Self Supervision

through Context, Color,
and Physical Interactions

Lama Affara

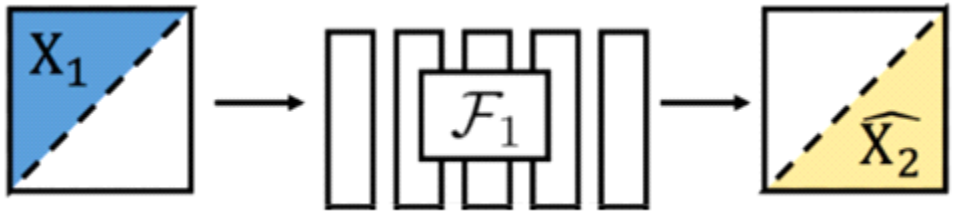
Basic Learning Types

- Supervised Learning
  - Classifying data into labels
- Unsupervised Learning
  - Clustering data, Anomaly detection
- Semi/Weak Supervised Learning
  - Missing/Incomplete labels
- Reinforcement Learning
  - Rewards from actions

- What about Self Supervision?

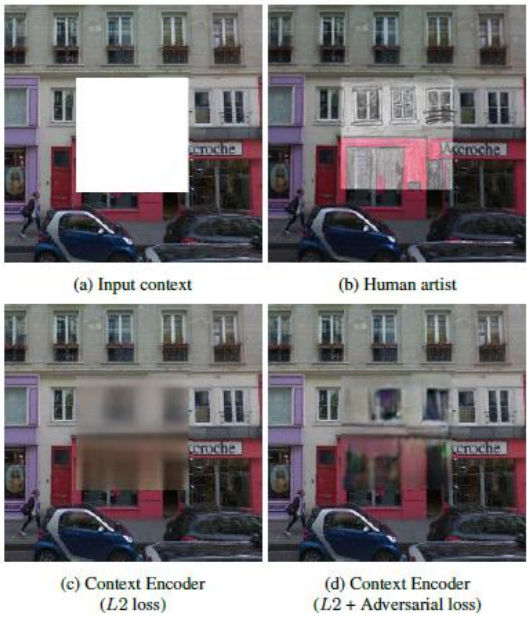
Self Supervision

- No supervision from other agents
- Supervision can come from other modalities or from time



Self Supervision Examples

Context

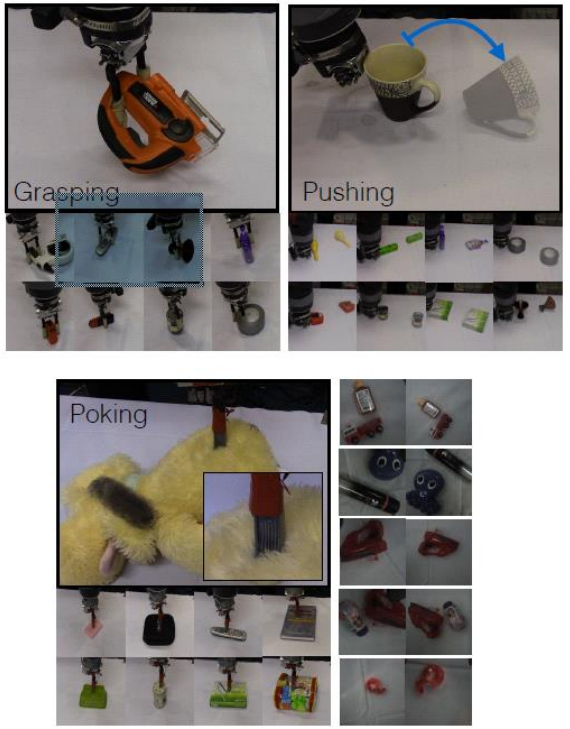


(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

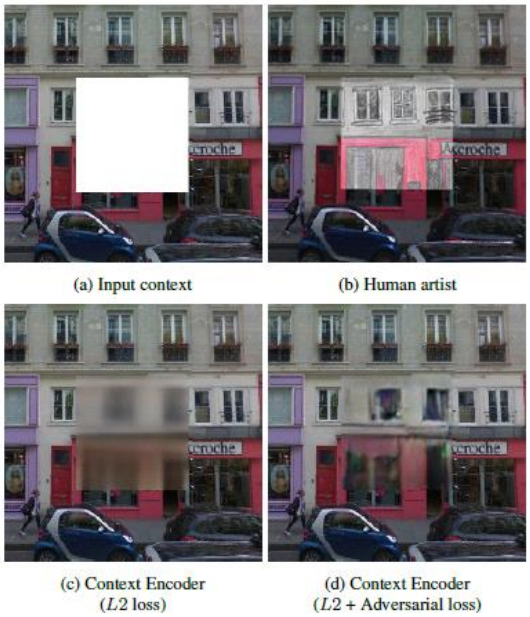
Color

Physical Interactions



Self Supervision Examples

Context



(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

Color

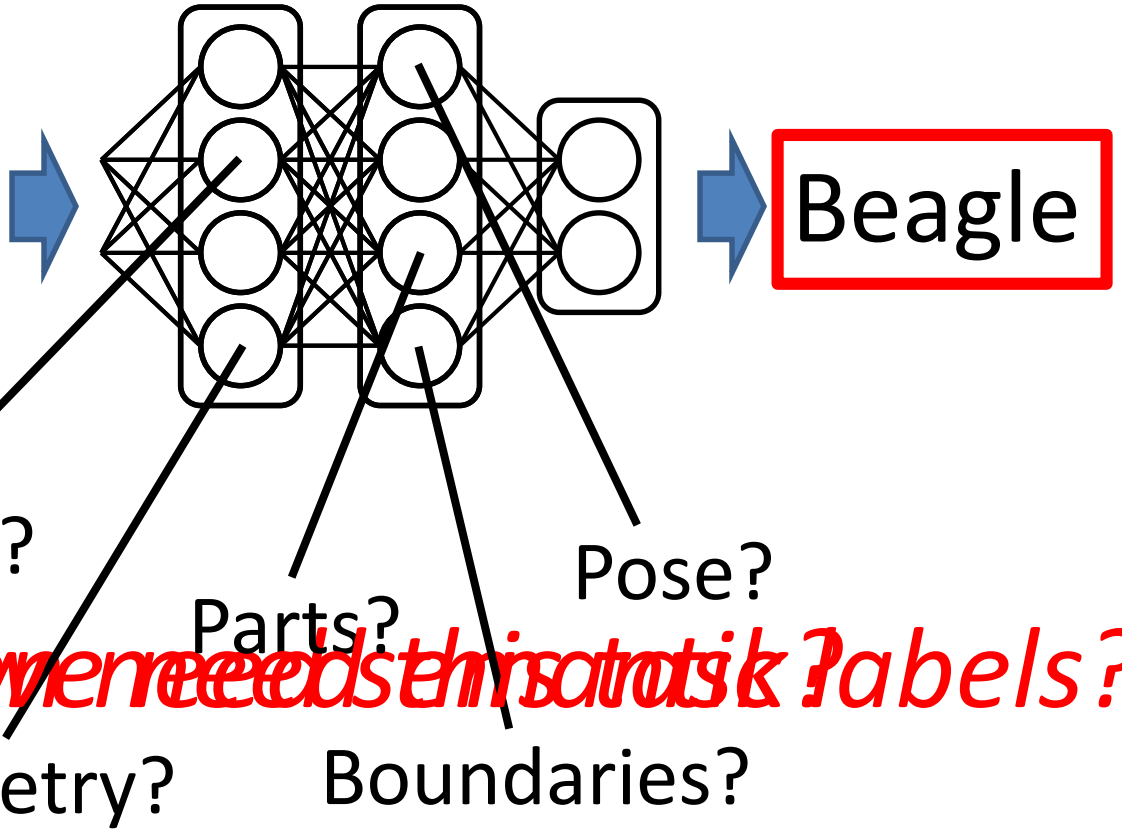
Physical Interactions

Using Patch **Context** for Self Supervision

Unsupervised Visual Representation Learning by Context Prediction

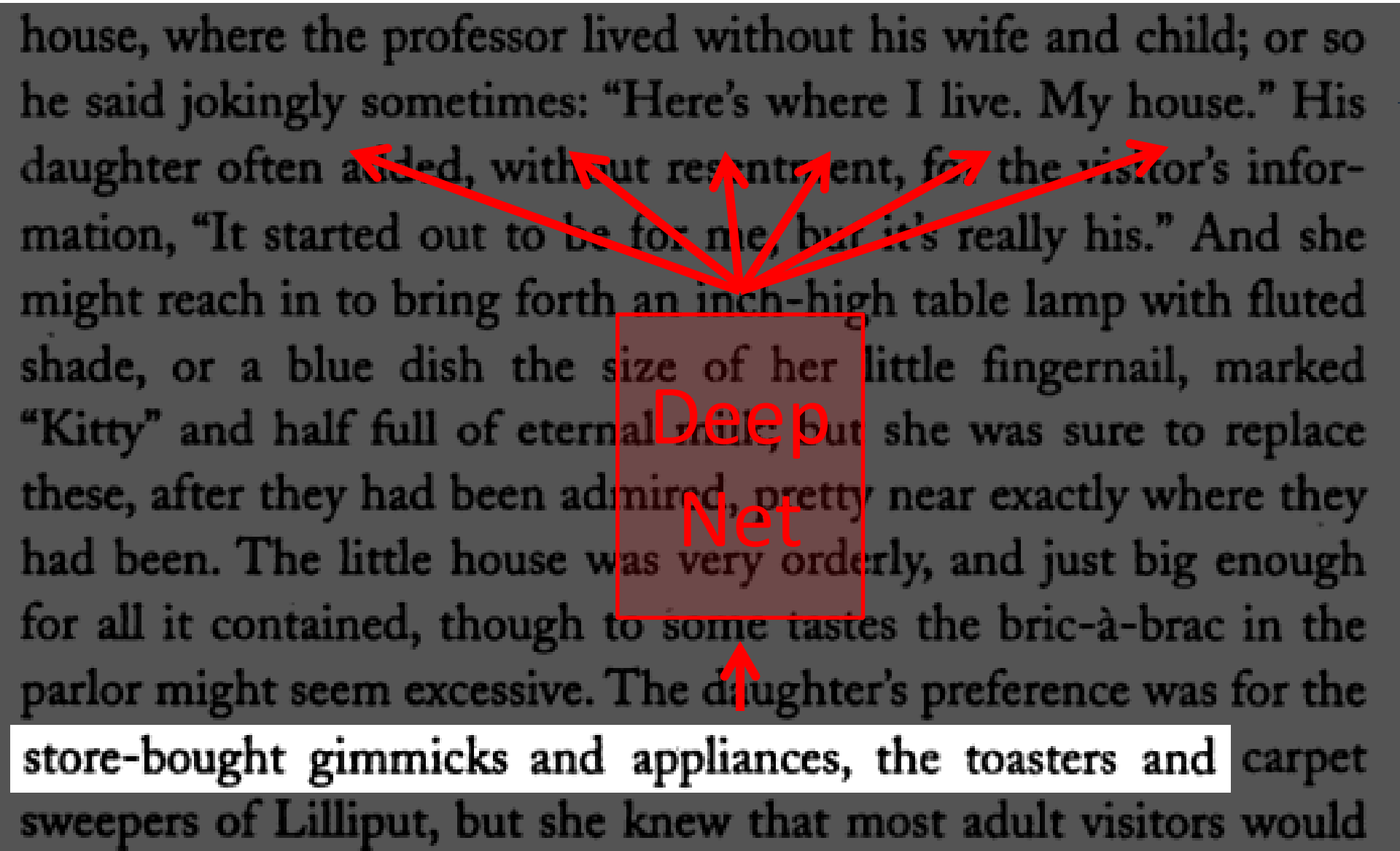
Carl Doersch[1,2] Abhinav Gupta[1] Alexei A. Efros[2]

ImageNet + Deep Learning

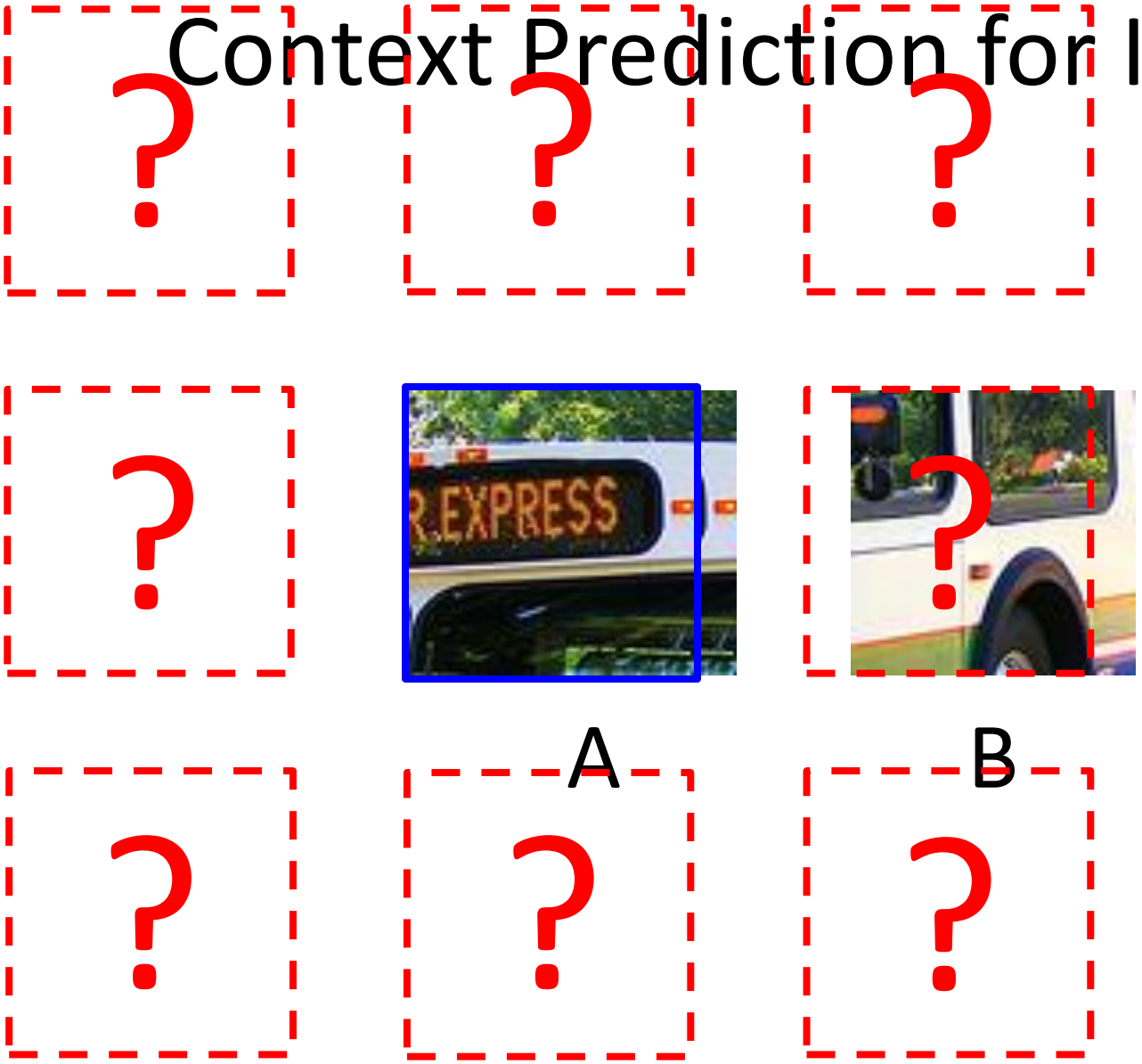


Context as Supervision

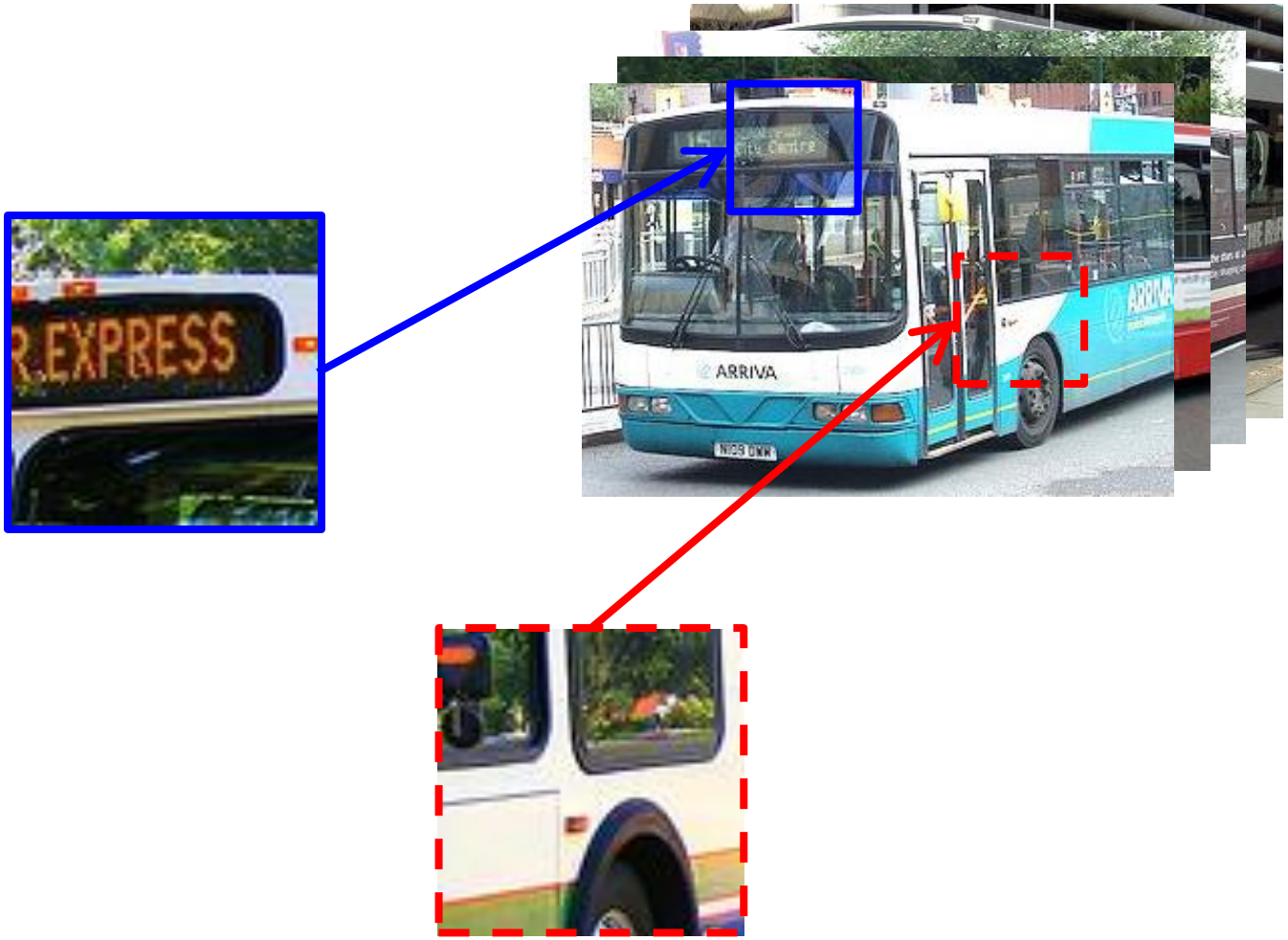
[Collobert & Weston 2008; Mikolov et al. 2013]



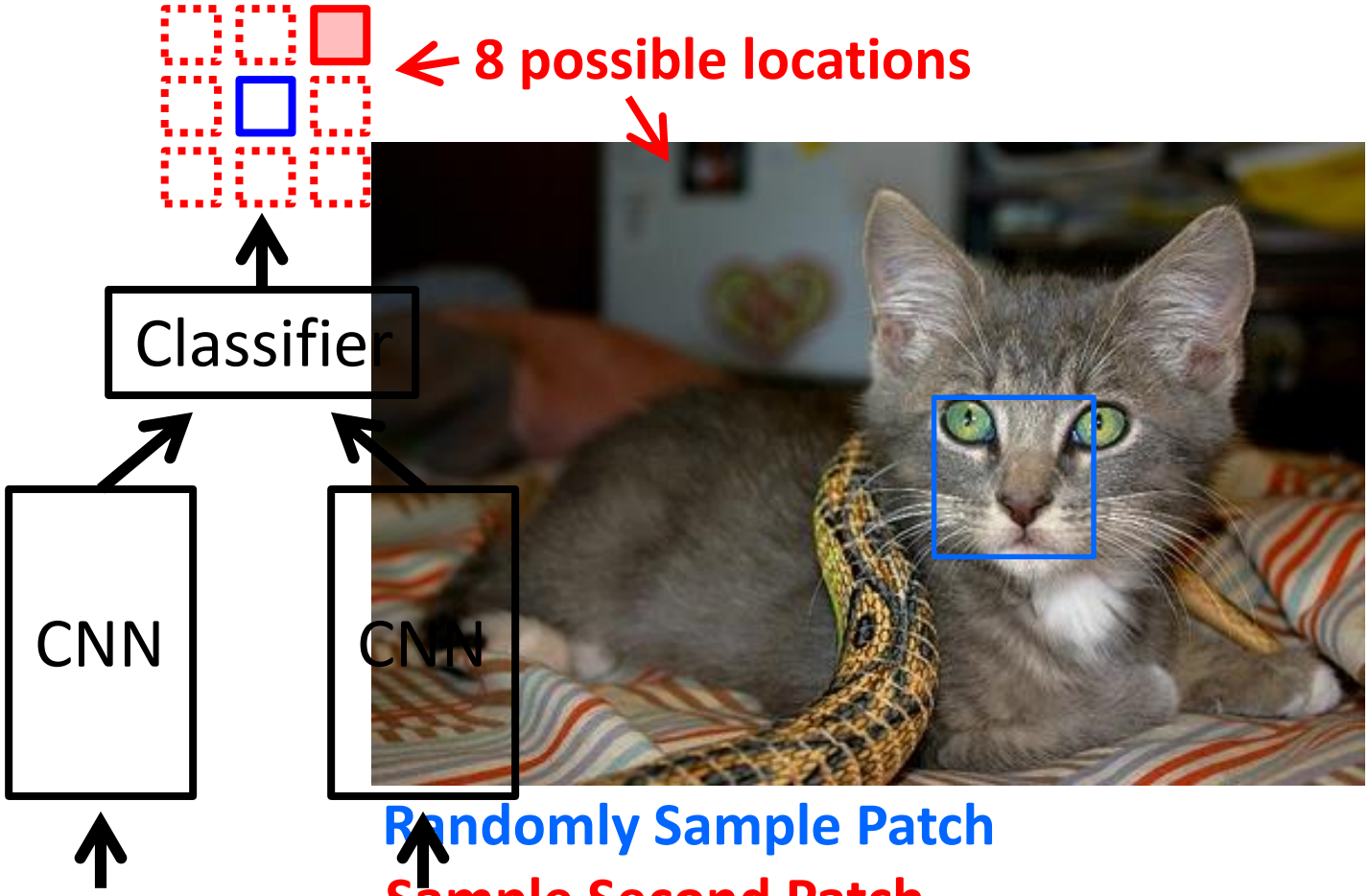
Context Prediction for Images

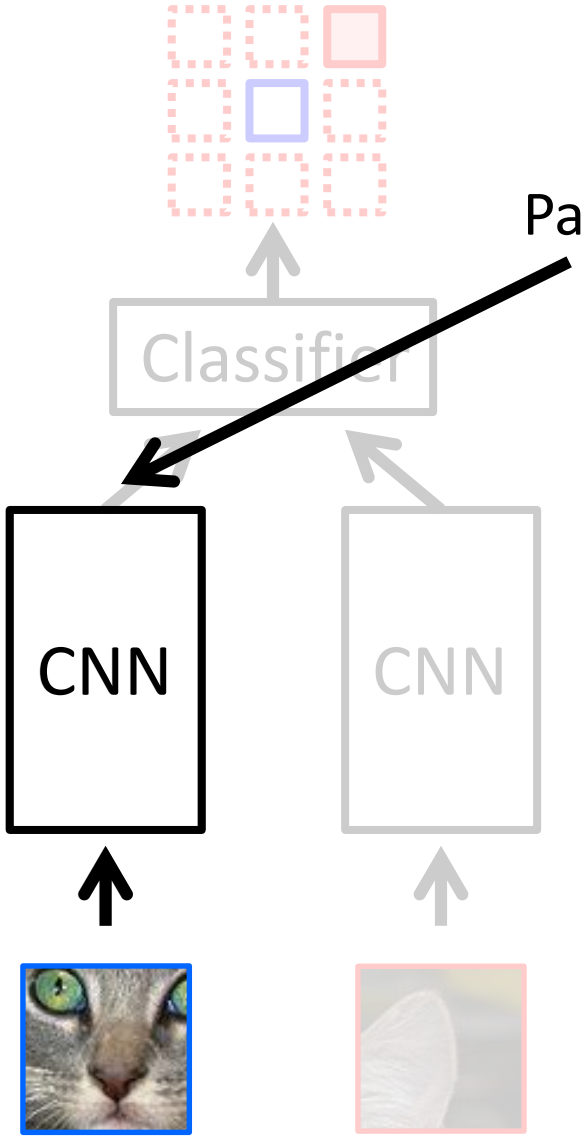


Semantics from a non-semantic task



Relative Position Task





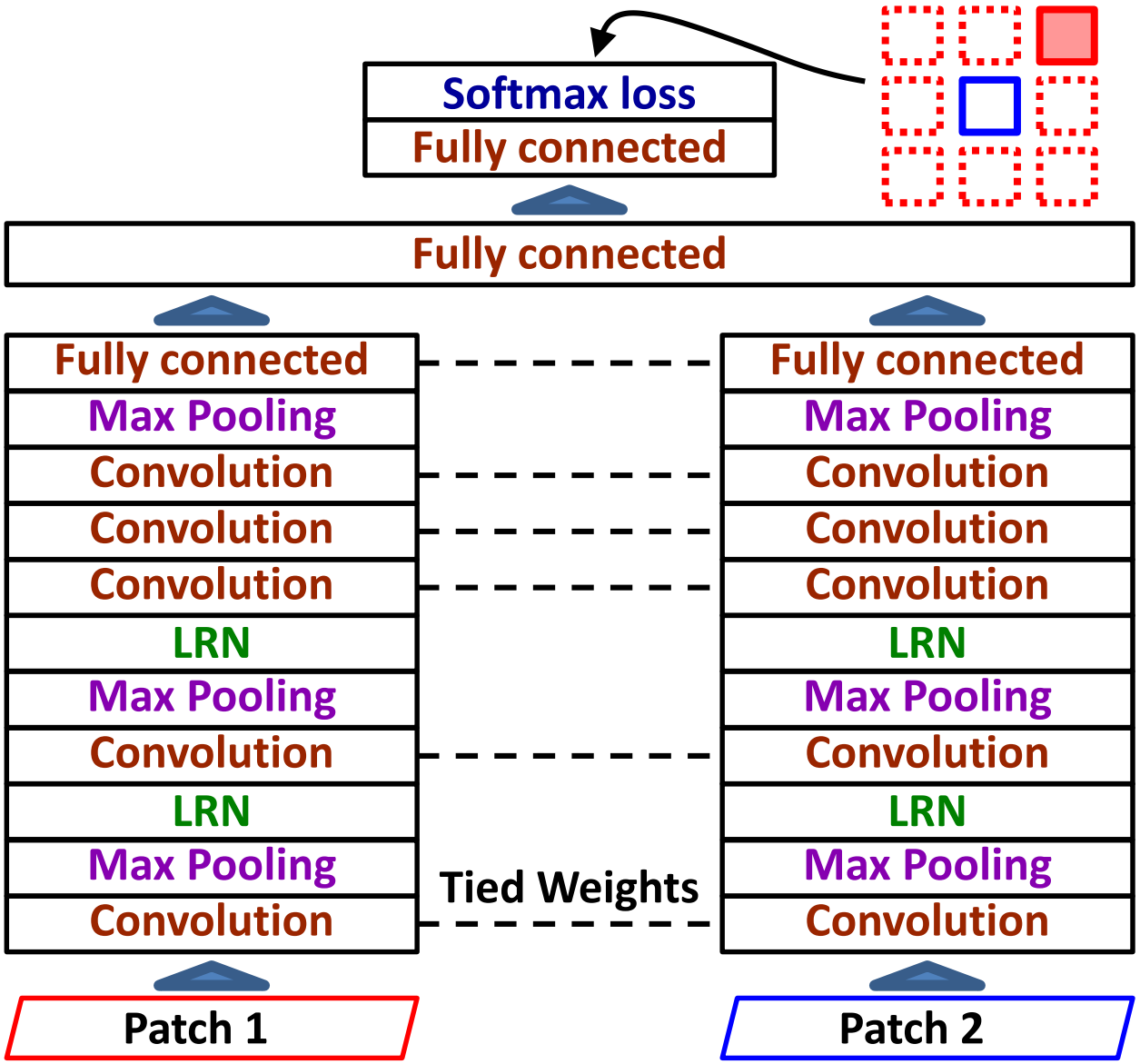
Patch Embedding

Input

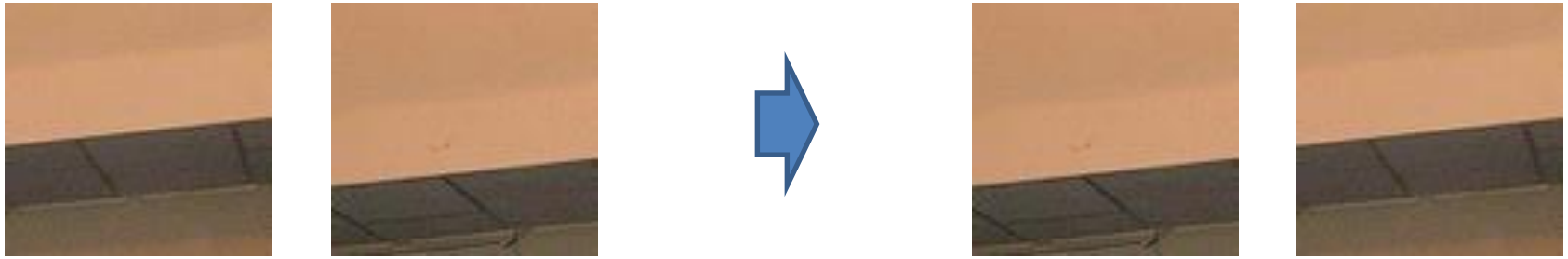
Nearest Neighbors

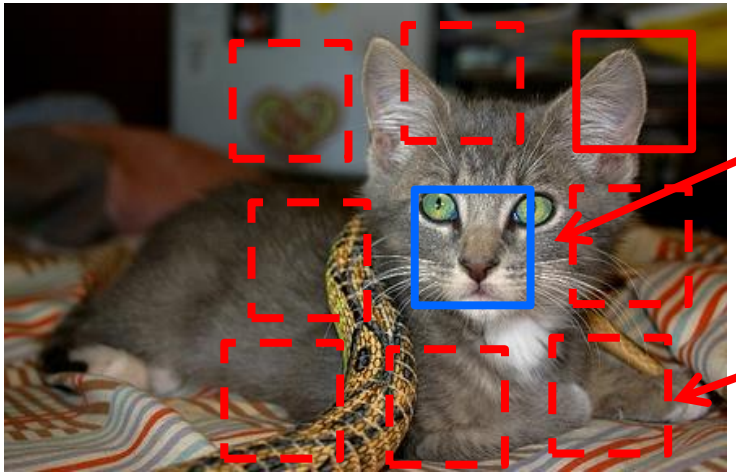
Note: connects ***across*** instances!

Architecture

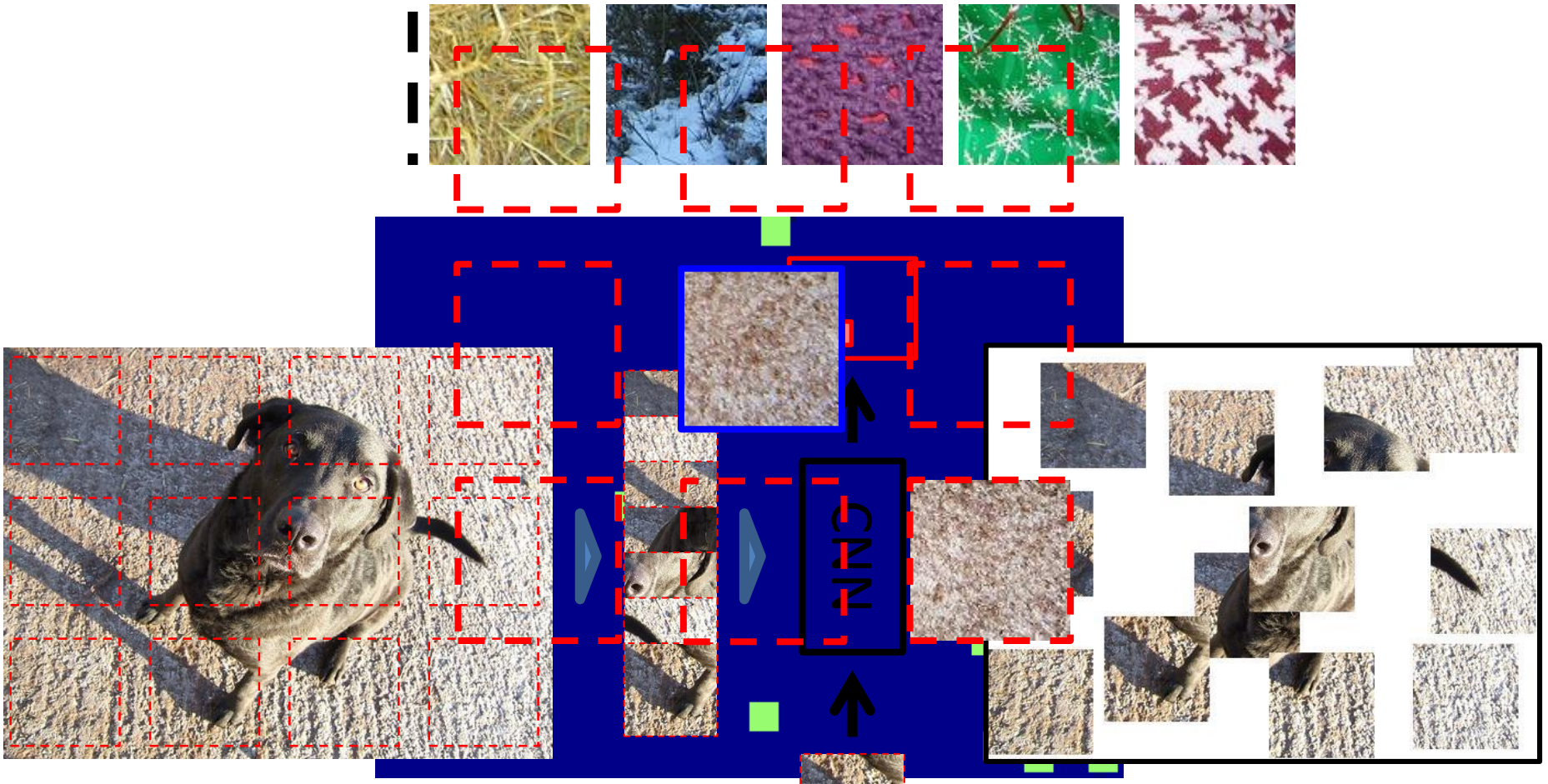


Avoiding Trivial Shortcuts

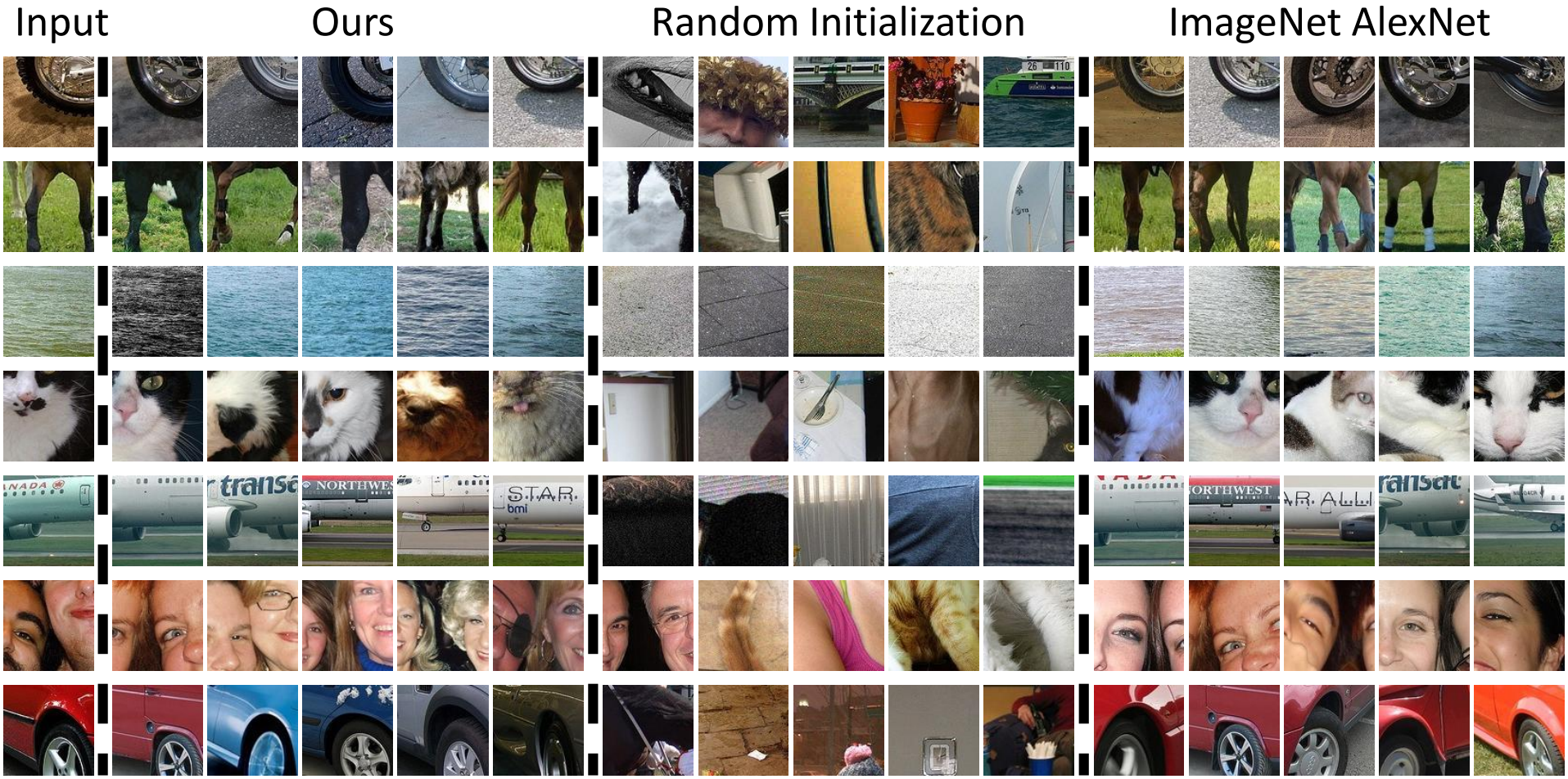




A Not-So “Trivial” Shortcut

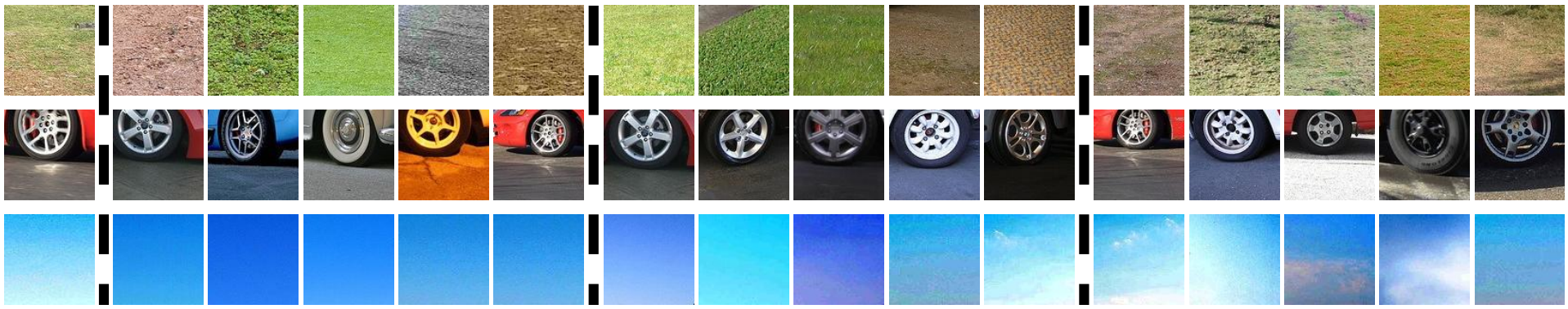


What is learned?

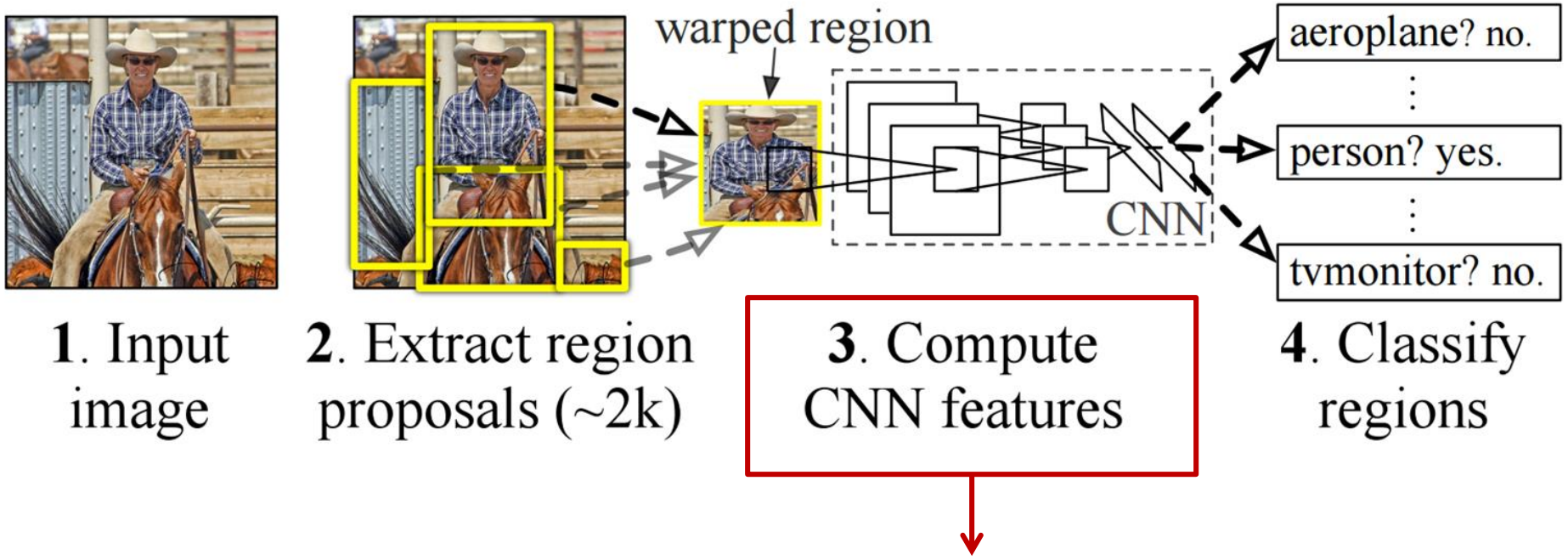


Still don’t capture everything

You don’t always need to learn!

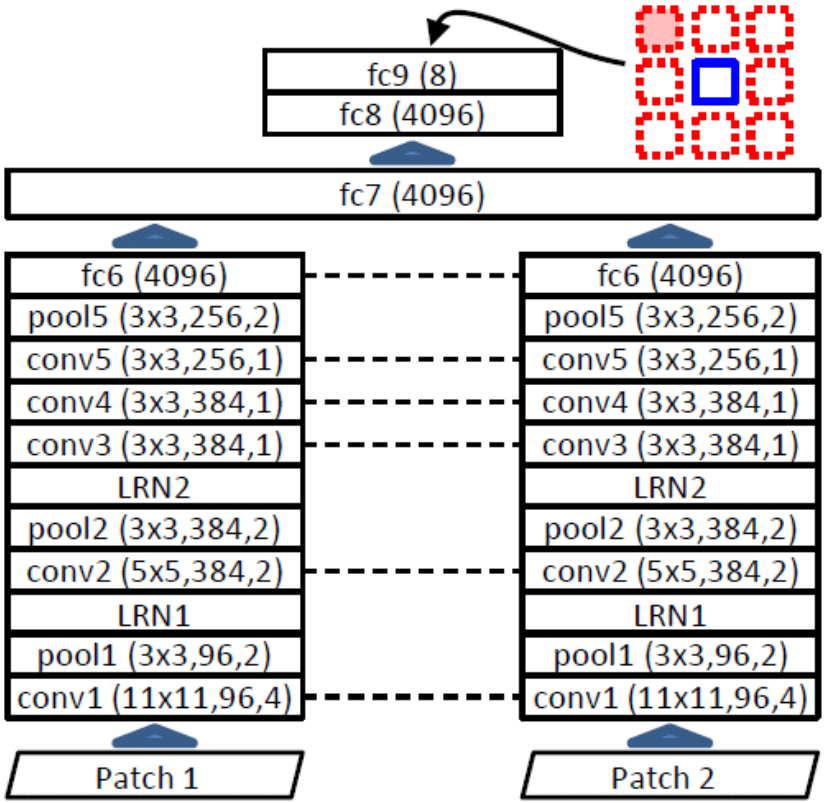


Pre-Training for R-CNN



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

[Girshick et al. 2014]



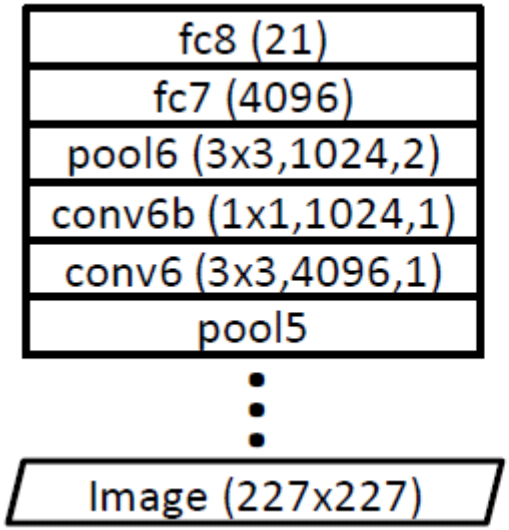
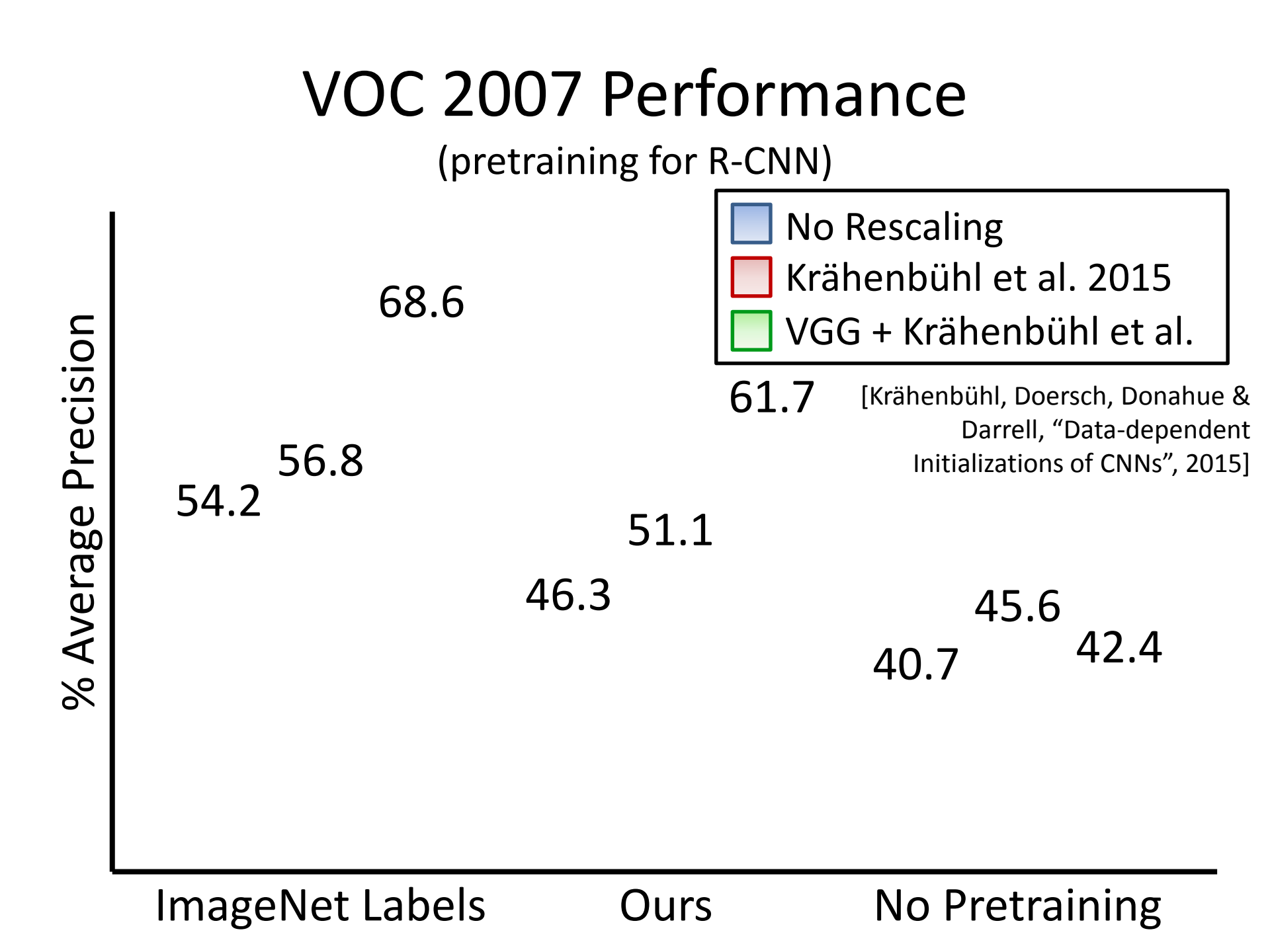


Figure 6. Our architecture for Pascal VOC detection. Layers from conv1 through pool5 are copied from our patch-based network (Figure 3). The new 'conv6' layer is created by converting the fc6 layer into a convolution layer. Kernel sizes, output units, and stride are given in parentheses, as in Figure 3.


VOC 2007 Performance
(pretraining for R-CNN)
No Rescaling
Krähenbühl et al. 2015
VGG + Krähenbühl et al.
[Krähenbühl, Doersch, Donahue & Darrell, "Data-dependent Initializations of CNNs", 2015]
% Average Precision
68.6
61.7
56.8
54.2
51.1
46.3
45.6
42.4
40.7
ImageNet Labels
Ours
No Pretraining

Self Supervision Examples

Context

Color

Physical Interactions

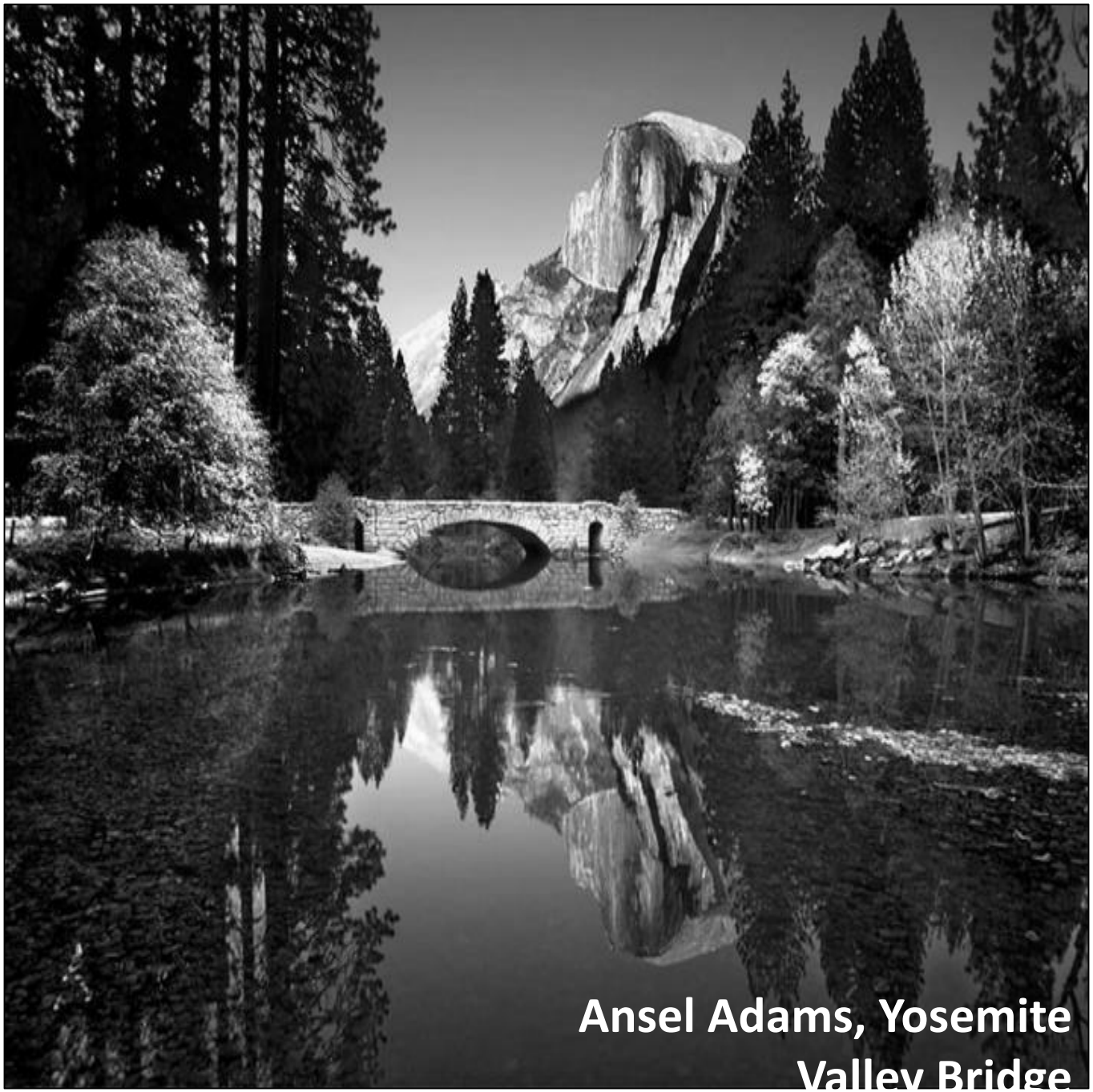
Using **Colorization** for Self Supervision

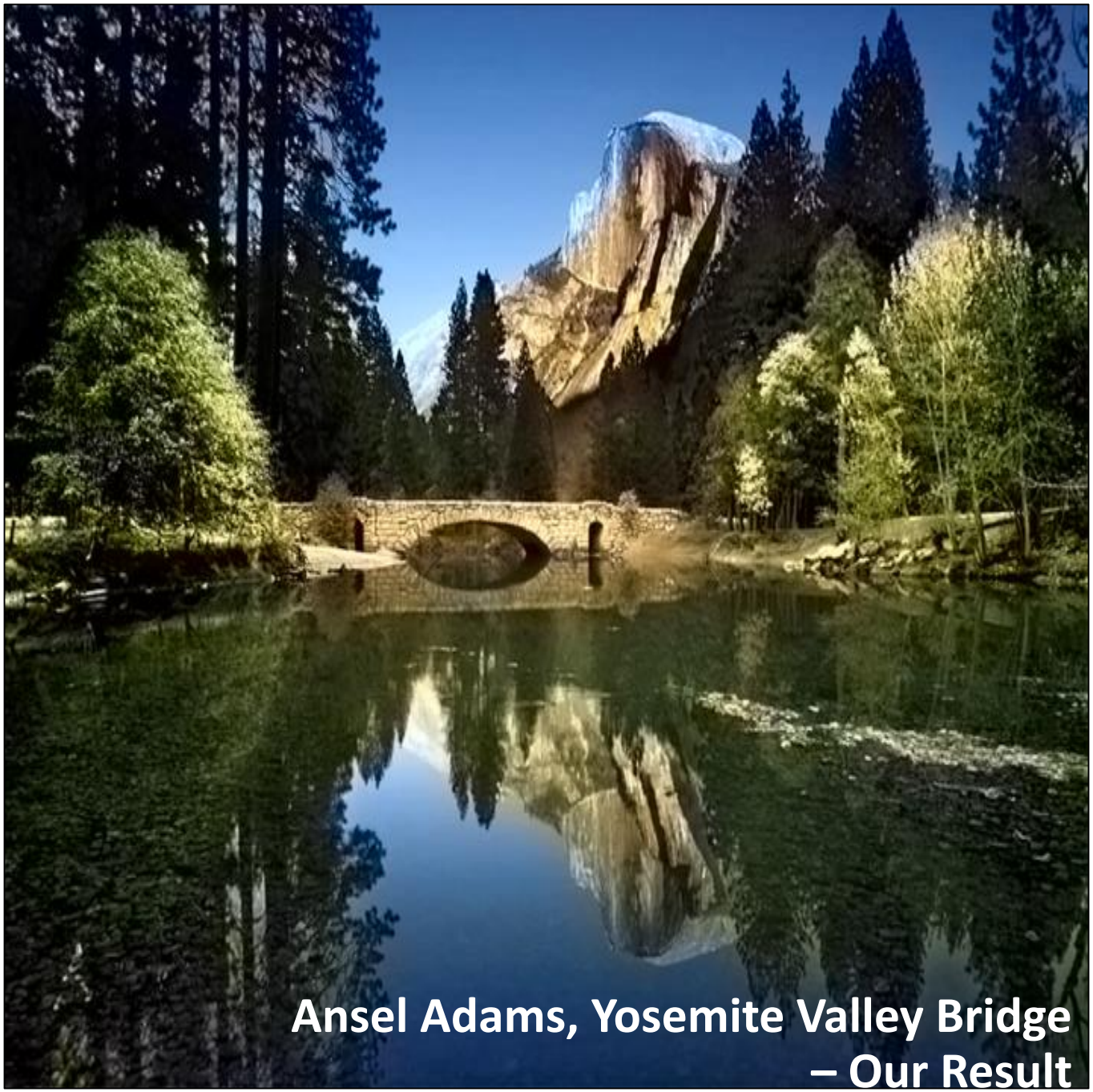
Colorful Image Colorization

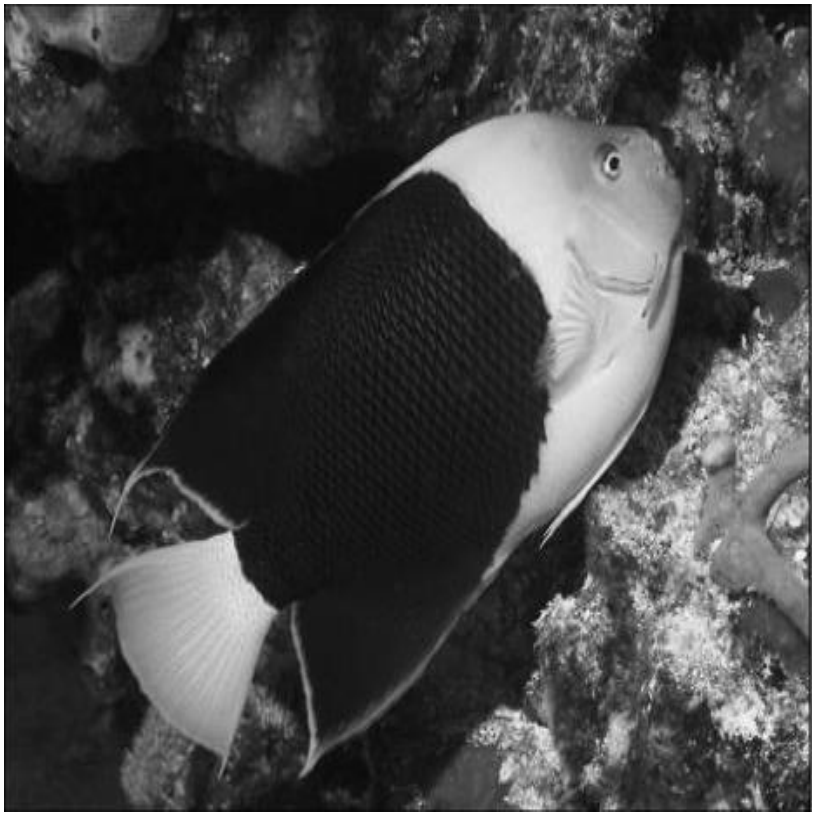
Richard Zhang Phillip Isola Alexei A. Efros

[Demo] [GitHub] [Talk] [Slides] [Paper]

Also check out our new work on Interactive Deep Colorization!


Ansel Adams, Yosemite
Valley Bridge

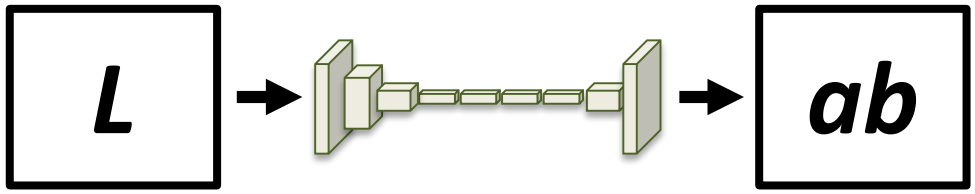

Ansel Adams, Yosemite Valley Bridge
– Our Result

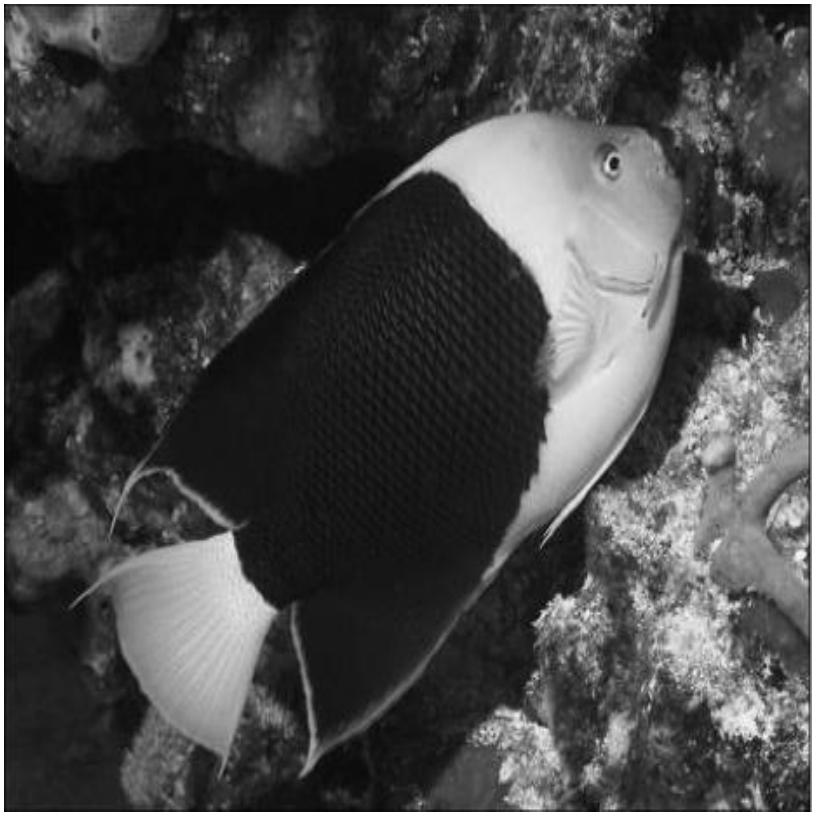


$\mathcal{F}$

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

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





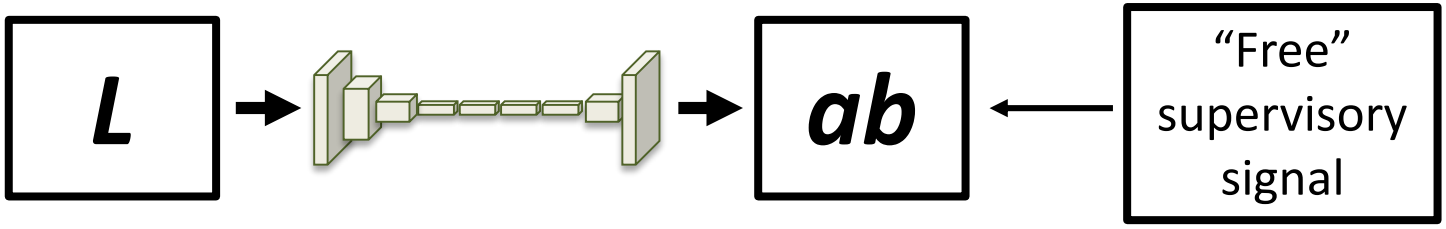
$\mathcal{F}$

Grayscale image: *L* channel

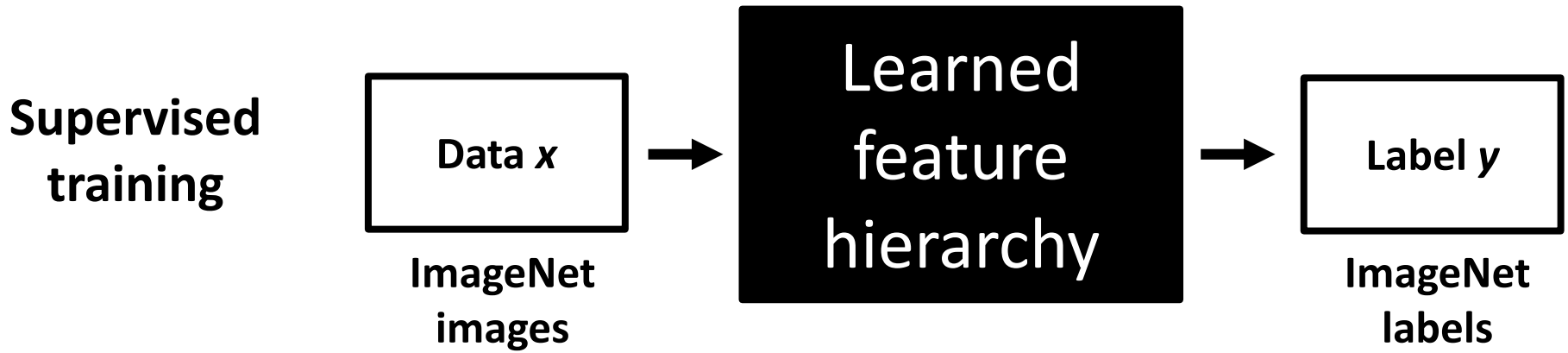
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Concatenate (L,ab)

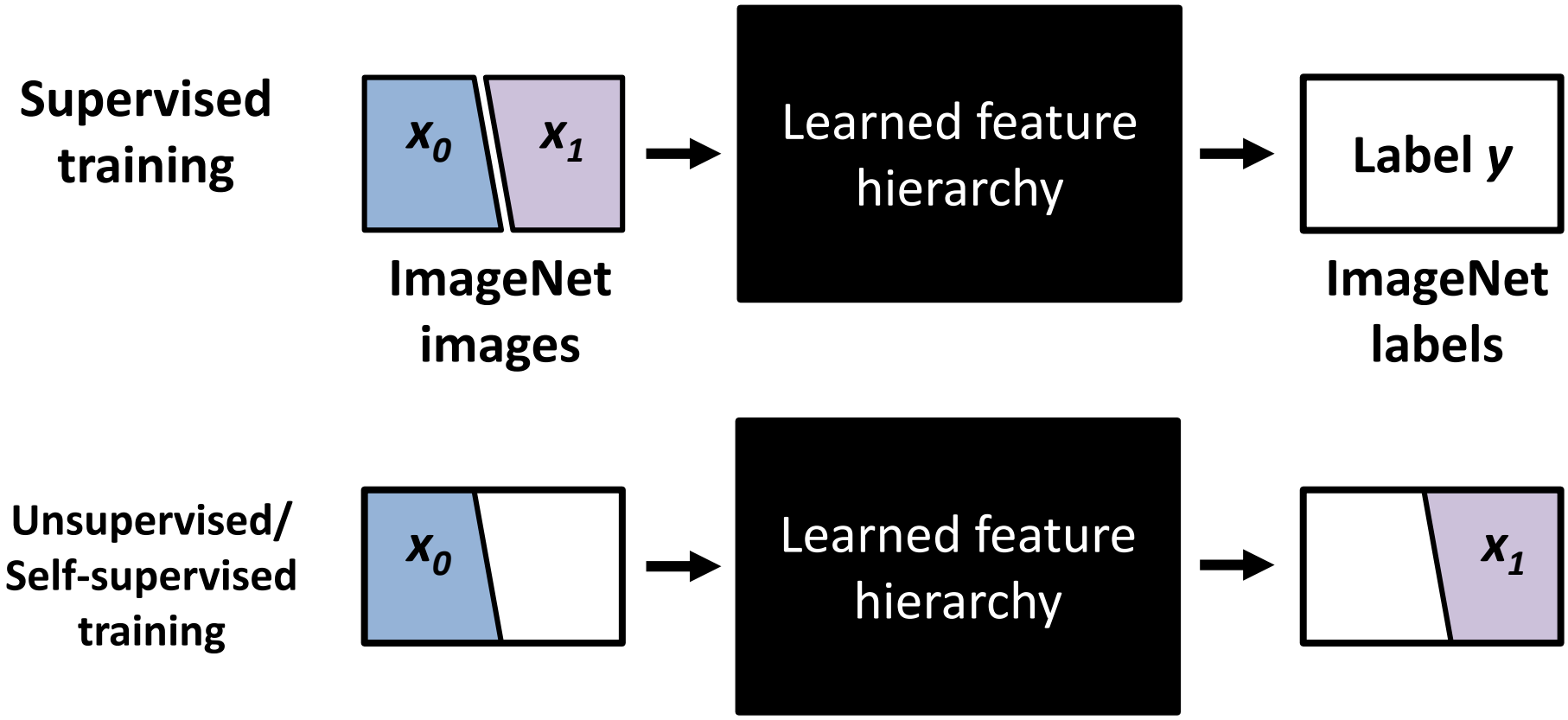
$(\mathbf{X}, \hat{\mathbf{Y}})$



Predicting Labels from Data



Predicting Data from Data



Inherent Ambiguity

Grayscale

Inherent Ambiguity

Our Output

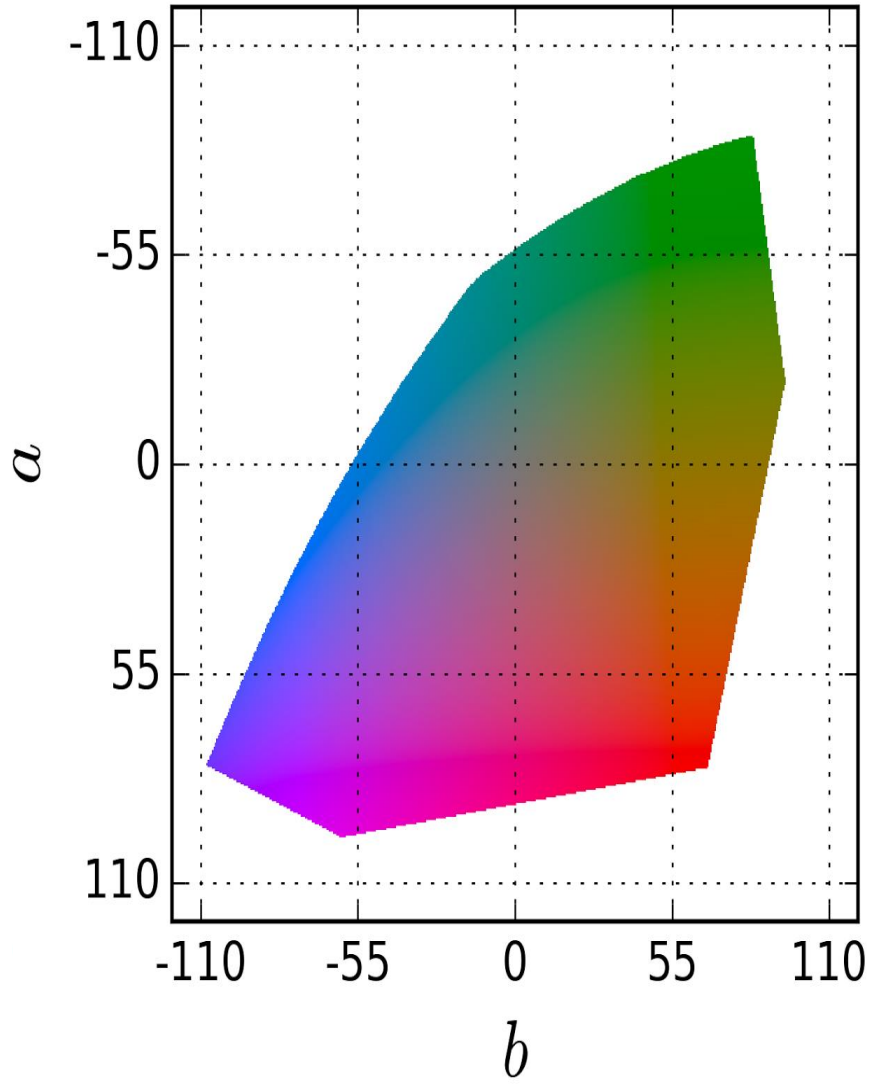
Ground Truth

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



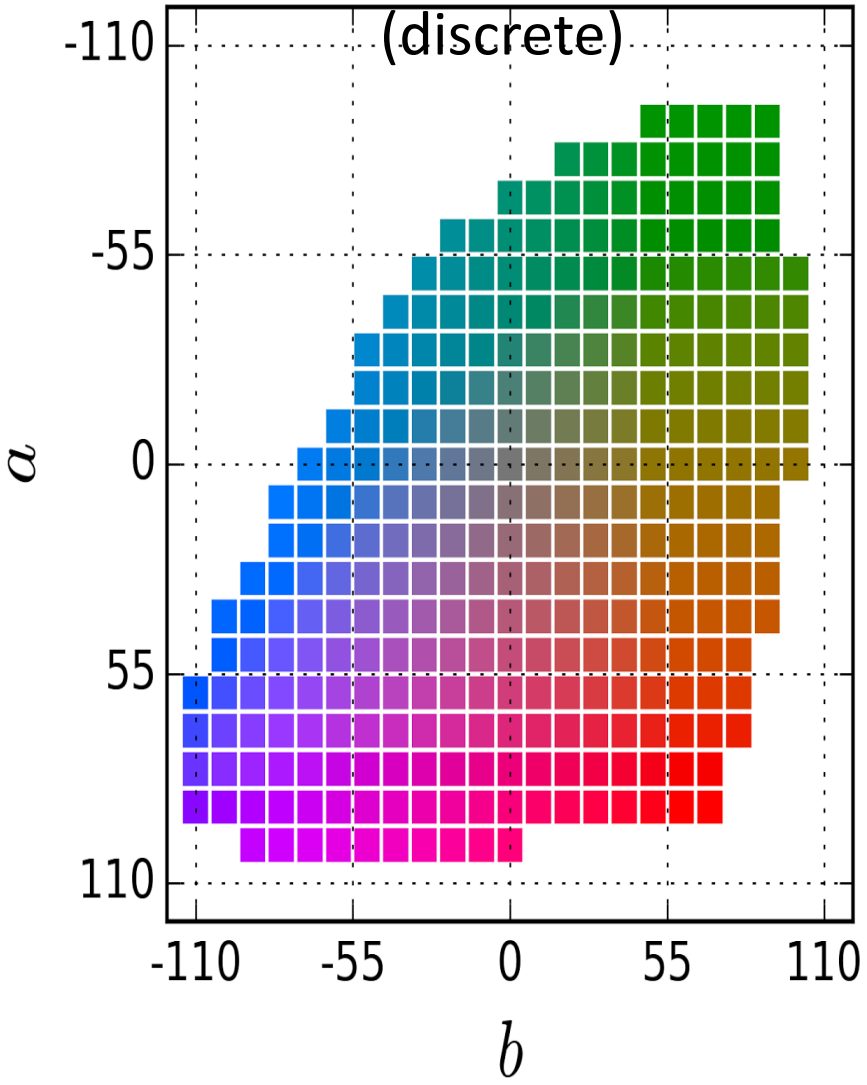
Better Loss Function

- Regression with L2 loss inadequate

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

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

Colors in *ab* space



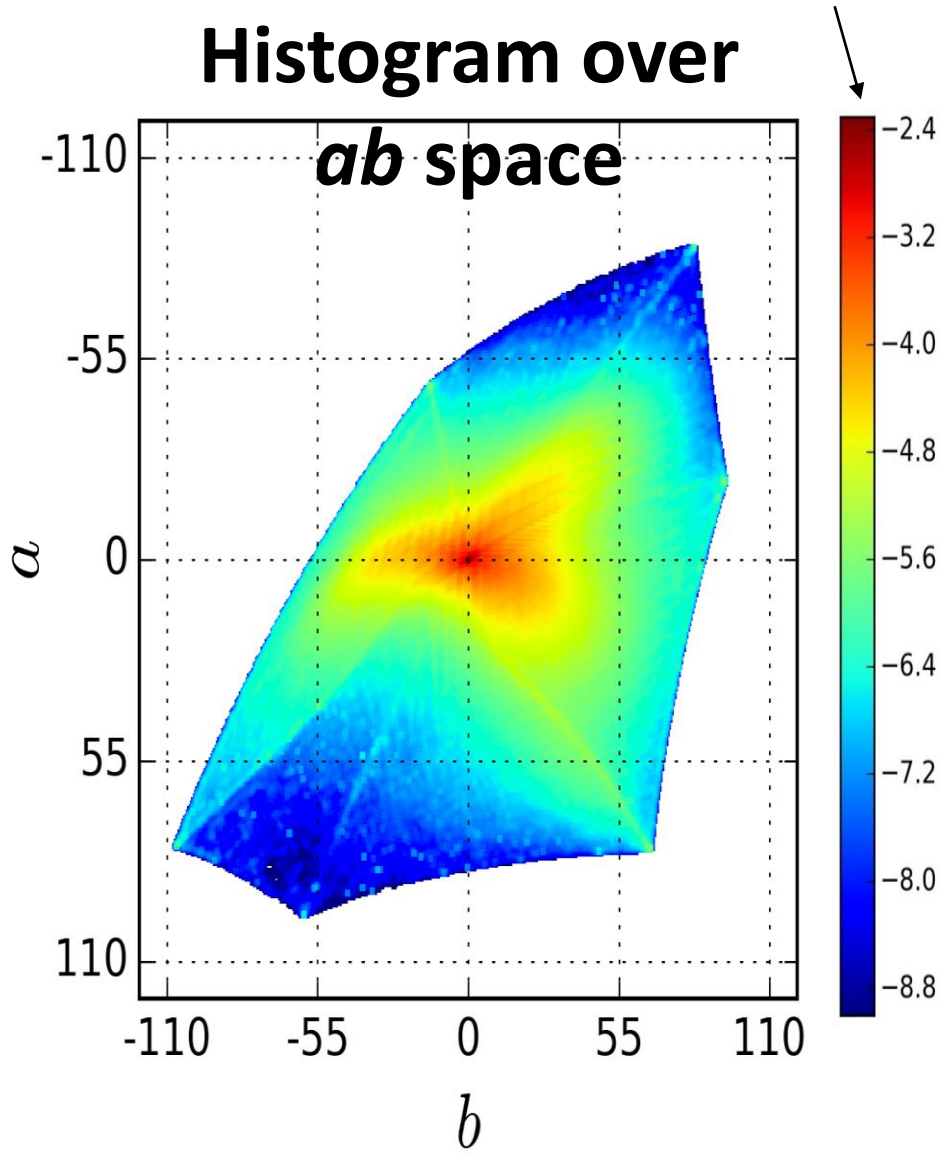
a
b

Better Loss Function

- Regression with L2 loss inadequate

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

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



Better Loss Function

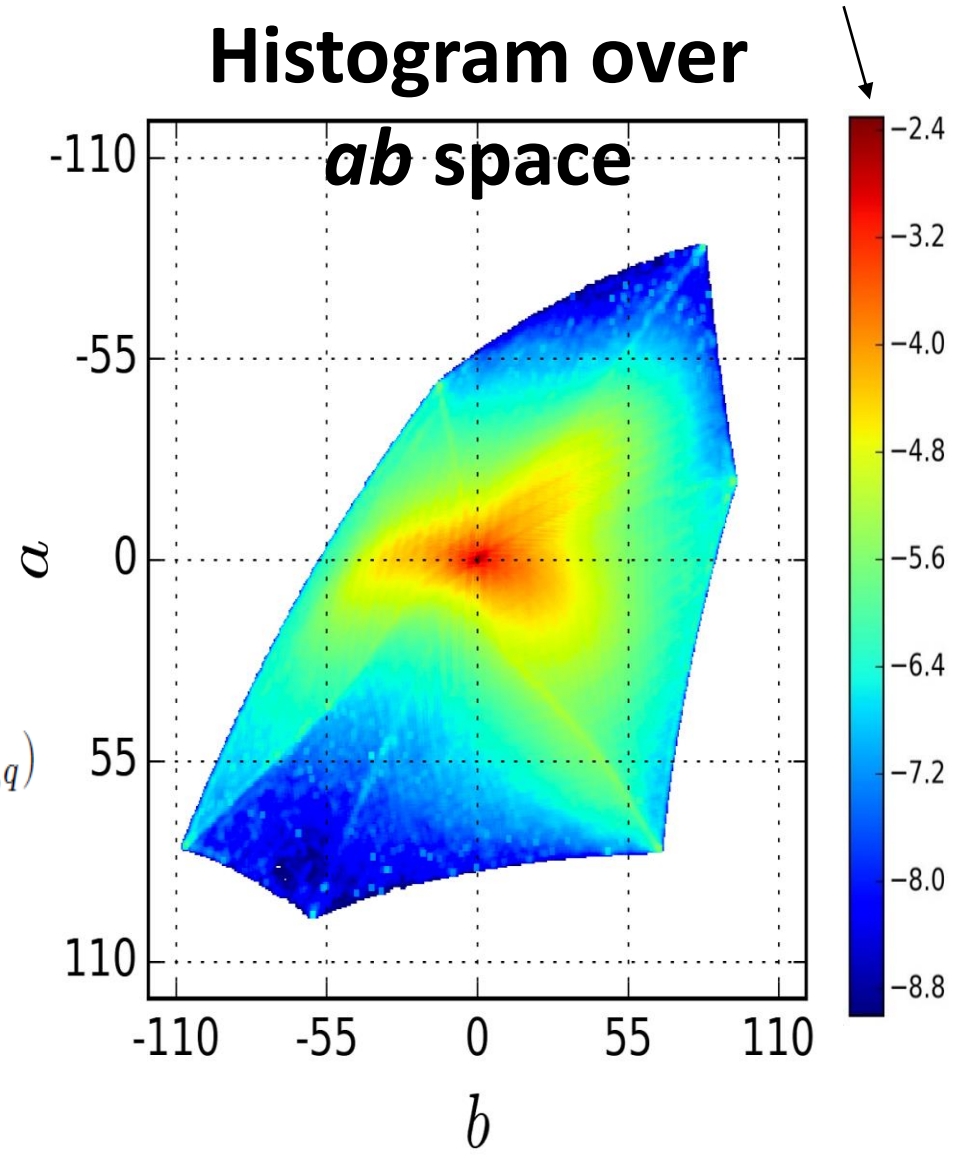
- Regression with L2 loss inadequate

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

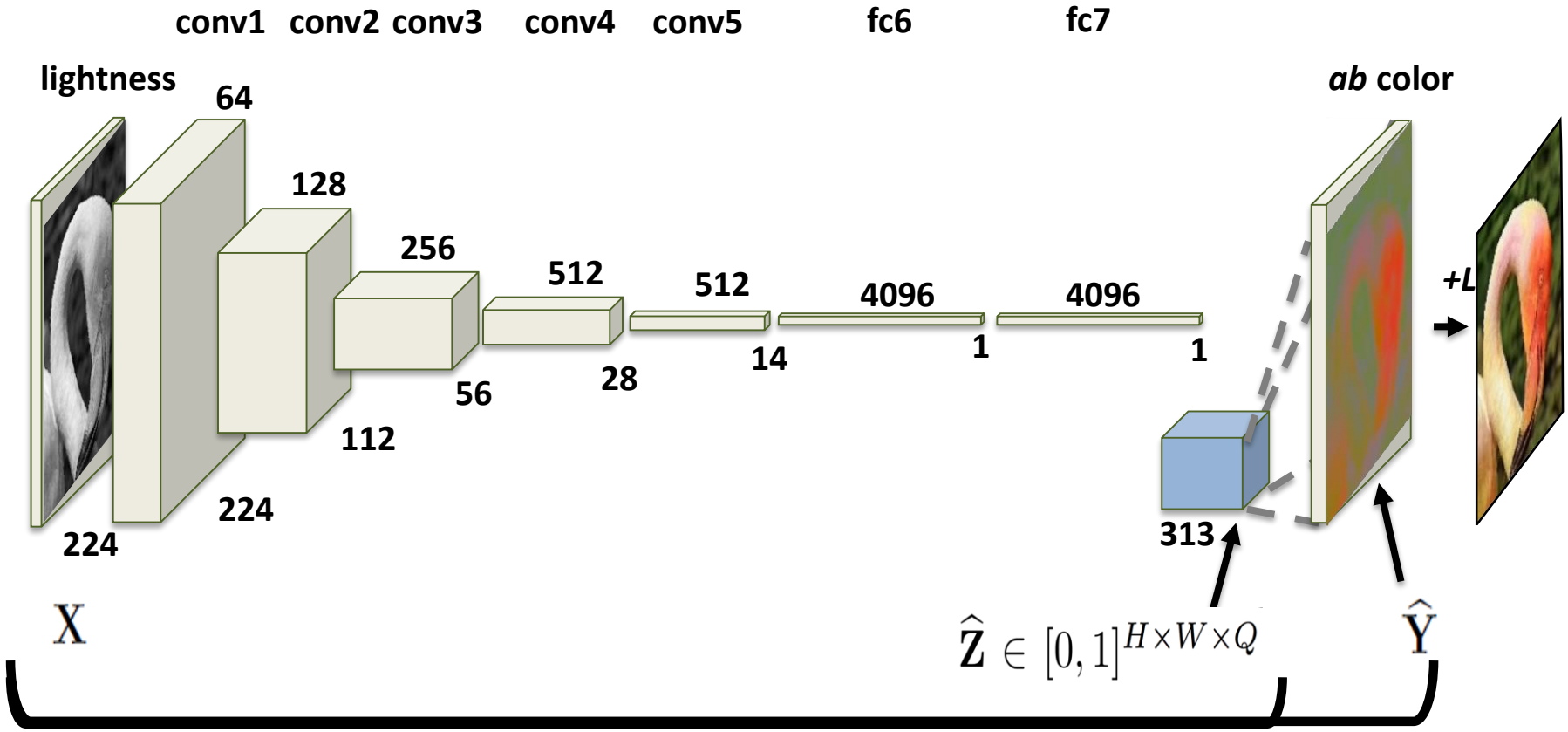
$$\mathrm{L}(\hat{\mathbf{Z}}, \mathbf{Z}) = -\frac{1}{HW}\sum_{h,w}\sum_{q} \mathbf{Z}_{h,w,q} \log(\hat{\mathbf{Z}}_{h,w,q})$$

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

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

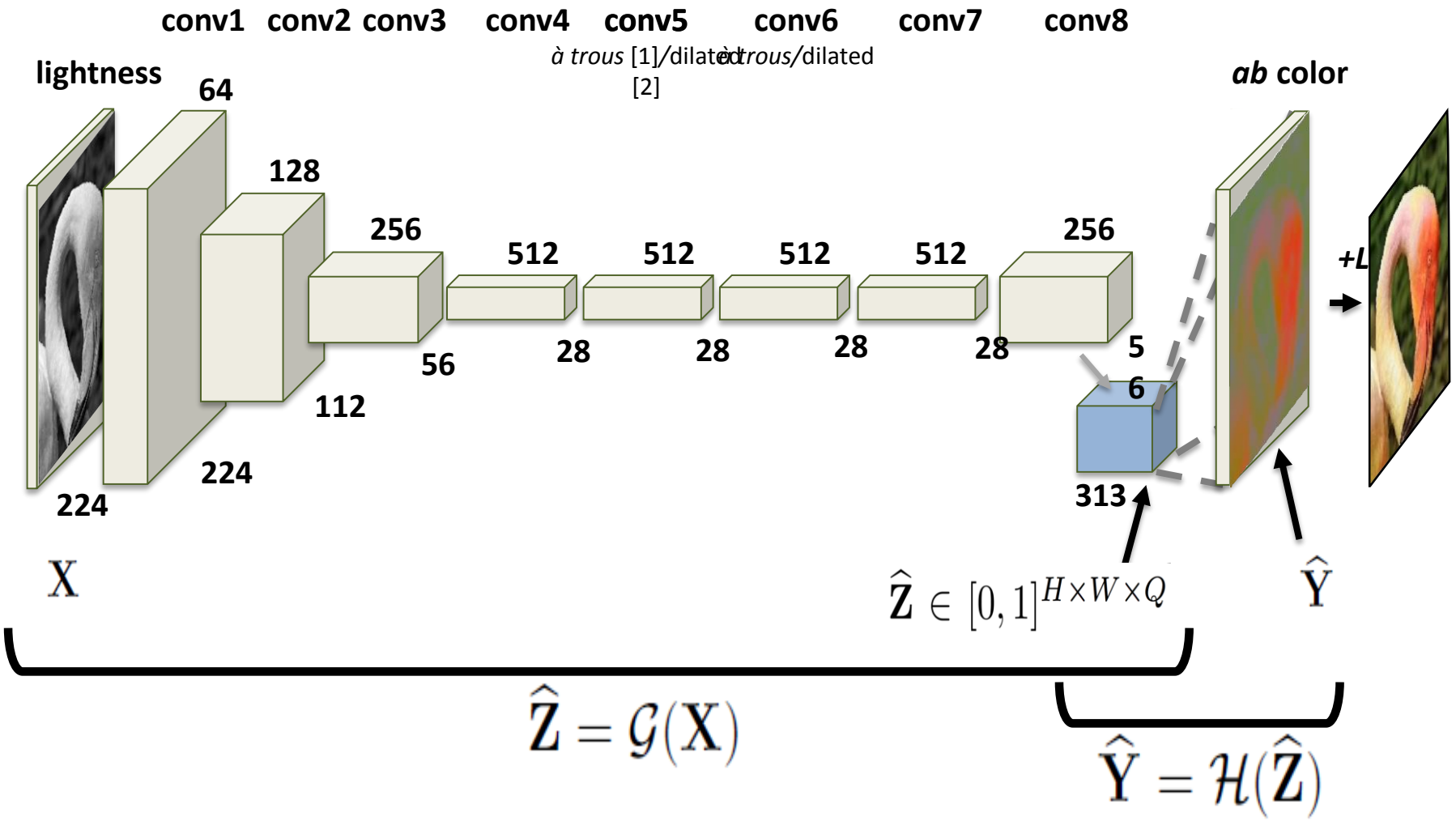


Network Architecture



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

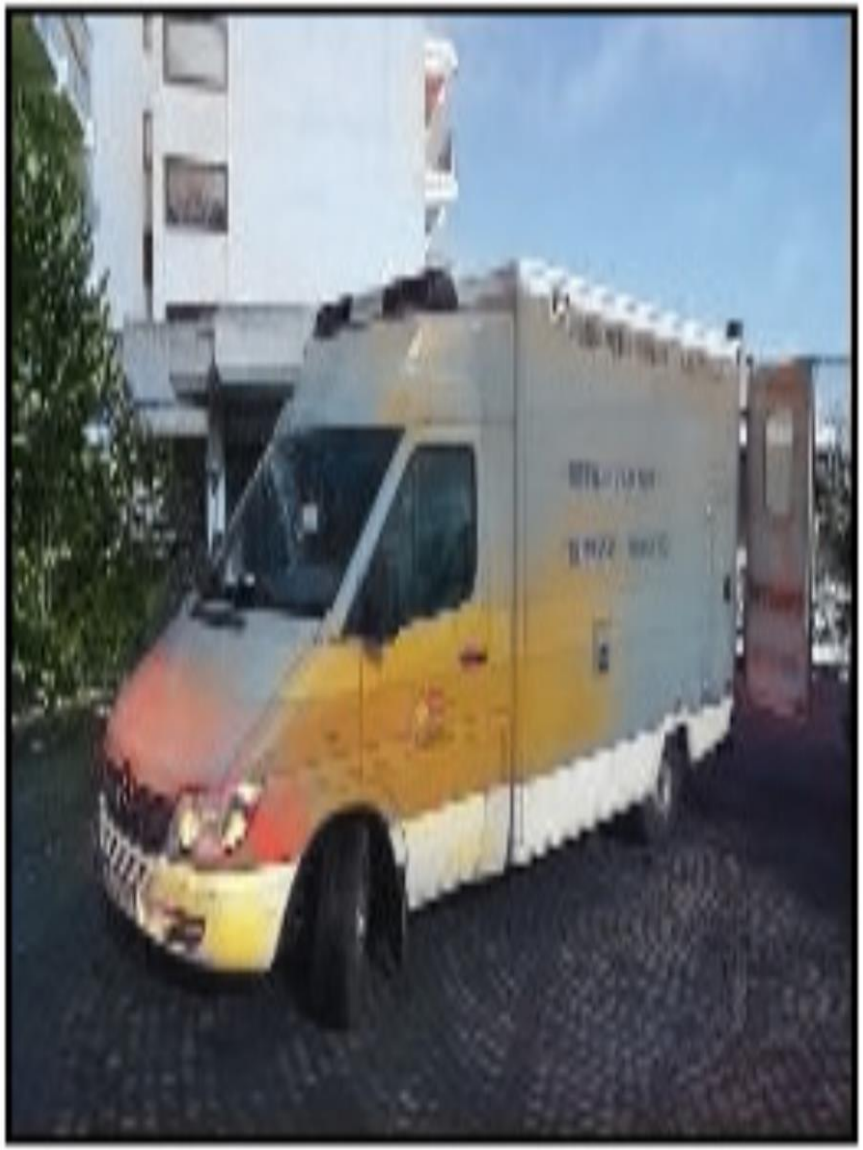
Network Architecture



Ground Truth L2 Regression Class w/ Rebalancing

Failure Cases

Biases



Fake, 0% fooled

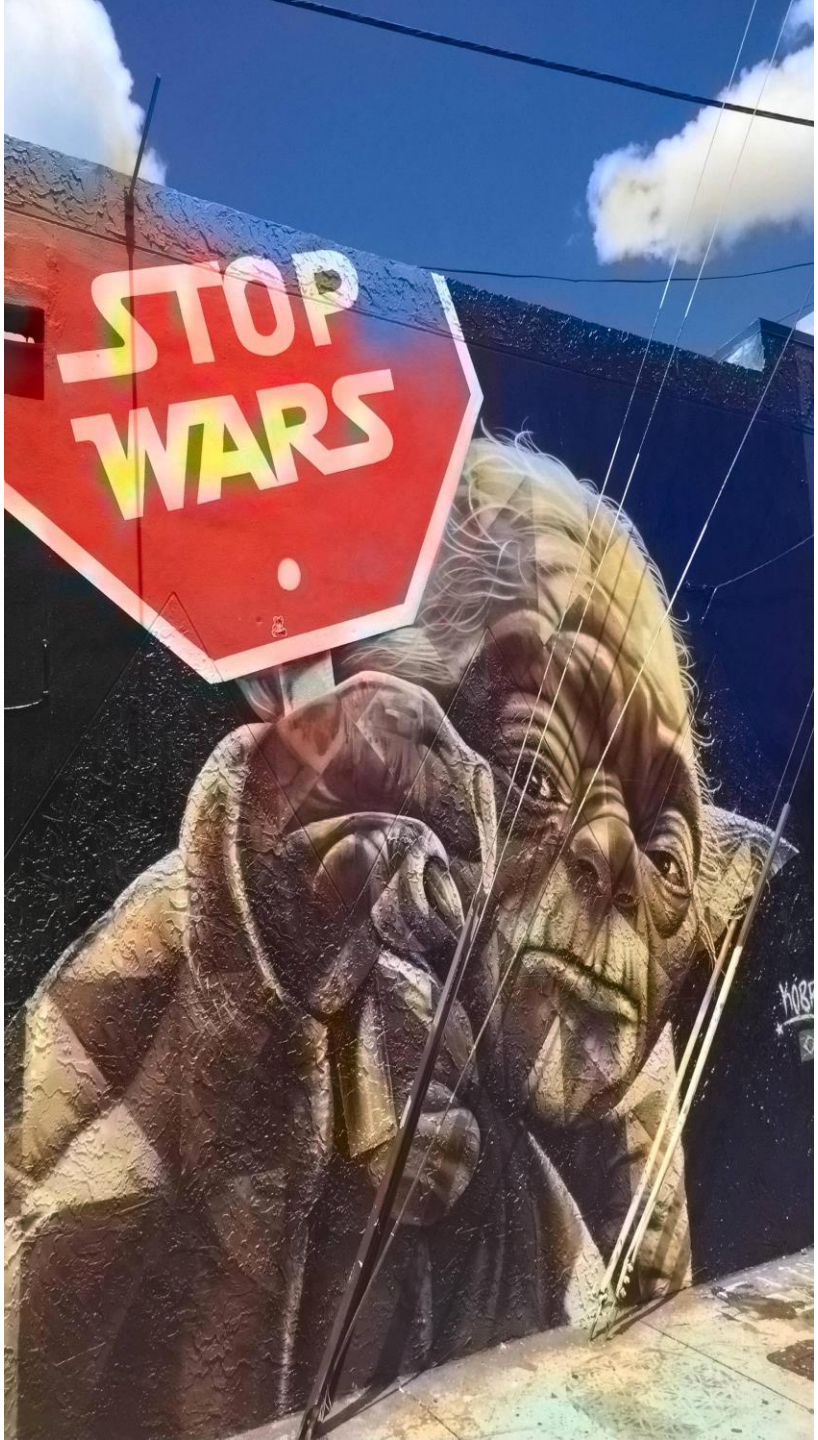
Fake, 55% fooled

Fake, 58% fooled

from Reddit /u/SherySantucci

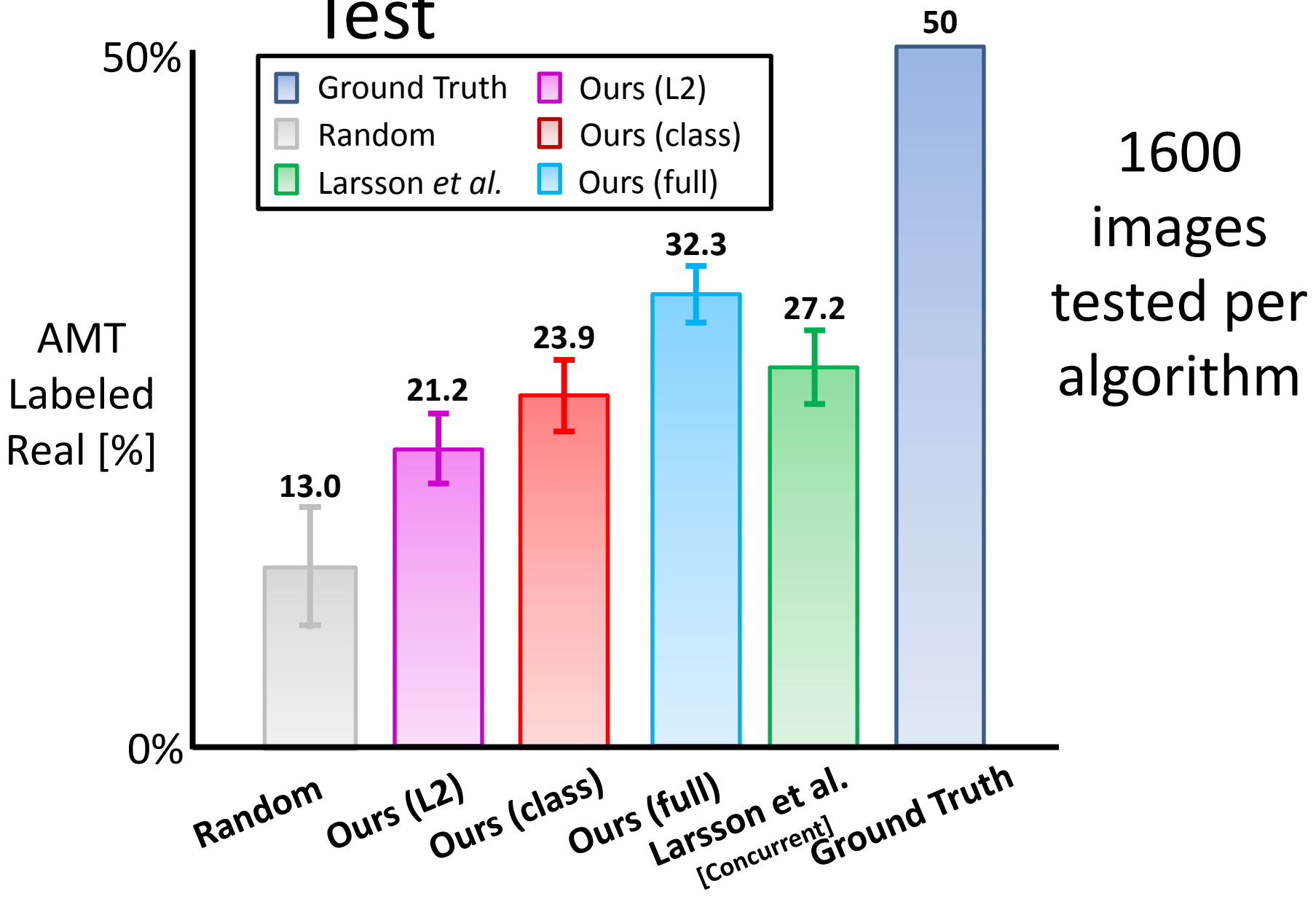
Recolorized by Reddit ColorizeBot

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



Recolorized by Reddit ColorizeBot

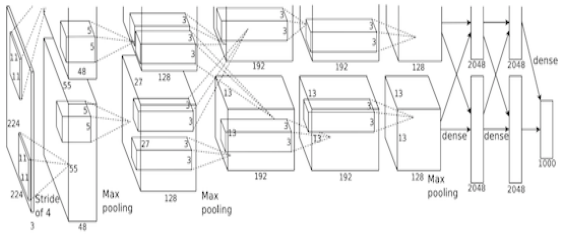
Perceptual Realism Test



1600 images tested per algorithm

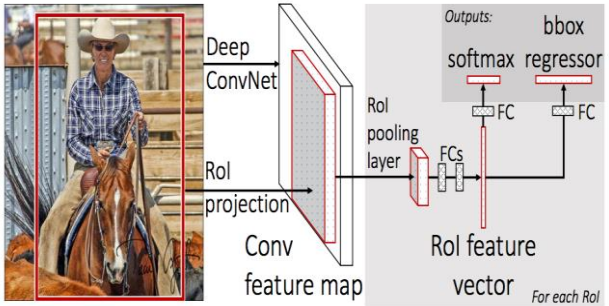
Dataset & Task Generalization on PASCAL VOC

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



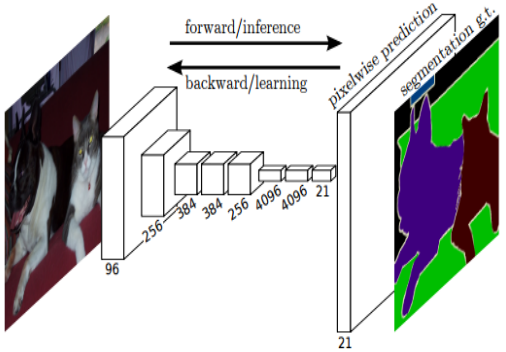
Classification

Krähenbühl et al. In ICLR, 2016.



Detection

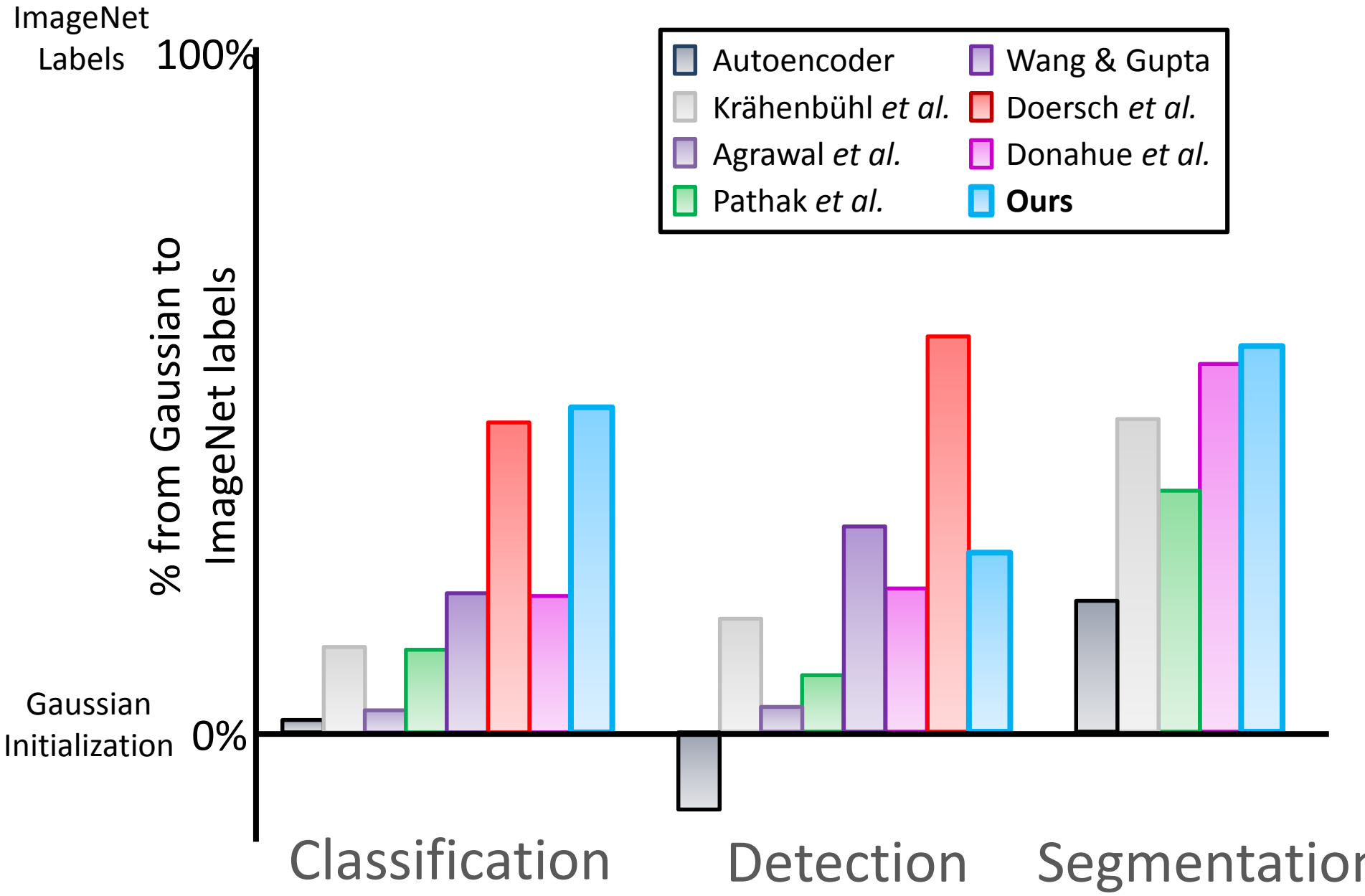
Fast R-CNN. Girshick. In ICCV, 2015.



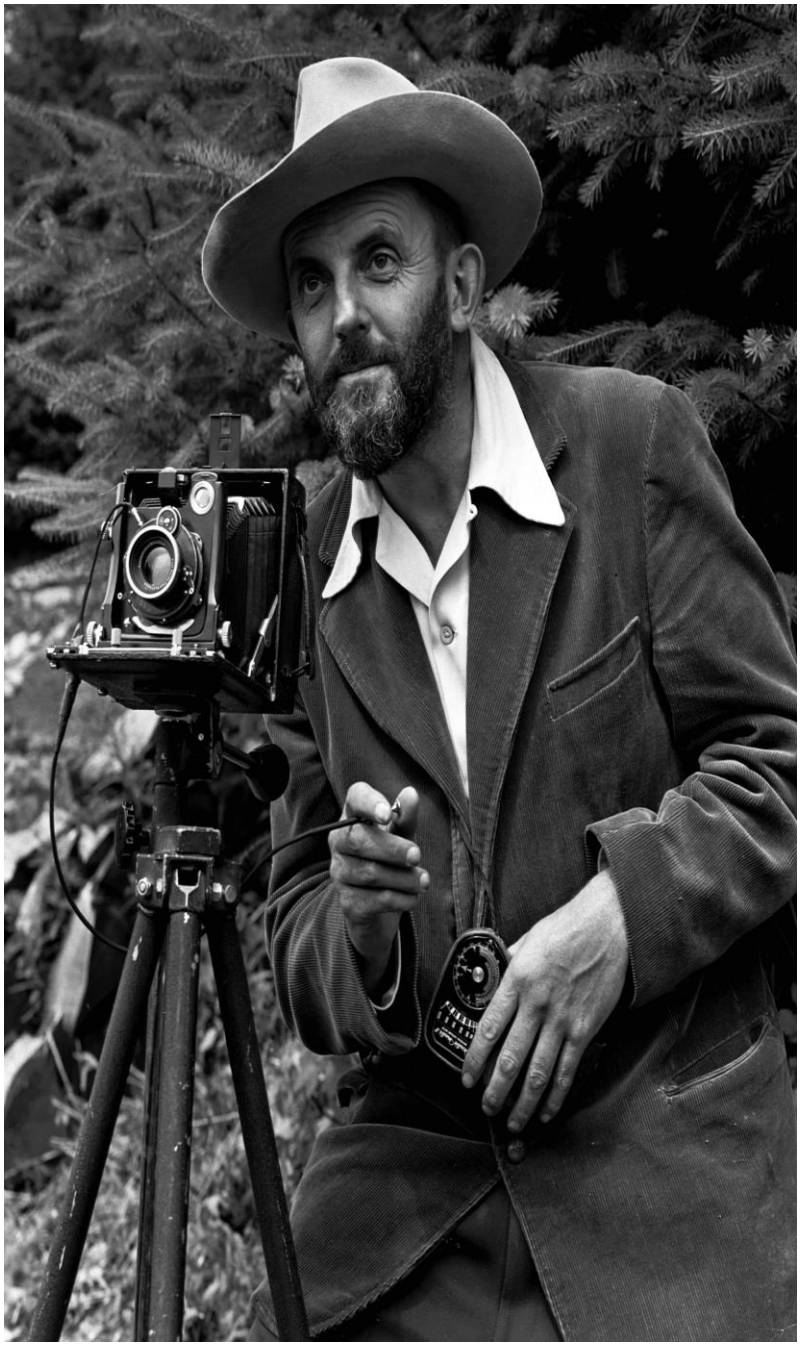
Segmentation

FCNs. Long et al. In CVPR, 2015.

Dataset & Task Generalization on PASCAL VOC

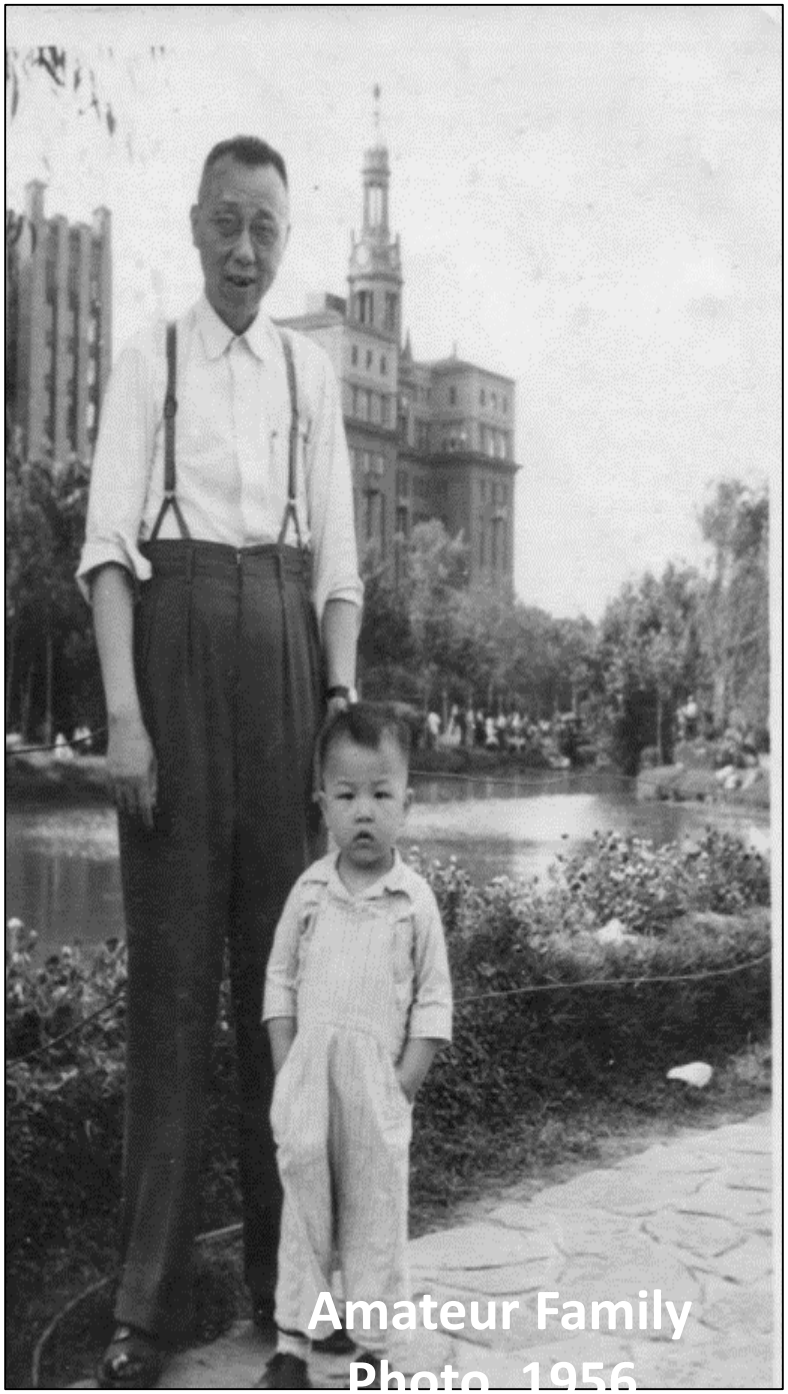


Does the method work on ***legacy*** black and white photos?




Thylacine, Dr. David Fleay,
extinct in 1936


Thylacine, Dr. David Fleay,
extinct in 1936



Amateur Family Photo, 1956

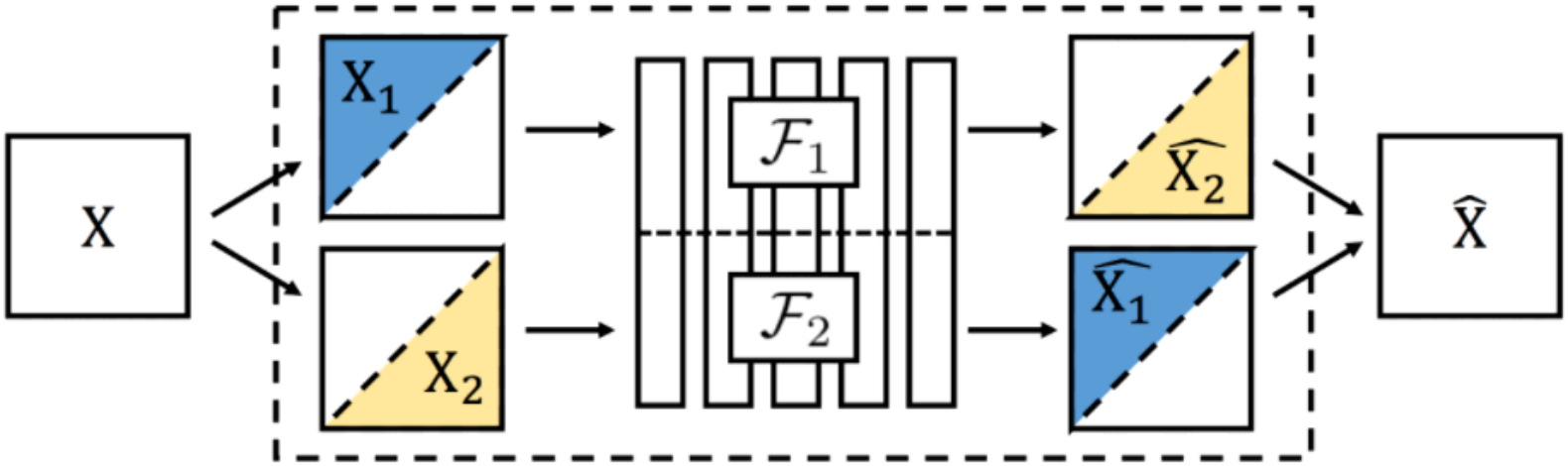
Amateur Family Photo, 1956

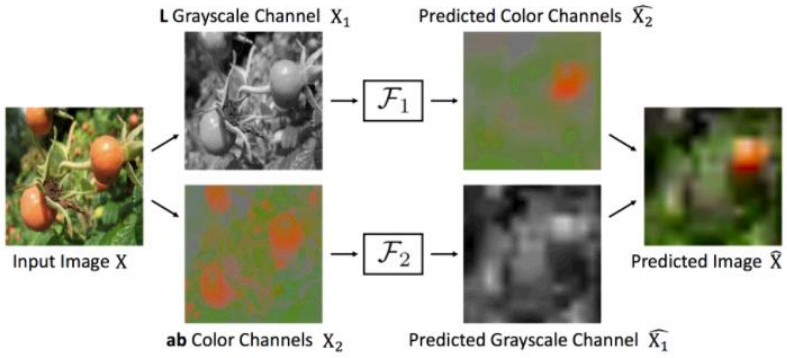
Henri Cartier-Bresson, Sunday on the Banks of the River Seine, 1938.

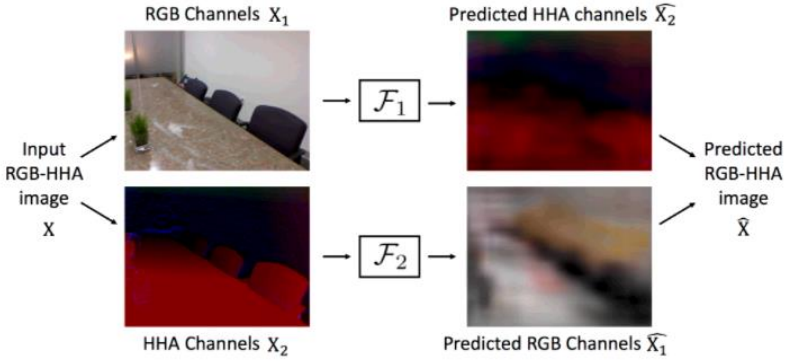
Henri Cartier-Bresson, Sunday on the Banks of the River Seine, 1938.

Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction

Richard Zhang Phillip Isola Alexei A. Efros
Department of EECS, University of California, Berkeley

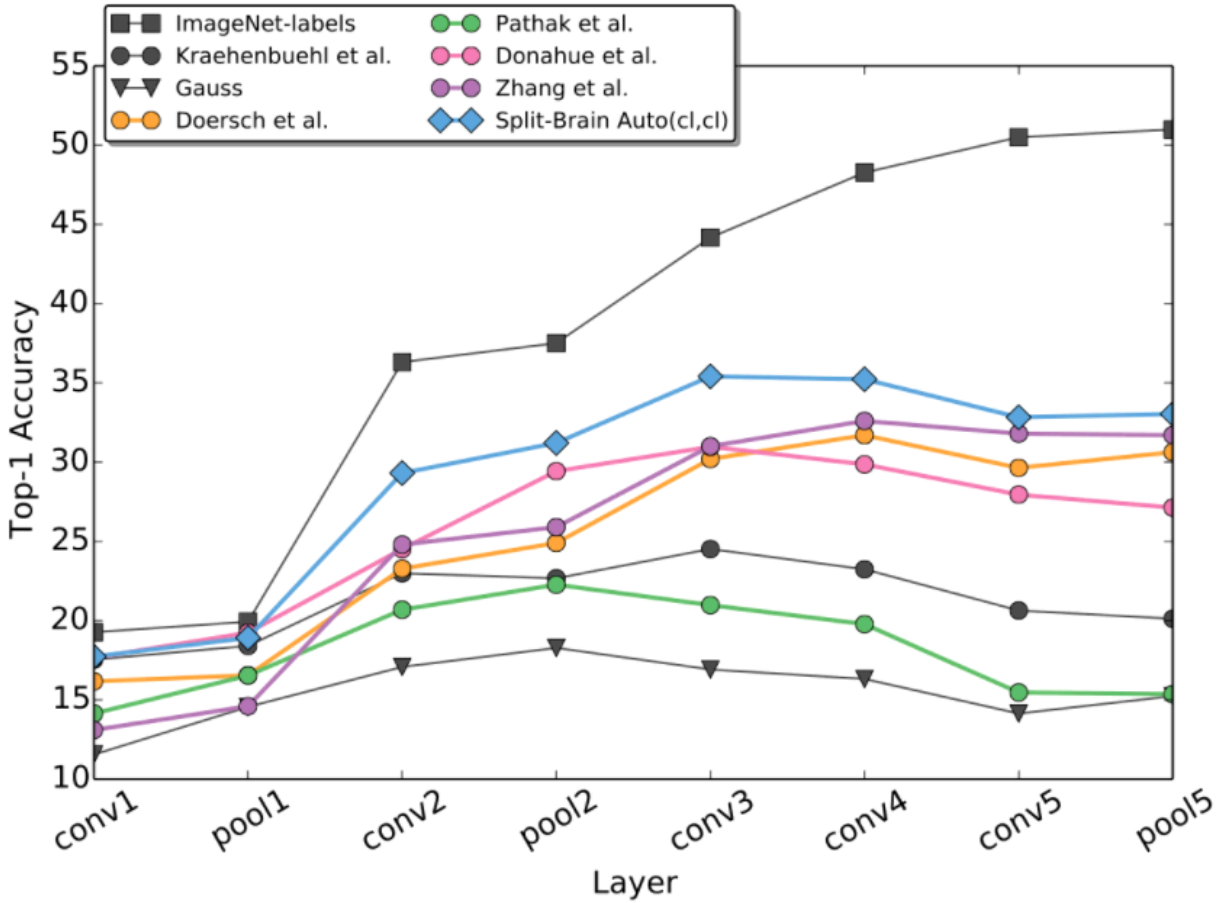






We show that we can learn features in an unsupervised framework, simply by predicting *channels of raw data* from other channels of raw data.

Feature Evaluation Results

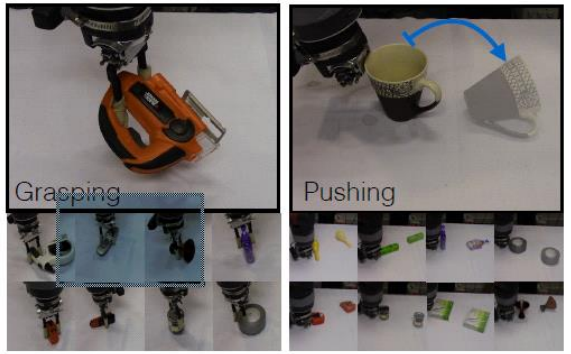


Self Supervision Examples

Context

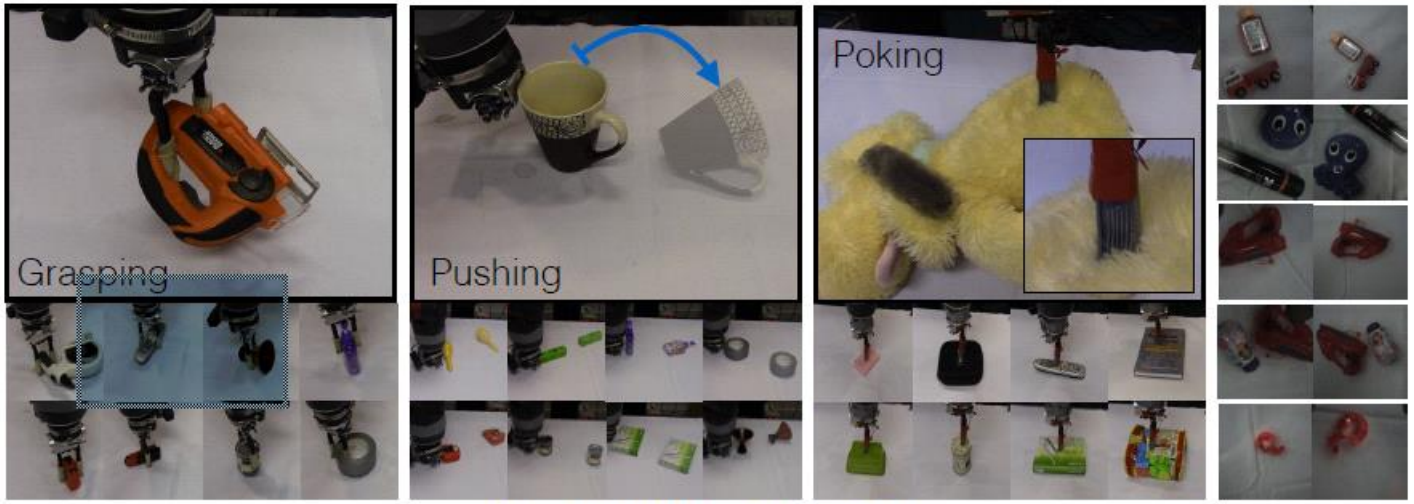
Color

Physical Interactions



Using **Physical Interactions** for Self Supervision

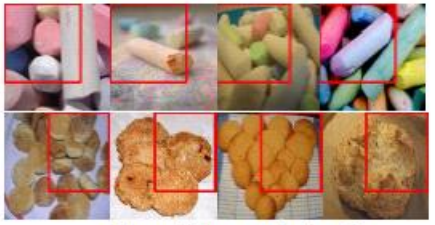
The Curious Robot: Learning Visual Representations via Physical Interactions



Physical Interaction Data

Conv Layer1 Filters

Conv3 Neuron Activations



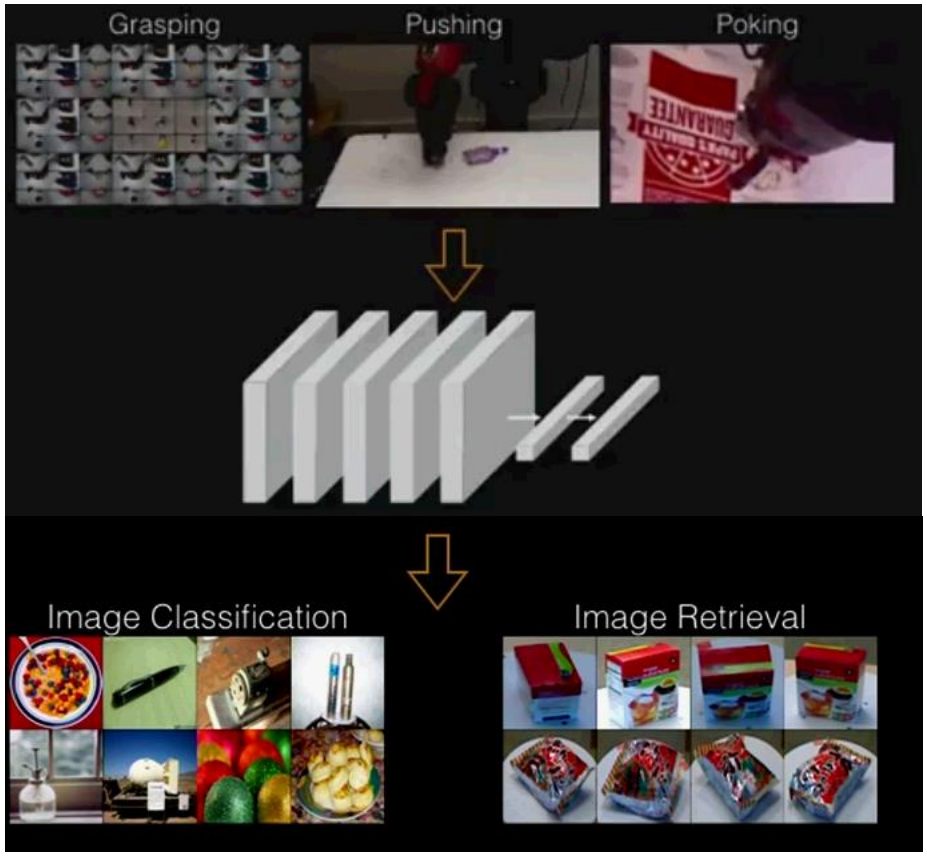
Conv5 Neuron Activations

Learned Visual Representation

Learning in Humans

- Use Actions and supervision effects on these actions to train our representations
- Build a similar idea
- Use a robot to learn visual features

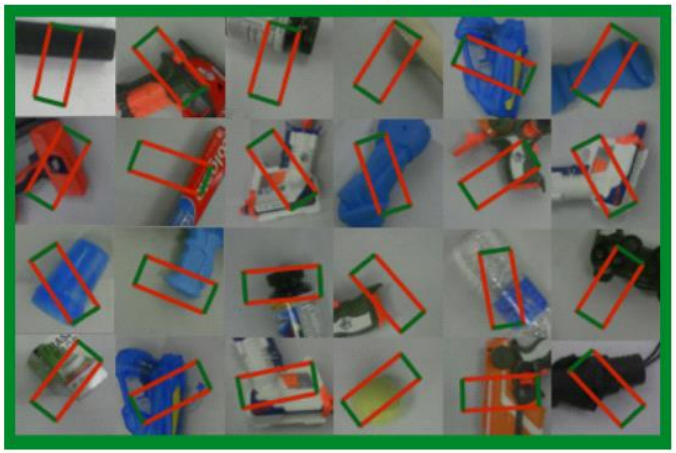
Using robot tasks to learn visual features



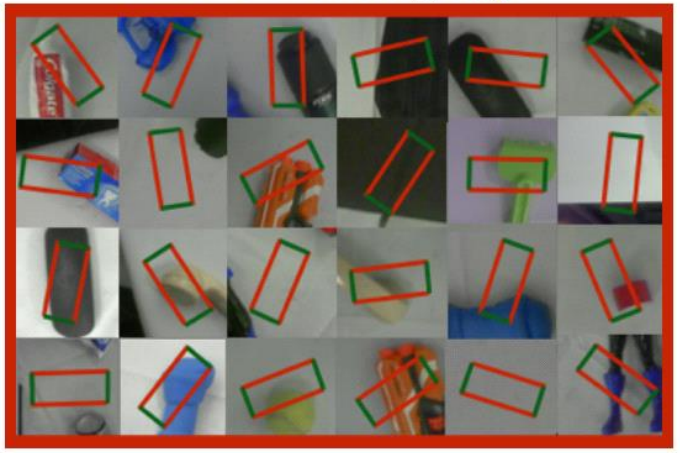
Grasping Task

- Robot continuously trying to grasp objects
- Grasp configuration lies in 3 dimensions:
  - position of grasp point on the surface of table
  - angle of grasp
- Record successful and unsuccessful grasps

Successful grasps

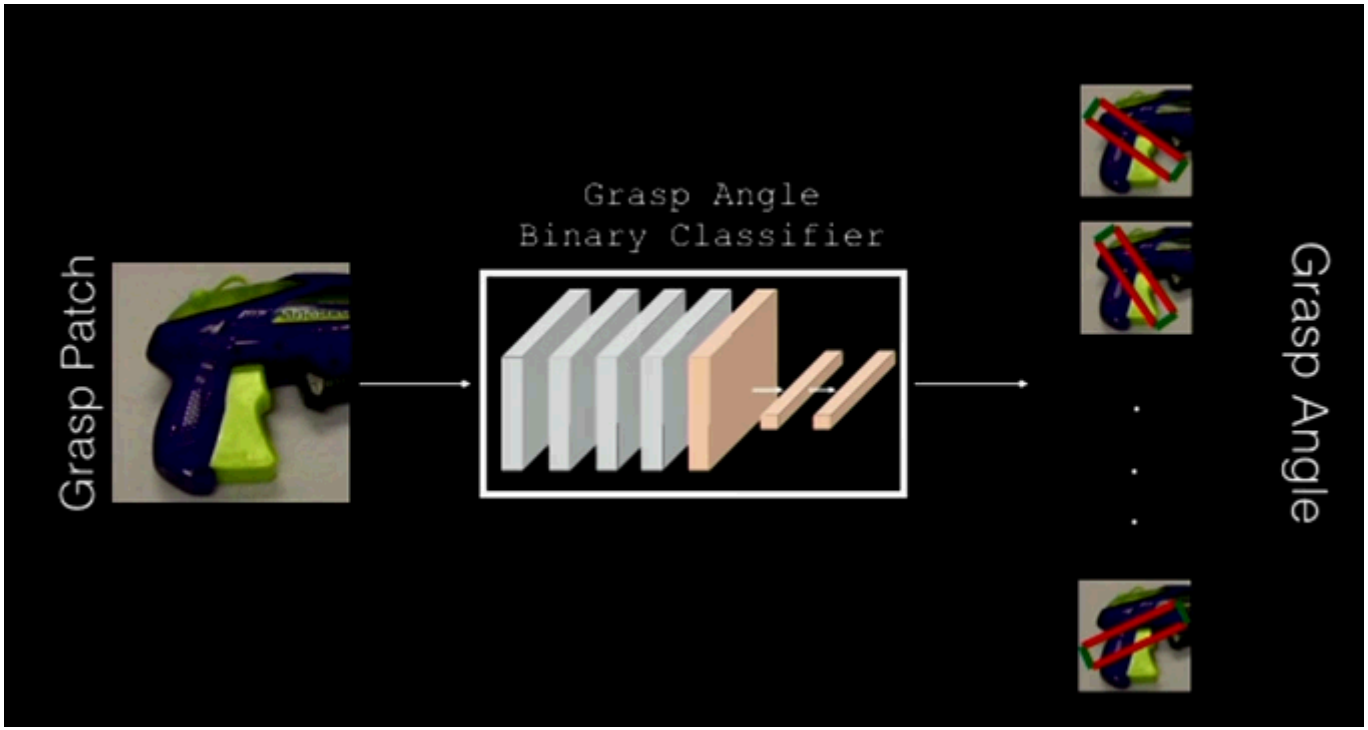


Unsuccessful grasps

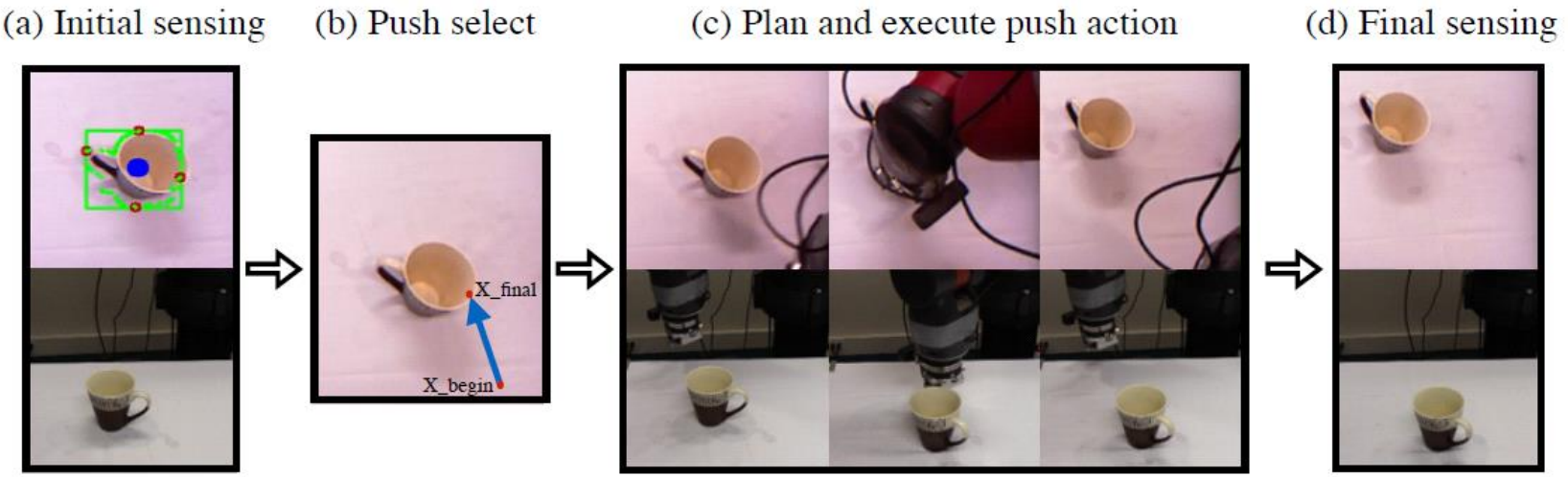


https://www.youtube.com/watch?v=oSqHc0nLkm8

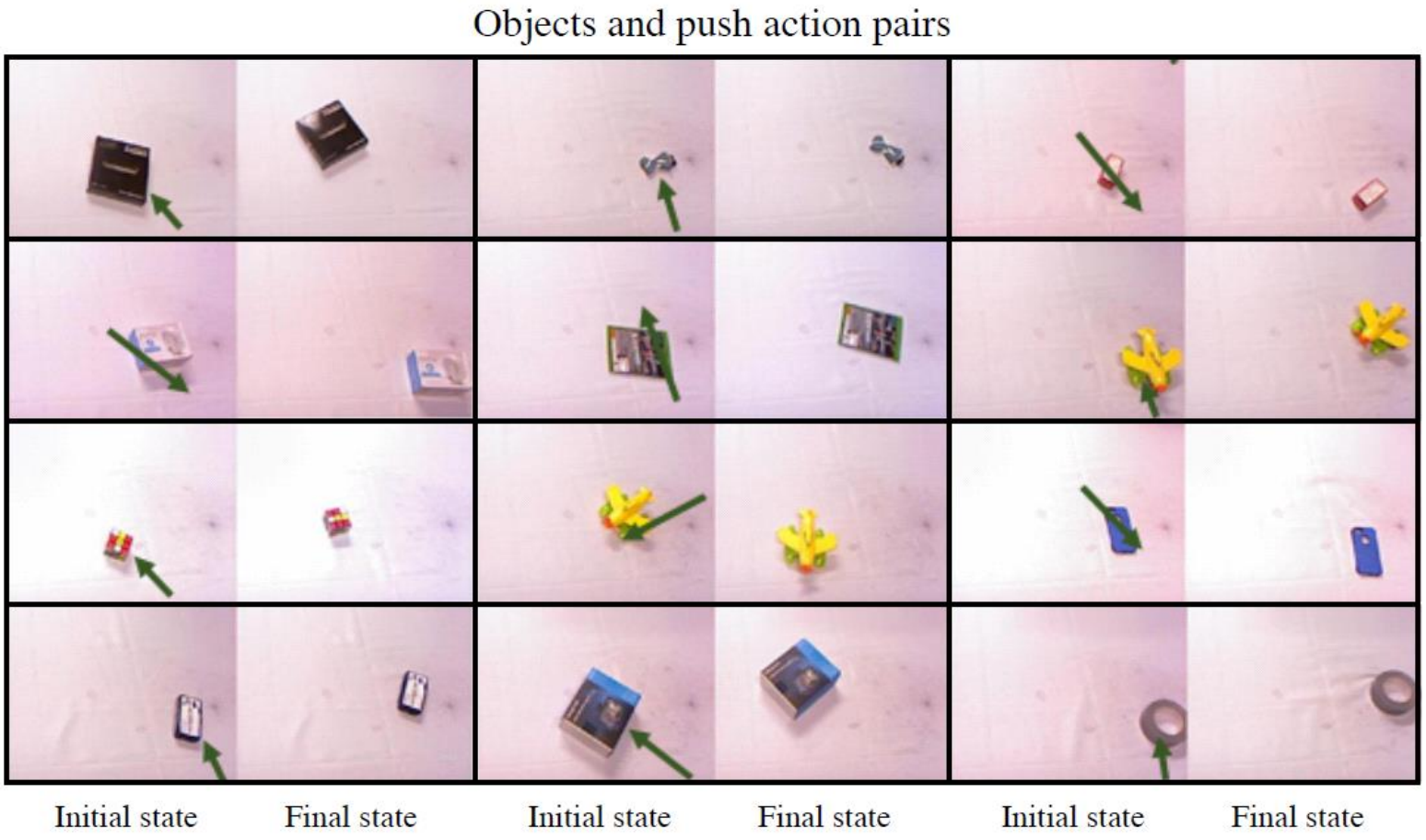
Grasping Formulation



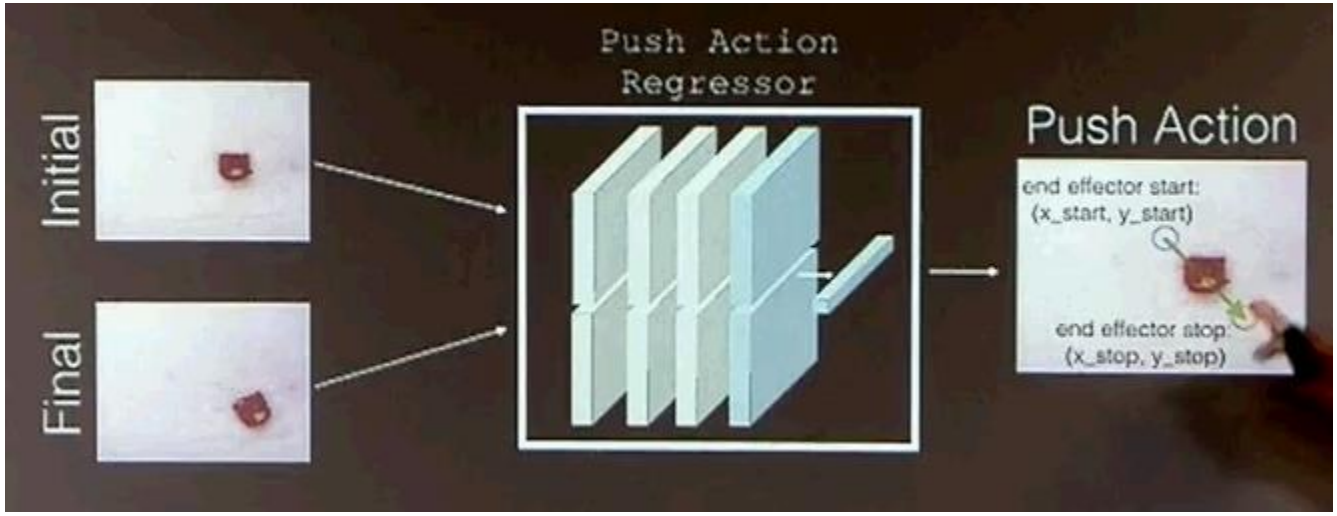
Pushing Task



Pushing Task

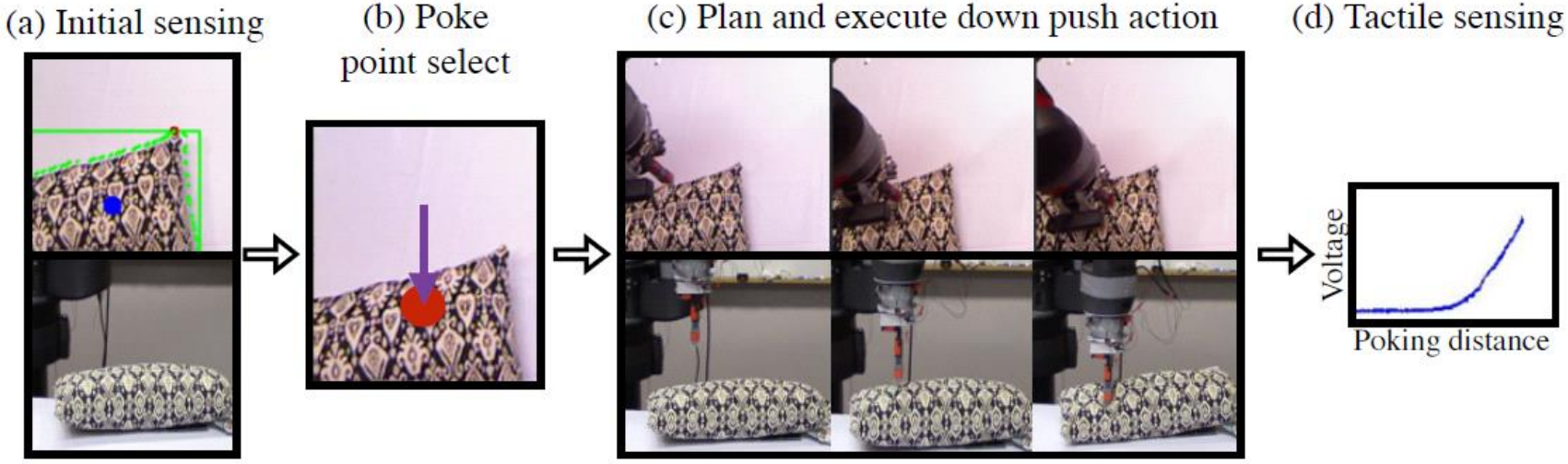


Planar Pushing Formulation

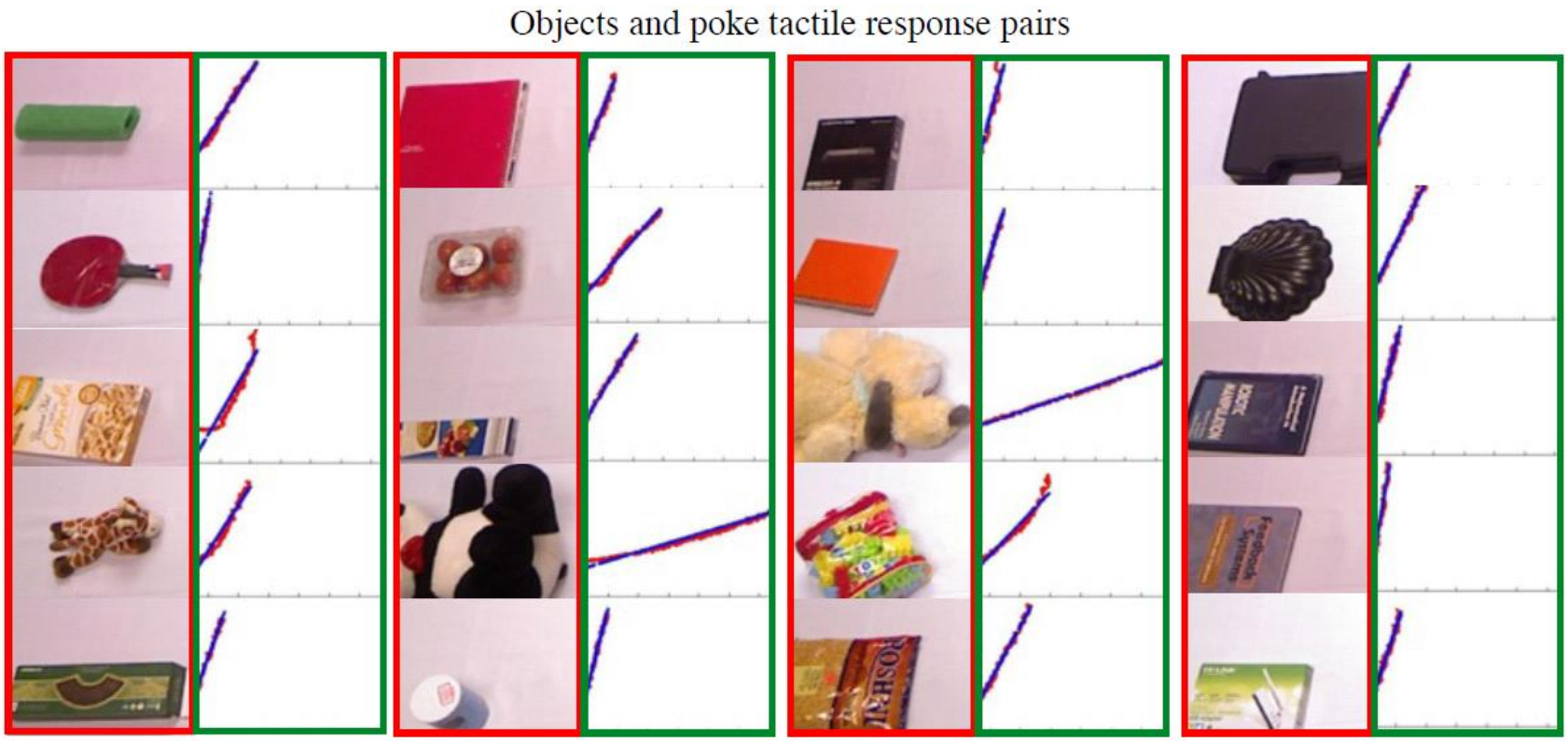


Poking Task

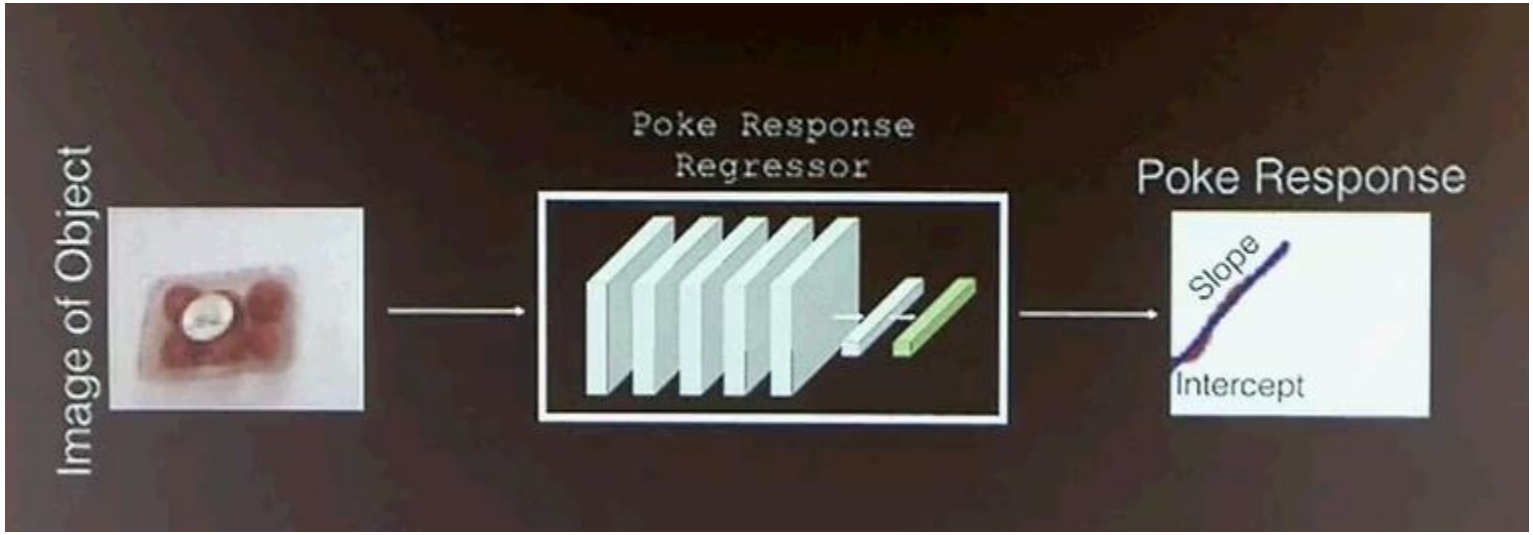
Learning Visual Representations via Physical Interactions



Poking Task

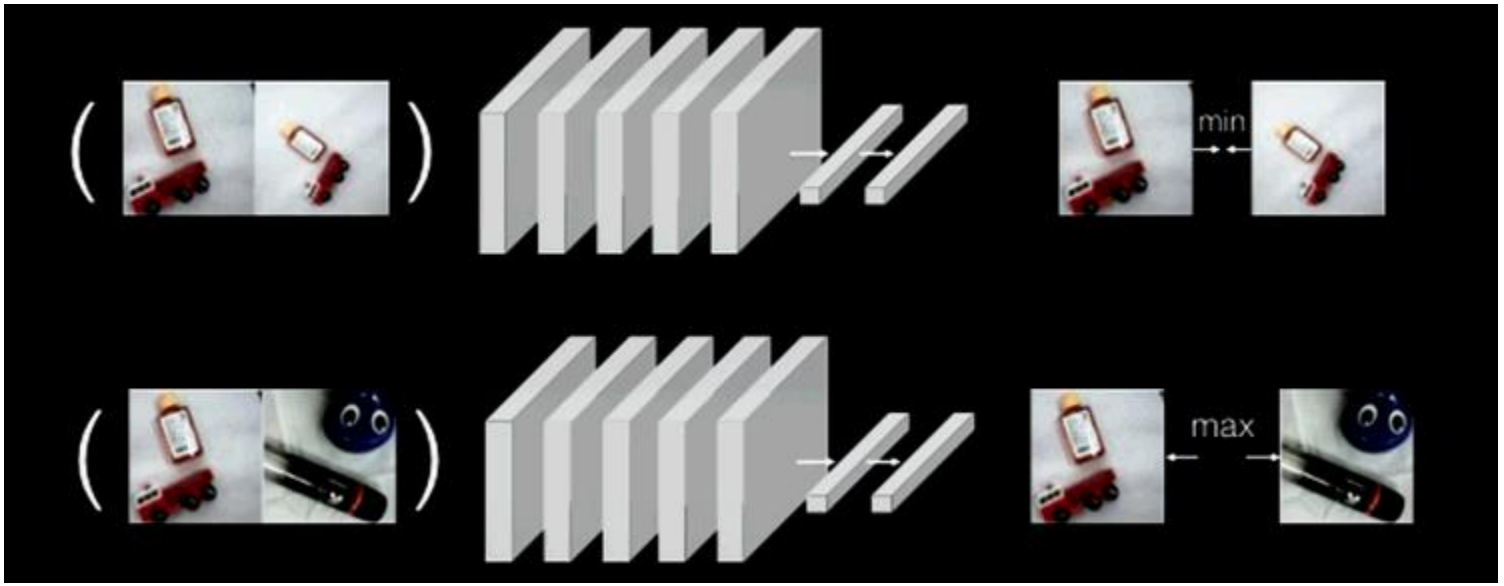


Poking Formulation

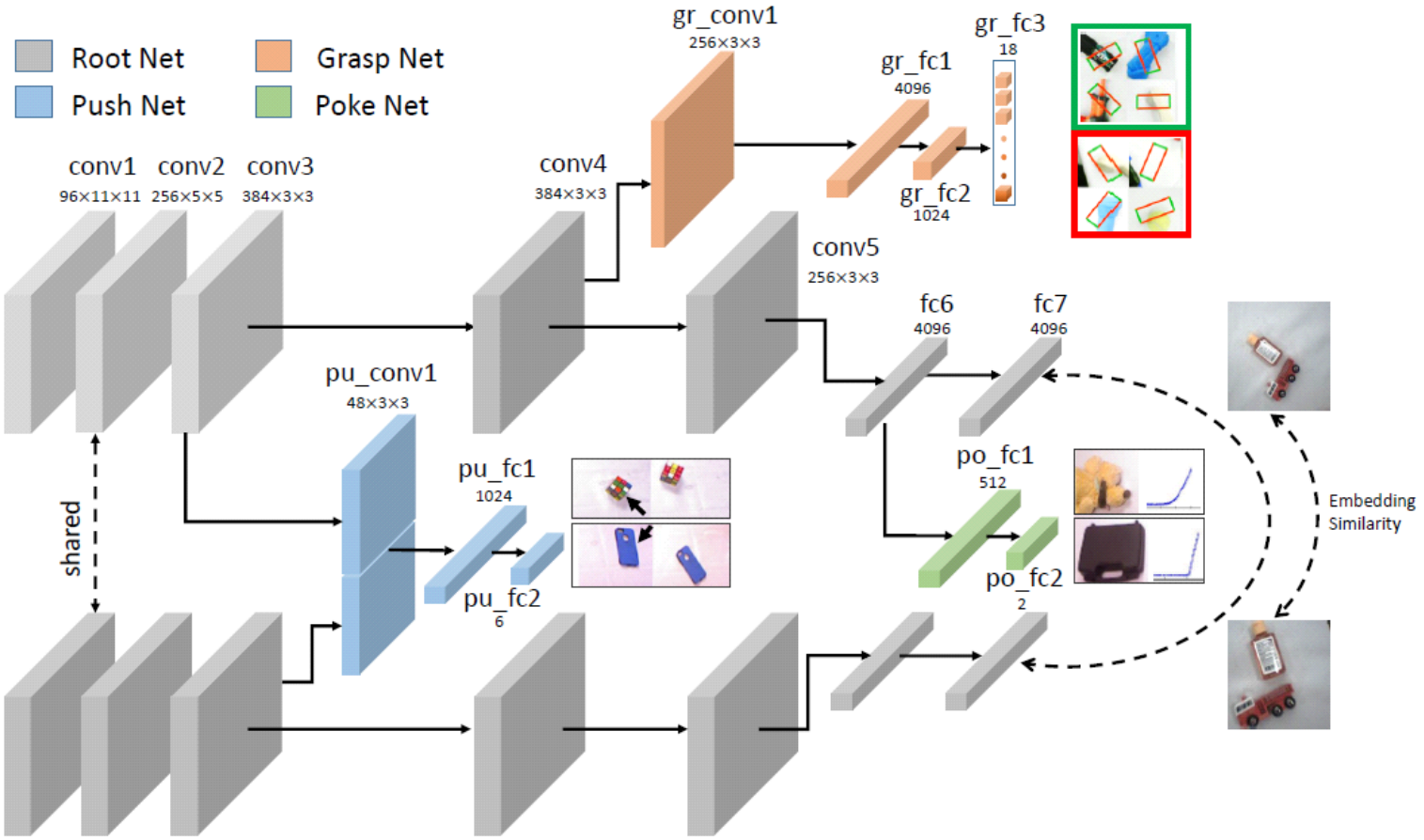


Pose and Scale Invariance

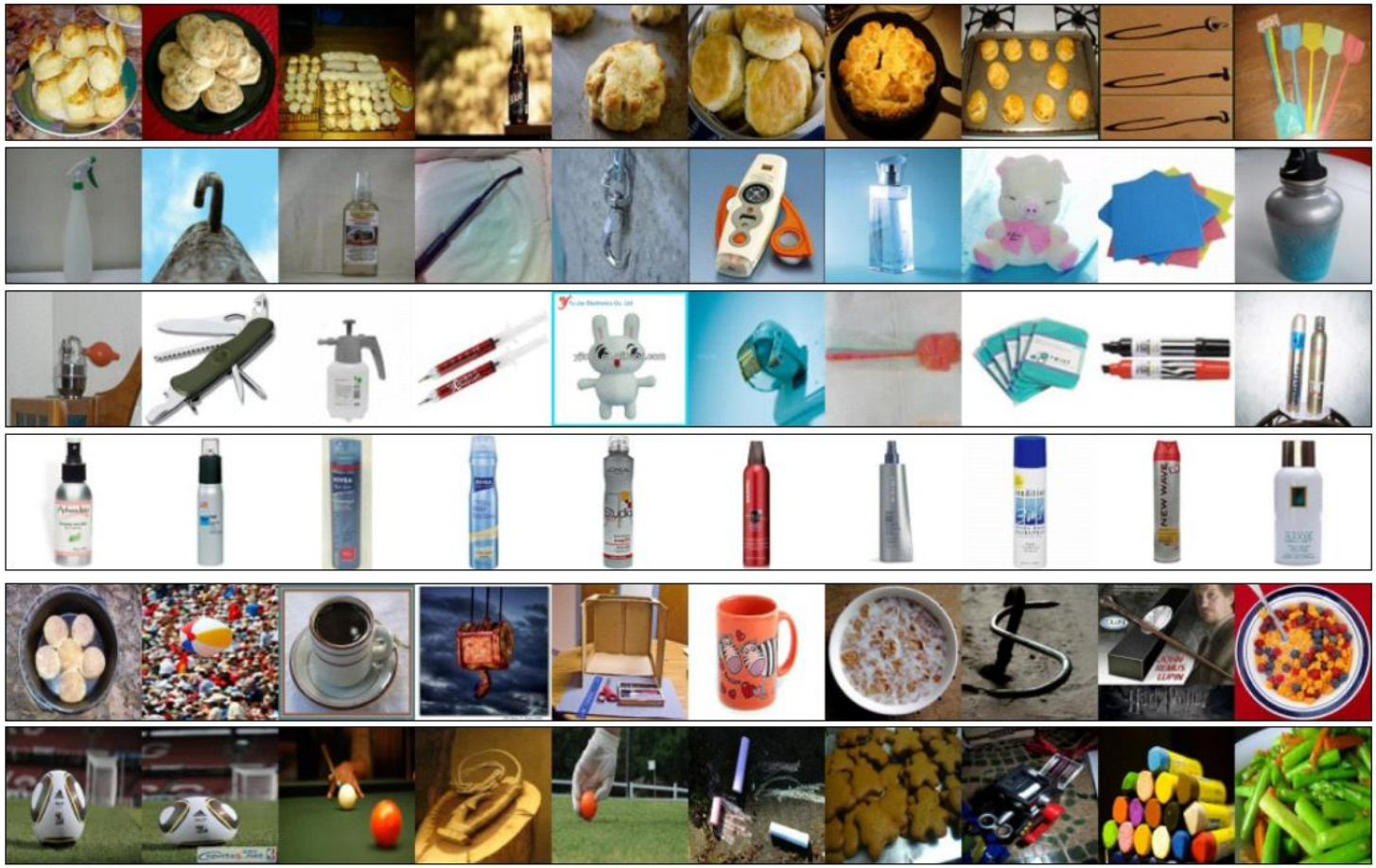
Pose and Scale Invariance Formulation



Convolutional Architecture



Nearest Neighbors



Classification Accuracy

	Household	UW RGBD	Caltech-256
Root network with random init.	0.250	0.468	0.242
Root network trained on robot tasks (**ours**)	0.354	0.693	0.317
AlexNet trained on ImageNet	0.625	0.820	0.656
Root network trained on identity data	0.315	0.660	0.252
Auto-encoder trained on all robot data	0.296	0.657	0.280

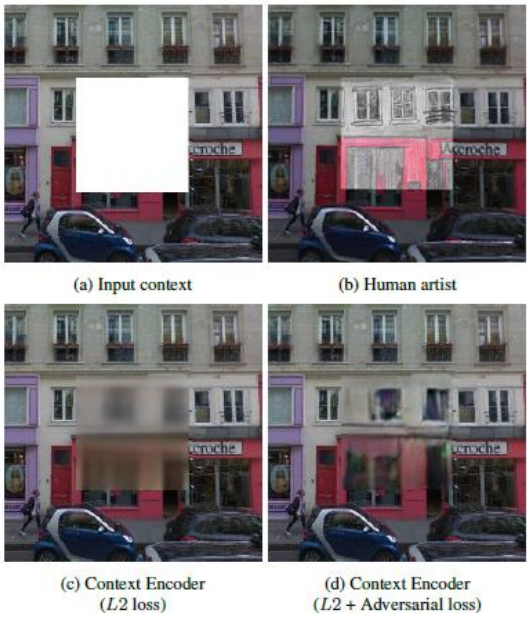
Task Ablation Analysis

	Household	UW RGB-D	Caltech-256
All robot tasks	0.354	0.693	0.317
Except Grasp	0.309	0.632	0.263
Except Push	0.356	0.710	0.279
Except Poke	0.342	0.684	0.289
Except Identity	0.324	0.711	0.297

Drone: Real World Poking

Self Supervision Examples

Context

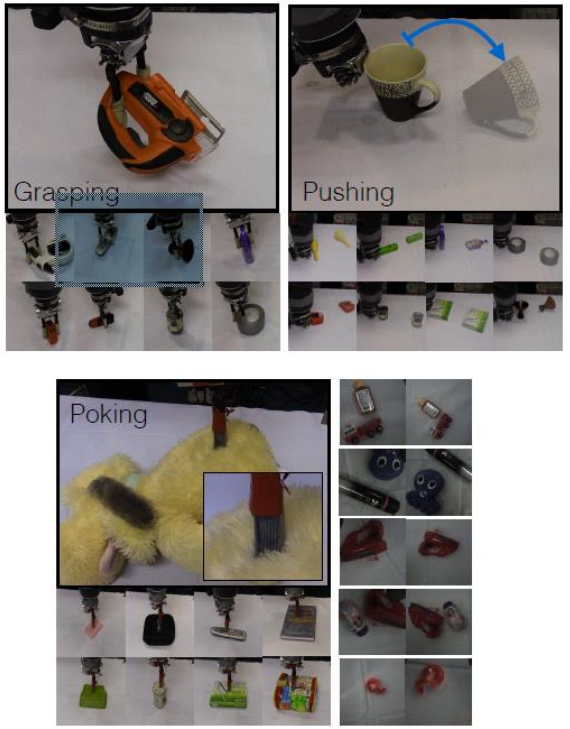


(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

Color

Physical Interactions



Thank You