Interpretability and Visualization of Deep Neural Networks

Aude Oliva
MIT

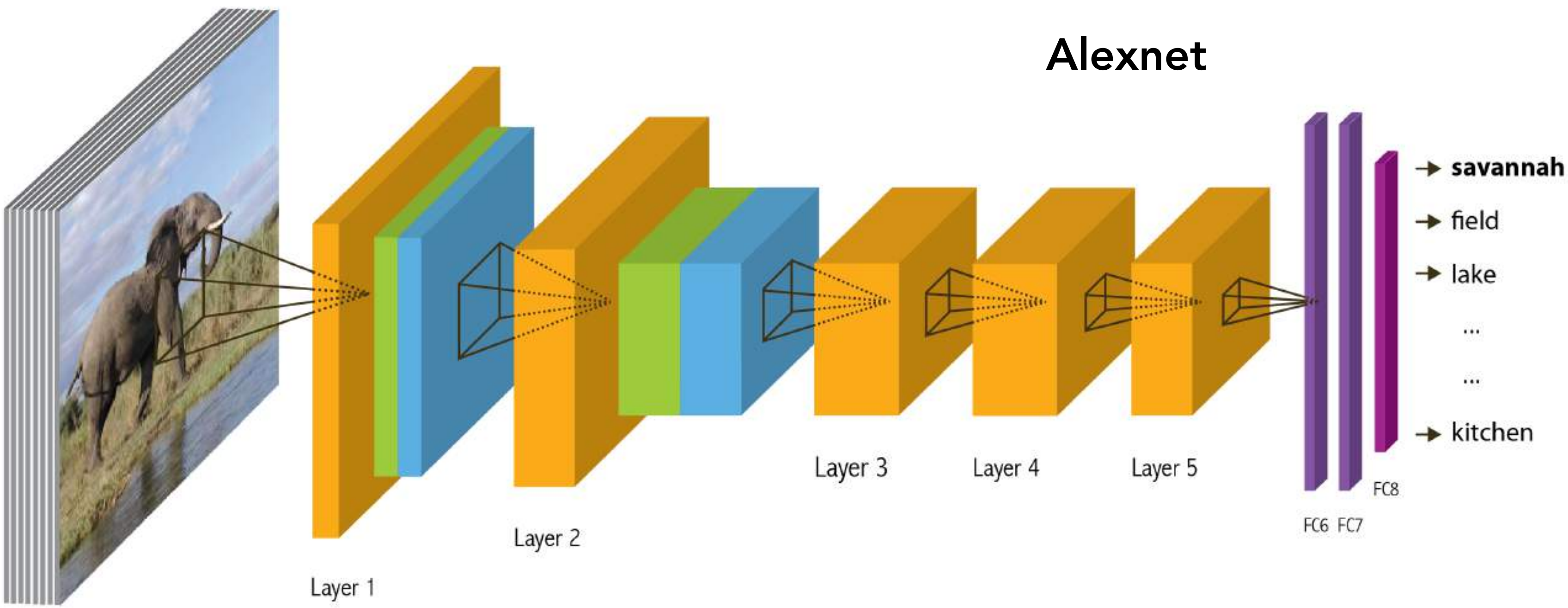
MIT Quest for Intelligence MIT-IBM Watson AI Lab CSAIL

RIGHT ISOMETRIC VIEW
OPA
RSC
Big-TOS
LOC
Small-LO
LOS
Big-PHC
Small-OTS
PPA
SAGITTAL
AXIAL
A
L
R
P


IMAGENET
places

Convolutional Neural Networks

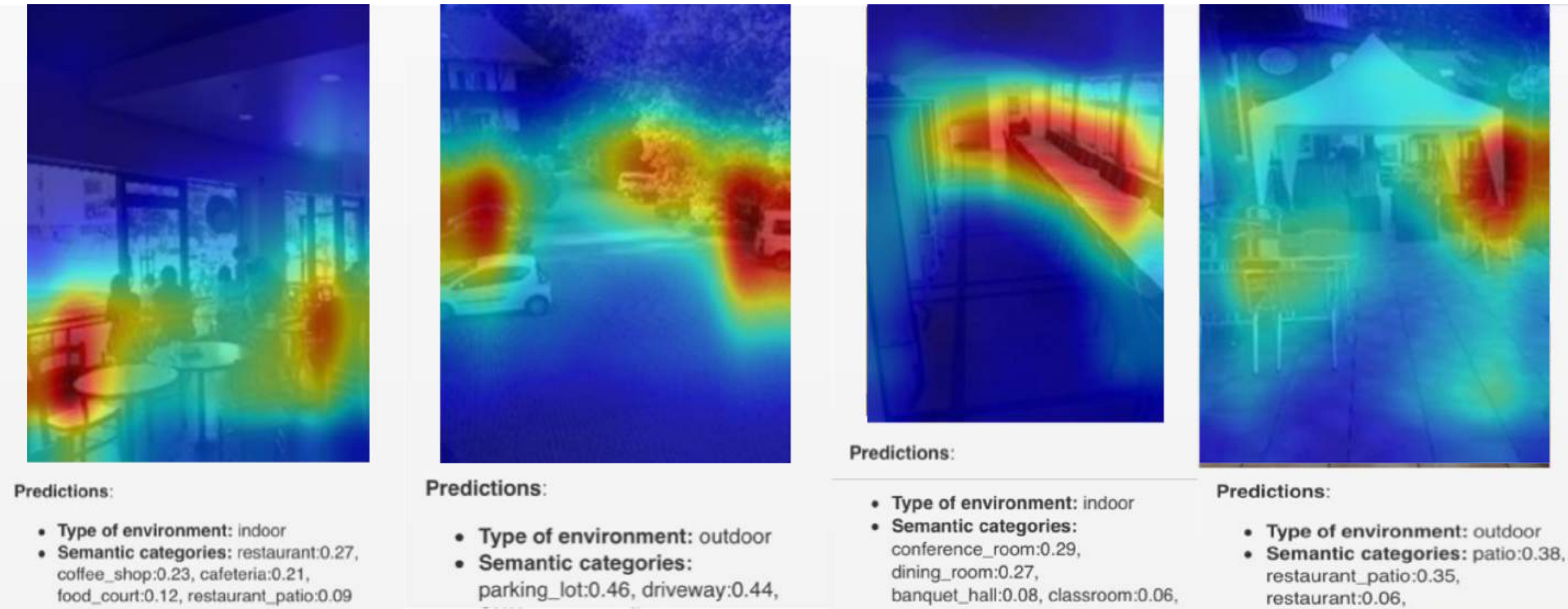
convolution max-pooling normalization full connected



Each layer learns progressively more complex features

Krizhevsky et al (2012).NIPS

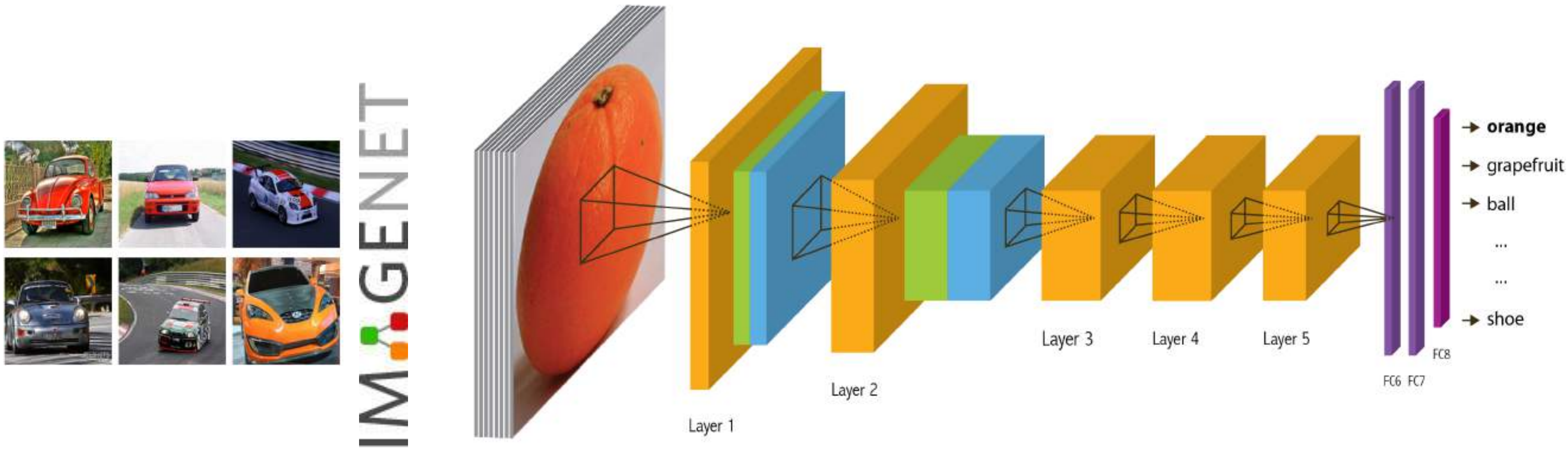
places

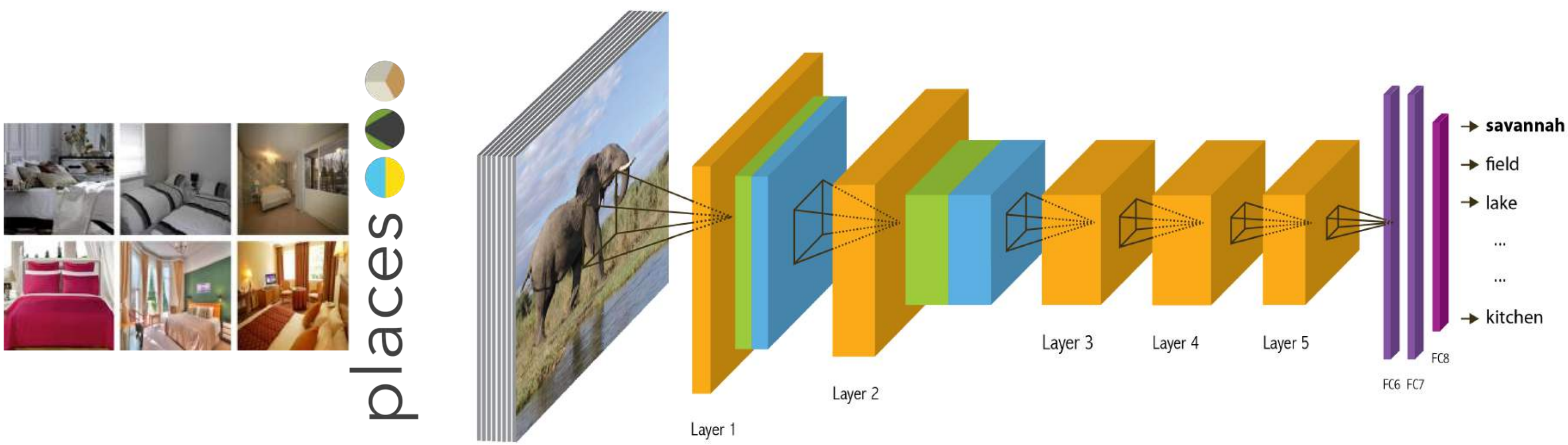


What did the network learn ?

places2.csail.mit.edu

Comparing Object and Scenes CNNs





Zhou, Khosla, at al (2015). ICLR

Data driven approach inspired by Neuroscience: Empirical receptive field

Pipeline for estimating the Receptive Fields:

Goal is to identify which regions of the image lead to the high unit activations.

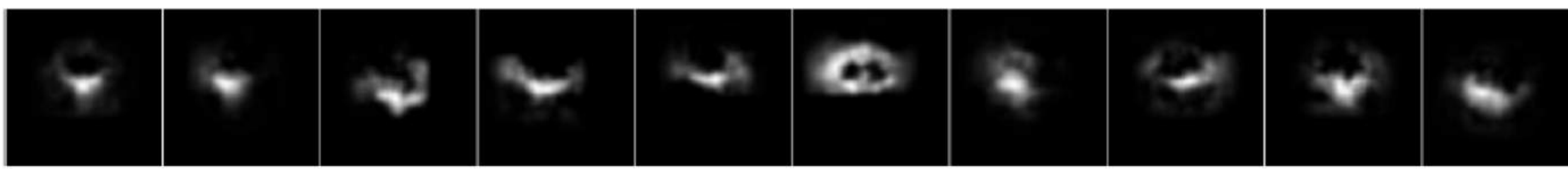
5000 occluded versions

Discrepancy map per unit

Zhou, Khosla, at al (2015). ICLR

Pipeline for estimating the Receptive Fields

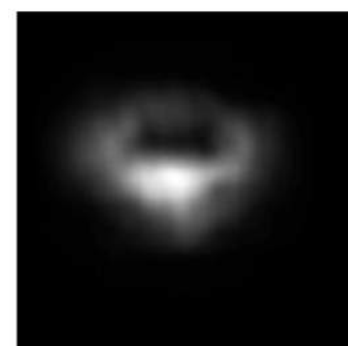
discrepancy maps for top 10 images



calibrated discrepancy maps

To consolidate the information from several images, we center the discrepancy map around the spatial location of the unit that caused the maximum activation for the given image.

Then we average the re-centered discrepancy maps to generate the final receptive field of each given unit.

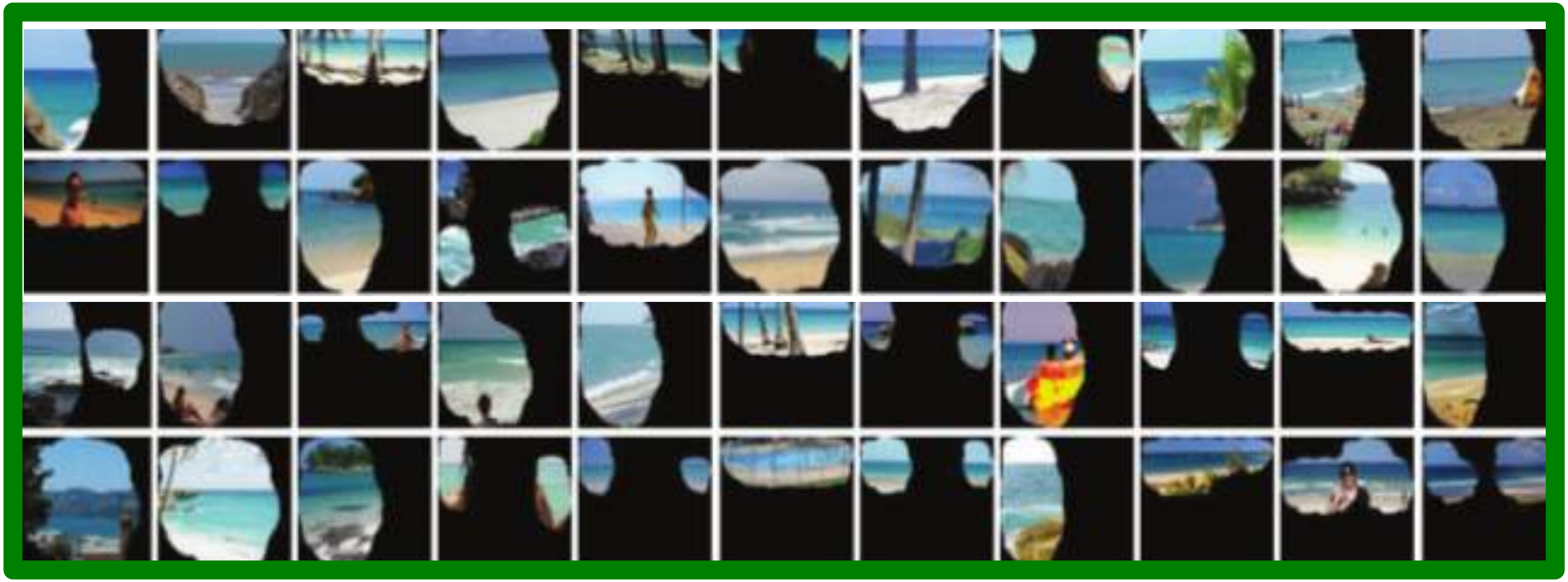


receptive field

Zhou, Khosla, at al (2015). ICLR

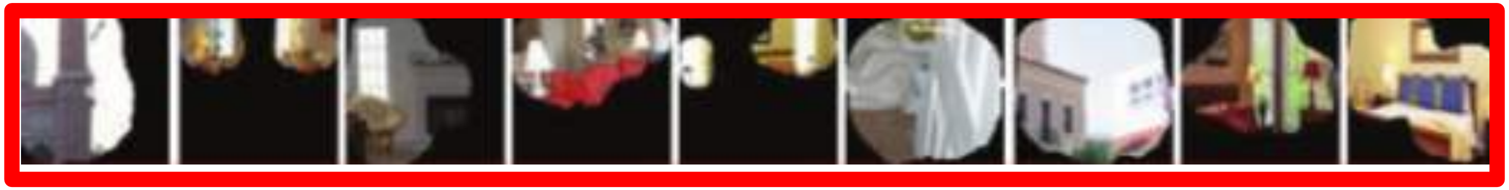
Annotating the Semantics of Units

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



Annotating the Semantics of Units

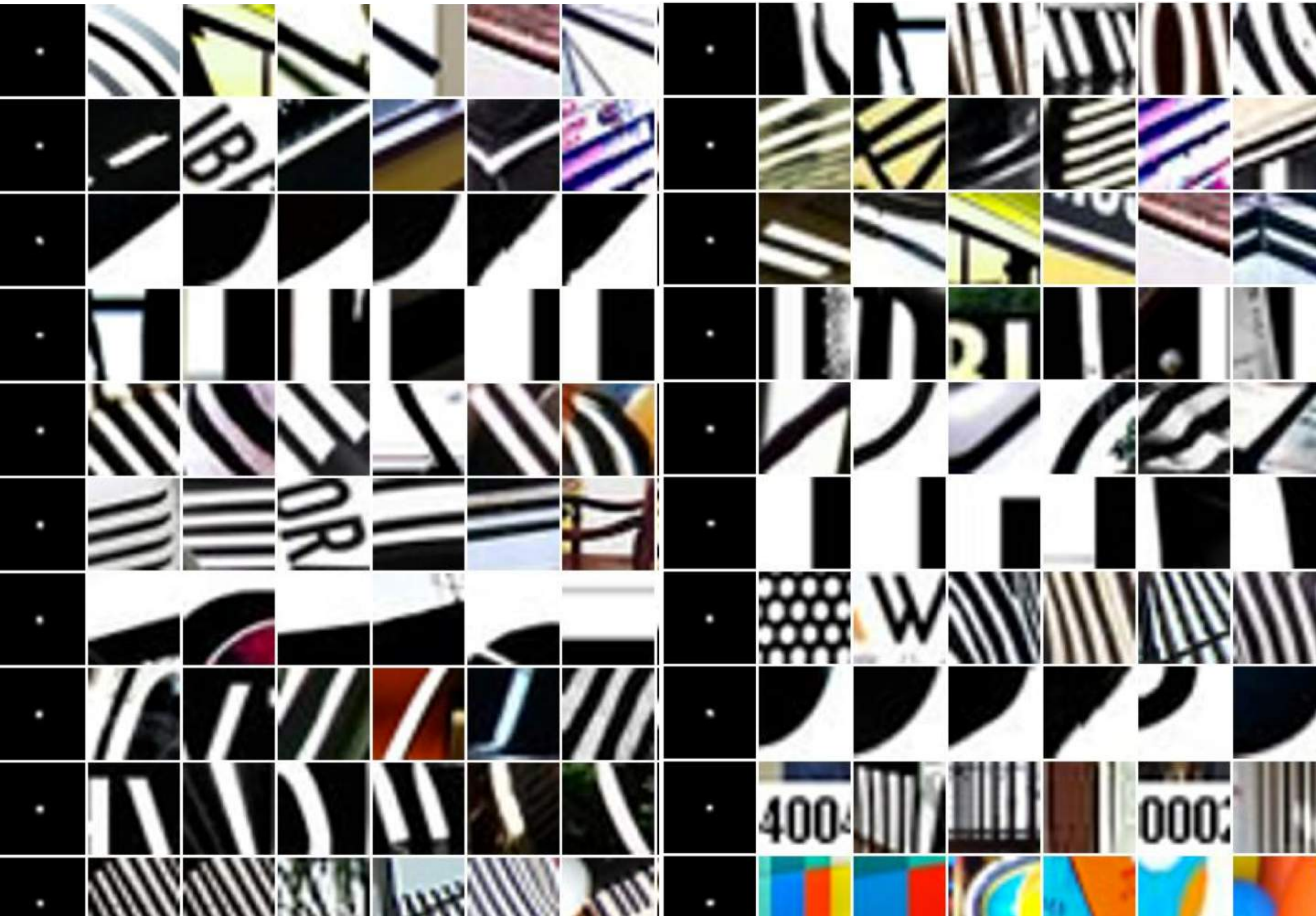
Pool5, unit 13; Label: Lamps; Type: object; **Precision: 84%**



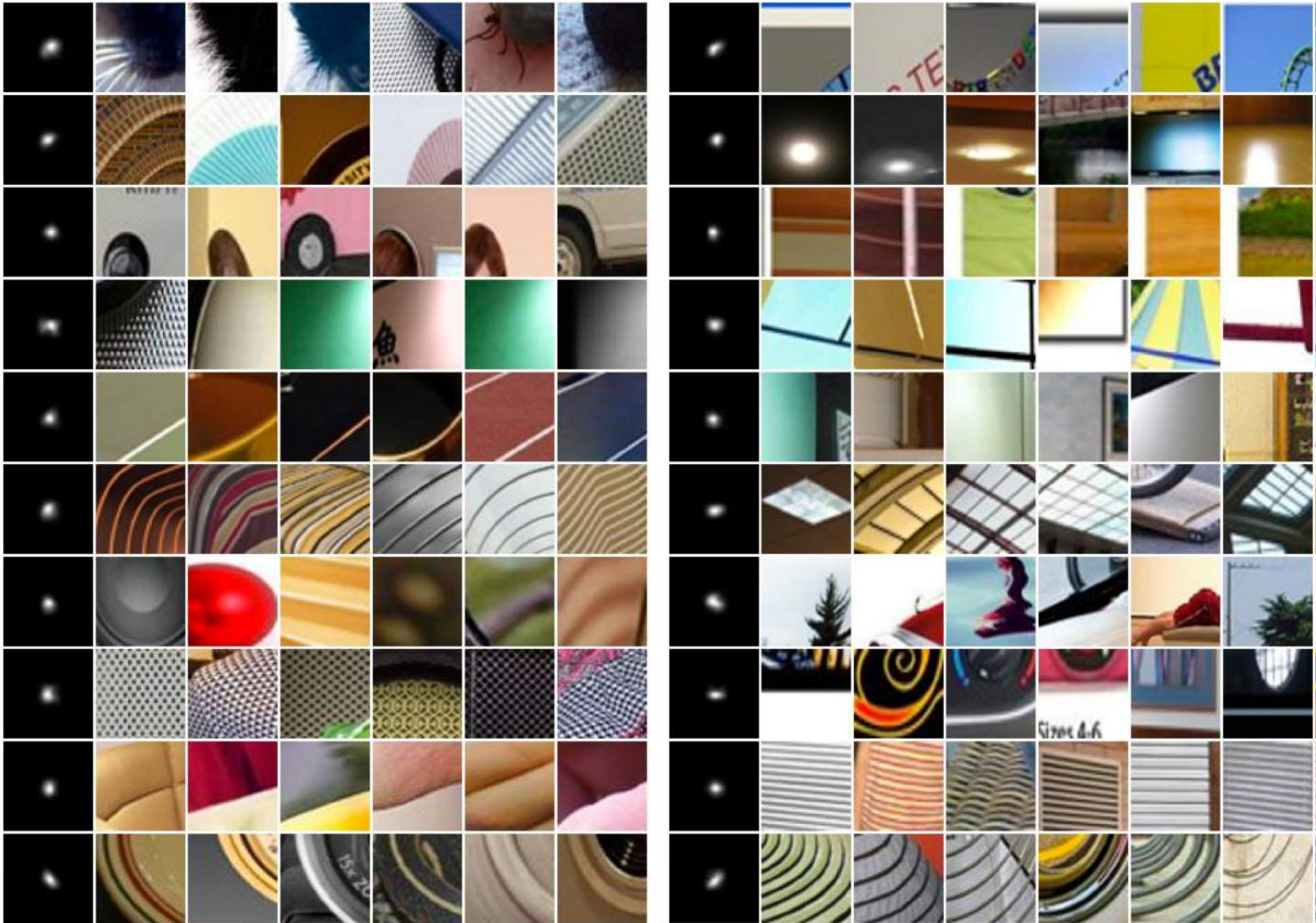
Annotating the Semantics of Units

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

IMAGENET **Layer 1** places



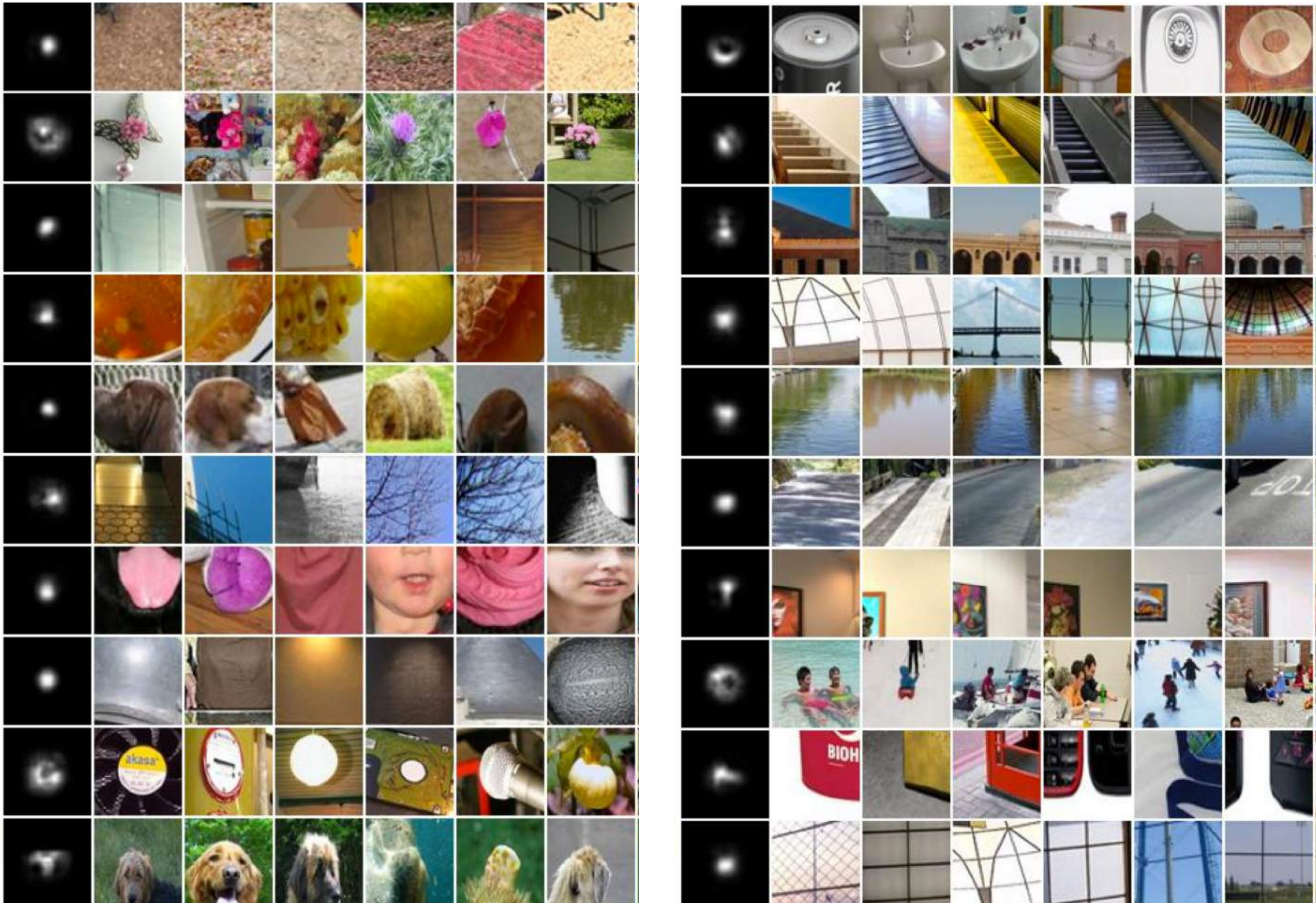
IMAGENET Layer 2 places



IMAGENET

Layer 4

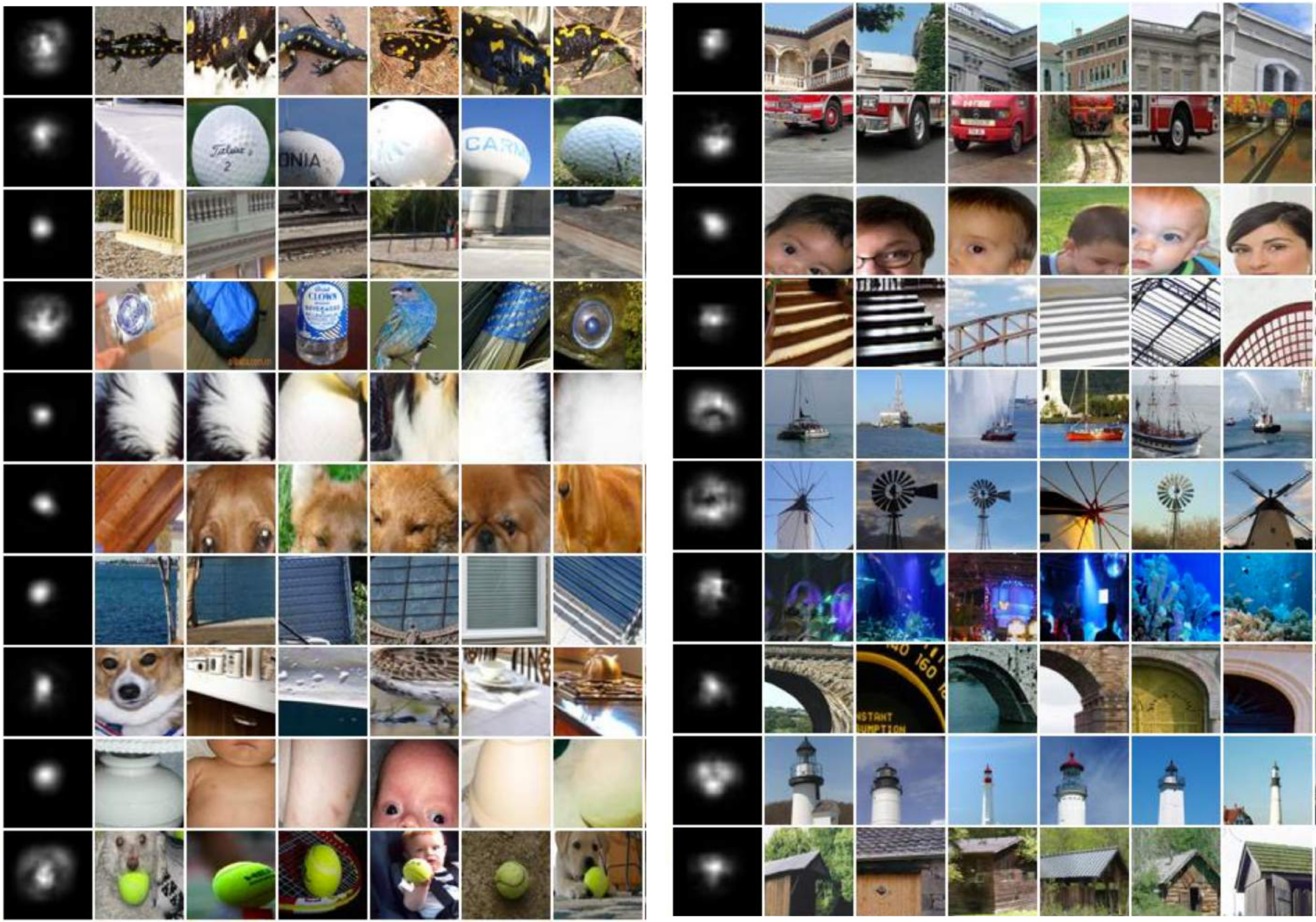
places

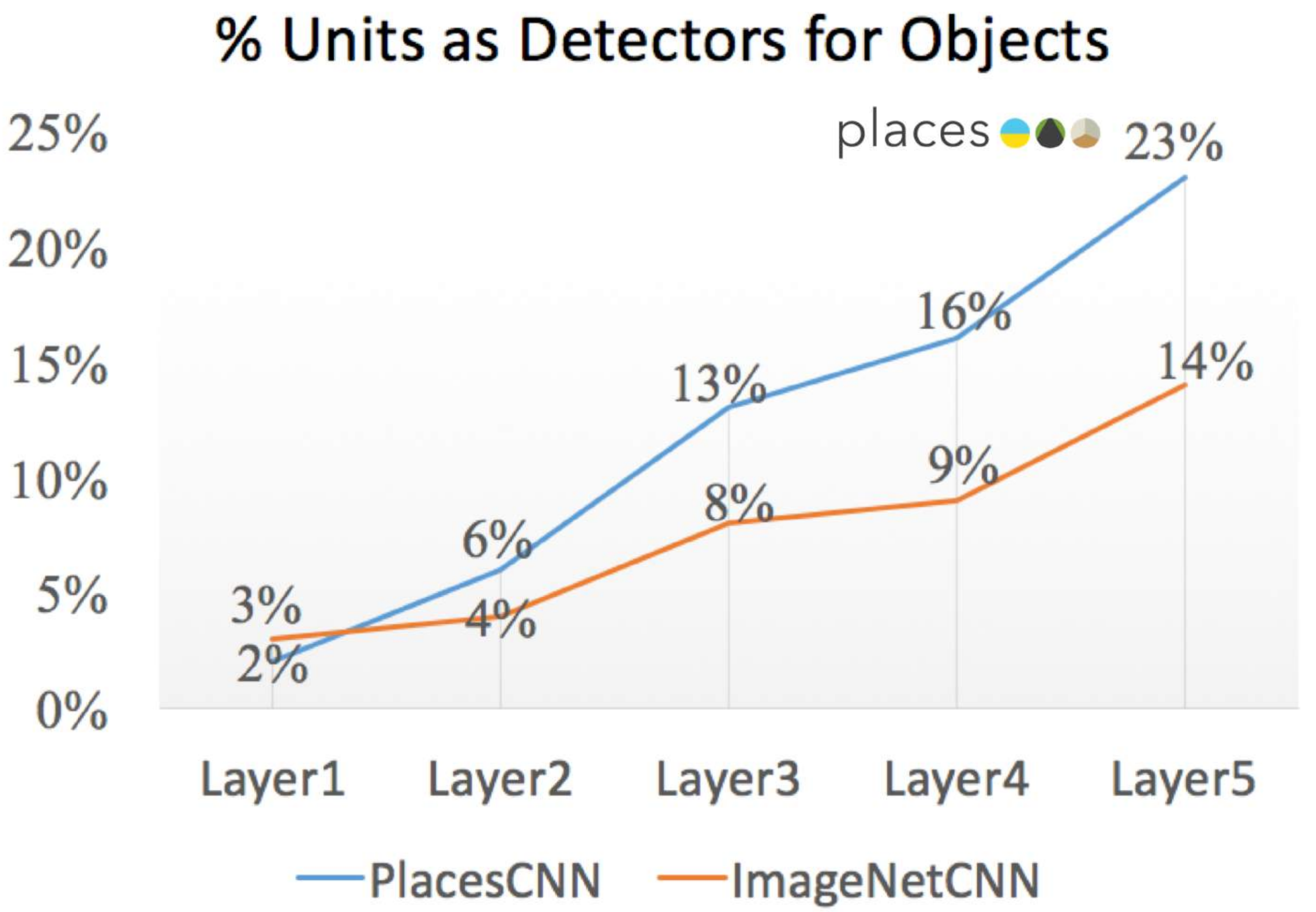


IMAGENET

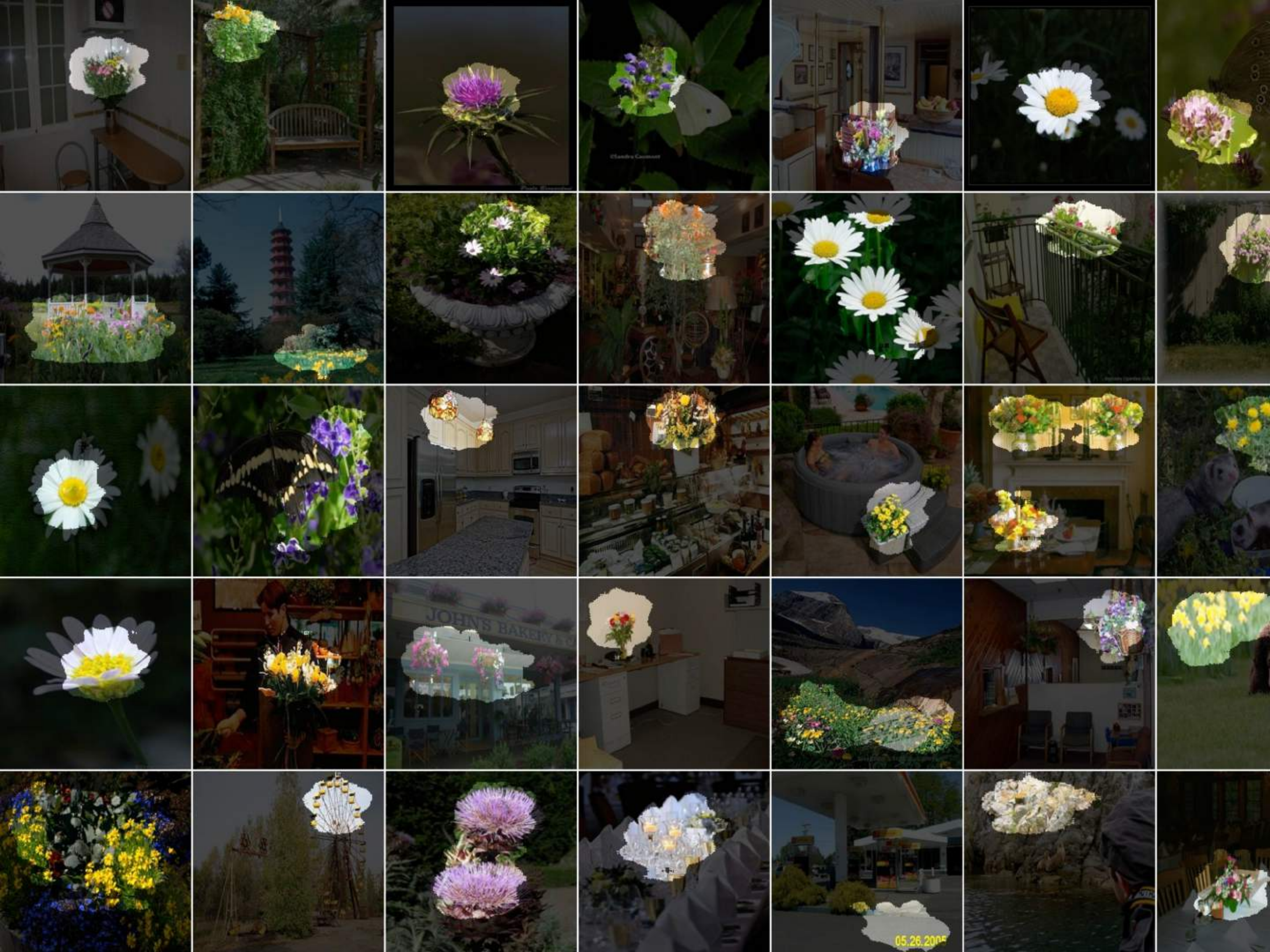
Layer 5

places

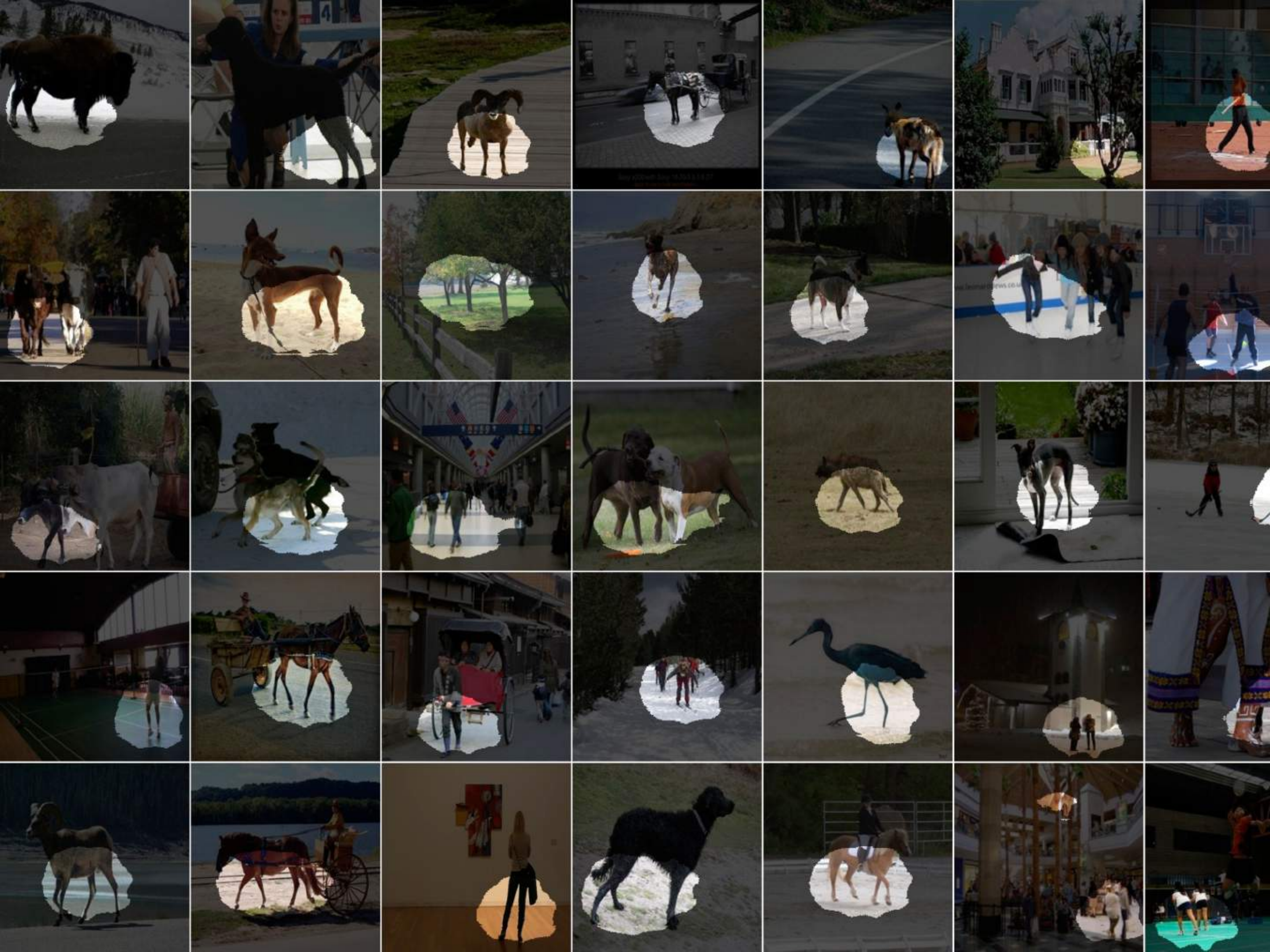


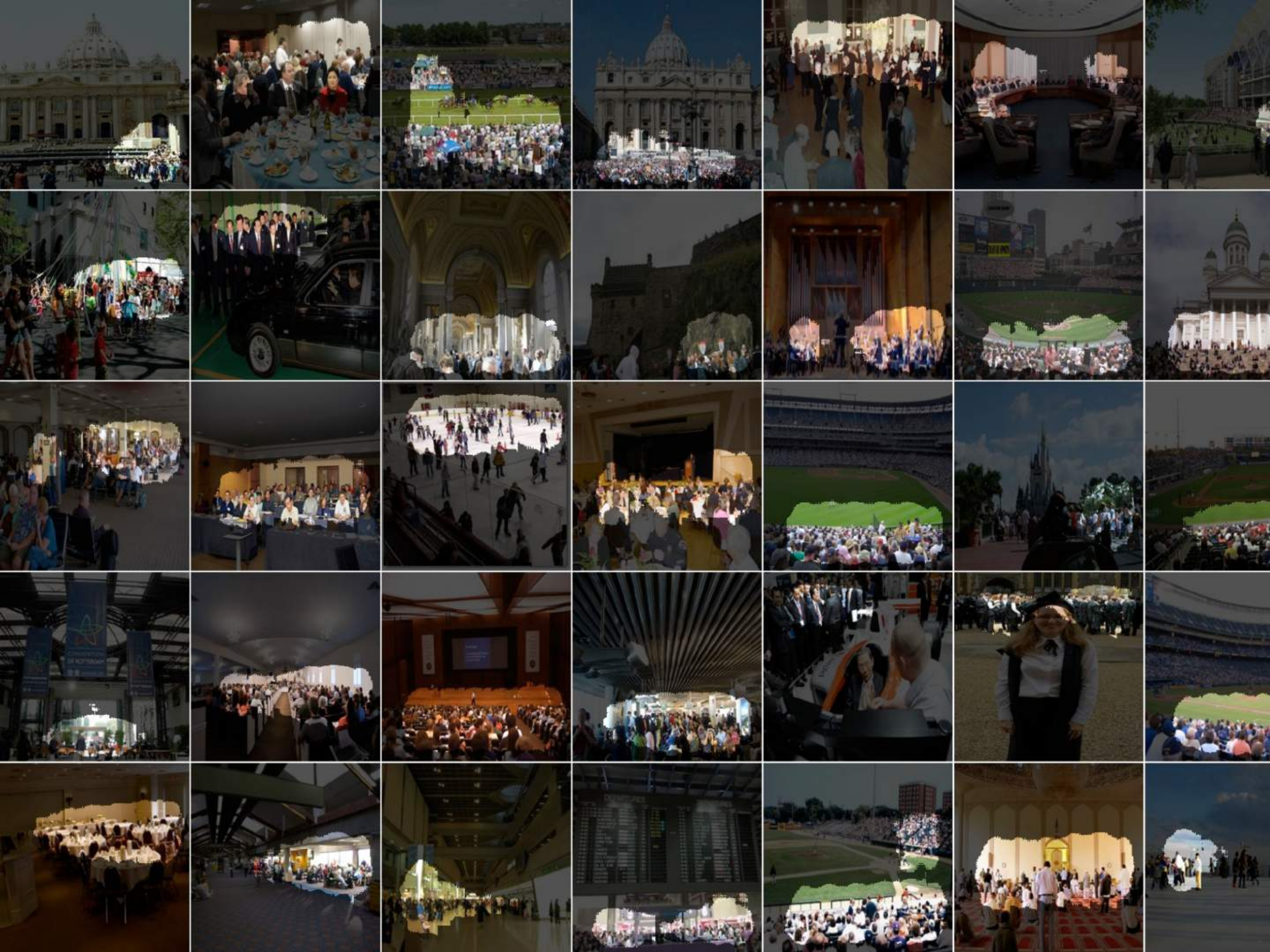

% Units as Detectors for Objects
places
25%
20%
15%
10%
5%
0%
23%
16%
14%
13%
9%
8%
6%
4%
3%
2%
Layer1
Layer2
Layer3
Layer4
Layer5
PlacesCNN
ImageNetCNN

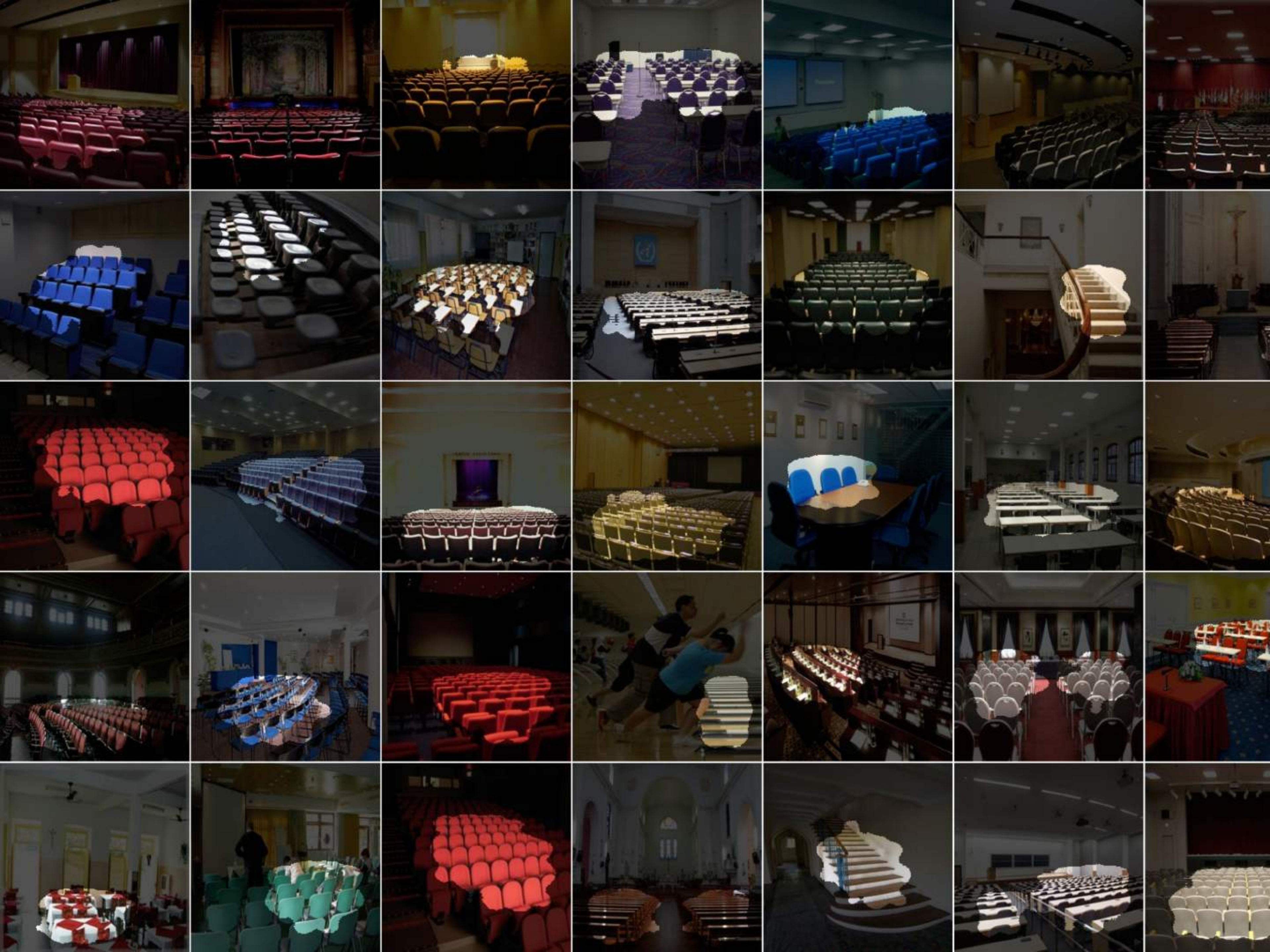

SPORT
50


JOHN'S BAKERY
05.26.2005

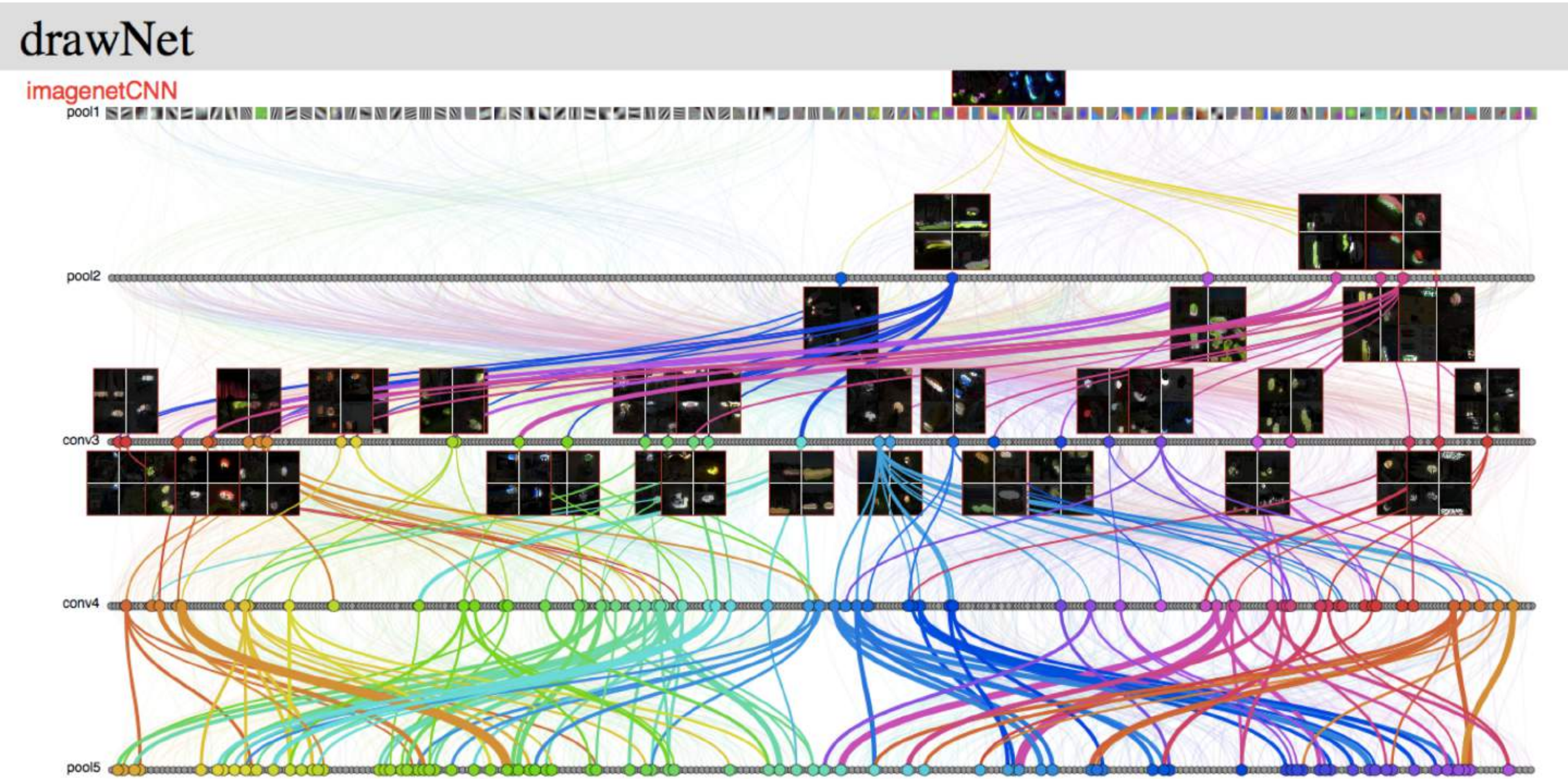

IVAN







Visualizing Units & Connections

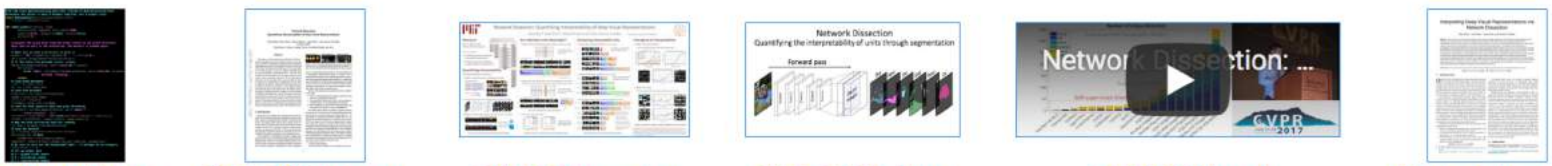


http://people.csail.mit.edu/torralba/research/drawCNN/drawNet.html

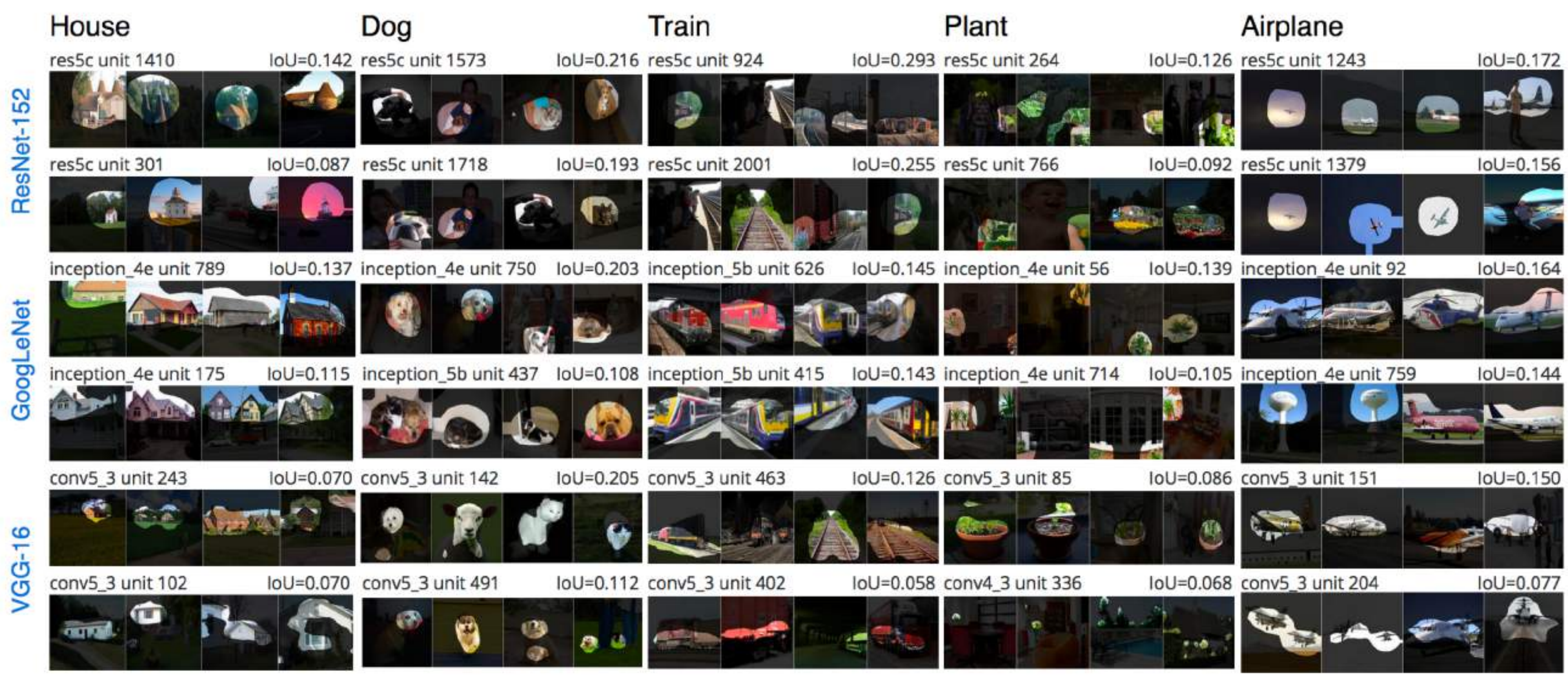
Network Dissection: Quantifying Interpretability of Deep Visual Representations

David Bau*, Bolei Zhou*, Aditya Khosla, Aude Oliva, Antonio Torralba
Massachusetts Institute of Technology

http://netdissect.csail.mit.edu/

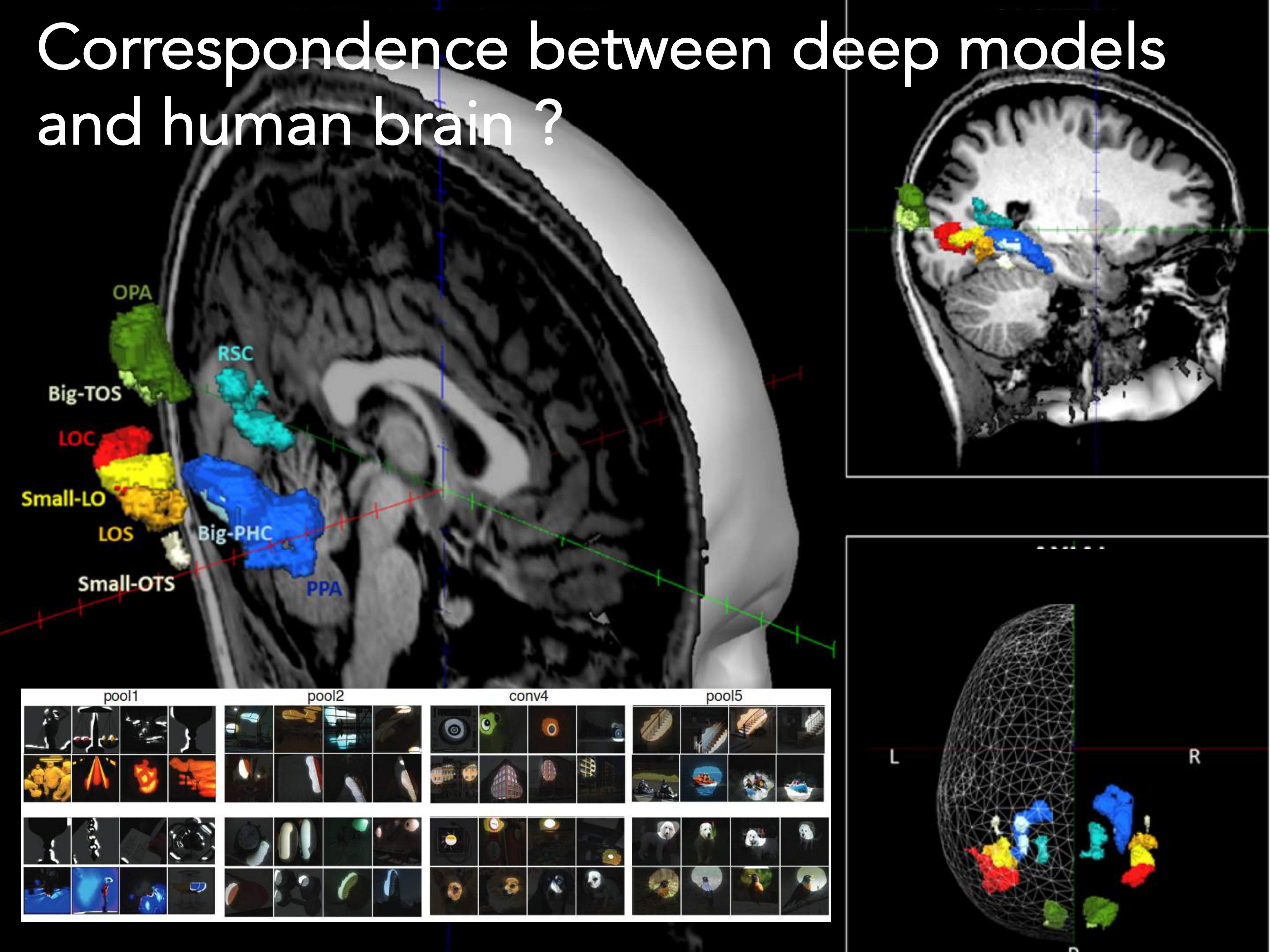


Code and Data | CVPR 2017 paper | CVPR 2017 poster | CVPR 2017 slides | CVPR 2017 oral | Extended paper



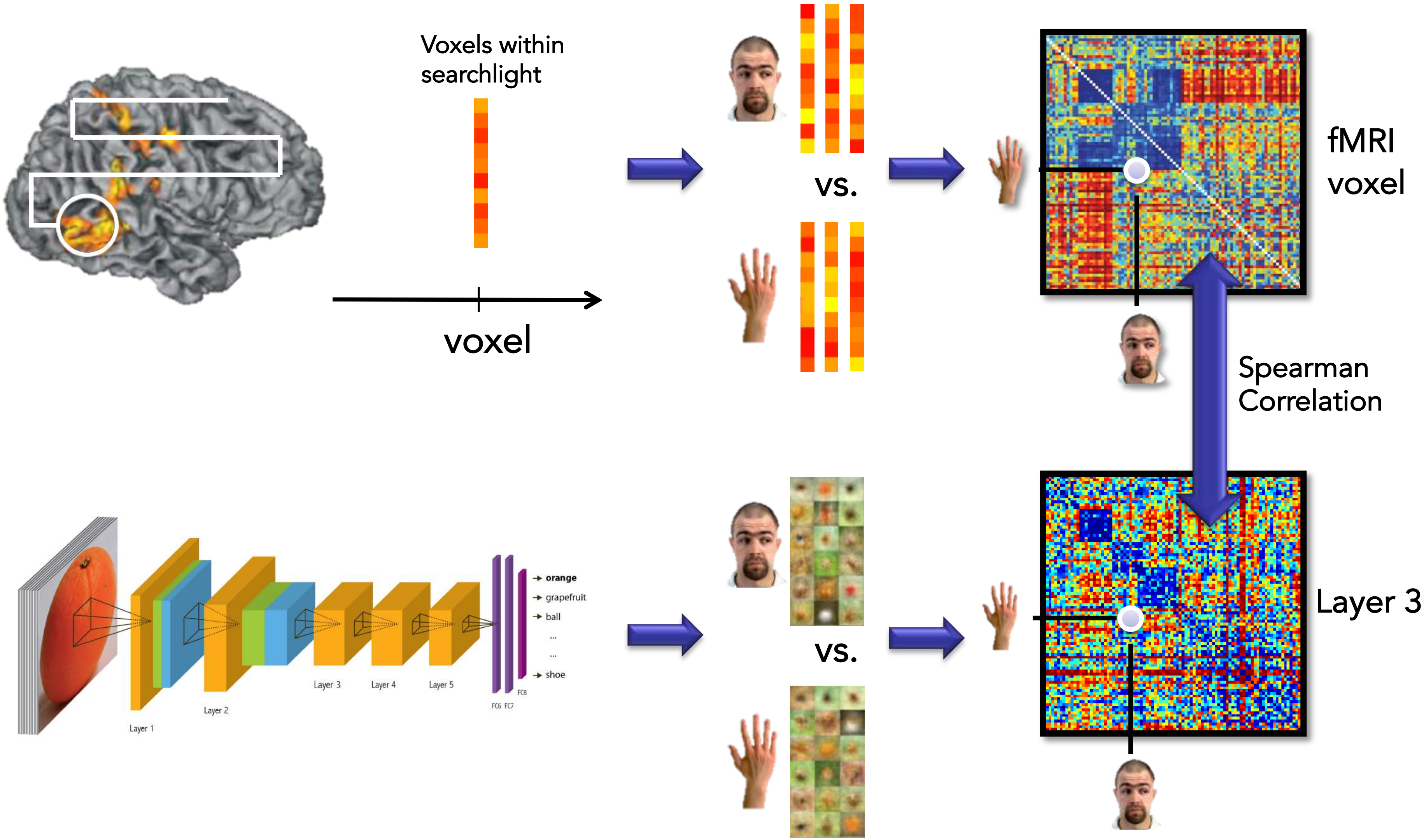
Selected units are shown from three state-of-the-art network architectures when trained to classify images of places (places-365). Many individual units respond to specific high-level concepts (object segmentations) that are not directly represented in the training set (scene classifications).

Correspondence between deep models and human brain ?



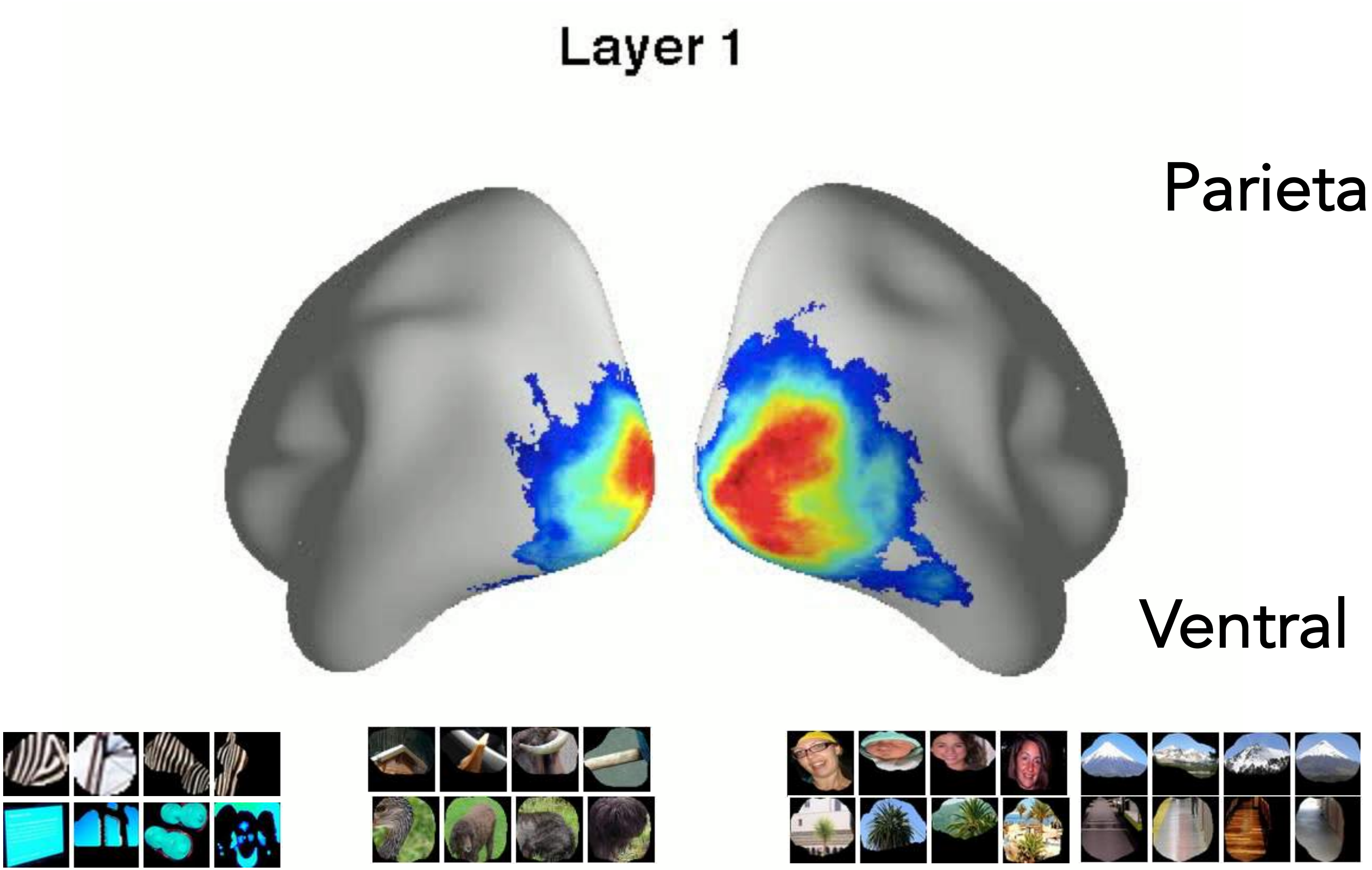
Algorithmic-specific fMRI searchlight analysis

A spatially unbiased view of the relations in similarity structure between models and fMRI



Cichy, Khosla, Pantazis, Torralba & Oliva, A. (2016). Scientific Reports.

Spatiotemporal maps of correlations between human brain and model layers



Layers 1-2 | Layers 2-4 | Layers 5-8

Cichy, Khosla, Pantazis, Torralba & Oliva, A. (2016). Scientific Reports.

Comparing Natural and Artificial Deep Neural Networks

- **New fields of expertise:** Cognitive / Clinical / Social / Perceptual **Computational Experimentalist**
- **Studying the implementation** that works best for performing specific tasks
- **Characterizing the network behavior** when it is adapting, compromised or enhanced
- **Exploring the alternatives** that have not been taken by biological systems

Bolei Zhou

David Bau

Aditya Khosla

Radoslaw Cichy

Antonio Torralba