Lecture 14: Representation learning and language

Announcements

- Project proposal info out
- Friday section: project office hours

- Ad: ECE Cider and Donuts: https://eecs.engin.umich.edu/event/ece-cider-and-donuts/ Tuesday at 9am in 3313 EECS

Supervised computer vision

Object recognition [Russakovsky et al., “ImageNet”, 2015]

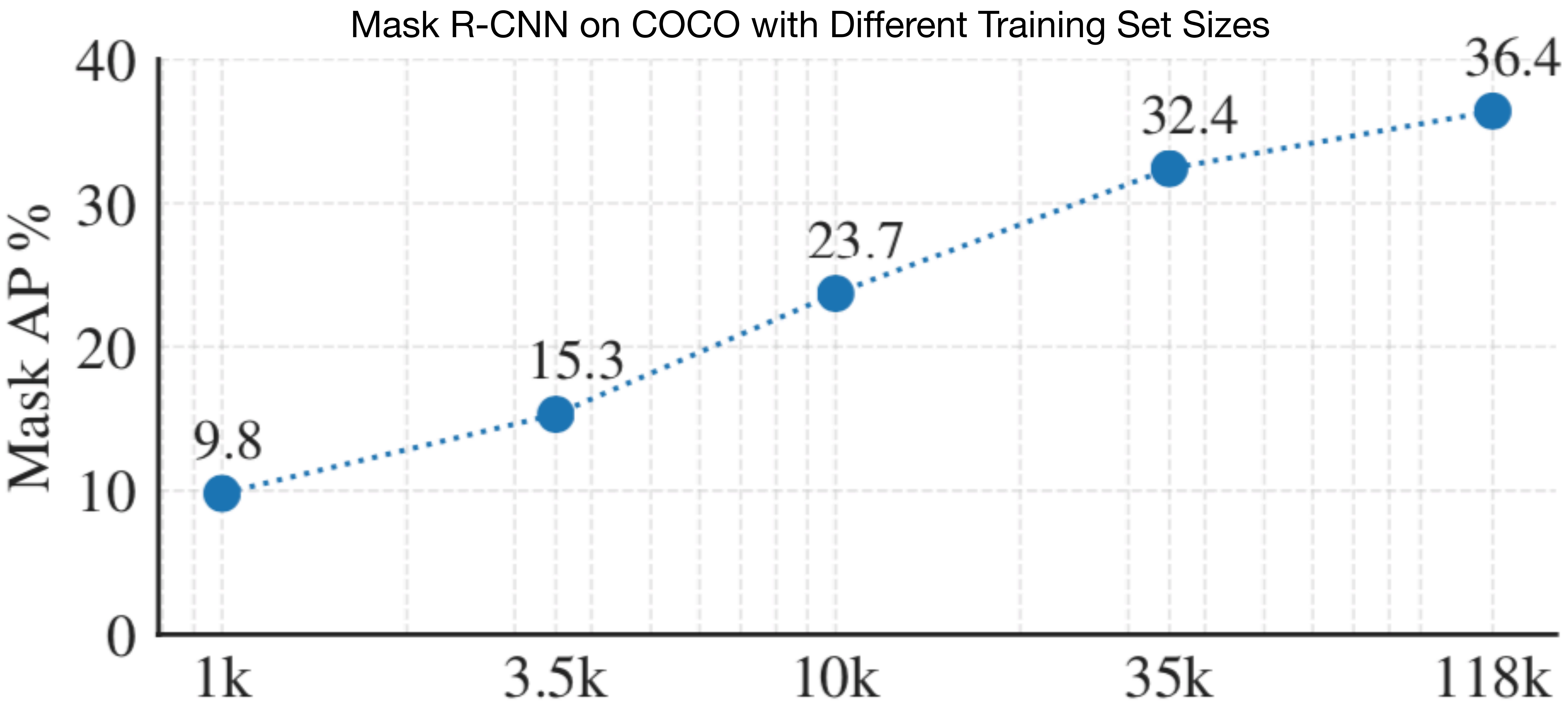
Object segmentation [Gupta et al., “LVIS”, 2019]

Supervised computer vision

These methods need lots of labeled training examples.

[Lin et al., COCO dataset]

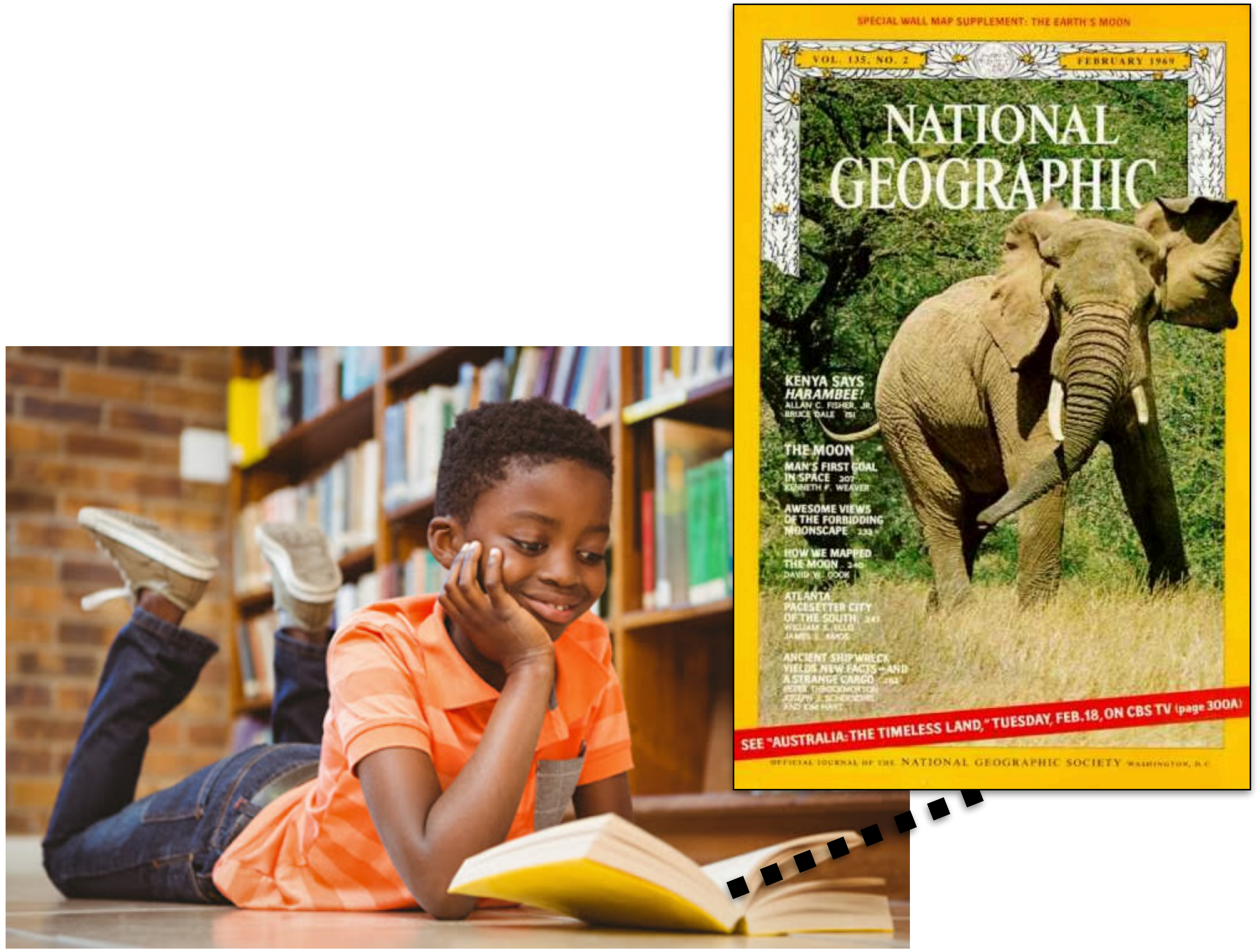
We still need *lots* of labeled examples



Object detection accuracy vs. dataset size

Image source: R. Girshick

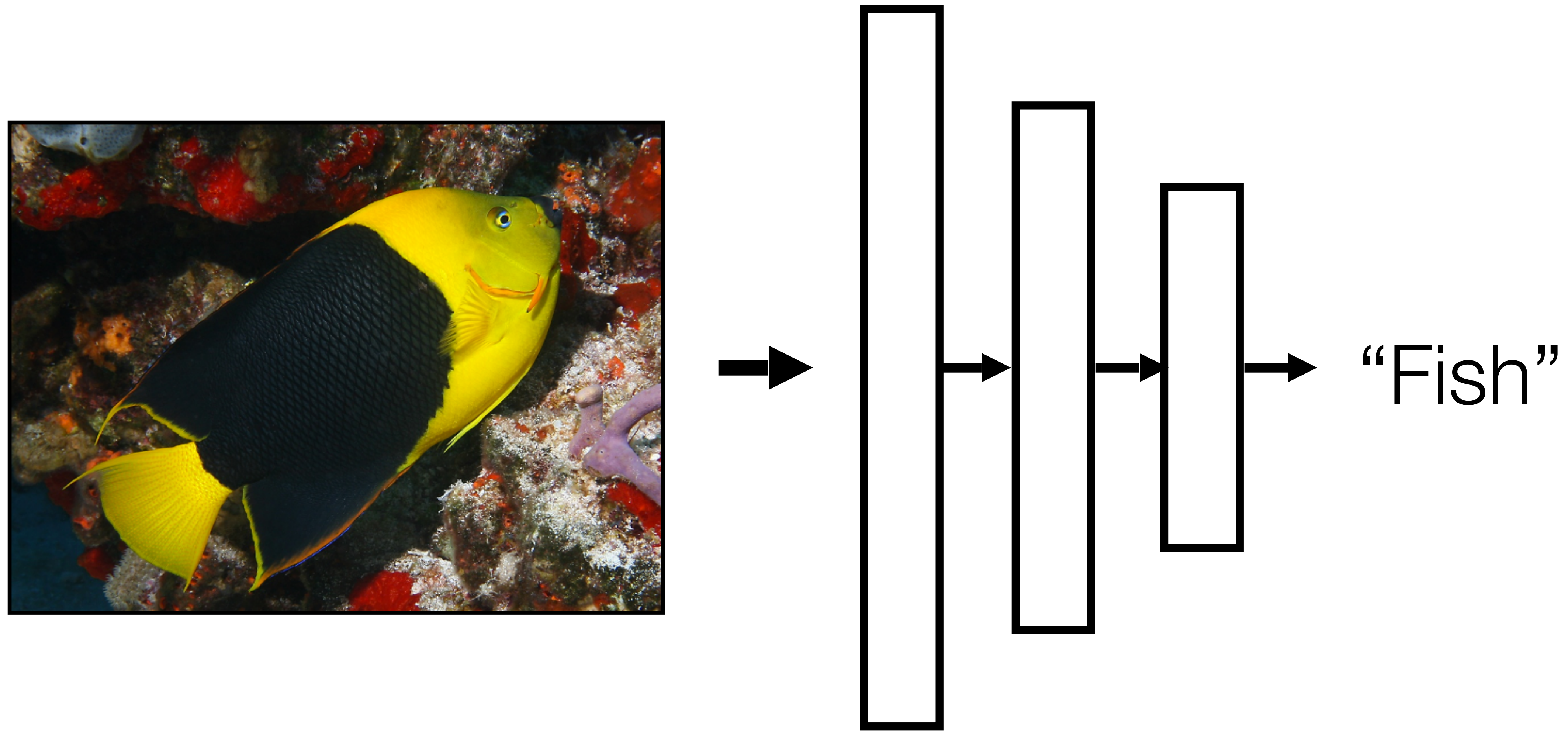
We want models that can generalize



Transfer learning

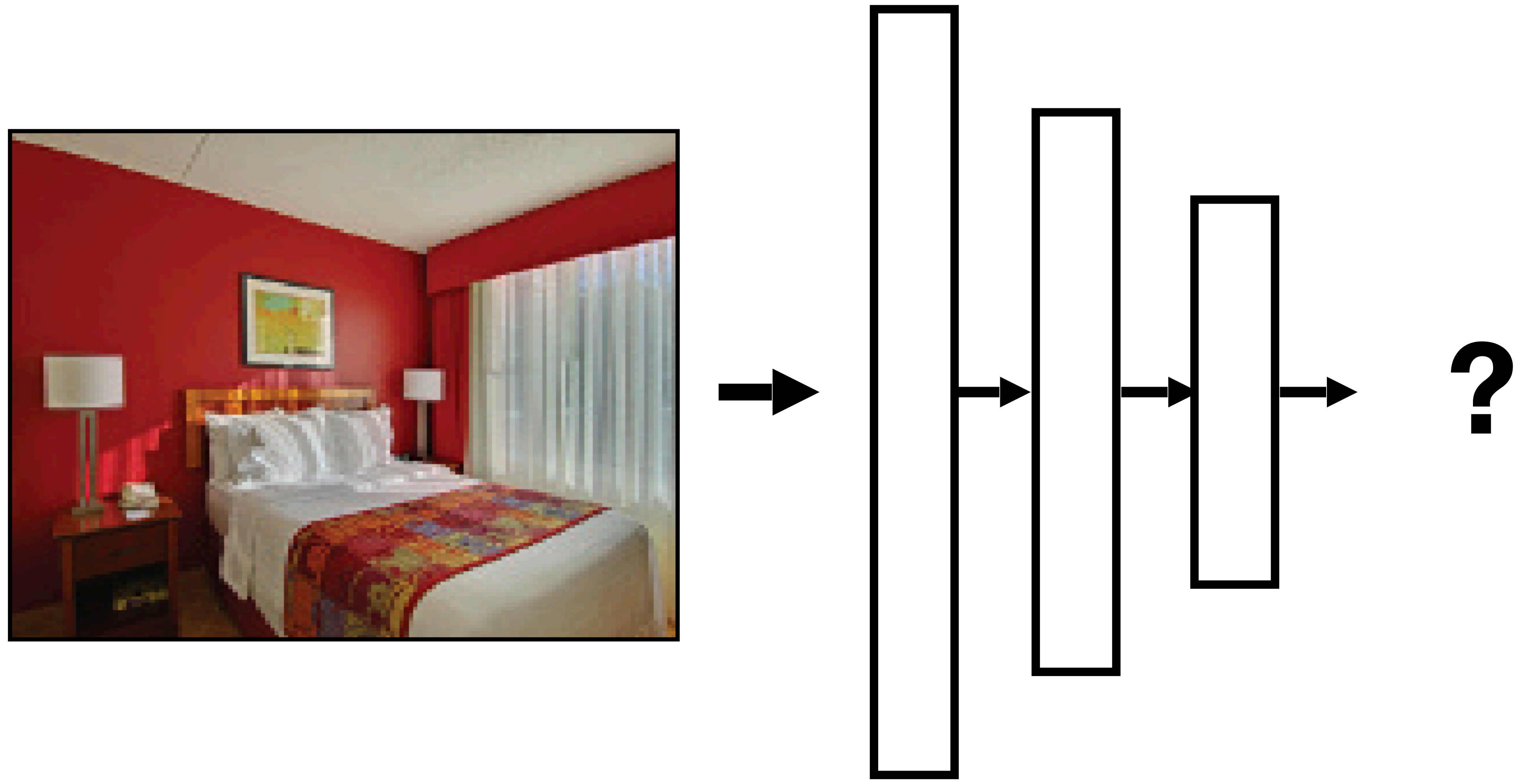
Training

Object recognition



Testing

Scene recognition

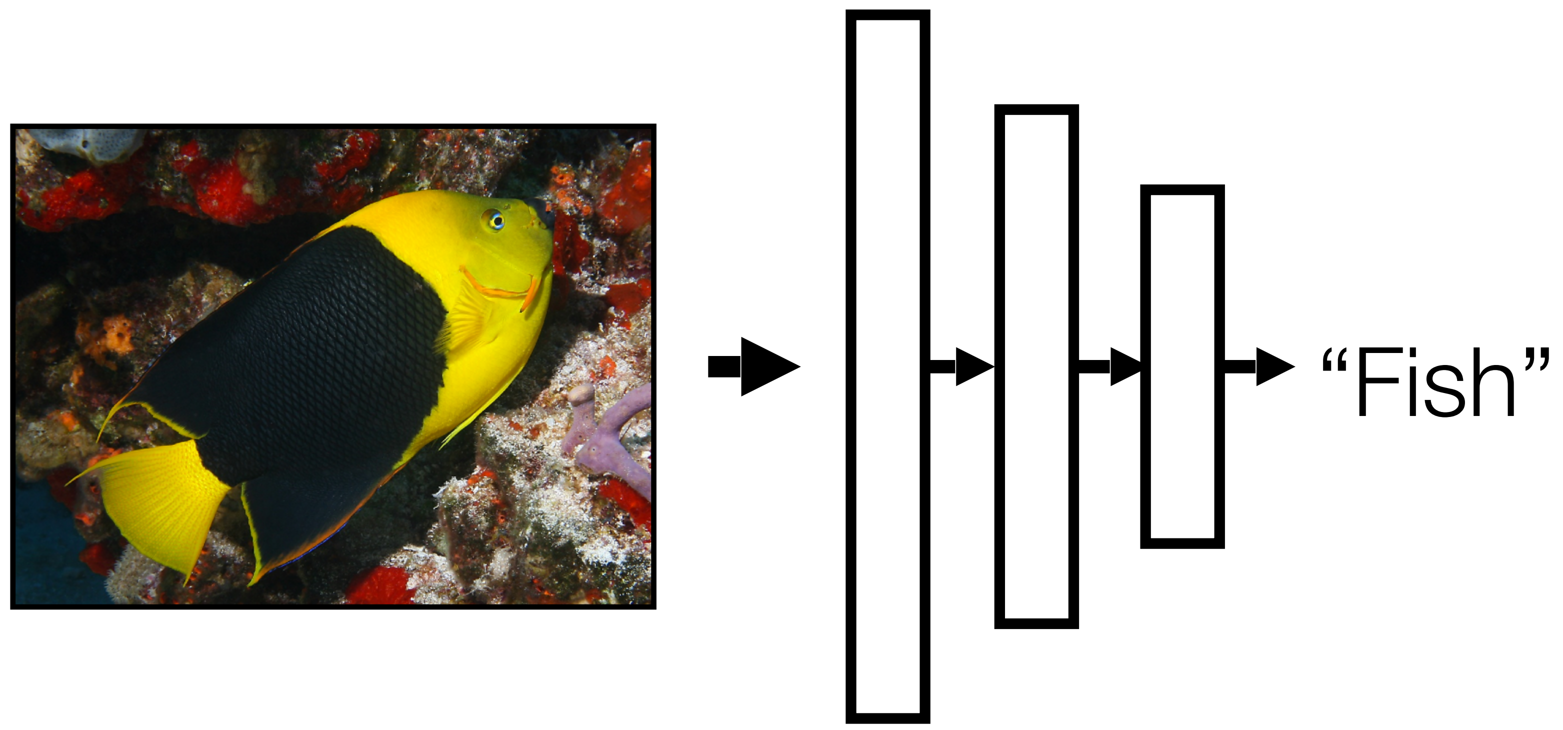


Often, what we will be “tested” on is to learn to do something new.

Source: Isola, Freeman, Torralba

Pretraining

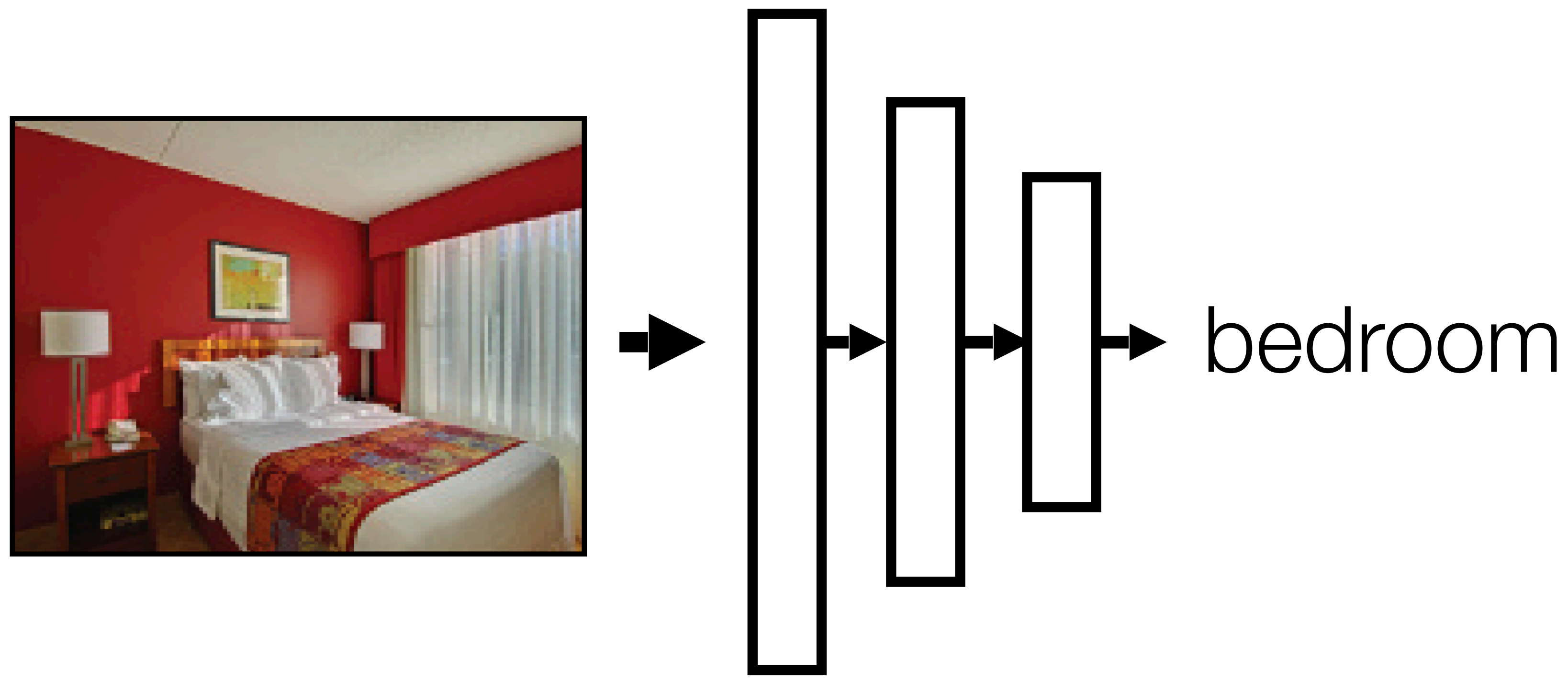
Object recognition



A lot of data

Finetuning

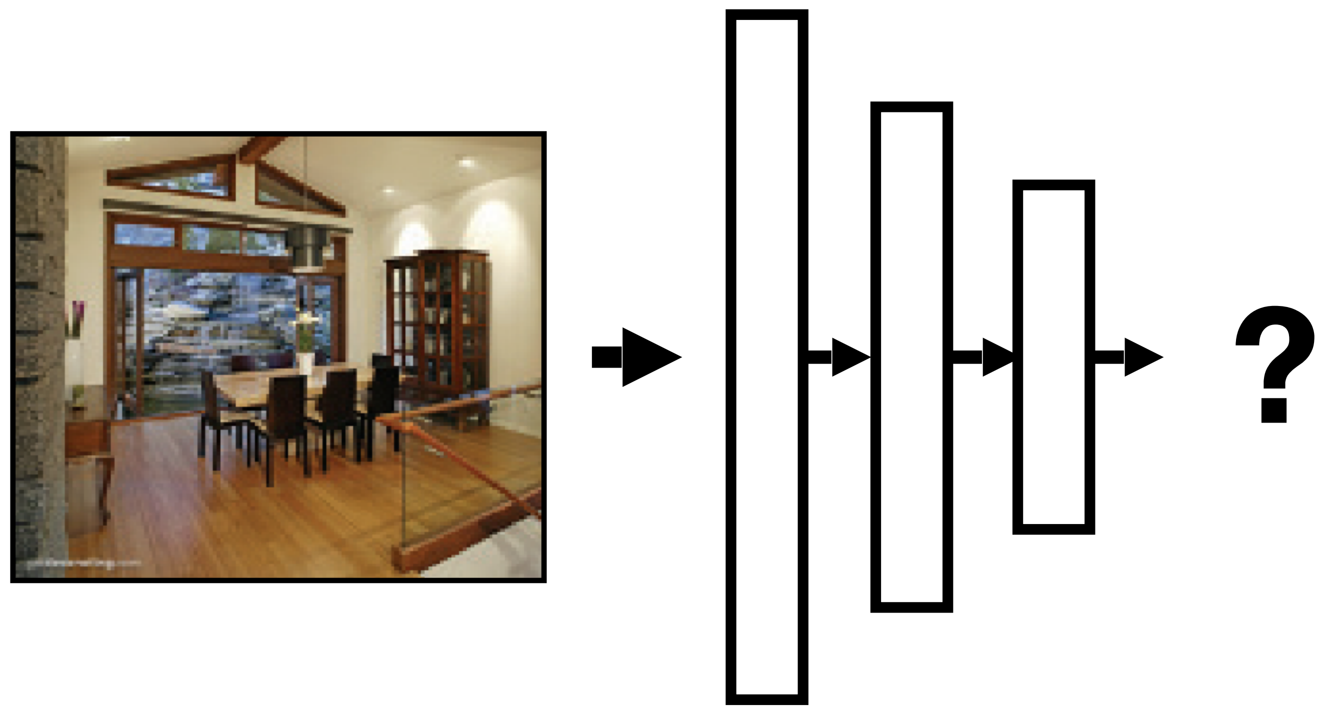
Scene recognition



A little data

Testing

Scene recognition



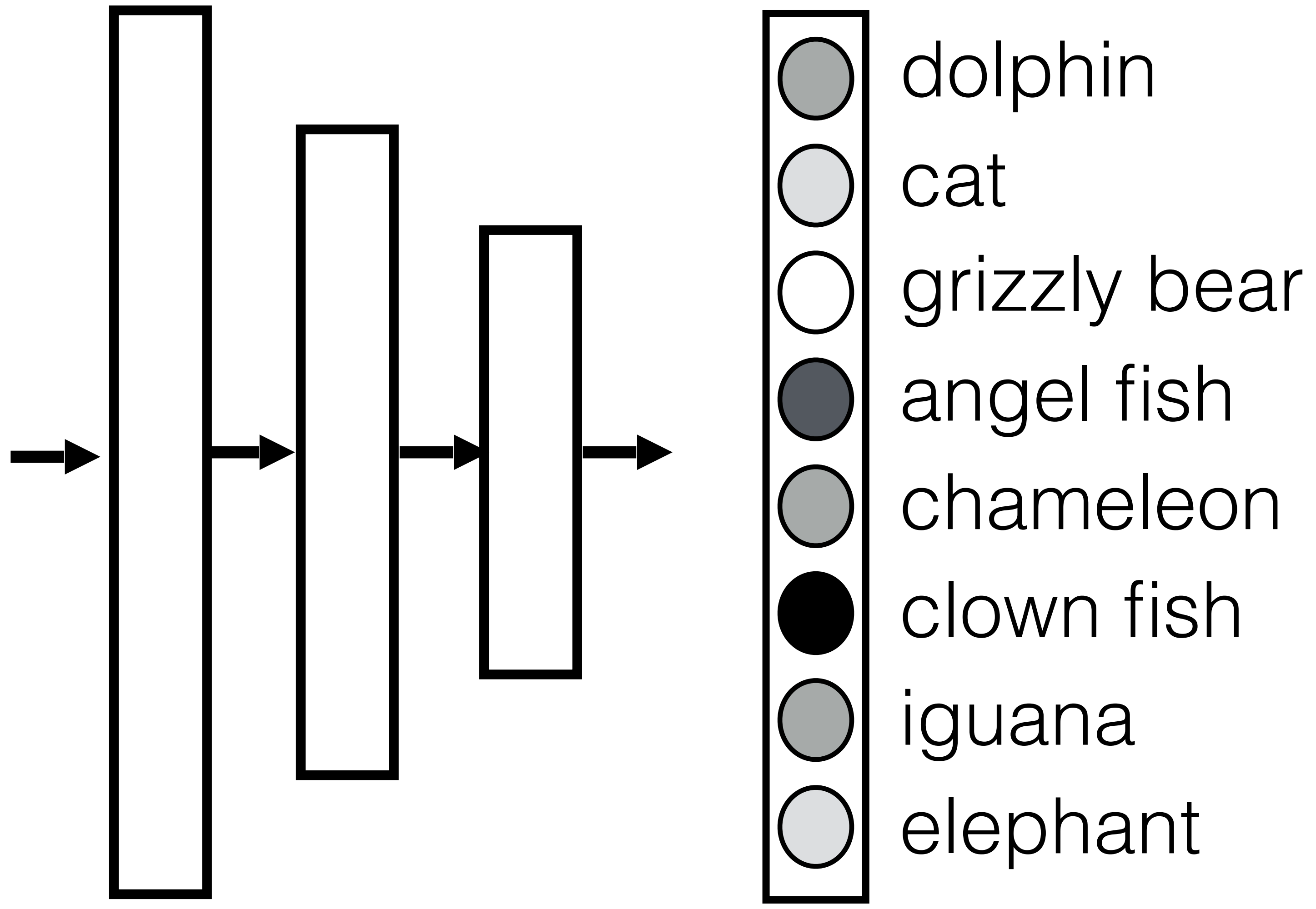
Finetuning starts with the representation learned on a previous task, and adapts it to perform well on a new task.

Source: Isola, Freeman, Torralba

Finetuning

Pretraining

Object recognition

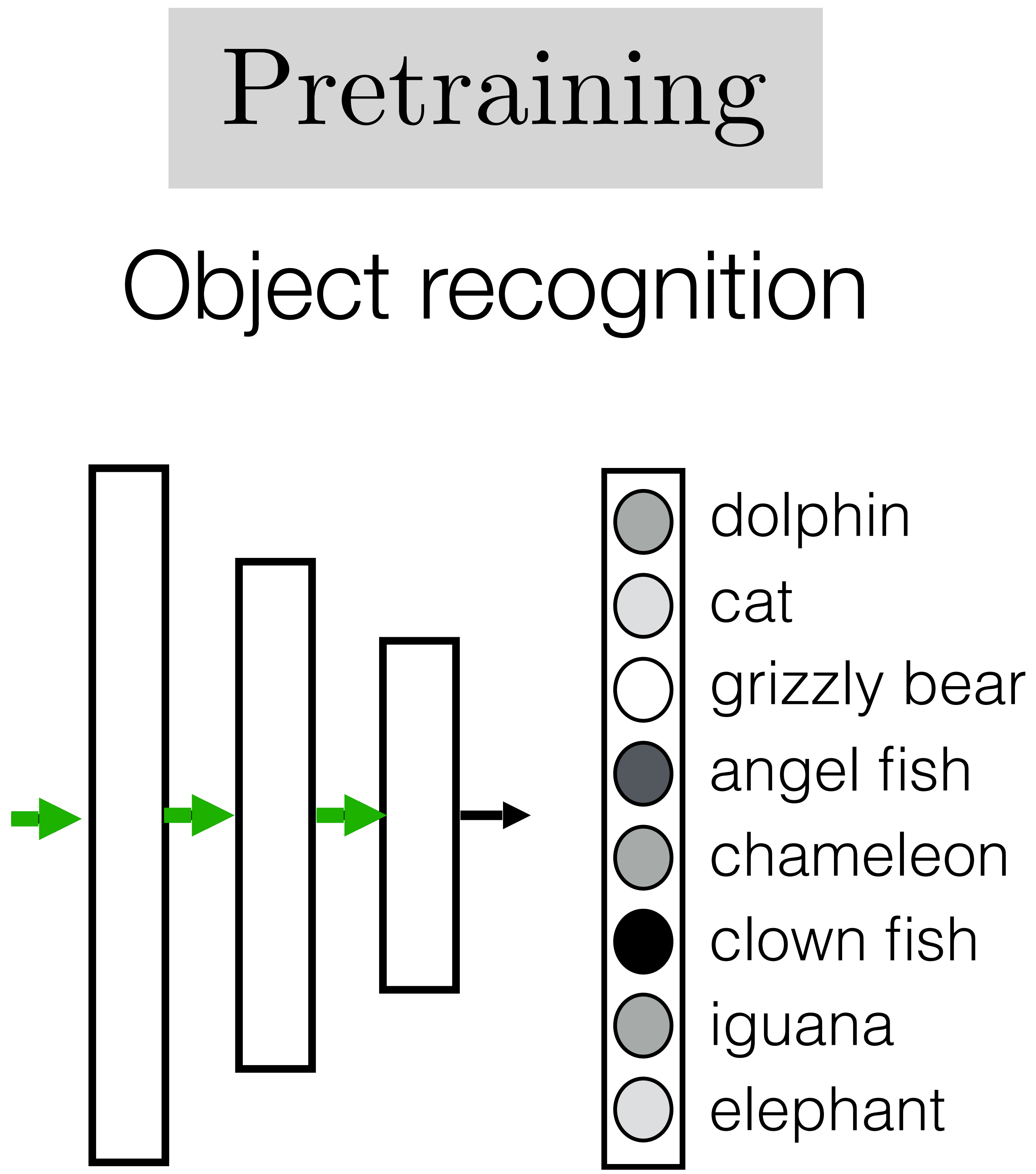


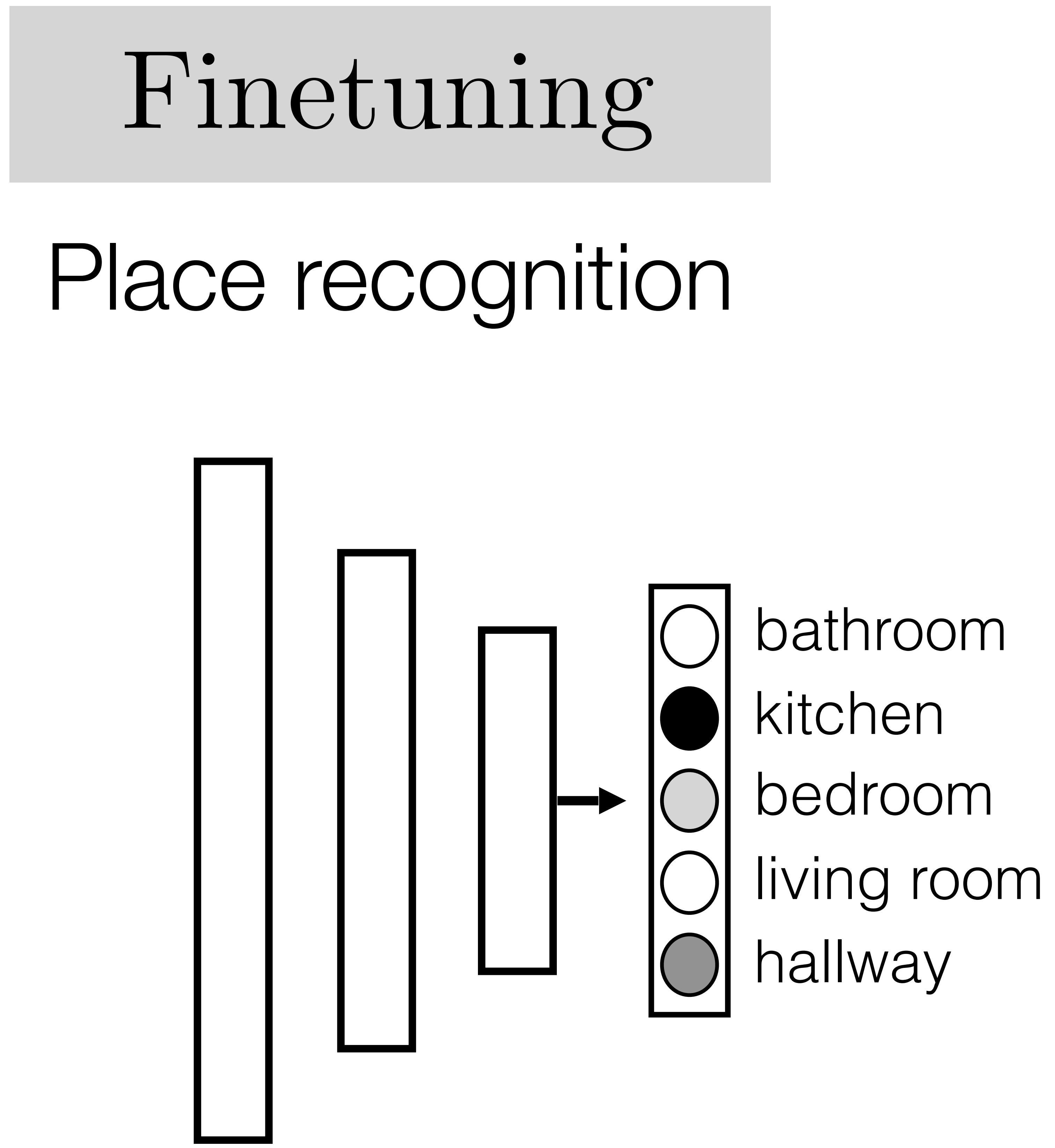
Finetuning

Scene recognition

Source: Isola, Freeman, Torralba

Finetuning





Initialize the weights using the pretraining task!

Source: Isola, Freeman, Torralba

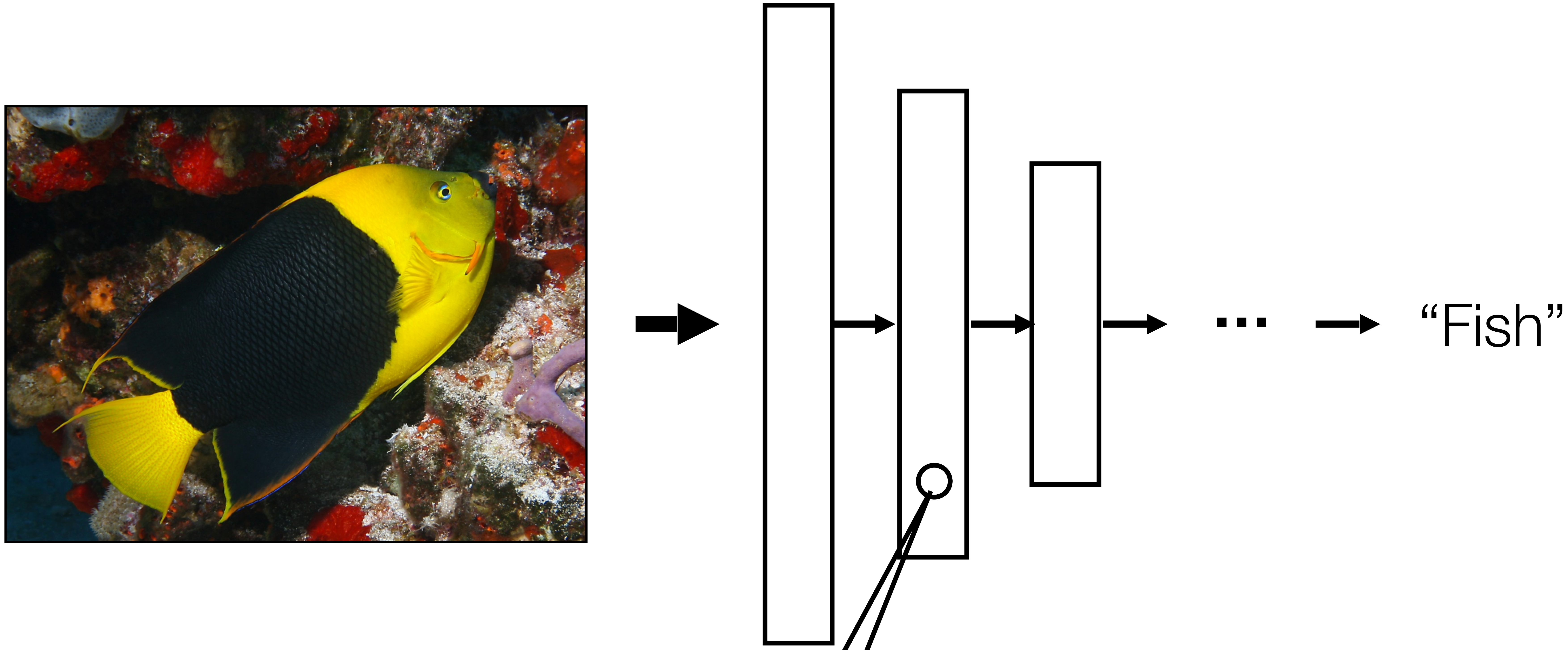
Finetuning

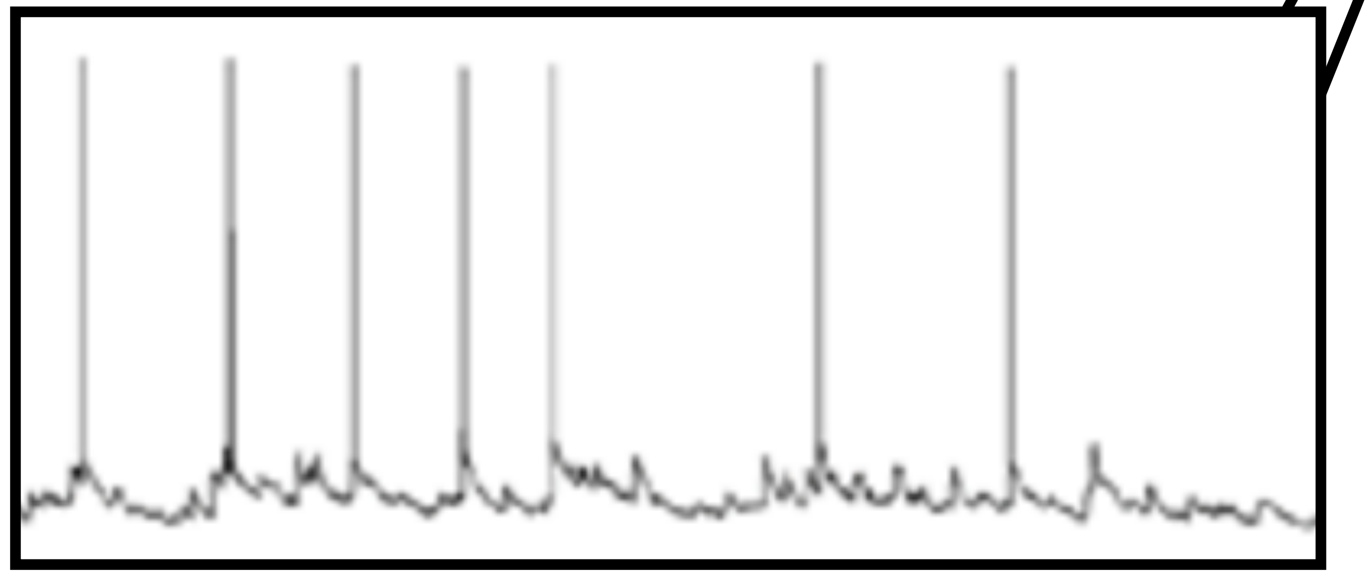
- Pretrain a network on task A (e.g., object recognition), resulting in parameters **W**.
- Initialize a second network with some or all of **W**.
- Train the second network on task B, resulting in parameters **W’**
- Why would we expect this to work?

Source: Isola, Freeman, Torralba

Visualizing representations

Deep net "electrophysiology"





[Zeiler & Fergus, ECCV 2014]
[Zhou et al., ICLR 2015]

Source: Isola, Freeman, Torralba

Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

Gabor-like filters learned by **layer 1**

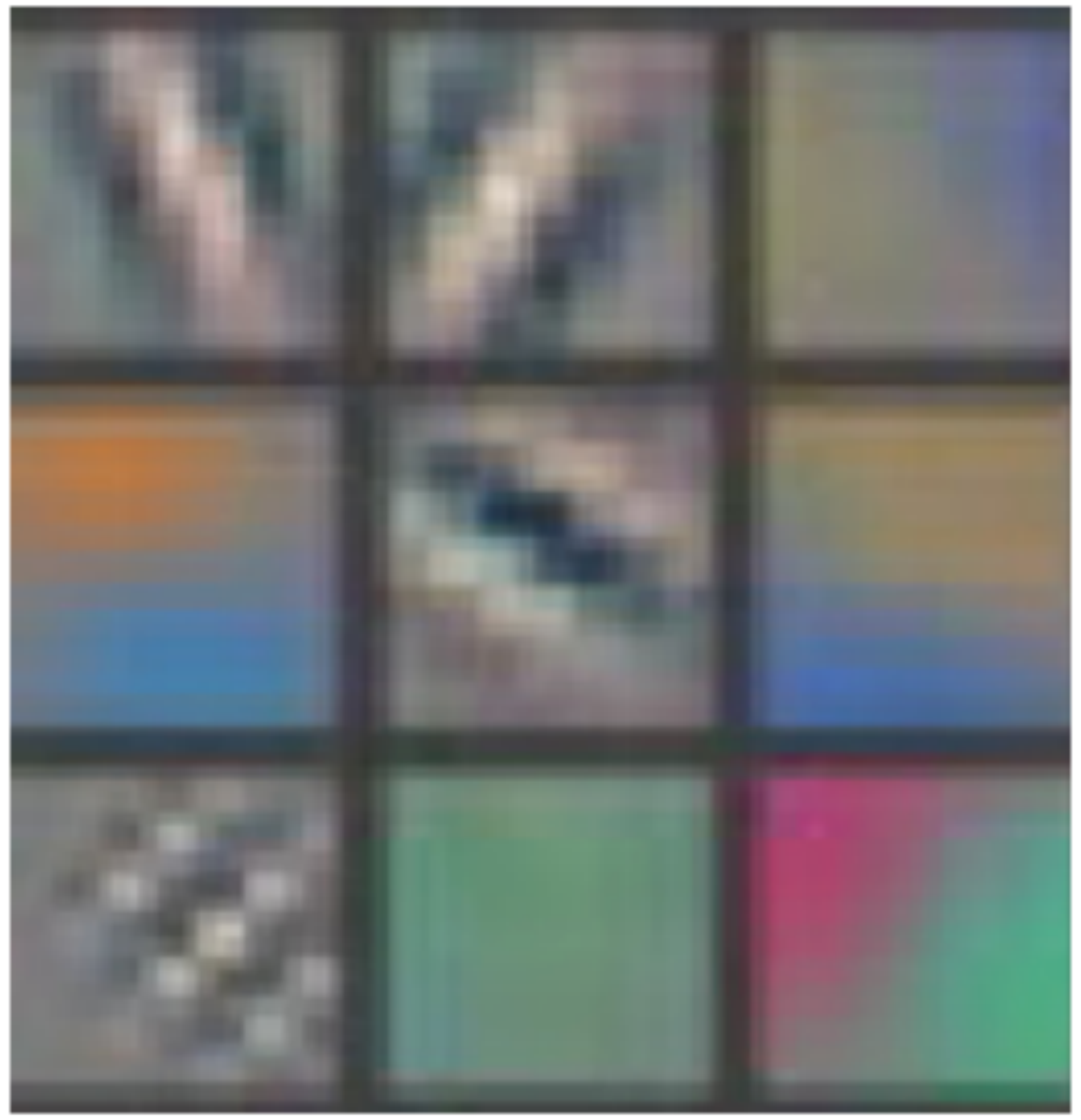
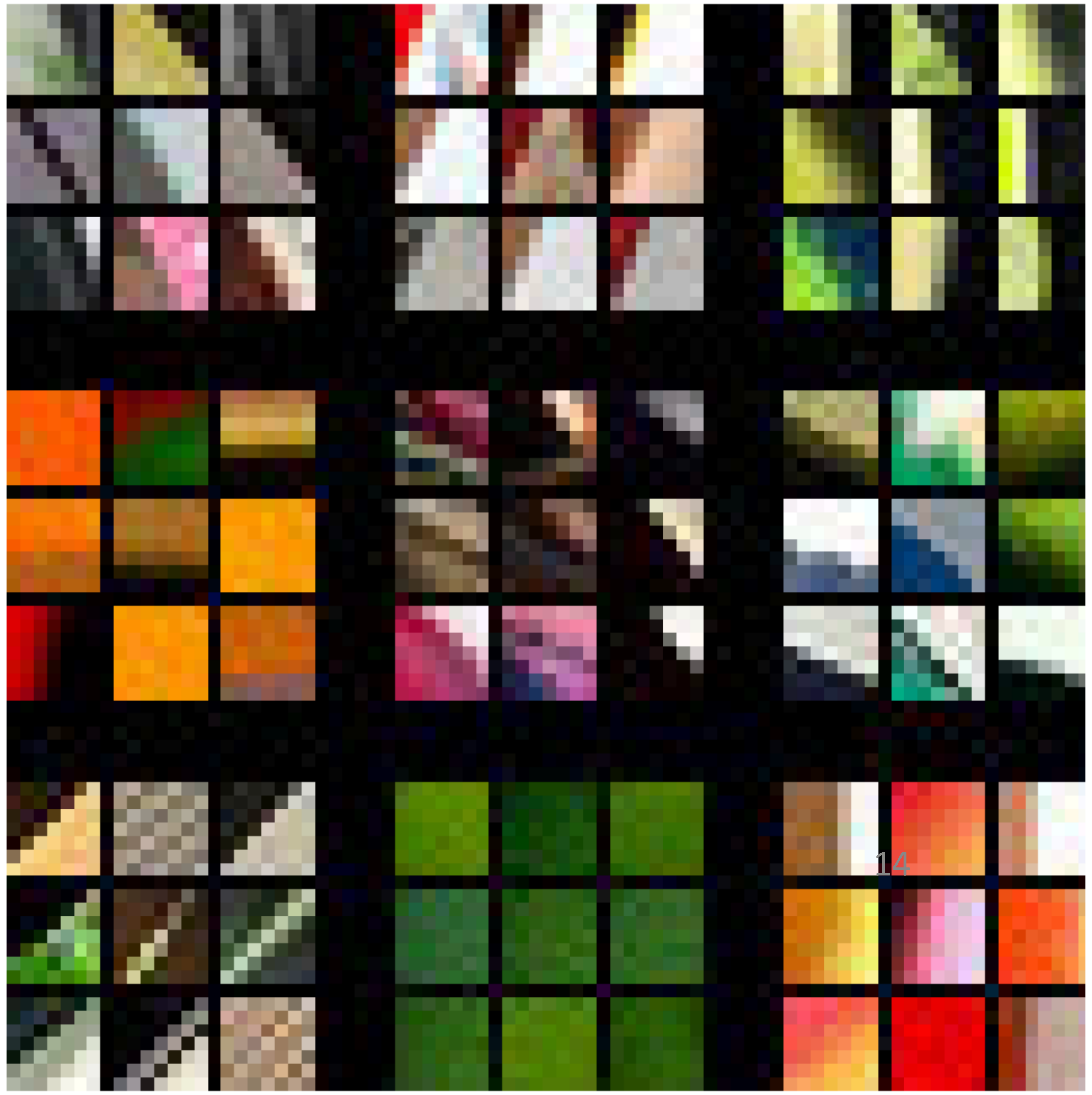


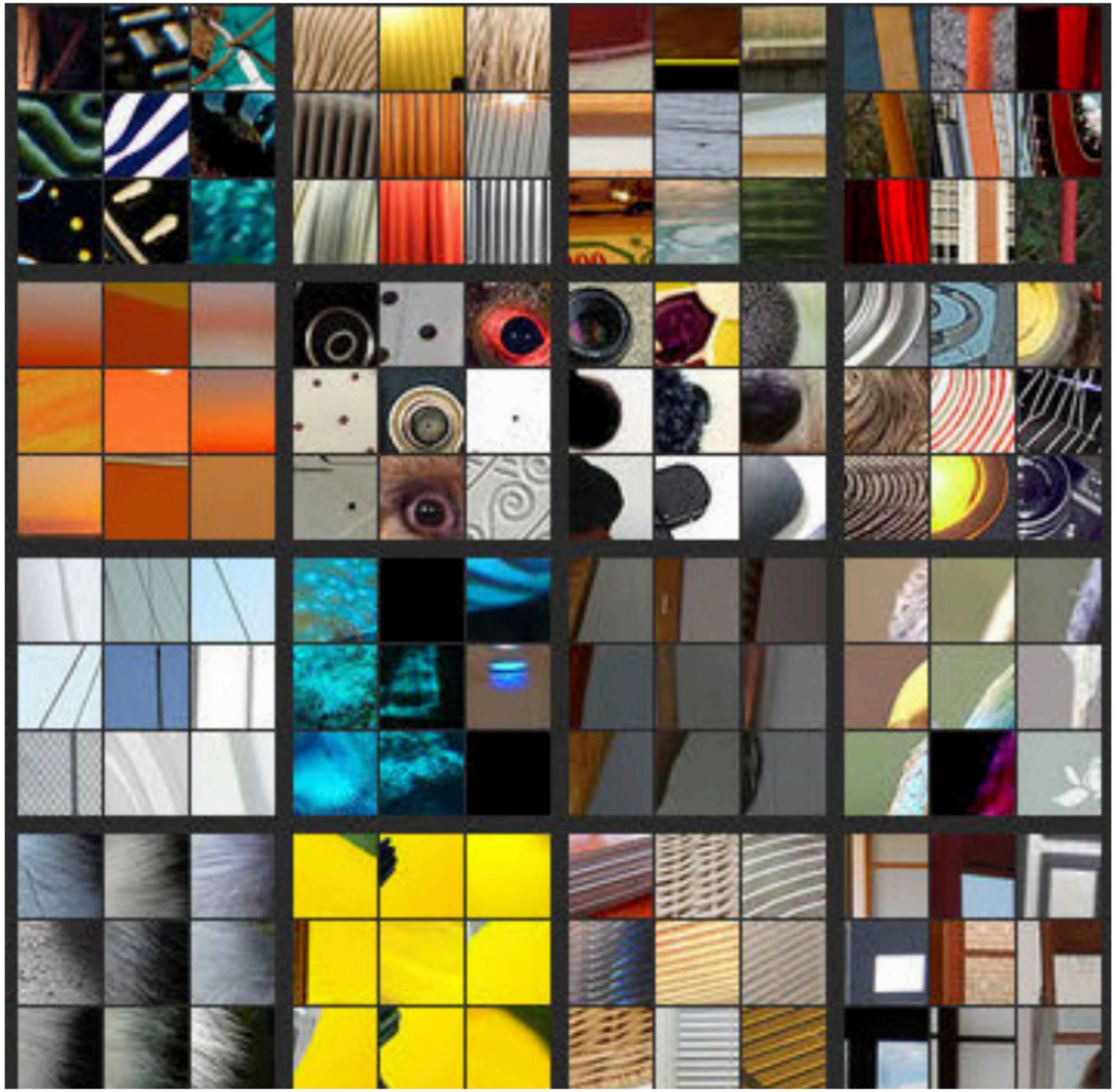
Image patches that activate each of the **layer 1** filters most strongly



Source: Isola, Freeman, Torralba

[Zeiler and Fergus, 2014]

Image patches that activate each of the **layer 2** neurons most strongly



Source: Isola, Freeman, Torralba

[Zeiler and Fergus, 2014]

Image patches that activate each of the **layer 3** neurons most strongly

Source: Isola, Freeman, Torralba

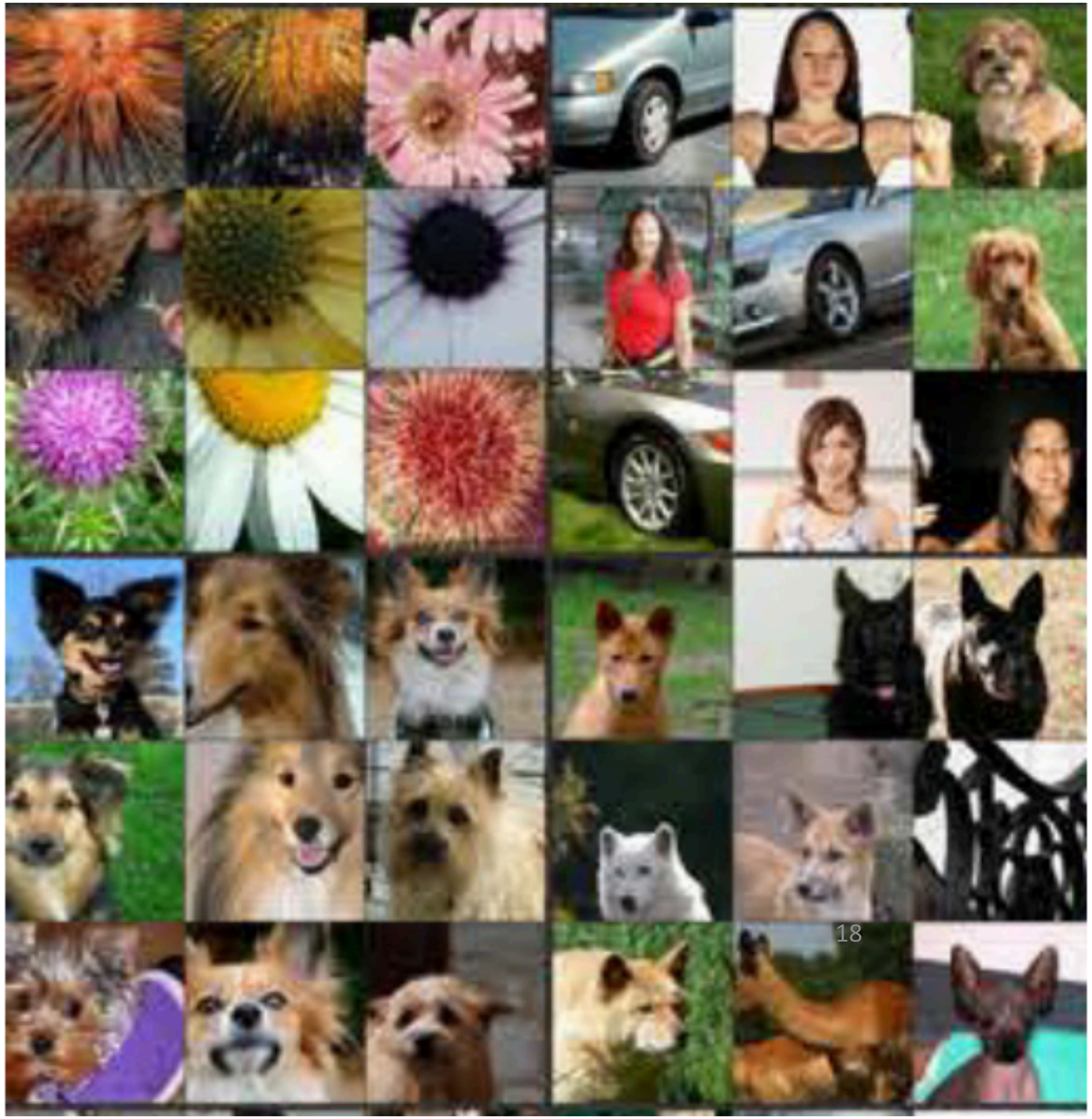
[Zeiler and Fergus, 2014]

Image patches that activate each of the **layer 4** neurons most strongly

Source: Isola, Freeman, Torralba

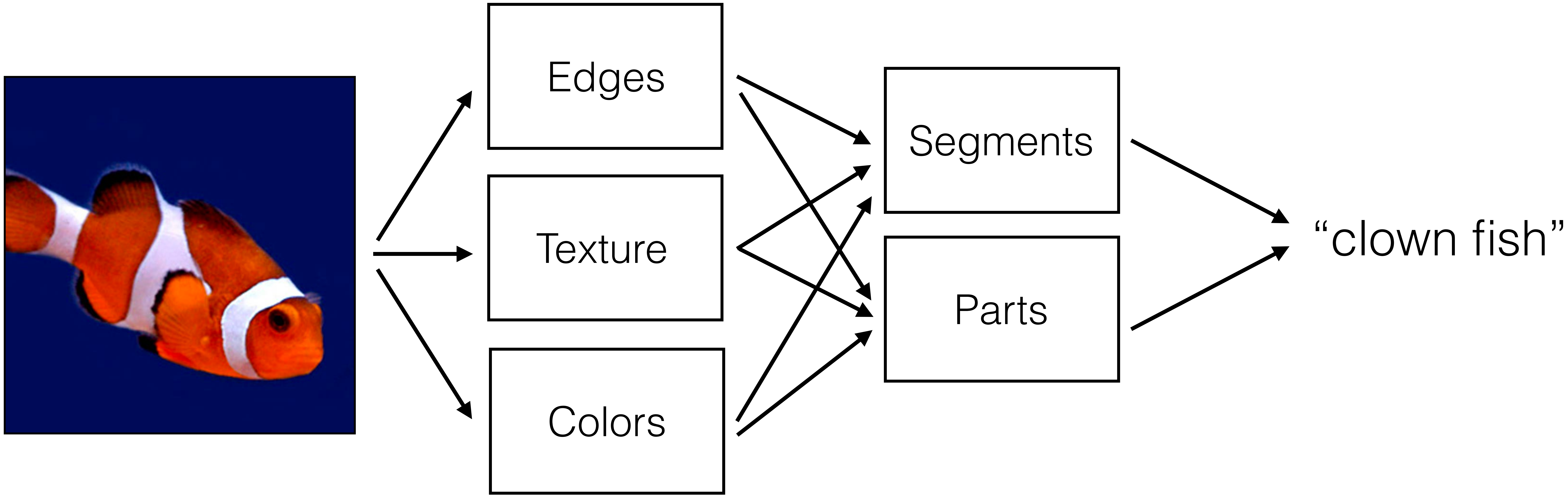
[Zeiler and Fergus, 2014]

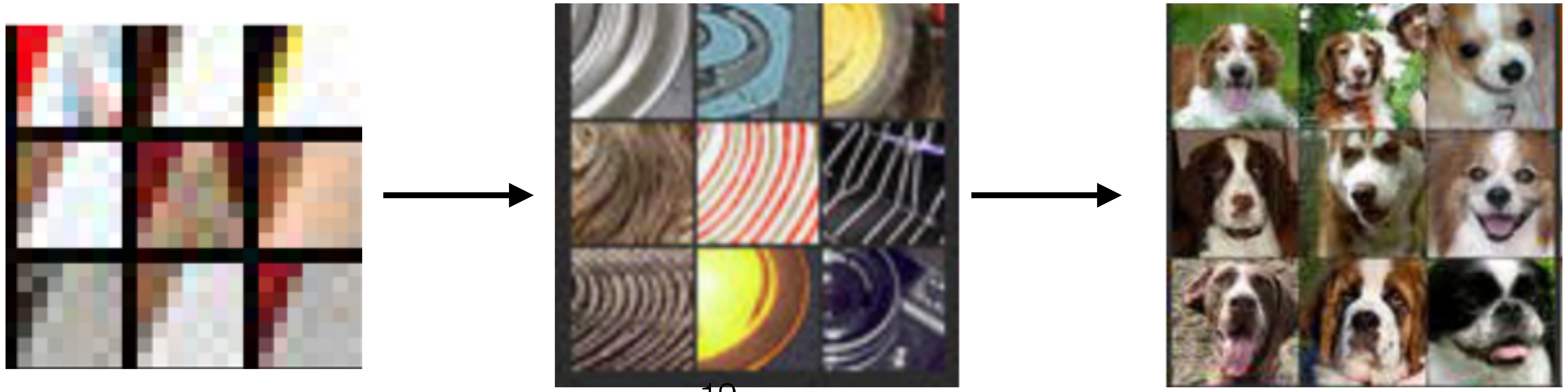
Image patches that activate each of the **layer 5** neurons most strongly



Source: Isola, Freeman, Torralba

CNNs learned the classical visual recognition pipeline





Source: Isola, Freeman, Torralba

Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]

pool 1

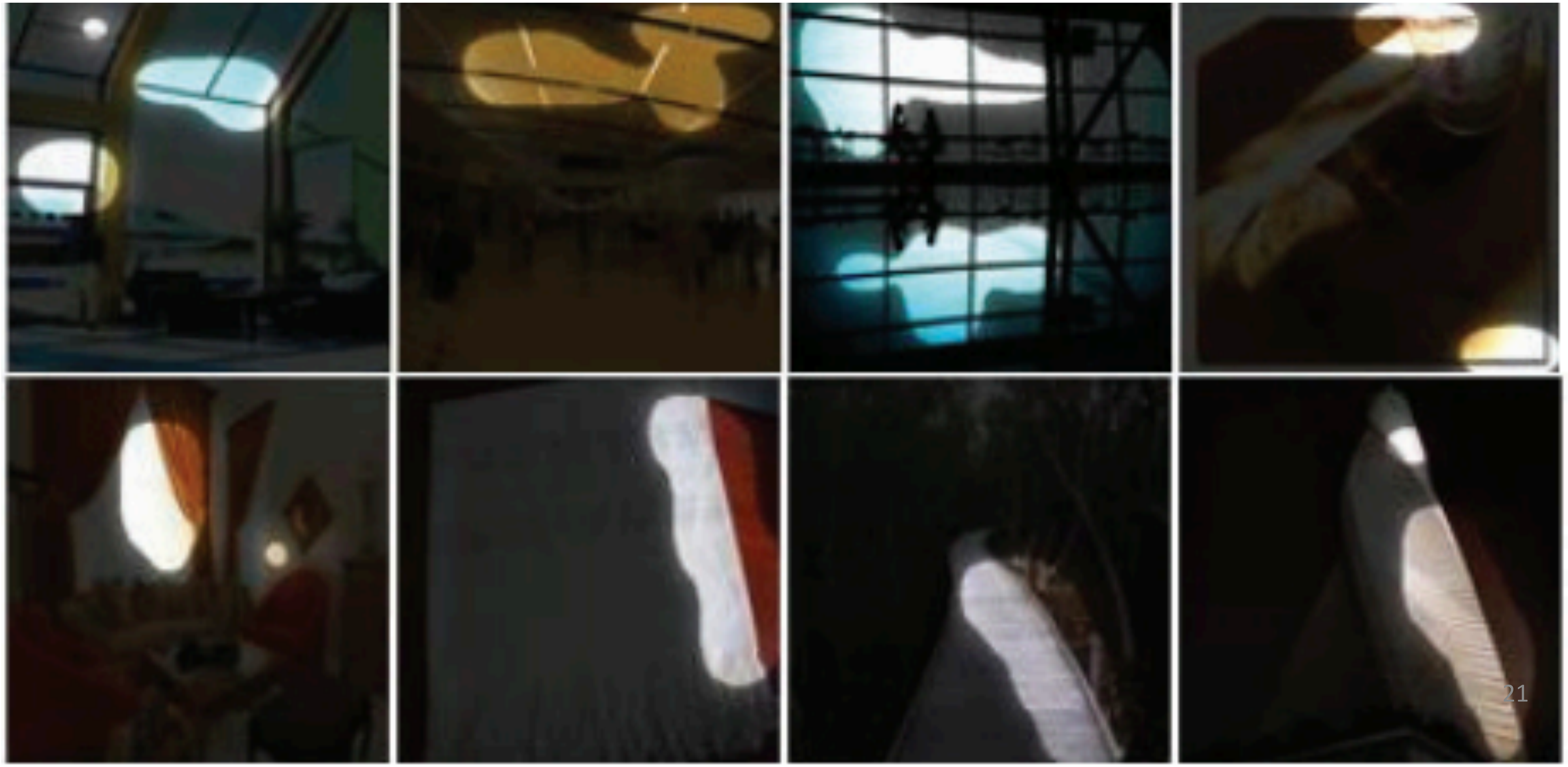
[http://people.csail.mit.edu/torralba/research/drawCNN/drawNet.html]

Source: Isola, Freeman, Torralba

Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]

pool 2



Source: Isola, Freeman, Torralba

Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]

conv 4

Source: Isola, Freeman, Torralba

Object Detectors Emerge in Deep Scene CNNs

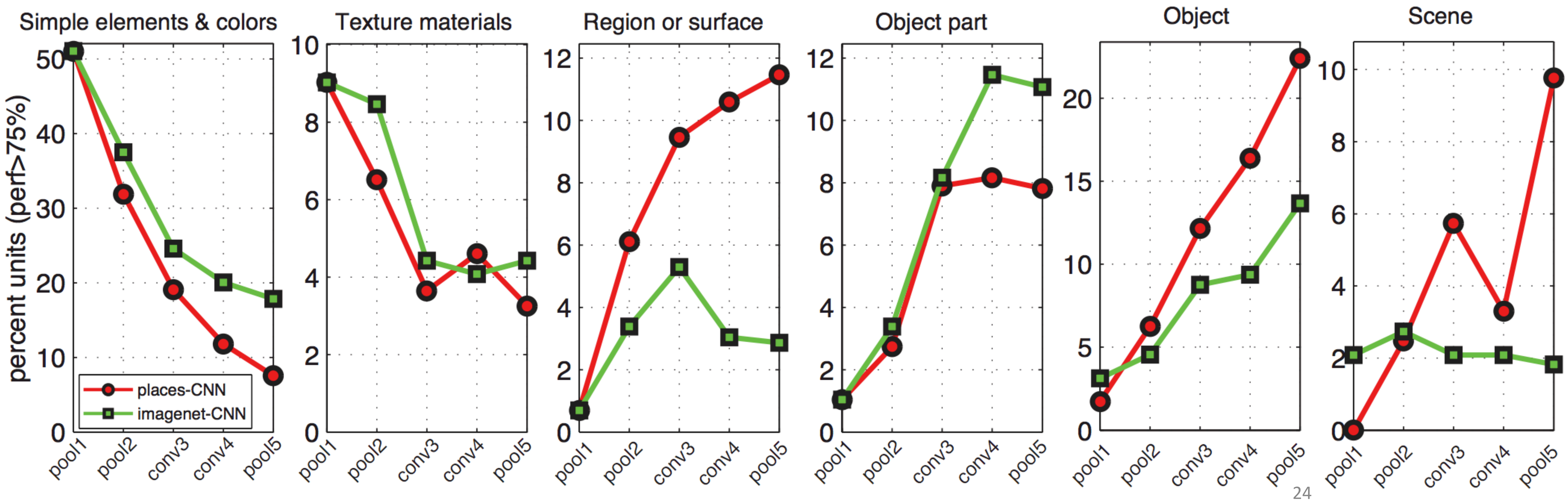
[Zhou et al., ICLR 2015]

pool 5

Source: Isola, Freeman, Torralba

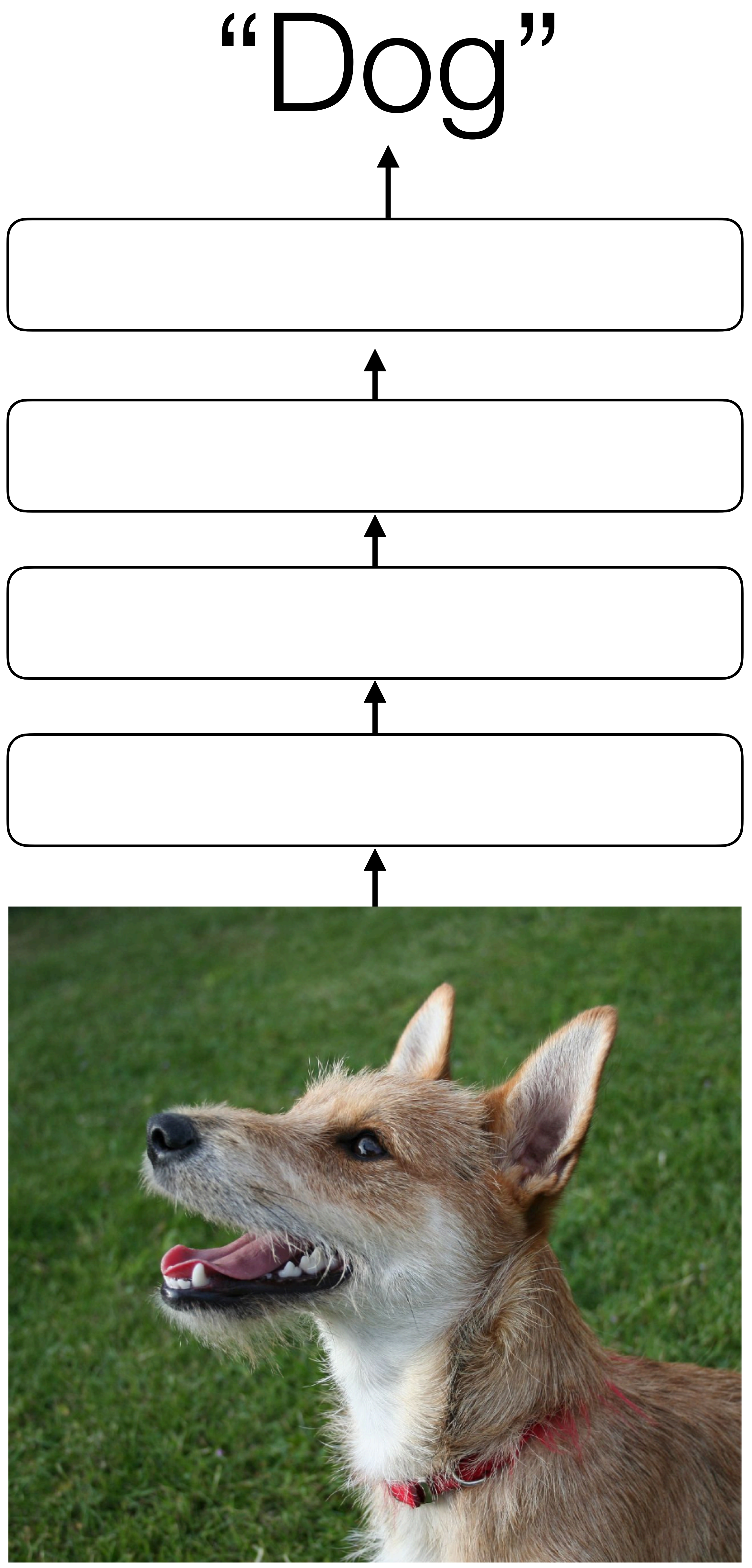
Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]



Source: Isola, Freeman, Torralba

Linear probe

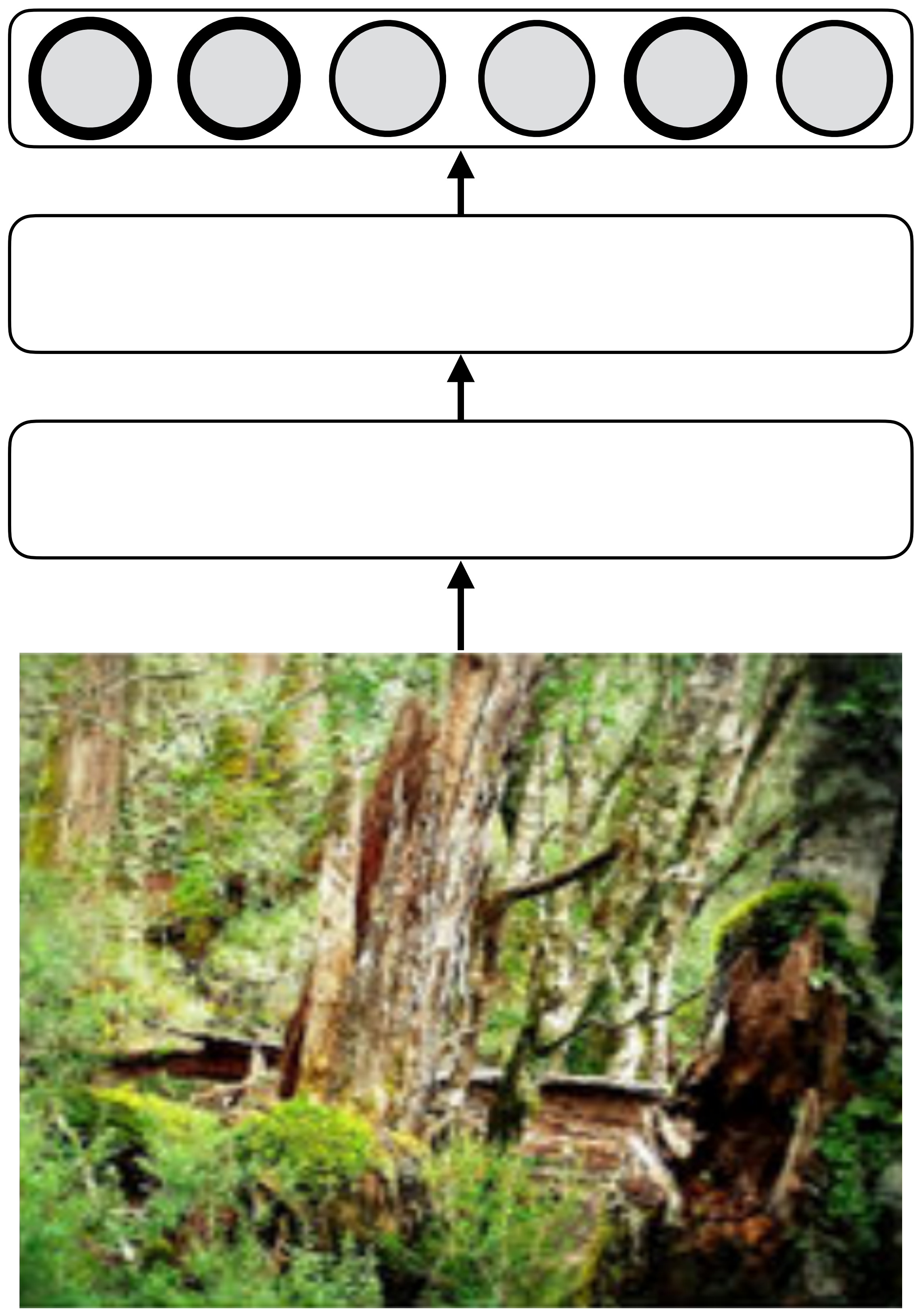


Object recognition net

Linear probe

Feature representation

$\mathbf{z}$



$(\mathbf{W}, \mathbf{b})$

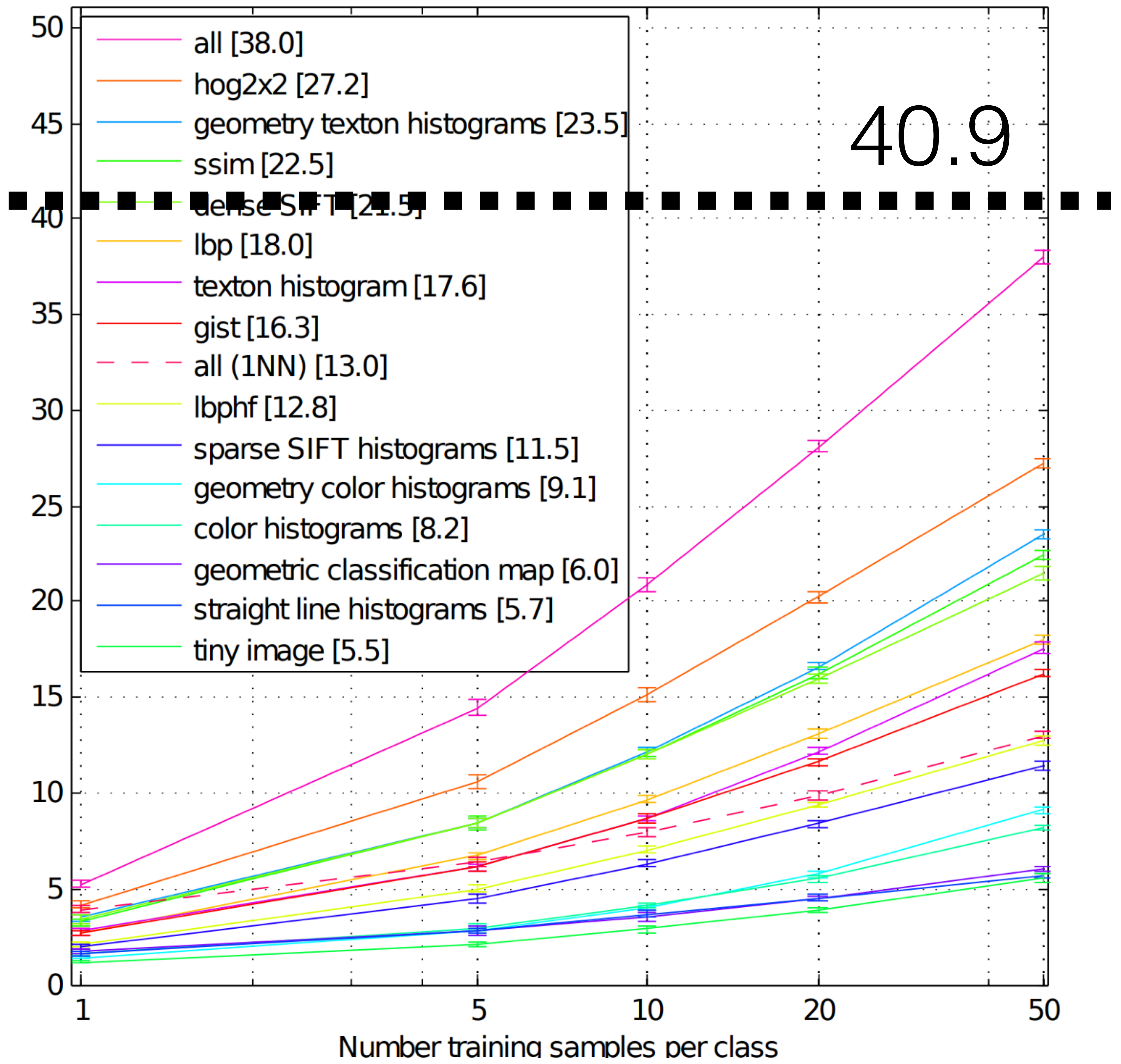
Linear Classifier → y "Rainforest"

Logistic regression:

$$y = \sigma(\mathbf{W}\mathbf{z} + \mathbf{b})$$

Transferring CNN features

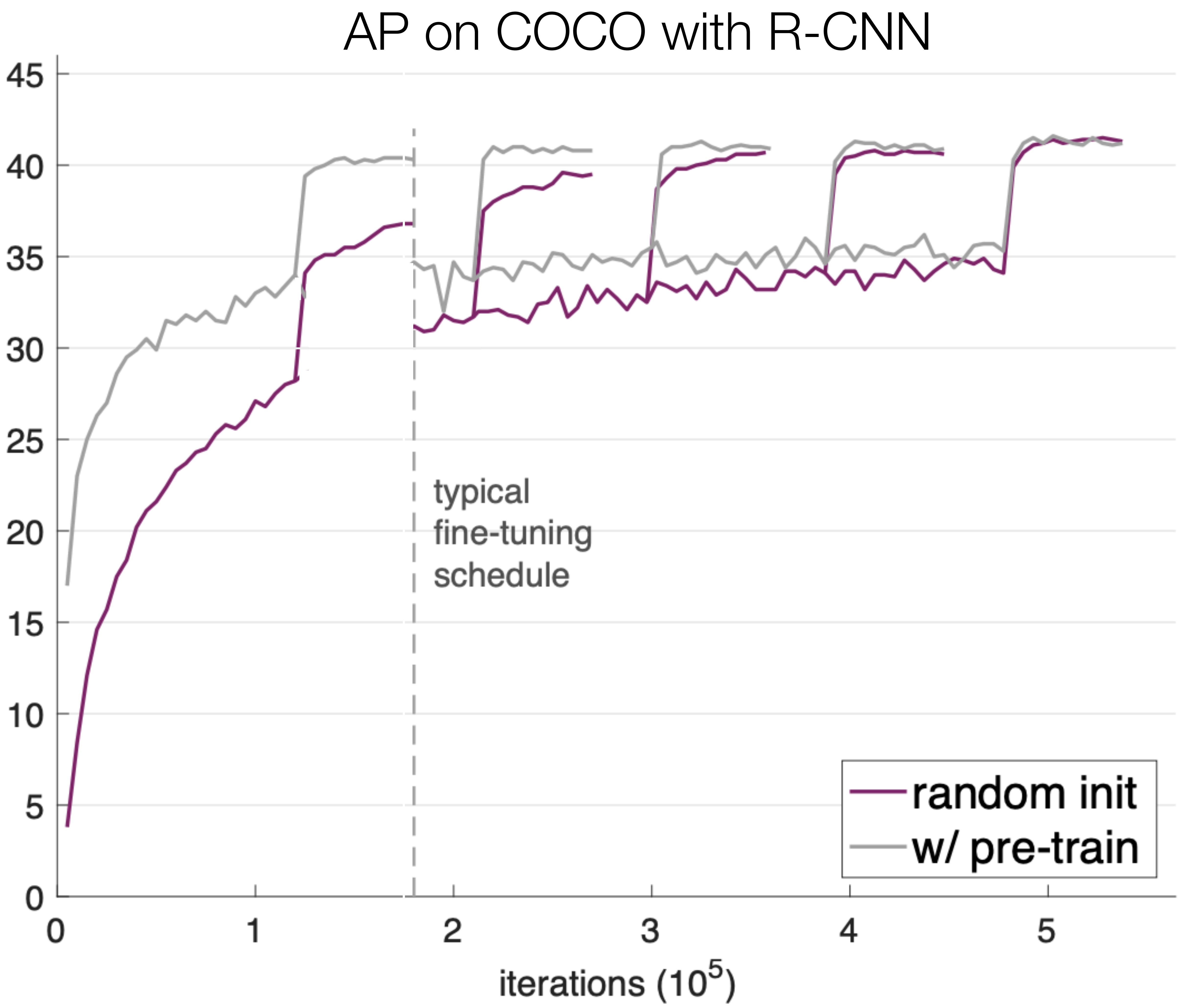
Hand-crafted features



[Xiao et al., CVPR 2010]

CNN features pretrained on ImageNet
+ linear classifier [Donahue et al. 2013]

Case study: fine-tuning for object detection



Observations:

- ImageNet pretraining speeds up object detection training by 5x
- No change in accuracy for this dataset — just training speed, perhaps because it is so large.
- Big performance gains for small/medium datasets (e.g. 1K examples per class)

[He et al. 2018]

Learning from examples

Training data

$$\{x_1, y_1\}$$
$$\{x_2, y_2\} \quad \rightarrow \quad \boxed{\text{Learner}} \quad \rightarrow \quad f : X \rightarrow Y$$
$$\{x_3, y_3\}$$
$$\dots$$

Source: Isola, Freeman, Torralba

Representation Learning

Data

$\{x_1\}$

$\{x_2\}$ $\rightarrow$ $\boxed{\text{Learner}}$ $\rightarrow$ Representations

$\{x_3\}$

$\dots$

Source: Isola, Freeman, Torralba

How do we learn good representations?

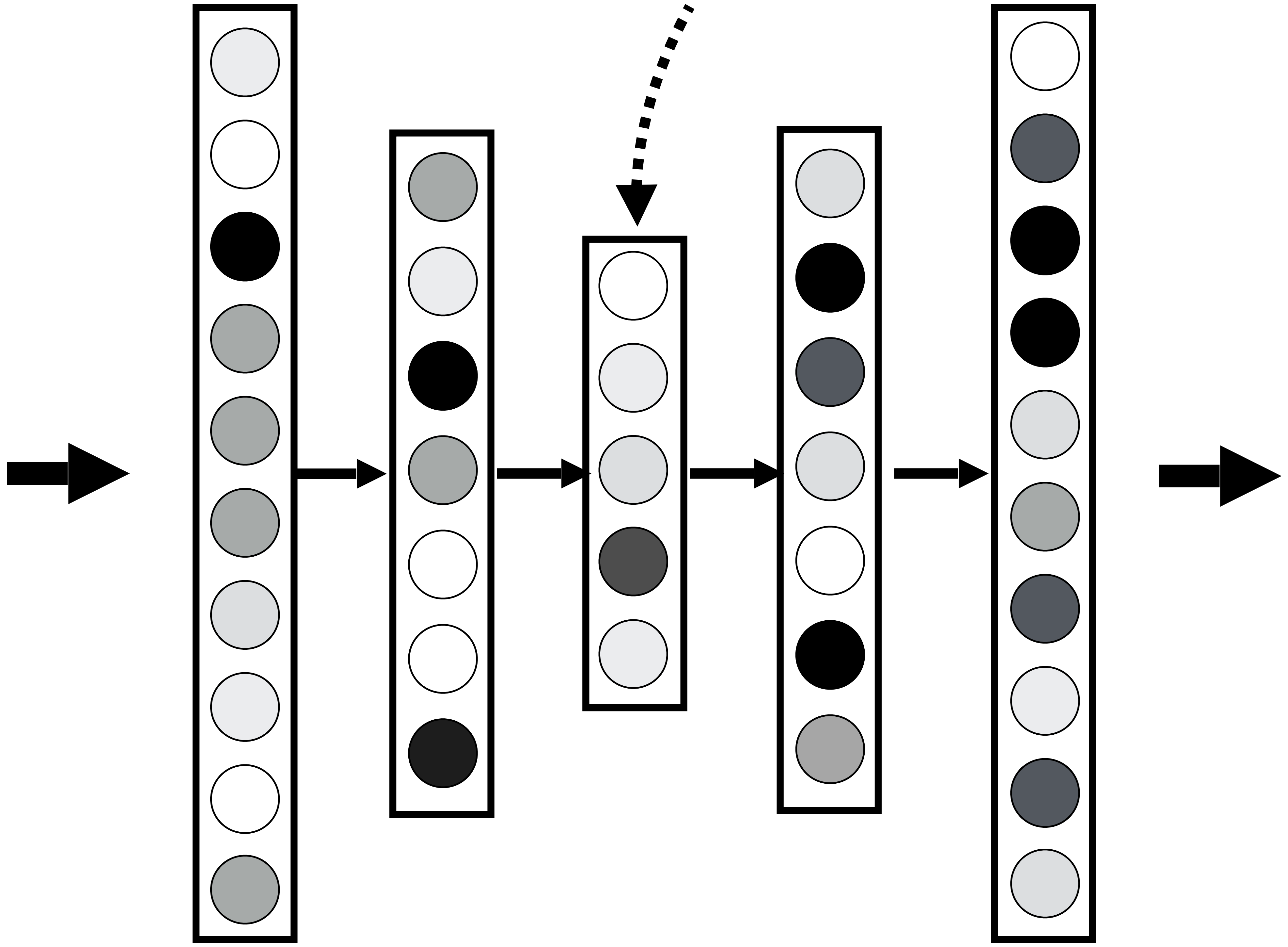
Self-supervised learning methods

Recall: autoencoder

compressed image code
(vector **z**)

$\mathbf{x}$

Image



$\hat{\mathbf{x}}$

Reconstructed image

[e.g., Hinton & Salakhutdinov, Science 2006]

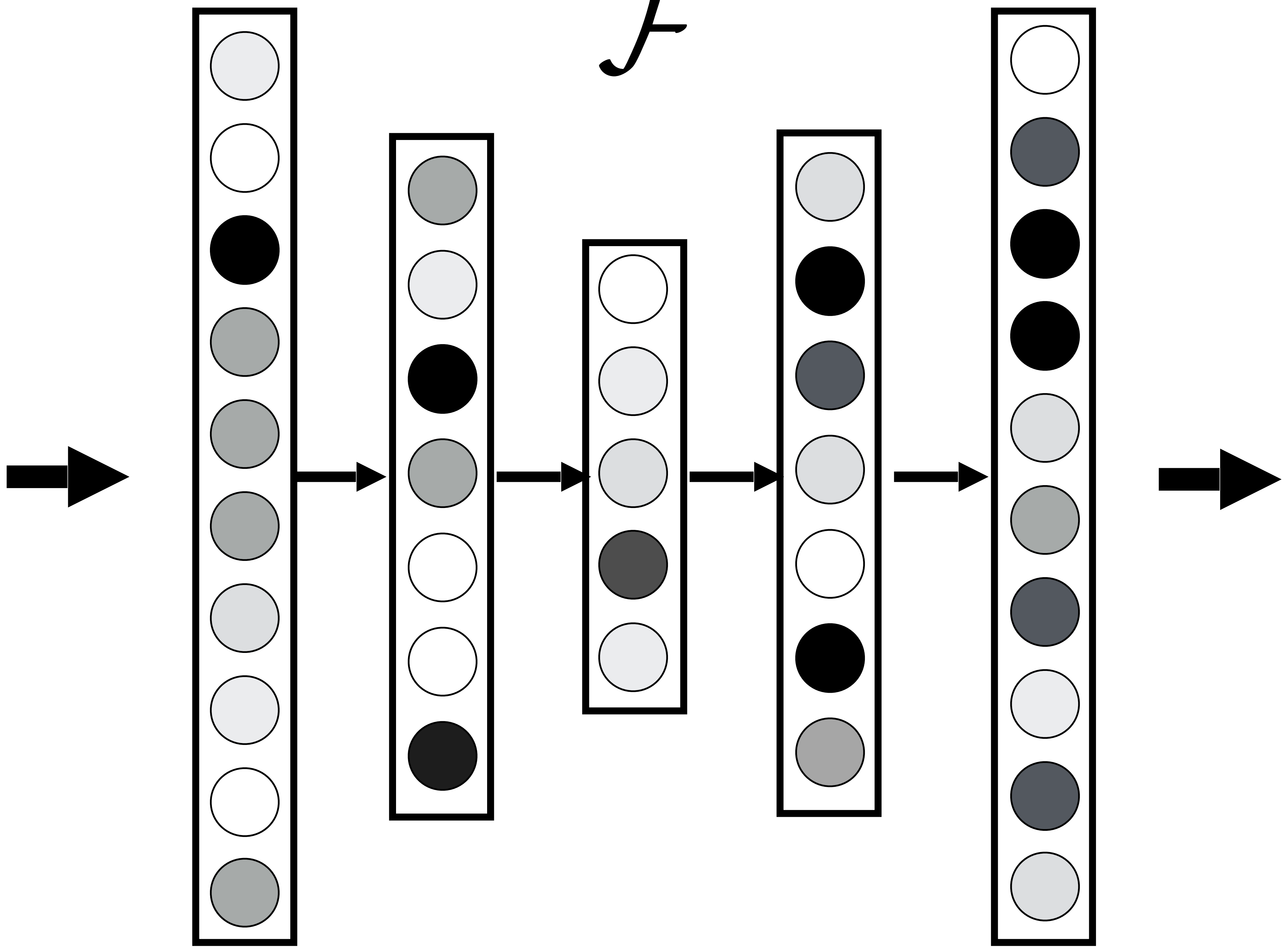
Source: Isola, Freeman, Torralba

Autoencoder

$\mathbf{X}$

Image

$\mathcal{F}$

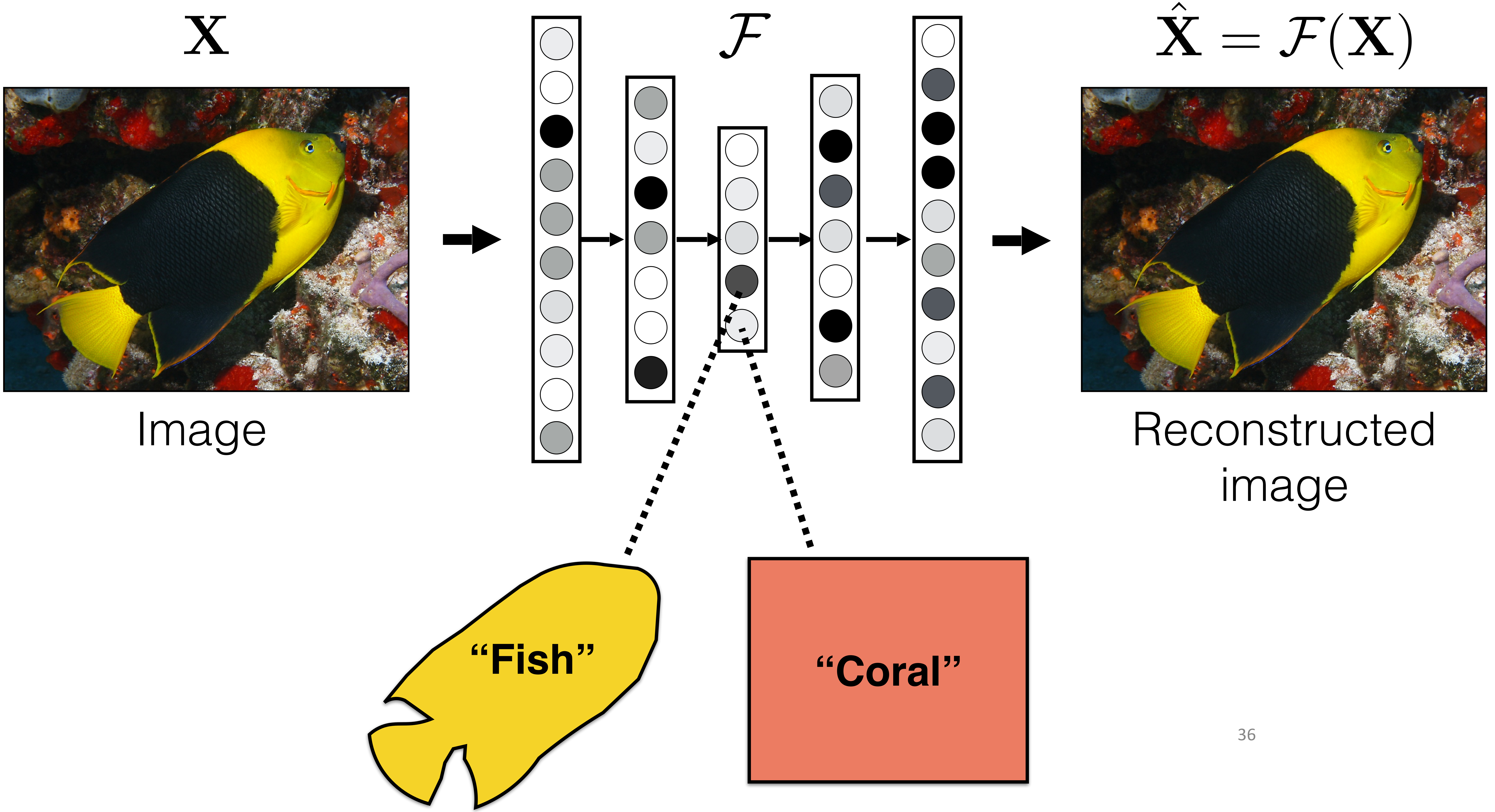


$\hat{\mathbf{X}} = \mathcal{F}(\mathbf{X})$

Reconstructed image

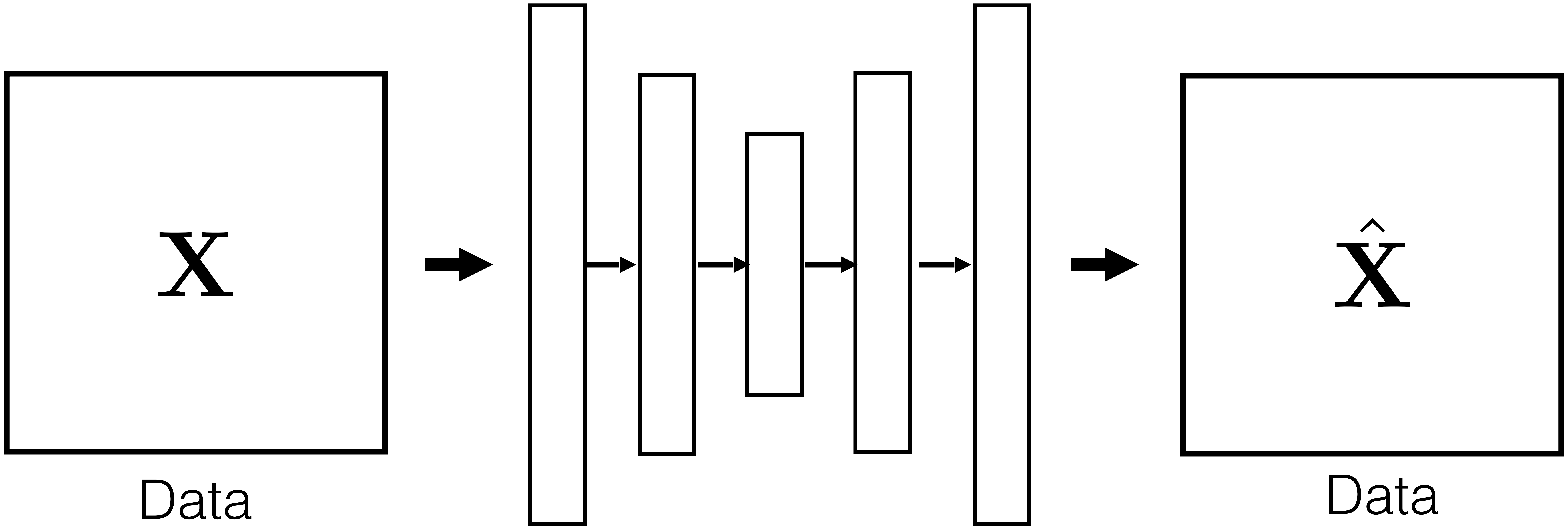
$$\arg\min_{\mathcal{F}} \mathbb{E}_{\mathbf{X}}[||\mathcal{F}(\mathbf{X}) - \mathbf{X}||]$$

Source: Isola, Freeman, Torralba



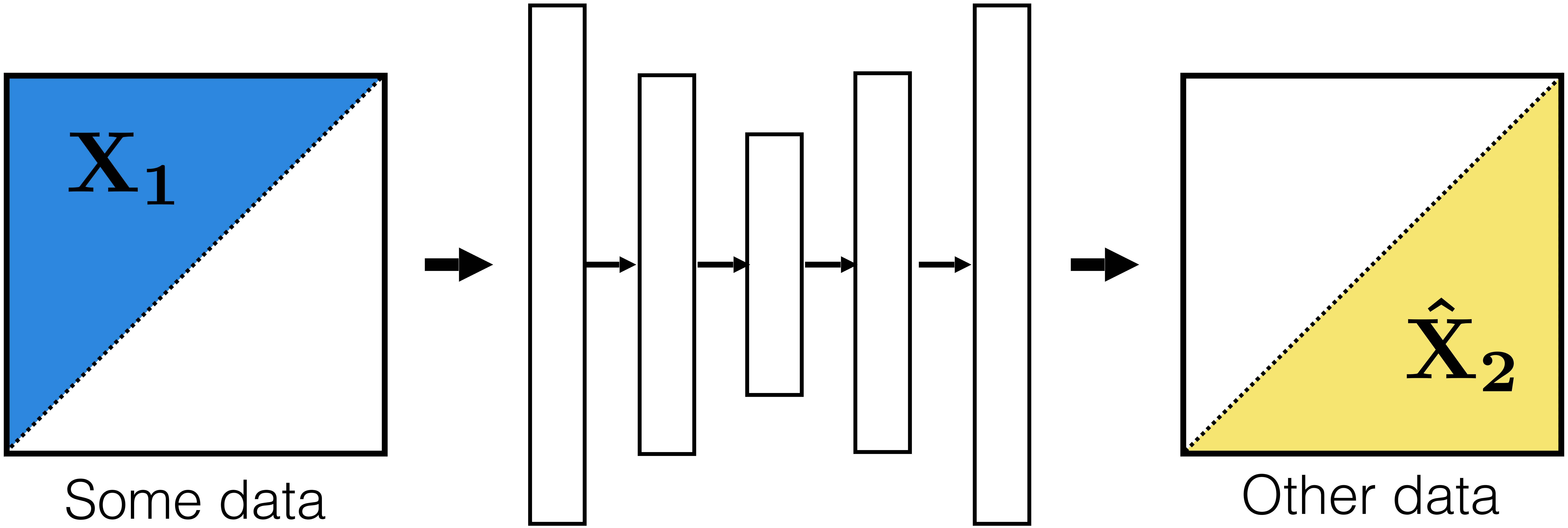
Source: Isola, Freeman, Torralba

Data compression



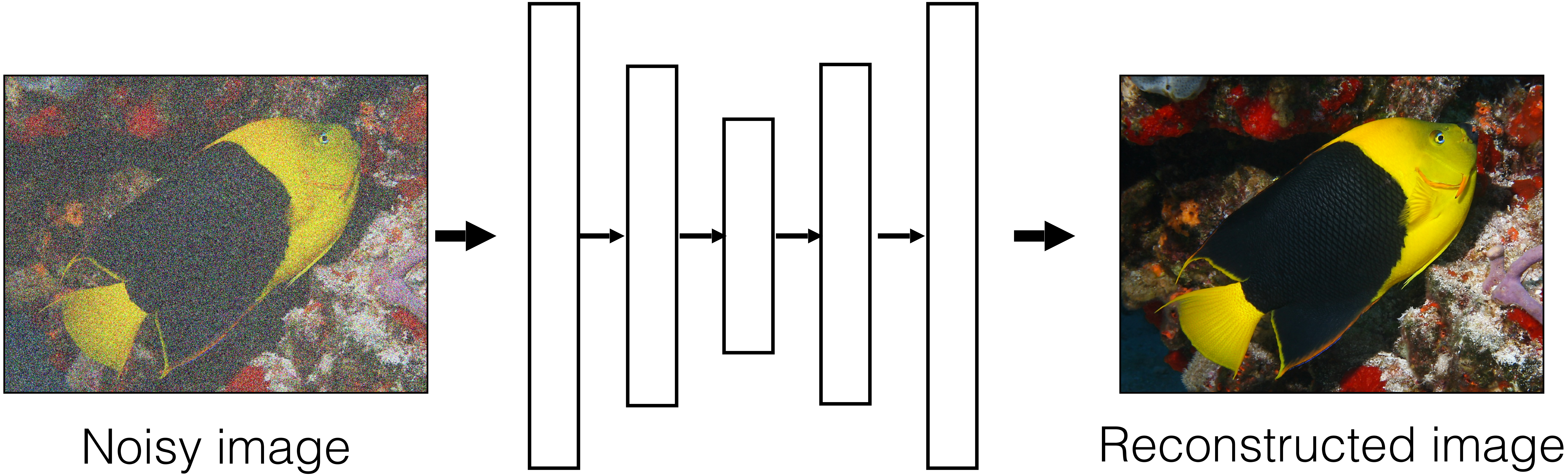
Source: Isola, Freeman, Torralba

Data prediction



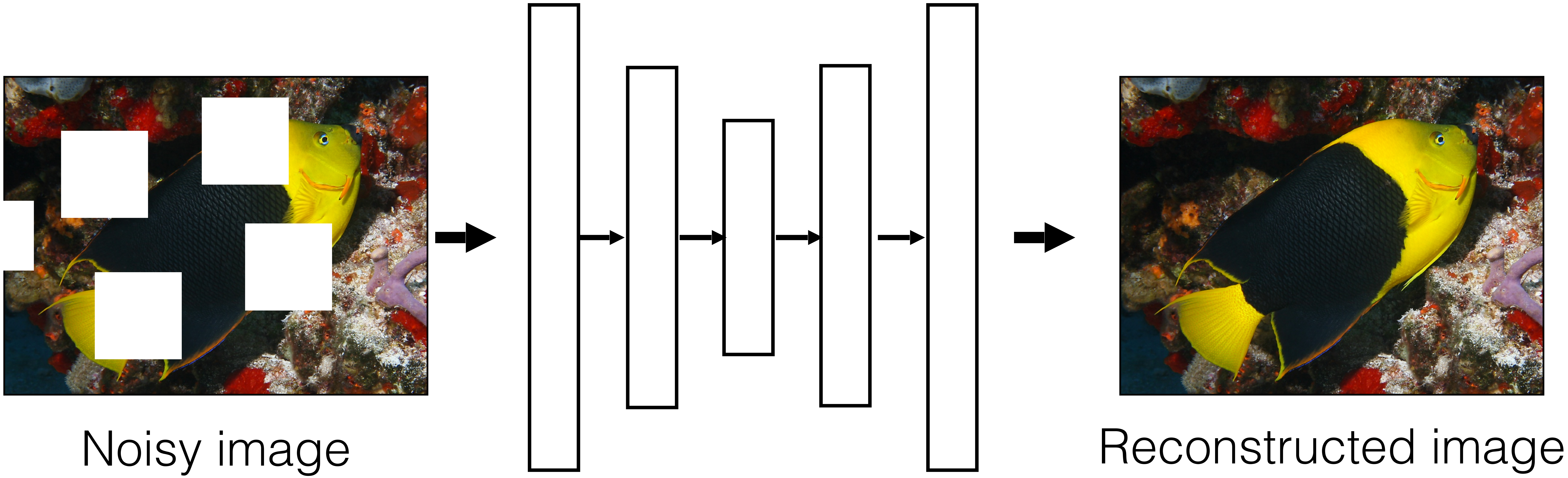
Source: Isola, Freeman, Torralba

Denoising autoencoder



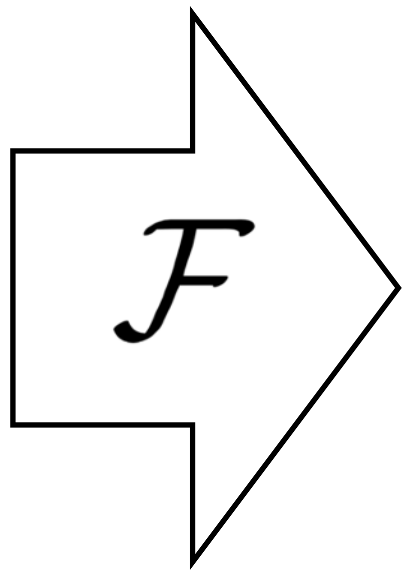
[Vincent et al., 2008]

Denoising autoencoder



Other types of “noise”.

[Vincent et al., 2008, Pathak et al. 2015, He, 2020]

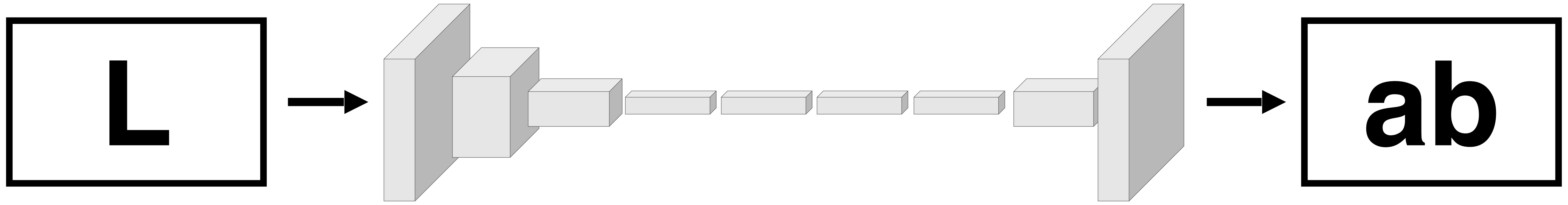


Grayscale image: L channel

$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Color information: ab channels

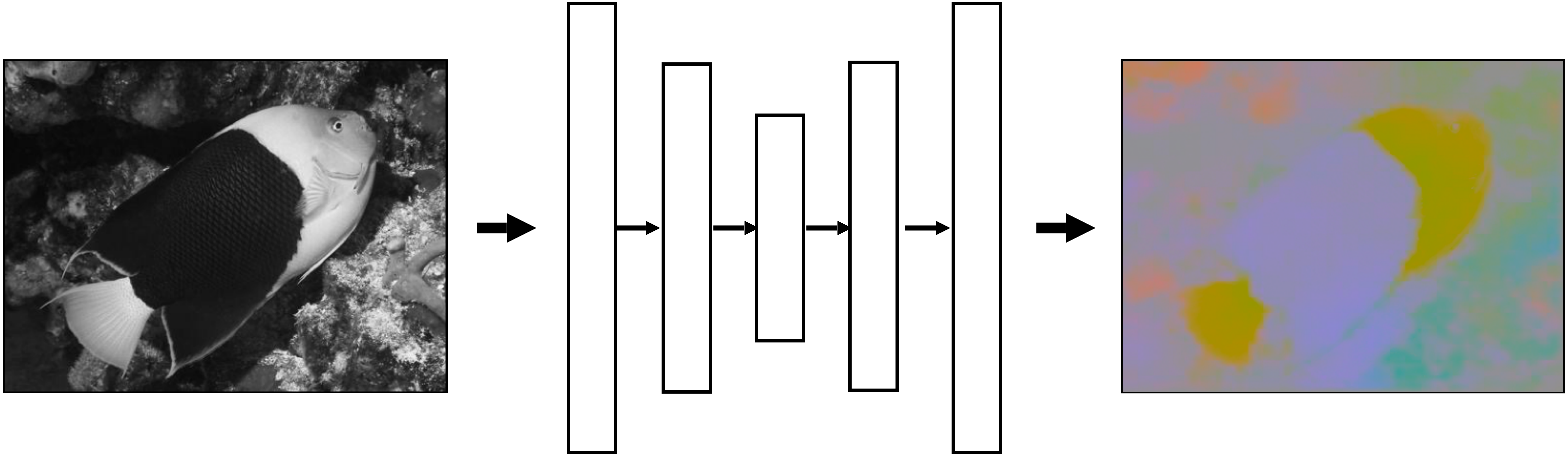
$\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$



[Zhang, Isola, Efros, ECCV 2016]

Source: Isola, Freeman, Torralba

Visualizing units



[Zeiler & Fergus, ECCV 2014]

[Zhou et al., ICLR 2015]

Source: Isola, Freeman, Torralba

Source: Isola, Torralba, Freeman

[“Colorful image colorization”, Zhang et al., ECCV 2016]

[“Colorful image colorization”, Zhang et al., ECCV 2016]

Source: Isola, Torralba, Freeman

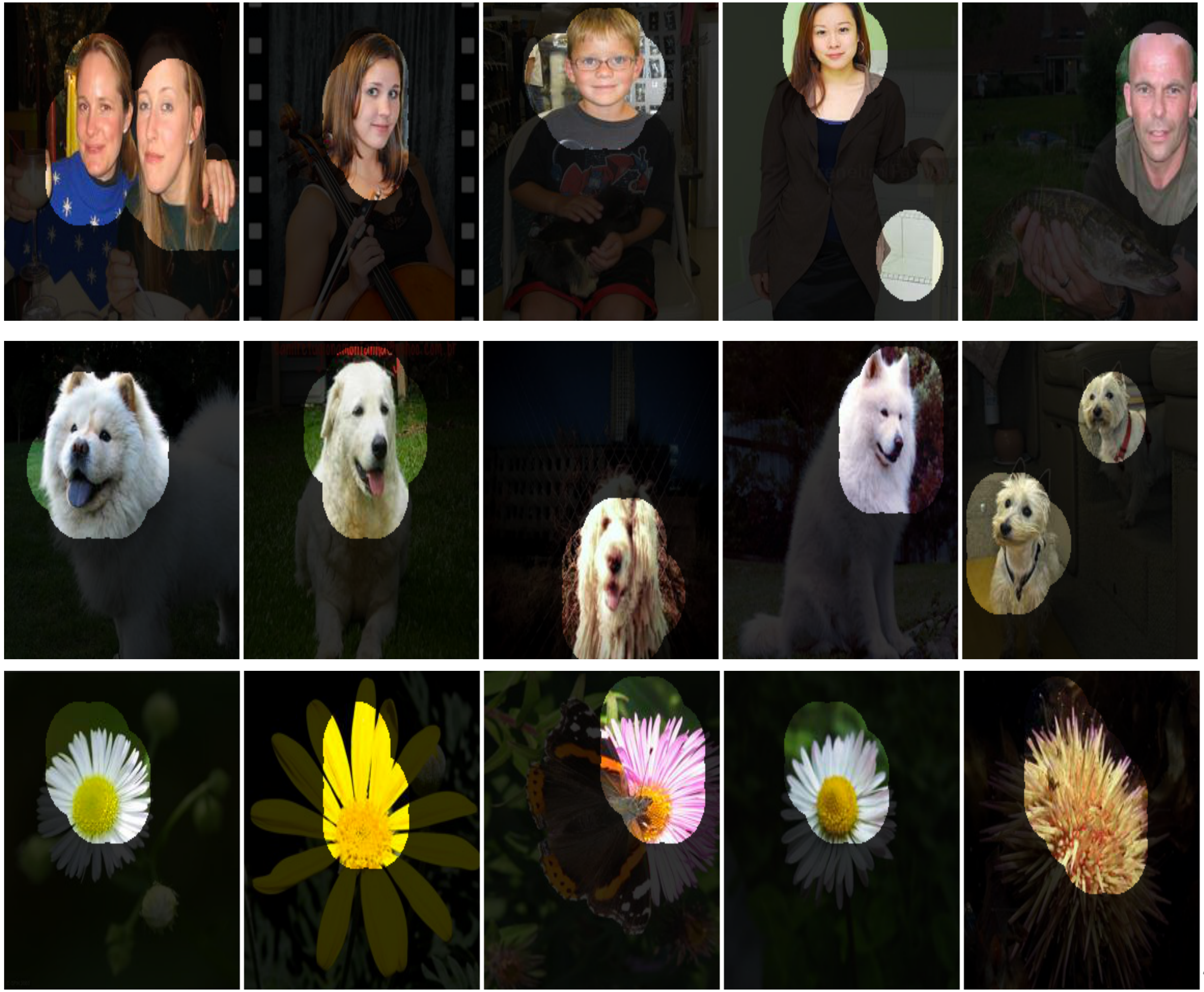
["Colorful image colorization", Zhang et al., ECCV 2016]

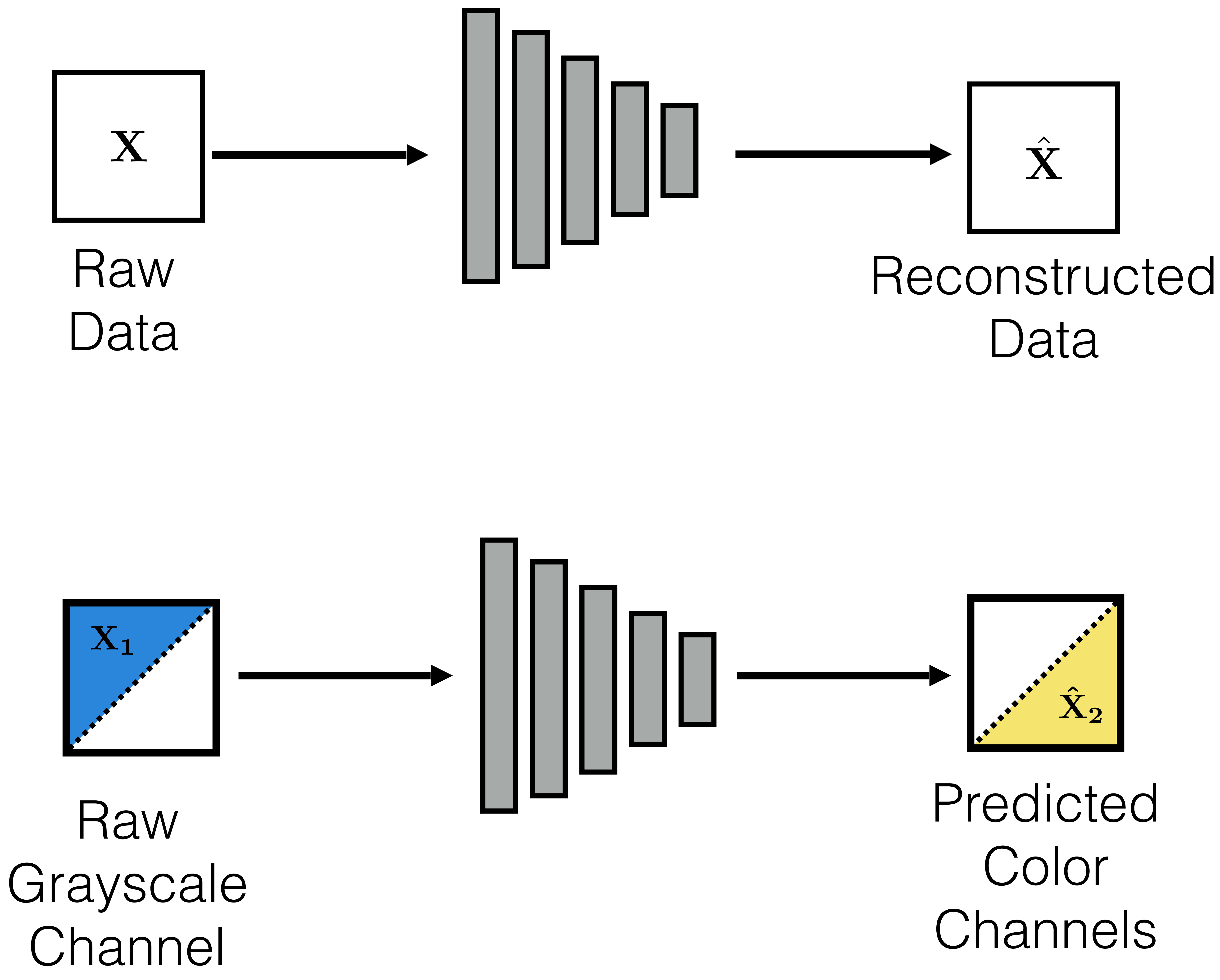
Stimuli that drive selected neurons (conv5 layer)

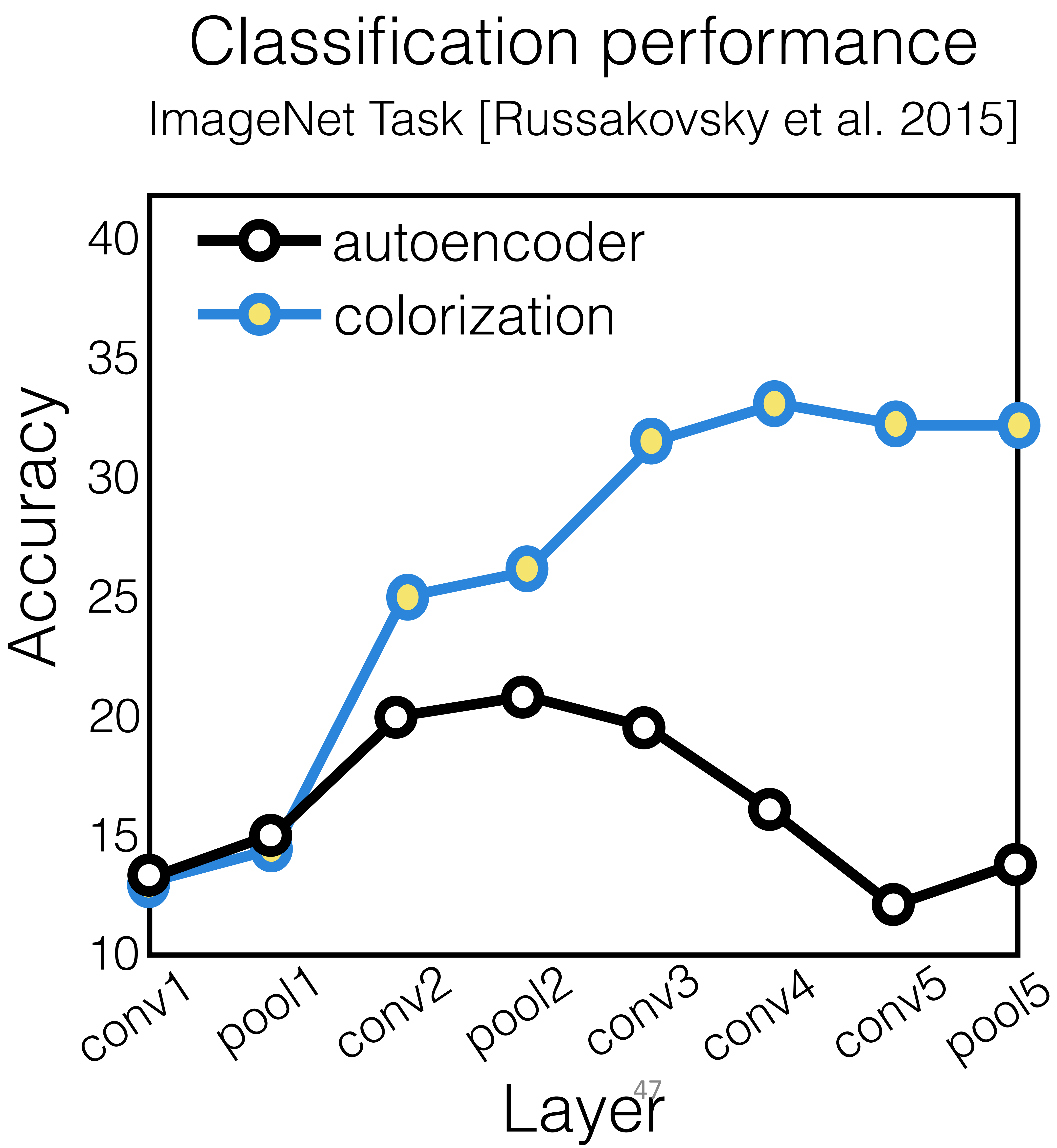
faces

dog faces

flowers





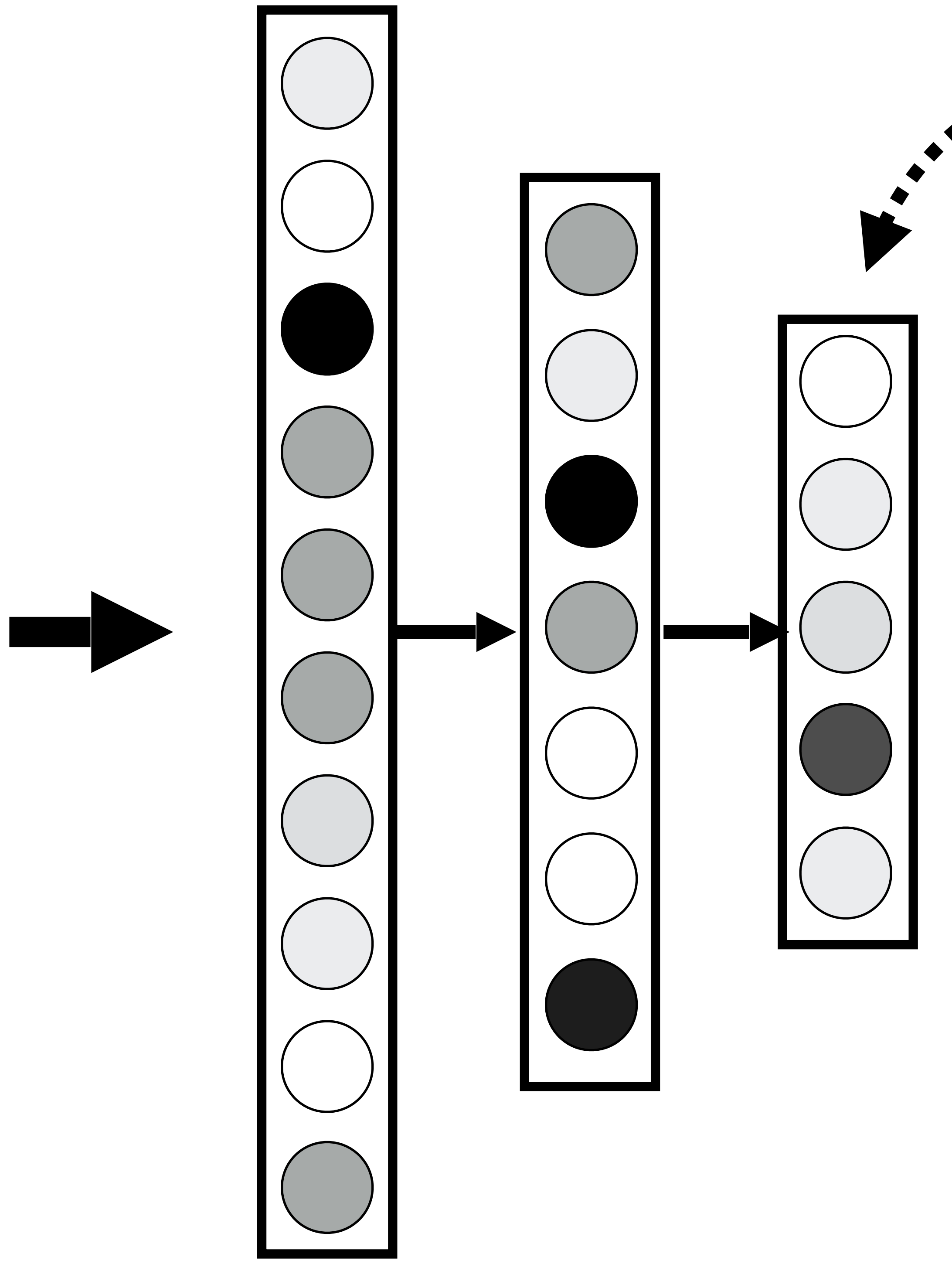


Source: Isola, Freeman, Torralba

Image representations

$\mathbf{x}$

Image



layer 3 representation of image

layer 1 representation of image

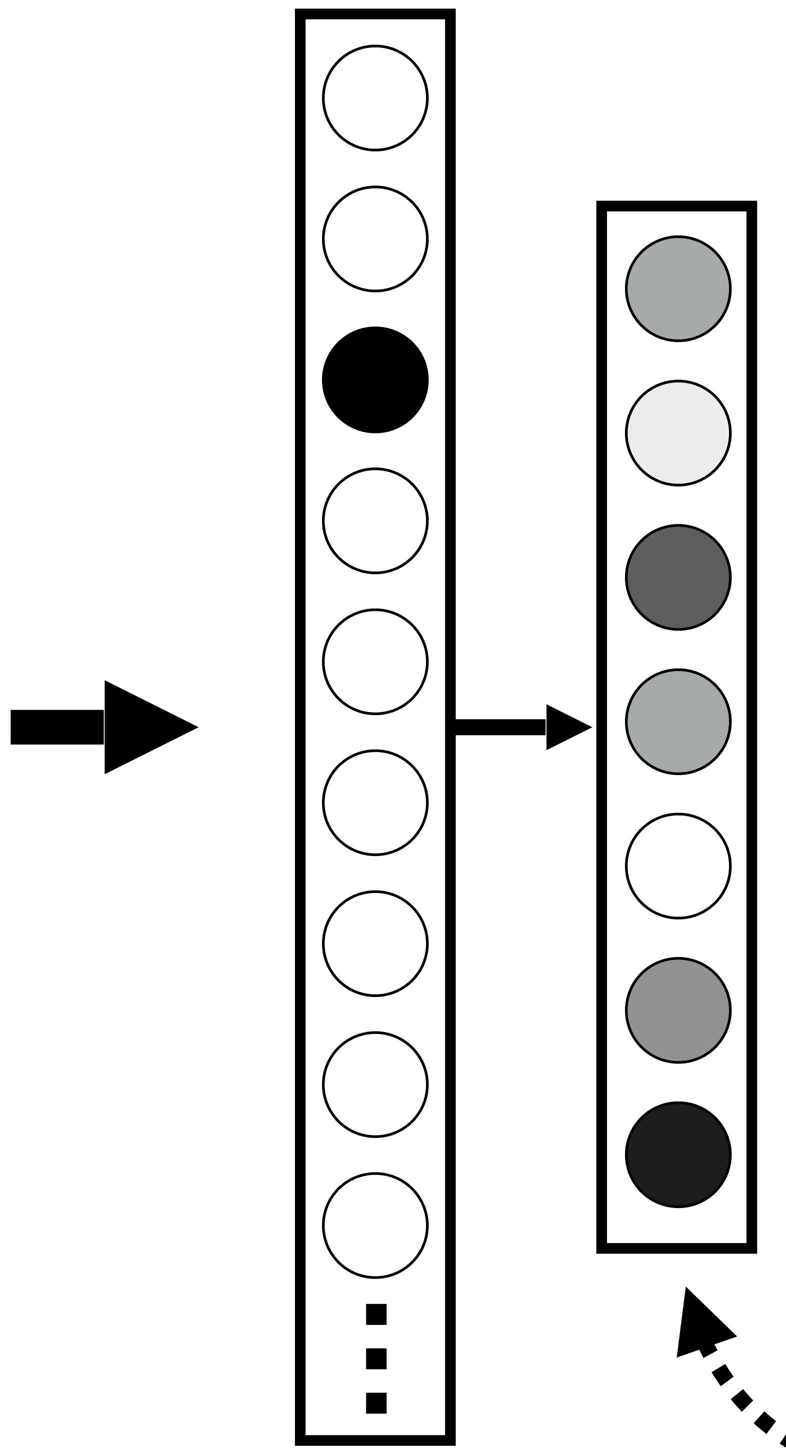
Represent image as a vector of neural activations

(perhaps representing a vector of detected texture patterns or object parts)

Source: Isola, Freeman, Torralba

Example from language: word2vec

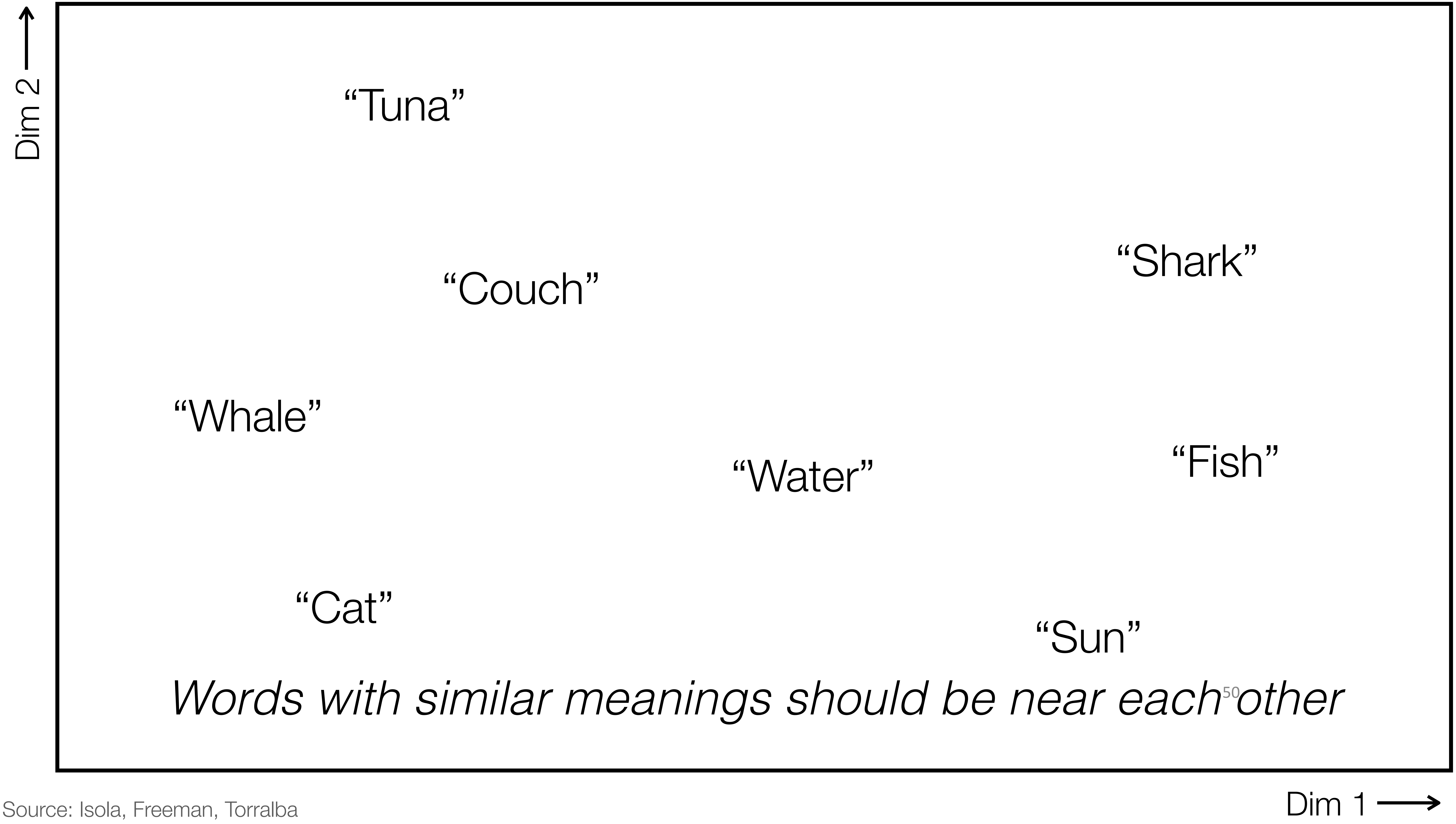
“Elephant”



dense vector representation of word,
called a **word embedding**.

one-hot vector representation of word

Source: Isola, Freeman, Torralba



Words with similar meanings should be near each other

Source: Isola, Freeman, Torralba

word2vec

Words with similar meanings should be near each other

Proxy: words that are used in the same context tend to have similar meanings

words with similar contexts should be near each other

Source: Isola, Freeman, Torralba

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'
…	…

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

word2vec

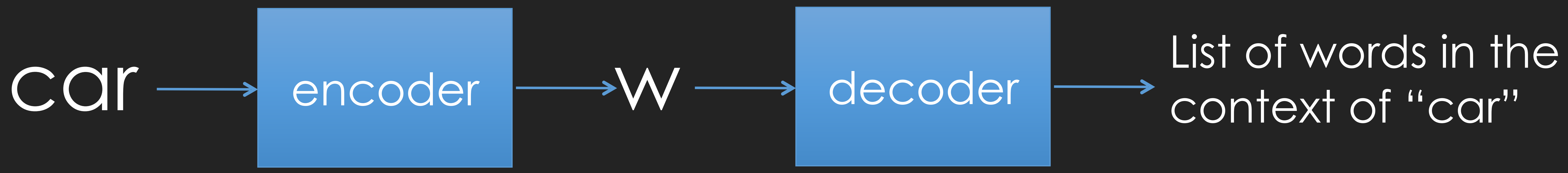
I parked the car in a nearby street. It is a red car with two doors, …

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

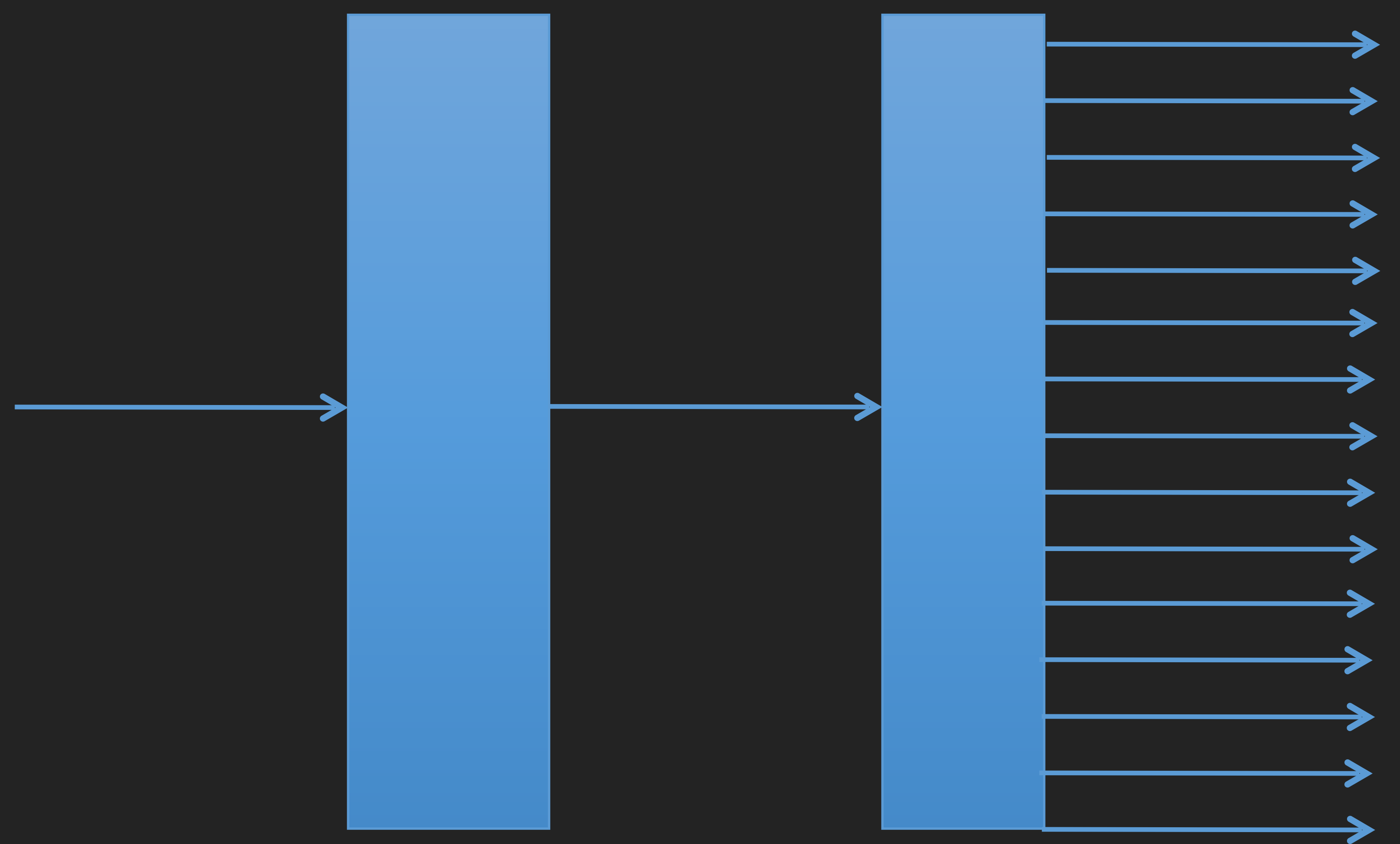
I parked the car in a nearby street. It is a red car with two doors, …



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

word2vec

Word
= ‘car’



Output prob. That
each word is in the
context of the input
word

Hidden layer

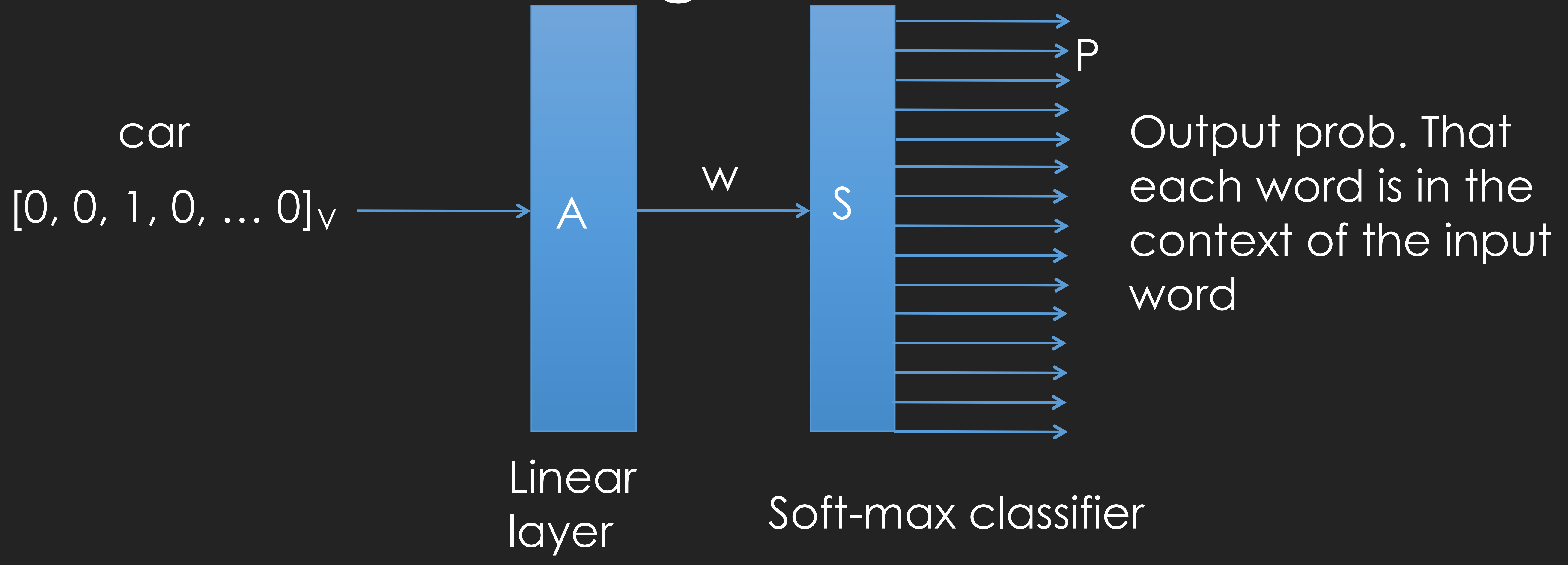
Soft-max classifier

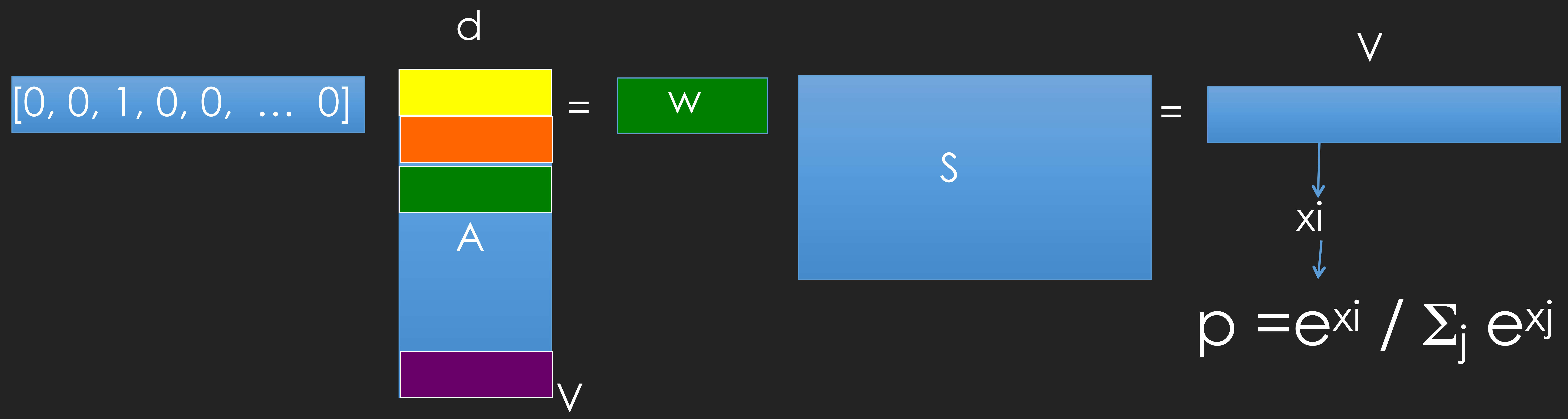
Encoder

Decoder

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

word2vec, training





$$p = e^{xi} / \Sigma_j e^{xj}$$

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

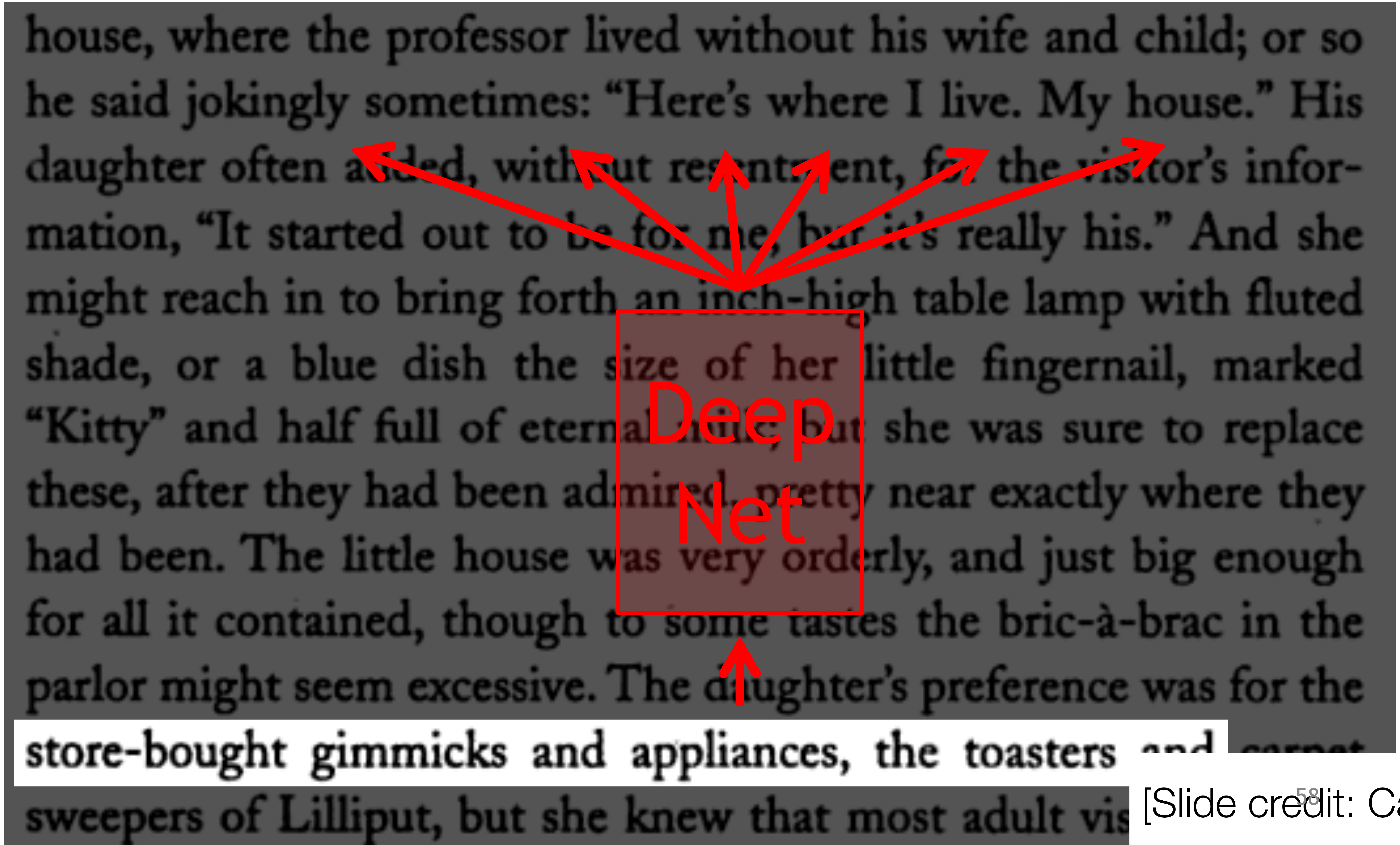
Algebraic operations with the vector representation of words

X = Vector("Paris") – vector("France") + vector("Italy")

Closest nearest neighbor to X is vector("Rome")

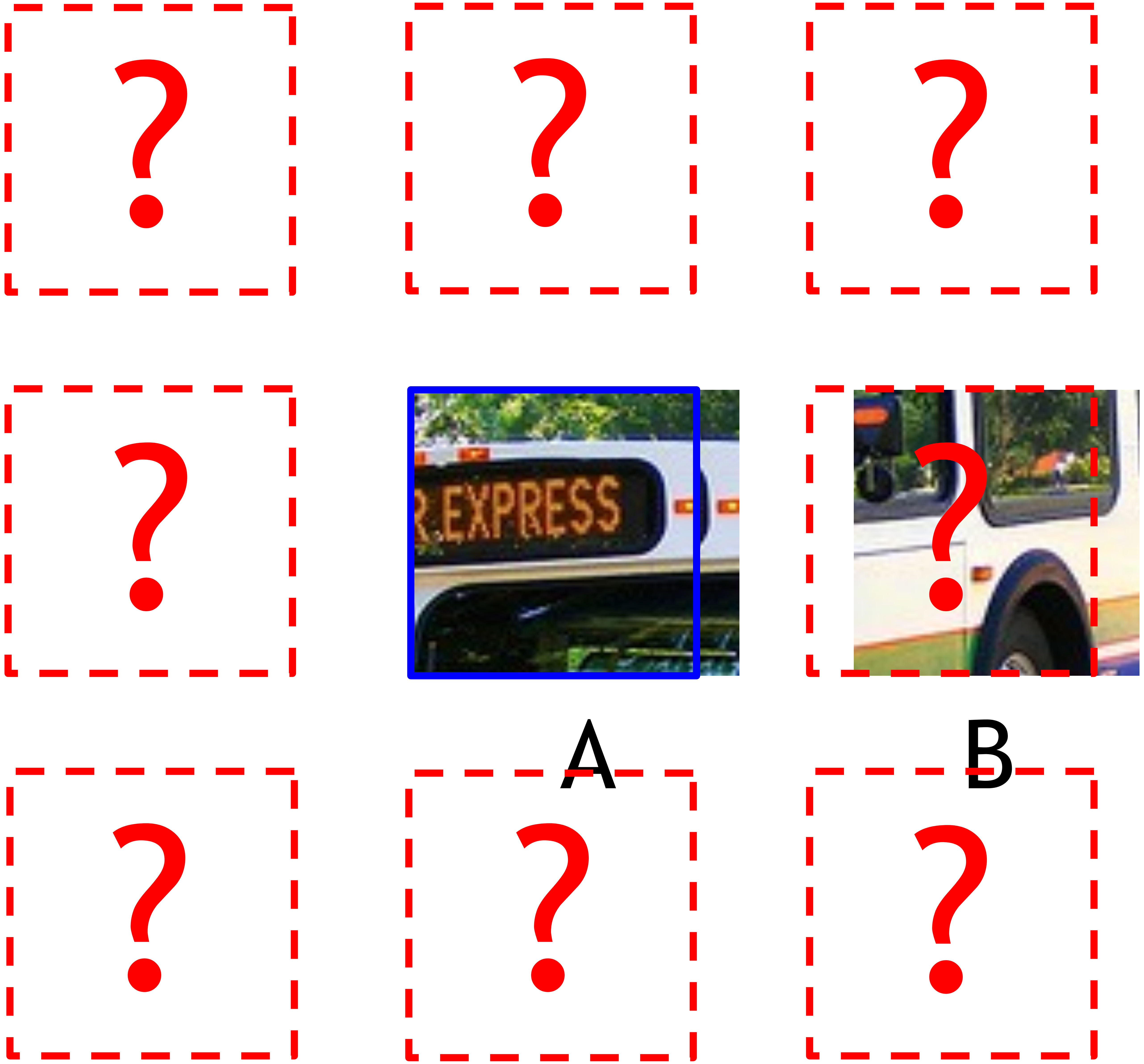
Context as Supervision

[Collobert & Weston 2008; Mikolov et al. 2013]



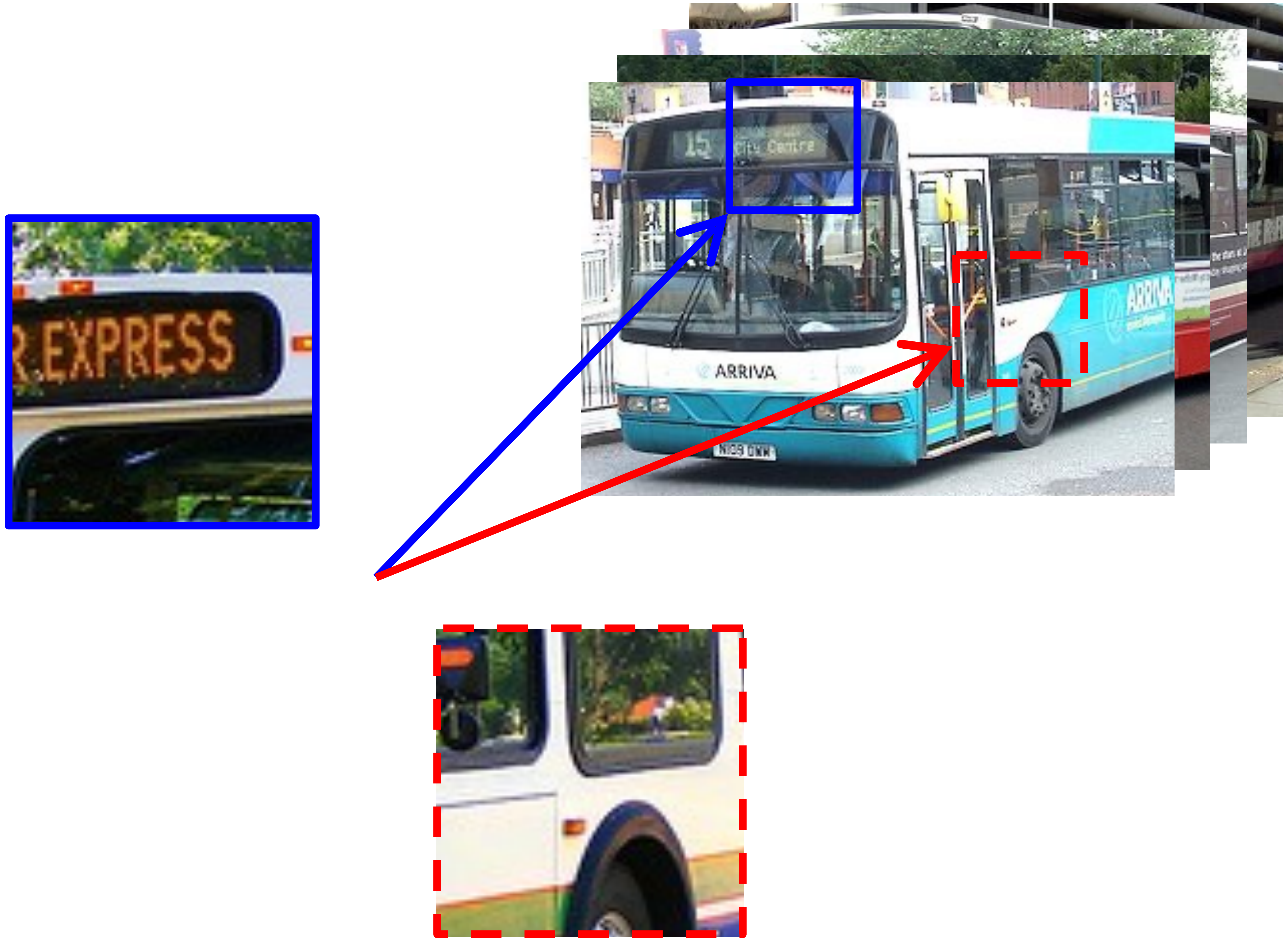
[Slide credit: Carl Doersch]

Context Prediction as Supervision



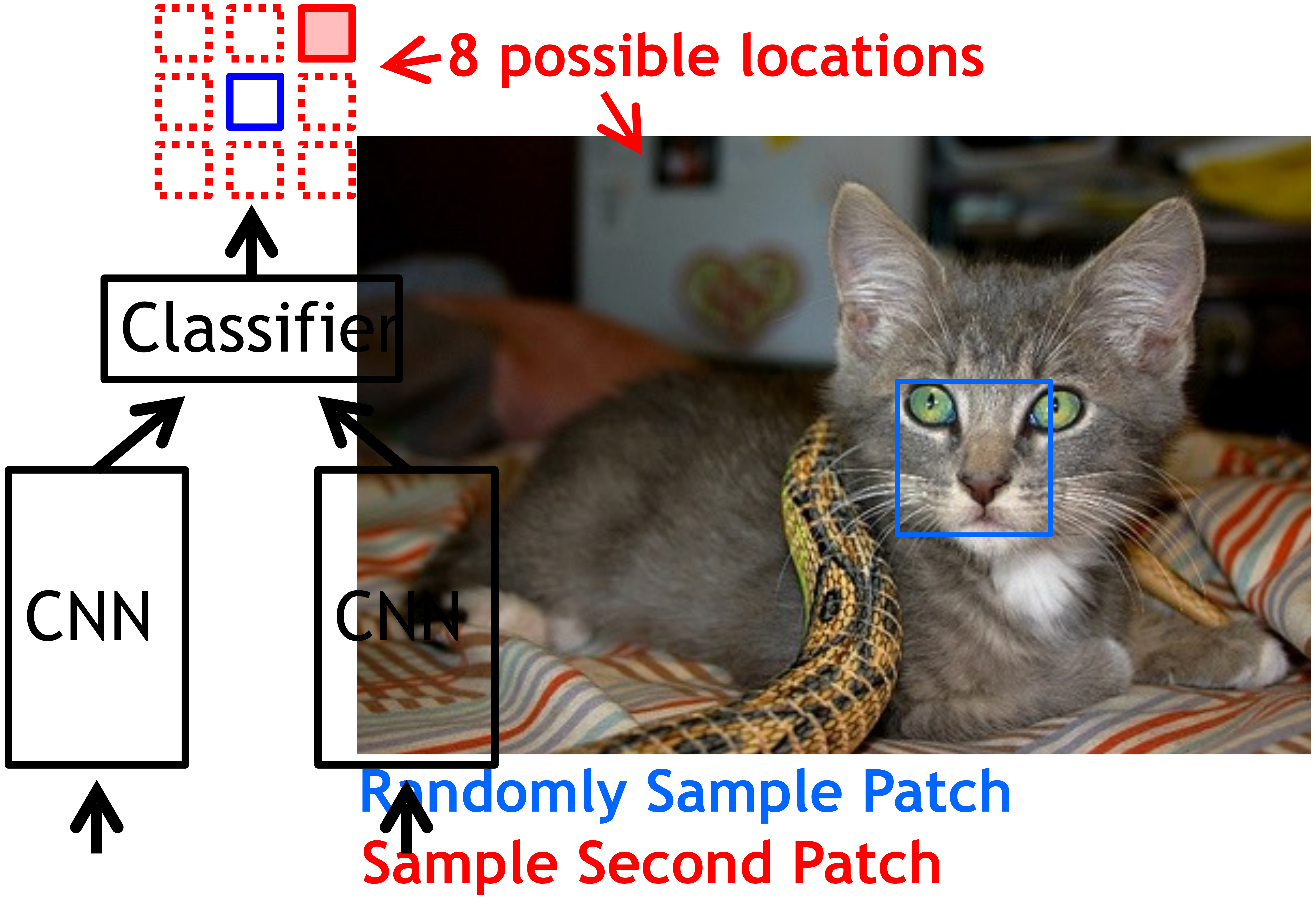
[Slide credit: Carl Doersch]

Semantics from a non-semantic task

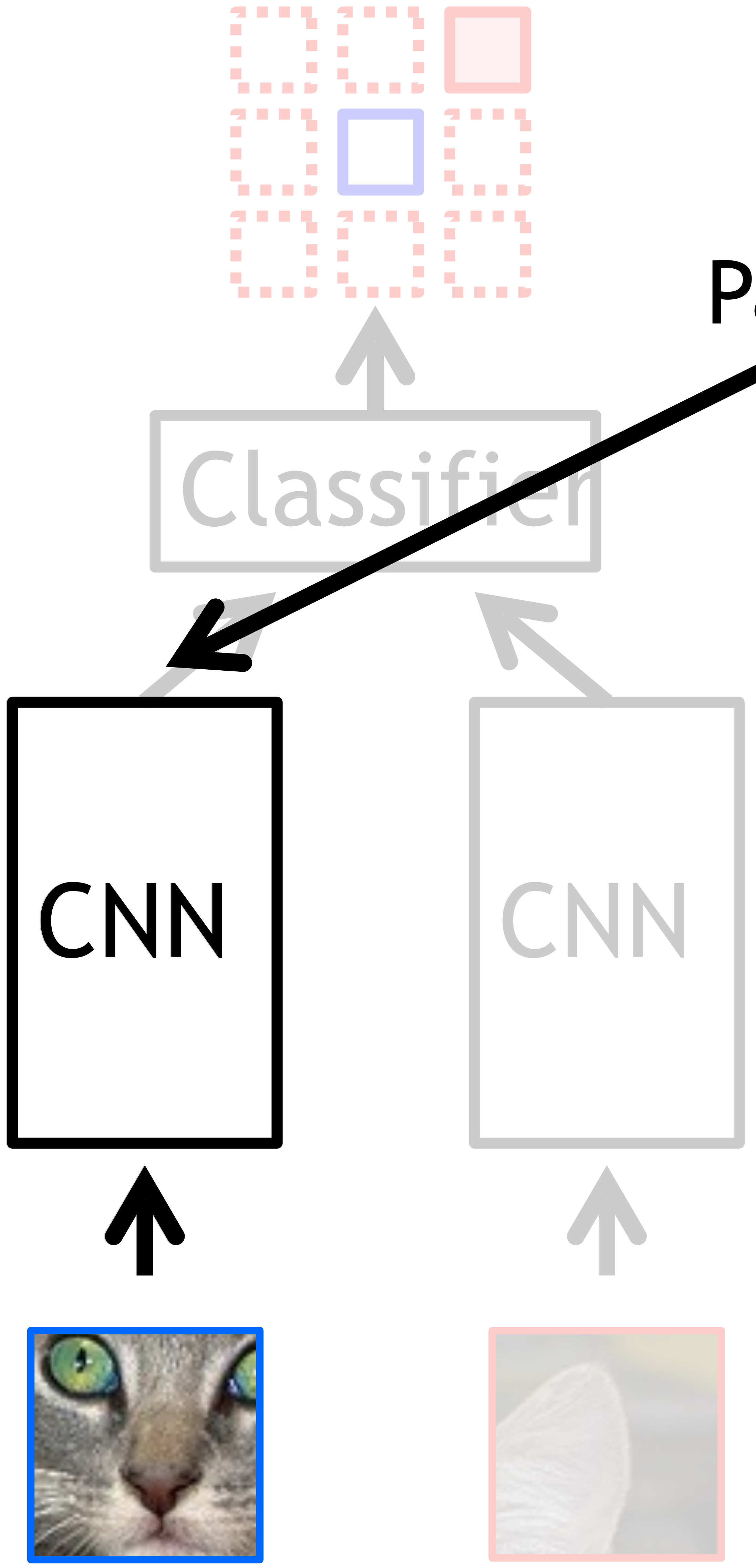


[Slide credit: Carl Doersch]

Relative Position Task



[Slide credit: Carl Doersch]



Patch Embedding (representation)

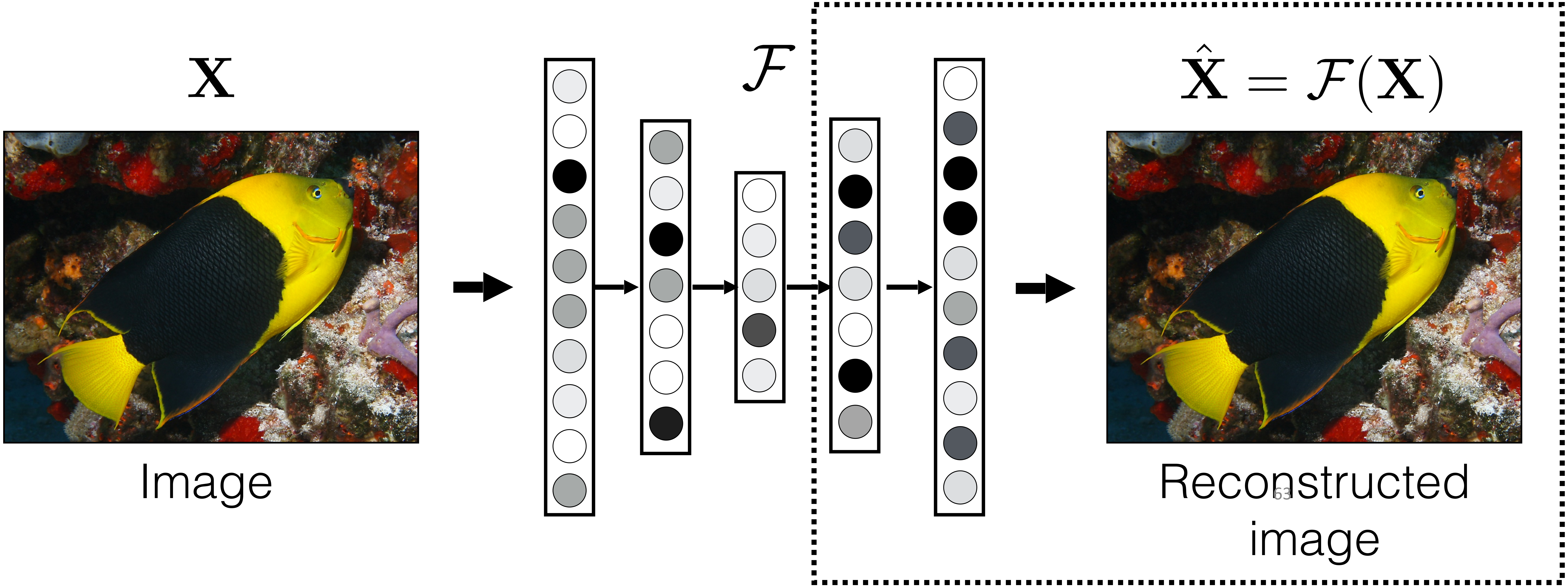
Input Nearest Neighbors

Note: connects ***across*** instances!

[Slide credit: Carl Doersch]

Revisiting autoencoders

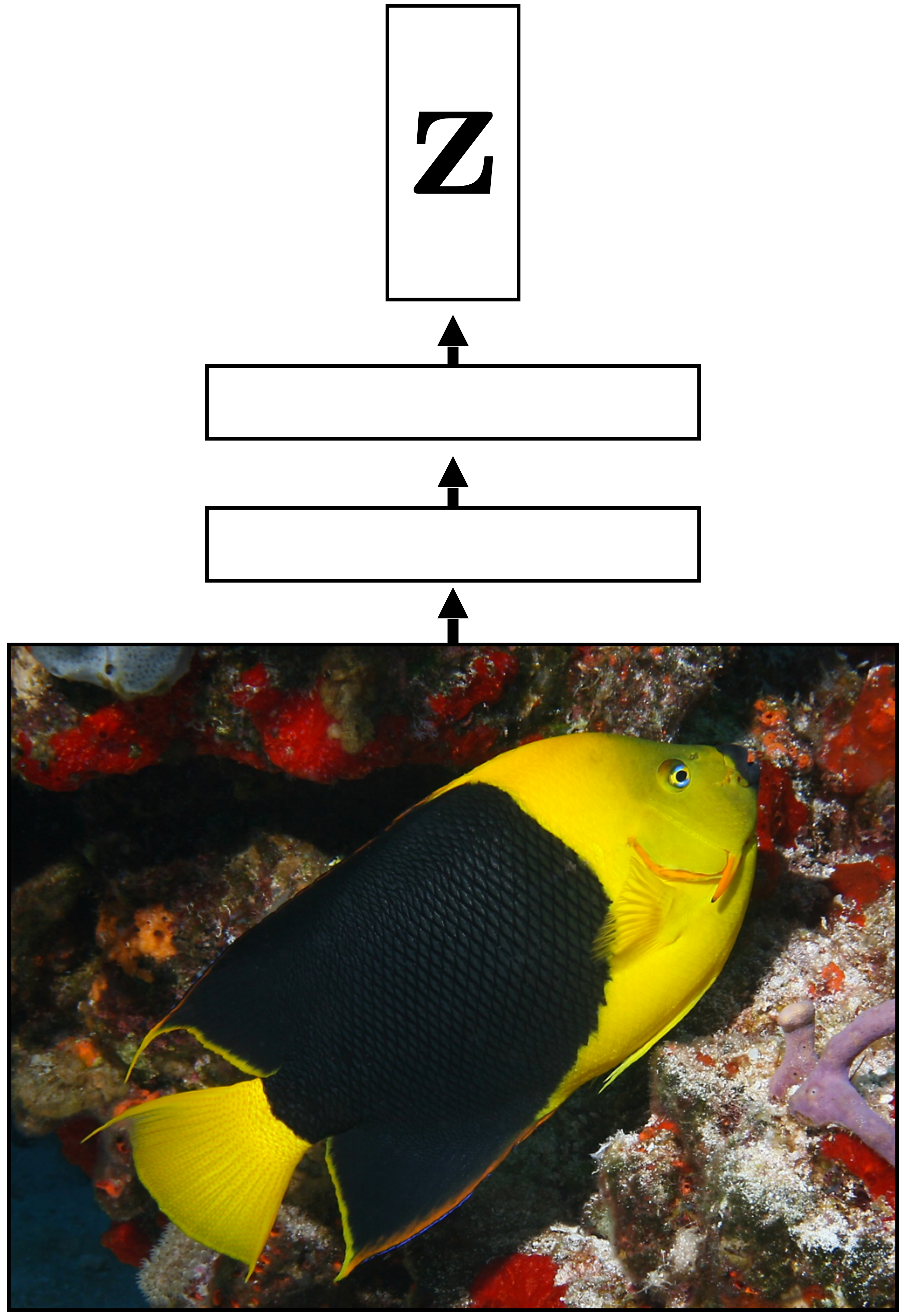
Is prediction necessary?

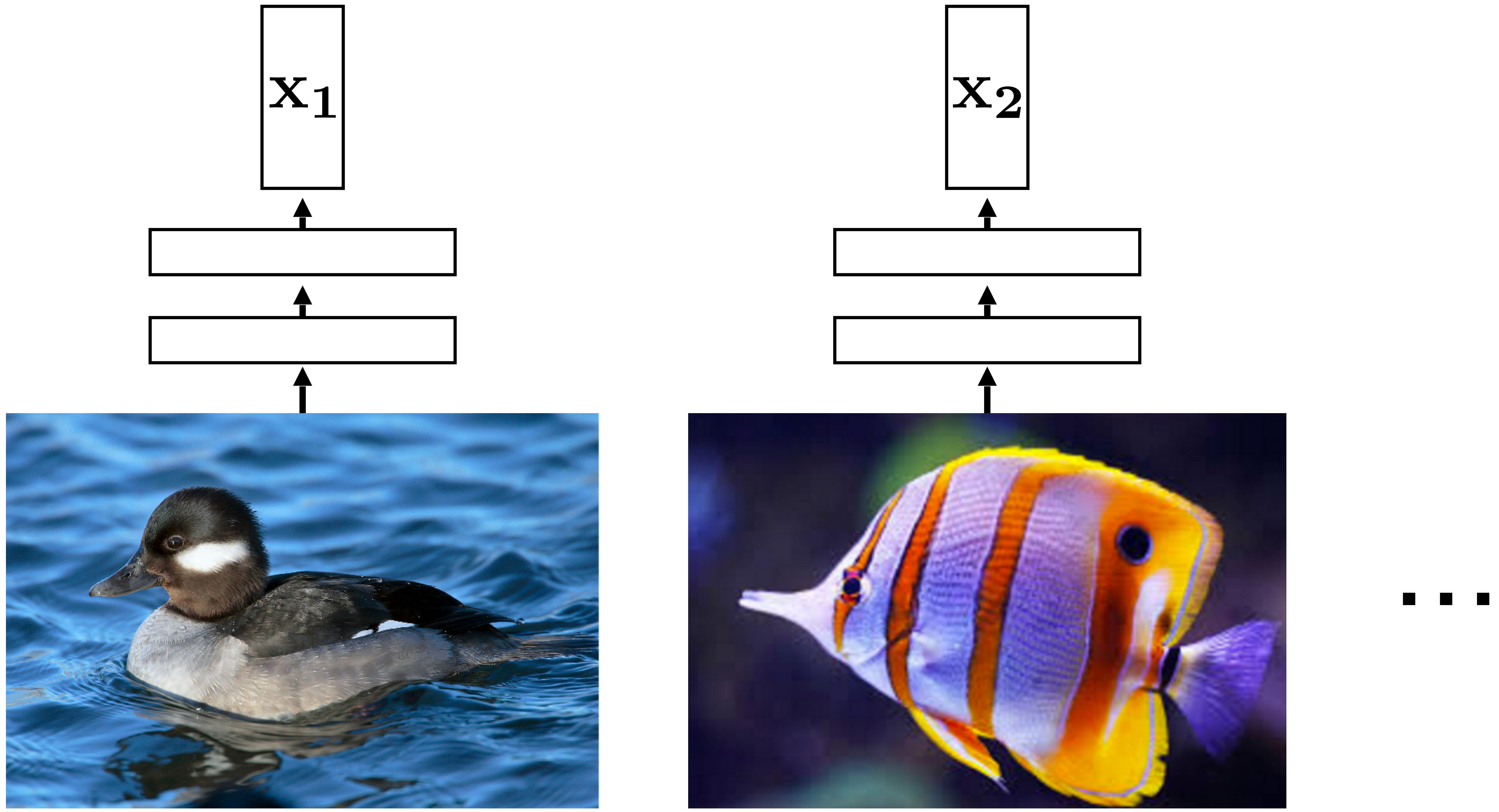


Contrastive learning

$\mathbf{z}^\top \mathbf{z} \longrightarrow$ High dot product with self

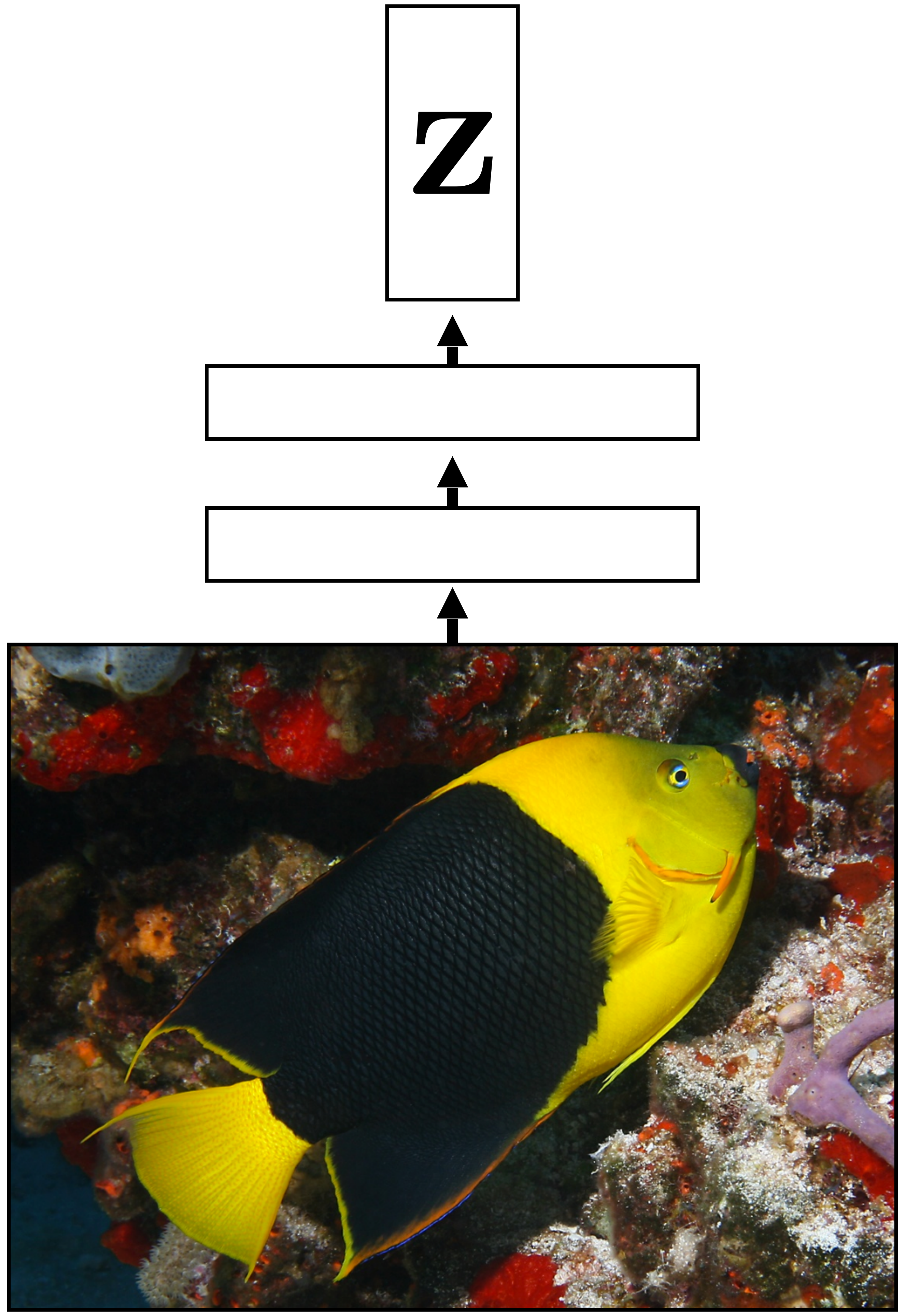
$\mathbf{z}^\top \mathbf{x}_1 \longrightarrow$ Low dot product with others





[Wu et al., Instance discrimination 2018], [He et al. Momentum contrastive learning 2019]

Contrastive learning



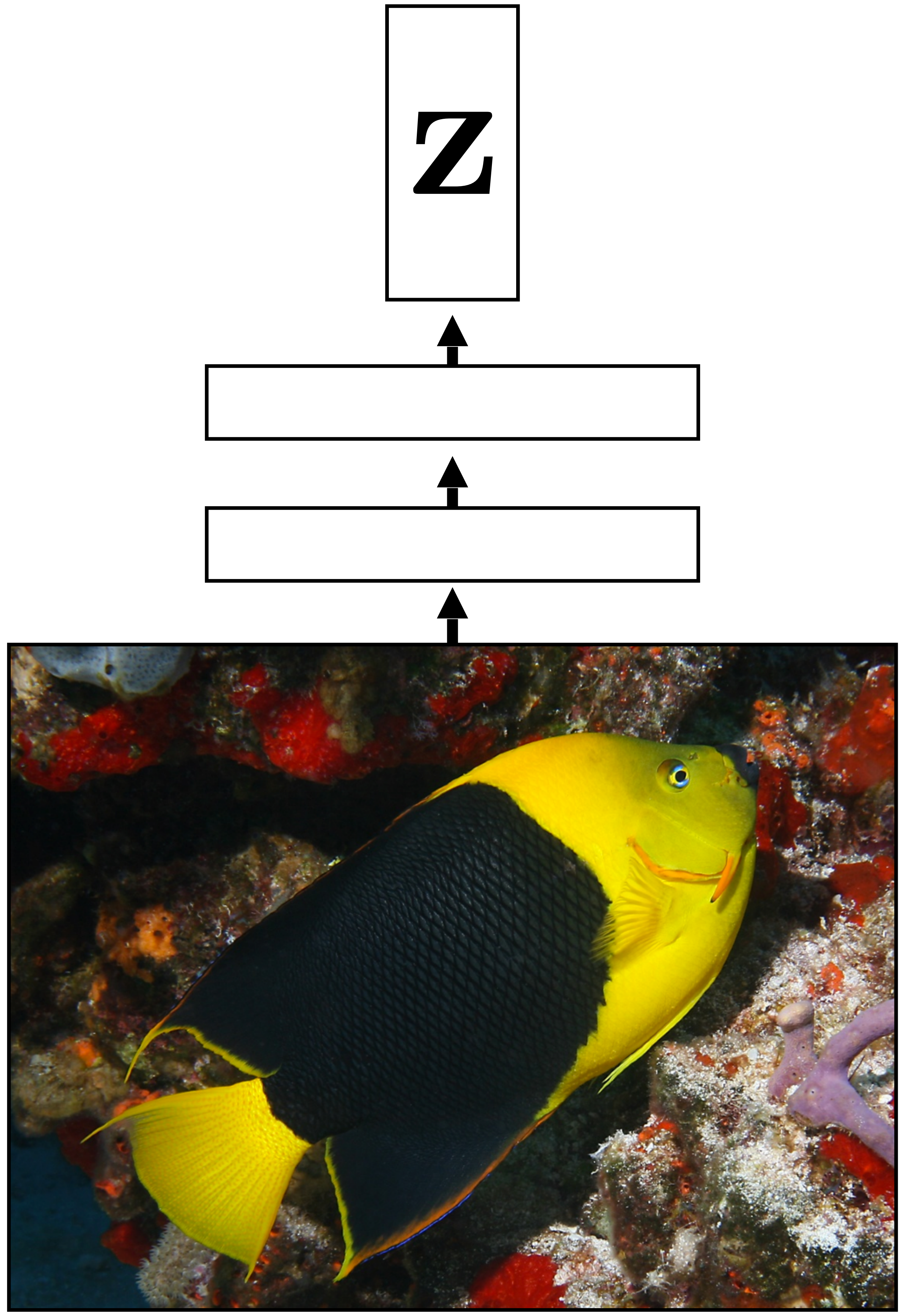
Minimize:

$$\mathscr{L} = -\log\left(\frac{\exp(\mathbf{z}^\top \mathbf{z})}{\sum_{i=1}^{n} \exp(z^\top \mathbf{x}_i)}\right)$$

Equivalent to softmax loss with each image as a category.

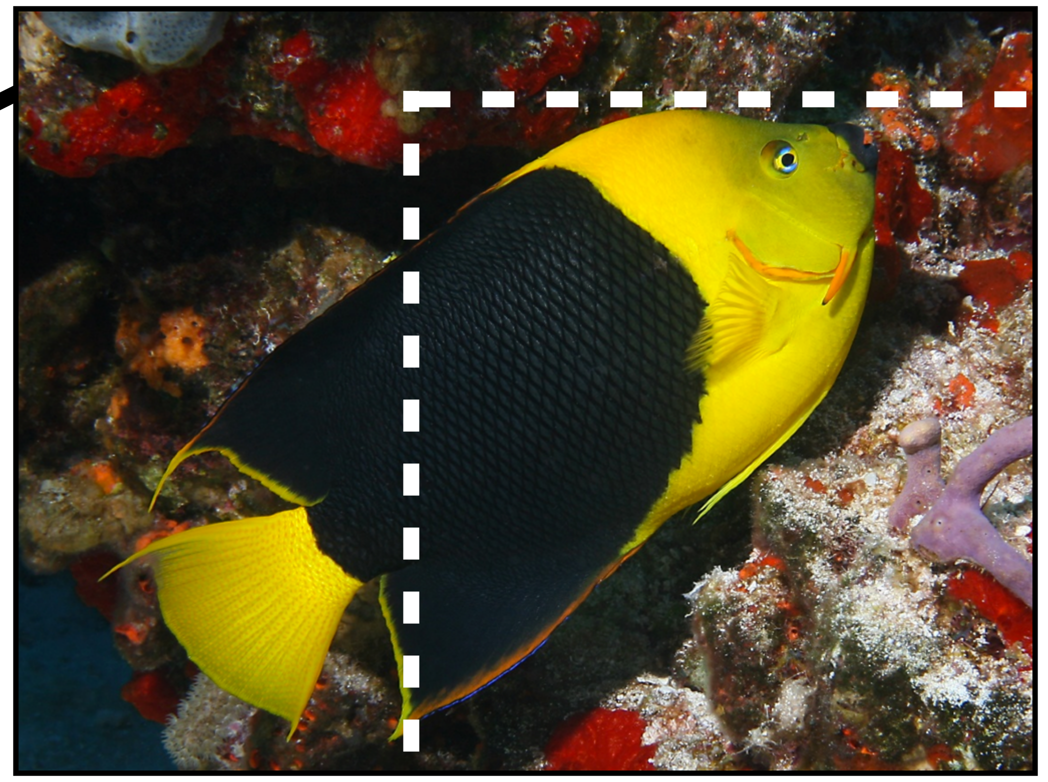
[Wu et al., Instance discrimination 2018], [He et al. Momentum contrastive learning 2019]

Contrastive learning



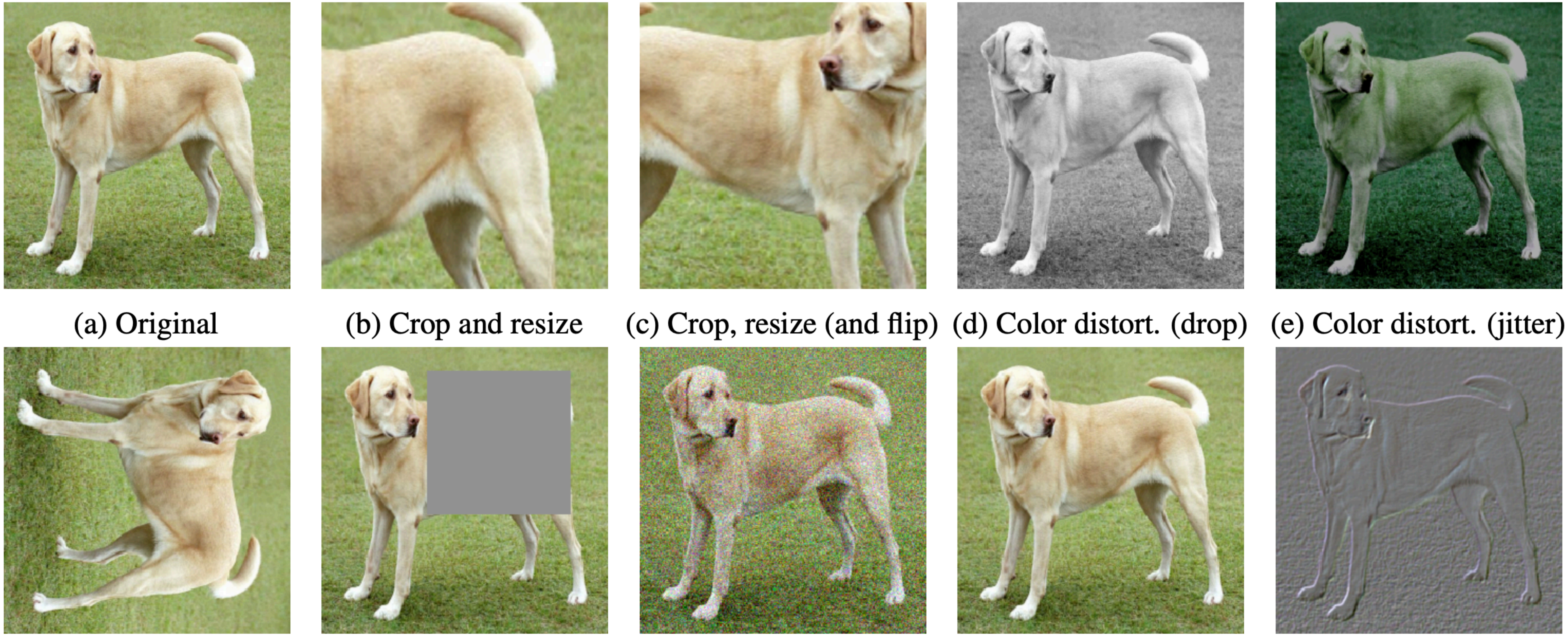
Build in invariance by comparing to **distorted** images.

$$\frac{\exp\{\mathbf{z}^\top \tilde{\mathbf{z}}\}}{\exp\{\mathbf{z}^\top \tilde{\mathbf{z}} + \sum_i \mathbf{z}^\top \mathbf{x}_i\}}$$



[Wu et al., Instance discrimination 2018], [He et al. Momentum contrastive learning 2019]

Data augmentation used in contrastive learning

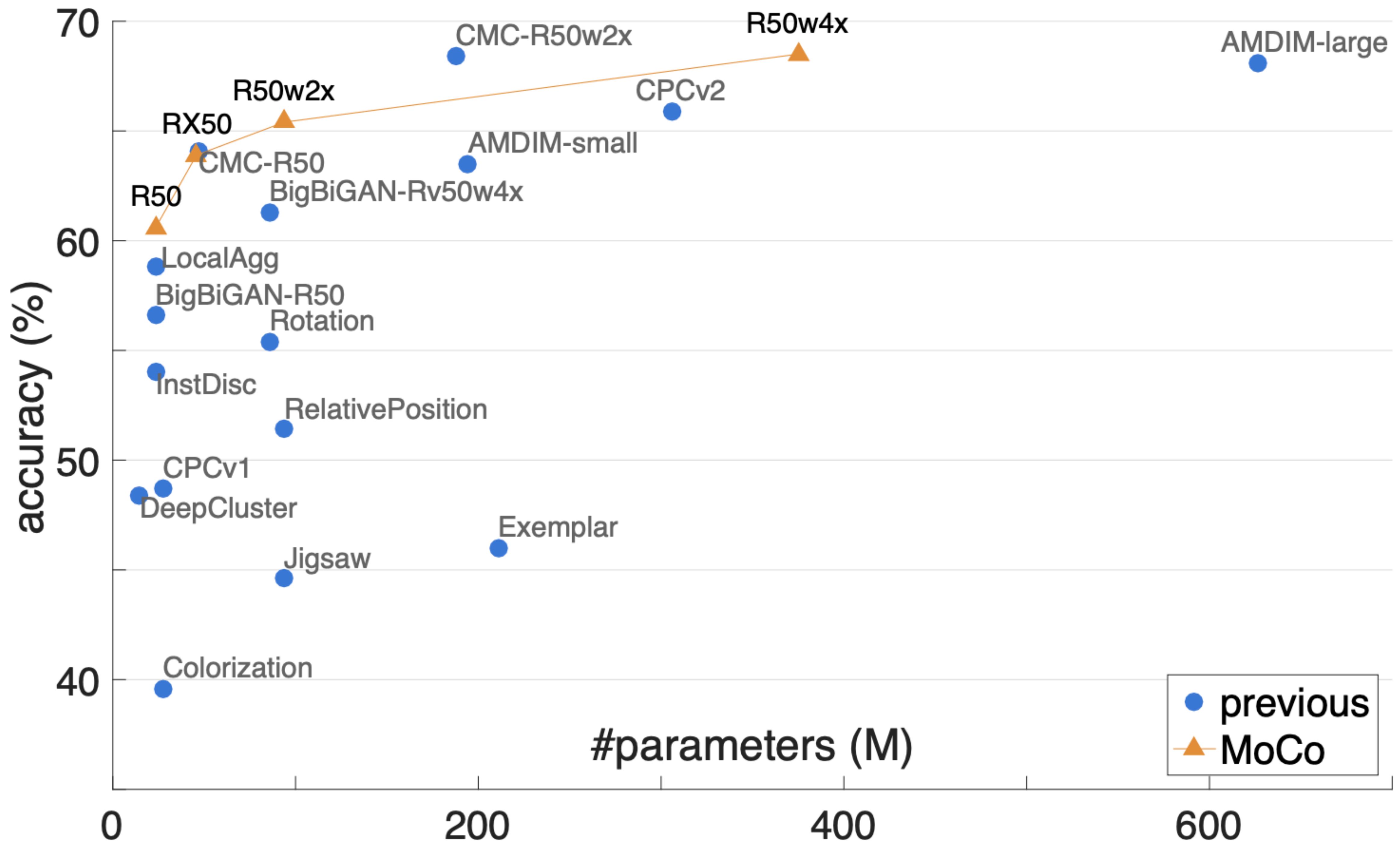


(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

[Chen et al., SimCLR, 2020]

Performance snapshot



ImageNet linear classification

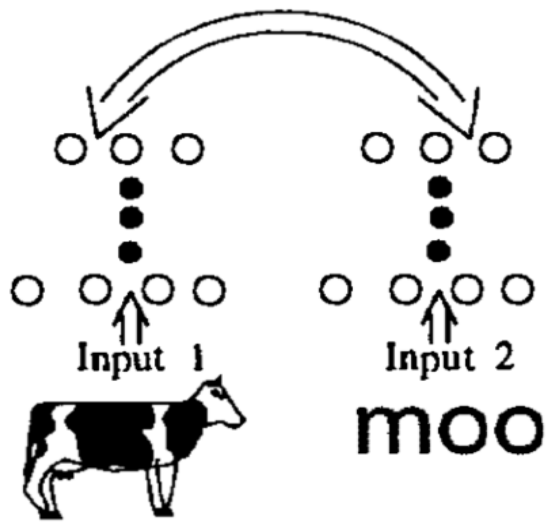
pre-train	AP_{50}
random init.	52.5
super. IN-1M	80.8
MoCo IN-1M	81.4 **(+0.6)**
MoCo IG-1B	82.1 **(+1.3)**

Object detection finetuning

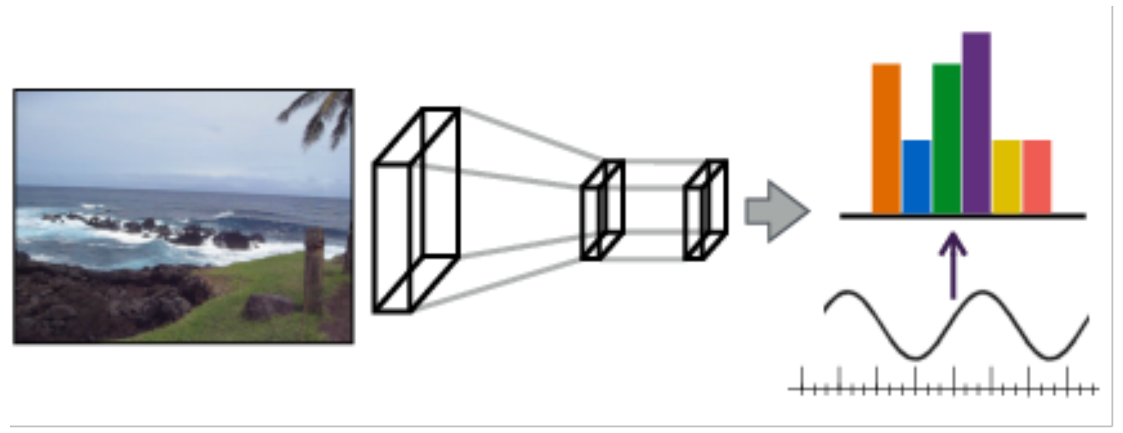
Comparable in many cases to supervised pretraining.

[He et al. Momentum contrastive learning 2019]

Audio

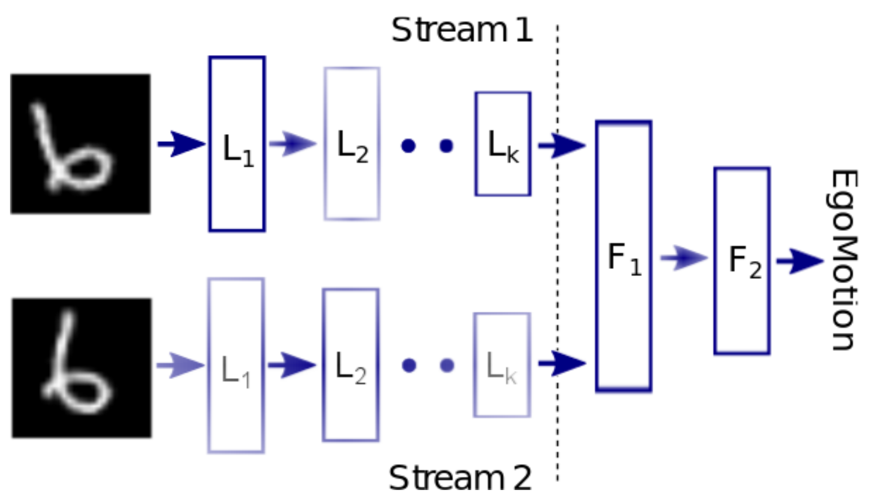


de Sa. NIPS 1994.



Owens et al. ECCV 2016.

Egomotion

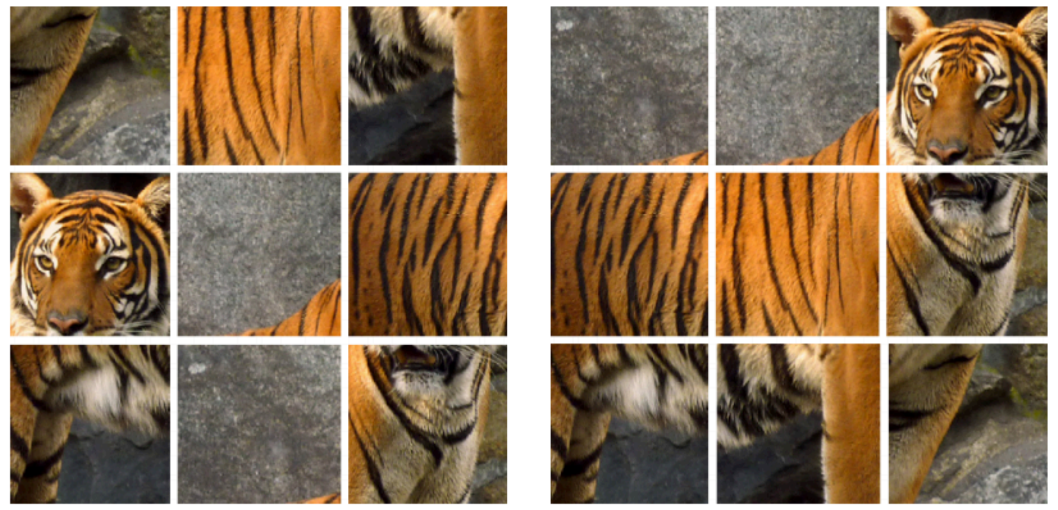


Agrawal et al. ICCV 2015.

Jayaraman et al. ICCV 2015.

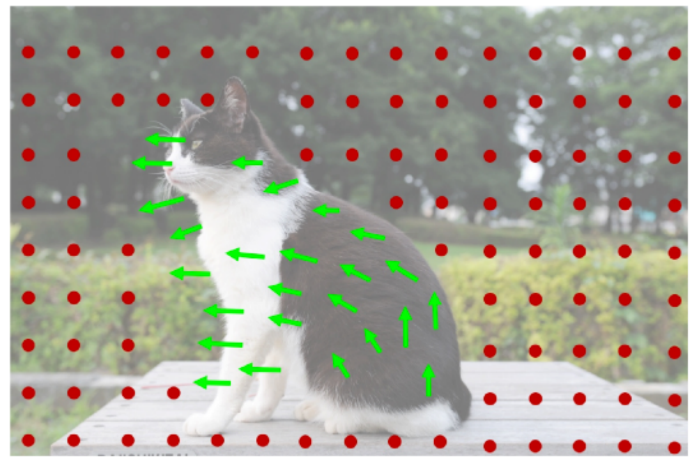
Context

Pathak et al. CVPR 2016.



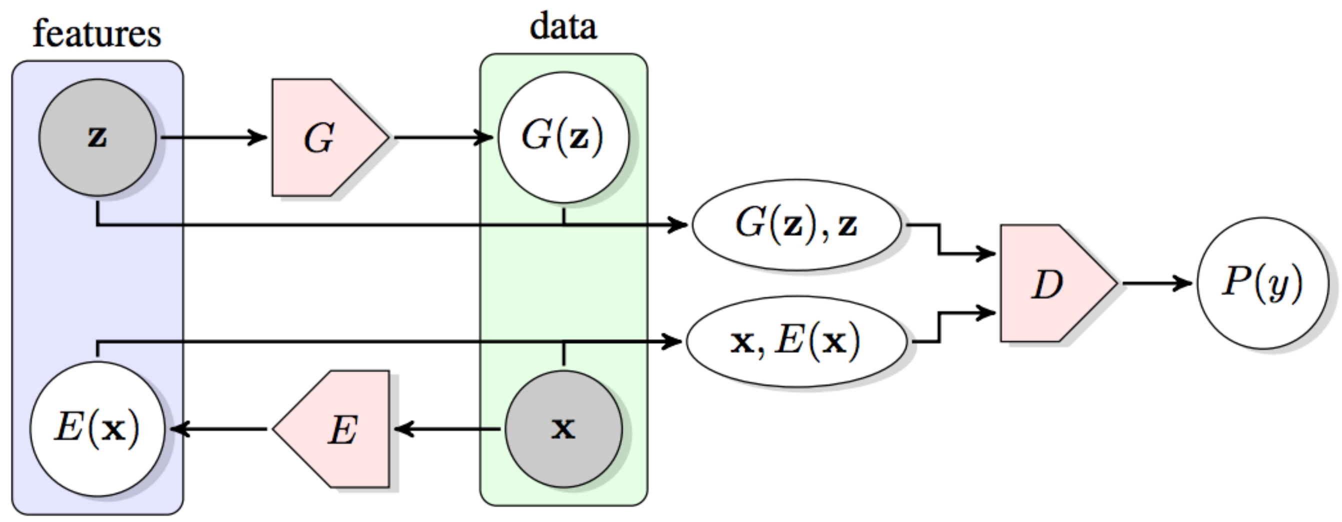
Noroozi and Favaro. ECCV 2016.
Doersch et al. ICCV 2015.

Video



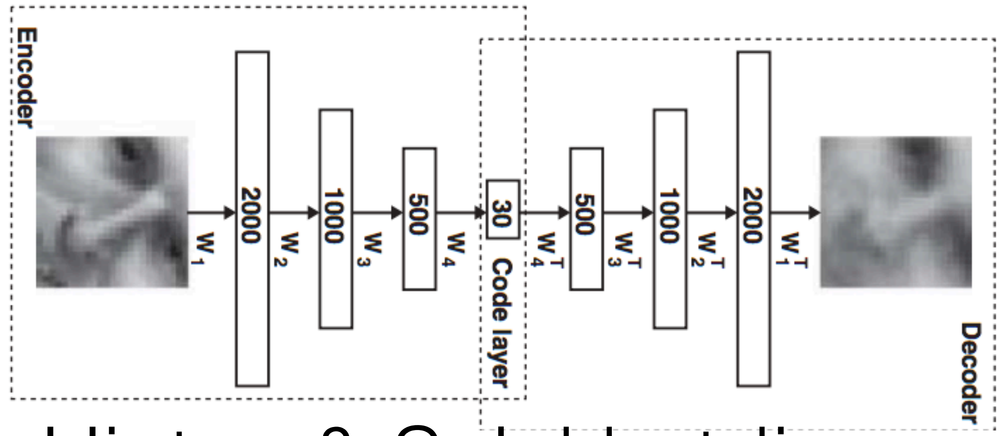
Wang et al. ICCV 2015. Pathak et al. CVPR 2017.
Misra et al. ECCV 2016.

Generative Modeling



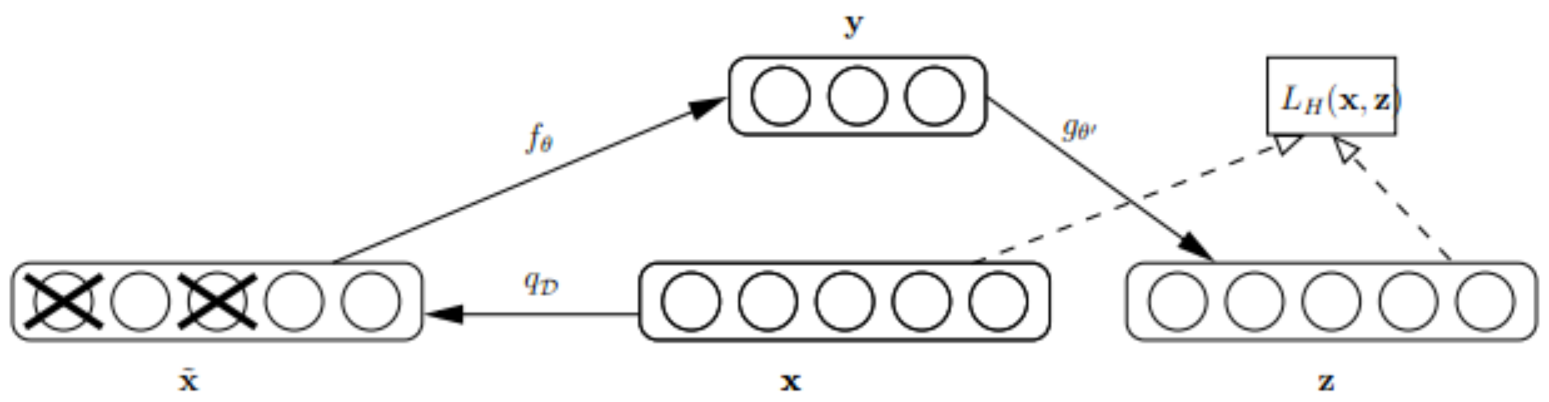
Donahue et al. Dumoulin et al. ICLR 2017.

Autoencoders



Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders



Vincent et al. ICML 2008.

Goal: Set up a pre-training scheme to induce a "useful" representation

Source: Richard Zhang

Language


Soap box derby
Bicycle
Person
Cart
Grate
Ramp

Language-based supervision

WIKIPEDIA

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

Languages
Français
한국어
Edit links

Article Talk

Soap Box Derby

Using standardized wheels with precision ball bearings, modern gravity-powered racers start at a ramp on top of a hill, attaining speeds of up to 35 miles per hour. Rally races and qualifying races in cities around the world use advanced timing systems that measure the time difference between the competing cars to the thousandth of a second to determine the winner of a heat. Each heat of a race lasts less than 30 seconds. Most races are double elimination races in which a racer that loses a heat can work their way through the Challenger's Bracket in an attempt to win the overall race. The annual World Championship race in Akron, however, is a single elimination race which uses overhead photography, triggered by a timing system, to determine the winner of each heat. Approximately 500 racers compete in two or three heats to determine a World Champion in each divisions.

There are three racing divisions in most locals and at the All-American competition.[10] The Stock division is designed to give the first-time builder a learning experience. Boys and girls, ages 7 through 13, compete in simplified cars built from kits purchased from the All-American. These kits assist the Derby novice by providing a step-by-step layout for construction of a basic lean forward style car. The Super Stock Car division, ages 9 through 18, gives the competitor an opportunity to expand their knowledge and build a more advanced model. Both of these beginner levels make use of kits and shells available from the All-American. These entry levels of racing are popular in race communities across the country, as youngsters are exposed to the Derby program for the first time. The Masters division offers boys and girls, ages 10 through 20, an advanced class of racer in which to try their creativity and design skills. Masters entrants may purchase a Scottie Masters Kit with a fiberglass body from the All-American Soap Box Derby.

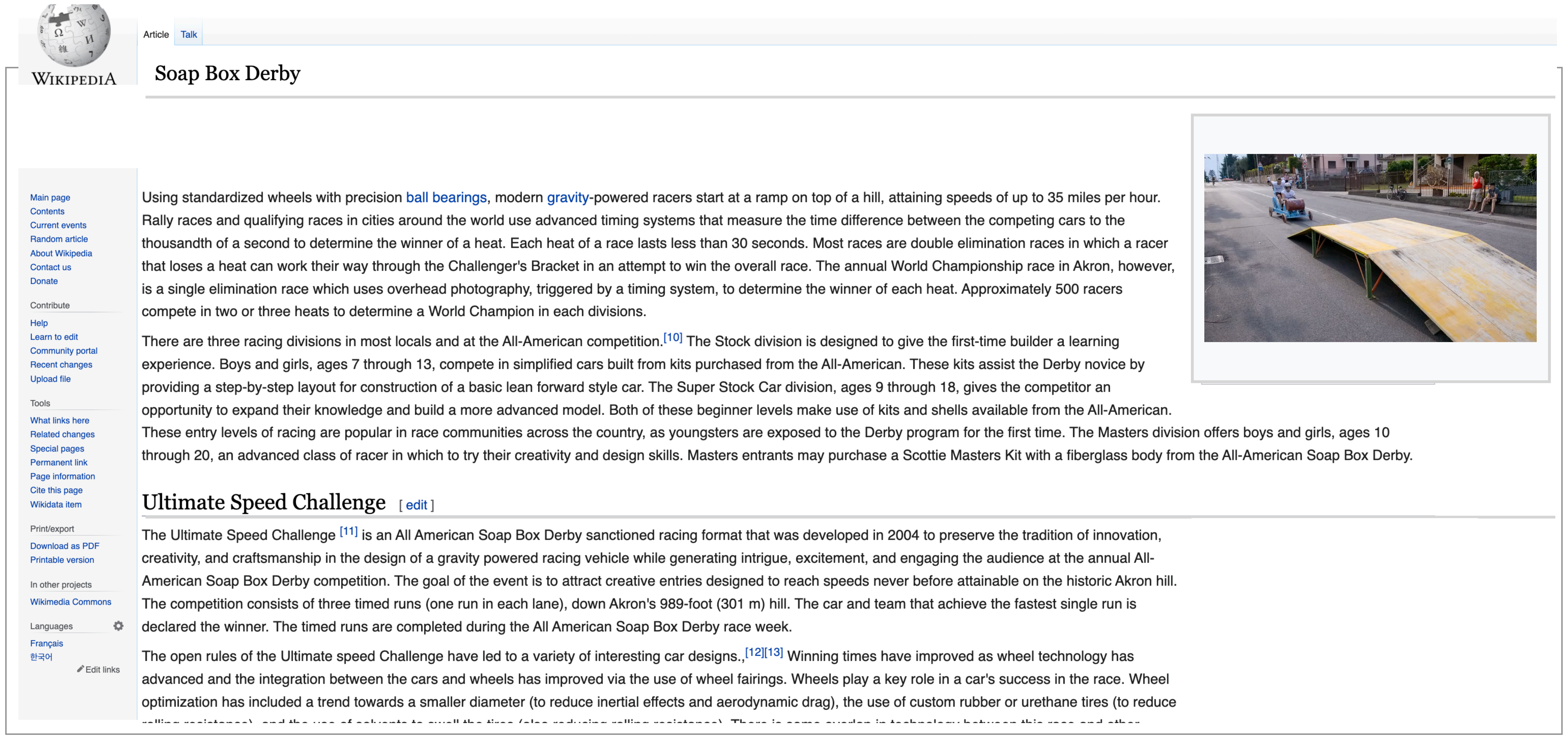
Ultimate Speed Challenge [edit]

The Ultimate Speed Challenge [11] is an All American Soap Box Derby sanctioned racing format that was developed in 2004 to preserve the tradition of innovation, creativity, and craftsmanship in the design of a gravity powered racing vehicle while generating intrigue, excitement, and engaging the audience at the annual All-American Soap Box Derby competition. The goal of the event is to attract creative entries designed to reach speeds never before attainable on the historic Akron hill. The competition consists of three timed runs (one run in each lane), down Akron's 989-foot (301 m) hill. The car and team that achieve the fastest single run is declared the winner. The timed runs are completed during the All American Soap Box Derby race week.

The open rules of the Ultimate speed Challenge have led to a variety of interesting car designs.,[12][13] Winning times have improved as wheel technology has advanced and the integration between the cars and wheels has improved via the use of wheel fairings. Wheels play a key role in a car's success in the race. Wheel optimization has included a trend towards a smaller diameter (to reduce inertial effects and aerodynamic drag), the use of custom rubber or urethane tires (to reduce

e.g., [Radford et al., "CLIP", 2021]

Language-based supervision



WIKIPEDIA

Article Talk

Soap Box Derby

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Learn to edit
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Wikidata item

Print/export
Download as PDF
Printable version

In other projects
Wikimedia Commons

Languages
Français
한국어
Edit links

Using standardized wheels with precision ball bearings, modern gravity-powered racers start at a ramp on top of a hill, attaining speeds of up to 35 miles per hour. Rally races and qualifying races in cities around the world use advanced timing systems that measure the time difference between the competing cars to the thousandth of a second to determine the winner of a heat. Each heat of a race lasts less than 30 seconds. Most races are double elimination races in which a racer that loses a heat can work their way through the Challenger's Bracket in an attempt to win the overall race. The annual World Championship race in Akron, however, is a single elimination race which uses overhead photography, triggered by a timing system, to determine the winner of each heat. Approximately 500 racers compete in two or three heats to determine a World Champion in each divisions.

There are three racing divisions in most locals and at the All-American competition.[10] The Stock division is designed to give the first-time builder a learning experience. Boys and girls, ages 7 through 13, compete in simplified cars built from kits purchased from the All-American. These kits assist the Derby novice by providing a step-by-step layout for construction of a basic lean forward style car. The Super Stock Car division, ages 9 through 18, gives the competitor an opportunity to expand their knowledge and build a more advanced model. Both of these beginner levels make use of kits and shells available from the All-American. These entry levels of racing are popular in race communities across the country, as youngsters are exposed to the Derby program for the first time. The Masters division offers boys and girls, ages 10 through 20, an advanced class of racer in which to try their creativity and design skills. Masters entrants may purchase a Scottie Masters Kit with a fiberglass body from the All-American Soap Box Derby.

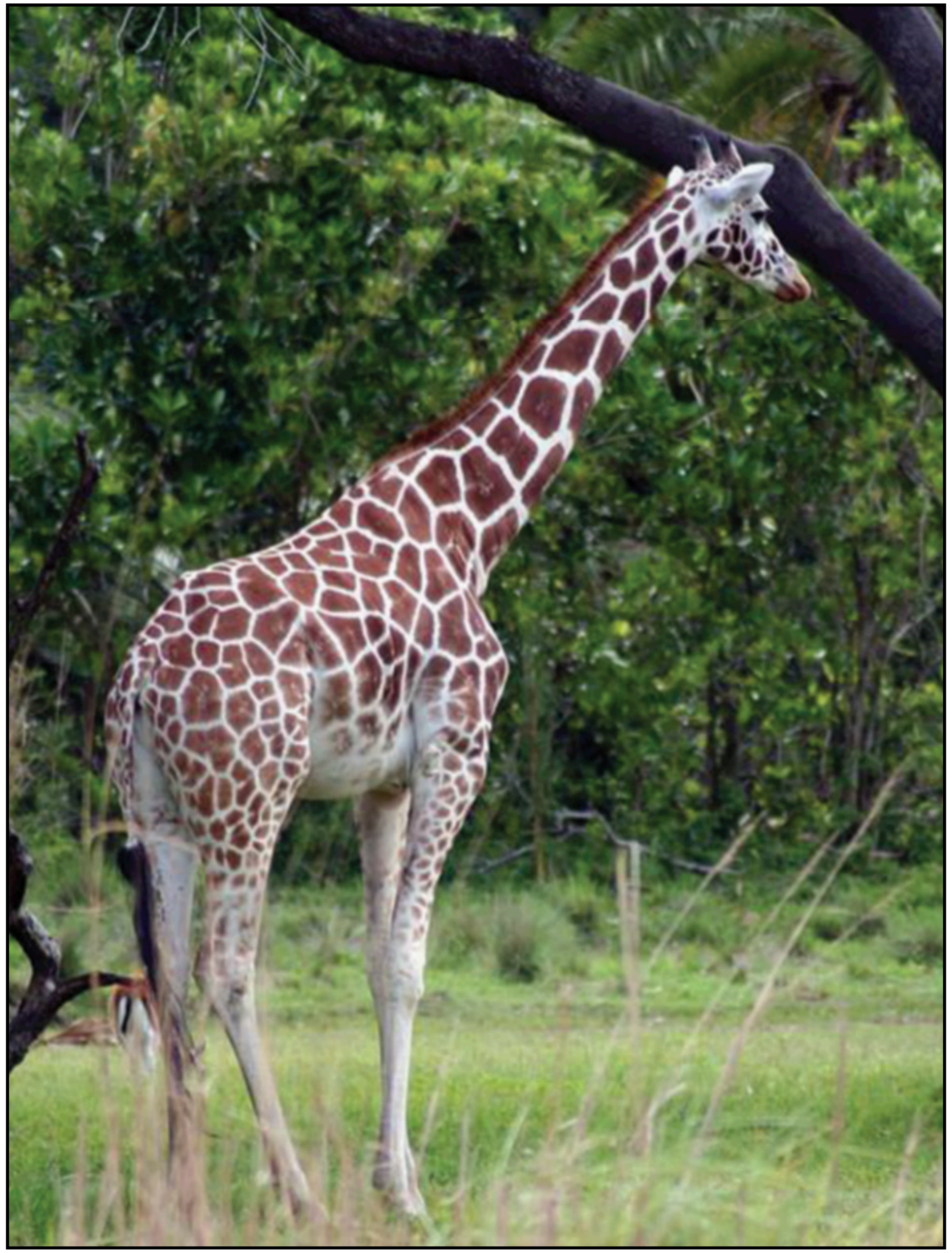
Ultimate Speed Challenge [edit]

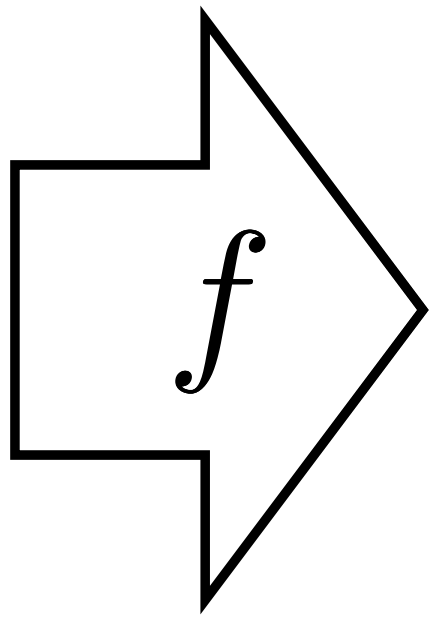
The Ultimate Speed Challenge [11] is an All American Soap Box Derby sanctioned racing format that was developed in 2004 to preserve the tradition of innovation, creativity, and craftsmanship in the design of a gravity powered racing vehicle while generating intrigue, excitement, and engaging the audience at the annual All-American Soap Box Derby competition. The goal of the event is to attract creative entries designed to reach speeds never before attainable on the historic Akron hill. The competition consists of three timed runs (one run in each lane), down Akron's 989-foot (301 m) hill. The car and team that achieve the fastest single run is declared the winner. The timed runs are completed during the All American Soap Box Derby race week.

The open rules of the Ultimate speed Challenge have led to a variety of interesting car designs.,[12][13] Winning times have improved as wheel technology has advanced and the integration between the cars and wheels has improved via the use of wheel fairings. Wheels play a key role in a car's success in the race. Wheel optimization has included a trend towards a smaller diameter (to reduce inertial effects and aerodynamic drag), the use of custom rubber or urethane tires (to reduce

e.g., [Radford et al., "CLIP", 2021]

What about language?



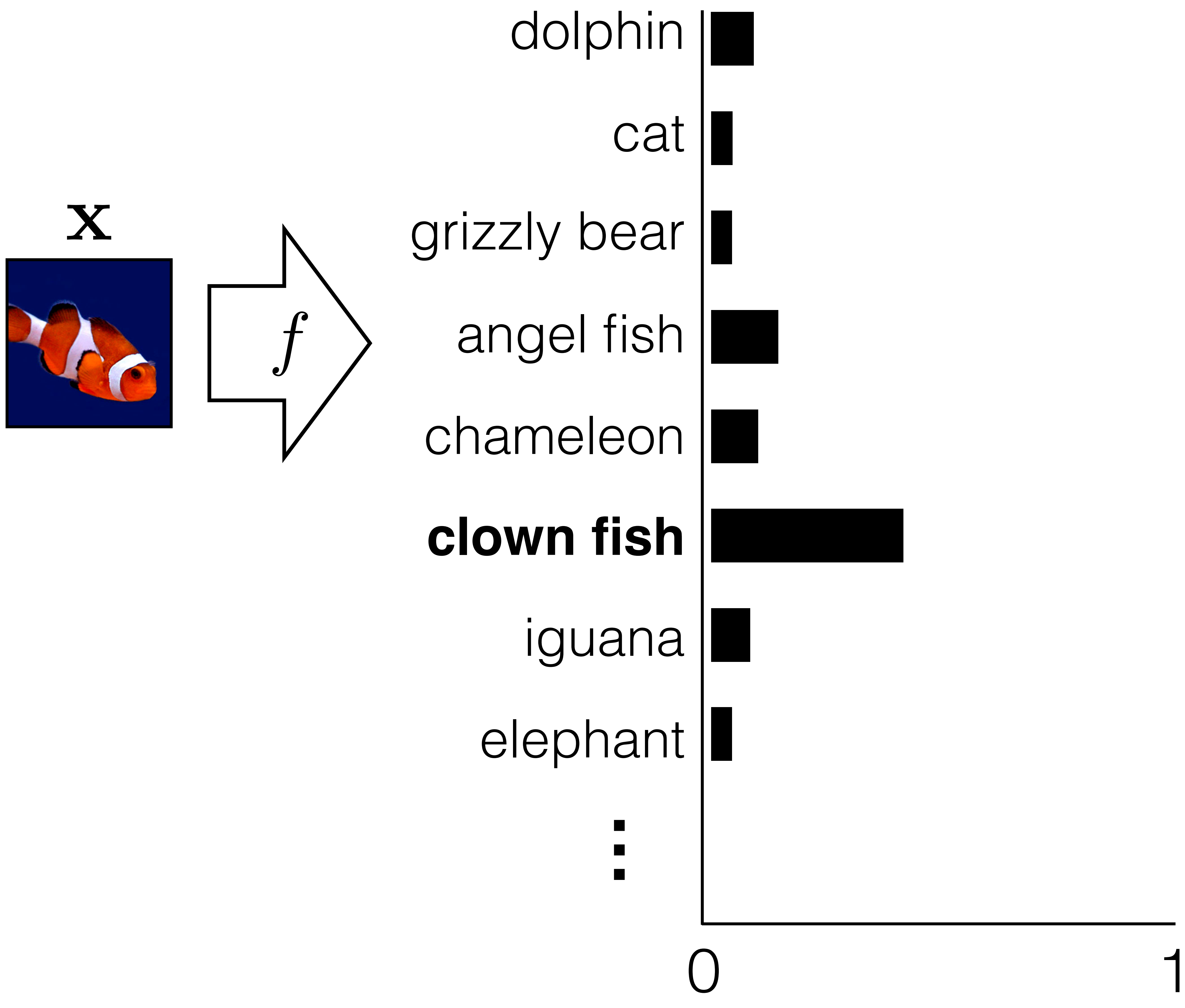


"A giraffe standing in the grass next to a tree"

How to represent words as numbers?

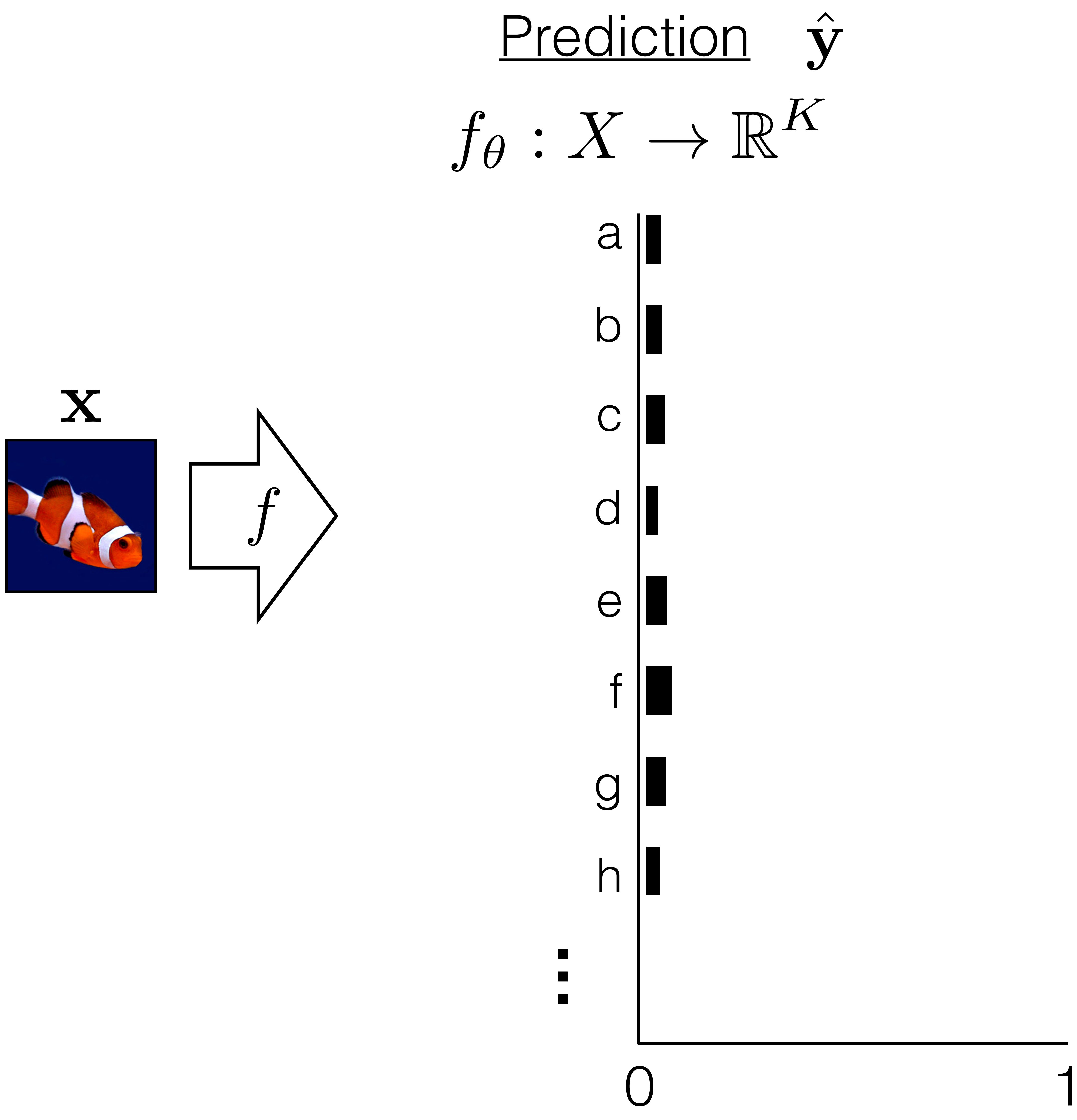
Prediction $\hat{\mathbf{y}}$

$$f_\theta : X \to \mathbb{R}^K$$



Source: Isola, Torralba, Freeman

How to represent words as numbers?



Or, represent each character as a class (e.g., K=26 for English letters),

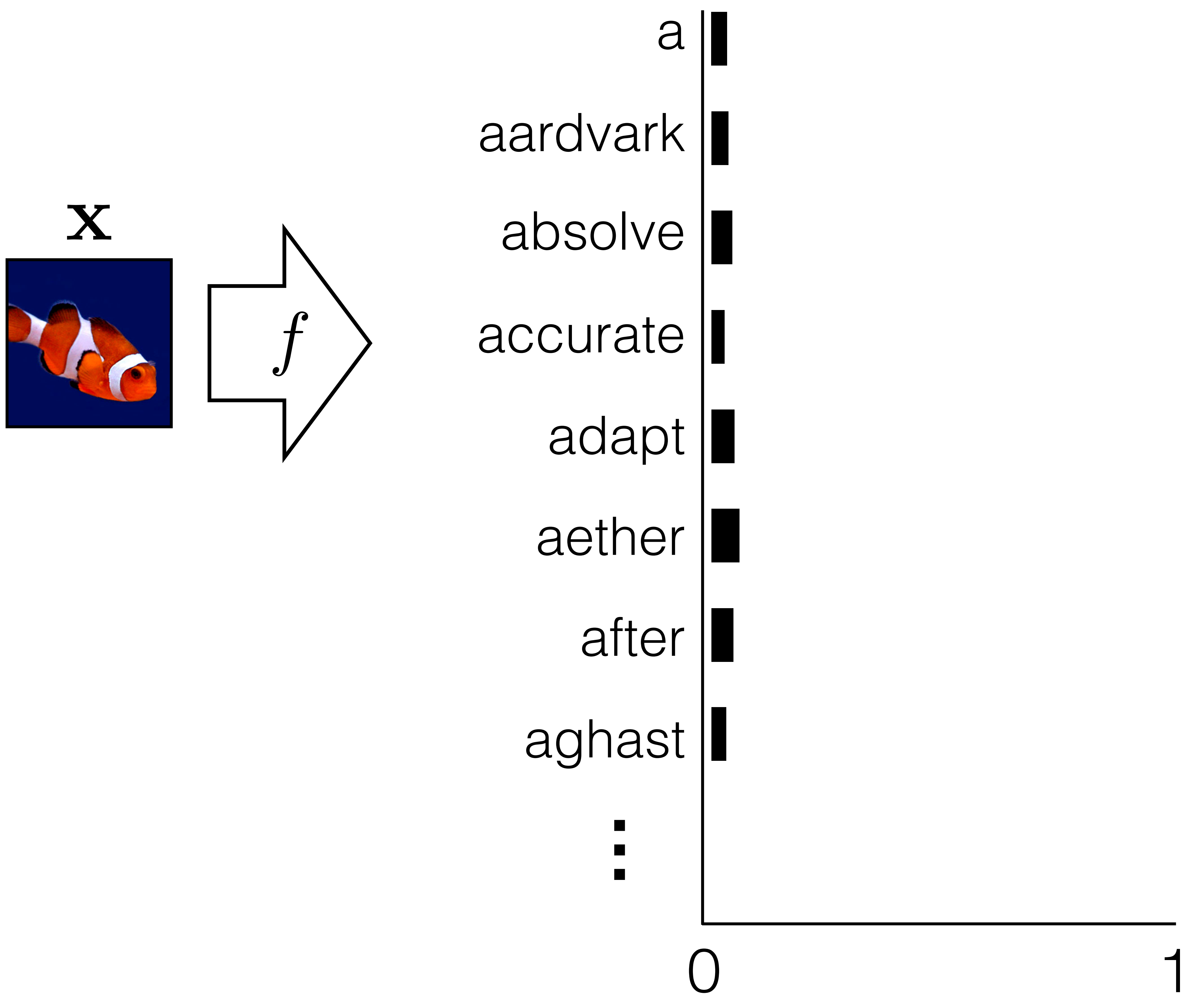
and represent words as a sequence of characters.

Source: Isola, Torralba, Freeman

How to represent words as numbers?

Prediction $\hat{\mathbf{y}}$

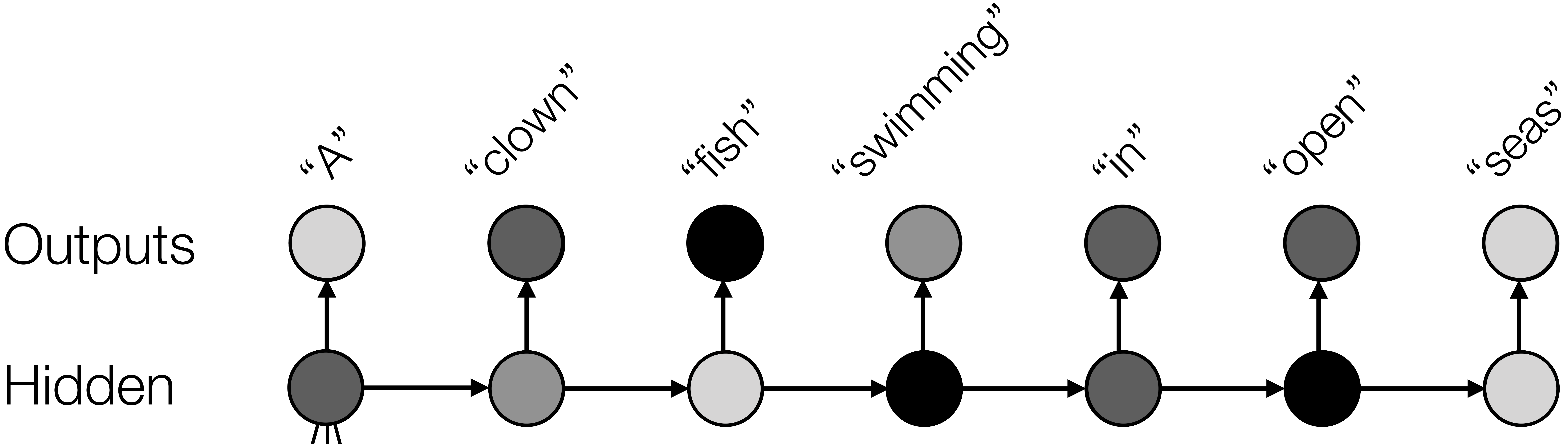
$f_\theta : X \to \mathbb{R}^K$

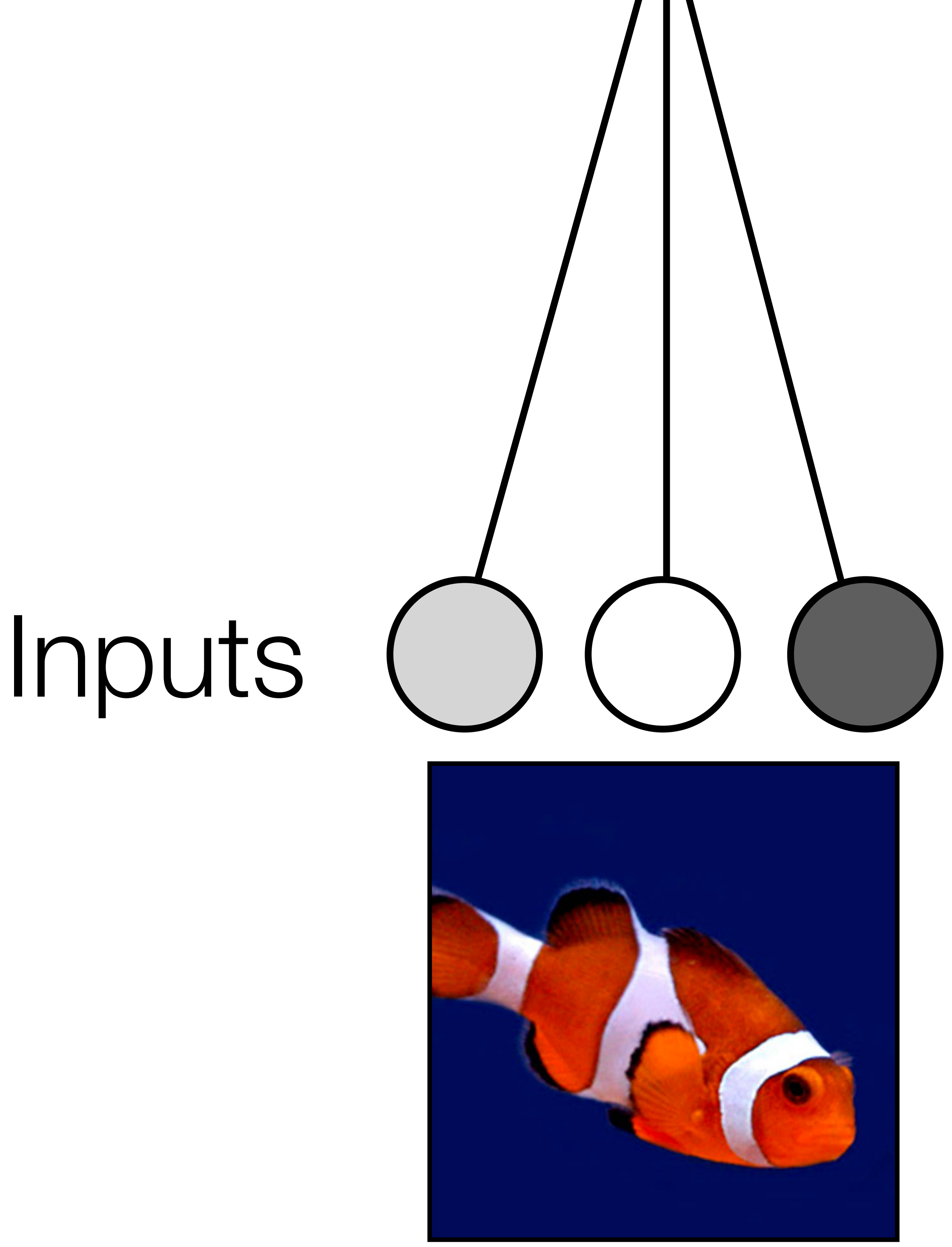


Rather than having just a handful of possible object classes, we can represent all words in a large vocabulary using a very large K (e.g., K=100,000).

Use “chunks” of characters instead.

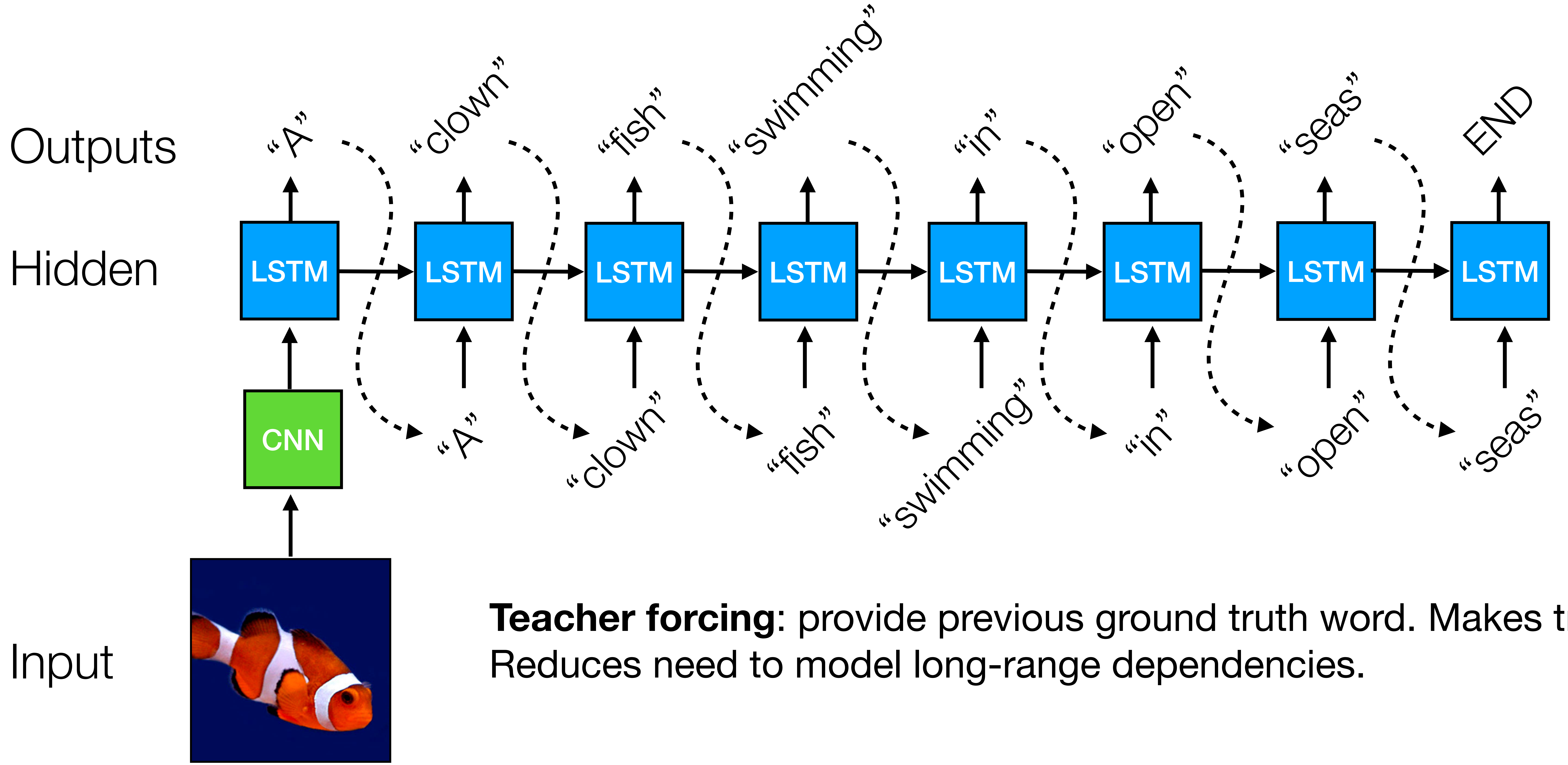
Source: Isola, Torralba, Freeman





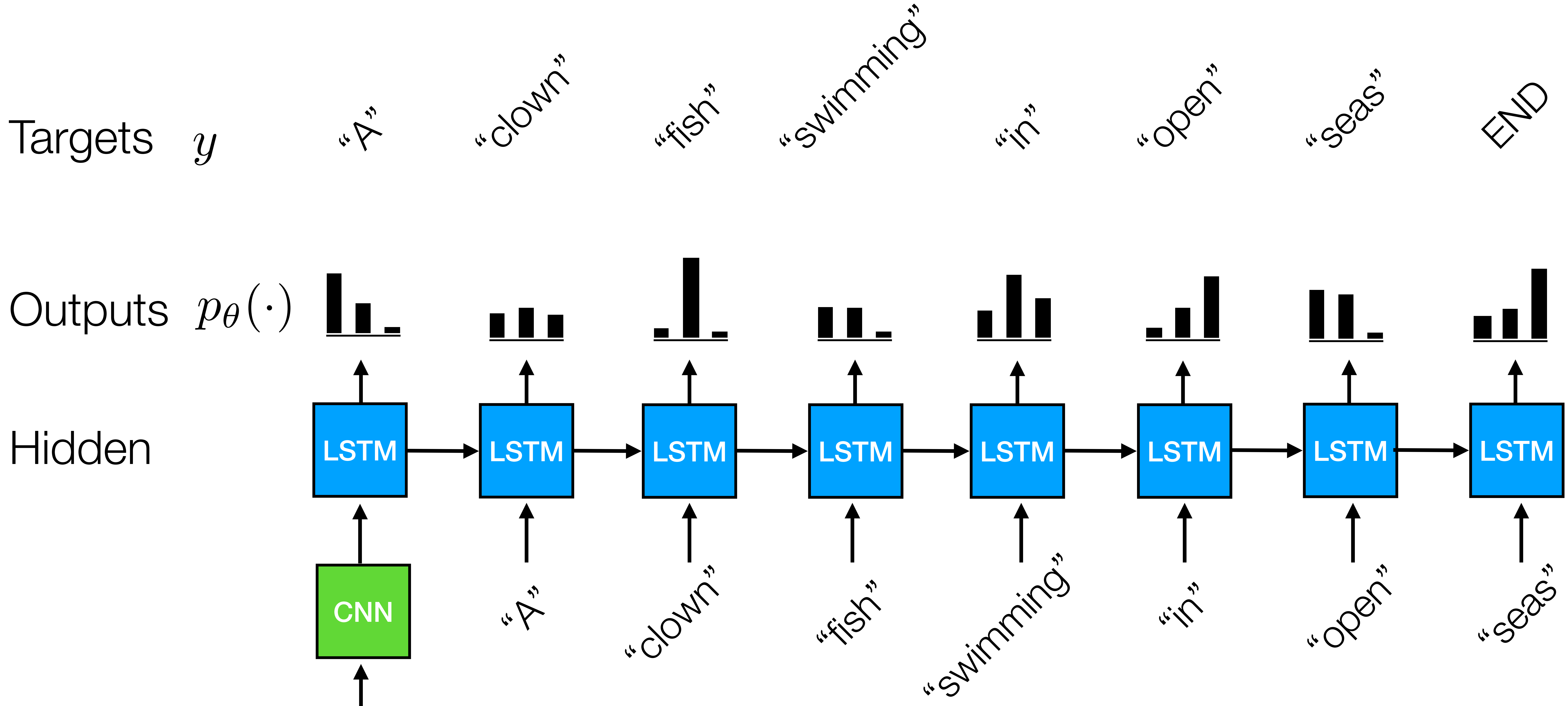
This problem is called **image captioning.**

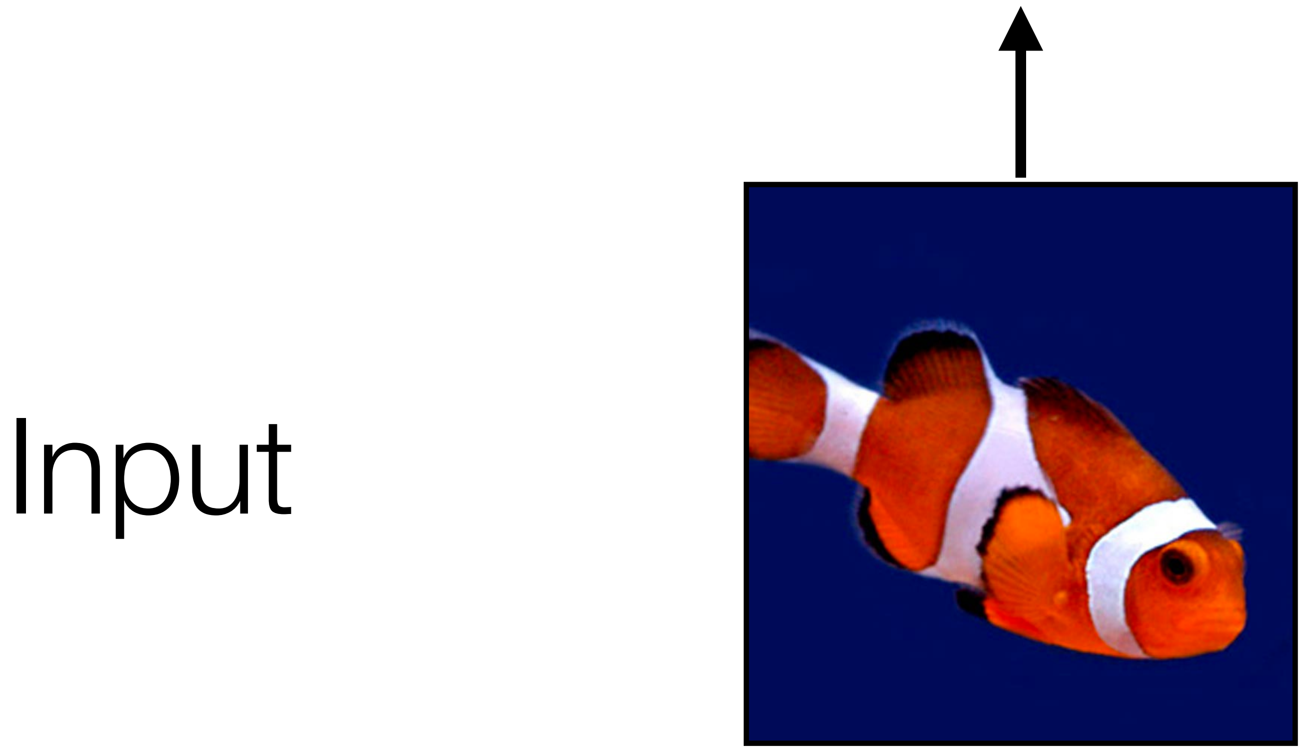
Source: Isola, Torralba, Freeman



Teacher forcing: provide previous ground truth word. Makes training easier. Reduces need to model long-range dependencies.

Source: Isola, Torralba, Freeman

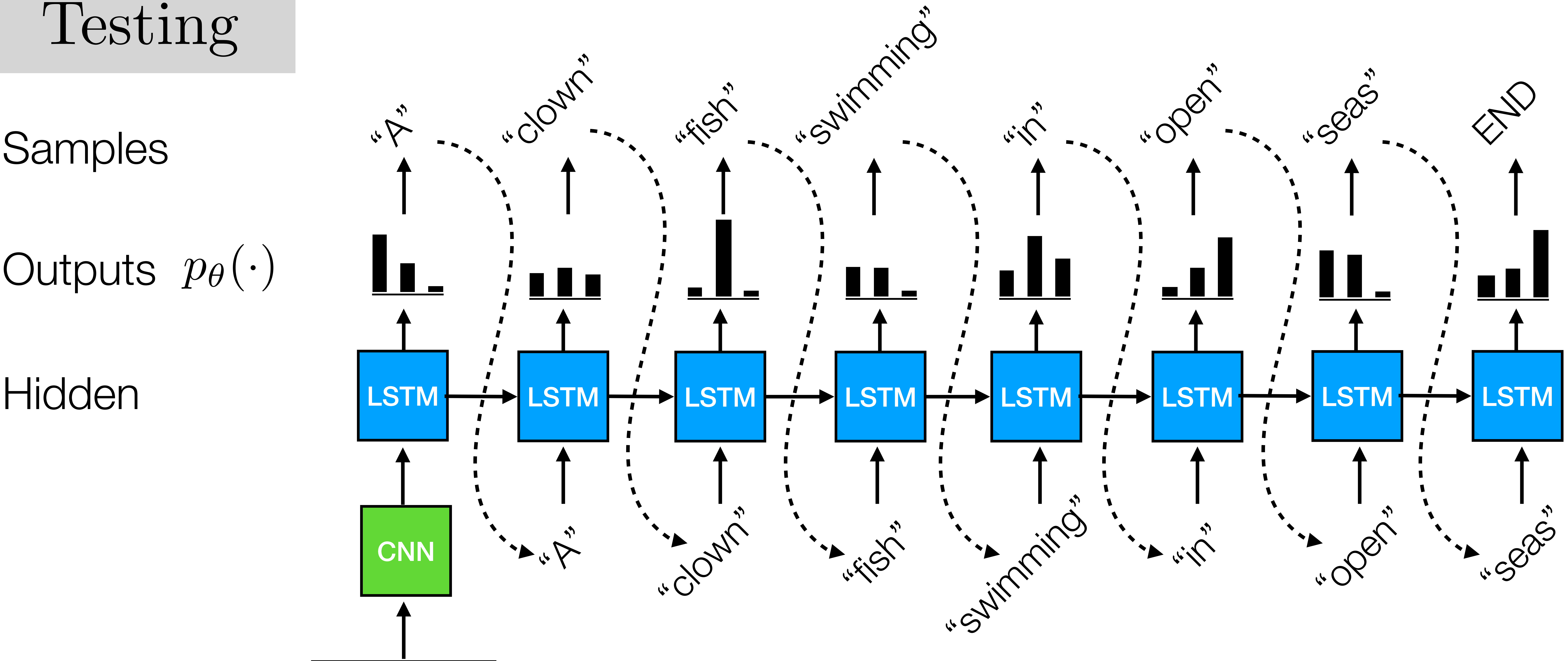




Maximum likelihood: maximize probability of all ground truth words.

Source: Isola, Torralba, Freeman

Testing



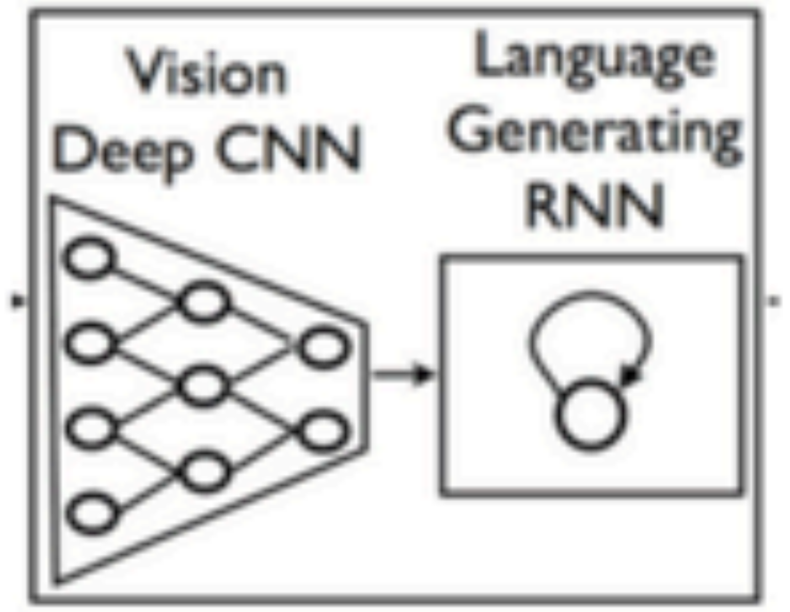
Input

Sample from predicted distribution over words.

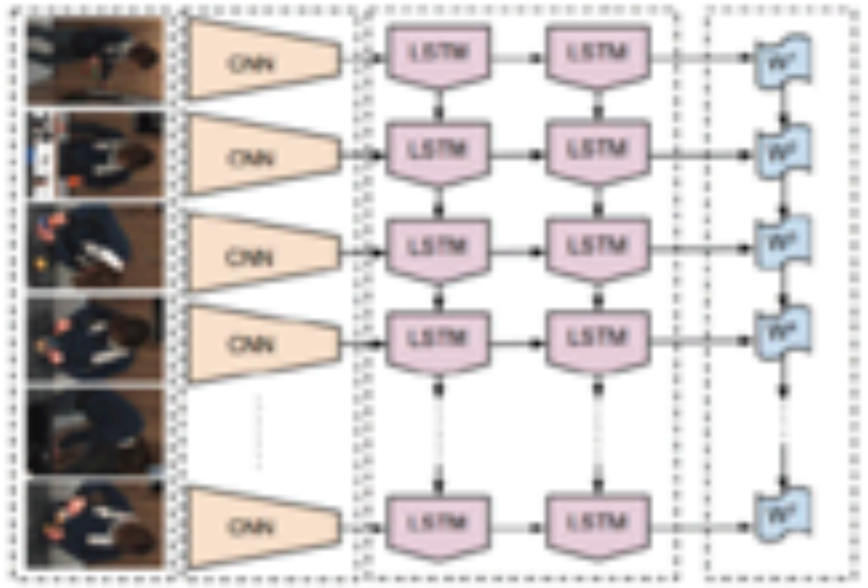
Alternatively, sample most likely word.

Source: Isola, Torralba, Freeman

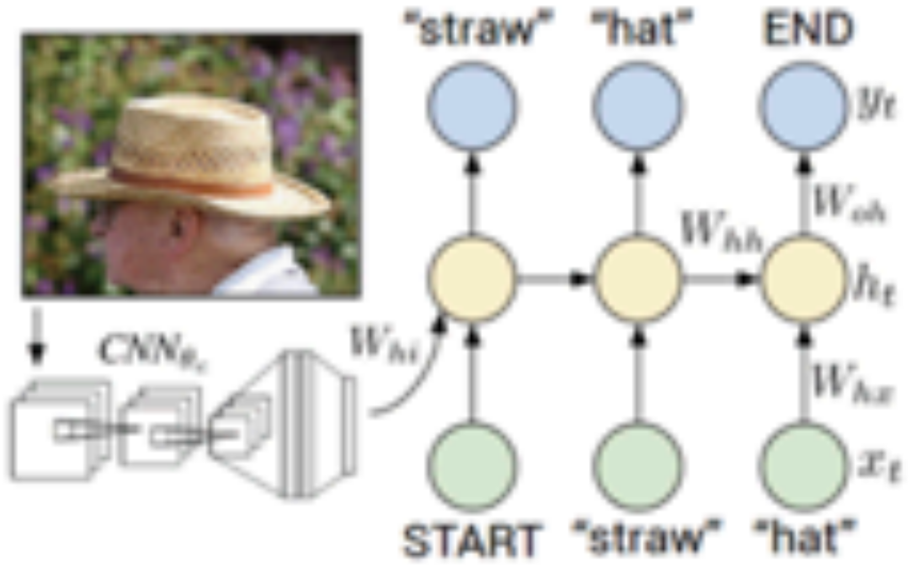
Captioning: popular topic circa 2015



Vinyals et al., 2015



Donahue et al., 2015



Karpathy and Fei-Fei, 2015

Hodosh et al., 2013

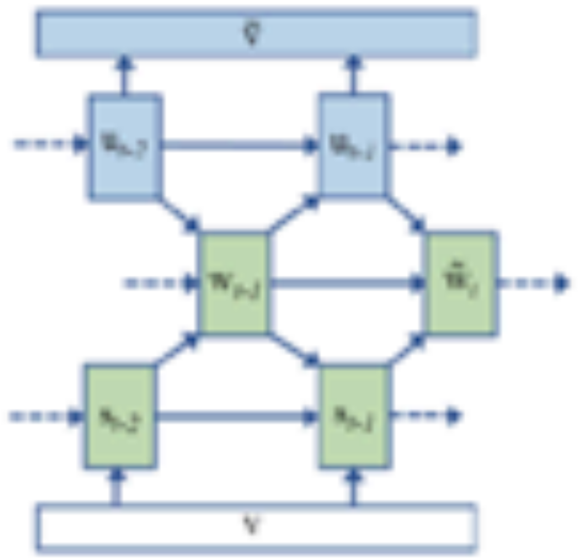
Fang et al., 2015

... and many more

Mao et al., 2015

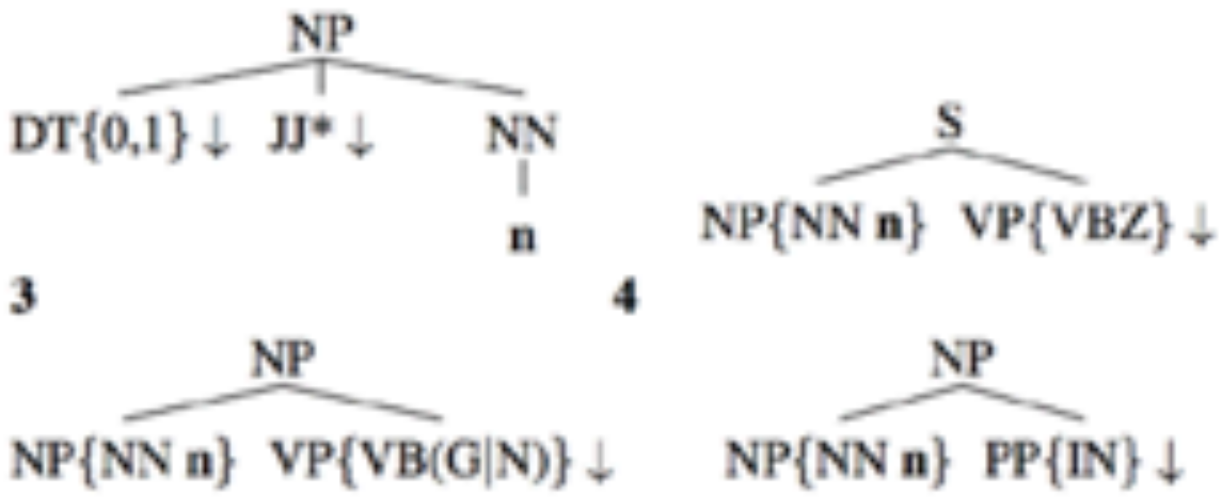
Ordonez et al., 2011

Kulkarni et al., 2011

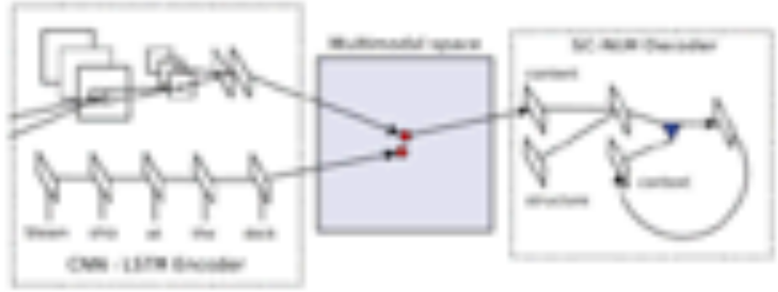


Chen and Zitnick, 2015

Farhadi et al., 2010



Mitchell et al., 2012

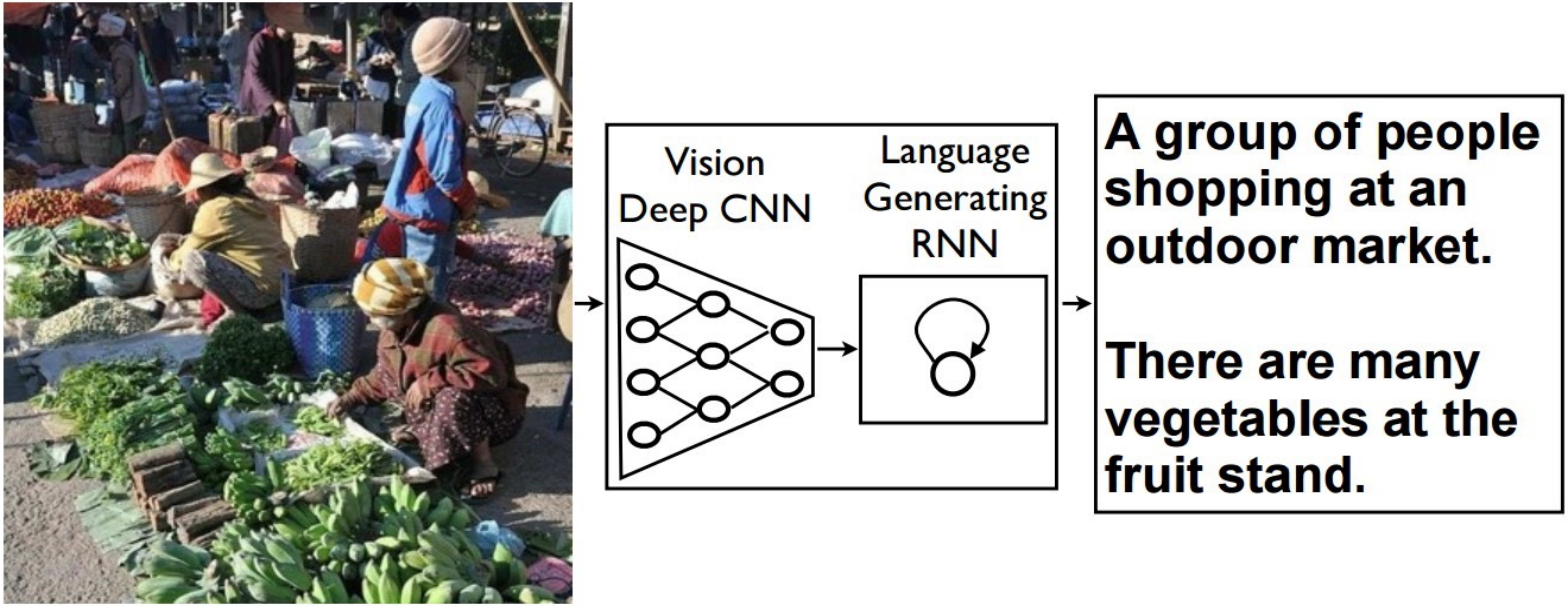


Kiros et al., 2015

Source: D. Parikh

Show and Tell: A Neural Image Caption Generator

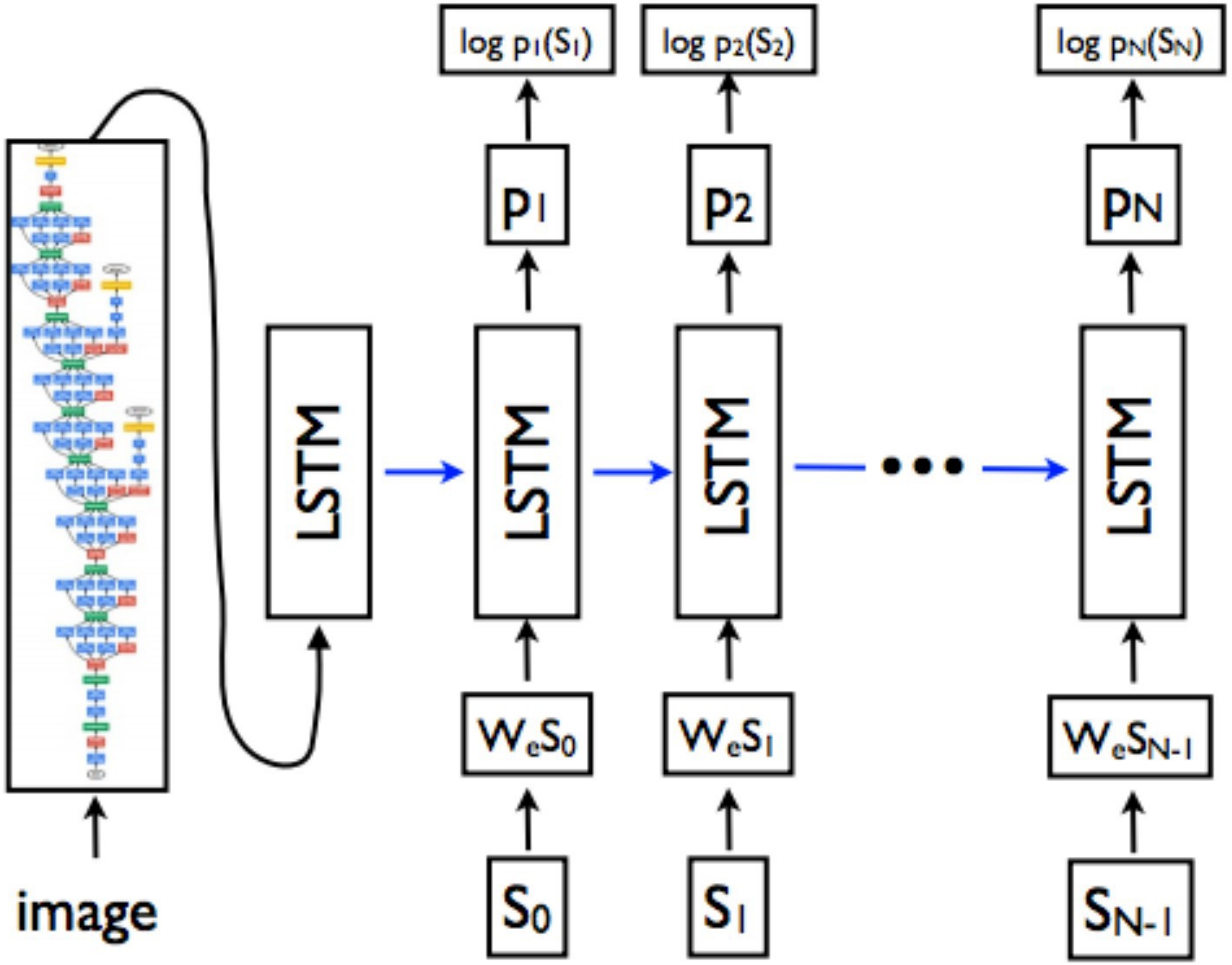
[Vinyals et. al., CVPR 2015]



Source: Isola, Torralba, Freeman

Show and Tell: A Neural Image Caption Generator

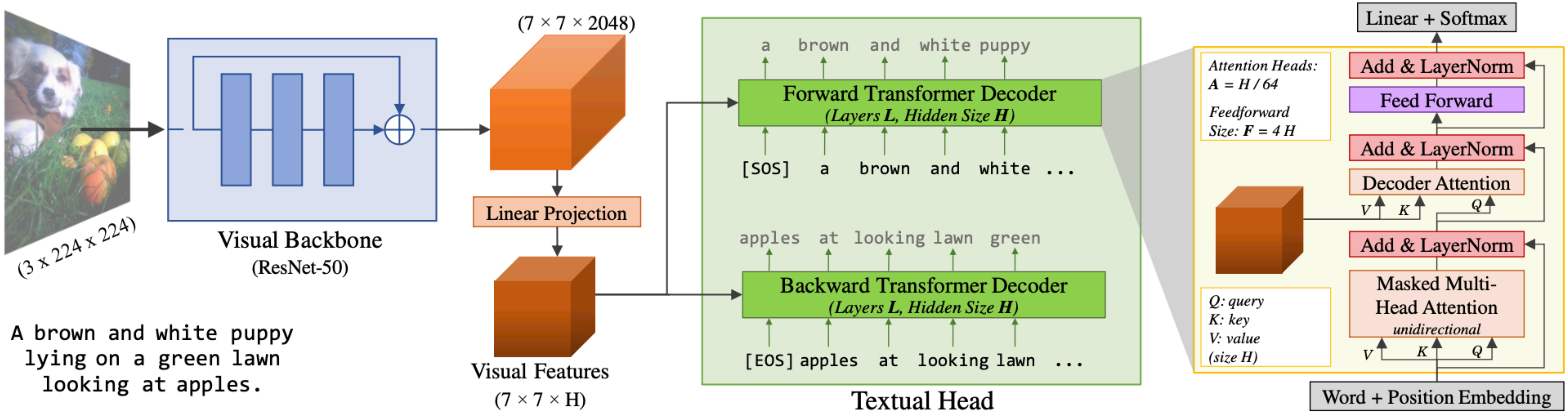
[Vinyals et. al., CVPR 2015]



A person riding a motorcycle on a dirt road.

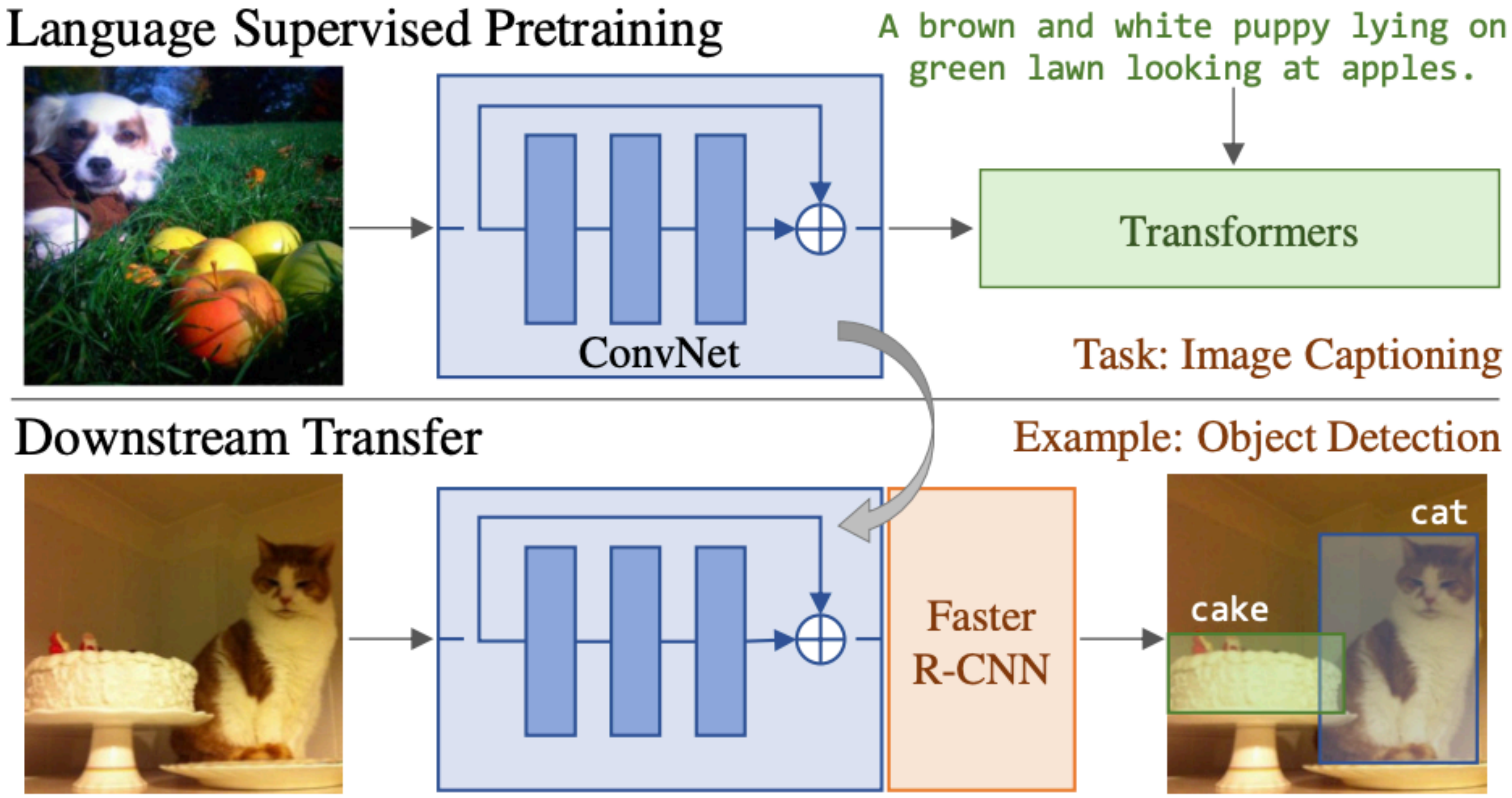

A person riding a motorcycle on a dirt road.
Two dogs play in the grass.
A skateboarder does a trick on a ramp.
A dog is jumping to catch a frisbee.
A group of young people playing a game of frisbee.
Two hockey players are fighting over the puck.
A little girl in a pink hat is blowing bubbles.
A refrigerator filled with lots of food and drinks.
A herd of elephants walking across a dry grass field.
A close up of a cat laying on a couch.
A red motorcycle parked on the side of the road.
A yellow school bus parked in a parking lot.
Describes without errors
Describes with minor errors
Somewhat related to the image
Unrelated to the image

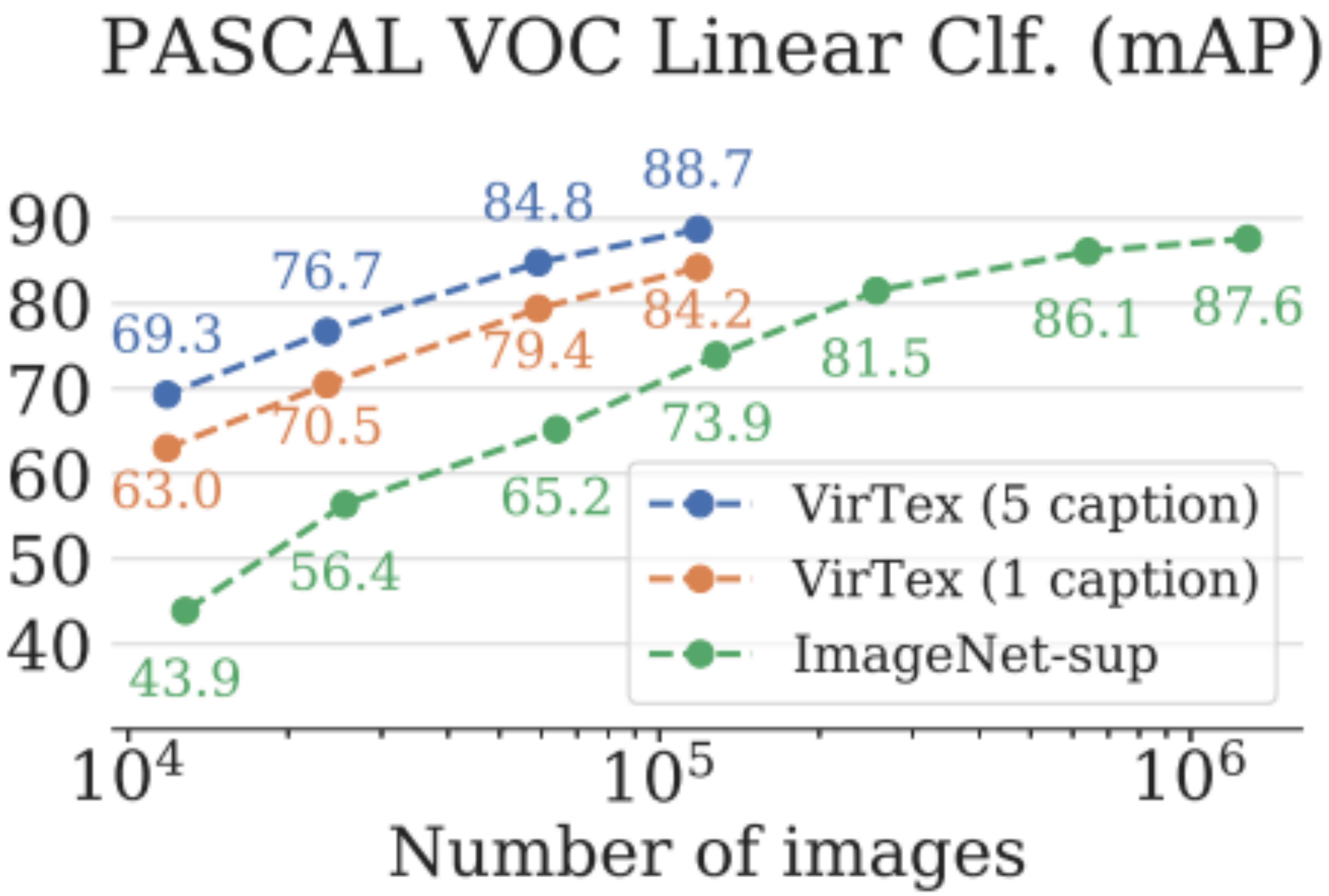
Transformer-based captioning



[Desai and Johnson, VirTex, 2020]

Good source of features





[Desai and Johnson, VirTex, 2020]

VQA: Visual Question Answering

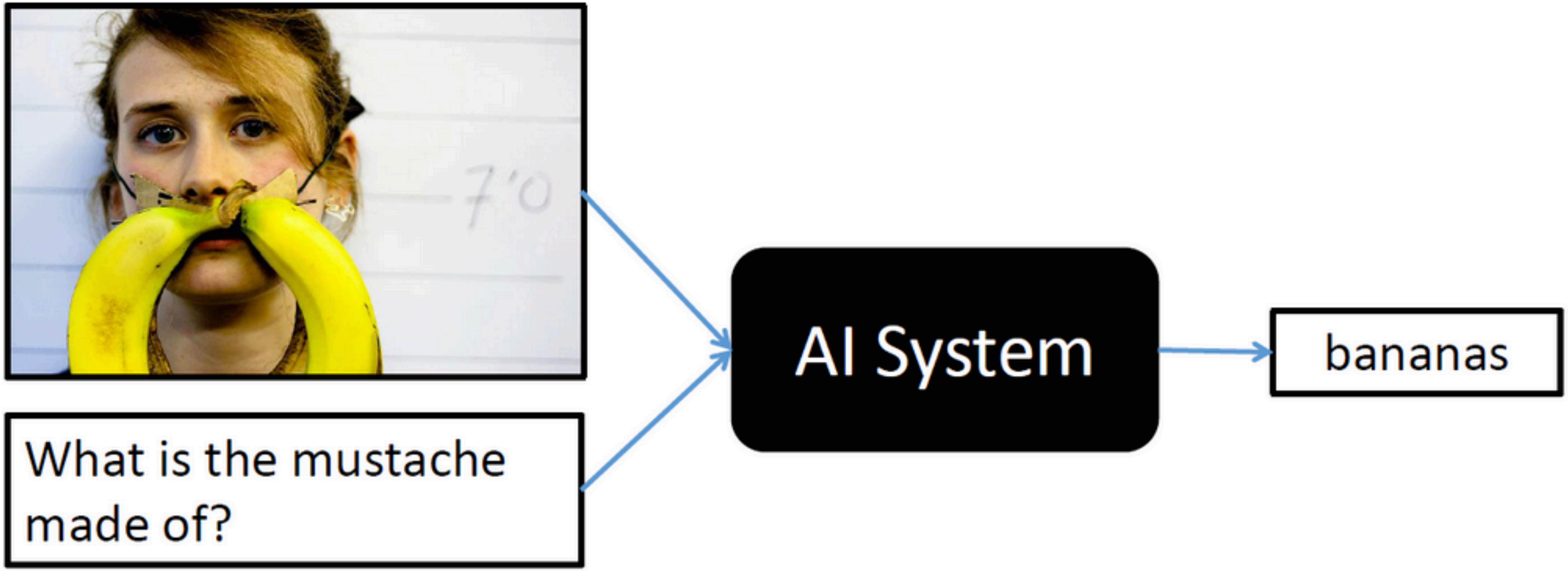
www.visualqa.org

Aishwarya Agrawal*, Jiasen Lu*, Stanislaw Antol*,
Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

Abstract—We propose the task of *free-form* and *open-ended* Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We provide a dataset containing ∼0.25M images, ∼0.76M questions, and ∼10M answers (www.visualqa.org), and discuss the information it provides. Numerous baselines and methods for VQA are provided and compared with human performance.

2016

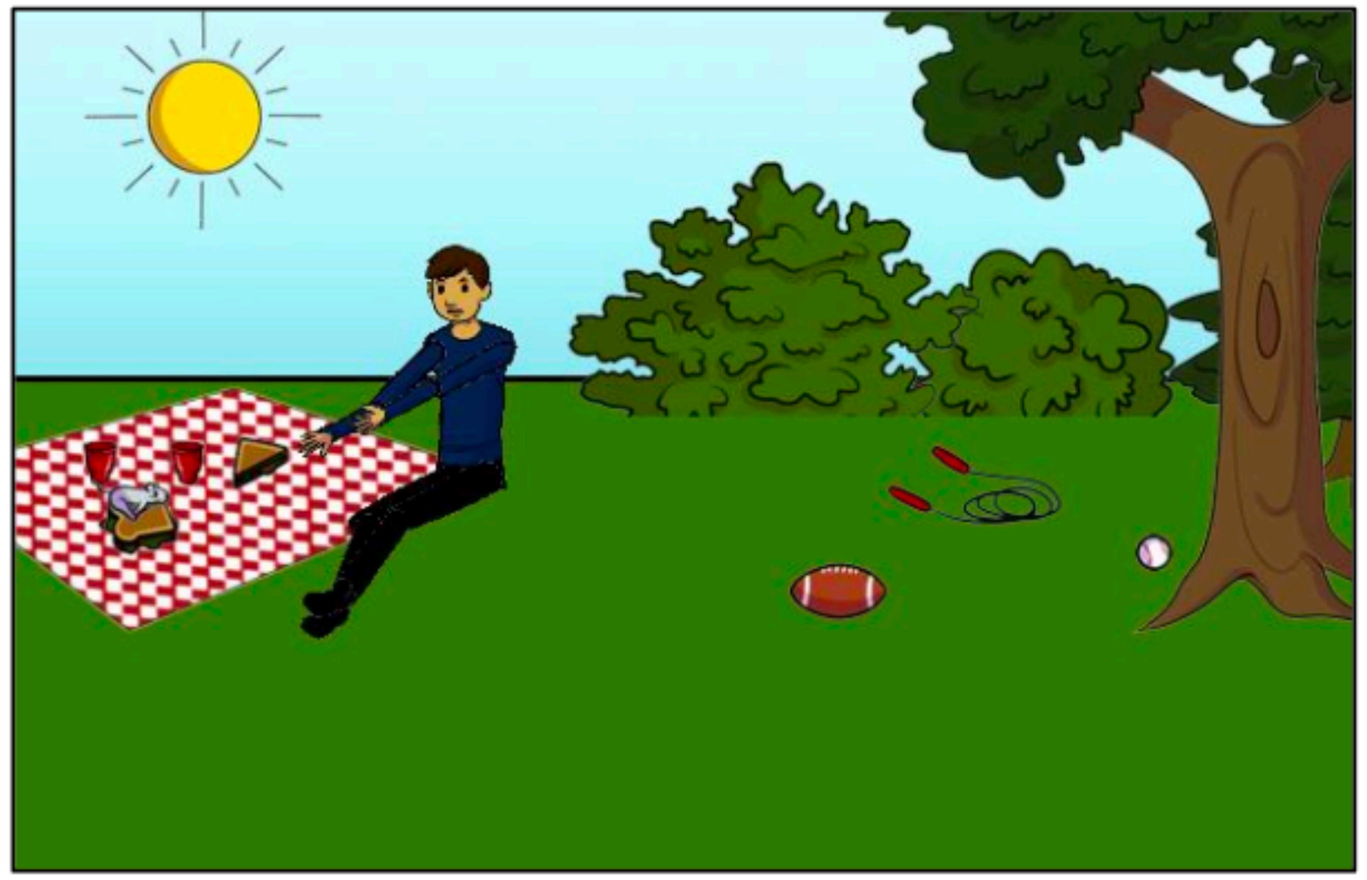
[https://arxiv.org/pdf/1505.00468v6.pdf]



[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?

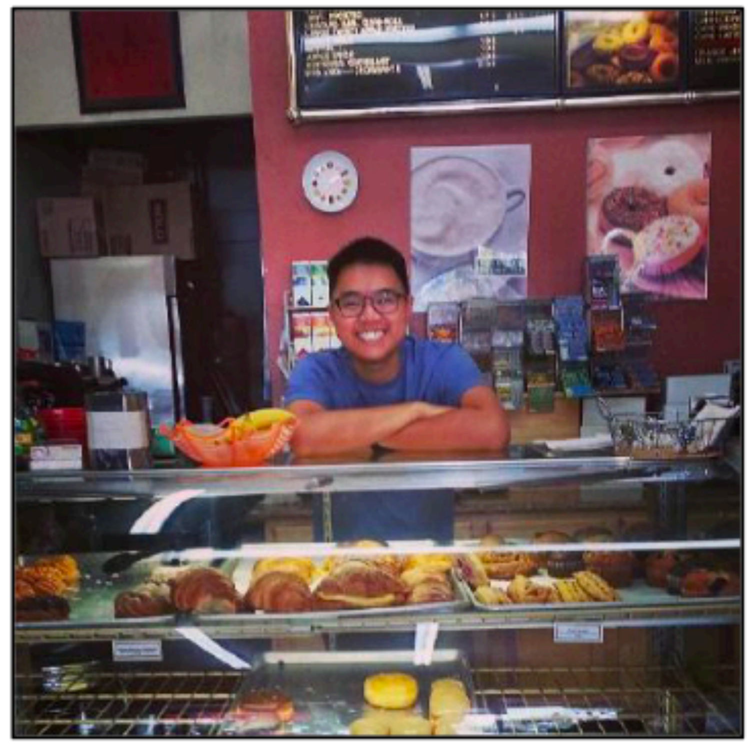
Does it appear to be rainy?
Does this person have 20/20 vision?

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.

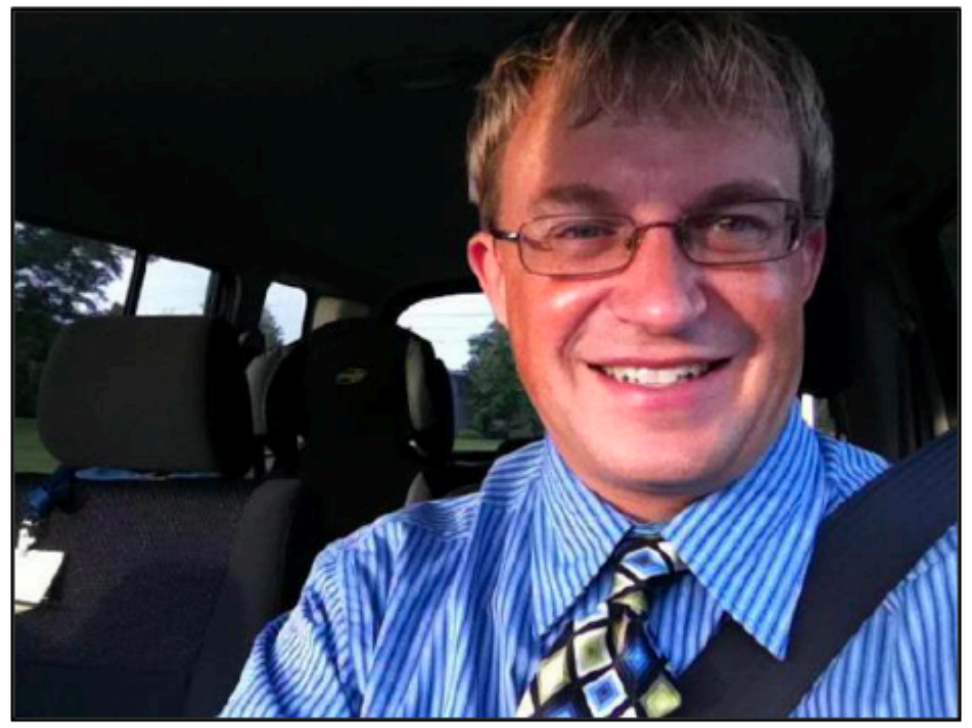
Questions and answers collected with Amazon Mechanical Turk

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

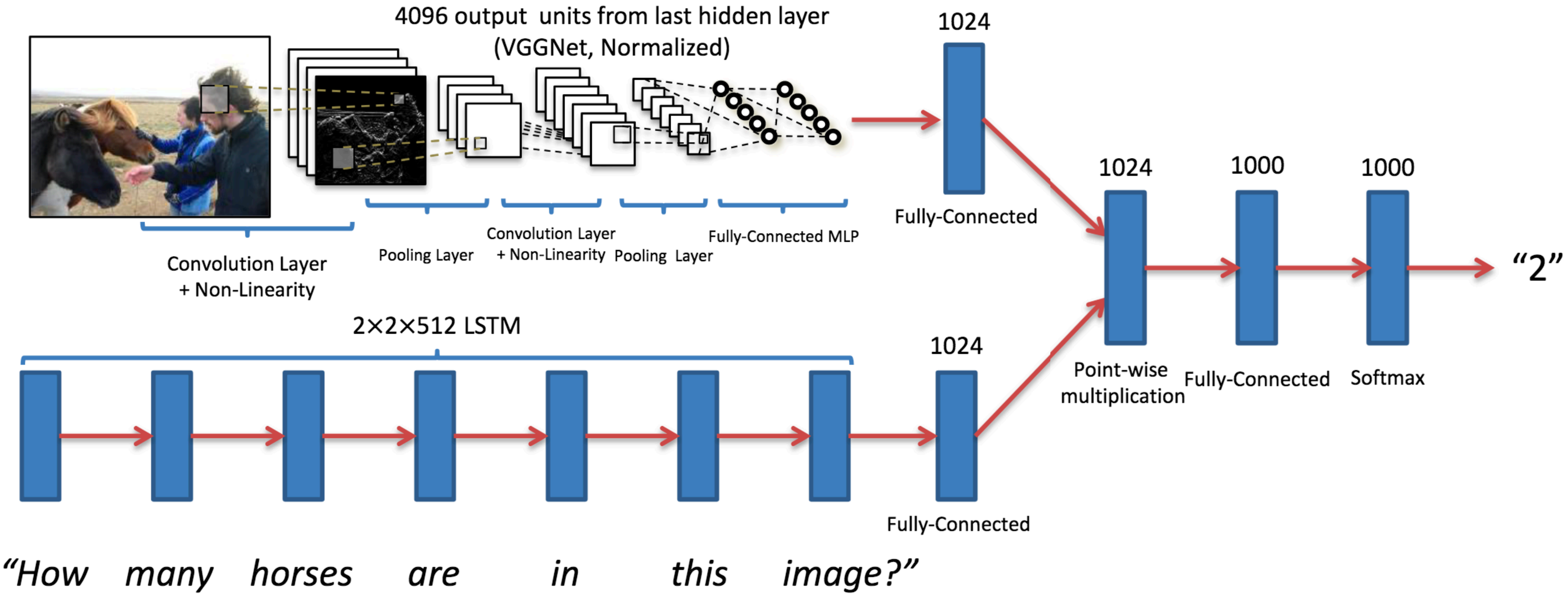
How many pickles are on the plate?	1 1 1	1 1 1
What is the shape of the plate?	circle round round	circle round round

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

Source: Isola, Torralba, Freeman

Architecture



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

[Agrawal et al., "VQA: Visual Question Answering" 2016]

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

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

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

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

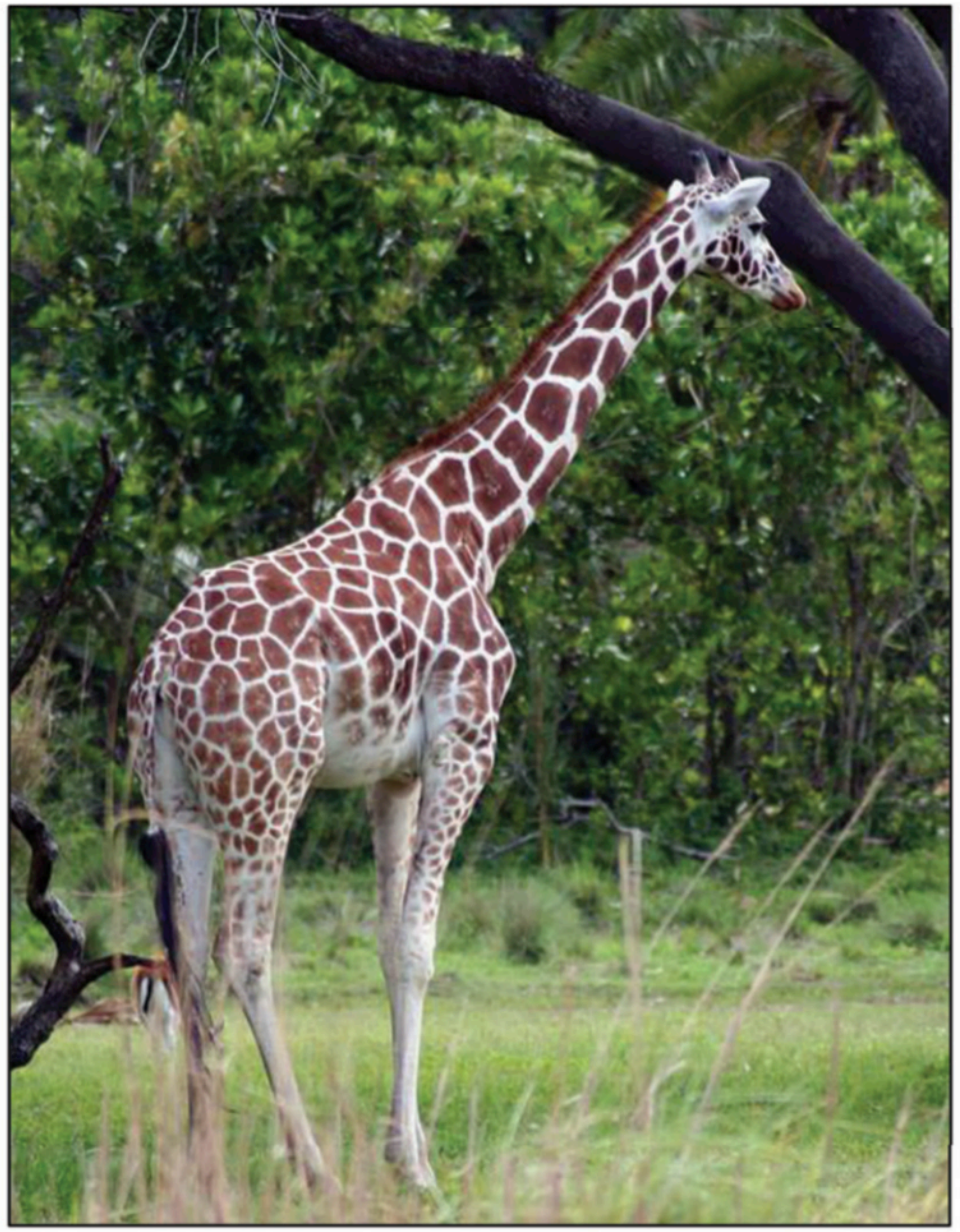
11.343%

6

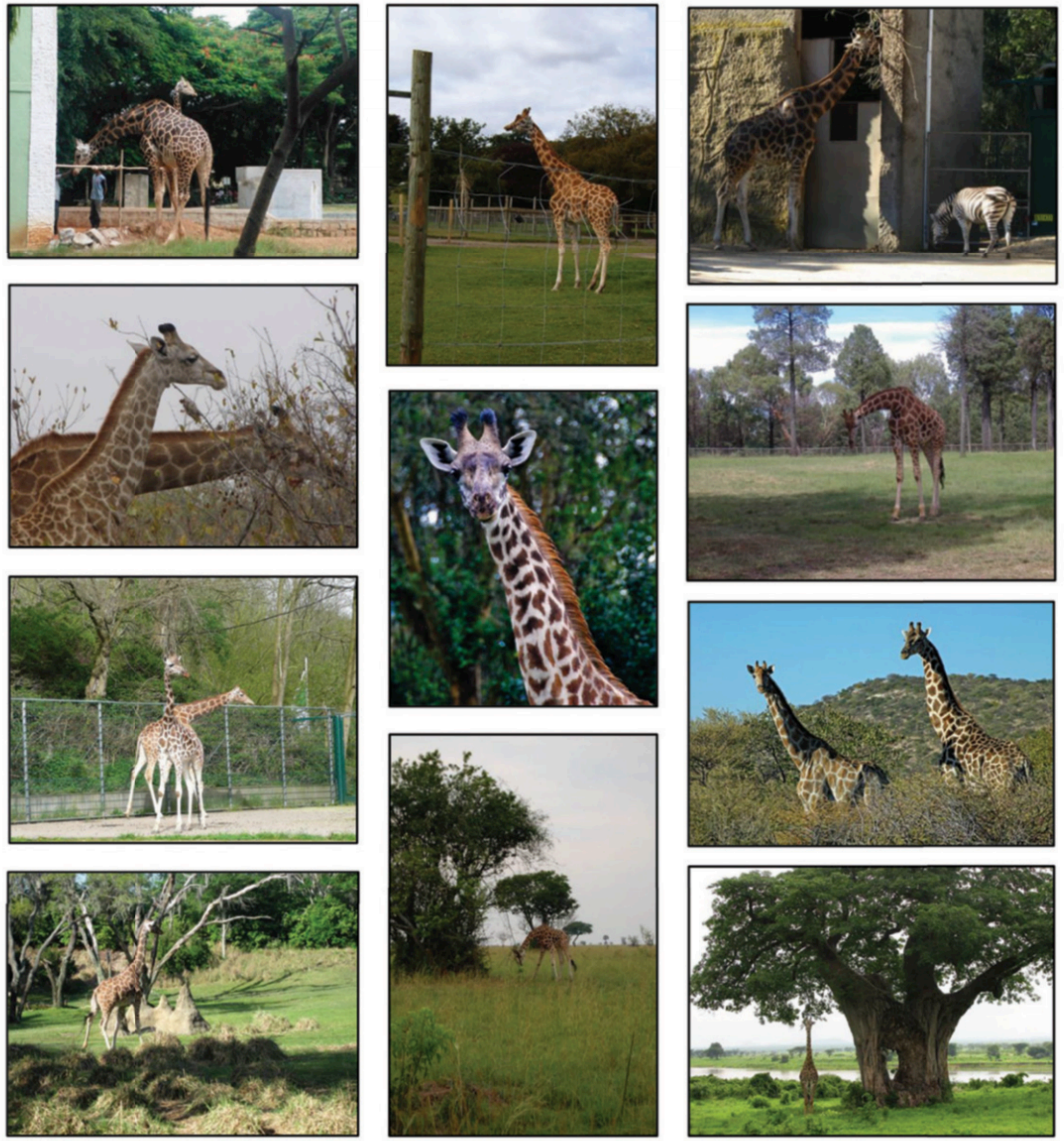
7.806%

What’s going on?

The Giraffe-Tree problem



A giraffe standing in the grass next to a tree.



["Measuring Machine Intelligence Through Visual Question Answering", Zitnick et al., 2016]

Nearest neighbor baseline

Test

Train

Source: L. Zitnick

Nearest Neighbor

A black and white cat sitting in a bathroom sink.

Two zebras and a giraffe in a field.

See mscoco.org for image information

Source: L. Zitnick

Image captioning

A man riding a motorcycle on a beach.

An airplane is parked on the tarmac at an airport.

Source: L. Zitnick

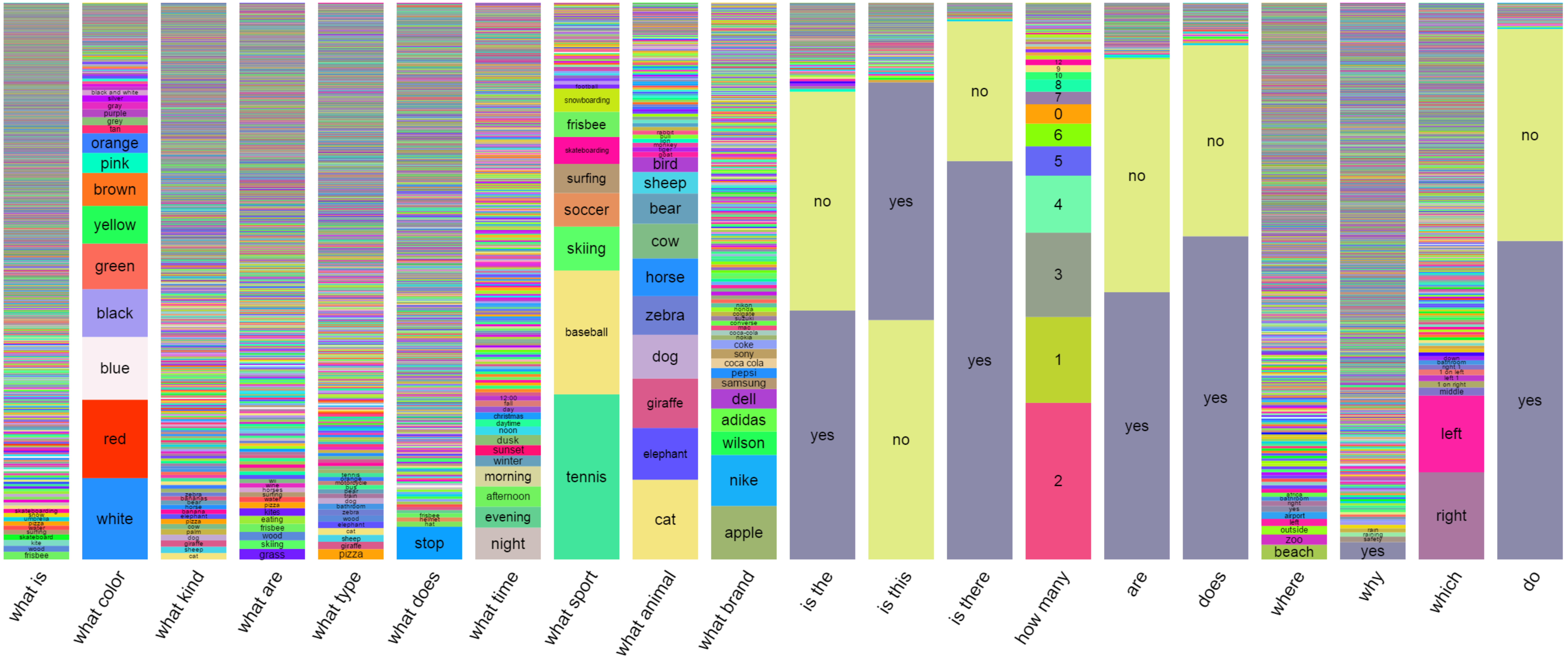
Results

COCO Caption Challenge

	CIDEr-D	Meteor	ROUGE-L	BLEU-4
Google[4]	0.943	0.254	0.53	0.309
MSR Captivator[9]	0.931	0.248	0.526	0.308
m-RNN[15]	0.917	0.242	0.521	0.299
MSR[8]	0.912	0.247	0.519	0.291
Nearest Neighbor[11]	0.886	0.237	0.507	0.280
m-RNN (Baidu/ UCLA)[16]	0.886	0.238	0.524	0.302
Berkeley LRCN[2]	0.869	0.242	0.517	0.277
Human[5]	0.854	0.252	0.484	0.217
Montreal/Toronto[10]	0.85	0.243	0.513	0.268
PicSOM[13]	0.833	0.231	0.505	0.281
MLBL[7]	0.74	0.219	0.499	0.26
ACVT[1]	0.709	0.213	0.483	0.246
NeuralTalk[12]	0.674	0.21	0.475	0.224
Tsinghua Bigeye[14]	0.673	0.207	0.49	0.241
MIL[6]	0.666	0.214	0.468	0.216
Brno University[3]	0.517	0.195	0.403	0.134

Source: L. Zitnick

Visual Question Answering Dataset



Source: L. Zitnick

Aside: biases in data collection

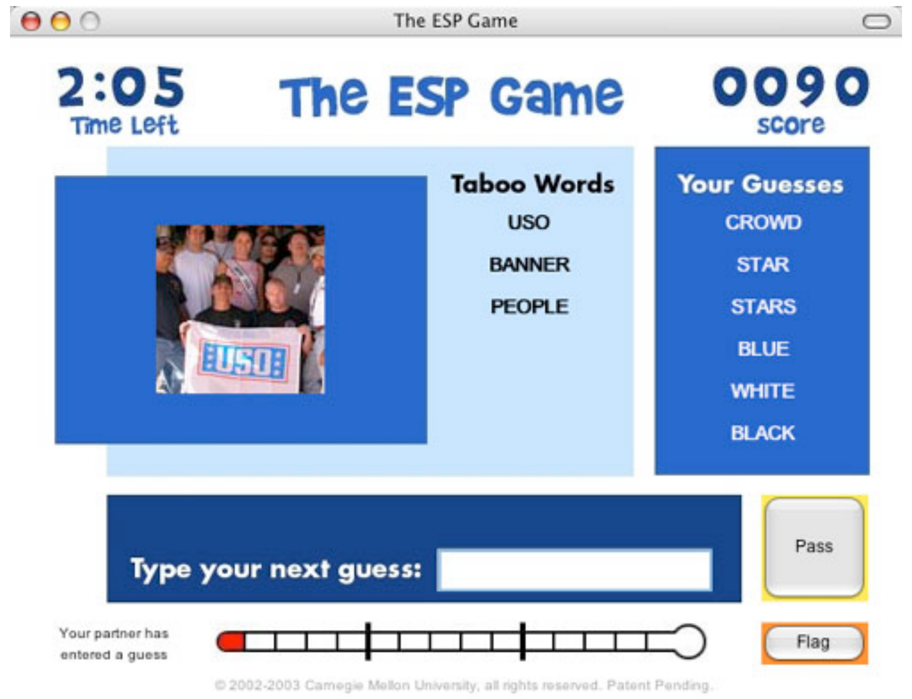
Source: Isola, Torralba, Freeman

Getting more humans in the annotation loop

Labeling to get a Ph.D.

Labeling for fun

Luis Von Ahn and Laura Dabbish 2004

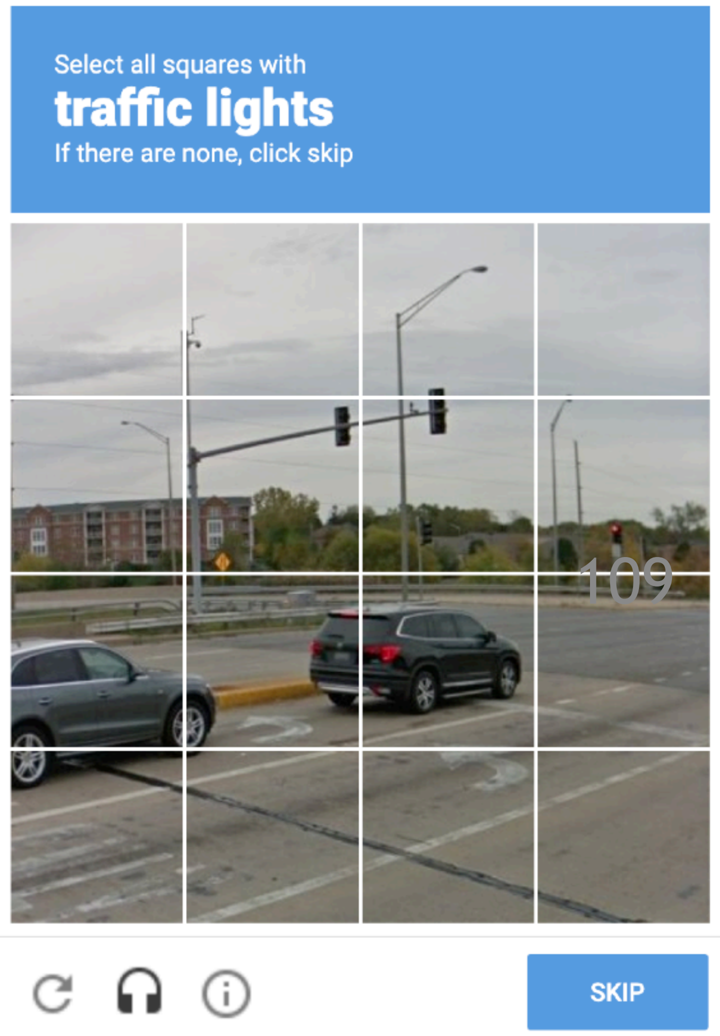


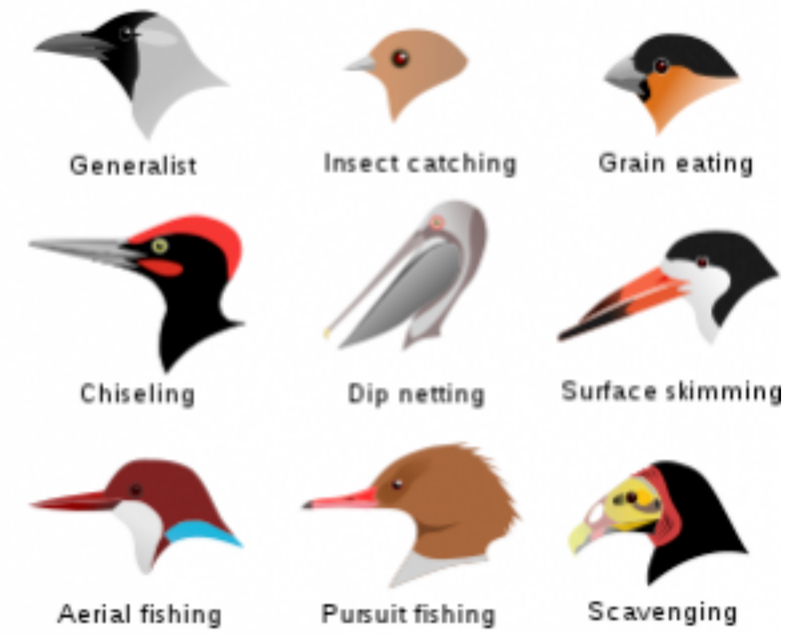
Labeling for money

(Sorokin, Forsyth, 2008)

amazonmechanical turk
beta
Artificial Artificial Intelligence

Labeling to prove you're human





Labeling because it gives you added value

Visipedia

(Belongie, Perona, et al)

Source: Isola, Torralba, Freeman

Beware of the human in your loop

- What do you know about them?
- Will they do the work you pay for?

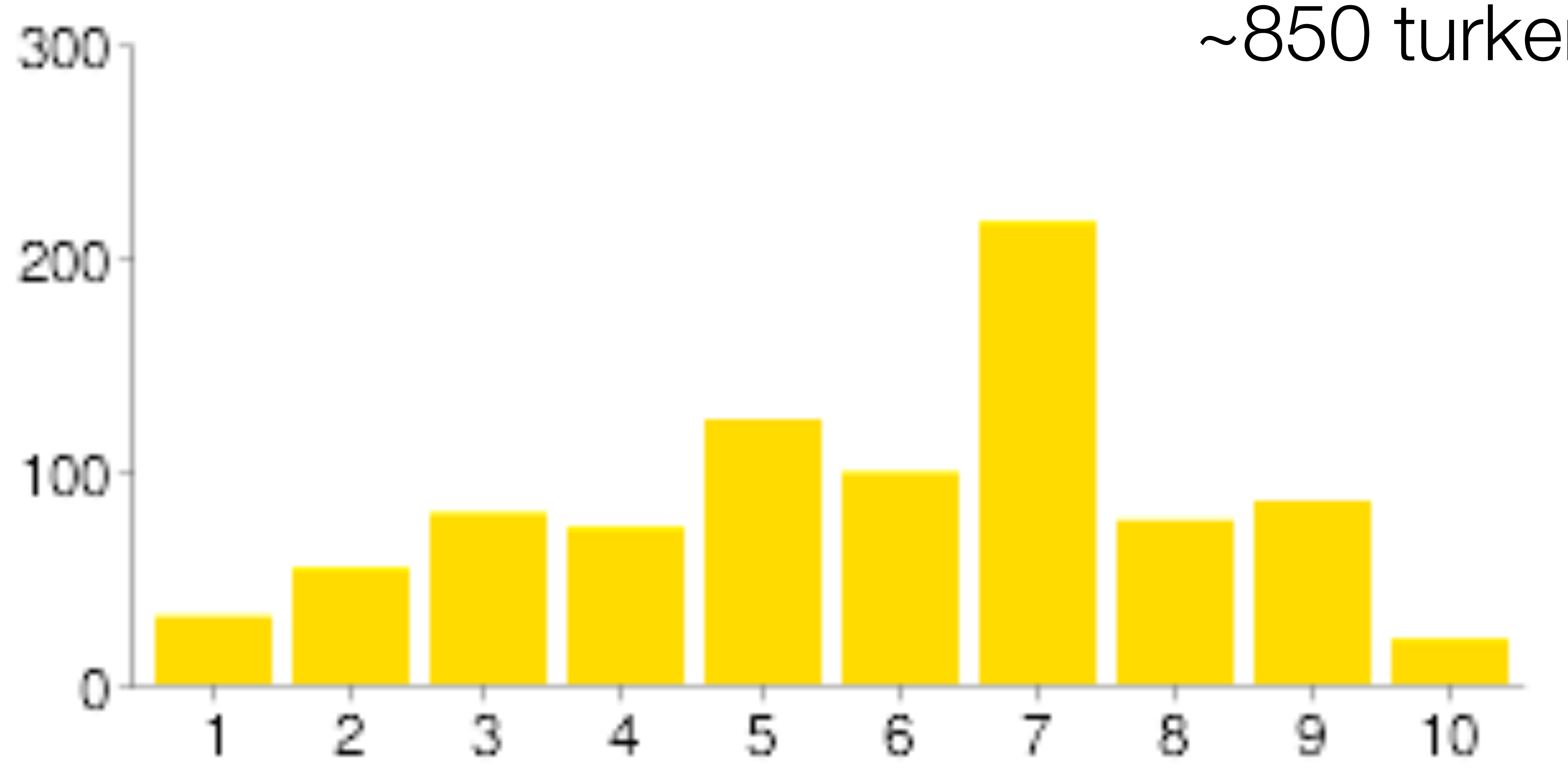
Let’s check a few simple experiments

Source: Isola, Torralba, Freeman

People have biases…

Turkers were offered 1 cent to pick a number from 1 to 10.

~850 turkers

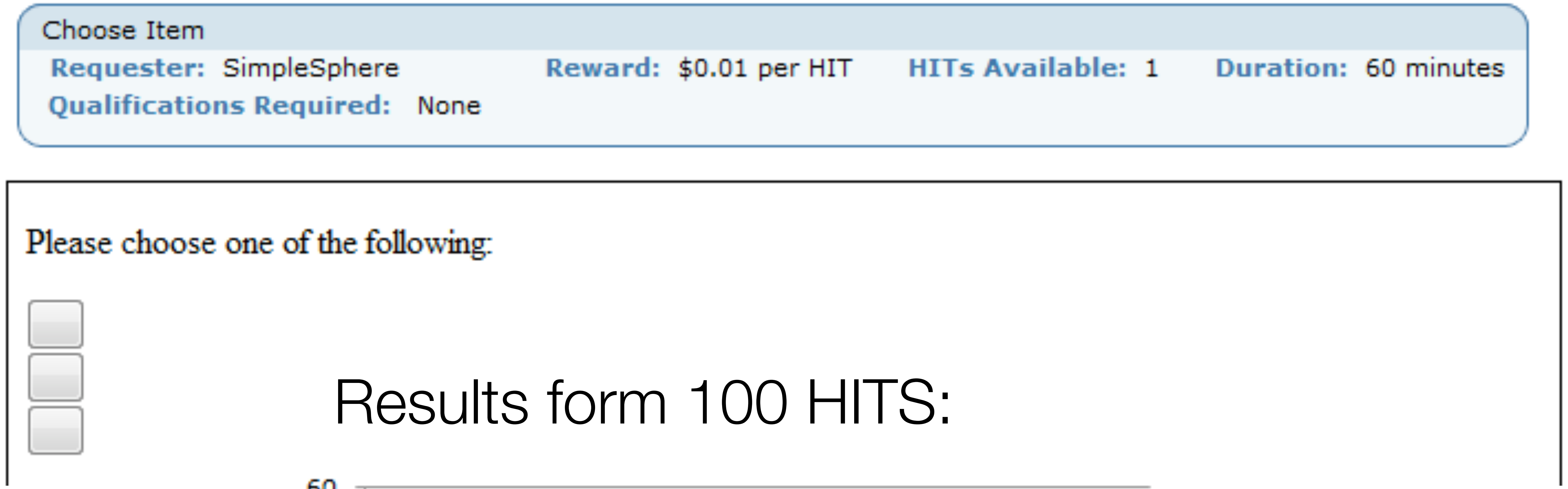


Experiment by Greg Little

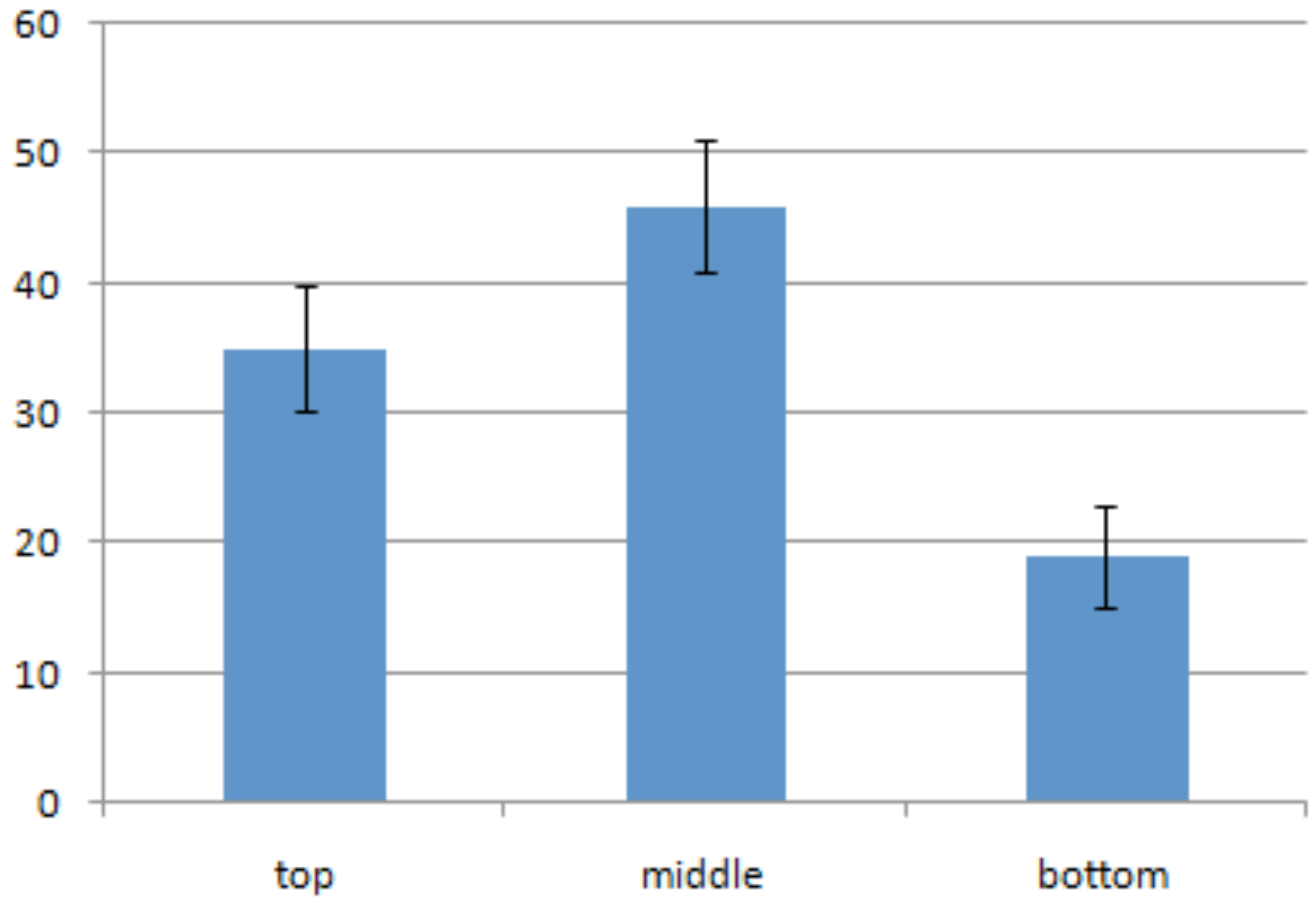
From http://groups.csail.mit.edu/uid/deneme/

Source: Isola, Torralba, Freeman

Do humans have consistent biases?



Results form 100 HITS:



Experiment by Greg Little

From http://groups.csail.mit.edu/uid/deneme/

Source: Isola, Torralba, Freeman

Are humans reliable even in simple tasks?

Choose the given item.

Requester: SimpleSphere **Reward:** $0.01 per HIT **HITs Available:** 1 **Duration:** 60 minutes

Qualifications Required: None

Please click button B:

B

C

A

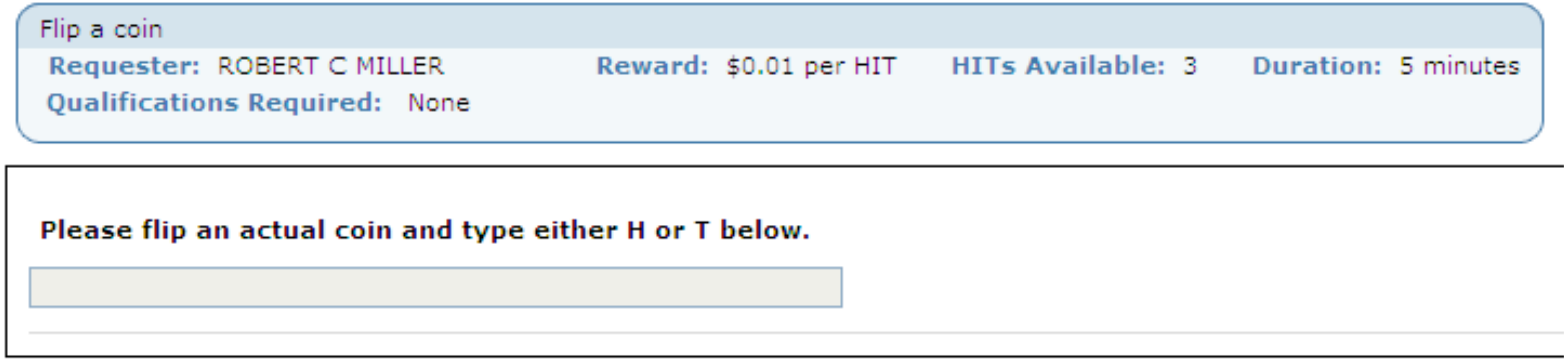
Results of 100 HITS:
A: 2
B: 96
C: 2

Experiment by Greg Little

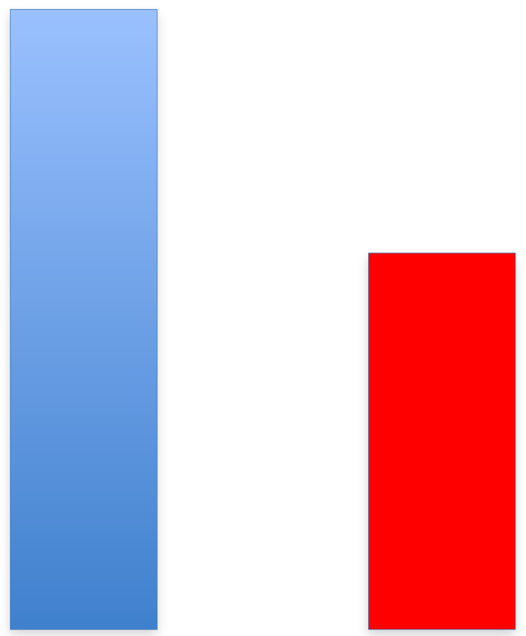
From http://groups.csail.mit.edu/uid/deneme/

Source: Isola, Torralba, Freeman

Do humans do what you ask for?

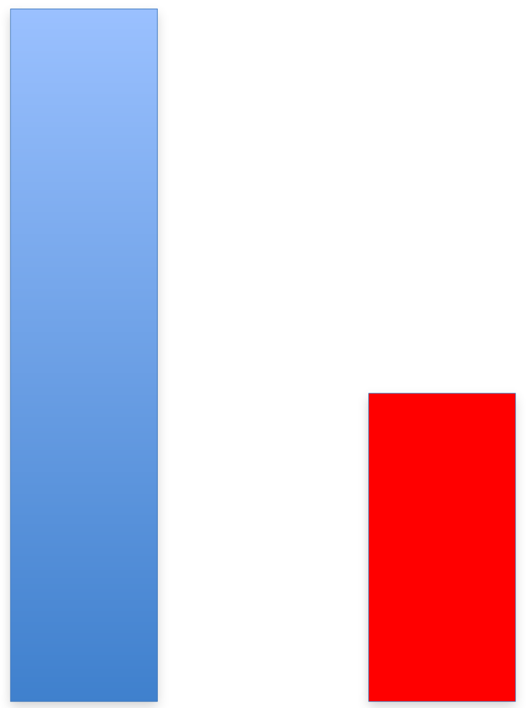


After 50 HITS:



31 heads, 19 tails

And 50 more:



34 heads, 16 tails

Experiment by Rob Miller

From http://groups.csail.mit.edu/uid/deneme/

Source: Isola, Torralba, Freeman

So we can sometimes collect good training data.

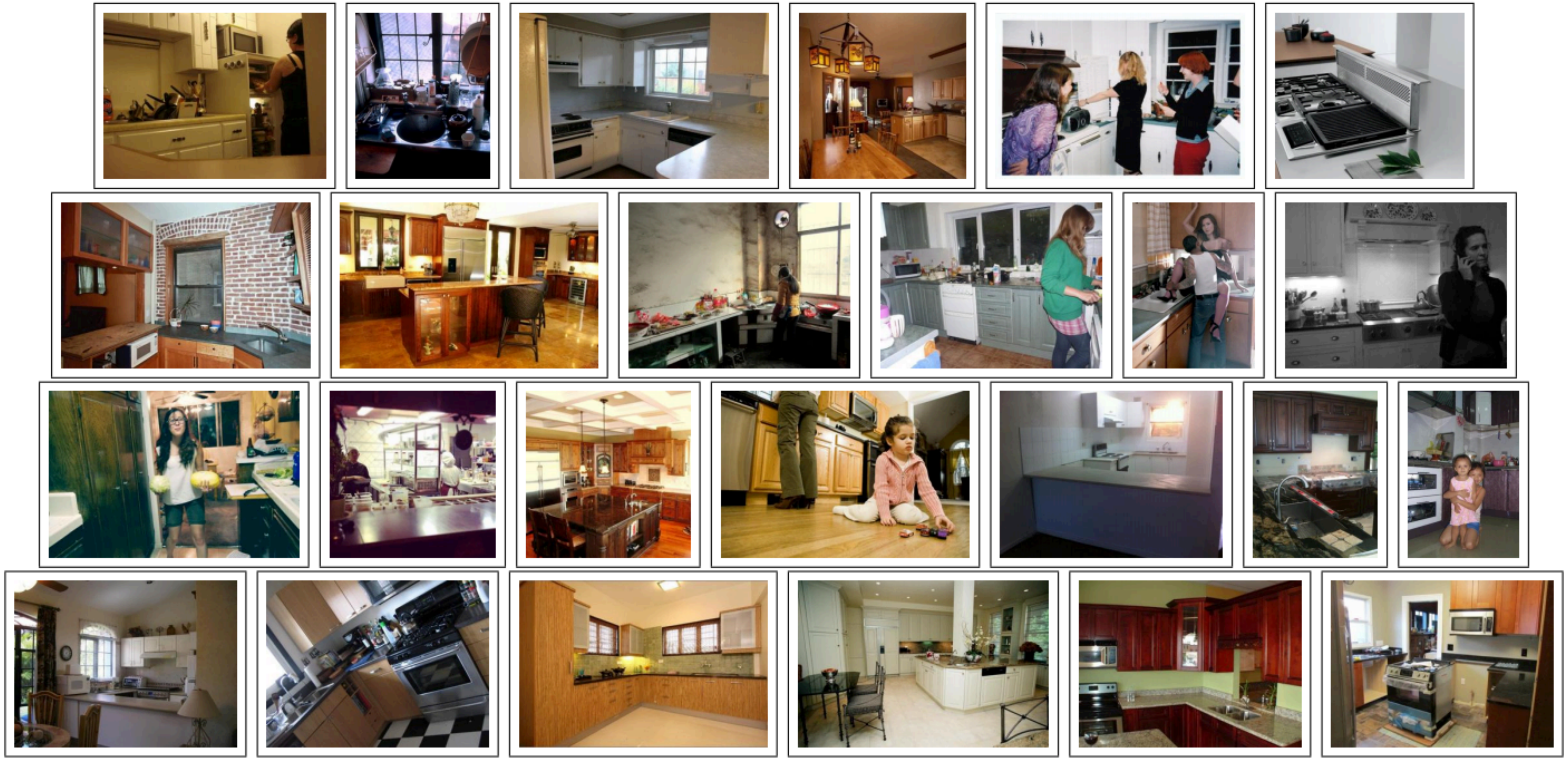
But suppose we messed up. Our test setting doesn't look like the training data!

How can we bridge the domain gap?

Source: Isola, Torralba, Freeman

Finding more representative images

Places365 Kitchen



[Fouhey et al., "From Lifestyle Vlogs to Everyday Actions", 2017]

Finding more representative images

VLOG Kitchen

[Fouhey et al., "From Lifestyle Vlogs to Everyday Actions", 2017]

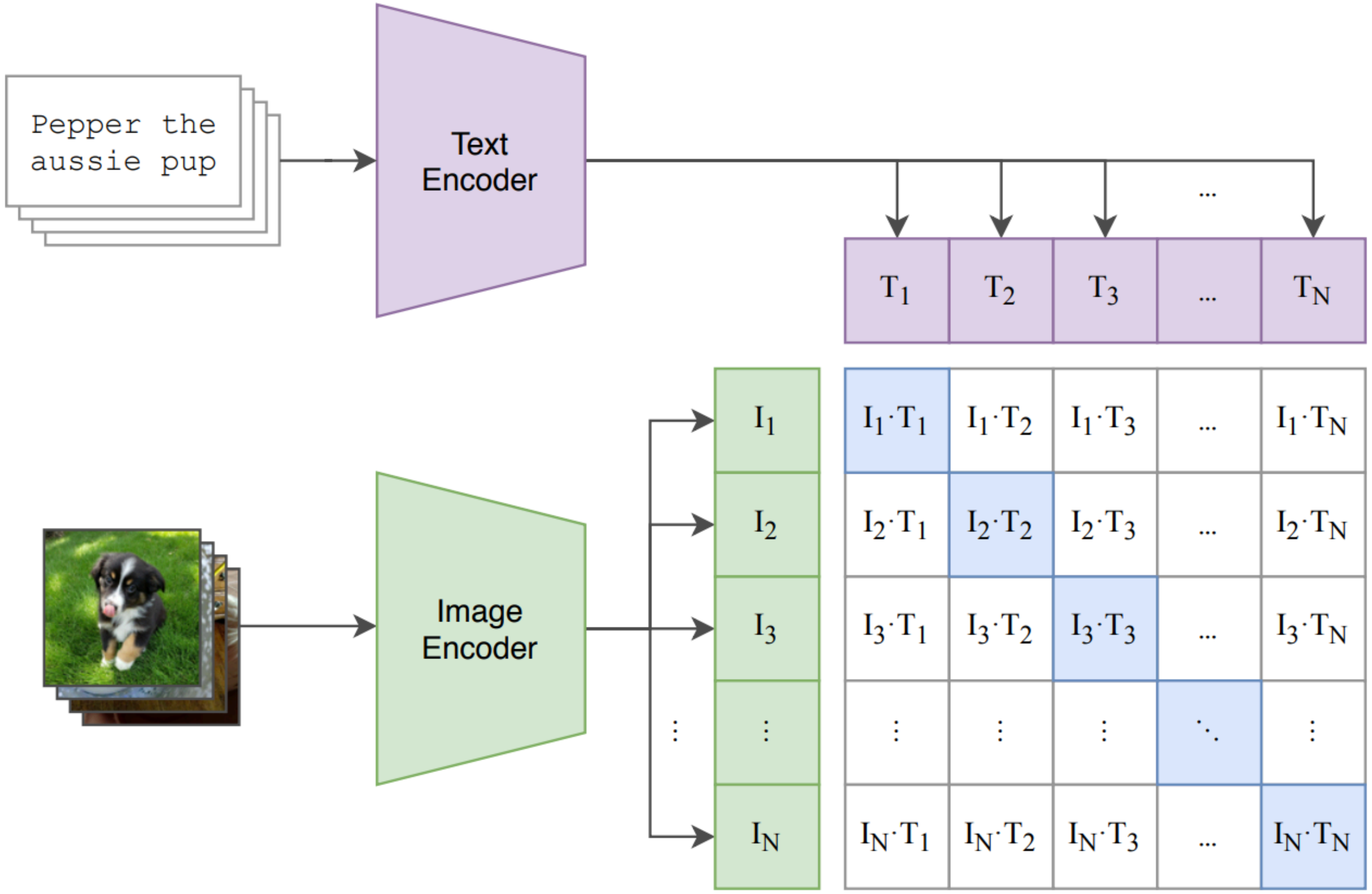
Name that dataset game

Caltech101 Tiny LabelMe 15 Scenes
MSRC Corel COIL–100 Caltech256
UIUC PASCAL 07 ImageNet SUN09

[Torralba and Efros, “An unbiased look at dataset bias,” 2011]

Some recent directions

Learning representations from language



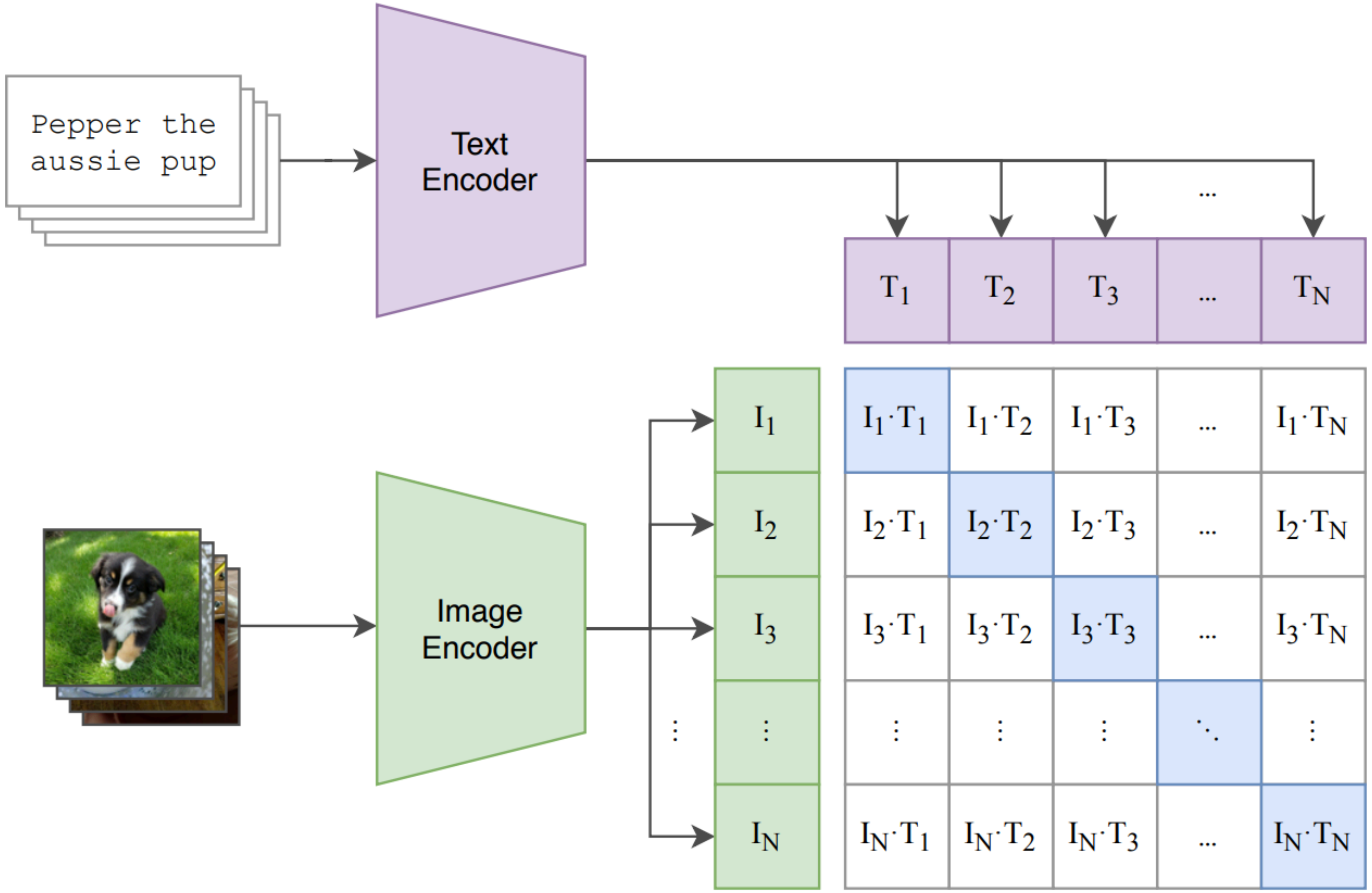
Contrastive learning

maximize:

$$\log\left(\frac{\exp(\mathbf{I_i}\cdot\mathbf{T_i})}{\sum_j \exp(\mathbf{I_i}\cdot\mathbf{T_j})}\right)$$

[Radford et al., "CLIP" , 2021]

Learning representations from language



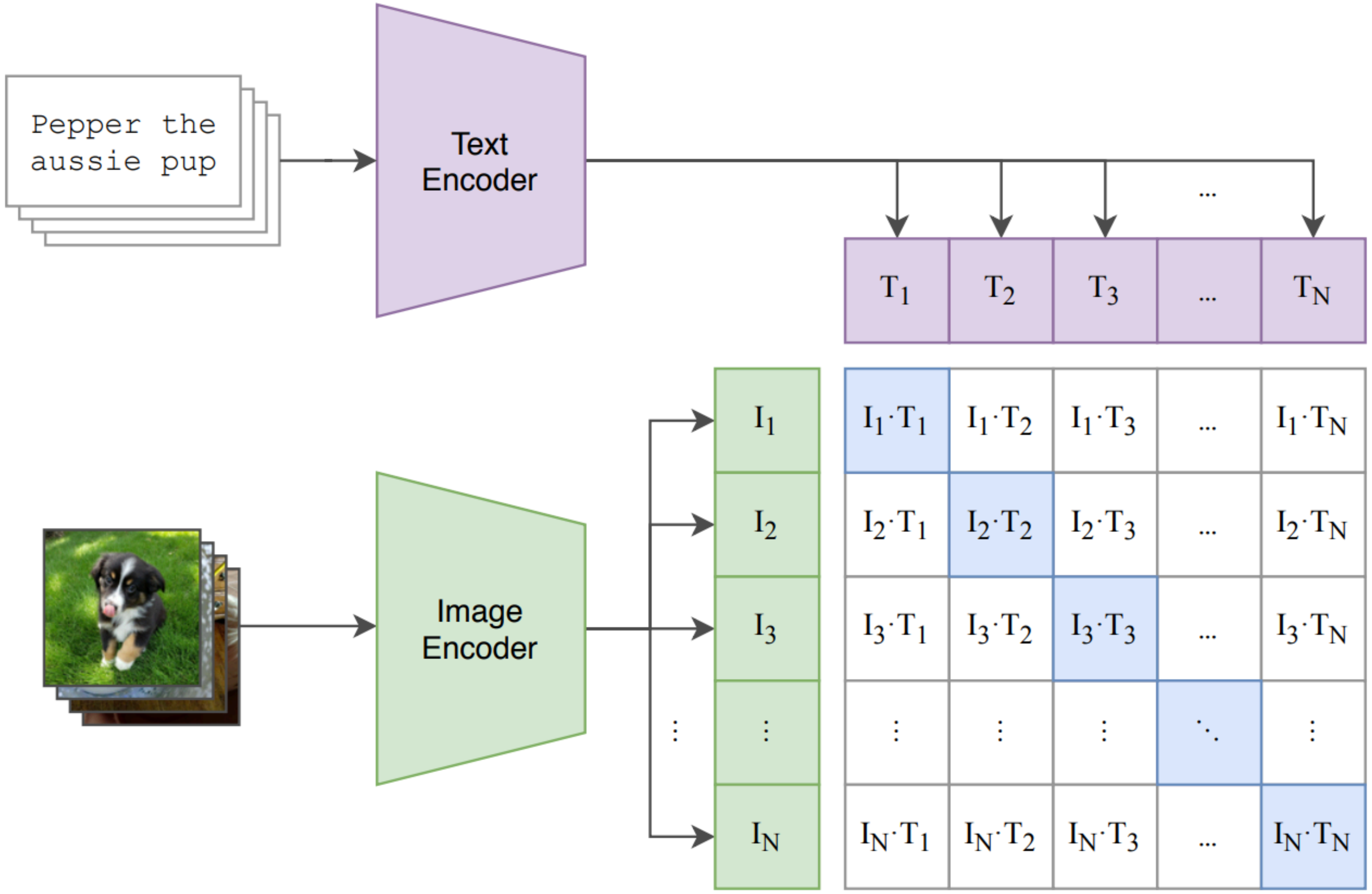
Contrastive learning

maximize:

$$\log\left(\frac{\exp(\mathbf{I_i} \cdot \mathbf{T_i})}{\sum_j \exp(\mathbf{I_i} \cdot \mathbf{T_j})}\right)$$

[Radford et al., "CLIP" , 2021]

Learning representations from language



Contrastive learning

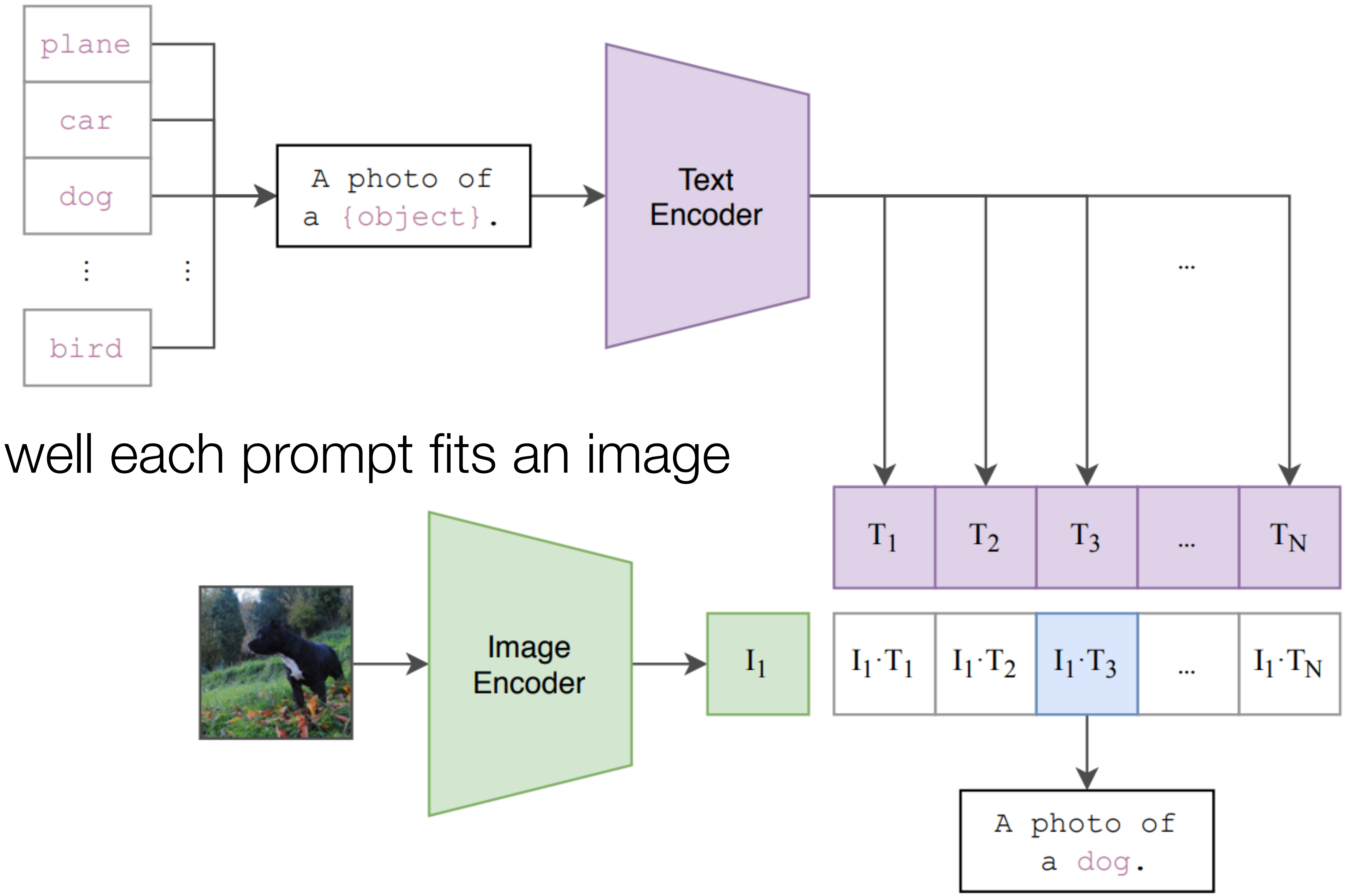
$$\log\left(\frac{\exp(\mathbf{I_i}\cdot\mathbf{T_i})}{\sum_j \exp(\mathbf{I_i}\cdot\mathbf{T_j})}\right)$$

[Radford et al., ”CLIP” , 2021]

“Zero-shot” classification

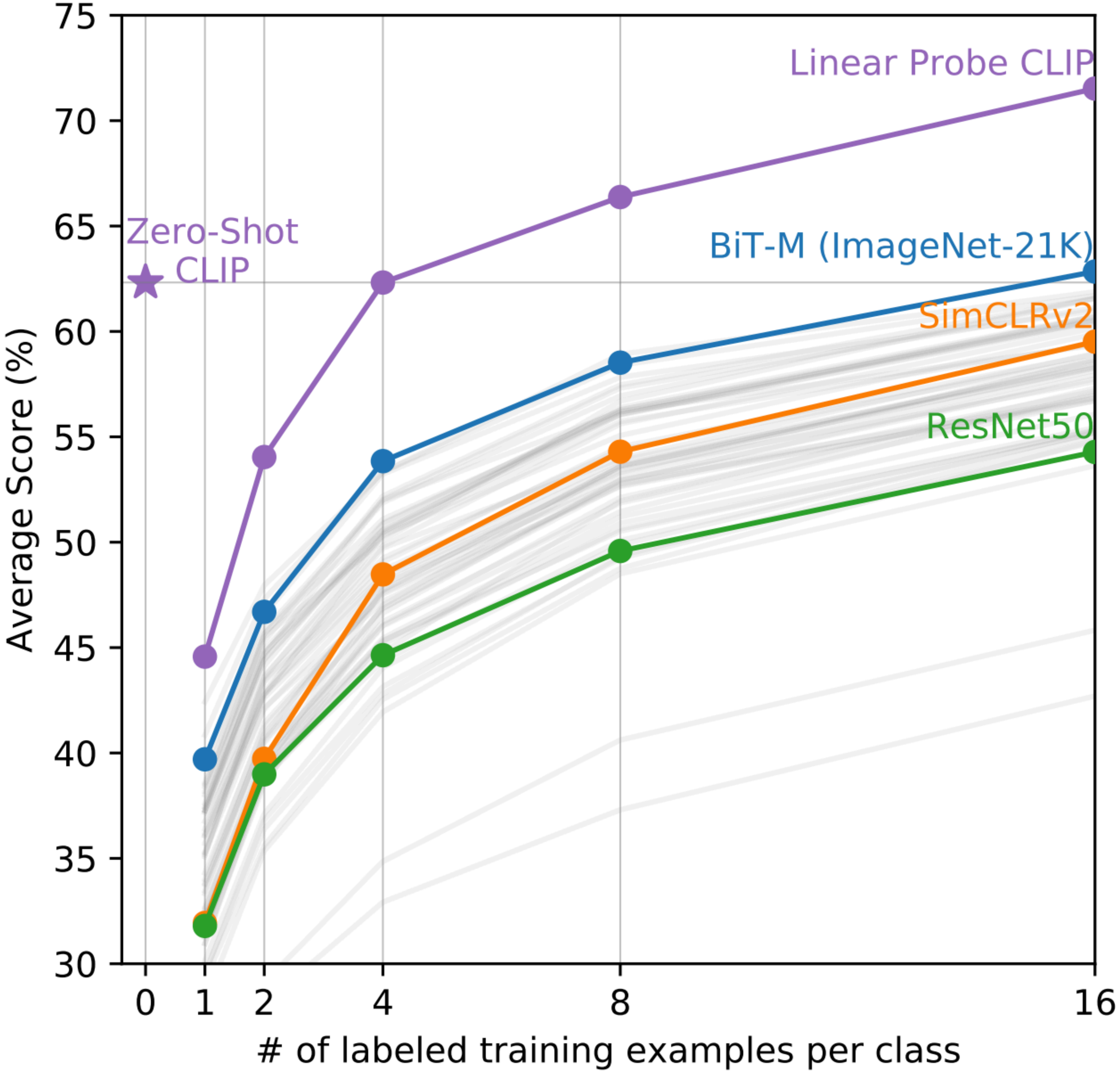
(1) Create classifier from label text

(2) Test how well each prompt fits an image



[Radford et al., ”CLIP” , 2021]

“Zero-shot” classification



[Radford et al., ”CLIP” , 2021]

“Zero-shot” classification

FOOD101

guacamole (90.1%) Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

× a photo of **ceviche**, a type of food.

× a photo of **edamame**, a type of food.

× a photo of **tuna tartare**, a type of food.

× a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397

✓ a photo of a **television studio**.

× a photo of a **podium indoor**.

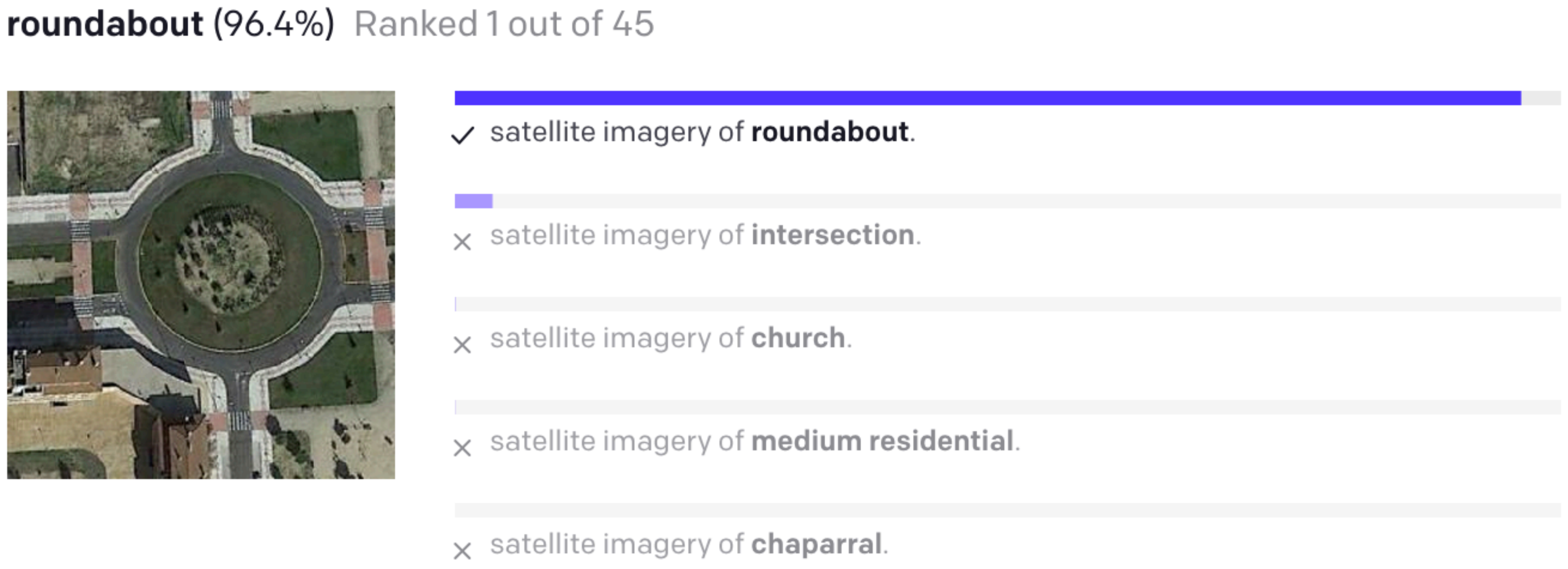
× a photo of a **conference room**.

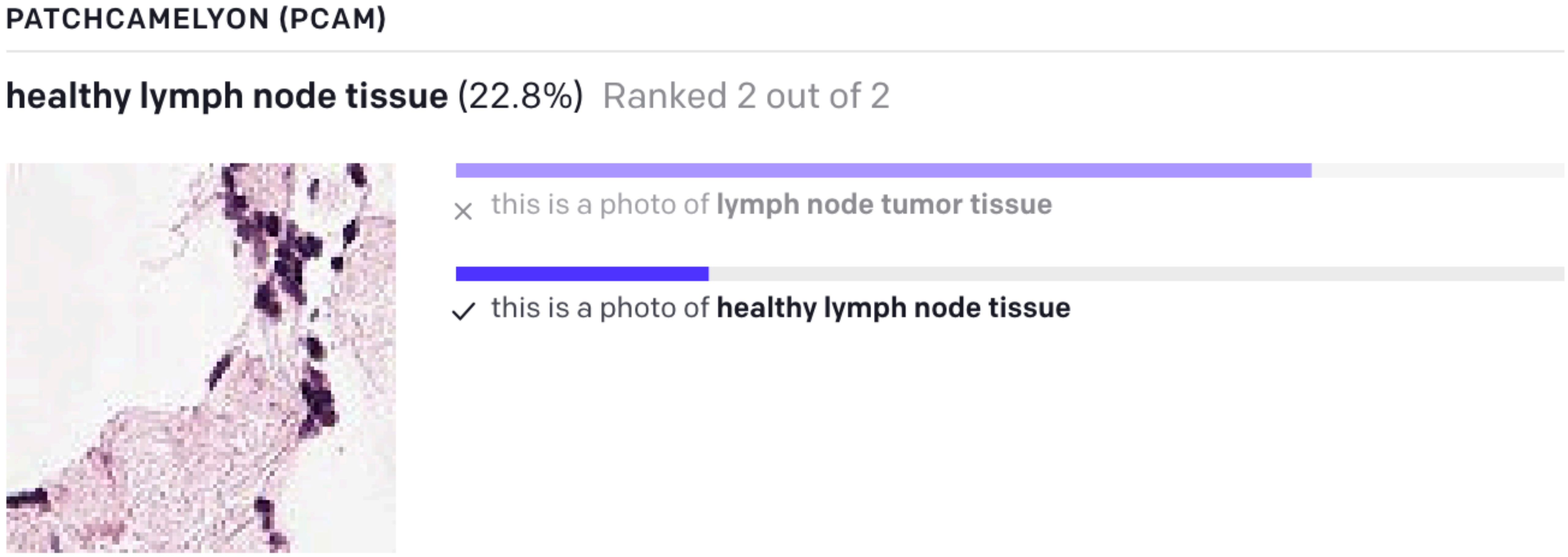
× a photo of a **lecture room**.

× a photo of a **control room**.

[Radford et al., ”CLIP” , 2021]

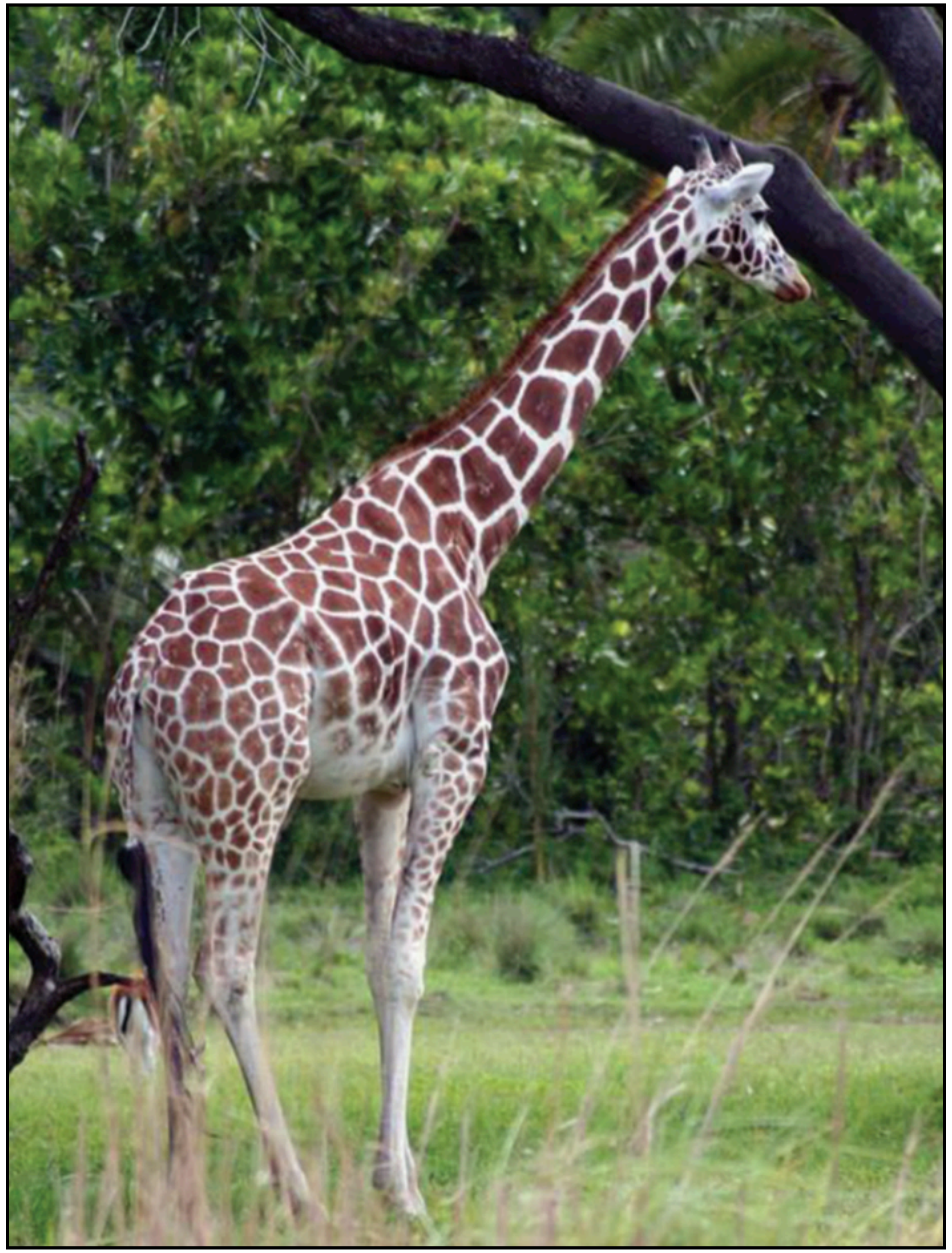
“Zero-shot” classification

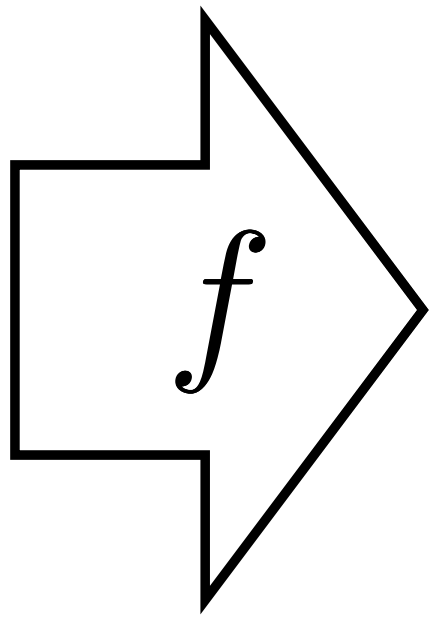




[Radford et al., ”CLIP” , 2021]

Image-to-text



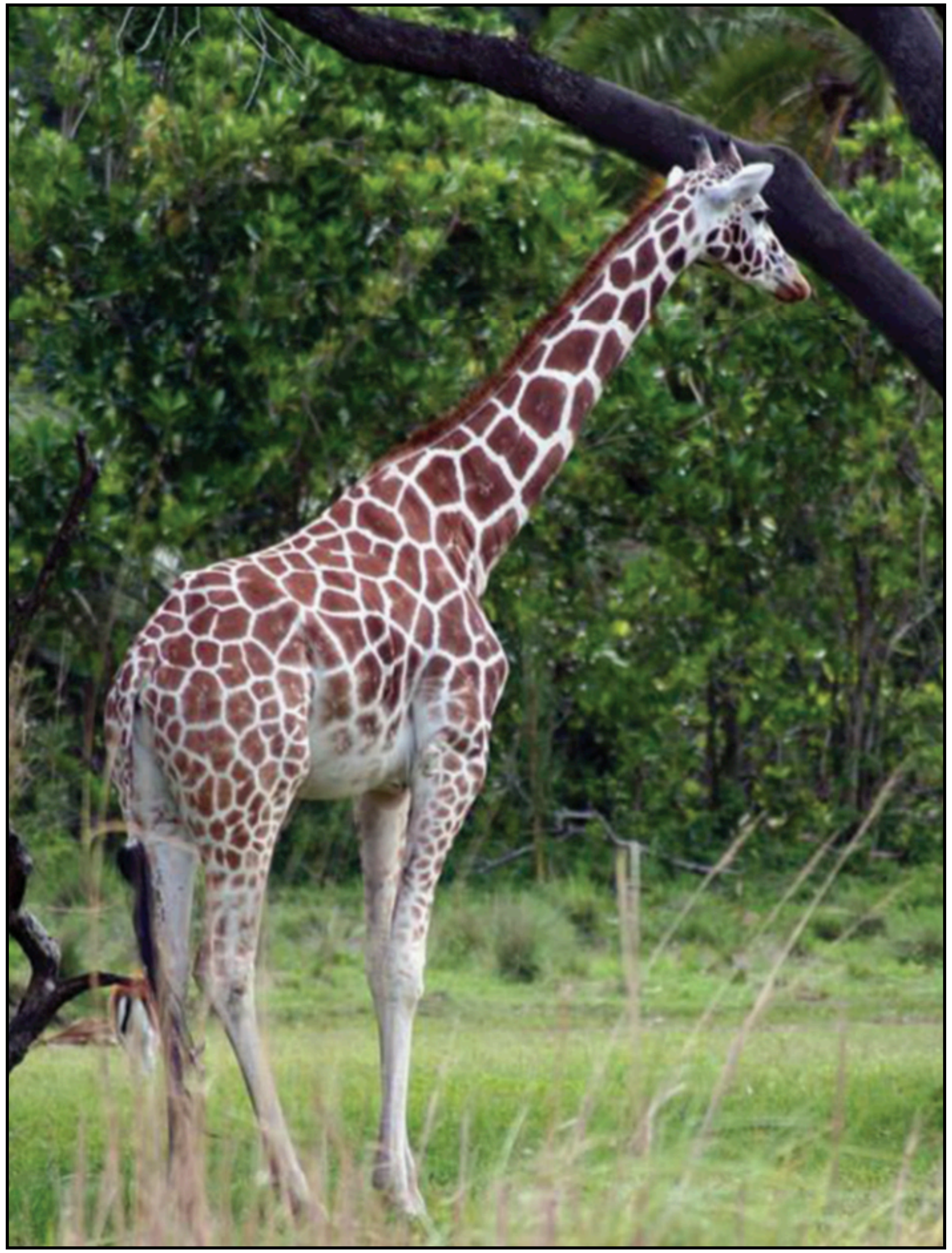


"A giraffe standing in the grass next to a tree"

Text-to-image

"A giraffe standing in the grass next to a tree"

f



Text-to-image

(a) a tapir made of accordion. a tapir with the texture of an accordion.

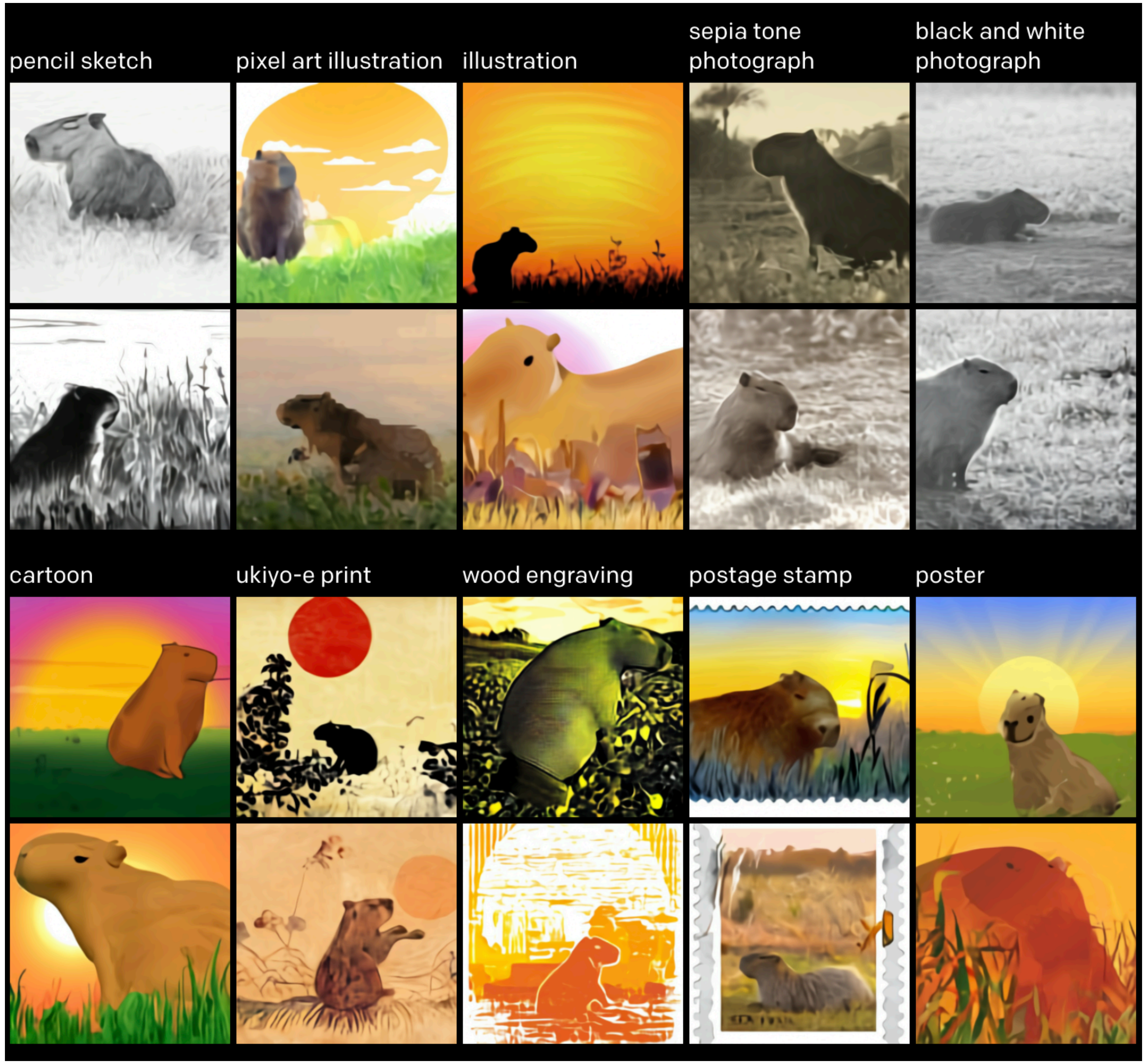
(b) an illustration of a baby hedgehog in a christmas sweater walking a dog

(c) a neon sign that reads “backprop”. a neon sign that reads “backprop”. backprop neon sign

(d) the exact same cat on the top as a sketch on the bottom

[Ramesh et al., “Zero-Shot Text-to-Image Generation”, 2021]

Text-to-image



[Ramesh et al., “Zero-Shot Text-to-Image Generation”, 2021]

Diffusion text-to-image synthesis

a teddy bear on a skateboard in times square

A photo of Michelangelo's sculpture of David wearing headphones djing

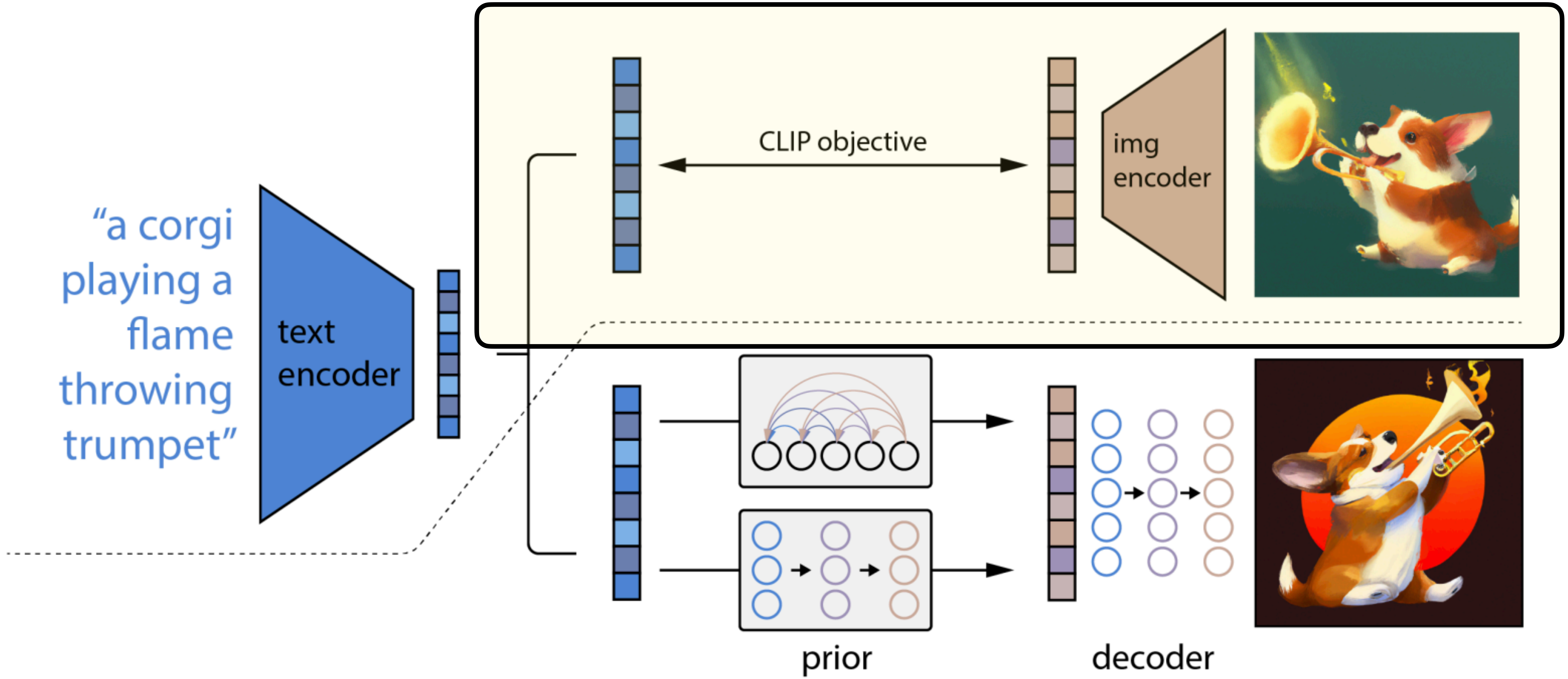
"A sea otter with a pearl earring" by Johannes Vermeer

3D render of a cute tropical fish in an aquarium on a dark blue background, digital art

["DALL-E 2", Ramesh et al., 2022]

Diffusion text-to-image synthesis

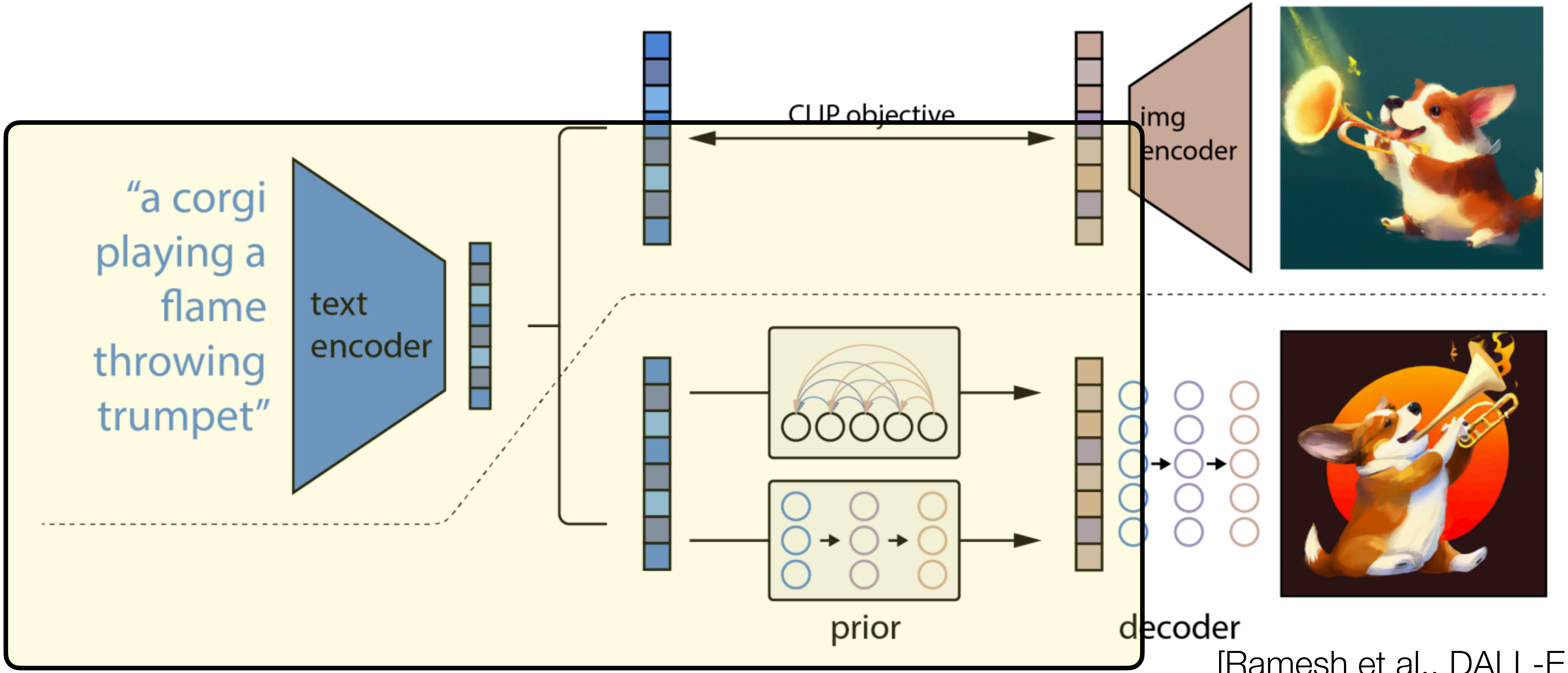
1. Train CLIP



[Ramesh et al., DALL-E 2, 2021]

Diffusion text-to-image synthesis

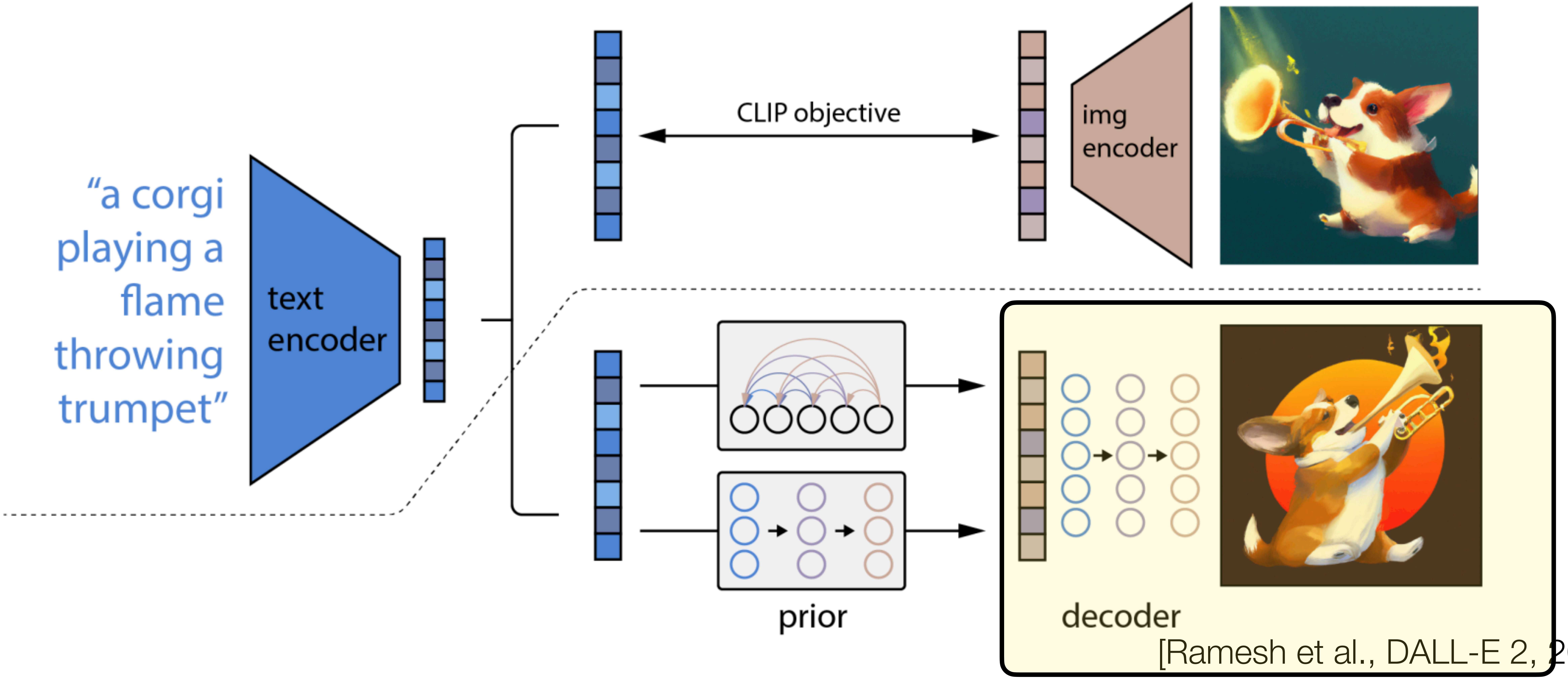
2. Estimate image embedding from text embedding



[Ramesh et al., DALL-E 2, 2021]

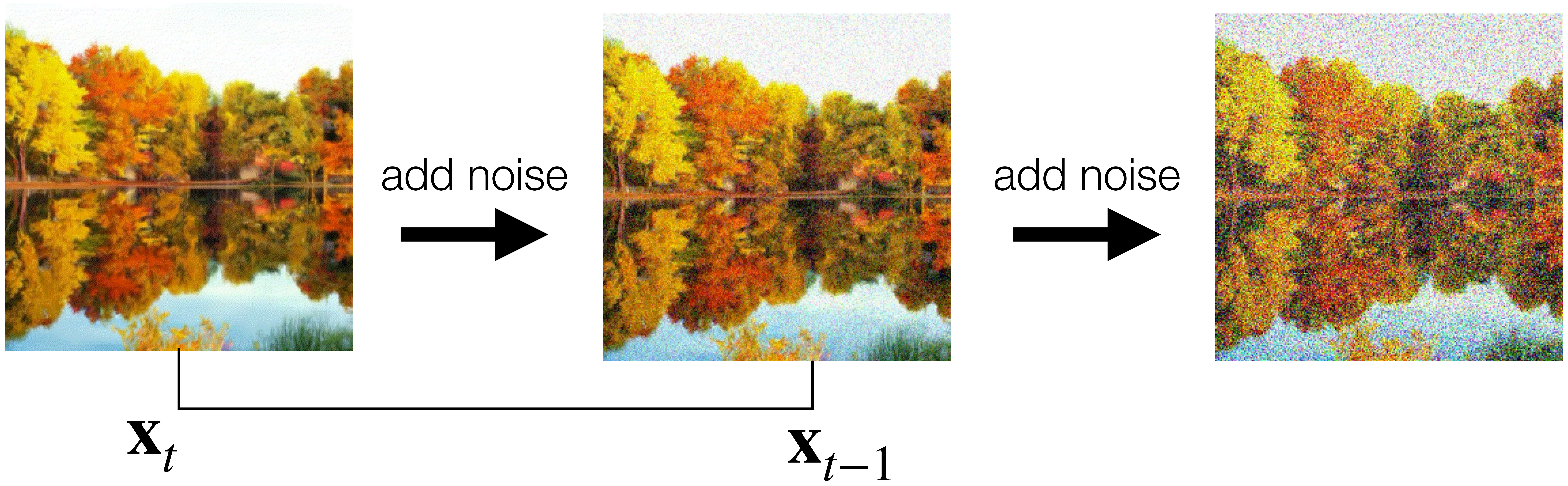
Diffusion text-to-image synthesis

3. Conditional model



[Ramesh et al., DALL-E 2, 2021]

Conditional diffusion



Basic idea

- Unconditional diffusion: predict noise at step t with neural net: $\epsilon_\theta(\mathbf{x}_t, t)$.
- Conditional diffusion: predict noise with: $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)$, where $\mathbf{c}$ conditional input.

Summary

1. Deep nets learn *representations*

2. This is useful because representations transfer — they act as prior knowledge that enables quick learning on new tasks

3. Representations can also be learned without labels

4. Without labels there are many ways to learn representations. We saw:

    1. representations as compressed codes

    2. representations that are predictive of their context

5. Language is a powerful form of supervision

6. Language is a natural “user interface” for computer vision systems

Adapted from Isola, Freeman, Torralba

Next class: sound and touch

Next class: vision and language