Self-supervised learning in computer vision

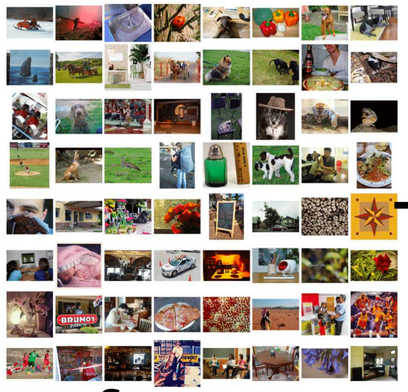
Ishan Misra

Facebook AI Research

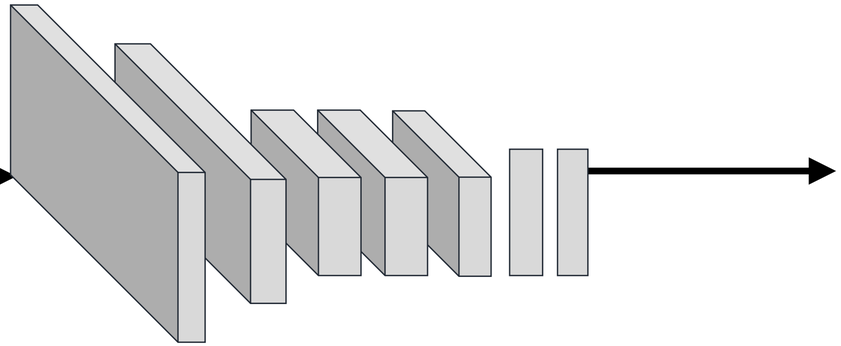
With slides from Andrew Zisserman, Carl Doersch

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

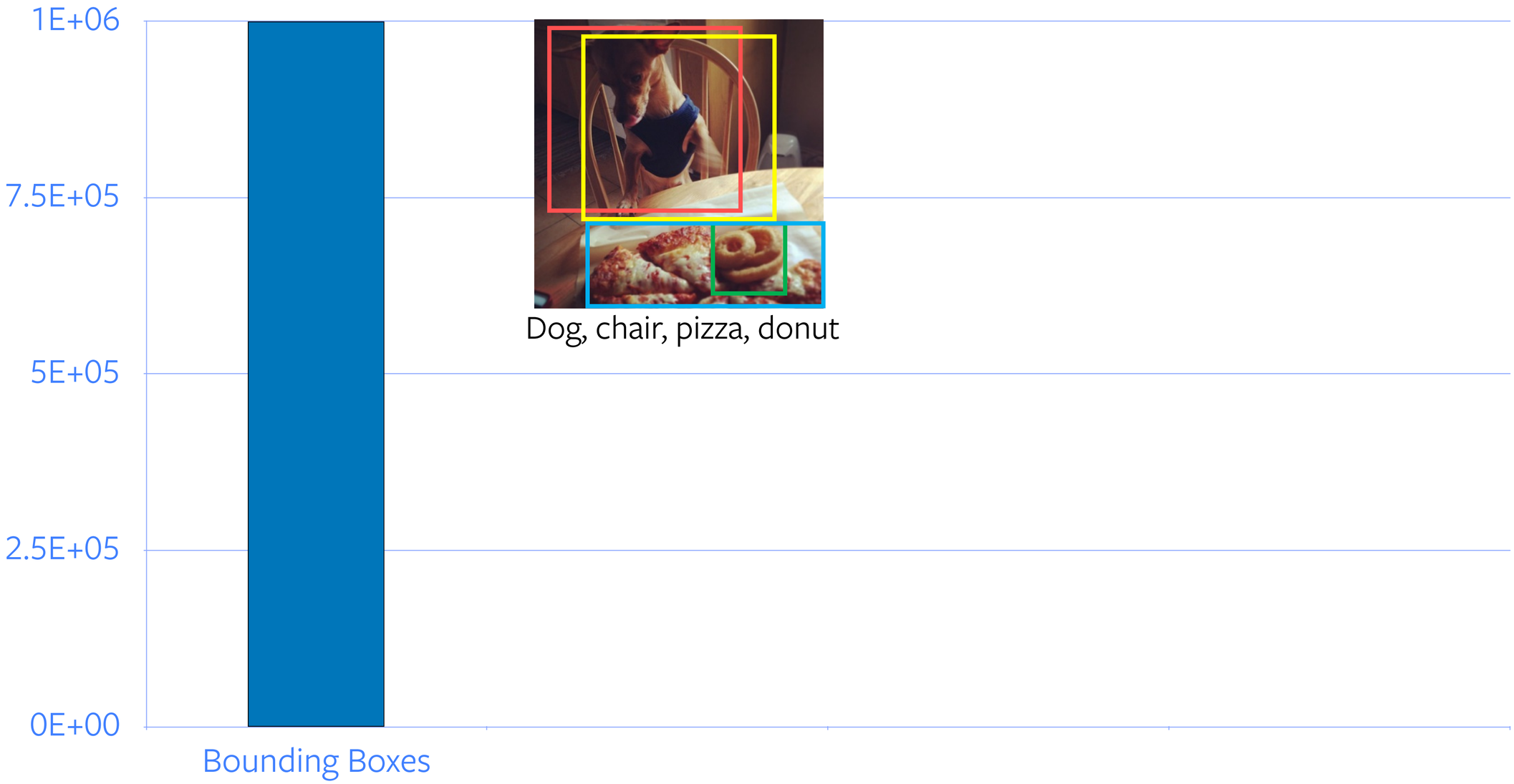
The promise of "alternative" supervision

- 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?

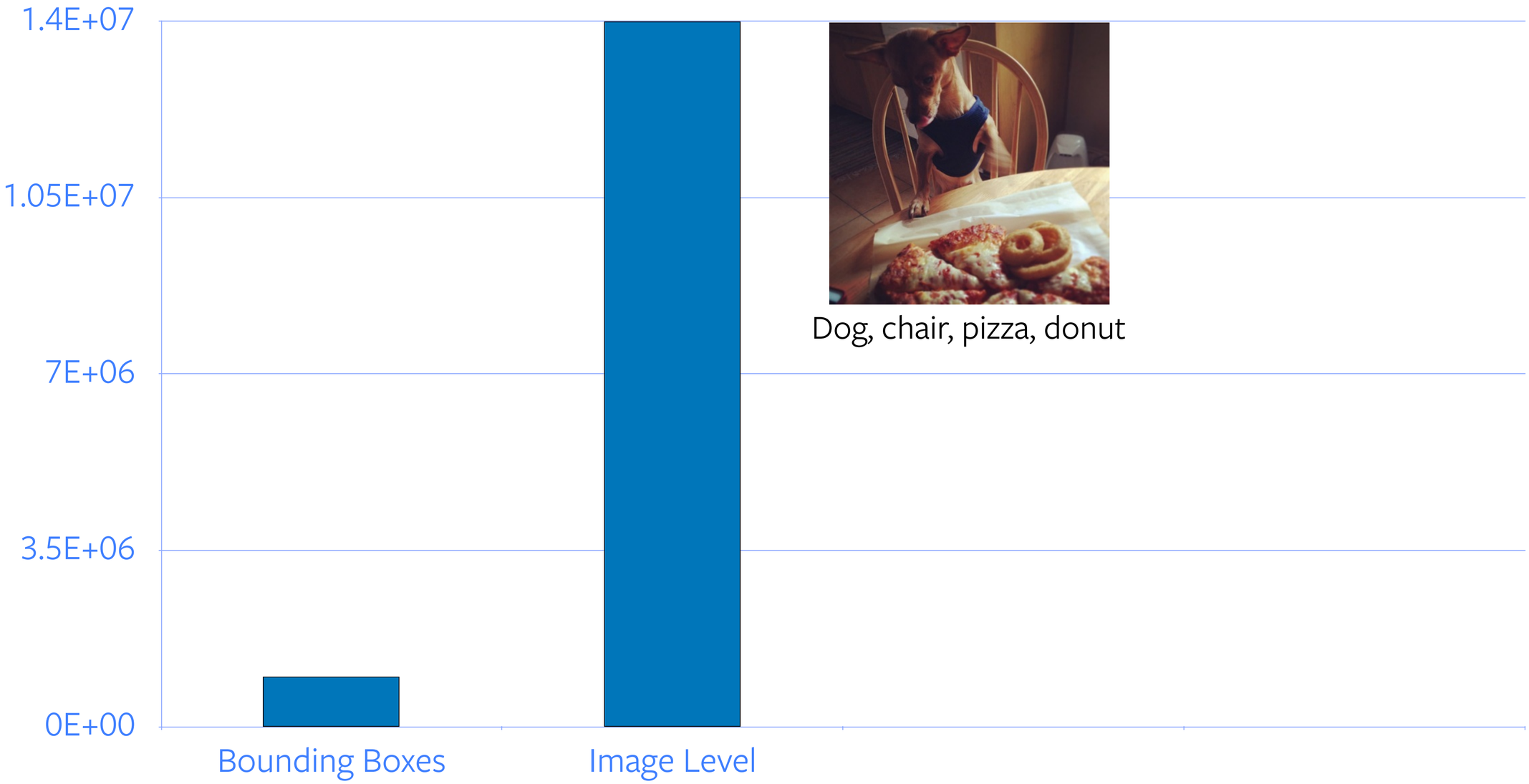
Can we get labels for all data?



Dog, chair, pizza, donut

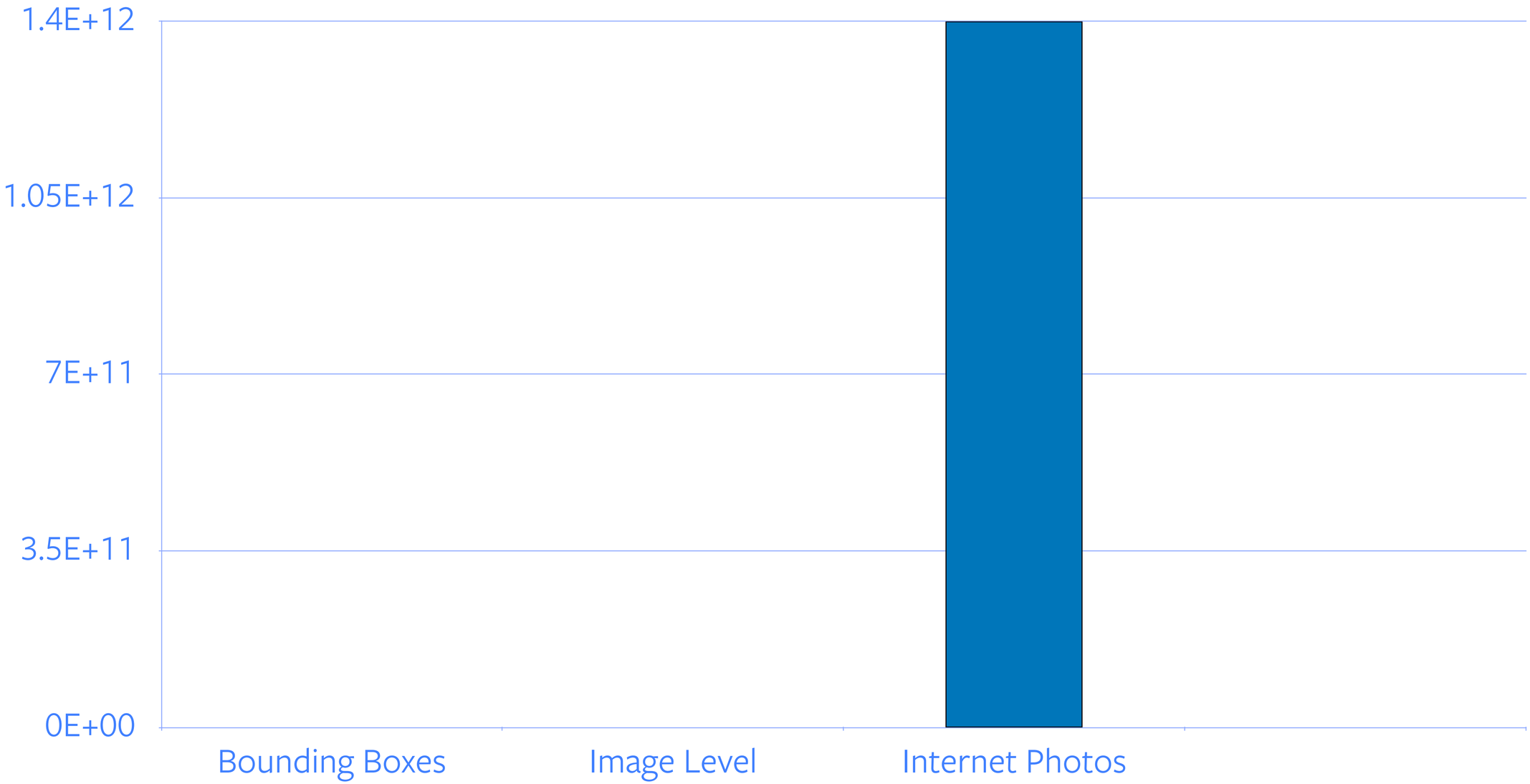
Stats from Pawan Kumar at Oxford

Can we get labels for all data?



Dog, chair, pizza, donut

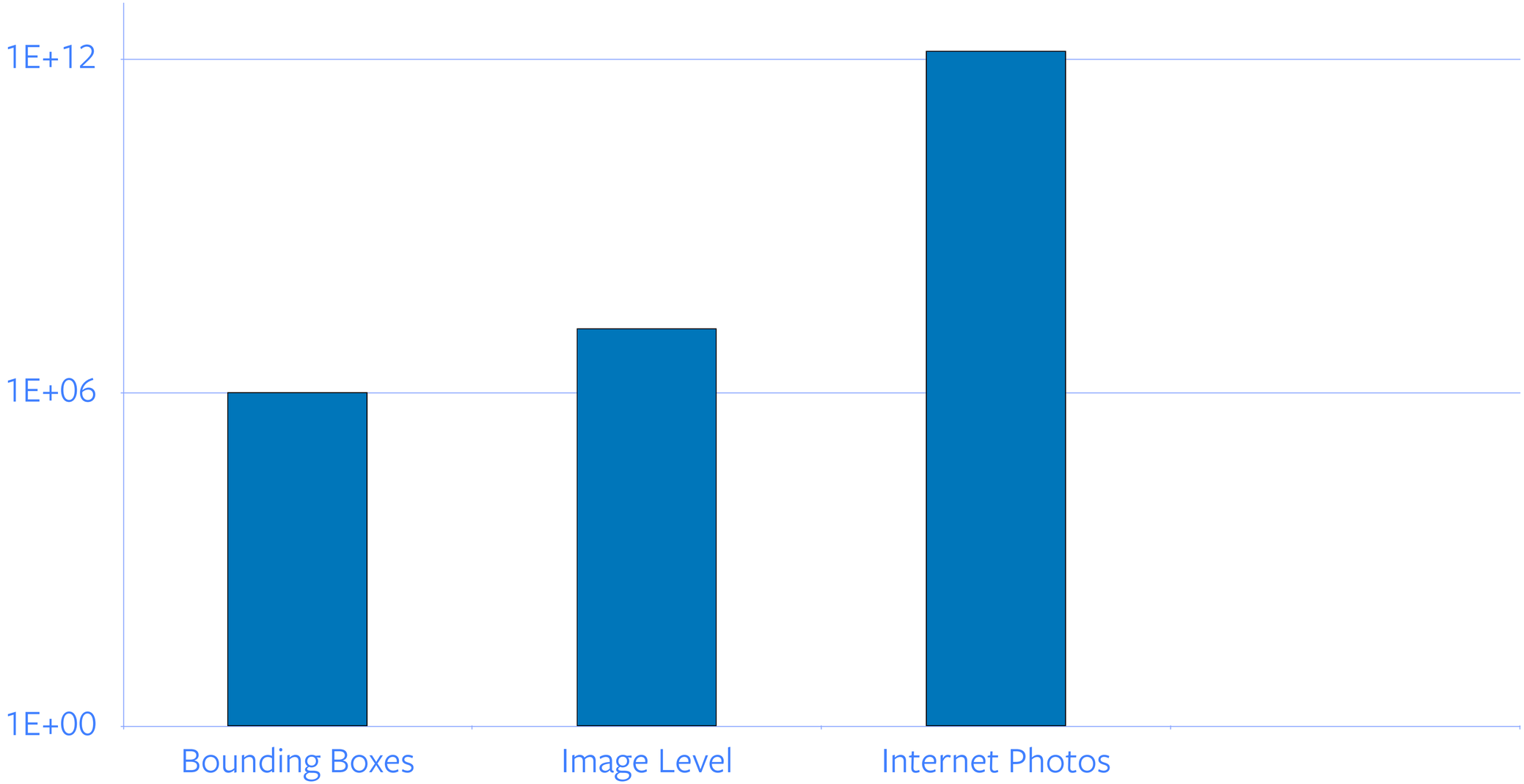
Can we get labels for all data?



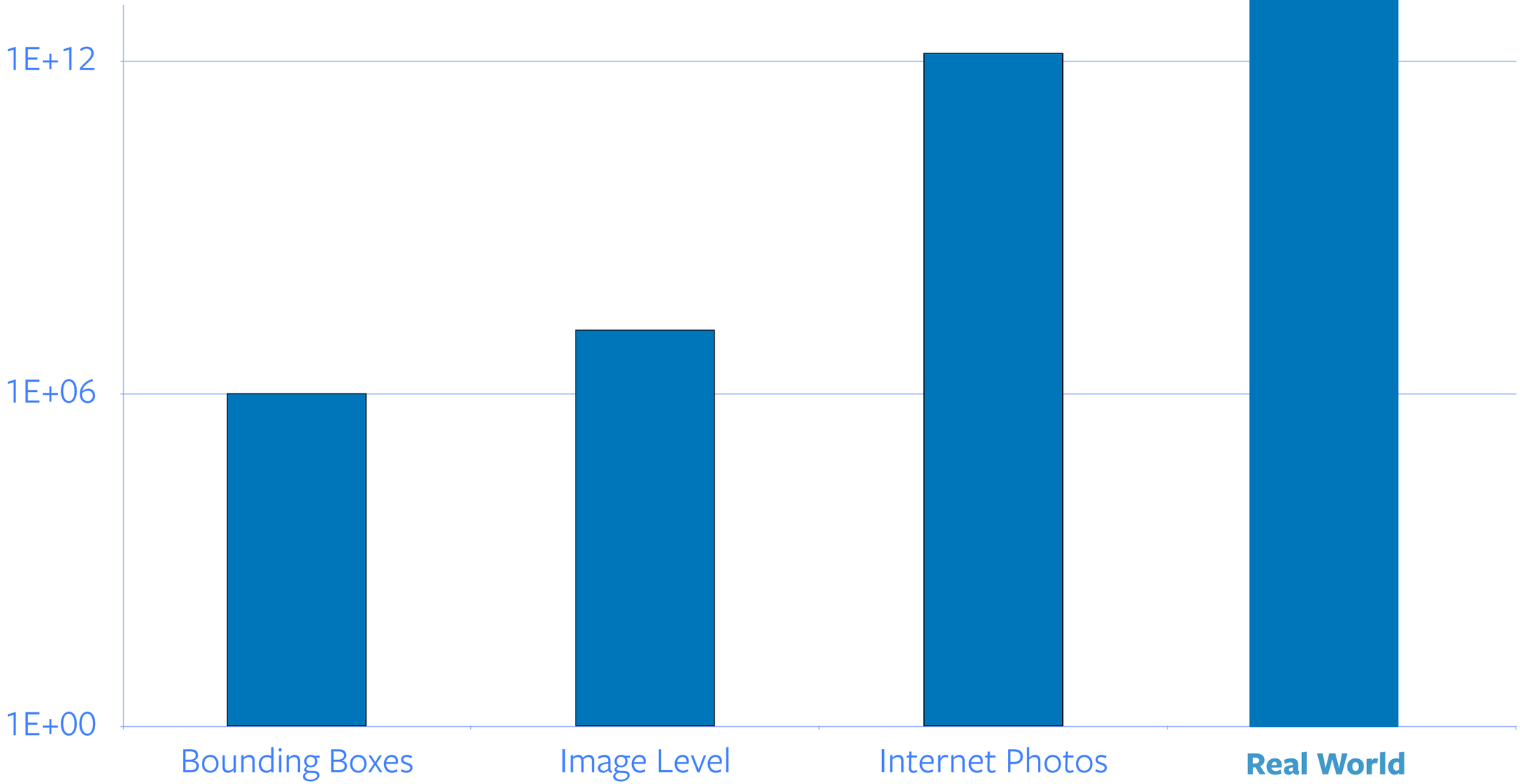
forbes.com

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?

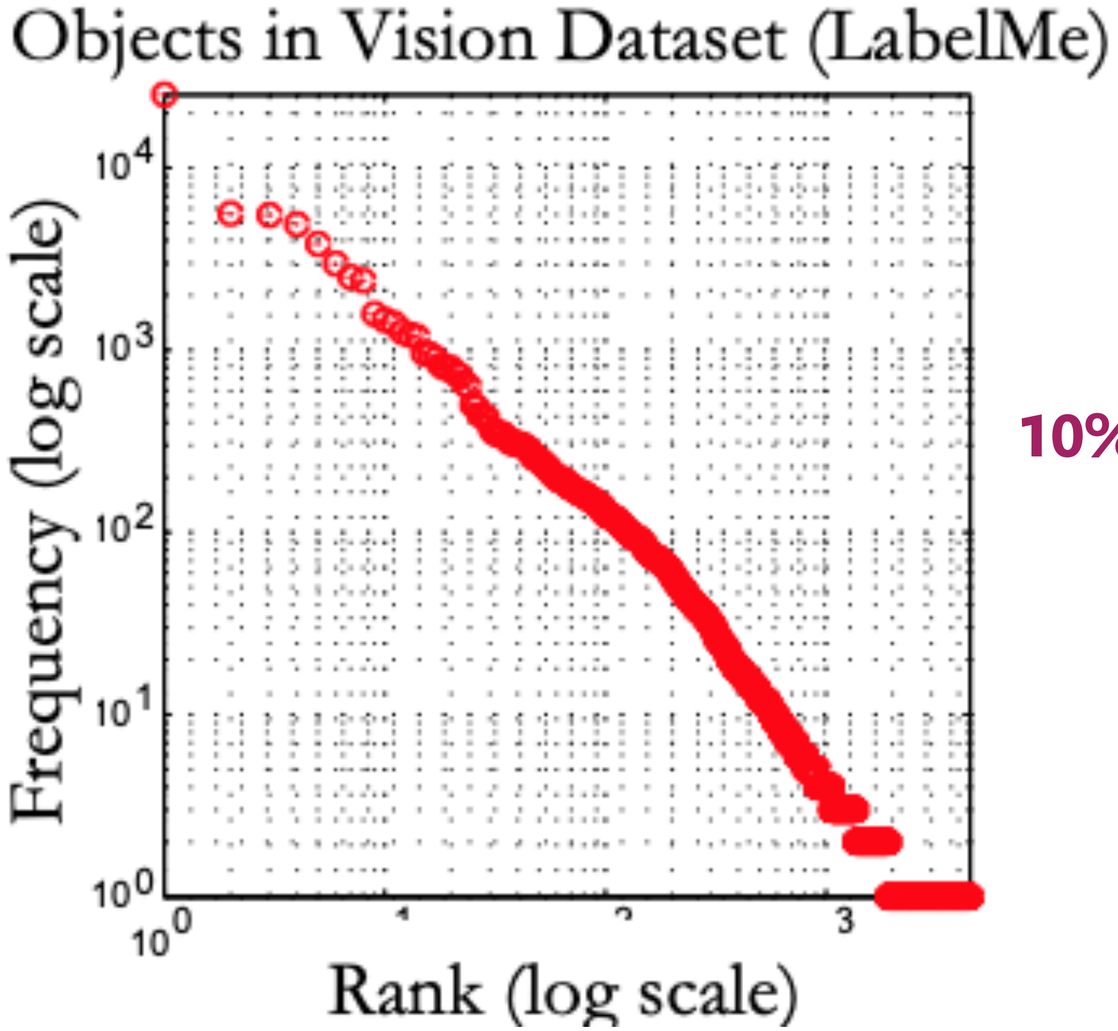


ImageNet (14 million images) needed 22 human years to label

Can we get labels for all data?

- What about complex concepts?
  - Video?
- Labelling cannot scale to the size of the data we generate

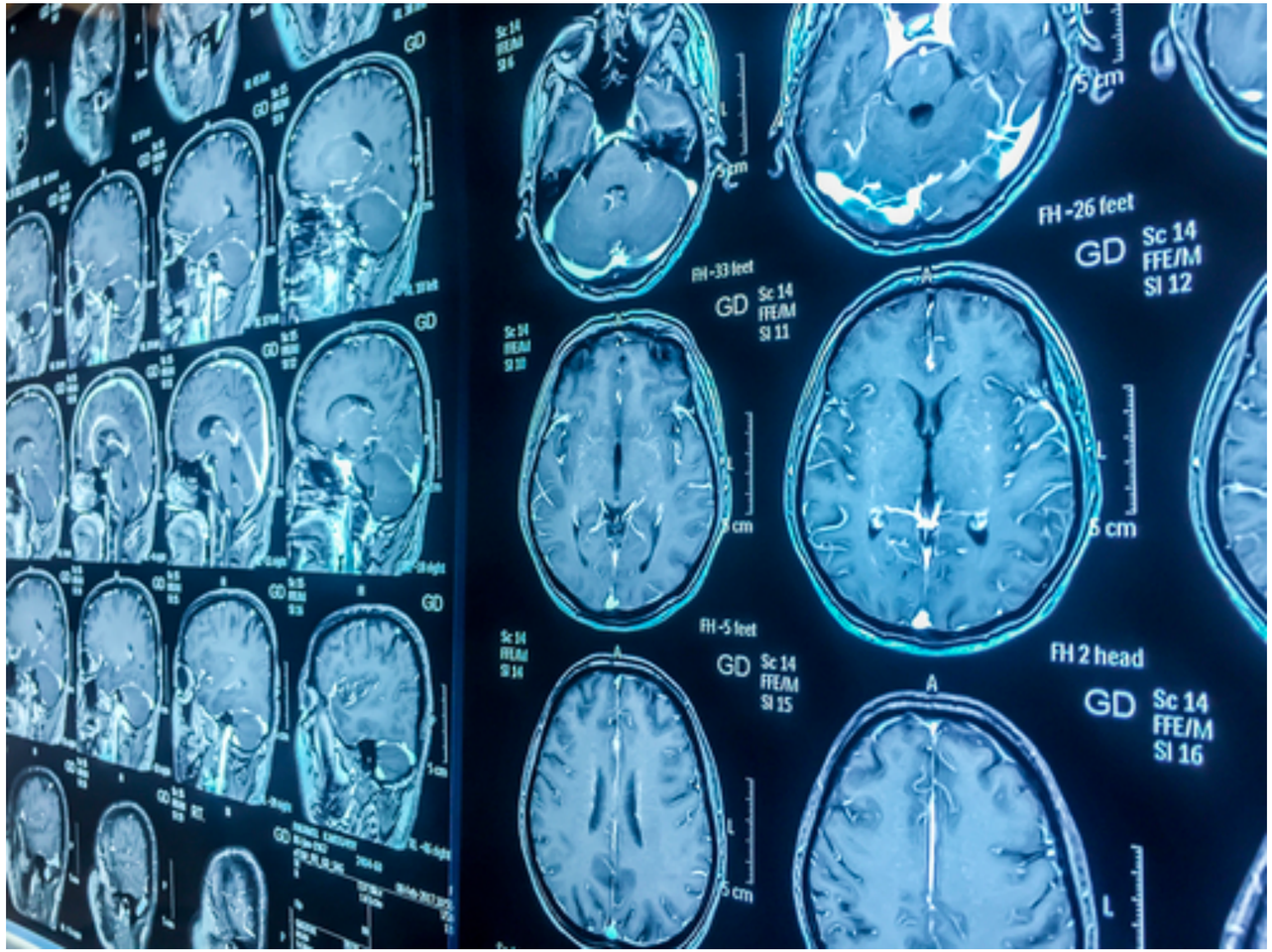
Rare concepts?



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

Slide credit: Rob Fergus

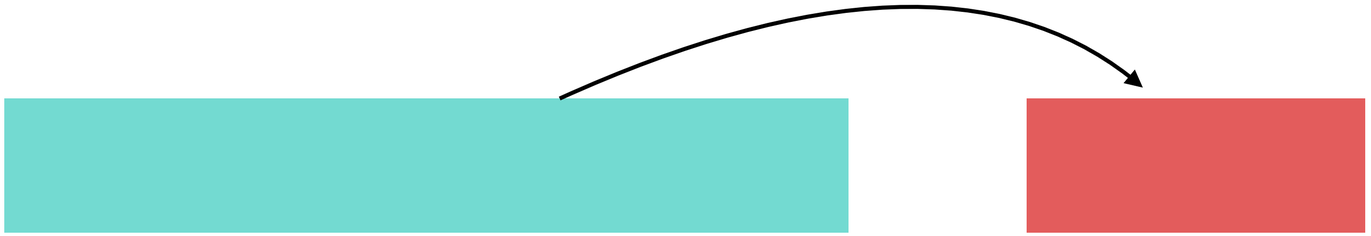
Different Domains?



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

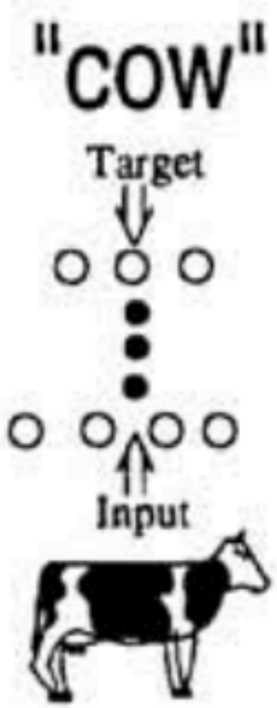
Hidden data

Hidden property of the data

What is "self" supervision?

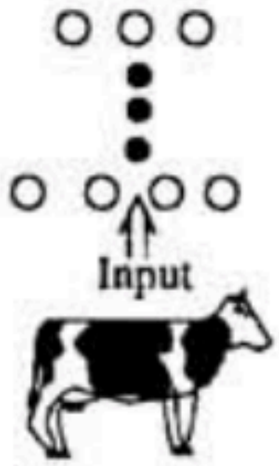
Supervised

- implausible label



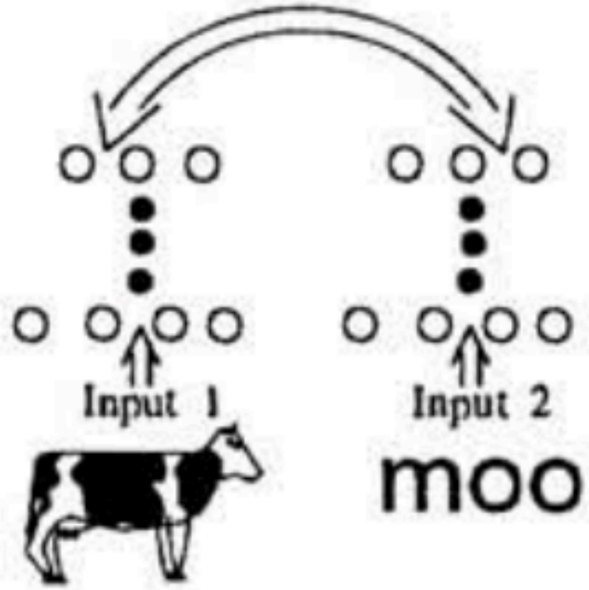
Unsupervised

- limited power



Self-Supervised

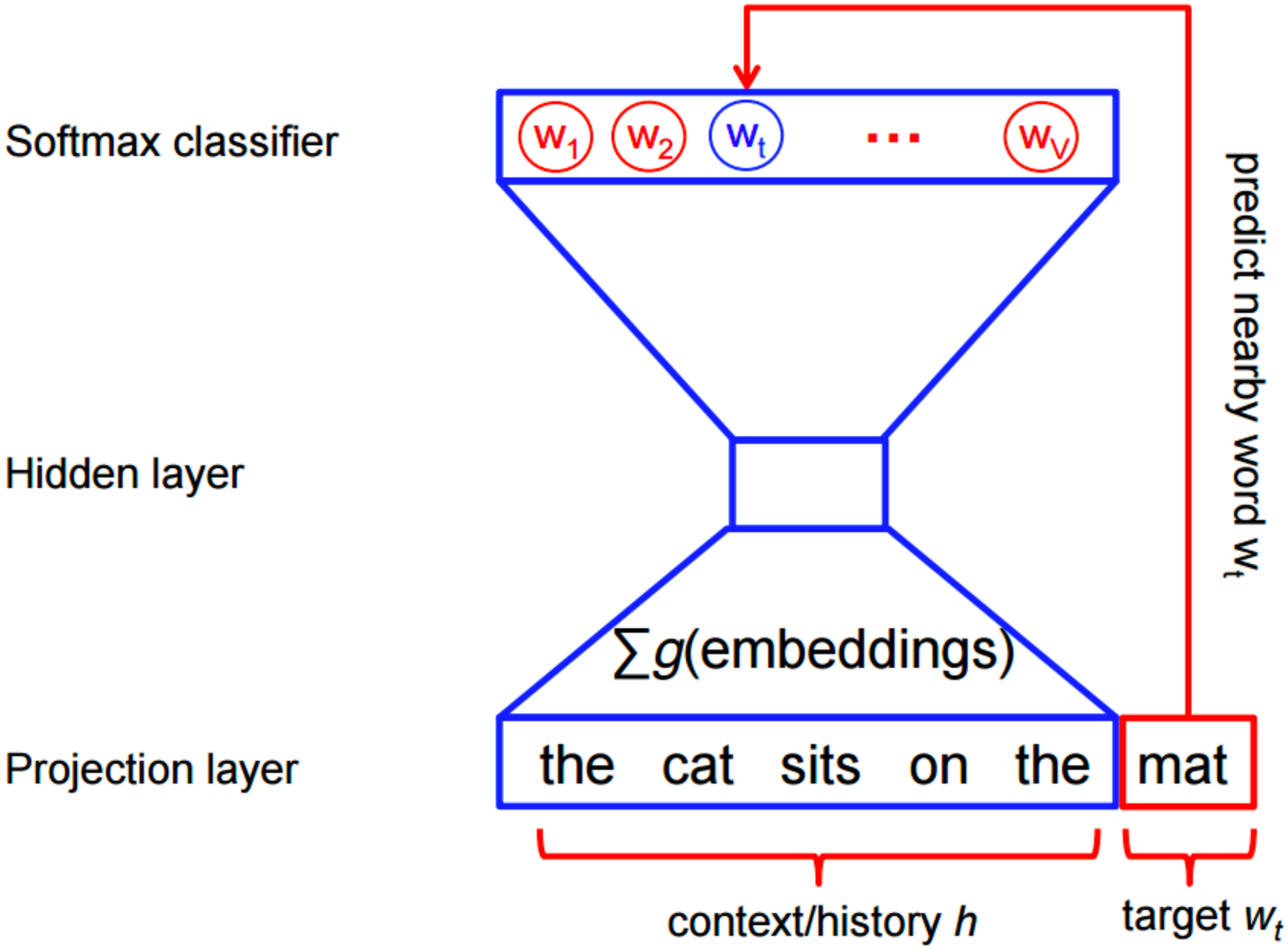
- derives label from a co-occuring input to another modality



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

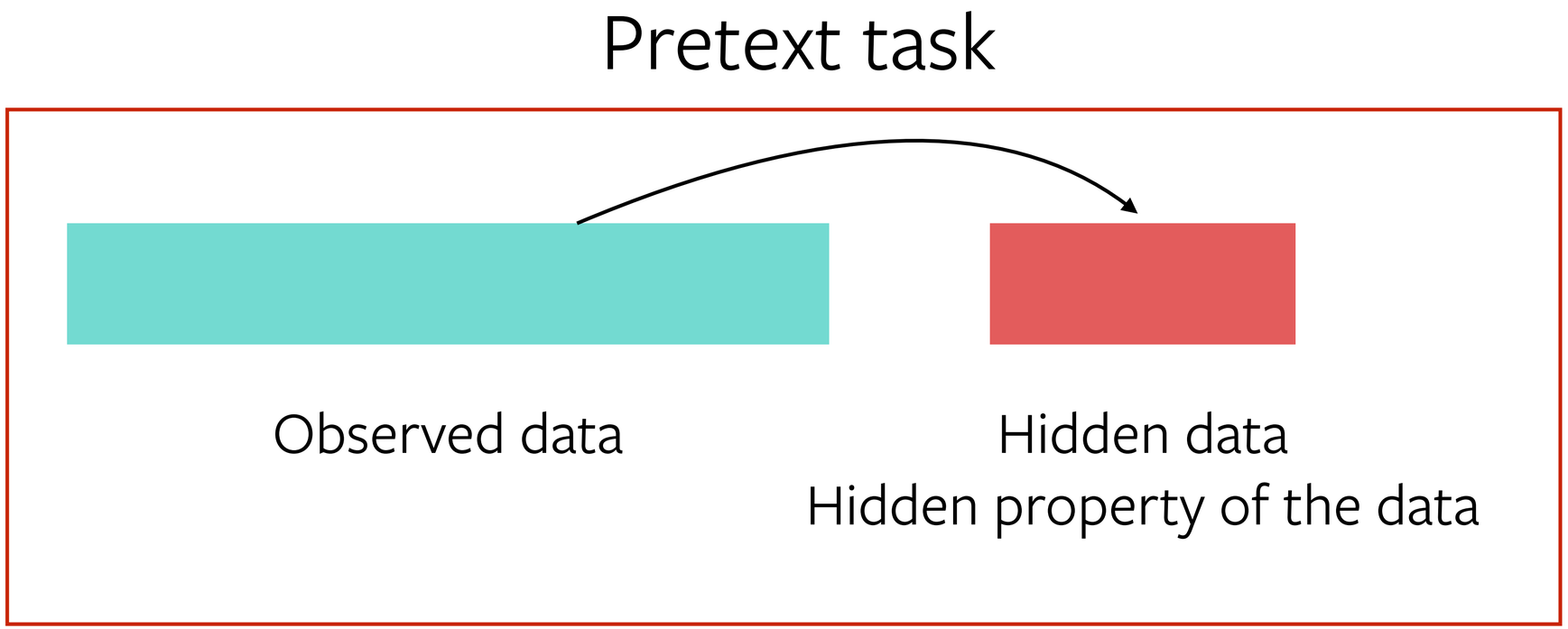
Why self supervision?

- Helps us learn using observations and interactions
- Does not require exhaustive annotation of concepts
- Leverage multiple modalities or structure in the domain

In the context of Computer Vision

Pretext task

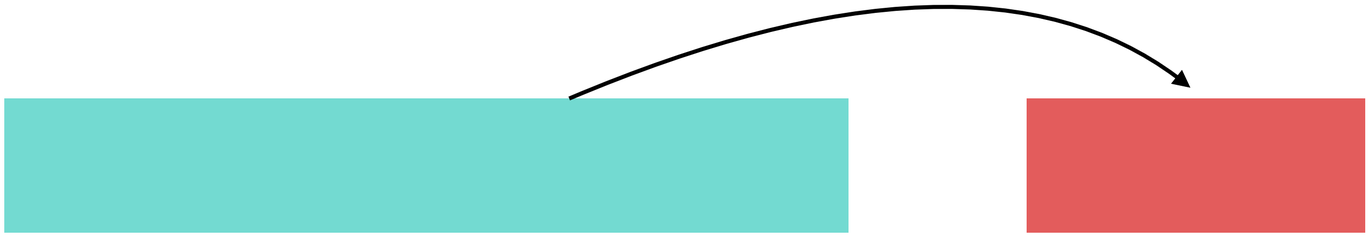
- Self-supervised task used for learning representations
- Often, not the "real" task (like image classification) we care about



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

Pretext task

- Using images
- Using video
- Using video and sound



Observed data

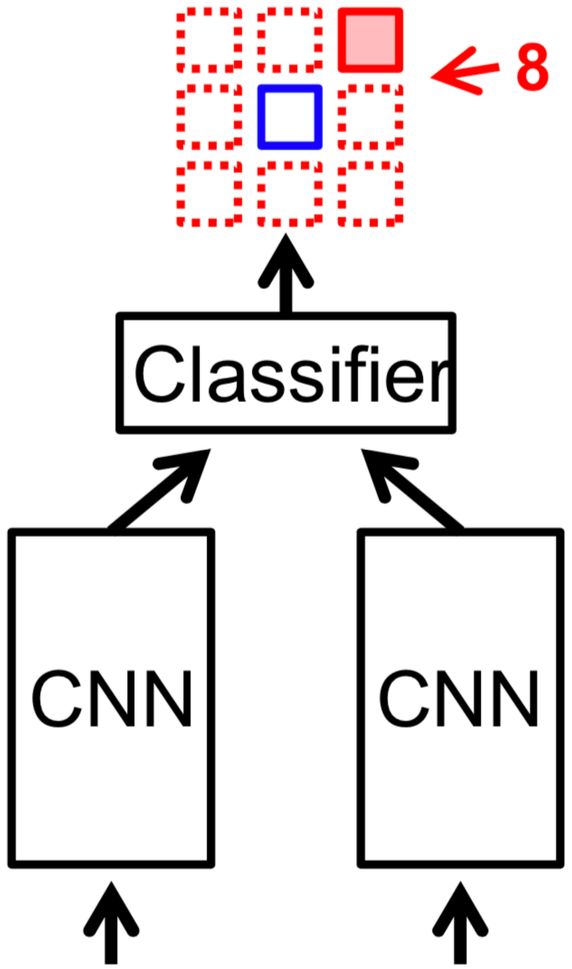
Hidden data

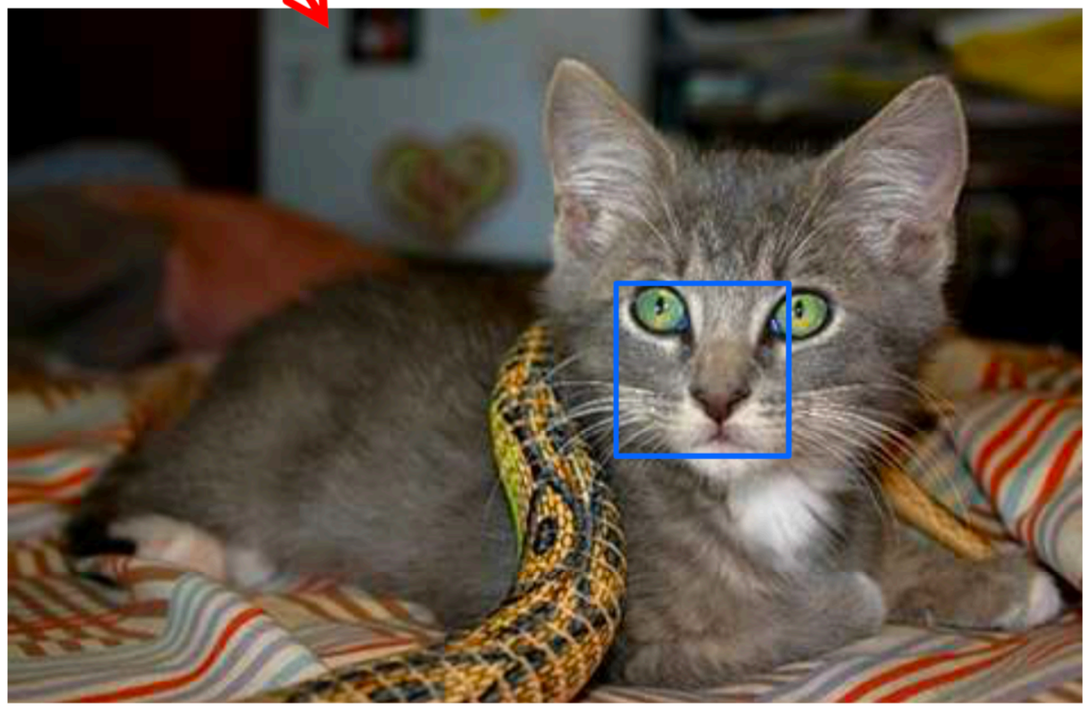
Hidden property of the data

Pretext task

- Using images
- Using video
- Using video and sound

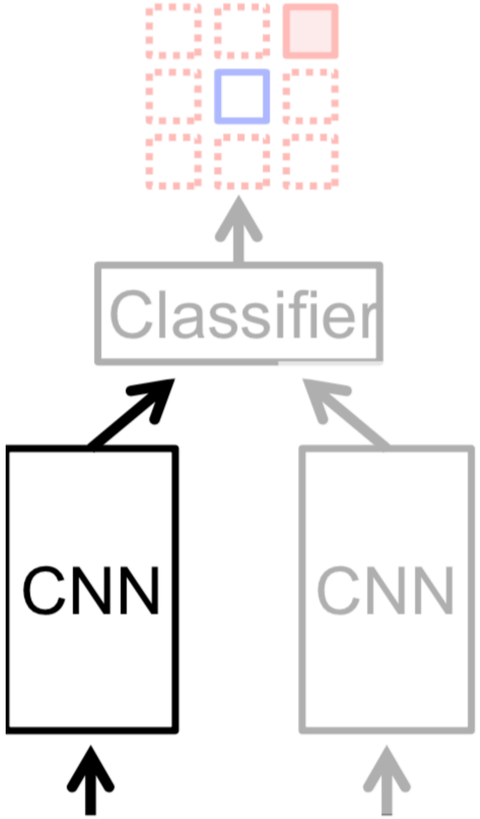
Relative Position of patches

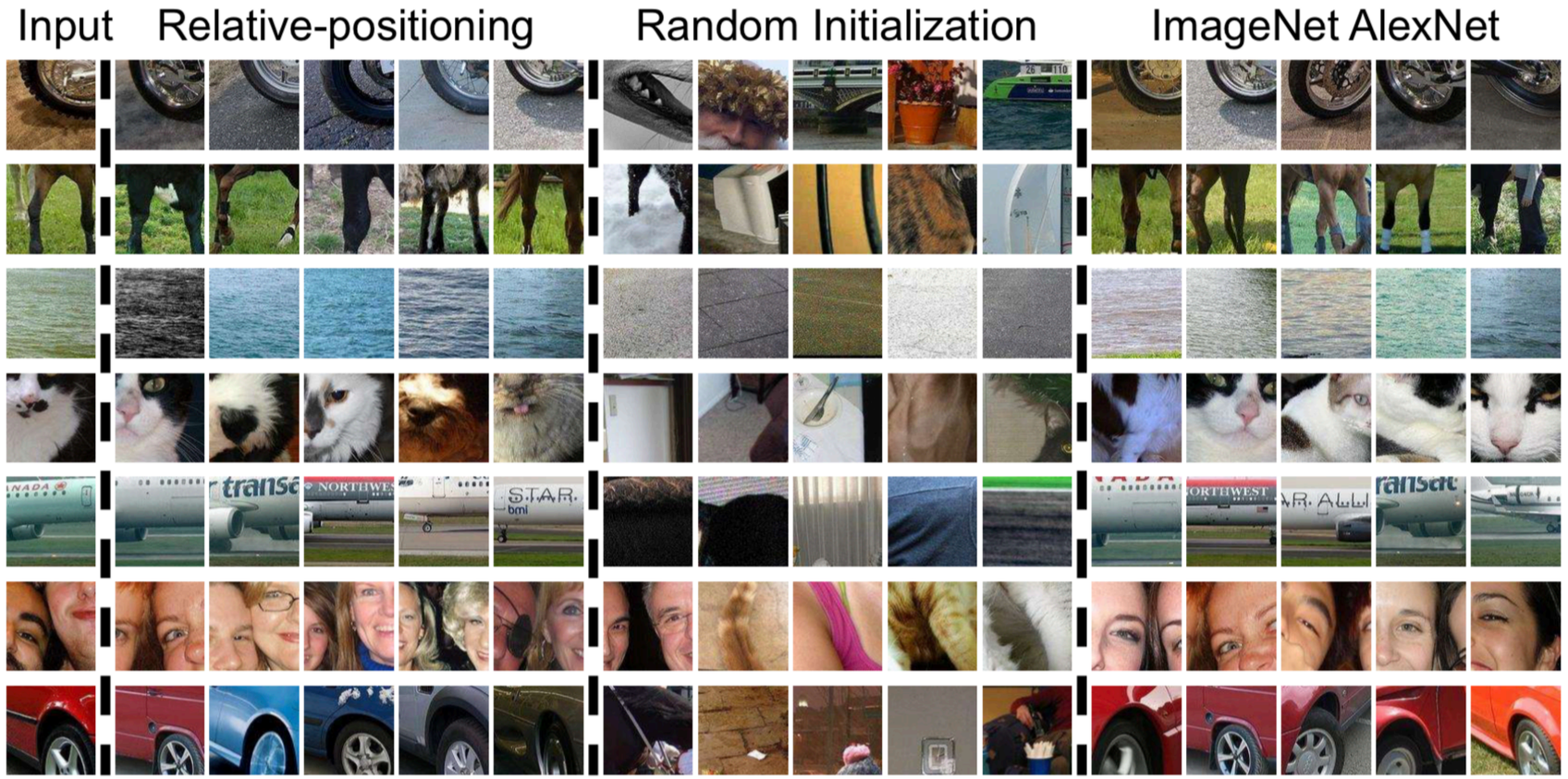




Doersch et al., 2015, Unsupervised visual representation learning by context prediction

Relative Position: Nearest Neighbors in features

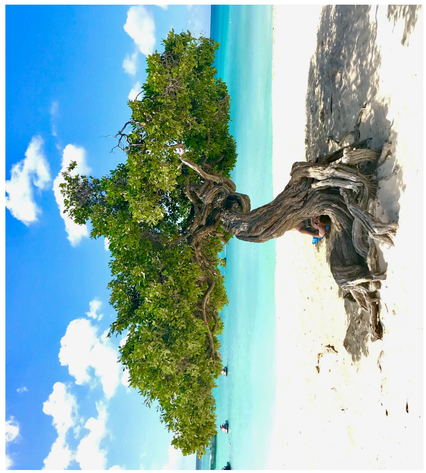




Doersch et al., 2015, Unsupervised visual representation learning by context prediction

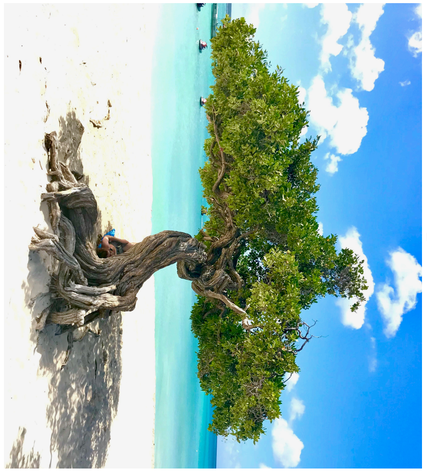
Predicting Rotations

0^0



90^0

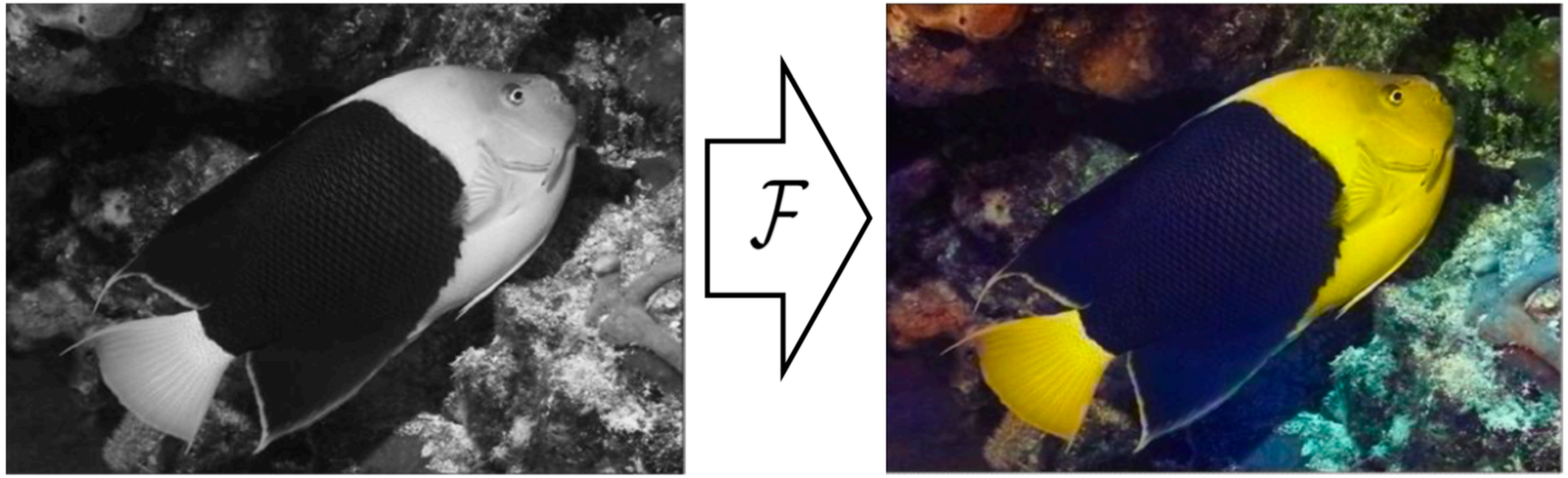
180^0



270^0

Gidaris et al., 2018, Predicting Image Rotations

Colorization

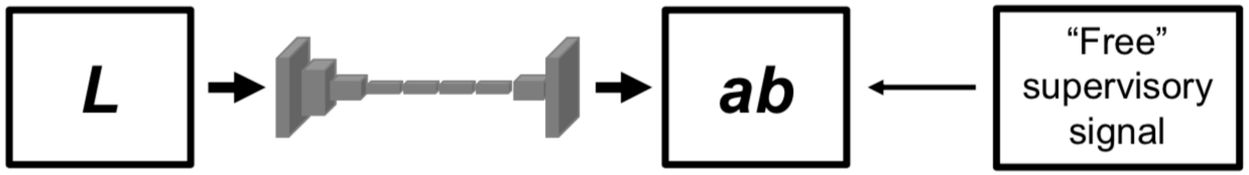


Grayscale image: *L* channel

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

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

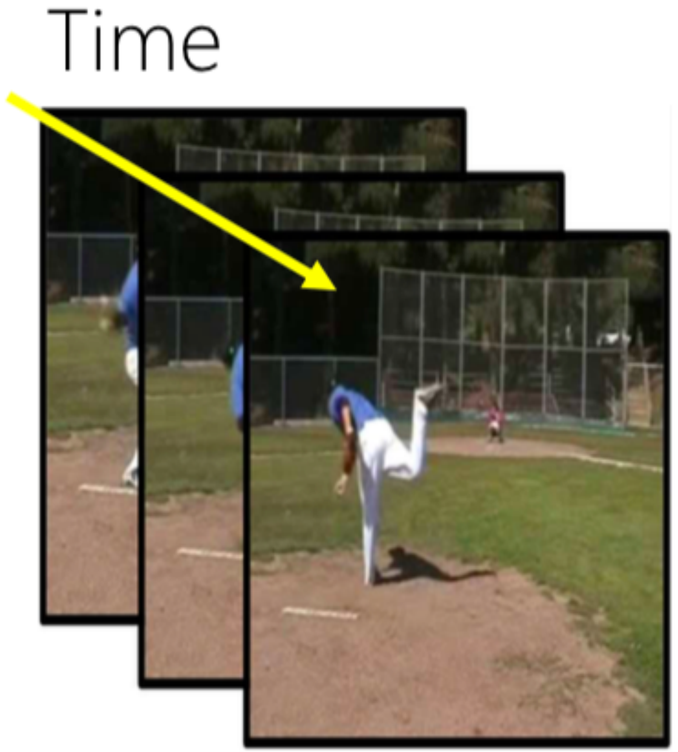
Self-supervision in computer vision

- Using images
- Using video
- Using video and sound

Video

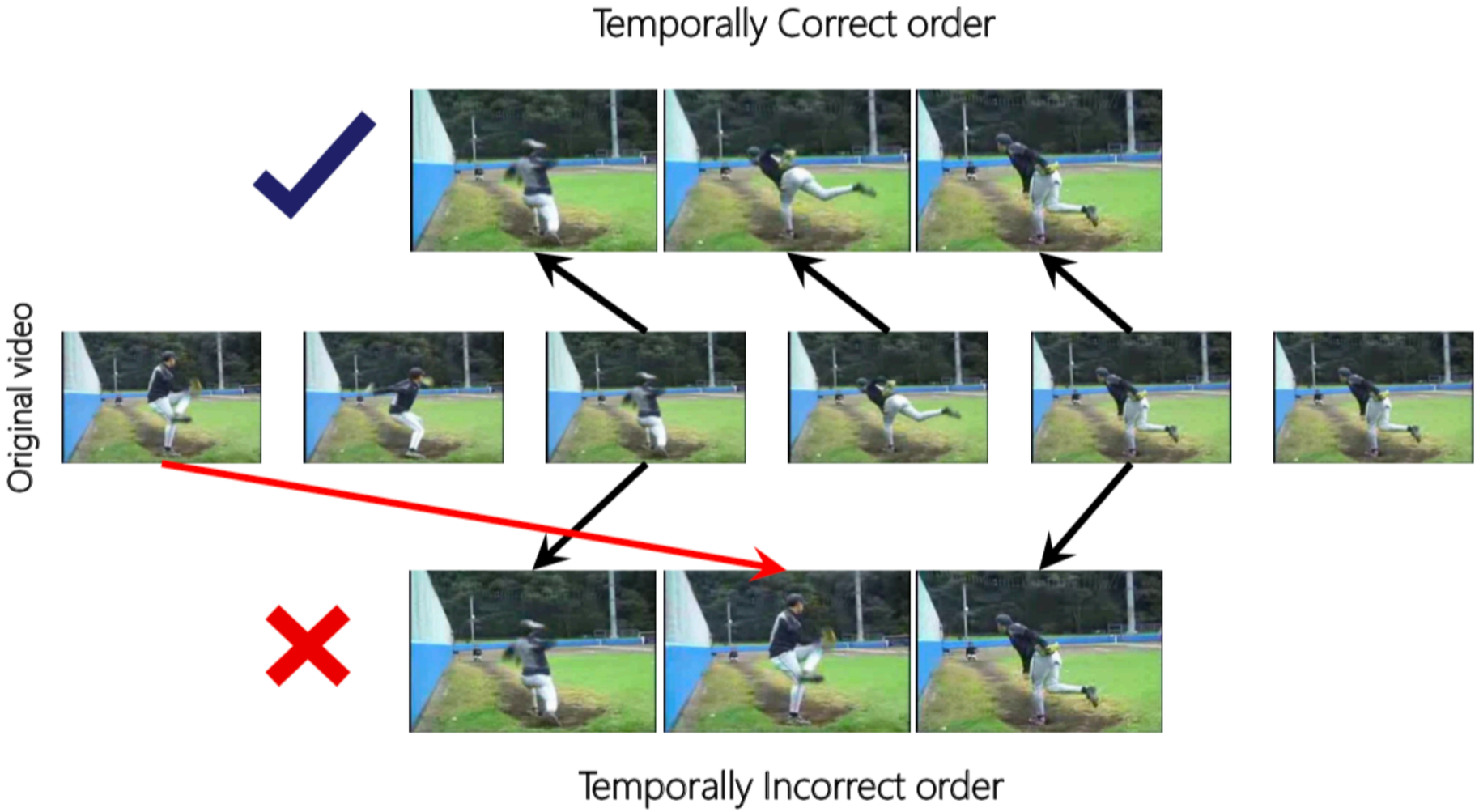
- Video is a "sequence" of frames
- How to get "self-supervision"?

- Predict order of frames
- Fill in the blanks
- Track objects and predict their position



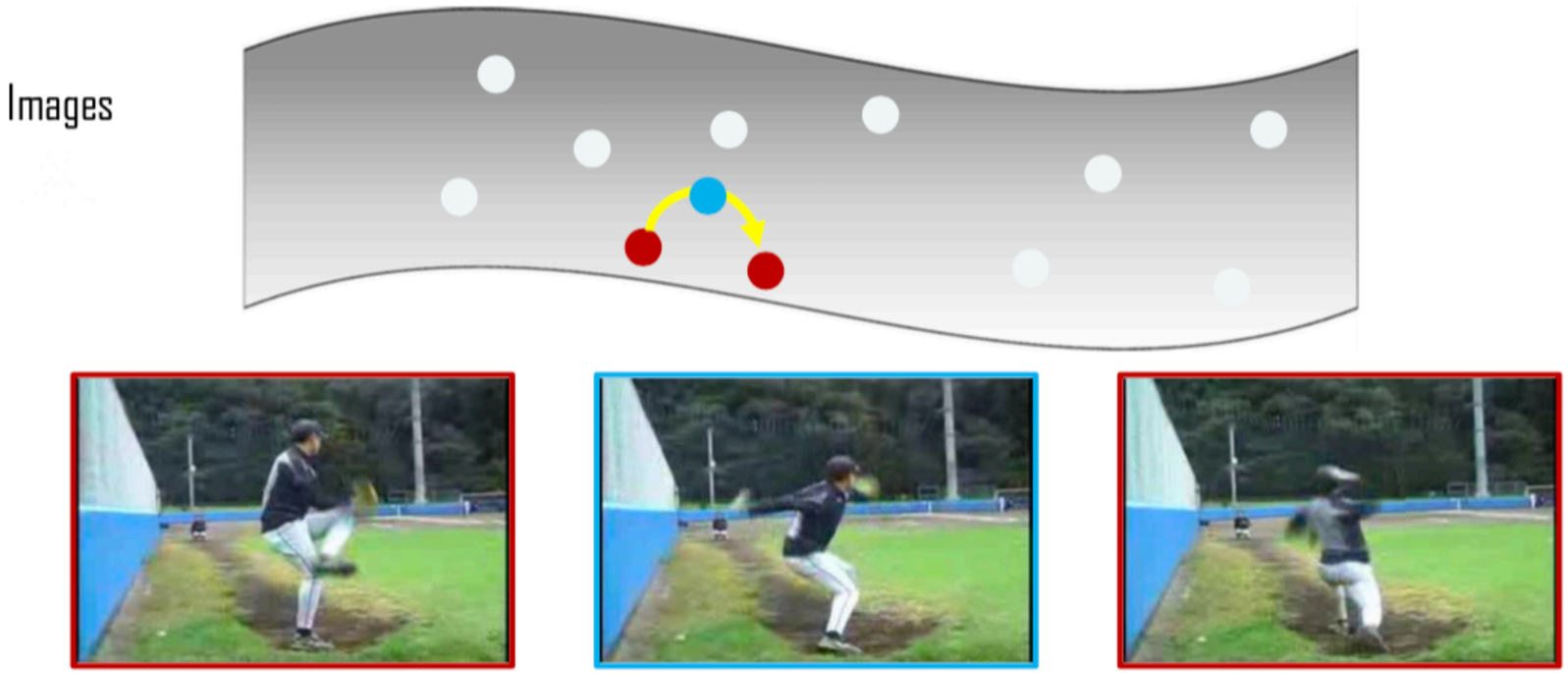
"Sequence" of data

Shuffle & Learn



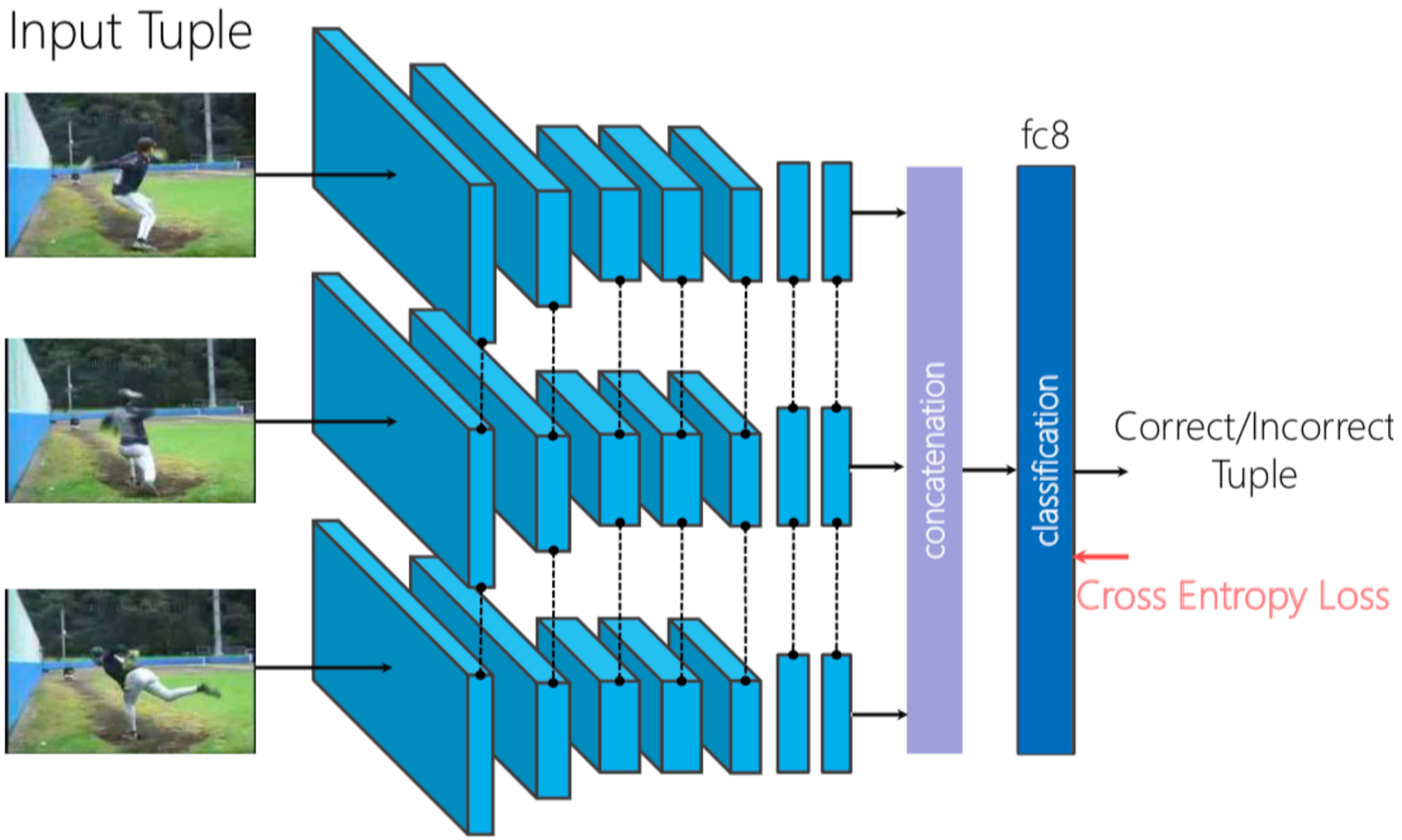
Misra et al., 2016, Shuffle and Learn

Shuffle & Learn

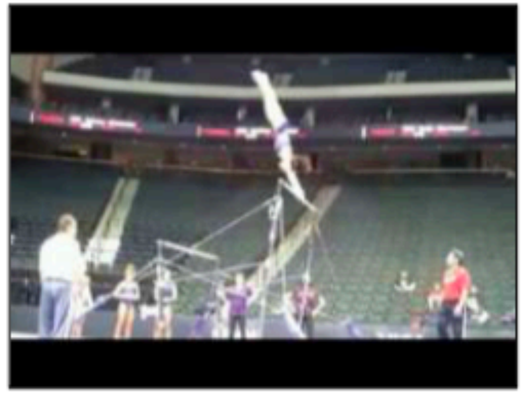


Given a start and an end, can this point lie in between?

Shuffle & Learn

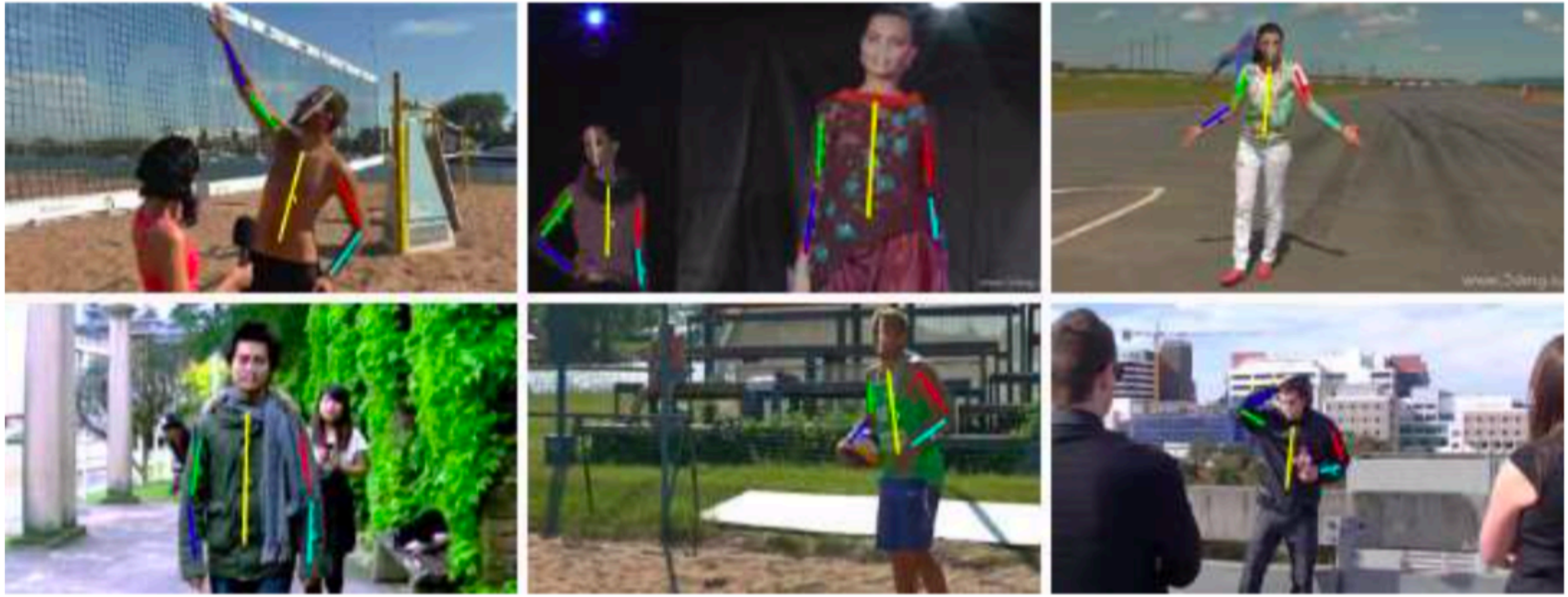


Nearest Neighbors of Query Frame (fc7 features)

Query	ImageNet	Shuffle & Learn	Random
			
			

Shuffle & Learn

Fine-tune on Human Keypoint Estimation

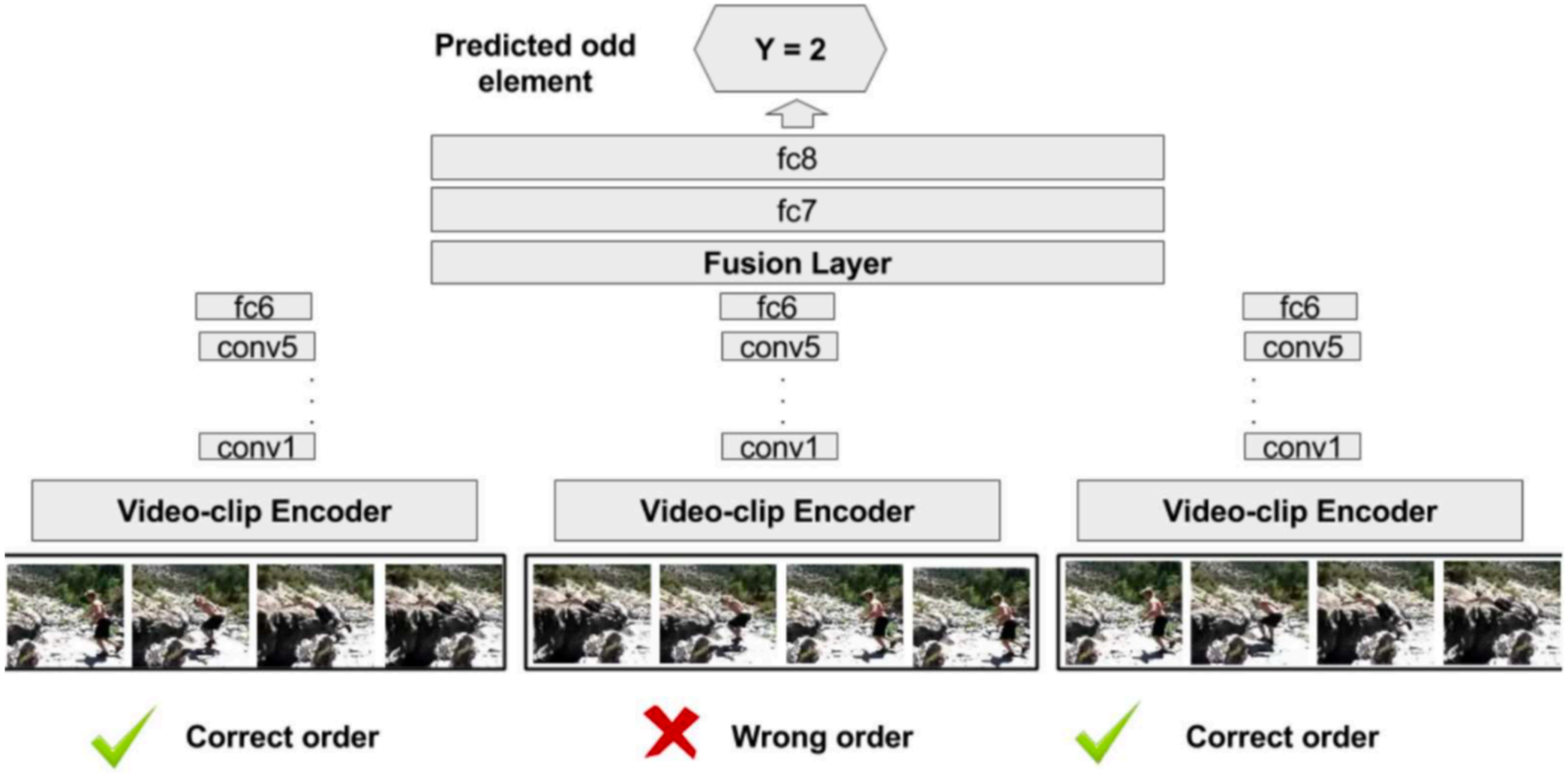


Shuffle & Learn

Fine-tune on Human Keypoint Estimation

Initialization (AlexNet)	End task	
	FLIC Dataset Keypoints AUC	MPII Dataset Keypoints AUC
ImageNet Supervised	**51.3**	47.2
Shuffle and Learn (Self-supervised)	49.6	**47.6**

Odd-one-out Networks



Fernando et al., 2017, Odd-one-out networks

Self-supervision in computer vision

- Using images
- Using video
- Using video and sound

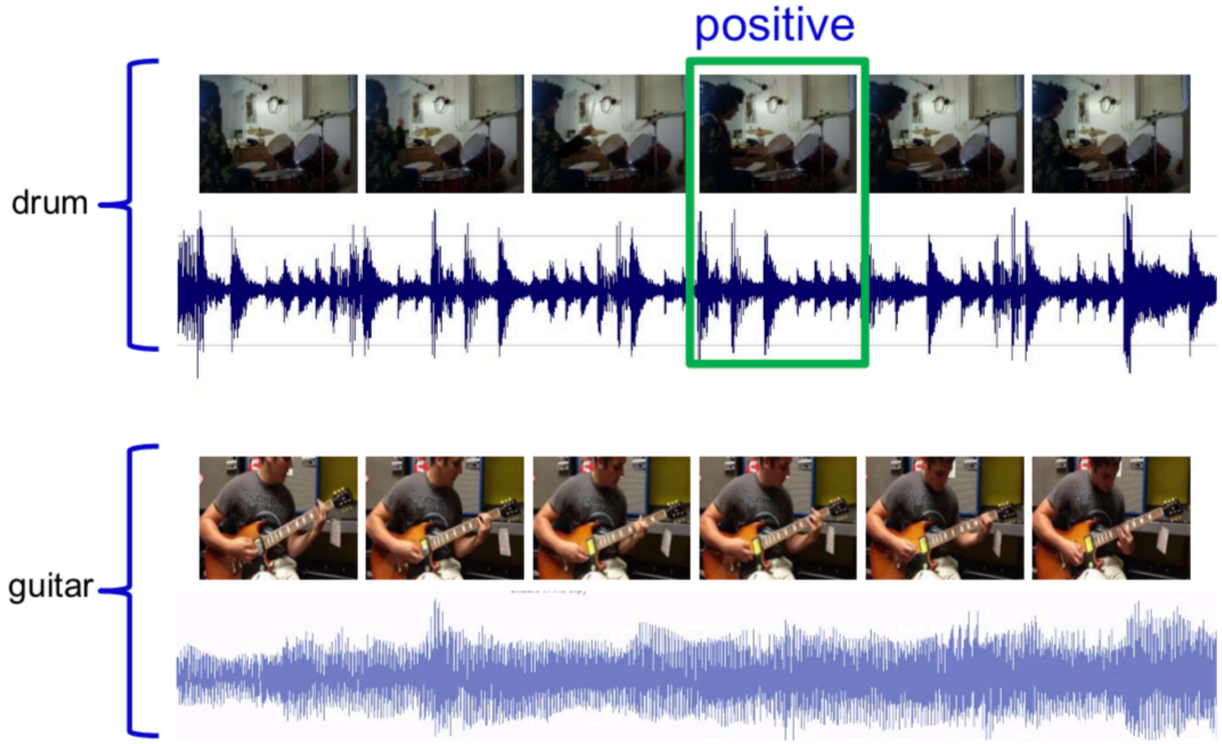
Audio-Visual co-supervision

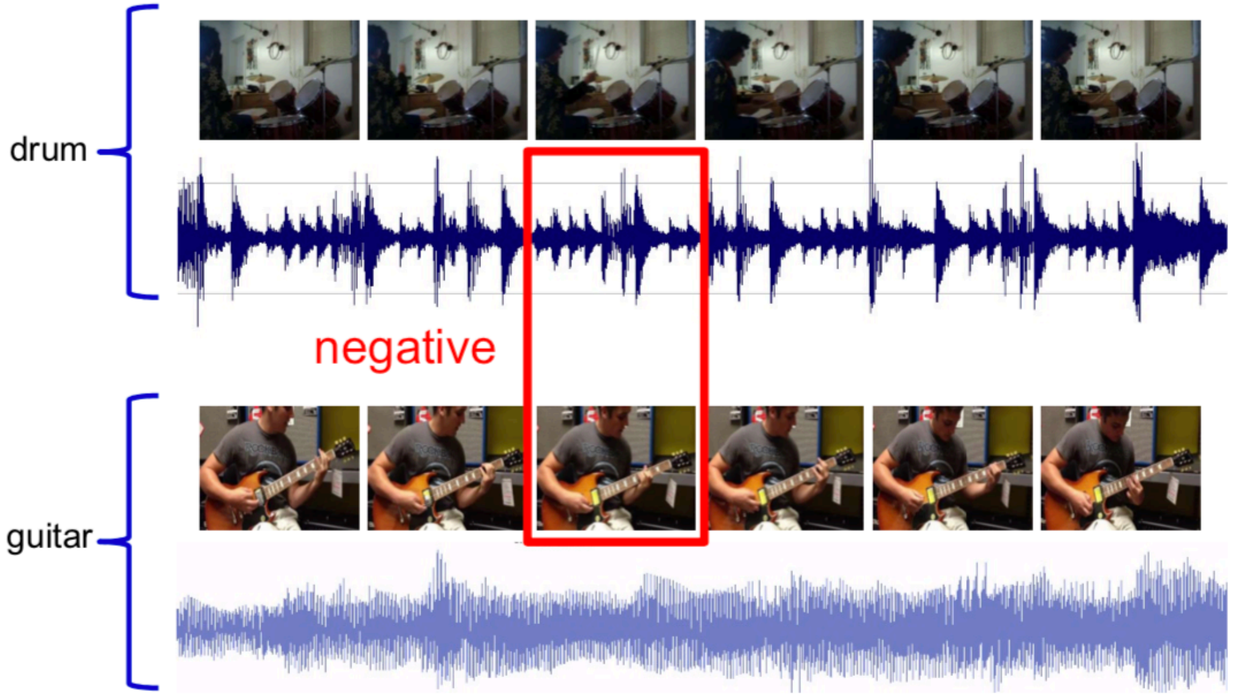
Train a network to predict if **image** and audio clip correspond

Correspond?

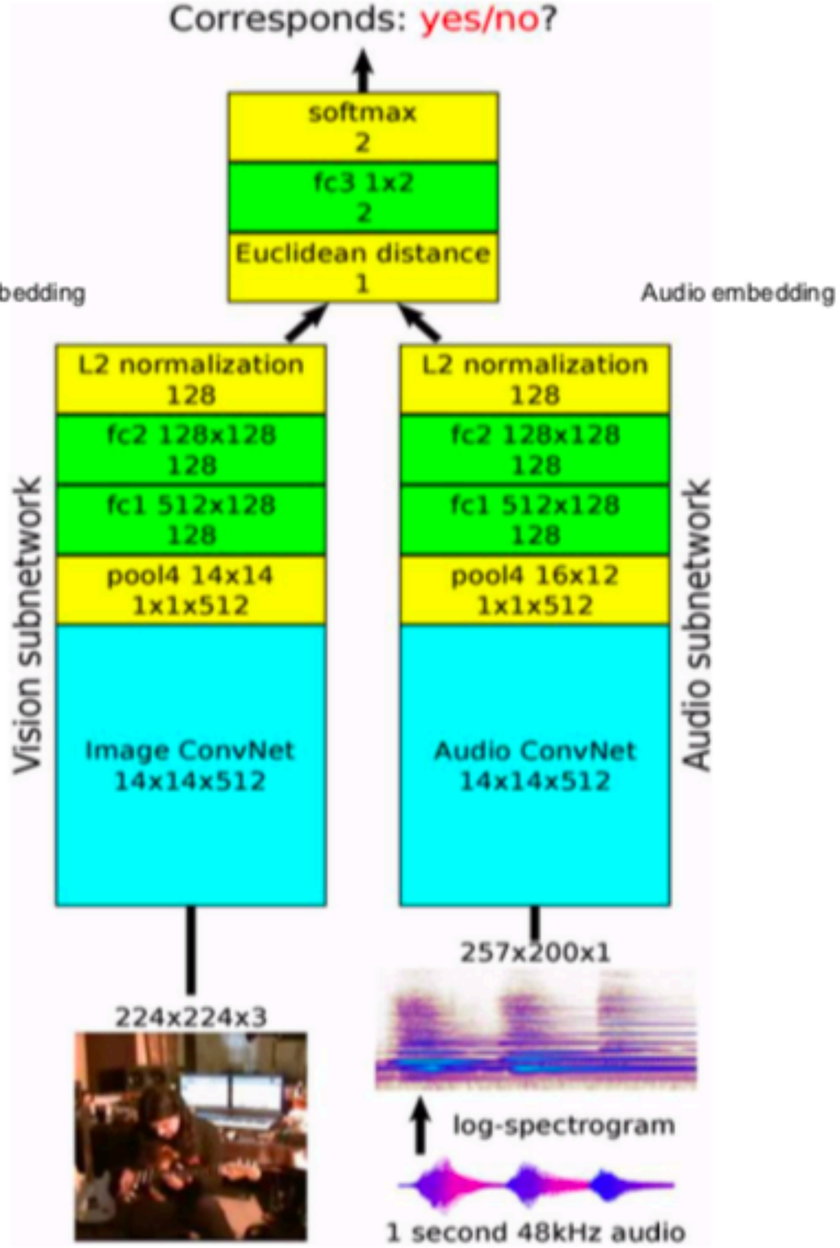
Arandjelović and Zisserman, 2017, “Objects that Sound”

Objects that Sound

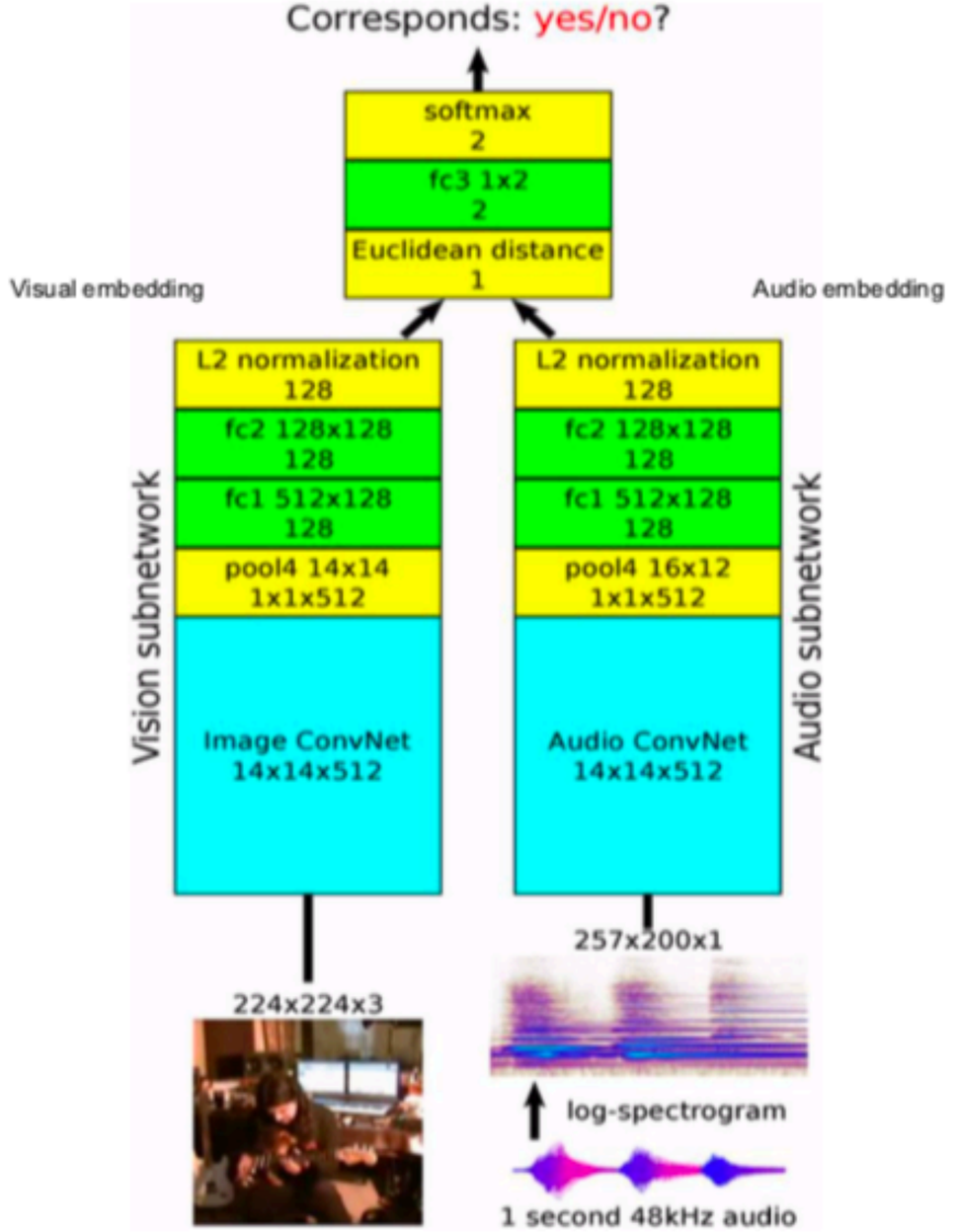




Objects that Sound



Objects that Sound



What can be learnt?

- Good representations – Visual features – Audio features

- Intra- and cross-modal retrieval – Aligned audio and visual embeddings

- "What is making the sound?" – Learn to localize objects that sound

Objects that Sound

What would make this sound?

Note, no video (motion) information is used

Understanding what the ″pretext″ task learns

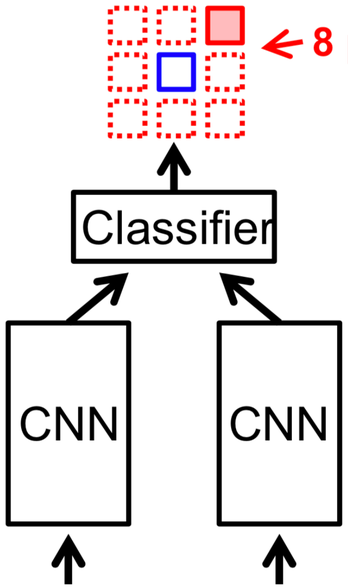
Are they complementary?

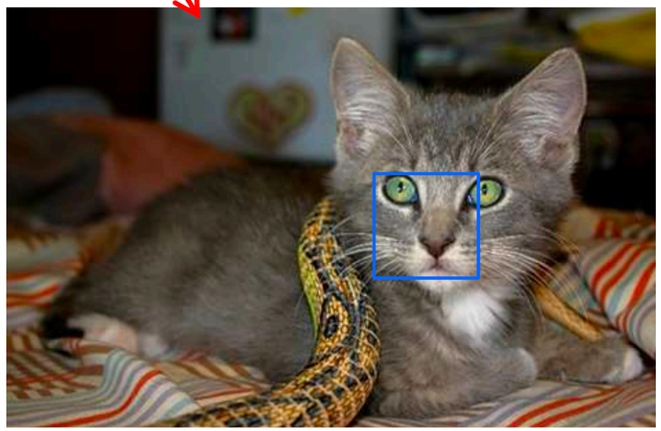
Initialization (ResNet101)	End task	
	ImageNet top-5 accuracy	VOC07 Detection mAP
Relative Position	59.2	66.8
Colorization	62.5	65.5
Relative Position + Colorization (Multi-task)	66.6	68.8

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

Information predicted: varies across tasks

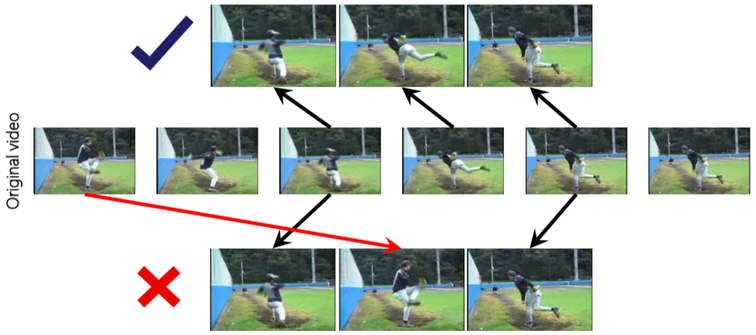
Less → More

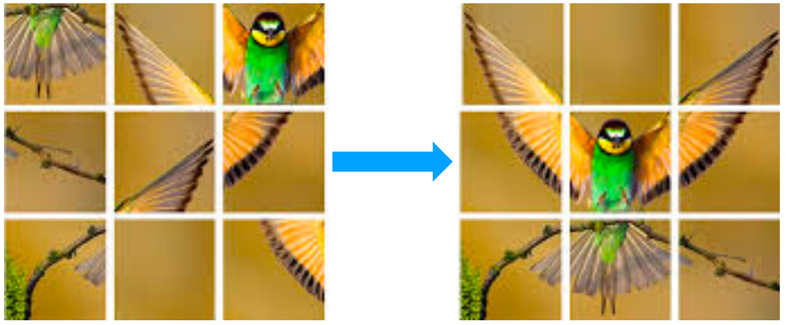




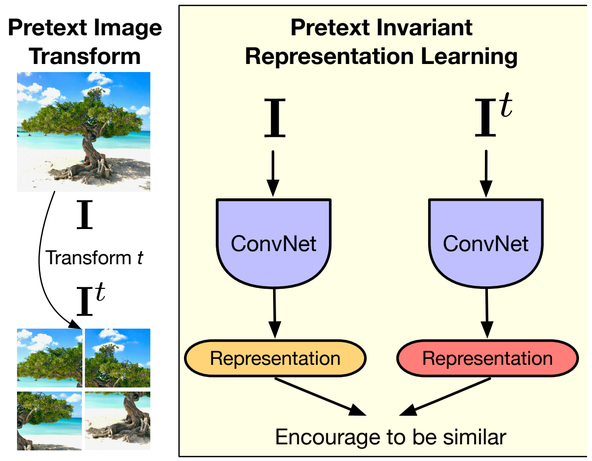
Randomly Sample Patch
Sample Second Patch

Pretext tasks

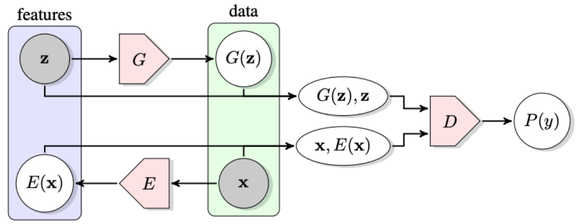




Contrastive



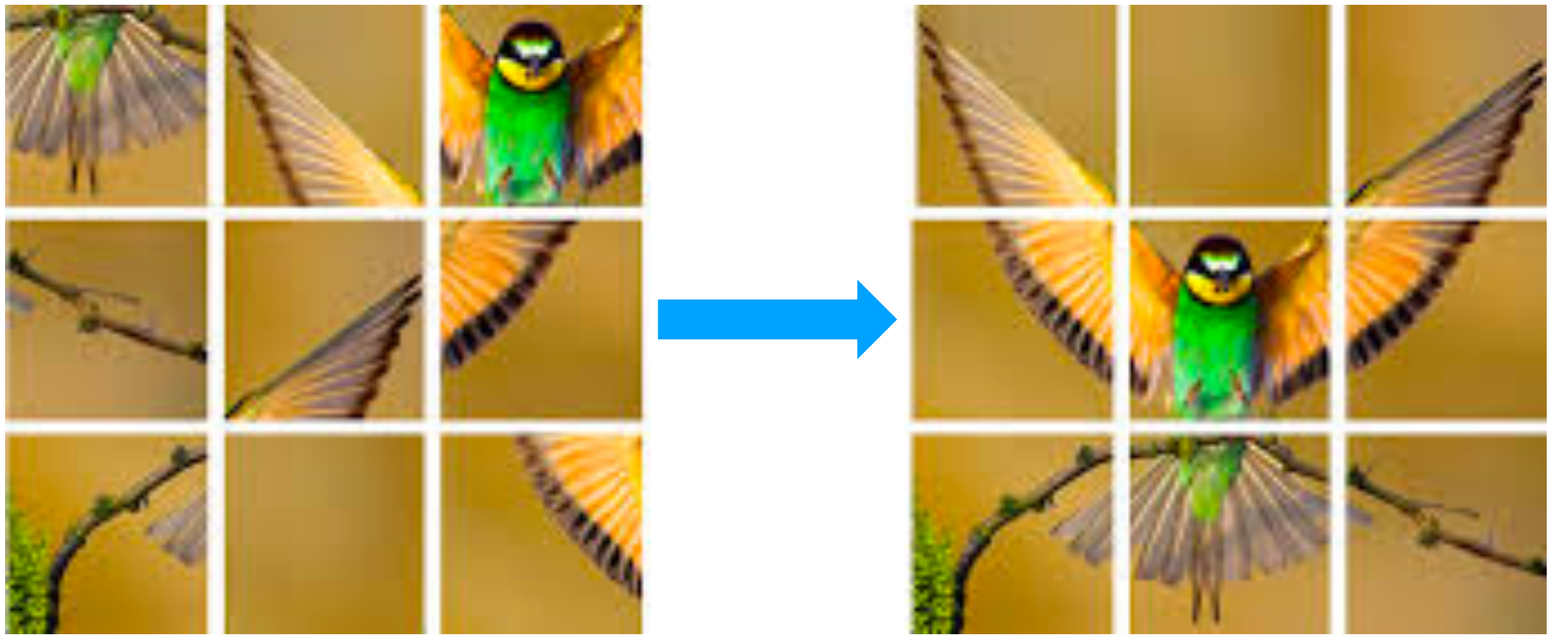
Generative



AutoEncoder,
VAE, GAN,
BiGAN

Predict more information

Scaling self-supervised learning

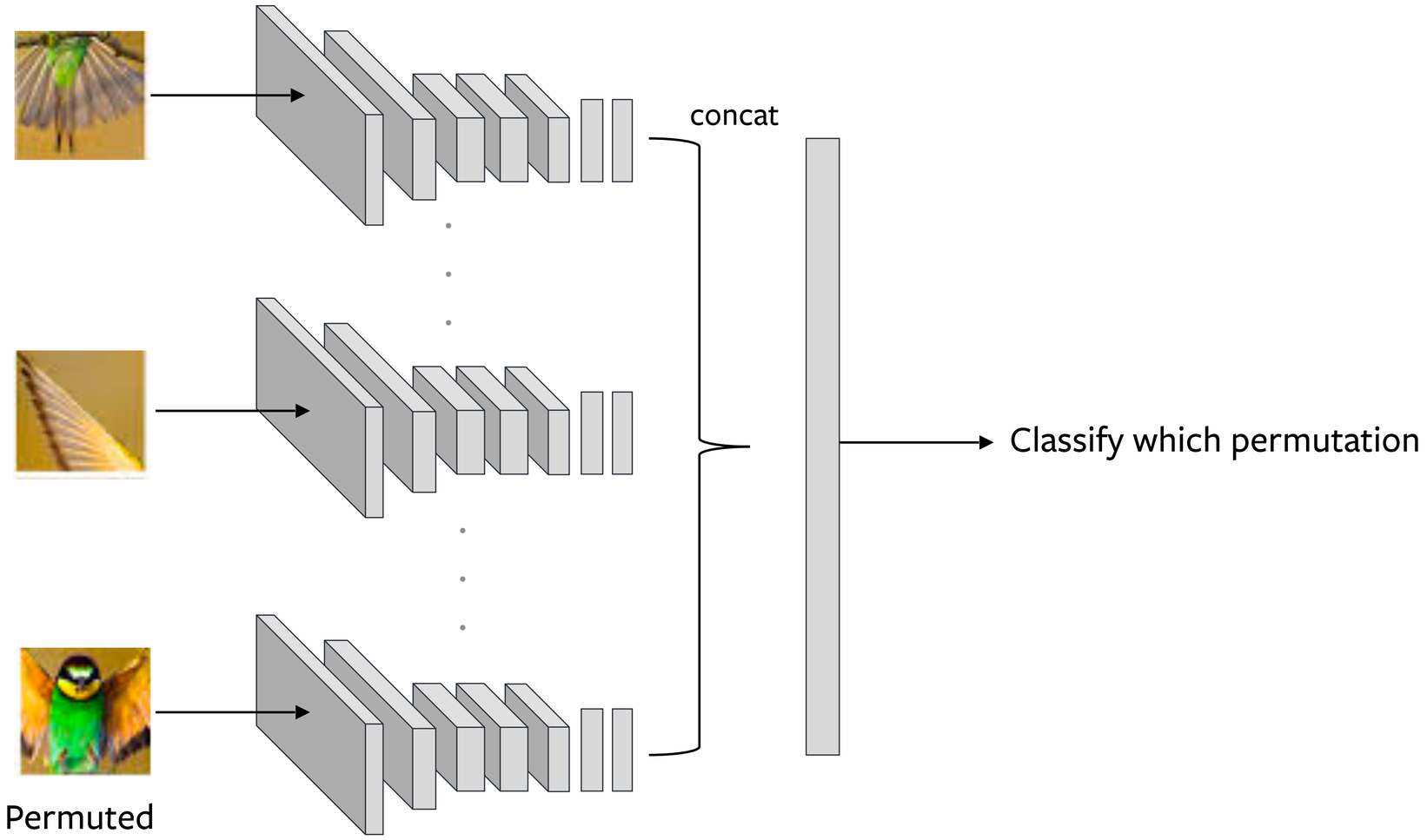


Jigsaw puzzles

(Noorozi & Favaro, 2016)

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

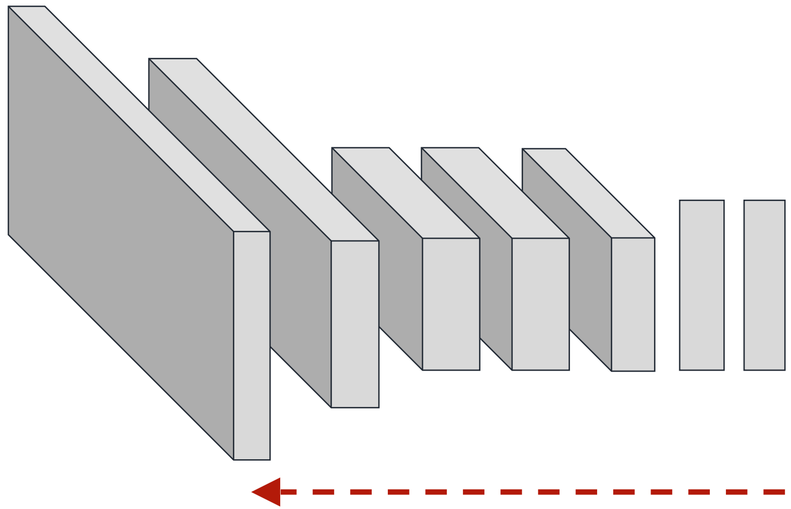
Jigsaw Puzzles



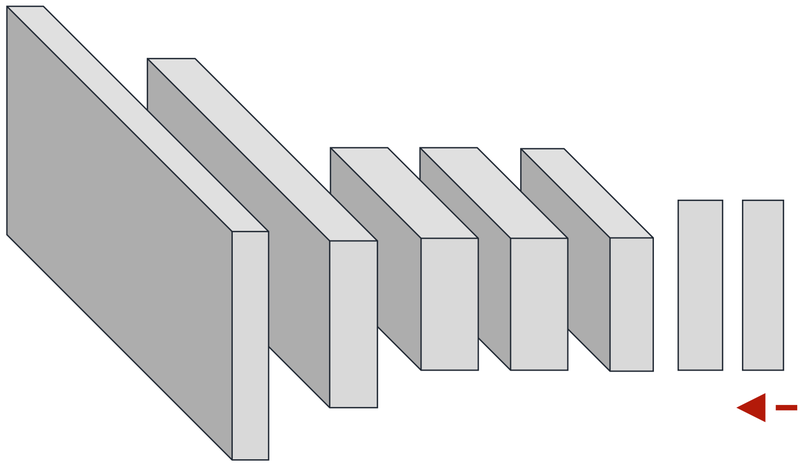
- Use N=9 patches
- In practice, use a subset of permutations
- E.g. 100 from 9!
- Each patch is processed independently
- N-way ConvNet (shared params)

- Problem Complexity
  - Size of subset

Evaluation – fine-tuning vs. linear classifier



Fine-tune all layers



Linear classifier

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

Evaluation – many tasks

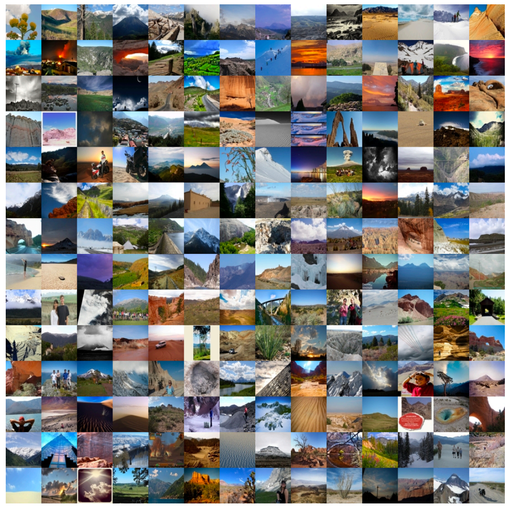
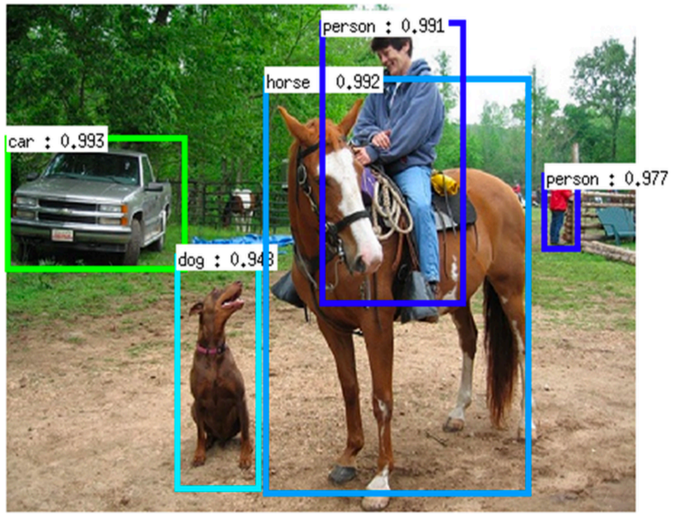


Image classification

Few-shot learning

ImageNet, Places-205, VOC'07, COCO



Object detection

VOC'07

3D Understanding

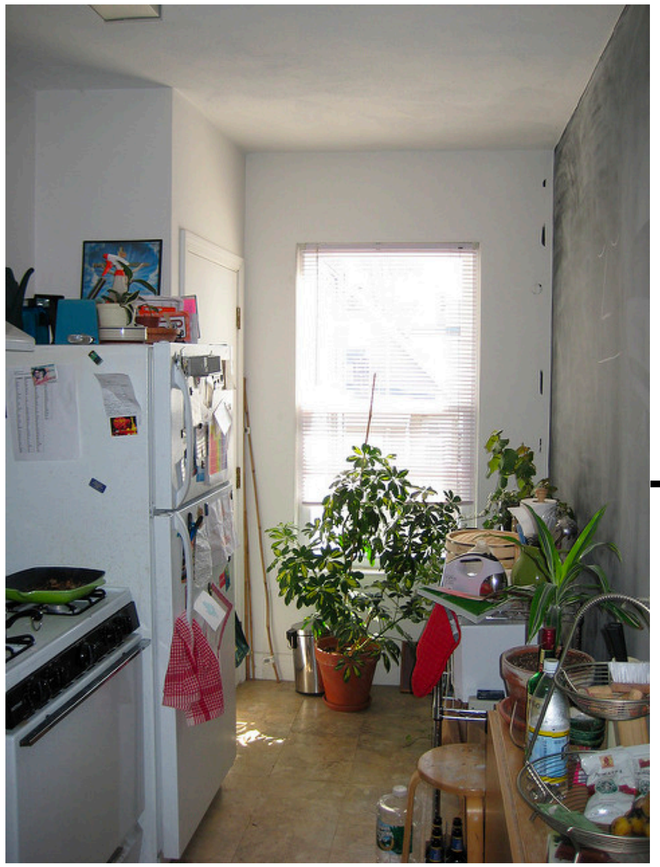
Surface Normals – NYUv2

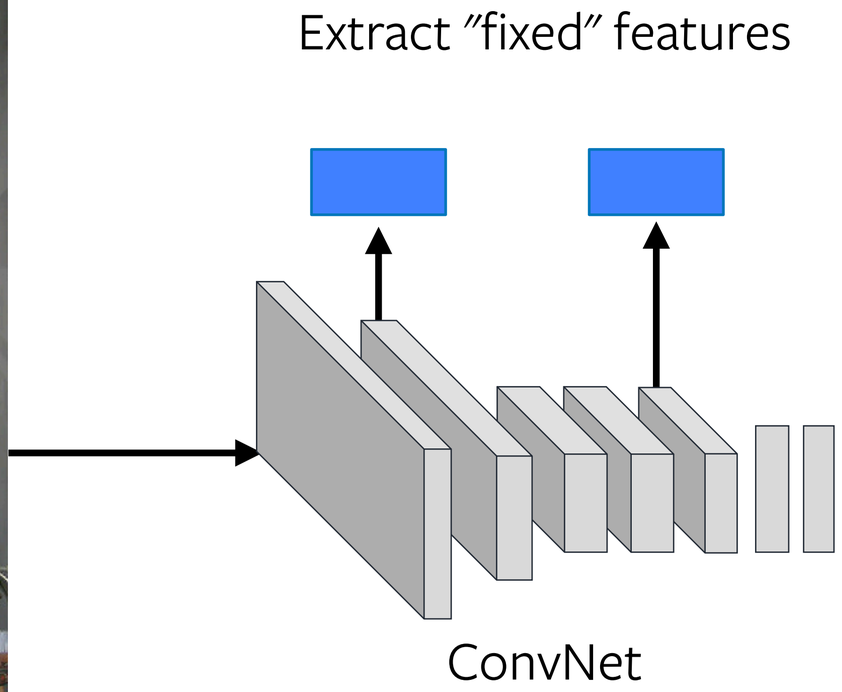
Navigation

Gibson environment

Images from the Places, VOC07, NYUv2 and Gibson datasets

Evaluating the representation

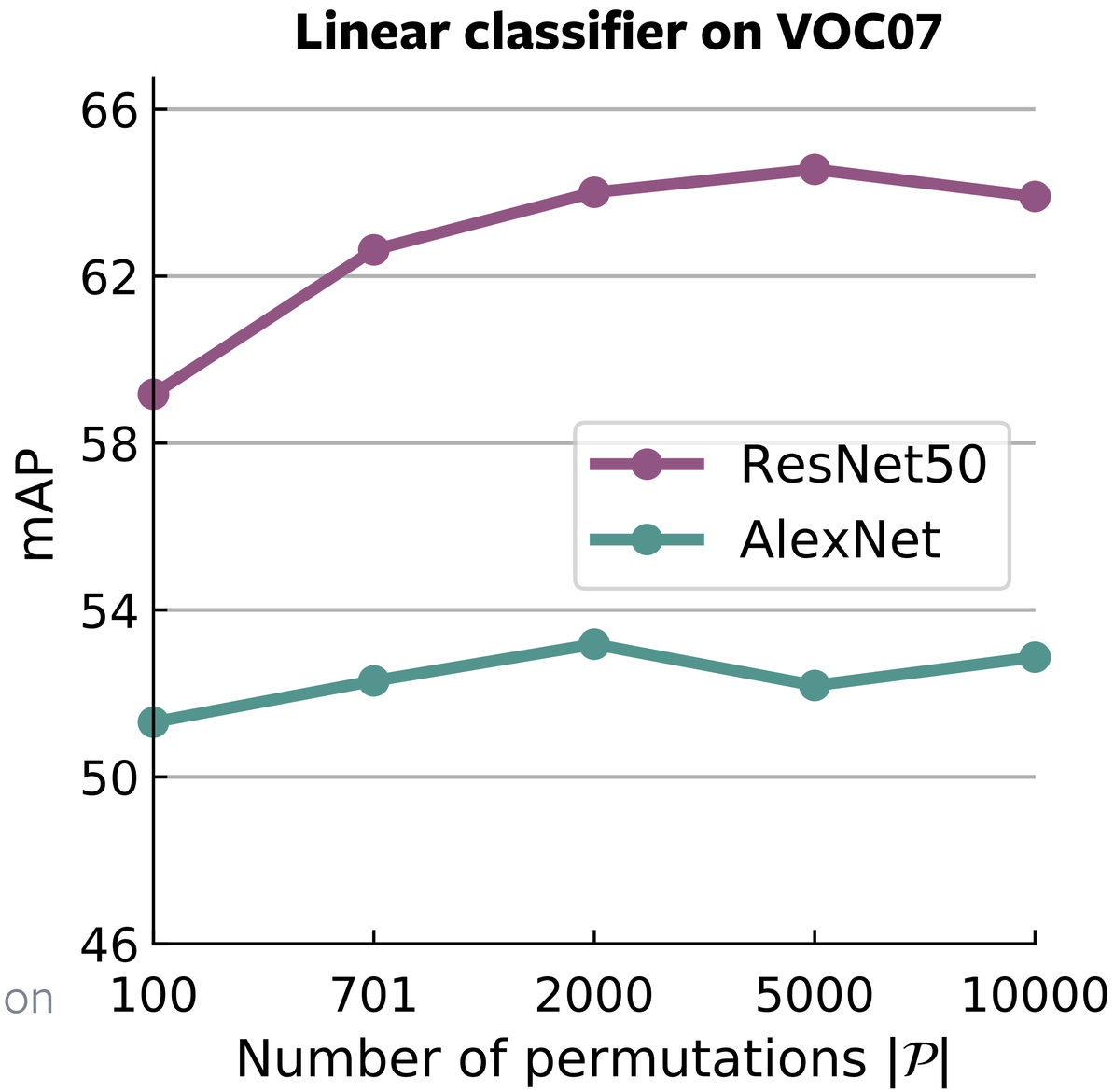




Evaluating the representation

- Train a Linear SVM on **fixed feature** representations
- Use the VOC07 image classification task

Increasing amount of information predicted



mAP = mean Average Precision
(Higher is better)

Object Detection

- Fast R-CNN (Girshick et al., 2015)
  - Same optimization parameters for all methods (including supervised)
  - Use VOC'07

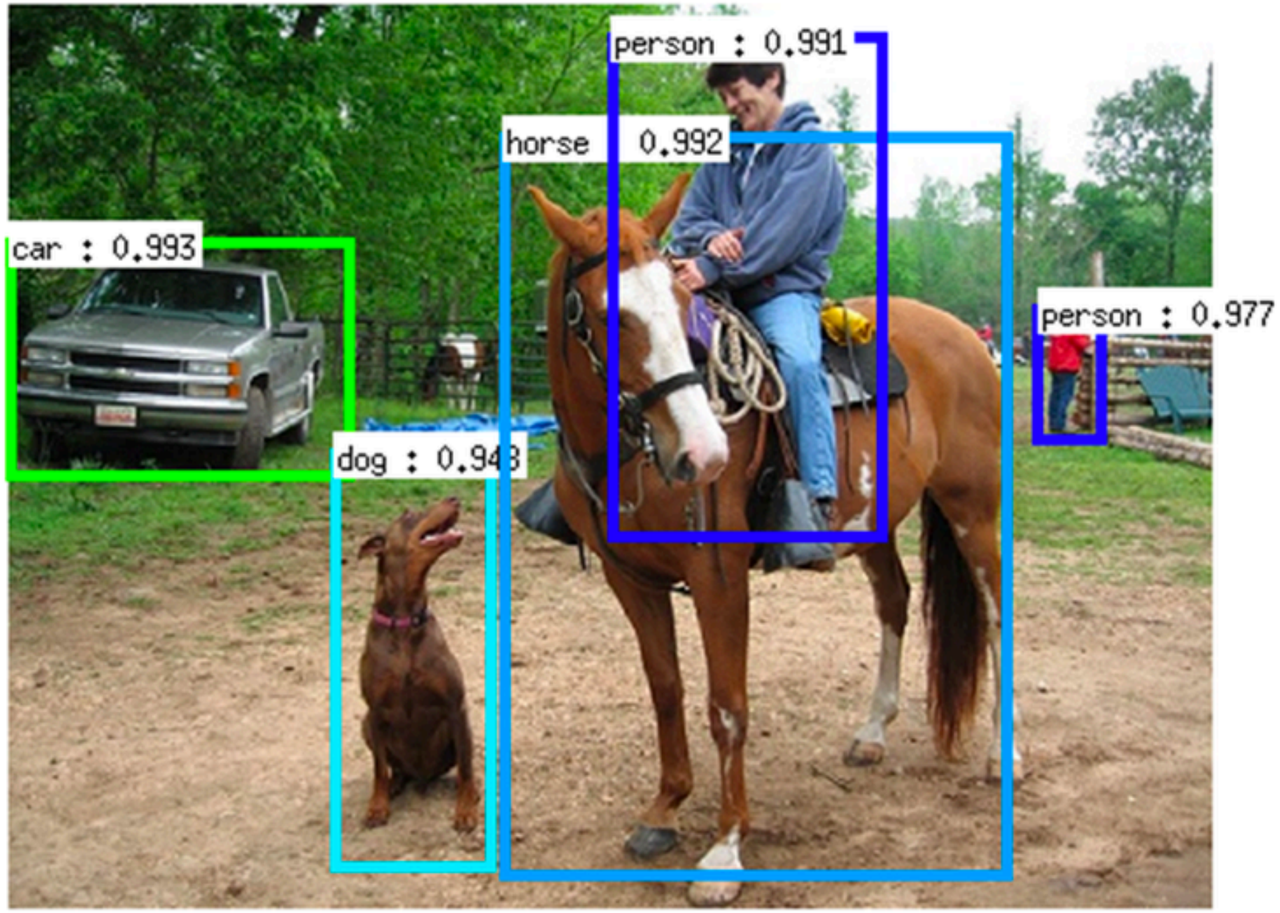
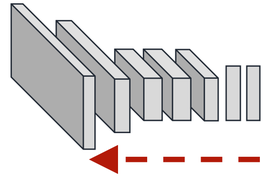


Image from the VOC07 dataset

Object Detection



Fine-tune all

VOC07 test set.
Fast R-CNN ResNet50

Initialization	Train Set	
	VOC07	VOC07+12
ImageNet Supervised	70.5	76.2
Jigsaw ImageNet 14M	69.2	75.4

Within error of ImageNet pre-trained network

Surface Normal Estimation

- Predict surface normals on NYU-v2
  - Same optimization parameters for all methods (including supervised)
  - PSPNet Architecture
  - Train last few layers only (res5 onwards)

Input

Output

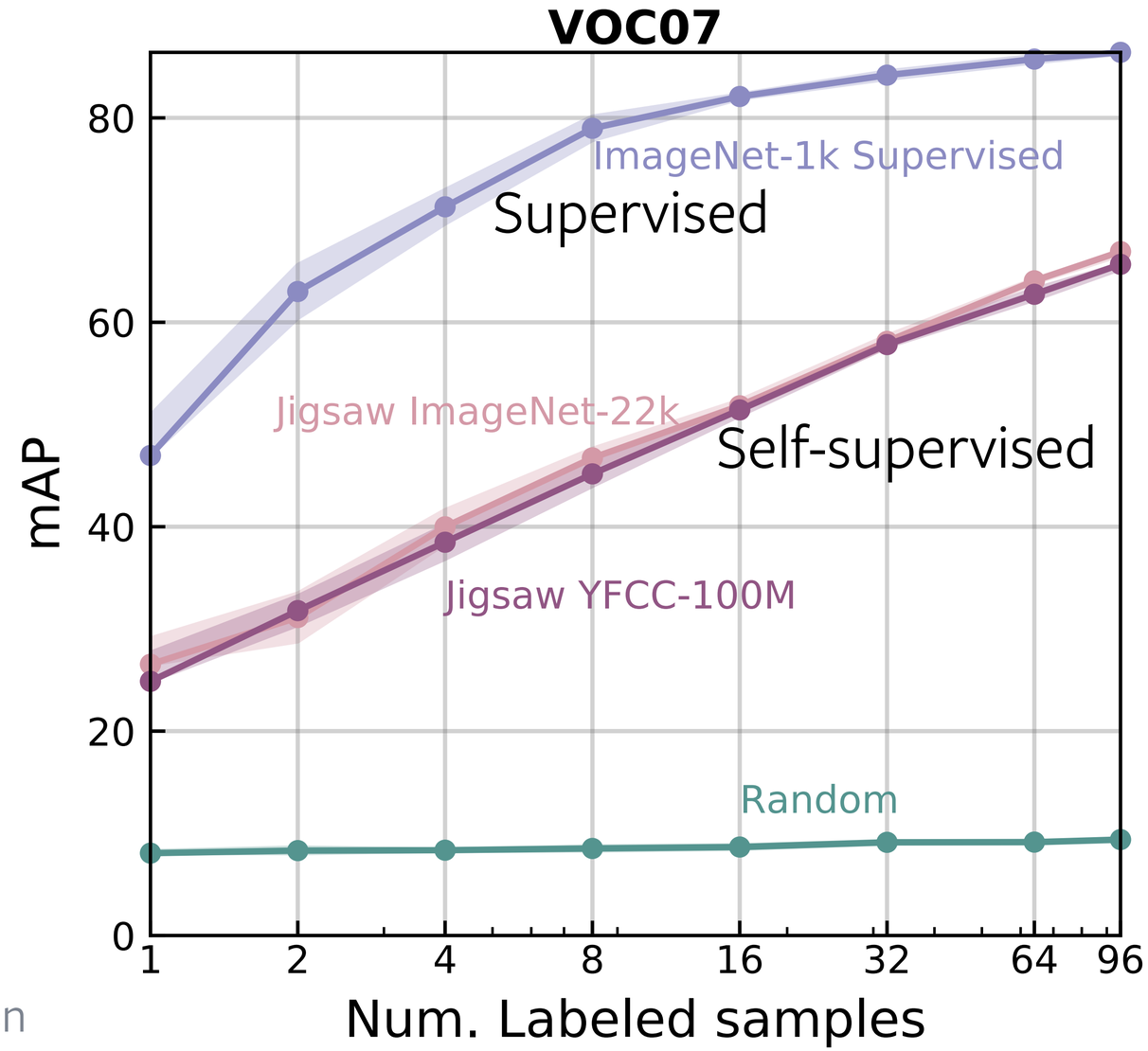
Image from the NYU dataset

Surface Normal Estimation

Initialization	Median Error (Lower better)	% correct within 11.25^0 (higher better)
ImageNet Supervised	17.1	36.1
Jigsaw Flickr 100M	**13.1**	**44.6**

Outperforms ImageNet supervised

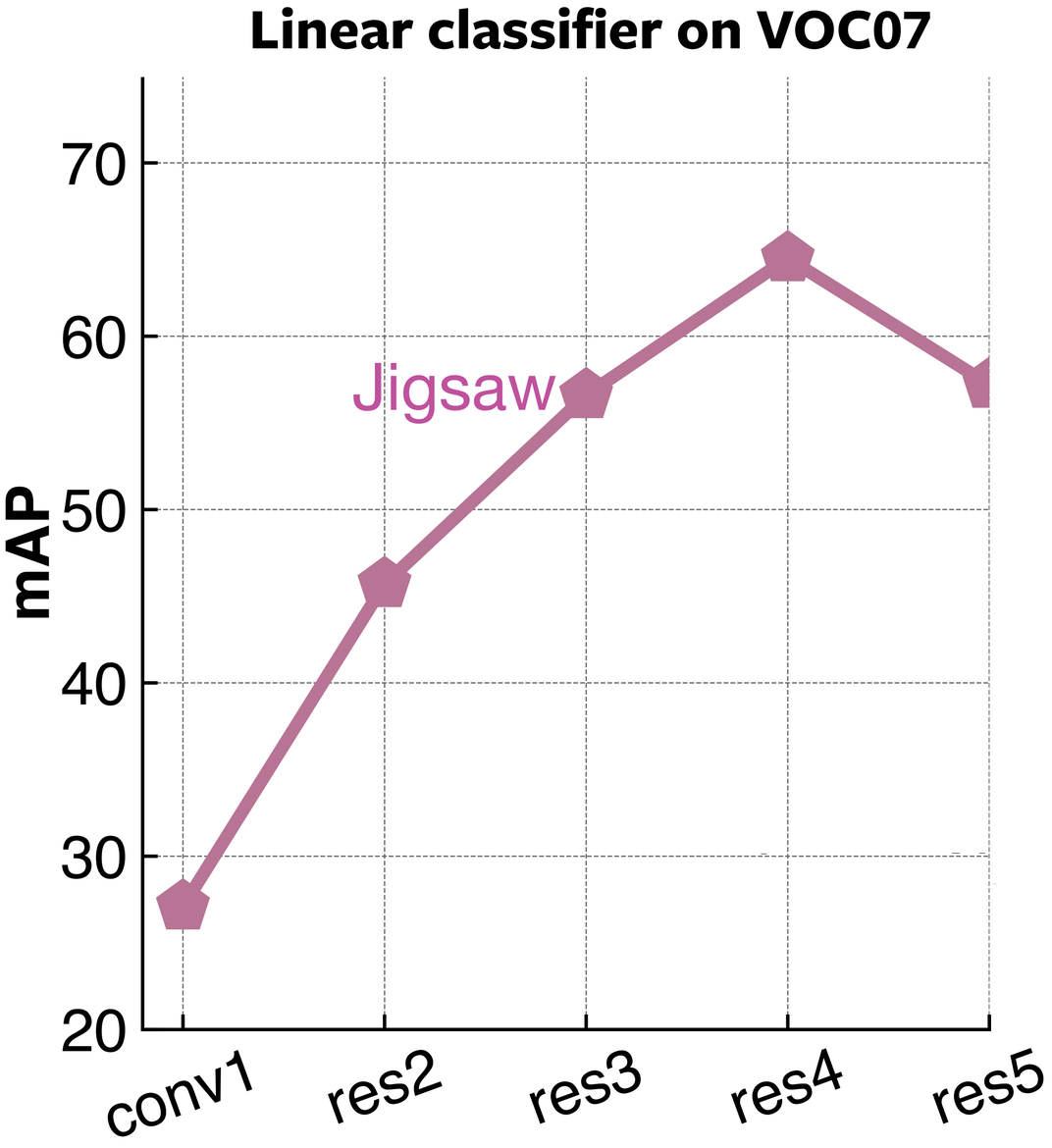
Few shot learning

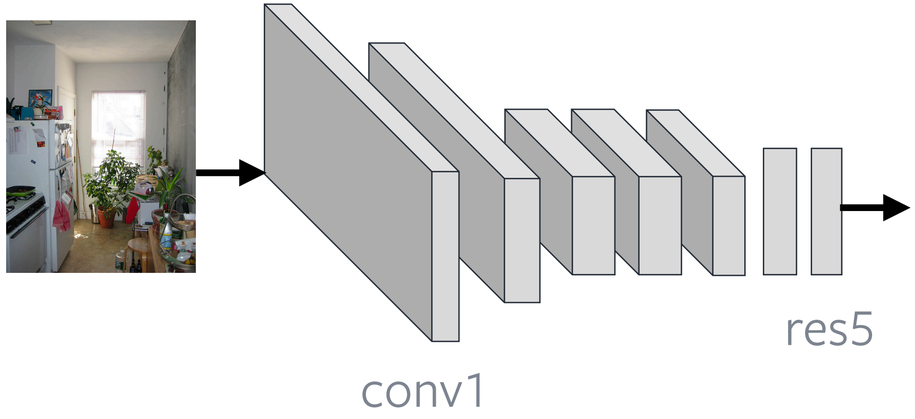


mAP = mean Average Precision
(Higher is better)

Self-supervised representations are not as sample efficient

What does each layer learn?





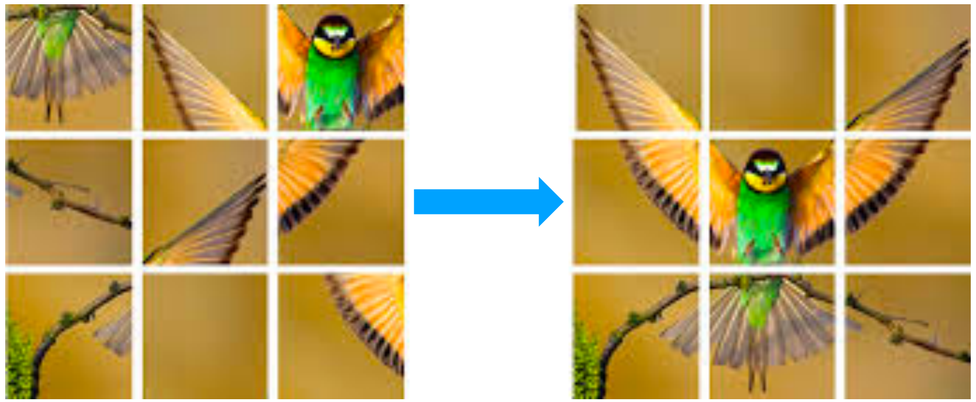
mAP = mean Average Precision
(Higher is better)

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

Pretext tasks

Rotation

(Gidaris et al., 2018)

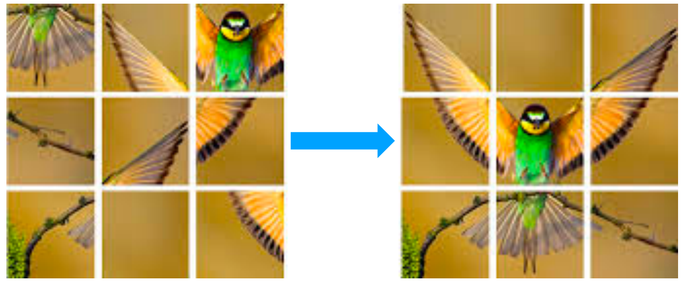


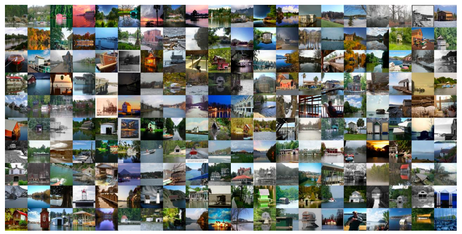
Jigsaw puzzles

(Noroozi et al., 2016)

The hope of generalization

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





Pre-training

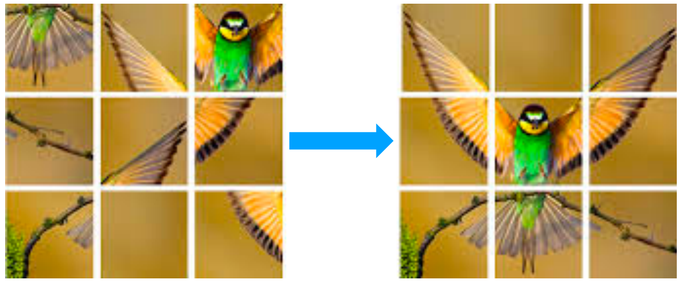
Self-supervised

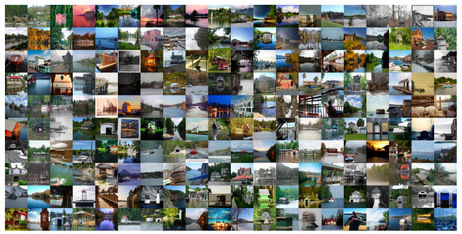
Transfer Tasks

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

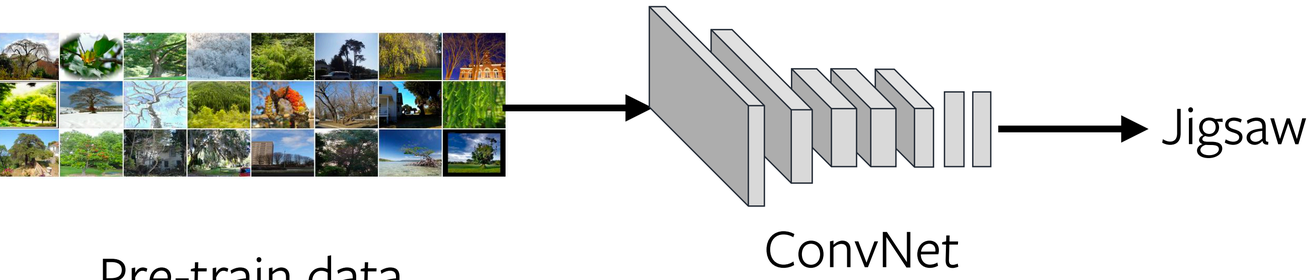
Weak or self-supervised

Transfer Tasks

Why should solving Jigsaw puzzles teach about "semantics"?

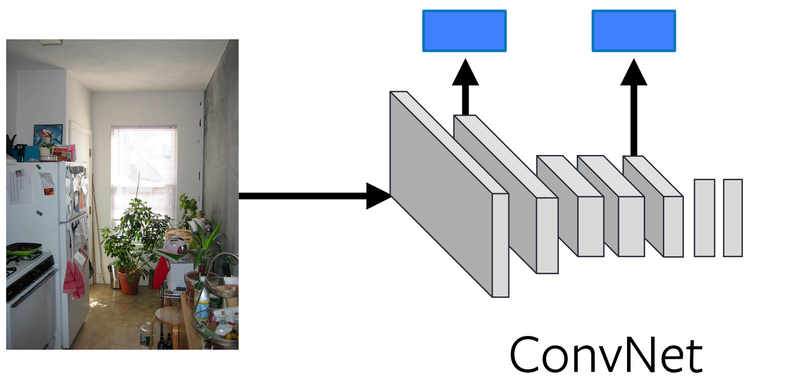
Why should performing a non semantic task produce good features?

The hope of generalization ... ?



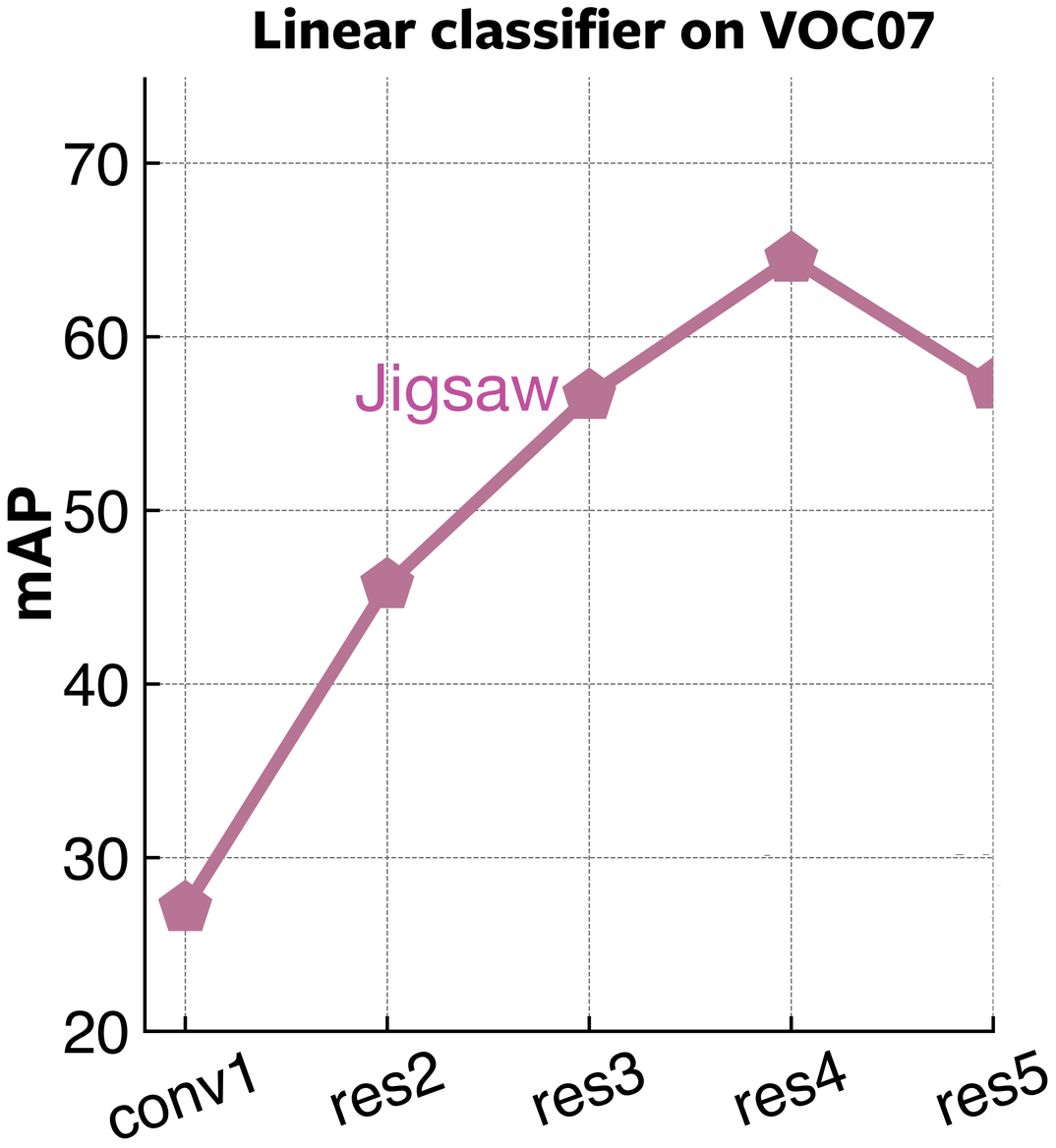
Pre-training

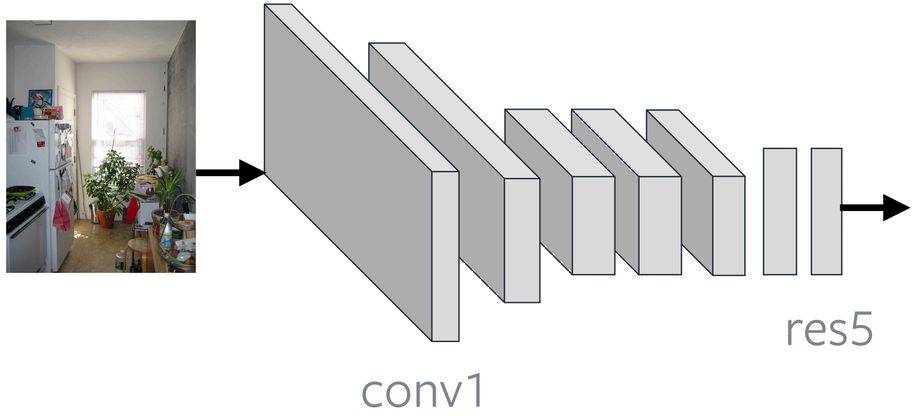
Weak or self-supervised



Transfer

Higher layers do not generalize ...





mAP = mean Average Precision
(Higher is better)

Pre-trained features should ...

- Represent how images relate to one another

- Be robust to "nuisance factors" -- Invariance
- e.g., exact location of objects, lighting, exact color

Pre-trained features should ...

- Represent how images relate to one another

- Be robust to "nuisance factors" -- Invariance
- e.g., exact location of objects, lighting, exact color

Clustering and Contrastive Learning are two ways to achieve the above

Pre-trained features should ...

- Represent how images relate to one another
  - **ClusterFit**: Improving Generalization of Visual Representations

- Be robust to invariances, e.g., data transforms of the input
  - **PIRL**: Self-supervised learning of Pre-text Invariant Representations

See also:
DrLim (Hadsell et al. 2006), Exemplar CNN (Dosovitsky et al, 2014, Bautista et al., 2016), Visual Groups (Isola et al., 2016), Cluster (Xie et al., 2016), NAT (Bojanowski et al., 2017) Transitive (Wang et al., 2017), Boosting Knowledge (Noroozi et al., 2018), DeepCluster (Caron et al., 2018, 2019), CPC (Oord et al., 2018, Henaff et al., 2019), CMC (Tian et al., 2019) MoCo (He et al., 2019)

ClusterFit: Improving Generalization of Visual Representations

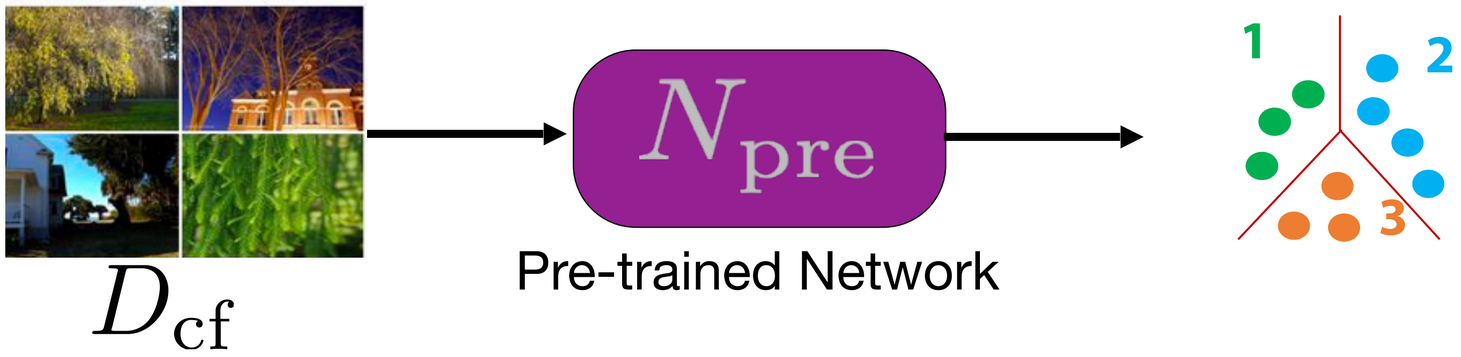
Yan*, Misra* et al., CVPR 2020

Understanding how images relate to each other

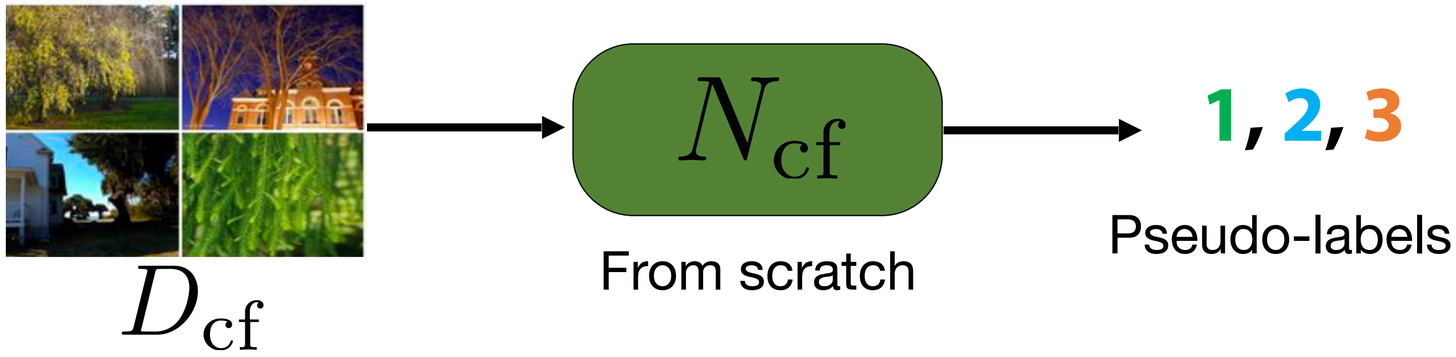
- Clustering the feature space is a way to see what images relate to one another

Main Idea

1. **Cluster**: Feature clustering

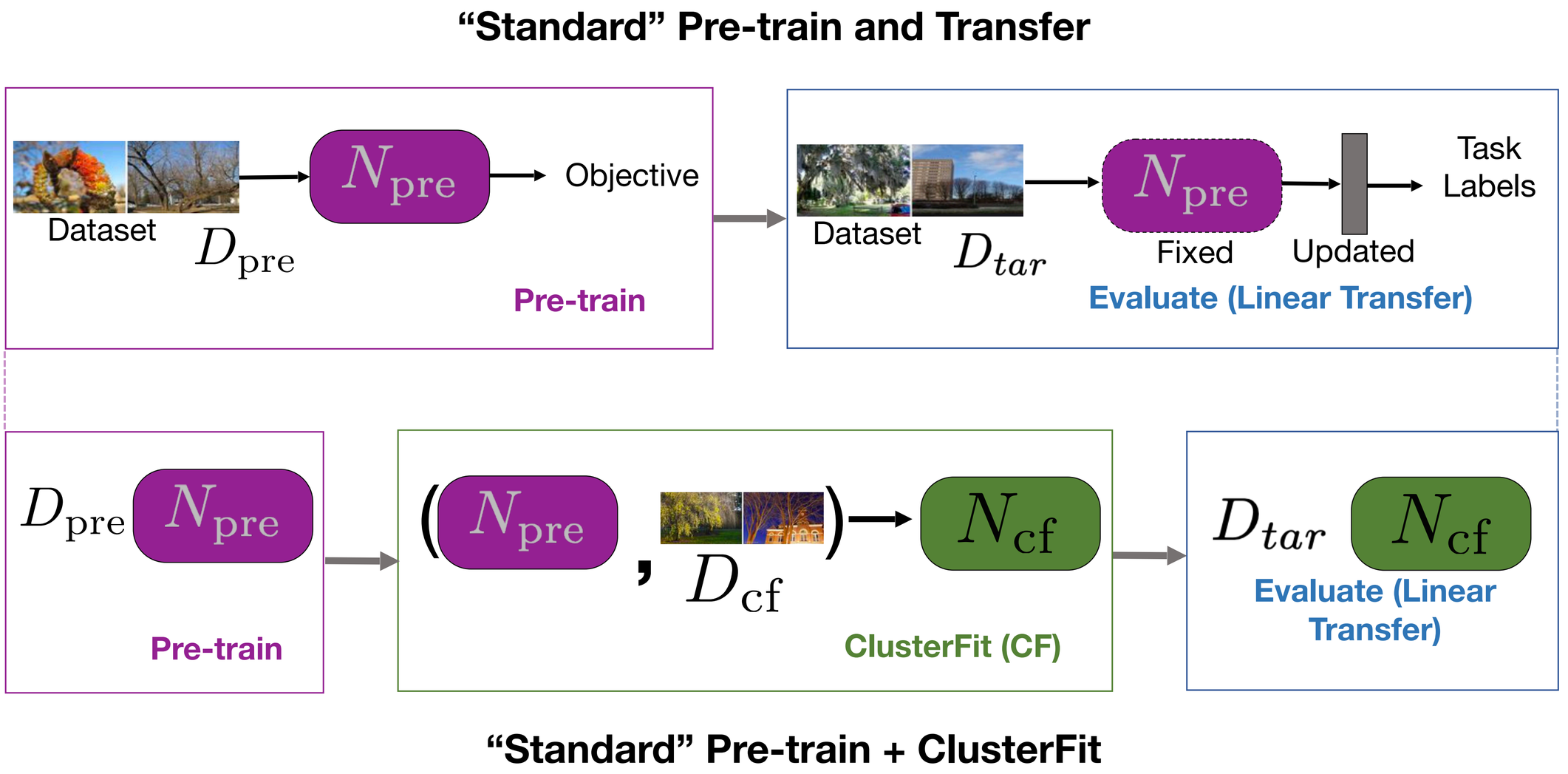


2. **Fit:** Predict Cluster Assignments



See also: Boosting Knowledge (Noroozi et al., 2018); DeepCluster (Caron et al., 2018); DeeperCluster (Caron et al., 2019)

Applied to any pre-trained network



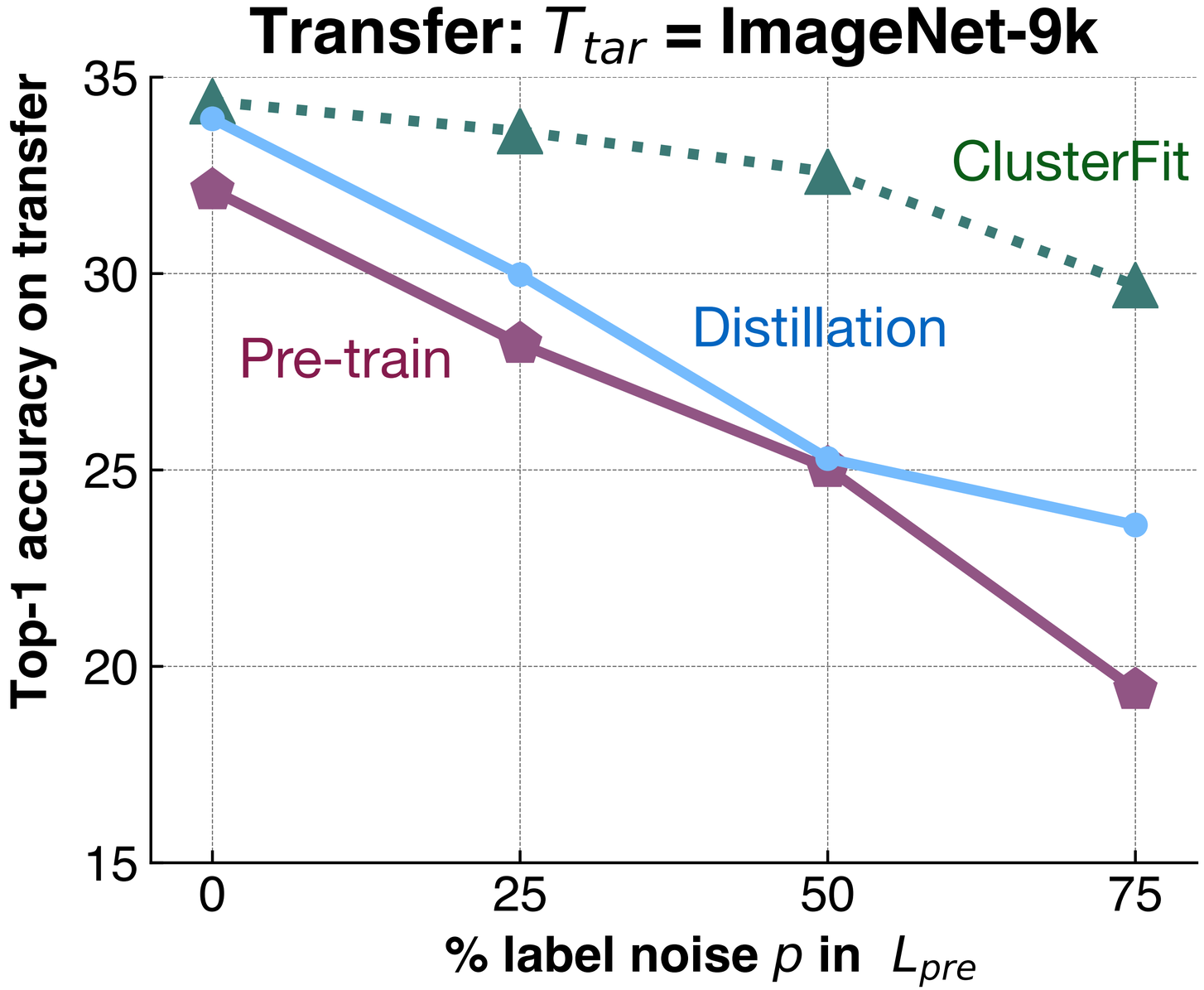
Why does it work?

- Clustering introduces a "bottleneck" and only the most important information is transferred

Synthetic noise

- Add label noise to ImageNet-1K (1000 classes)
- Measure transfer performance on ImageNet-9K (9000 classes)

Synthetic noise



Self-supervised Images

Method	Transfer Dataset: ImageNet-1M	VOC07	Places205	iNaturalist
Jigsaw Pretrain	46.0	66.1	39.9	22.1
Pretrain 2x	45.1	65.4	38.7	21.8
ClusterFit	**55.2** **+9**	**69.5** **+3**	**45.0** **+5**	**29.8** **+7**

ResNet50, ImageNet-1M

Similar gains for RotNet

Applied to any pre-trained network

Pre-training Setup	Network	Dataset	Gain of ClusterFit Top-1 Accuracy
Fully Supervised	ResNet50	ImageNet1K	2.1% on ImageNet-9K
Weakly Supervised Images	ResNet50	IG 1B	4.6% on ImageNet-9K 5.8% on iNaturalist
Weakly Supervised Videos	R(2+1)D-34	IG 19M	3.2% on Kinetics 4.3% on Sports1M
Self-supervised (Jigsaw, RotNet)	ResNet50	ImageNet1K	7-9% on ImageNet-1K 3-5% on Places-205

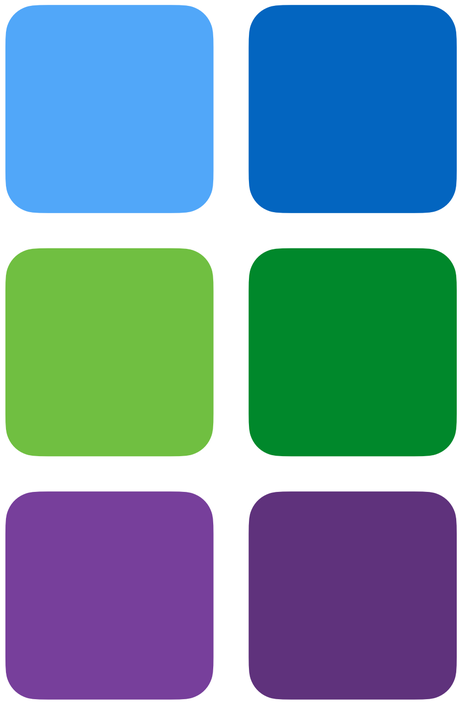
Gains without extra data, labels or changes in architecture

Self-supervised Learning of Pretext Invariant Representations (PIRL)

Misra & van der Maaten, CVPR 2020

Contrastive Learning

Groups of
Related and Unrelated
Images

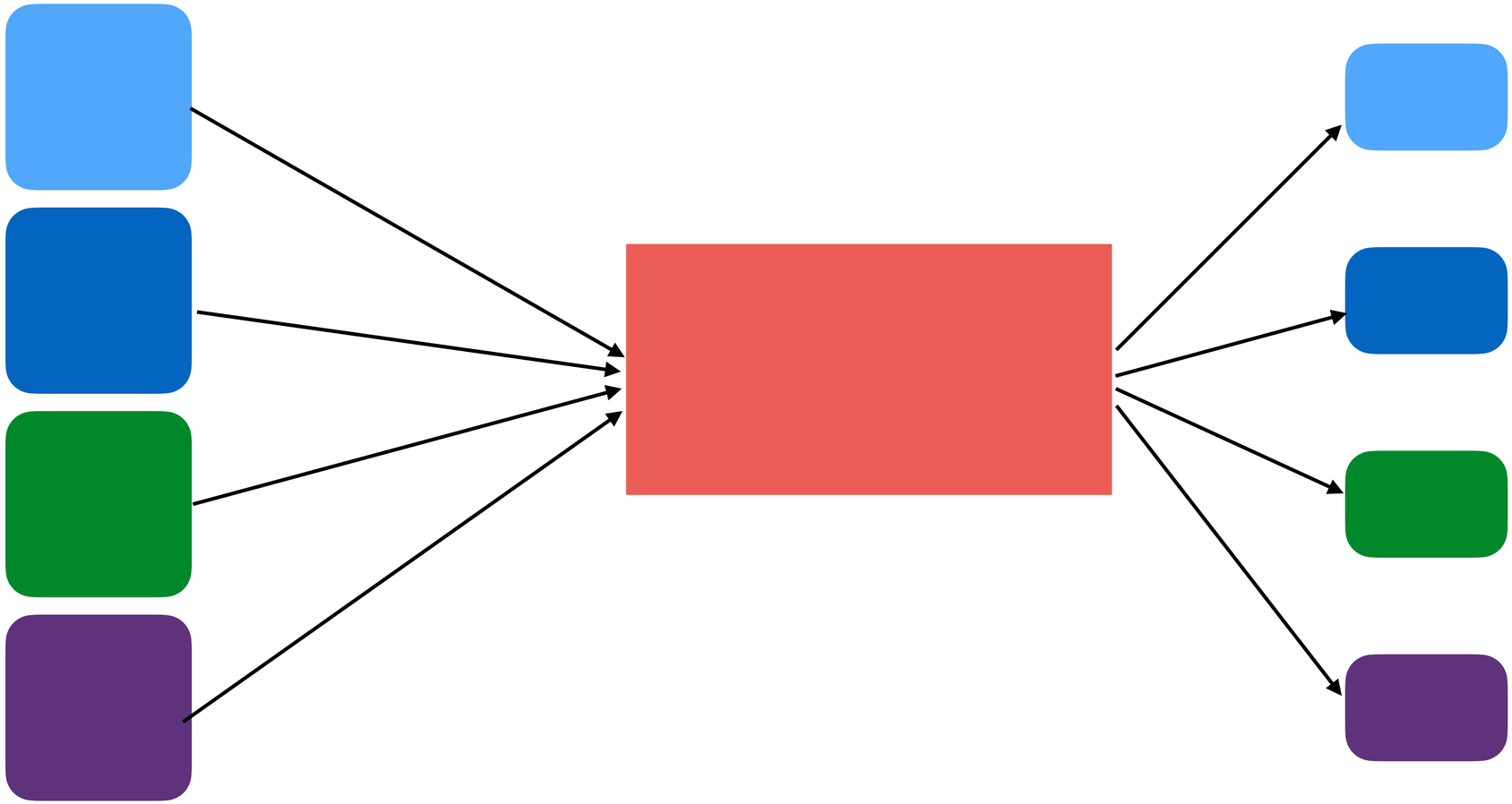


Contrastive Learning

Groups of Related and Unrelated Images

Shared network (Siamese Net)

Image Features (Embeddings)

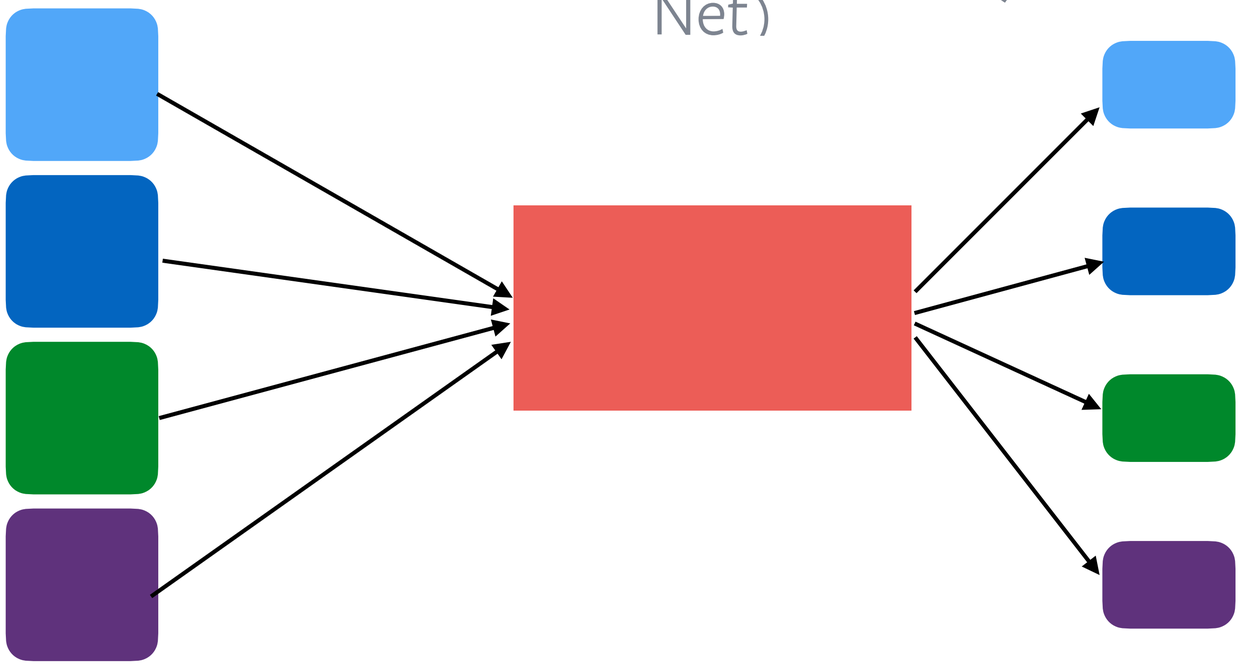


Contrastive Learning

Related and Unrelated Images

Shared network (Siamese Net)

Image Features (Embeddings)



Loss Function

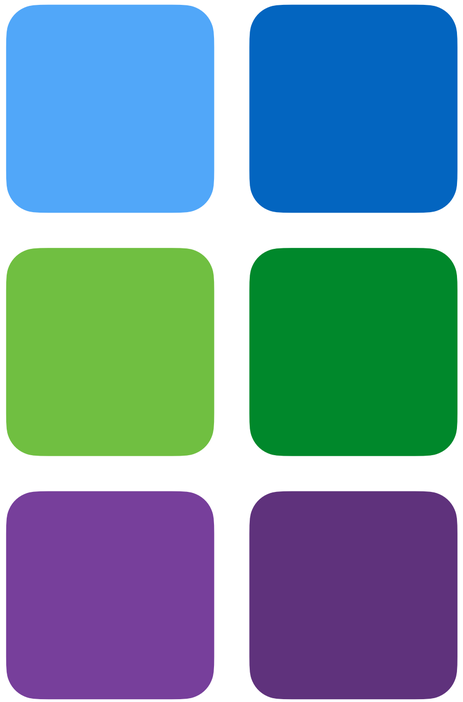
Embeddings from related images should be closer than embeddings from unrelated images

Hadsell et al., 2005, DrLim

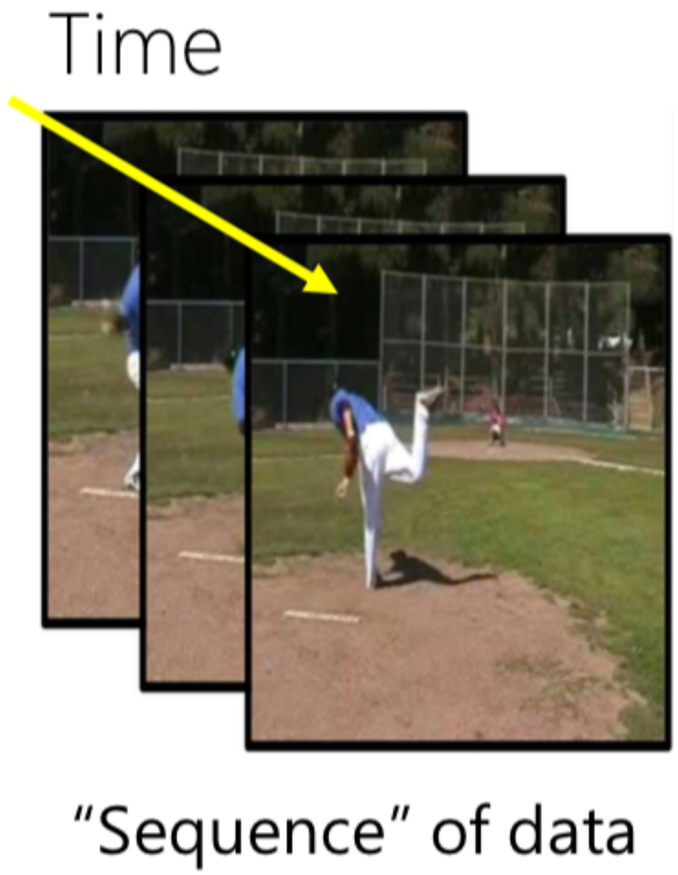
Contrastive Learning

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

Related and Unrelated Images



Frames of a video

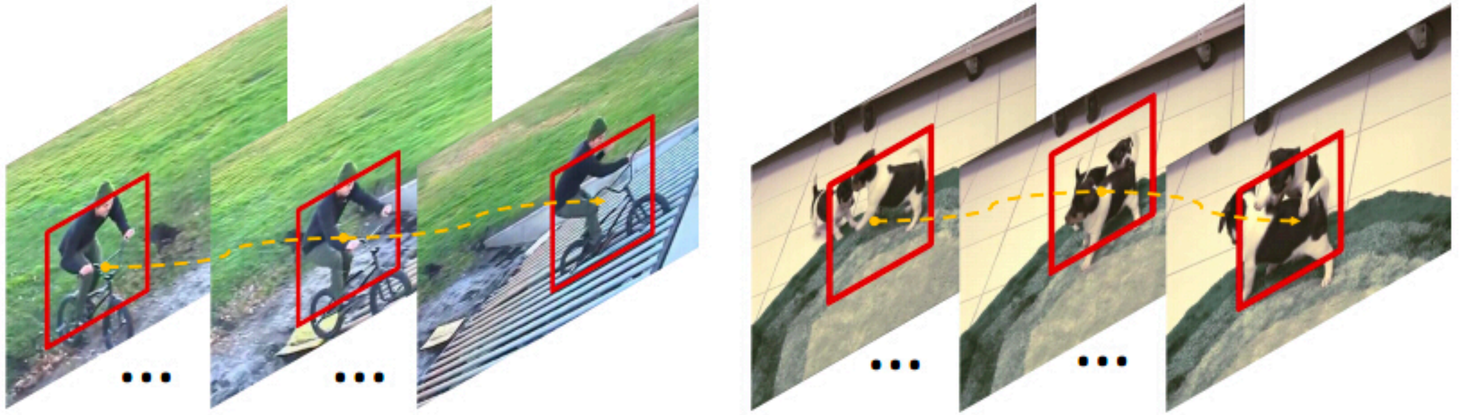


Hadsell et al., 2005, DrLim
van der Oord et al., 2018, CPC

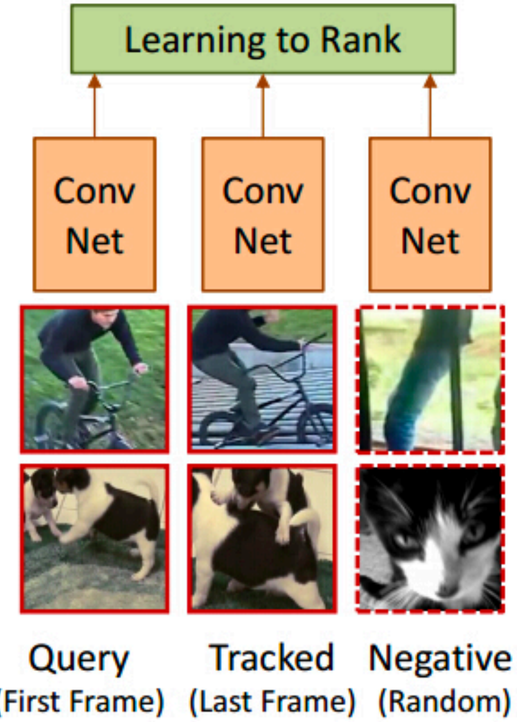
Video & Audio

Coming soon

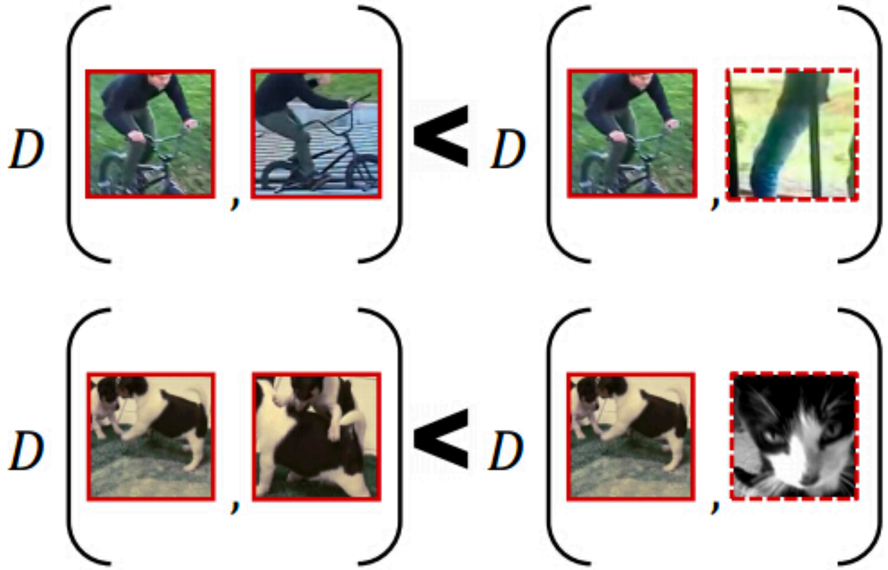
Tracking Objects



(a) Unsupervised Tracking in Videos



(b) Siamese-triplet Network

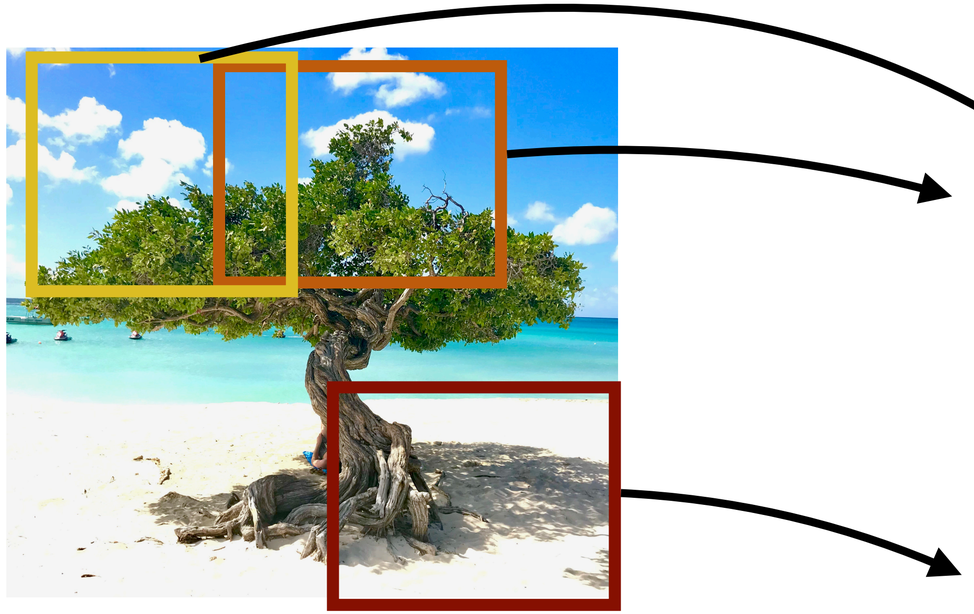


D: Distance in deep feature space

(c) Ranking Objective

Wang & Gupta, 2015, Unsupervised Learning of Visual Representations using Videos

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

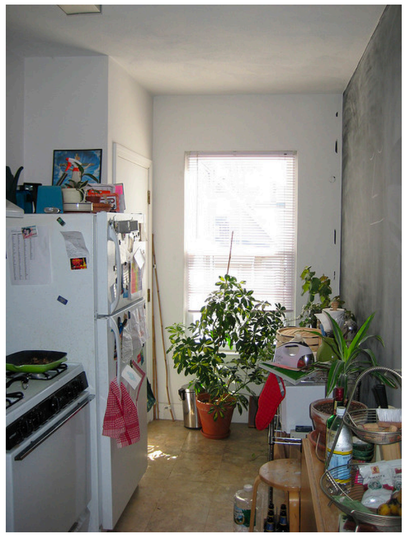
Related
(Positives)

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)

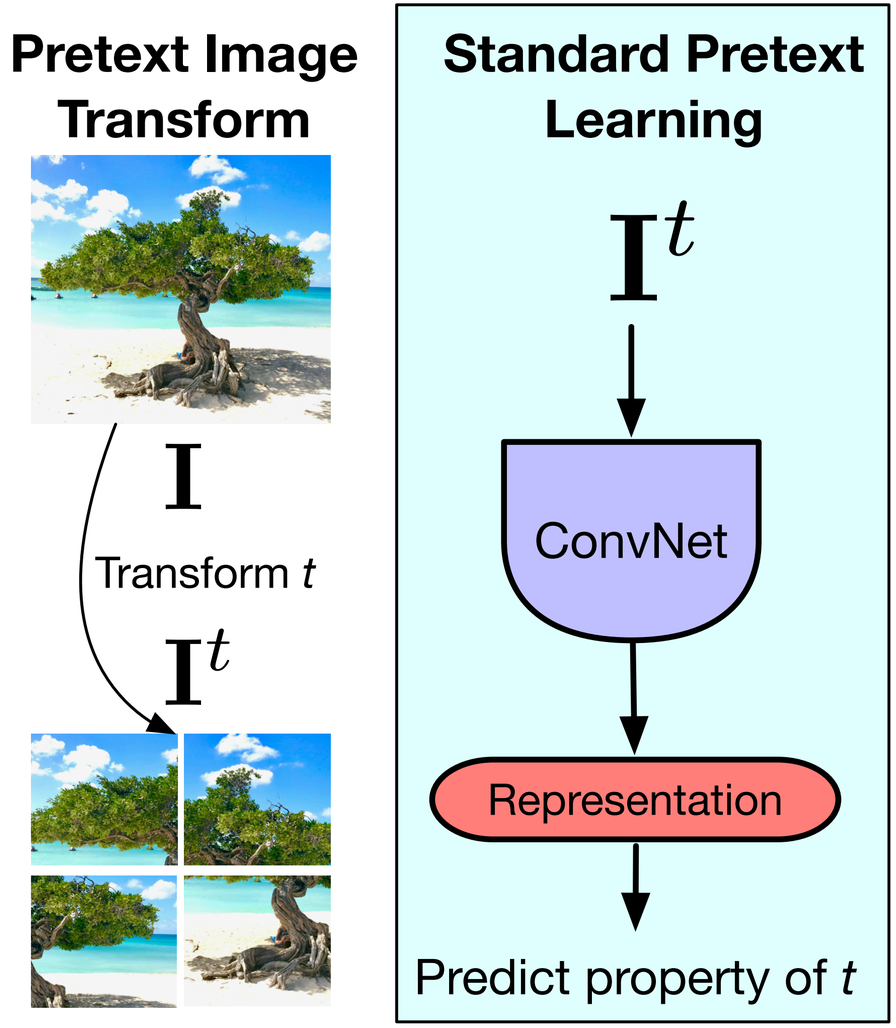


Unrelated
(Negative)

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**

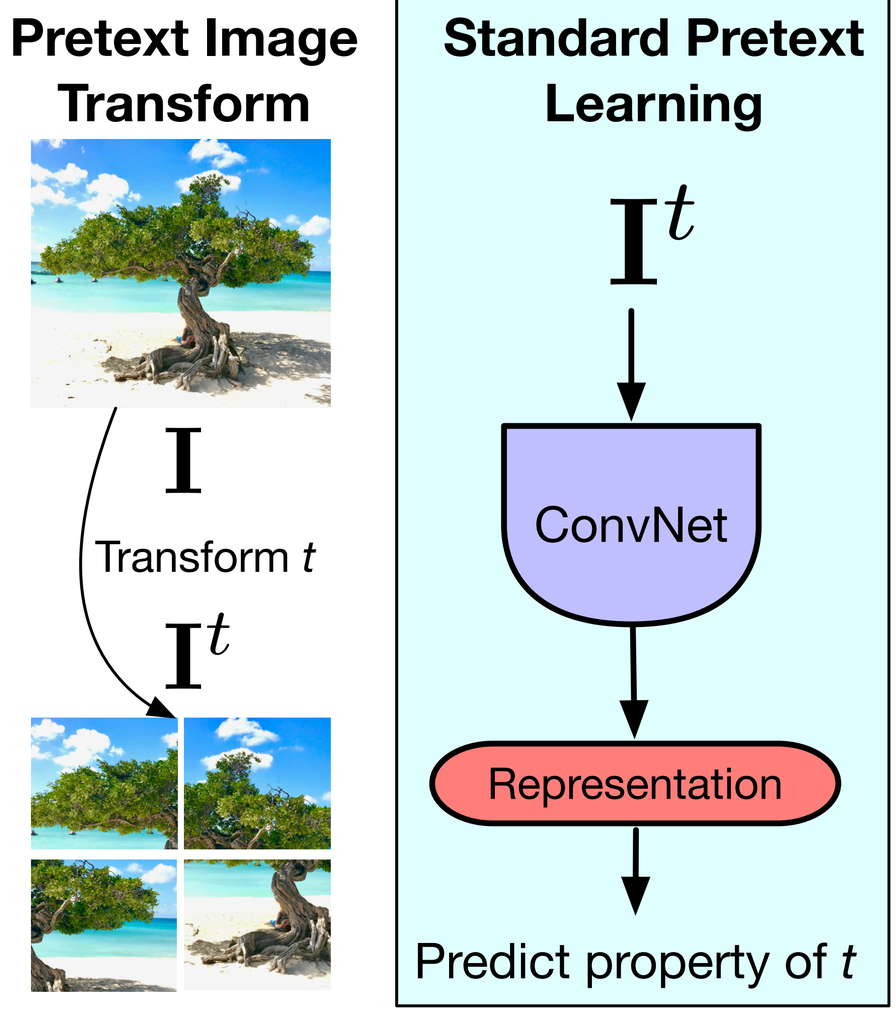


Underlying Principle for Pretext Tasks

- 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?



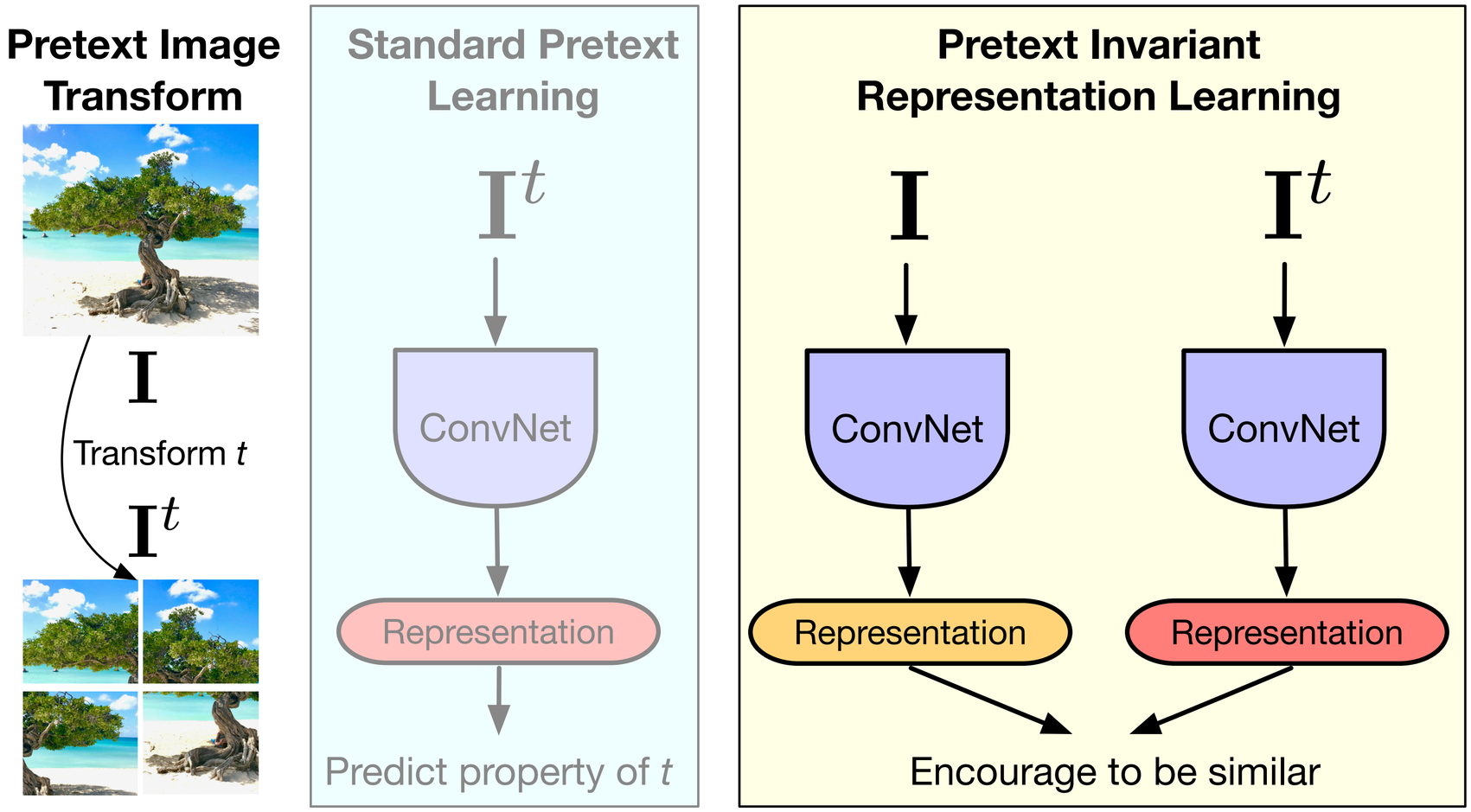
There are many exceptions: e.g., in 3D (Kanazawa et al., 2016; Rocco et al., 2017); Novotny et al., 2018)

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"

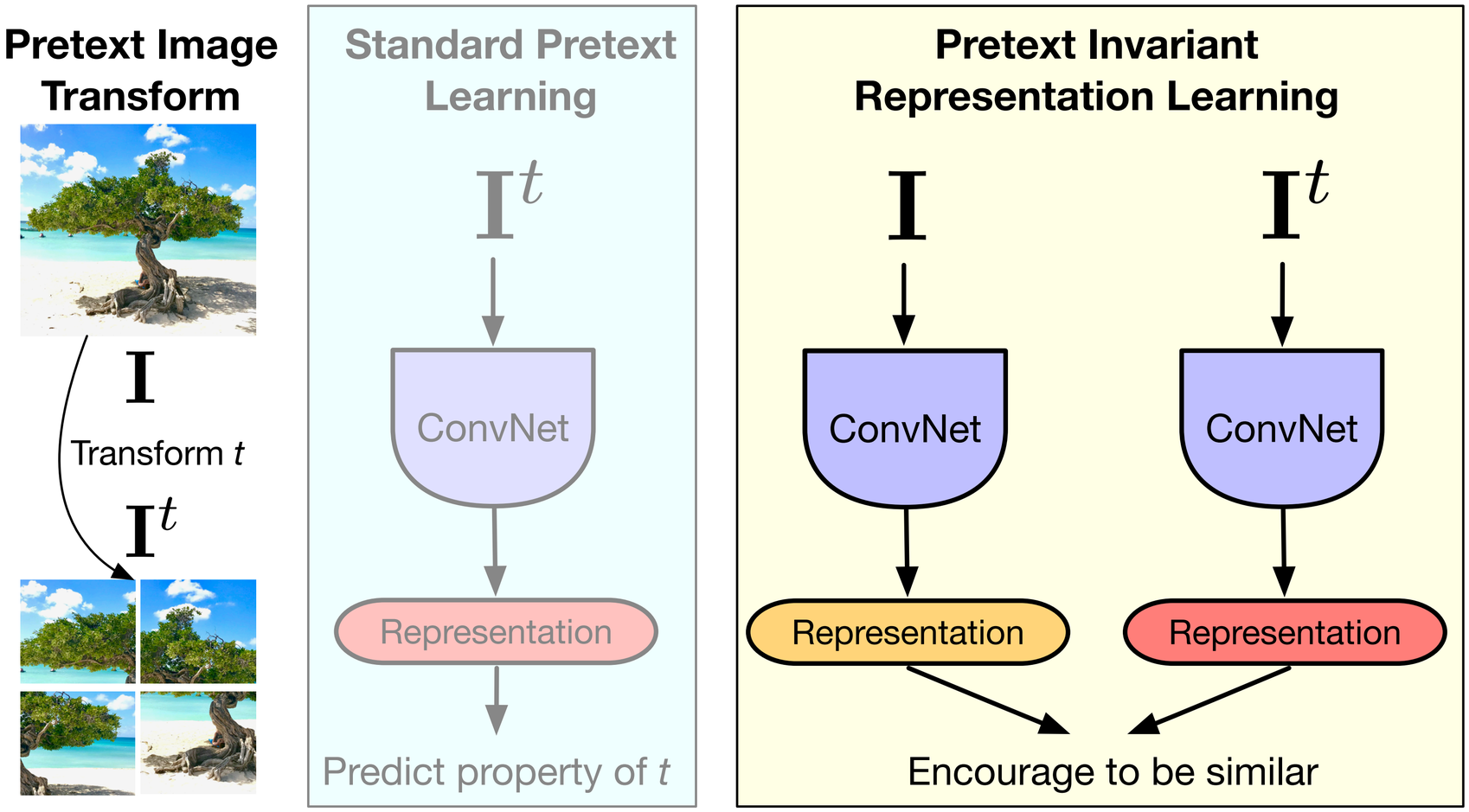
Pretext-Invariant Representation Learning (PIRL)

- Be invariant to **t**



Pretext-Invariant Representation Learning (PIRL)

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

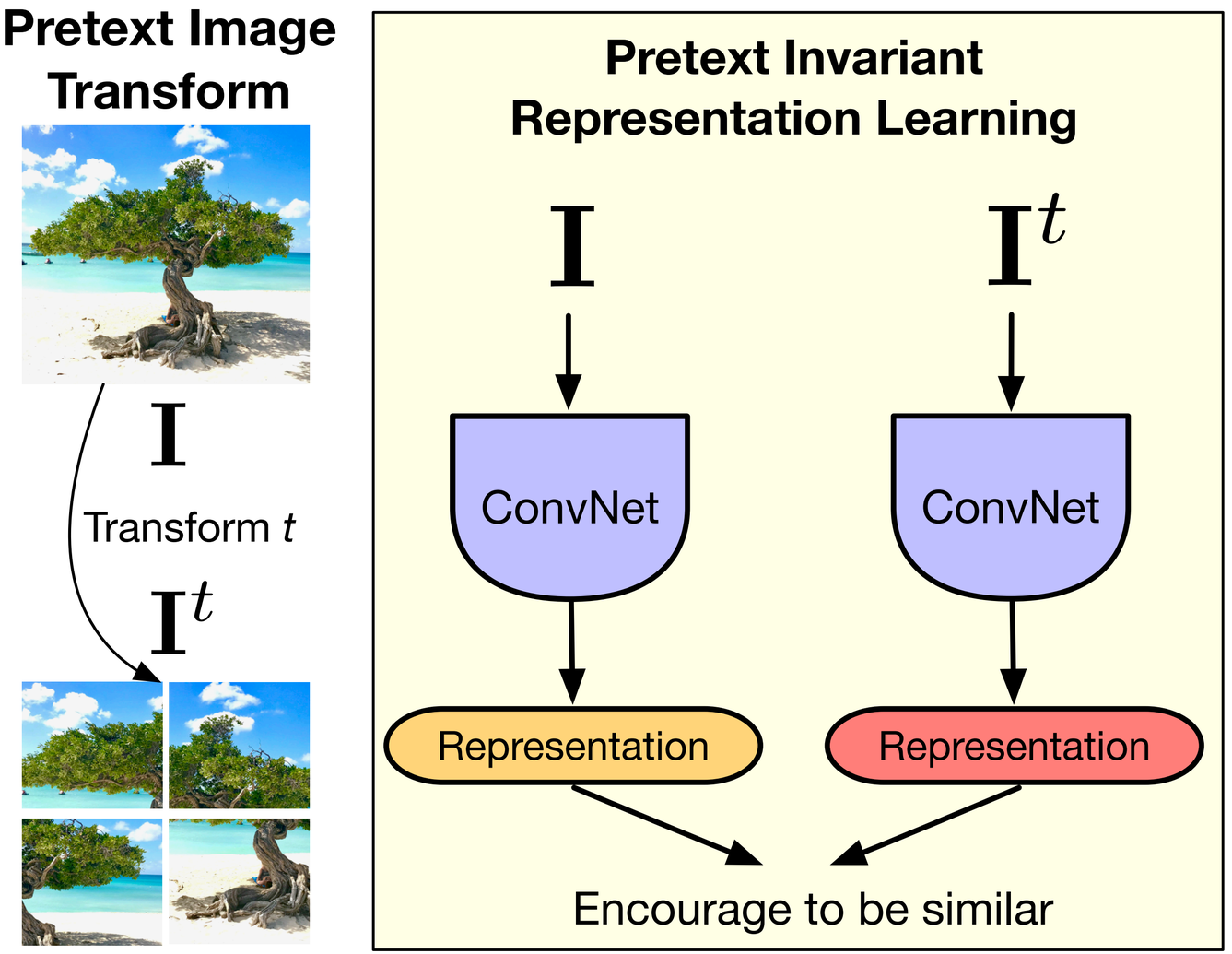


PIRL

- Representations from **I** and **I**^{t} should be similar
- **t** = Pretext Transforms (Jigsaw/ Rotation, combinations etc.)

- Use a contrastive loss to enforce similarity of features

$$L_{\text{contrastive}}(\mathbf{v_I}, \mathbf{v_{I^t}})$$

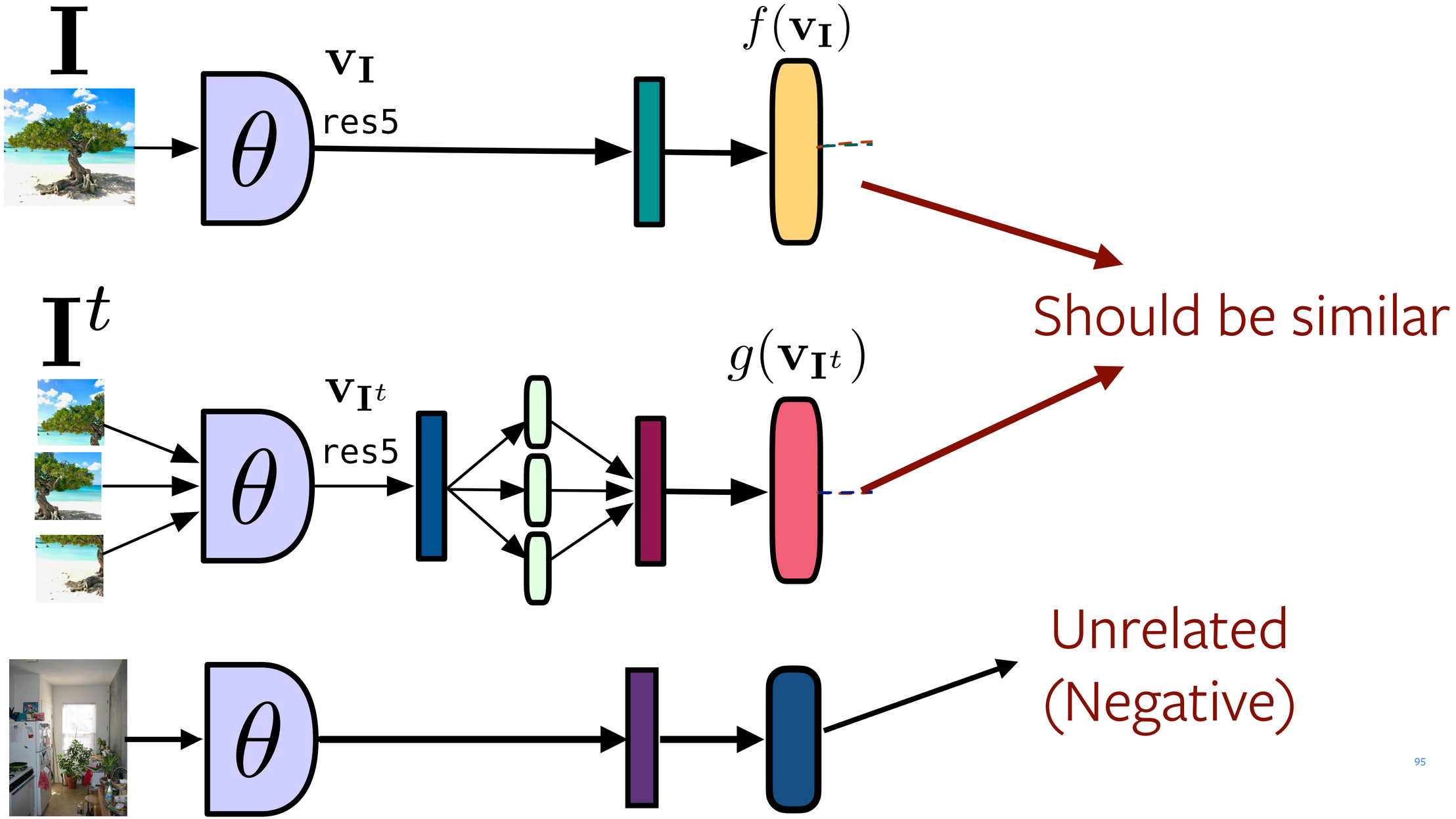


PIRL - Using Large Number of Negatives

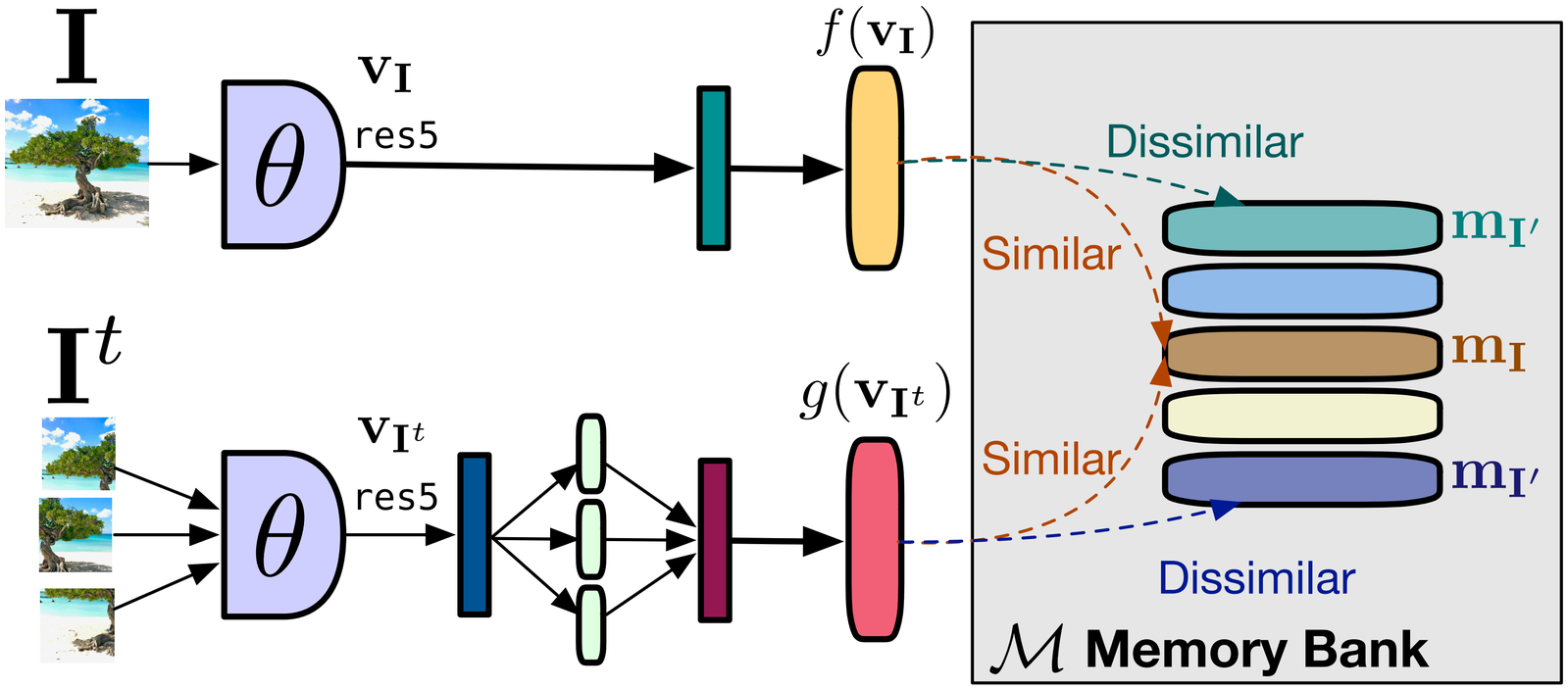
- Large number of negatives is crucial in contrastive learning
- Following Wu et al., PIRL uses a **memory bank** of negatives
- Memory bank stores a feature vector per image in the dataset

Wu et al., 2018, Unsupervised Feature Learning via Non-Parametric Instance Discrimination

How it works


$\mathbf{I}$
θ
$\mathbf{v_I}$
res5
$f(\mathbf{v_I})$
Should be similar
$\mathbf{I}^t$
θ
$\mathbf{v_{I^t}}$
res5
$g(\mathbf{v_{I^t}})$
θ
Unrelated
(Negative)

How it works



$$L_{\text{NCE}}(g(\mathbf{v_{I^t}}), \mathbf{m_I}) + L_{\text{NCE}}(f(\mathbf{v_I}), \mathbf{m_I})$$

Rep. $\mathbf{I^t}$ and $\mathbf{m_I}$ should be similar

Rep. $\mathbf{I}$ and $\mathbf{m_I}$ should be similar

NCE - Gutmann et al., 2010

PIRL Pre-training

- Pre-trained on images without labels
  - ImageNet-1M
  - YFCC ("in-the-wild"/uncurated images)

Evaluation

- Evaluate initialization: Full fine-tuning
- Evaluate features: Train linear classifiers on fixed features

Object Detection

- **Outperforms** ImageNet supervised pre-trained networks
- Full fine-tuning, no bells & whistles
- No extra data, changes in model architecture, fine-tuning schedule

Initialization	VOC07+12			VOC07		
	AP^{all}	**AP^{50}**	**AP^{75}**	**AP^{all}**	**AP^{50}**	**AP^{75}**
ImageNet Supervised	52.6	**81.1**	57.4	43.8	**74.5**	45.9
PIRL	**54.0** (+1.4)	80.7	**59.7** (+2.3)	**44.7**	73.4	**47.0** (+1.1)

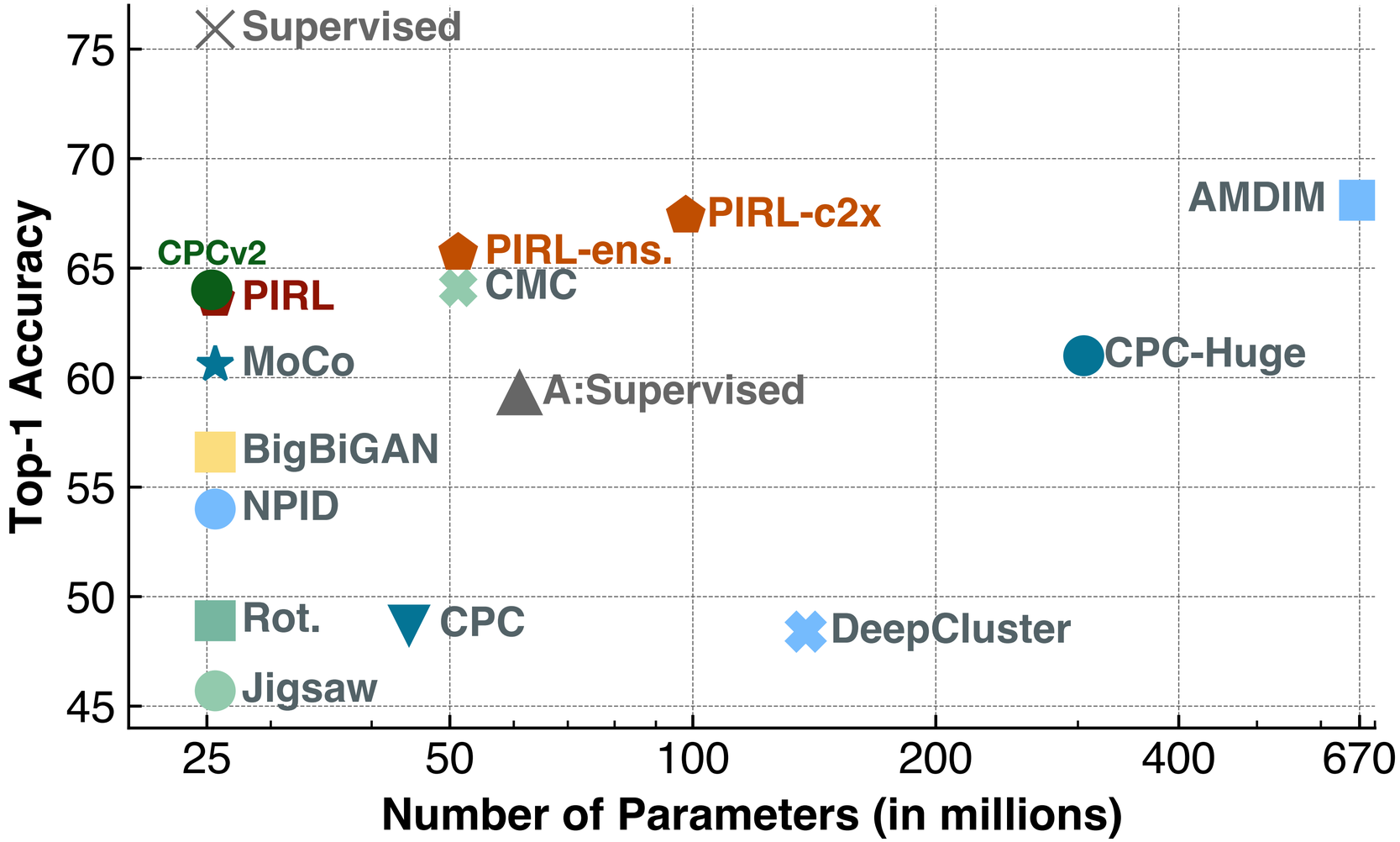
Semi-supervised Learning

- 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**

Linear Classification

- Linear classifiers on fixed features. Evaluate on **ImageNet-1K**

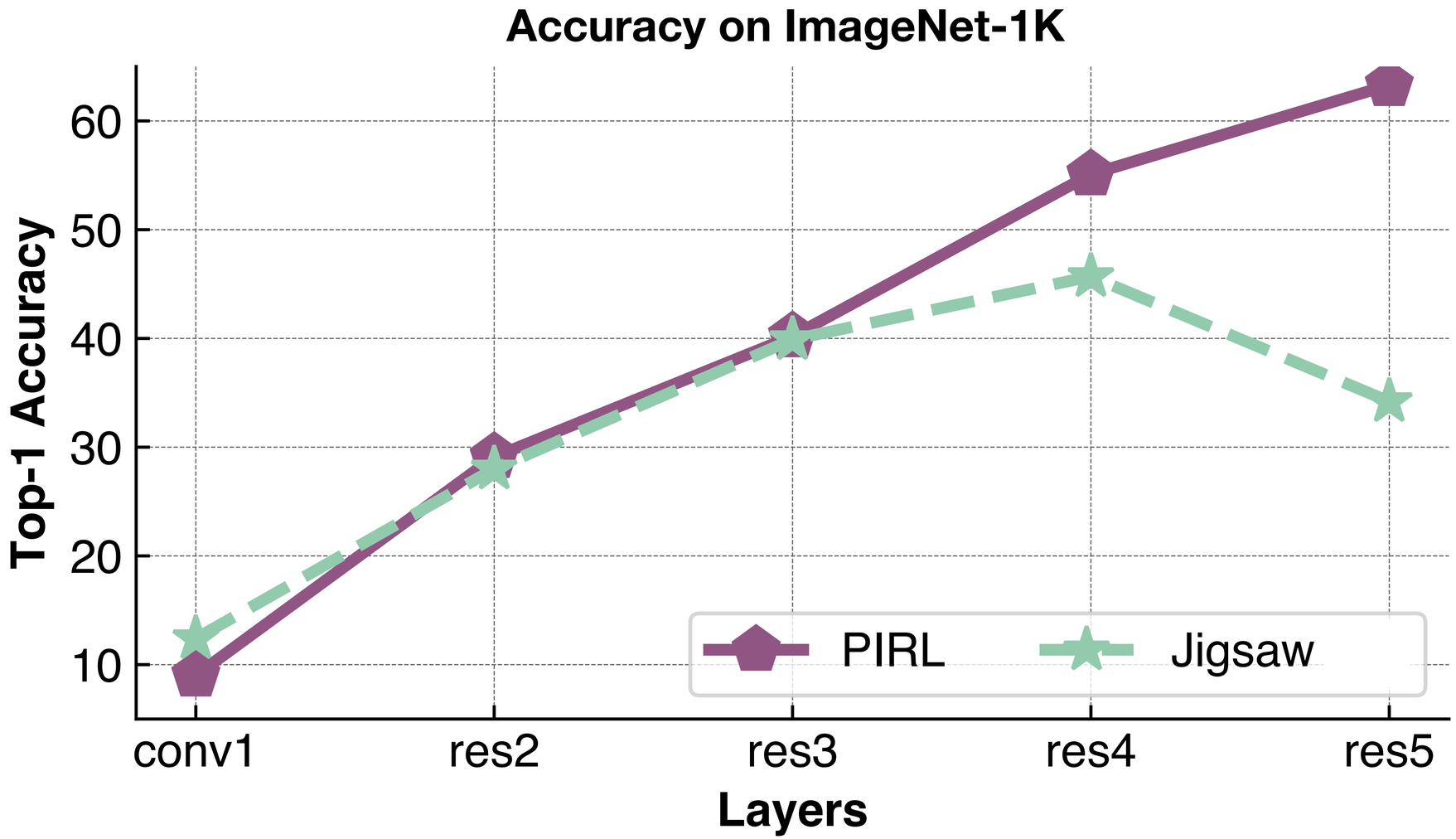


"In-the-wild" Flickr images

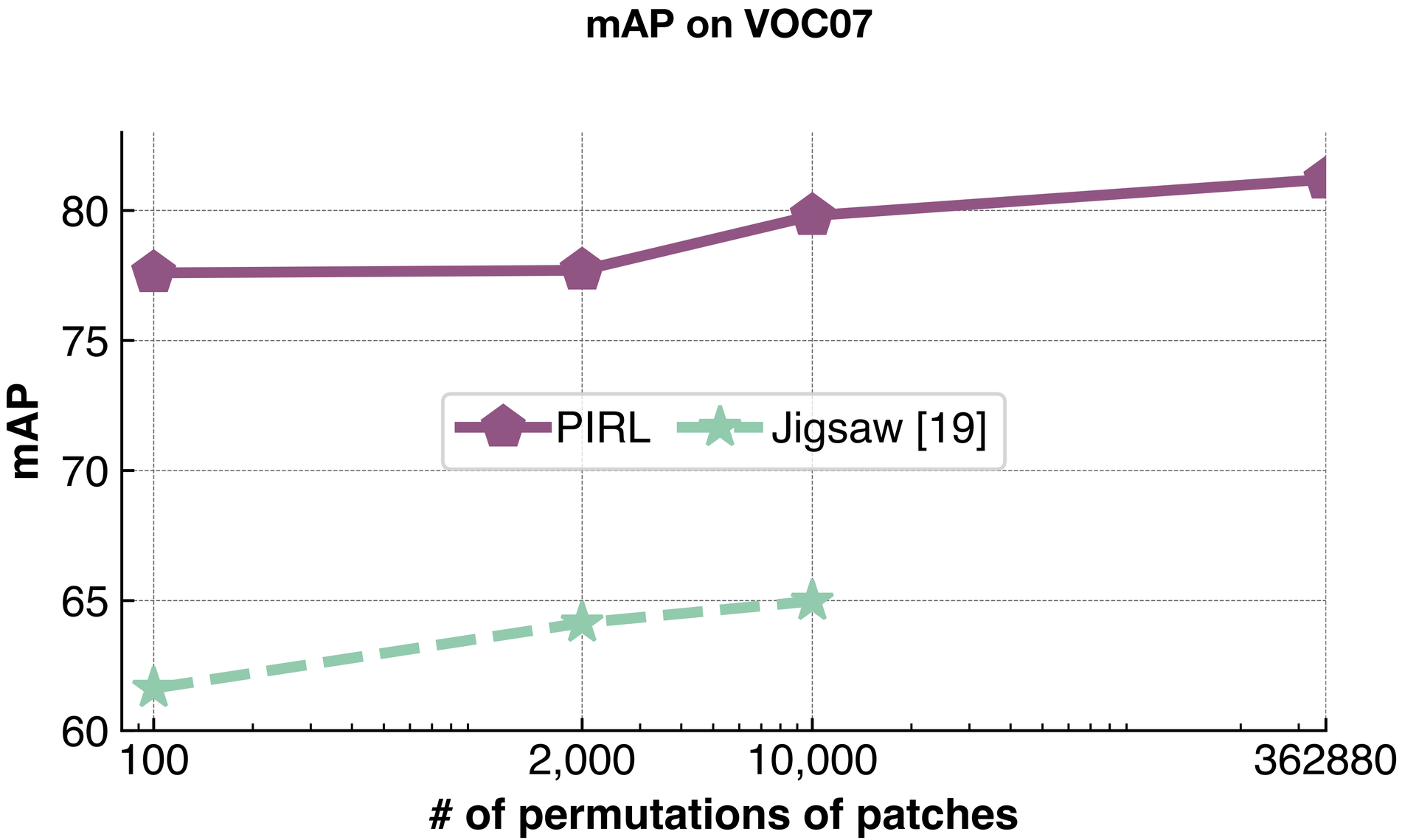
- Yahoo Flickr Creative Commons (YFCC) images. No labels.
- Linear classifiers on fixed features

Method	# Pretrain Images	Evaluation	
		ImageNet	Places-205
DeeperCluster (Caron et al., 2019)	100M	45.6	42.1
Jigsaw (Goyal et al., 2019)	100M	48.3	44.8
PIRL	1M $\frac{1}{100}$	**57.8** +9	**51.0** +6

Semantic Features?



Easily scale to large number of permutations



Easily Extend to other tasks

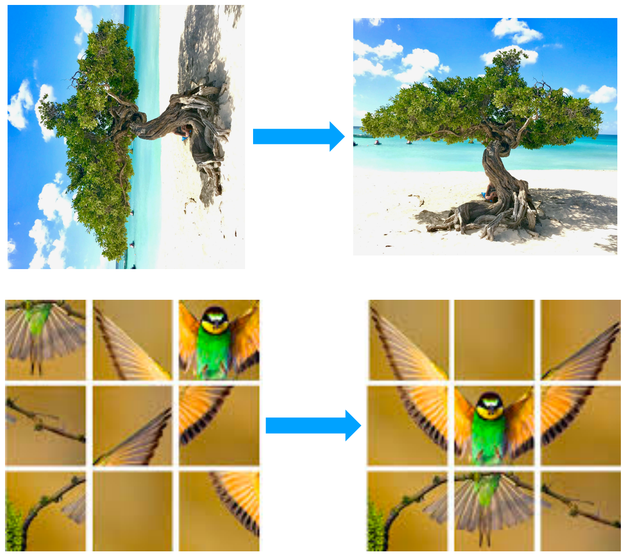
Method	Transfer Dataset			
	ImageNet-1M	VOC07	Places205	iNaturalist
Jigsaw	46.0	66.1	41.4	22.1
Rotation	48.9	63.9	47.6	23
PIRL (Rot)	**60.2**	**77.1**	**47.6**	**31.2**

Easily Multi-task

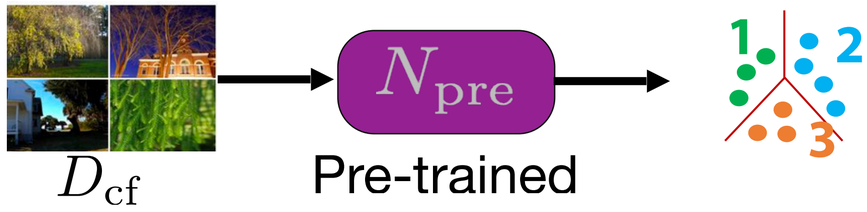
Method	Transfer Dataset			
	ImageNet-1M	VOC07	Places205	iNaturalist
Jigsaw	46.0	66.1	41.4	22.1
Rotation	48.9	63.9	47.6	23
PIRL (Rot)	60.2	77.1	47.6	31.2
PIRL (Jigsaw + Rot)	**63.1**	**80.3**	**49.7**	**33.6**

More invariance

Pretext tasks

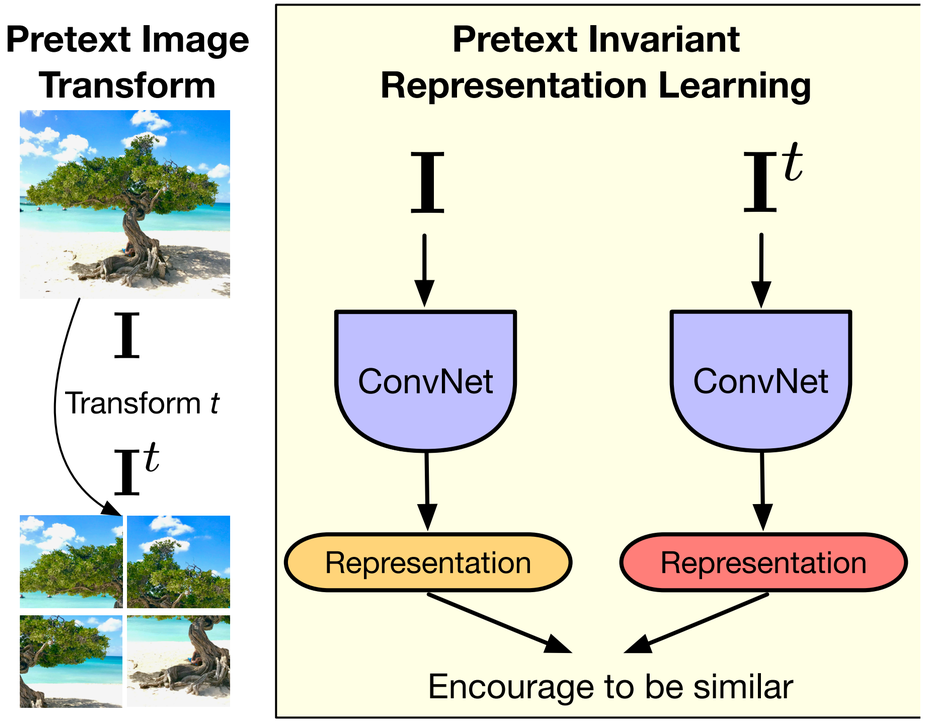


Clustering



D_{cf}

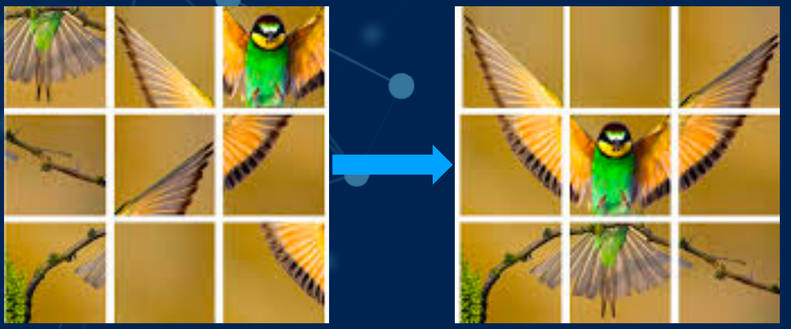
PIRL



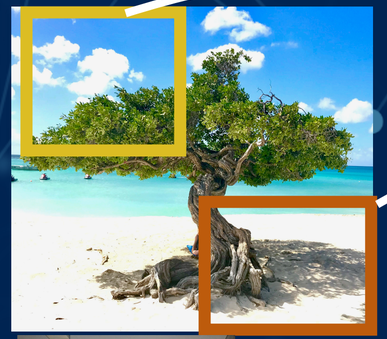
Shortcomings

- Set of data transforms matters a lot
- Saturation with model size and data size
- What invariances matter?

Pretext tasks

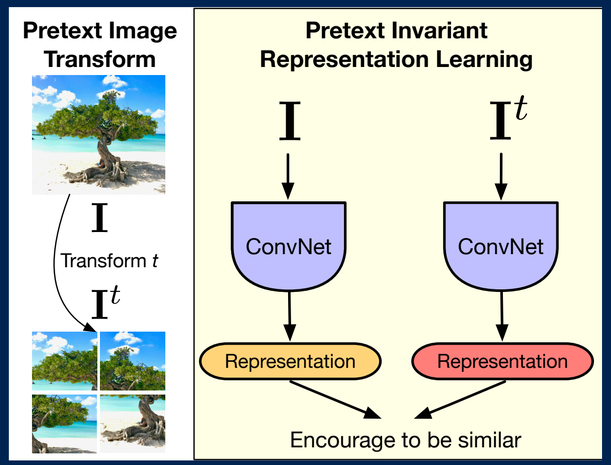


Contrastive

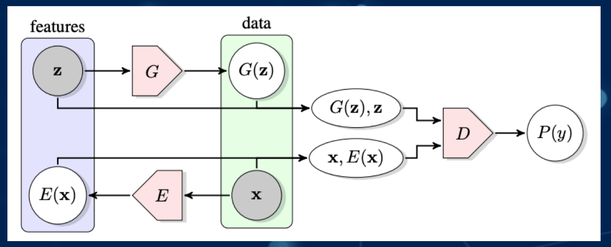


Related

Unrelated



Generative



AutoEncoder,
VAE, GAN,
BiGAN

Predict more information