“~~Unsupervised~~” *Self Supervised* Deep Learning

James Hays
slides from Carl Doersch and Richard Zhang

Recap

- Big Data
  - The Unreasonable Effectiveness of Data
  - Scene Completion
  - Im2gps
  - Recognition via Tiny Images
- Crowdsourcing
  - “Wisdom of the Crowds” / consensus
  - Find good annotators through grading
  - Pricing affects throughput but not quality
  - User interface and instructions matter a lot

Today’s Lecture

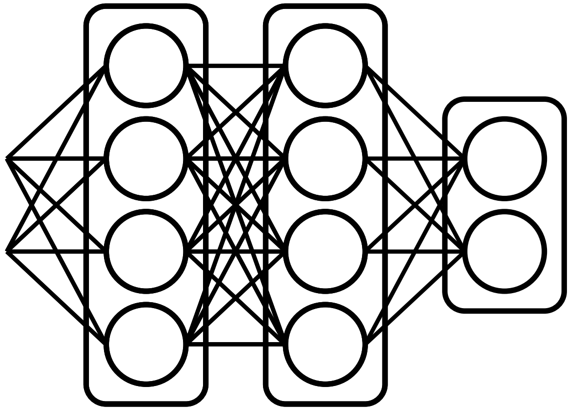
- Three methods for “unsupervised” deep learning
  - Context Prediction. Doersch et al. ICCV 2015
  - Colorful Image Colorization. Zhang et al. ECCV 2016
  - SimCLR. Chen et al. ICML 2020
  - Masked Autoencoders. He et al. CVPR 2022
- Big picture: do we need big, labeled datasets like ImageNet to make deep learning worthwhile? Can we learn from something else?

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei A. Efros, and Abhinav Gupta

ICCV 2015

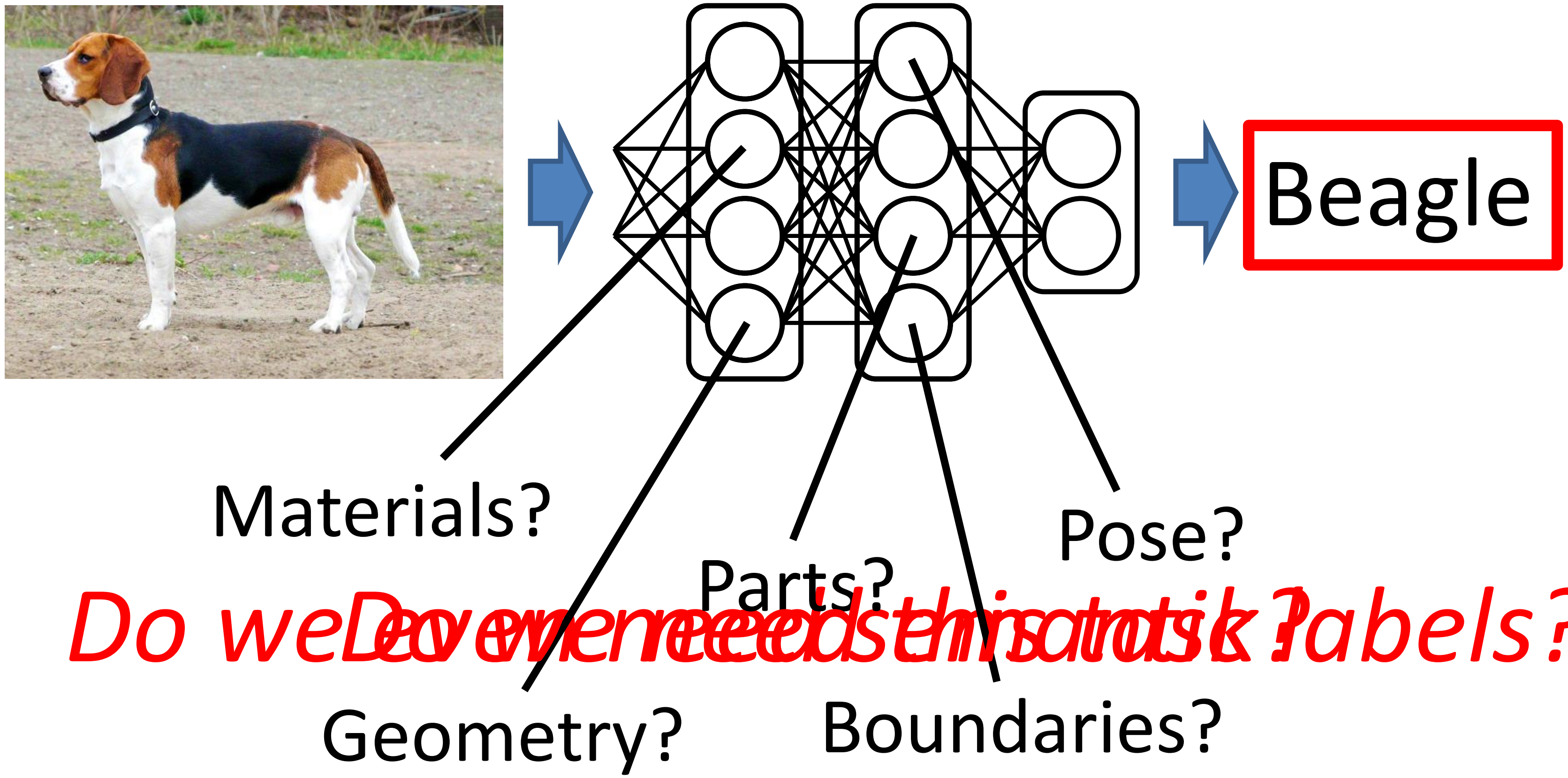
ImageNet + Deep Learning



Beagle

- Image Retrieval
- Detection (RCNN)
- Segmentation (FCN)
- Depth Estimation
- …

ImageNet + Deep Learning

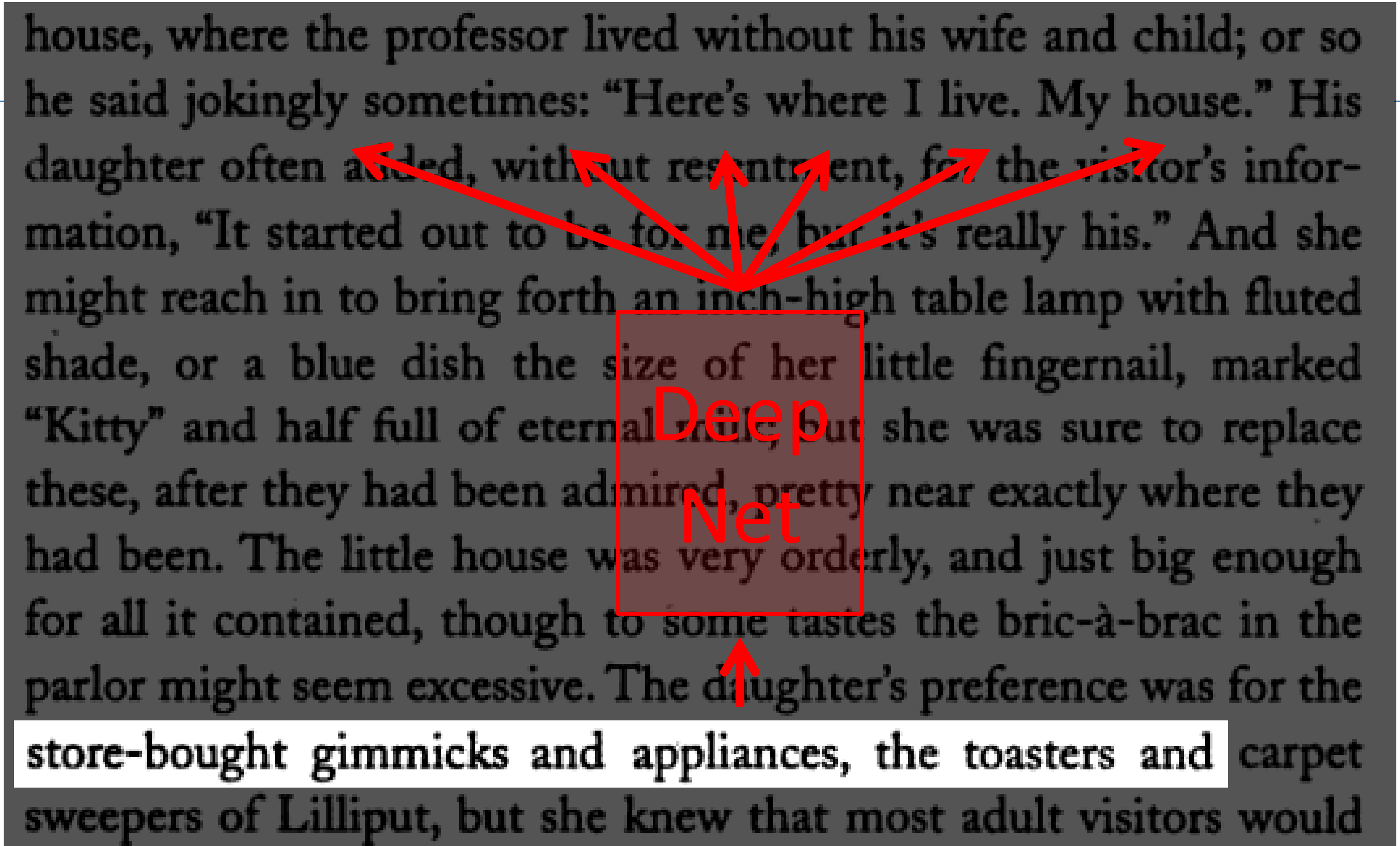


Do we even need semantic labels?

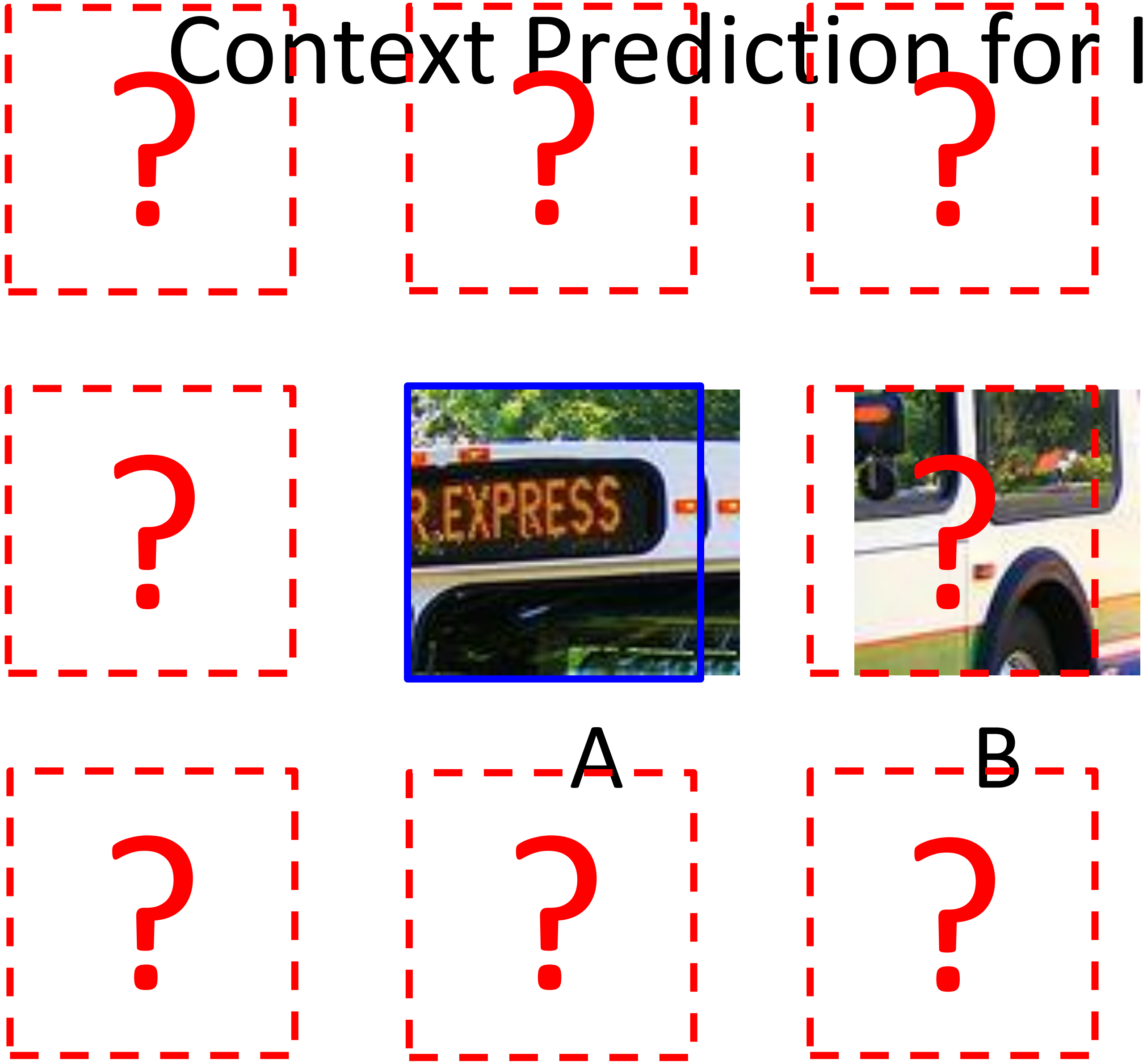
Do we need this task?

Context as Supervision

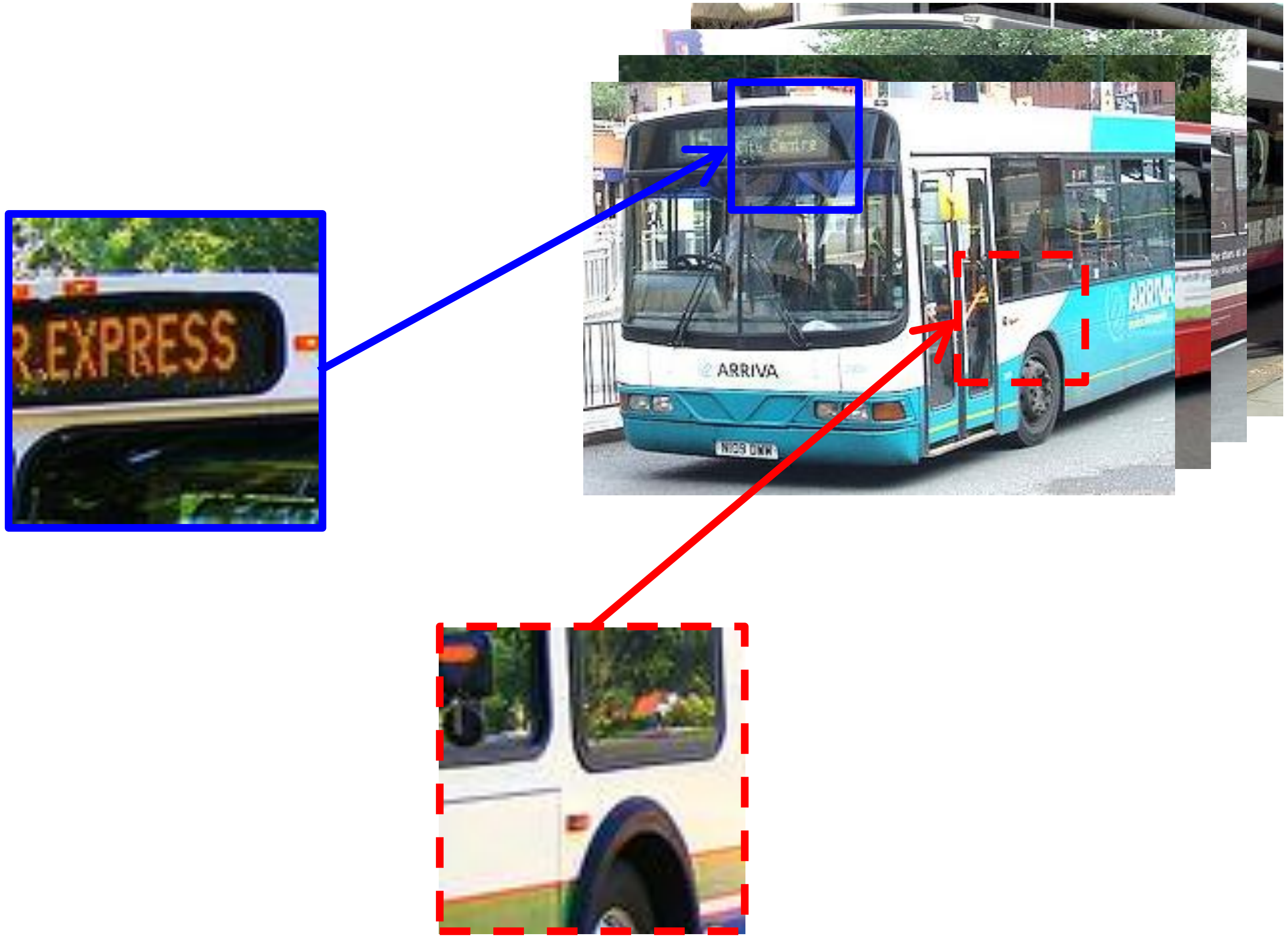
[Collobert & Weston 2008; Mikolov et al. 2013 (Word2Vec)]



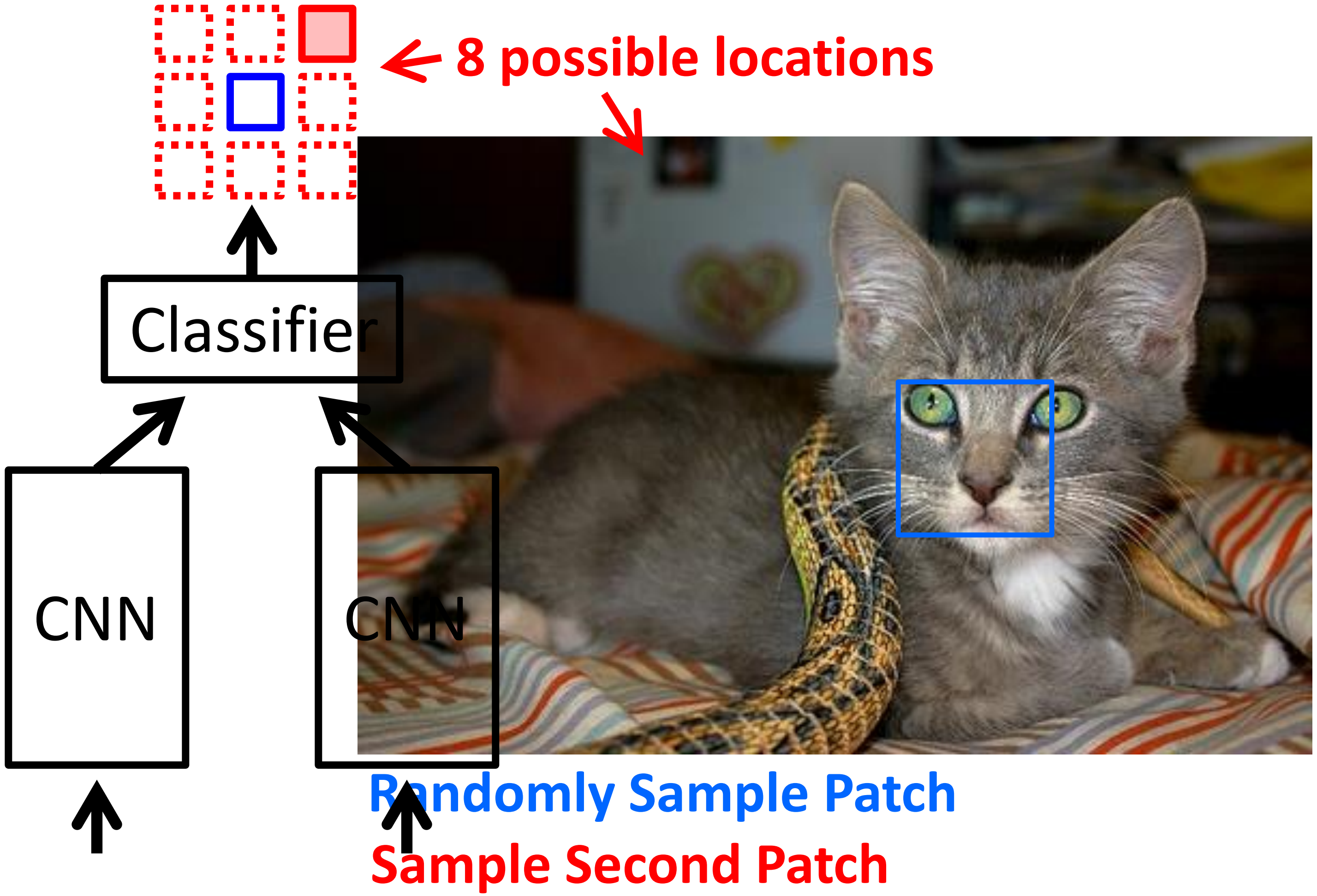
Context Prediction for Images

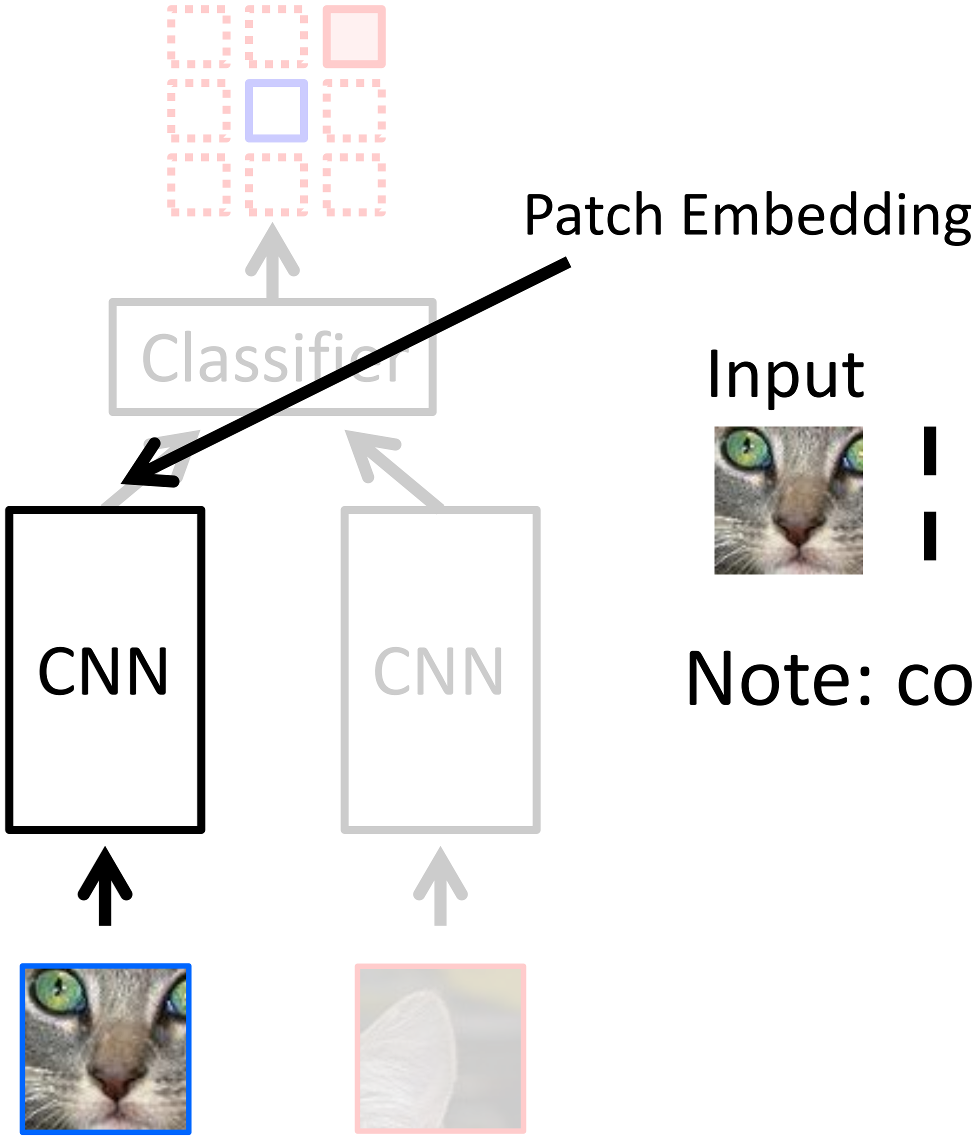


Semantics from a non-semantic task



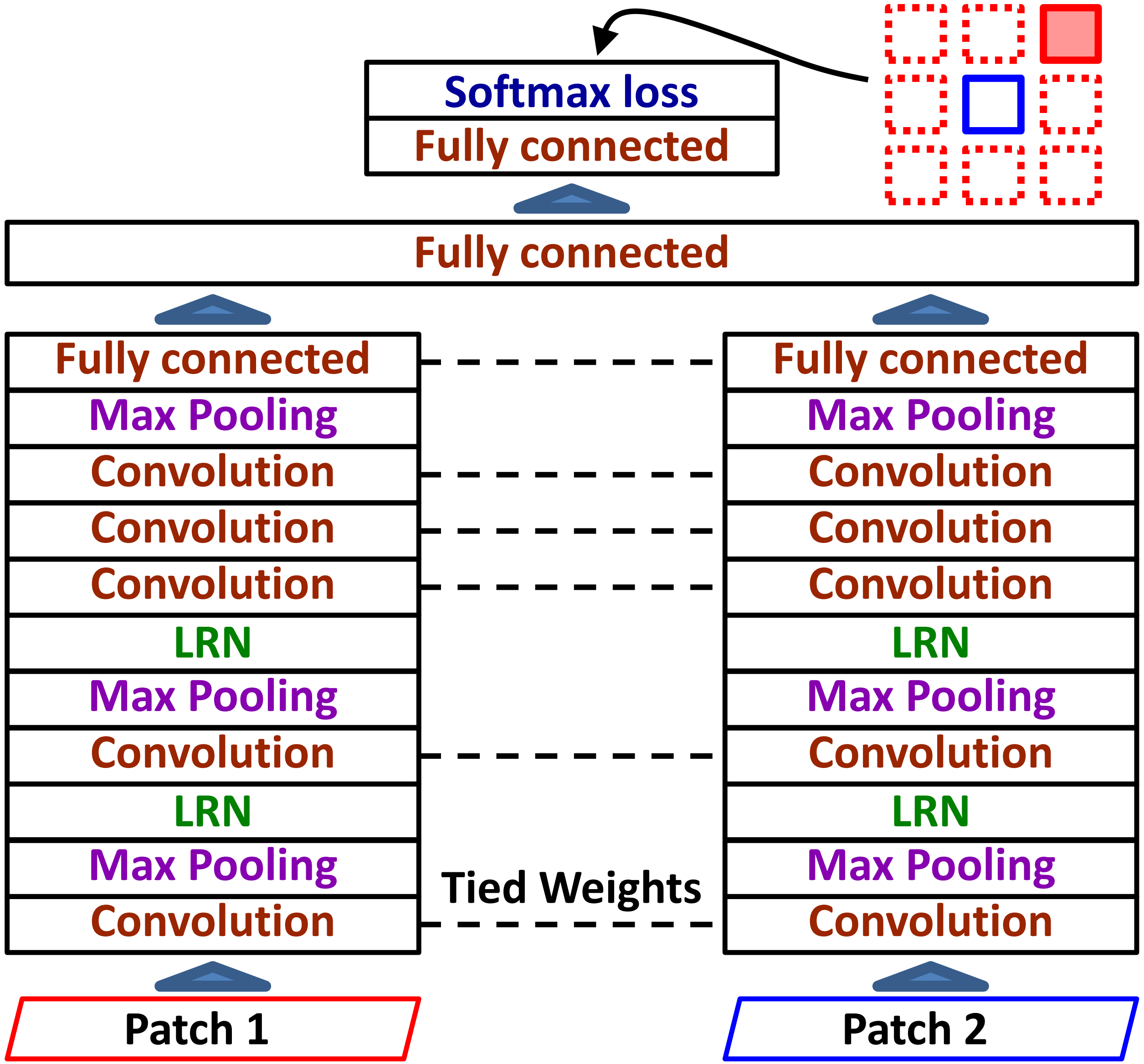
Relative Position Task



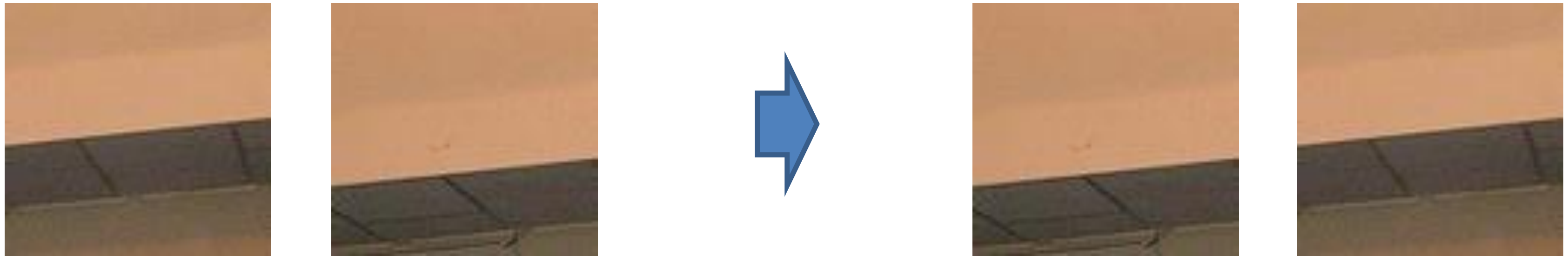


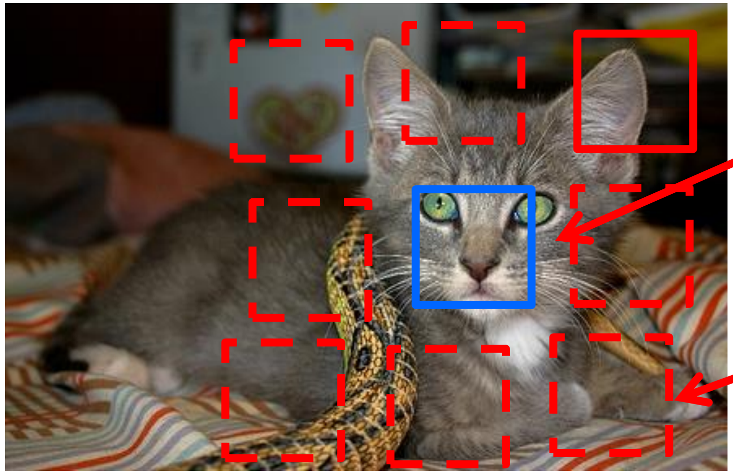
Note: connects ***across*** instances!

Architecture



Avoiding Trivial Shortcuts

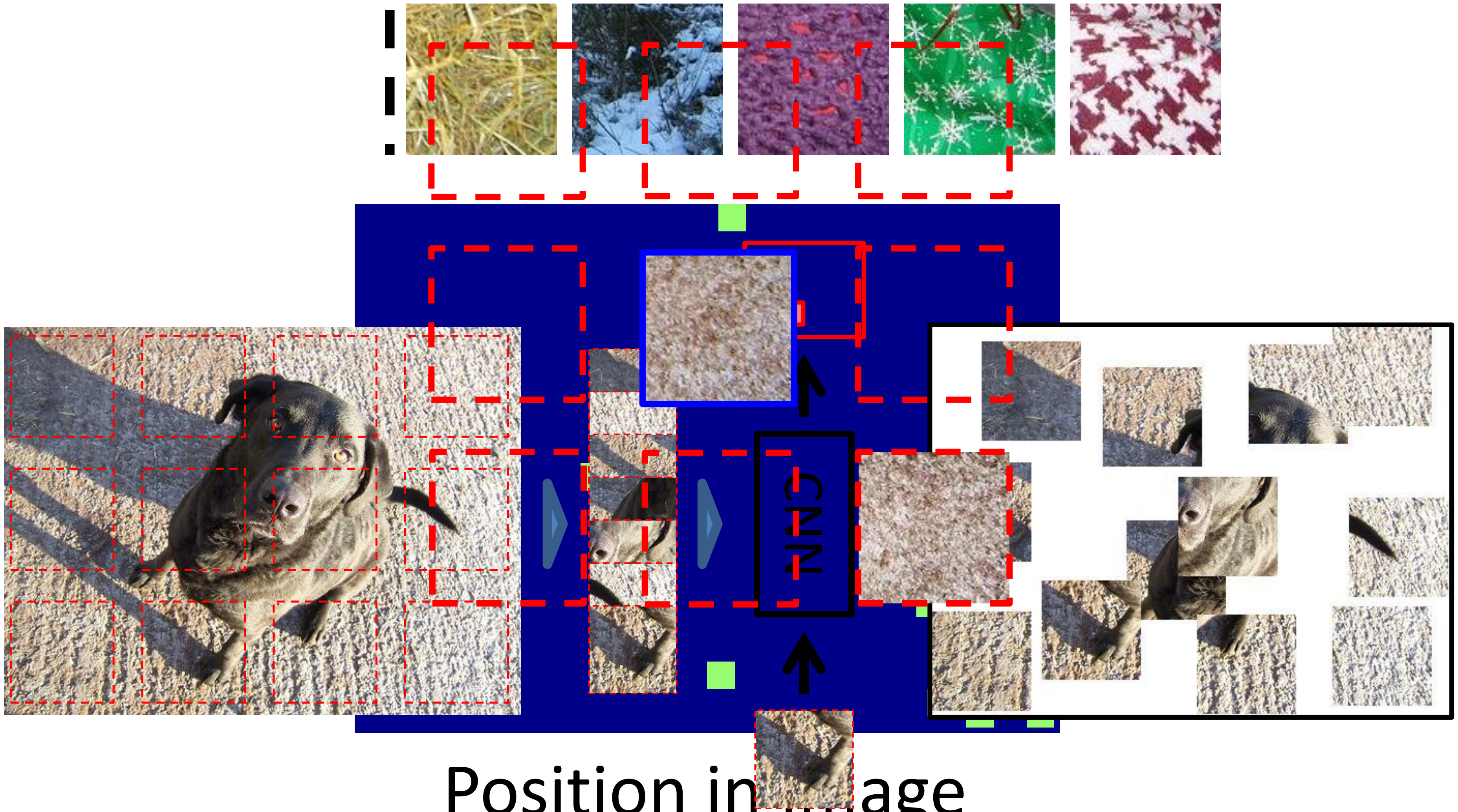




Include a gap

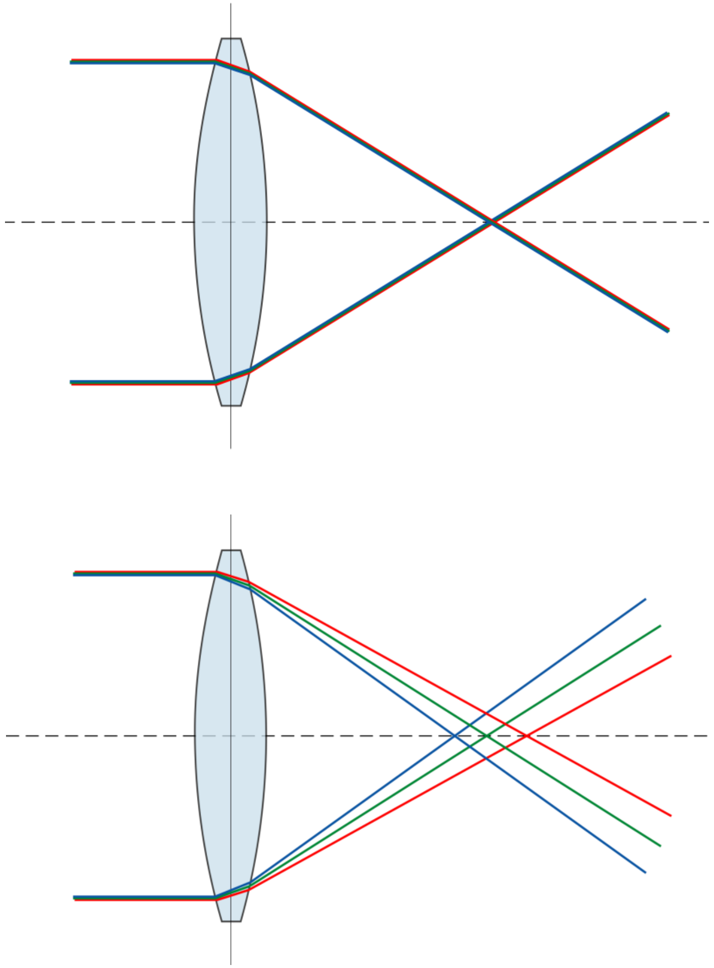
Jitter the patch locations

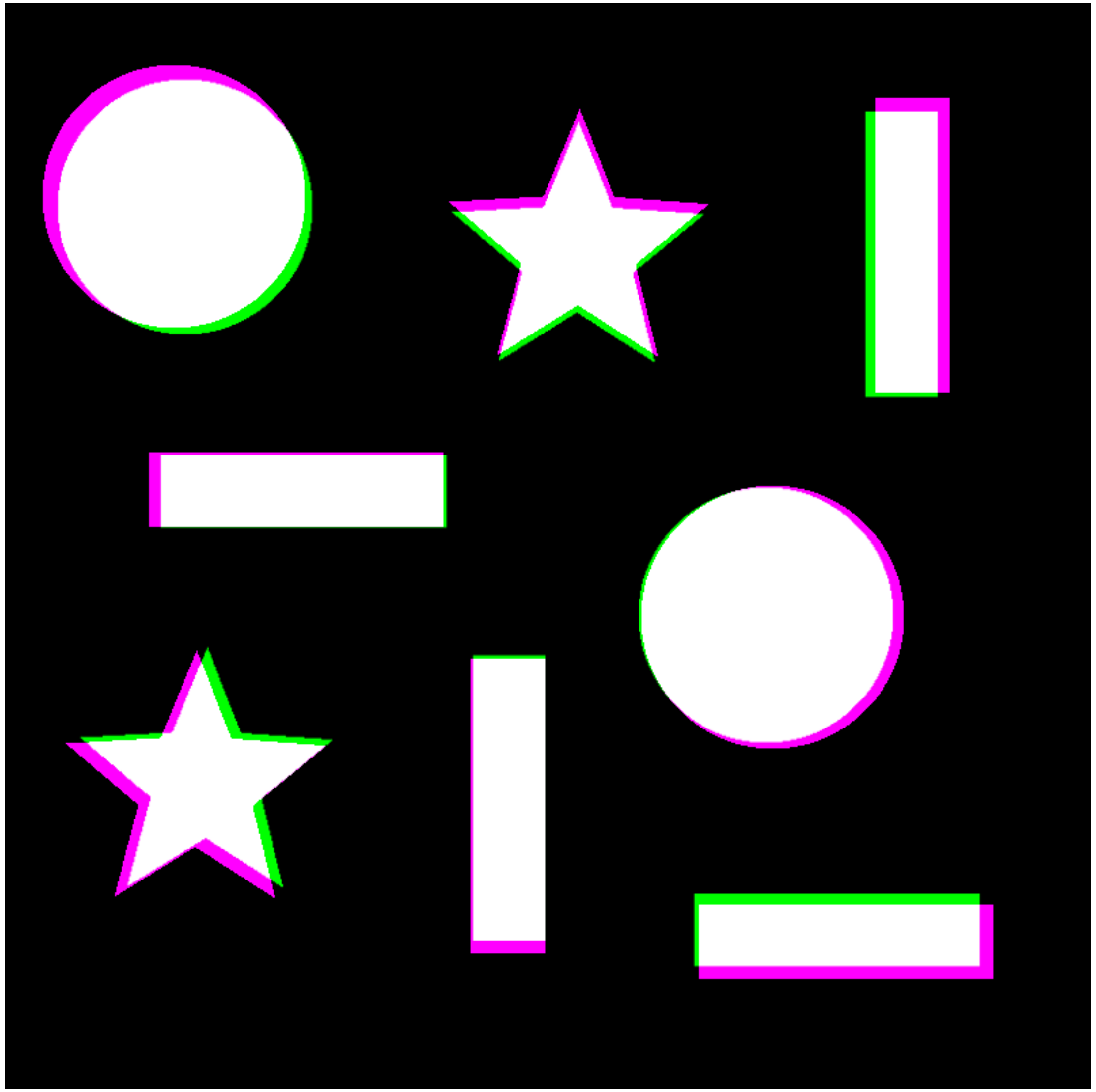
A Not-So "Trivial" Shortcut



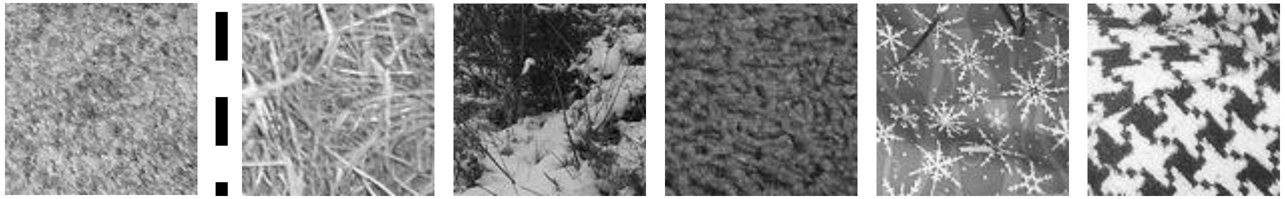
Position in Image

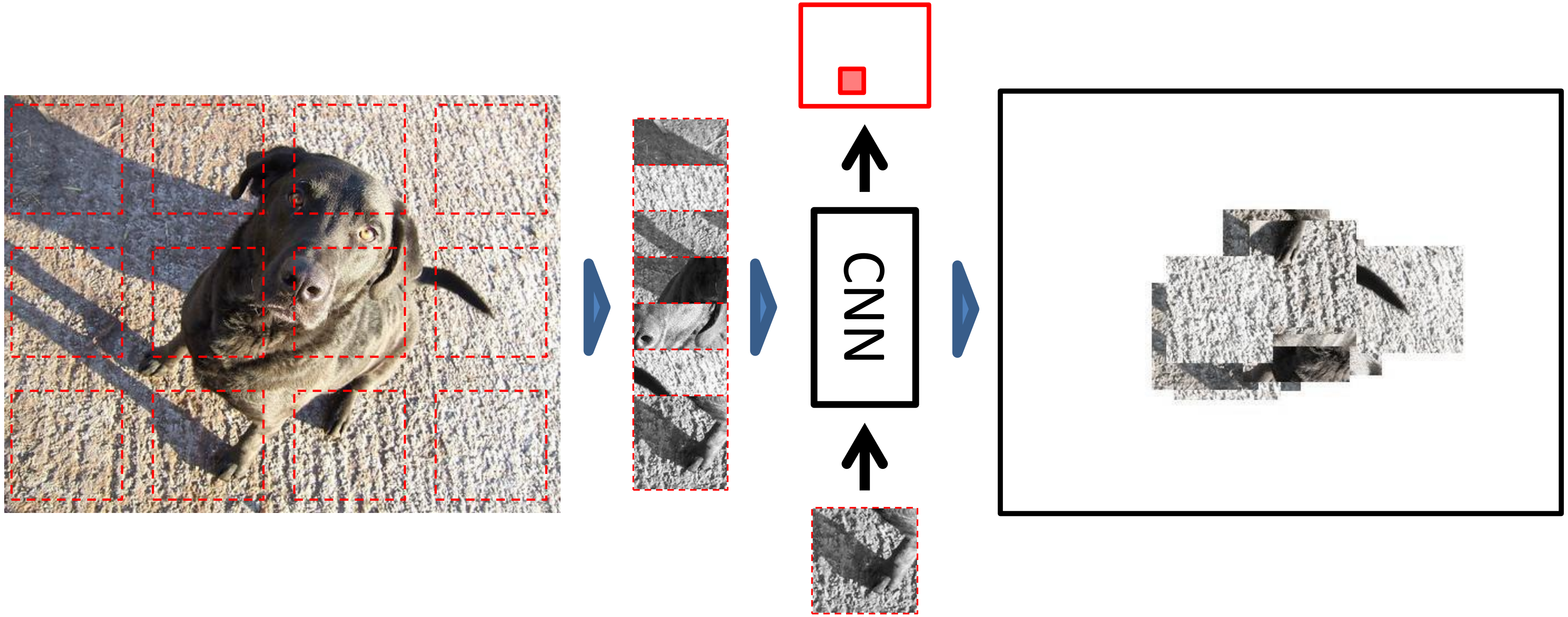
Chromatic Aberration



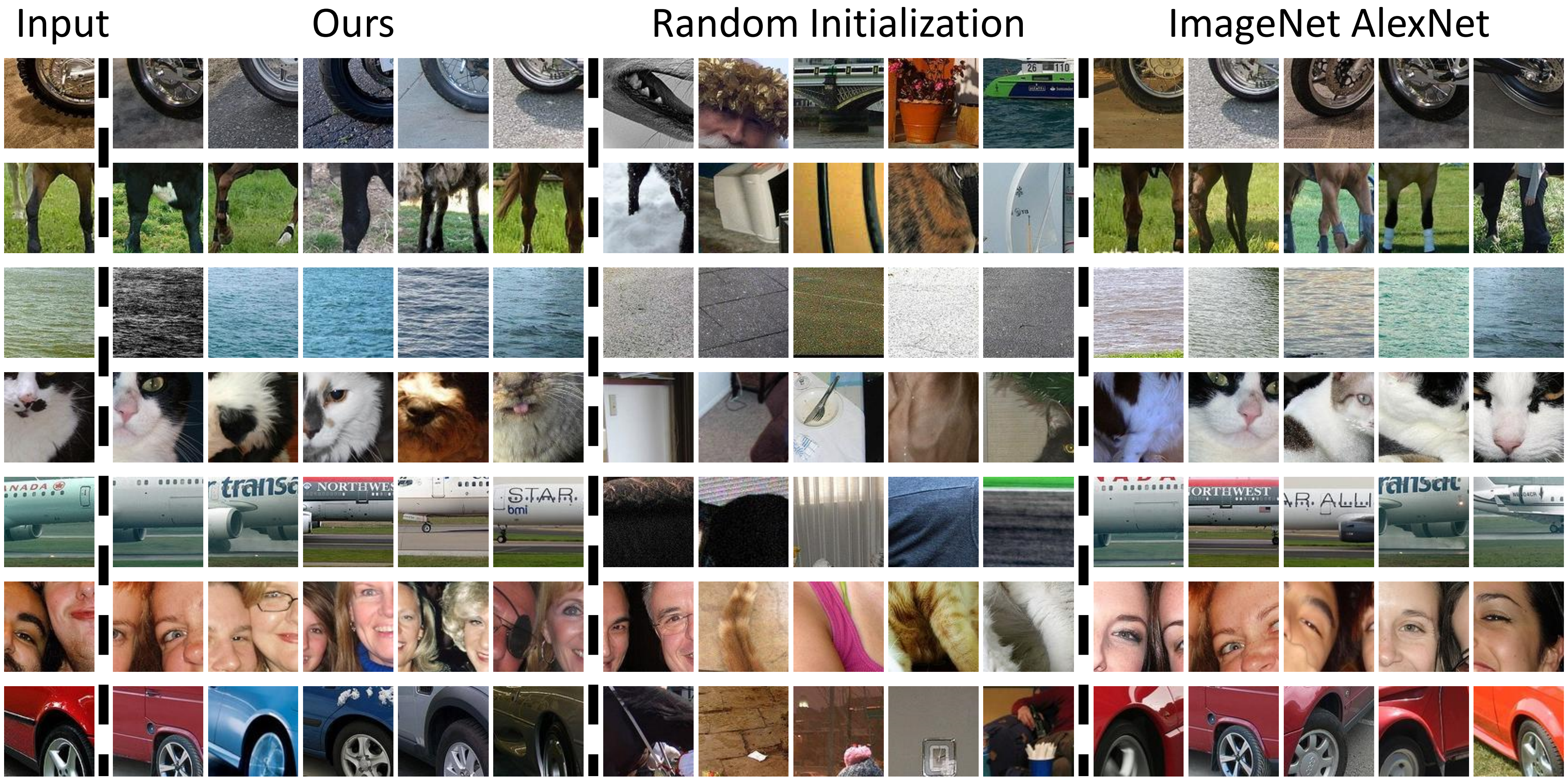


Chromatic Aberration



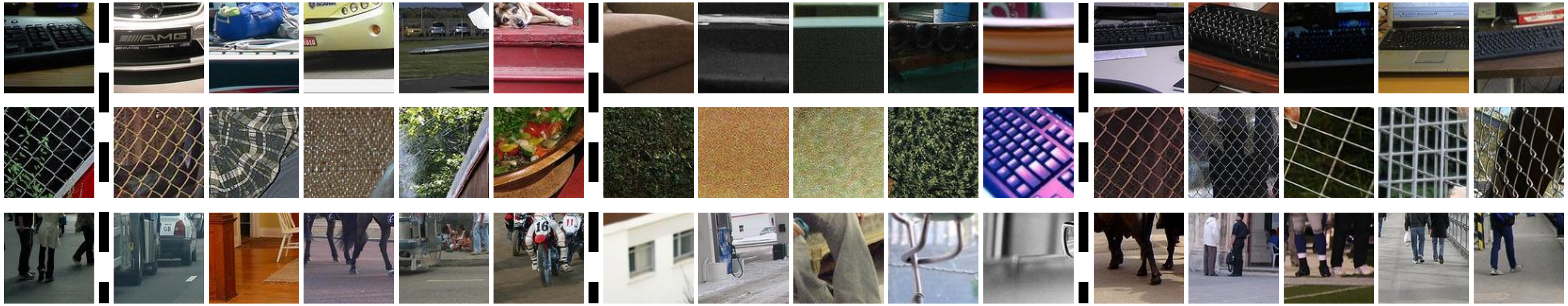


What is learned?



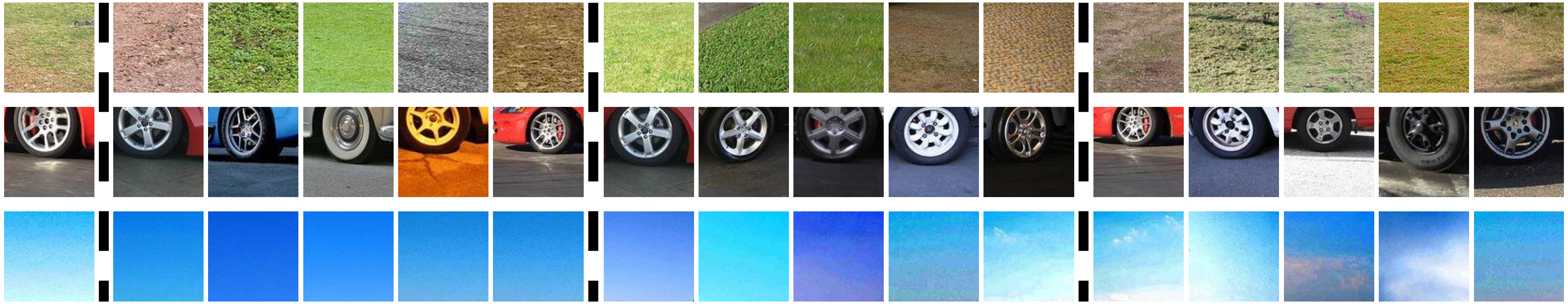
Still don't capture everything

Input Ours Random Initialization ImageNet AlexNet

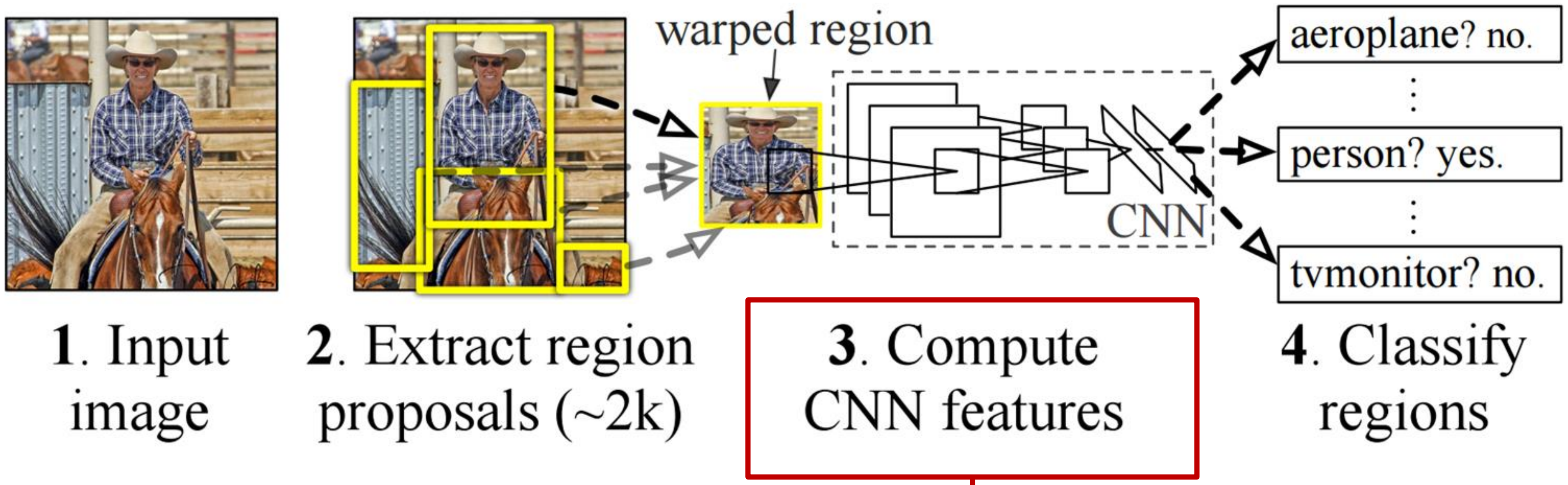


You don't always need to learn!

Input Ours Random Initialization ImageNet AlexNet



Pre-Training for R-CNN

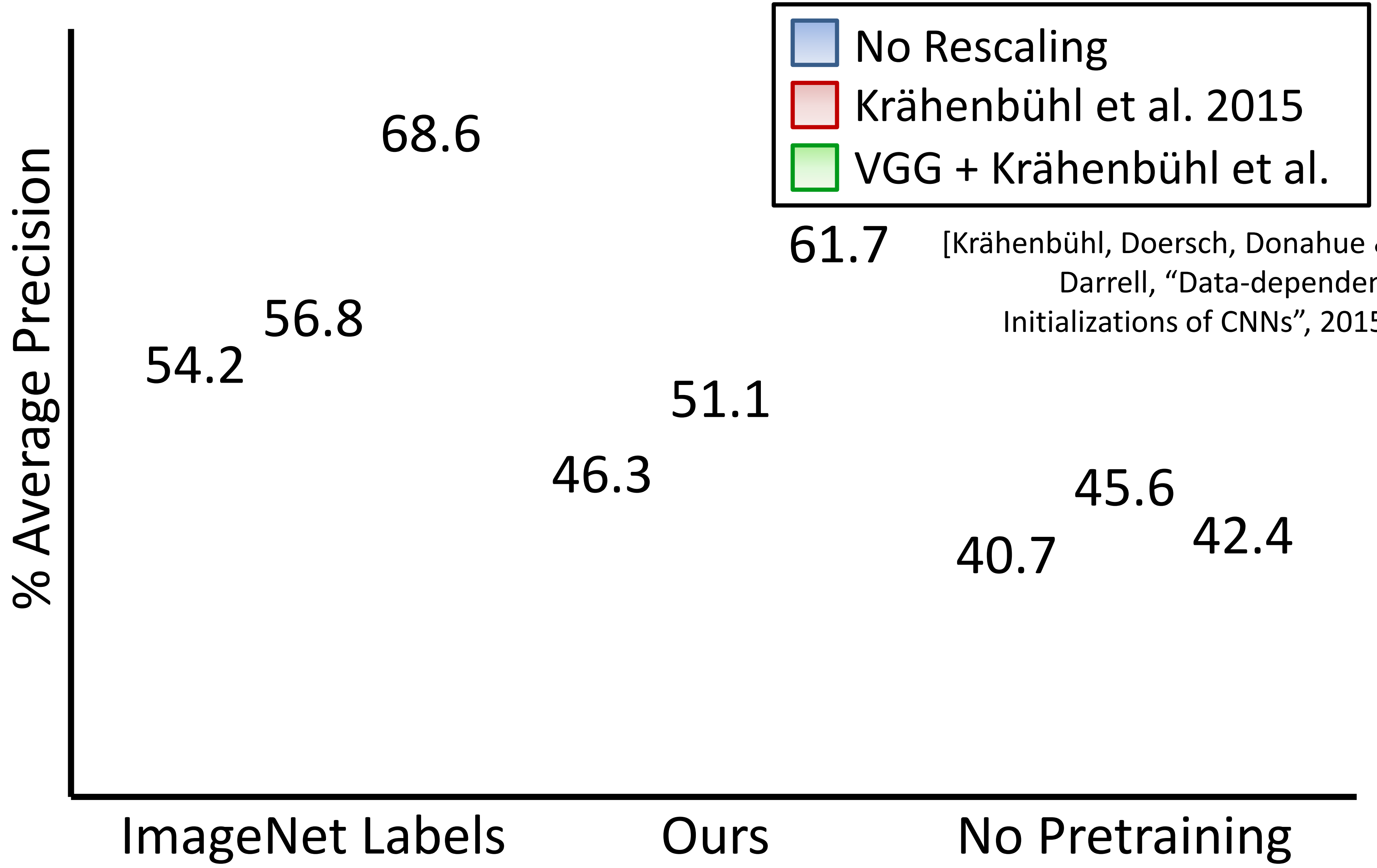


Pre-train on relative-position task, w/o labels

[Girshick et al. 2014]

VOC 2007 Performance

(pretraining for R-CNN)

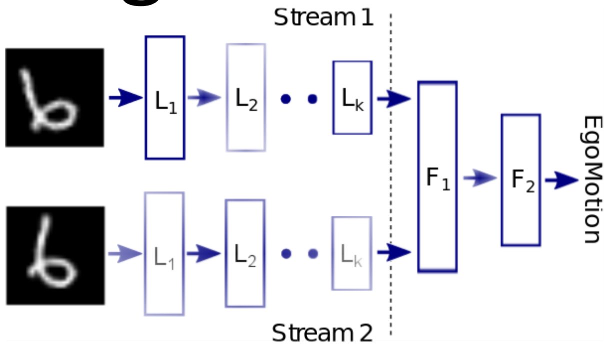


[Krähenbühl, Doersch, Donahue & Darrell, "Data-dependent Initializations of CNNs", 2015]

So, do we need semantic labels?

“Self-Supervision” and the Future

Ego-Motion

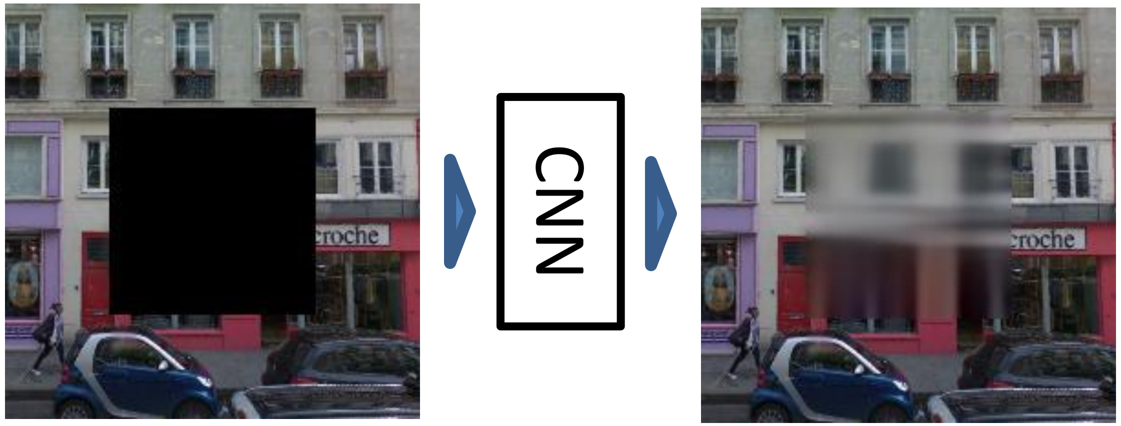


[Agrawal et al. 2015; Jayaraman et al. 2015]

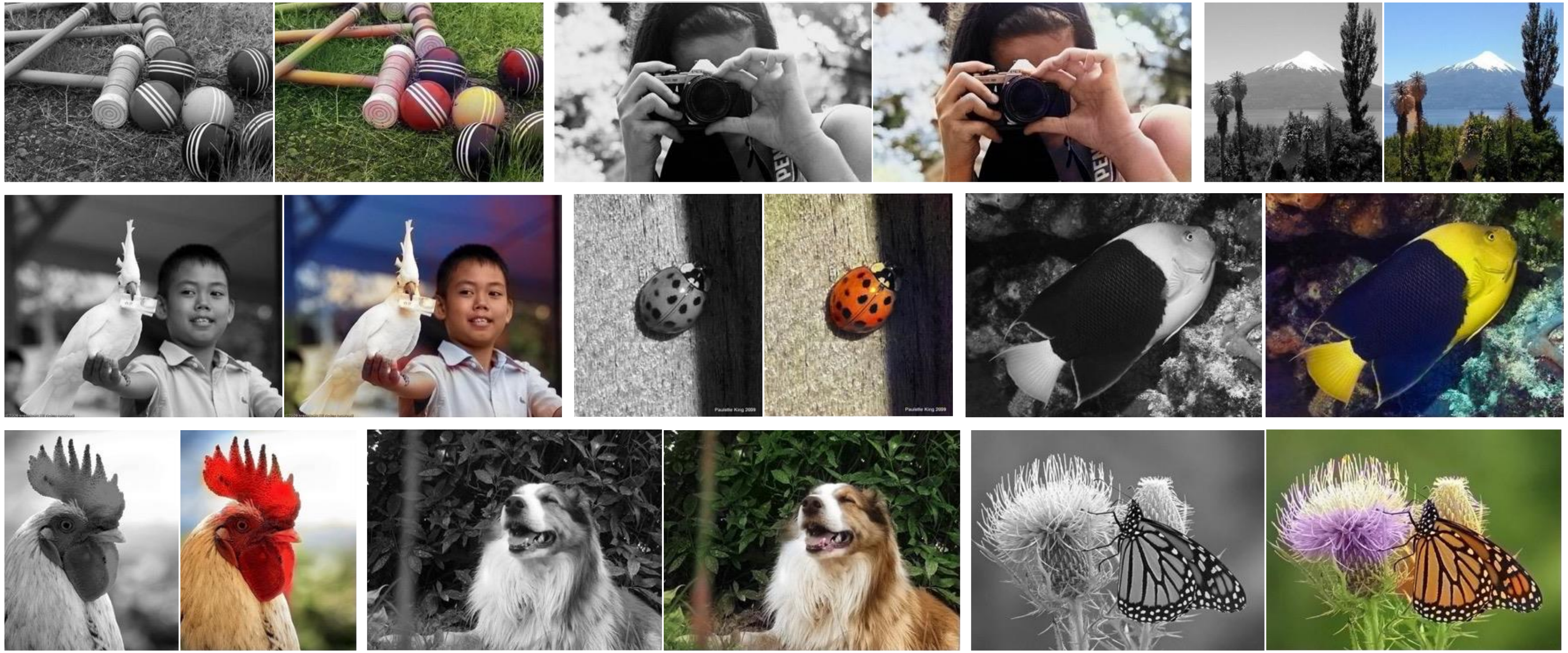
Video

[Wang et al. 2015; Srivastava et al 2015; …]

Context



[Doersch et al. 2014; Pathak et al. 2015; Isola et al. 2015]

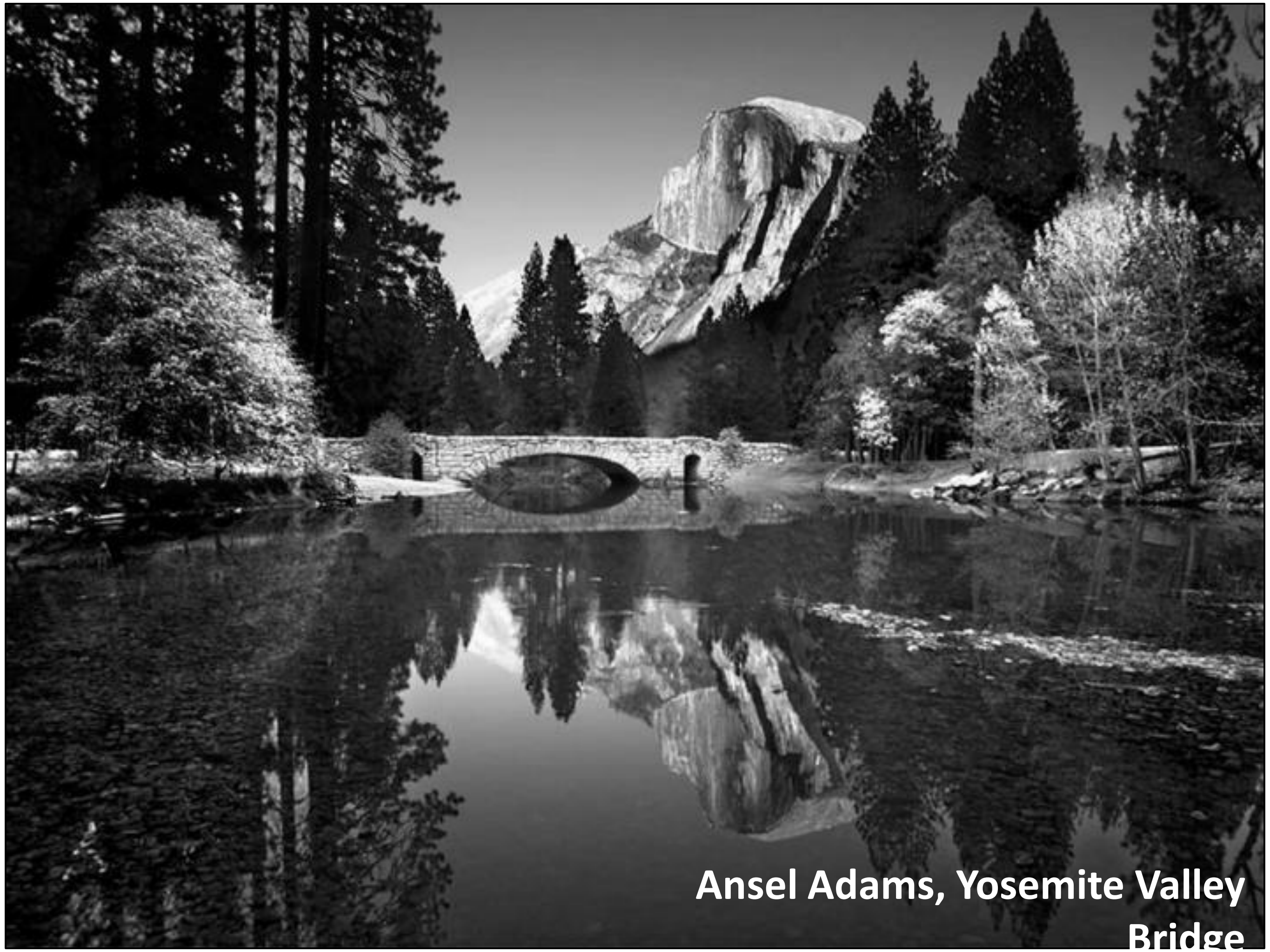


Colorful Image Colorization

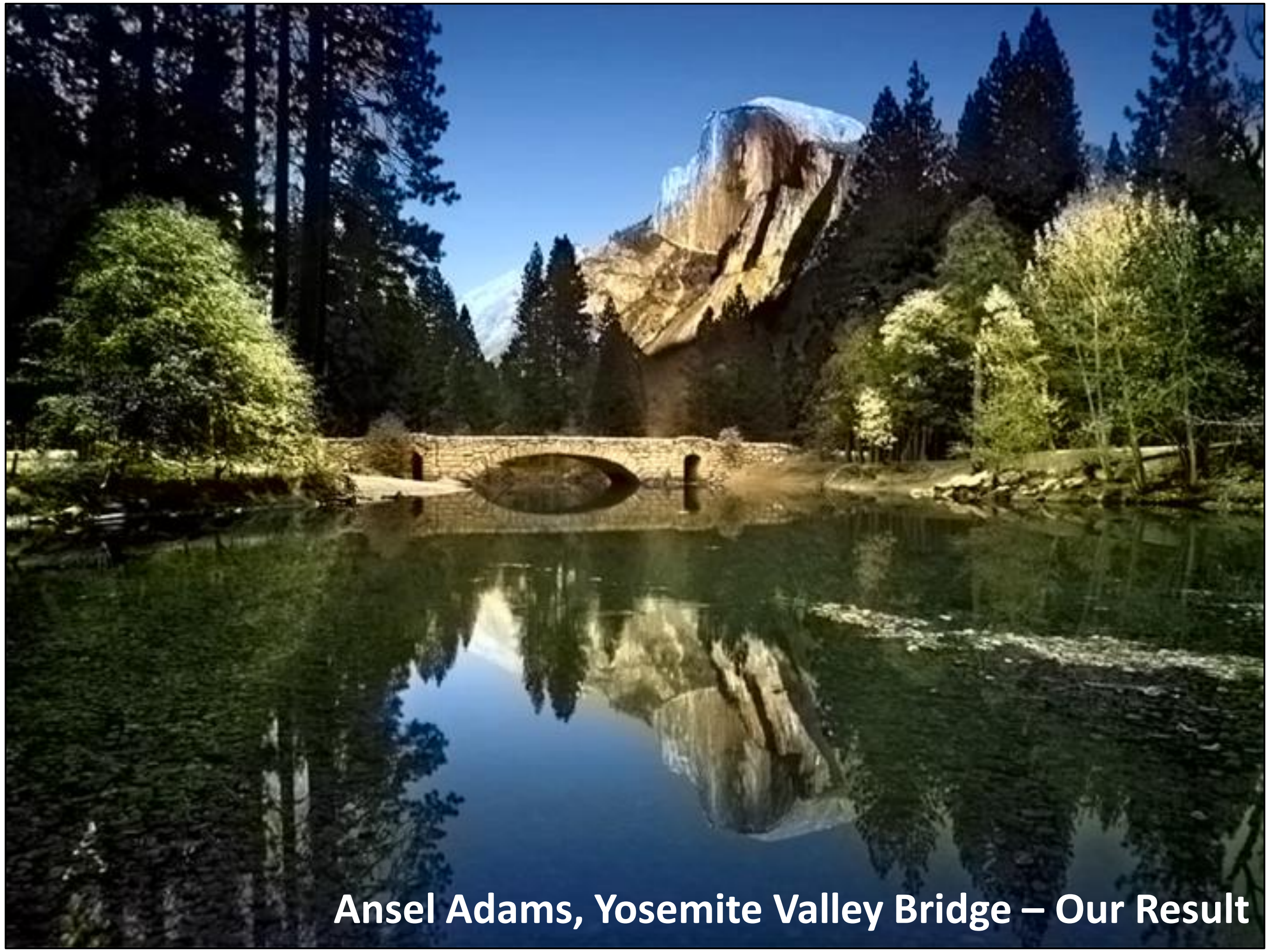
Richard Zhang, Phillip Isola, Alexei (Alyosha) Efros

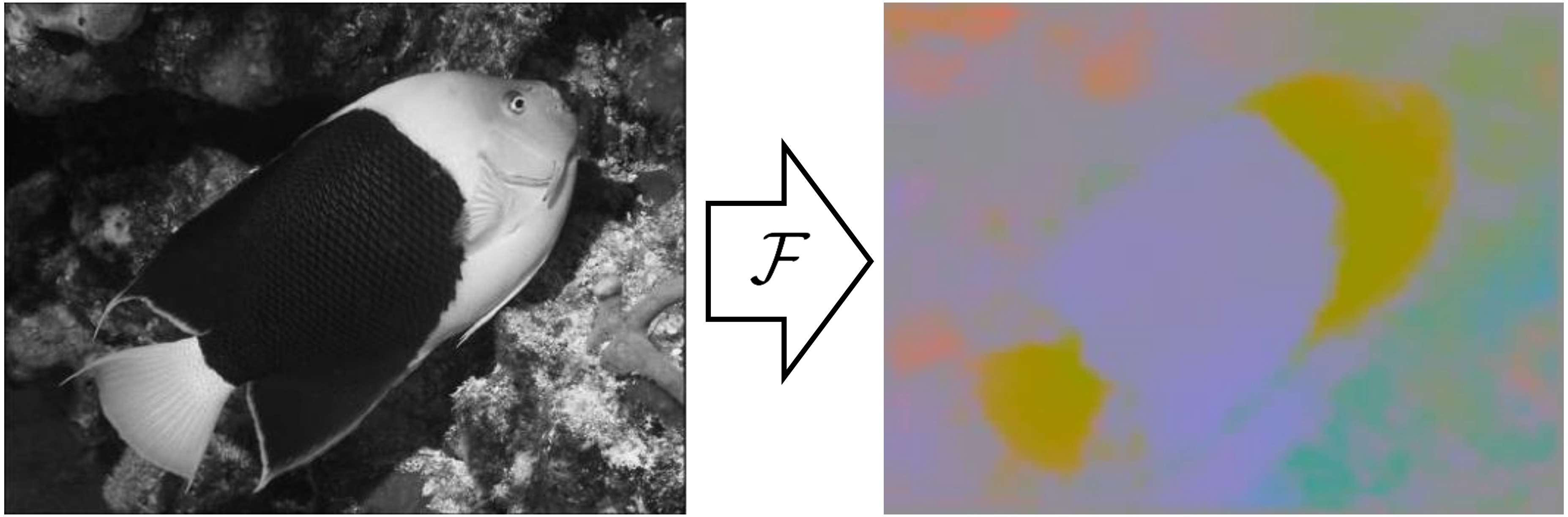
richzhang.github.io/colorization

ECCV 2016



Ansel Adams, Yosemite Valley Bridge


Ansel Adams, Yosemite Valley Bridge – Our Result

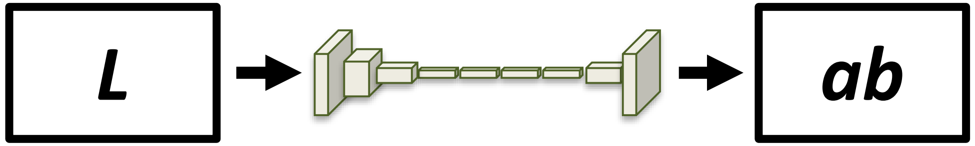


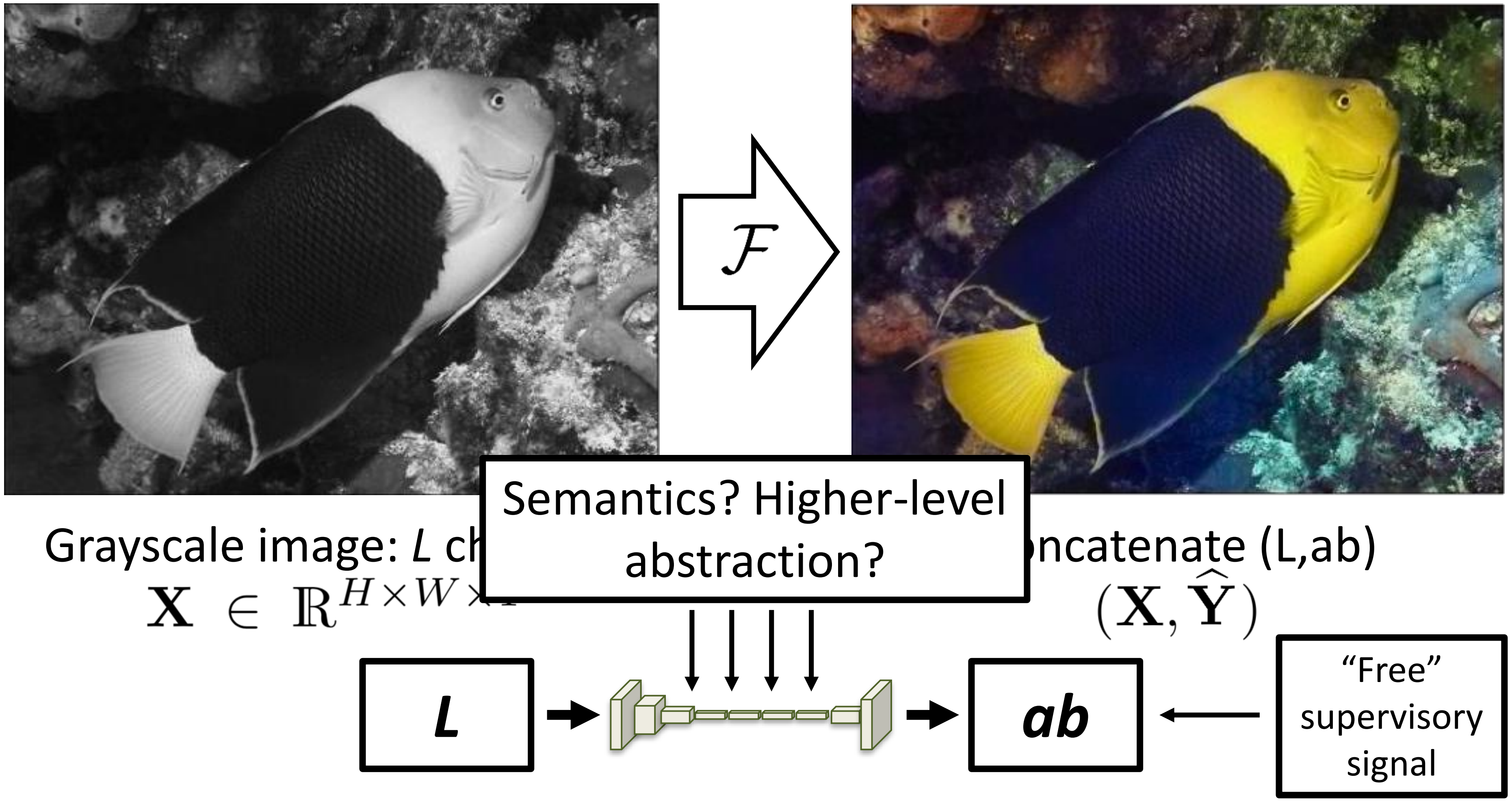
Grayscale image: *L* channel

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

Color information: *ab* channels

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




$\mathcal{F}$
Semantics? Higher-level abstraction?
Grayscale image: *L* ch
$\mathbf{X} \in \mathbb{R}^{H \times W \times}$
ncatenate (L,ab)
$(\mathbf{X}, \widehat{\mathbf{Y}})$
L
ab
“Free” supervisory signal

Inherent Ambiguity

Grayscale

Inherent Ambiguity

Our Output

Ground Truth

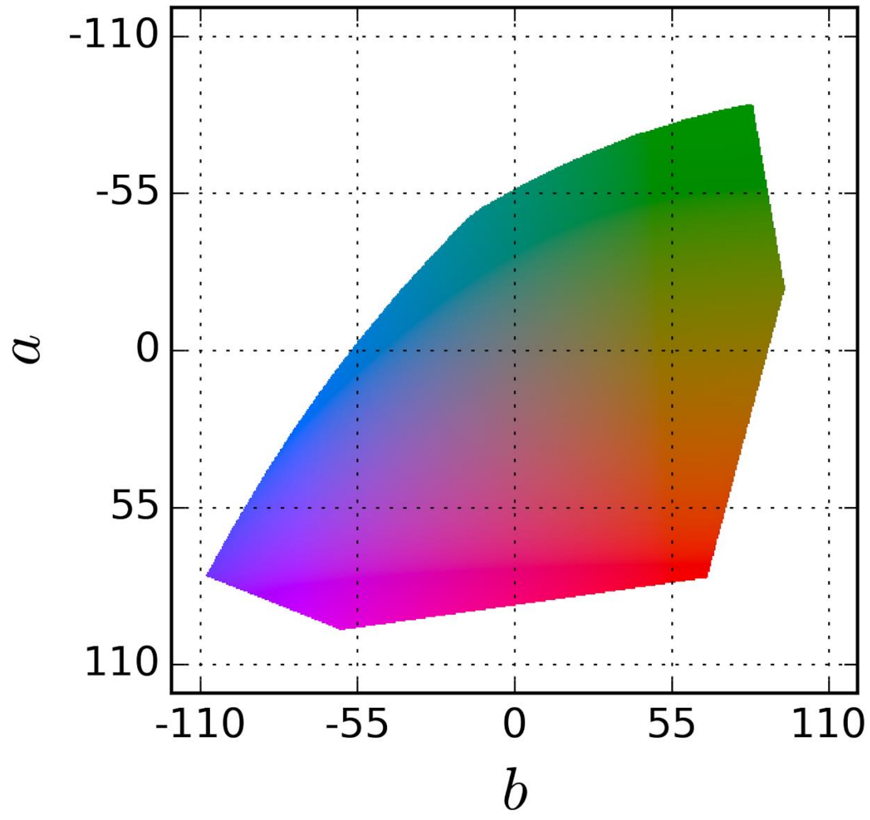
Better Loss Function

- Regression with L2 loss inadequate

$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2}\sum_{h,w}\|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$

Colors in *ab* space

(continuous)



Better Loss Function

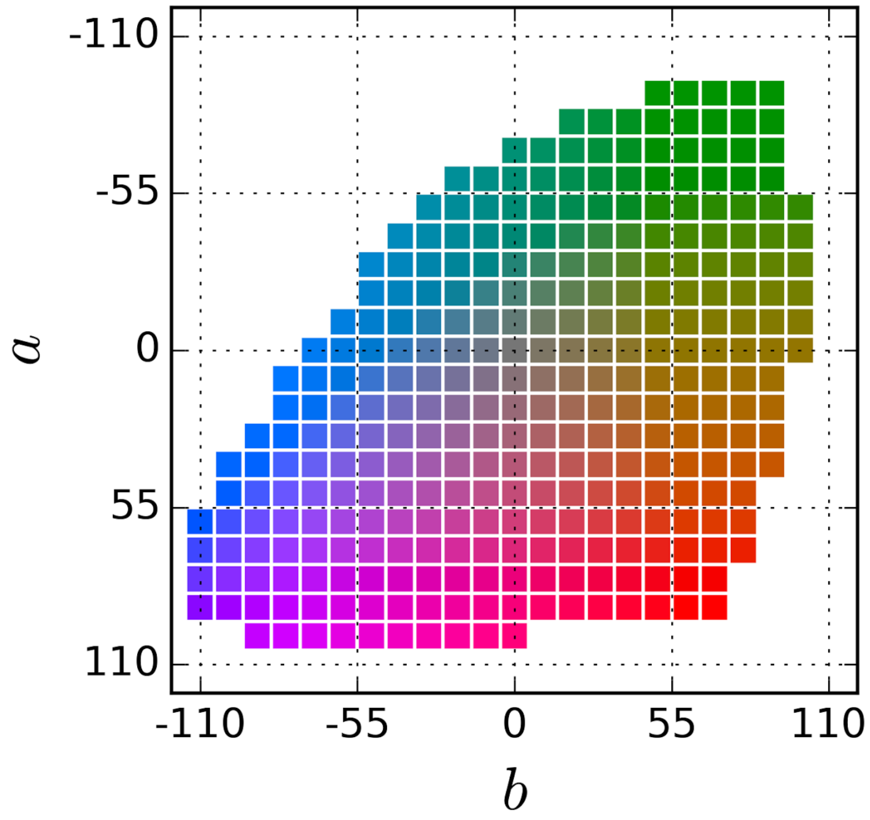
- Regression with L2 loss inadequate

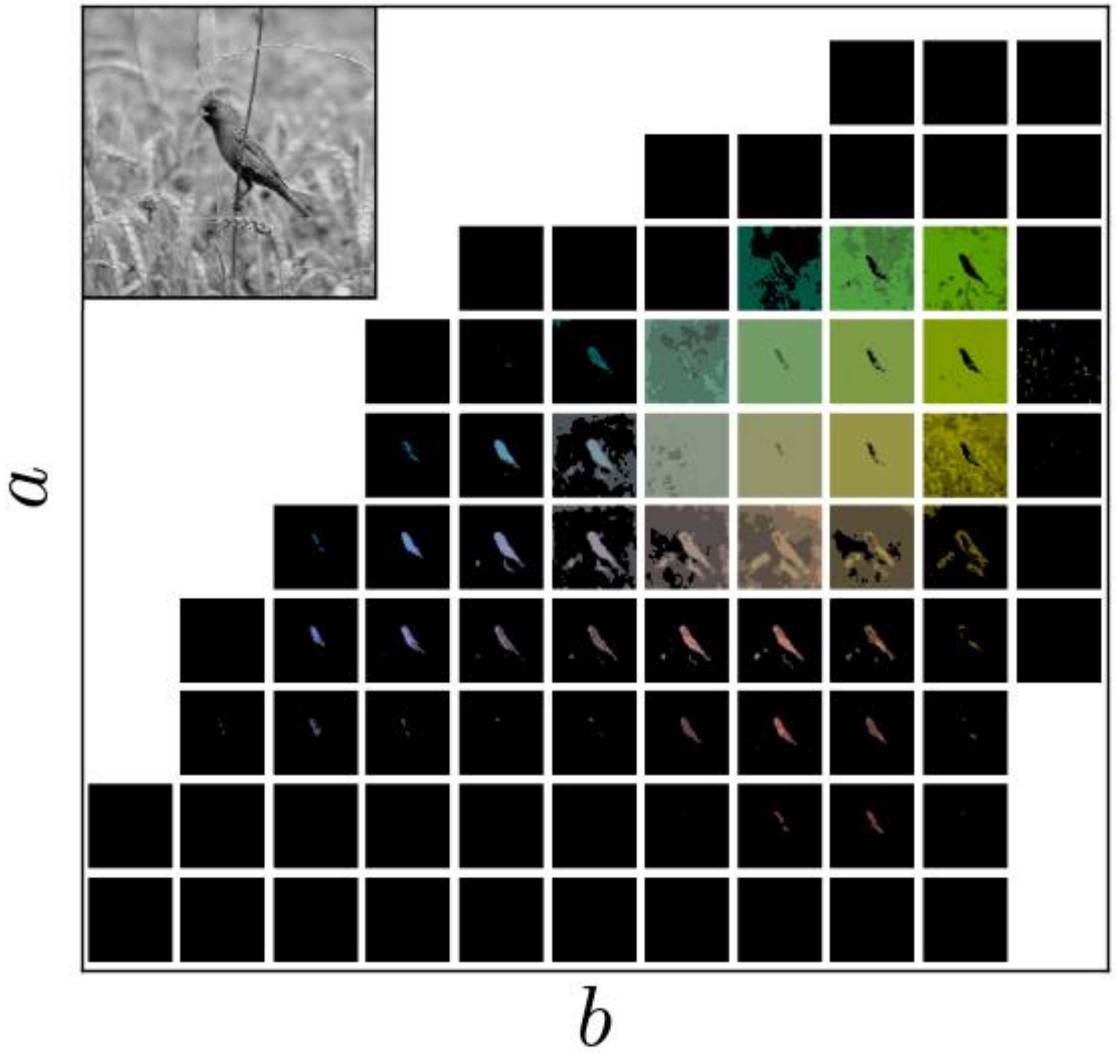
$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$

- Use **multinomial classification**

$$\mathrm{L}(\widehat{\mathbf{Z}}, \mathbf{Z}) = -\frac{1}{HW} \sum_{h,w} \sum_q \mathbf{Z}_{h,w,q} \log(\widehat{\mathbf{Z}}_{h,w,q})$$

Colors in *ab* space
(discrete)





a
b

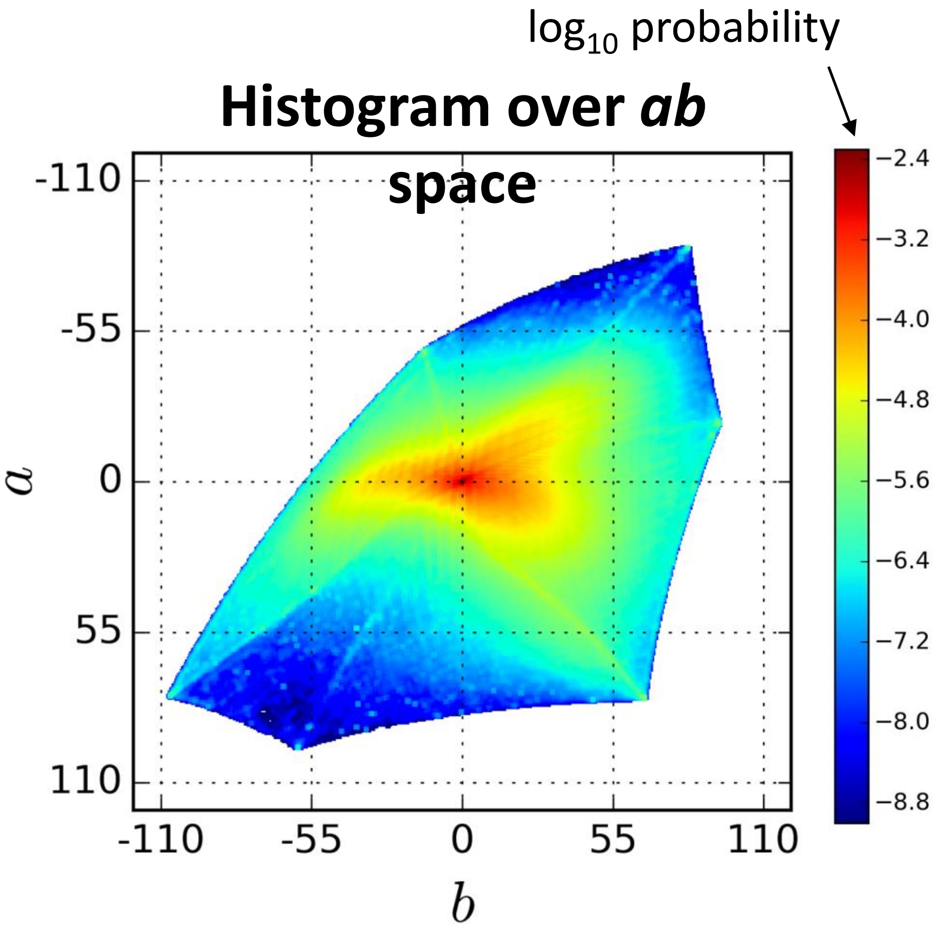
Better Loss Function

- Regression with L2 loss inadequate

$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2}\sum_{h,w}\|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$

- Use **multinomial classification**

$$\mathrm{L}(\widehat{\mathbf{Z}}, \mathbf{Z}) = -\frac{1}{HW}\sum_{h,w}\sum_{q}\mathbf{Z}_{h,w,q}\log(\widehat{\mathbf{Z}}_{h,w,q})$$



Better Loss Function

- Regression with L2 loss inadequate

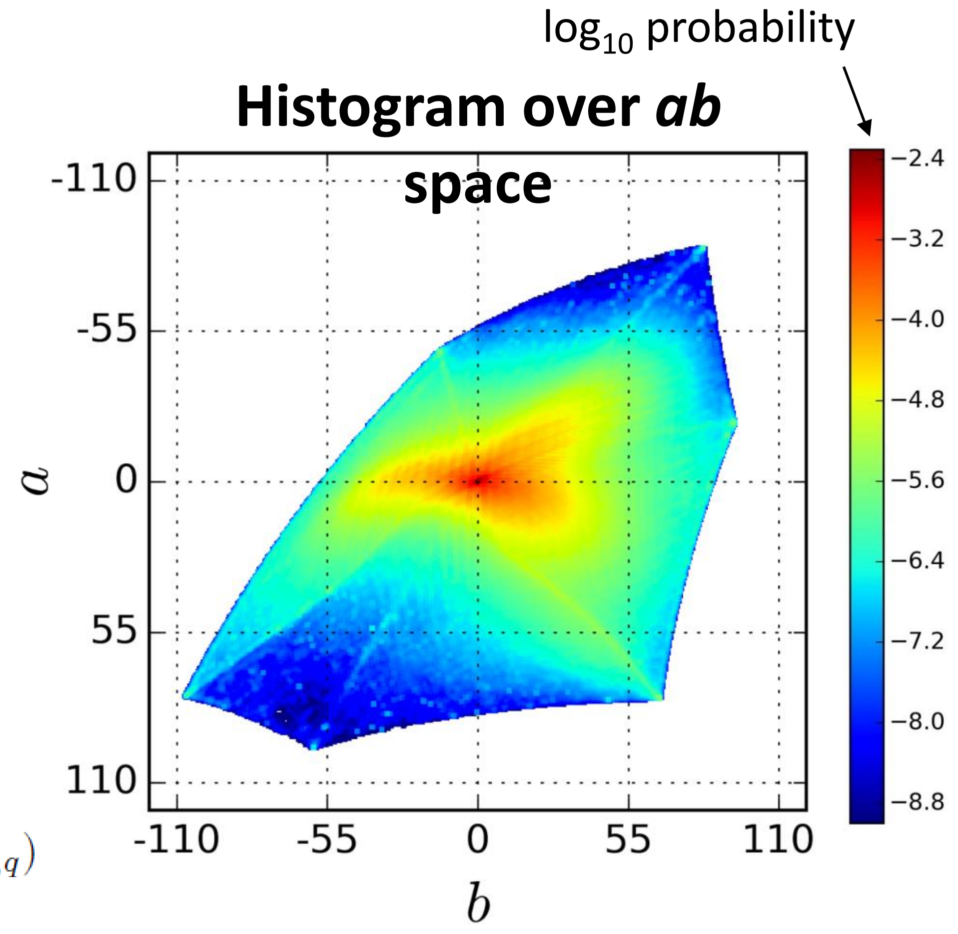
$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$

- Use **multinomial classification**

$$\mathrm{L}(\widehat{\mathbf{Z}}, \mathbf{Z}) = -\frac{1}{HW} \sum_{h,w} \sum_{q} \mathbf{Z}_{h,w,q} \log(\widehat{\mathbf{Z}}_{h,w,q})$$

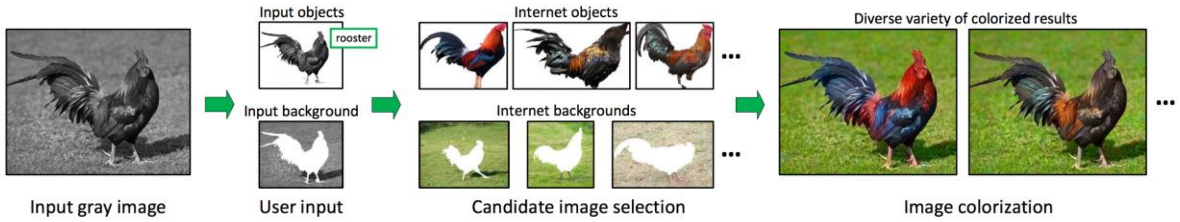
- **Class rebalancing** to encourage learning of *rare* colors

$$\mathrm{L}(\widehat{\mathbf{Z}}, \mathbf{Z}) = -\frac{1}{HW} \sum_{h,w} v(\mathbf{Z}_{h,w}) \sum_{q} \mathbf{Z}_{h,w,q} \log(\widehat{\mathbf{Z}}_{h,w,q})$$



Non-parametric

Hertzmann et al. In SIGGRAPH, 2001.
Welsh et al. In TOG, 2002.
Irony et al. In Eurographics, 2005.
Liu et al. In TOG, 2008.
Chia et al. In ACM 2011.
Gupta et al. In ACM, 2012.

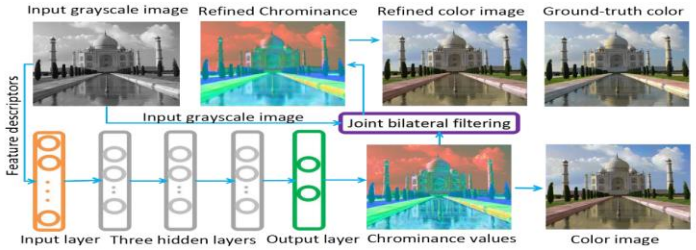


Parametric

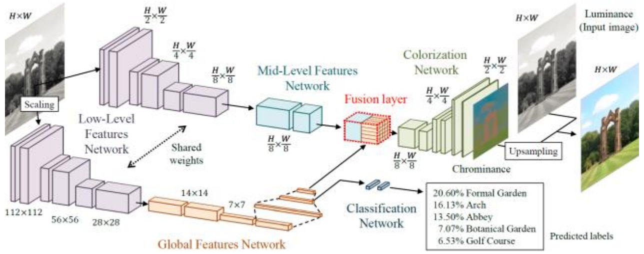
Hand-engineered Features

Deep Networks

L2 Regression

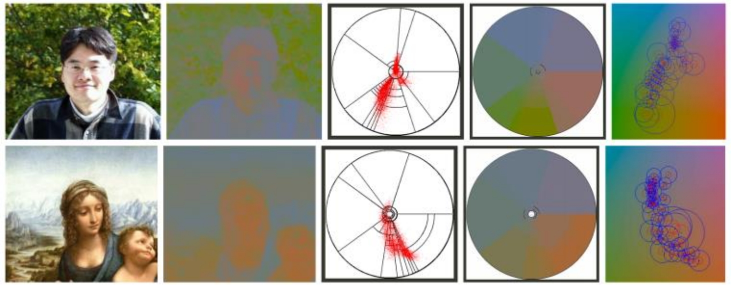


Deshpande et al. Cheng et al. In ICCV 2015.

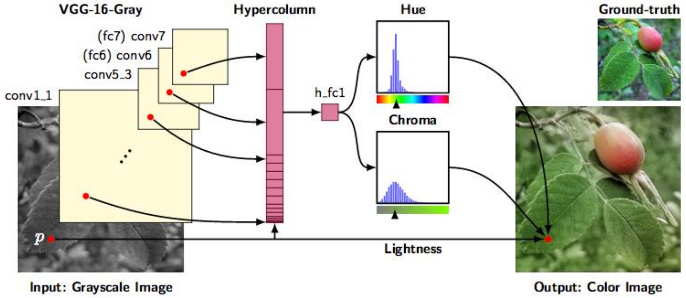


Dahl. Jan 2016. Iizuka et al. In SIGGRAPH, 2016.

Classification

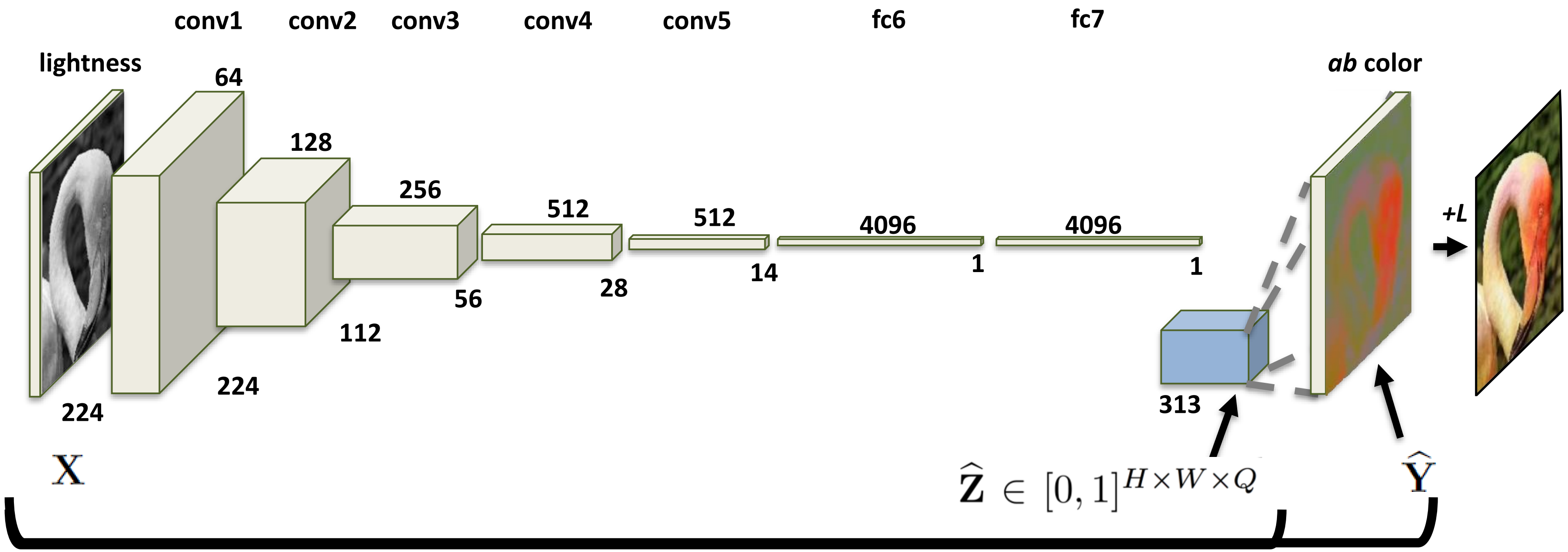


Charpiat et al. In ECCV 2008.



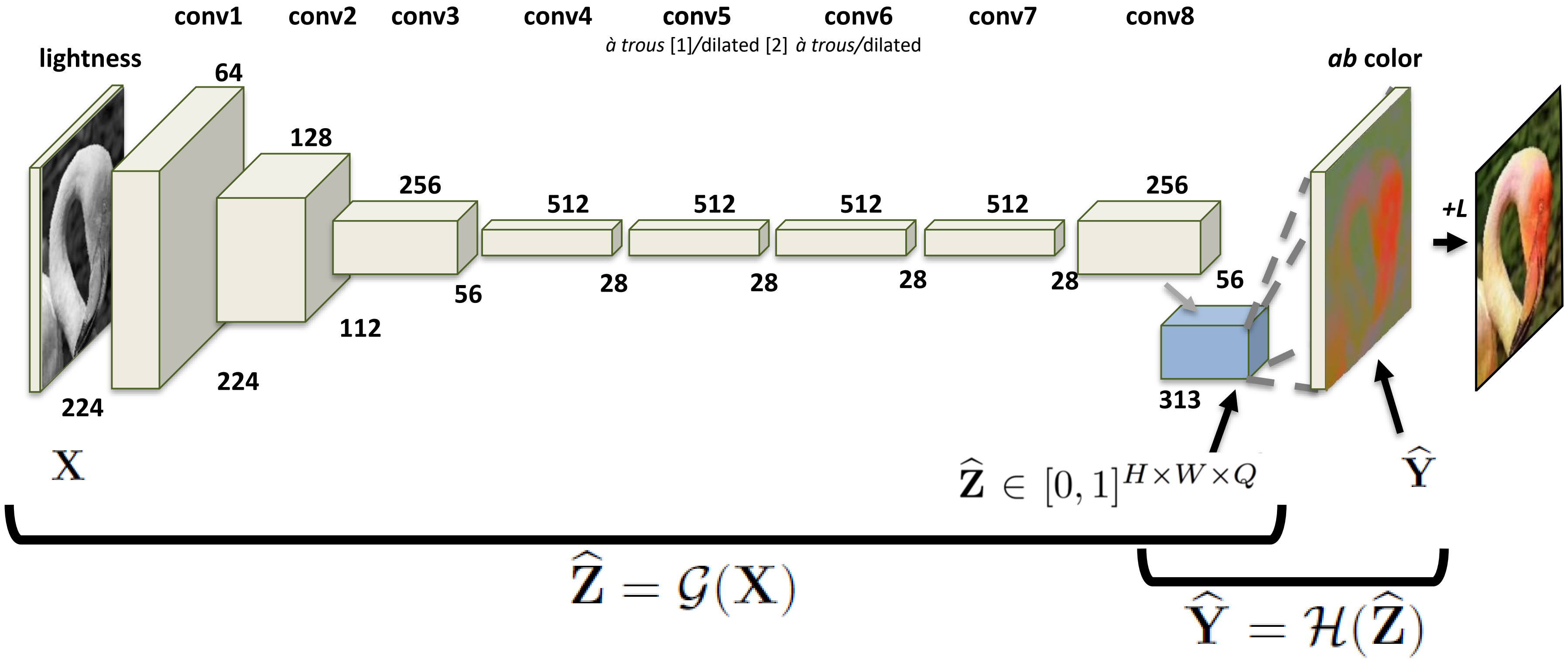
Larsson et al. In ECCV 2016. [Concurrent]

Network Architecture



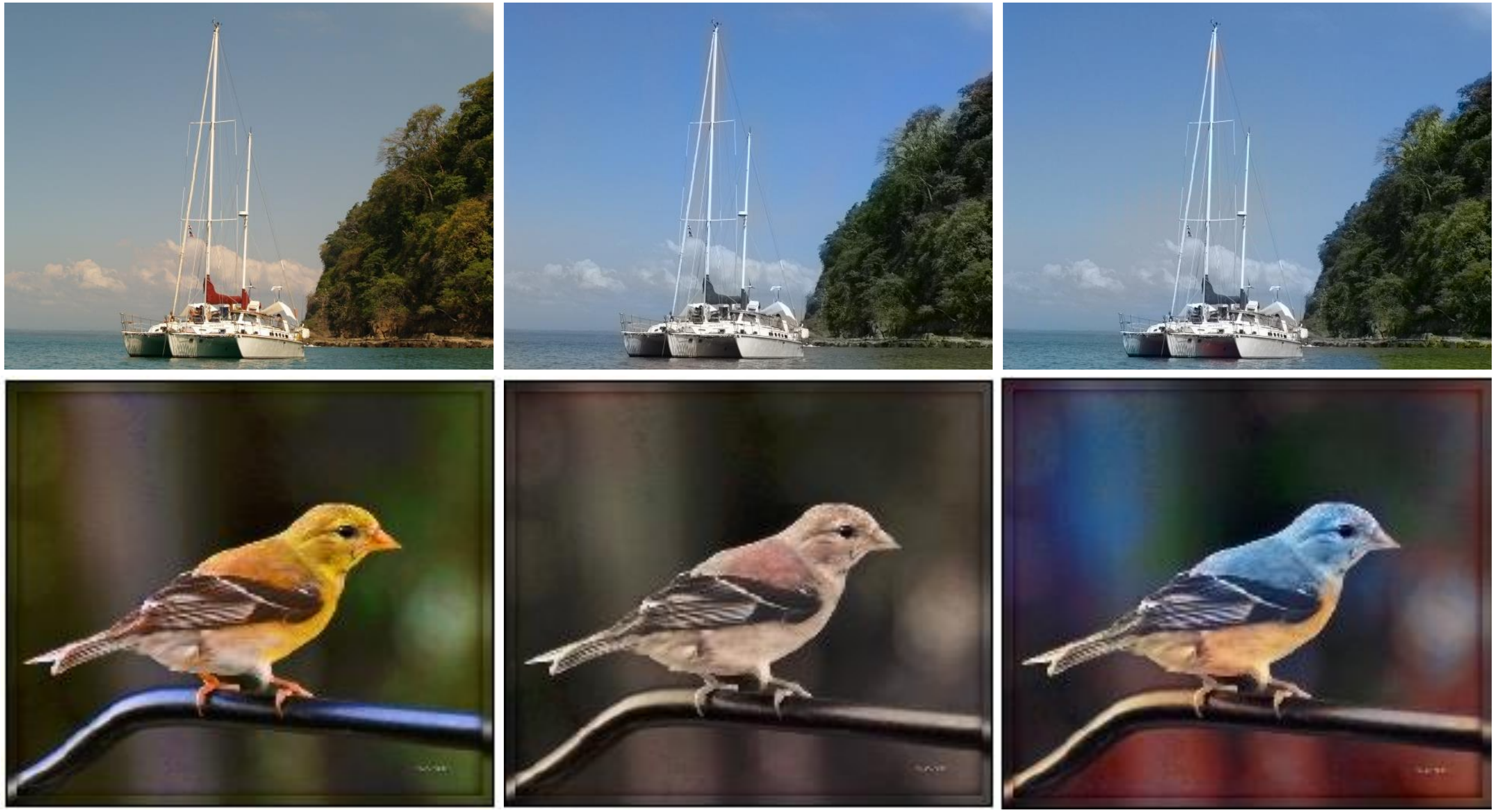
$$\widehat{\widehat{\mathbf{Y}}} = \mathcal{F}(\mathbf{X})$$

Network Architecture



[1] Chen *et al.* In arXiv, 2016.
[2] Yu and Koltun. In ICLR, 2016

Ground Truth / Input | L2 Regression | Class w/ Rebalancing



Failure Cases

Biases

Evaluation

	Visual Quality
	Per-pixel accuracy
Quantitative	

Evaluation

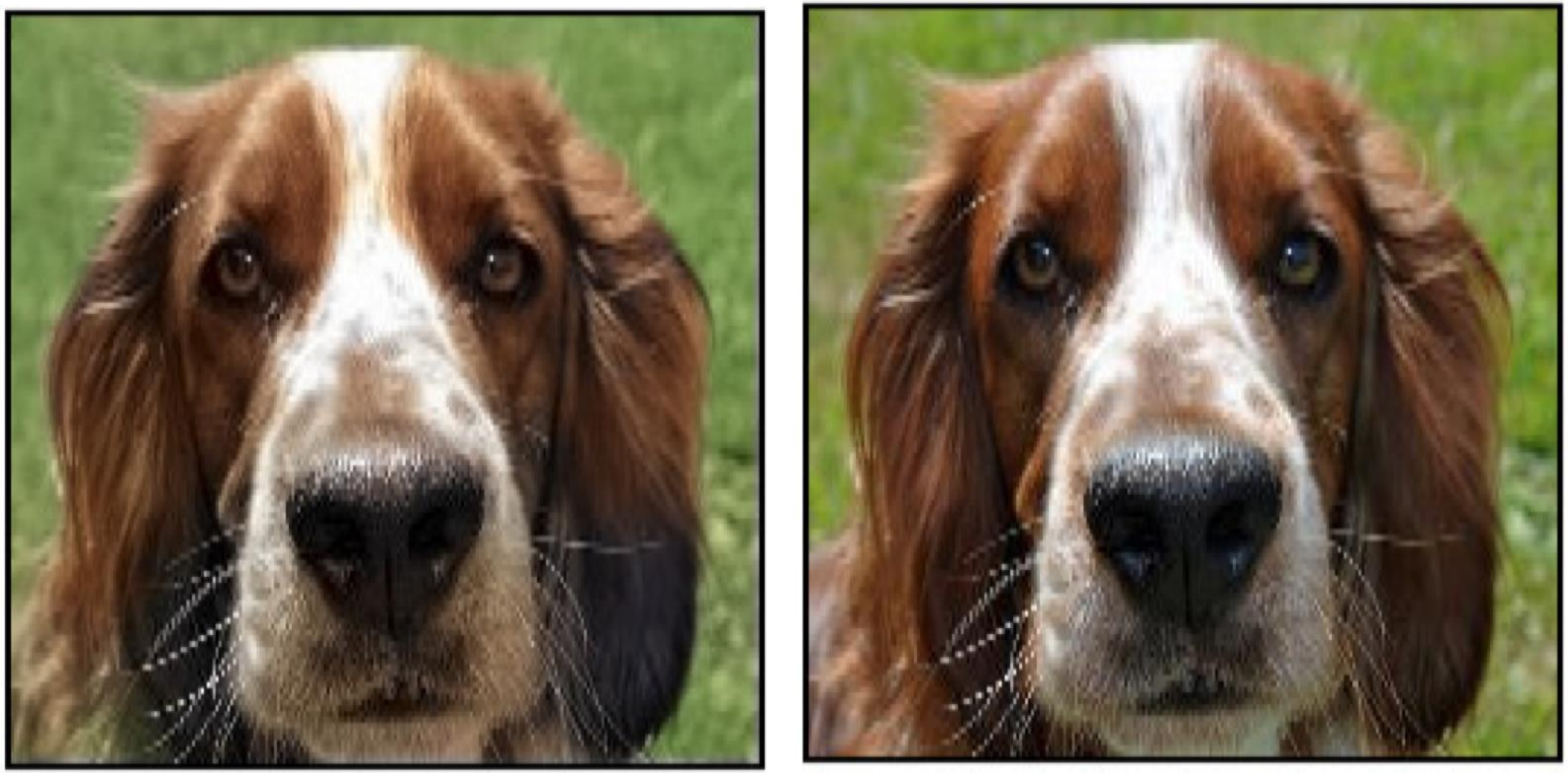
	Visual Quality	Representation Learning
Quantitative	Per-pixel accuracy Perceptual realism Semantic interpretability	Task generalization ImageNet classification Task & dataset generalization PASCAL classification, detection, segmentation
Qualitative	Low-level stimuli Legacy grayscale photos	Hidden unit activations

Evaluation

	Visual Quality	Representation Learning
Quantitative	Per-pixel accuracy **Perceptual realism** Semantic interpretability	Task generalization ImageNet classification Task & dataset generalization PASCAL classification, detection, segmentation
Qualitative	Low-level stimuli Legacy grayscale photos	Hidden unit activations

Perceptual Realism / Amazon Mechanical Turk Test

Fake, 0% fooled



Fake, 55% fooled


Fake, 58% fooled

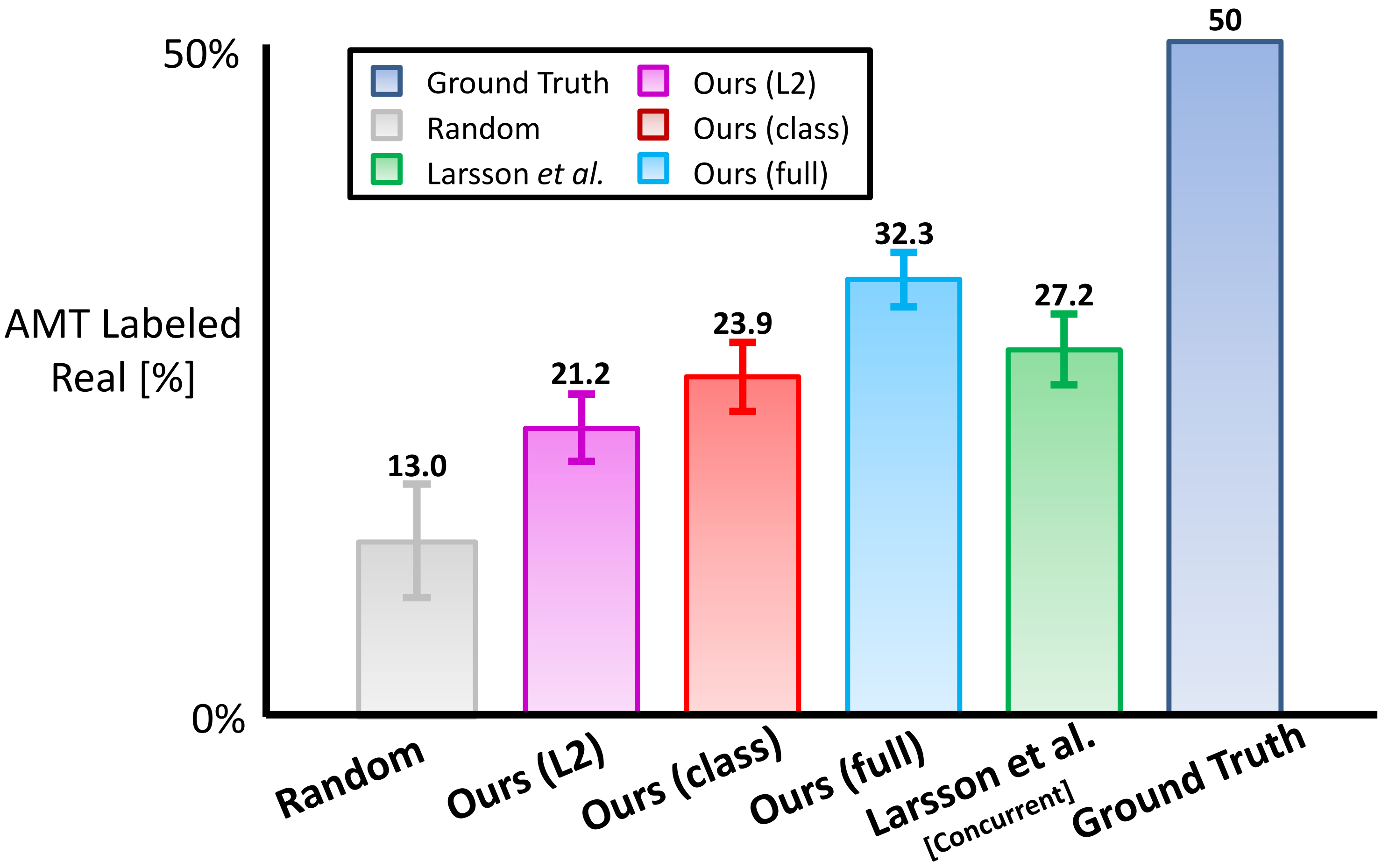
from Reddit /u/SherySantucci

Recolorized by Reddit ColorizeBot

Photo taken by Reddit /u/Timteroo, Mural from street artist Eduardo Kobra

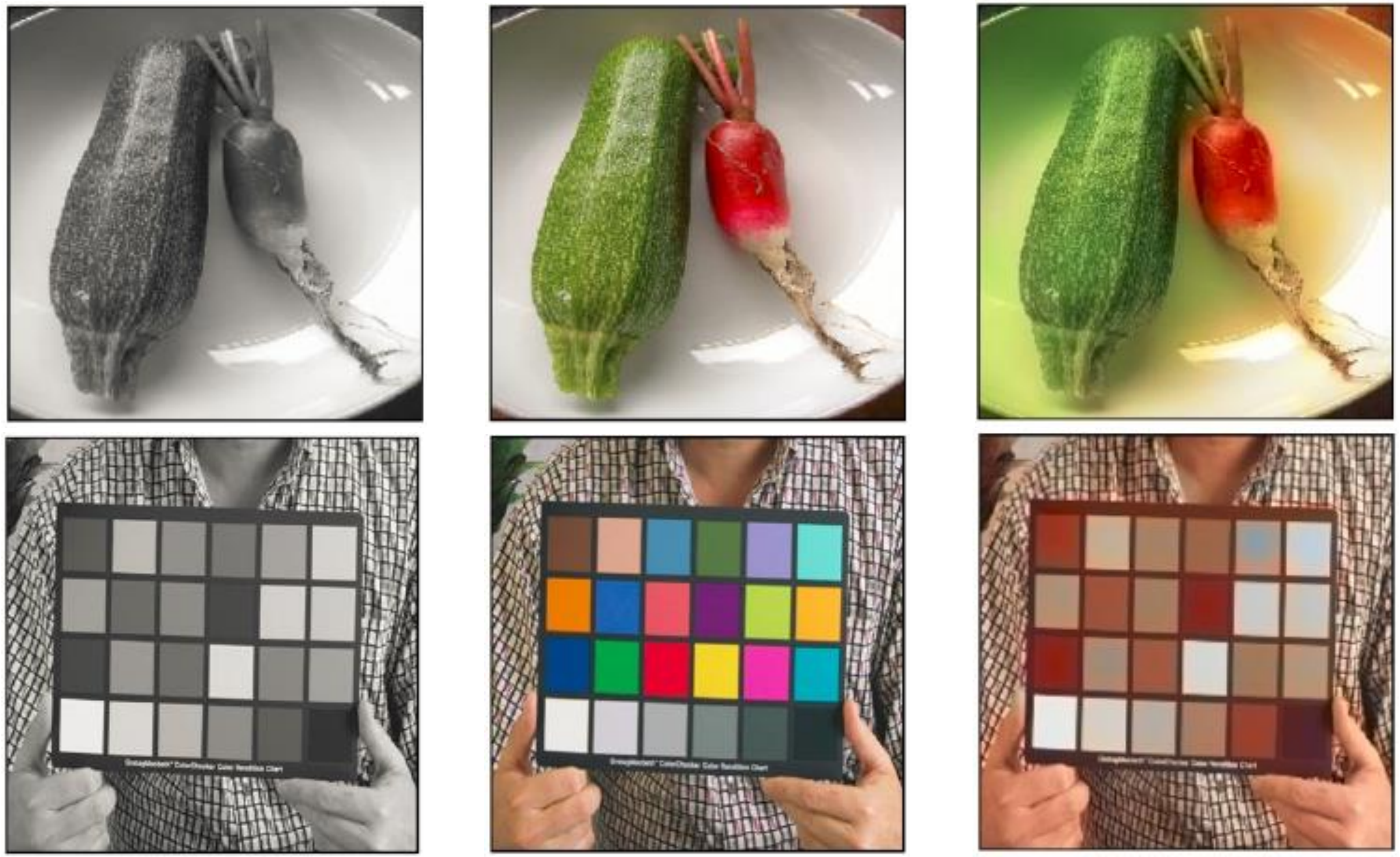
Recolorized by Reddit ColorizeBot

Perceptual Realism Test

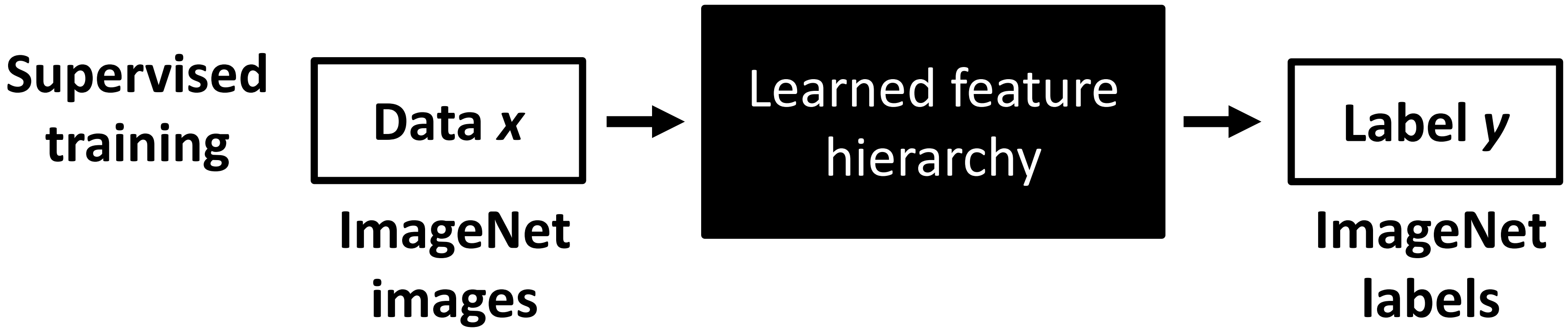


1600 images tested per algorithm

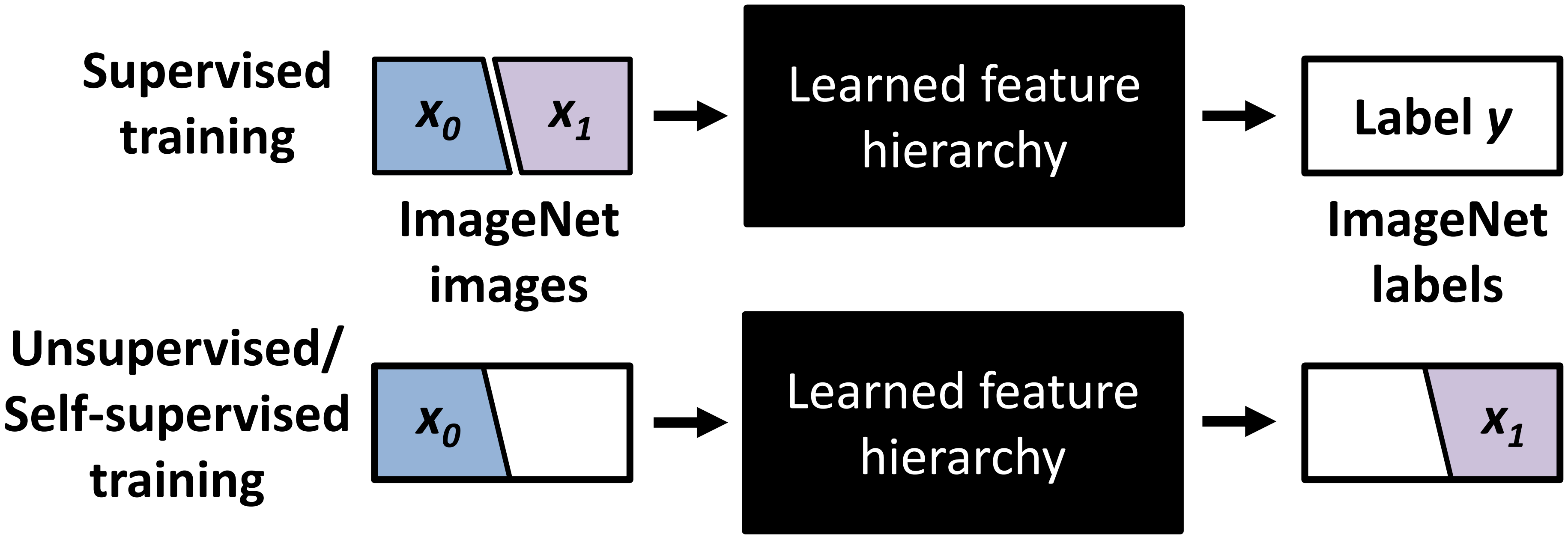
Input **Ground Truth** **Output**

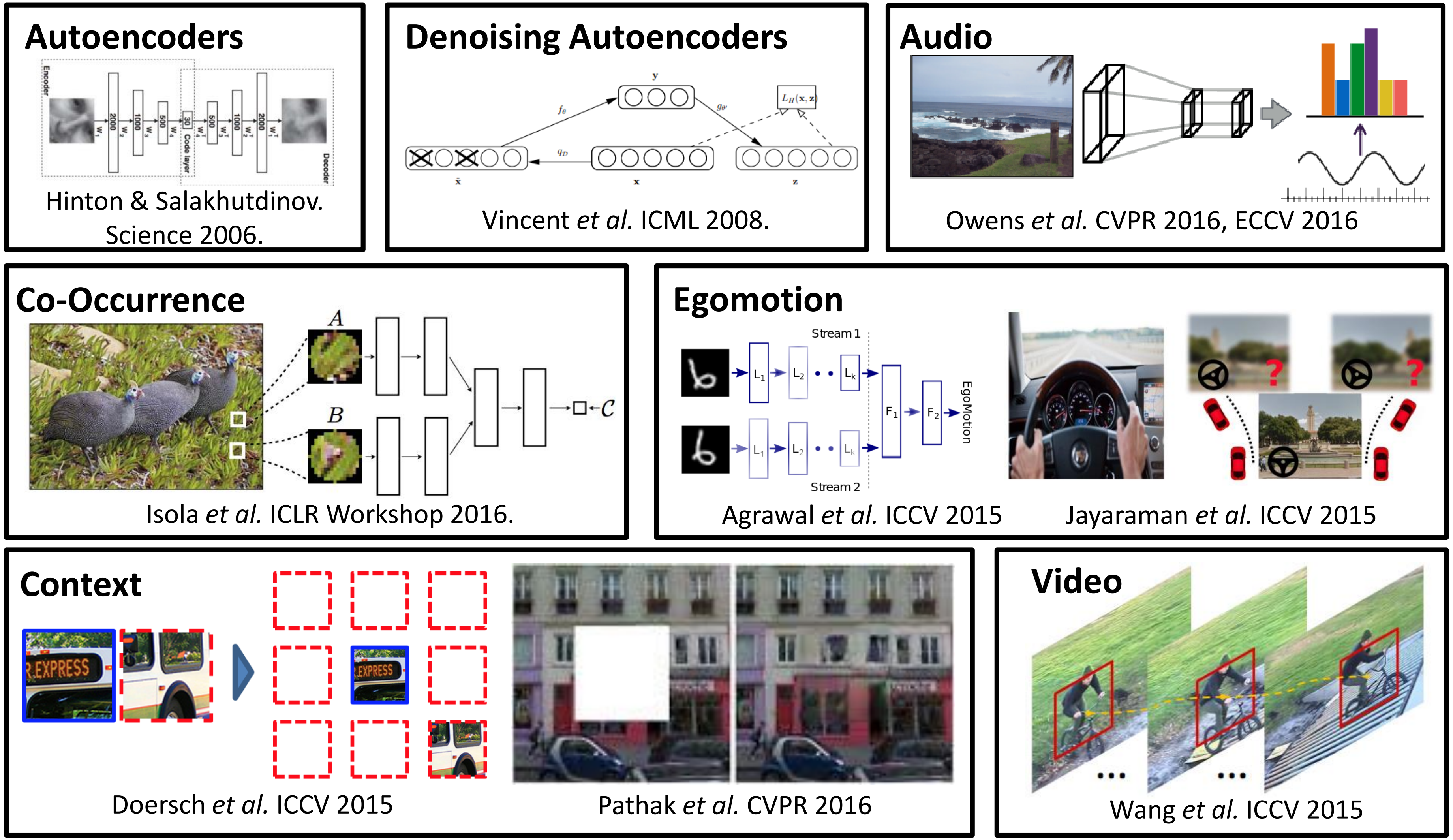


Predicting Labels from Data

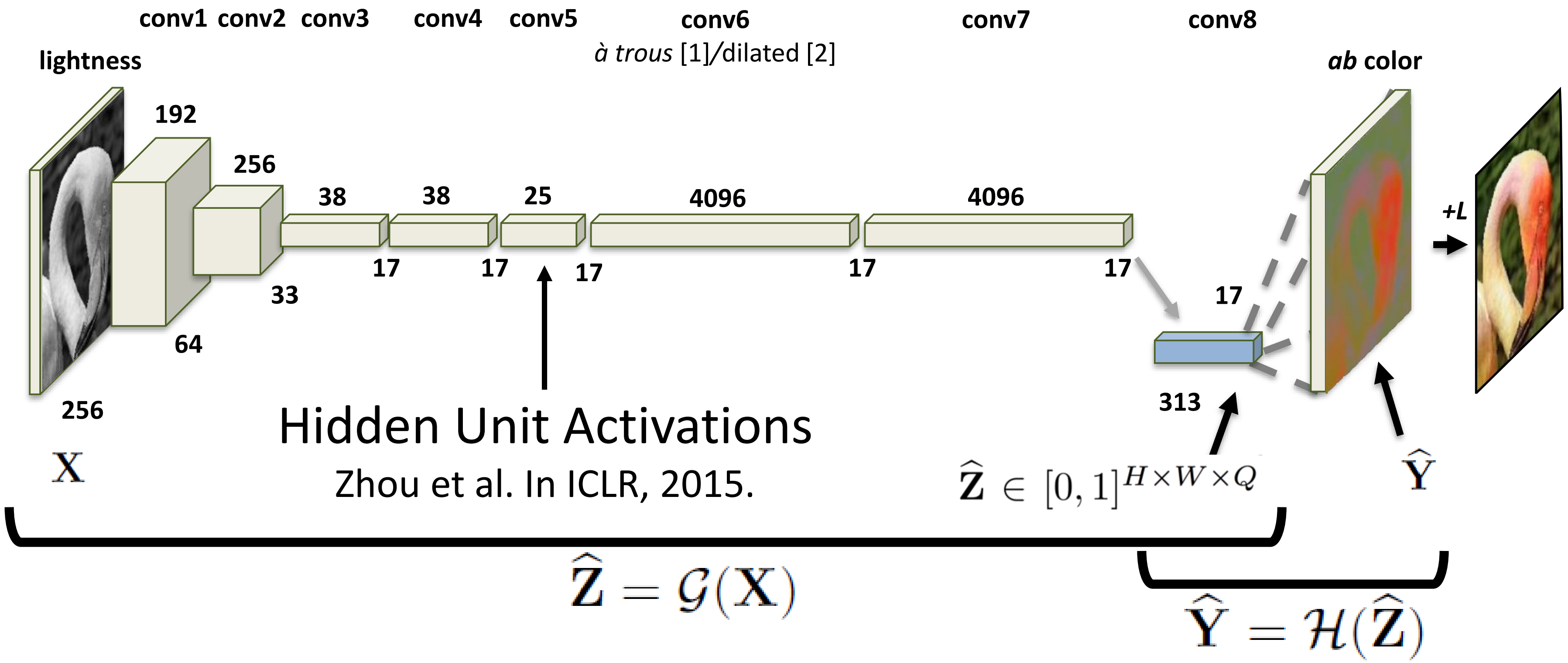


Predicting Data from Data



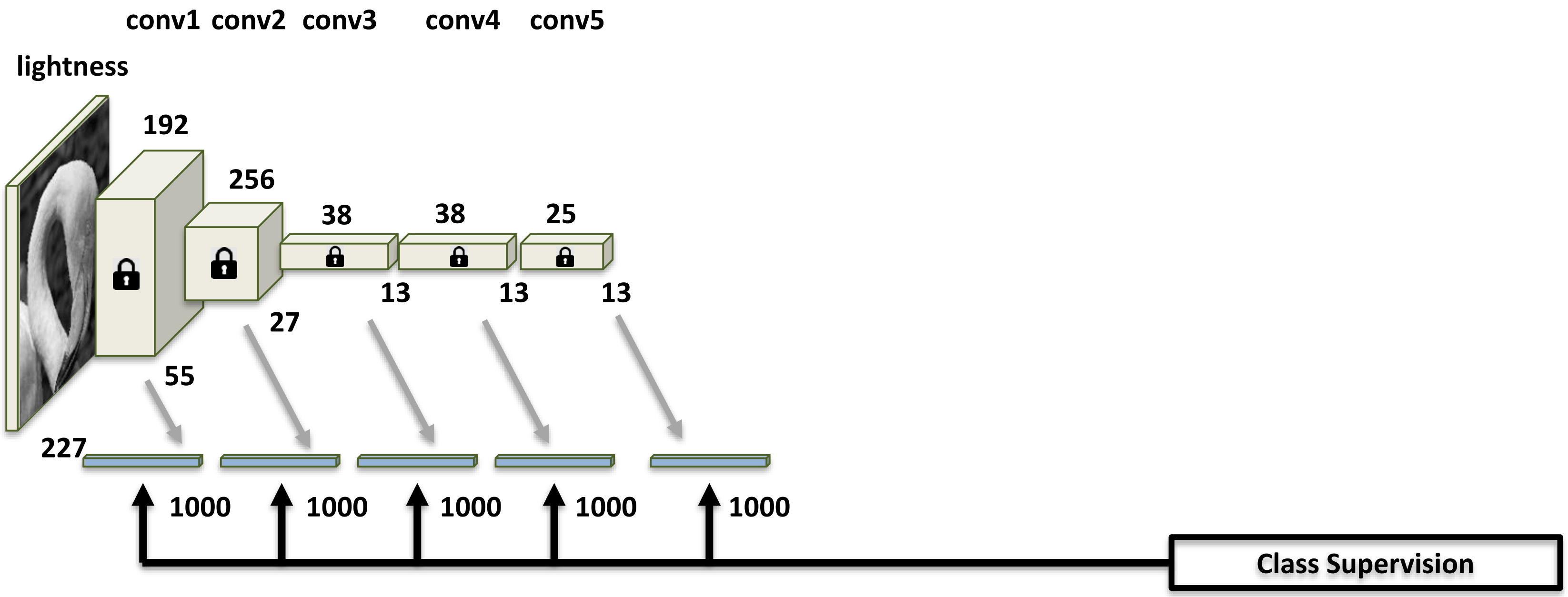

Autoencoders
Encoder
2000
1000
500
30
Code layer
Decoder
Hinton & Salakhutdinov. Science 2006.
Denoising Autoencoders
y
x
z
Vincent et al. ICML 2008.
Audio
Owens et al. CVPR 2016, ECCV 2016
Co-Occurrence
A
B
C
Isola et al. ICLR Workshop 2016.
Egomotion
Stream 1
Stream 2
EgoMotion
Agrawal et al. ICCV 2015
Jayaraman et al. ICCV 2015
Context
EXPRESS
Doersch et al. ICCV 2015
Pathak et al. CVPR 2016
Video
Wang et al. ICCV 2015

Cross-Channel Encoder



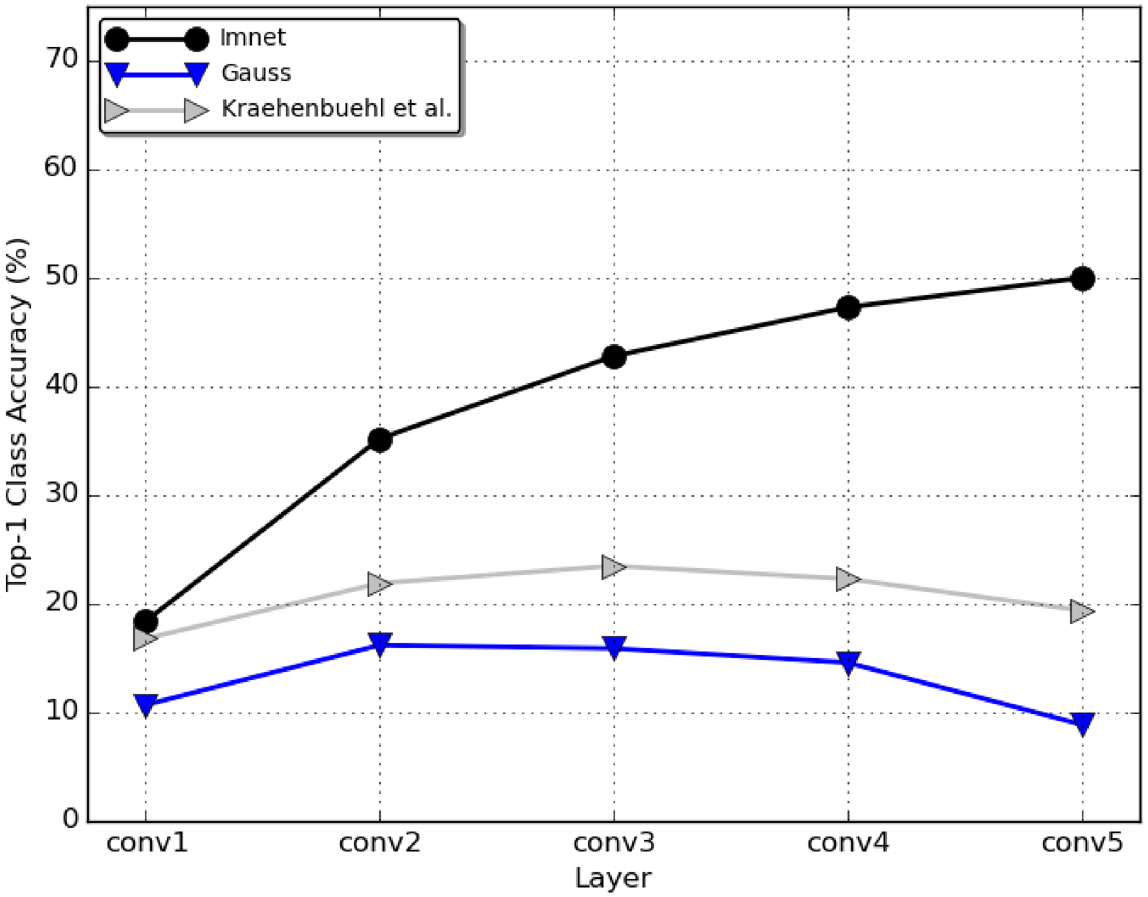
[1] Chen *et al.* In arXiv, 2016.
[2] Yu and Koltun. In ICLR, 2016

Task Generalization: ILSVRC linear classification

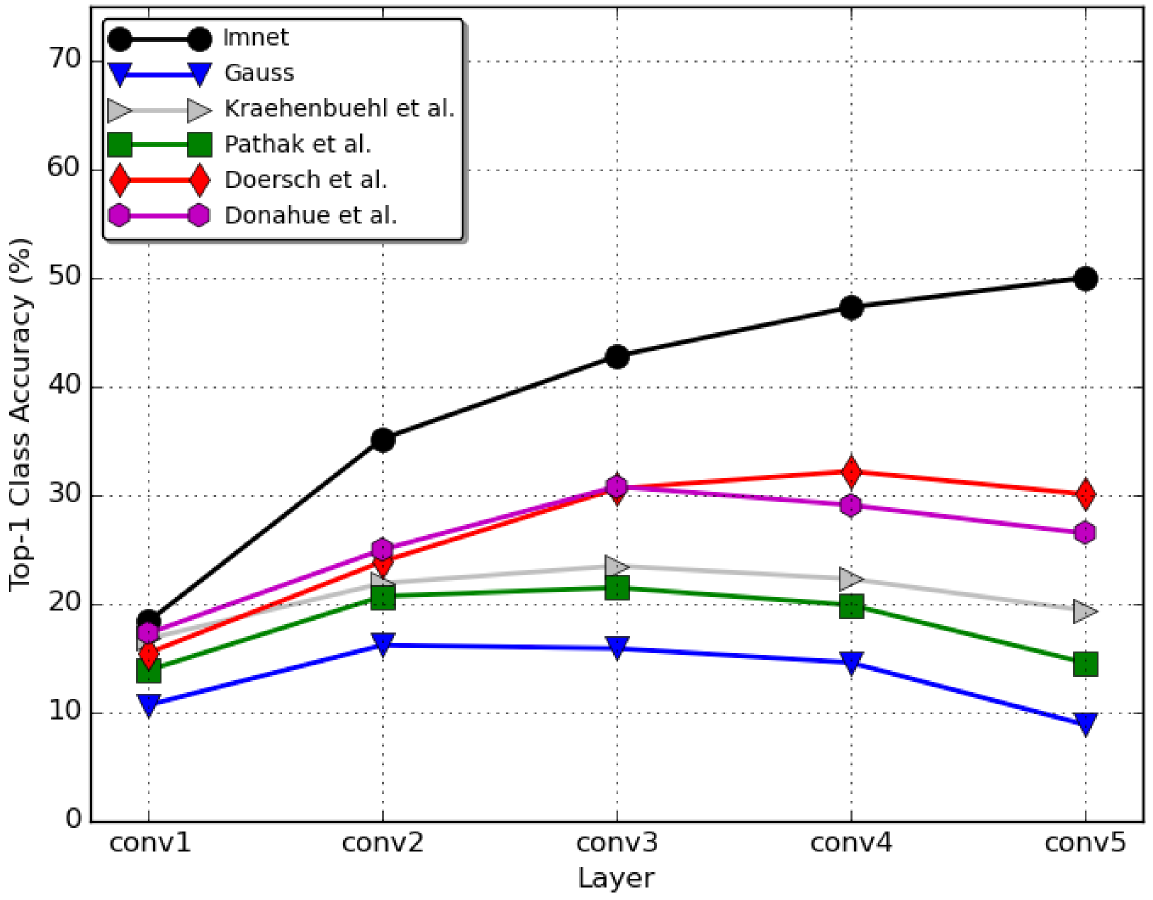


Are semantic classes ***linearly separable***
in the learned feature space?

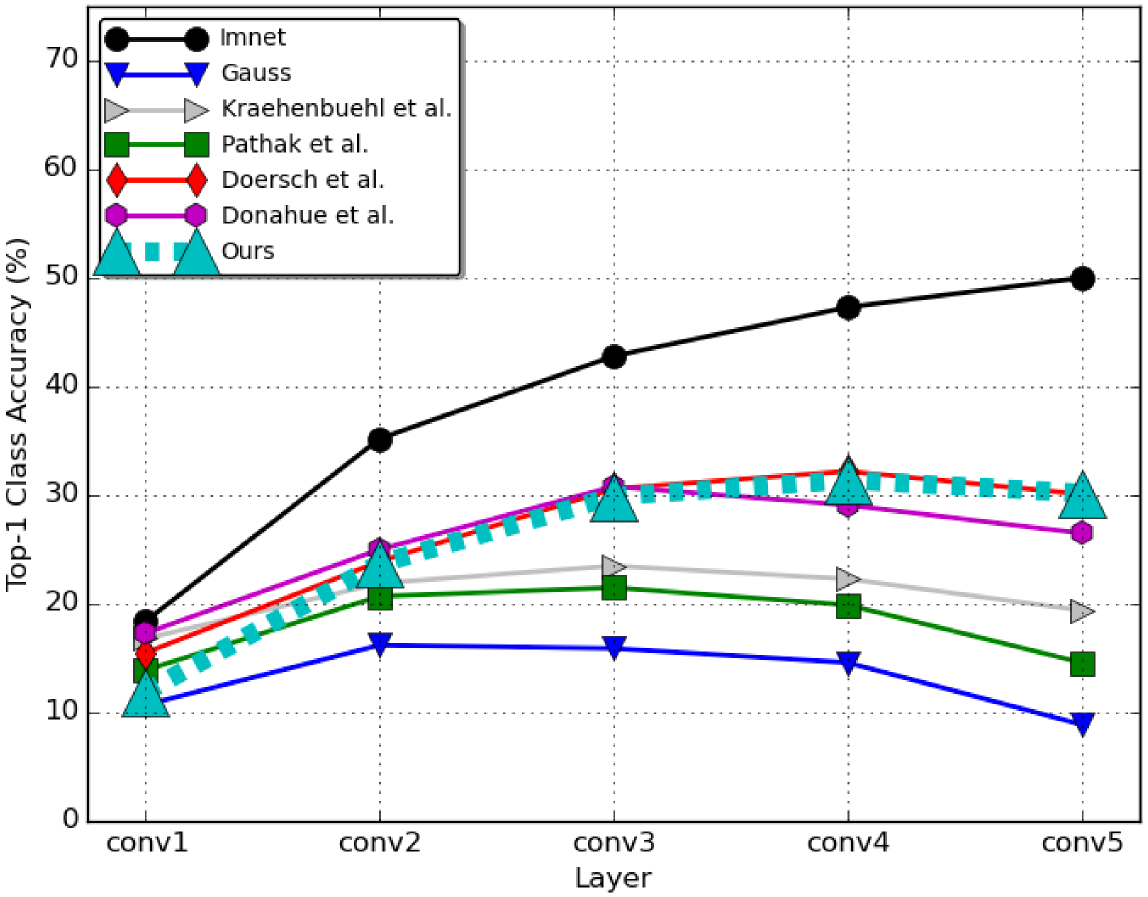
Task Generalization: ILSVRC linear classification



Task Generalization: ILSVRC linear classification

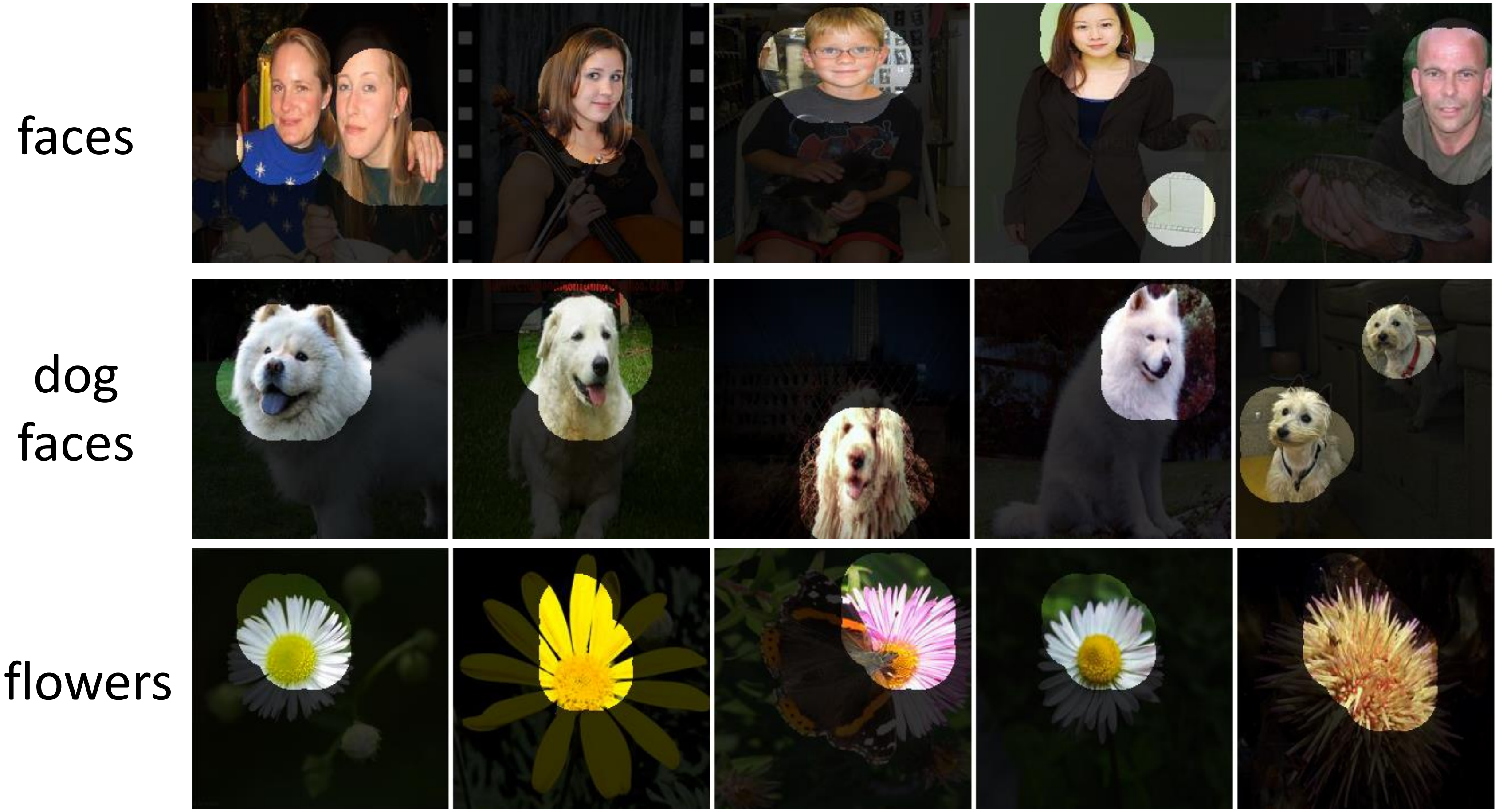


Task Generalization: ILSVRC linear classification



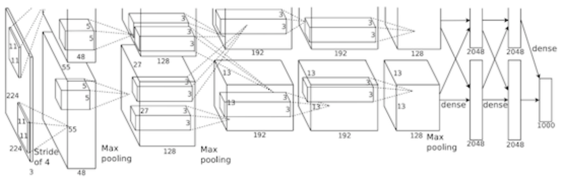
Hidden Unit (conv5) Activations

Hidden Unit (conv5) Activations



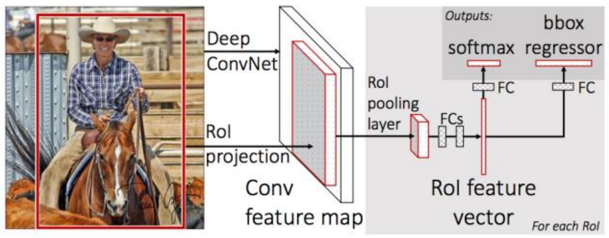
Dataset & Task Generalization on PASCAL VOC

Does the feature representation ***transfer*** to other datasets and tasks?



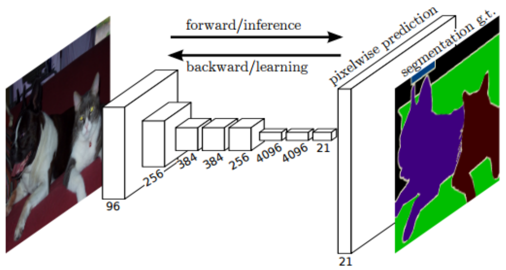
Classification

Krähenbühl et al. In ICLR, 2016.



Detection

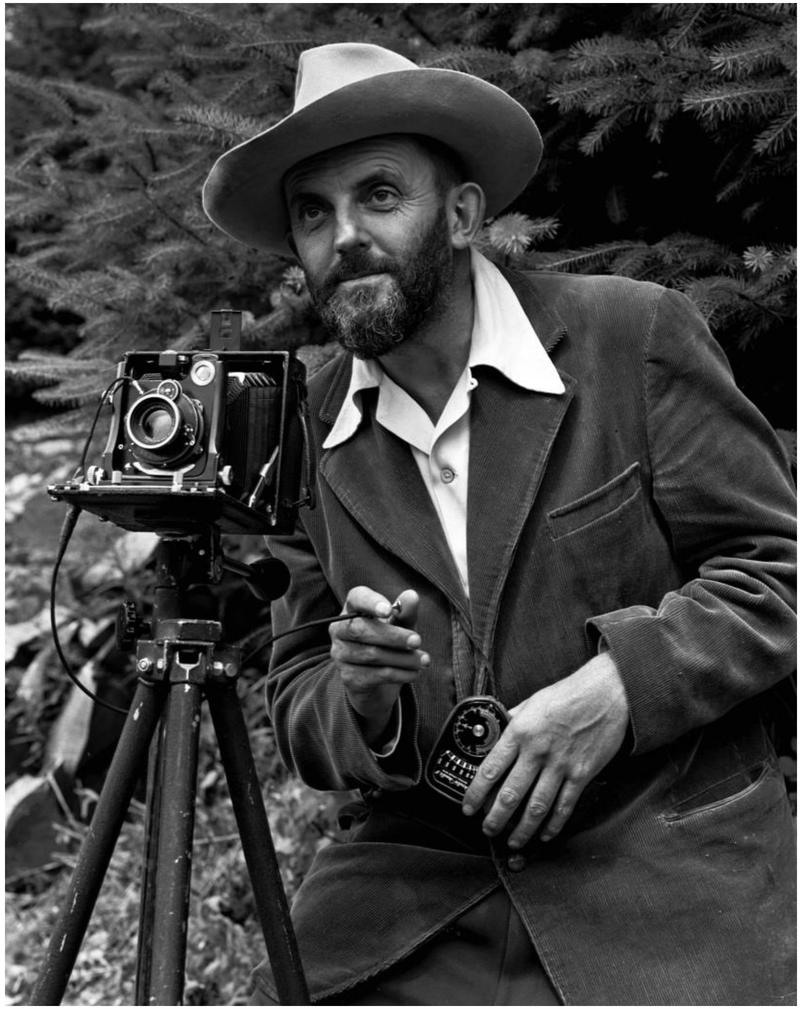
Fast R-CNN. Girshick. In ICCV, 2015.

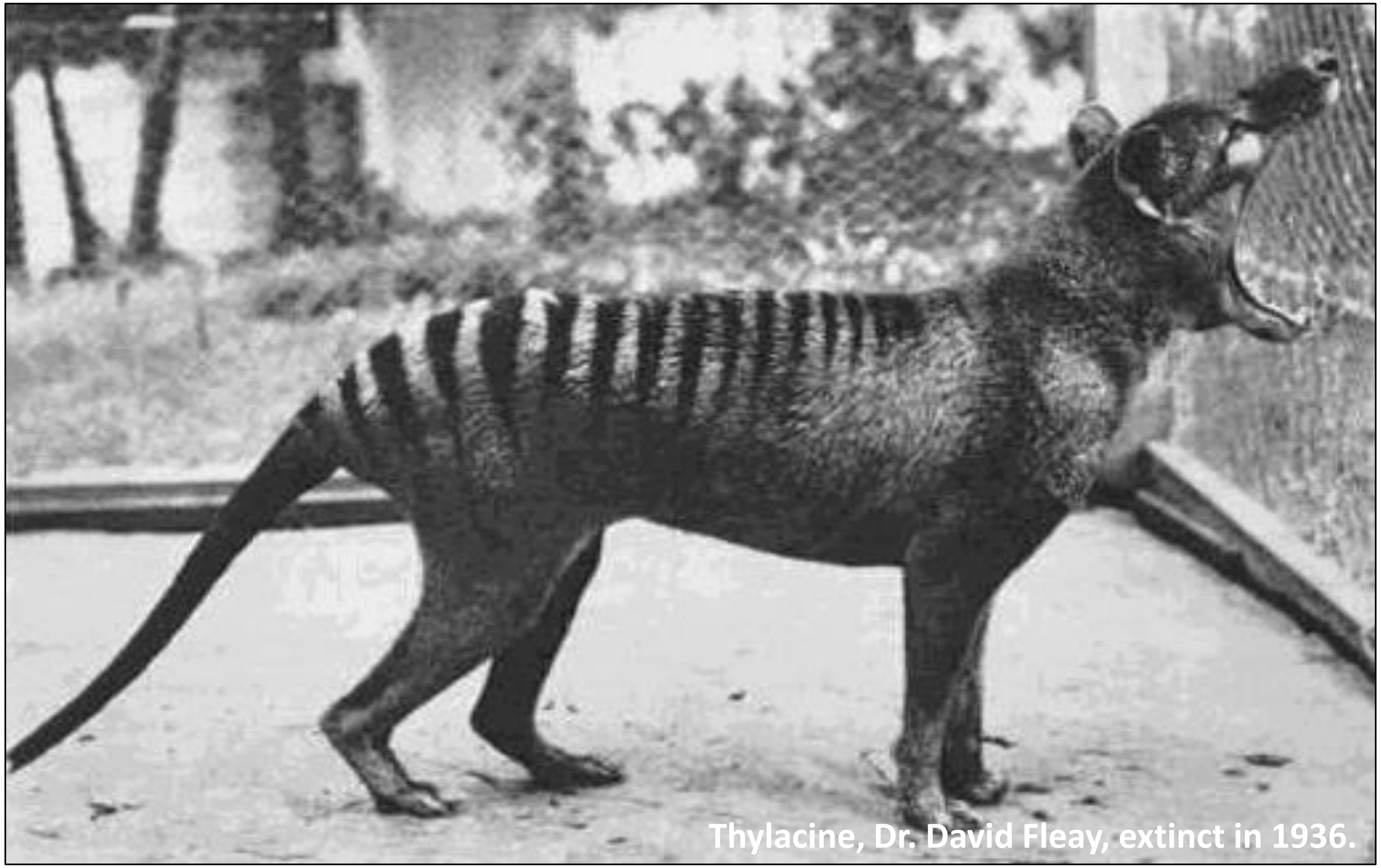


Segmentation

FCNs. Long et al. In CVPR, 2015.

Does the method work on ***legacy*** black and white photos?



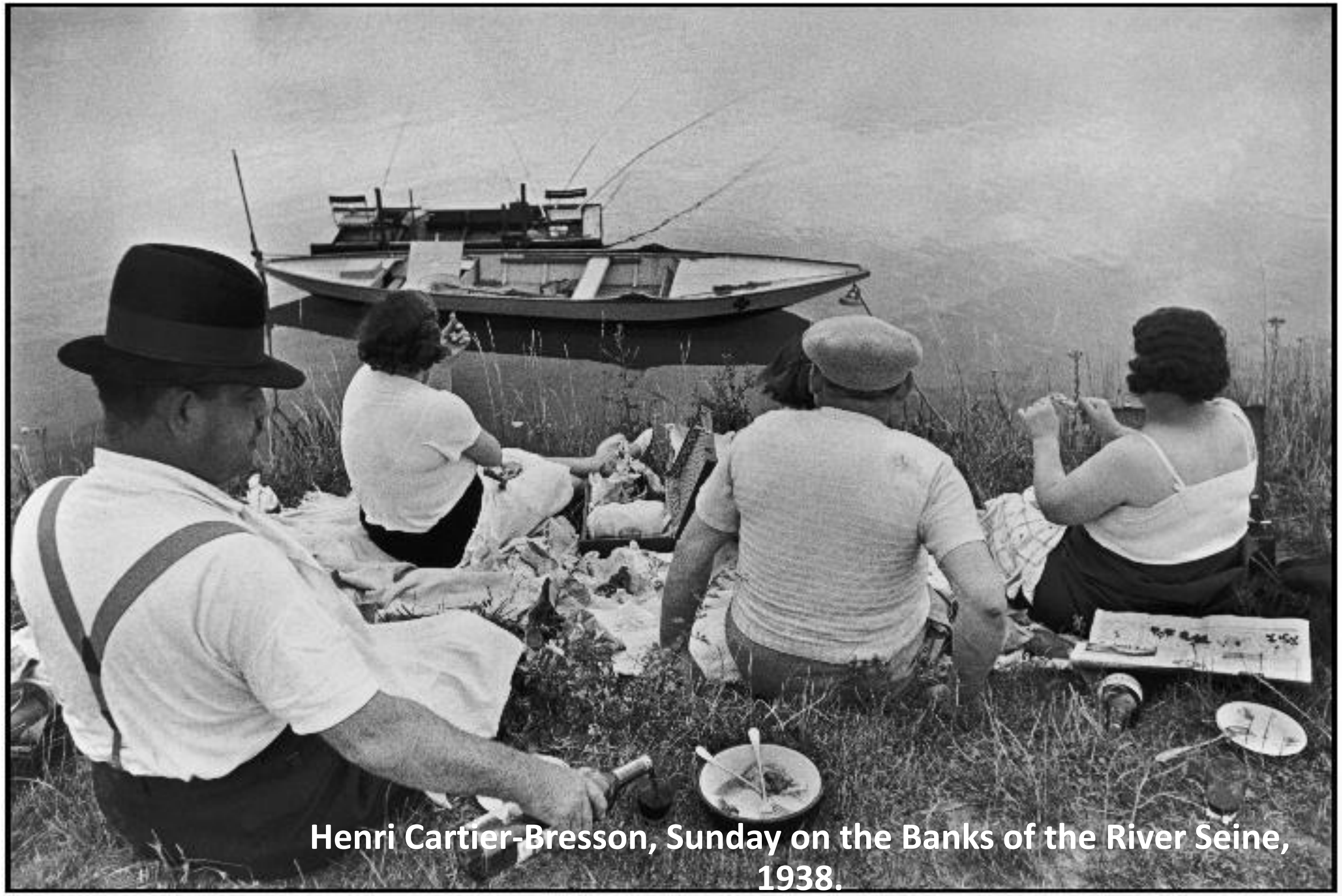


Thylacine, Dr. David Fleay, extinct in 1936.


Thylacine, Dr. David Fleay, extinct in 1936.

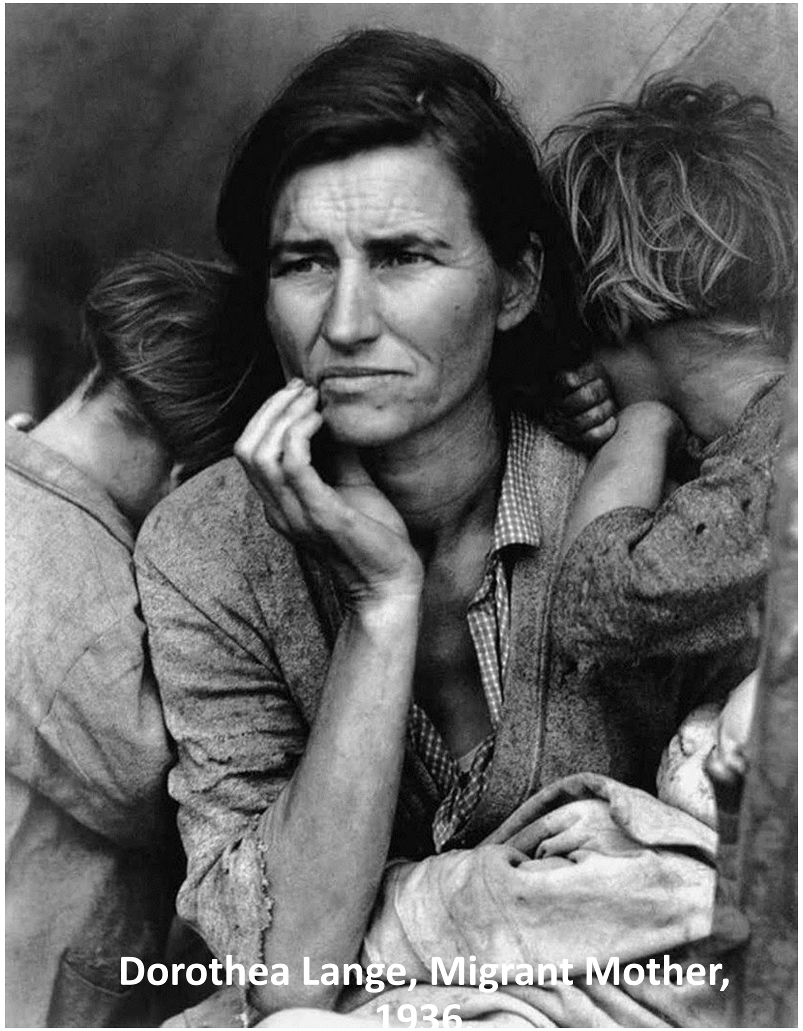
Amateur Family Photo, 1956

Amateur Family Photo, 1956

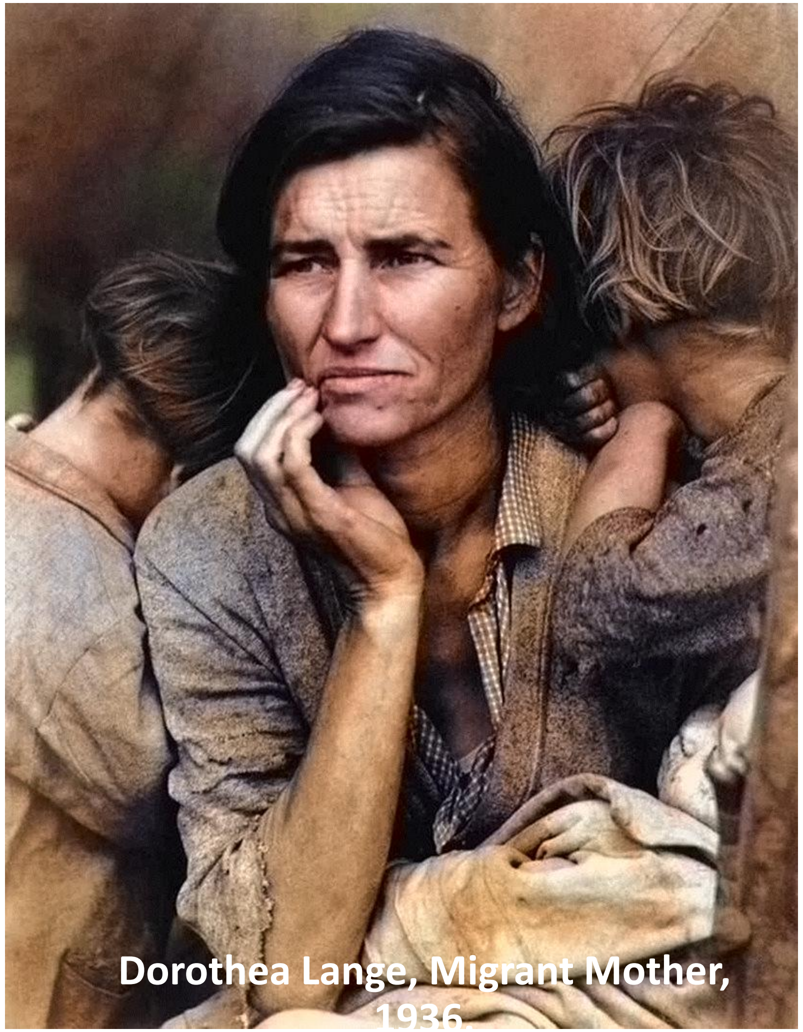


Henri Cartier-Bresson, Sunday on the Banks of the River Seine, 1938.

Henri Cartier-Bresson, Sunday on the Banks of the River Seine, 1938.



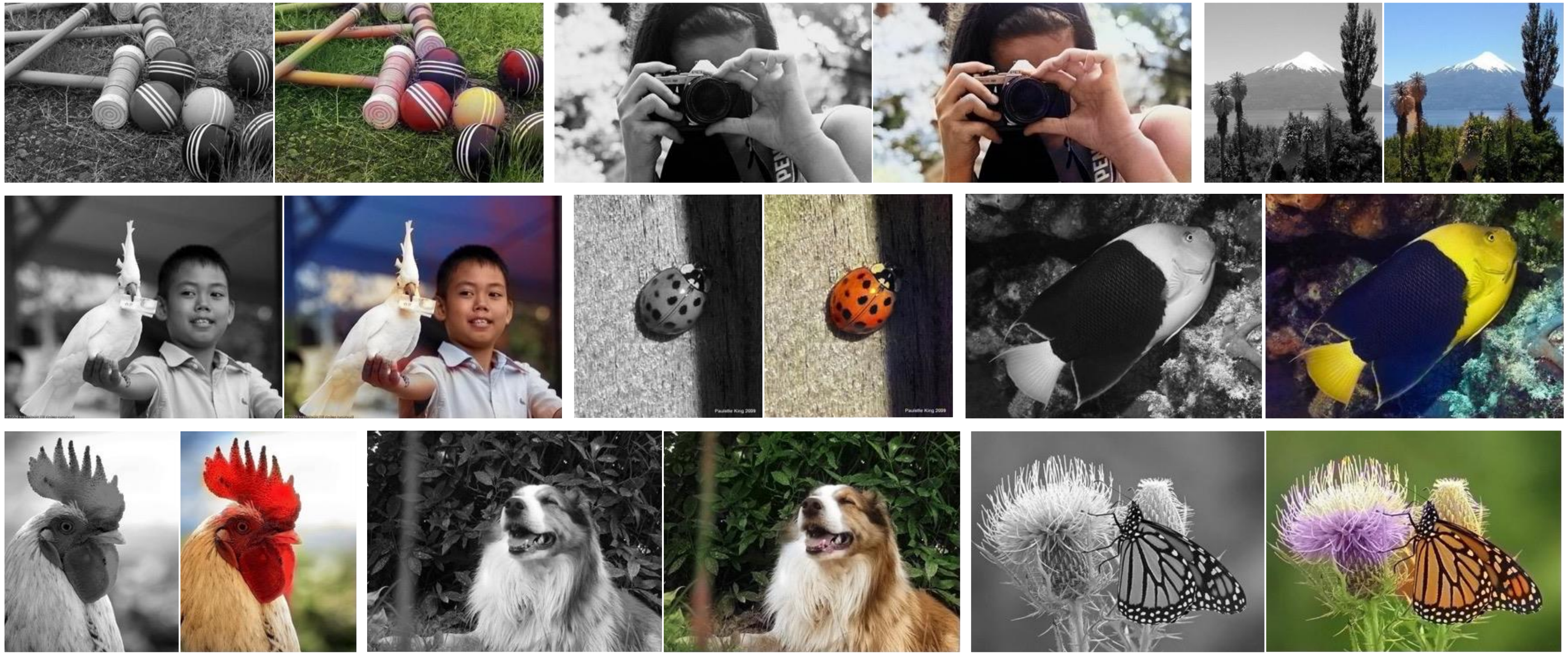
Dorothea Lange, Migrant Mother, 1936



Dorothea Lange, Migrant Mother, 1936.

Additional Information

- Demo
  - http://demos.algorithmia.com/colorize-photos/
- Reddit ColorizeBot
  - Type “colorizebot” under any image post
- Code
  - https://github.com/richzhang/colorization
- Website – full paper, user examples, visualizations
  - http://richzhang.github.io/colorization



For the full paper, additional examples and our model:
richzhang.github.io/colorization

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

SimCLR, IMCL 2020

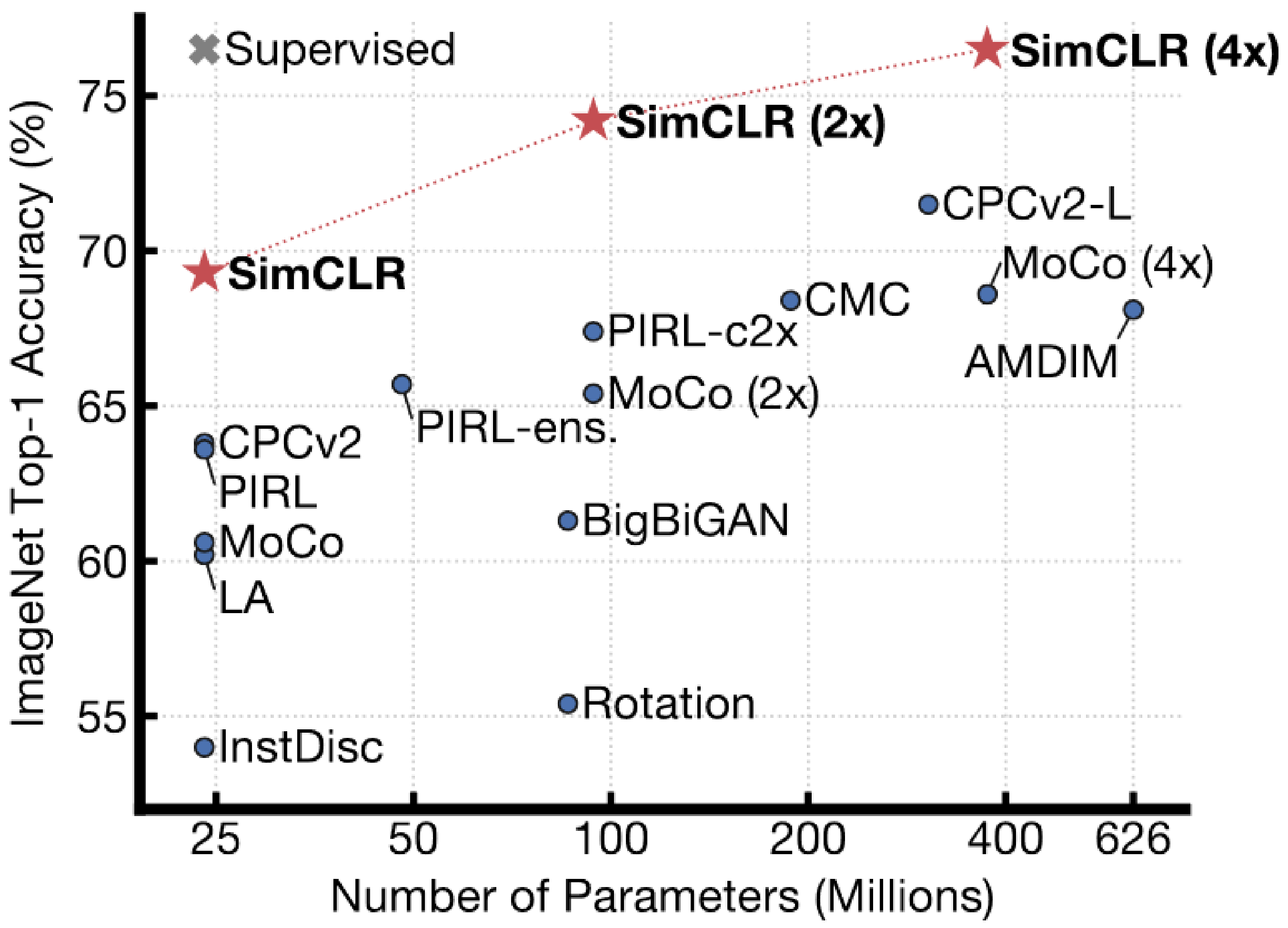
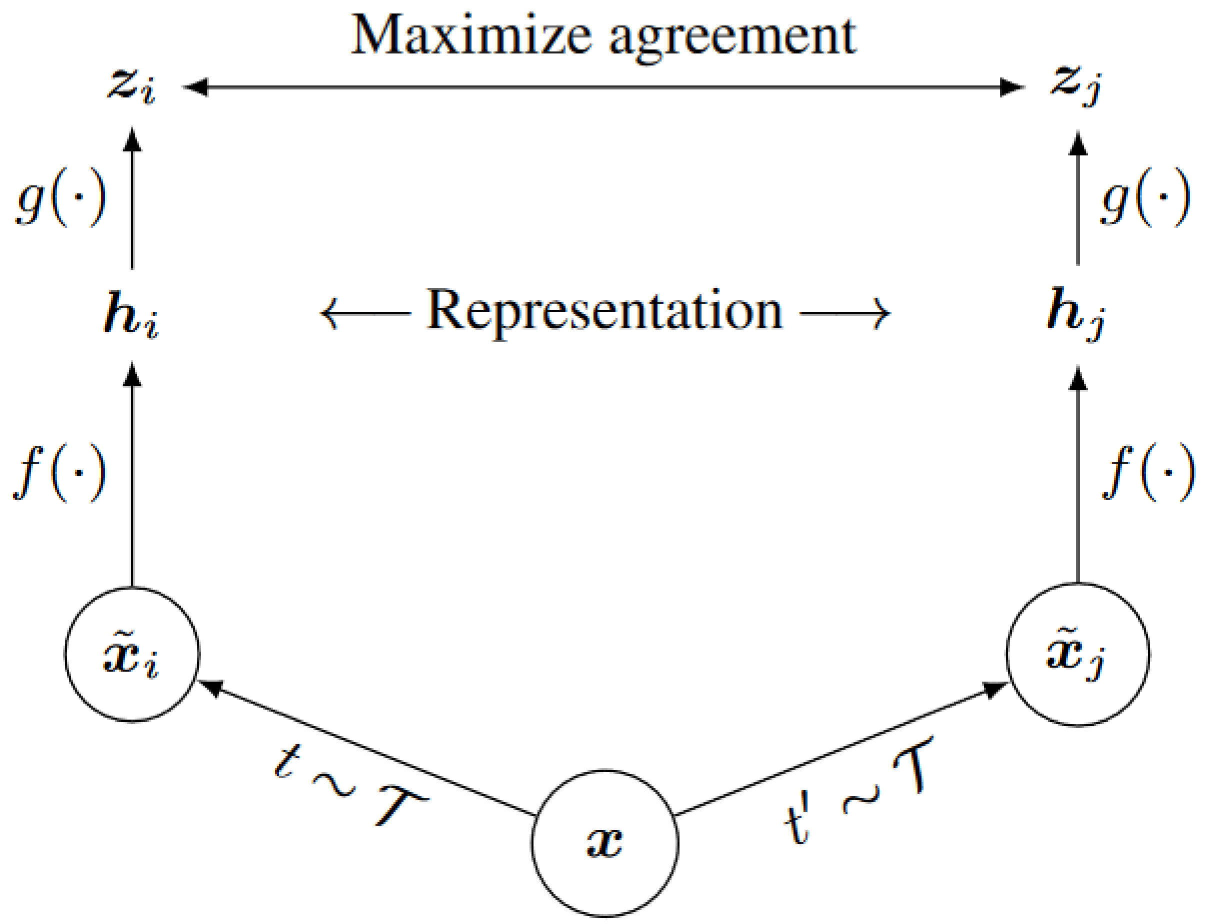
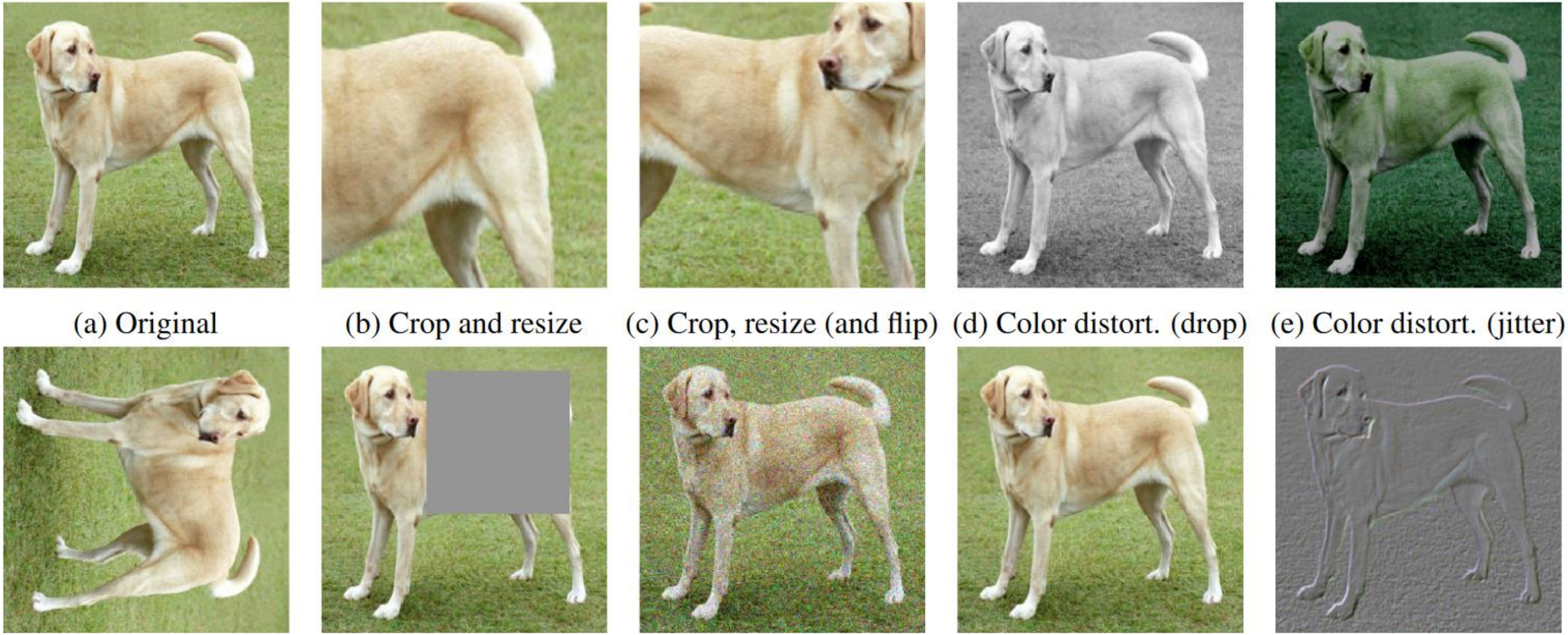


Figure 1. ImageNet Top-1 accuracy of linear classifiers trained on representations learned with different self-supervised methods (pretrained on ImageNet). Gray cross indicates supervised ResNet-50. Our method, SimCLR, is shown in bold.



Maximize agreement
z_i
z_j
$g(\cdot)$
$g(\cdot)$
h_i
$\longleftarrow$ Representation $\longrightarrow$
h_j
$f(\cdot)$
$f(\cdot)$
$\tilde{x}_i$
$\tilde{x}_j$
$t \sim \mathcal{T}$
$t' \sim \mathcal{T}$
x



(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

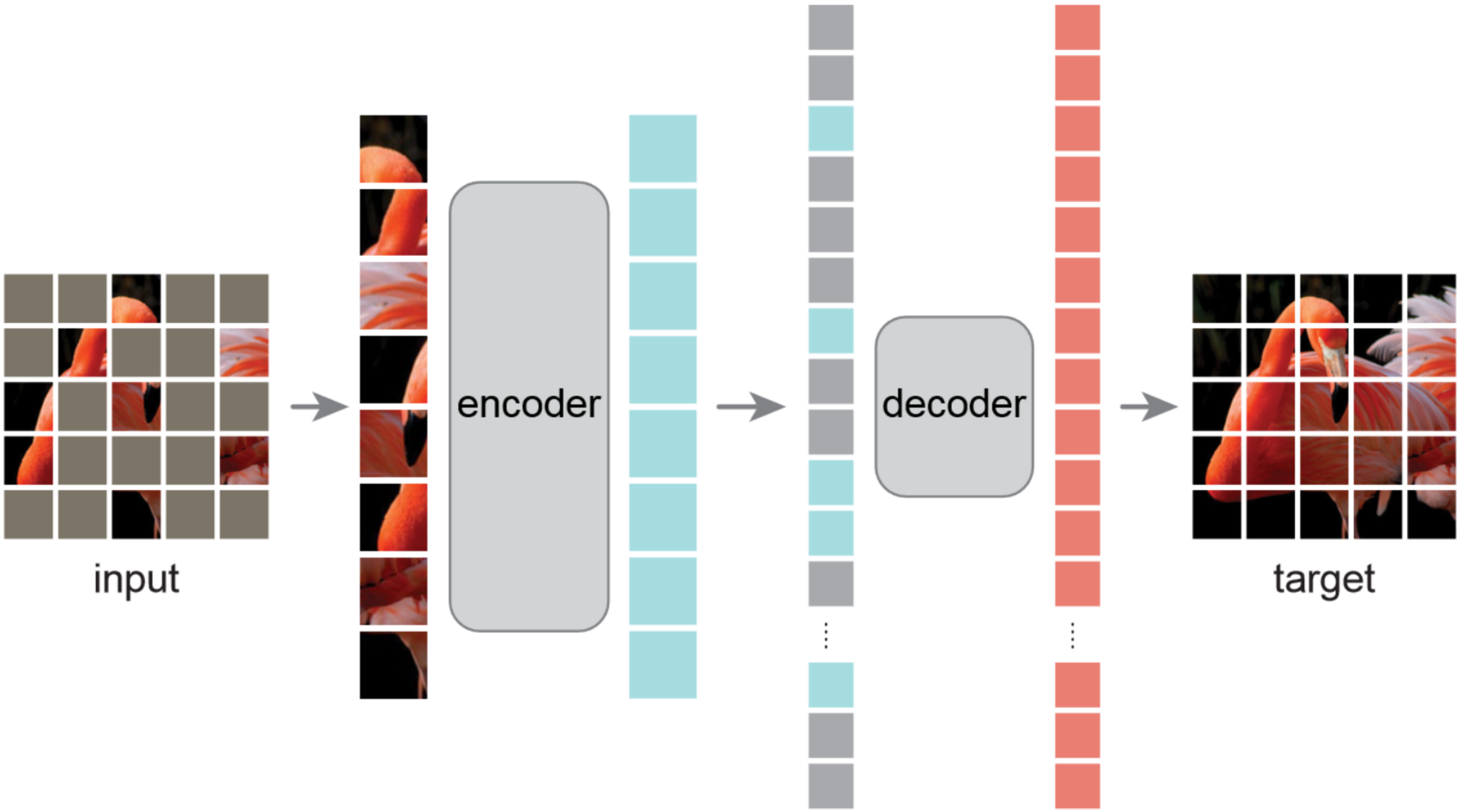
Masked Autoencoders Are Scalable Vision Learners

Kaiming He[*,†] Xinlei Chen[*] Saining Xie Yanghao Li Piotr Dollár Ross Girshick

[*]equal technical contribution [†]project lead

Facebook AI Research (FAIR)

CVPR 2022



input
encoder
decoder
target

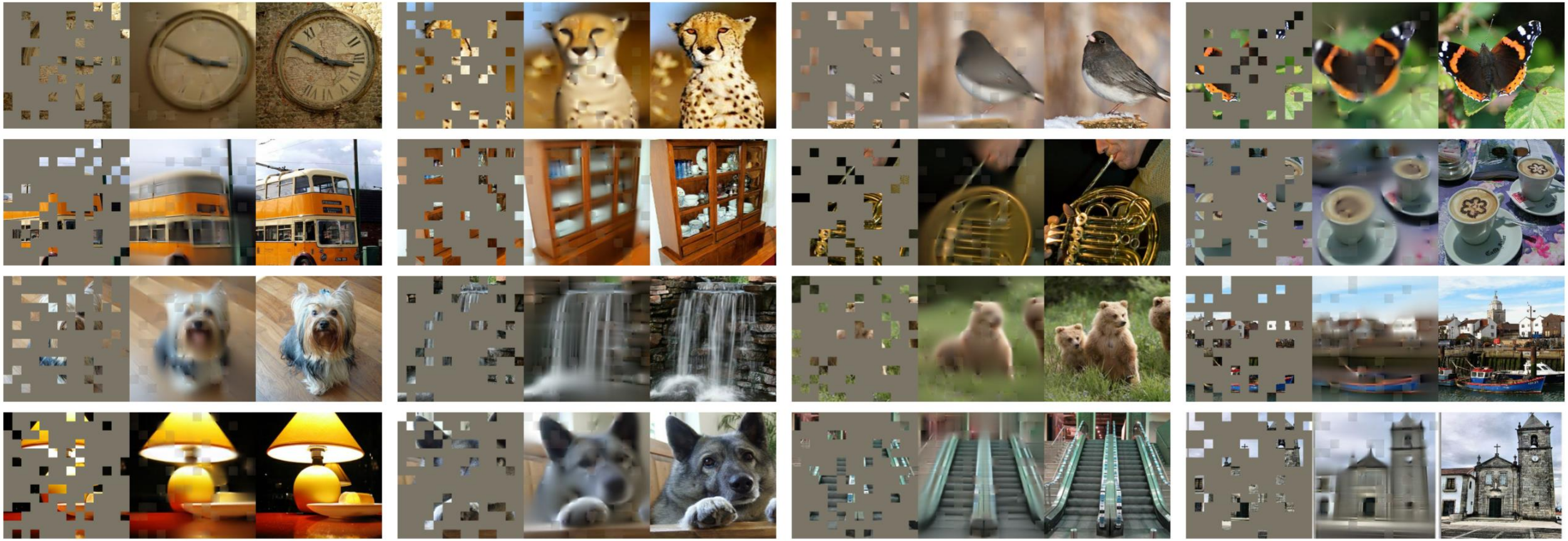
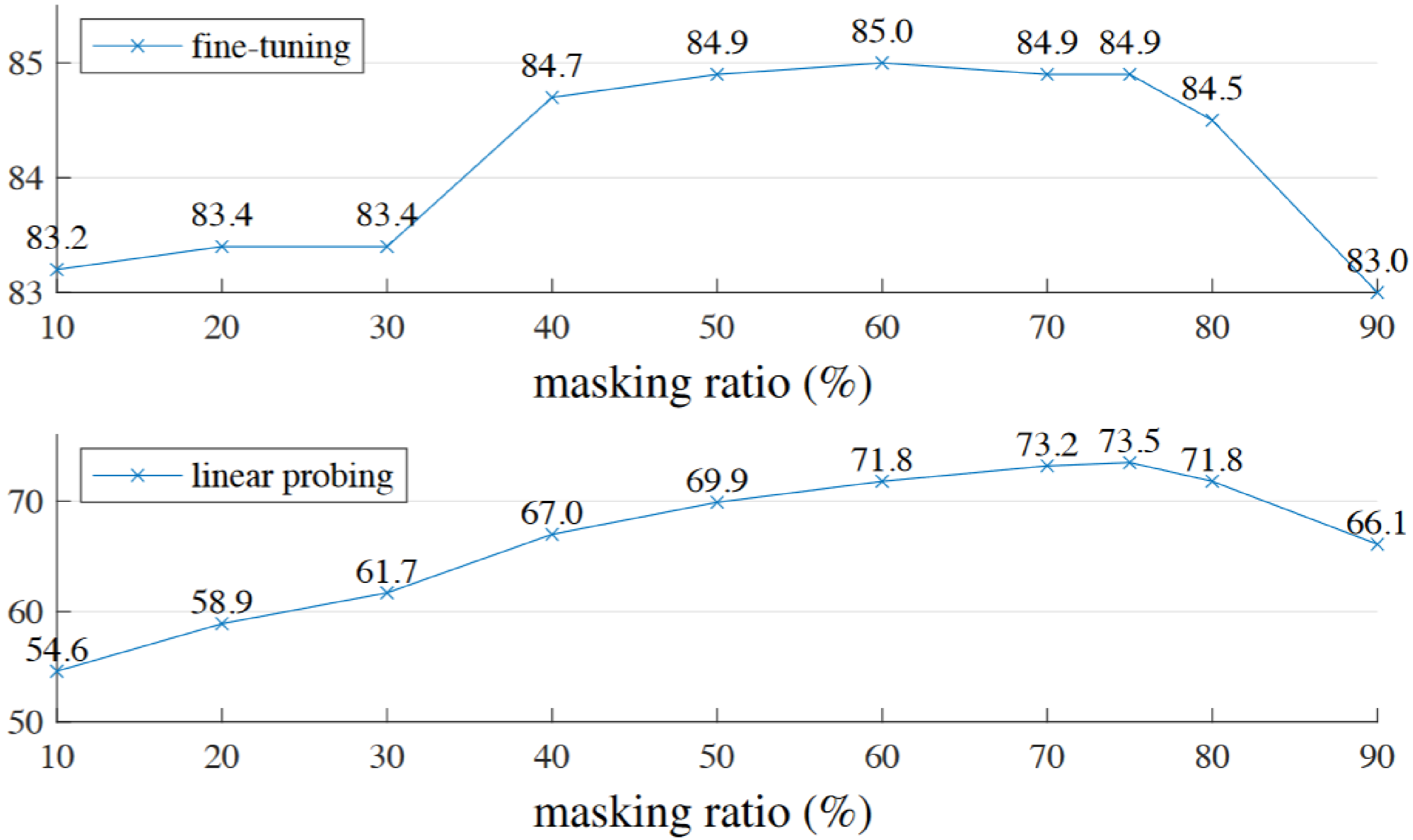


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix. †*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

Figure 3. Example results on COCO validation images, using an MAE trained on ImageNet (the same model weights as in Figure 2). Observe the reconstructions on the two right-most examples, which, although different from the ground truth, are semantically plausible.

Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.


fine-tuning
85
84
83
83.2
83.4
83.4
84.7
84.9
85.0
84.9
84.9
84.5
83.0
10
20
30
40
50
60
70
80
90
masking ratio (%)
linear probing
70
60
50
54.6
58.9
61.7
67.0
69.9
71.8
73.2
73.5
71.8
66.1
10
20
30
40
50
60
70
80
90
masking ratio (%)

How Useful is Self-Supervised Pretraining for Visual Tasks?

Alejandro Newell   Jia Deng
Princeton University
{anewell,jiadeng}@cs.princeton.edu

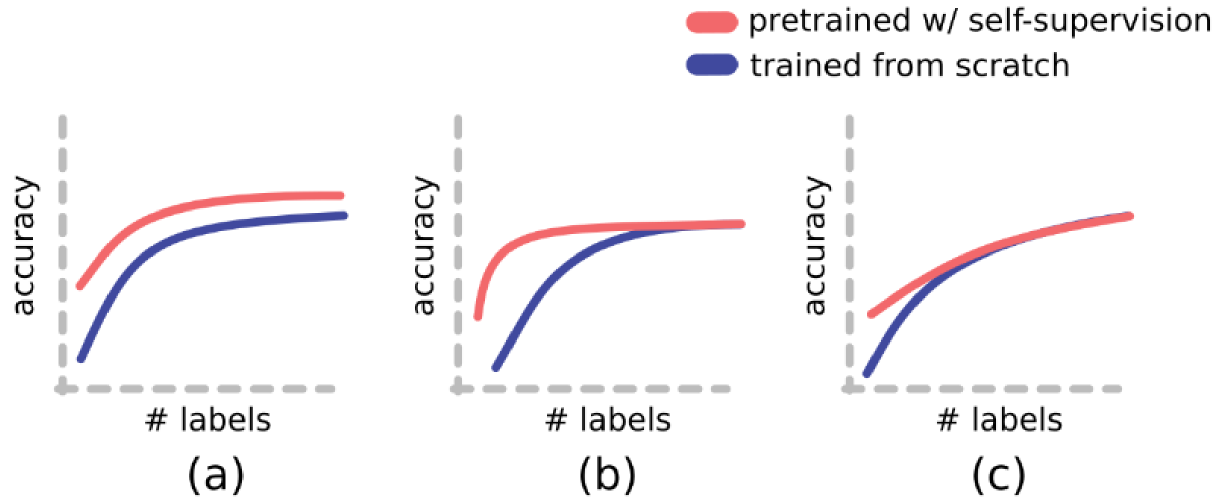


Figure 1. We highlight three possible outcomes when using self-supervised pretraining, the pretrained model either: a) always provides an improvement over the the model trained from scratch even as the amount of labeled data increases, b) reaches higher accuracy with fewer labels but plateaus to the same accuracy as the baseline, c) converges to baseline performance before accuracy plateaus. In our experiments we find option (c) to be the most common outcome.