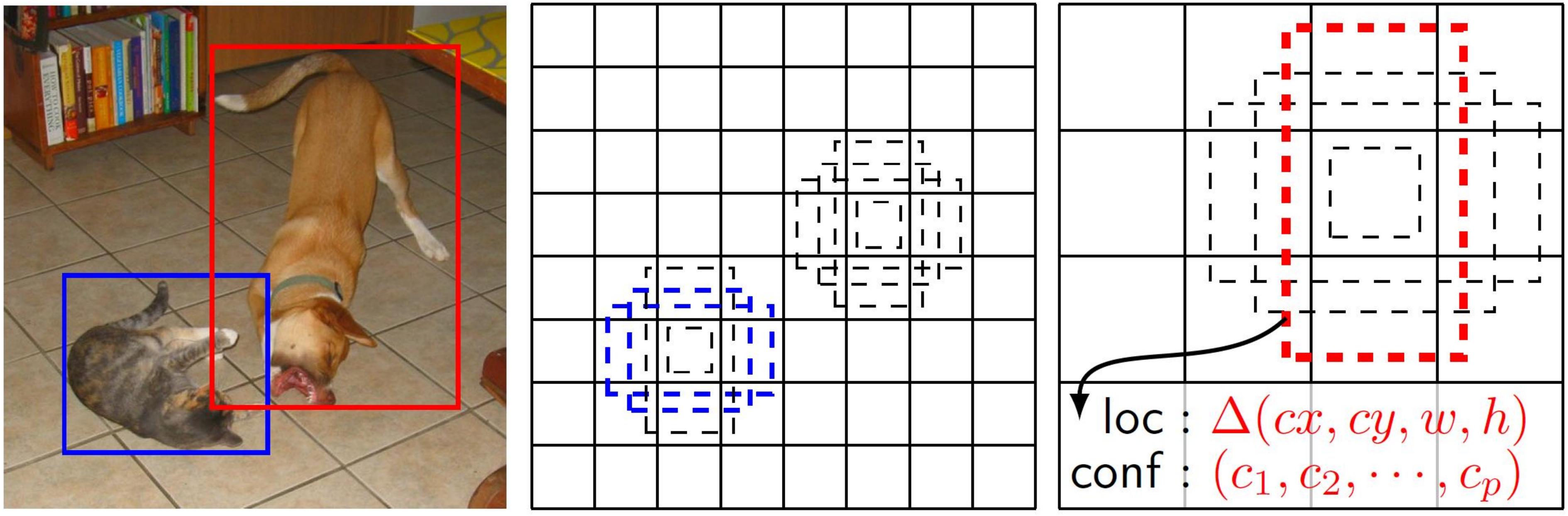
“Unsupervised” Deep Learning

James Hays
slides from Carl Doersch and Richard Zhang

Recap from Previous Lecture

- We saw two strategies to get *structured* output while using deep learning
  – With object detection, one strategy is brute force: detect everywhere at once



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

Recap from Previous Lecture

- We saw two strategies to get *structured* output while using deep learning
  - With pose estimation / keypoint detection, the network produces an image-based intermediate representation

Part Detection

Part Association

Recap from Previous Lecture

- More generally, it can pay off to get creative. Even if Deep ConvNets aren’t a natural fit for an image-related task, they might be able to learn a subtask or create a useful intermediate representation.

Today's Lecture

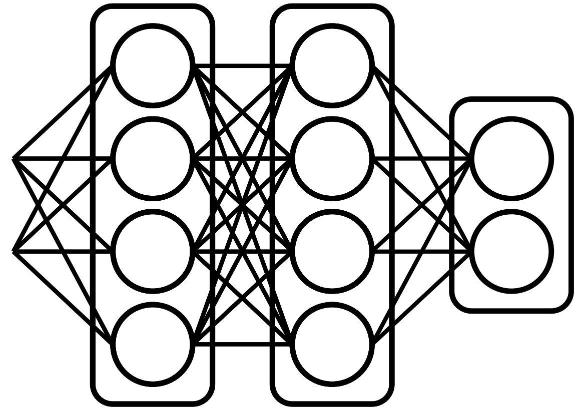
- Two methods for "unsupervised" deep learning
  - Context Prediction. Doersch et al. ICCV 2015
  - Colorful Image Colorization. Zhang et al. ECCV 2016
- Big picture: do we need big datasets like ImageNet to make deep learning worthwhile? Can we learn from something else?

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch
Joint work with Alexei A. Efros & Abhinav Gupta

ICCV 2015

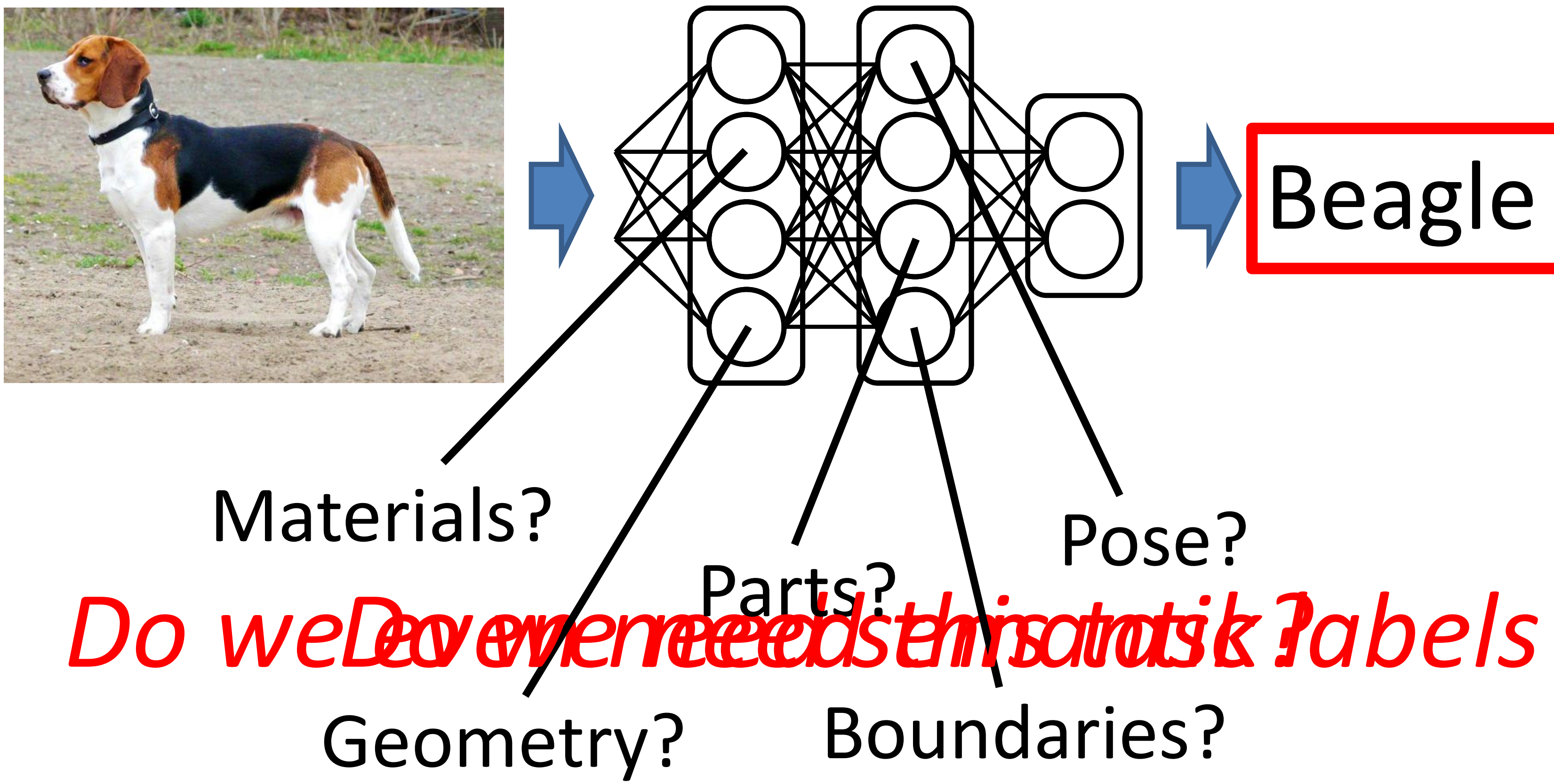
ImageNet + Deep Learning



Beagle

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

ImageNet + Deep Learning

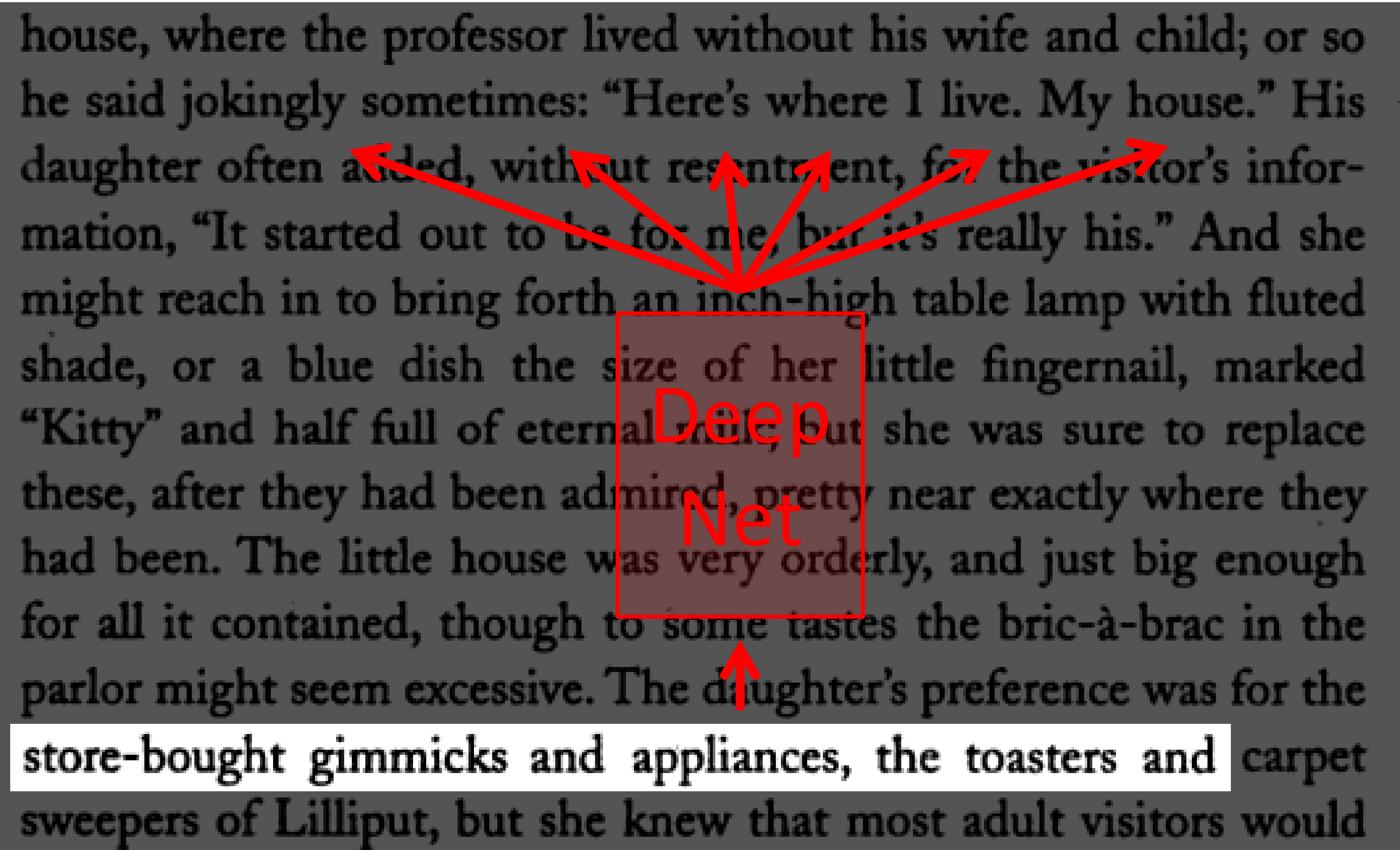


Do we even need semantic labels?

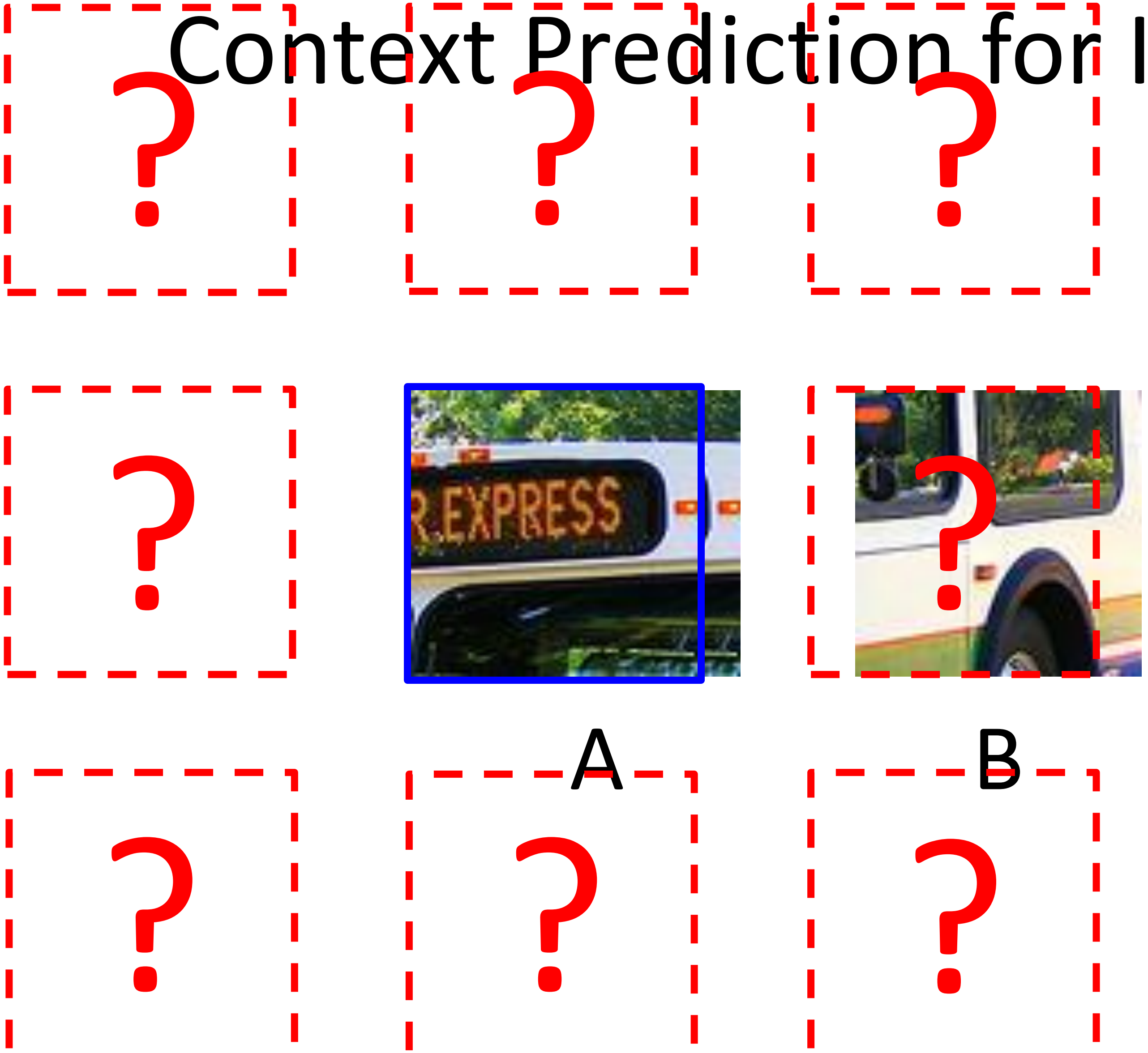
Do we need this task?

Context as Supervision

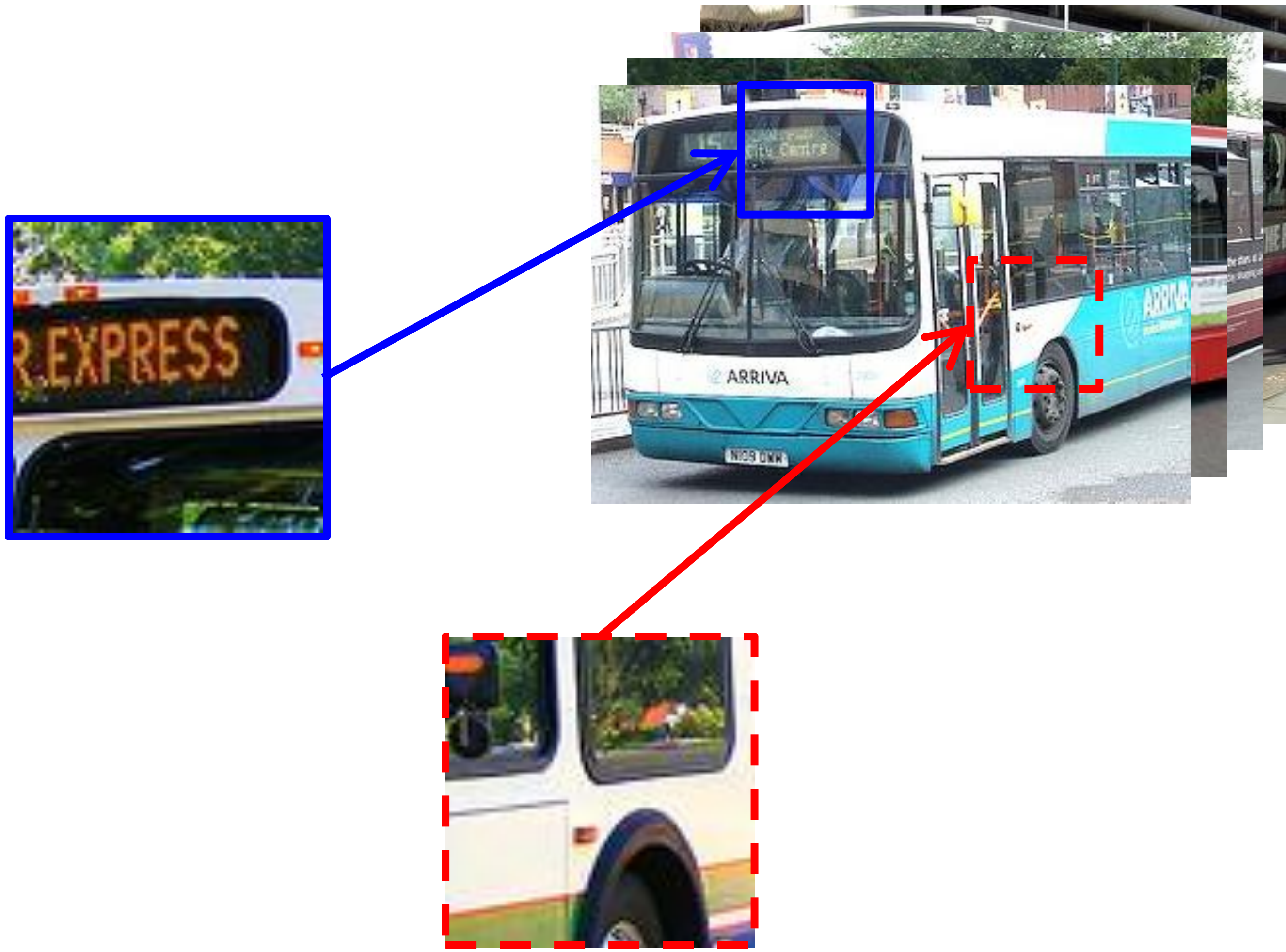
[Collobert & Weston 2008; Mikolov et al. 2013]


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

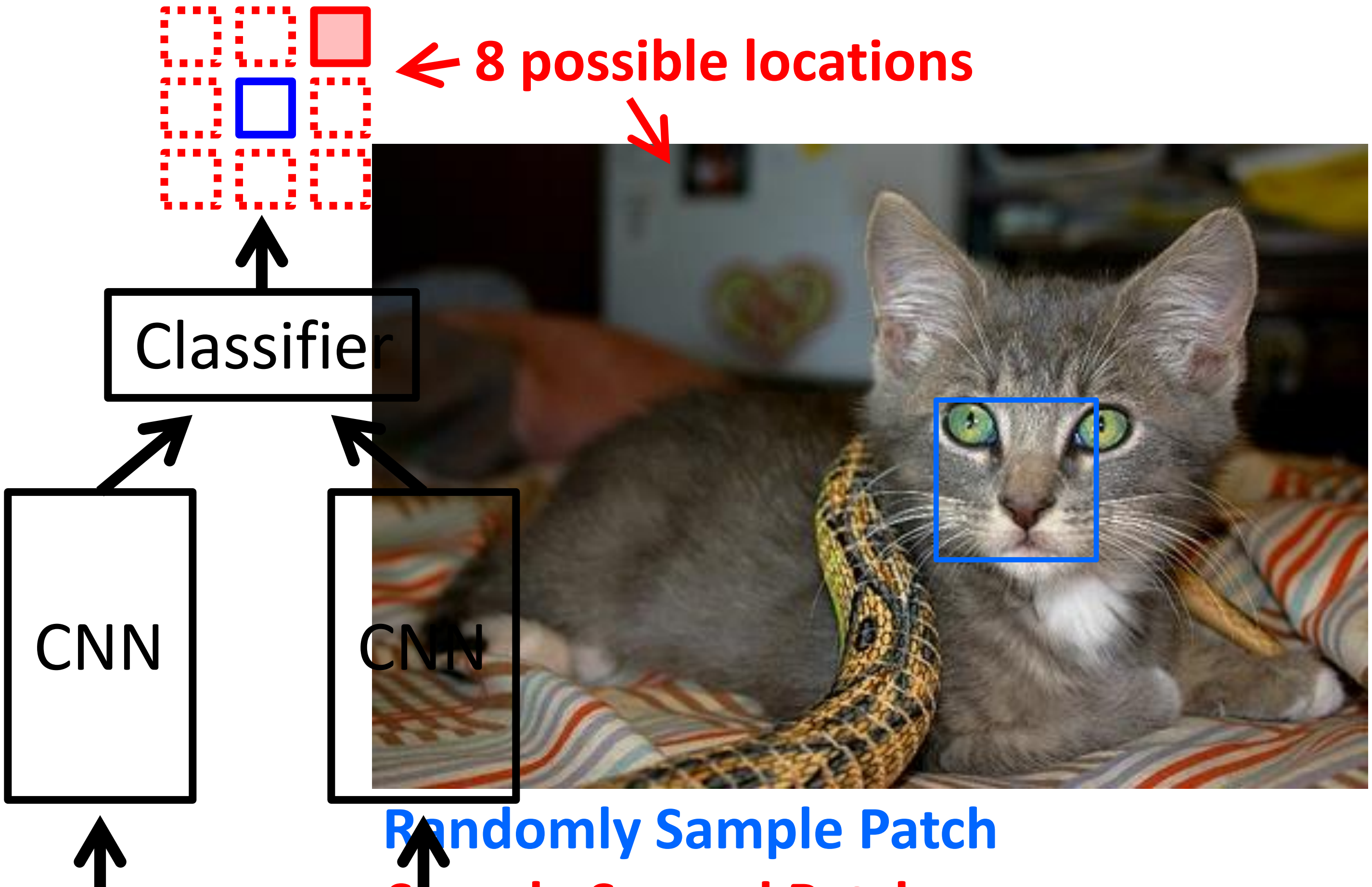
Context Prediction for Images

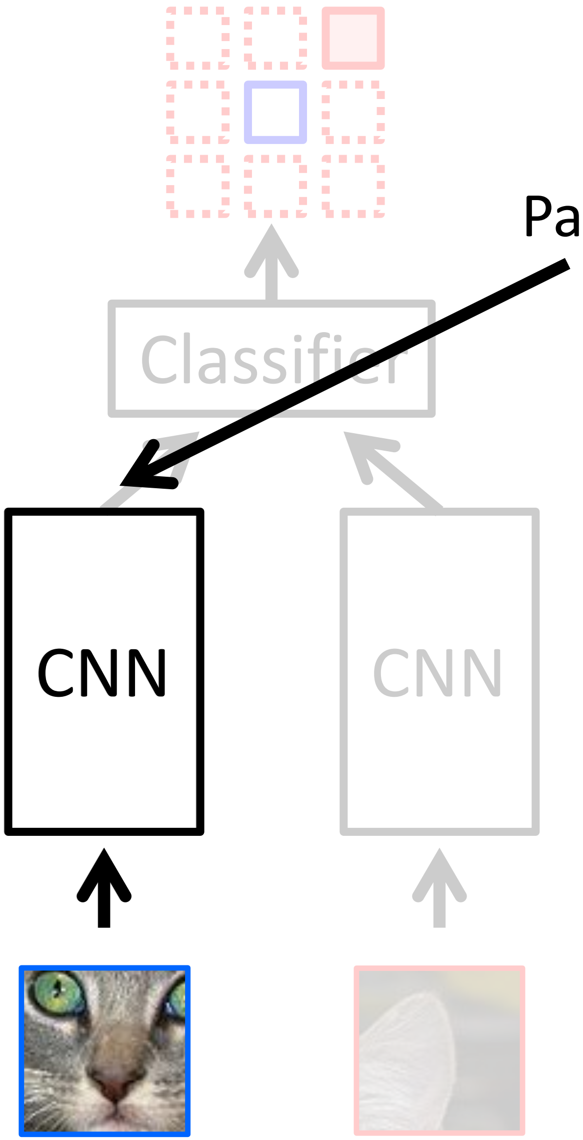


Semantics from a non-semantic task



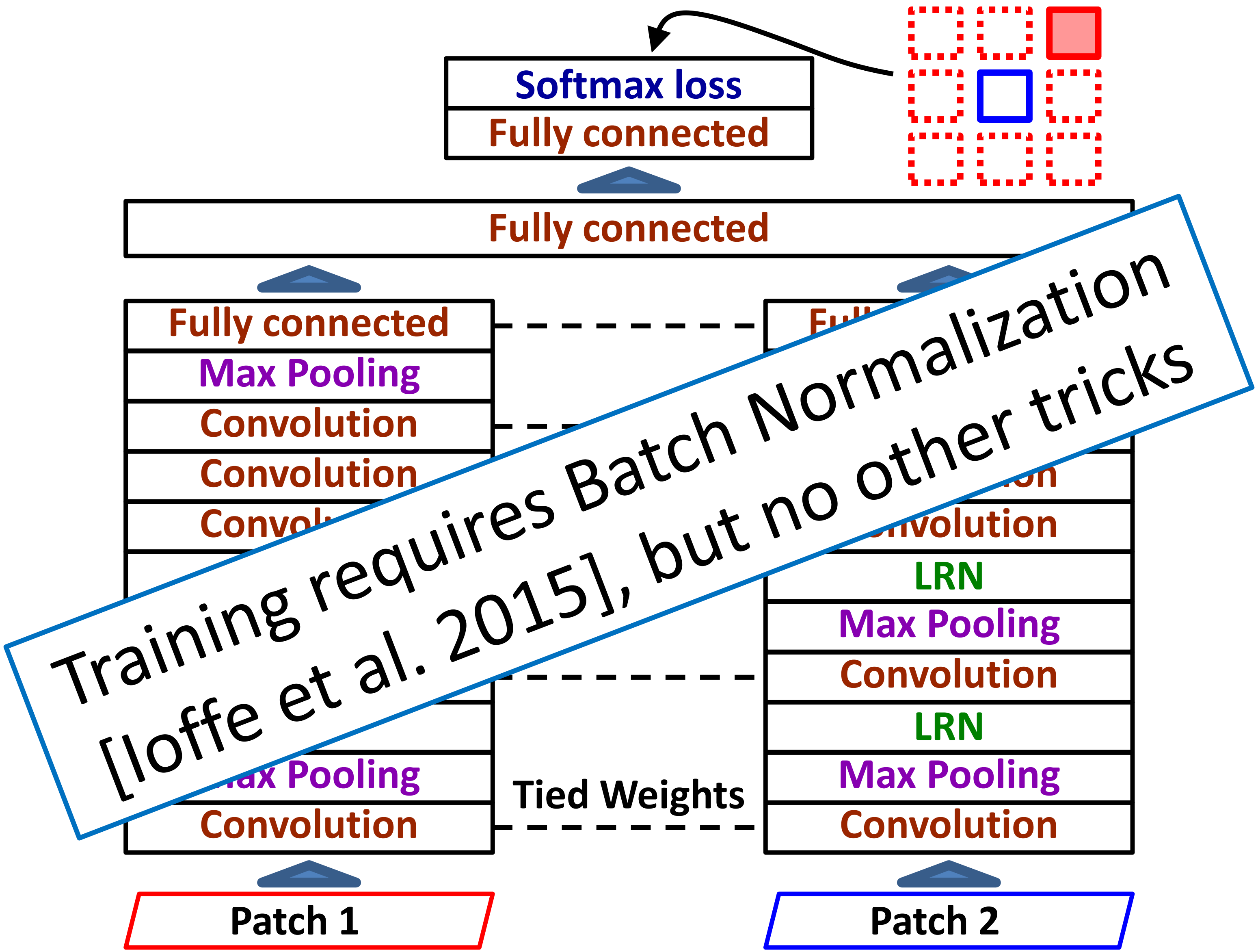
Relative Position Task



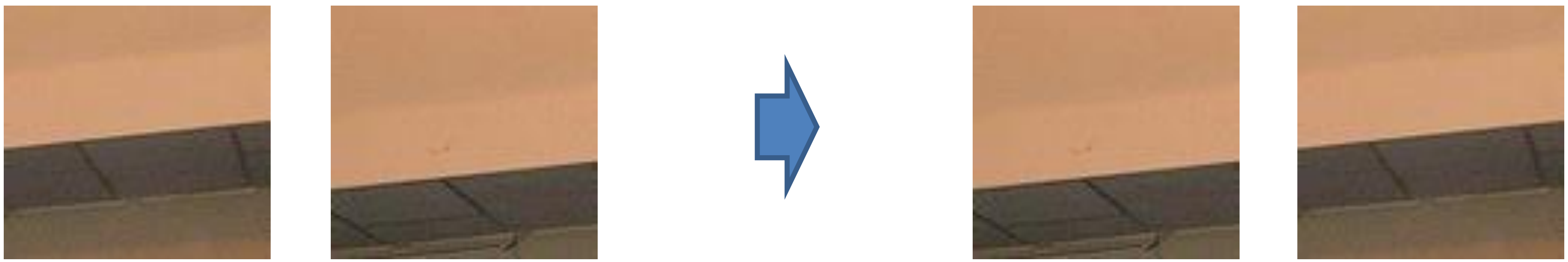


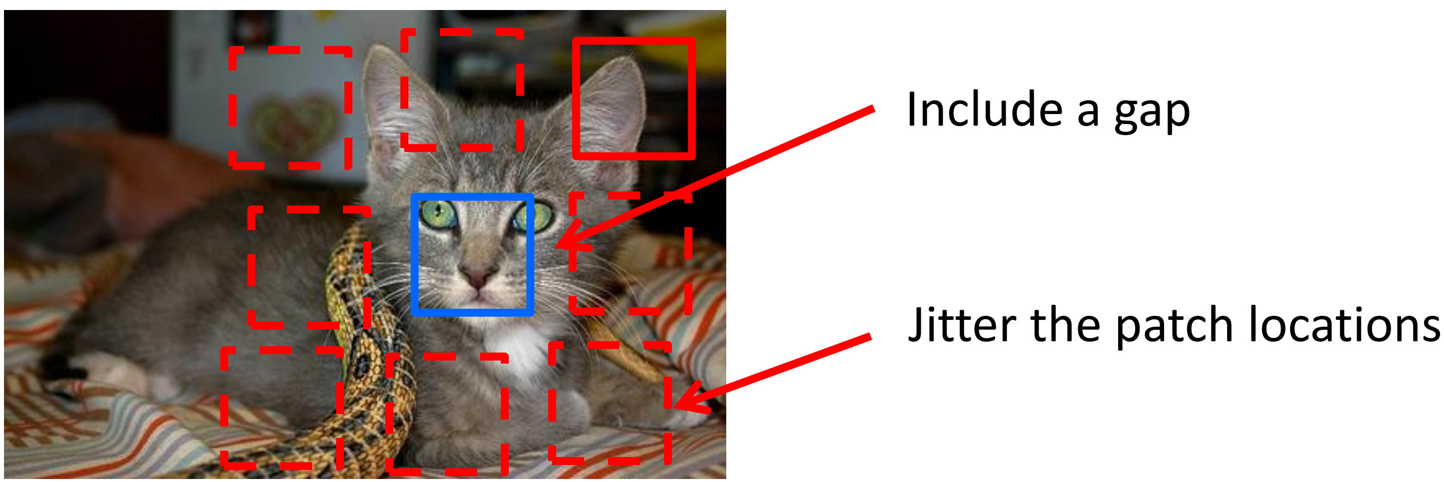
Note: connects ***across*** instances!

Architecture

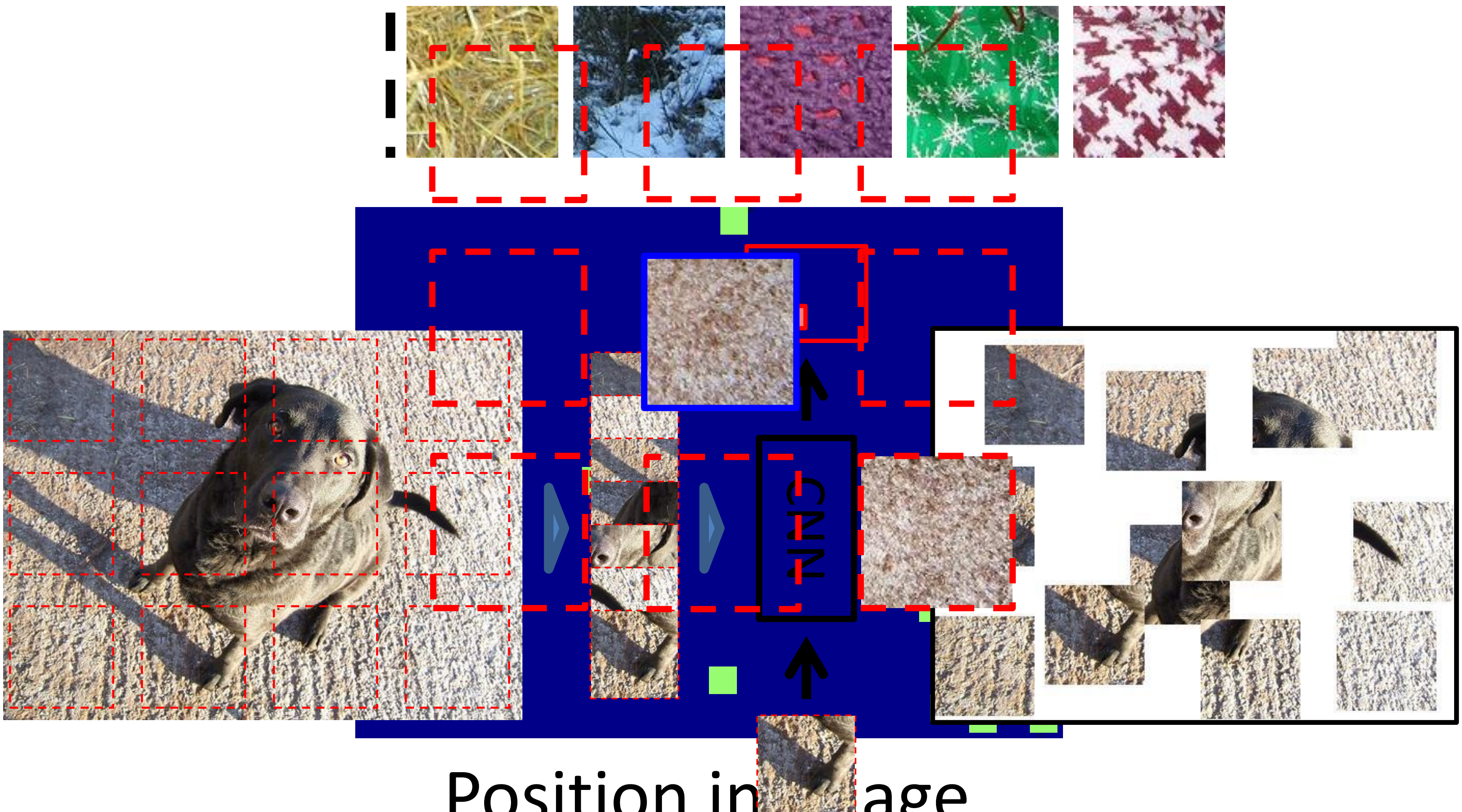


Avoiding Trivial Shortcuts

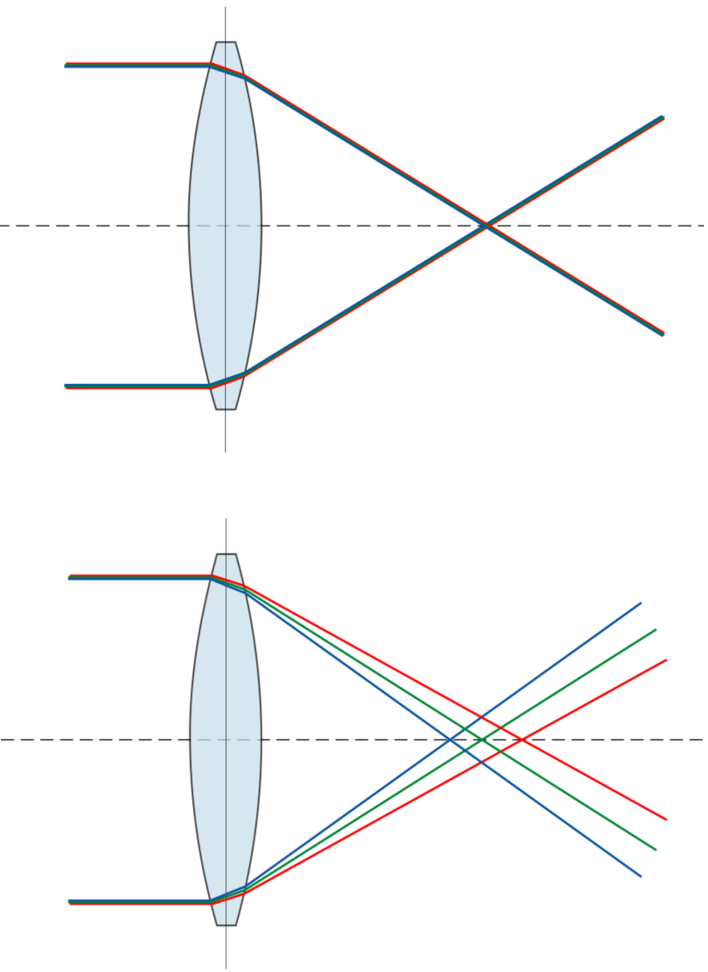


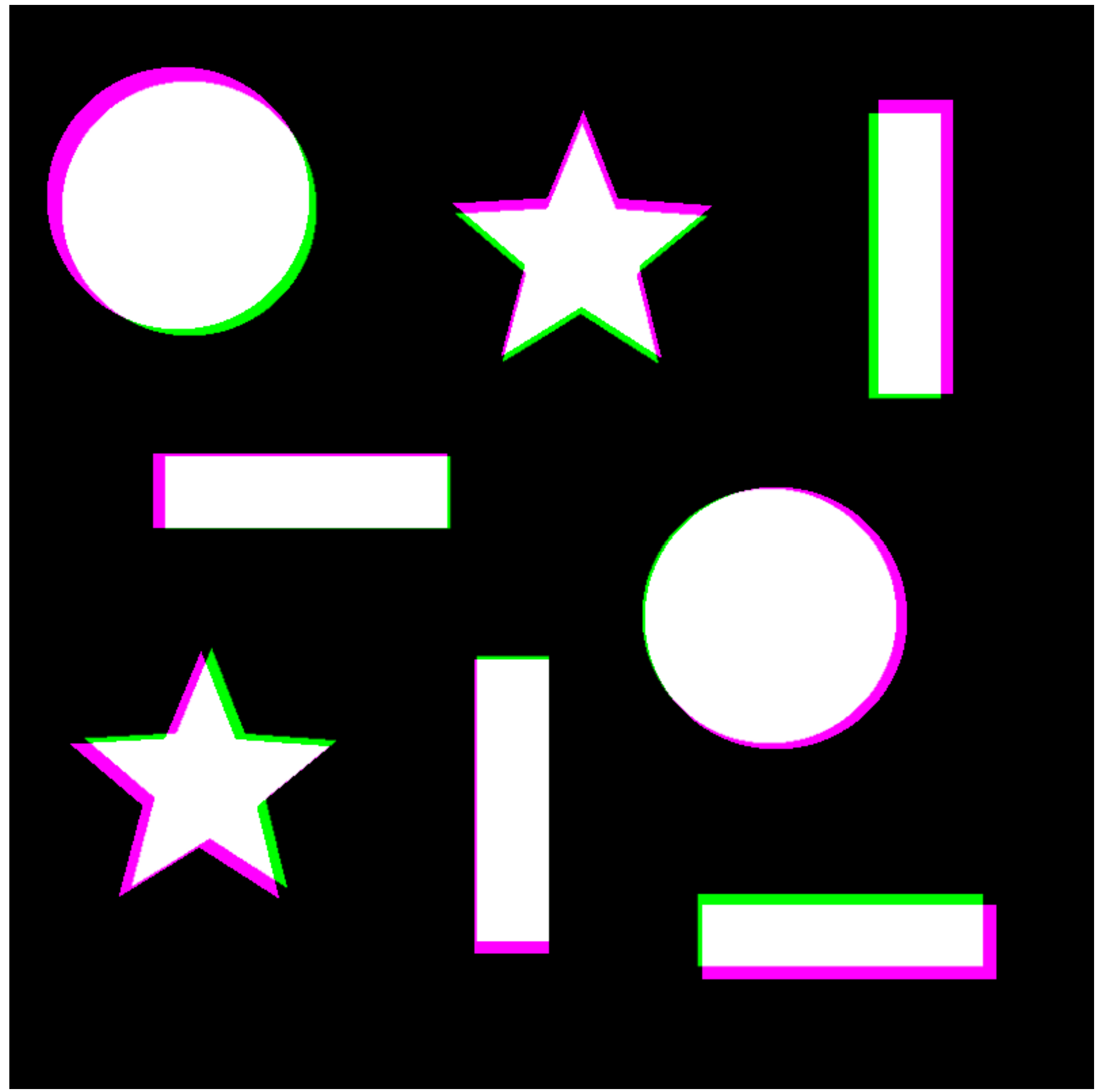


A Not-So “Trivial” Shortcut

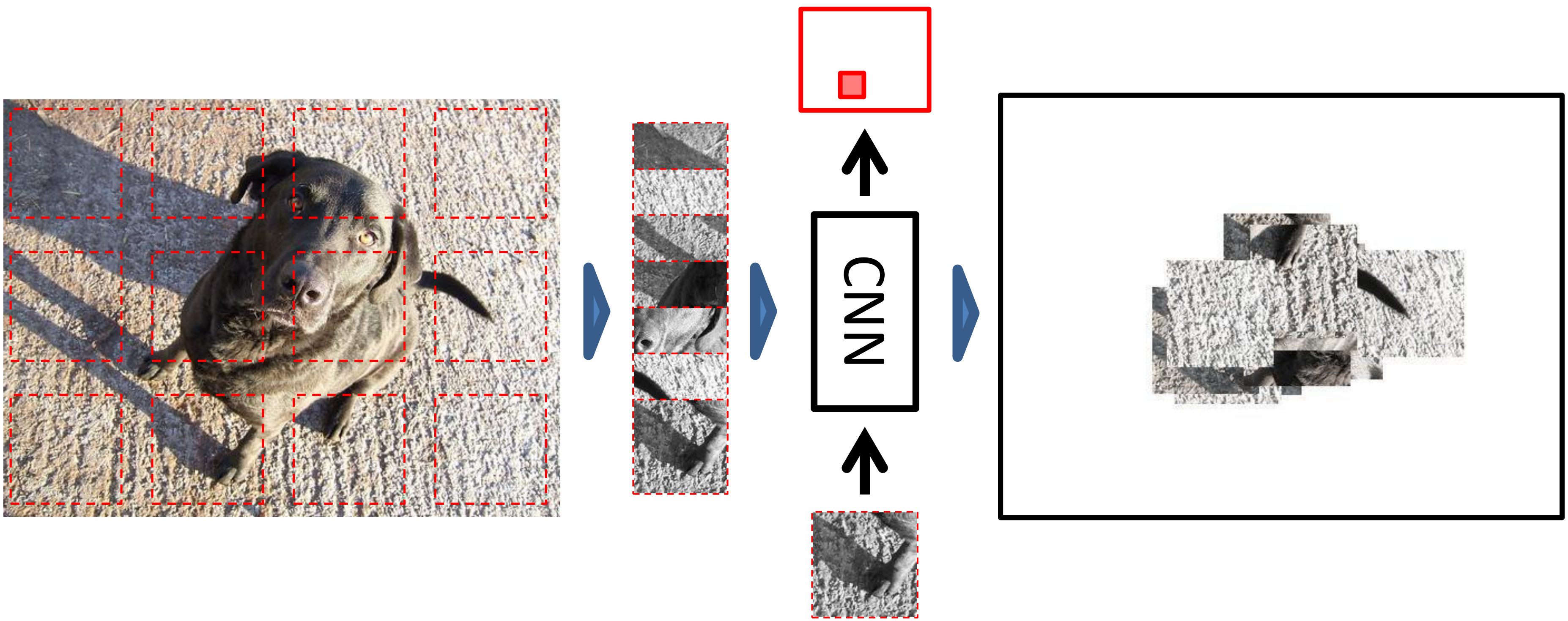


Chromatic Aberration



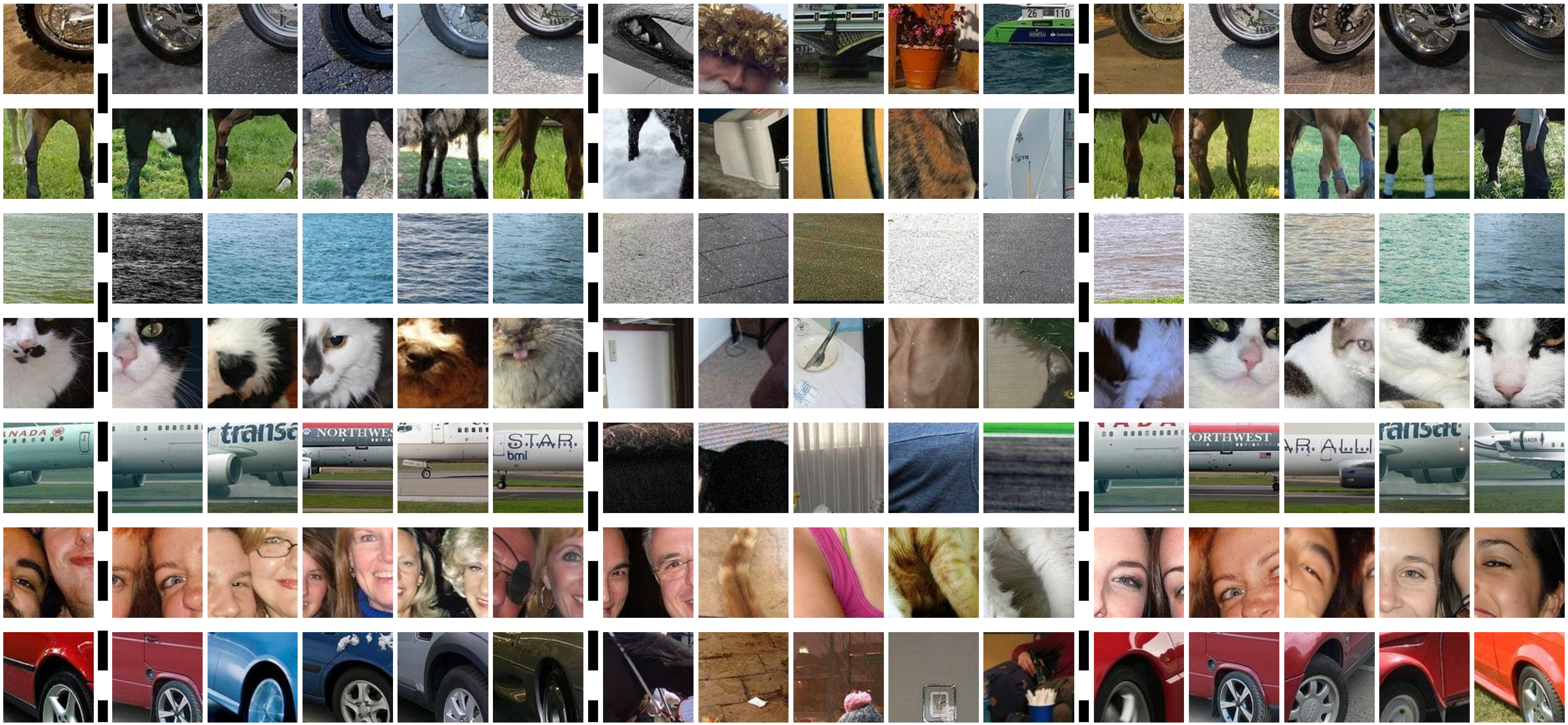


Chromatic Aberration



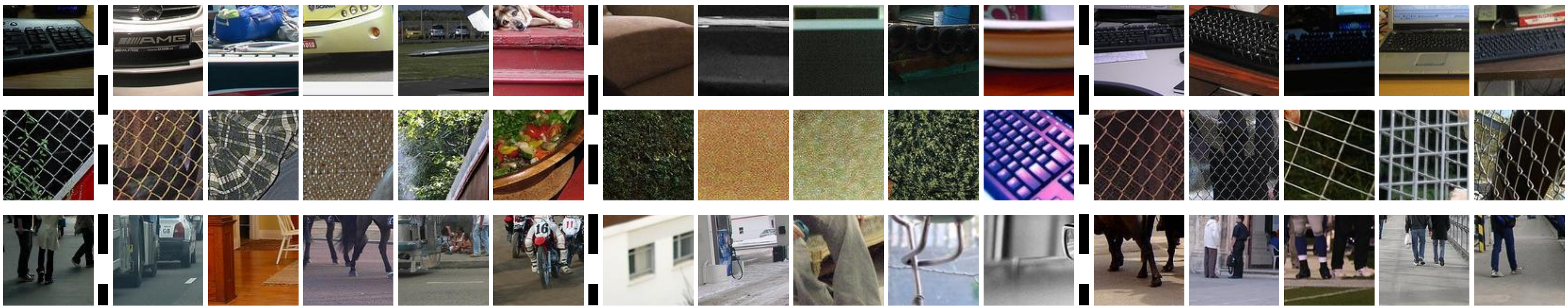
What is learned?

Input Ours Random Initialization ImageNet AlexNet



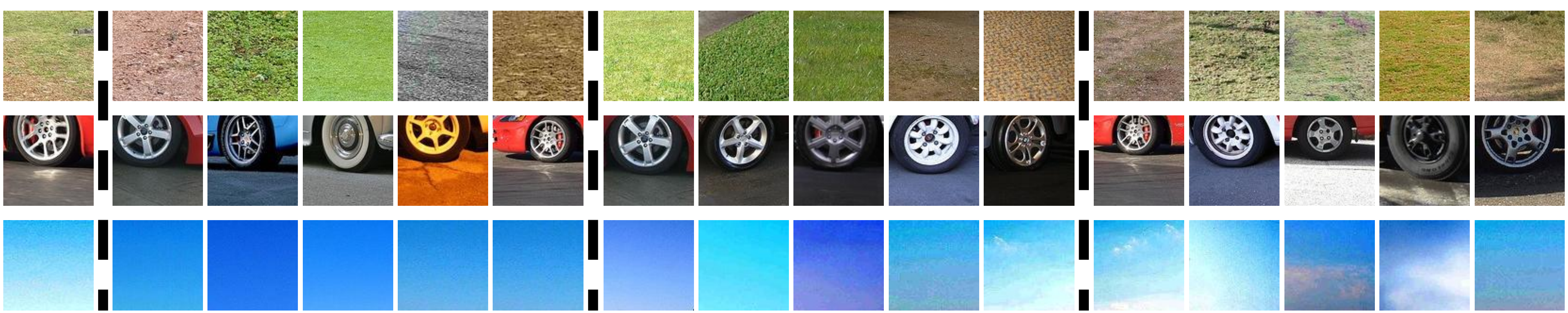
Still don't capture everything

Input | Ours | Random Initialization | ImageNet AlexNet

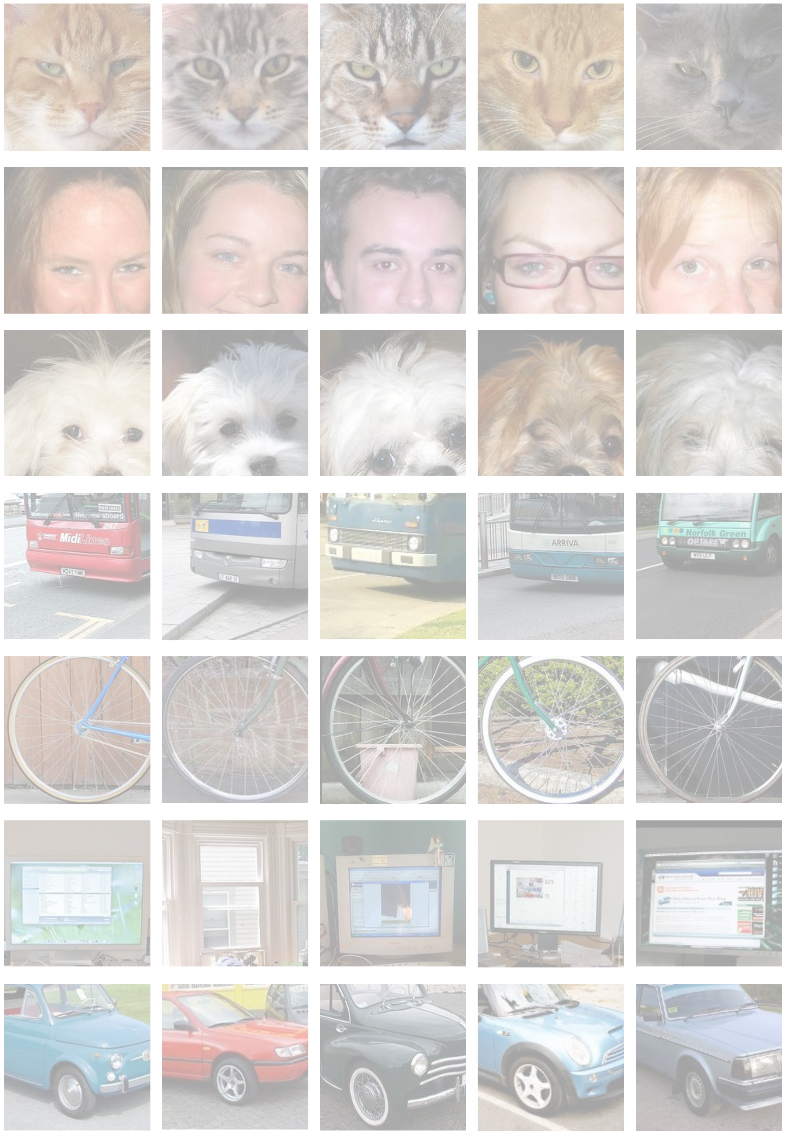


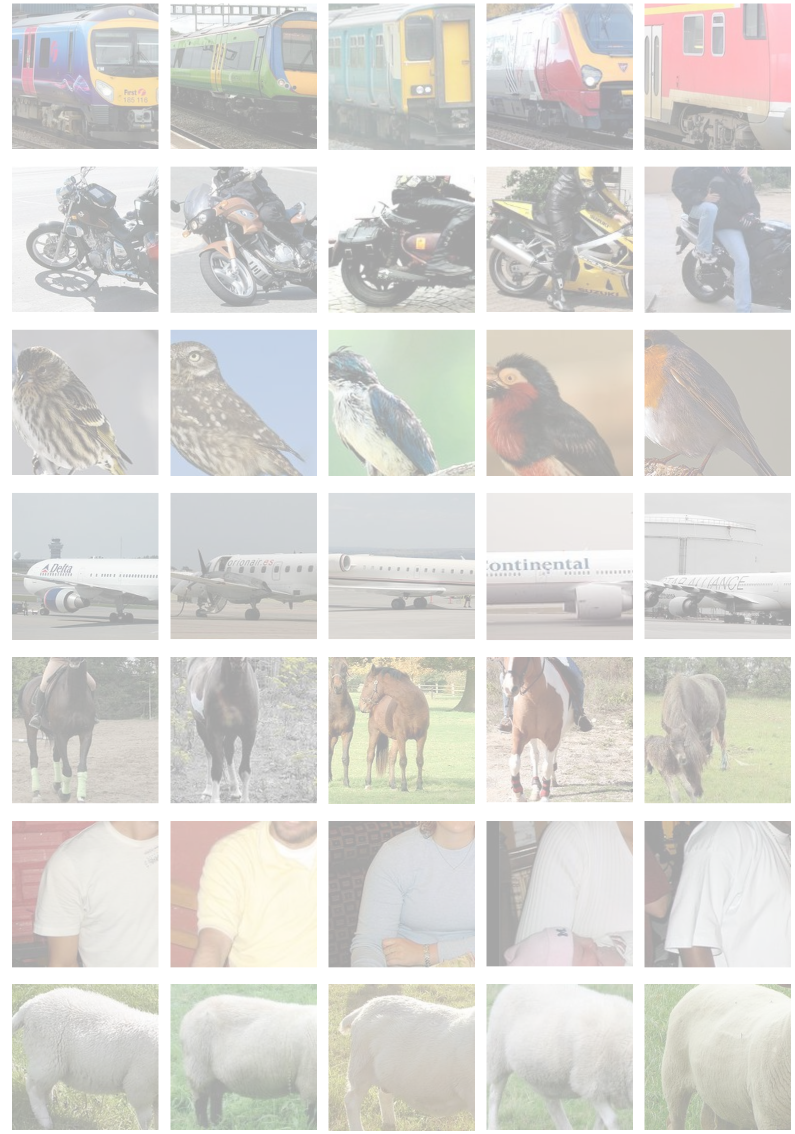
You don't always need to learn!

Input | Ours | Random Initialization | ImageNet AlexNet

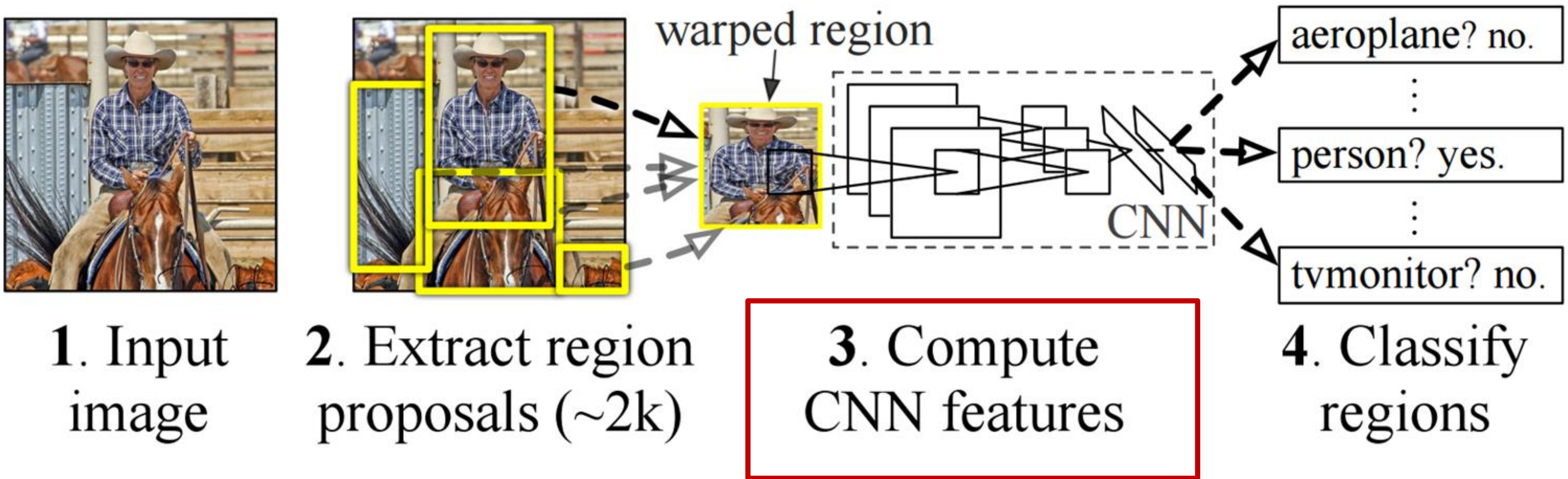


Mined from Pascal VOC2011





Pre-Training for R-CNN

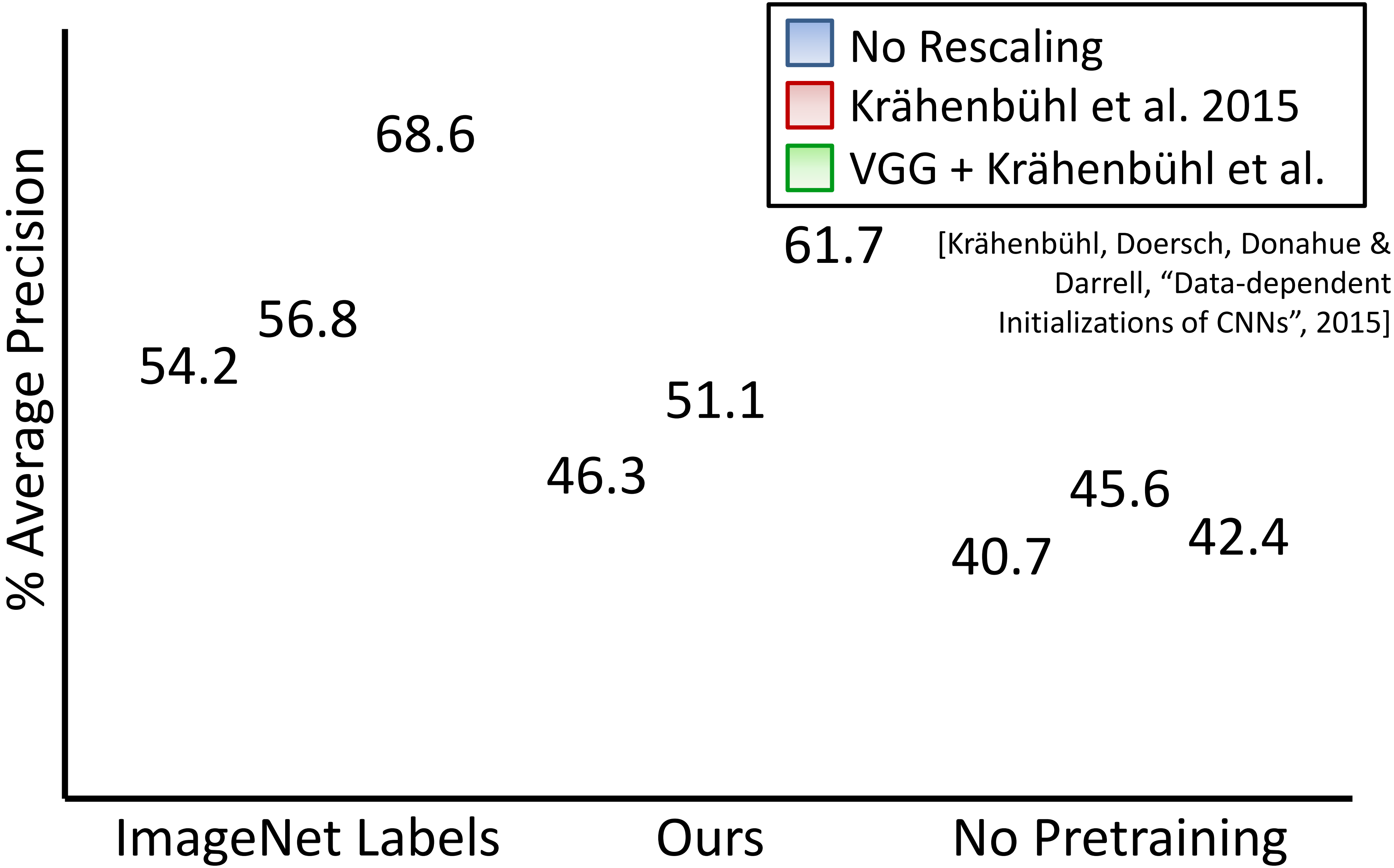


Pre-train on relative-position task, w/o labels

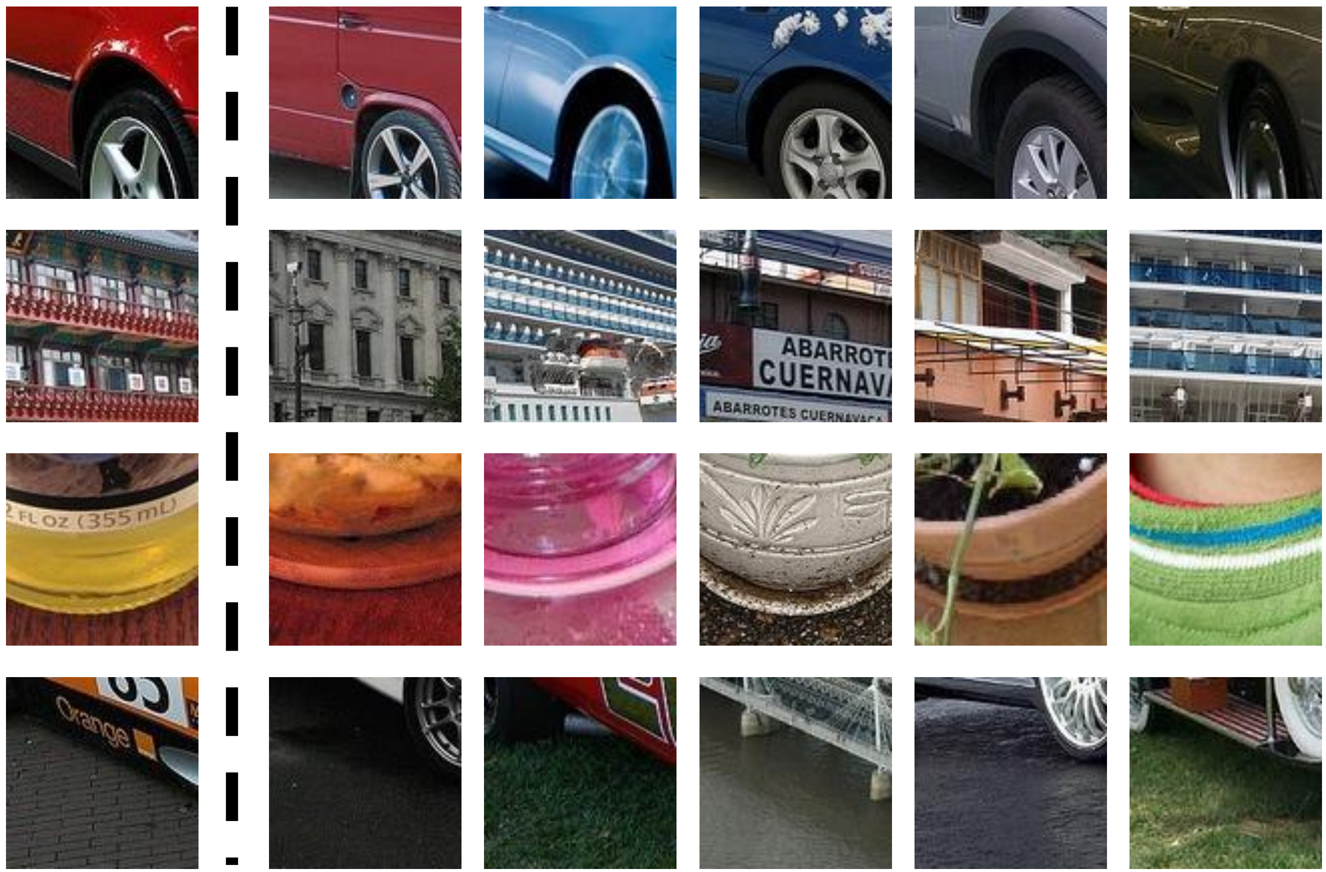
[Girshick et al. 2014]

VOC 2007 Performance

(pretraining for R-CNN)



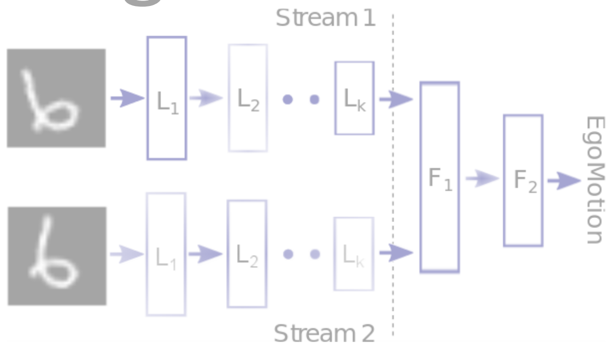
Capturing Geometry?



So, do we need semantic labels?

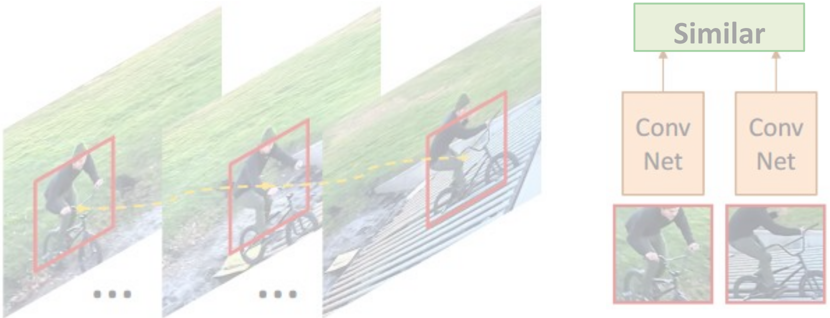
“Self-Supervision” and the Future

Ego-Motion



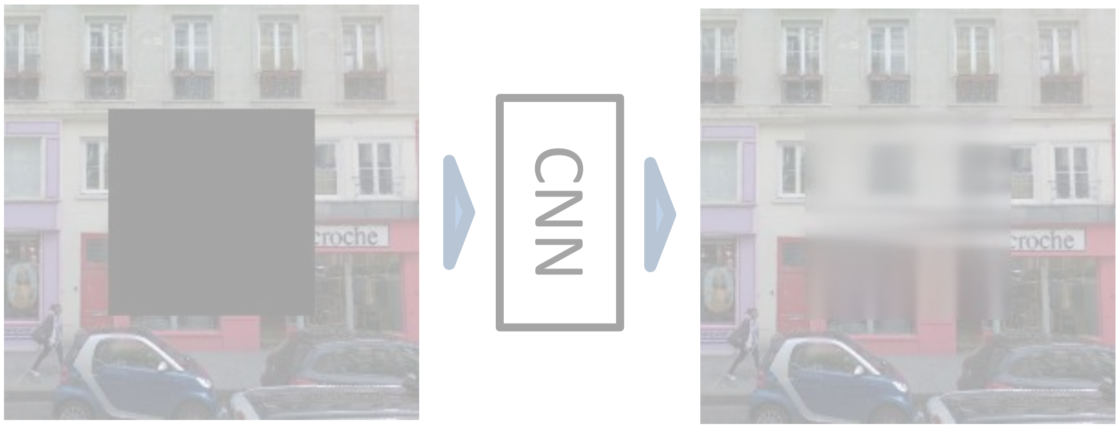
[Agrawal et al. 2015; Jayaraman et al. 2015]

Video

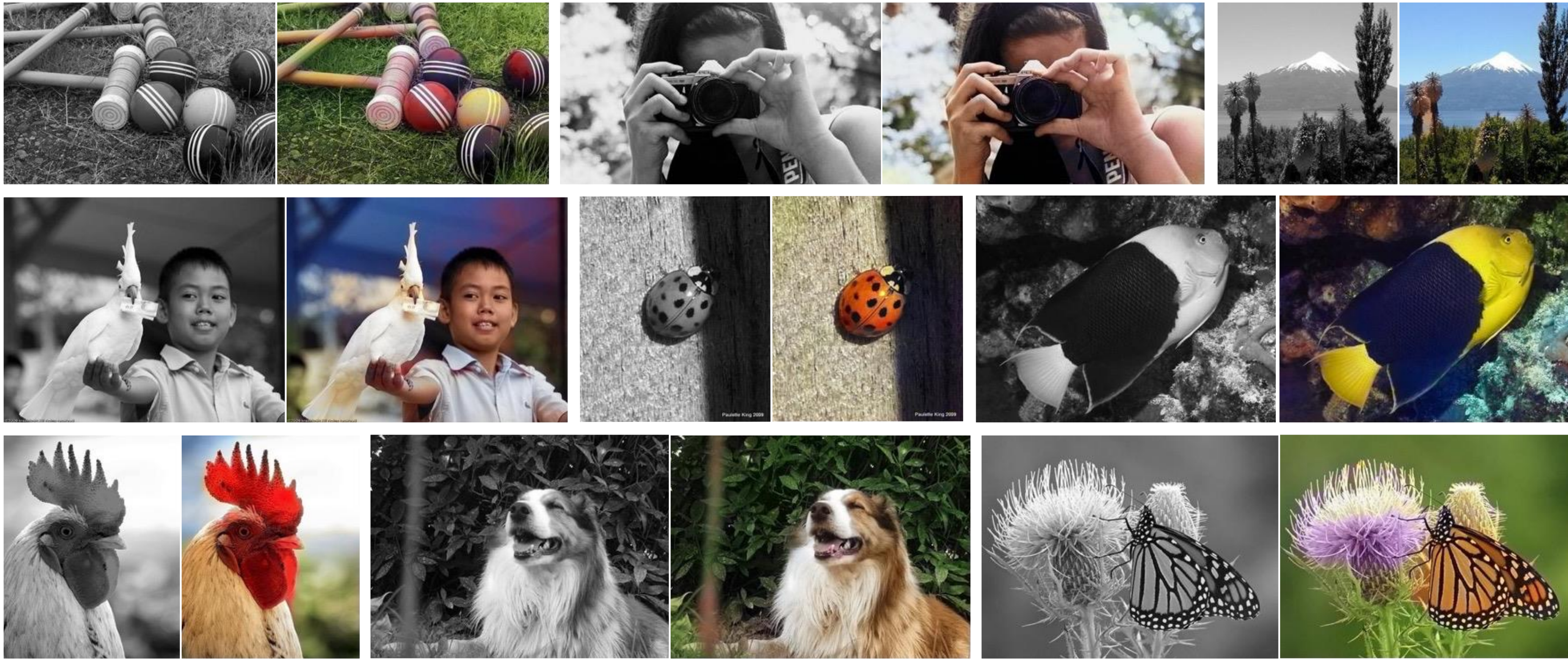


[Wang et al. 2015; Srivastava et al 2015; …]

Context



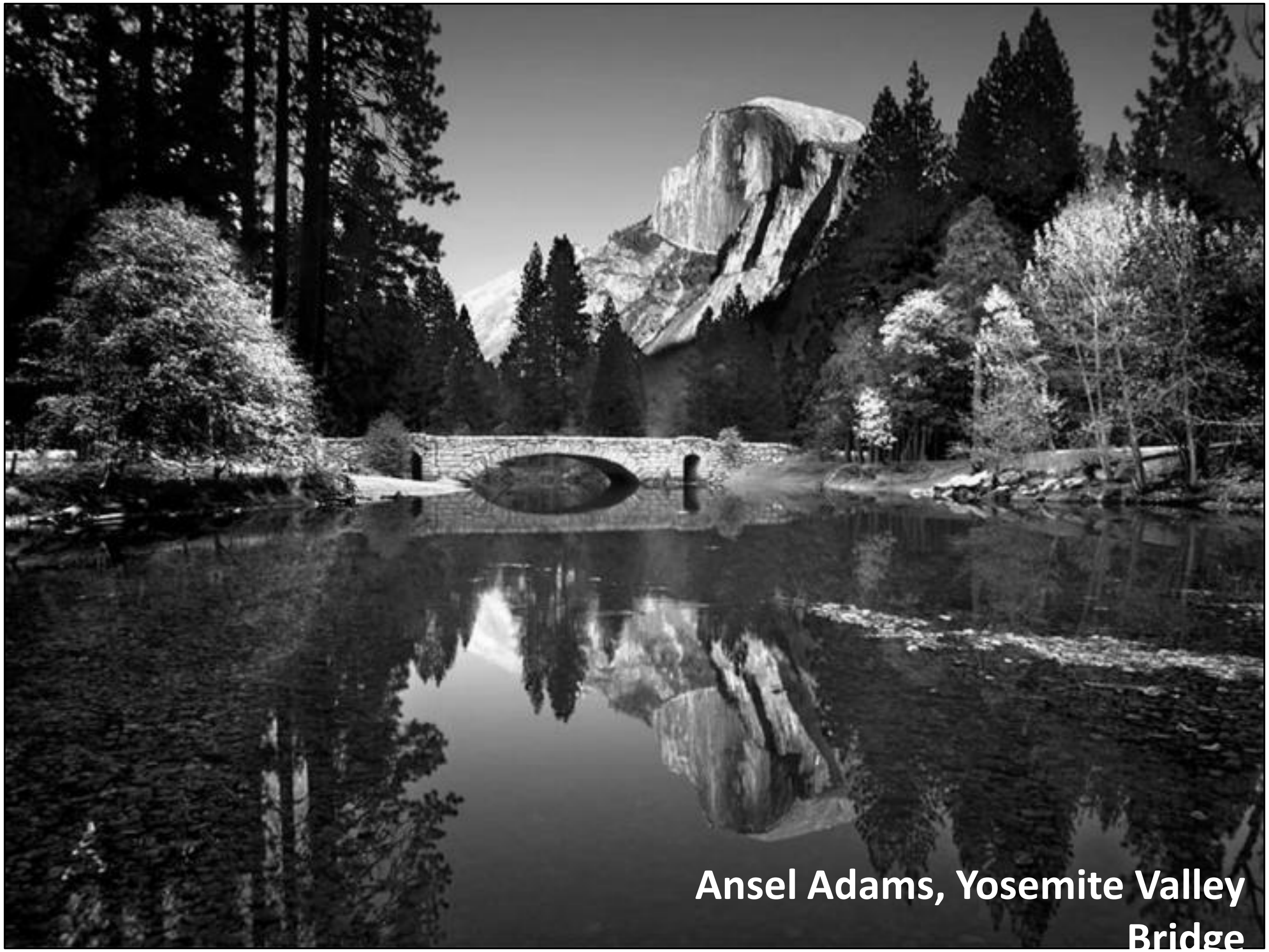
[Doersch et al. 2014; Pathak et al. 2015; Isola et al. 2015]



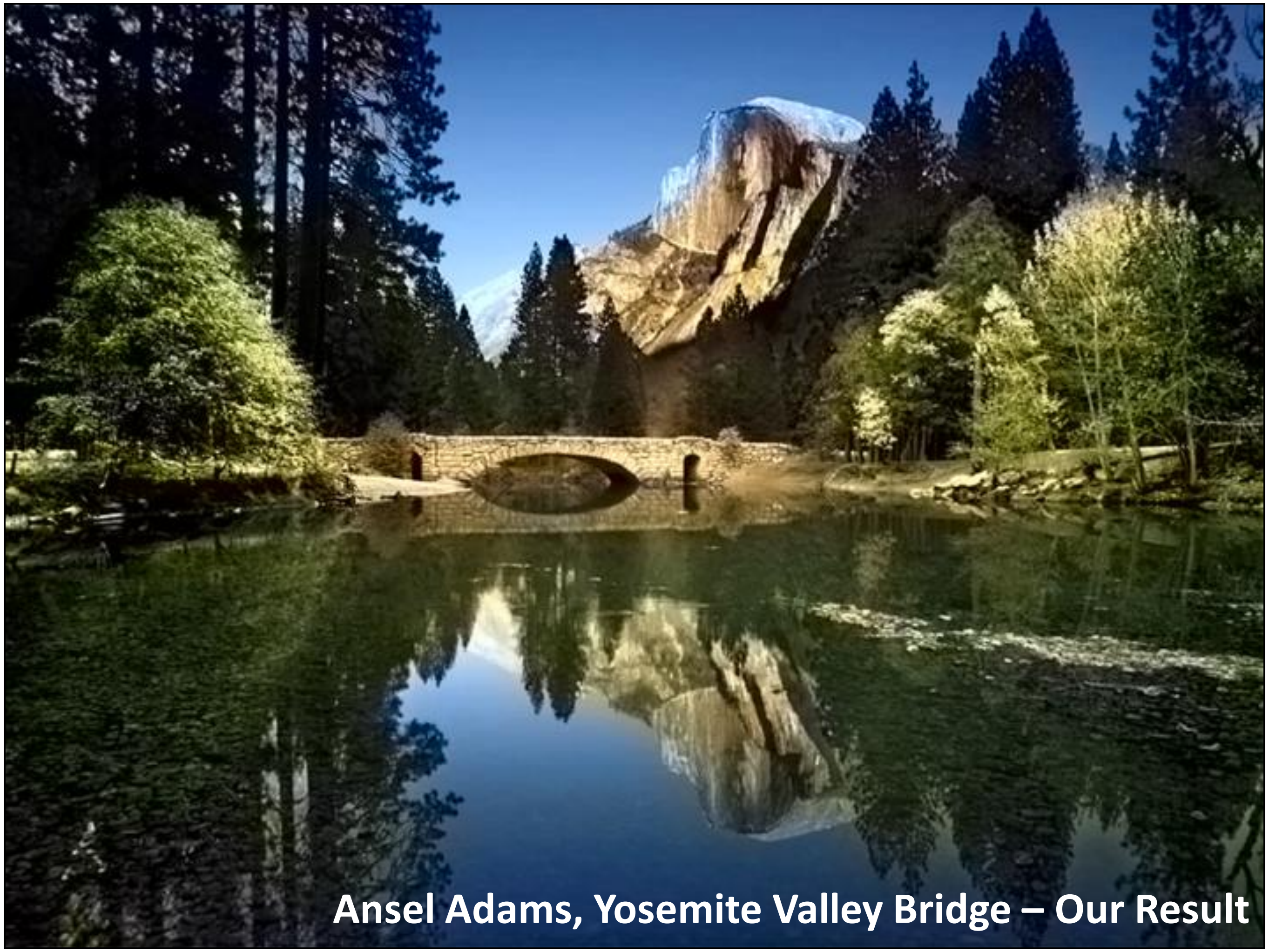
Colorful Image Colorization

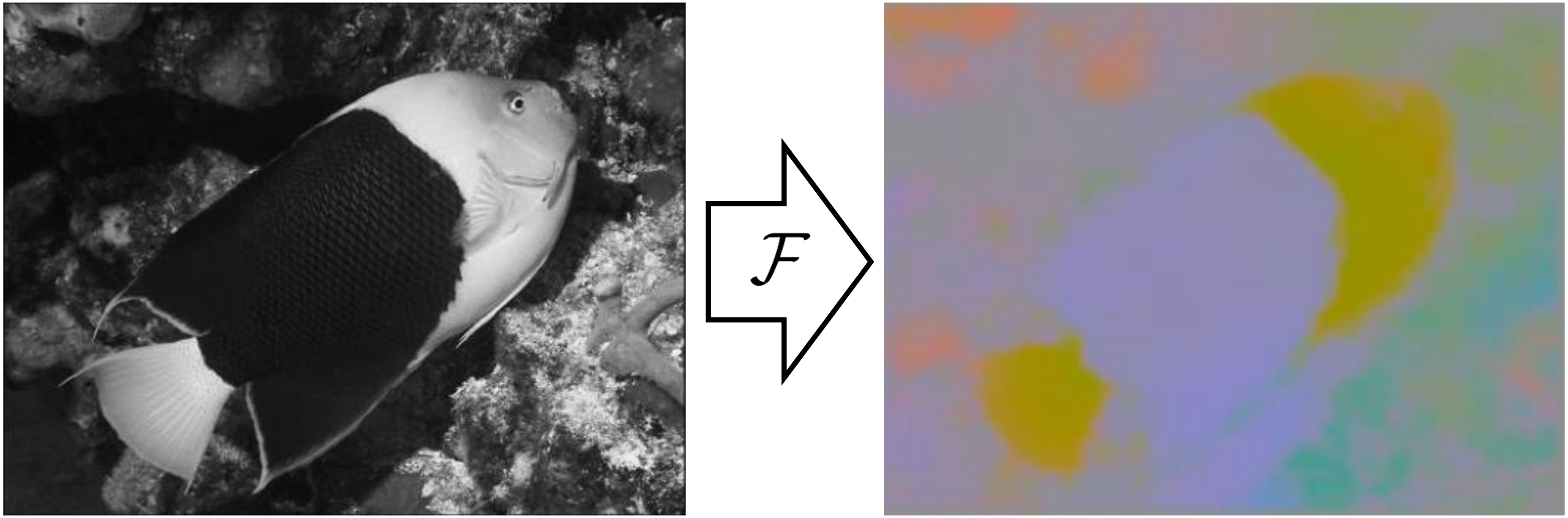
Richard Zhang, Phillip Isola, Alexei (Alyosha) Efros

richzhang.github.io/colorization



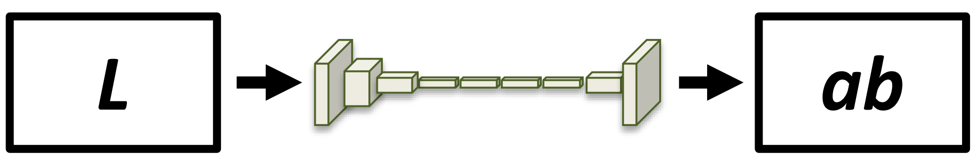
Ansel Adams, Yosemite Valley Bridge

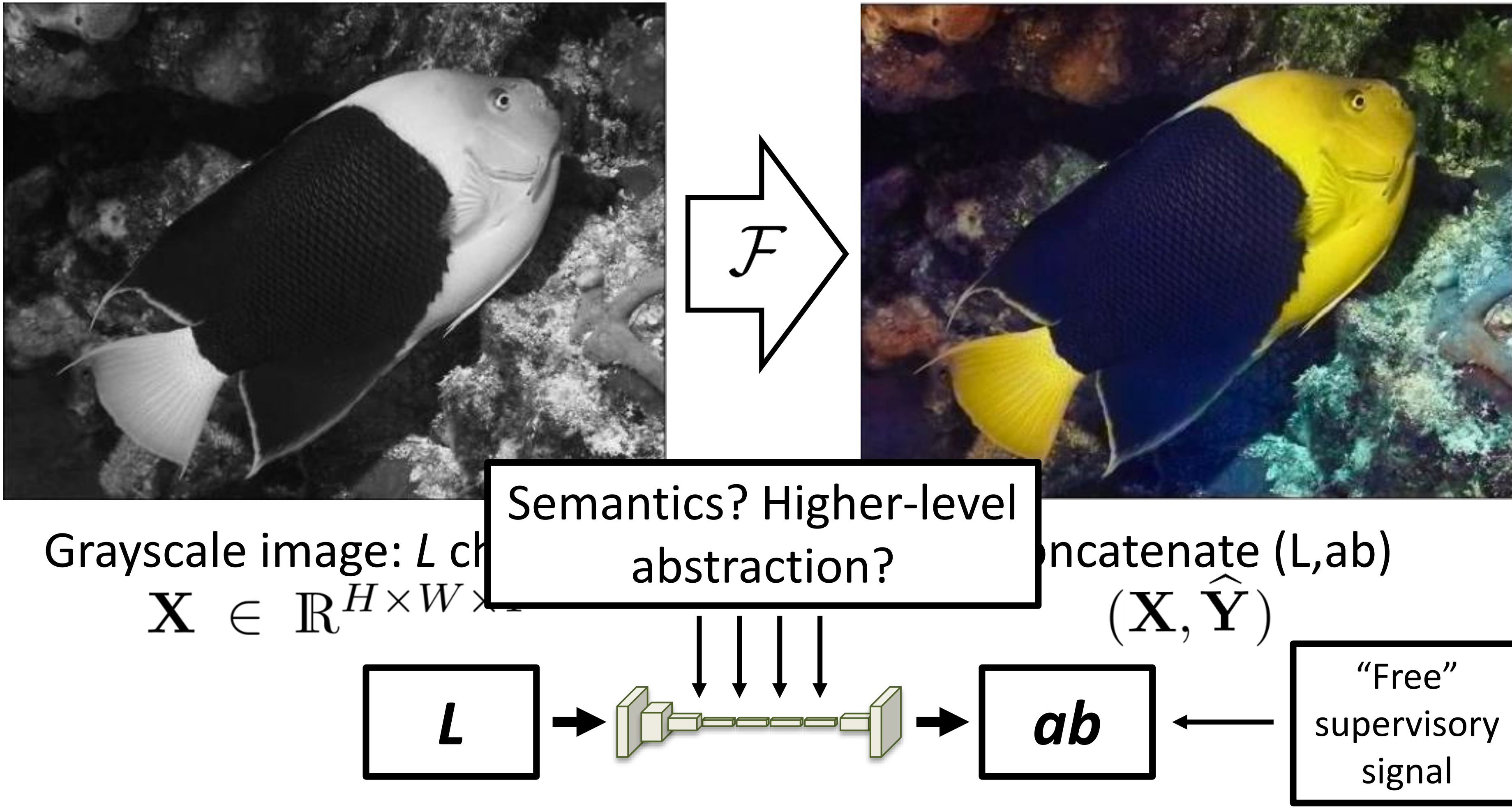

Ansel Adams, Yosemite Valley Bridge – Our Result



Grayscale image: *L* channel
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Color information: *ab* channels
$\widehat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$





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

Inherent Ambiguity

Grayscale

Inherent Ambiguity

Our Output

Ground Truth

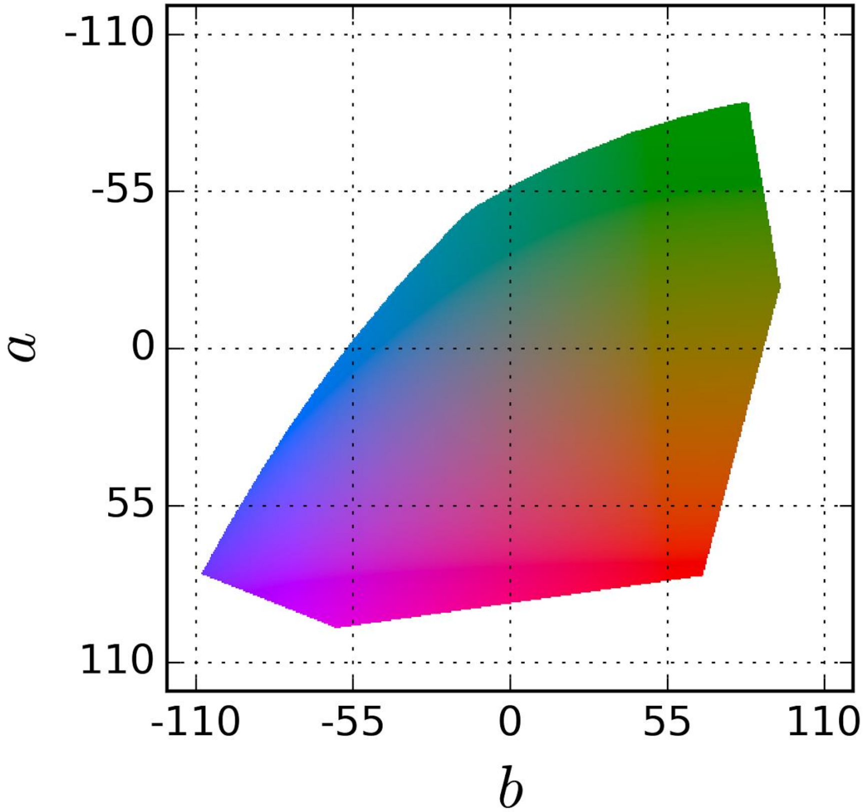
Better Loss Function

- 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

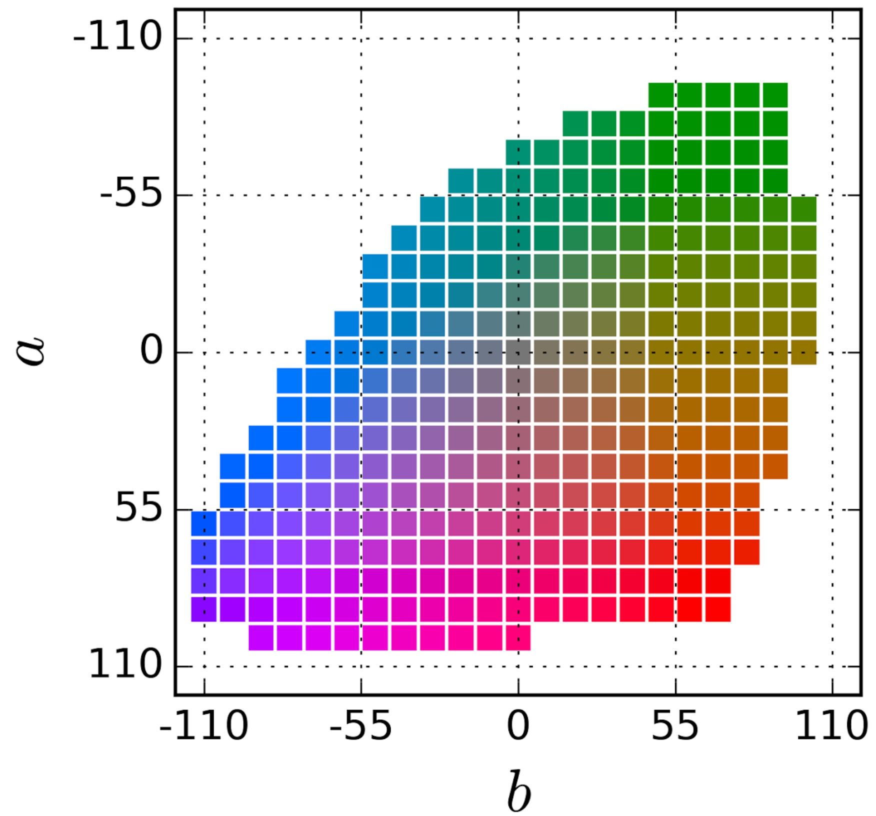
- Regression with L2 loss inadequate

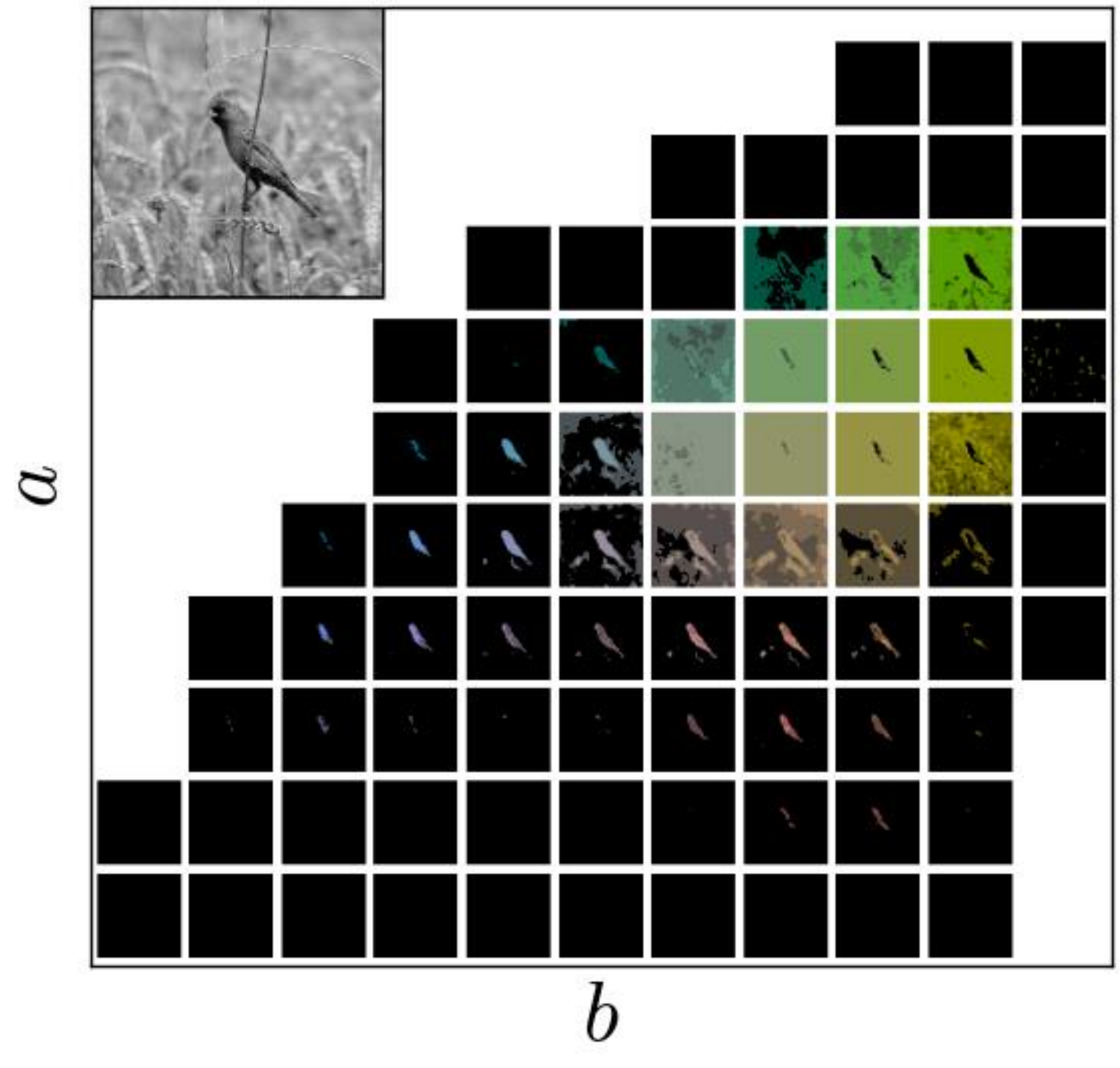
$$\mathrm{L}_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2} \sum_{h,w} \|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$

- 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

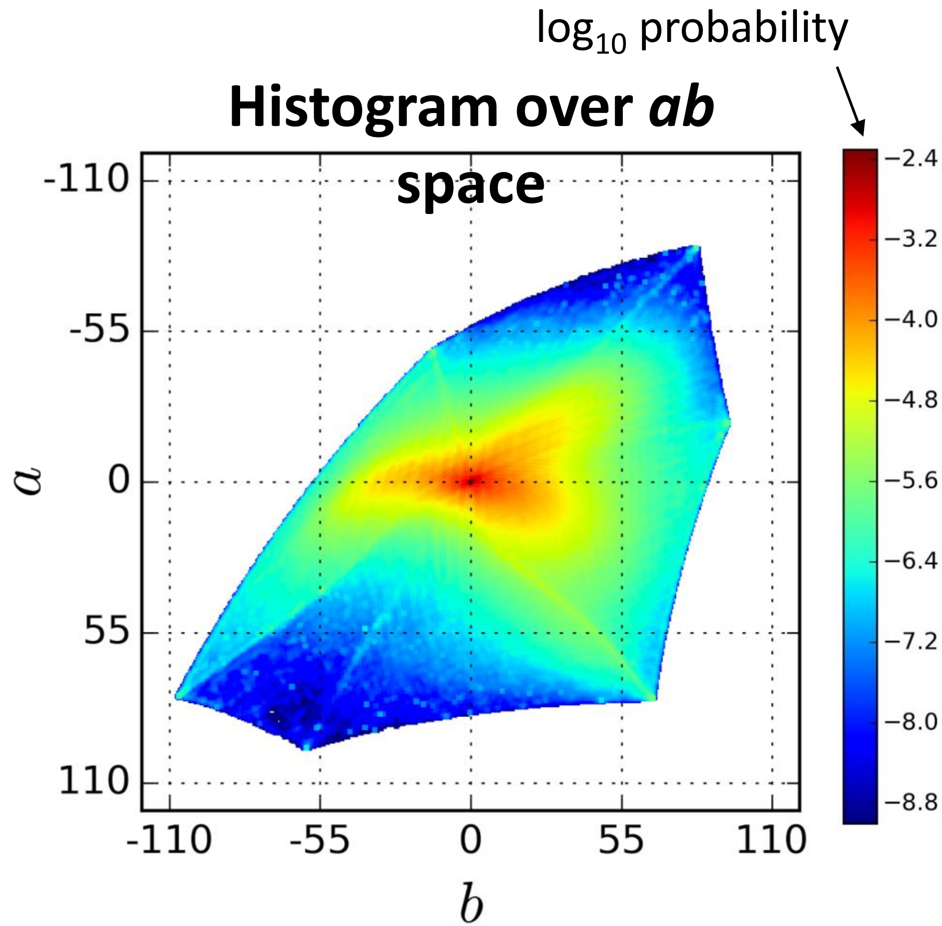
Better Loss Function

- Regression with L2 loss inadequate

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

- Use **multinomial classification**

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



Better Loss Function

- Regression with L2 loss inadequate

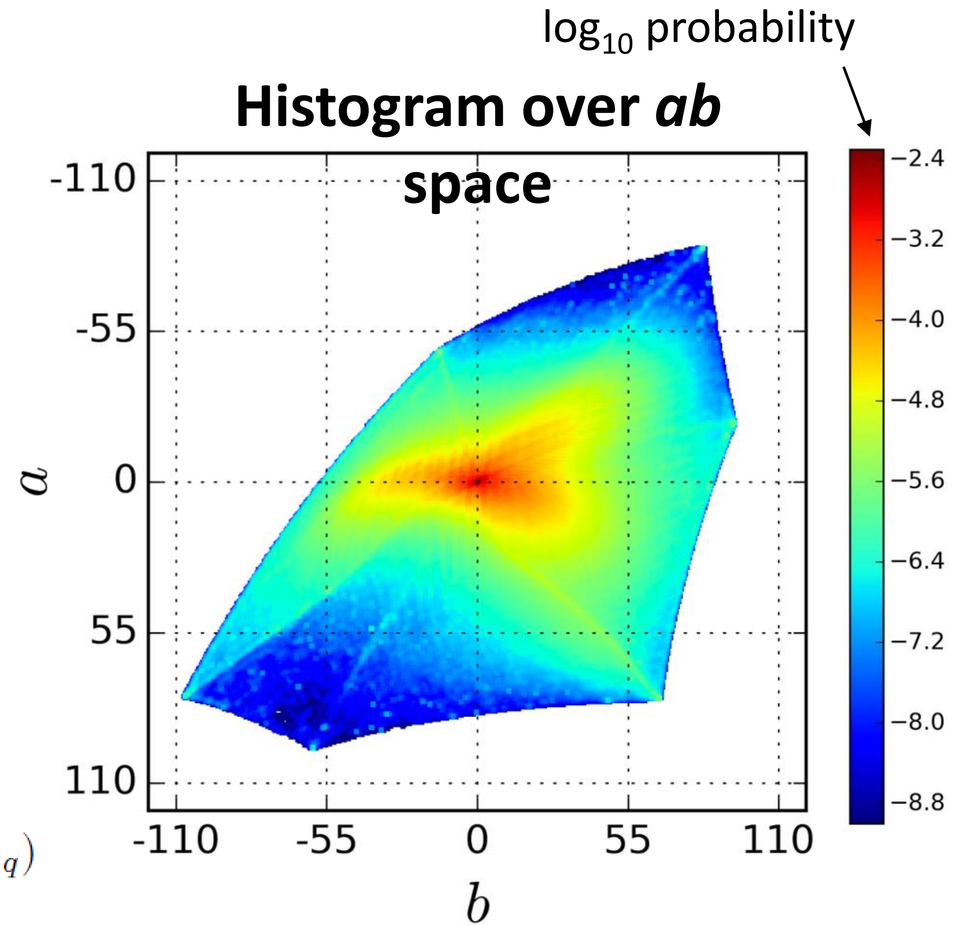
$$L_2(\widehat{\mathbf{Y}}, \mathbf{Y}) = \frac{1}{2}\sum_{h,w}\|\mathbf{Y}_{h,w} - \widehat{\mathbf{Y}}_{h,w}\|_2^2$$

- Use **multinomial classification**

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

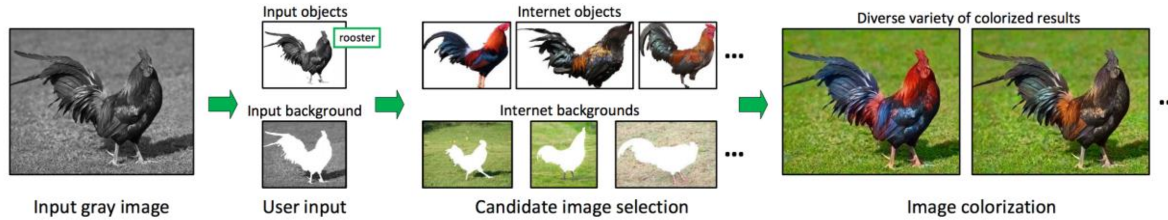
- **Class rebalancing** to encourage learning of *rare* colors

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



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.

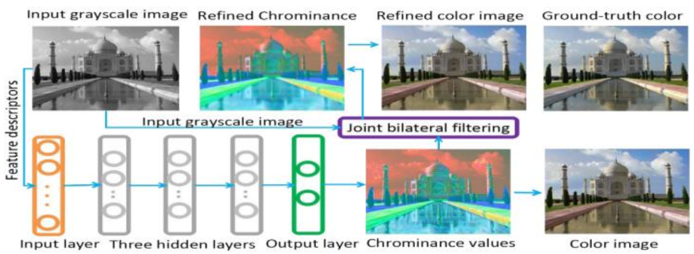


Parametric

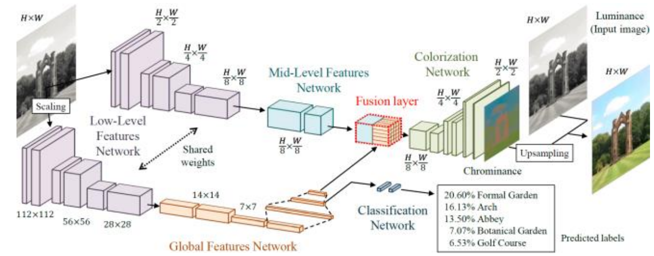
Hand-engineered Features

Deep Networks

L2 Regression

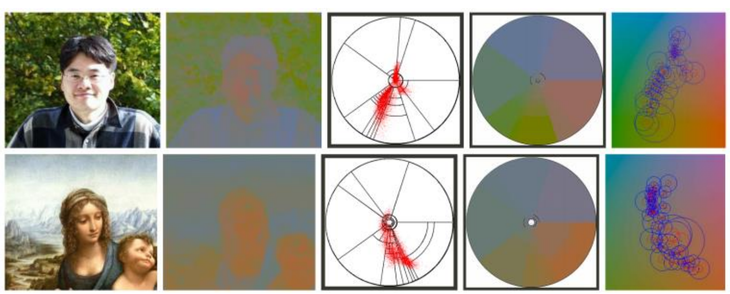


Deshpande et al. Cheng et al. In ICCV 2015.

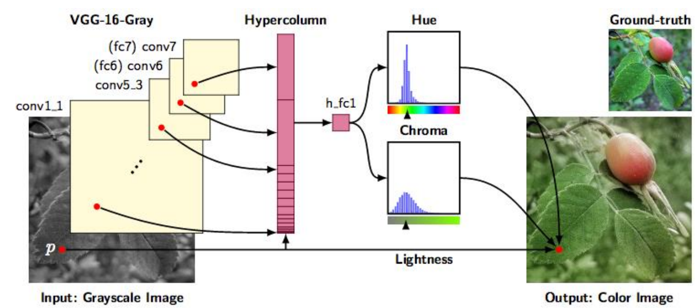


Dahl. Jan 2016. Iizuka et al. In SIGGRAPH, 2016.

Classification

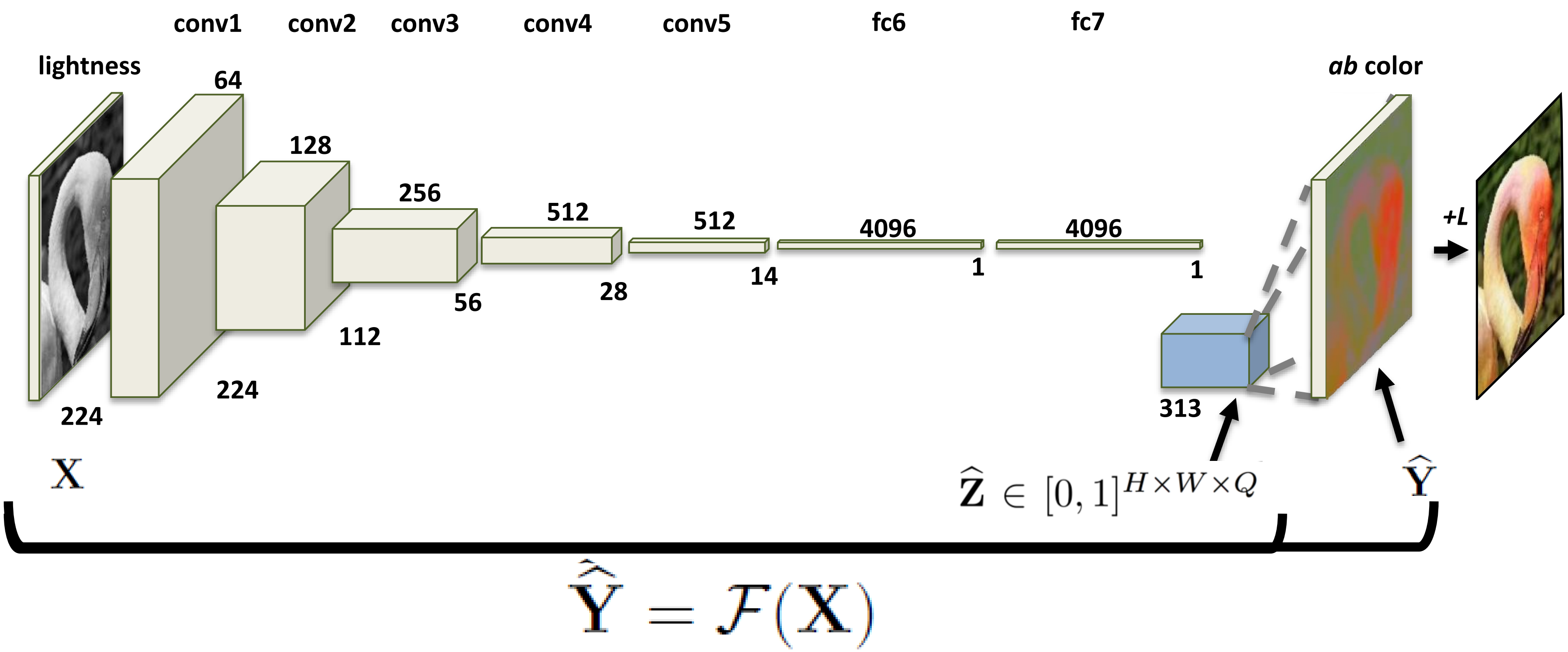


Charpiat et al. In ECCV 2008.

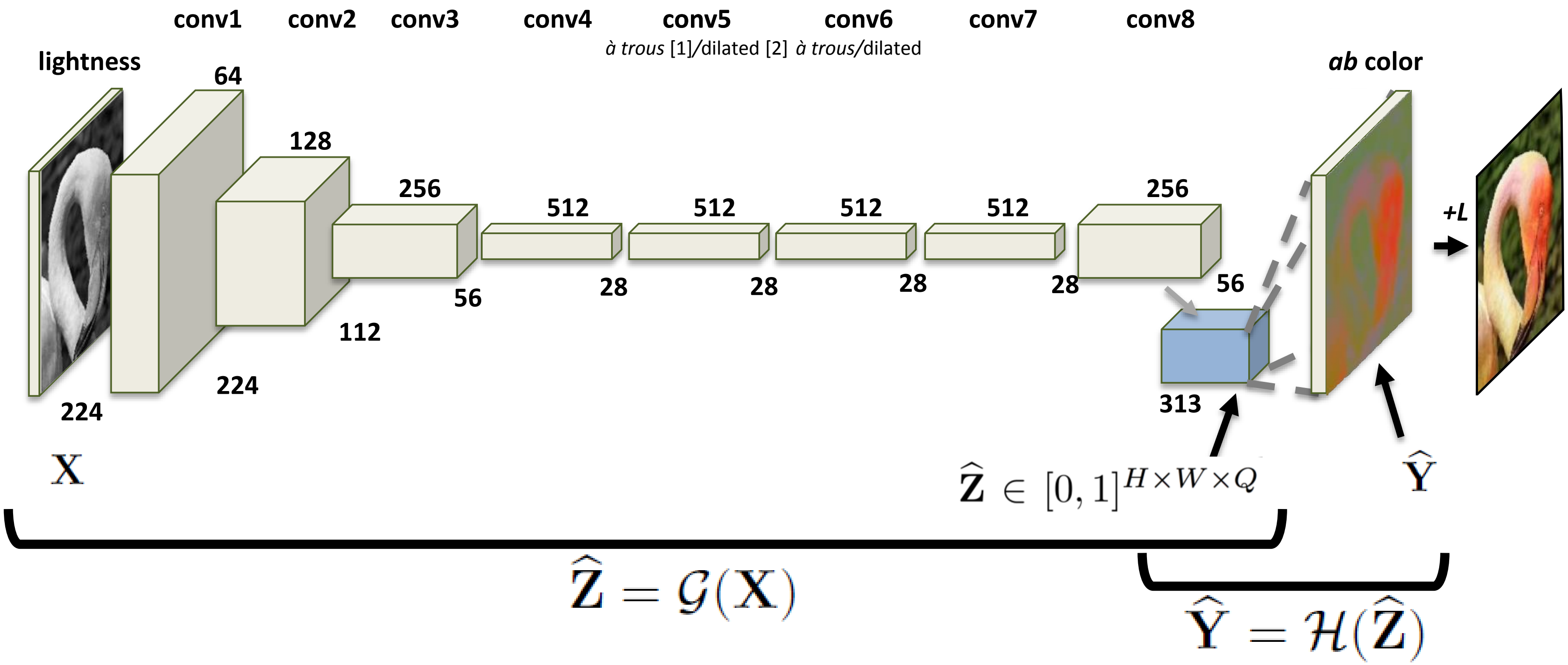


Larsson et al. In ECCV 2016. [Concurrent]

Network Architecture



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

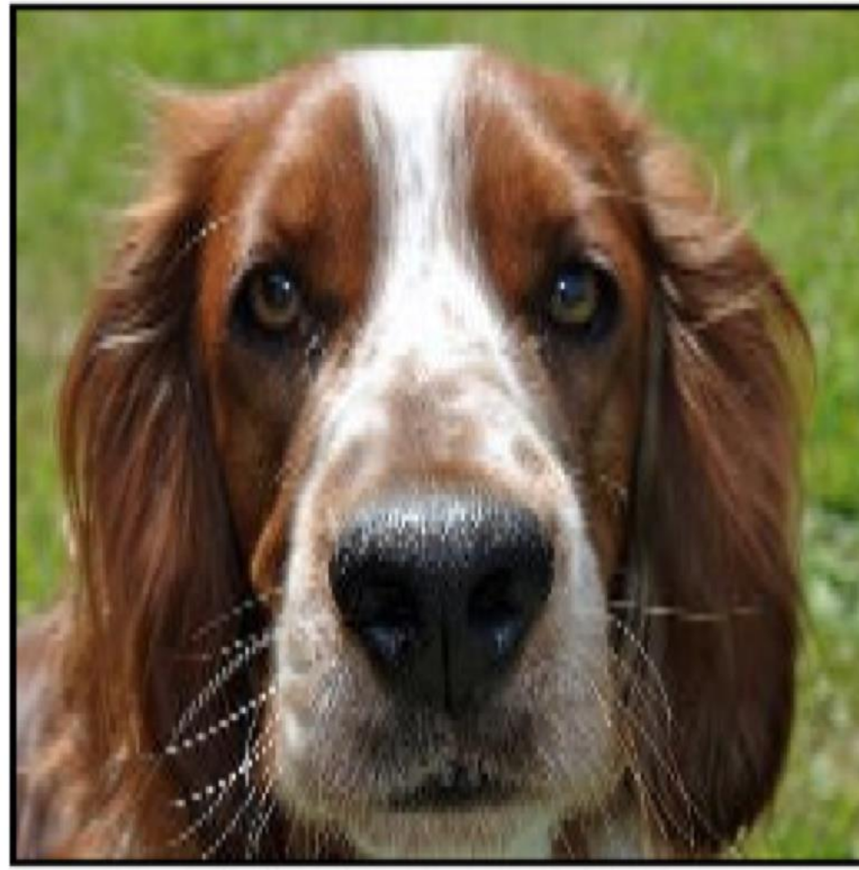
	Visual Quality	Representation Learning
Quantitative	Per-pixel accuracy **Perceptual realism** Semantic interpretability	Task generalization ImageNet classification Task & dataset generalization PASCAL classification, detection, segmentation
Qualitative	Low-level stimuli Legacy grayscale photos	Hidden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

Perceptual Realism / Amazon Mechanical Turk Test

clap if “fake”

clap if “fake”

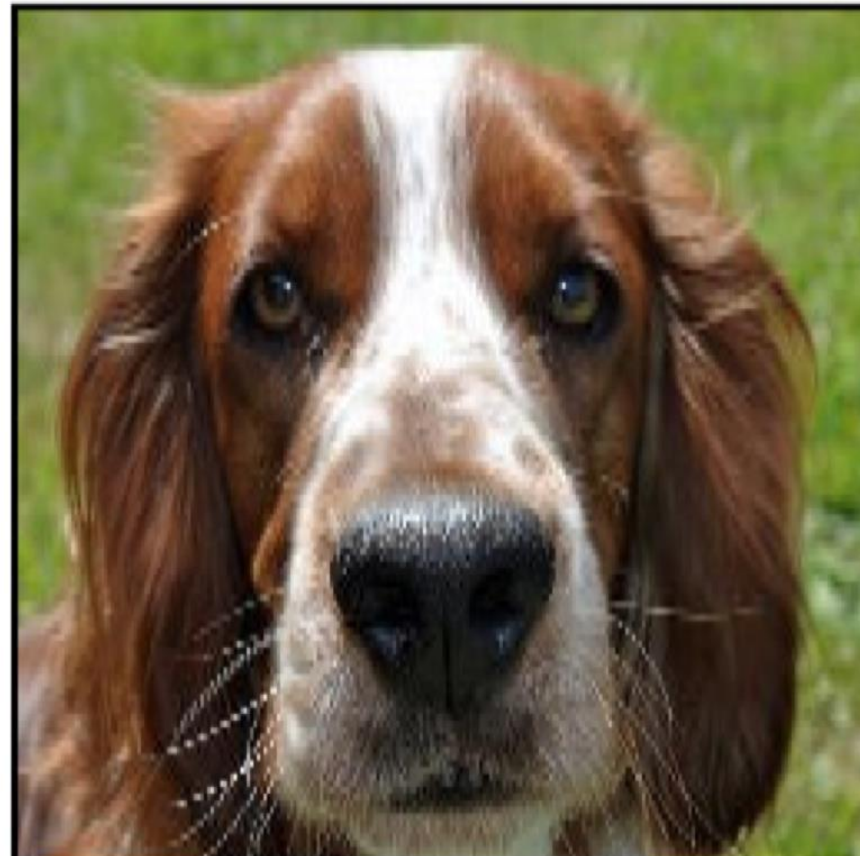
Fake, 0% fooled

clap if “fake”

clap if “fake”

Fake, 55% fooled



clap if “fake”

clap if “fake”

Fake, 58% fooled

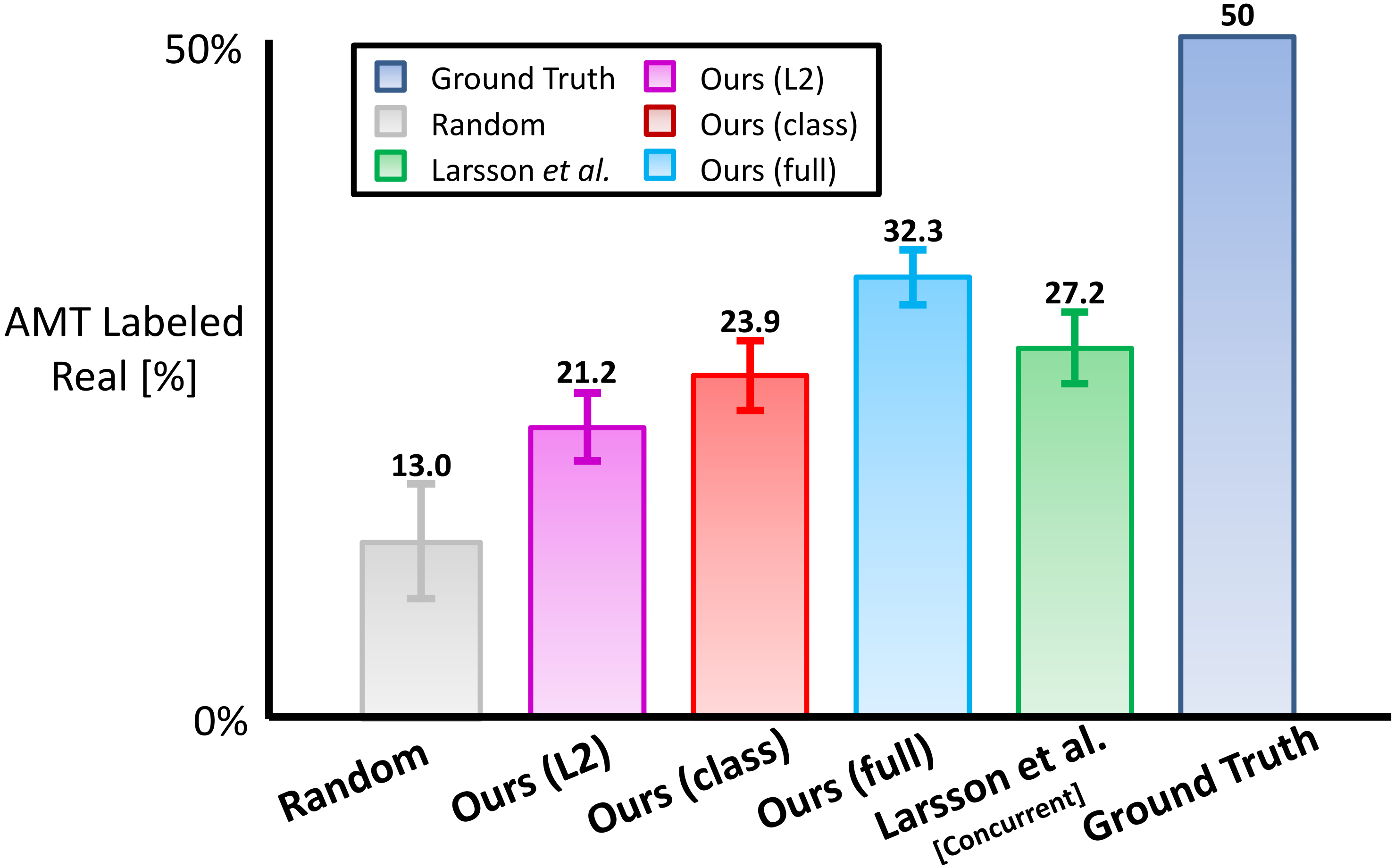
from Reddit /u/SherySantucci

Recolorized by Reddit ColorizeBot

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

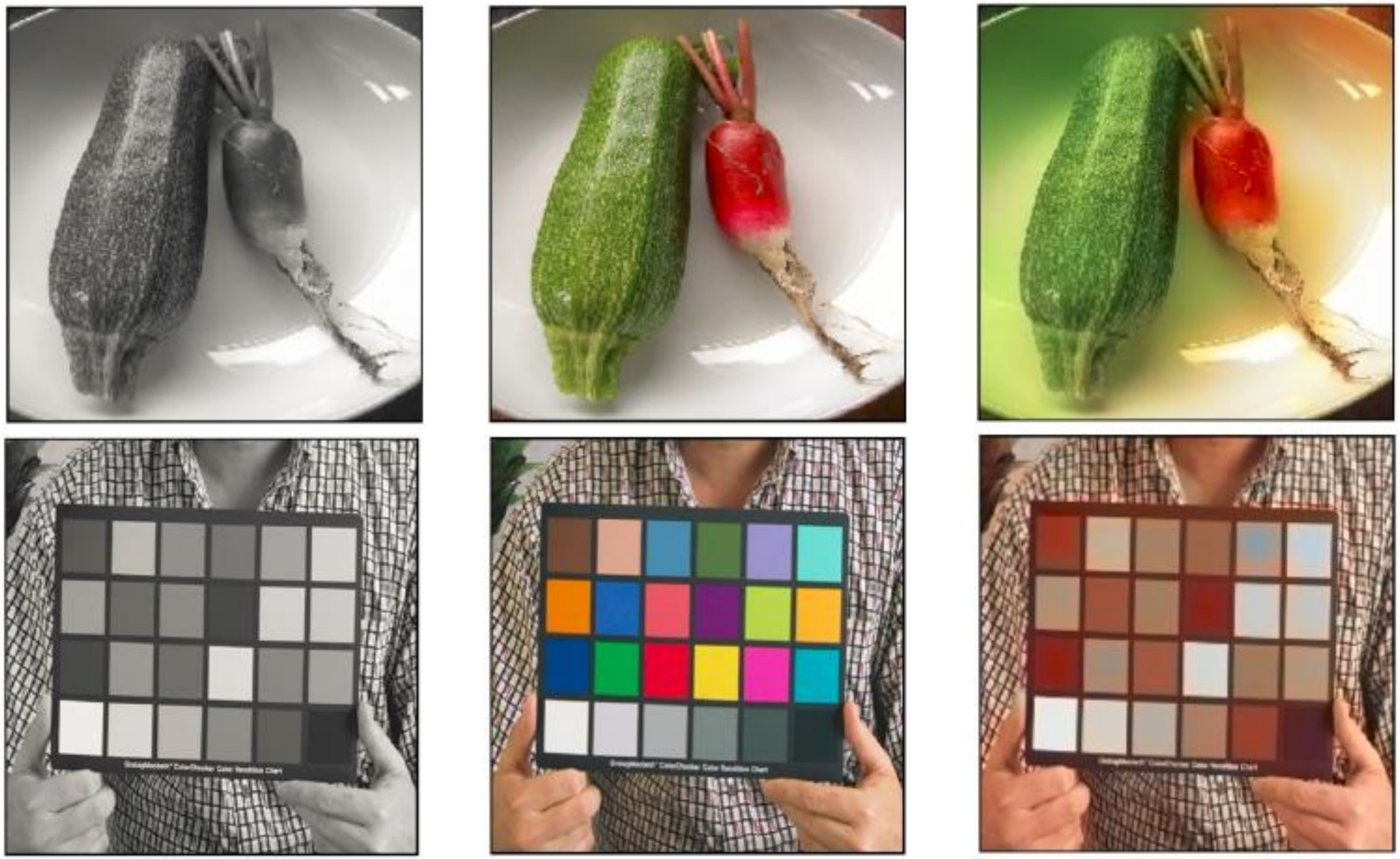
Recolorized by Reddit ColorizeBot

Perceptual Realism Test

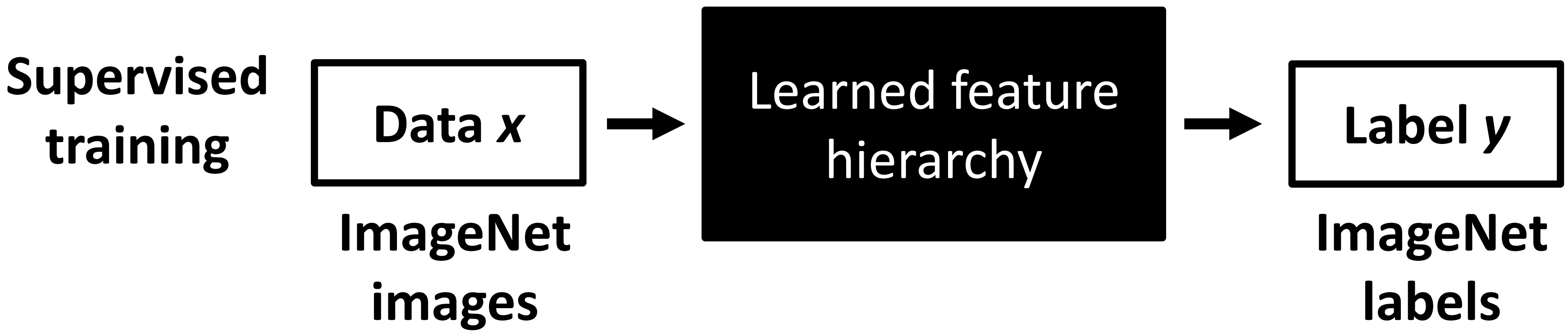


1600 images tested per algorithm

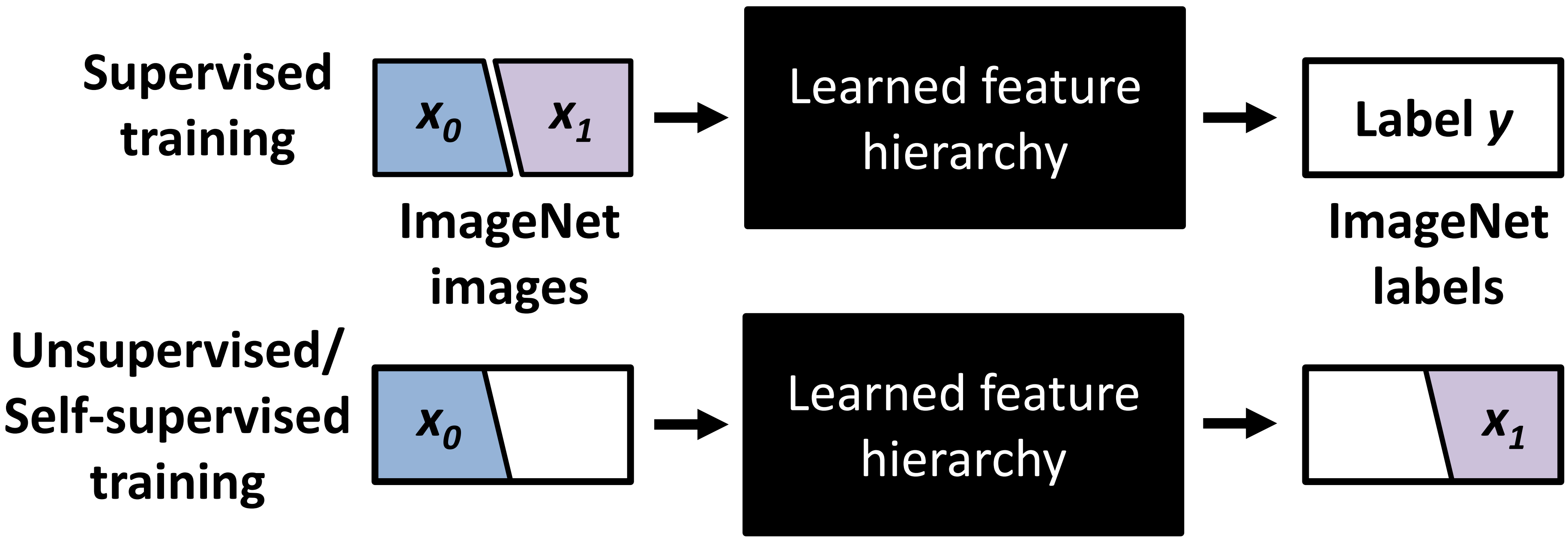
Input **Ground Truth** **Output**



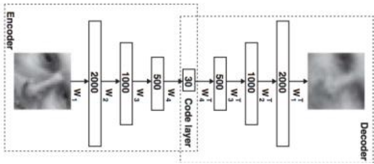
Predicting Labels from Data



Predicting Data from Data

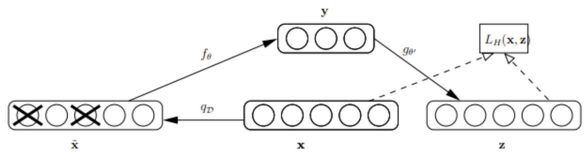


Autoencoders



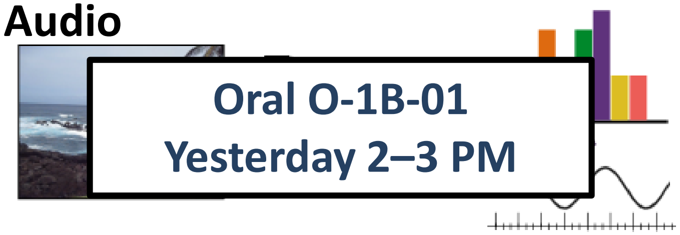
Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders



Vincent *et al.* ICML 2008.

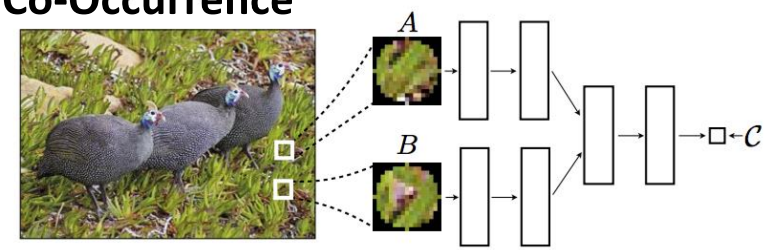
Audio



Oral O-1B-01
Yesterday 2–3 PM

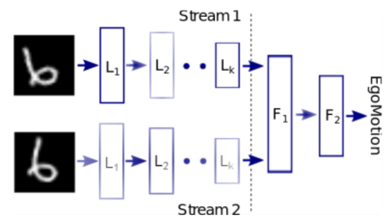
Owens *et al.* CVPR 2016, ECCV 2016

Co-Occurrence



Isola *et al.* ICLR Workshop 2016.

Egomotion



Agrawal *et al.* ICCV 2015

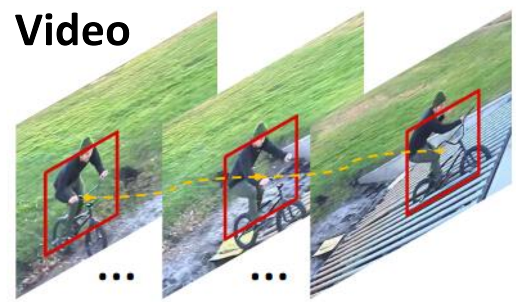
Jayaraman *et al.* ICCV 2015

Context

Doersch *et al.* ICCV 2015

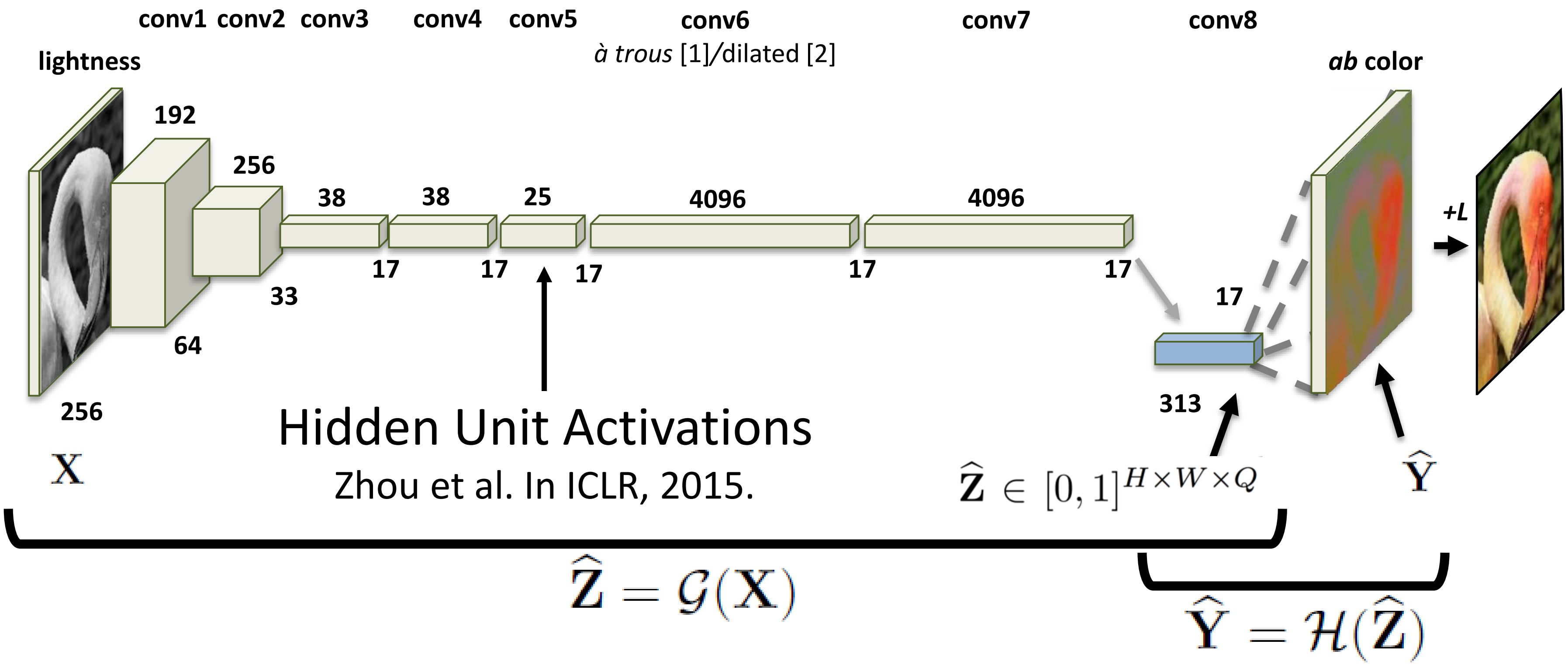
Pathak *et al.* CVPR 2016

Video



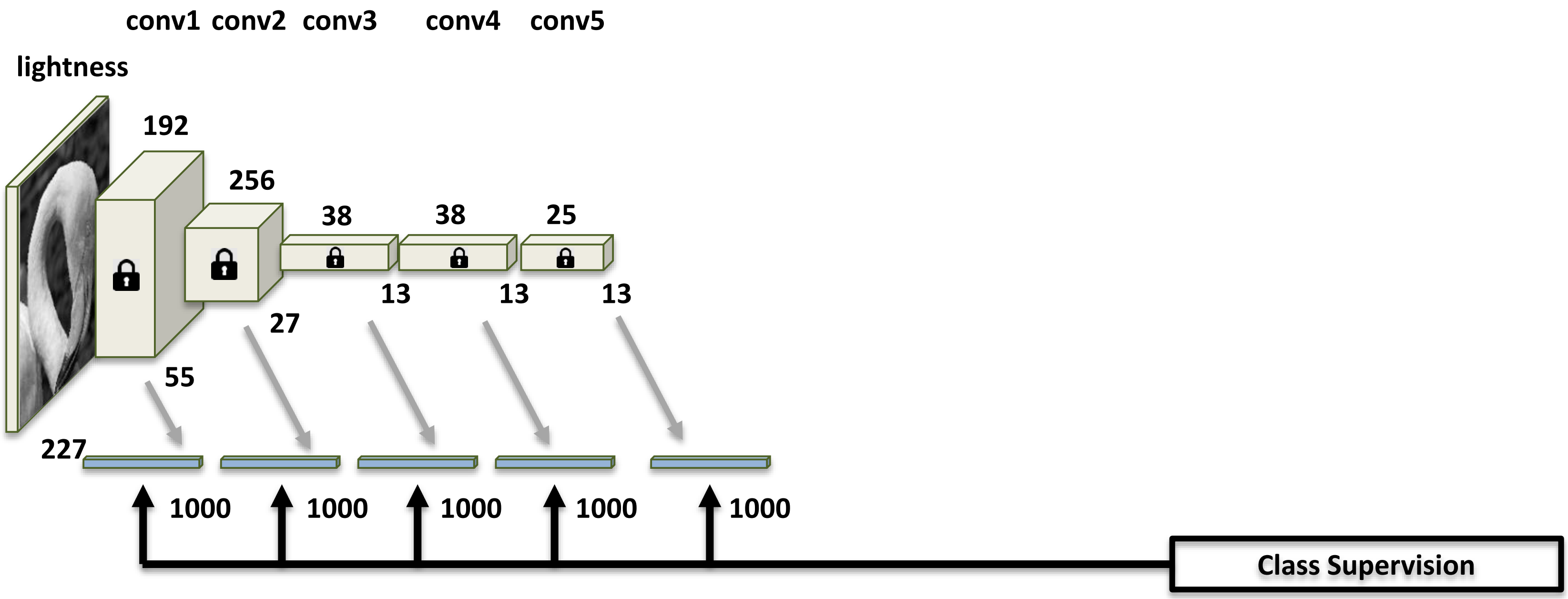
Wang *et al.* ICCV 2015

Cross-Channel Encoder



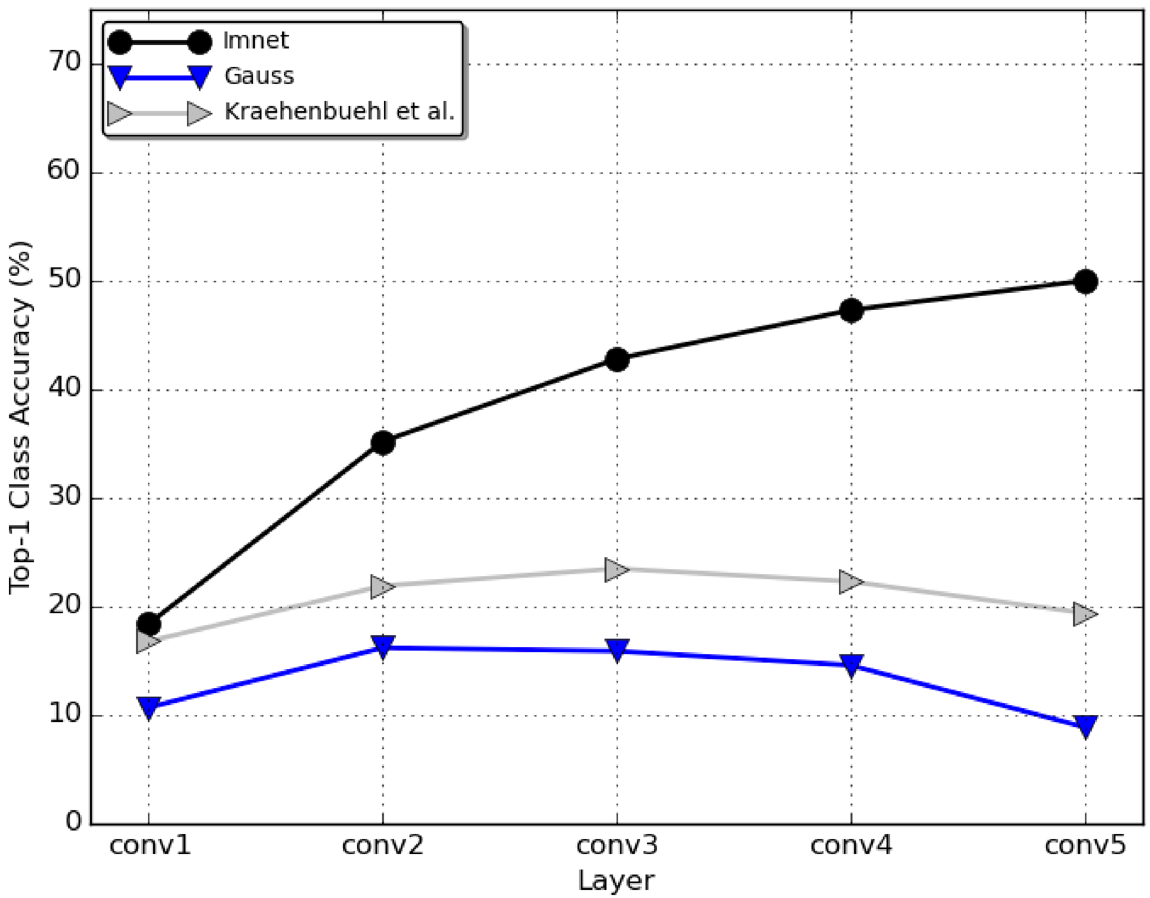
[1] Chen *et al.* In arXiv, 2016.
[2] Yu and Koltun. In ICLR, 2016

Task Generalization: ILSVRC linear classification

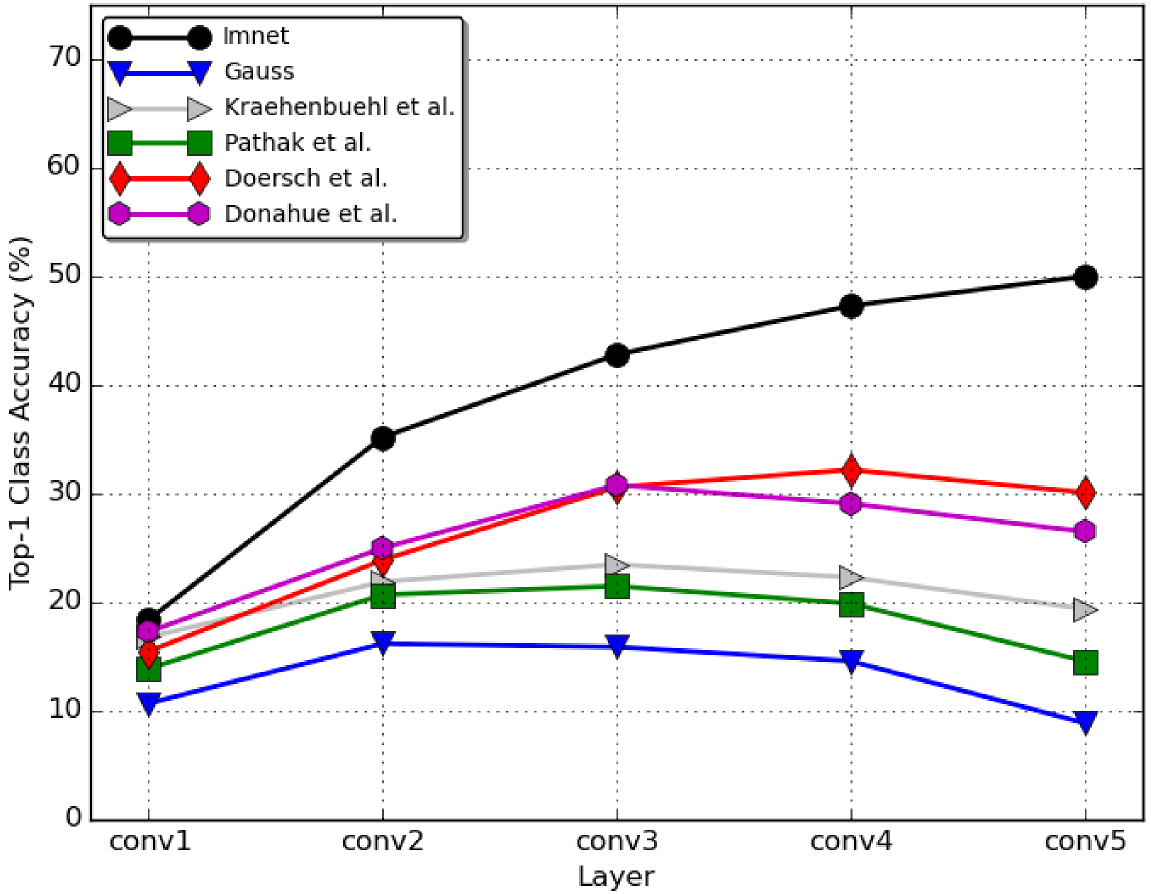


Are semantic classes ***linearly separable***
in the learned feature space?

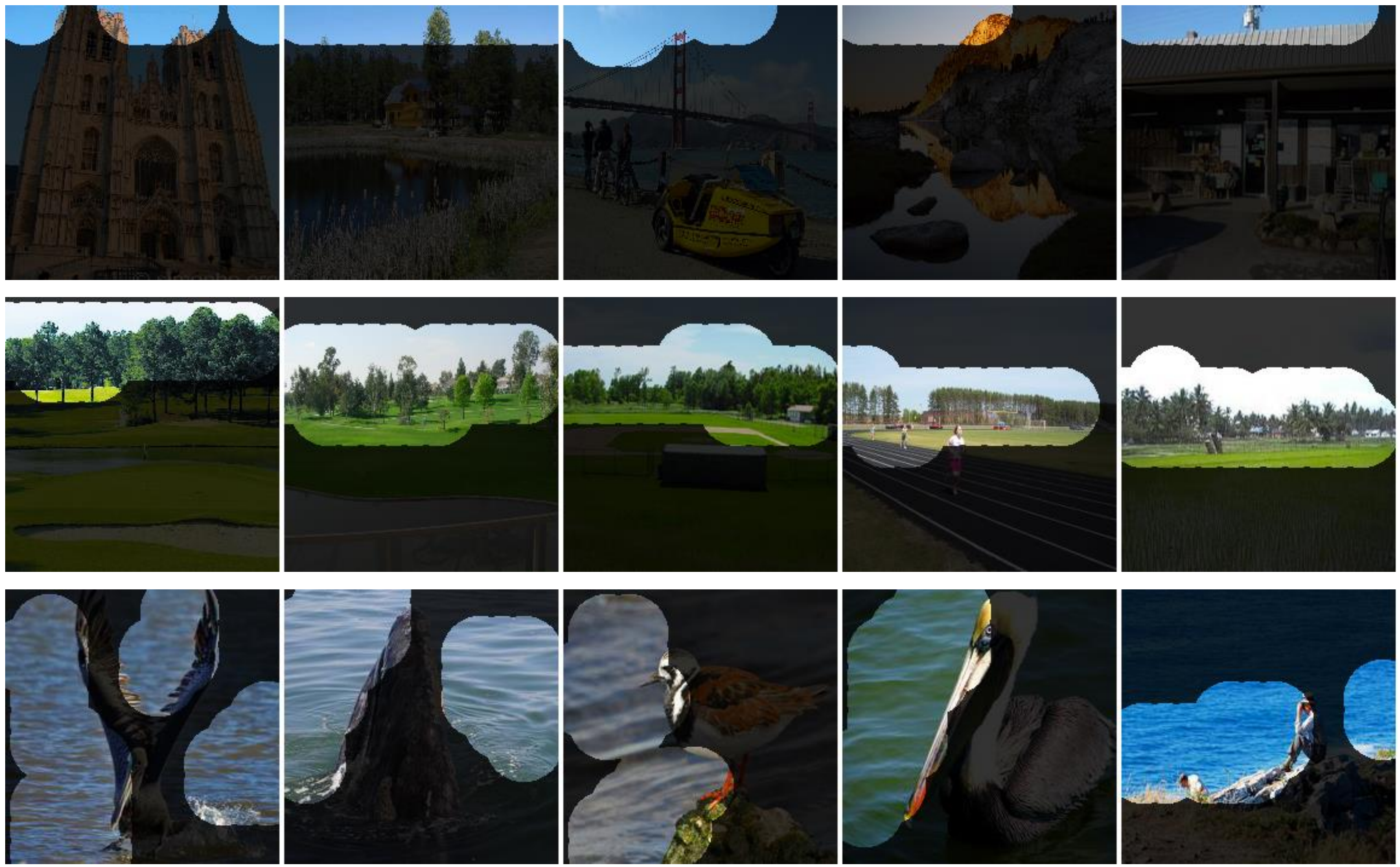
Task Generalization: ILSVRC linear classification



Task Generalization: ILSVRC linear classification



Hidden Unit (conv5) Activations

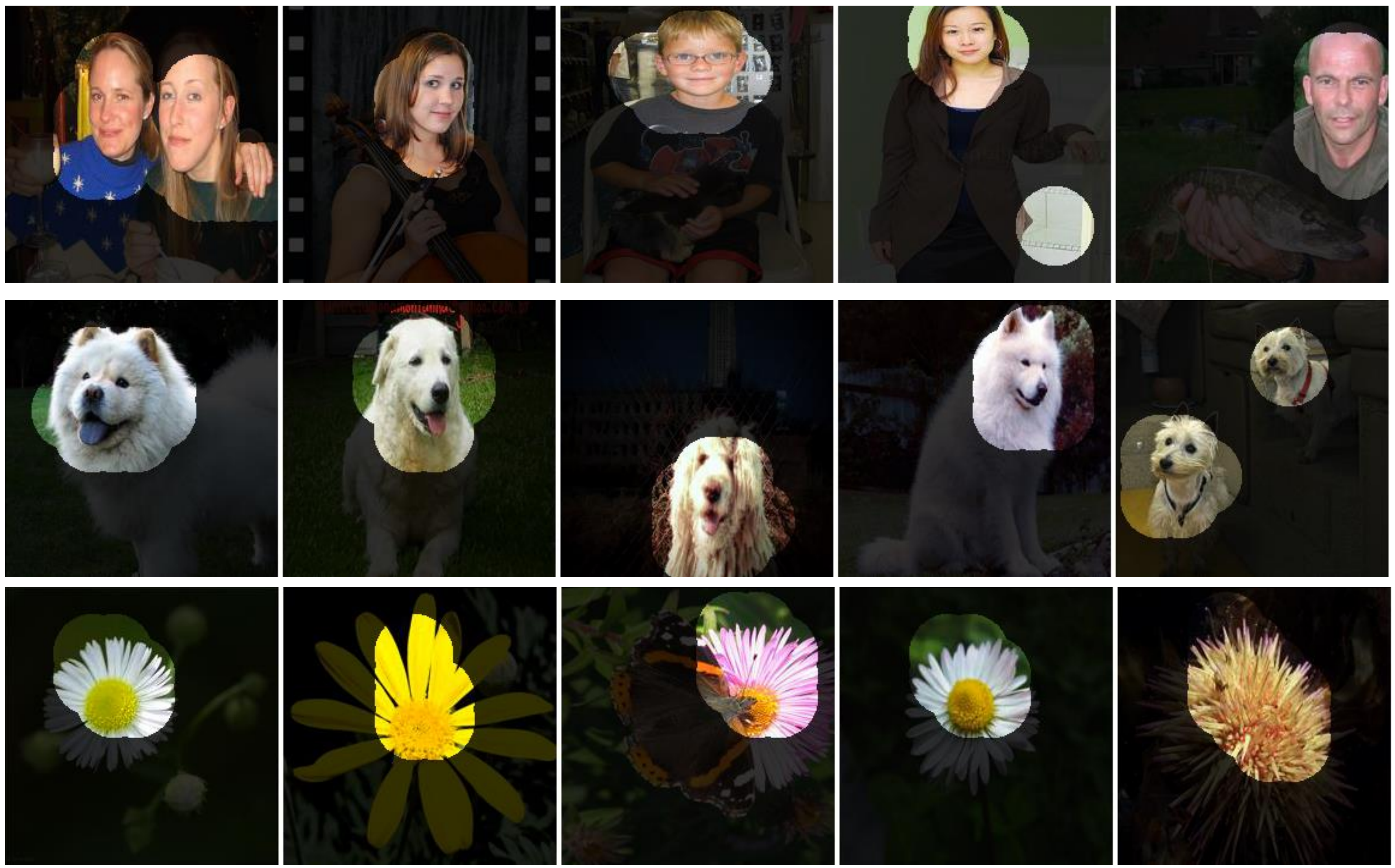


sky

trees

water

Hidden Unit (conv5) Activations



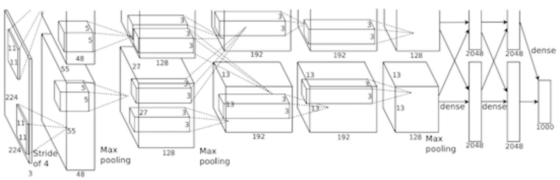
faces

dog faces

flowers

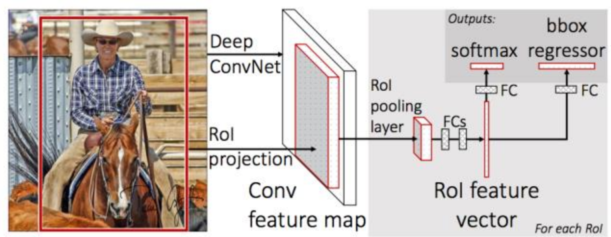
Dataset & Task Generalization on PASCAL VOC

Does the feature representation
transfer to other datasets and tasks?



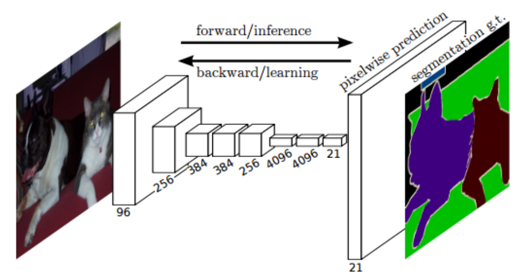
Classification

Krähenbühl et al. In ICLR, 2016.



Detection

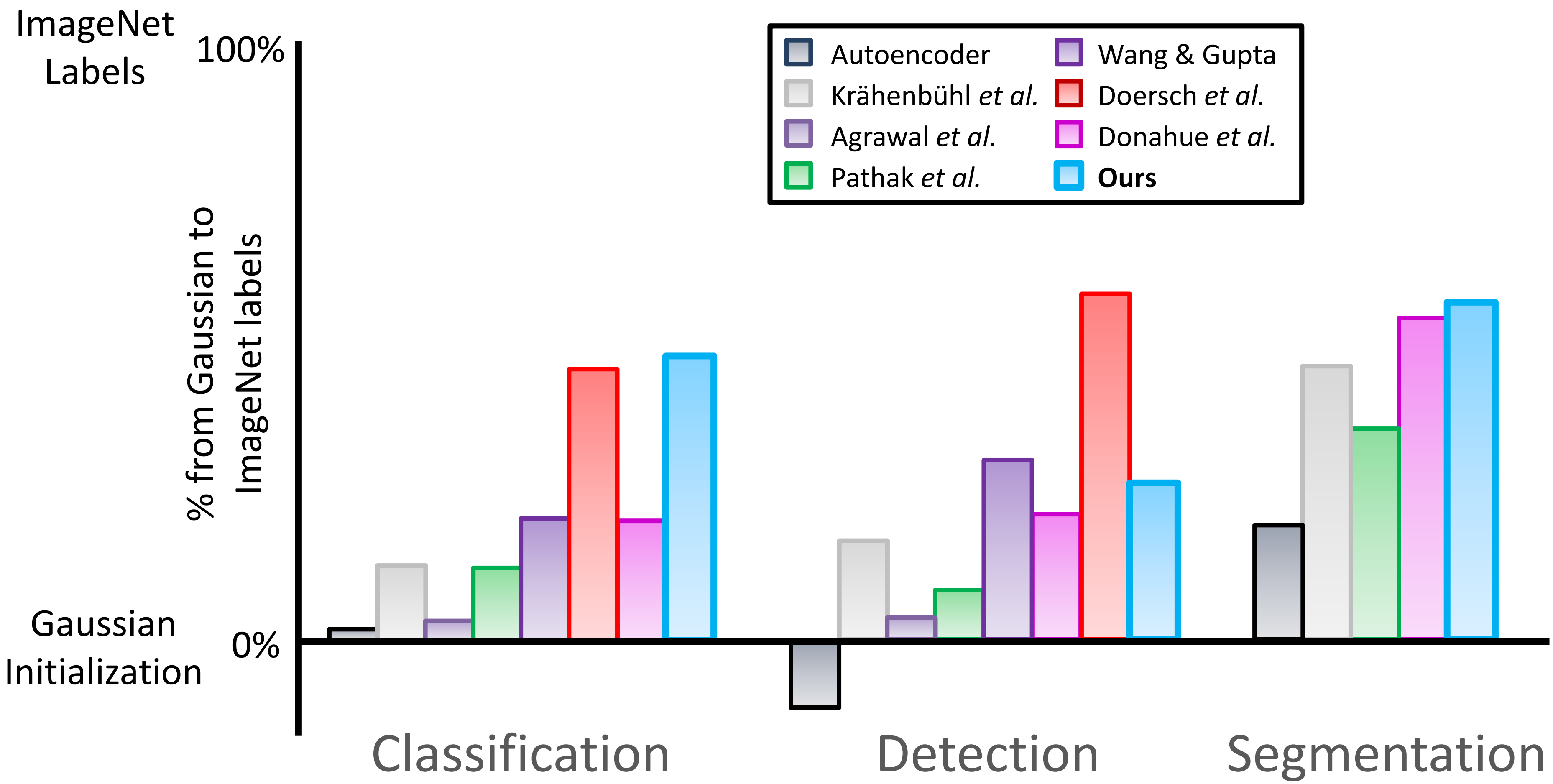
Fast R-CNN. Girshick. In ICCV, 2015.



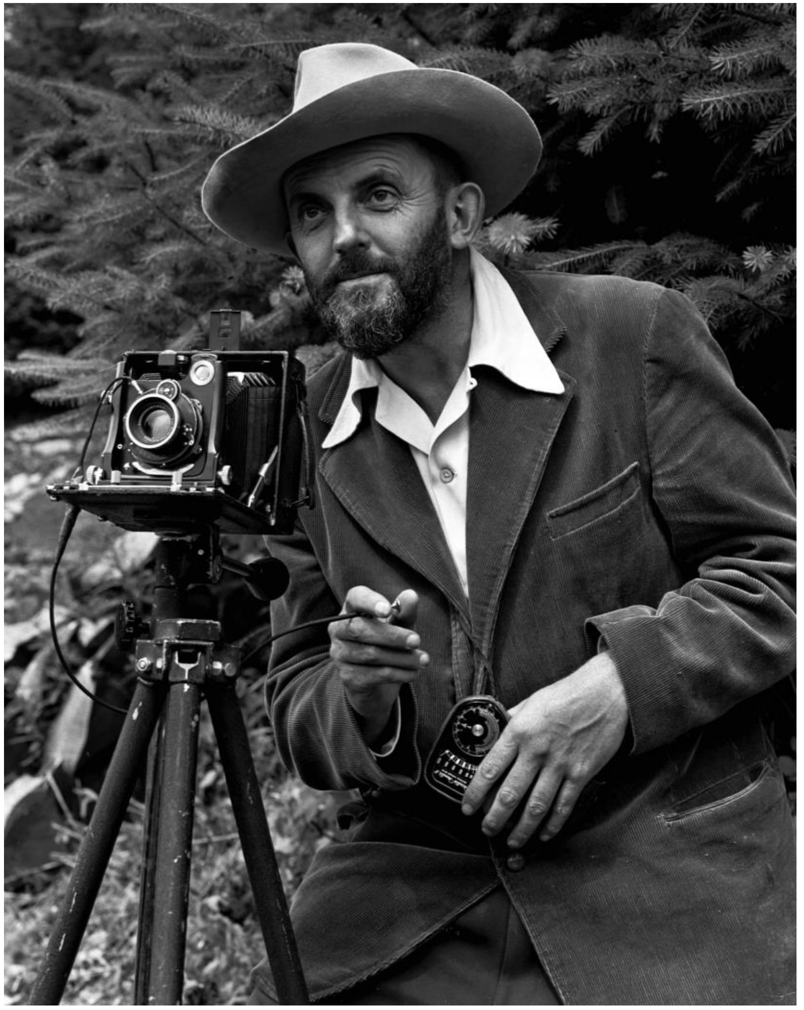
Segmentation

FCNs. Long et al. In CVPR, 2015.

Dataset & Task Generalization on PASCAL VOC



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

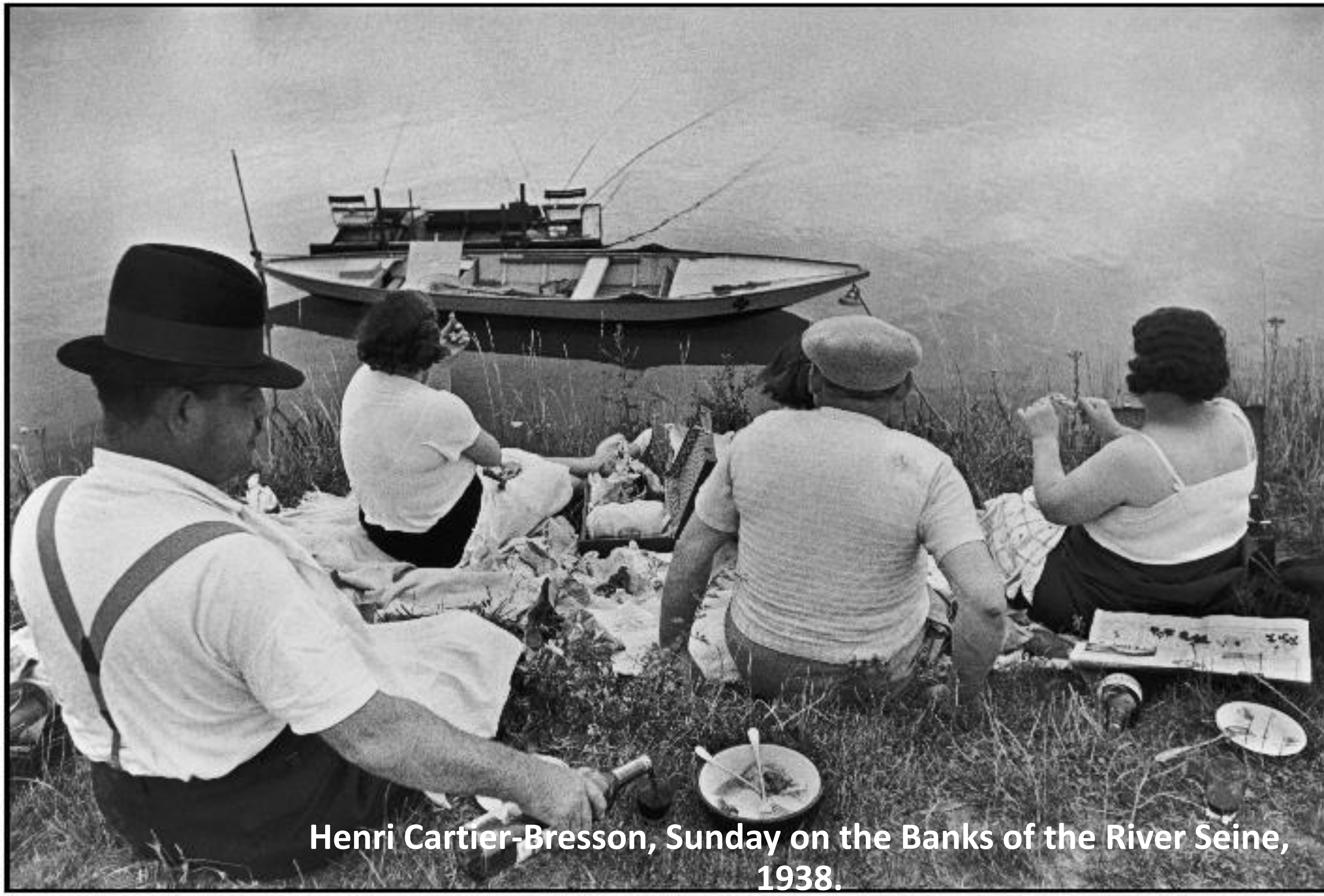



Thylacine, Dr. David Fleay, extinct in 1936.


Thylacine, Dr. David Fleay, extinct in 1936.

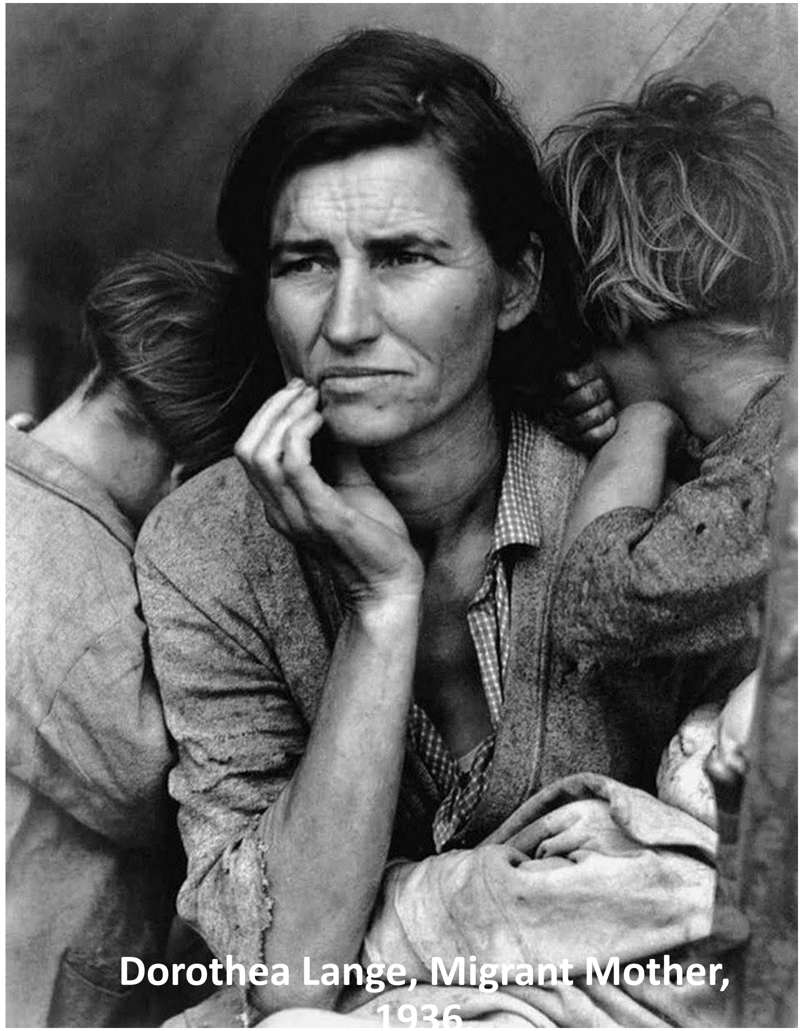
Amateur Family Photo, 1956

Amateur Family Photo, 1956

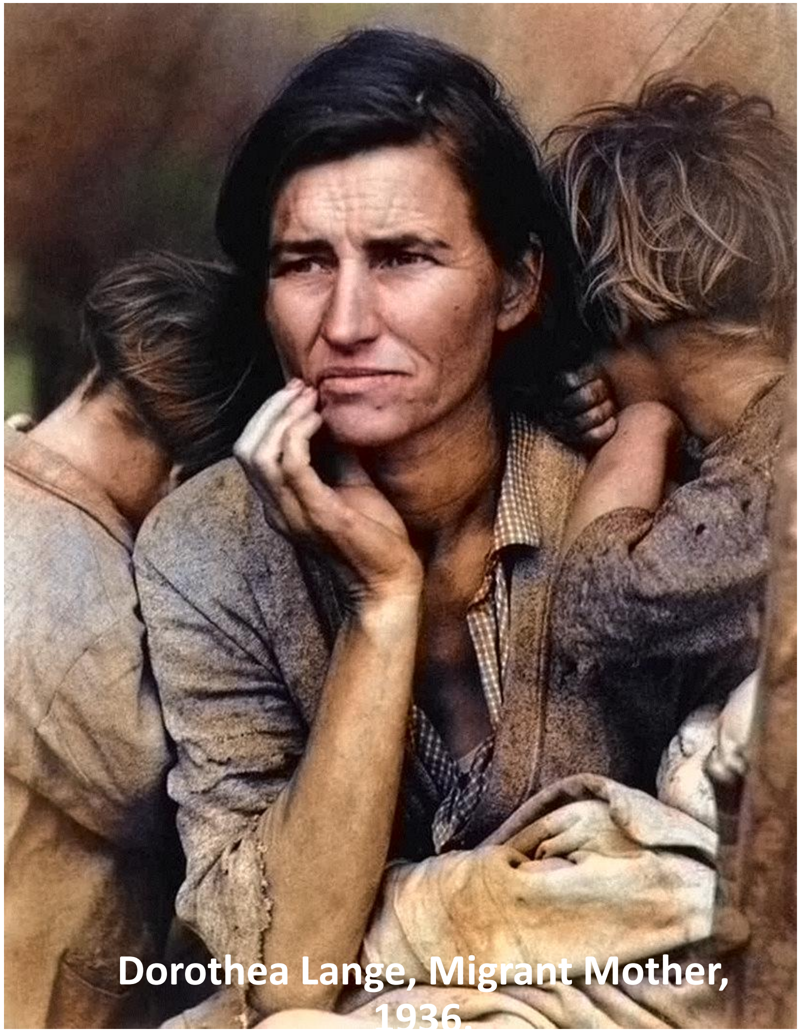


Henri Cartier-Bresson, Sunday on the Banks of the River Seine, 1938.

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