Learning to see

Antonio Torralba

Computer Science and Artificial Intelligence Laboratory (CSAIL)
Department of Electrical Engineering and Computer Science

Exciting times for computer vision

A bit of history…

The early optimism (1960-1970)

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PROJECT MAC

Artificial Intelligence Group July 7, 1966

Vision Memo. No. 100.

THE SUMMER VISION PROJECT

Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

50 years ago…

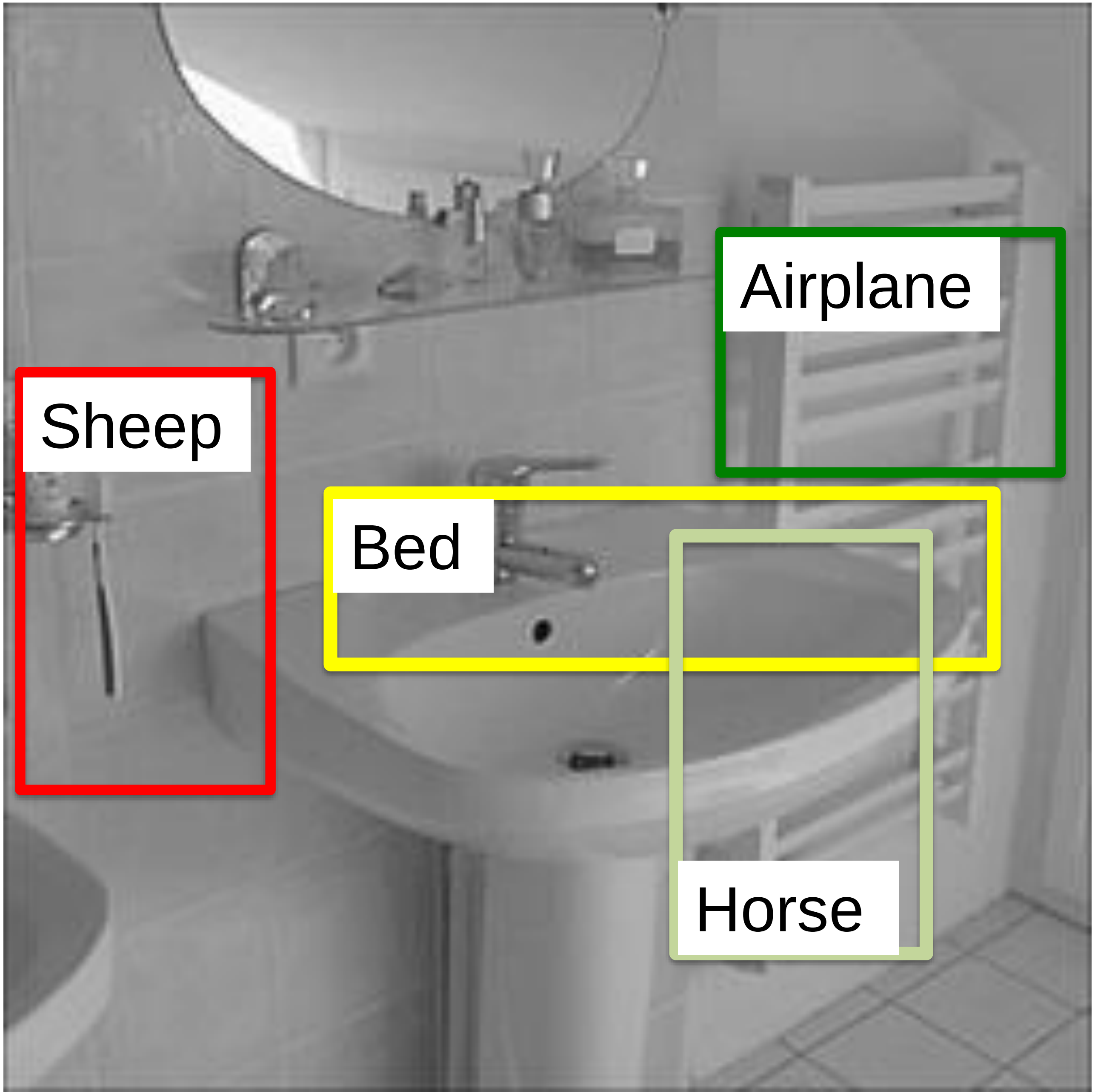
50 years ago…

50 years ago…

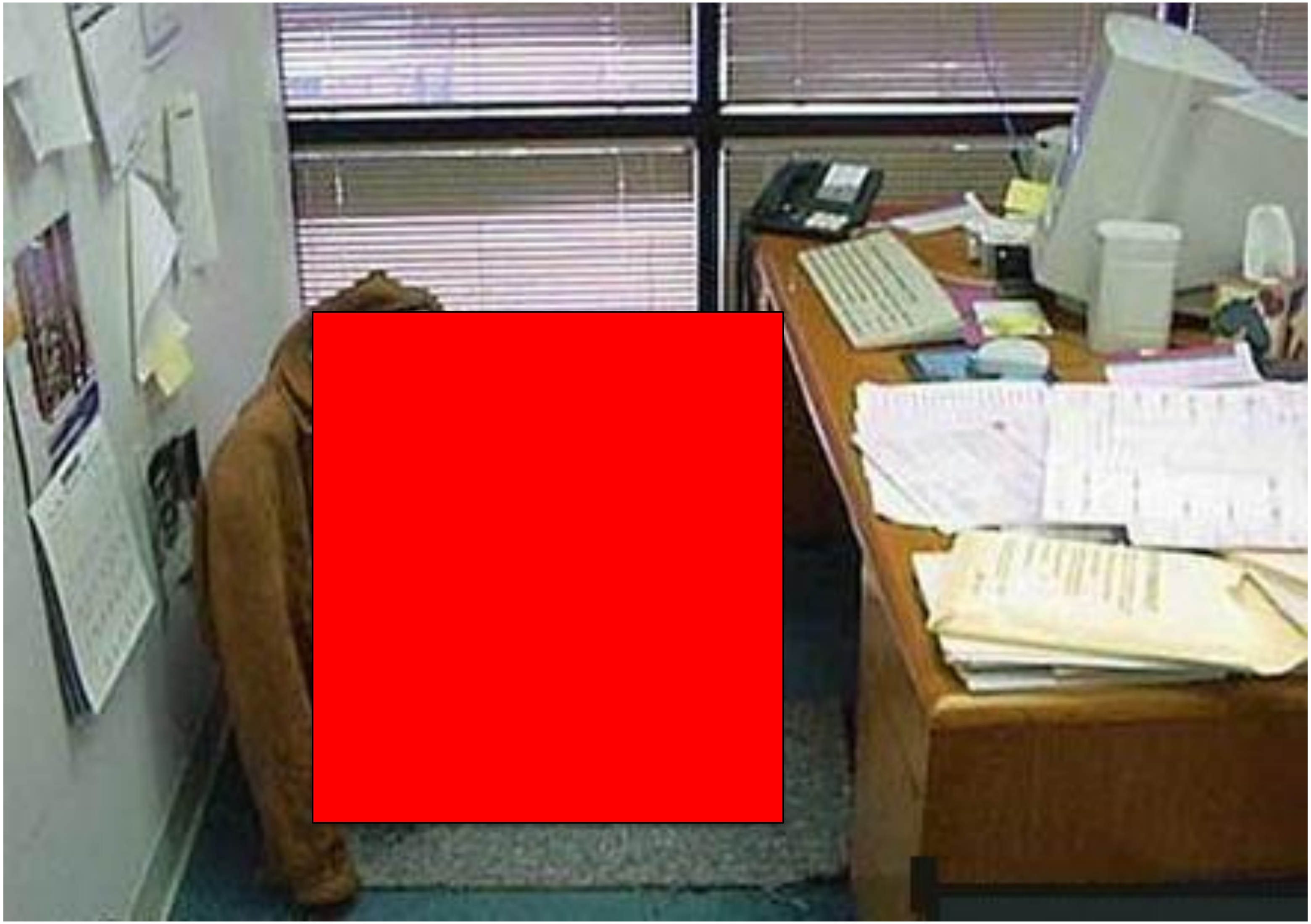
25 years ago…

25 years ago…

25 years ago...

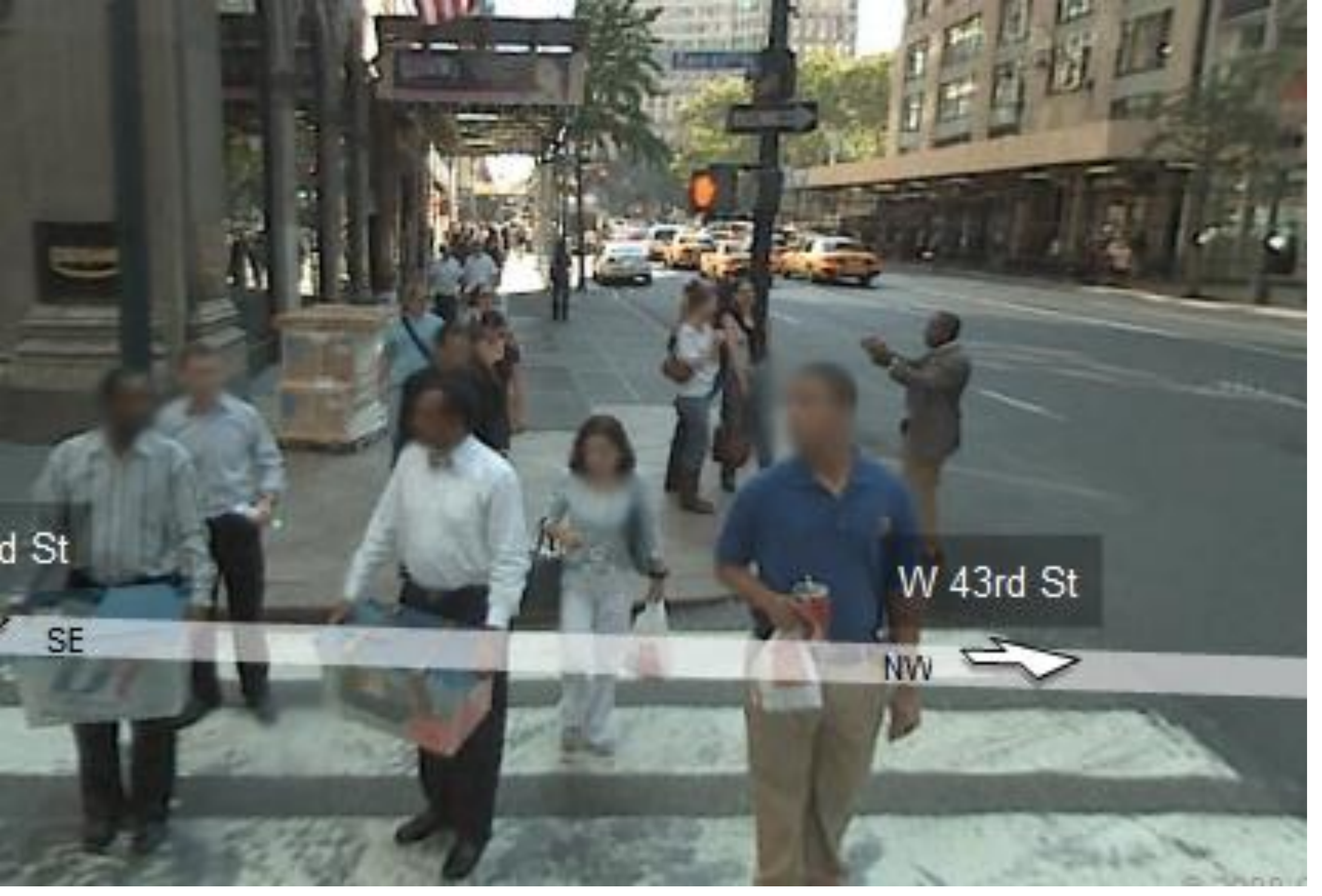


The vision crisis (1970-2000)



But 15 years ago…

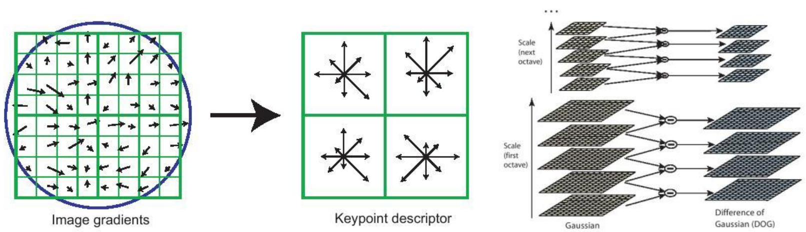
• The representation and matching of pictorial structures Fischler, Elschlager (1973).
• Face recognition using eigenfaces M. Turk and A. Pentland (1991).
• Human Face Detection in Visual Scenes - Rowley, Baluja, Kanade (1995)
• Graded Learning for Object Detection - Fleuret, Geman (1999)
• Robust Real-time Object Detection - Viola, Jones (2001)
• Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images - Heisele, Serre, Mukherjee, Poggio (2001)
•….



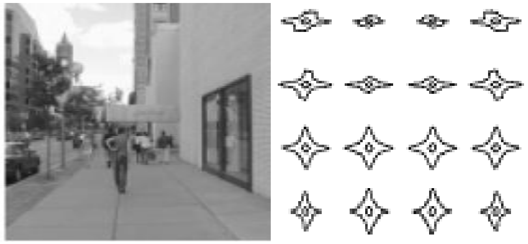
- The representation and matching of pictorial structures Fischler, Elschlager (1973).
- Face recognition using eigenfaces M. Turk and A. Pentland (1991).
- Human Face Detection in Visual Scenes - Rowley, Baluja, Kanade (1995)
- Graded Learning for Object Detection - Fleuret, Geman (1999)
- Robust Real-time Object Detection - Viola, Jones (2001)
- Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images - Heisele, Serre, Mukherjee, Poggio (2001)
- ….

- The representation and matching of pictorial structures Fischler, Elschlager (1973).
- Face recognition using eigenfaces M. Turk and A. Pentland (1991).
- Human Face Detection in Visual Scenes - Rowley, Baluja, Kanade (1995)
- Graded Learning for Object Detection - Fleuret, Geman (1999)
- Robust Real-time Object Detection - Viola, Jones (2001)
- Feature Reduction and Hierarchy of Classifiers for Fast Object Detection in Video Images - Heisele, Serre, Mukherjee, Poggio (2001)
- ….

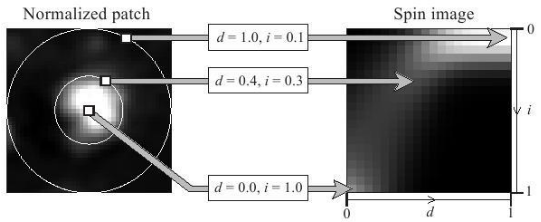
Advances in computer vision



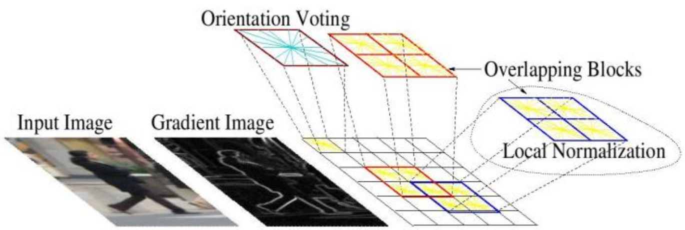
SIFT



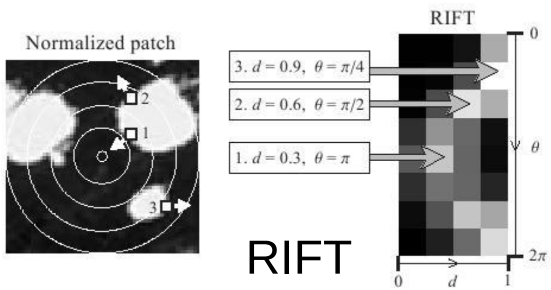
GIST



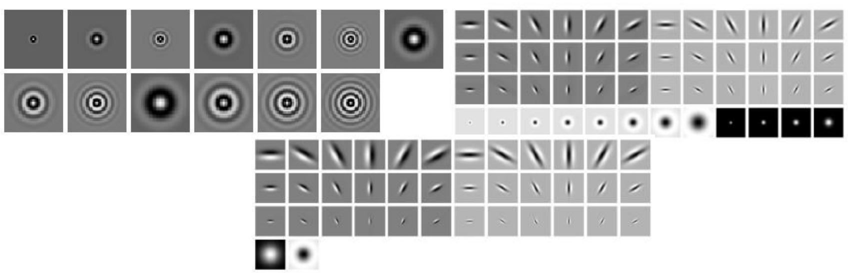
Spin image



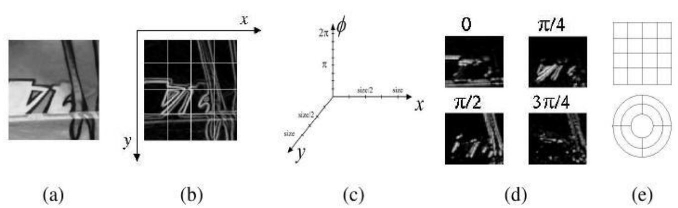
HoG



RIFT



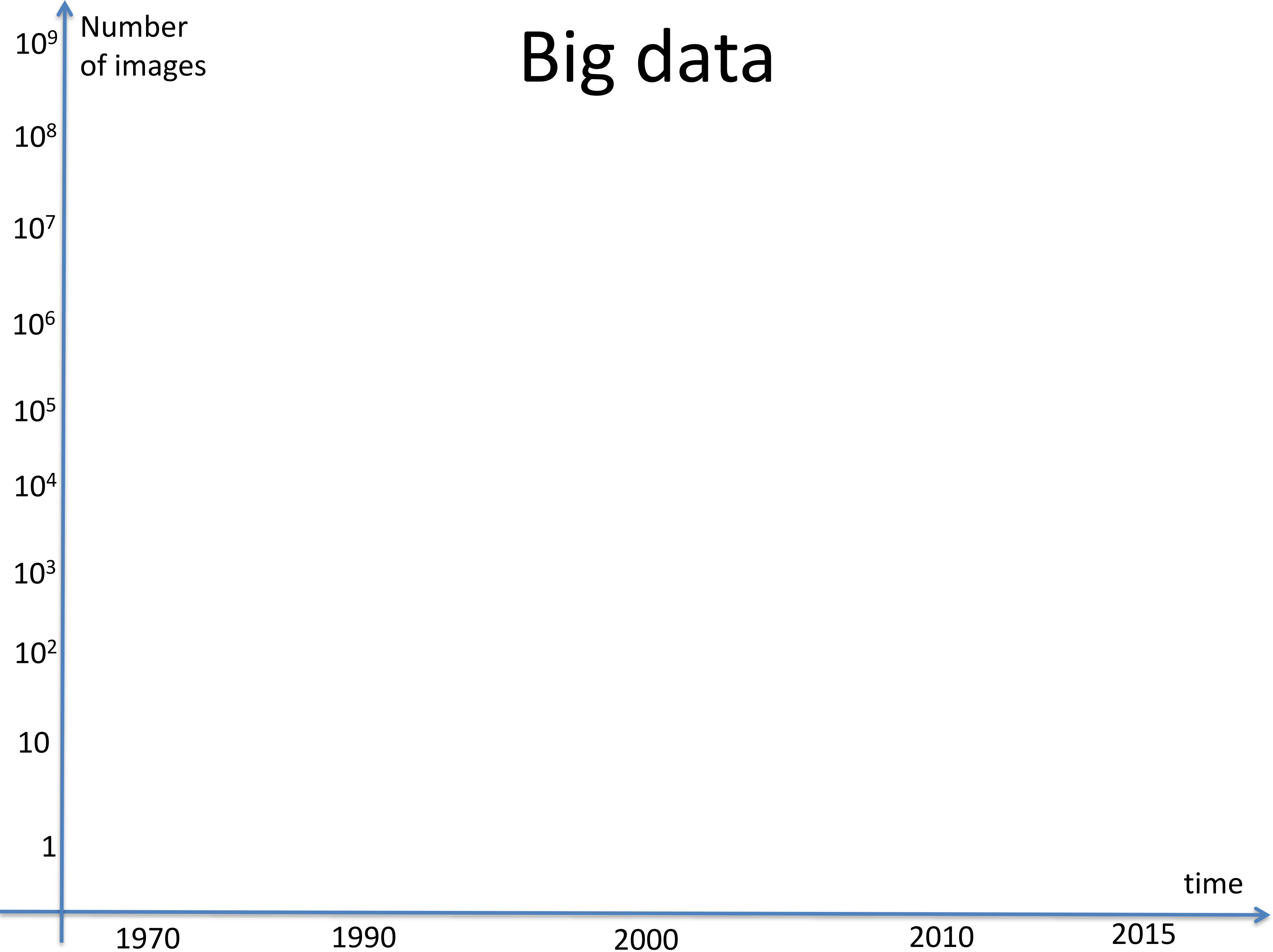
Textons



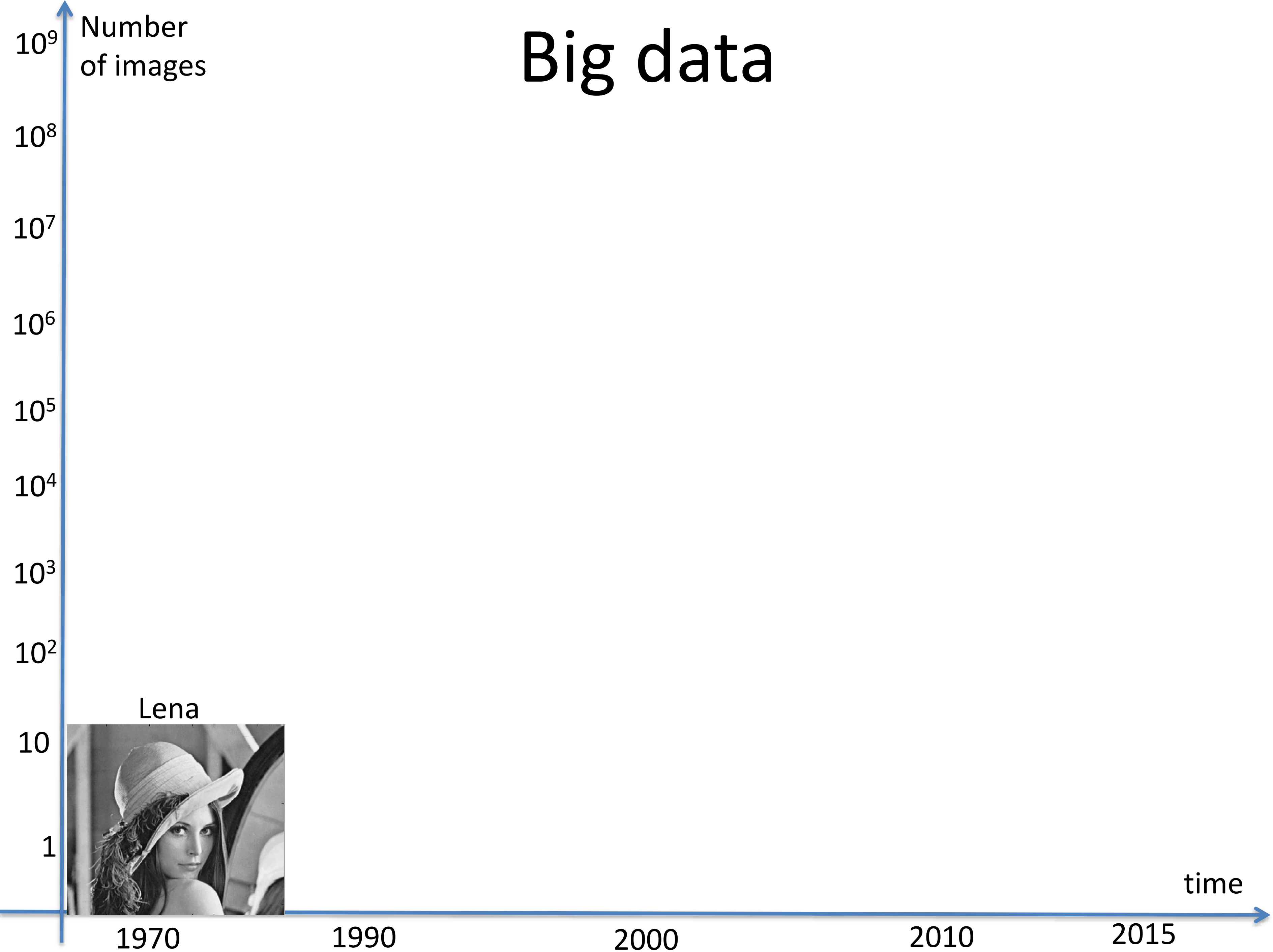
GLOH

A short story of image databases

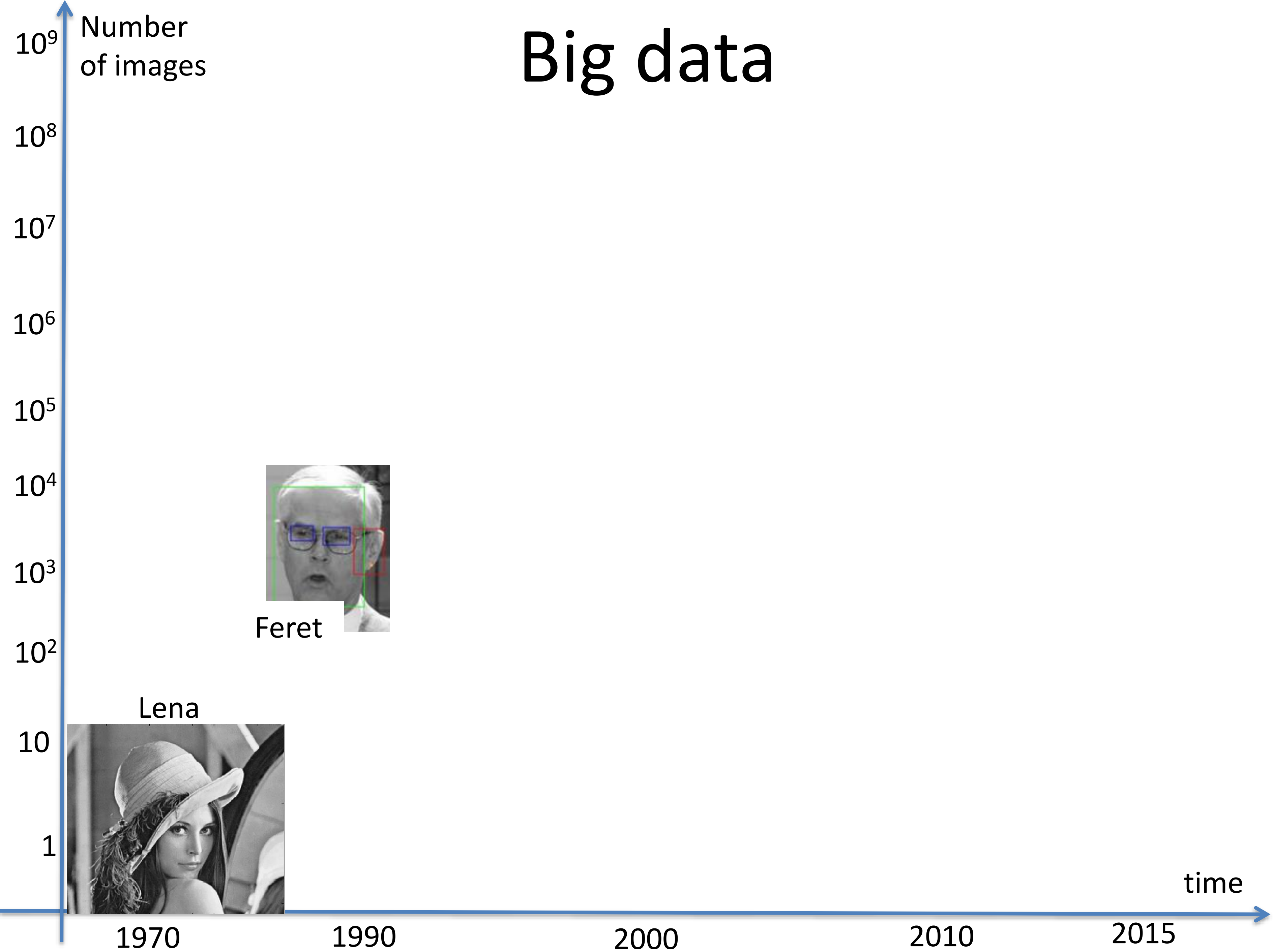
Big data


Number
of images
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
time
1970
1990
2000
2010
2015

Big data

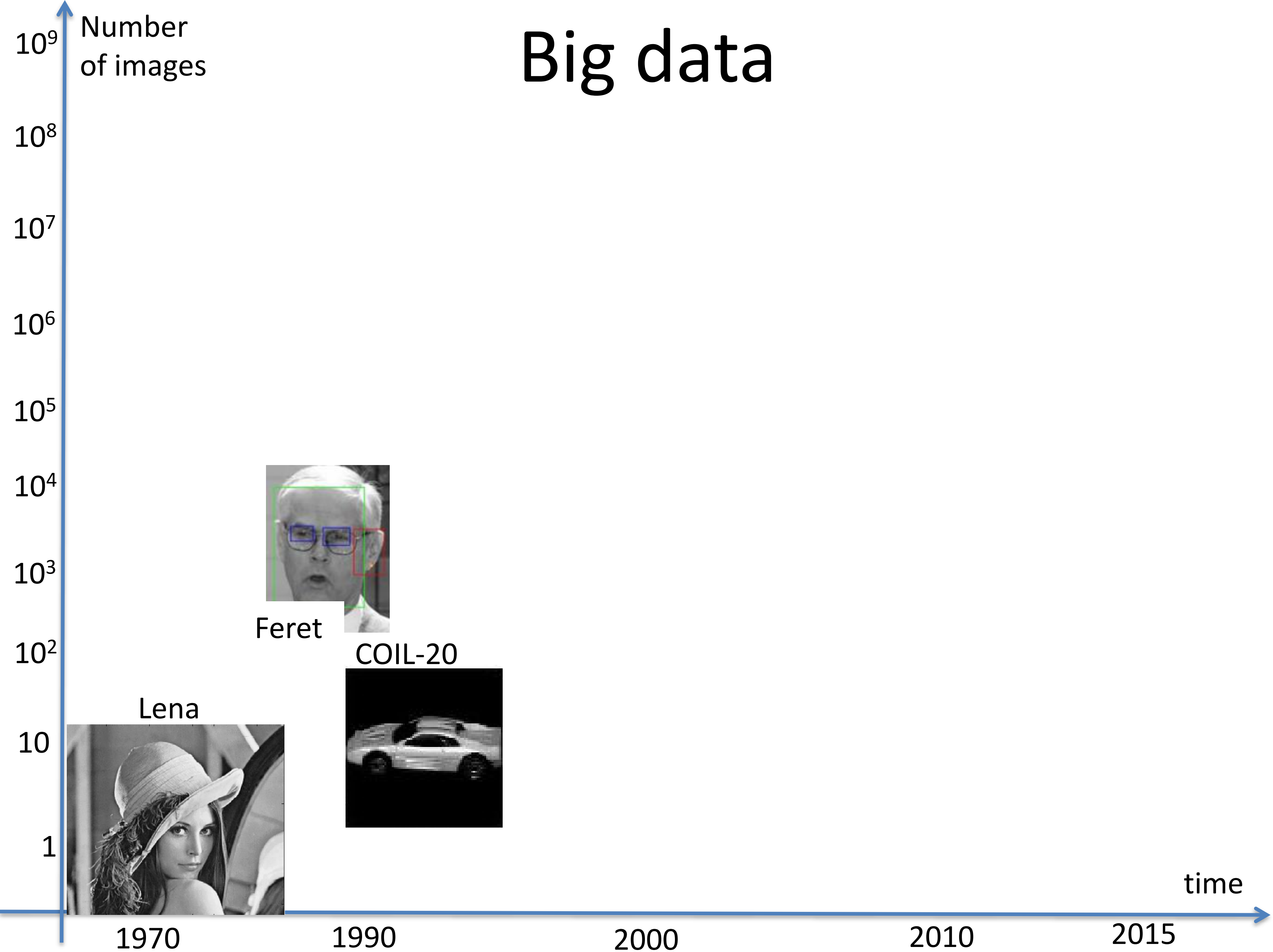

Number
of images
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
Lena
time
1970
1990
2000
2010
2015

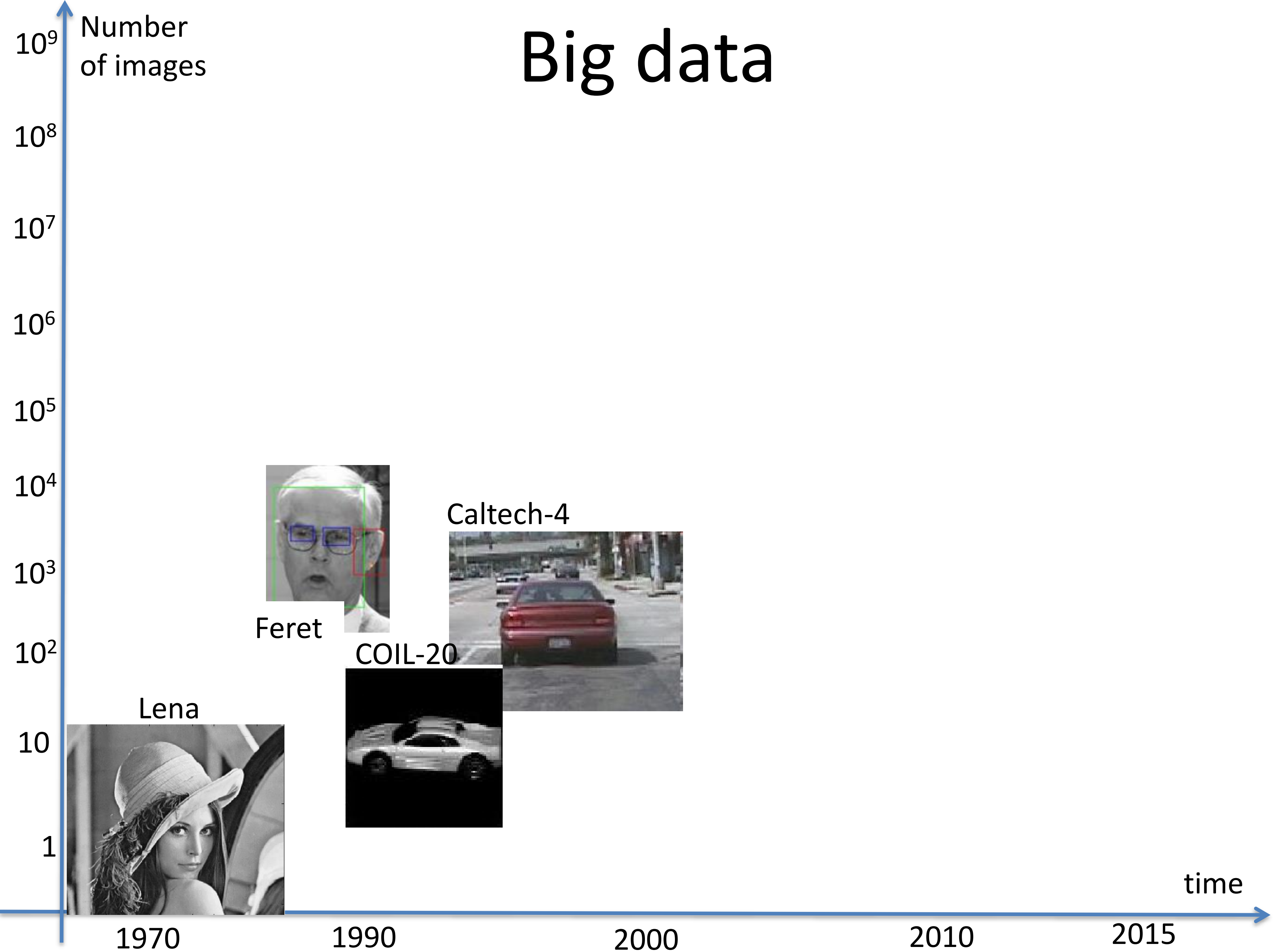
Big data



Number of images
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
Feret
Lena
time
1970
1990
2000
2010
2015

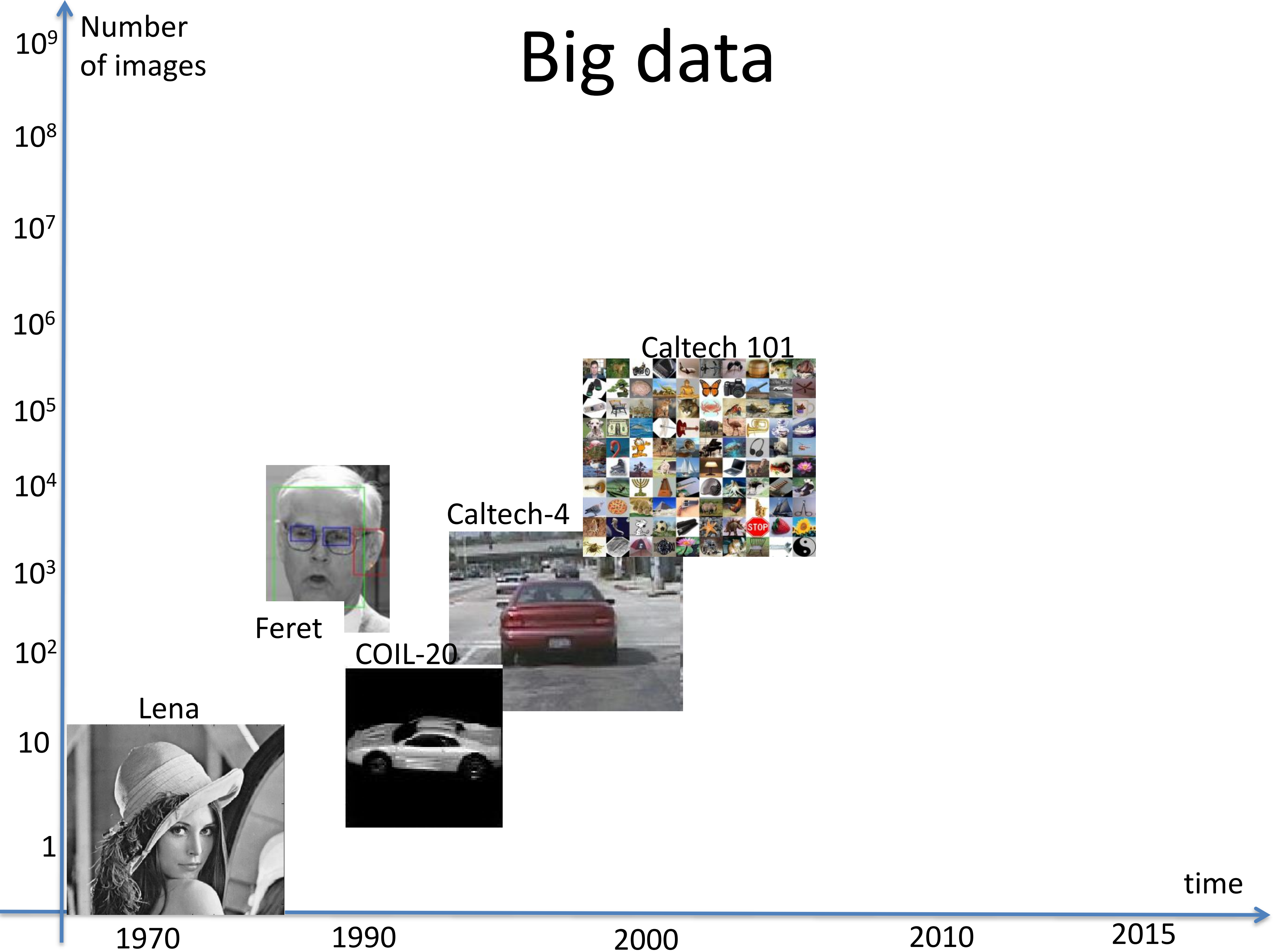
Big data


Number of images
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
Feret
COIL-20
Lena
time
1970
1990
2000
2010
2015

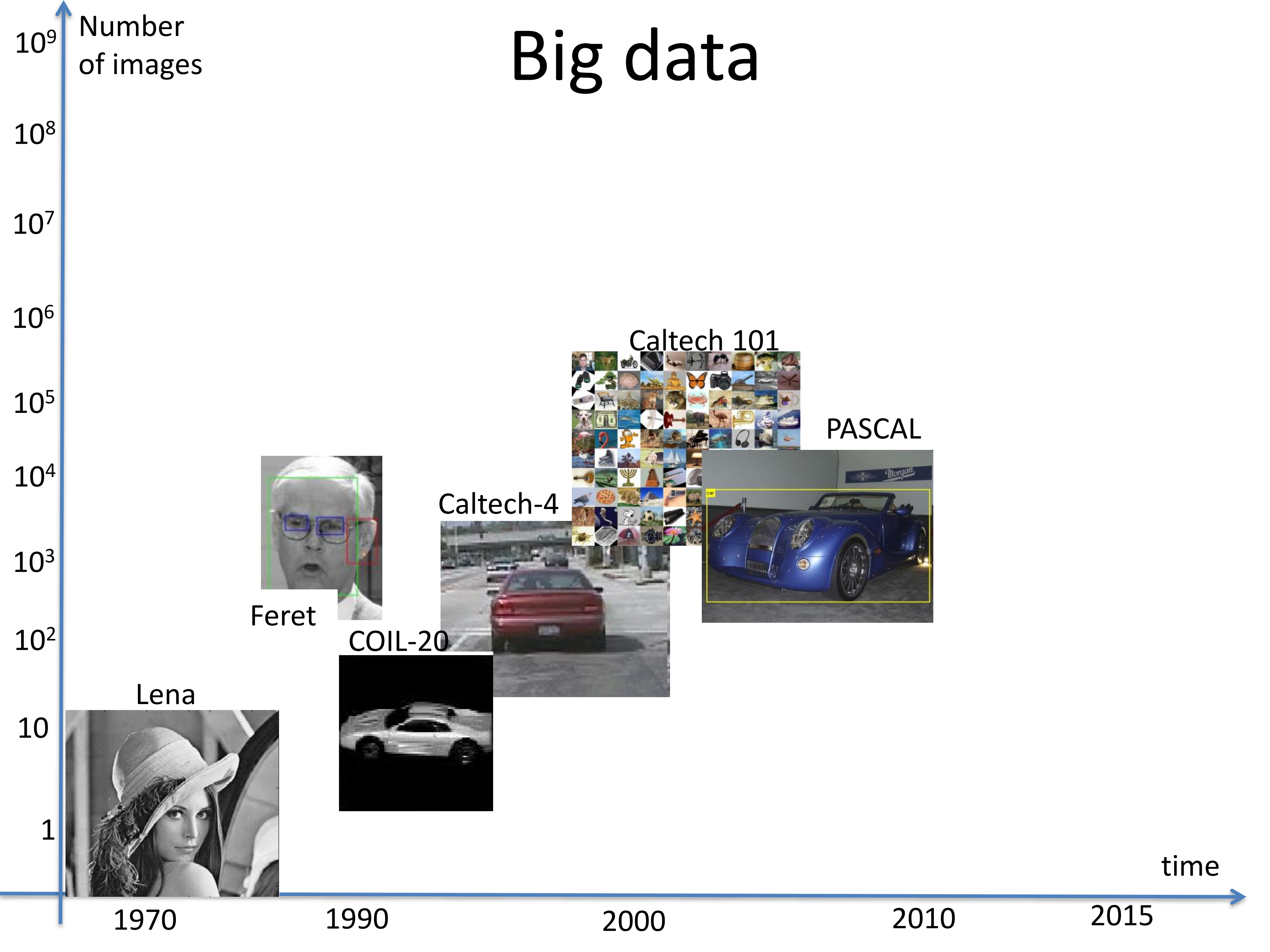


Big data
Number
of images
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
Caltech-4
Feret
COIL-20
Lena
time
1970
1990
2000
2010
2015

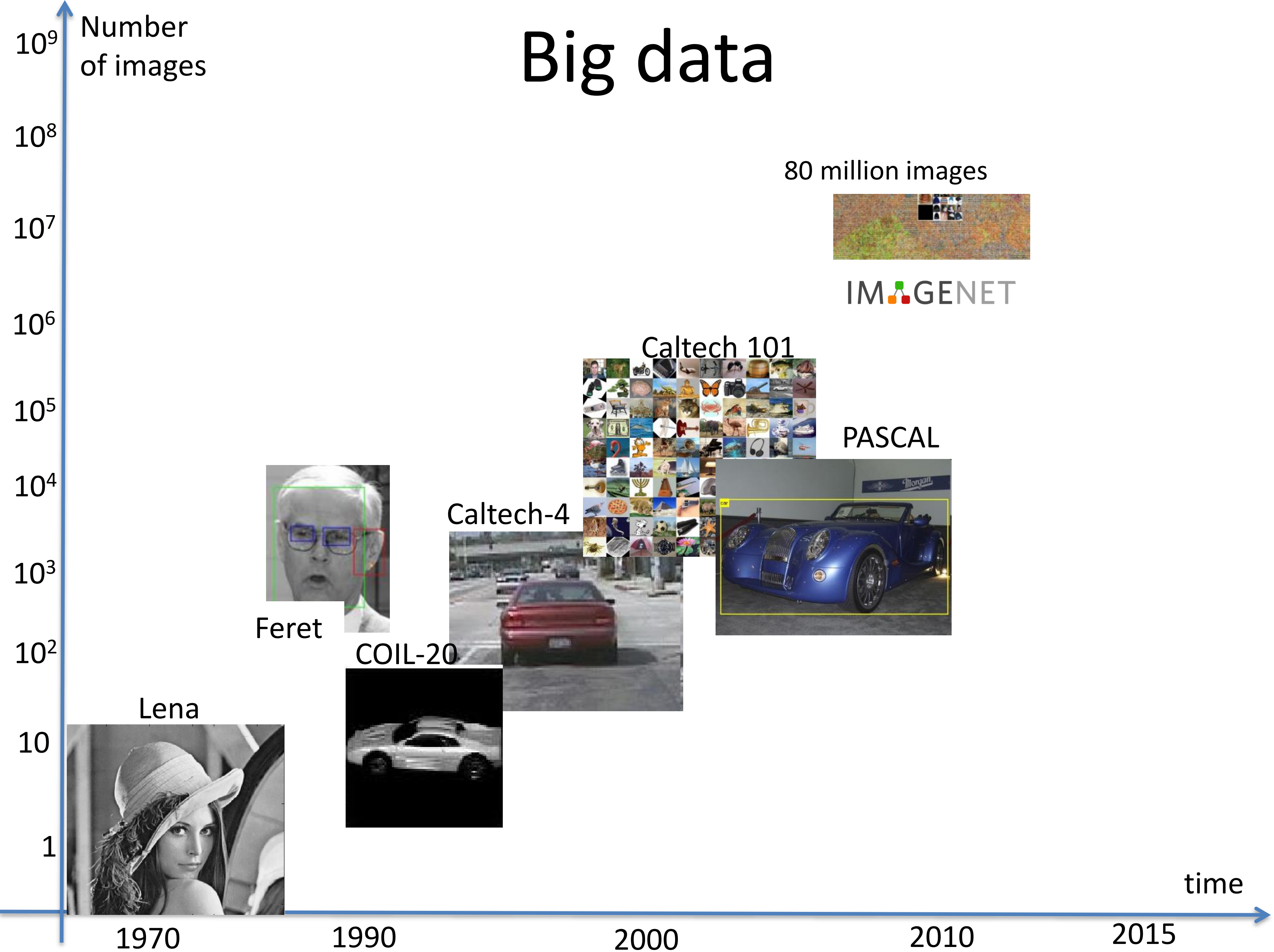
Big data

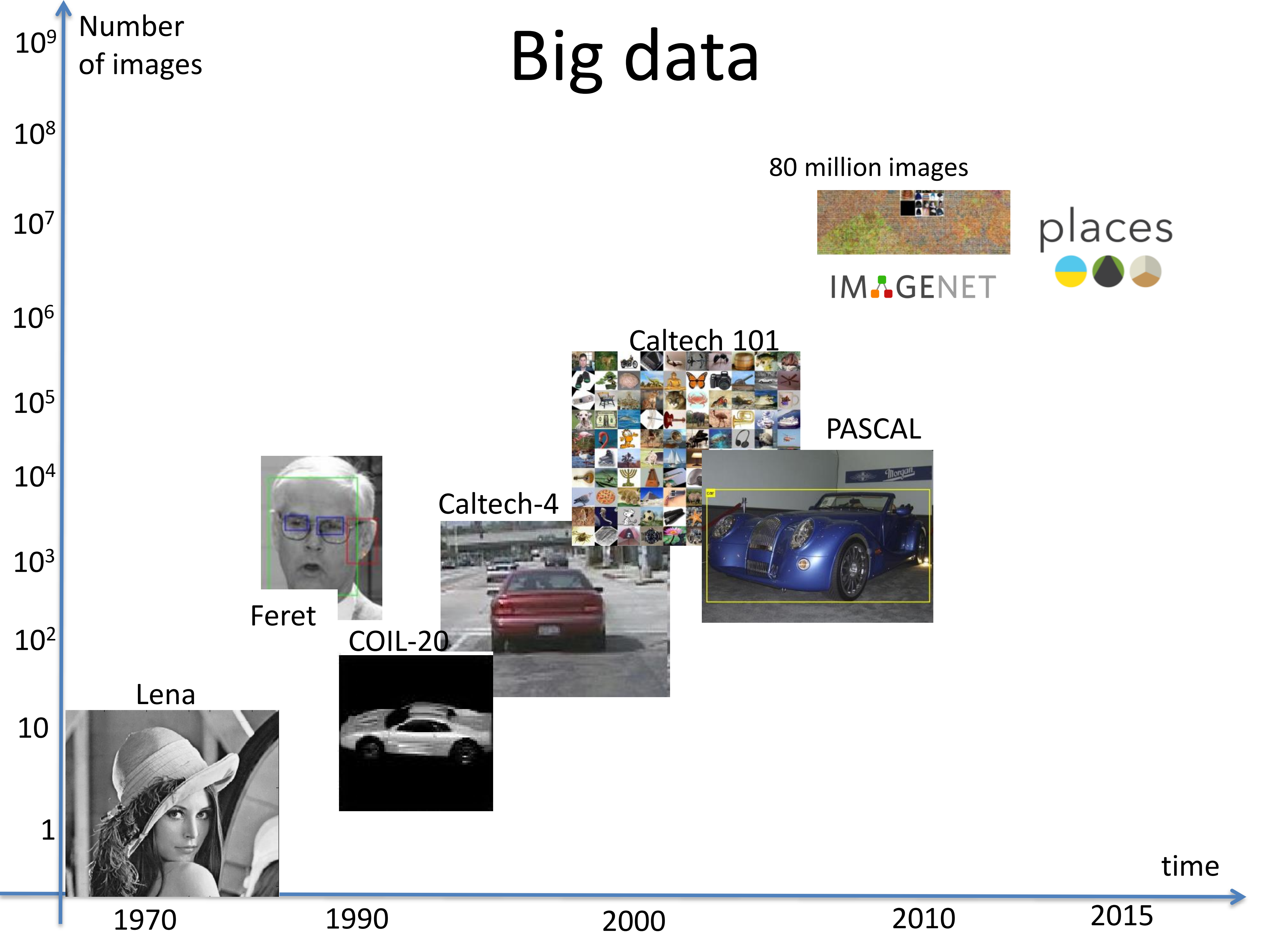

Number
of images
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
Caltech 101
Caltech-4
Feret
COIL-20
Lena
STOP
time
1970
1990
2000
2010
2015

Big data


Number
of images
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
Caltech 101
PASCAL
Caltech-4
Feret
COIL-20
Lena
time
1970
1990
2000
2010
2015

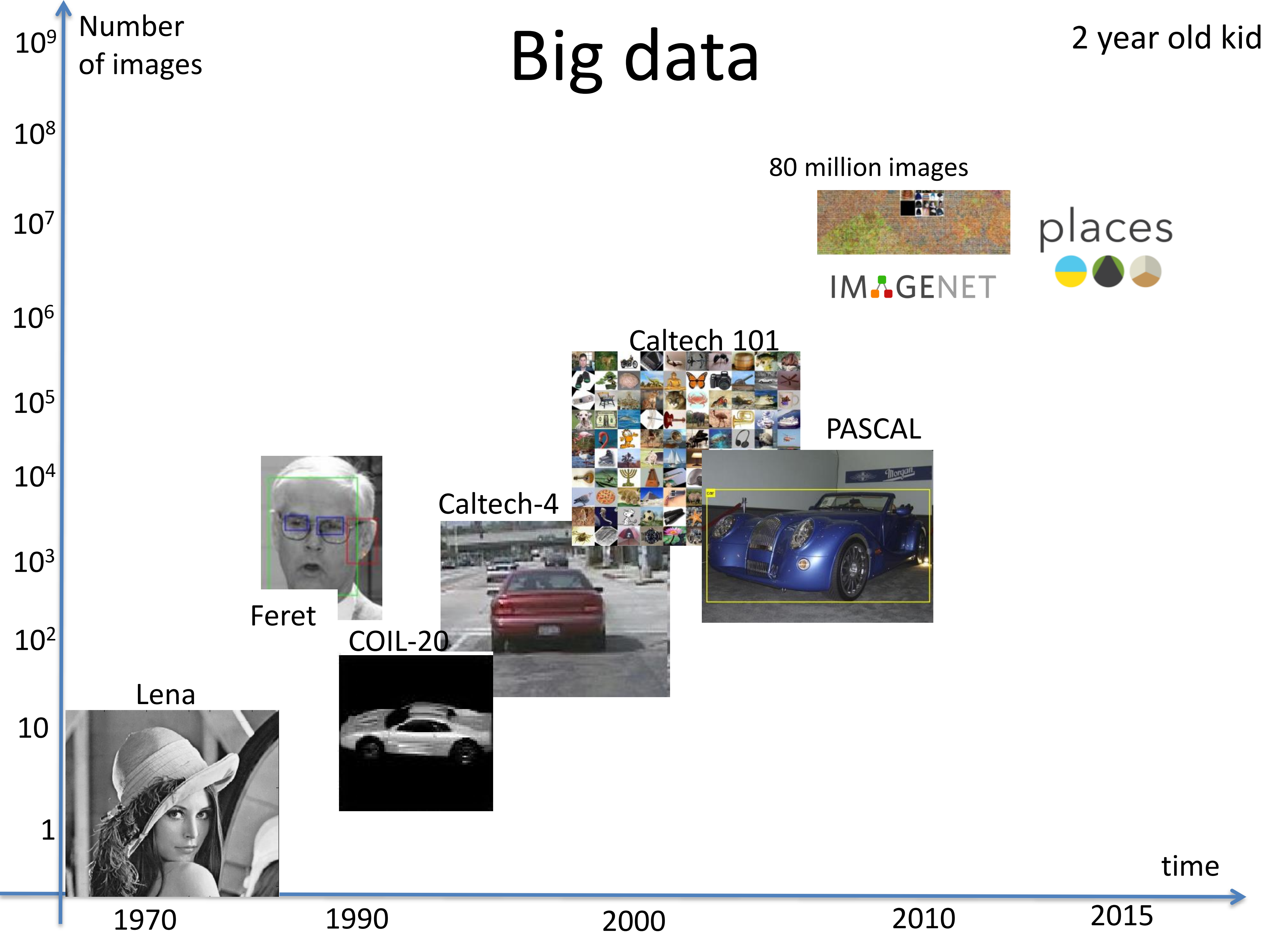
Big data


Number
of images
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
80 million images
IMAGENET
Caltech 101
PASCAL
Caltech-4
Feret
COIL-20
Lena
time
1970
1990
2000
2010
2015



Big data
Number
of images
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
80 million images
IMAGENET
places
Caltech 101
PASCAL
Caltech-4
Feret
COIL-20
Lena
time
1970
1990
2000
2010
2015

Big data


Number
of images
2 year old kid
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2
10
1
80 million images
places
IMAGENET
Caltech 101
PASCAL
Caltech-4
Feret
COIL-20
Lena
time
1970
1990
2000
2010
2015

The time of big data

Ask Images

Cydral Image & Site Search

Google Image Search

webshots beta

picsearch

altavista

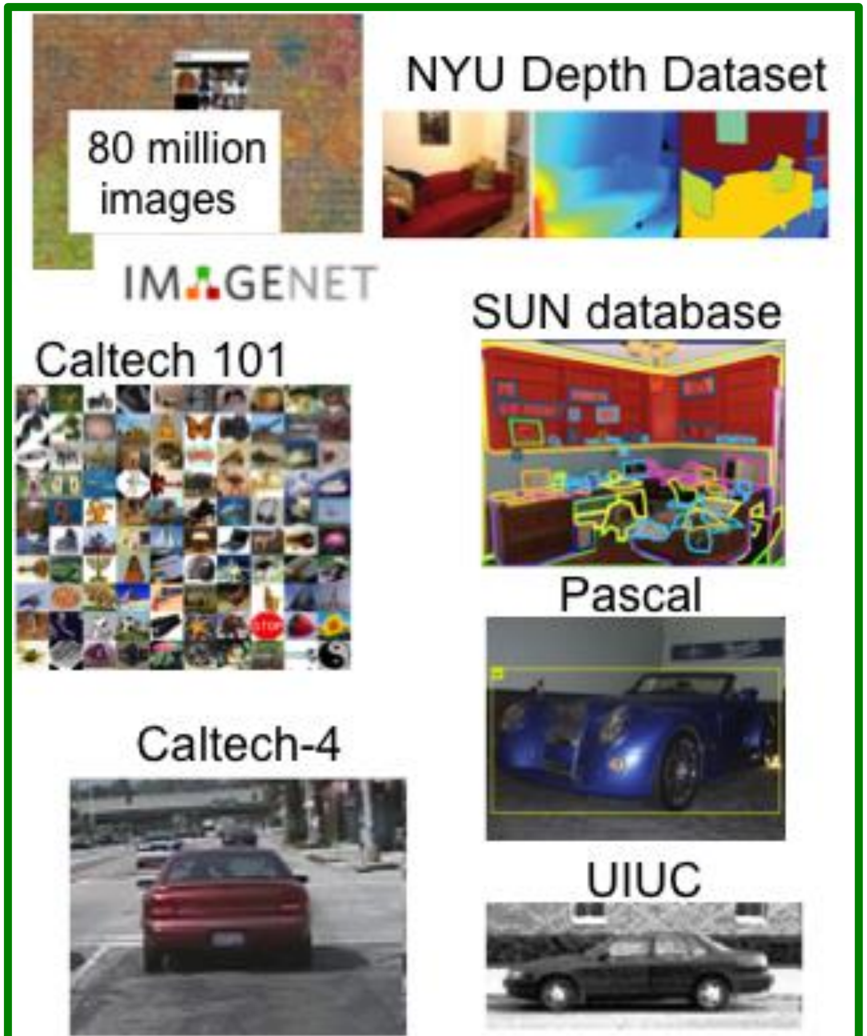
flickr GAMMA

facebook

In 2010, a new student gets into computer vision…

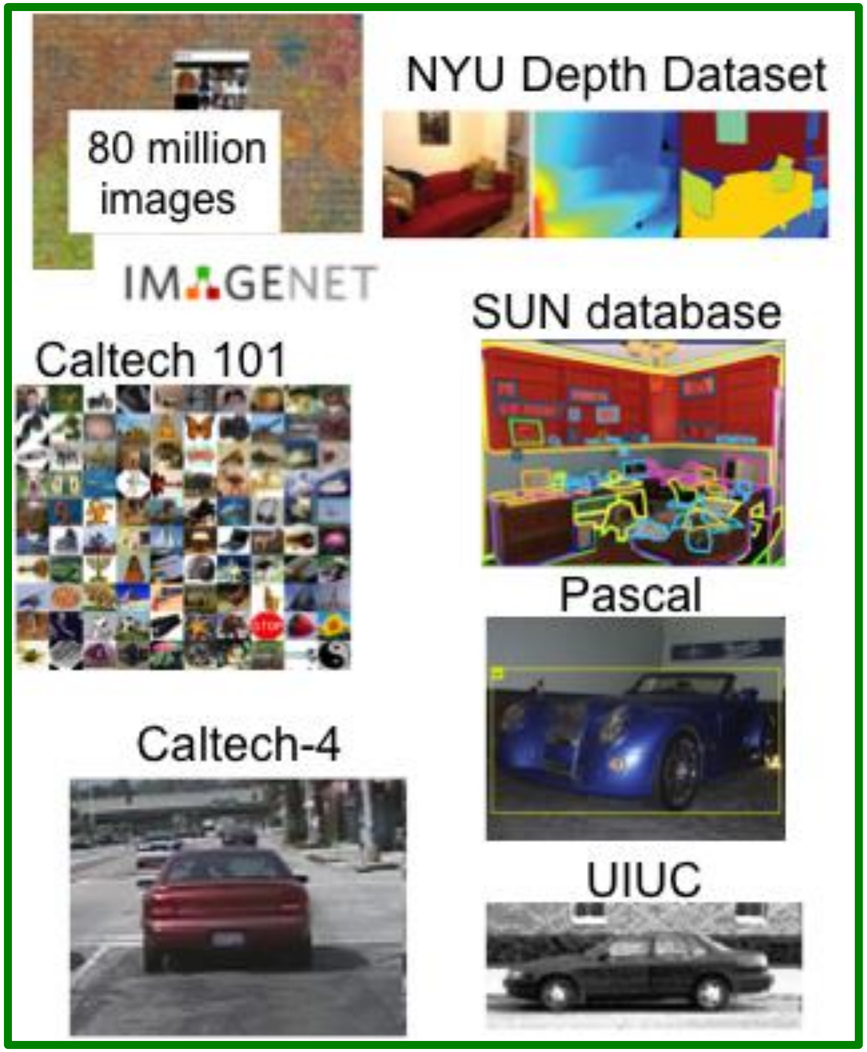
In 2010, a new student gets into computer vision…

Pick one dataset

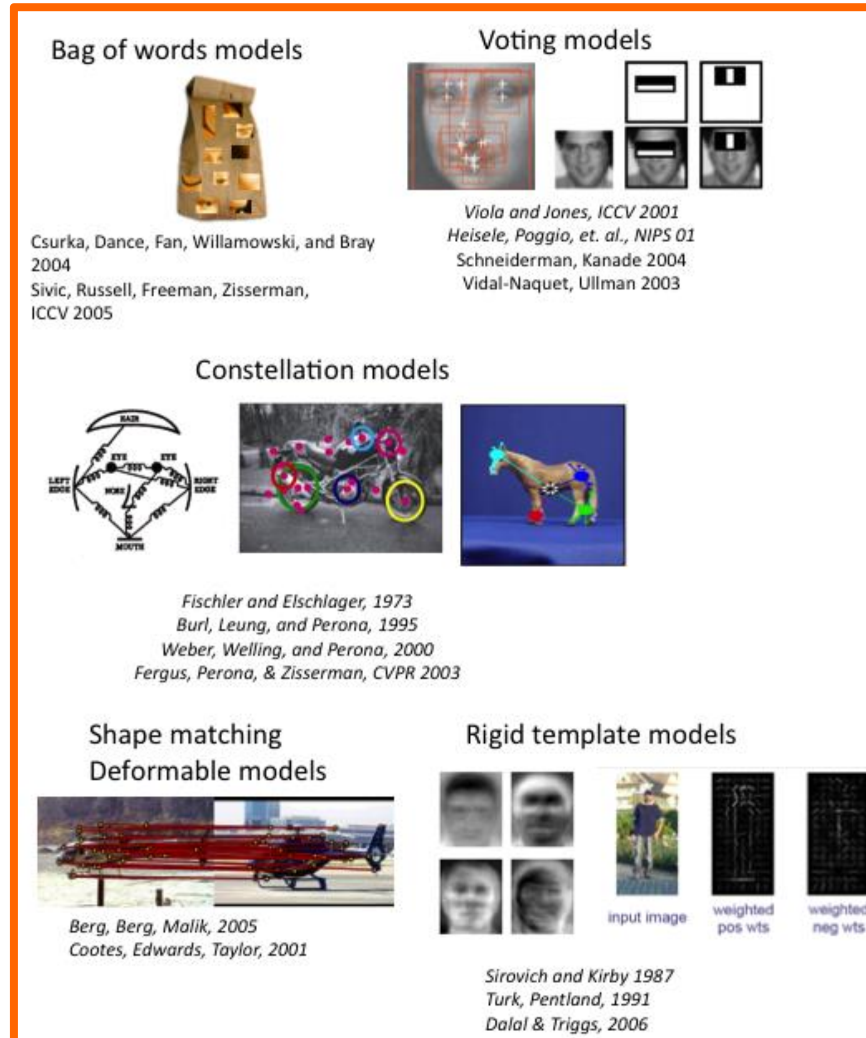


In 2010, a new student gets into computer vision…

Pick one dataset



Pick one model

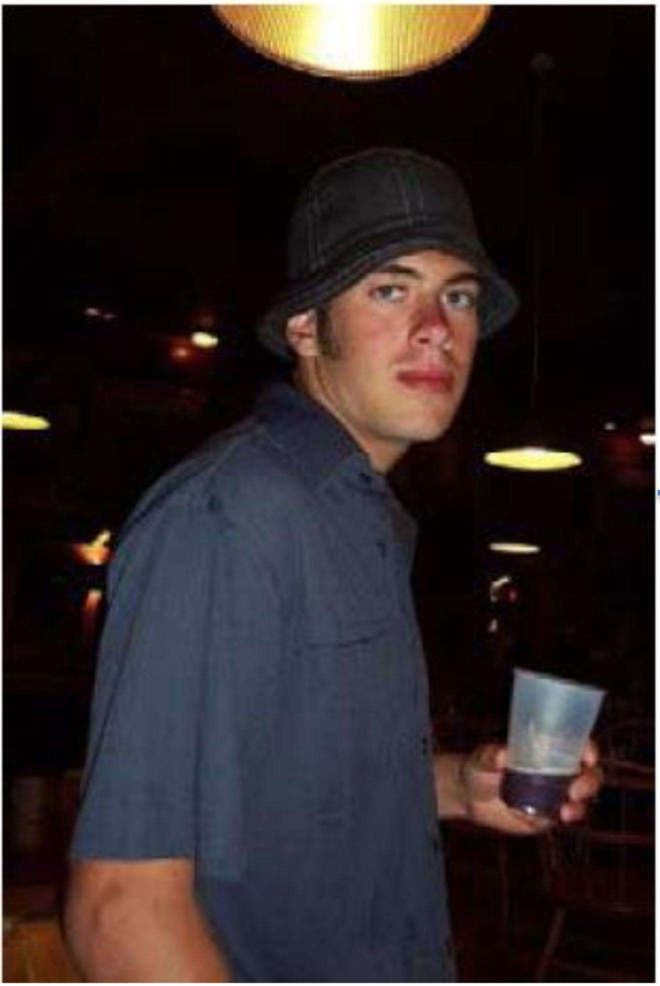



car

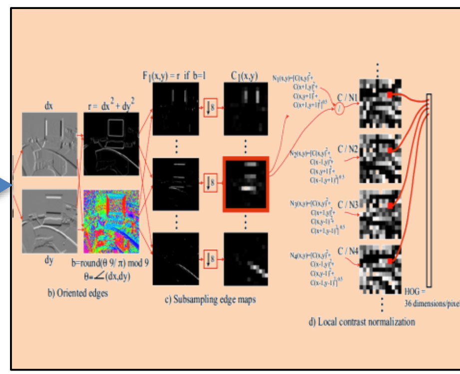
Who’s to blame?

- The data
- The features
- The student

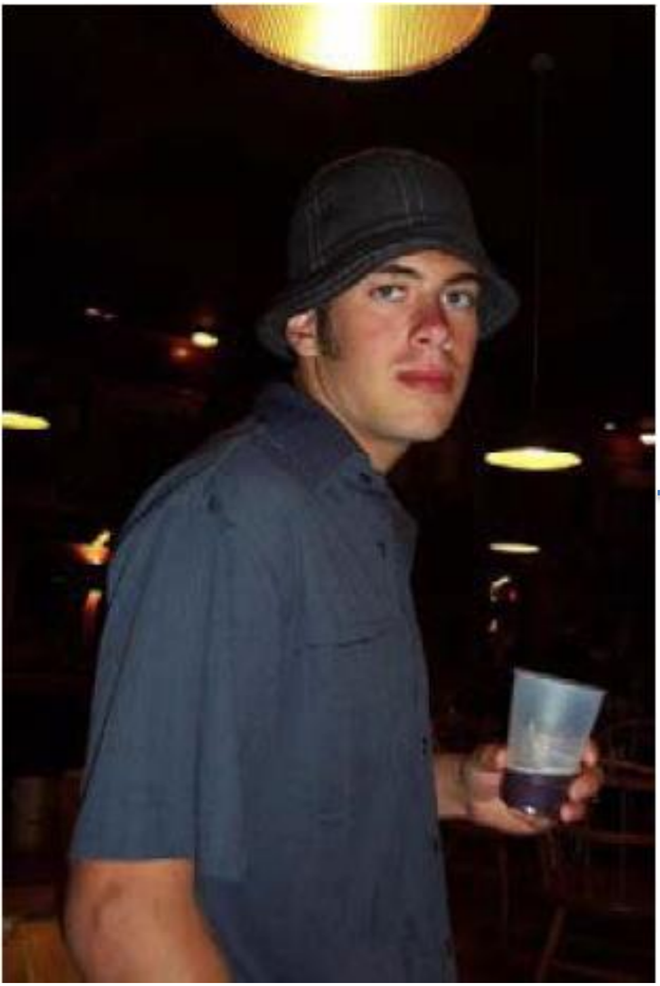
Features for object detection



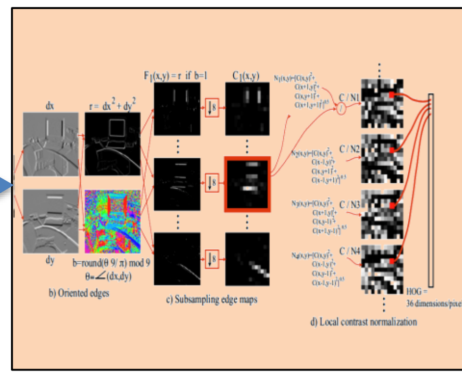
HOG



What does a detector sees?



HOG

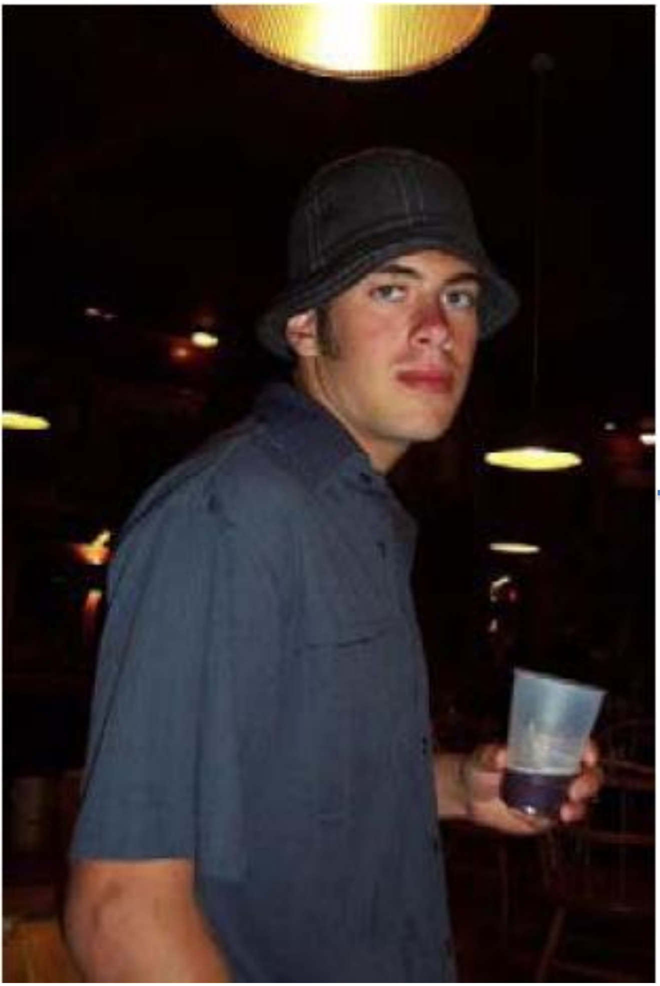


Can we visualize this output?

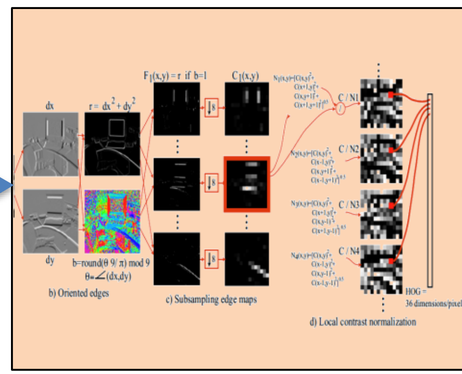
Carl Vondrick

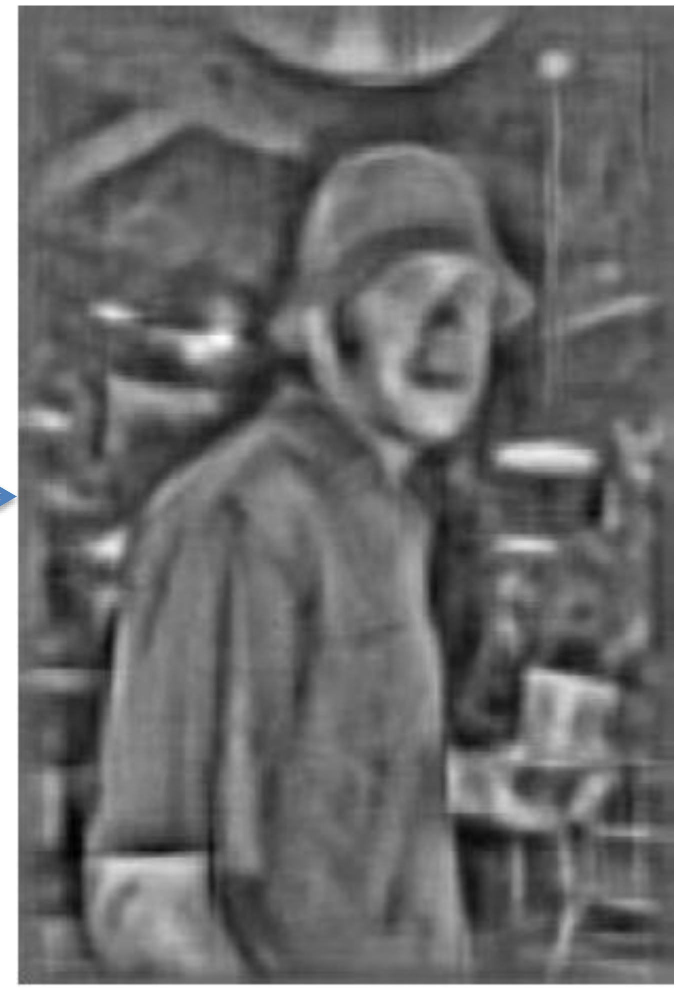
Aditya Khosla

What does a detector sees?



HOG

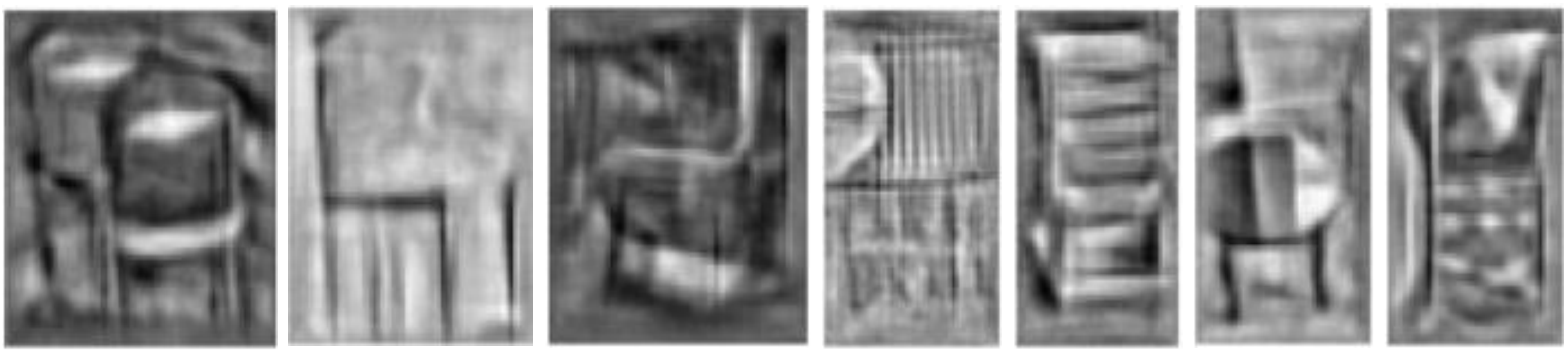




Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”, ICCV 2013

Person

Chair



Car

Can you tell which ones are not the object?

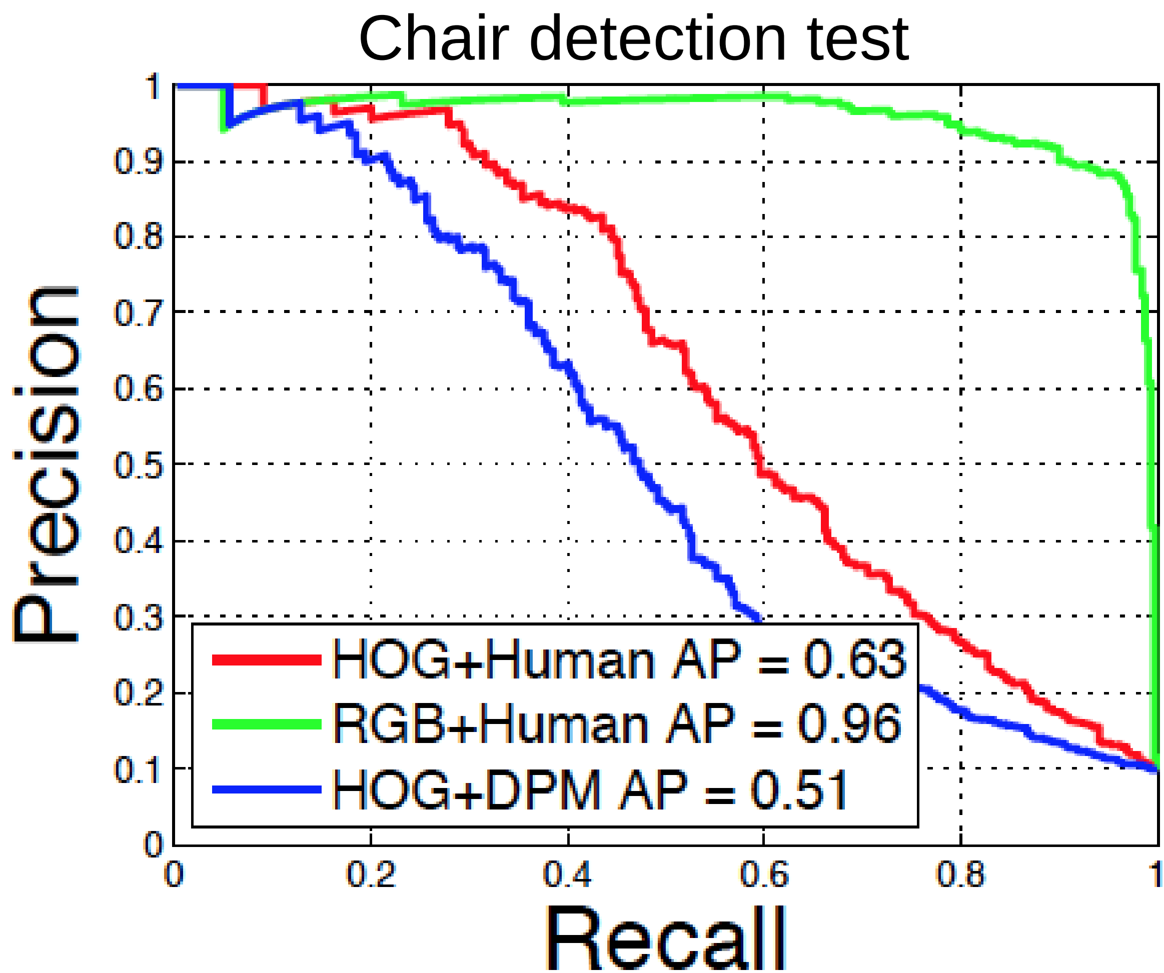
Person

Chair

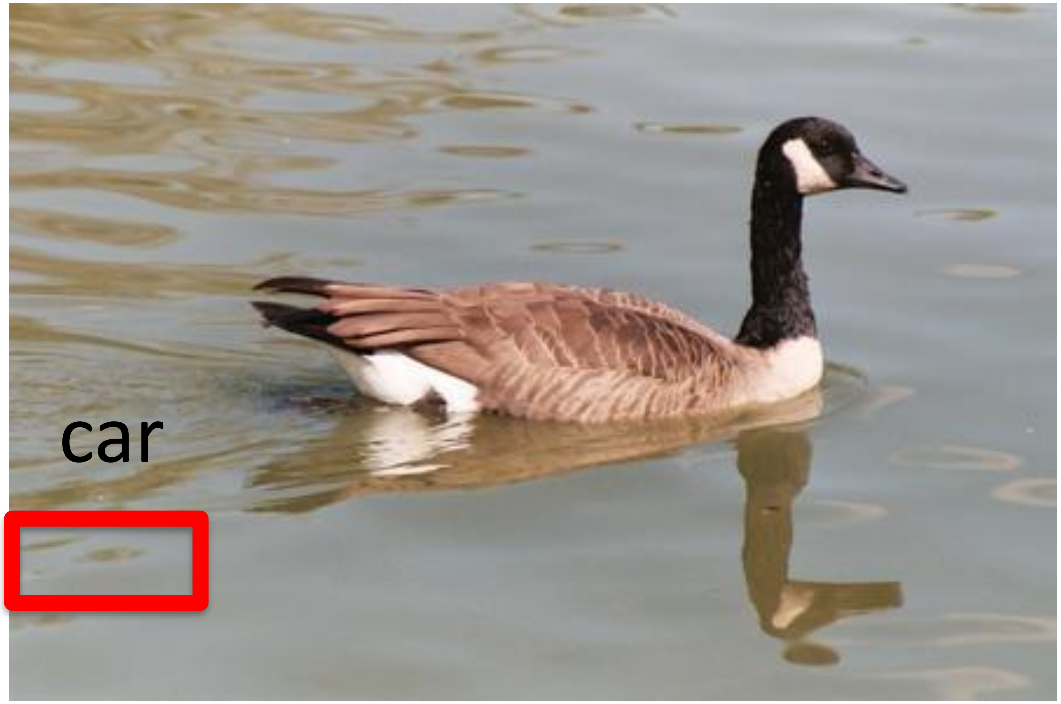
Car

Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”

HOG visualization predicts SVM performance

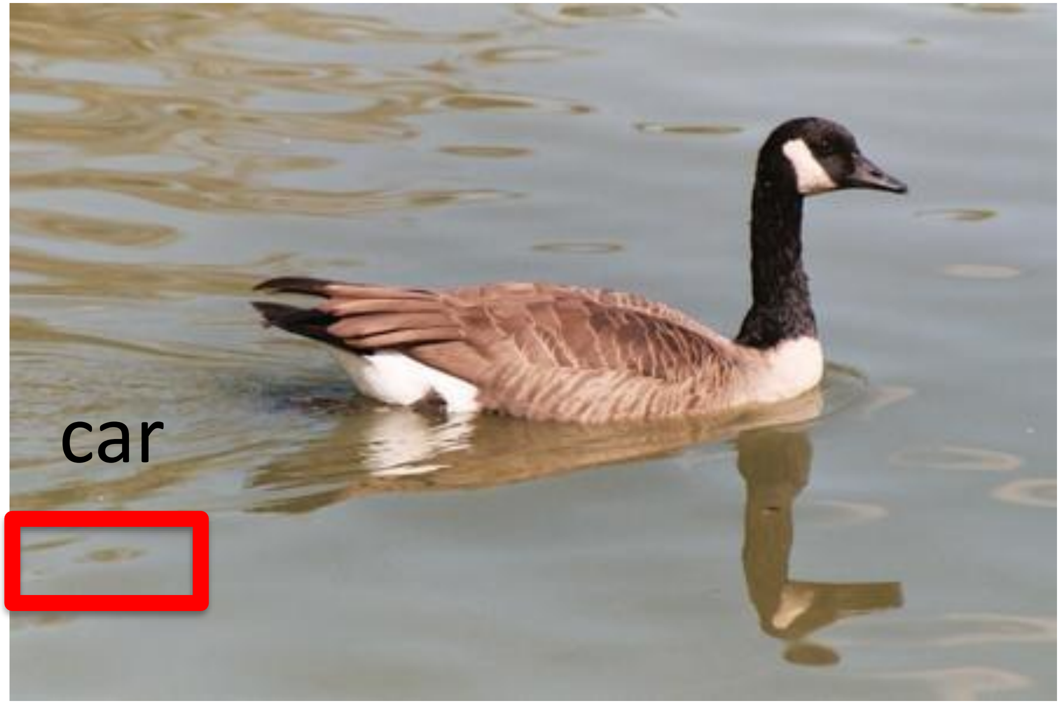


Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”



http://mit.edu/vondrick/ihog/

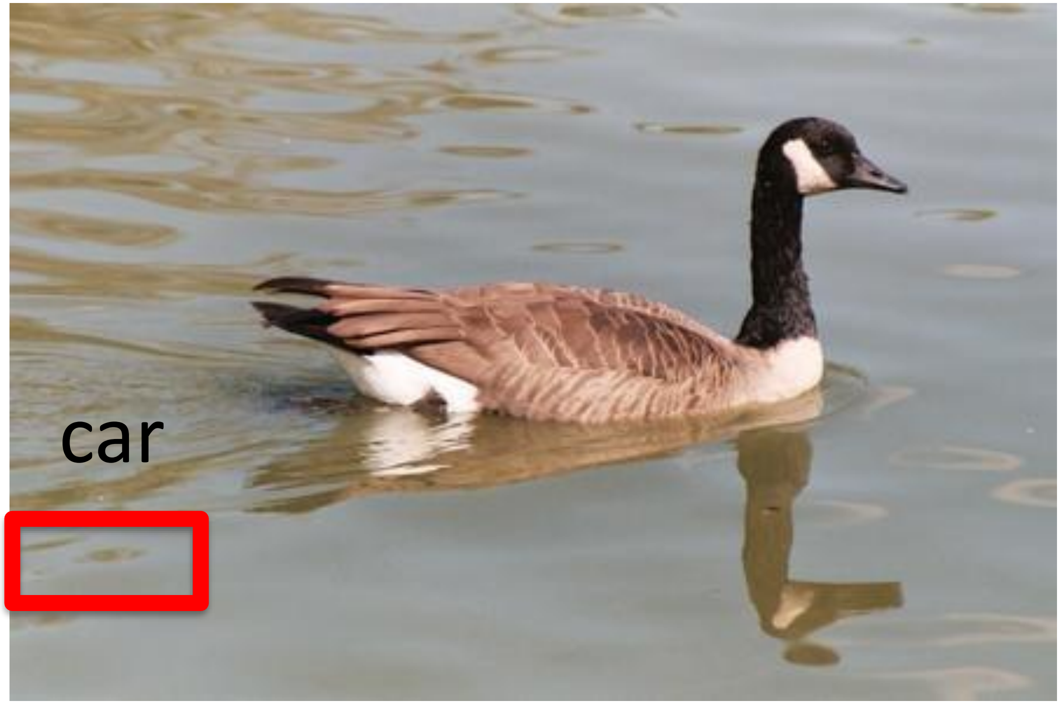
Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”



The image patch

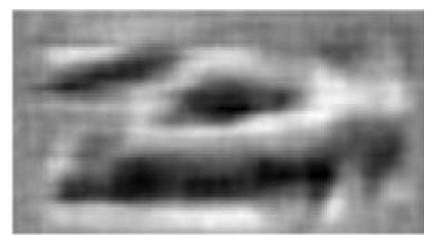
http://mit.edu/vondrick/ihog/

Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”



The image patch

What the detector sees

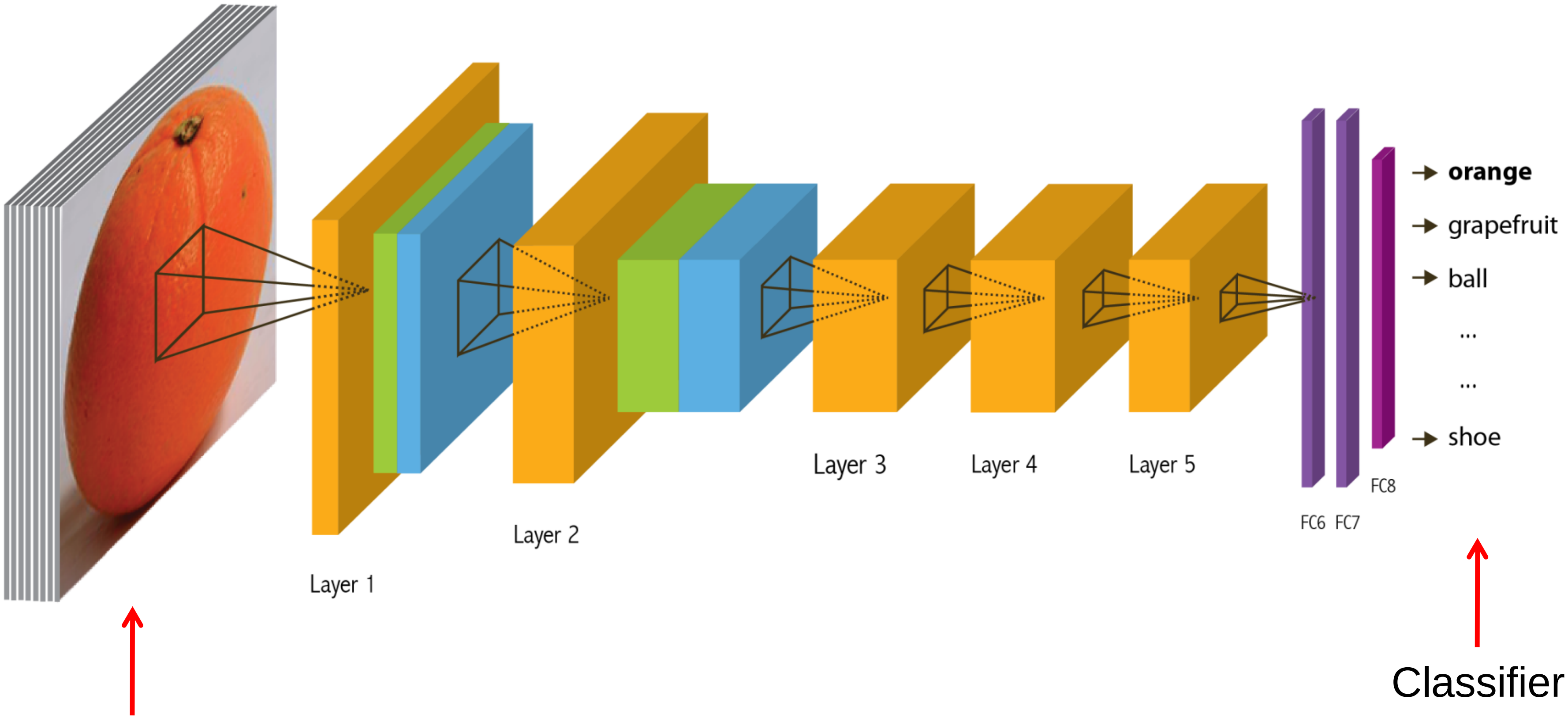


http://mit.edu/vondrick/ihog/

Vondrick, Khosla, Malisiewicz, Torralba. “Inverting and Visualizing Features for Object Detection.”

Deep architectures

Geoffrey Hinton, Yann LeCun

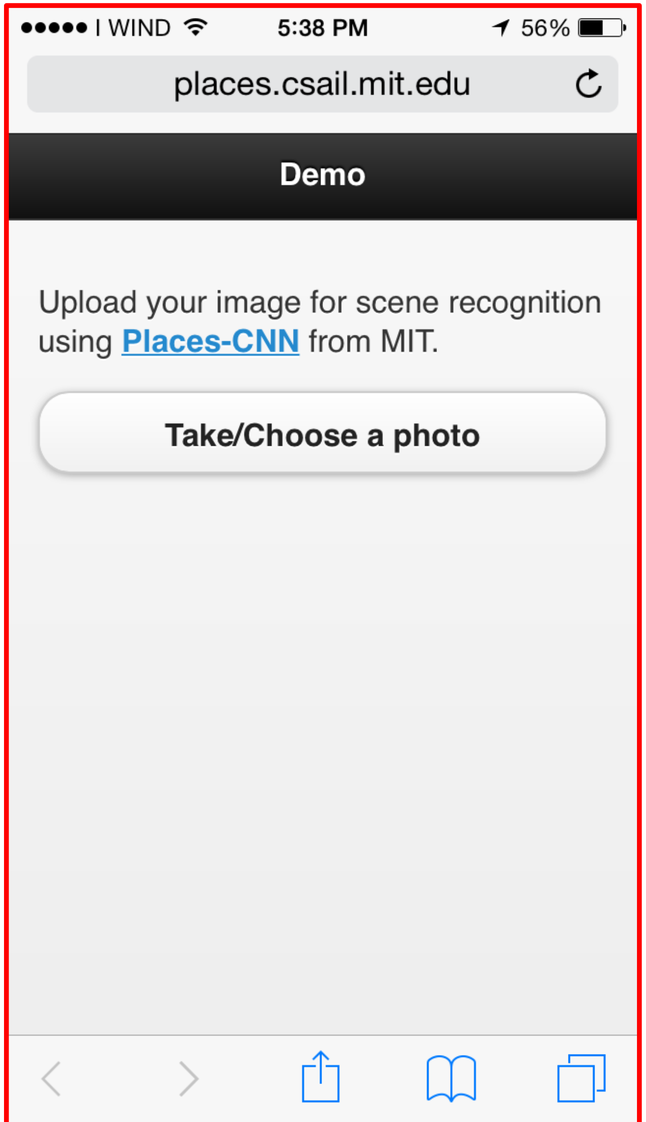


Input image

Classifier output

Scene recognition demo

http://places.csail.mit.edu/demo.html



Users report 78% correct results

Zhou, Lapedriza, Xiao, Oliva & Torralba (NIPS 2014)

http://places.csail.mit.edu/demo.html

places.csail.mit.edu

Take/Choose a photo

Predictions:

- **Type of environment:** outdoor
- **Semantic categories:** swimming_pool/outdoor:0.74, sandbar:0.11,

●●●○○ ROGERS LTE 6:24 PM 45%
places.csail.mit.edu

Predictions:

- **Type of environment:** indoor
- **Semantic categories:** airport_terminal:0.70,
- **SUN scene attributes:** enclosedarea, electricindoorlighting, nohorizon, man-made, congregating, cloth, glass, socializing, glossy, waitinginlinequeuing

●●●●● ROGERS 3G 6:47 PM 21%
places.csail.mit.edu

the image uploaded follow Creative Commons licenses.

Take/Choose a photo

Predictions:

- **Type of environment:** indoor
- **Semantic categories:** cockpit:0.08, parking_lot:0.06, playground:0.05,
- **SUN scene attributes:** nohorizon, enclosedarea, cloth, man-made, electricindoorlighting, working, stressful, dry, competing, waitinginlinequeuing

http://places.csail.mit.edu/demo.html

Predictions:

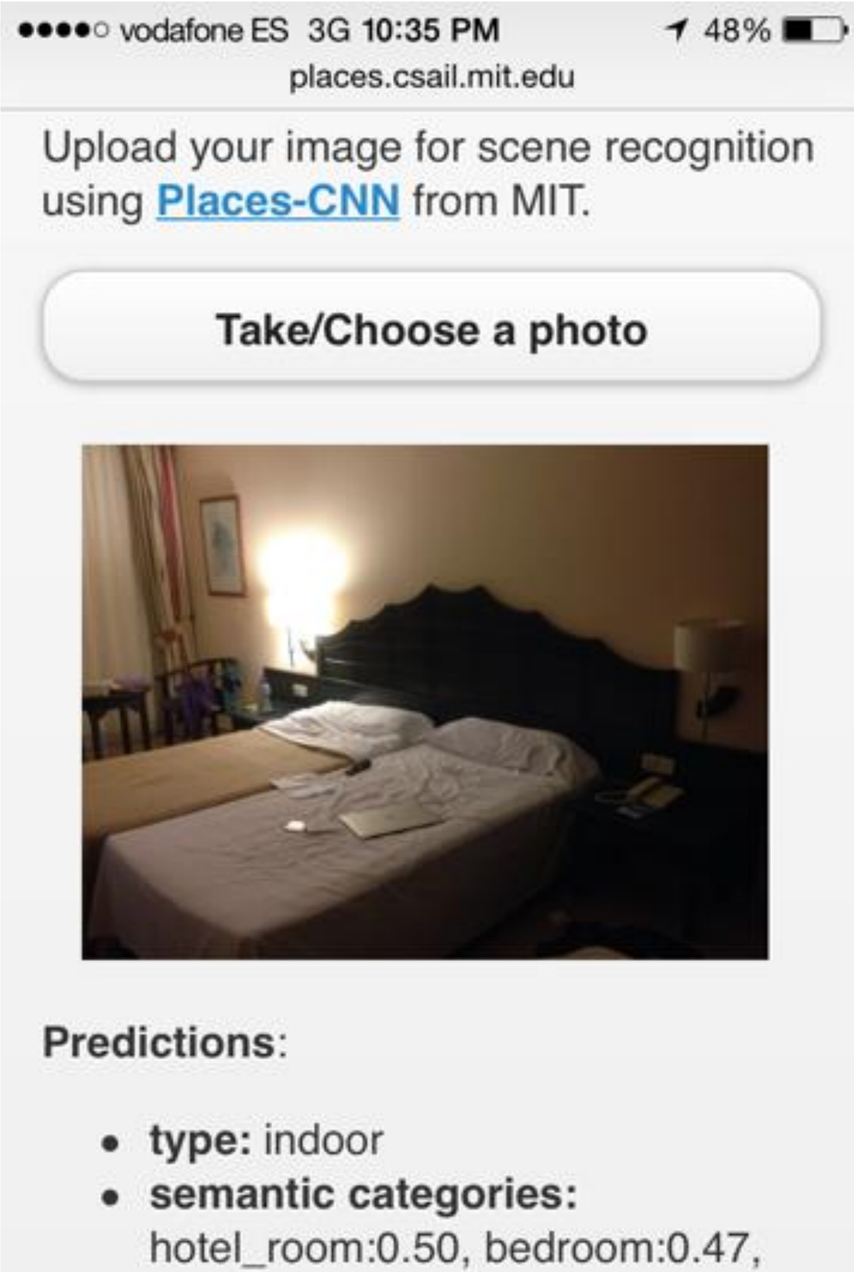
- **Type of environment:** indoor
- **Semantic categories:** auditorium:0.61, conference_center:0.34,

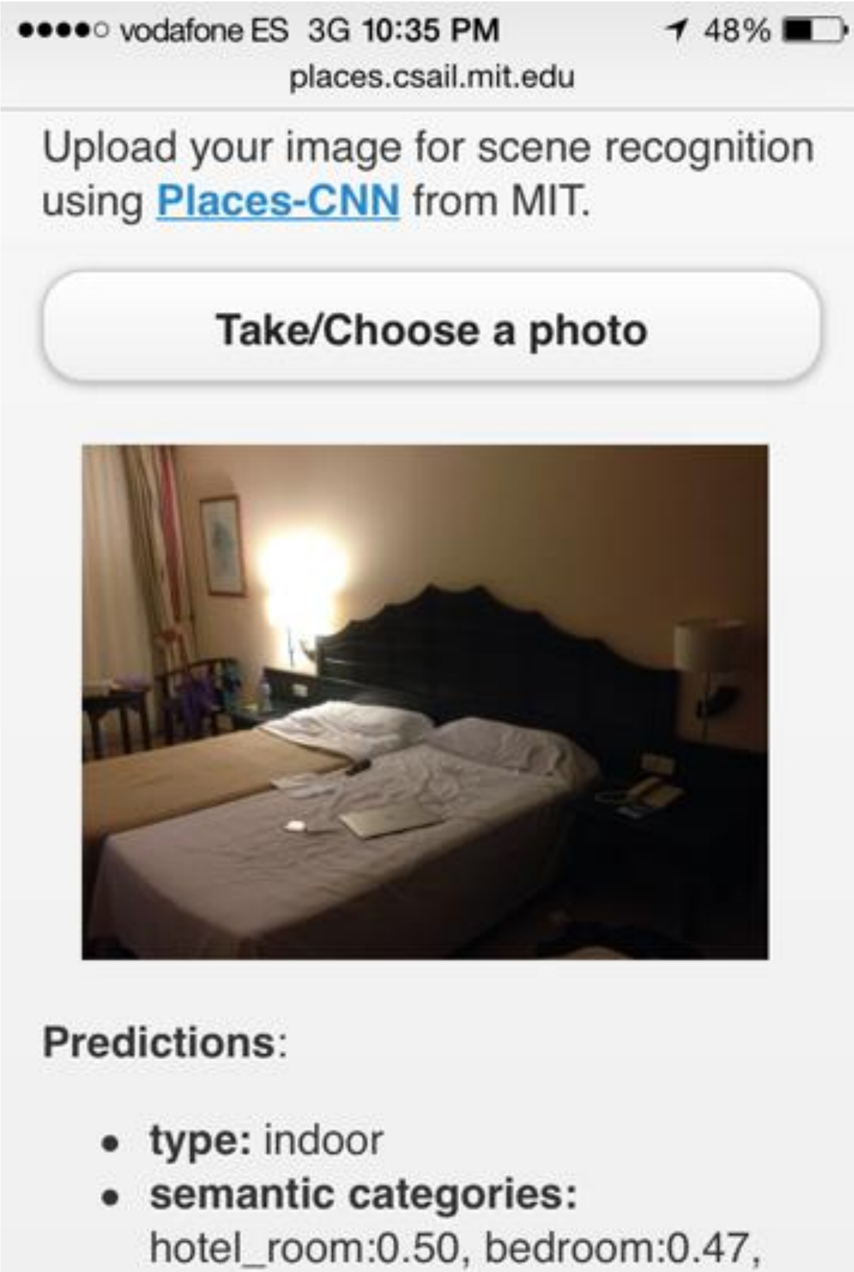
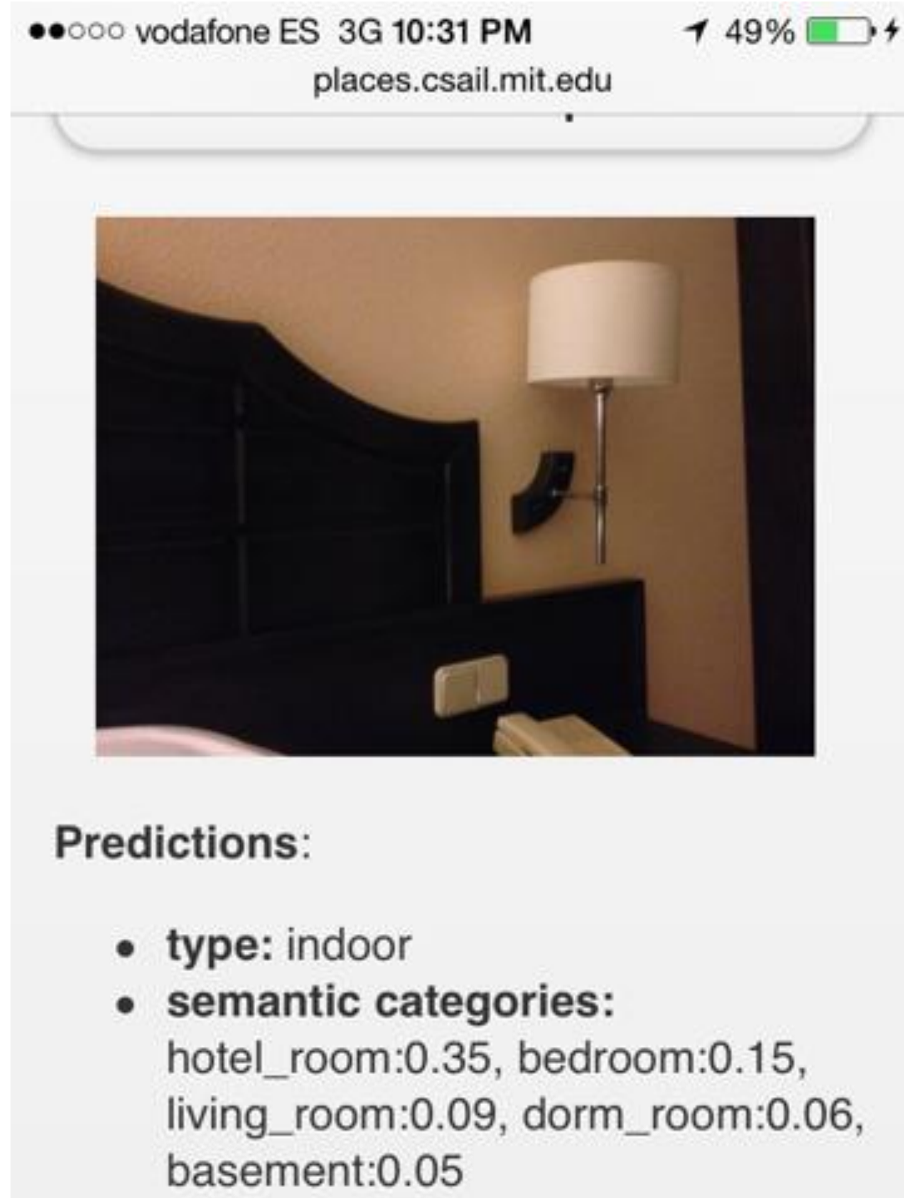
●○○○○ ROGERS LTE 9:17 AM 49%

places.csail.mit.edu

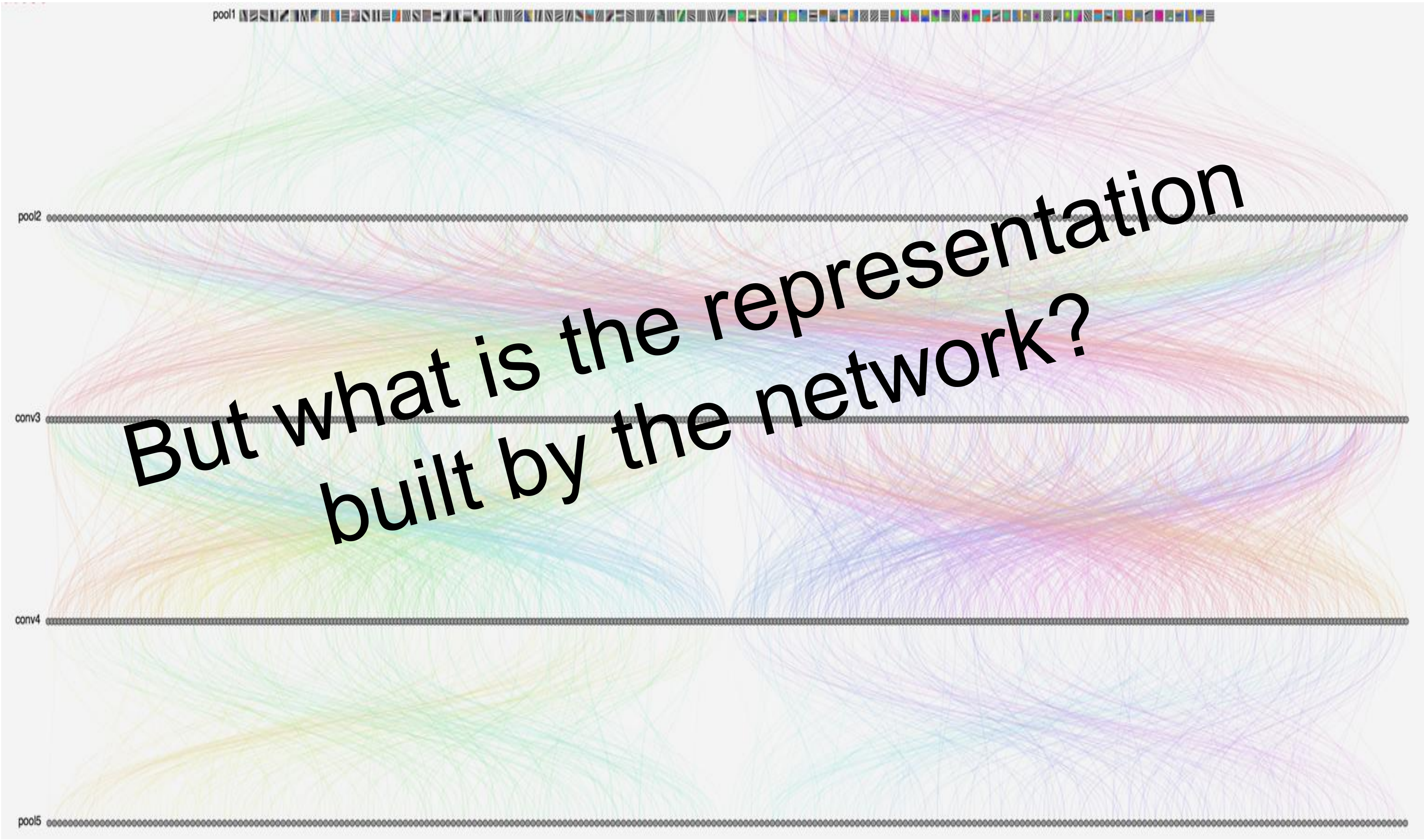
Predictions:

- **Type of environment:** indoor
- **Semantic categories:** bar:0.25, auditorium:0.20, restaurant_kitchen:0.07, coffee_shop:0.05,
- **SUN scene attributes:** enclosedarea, nohorizon, man-made, electricindoorlighting, wood(notpartofatree), working, matte, glass, cloth, conductingbusiness


●●●●○ vodafone ES 3G 10:35 PM 48%
places.csail.mit.edu
Upload your image for scene recognition using Places-CNN from MIT.
Take/Choose a photo
Predictions:
type: indoor
semantic categories:
hotel_room:0.50, bedroom:0.47,


●●●●○ vodafone ES 3G 10:35 PM 48%
places.csail.mit.edu
Upload your image for scene recognition using Places-CNN from MIT.
Take/Choose a photo
Predictions:
type: indoor
semantic categories:
hotel_room:0.50, bedroom:0.47,

●●○○○ vodafone ES 3G 10:31 PM 49%
places.csail.mit.edu
Predictions:
type: indoor
semantic categories:
hotel_room:0.35, bedroom:0.15, living_room:0.09, dorm_room:0.06, basement:0.05

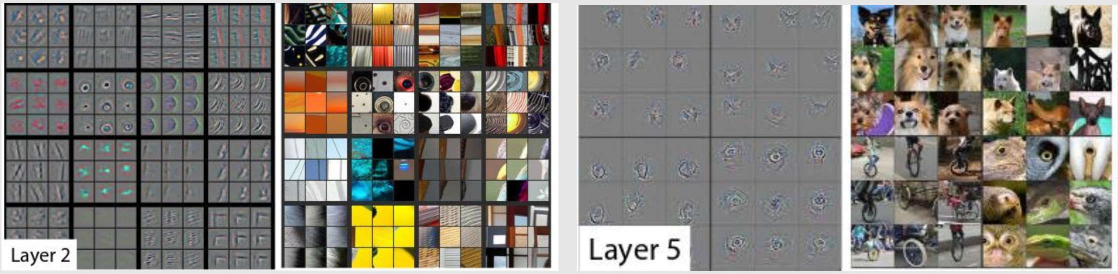
Why is working so well?



But what is the representation built by the network?

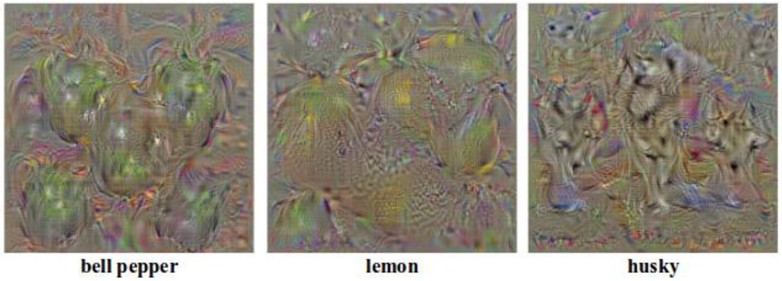
Visualizing the internal representation

Deconvolution



Zeiler & Fergus, Visualizing and Understanding Convolutional Networks, ECCV 2014.

Backpropagation



Simonyan et al. Visualizing image classification models and saliency maps. ICLRW, 2014.

Strong activations

Girshick, et al, Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014.

Visualizing and Understanding Convolutional Networks

Matthew D. Zeiler and Rob Fergus

Dept. of Computer Science,
New York University, USA
{zeiler,fergus}@cs.nyu.edu

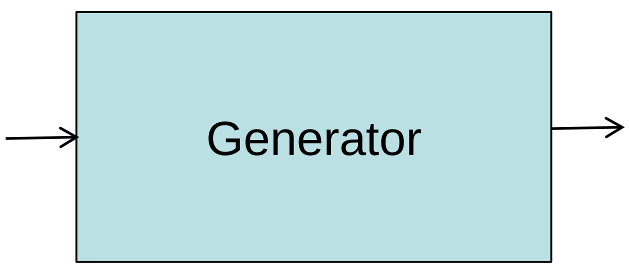
Zeiler & Fergus, Visualizing and Understanding Convolutional Networks, ECCV 2014.

Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley,
Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

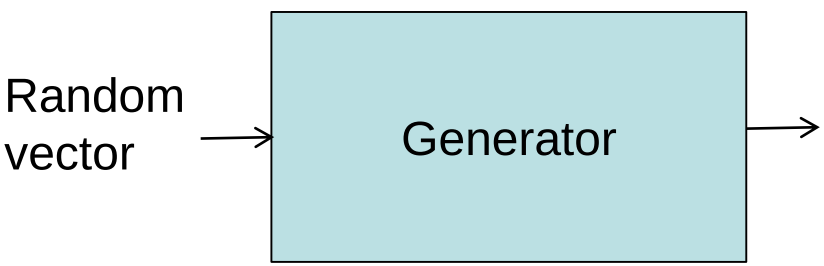
Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7



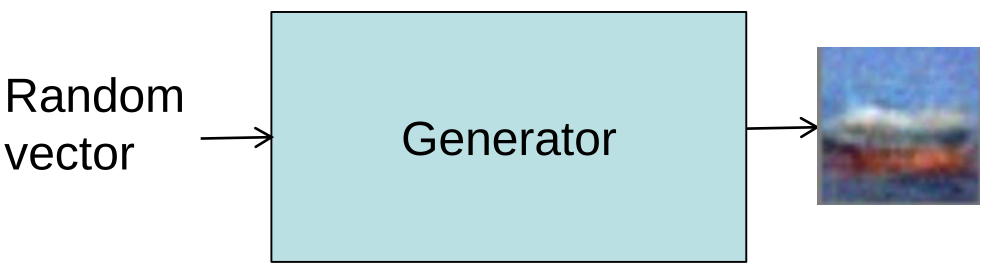
Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7



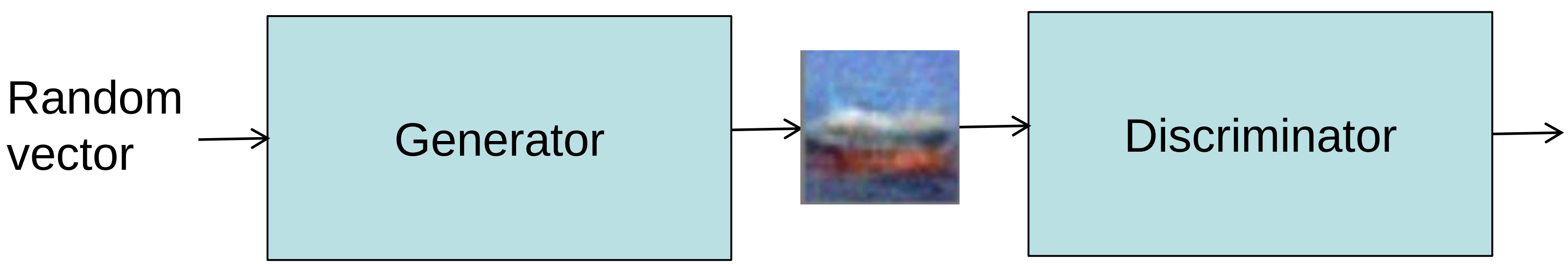
Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7



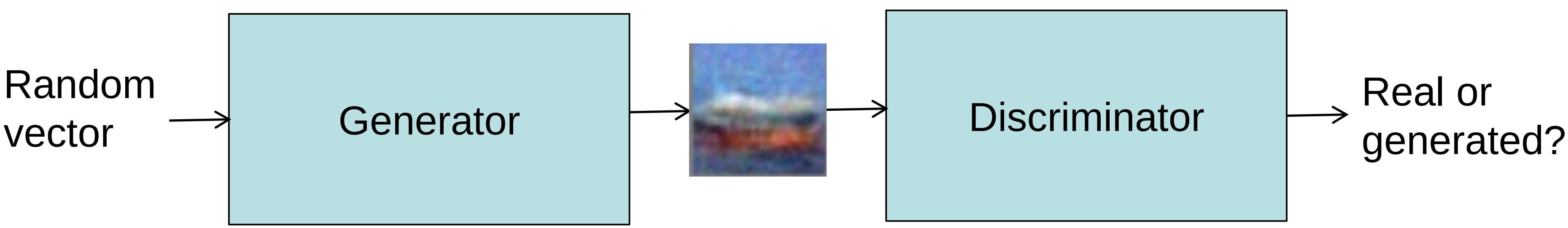
Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7



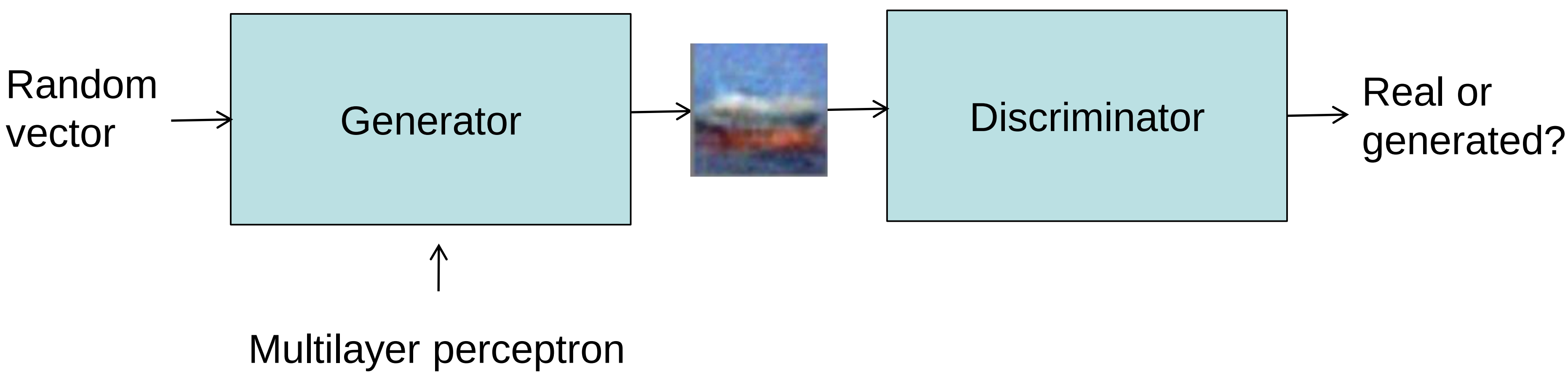
Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7



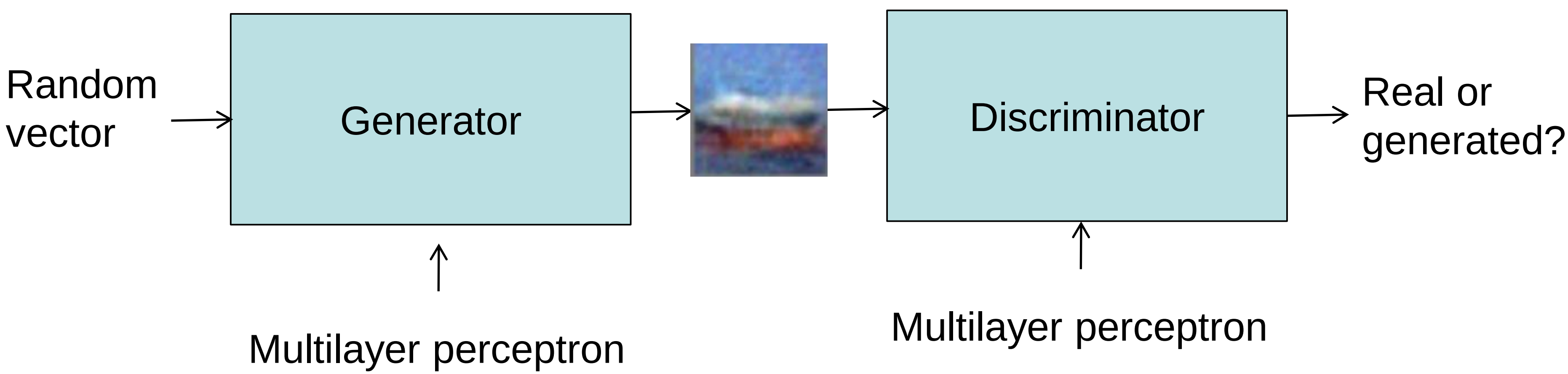
Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie[*], Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair[†], Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

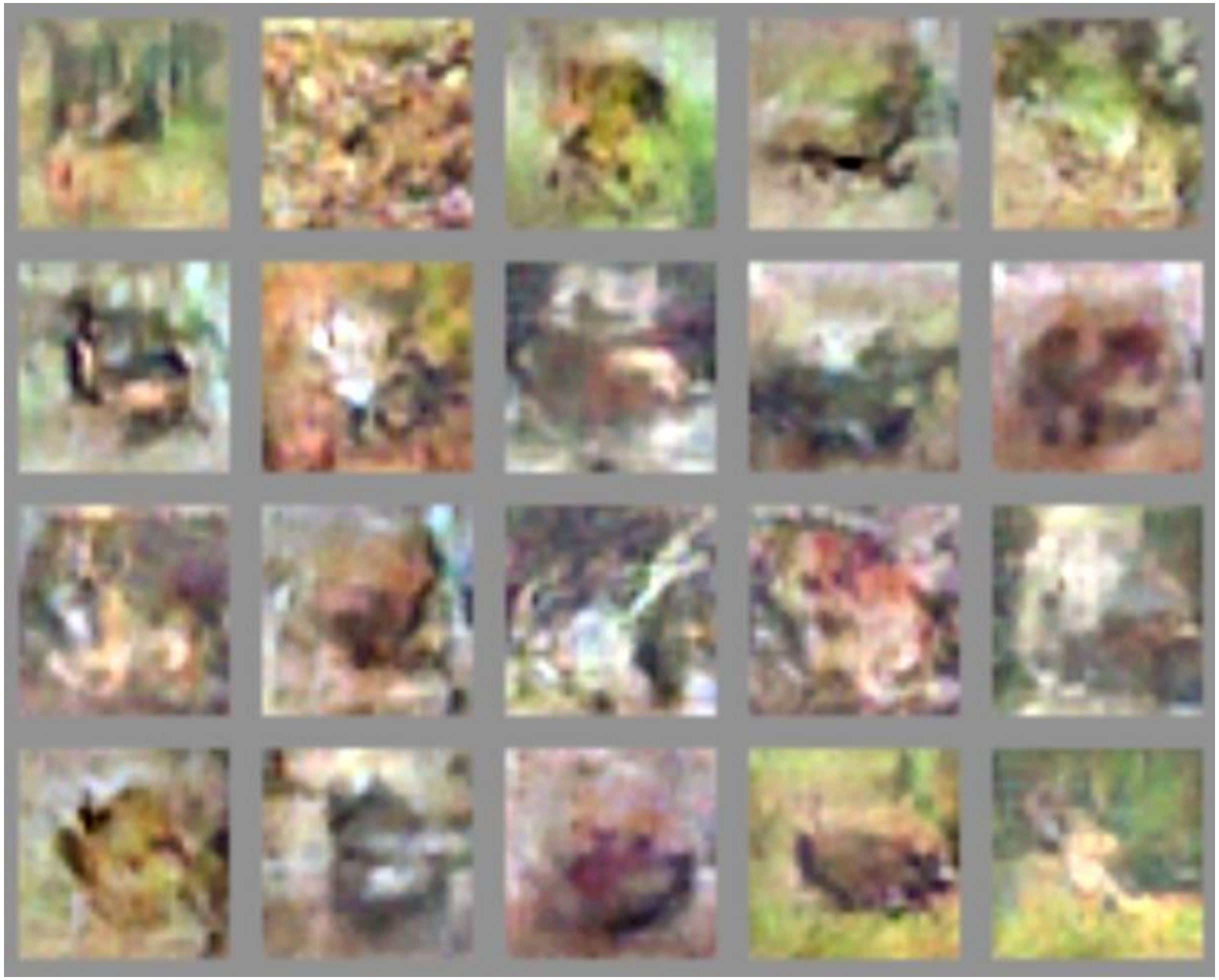


Generative Adversarial Nets

Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7



Generated images



Trained with CIFAR-10

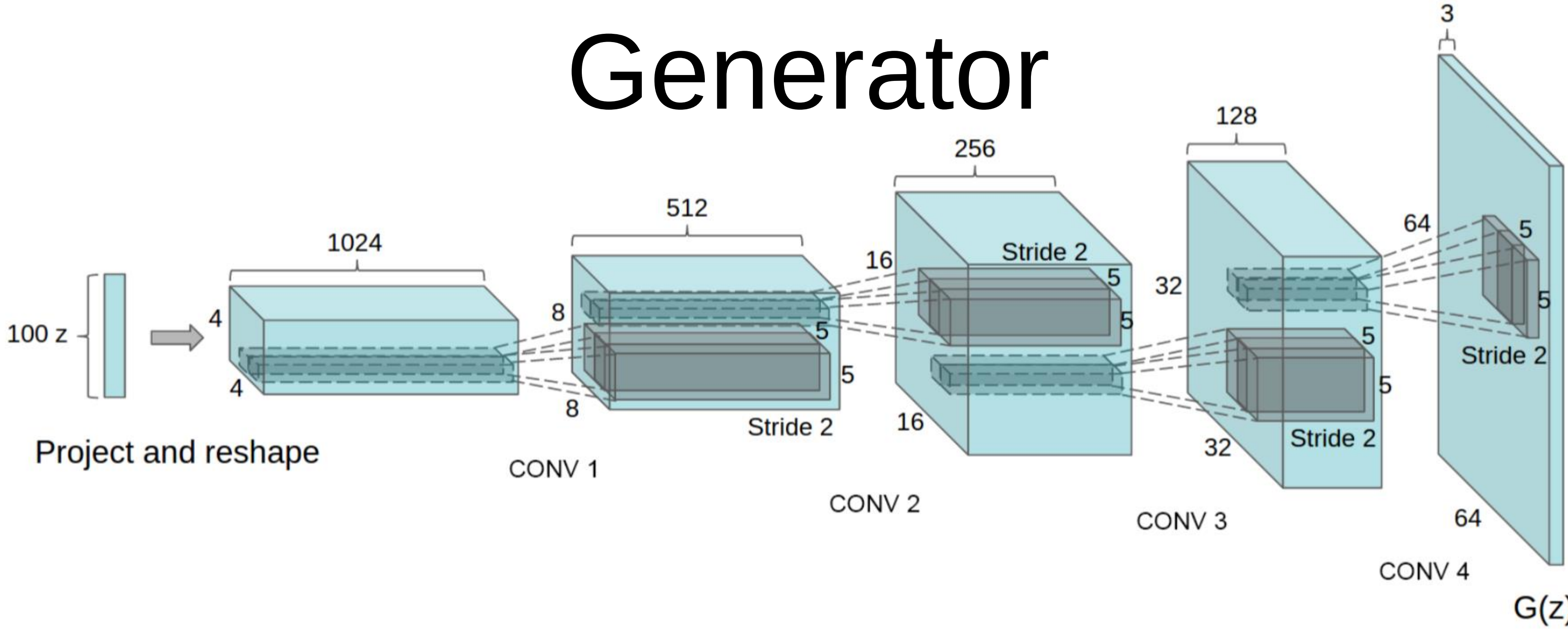
UNSUPERVISED REPRESENTATION LEARNING WITH DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS

Alec Radford & Luke Metz
indico Research
Boston, MA
{alec,luke}@indico.io

Soumith Chintala
Facebook AI Research
New York, NY
soumith@fb.com

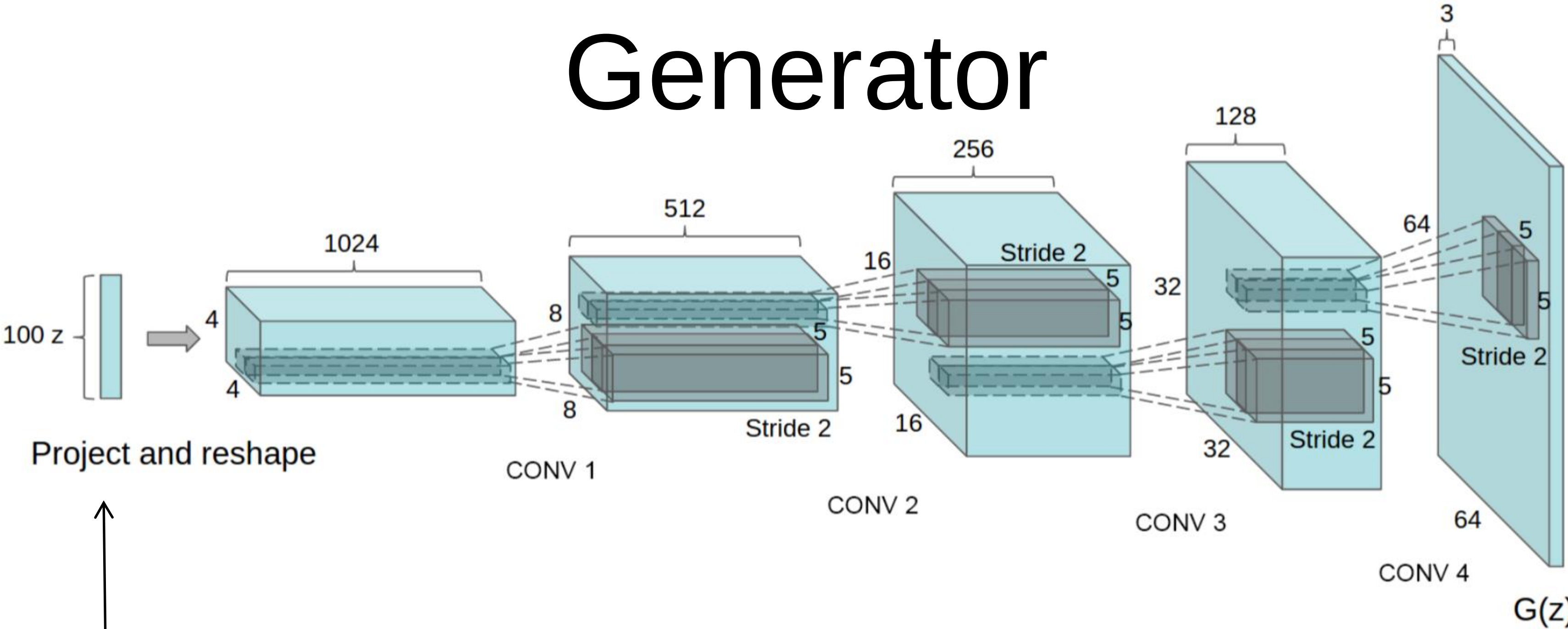
Introduced a form of ConvNet more stable under adversarial training than previous attempts.

Generator



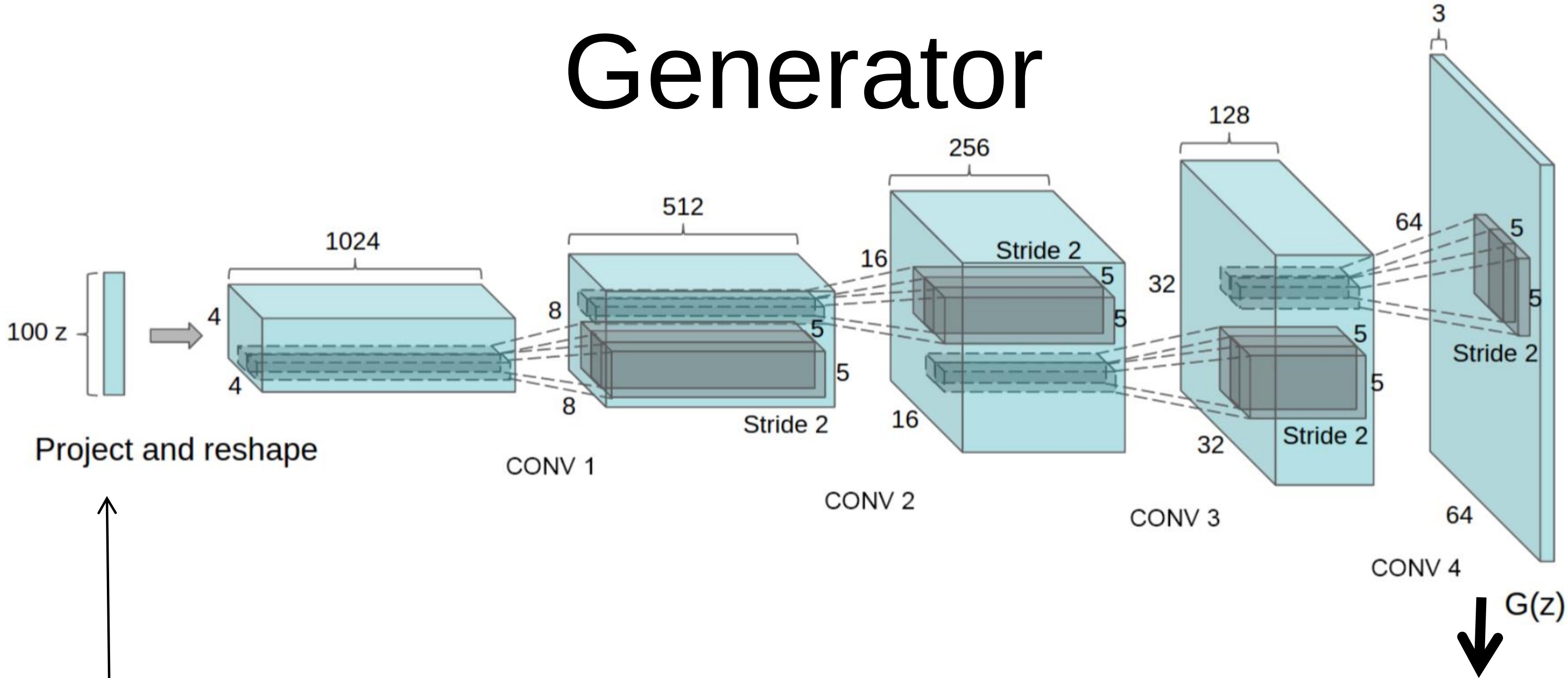
100 z
Project and reshape
1024
4
4
512
8
8
5
5
Stride 2
CONV 1
256
16
16
Stride 2
5
5
CONV 2
128
32
32
5
5
Stride 2
CONV 3
3
64
64
5
5
Stride 2
CONV 4
G(z)

Generator



Random uniform
vector (100 numbers)

Generator



Random uniform
vector (100 numbers)

Synthesizing the preferred inputs for neurons in neural networks via deep generator networks

Anh Nguyen
anguyen8@uwyo.edu

Alexey Dosovitskiy
dosovits@cs.uni-freiburg.de

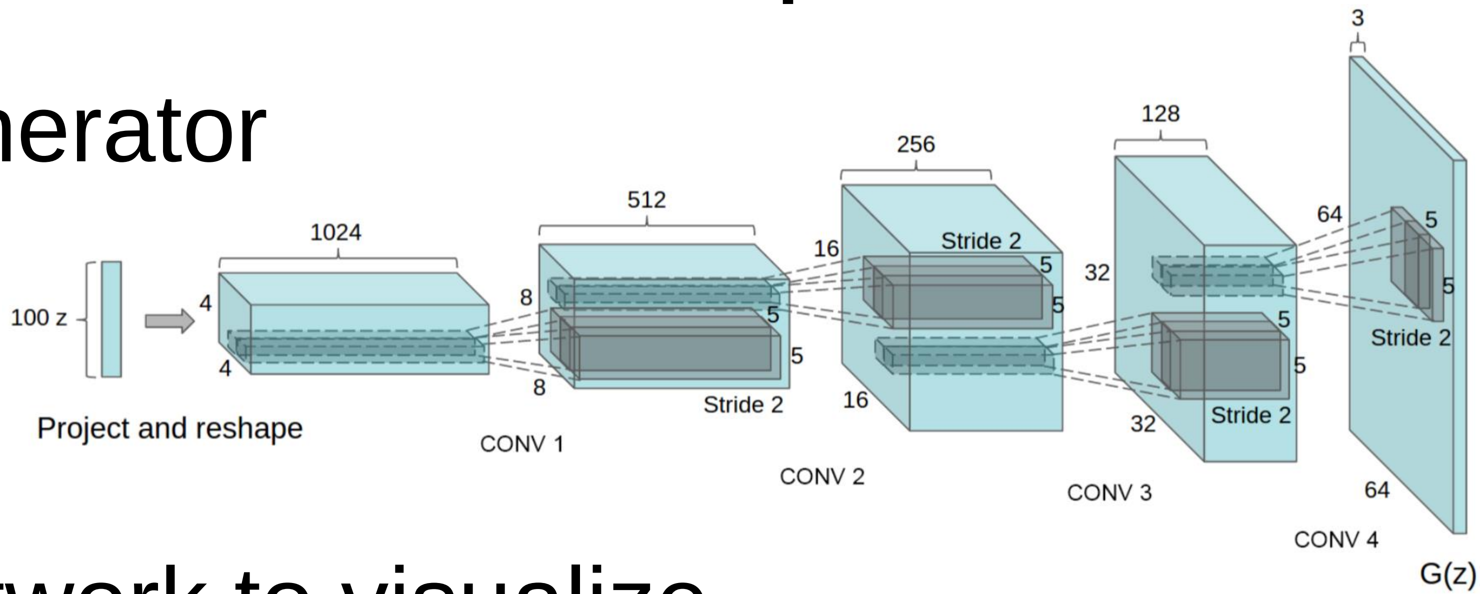
Jason Yosinski
jason@geometricintelligence.com

Thomas Brox
brox@cs.uni-freiburg.de

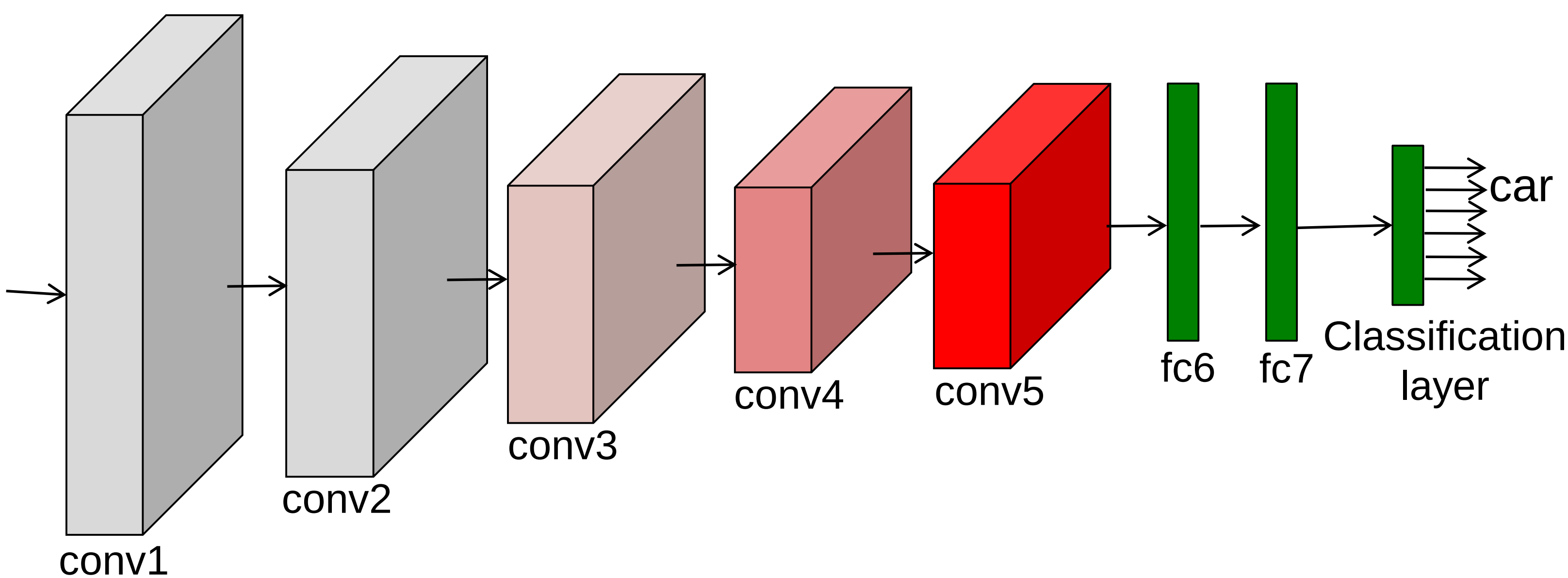
Jeff Clune
jeffclune@uwyo.edu

Two components

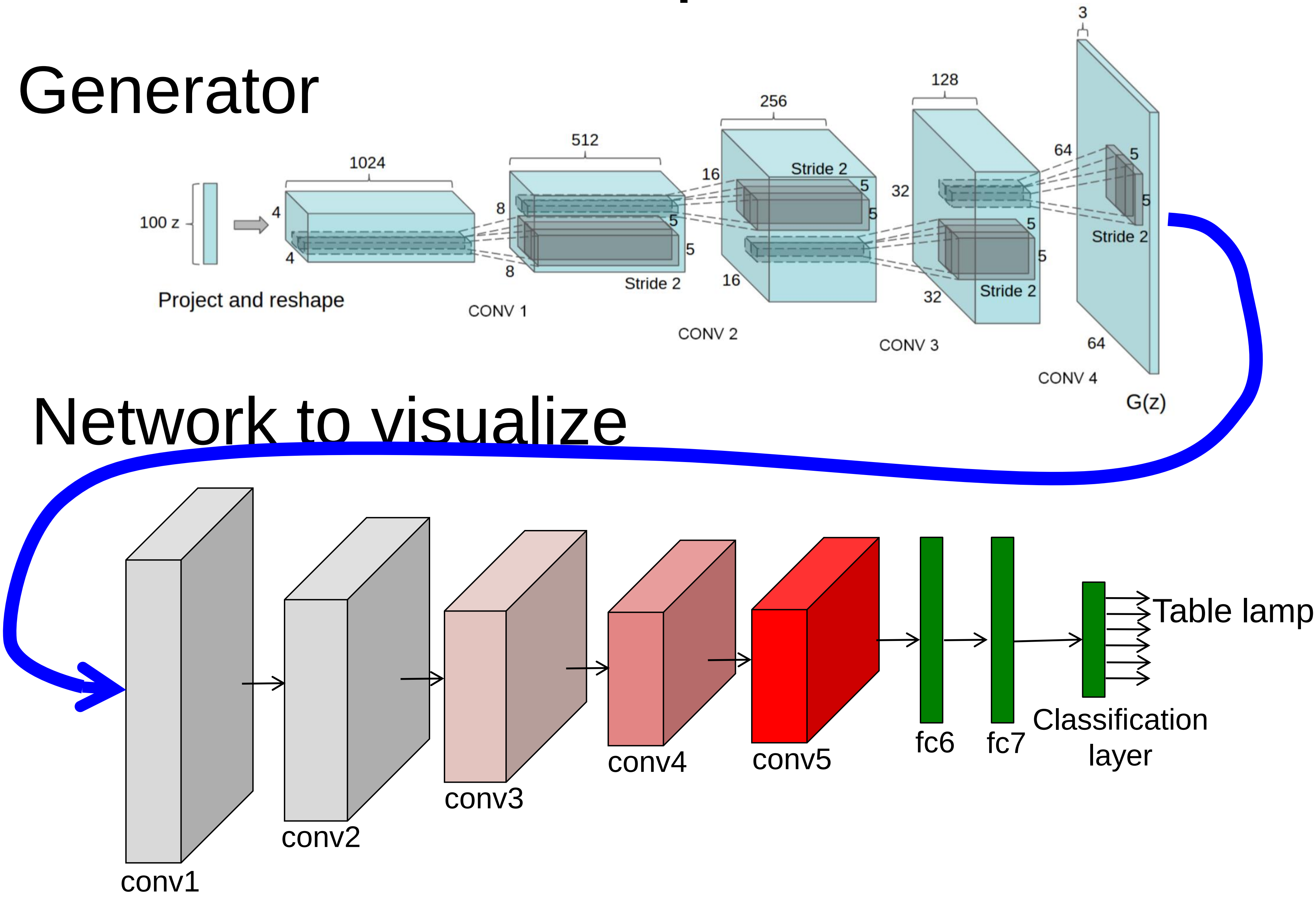
Generator



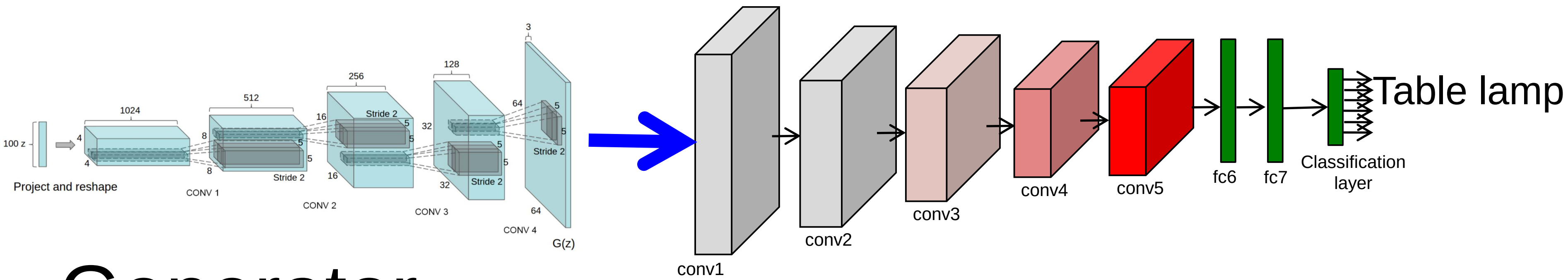
Network to visualize



Two components

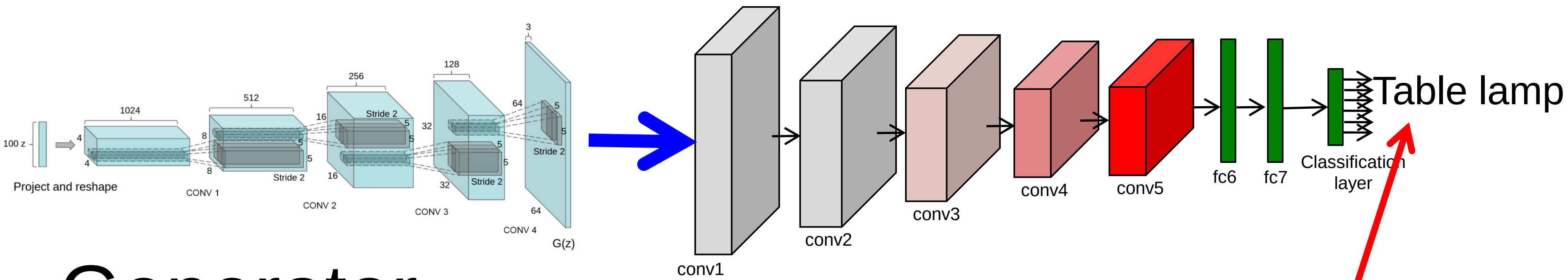


Two components



Generator

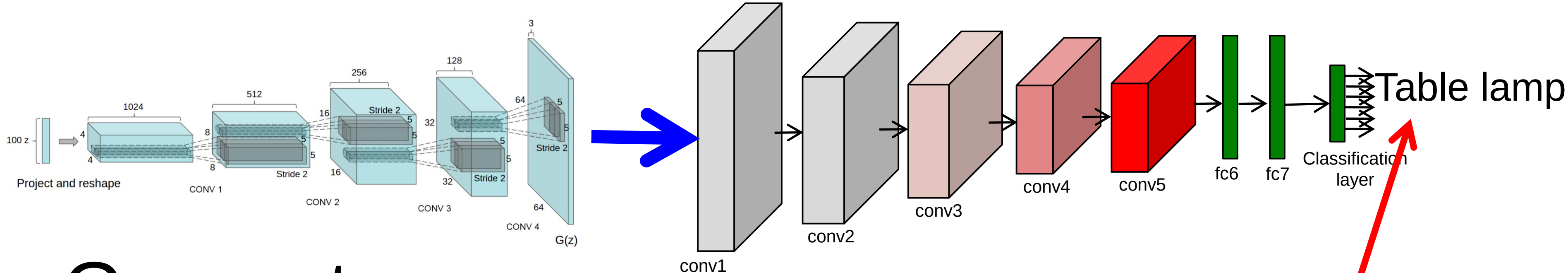
Two components



Generator

Unit to visualize

Two components



Generator

Unit to visualize

Synthesizing Images Preferred by CNN

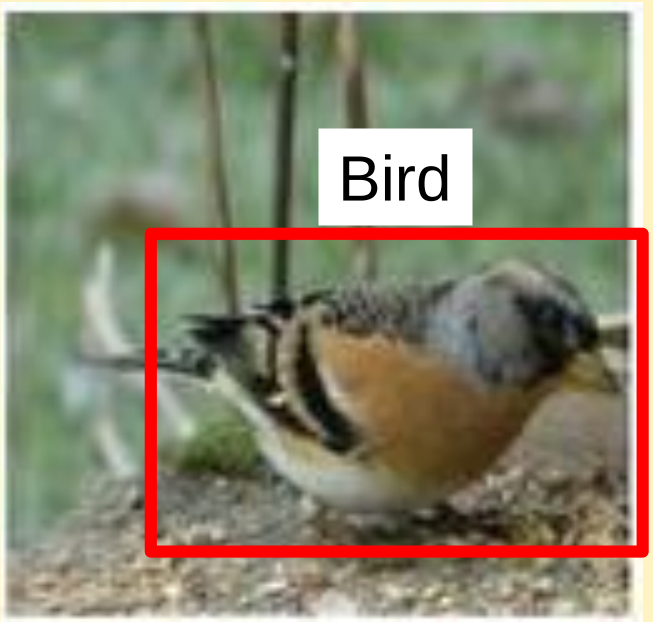
ImageNet-Alexnet-final units (class units)

Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J. (2016). "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks.". arXiv:1605.09304.

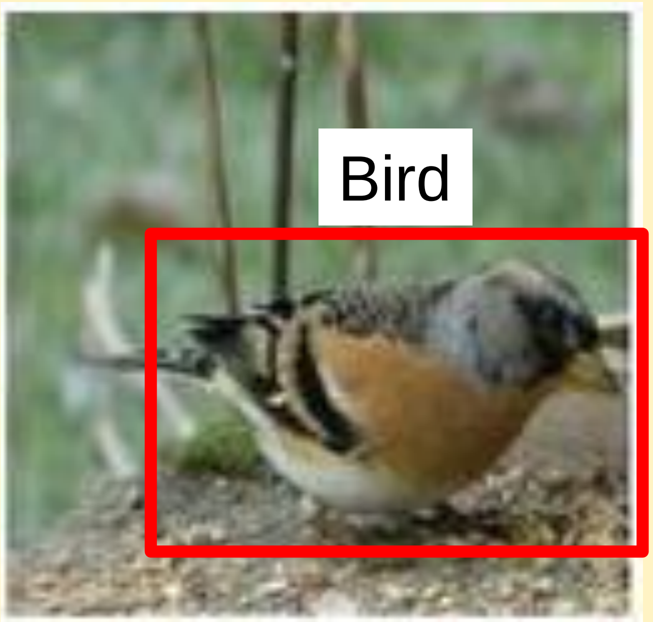
Object detection vs. Scene recognition

Object detection vs. Scene recognition

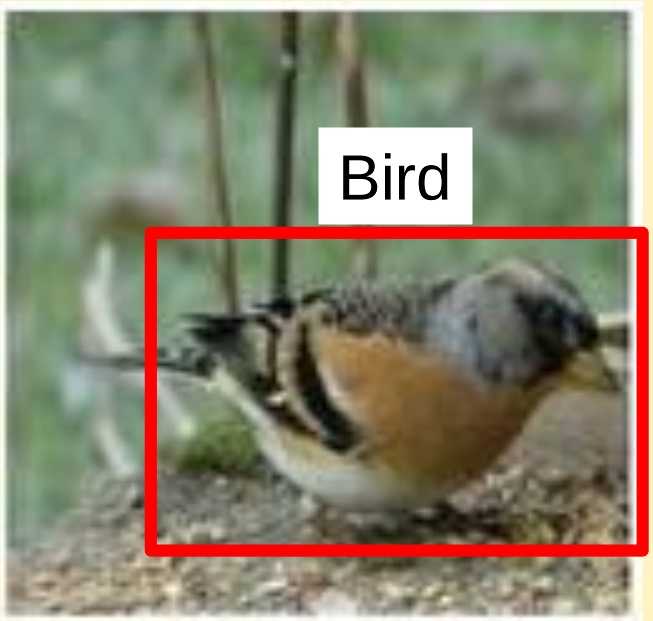
Object detection vs. Scene recognition



Object detection vs. Scene recognition

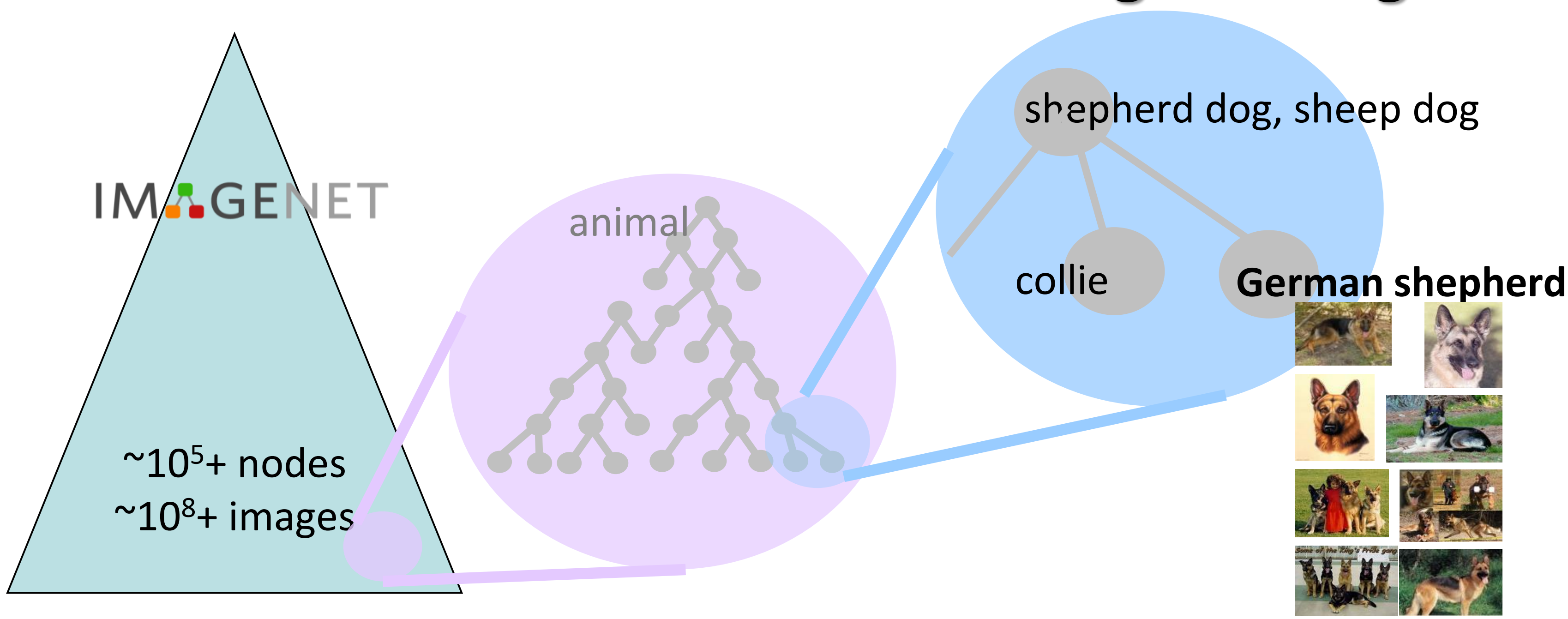


Object detection vs. Scene recognition



IMAGENET

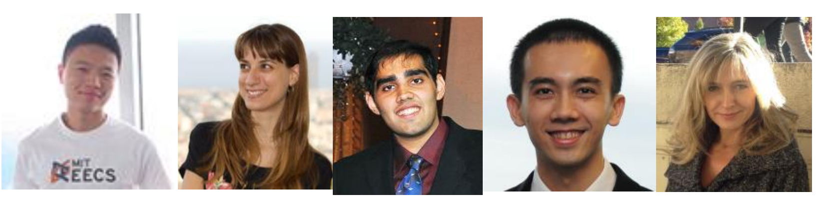
- An ontology of images based on WordNet
- ImageNet currently has
  - 13,000+ categories of visual concepts
  - 10 million human-cleaned images (~700im/categ)
  - 1/3+ is released online @ **www.image-net.org**



Deng, Dong, Socher, Li & Fei-Fei, CVPR 2009

places

places.csail.mit.edu



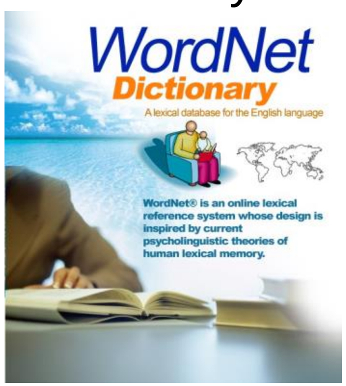
Zhou Lapedriz Khosla Xiao Oliva

Zhou, Lapedriza, Xiao, Oliva & Torralba (NIPS 2014)

places

places.csail.mit.edu

1. We take all scene words from a dictionary



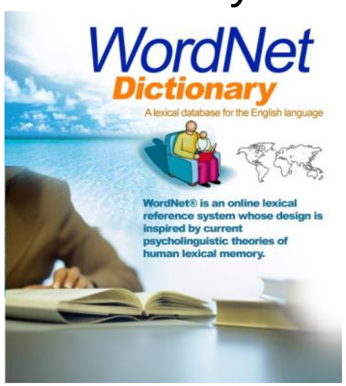
Zhou Lapedriz Khosla Xiao Oliva

Zhou, Lapedriza, Xiao, Oliva & Torralba (NIPS 2014)

places

places.csail.mit.edu

1. We take all scene words from a dictionary



2. We download images and clean the categories

Google Image Search

altavista

flickr


Zhou Lapedriz Khosla Xiao Oliva

Zhou, Lapedriza, Xiao, Oliva & Torralba (NIPS 2014)

Two large databases, two tasks

IMAGENET

brambling

terrier

Places Database

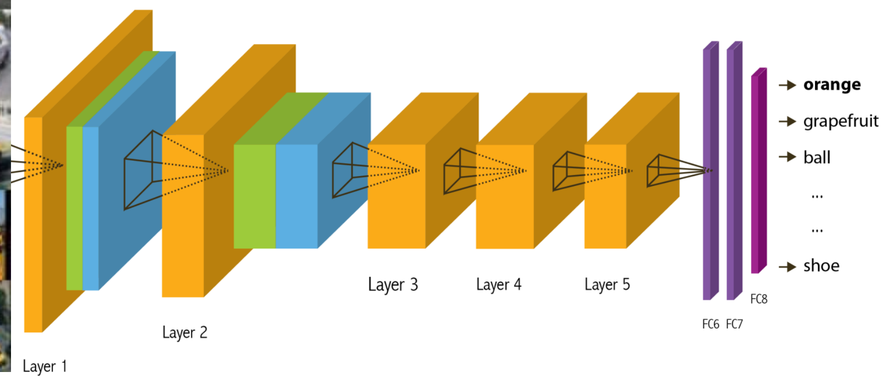
bedroom

mountain

ImageNet CNN and Places CNN

ImageNet CNN and Places CNN

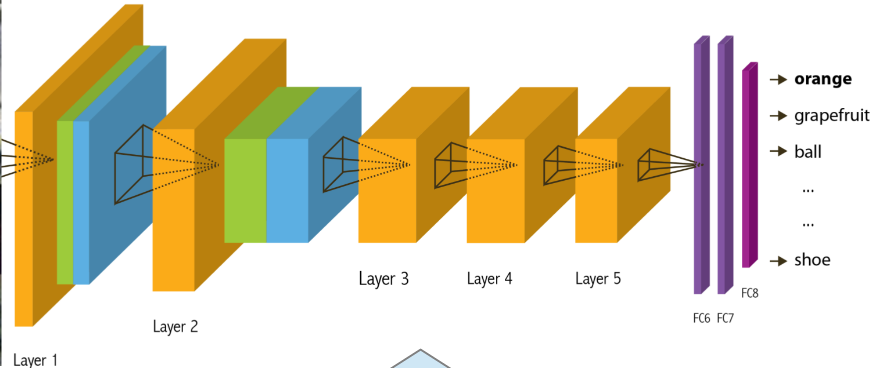
ImageNet CNN for Object Classification



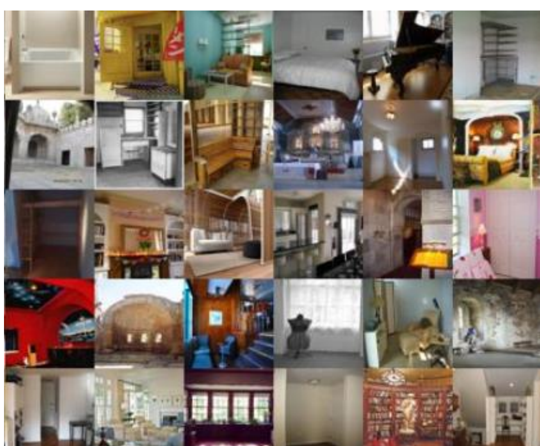
ImageNet CNN and Places CNN

IMAGENET

ImageNet CNN for Object Classification

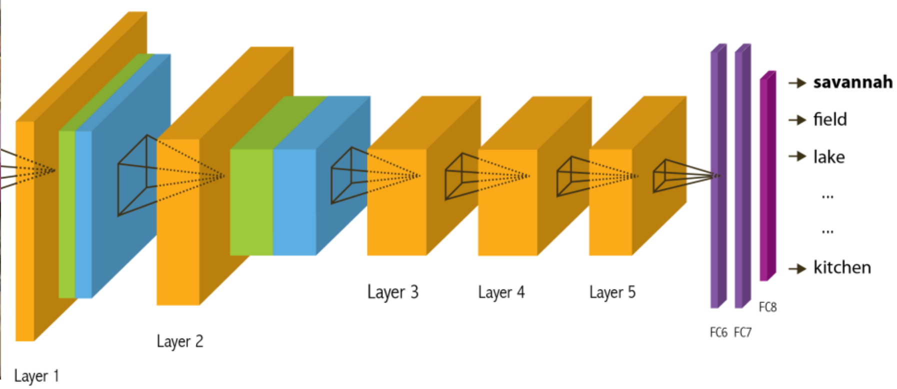


Same architecture: AlexNet



Places

Places CNN for Scene Classification



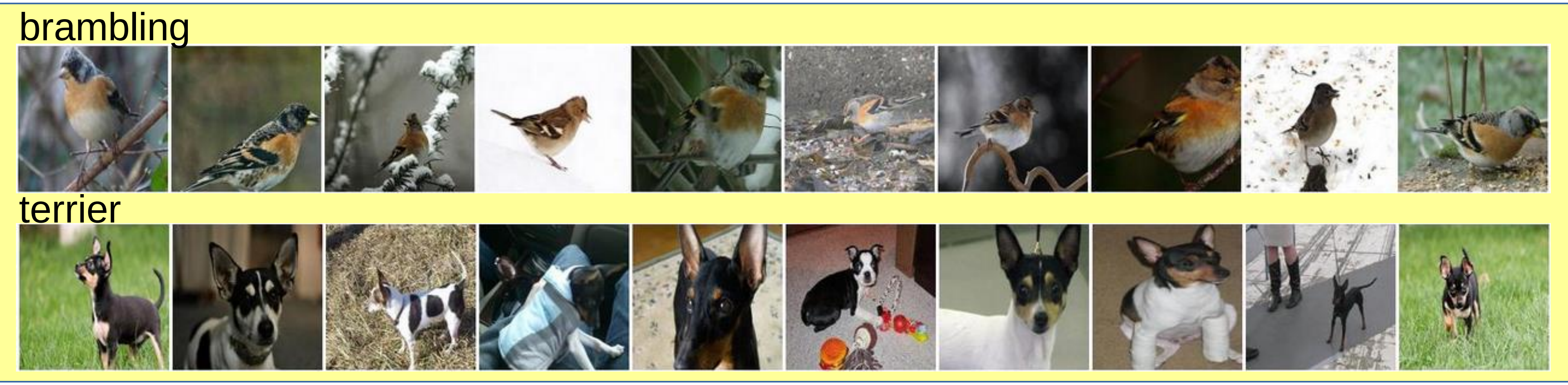
Possible internal representations

IMAGENET
[Deng et al. CVPR 2009]

PLACES

Learning to Recognize Objects

IMAGENET



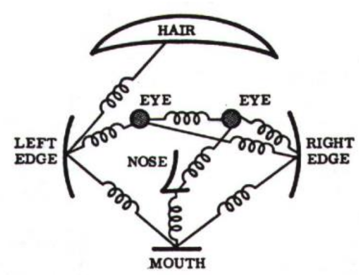
Learning to Recognize Objects

brambling

terrier

Possible internal representations:

- Object parts
- Textures
- Attributes



Learning to Recognize Scenes

bedroom

mountain

Learning to Recognize Scenes

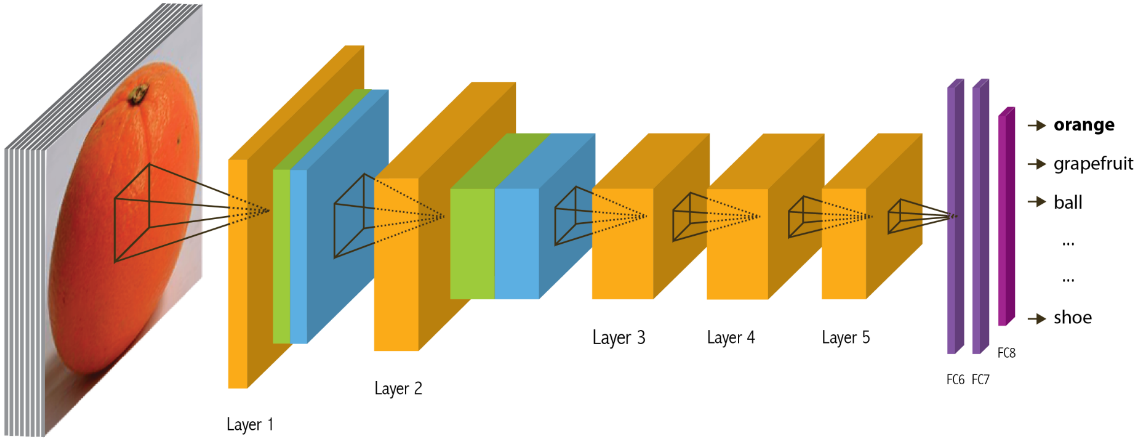
bedroom

mountain

Possible internal representations:

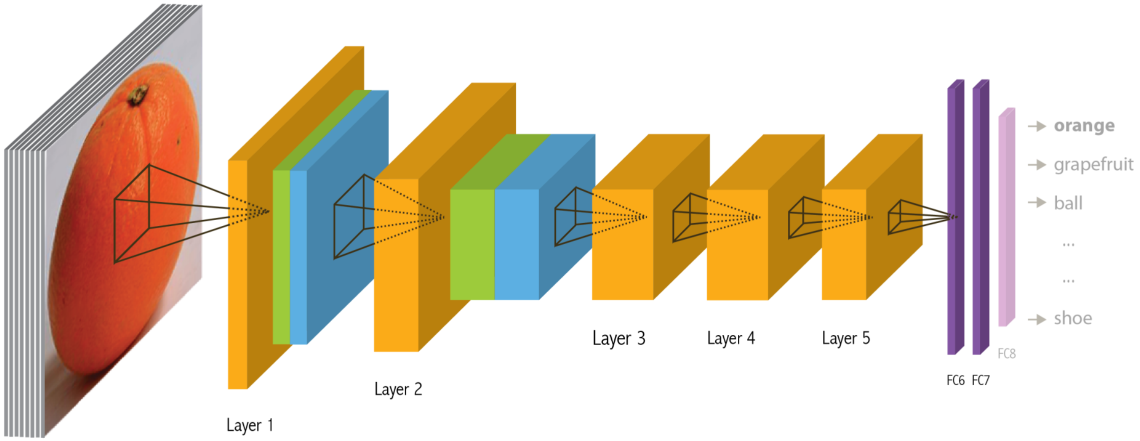
- Scene parts
- Objects
- Scene attributes
- Object parts
- Textures

Places and objects



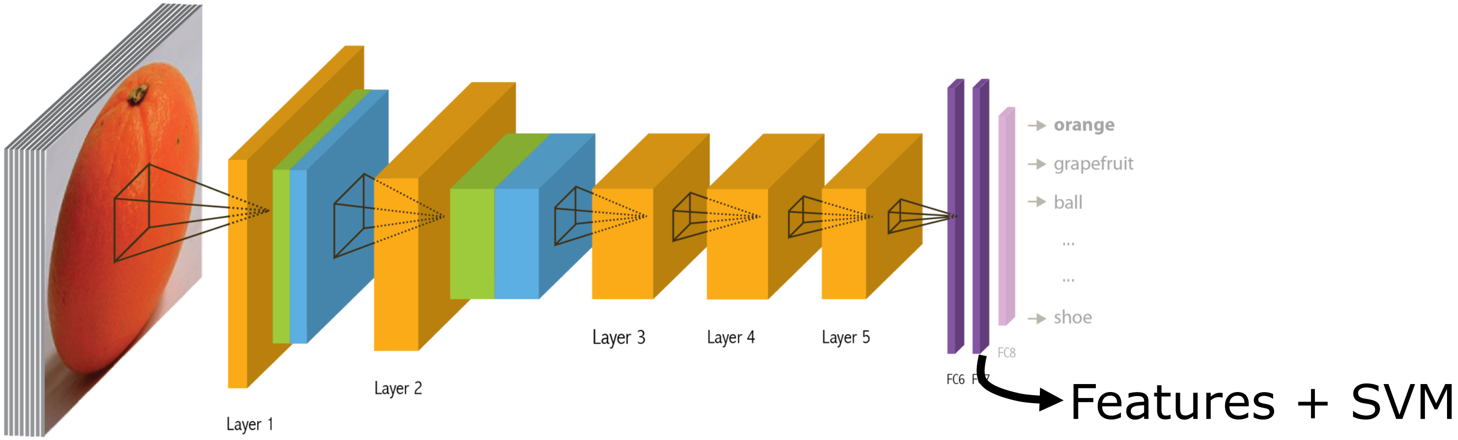
Zhou, Lapedriza, Xiao, Torralba & Oliva (NIPS 2014)

Places and objects



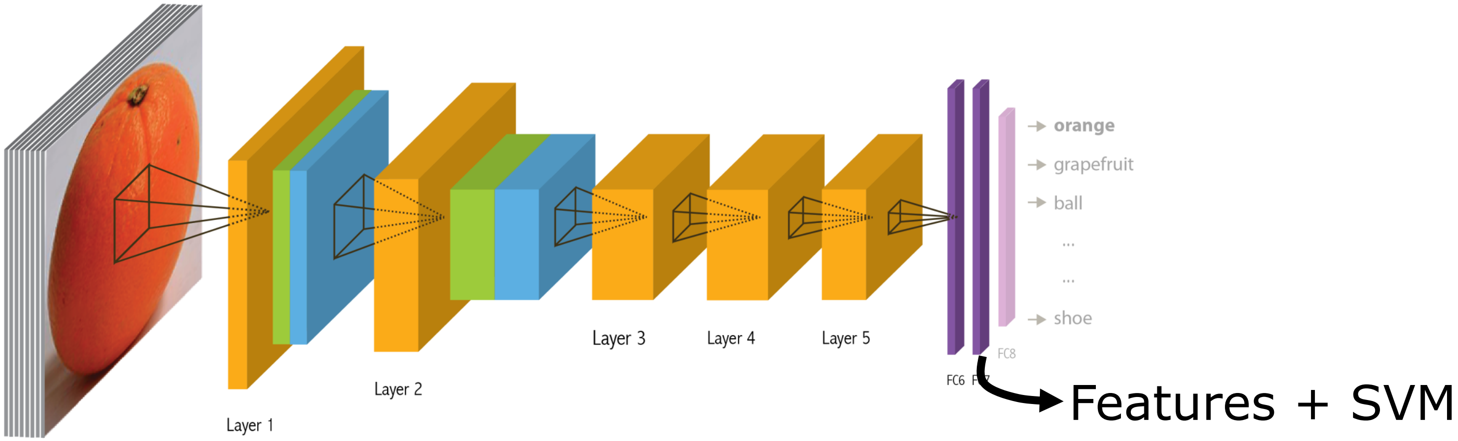
Zhou, Lapedriza, Xiao, Torralba & Oliva (NIPS 2014)

Places and objects



Zhou, Lapedriza, Xiao, Torralba & Oliva (NIPS 2014)

Places and objects

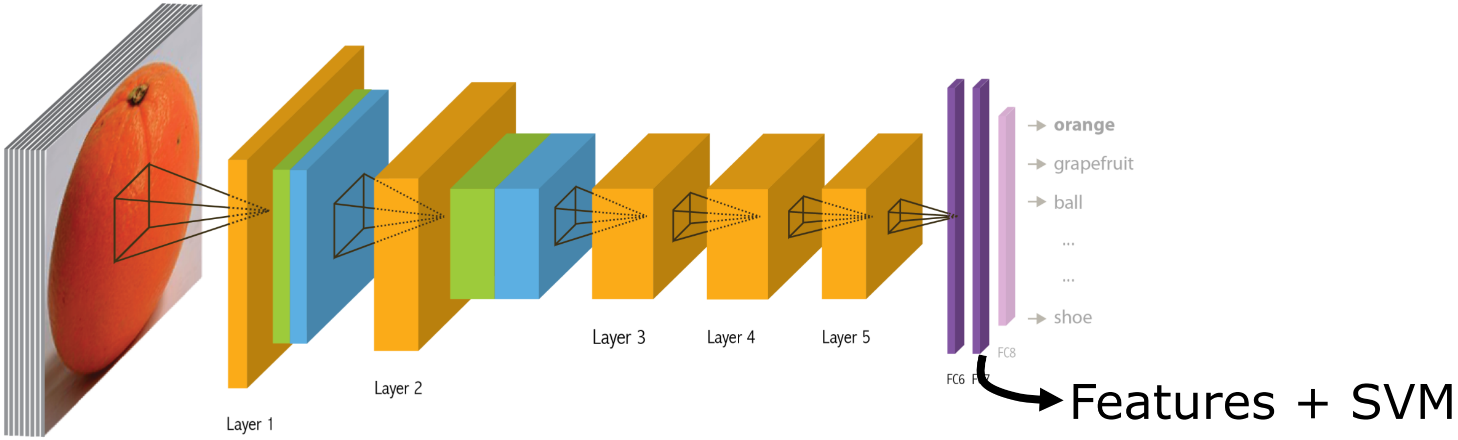


Scene datasets

	SUN397	MIT Indoor67	Scene15	SUN Attribute
Places-CNN feature	**54.32±0.14**	**68.24**	**90.19±0.34**	**91.29**
ImageNet-CNN feature	42.61±0.16	56.79	84.23±0.37	89.85

Zhou, Lapedriza, Xiao, Torralba & Oliva (NIPS 2014)

Places and objects



Scene datasets

	SUN397	MIT Indoor67	Scene15	SUN Attribute
Places-CNN feature	**54.32±0.14**	**68.24**	**90.19±0.34**	**91.29**
ImageNet-CNN feature	42.61±0.16	56.79	84.23±0.37	89.85

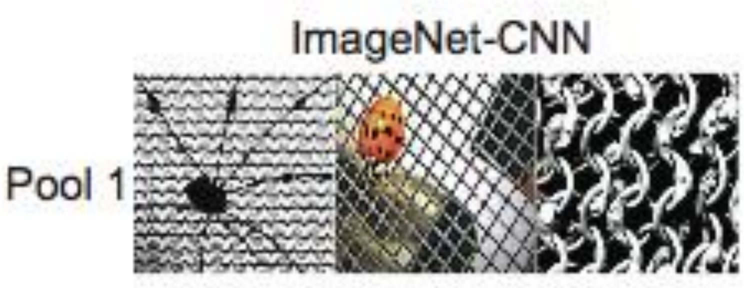
Object datasets

	Caltech101	Caltech256	Action40	Event8
Places-CNN feature	65.18±0.88	45.59±0.31	42.86±0.25	94.12±0.99
ImageNet-CNN feature	**87.22±0.92**	**67.23±0.27**	**54.92±0.33**	**94.42±0.76**

Zhou, Lapedriza, Xiao, Torralba & Oliva (NIPS 2014)

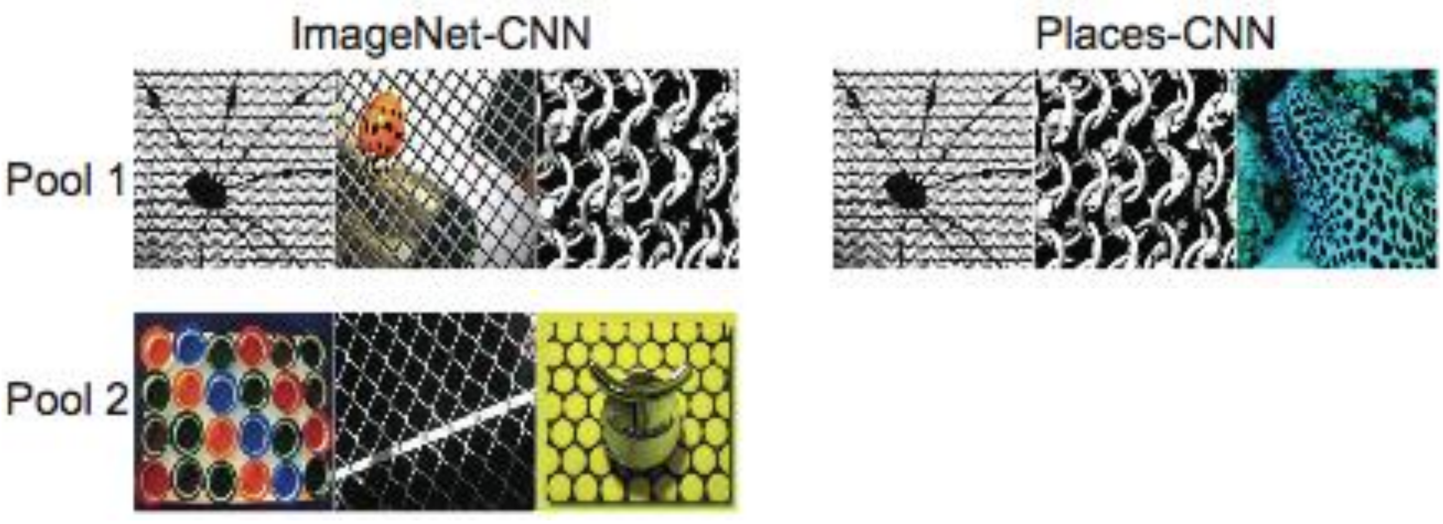
Preferred images

Preferred images

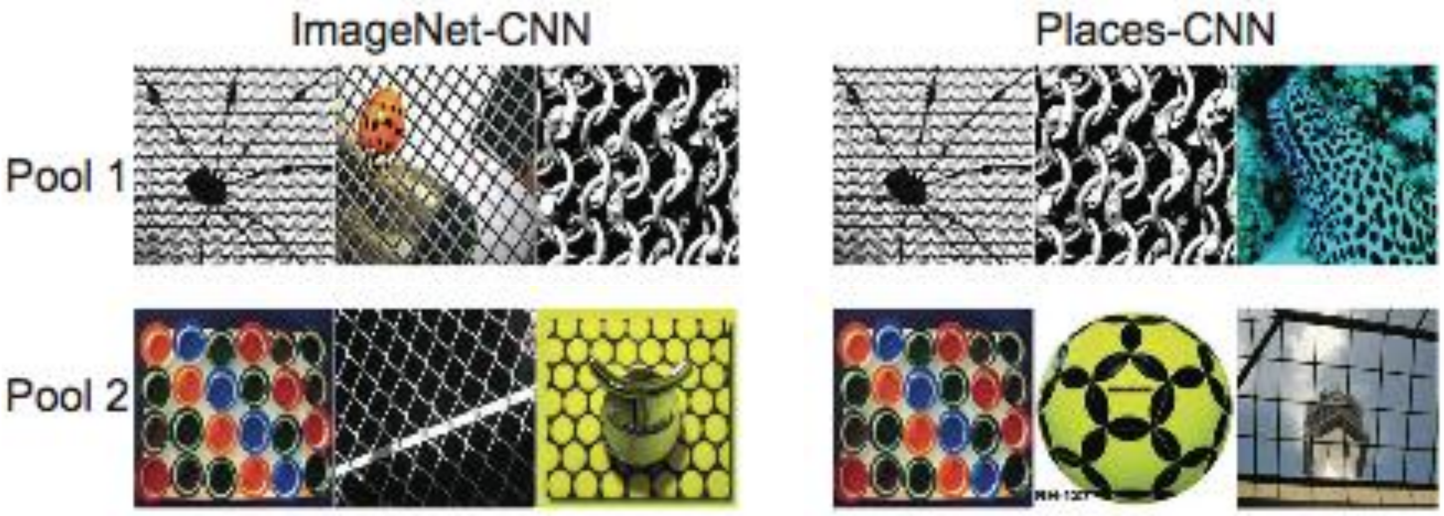


Preferred images

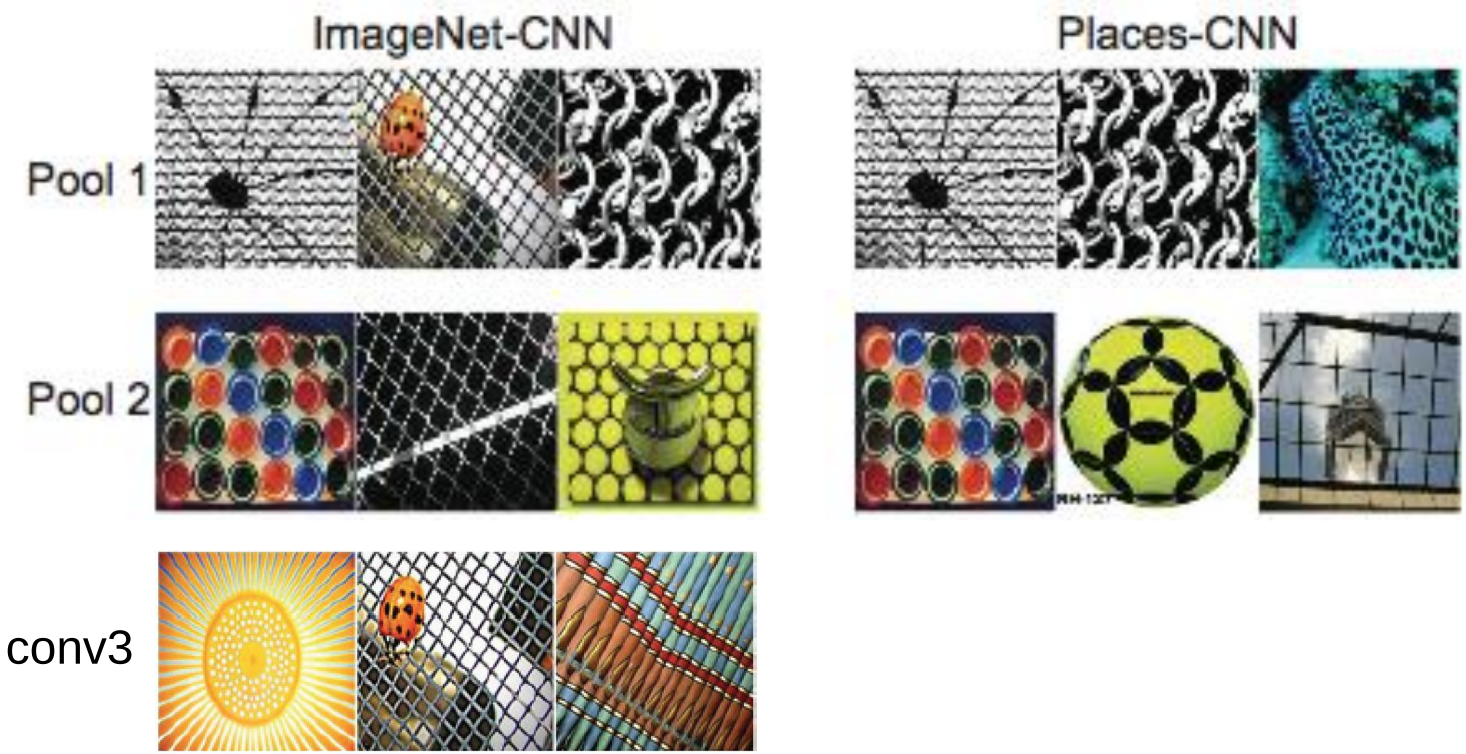
Preferred images



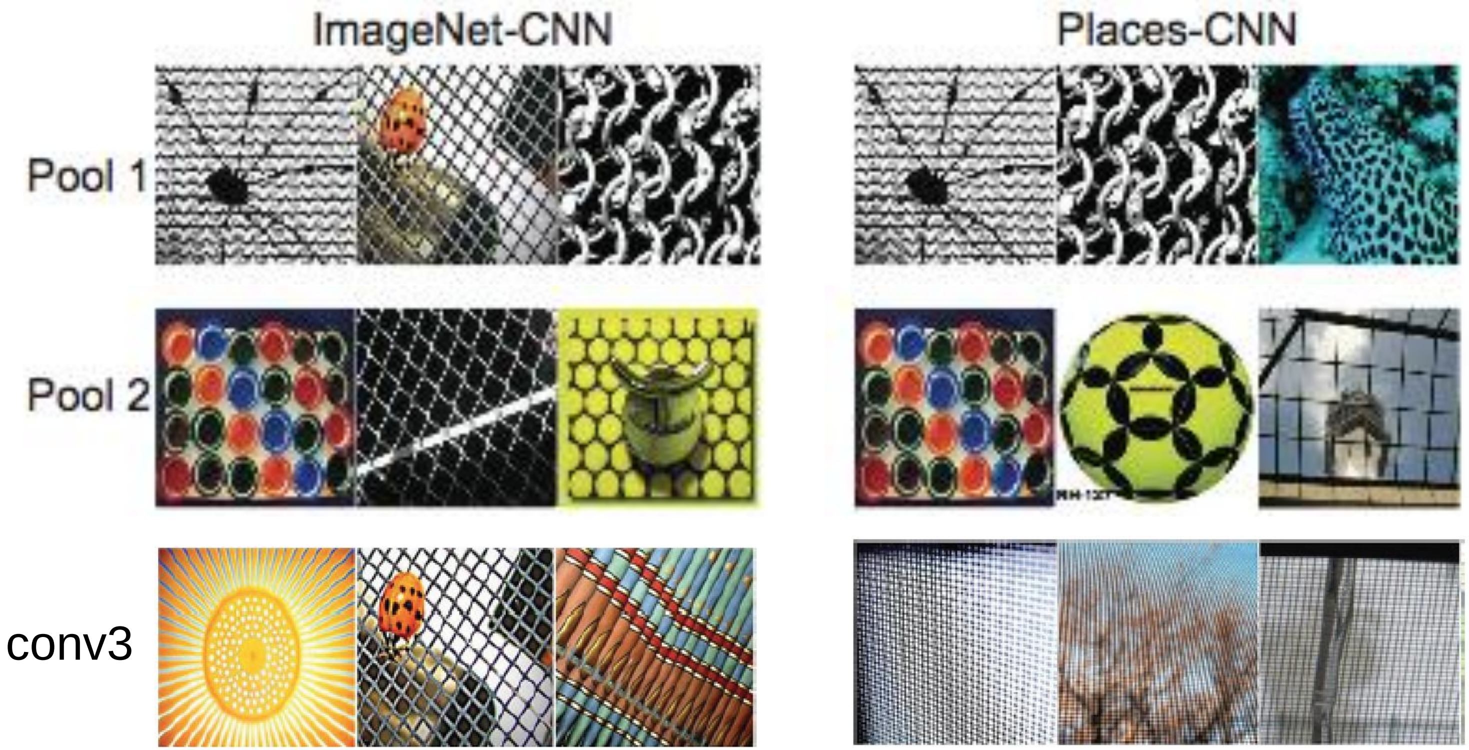
Preferred images



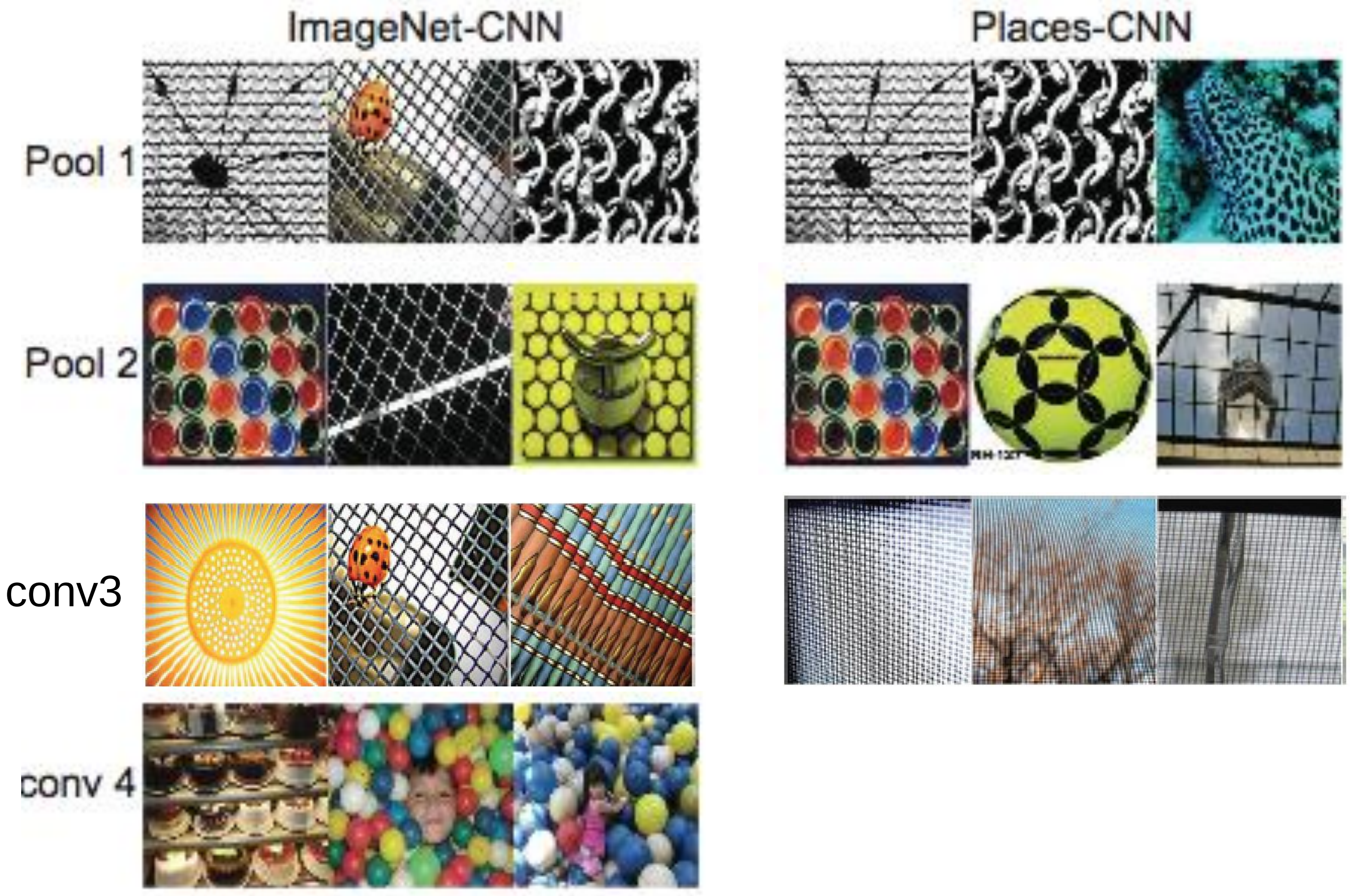
Preferred images



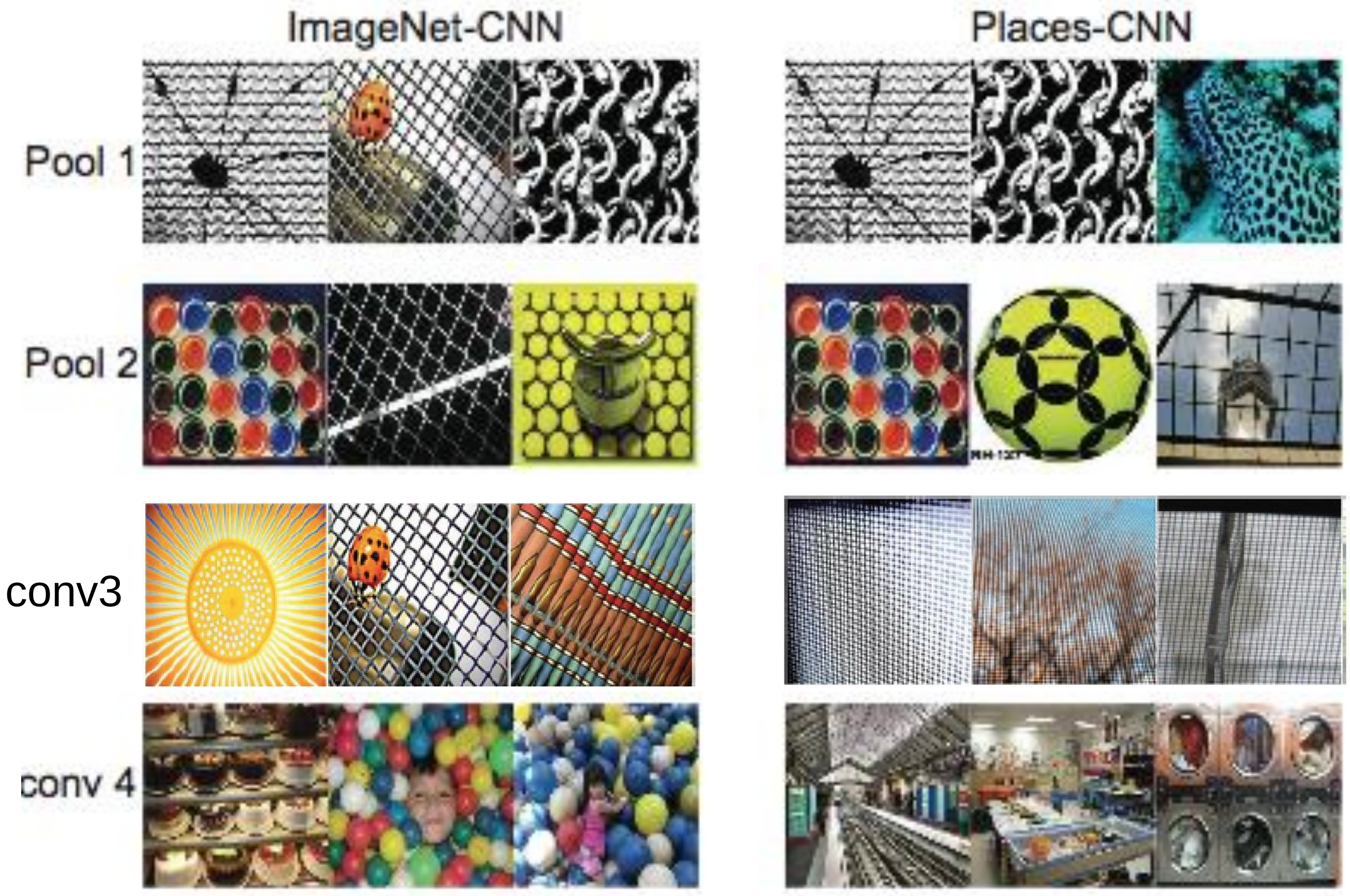
Preferred images



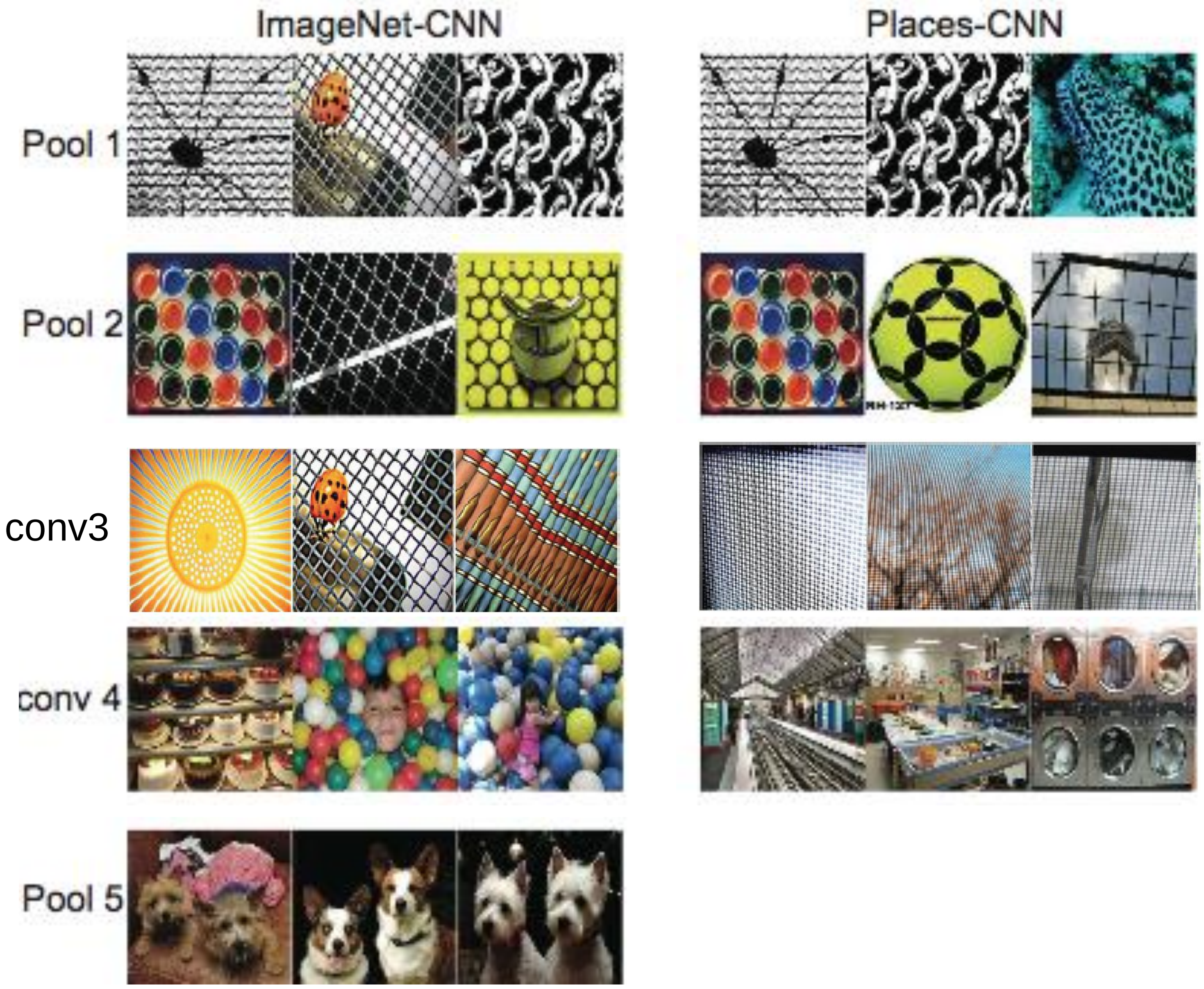
Preferred images



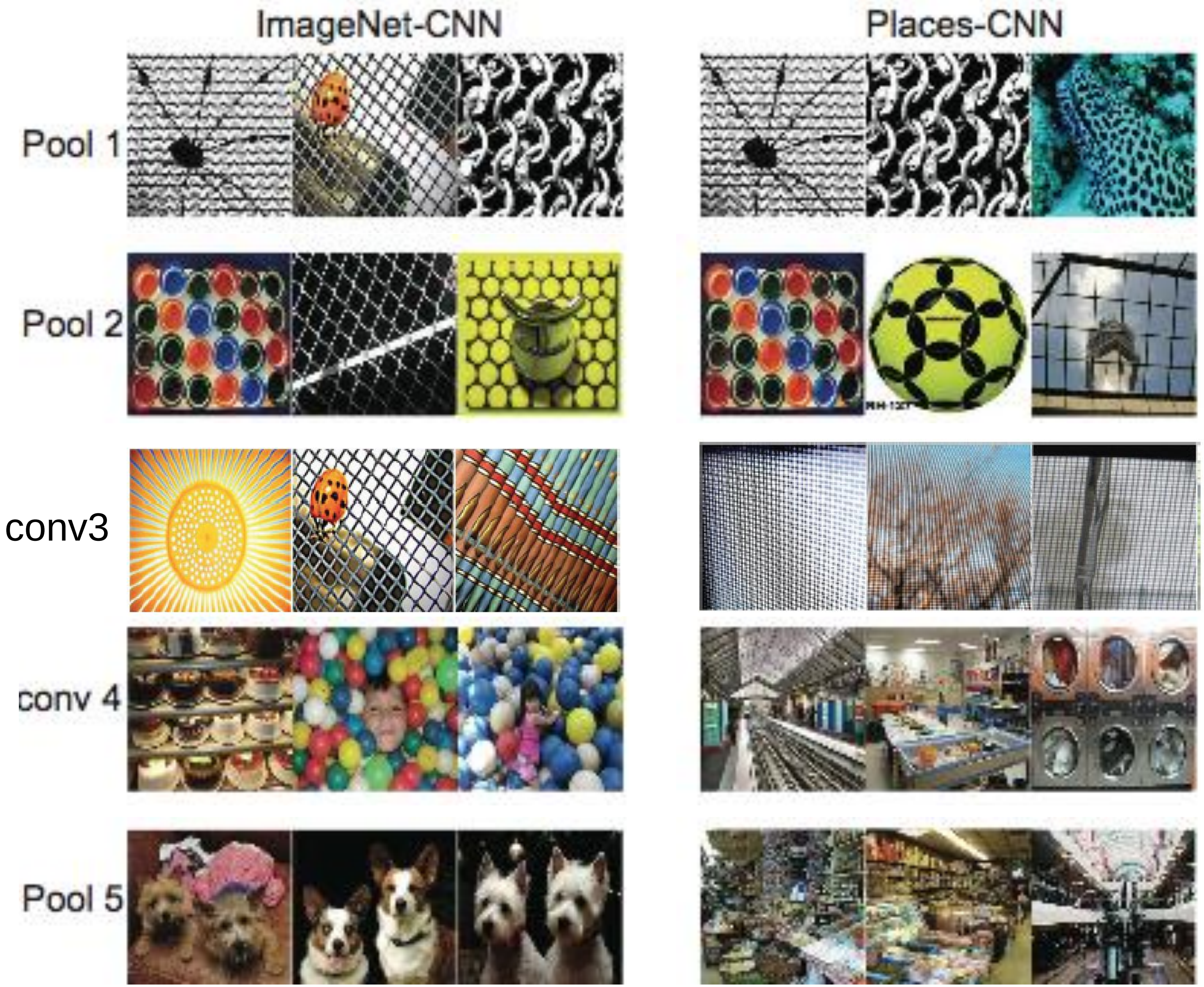
Preferred images



Preferred images

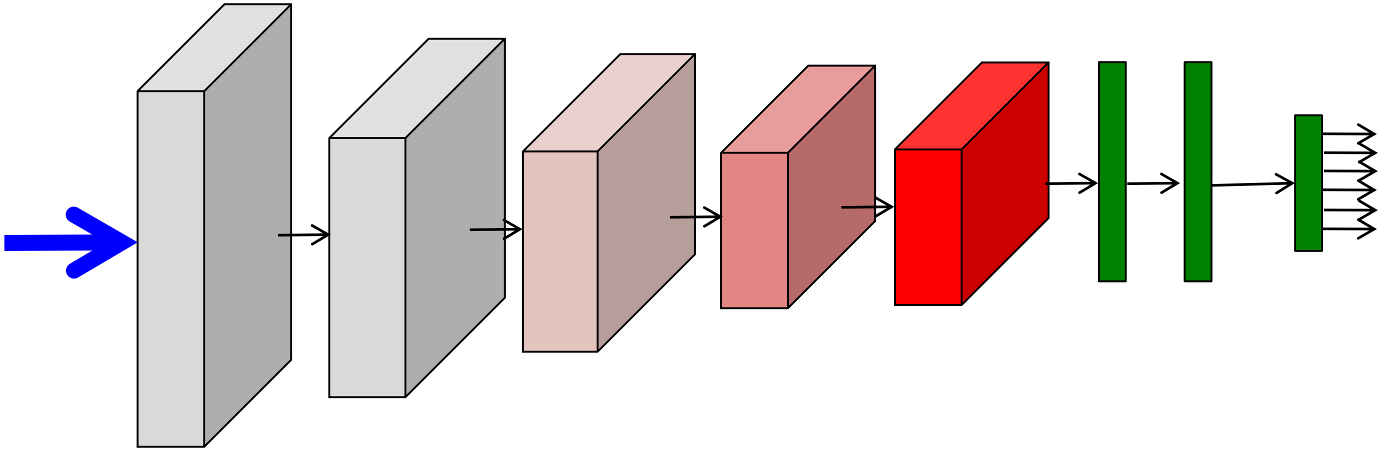


Preferred images



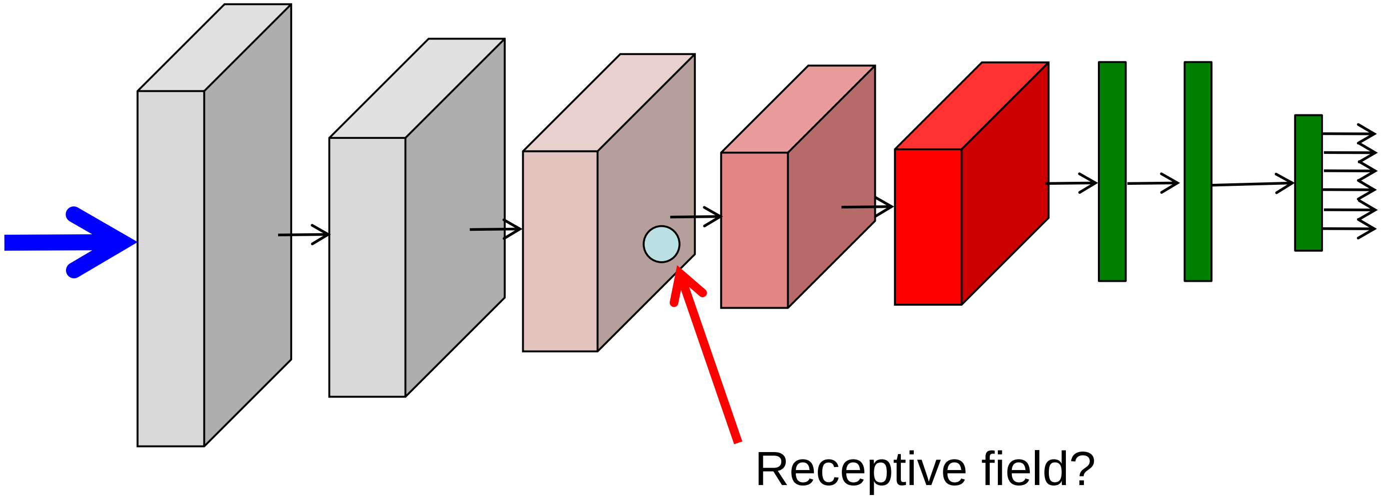
Estimating the receptive field

■

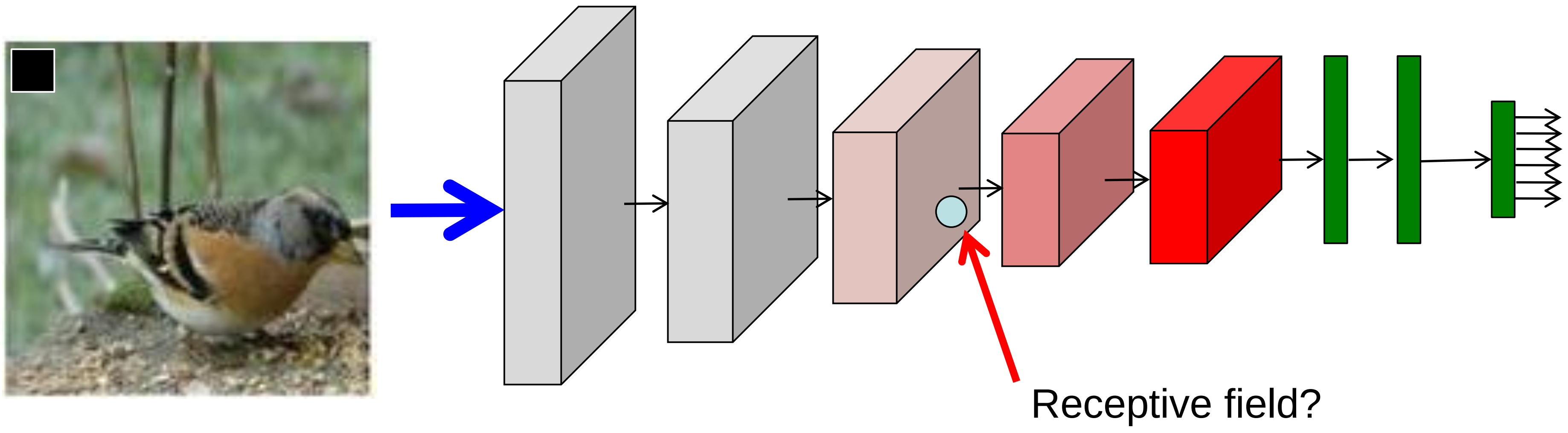


Estimating the receptive field

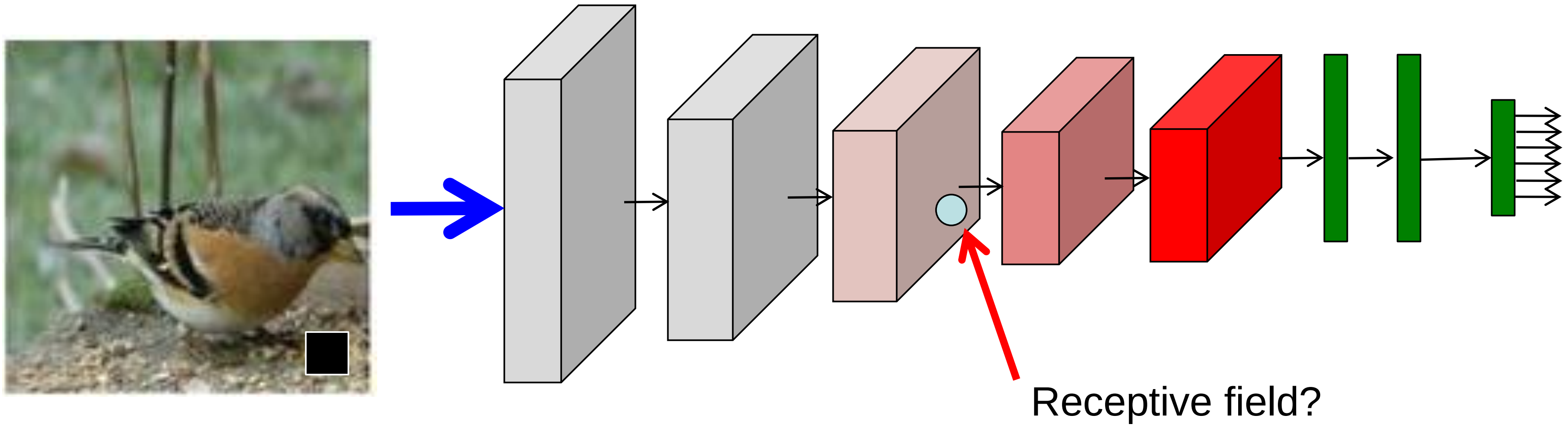
■



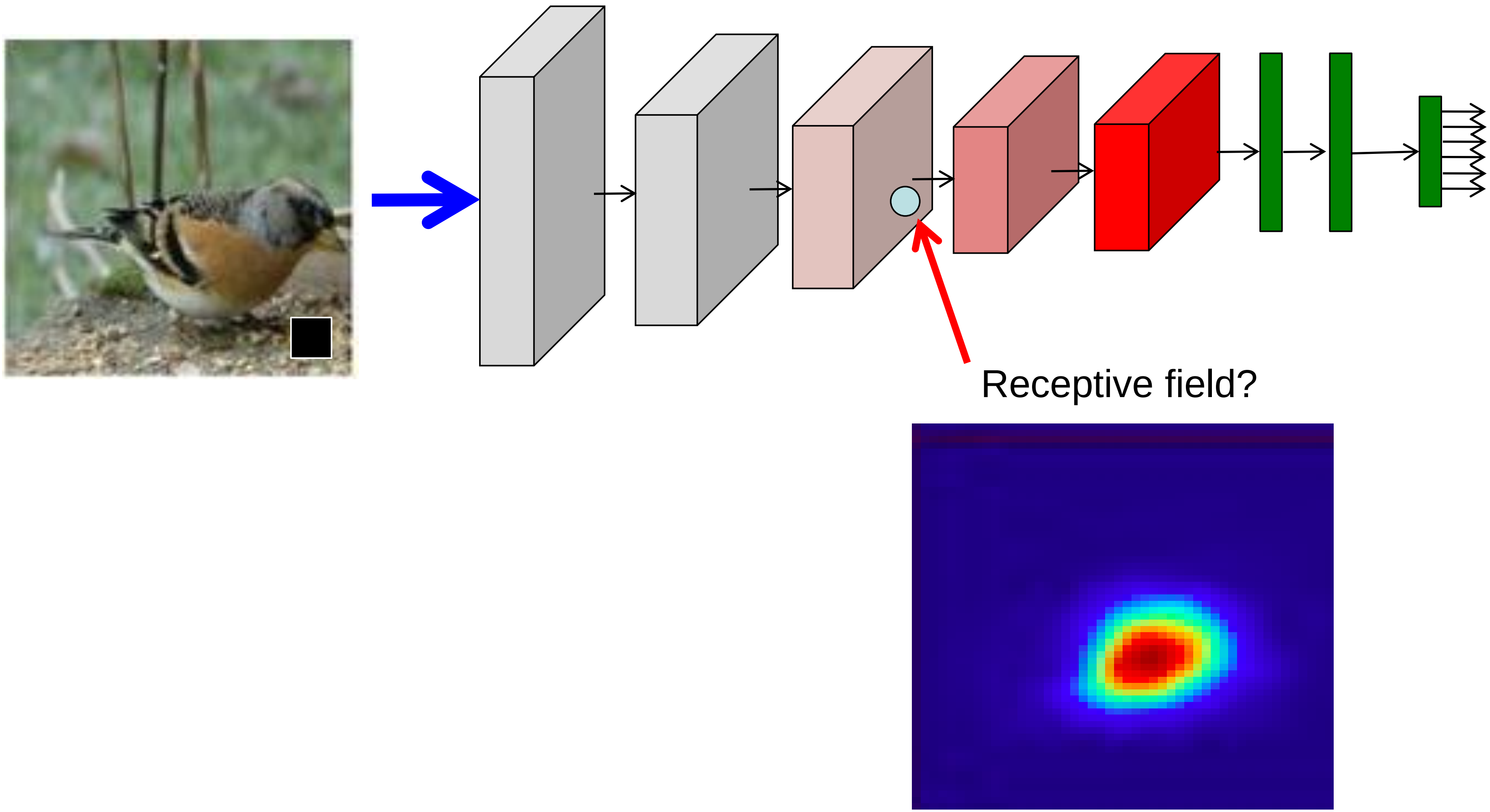
Estimating the receptive field



Estimating the receptive field



Estimating the receptive field



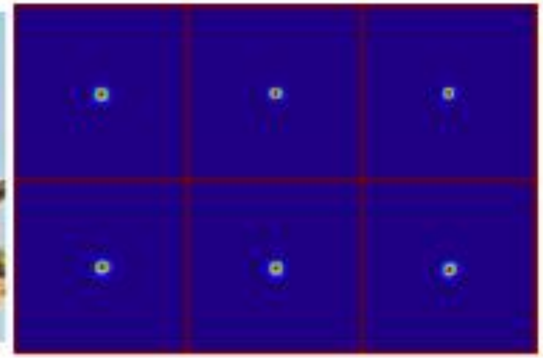
Estimating the receptive field

Theoretical size

Actual size

Estimating the receptive field

Layer 1



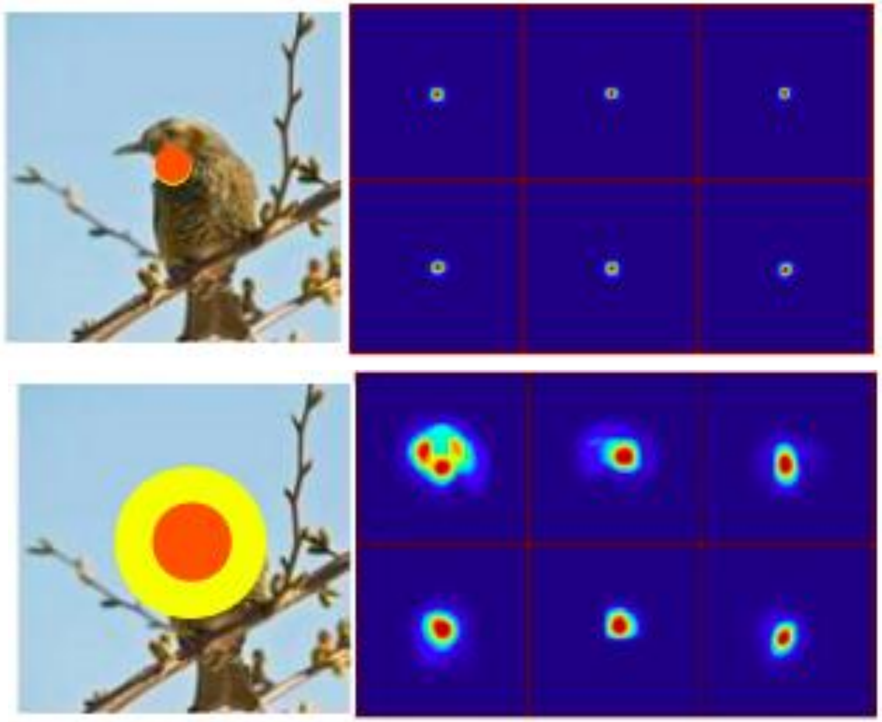
Theoretical size

Actual size

Estimating the receptive field

Layer 1

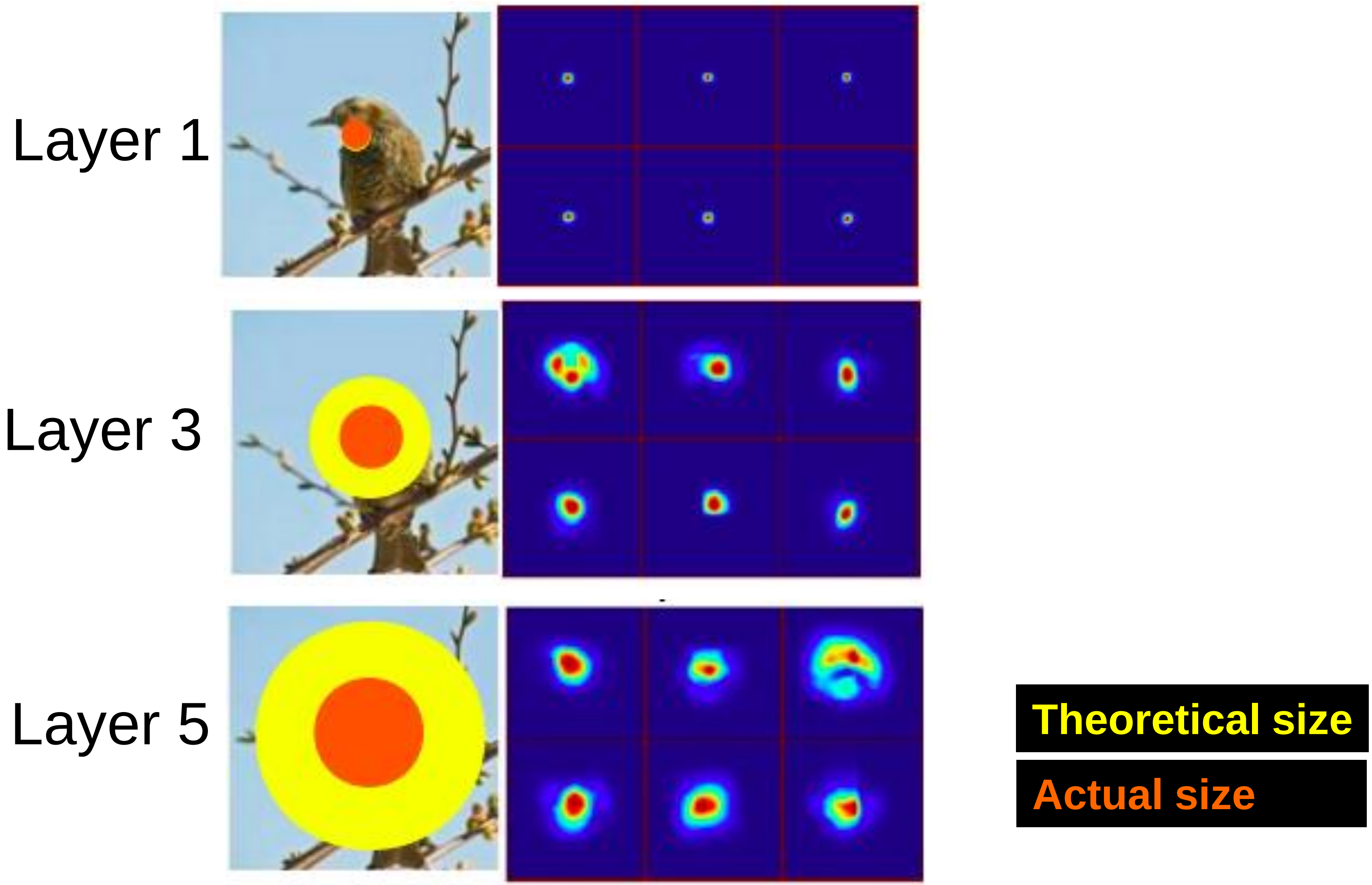
Layer 3



Theoretical size

Actual size

Estimating the receptive field



Generating segmentations

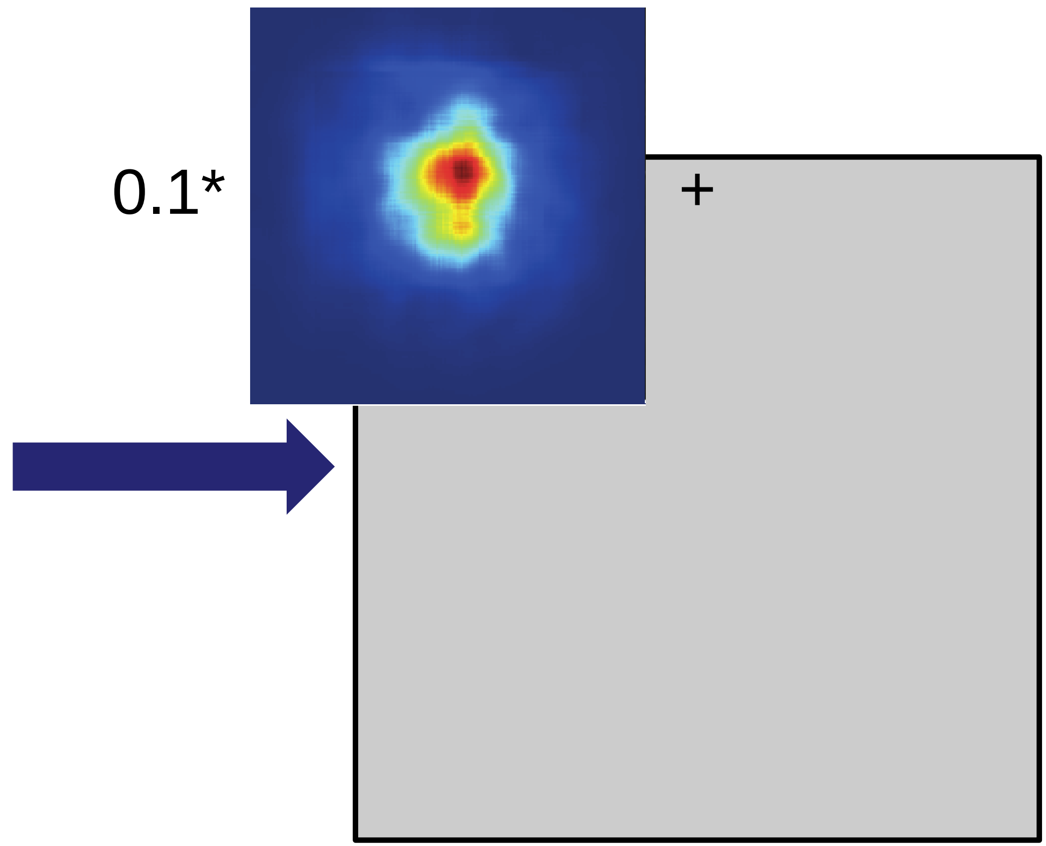
0.1	0.7	0.5
0.3	0.6	0.4
0	0.1	0
0	0	0

feature map

Generating segmentations

0.1	0.7	0.5
0.3	0.6	0.4
0	0.1	0
0	0	0

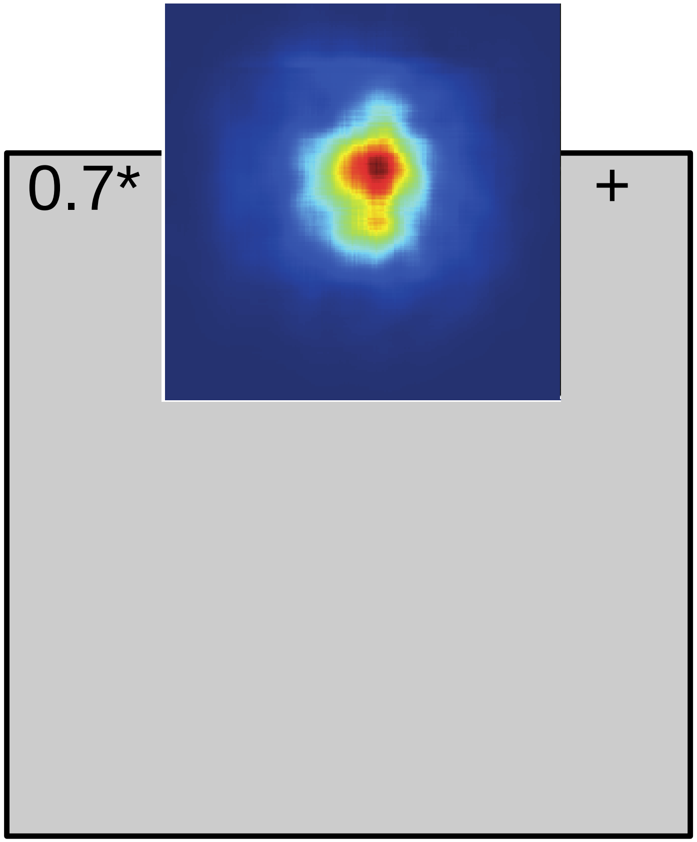
feature map



Generating segmentations

0.1	0.7	0.5
0.3	0.6	0.4
0	0.1	0
0	0	0

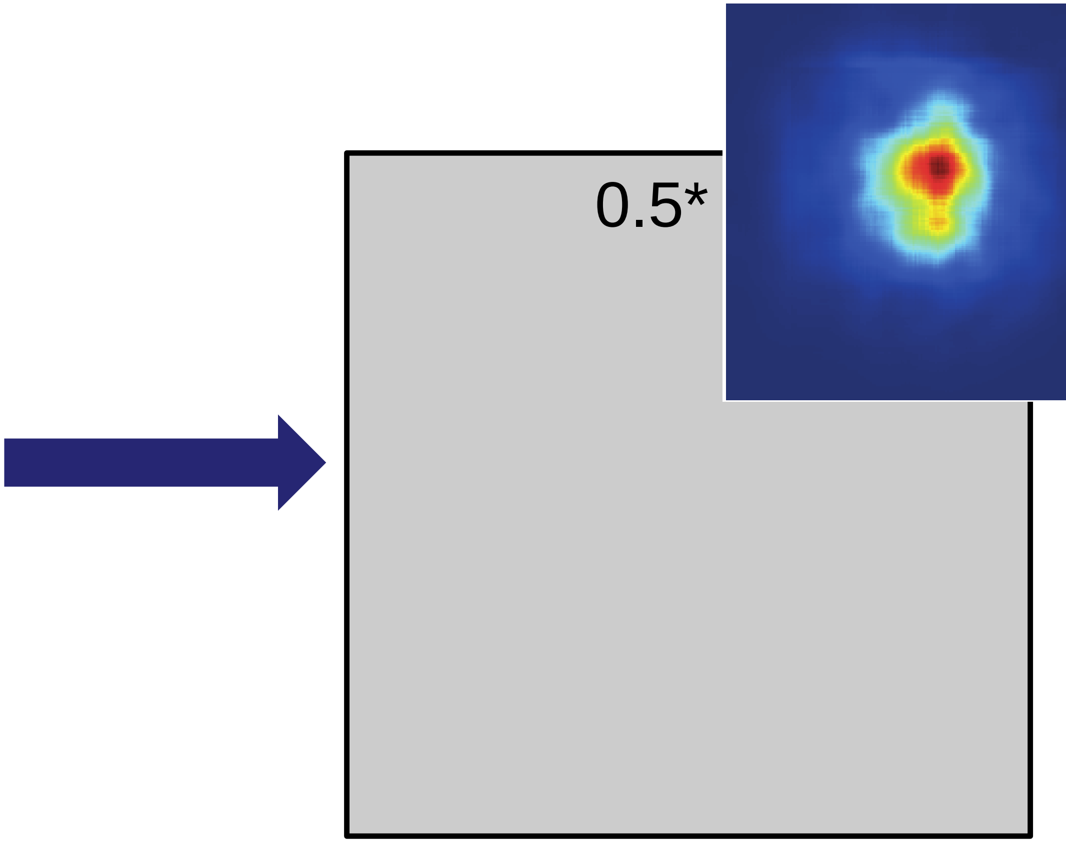
feature map



Generating segmentations

0.1	0.7	0.5
0.3	0.6	0.4
0	0.1	0
0	0	0

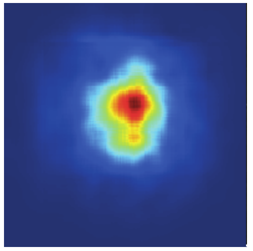
feature map



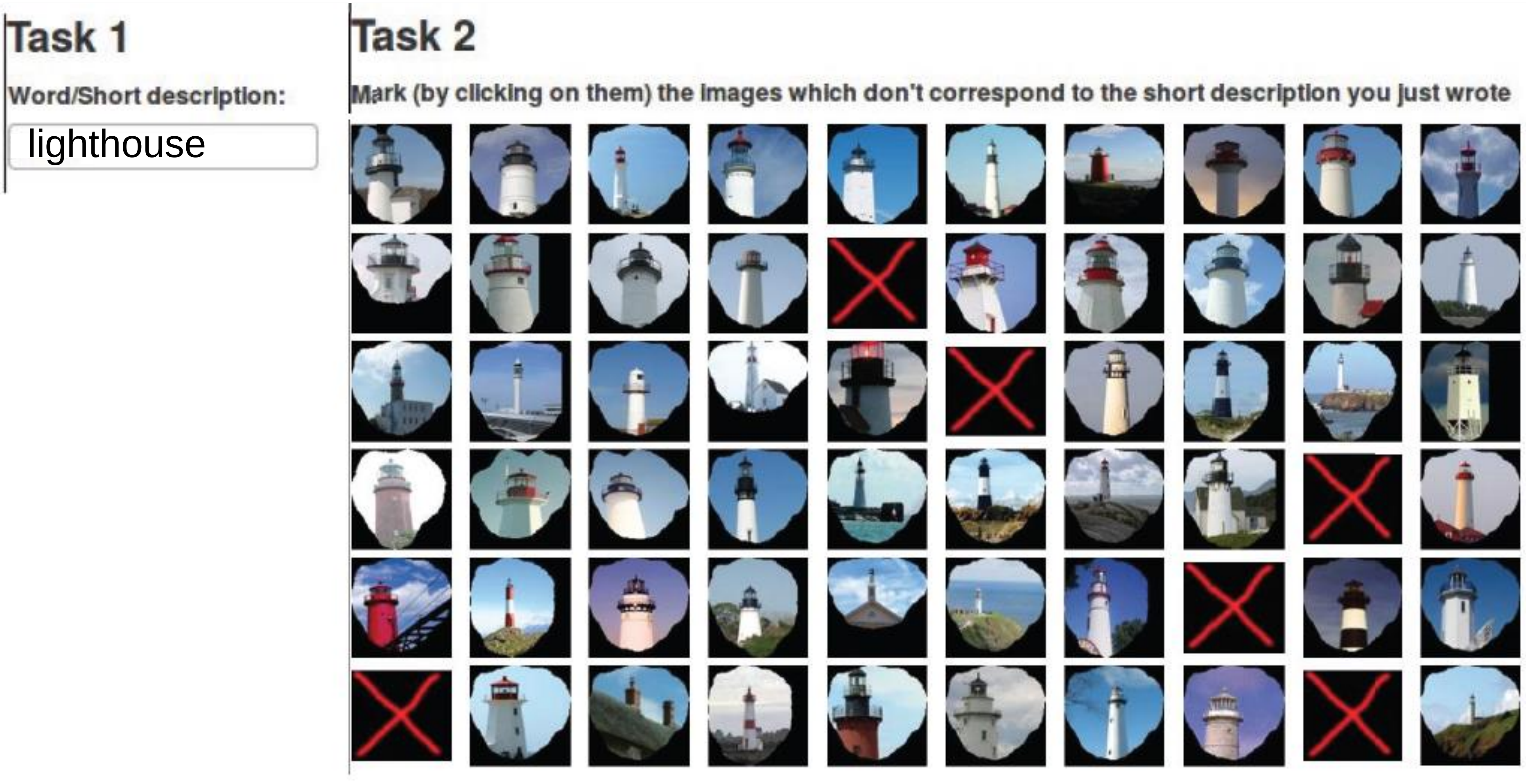
Generating segmentations

0.1	0.7	0.5
0.3	0.6	0.4
0	0.1	0
0	0	0

feature map



Crowdsourcing units



Crowdsourcing units

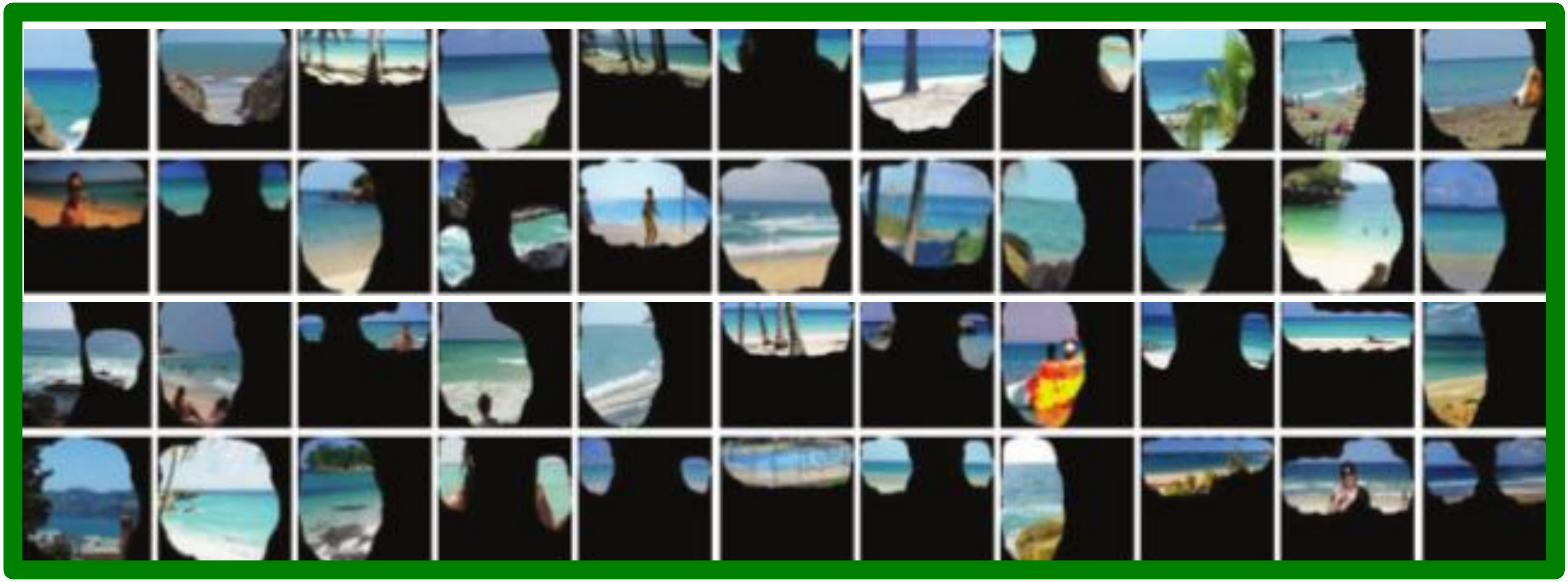
Task 3

Which category does your short description mostly belong to?

- ○ **Scene (kitchen, corridor, street, beach, ...)**
- ○ **Region or surface (road, grass, wall, floor, sky, ...)**
- ● **Object (bed, car, building, tree, ...)**
- ○ **Object part (leg, head, wheel, roof, ...)**
- ○ **Texture or material (striped, rugged, wooden, plastic, ...)**
- ○ **Simple elements or colors (vertical line, curved line, color blue, ...)**

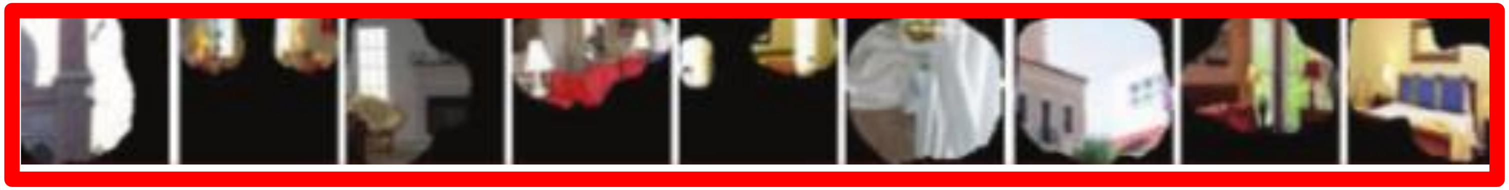
Annotating the Semantics of Units

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



Annotating the Semantics of Units

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

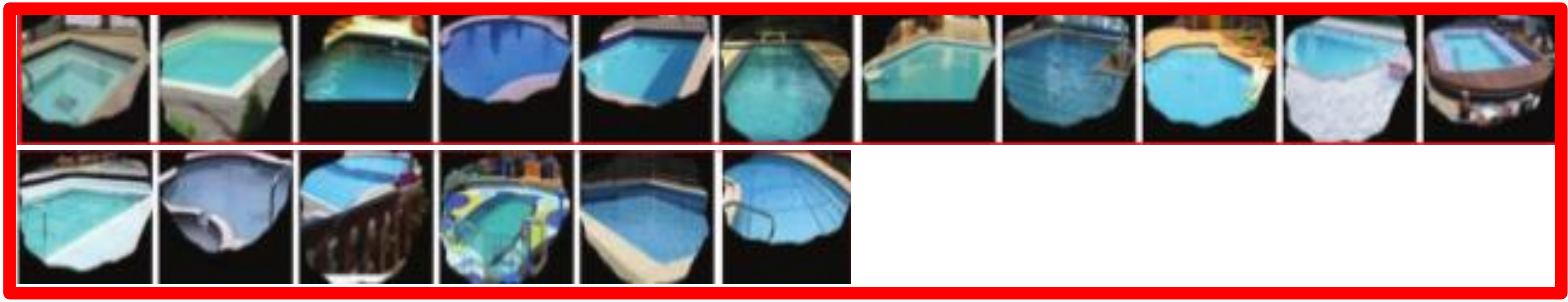


Annotating the Semantics of Units

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

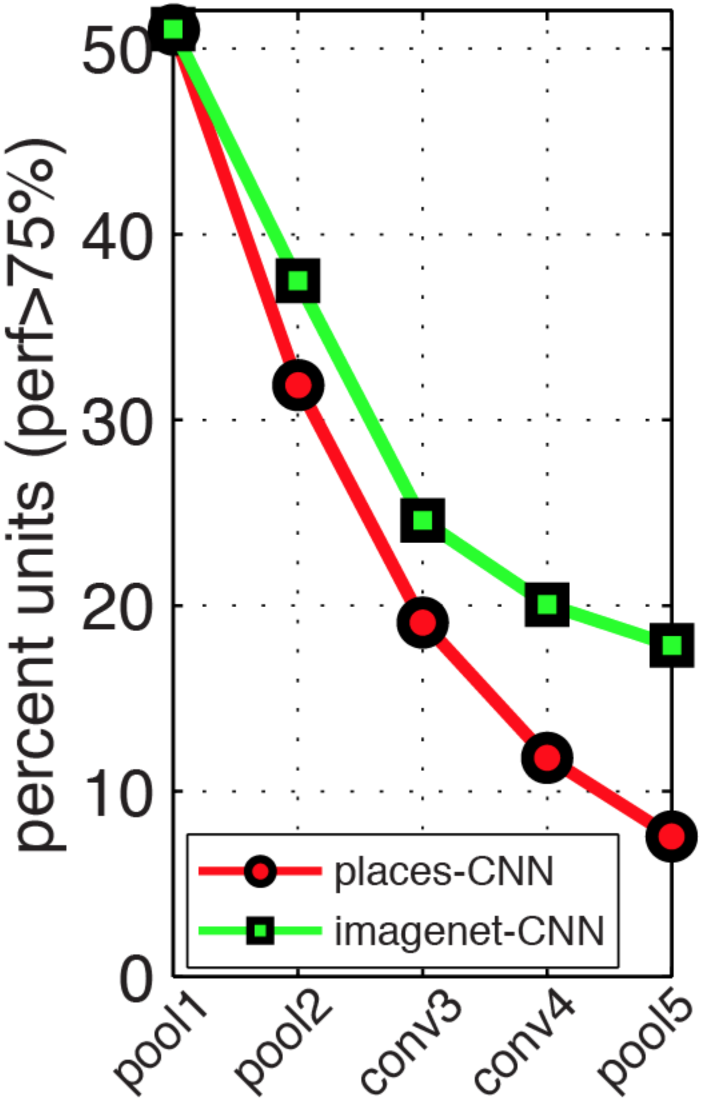
1 - Simple elements and colors

Ex: vertical line, curved line, color blue, ….

Distribution of semantic types at each layer

1 - Simple elements and colors

Ex: vertical line, curved line, color blue, ….



Distribution of semantic types at each layer

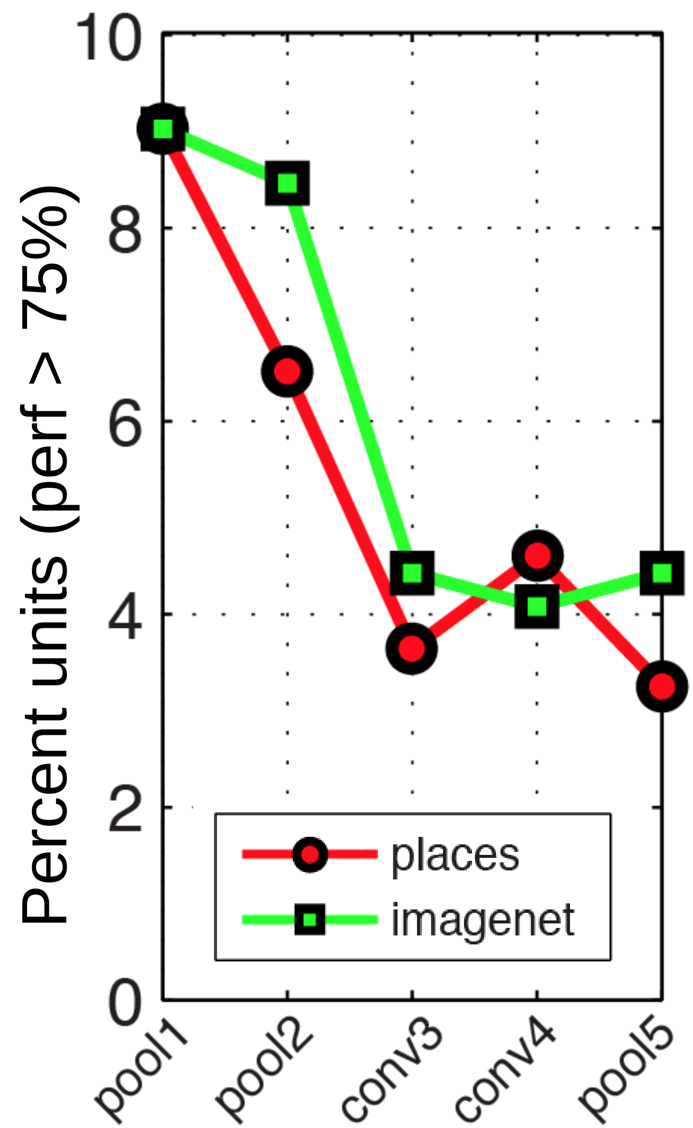
2 - Texture or materials

Ex: stripes, wooden, plastic, ...

Distribution of semantic types at each layer

2 - Texture or materials

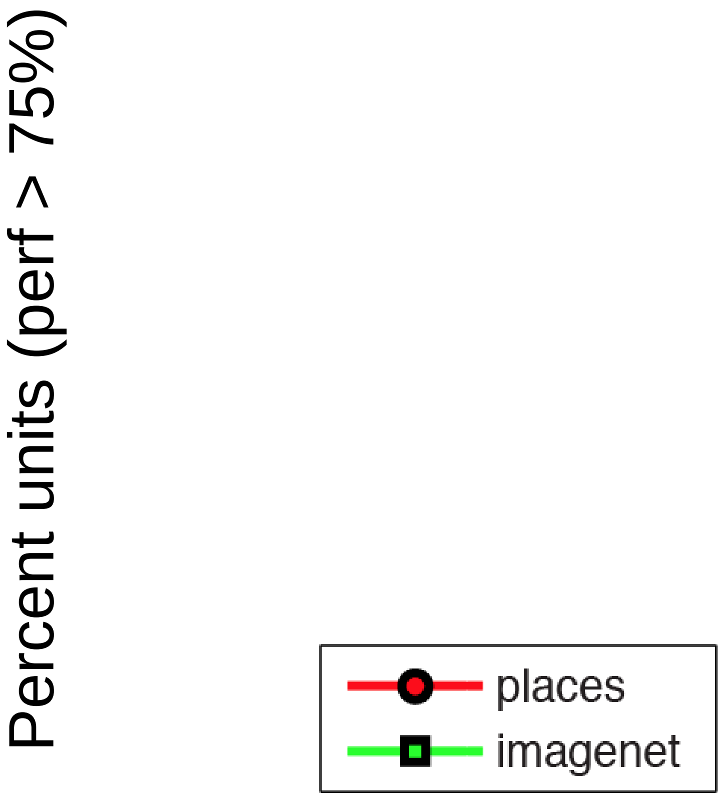
Ex: stripes, wooden, plastic, ...



Distribution of semantic types at each layer

3 - Regions and surfaces

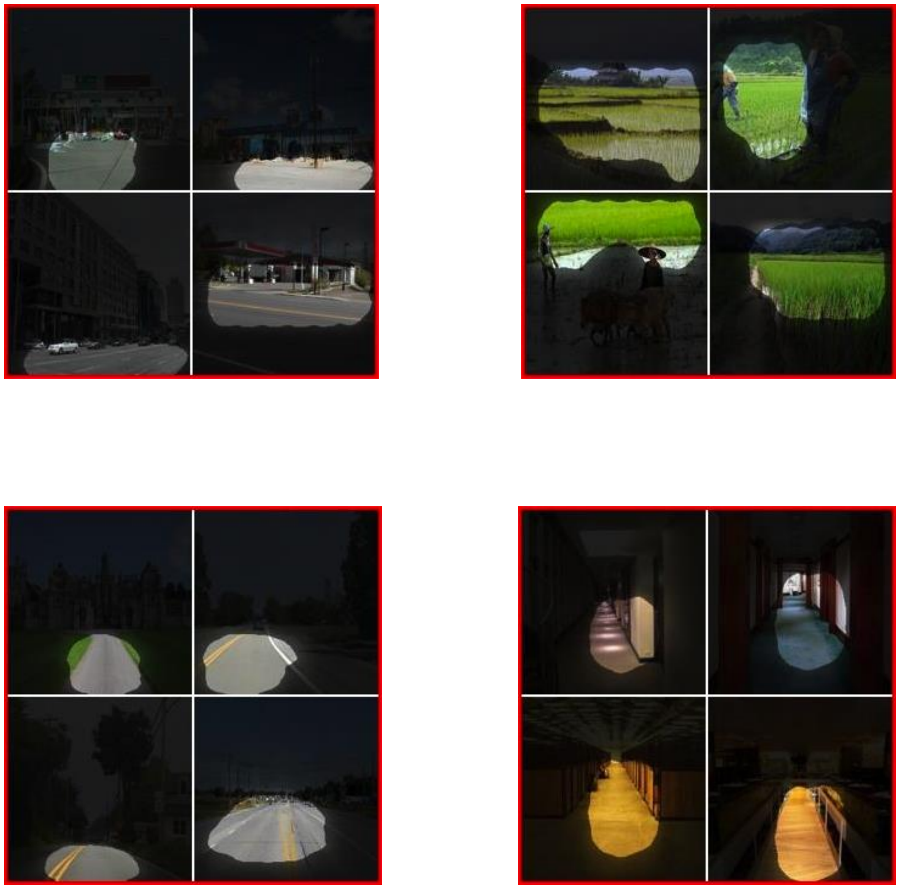
Ex: Road, grass, wall, floor, sky, ….

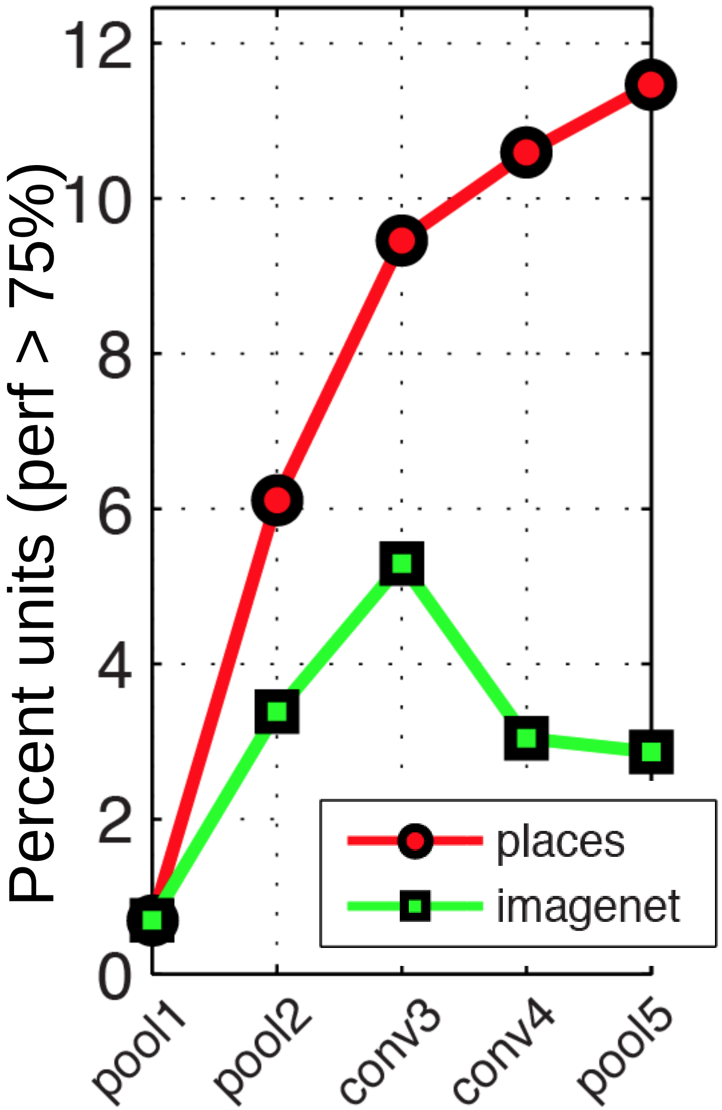


Distribution of semantic types at each layer

3 - Regions and surfaces

Ex: Road, grass, wall, floor, sky, ….

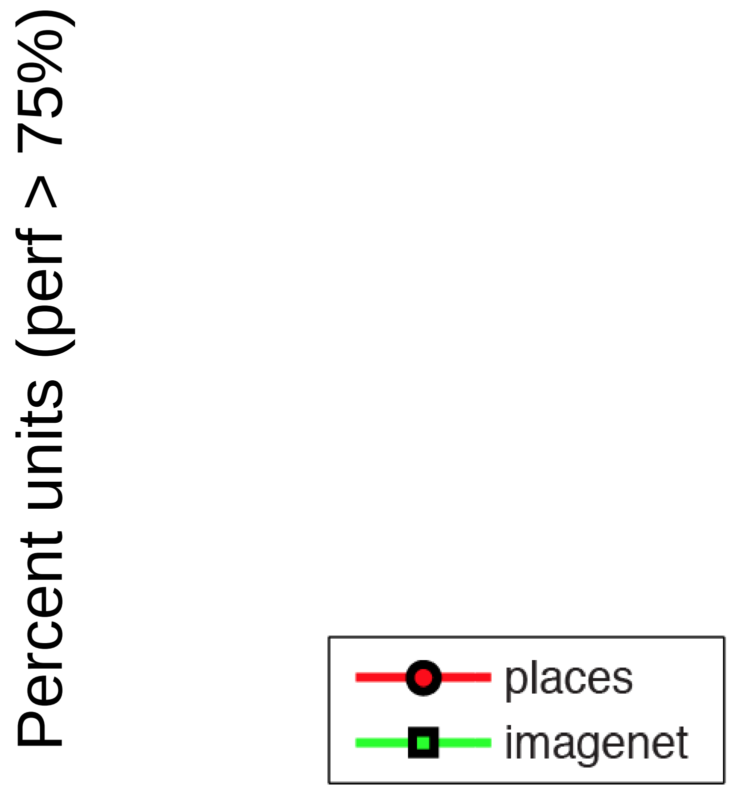




Distribution of semantic types at each layer

4 - Object parts

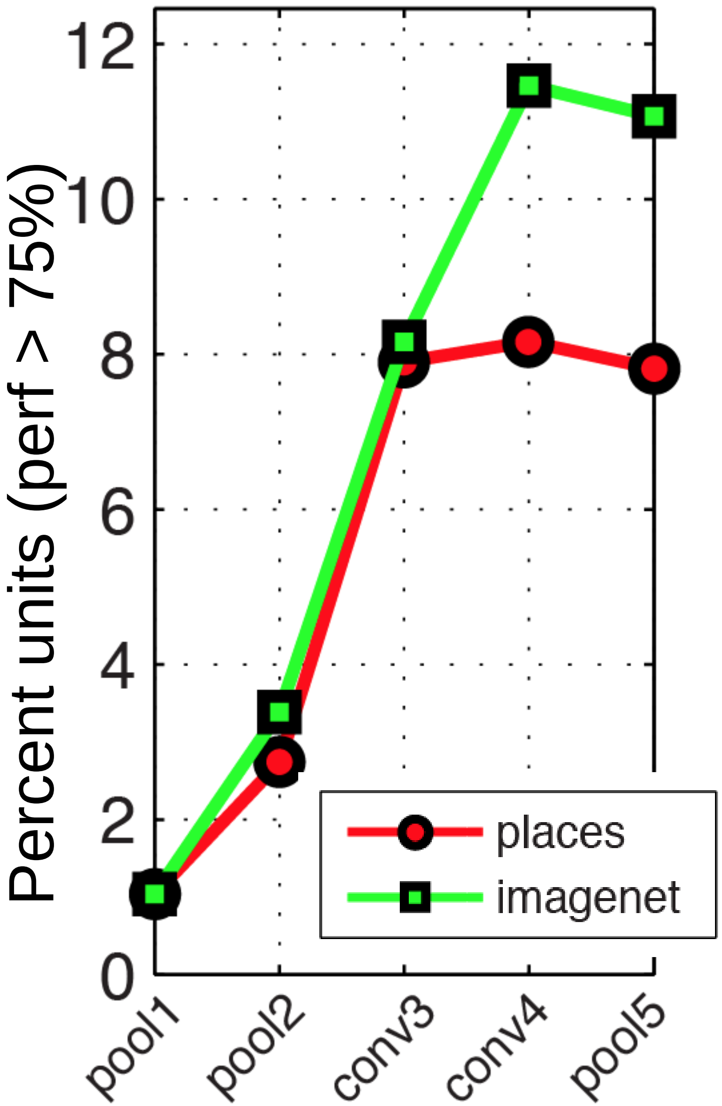
Ex: leg, head, wheel, roof, ….



Distribution of semantic types at each layer

4 - Object parts

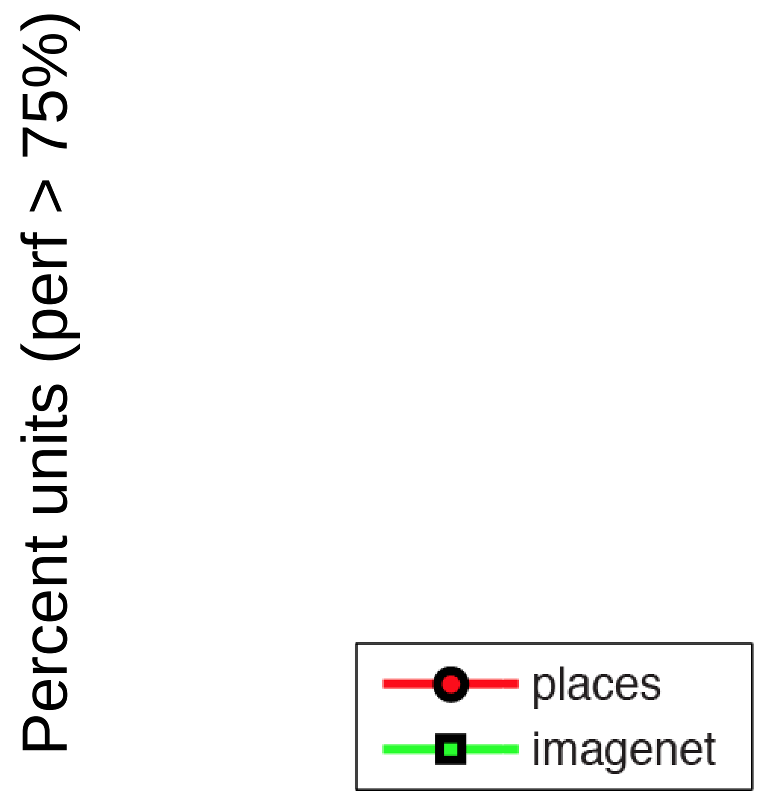
Ex: leg, head, wheel, roof, ….



Distribution of semantic types at each layer

5 - Objects

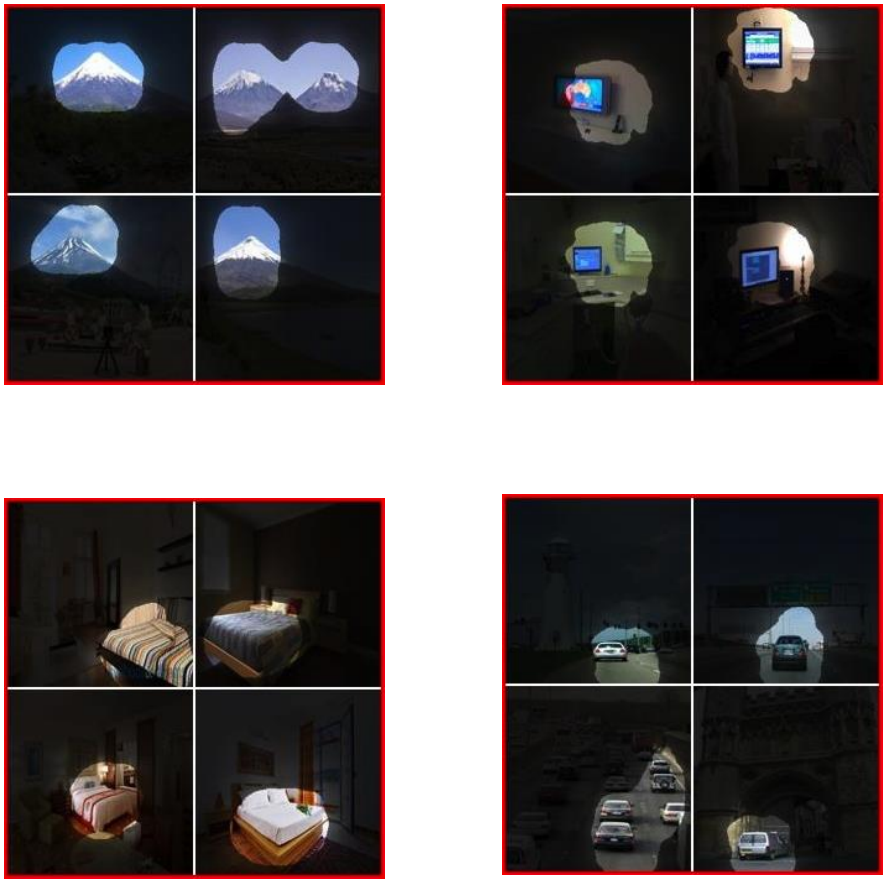
Ex: bed, car, building, tree, ….

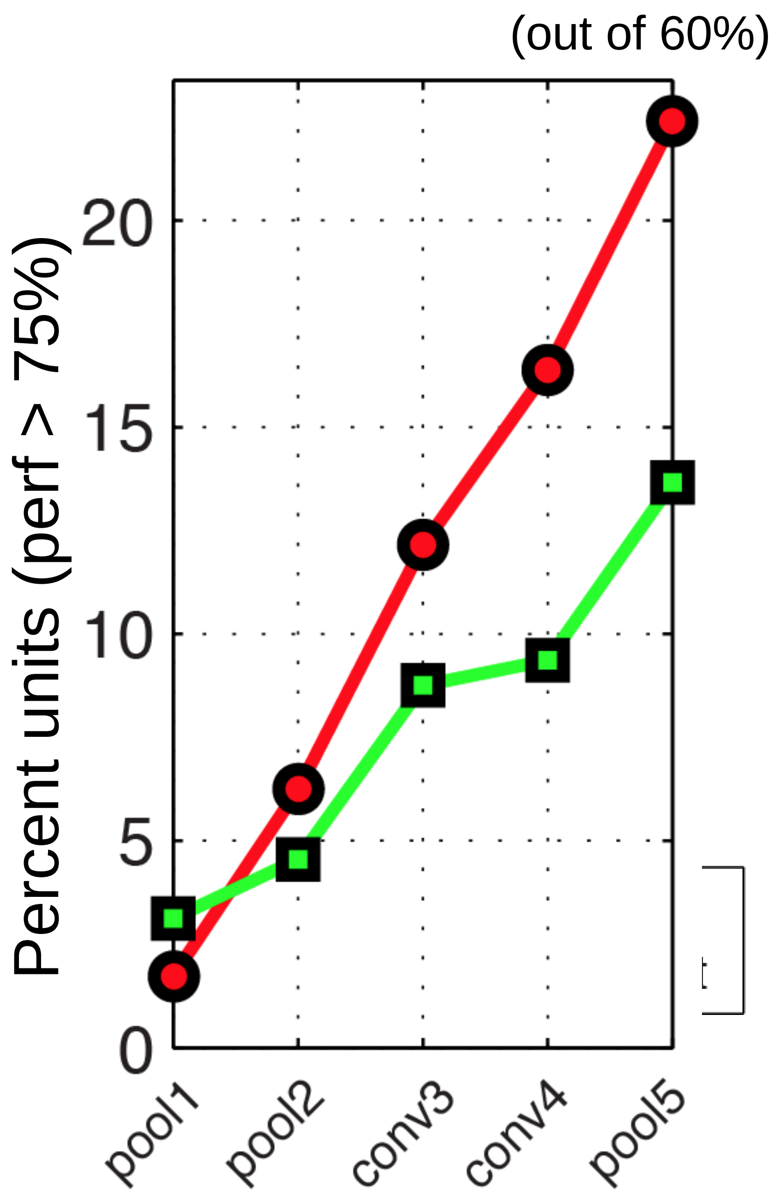


Distribution of semantic types at each layer

5 - Objects

Ex: bed, car, building, tree, ….





Distribution of semantic types at each layer

6 - Scenes

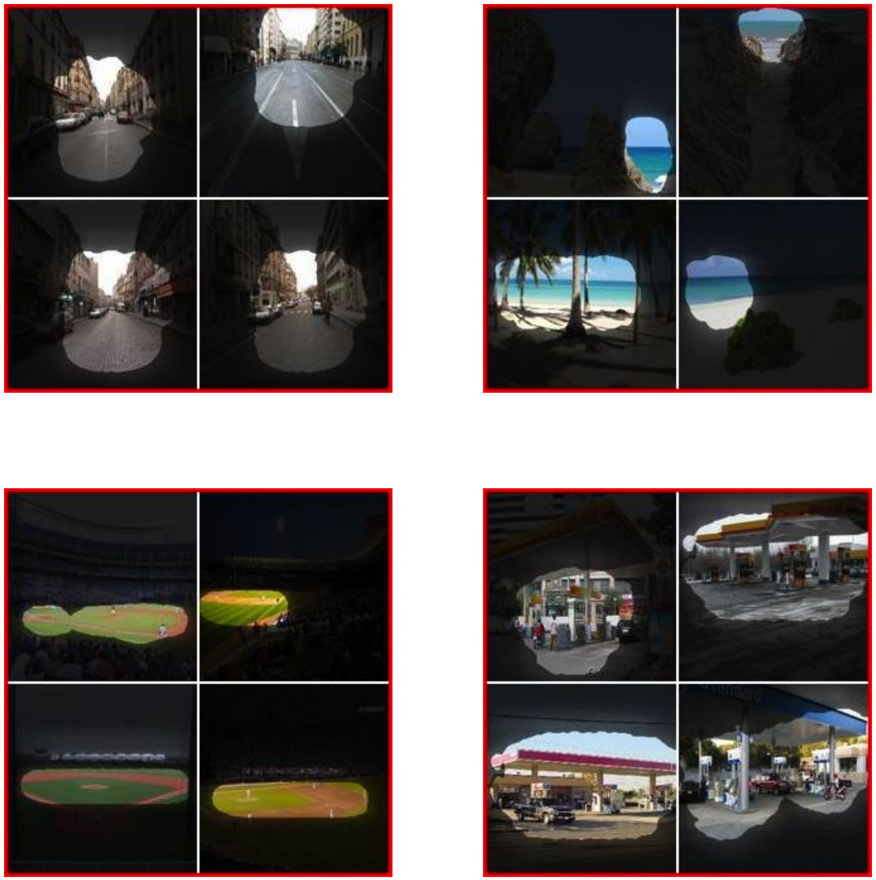
Ex: kitchen, corridor, street, beach, ….

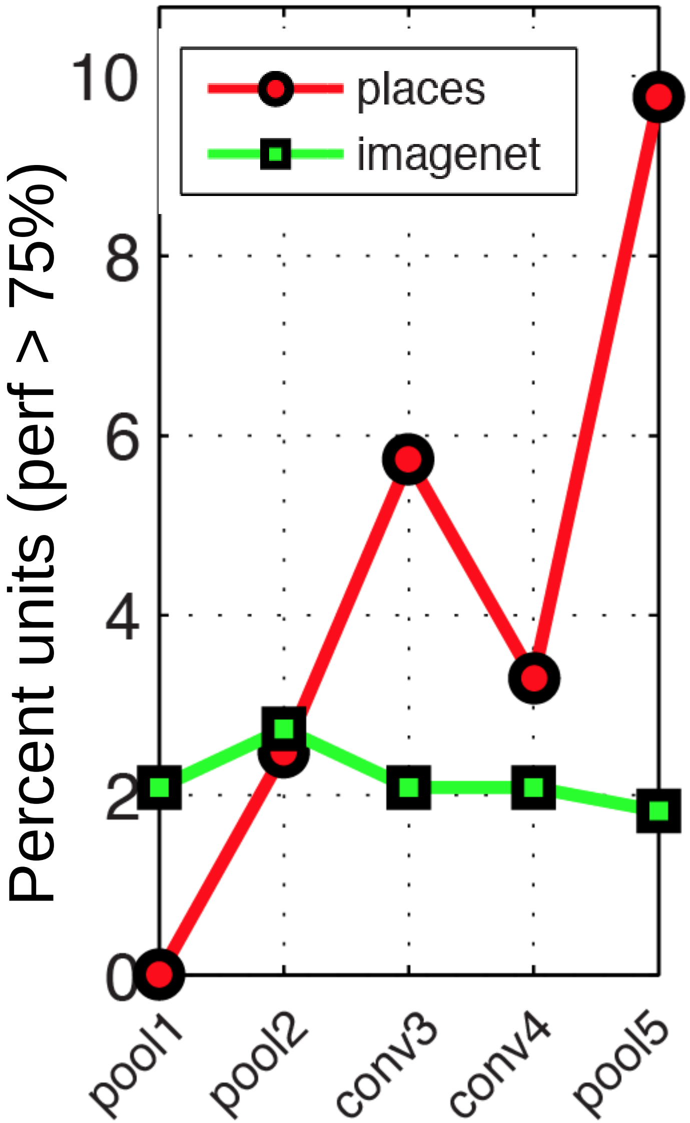
Percent units (perf > 75%)

Distribution of semantic types at each layer

6 - Scenes

Ex: kitchen, corridor, street, beach, ….





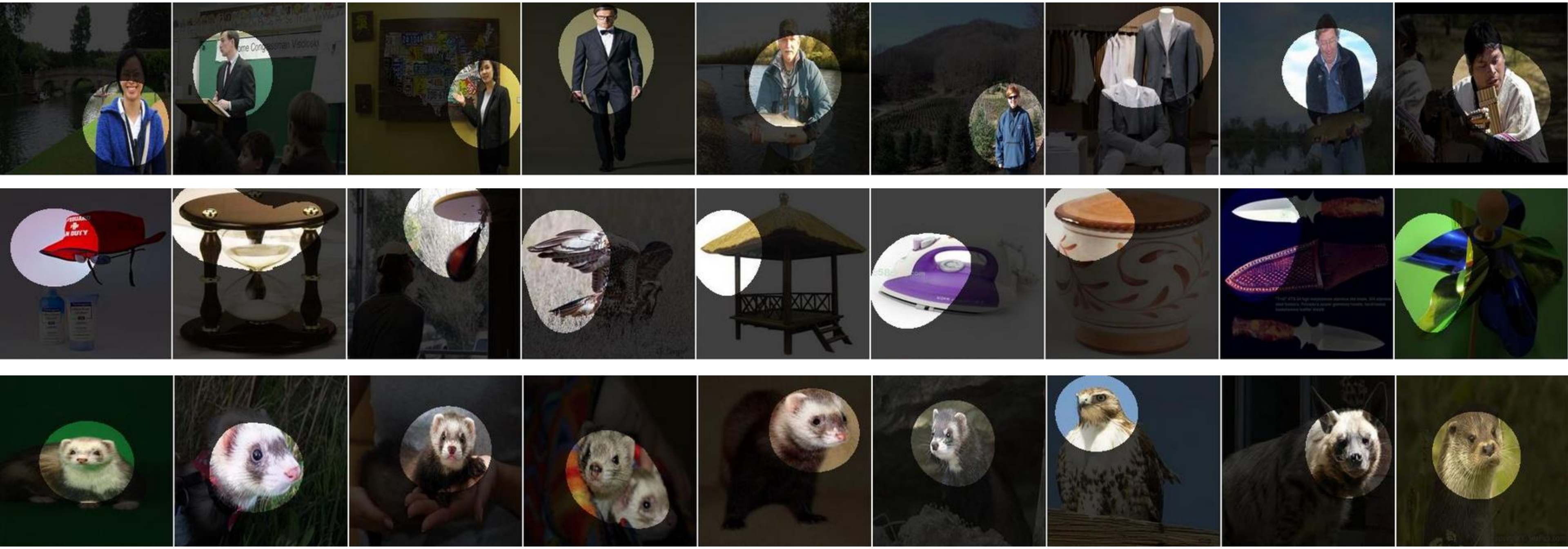
What objects are found?

ImageNet-CNN Units

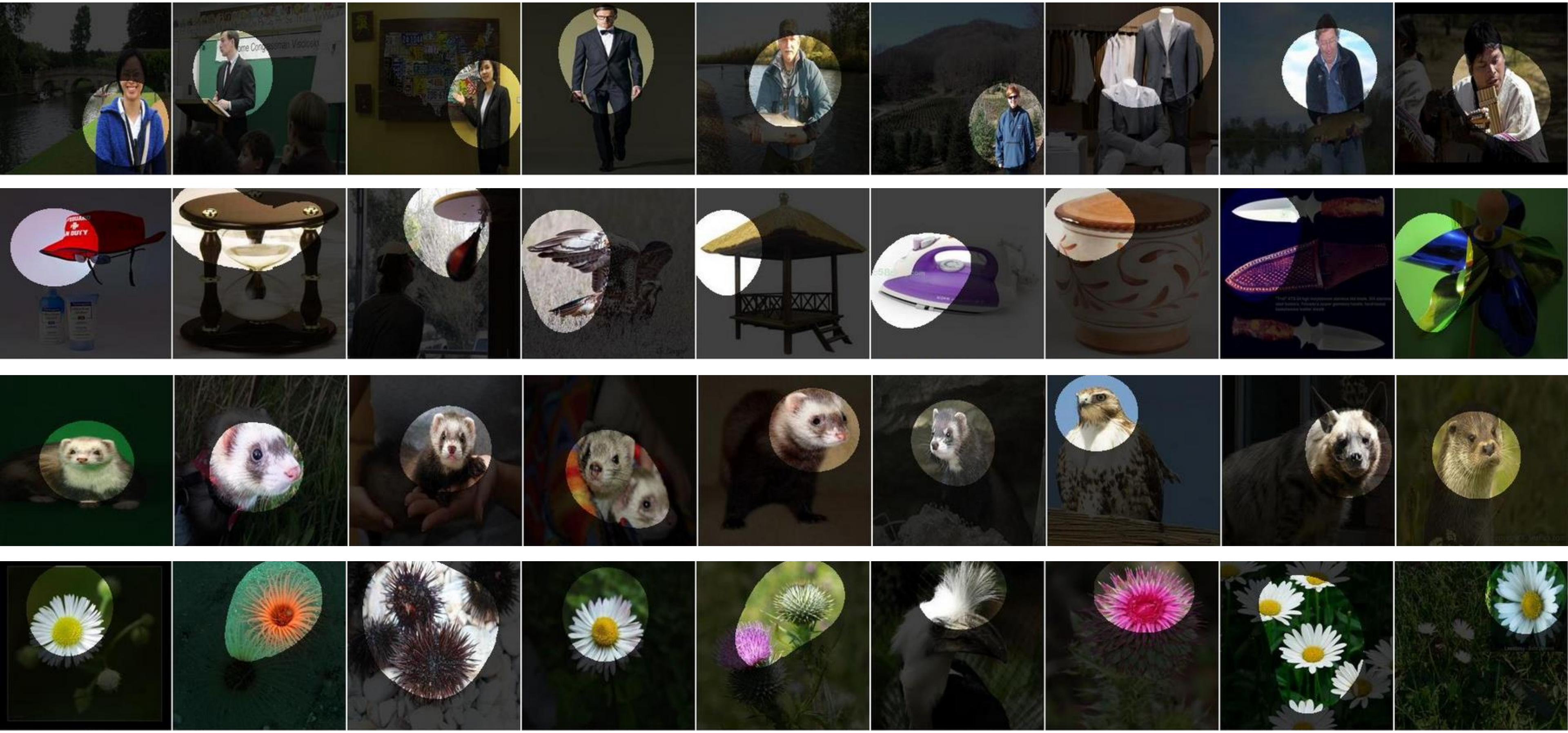
ImageNet-CNN Units

ImageNet-CNN Units

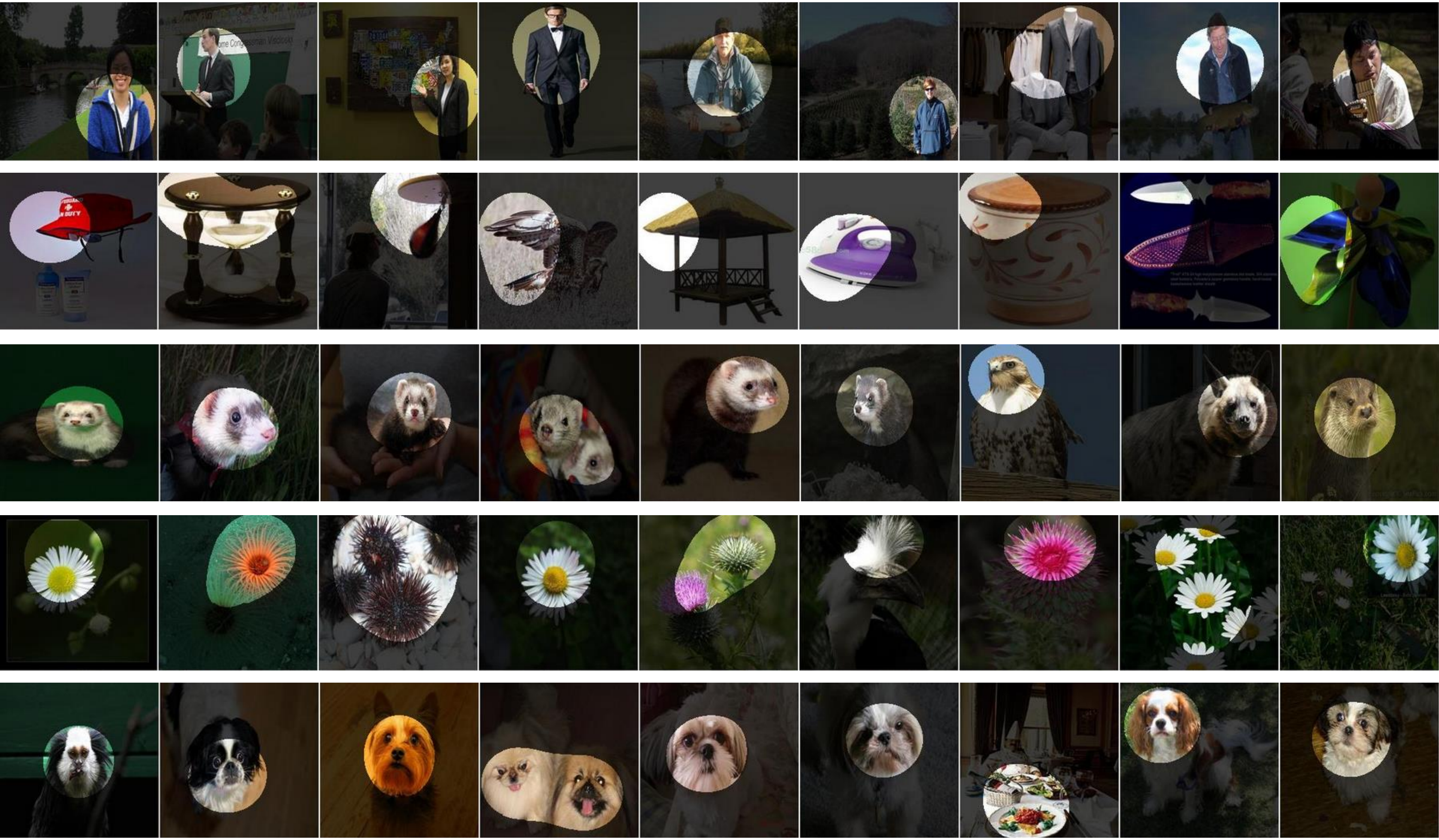
ImageNet-CNN Units



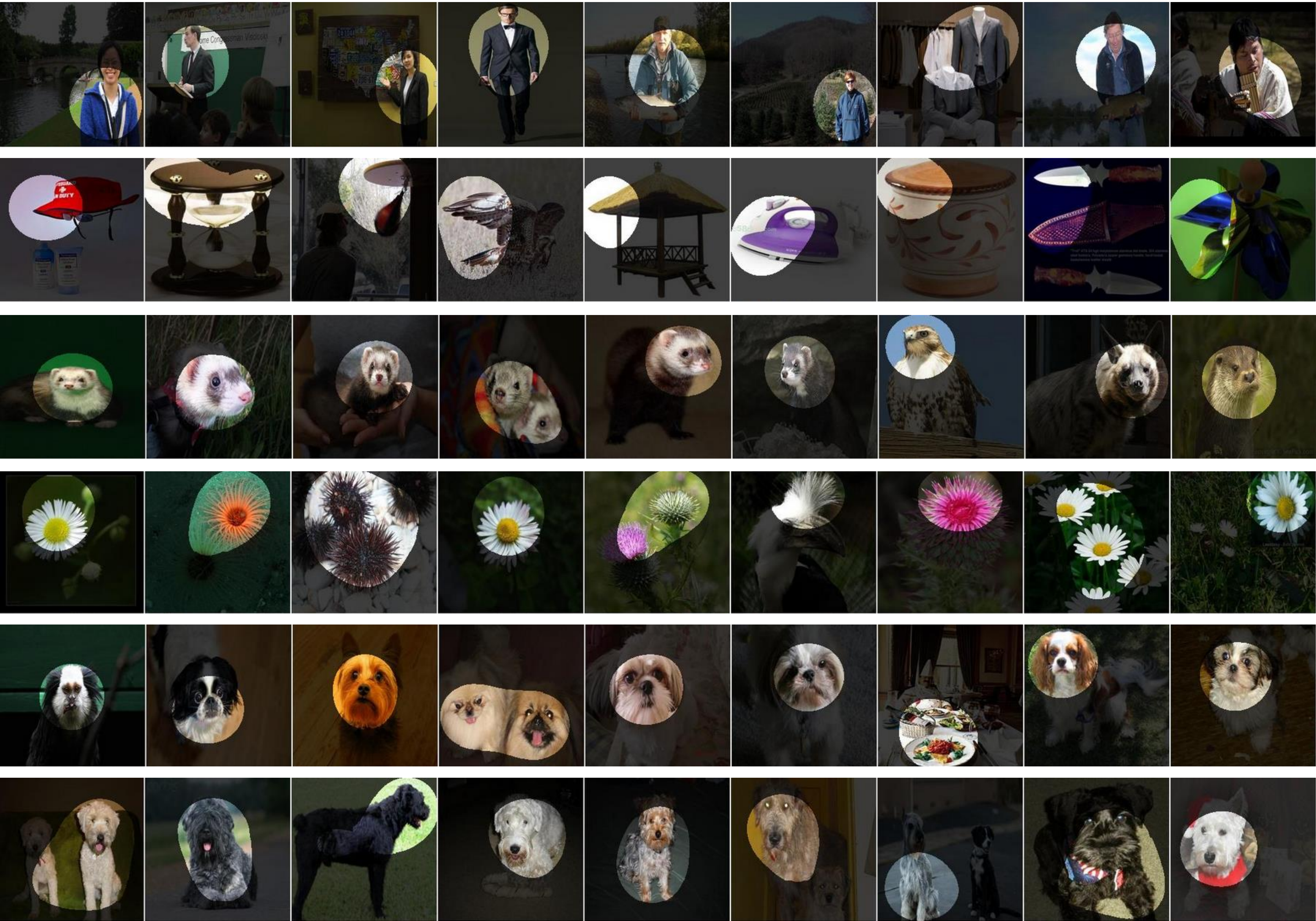
ImageNet-CNN Units



ImageNet-CNN Units



ImageNet-CNN Units



Places-CNN Units

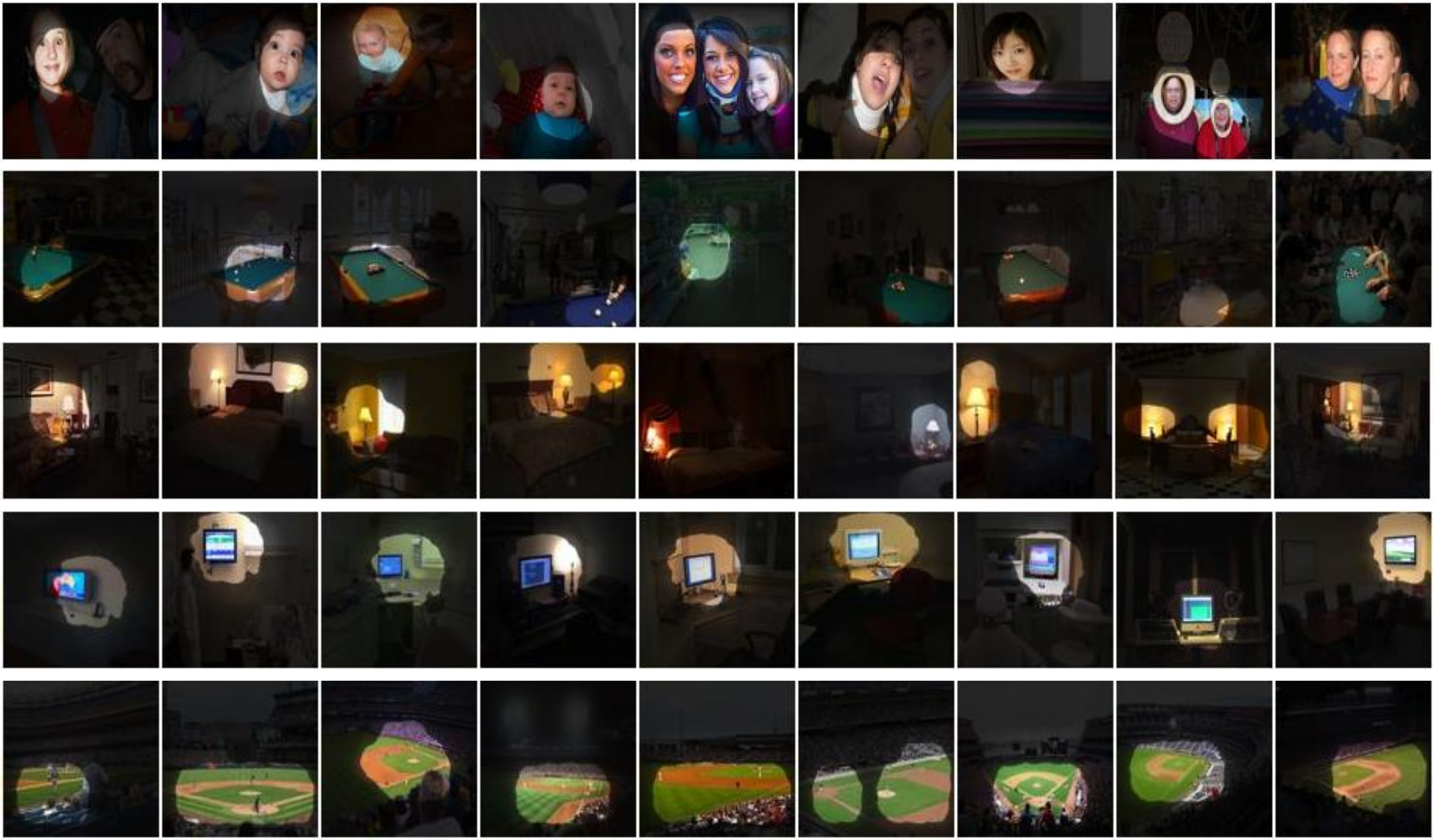
Places-CNN Units

Places-CNN Units

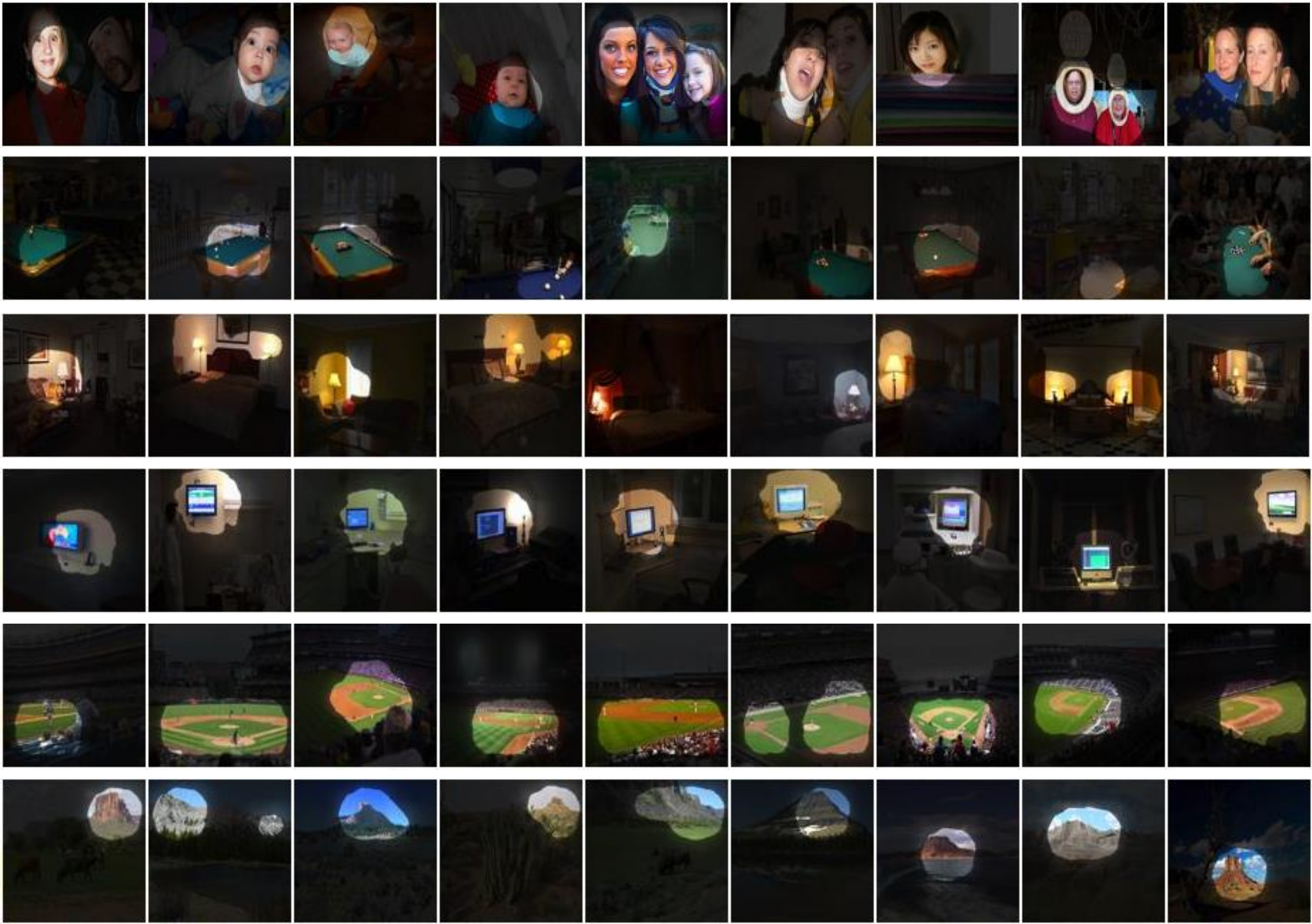
Places-CNN Units

Places-CNN Units

Places-CNN Units

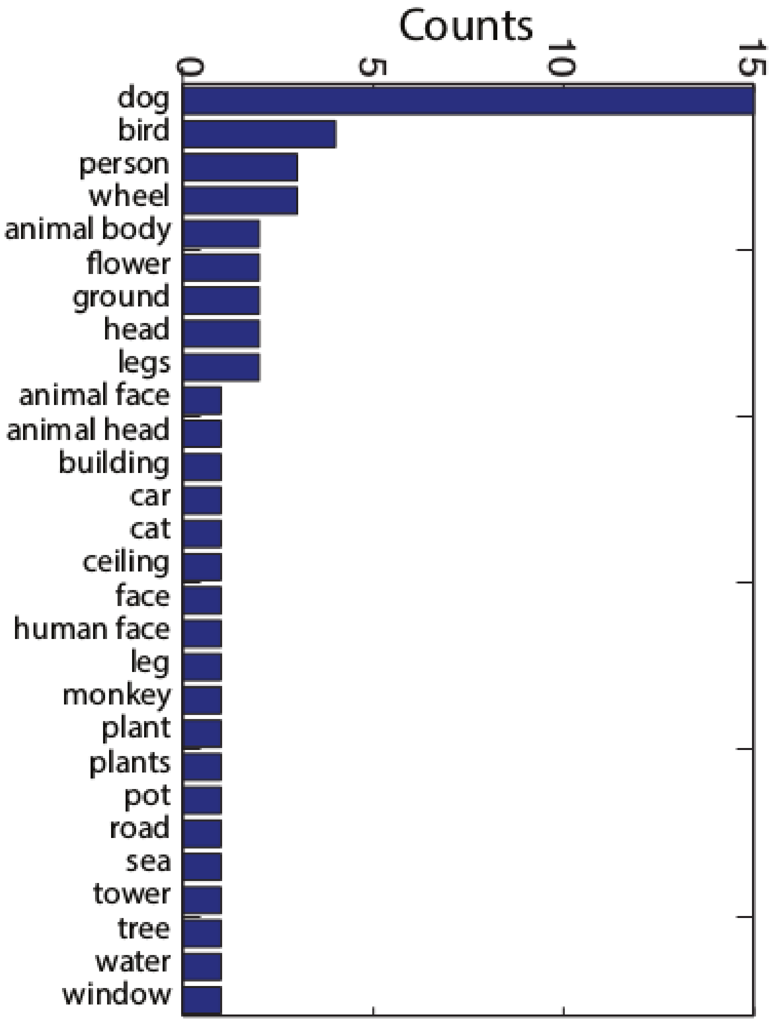


Places-CNN Units



Histogram of Emerged Objects in Pool5

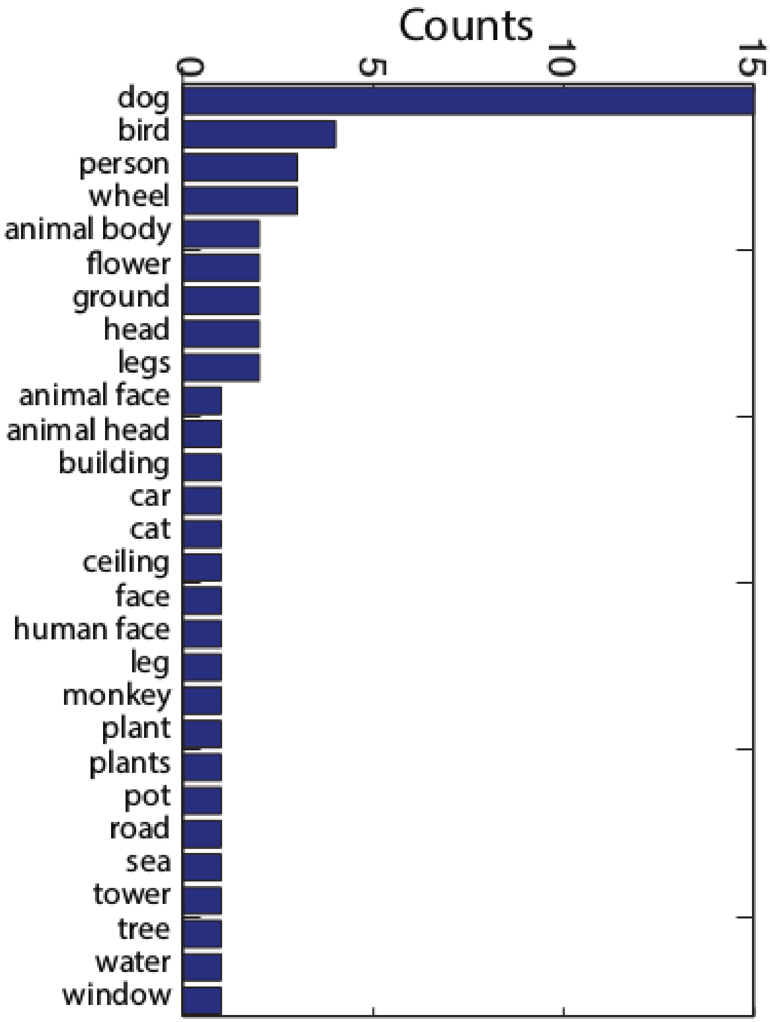
ImageNet-CNN (59/256)



Includes: Objects, nameable parts, and regions

Histogram of Emerged Objects in Pool5

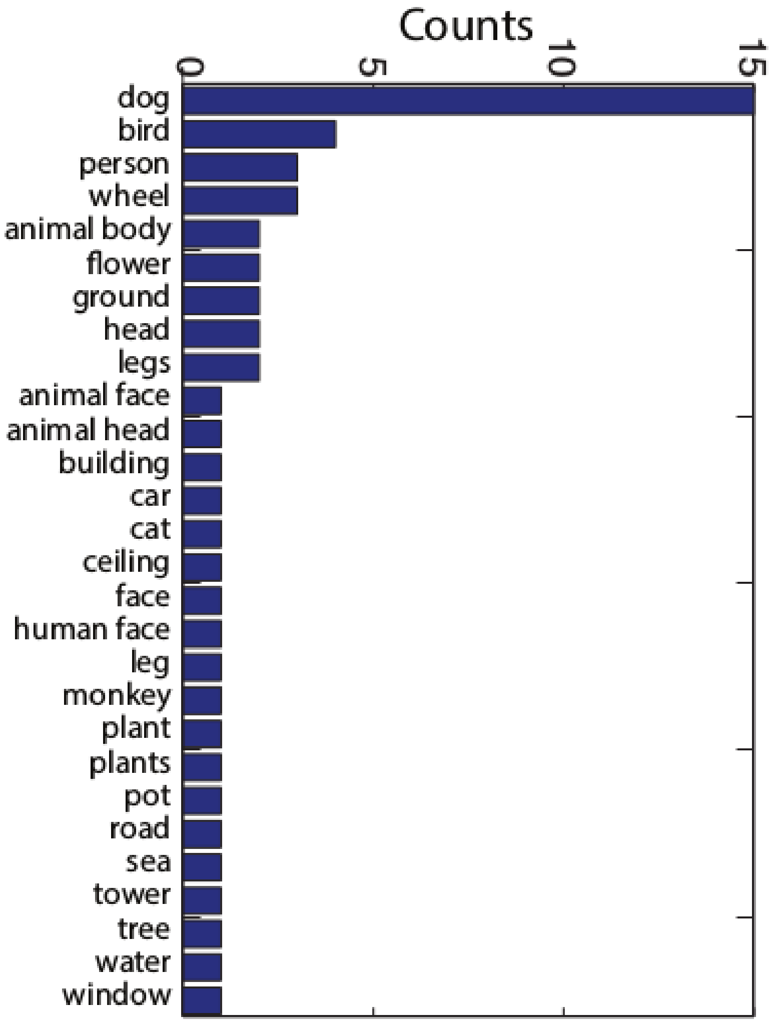
ImageNet-CNN (59/256)



Includes: Objects, nameable parts, and regions

Histogram of Emerged Objects in Pool5

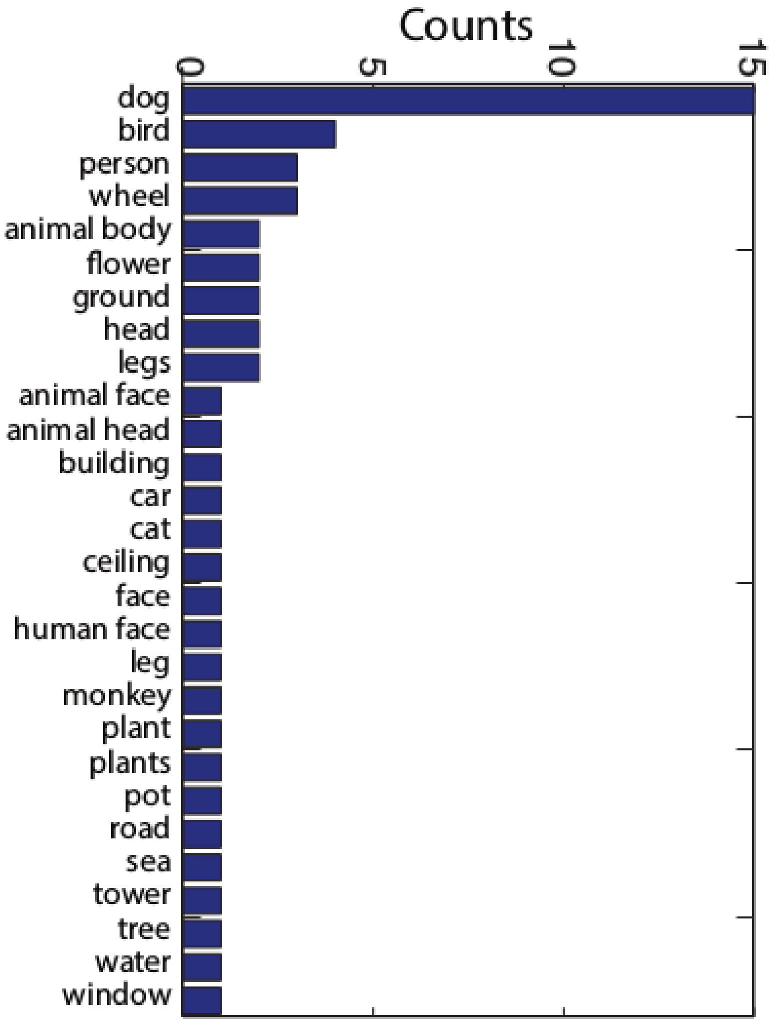
ImageNet-CNN (59/256)



Includes: Objects, nameable parts, and regions

Histogram of Emerged Objects in Pool5

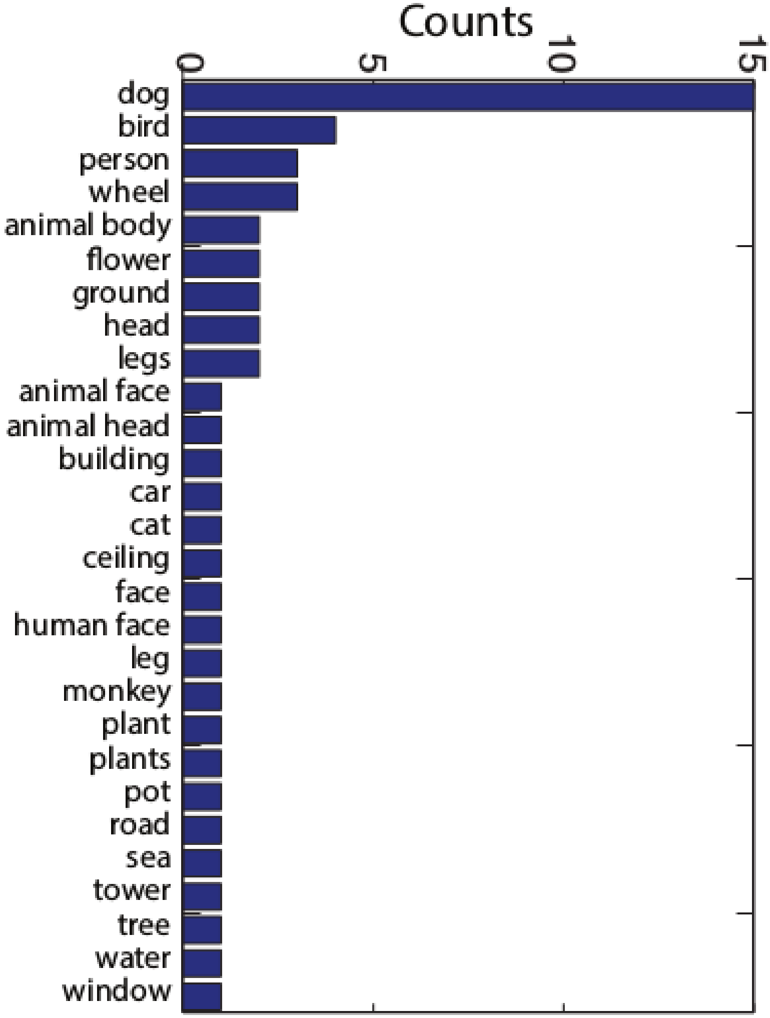
ImageNet-CNN (59/256)



Includes: Objects, nameable parts, and regions

Histogram of Emerged Objects in Pool5

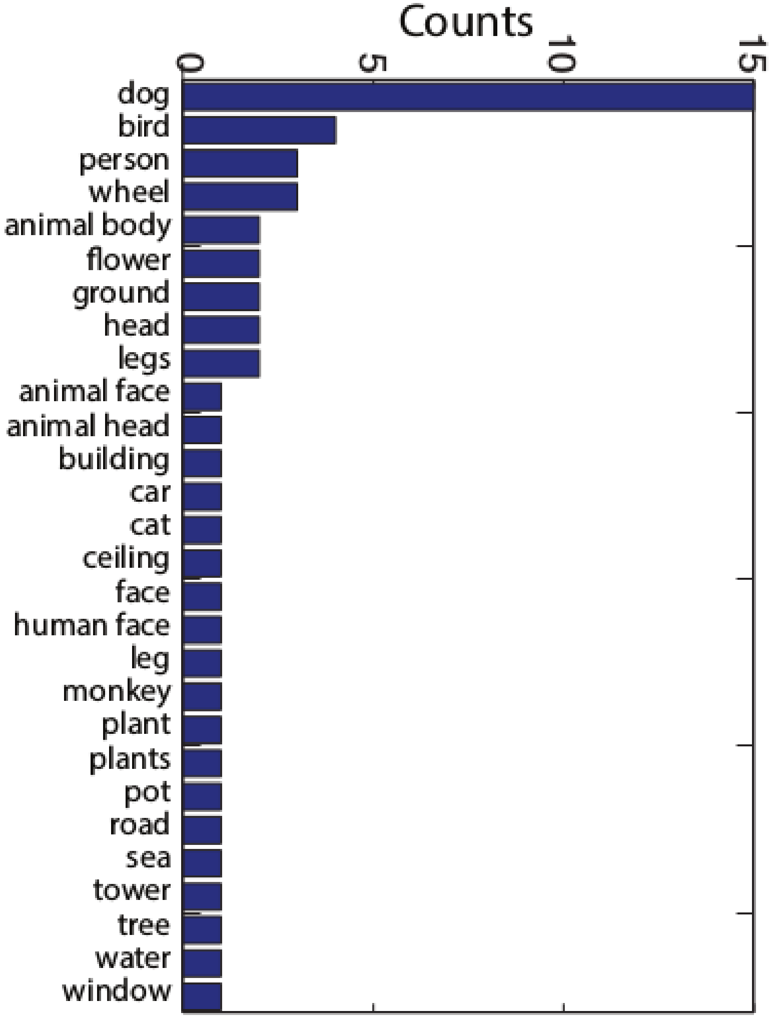
ImageNet-CNN (59/256)



Includes: Objects, nameable parts, and regions

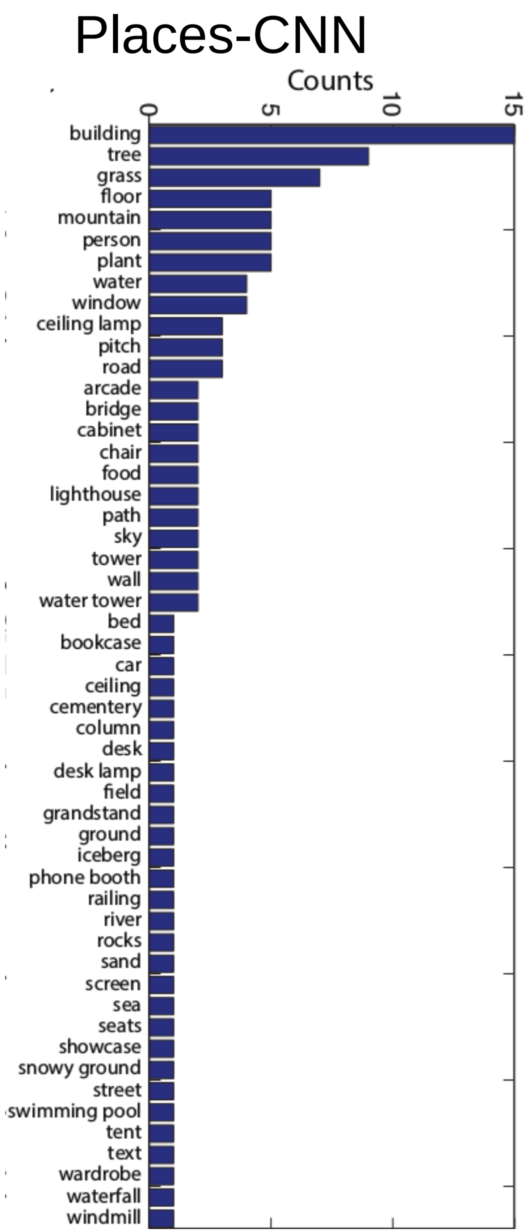
Histogram of Emerged Objects in Pool5

ImageNet-CNN (59/256)



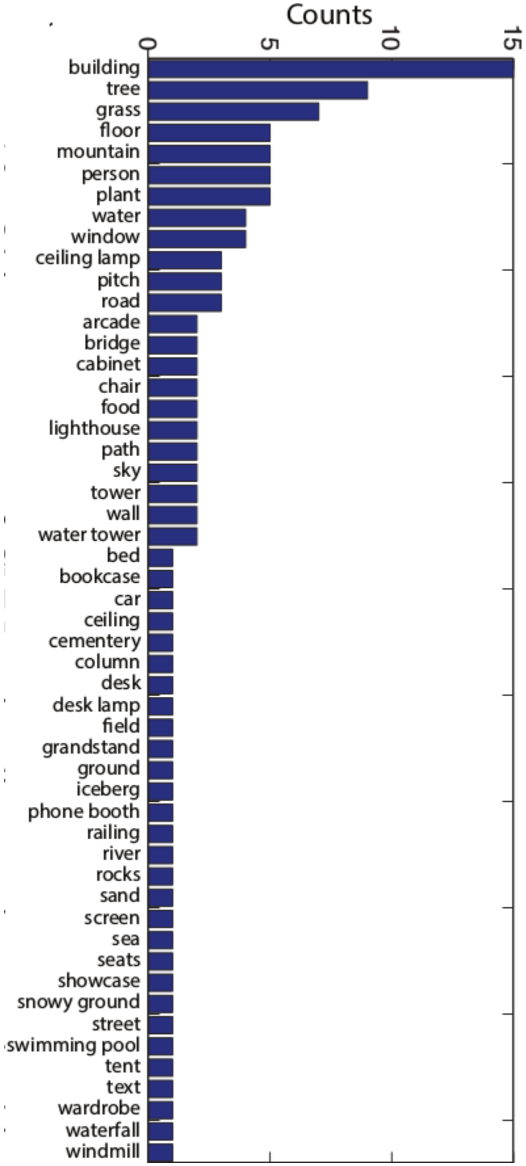
Includes: Objects, nameable parts, and regions

Histogram of Emerged Objects in Pool5



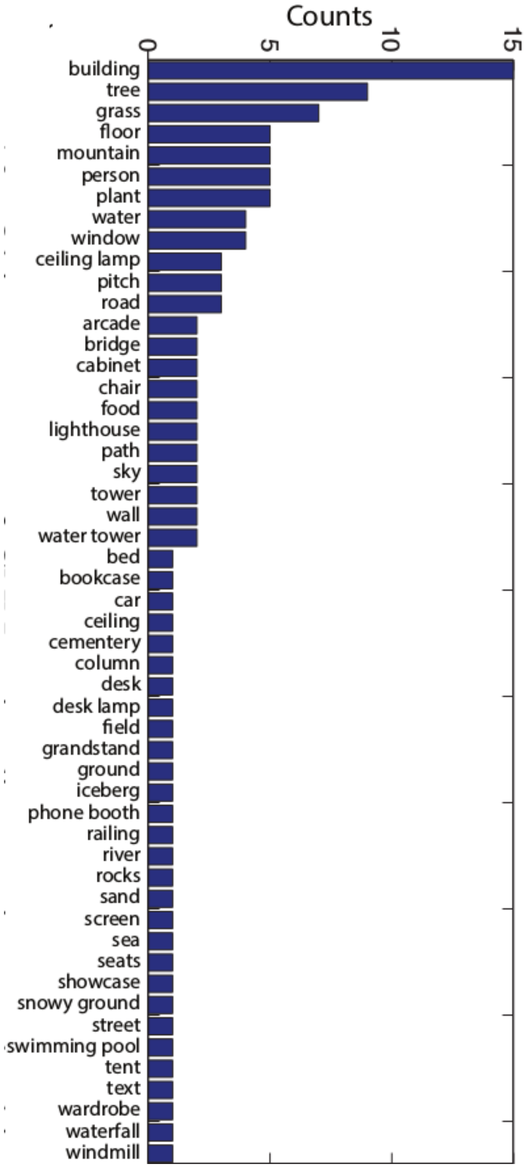
Histogram of Emerged Objects in Pool5

Places-CNN



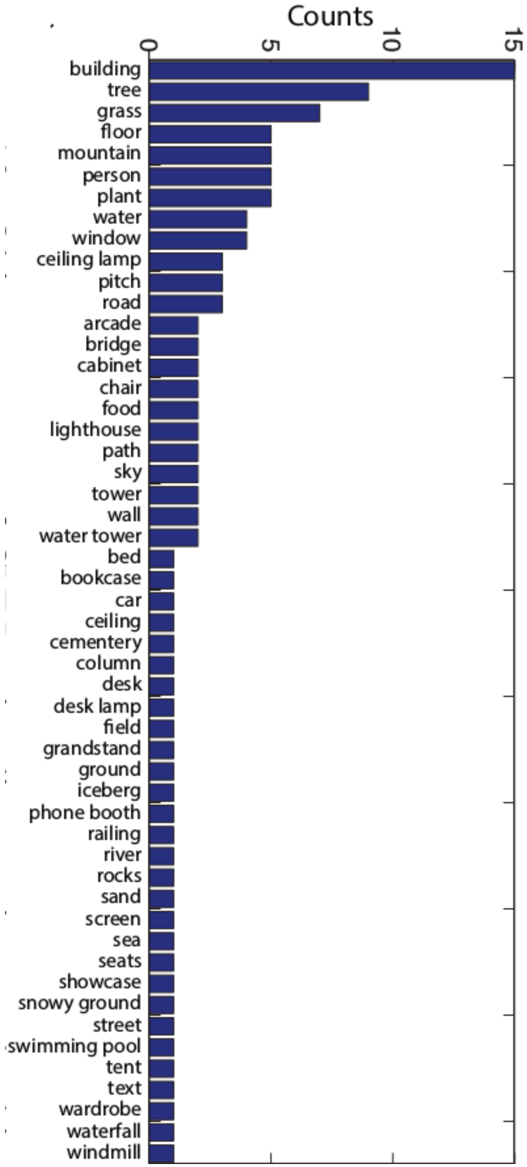
Histogram of Emerged Objects in Pool5

Places-CNN



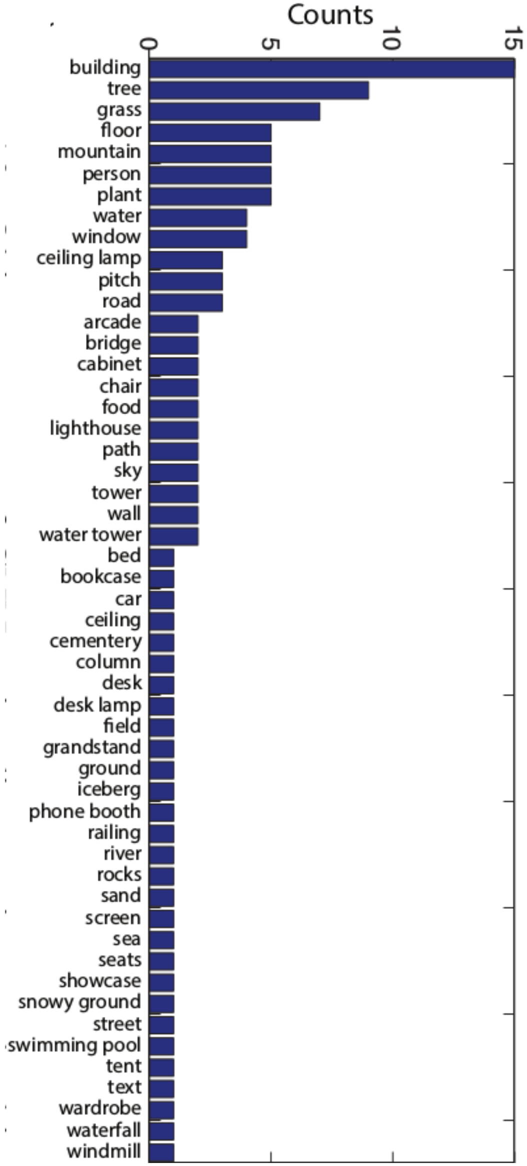
Histogram of Emerged Objects in Pool5

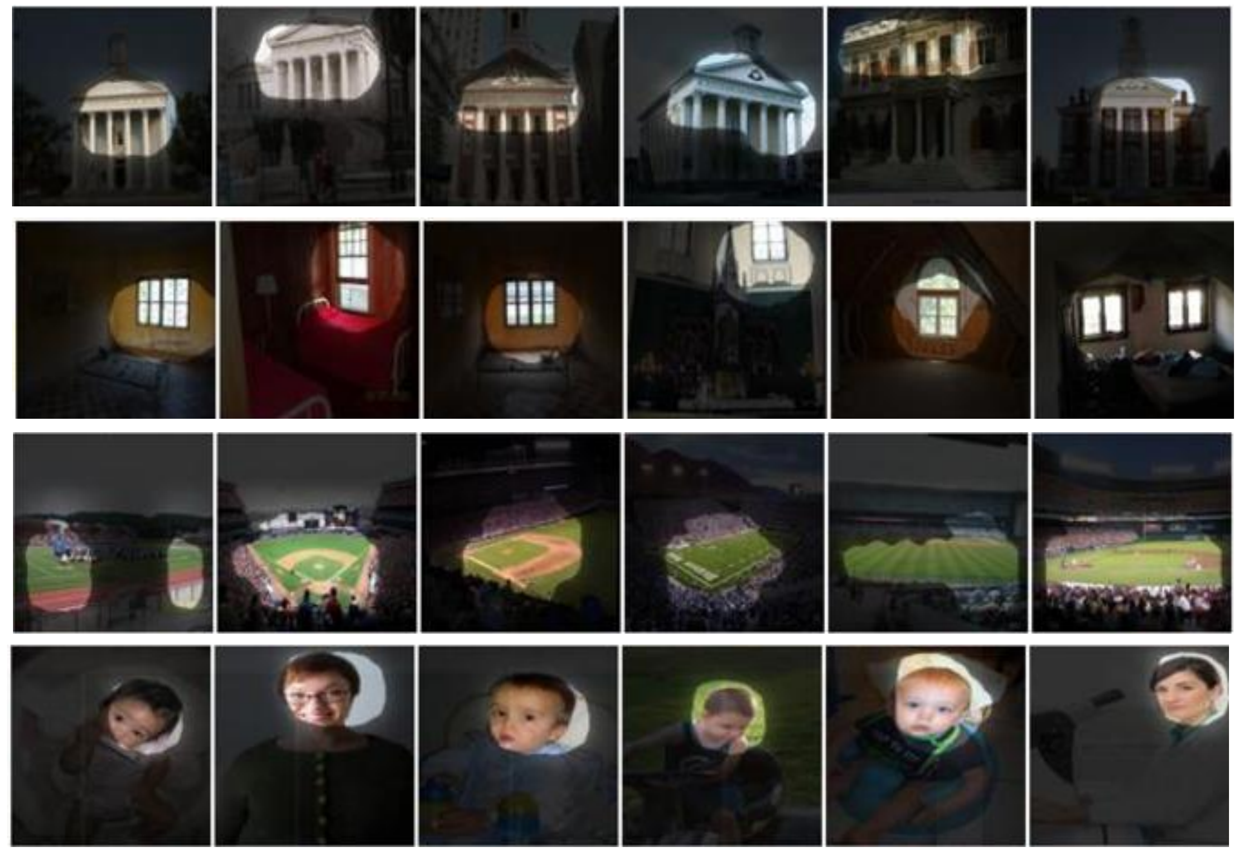
Places-CNN



Histogram of Emerged Objects in Pool5

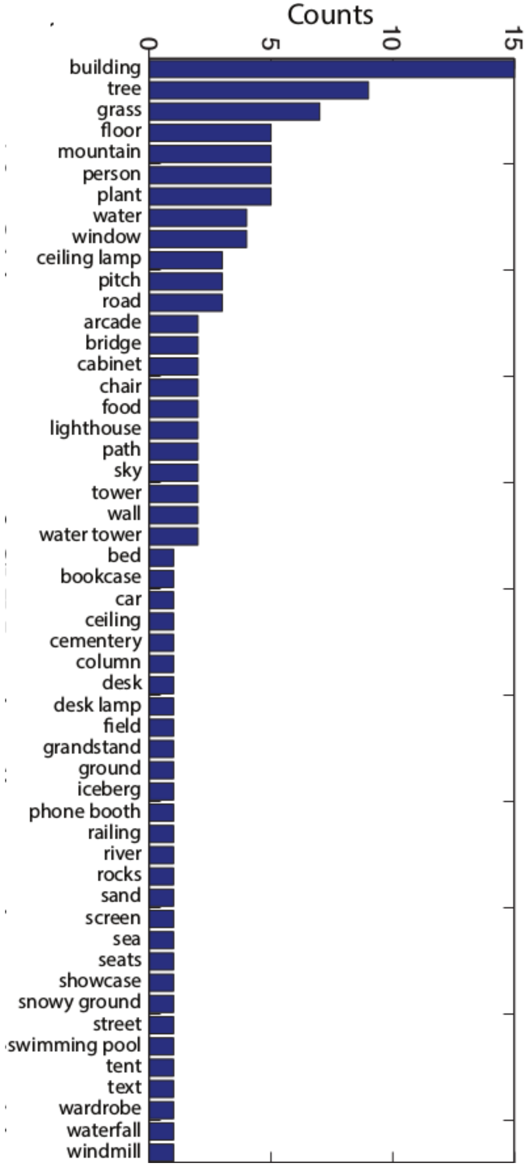
Places-CNN

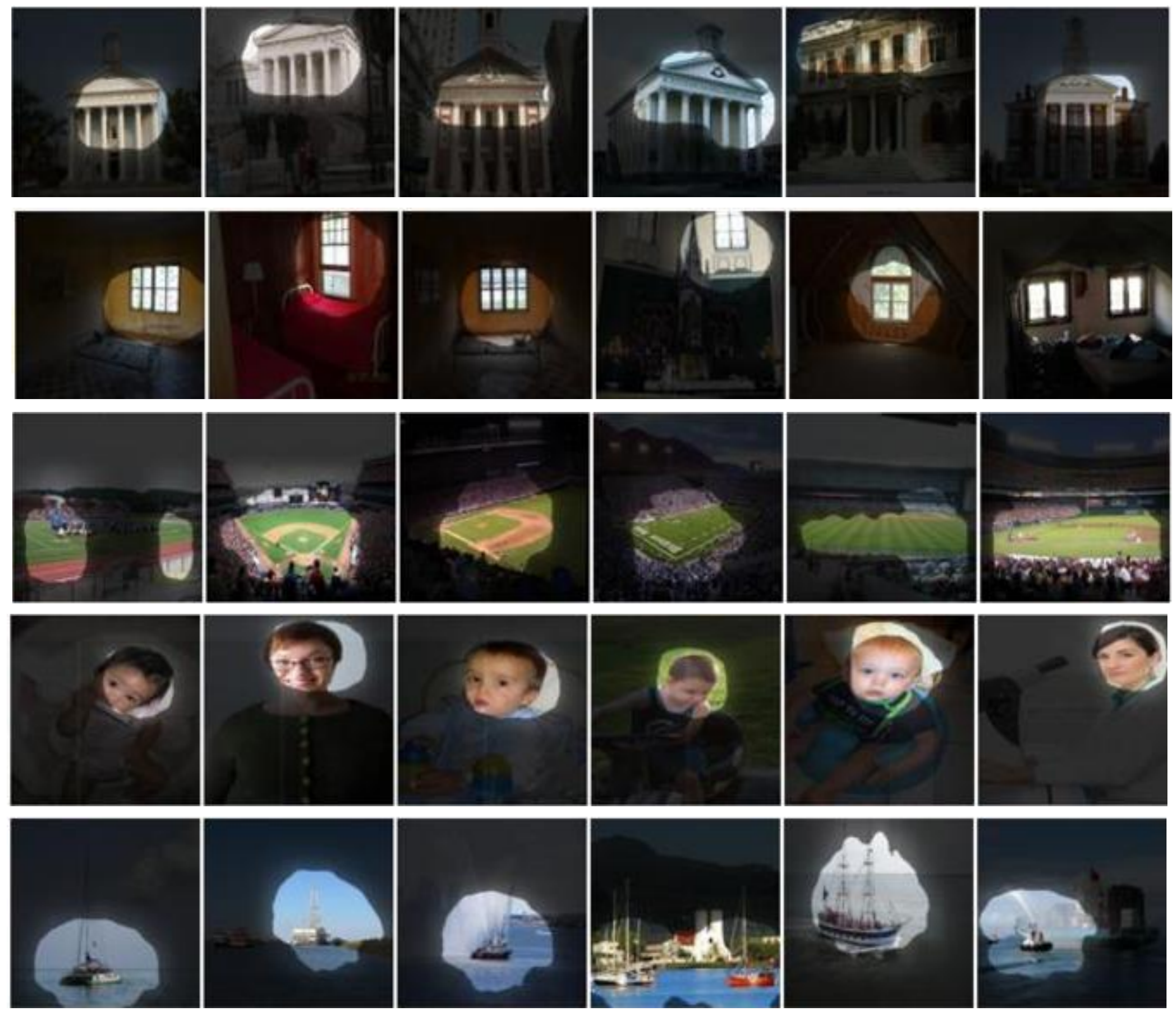




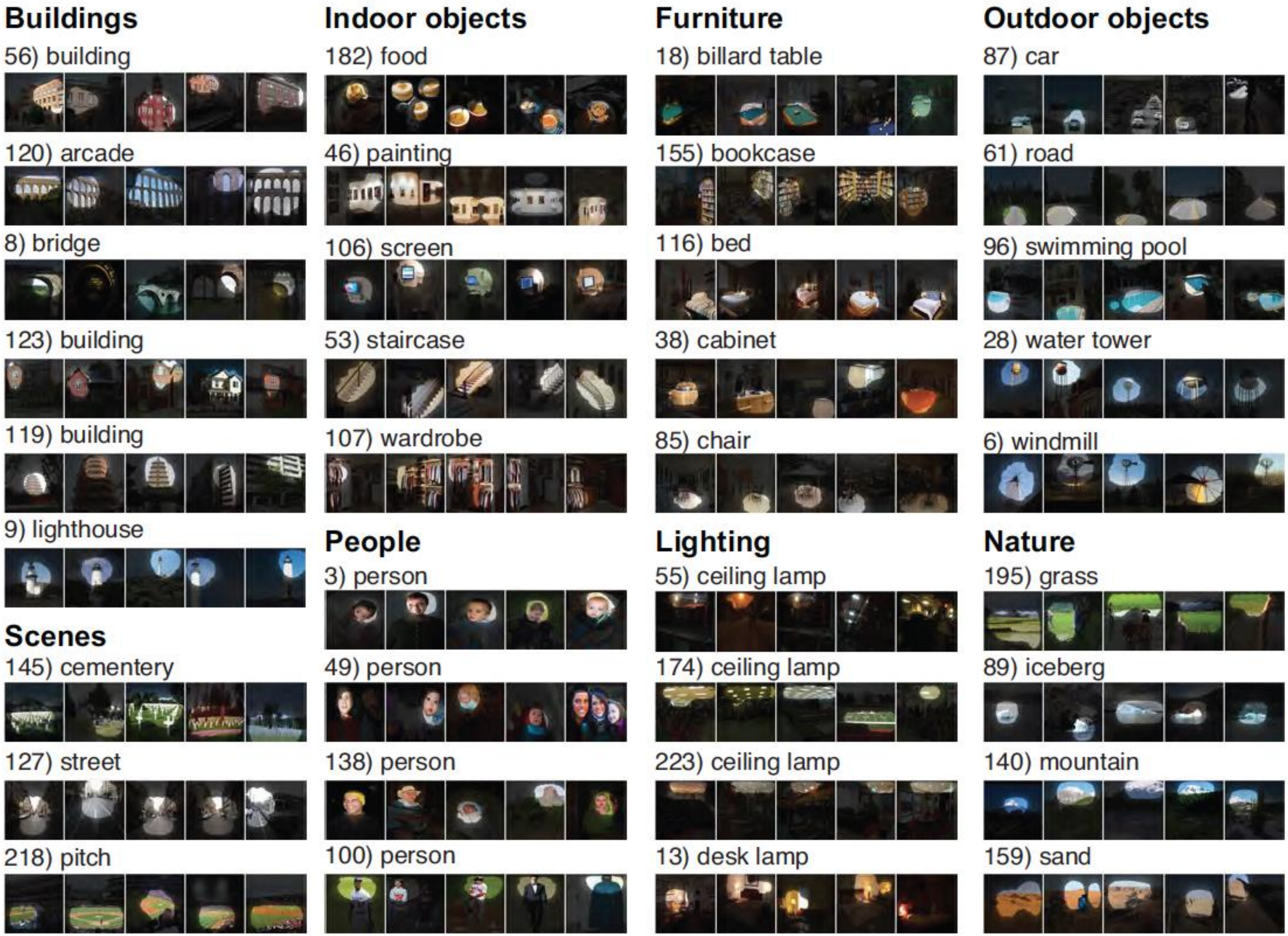
Histogram of Emerged Objects in Pool5

Places-CNN

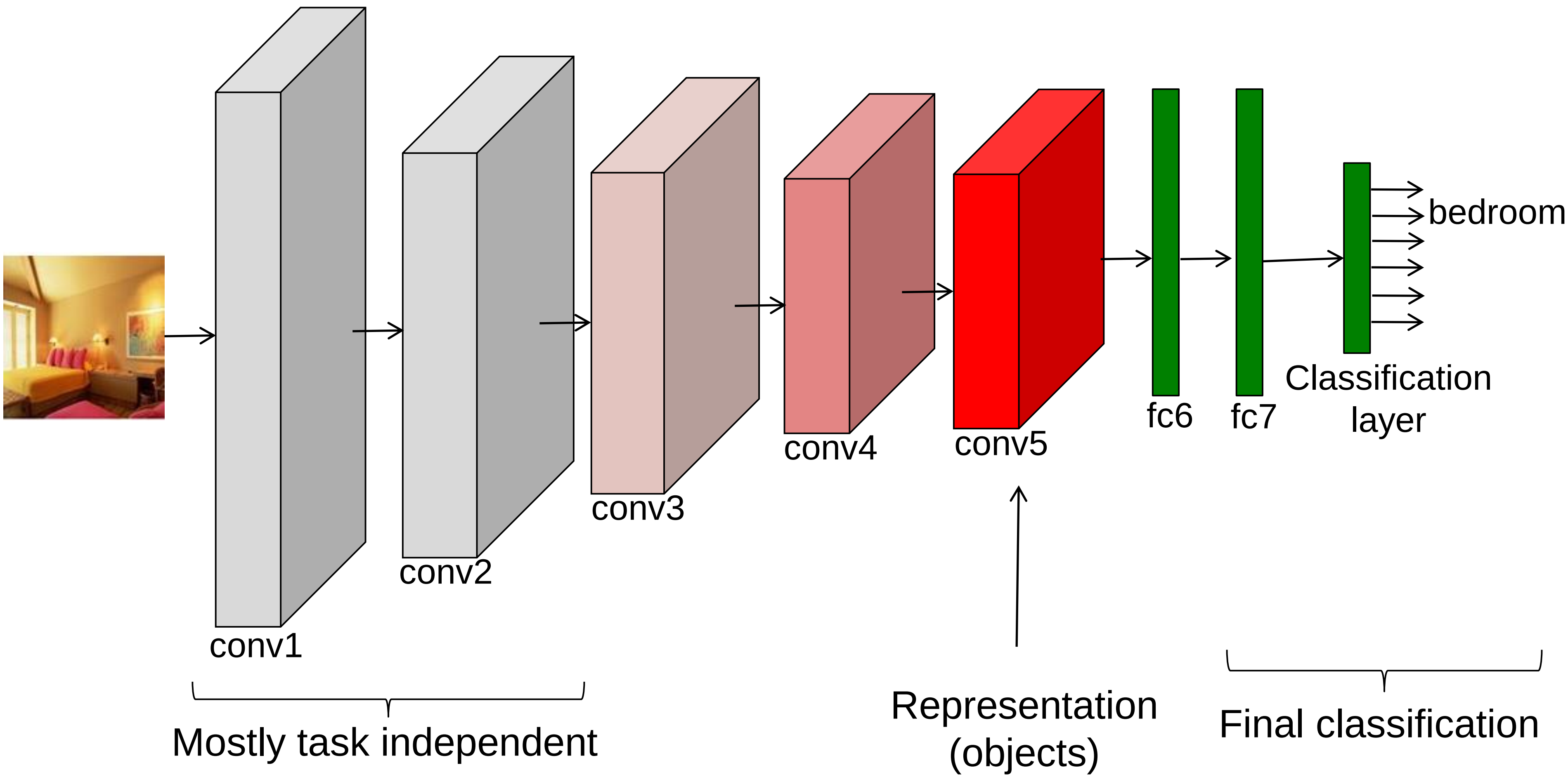




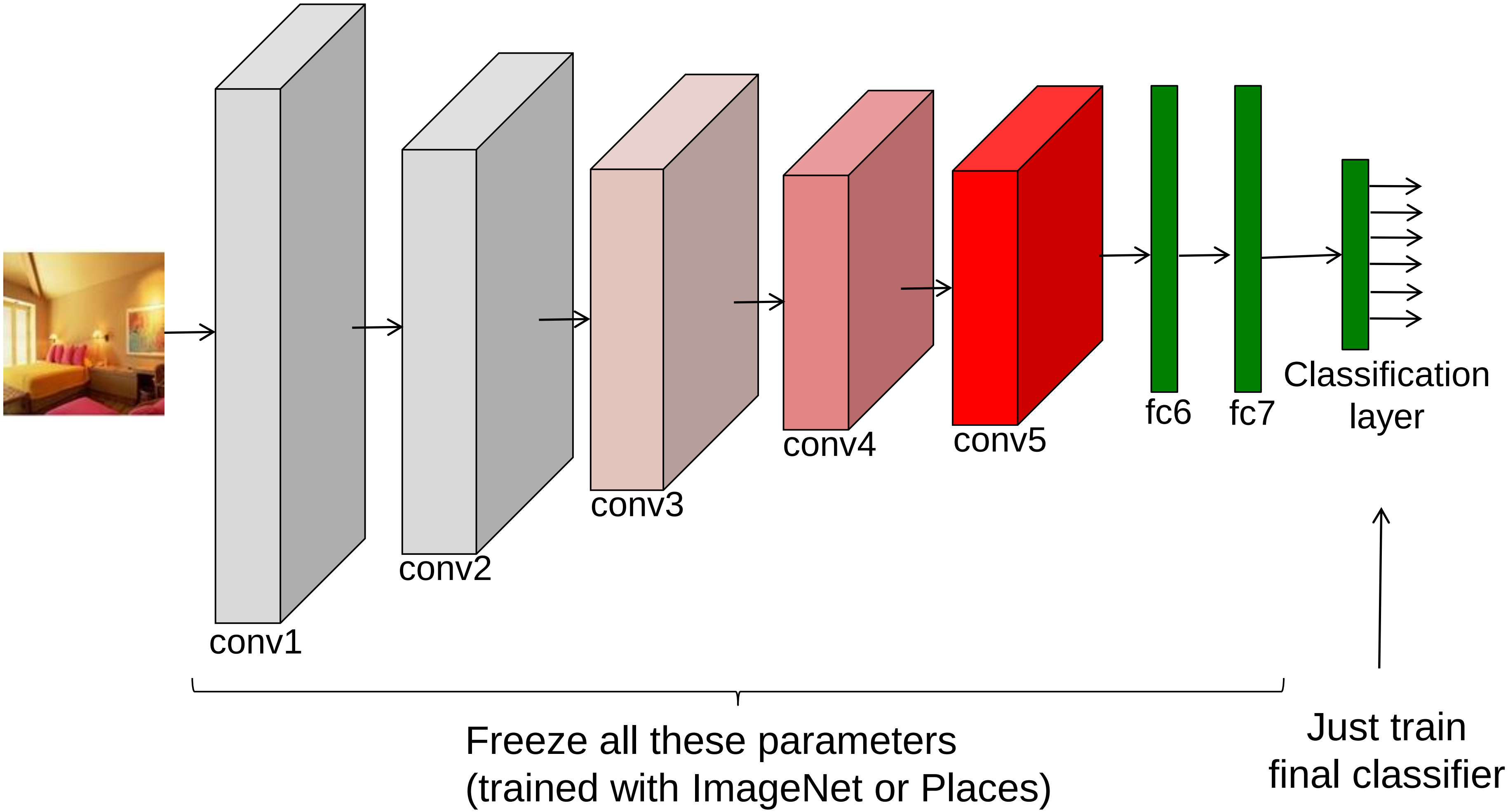
Object detectors emerge inside the CNN



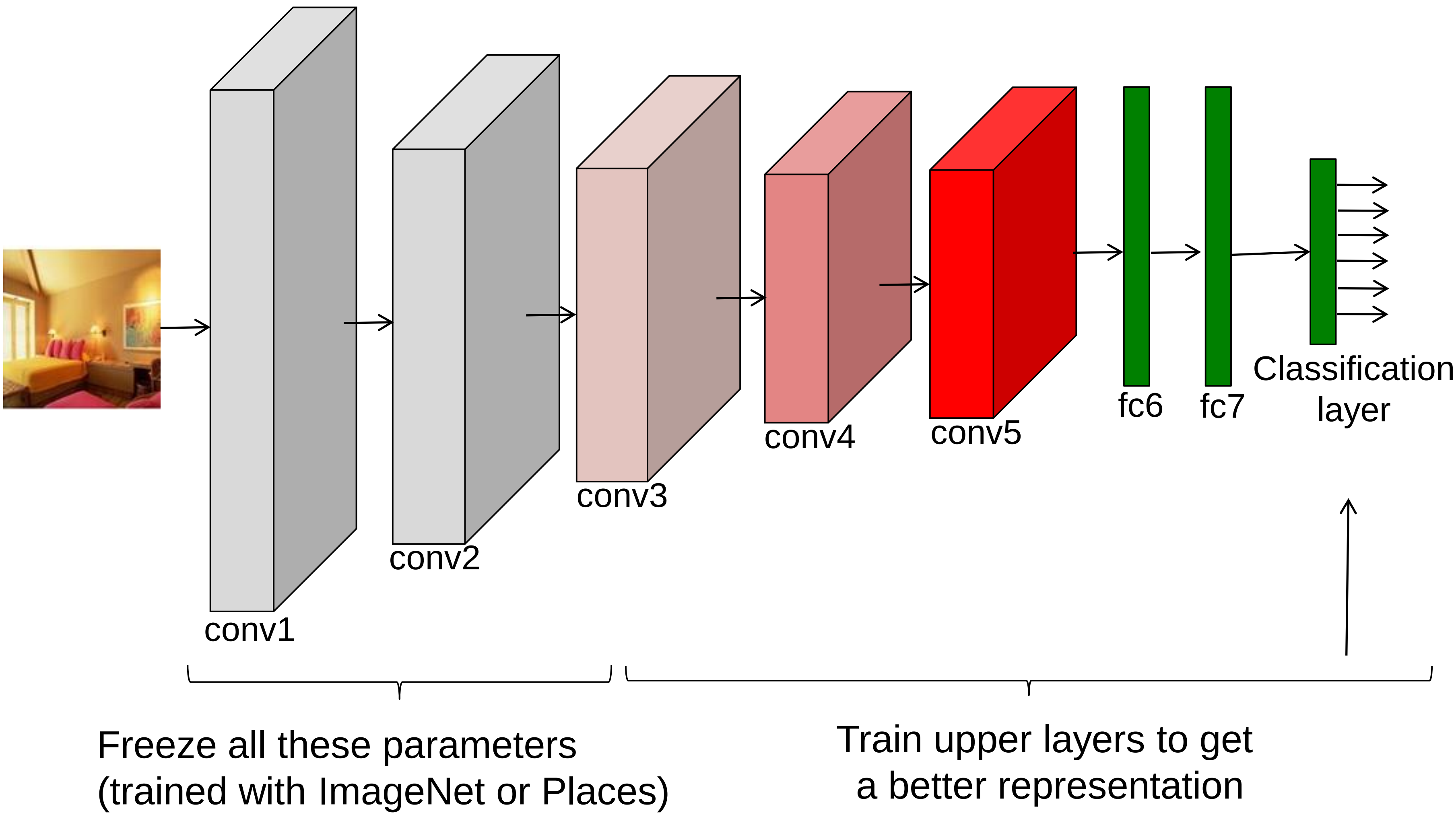
Nguyen et al, 2016


bedroom
conv1
conv2
conv3
conv4
conv5
fc6
fc7
Classification layer
Mostly task independent
Representation (objects)
Final classification

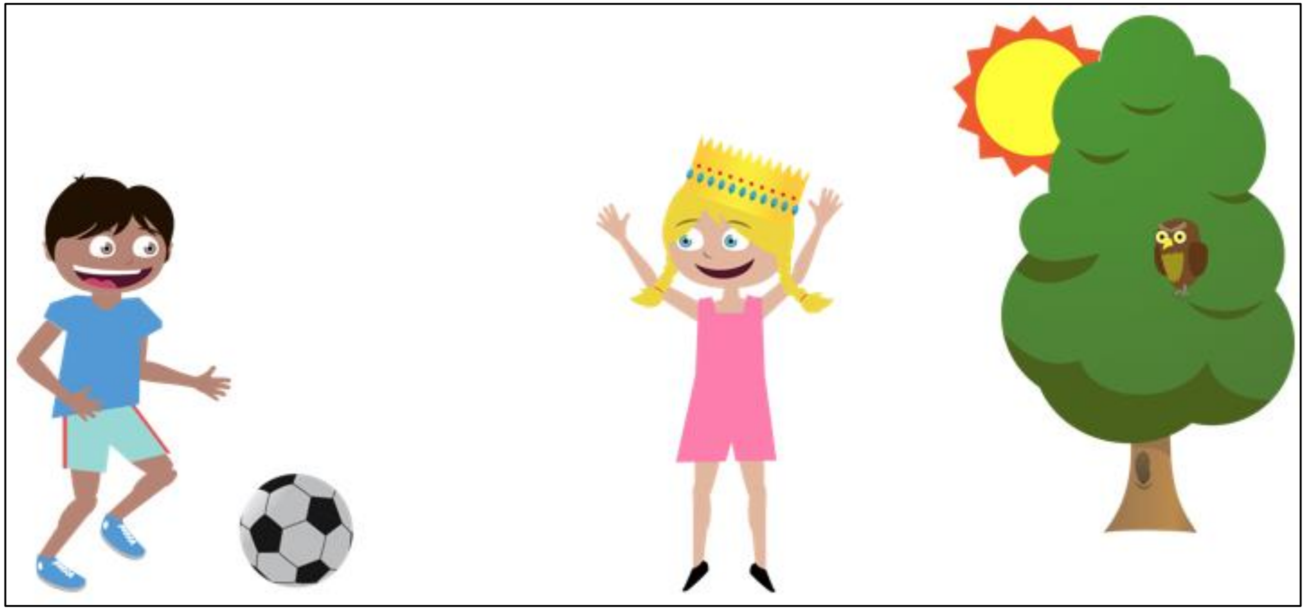
Strategies for training for new tasks



Strategies for training for new tasks

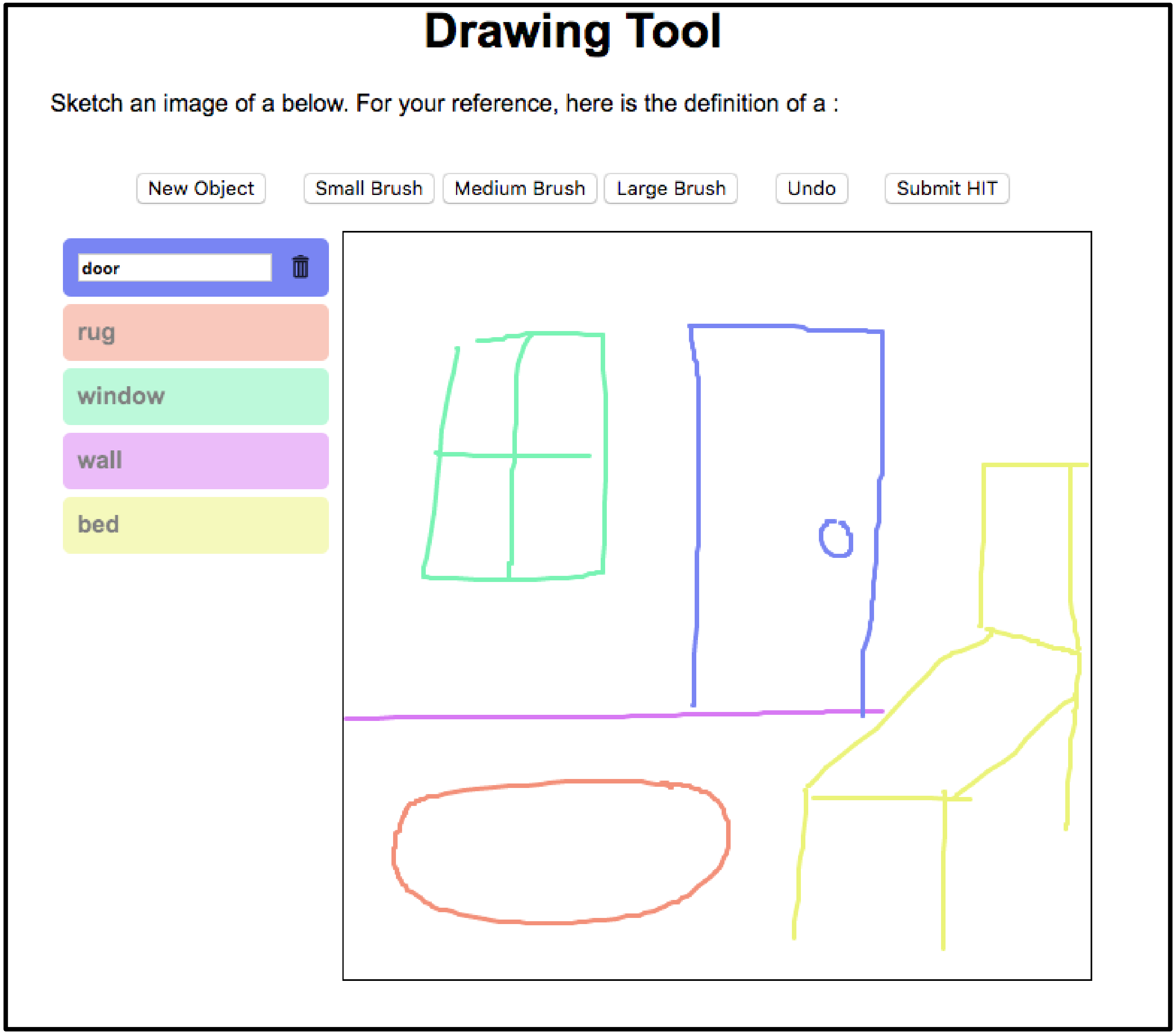


But what if you keep the task but change the input modality?

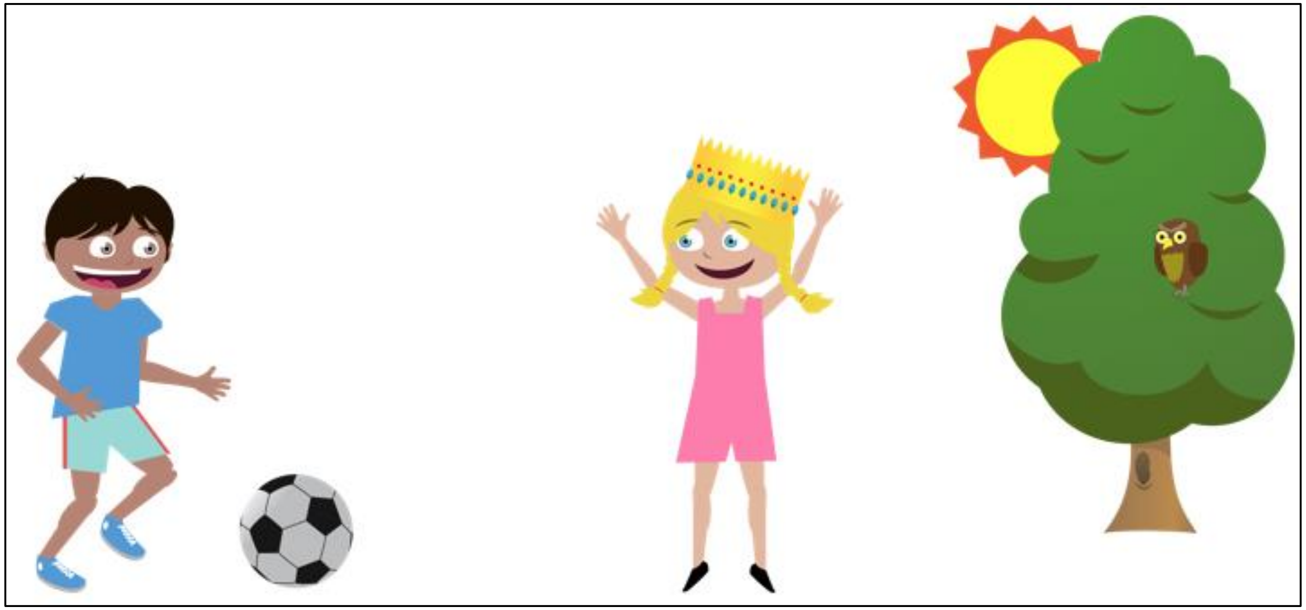


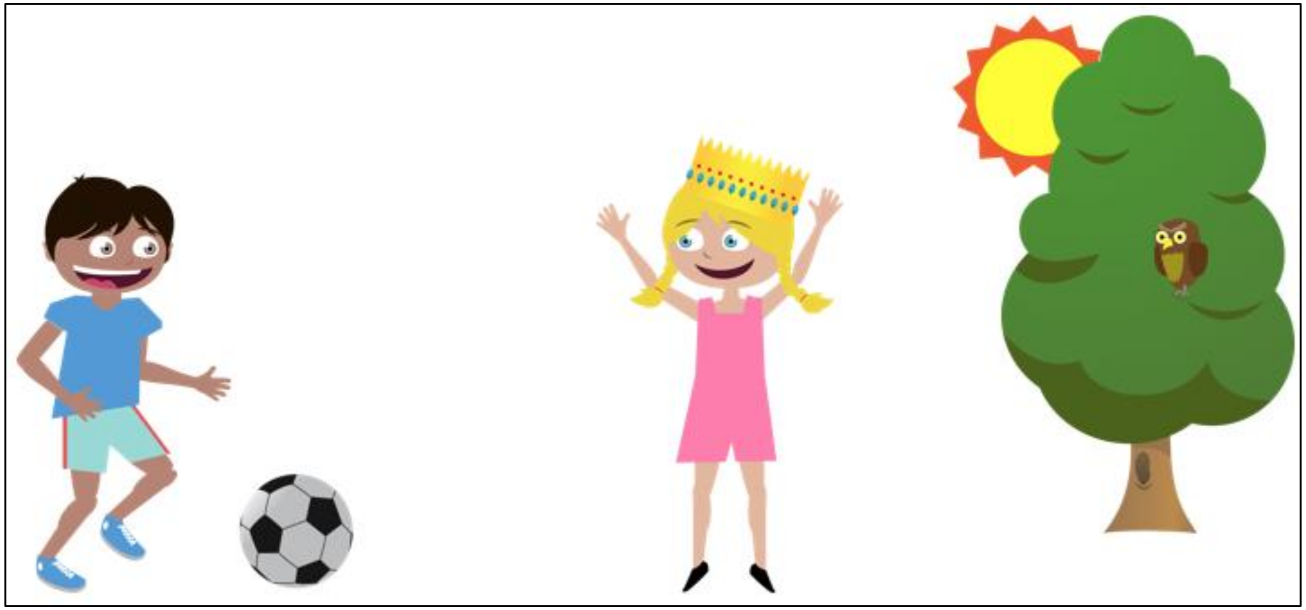
From Devi's webpage: "Abstract images provide several advantages. They allow for the direct study of how to infer high-level semantic information, since they remove the reliance on noisy low-level object, attribute and relation detectors, or the tedious hand-labeling of images."

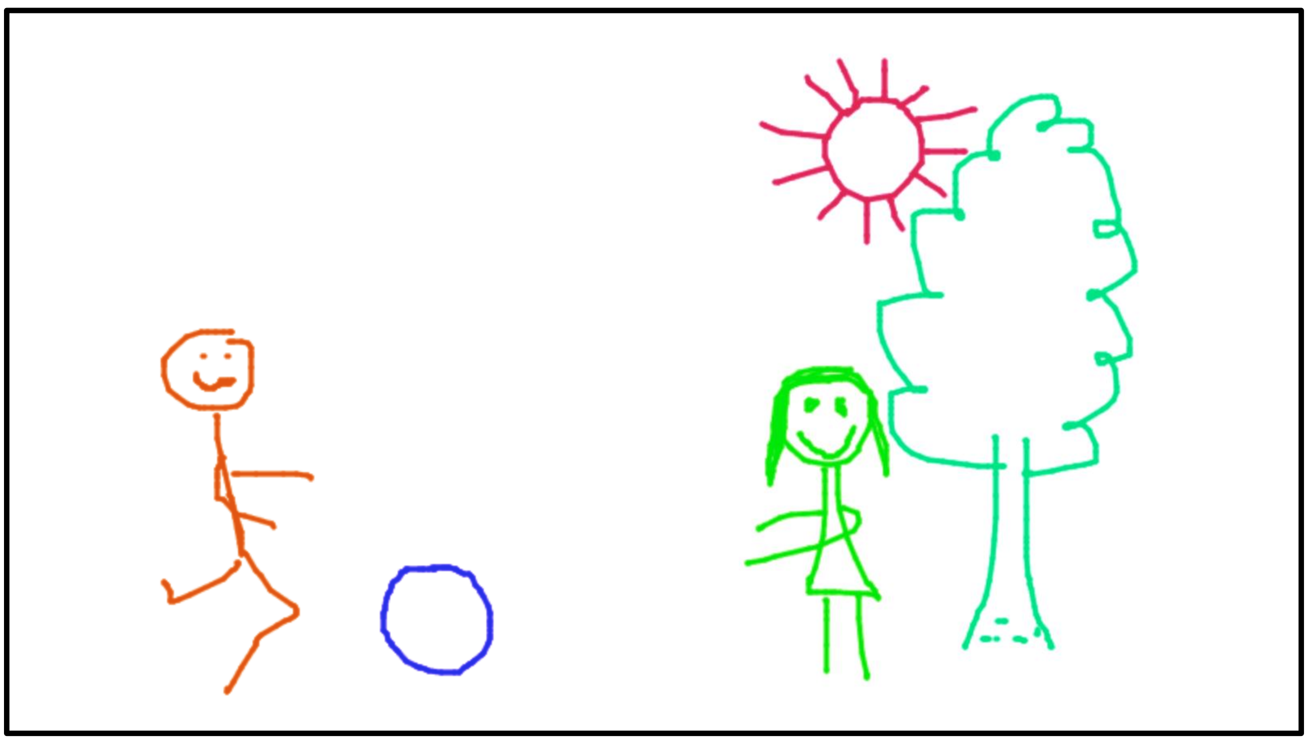
Bringing Semantics Into Focus Using Visual Abstraction (CVPR), 2013. Zitnick and Parikh.
Learning the Visual Interpretation of Sentences (ICCV), 2013. Zitnick, Parikh, and Vanderwende
Adopting Abstract Images for Semantic Scene Understanding (PAMI), 2015. Zitnick, Vedantam and Parikh



Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016







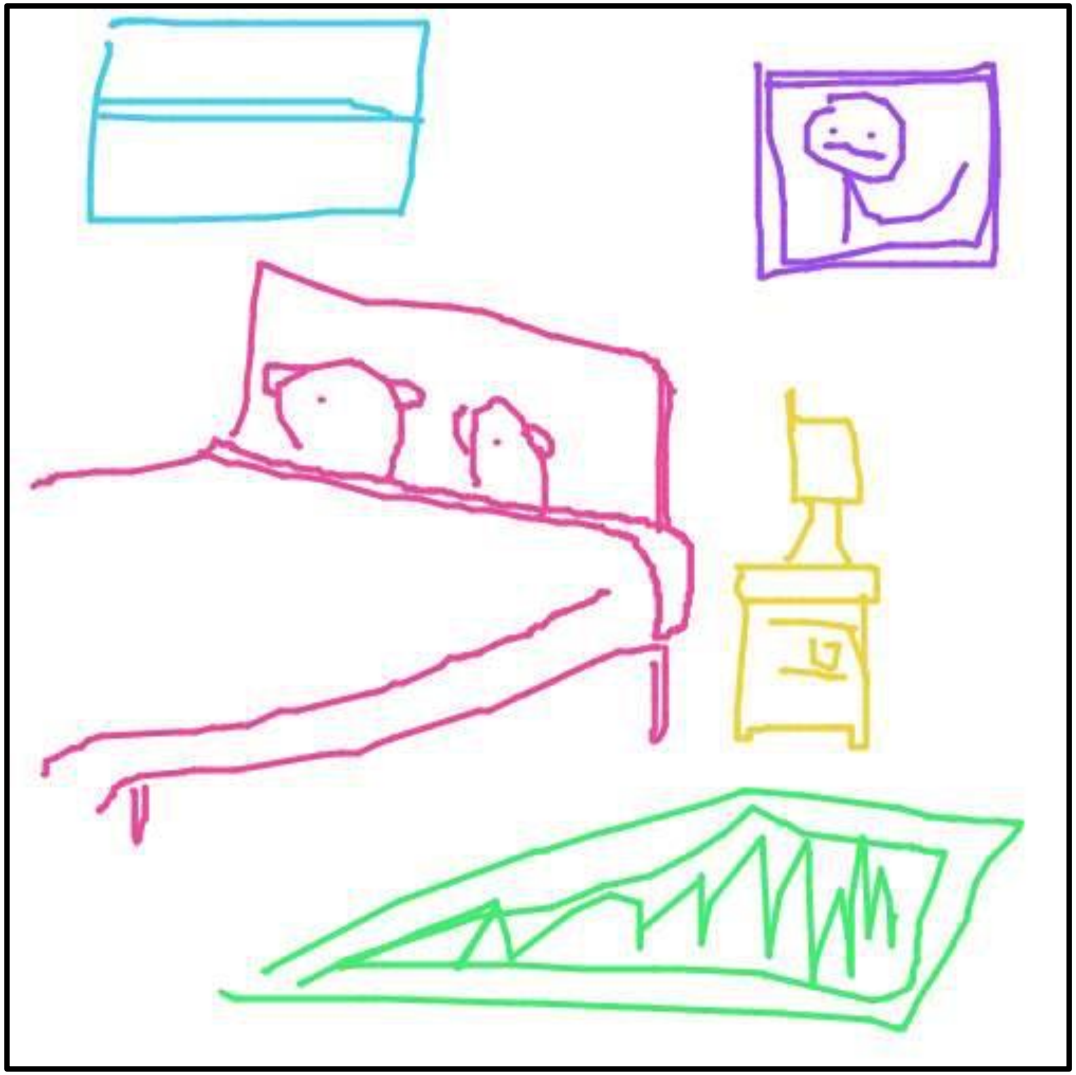
From crowdsourcing

Line drawings

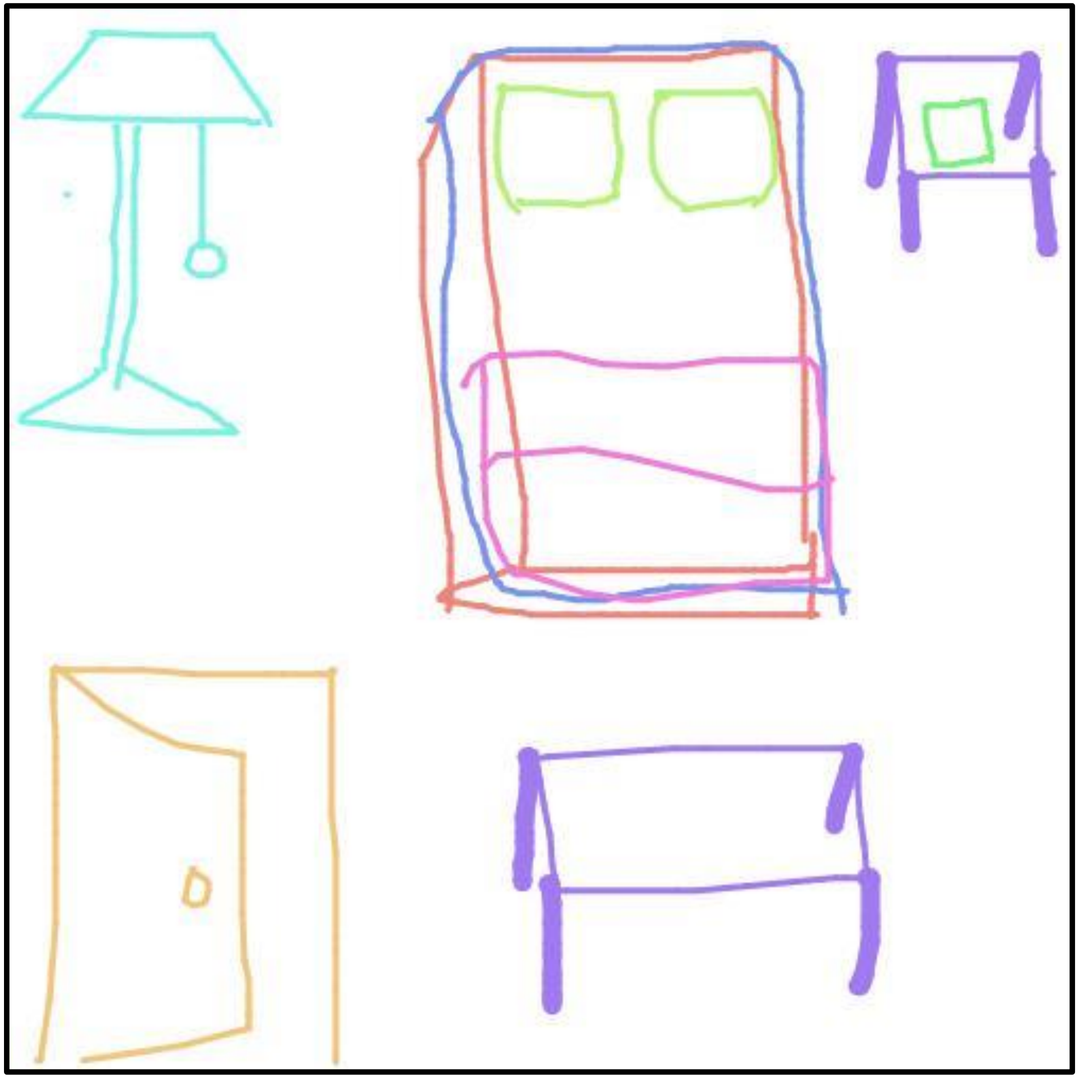
Line drawings

Line drawings

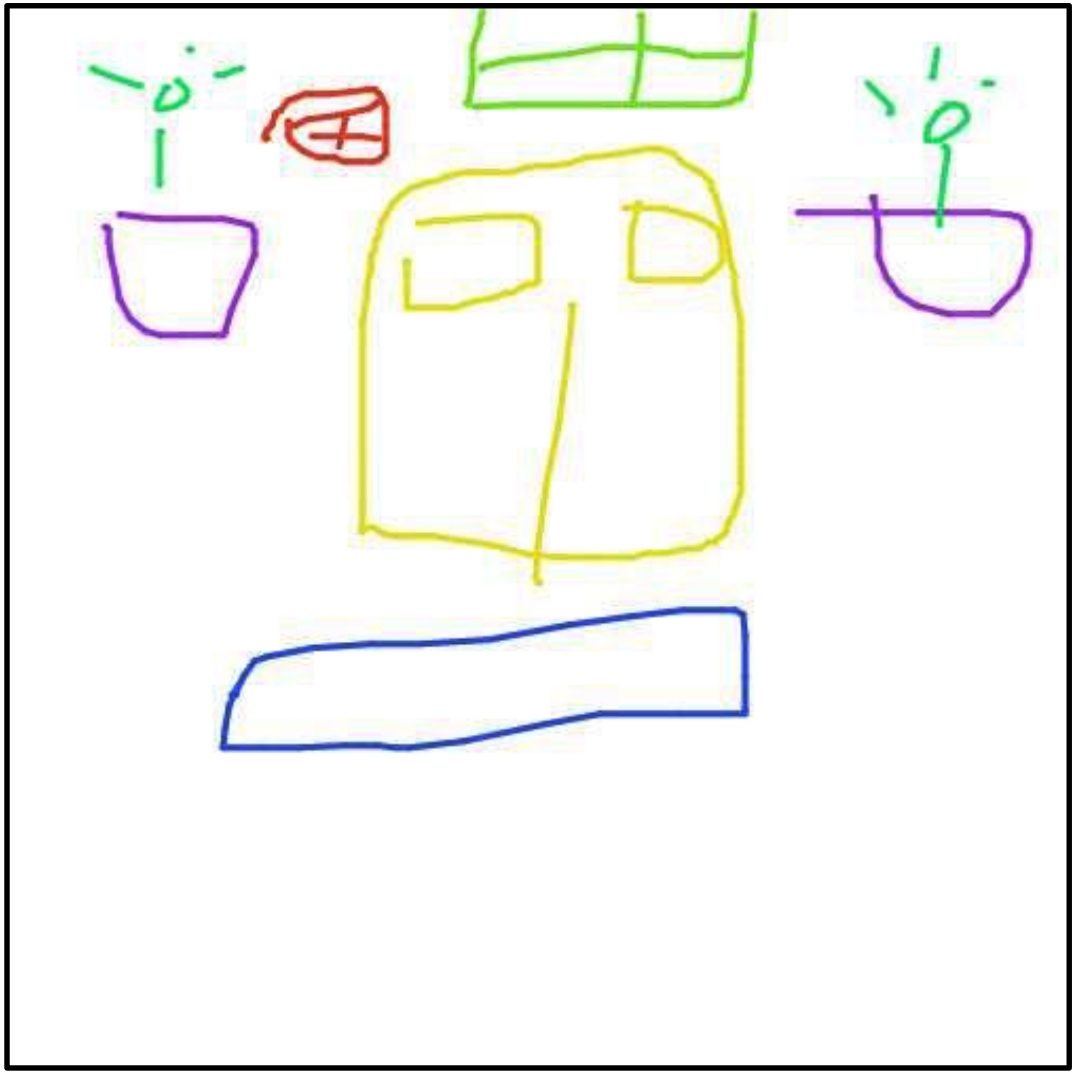
Line drawings



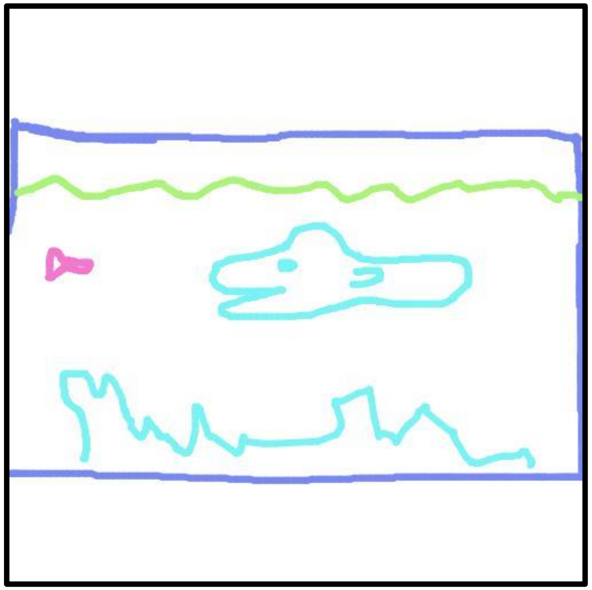
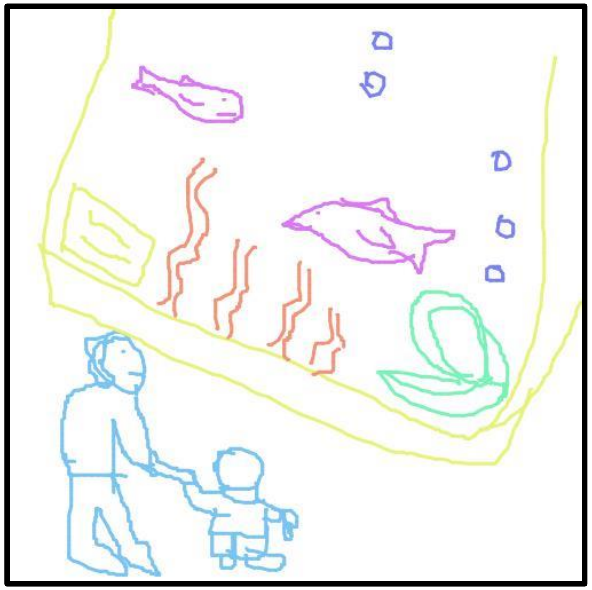
Line drawings



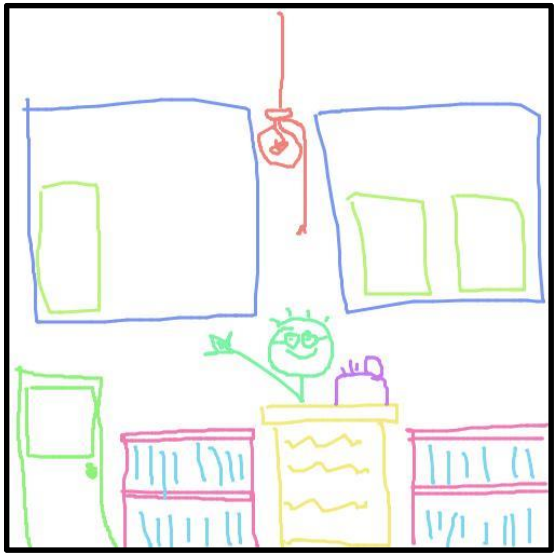
Line drawings



Aquarium

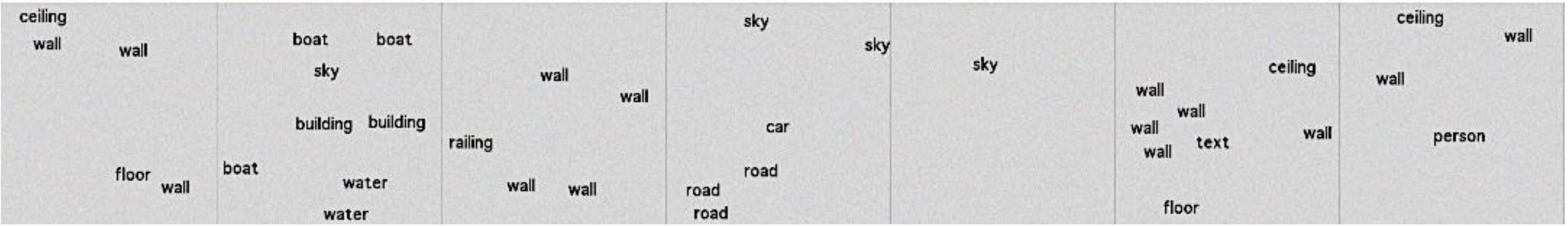



Library



Localized words

Localized words



or descriptions

There is a bed with a striped bedspread. Beside this is a nightstand with a drawer. There is also a tall dresser and a chair with a blue cushion. On the dresser is a jewelry box and a clock.

Descriptions

(Auditorium)

I'm looking forward to seeing this speaker and hearing his story today. I want to get in before all the seats are filled, because he is quite popular with the students and faculty. I don't want to sit way in the back where the sound may not carry as well to.

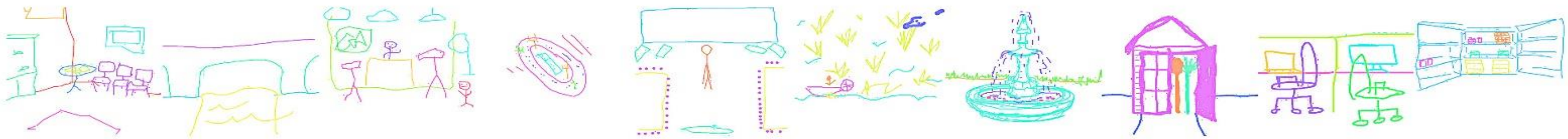
Descriptions

(Classroom)

This room is where students attend and are taught by a teacher on a variety of subjects. Each student seats in a desk which allows him to place books, and write on notebooks or sheets of paper. The teacher presides this room, and usually writes on a blackboard which occupies most of the front wall.

We collected a dataset formed by examples of 205 scene types in five different modalities:

Line drawings: 6,644 training – 2,050 validation examples



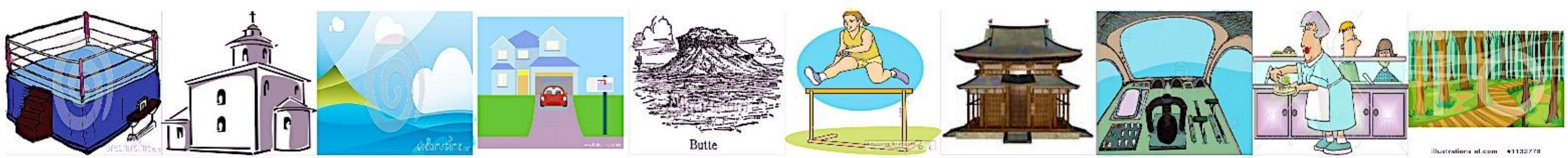
Descriptions: 4,307 training – 2,050 validation examples

There is a bed with a striped bedspread. Beside this is a nightstand with a drawer. There is also a tall dresser and a chair with a blue cushion. On the dresser is a jewelry box and a clock.

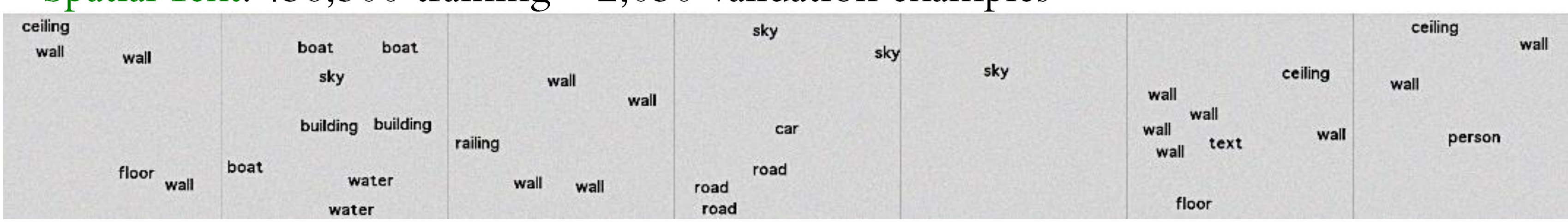
I am inside a room surrounded by my favorite things. This room is filled with pillows and a comfortable bed. There are stuffed animals everywhere. I have posters on the walls. My jewelry box is on the dresser.

There are brightly colored wooden tables with little chairs. There is a rug in one corner with ABC blocks on it. There is a bookcase with picture books, a larger teacher's desk and a chalkboard.

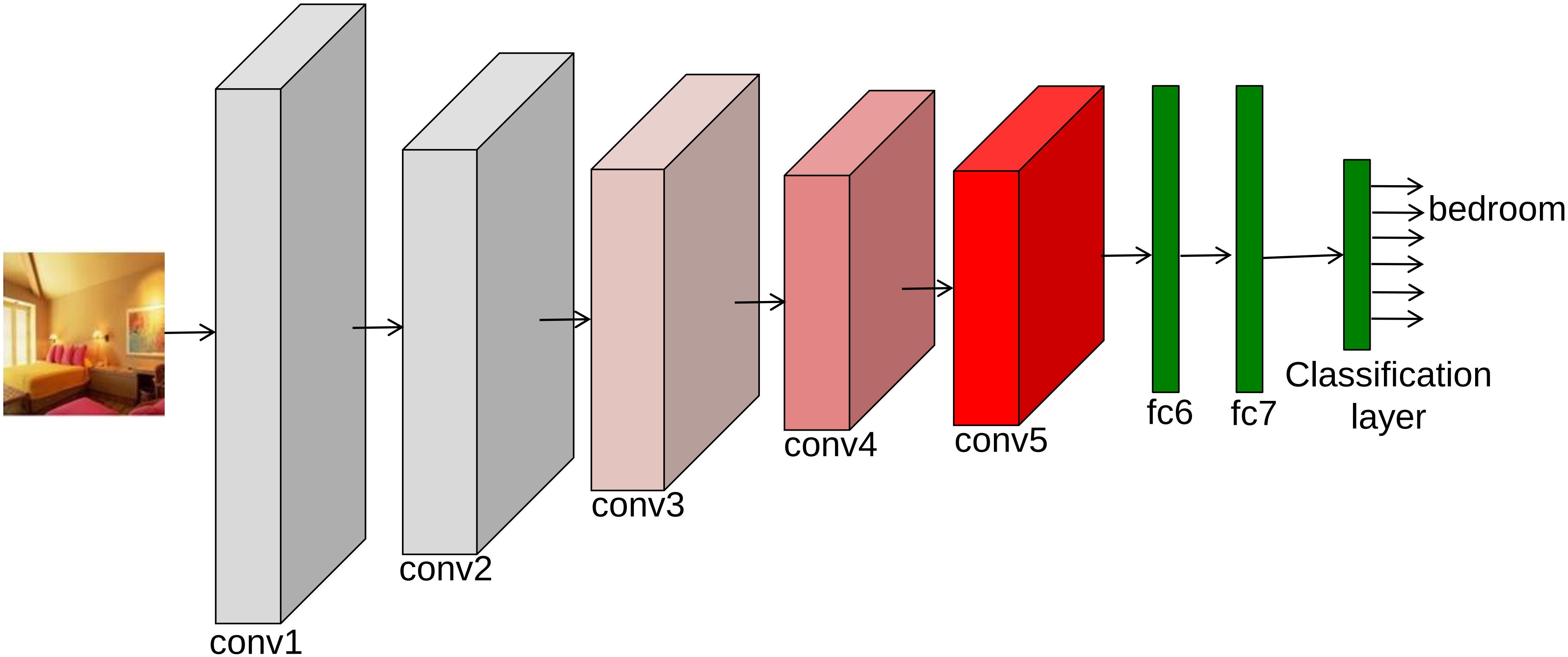
Clipart: 11,372 training – 1,954 validation examples



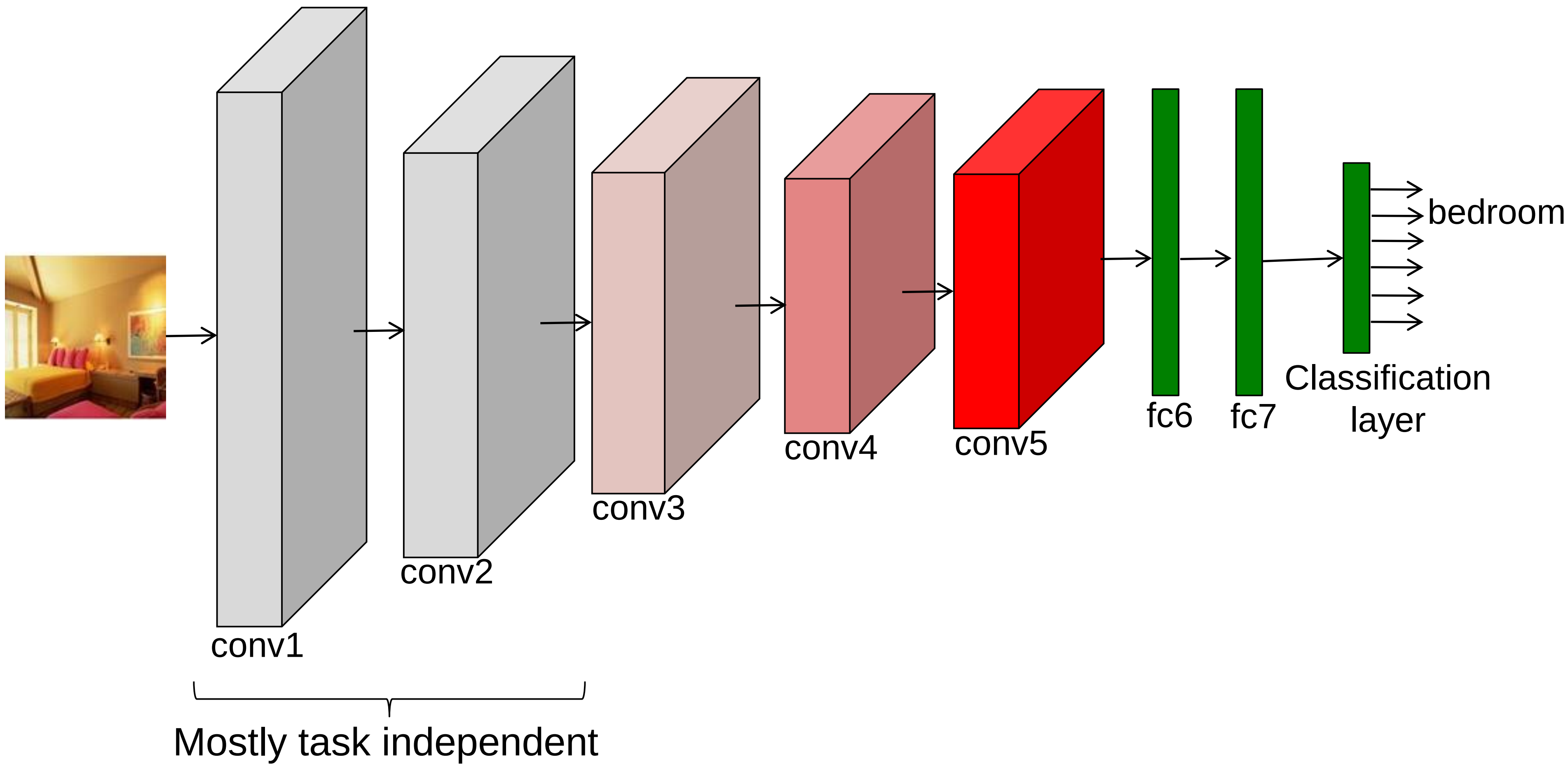
Spatial Text: 456,300 training – 2,050 validation examples

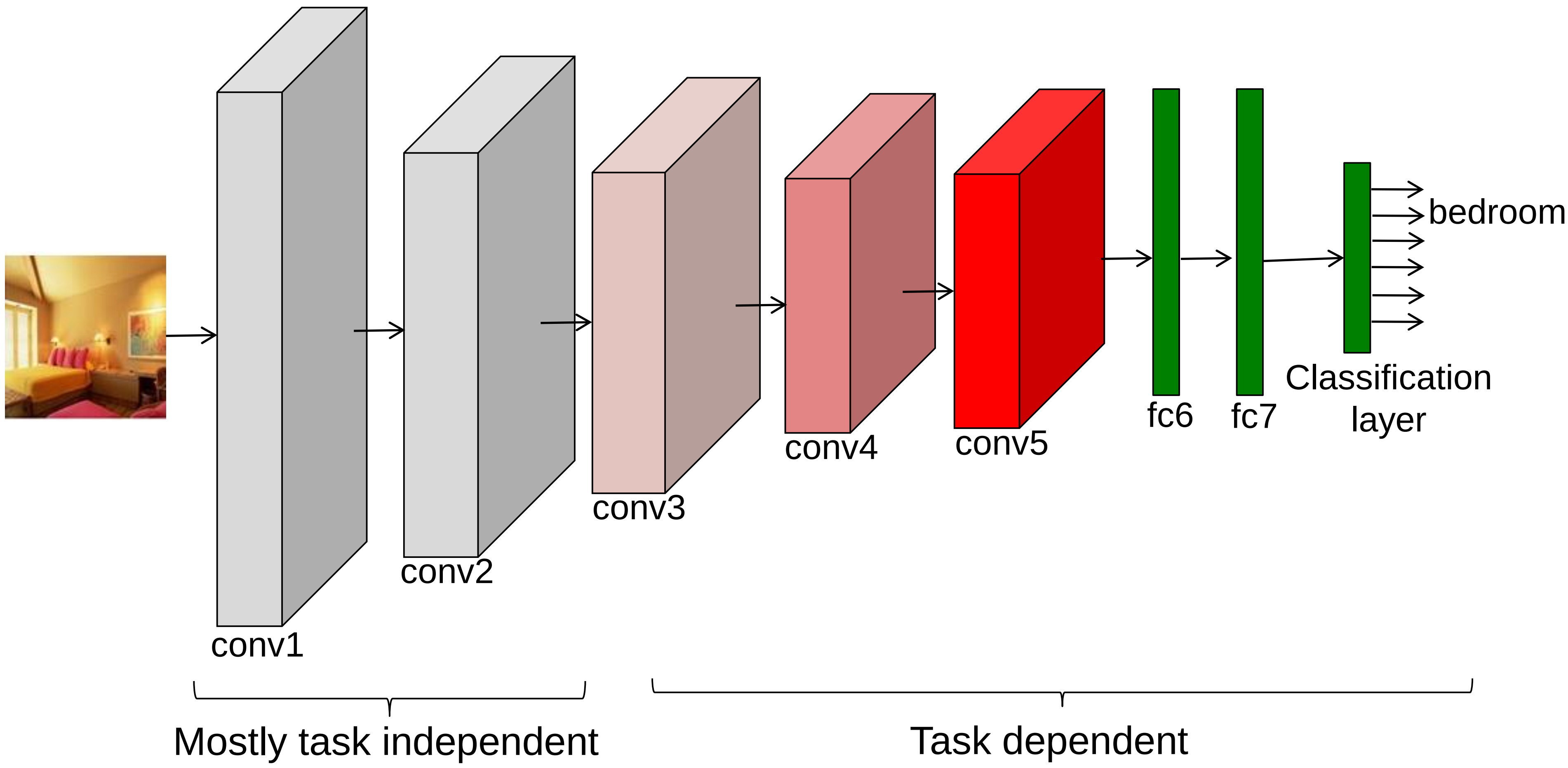


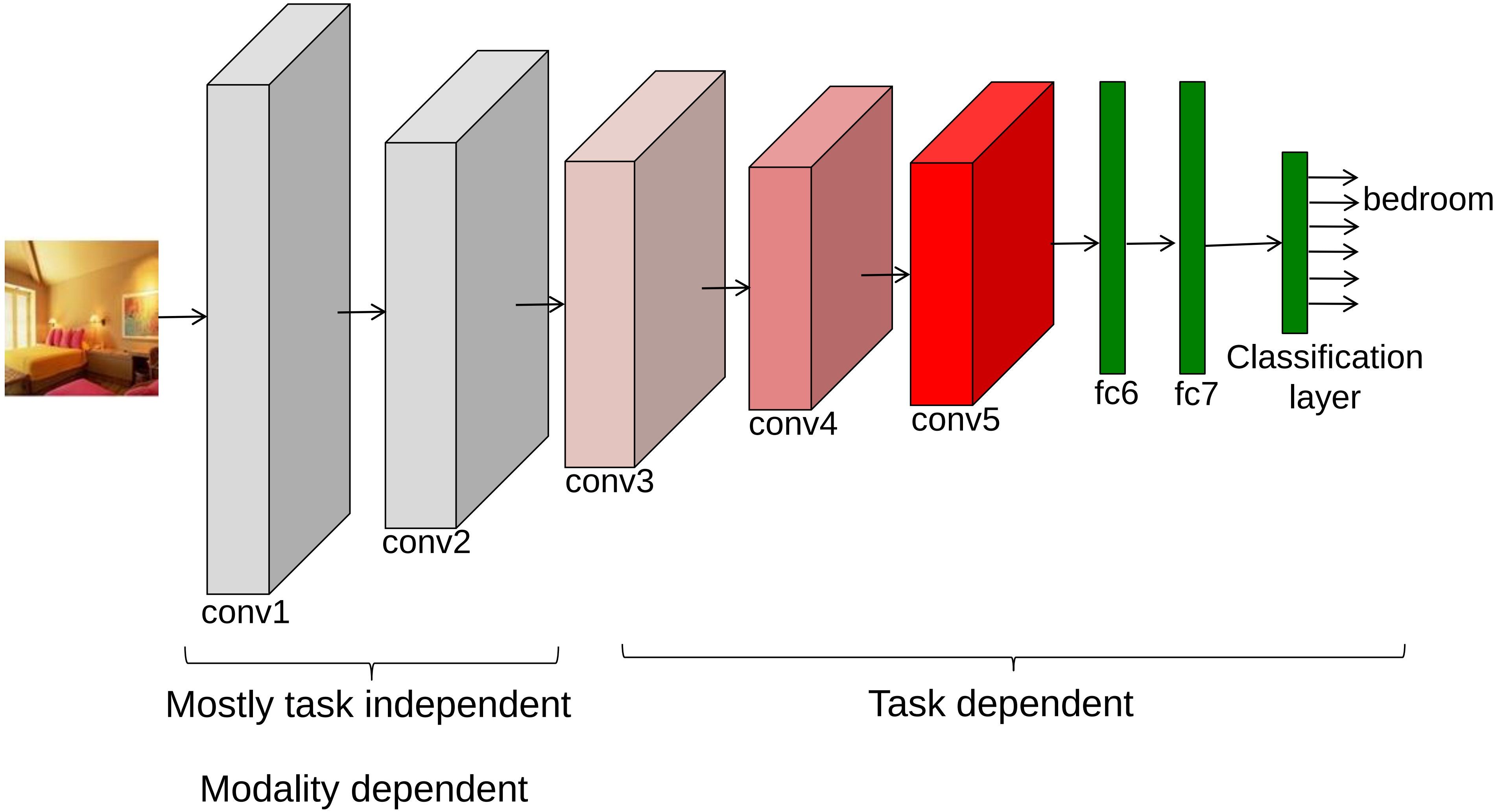
Natural images (Places dataset): ~ 2M training – 20,500 validation examples



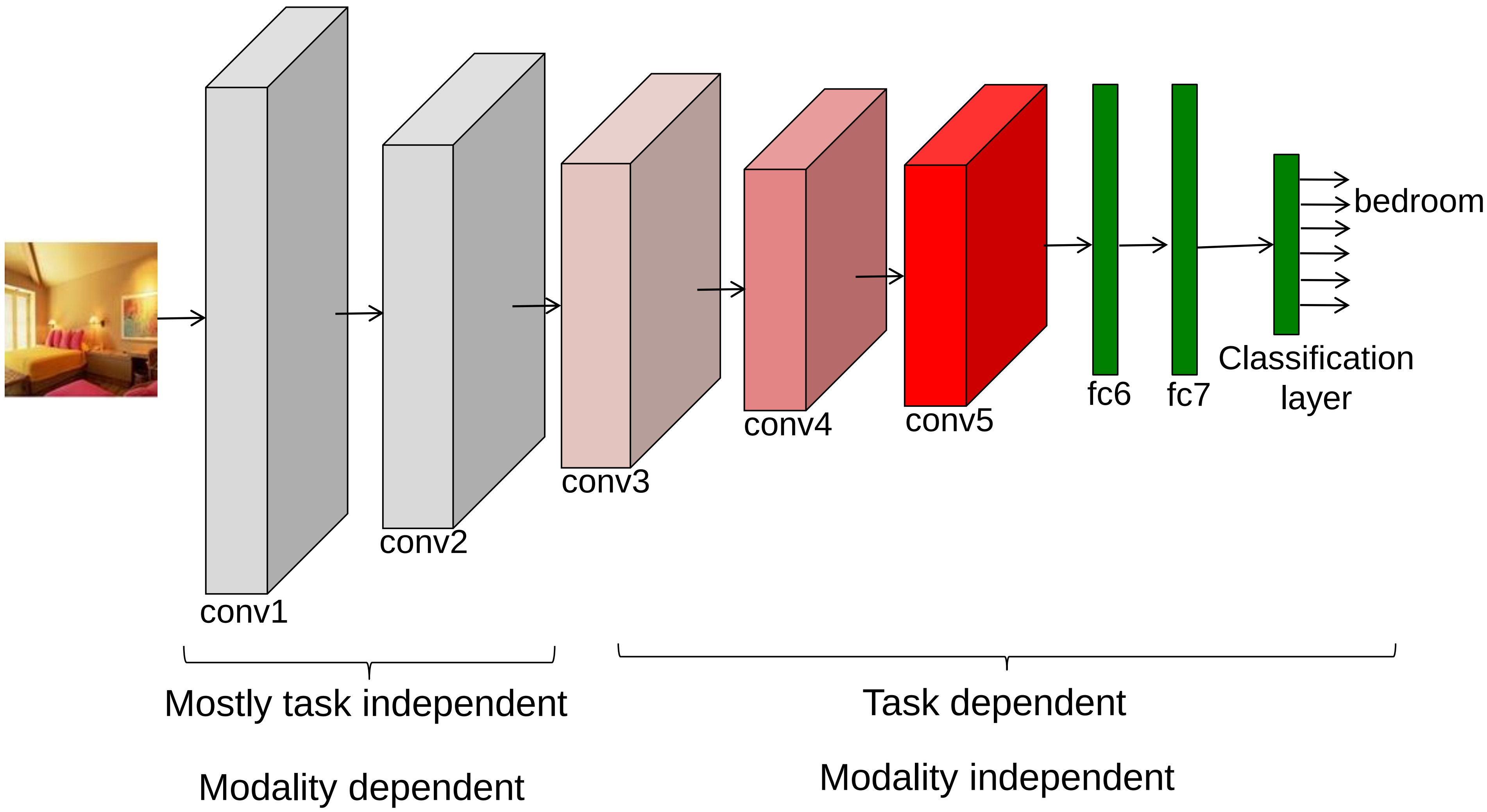
bedroom
conv1
conv2
conv3
conv4
conv5
fc6
fc7
Classification
layer

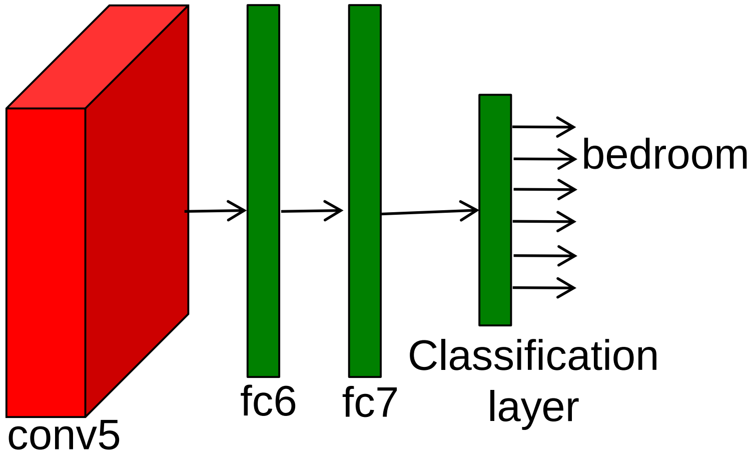

conv1
conv2
conv3
conv4
conv5
fc6
fc7
Classification layer
bedroom
Mostly task independent


bedroom
conv1
conv2
conv3
conv4
conv5
fc6
fc7
Classification layer
Mostly task independent
Task dependent

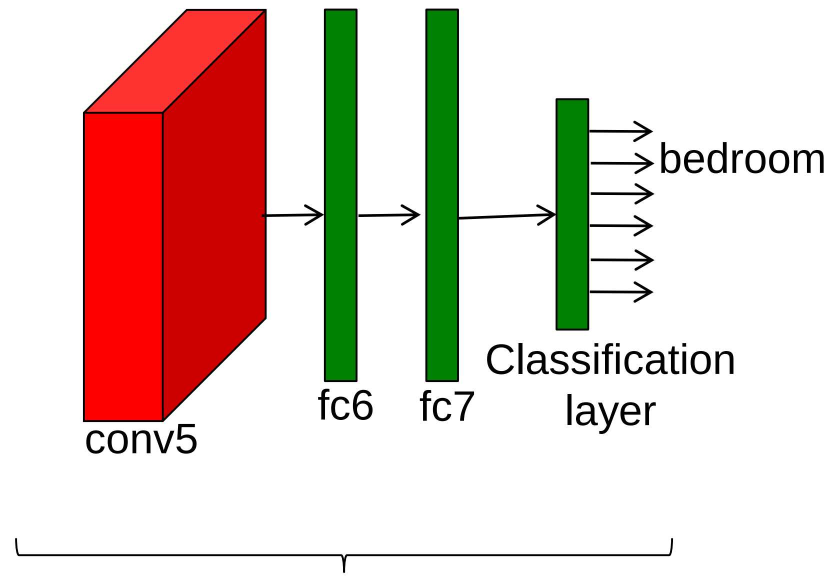


bedroom
Classification layer
fc6
fc7
conv1
conv2
conv3
conv4
conv5
Mostly task independent
Task dependent
Modality dependent


bedroom
conv1
conv2
conv3
conv4
conv5
fc6
fc7
Classification layer
Mostly task independent
Modality dependent
Task dependent
Modality independent

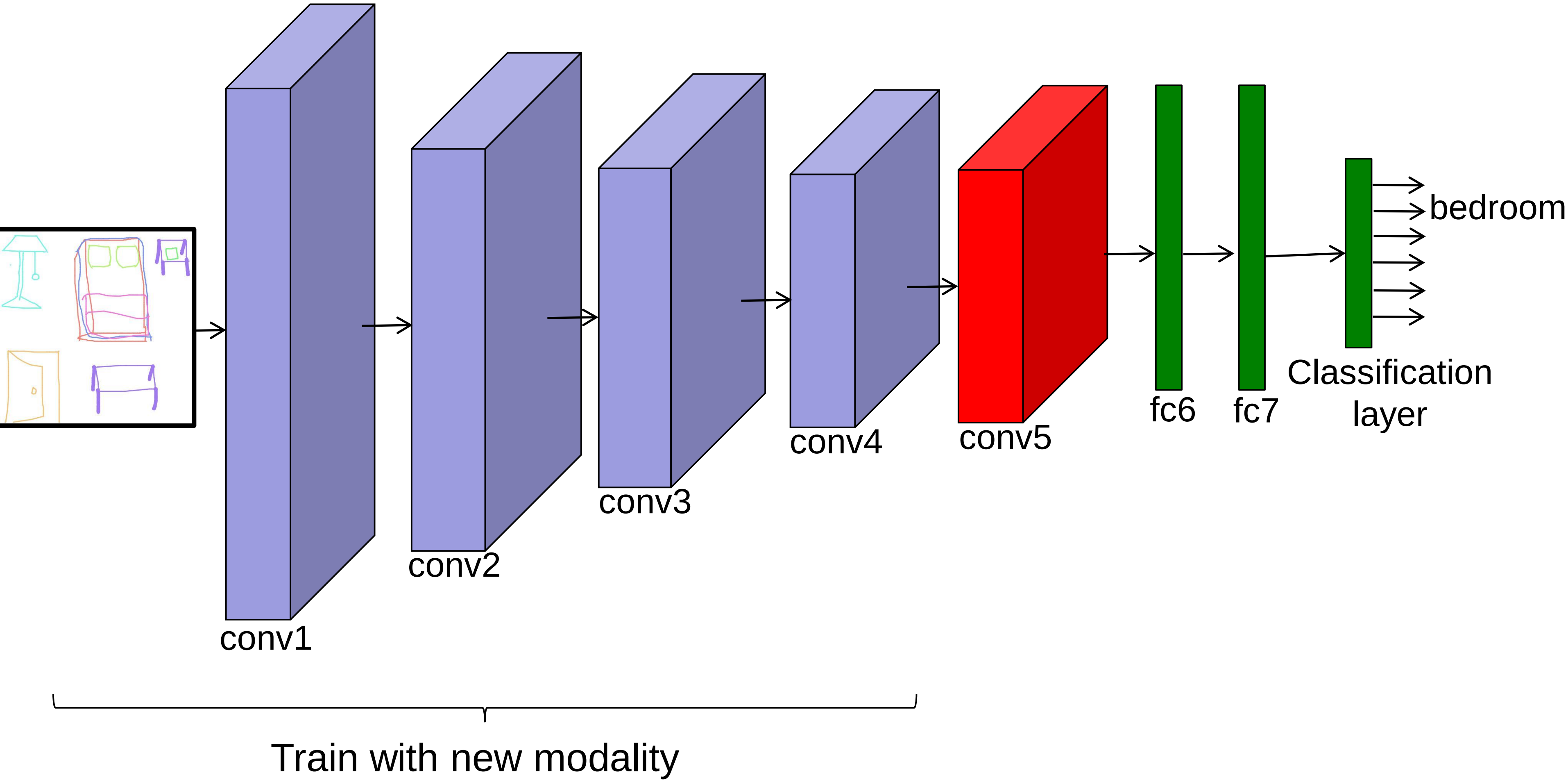


Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016

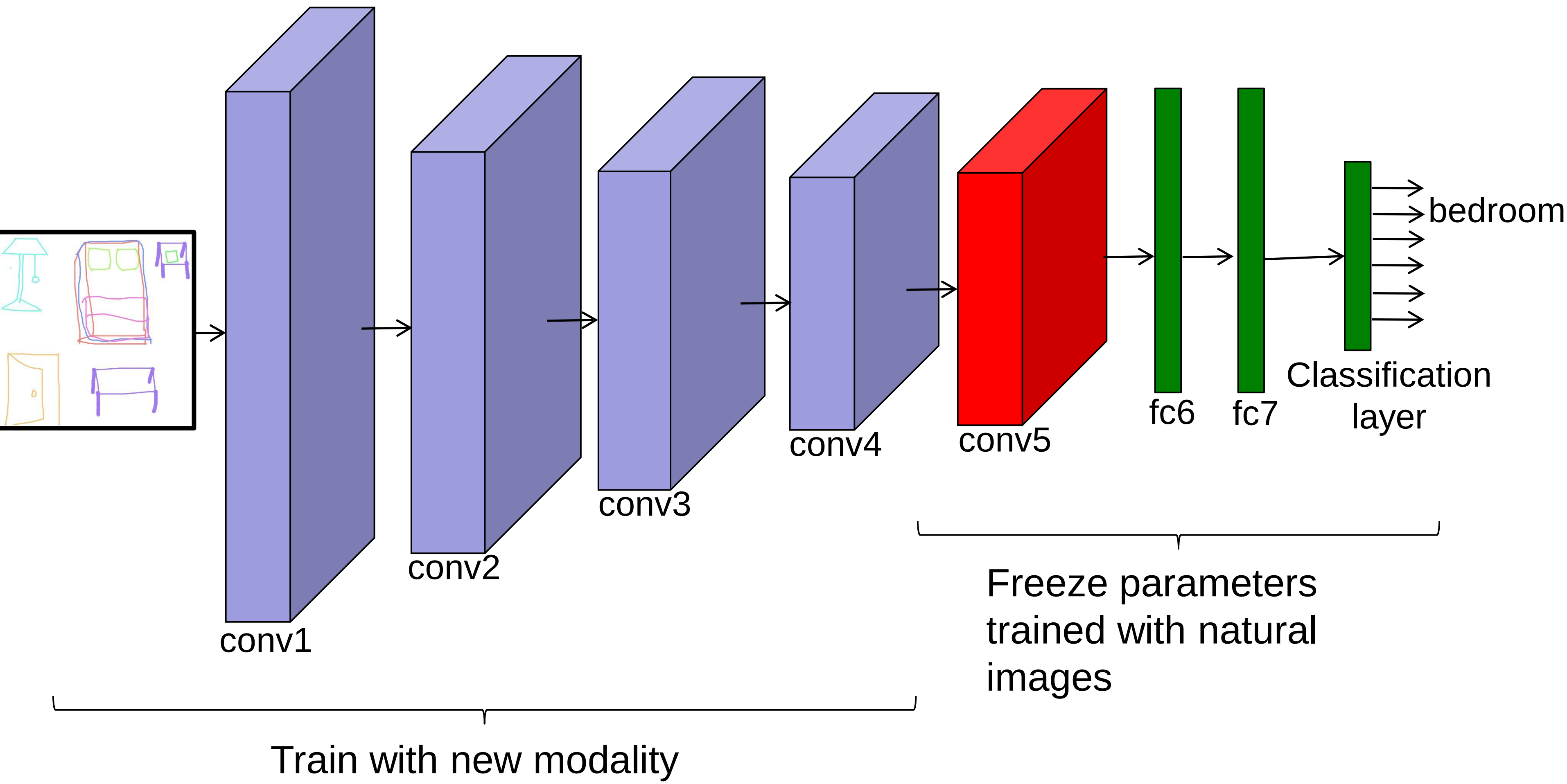


Freeze parameters trained with natural images

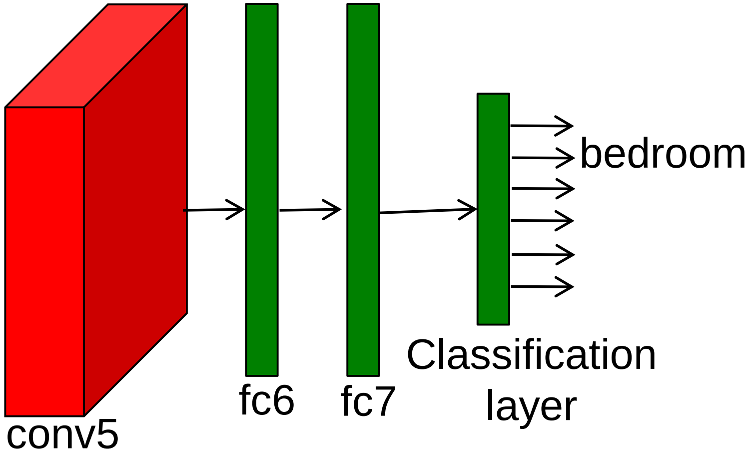
Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016



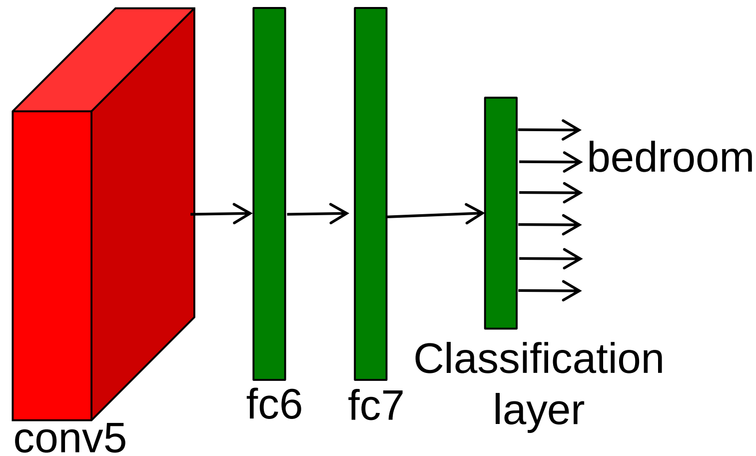
Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016



Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016

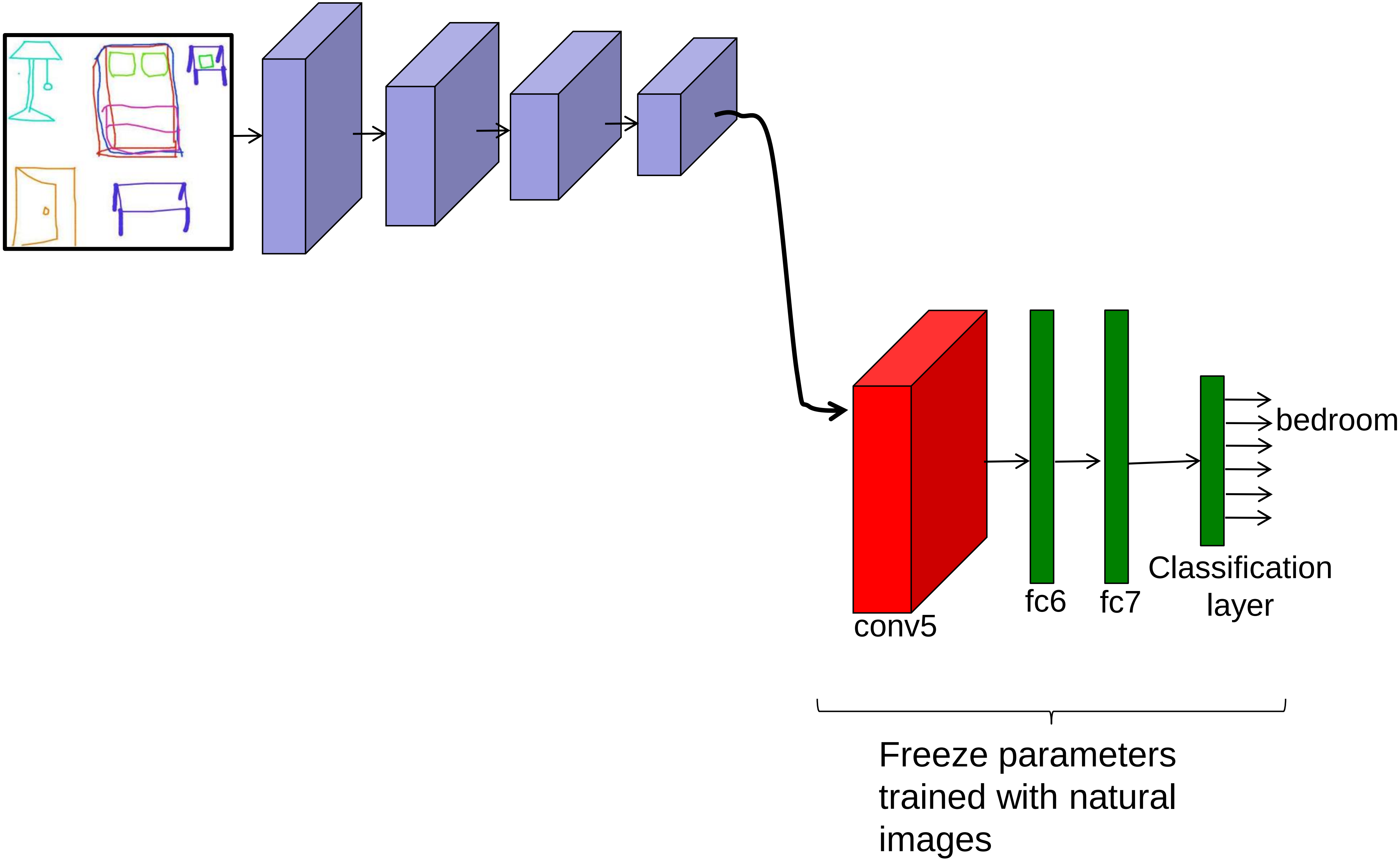


Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016

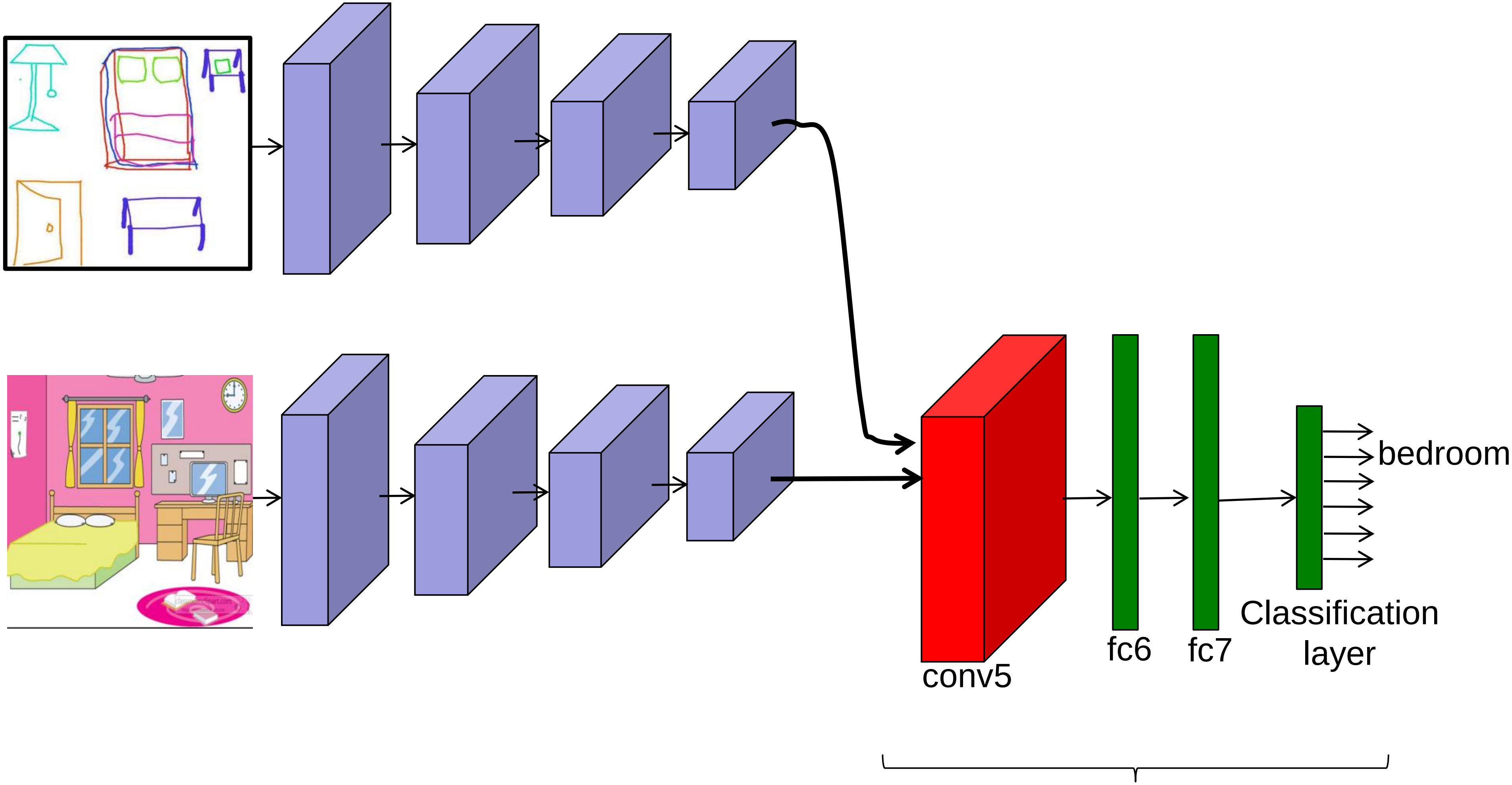


Freeze parameters trained with natural images

Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016

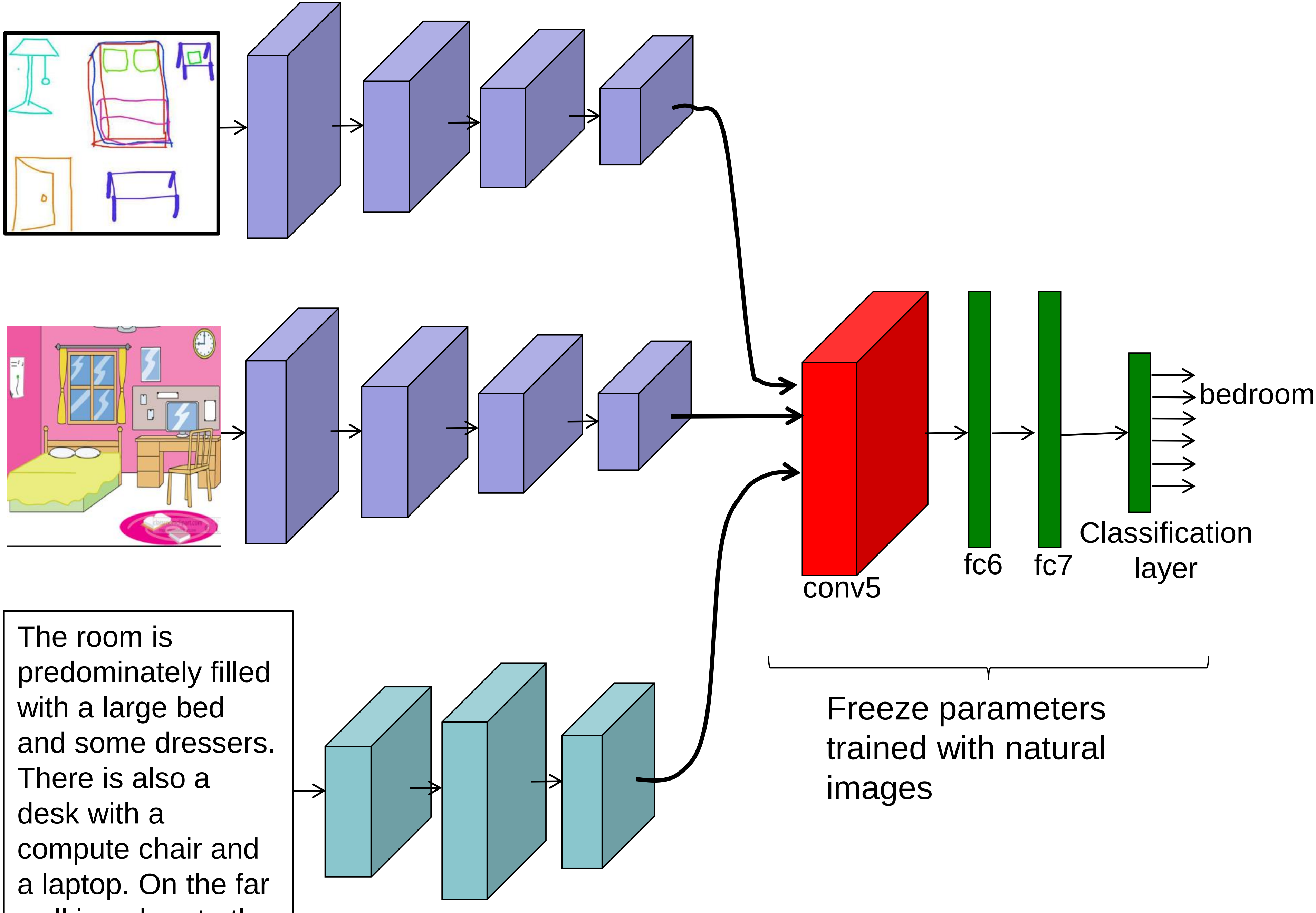


Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016



Freeze parameters trained with natural images

Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016



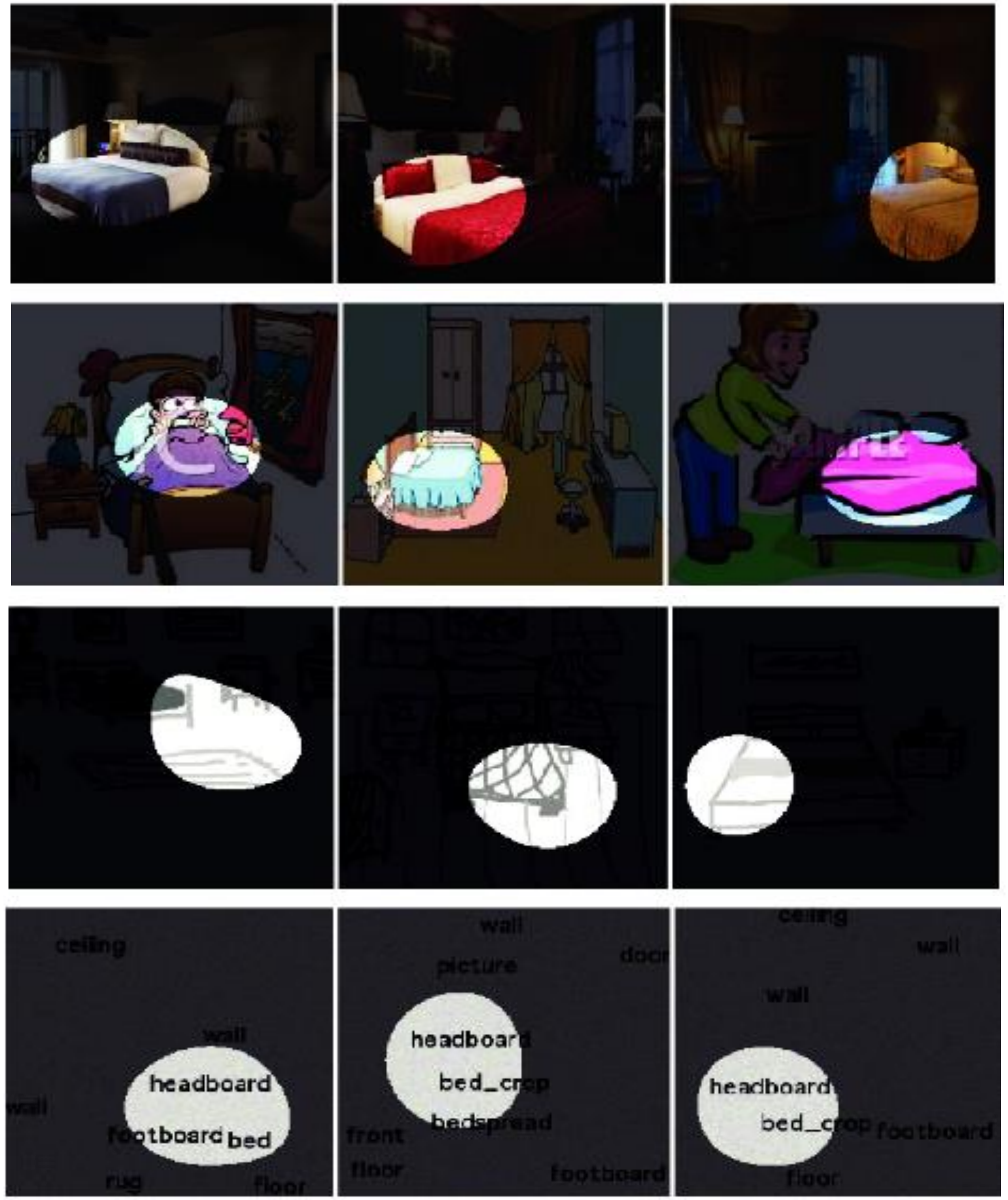
Learning Aligned Cross-Modal Representations from Weakly Aligned Data. Ll. Castrejón*, Y. Aytar*, C. Vondrick, H. Pirsiavash and A. Torralba. CVPR 2016

Unit 115
(Bed)

Unit 115
(Bed)

Unit 115
(Bed)

Unit 115
(Bed)

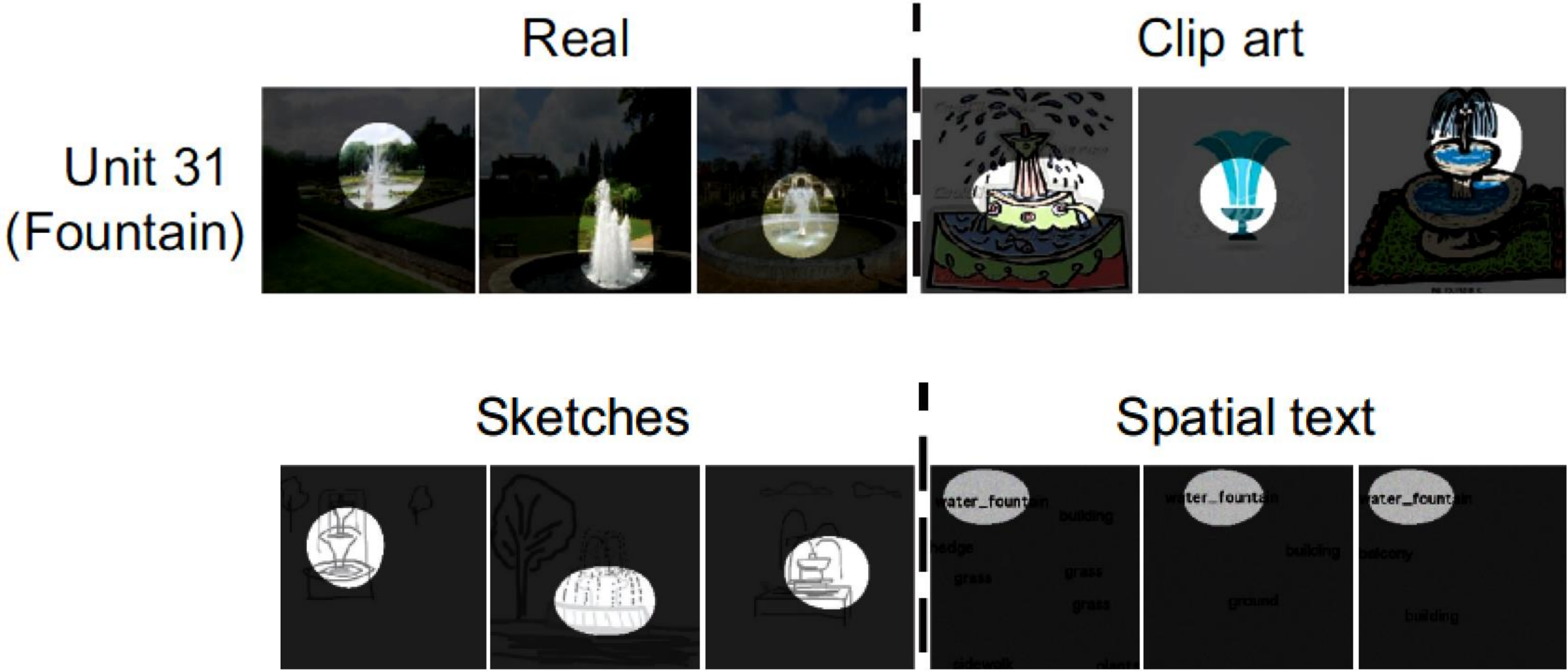


ceiling
wall
headboard
wall
footboard
bed
rug
floor
wall
picture
door
headboard
bed_crop
bedspread
front
floor
footboard
ceiling
wall
wall
headboard
bed_crop
footboard
floor

Unit 115
(Bed)

ice, terrain, plane, cold, i, nightstand, inside, beds, two, movement

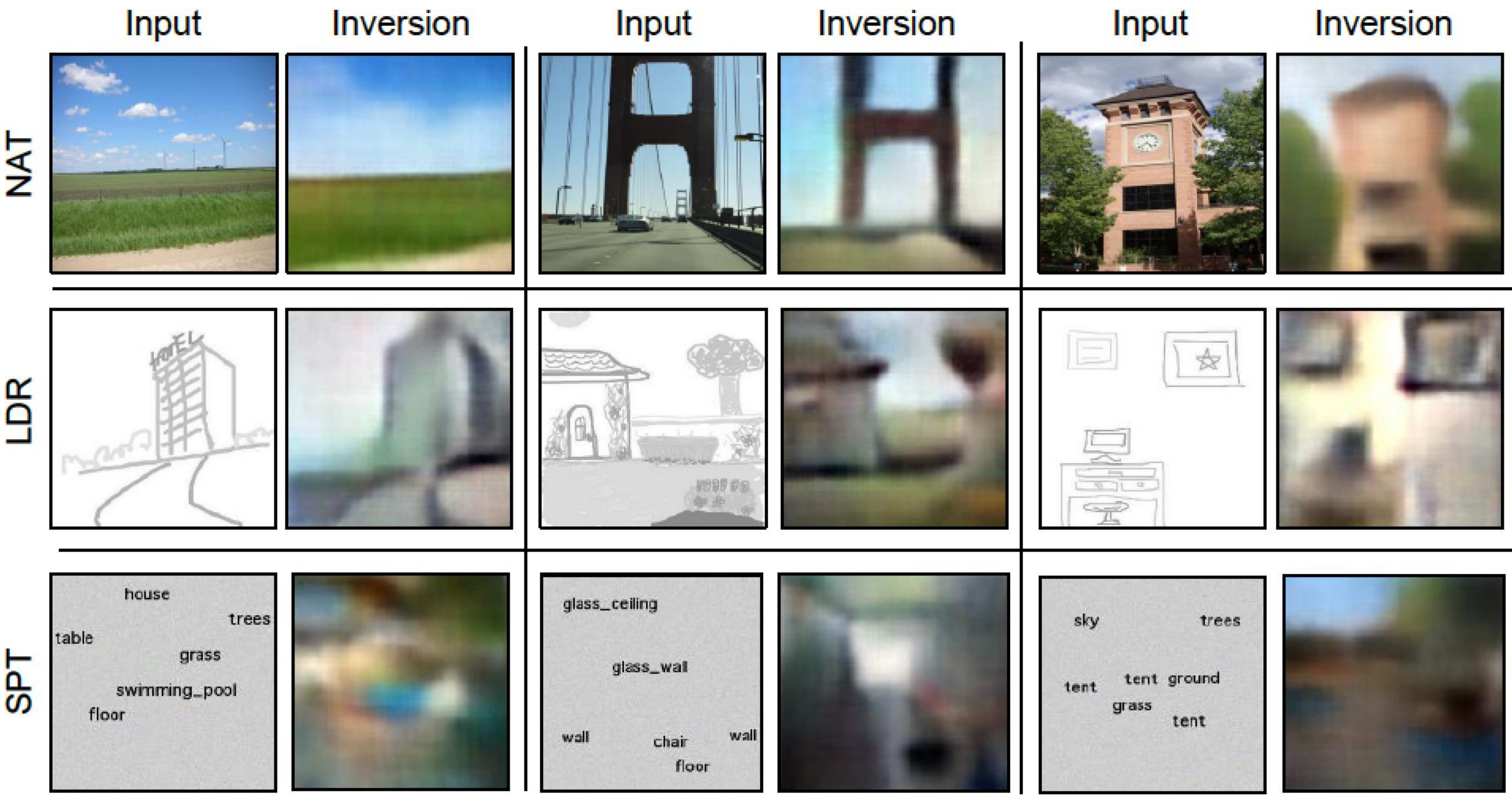
Units in pool5 become multimodal



Descriptions

we, water, fishes, you,
drink, formed, greek,
would, ball, have

Generating across modalities



A. Dosovitskiy and T. Brox. Inverting convolutional networks with convolutional networks. arXiv, 2015

Cross-modal learning

Description (eg, Wikipedia article)

Snares penguin

From Wikipedia, the free encyclopedia

The **Snares penguin** (*Eudyptes robustus*), also known as the **Snares crested penguin** and the **Snares Islands penguin**, is a penguin from New Zealand. The species breeds on The Snares, a group of islands off the southern coast of the South Island. This is a medium-small, yellow-crested penguin, at a size of 50–70 cm (19.5–27.5 in) and a weight of 2.5–4 kg (5.5–8.8 lb). It has dark blue-black upperparts and white underparts. It has a bright yellow eyebrow-stripe which extends over the eye to form a drooping, bushy crest. It has bare pink skin at the base of its large red-brown bill.

- Lots of descriptions/entries in Wikipedia available

Images





Zero-shot Learning

Description (eg, Wikipedia article)

Cardinal (bird)

From Wikipedia, the free encyclopedia

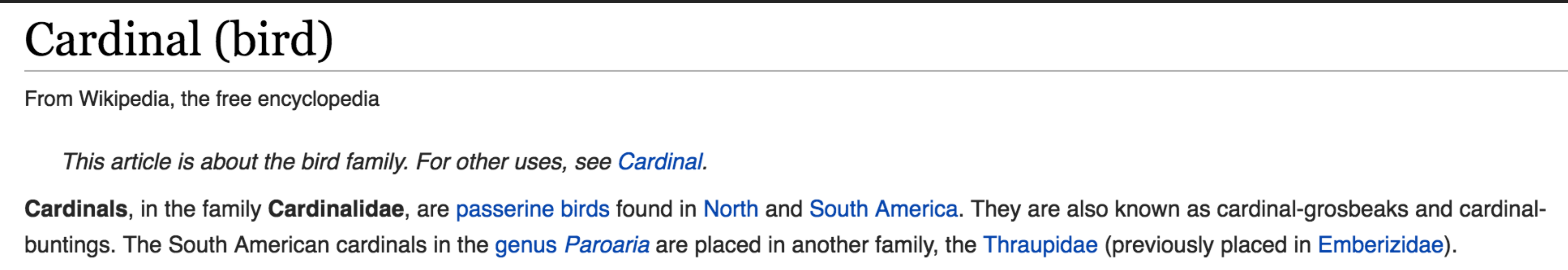
This article is about the bird family. For other uses, see Cardinal.

Cardinals, in the family **Cardinalidae**, are passerine birds found in North and South America. They are also known as cardinal-grosbeaks and cardinal-buntings. The South American cardinals in the genus *Paroaria* are placed in another family, the Thraupidae (previously placed in Emberizidae).

Can we predict an image classifier from a description alone?

Zero-shot Learning

Description (eg, Wikipedia article)



Can we predict an image classifier from a description alone?

Assume:

- In training we have access to wiki articles and labeled images
- For test classes we only have wiki articles
- We want to classify a new image (it can belong to any class)

Zero-shot Learning

- Goal: learn to predict an image classifier from a description
- Linear binary 1-vs-all classifier:

$$y_c = w_c^T \, x$$

- x … image feature vector
- w_c ... classifier weight vector for class c

Zero-shot Learning

- Goal: learn to predict an image classifier from a description
- Linear binary 1-vs-all classifier:

$$y_c = w_c^T x$$

- x … image feature vector
- w_c ... classifier weight vector for class c

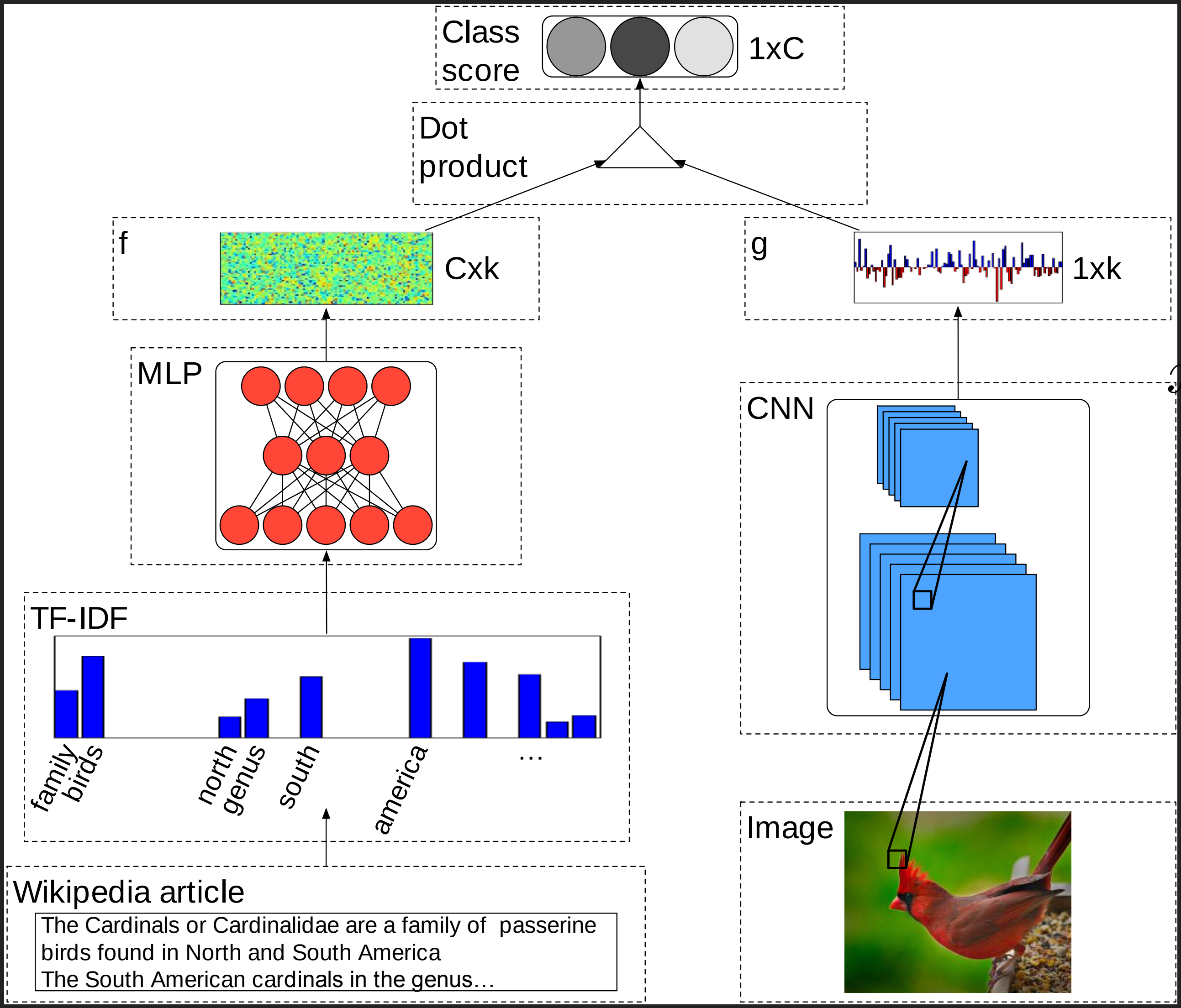
- We are also given t_c, a vector representing a textual description about class c
- We want:

$$w_c = f_t(t_c)$$

- f_c … a mapping $\mathbb{R}^p \rightarrow \mathbb{R}^d$ that transforms text features to the visual image feature space

Zero-shot Learning

- f_t can be a neural network



g used to compress x to a k<<d dim

Red faced Cormorant

The Red-faced Cormorant, Red-faced Shag or Violet Shag, Phalacrocorax urile, is a species of cormorant that is found in the far north of the Pacific Ocean and Bering Sea, from the eastern tip of Hokkaidō in Japan, via the Kuril Islands, the southern tip of the Kamchatka Peninsula and the Aleutian Islands to the Alaska Peninsula and Gulf of Alaska. The Red-faced Cormorant is closely related to the Pelagic Cormorant P. pelagicus, which has a similar range, and like the Pelagic Cormorant is placed by some authors (e.g. Johnsgaard) in a genus Leucocarbo. Where it nests alongside the Pelagic Cormorant, the Red-faced Cormorant generally breeds the more successfully of the two species, and it is currently increasing in numbers, at least in the easterly parts of its range. It is however listed as being of conservation concern{Verify source|date=September 2009}, partly because relatively little is so far known about it.

The adult bird has glossy plumage that is a deep greenish blue in colour, becoming purplish or bronze on the back and sides. In breeding condition it has a double crest,

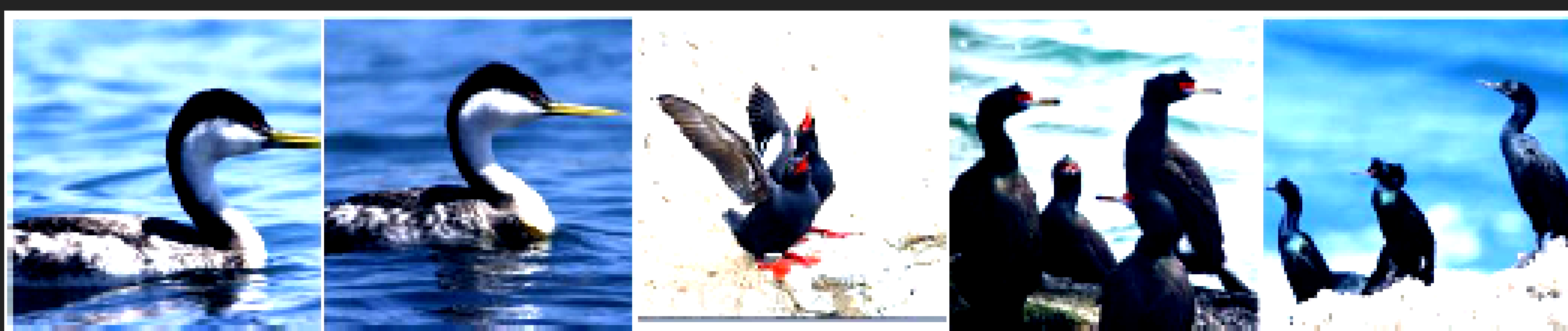
……

Red faced Cormorant

The Red-faced Cormorant, Red-faced Shag or Violet Shag, Phalacrocorax urile, is a species of cormorant that is found in the far north of the Pacific Ocean and Bering Sea, from the eastern tip of Hokkaidō in Japan, via the Kuril Islands, the southern tip of the Kamchatka Peninsula and the Aleutian Islands to the Alaska Peninsula and Gulf of Alaska. The Red-faced Cormorant is closely related to the Pelagic Cormorant P. pelagicus, which has a similar range, and like the Pelagic Cormorant is placed by some authors (e.g. Johnsgaard) in a genus Leucocarbo. Where it nests alongside the Pelagic Cormorant, the Red-faced Cormorant generally breeds the more successfully of the two species, and it is currently increasing in numbers, at least in the easterly parts of its range. It is however listed as being of conservation concern{Verify source|date=September 2009}, partly because relatively little is so far known about it.

The adult bird has glossy plumage that is a deep greenish blue in colour, becoming purplish or bronze on the back and sides. In breeding condition it has a double crest,

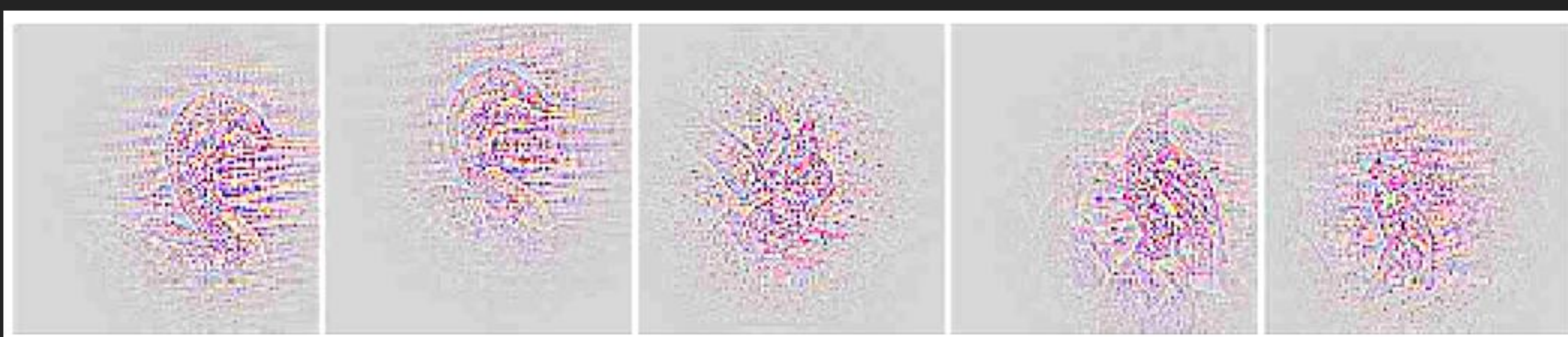
......



Red faced Cormorant

The Red-faced Cormorant, Red-faced Shag or Violet Shag, Phalacrocorax urile, is a species of cormorant that is found in the far north of the Pacific Ocean and Bering Sea, from the eastern tip of Hokkaidō in Japan, via the Kuril Islands, the southern tip of the Kamchatka Peninsula and the Aleutian Islands to the Alaska Peninsula and Gulf of Alaska. The Red-faced Cormorant is closely related to the Pelagic Cormorant P. pelagicus, which has a similar range, and like the Pelagic Cormorant is placed by some authors (e.g. Johnsgaard) in a genus Leucocarbo. Where it nests alongside the Pelagic Cormorant, the Red-faced Cormorant generally breeds the more successfully of the two species, and it is currently increasing in numbers, at least in the easterly parts of its range. It is however listed as being of conservation concern{Verify source|date=September 2009}, partly because relatively little is so far known about it.

The adult bird has glossy plumage that is a deep greenish blue in colour, becoming purplish or bronze on the back and sides. In breeding condition it has a double crest,
......

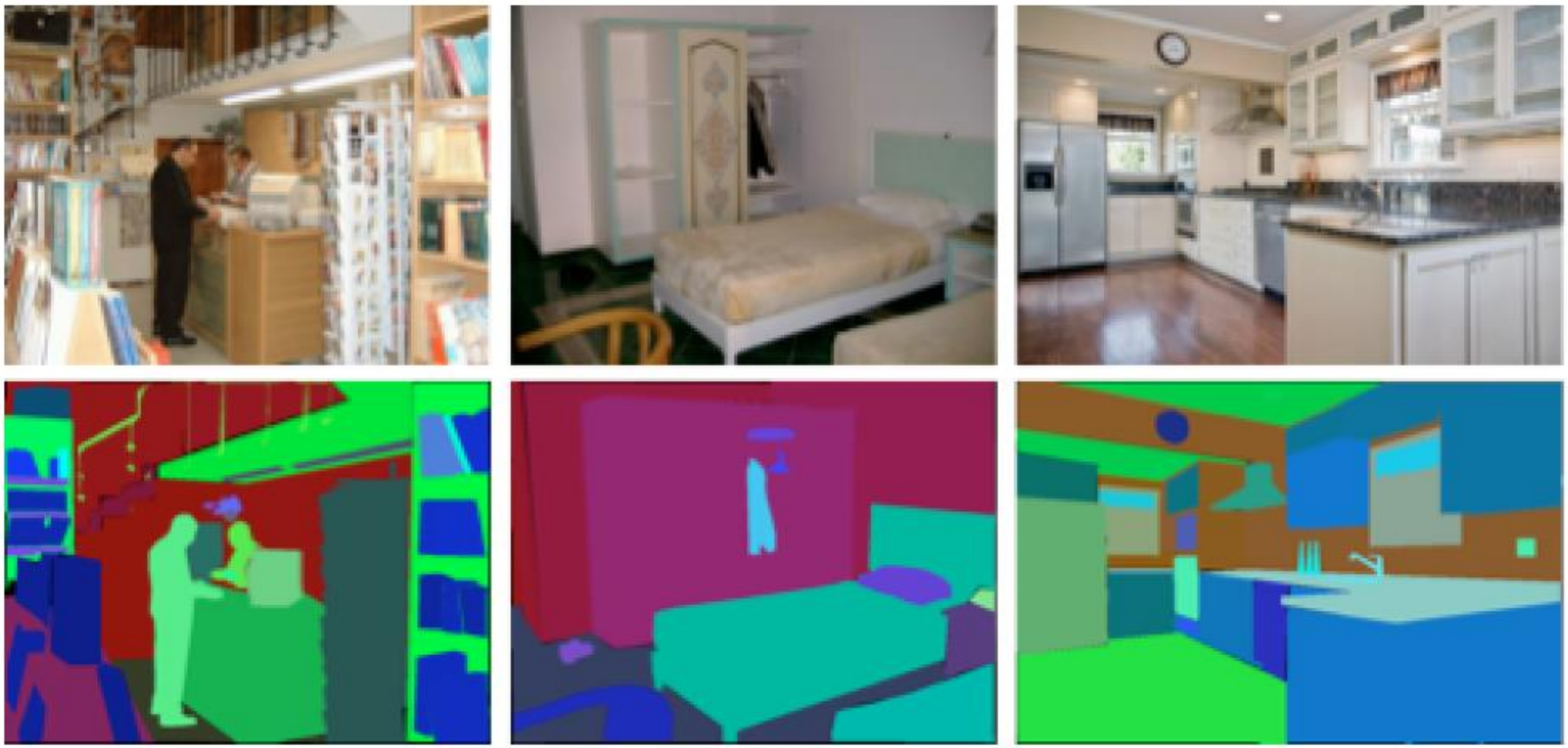


visualization by Zeiler & Fergus, ECCV’14.

Learning to see

It is all about the data…

Strong supervision



Pixel wise labeling

Learning to see

It is all about the data…

Weak supervision

Bird

Bedroom

Short captions

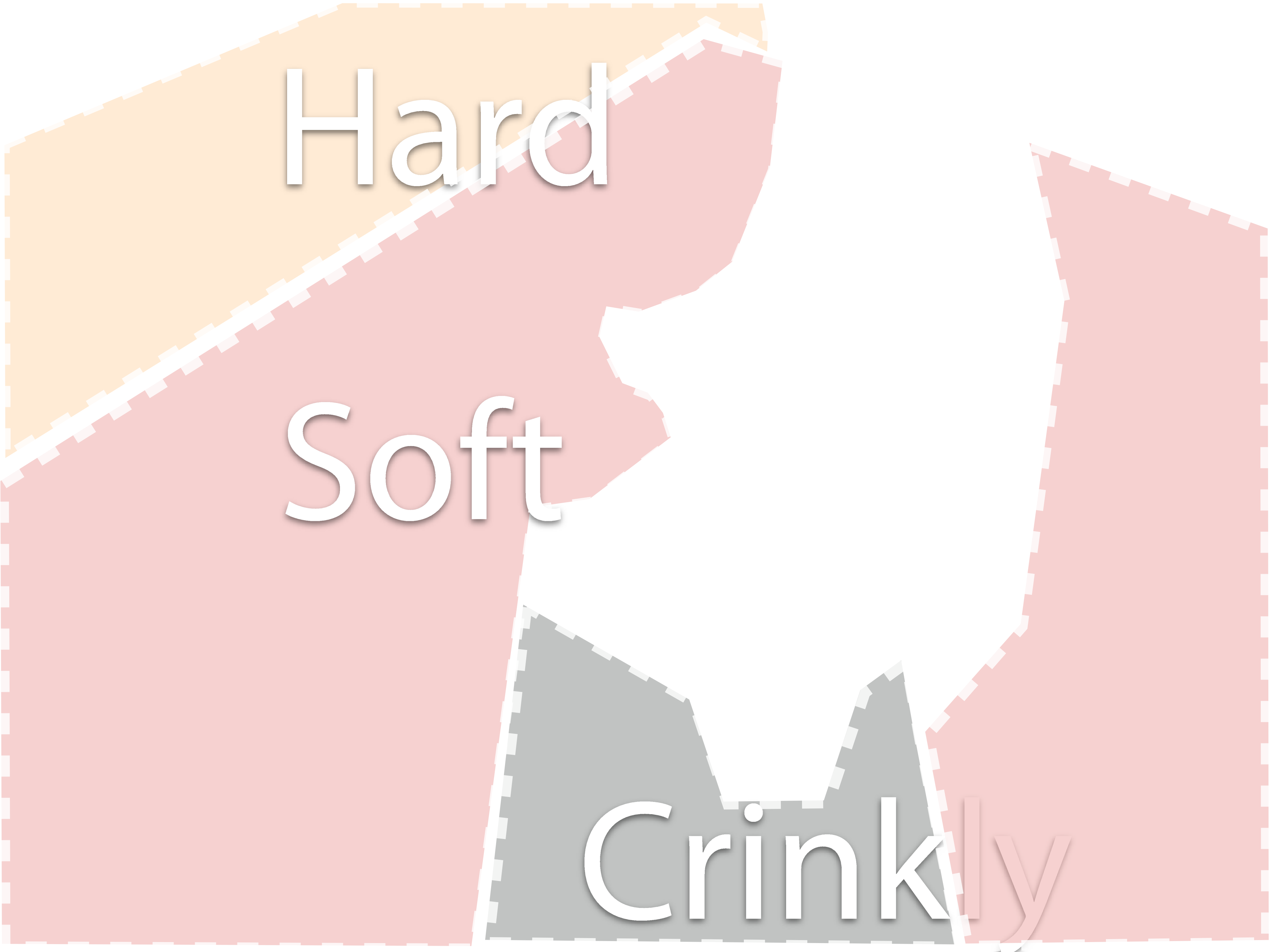
Cross modal: text and images

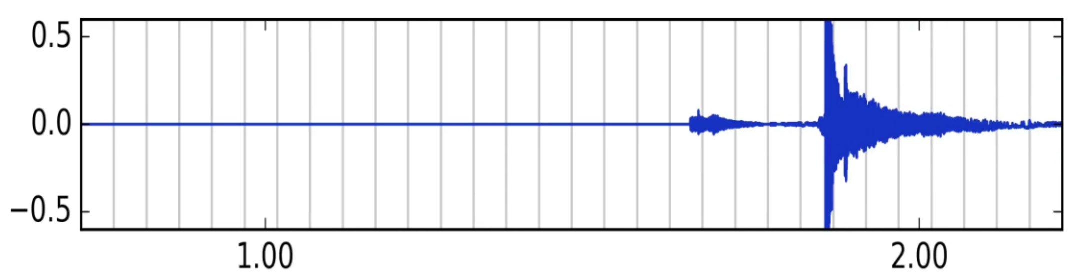
Man holding a metal bowl at the table.

from Microsoft CoCo

Q: Is everyone of these four holding a wine glass? A: No
Q: How many men are there? A: 3
Q: Does the window have blinds? A: yes

From http://visualqa.org/index.html


Hard
Soft
Crinkly


0.5
0.0
−0.5
1.00
2.00


Soft
Hard
Rough

Visually Indicated Sounds

Andrew Owens

Phillip Isola

Josh McDermott

Antonio Torralba

Ted Adelson

Bill Freeman

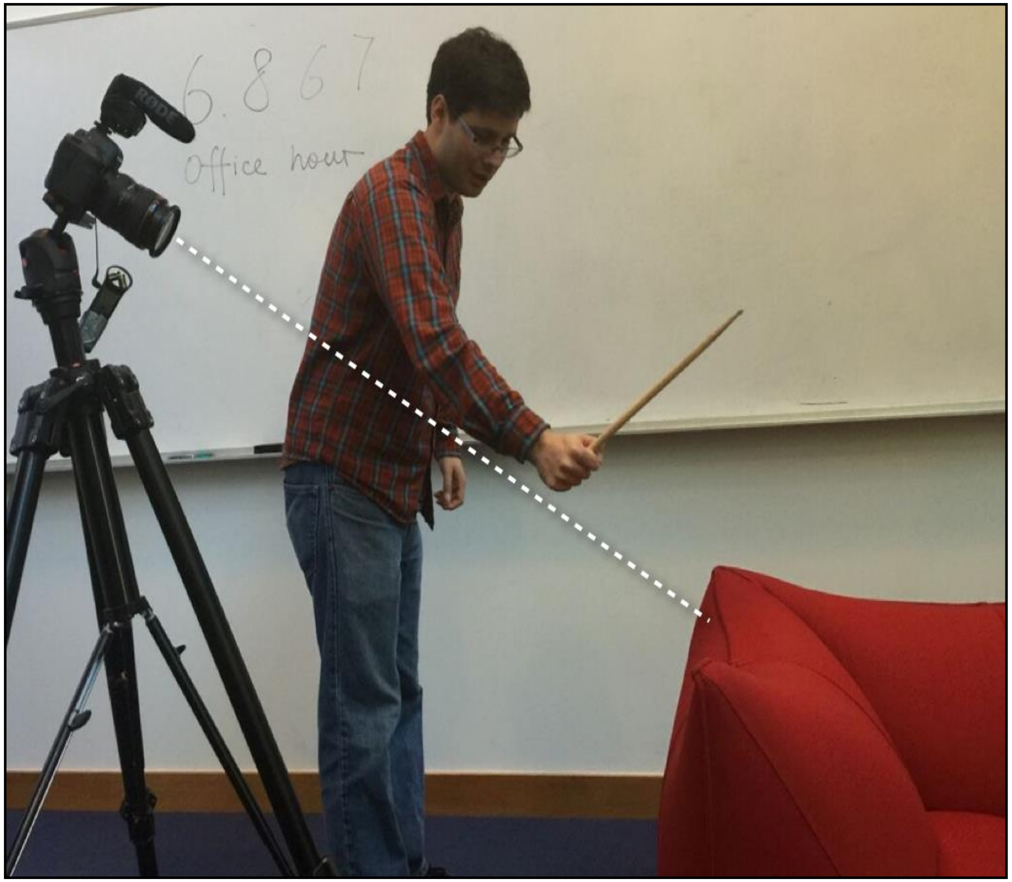
MIT CSAIL

brain+cognitive sciences

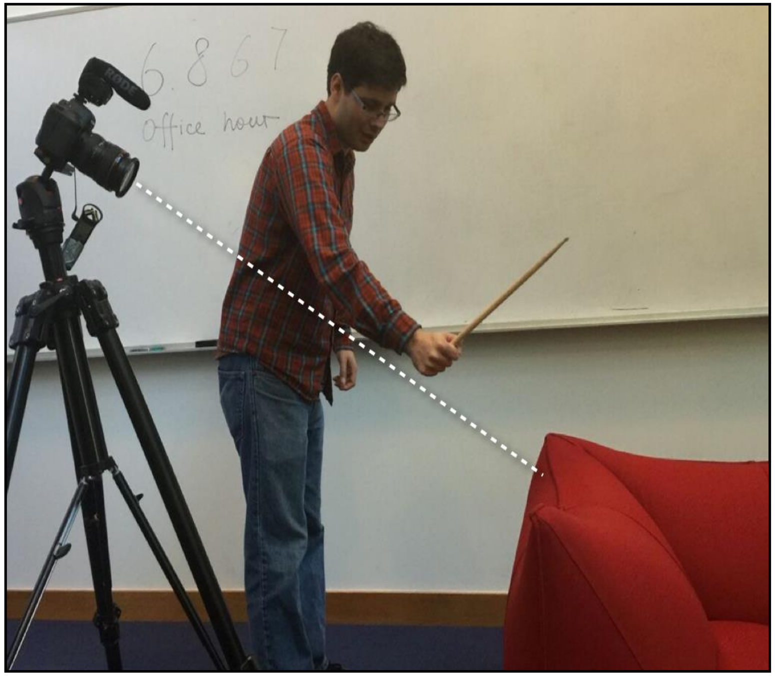
Berkeley EECS
ELECTRICAL ENGINEERING & COMPUTER SCIENCES

Research at Google

Collecting a dataset of physical interactions



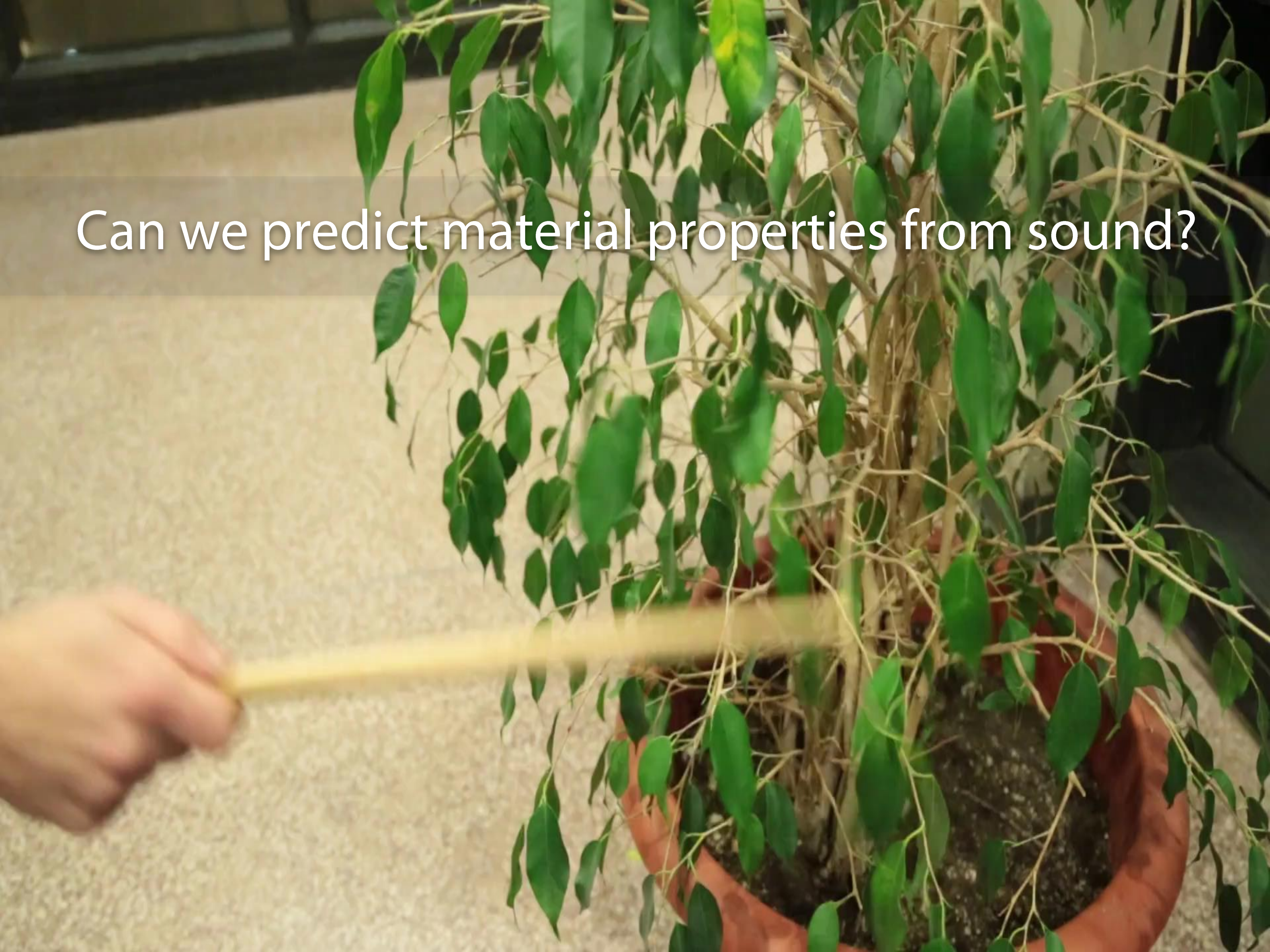
Collecting a dataset of physical interactions



The Greatest Hits dataset

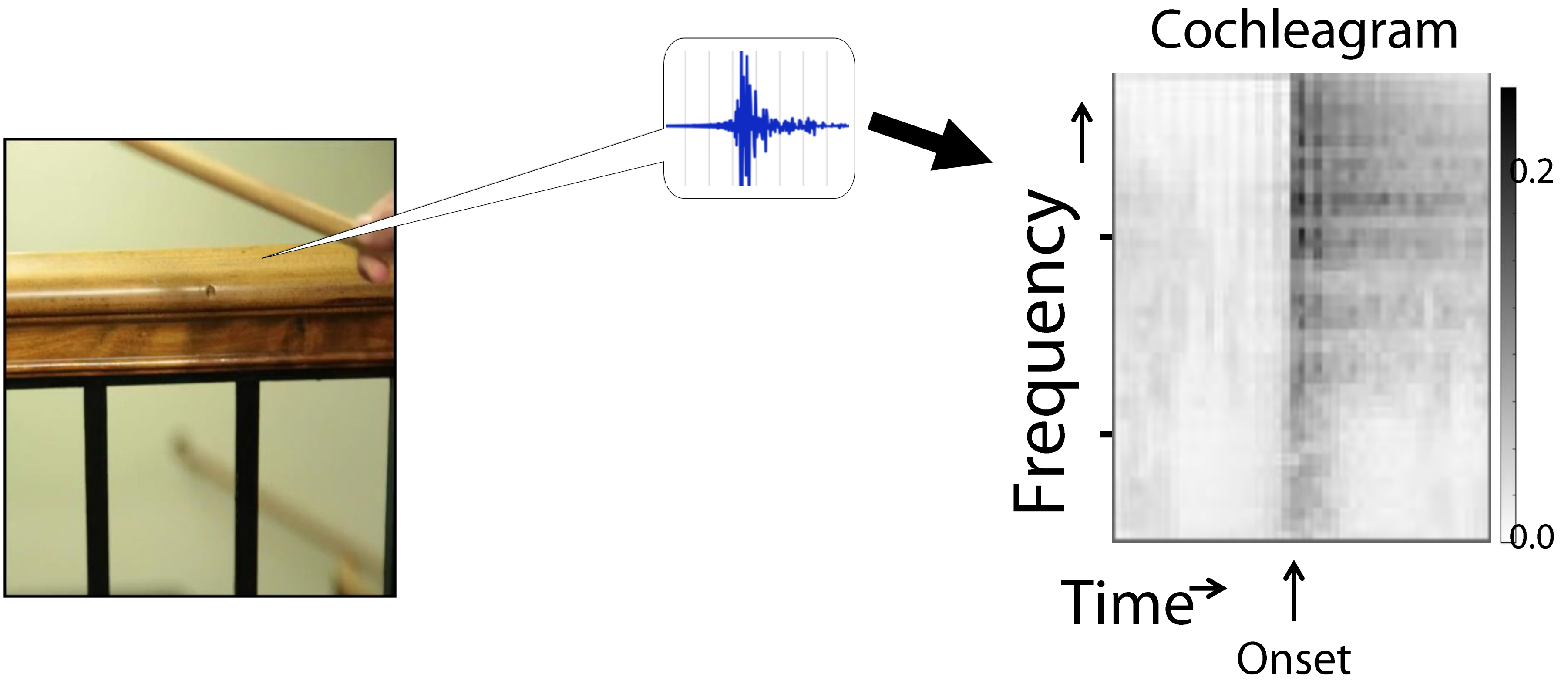
- 977 videos, 35 sec. long
- 46,577 segmented hits and scratches
- Material, action, reaction labels


The *Greatest Hits* Dataset


Can we predict material properties from sound?

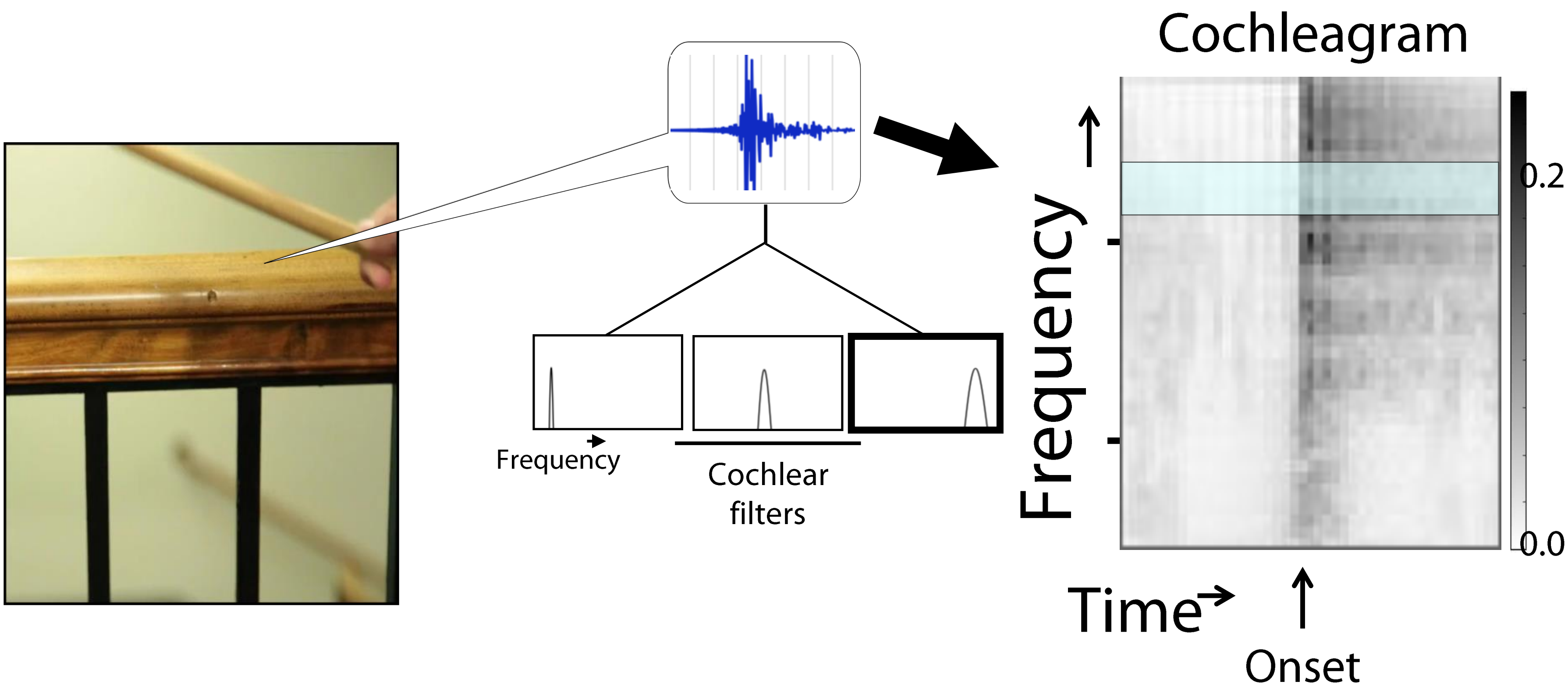
Can we predict material properties from sound?

Can we predict material properties from sound?



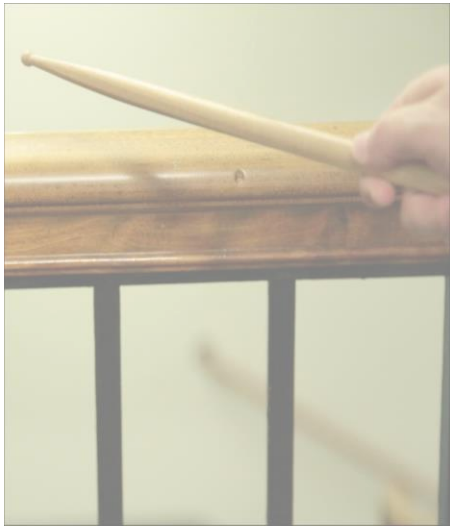
- 40 bandpass filters (+ high/low pass)
- 3 samples per frame (90 Hz)

Can we predict material properties from sound?

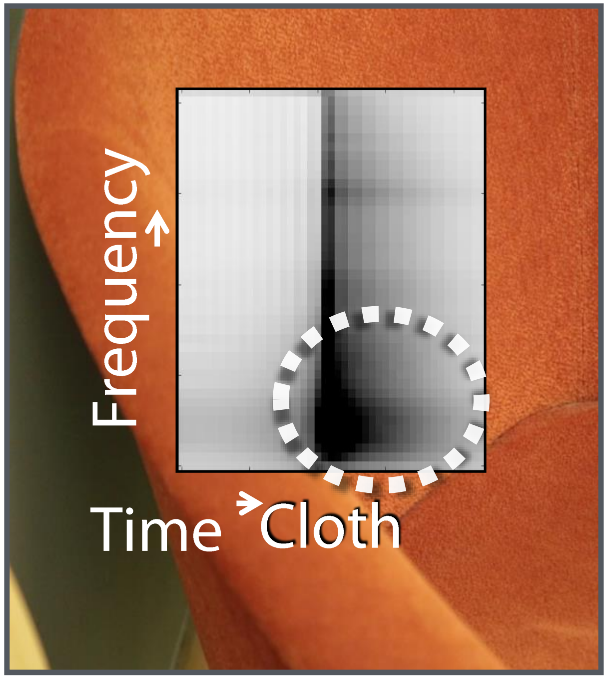


- 40 bandpass filters (+ high/low pass)
- 3 samples per frame (90 Hz)

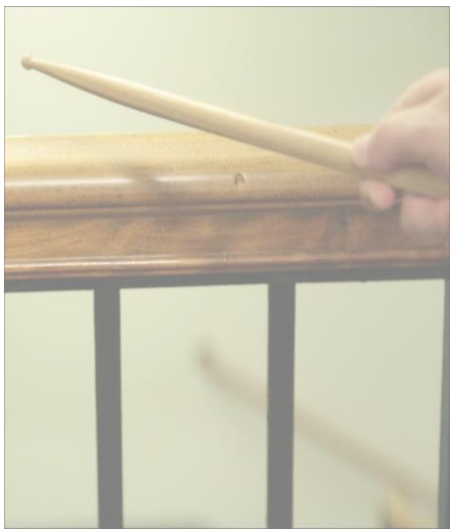
Can we predict material properties from sound?



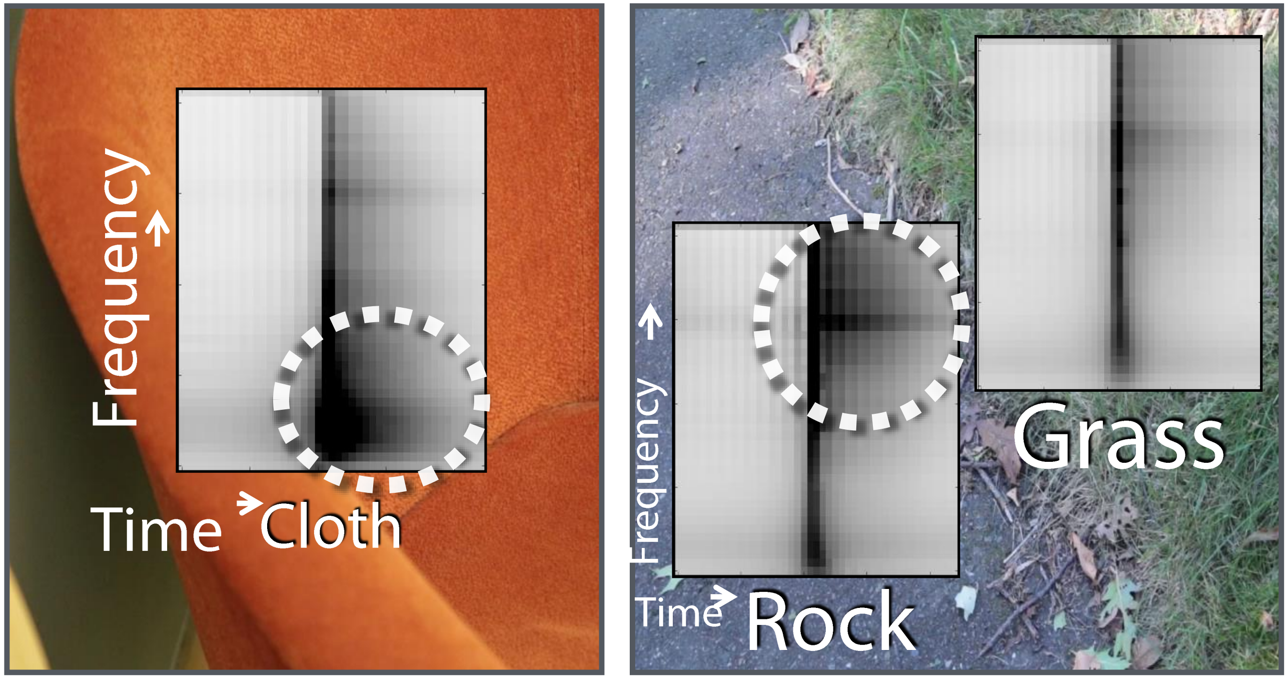
Mean sound features per category



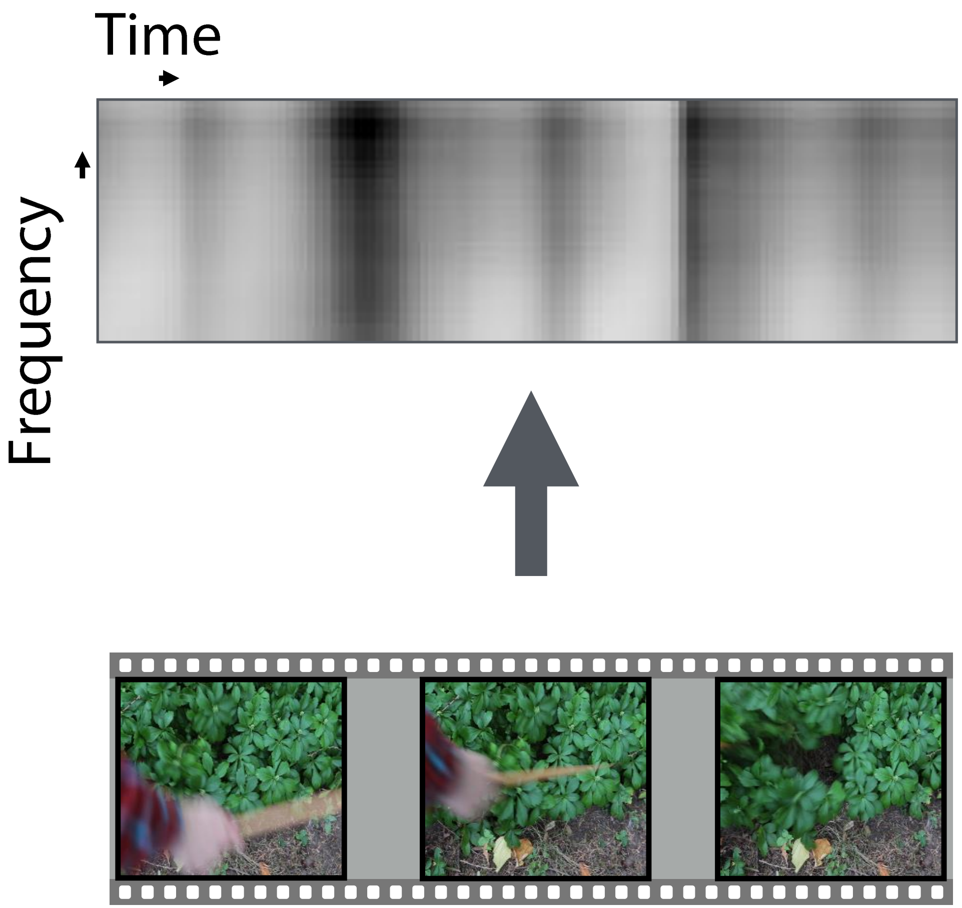
Can we predict material properties from sound?



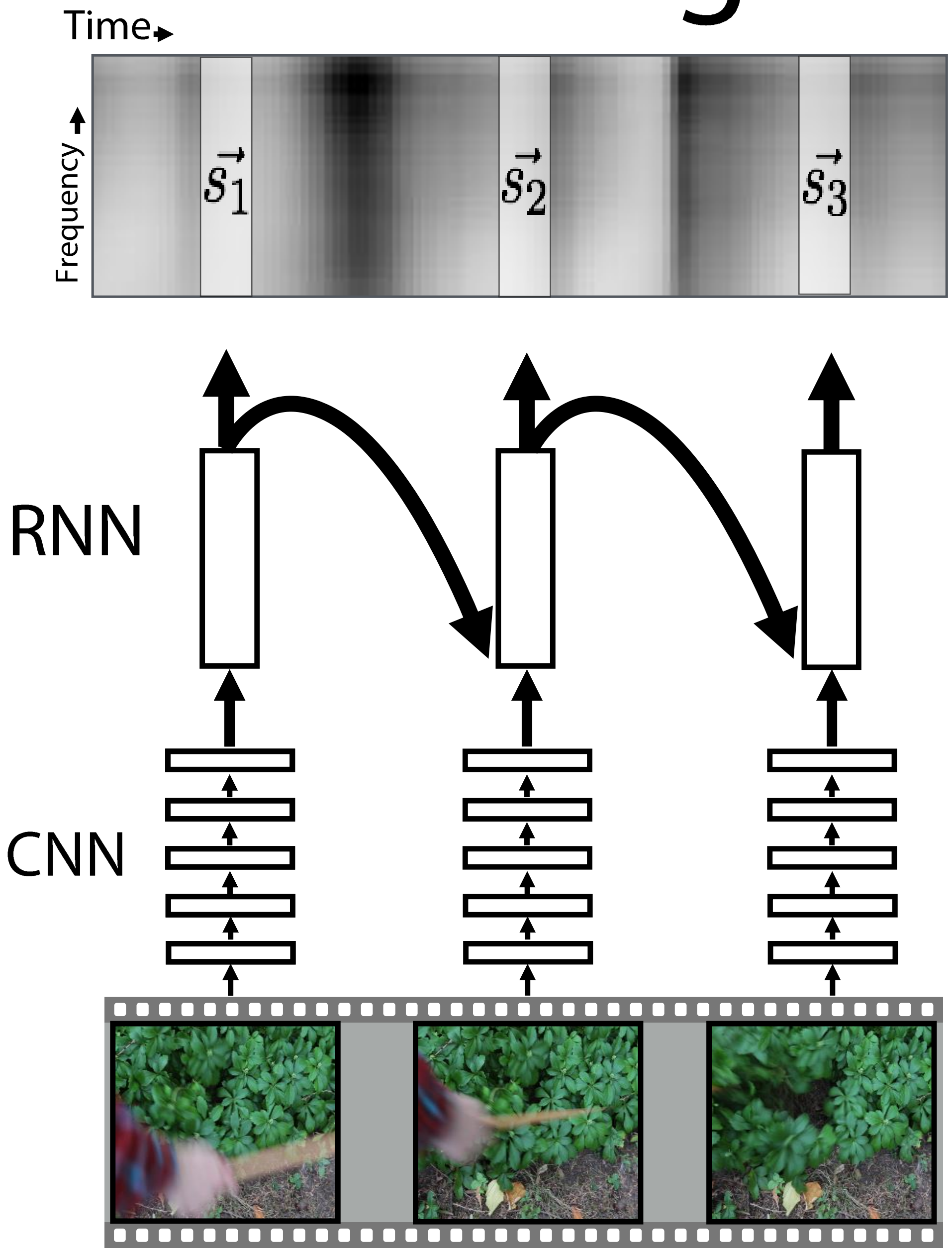
Mean sound features per category



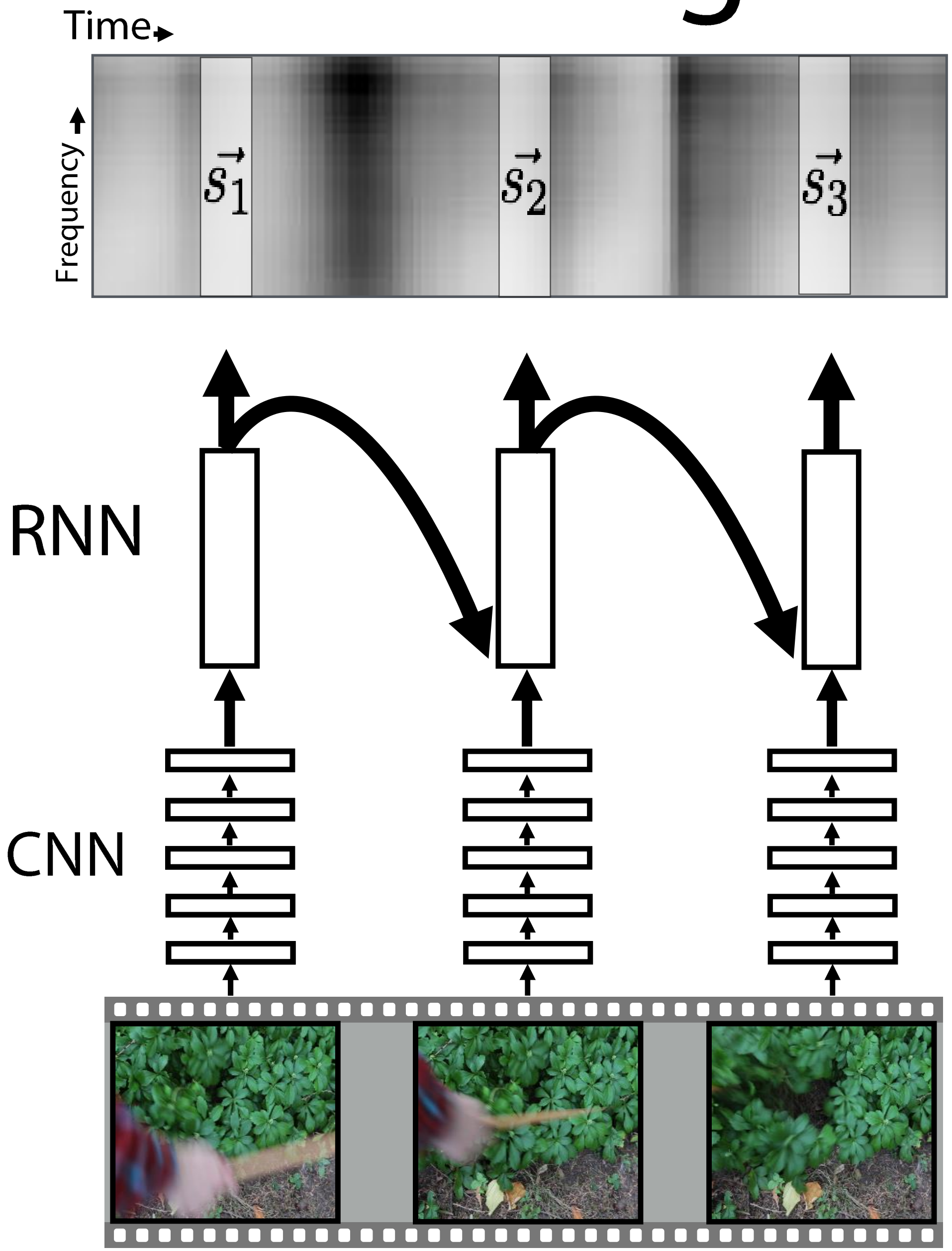
Predicting audio features



Predicting audio features



Predicting audio features



Regression loss

Ground truth

$$\sum_{t=1}^{T} \rho(\|\vec{s}_t - \tilde{\vec{s}}_t\|)$$

where $\rho(r) = \log(\epsilon + r^2)$

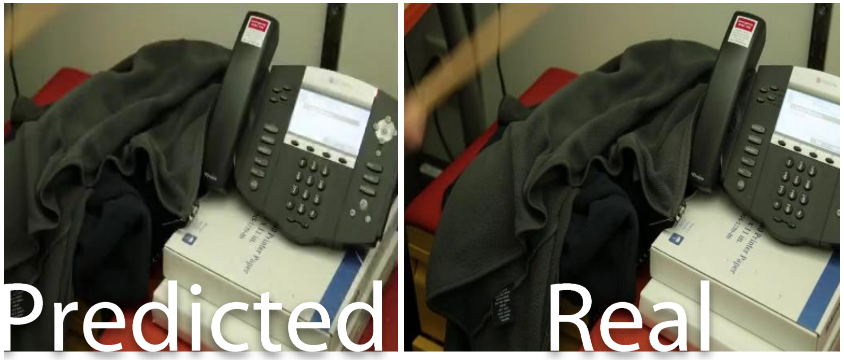
- 3D CNN in time domain
- Pretrain from ImageNet
- Long short-term memory

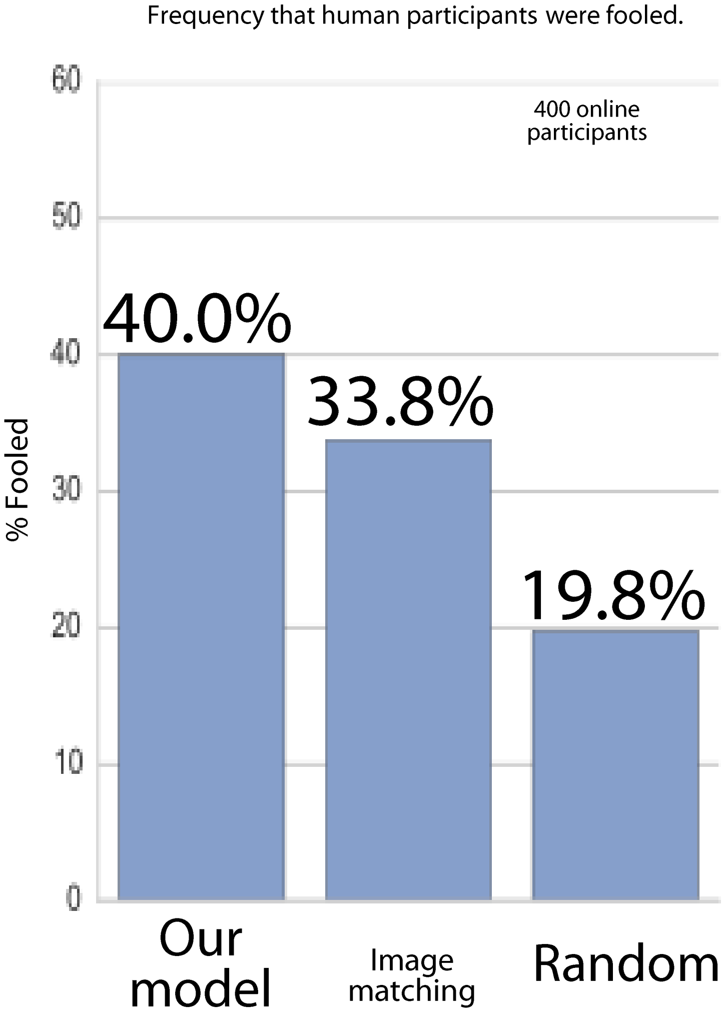
Real-or-fake study

Real-or-fake study

Real-or-fake study

Real-or-fake study





Adding soundtracks to silent videos

Predicted
cochleagram

Predicted sound


Predicted sound
Predicted cochleagram

Input video

Transferred audio clips


Predicted sound

Input video

Transferred audio clips


Predicted sound
Predicted cochleagra m

Input video

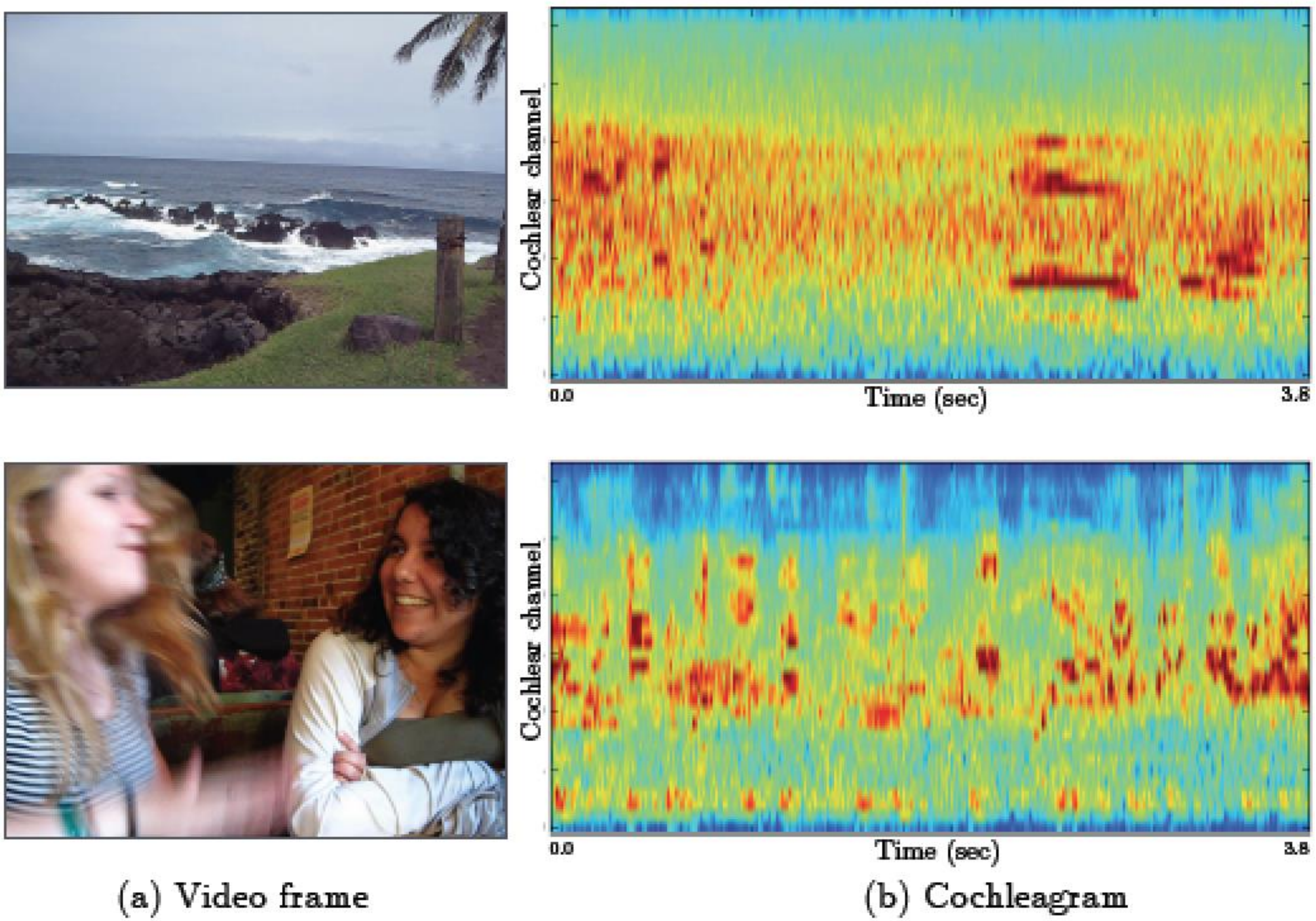
Transferred audio clips

Predicted
cochleagra
m

Predicted sound


Predicted sound
Predicted cochleagra m

Ambient sound



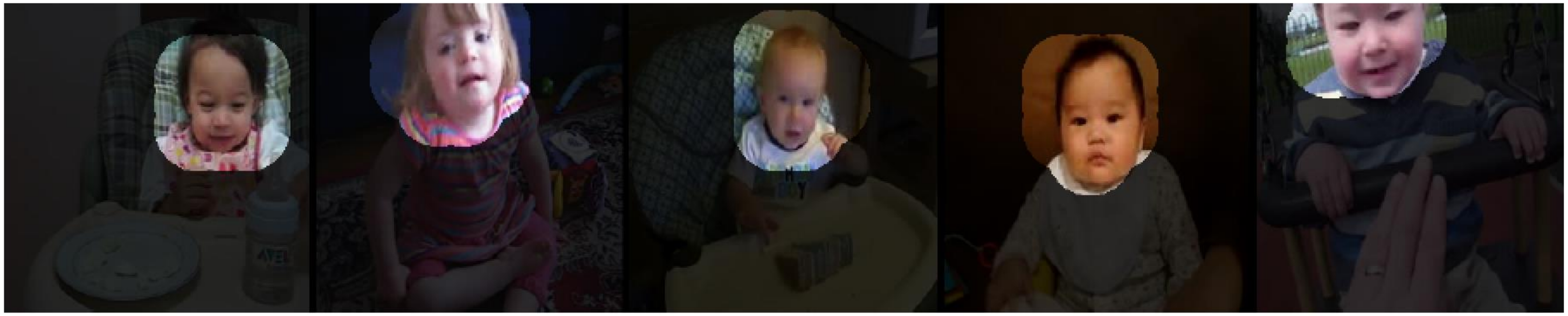
(a) Video frame (b) Cochleagram

99 waterfall

99 waterfall

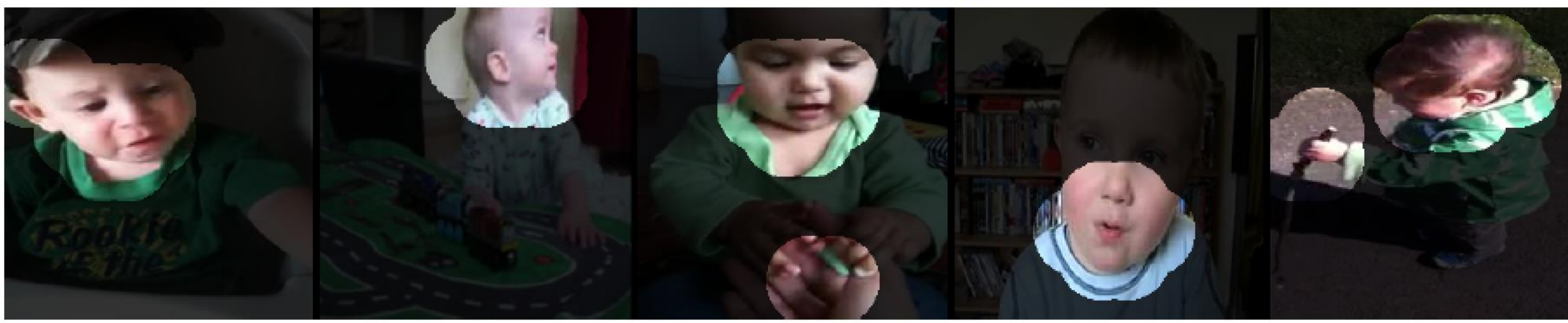
194 crowd

111 baby



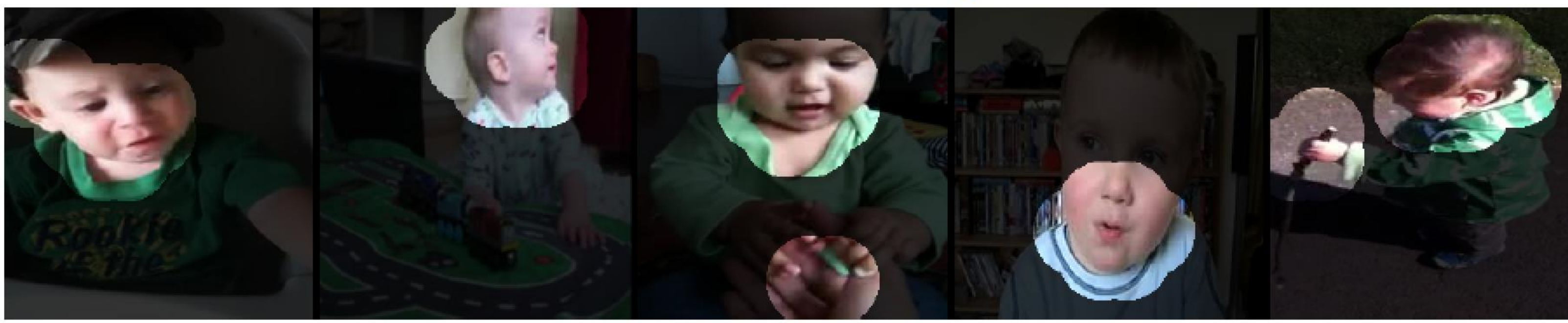
111 baby

171 baby

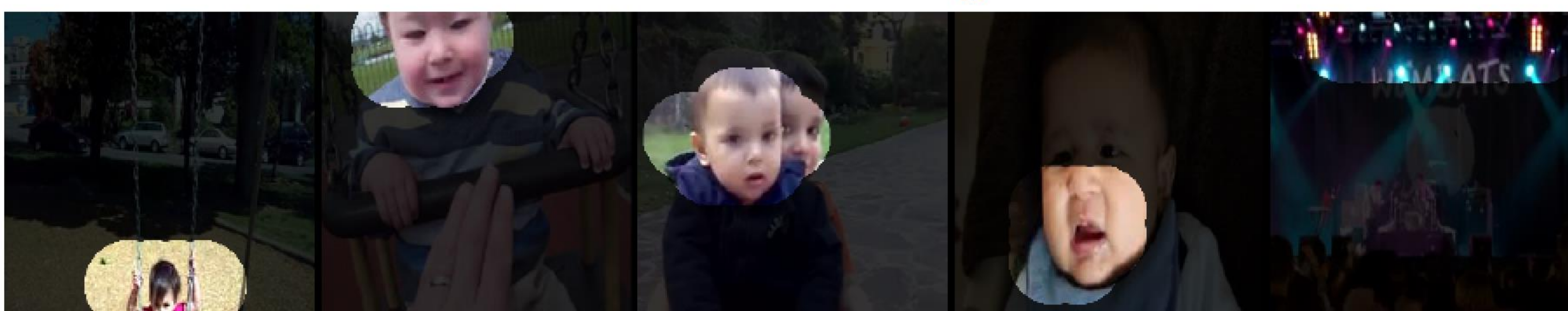


111 baby

171 baby



153 baby



Neuron visualizations of the network trained by sound

Learning about the world by hitting things with a drumstick and listening

- Sound is a ubiquitous training signal
- Predicted sounds convey material properties
- Objects make characteristic noises