Self-supervised learning in computer vision

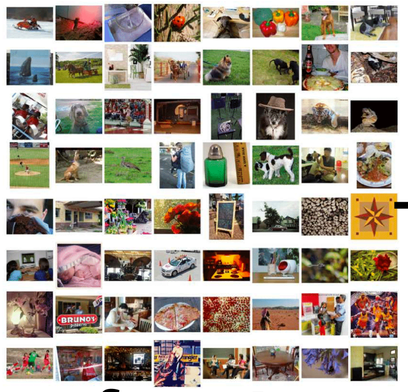
Ishan Misra

Facebook AI Research

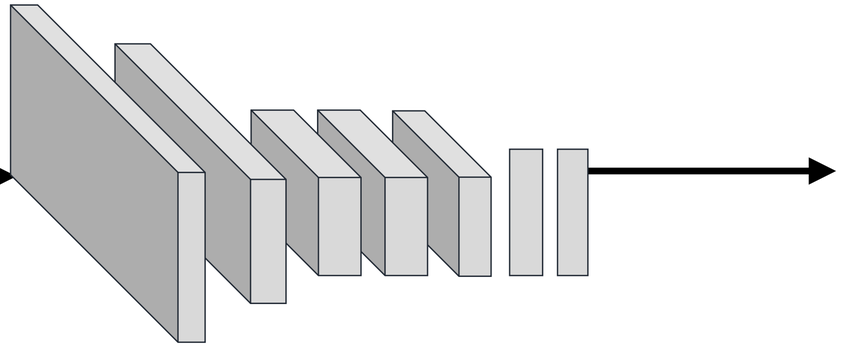
With slides from Andrew Zisserman, Carl Doersch

Success story of supervision: Pre-training

- Features from networks pre-trained on ImageNet can be used for a variety of different downstream tasks



Images from ImageNet (Pre-train)



ConvNet

husky, terrier, tench, ...

Learn a representation

Success story of supervision: Recipe for good solutions

- Pre-train on a large supervised dataset.
- Collect a dataset of ″supervised″ images
- Train a ConvNet

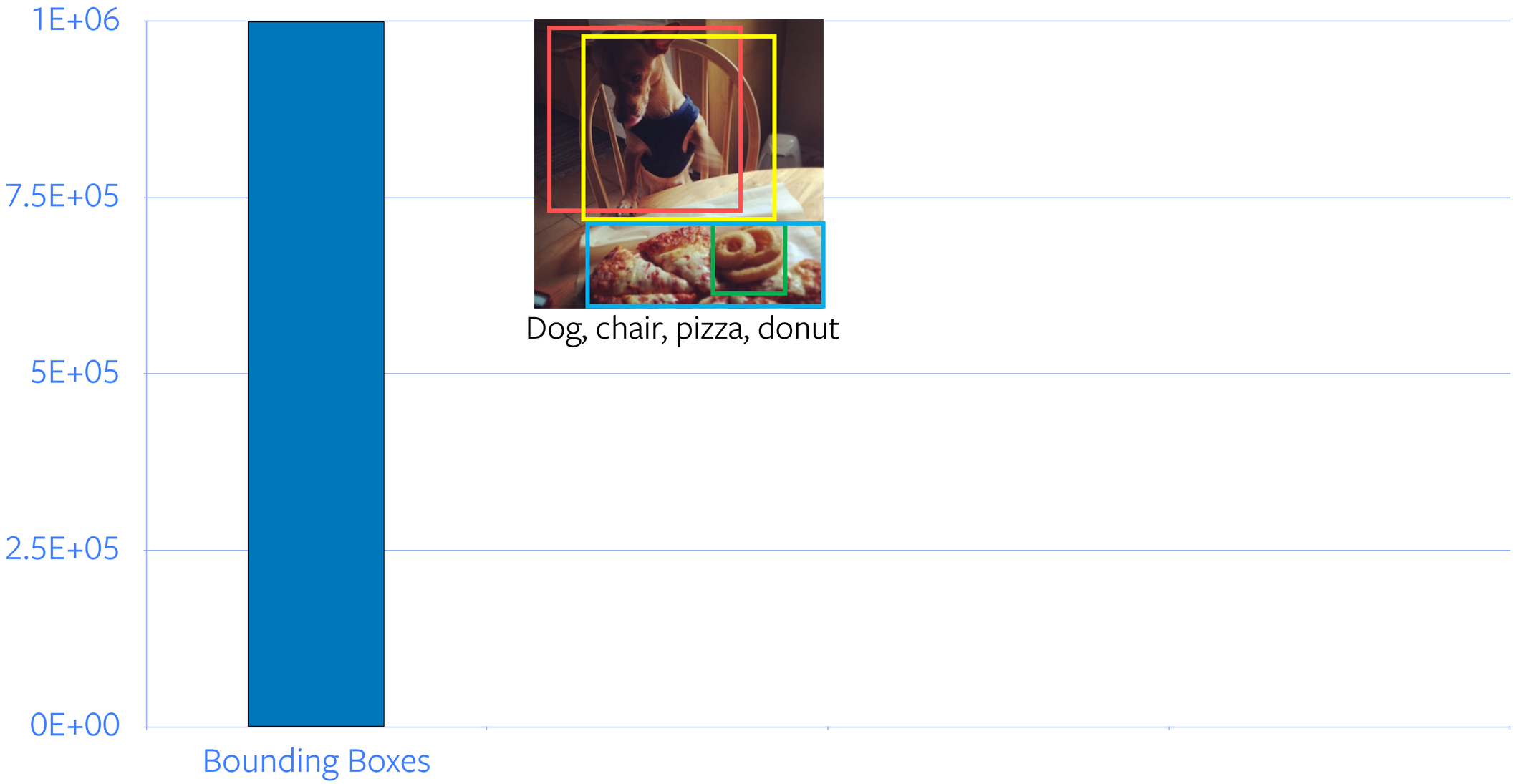
The promise of "alternative" supervision

- Getting "real" labels is difficult and expensive
  - ImageNet with 14M images took 22 human years.

- Obtain labels using a "semi-automatic" process
  - Hashtags
  - GPS locations
  - Using the data itself: "self"-supervised

Can we get labels for all data?

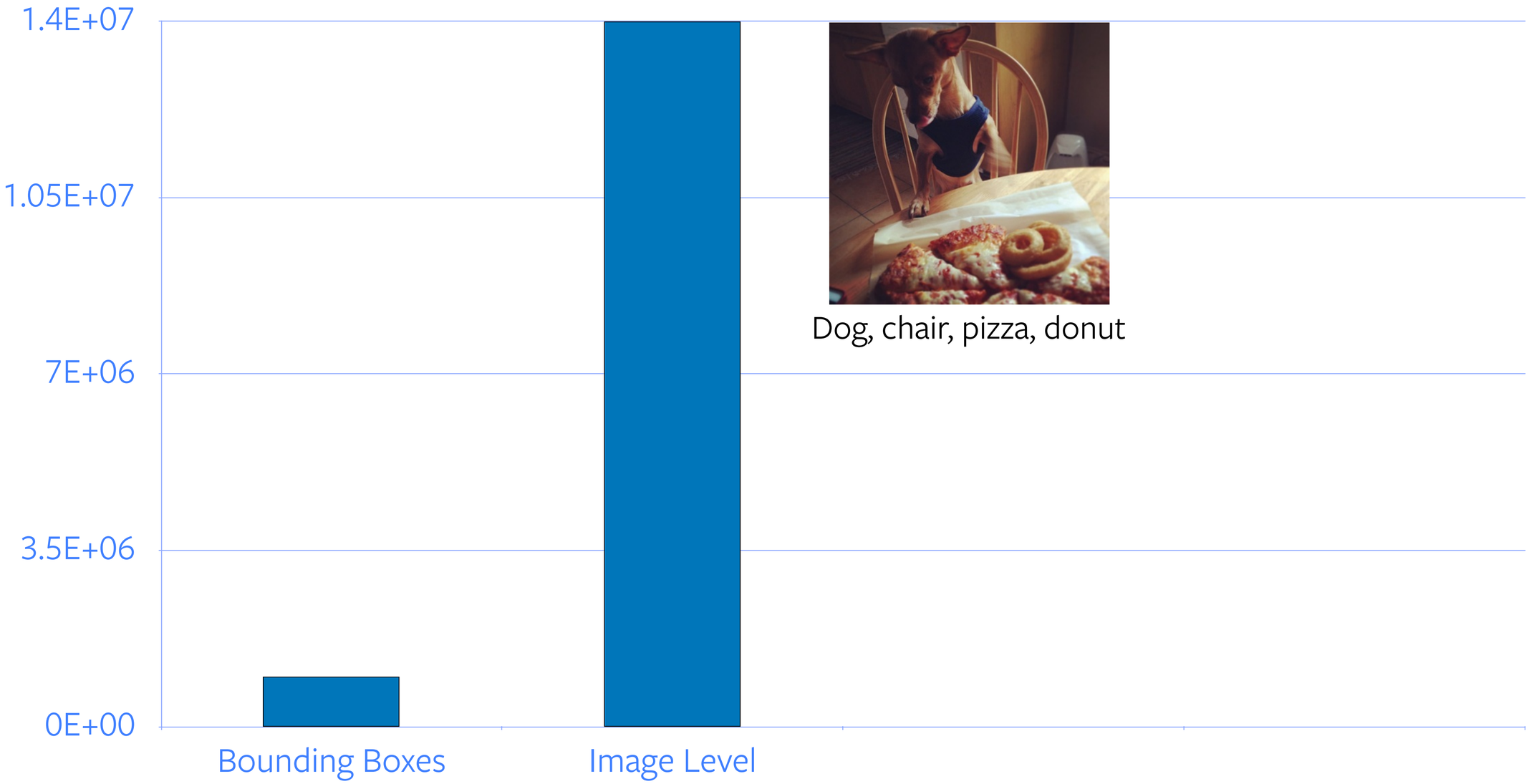
Can we get labels for all data?



Dog, chair, pizza, donut

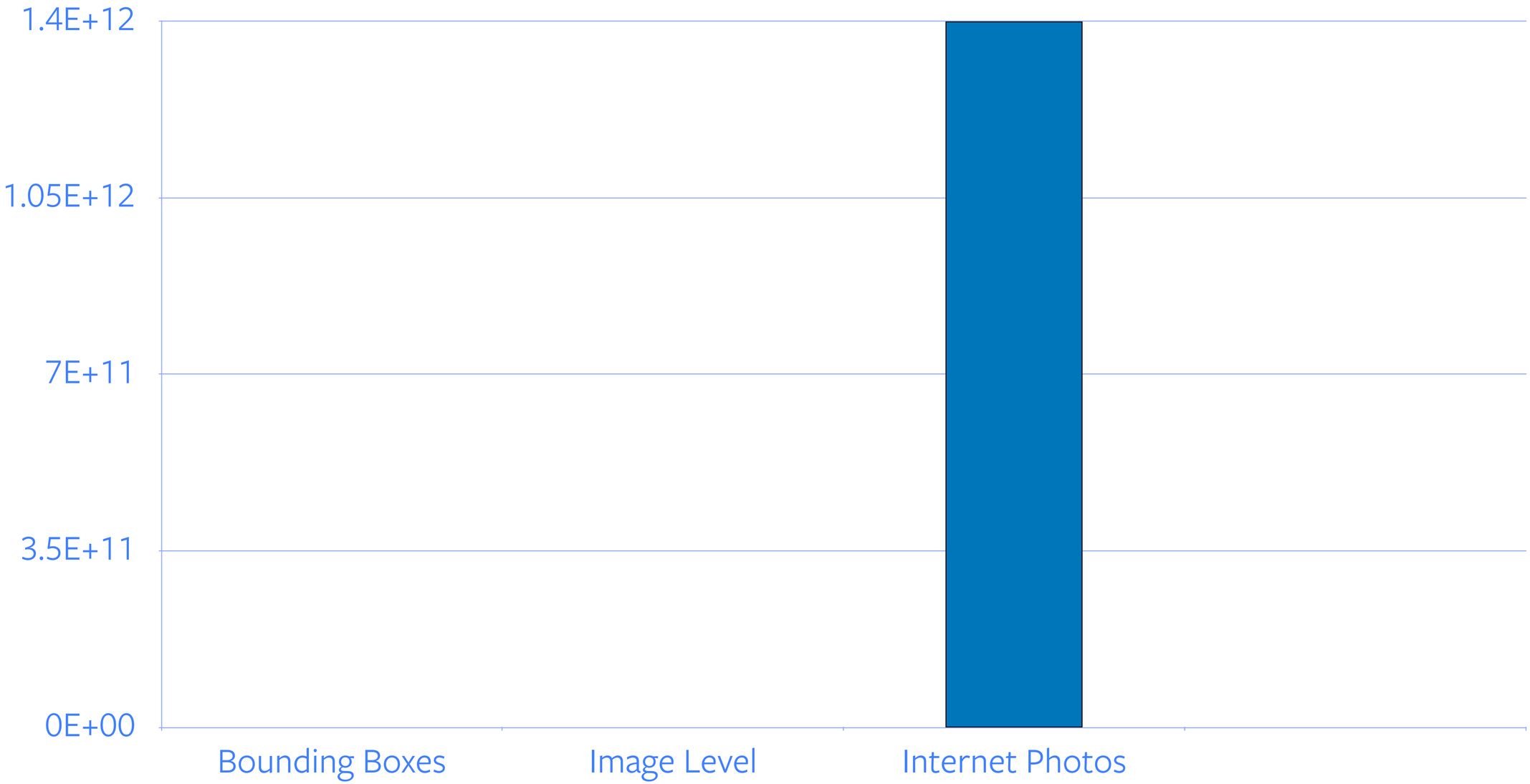
Stats from Pawan Kumar at Oxford

Can we get labels for all data?



Dog, chair, pizza, donut

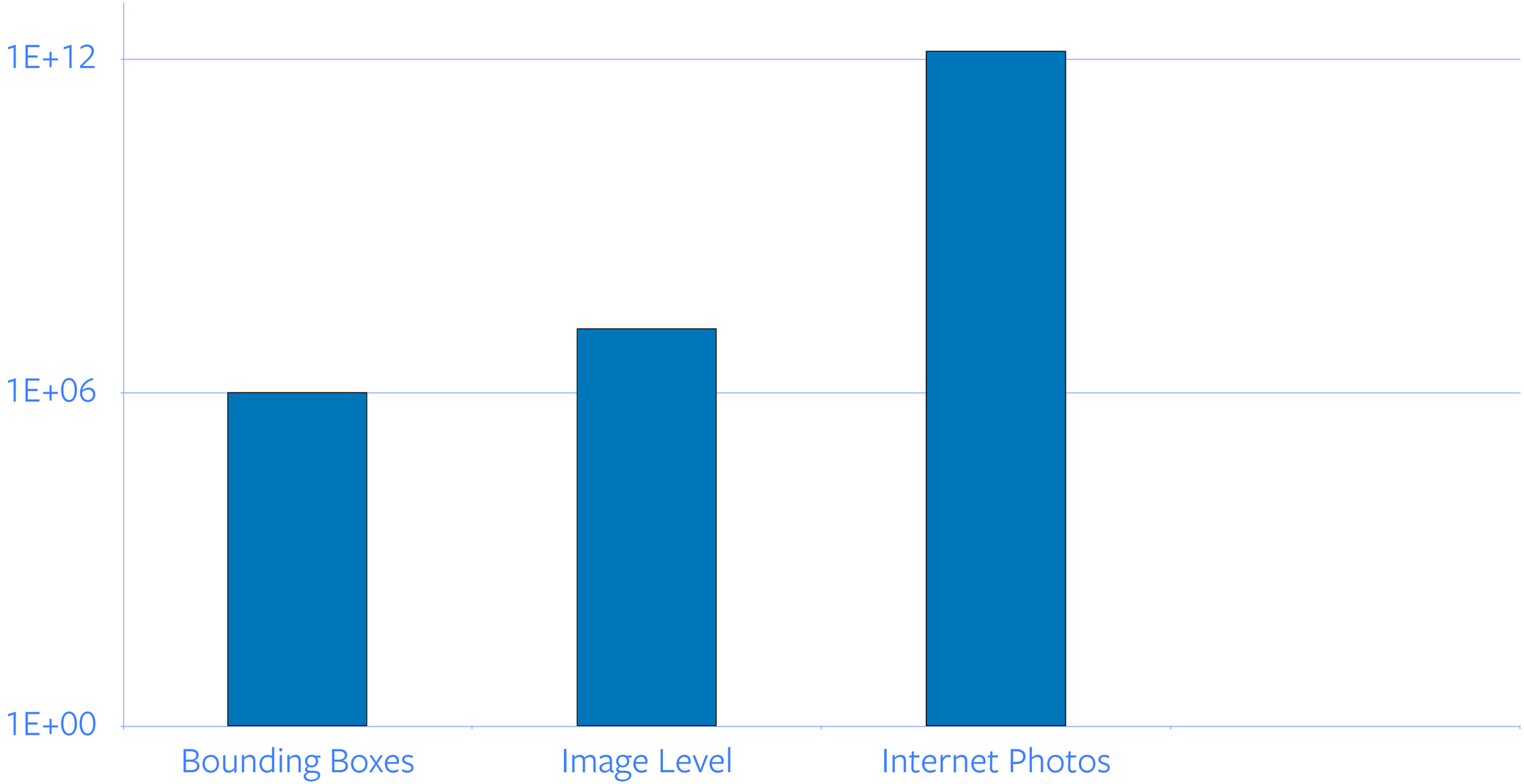
Can we get labels for all data?



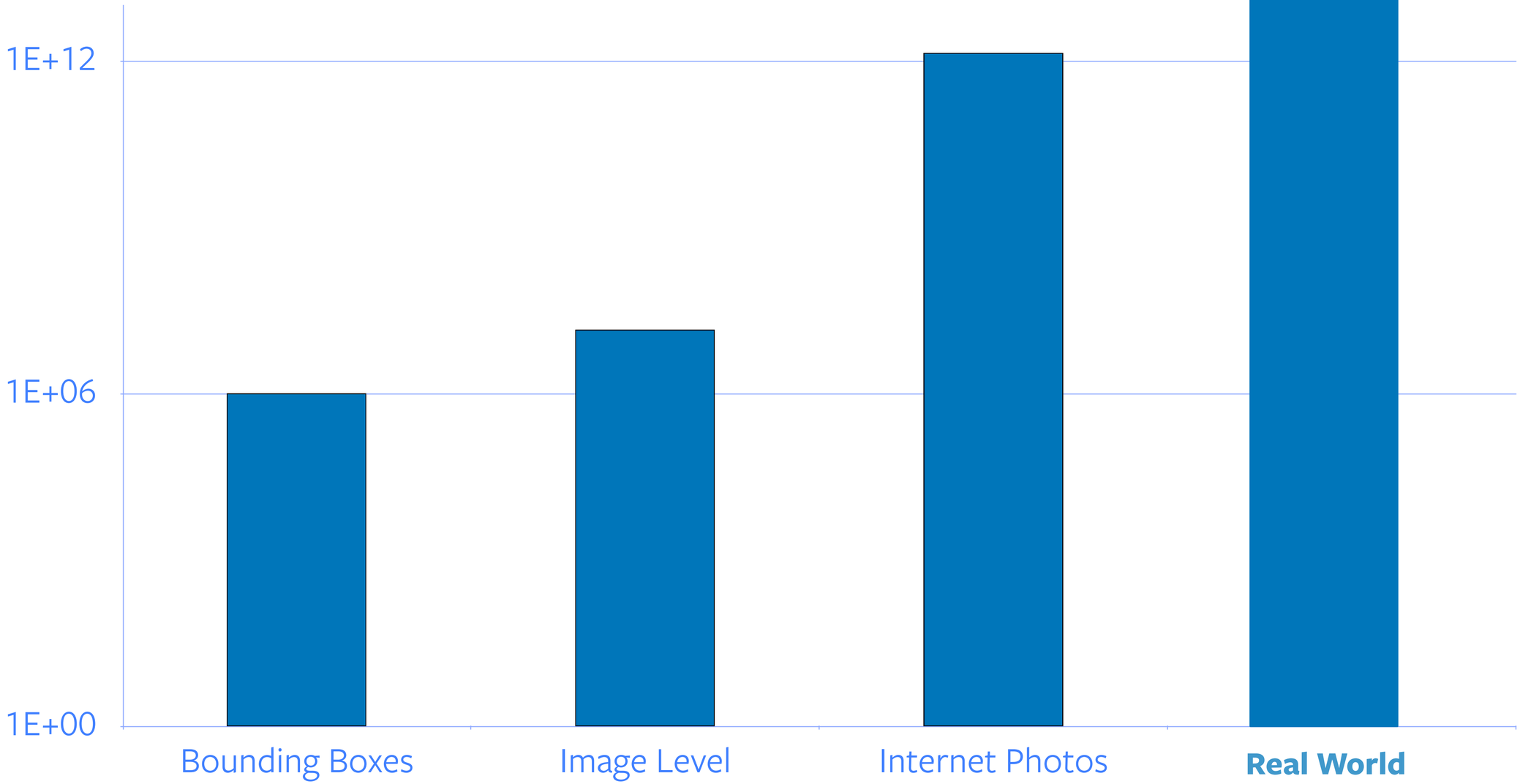
forbes.com

https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

Can we get labels for all data?

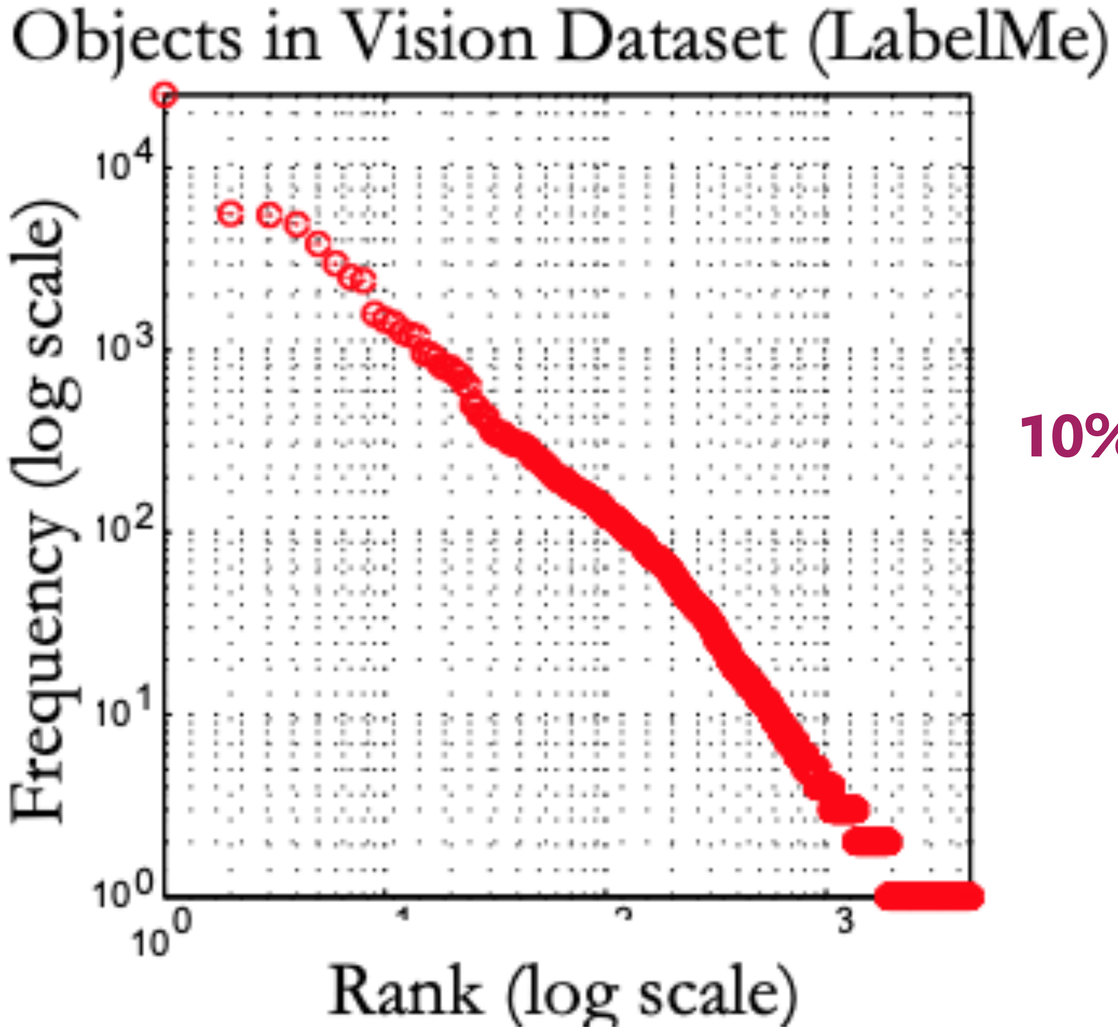


Can we get labels for all data?



ImageNet (14 million images) needed 22 human years to label

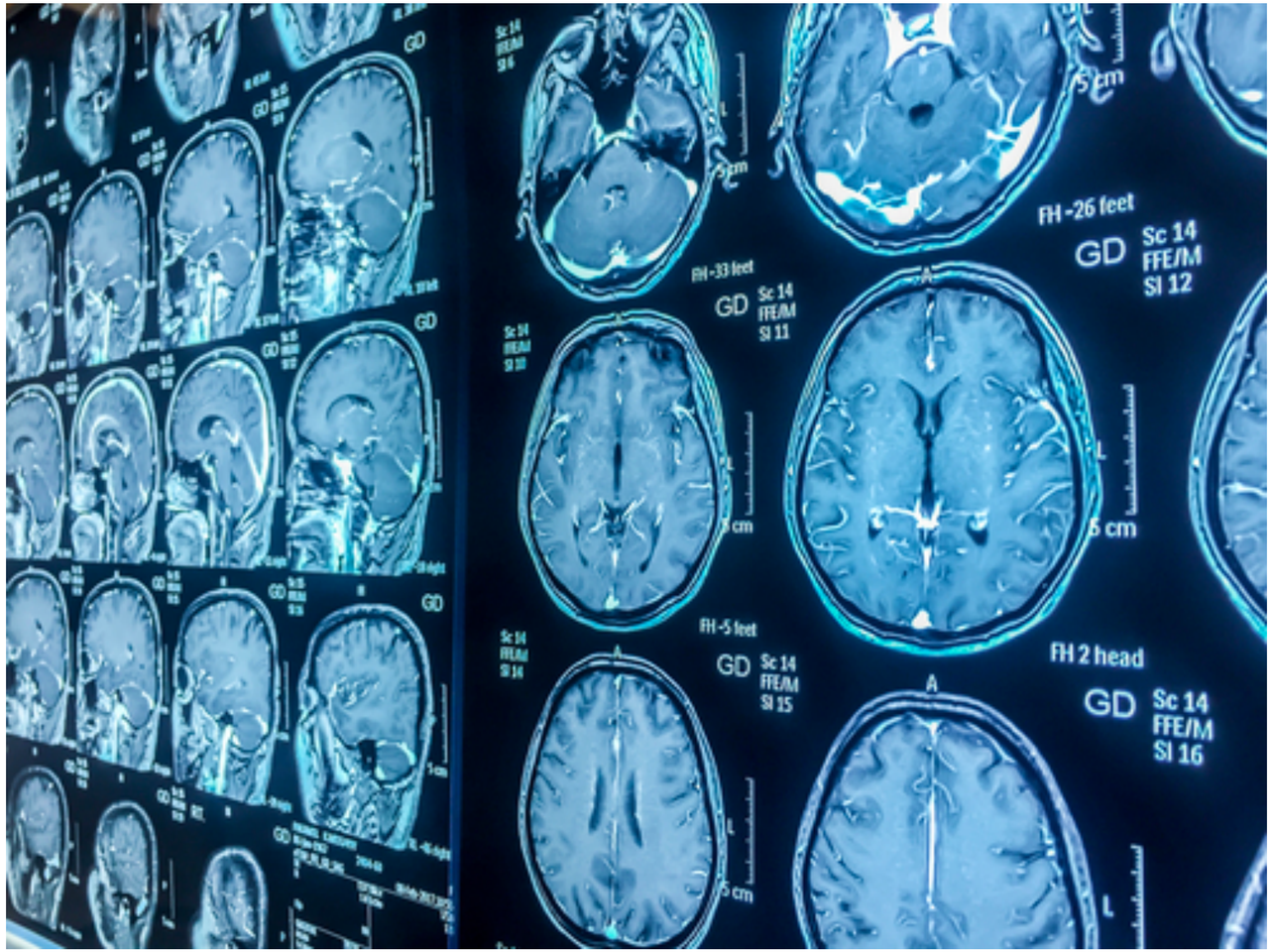
Rare concepts?



10% of the classes account for 93% of the data

Slide credit: Rob Fergus

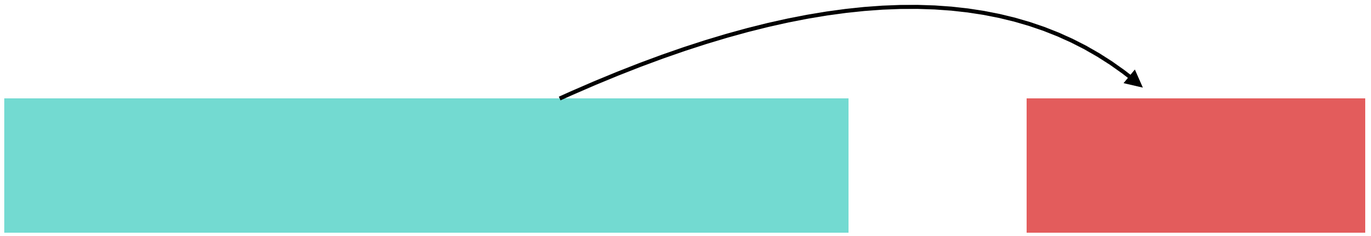
Different Domains?



ImageNet pre-training may not work

What is "self" supervision?

- Obtain "labels" from the data itself by using a "semi-automatic" process
- Predict part of the data from other parts



Observed data

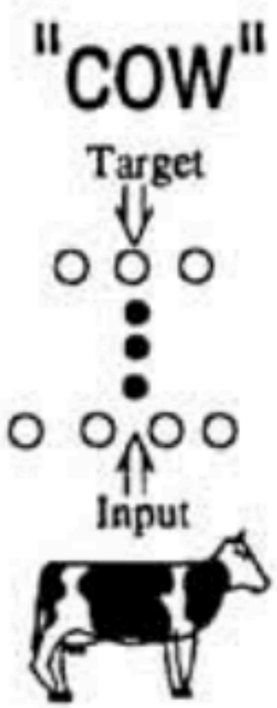
Hidden data

Hidden property of the data

What is "self" supervision?

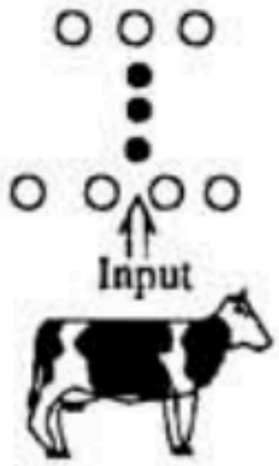
Supervised

- implausible label



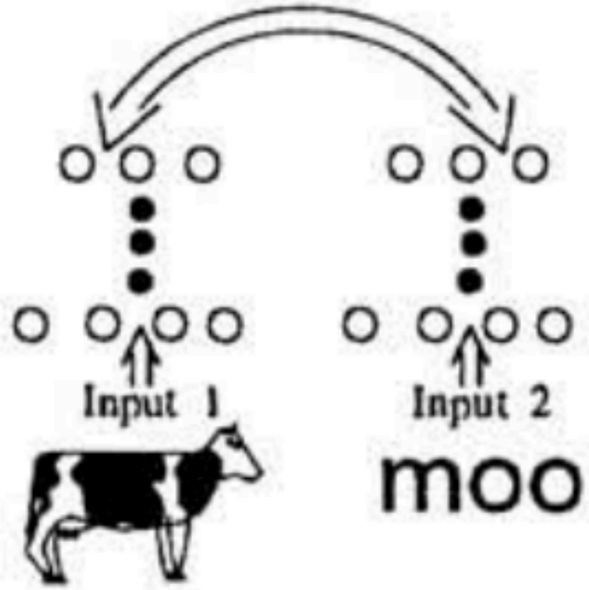
Unsupervised

- limited power



Self-Supervised

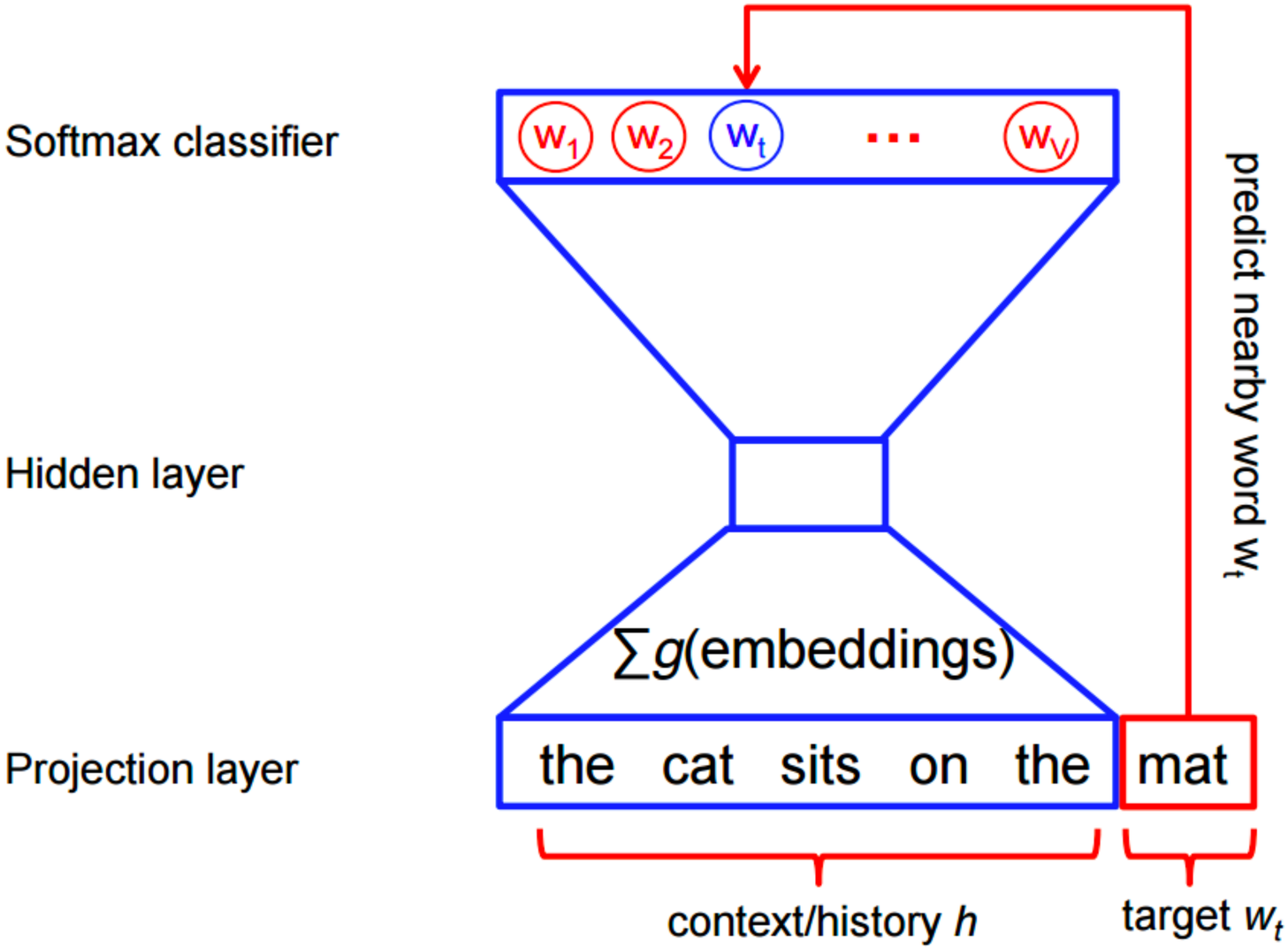
- derives label from a co-occuring input to another modality



Virginia de Sa, 1994, Image: Learning classification with Unlabeled Data

Word2vec

- Fill in the blanks



word2vec - Mikolov et al.
Image by Julian Gilyadov

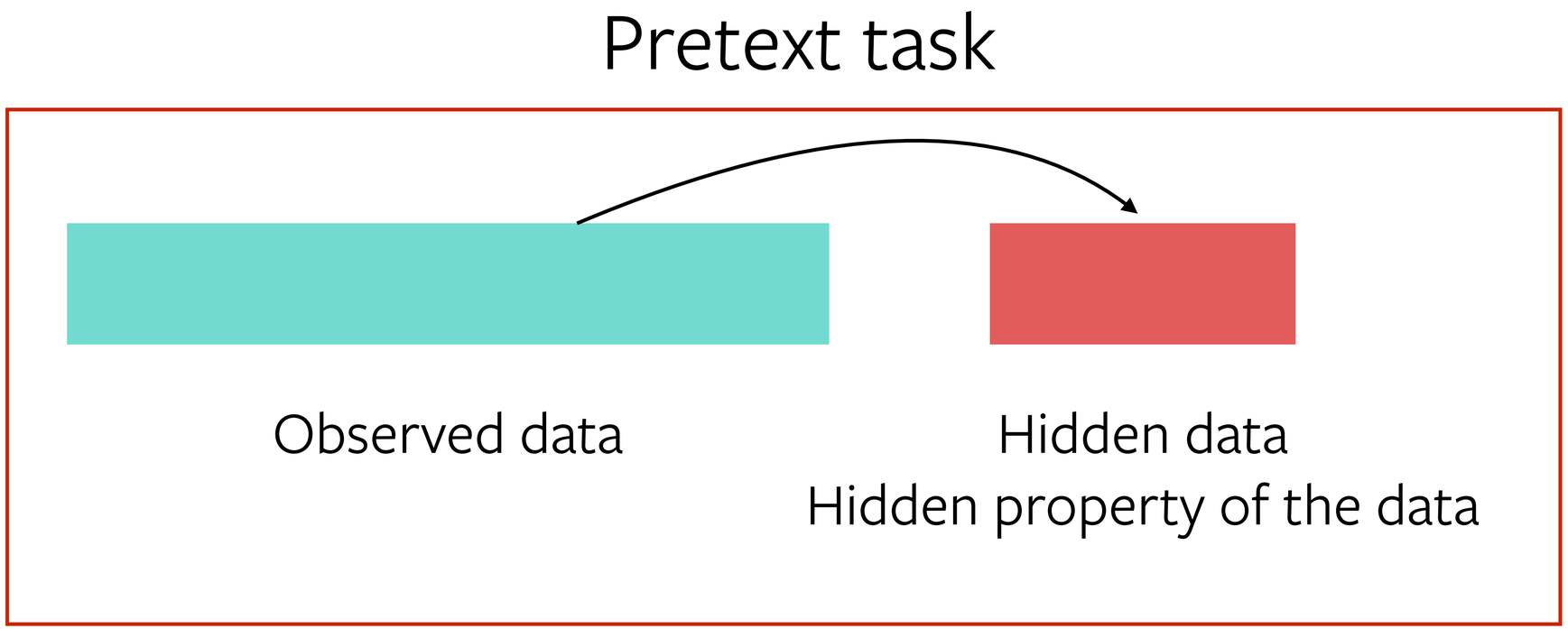
Success of self-supervised learning in NLP

- Fill in the blanks is a powerful signal to learn representations

- Sentence/Word representations: BERT - Devlin et al., 2018

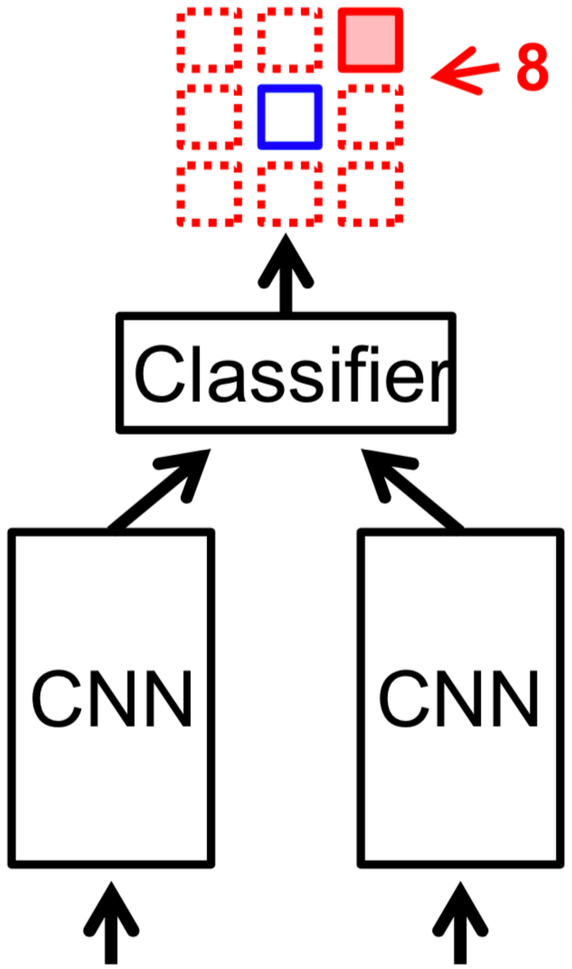
Pretext task

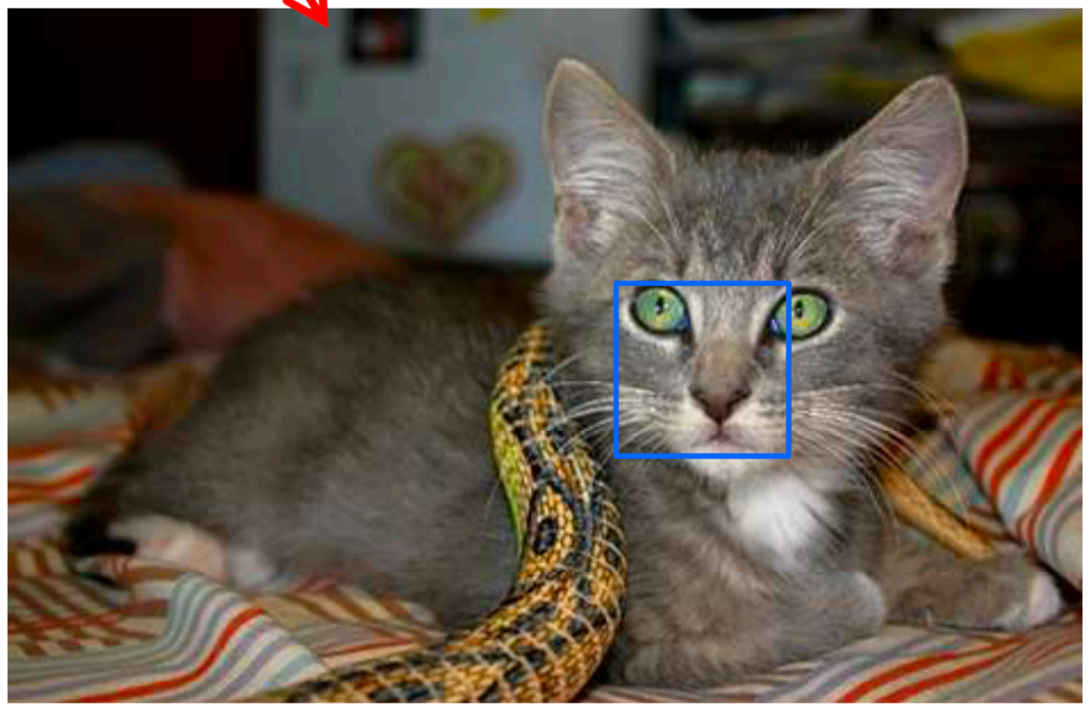
- Self-supervised task used for learning representations
- Often, not the "real" task (like image classification) we care about



Pretext task - Doersch et al., 2015, Unsupervised visual representation learning by context prediction

Relative Position of patches

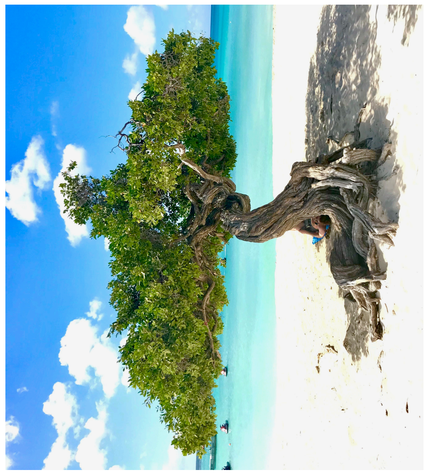




Doersch et al., 2015, Unsupervised visual representation learning by context prediction

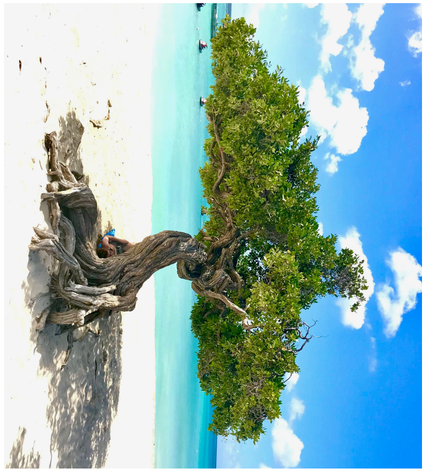
Predicting Rotations

0^0



90^0

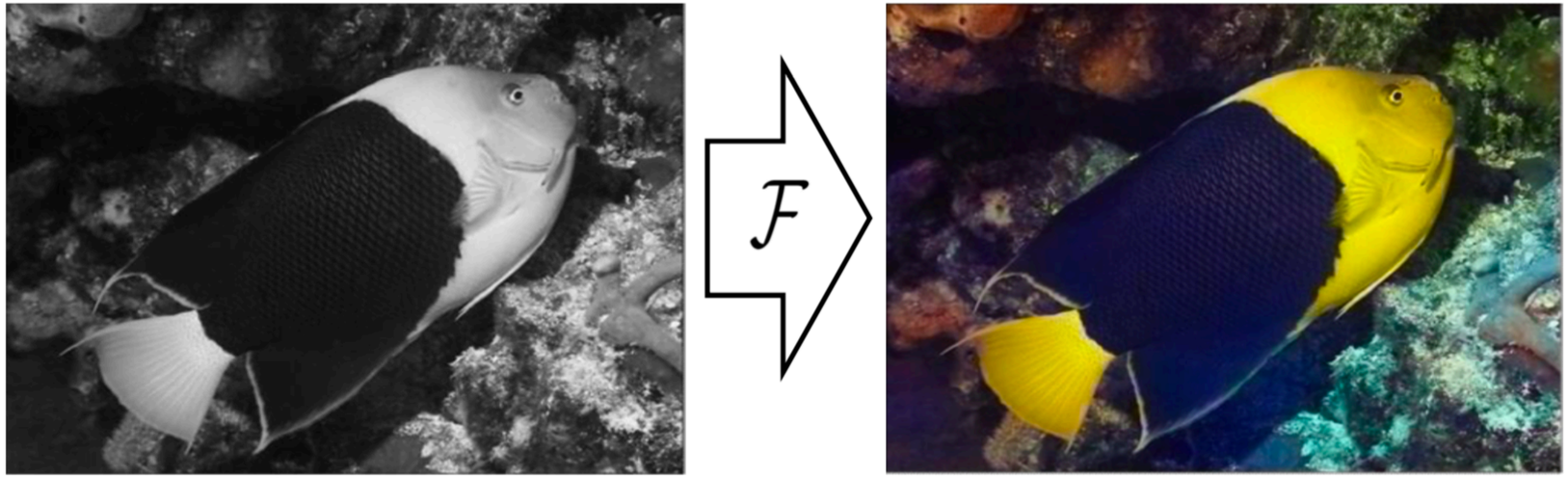
180^0



270^0

Gidaris et al., 2018, Predicting Image Rotations

Colorization

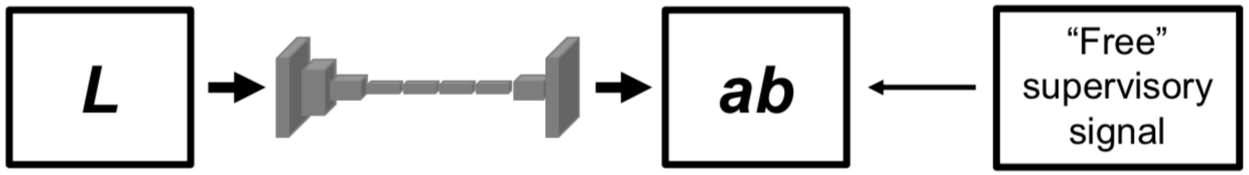


Grayscale image: *L* channel

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

Concatenate (L,ab)

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



Zhang and Efros, 2016, Colorful image colorization

Fill in the blanks

Pathak et al., 2016, Context auto encoders

Understanding what the ″pretext″ task learns

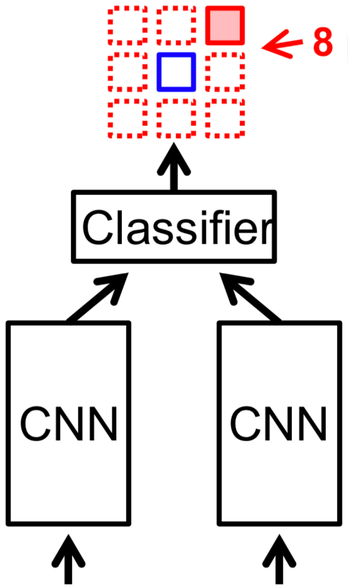
Are they complementary?

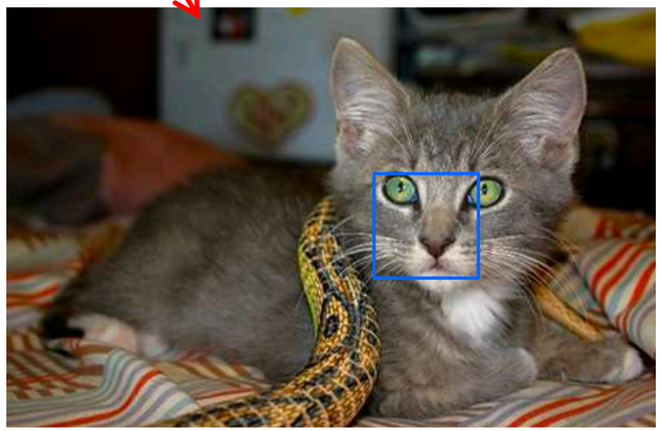
Initialization (ResNet101)	End task	
	ImageNet top-5 accuracy	VOC07 Detection mAP
Relative Position	59.2	66.8
Colorization	62.5	65.5
Relative Position + Colorization (Multi-task)	66.6	68.8

Doersch & Zisserman, 2017, Multi-task self-supervised visual learning

Information predicted: varies across tasks

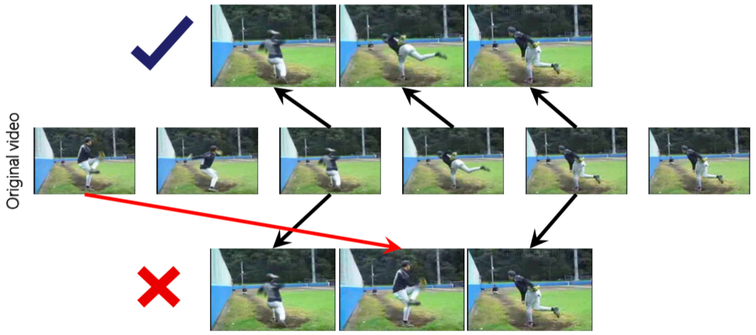
Less → More

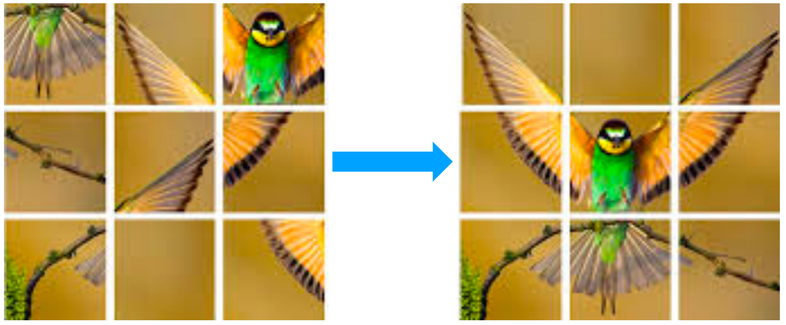




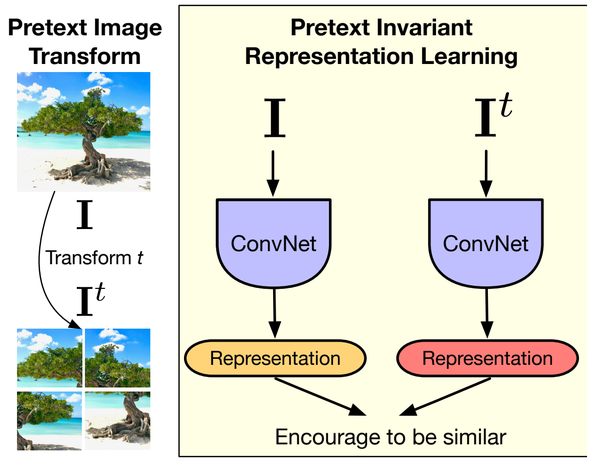
Randomly Sample Patch
Sample Second Patch

Pretext tasks

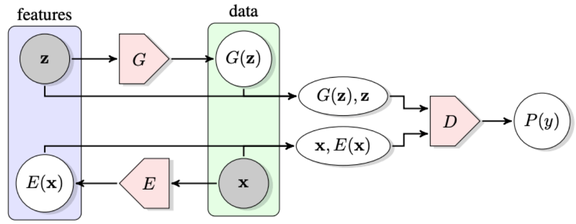




Contrastive



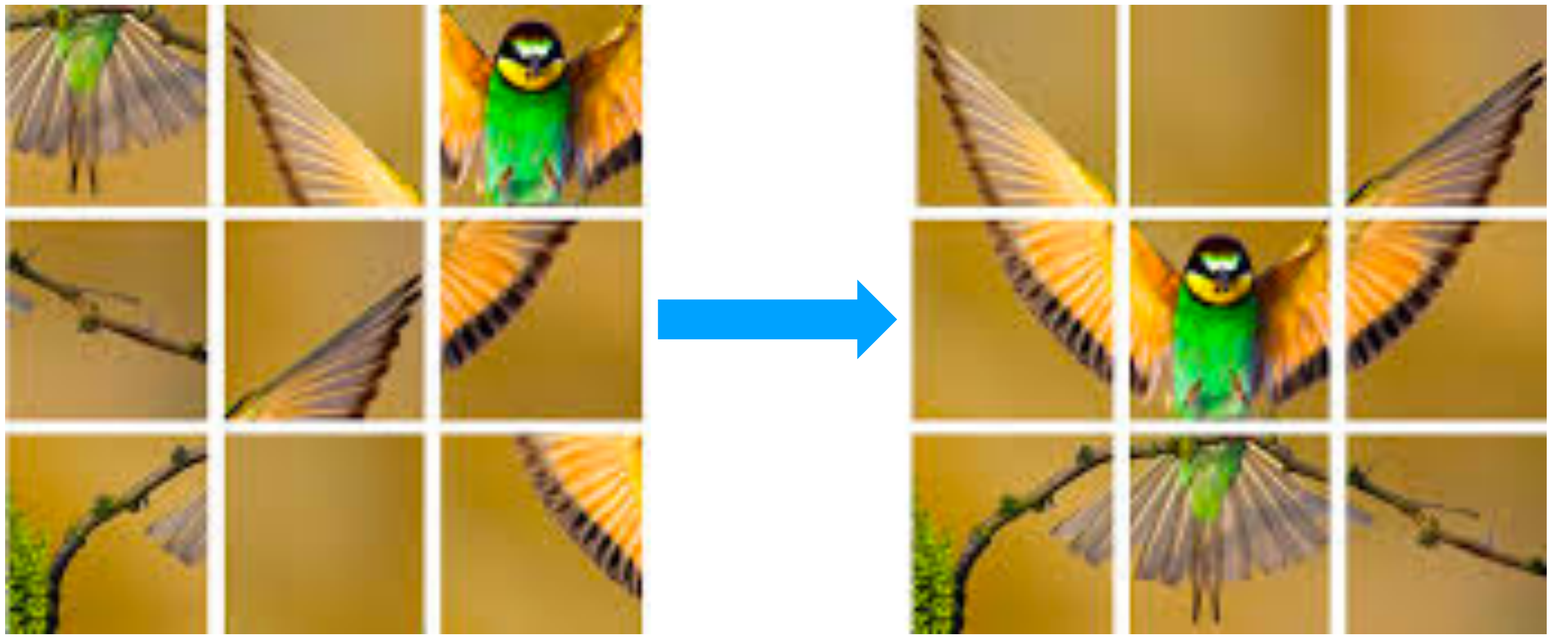
Generative



AutoEncoder,
VAE, GAN,
BiGAN

Predict more information

Scaling self-supervised learning

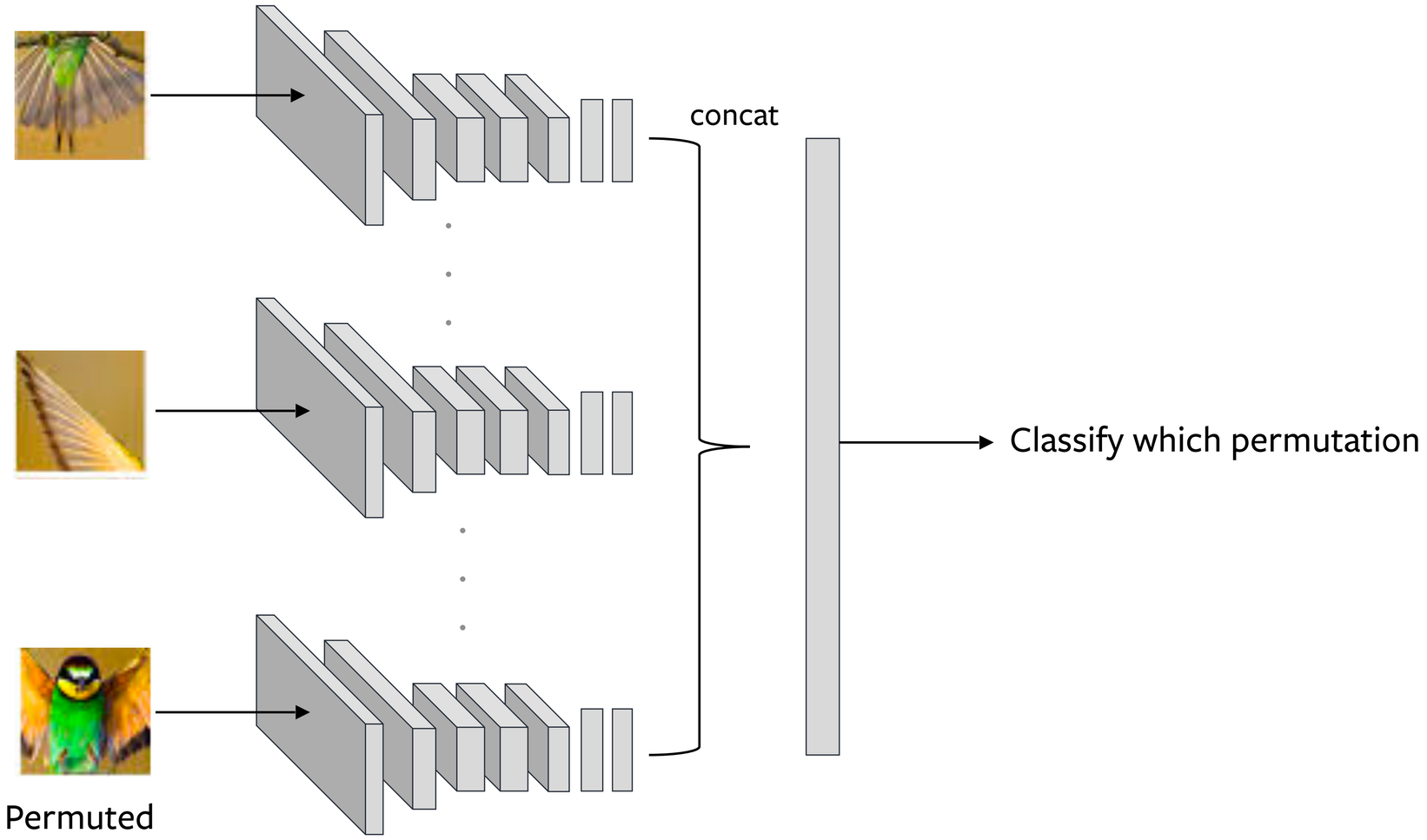


Jigsaw puzzles

(Noorozi & Favaro, 2016)

Goyal et al., 2019, Scaling and benchmarking self-supervised visual representation learning

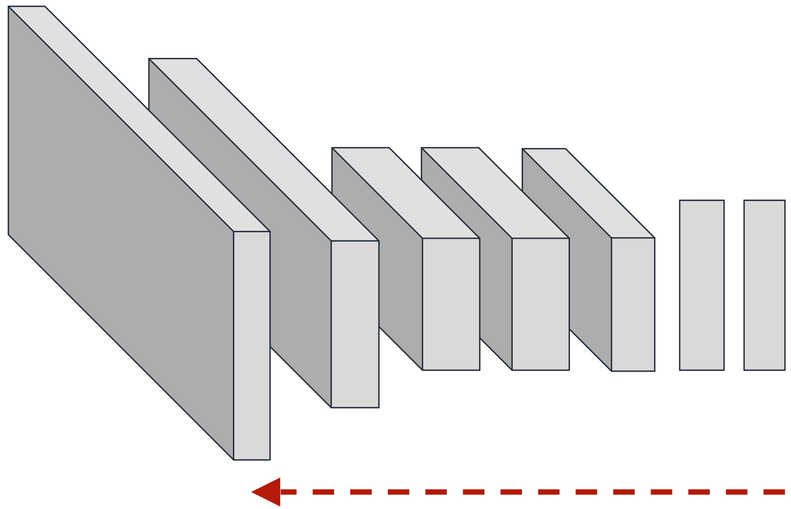
Jigsaw Puzzles



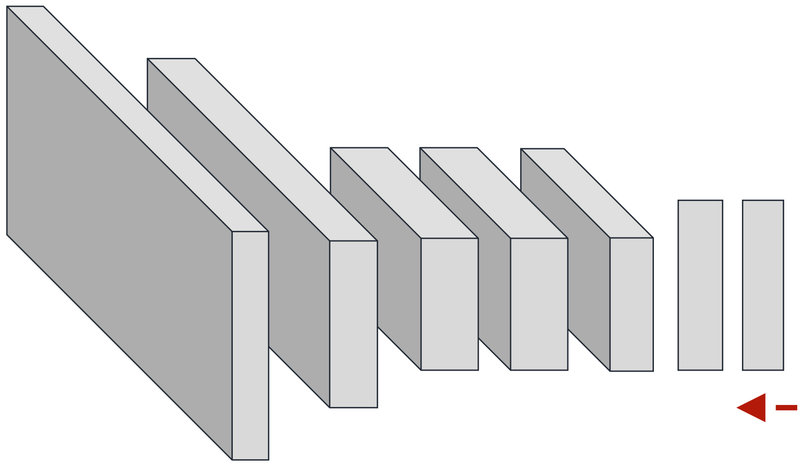
- Use N=9 patches
- In practice, use a subset of permutations
- E.g. 100 from 9!
- Each patch is processed independently
- N-way ConvNet (shared params)

- Problem Complexity
  - Size of subset

Evaluation – fine-tuning vs. linear classifier



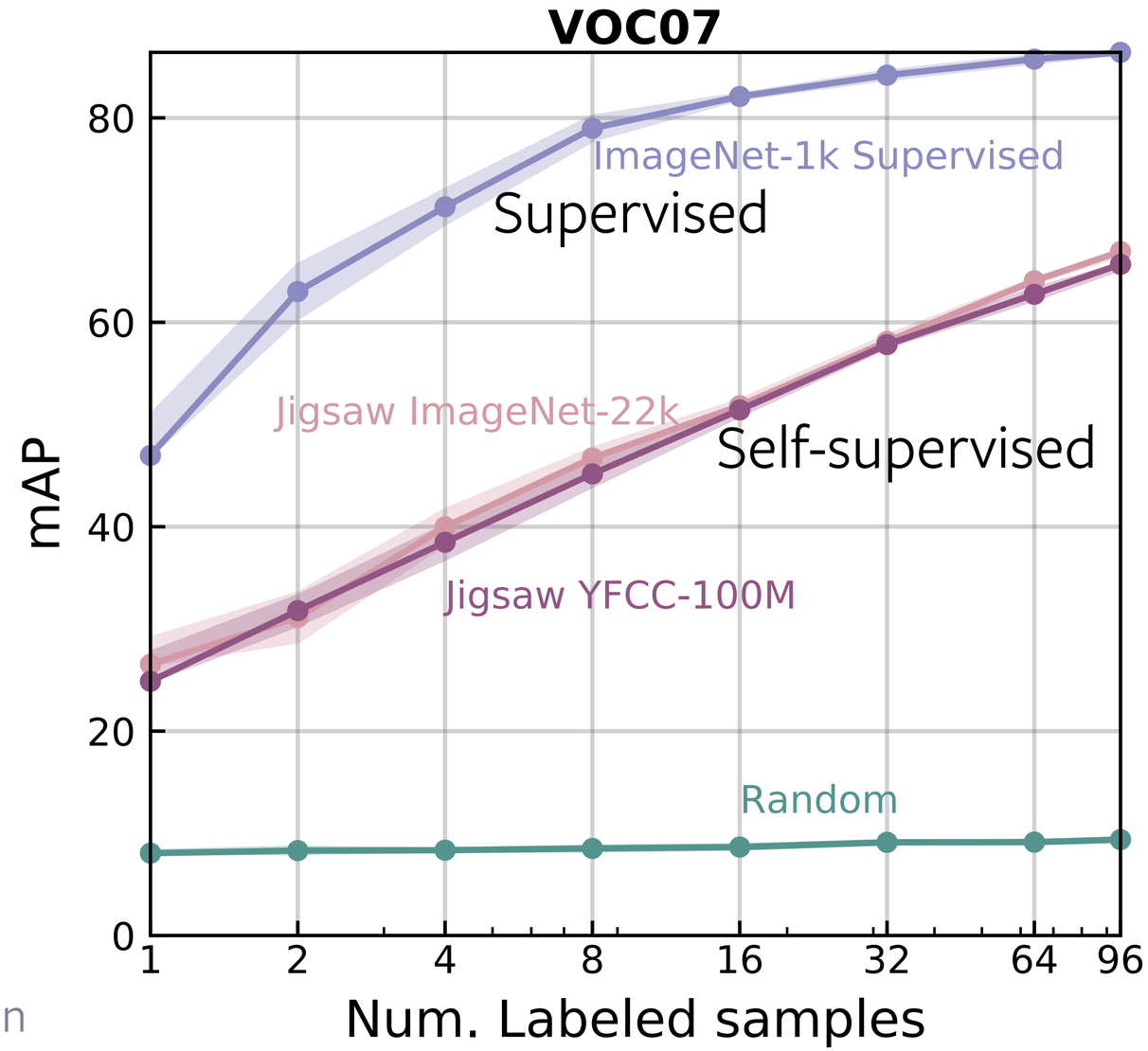
Fine-tune all layers



Linear classifier

*A good representation transfers with **little training***

Few shot learning



mAP = mean Average Precision
(Higher is better)

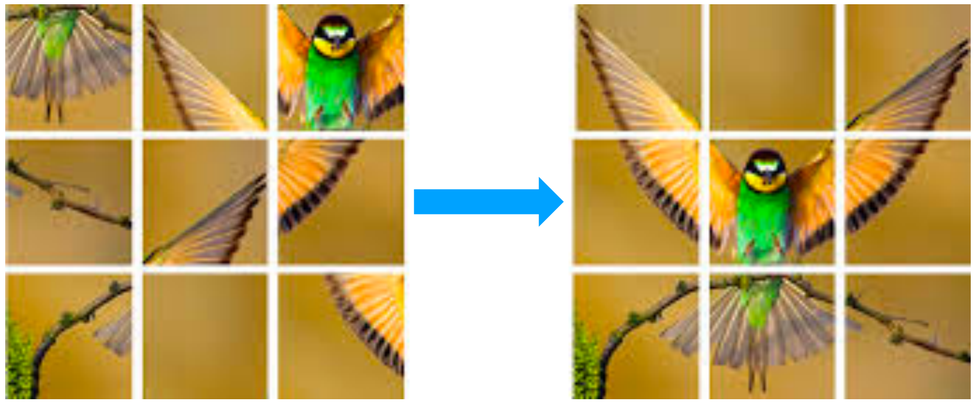
Self-supervised representations are not as sample efficient

What is missing from ″pretext″ tasks?
Or in general ″proxy″ tasks

Pretext tasks

Rotation

(Gidaris et al., 2018)

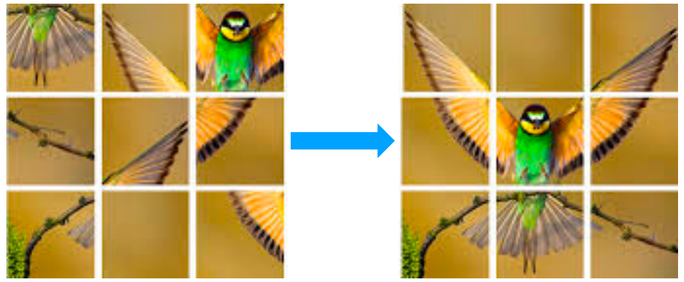


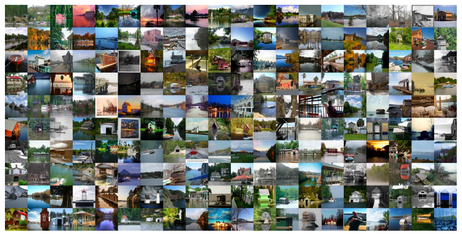
Jigsaw puzzles

(Noroozi et al., 2016)

The hope of generalization

- We really **hope** that the pre-training task and the transfer task are "aligned"





Pre-training

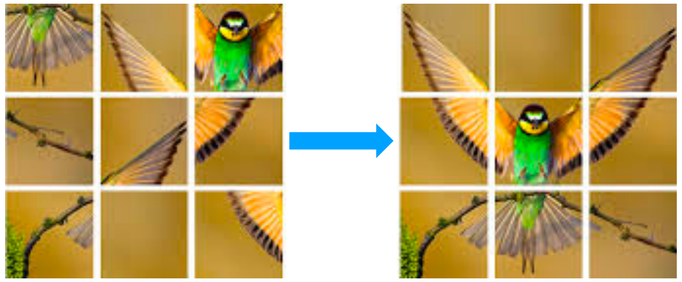
Self-supervised

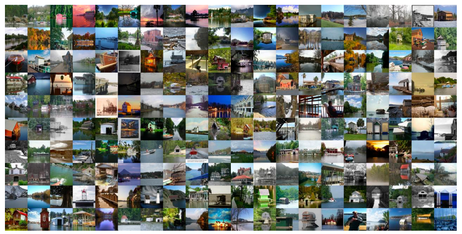
Transfer Tasks

The hope of generalization

- We really **hope** that the pre-training task and the transfer task are "aligned"

#sun #nofilter #fun
#tree #aruba





Pre-training

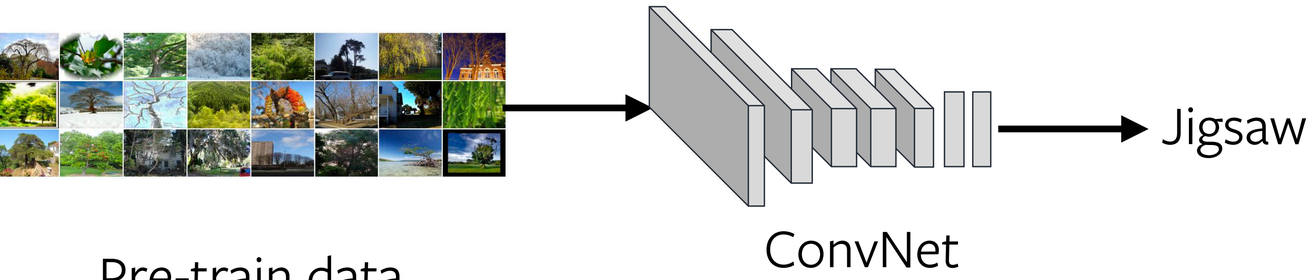
Weak or self-supervised

Transfer Tasks

Why should solving Jigsaw puzzles teach about "semantics"?

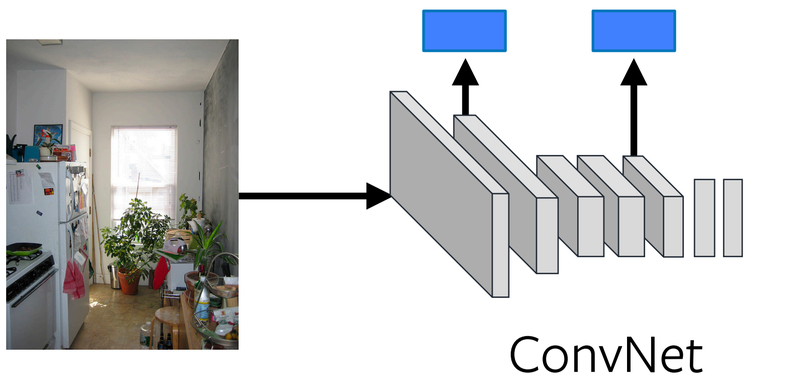
Why should performing a non semantic task produce good features?

The hope of generalization ... ?



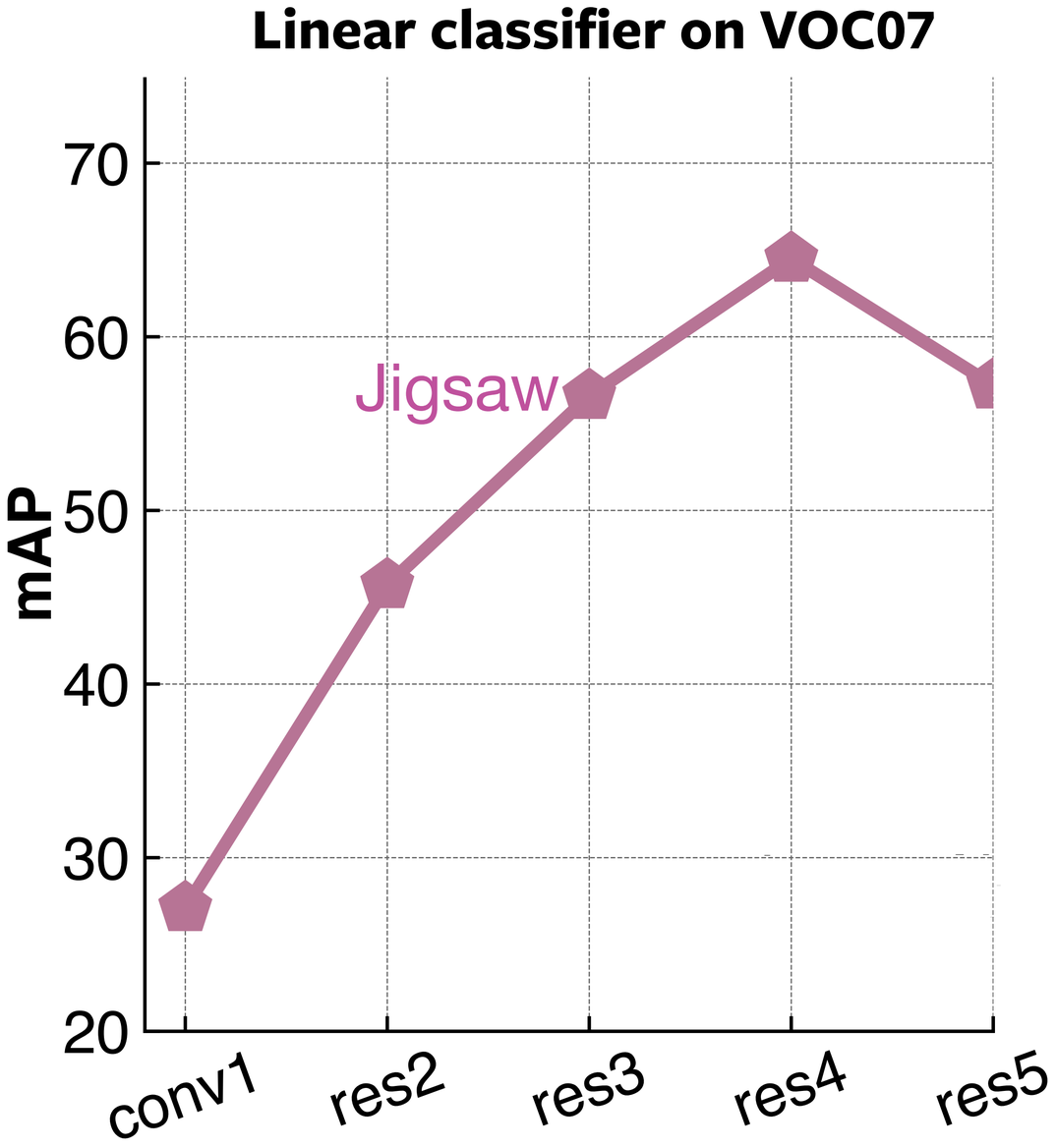
Pre-training

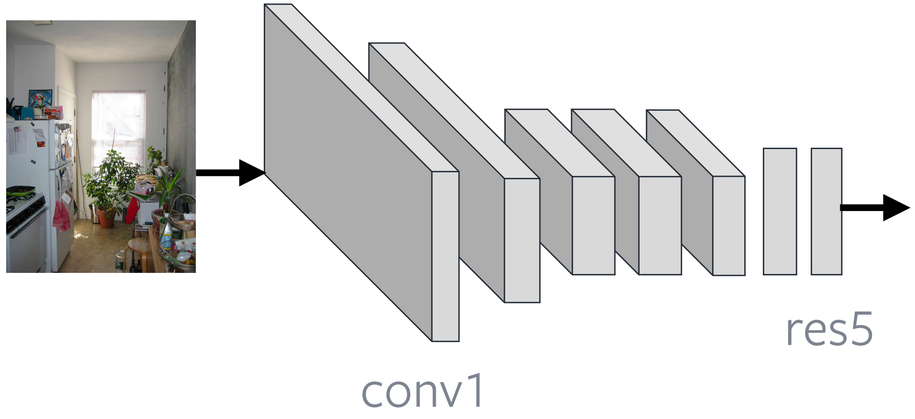
Weak or self-supervised



Transfer

Higher layers do not generalize ...





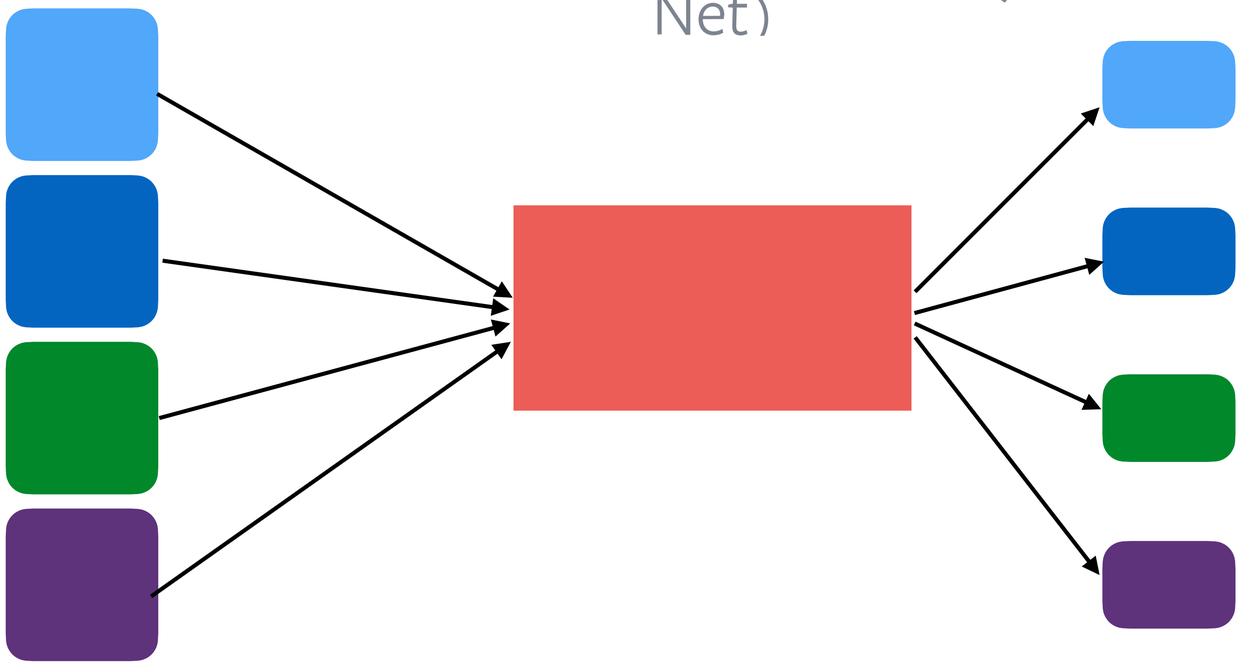
mAP = mean Average Precision
(Higher is better)

Contrastive Learning

Related and Unrelated Images

Shared network (Siamese Net)

Image Features (Embeddings)



Loss Function

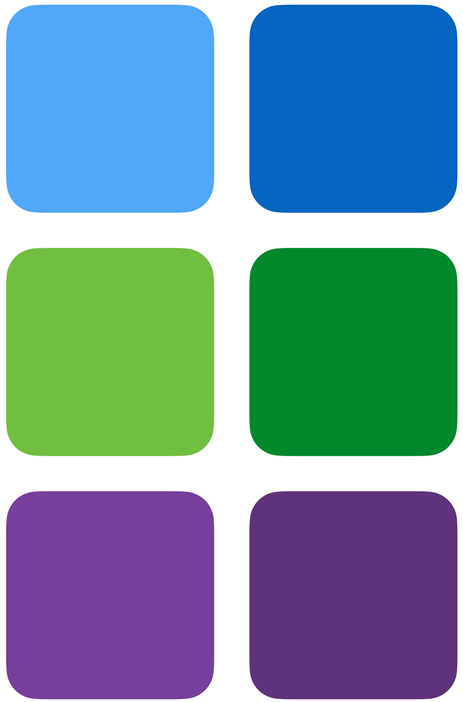
Embeddings from related images should be closer than embeddings from unrelated images

Hadsell et al., 2005, DrLim

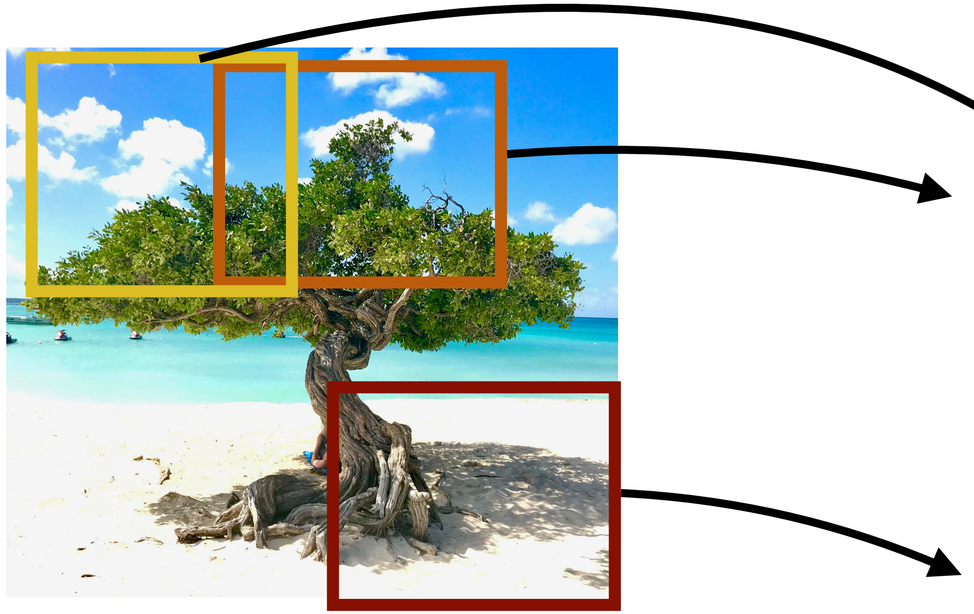
Contrastive Learning

- How to define what images are "related" and "unrelated"?

Related and Unrelated Images



Nearby patches vs. distant patches of an Image



Related
(Positives)

Unrelated
(Negative)

van der Oord et al., 2018,
Henaff et al., 2019
Contrastive Predictive Coding

Patches of an image vs. patches of other images

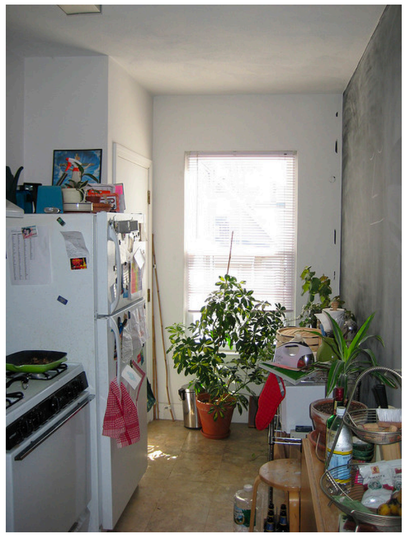
Related
(Positives)

Unrelated
(Negative)

Wu et al., 2018, Instance Discrimination
He et al., 2019, MoCo
Misra & van der Maaten, 2019, PIRL
Chen et al., 2020, SimCLR

Data Augmentations of each patch

Related
(Positives)

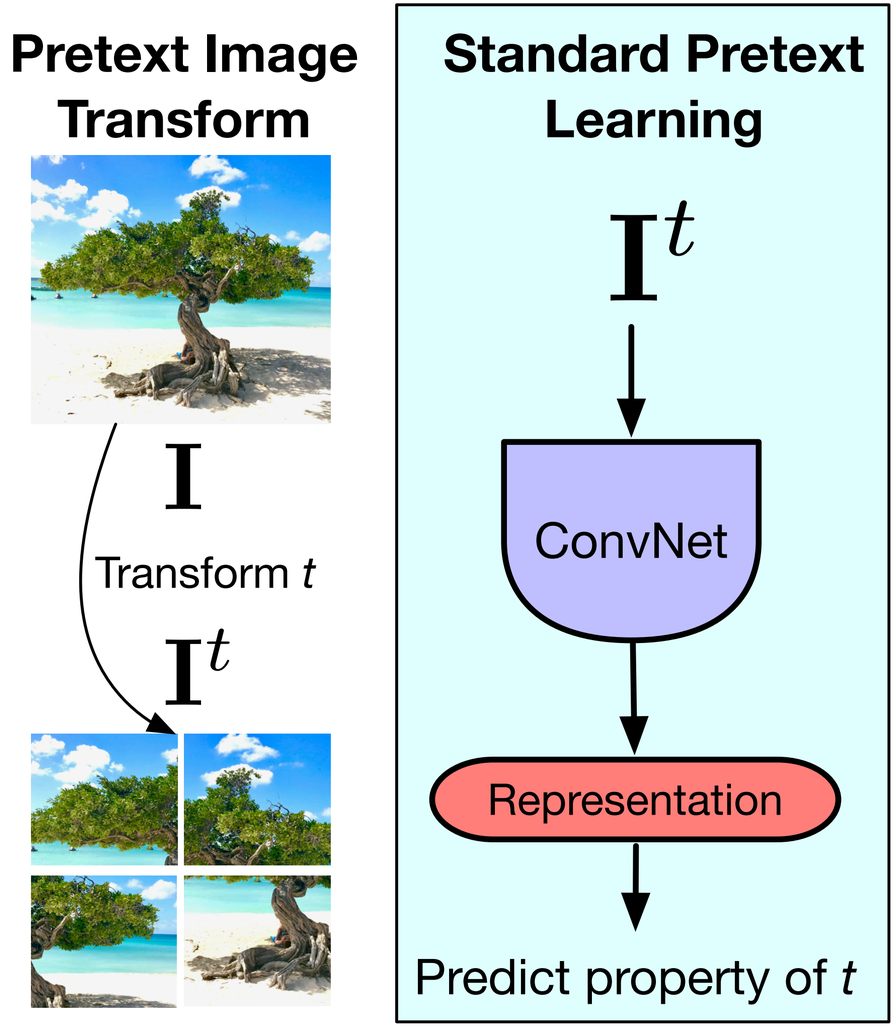


Unrelated
(Negative)

Underlying Principle for Pretext Tasks

- Apply known image transform **t**
- Construct task to predict **t** from transformed Image ($\mathbf{I}^{\mathbf{t}}$)

- Final layer representations must carry information about **t**
- Representations "covary" with **t**

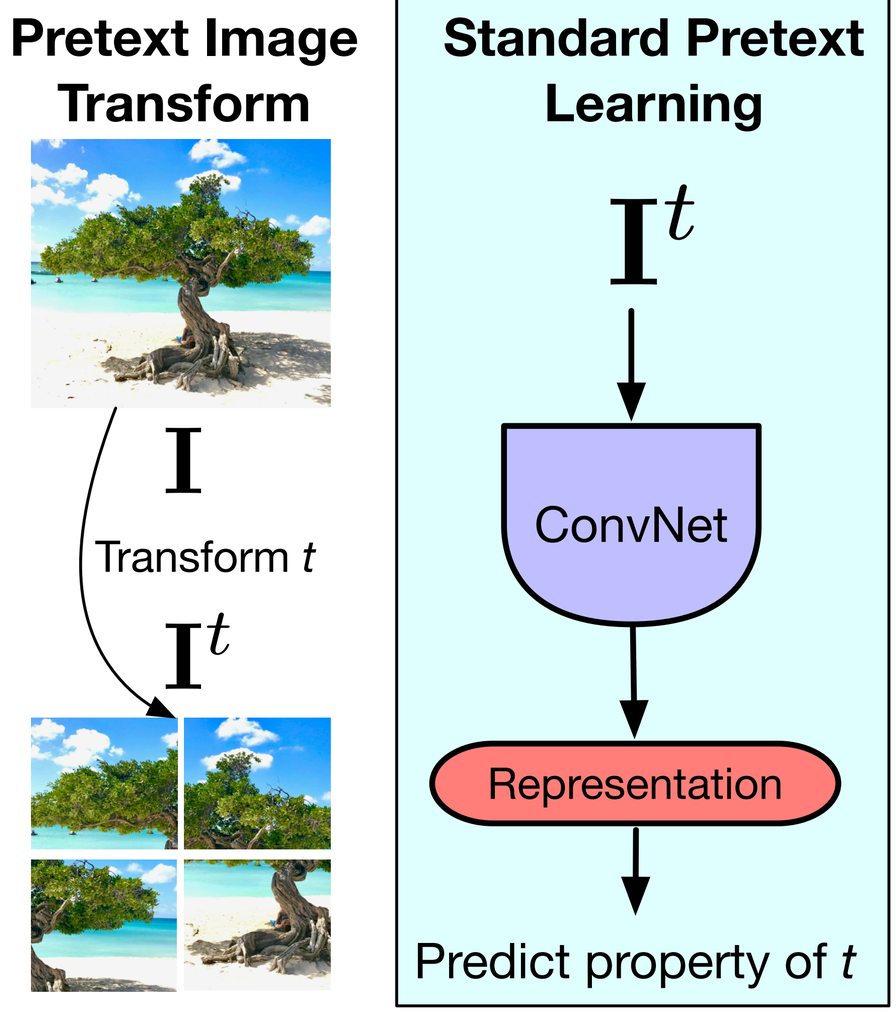


Underlying Principle for Pretext Tasks

- Apply known image transform **t**
- Construct task to predict **t** from transformed Image ($\mathbf{I}^t$)

- Final layer representations must carry information about **t**
- Representations "covary" with **t**

But .... shouldn't representations be invariant to low-level image transforms?



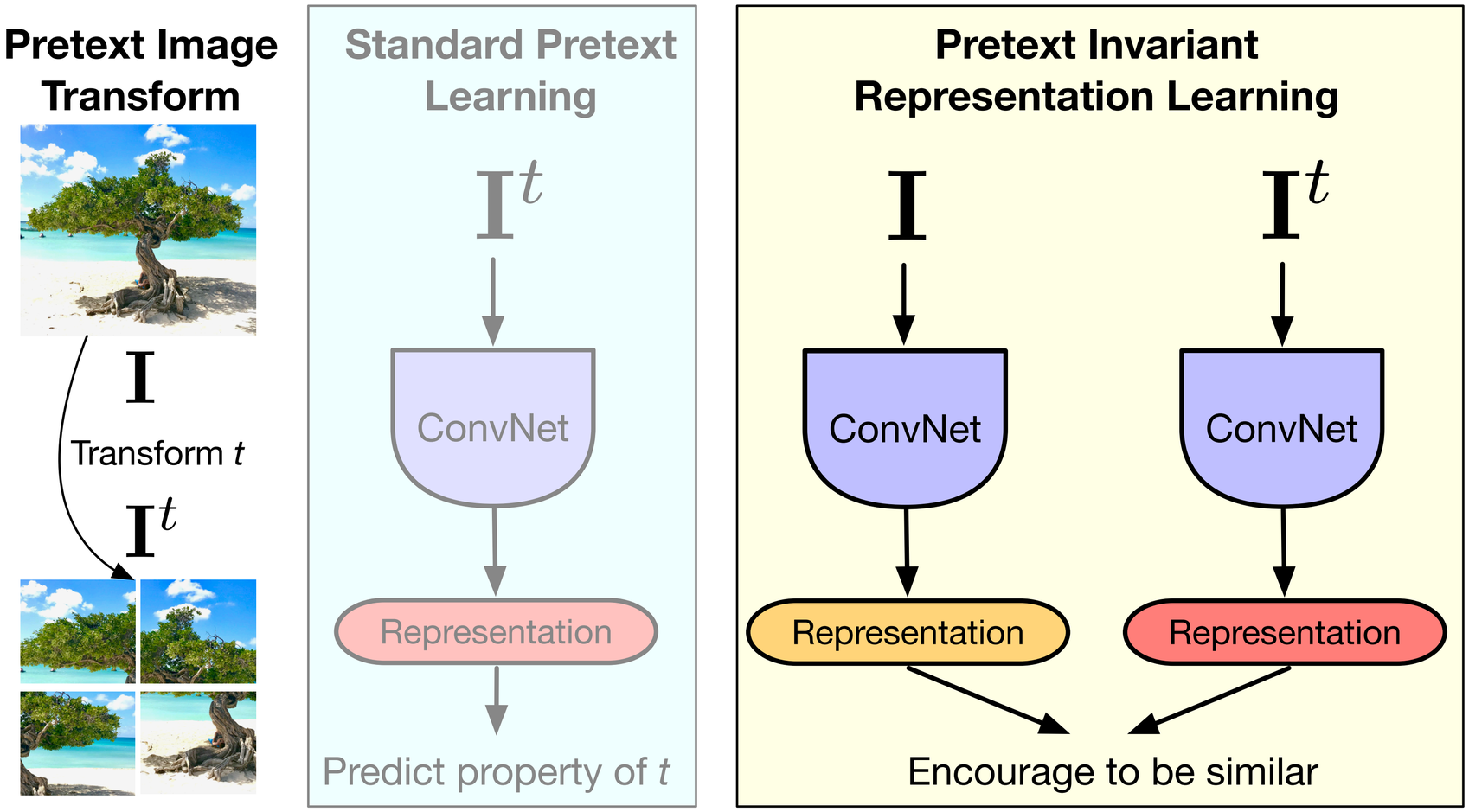
There are many exceptions: e.g., in 3D (Kanazawa et al., 2016; Rocco et al., 2017); Novotny et al., 2018)

How important has invariance been?

- Hand-crafted features like SIFT and HOG
- SIFT - Scale **Invariant** Feature Transform
- Supervised systems are trained to be invariant to "data augmentation"

Pretext-Invariant Representation Learning (PIRL)

- Be invariant to **t**
- Representation contains no information about **t**

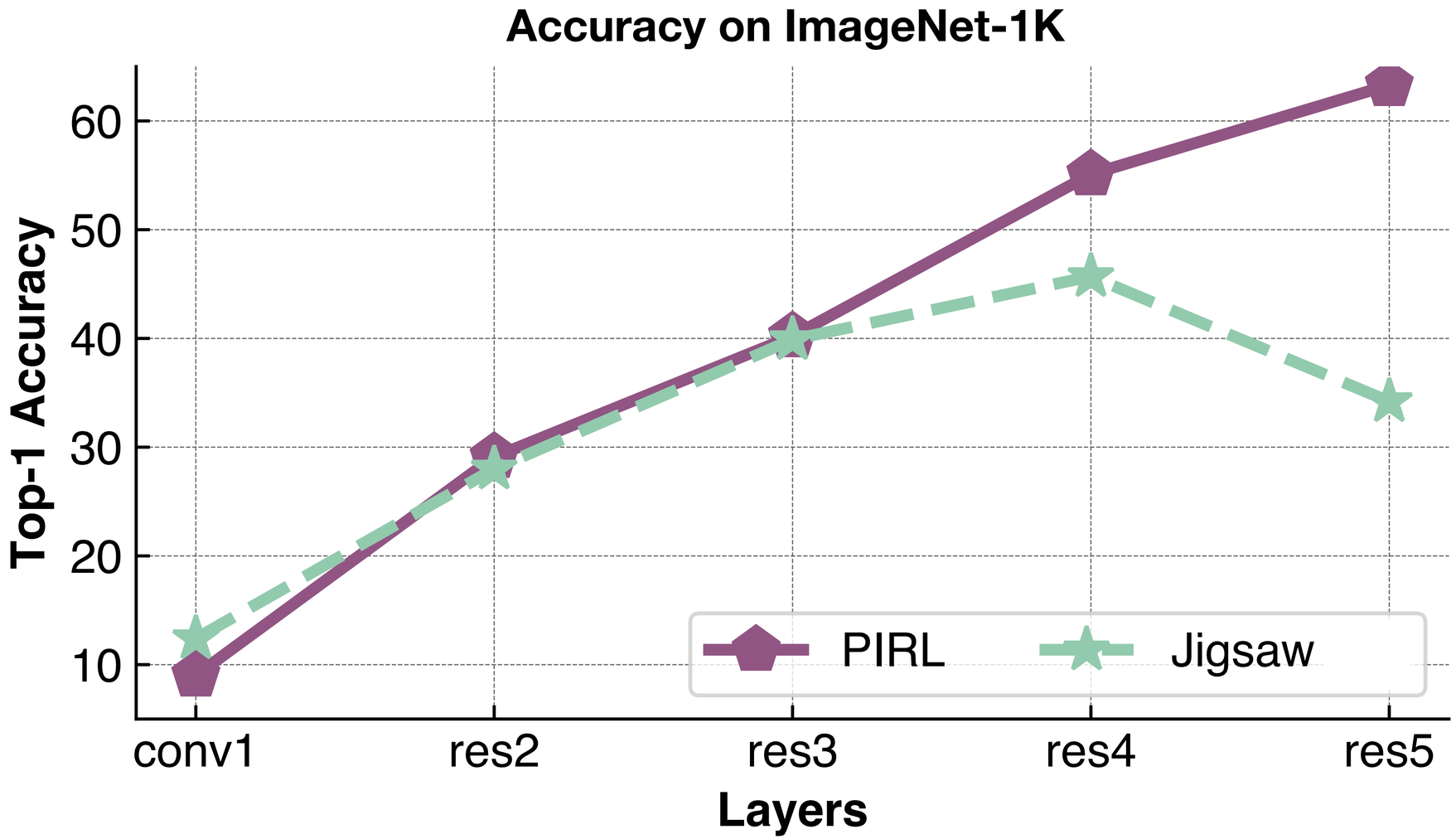


Semi-supervised Learning

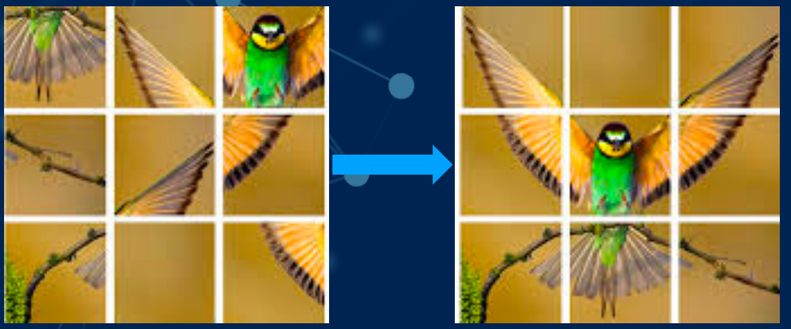
- Fine-tune on fraction of labeled data from **ImageNet-1K**

Method	Top-5 Accuracy	
Fraction of Data ⟶	1%	10%
Jigsaw (Goyal et al., 2019)	45.3	79.3
VAT + Ent Min (Grandvalet et al., Miyato et al.)	47.0	83.4
S4L Rotation (Zhai et al., 2019)	53.4	**83.8**
PIRL	**57.2**	**83.8**

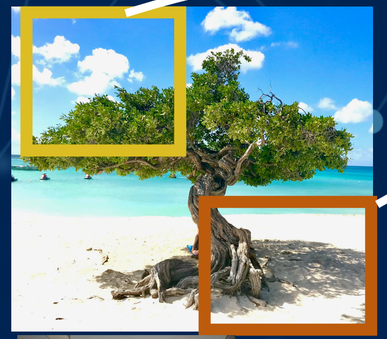
Semantic Features?



Pretext tasks

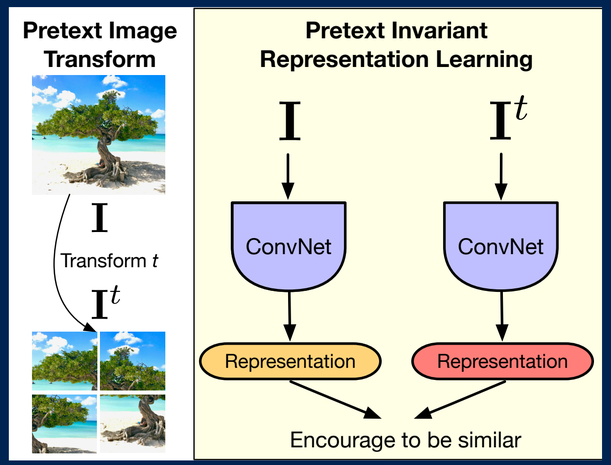


Contrastive

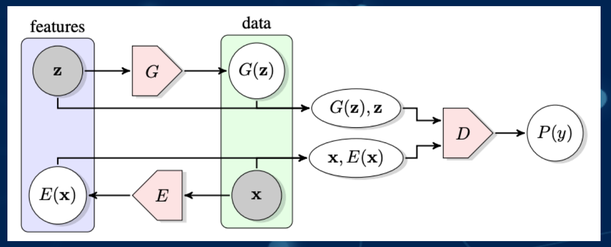


Related

Unrelated



Generative



AutoEncoder,
VAE, GAN,
BiGAN

Predict more information