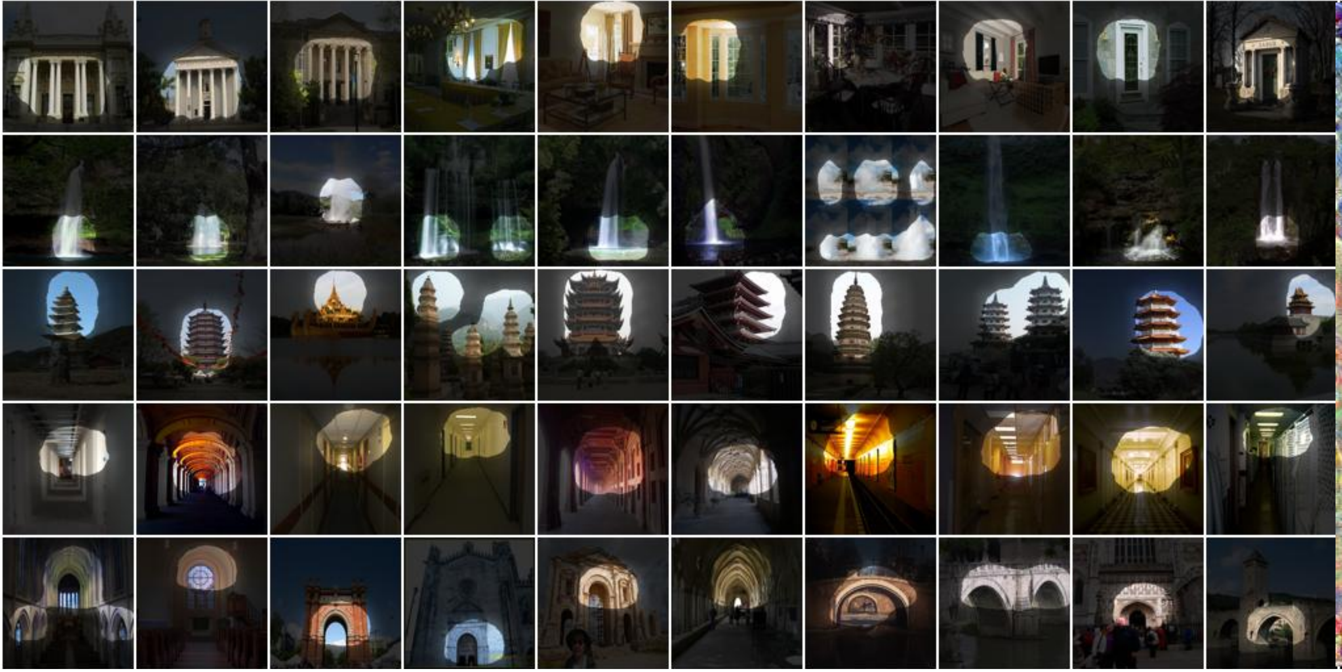


Understand and Leverage the Internal Representations of Convolutional Neural Networks

Bolei Zhou
MIT

CNN for Image Classification

Large-scale image classification result on ImageNet

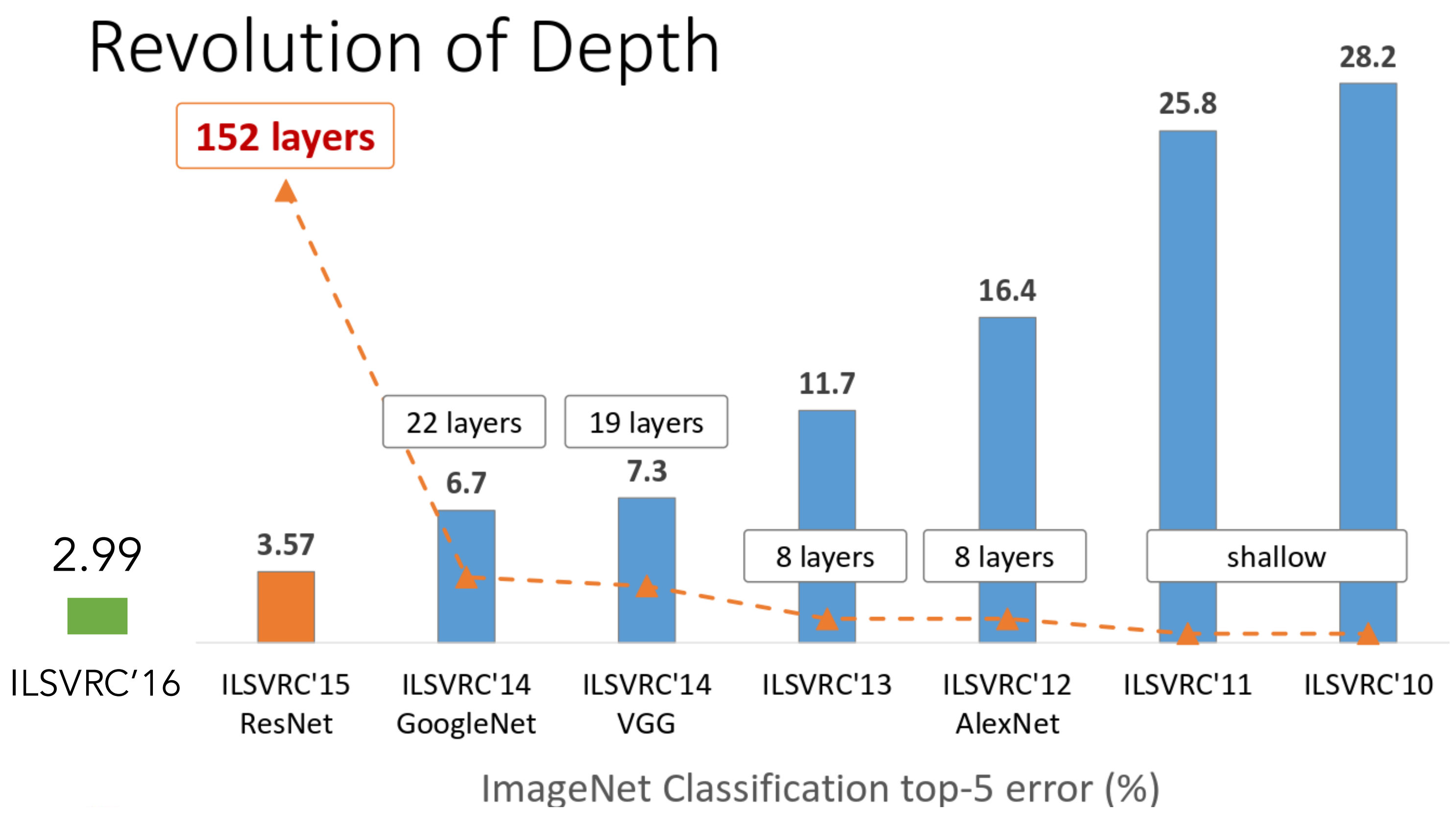
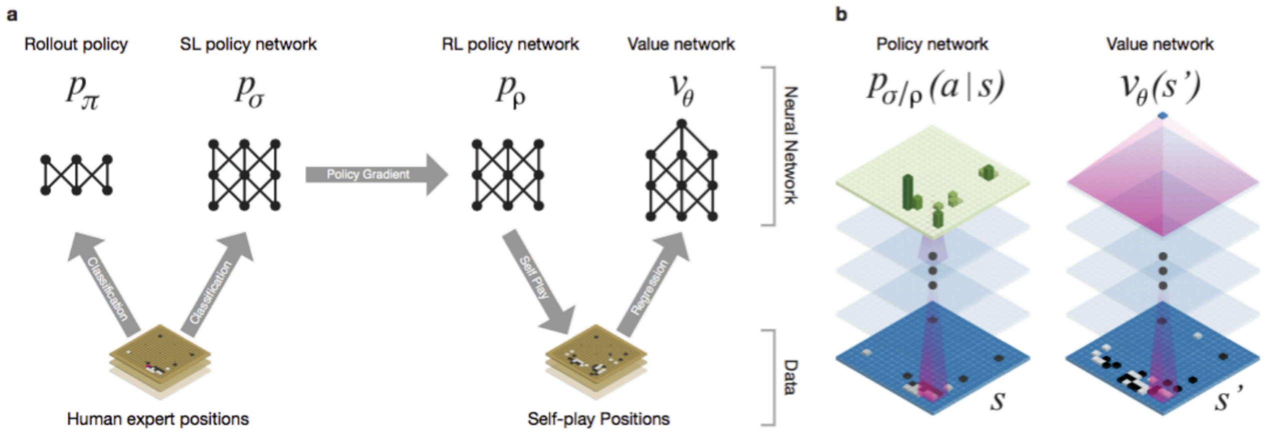


Figure from Kaiming He's ICCV workshop

CNN for General AI

- Alpha Go

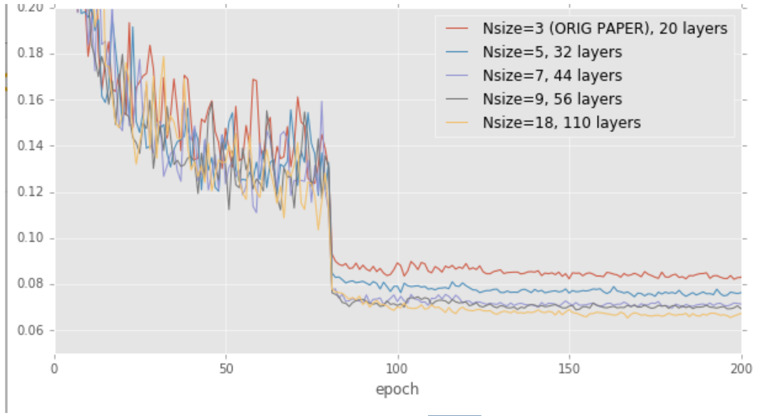


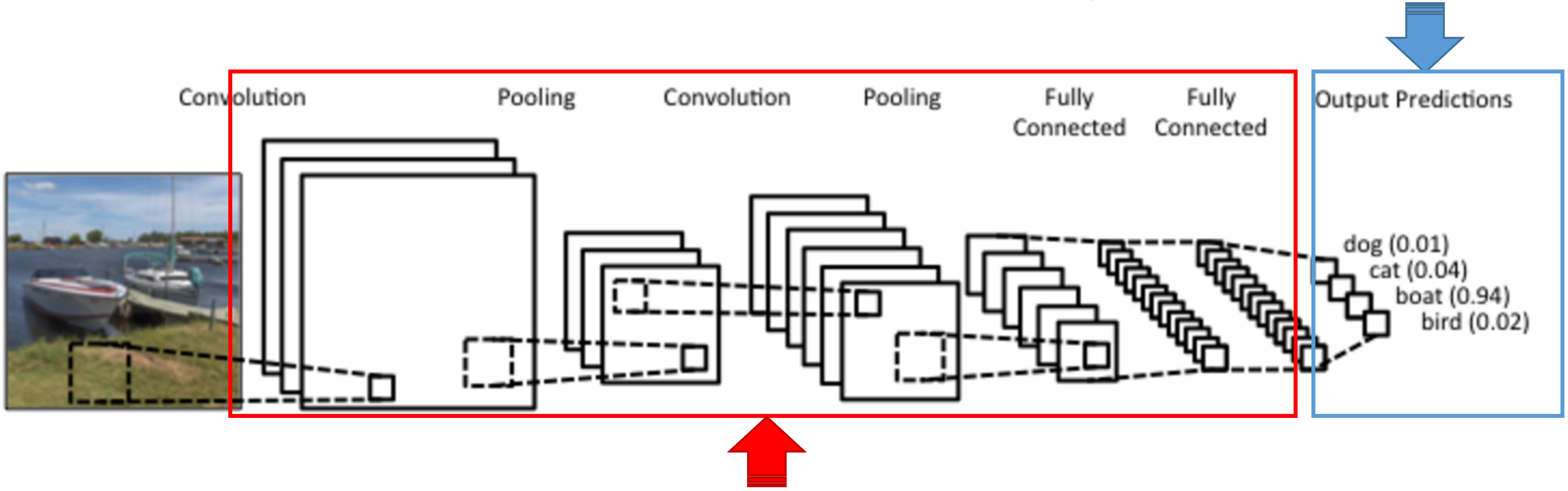
- SL policy network is 13 layer-CNN
- Training: 29.4 million positions from 160,000 human professional games.
- CNN beats human professional, can we discover the inside knowledge?

Mastering the game of Go with deep neural networks and tree search
D Silver et al. Nature, 2016

final output is a small part of the story

Secret of CNN





Understand and leverage the internal units/representation

Outline

- Visualizing and annotating the internal units
- Application: weakly supervised localization

Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR'15
Zhou et al. Learning Deep Features for Discriminative Localization. CVPR'16
Bau*, and Zhou*, et al. Network Dissection. CVPR'17

Joint work with: David Bau, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba

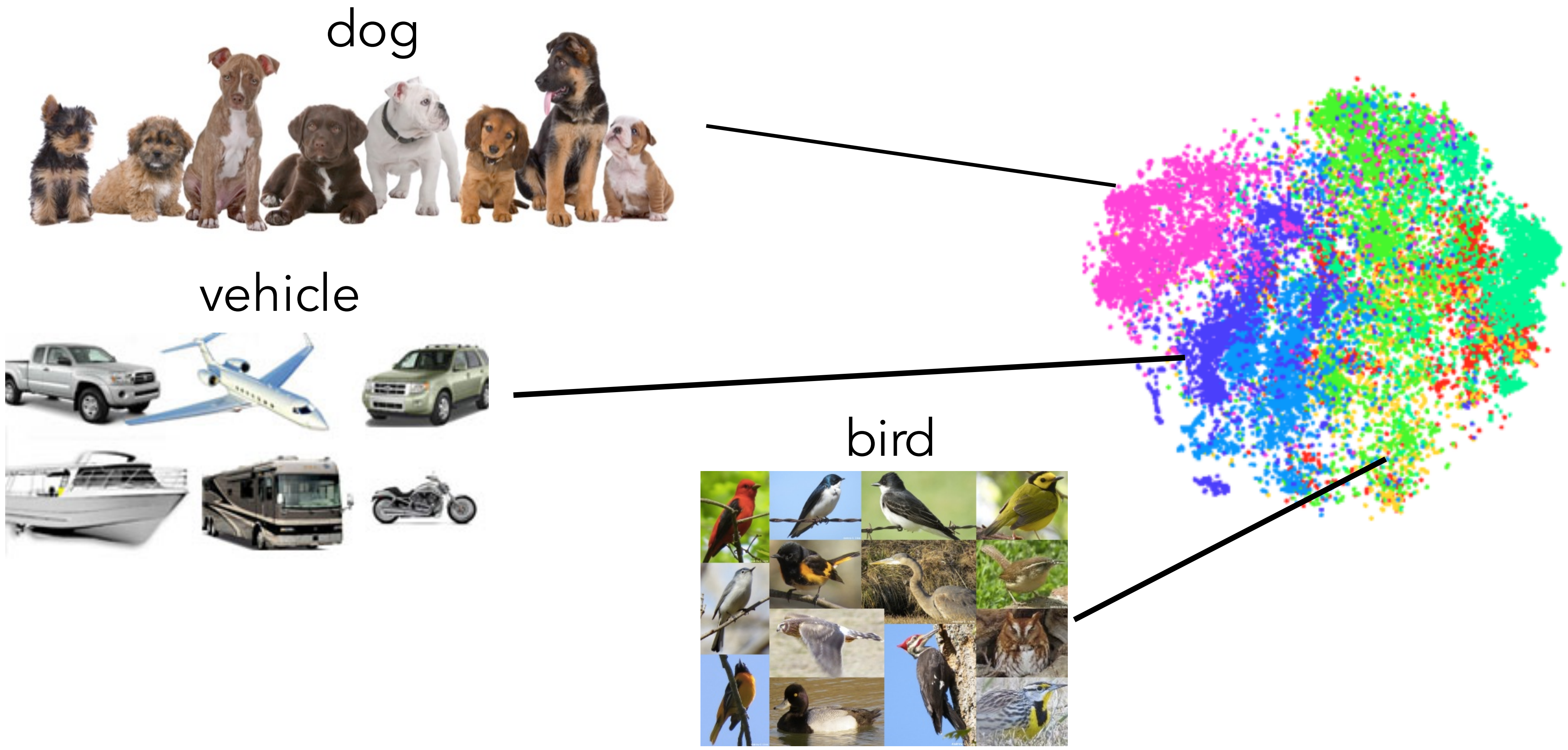
Object Representations in Computer Vision

dog

bird

vehicle

Object Representations in Computer Vision

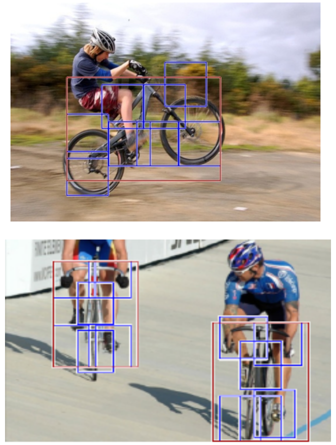


Object Representations in Computer Vision

Constellation model

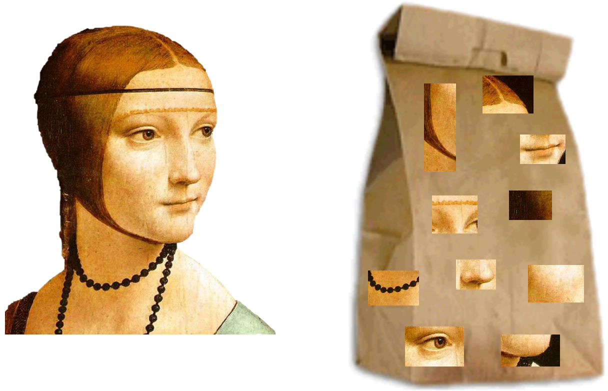
Weber, Welling & Perona (2000),
Fergus, Perona & Zisserman (2003)

Deformable Part model



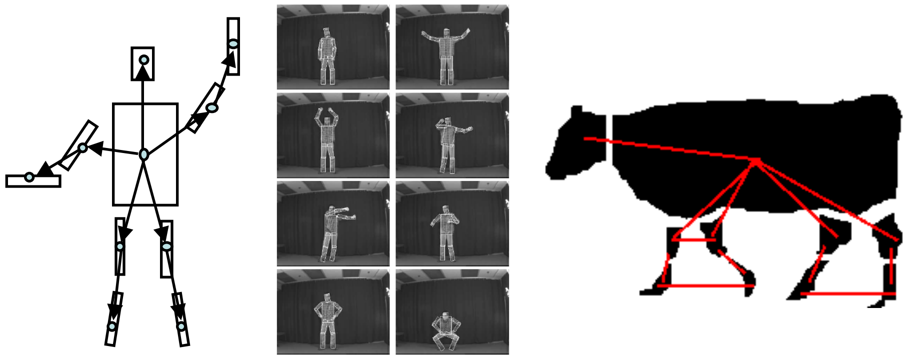
P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan (2010)

Bag-of-word model



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

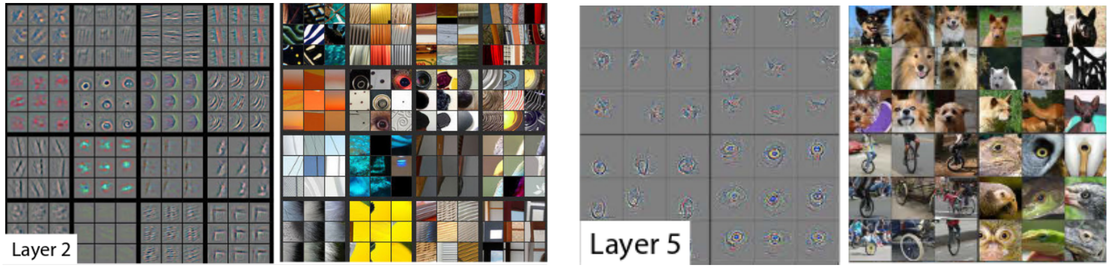
Class-specific graph model



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

Object Representations in CNN

Deconvolution

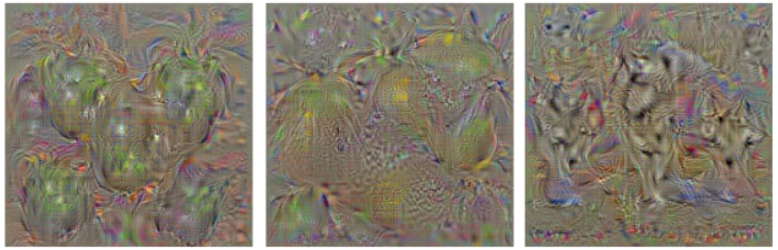


Zeiler, M. et al. Visualizing and Understanding Convolutional Networks,ECCV 2014.

Strong activation image

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

Back-propagation

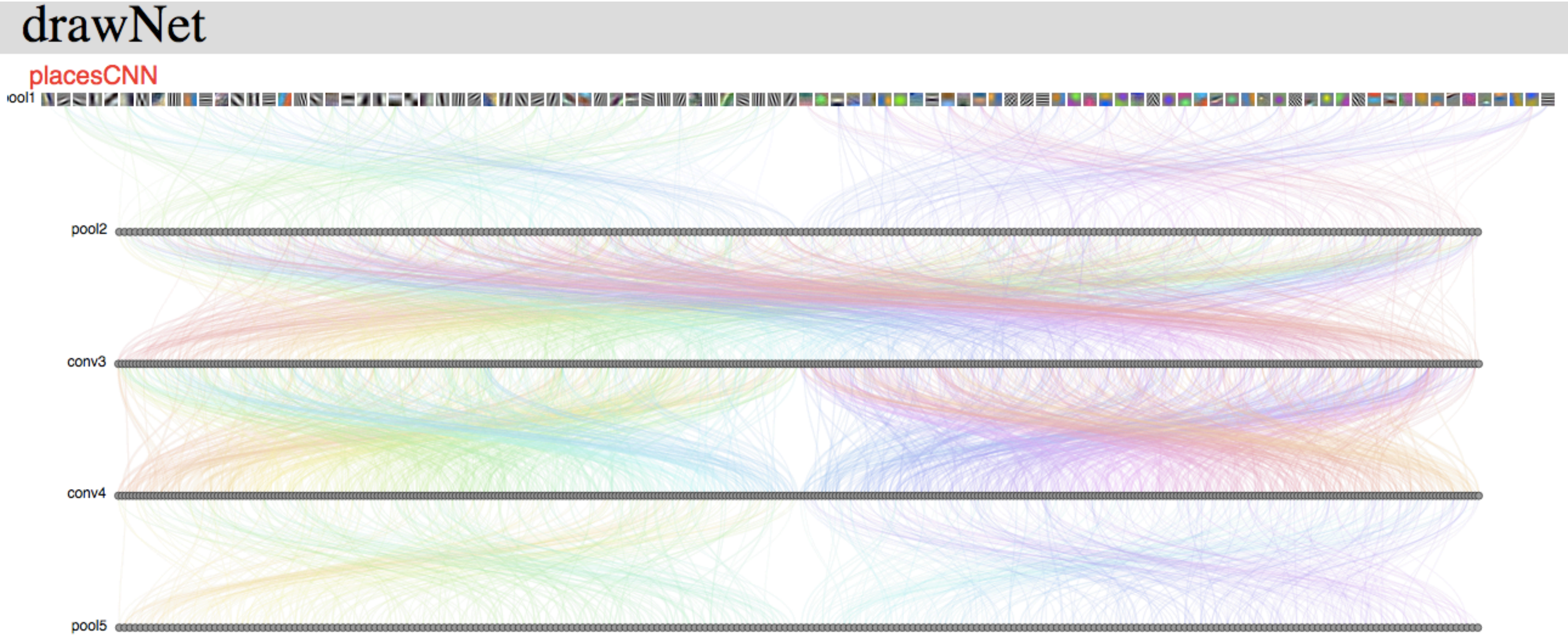


bell pepper lemon husky

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

Object Representations in CNN

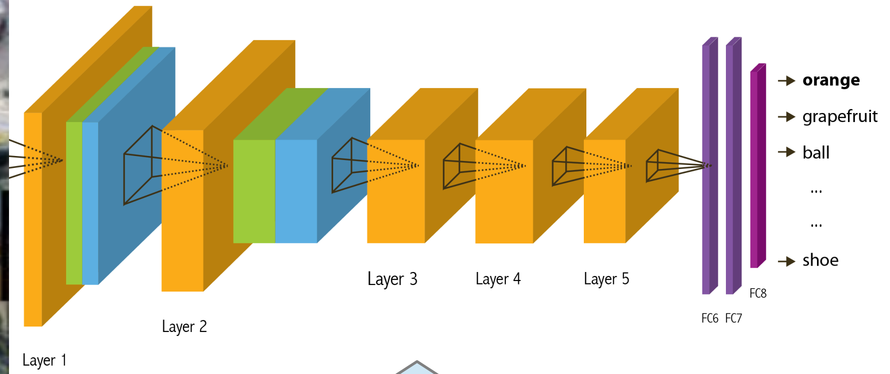
http://people.csail.mit.edu/torralba/research/drawCNN/drawNet.html



A Comparison Study on CNNs

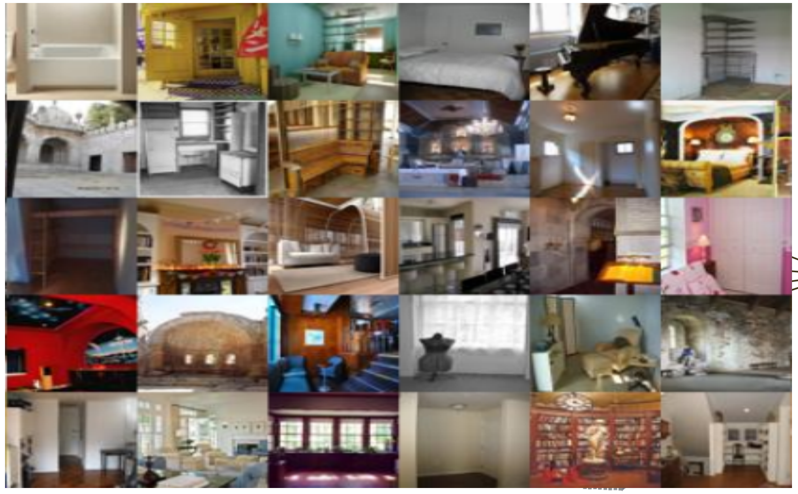
IMAGENET

ImageNet CNN for Object Classification

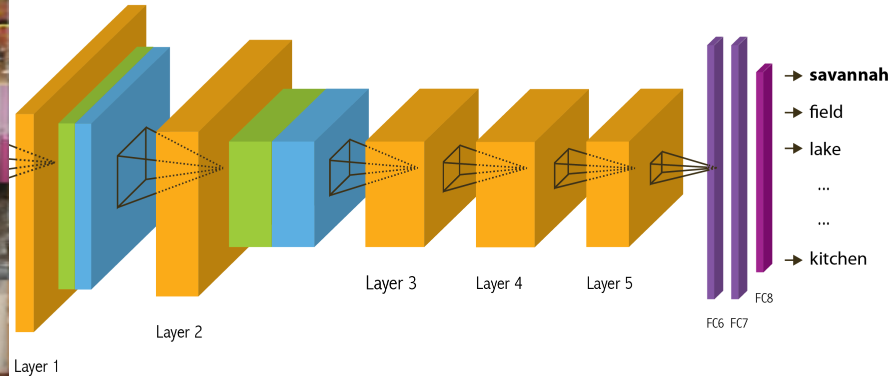


Same architecture: AlexNet

places
THE SCENE RECOGNITION DATABASE



Places CNN for Scene Classification



Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR'15

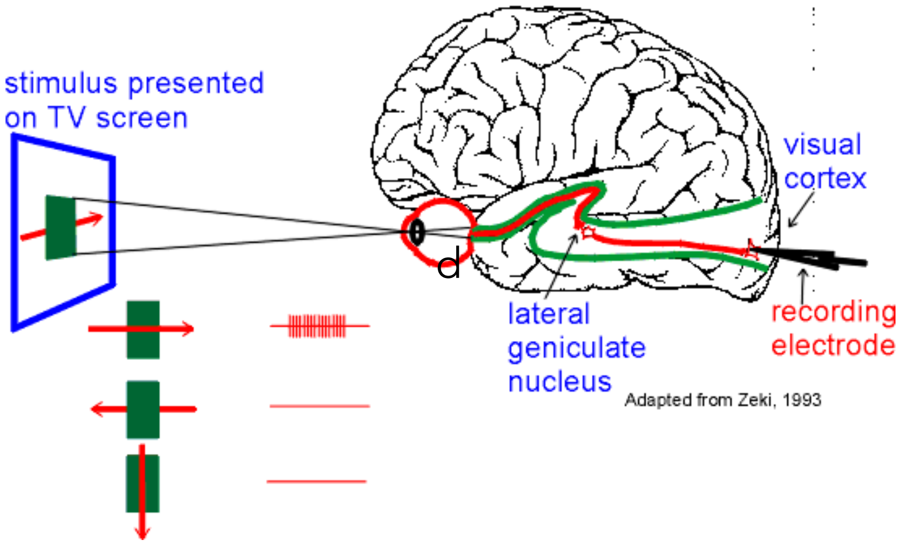
Places: Large-scale Scene Recognition Database

- Places contains 10 million images from ~400 scene categories.
- Data and models are available at http://places.csail.mit.edu

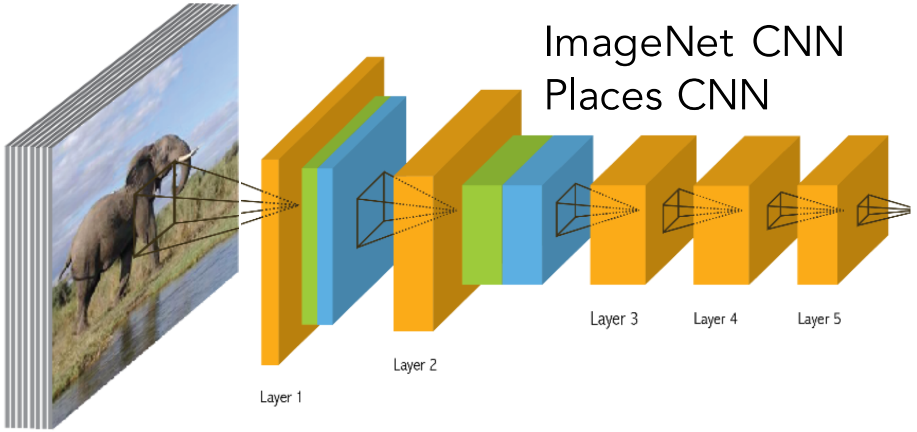
Zhou, et al. "Learning Deep Features for Scene Recognition using Places Database." NIPS'14

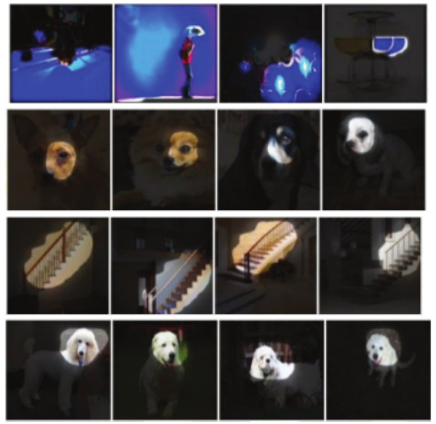
Data-Driven Approach to Visualize CNN

Neuroscientists study brain



200,000 image stimuli of objects and scenes



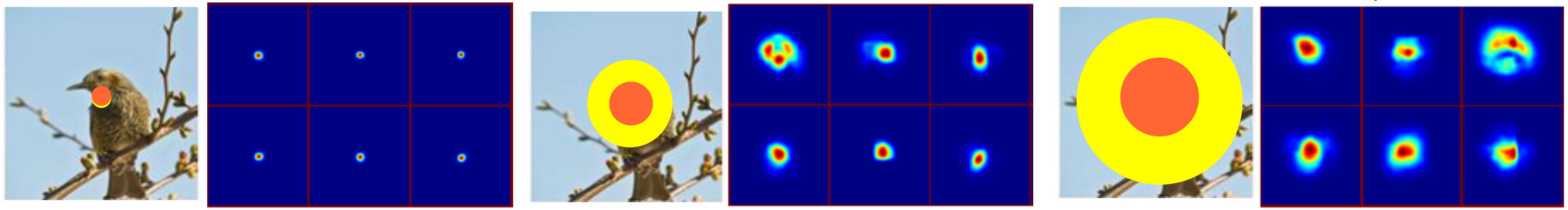


Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR'15

Estimating the Receptive Field of Unit

Estimated receptive fields

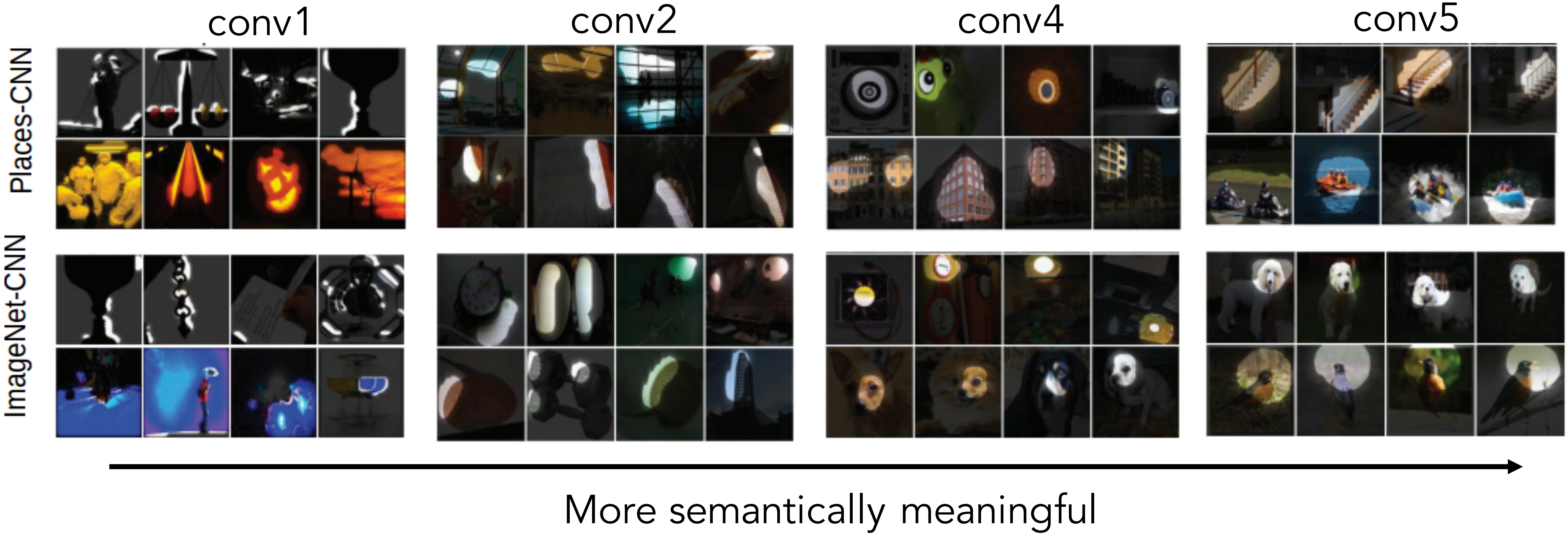
Actual size of RF is much smaller than the theoretic size



Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR'15

Segmenting Images by Units' Receptive Fields

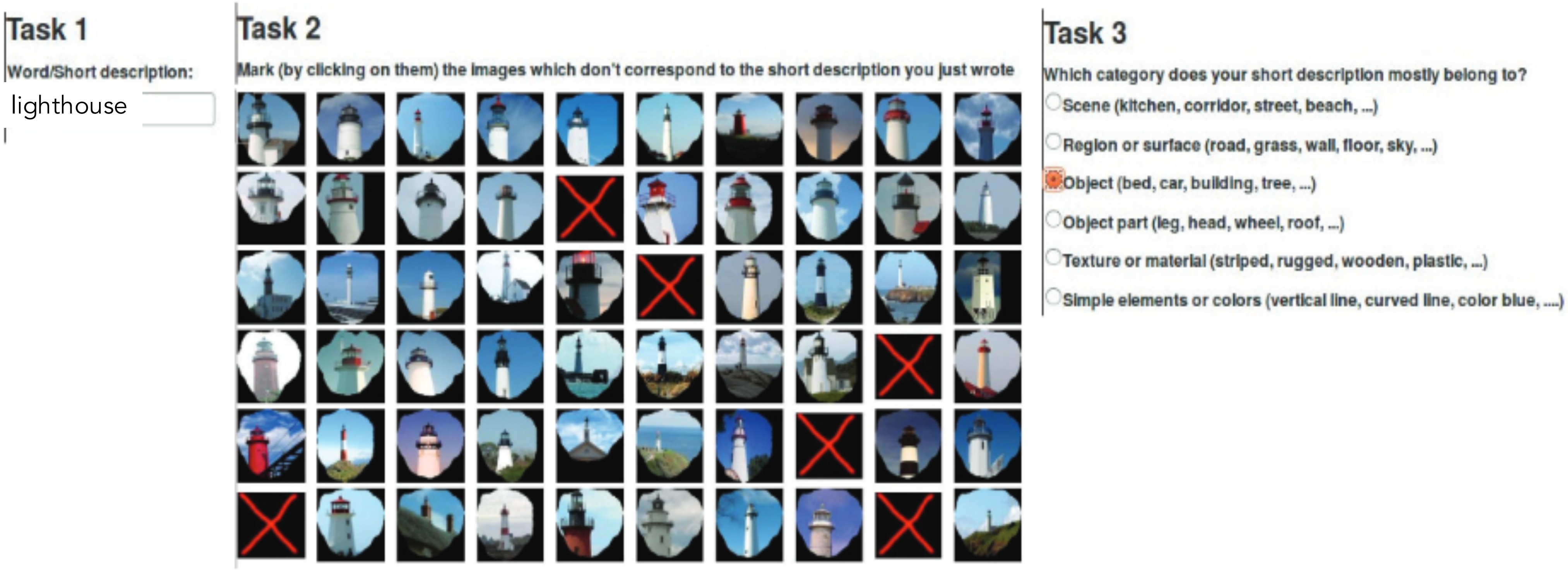
Image segmentation using units at different layers:



Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR'15

Annotating the Semantics of Units

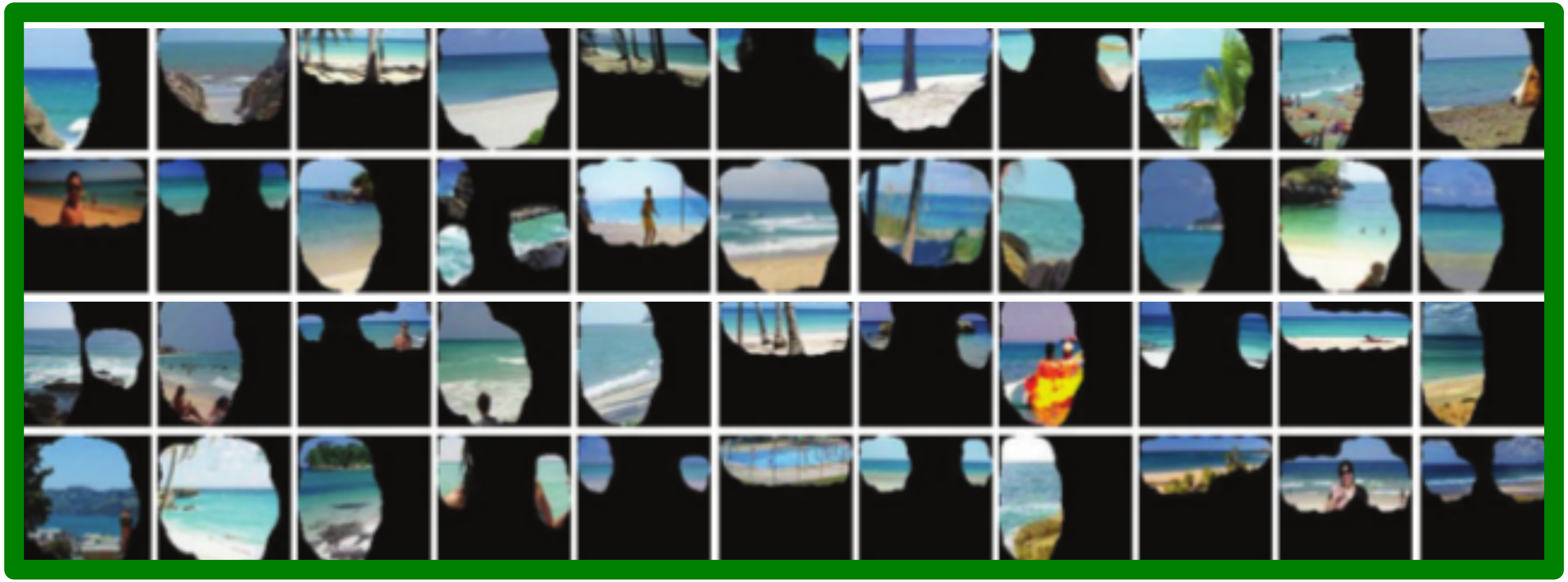
Top ranked segmented images are cropped and sent to Amazon Turk for annotation.

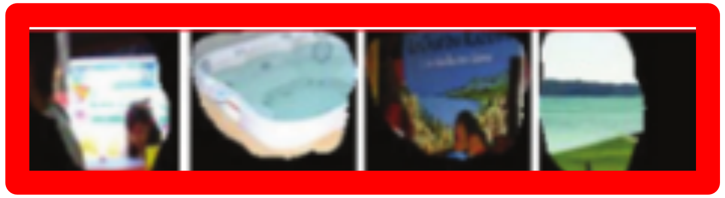


Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR'15

Annotating the Semantics of Units

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



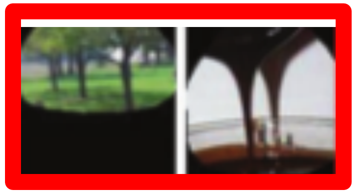


Annotating the Semantics of Units

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

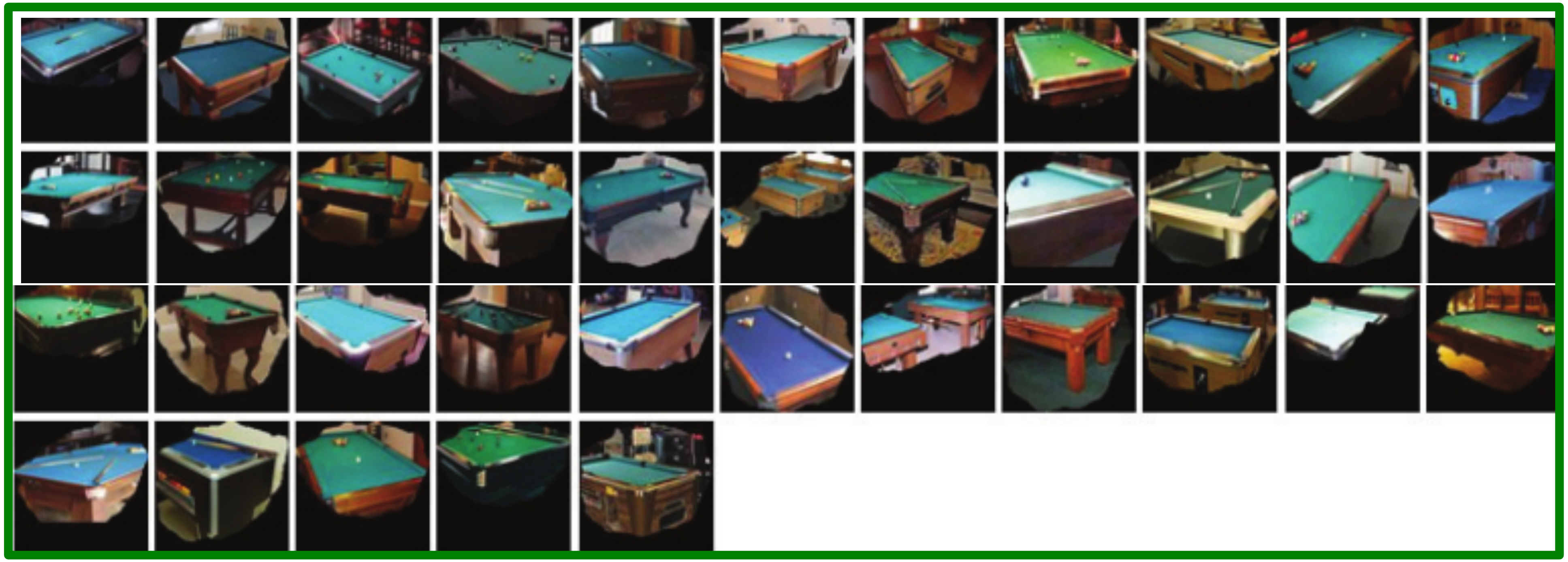
Annotating the Semantics of Units

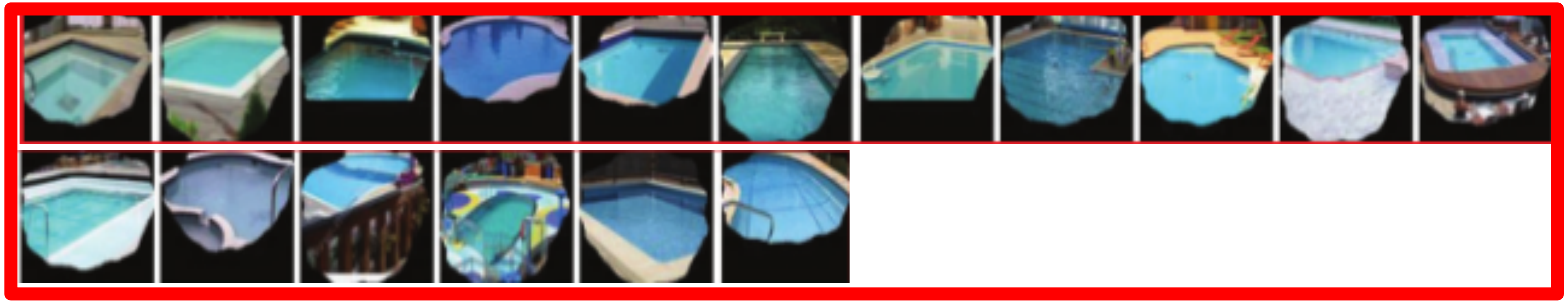
Pool5, unit 77; Label:legs; Type: object part; Precision: 96%



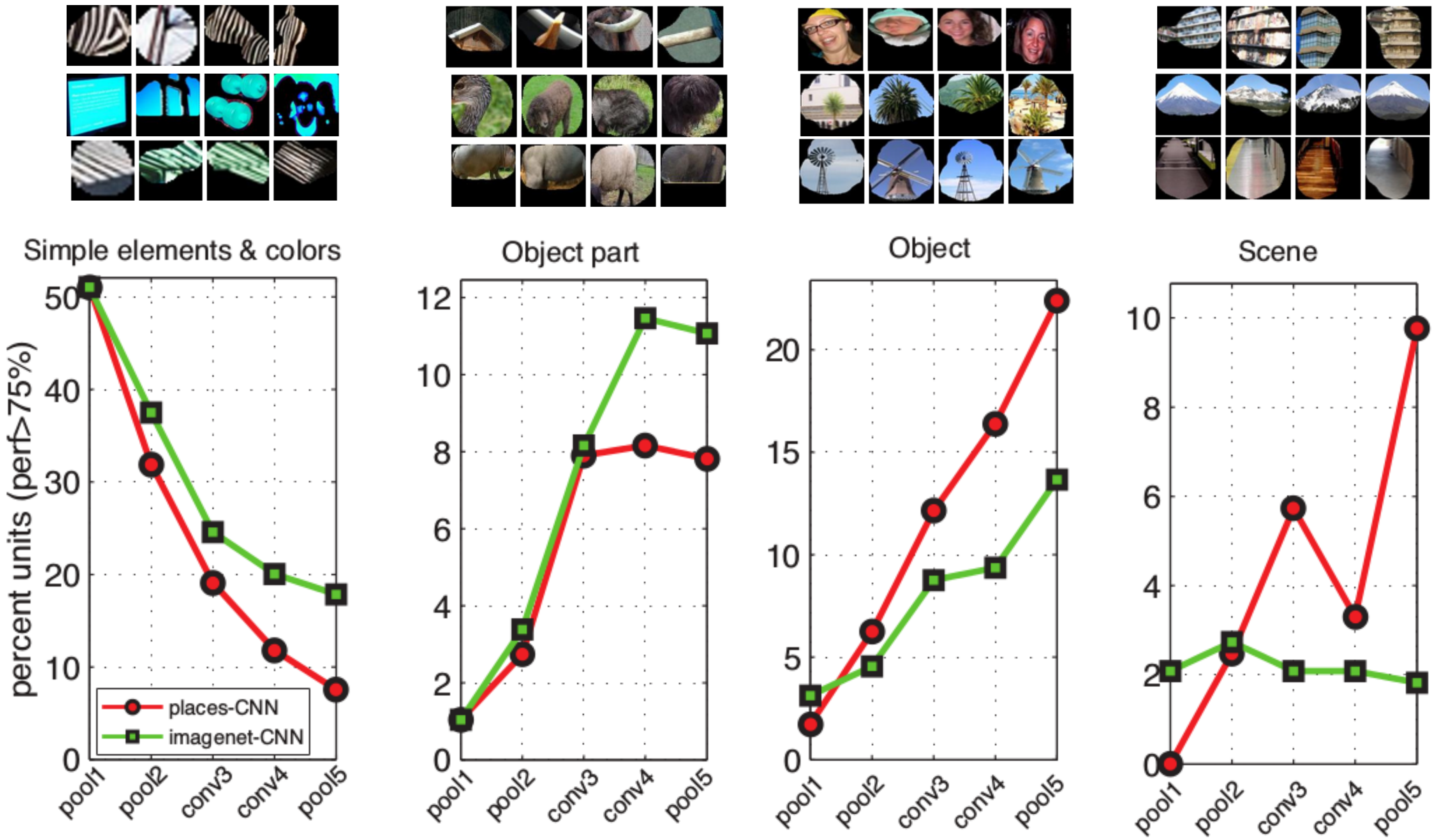
Annotating the Semantics of Units

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



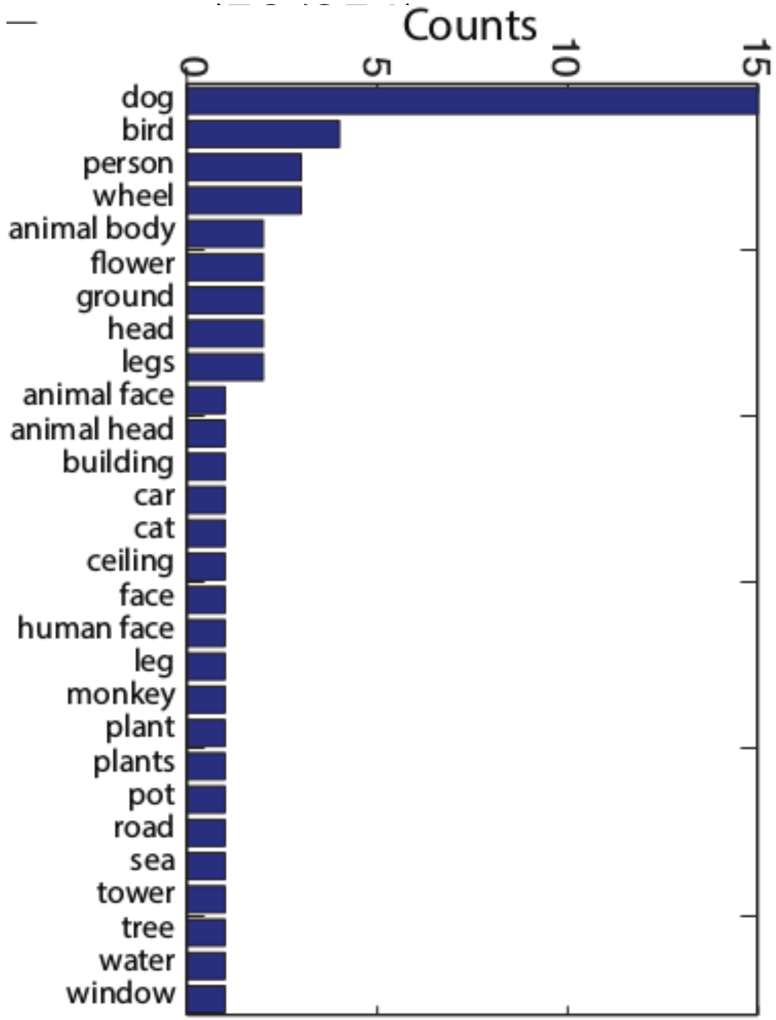


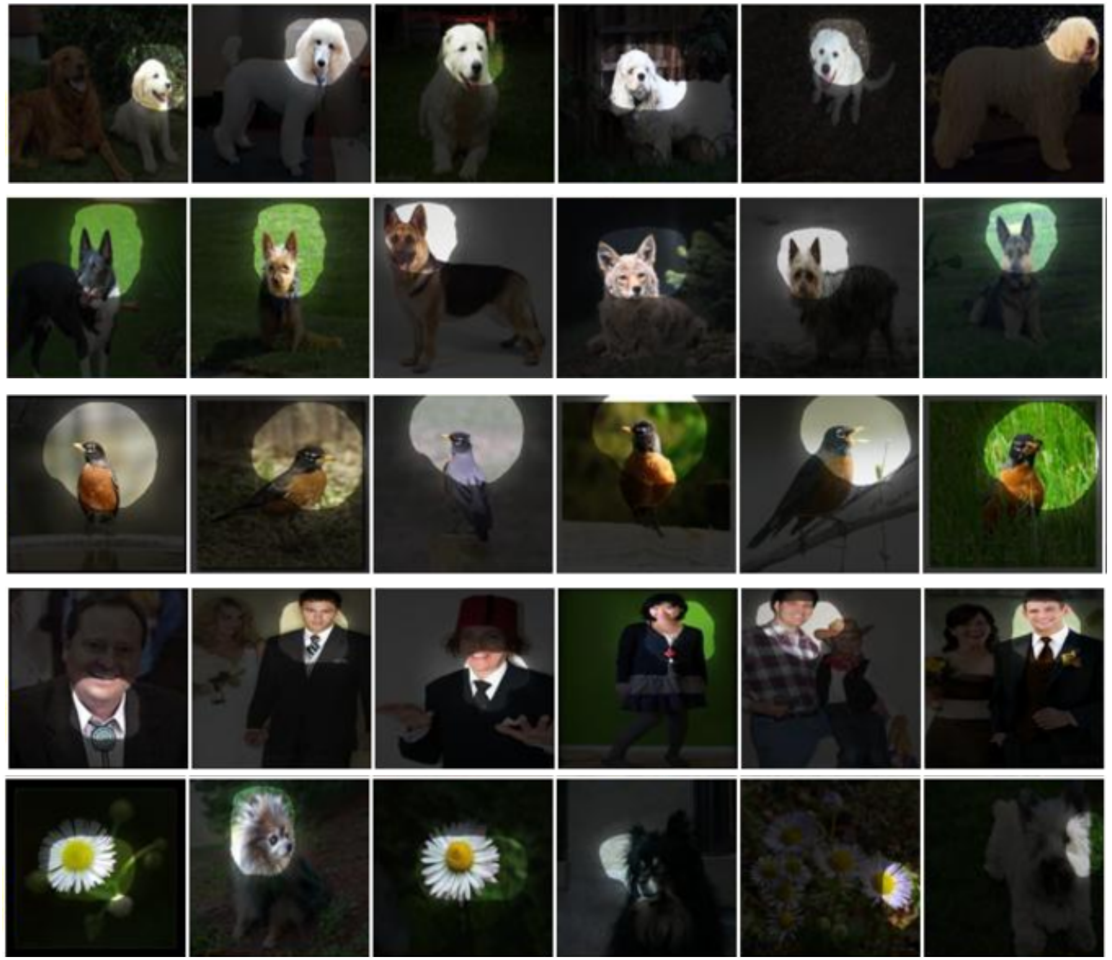
Distribution of Semantic Types at Each Layer



Histogram of Object Detectors in Pool5

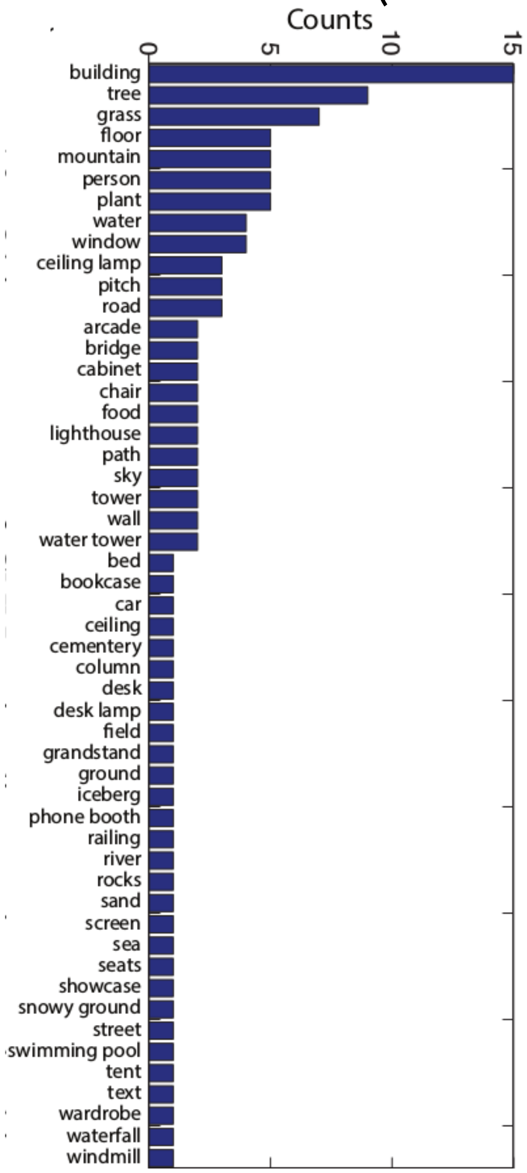
ImageNet-CNN

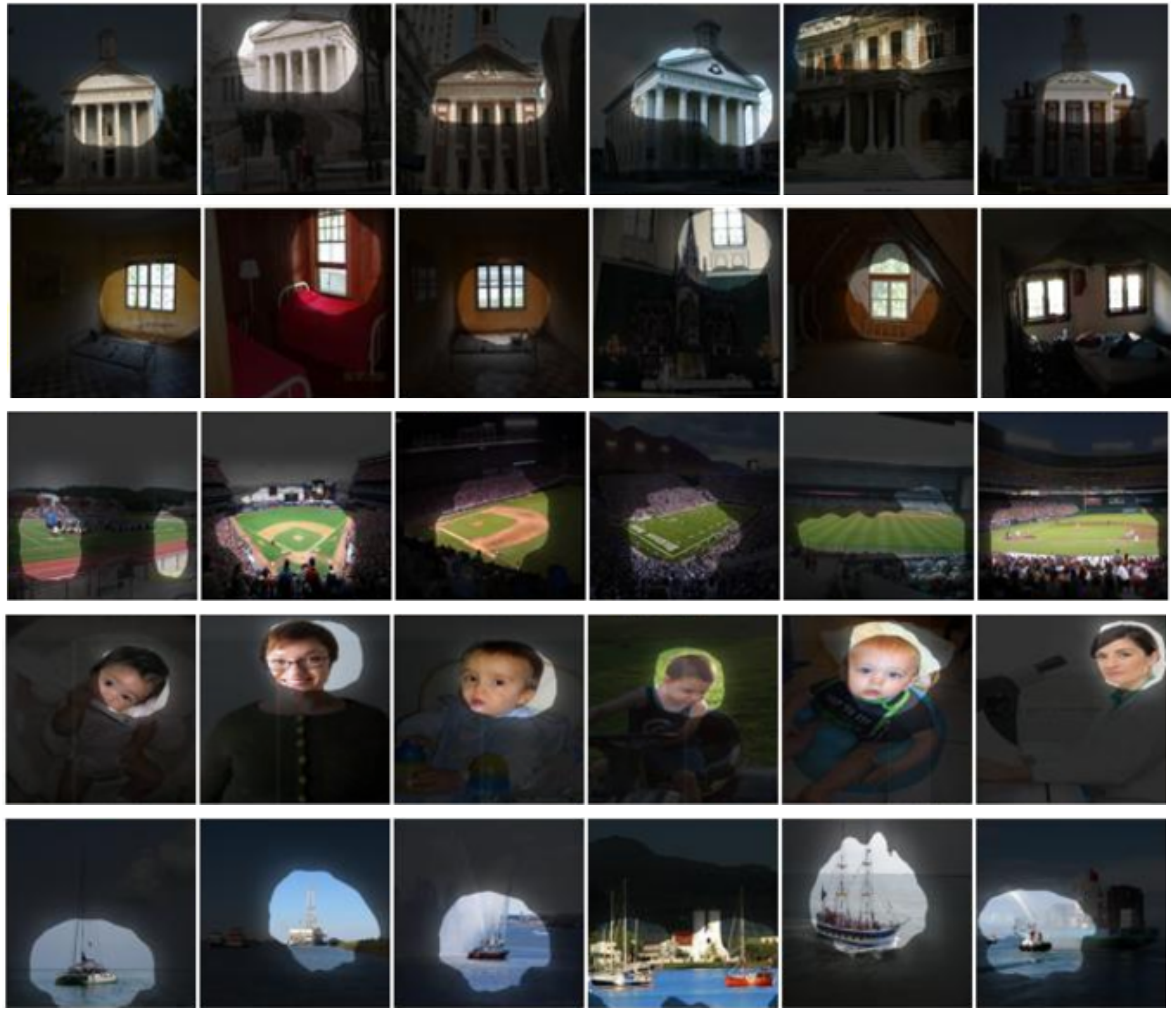


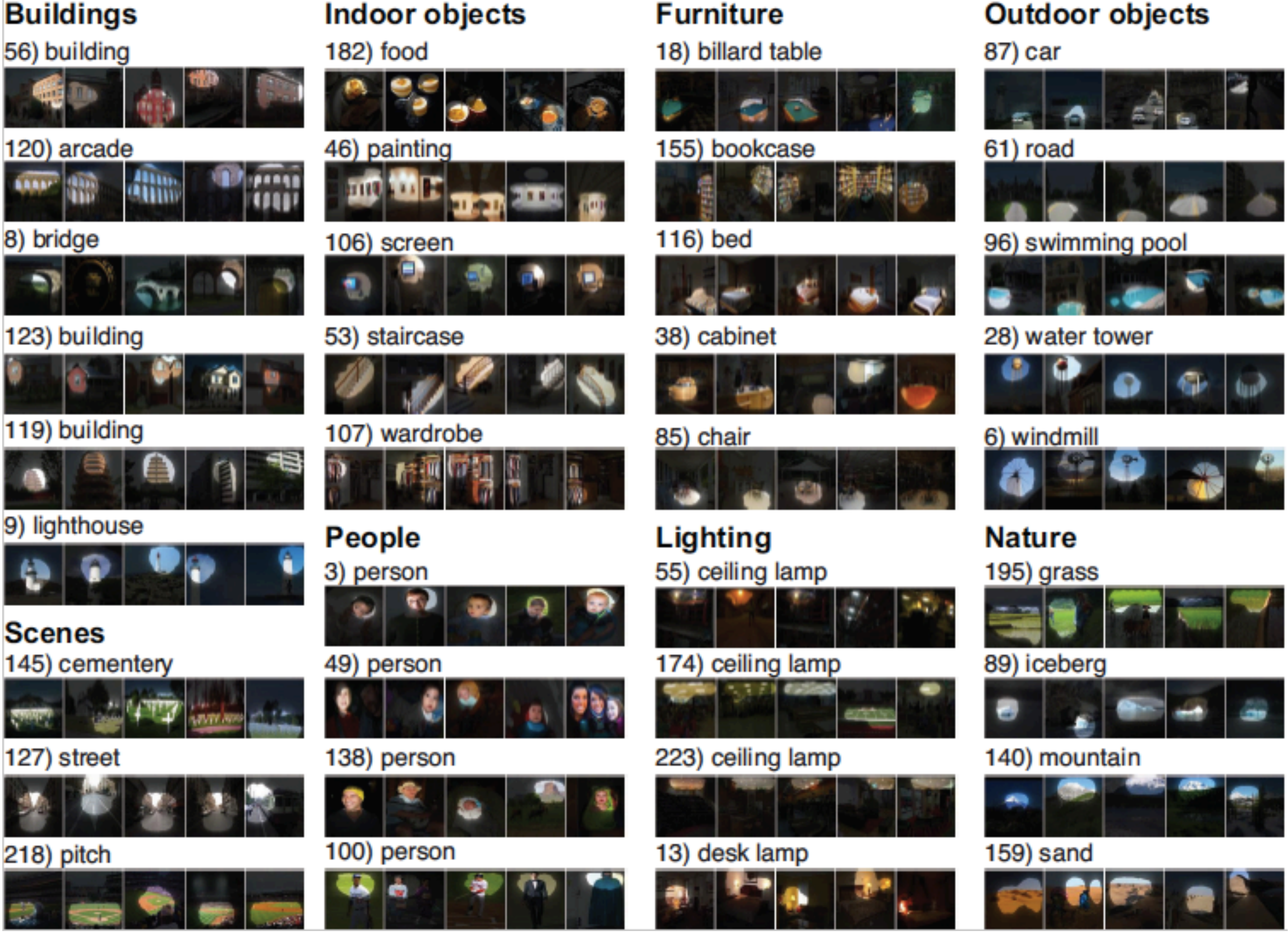


Histogram of Object Detectors in Pool5

Places-CNN (151/256)







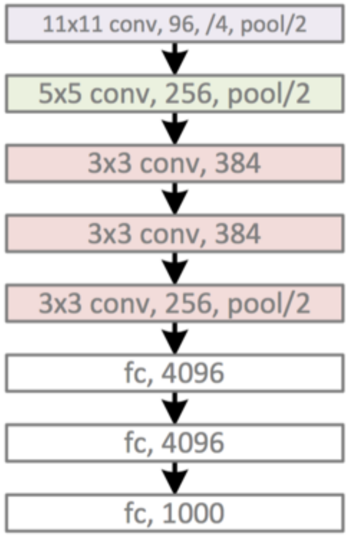
Buildings
56) building
120) arcade
8) bridge
123) building
119) building
9) lighthouse
Scenes
145) cementery
127) street
218) pitch
Indoor objects
182) food
46) painting
106) screen
53) staircase
107) wardrobe
People
3) person
49) person
138) person
100) person
Furniture
18) billard table
155) bookcase
116) bed
38) cabinet
85) chair
Lighting
55) ceiling lamp
174) ceiling lamp
223) ceiling lamp
13) desk lamp
Outdoor objects
87) car
61) road
96) swimming pool
28) water tower
6) windmill
Nature
195) grass
89) iceberg
140) mountain
159) sand

Issue: Manually annotating units is not scalable

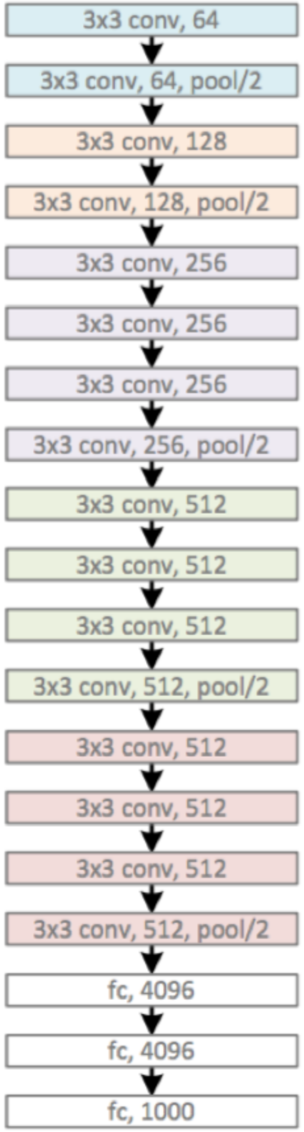
Top ranked segmented images are cropped and sent to Amazon Turk for annotation.

Zhou et al. Object Detectors Emerge from Deep Scene CNN. ICLR'15

AlexNet
5 conv layers



VGG
16 conv layers



GoogLeNet
~20 conv layers

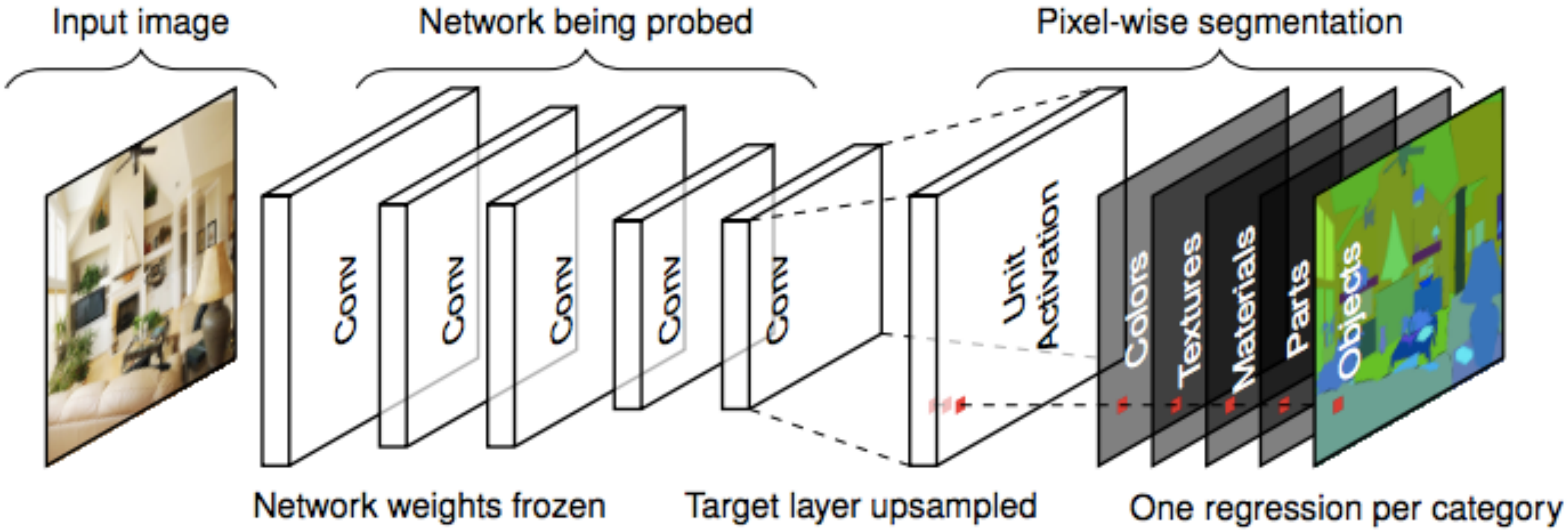
ResNet
>100 layers

Solution: Automatic annotation for unit semantics

Corpus of color dataset, texture dataset, shape dataset, object dataset, scene dataset

Bau* and Zhou*, et al. Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR'17

Solution: Automatic annotation for unit semantics



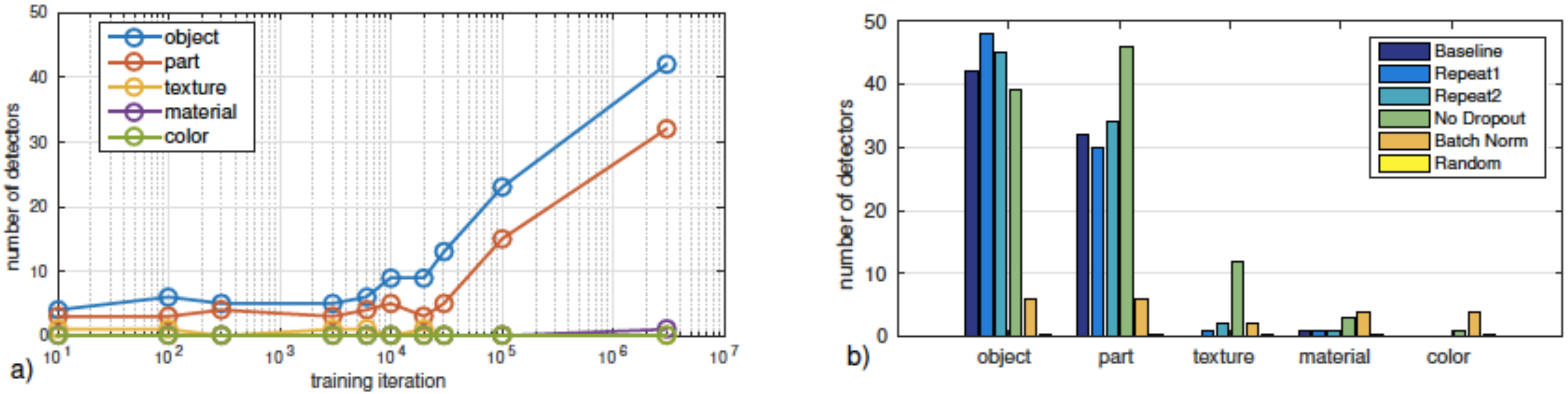
Bau* and Zhou*, et al. Network Dissection: Quantifying Interpretability of Deep Visual Representations. CVPR'17

Automatically Annotating Internal Units

Units annotated as concept detectors in the Places-AlexNet

Automatically Annotating Internal Units

Analyzing the effect of training tricks for network interpretability



A zoo of CNN models

Training	Network	Dataset or task
N/A	AlexNet	random
Supervised	AlexNet	ImageNet, Places205, Places365, Hybrid.
	GoogLeNet	ImageNet, Places205, Places365.
	VGG	ImageNet, Places205, Places365, Hybrid.
	ResNet	ImageNet, Places365.
Self	AlexNet	`context`, `puzzle`, `egomotion`, `tracking`, `moving`, `videoorder`, `audio`, `crosschannel`,`colorization`. `objectcentric`.

Supervised CNN on ImageNet/Places

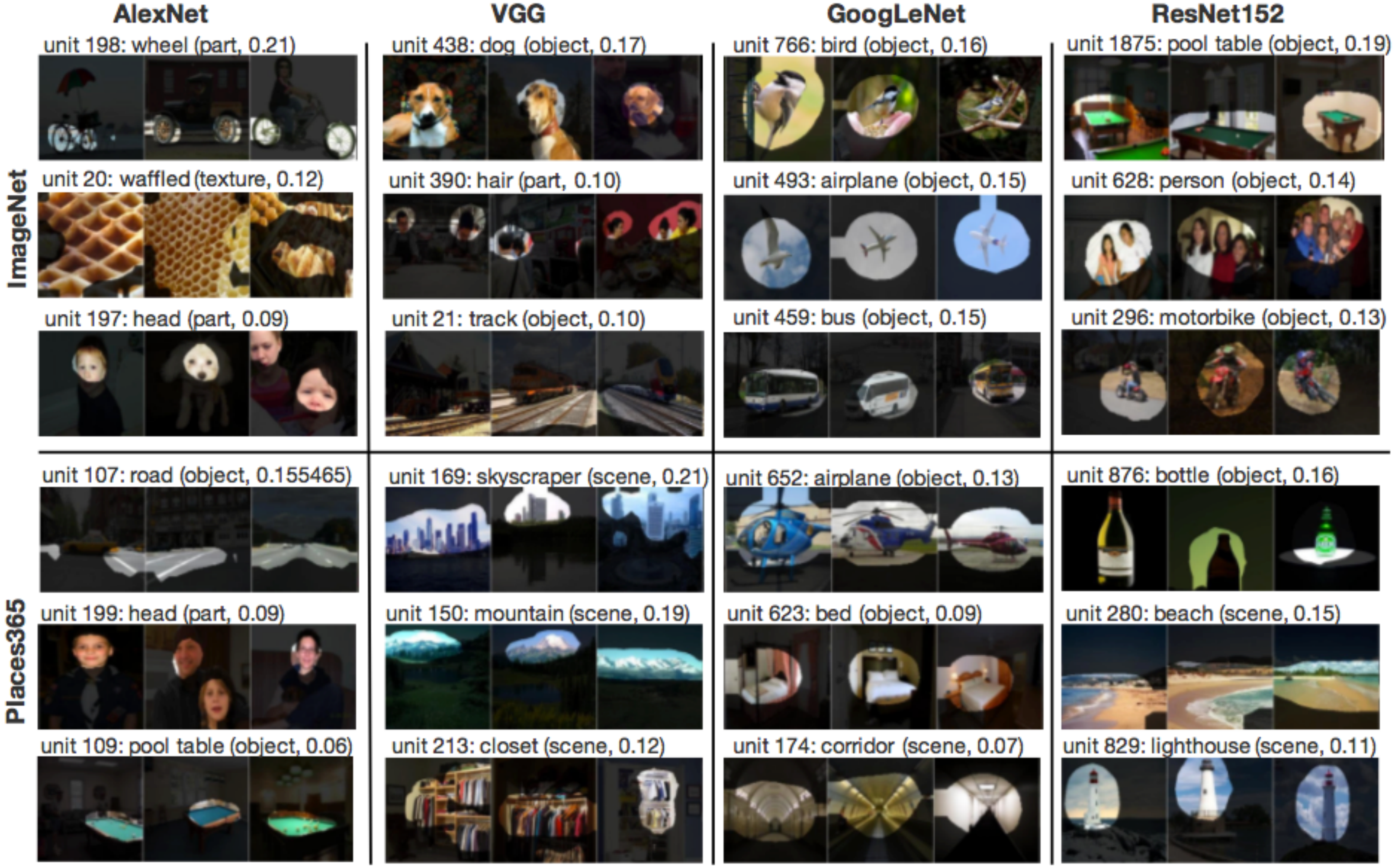
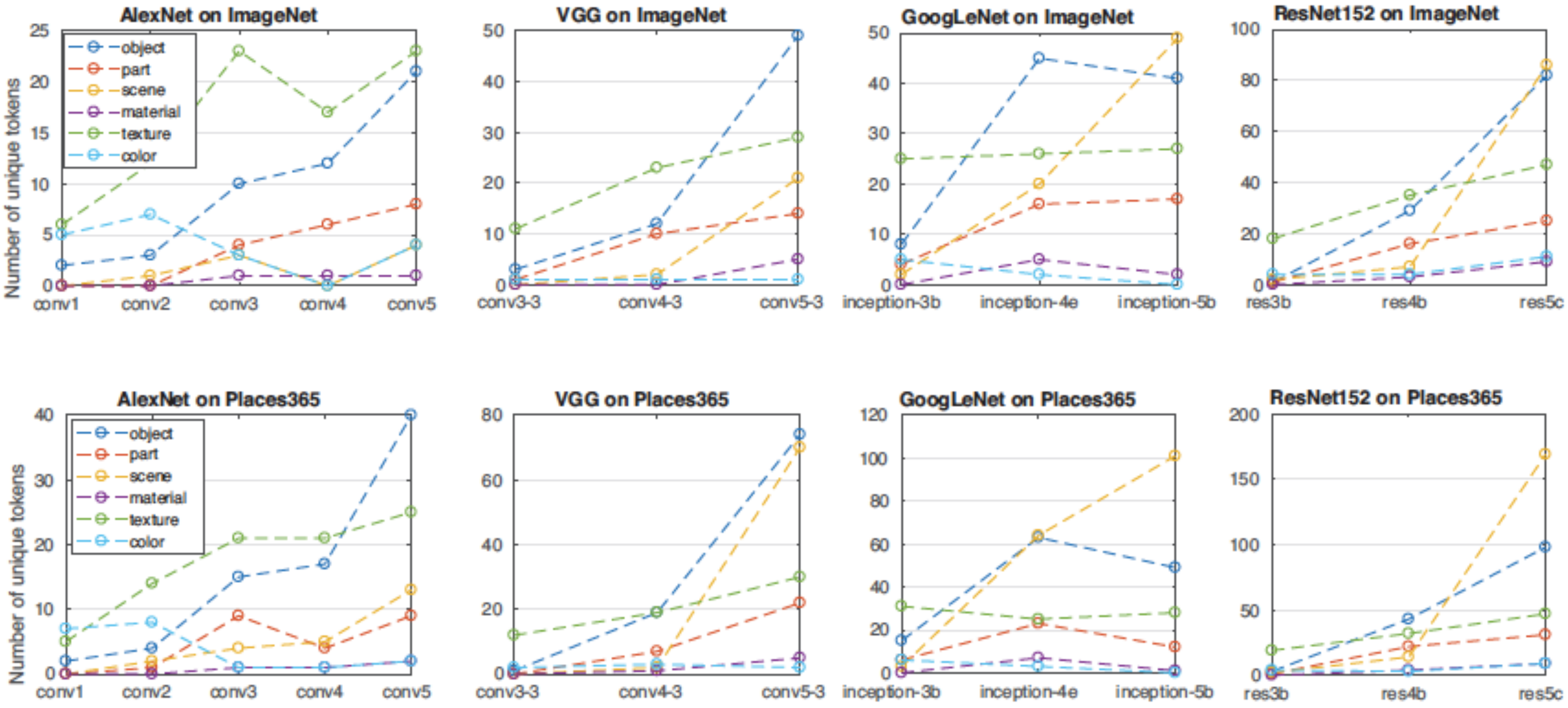


Figure 4: The top ranked tokens identified in the AlexNet, VGG, GoogLeNet, and ResNet on ImageNet and Places365.

Supervised CNN on ImageNet and Places

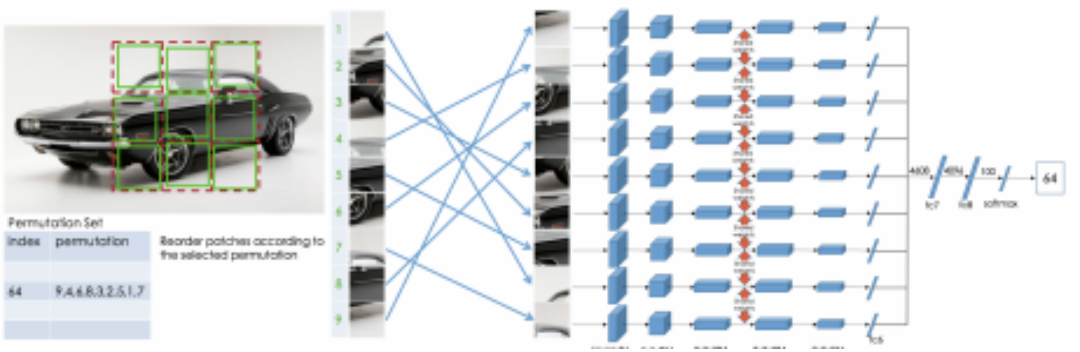
- Analyzing concept detectors change over layers



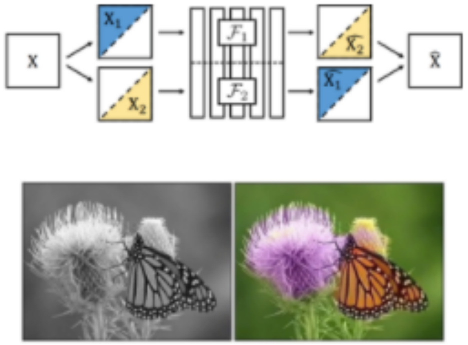
Self-supervised CNNs

- Examples of self-supervised training tasks:

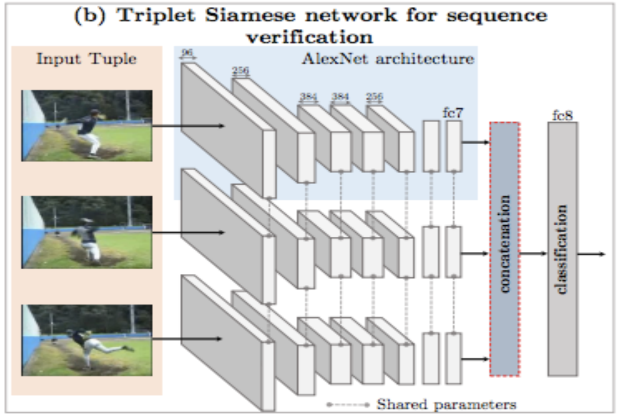
Context prediction, ICCV'15



Solving puzzle, ECCV'16



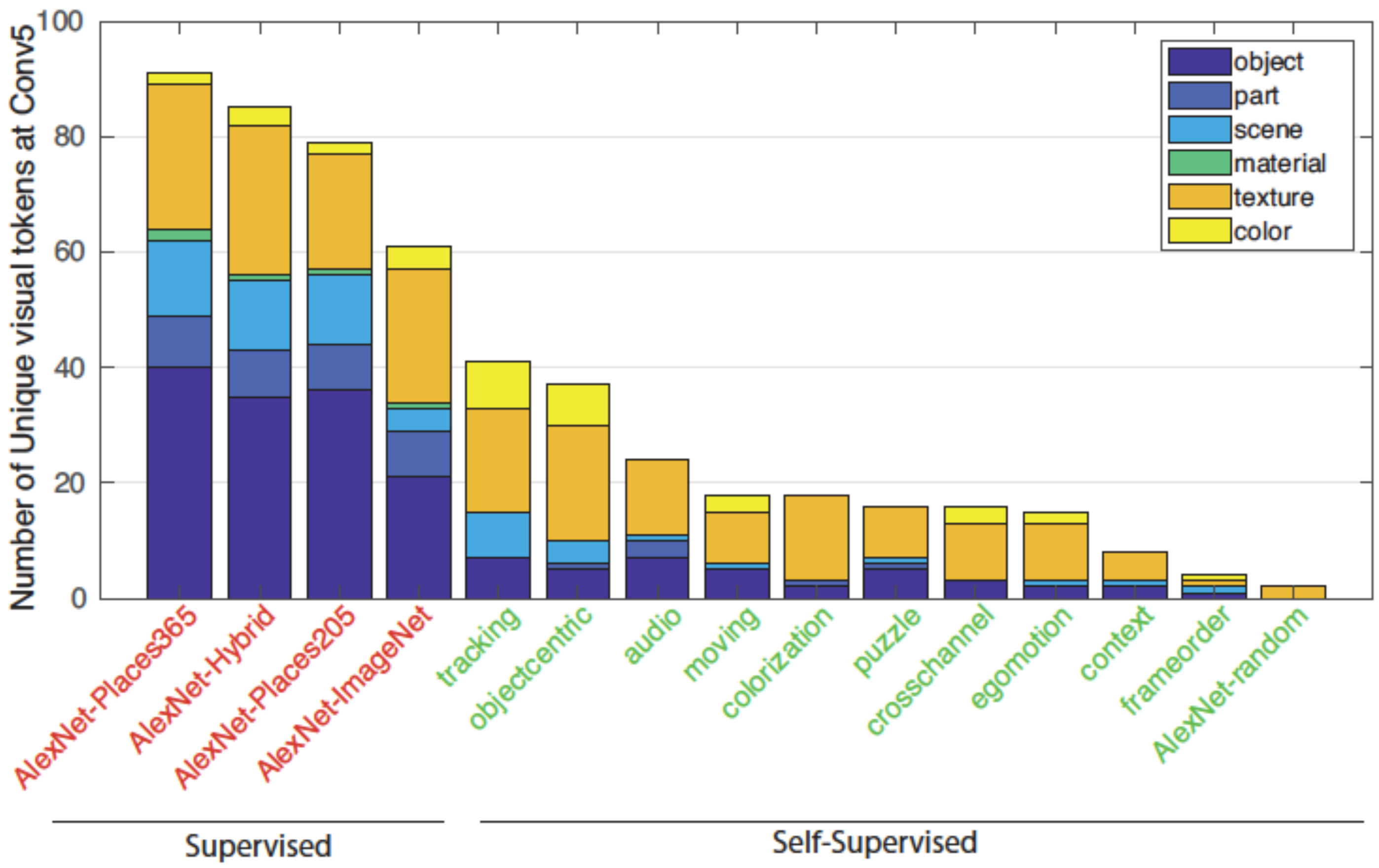
Colorization, ECCV'16 and CVPR'17



Predicting video order, ECCV'16

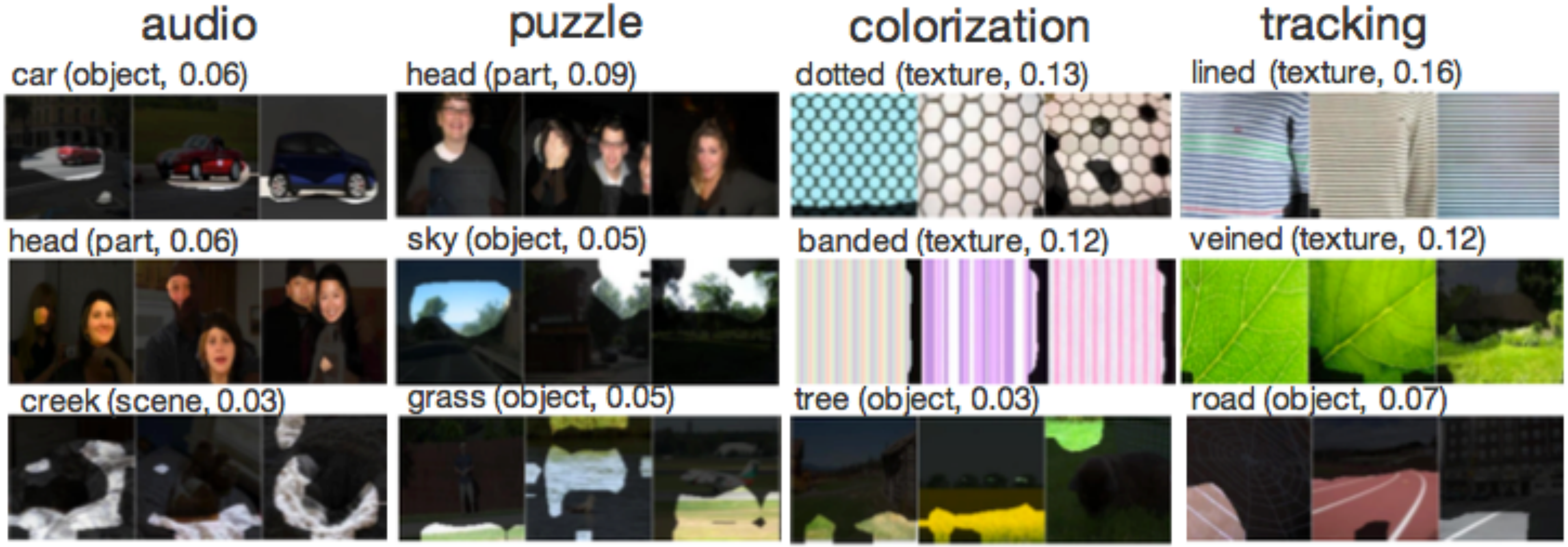
Self-supervised CNNs

- Comparison of supervised CNNs and self-supervised CNNs



Self-supervised CNNs

- Examples of detectors in self-supervised CNNs:



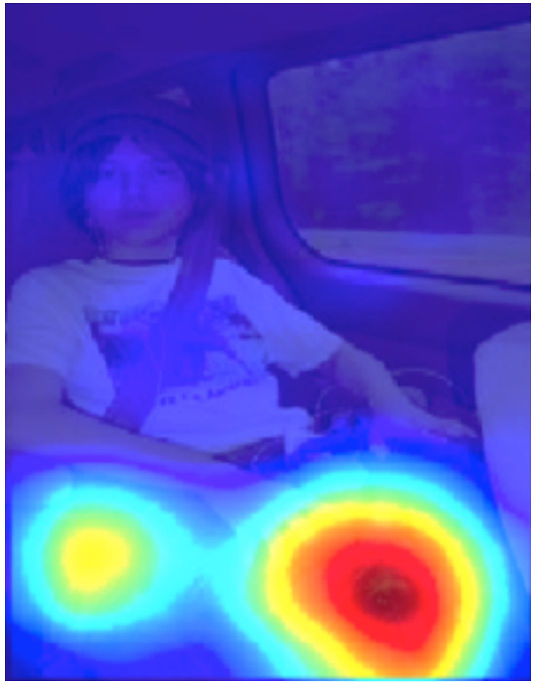
Explanatory factors in deep features

Leveraging the Internal Representations of CNNs

Zhou et al. **Learning Deep Features for Discriminative Localization.**
Computer Vision and Pattern Recognition (CVPR), 2016

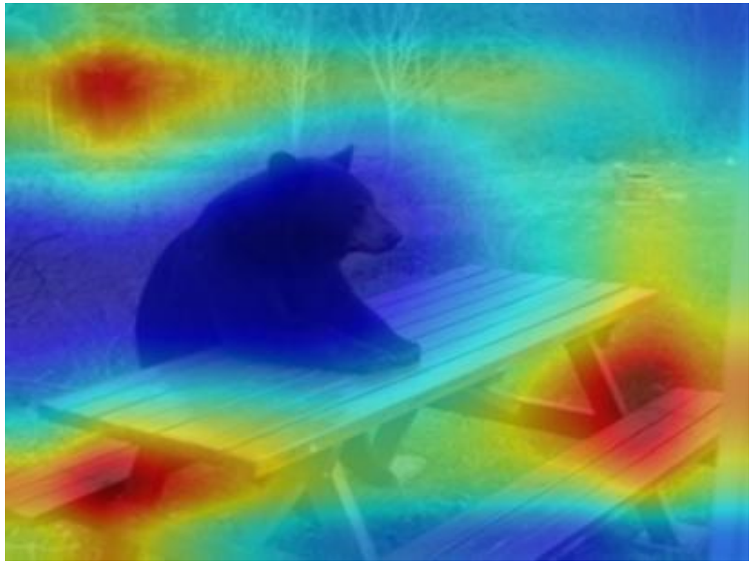
Why CNN makes the prediction?

Prediction from ImageNet-CNN:
Australian terrier:0.75



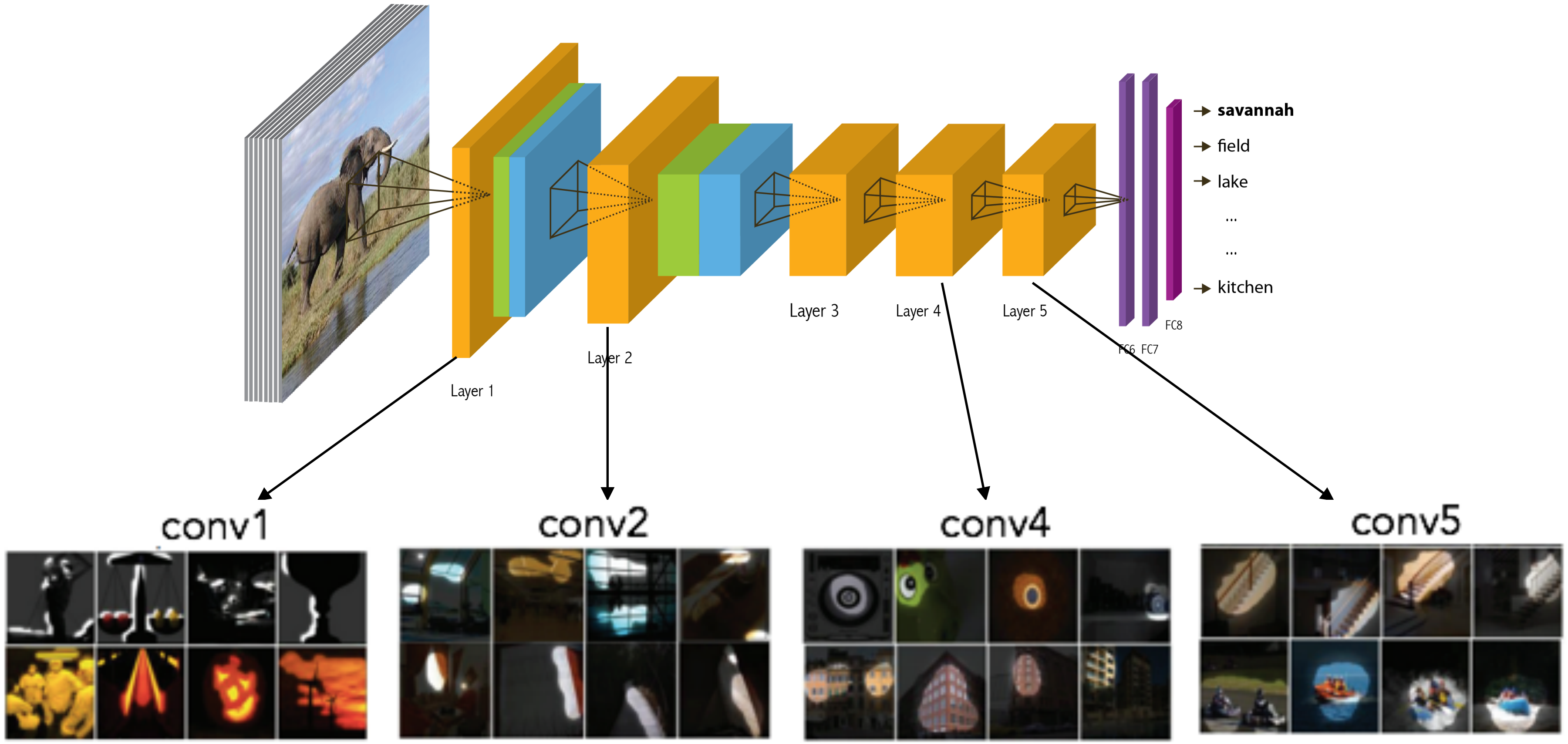
Why CNN makes the prediction?

Prediction from Places-CNN:
    Picnic area:0.64



Why CNN makes the prediction?

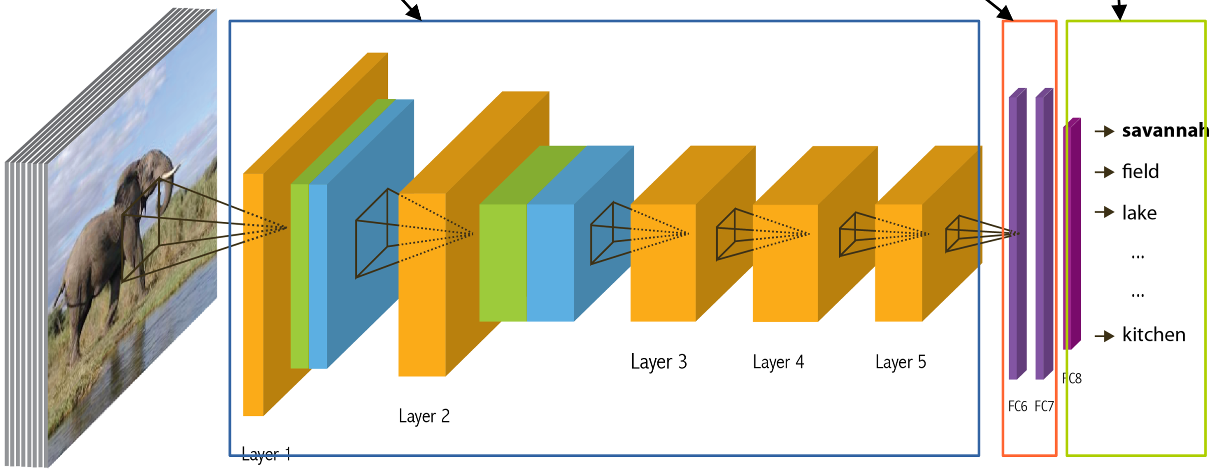
Previous work:
convolutional units as concept detectors at different layers

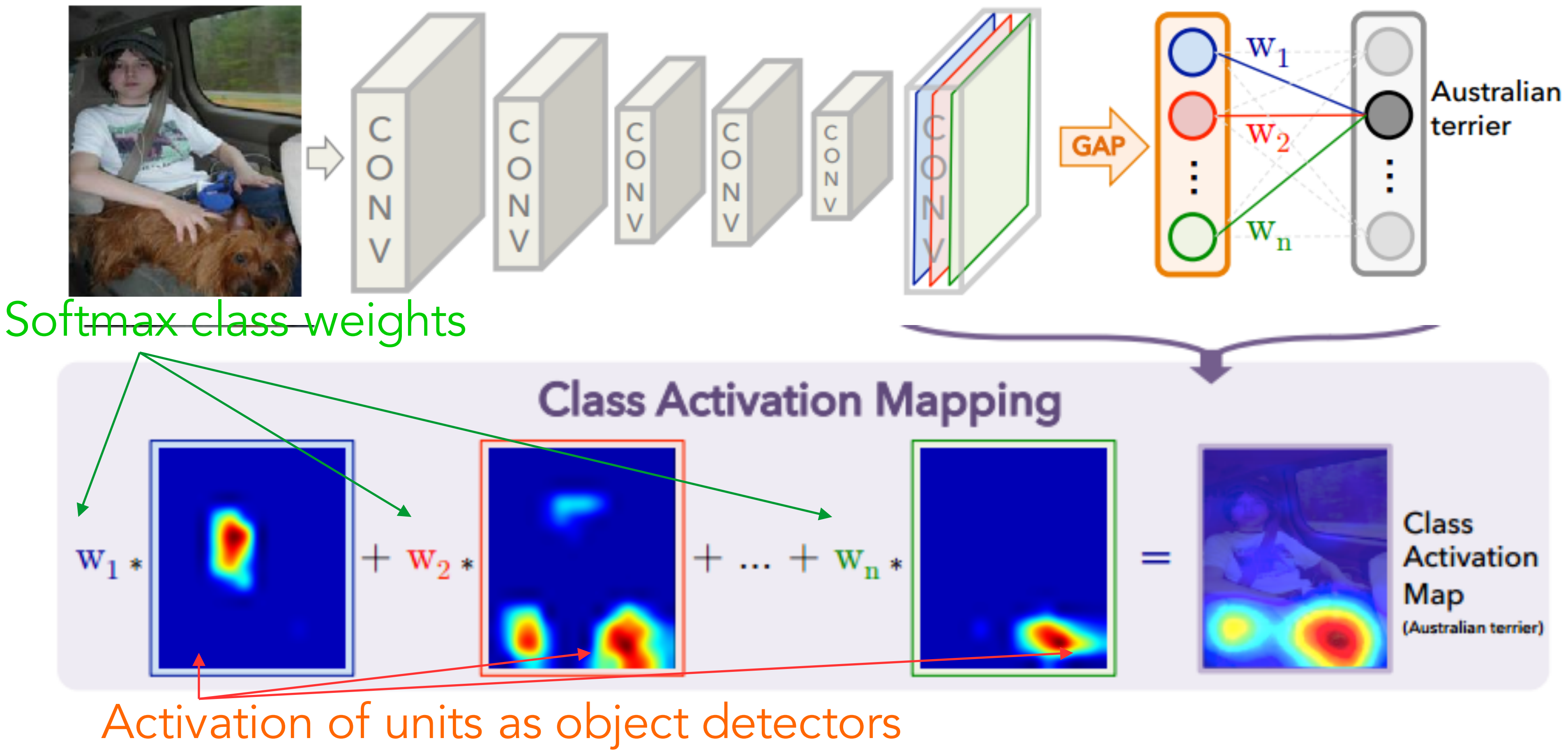


Simplifying the Network Architecture

Global Average Pooling

conv layers + ~~FC layers~~ + softmax layer

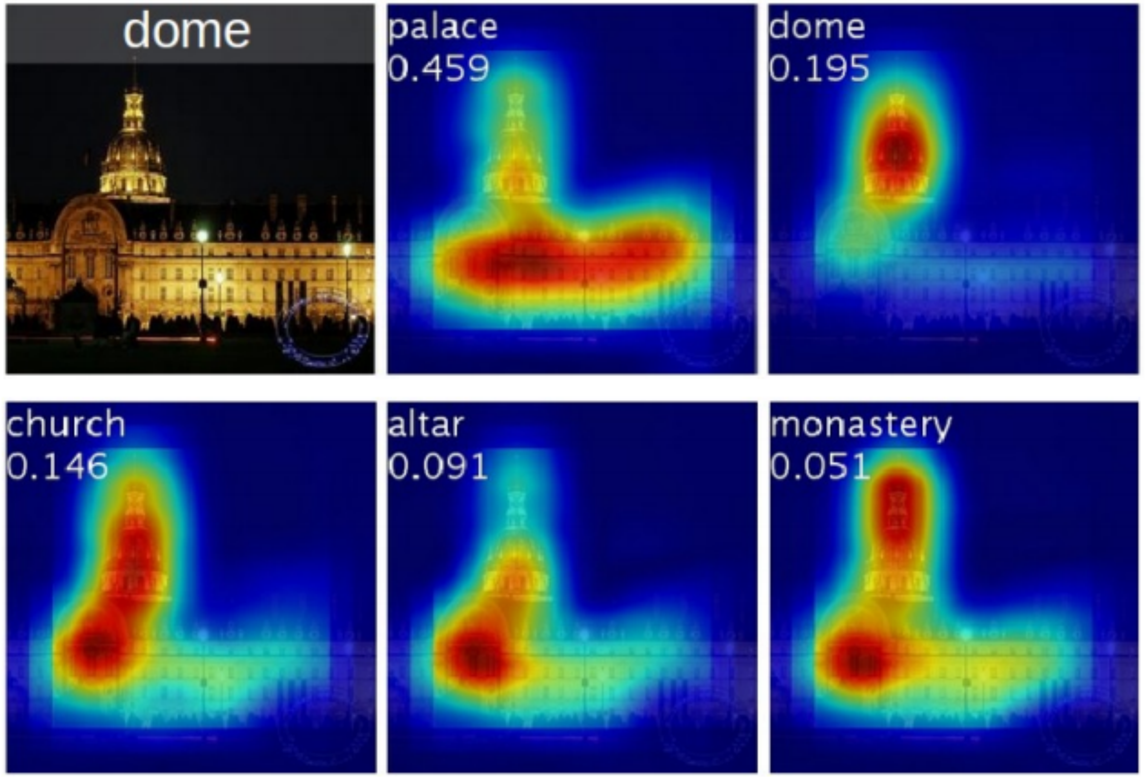




CONV
GAP
w_1
w_2
w_n
Australian terrier
Softmax class weights
Class Activation Mapping
$w_1 * \ + \ w_2 * \ + \ \ldots \ + \ w_n * \ =$
Class Activation Map
(Australian terrier)
Activation of units as object detectors

Class Activation Mapping

- Different classes have different class activation maps
- Top5 predictions: palace, dome, church, altar, monastery



Effect from Removing the FC layers

Classification accuracy drops 2~3%, but with 90% less model parameters

Table 1. Classification error on the ILSVRC validation set.

Networks	top-1 val. error	top-5 val. error
VGGnet-GAP	33.4	12.2
GoogLeNet-GAP	35.0	13.2
AlexNet*-GAP	44.9	20.9
AlexNet-GAP	51.1	26.3
GoogLeNet	31.9	11.3
VGGnet	31.2	11.4
AlexNet	42.6	19.5
NIN	41.9	19.6
GoogLeNet-GMP	35.6	13.9

Weakly-supervised object localization

CNN-GAP is used for object localization, without training with bounding box annotation.

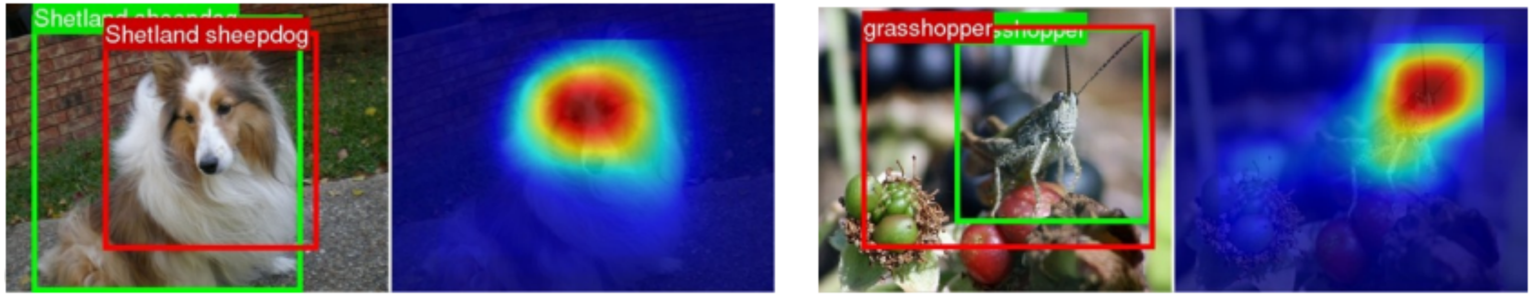


Table 3. Localization error on the ILSVRC test set for various weakly- and fully- supervised methods.

Method	supervision	top-5 test error
GoogLeNet-GAP (heuristics)	weakly	**37.1**
GoogLeNet-GAP	weakly	42.9
Backprop [22]	weakly	46.4
GoogLeNet [24]	full	26.7
OverFeat [21]	full	29.9
AlexNet [24]	full	34.2

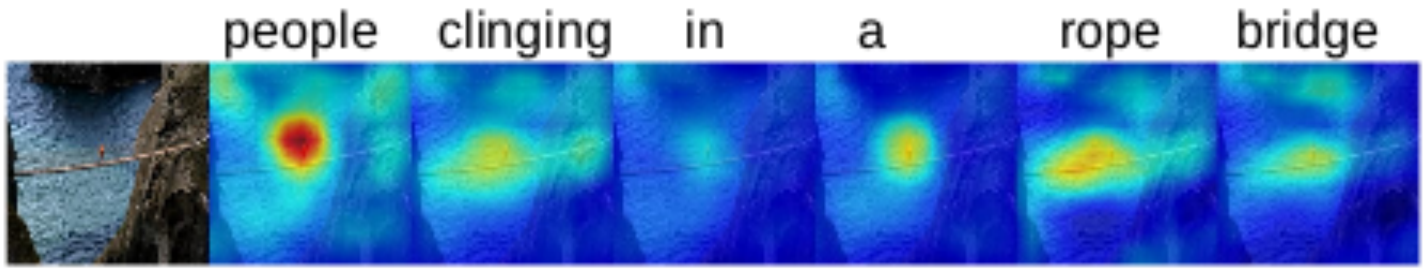
Localizable Visual Features

Deep CAM feature + linear SVM: localize informative regions

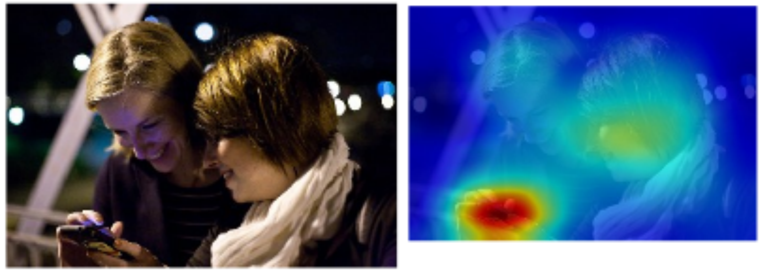
Or any other tasks with any loss functions, like regression, clustering, etc.

Localizable Visual Features

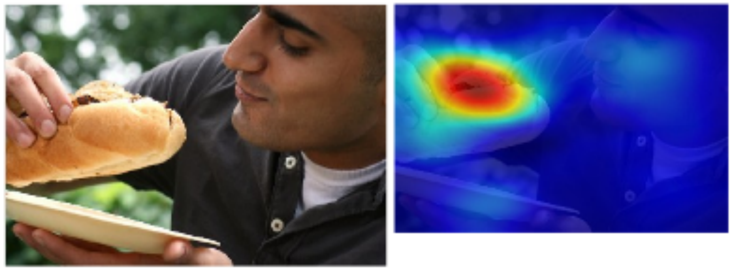
Image captioning using LSTM



Visual question answering



Question: What are they doing?
Prediction: texting (score: 12.02=3.78 [image] + 8.24 [word])



Question: What is he eating?
Prediction: hot dog (score: 13.01=5.02 [image] + 7.99 [word])

Zhou, et al, Simple Baseline for Visual Question Answering, arXiv1512

Demo video

https://www.youtube.com/watch?v=fZvOy0VXWAI
http://cnnlocalization.csail.mit.edu

- We analyzed the internal representation of CNNs, and leveraged them for weakly-supervised localization.
- The papers, the code, and pre-trained models are at

http://places.csail.mit.edu

http://cnnlocalization.csail.mit.edu