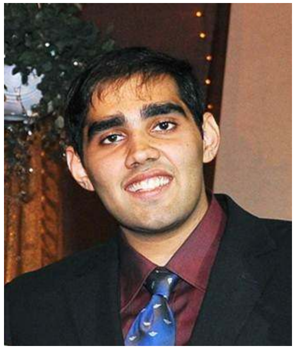
Object Detectors Emerge in Deep Scene CNNs

Bolei Zhou,
Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba



Massachusetts Institute of Technology

CNN for Object Recognition

Large-scale image classification result on ImageNet

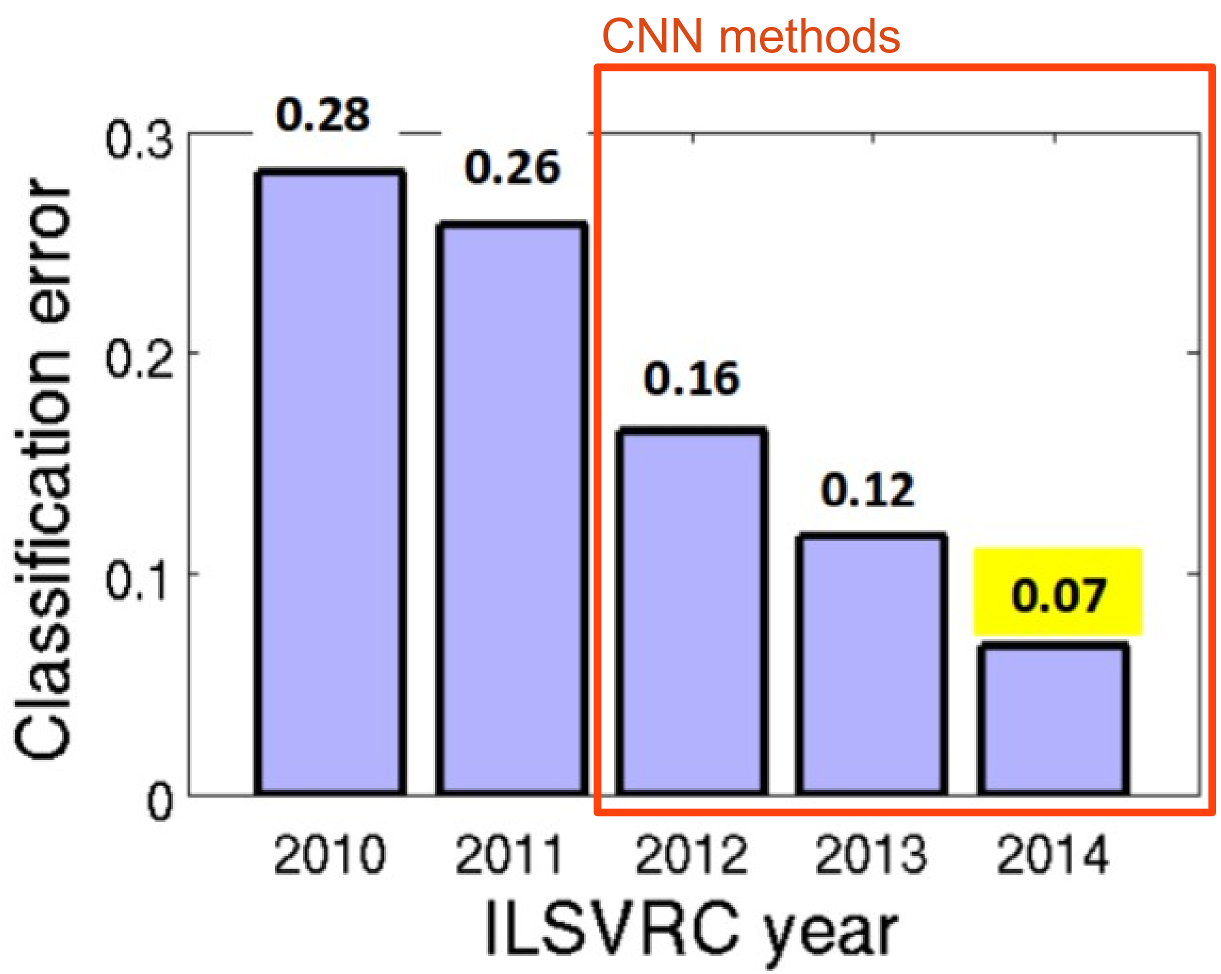
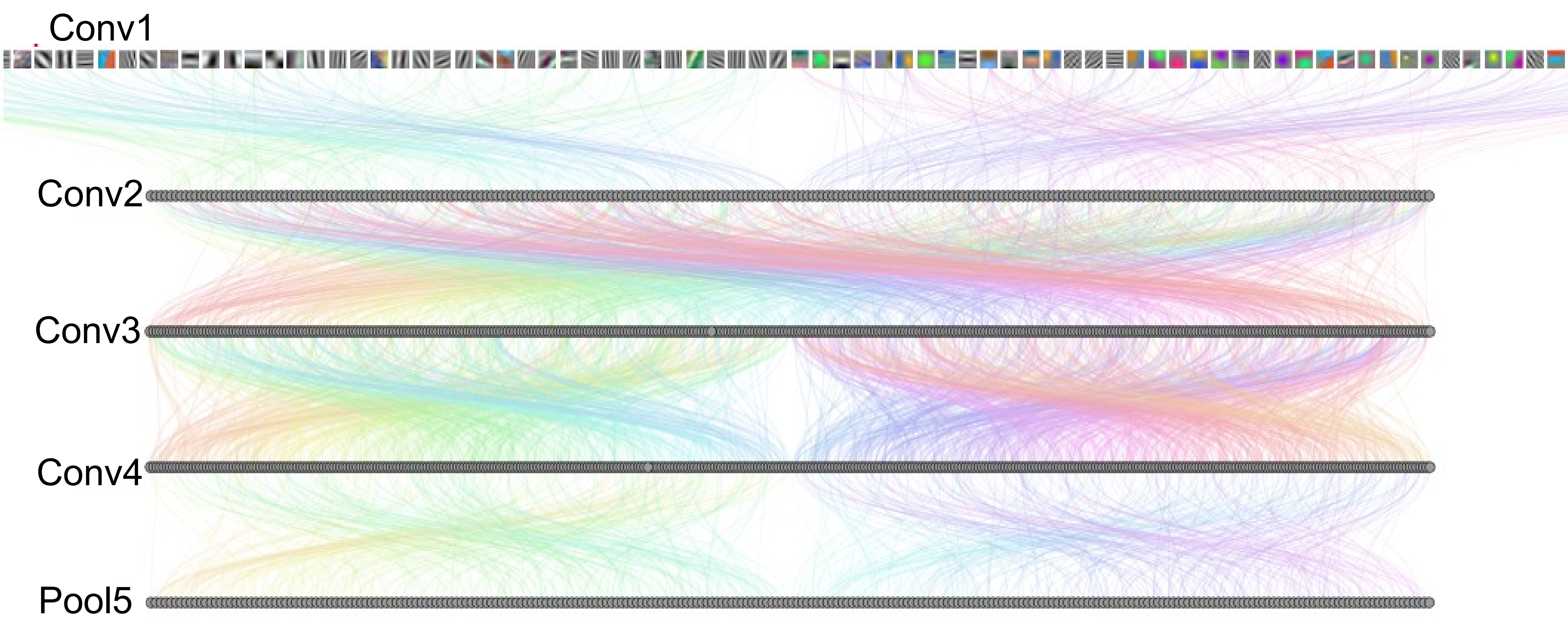


Figure from Olga Russakovsky ECCV'14 workshop

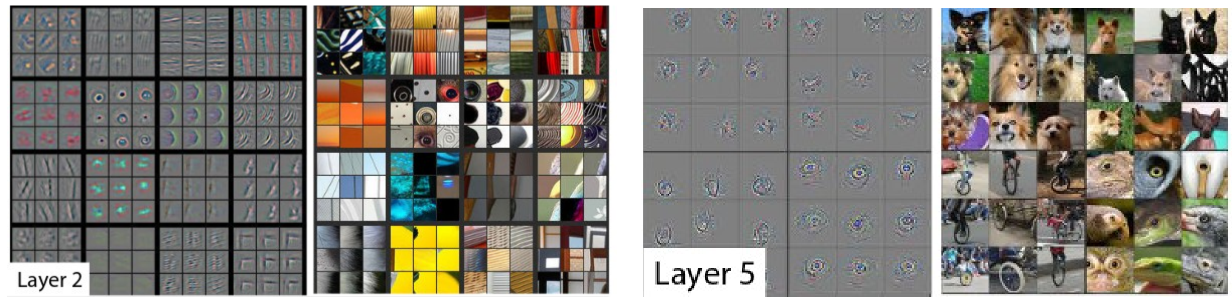
How Objects are Represented in CNN?



DrawCNN: visualizing the units' connections

How Objects are Represented in CNN?

Deconvolution

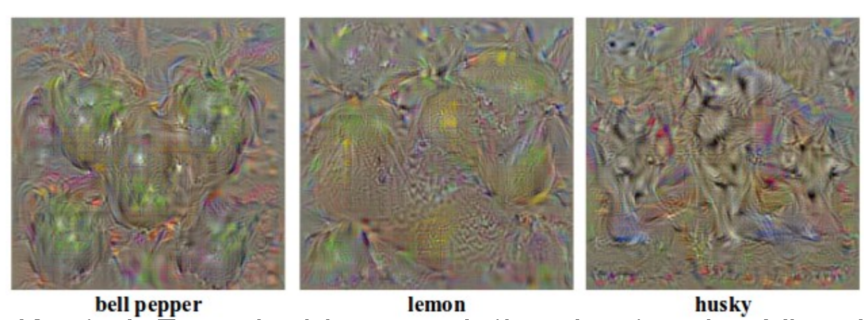


Zeiler, M. et al. Visualizing and Understanding Convolutional Networks,ECCV 2014.

Strong activation image

Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accu-rate object detection and semantic segmentation. CVPR 2014

Back-propagation



Simonyan, K. et al. Deep inside convolutional networks: Visualising image classification models and saliency maps. ICLR workshop, 2014

Object Representations in Computer Vision

Part-based models are used to represent objects and visual patterns.

-Object as a set of parts

-Relative locations between parts

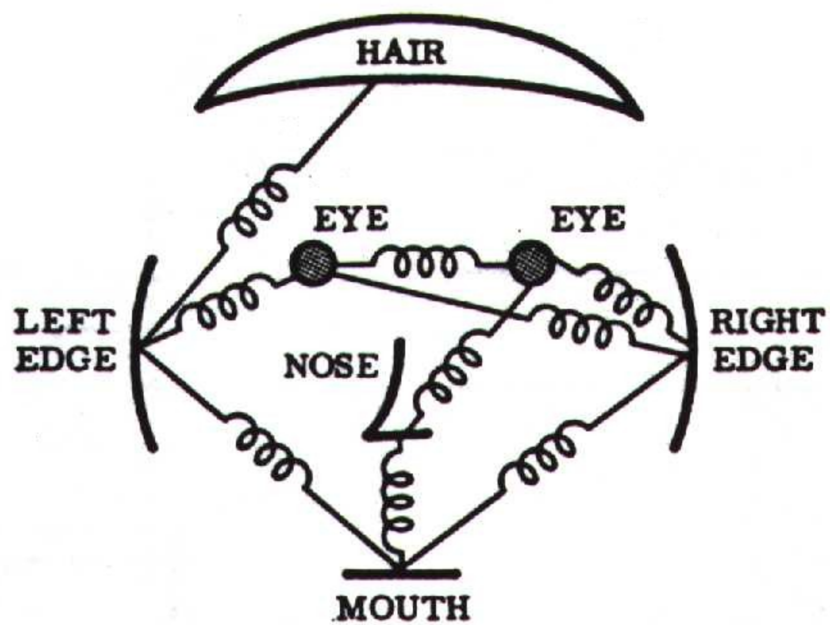


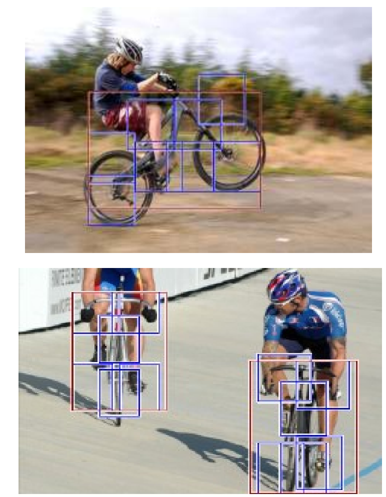
Figure from Fischler & Elschlager (1973)

Object Representations in Computer Vision

Constellation model

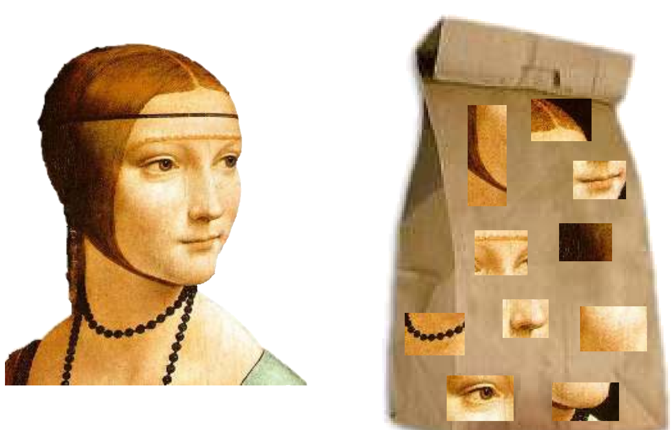
Weber, Welling & Perona (2000),
Fergus, Perona & Zisserman (2003)

Deformable Part model



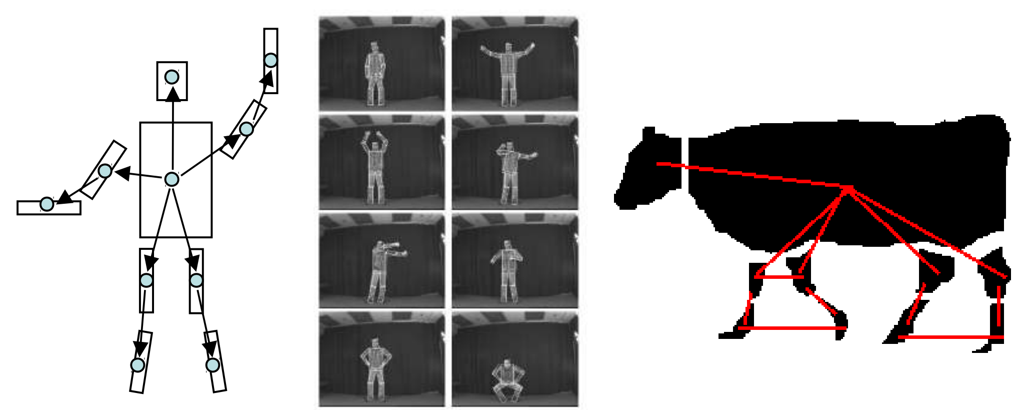
P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan (2010)

Bag-of-word model



Lazebnik, Schmid & Ponce(2003), Fei-Fei Perona (2005)

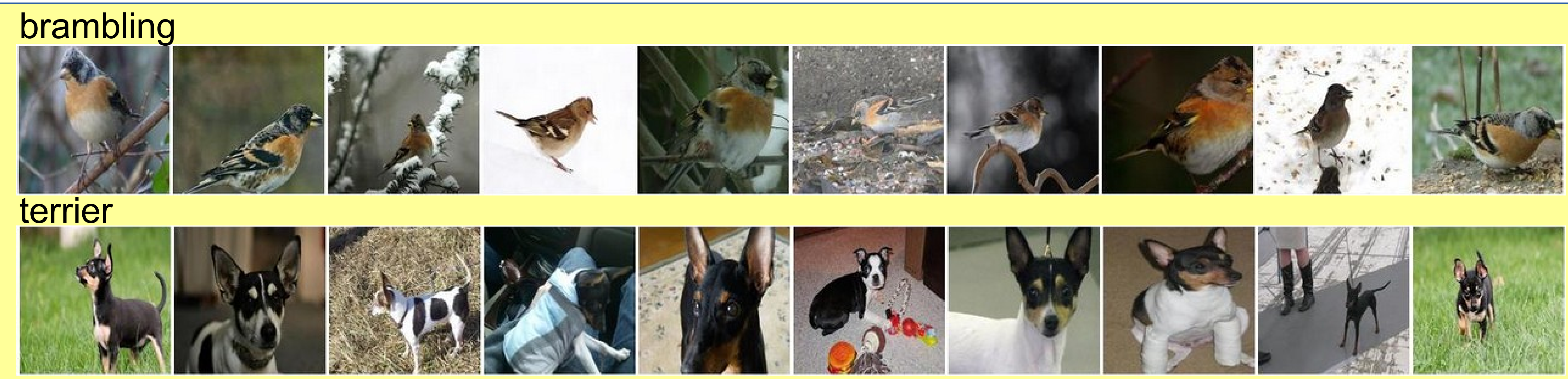
Class-specific graph model



Kumar, Torr and Zisserman (2005), Felzenszwalb & Huttenlocher (2005)

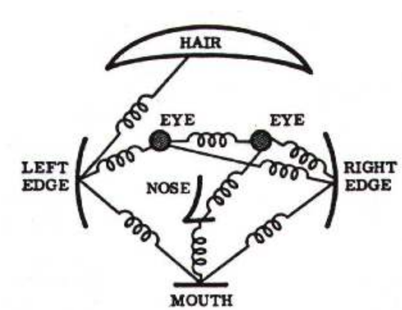
Learning to Recognize Objects

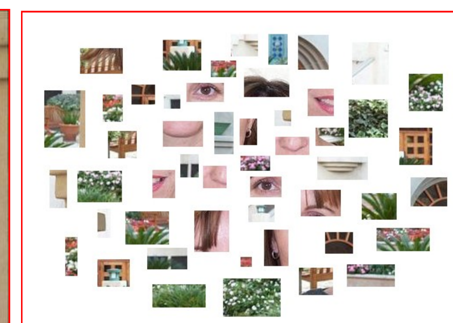
IMAGENET



Possible internal representations:

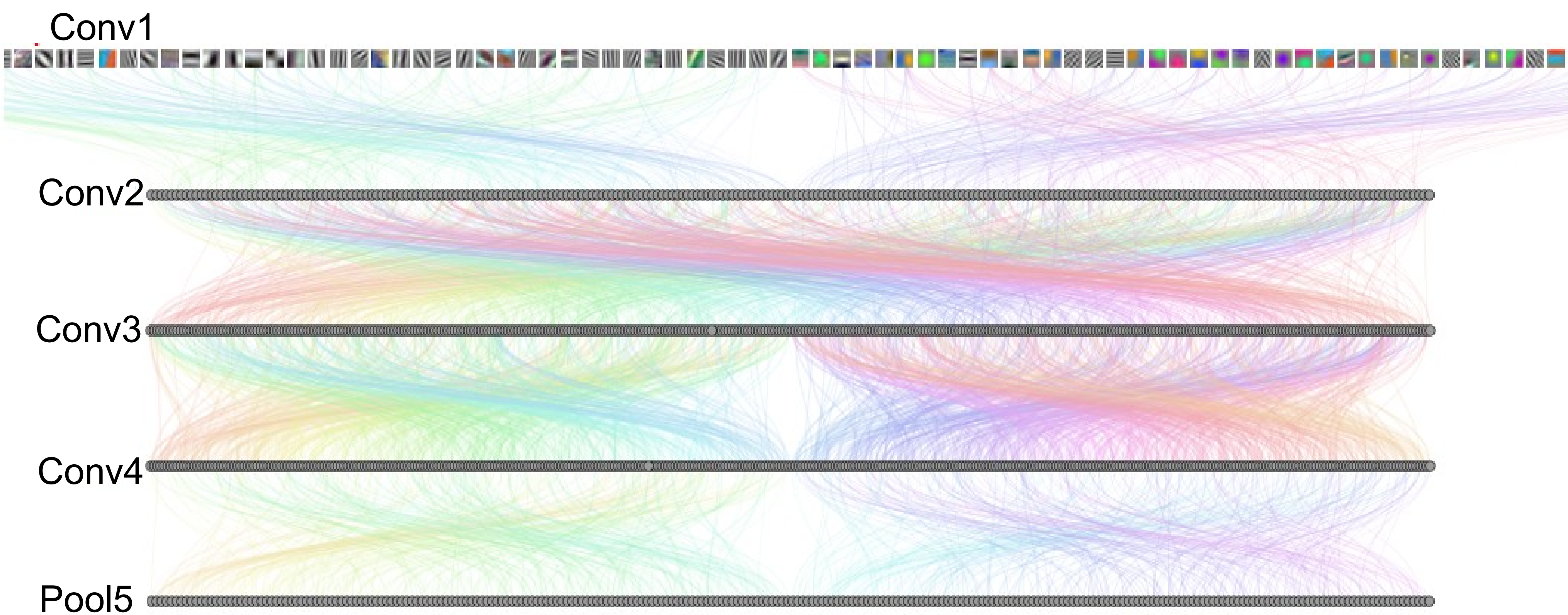
- Object parts
- Textures
- Attributes





How Objects are Represented in CNN?

CNN uses **distributed code** to represent objects.



Agrawal, et al. Analyzing the performance of multilayer neural networks for object recognition. ECCV, 2014
Szegedy, et al. Intriguing properties of neural networks.arXiv preprint arXiv:1312.6199, 2013.
Zeiler, M. et al. Visualizing and Understanding Convolutional Networks, ECCV 2014.

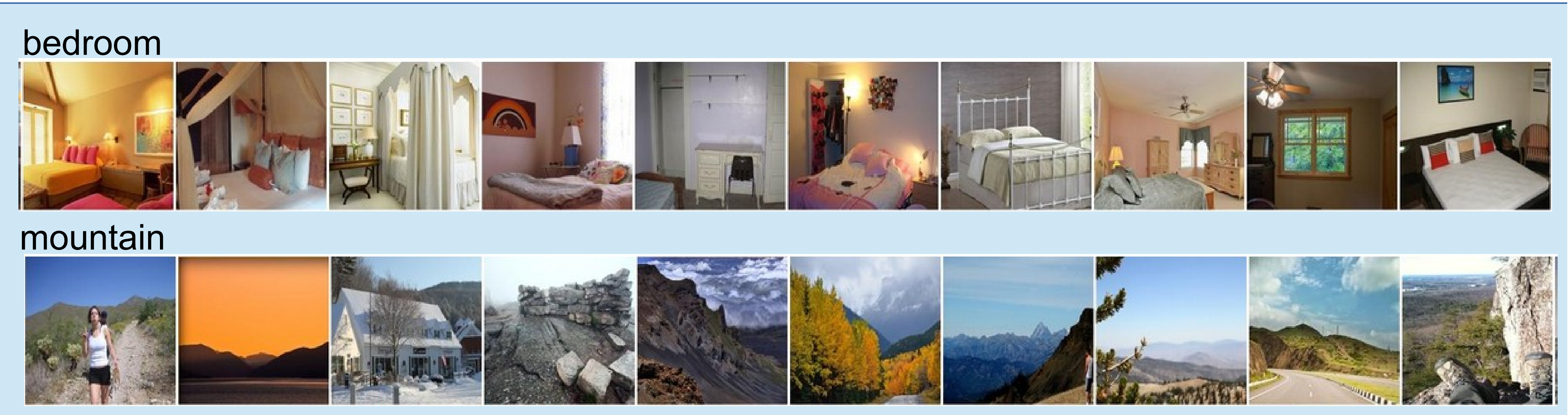
Scene Recognition

Given an image, predict which place we are in.

Bedroom

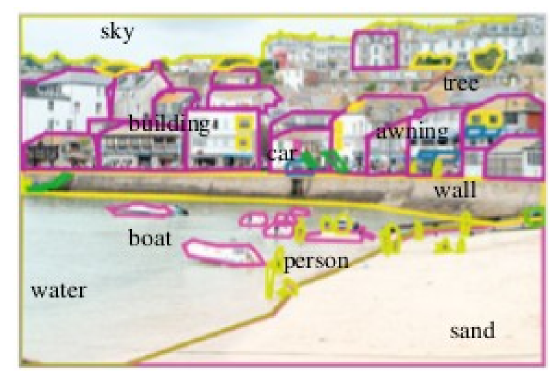
Harbor

Learning to Recognize Scenes



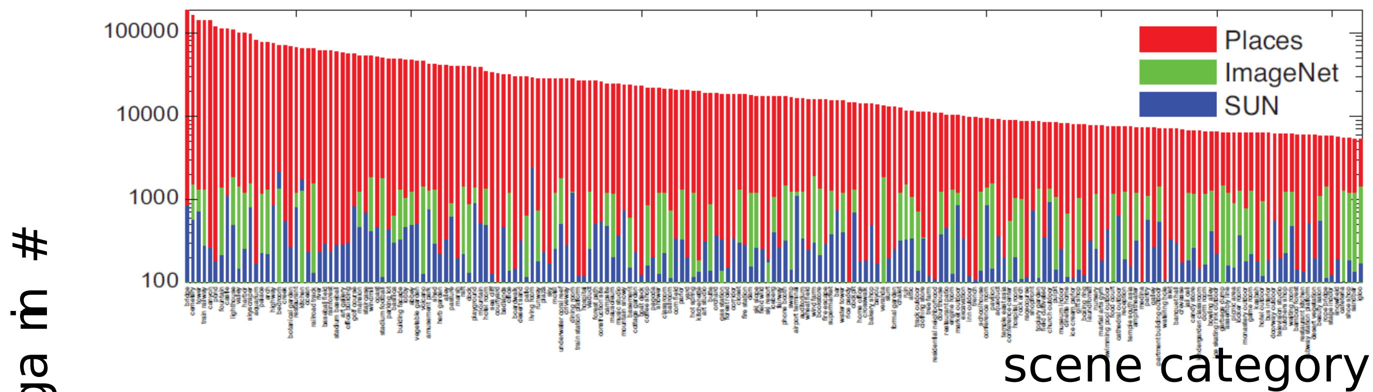
Possible internal representations:

- Objects (scene parts?)
- Scene attributes
- Object parts
- Textures



CNN for Scene Recognition

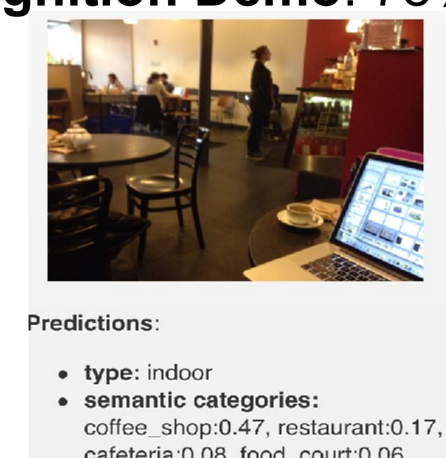
Places Database: 7 million images from 400 scene categories

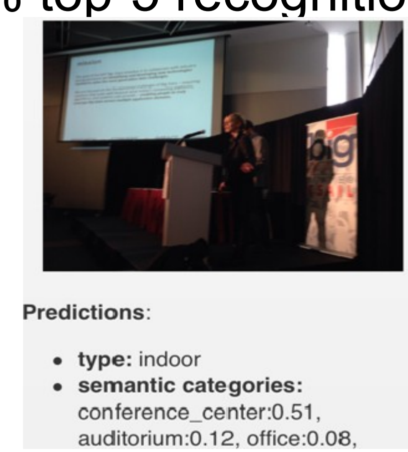


Places-CNN: AlexNet CNN on 2.5 million images from 205 scene categories.

	Places 205	SUN 205
Places-CNN	**50.0%**	**66.2%**
ImageNet CNN feature+SVM	40.8%	49.6%

Scene Recognition Demo: 78% top-5 recognition accuracy in the wild



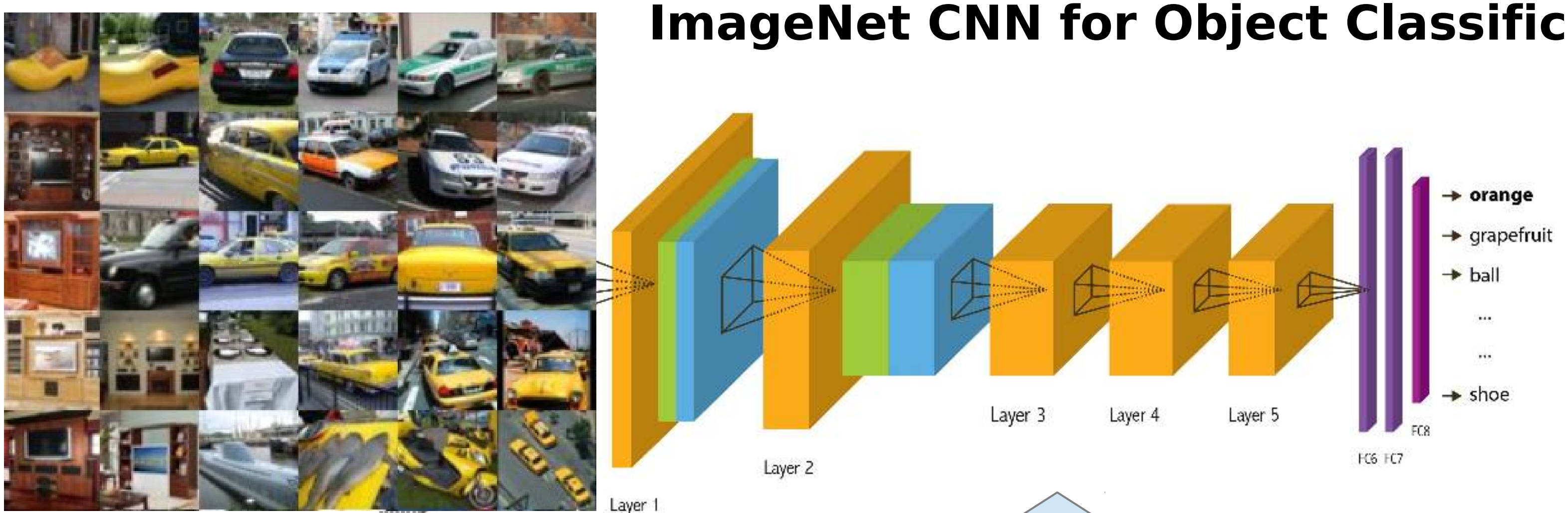


http://places.csail.mit.edu

Zhou, et al. NIPS, 2014.

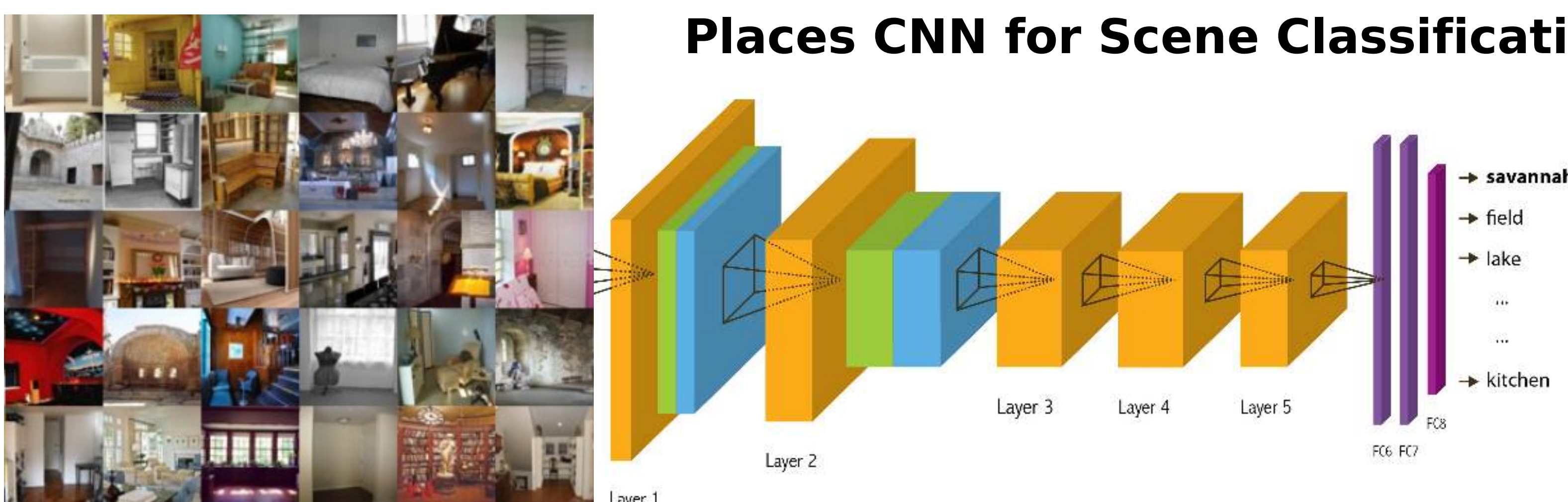
ImageNet CNN and Places CNN

ImageNet CNN for Object Classification



Same architecture: AlexNet

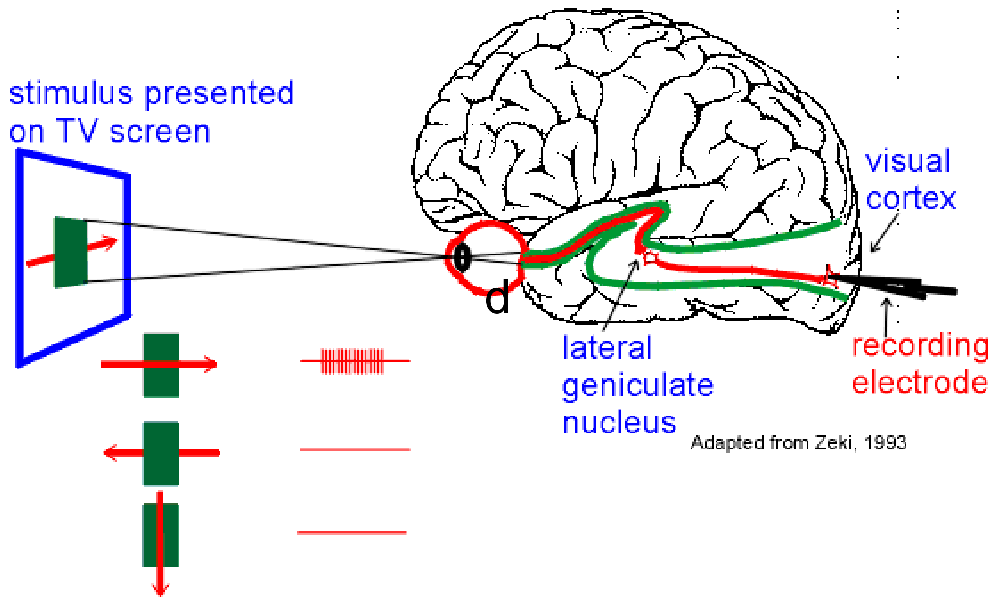
Places CNN for Scene Classification



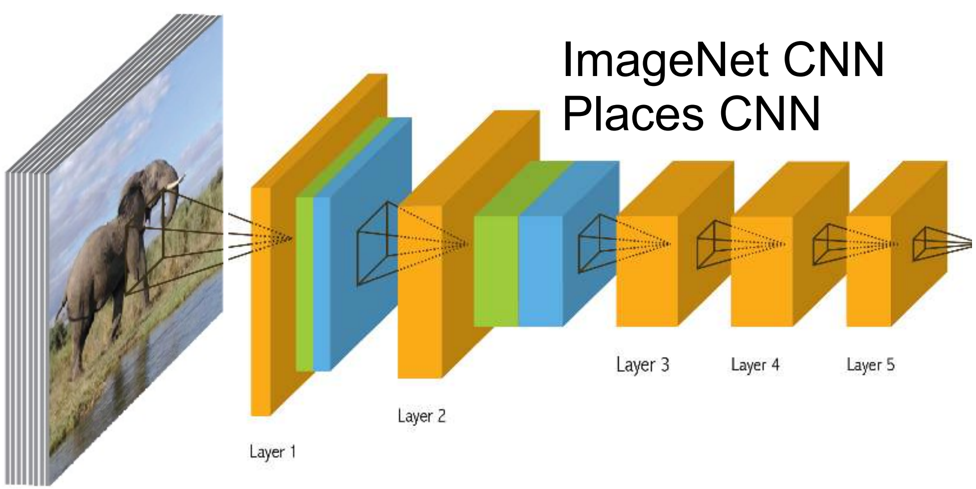
Places

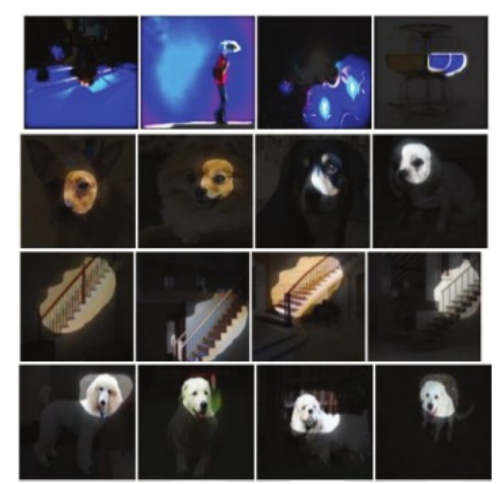
Data-Driven Approach to Study CNN

Neuroscientists study brain



200,000 image stimuli of objects and scene categories (ImageNet TestSet+SUN database)

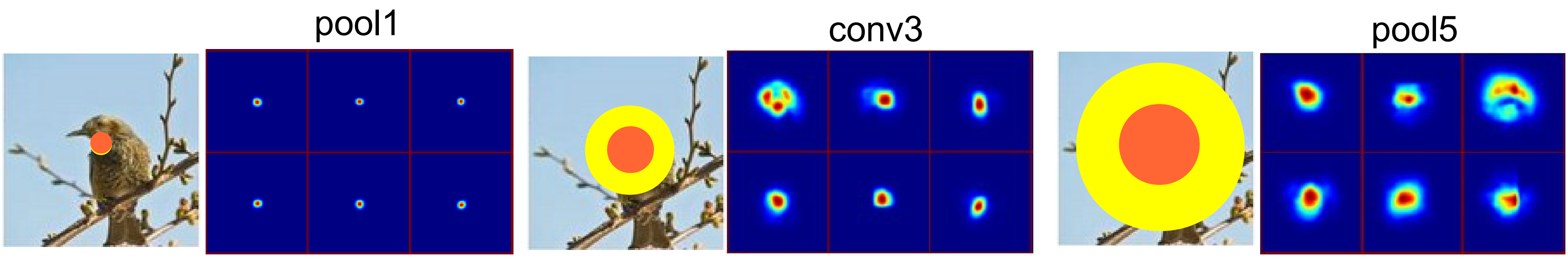




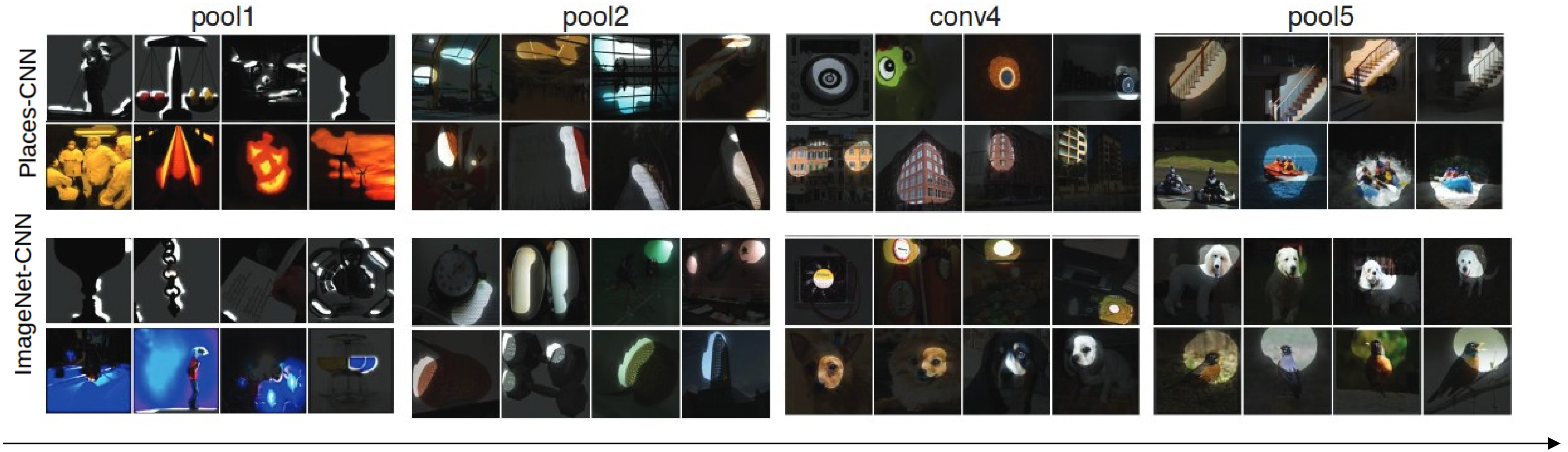
Estimating the Receptive Fields

Estimated receptive fields

Actual size of RF is much smaller than the theoretic size



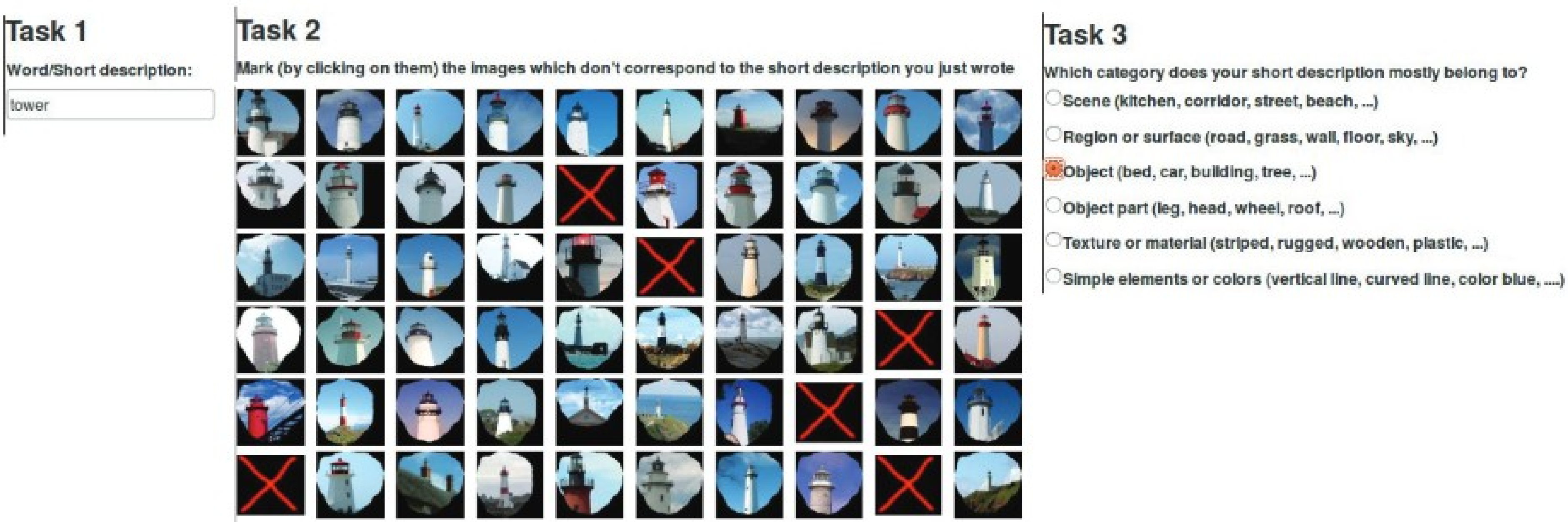
Segmentation using the RF of Units



More semantically meaningful

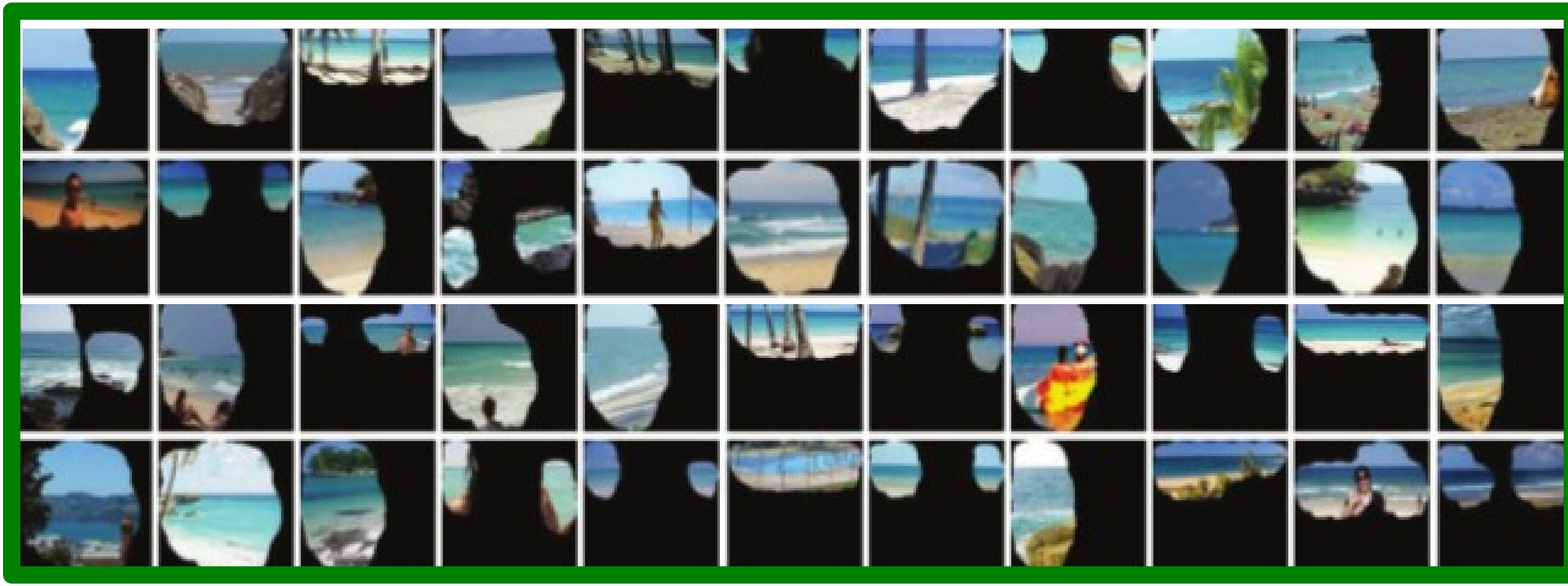
Annotating the Semantics of Units

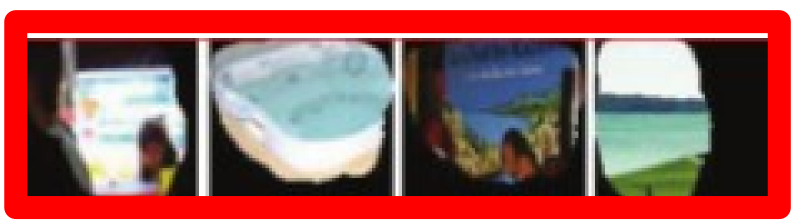
Top ranked segmented images are cropped and sent to Amazon Turk for annotation.



Annotating the Semantics of Units

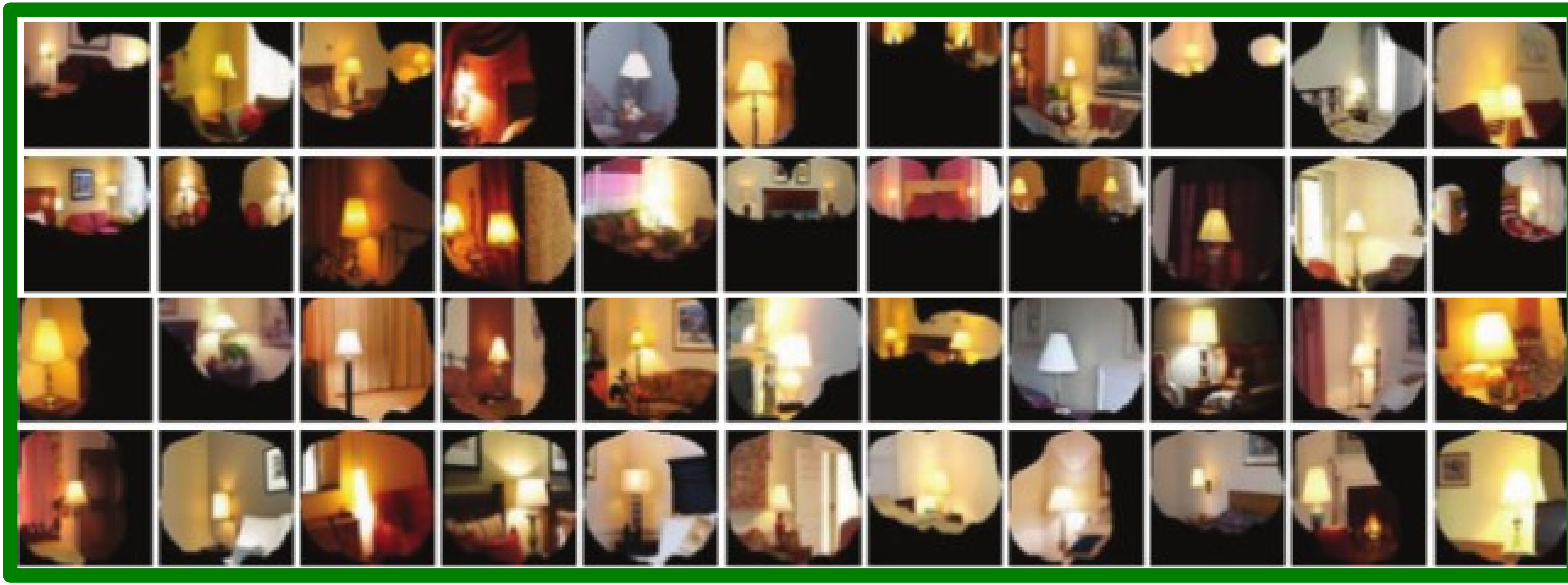
Pool5, unit 76; Label: ocean; Type: scene; Precision: 93%

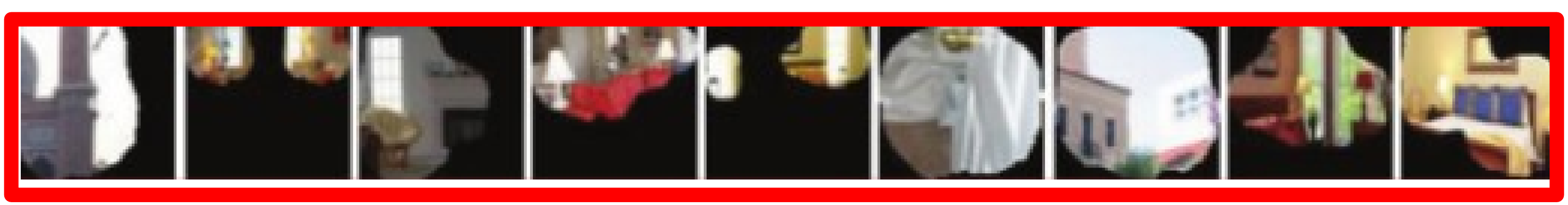




Annotating the Semantics of Units

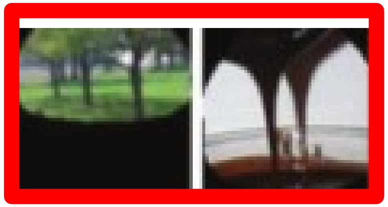
Pool5, unit 13; Label: Lamps; Type: object; Precision: 84%





Annotating the Semantics of Units

Pool5, unit 77; Label:legs; Type: object part; Precision: 96%



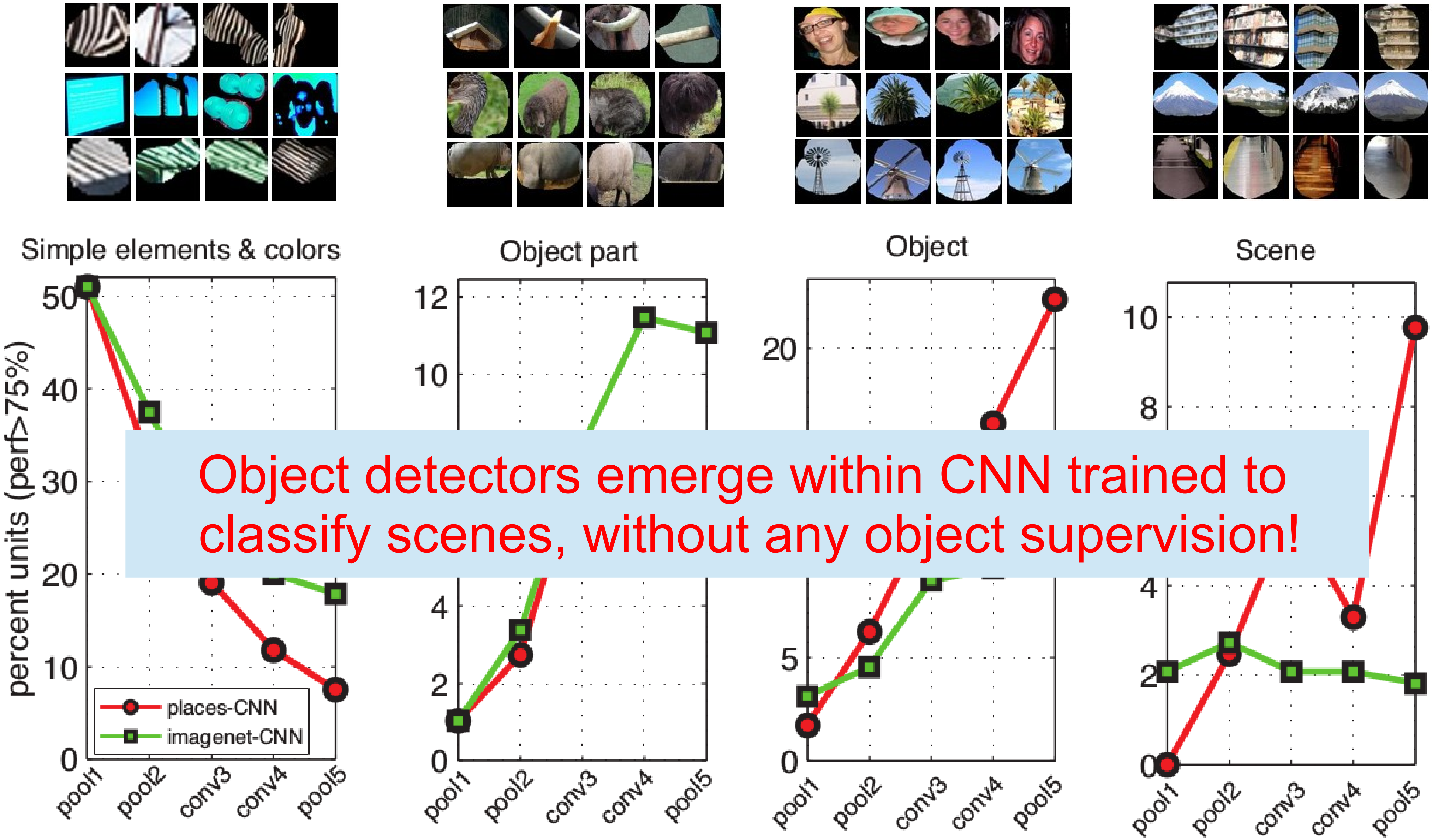
Annotating the Semantics of Units

Pool5, unit 112; Label: pool table; Type: object; Precision: 70%

Annotating the Semantics of Units

Pool5, unit 22; Label: dinner table; Type: scene; Precision: 60%

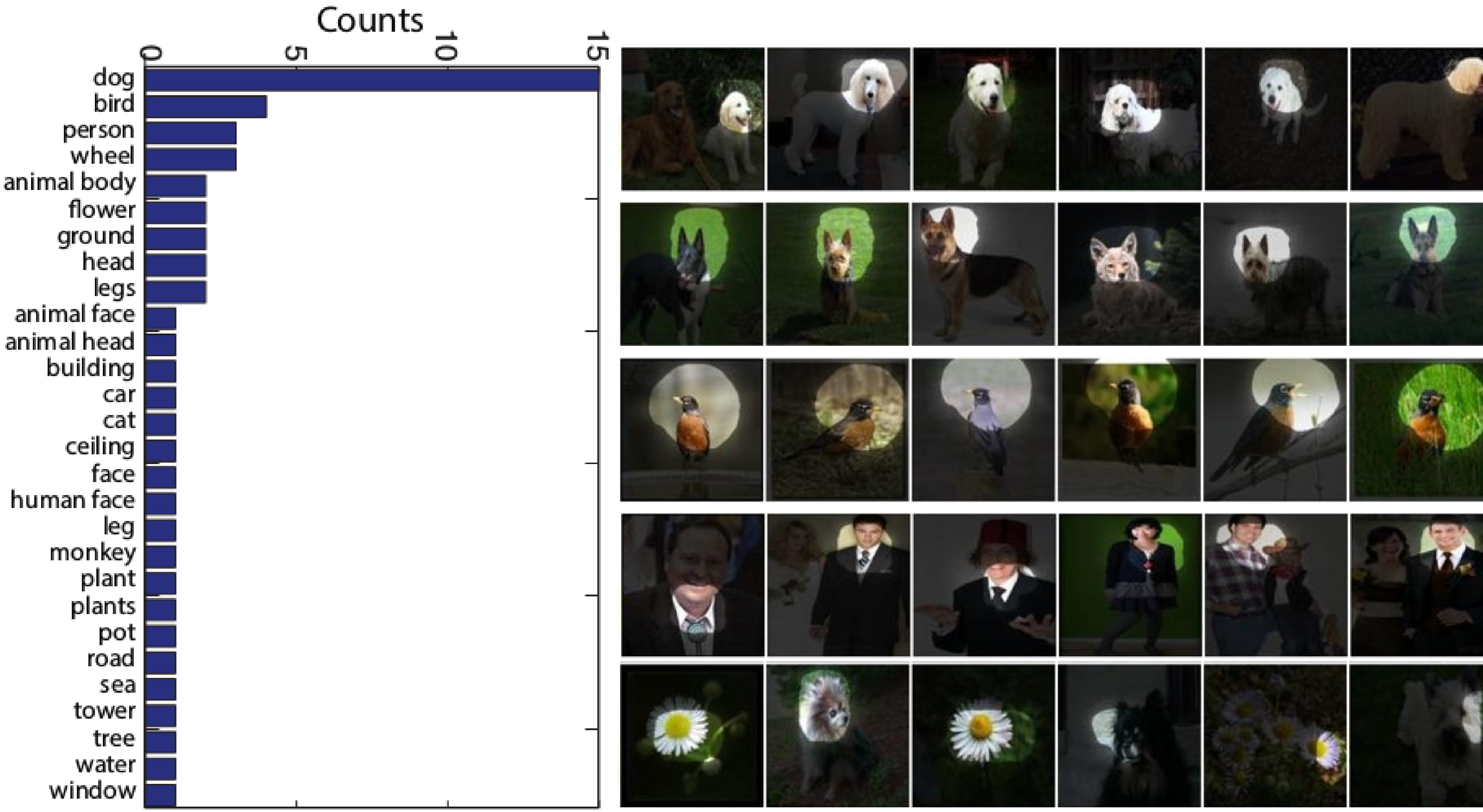
Distribution of Semantic Types at Each Layer



Object detectors emerge within CNN trained to classify scenes, without any object supervision!

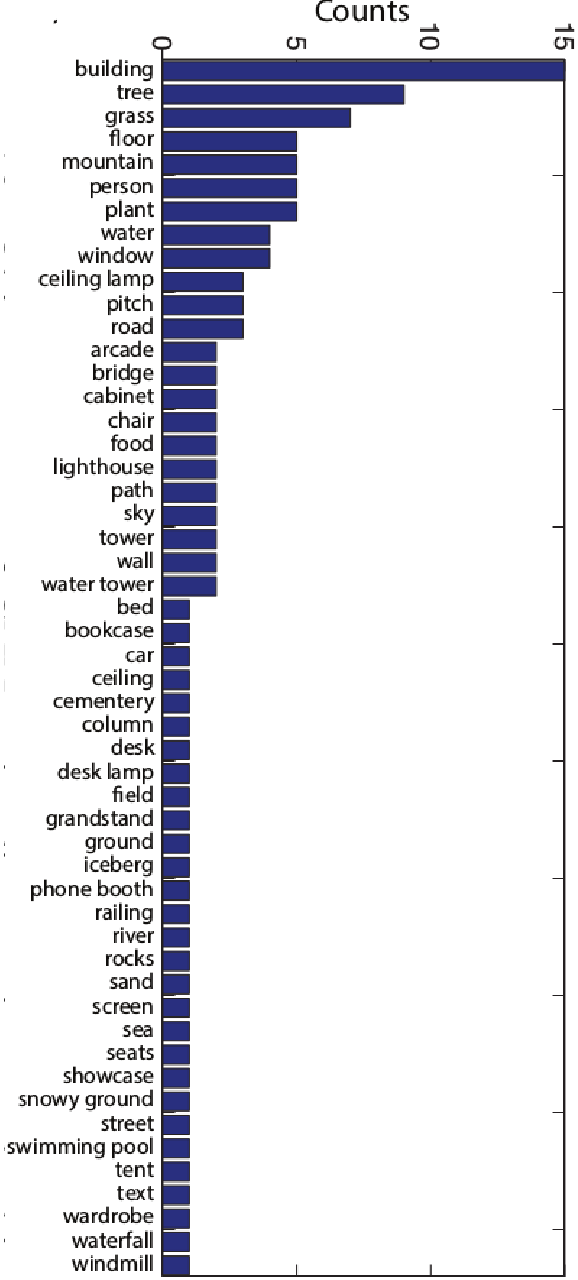
Histogram of Emerged Objects in Pool5

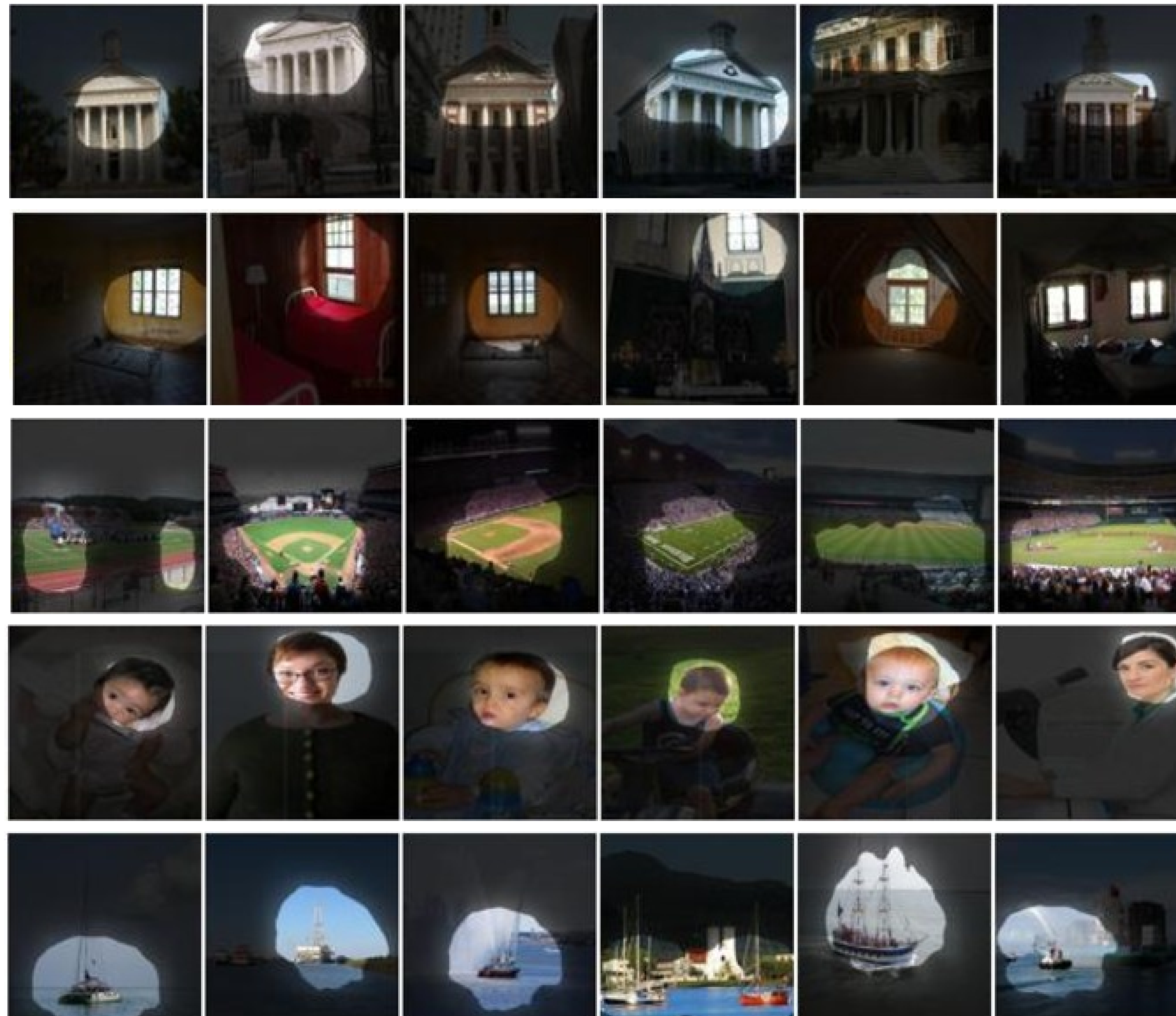
ImageNet-CNN (59/256)

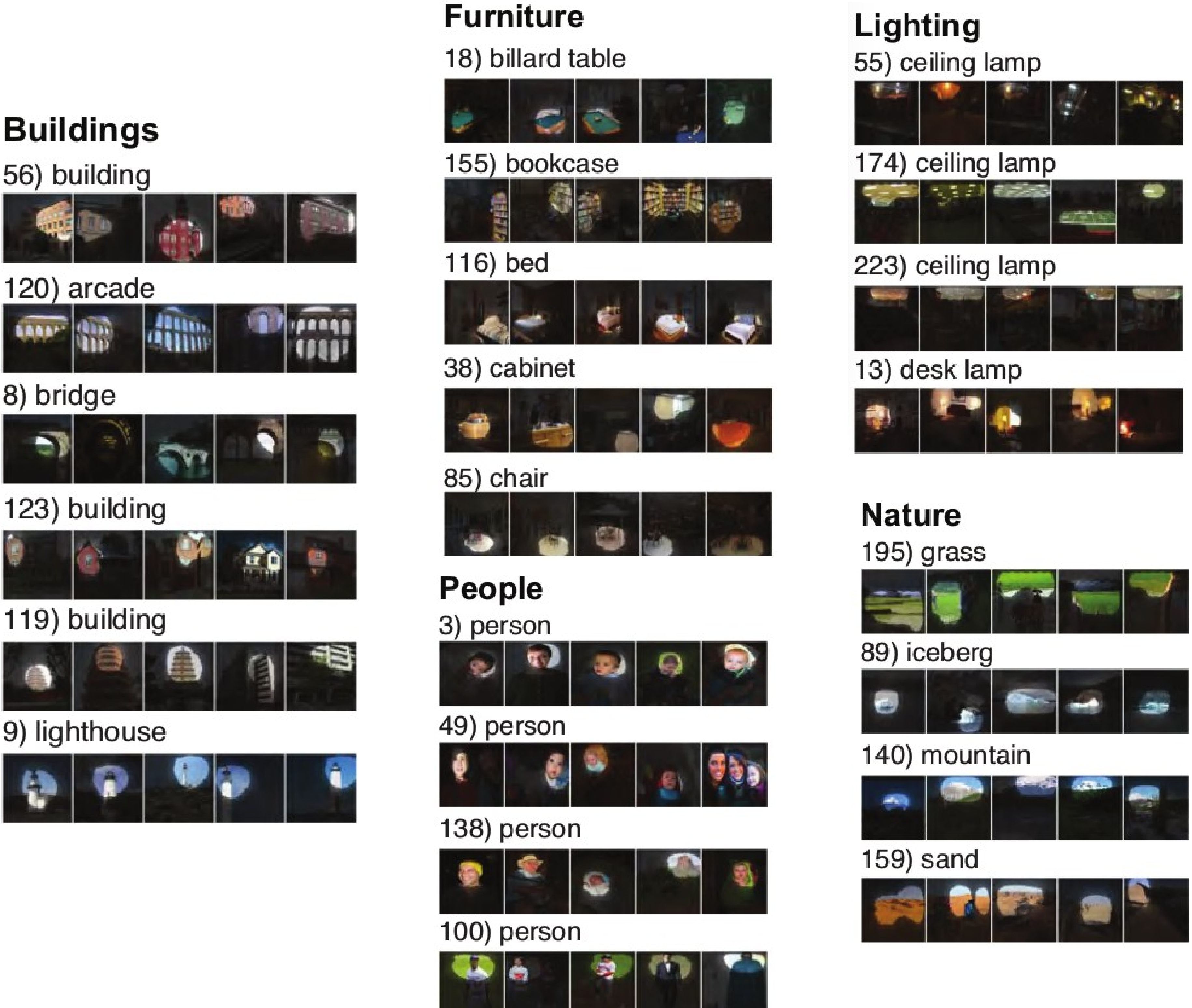


Histogram of Emerged Objects in Pool5

Places-CNN (151/256)



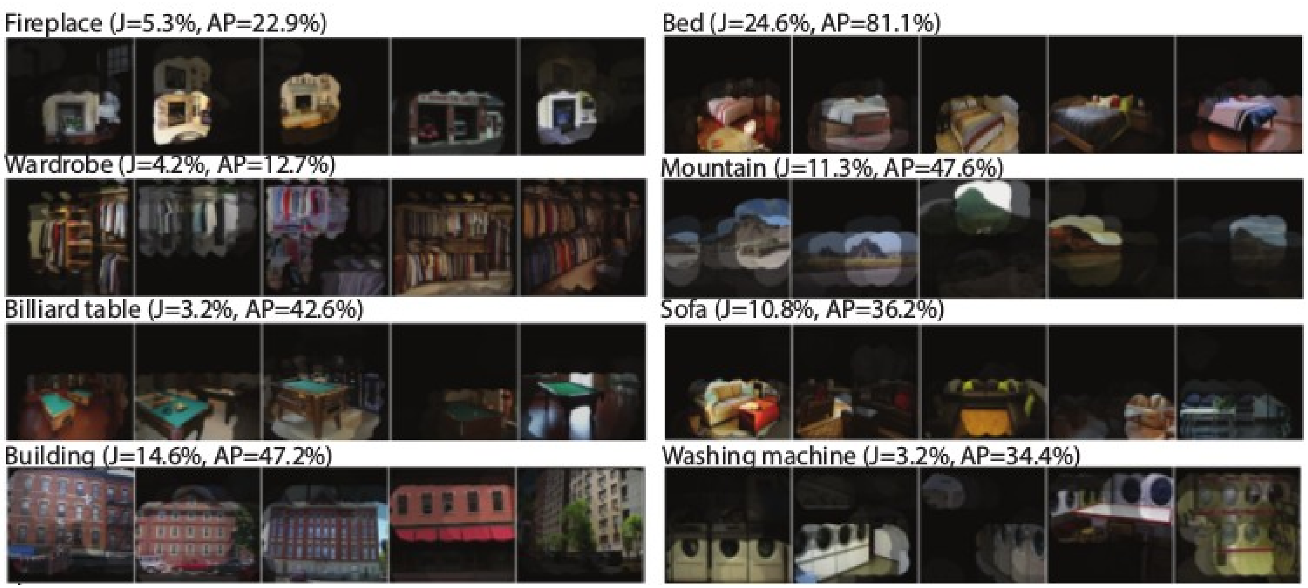


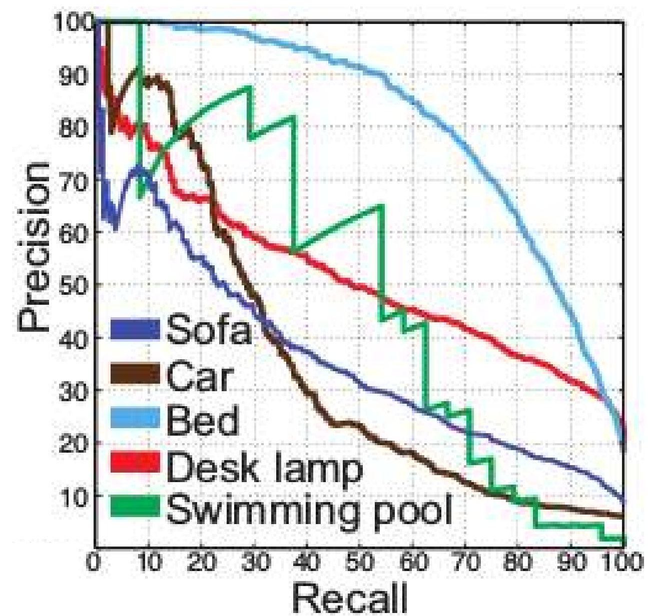


Buildings
56) building
120) arcade
8) bridge
123) building
119) building
9) lighthouse
Furniture
18) billard table
155) bookcase
116) bed
38) cabinet
85) chair
People
3) person
49) person
138) person
100) person
Lighting
55) ceiling lamp
174) ceiling lamp
223) ceiling lamp
13) desk lamp
Nature
195) grass
89) iceberg
140) mountain
159) sand

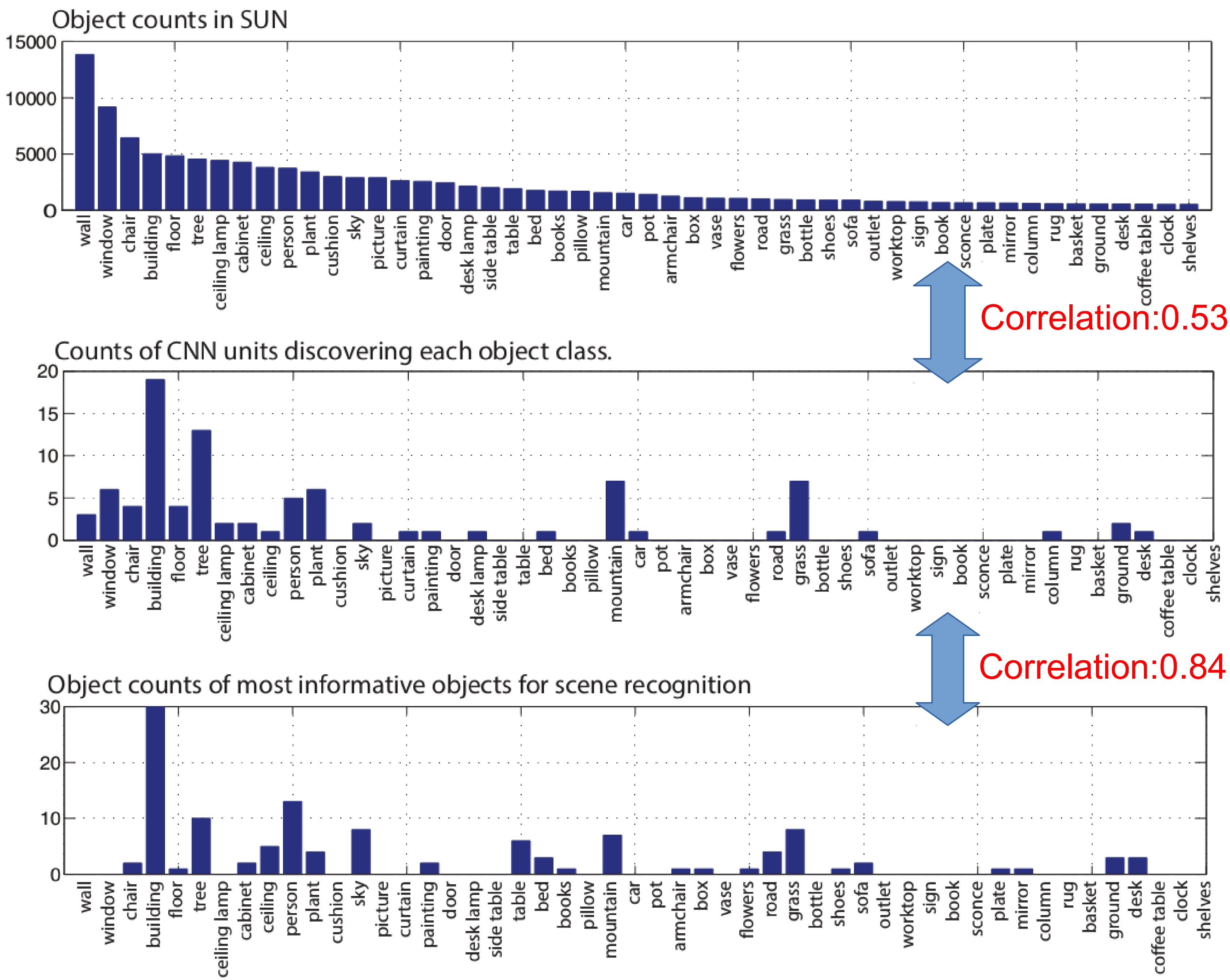
Evaluation on SUN Database

Evaluate the performance of the emerged object detectors





Evaluation on SUN Database



Conclusion

We show that object detectors emerge inside a CNN trained to classify scenes, without any object supervision.

Object detectors for free!

Places database, Places CNN, and unit annotations could be downloaded at

http://places.csail.mit.edu