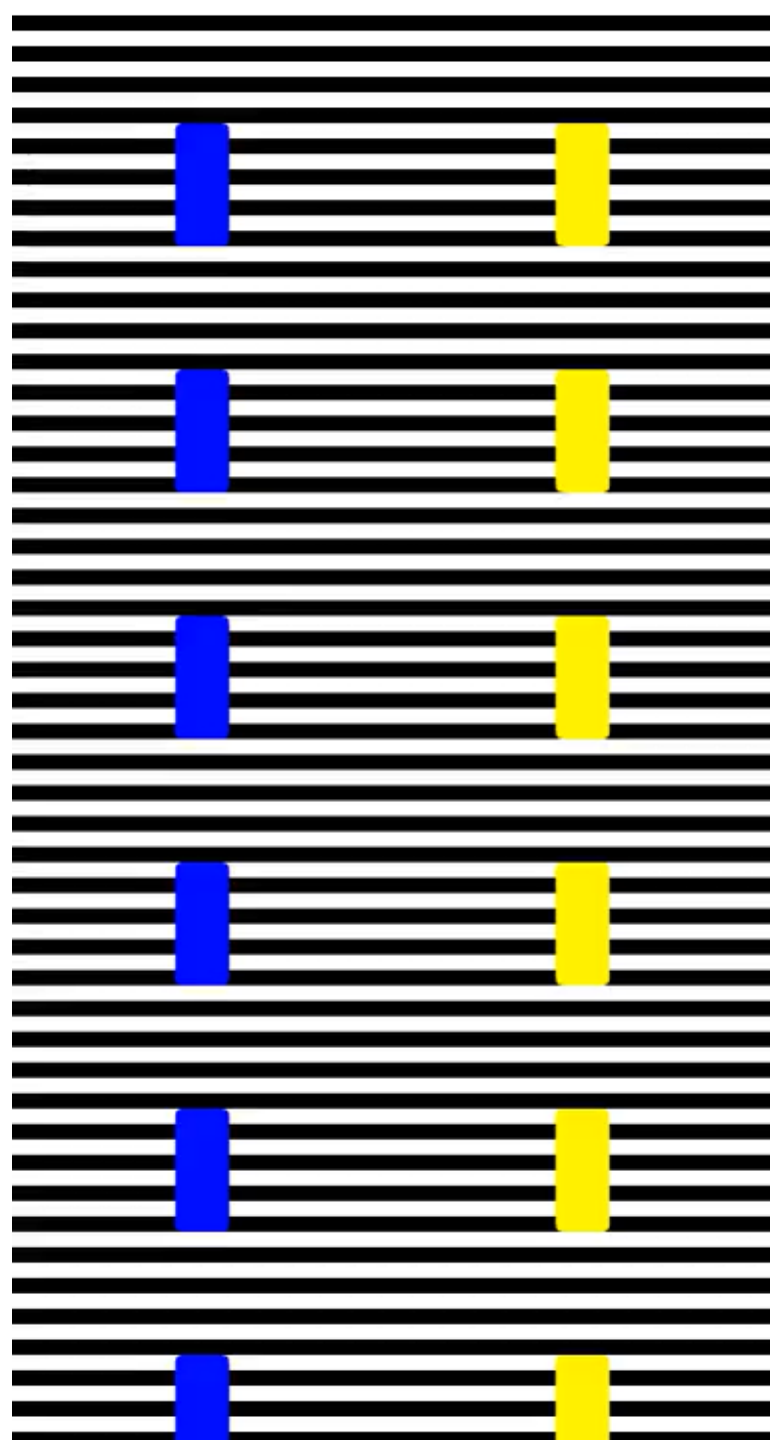
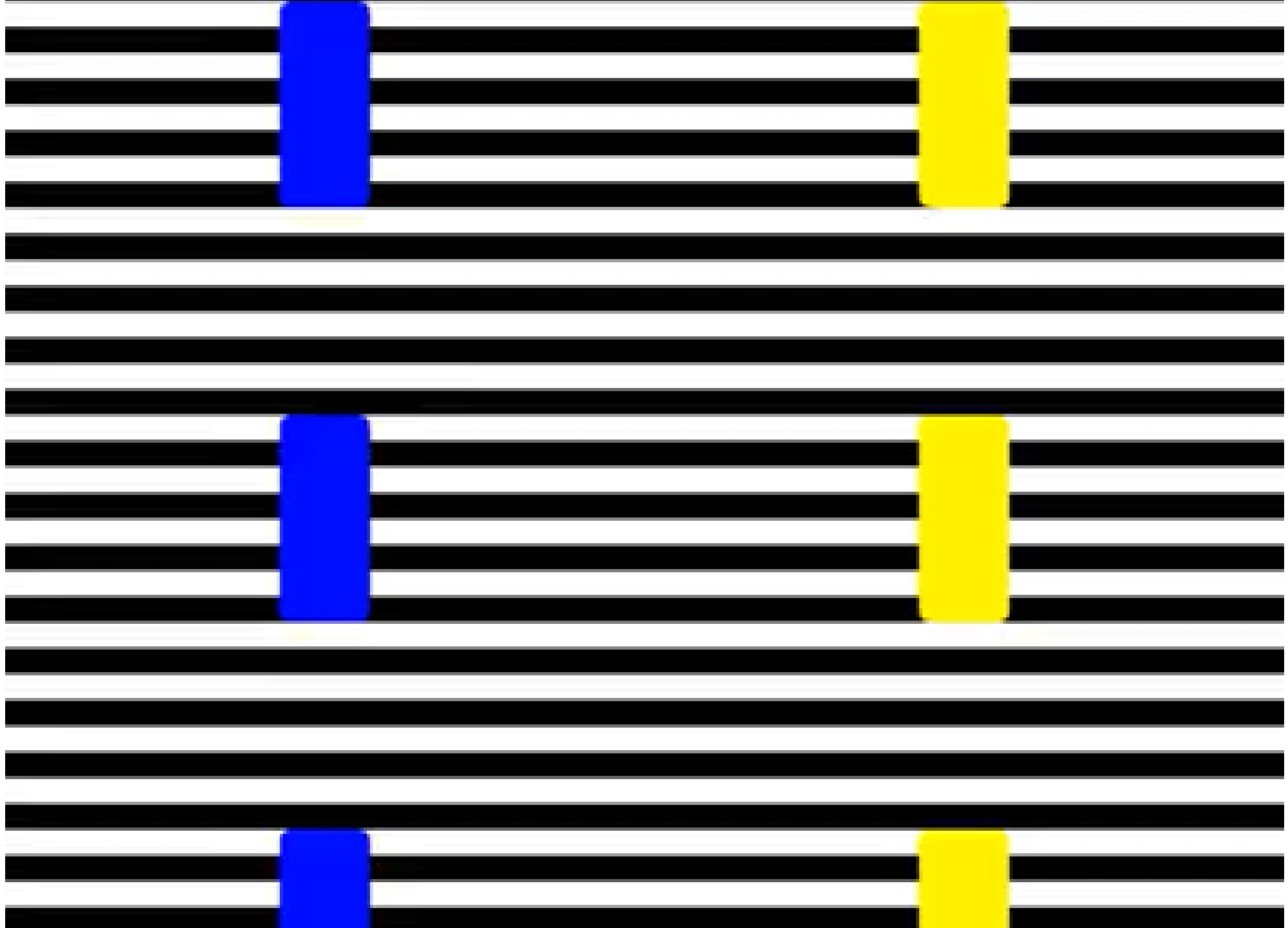






~~“Unsupervised”~~ *Self Supervised* Deep Learning

James Hays

slides from Carl Doersch and Richard Zhang

Recap

Big Data

- The Unreasonable Effectiveness of Data
- Scene Completion
- Im2gps

Crowdsourcing

- “Wisdom of the Crowds” / consensus
- Find good annotators through grading
- Pricing affects throughput but not quality
- User interface and instructions matter a lot

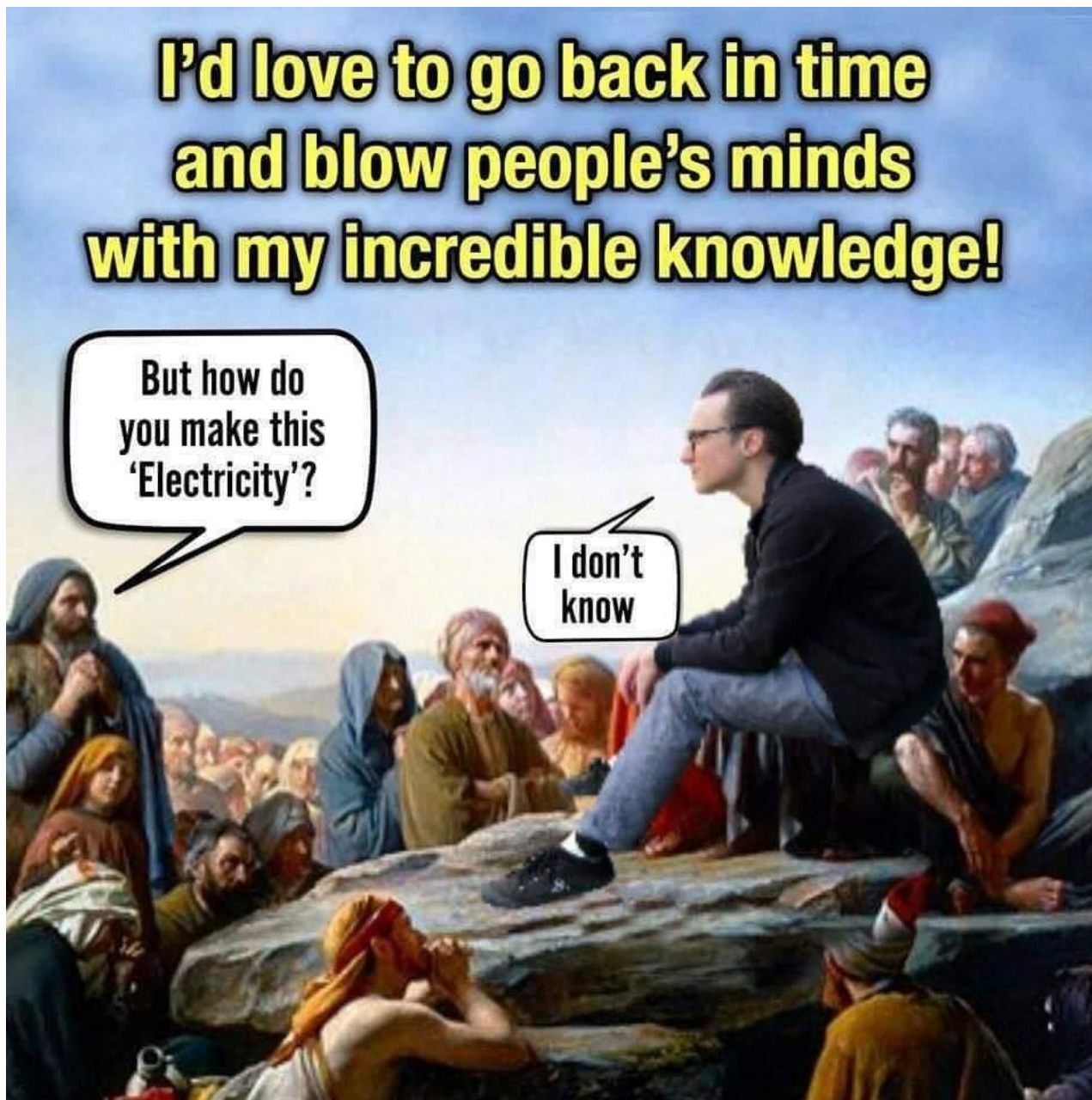
Today's Lecture

- Four 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?

**I'd love to go back in time
and blow people's minds
with my incredible knowledge!**

But how do
you make this
'Electricity'?

I don't
know



**I'd love to go back in time
and blow people's minds
with my incredible knowledge!**

But how do
you make this

Large Pretrained Model?

I don't
know



Some of Science's Greatest Discoveries

The Circumference of the Earth (Eratosthenes, ~240 BCE) Calculated the Earth's circumference with remarkable accuracy using the angles of shadows in two different cities.

Heliocentrism & Planetary Motion (Copernicus 1543 / Kepler 1609) The sun is the center of the solar system, and planets move in elliptical orbits.

The Gravitational Constant, G (Henry Cavendish, 1798) Measuring the incredibly weak gravitational pull between two lead spheres to find G (and thus the mass of the Earth).

Electromagnetism (James Clerk Maxwell, 1861–1862) Maxwell's equations, unifying electricity, magnetism, and light.

Special Relativity (Albert Einstein, 1905) The laws of physics are the same for all non-accelerating observers, and the speed of light in a vacuum is independent of the motion of all observers.

Expansion of the Universe (Edwin Hubble, 1929) Galaxies are moving away from us, meaning the universe is expanding.

Structure of DNA (Watson, Crick, and Franklin, 1953) The double-helix structure of DNA.

Age of the Earth (Clair Patterson, 1953) Pinning the age of the Earth at roughly 4.5 billion years using lead isotope dating on meteorites.

Could they have been verified earlier?

- ✓ **The Circumference of the Earth (Eratosthenes, ~240 BCE)** Calculated the Earth's circumference with remarkable accuracy using the angles of shadows in two different cities.
- ✗ **Heliocentrism & Planetary Motion (Copernicus 1543 / Kepler 1609)** The sun is the center of the solar system, and planets move in elliptical orbits.
- ✗ **The Gravitational Constant, G (Henry Cavendish, 1798)** Measuring the incredibly weak gravitational pull between two lead spheres to find G (and thus the mass of the Earth).
- ✗ **Electromagnetism (James Clerk Maxwell, 1861–1862)** Maxwell's equations, unifying electricity, magnetism, and light.
- ✓ **Special Relativity (Albert Einstein, 1905)** The laws of physics are the same for all non-accelerating observers, and the speed of light in a vacuum is independent of the motion of all observers.
- ✗ **Expansion of the Universe (Edwin Hubble, 1929)** Galaxies are moving away from us, meaning the universe is expanding.
- ✗ **Structure of DNA (Watson, Crick, and Franklin, 1953)** The double-helix structure of DNA.
- ✗ **Age of the Earth (Clair Patterson, 1953)** Pinning the age of the Earth at roughly 4.5 billion years using lead isotope dating on meteorites.

The Gelato Bet

"If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, Mr. Malik promises to buy Mr. Efros one (1) gelato"



R-CNN: *Regions with CNN features*

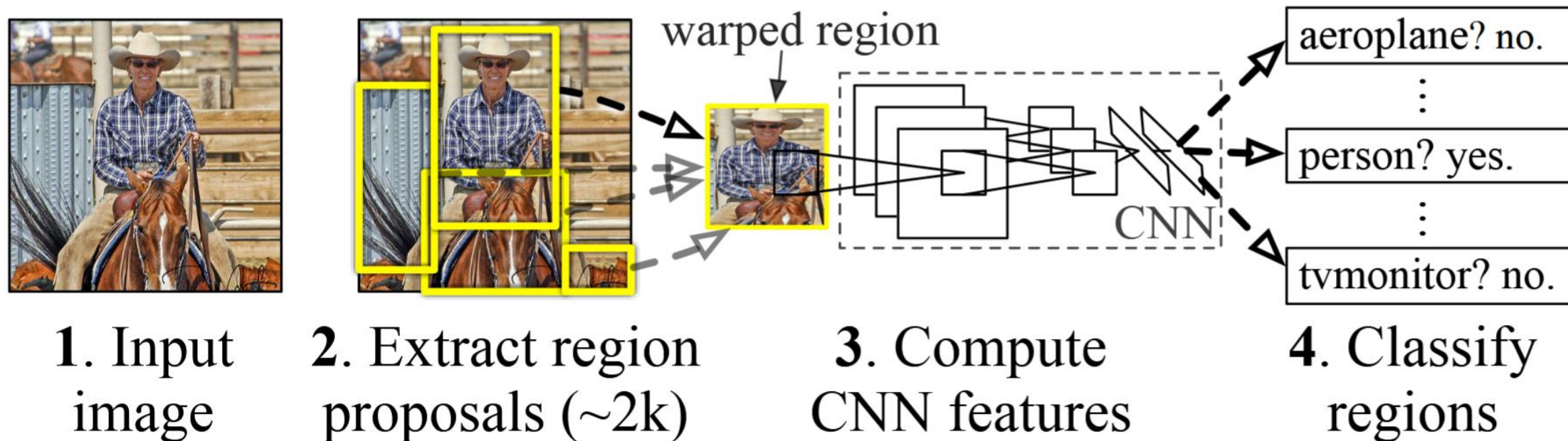


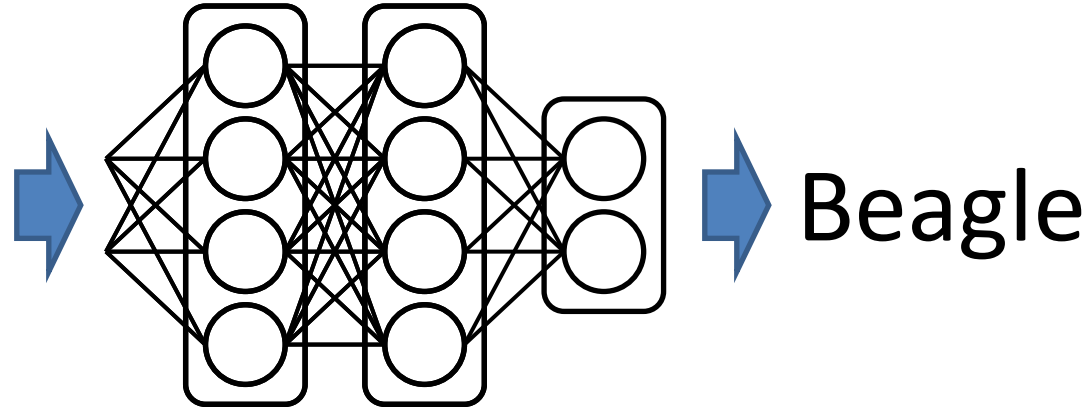
Figure 1: Object detection system overview. Our system (1) takes an input image, (2) extracts around 2000 bottom-up region proposals, (3) computes features for each proposal using a large convolutional neural network (CNN), and then (4) classifies each region using class-specific linear SVMs. R-CNN achieves a mean average precision (mAP) of **53.7% on PASCAL VOC 2010**. For

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Alexei A. Efros, and Abhinav Gupta

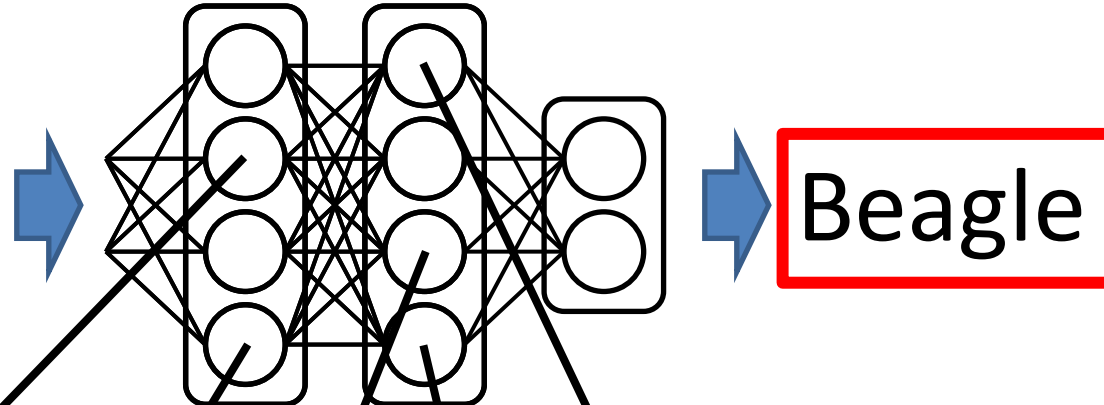
ICCV 2015

ImageNet + Deep Learning



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

ImageNet + Deep Learning



Materials?

Parts?

Pose?

Do we ever need this sort of labels?

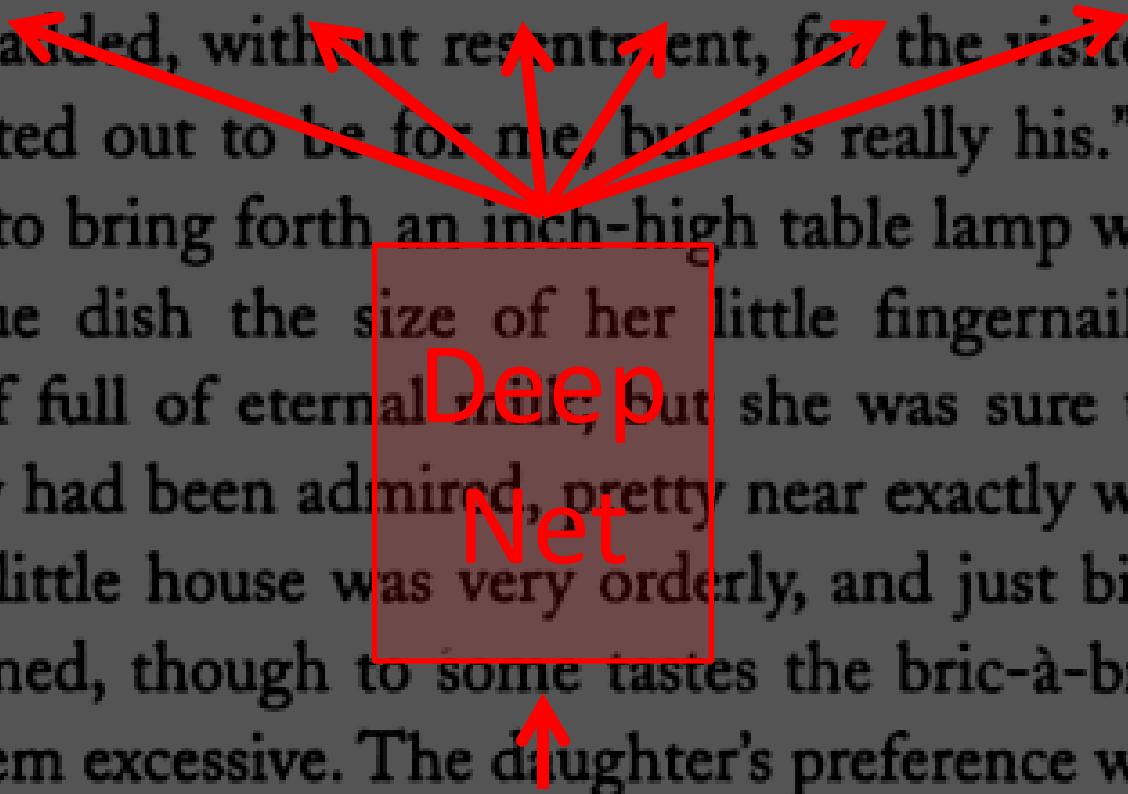
Geometry?

Boundaries?

Context as Supervision

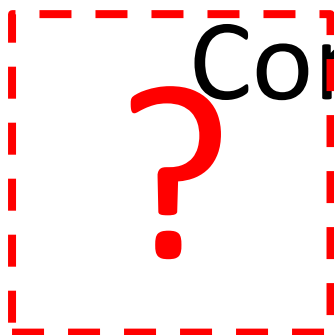
[Collobert & Weston 2008; Mikolov et al. 2013 (Word2Vec)]

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 visitors would



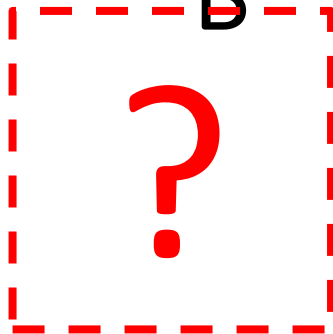
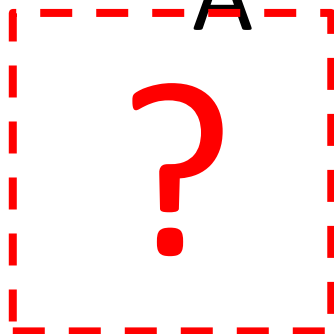
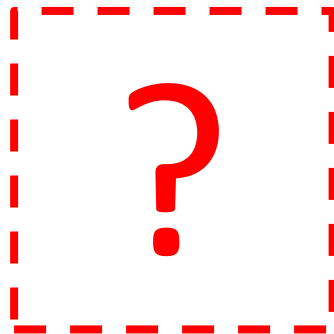
Deep
Net

Context Prediction for Images

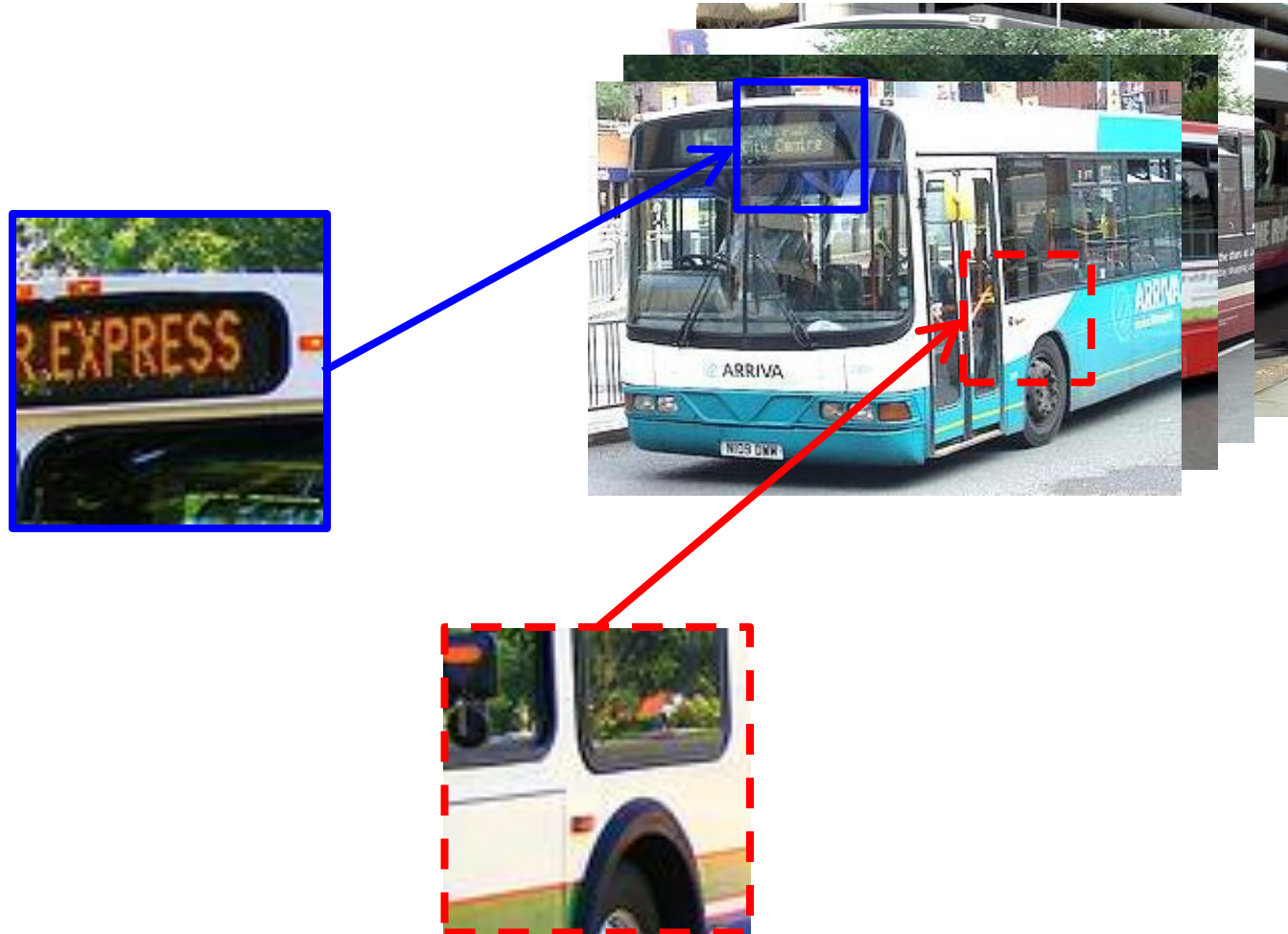


A

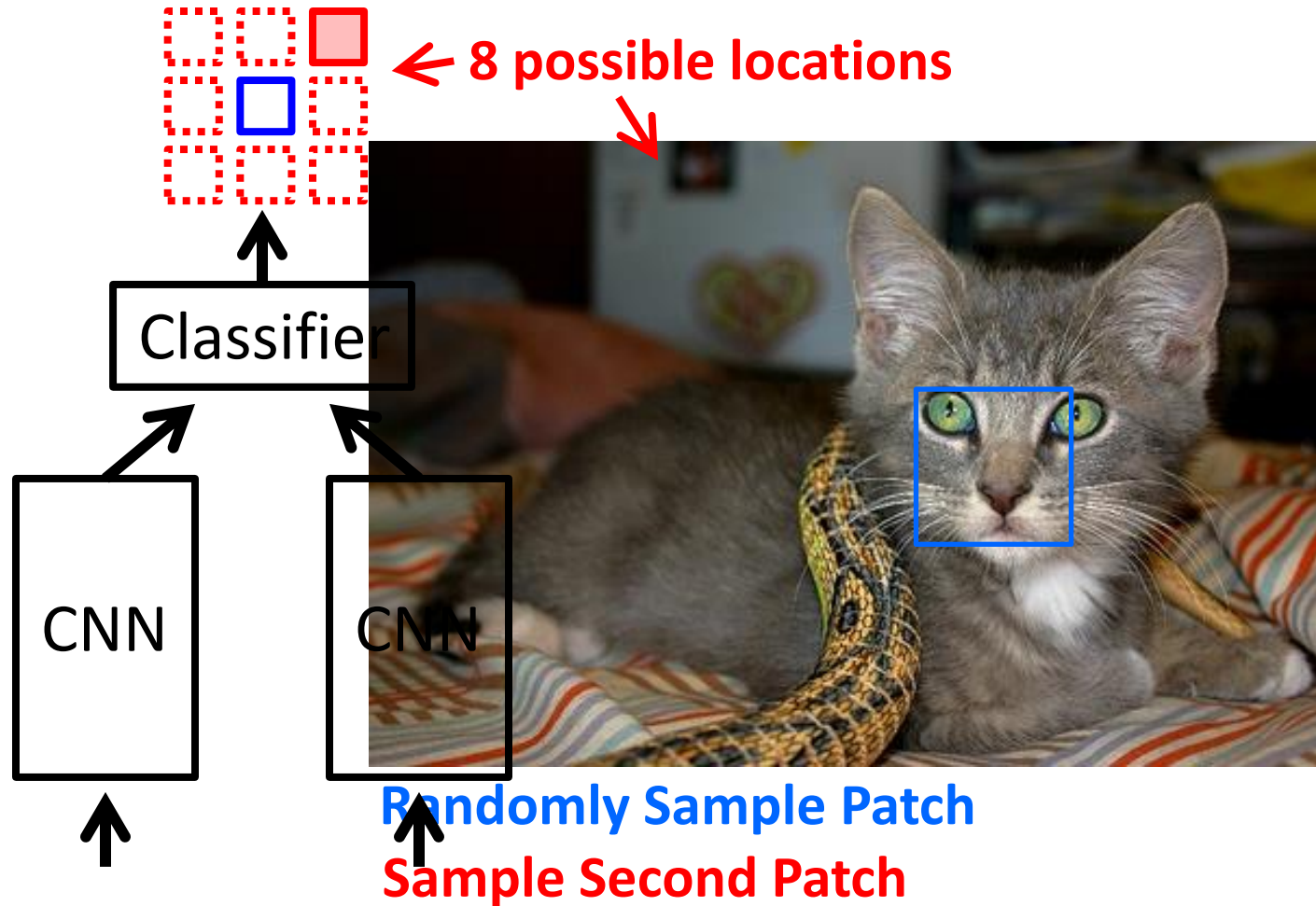
B

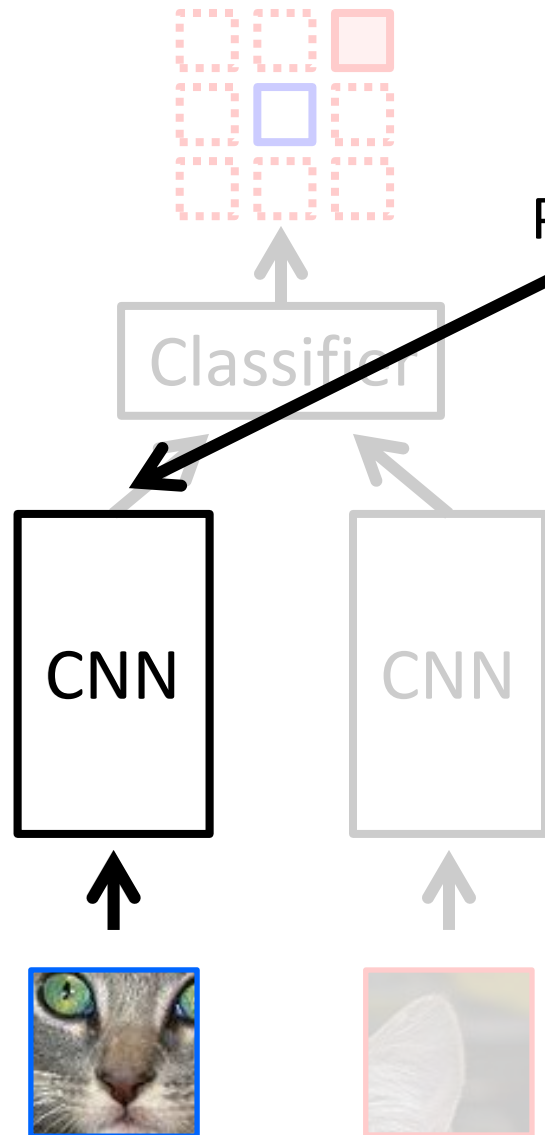


Semantics from a non-semantic task



Relative Position Task





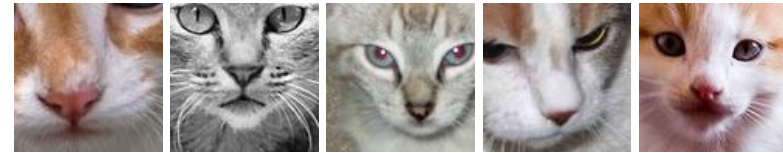
Patch Embedding

Input



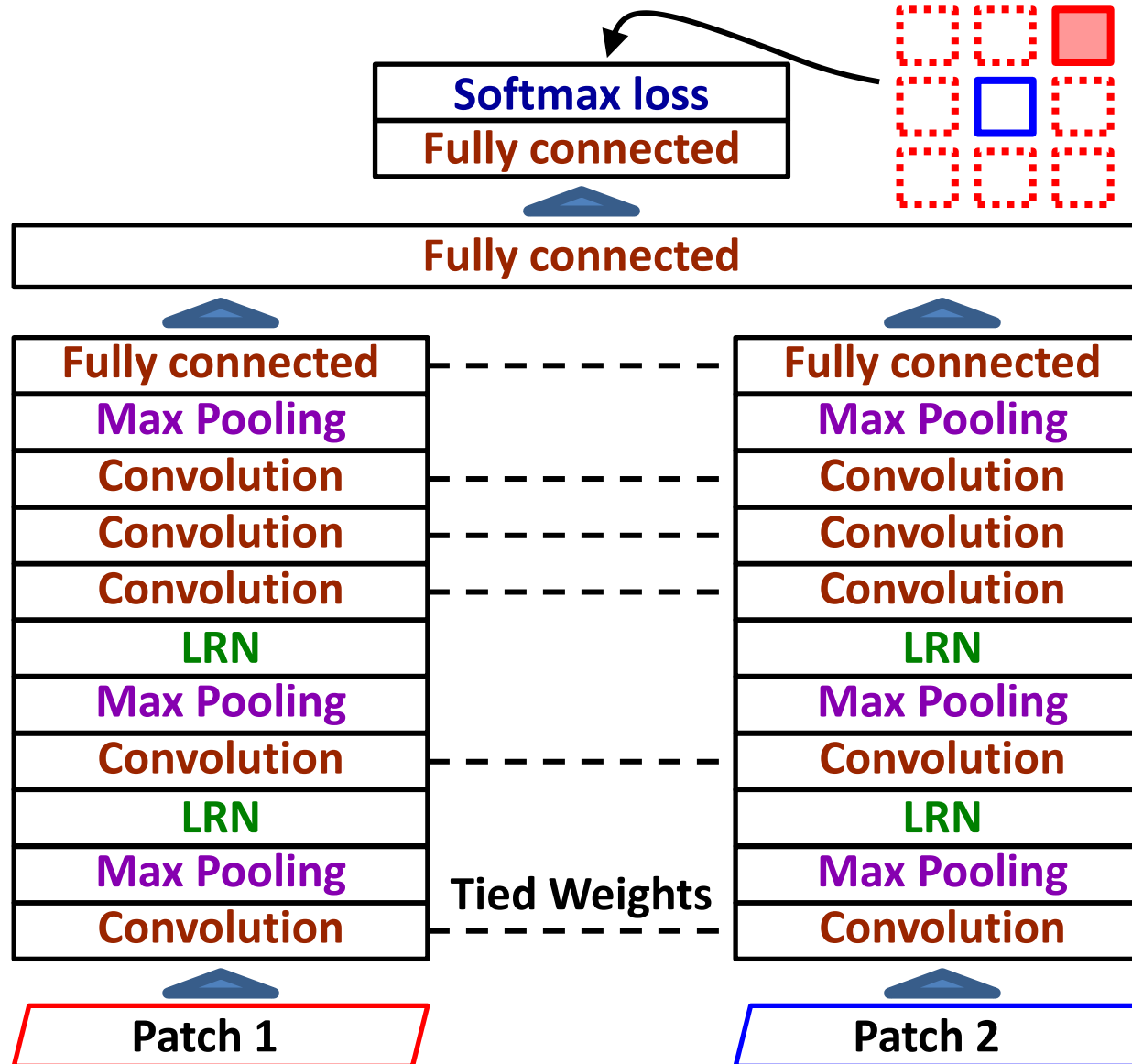
!

Nearest Neighbors

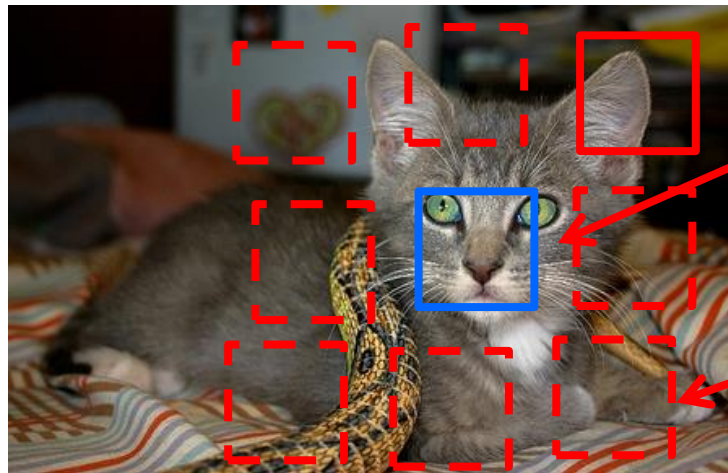
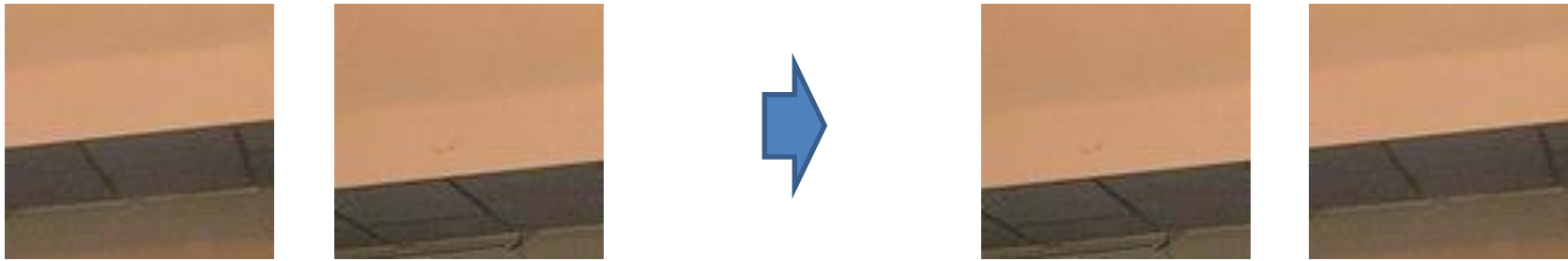


Note: connects ***across*** instances!

Architecture



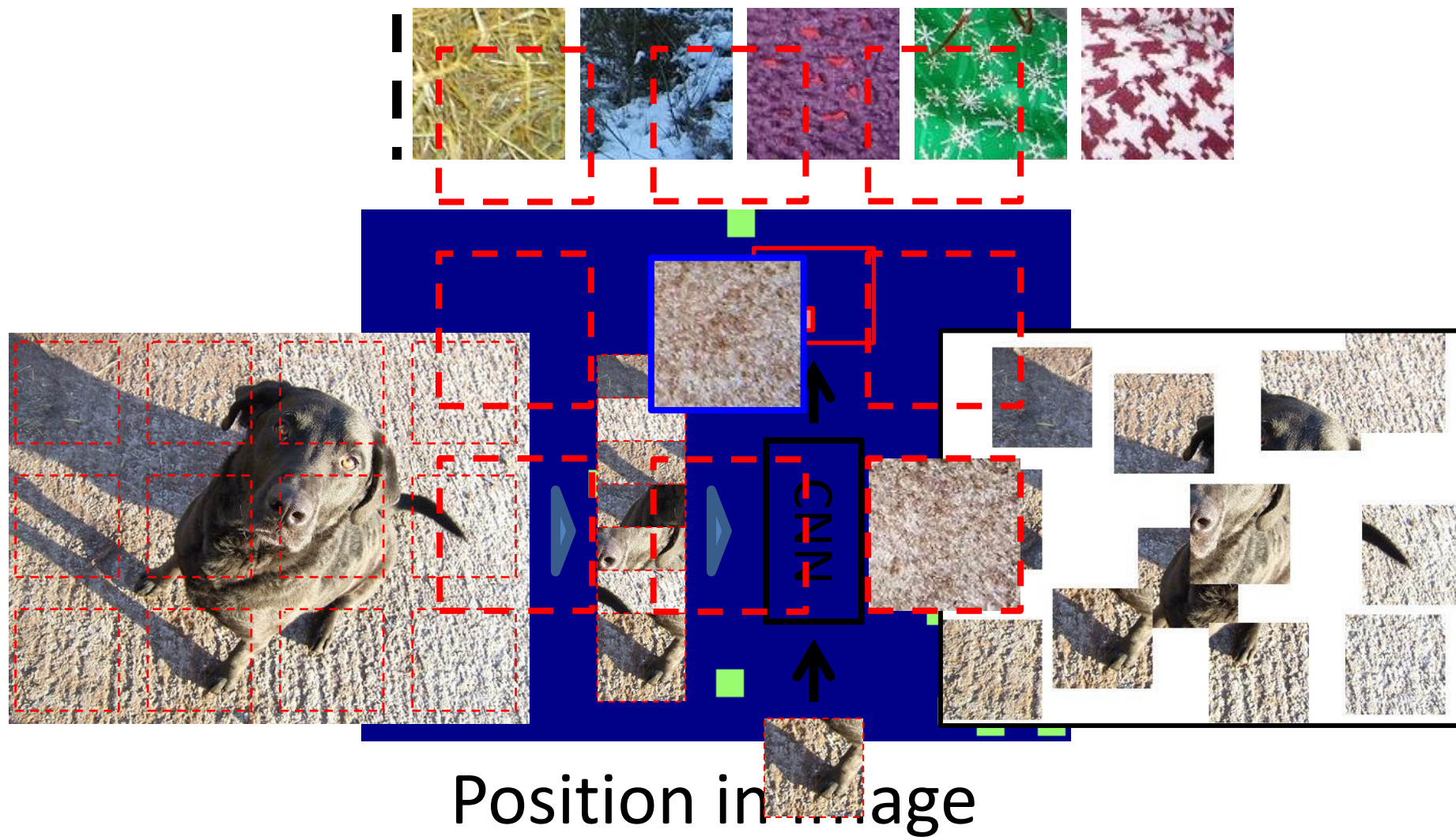
Avoiding Trivial Shortcuts



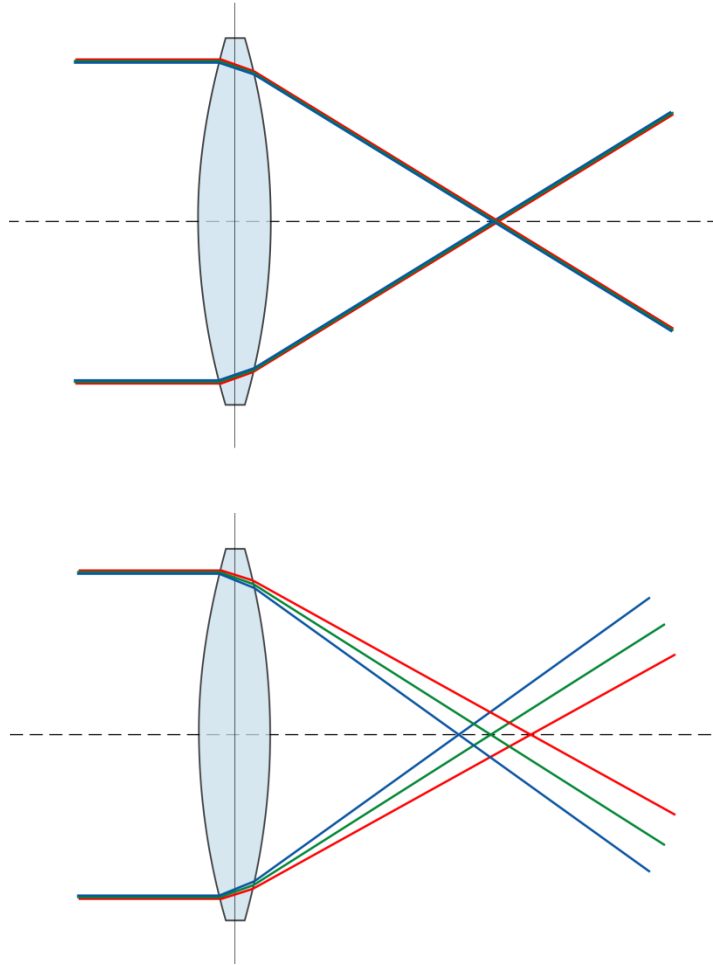
Include a gap

Jitter the patch locations

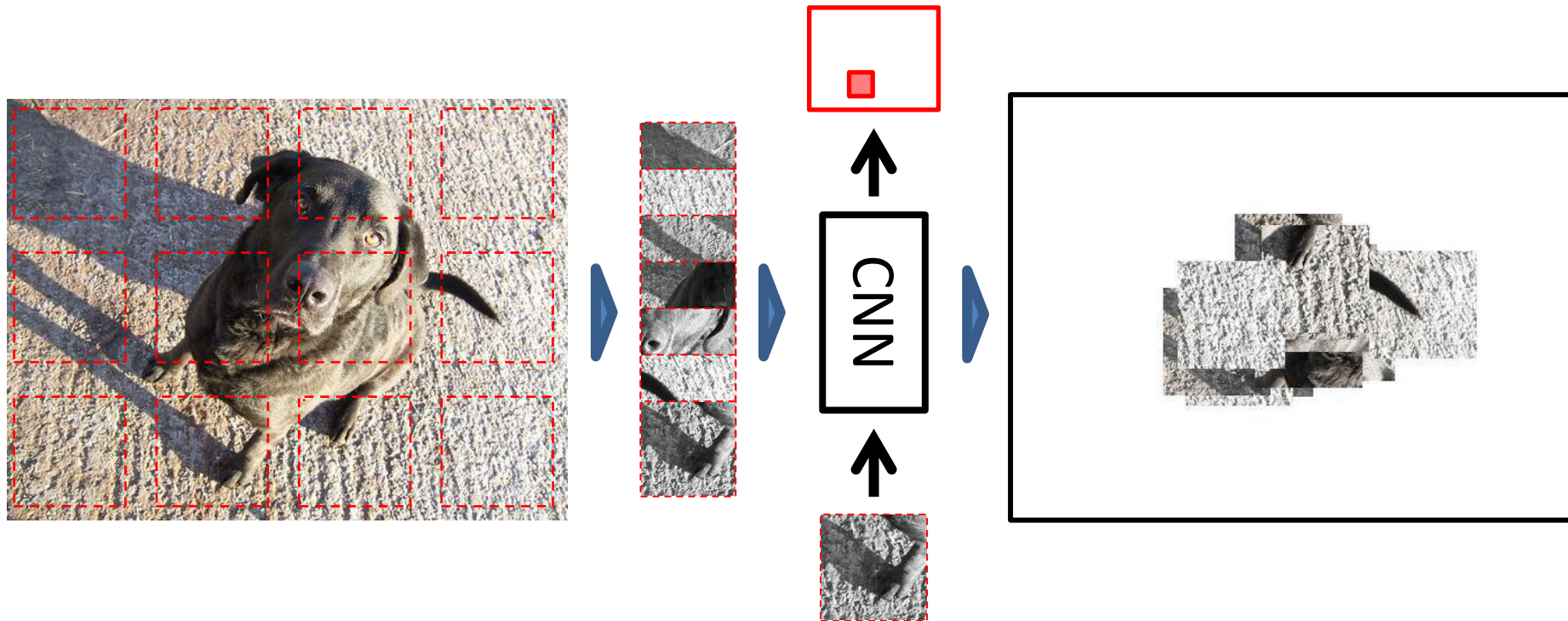
A Not-So “Trivial” Shortcut



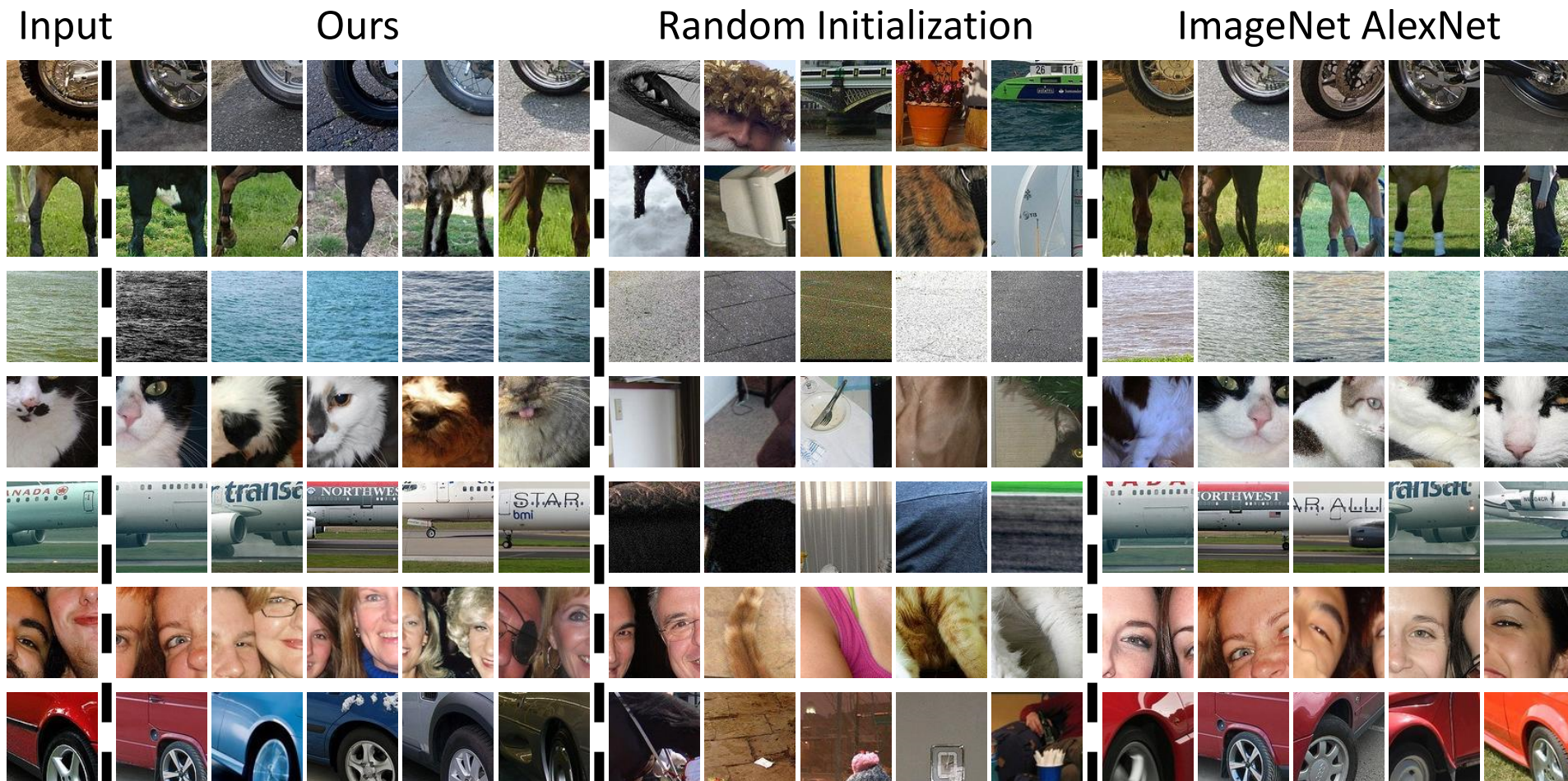
Chromatic Aberration



Chromatic Aberration



What is learned?



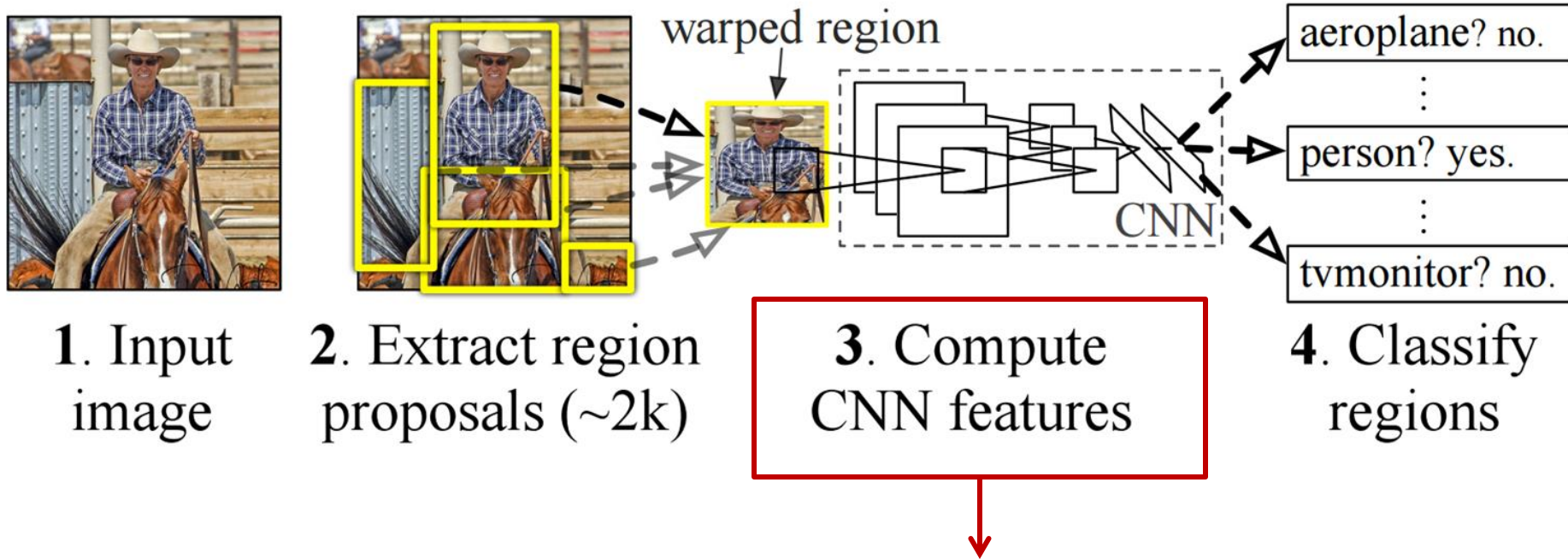
Still don't capture everything



You don't always need to learn!



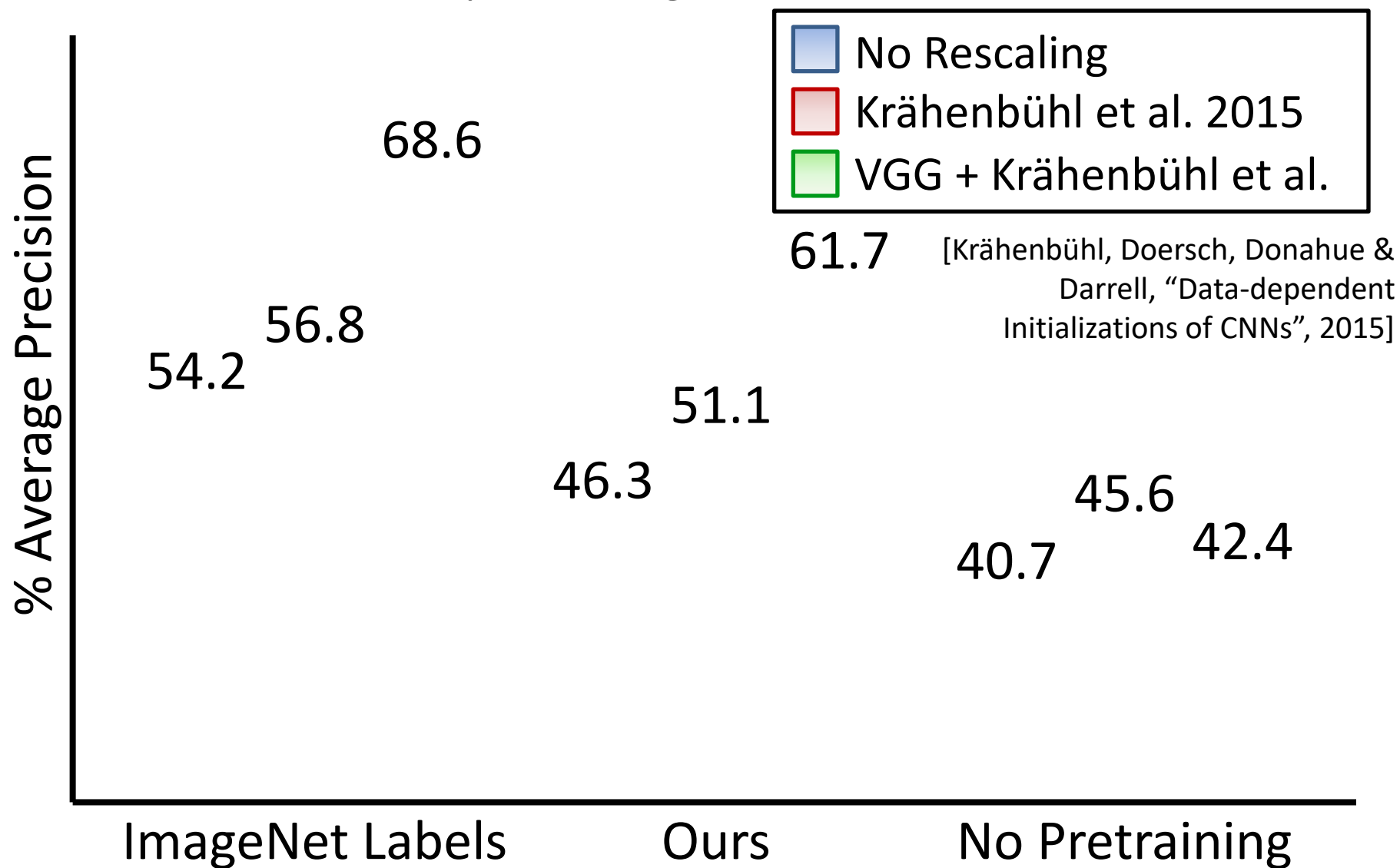
Pre-Training for R-CNN



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

VOC 2007 Performance

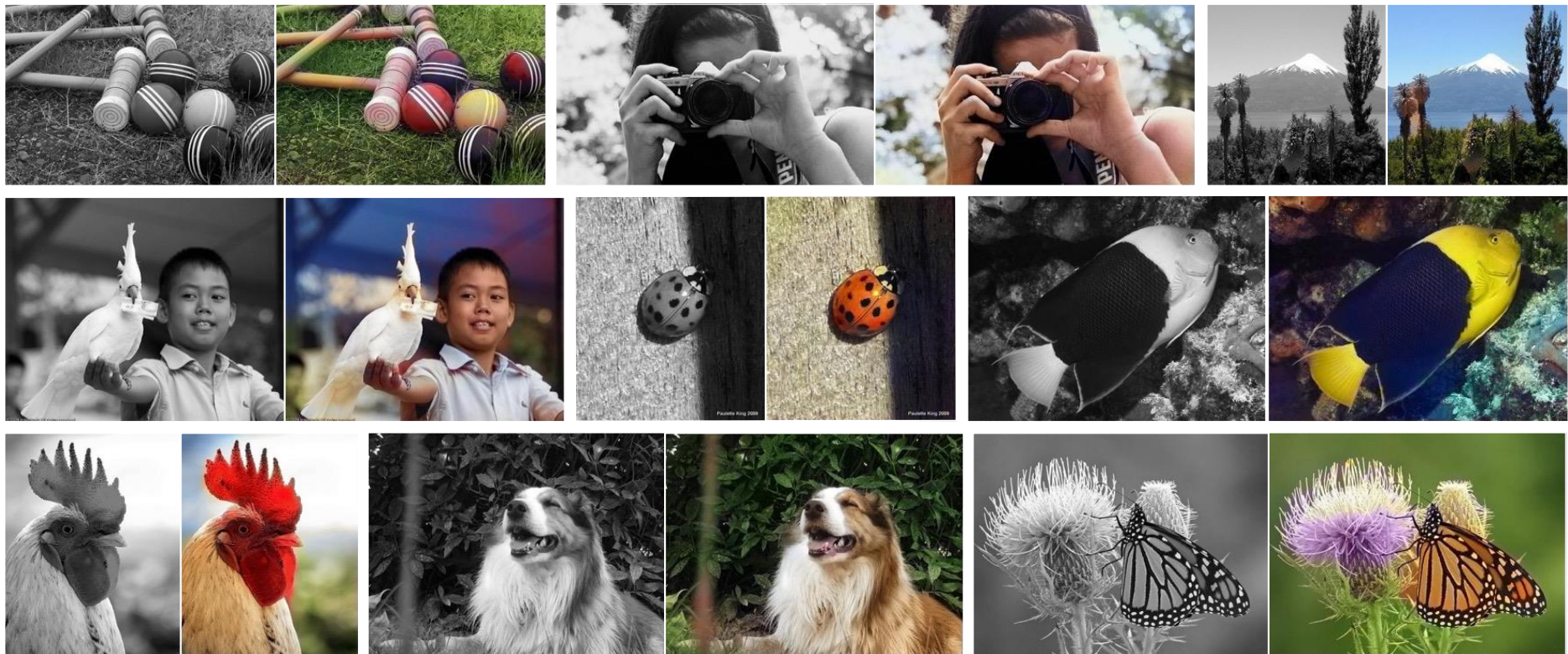
(pretraining for R-CNN)



Clever ideas to keep in mind

- “Pretext” tasks to train networks
- Avoiding shortcuts in learning by
 - Dropping out modalities
 - Randomizing patch offsets
 - Including gaps in patch offsets

So, do we need semantic labels?

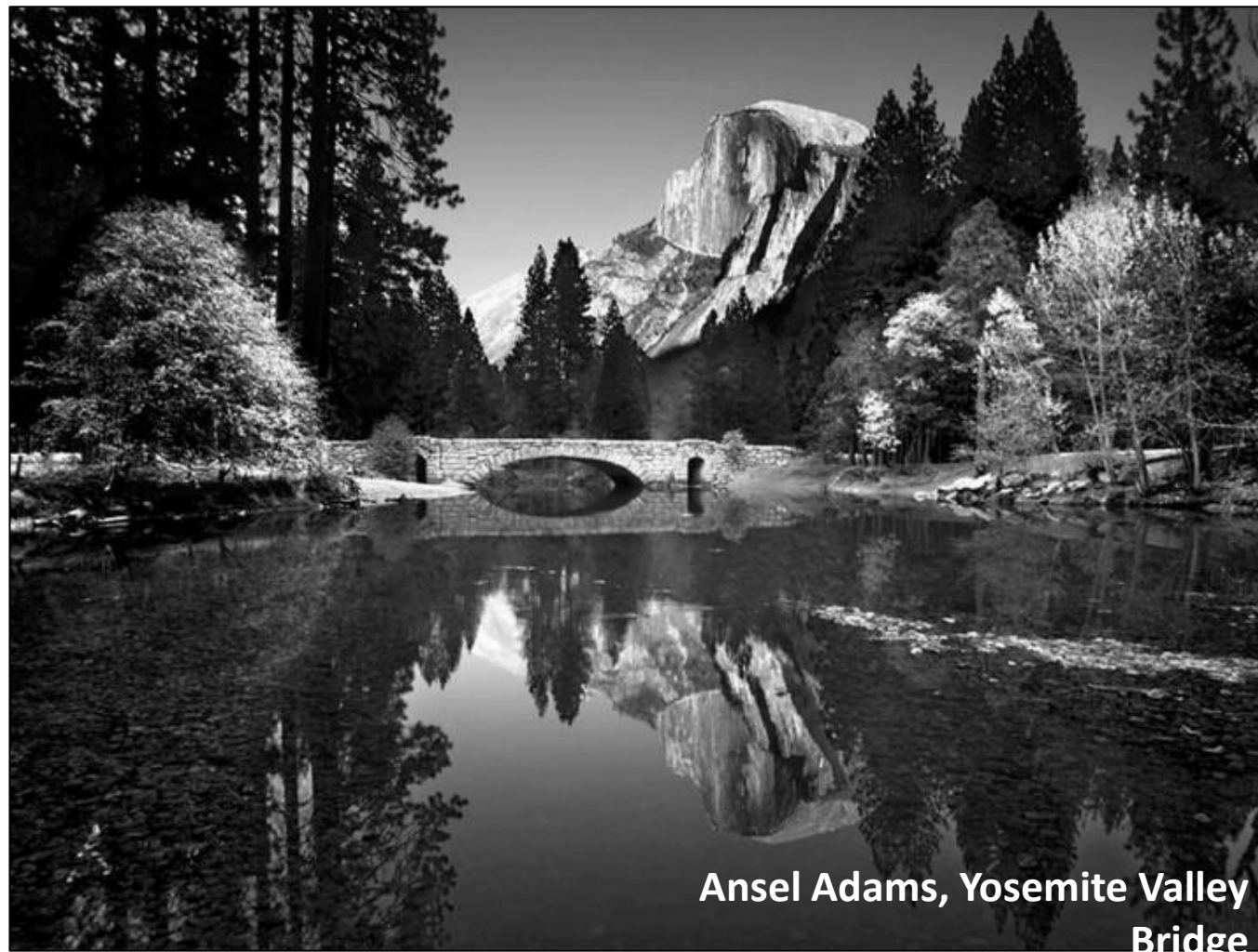


Colorful Image Colorization

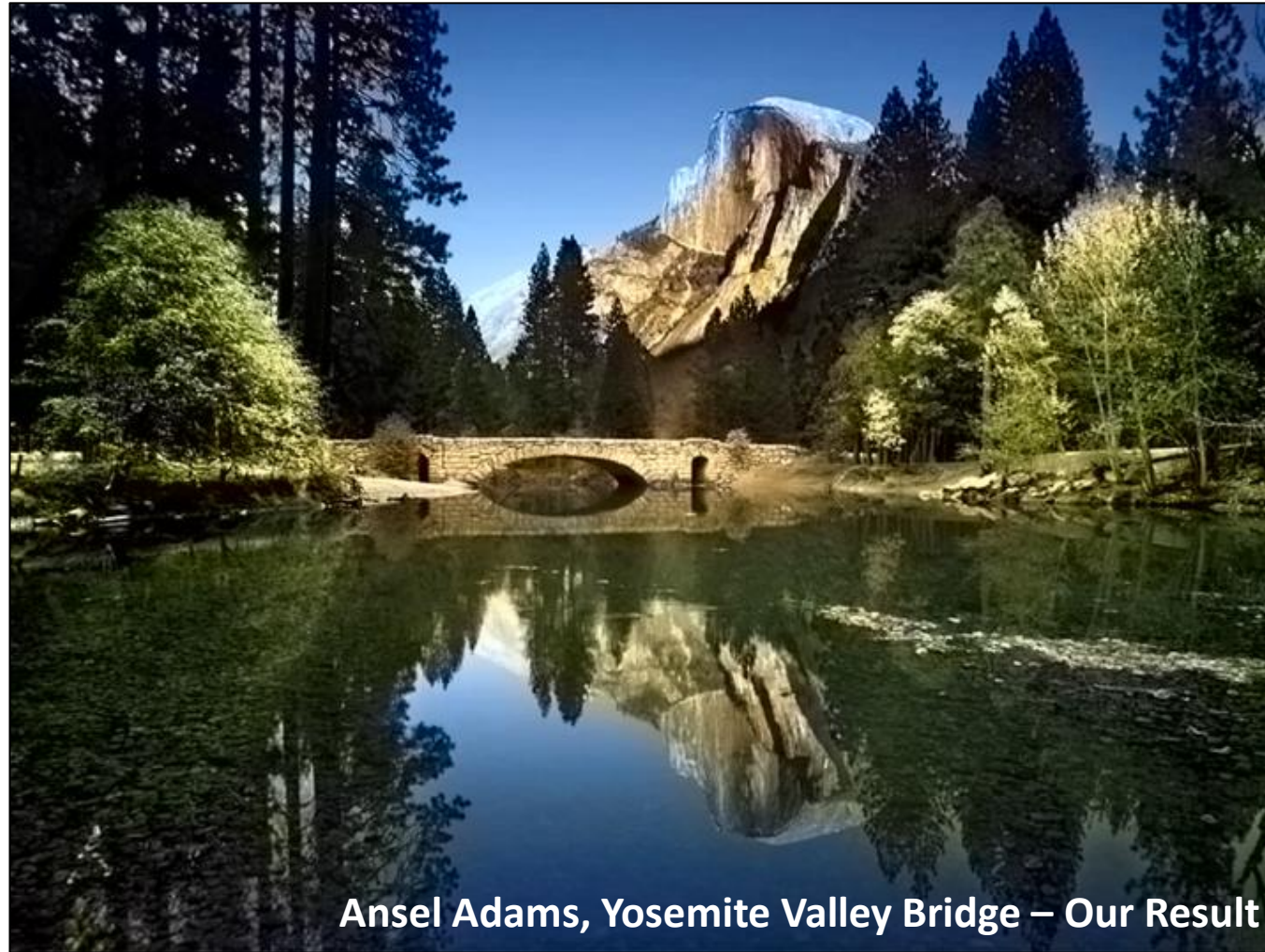
Richard Zhang, Phillip Isola, Alexei (Alyosha) Efros

richzhang.github.io/colorization

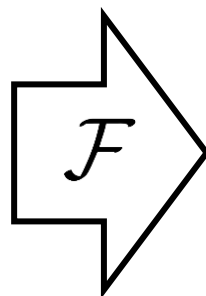
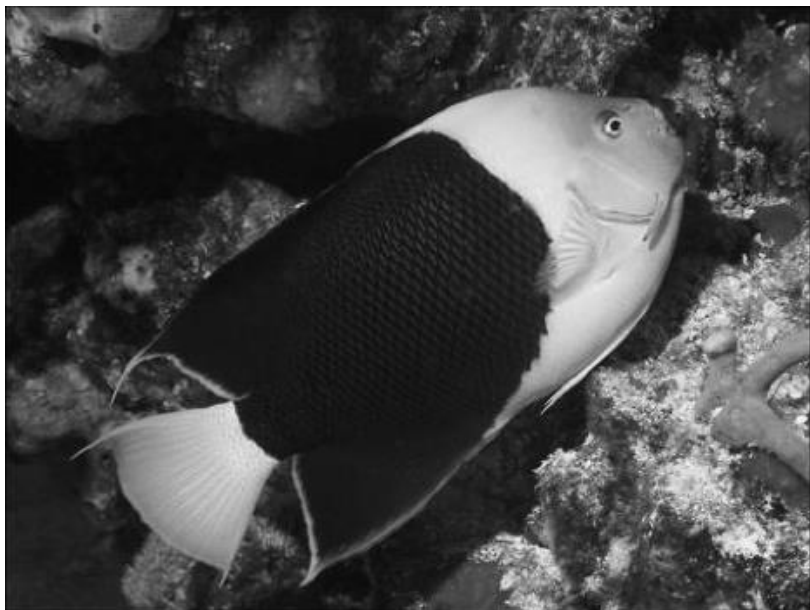
ECCV 2016



Ansel Adams, Yosemite Valley
Bridge



Ansel Adams, Yosemite Valley Bridge – Our Result

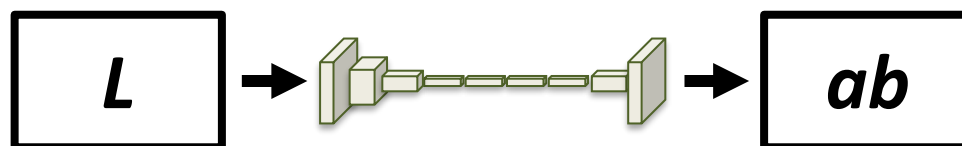


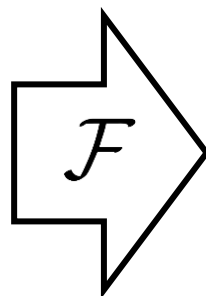
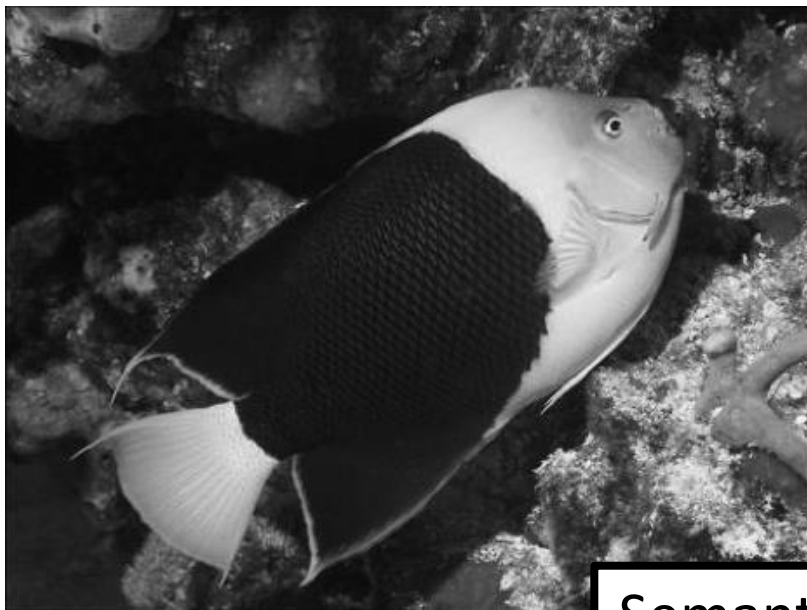
Grayscale image: L channel

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

Color information: ab channels

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





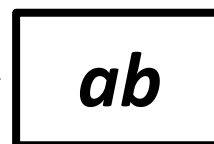
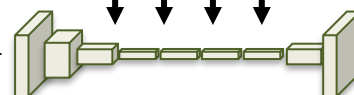
Grayscale image: L ch

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

Semantics? Higher-level abstraction?

Concatenate (L, ab)

$$(\mathbf{X}, \hat{\mathbf{Y}})$$



"Free"
supervisory
signal

Inherent Ambiguity



Grayscale

Inherent Ambiguity



Our Output



Ground Truth

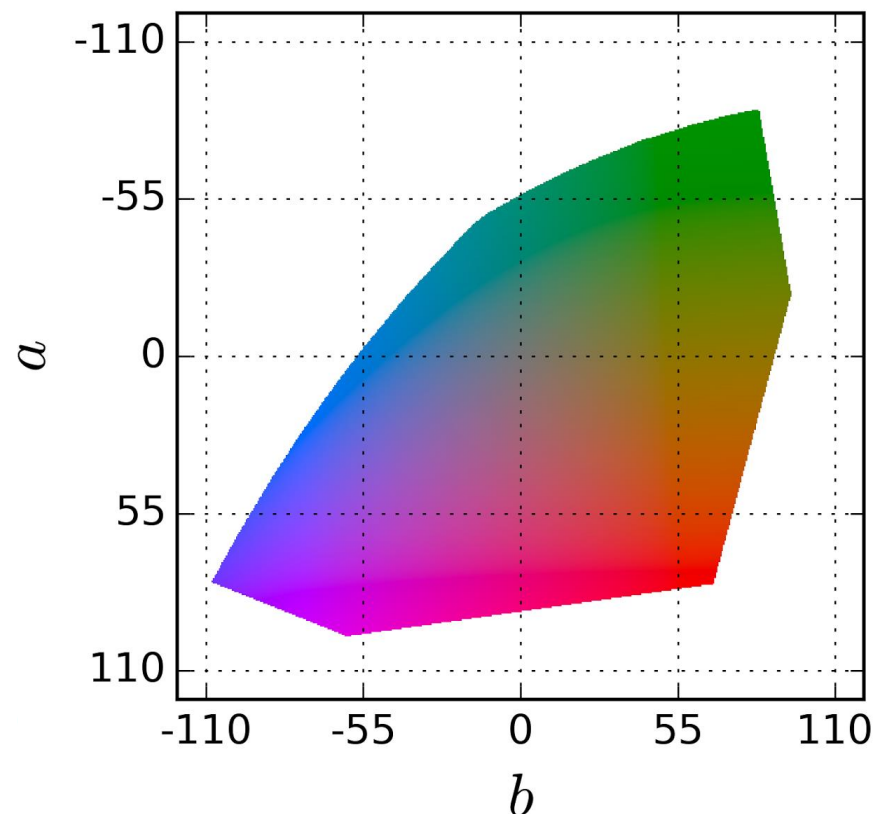
Better Loss Function

- Regression with L2 loss inadequate

$$L_2(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w}\|_2^2$$

Colors in *ab* space

(continuous)



Better Loss Function

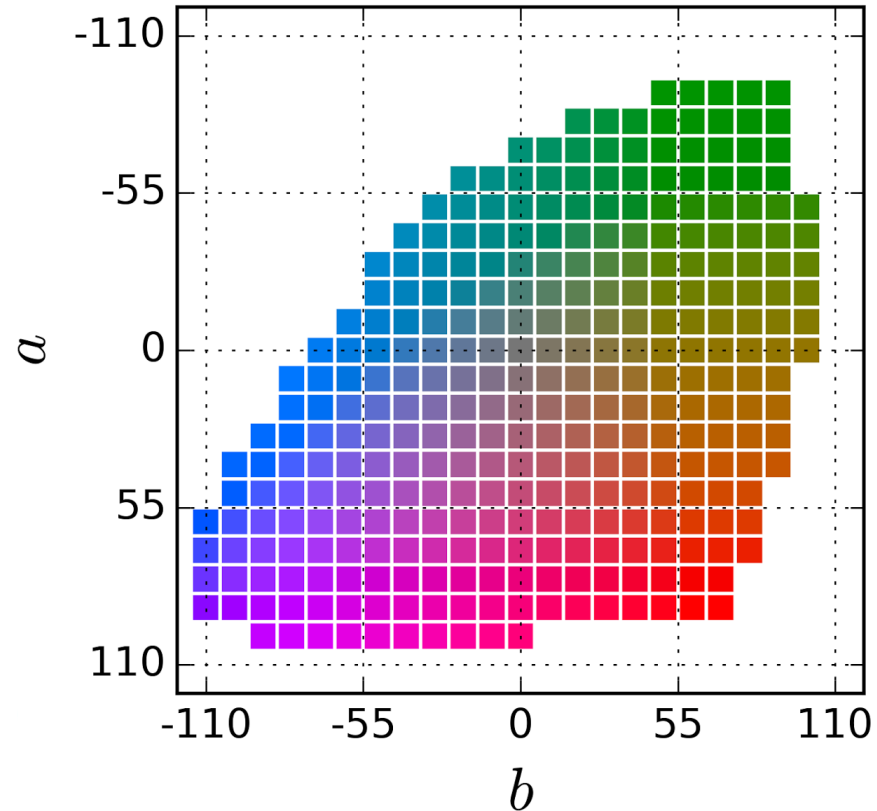
- Regression with L2 loss inadequate

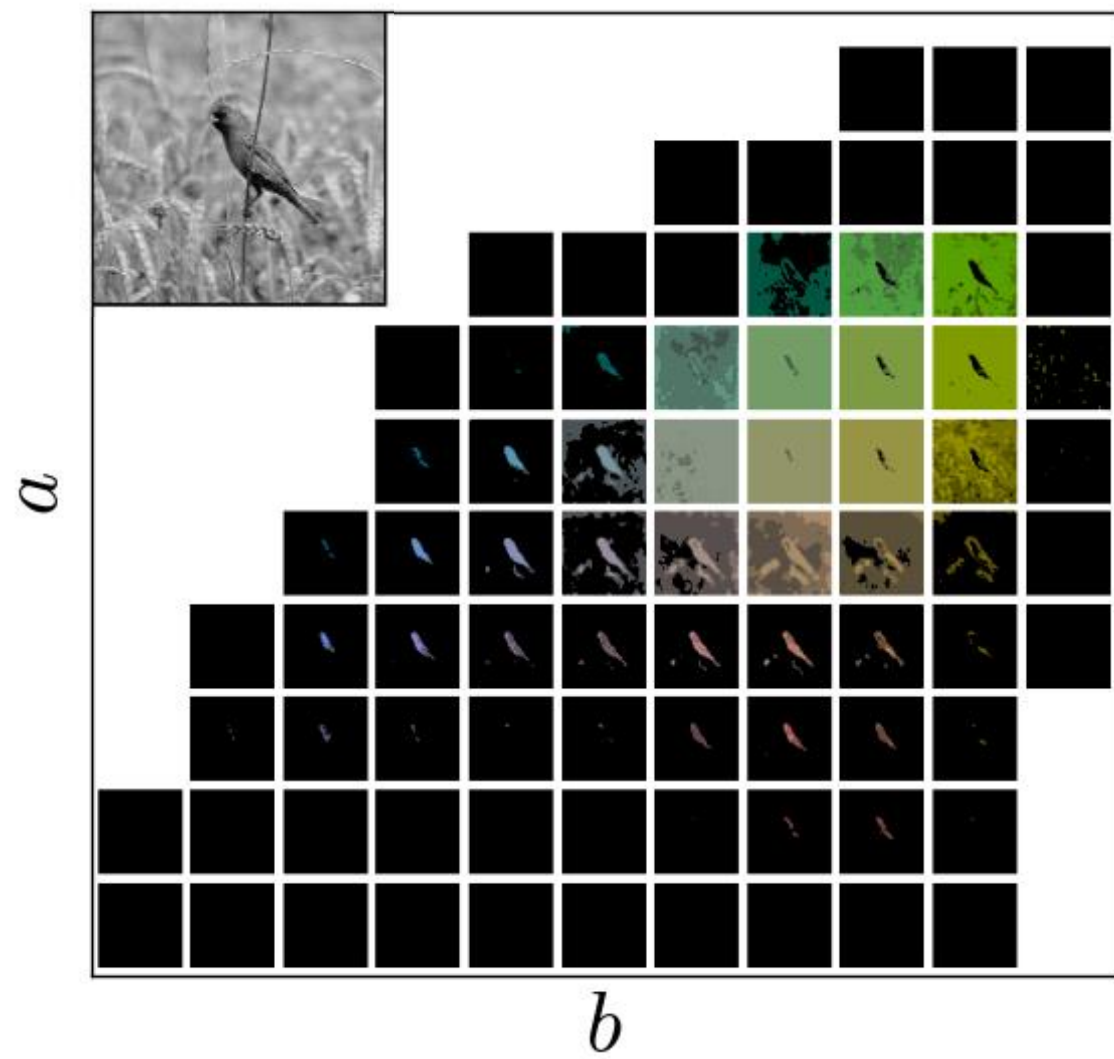
$$L_2(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w}\|_2^2$$

- Use **multinomial classification**

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

Colors in *ab* space
(discrete)





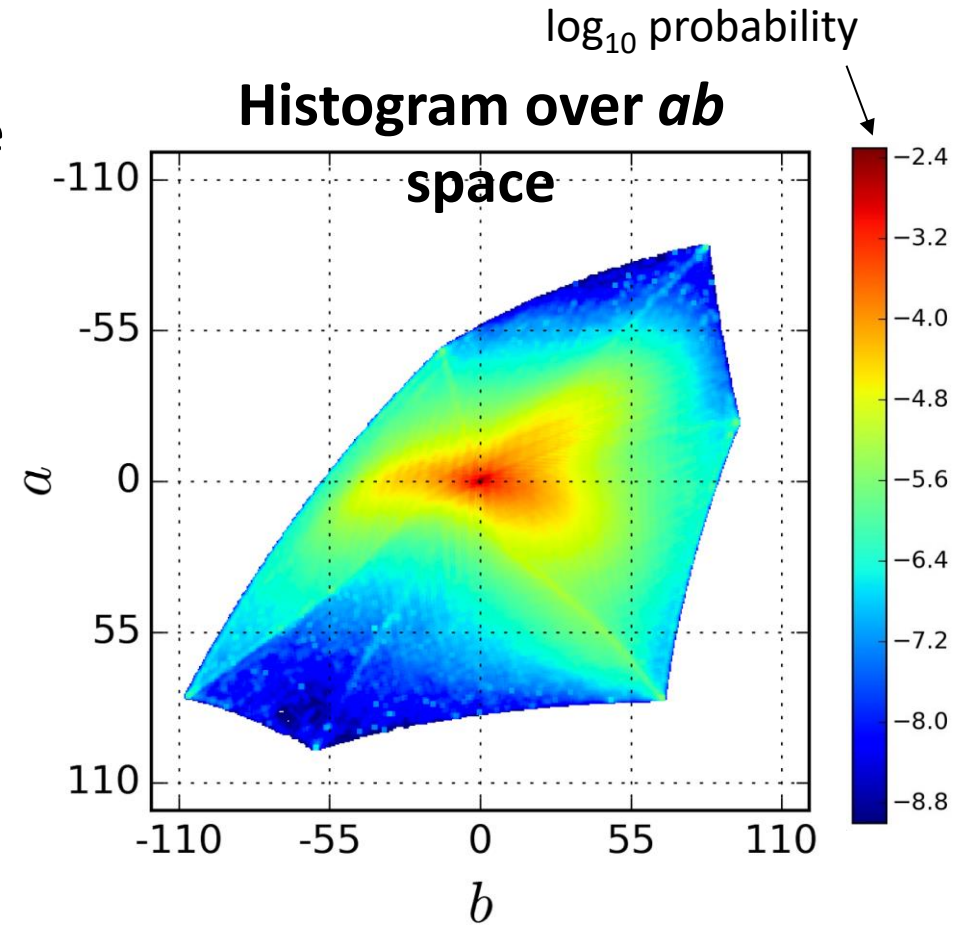
Better Loss Function

- Regression with L2 loss inadequate

$$L_2(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w}\|_2^2$$

- Use **multinomial classification**

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



Better Loss Function

- Regression with L2 loss inadequate

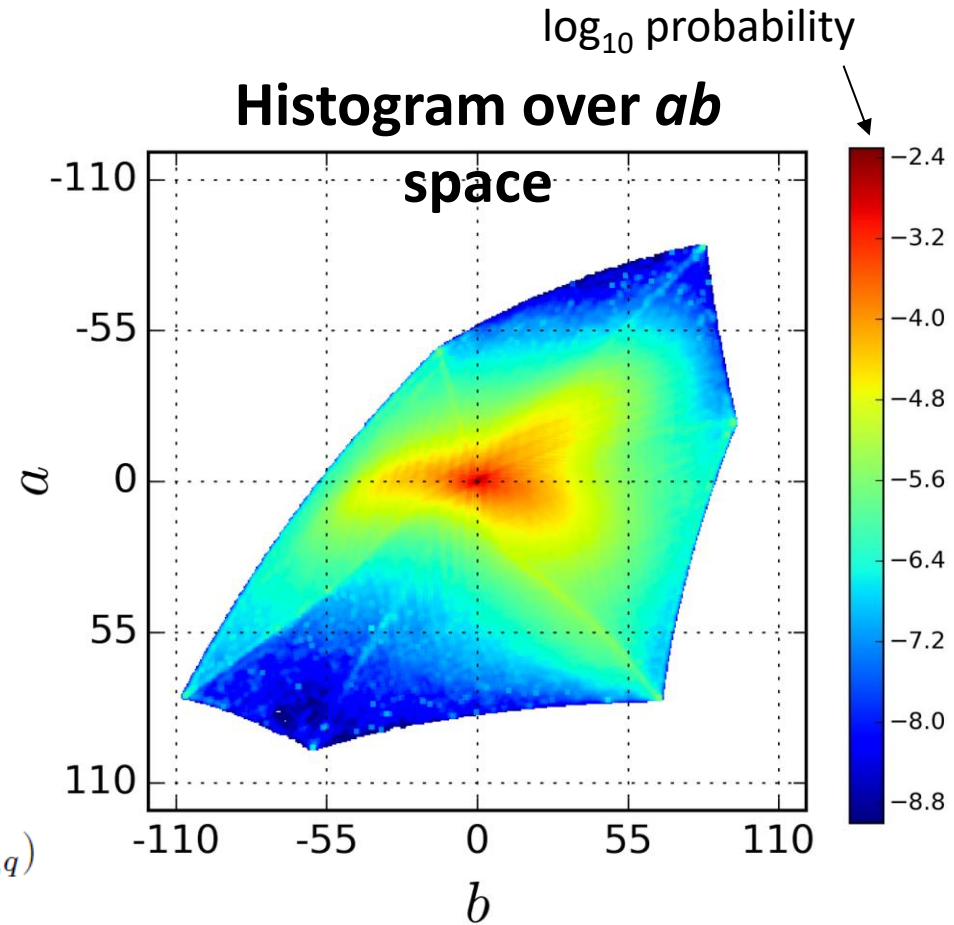
$$L_2(\hat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \hat{\mathbf{Y}}_{h,w}\|_2^2$$

- Use **multinomial classification**

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

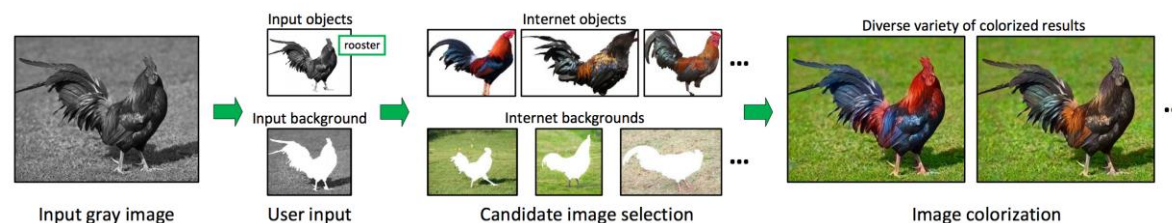
- Class rebalancing** to encourage learning of *rare* colors

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



Non-parametric

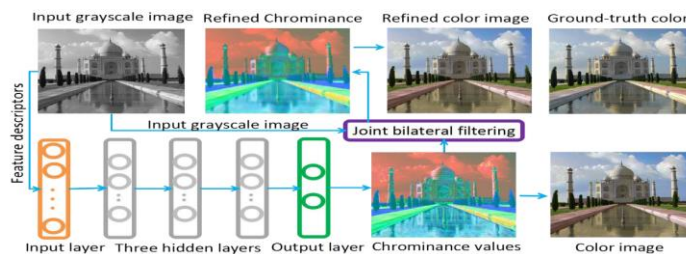
Hertzmann et al. In SIGGRAPH, 2001.
 Welsh et al. In TOG, 2002.
 Irony et al. In Eurographics, 2005.
 Liu et al. In TOG, 2008.
 Chia et al. In ACM 2011.
 Gupta et al. In ACM, 2012.



Parametric

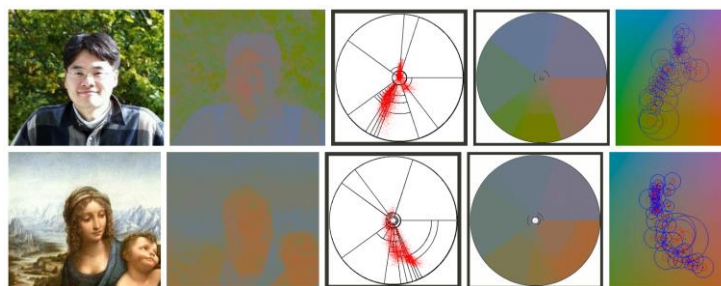
L2 Regression

Hand-engineered Features



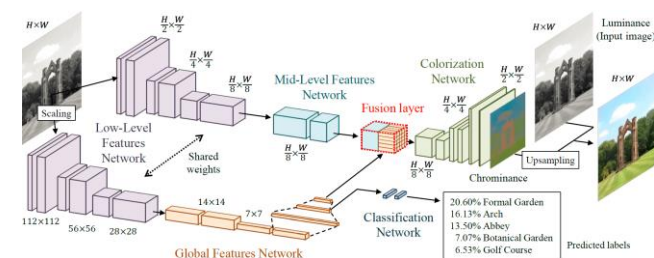
Deshpande et al. Cheng et al. In ICCV 2015.

Classification

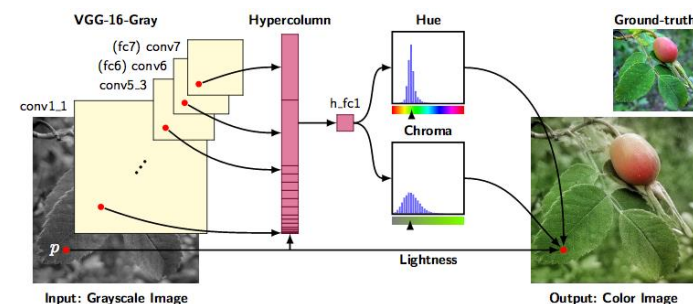


Charpiat et al. In ECCV 2008.

Deep Networks

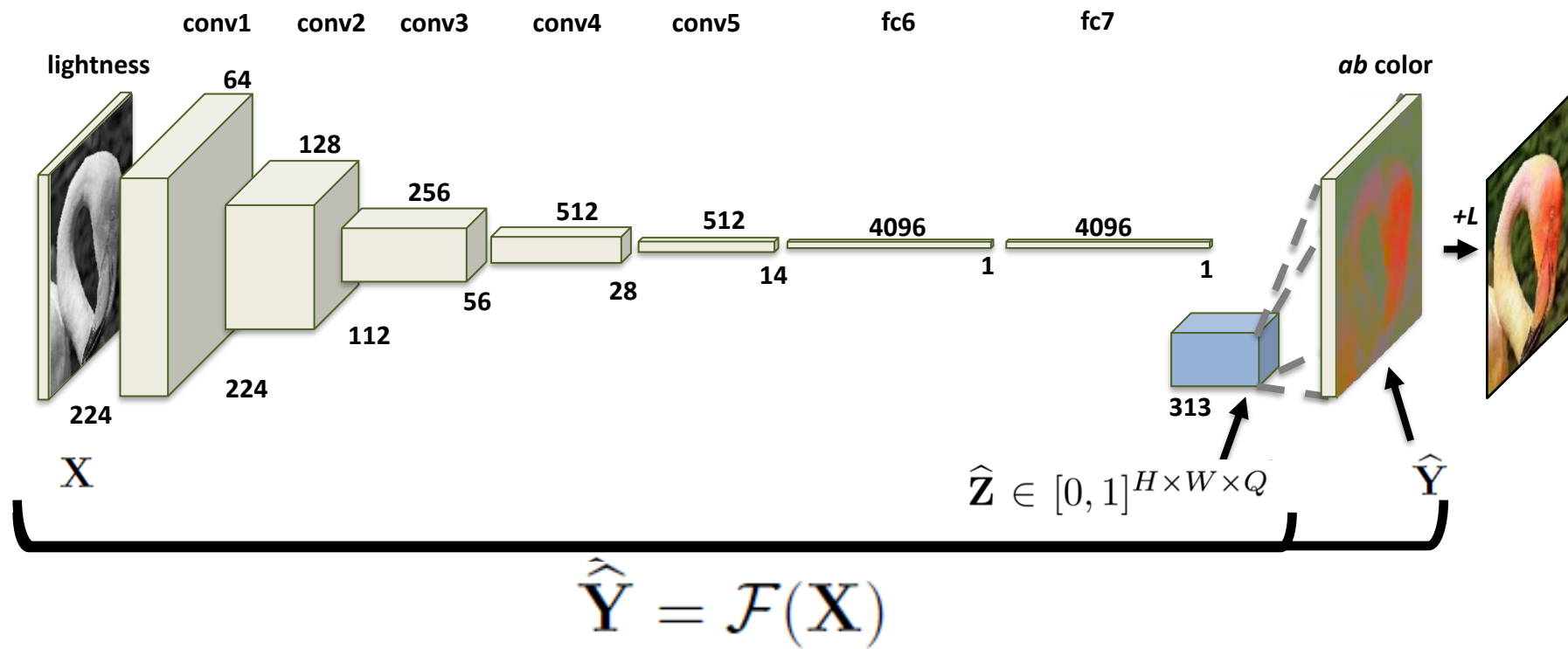


Dahl. Jan 2016. Iizuka et al. In SIGGRAPH, 2016.

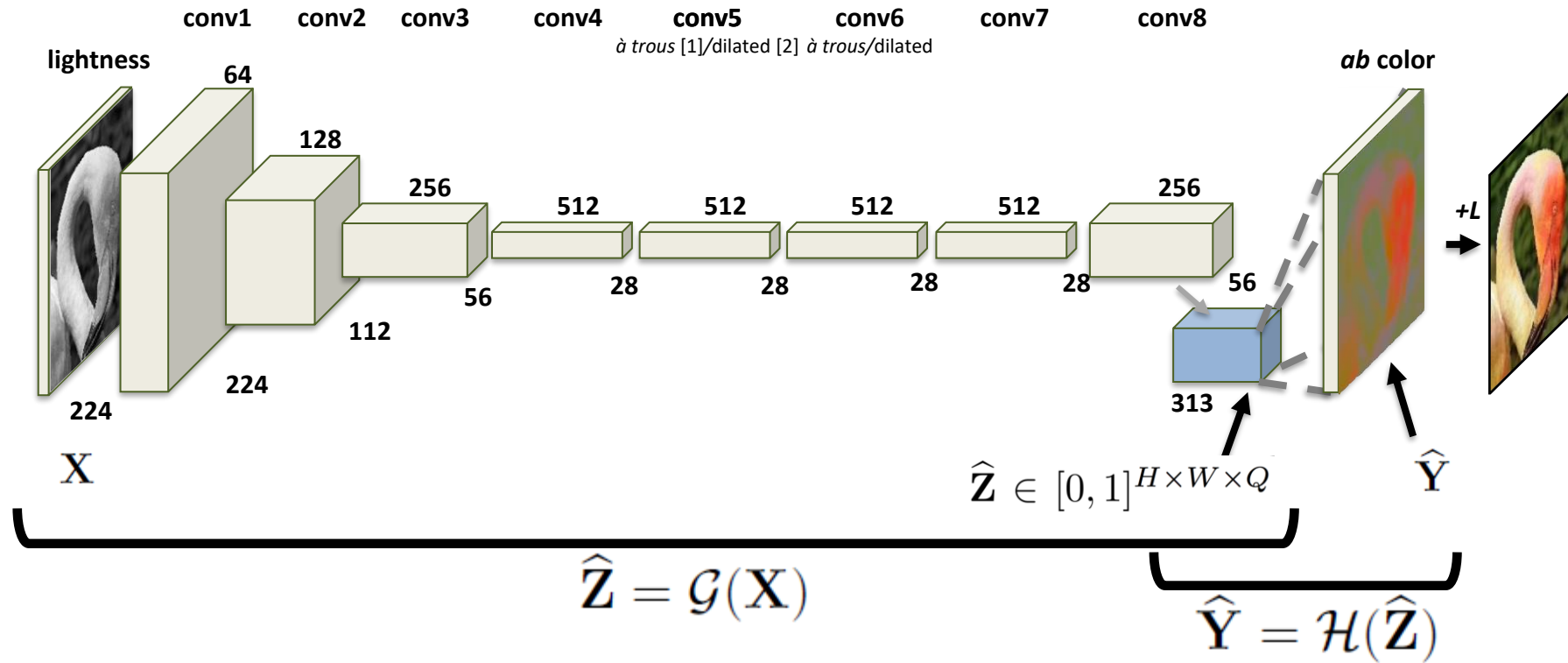


Larsson et al. In ECCV 2016. [Concurrent]

Network Architecture



Network Architecture



[1] Chen *et al.* In arXiv, 2016.

[2] Yu and Koltun. In ICLR, 2016

Ground Truth



L2 Regression



Class w/ Rebalancing



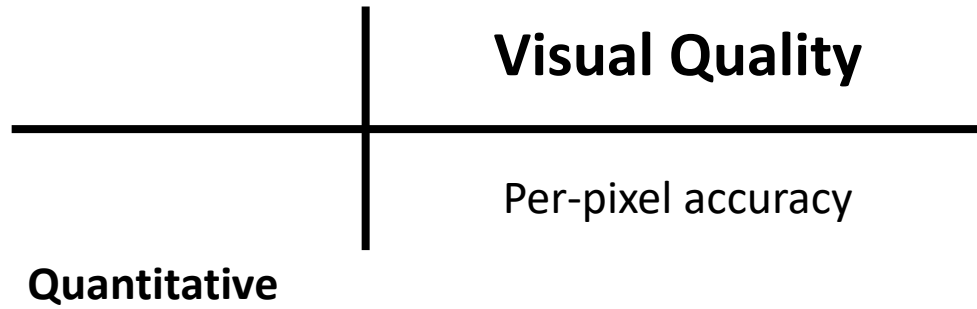
Failure Cases



Biases



Evaluation



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

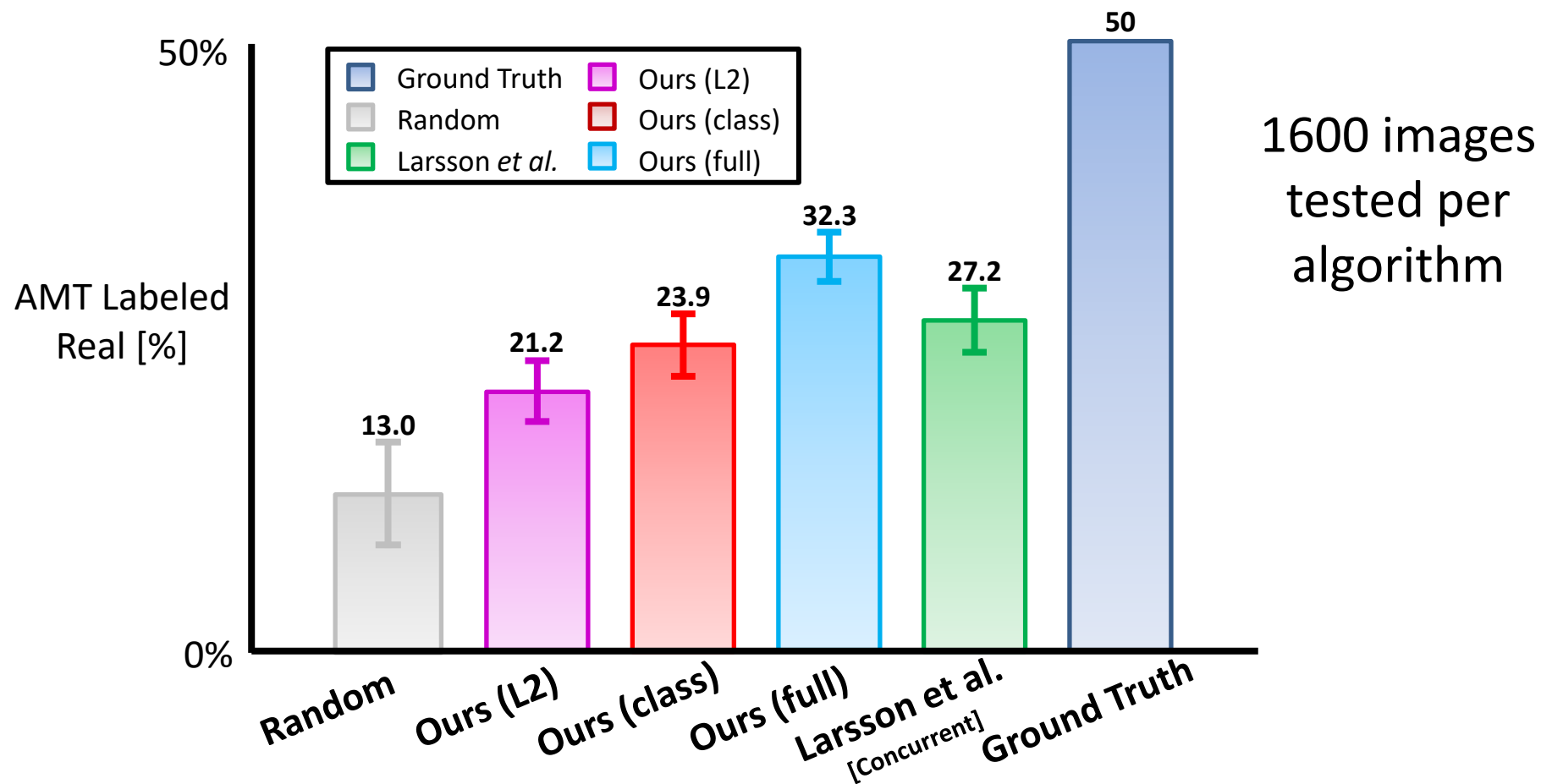


**Photo taken by
Reddit /u/Timteroo,
Mural from street
artist Eduardo Kobra**



**Recolorized
by Reddit
ColorizeBot**

Perceptual Realism Test



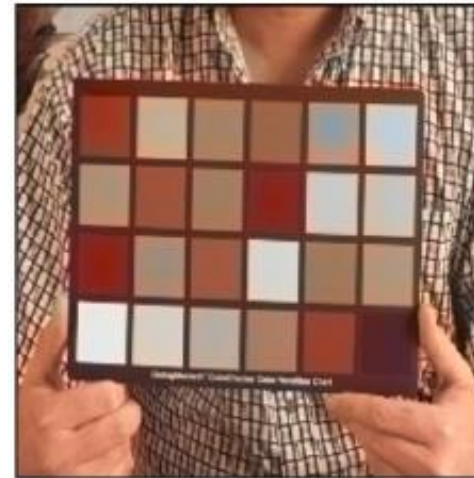
Input



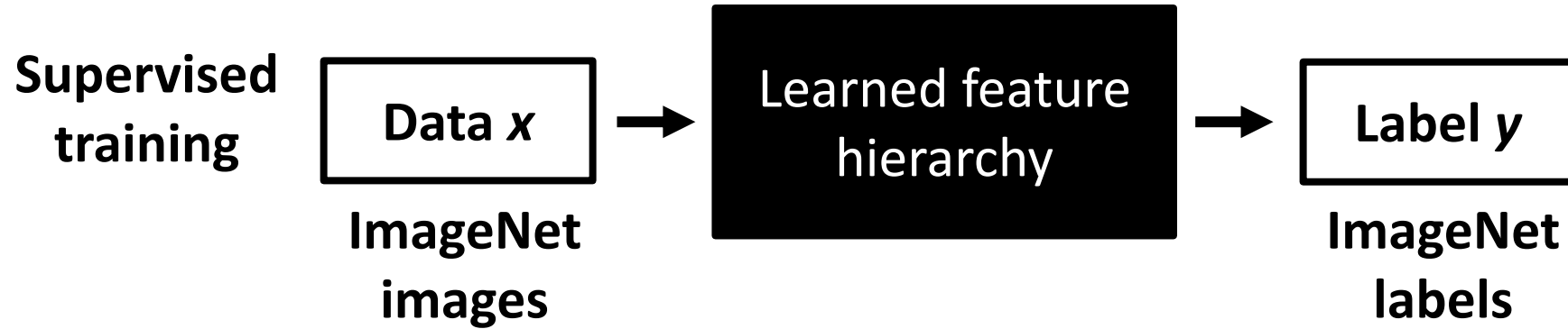
Ground Truth



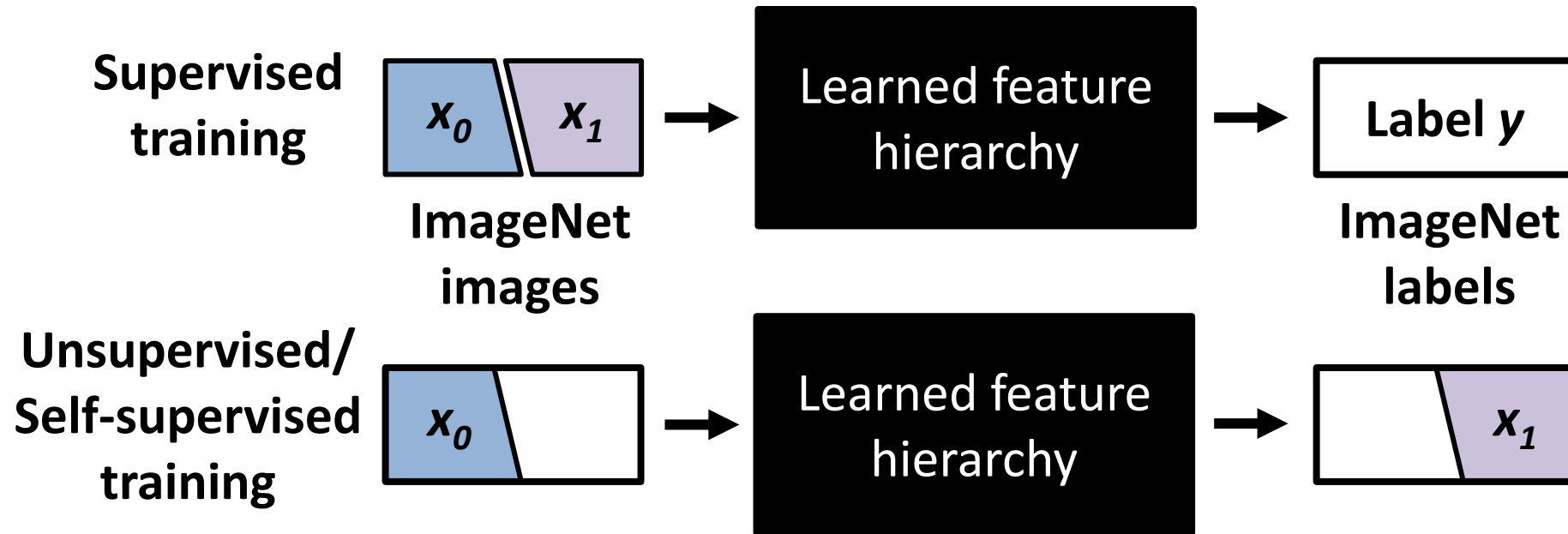
Output



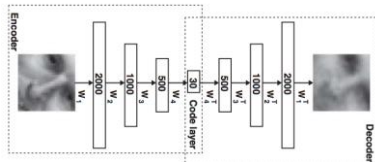
Predicting Labels from Data



Predicting Data from Data

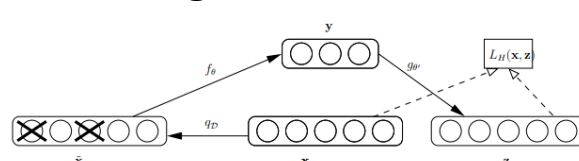


Autoencoders



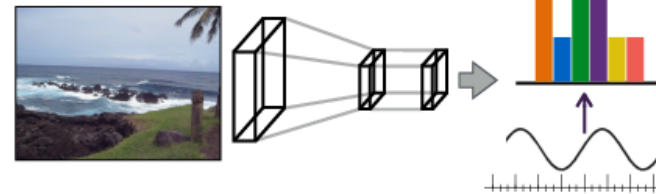
Hinton & Salakhutdinov.
Science 2006.

Denoising Autoencoders



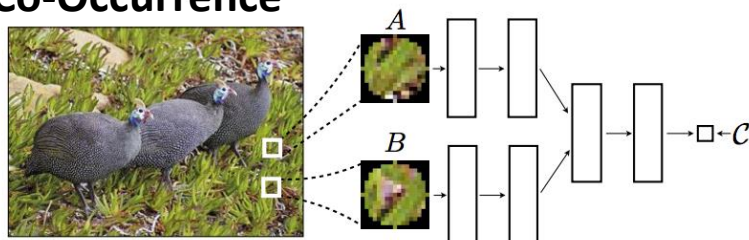
Vincent *et al.* ICML 2008.

Audio



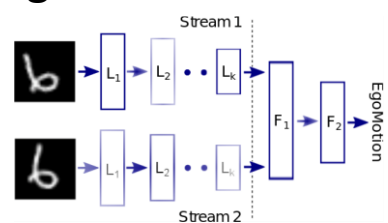
Owens *et al.* CVPR 2016, ECCV 2016

Co-Occurrence



Isola *et al.* ICLR Workshop 2016.

Egomotion

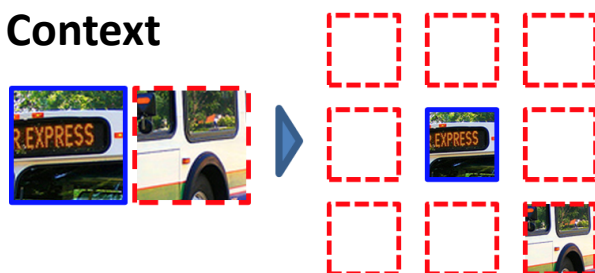


Agrawal *et al.* ICCV 2015



Jayaraman *et al.* ICCV 2015

Context

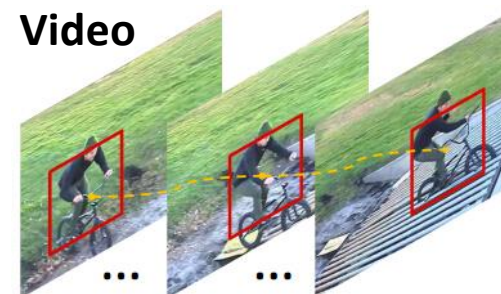


Doersch *et al.* ICCV 2015



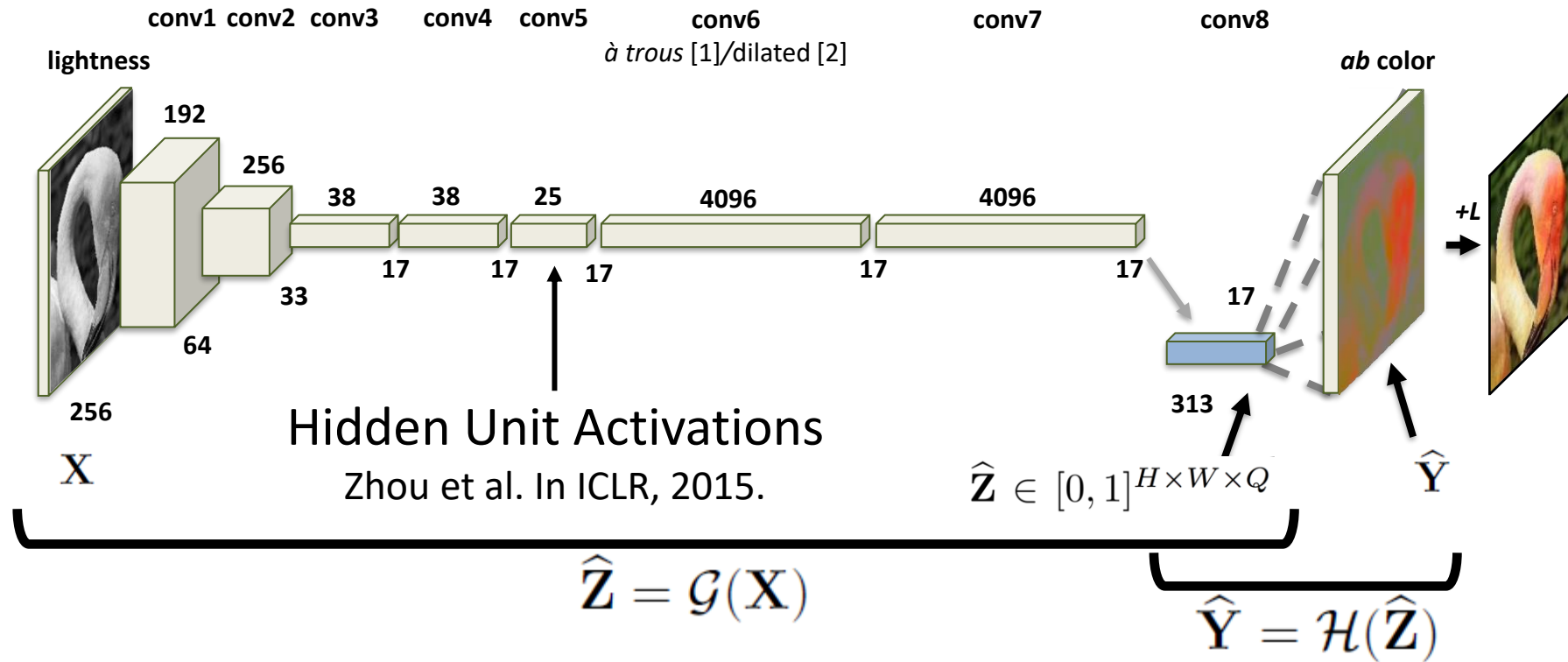
Pathak *et al.* CVPR 2016

Video



Wang *et al.* ICCV 2015

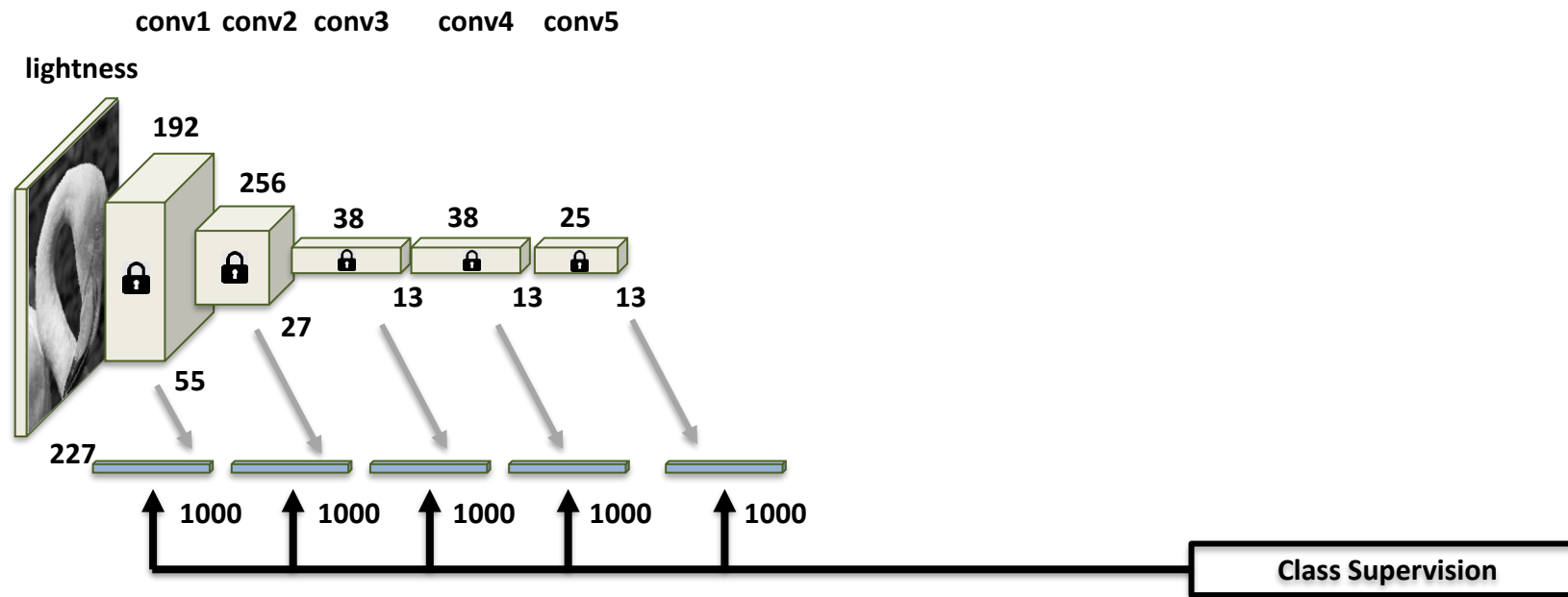
Cross-Channel Encoder



[1] Chen *et al.* In arXiv, 2016.

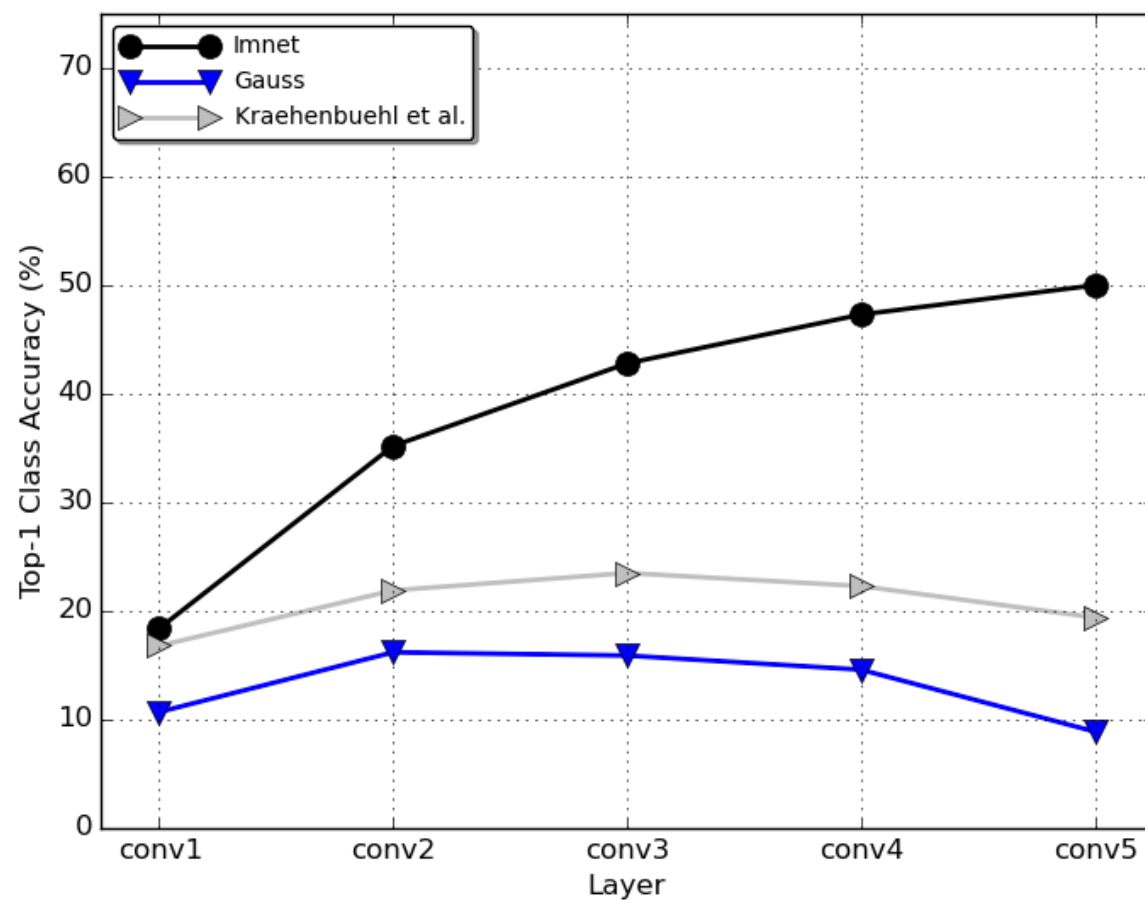
[2] Yu and Koltun. In ICLR, 2016

Task Generalization: ILSVRC linear classification

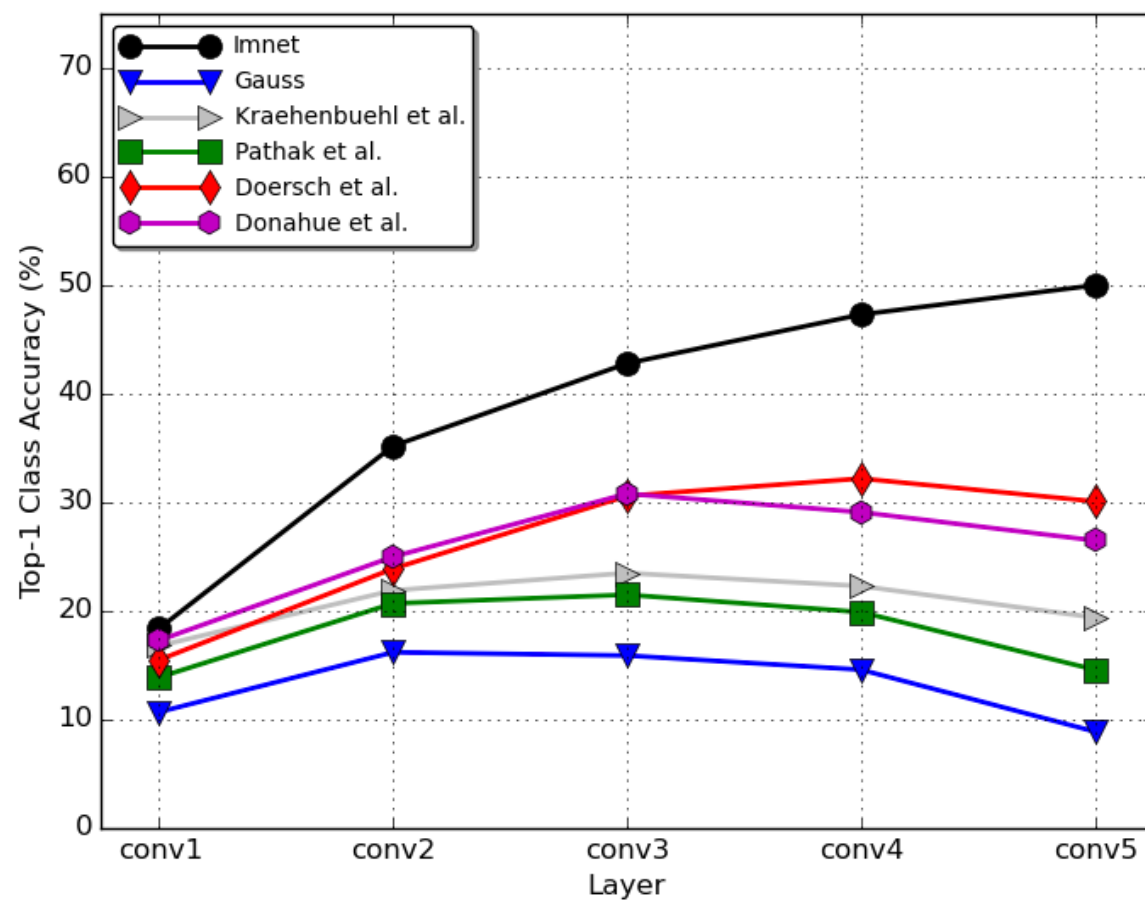


Are semantic classes *linearly separable*
in the learned feature space?

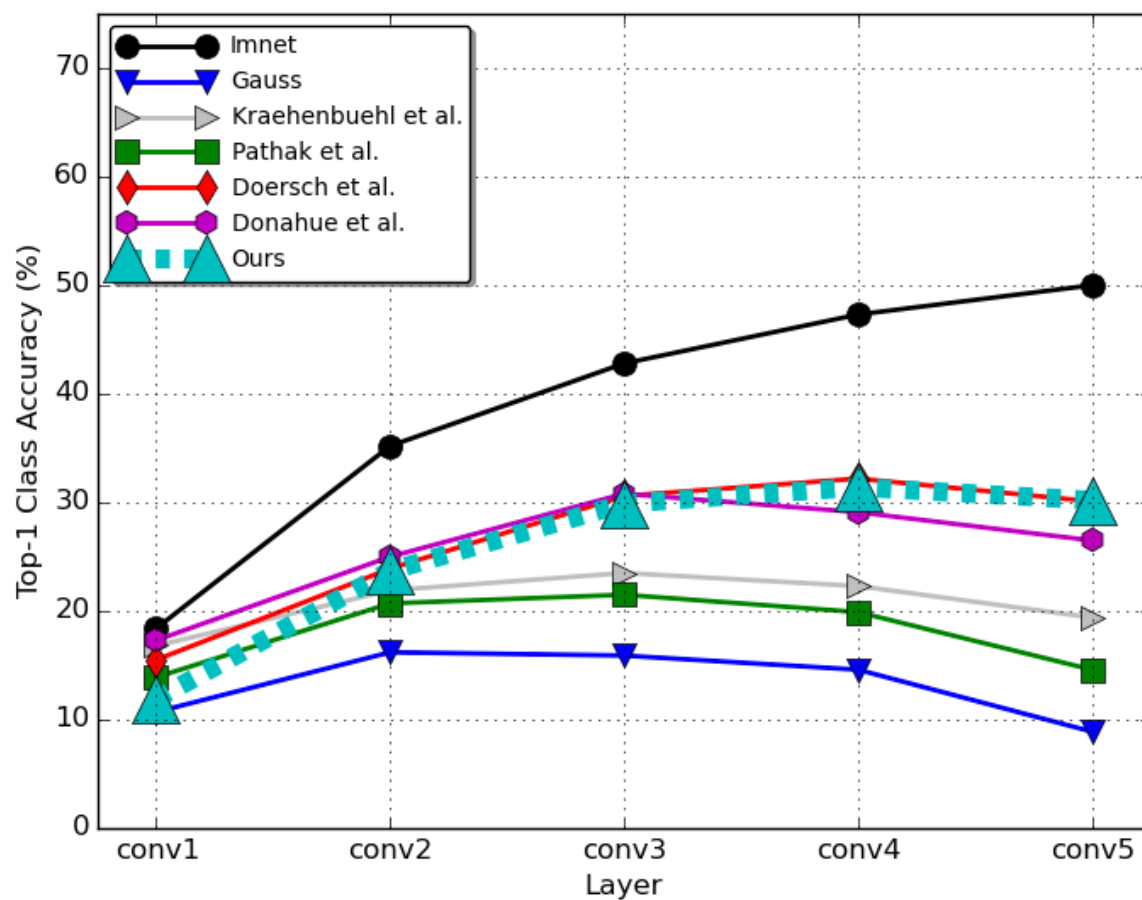
Task Generalization: ILSVRC linear classification



Task Generalization: ILSVRC linear classification

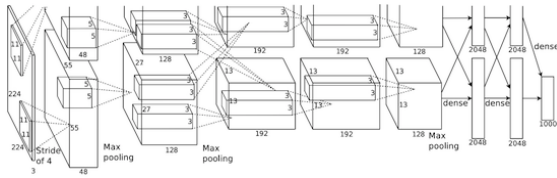


Task Generalization: ILSVRC linear classification



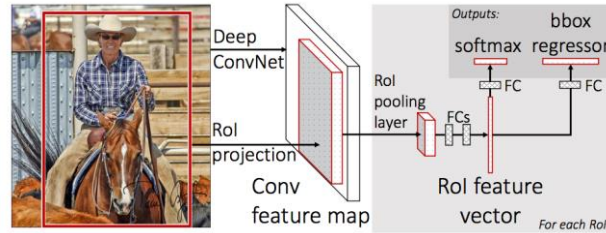
Dataset & Task Generalization on PASCAL VOC

Does the feature representation
transfer to other datasets and tasks?



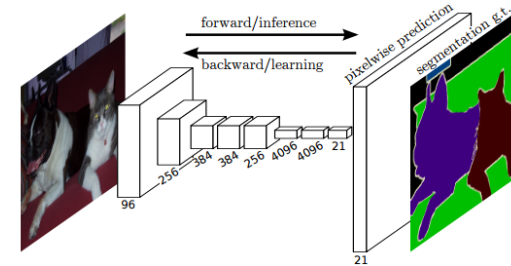
Classification

Krähenbühl et al. In ICLR, 2016.



Detection

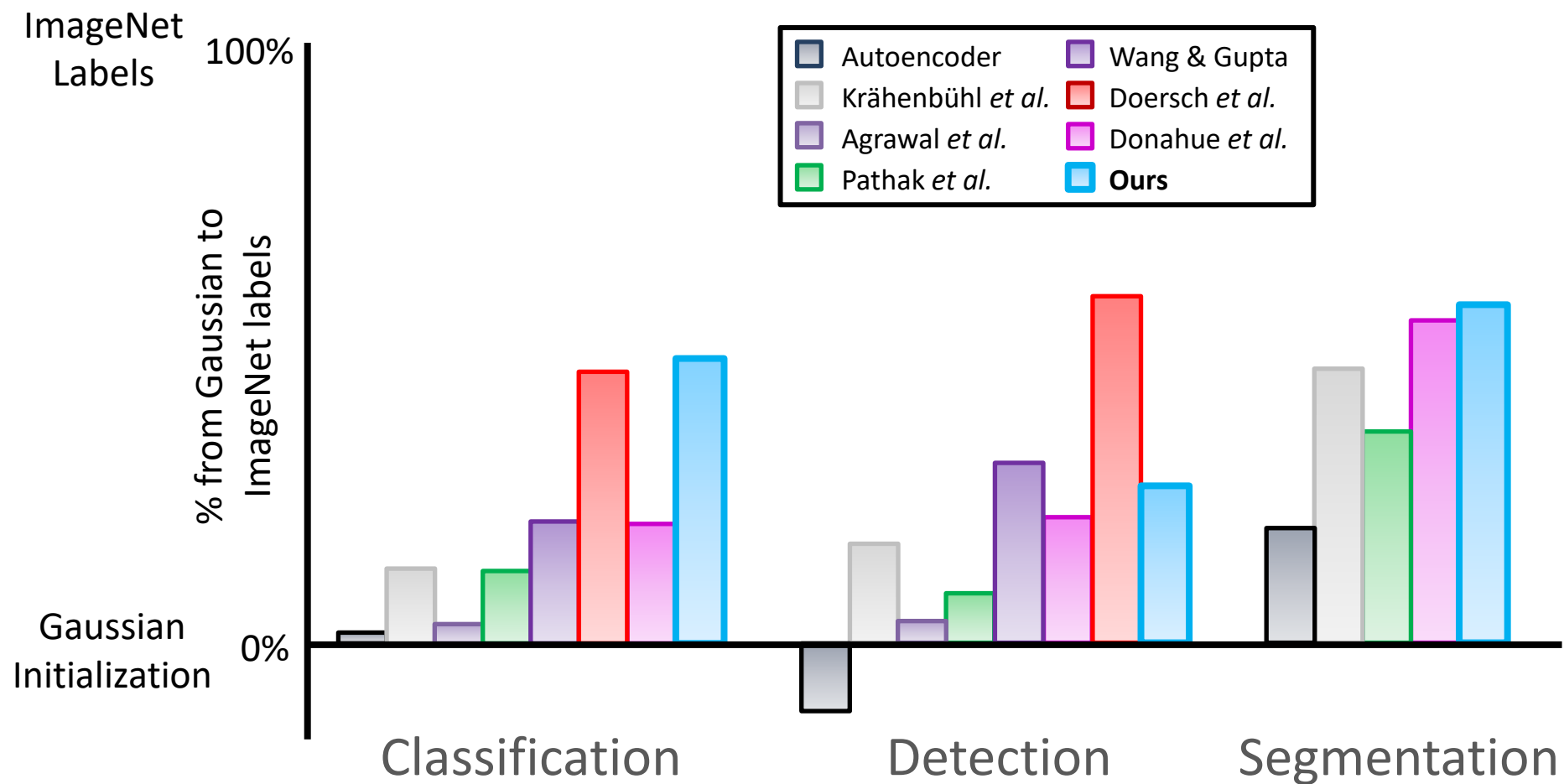
Fast R-CNN. Girshick. In ICCV, 2015.

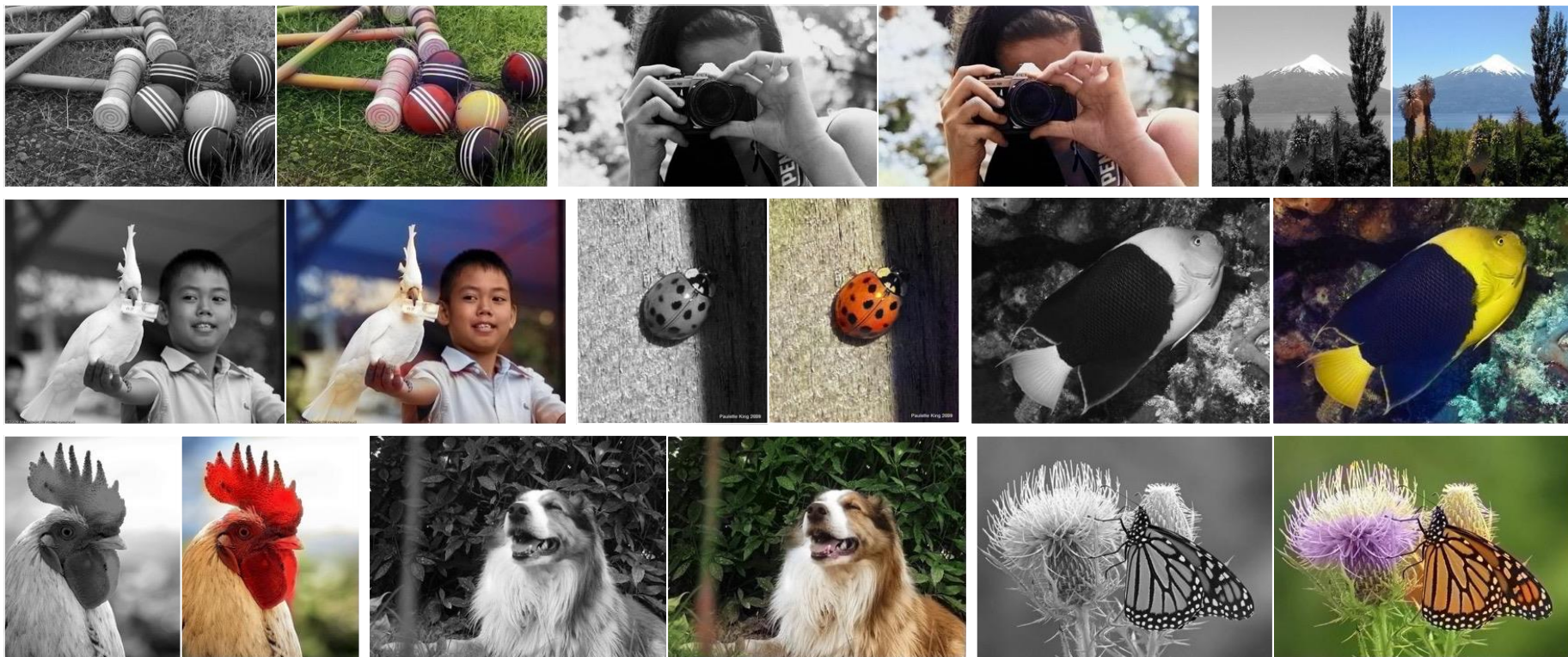


Segmentation

FCNs. Long et al. In CVPR, 2015.

Dataset & Task Generalization on PASCAL VOC





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

The Gelato Bet, Resolved

"If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, Mr. Malik promises to buy Mr. Efros one (1) gelato"



Clever ideas to keep in mind

- Turn a regression task into a classification task
- Fight back against class imbalance
 - Loss functions that emphasize rare classes
 - Replay rare classes more during training

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen¹ Simon Kornblith¹ Mohammad Norouzi¹ Geoffrey Hinton¹

SimCLR, IMCL 2020

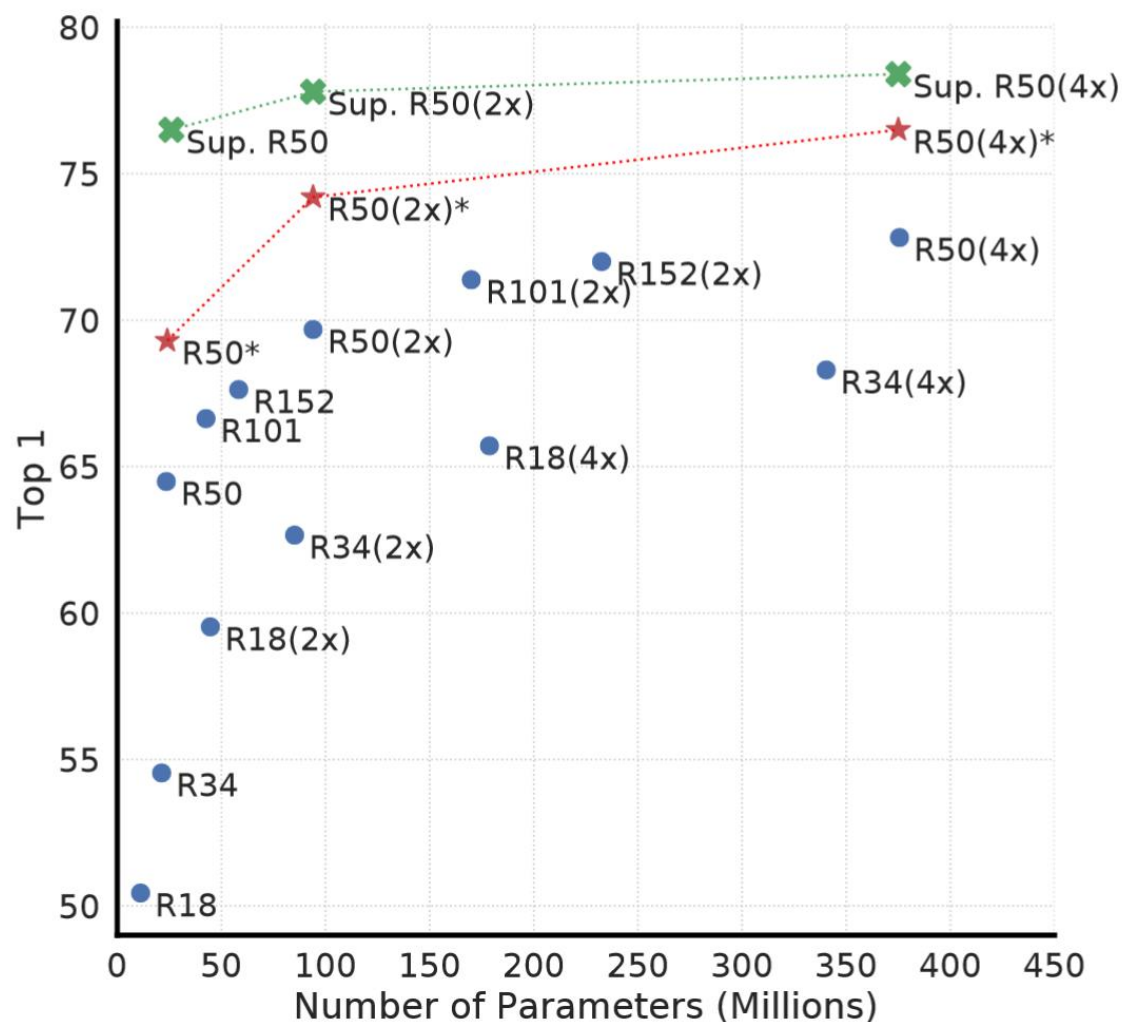
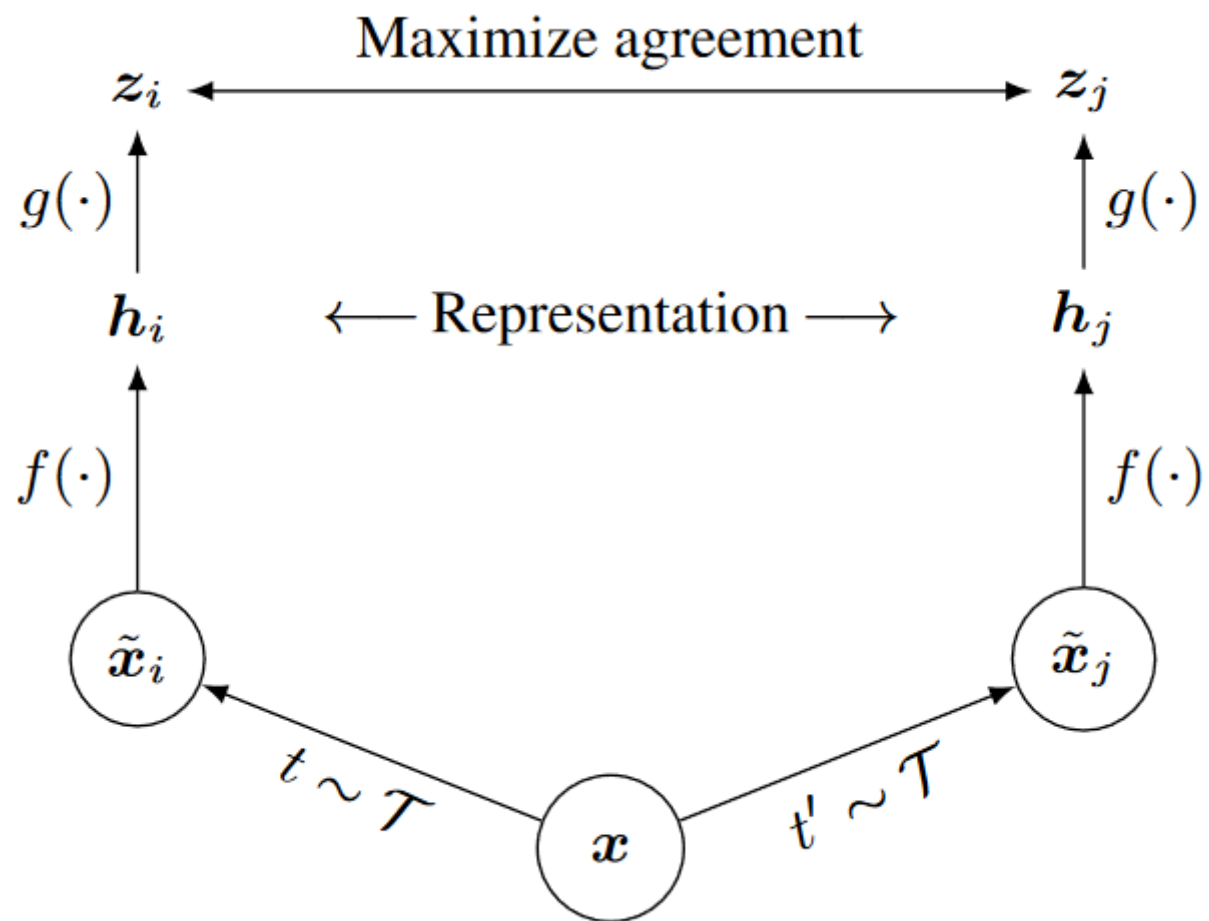


Figure 7. Linear evaluation of models with varied depth and width. Models in blue dots are ours trained for 100 epochs, models in red stars are ours trained for 1000 epochs, and models in green crosses are supervised ResNets trained for 90 epochs⁷ (He et al., 2016).





(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

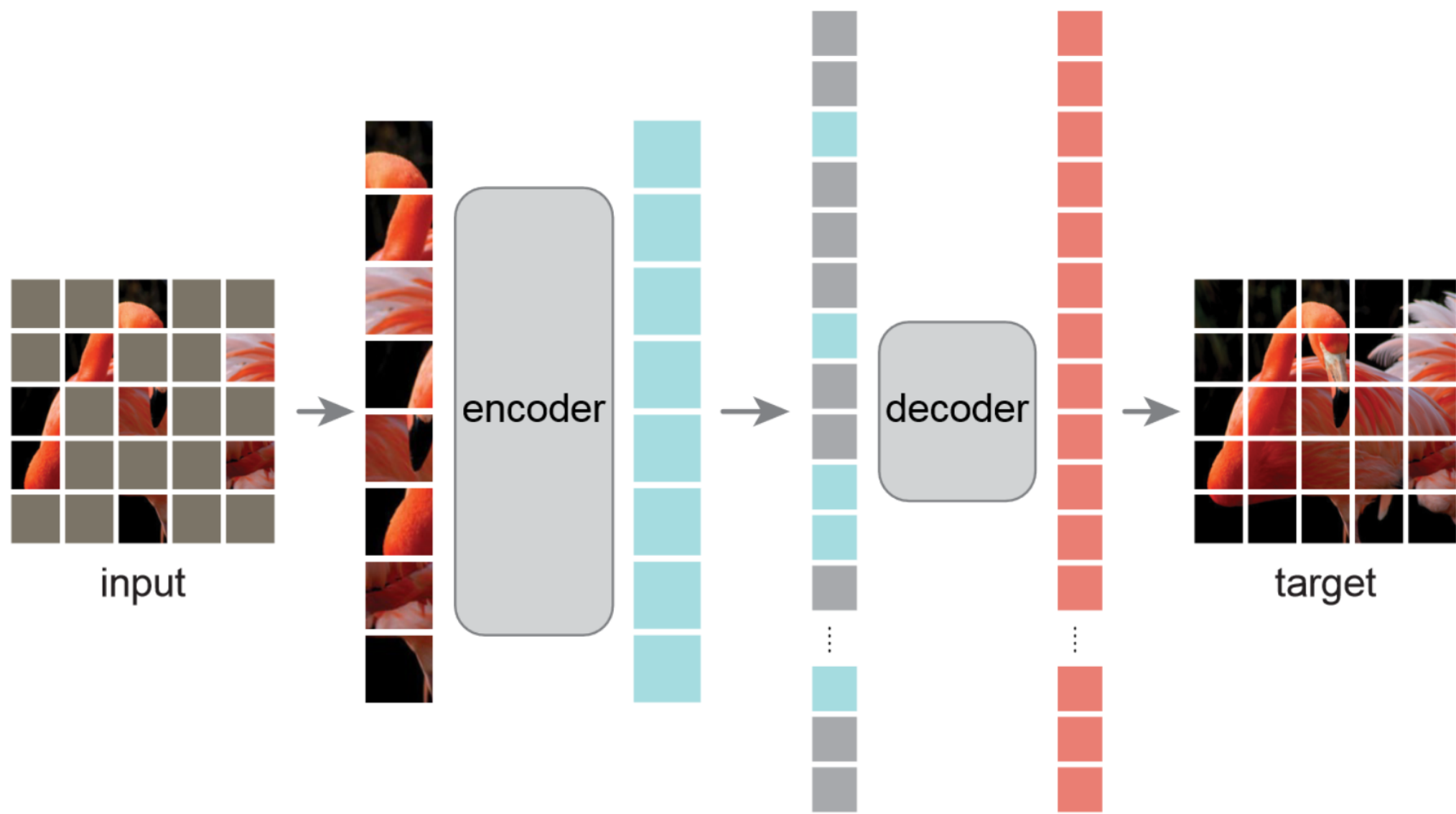
Masked Autoencoders Are Scalable Vision Learners

Kaiming He^{*,†} Xinlei Chen^{*} Saining Xie Yanghao Li Piotr Dollár Ross Girshick

^{*}equal technical contribution [†]project lead

Facebook AI Research (FAIR)

CVPR 2022



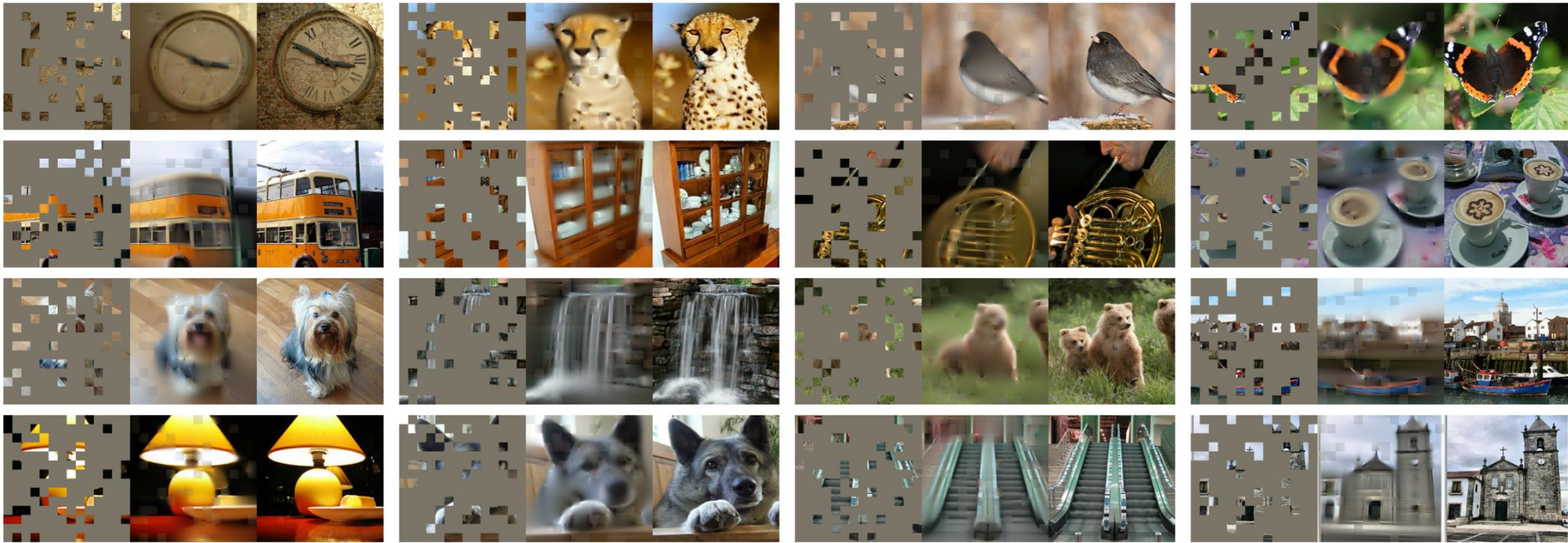


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction[†] (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.

[†]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.

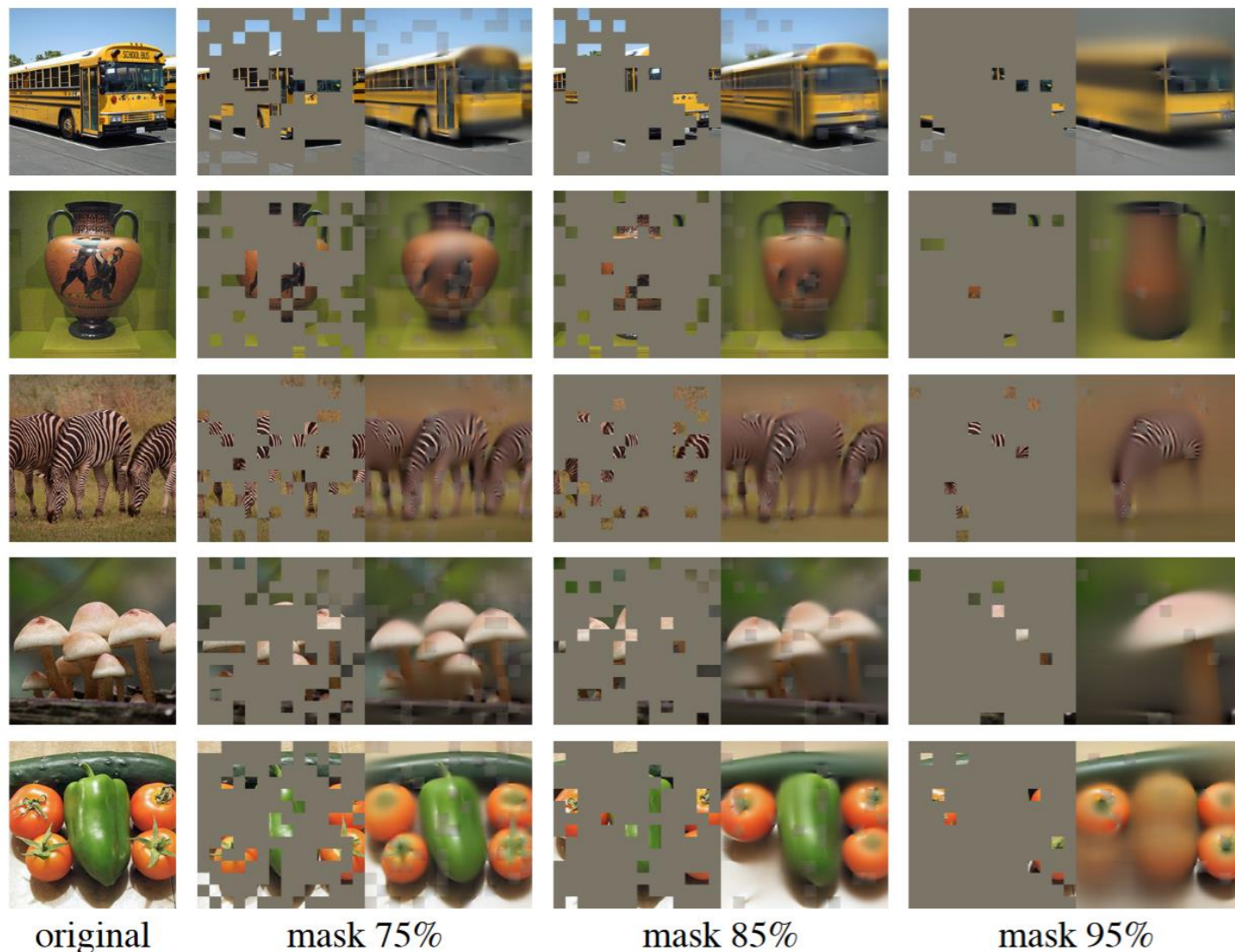
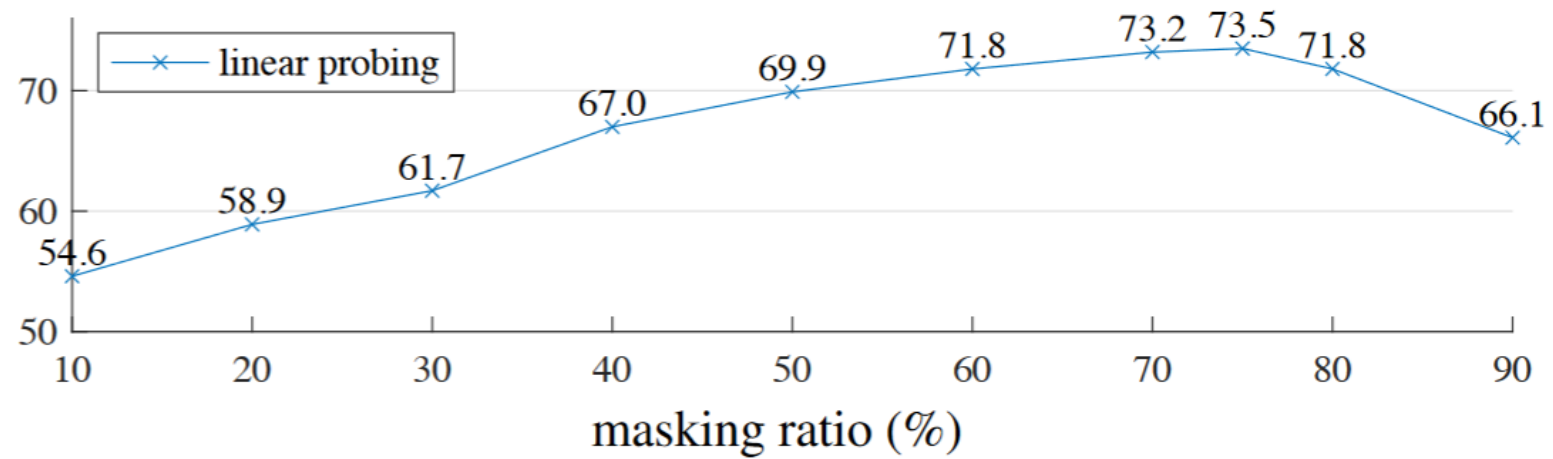
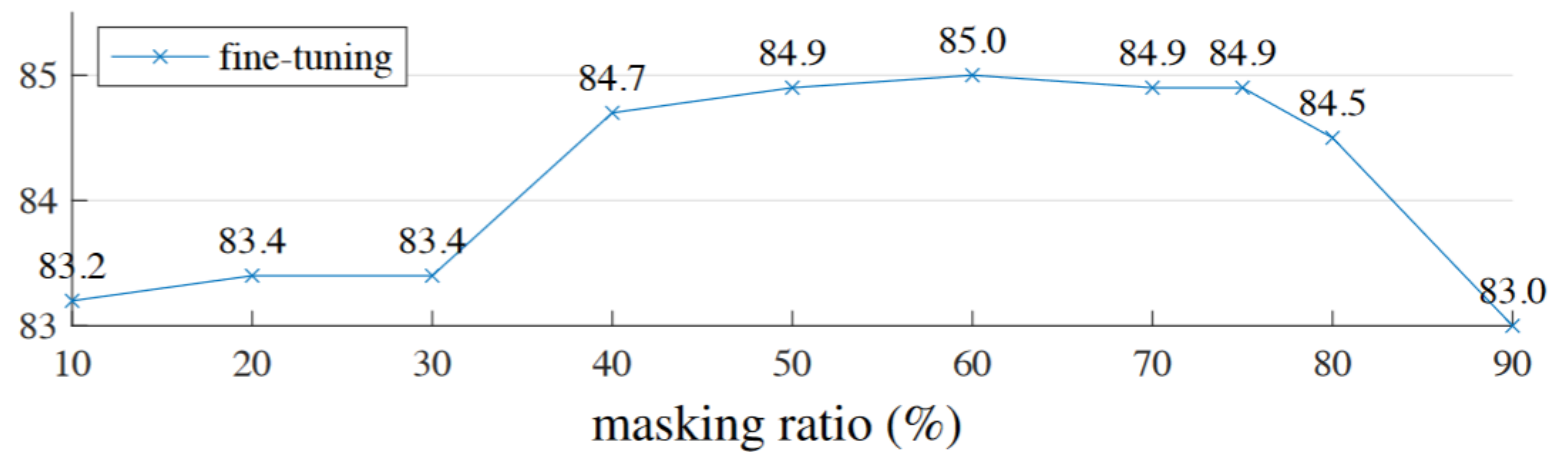


Figure 4. Reconstructions of ImageNet *validation* images using an MAE pre-trained with a masking ratio of 75% but applied on inputs with higher masking ratios. The predictions differ plausibly from the original images, showing that the method can generalize.



method	pre-train data	AP^{box}		AP^{mask}	
		ViT-B	ViT-L	ViT-B	ViT-L
supervised	IN1K w/ labels	47.9	49.3	42.9	43.9
MoCo v3	IN1K	47.9	49.3	42.7	44.0
BEiT	IN1K+DALLE	49.8	53.3	44.4	47.1
MAE	IN1K	50.3	53.3	44.9	47.2

Table 4. **COCO object detection and segmentation** using a ViT Mask R-CNN baseline. All entries are based on our implementation. Self-supervised entries use IN1K data *without* labels. Mask AP follows a similar trend as box AP.

Conclusion

- With the right “pretext” tasks and architectures, we are pretty close to matching supervised performance with self-supervised approaches. But it takes some work (longer training, bigger models, precise hyperparameter tuning)
- SimCLR and Masked AutoEncoder only train on ImageNet images. But couldn't you use a lot more data if you don't need human labels?
- The gelato bet was just a bit premature

Conclusion

- Supervised Deep Learning (2012-2020): Similar to Kepler or Hubble. The algorithms (CNNs) existed in the 90s, but the field was fundamentally stalled waiting for the "telescope" (GPUs) and the "star catalogs" (ImageNet—millions of human-annotated data points).
- Self-Supervised Learning (Present): Similar to Eratosthenes or Einstein. We realize the data we already have (pixels, unlabeled images) contains its own inherent structure. We just needed the algorithmic cleverness to set up pretext tasks (like masking or contrastive learning) to force the network to learn physics, geometry, and semantics without a human explicitly pointing it out.