Object Recognition

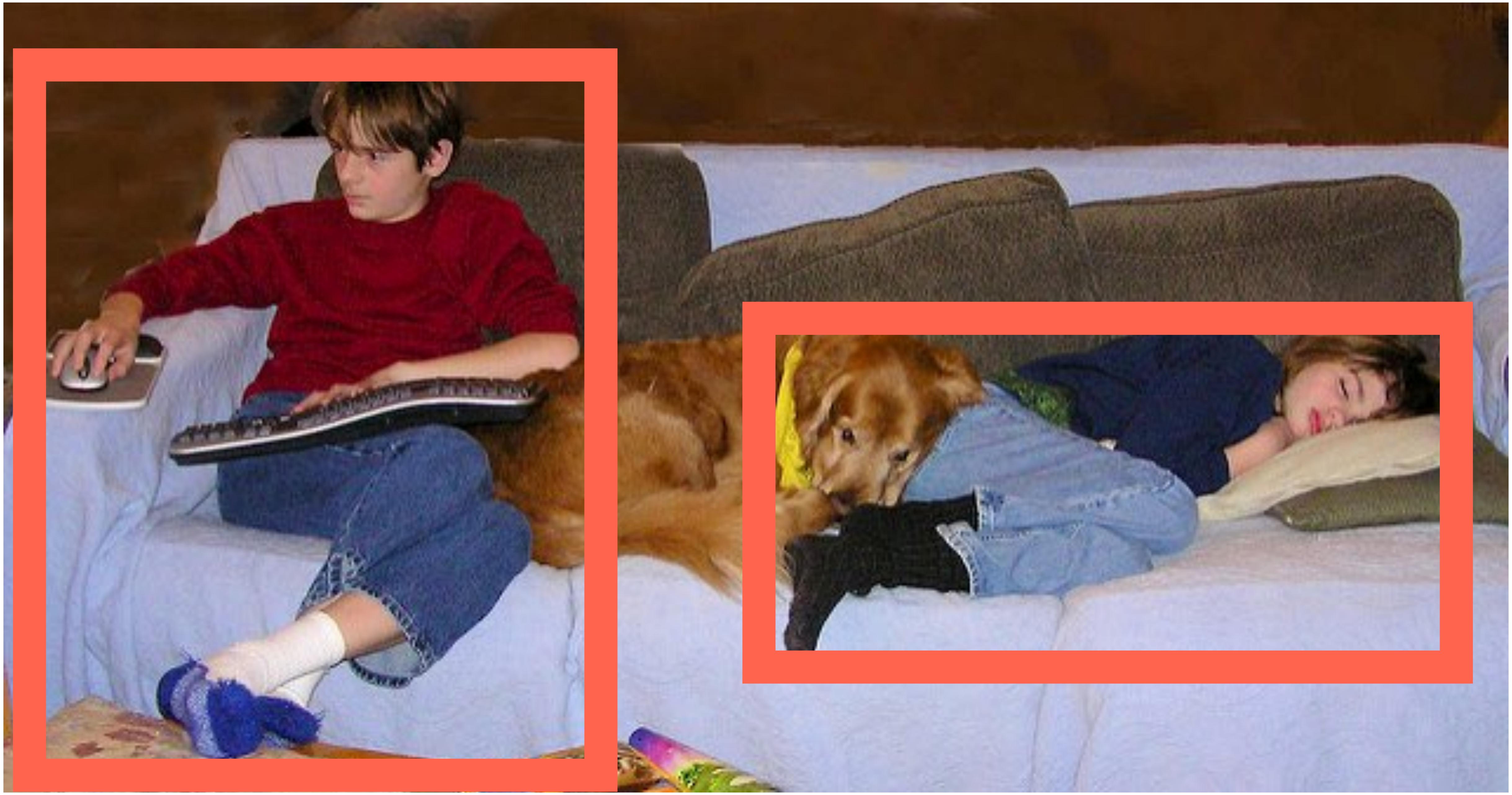
Computer Vision
Fall 2018
Columbia University

Project Proposals and Homework 3

- How’s it going?
- Questions?

Quick Experiment

- Get pen and paper
- Draw a coffee cup



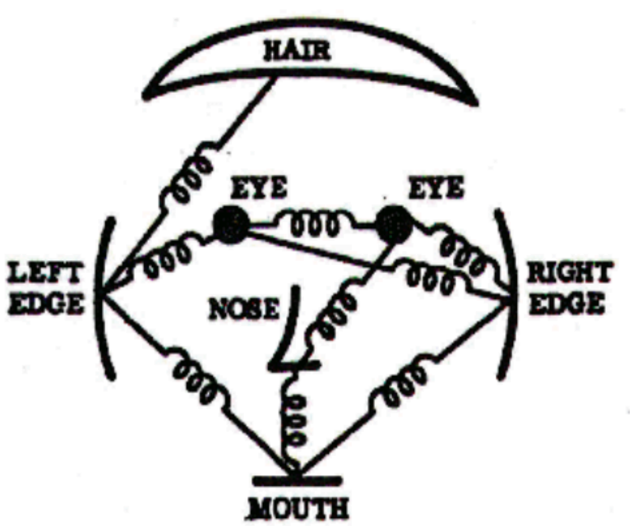
Two Extremes of Vision

Extrapolation problem
Generalization
Diagnostic features

Interpolation problem
Correspondence
Finding the differences

1 | 10 | 10^2 | 10^3 | 10^4 | 10^5 | 10^6 | ∞

Number of training samples



Slide credit: Aude Oliva

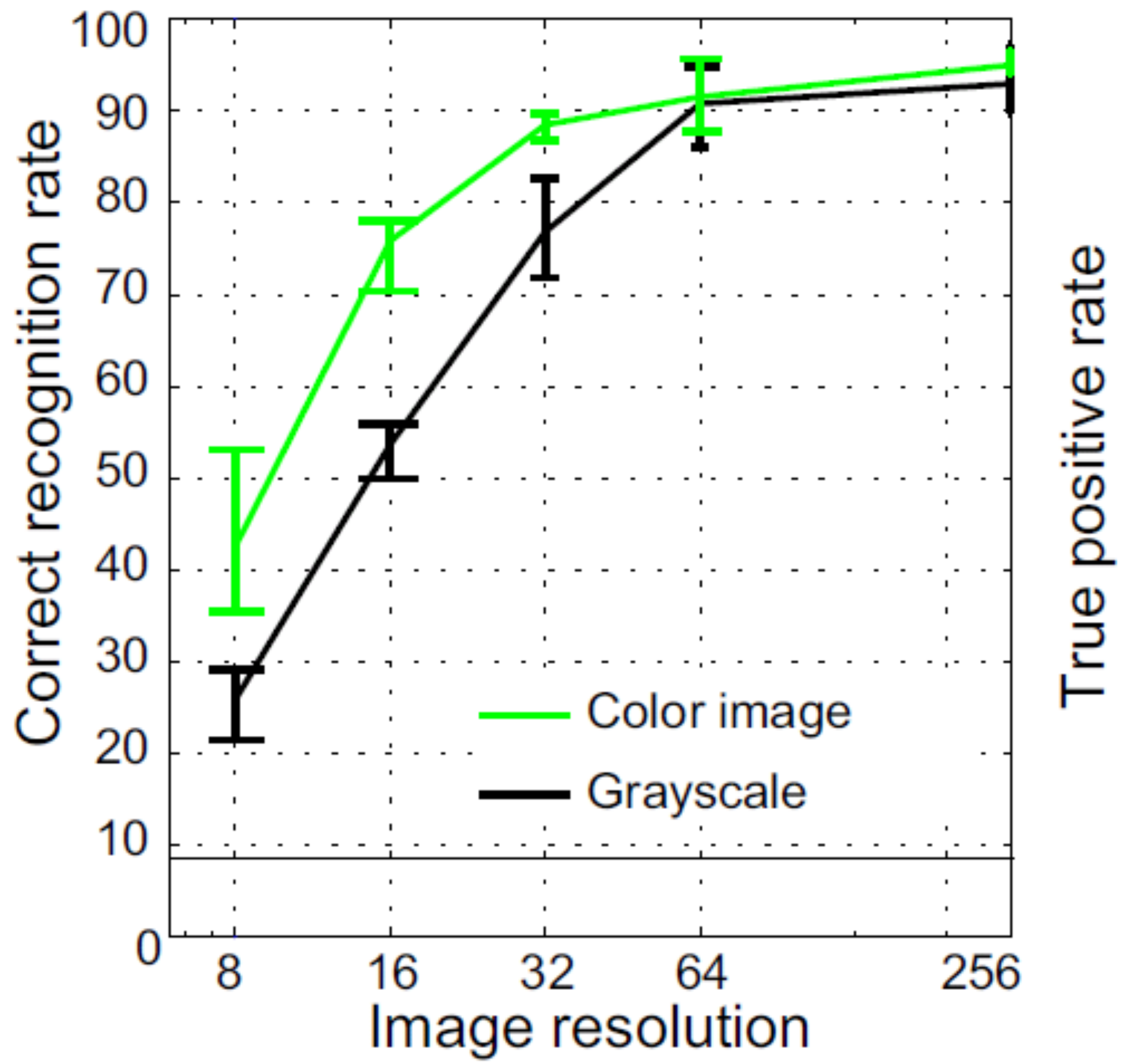
Tiny Images

80 million tiny images: a large dataset for non-parametric object and scene recognition

Antonio Torralba, Rob Fergus and William T. Freeman. PAMI 2008.
http://groups.csail.mit.edu/vision/TinyImages/

c) Segmentation of 32x32 images

Given a benchmark, resolution and human scene recognition accuracy increase to a limit

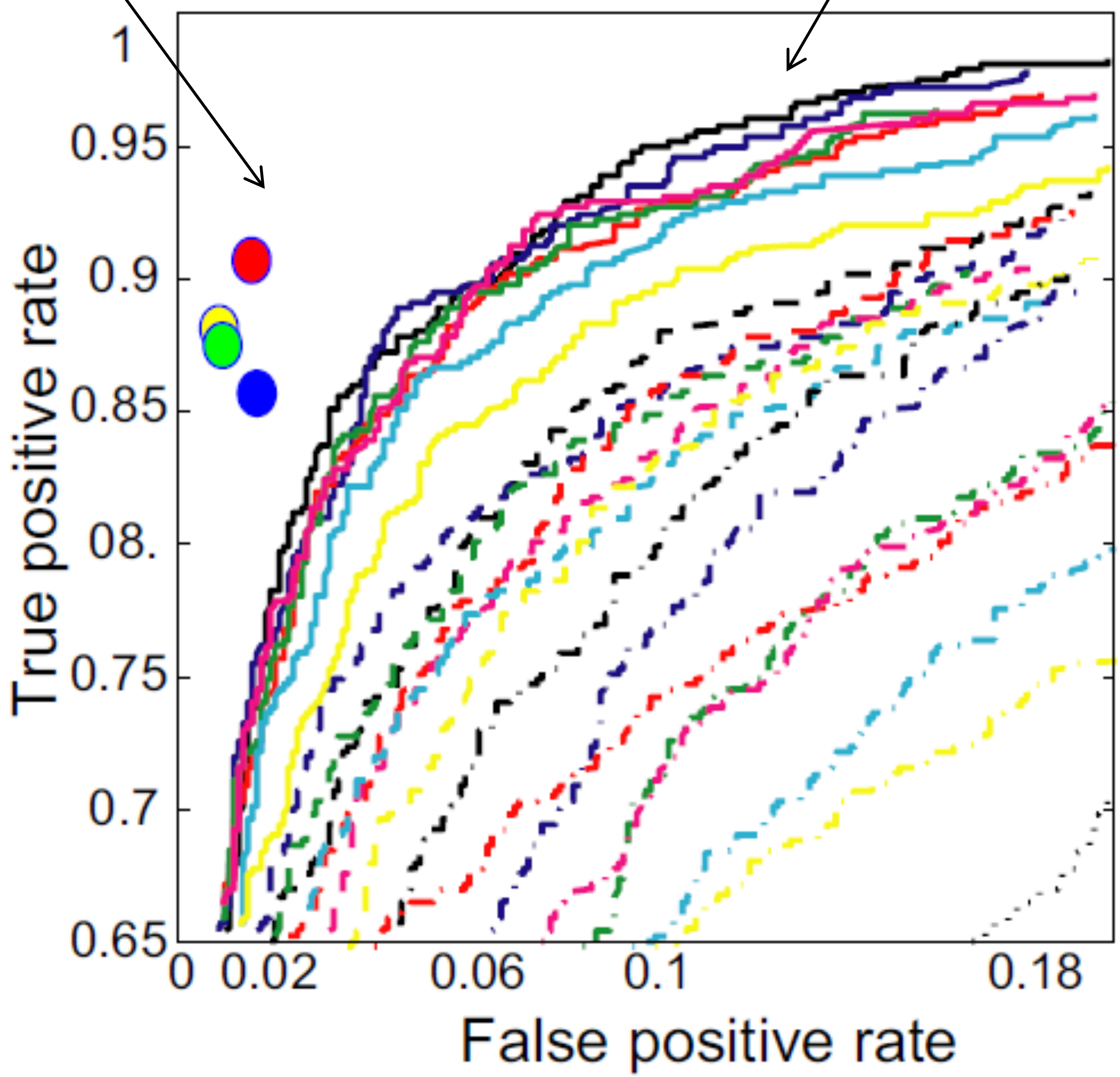


Torralba et al.

Humans vs. Computers: Car Classification

Humans for 32 pixel tall images

Various computer vision algorithms for *full* resolution images



Example test images

Torralba et al.

Powers of 10

Number of images on my hard drive: 10^6

Number of images seen during my first 10 years: 10^8

(3 images/second * 60 * 60 * 16 * 365 * 10 = 630,720,000)

Number of images seen by all humanity: 10^{20}

106,456,367,669 humans[1] * 60 years * 3 images/second * 60 * 60 * 16 * 365 =

1 from http://www.prb.org/Articles/2002/HowManyPeopleHaveEverLivedonEarth.aspx

Number of photons in the universe: 10^{88}

Number of all 32x32 images: 10^{7373}

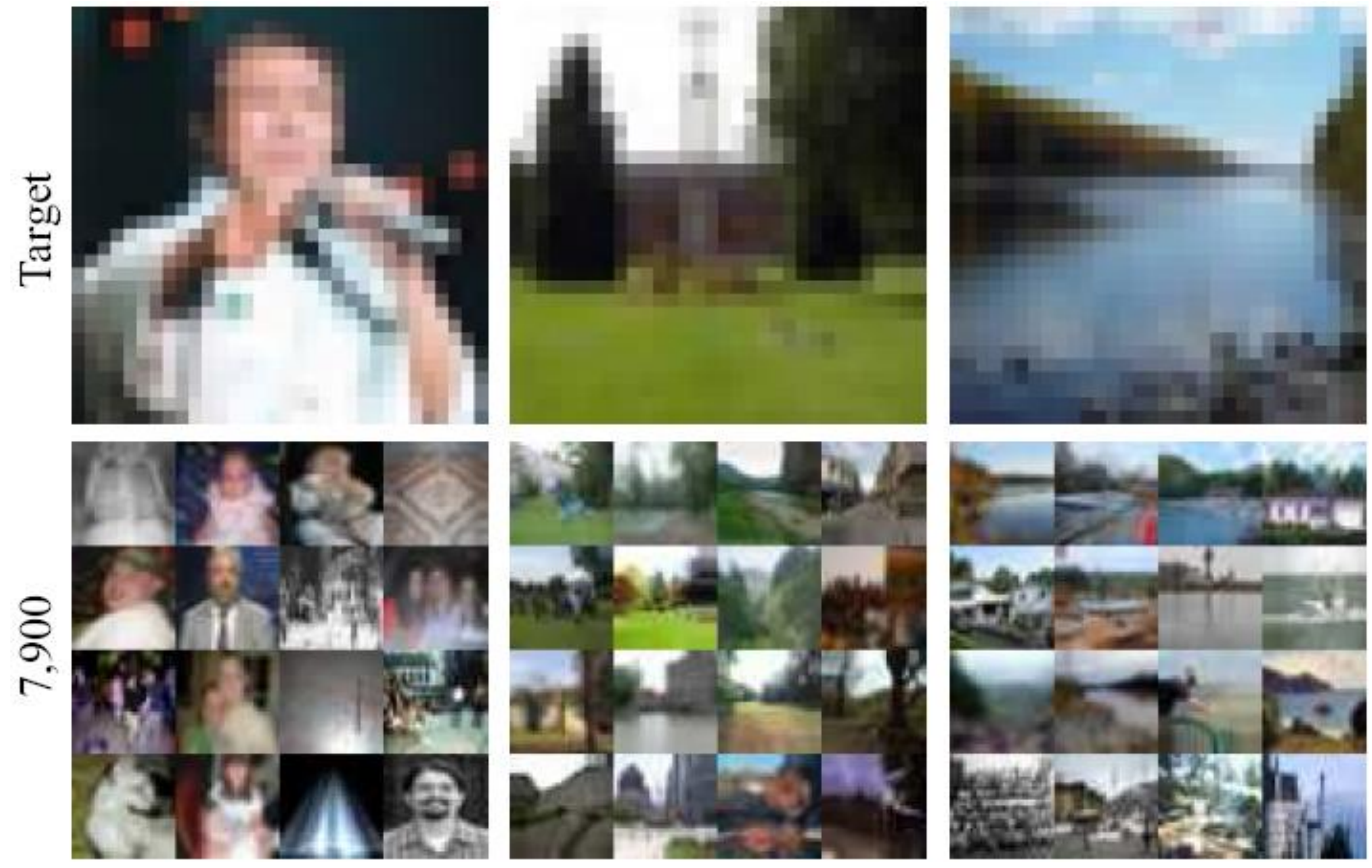
$256^{32*32*3} \sim 10^{7373}$

But not all scenes are so original

Lots

Of

Images

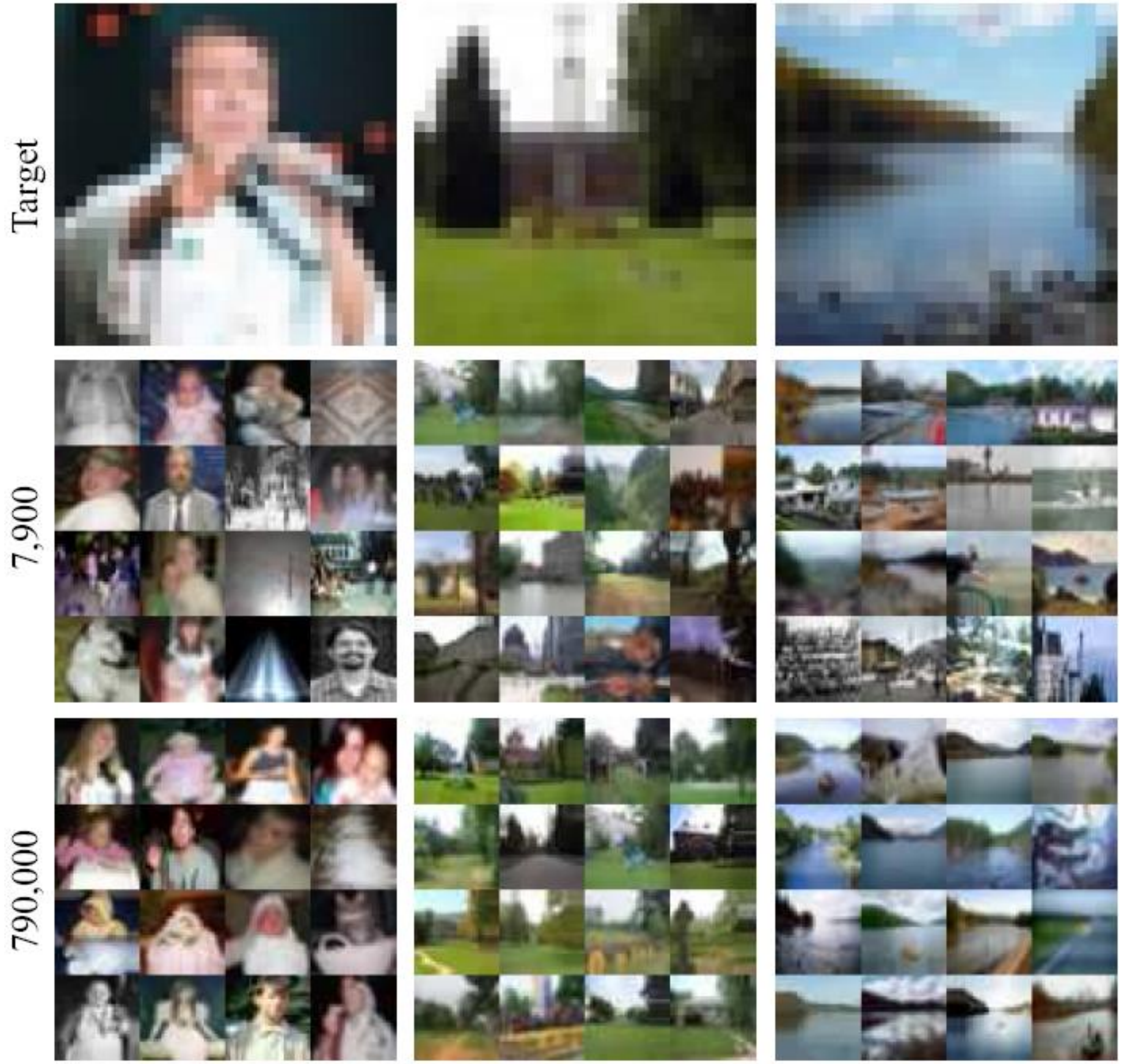


A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

Of

Images

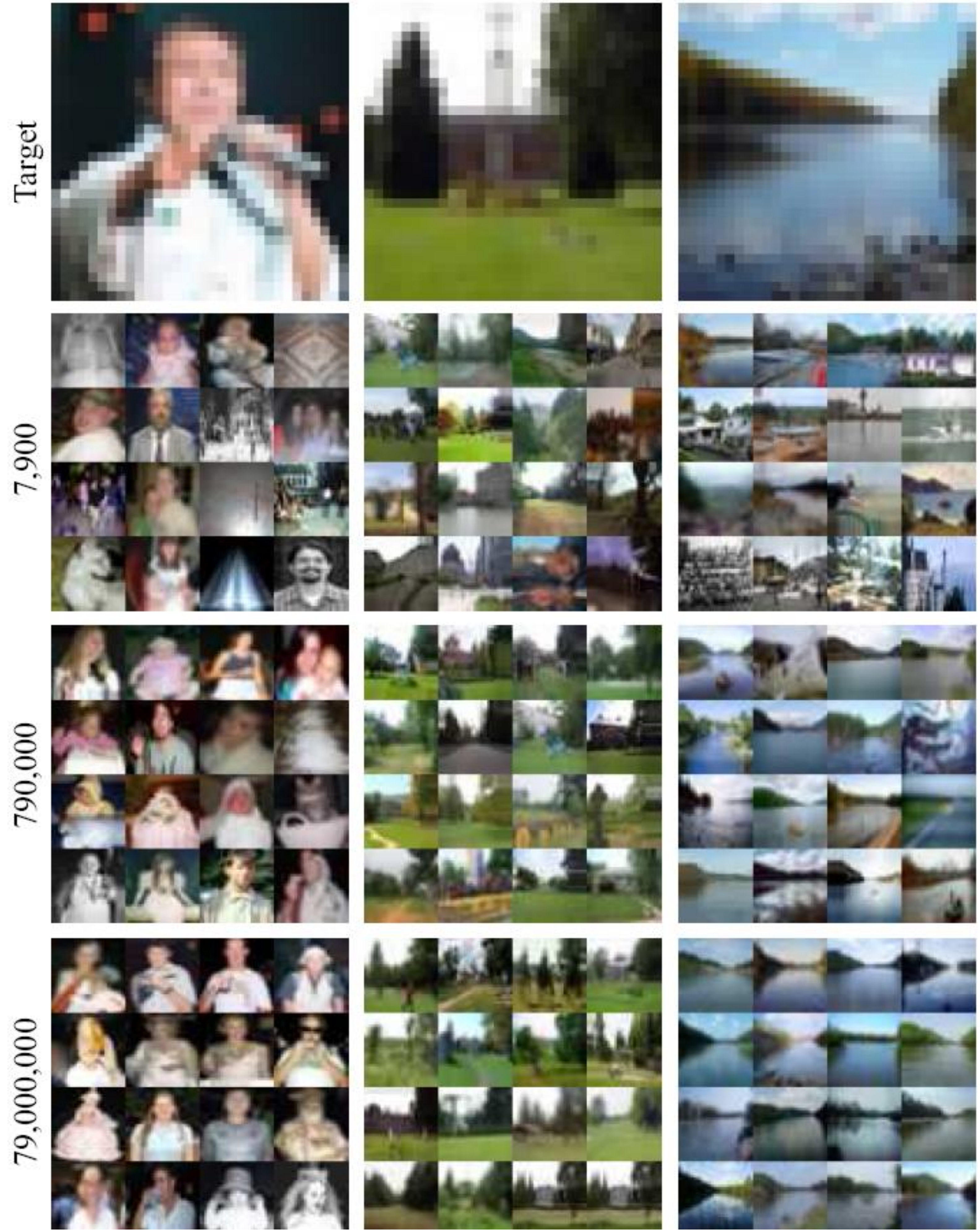


A. Torralba, R. Fergus, W.T.Freeman. PAMI 2008

Lots

Of

Images



Application: Automatic Colorization

Input

Color Transfer

Color Transfer

Matches (gray)

Matches (w/ color)

Avg Color of Match

Application: Automatic Colorization

Input

Color Transfer

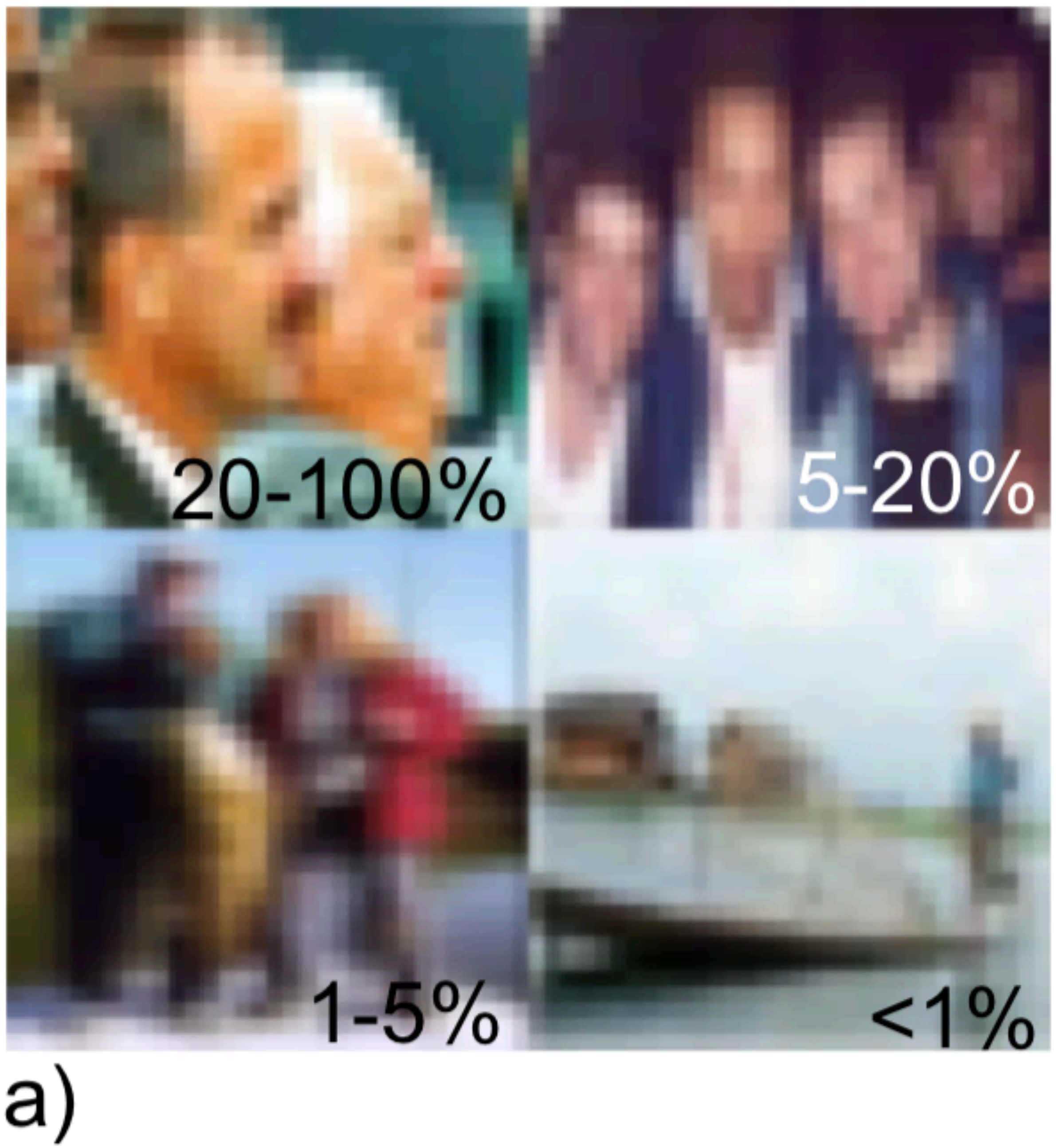
Color Transfer

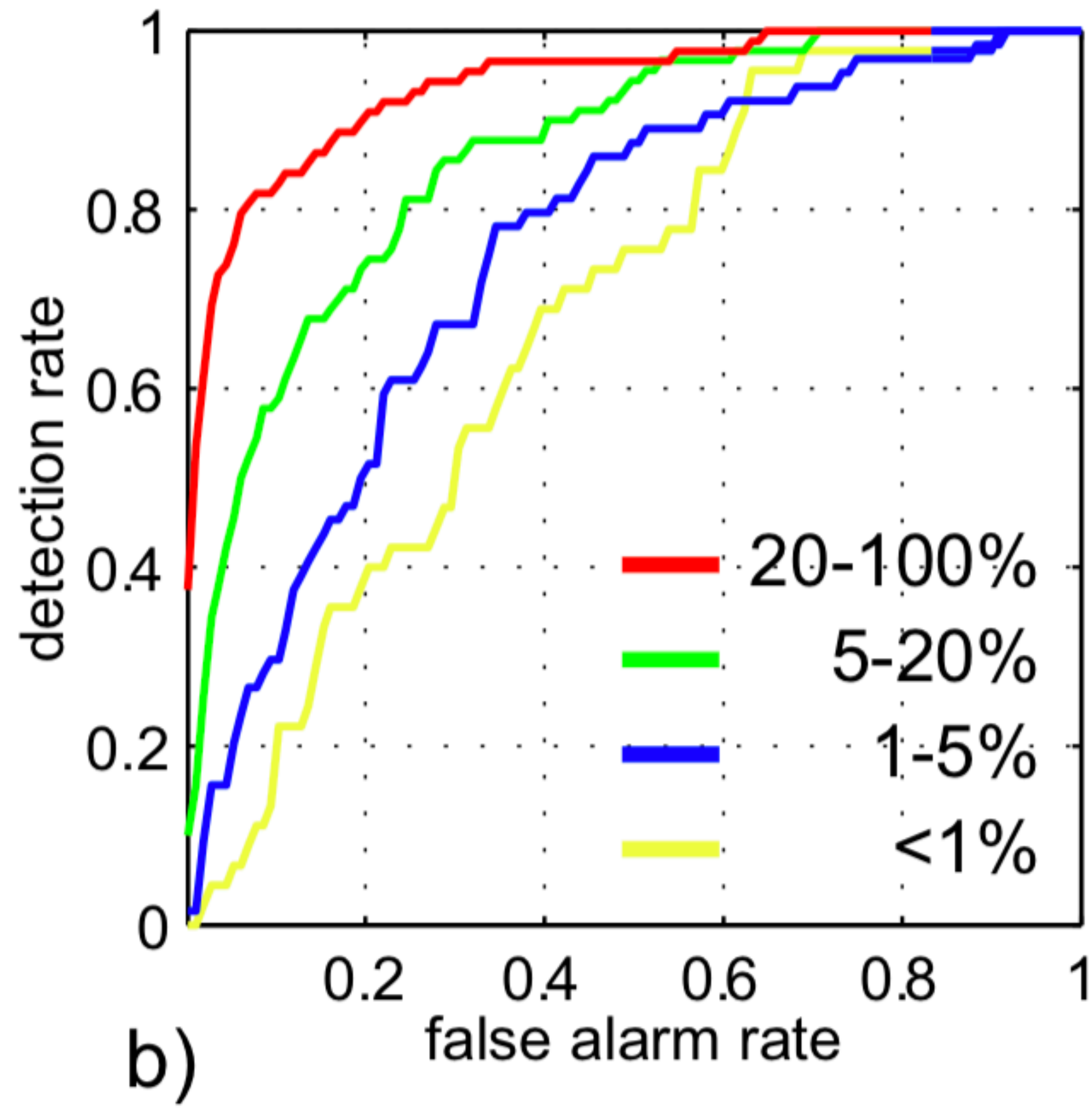
Matches (gray)

Matches (w/ color)

Avg Color of Match

Person Recognition





Exploring the Limits of Weakly Supervised Pretraining

Laurens van der Maaten

ECCV 2018

Dhruv Mahajan

Ross Girshick

Vignesh Ramanathan

Kaiming He

Manohar Paluri

Yixuan Li

Ashwin Bharambe

facebook
Artificial Intelligence Research

https://arxiv.org/pdf/1805.00932.pdf

Research question

Can we use large amounts of weakly supervised images for pretraining?

Highlights

- We pretrain models by predicting relevant hashtags for images
- We pretrain models to predict 17.5K hashtags for 3.5B images
- After finetuning, we beat the state-of-the-art on, e.g., ImageNet

Hashtag Supervision

- It is easy to get billions of public images and hashtags
- Hashtags are more structured than captions
- Hashtags were often assigned to make images "searchable"

#cheesecake #birthday

Hashtag Supervision

- But hashtags are not perfect supervision
- Some hashtags are not visually relevant
- Other hashtags are not in the photo
- And there are many false negatives

#cat #travel #thailand #family
#building #fence #...

3,500,000,000 images!

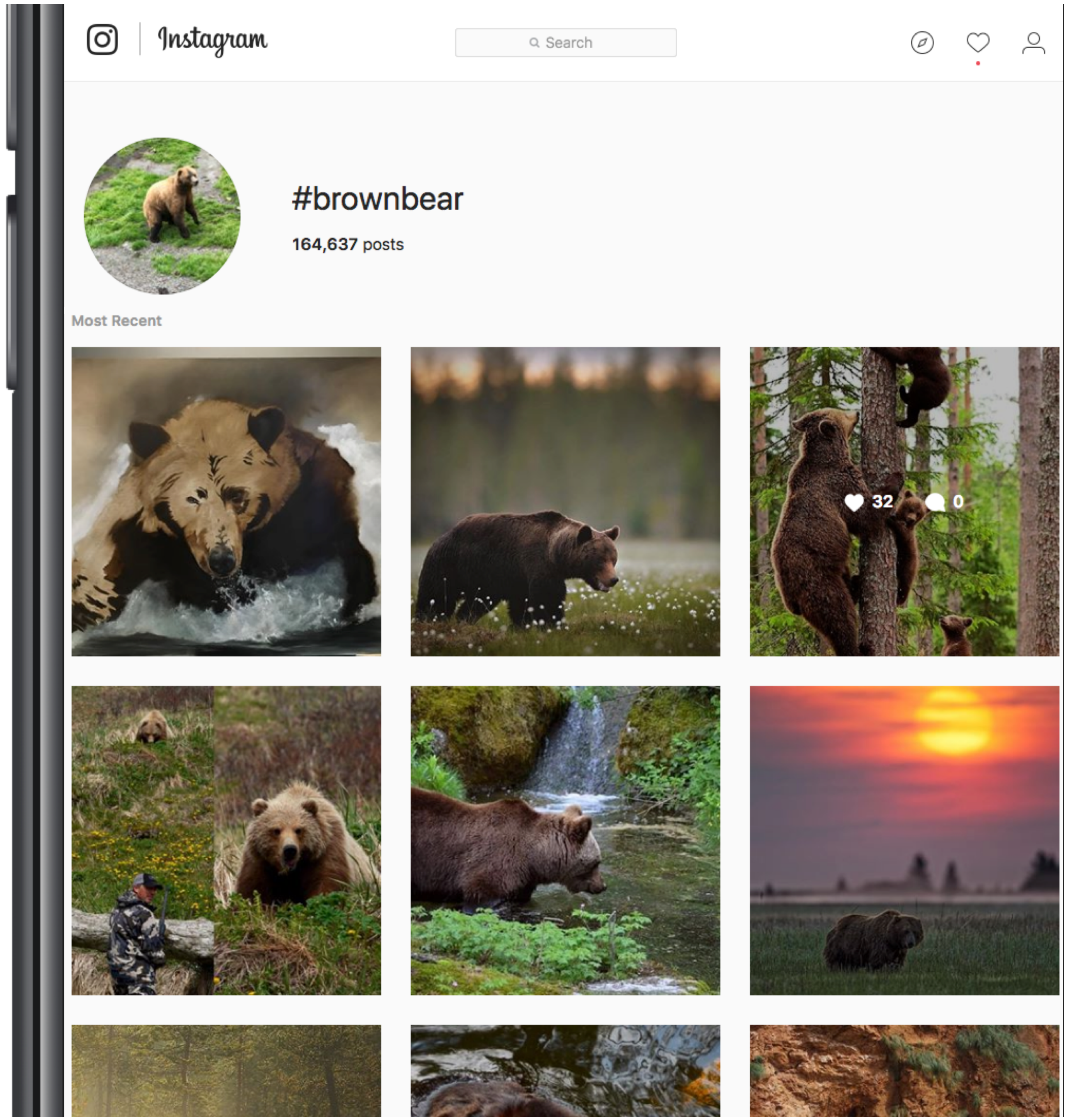
Experiments

- Select a set of hashtags
- Download all public Instagram images that has at least one of these hashtags
- Use WordNet synsets to merge hashtags into canonical form (merge **#brownbear** and **#ursusarctos**)
- The final list has 17,517 hashtags

1	aar
2	aardvark
3	aardwolf
4	aba
5	abaca
6	abacus
7	abalone
8	abatis
9	abaya
10	abbey
11	abele
12	abelia
13	abies
14	abila
15	abm
16	abortus
17	abronia
18	absinth
19	absinthe
20	abstraction
21	abstractionism
22	abutilon
23	abutment
24	abyss
25	abyssinian
26	acacia
27	acaciadealbata
28	academy
29	acalypha
30	acanthaceae
31	acanthurus
32	acanthus
33	acanthusmollis
34	acapulcogold
35	acarus
36	accelerator
37	accelerometer
38	access
39	accessory
40	accident
41	accipiter
42	accipiternisus
43	accipitridae
44	accommodation
45	accompaniment
46	accordion
47	accoutrement
48	accumulator
49	ace
50	aceofclubs
51	aceofdiamonds
52	aceofhearts
53	aceofspades
54	acer
55	acerjaponicum
56	acerola
57	acerpalmatum
58	acerrubrum
59	acetaminophen
60	acetate
61	acheron
62	acherontia
63	acherontiaatropos
64	achillea
65	achilleamillefolium
66	achimenes
67	acid
68	acidophilus
69	acinonyxjubatus
70	acinus
71	ackee
72	aconcagua
73	aconite
74	aconitum
75	acorn
76	acornsquash
77	acousticguitar
78	acoustics
79	acrididae
80	acrobates
81	acropolis
82	acropora
83	acrylic
84	acrylicpaints
85	actias
86	actiasluna
...	
17474	yurt
17475	zabaglione
17476	zambeziriver
17477	zamboni
17478	zamia
17479	zantac
17480	zantedeschia
17481	zap
17482	zapper
17483	zarf
17484	zea
17485	zebra
17486	zebrafinch
17487	zebrawood
17488	zebu
17489	zero
17490	zeus
17491	zhujiang
17492	ziggurat
17493	zill
17494	zimmerframe
17495	zinfandel
17496	zing
17497	zingiber
17498	zinnia
17499	zipgun
17500	zipper
17501	zither
17502	ziti
17503	ziziphus
17504	zizz
17505	zodiac
17506	zoloft
17507	zombi
17508	zoologicalgarden
17509	zoom
17510	zooplankton
17511	zootsuit
17512	zori
17513	zoysia
17514	zuiderzee
17515	zygnema
17516	zygocactus
17517	zygoptera

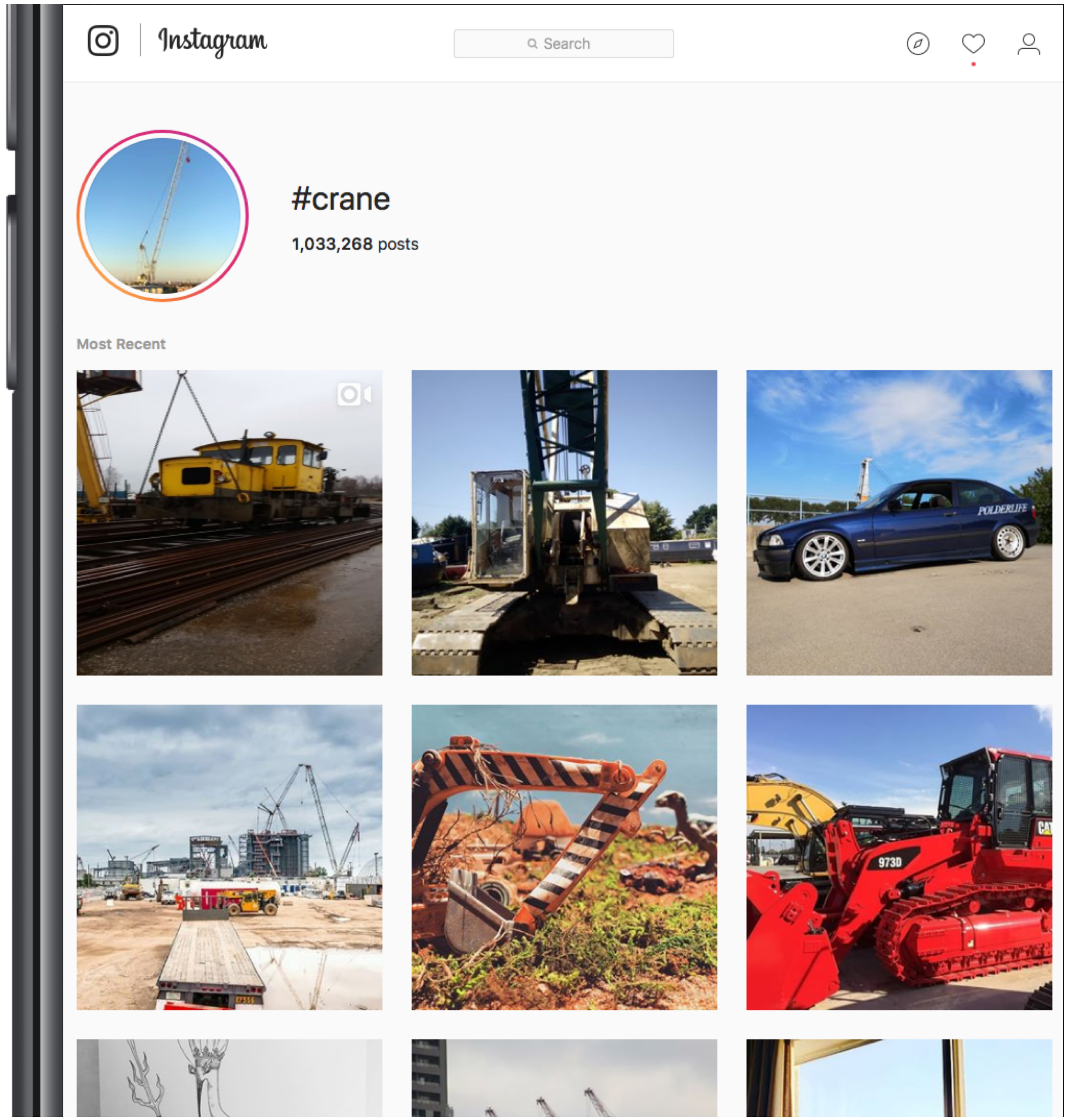
Experiments

- Select a set of hashtags
- Download all public Instagram images that has at least one of these hashtags
- Use WordNet synsets to merge hashtags into canonical form (merge **#brownbear** and **#ursusarctos**)
- Final dataset has ~3.5 **billion** images



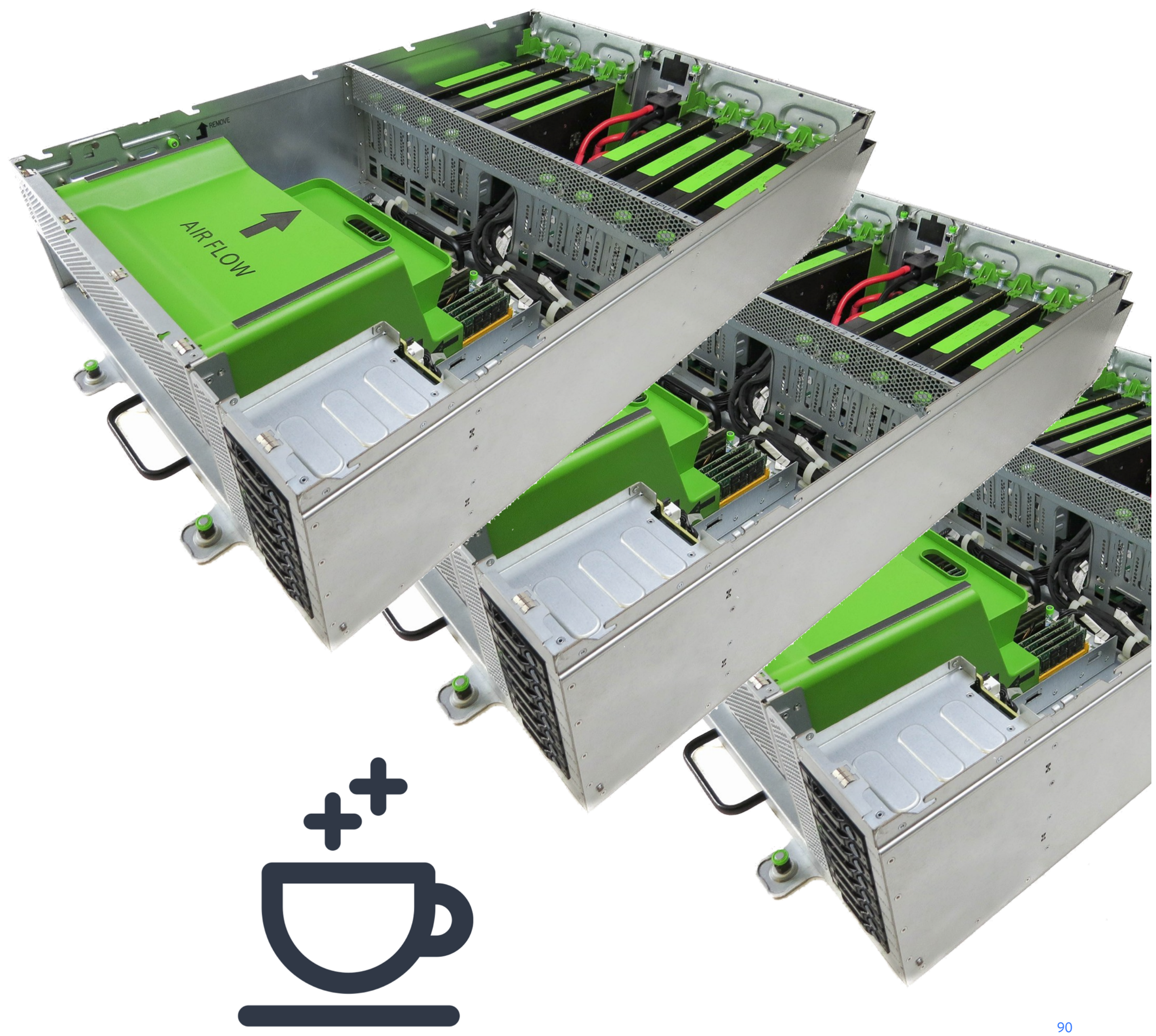
Experiments

- Select a set of hashtags
- Download all public Instagram images that has at least one of these hashtags
- Use WordNet synsets to merge hashtags into canonical form (merge **#brownbear** and **#ursusarctos**)
- Final dataset has ~3.5 **billion** images



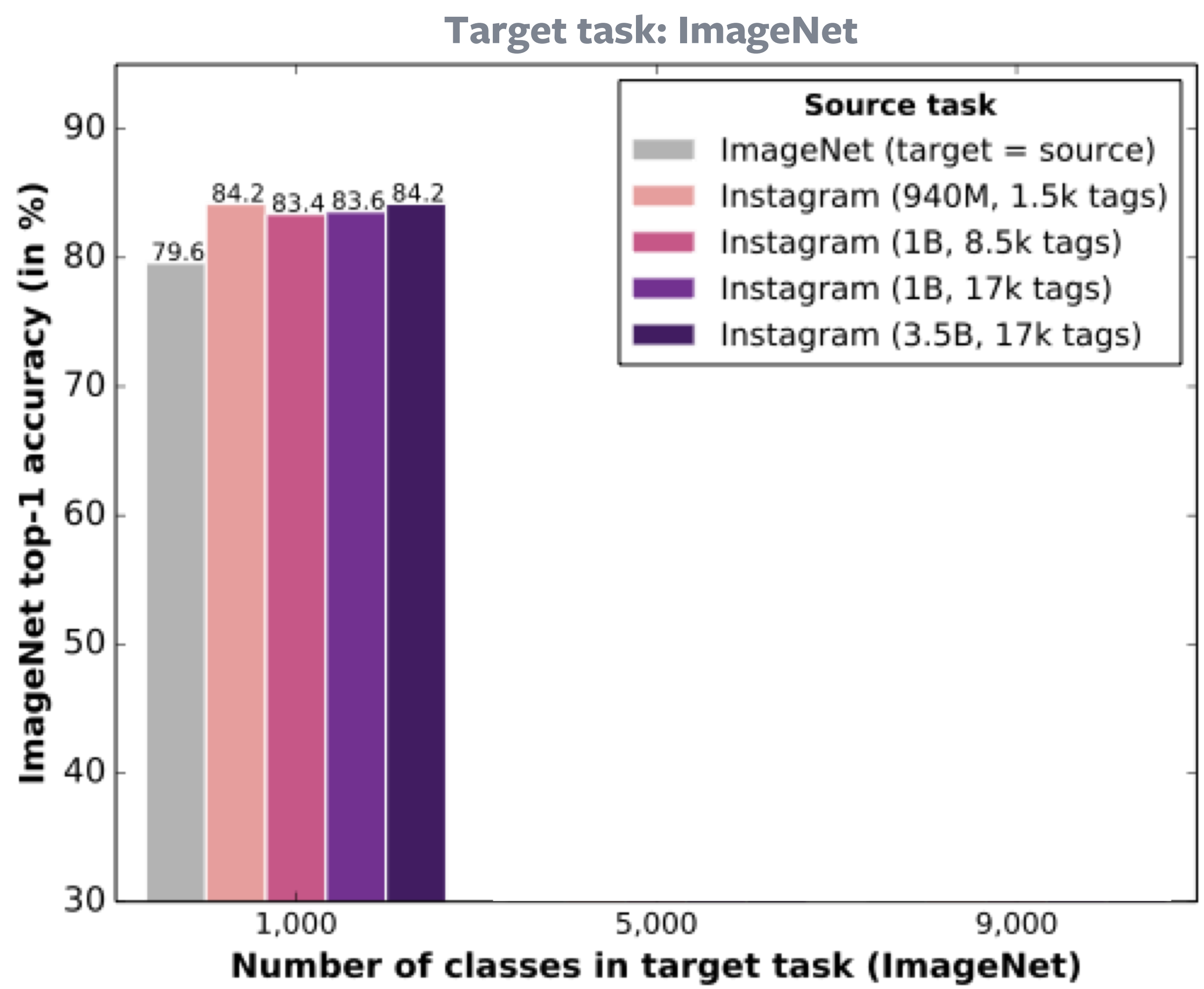
Experiments

- Train ResNeXt-32xCd convolutional networks
- Use *c*-of-*K* vector to represent multiple labels
- Train to minimize multi-class logistic loss
- Distribute training batches across 336 GPUs
- Scale learning rate by batch size (*N=8,064*) after learning rate "warm-up" (Goyal *et al.*, 2017)



Fix Model; Vary Data

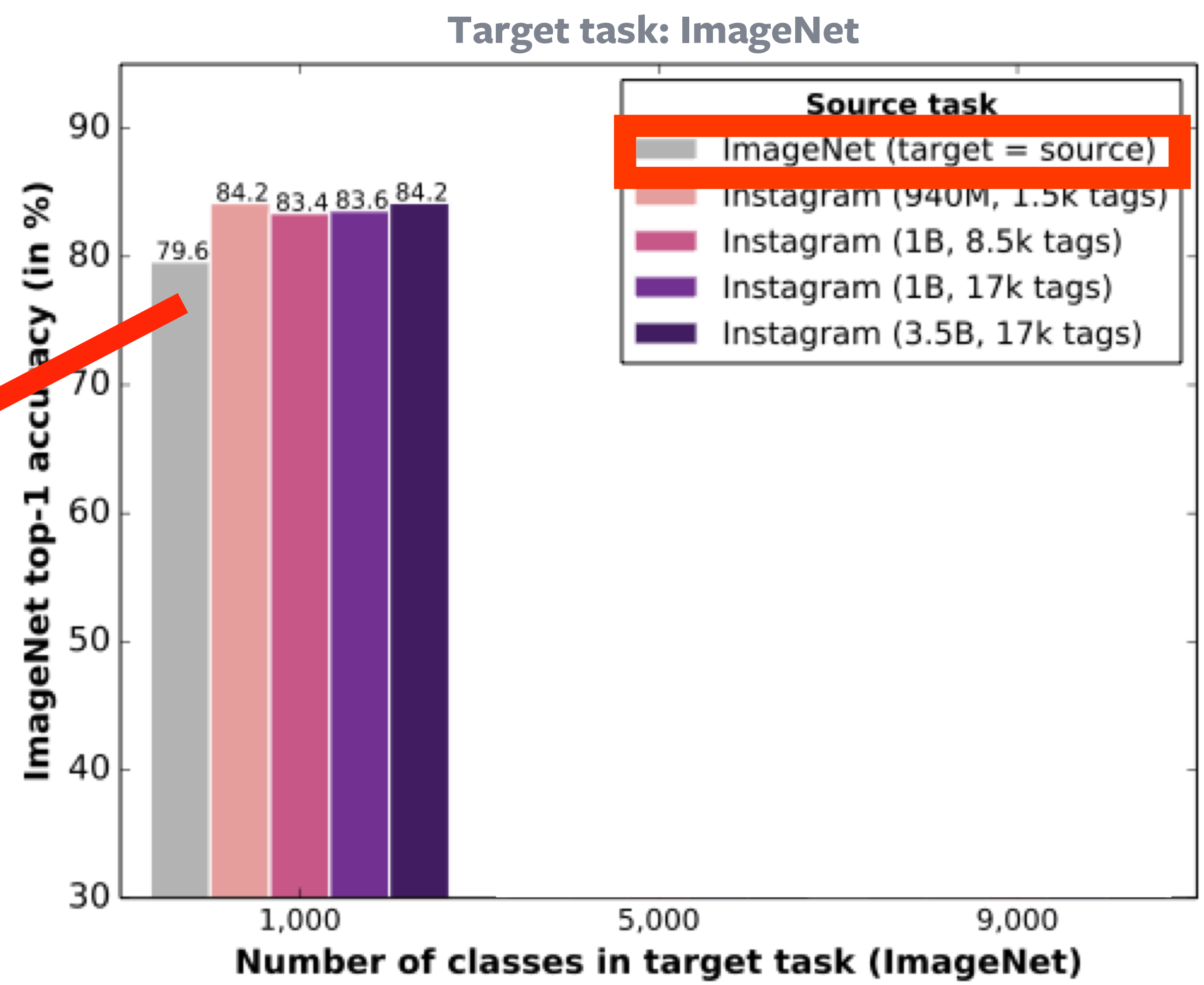
- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet



Fix Model;
Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

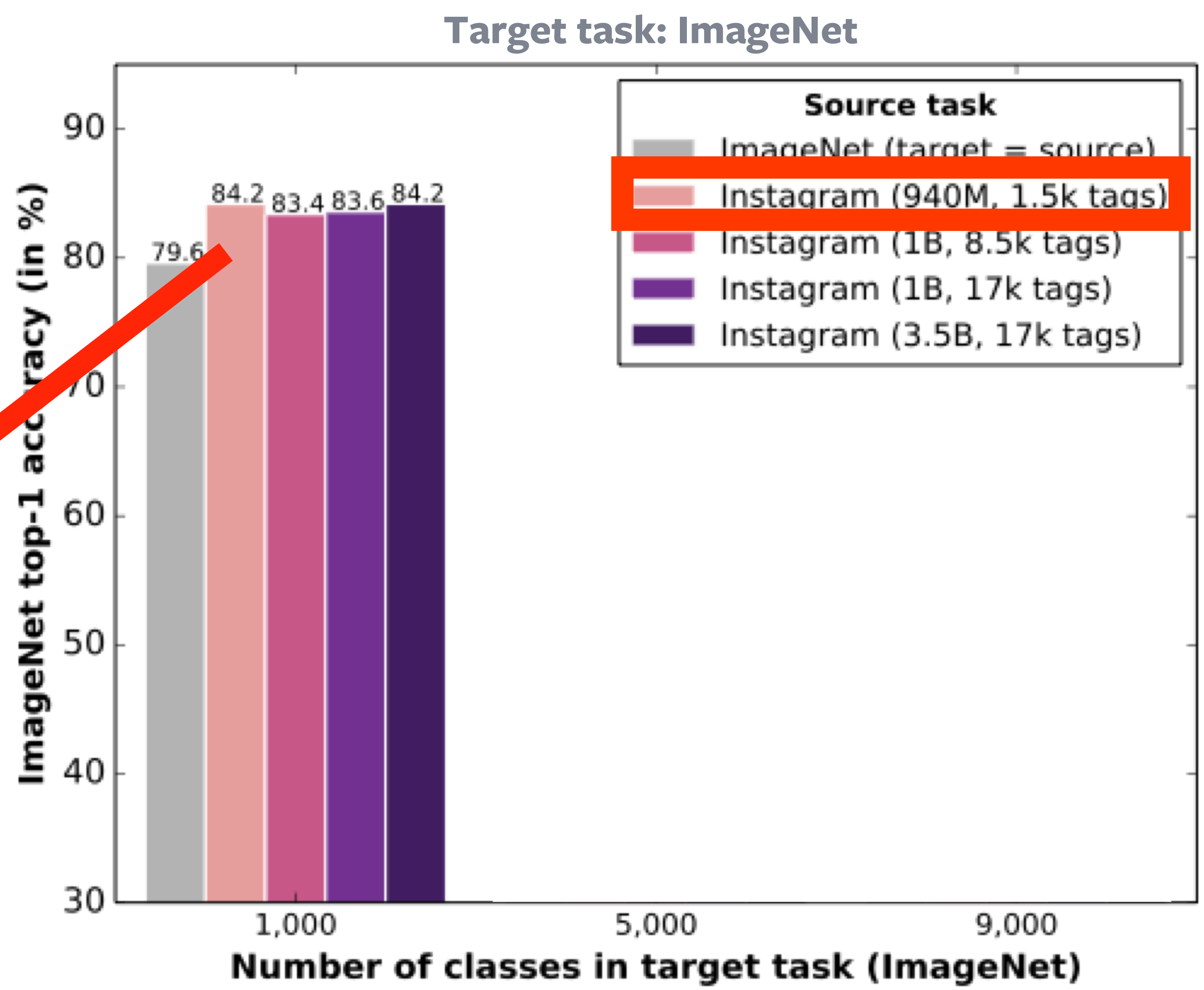
"standard" ImageNet training



Fix Model;
Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

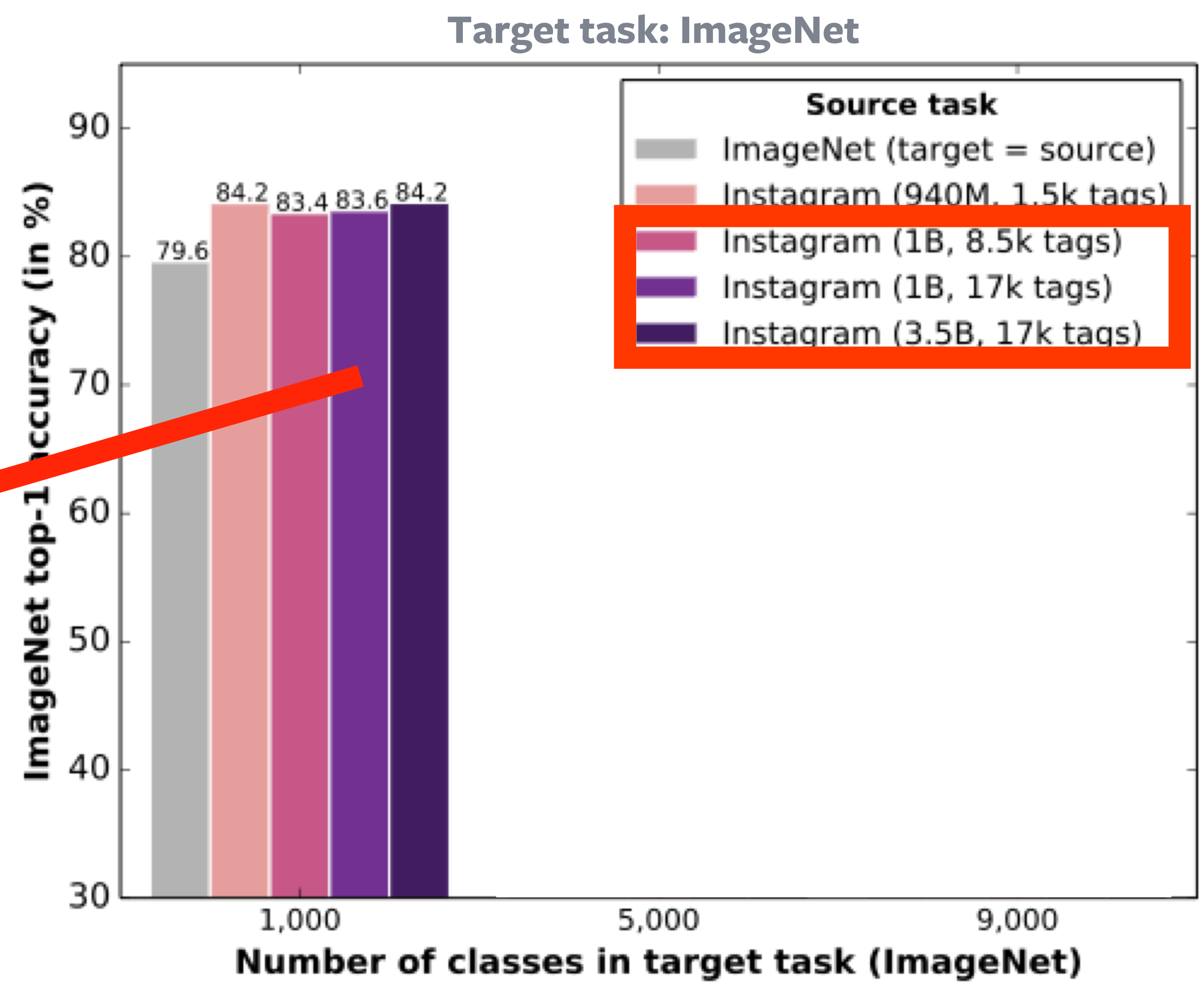
pre-training on 1B Instagram images, selected to match ImageNet classes



Fix Model;
Vary Data

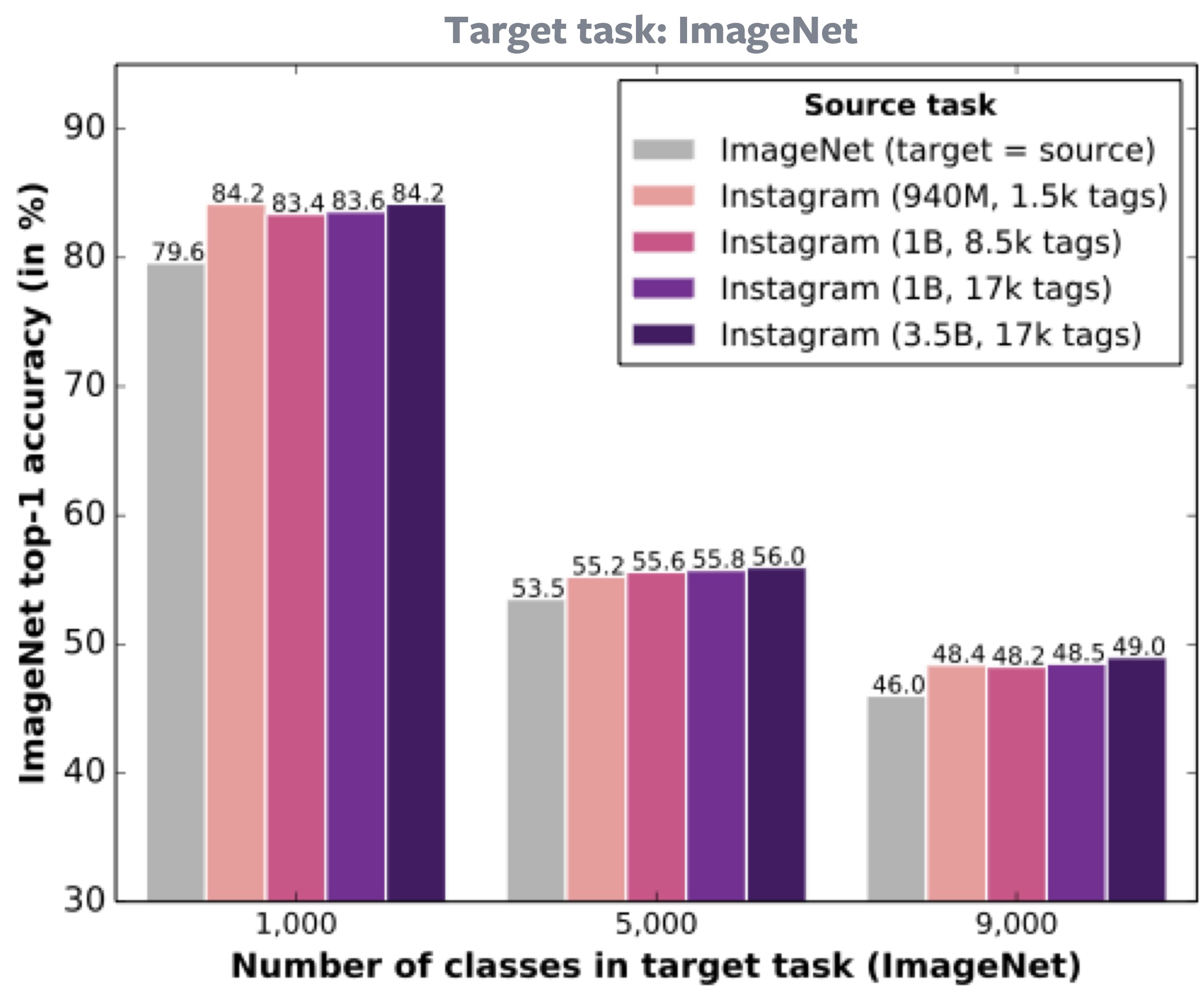
- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet

pretraining on 1-3.5B Instagram images, without selection



Fix Model; Vary Data

- Pretrain model on ImageNet or Instagram
- Finetune on ImageNet
- Similar results on larger versions of ImageNet



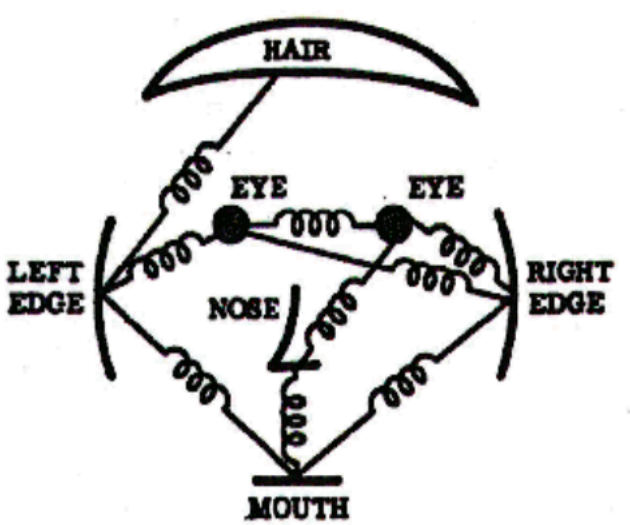
Two Extremes of Vision

Extrapolation problem
Generalization
Diagnostic features

Interpolation problem
Correspondence
Finding the differences

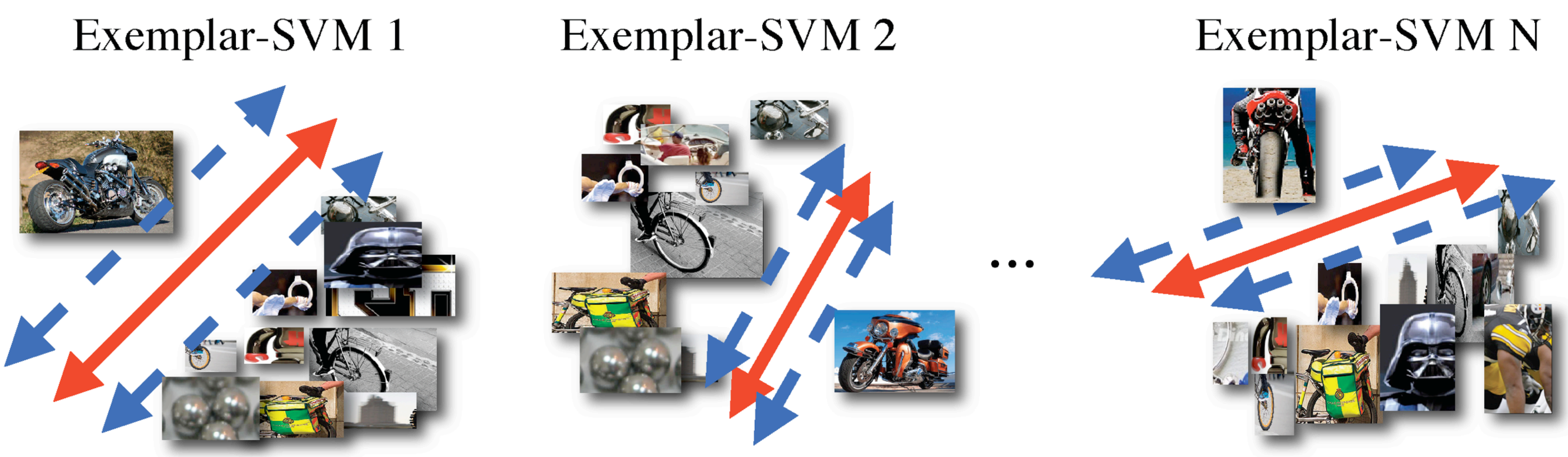
1 — 10 — 10^2 — 10^3 — 10^4 — 10^5 — 10^6 — ∞

Number of training samples



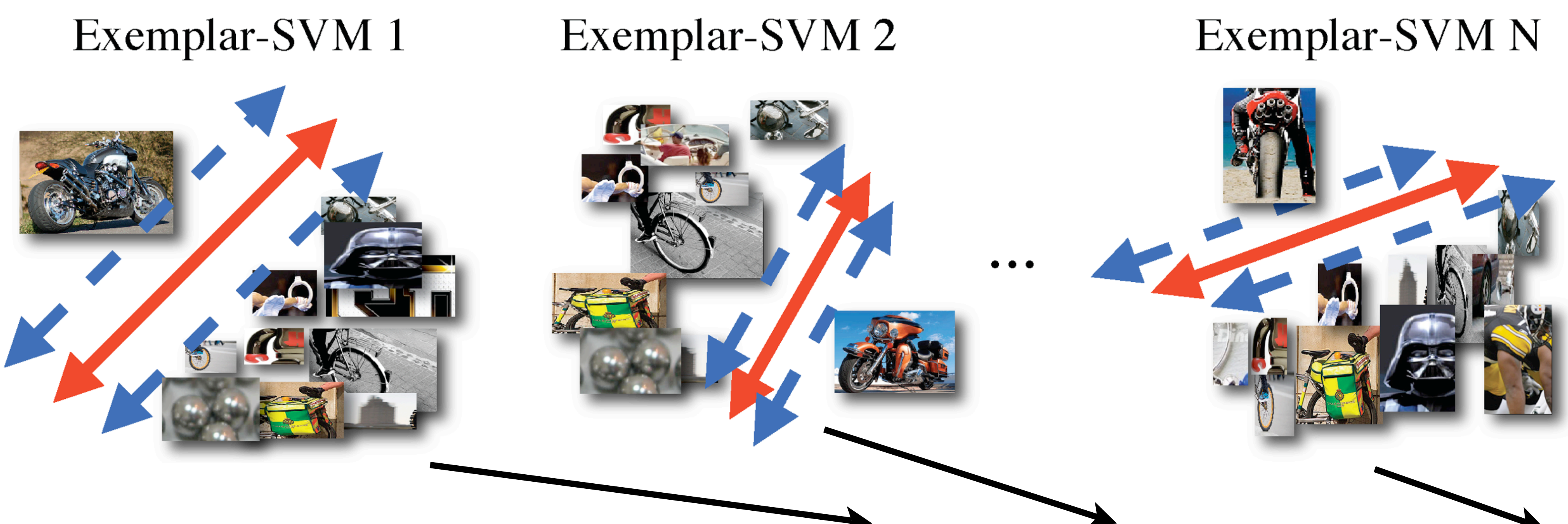
Slide credit: Aude Oliva

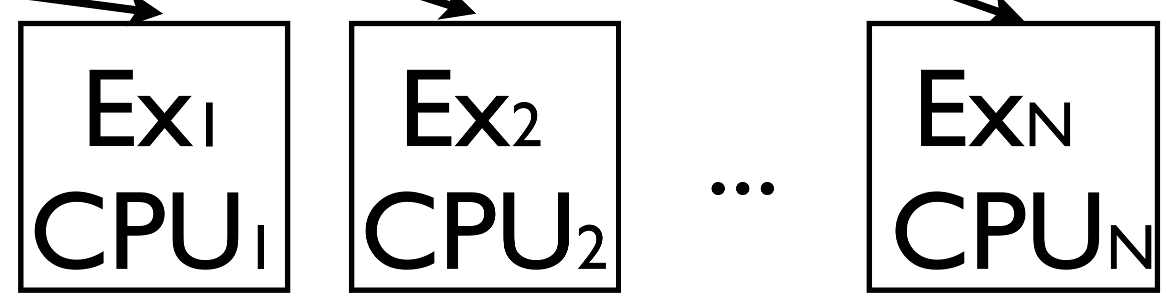
Exemplar-SVMs



- Learn a separate linear SVM for each instance (exemplar) in the dataset (PASCAL VOC)
- Each Exemplar-SVM is trained with a **single** positive instance
- Each Exemplar-SVM is more defined by "*what it is not*" vs. "*what it is similar to*"

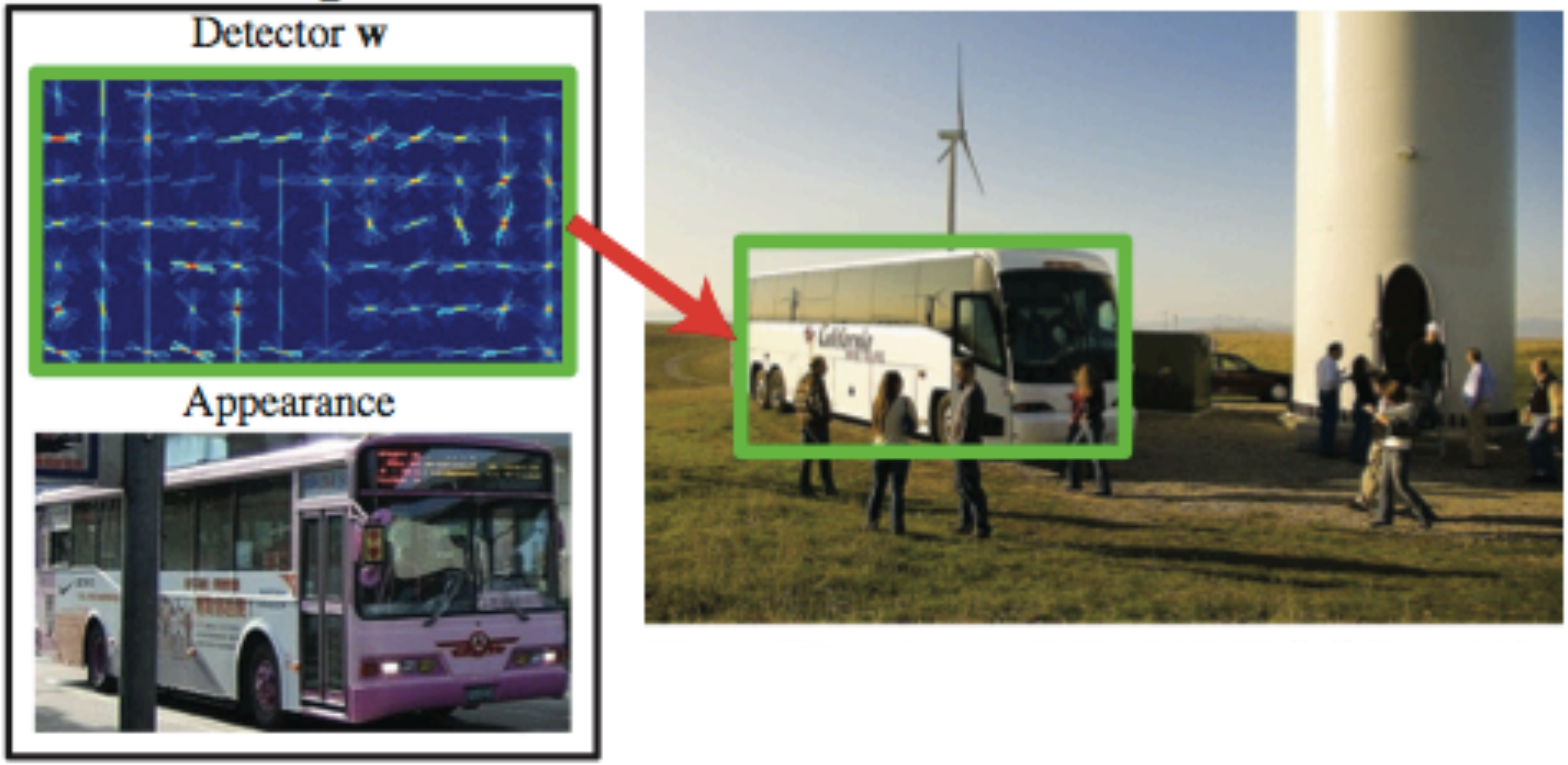
Large-scale training

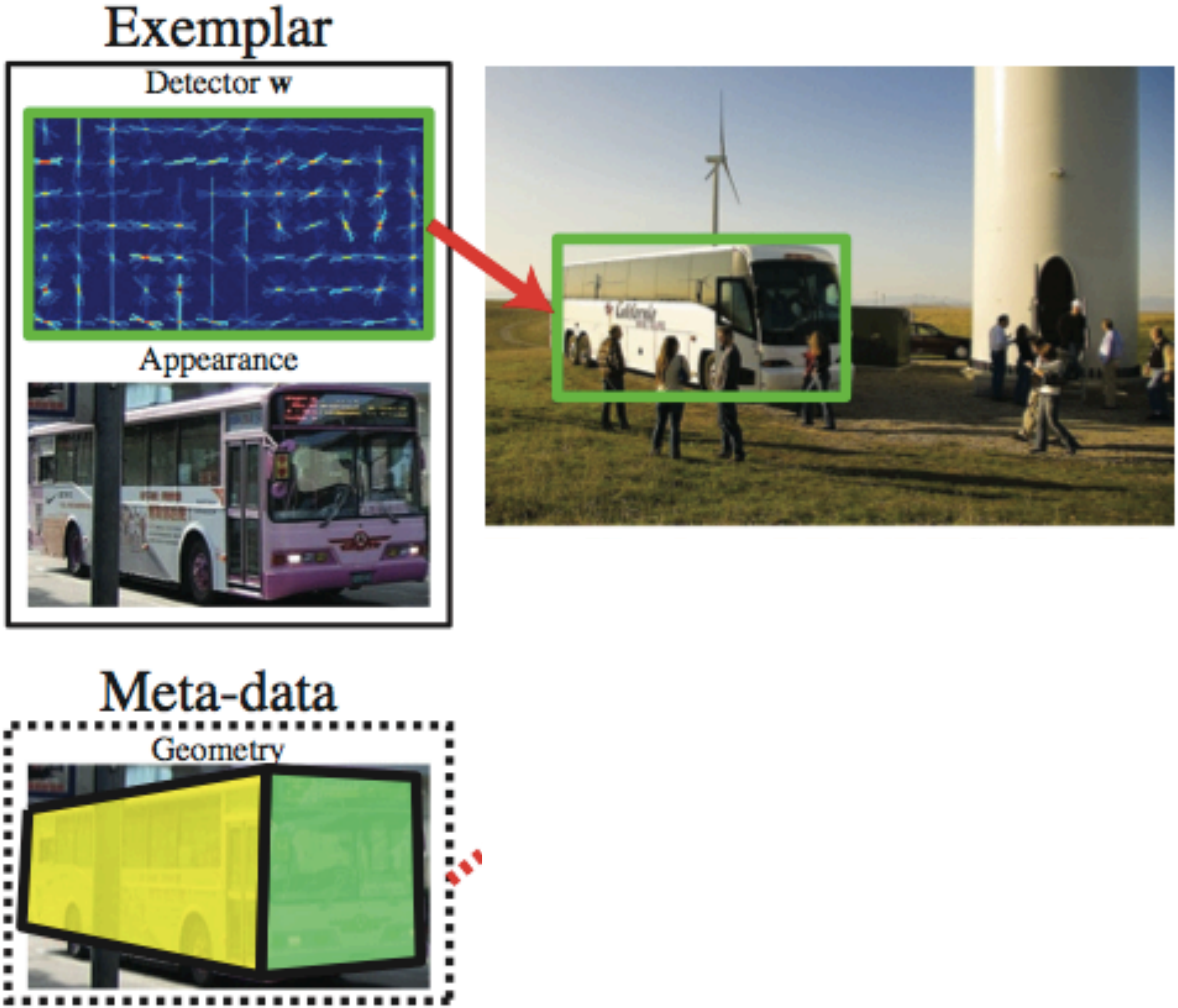


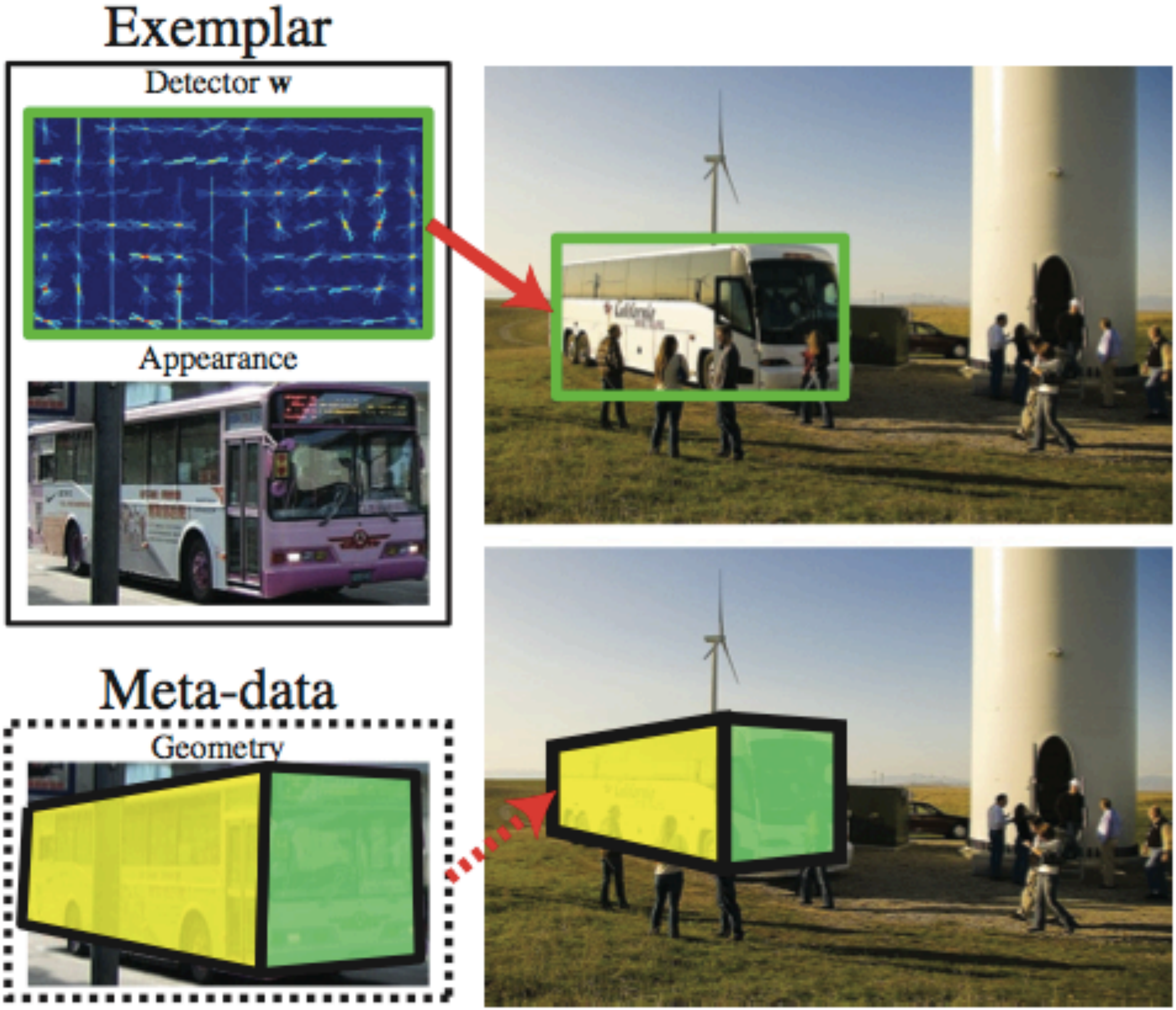


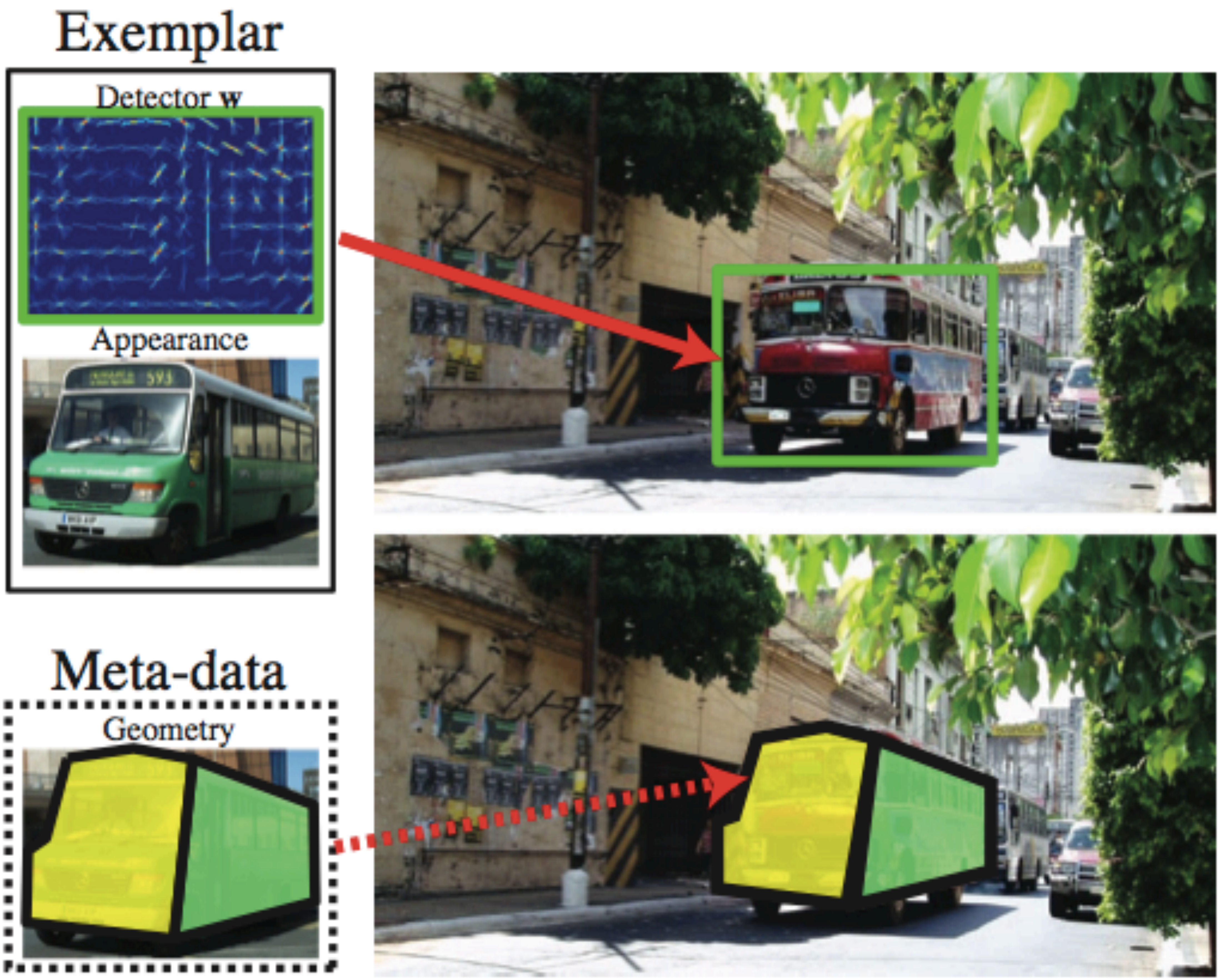
- Each exemplar performs its own hard negative mining
- Solve many convex learning problems
- Parallel training on cluster

Exemplar

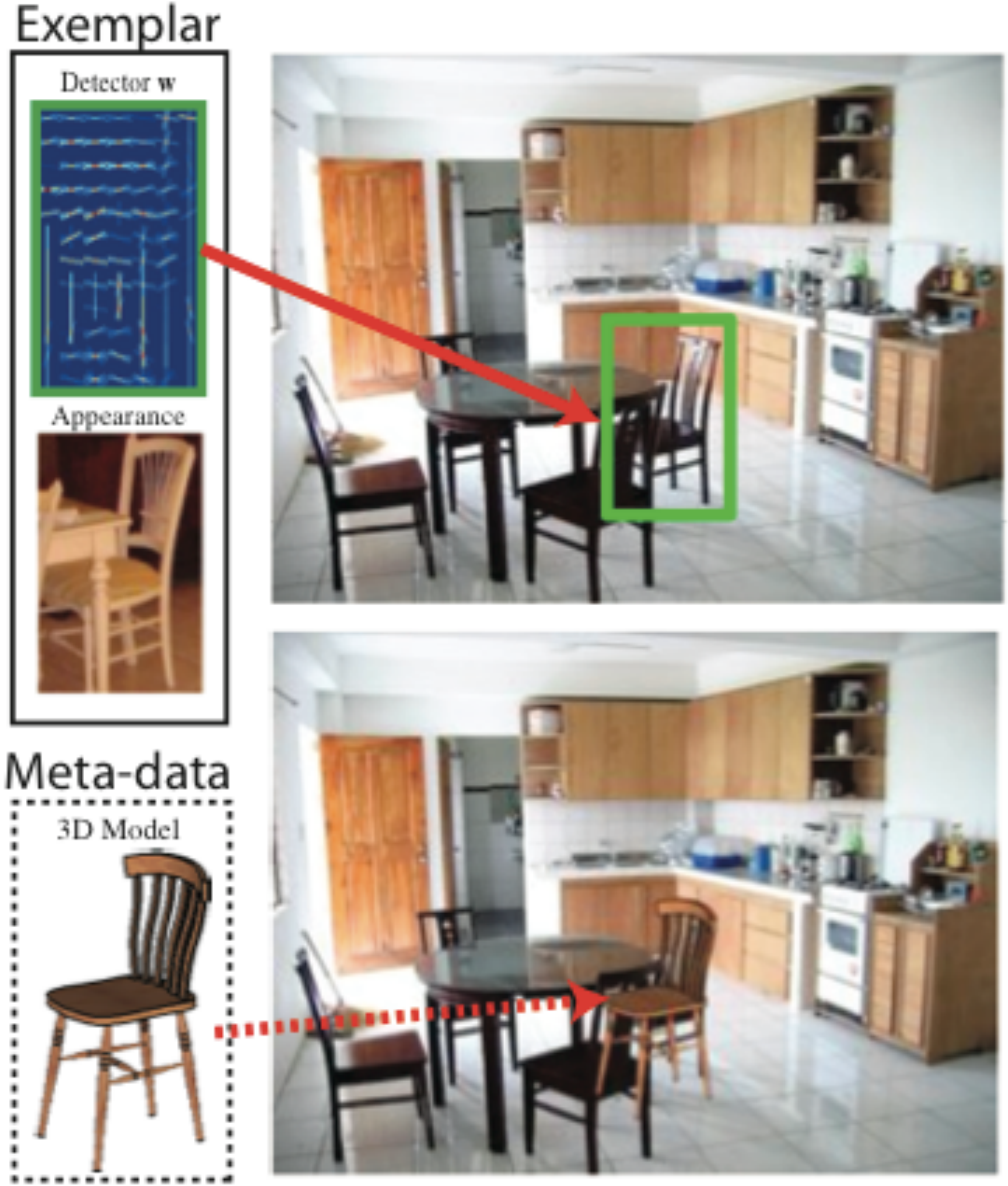


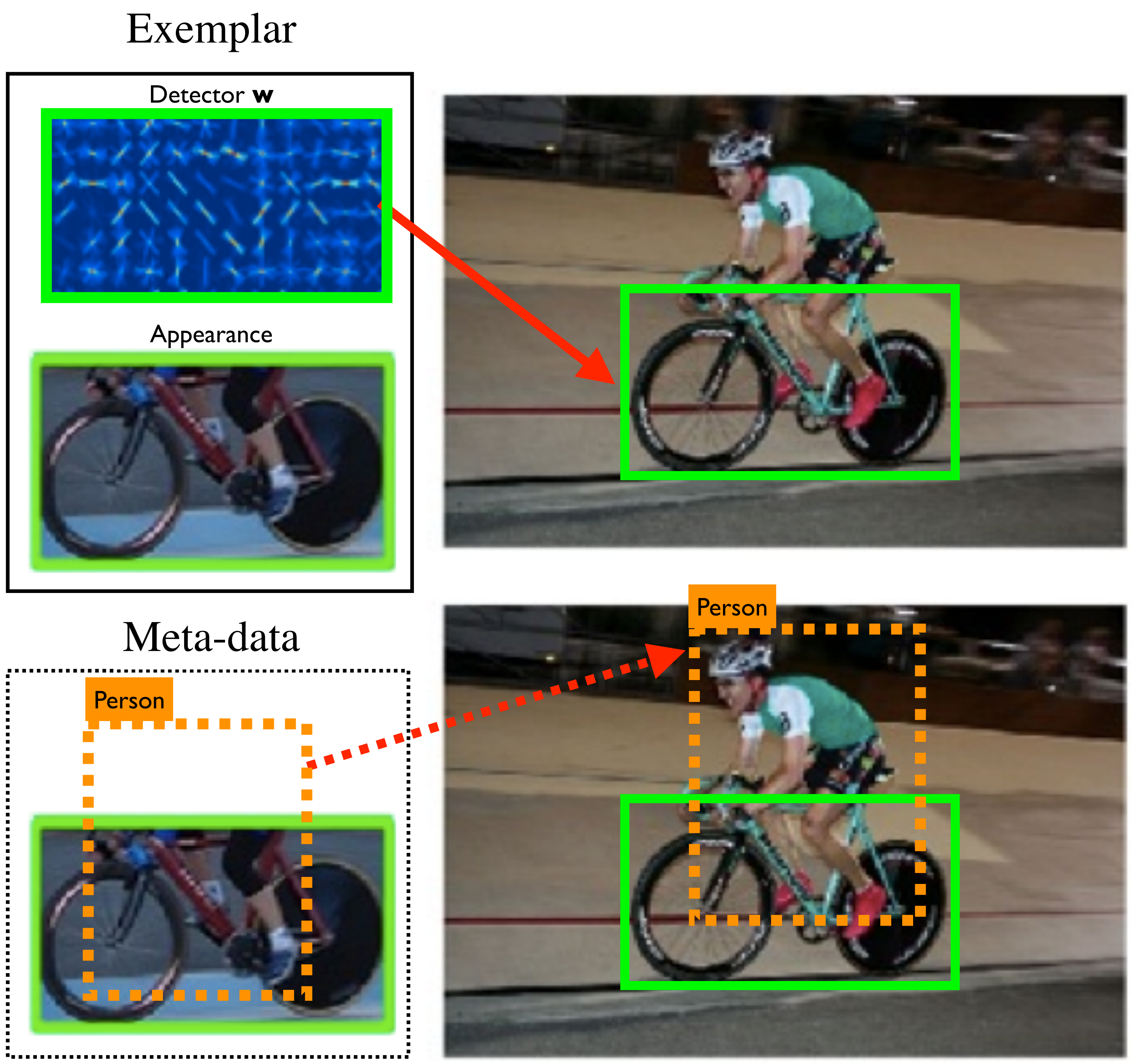

Exemplar
Detector w
Appearance
Meta-data
Geometry

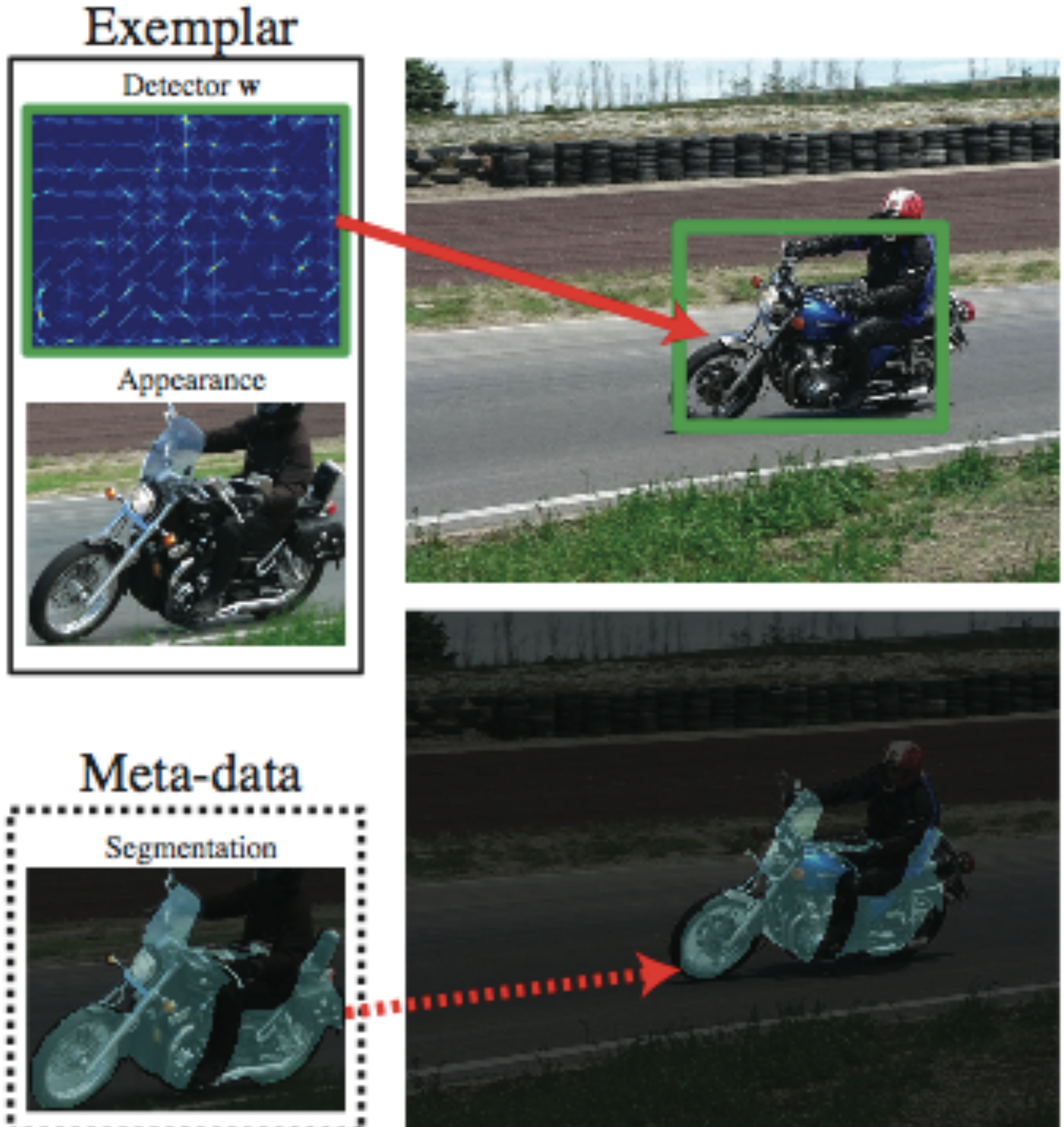

Exemplar
Detector w
Appearance
Meta-data
Geometry

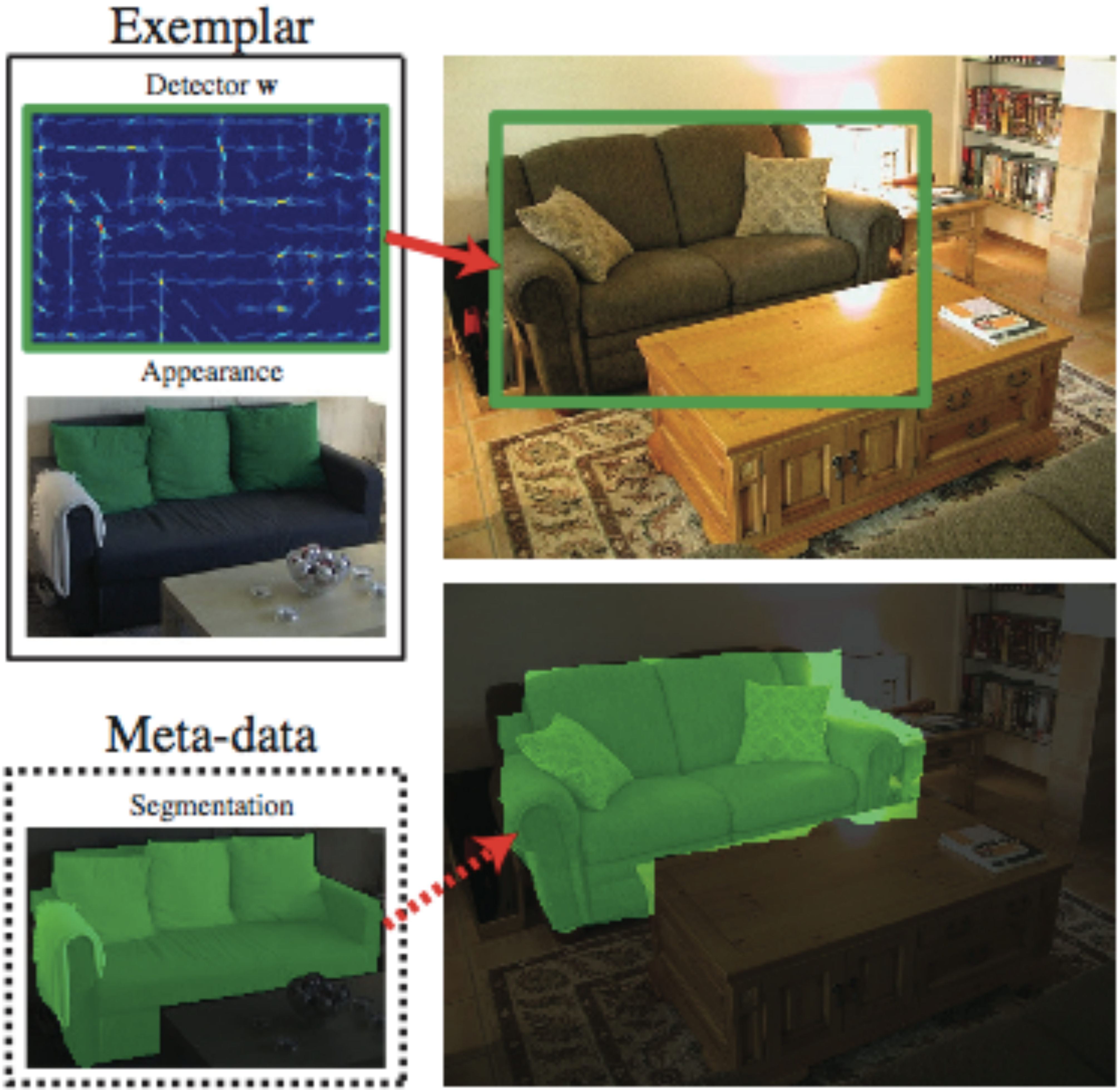

Exemplar
Detector w
Appearance
Meta-data
Geometry


Exemplar
Detector w
Appearance
Meta-data
3D Model


Exemplar
Detector w
Appearance
Meta-data
3D Model


Exemplar
Detector w
Appearance
Meta-data
Person
Person


Exemplar
Detector w
Appearance
Meta-data
Segmentation


Exemplar
Detector **w**
Appearance
Meta-data
Segmentation

What’s this?

Photo from Coffee Creek Watershed Preserve

What’s this?

Entry-level categories
(Jolicoeur, Gluck, Kosslyn 1984)

- Typical member of a basic-level category are categorized at the expected level
- Atypical members tend to be classified at a subordinate level.

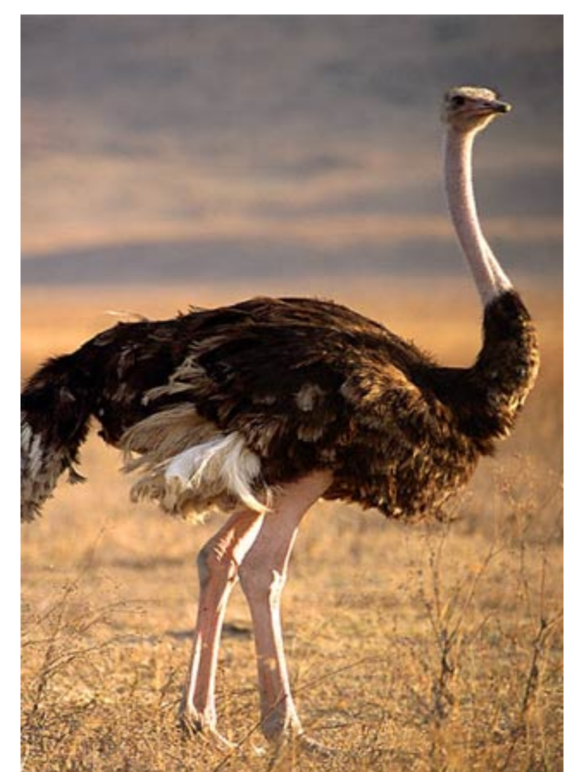
A bird


An ostrich

Classical Categorization

- Group objects by common properties
- What are birds?
  - animals, has wings, has feathers, can fly, chirps

Photo from Coffee Creek Watershed Preserve



Prototype Theory

Rosch and Lakoff

- According to the prototype view, an object will be classified as an instance of a category if it is sufficiently similar to the prototype.
- **Evidence for Prototype**:
- **Typicality ratings**: how good are robins as an example of birds
- **Production order of exemplars**: Name all the kinds of bird you can think of
- **Time to verify categorical statements**: True or false: a robin is a bird

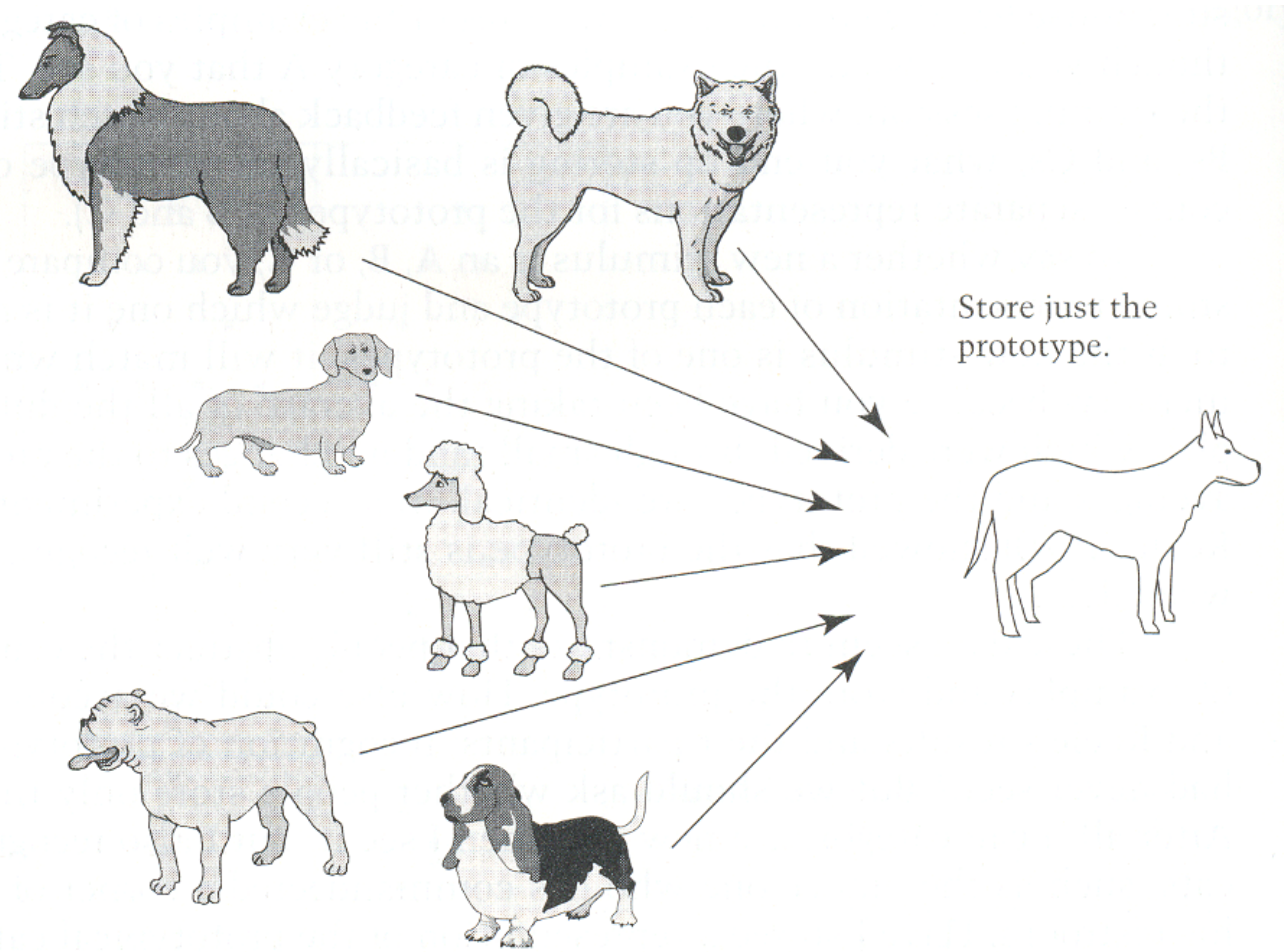


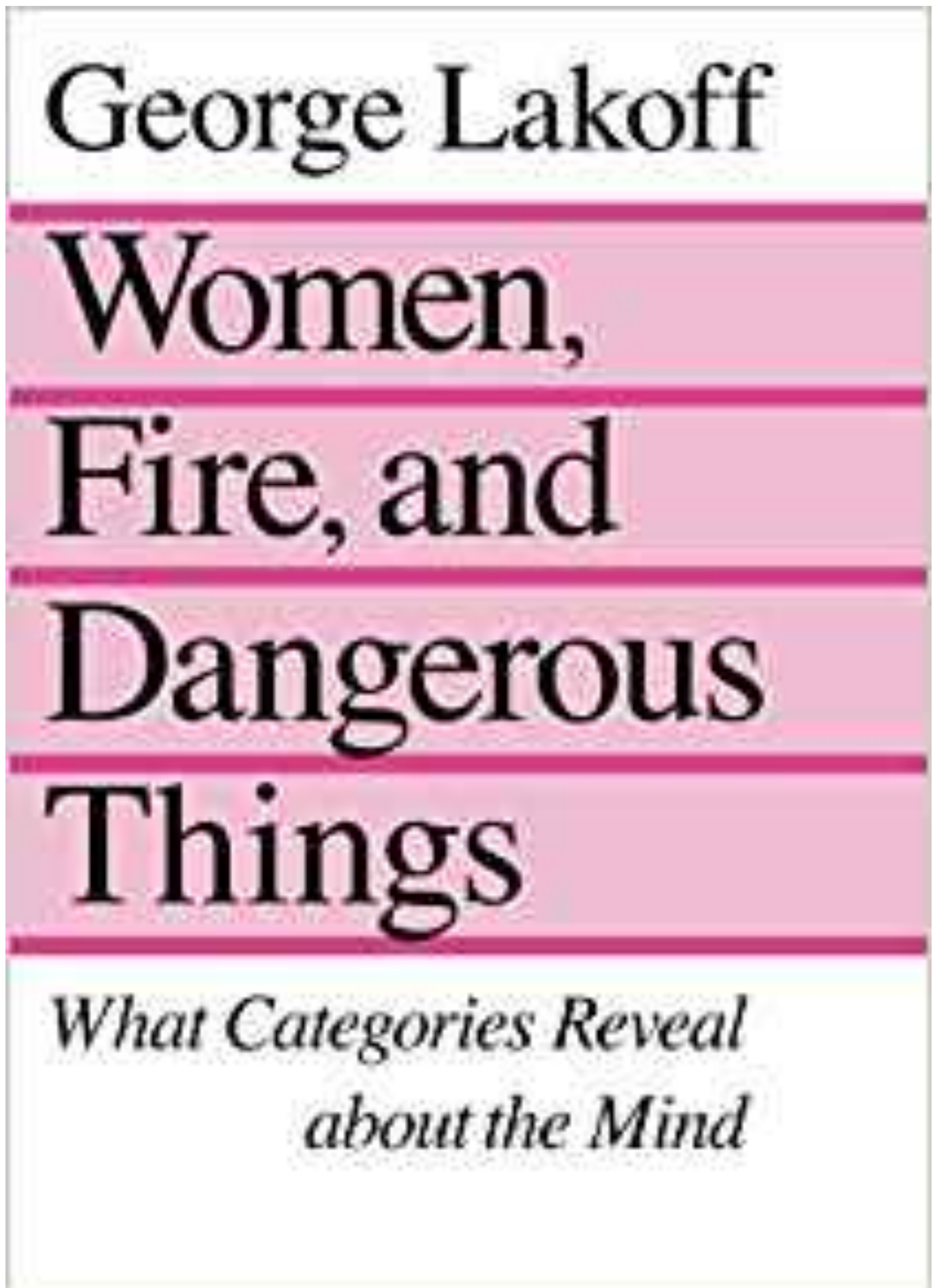
Figure 7.3. *Schematic of the prototype model. Although many exemplars are seen, only the prototype is stored. The prototype is updated continually to incorporate more experience with new exemplars.*

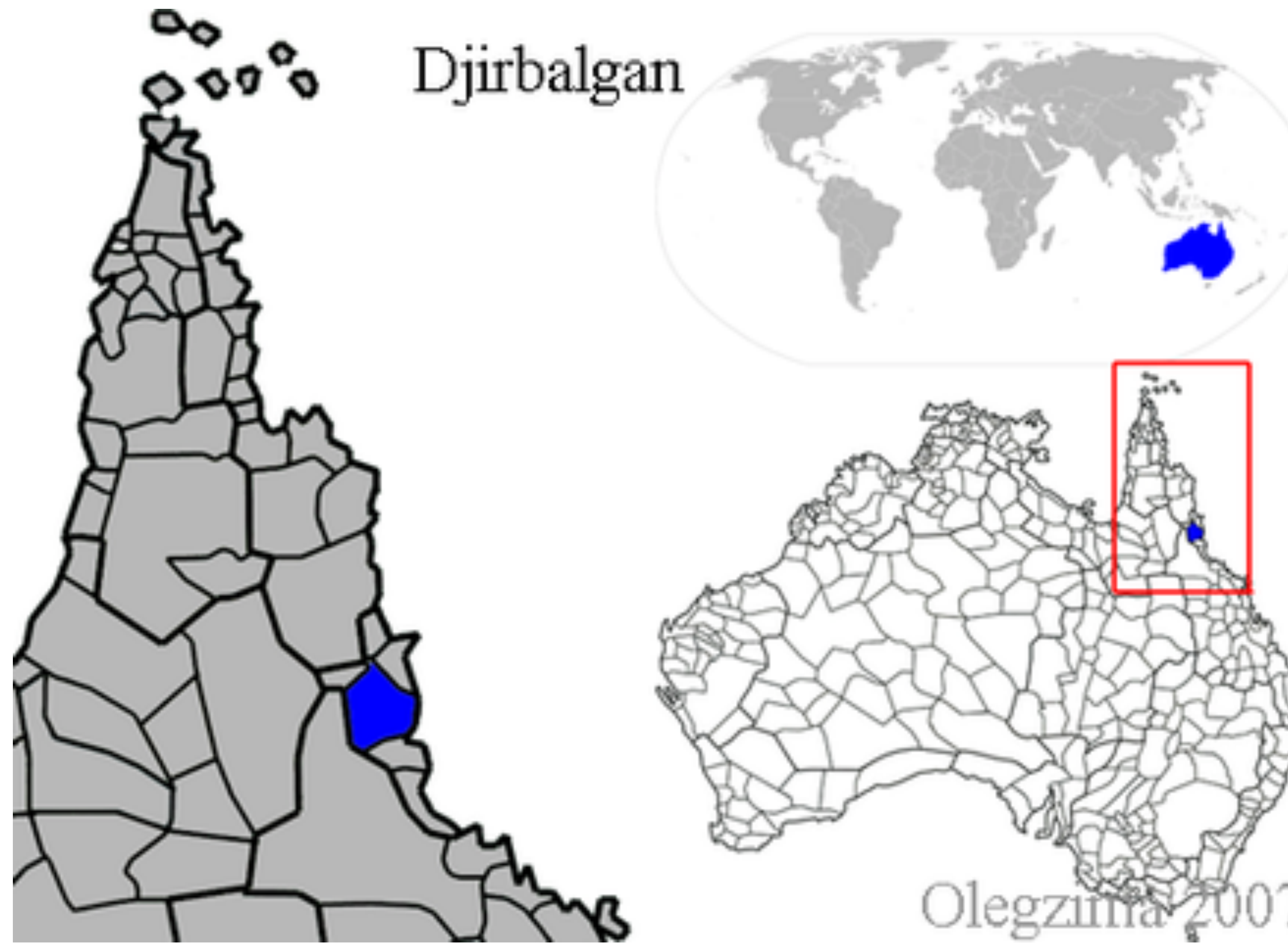
Canonical Perspective

The "best," most easily identified view of an object.
(Palmer, Rosch & Chase, 1981)

Slide by Palmer

Dyirbal Indigenous People





The perception of function

- Direct perception (affordances): Gibson

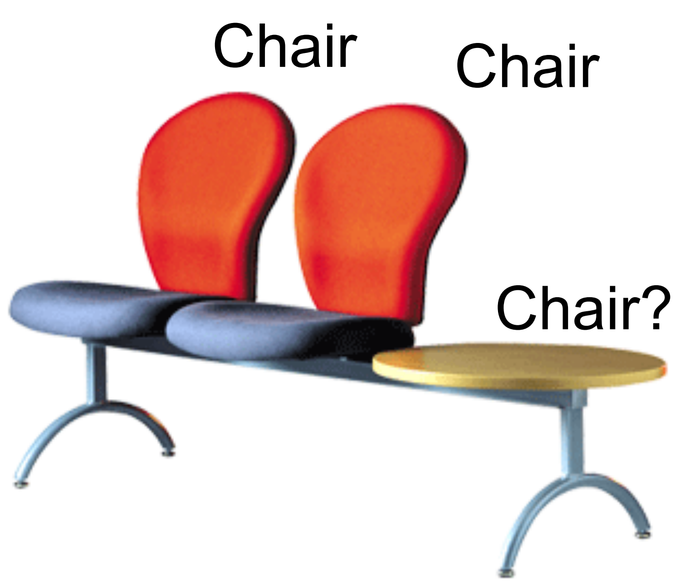
Flat surface
Horizontal
Knee-high
…

→ Sittable upon

- Mediated perception (Categorization)

Flat surface
Horizontal
Knee-high
…

→ Chair → Sittable upon



Direct perception

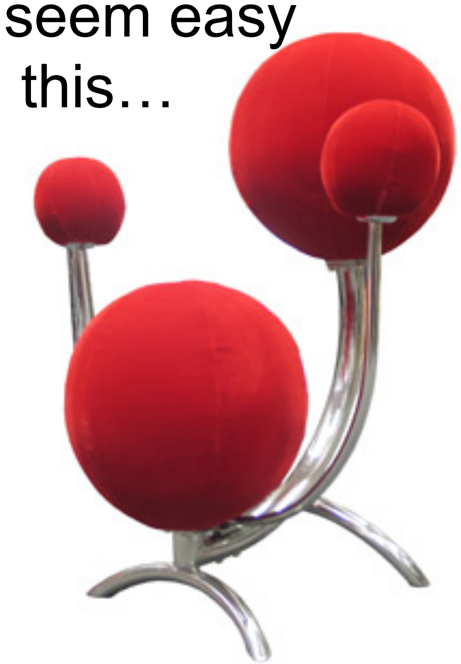
Some aspects of an object function can be perceived directly

- Functional form: Some forms clearly indicate to a function (“sittable-upon”, container, cutting device, …)

Sittable-upon Sittable-upon

Sittable-upon

It does not seem easy to sit-upon this…



Direct perception

Some aspects of an object function can be perceived directly

- Observer relativity: Function is observer dependent

Limitations of Direct Perception

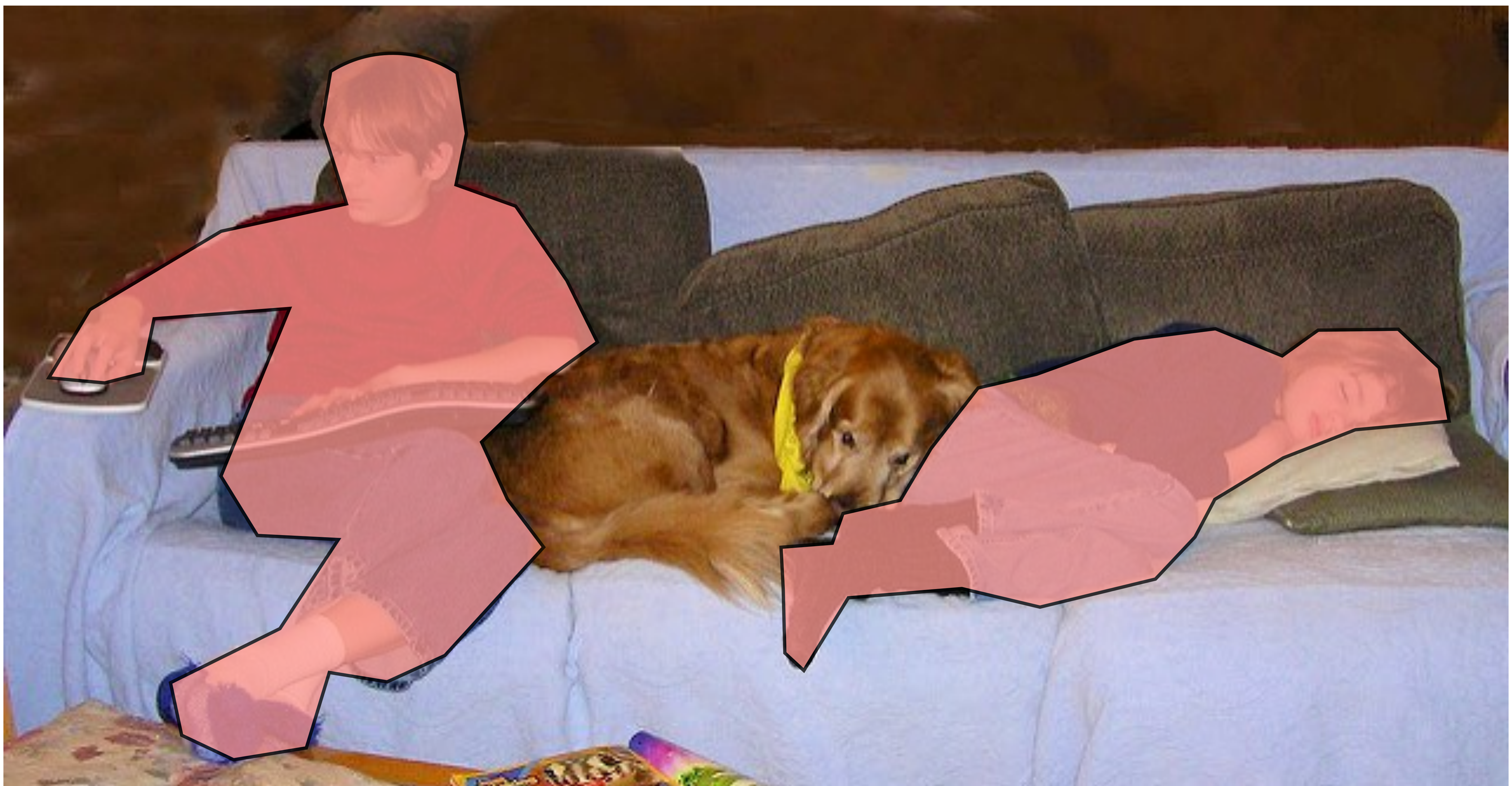
Objects of similar structure might have very different functions

Figure 9.1.2 Objects with similar structure but different functions. Mailboxes afford letter mailing, whereas trash cans do not, even though they have many similar physical features, such as size, location, and presence of an opening large enough to insert letters and medium-sized packages.

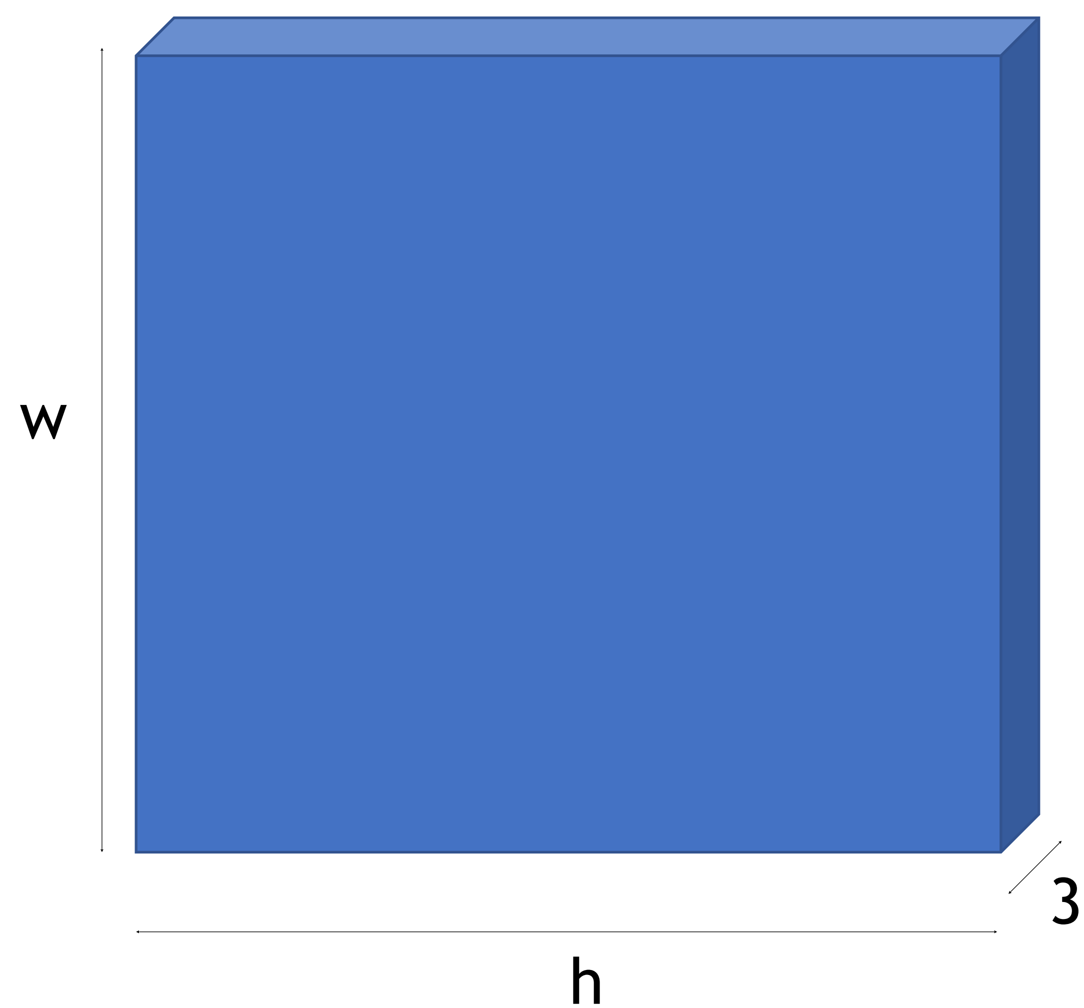
Not all functions seem to be available from direct visual information only.

The functions are the same at some level of description: we can put things inside in both and somebody will come later to empty them. However, we are not expected to put inside the same kinds of things…

Segmentation:
Where *really* are the people?

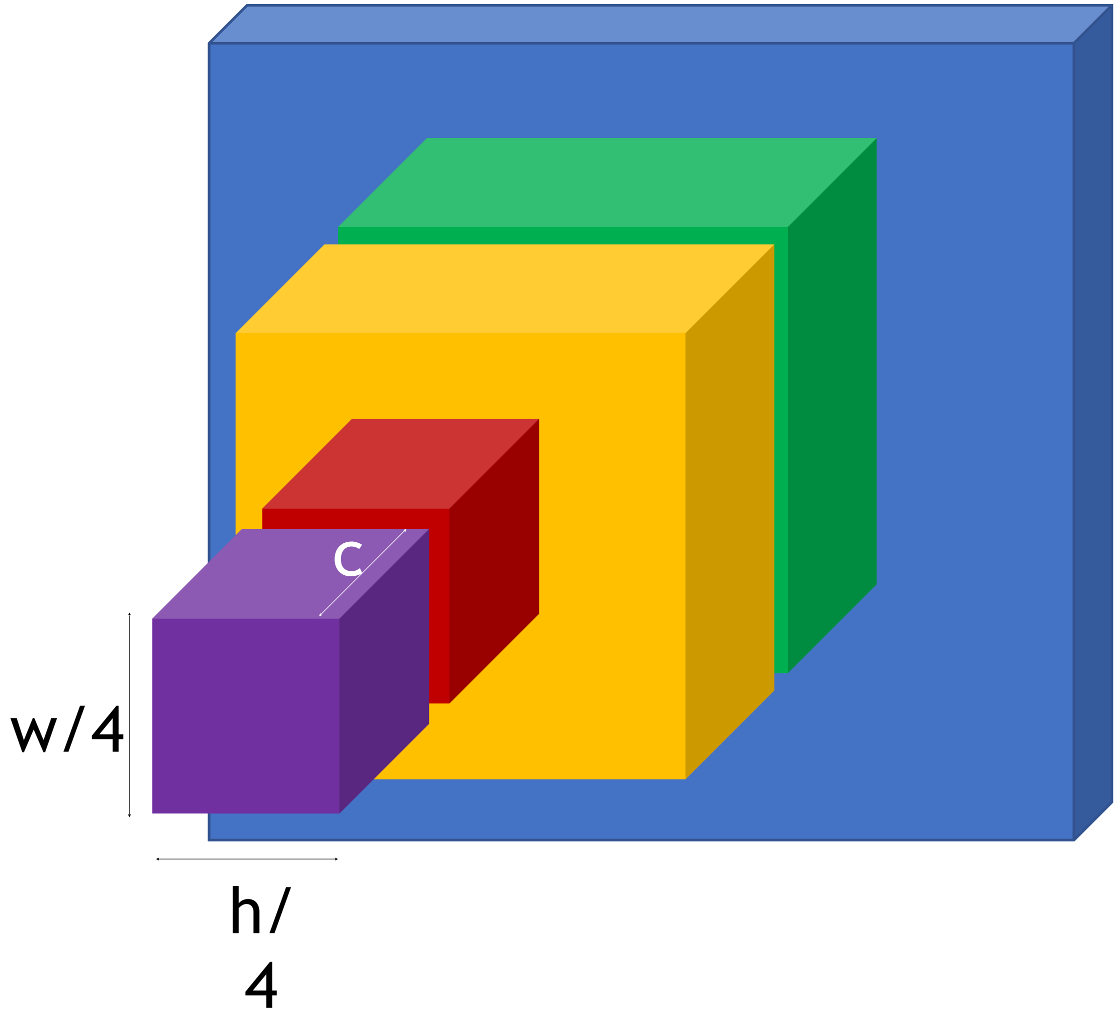


Semantic segmentation using convolutional networks



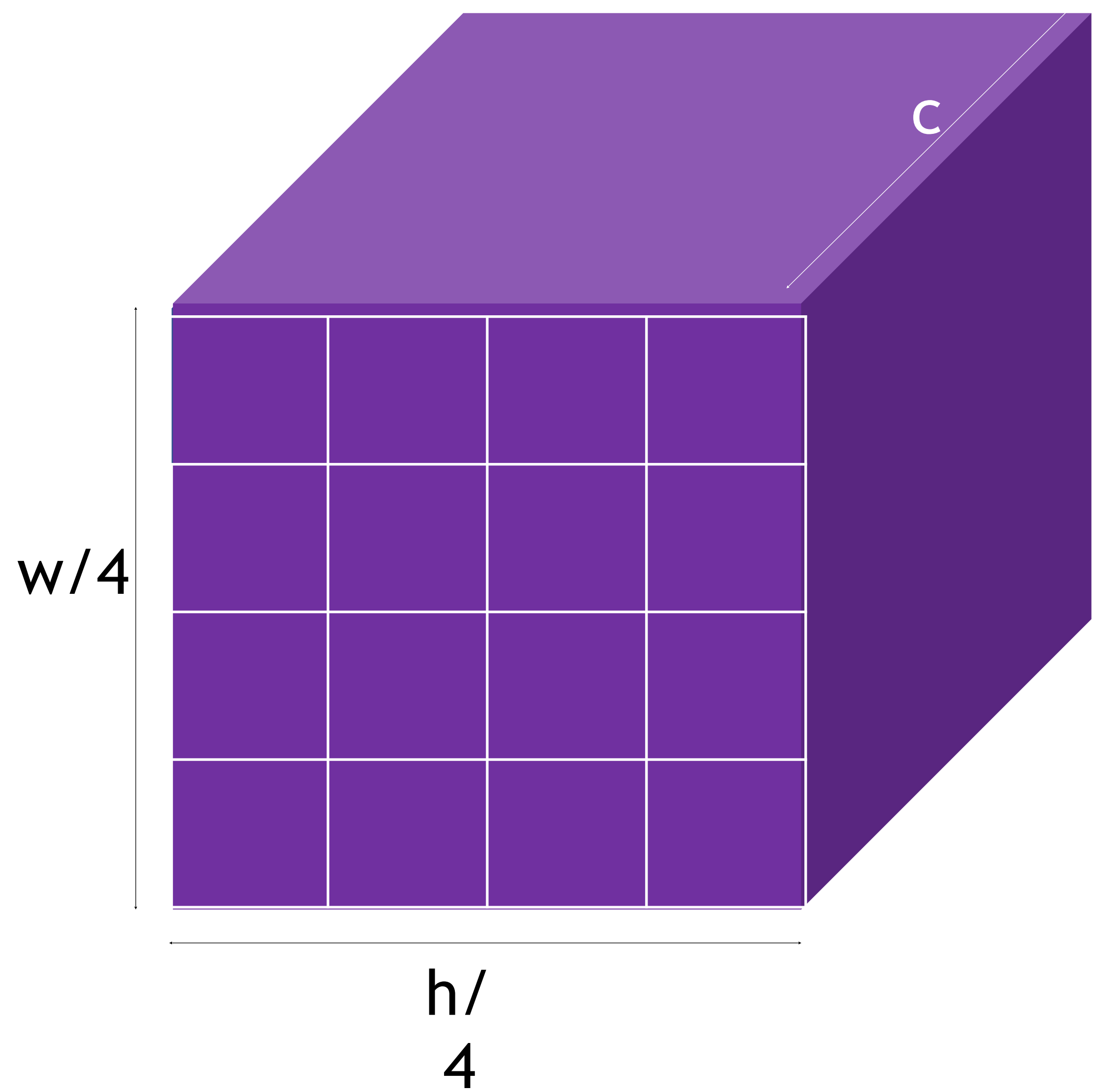
Slide credit: Bharath Hariharan

Semantic segmentation using convolutional networks



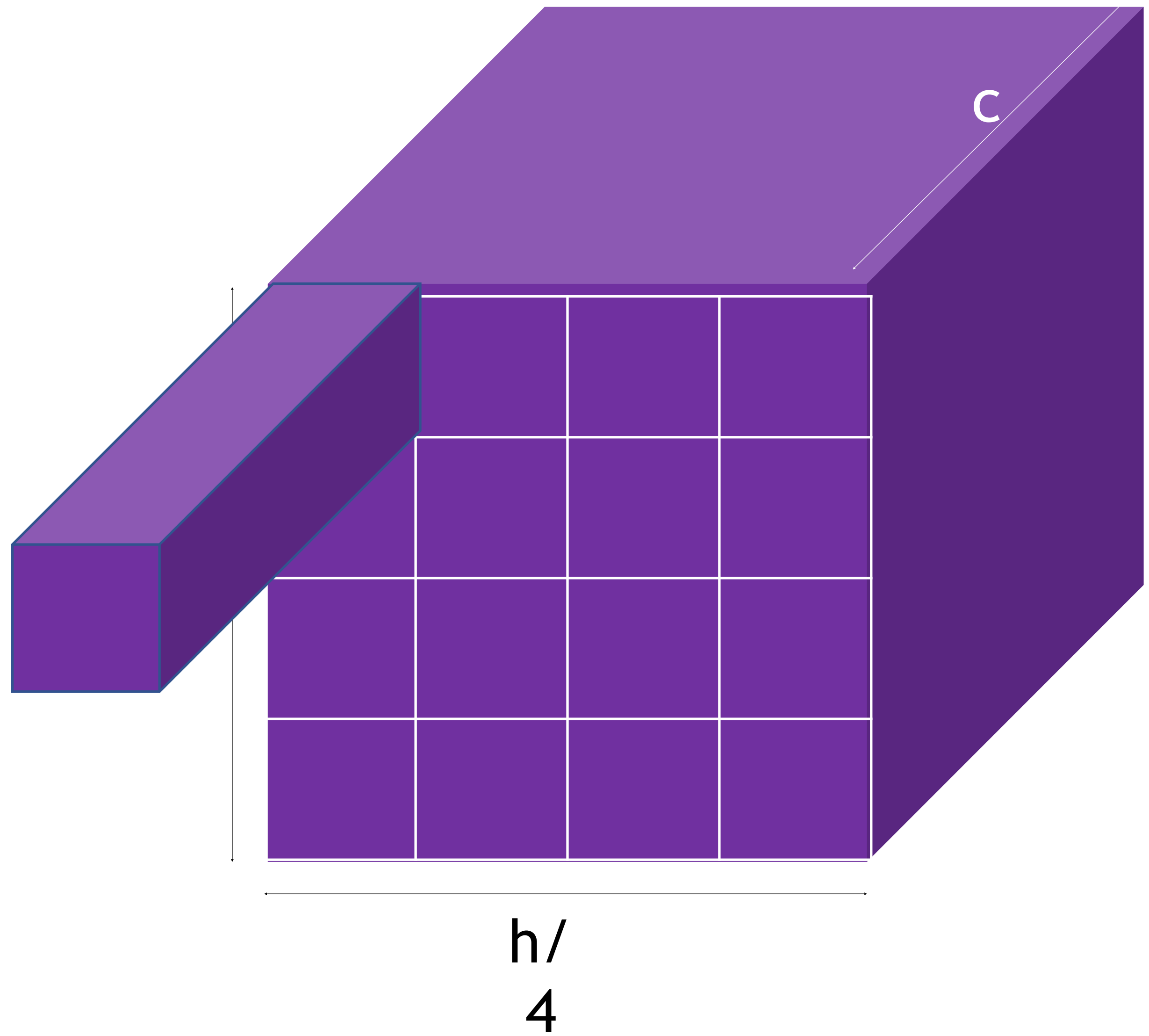
Slide credit: Bharath Hariharan

Semantic segmentation using convolutional networks



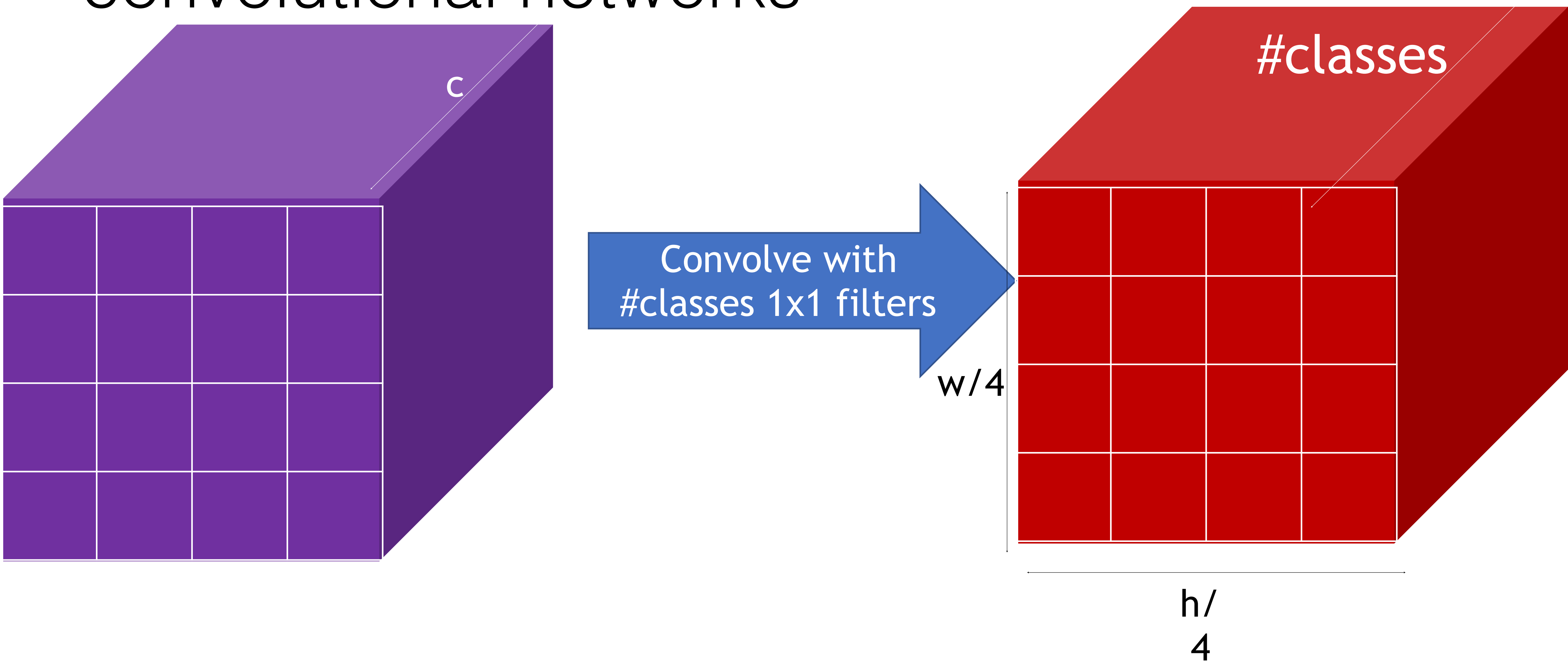
Slide credit: Bharath Hariharan

Semantic segmentation using convolutional networks



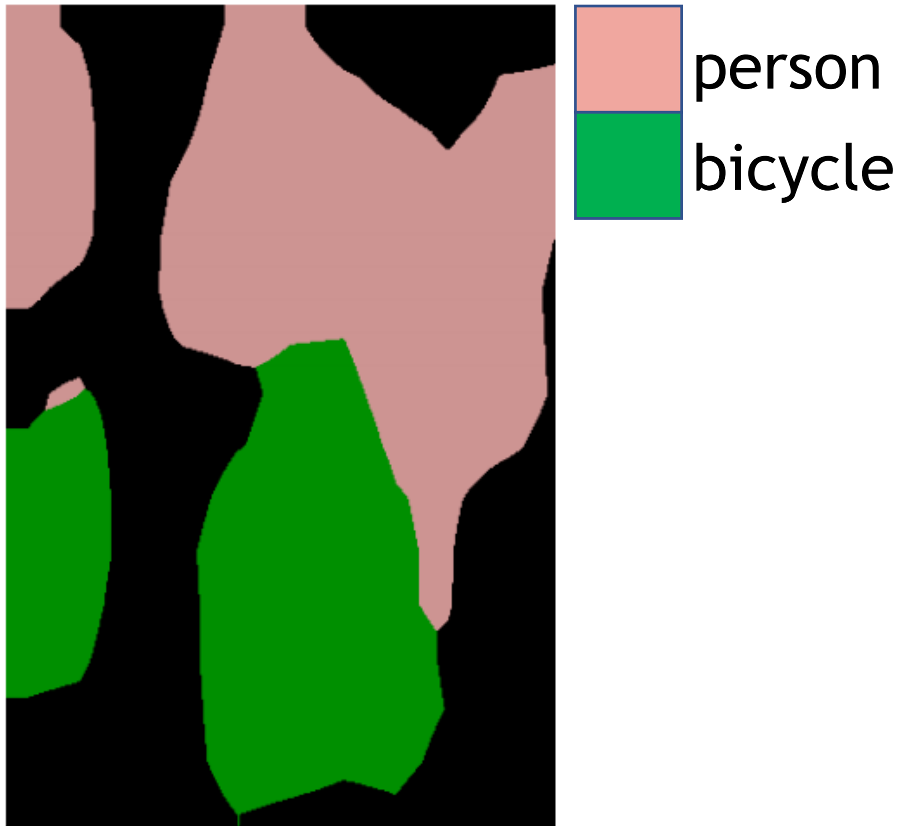
Slide credit: Bharath Hariharan

Semantic segmentation using convolutional networks



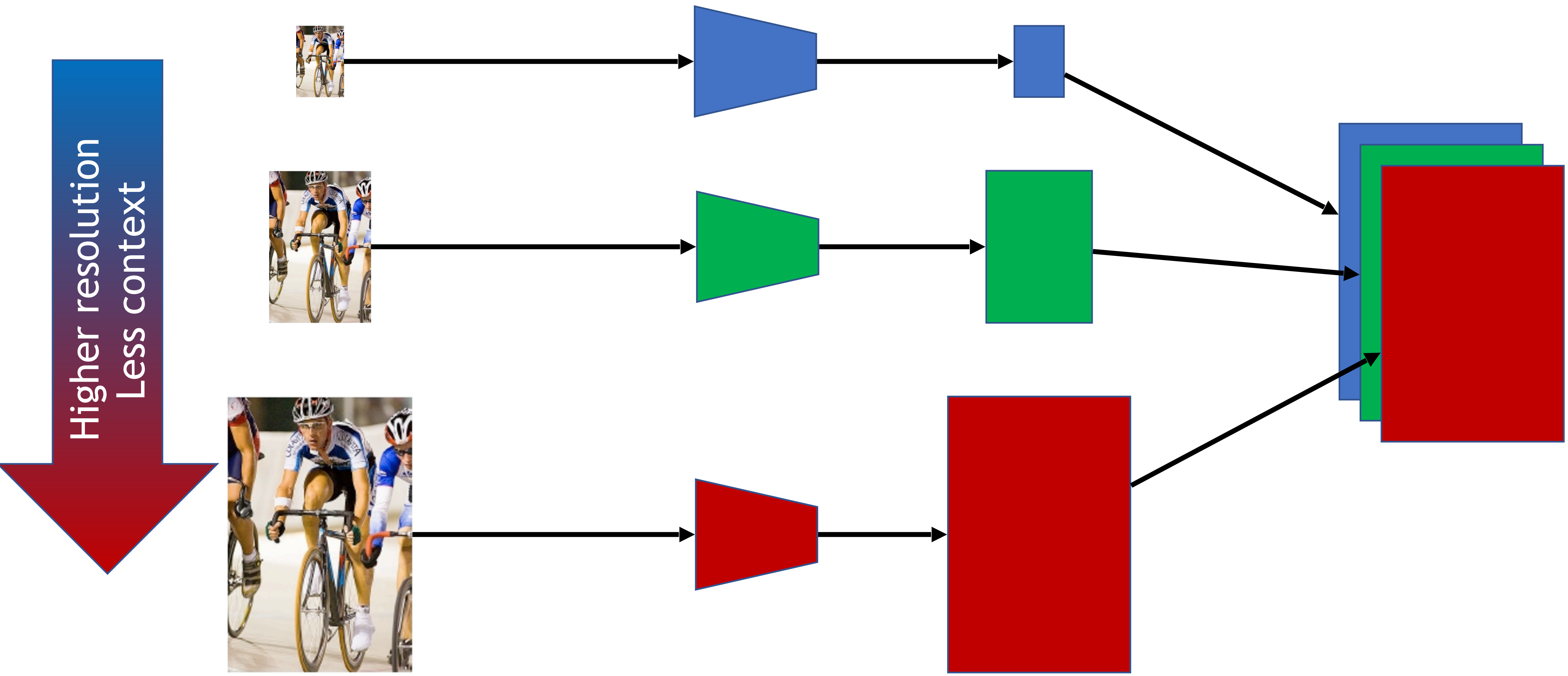
Slide credit: Bharath Hariharan

Semantic segmentation using convolutional networks



Slide credit: Bharath Hariharan

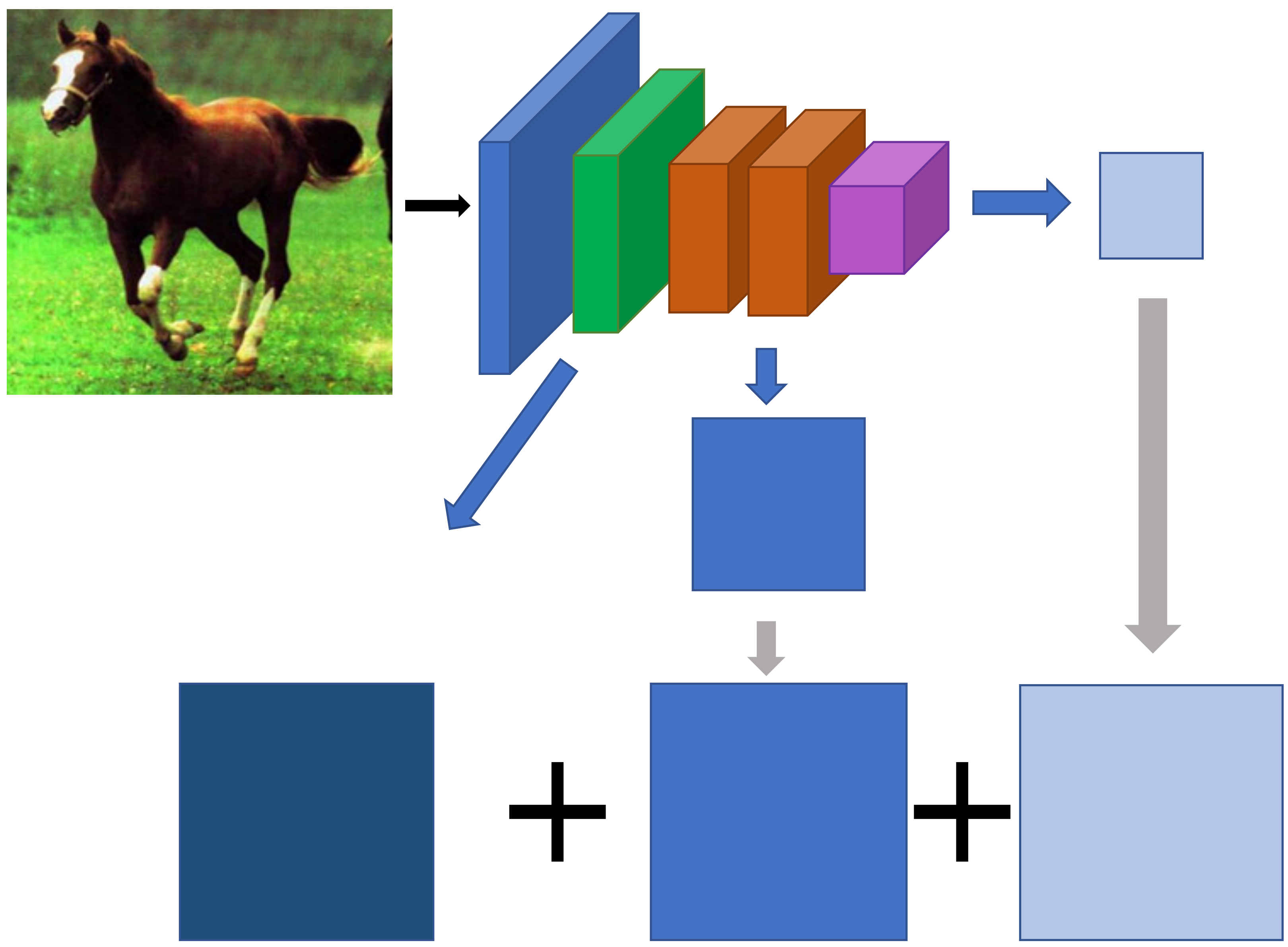
Solution 1: Image pyramids



Learning Hierarchical Features for Scene Labeling. Clement Farabet, Camille Couprie, Laurent Najman, Yann LeCun. In *TPAMI*, 2013.

Slide credit: Bharath Hariharan

Solution 2: Skip connections



Slide credit: Bharath Hariharan

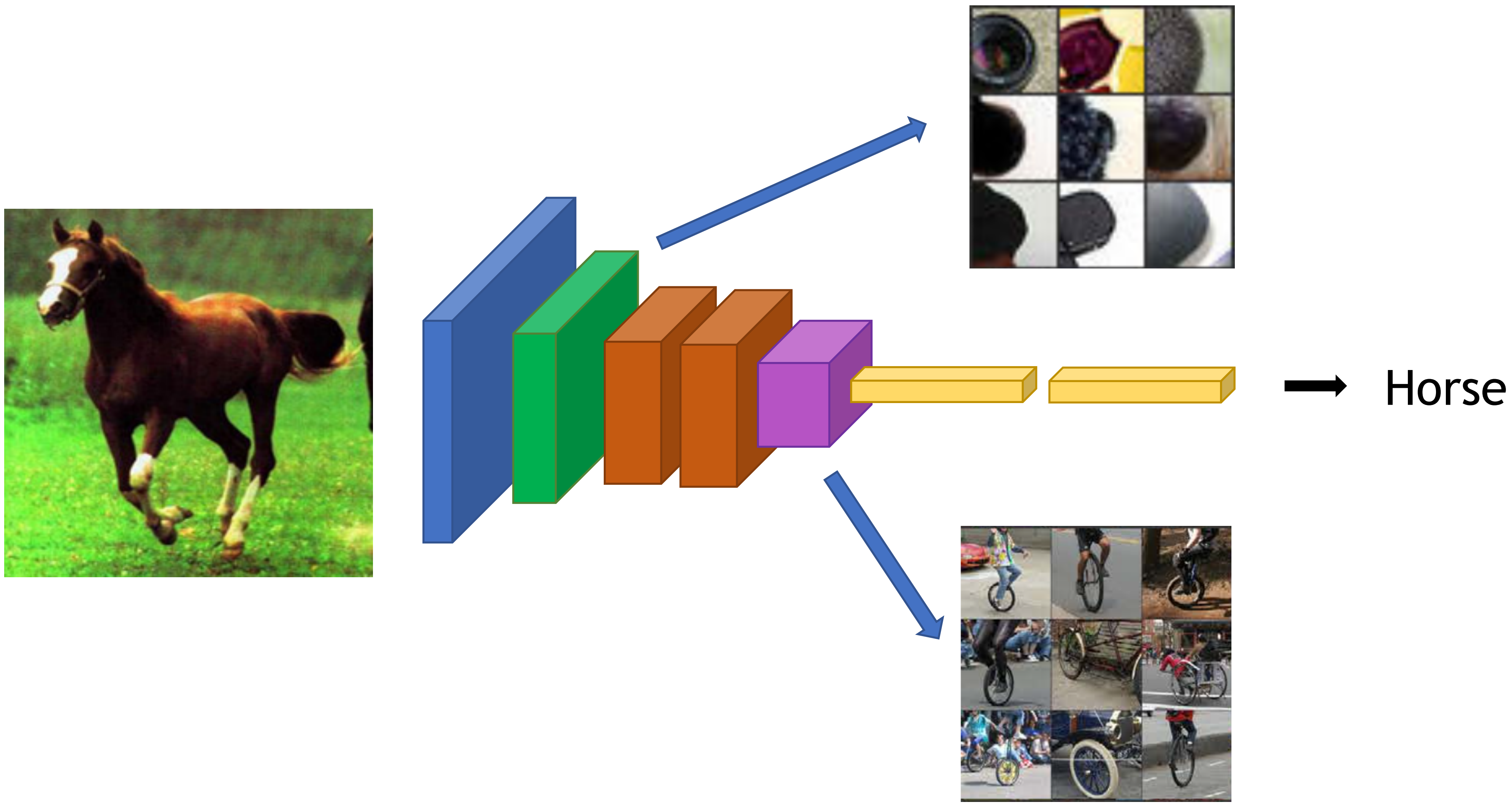
Skip connections

Fully convolutional networks for semantic segmentation. Evan Shelhamer, Jon Long, Trevor Darrell. In *CVPR* 2015

Slide credit: Bharath Hariharan

Skip connections

- Problem: early layers not semantic



Visualizations from : M. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In *ECCV* 2014.

Slide credit: Bharath Hariharan

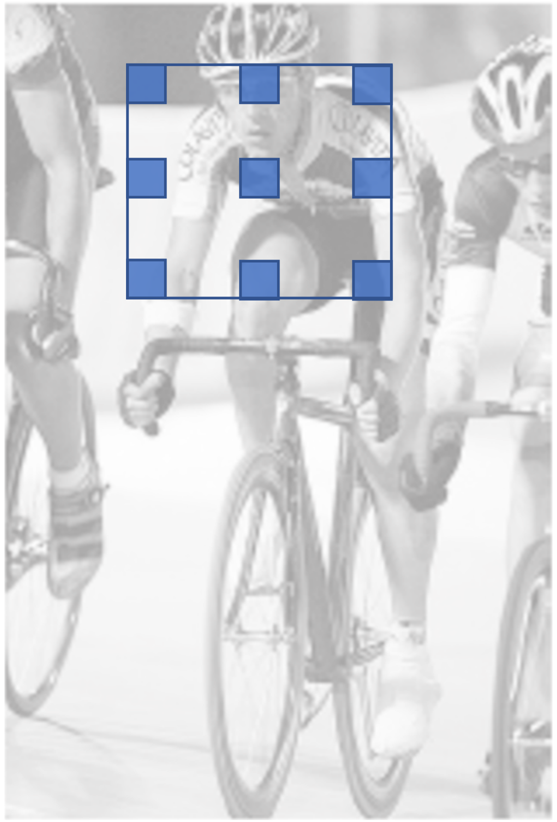
Solution 3: Dilation

- Need subsampling to allow convolutional layers to capture large regions with small filters
  - Can we do this without subsampling?

Slide credit: Bharath Hariharan

Solution 3: Dilation

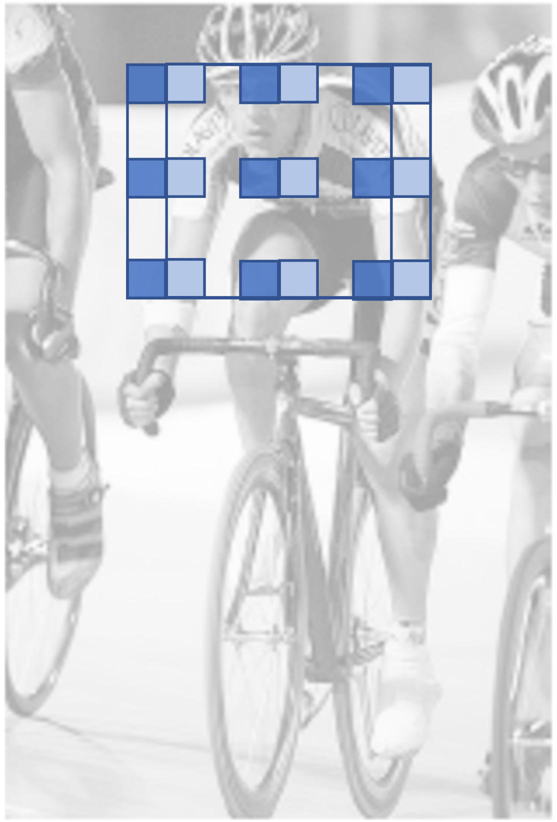
- Need subsampling to allow convolutional layers to capture large regions with small filters
  - Can we do this without subsampling?



Slide credit: Bharath Hariharan

Solution 3: Dilation

- Need subsampling to allow convolutional layers to capture large regions with small filters
  - Can we do this without subsampling?



Slide credit: Bharath Hariharan

Solution 4: Conditional Random Fields

- Idea: take convolutional network prediction and sharpen using classic techniques
- *Conditional Random Field*

$$\mathbf{y}^* = \arg\min_{\mathbf{y}} \sum_{(i,j)} E_{data}(y(i,j)) + \sum_{(i,j),(k,l)\in\mathcal{N}} E_{smooth}(y(i,j), y(k,l))$$

$$E_{smooth}(y(i,j), y(k,l)) = \mathbb{I}(y(i,j) \neq y(k,l))w(i,j,k,l)$$

Slide credit: Bharath Hariharan

Fully Connected CRFs

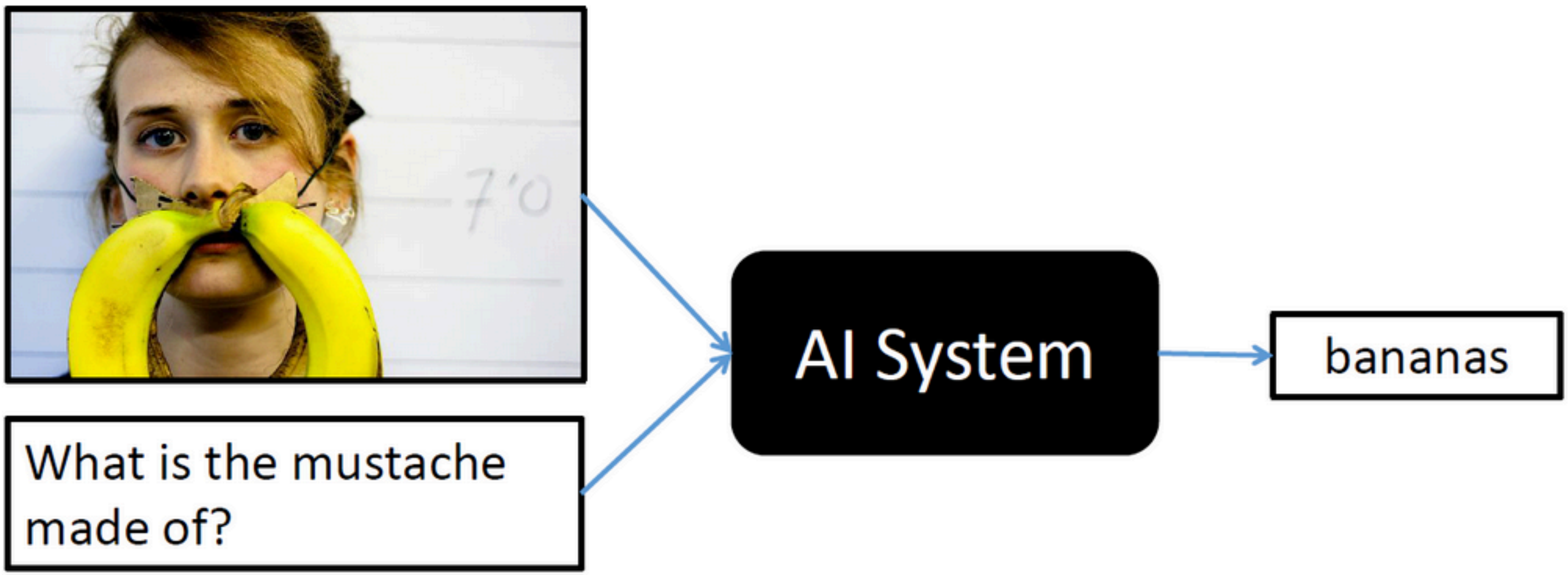
- Typically, only adjacent pixels connected
  - Fewer connections => Easier to optimize
- Dense connectivity: every pixel connected to everything else
- Intractable to optimize except if pairwise potential takes specific form

$$E_{smooth}(y(i,j), y(k,l)) = \mathbb{I}(y(i,j) \neq y(k,l))w(i,j,k,l)$$

$$w(i,j,k,l) = \sum_m w_m e^{-\|\mathbf{f}_m(i,j) - \mathbf{f}_m(k,l)\|^2}$$

Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. Philipp Krahenbuhl, Vladlen Koltun. In *NIPS*, 2011.

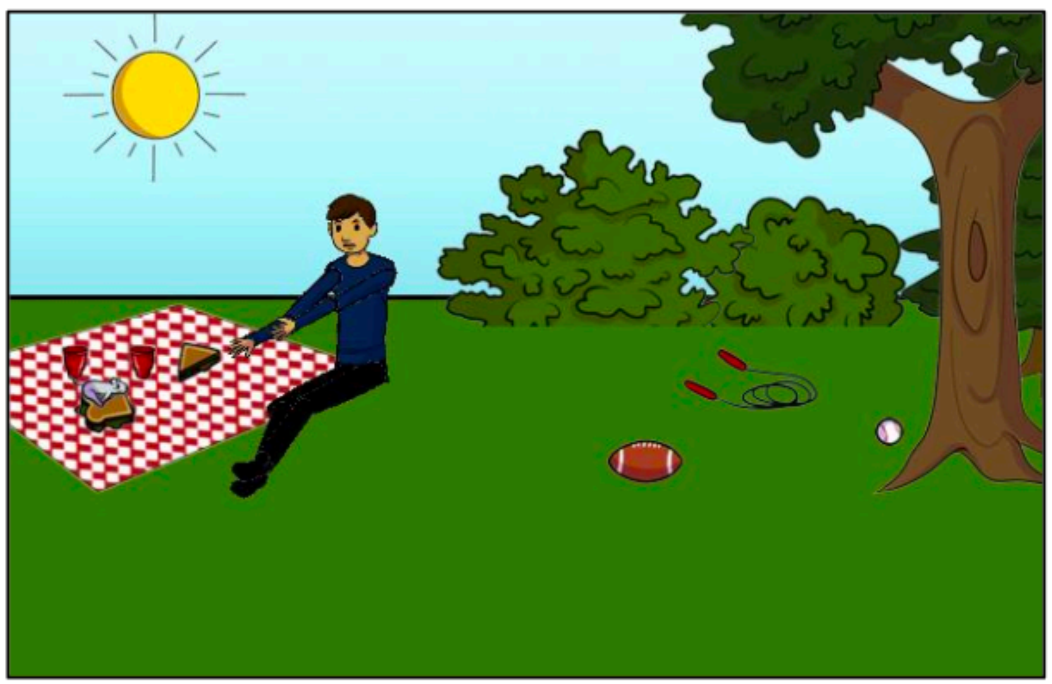
Slide credit: Bharath Hariharan



http://www.visualqa.org/challenge.html

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



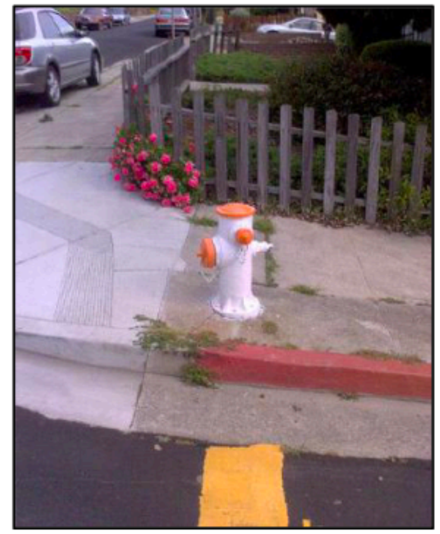
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

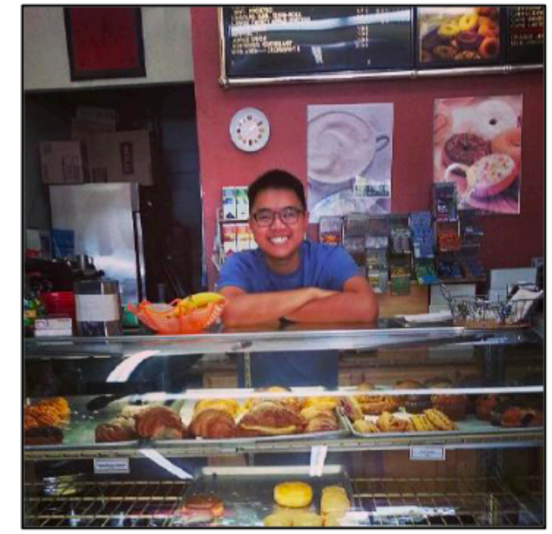
Fig. 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

Questions and answers collected with Amazon Mechanical Turk

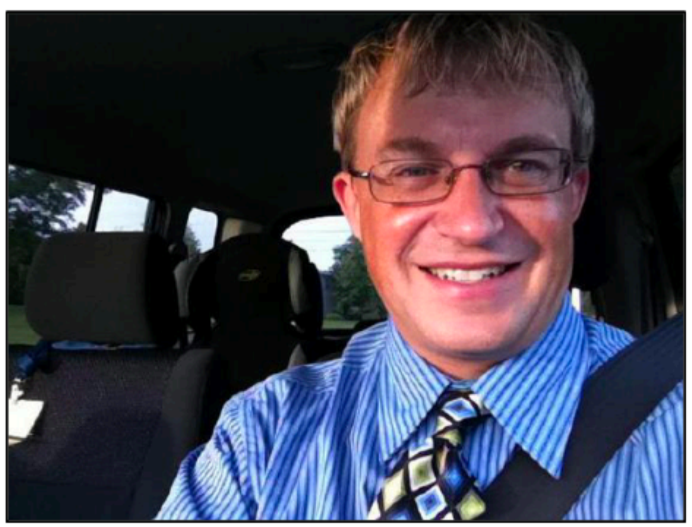
Is something under the sink broken?	yes yes yes	no no no
What number do you see?	33 33 33	5 6 7



Can you park here?	no no no	no no yes
What color is the hydrant?	white and orange white and orange white and orange	red red yellow



What kind of store is this?	bakery bakery pastry	art supplies grocery grocery
Is the display case as full as it could be?	no no no	no yes yes



Does this man have children?	yes yes yes	yes yes yes
Is this man crying?	no no no	no yes yes

Has the pizza been baked?	yes yes yes	yes yes yes
What kind of cheese is topped on this pizza?	feta feta ricotta	mozzarella mozzarella mozzarella

How many pickles are on the plate?	1 1 1	1 1 1
What is the shape of the plate?	circle round round	circle round round

Fig. 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples.

Architecture

Words

- Need ways to compare words

Next to the 'sofa' is a desk, and a 'person' is sitting behind it.

'armchair'	'man'
'bench'	'woman'
'chair'	'child'
'deck chair'	'teenager'
'ottoman'	'girl'
'seat'	'boy'
'stool'	'baby'
'swivel chair'	'daughter'
'loveseat'	'son'
…	…

word2vec

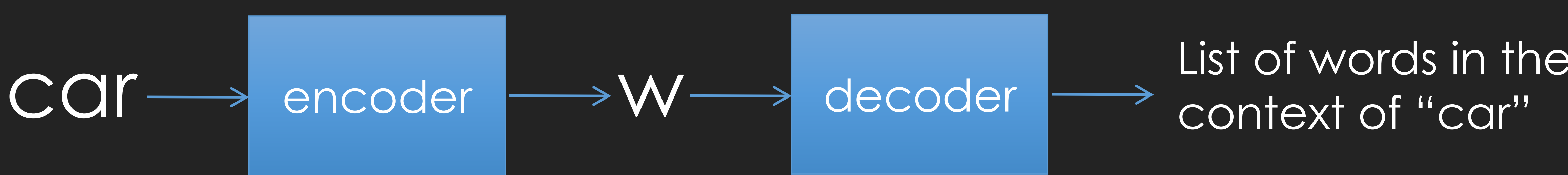
I parked the car in a nearby street. It is a red car with two doors, …

I parked the vehicle in a nearby street…

T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

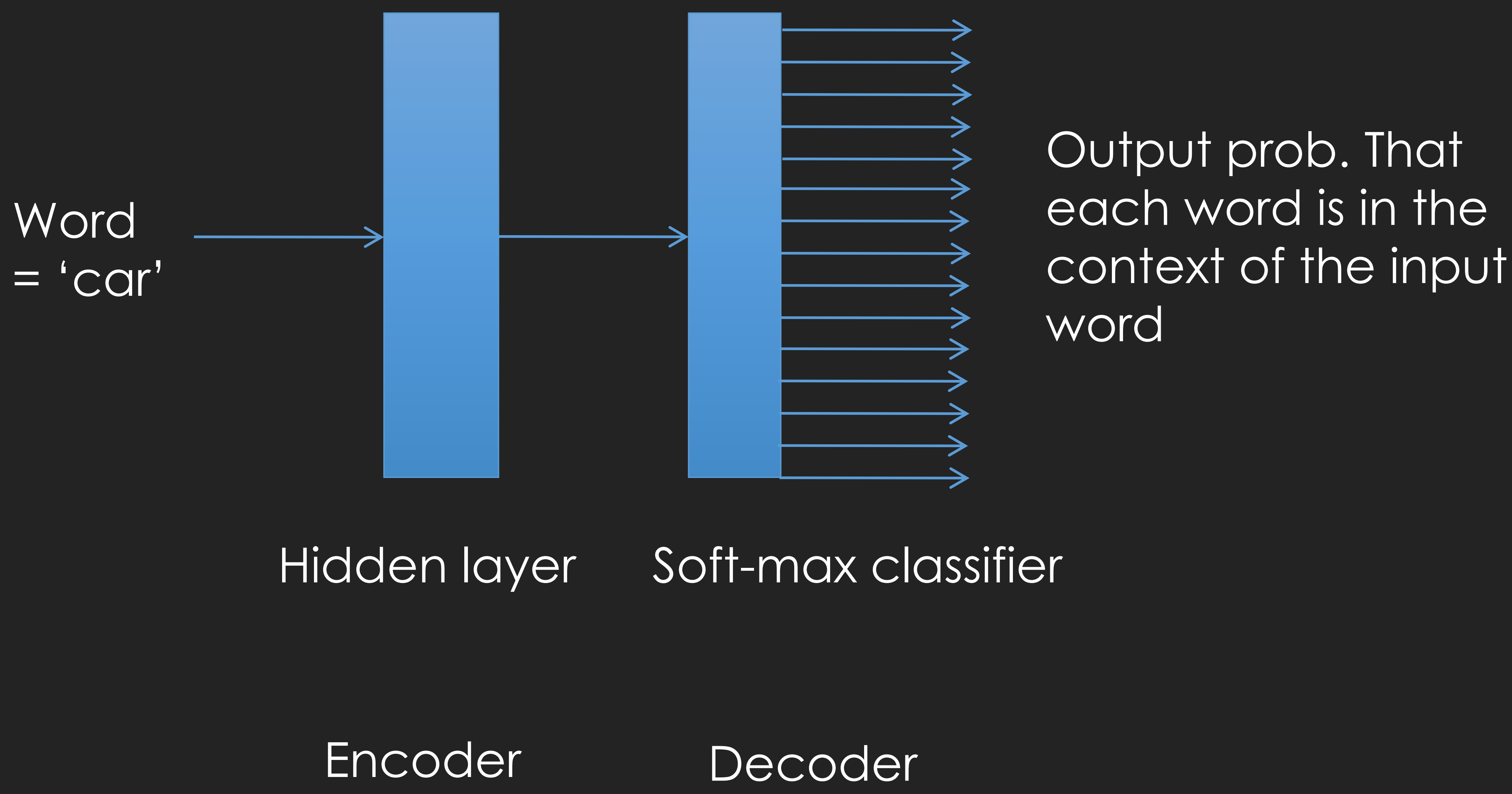
word2vec

I parked the car in a nearby street. It is a red car with two doors, ...



T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

word2vec



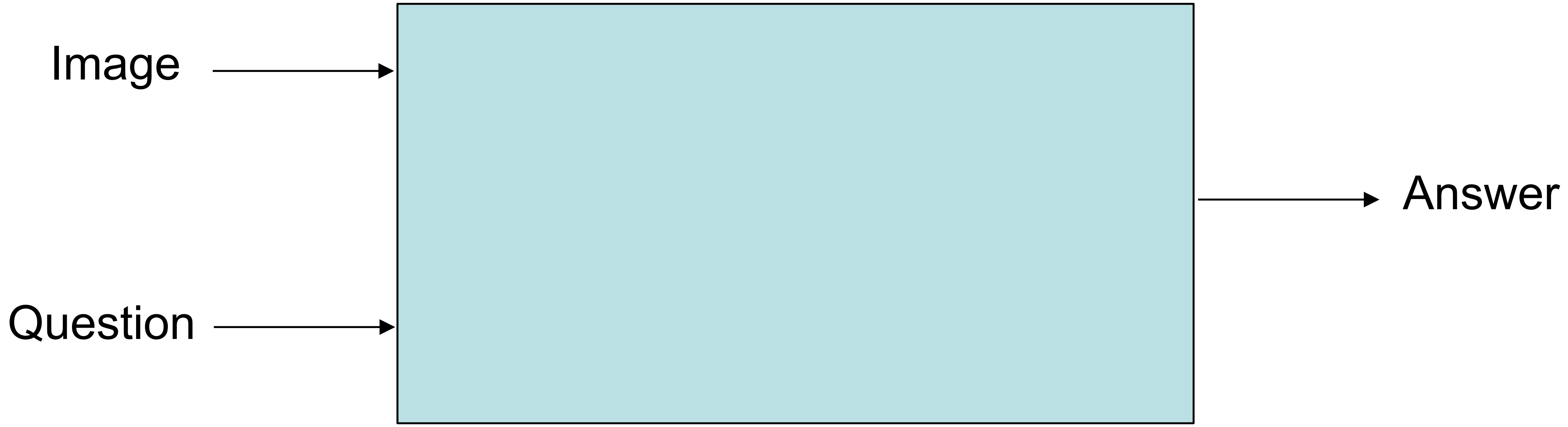
T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

Algebraic operations with the vector representation of words

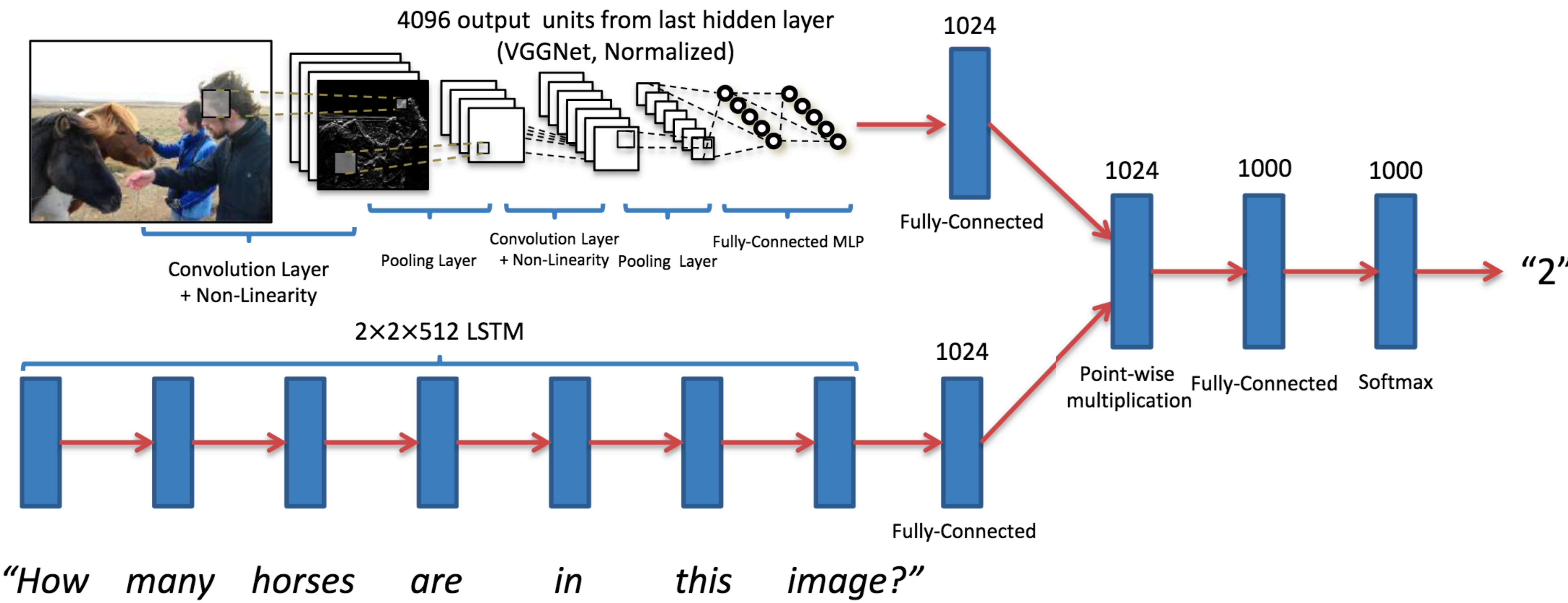
X = Vector(“Paris”) – vector(“France”) + vector(“Italy”)

Closest nearest neighbor to X is vector(“Rome”)

Architecture



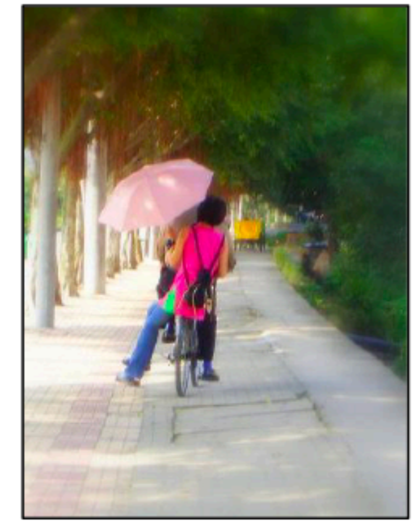
Architecture



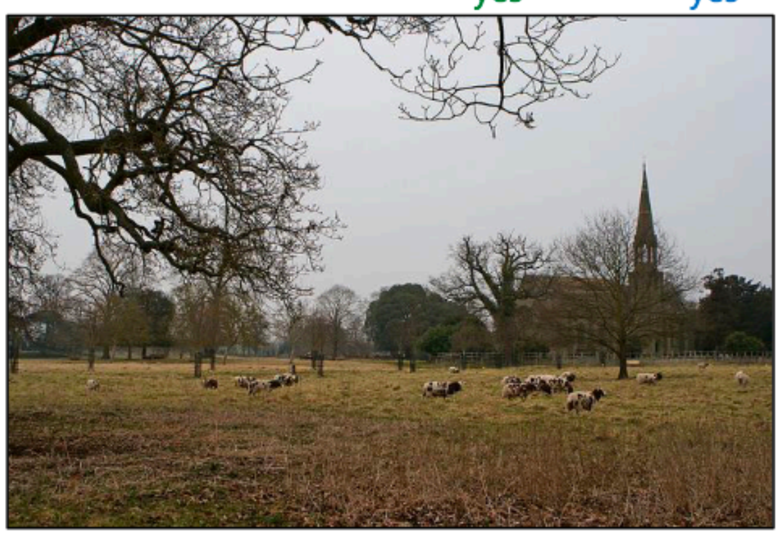
There are 1000 possible answers in this system.
Questions are unlimited.

What red objects in front are almost covered by snow?	meters parking meters parking meters	car cars shoes
Is it winter?	yes yes yes	no yes yes

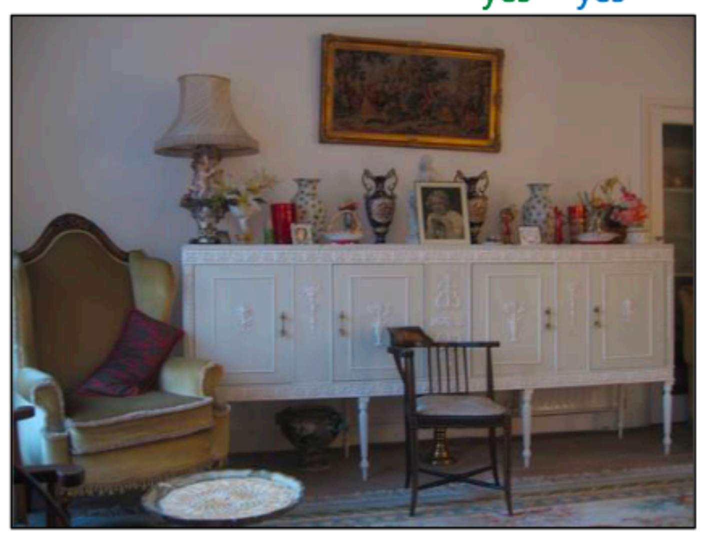
Does the car have a license plate?	yes yes yes	yes yes yes
Could the truck have a camper?	yes yes yes	yes yes yes



Is the woman on the back of the bicycle pedaling?	no no yes	no no yes
Why is the woman holding an umbrella?	sunny to block sun uncertain	it's raining it's raining to stay dry



Is this photo taken in Antarctica?	no no no	no yes yes
Overcast or sunny?	overcast overcast overcast	overcast overcast sunny



Is the picture hanging straight?	no yes yes	no yes yes
How many cabinets are on the piece of furniture?	4 4 4	3 3 6

What type of trees are here?	palm palm palm	ash oak pine
Is the skateboard airborne?	yes yes yes	no yes yes

Fig. 27: Random examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the real image dataset.

what is on the ground?

Submit

what is on the ground?

Submit

Predicted top-5 answers with confidence:

sand

90.748%

snow

2.858%

beach

1.418%

surfboards

0.677%

water

0.528%

what color is the umbrella?

Submit

what color is the umbrella?

Submit

Predicted top-5 answers with confidence:

yellow

95.090%

white

1.811%

black

0.663%

blue

0.541%

gray

0.362%

are we alone in the universe?

Submit

are we alone in the universe?

Submit

Predicted top-5 answers with confidence:

no

78.234%

yes

21.763%

people

0.001%

birds

0.000%

out

0.000%

what is the meaning of life?

Submit

what is the meaning of life?

Submit

Predicted top-5 answers with confidence:

beach
15.262%

sand
8.537%

seagull
4.708%

tower
2.393%

rocks
1.746%

what is the yellow thing?

Submit

Predicted top-5 answers with confidence:

frisbee

79.844%

surfboard

7.319%

banana

2.844%

lemon

2.438%

surfboards

1.252%

how many trains are in the picture?

Submit

Predicted top-5 answers with confidence:

3

30.233%

5

18.270%

4

17.000%

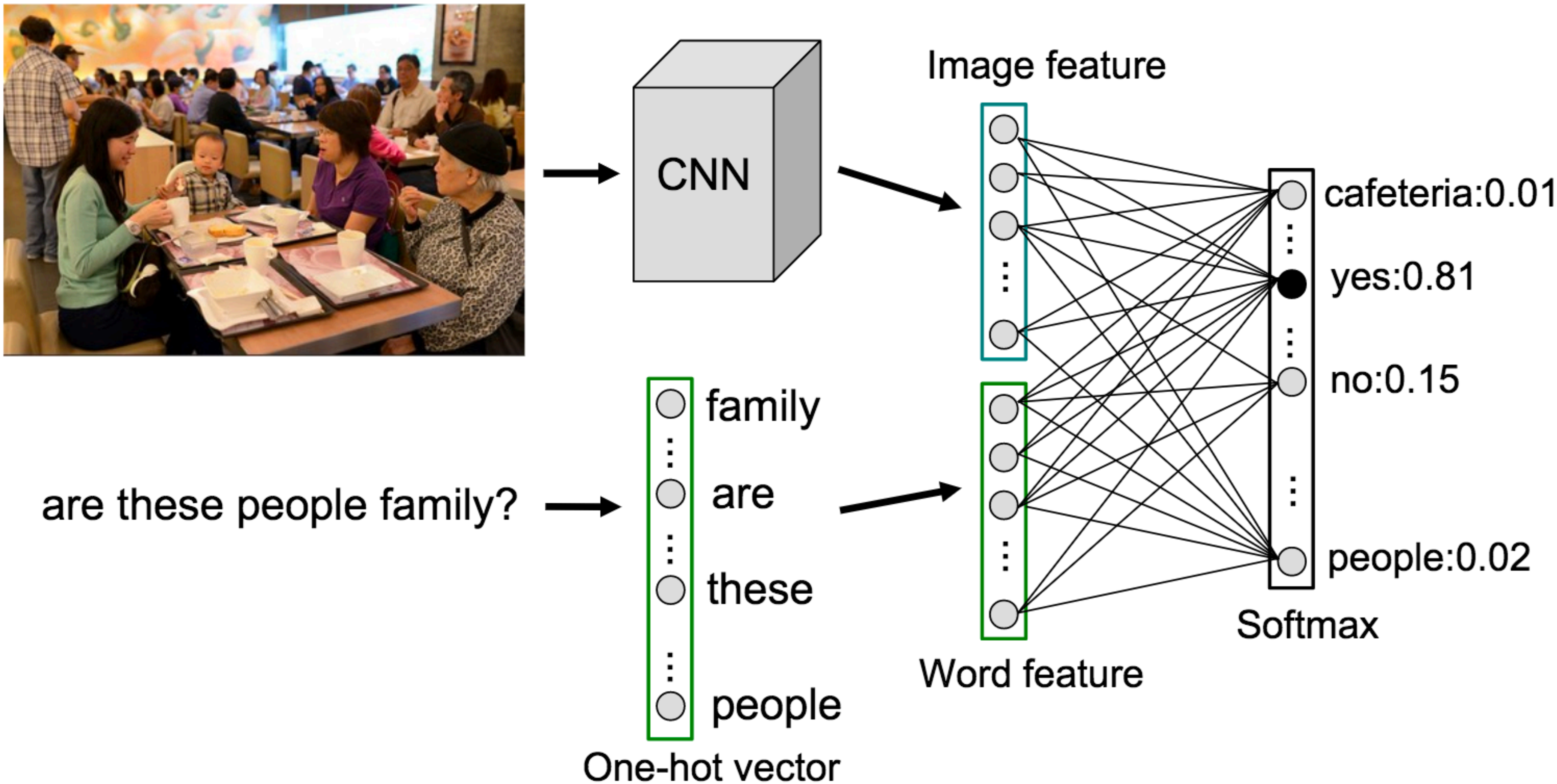
2

11.343%

6

7.806%

iBOWIMG


CNN
Image feature
cafeteria:0.01
yes:0.81
no:0.15
people:0.02
Softmax
are these people family?
family
are
these
people
One-hot vector
Word feature

iBOWIMG

Table 2: Performance comparison on test-standard.

	Open-Ended				Multiple-Choice			
	Overall	yes/no	number	others	**Overall**	yes/no	number	others
LSTMIMG [2]	54.06	-	-	-	-	-	-	-
NMN+LSTM [1]	55.10	-	-	-	-	-	-	-
ACK [16]	55.98	79.05	36.10	40.61	-	-	-	-
DPPnet [11]	**57.36**	80.28	36.92	42.24	**62.69**	80.35	38.79	52.79
iBOWIMG	55.89	76.76	34.98	42.62	61.97	76.86	37.30	54.60

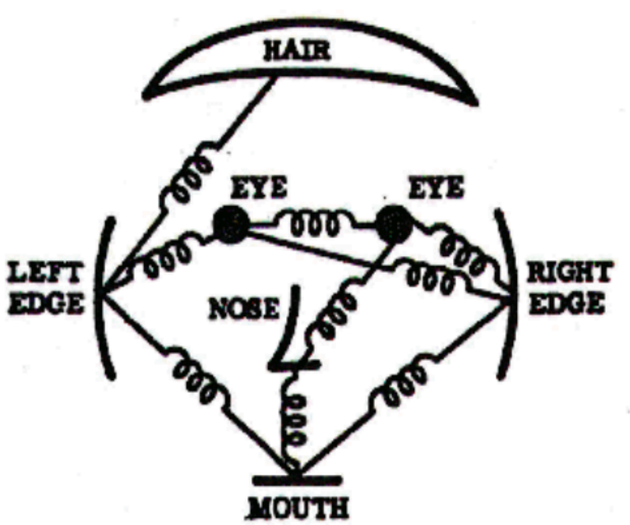
Two Extremes of Vision

Extrapolation problem
Generalization
Diagnostic features

Interpolation problem
Correspondence
Finding the differences

1 | 10 | 10^2 | 10^3 | 10^4 | 10^5 | 10^6 | ∞

Number of training samples



Slide credit: Aude Oliva