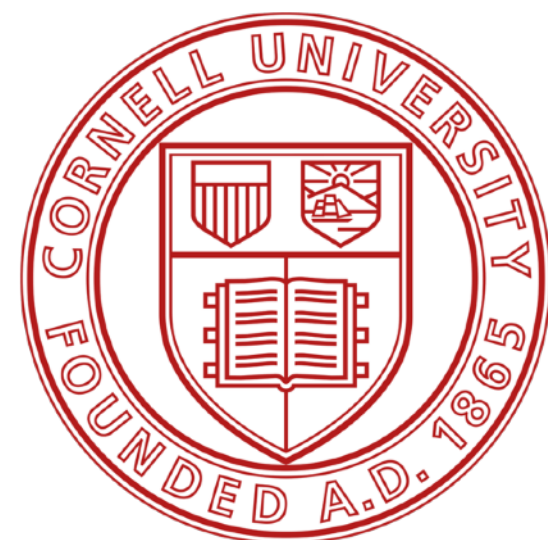


Lecture 12: Representation learning

CS 5670: Introduction to Computer Vision



Announcements

- PS3 out (due on a Friday, Oct. 17th)

Supervised computer vision



Object recognition [Russakovsky et al., "ImageNet", 2015]



Object segmentation [Gupta et al., "LVIS", 2019]

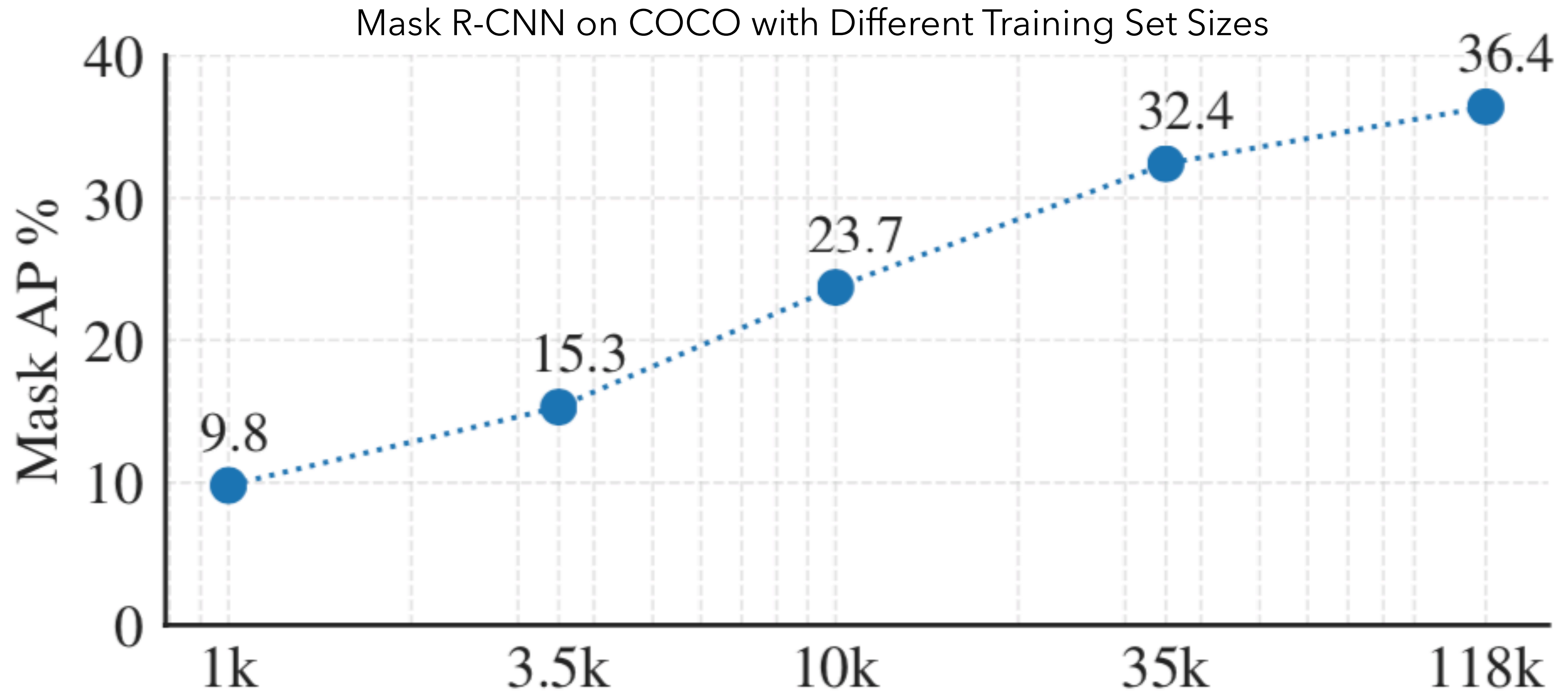
Supervised computer vision



These methods need *lots* of labeled training examples!

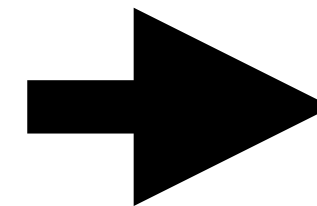
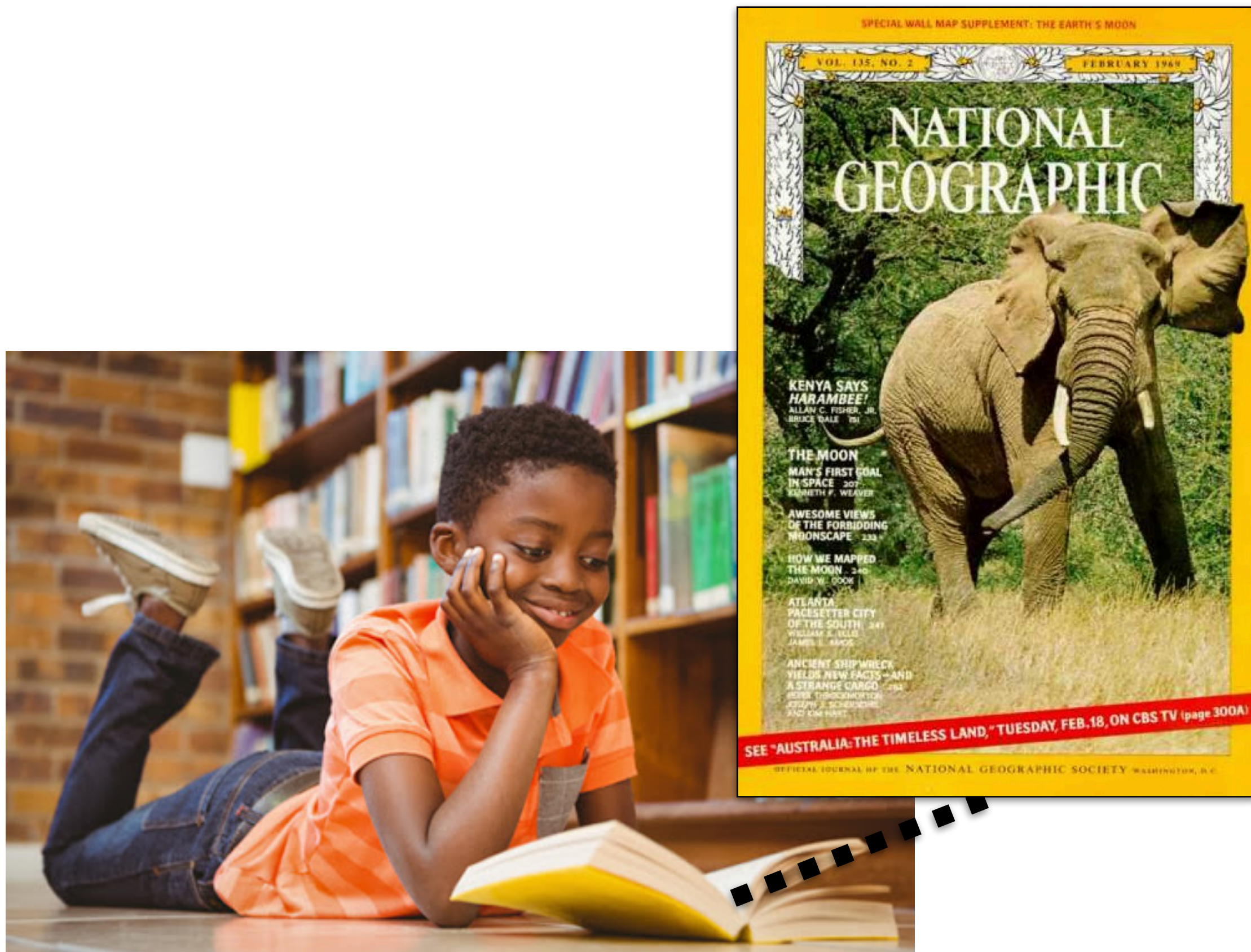
[Lin et al., COCO dataset]

The need for labeled training data



Object detection accuracy vs. dataset size

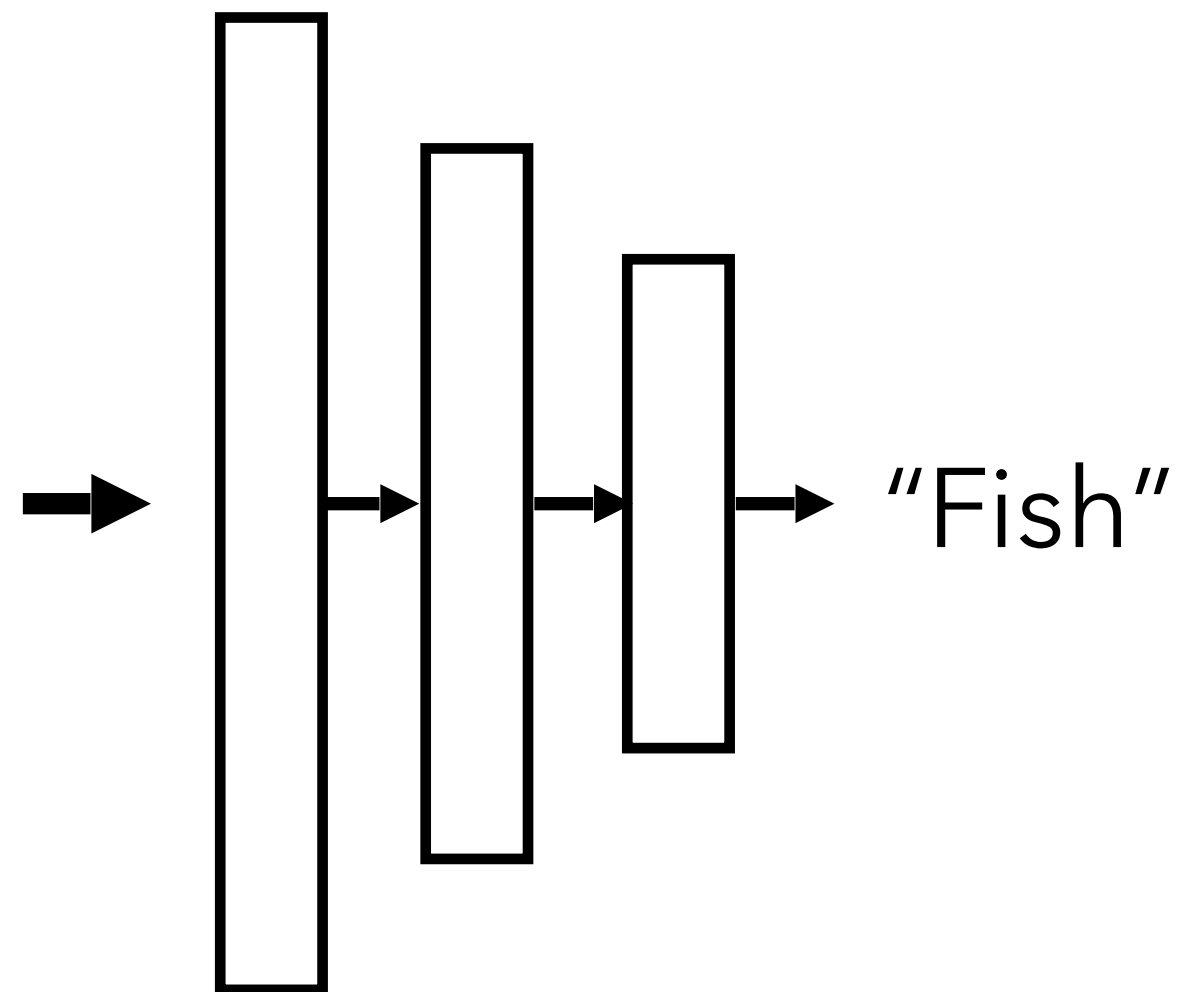
We want models that can generalize



Transfer learning

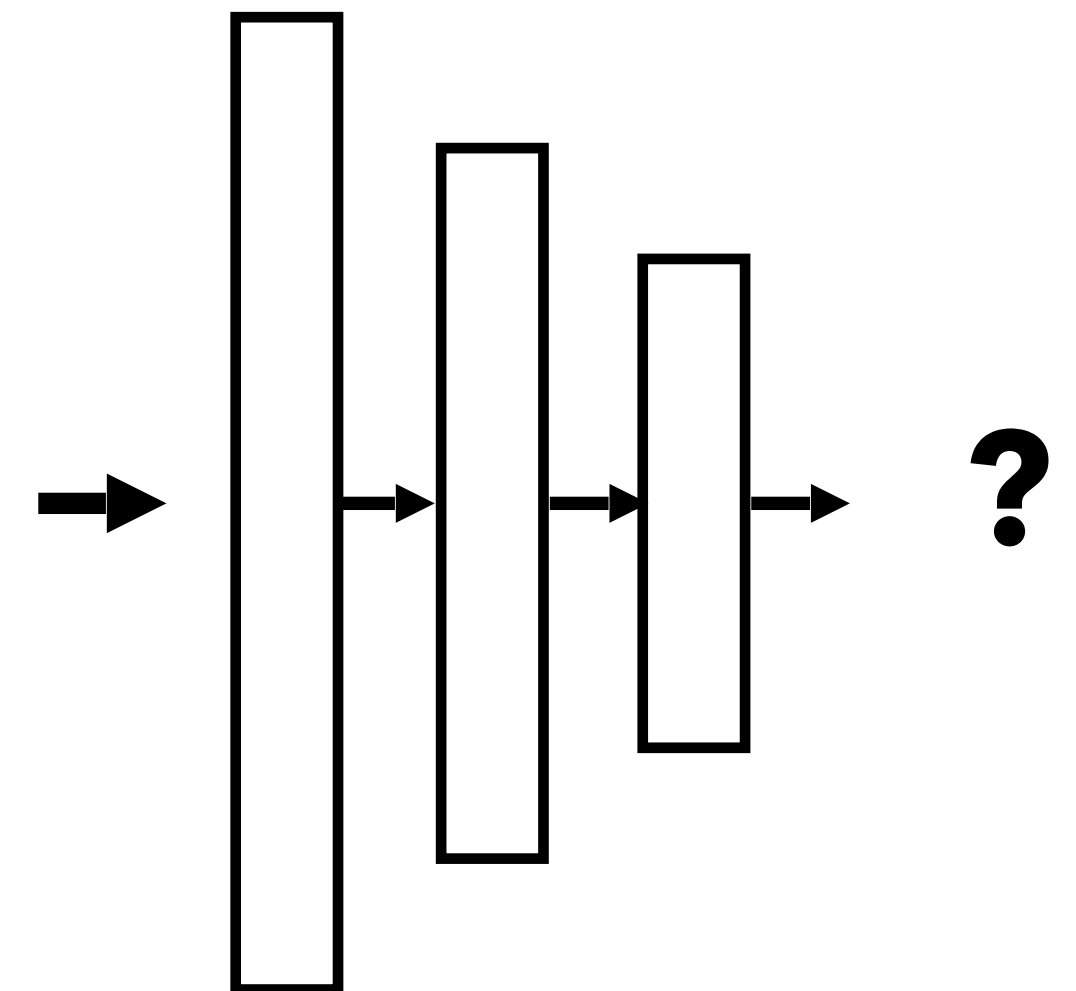
Training

Object recognition



Testing

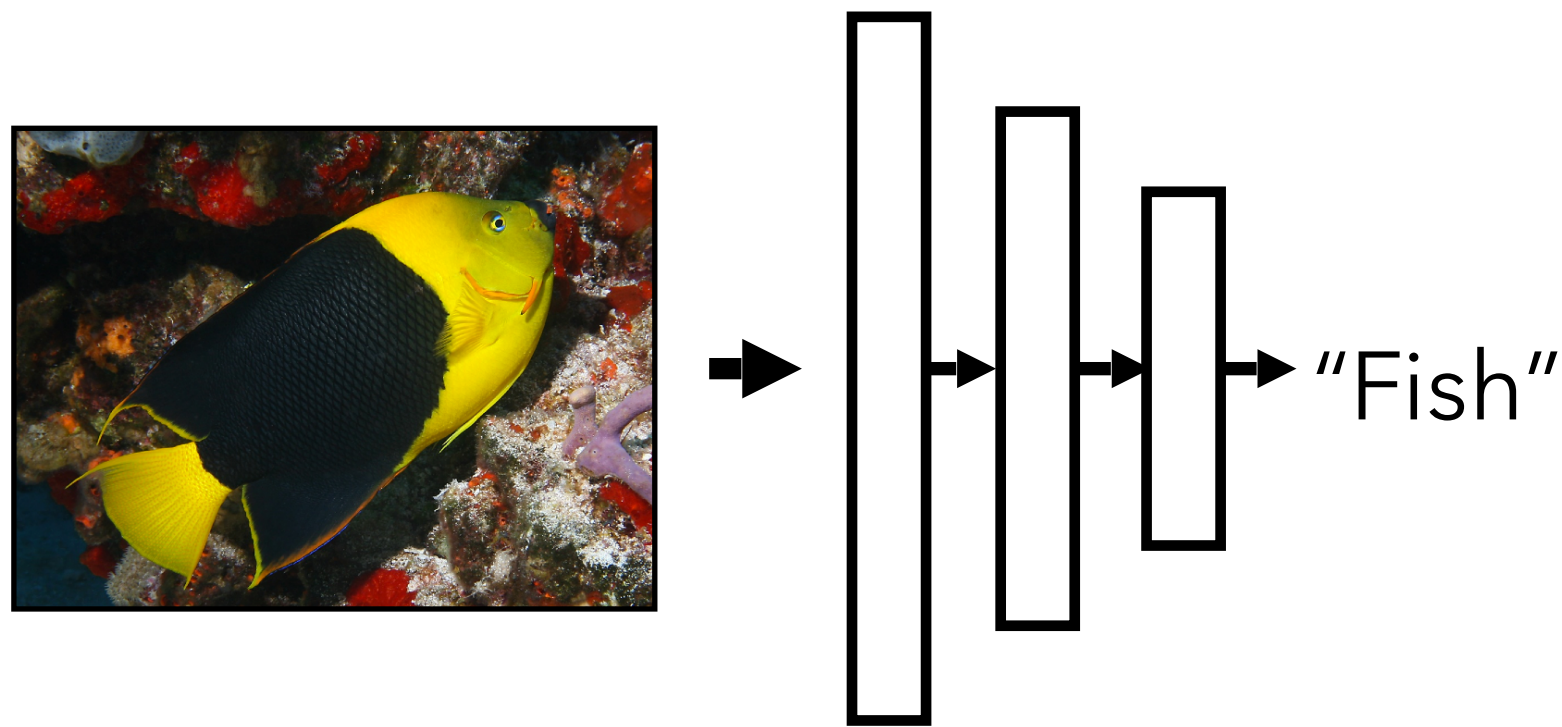
Scene recognition



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

Pretraining

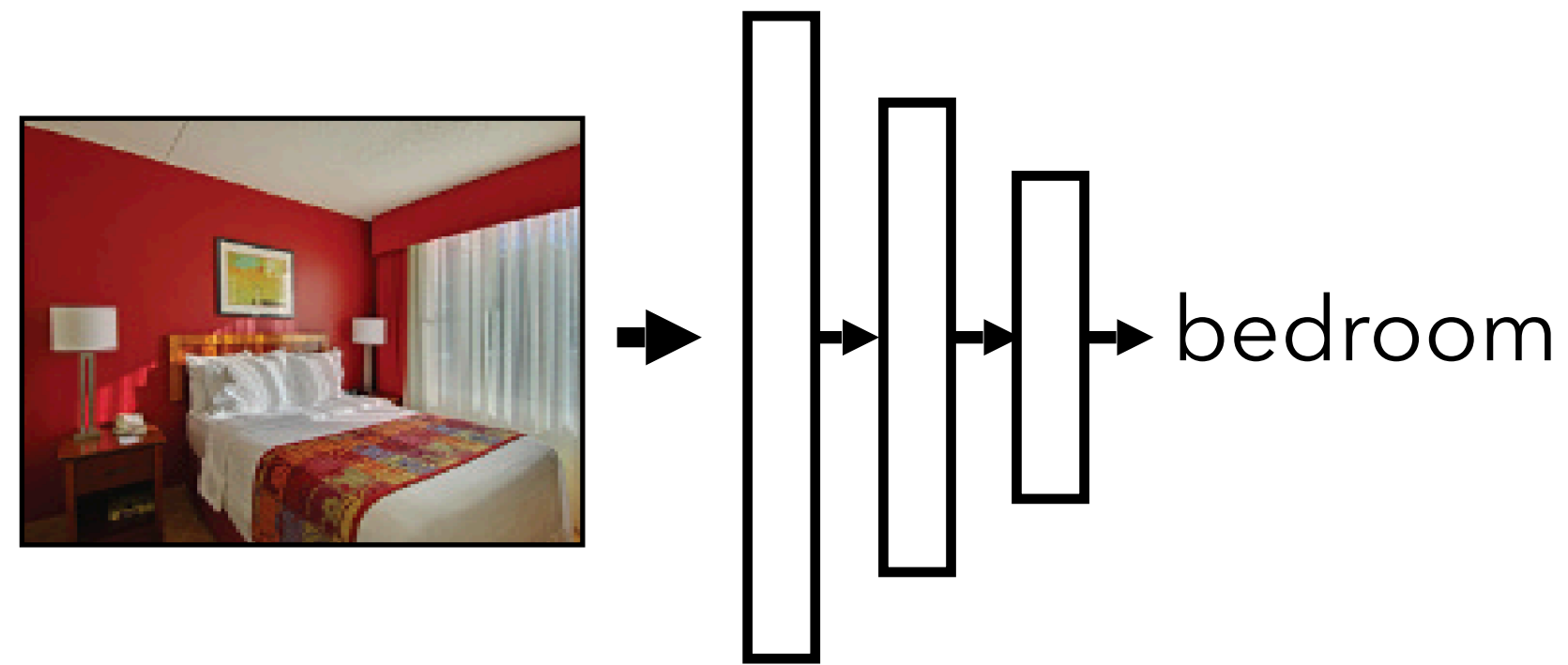
Object recognition



A lot of data

Finetuning

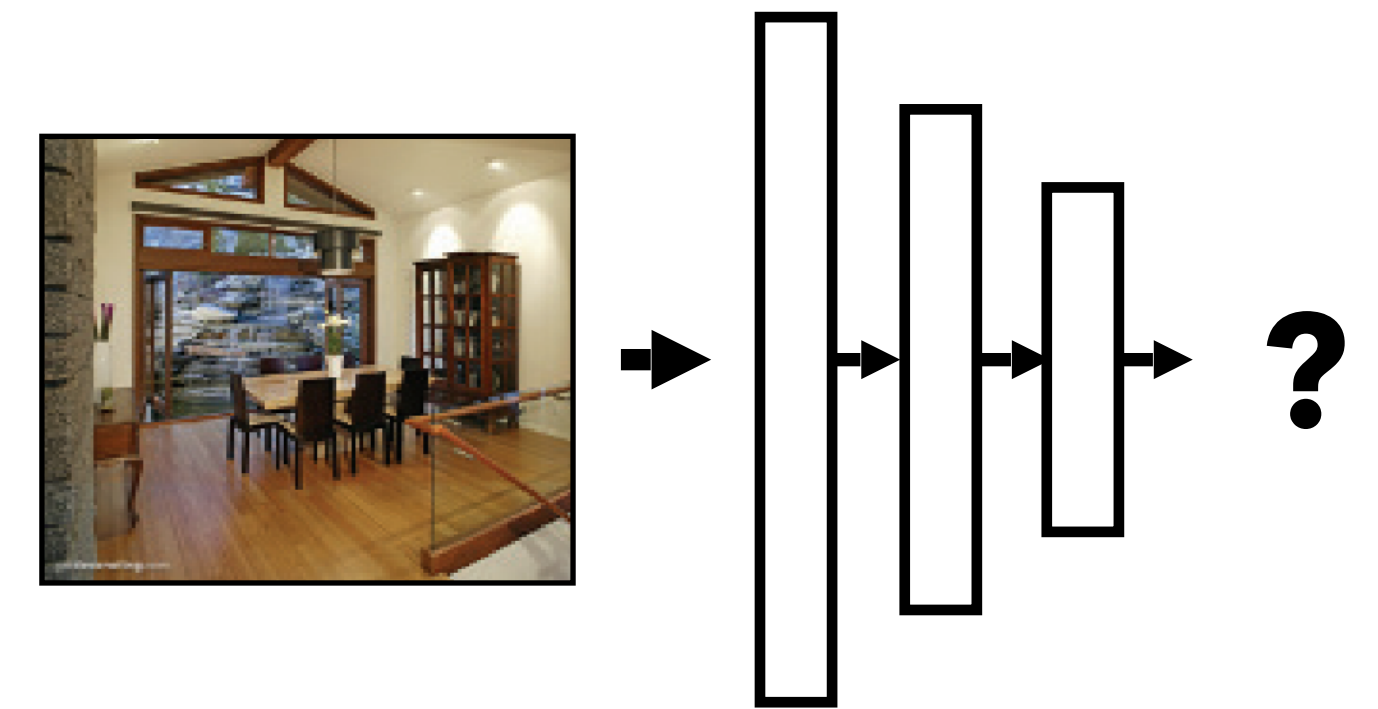
Scene recognition



A little data

Testing

Scene recognition

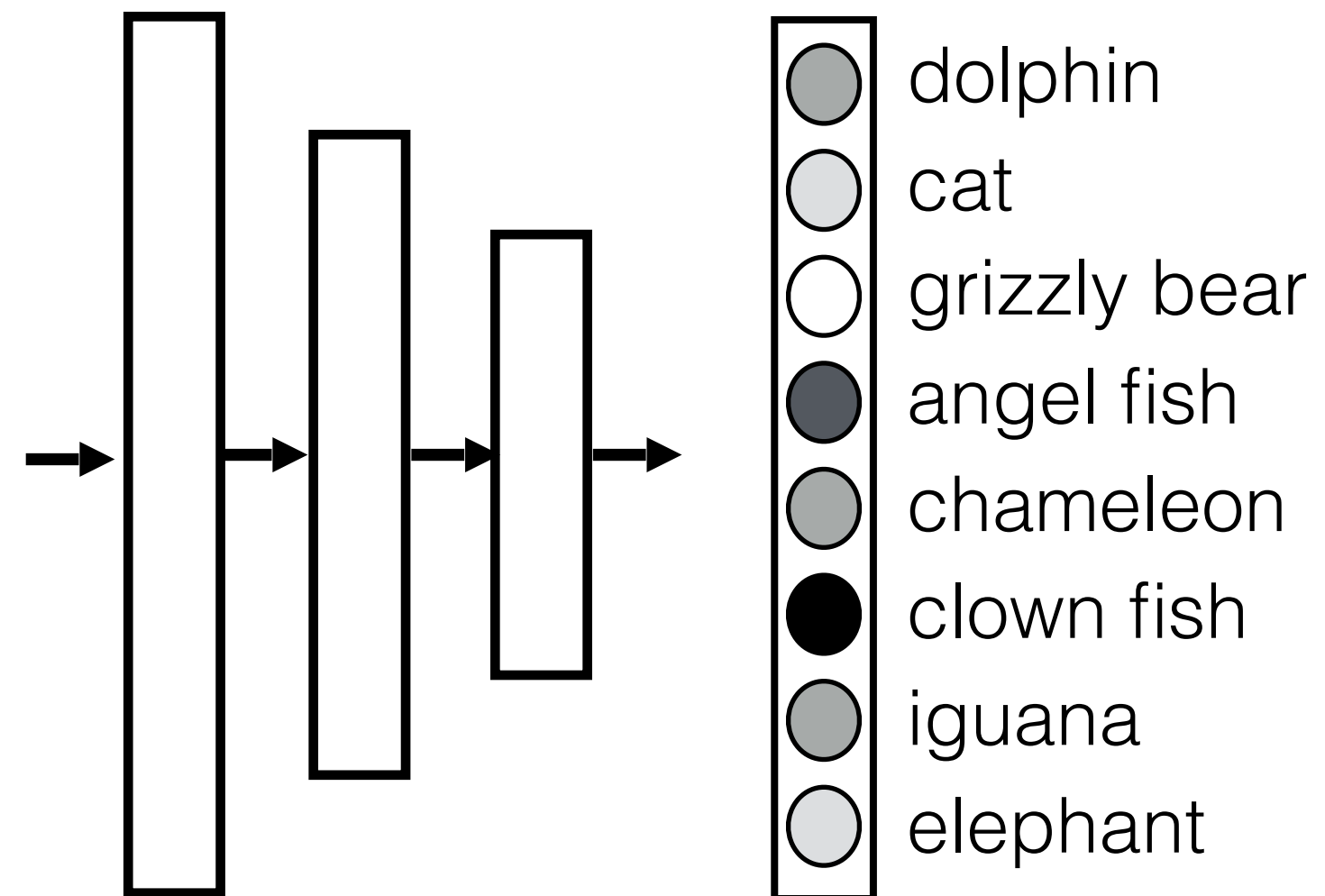


Finetuning: take a model trained on one task and retrain it for another.

Finetuning

Pretraining

Object recognition



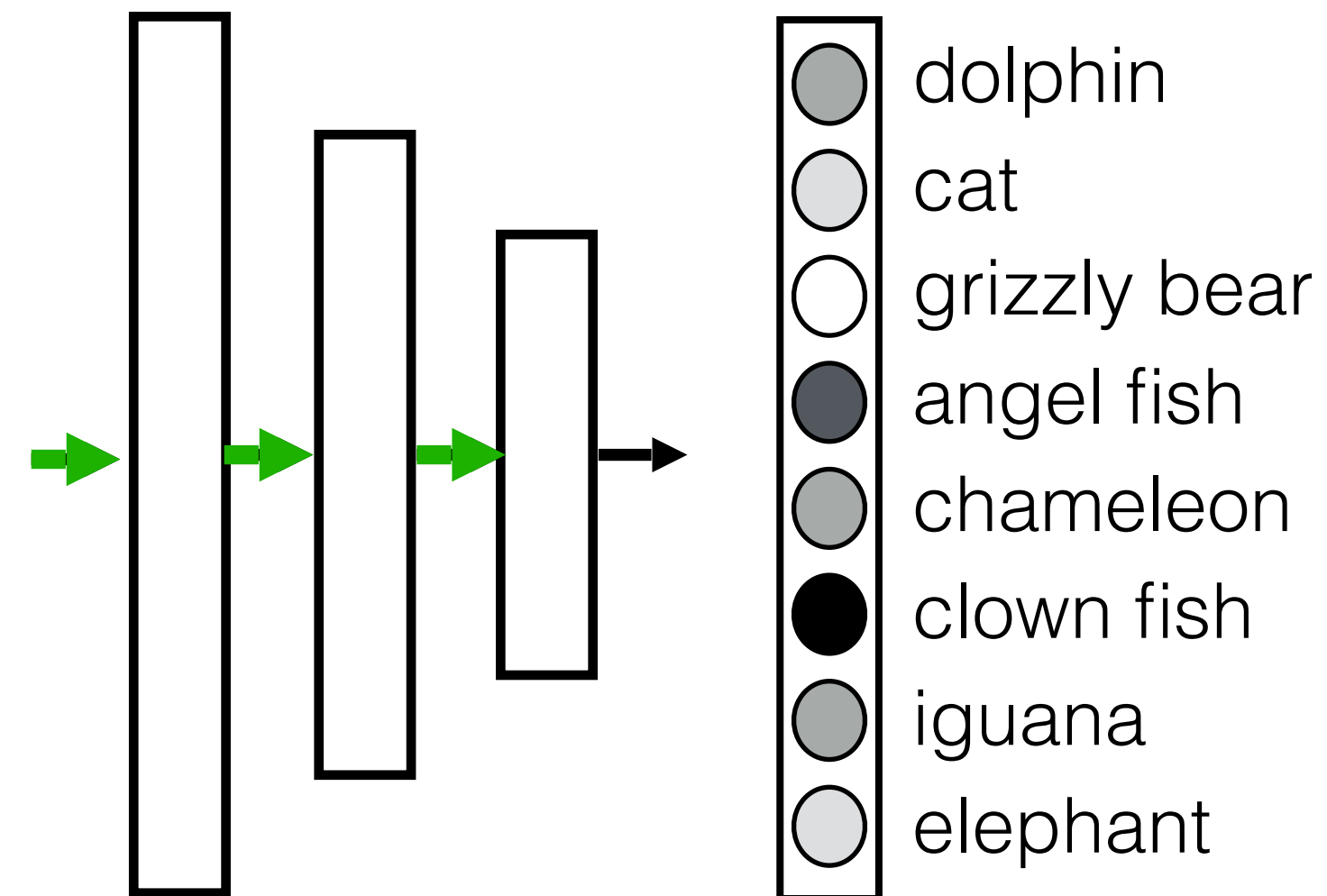
Finetuning

Scene recognition

Finetuning

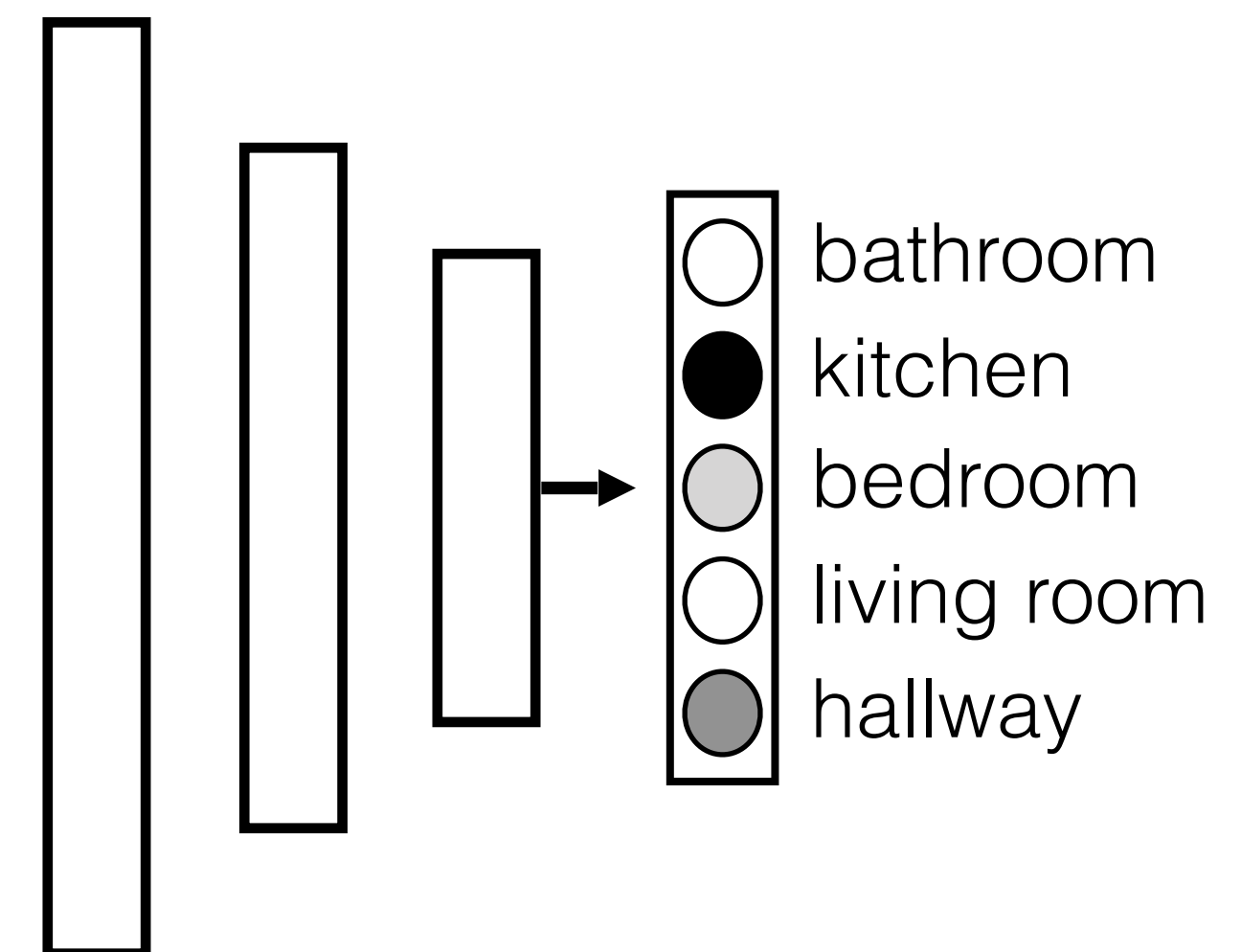
Pretraining

Object recognition



Finetuning

Scene recognition



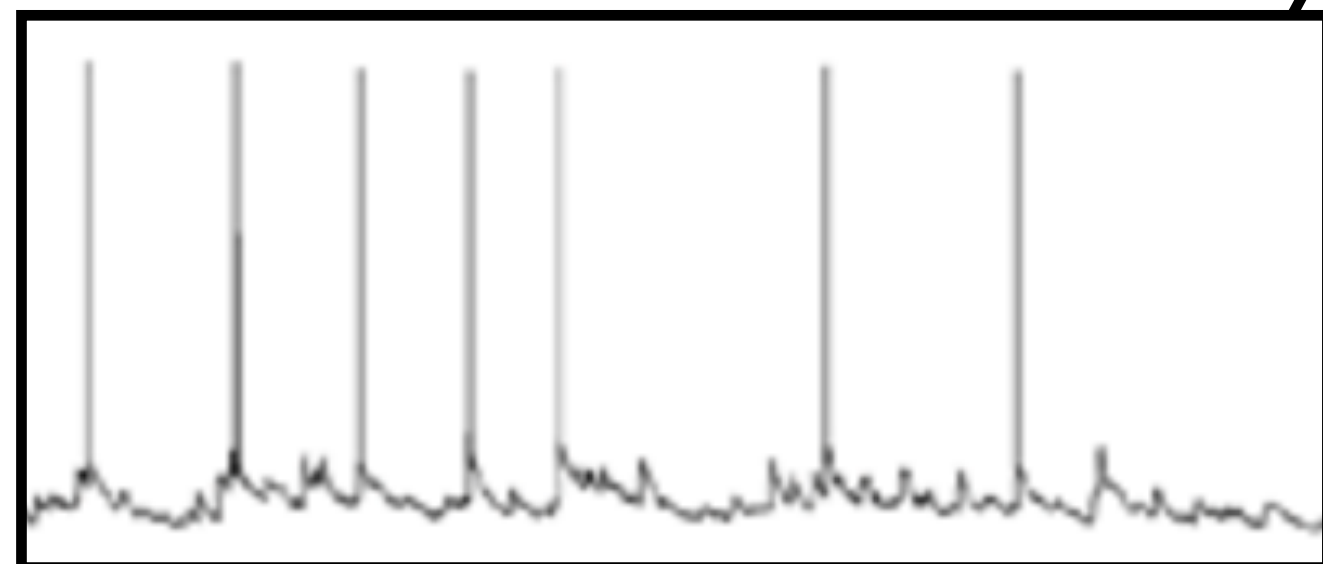
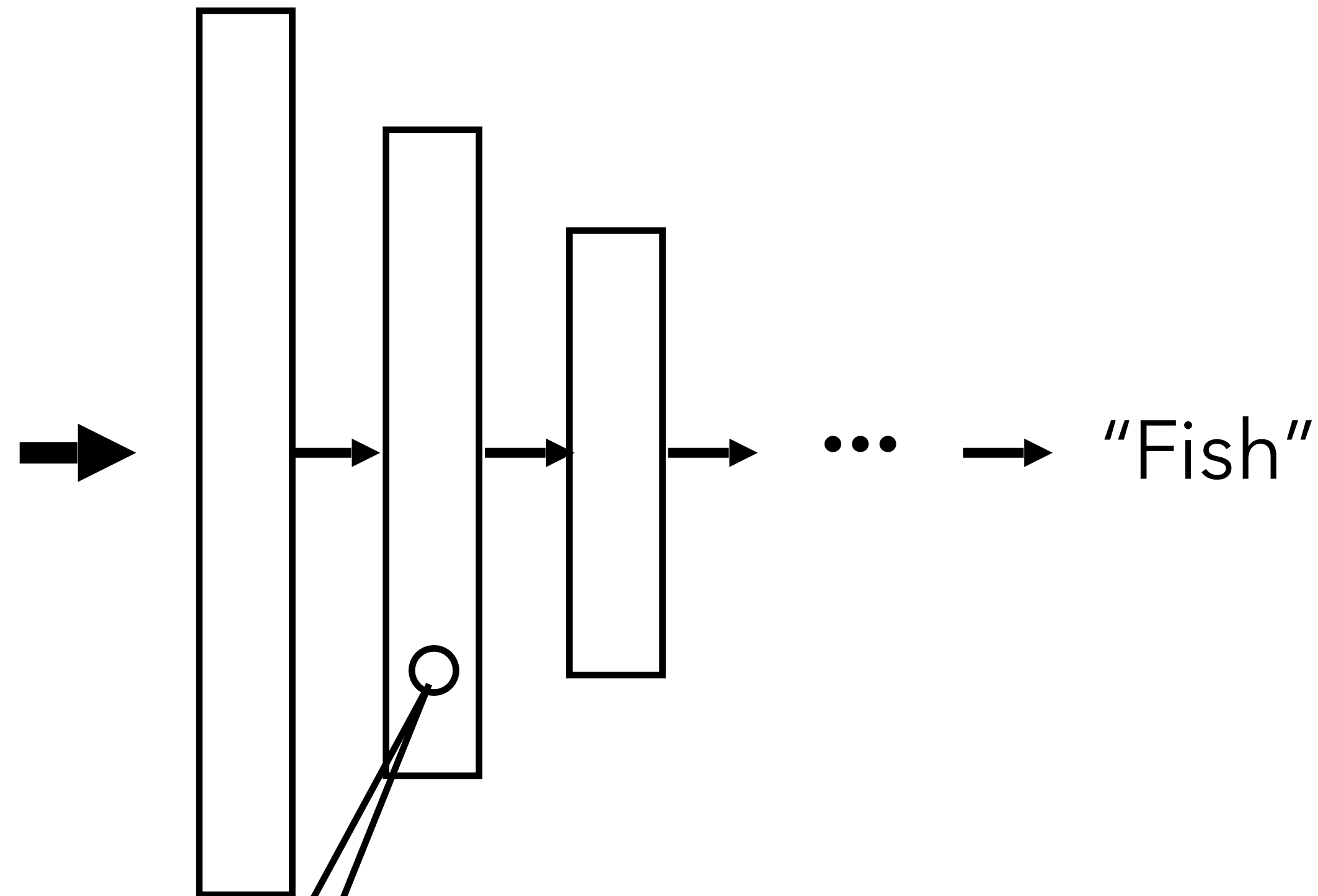
Initialize the weights using the pretraining task!

Finetuning

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

Visualizing representations

Deep net “electrophysiology”



[Zeiler & Fergus, ECCV 2014]

[Zhou et al., ICLR 2015]

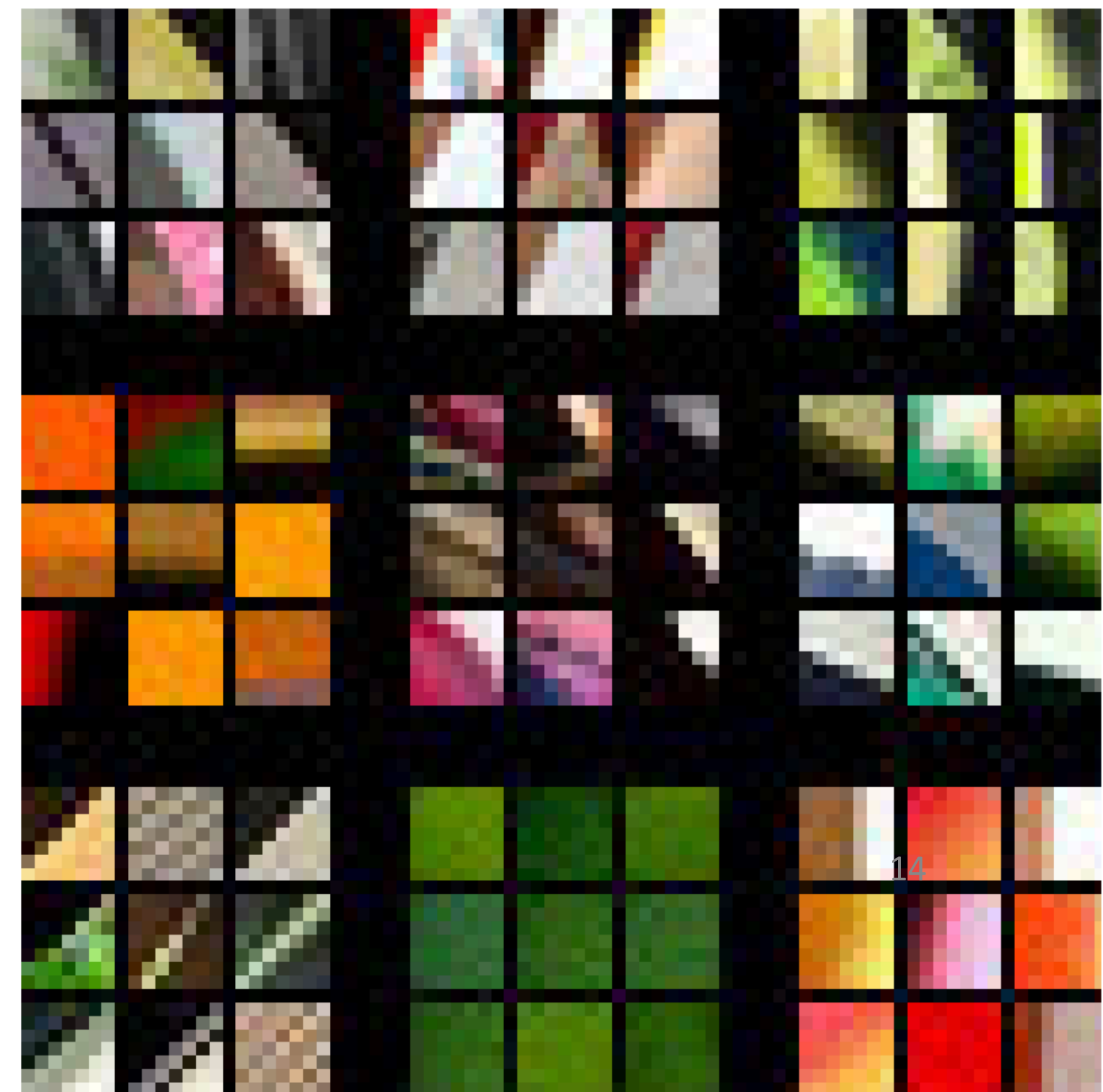
Visualizing and Understanding CNNs

[Zeiler and Fergus, 2014]

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

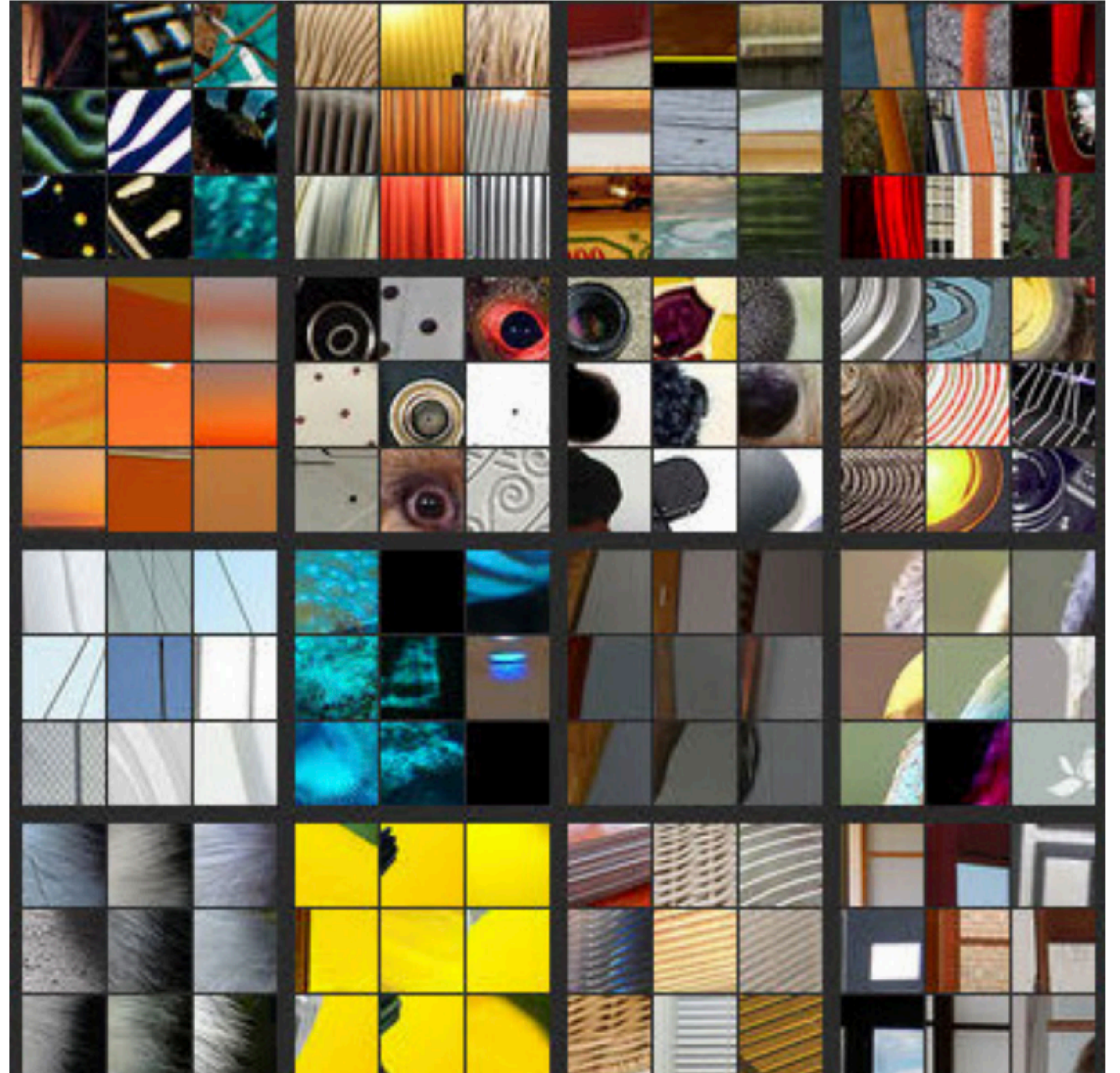


Image patches that activate each of layer 1's filters most strongly



[Zeiler and Fergus, 2014]

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



[Zeiler and Fergus, 2014]

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



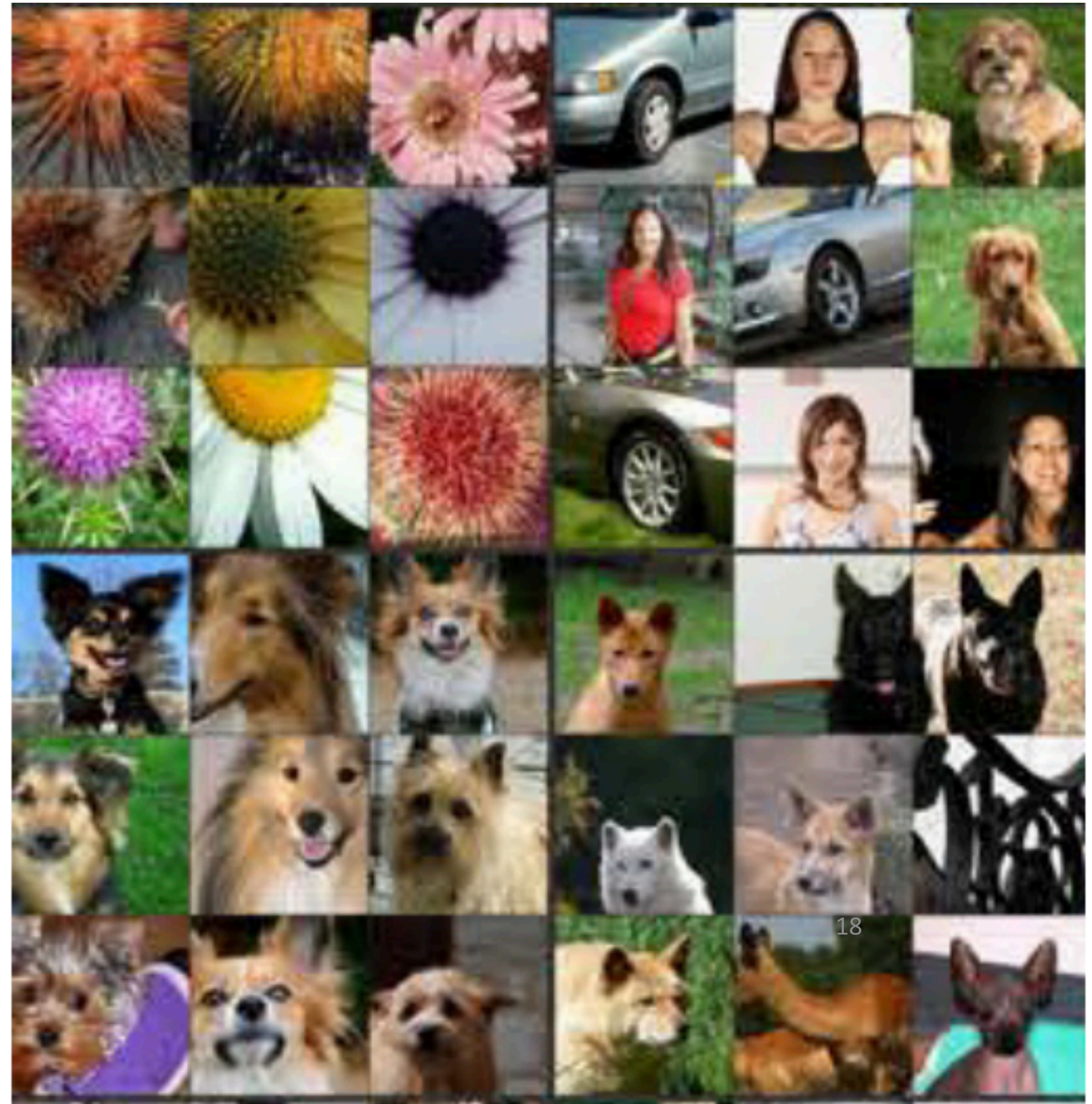
[Zeiler and Fergus, 2014]

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

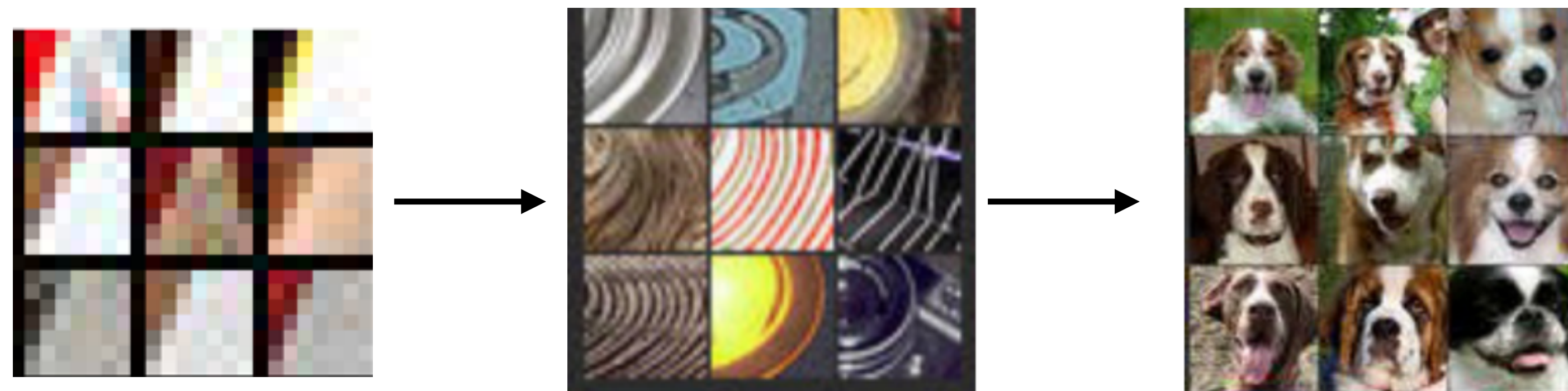
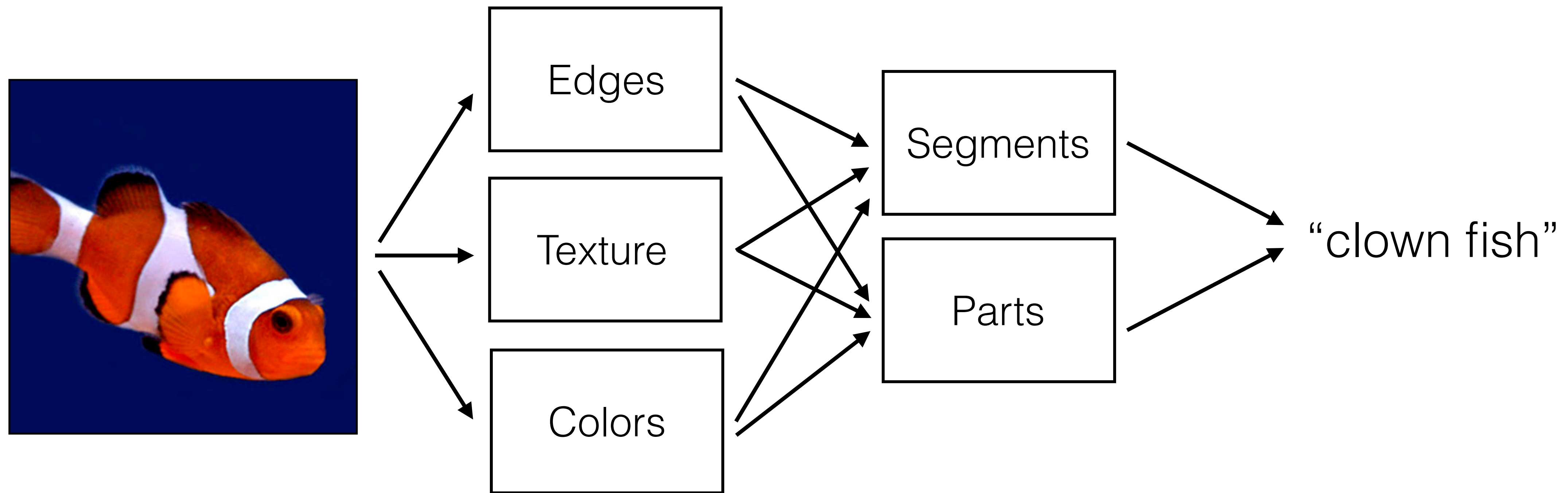


[Zeiler and Fergus, 2014]

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



CNNs learned the classical visual recognition pipeline



Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]

pool 1



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

Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]

pool 2



Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]

conv 4



22

Object Detectors Emerge in Deep Scene CNNs

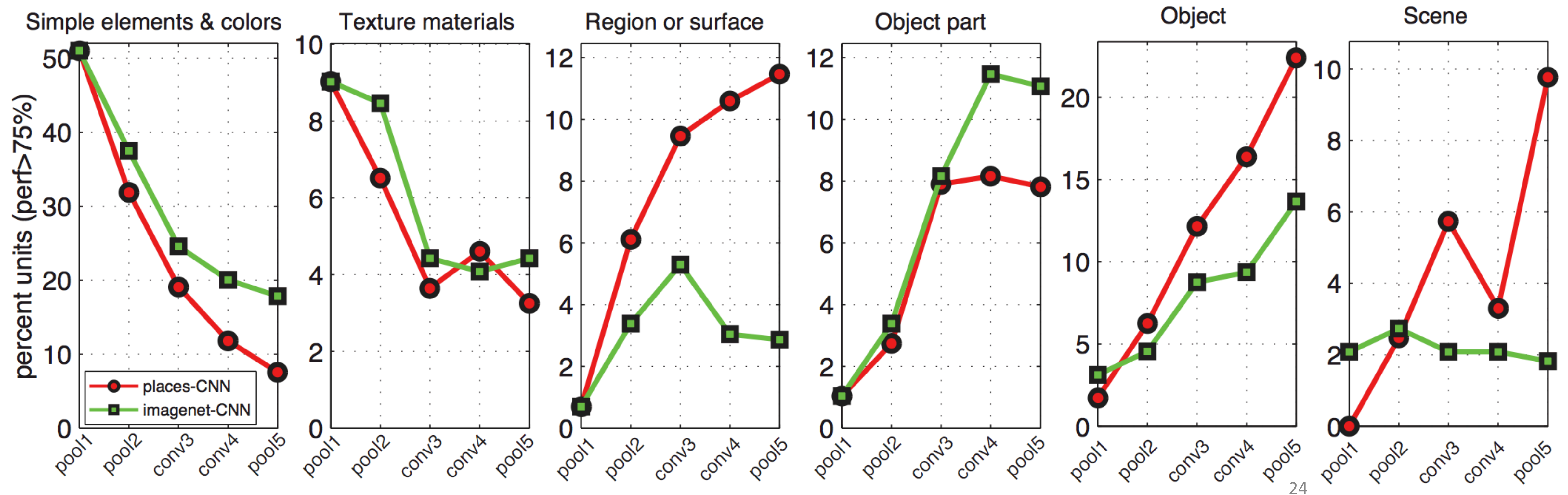
[Zhou et al., ICLR 2015]

pool 5



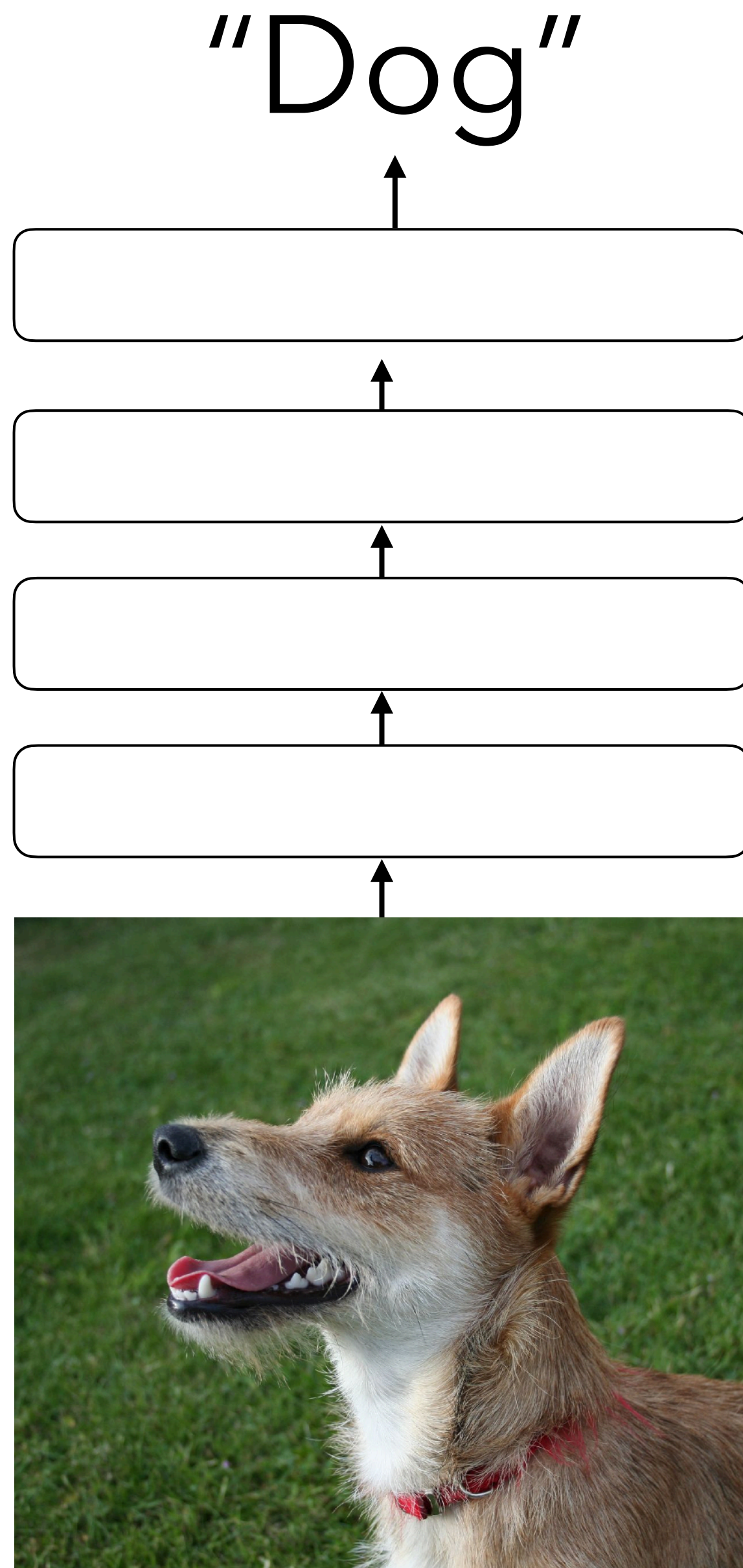
Object Detectors Emerge in Deep Scene CNNs

[Zhou et al., ICLR 2015]



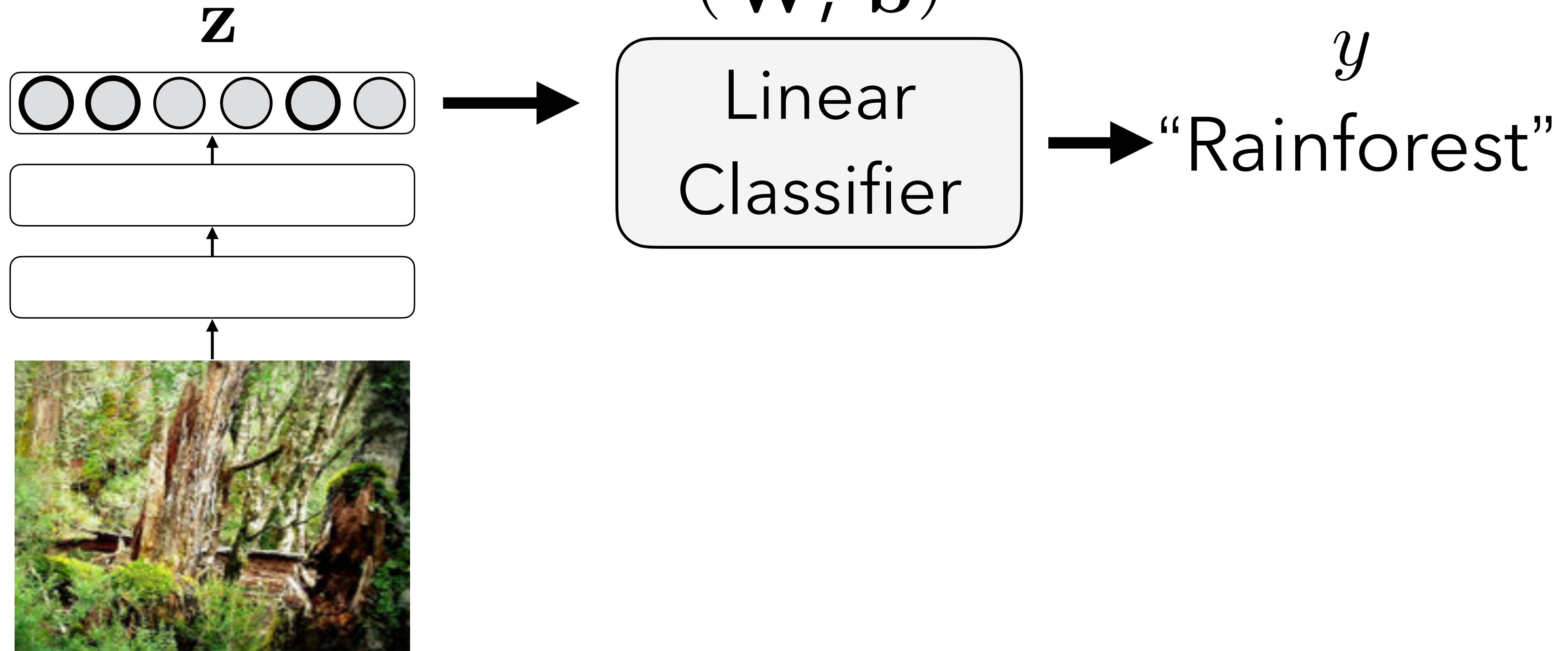
Linear probe

Object recognition
network



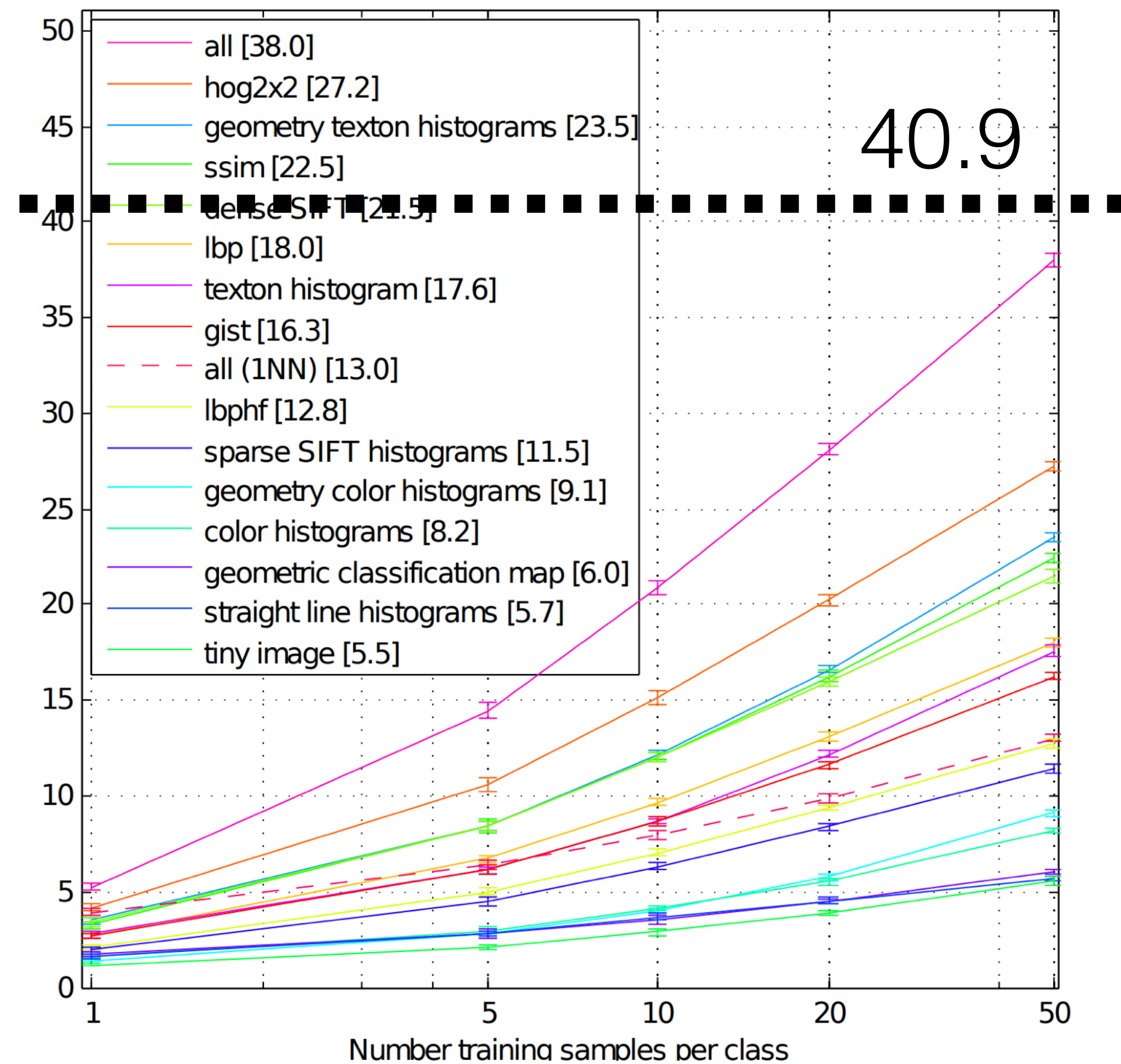
Linear probe

Feature representation



Transferring CNN features

Hand-crafted features

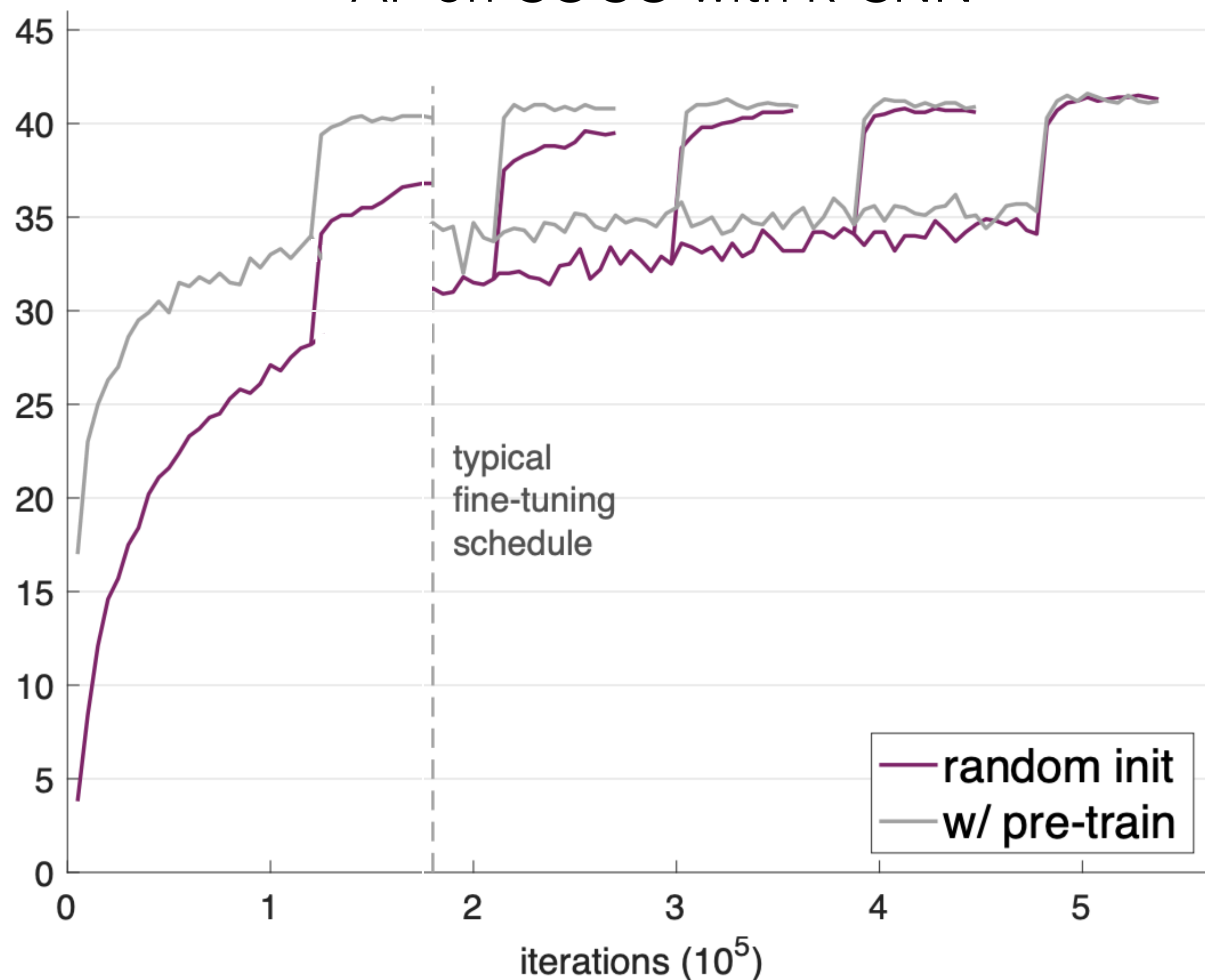


CNN features pretrained on ImageNet
+ linear classifier [Donahue et al. 2013]

[Xiao et al., CVPR 2010]

Case study: object detection

AP on COCO with R-CNN

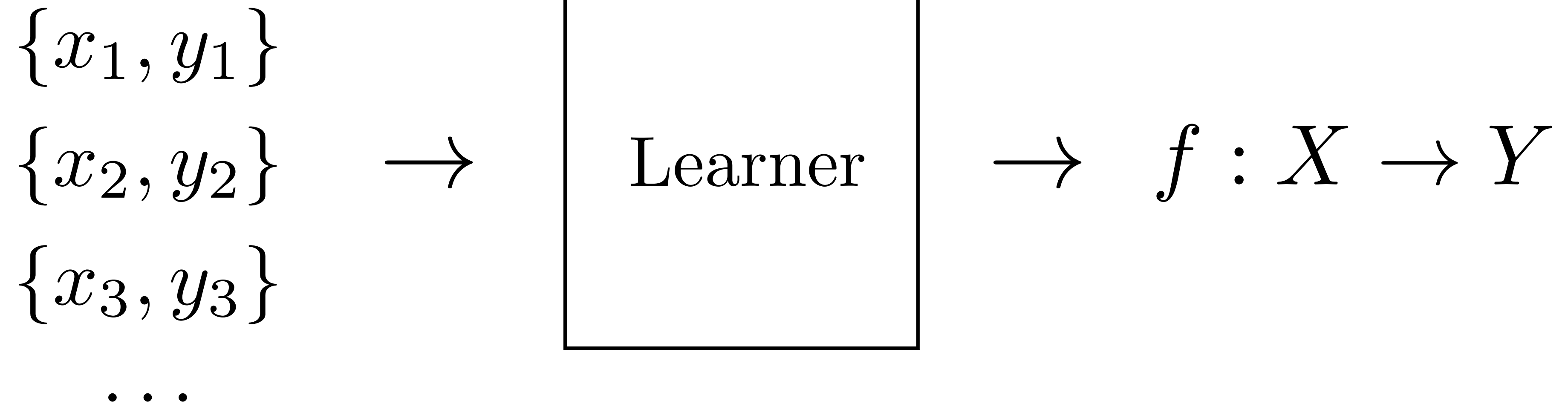


Observations:

- ImageNet pretraining speeds up object detection training by 5x
- Big performance gains for small/medium datasets (e.g. 1K examples per class)
- No benefit for large datasets

Learning from examples

Training data



Representation Learning

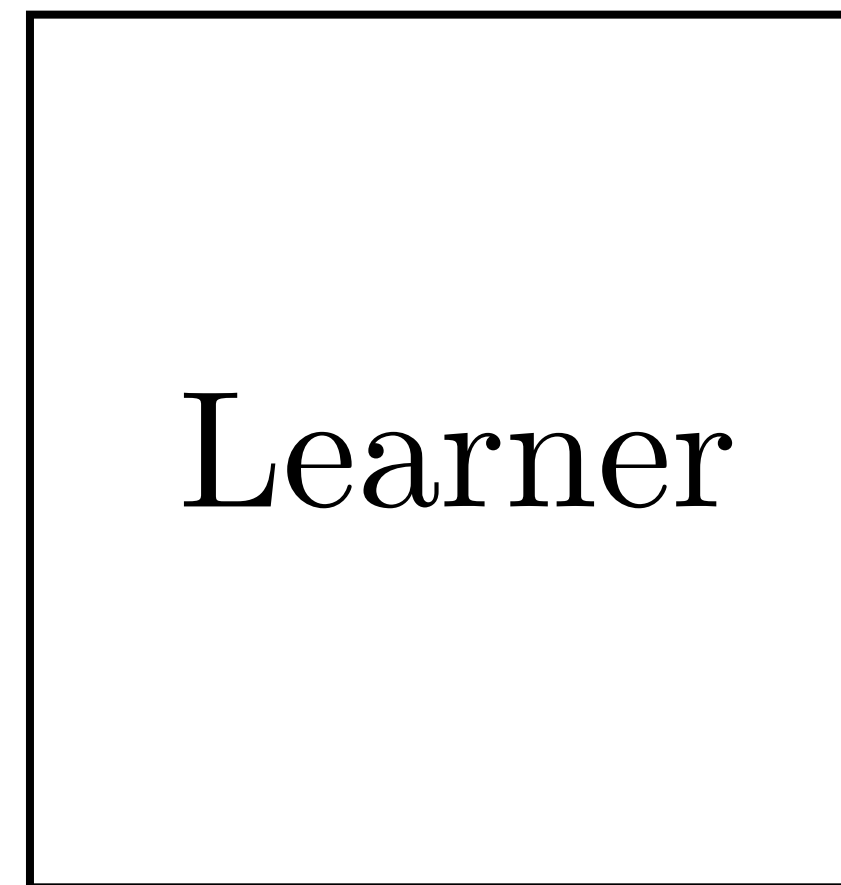
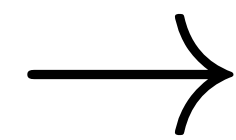
Data

$\{x_1\}$

$\{x_2\}$

$\{x_3\}$

...



Representations

How do we learn good representations?



Self-supervised learning methods

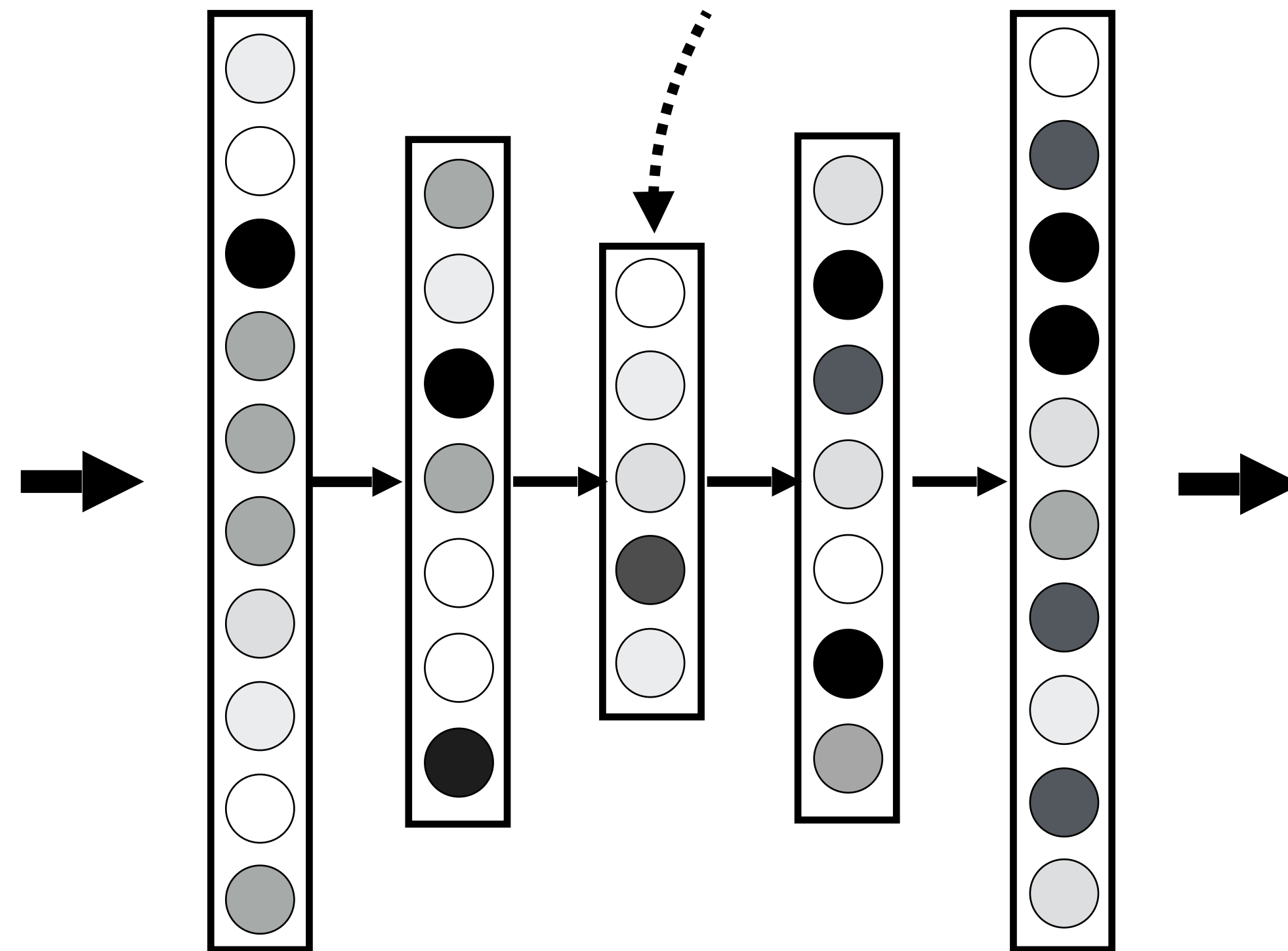
Autoencoder

compressed image code
(vector \mathbf{z})

\mathbf{X}



Image



$\hat{\mathbf{X}}$



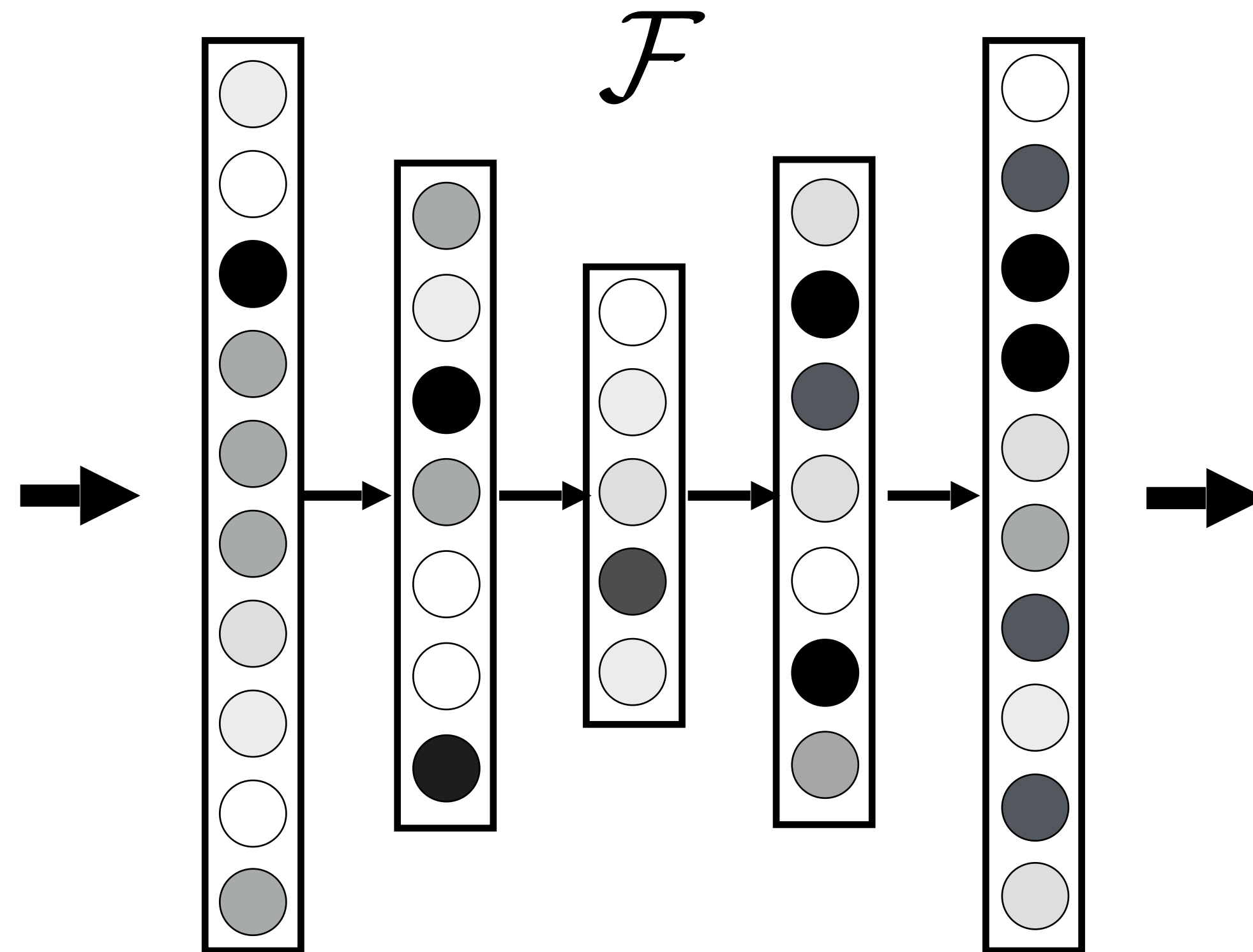
Reconstructed
image

Autoencoder

\mathbf{X}



Image



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



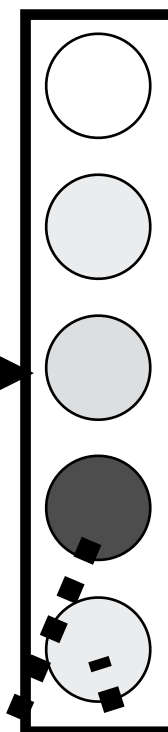
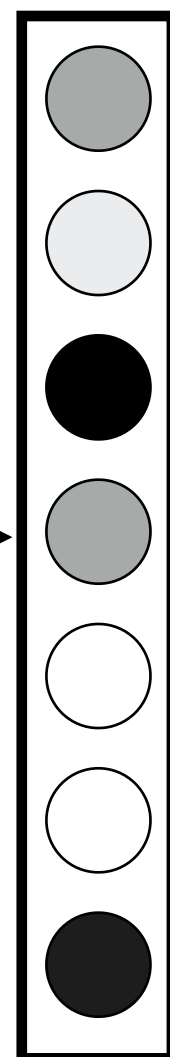
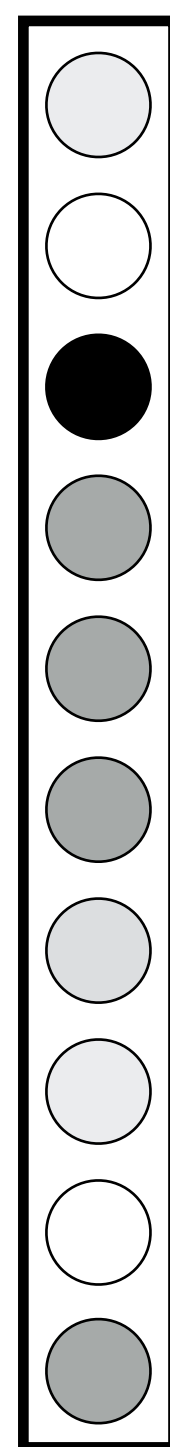
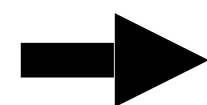
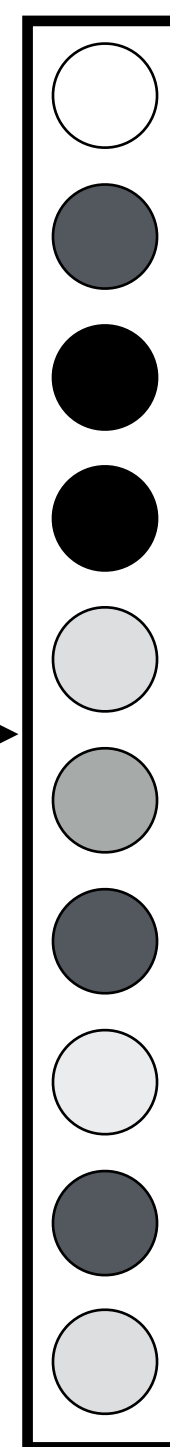
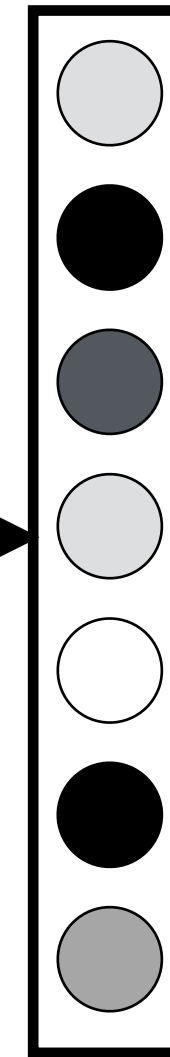
Reconstructed
image

Loss:

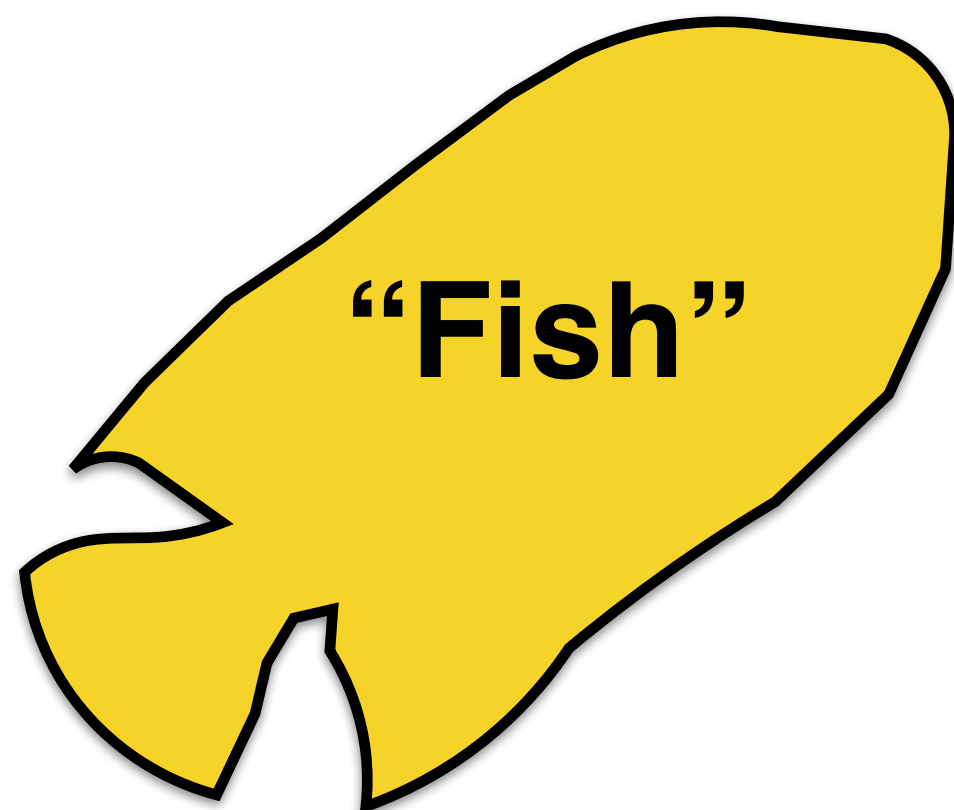
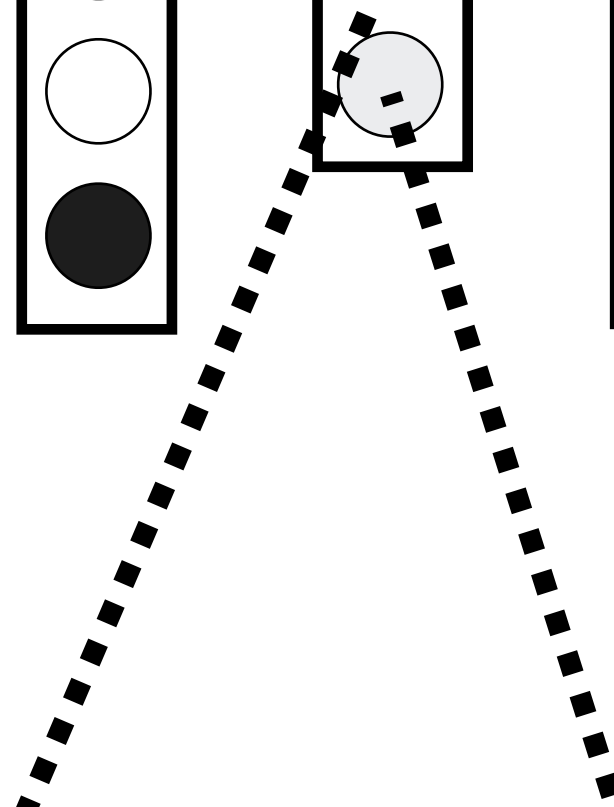
$$\mathcal{L}_{\theta} = \|\mathbf{X} - \hat{\mathbf{X}}\|^2$$

\mathbf{X} 

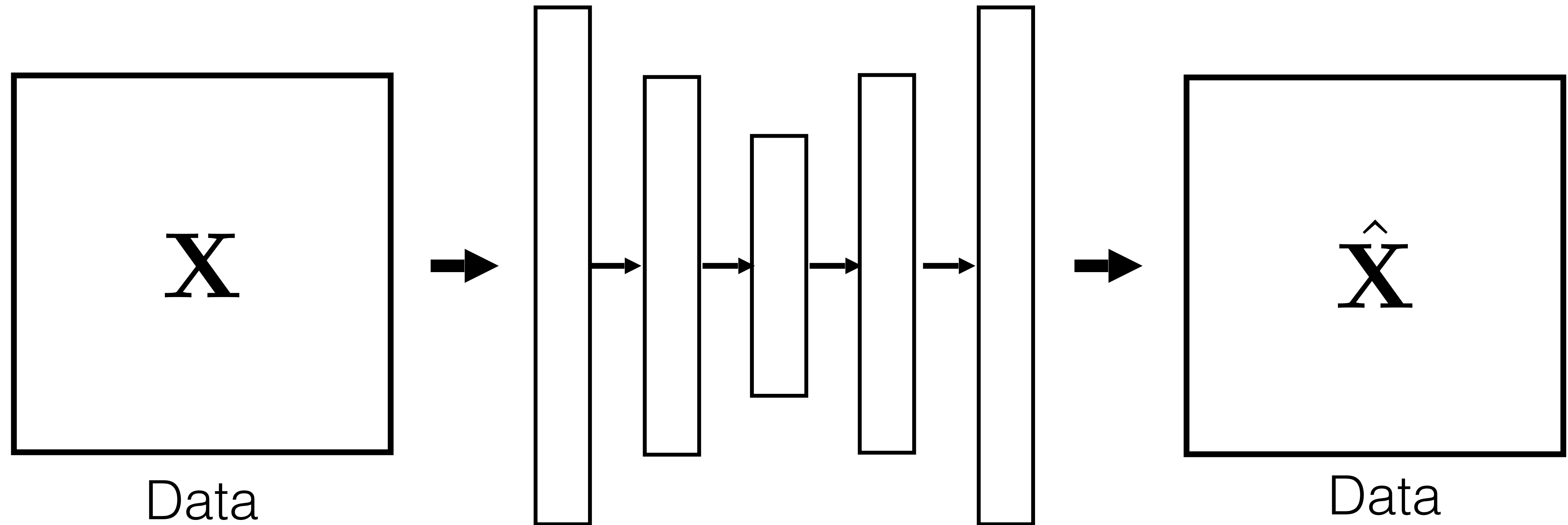
Image

 \mathcal{F} 

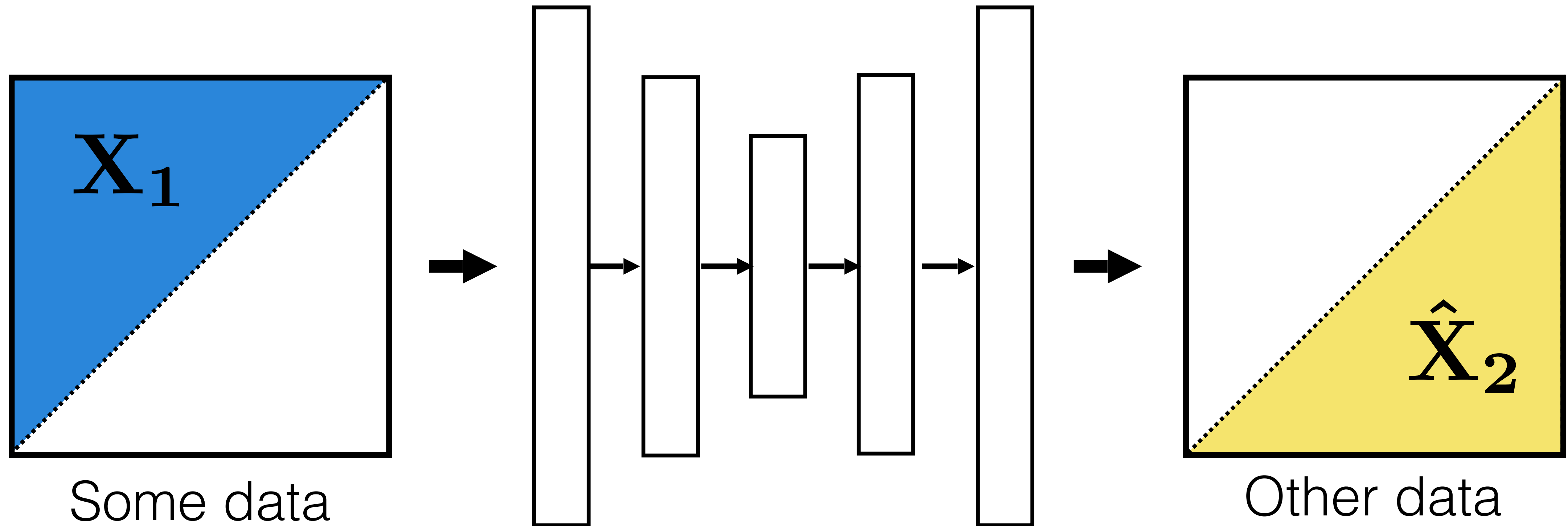
$$\hat{\mathbf{X}} = \mathcal{F}(\mathbf{X})$$

Reconstructed
image

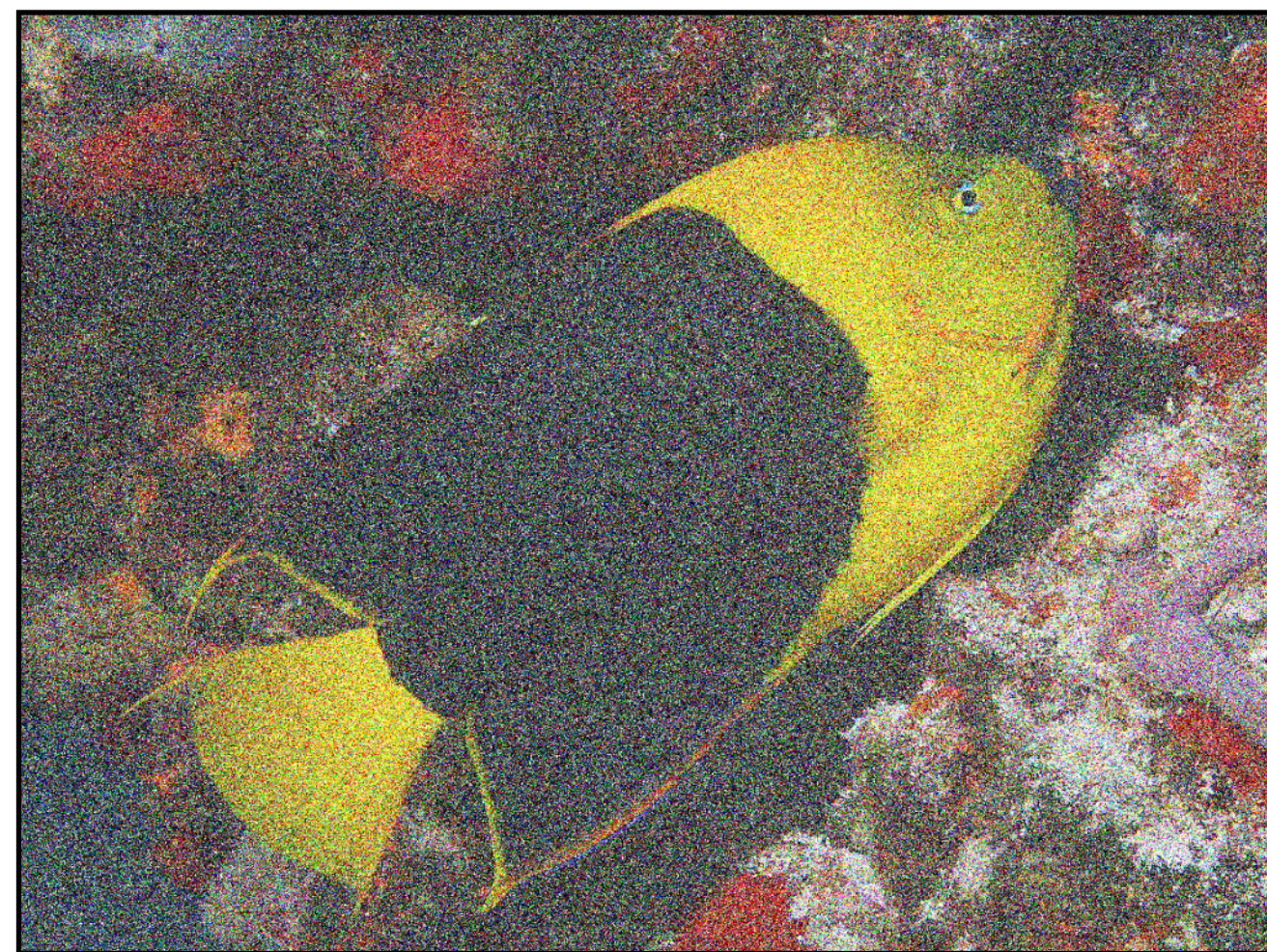
Data compression



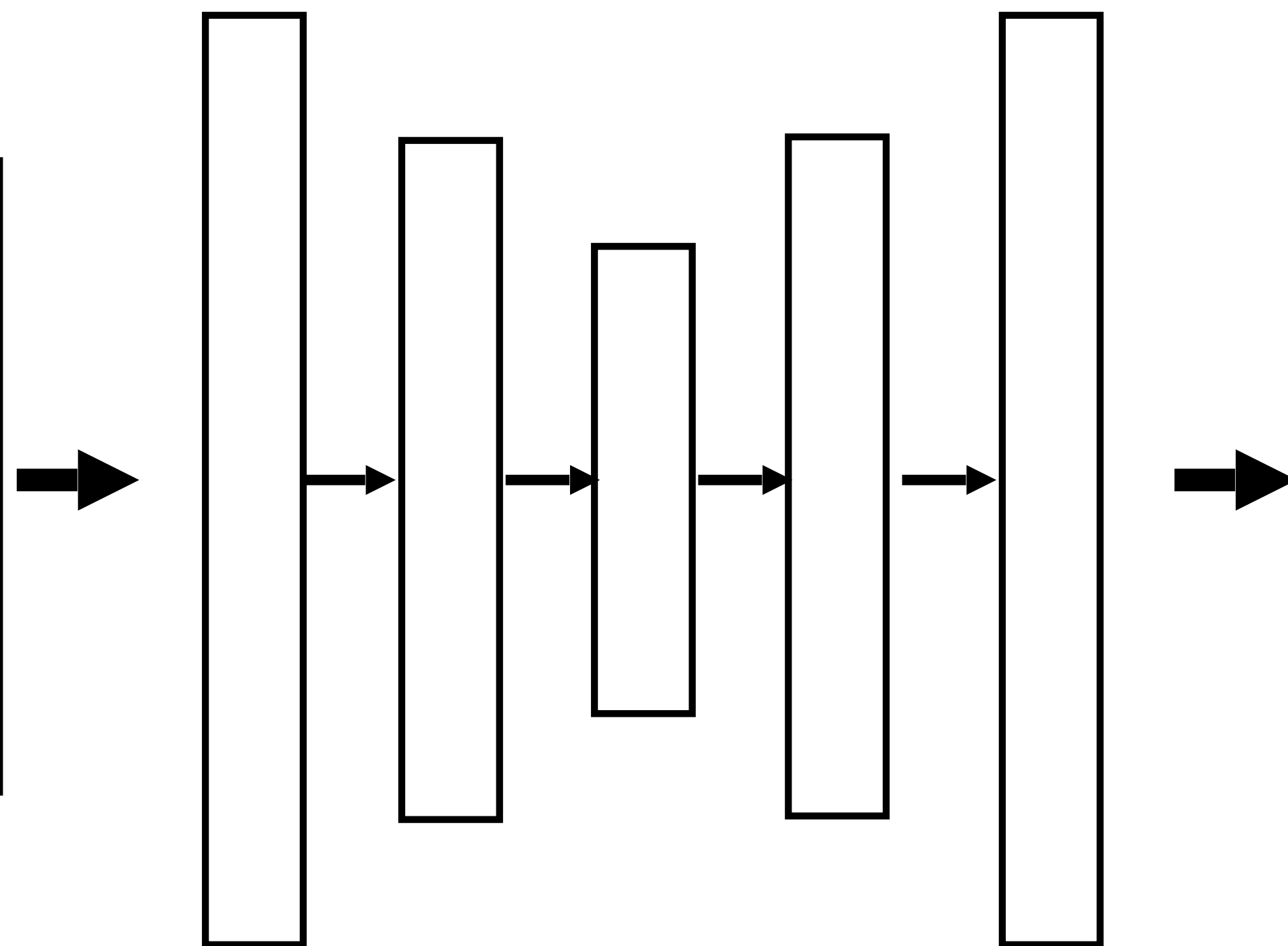
Data prediction



Denoising autoencoder

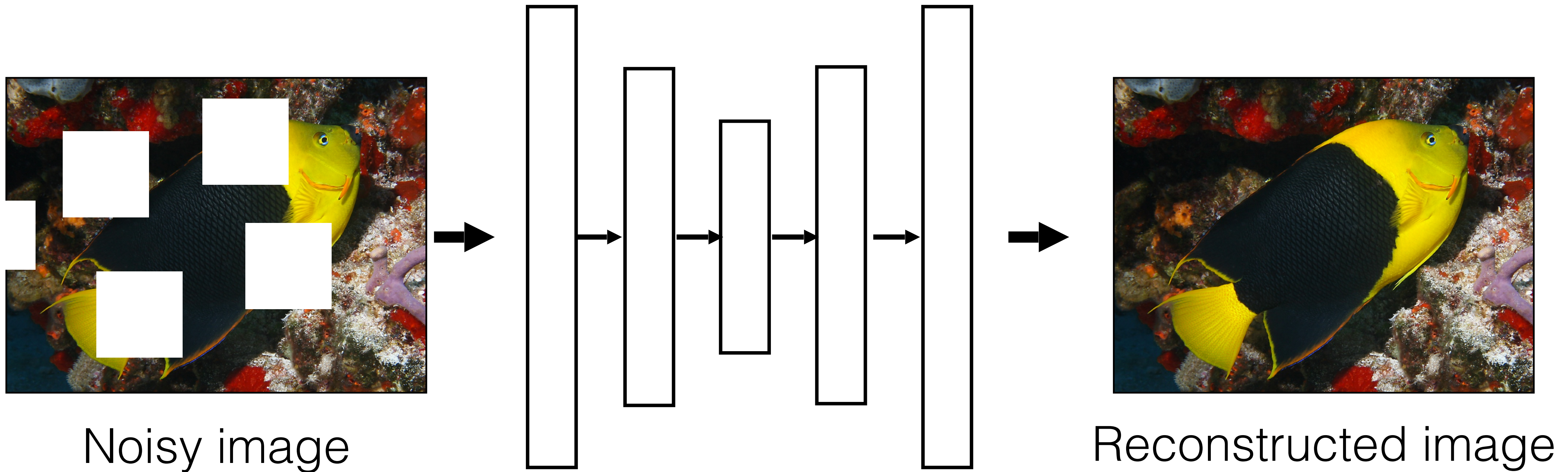


Noisy image



Reconstructed image

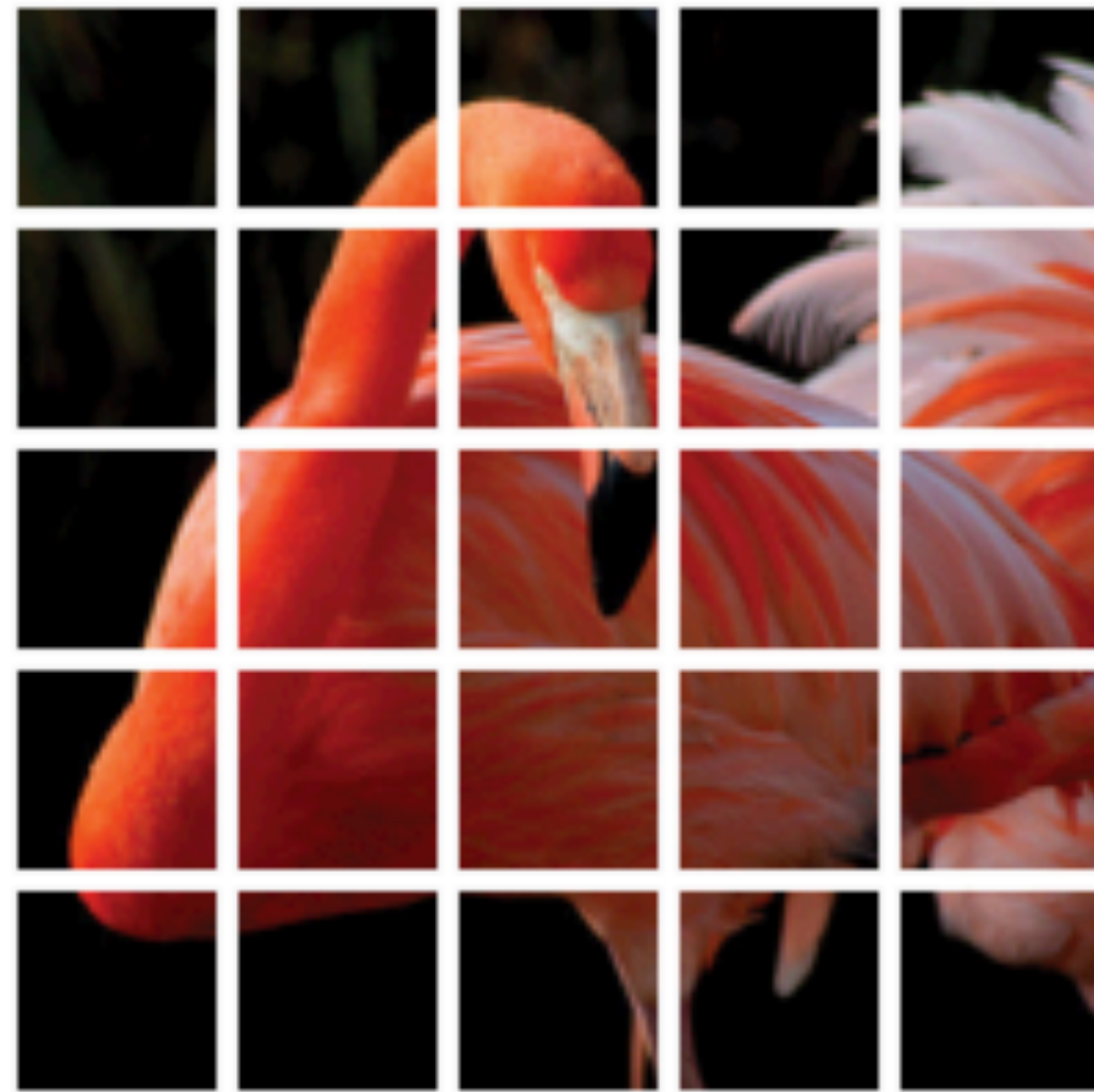
Denoising autoencoder



Other types of “noise”.

40

Masked autoencoders with transformers



image

In PS3!

[Kaiming He et al., "Masked Autoencoders Are Scalable Vision Learners", 2021]

Masked autoencoders with transformers

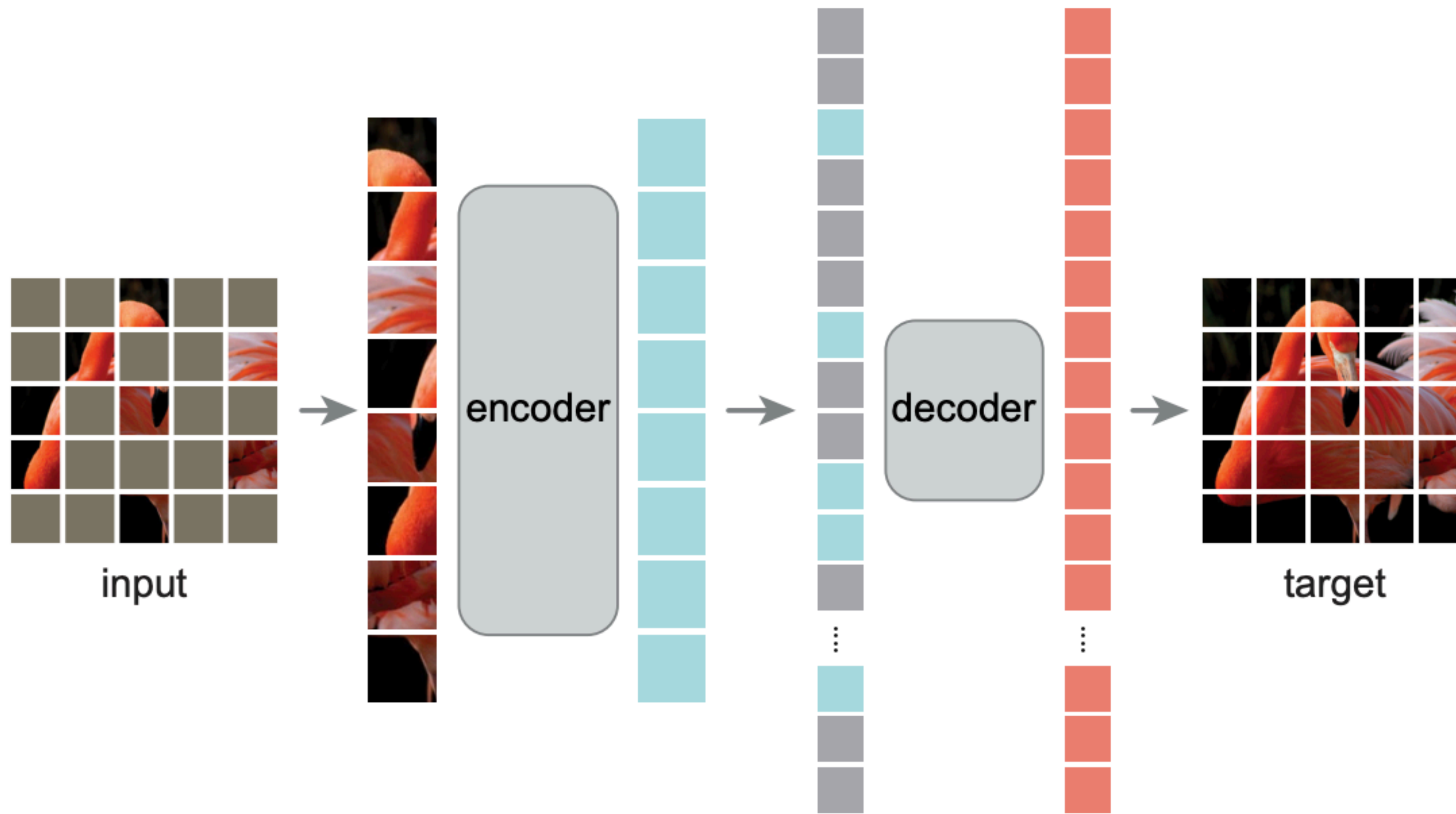


masked image

In PS3!

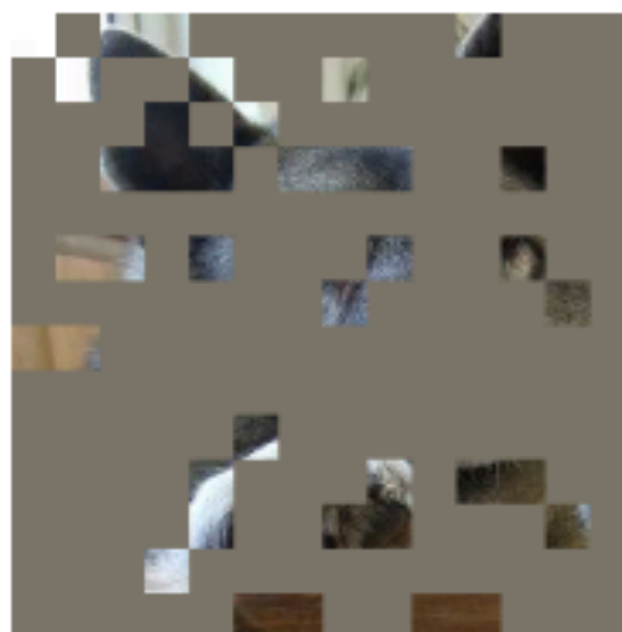
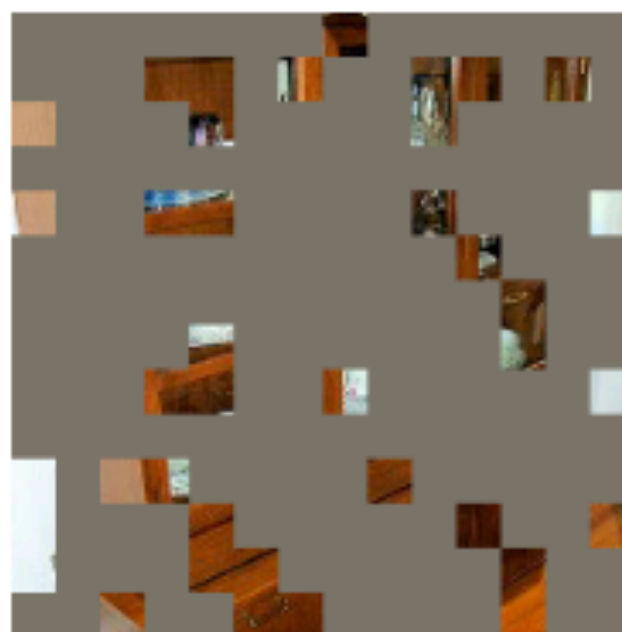
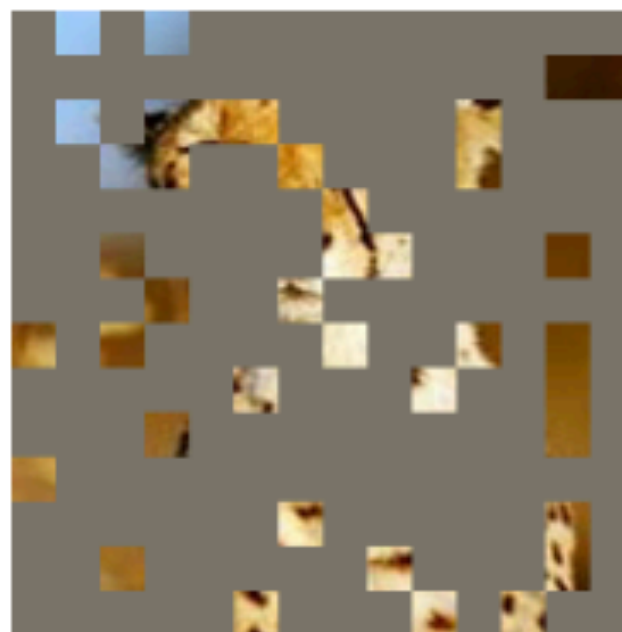
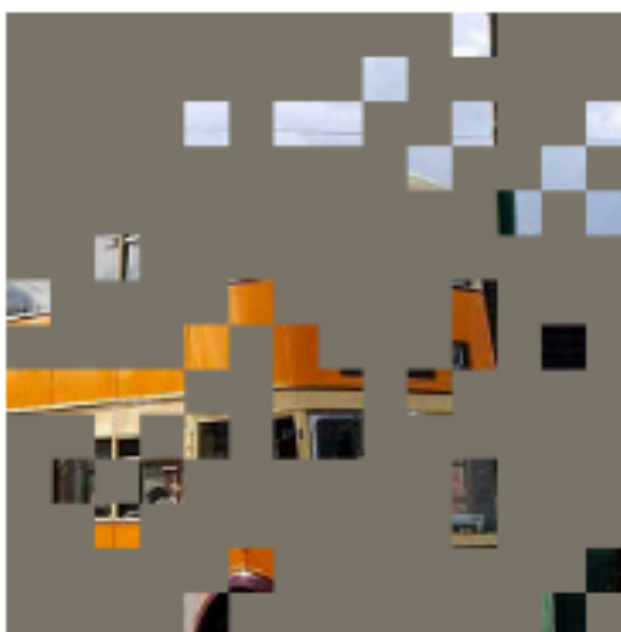
[Kaiming He et al., "Masked Autoencoders Are Scalable Vision Learners", 2021]

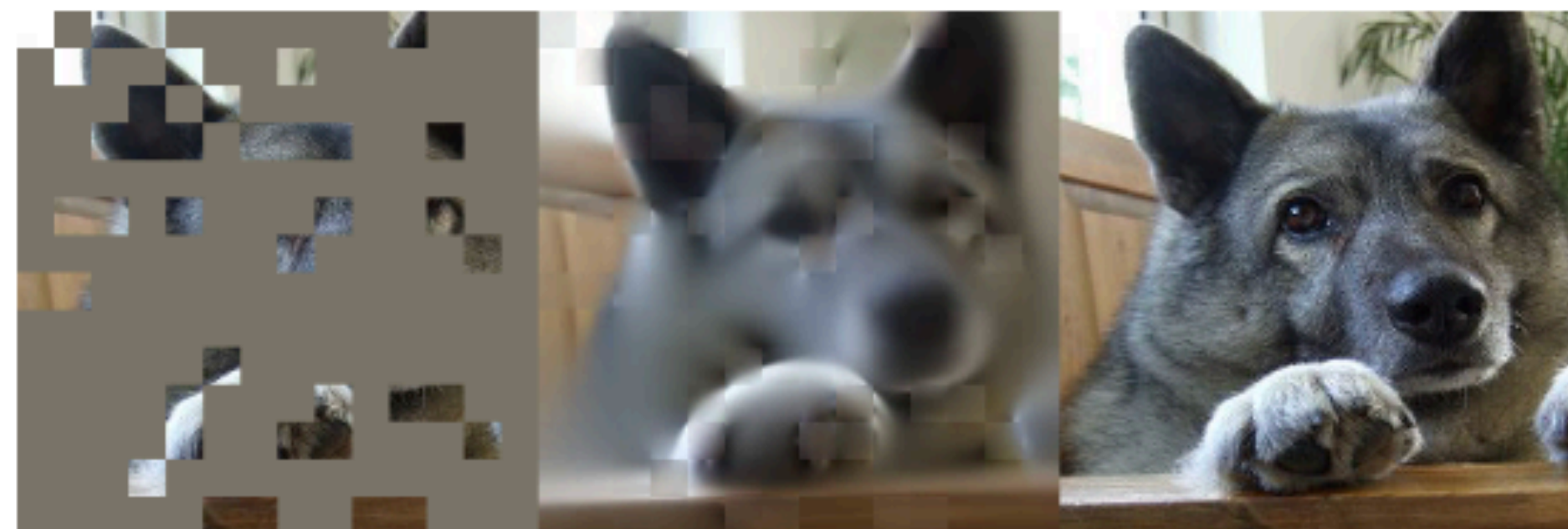
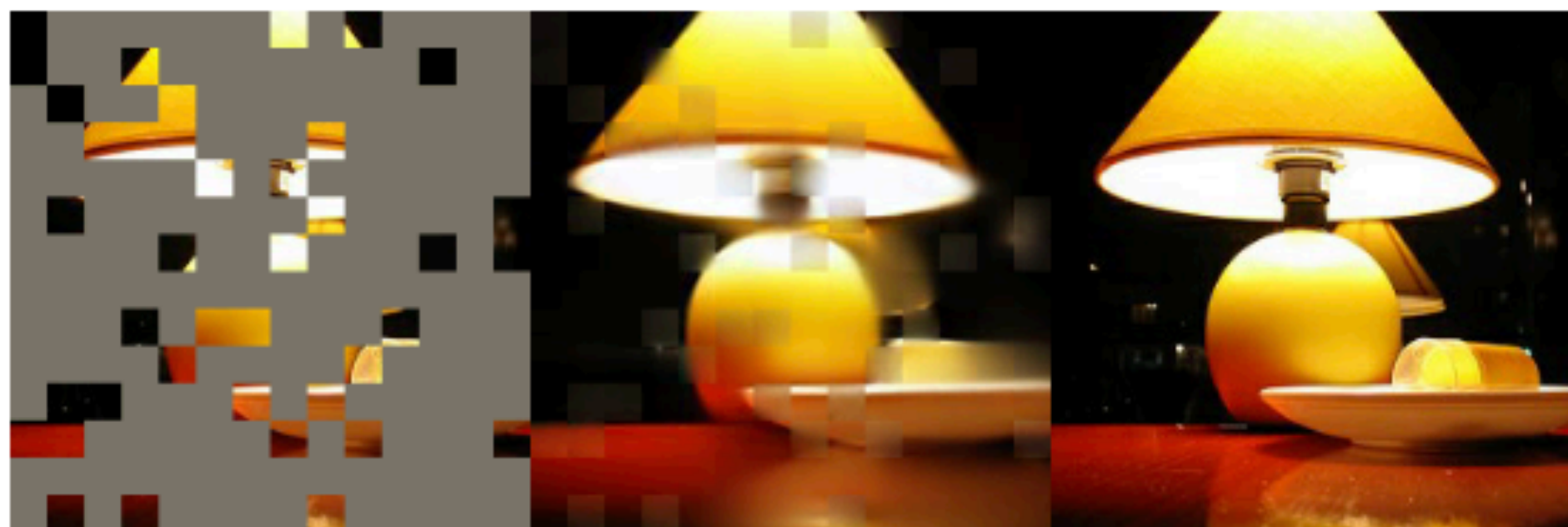
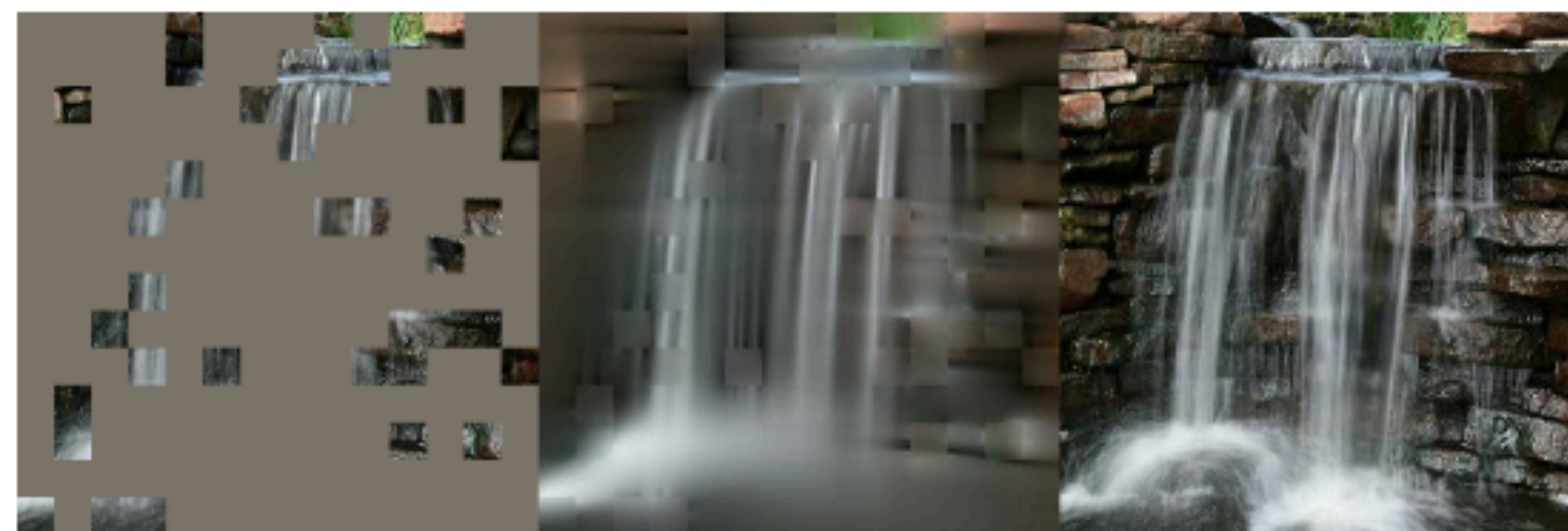
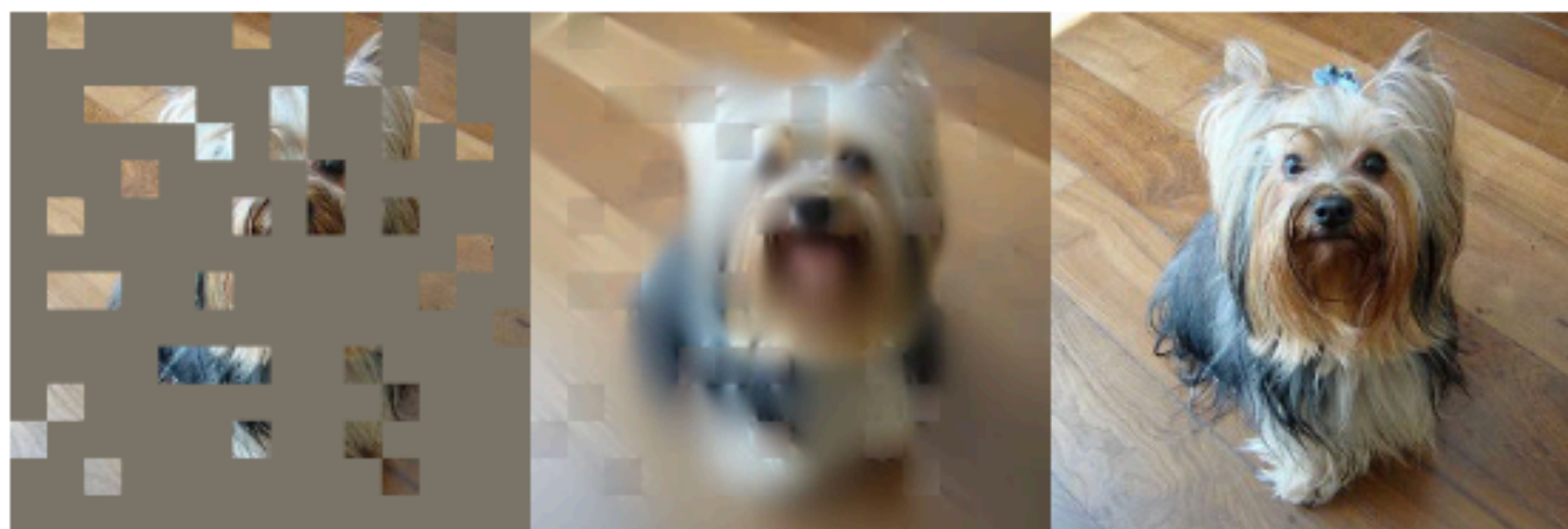
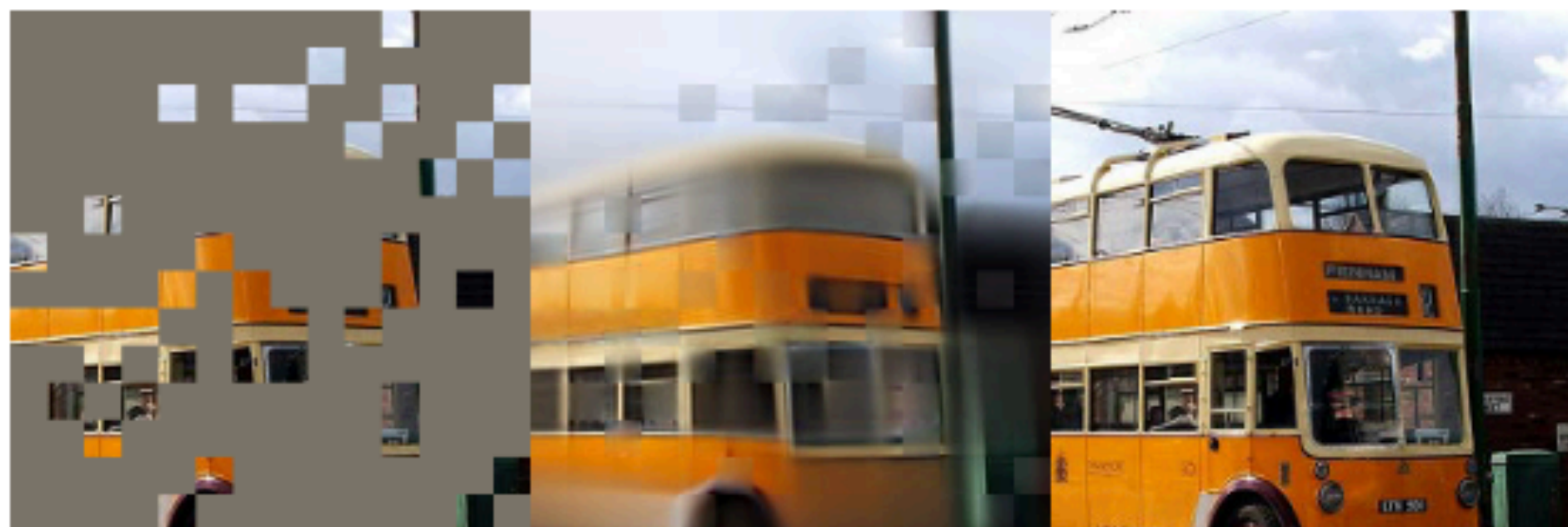
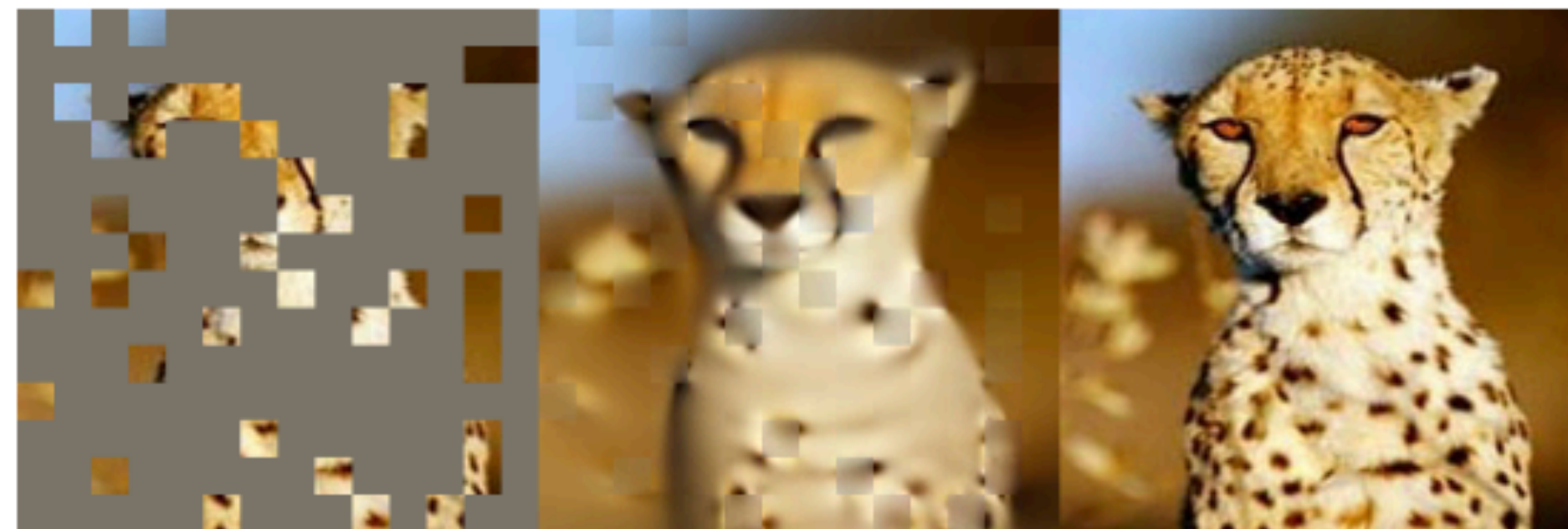
Masked autoencoders with transformers



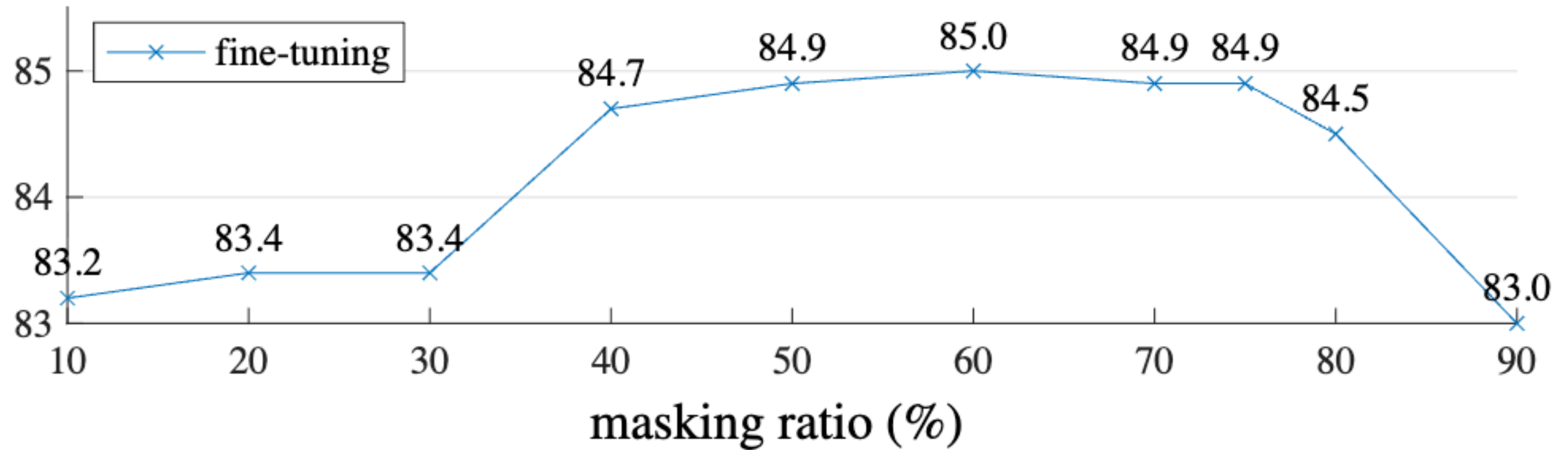
Autoencoder loss:

$$\mathcal{L}_{\theta} = \|\mathbf{X} - \hat{\mathbf{X}}\|^2$$



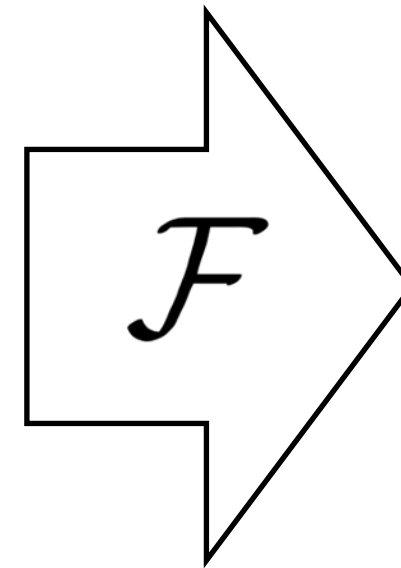
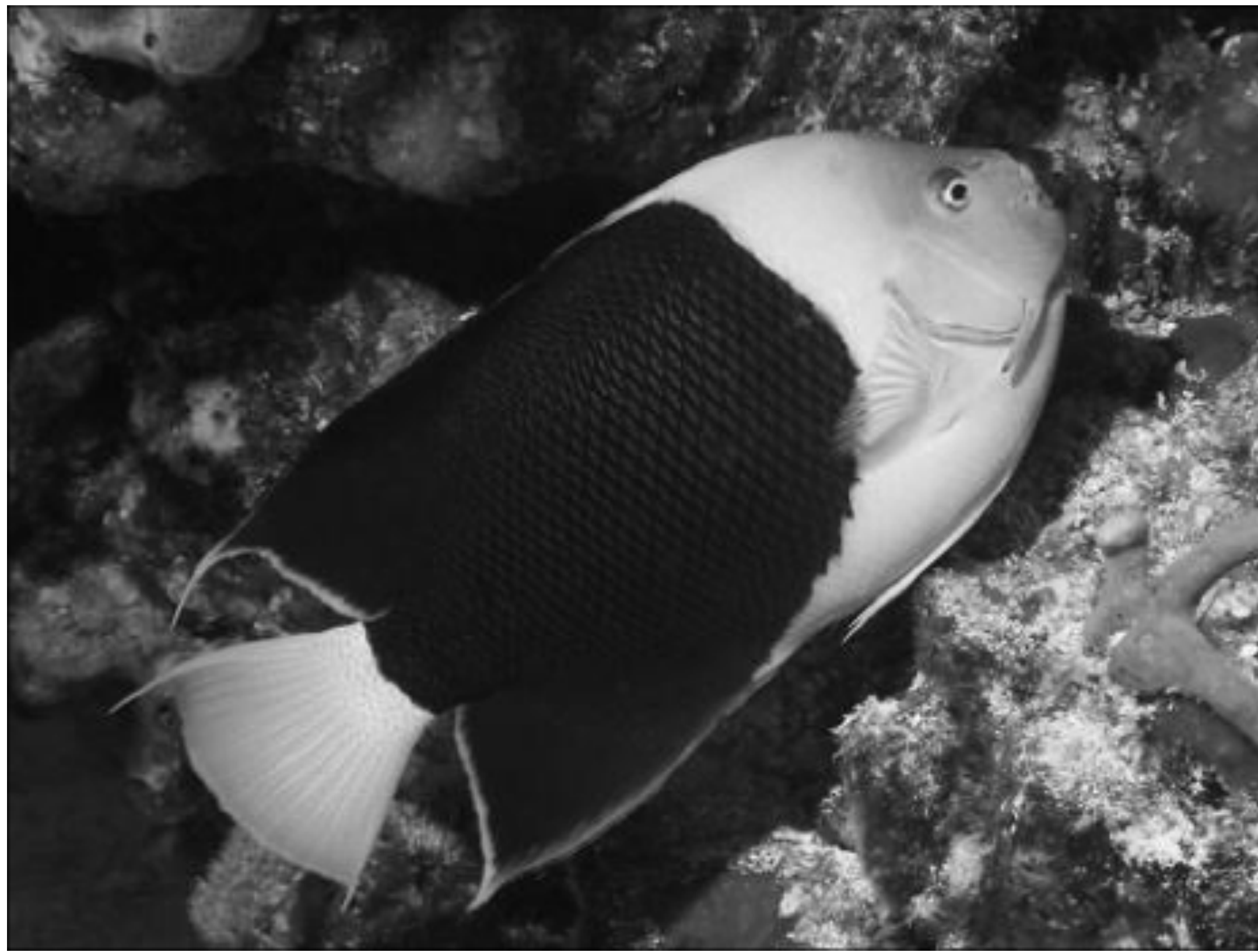


Feature learning performance



Downstream ImageNet recognition performance

Other self-supervised learning tasks

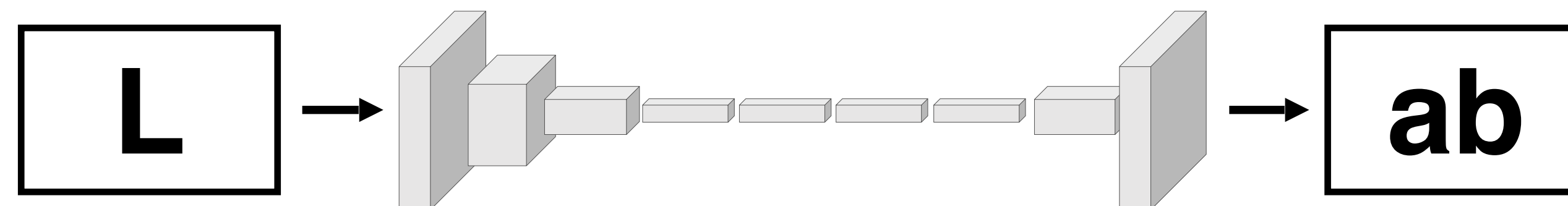


Grayscale image: L channel

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

Color information: ab channels

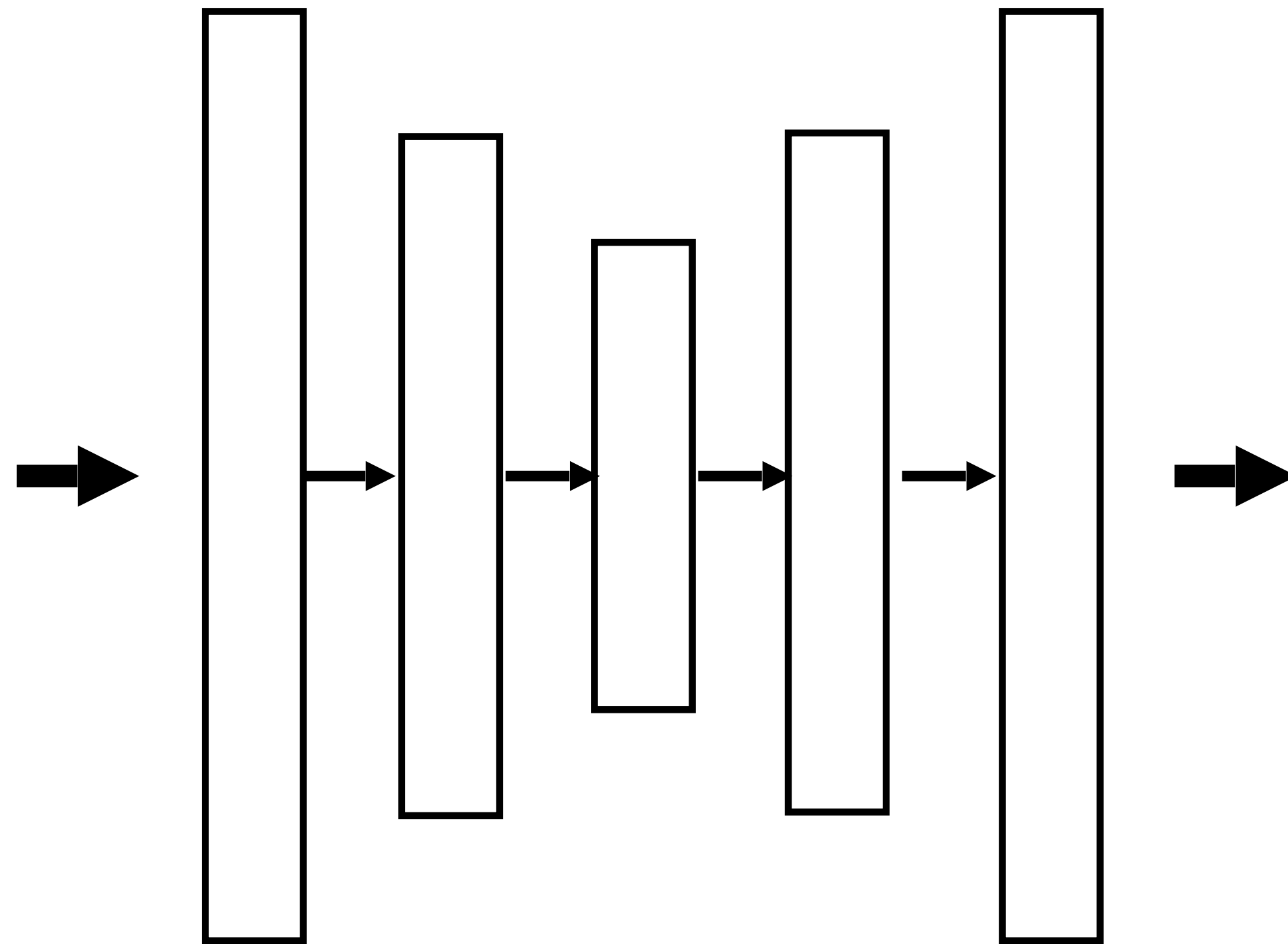
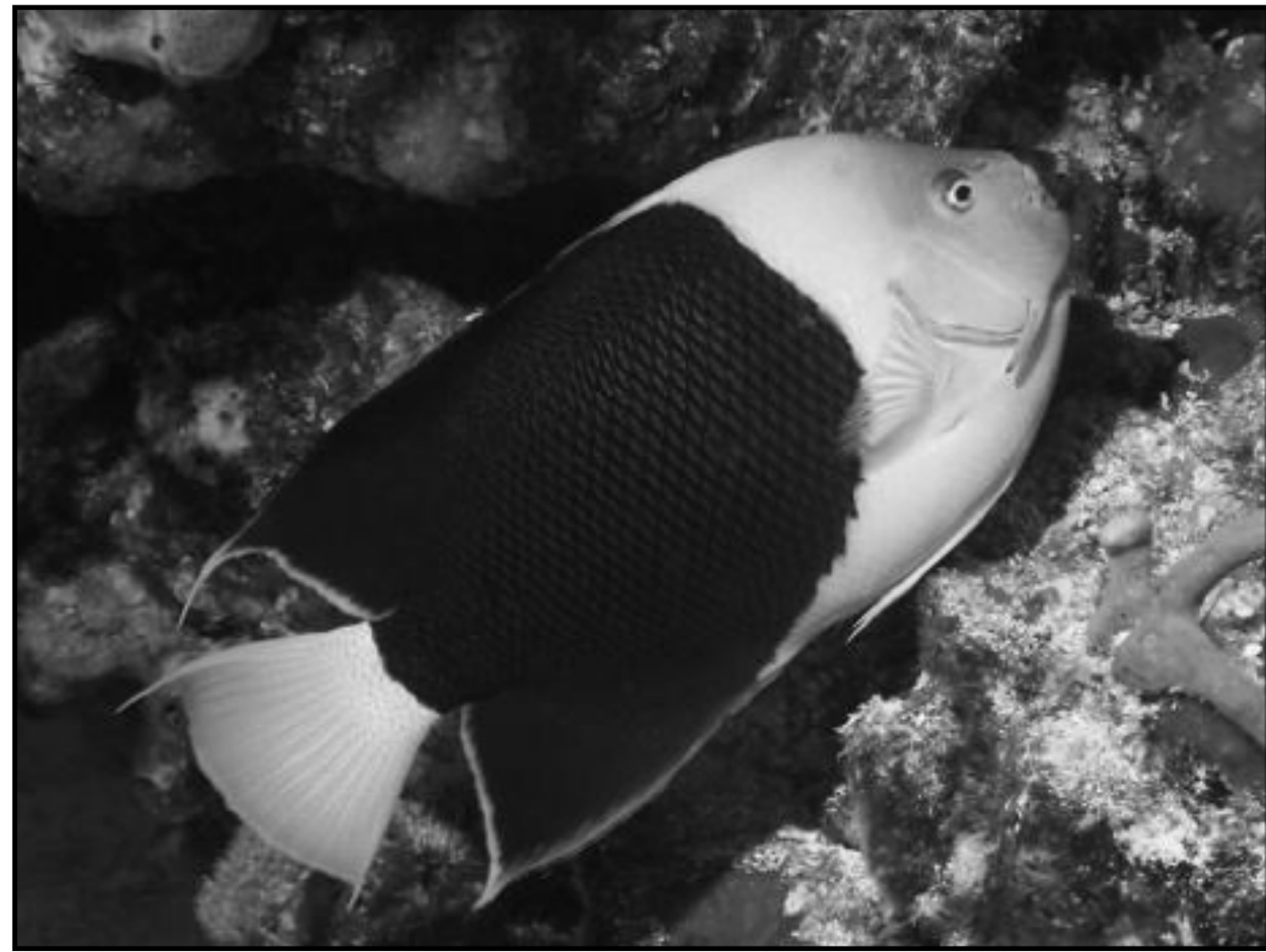
$$\hat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$$



48

[Zhang, Isola, Efros, ECCV 2016]

Visualizing units



49

[Zeiler & Fergus, ECCV 2014]

[Zhou et al., ICLR 2015]



Source: Isola, Torralba, Freeman

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

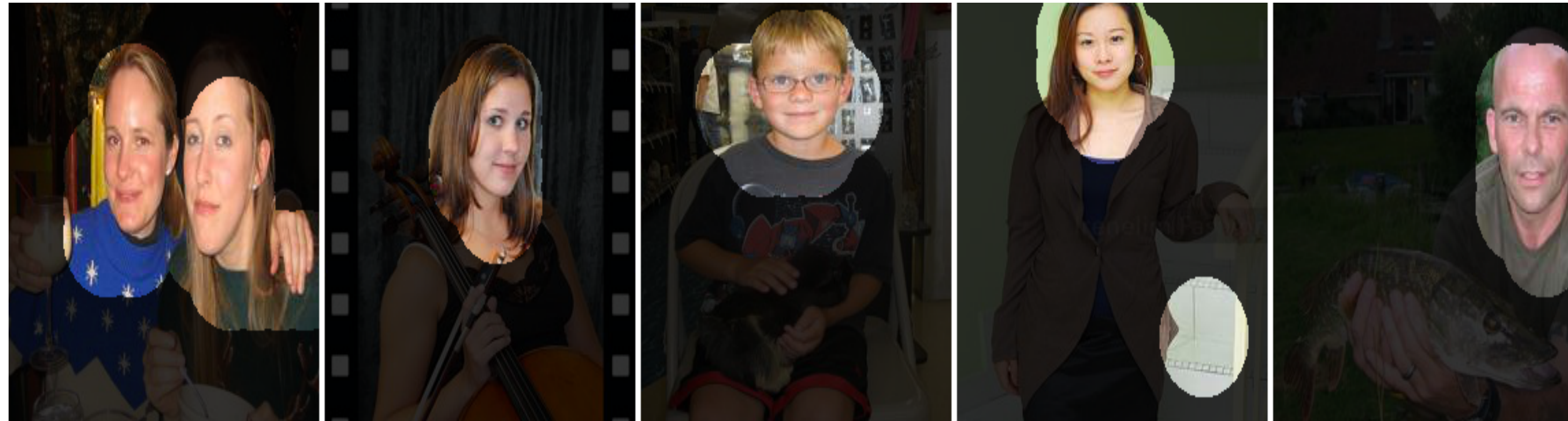


["Colorful image colorization", Zhang et al., ECCV 2016]

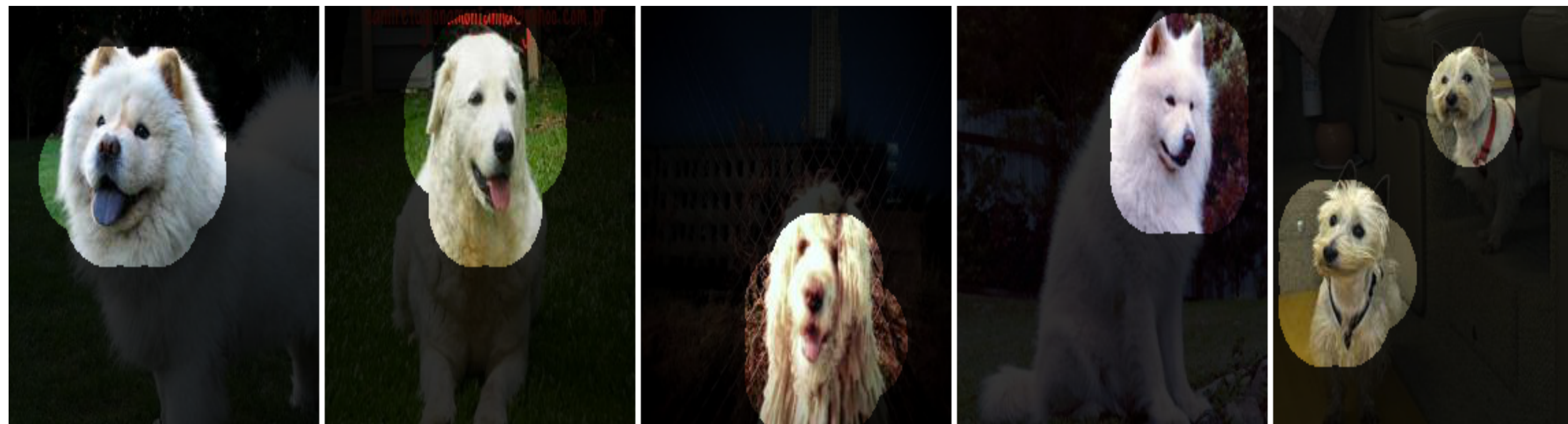


Stimuli that drive selected neurons (conv5 layer)

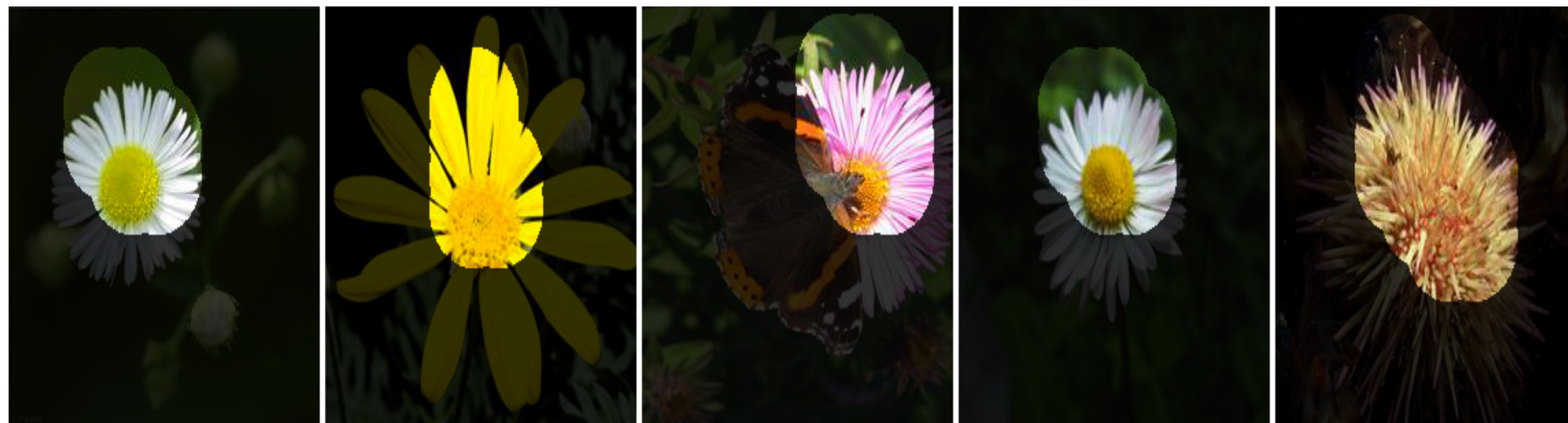
faces

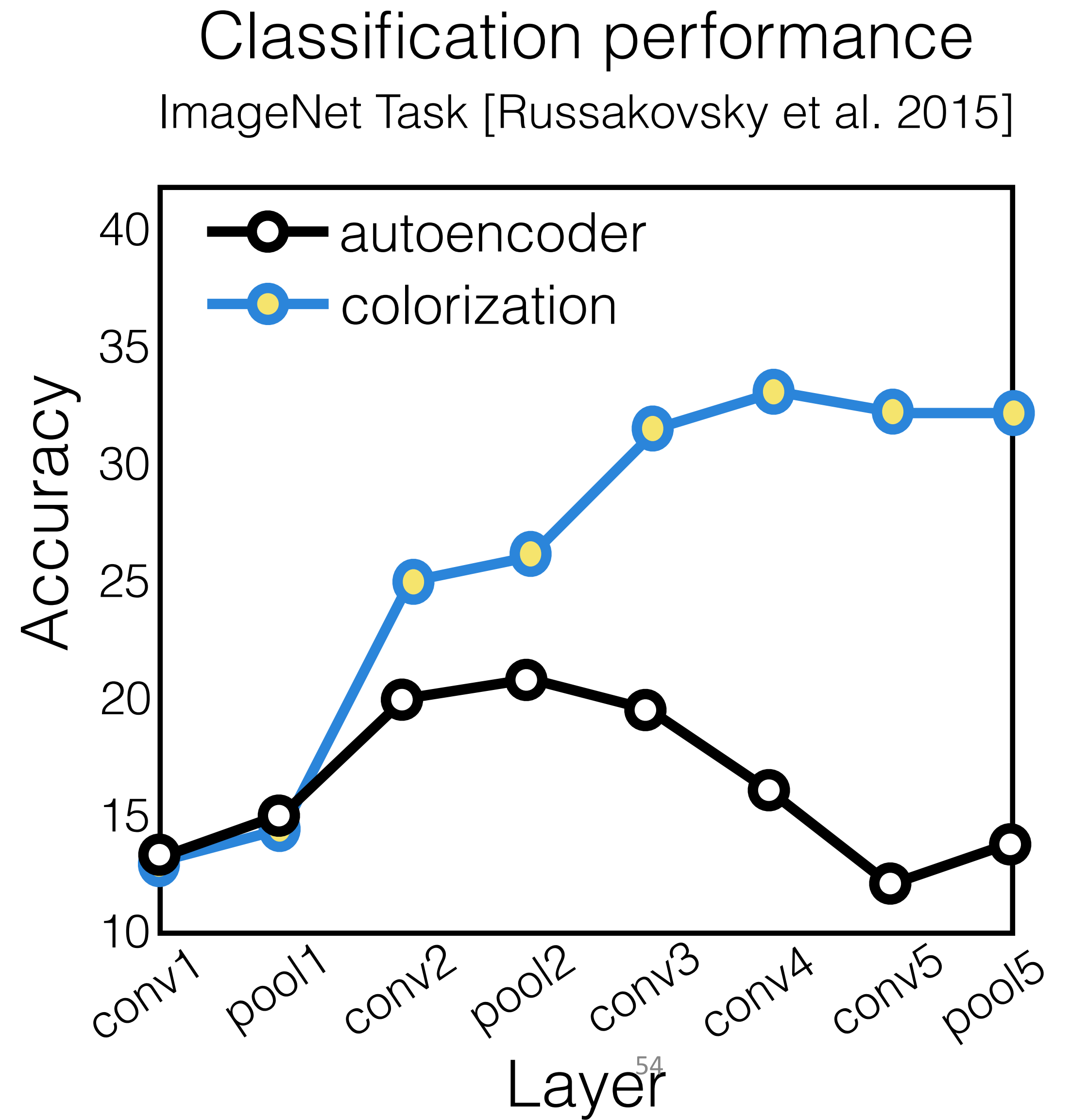
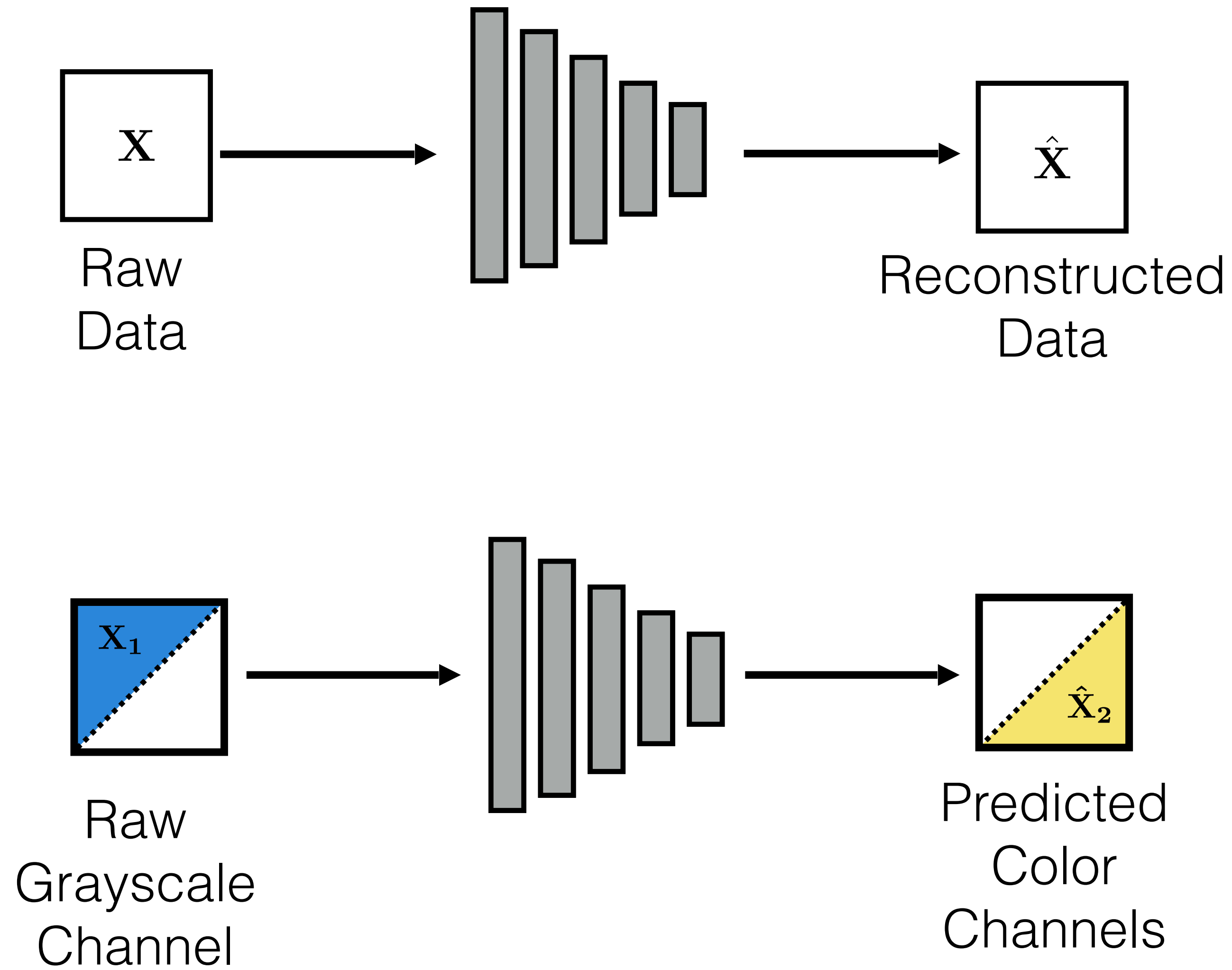


dog
faces



flowers





Context as Supervision

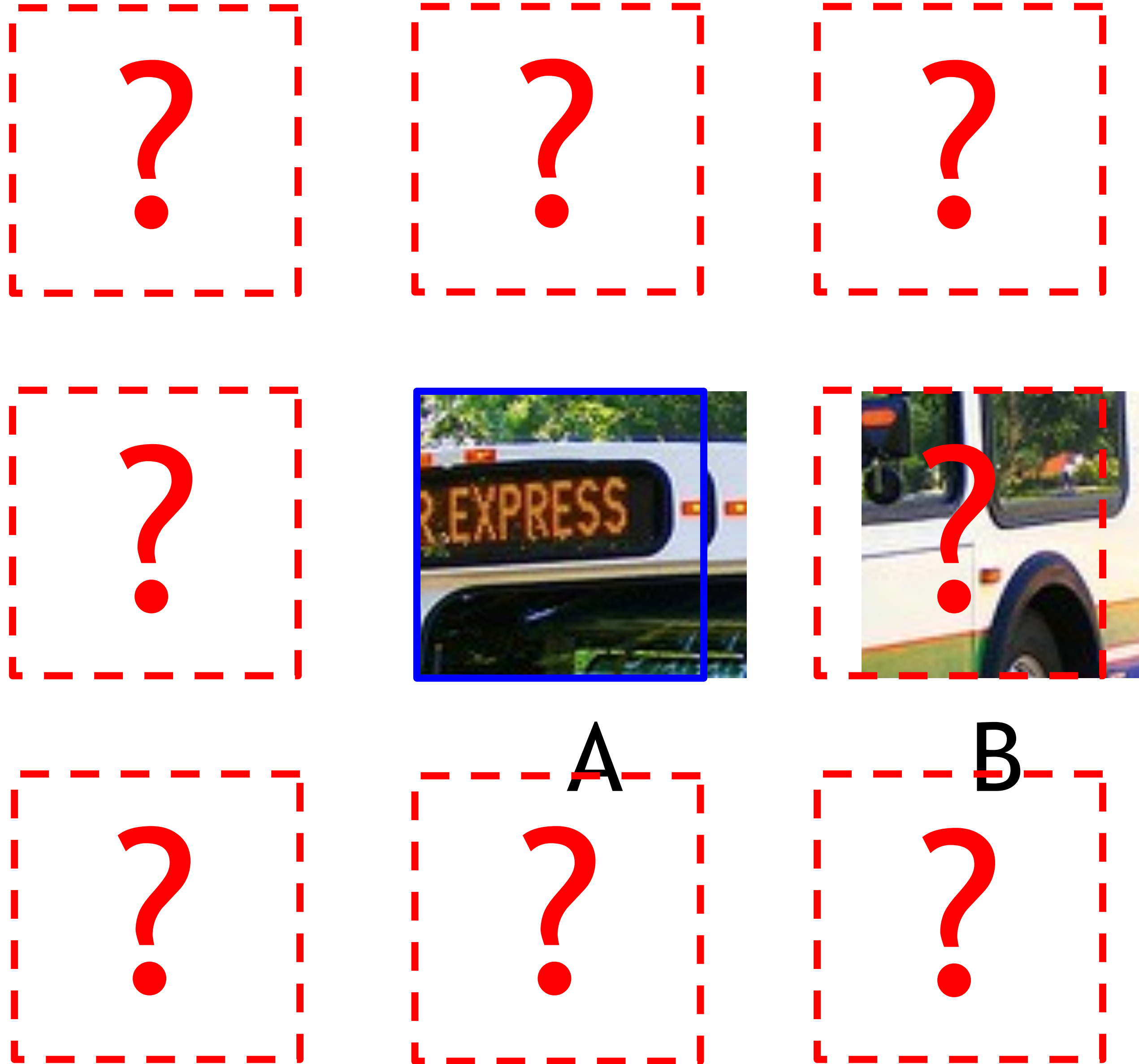
[Collobert & Weston 2008; Mikolov et al. 2013]

house, where the professor lived without his wife and child; or so he said jokingly sometimes: "Here's where I live. My house." His daughter often added, without resentment, for the visitor's information, "It started out to be for me, but it's really his." And she might reach in to bring forth an inch-high table lamp with fluted shade, or a blue dish the size of her little fingernail, marked "Kitty" and half full of eternal milk, but she was sure to replace these, after they had been admired, pretty near exactly where they had been. The little house was very orderly, and just big enough for all it contained, though to some tastes the bric-à-brac in the parlor might seem excessive. The daughter's preference was for the store-bought gimmicks and appliances, the toasters and carpet sweepers of Lilliput, but she knew that most adult vis



Deep
Net

Context Prediction as Supervision

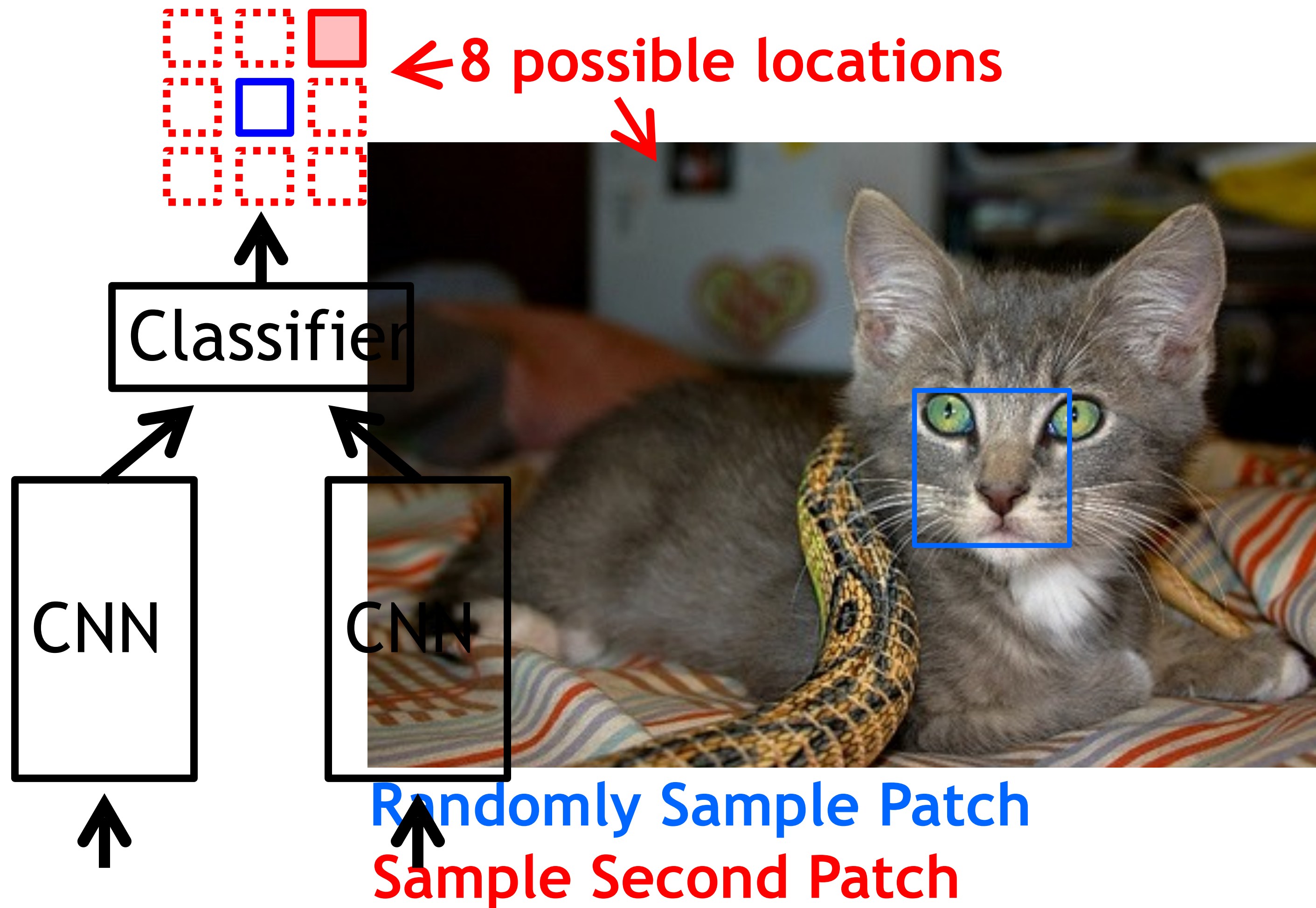


Semantics from a non-semantic task

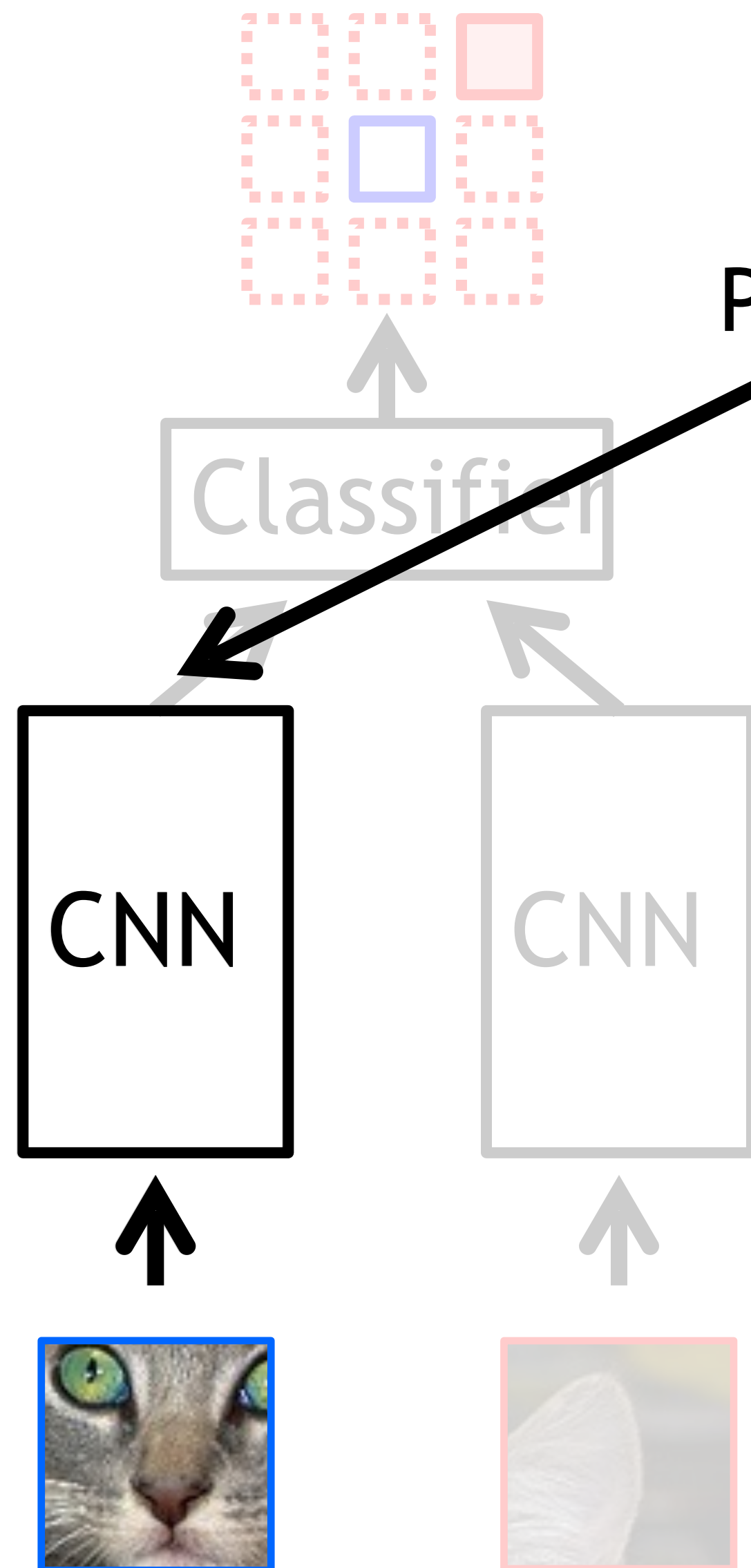


[Slide credit: Carl Doersch]

Relative Position Task



[Slide credit: Carl Doersch]



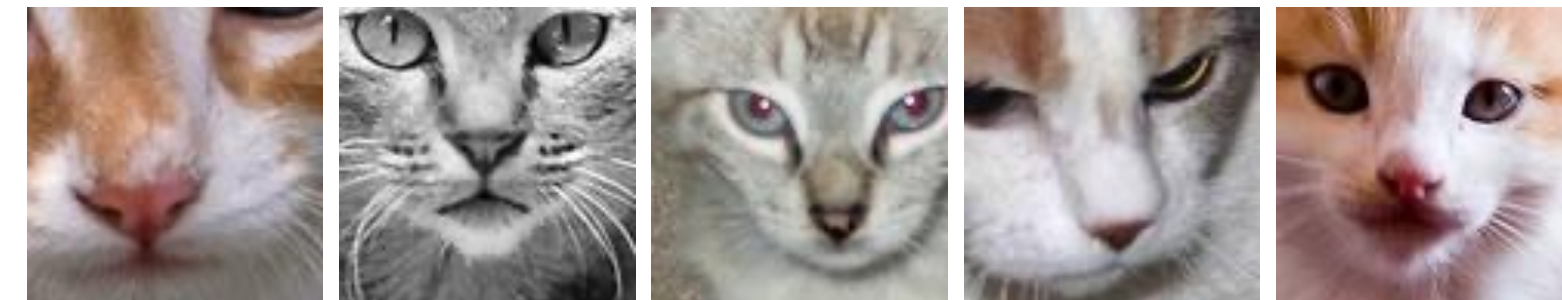
Patch Embedding (representation)

Input



!

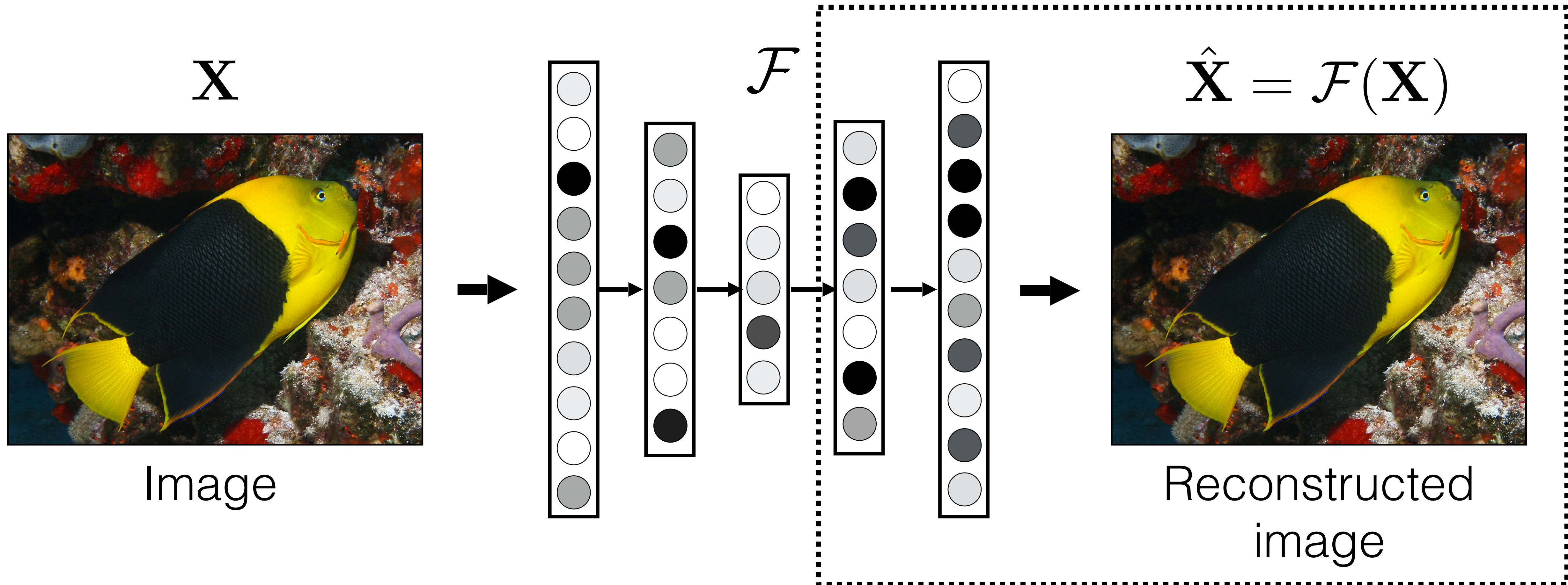
Nearest Neighbors



[Slide credit: Carl Doersch]

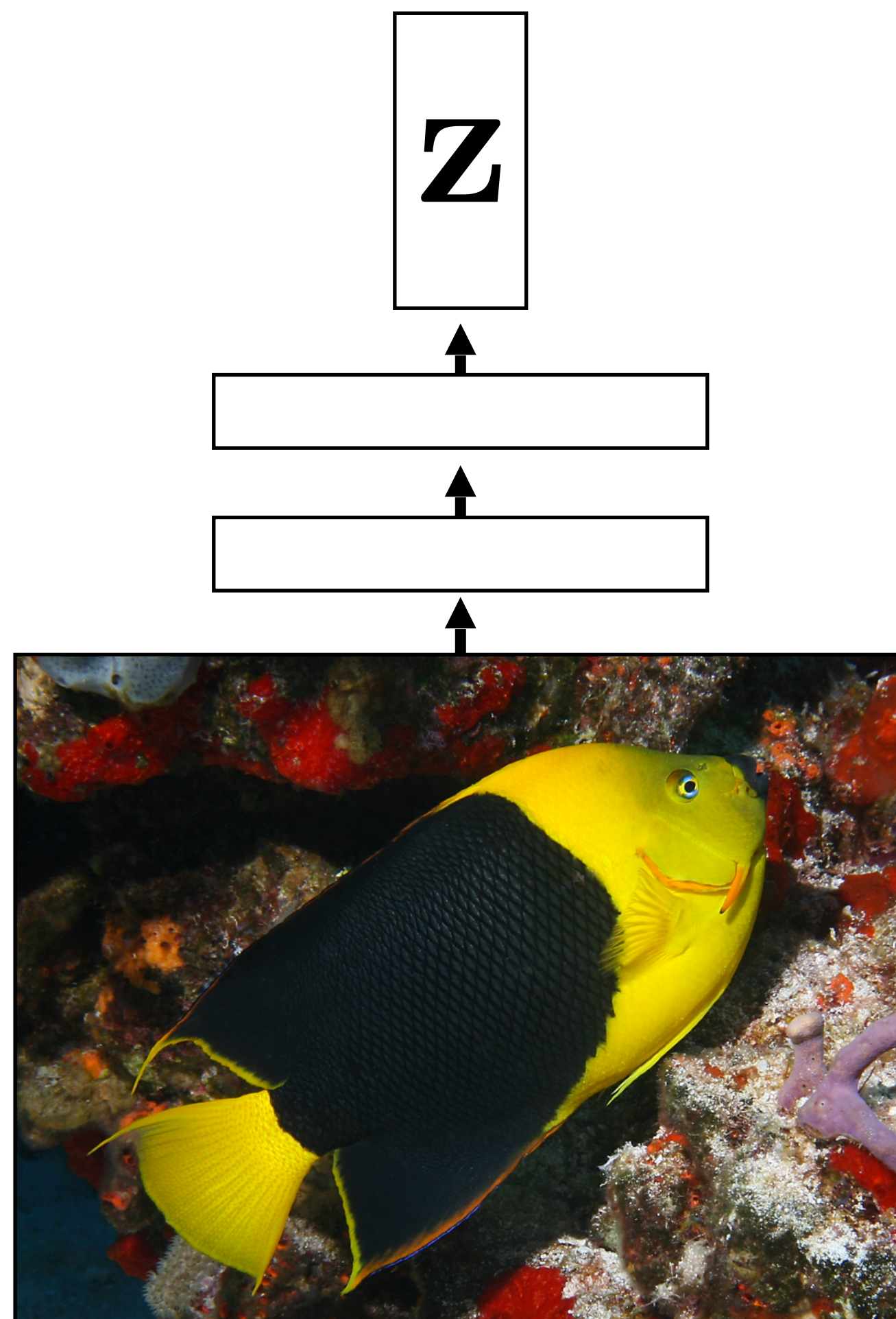
Revisiting autoencoders

Is prediction necessary?

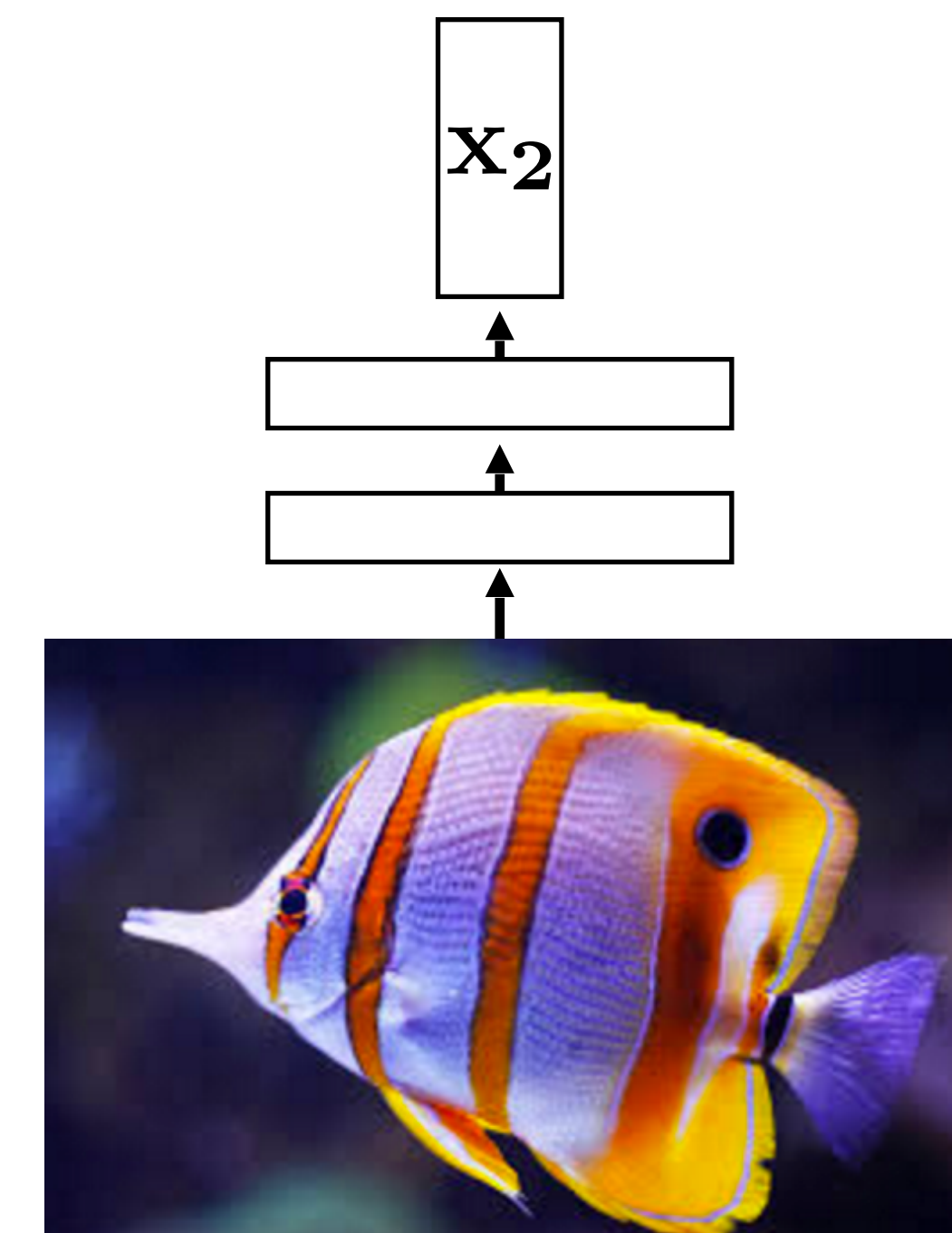
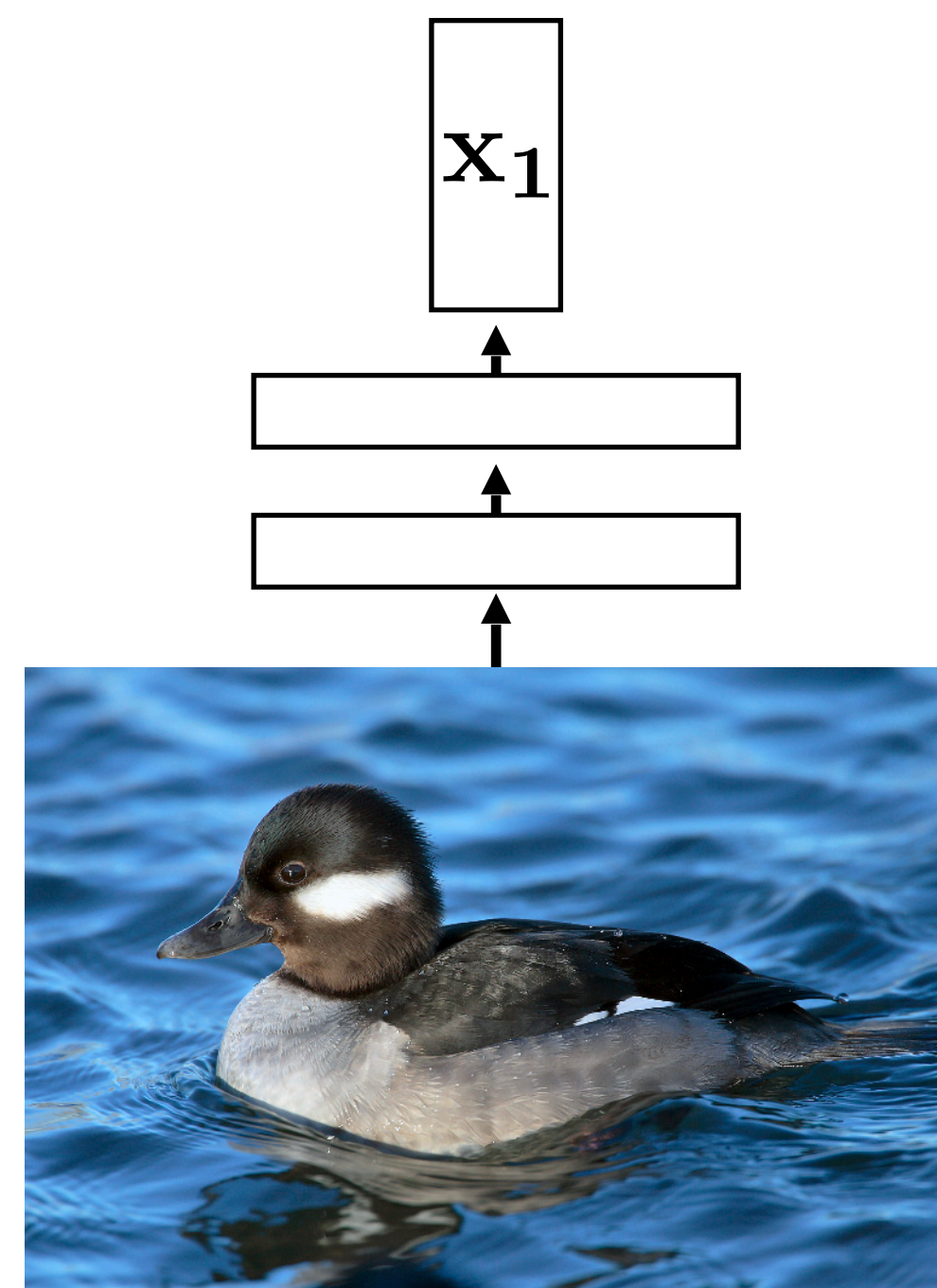


Contrastive learning

Feature embedding



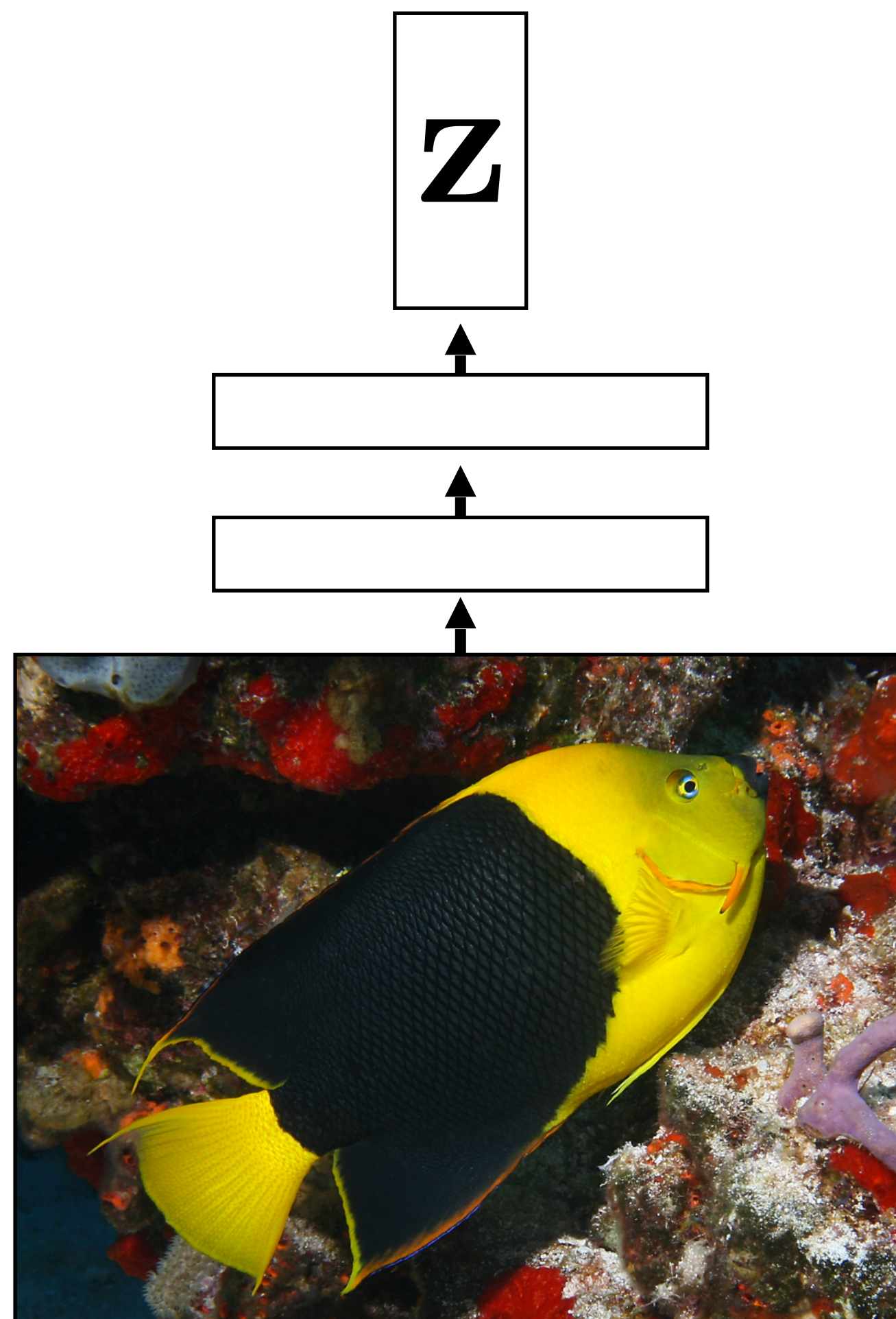
Other images in dataset



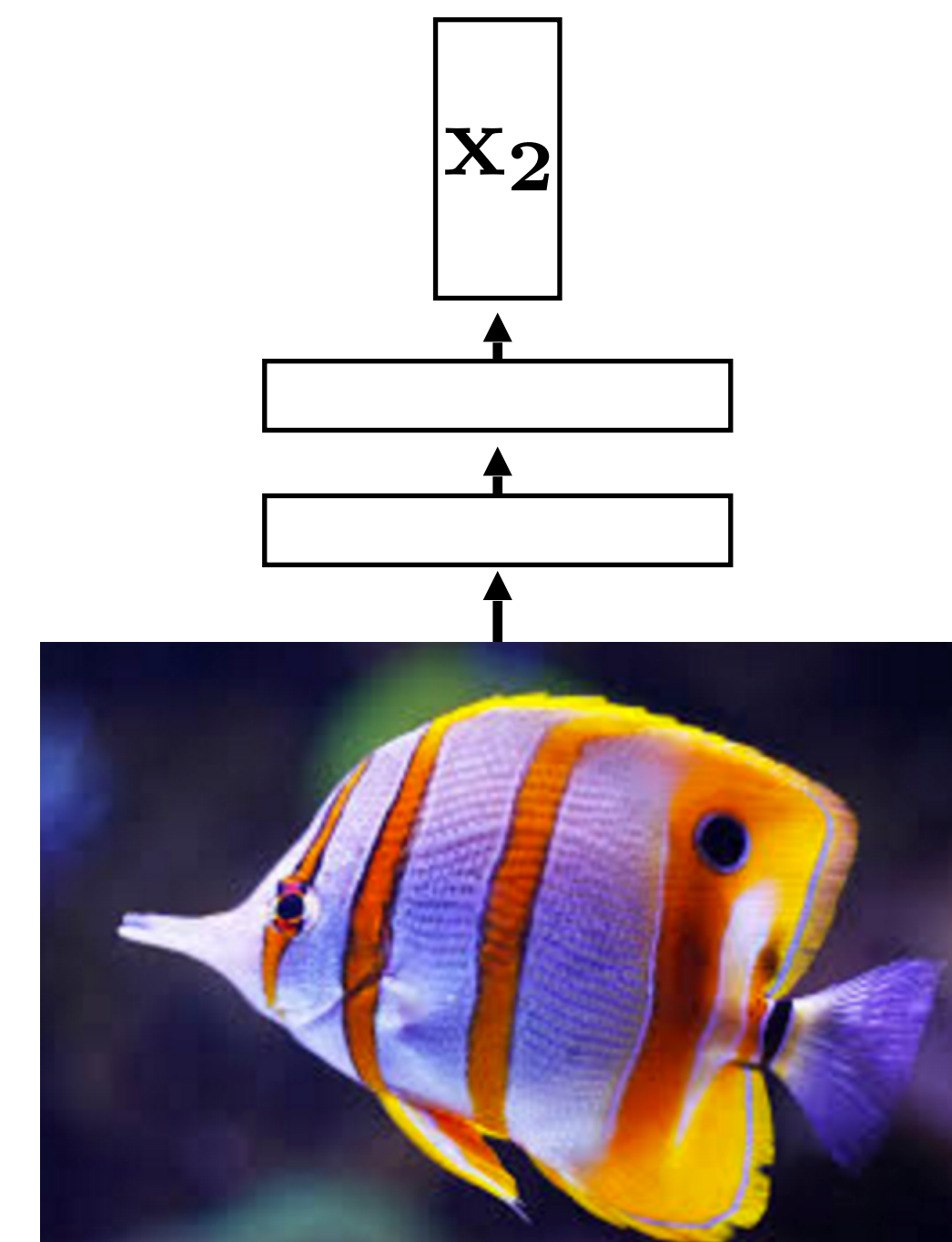
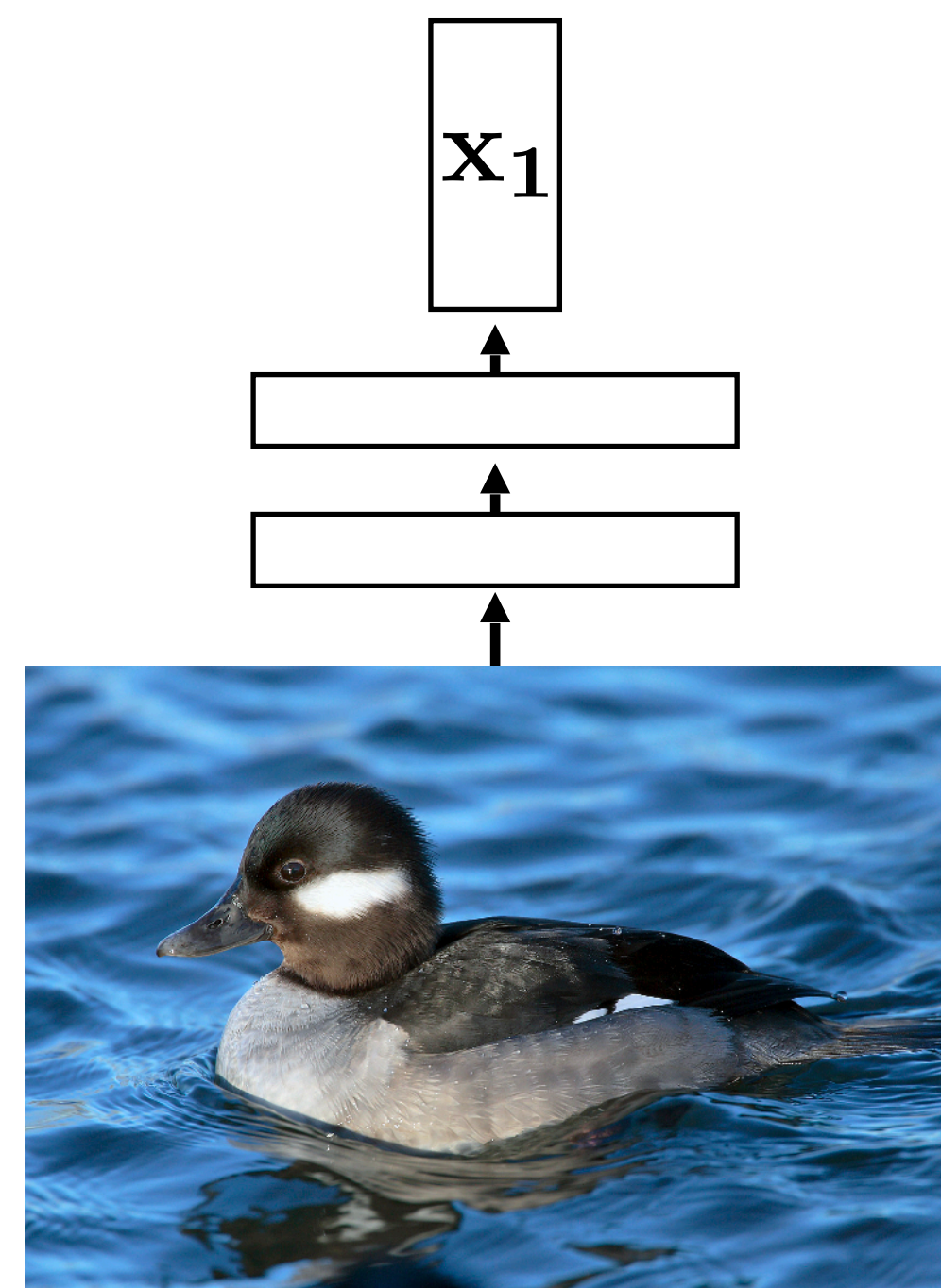
...

Contrastive learning

Feature embedding

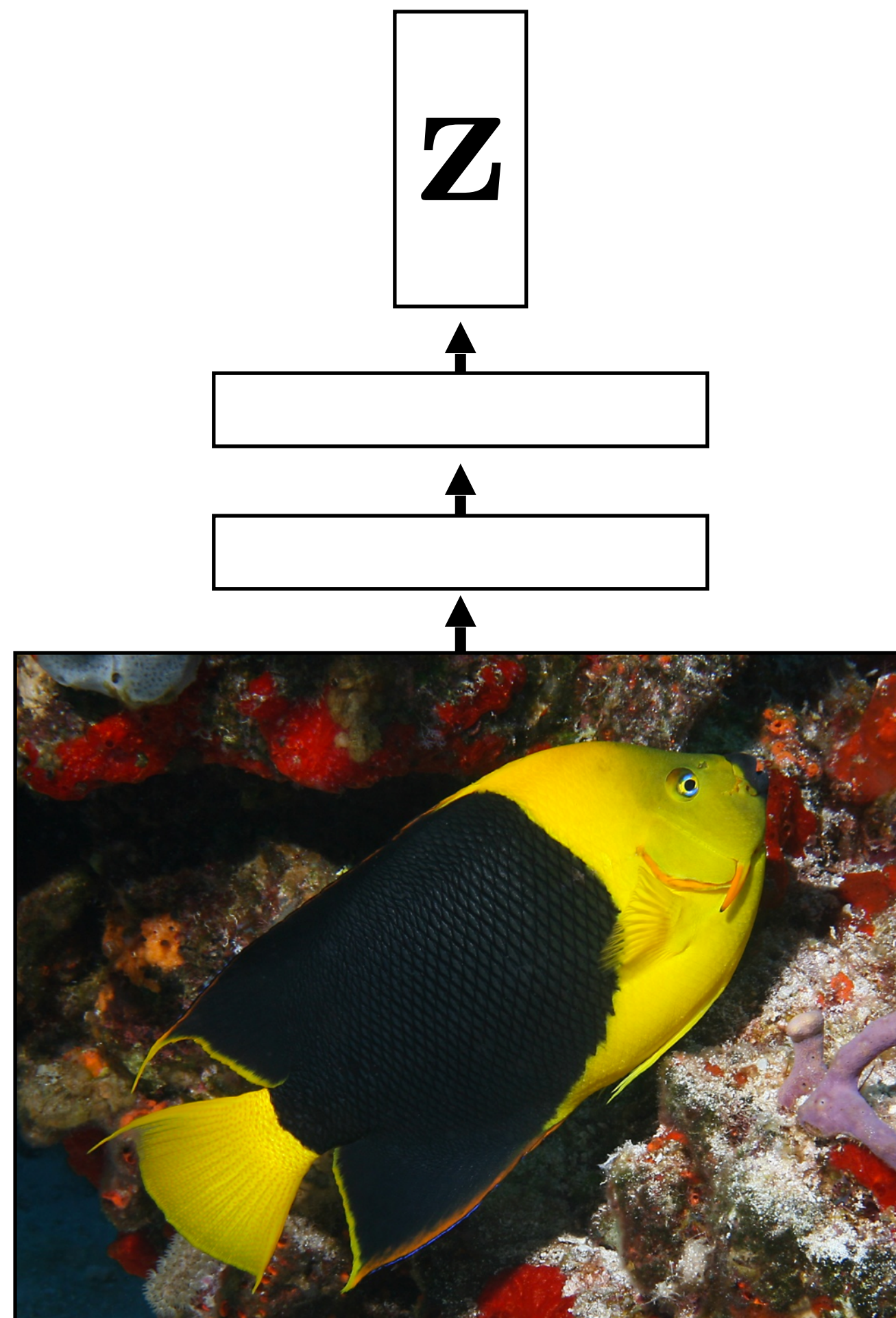


$\mathbf{z}^\top \mathbf{z} \rightarrow$ High dot product with self
 $\mathbf{z}^\top \mathbf{x}_1 \rightarrow$ Low dot product with others



...

Contrastive learning



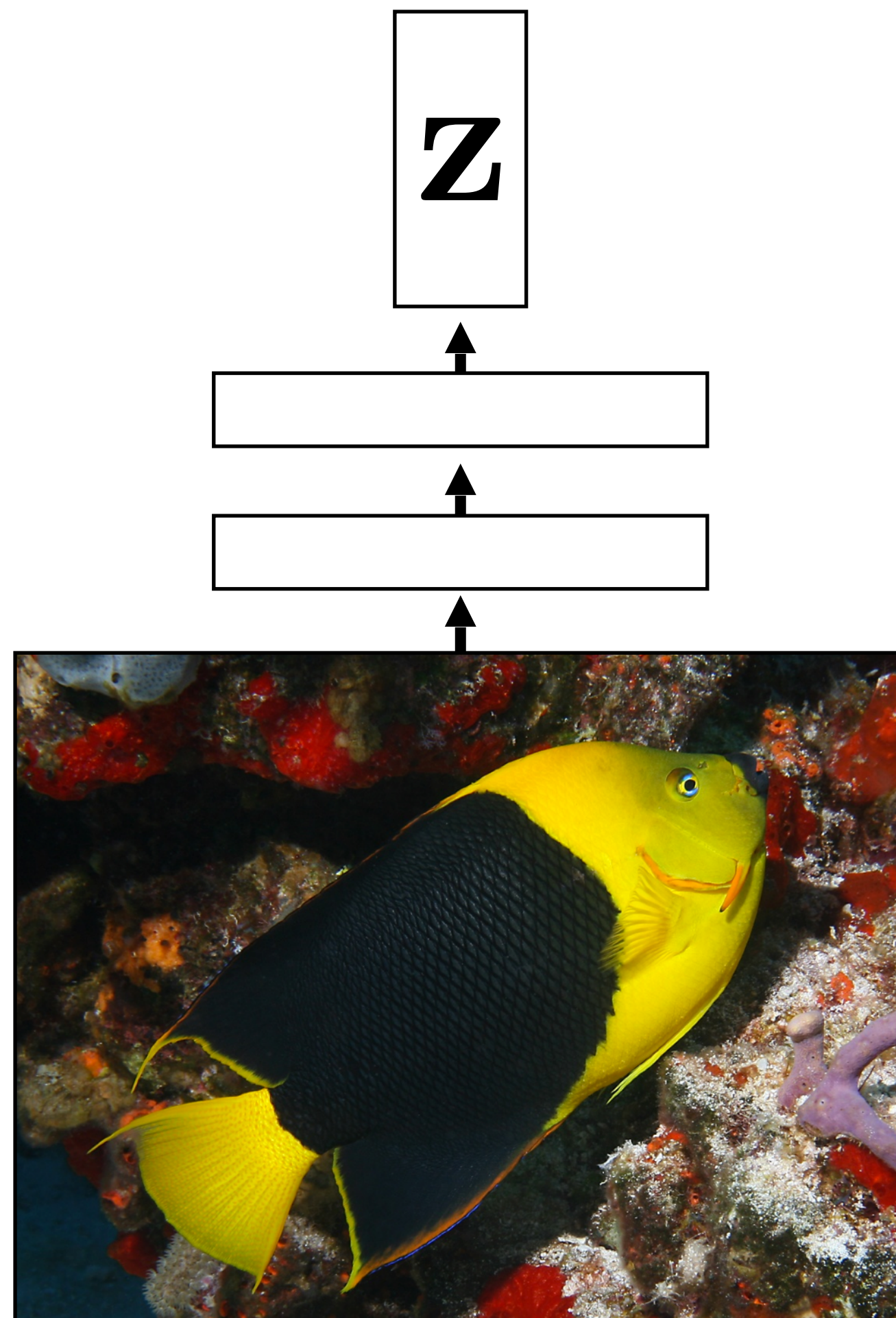
Minimize InfoNCE loss:

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

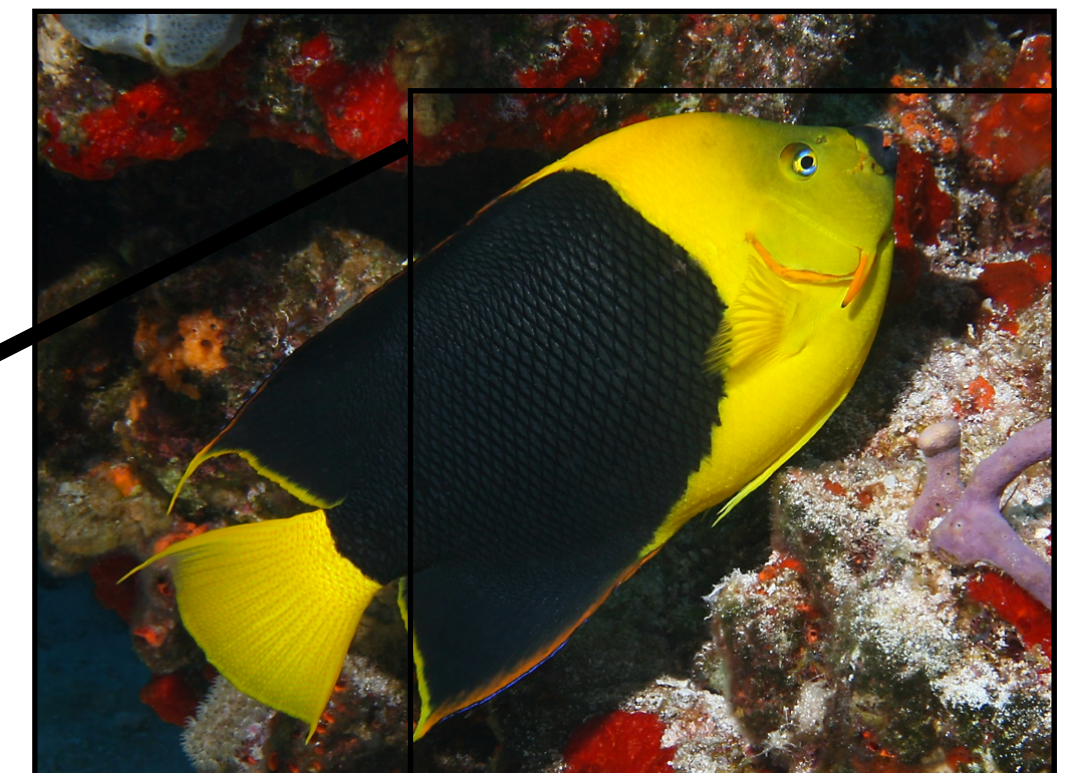
Equivalent to softmax loss with each image in the batch as a category.

Contrastive learning

Build in invariance by comparing to augmented images.



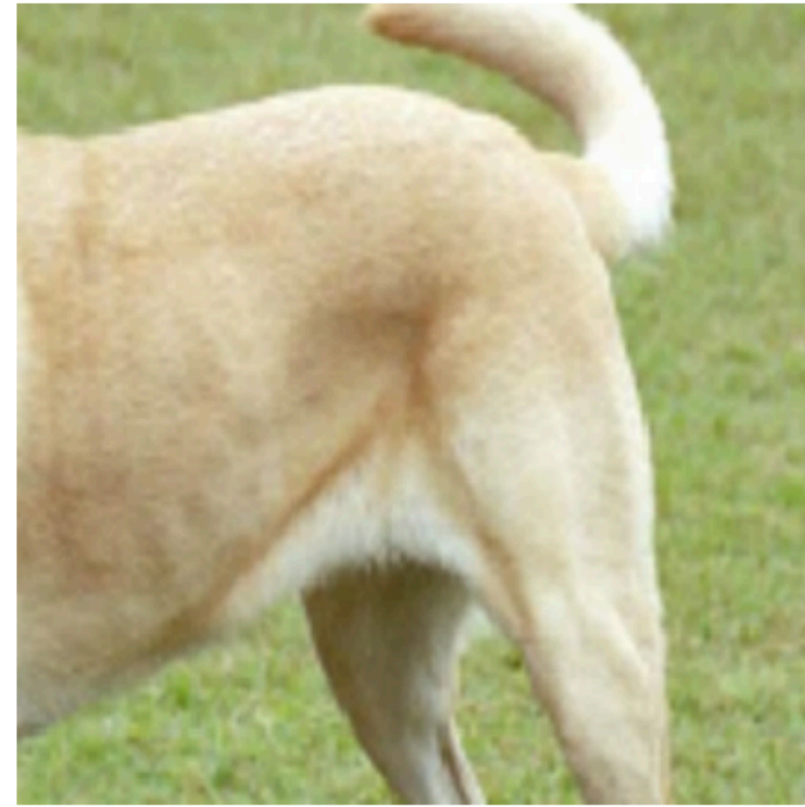
$$\mathcal{L} = -\log \left(\frac{\exp\{\mathbf{z}^\top \mathbf{z}_{aug}\}}{\exp\{\mathbf{z}^\top \mathbf{z}_{aug}\} + \sum_i \mathbf{z}^\top \mathbf{x}_i} \right)$$



Other augmentations for contrastive learning



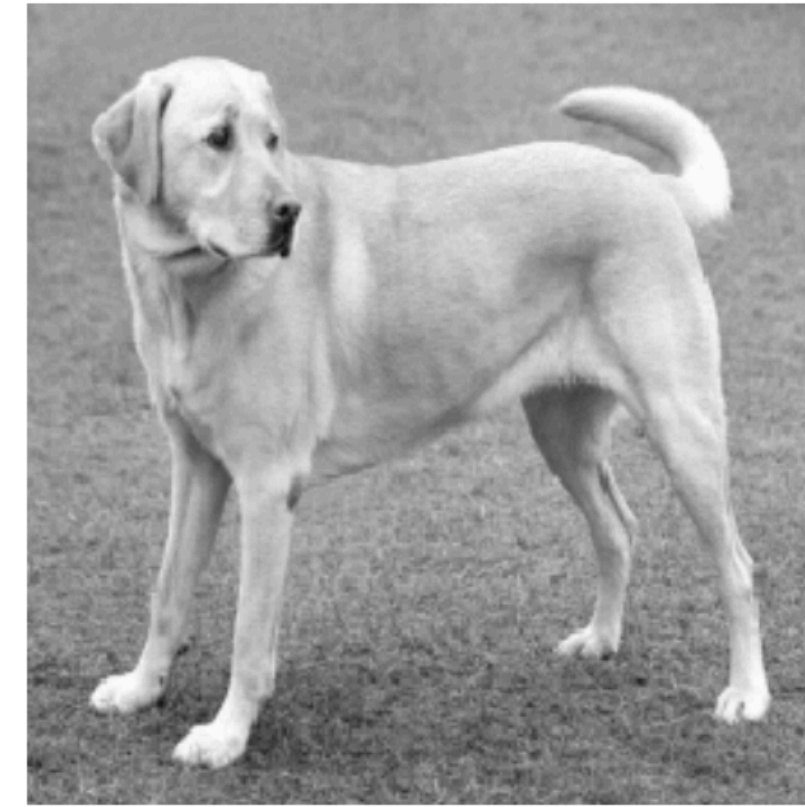
(a) Original



(b) Crop and resize



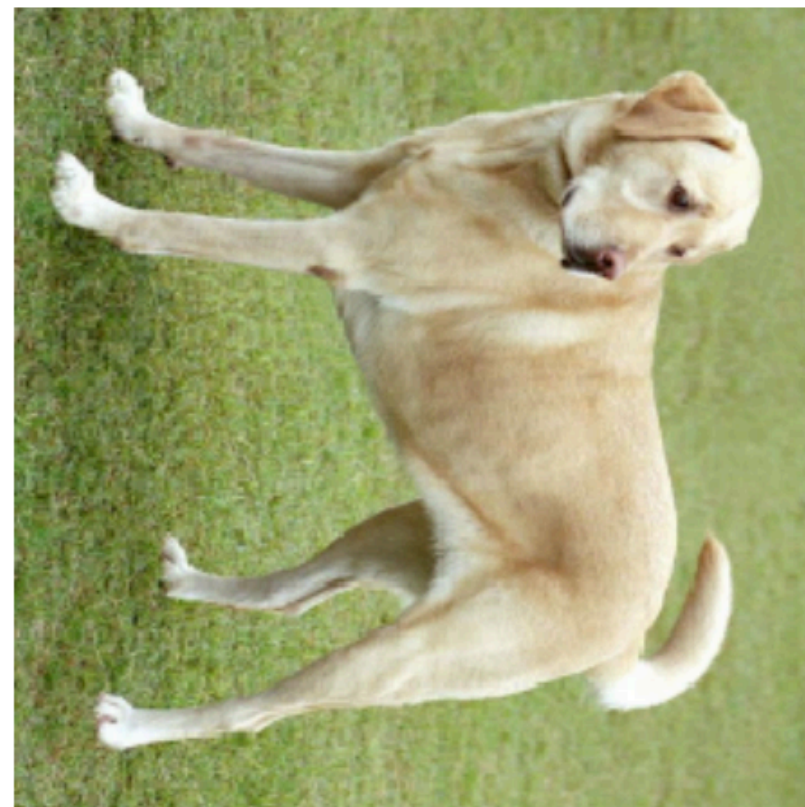
(c) Crop, resize (and flip)



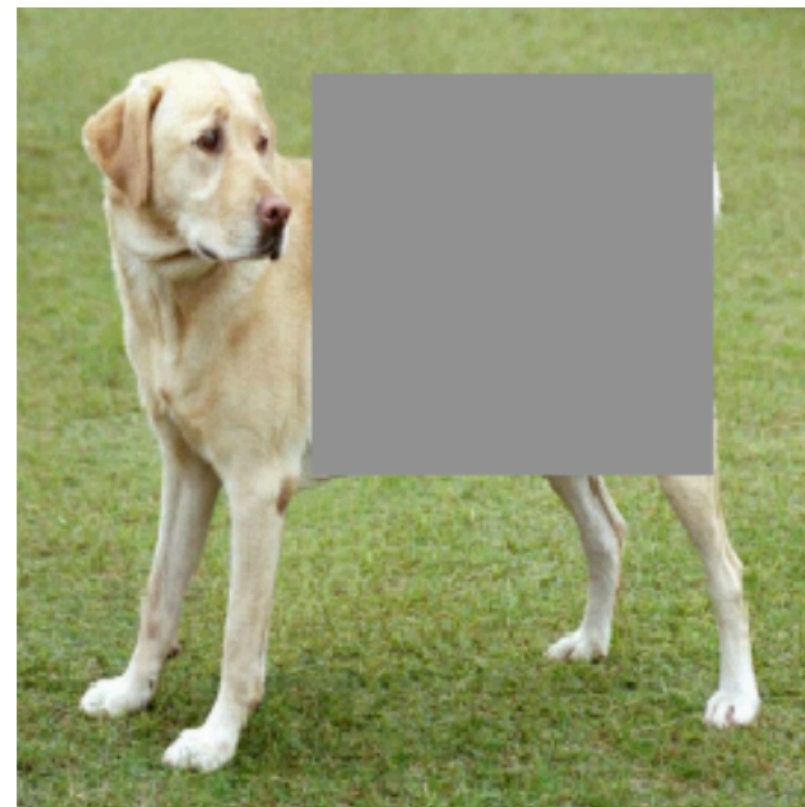
(d) Color distort. (drop)



(e) Color distort. (jitter)



(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$



(g) Cutout



(h) Gaussian noise

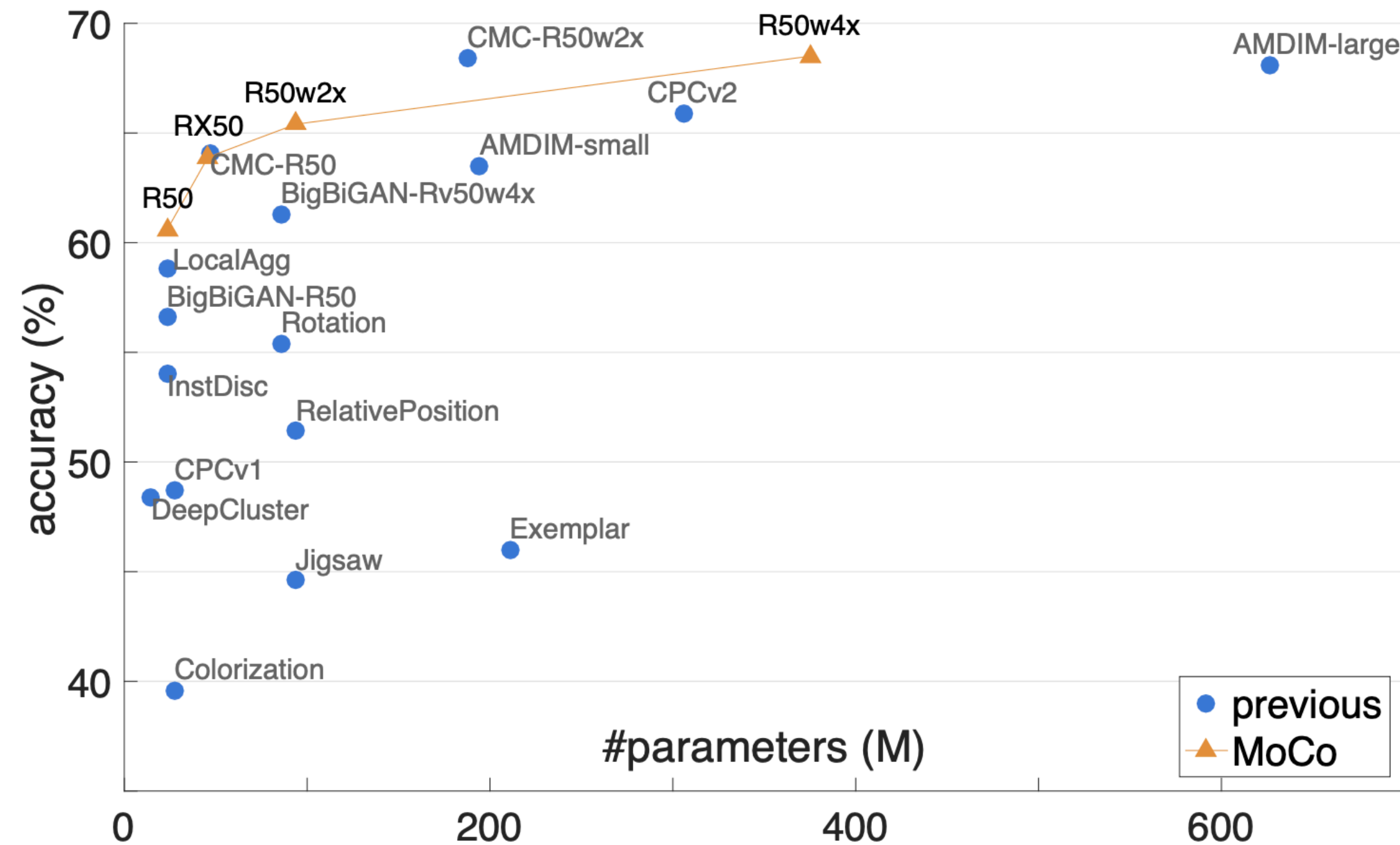


(i) Gaussian blur



(j) Sobel filtering

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.

Representation learning with language

Soap box derby



Cart



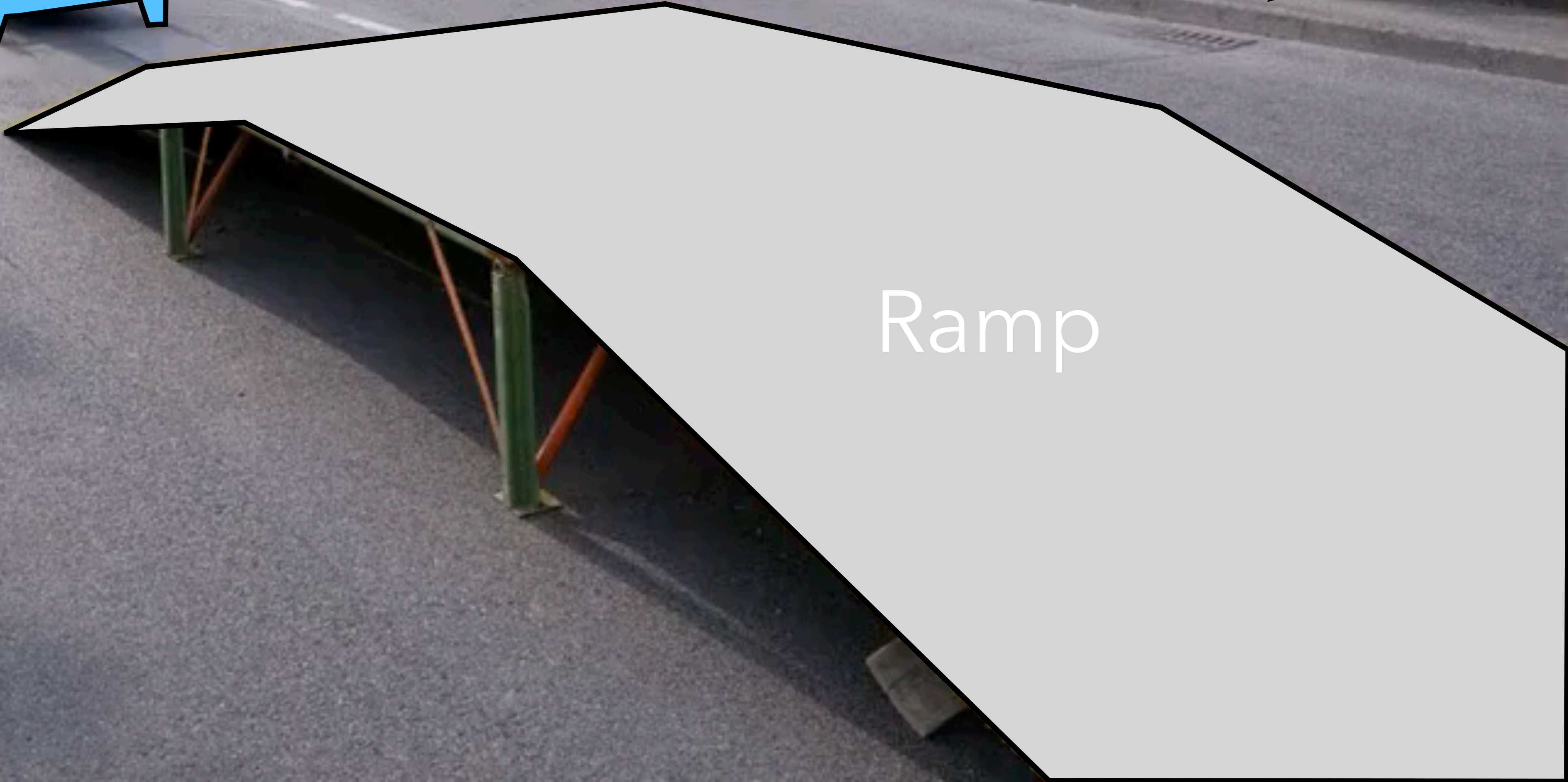
Bicycle



Person



Grate



Ramp

Language-based supervision



WIKIPEDIA

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 rolling resistance), and the use of solvents to swell the tires (also reducing rolling resistance). There is some overlap in technology between this race and other

- [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

Language-based supervision



WIKIPEDIA

Article Talk

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

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

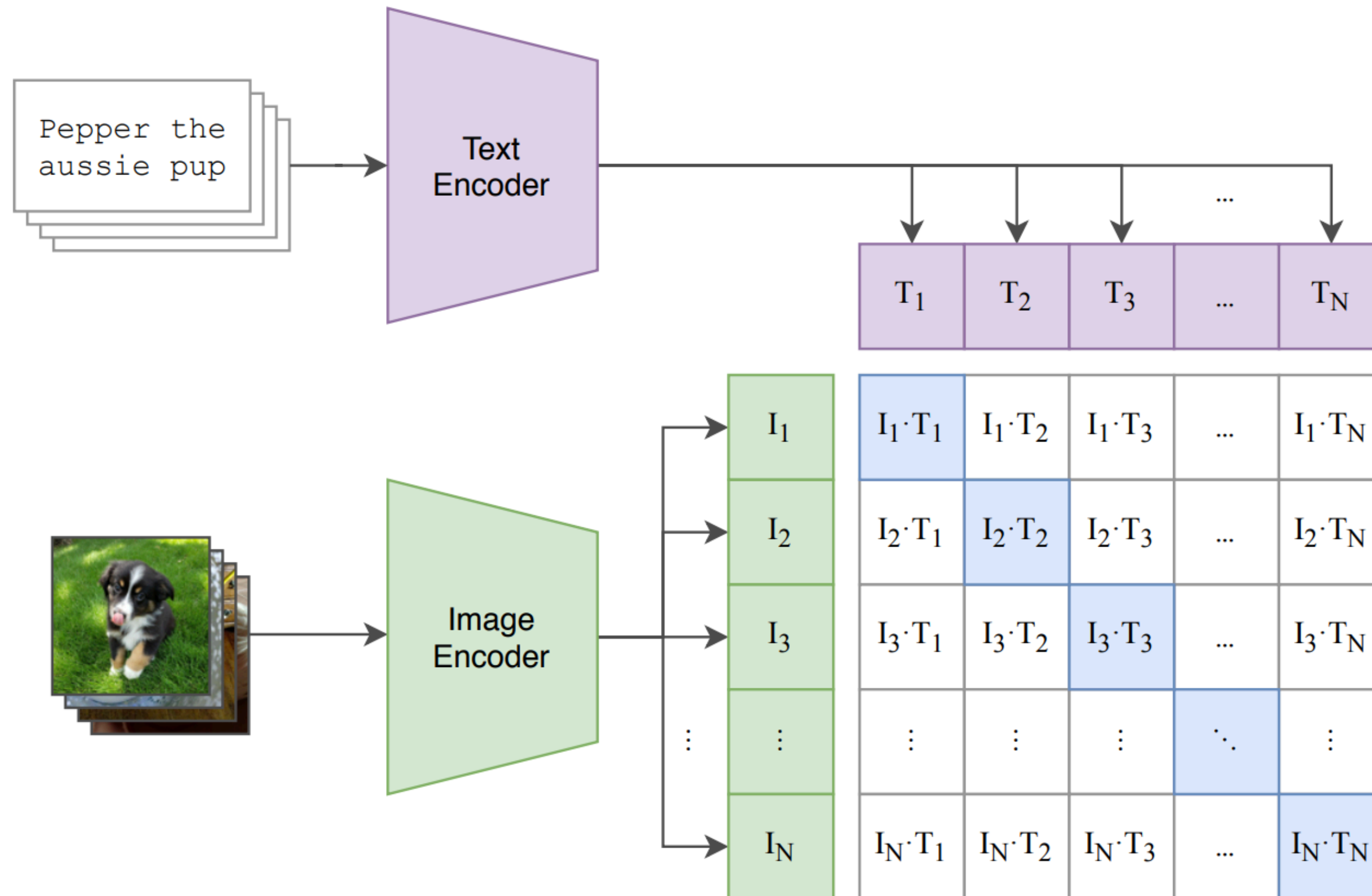
[edit]

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 rolling resistance), and the use of spherulites to swell the tires (also reducing rolling resistance). There is some overlap in technology between this race and other



Contrastive learning with language

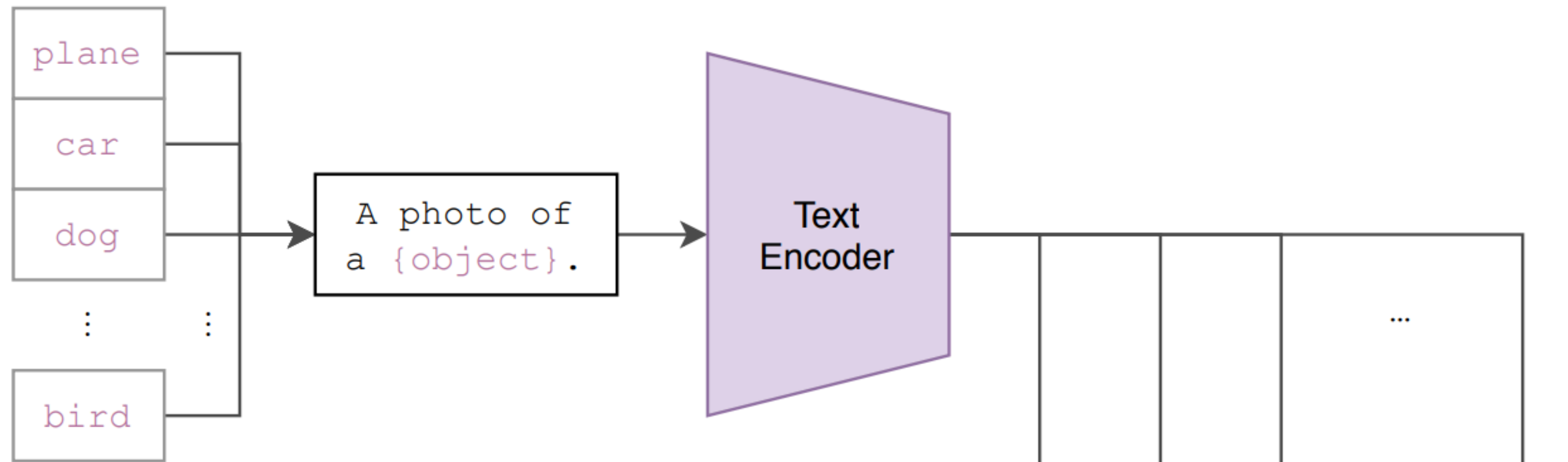


InfoNCE loss:

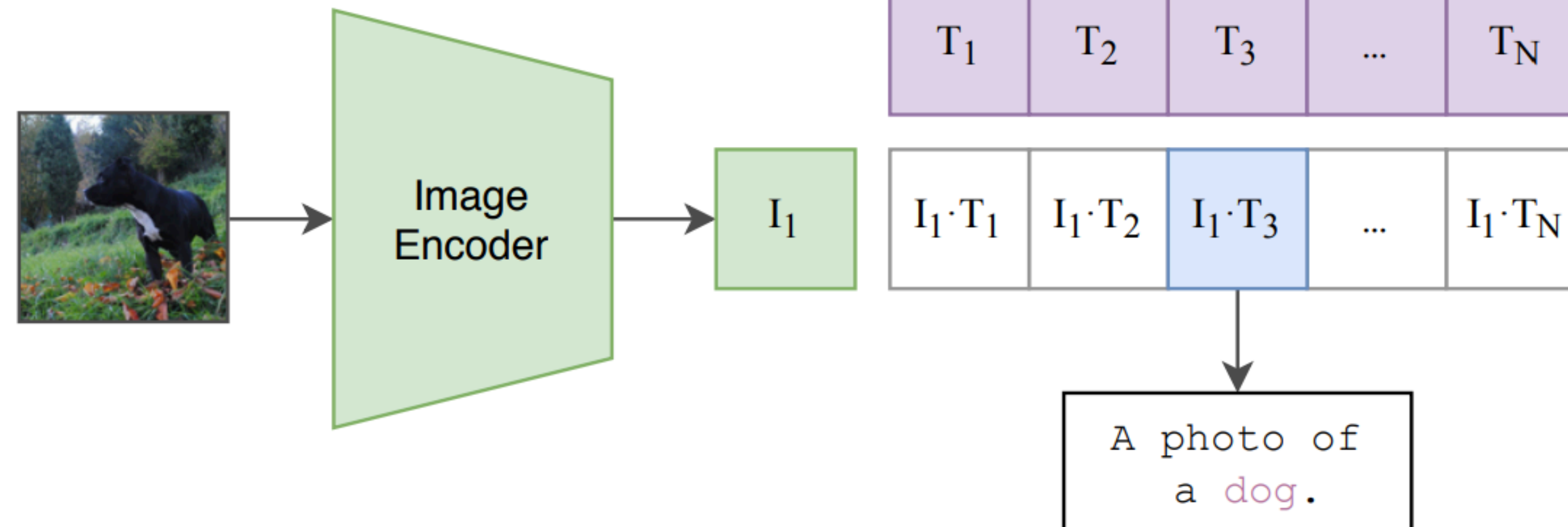
$$\mathcal{L}_{\theta} = -\log \left(\frac{\exp(\mathbf{I}_i \cdot \mathbf{T}_i)}{\sum_j \exp(\mathbf{I}_i \cdot \mathbf{T}_j)} \right)$$

"Zero-shot" classification

(1) Create classifier from label text



(2) Test how well each prompt fits an image



"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**.

"Zero-shot" classification

roundabout (96.4%) Ranked 1 out of 45



✓ satellite imagery of **roundabout**.

✗ satellite imagery of **intersection**.

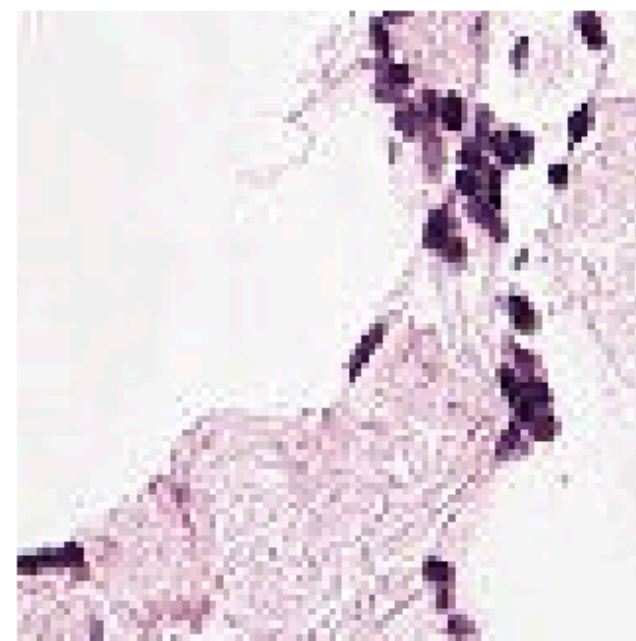
✗ satellite imagery of **church**.

✗ satellite imagery of **medium residential**.

✗ satellite imagery of **chaparral**.

PATCHCAMELYON (PCAM)

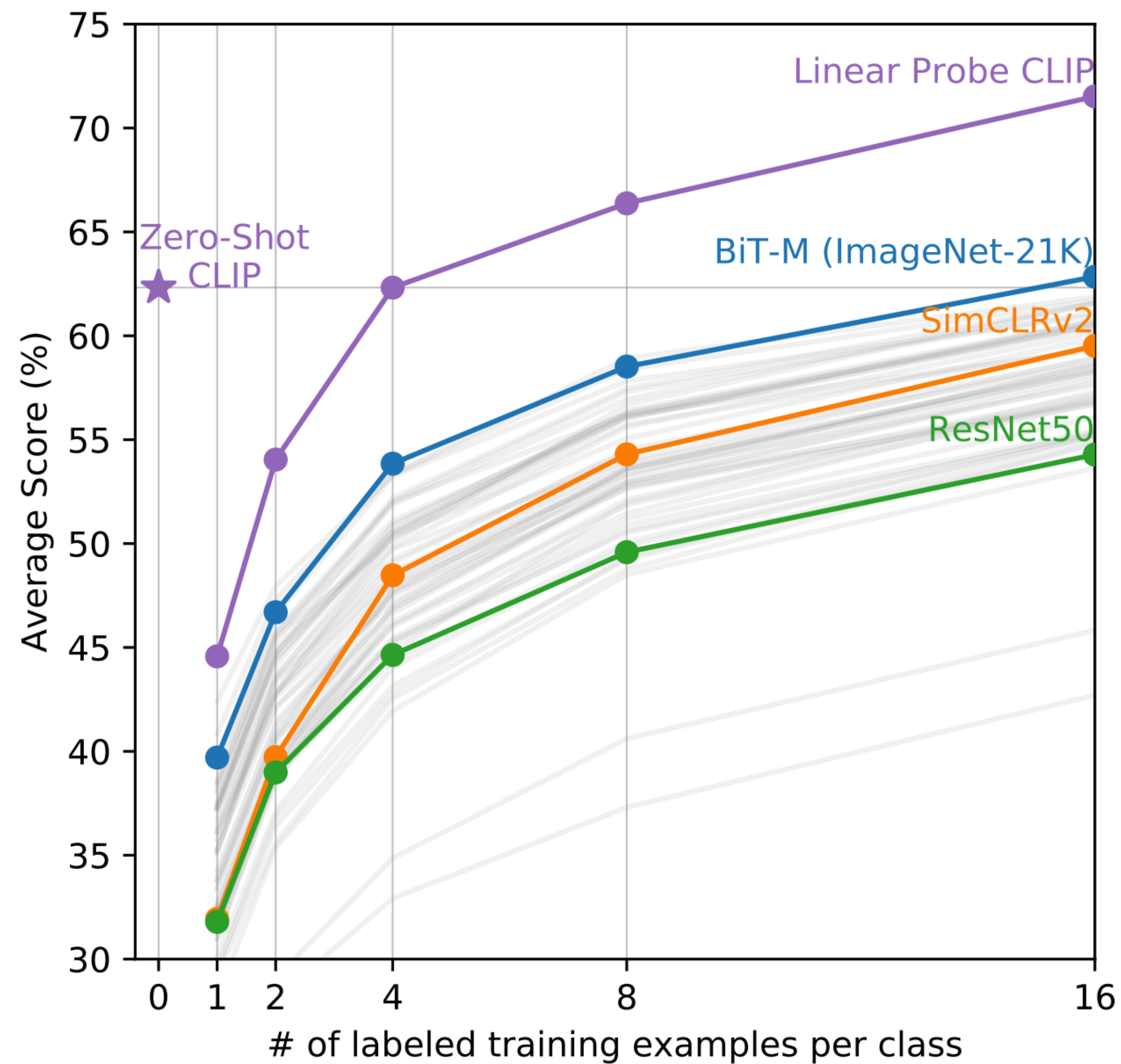
healthy lymph node tissue (22.8%) Ranked 2 out of 2



✗ this is a photo of **lymph node tumor tissue**

✓ this is a photo of **healthy lymph node tissue**

"Zero-shot" classification



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

Next classes: generative models