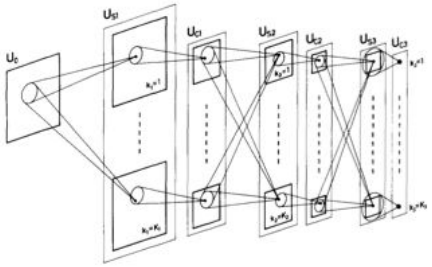


PASCAL detection
Girshick et al. 2013
Szegedy, Toshev, Erhan 2013

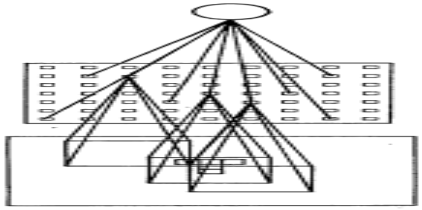
ImageNet detection
Girshick et al. 2014 (R-CNN)
Sermanet et al. 2014 (OverFeat)

slide: Girshick/Sermanet

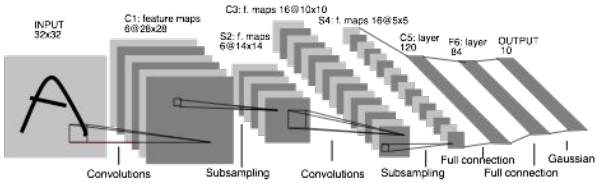
History of ConvNets



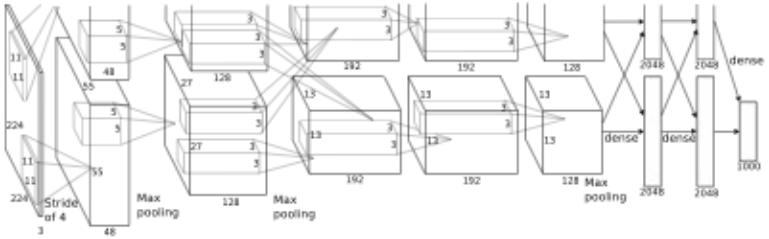
Fukushima 1980
Neocognitron



Rumelhart, Hinton, Williams 1986
"T" versus "C" problem



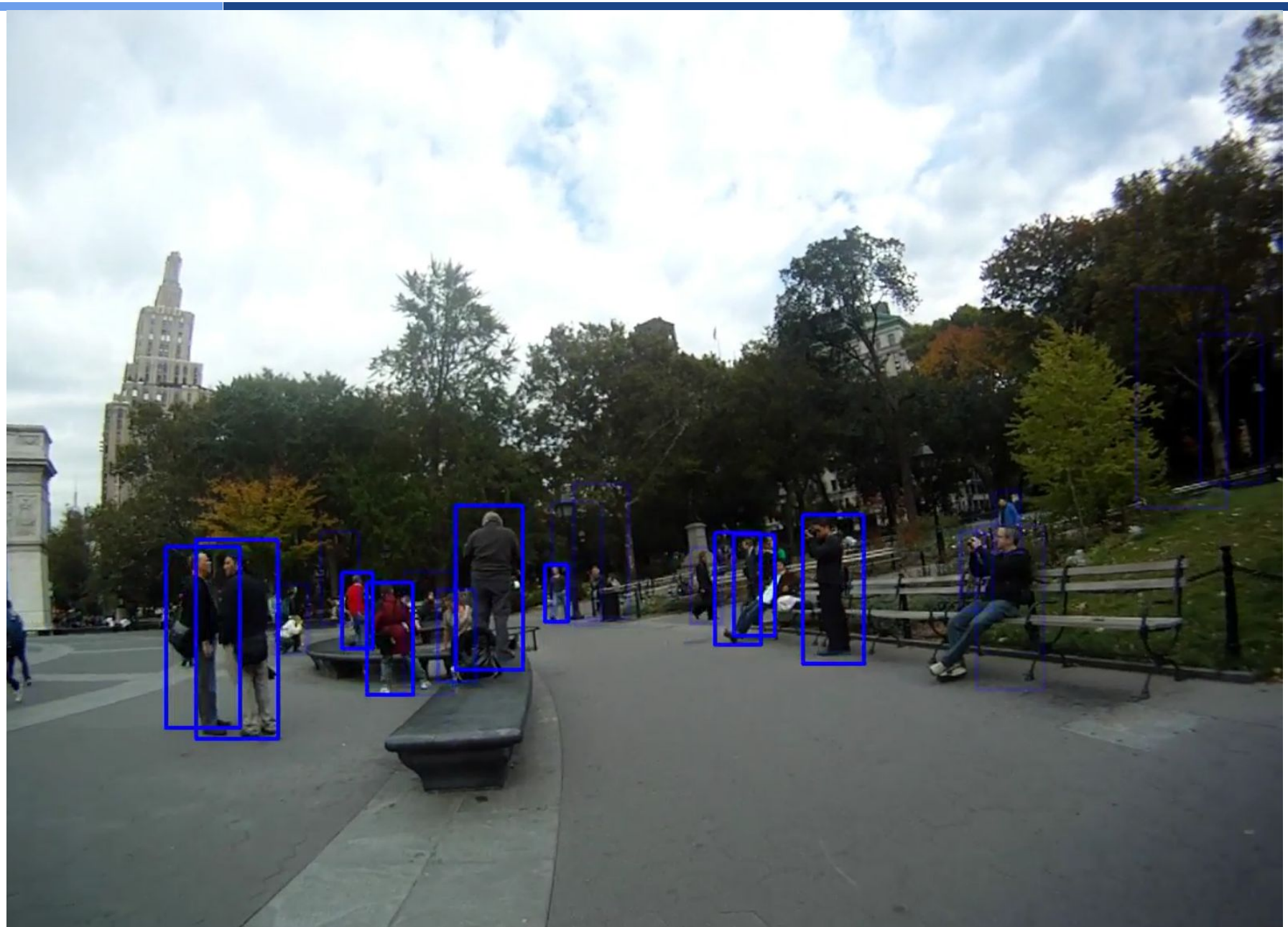
LeCun et al. 1989-1998
Hand-written digit reading



Krizhevksy, Sutskever, Hinton 2012
ImageNet classification breakthrough
"SuperVision" CNN

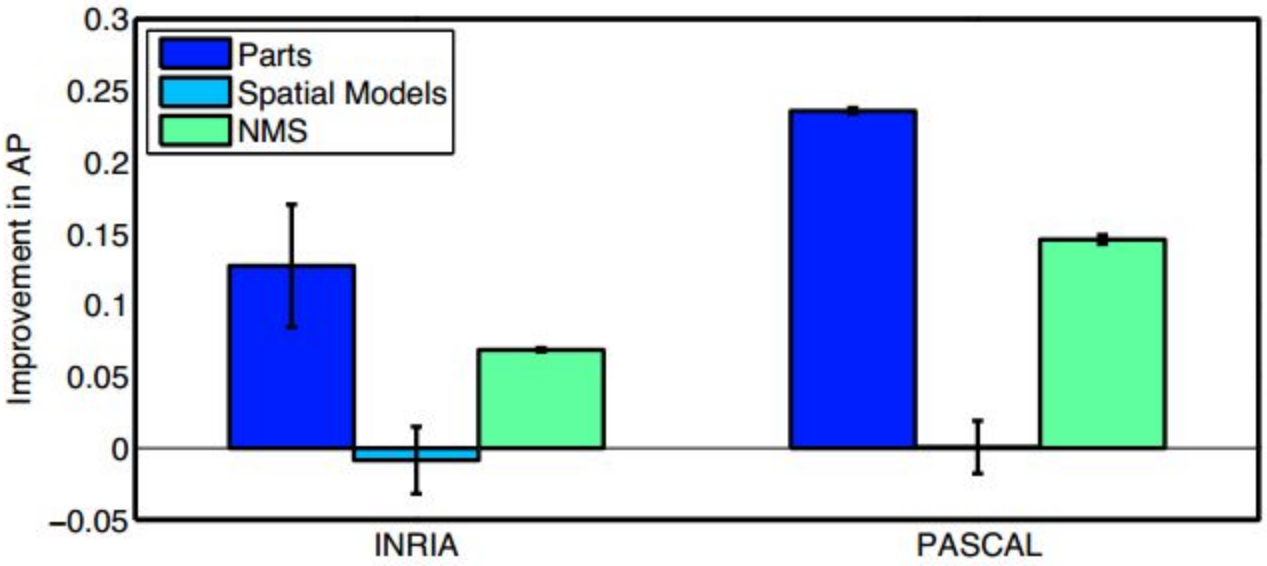
slide: Girshick

Pedestrian detection with ConvNets (video)



What are the weakest links in detection?

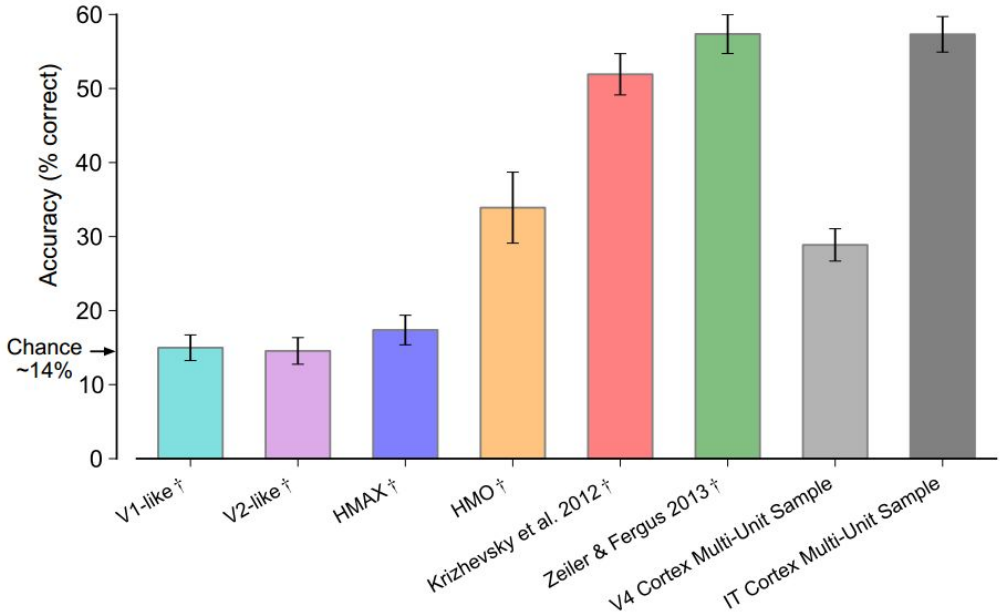
- [Parikh & Zitnick CVPR'10] **replaced each component with humans** (Amazon Turk) and show which ones gain the most:
  - **features**
  - **NMS**



Finding the weakest link in person detectors. Parikh, Devi, and C. Lawrence Zitnick. *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011.

ConvNets vs Primates

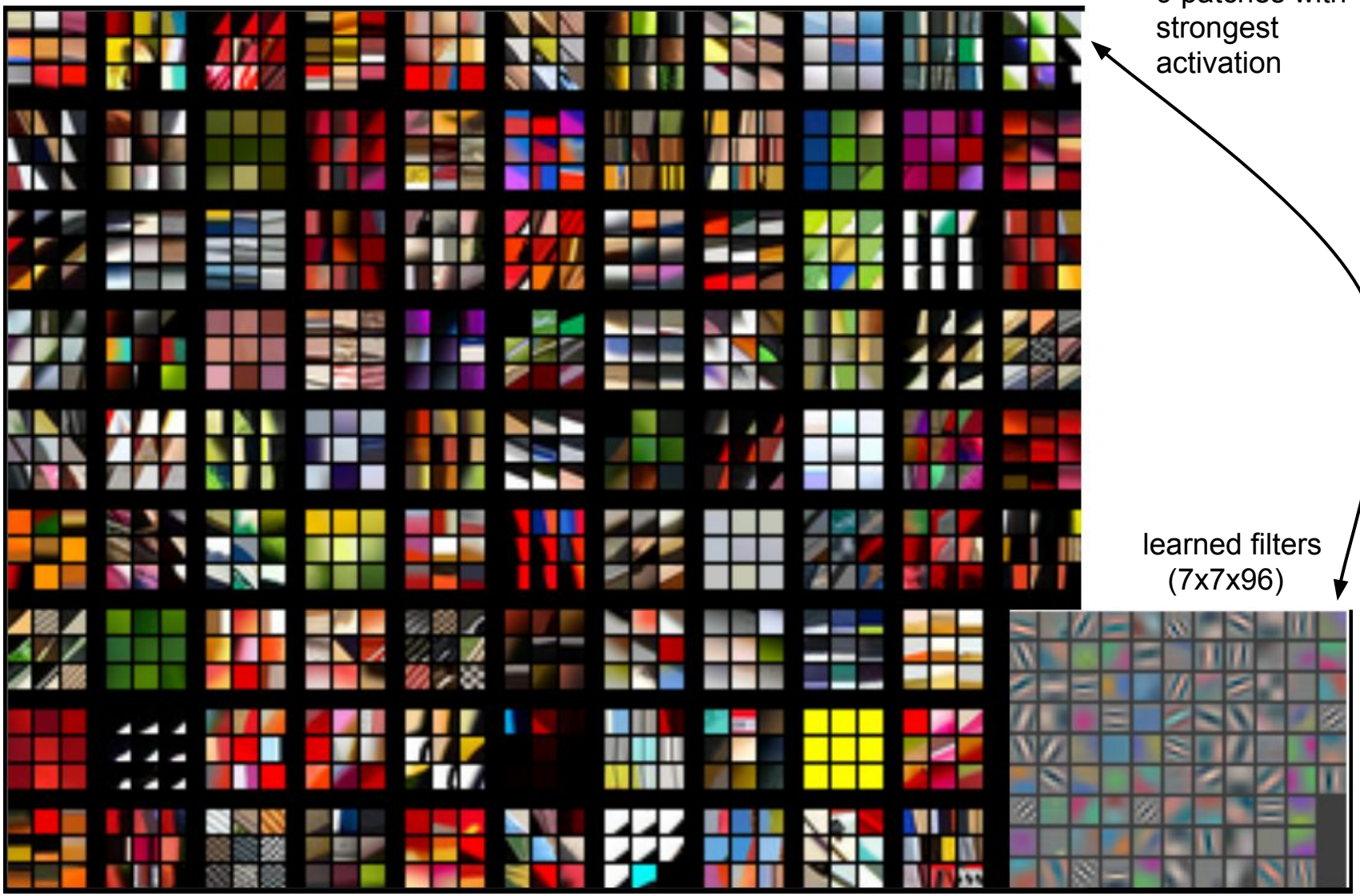
- **[Cadieu'14] ConvNets rival primates "core visual object recognition"**
  - rapid (100ms) feed-forward processing



- **Feed-forward part ~solved?**
- Remains:
  - lateral and feedback loops
  - better spatial exploration for classification/localization/detection

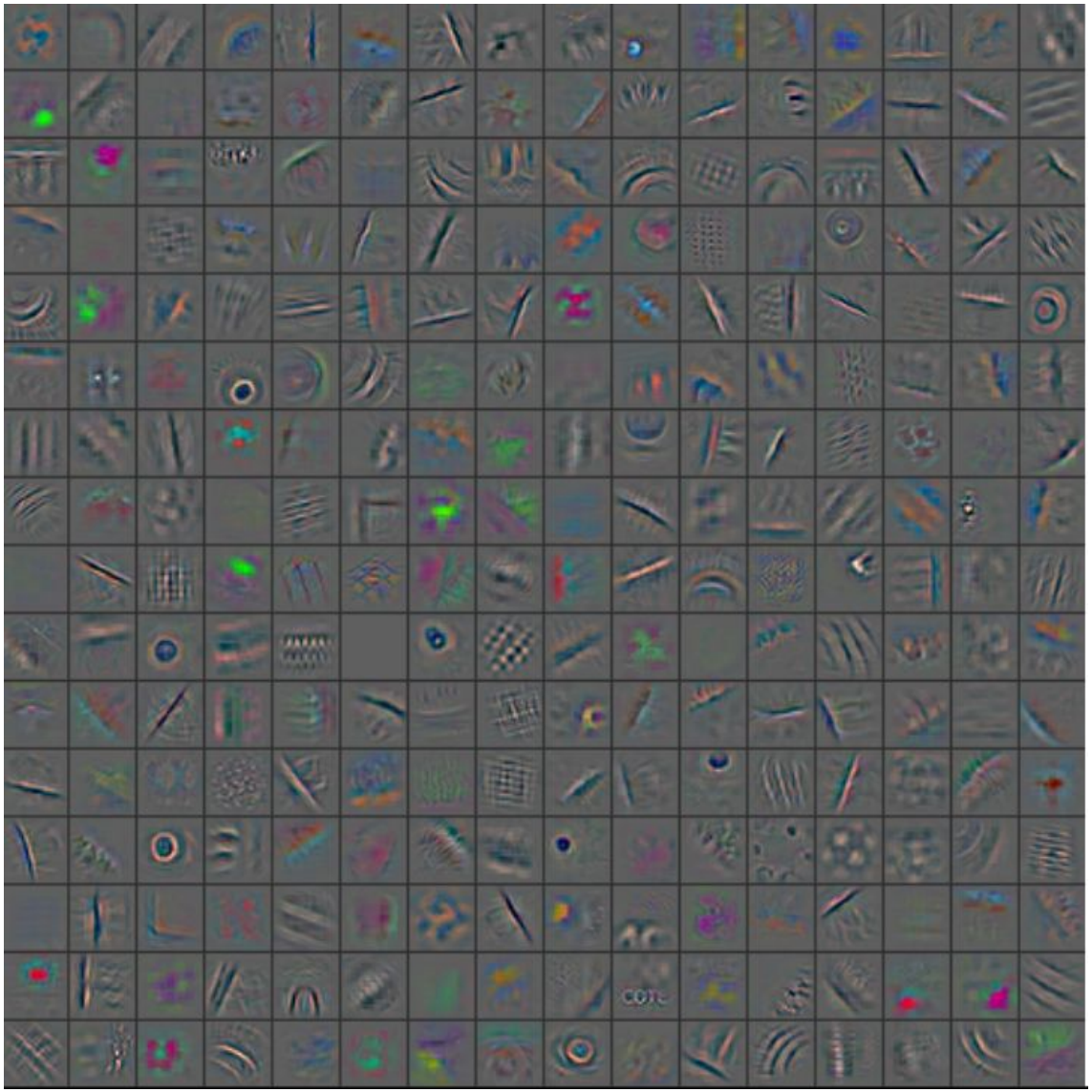
Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. Cadieu, Charles F., Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. *arXiv preprint arXiv:1406.3284* (2014).

Learned convolutional filters: Stage 1



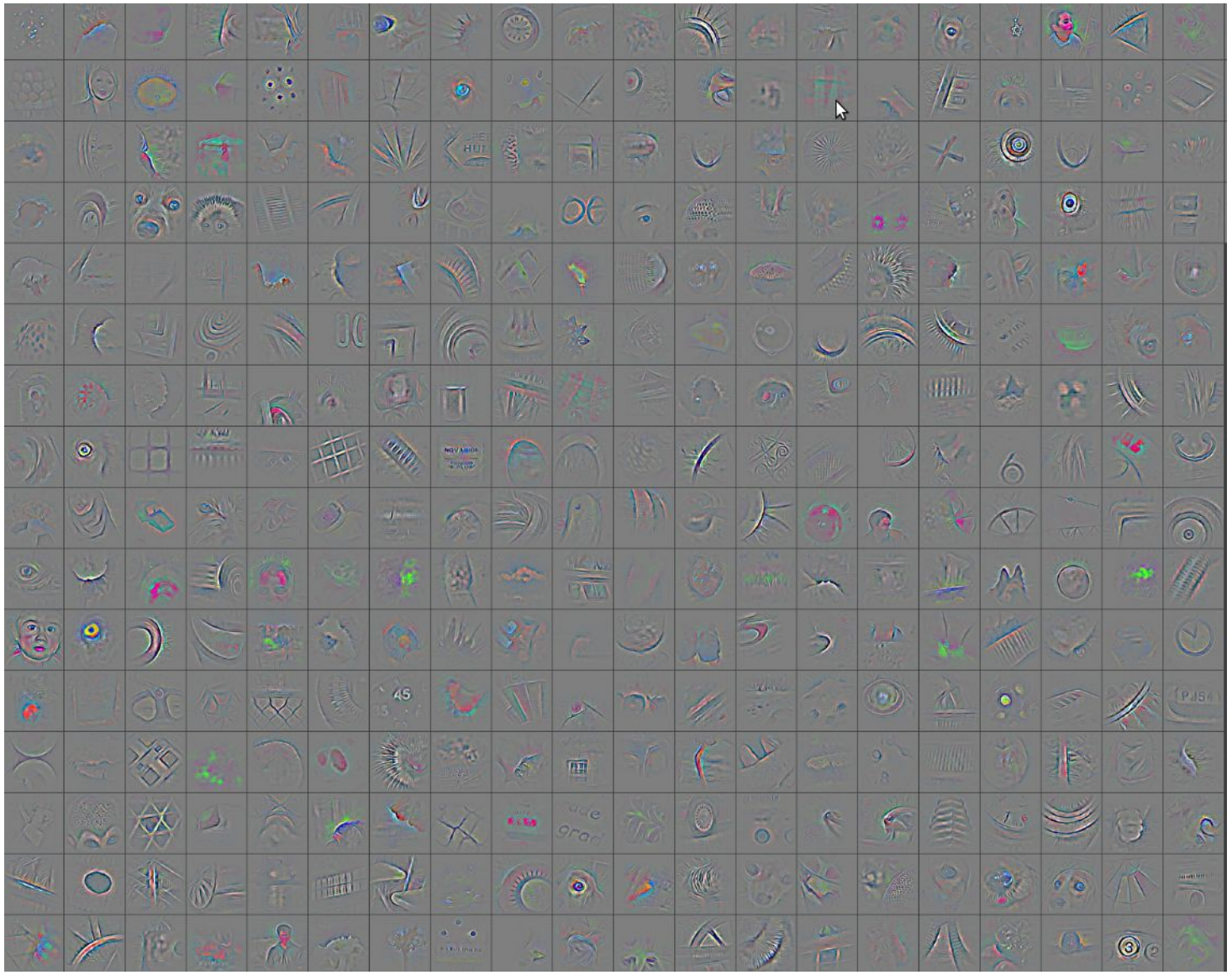
Visualizing and understanding convolutional neural networks. Zeiler, Matthew D., and Rob Fergus. *arXiv preprint arXiv:1311.2901* (2013).

Strongest activations: Stage 2



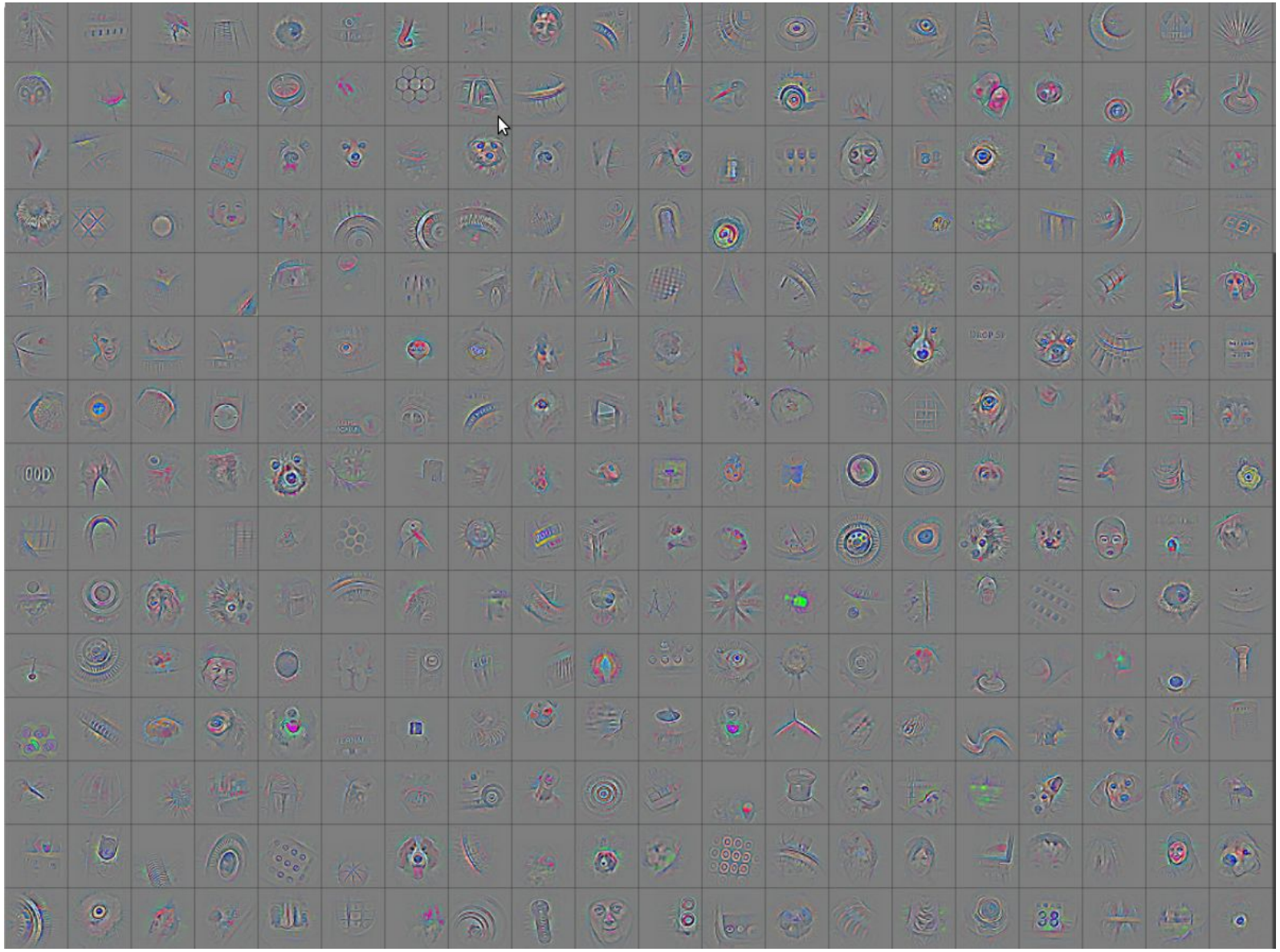
Visualizing and understanding convolutional neural networks. Zeiler, Matthew D., and Rob Fergus. *arXiv preprint arXiv:1311.2901* (2013).

Strongest activations: Stage 3



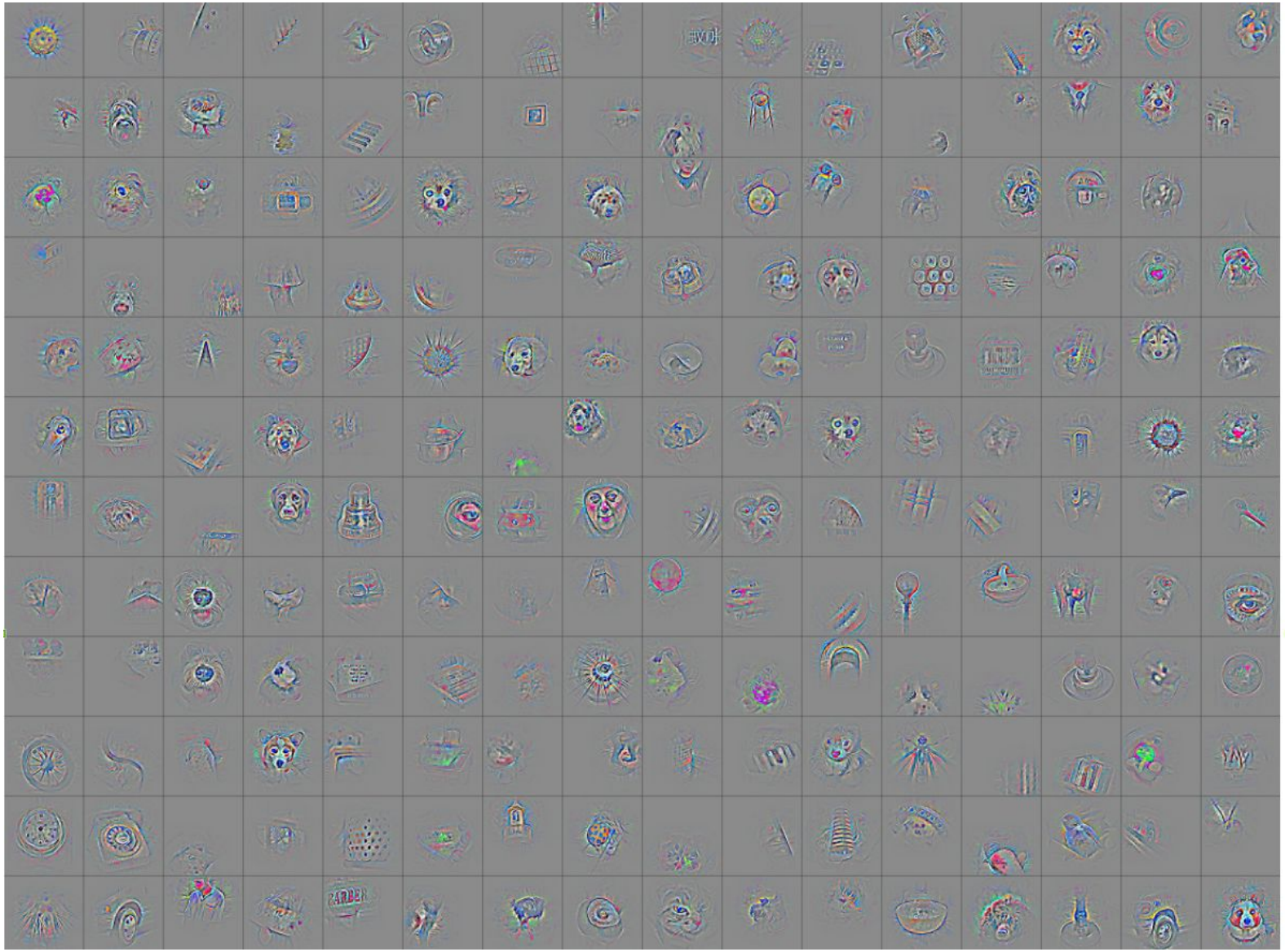
Visualizing and understanding convolutional neural networks. Zeiler, Matthew D., and Rob Fergus. *arXiv preprint arXiv:1311.2901* (2013).

Strongest activations: Stage 4



Visualizing and understanding convolutional neural networks. Zeiler, Matthew D., and Rob Fergus. *arXiv preprint arXiv:1311.2901* (2013).

Strongest activations: Stage 5



Visualizing and understanding convolutional neural networks. Zeiler, Matthew D., and Rob Fergus. *arXiv preprint arXiv:1311.2901* (2013).

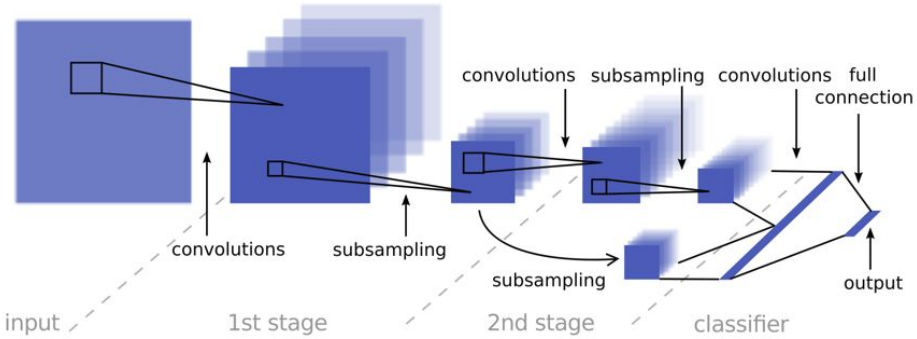
What is a ConvNet?

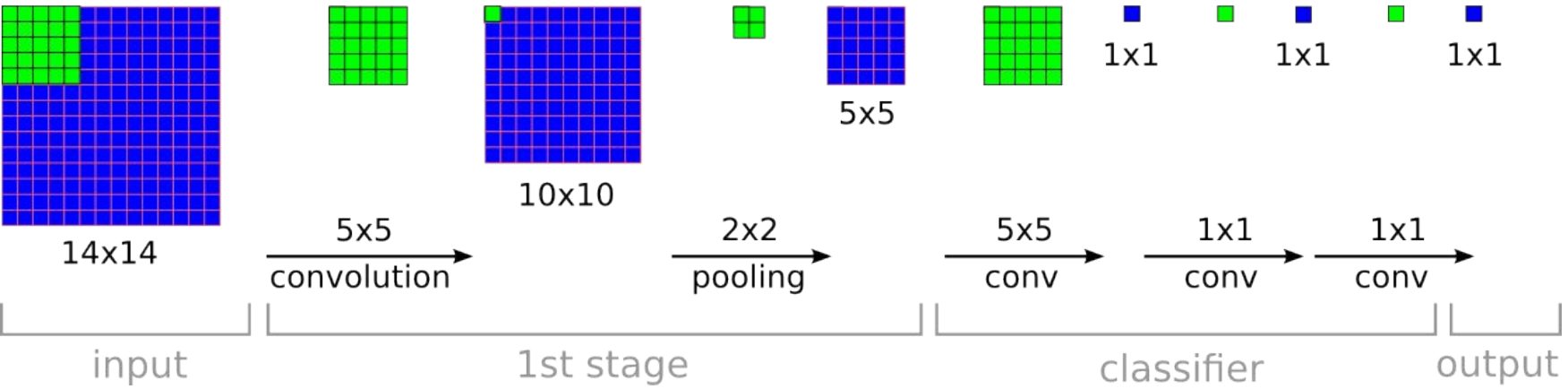
- A special type of Neural Net that incorporates **priors about continuous signals**
  - sound / speech (2D signal)
  - images (3D signal)
  - videos (4D signal)

- **Parameters sharing** and **pooling** take advantage of local coherence to learn invariant features

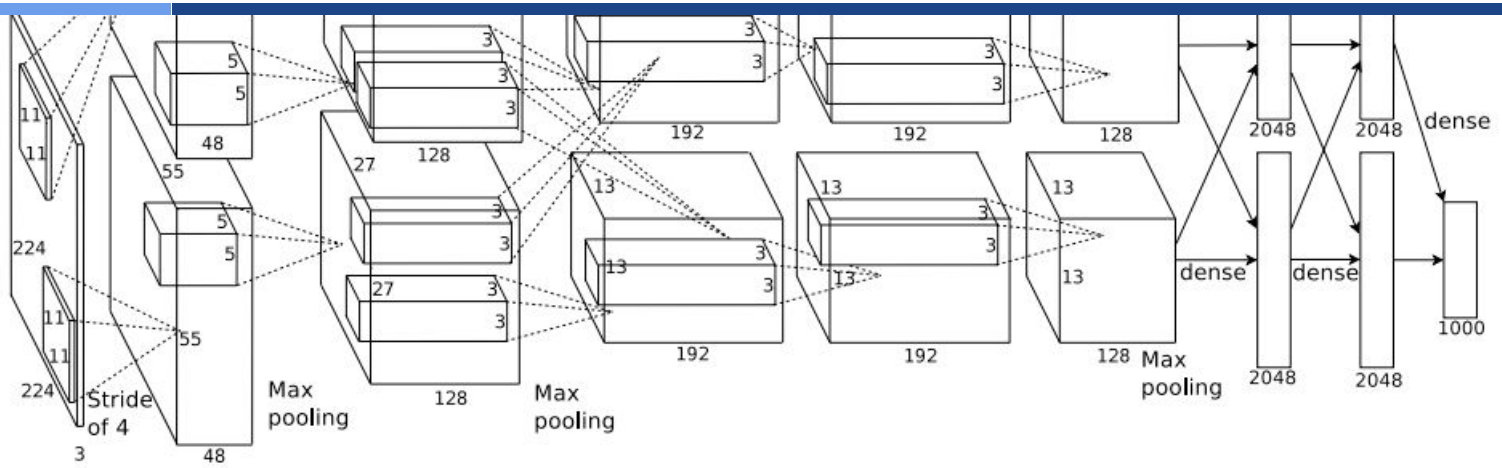
- In its **simplest form**, a ConvNet is just a series of stages of the form:
  - convolution
  - bias
  - non-linearity (ReLu or sigmoid functions)
  - pooling

- Normalization layers (LCN) may be added





ConvNets 2.0



- [Krizhevsky'12] win 2012 ImageNet classification with a **much bigger ConvNet** than before:
  - **deeper**: 7 stages vs 3 before
  - **larger**: 60 million parameters vs 1 million before

- This was made possible by:
  - **fast hardware**: GPU-optimized code
  - **big dataset**: 1.2 million images vs thousands before
  - **better regularization**: dropout

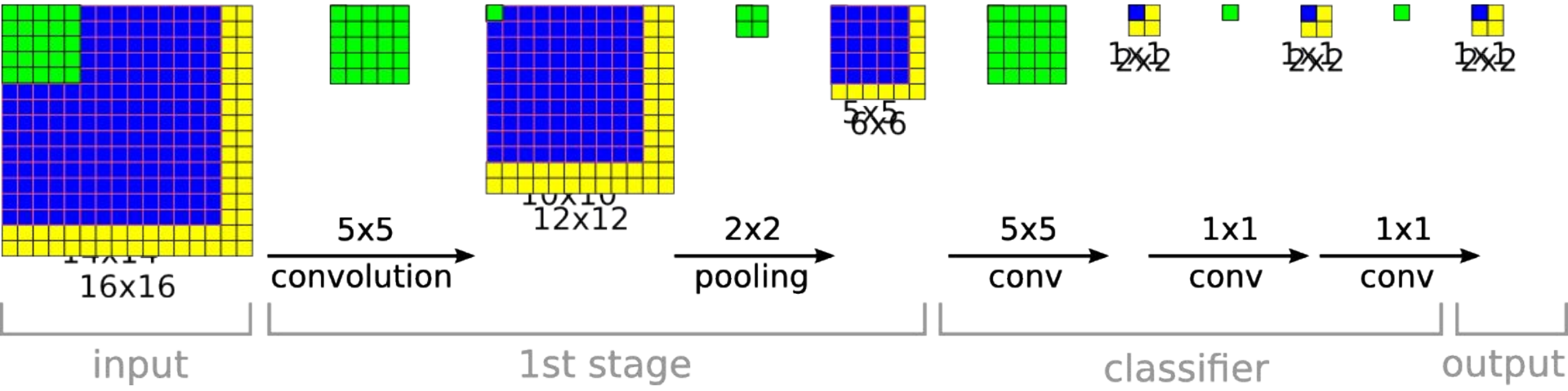
Imagenet classification with deep convolutional neural networks. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. *Advances in neural information processing systems*. 2012.

Why are ConvNets good for detection?

- **Sharing parameters** is good
  - taking advantage of local coherence to learn a more efficient representation:
    - no redundancy
    - translation invariance
    - slight rotation invariance with pooling

- **Efficient for detection:**
  - all computations are shared
  - can handle varying input sizes (no need to relearn weights for new sizes)

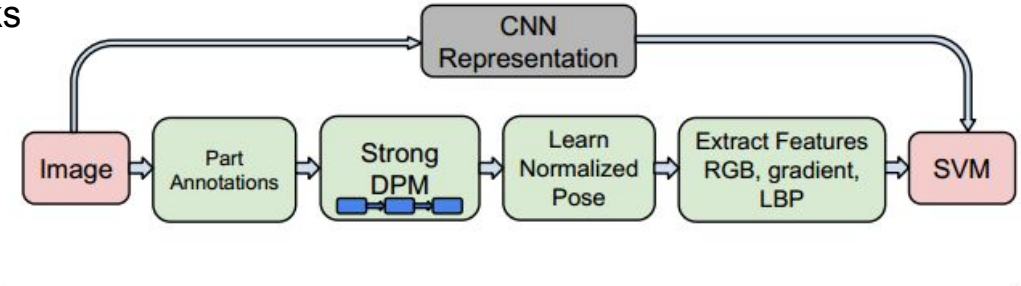
- **ConvNets are convolutional all the way up** including fully connected layers

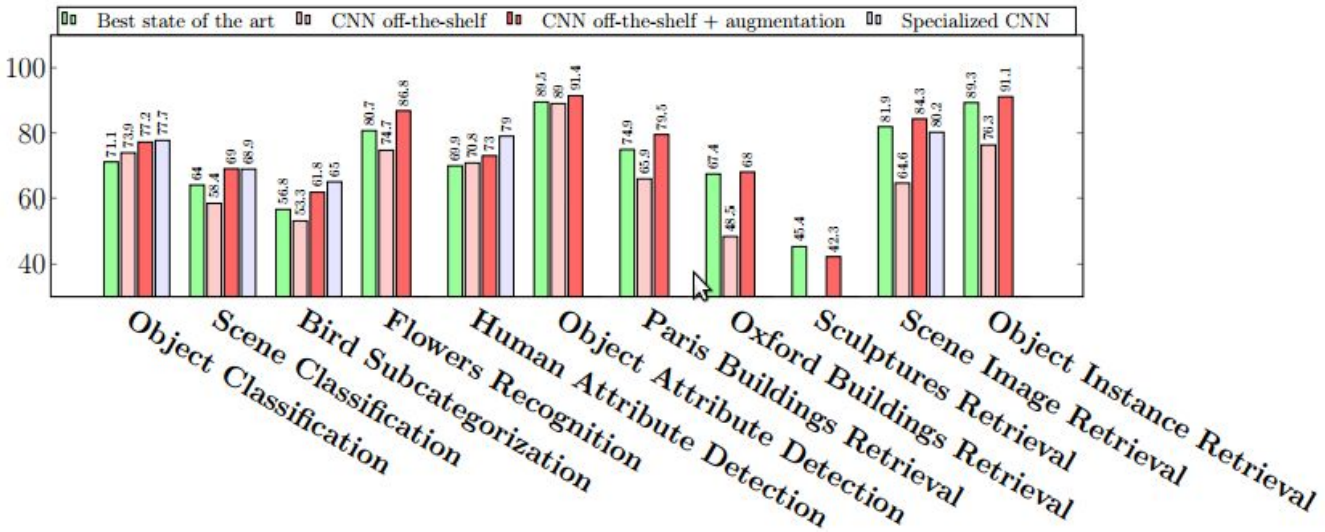


ImageNet pre-training

- **Labeled data is rare for detection:** leverage large classification labeled datasets for pre-training.

- **ImageNet Classification pretraining + fine-tuning on a different task** has been shown to work very well by many people.
  - in particular [Razavian'14] took the off-the-shelf convnet **OverFeat + SVM classifier** on top and obtained many state-of-the-art or competitive results on 10+ datasets and visual tasks





CNN Features off-the-shelf: an Astounding Baseline for Recognition. Razavian, Ali Sharif, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. *arXiv preprint arXiv:1403.6382* (2014).

ImageNet pre-training

- **Capacity must match the problem at hand**
  - 60M-parameters model has a capacity designed for ImageNet-scale data
  - one cannot train such model on a small dataset: e.g. 6k bird dataset in [Branson'14]
  - ImageNet pre-training will ensure **general features as a starting point**

- **Fine-tuning**
  - requires **lowering the learning rate** to avoid forgetting pre-training or use **different learning rates for the pre-trained and new layers**
  - [Branson'14] propose a **2-step fine-tuning method** that improves accuracy
    i. only train the weights of the new layer(s)
    ii. train all weights

Bird Species Categorization Using Pose Normalized Deep Convolutional Nets. Branson, Steve, Grant Van Horn, Serge Belongie, and Pietro Perona. *arXiv preprint arXiv:1406.2952* (2014).

How much does fine-tuning matter?

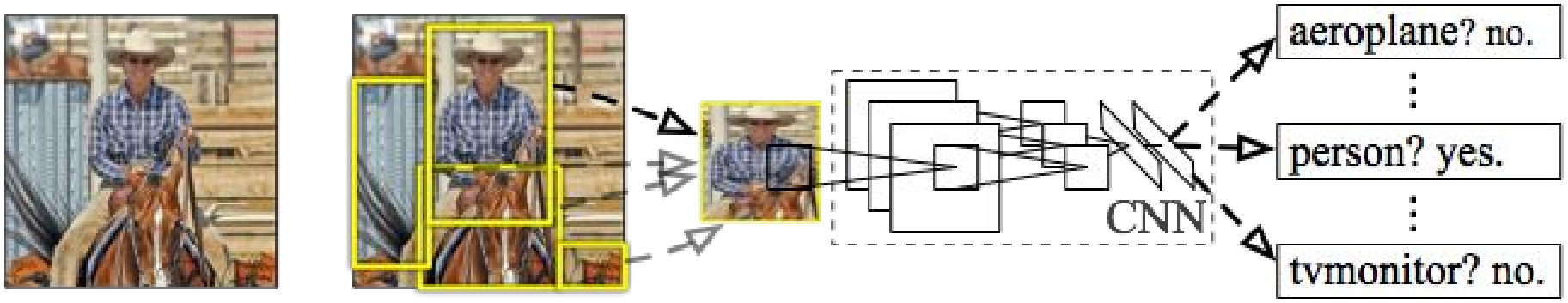
- [Razavian'14] showed consistent gains using fixed ConvNet weights
- **Fine-tuning always improves**, but how much?
- However [Girshick'14] shows **substantial improvements with fine-tuning** on PASCAL detection: 44.7% to 54.2% mAP

	VOC 2007	VOC 2010
Regionlets (Wang et al. 2013)	41.7%	39.7%
SegDPM (Fidler et al. 2013)		40.4%
R-CNN $pool_5$	44.2%	
R-CNN fc_6	46.2%	
R-CNN fc_7	44.7%	
fine-tuned: R-CNN FT $pool_5$	47.3%	
fine-tuned: R-CNN FT fc_6	53.1%	
fine-tuned: R-CNN FT fc_7	**54.2%**	**50.2%**

metric: mean average precision (higher is better)

CNN Features off-the-shelf: an Astounding Baseline for Recognition. Razavian, Ali Sharif, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. *arXiv preprint arXiv:1403.6382* (2014).
Rich feature hierarchies for accurate object detection and semantic segmentation. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *arXiv preprint arXiv:1311.2524* (2013).

R-CNN: Regions with CNN features



- **Proposal-method agnostic, many choices**
  - Selective Search [Uijlings et al. 2013] (Used in this work)
  - Objectness [Alexe et al. 2012]
  - Category independent object proposals [Endres & Hoiem 2010]
- **Active area, at this CVPR**
  - BING [Ming et al.] – *fast*
  - MCG [Arbelaez et al.] – *high-quality segmentation*

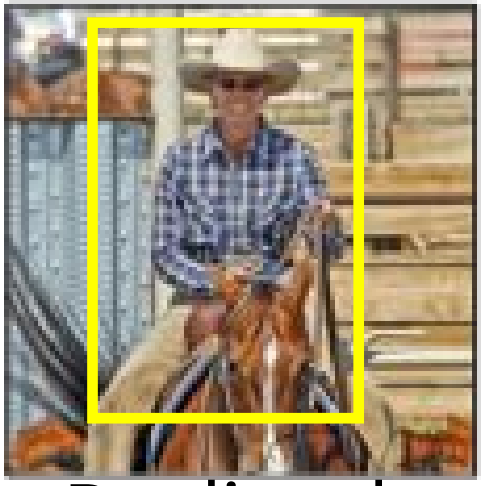
Rich feature hierarchies for accurate object detection and semantic segmentation. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *arXiv preprint arXiv:1311.2524* (2013).

R-CNN: bounding-box regression

Original proposal

Linear regression

on CNN features

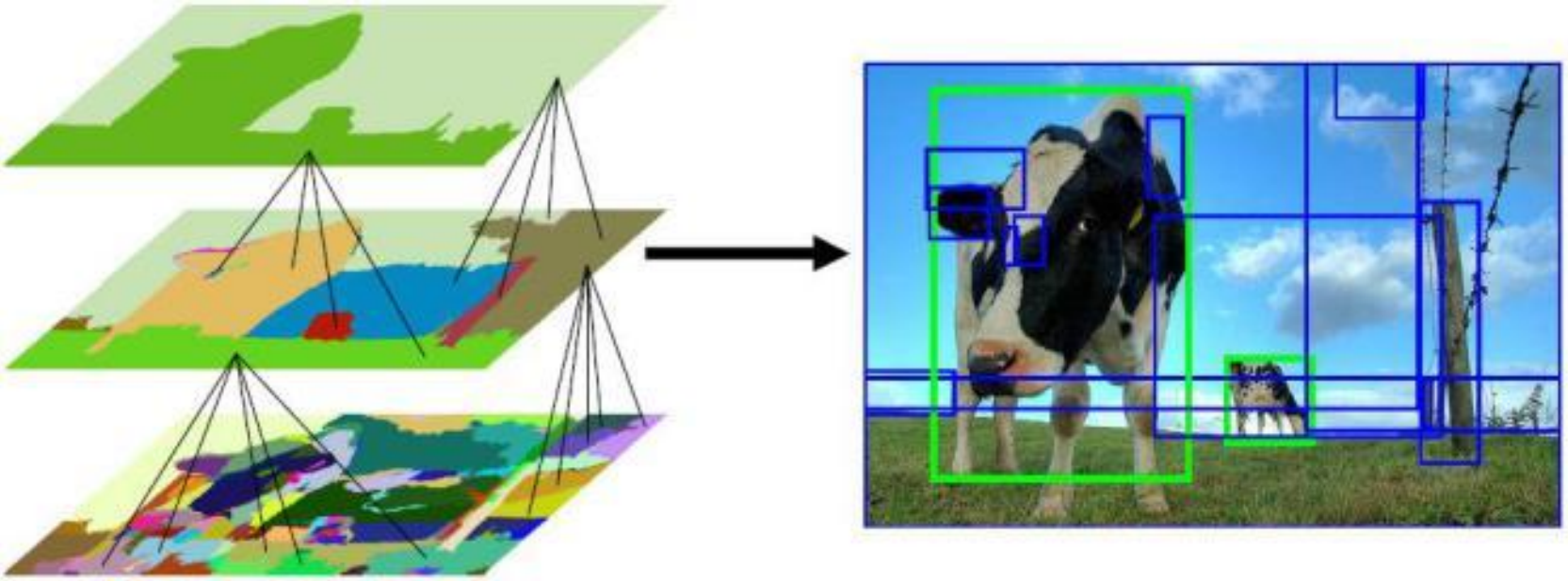


Predicted object bounding box

Rich feature hierarchies for accurate object detection and semantic segmentation. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *arXiv preprint arXiv:1311.2524* (2013).

Objectness / Selective Search

- **fast dense generic detection + slow sparse classification**
  (old idea used for detection of faces, traffic signs, etc)
  - segmentation [van de Sande 2011], ConvNets [Erhan 2014]
  - speeds up
  - reduces false positives
  - bounding boxes with any aspect ratio

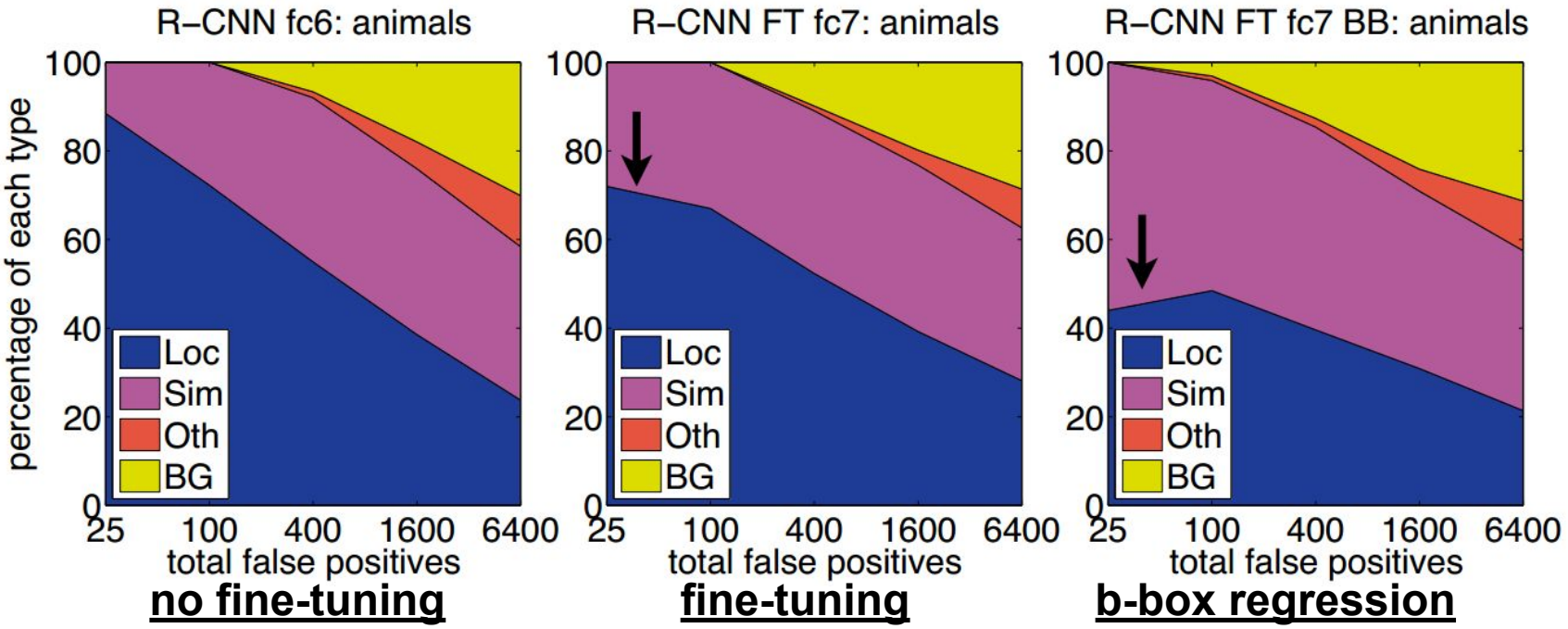


van de Sande, K. E., Uijlings, J. R., Gevers, T., & Smeulders, A. W. (2011, November). **Segmentation as selective search for object recognition**. ICCV'11

Dumitru Erhan, Christian Szegedy, Alexander Toshev, Dragomir Anguelov. **Scalable Object Detection using Deep Neural Networks**. CVPR'14

How to debug object detection?

- [Hoiem'12] provide **analysis software for detection**
  - http://www.cs.illinois.edu/~dhoiem/projects/properties_project.html
- **Visualize the proportions of**
  - localization errors
  - confusion with similar classes
  - confusion with other classes
  - confusion with background
- [Girshick'14] show how **fine-tuning and bbox-regression decrease localization errors**

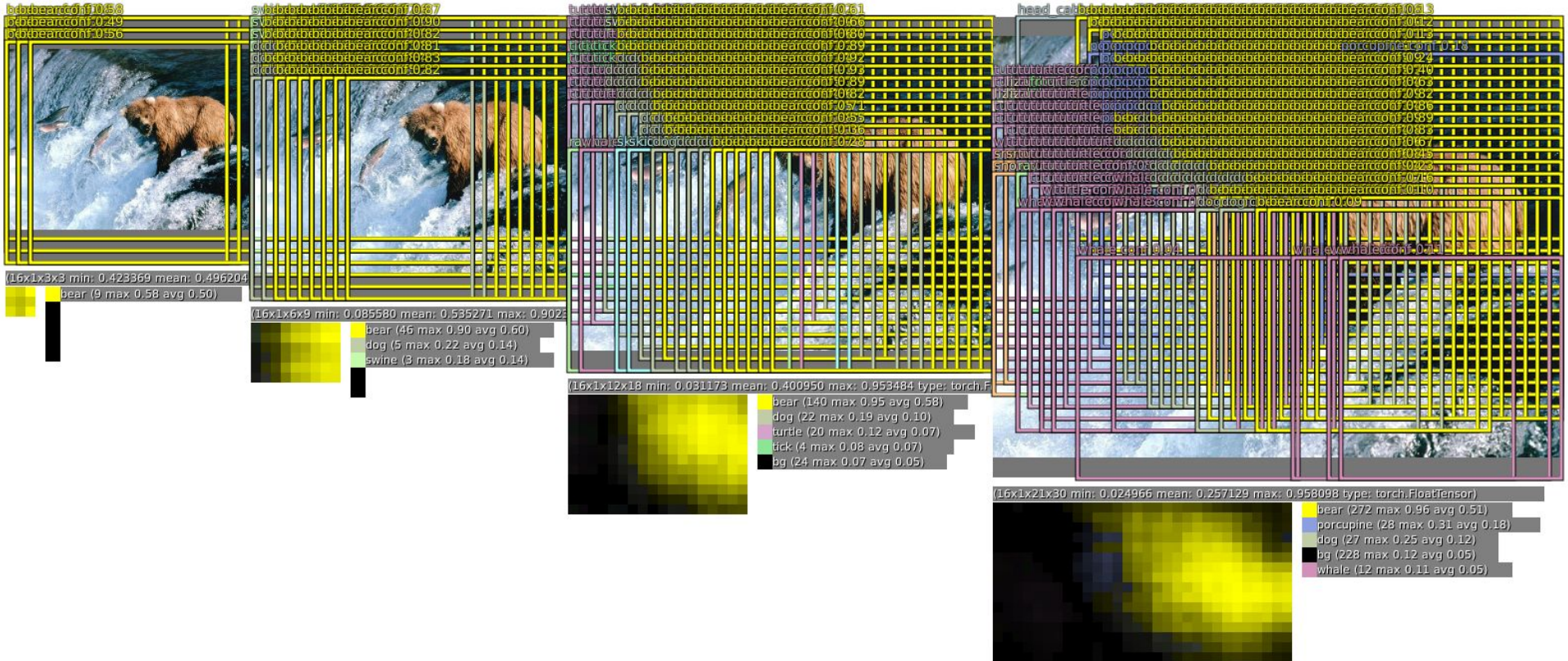


Diagnosing error in object detectors. Hoiem, Derek, Yodsawalai Chodpathumwan, and Qieyun Dai. *Computer Vision–ECCV 2012*. Springer Berlin Heidelberg, 2012. 340-353.

Rich feature hierarchies for accurate object detection and semantic segmentation. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *arXiv preprint arXiv:1311.2524* (2013).

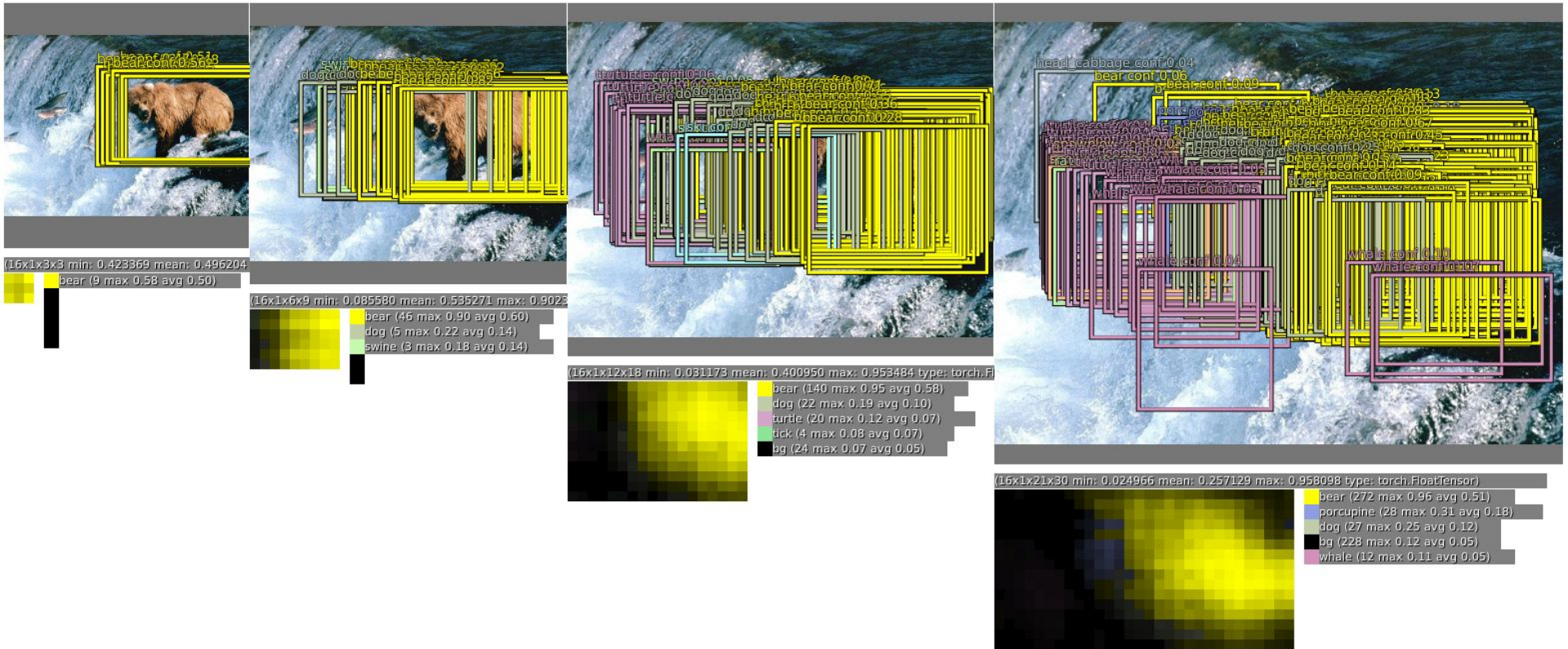
OverFeat: dense detection

- **2-headed network:** shared feature extraction part + prediction for:
  - **bounding box** (trained by regression)
  - **class** (trained with softmax classifier)
- **sliding window**: dense estimation of (bbox,class) at each location
- **accumulate** rather than suppress
- another form of **voting**



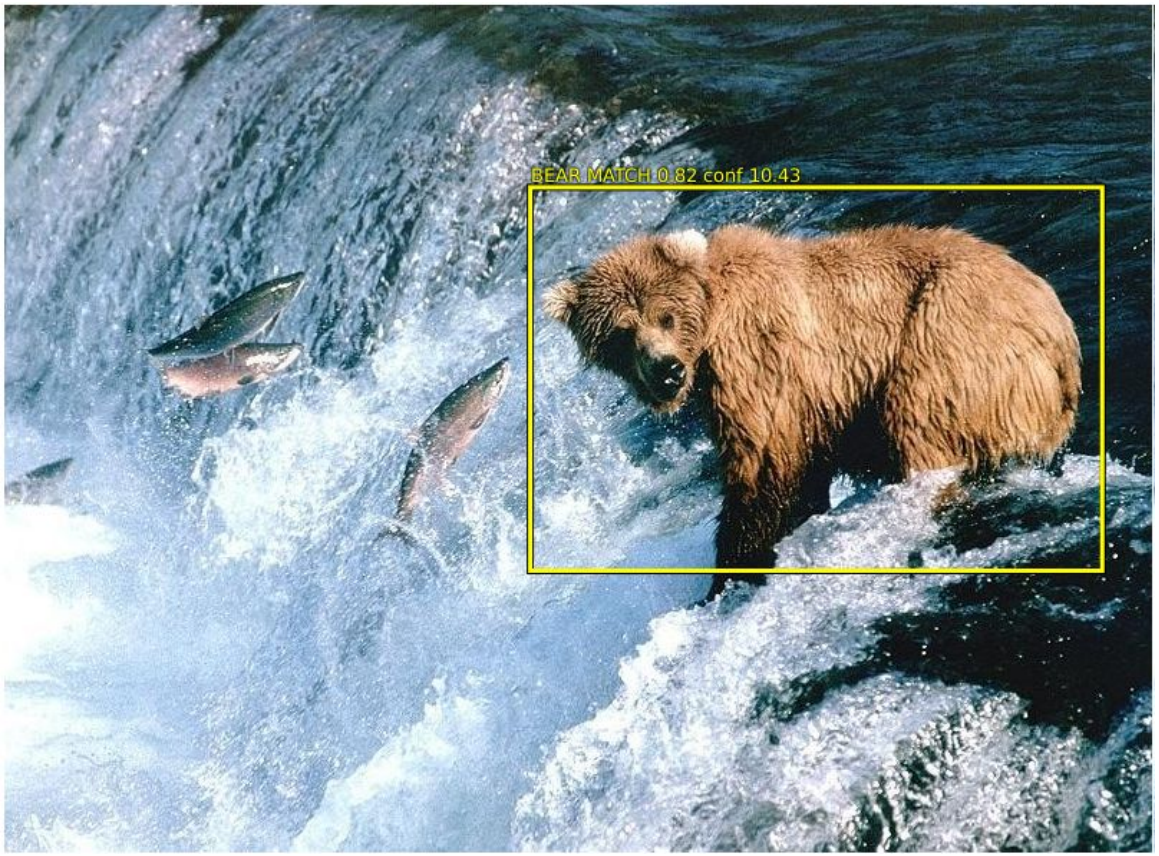
OverFeat: dense detection

- **2-headed network:** shared feature extraction part + prediction for:
  - **bounding box** (trained by regression)
  - **class** (trained with softmax classifier)
- **sliding window**: dense estimation of (bbox,class) at each location
- **accumulate** rather than suppress
- another form of **voting**



OverFeat: dense detection

- **Bounding boxes voting**:
  - **voting is good** (classification: views voting + model voting)
  - boosts confidence **high above false positives** ([0,1] up to 10.43 here)
  - more robust to individual localization errors
  - relying less on an accurate background class



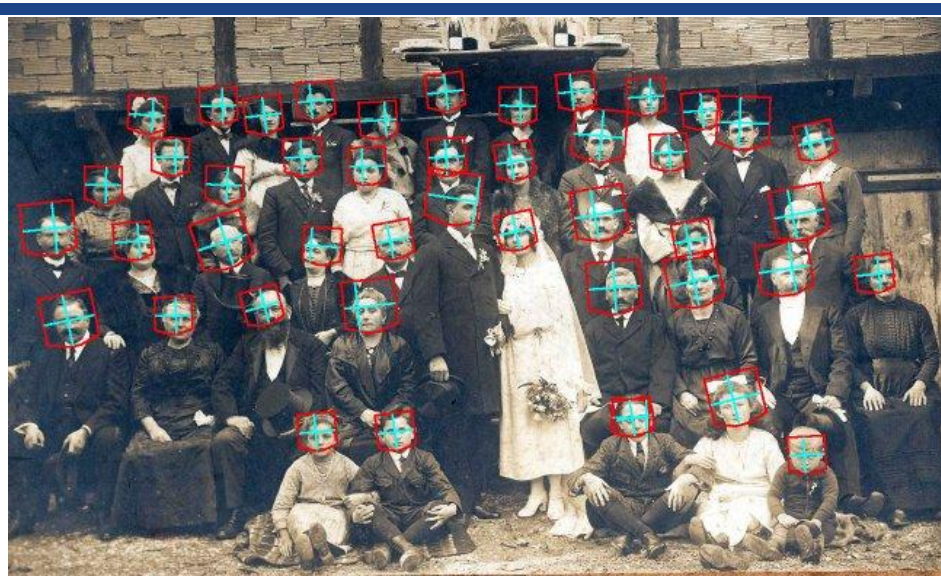
Localization regression

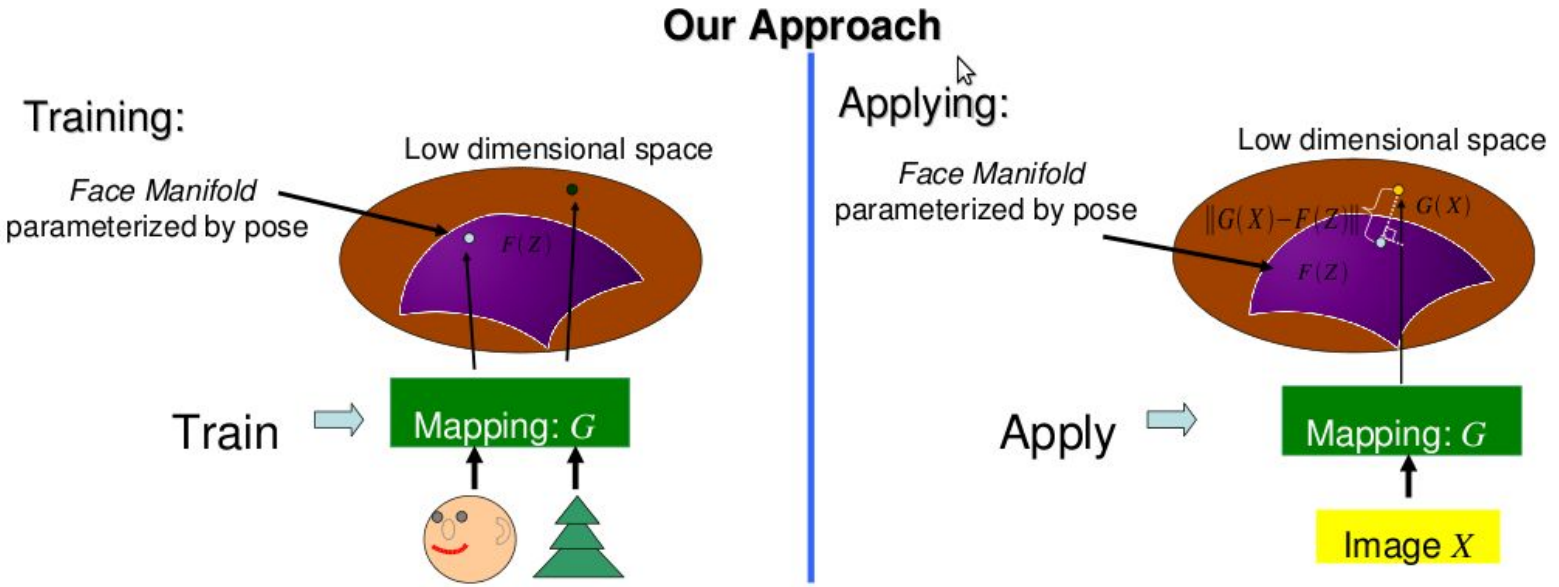
- **Pedestrian localization -> detection (Sermanet 2011)**
  - devised to emphasize separation with false positives

Pose regression with Convnet

[Osadchy'04]:

- estimating presence and pose
- distance away from manifold indicates confidence into presence of a face
- position on the manifold indicates face pose





Osadchy, Margarita, Yann Le Cun, and Matthew L. Miller. **Synergistic face detection and pose estimation with energy-based models.** *The Journal of Machine Learning Research* 8 (2007): 1197-1215.

Detection: localization voting for multiple objects

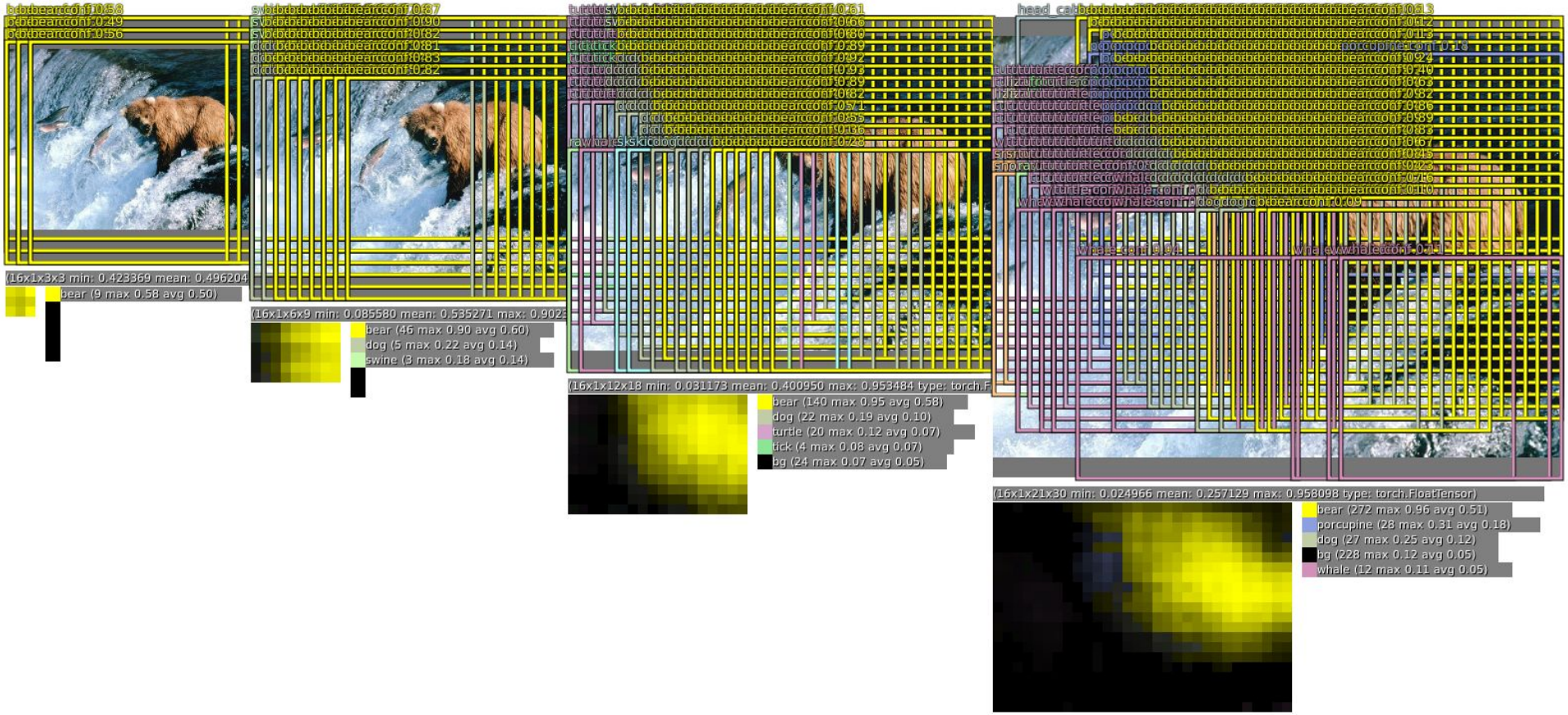
Augmenting sliding density of a ConvNet

- the more subsampling, the larger the output stride
- larger output stride means less views

- e.g.: subsampling x2, x3, x2, x3 => 36 pixels stride
- 1 pixel shift in output space corresponds to 36 pixels shift in input space

Augmenting sliding density of a ConvNet

- 9x more bounding boxes (with last pooling 3x3)
- idea introduced by [Giusti'13]

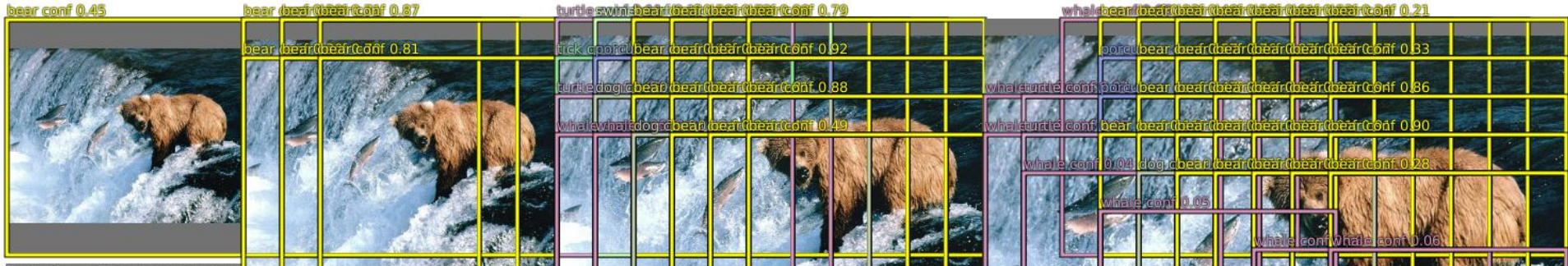


A. Giusti, D. C. Ciresan, J. Masci, L. M. Gambardella, and J. Schmidhuber. **Fast image scanning with deep max-pooling convolutional neural networks**. In International Conference on Image Processing (ICIP), 2013.

Online bootstrapping

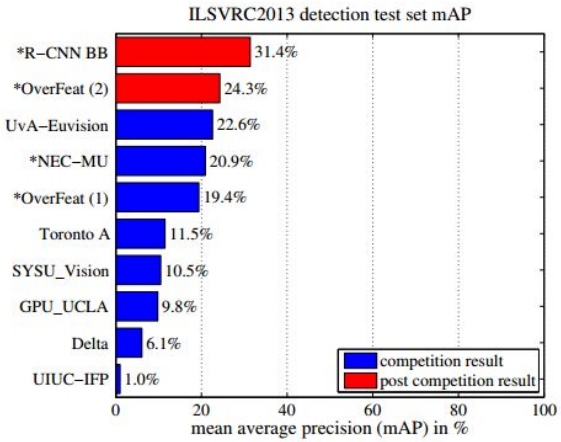
- **traditional offline bootstrapping:**
  - complicated and not very practical for large datasets
  - potential mismatch between false positives extraction and retraining
  - tuning of bootstrapping passes

- **online bootstrapping:**
  - constantly adapting
  - no storage required
  - backprop from multiple outputs per image, e.g.:
    - 1 worst false positive
    - 1 worst false negative
    - 1 random negative
    - 1 random positive
    - (backprop all outputs will bias training too much for each image)
  - slower to train entire image rather than single windows, but training time is fast enough starting from classification features



How do R-CNN and OverFeat differ?

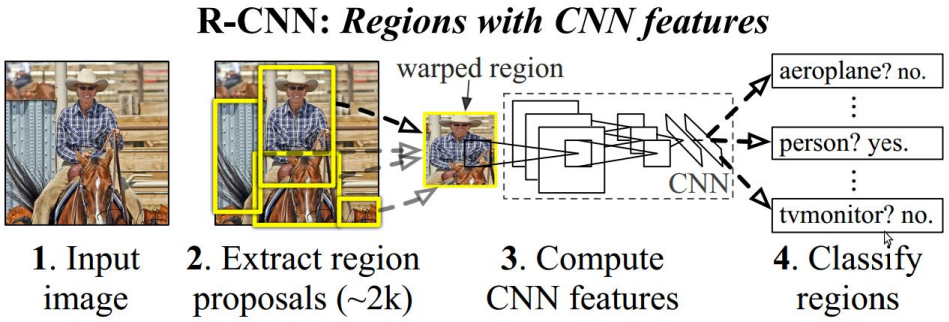
- **SVM classifier** instead of softmax: ~4 points increase in mAP on PASCAL
- **Warping:** ~3-5 points increase
- **Training on ImageNet validation + subset of training set** (R-CNN) while OverFeat is trained on training data only. Validation/Test are drawn from a different distribution than Training set, so including validation during training increases accuracy.
- **Sparse evaluation** (R-CNN) as opposed to dense evaluation (OverFeat) increases chances of false positives, separating "objectness" prediction from classification and localization probably reduces false positives.



Rich feature hierarchies for accurate object detection and semantic segmentation. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *arXiv preprint arXiv:1311.2524* (2013).
Overfeat: Integrated recognition, localization and detection using convolutional networks. Sermanet, Pierre, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. *arXiv preprint arXiv:1312.6229* (2013), International Conference on Learning Representations (ICLR)` 2014.

Warping

- ConvNets expect a **fixed size input**
- **Warping improves detection** by about 3-5 mAP points in [Girshick'13]
  - **likely simplifies the learning problem by normalizing poses and shapes**
- Overall better, but probably poses problems for some classes, e.g. cars/trucks or circle/elipses



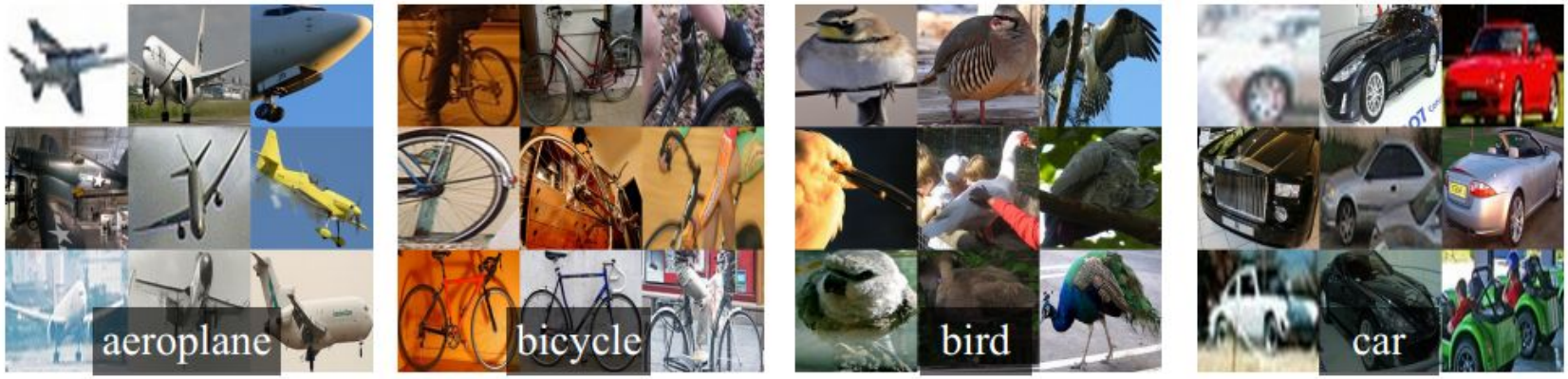
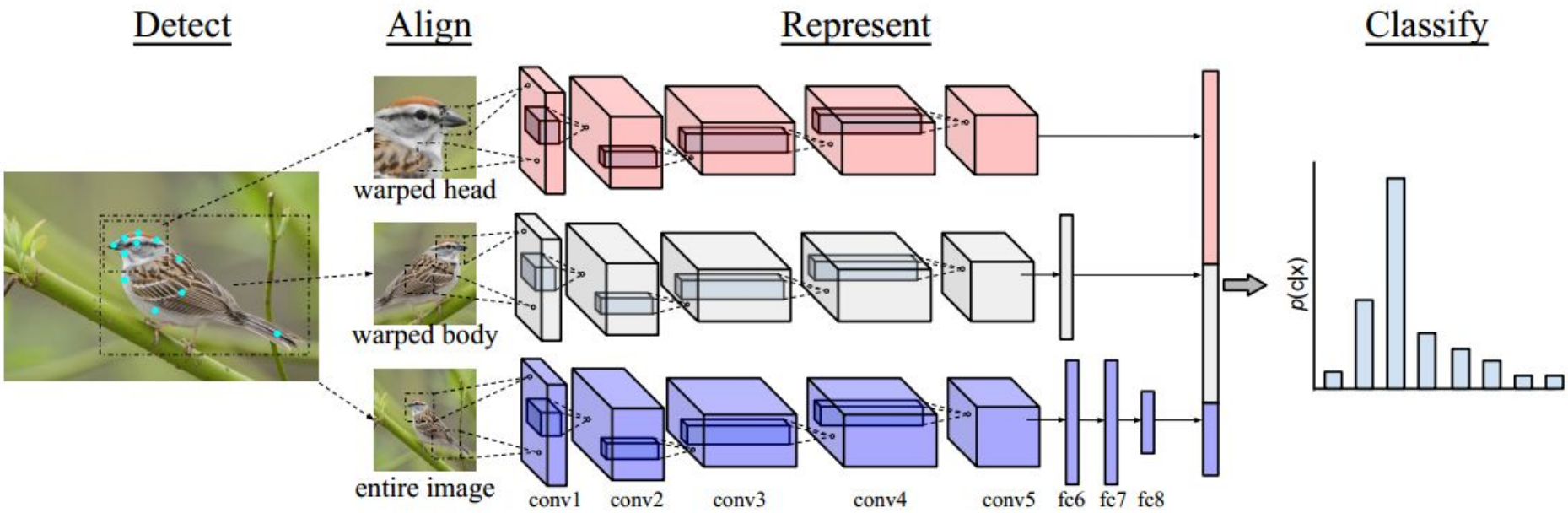


Figure 2: Warped training samples from VOC 2007 train.

Rich feature hierarchies for accurate object detection and semantic segmentation. Girshick, Ross, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *arXiv preprint arXiv:1311.2524* (2013).

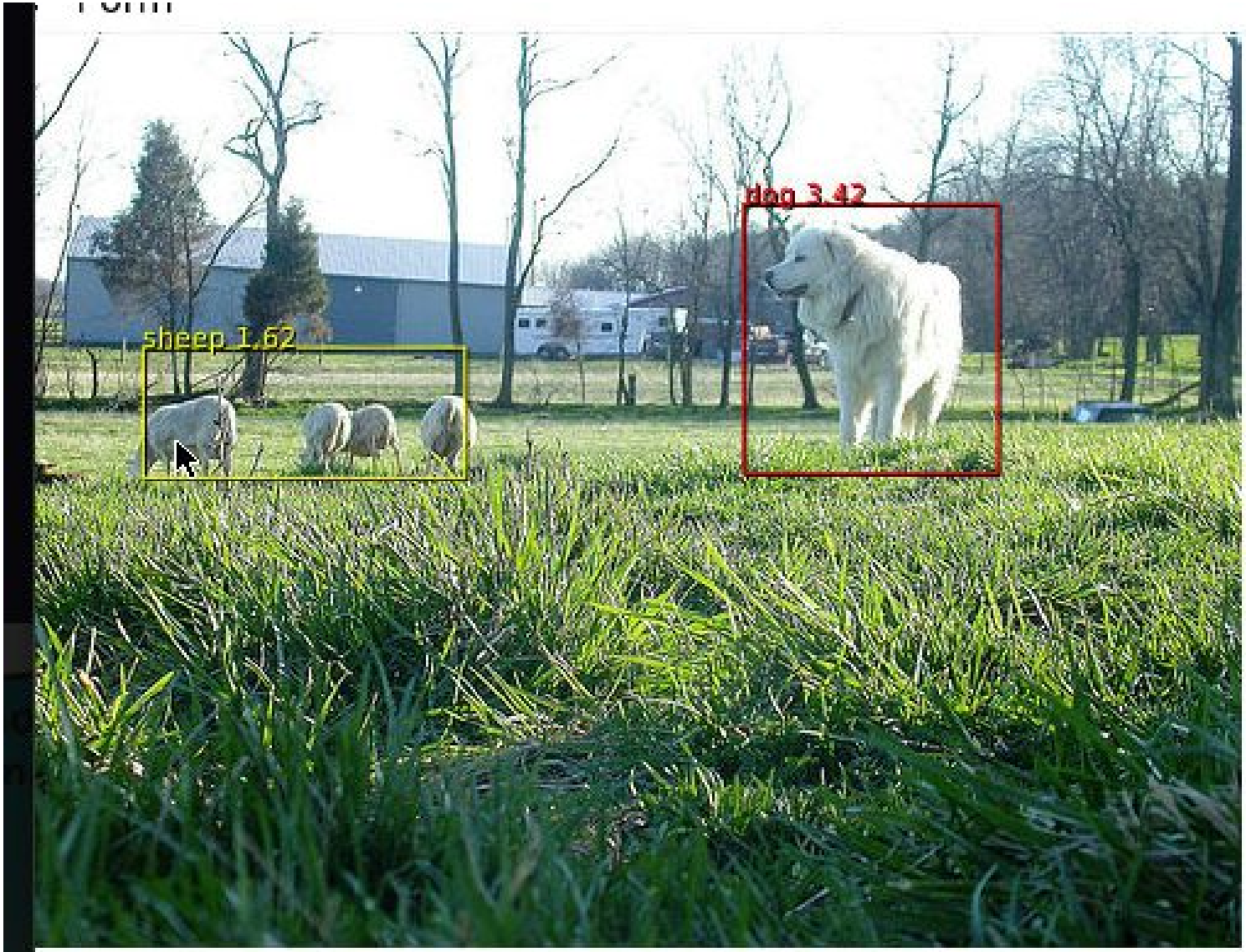
Warping

- Warping improves results for fine-grained recognition

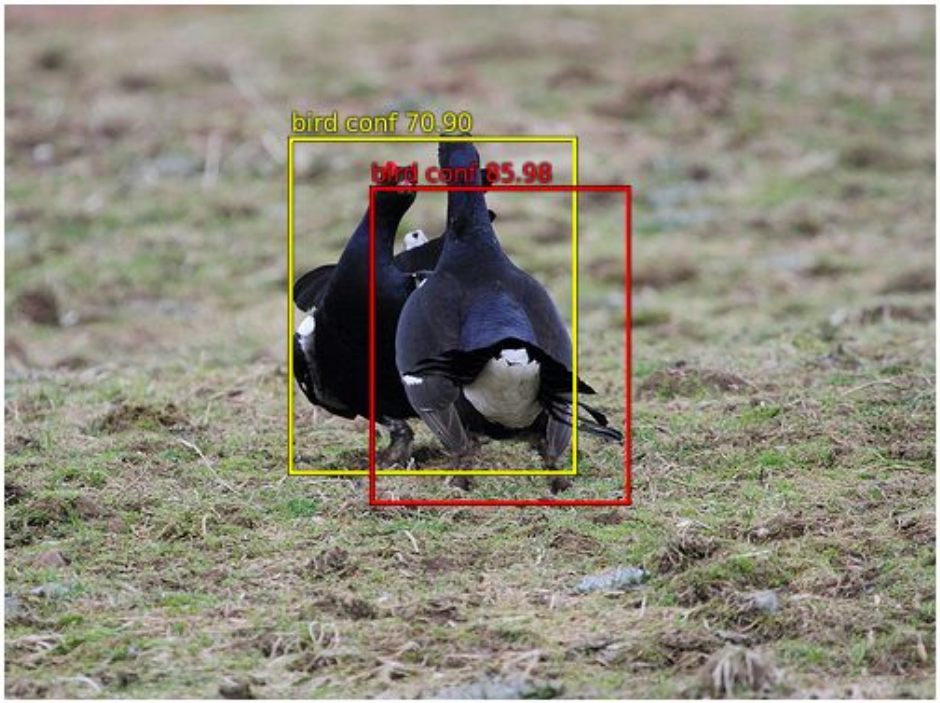


Bird Species Categorization Using Pose Normalized Deep Convolutional Nets. Branson, Steve, Grant Van Horn, Serge Belongie, and Pietro Perona. *arXiv preprint arXiv:1406.2952* (2014).

ImageNet detection examples



ImageNet detection examples: occlusions

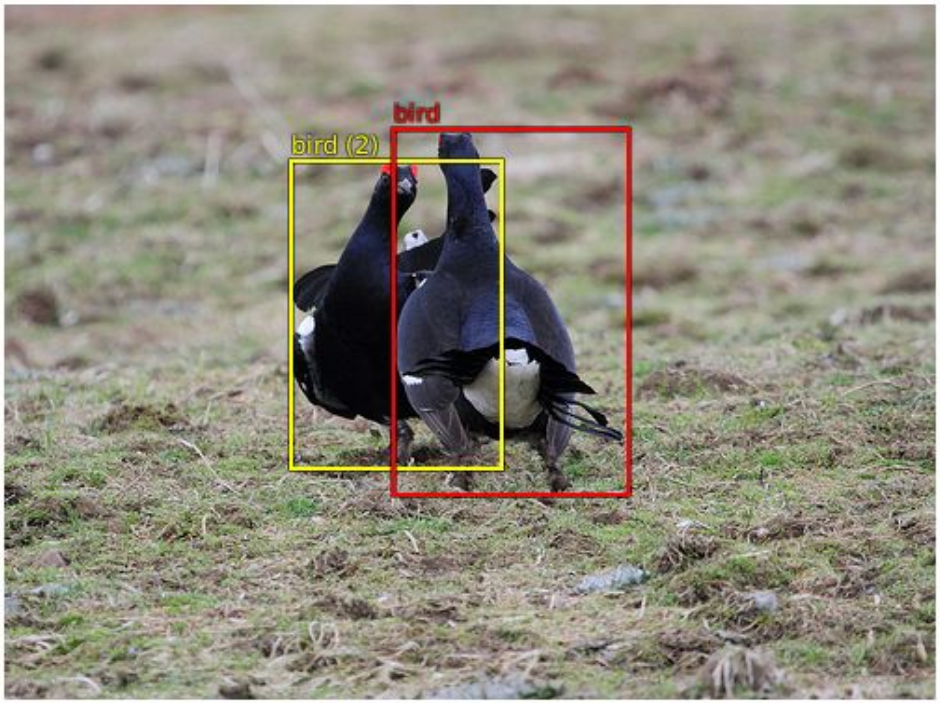


Top predictions:

bird (confidence 86.0)

bird (confidence 70.9)

ILSVRC2012_val_00001136.JPEG

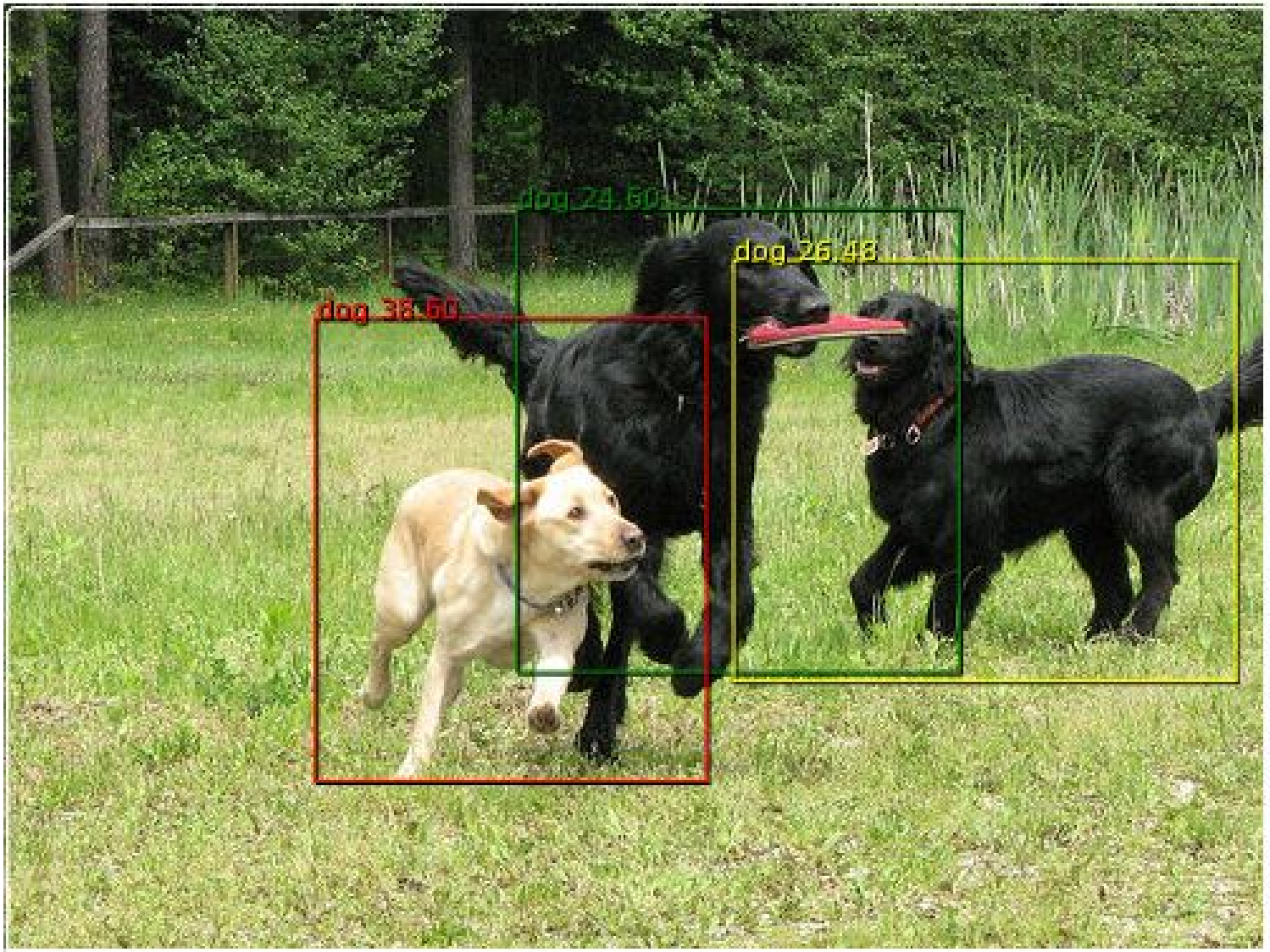


Groundtruth:

bird

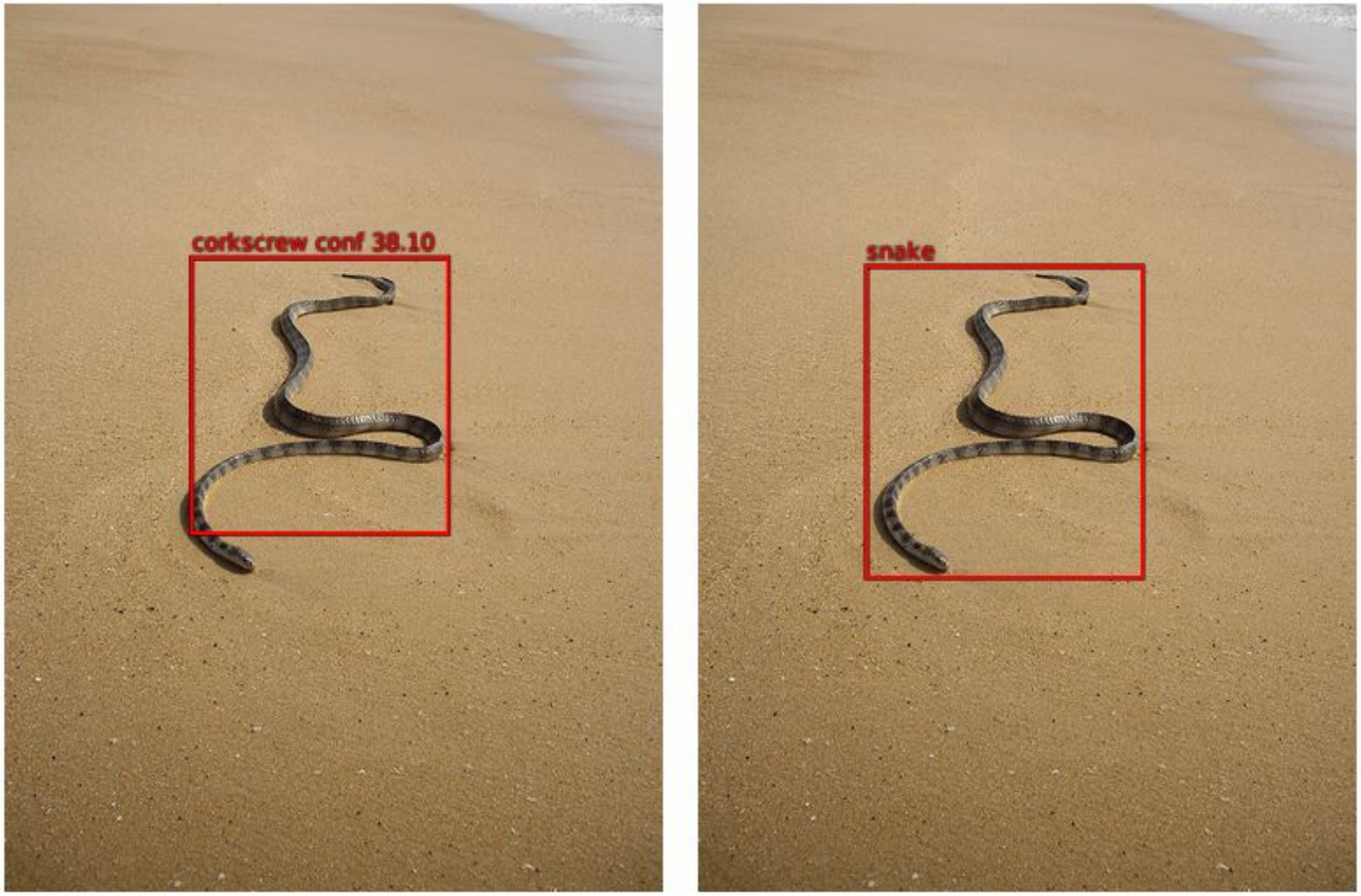
bird (2)

ImageNet detection examples: occlusions



ImageNet detection examples

ImageNet detection failures that make sense

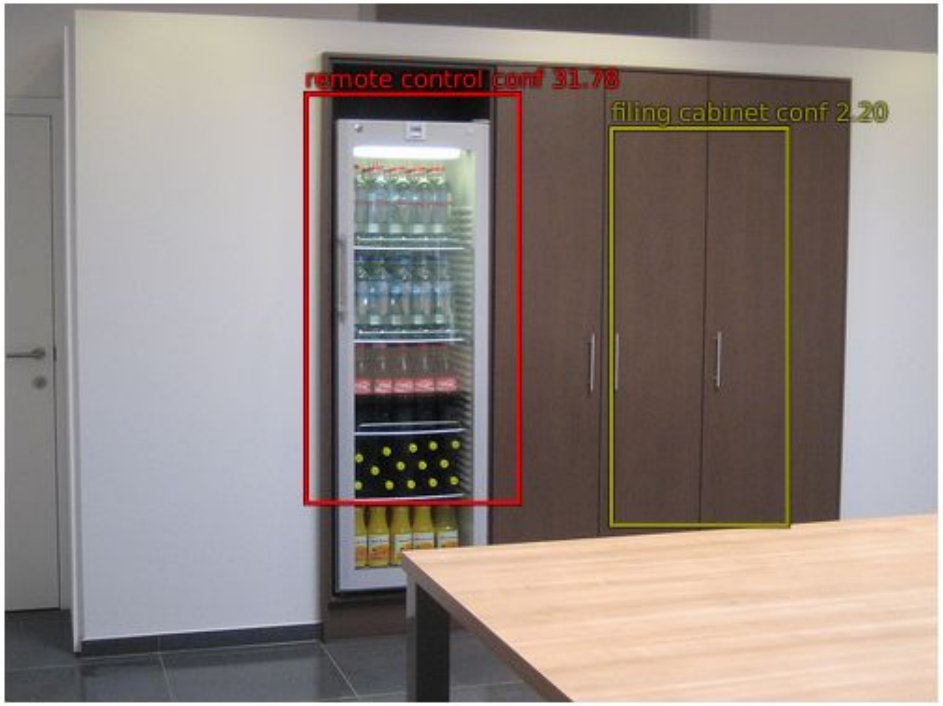


Top predictions:
corkscrew (confidence 38.1)

Groundtruth:
snake

ILSVRC2012_val_00000324.JPEG

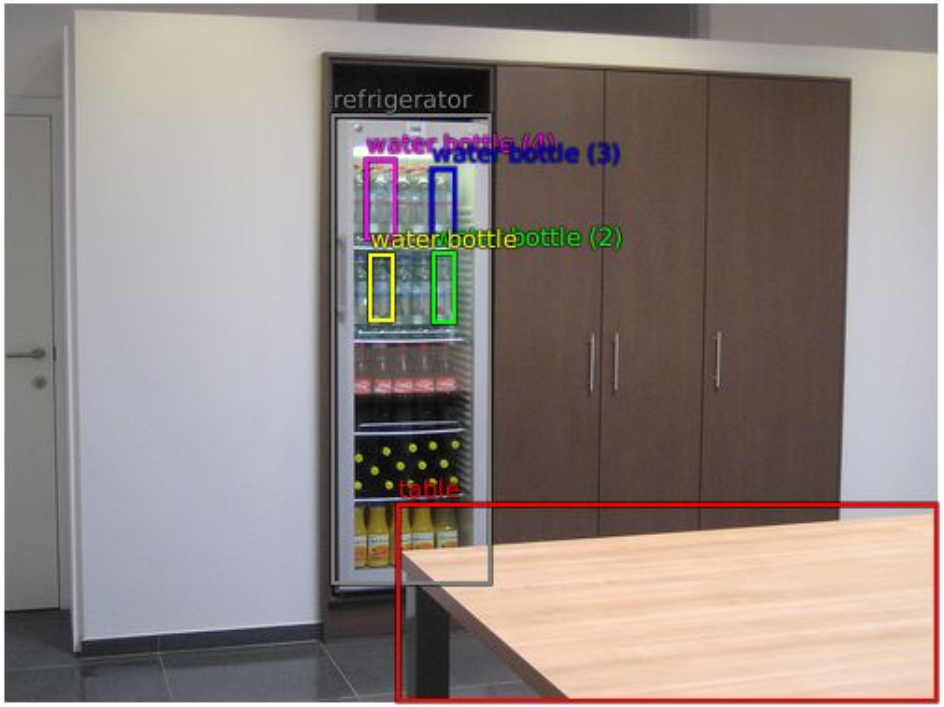
ImageNet detection failures that make sense



Top predictions:

remote control (confidence 31.8)
filing cabinet (confidence 2.2)

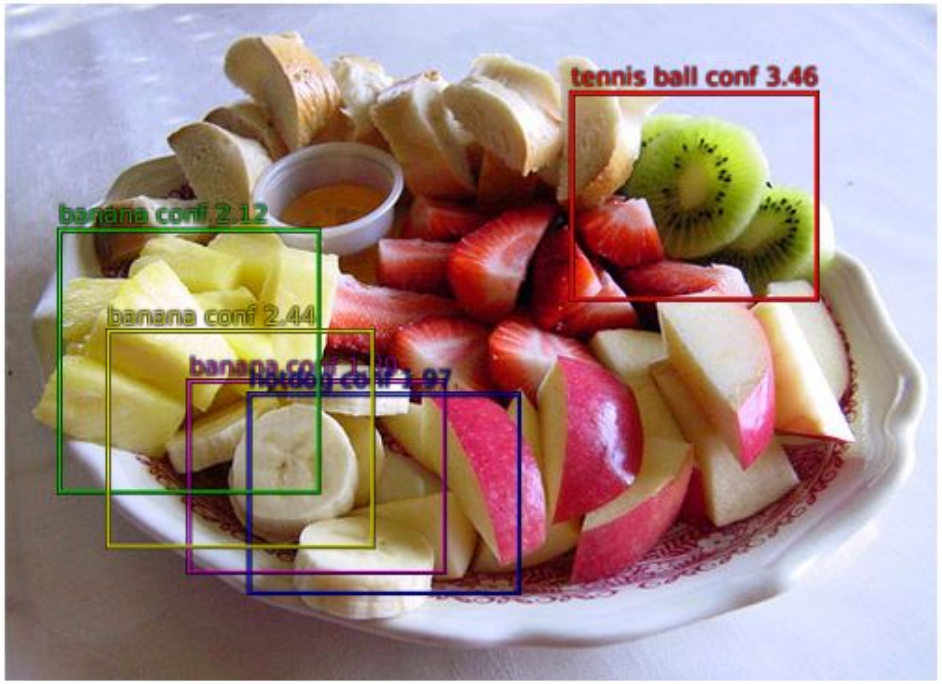
ILSVRC2012_val_00000331.JPEG



Groundtruth:

table
water bottle
water bottle (2)
water bottle (3)
water bottle (4)
refrigerator

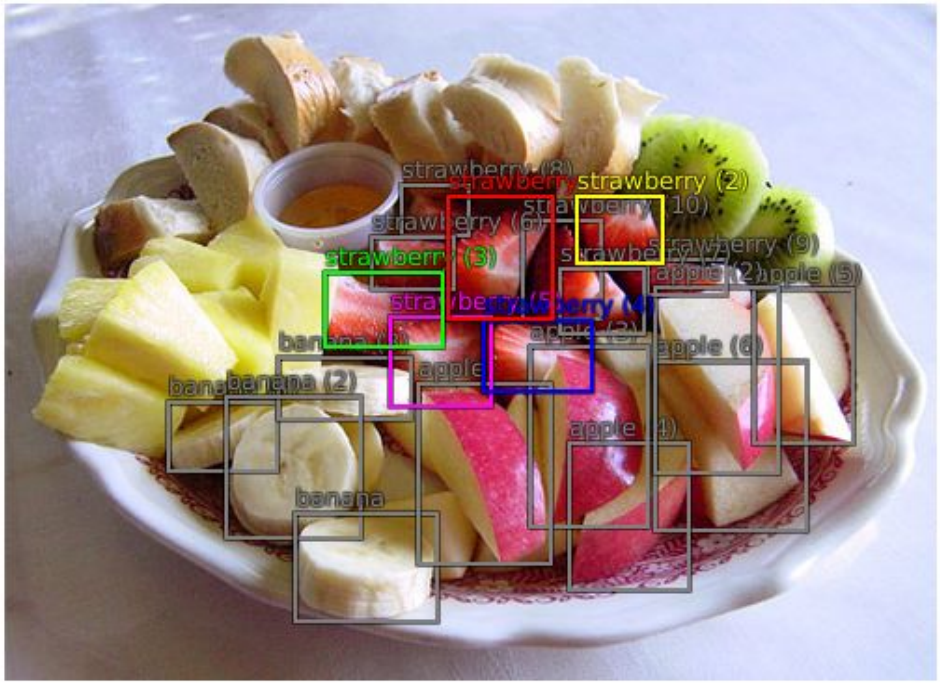
ImageNet detection: some hard ones



Top predictions:

tennis ball (confidence 3.5)
banana (confidence 2.4)
banana (confidence 2.1)
hotdog (confidence 2.0)
banana (confidence 1.9)

ILSVRC2012_val_00000320.JPEG



Groundtruth:

strawberry
strawberry (2)
strawberry (3)
strawberry (4)
strawberry (5)
strawberry (6)
strawberry (7)
strawberry (8)
strawberry (9)
strawberry (10)

Conclusions

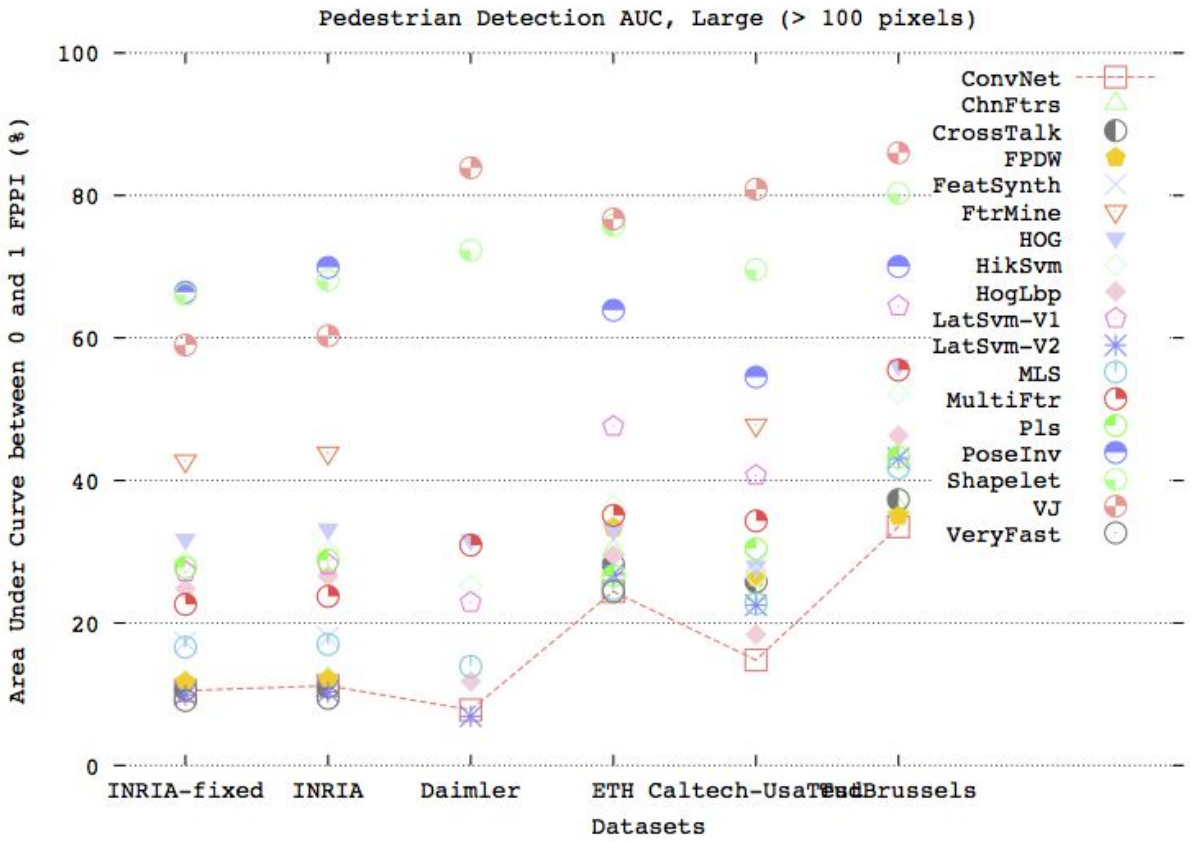
- **Deep ConvNets are the new feature extraction baseline** for computer vision
  - low hanging fruits: revisit traditional methods by swapping in deep ConvNets
- **Recent breakthroughs in object detection** using ConvNets
- We can **learn state of the art object localization** from features pre-trained on classification
- ConvNets are **efficient for detection**
- **Accumulation / voting of many predicted bounding boxes increases robustness** and yields the best object detection results
- **Density of sliding window** is important, the denser the better (best alignment gets best confidence)
  - **learning attention** with deep learning will improve accuracy while increasing efficiency

Questions

Additional Information

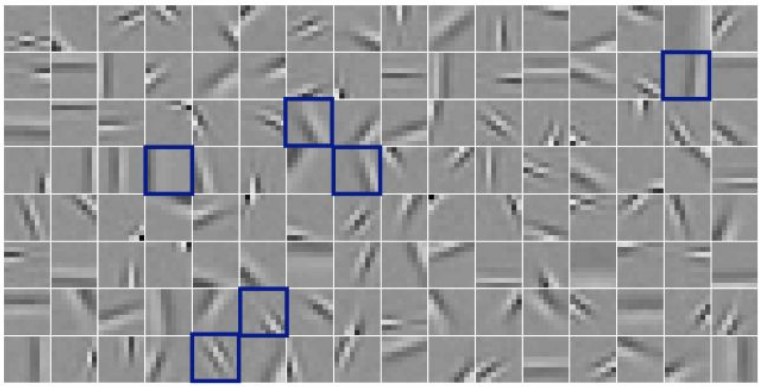
Pedestrian detection with ConvNets

- Detection with ConvNets in the **pre-Krizhevsky era**:
  - **State of the art / competitive** on major pedestrians datasets by [Sermanet'13]
  - **Smaller** models (1M) / datasets (thousands of samples)
  - Using **unsupervised learning for pre-training**

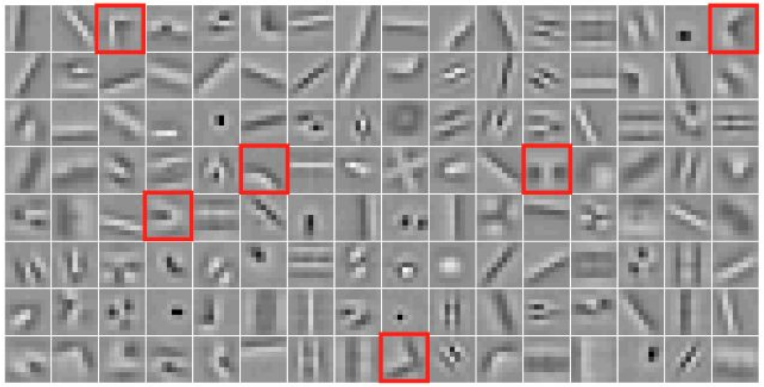


Pedestrian detection with unsupervised multi-stage feature learning. Sermanet, Pierre, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 3626-3633. IEEE, 2013.

Unsupervised learning for detection

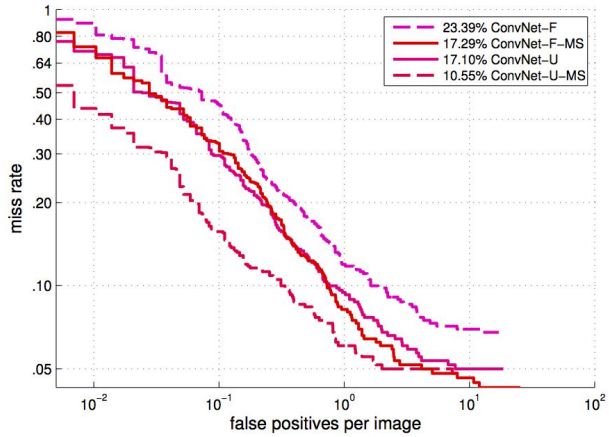


Sparse Coding



Convolutional Sparse Coding (CPSD)

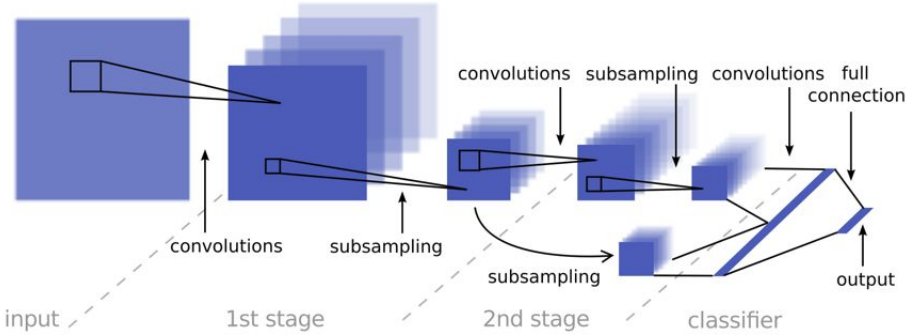
- INRIA pedestrian: **small dataset, unsupervised learning can help** [Sermanet'13]
- **CPSD improves** from 23.4% to 17.1% error
- **Multi-stage (MS) features improve** from 23.4% to 17.3%
- **CPSD + MS** = 10.55%



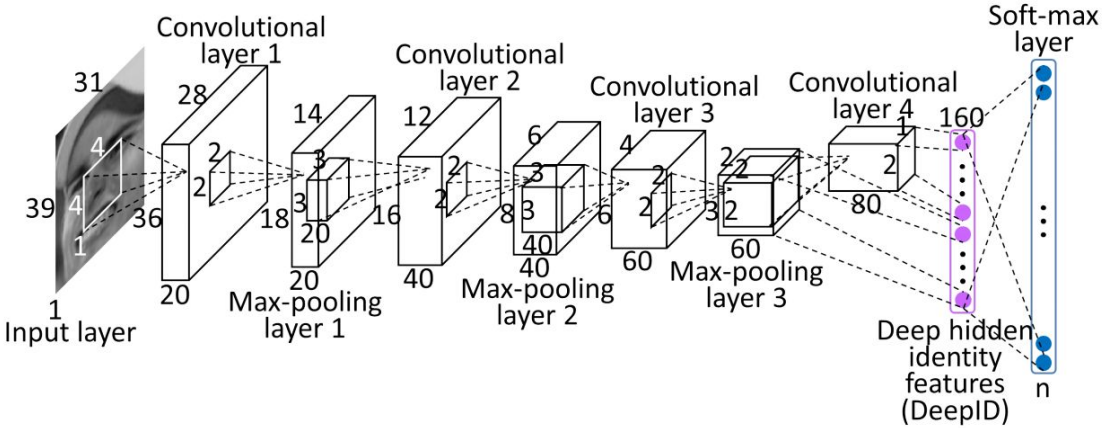
Pedestrian detection with unsupervised multi-stage feature learning. Sermanet, Pierre, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 3626-3633. IEEE, 2013.

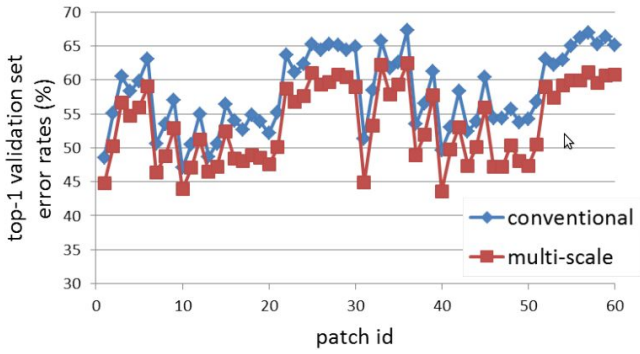
Multi-stage features (skip layer)

- **Skip connections can bring substantial improvements** [Sermanet'11'12'13] [Sun'14] by retaining lower-level information



Task	Single-Stage features	Multi-Stage features	Improvement %
Pedestrians detection (INRIA) [54]	14.26%	9.85%	31%
Traffic Signs classification (GTSRB) [53]	1.80%	0.83%	54%
House Numbers classification (SVHN) [55]	5.54%	5.36%	3.2%





Pedestrian detection with unsupervised multi-stage feature learning. Sermanet, P., Kavukcuoglu, K., Chintala, S., & LeCun, Y. (2013, June). In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*(pp. 3626-3633). IEEE.

Traffic sign recognition with multi-scale convolutional networks. Sermanet, Pierre, and Yann LeCun. *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011.

Convolutional neural networks applied to house numbers digit classification. Sermanet, Pierre, Soumith Chintala, and Yann LeCun. *Pattern Recognition (ICPR), 2012 21st International Conference on*. IEEE, 2012.

Deep Learning Face Representation from Predicting 10,000 Classes. Sun, Yi, Xiaogang Wang, and Xiaoou Tang.

Dogs vs Cats: results and approaches

Team	Error %	Software	Approach	ImageNet pre-training	deep learning
Pierre Sermanet	**1.09**	OverFeat	7 models average + multi-scale + drop fully connected layers	yes	yes
anton **(post competition)**	**1.68**	OverFeat	2 models average "about 15min of coding + 20 bucks to Amazon to get the images through the nets"	yes	yes
orchid	**1.70**	OverFeat + Decaf	OverFeat & Decaf models + hand-crafted features (haralick/zernikemoments/lbp/pftas/tas/surf)	yes	yes
Owen	**1.83**		?		
Paul Covington	**1.83**		?		
Maxim Milakov	**1.86**	nnForge	"rather deep" ConvNet with dropout and enriching training data by various distortions.	**no**	yes
we've been in KAIST	**1.90**	Caffe (Decaf)	L2-SVM hinge loss instead of softmax loss + fine tuning entire caffe model except for 1st convolution	yes	yes
Doug Koch	**1.94**		?		
fastml.com/ cats-and-dogs	**2.00**	OverFeat + Decaf	9 models ensemble	yes	yes