CS 1678/2078: Intro to Deep Learning

Self-supervised Learning

Prof. Adriana Kovashka
University of Pittsburgh
March 25, 2024

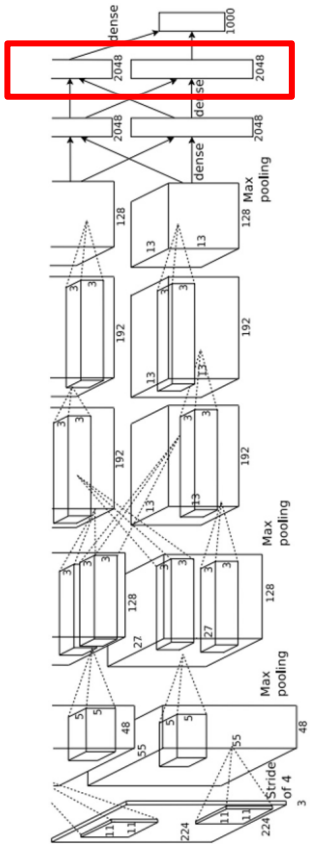
Plan for this lecture

- What is it? Learning representations from context in raw data
- Language – predict nearby words [*already covered]*
  - Word2Vec, transformers, BERT
- Vision – predict pixels from other pixels
  - Predict nearby patches in an image
  - Predict order of frames in a video
  - Predict what you will see as you move
  - Predict physics
  - More general formulation – contrastive learning

Jitendra Malik: "**Supervision** is the opium of the AI researcher"
Alyosha Efros: "The AI revolution will not be **supervised**"
Yann LeCun: "**Self-supervised** learning is the cake, **supervised** learning is the icing on the cake, **reinforcement learning** is the cherry on the cake"

Learned Representations

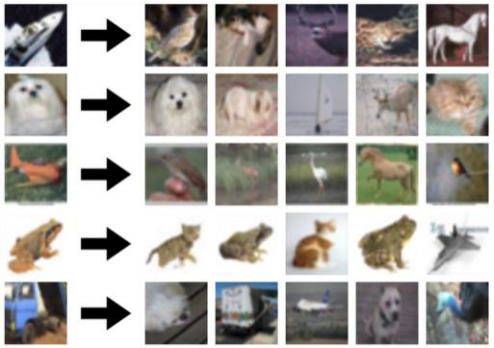
4096-dim vector



Test image

L2 Nearest neighbors in <u>feature</u> space

Recall: Nearest neighbors in <u>pixel</u> space



Krizhevsky et al, "ImageNet Classification with Deep Convolutional Neural Networks", NIPS 2012.
Figures reproduced with permission.

Self-supervised Learning

- Aims to learn from data without manual label annotation.
- Self-supervised learning methods solve “pretext” tasks that produce **good features** for downstream tasks.
  - Learn with supervised learning objectives, e.g., classification, regression.
  - Labels of these pretext tasks are generated *automatically*

Self-supervised pretext tasks

Example: learn to predict image transformations / complete corrupted images

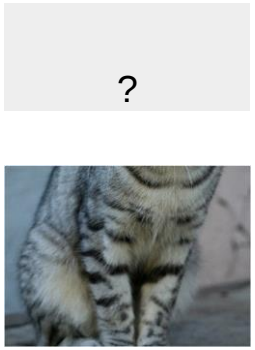
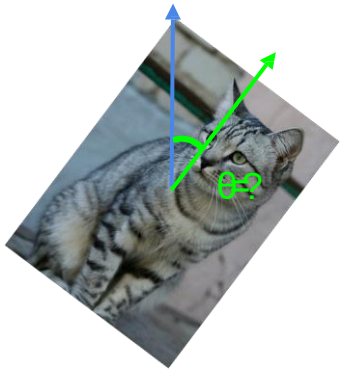


image completion



rotation prediction

“jigsaw puzzle”

colorization

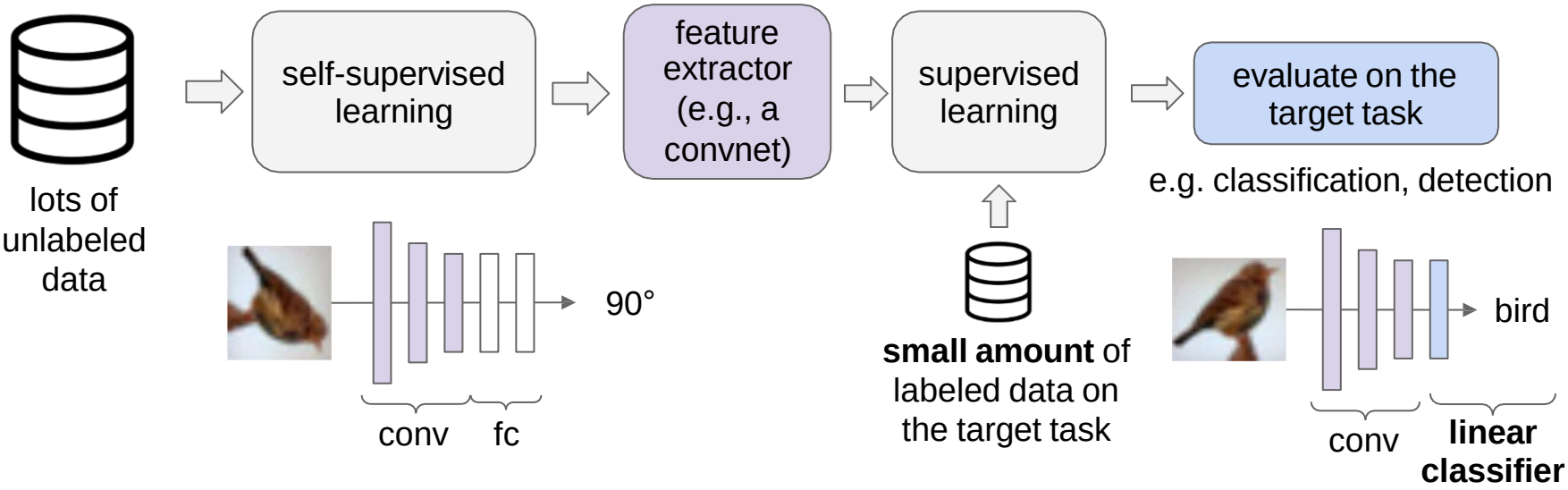
1. Solving the pretext tasks allow the model to learn good features.
2. We can automatically generate labels for the pretext tasks.

How to evaluate a self-supervised learning method?

We usually don't care about the performance of the self-supervised learning task, e.g., we don't care if the model learns to predict image rotation perfectly.

Evaluate the learned feature encoders on downstream *target tasks*

How to evaluate a self-supervised learning method?



1. Learn good feature extractors from self-supervised pretext tasks, e.g., predicting image rotations

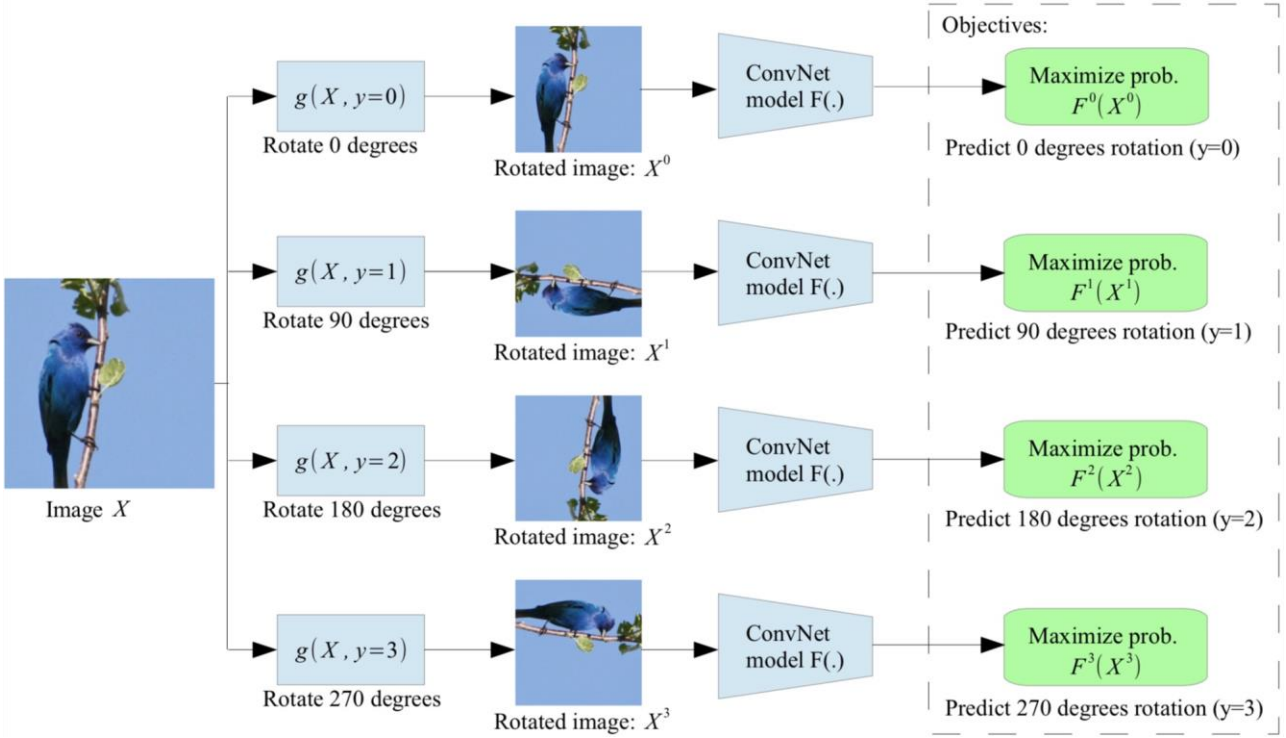
2. Attach a shallow network on the feature extractor; train the shallow network on the target task with small amount of labeled data

Pretext task: predict rotations

Hypothesis: a model could recognize the correct rotation of an object only if it has the "visual commonsense" of what the object should look like unperturbed.

(Image source: Gidaris et al. 2018)

Pretext task: predict rotations

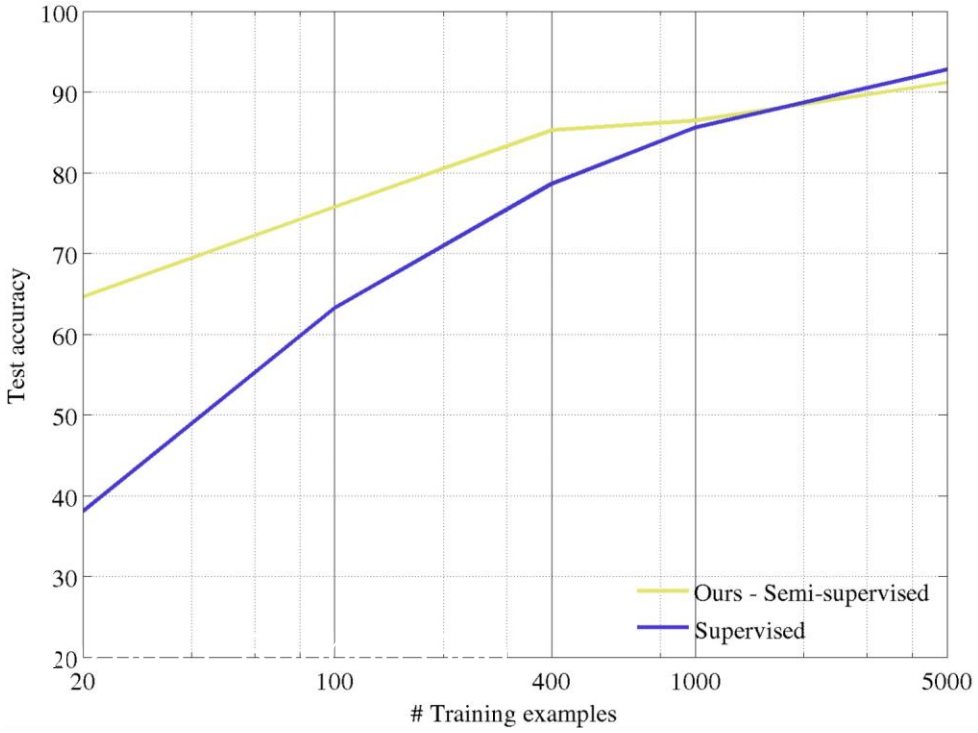


Self-supervised learning by rotating the entire input images.

The model learns to predict which rotation is applied (4-way classification)

(Image source: Gidaris et al. 2018)

Evaluation on semi-supervised learning



Self-supervised learning on **CIFAR10** (entire training set).

Freeze conv1 + conv2
Learn **conv3 + linear** layers with subset of labeled CIFAR10 data (classification).

(Image source: Gidaris et al. 2018)

Transfer learned features to supervised learning

	Classification (%mAP)		Detection (%mAP)	Segmentation (%mIoU)
Trained layers	fc6-8	all	all	all
ImageNet labels	78.9	79.9	56.8	48.0
Random		53.3	43.4	19.8
Random rescaled Krähenbühl et al. (2015)	39.2	56.6	45.6	32.6
Egomotion (Agrawal et al., 2015)	31.0	54.2	43.9	
Context Encoders (Pathak et al., 2016b)	34.6	56.5	44.5	29.7
Tracking (Wang & Gupta, 2015)	55.6	63.1	47.4	
Context (Doersch et al., 2015)	55.1	65.3	51.1	
Colorization (Zhang et al., 2016a)	61.5	65.6	46.9	35.6
BIGAN (Donahue et al., 2016)	52.3	60.1	46.9	34.9
Jigsaw Puzzles (Noroozi & Favaro, 2016)	-	67.6	53.2	37.6
NAT (Bojanowski & Joulin, 2017)	56.7	65.3	49.4	
Split-Brain (Zhang et al., 2016b)	63.0	67.1	46.7	36.0
ColorProxy (Larsson et al., 2017)		65.9		38.4
Counting (Noroozi et al., 2017)	-	67.7	51.4	36.6
(Ours) RotNet	**70.87**	**72.97**	**54.4**	**39.1**

Pretrained with full ImageNet supervision

No pretraining

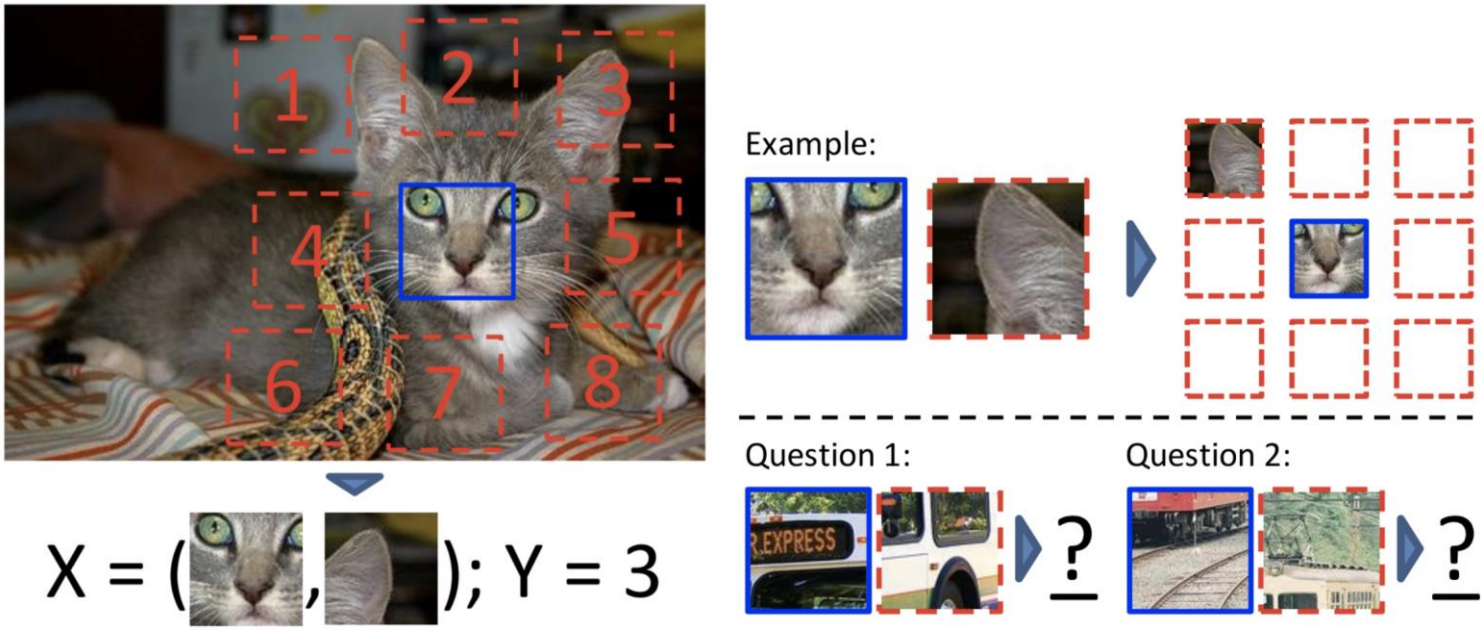
Self-supervised learning with rotation prediction

Self-supervised learning on **ImageNet** (entire training set) with AlexNet.

Finetune on labeled data from **Pascal VOC 2007**.

source: Gidaris et al. 2018

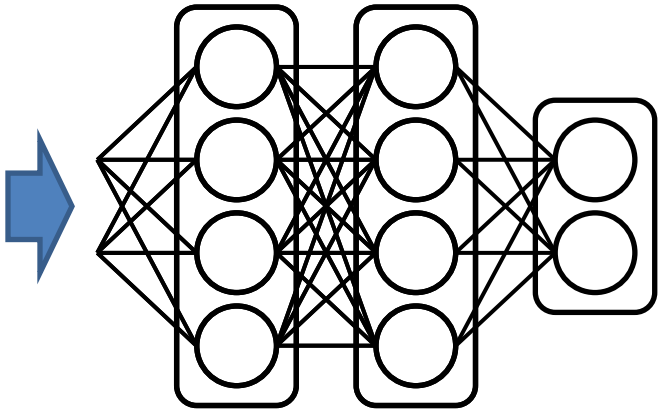
Pretext task: predict relative patch locations



(Image source: Doersch et al., 2015)

Unsupervised Visual Representation Learning by Context Prediction
Carl Doersch, Alexei Efros and Abhinav Gupta
ICCV 2015

ImageNet + Deep Learning

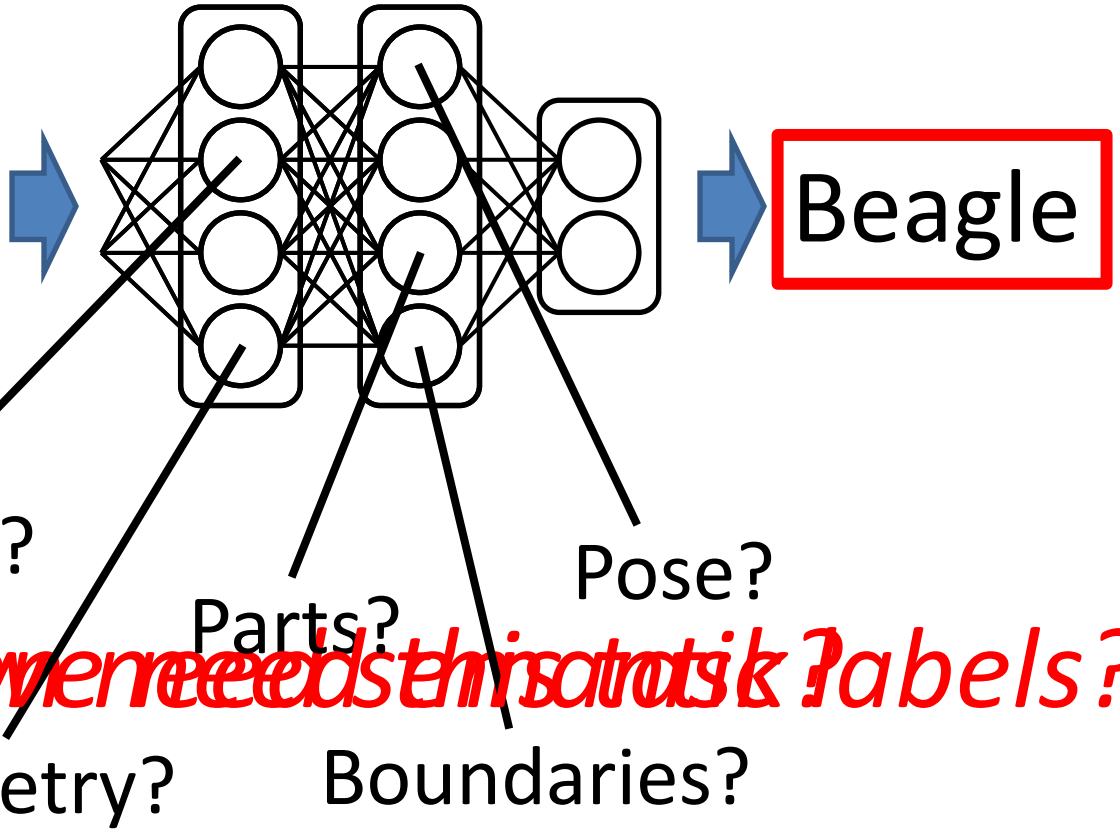


Beagle

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

Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

ImageNet + Deep Learning



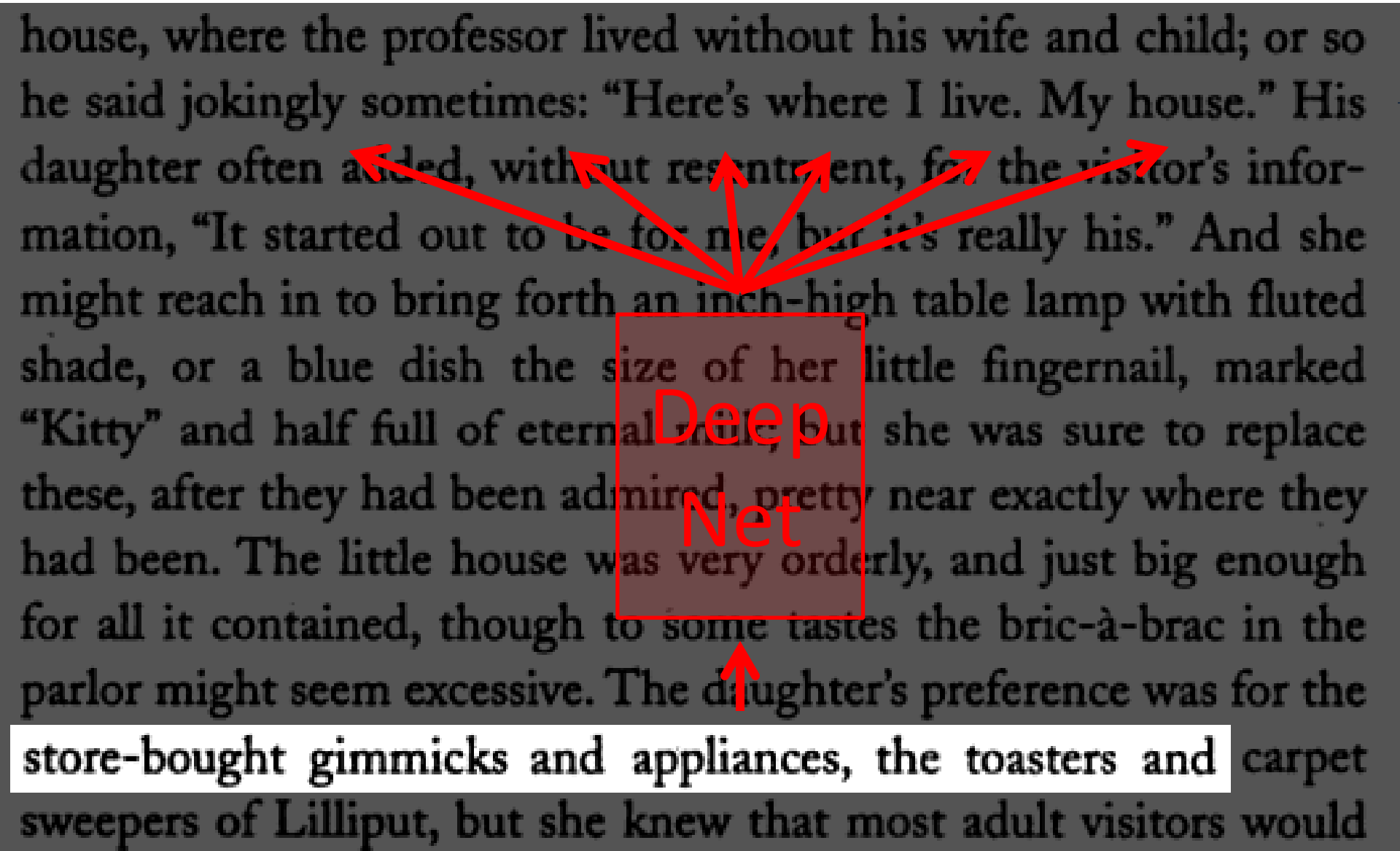
Do we even need this task?

Do we even need labels?

Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

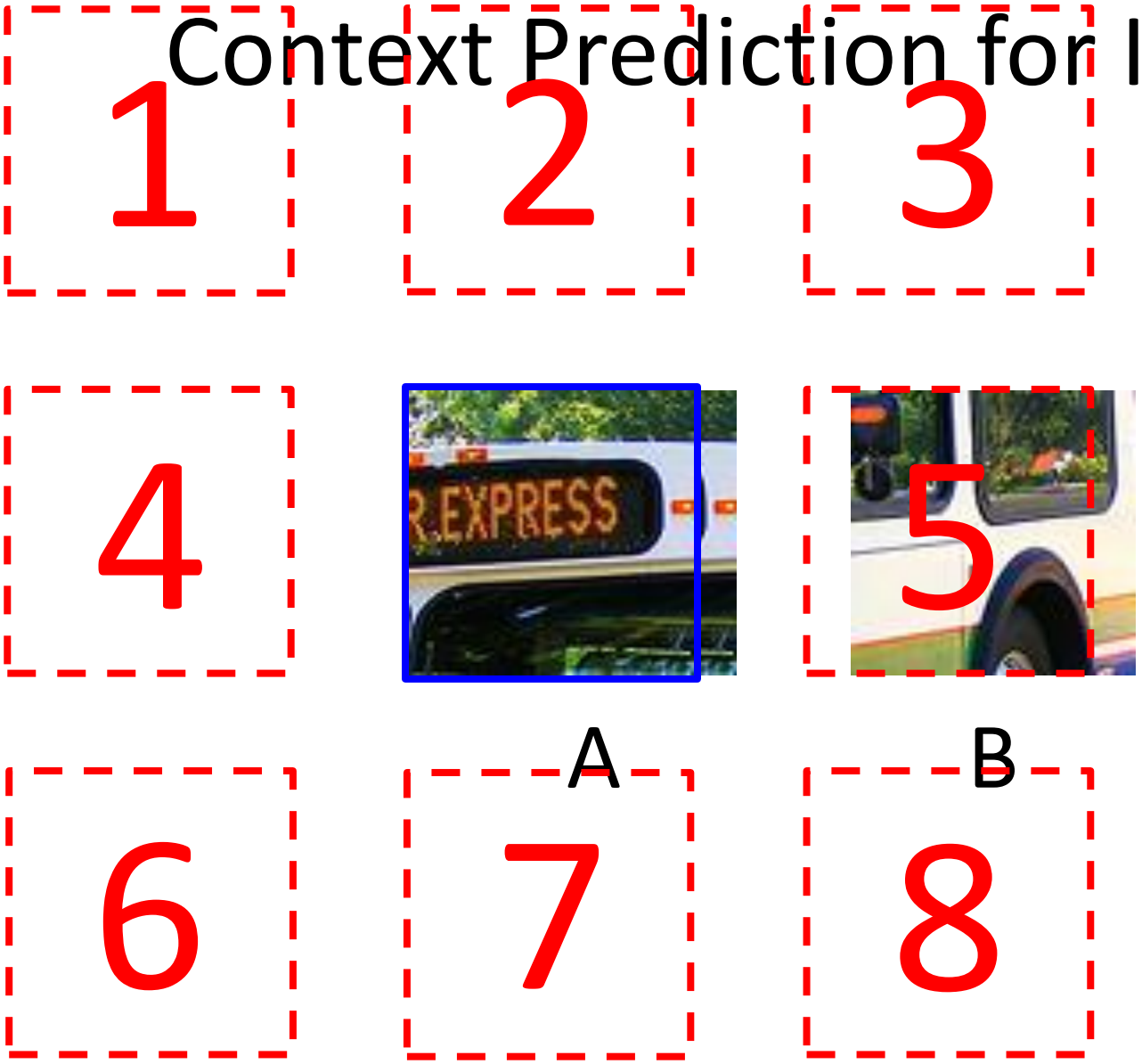
Context as Supervision

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



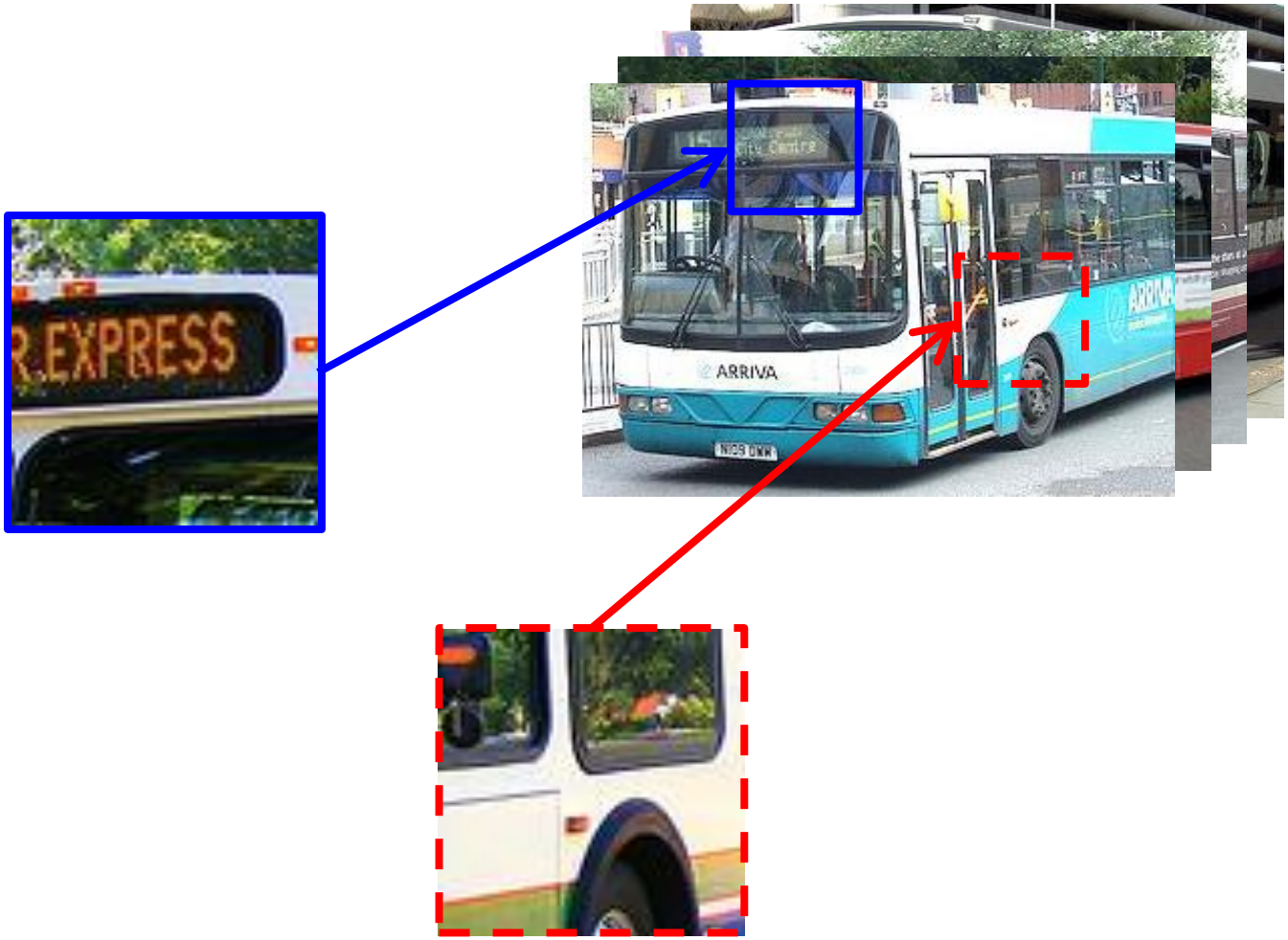
house, where the professor lived without his wife and child; or so he said jokingly sometimes: "Here's where I live. My house." His daughter often added, without resentment, for the visitor's information, "It started out to be for me, but it's really his." And she might reach in to bring forth an inch-high table lamp with fluted shade, or a blue dish the size of her little fingernail, marked "Kitty" and half full of eternal sweets, 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



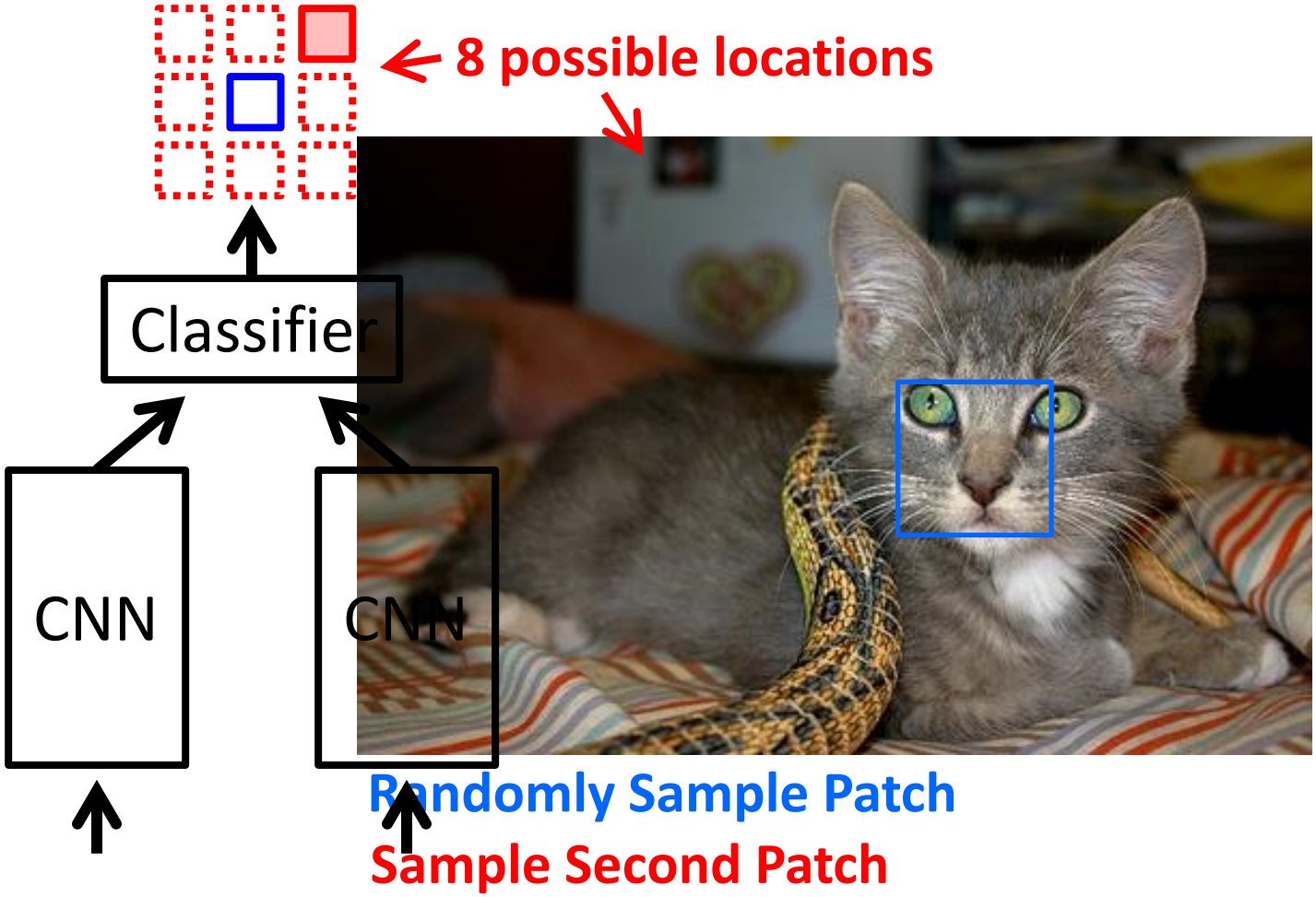
Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

Semantics from a non-semantic task



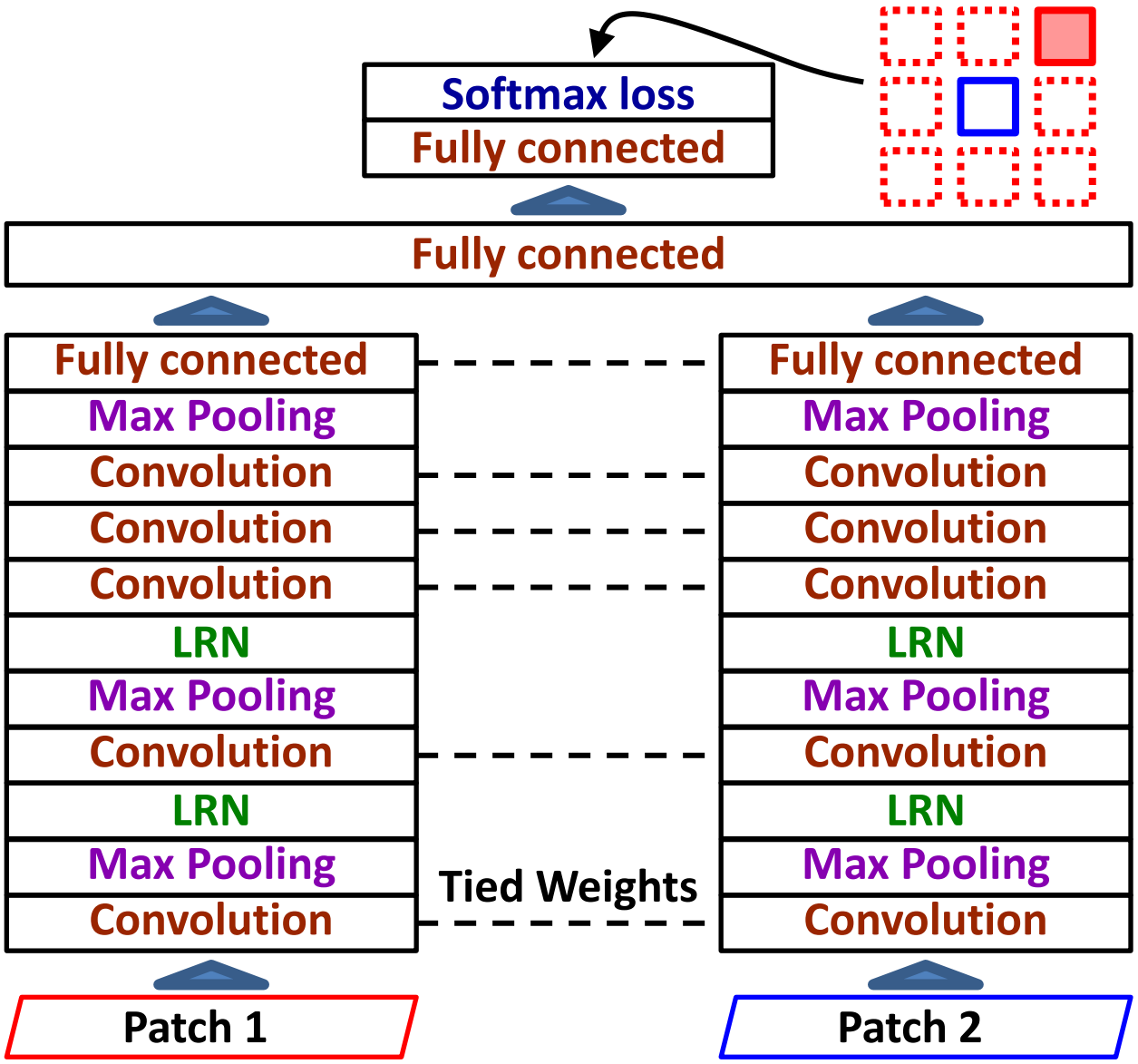
Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

Relative Position Task

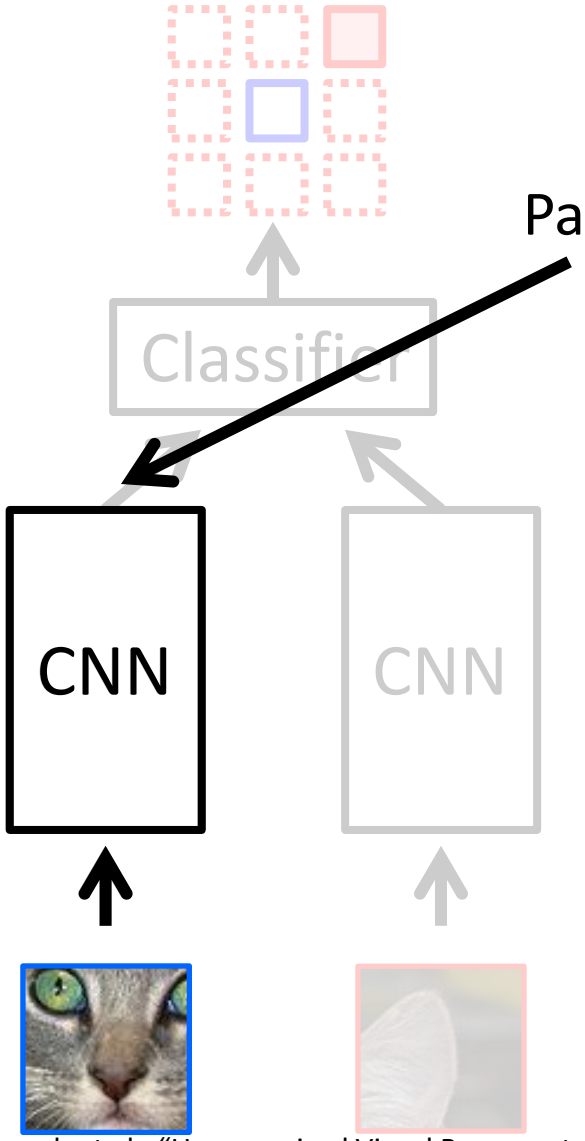


Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

Architecture



Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015



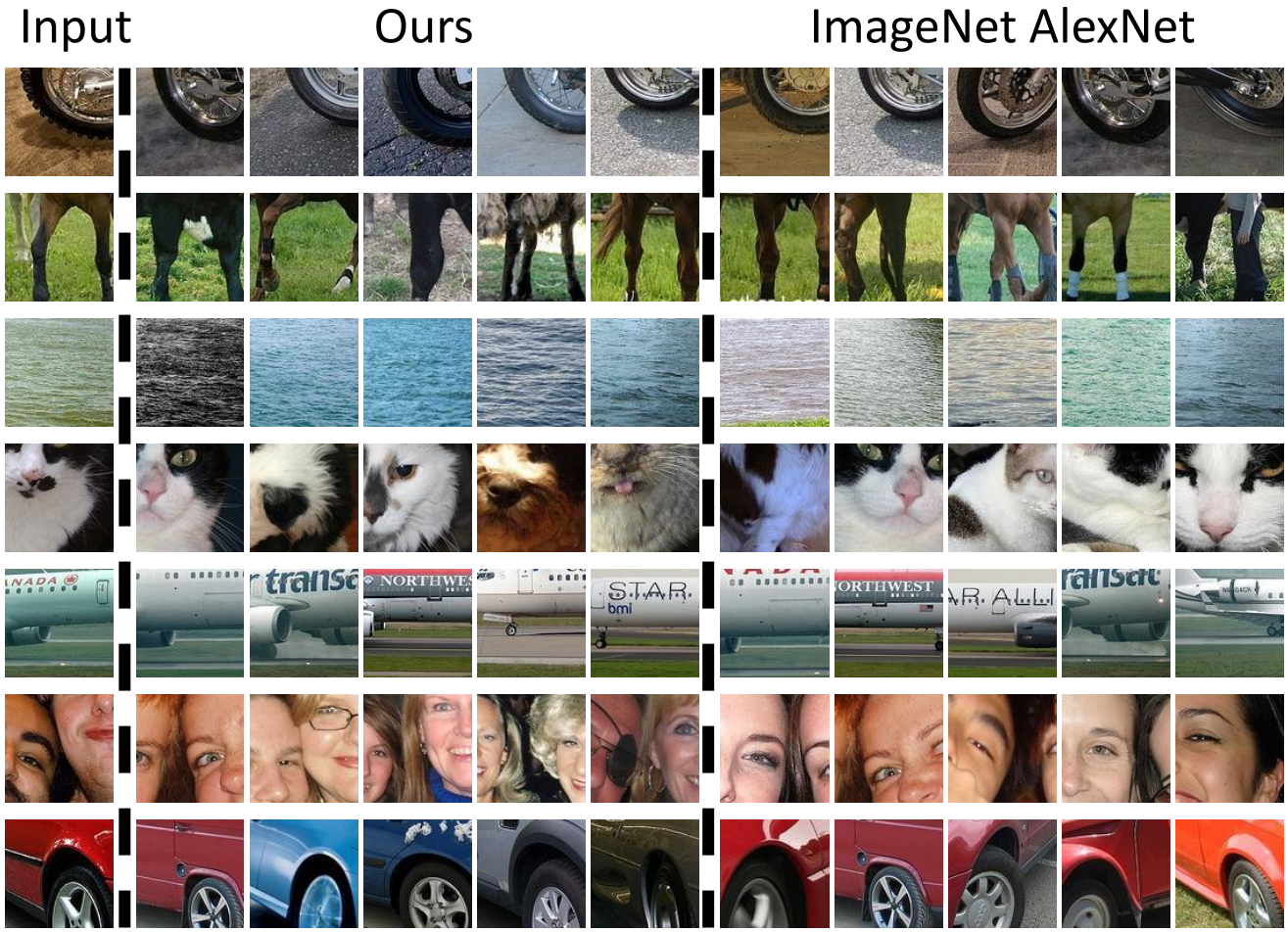
Input

Nearest Neighbors

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

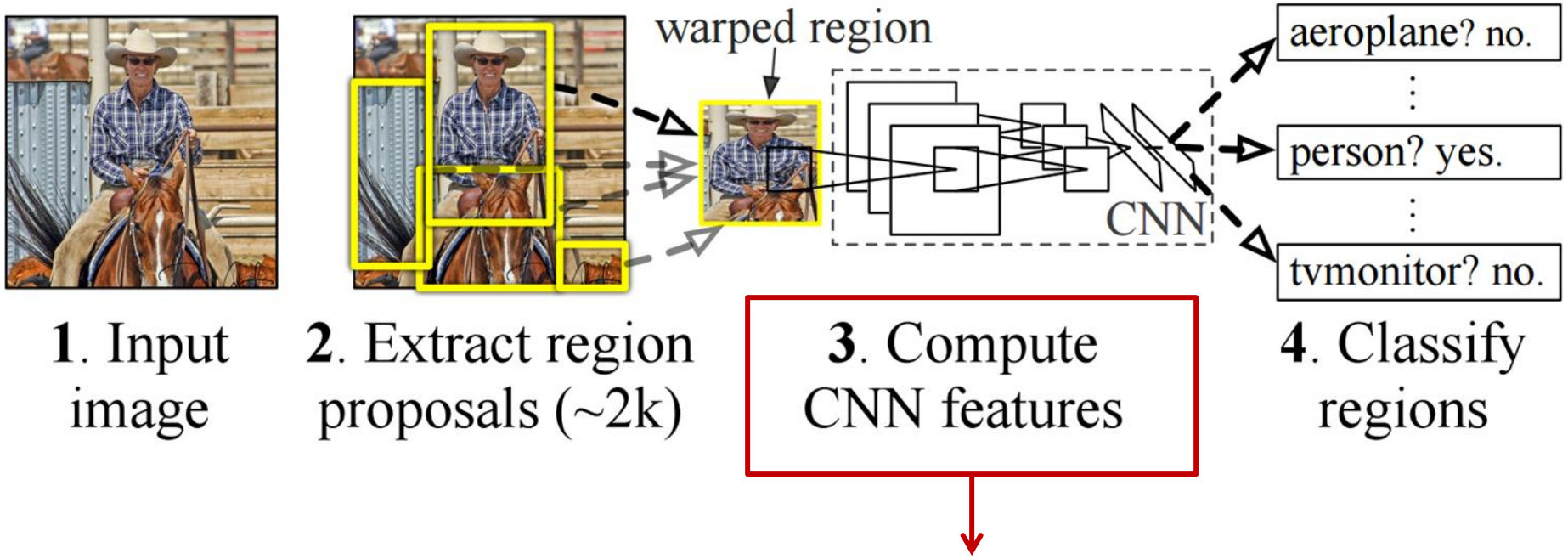
Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

What is learned?



Doersch et al., “Unsupervised Visual Representation Learning by Context Prediction”, ICCV 2015

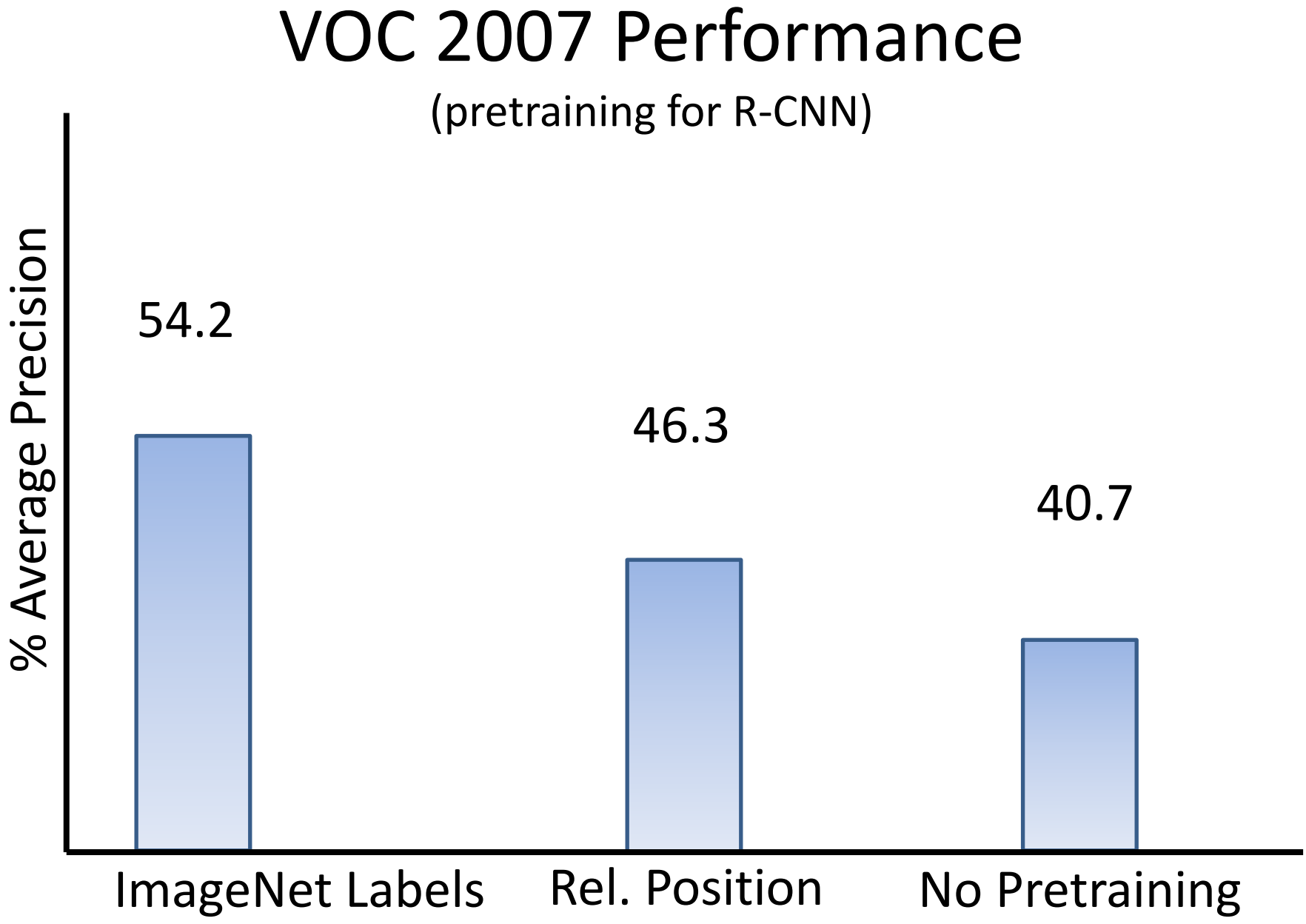
Pre-Training for R-CNN



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

Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

[Girshick et al. 2014]

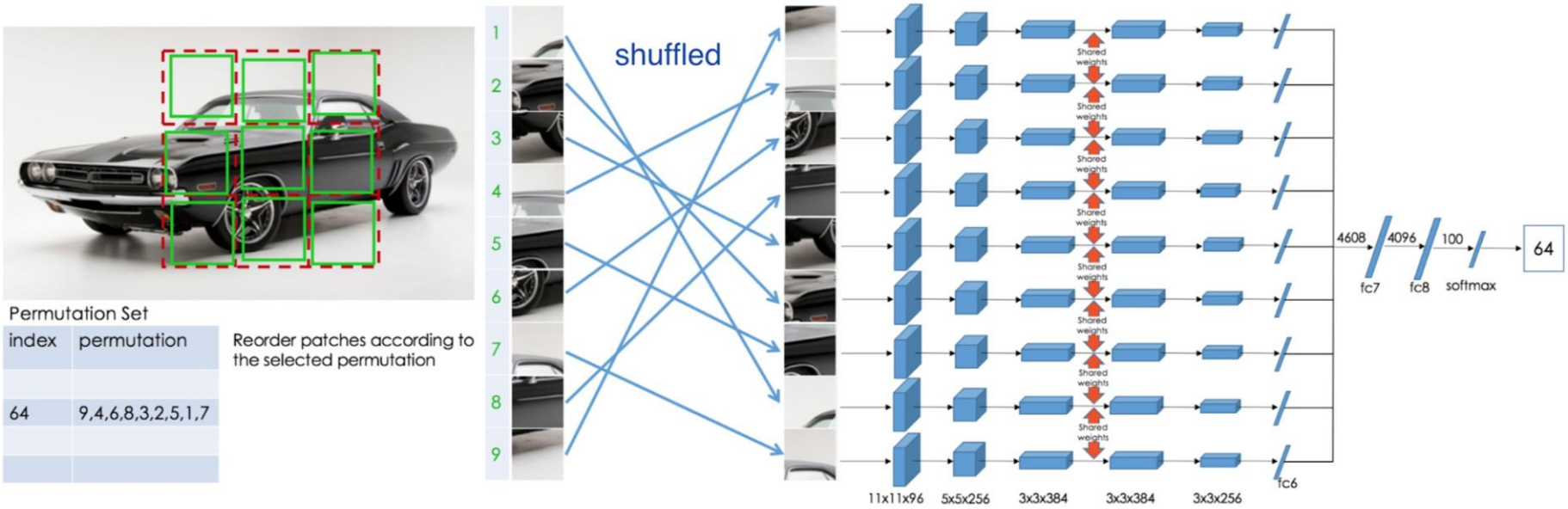


Doersch et al., "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015

Which will be better?

- Option 1: pretrain (unsup) on dataset B
- Option 2: pretrain (sup) on dataset A
- Test on dataset B

Pretext task: solving “jigsaw puzzles”



(Image source: Noroozi & Favaro, 2016)

Transfer learned features to supervised learning

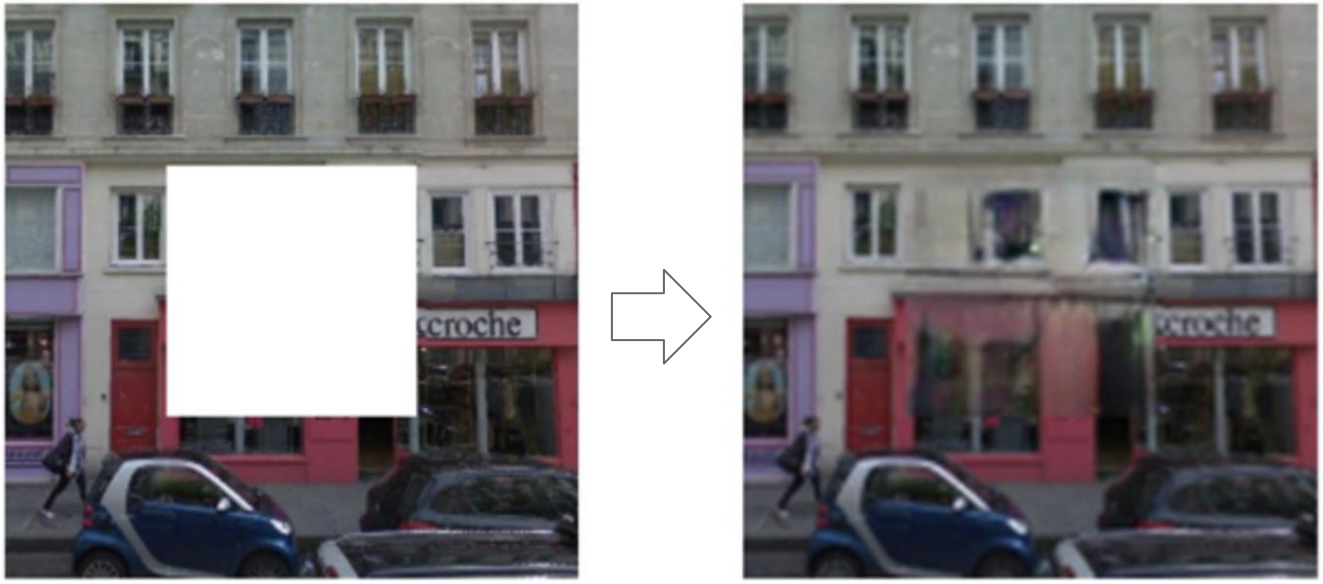
Table 1: Results on PASCAL VOC 2007 Detection and Classification. The results of the other methods are taken from Pathak *et al.* [30].

Method	Pretraining time	Supervision	Classification	Detection	Segmentation
Krizhevsky*et al.* [25]	3 days	1000 class labels	**78.2%**	**56.8%**	**48.0%**
Wang and Gupta[39]	1 week	motion	58.4%	44.0%	-
Doersch *et al.* [10]	4 weeks	context	55.3%	46.6%	-
Pathak *et al.* [30]	14 hours	context	56.5%	44.5%	29.7%
Ours	2.5 days	context	**67.6%**	**53.2%**	**37.6%**

"Ours" is feature learned from solving image Jigsaw puzzles (Noroozi & Favaro, 2016). Doersch et al. is the method with relative patch location

(source: Noroozi & Favaro, 2016)

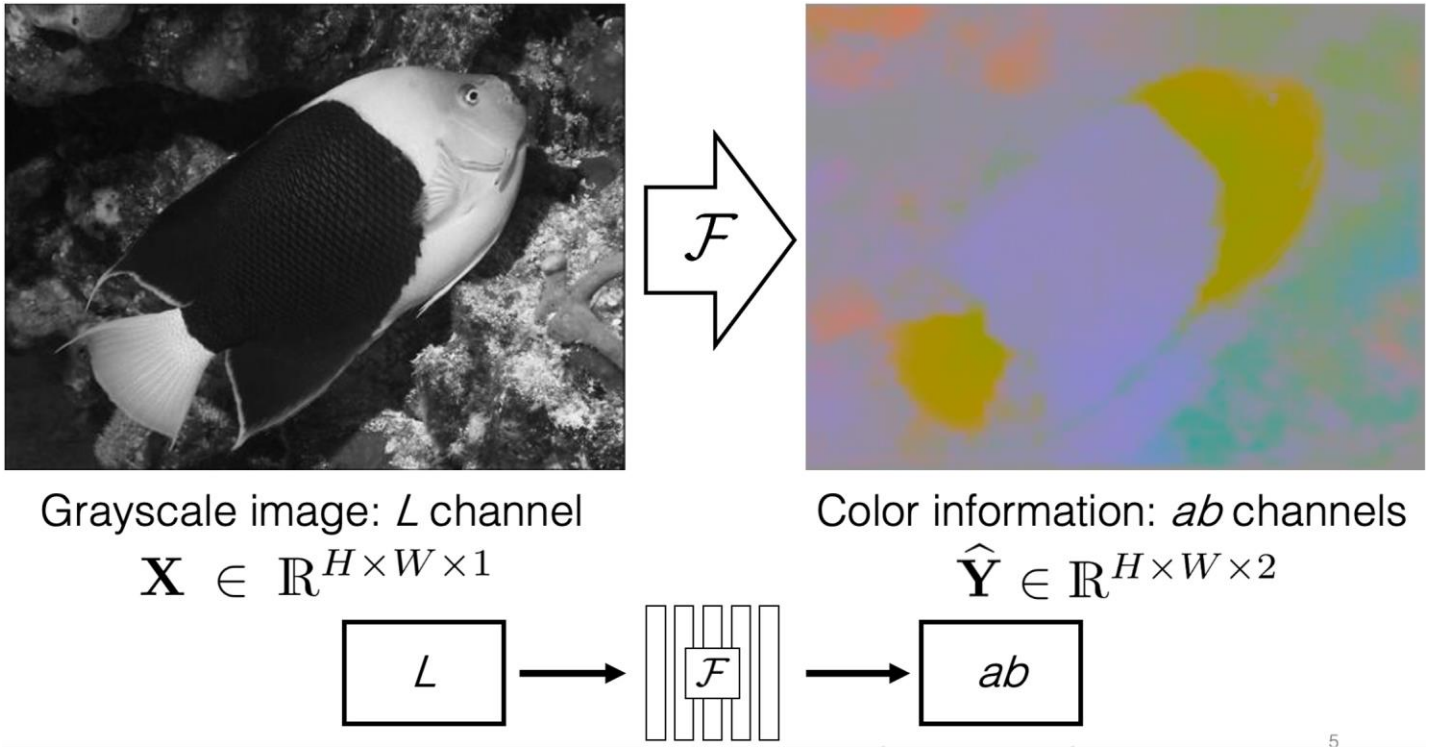
Pretext task: predict missing pixels (inpainting)



Context Encoders: Feature Learning by Inpainting (Pathak et al., 2016)

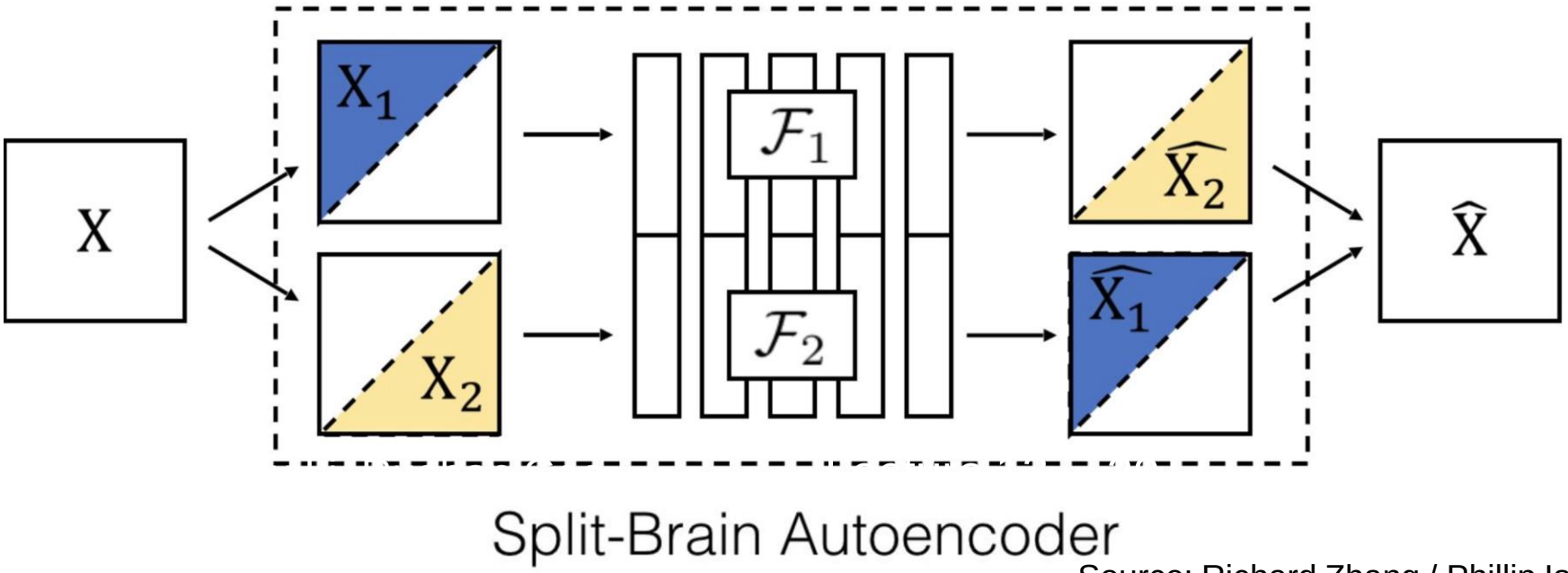
Source: Pathak et al., 2016

Pretext task: image coloring



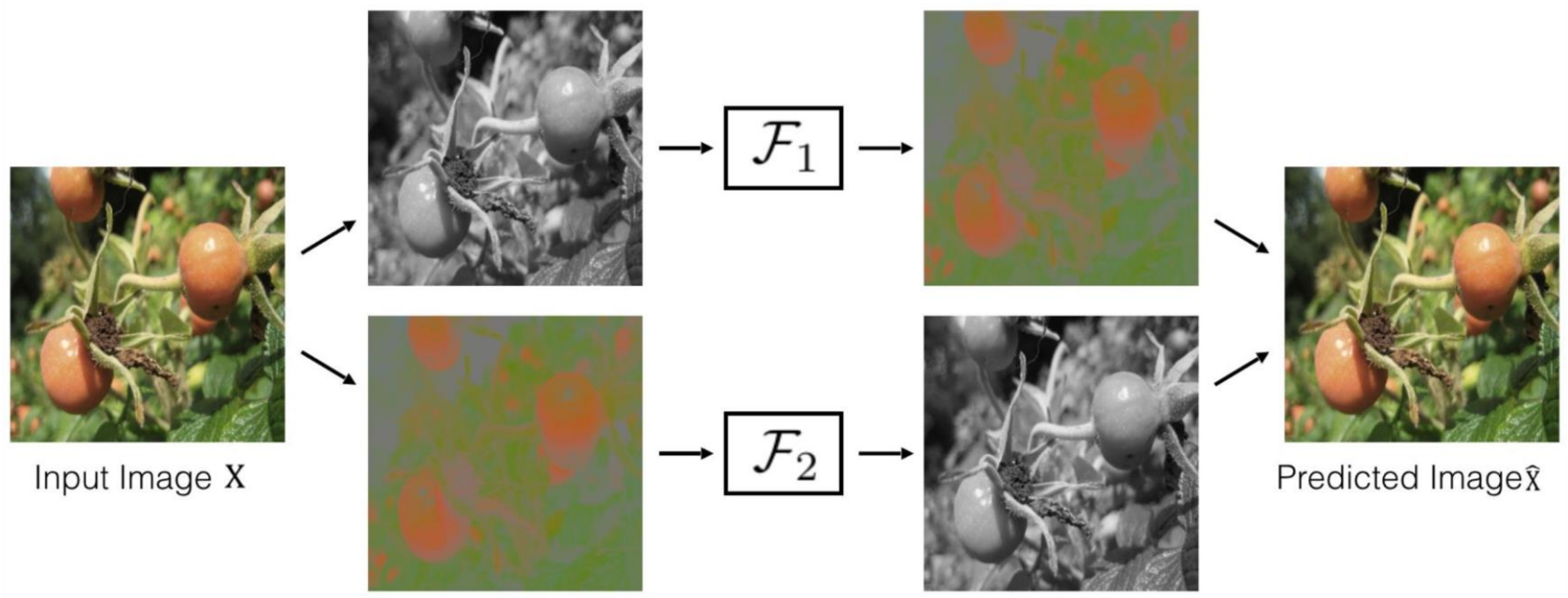
Source: Richard Zhang / Phillip Isola

Learning features from colorization: Split-brain Autoencoder



Source: Richard Zhang / Phillip Isola

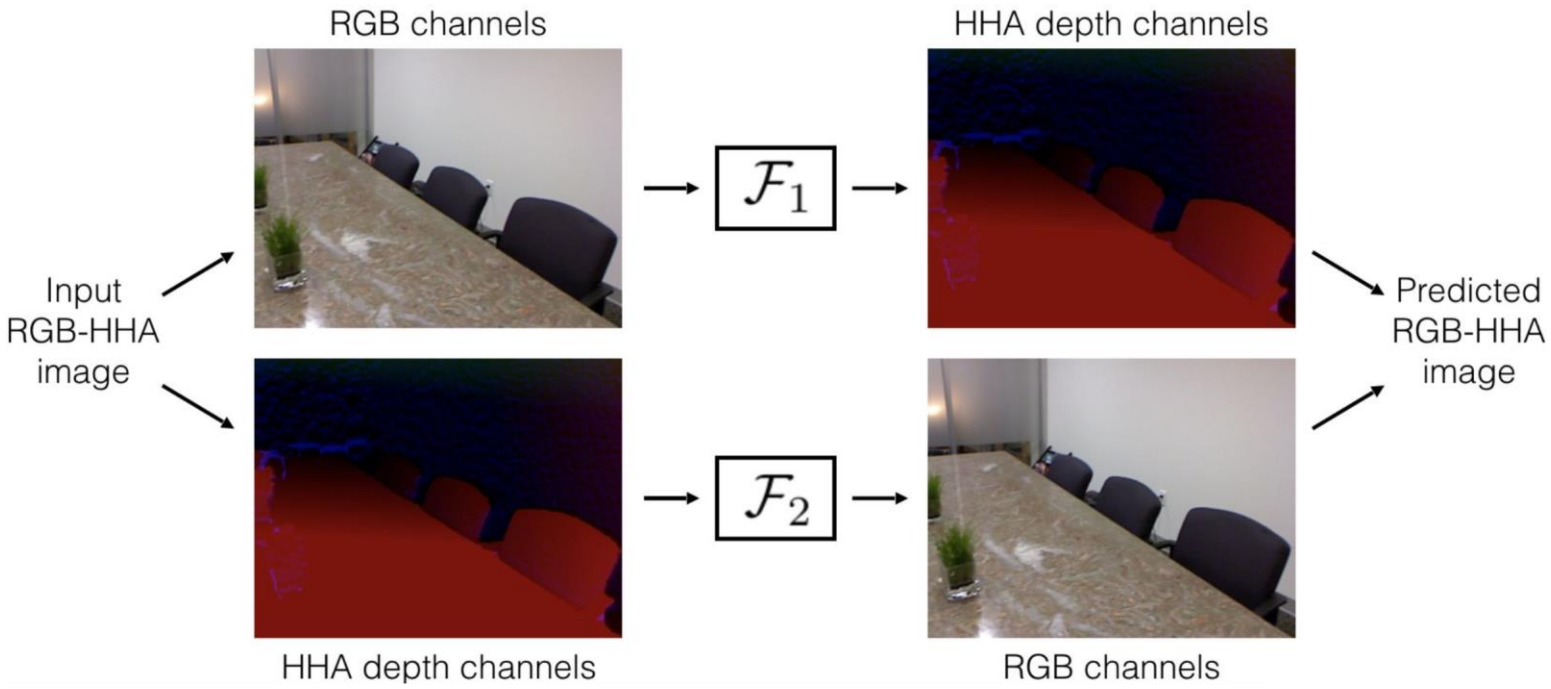
Learning features from colorization: Split-brain Autoencoder



Source: Richard Zhang / Phillip Isola

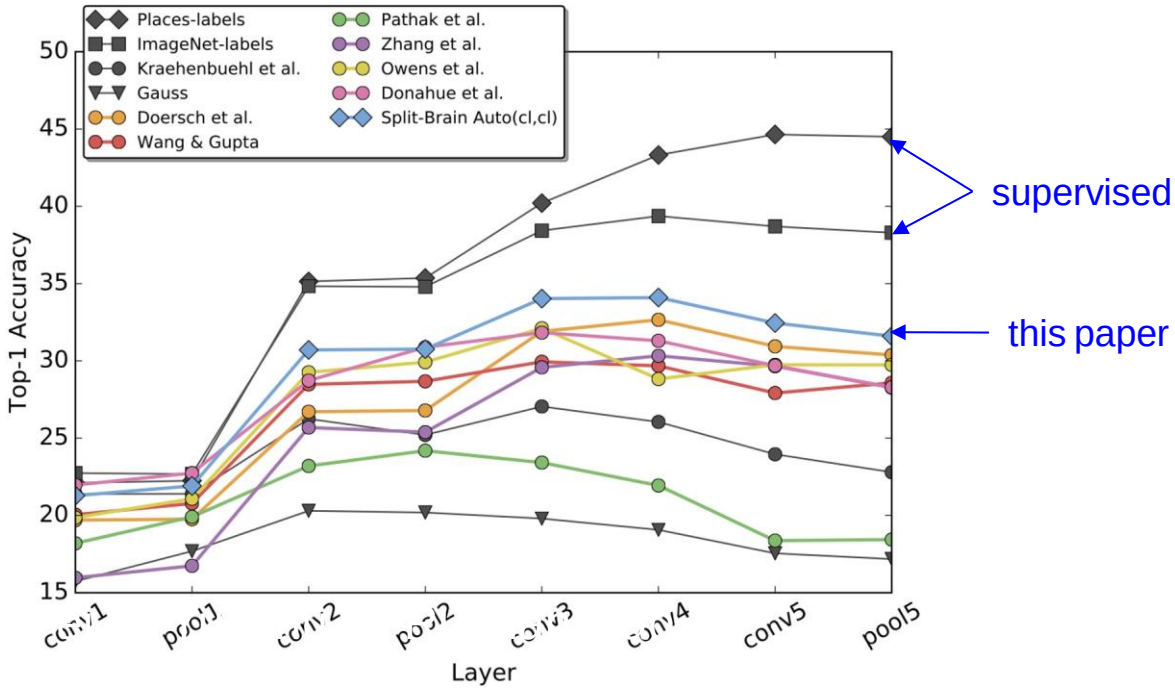
Yunzhu Li, Ruohan Guo, Justin Johnson, Fei-Fei Li

Learning features from colorization: Split-brain Autoencoder



Source: Richard Zhang / Phillip Isola

Transfer learned features to supervised learning



Self-supervised learning on **ImageNet** (entire training set).

Use *concatenated features* from F_1 and F_2

Labeled data is from the **Places** (Zhou 2016).

Source: Zhang et al., 2017

Pretext task:
Temporal Order Verification

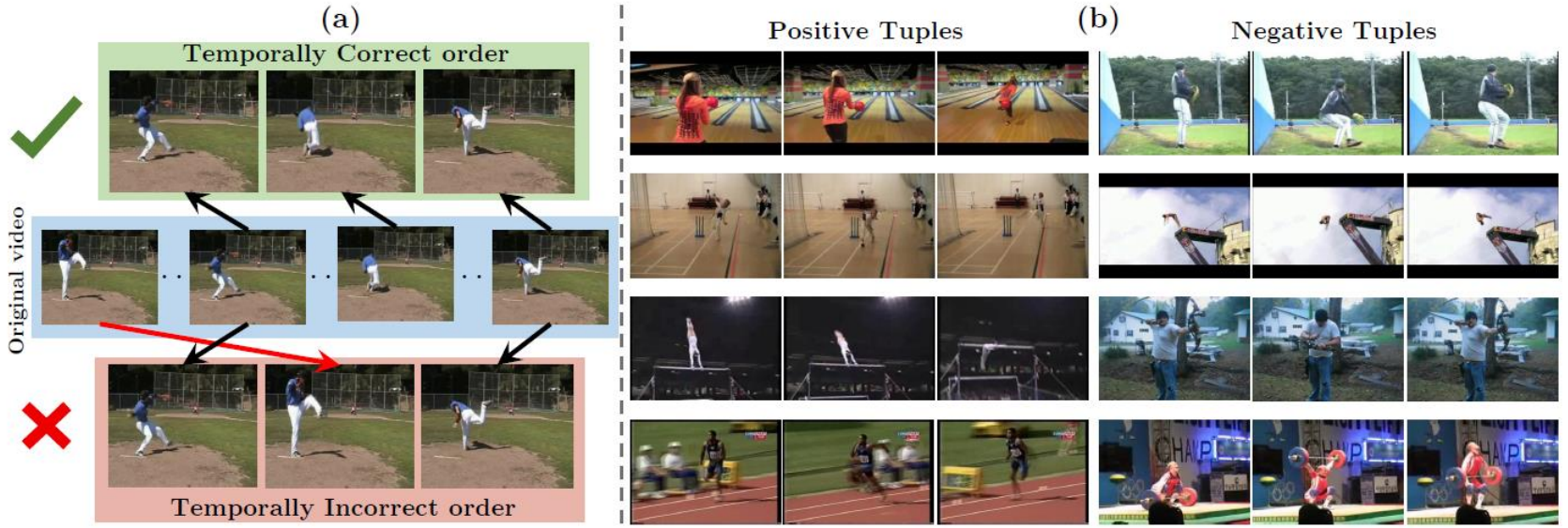


Fig. 1: **(a)** A video imposes a natural temporal structure for visual data. In many cases, one can easily verify whether frames are in the correct temporal order (shuffled or not). Such a simple sequential verification task captures important spatiotemporal signals in videos. We use this task for unsupervised pre-training of a Convolutional Neural Network (CNN). **(b)** Some examples of the automatically extracted positive and negative tuples used to formulate a classification task for a CNN.

Misra et al., "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", ECCV 2016

Pretext task: Temporal Order Verification

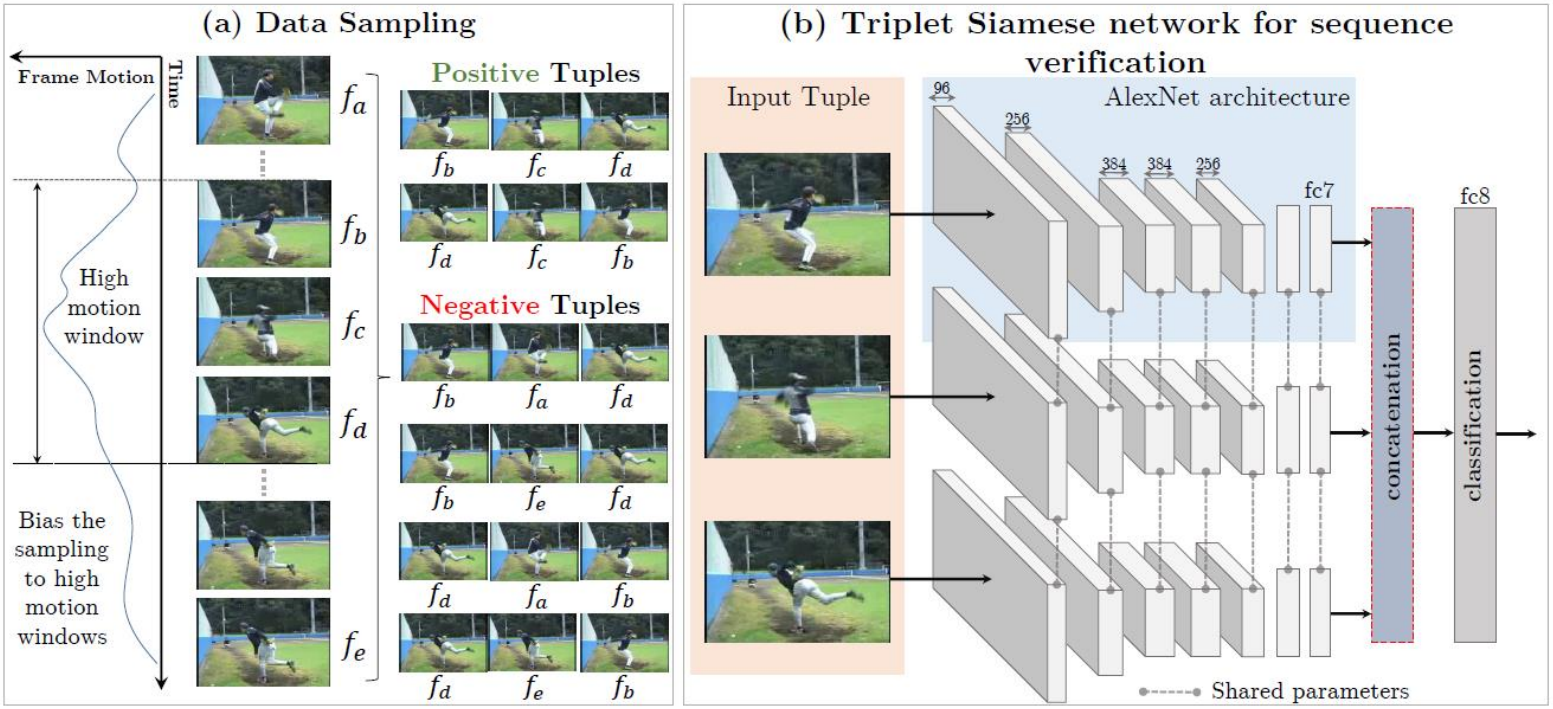


Fig. 2: **(a)** We sample tuples of frames from high motion windows in a video. We form positive and negative tuples based on whether the three input frames are in the correct temporal order. **(b)** Our triplet Siamese network architecture has three parallel network stacks with shared weights upto the `fc7` layer. Each stack takes a frame as input, and produces a representation at the `fc7` layer. The concatenated `fc7` representations are used to predict whether the input tuple is in the correct temporal order.

Misra et al., "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", ECCV 2016

Pretext task: Temporal Order Verification

Table 2: Mean classification accuracies over the 3 splits of UCF101 and HMDB51 datasets. We compare different initializations and finetune them for action recognition.

Dataset	Initialization	Mean Accuracy
UCF101	Random	38.6
	(Ours) Tuple verification	**50.2**
HMDB51	Random	13.3
	UCF Supervised	15.2
	(Ours) Tuple verification	**18.1**

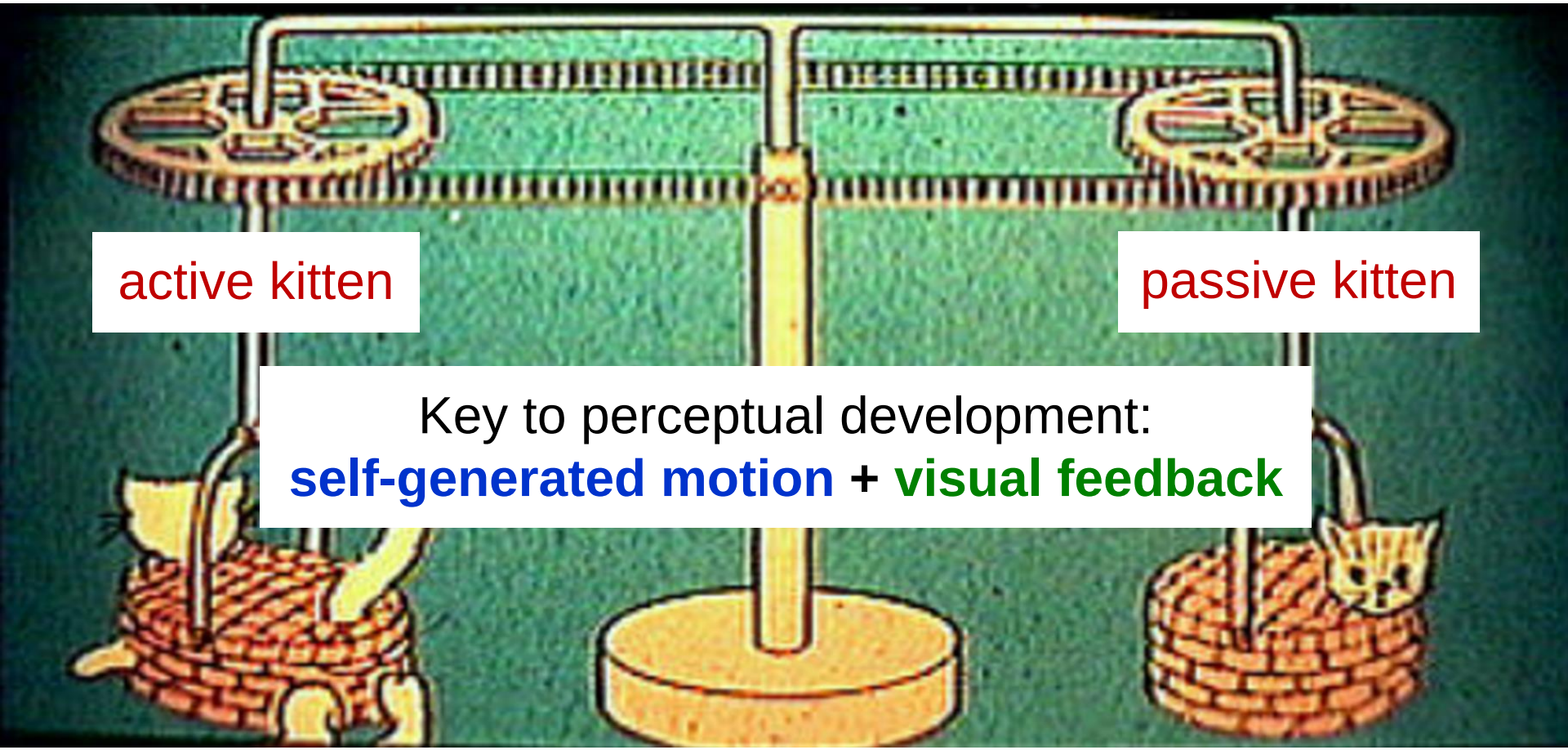
Misra et al., "Shuffle and Learn: Unsupervised Learning using Temporal Order Verification", ECCV 2016

Learning image representations tied to ego-motion

Dinesh Jayaraman and Kristen Grauman

ICCV 2015

The kitten carousel experiment
[Held & Hein, 1963]



Jayaraman and Grauman, “Learning image representations tied to ego-motion”, ICCV 2015

Problem with today's visual learning

Status quo: Learn from "disembodied" bag of labeled snapshots.

Our goal: Learn in the context of acting and moving in the world.

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Our idea: Ego-motion ↔ vision

Goal: Teach computer vision system the connection:
"how I move" ↔ "how my visual surroundings change"

Ego-motion motor signals

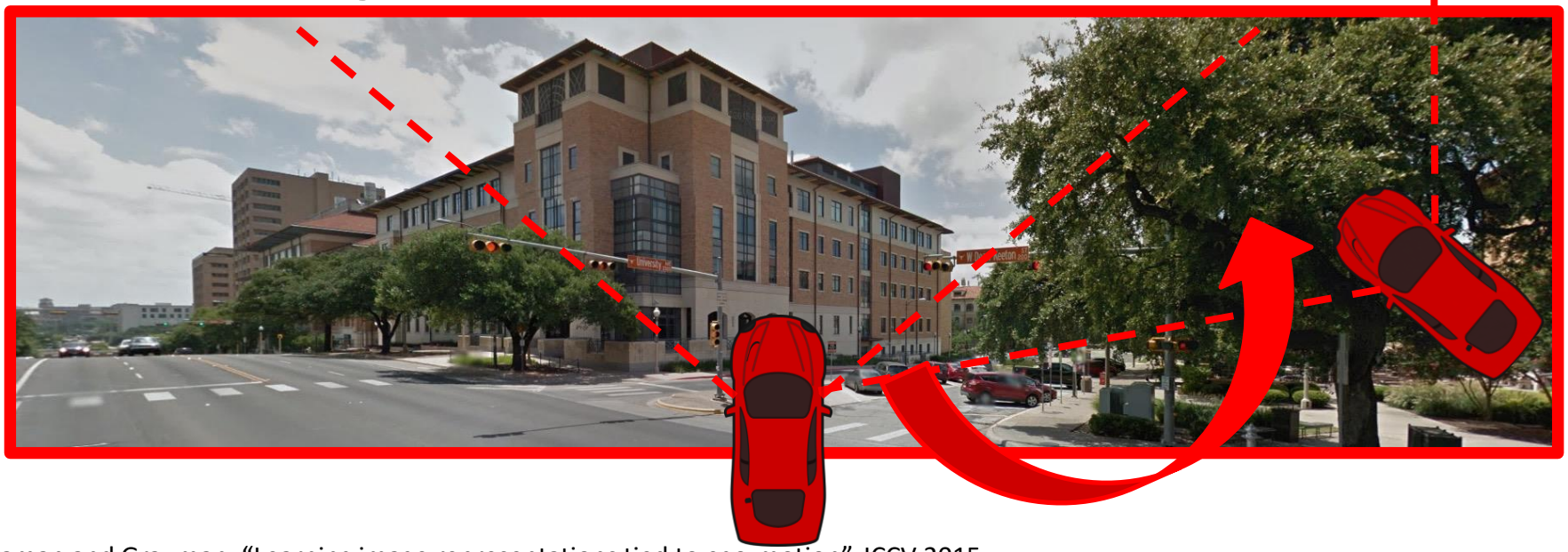
+

Unlabeled video

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Ego-motion ↔ vision: view prediction

After moving:



Jayaraman and Grauman, “Learning image representations tied to ego-motion”, ICCV 2015

Ego-motion ↔ vision for recognition

Learning this connection requires:

- Depth, 3D geometry
- Semantics
- Context

Also key to recognition!

Can be learned without manual labels!

Our approach: unsupervised feature learning using egocentric video + motor signals

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Approach idea: Ego-motion equivariance

Invariant features: unresponsive to some classes of transformations

$$\mathbf{z}(g\mathbf{x}) \approx \mathbf{z}(\mathbf{x})$$

Equivariant features : *predictably* responsive to some classes of transformations, through simple mappings (e.g., linear)

"equivariance map"

$$\mathbf{z}(g\mathbf{x}) \approx M_g \mathbf{z}(\mathbf{x})$$

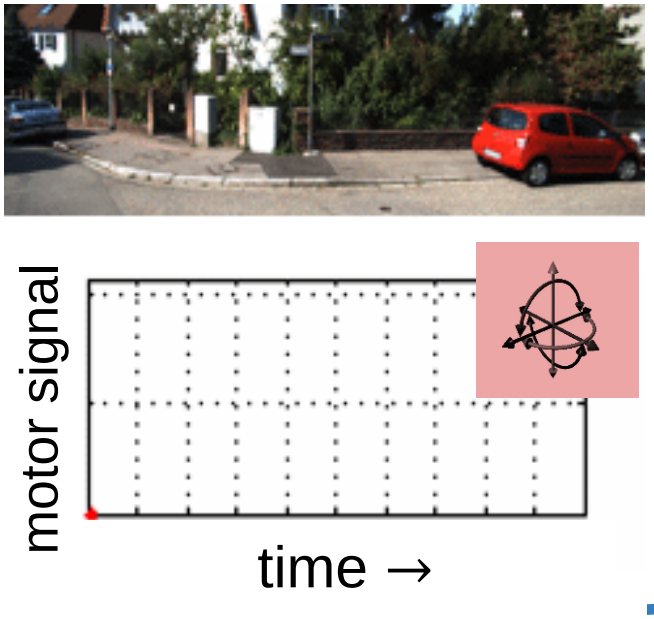
Invariance *discards* information;
equivariance *organizes* it.

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Approach idea: Ego-motion equivariance

Training data

Unlabeled video + motor signals



Learn

Equivariant embedding

organized by ego-motions

Pairs of frames related by similar ego-motion should be related by same feature transformation

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

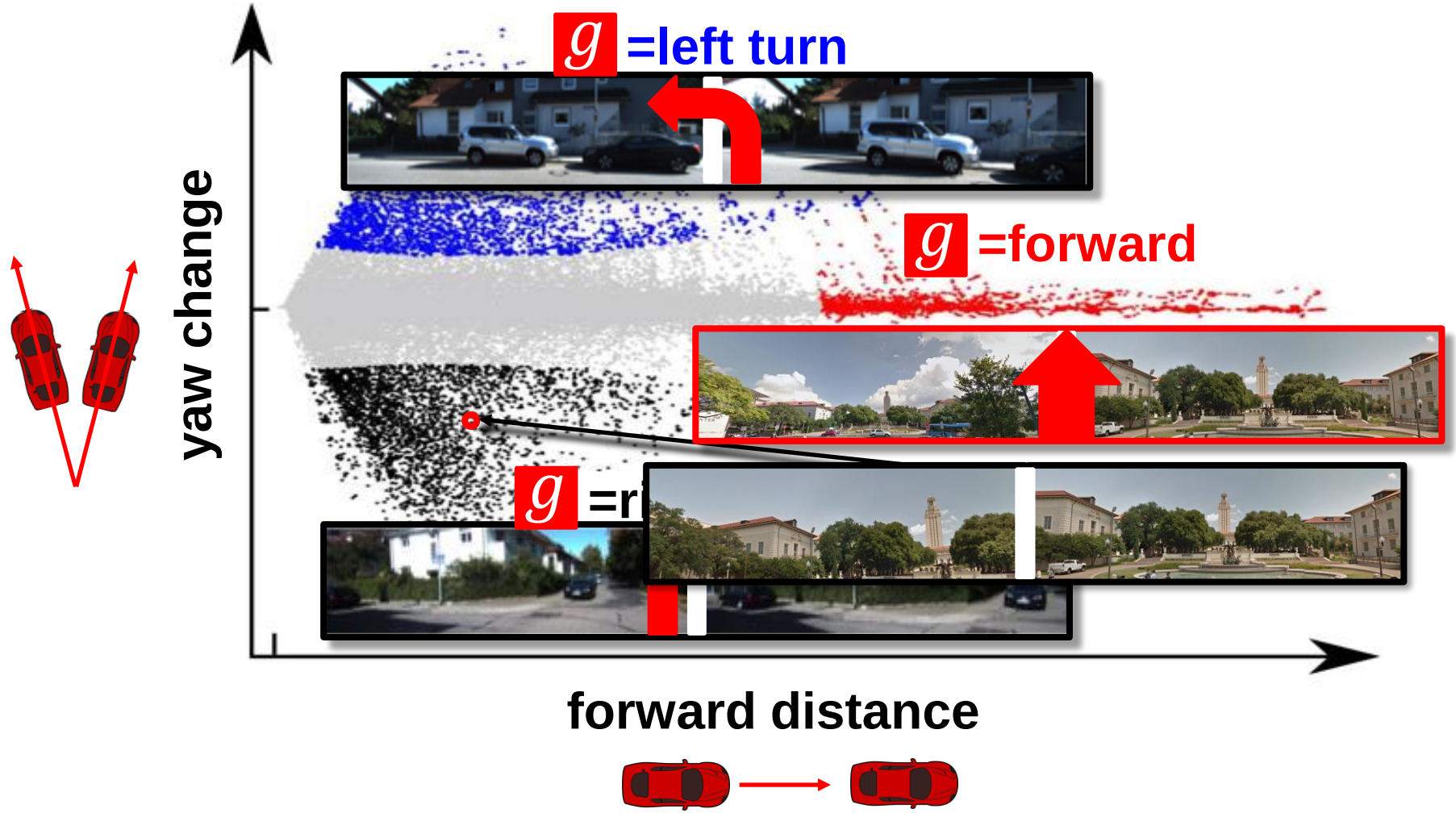
Approach overview

Our approach: unsupervised feature learning using egocentric video + motor signals

1. Extract training frame pairs from video
2. Learn ego-motion-equivariant image features
3. Train on target recognition task in parallel

Training frame pair mining

Discovery of ego-motion clusters



Jayaraman and Grauman, “Learning image representations tied to ego-motion”, ICCV 2015

Ego-motion equivariant feature learning

Given:

$\mathbf{x}_i$

$g\mathbf{x}_i$

$\mathbf{x}_k$

class y_k

Desired: for all motions g and all images $\mathbf{x}$,

$$\mathbf{z_\theta}(g\mathbf{x}) \approx M_g \mathbf{z_\theta}(\mathbf{x})$$

Unsupervised training

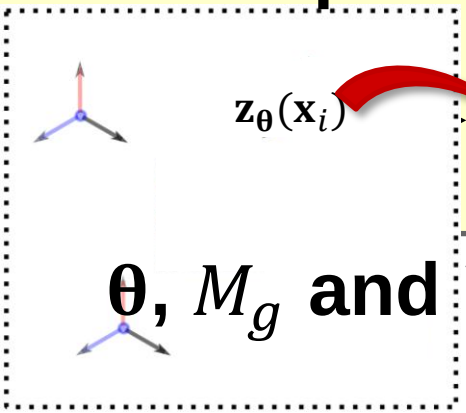
$\mathbf{x}_i$ → θ → $\mathbf{z_\theta}(\mathbf{x}_i)$ → M_g

$g\mathbf{x}_i$ → θ → $\mathbf{z_\theta}(g\mathbf{x}_i)$

$$\| M_g \mathbf{z_\theta}(\mathbf{x}_i) - \mathbf{z_\theta}(g\mathbf{x}_i) \|_2$$

Supervised training

$\mathbf{x}_k$ → θ → $\mathbf{z_\theta}(\mathbf{x}_i)$ → M_g → max loss $L_C(\mathbf{x}_k, y_k)$



θ, M_g and W jointly trained

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

Results: Recognition

Learn from **unlabeled car video** (KITTI)

Geiger et al, IJRR ’13

Exploit features for **static scene classification** (SUN, 397 classes)

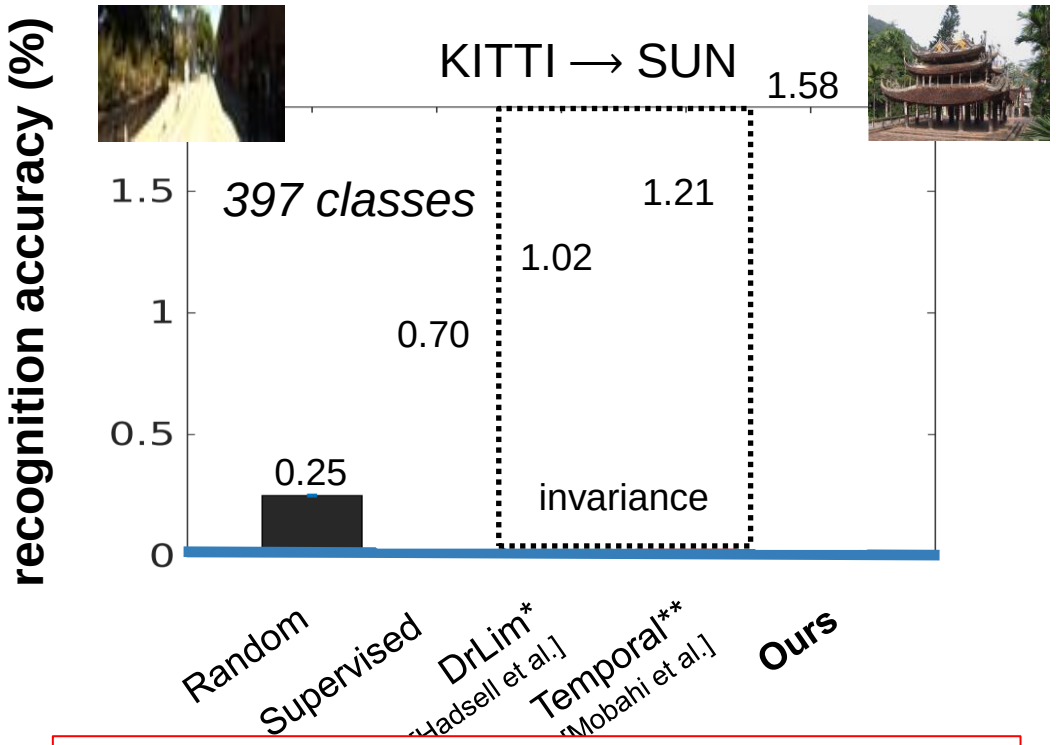
Apse | Window seat | Art school | Library | Auditorium | Bus interior | Cathedral | Freeway | Guardhouse

Xiao et al, CVPR ’10

Jayaraman and Grauman, “Learning image representations tied to ego-motion”, ICCV 2015

Results: Recognition

Do ego-motion equivariant features improve recognition?



6 labeled training examples per class

Up to 30% accuracy increase over state of the art!

Jayaraman and Grauman, "Learning image representations tied to ego-motion", ICCV 2015

The Curious Robot: Learning Visual Representations via Physical Interactions

Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han,
Yong-Lae Park, and Abhinav Gupta

ECCV 2016

Grasping

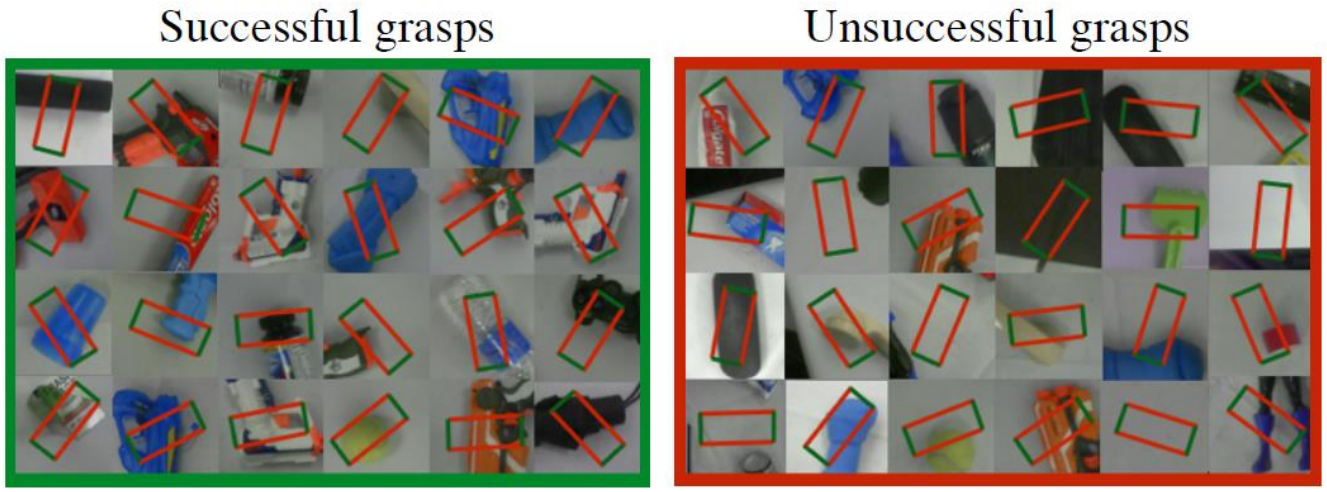


Fig. 2. Examples of successful (left) and unsuccesful grasps (right). We use a patch based representation: given an input patch we predict 18-dim vector which represents whether the center location of the patch is graspable at 0°, 10°, ... 170°.

Pinto et al., "The Curious Robot: Learning Visual Representations via Physical Interactions", ECCV 2016

Pushing

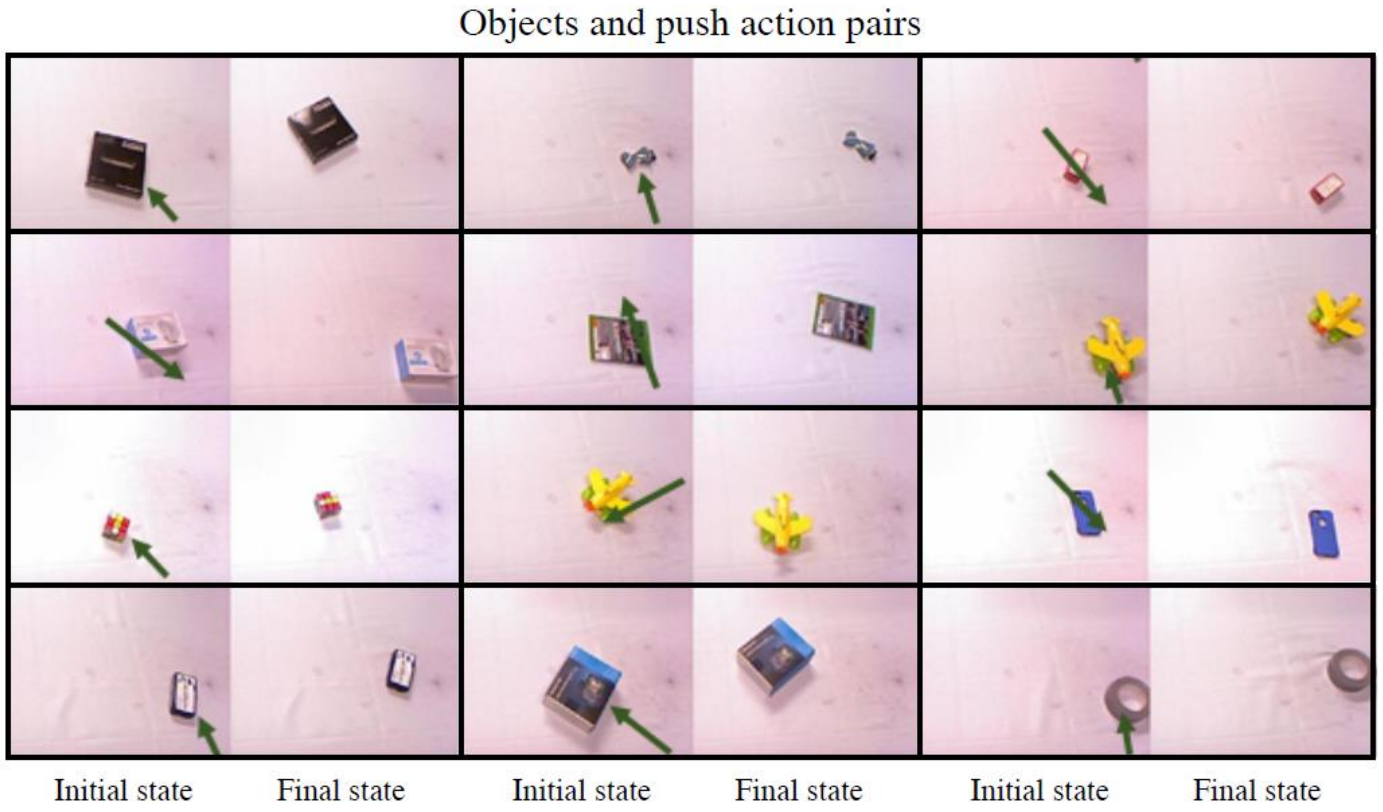


Fig. 4. Examples of initial state and final state images taken for the push action. The arrows demonstrate the direction and magnitude of the push action.

Pinto et al., "The Curious Robot: Learning Visual Representations via Physical Interactions", ECCV 2016

Poking

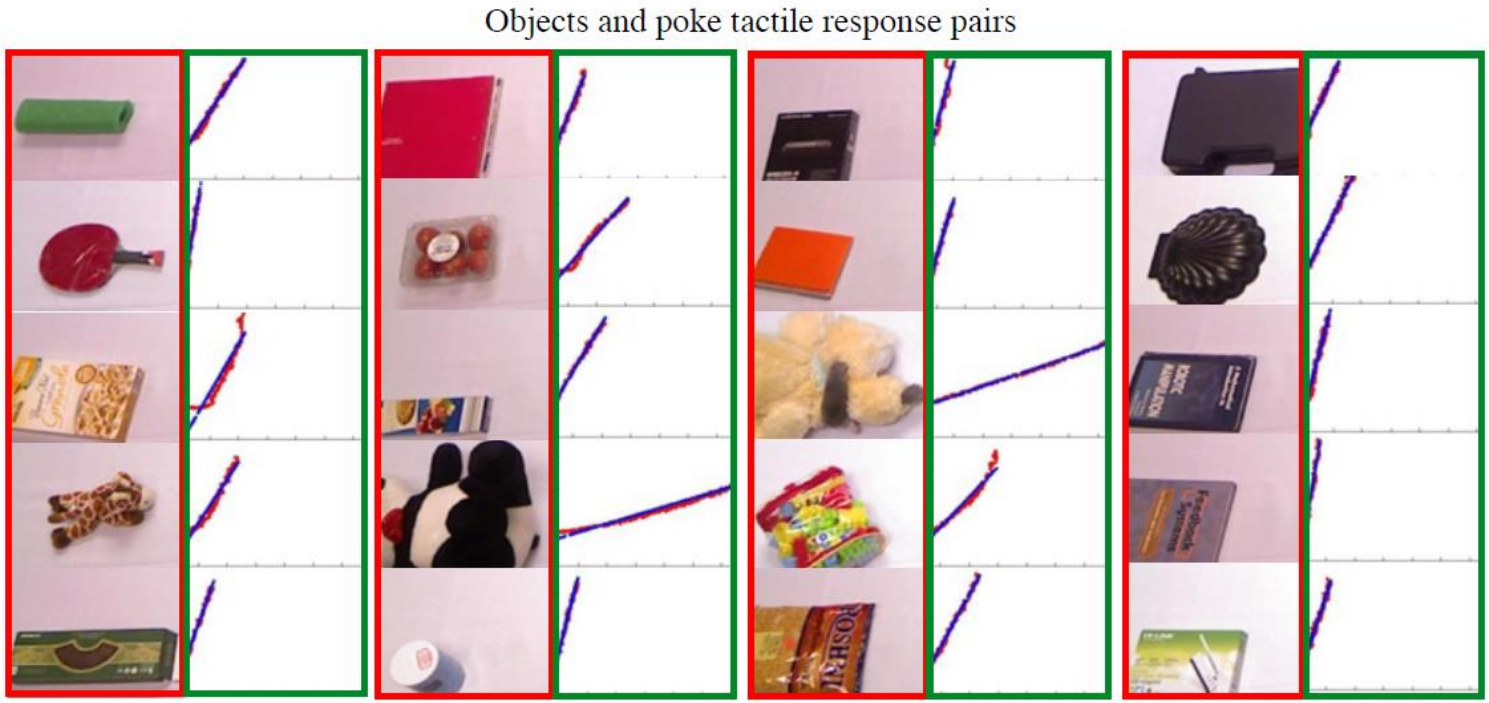


Fig. 6. Examples of the data collected by the poking action. On the left we show the object poked, and on the right we show force profiles as observed by the tactile sensor.

Pinto et al., "The Curious Robot: Learning Visual Representations via Physical Interactions", ECCV 2016

Representations from interactions

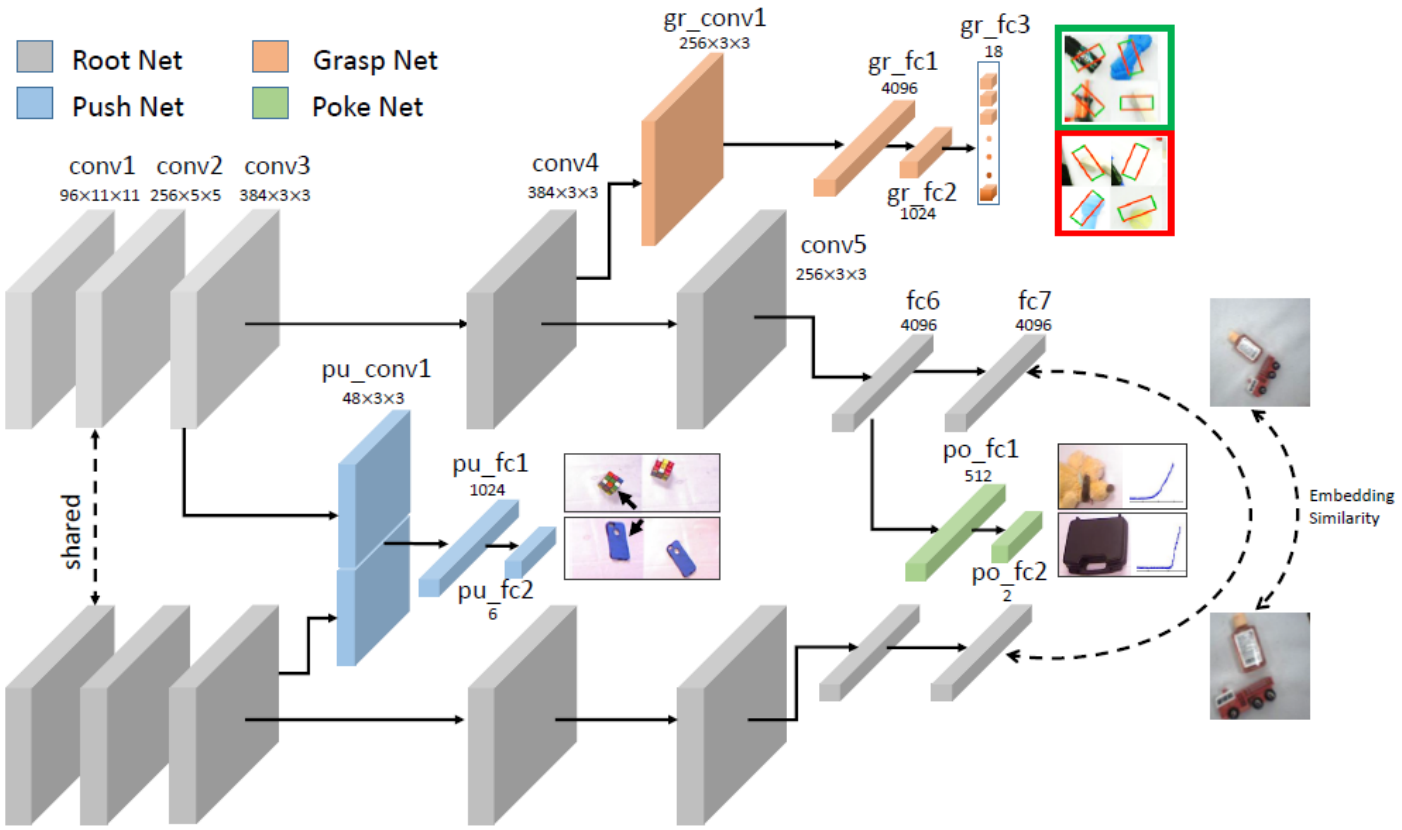


Fig. 8. Our shared convolutional architecture for four different tasks.

Pinto et al., "The Curious Robot: Learning Visual Representations via Physical Interactions", ECCV 2016

Classification/retrieval performance

Fig. 10. The first column corresponds to query image and rest show the retrieval. Note how the network learns that cups and bowls are similar (row 5).

Pinto et al., "The Curious Robot: Learning Visual Representations via Physical Interactions", ECCV 2016

Classification/retrieval performance

Table 1. Classification accuracy on ImageNet Household, UW RGBD and Caltech-256

	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

Table 2. Image Retrieval with Recall@k metric

	Instance level				Category level			
	k=1	k=5	k=10	k=20	k=1	k=5	k=10	k=20
Random Network	0.062	0.219	0.331	0.475	0.150	0.466	0.652	0.800
Our Network	0.720	0.831	0.875	0.909	0.833	0.918	0.946	0.966
AlexNet	0.686	0.857	0.903	0.941	0.854	0.953	0.969	0.982

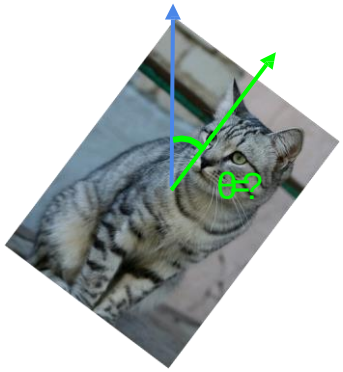
Pinto et al., "The Curious Robot: Learning Visual Representations via Physical Interactions", ECCV 2016

Summary: pretext tasks from image transformations

- Pretext tasks focus on “visual common sense”, e.g., predict rotations, inpainting, rearrangement, and colorization.
- The models are forced learn good features about natural images, e.g., semantic representation of an object category, in order to solve the pretext tasks.
- We don’t care about the performance of these pretext tasks, but rather how useful the learned features are for downstream tasks (classification, detection, segmentation).
- Problems: 1) coming up with individual pretext tasks is tedious, and 2) the learned representations may not be general.

Pretext tasks from image transformations

image completion



rotation prediction

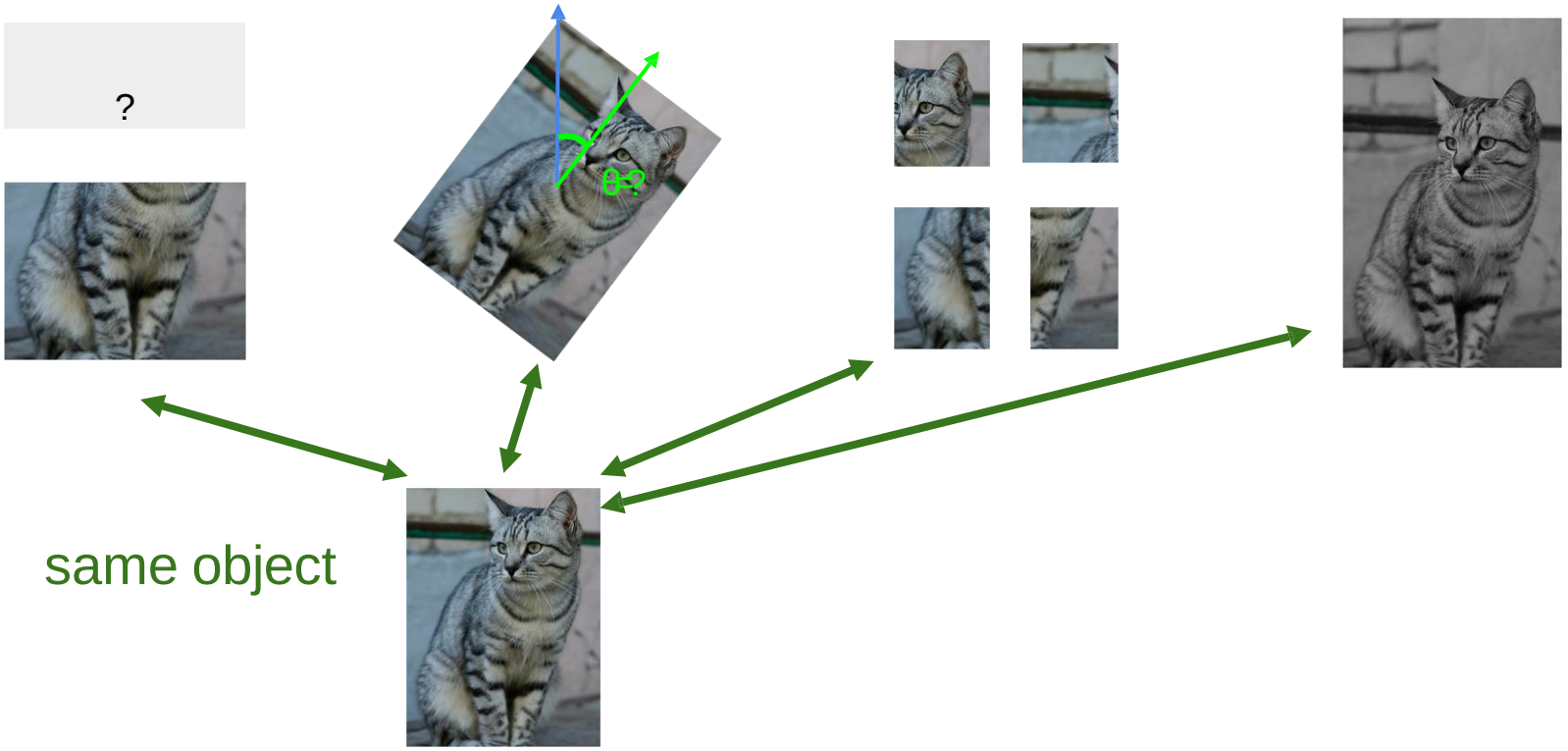
“jigsaw puzzle”

colorization

Learned representations may be tied to a specific pretext task!

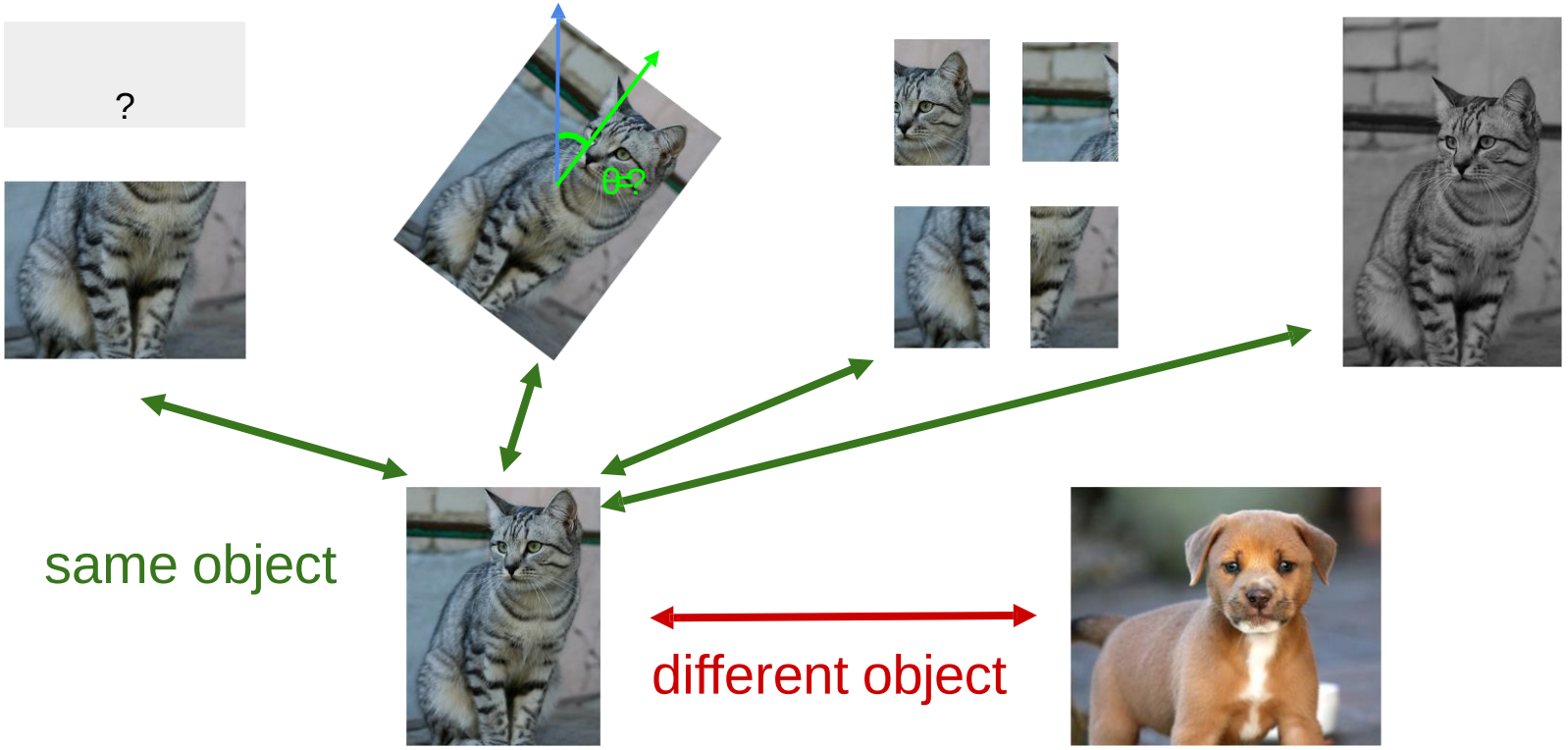
Can we come up with a more general pretext task?

A more general pretext task?

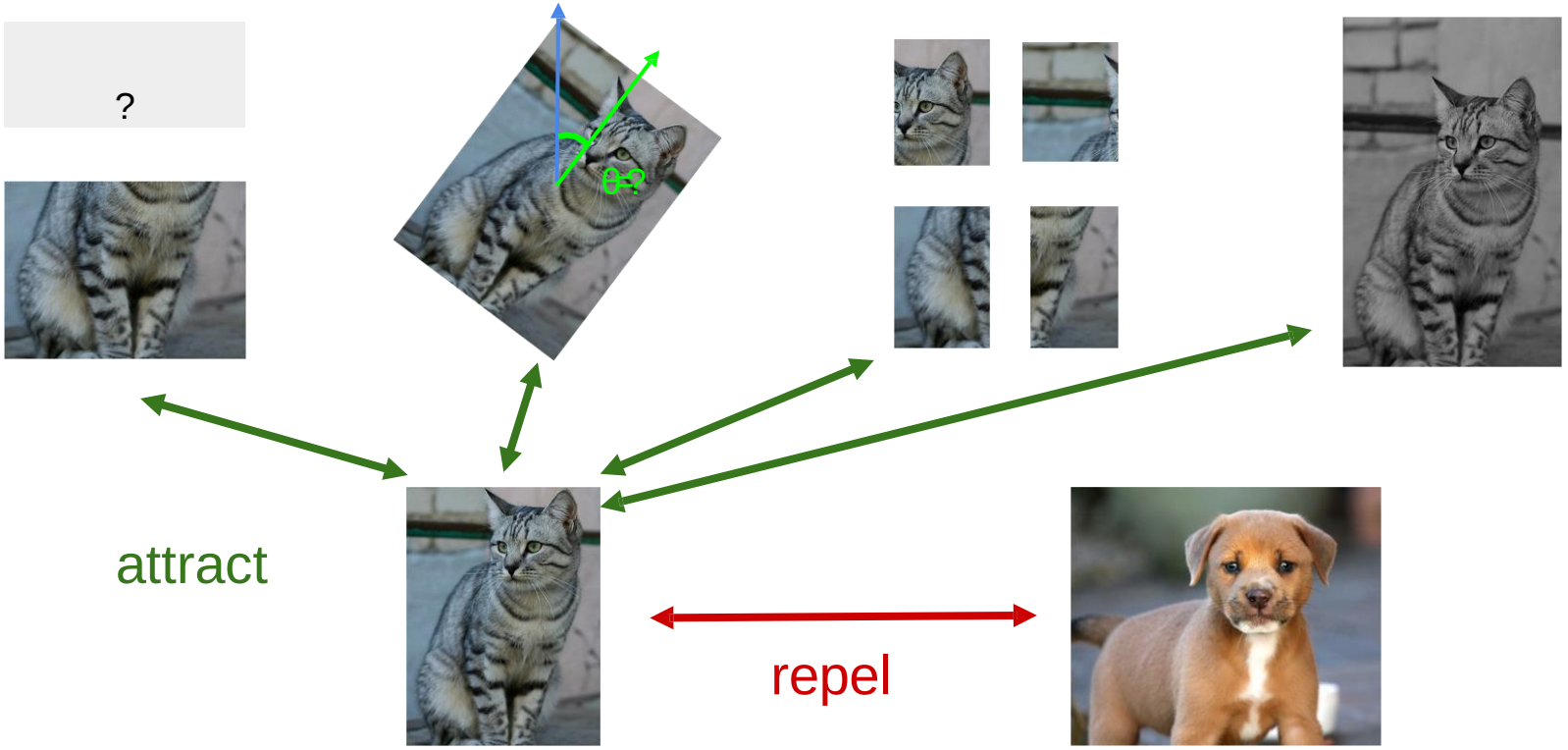


Yunzhu Li, Ruohan Guo, Justin Johnson, Fei-Fei Li

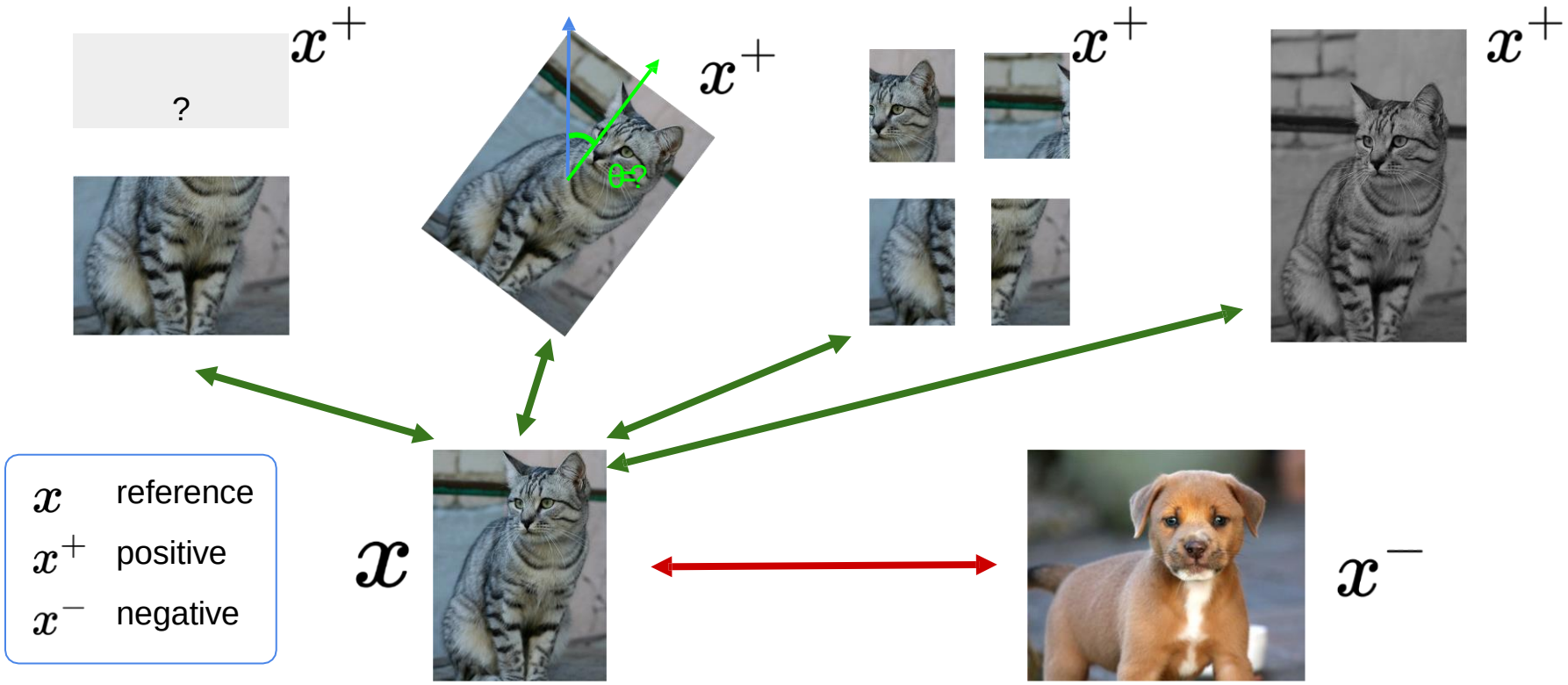
A more general pretext task?



Contrastive Representation Learning



Contrastive Representation Learning



Yunzhu Li, Ruohan Guo, Justin Johnson, Fei-Fei Li

A formulation of contrastive learning

What we want:

$$\text{score}(f(x), f(x^+)) >> \text{score}(f(x), f(x^-))$$

x: reference sample; x^+ positive sample; x^- negative sample

Given a chosen score function, we aim to learn an **encoder function** *f* that yields high score for positive pairs (x, x^+) and low scores for negative pairs (x, x^-).

A formulation of contrastive learning

Loss function given 1 positive sample and N - 1 negative samples:

$$L = -\mathbb{E}_X \left[\log \frac{\exp(s(f(x), f(x^+))}{\exp(s(f(x), f(x^+)) + \sum_{j=1}^{N-1} \exp(s(f(x), f(x_j^-))} \right]$$

x

x^+

x

x_1^-

x_2^-

x_3^-

...

A formulation of contrastive learning

Loss function given 1 positive sample and N - 1 negative samples:

$$L = -\mathbb{E}_X \left[\log \frac{\exp(s(f(x), f(x^+))}{\exp(s(f(x), f(x^+)) + \sum_{j=1}^{N-1} \exp(s(f(x), f(x_j^-))} \right]$$

score for the positive pair

score for the N-1 negative pairs

This seems familiar …
Cross entropy loss for a N-way softmax classifier!
I.e., learn to find the positive sample from the N samples

A formulation of contrastive learning

Loss function given 1 positive sample and N - 1 negative samples:

$$L = -\mathbb{E}_X \left[\log \frac{\exp(s(f(x), f(x^+))}{\exp(s(f(x), f(x^+)) + \sum_{j=1}^{N-1} \exp(s(f(x), f(x_j^-))} \right]$$

Commonly known as the InfoNCE loss (van den Oord et al., 2018)

A *lower bound* on the mutual information between *f(x)* and $f(x^+)$

$$MI[f(x), f(x^+)] - \log(N) \geq -L$$

The larger the negative sample size (*N*), the tighter the bound

Detailed derivation: Poole et al., 2019

SimCLR: A Simple Framework for Contrastive Learning

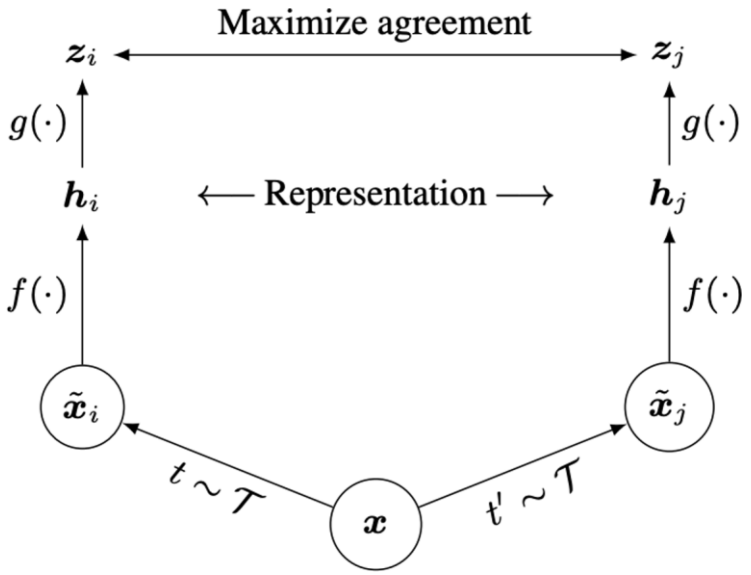
Cosine similarity as the score function:

$$s(u, v) = \frac{u^T v}{||u||||v||}$$

Use a projection network ***g(·)*** to project features to a space where contrastive learning is applied

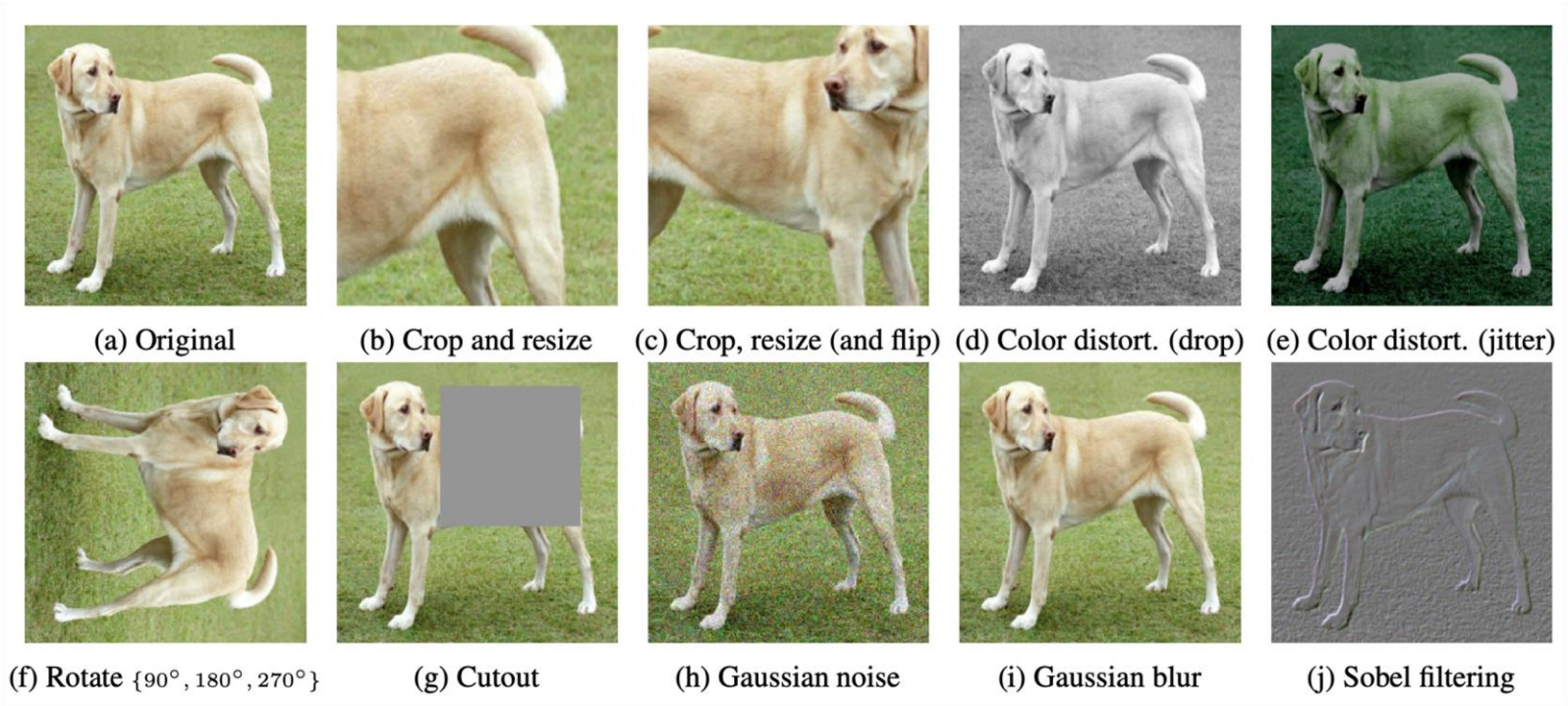
Generate positive samples through data augmentation:

- random cropping, random color distortion, and random blur.



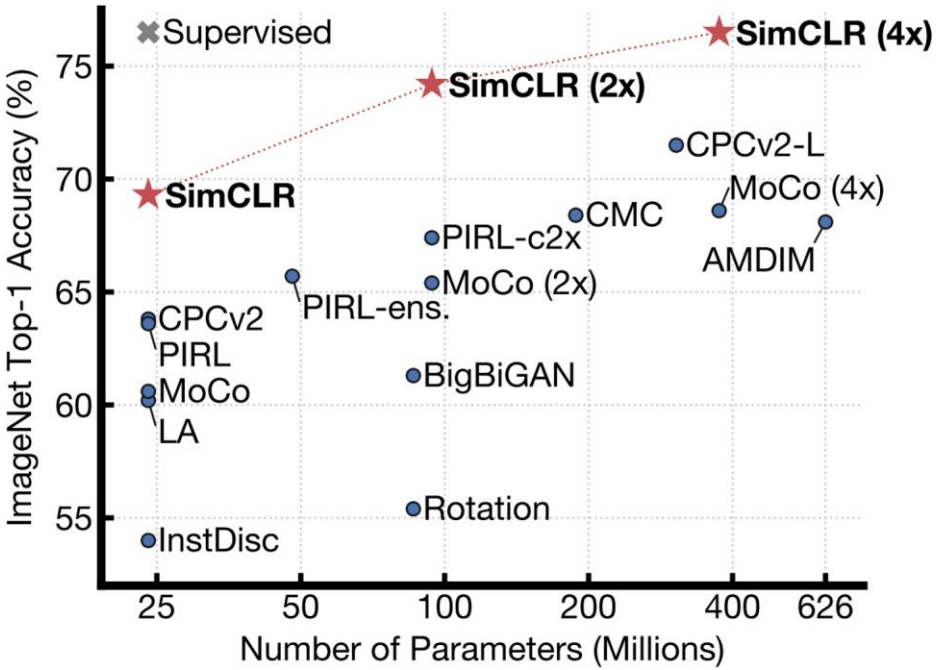
Source: Chen et al., 2020

SimCLR: generating positive samples from data augmentation



Source: Chen et al., 2020

Training linear classifier on SimCLR features



Train feature encoder on **ImageNet** (entire training set) using SimCLR.

Freeze feature encoder, train a linear classifier on top with labeled data.

Source: [Chen et al., 2020]

Semi-supervised learning on SimCLR features

Method	Architecture	Label fraction 1%	Label fraction 10%
		Top 5	
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(*)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

Table 7. ImageNet accuracy of models trained with few labels.

Train feature encoder on **ImageNet** (entire training set) using SimCLR.

Finetune the encoder with 1% / 10% of labeled data on ImageNet.

Source: Chen et al., 2020

SimCLR design choices: large batch size

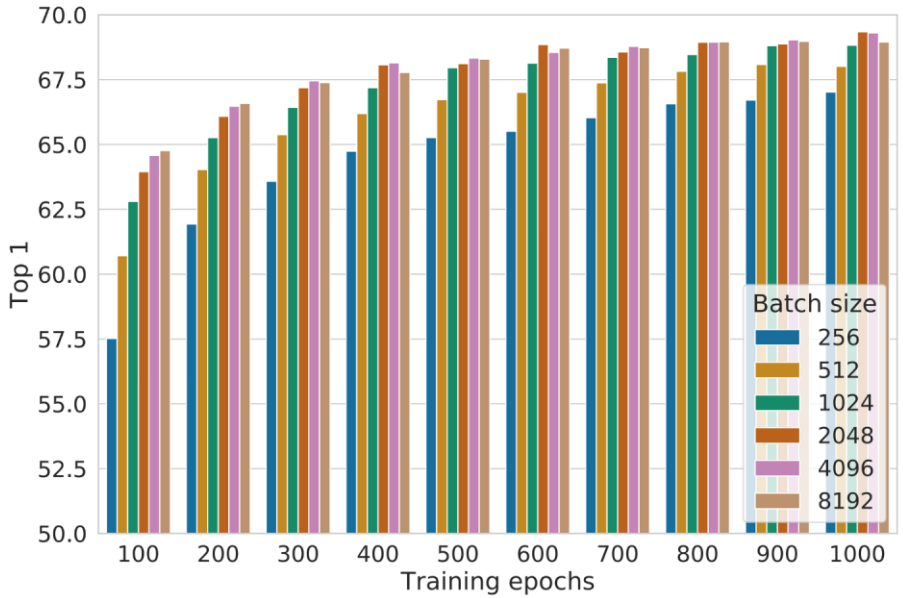


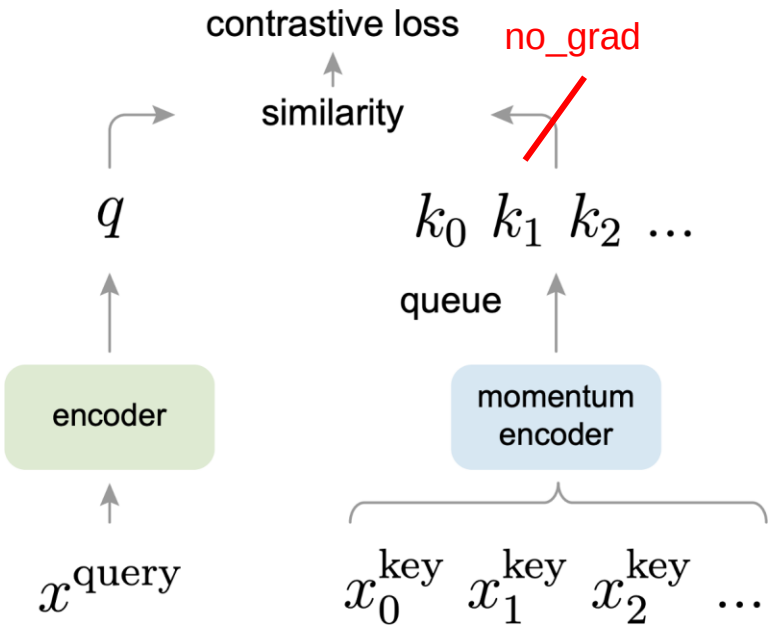
Figure 9. Linear evaluation models (ResNet-50) trained with different batch size and epochs. Each bar is a single run from scratch.[10]

Large training batch size is crucial for SimCLR!

Large batch size causes large memory footprint during backpropagation: requires distributed training on TPUs (ImageNet experiments)

Source: Chen et al., 2020

Momentum Contrastive Learning (MoCo)



Key differences to SimCLR:

- Keep a running queue of keys (negative samples).
- Compute gradients and update the encoder only through the queries.
- Decouple min-batch size with the number of keys: can support a large number of negative samples.
- The key encoder is slowly progressing through the momentum update rules:

$$\theta_{\text{k}} \leftarrow m\theta_{\text{k}} + (1-m)\theta_{\text{q}}$$

Source: He et al., 2020

MoCo Results: Transferring Features

VOC 2007 Detection, Faster R-CNN, ResNet-50

pre-train	AP_{50} RelPos, by [10]	Multi-task [10]	Jigsaw, by [22]	LocalAgg [60]	**MoCo**
super. IN-1M	74.2	74.2	70.5	74.6	74.4
unsup. IN-1M	66.8 (−7.4)	70.5 (−3.7)	61.4 (−9.1)	69.1 (−5.5)	74.9 (+**0.5**)
unsup. IN-14M	-	-	69.2 (−1.3)	-	75.2 (+**0.8**)
unsup. IG-1B	-	-	-	-	75.6 (+**1.2**)

MoCo:
benefit from 1 billion images

He et al., "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020

Other examples: DINO

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron[1,2] Hugo Touvron[1,3] Ishan Misra[1] Hervé Jegou[1]
Julien Mairal[2] Piotr Bojanowski[1] Armand Joulin[1]

[1] Facebook AI Research [2] Inria* [3] Sorbonne University

Figure 1: **Self-attention from a Vision Transformer with 8×8 patches trained with no supervision.** We look at the self-attention of the `[CLS]` token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

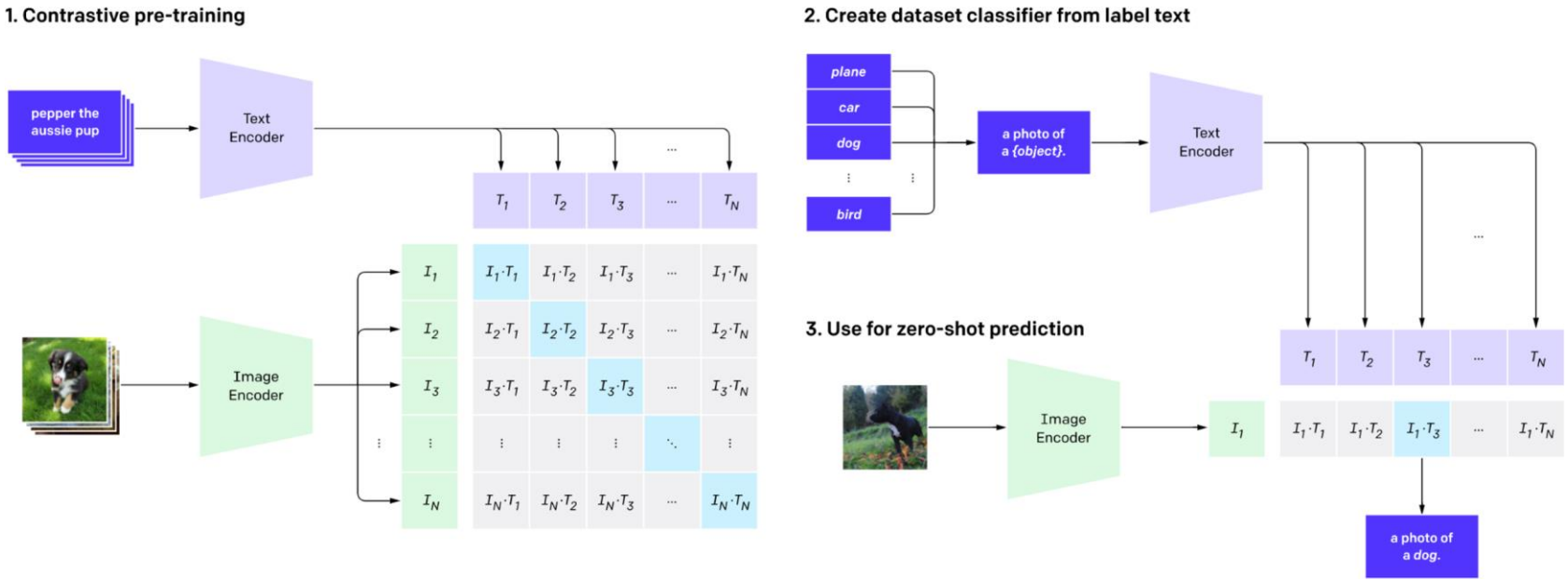
Other examples: Masked Autoencoder

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

He et al., Masked Autoencoders Are Scalable Vision Learners, FAIR

Other examples: CLIP

Contrastive learning between image and natural language sentences



CLIP (*Contrastive Language–Image Pre-training*) Radford *et al.*, 2021