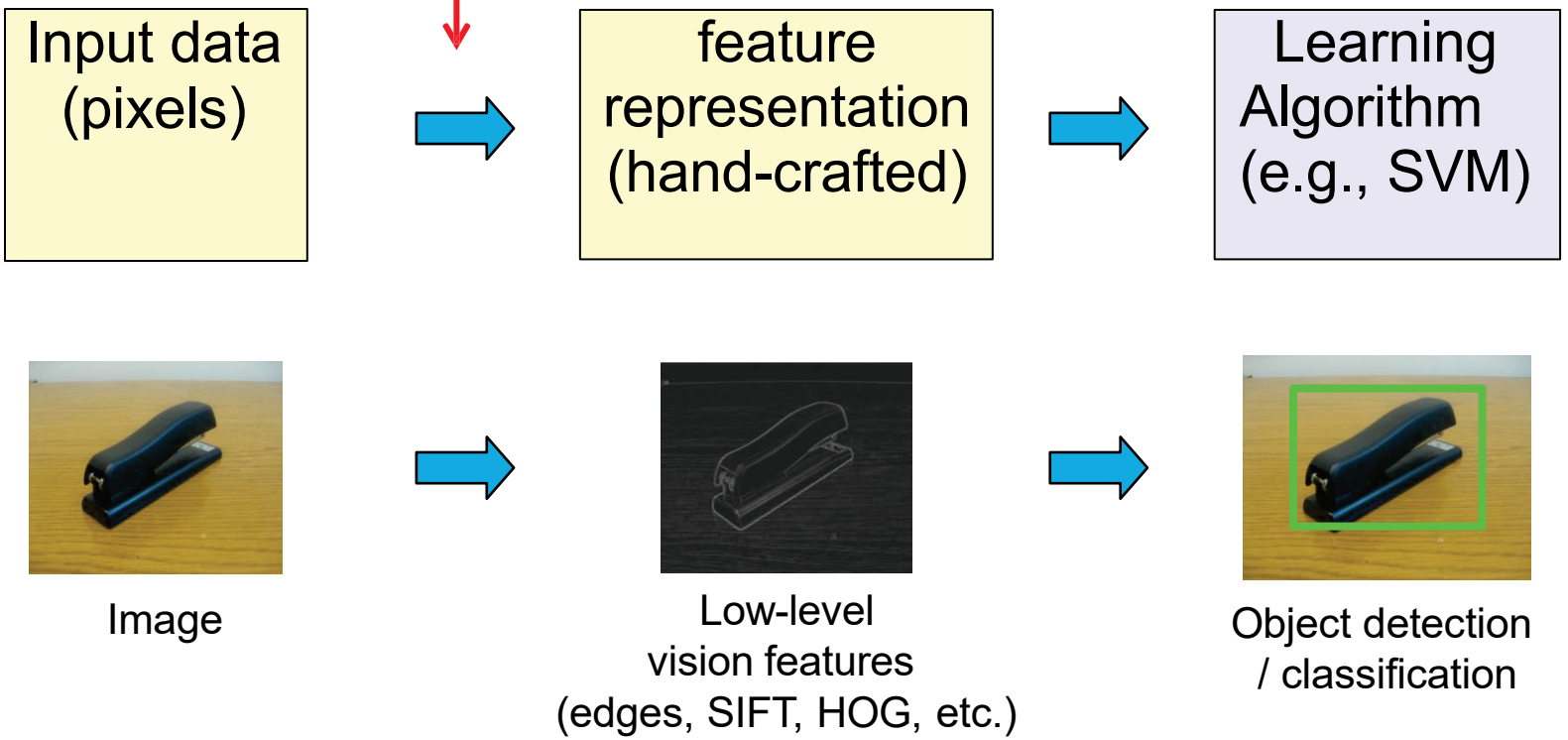
# Introduction to Deep Learning for Computer Vision

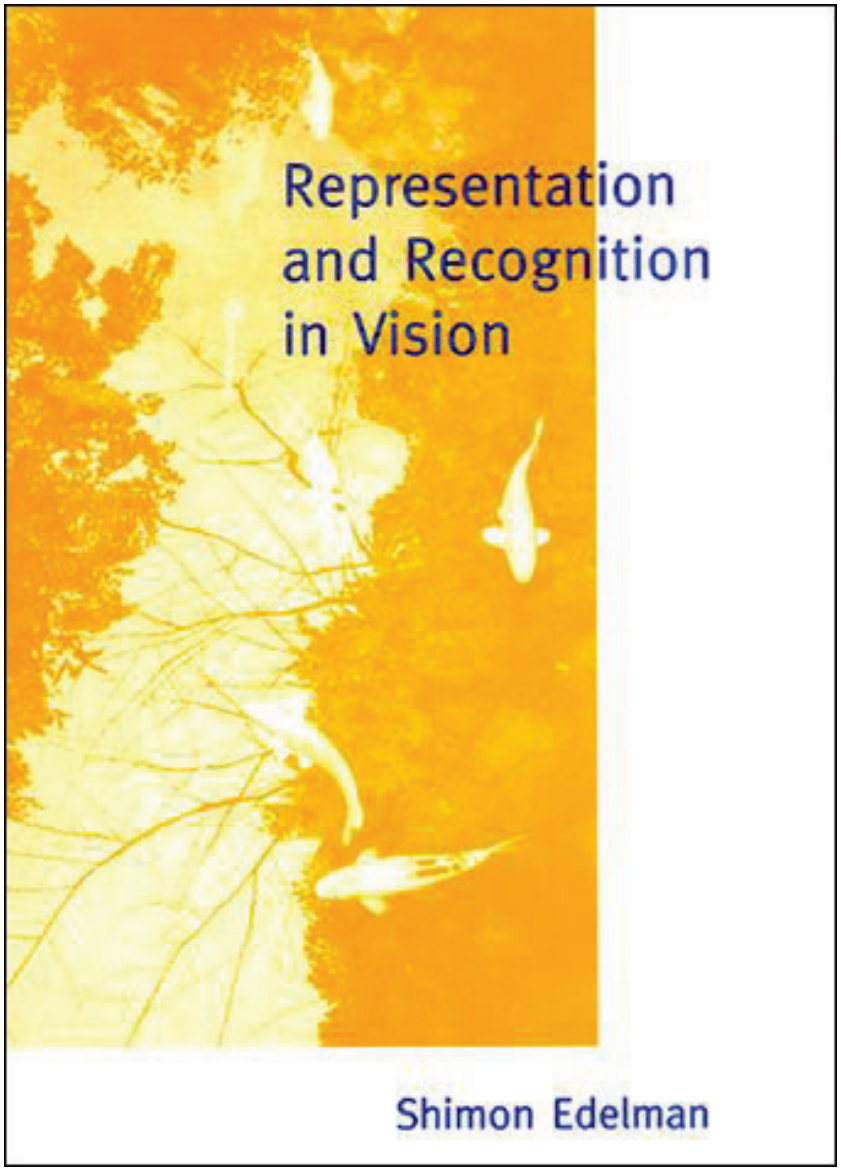
Adopted from CVPR'14 Tutorial given by
Graham Taylor (University of Guelph)
Marc Aurelio Ranzato (Facebook)
Honglak Lee (University of Michigan)
Pierre Sermanet (Google Research)

https://sites.google.com/site/deeplearningcvpr2014

# Traditional Recognition Approach

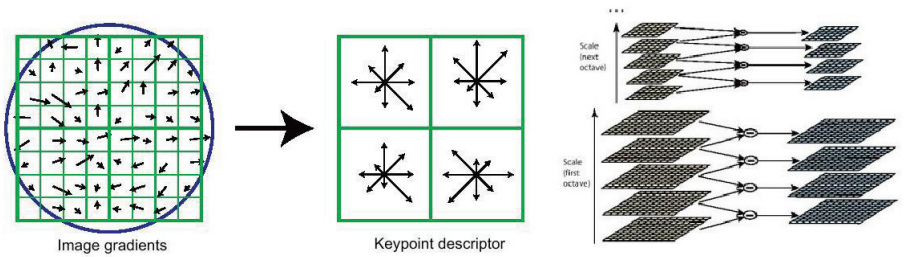
Features are not learned



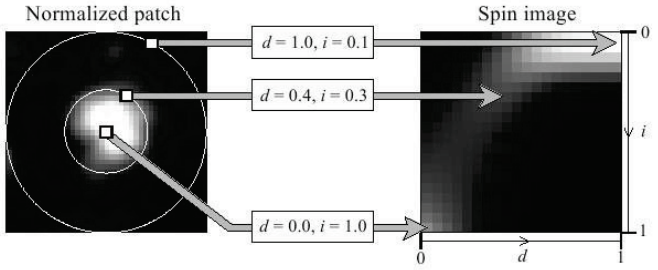



MIT Press 1999;
Behavioural and Brain Sciences 1998.

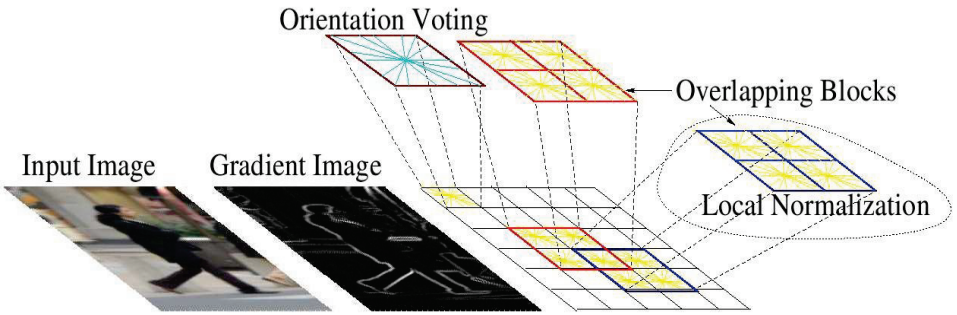
# Computer vision features



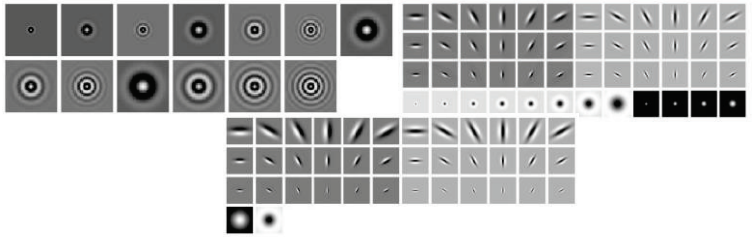

SIFT




Spin image




HoG



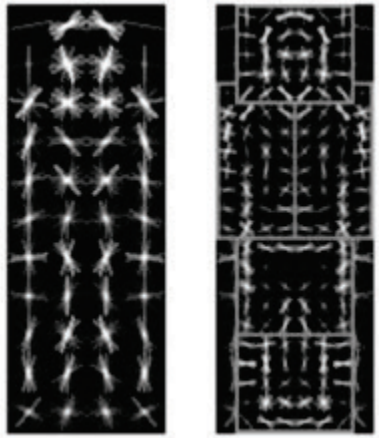
Textons

and many others:

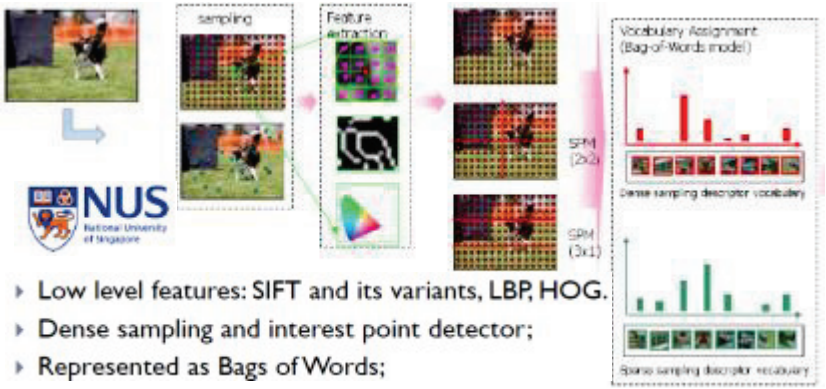
SURF, MSER, LBP, Color-SIFT, Color histogram, GLOH, …..

# Motivation

- Features are key to recent progress in recognition
- Multitude of hand-designed features currently in use
- Where next? Better classifiers? building better features?



Felzenszwalb, Girshick, McAllester and Ramanan, PAMI 2007




Yan & Huang
(Winner of PASCAL 2010 classification competition)

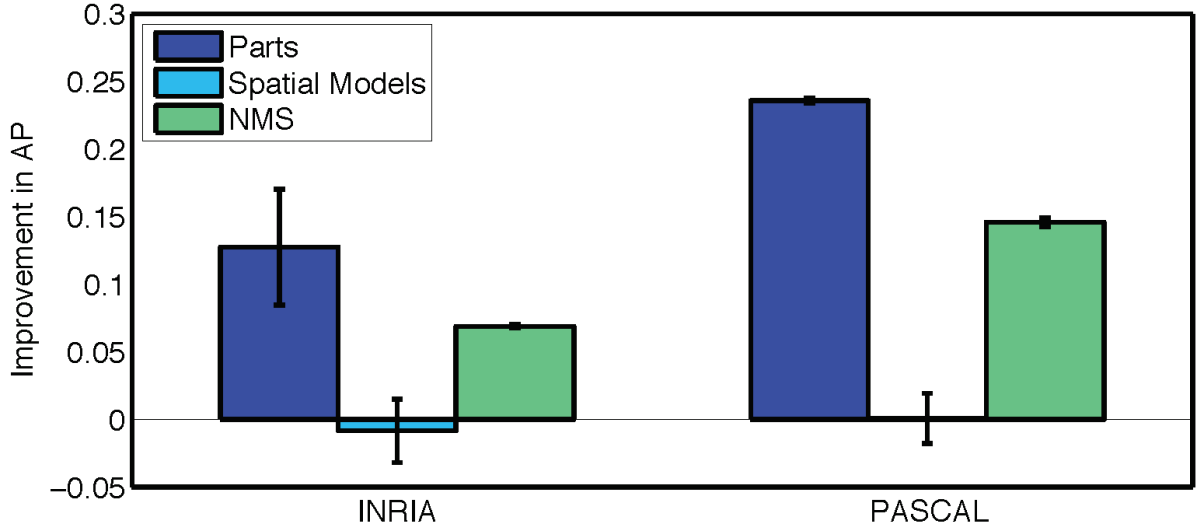
Slide: R. Fergus


Northbound platform 5

# What Limits Current Performance?

- Ablation studies on Deformable Parts Model
  - Felzenszwalb, Girshick, McAllester, Ramanan, PAMI'10

- Replace each part with humans (Amazon Turk):




Parikh & Zitnick, CVPR'10

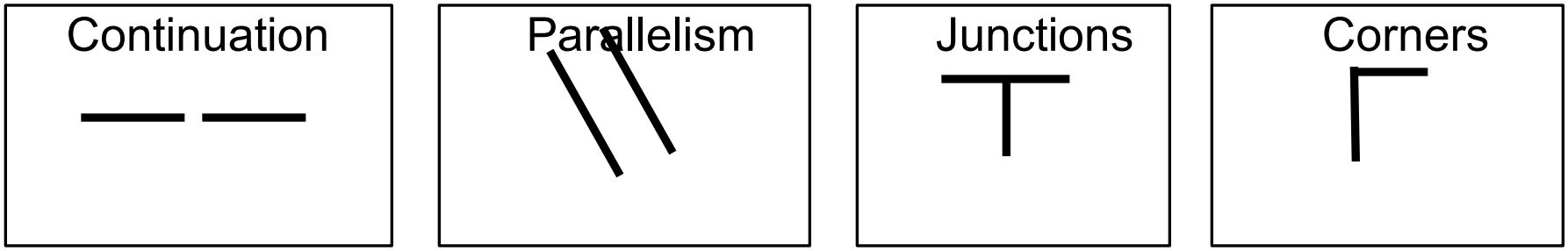
(NMS: Non-Maximum Suppression in DPM)

- Also removal of part deformations has small (<2%) effect.
  - Are "Deformable Parts" necessary in the Deformable Parts Model? Divvala, Hebert, Efros, ECCV 2012

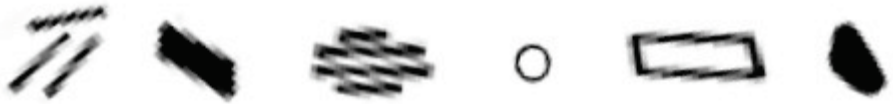
Slide: R. Fergus

# Mid-Level Representations

- Mid-level cues




"Tokens" from Vision by D.Marr:



- Object parts:

- Difficult to hand-engineer -> What about learning them?

Slide: R. Fergus

# Learning Feature Hierarchy

- Learn hierarchy
- All the way from pixels -> classifier
- One layer extracts features from output of previous layer

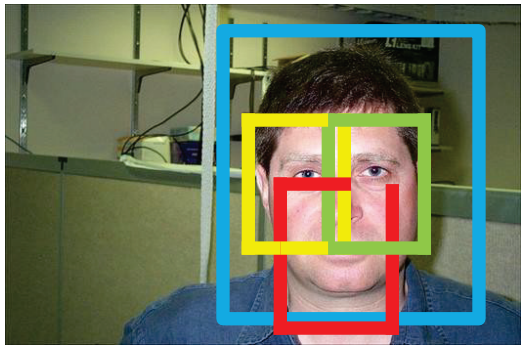

- **Train all layers jointly**

Slide: R. Fergus

# Learning Feature Hierarchy

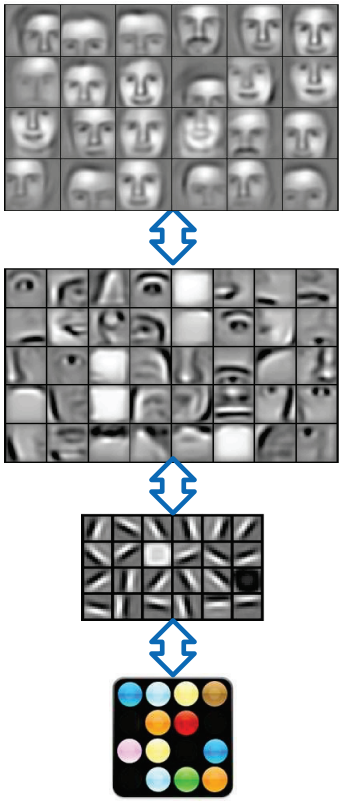
1. Learn **useful higher-level features** from images

Input data



Lee et al., ICML 2009;
CACM 2011

Feature representation



3rd layer
"Objects"

2nd layer
"Object parts"

1st layer
"Edges"

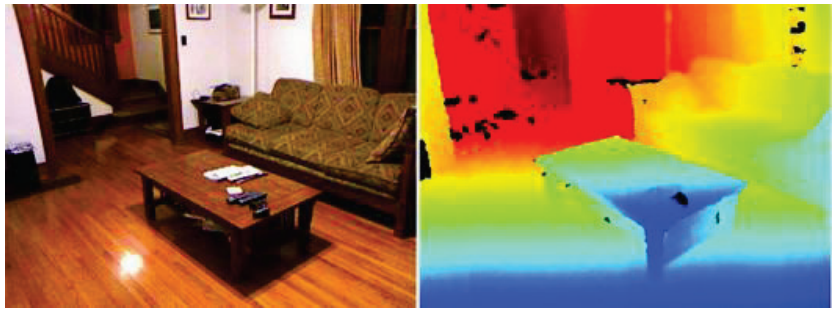
Pixels

2. **Fill in representation gap** in recognition

# Learning Feature Hierarchy

- Better performance

- Other domains (unclear how to hand engineer):
  - Kinect
  - Video
  - Multi spectral



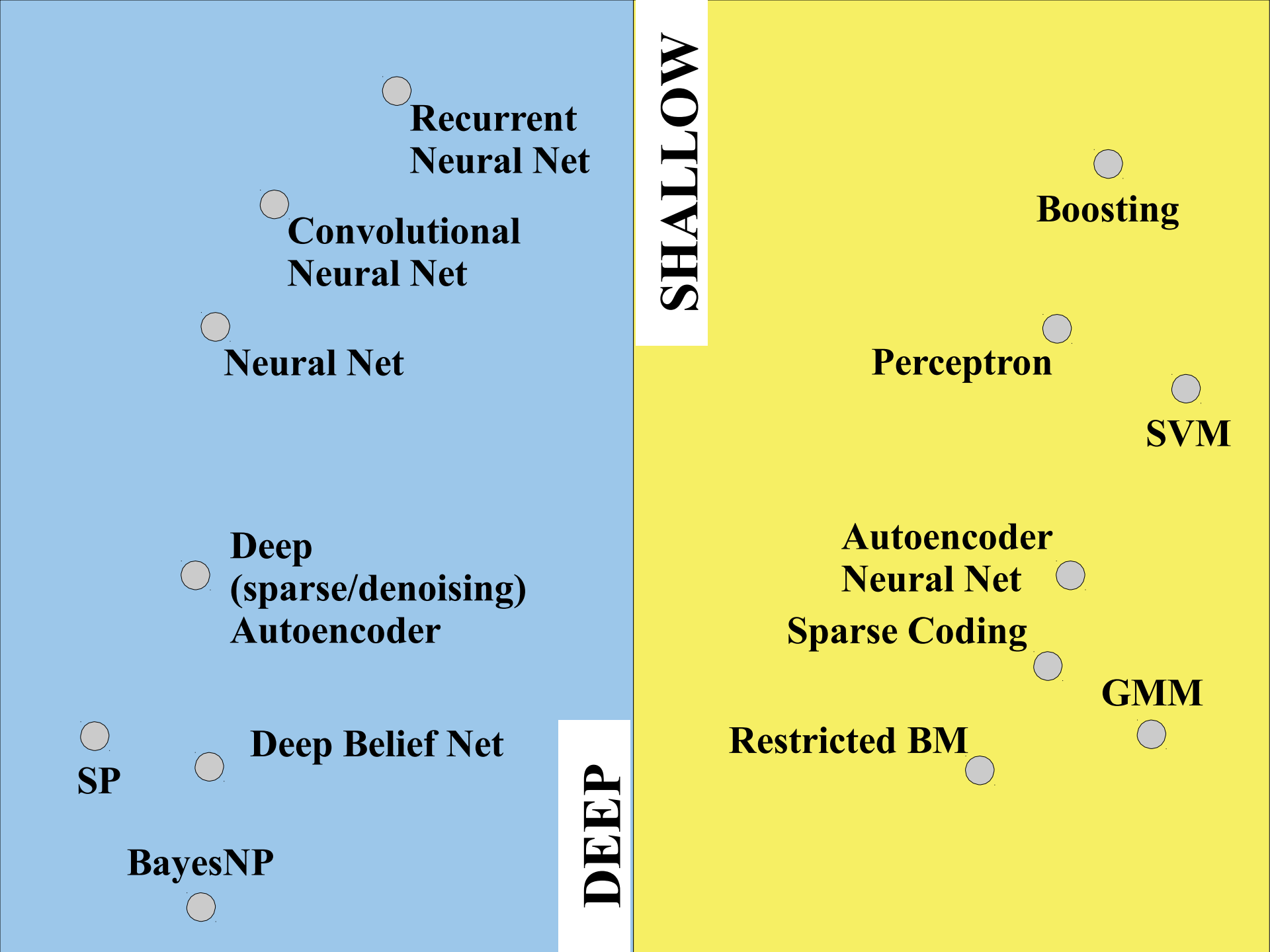
- Feature computation time
  - Dozens of features now regularly used [e.g., MKL]
  - Getting prohibitive for large datasets (10's sec /image)

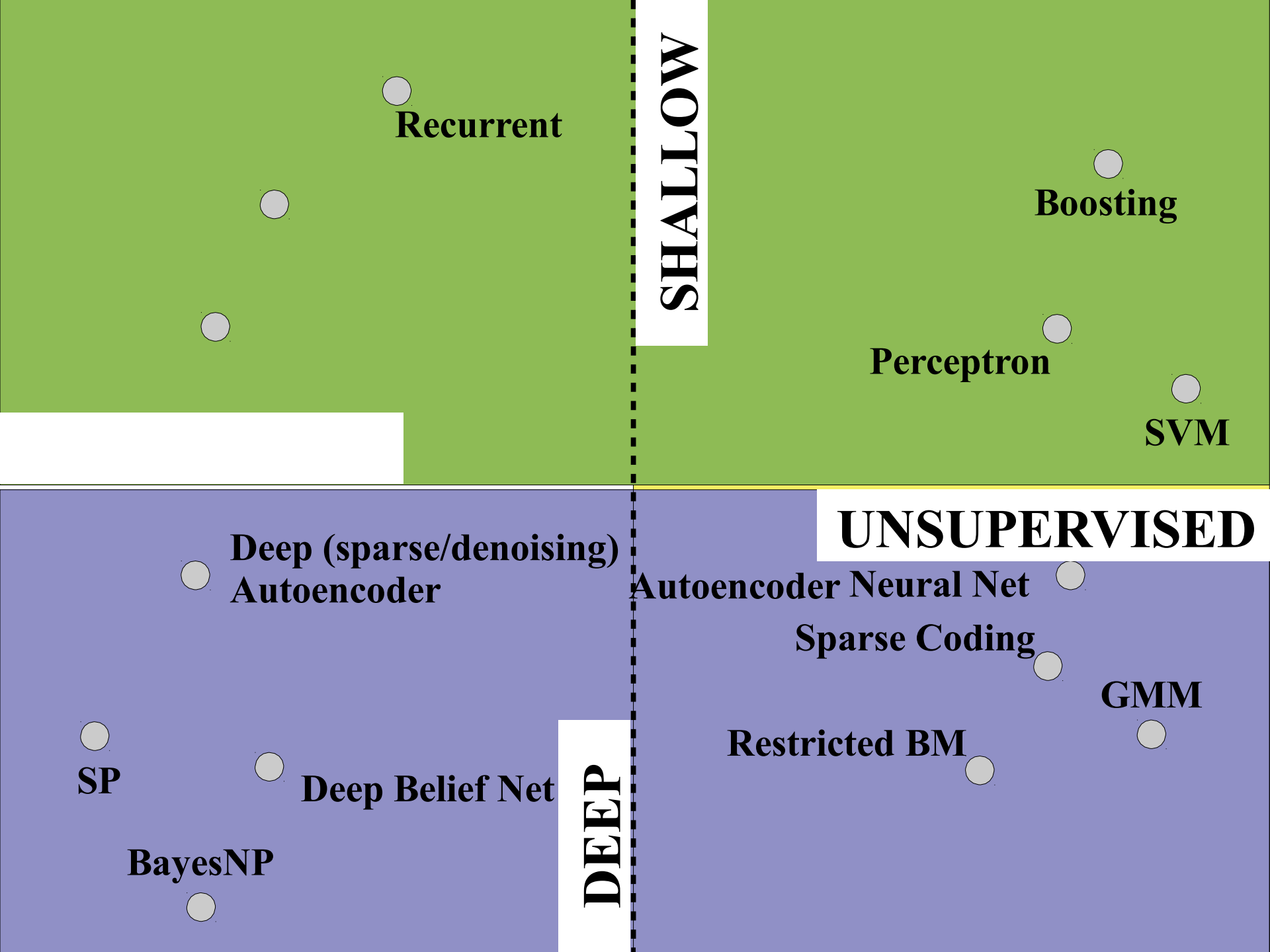
Slide: R. Fergus

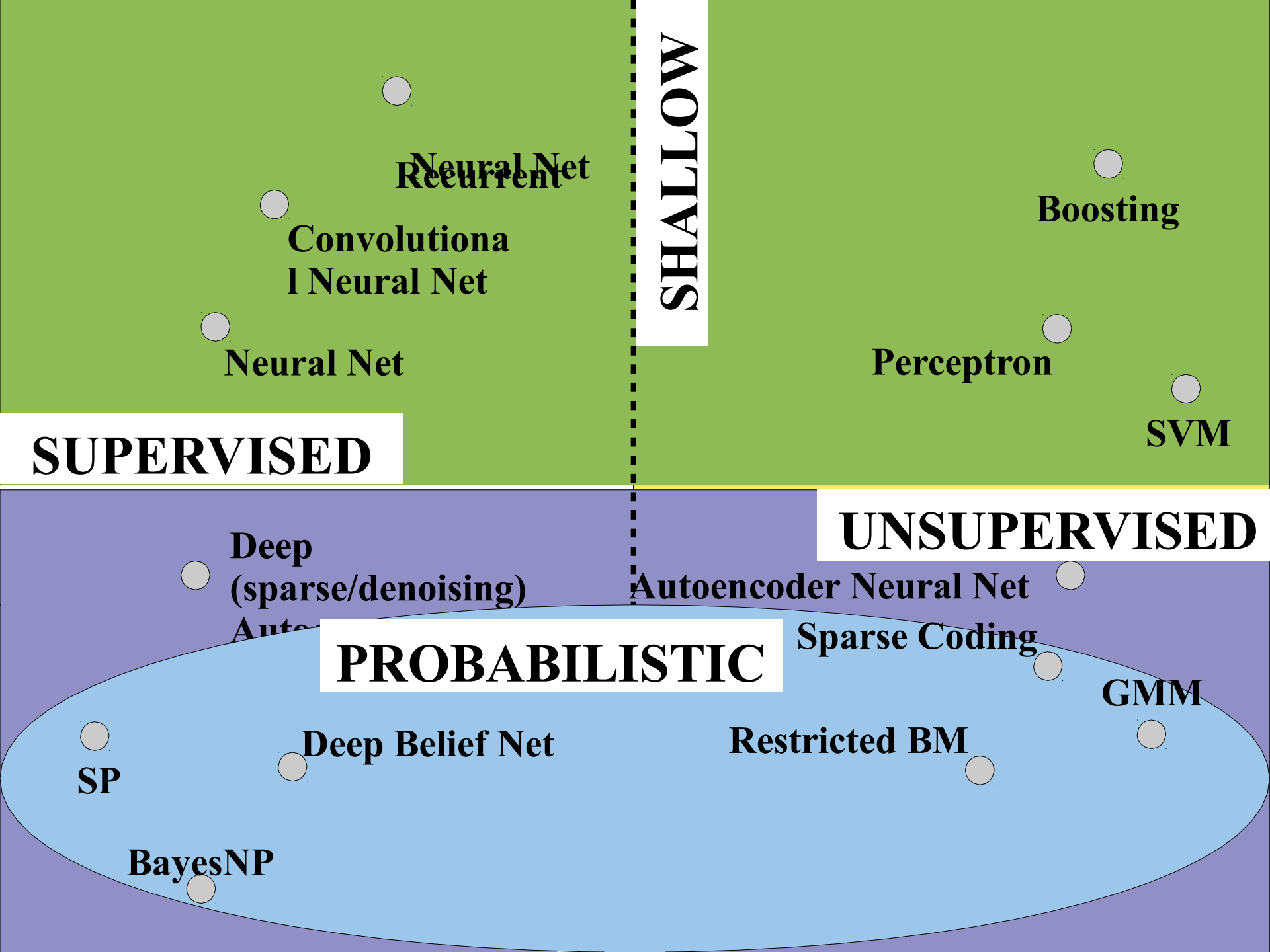
# Approaches to learning

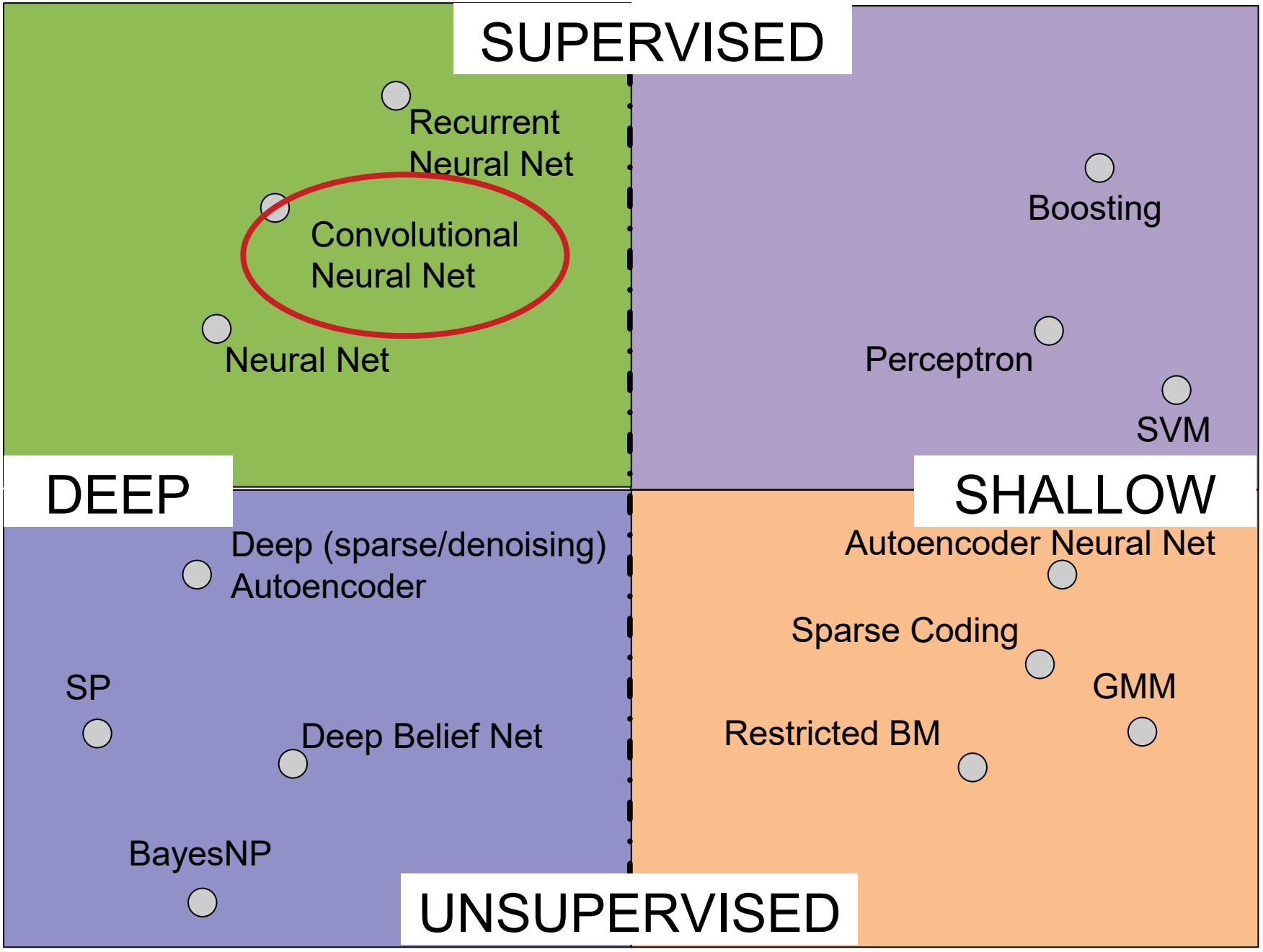
- Supervised Learning
  - End-to-end learning of deep architectures (e.g., deep neural networks) with back-propagation
  - Works well when the amounts of labels is large
  - Structure of the model is important (e.g. convolutional structure)

- Unsupervised Learning
  - Learn statistical structure or dependencies of the data from unlabeled data
  - Layer-wise training
  - Useful when the amount of labels is not large


Recurrent
Neural Net
Convolutional
Neural Net
Neural Net
Deep
(sparse/denoising)
Autoencoder
Deep Belief Net
SP
BayesNP
DEEP
SHALLOW
Boosting
Perceptron
SVM
Autoencoder
Neural Net
Sparse Coding
GMM
Restricted BM


Recurrent
SHALLOW
Boosting
Perceptron
SVM
UNSUPERVISED
Deep (sparse/denoising)
Autoencoder
Autoencoder Neural Net
Sparse Coding
GMM
Restricted BM
SP
Deep Belief Net
DEEP
BayesNP


SHALLOW
Neural Net
Recurrent
Convolutiona
l Neural Net
Neural Net
Boosting
Perceptron
SVM
SUPERVISED
UNSUPERVISED
Deep
(sparse/denoising)
Autoencoder Neural Net
Sparse Coding
PROBABILISTIC
GMM
Deep Belief Net
Restricted BM
SP
BayesNP




Slide: M. Ranzato

# Object Detection and Classification (Object Recognition)

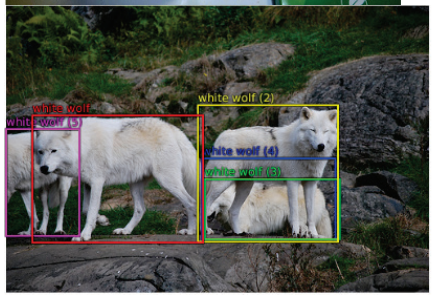
# What is object detection?

- classification
- localization
- detection
- segmentation


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

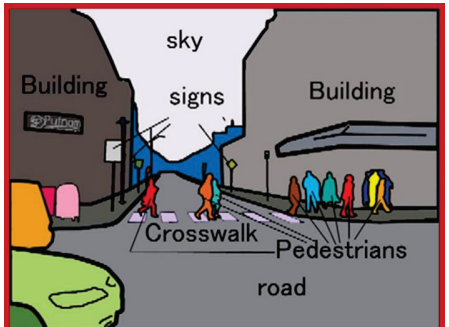
**Groundtruth:**
pencil sharpener

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 (Facebook, Microsoft, Lenovo)
    - 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

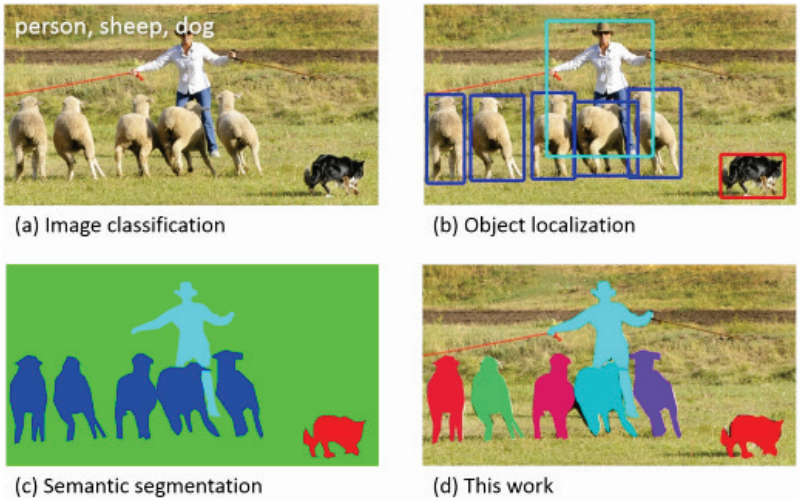


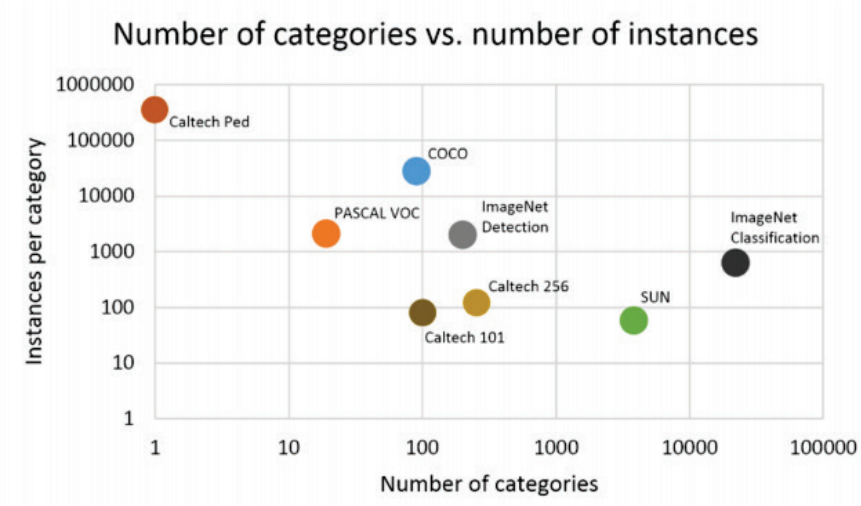


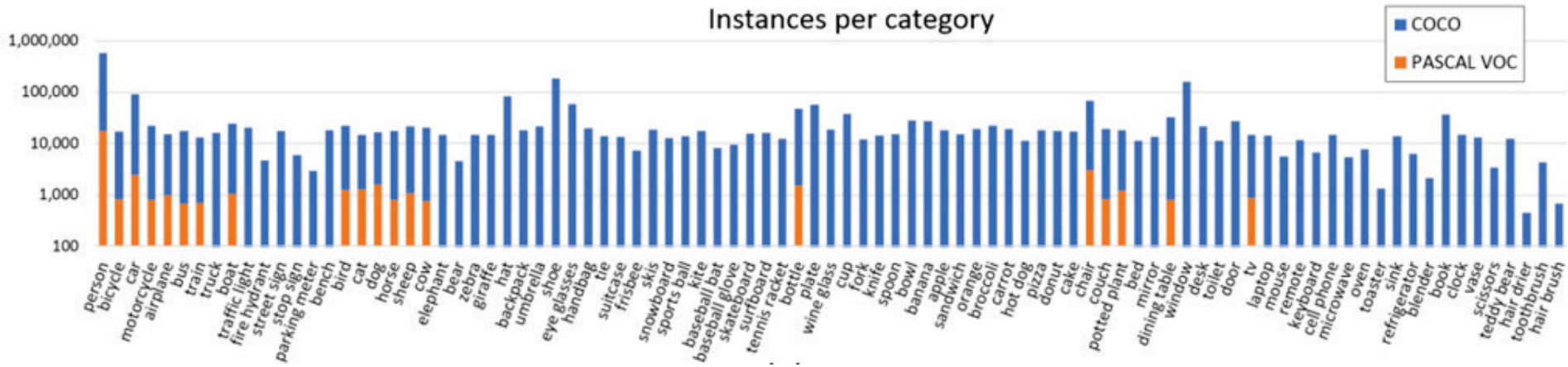
Fig. 2: Example of (a) iconic object images, (b) iconic scene images, and (c) non-iconic images.

**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 – the *imbalanced data distribution* problem)

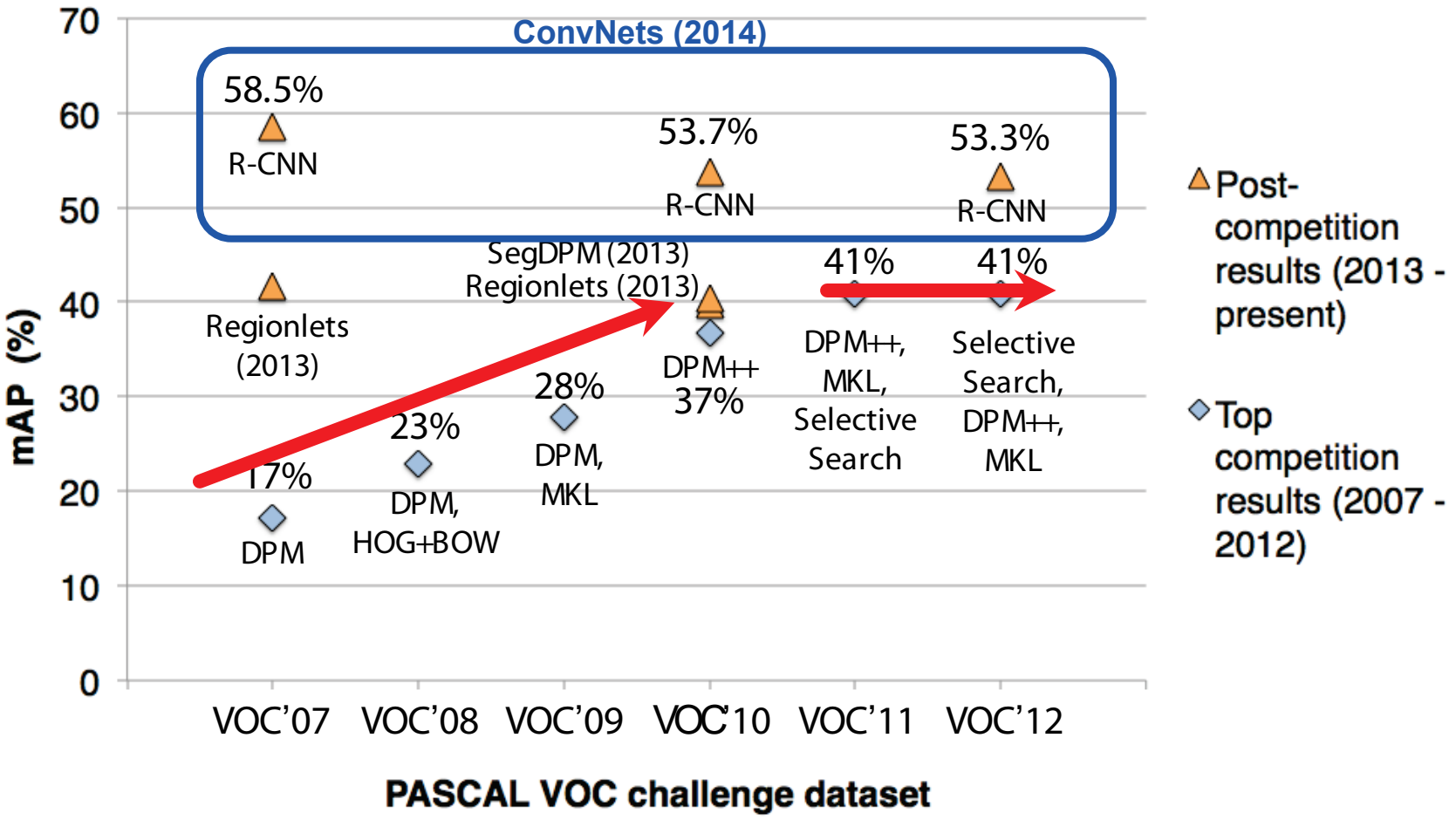



**Microsoft COCO: Common Objects in Context,** Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick, http://arxiv.org/abs/1405.0312, May 2014

# Recent history of object detection

- **Large improvements** using Deep Learning [Girshick'13/14]
  ( **mAP** – mean average precision of all classes )




**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).

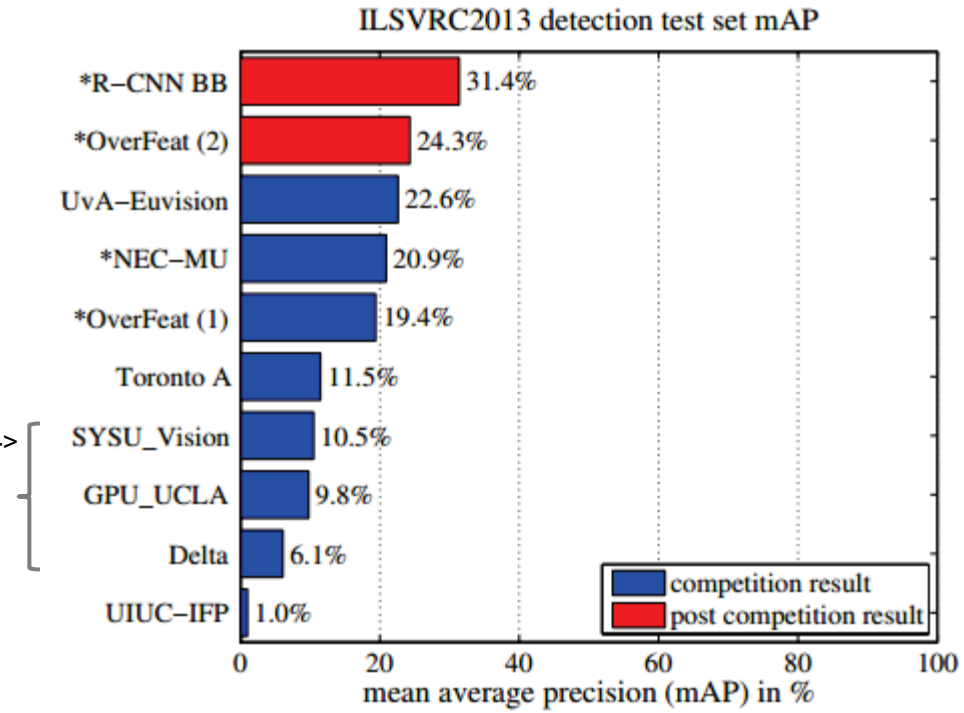
**The PASCAL Visual Object Classes Challenge - a Retrospective**, Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A. Accepted for International Journal of Computer Vision, 2014

# 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]**<br>• shadow detection | UCF<br>CMU<br>UIUC | **state of the art**<br>**state of the art**<br>**state of the art** | 90.56%<br>88.79%<br>93.16% |
| **[Sander Dieleman, 2014]**<br>• image attributes | Kaggle Galaxy Zoo challenge | **state of the art** | 0.07492 |

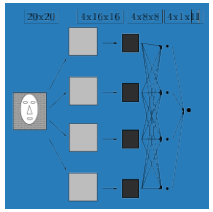
S. H. Khan, M. Bennamoun, F. Sohel, R. Togneri. **Automatic Feature Learning for Robust Shadow Detection,** CVPR 2014
Sander Dieleman, Kaggle Galaxy Zoo challenge 2014 http://benanne.github.io/2014/04/05/galaxy-zoo.html

# 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**."

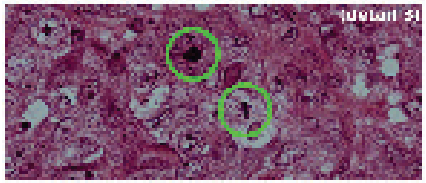
Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson, **CNN Features off-the-shelf: an Astounding Baseline for Recognition**, http://arxiv.org/abs/1403.6382, DeepVision CVPR 2014 workshop

# History of detection with ConvNets

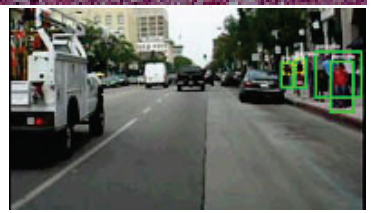


Vaillant, Monrocq, LeCun 1994
Osadchy, LeCun, Miller 2004
**Face detection with pose estimation!**

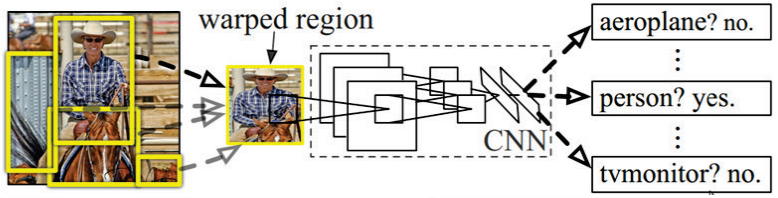
LeCun, Huang, Bottou 2004
**NORB dataset**

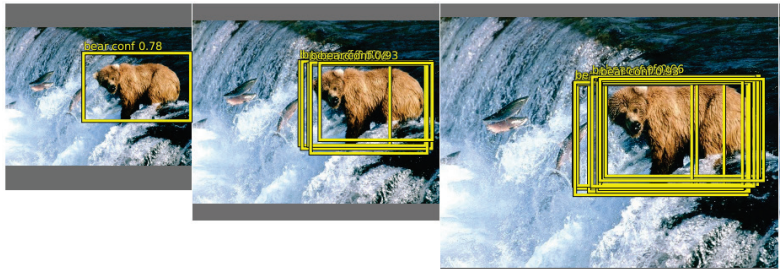


Cireşan et al. 2013
**Mitosis detection**



Sermanet et al. 2013
**Pedestrian detection**



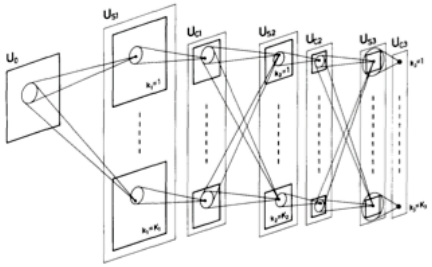



**PASCAL detection**
Girshick et al. 2013
Szegedy, Toshev, Erhan 2013

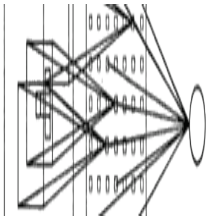
**ImageNet detection**
Girshick et al. 2014 (R-CNN), 2015 (Fast R-CNN)
Sermanet et al. 2014 (OverFeat)

**MS COCO**
Liu et al 2015 (SSD), Redmon et al. 2015 (YOLO)

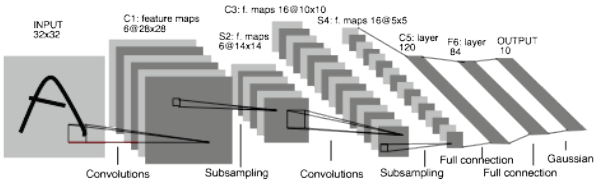
# History of ConvNets



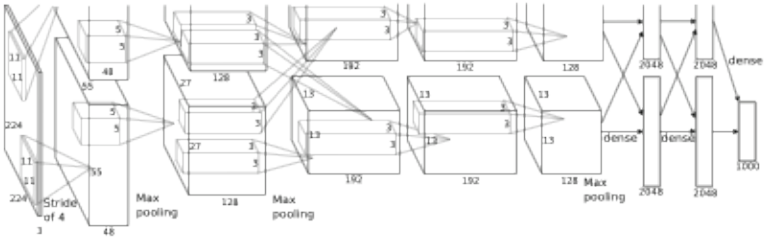
Fukushima 1980
**Neocognitron**



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

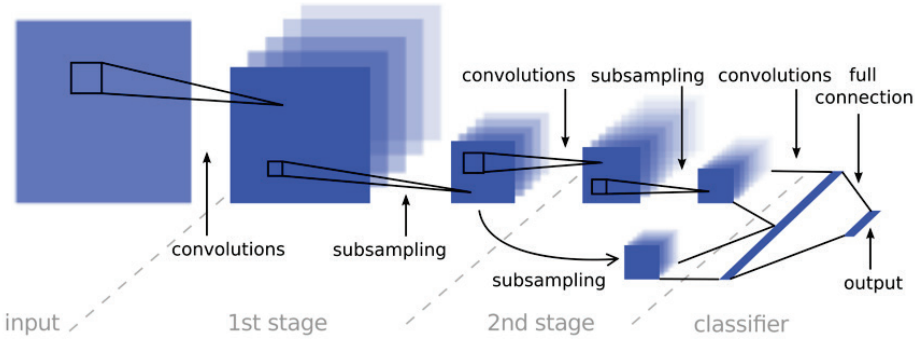
# What is a ConvNet?

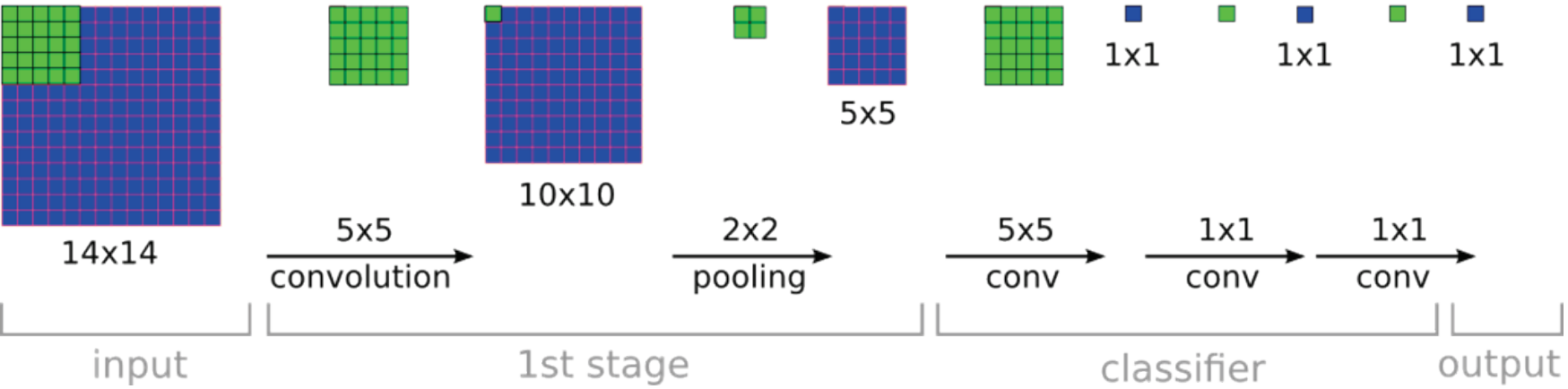
- A special type of Neural Net that incorporates **priors about continuous signals**
  - sound / speech (2D signal)
  - images (3D signal)
  - videos (4D signal)

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

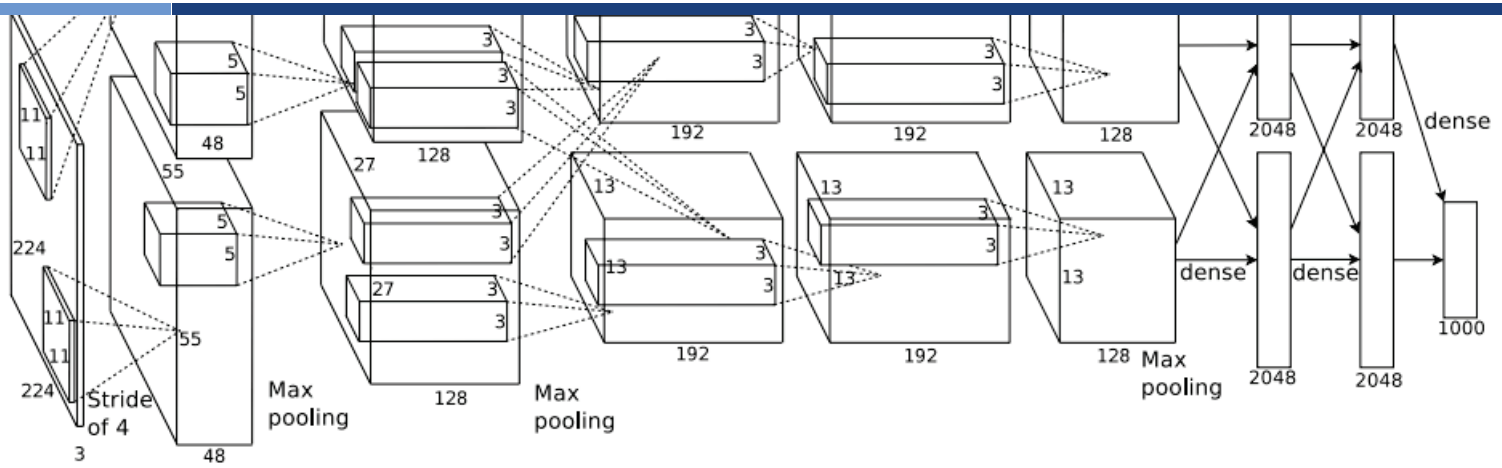
- In its **simplest form**, a ConvNet is just a series of stages of the form:
  - convolution
  - bias
  - non-linearity (ReLu or sigmoid functions)
  - pooling

- Normalization layers (LCN) may be added



# ConvNets 2.0

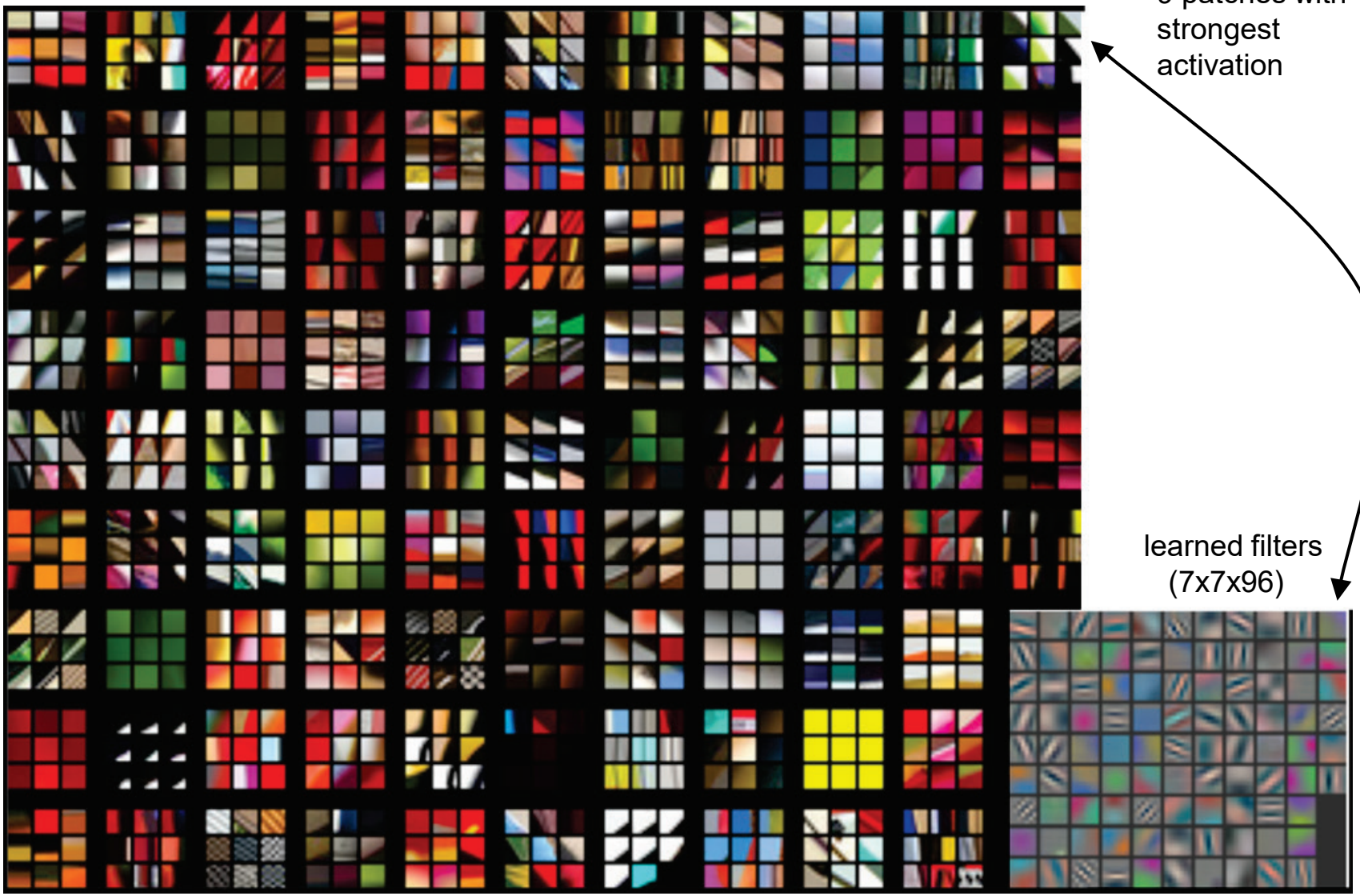



- [Krizhevsky'12] win 2012 ImageNet classification with a **much bigger ConvNet** than before:
  - **deeper**: 7 stages vs 3 before
  - **larger**: 60 million parameters vs 1 million before

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

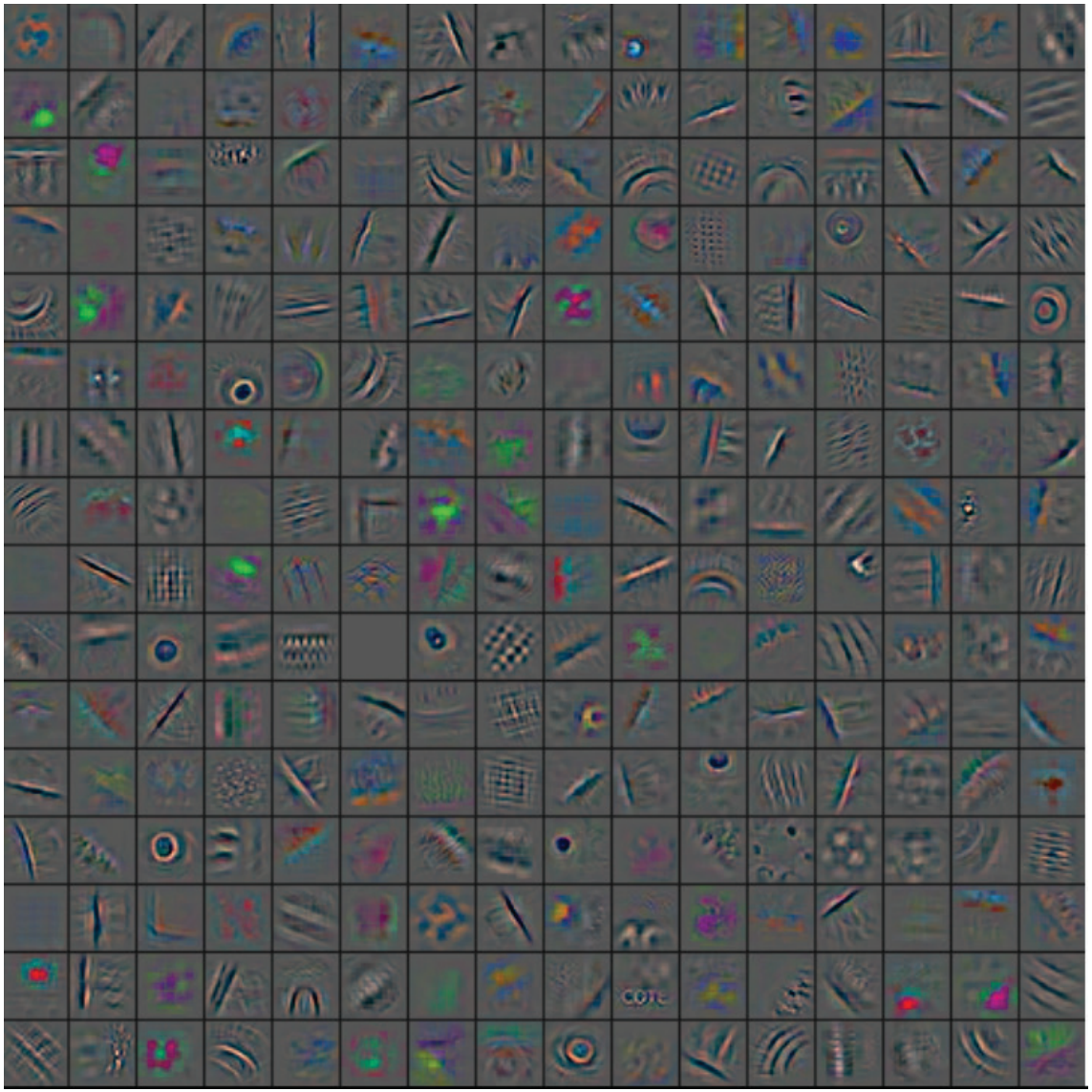
**Imagenet classification with deep convolutional neural networks.** Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. *Advances in neural information processing systems*. 2012.

# Learned convolutional filters: Stage 1



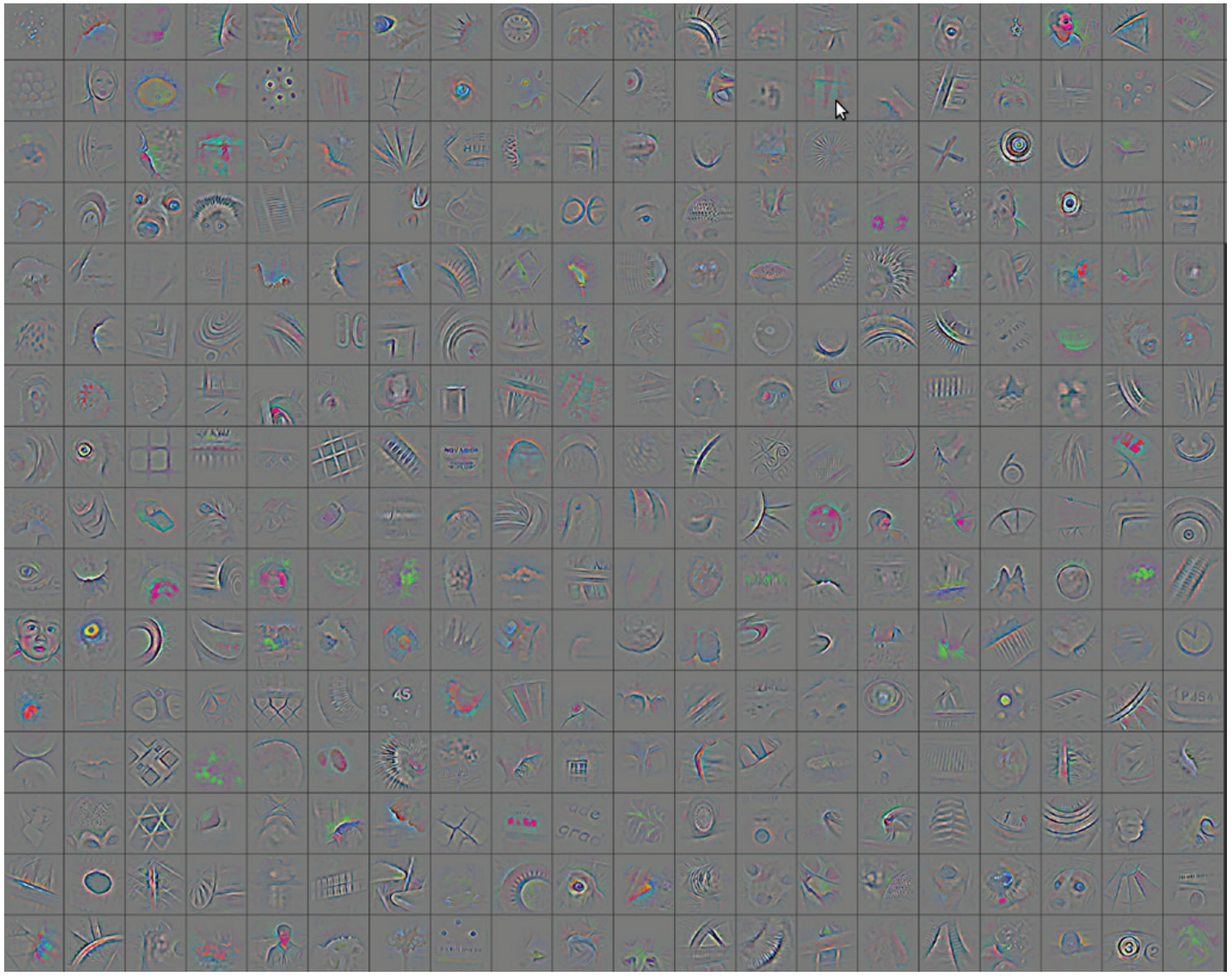

**Visualizing and understanding convolutional neural networks.** Zeiler, Matthew D., and Rob Fergus. *arXiv preprint arXiv:1311.2901* (2013).

# Strongest activations: Stage 2



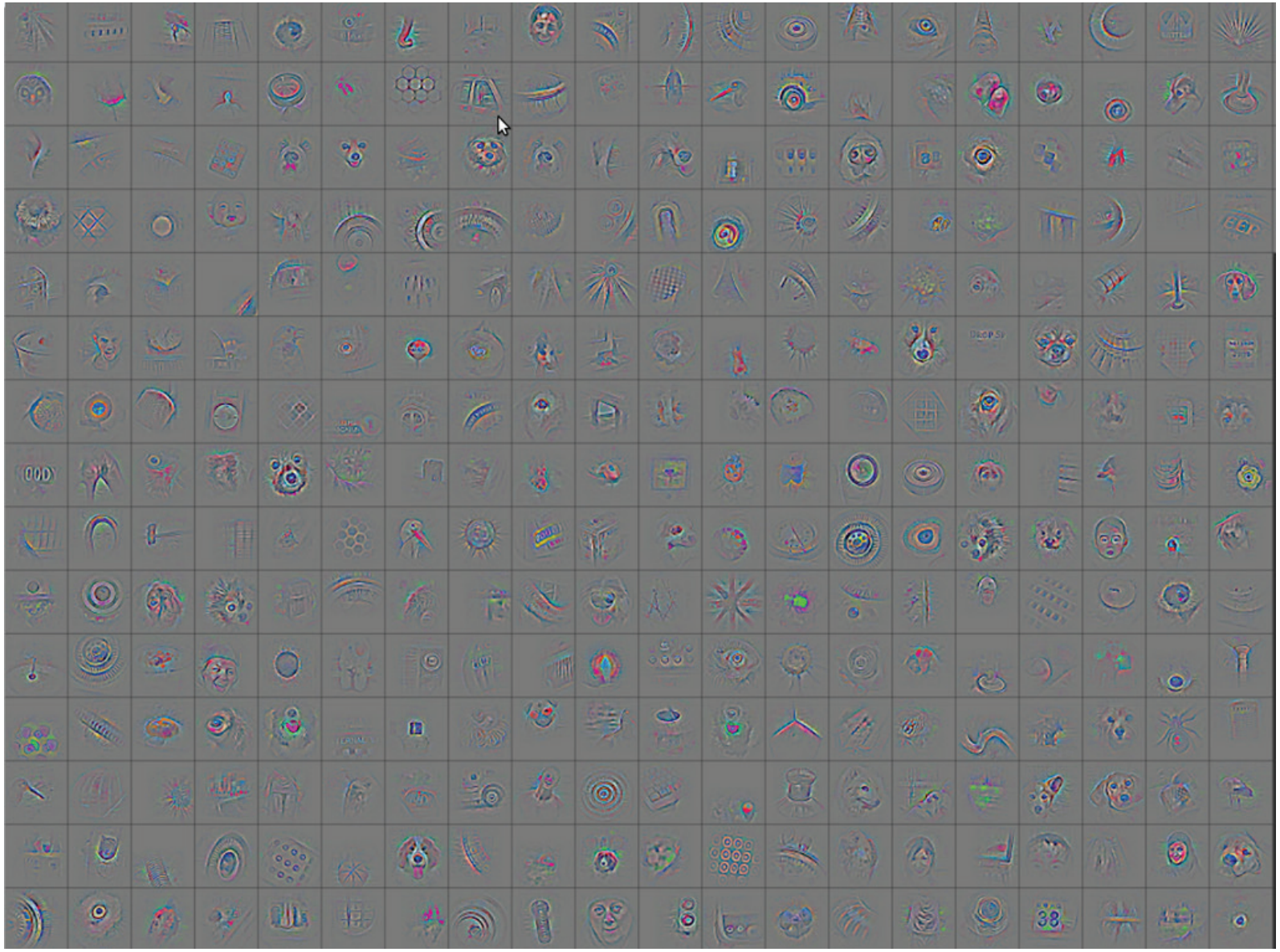
**Visualizing and understanding convolutional neural networks.** Zeiler, Matthew D., and Rob Fergus. *arXiv preprint arXiv:1311.2901* (2013).

# Strongest activations: Stage 3



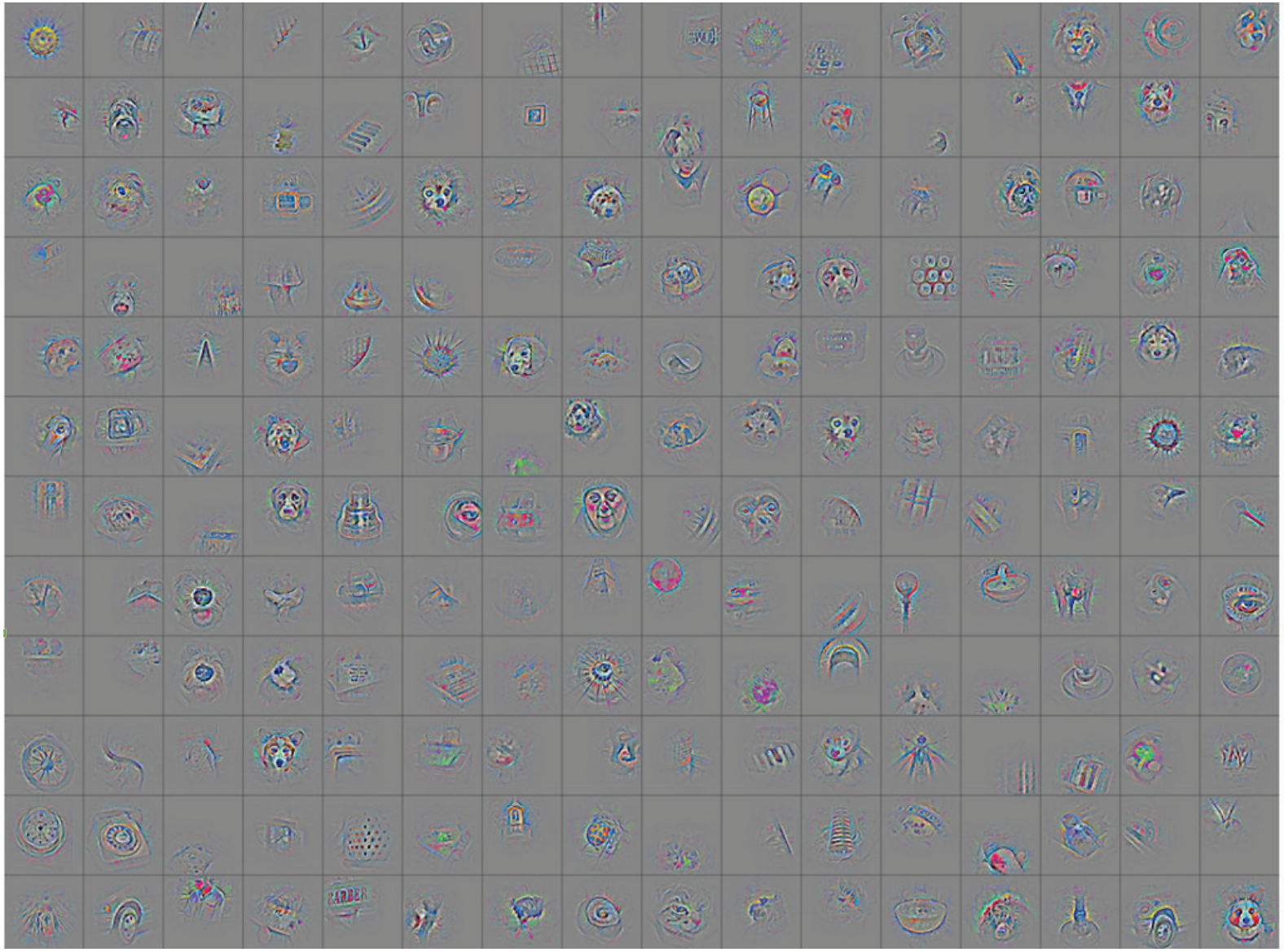
**Visualizing and understanding convolutional neural networks.** Zeiler, Matthew D., and Rob Fergus. *arXiv preprint arXiv:1311.2901* (2013).

# 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



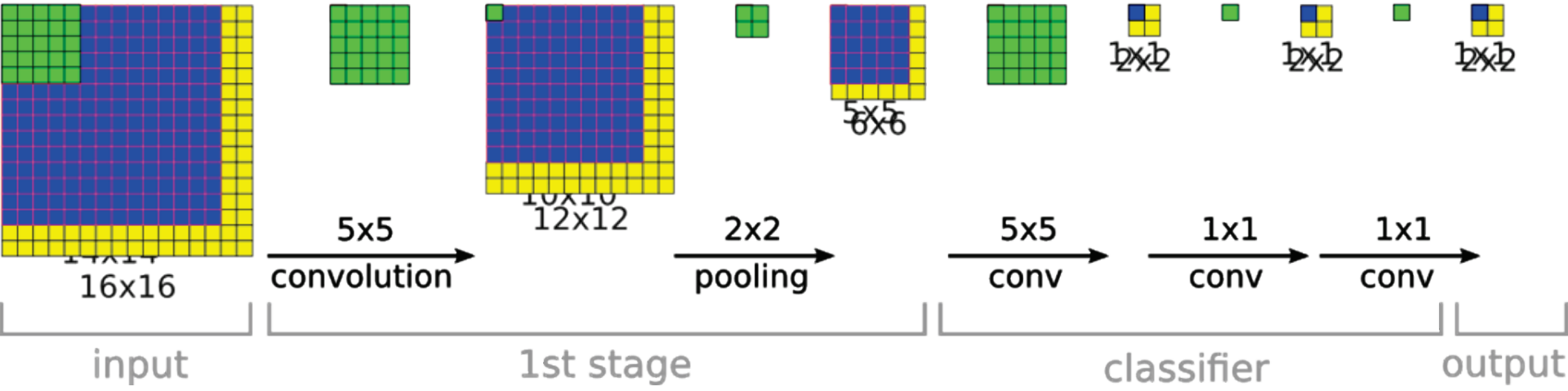
**Visualizing and understanding convolutional neural networks.** Zeiler, Matthew D., and Rob Fergus. *arXiv preprint arXiv:1311.2901* (2013).

# 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)

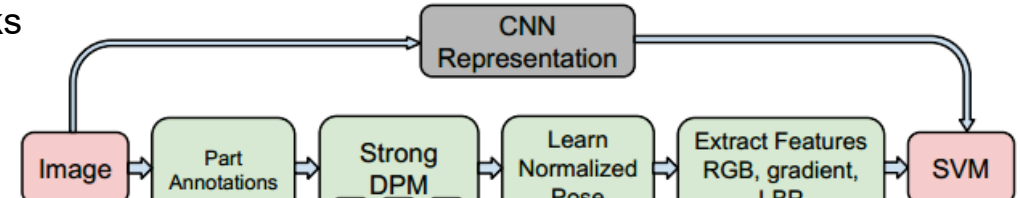
- **ConvNets are convolutional all the way up** including fully connected layers

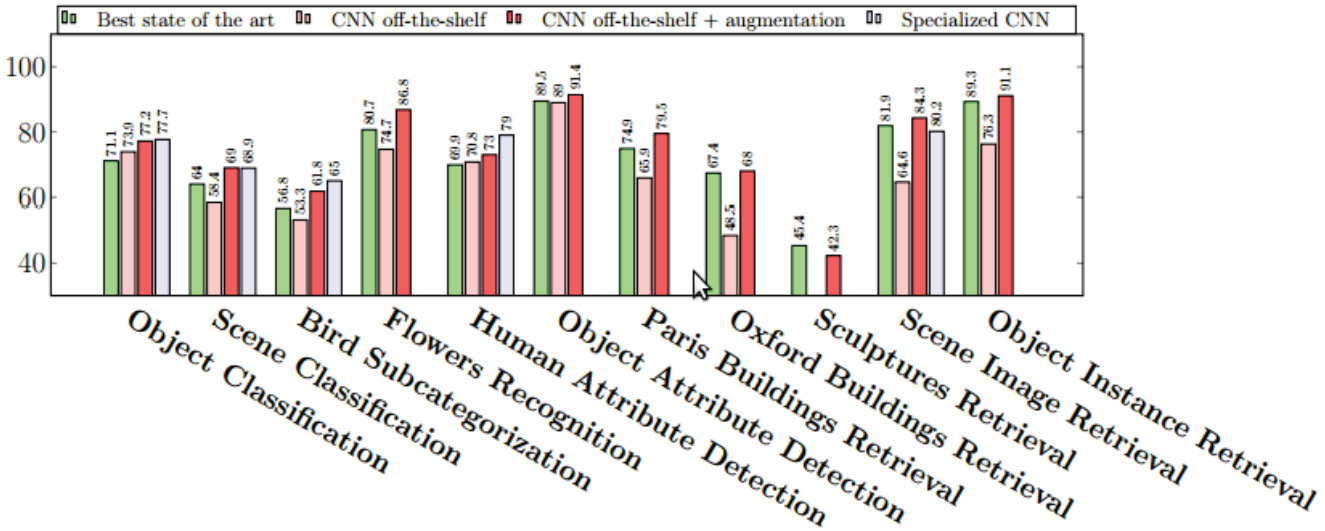


# ImageNet pre-training

- **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).

# Questions

1. Why Convolutional Neural Networks (CNN) for computer vision
2. What is a CNN and how does it work
3. What is Max Pooling and what is it for
4. What is Feedforward in CNN and what is it for
5. What is Backpropagation in CNN and what is it for
6. What is dropout in CNN and what is it for