Self-supervised learning in computer vision

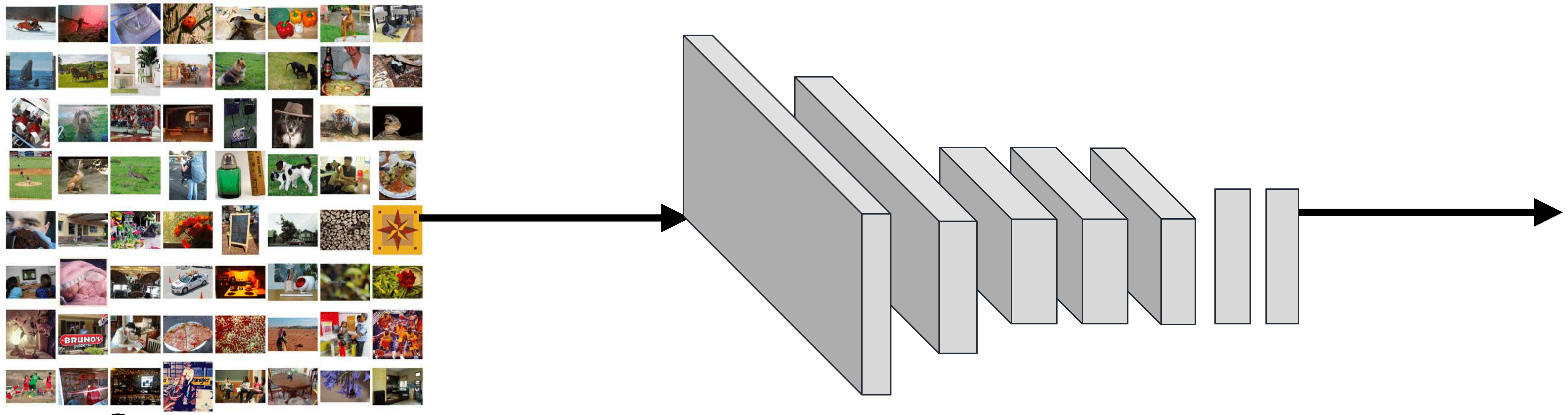
Ishan Misra

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



husky, terrier, tench, ...

Images from ImageNet (Pre-train)

ConvNet

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

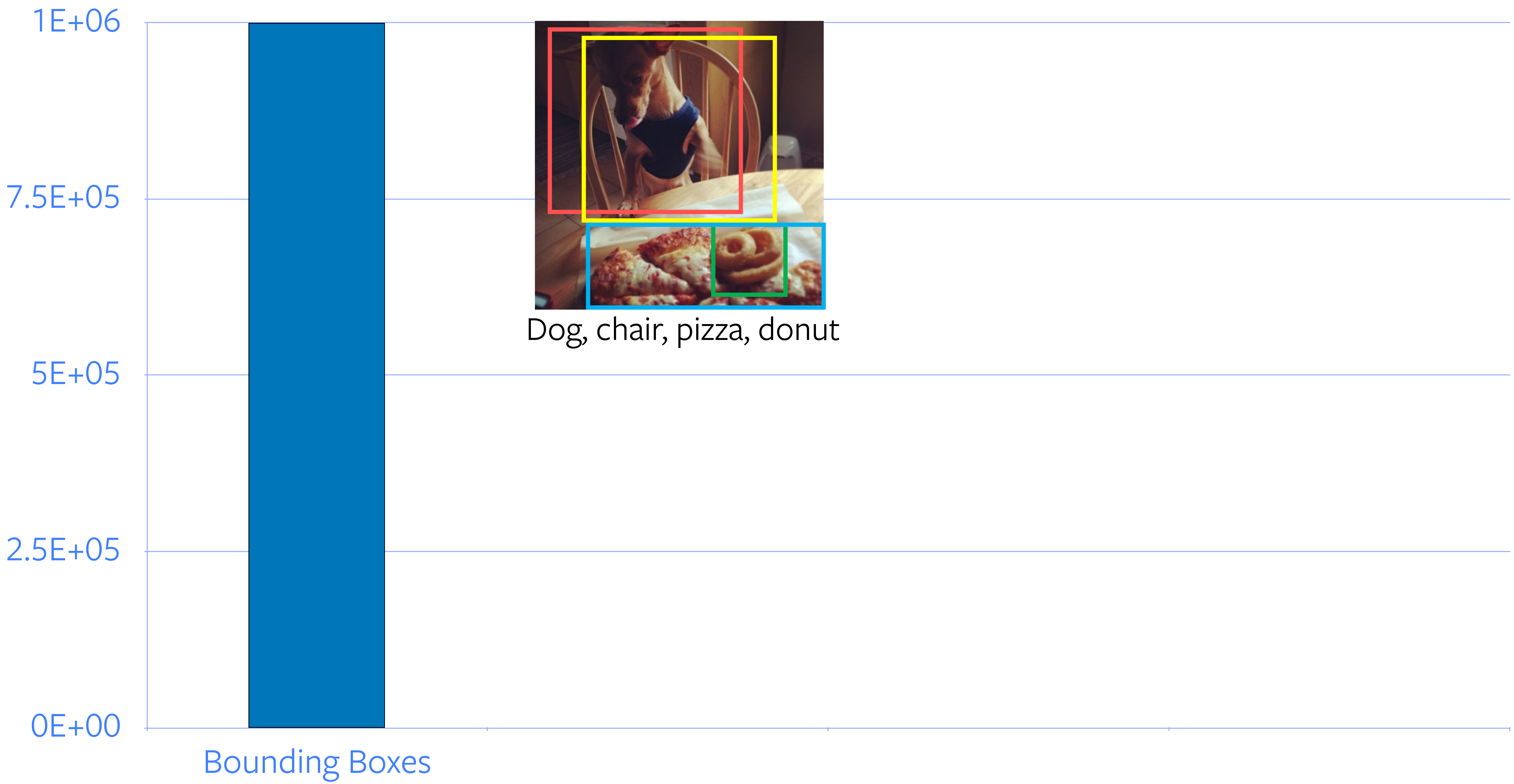
The promise of "alternative" supervision

- 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

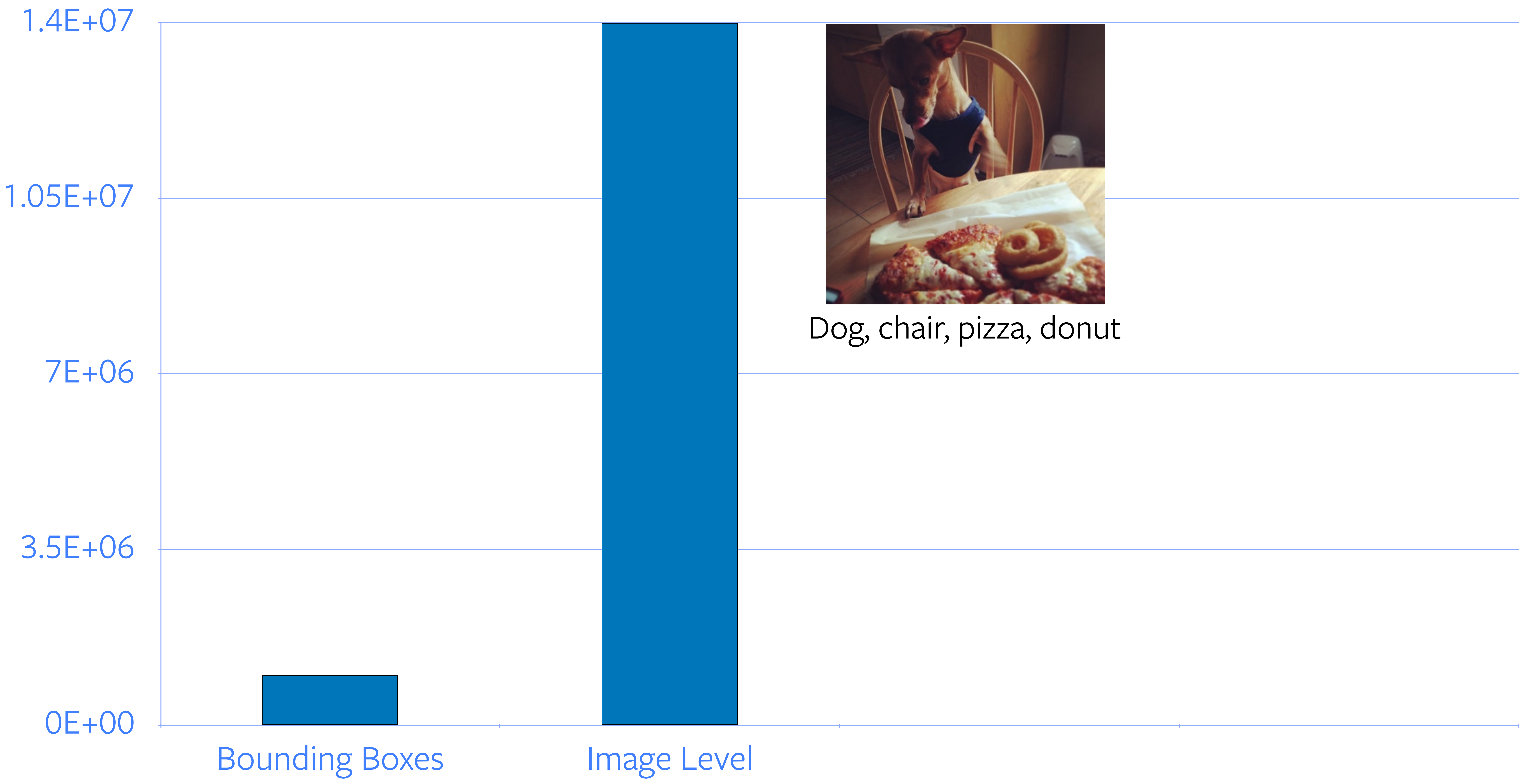
Can we get labels for all data?

Can we get labels for all data?



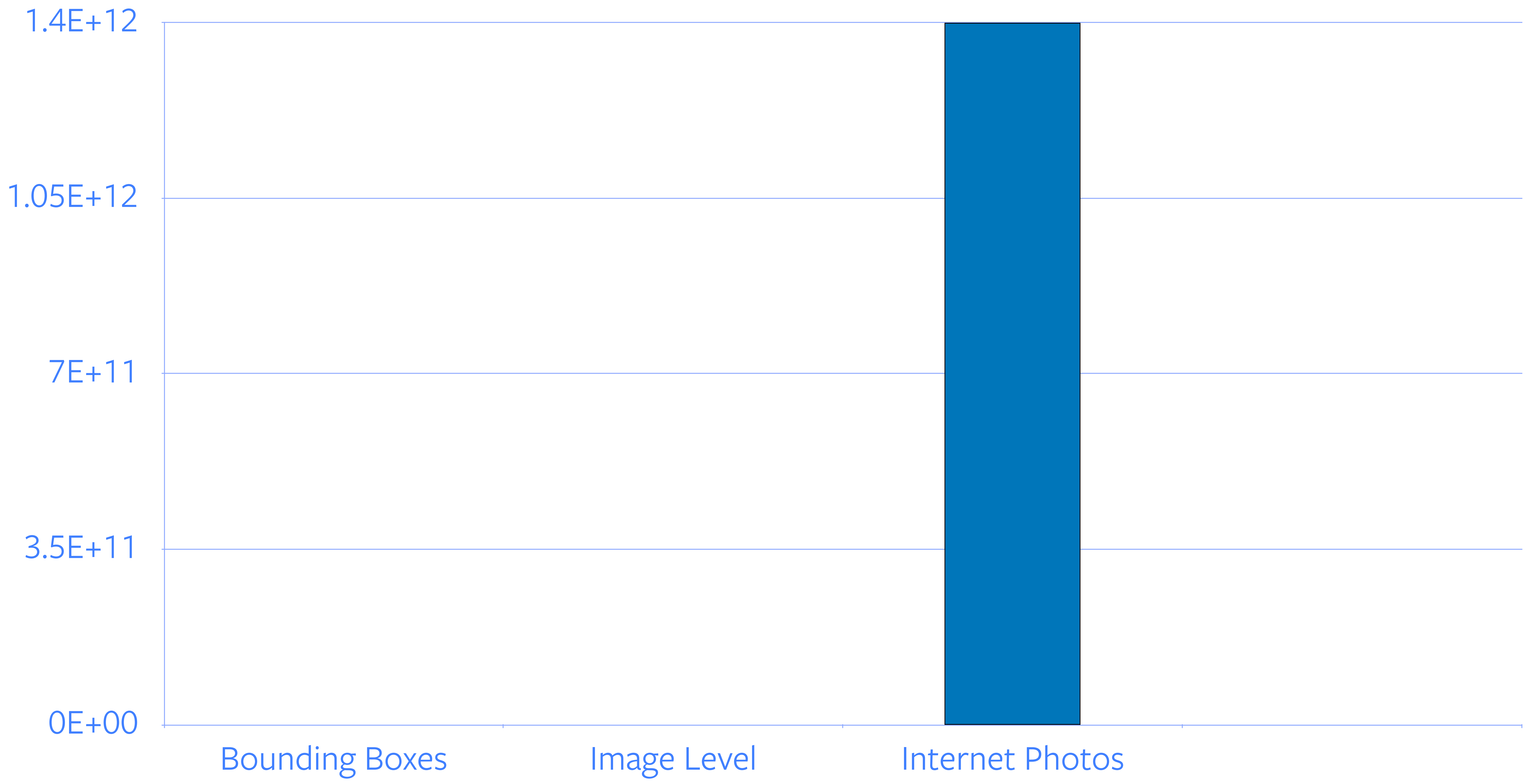
Stats from Pawan Kumar at Oxford

Can we get labels for all data?



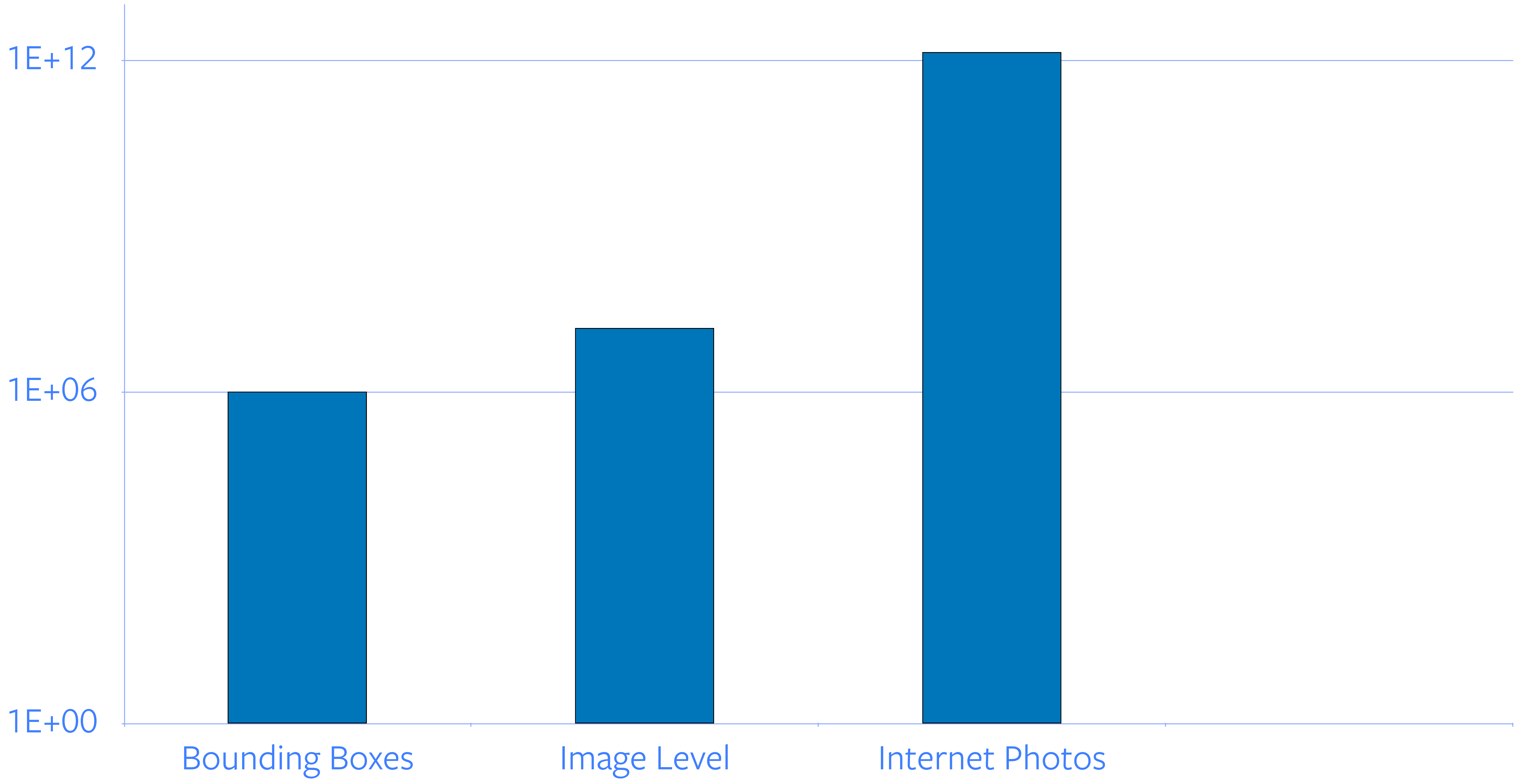
Dog, chair, pizza, donut

Can we get labels for all data?

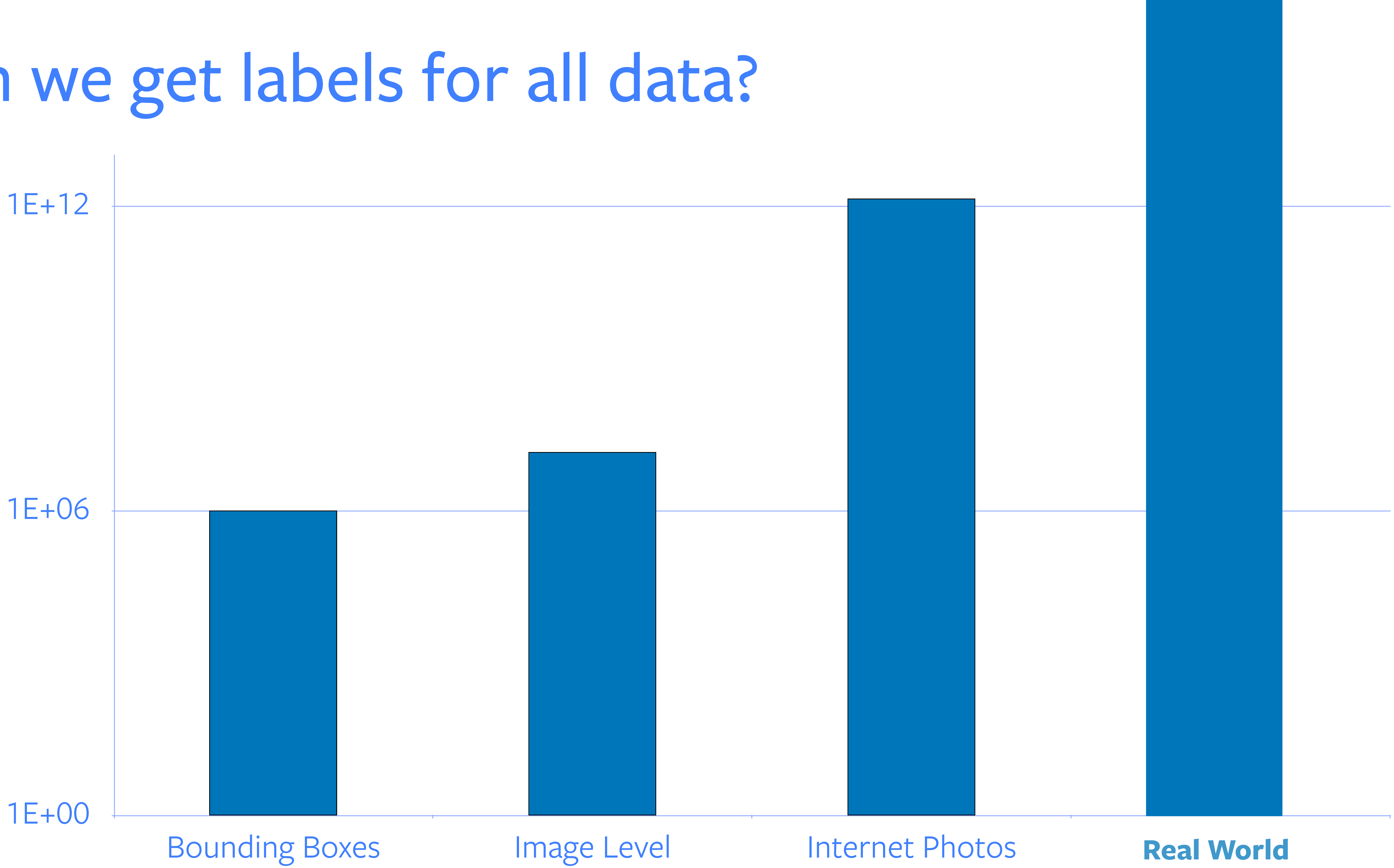


[forbes.com](https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/)
https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

Can we get labels for all data?



Can we get labels for all data?

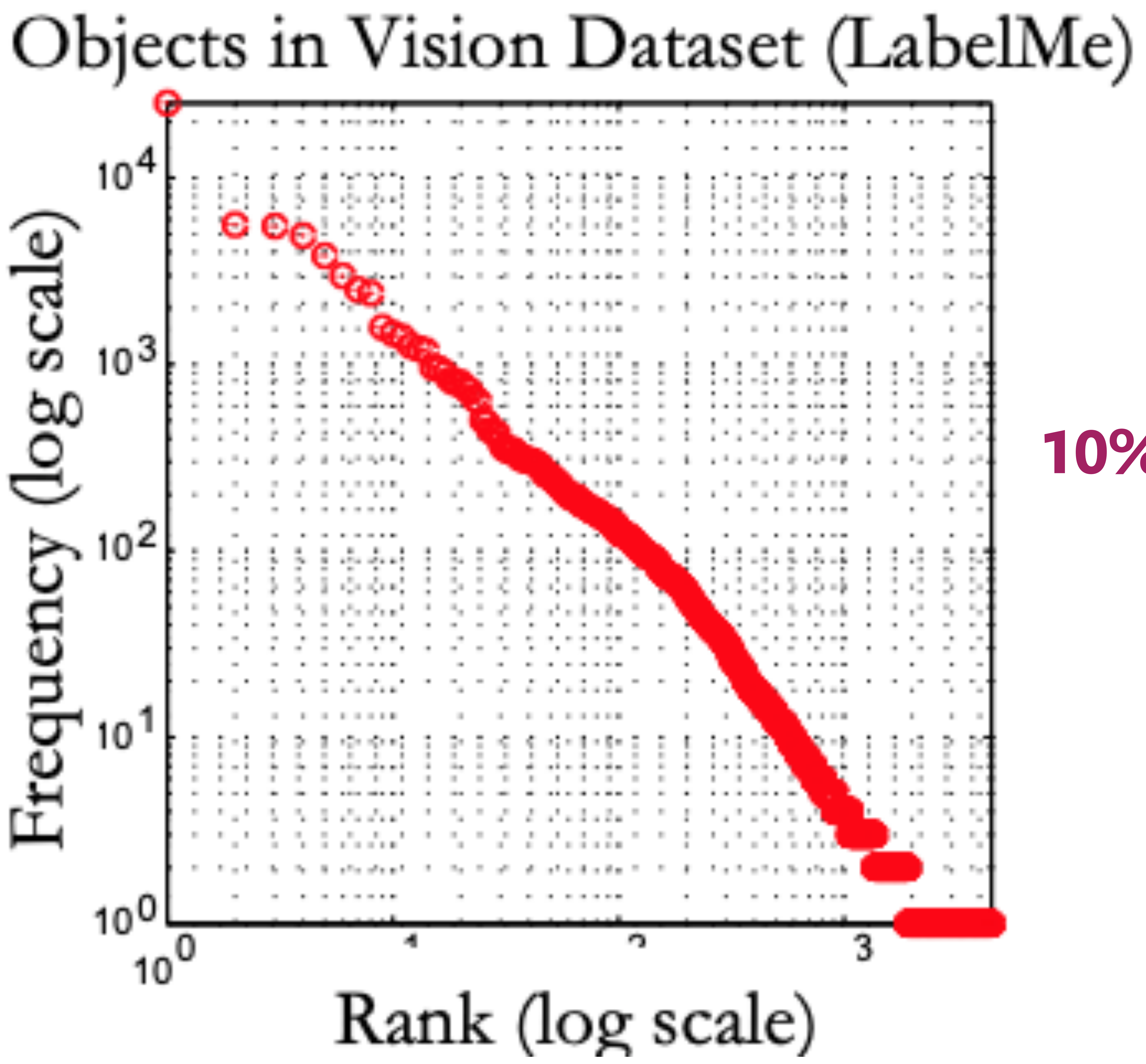


ImageNet (14 million images) needed 22 human years to label

Can we get labels for all data?

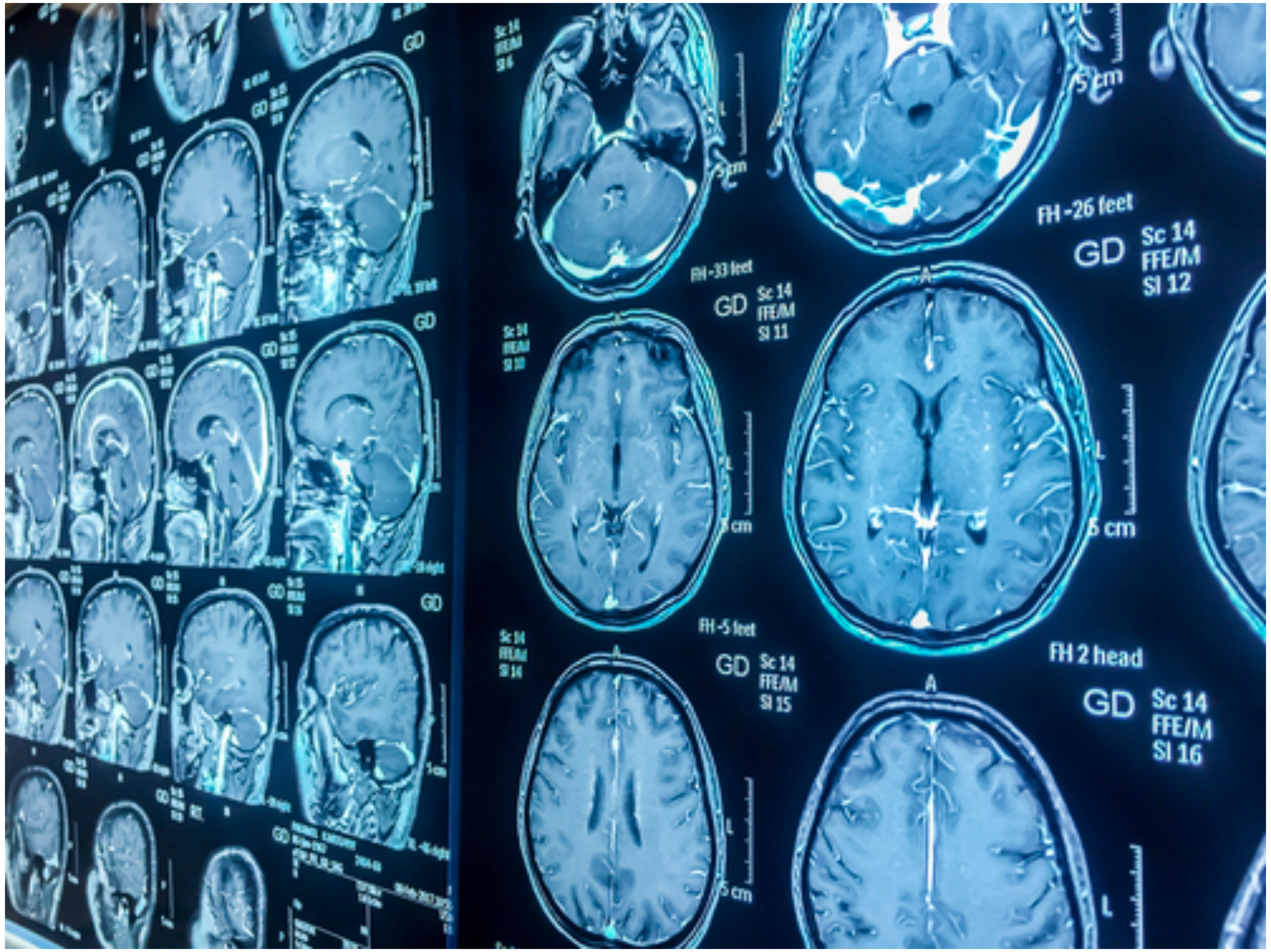
- What about complex concepts?
  - Video?
- Labelling cannot scale to the size of the data we generate

Rare concepts?



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

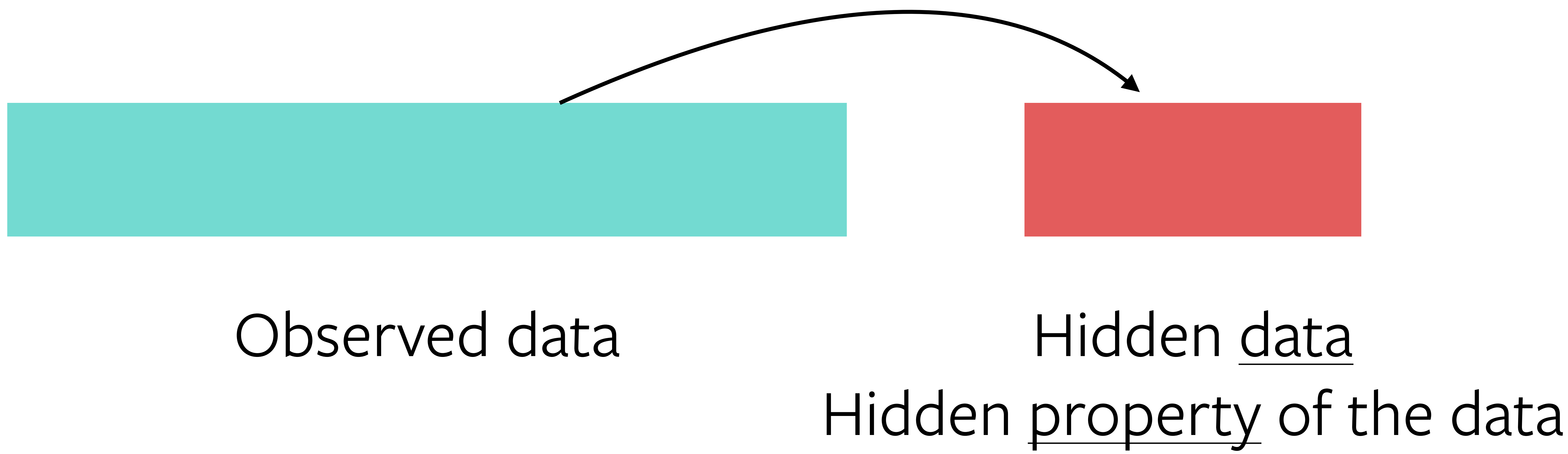
Different Domains?



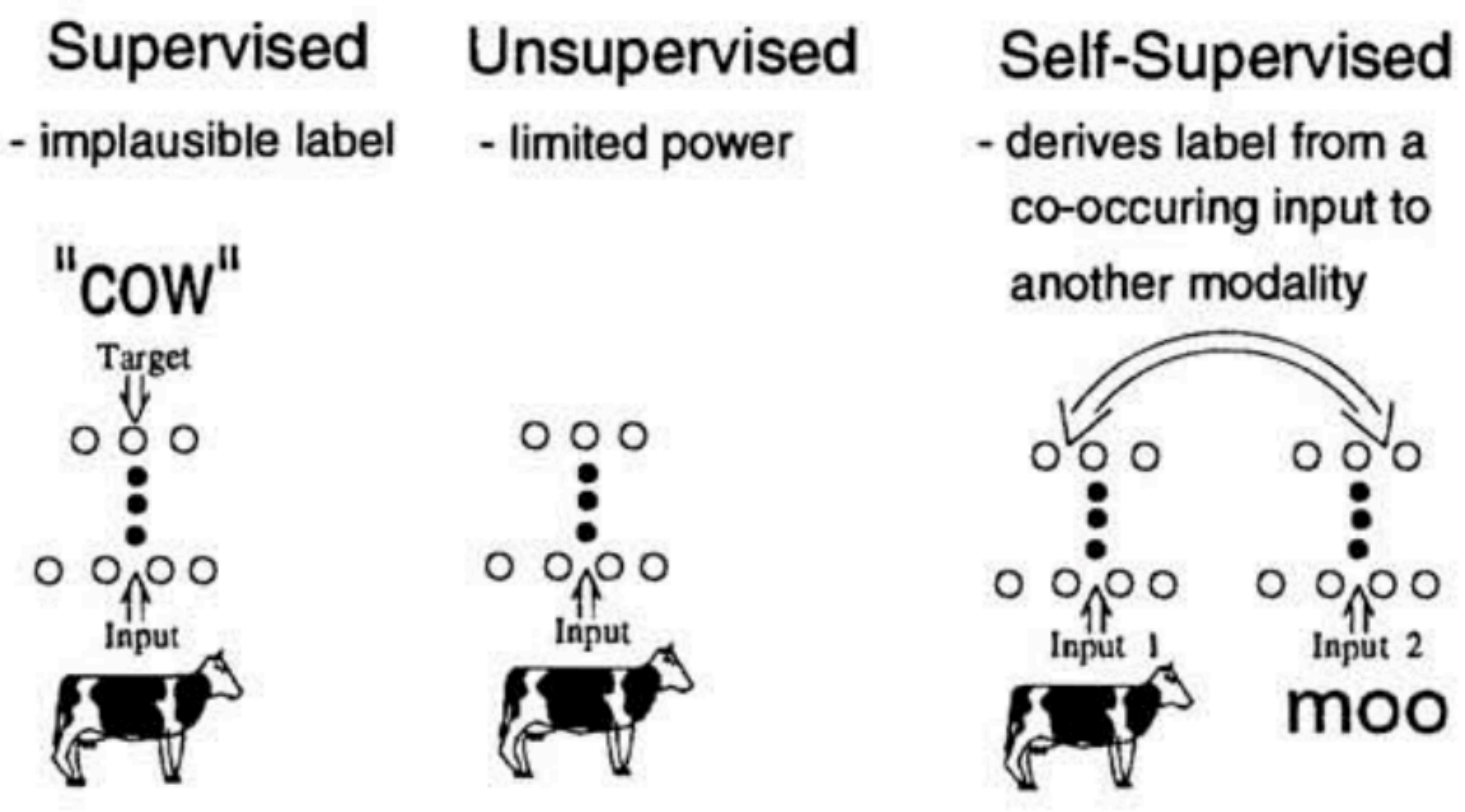
ImageNet pre-training may not work

What is "self" supervision?

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



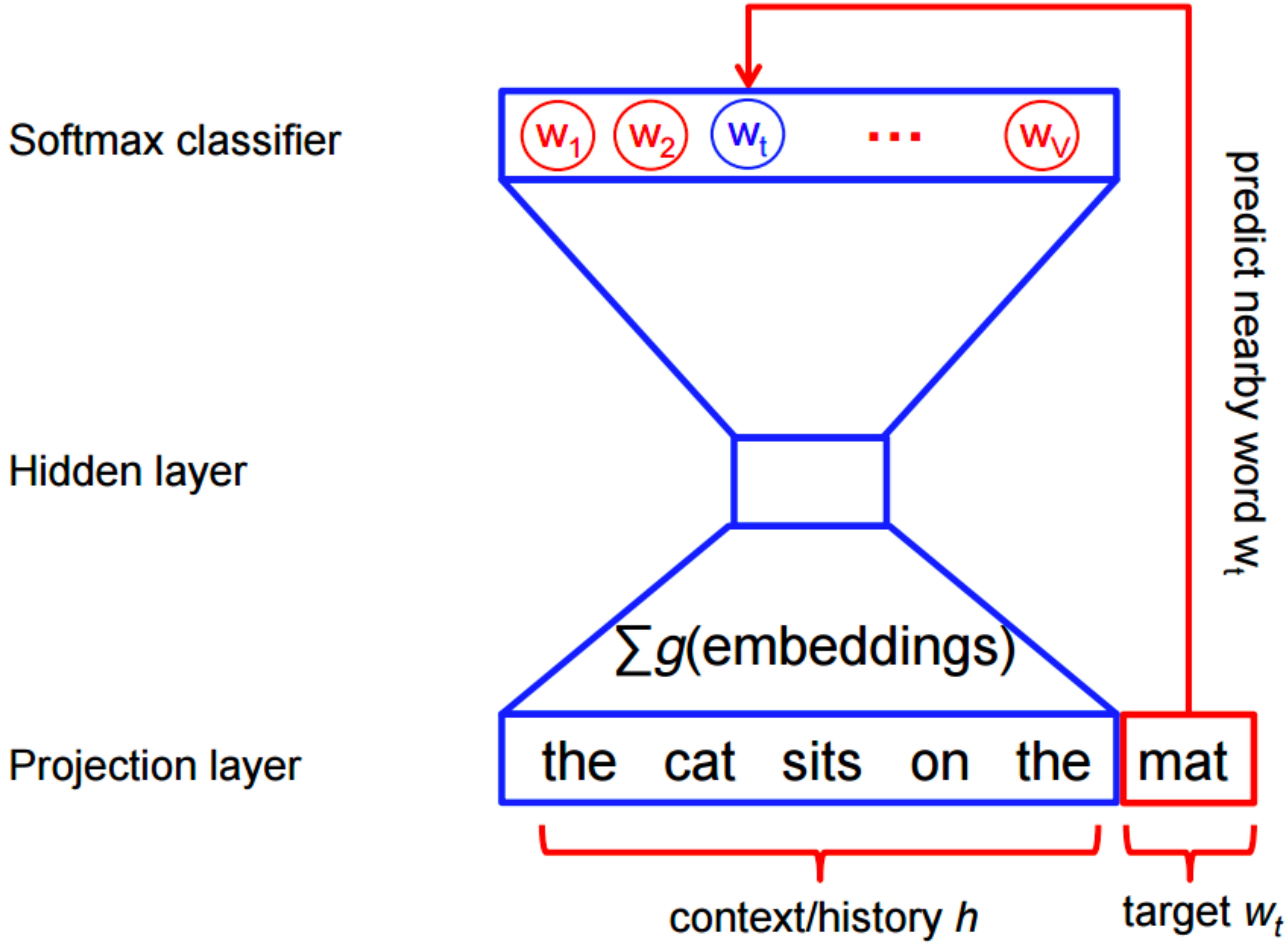
What is "self" supervision?



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?

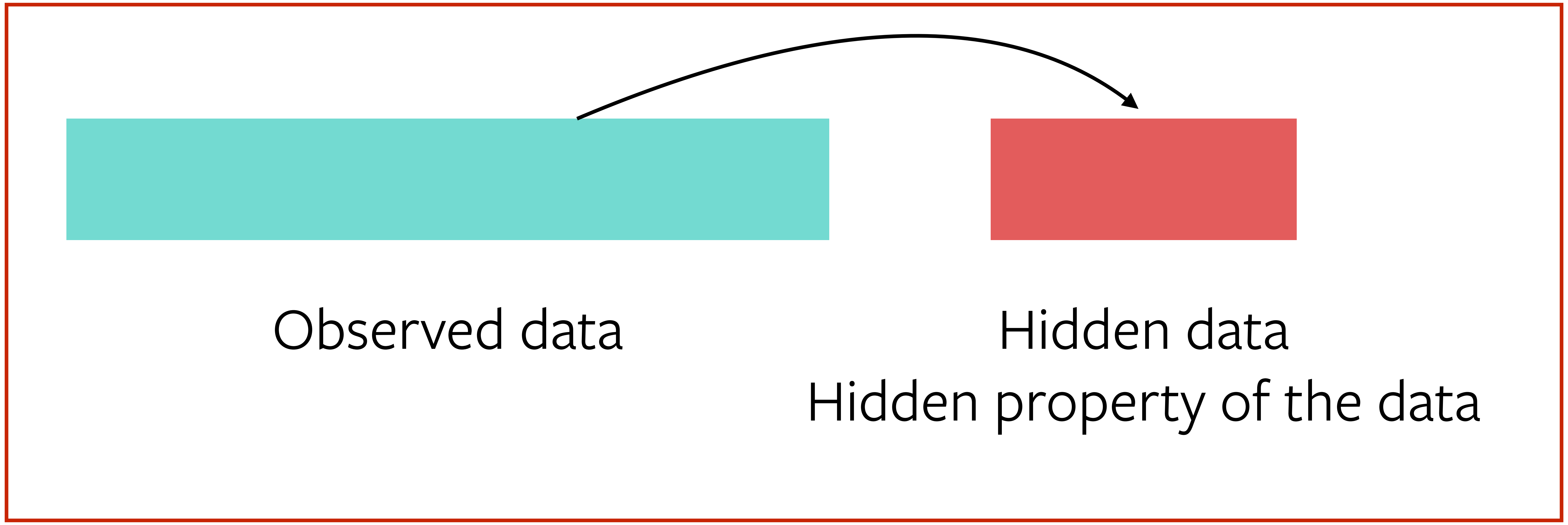
- Helps us learn using observations and interactions
- Does not require exhaustive annotation of concepts
- Leverage multiple modalities or structure in the domain

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

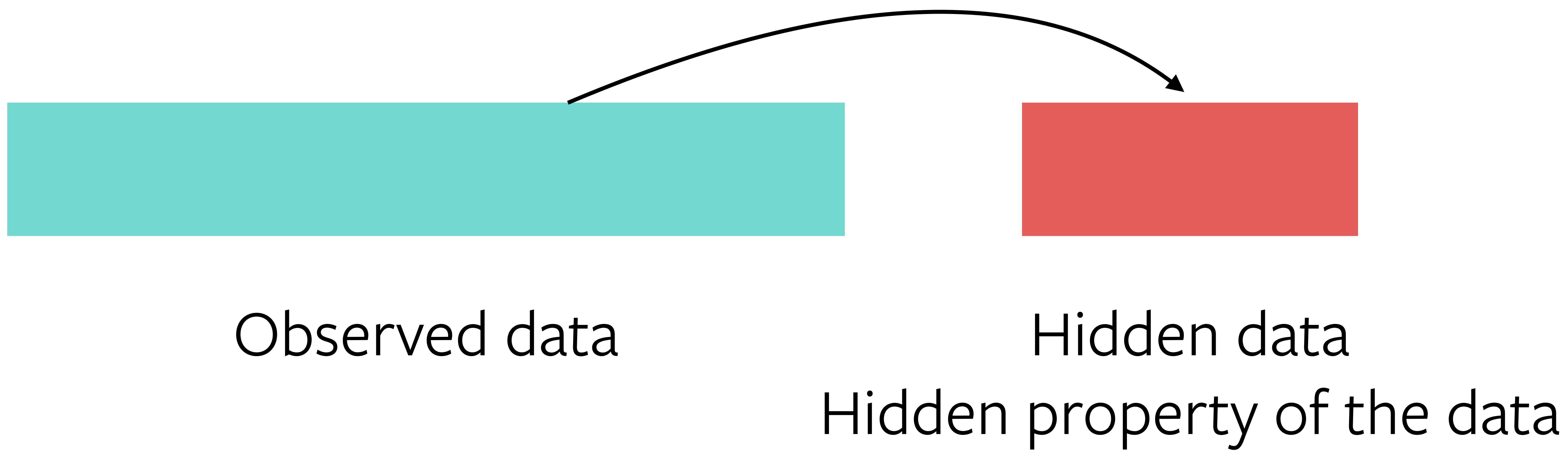
Pretext task



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

Pretext task

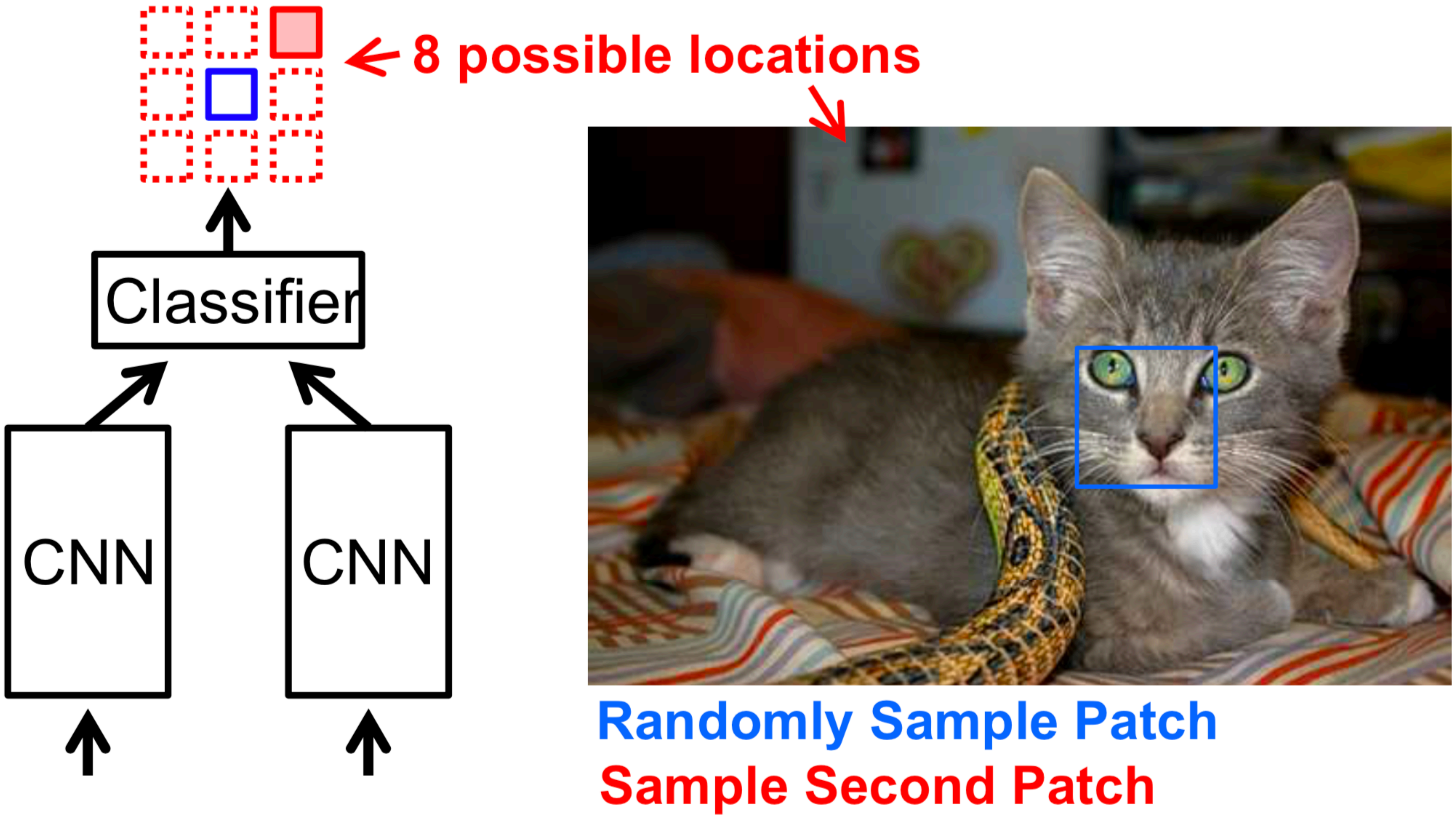
- Using images
- Using video
- Using video and sound



Pretext task

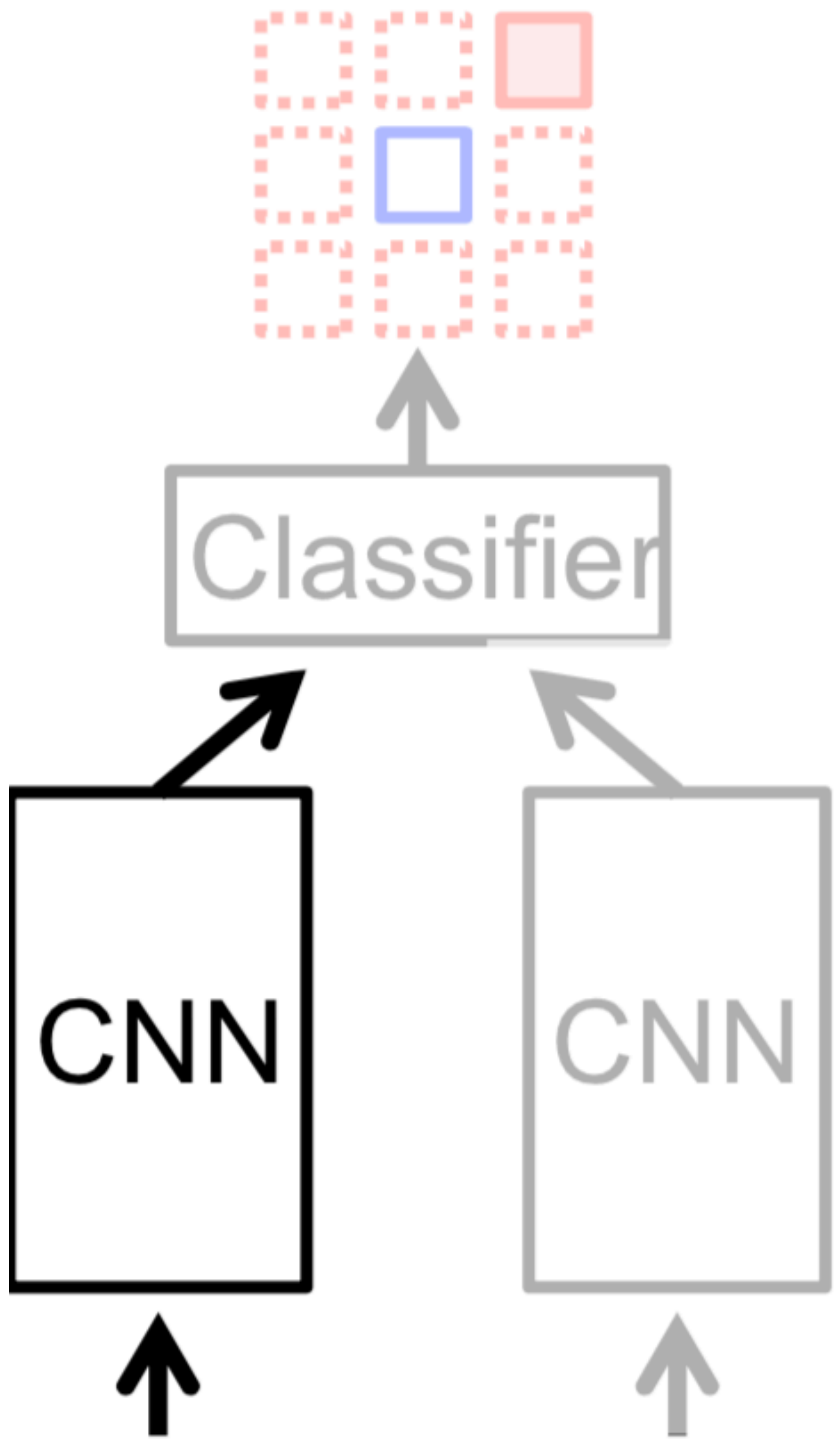
- Using images
- Using video
- Using video and sound

Relative Position of patches



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

Relative Position: Nearest Neighbors in features



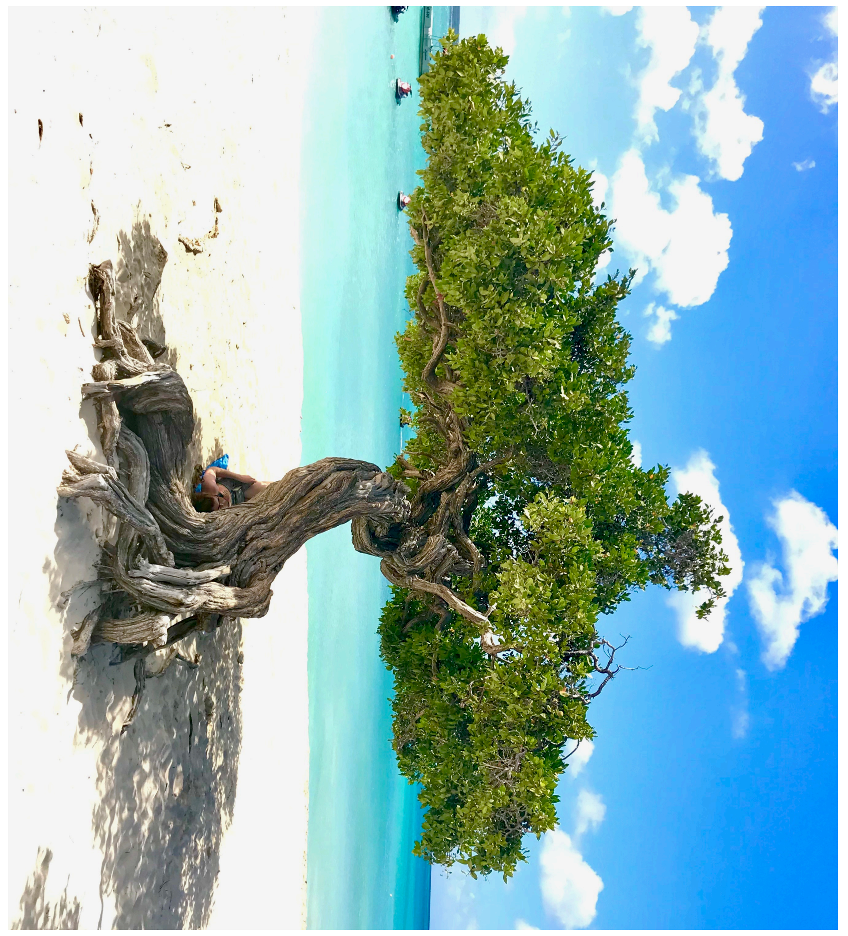
Doersch et al., 2015, Unsupervised visual representation learning by context prediction

Predicting Rotations

→ 0^0

→ 90^0

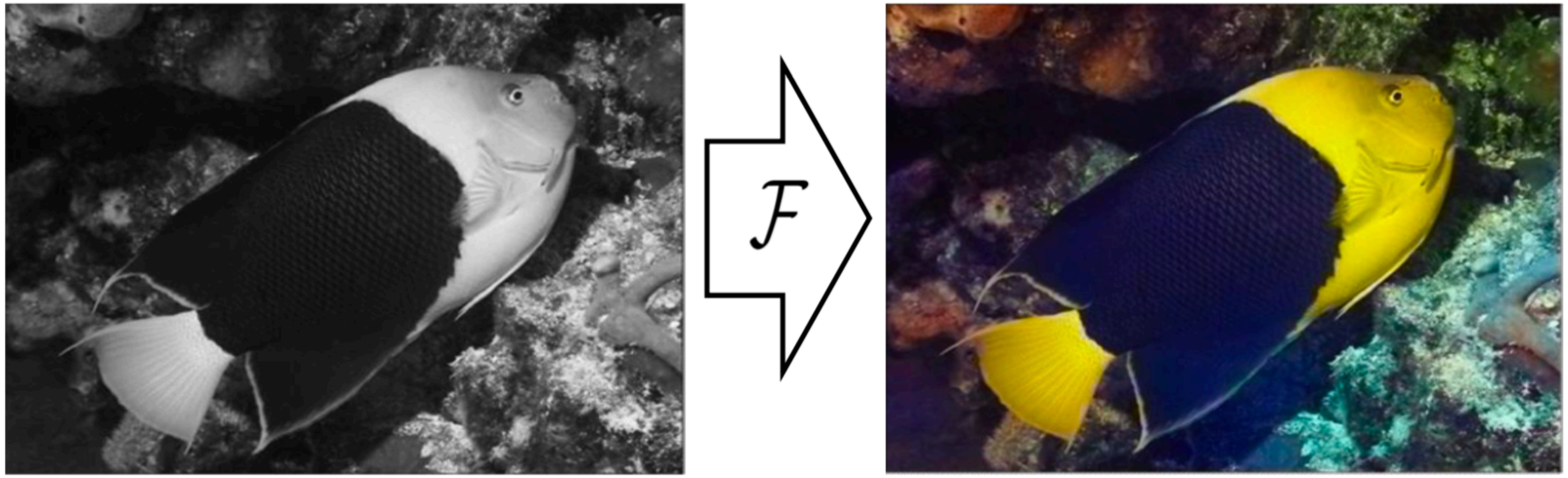
→ 180^0



→ 270^0

Gidaris et al., 2018, Predicting Image Rotations

Colorization

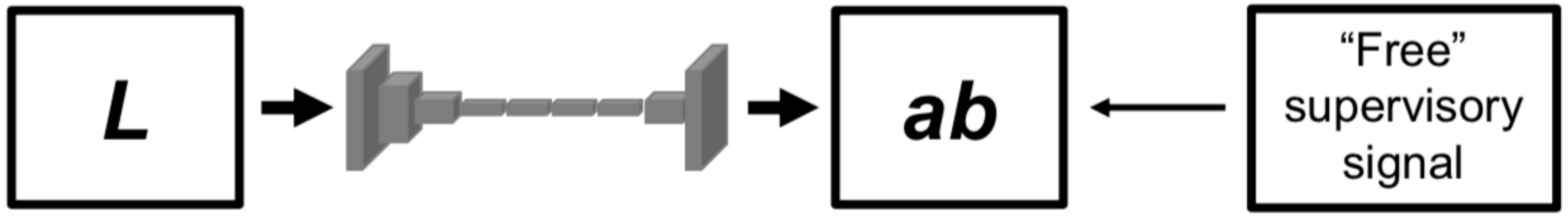


Grayscale image: *L* channel

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

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

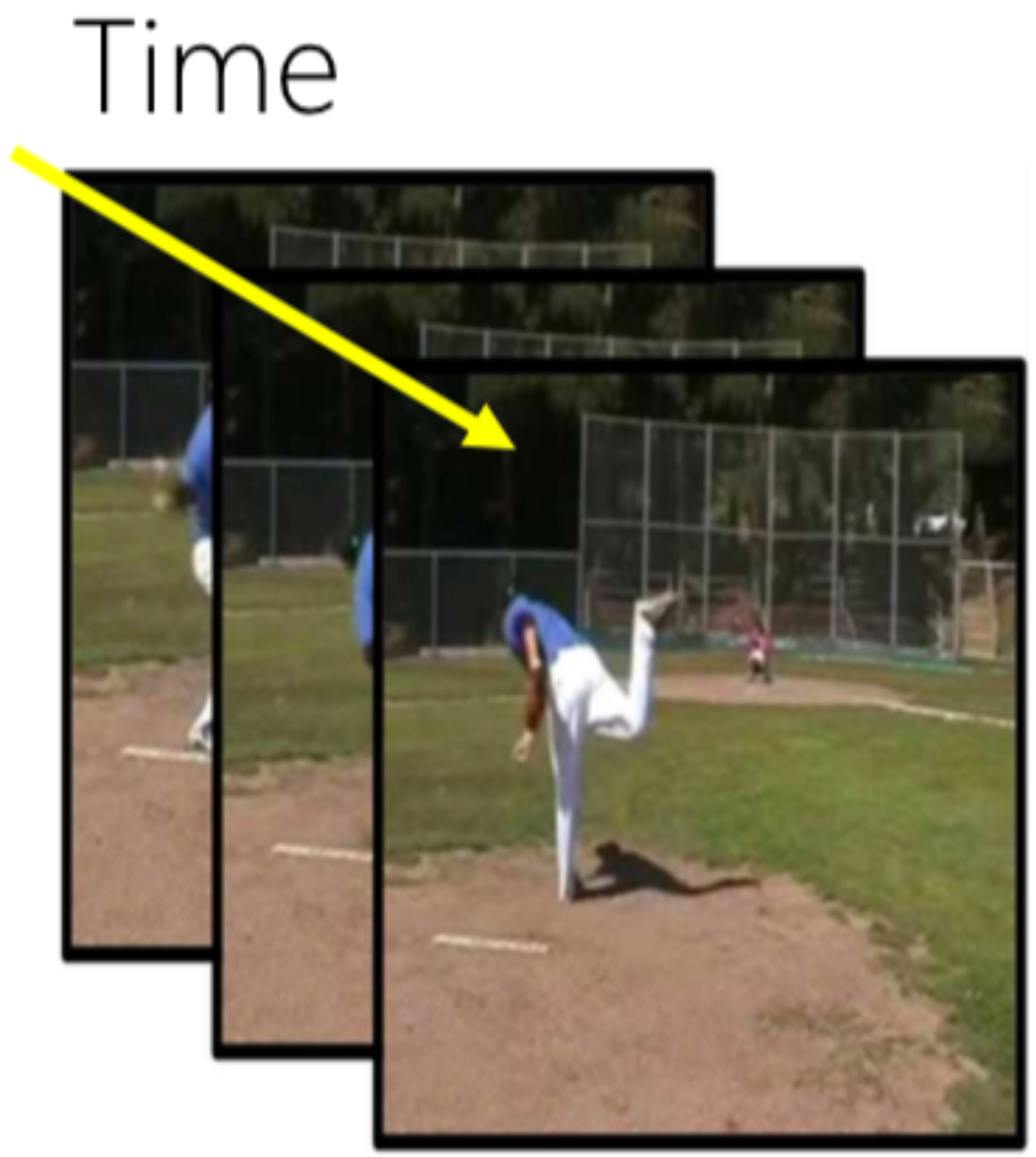
Self-supervision in computer vision

- Using images
- Using video
- Using video and sound

Video

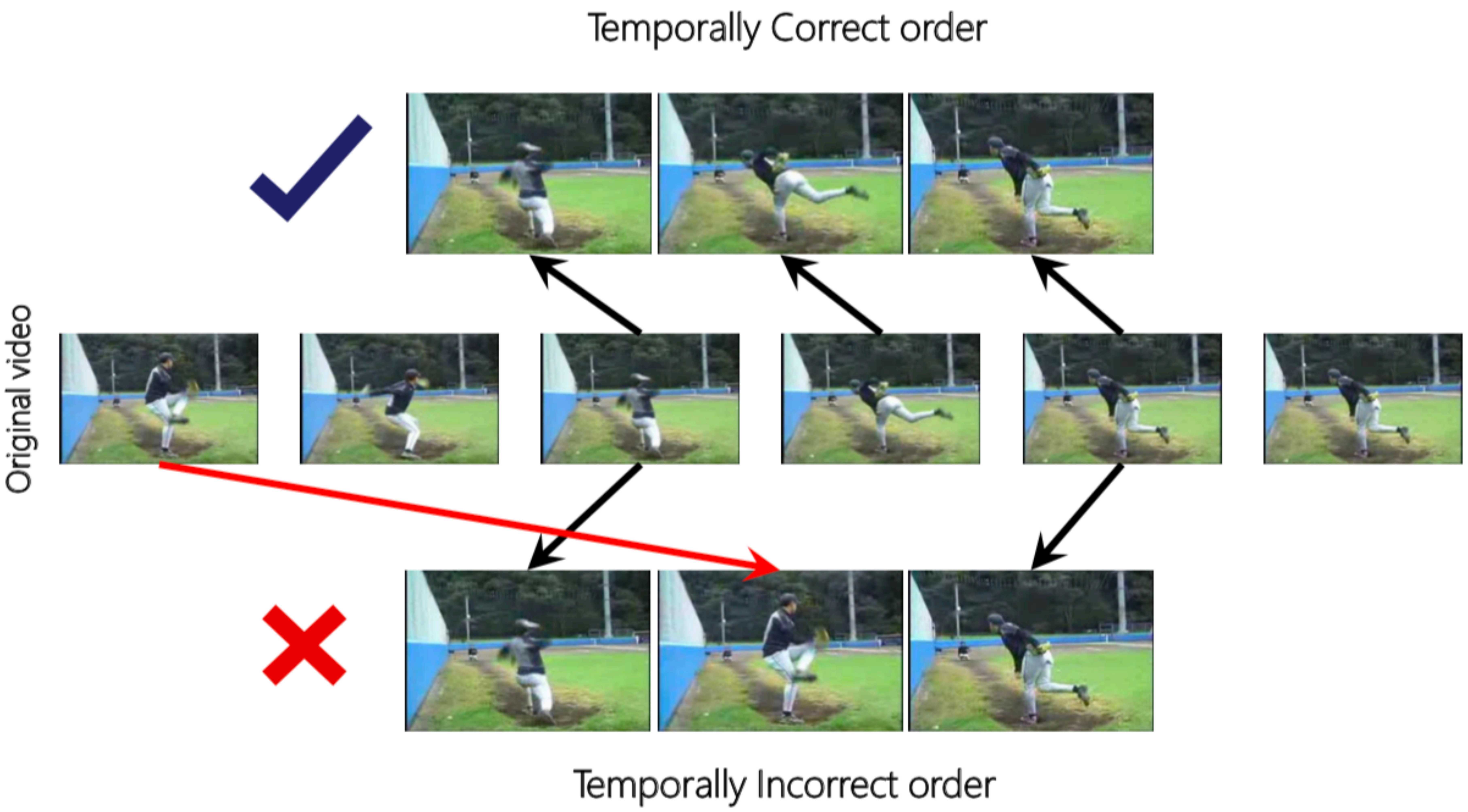
- Video is a ″sequence″ of frames
- How to get ″self-supervision″?

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



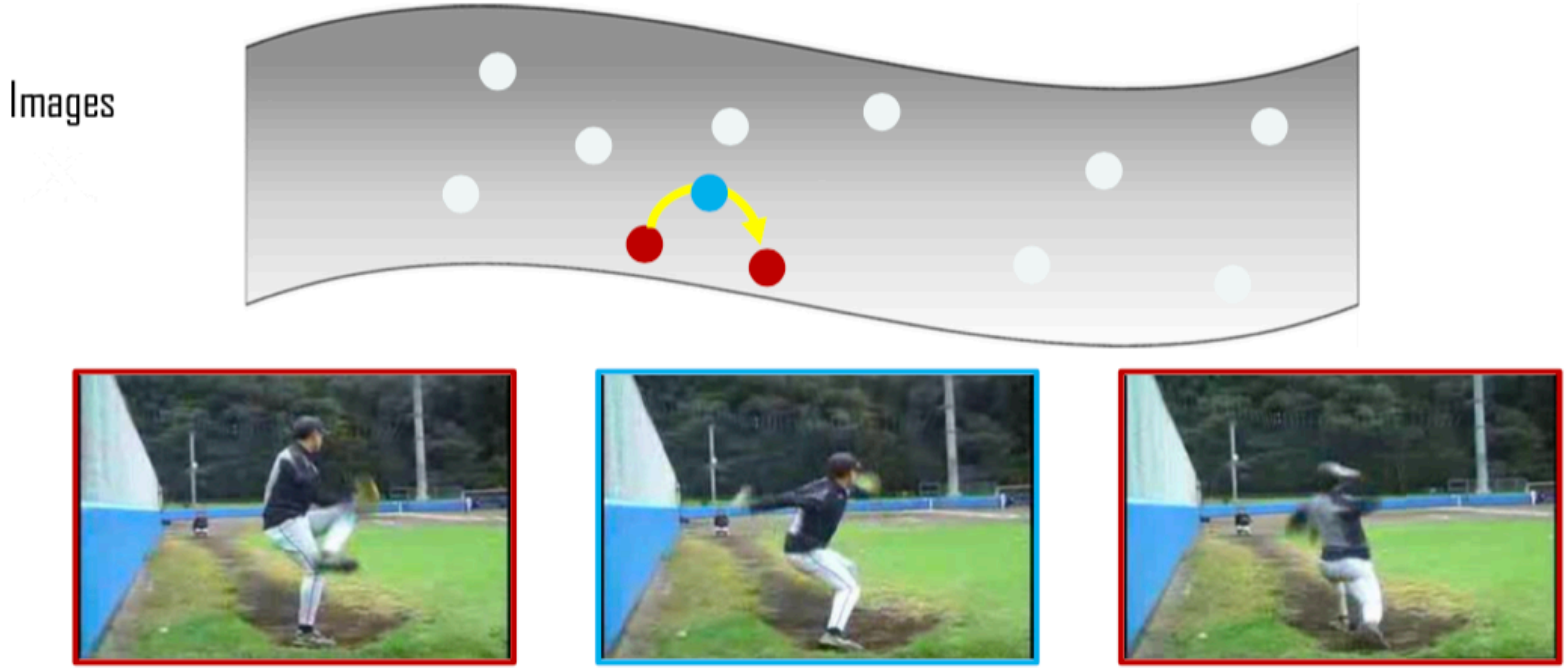
“Sequence” of data

Shuffle & Learn



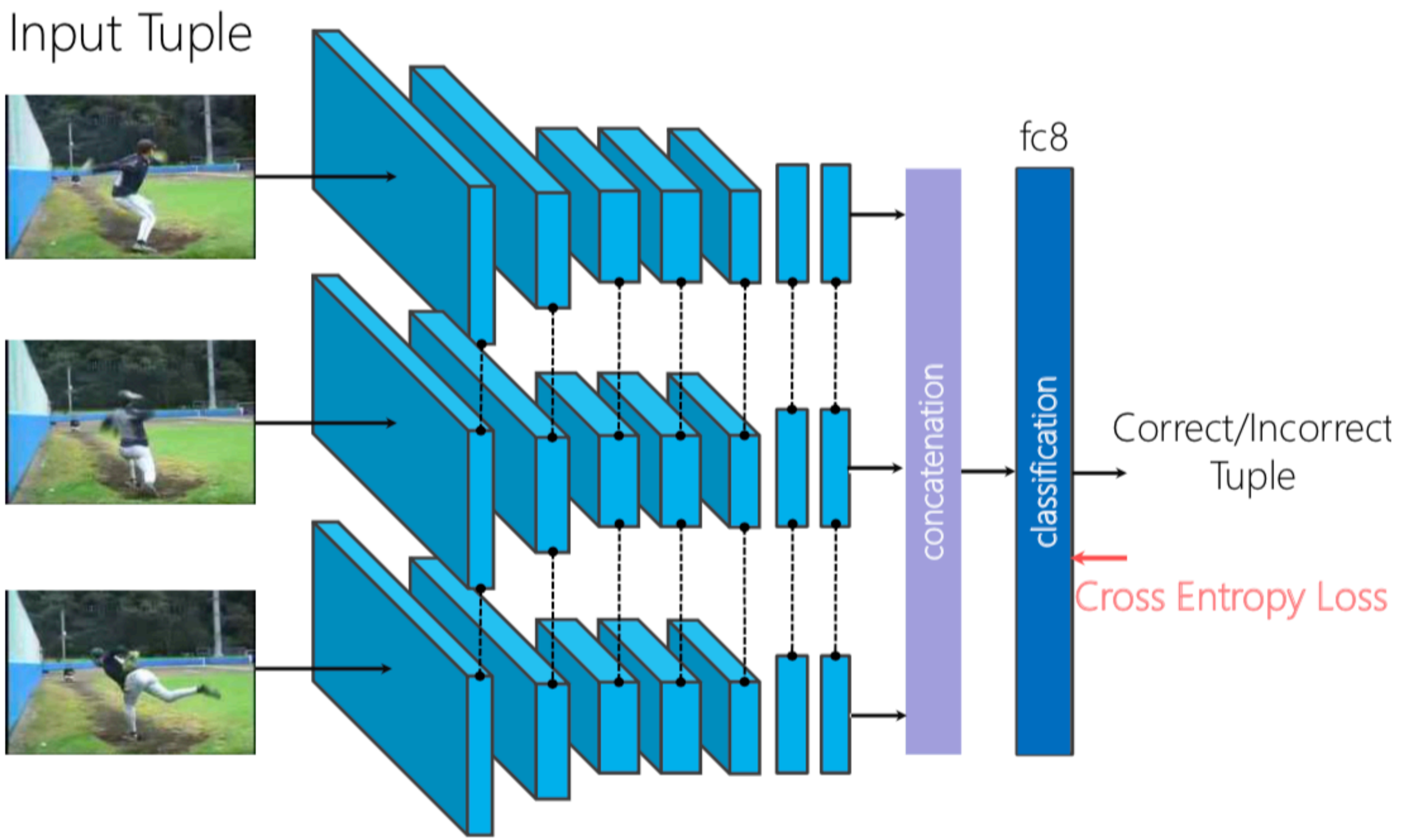
Misra et al., 2016, Shuffle and Learn

Shuffle & Learn



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

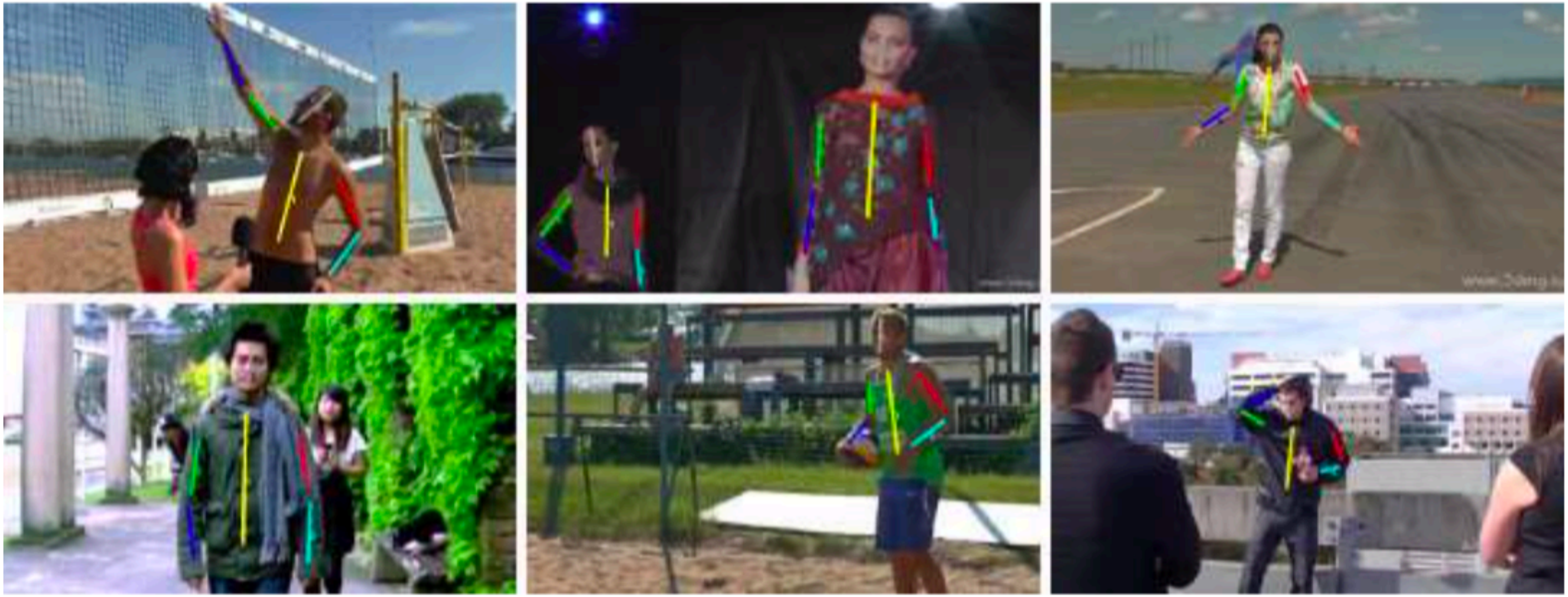
Shuffle & Learn



Nearest Neighbors of Query Frame (fc7 features)

Shuffle & Learn

Fine-tune on Human Keypoint Estimation

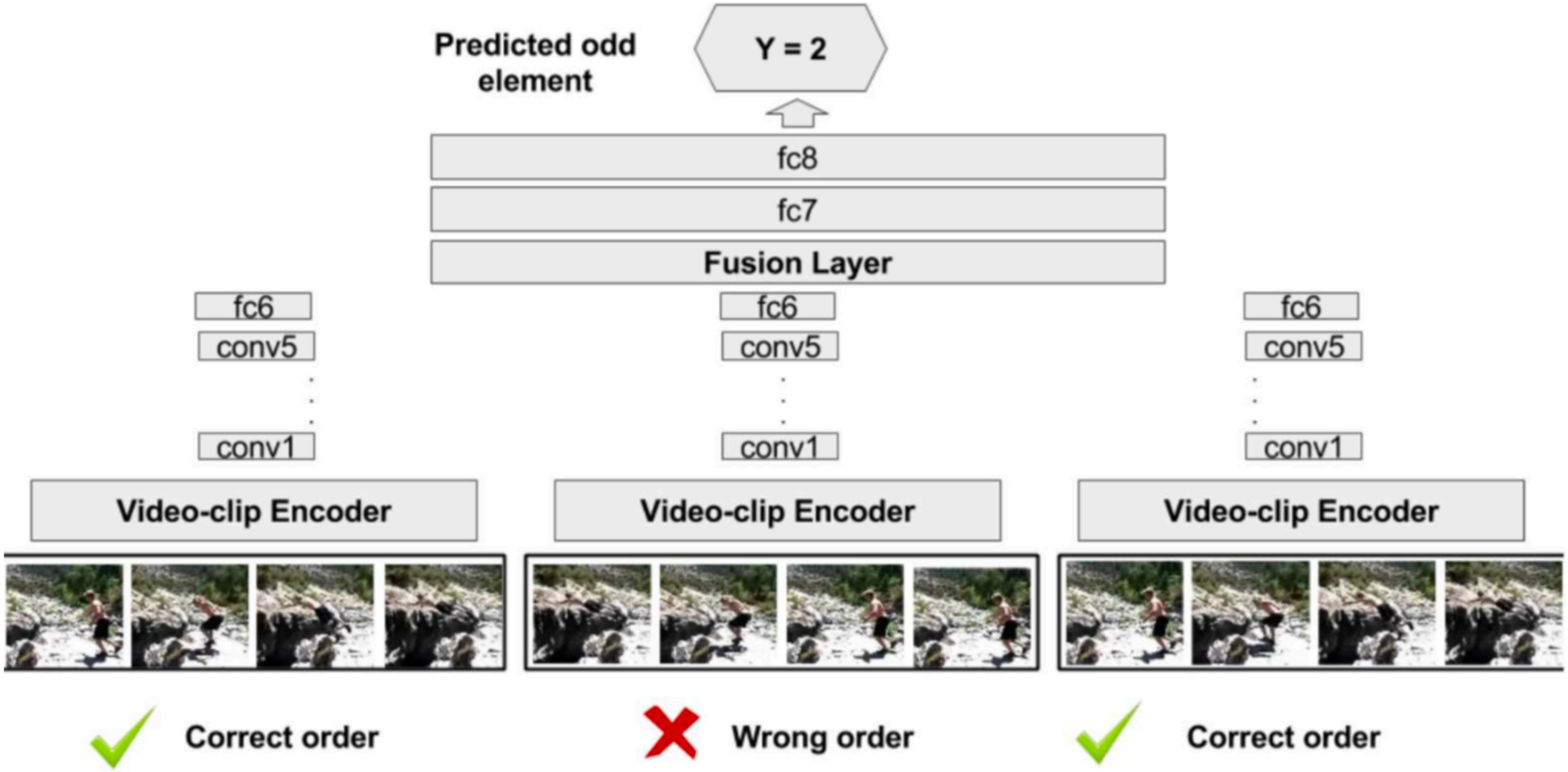


Shuffle & Learn

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

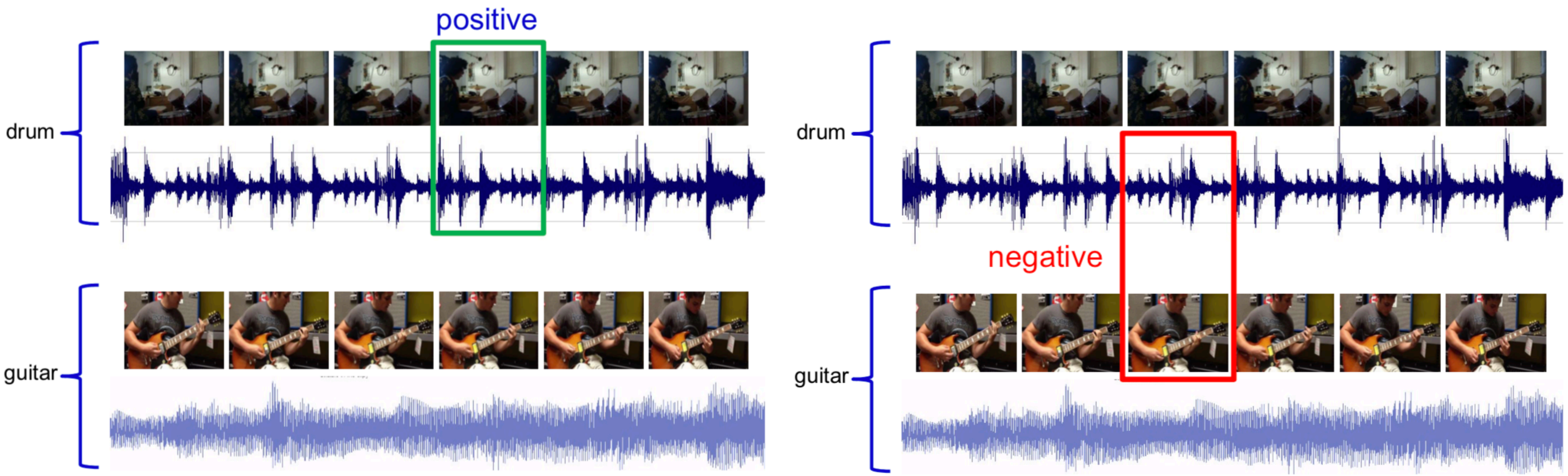
Audio-Visual co-supervision

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

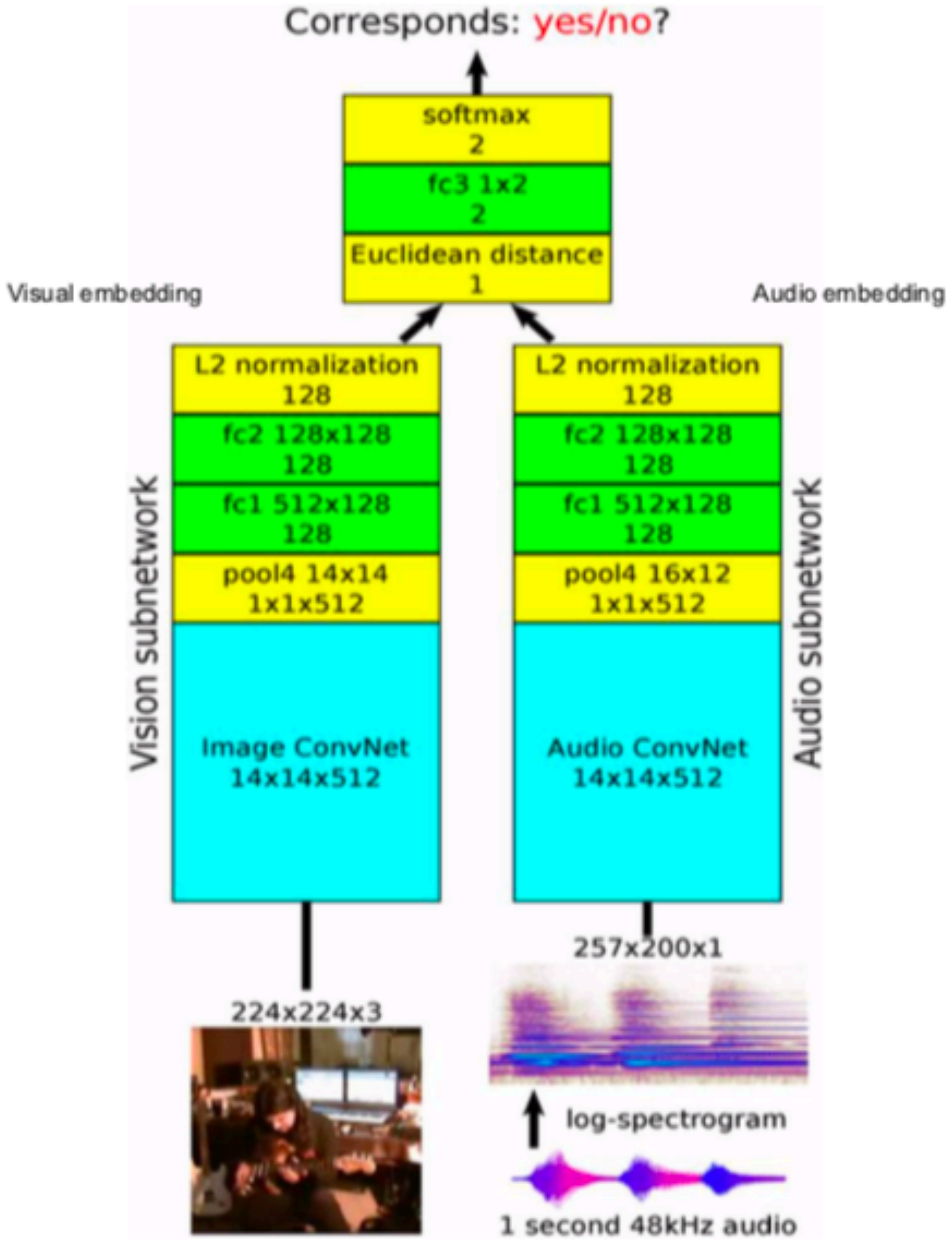
Correspond?

Arandjelović and Zisserman, 2017, “Objects that Sound”

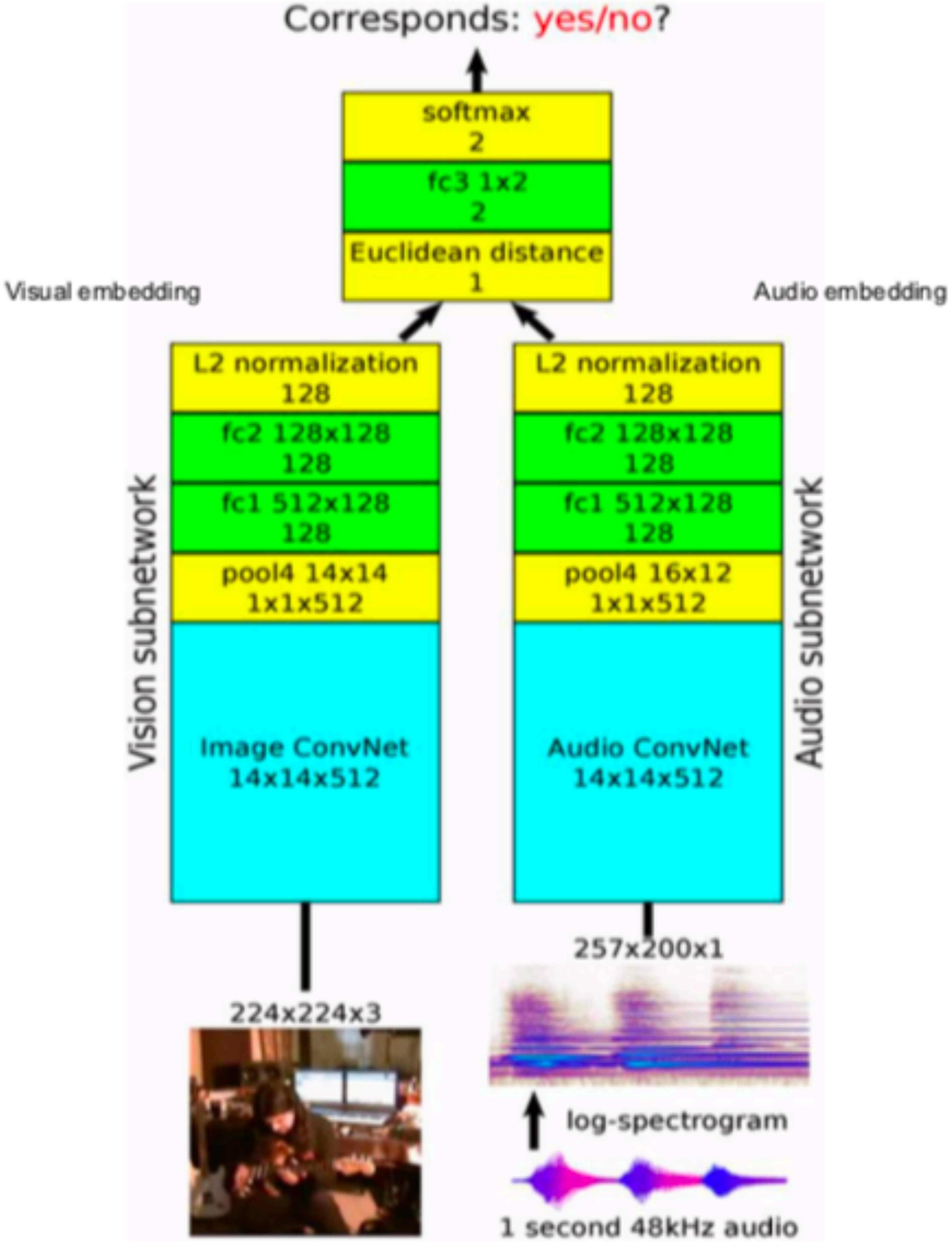
Objects that Sound



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

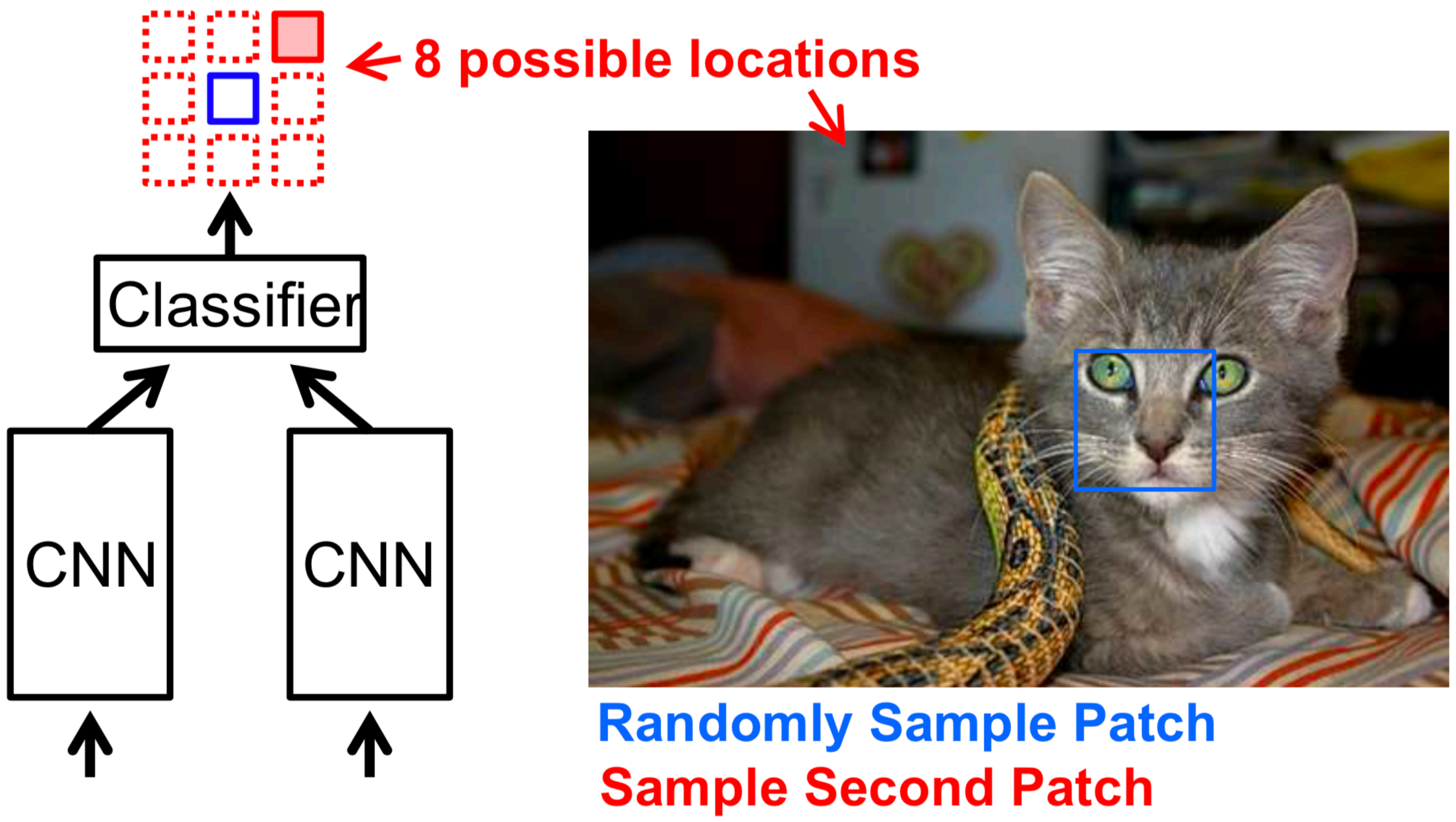
Are they complementary?

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

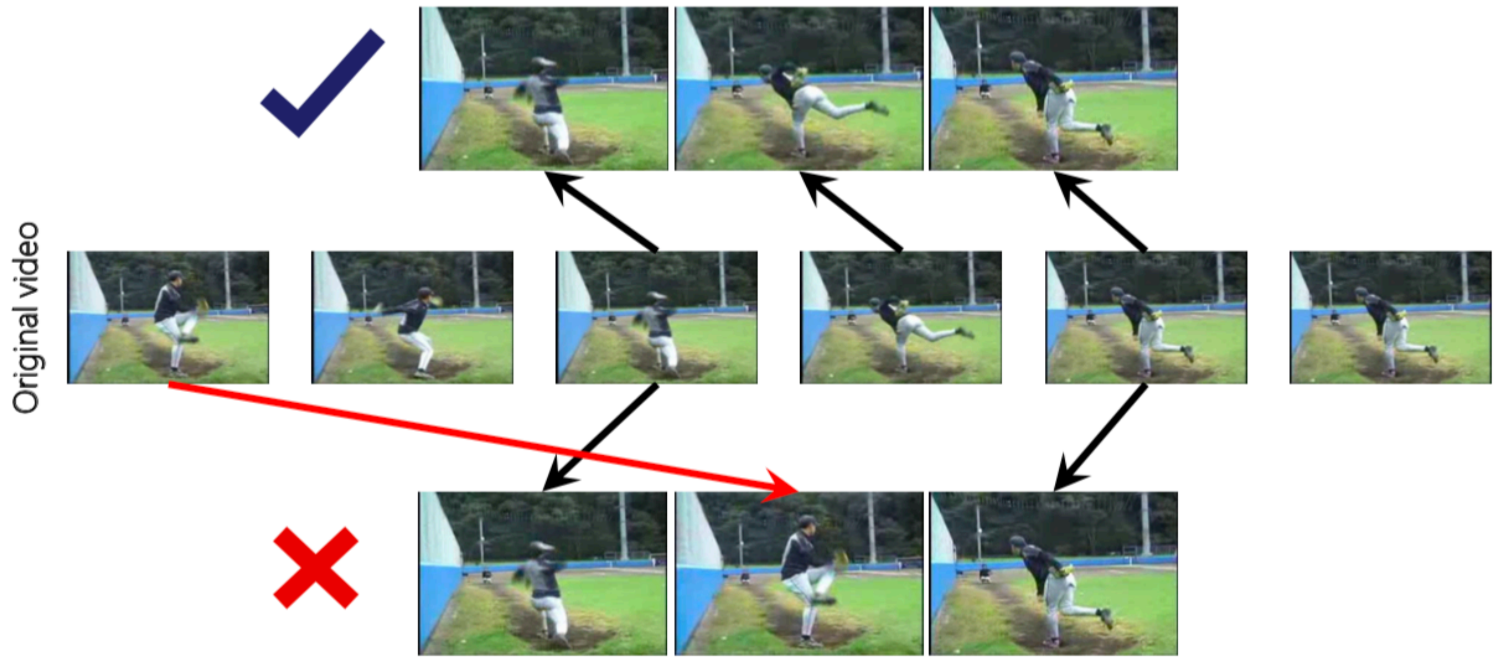
Doersch & Zisserman, 2017, Multi-task self-supervised visual learning

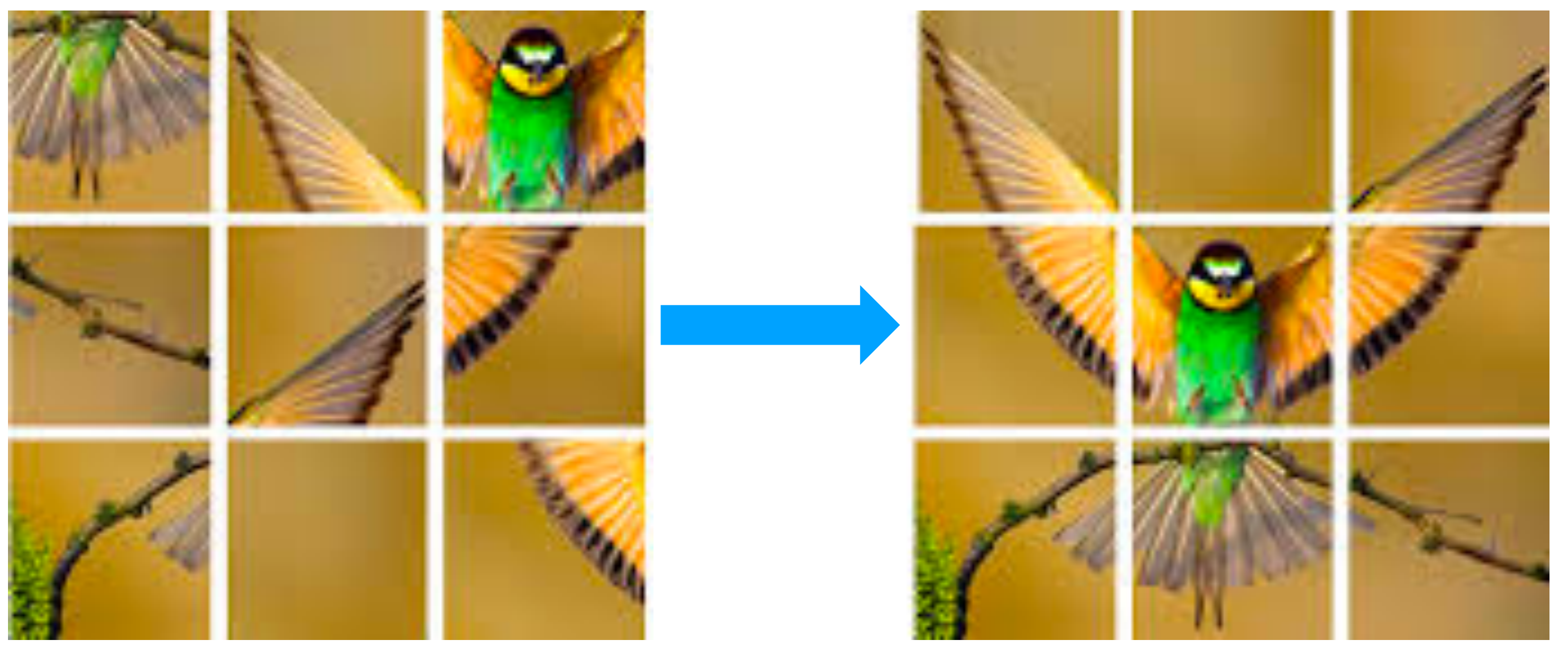
Information predicted: varies across tasks

Less More

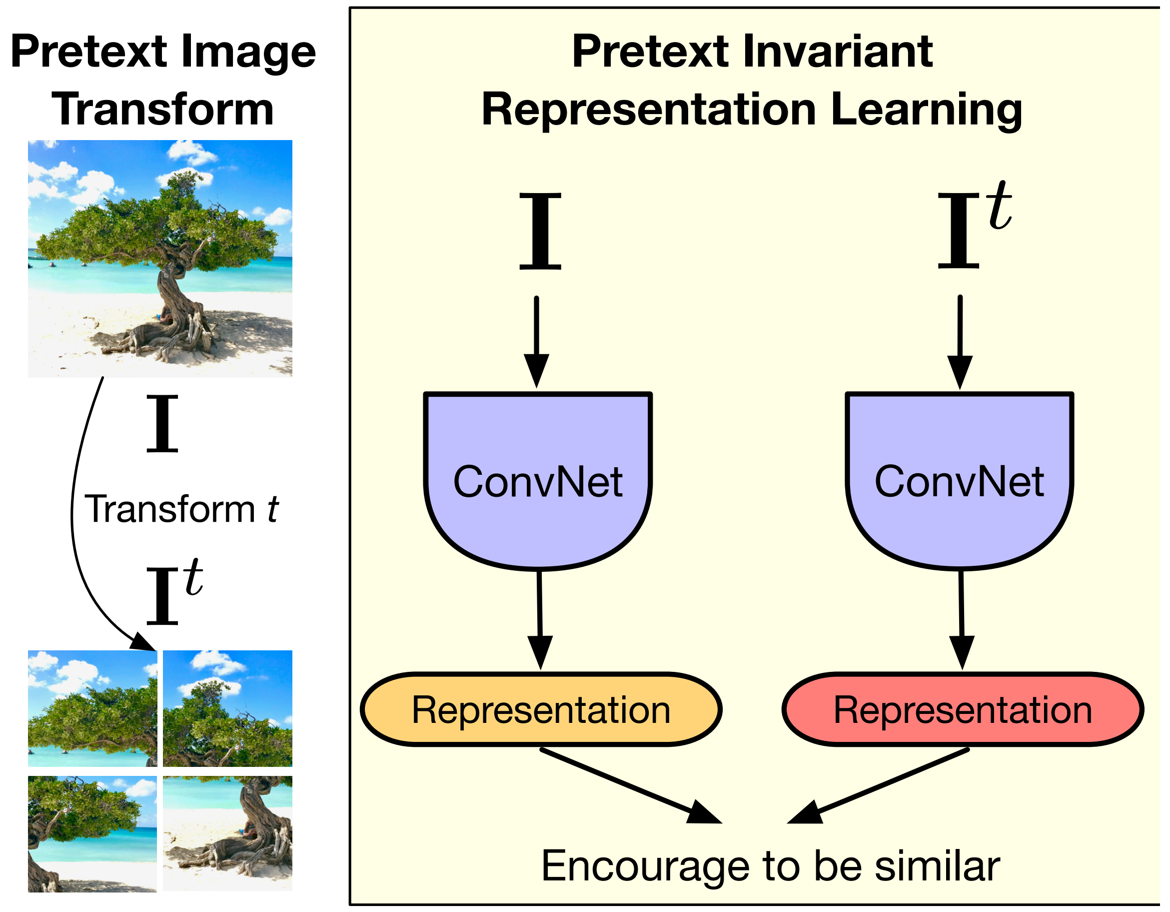


Pretext tasks

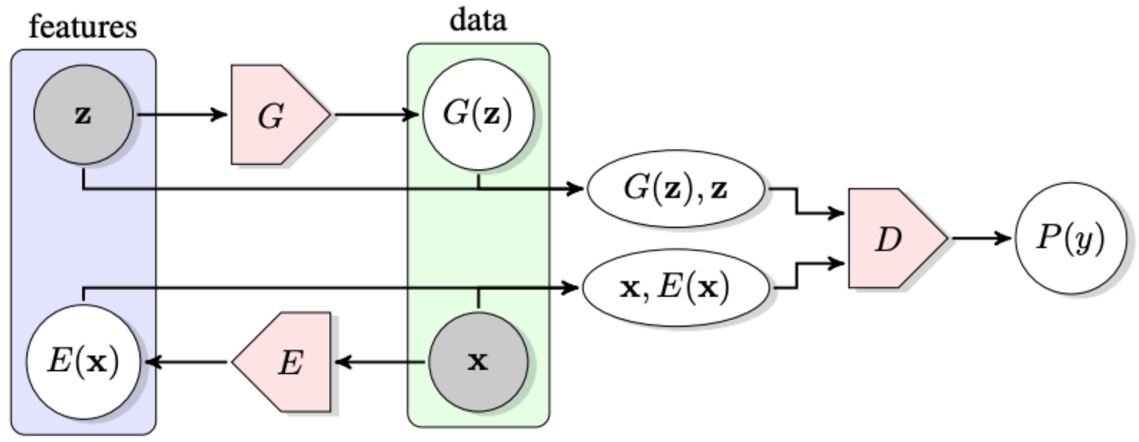




Contrastive/Clustering



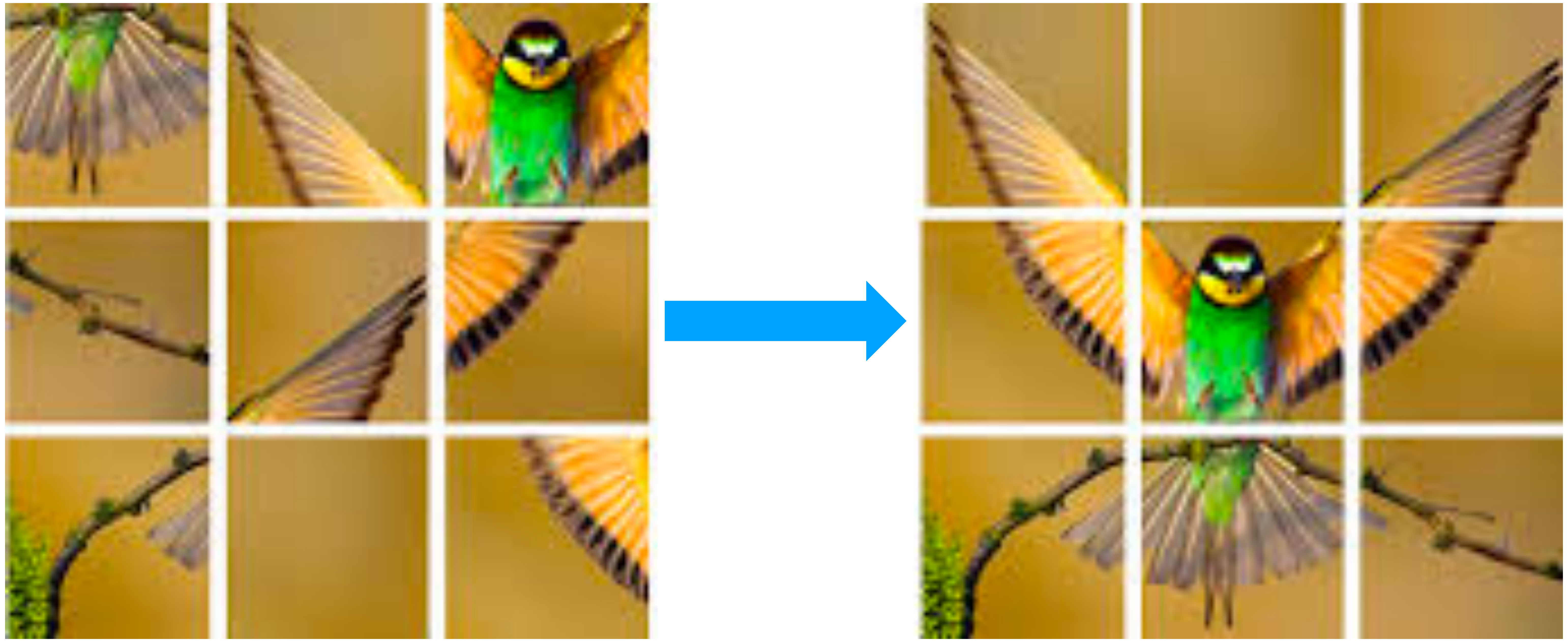
Generative



AutoEncoder, VAE, GAN, BiGAN

Predict more information

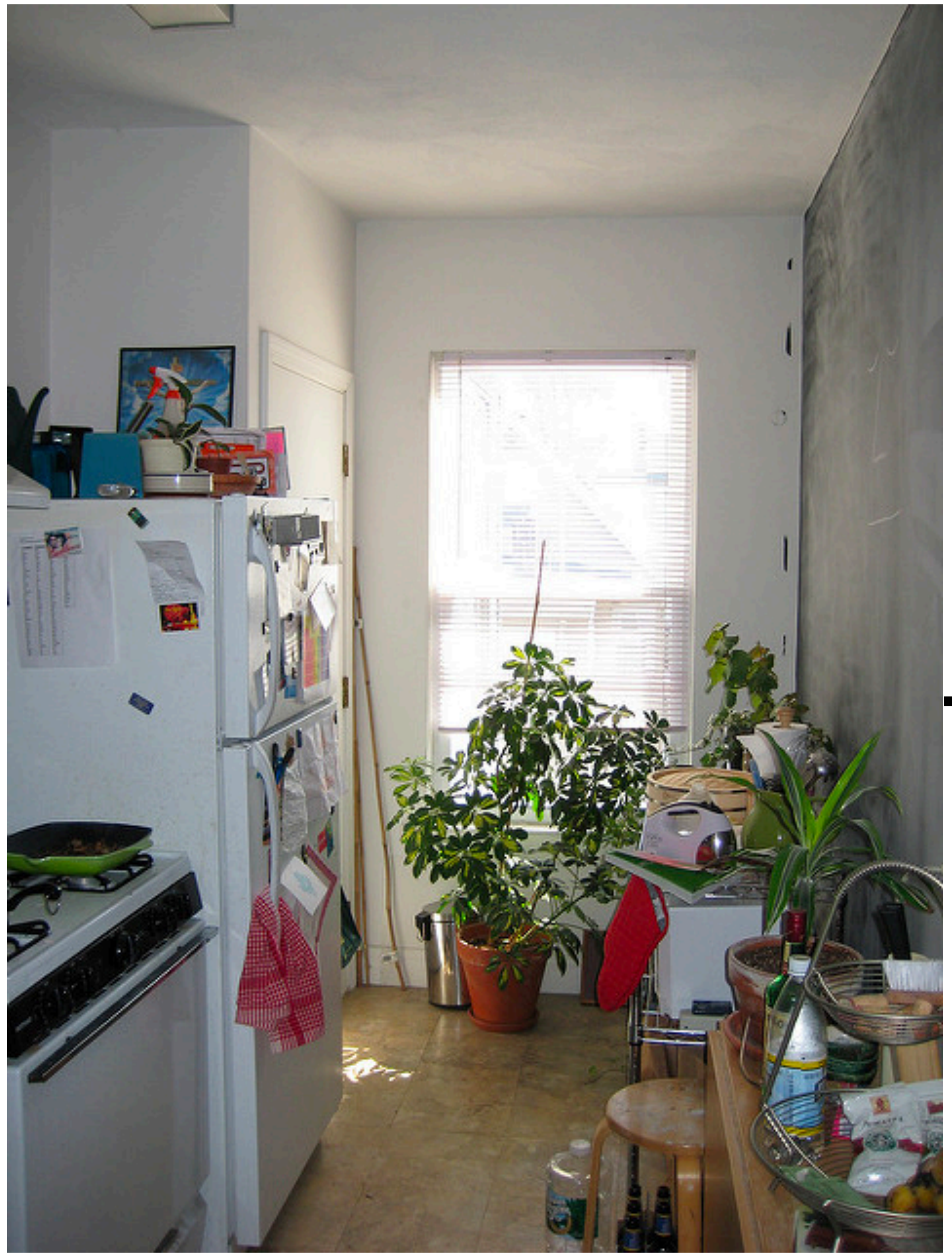
Scaling self-supervised learning

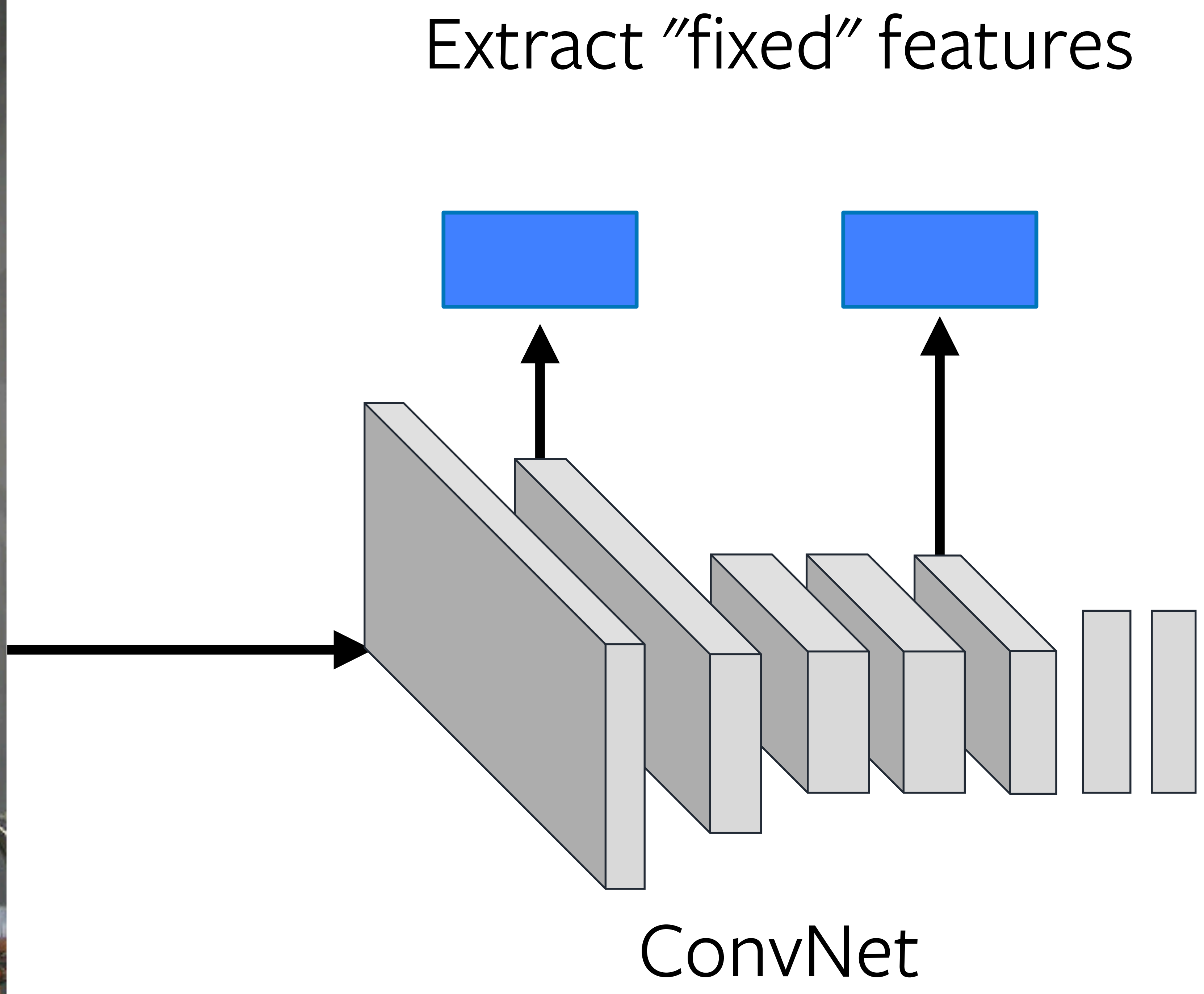


Jigsaw puzzles
(Noorozi & Favaro, 2016)

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

Evaluating the representation

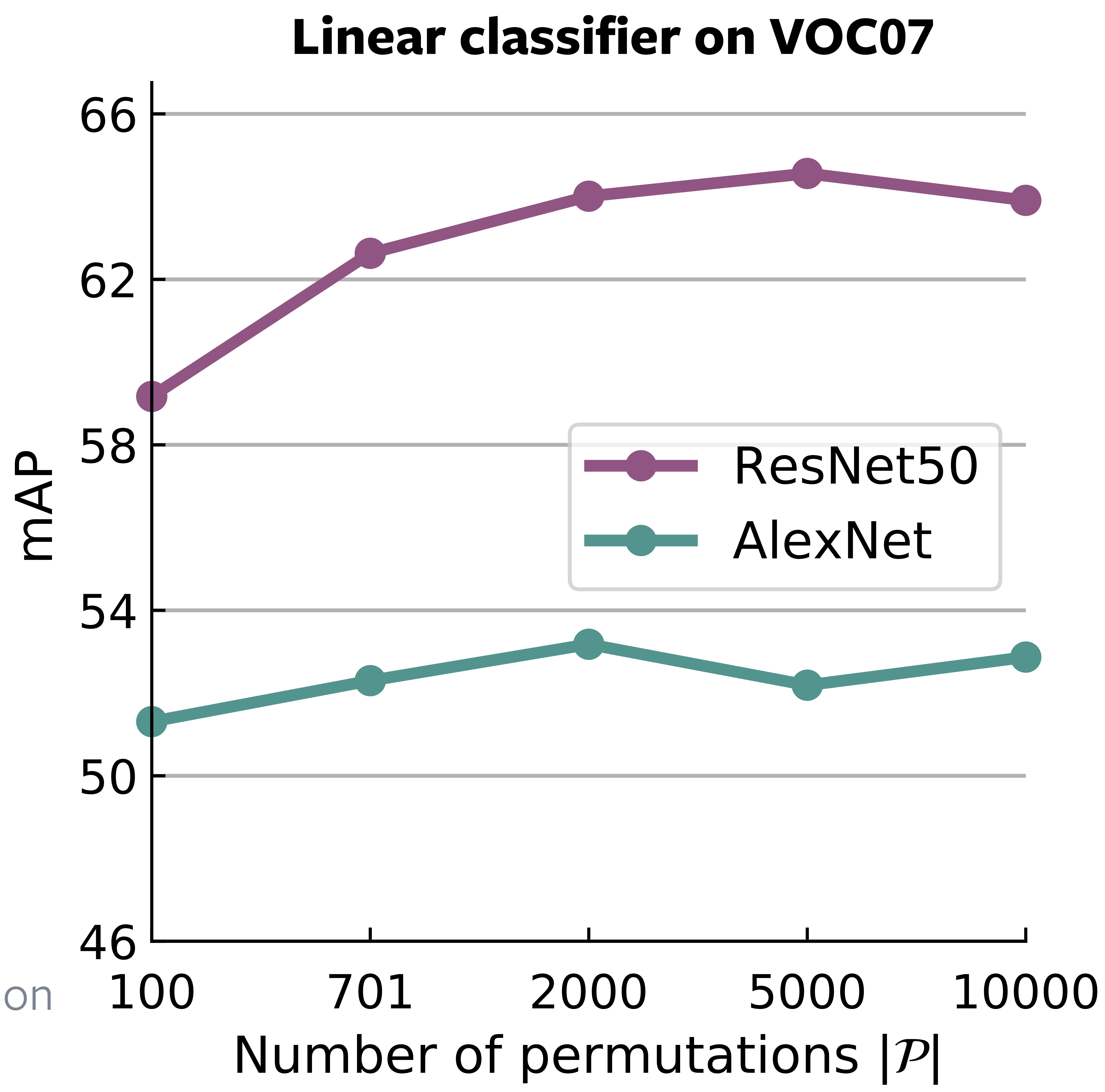




Evaluating the representation

- 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)

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**

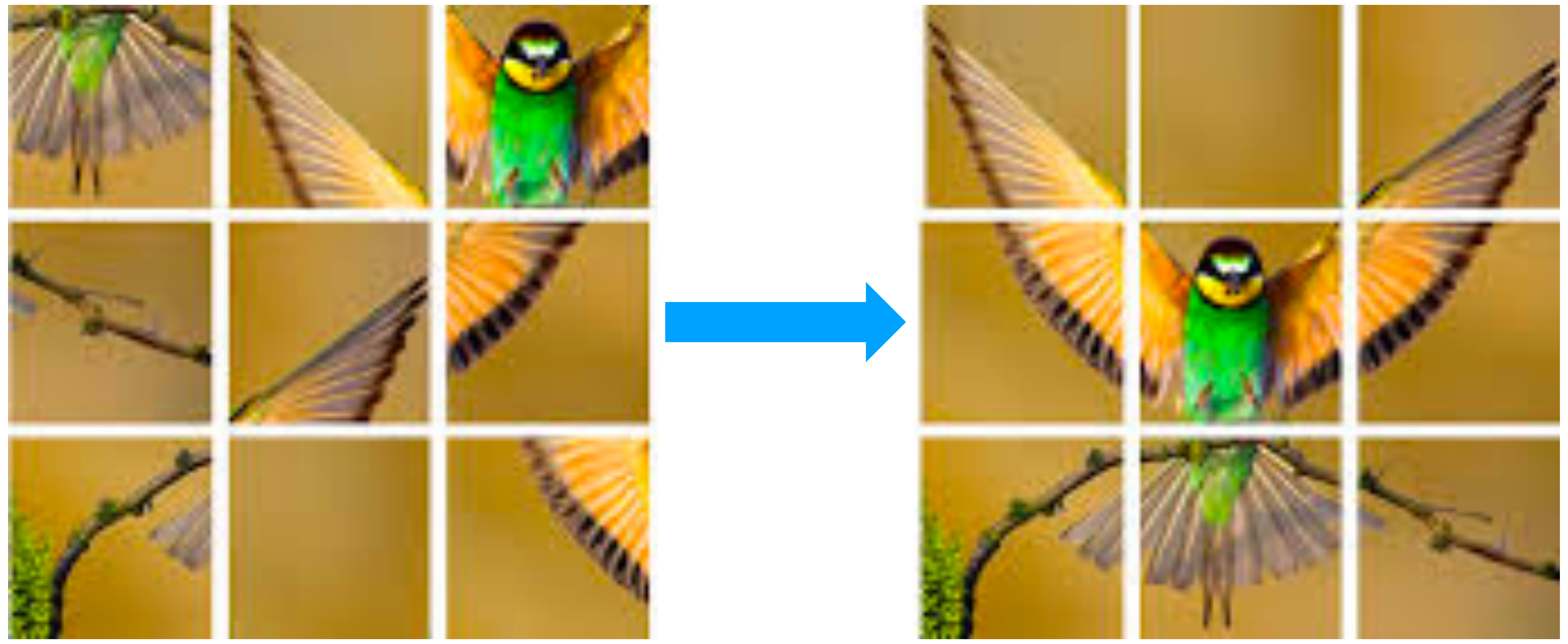
Outperforms ImageNet supervised

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

Pretext tasks

Rotation

(Gidaris et al., 2018)

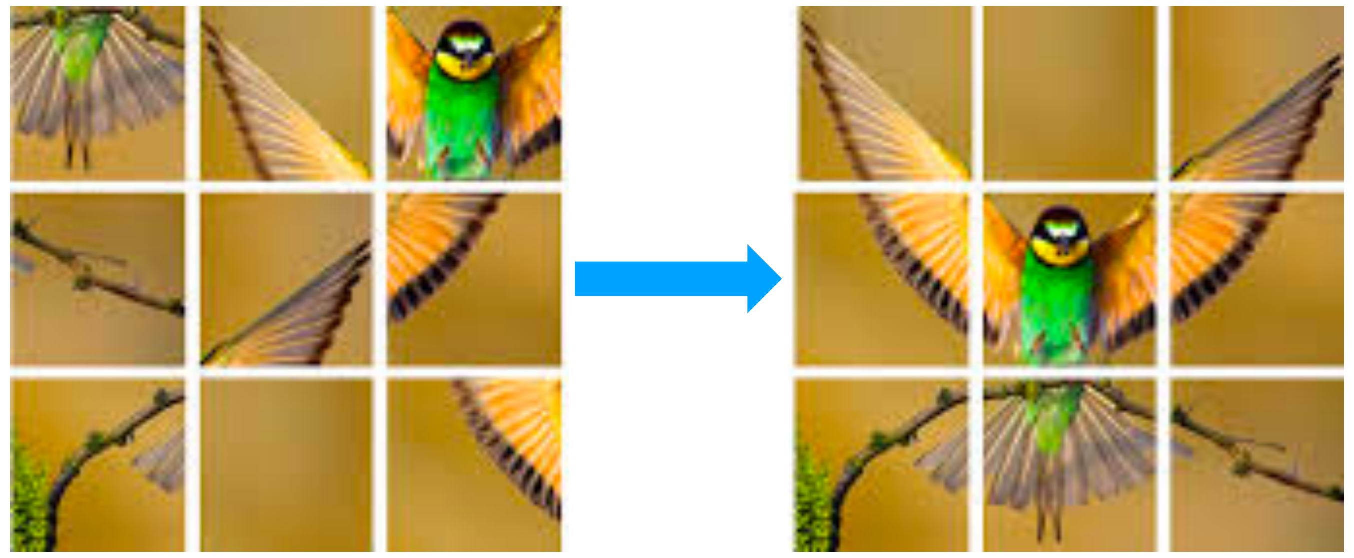


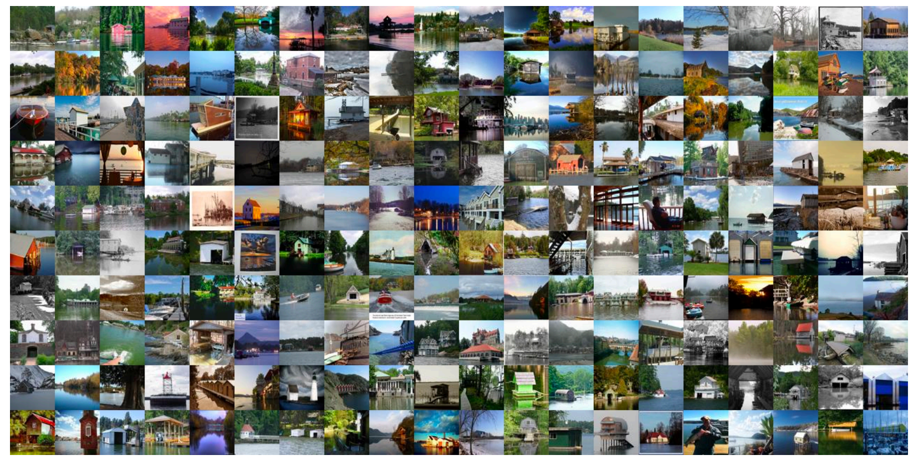
Jigsaw puzzles

(Noroozi et al., 2016)

The hope of generalization

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





Pre-training

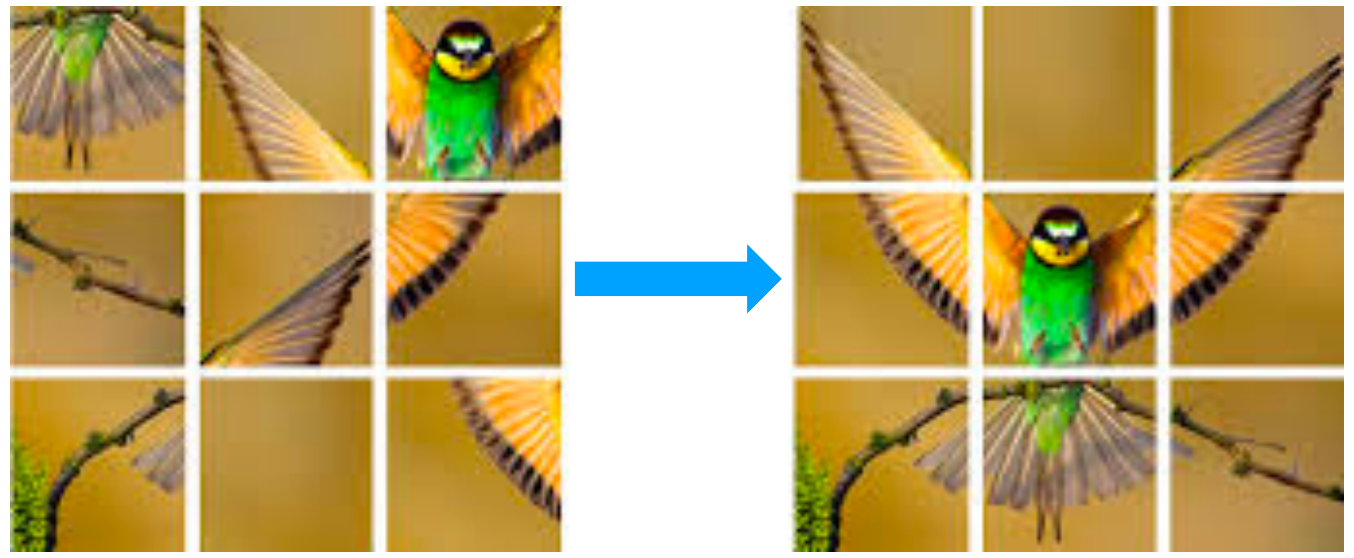
Self-supervised

Transfer Tasks

The hope of generalization

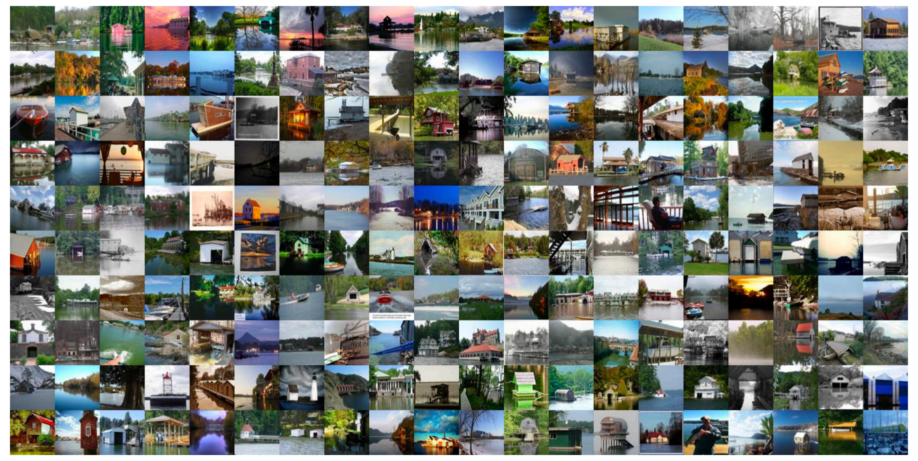
- We really **hope** that the pre-training task and the transfer task are "aligned"

#sun #nofilter #fun
#tree #aruba



Pre-training

Weak or self-supervised

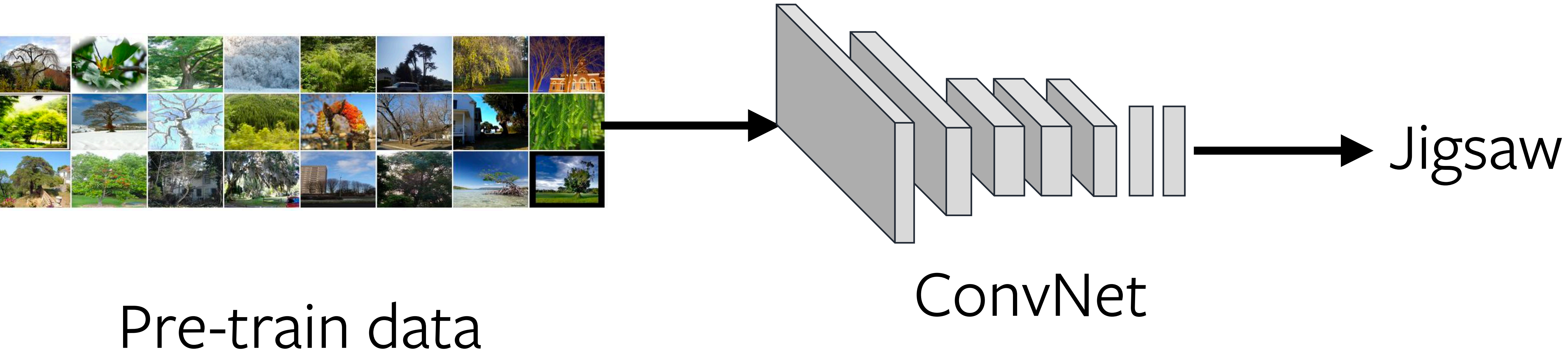


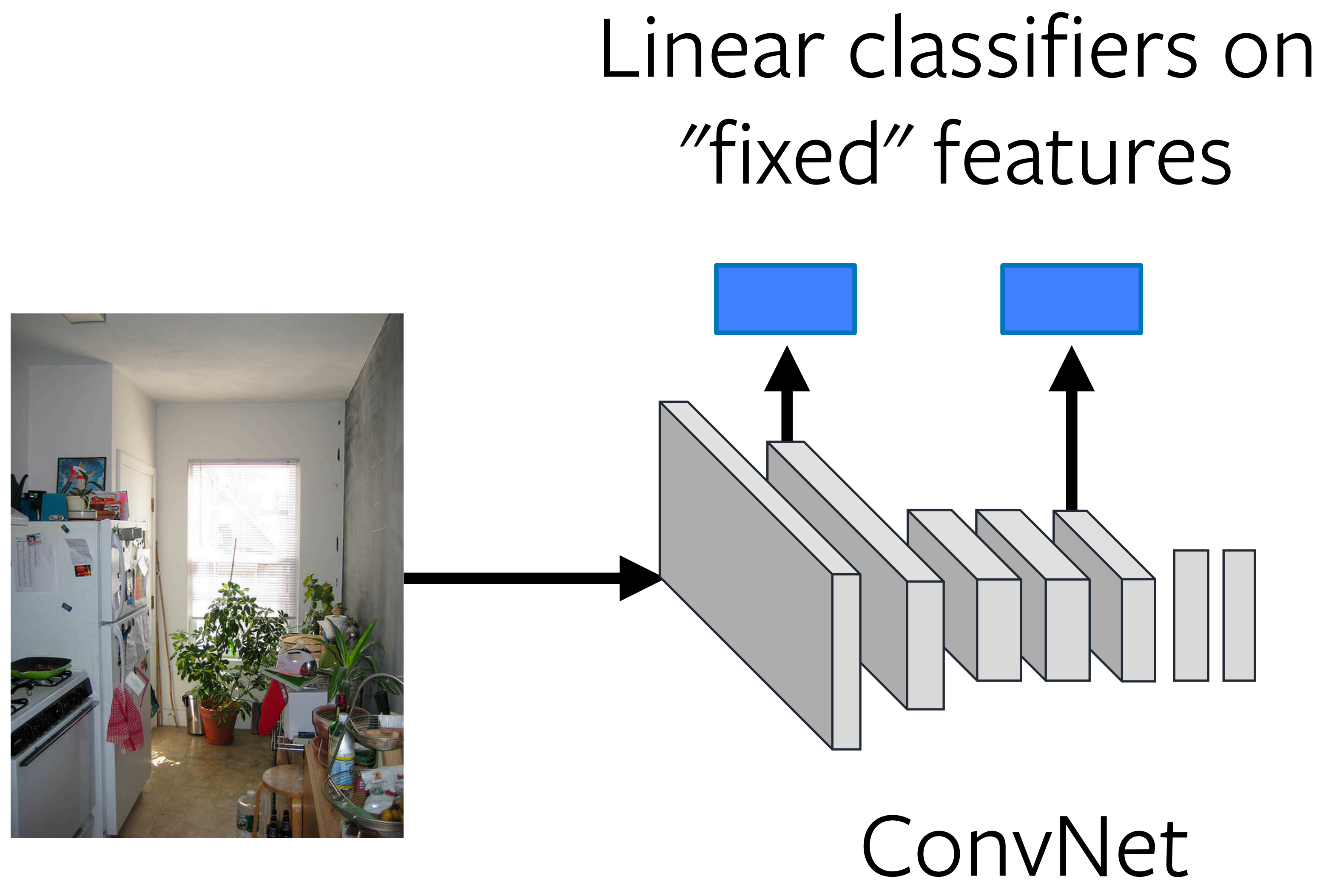
Transfer Tasks

Why should solving Jigsaw puzzles teach about "semantics"?

Why should performing a non semantic task produce good features?

The hope of generalization ... ?



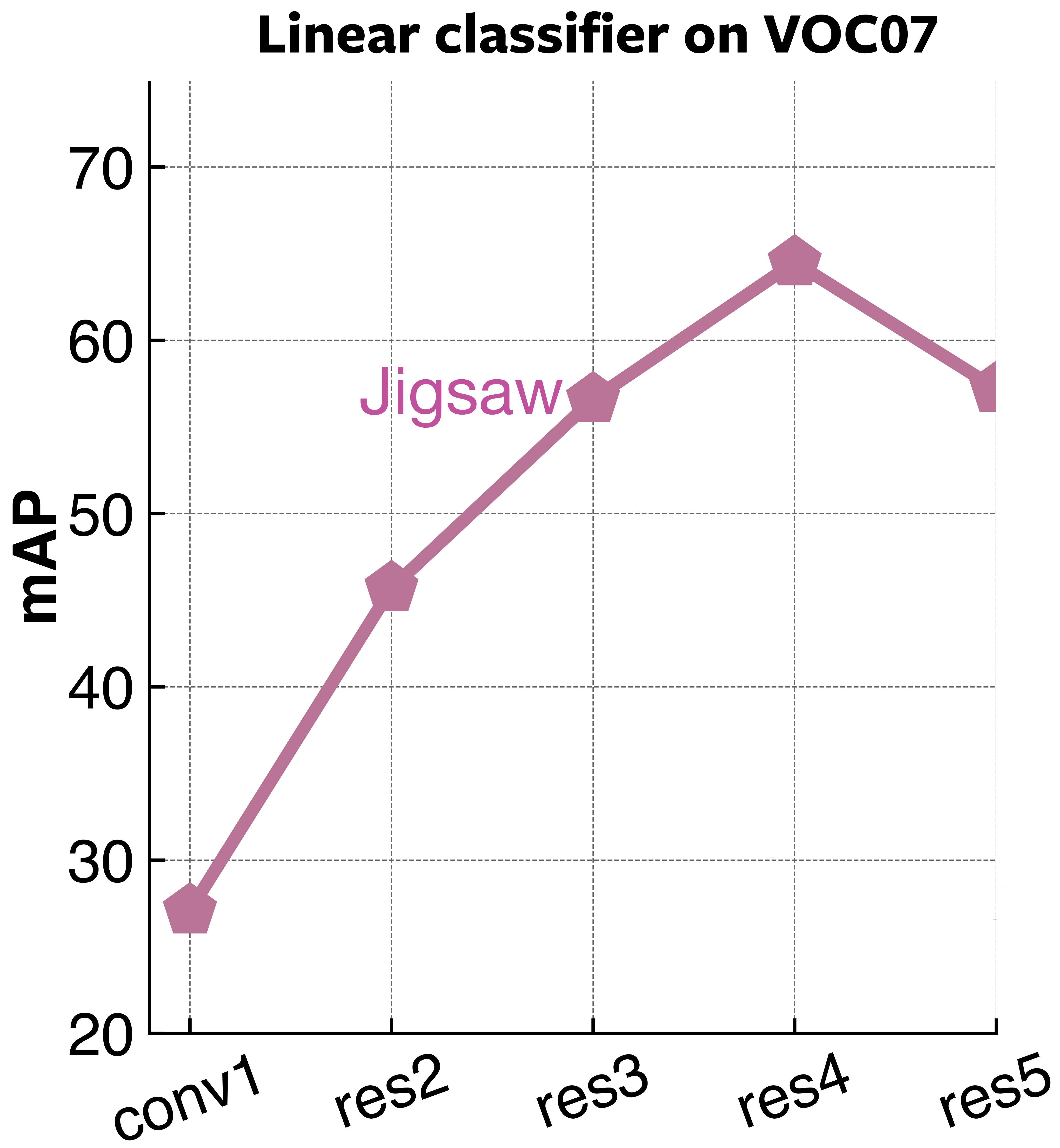


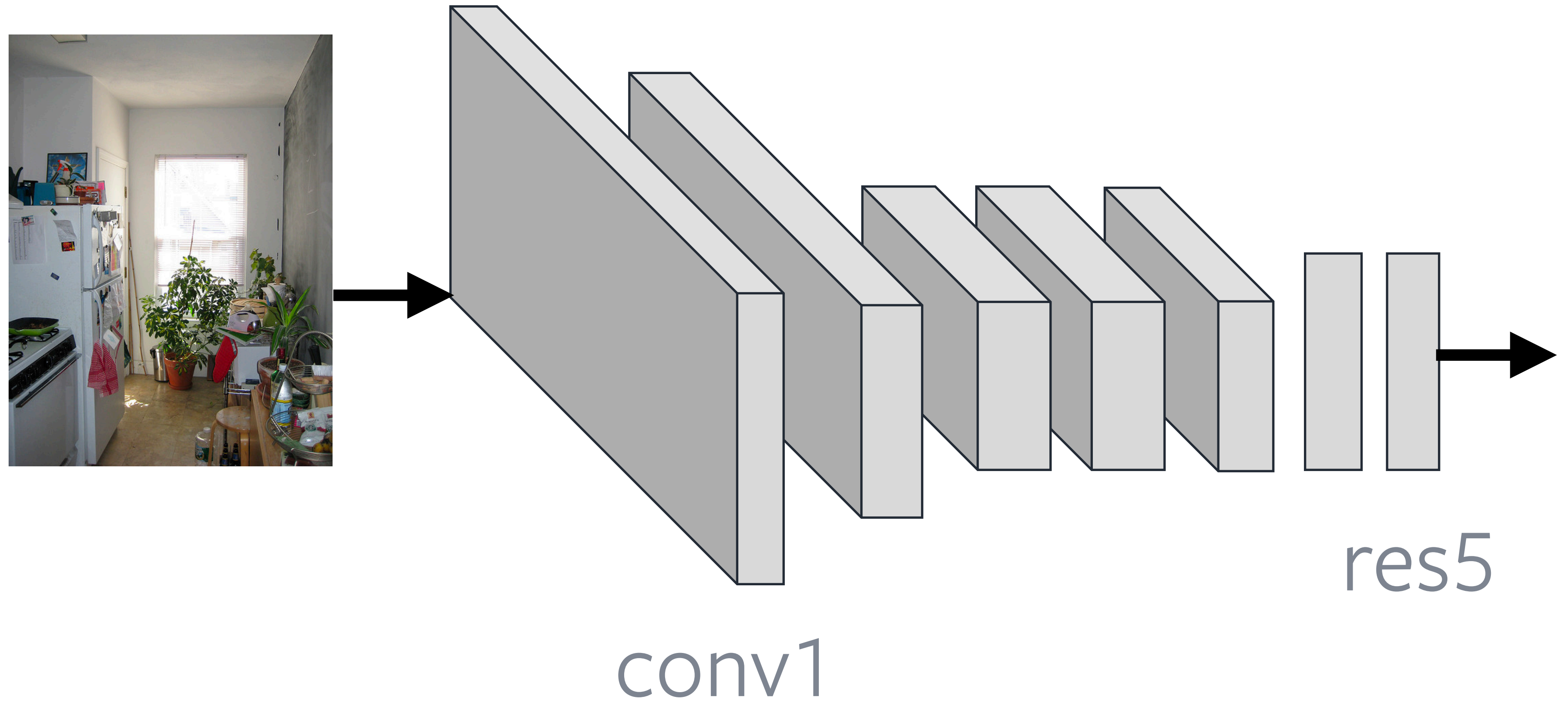
Pre-training

Weak or self-supervised

Transfer

Higher layers do not generalize ...



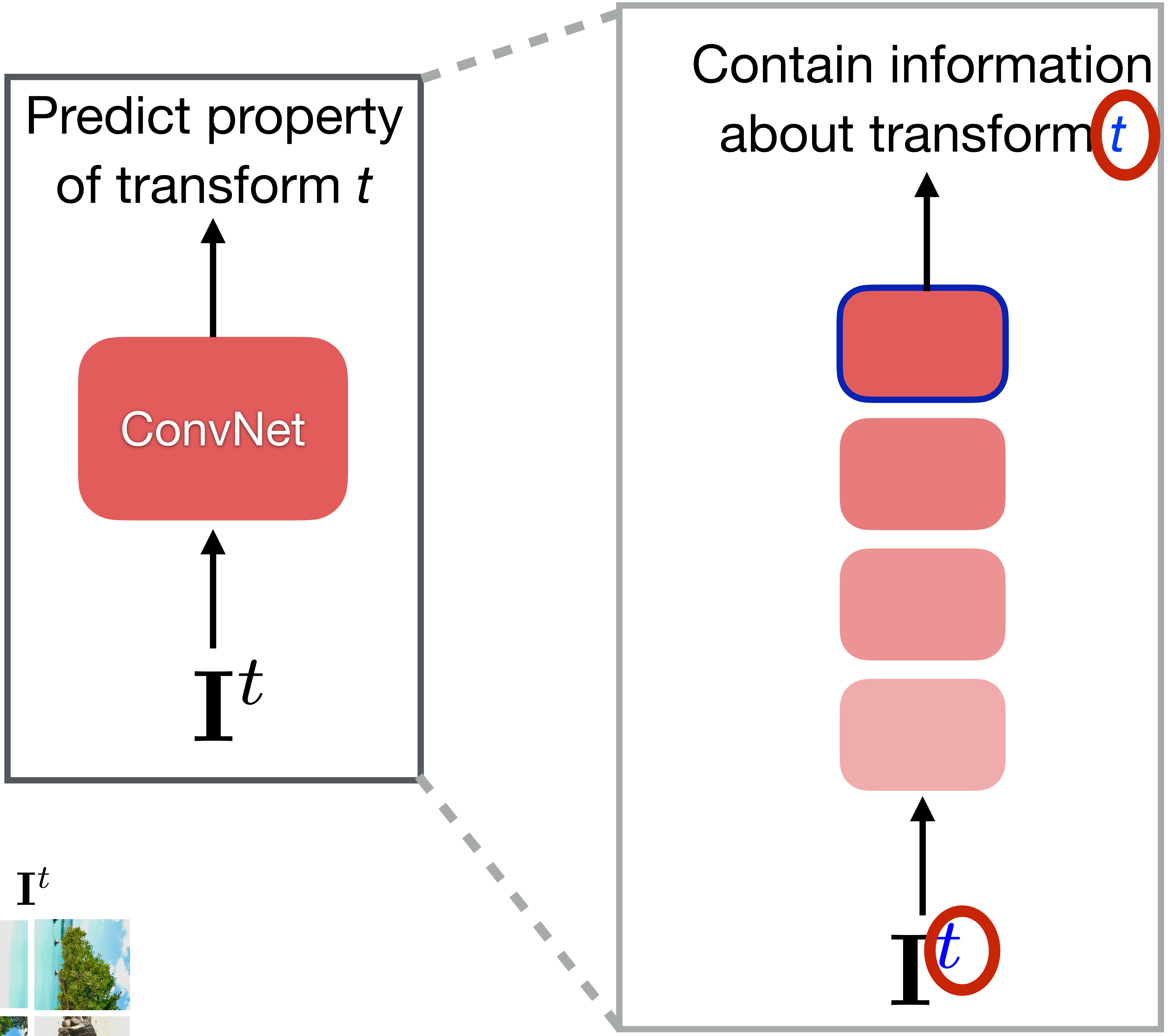


mAP = mean Average Precision
(Higher is better)

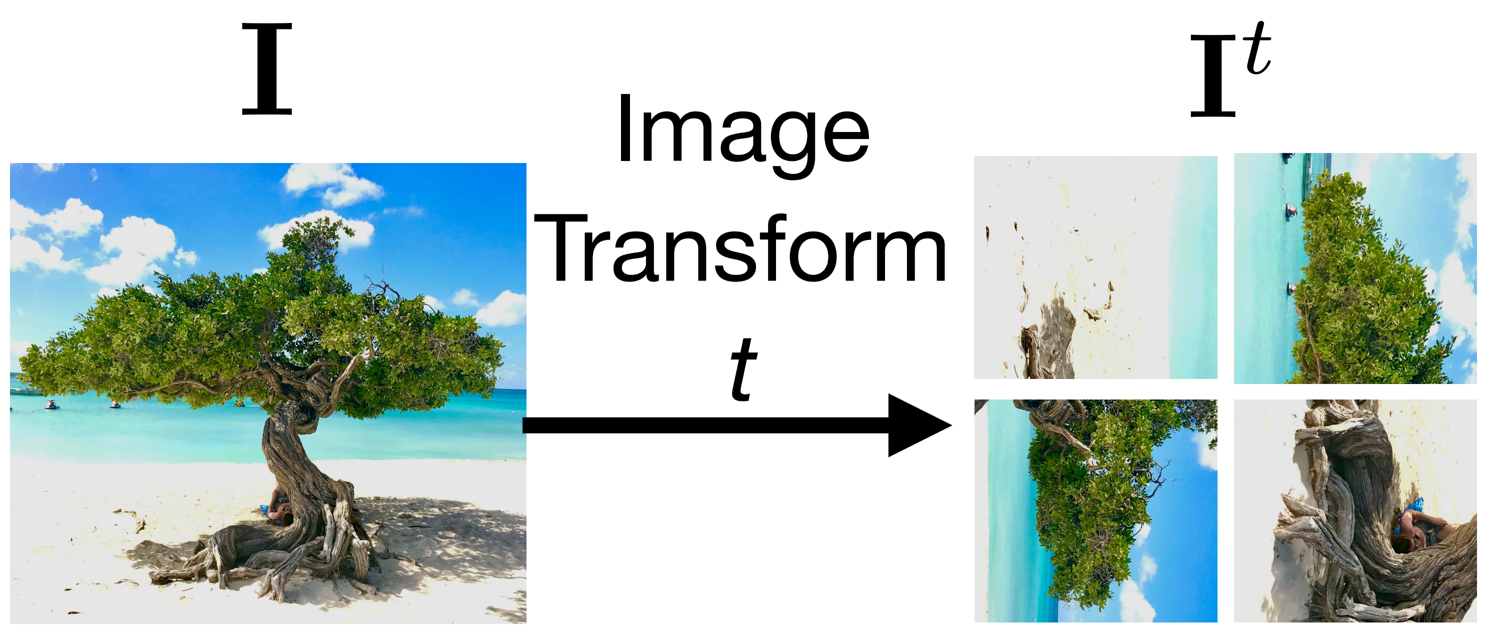
Pretext-Invariant Representation Learning (PIRL)

Ishan Misra, Laurens van der Maaten


Pretext task



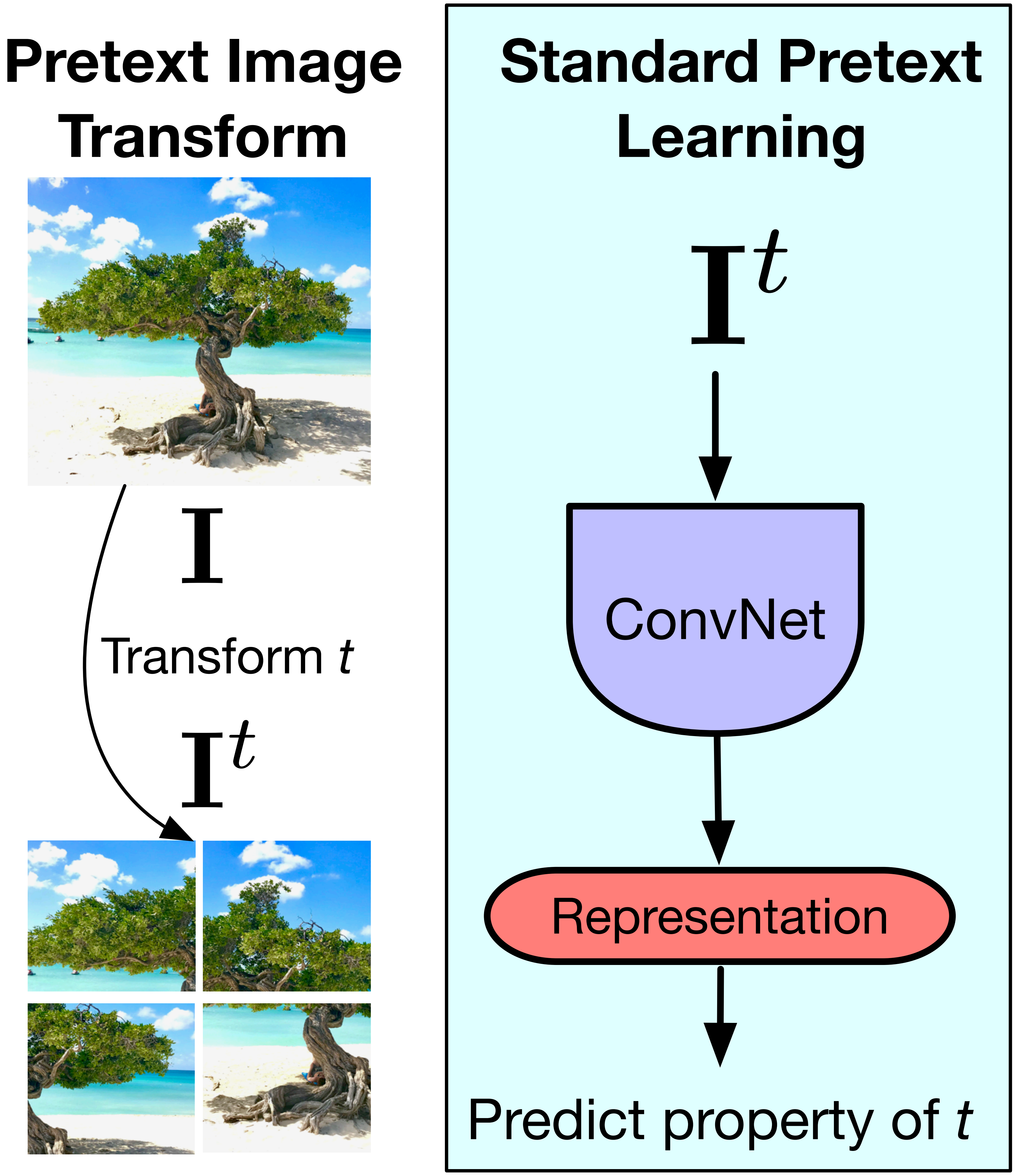
Less Semantic Features



Underlying Principle for Pretext Tasks

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

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

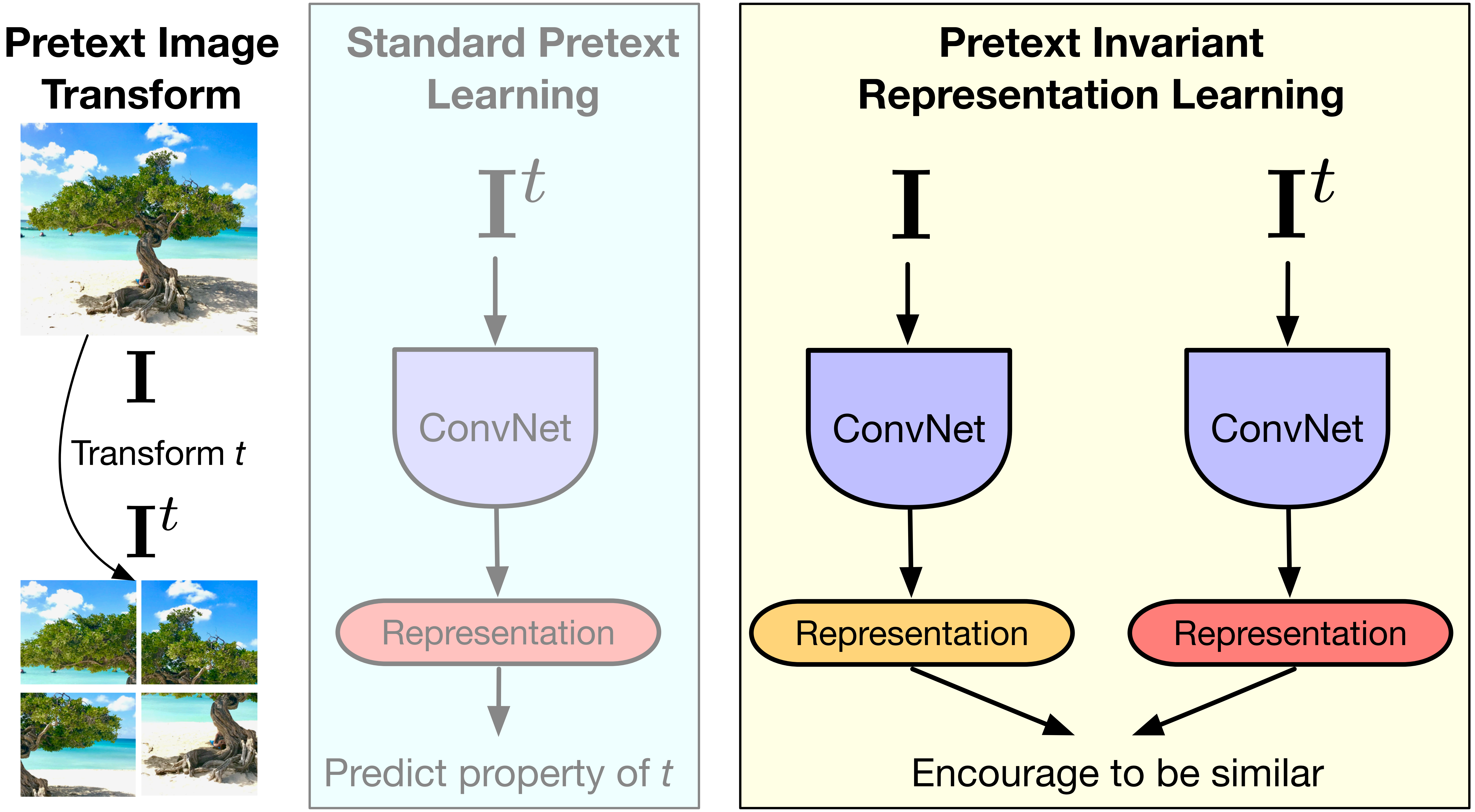


How important has invariance been?

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

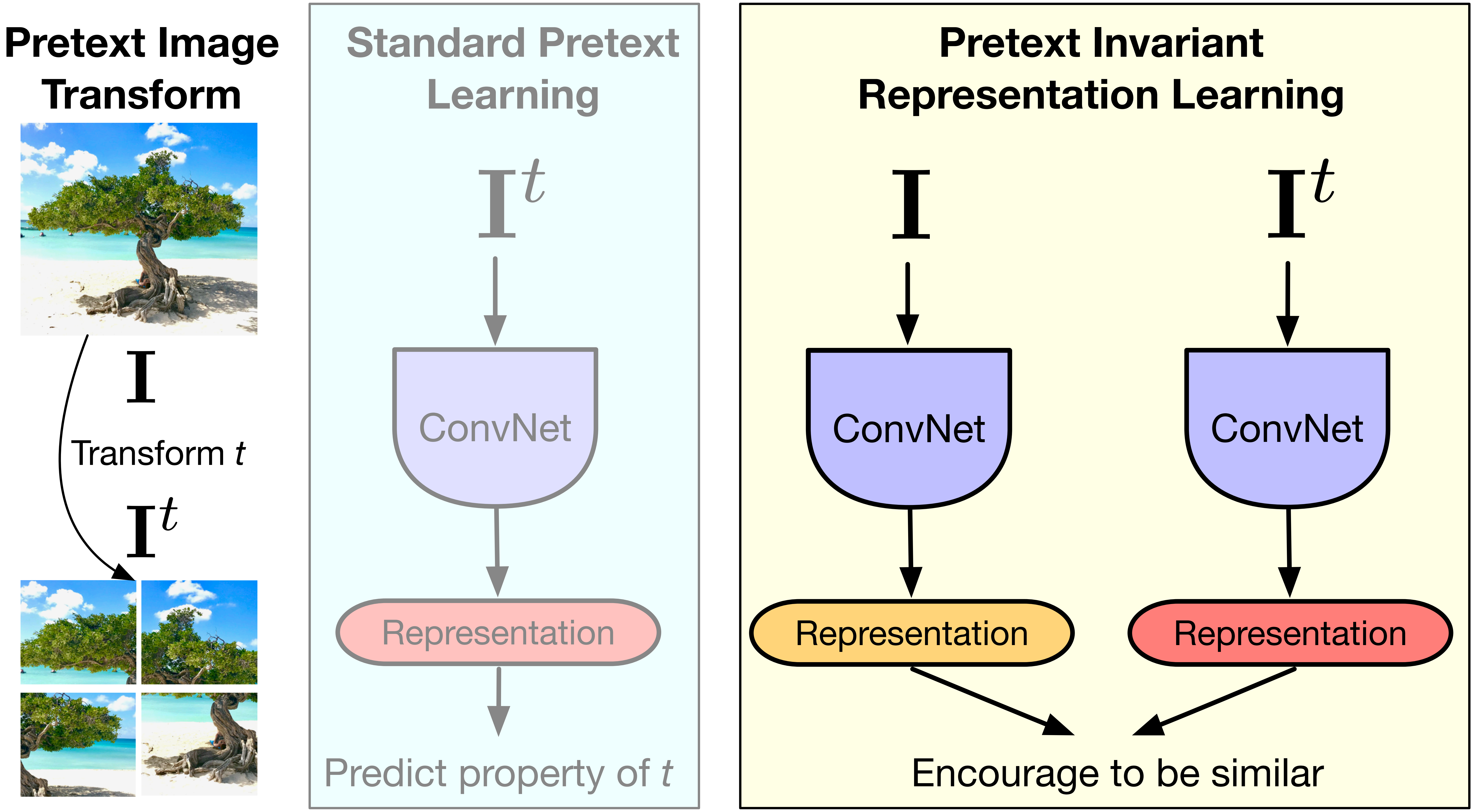
Pretext-Invariant Representation Learning (PIRL)

- Be invariant to **t**



Pretext-Invariant Representation Learning (PIRL)

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

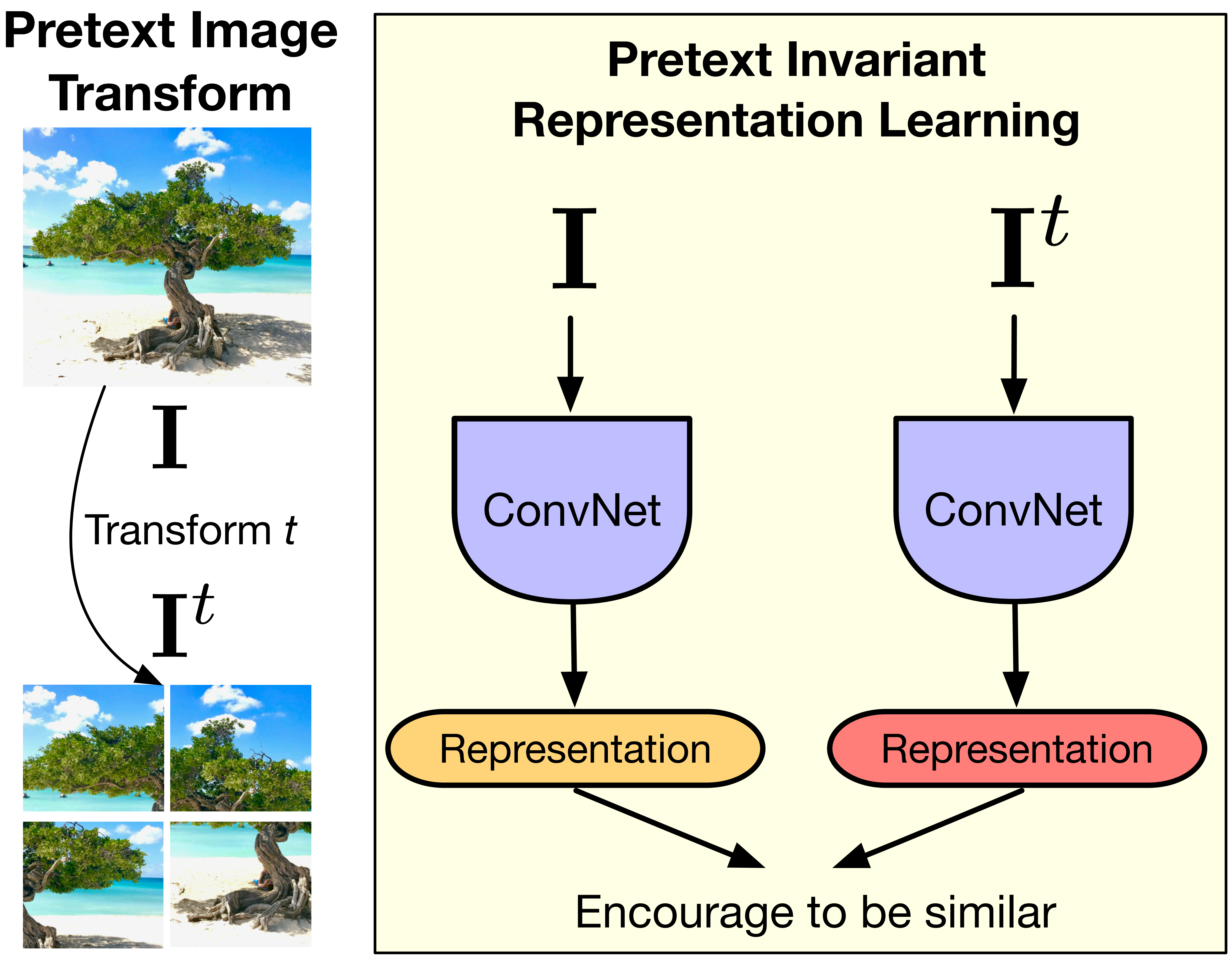


PIRL

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

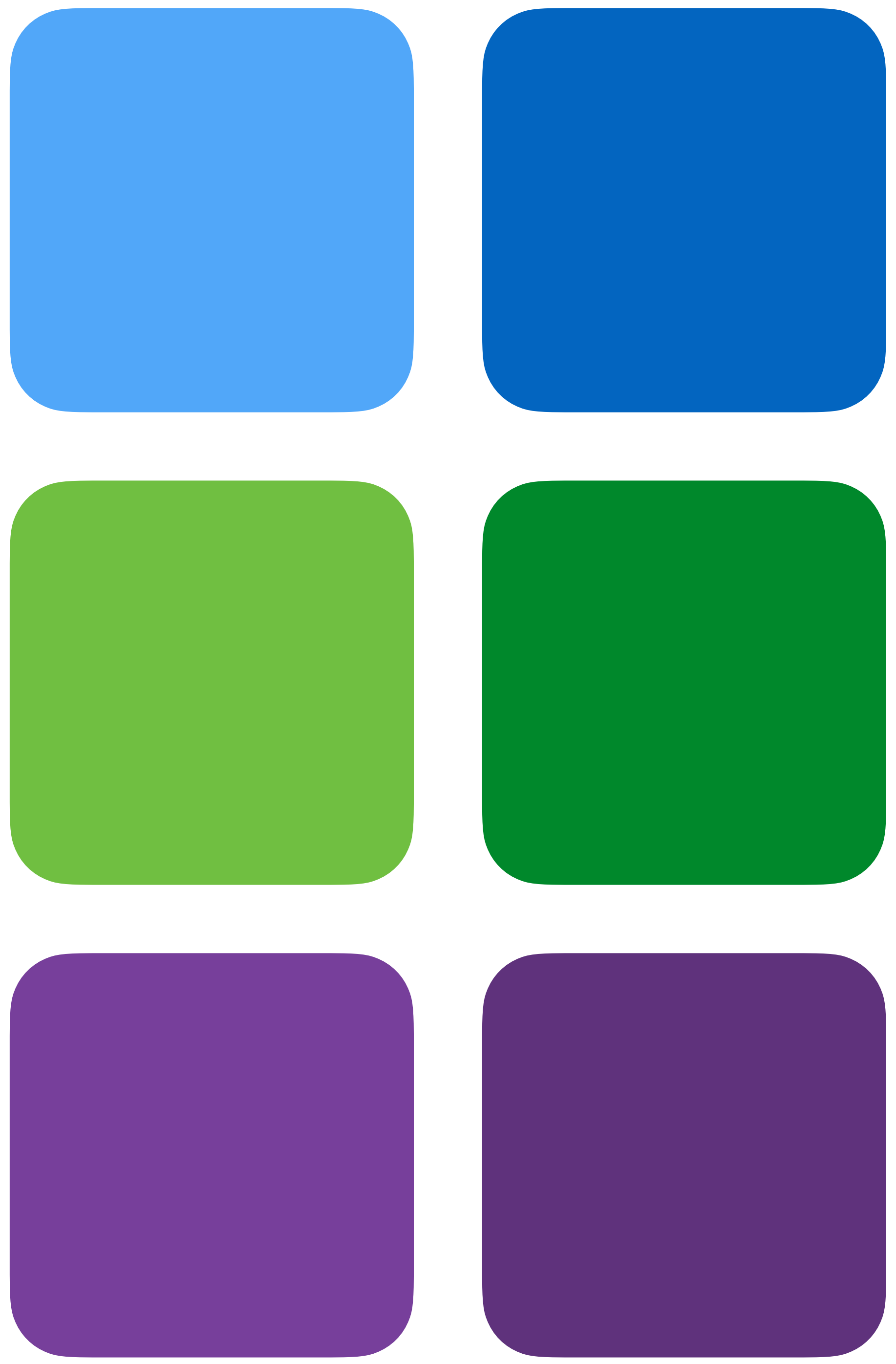
- Use a contrastive loss to enforce similarity of features

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

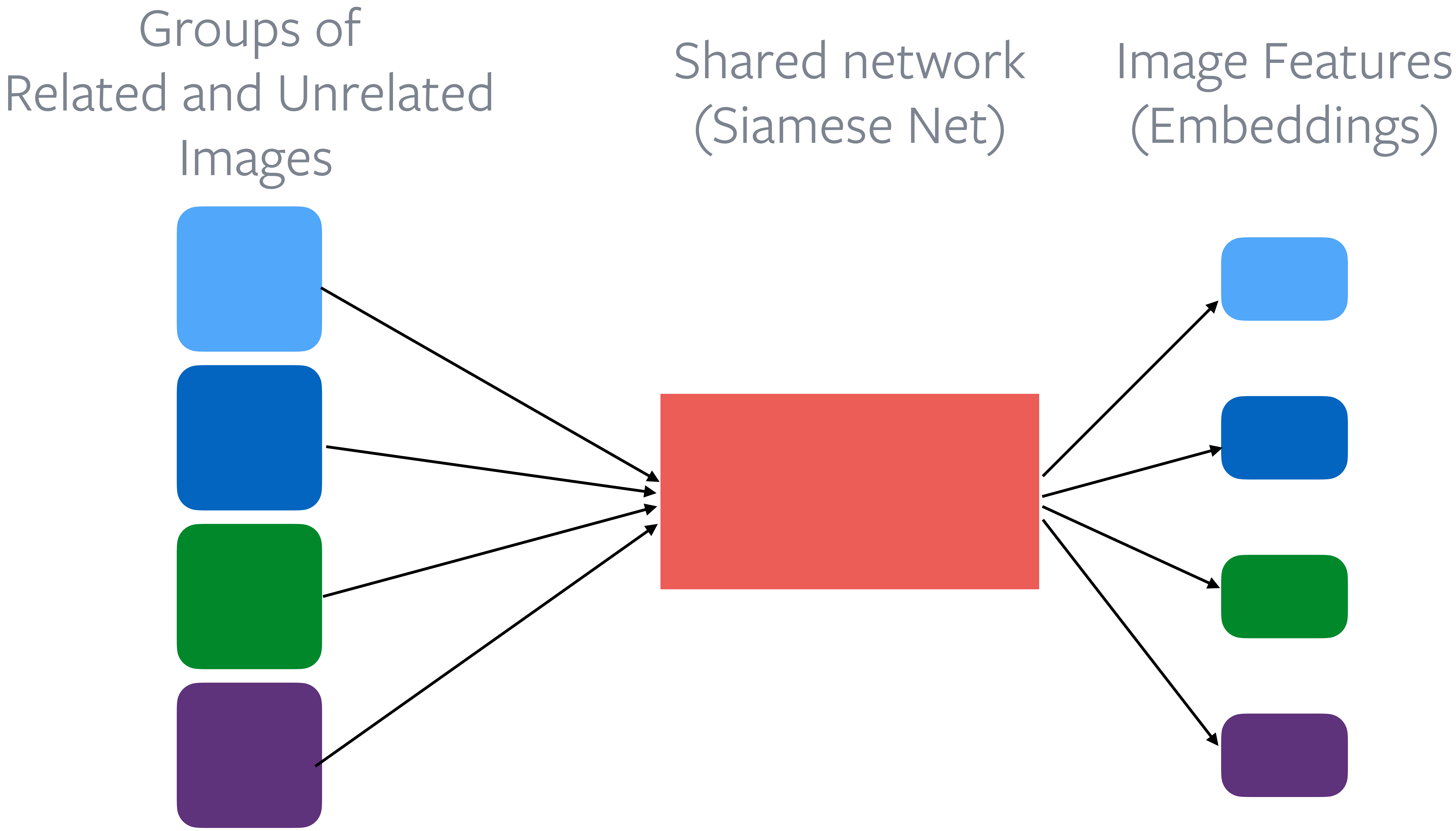


Contrastive Learning

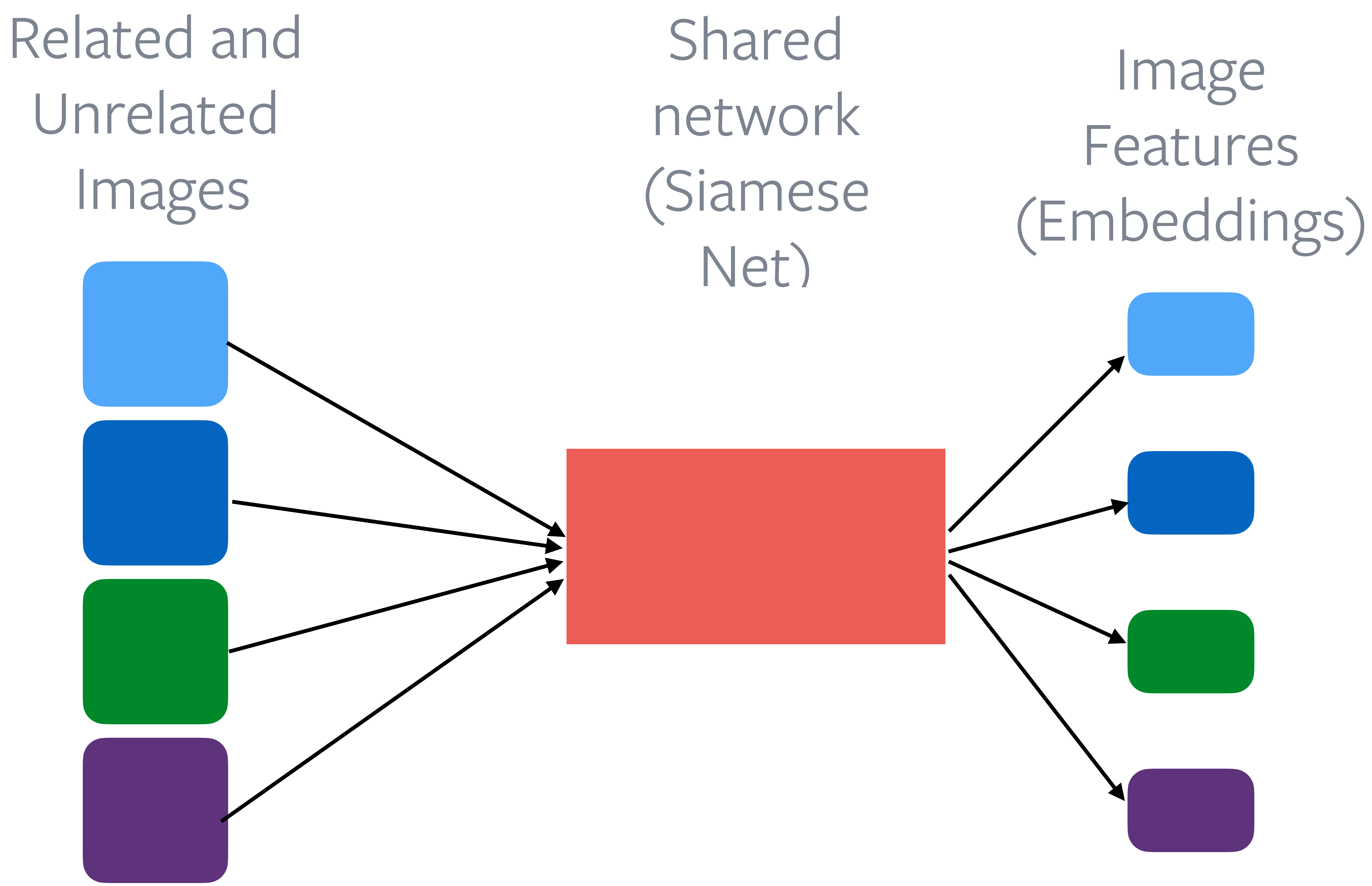
Groups of
Related and Unrelated
Images



Contrastive Learning

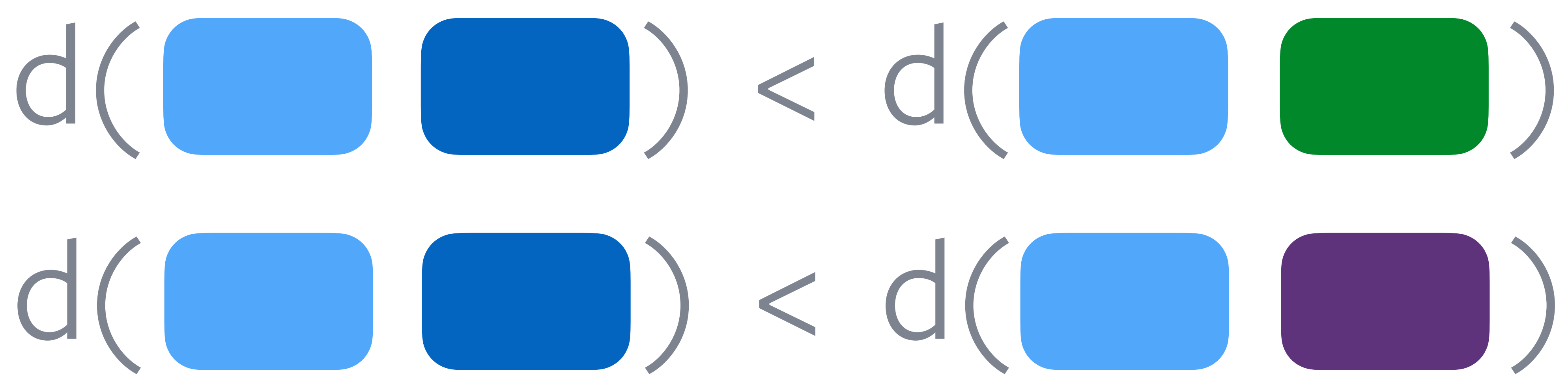


Contrastive Learning



Loss Function

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

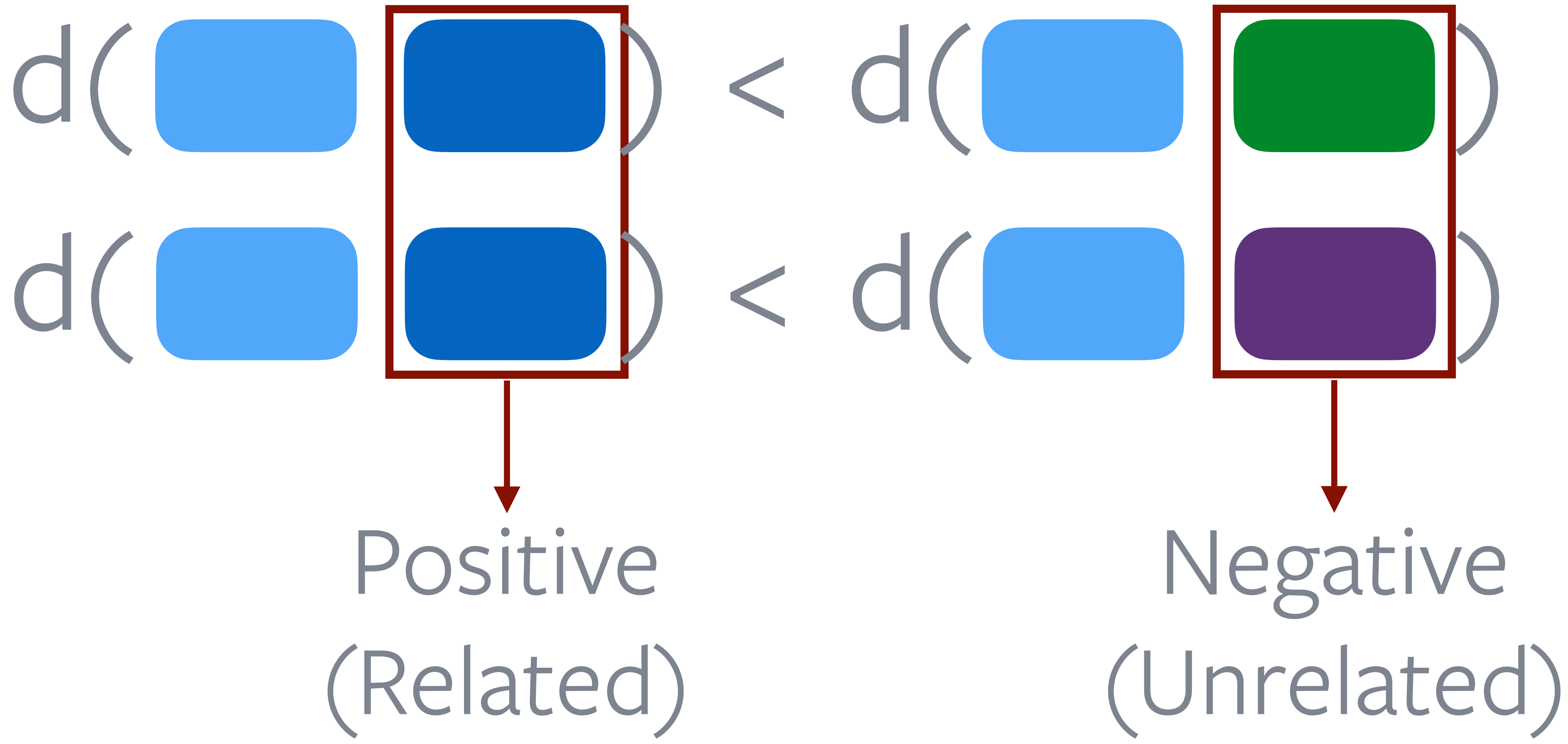


Hadsell et al., 2005, DrLim

Contrastive Loss Function

Loss Function

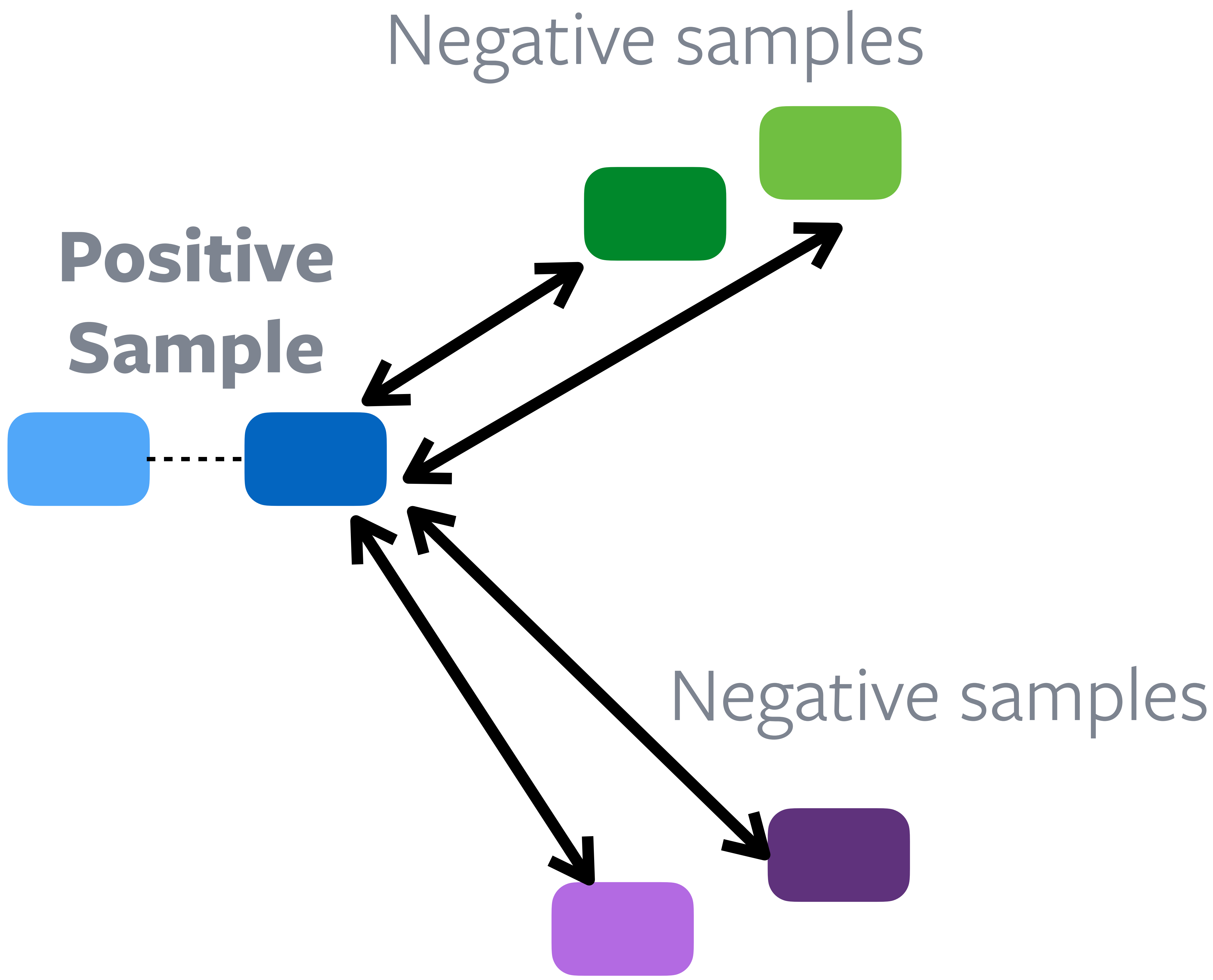
Embeddings from related images should be
closer than embeddings from unrelated images



Good negatives are *very* important in contrastive learning

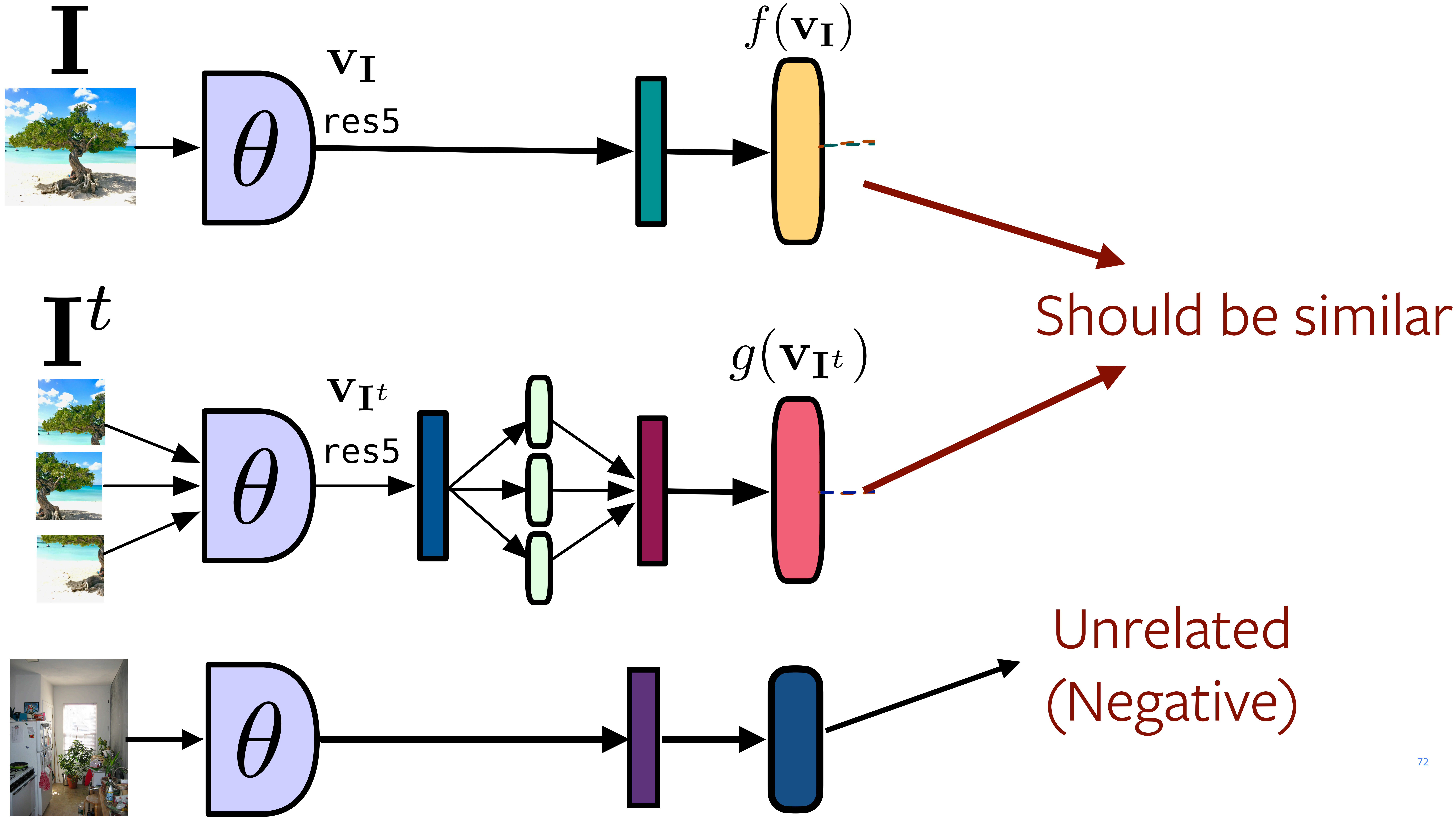
Hadsell et al., 2005, DrLim

Contrastive learning -- what does it do?

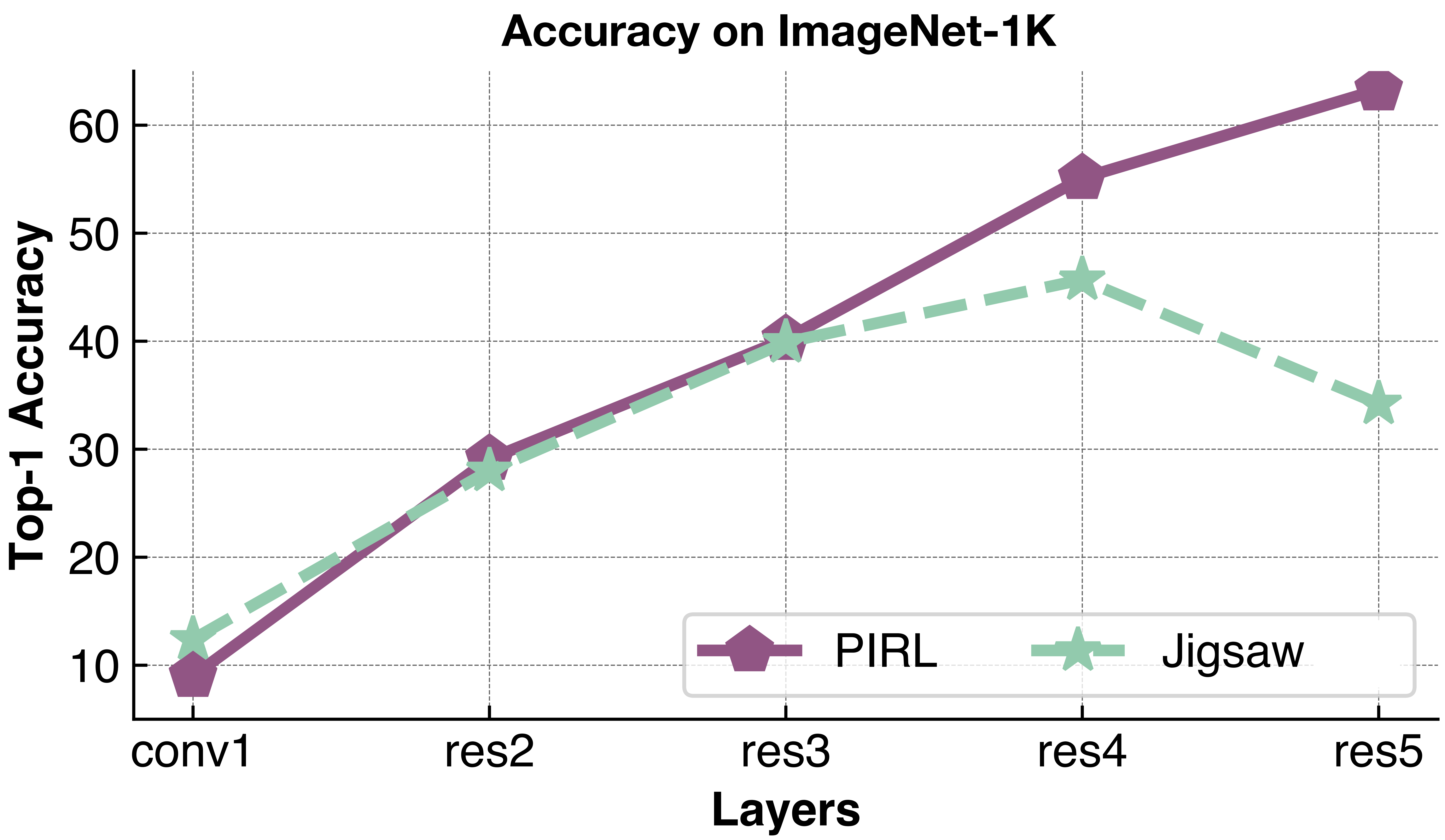


How does this relate to ”pretext” tasks?

PIRL - How it works



Better self-supervised learning objective



Object Detection

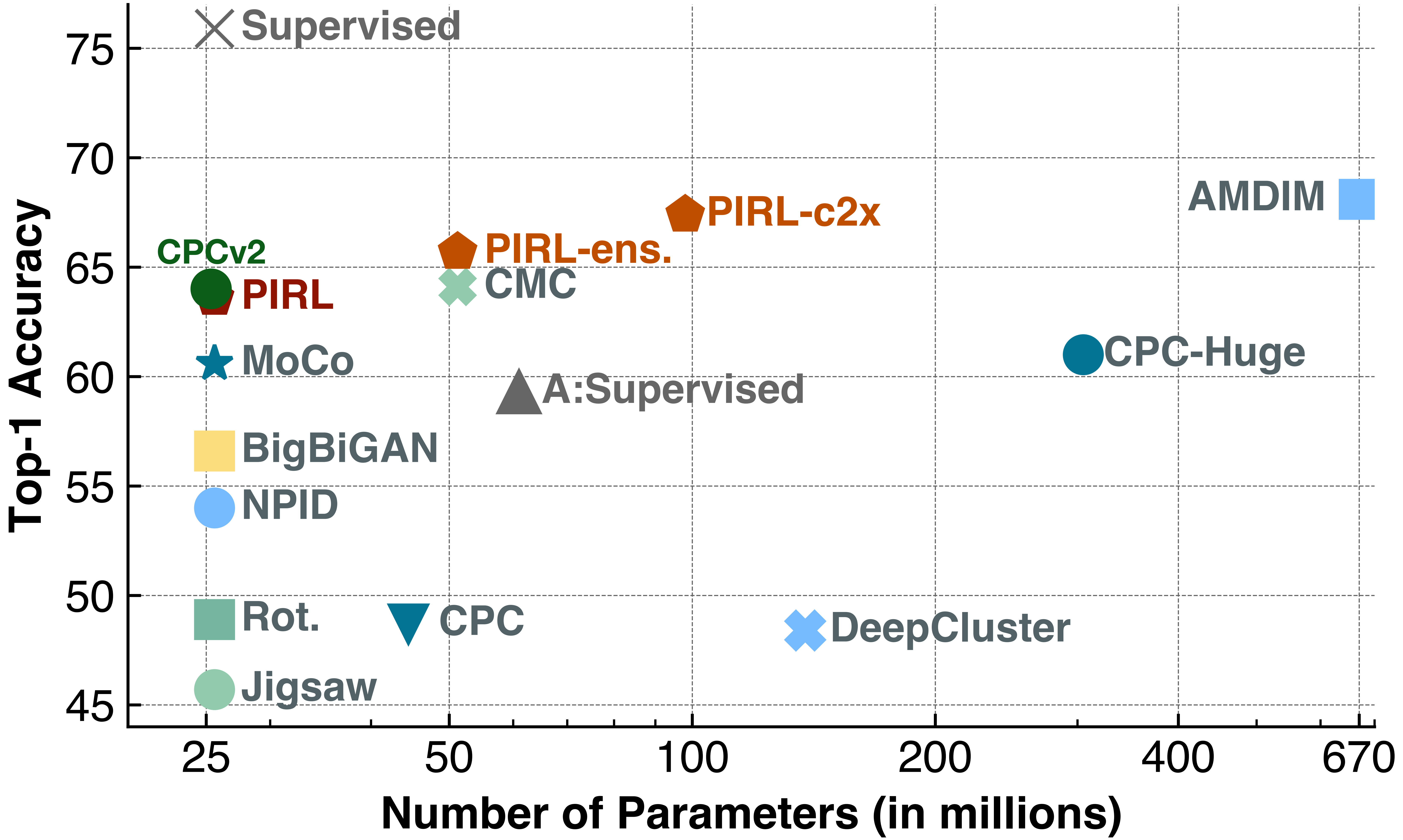
- **Outperforms** ImageNet supervised pre-trained networks
- Full fine-tuning, no bells & whistles
- No extra data, changes in model architecture, fine-tuning schedule

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**	80.7	**59.7**	**44.7**	73.4	**47.0**

+1.4 +2.3 +1.1

Linear Classification

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



Easily Multi-task

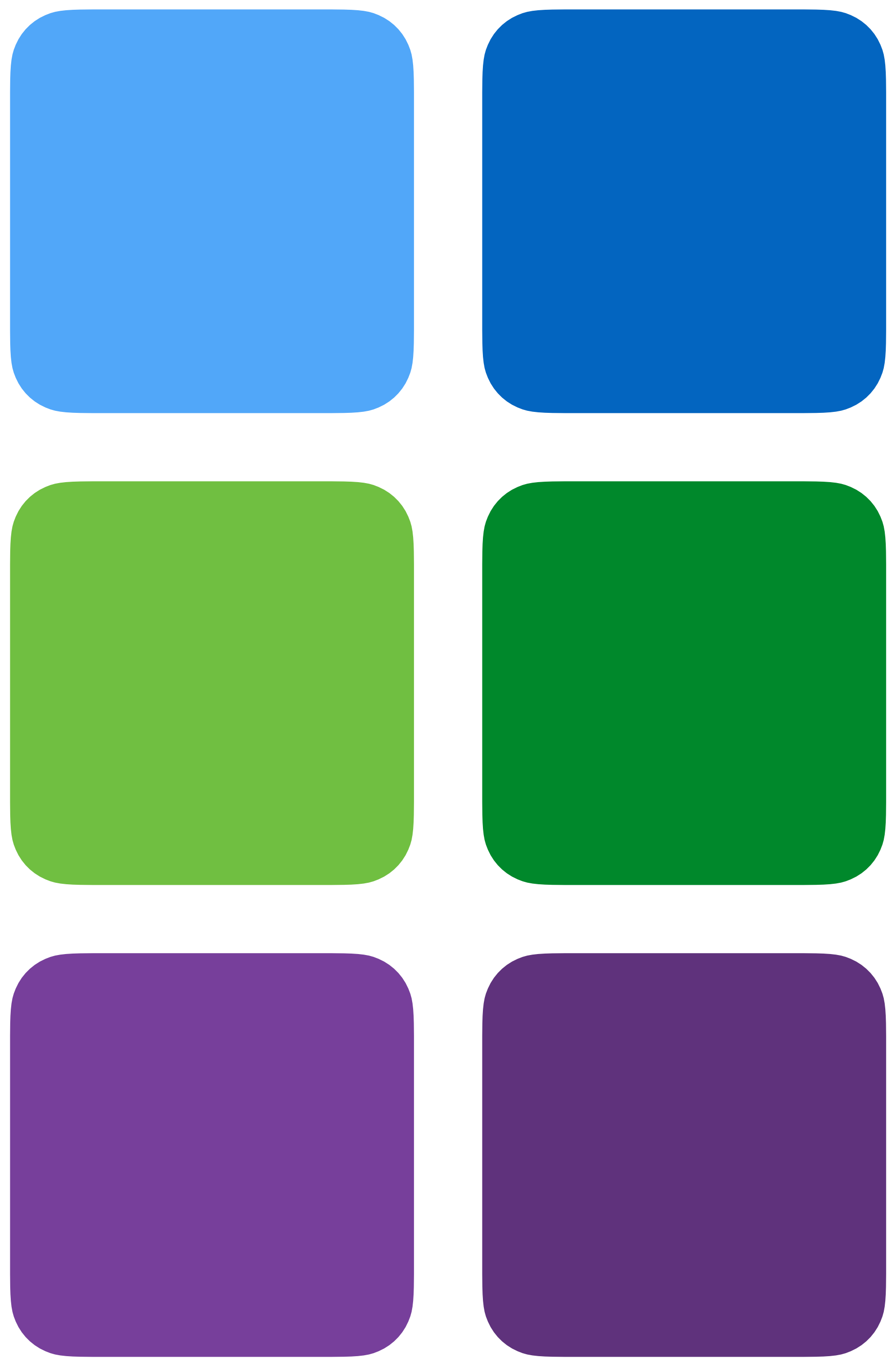
Method	Transfer Dataset			
	ImageNet-1M	VOC07	Places205	iNaturalist
Jigsaw	46.0	66.1	41.4	22.1
Rotation	48.9	63.9	47.6	23
PIRL (Rot)	60.2	77.1	47.6	31.2
PIRL (Jigsaw + Rot)	**63.1**	**80.3**	**49.7**	**33.6**

The rise of contrastive learning

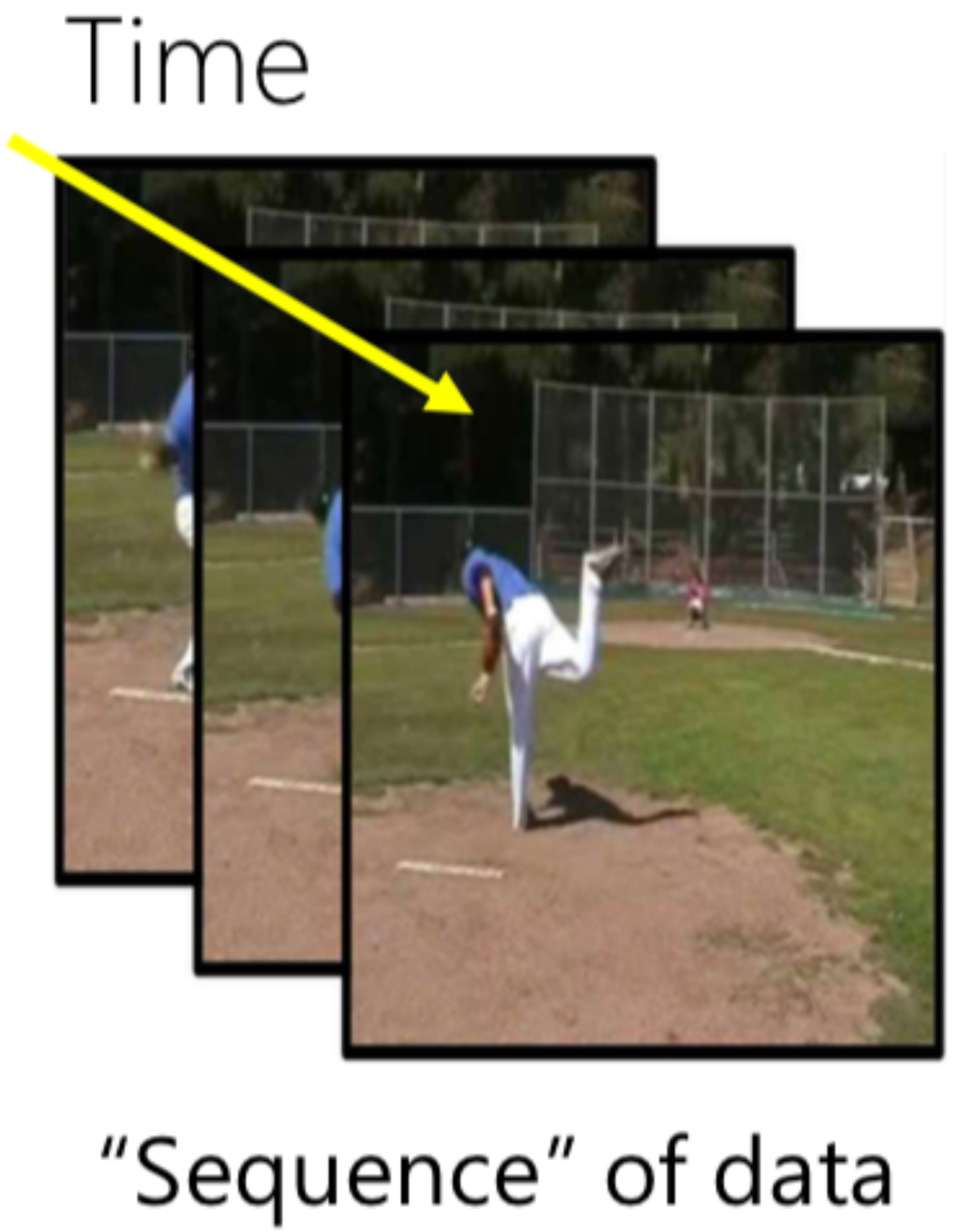
Contrastive Learning

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

Related and Unrelated
Images



Frames of a video

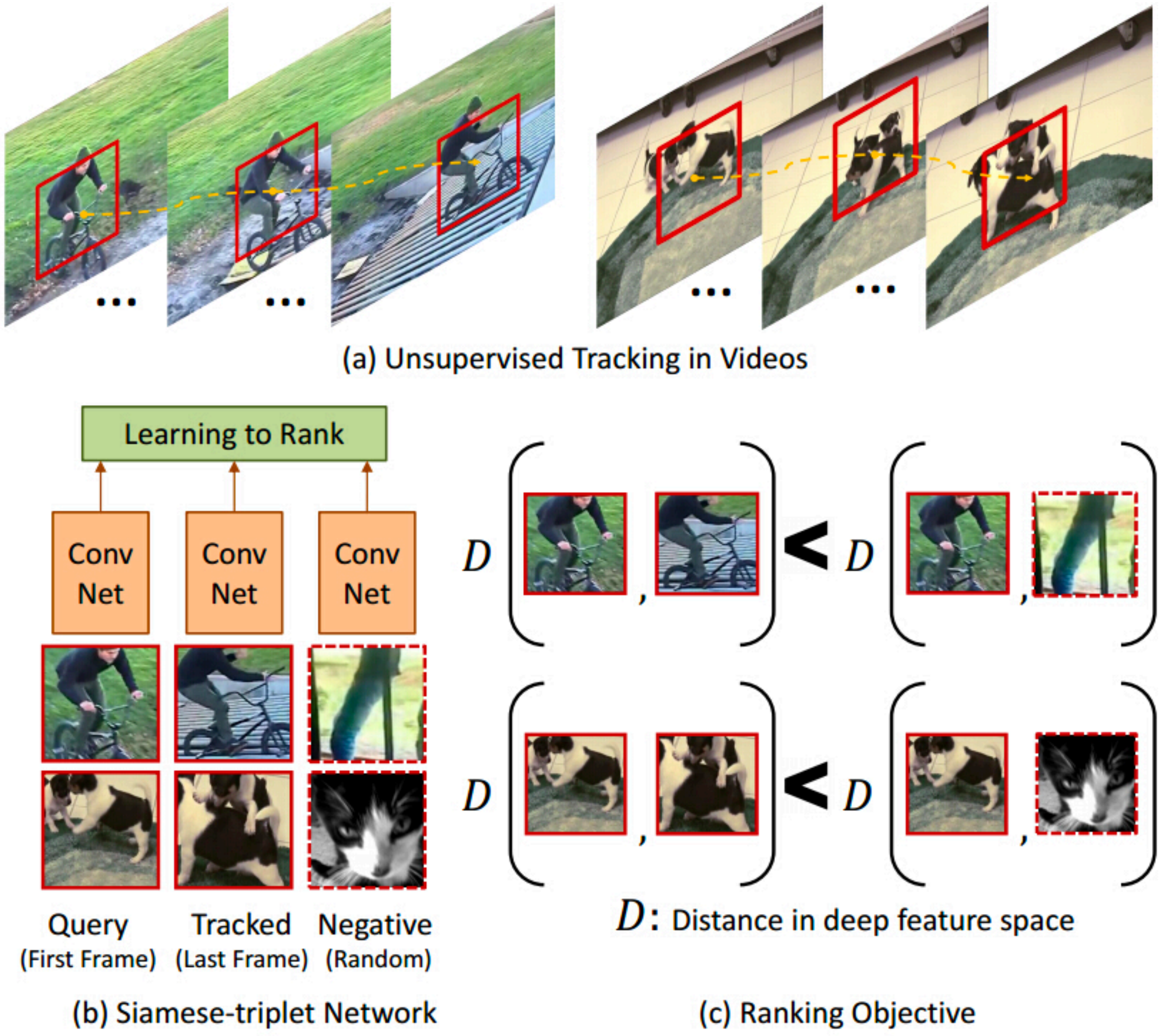


Hadsell et al., 2005, DrLim
van der Oord et al., 2018, CPC

Video & Audio

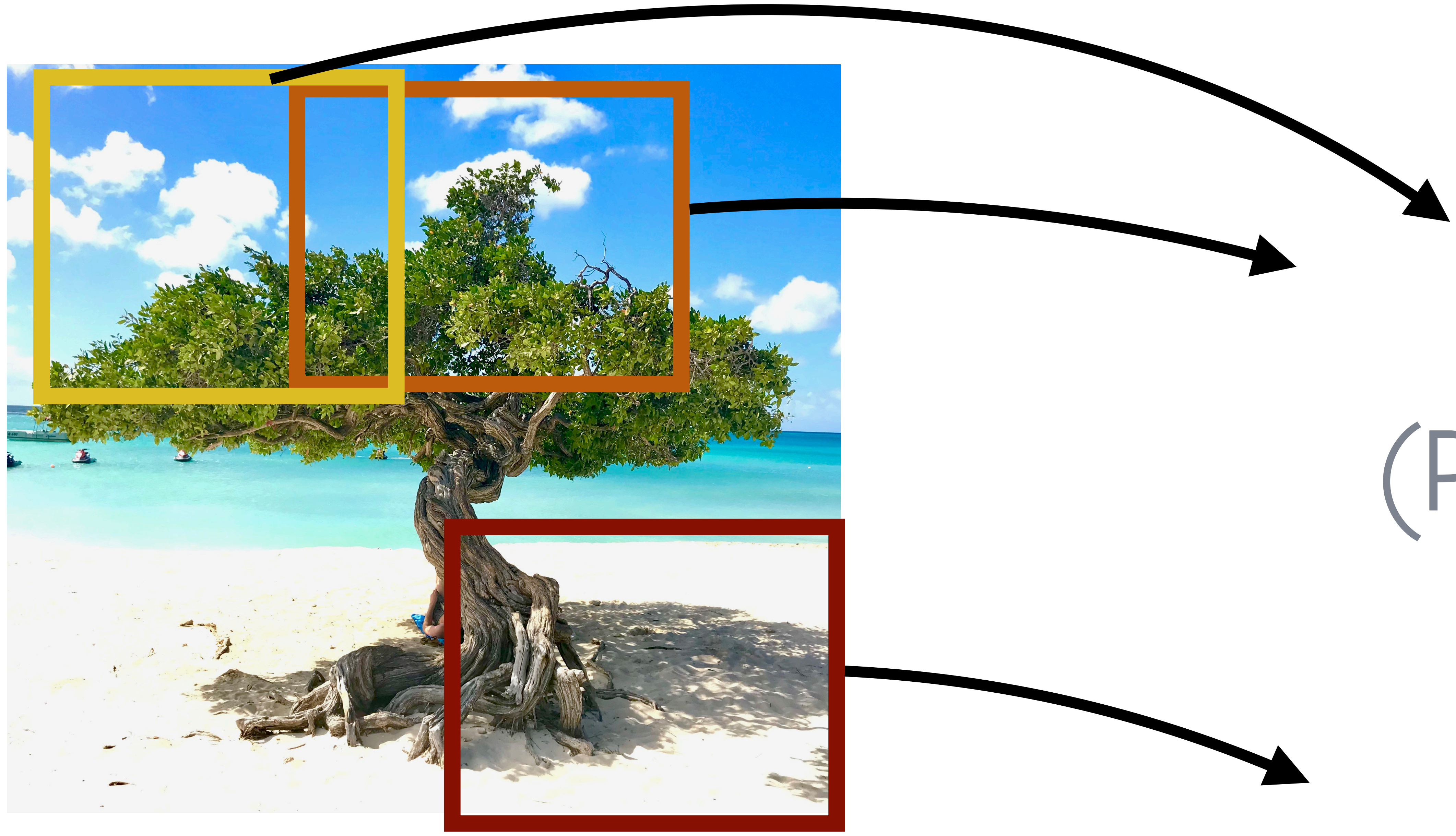
AVID - Morgado et al., ECCV 2020
GDT - Patrick et al., 2020

Tracking Objects



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)

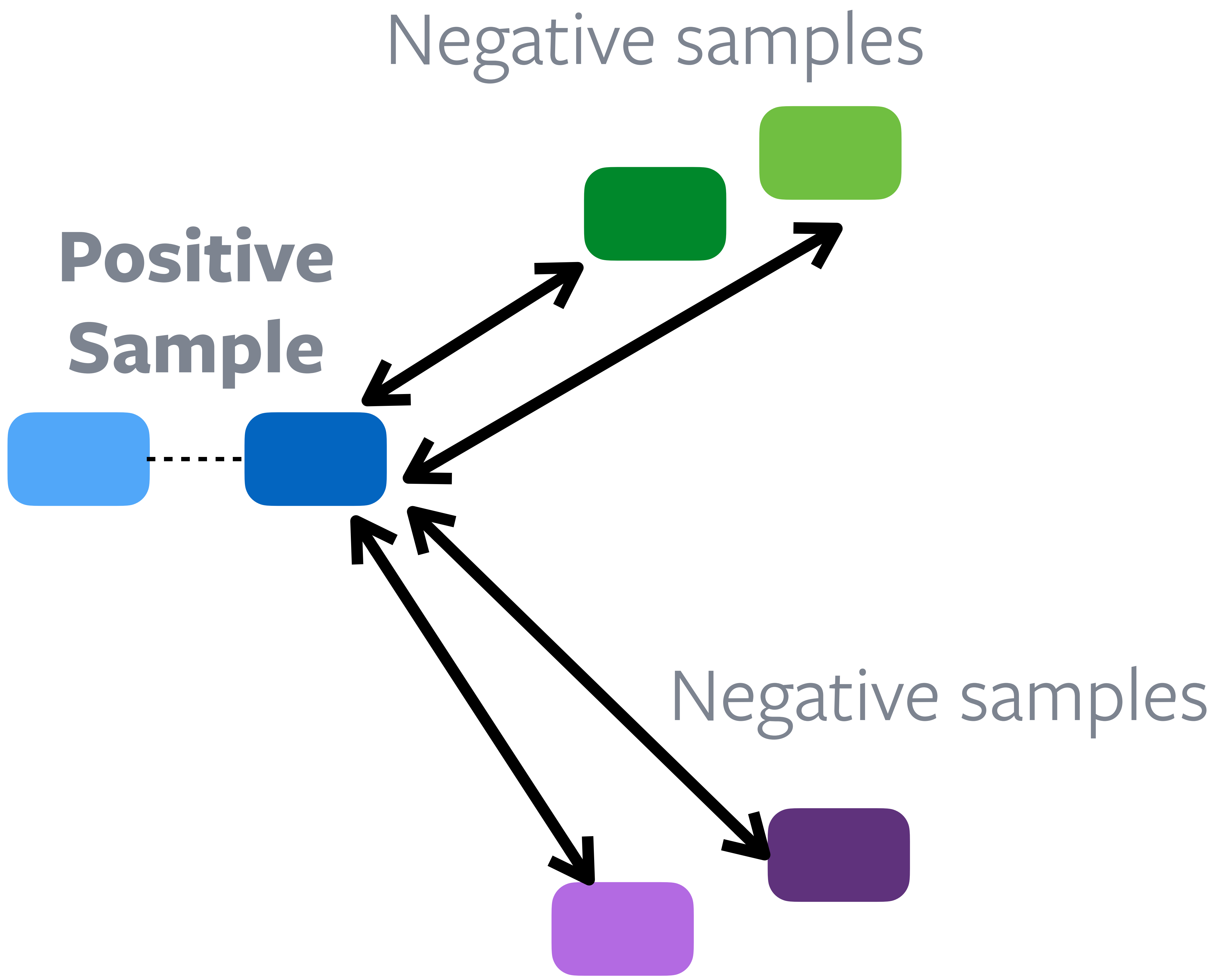
Unrelated
(Negative)

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

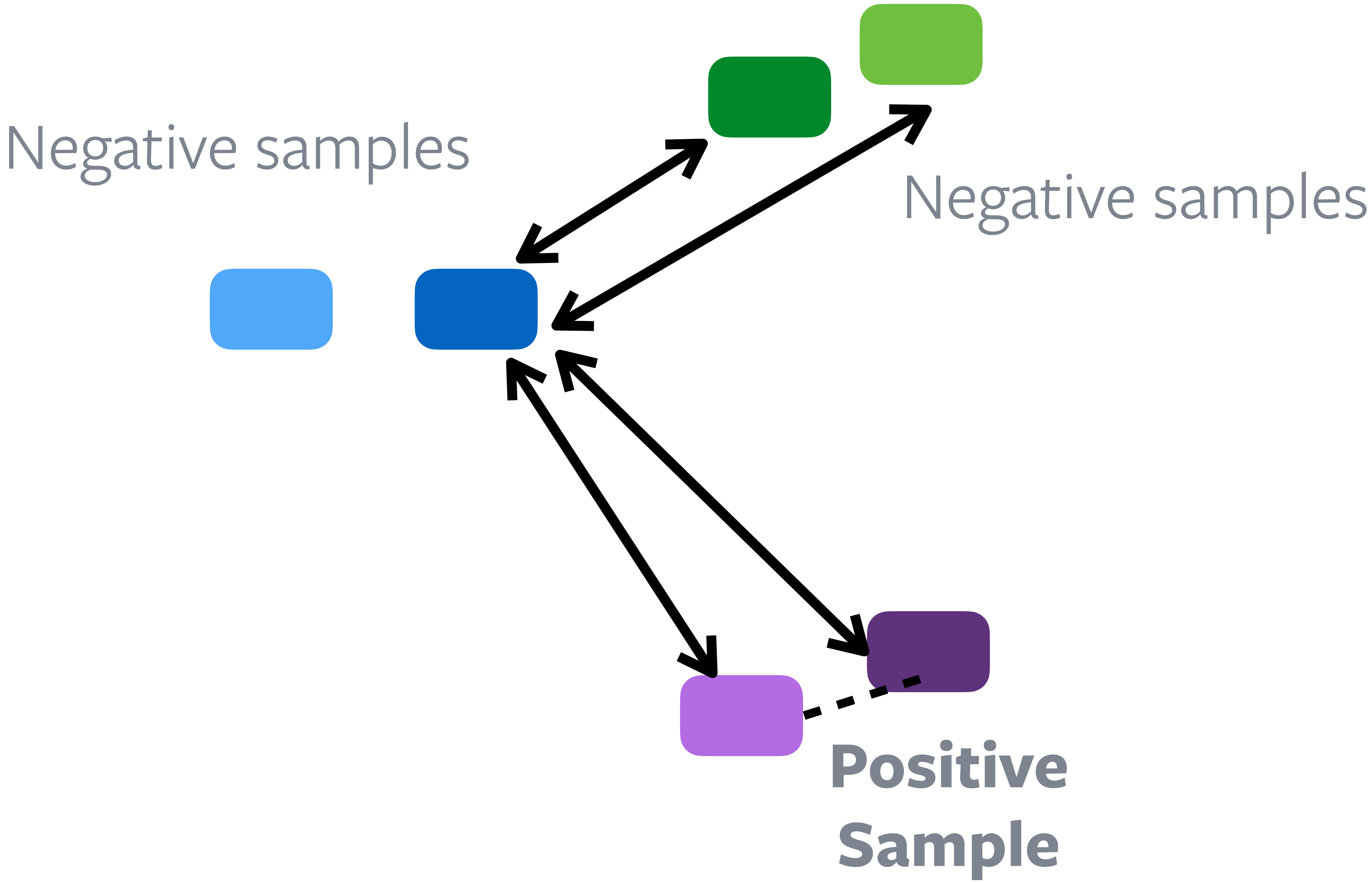
and lots more ....

Is "contrastive" really important?

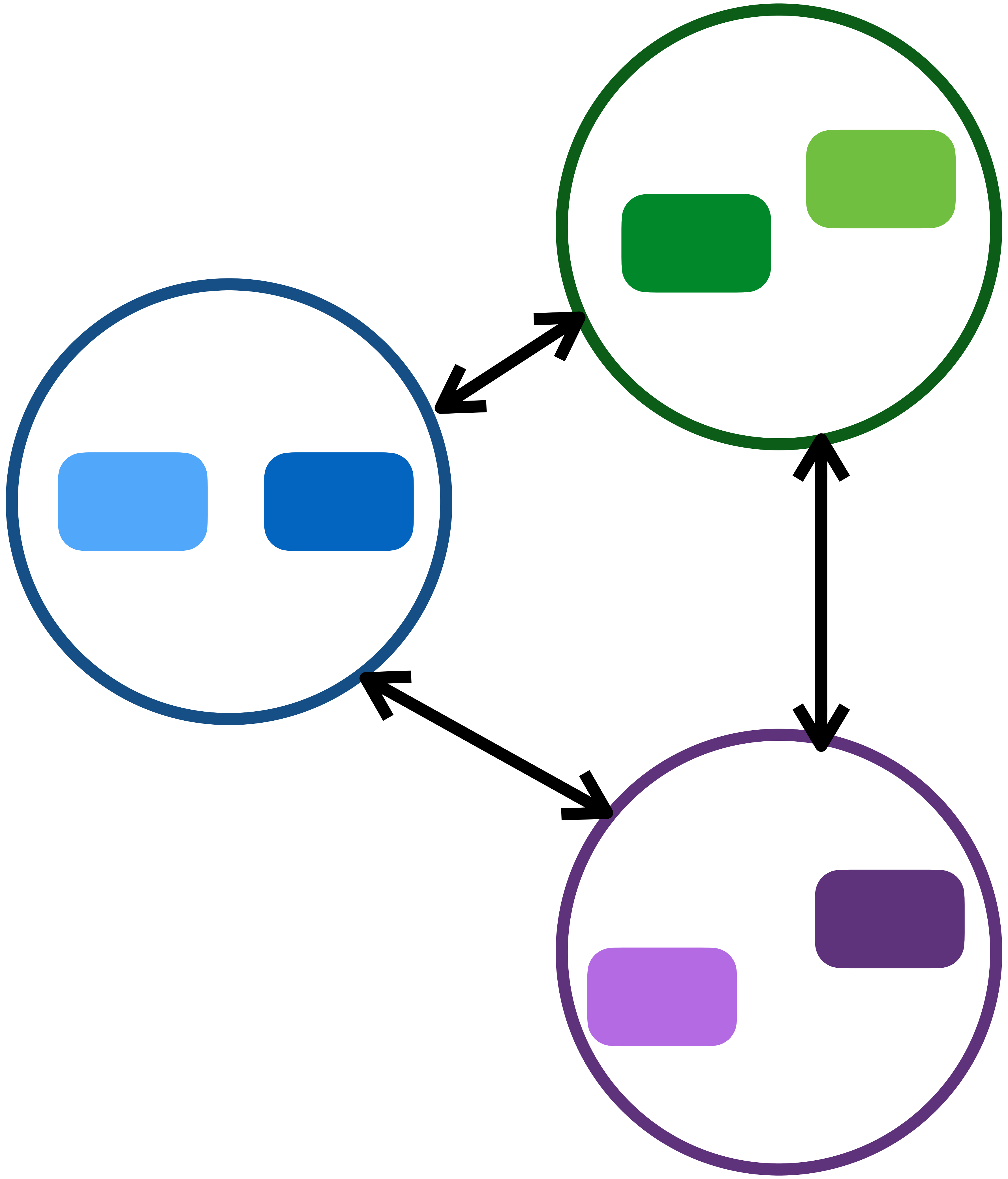
Contrastive learning -- what does it do?



Contrastive learning -- what does it do?

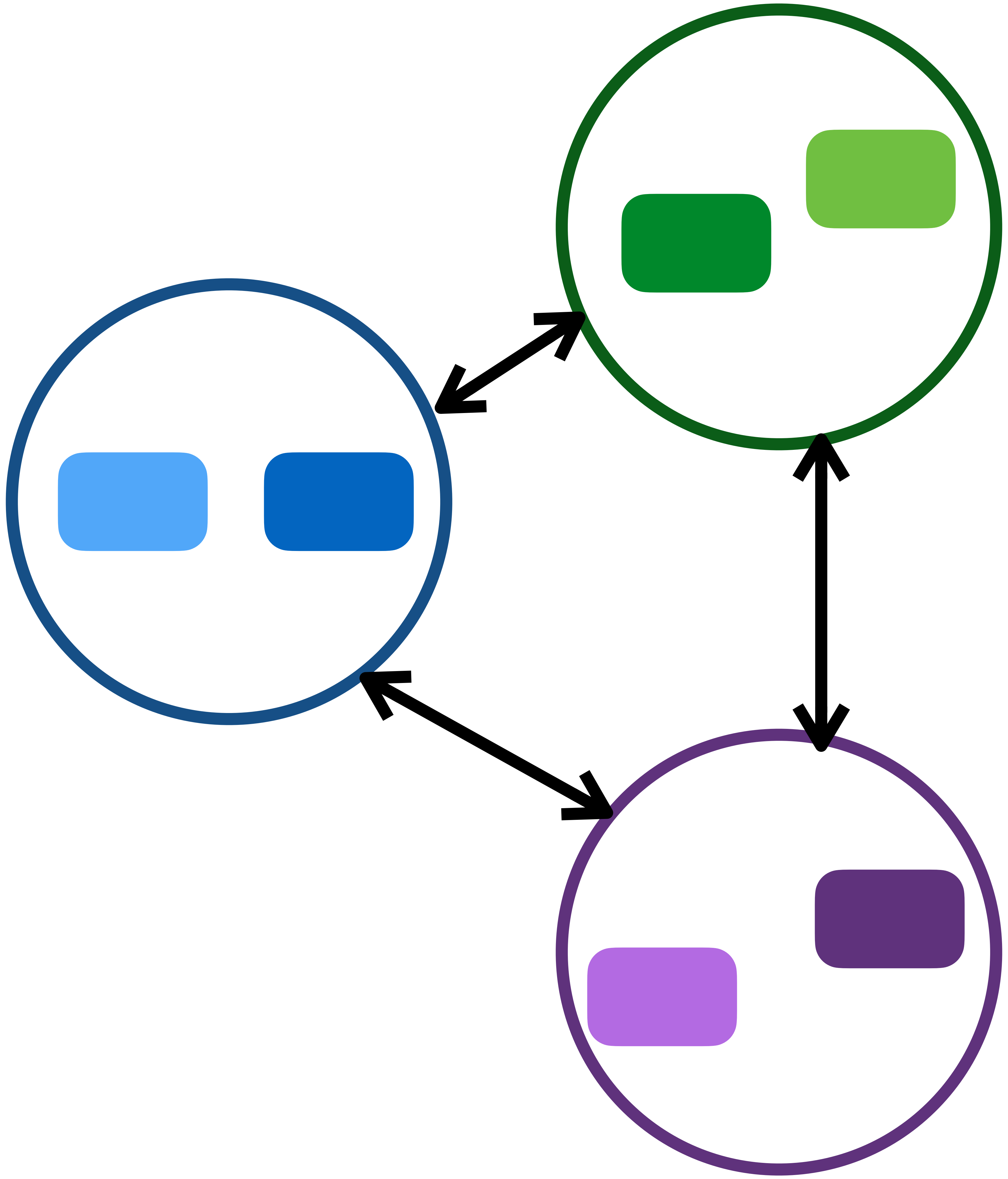


Contrastive learning -- what does it do?



Creates groups
in the feature space

Contrastive learning -- what does it do?



Creates groups
in the feature space

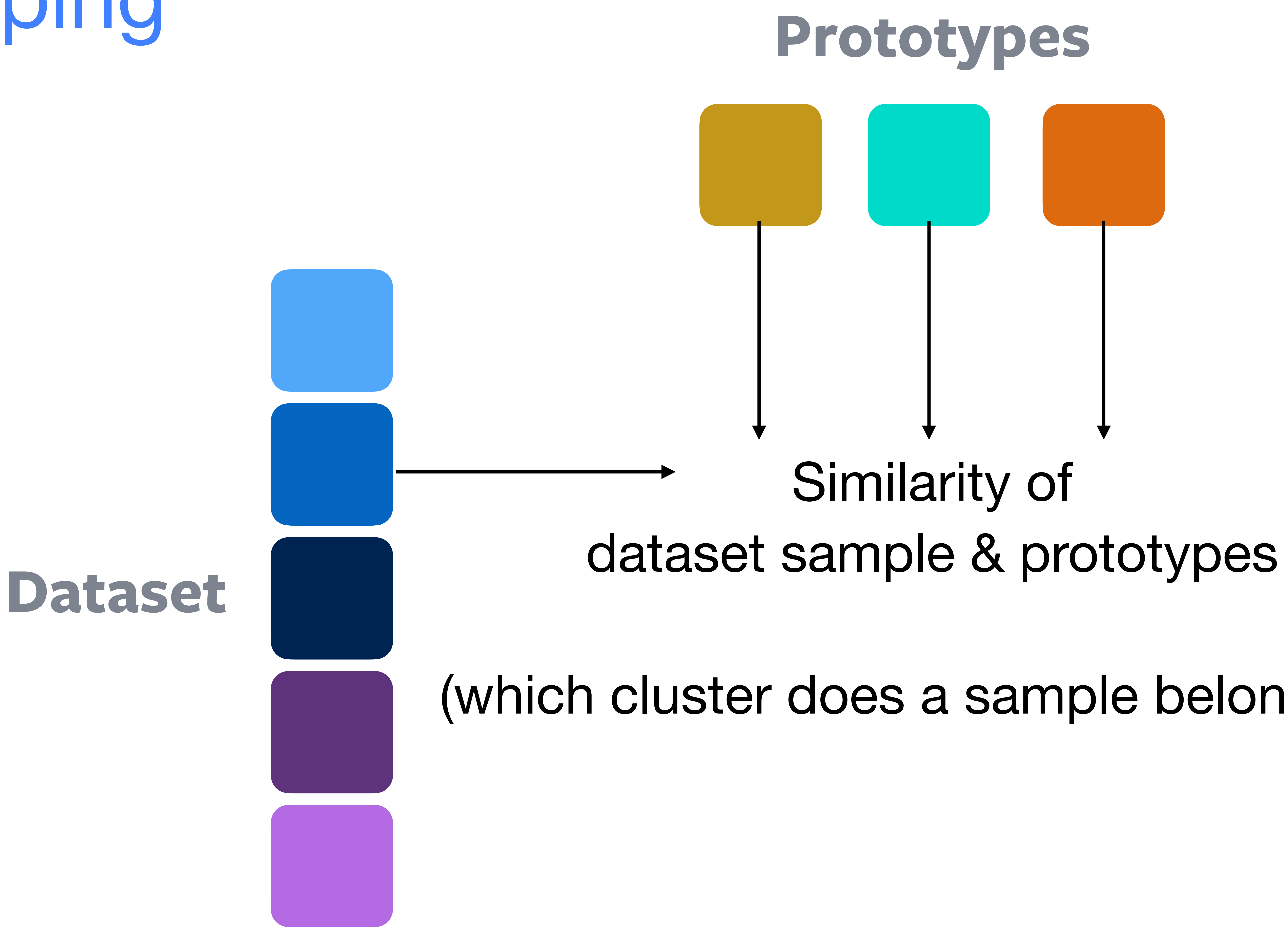
So does **clustering**?!

Swapping Assignments between Views (SwAV)

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin

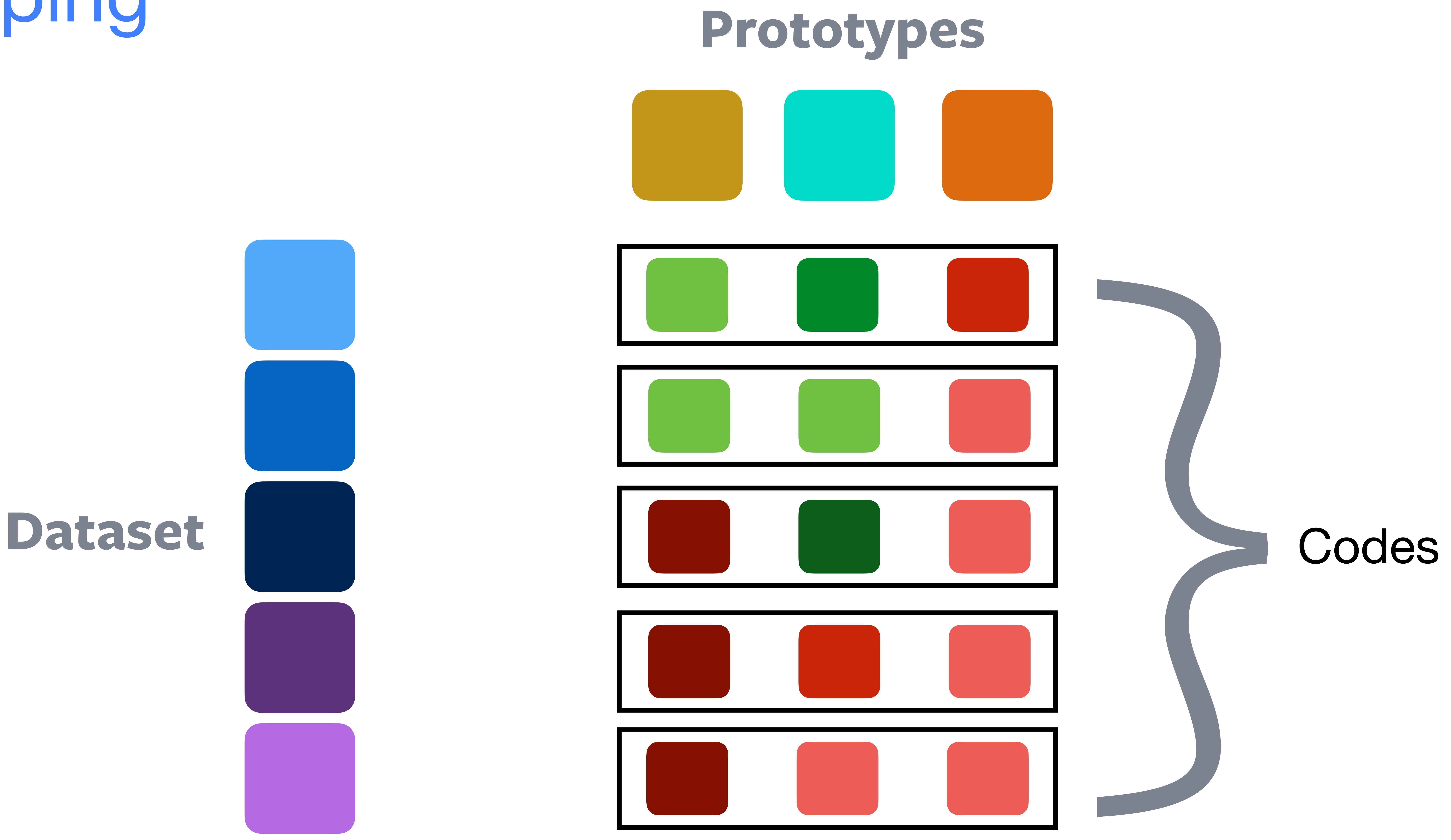
Grouping

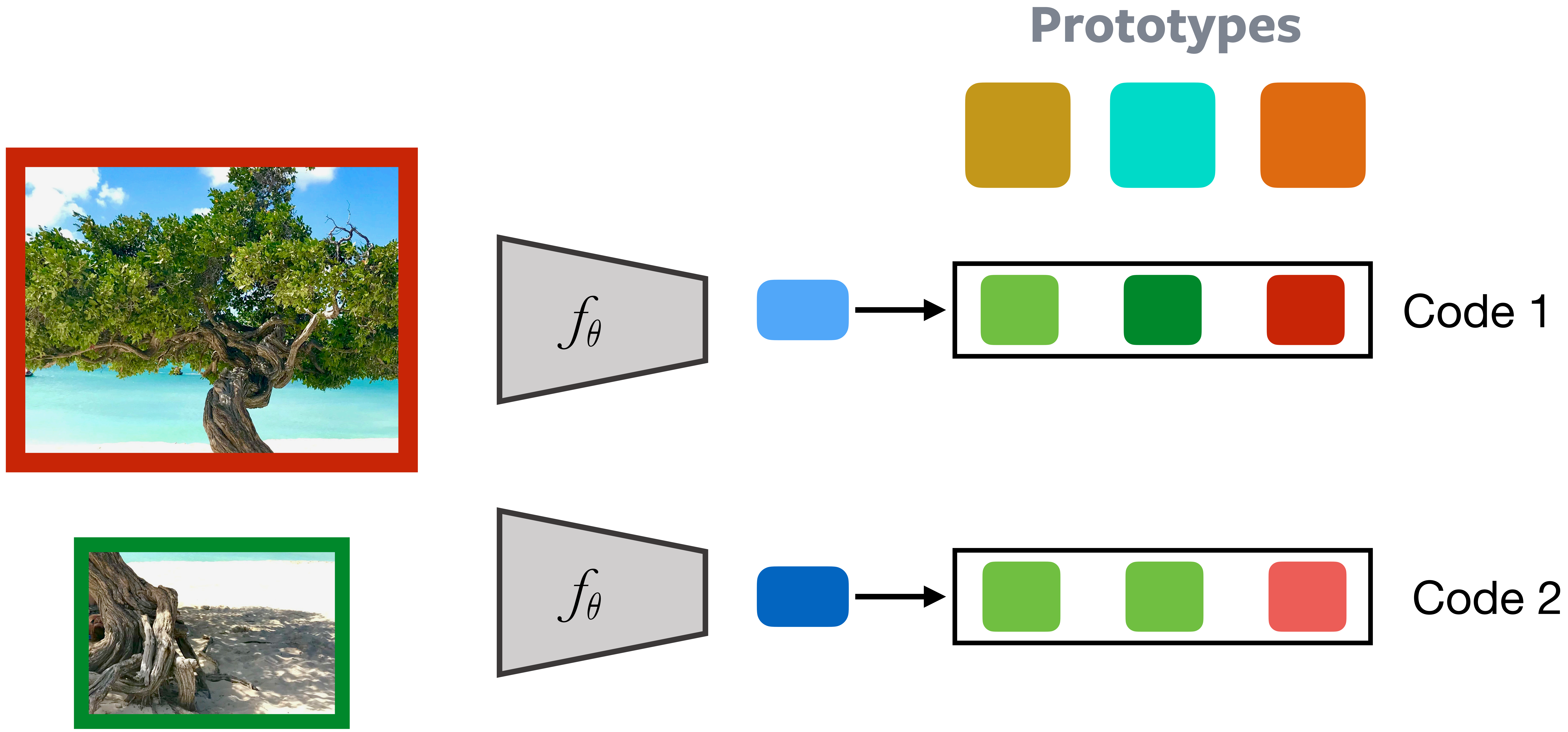


(which cluster does a sample belong to?)

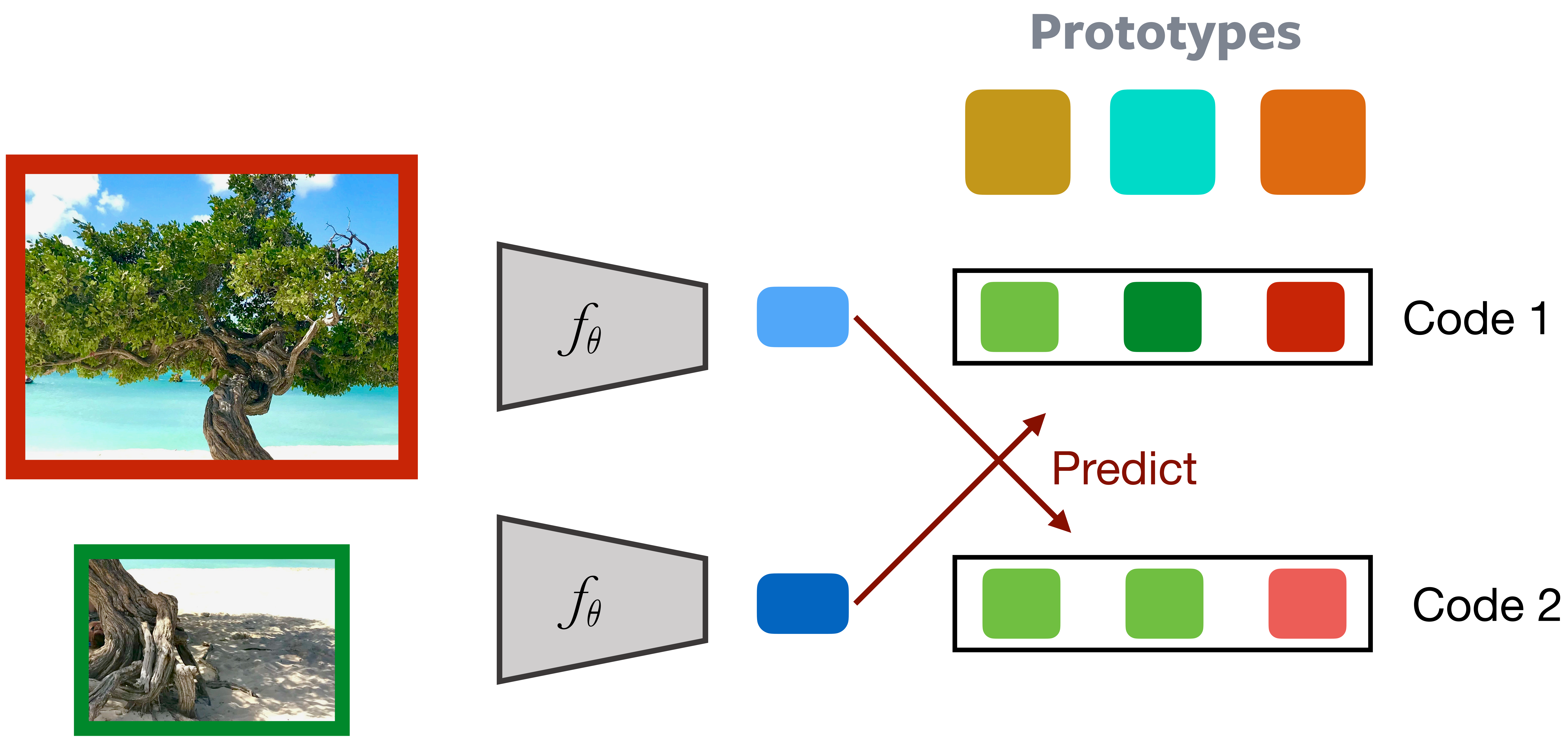
See also - *SeLa* by Asano et al., 2019

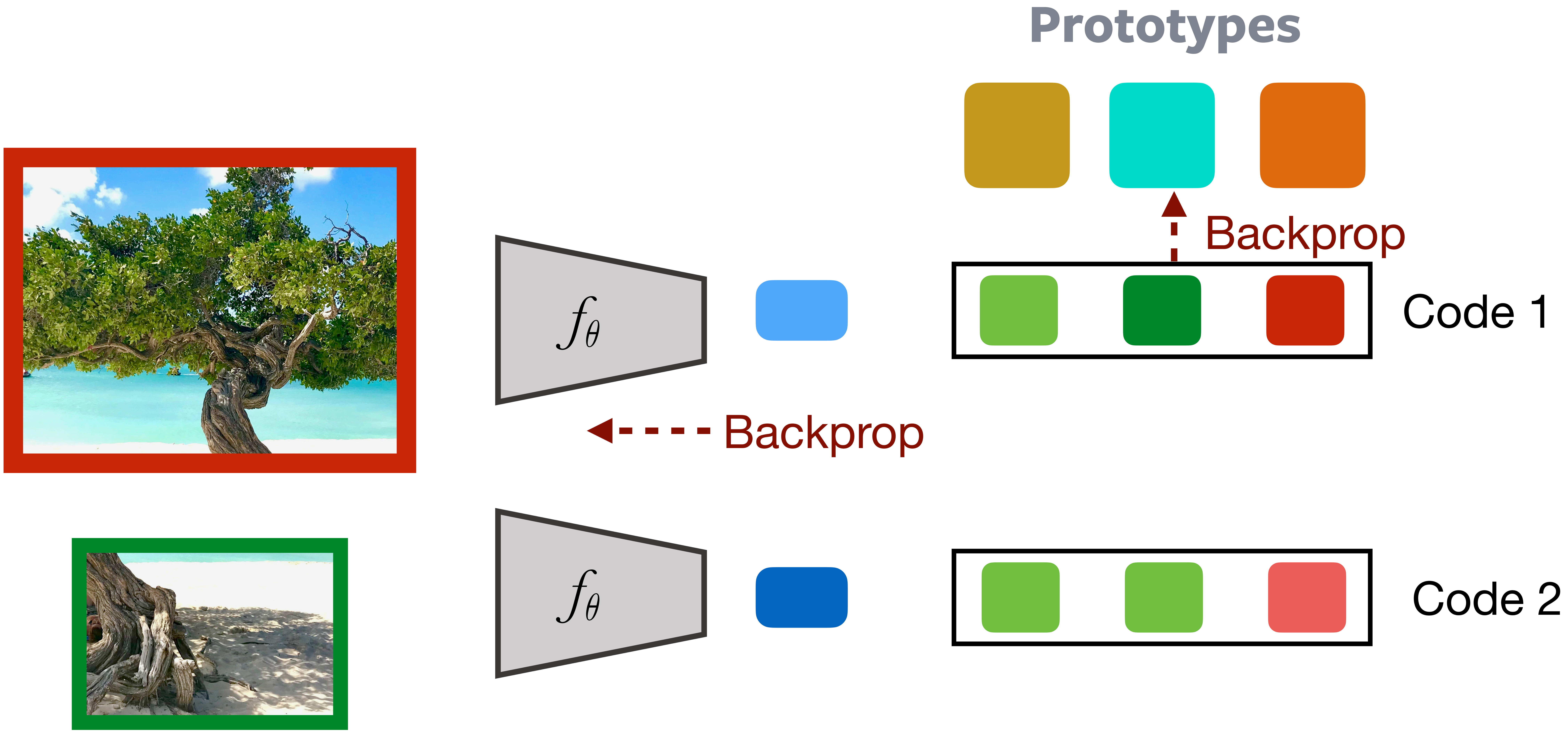
Grouping





Prototypes
f_θ
f_θ
Code 1
Code 2


Prototypes
f_θ
f_θ
Code 1
Predict
Code 2



Not contrastive!

Key Results

	Linear Classifier (Fixed Features)			Detection	
	ImageNet	Places	iNaturalist	VOC07+12	COCO
Supervised	76.5	53.2	46.7	81.3	40.8
Prior self-supervised	71.1 (-5.4)	52.1	38.9	82.5	42.0
SwAV	75.3 (-1.2)	**56.7**	**48.6**	**82.6**	**42.1**

Practical advantages of SwAV

- Trains on 4-8 GPUs
- **Faster convergence** than prior work (SimCLR, MoCov2)
  - Smaller compute requirements.
  - **2x faster** than MoCo-v2 on 8 GPUs
    - 72% after 100h vs. 71% after 200h

- Better results

Code & Models - https://github.com/facebookresearch/swav

PyTorch Lightning implementation on the way

Combining clustering with contrastive learning

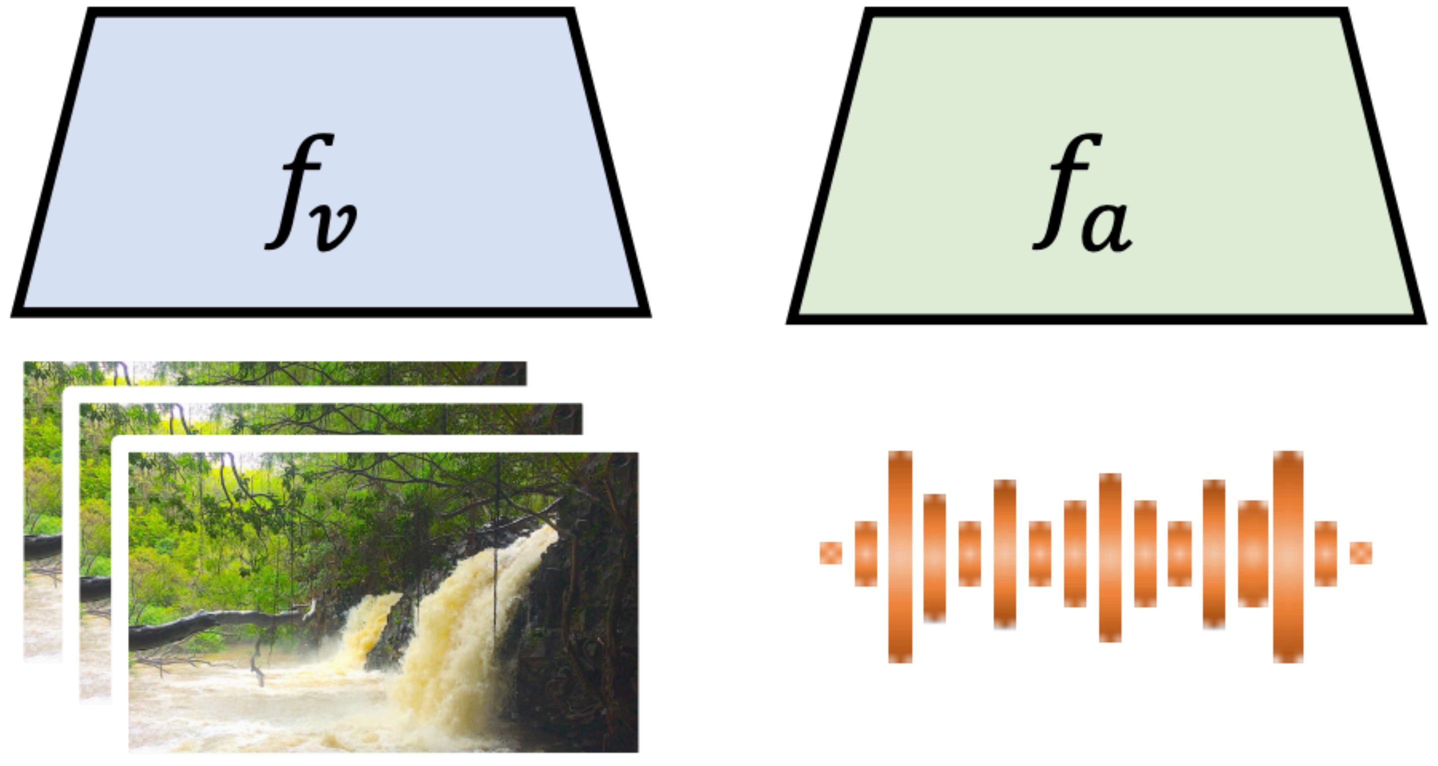
Audio Visual Instance Discrimination with Cross Modal Agreement (AVID + CMA)

Pedro Morgado, Nuno Vasconcelos, Ishan Misra

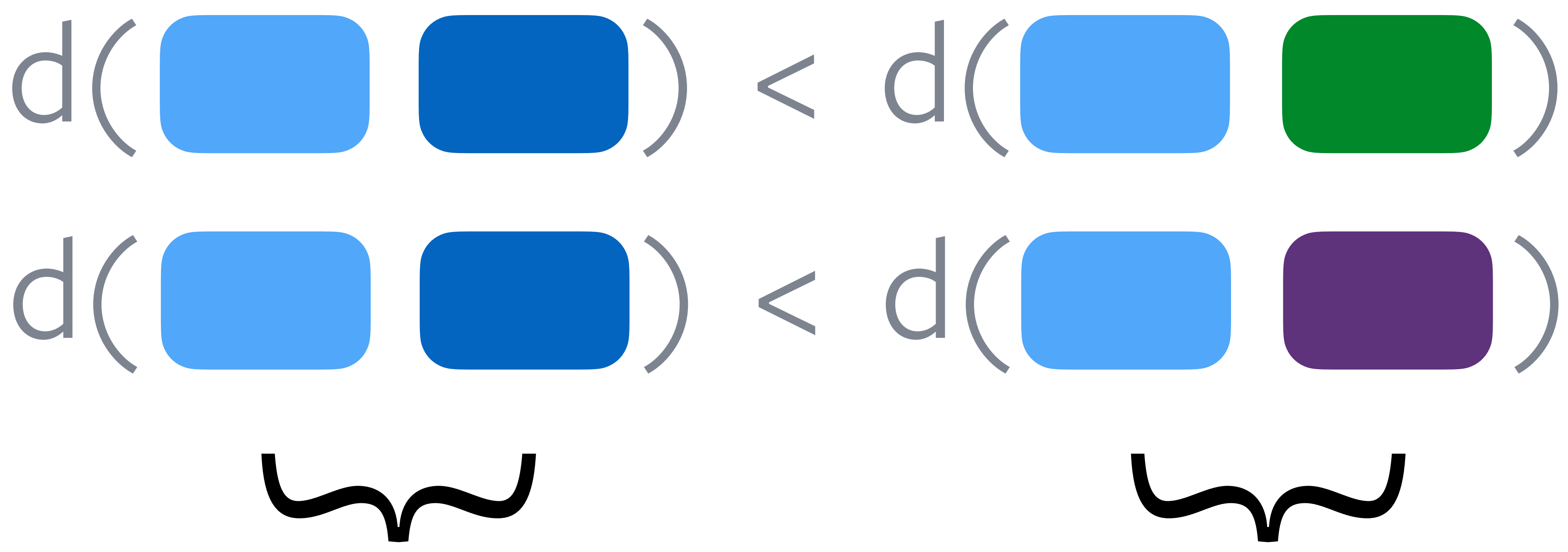
  

https://github.com/facebookresearch/AVID-CMA

Contrastive (Audio Video Instance Discrimination)



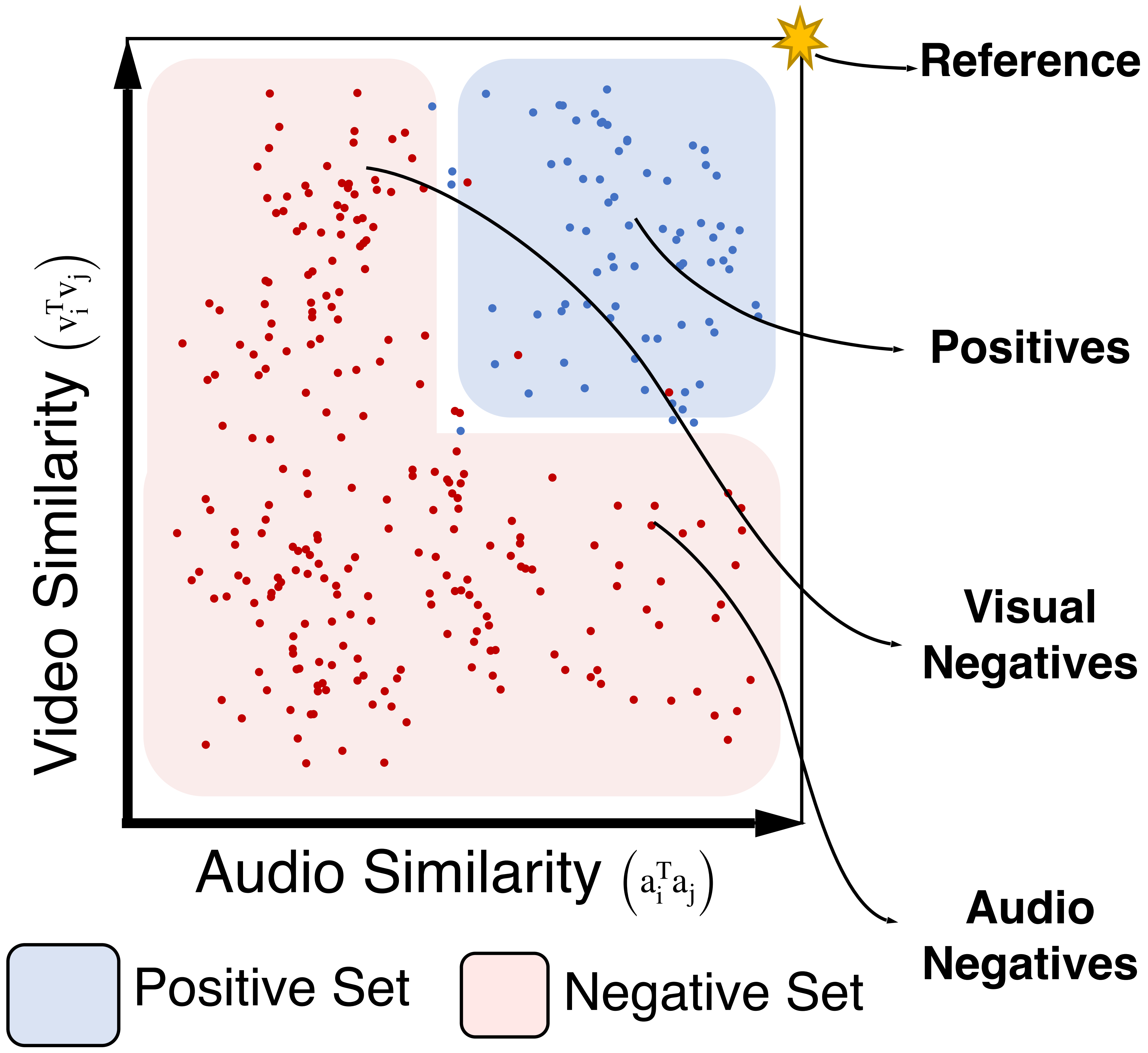
Positives **Negatives**

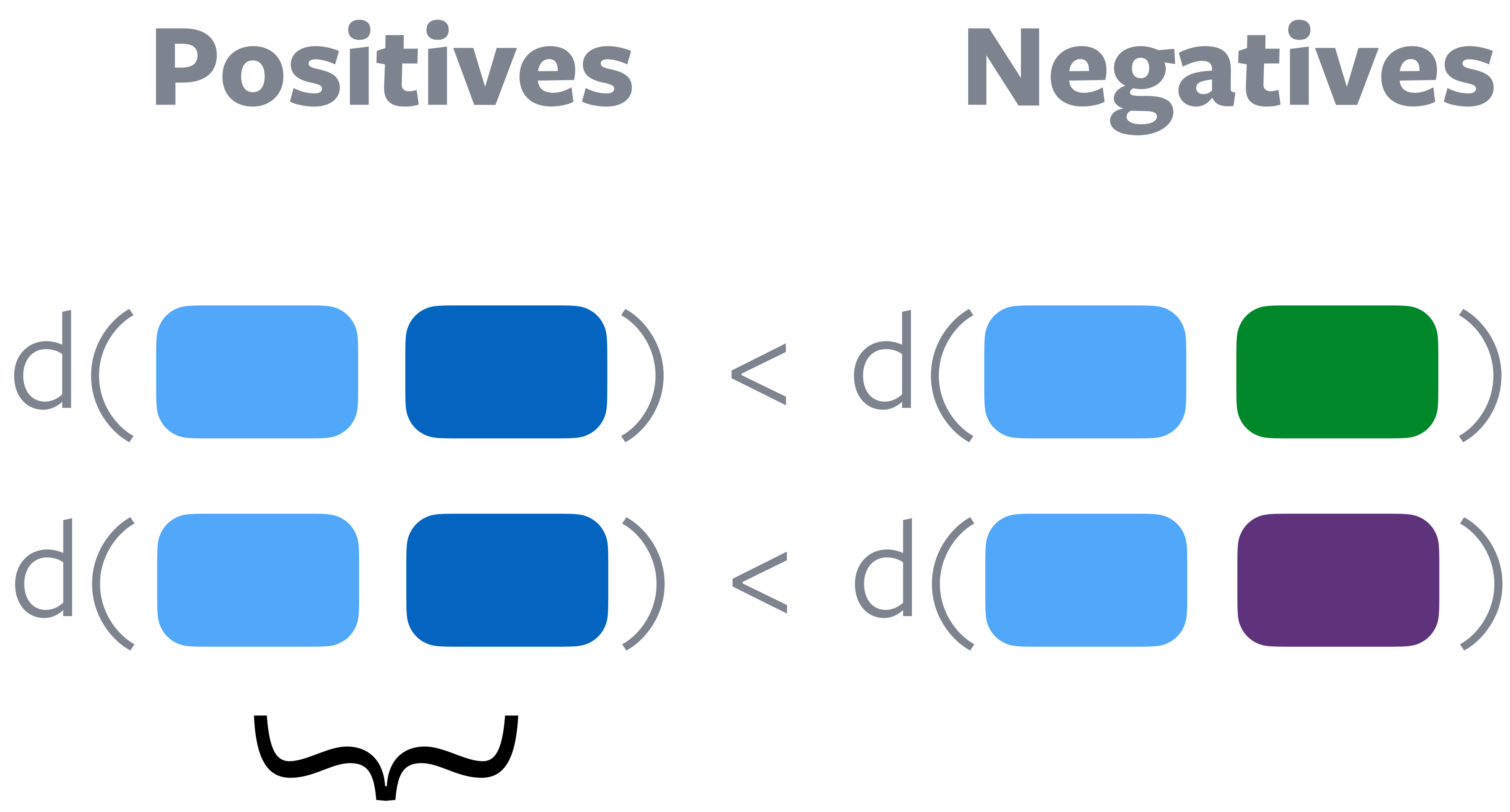


Audio & Video
(same sample)

Relate to other video/audio
using negatives

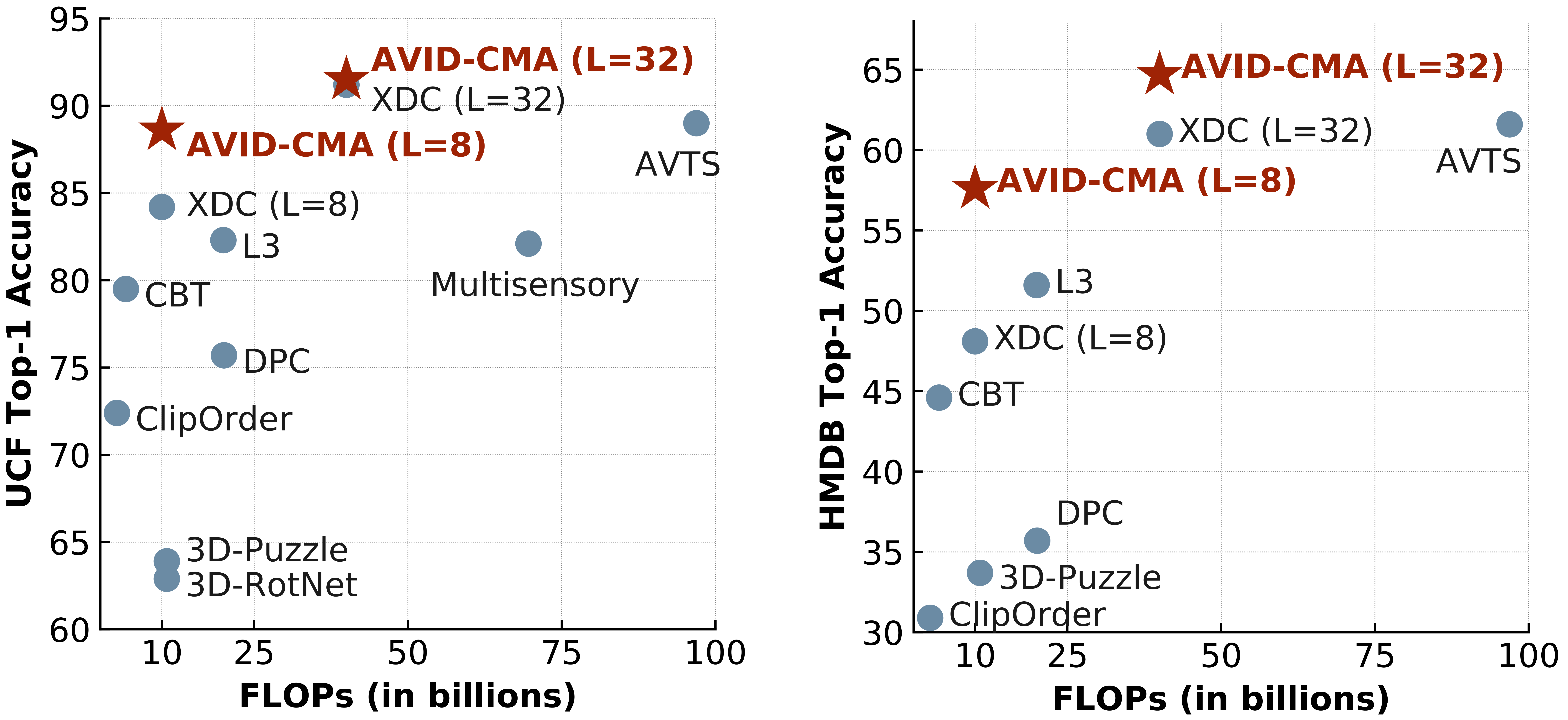
Grouping using Audio-visual Agreements (CMA)





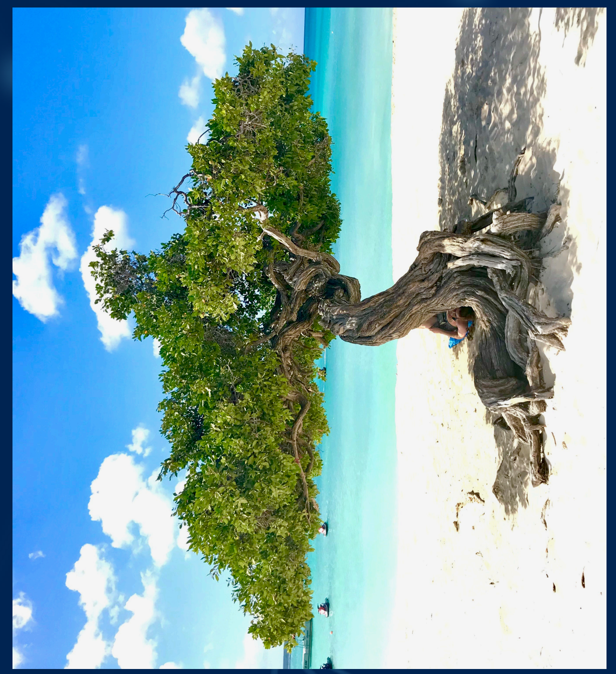
Videos that are similar in
audio & video features

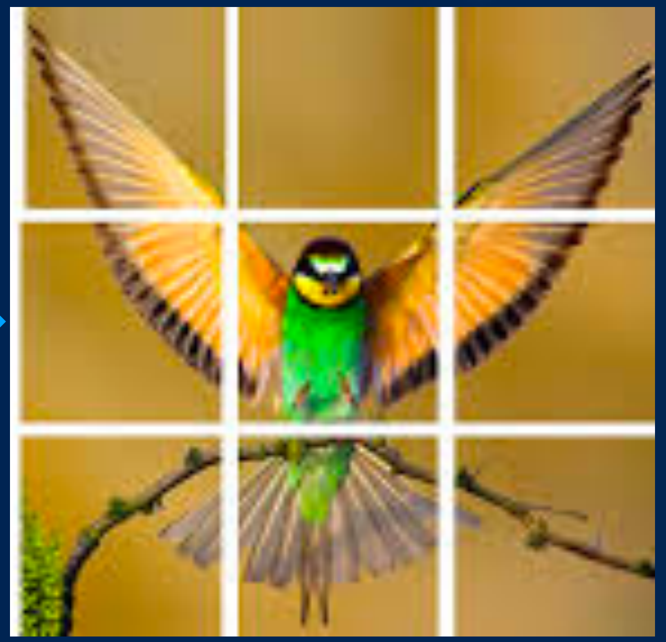
Grouping using Audio-visual Agreements (CMA)



UCF Top-1 Accuracy
95
90
85
80
75
70
65
60
10
25
50
75
100
FLOPs (in billions)
AVID-CMA (L=32)
XDC (L=32)
AVID-CMA (L=8)
AVTS
XDC (L=8)
L3
Multisensory
CBT
DPC
ClipOrder
3D-Puzzle
3D-RotNet
HMDB Top-1 Accuracy
65
60
55
50
45
40
35
30
10
25
50
75
100
FLOPs (in billions)
AVID-CMA (L=32)
XDC (L=32)
AVTS
AVID-CMA (L=8)
L3
XDC (L=8)
CBT
DPC
3D-Puzzle
ClipOrder

Pretext tasks

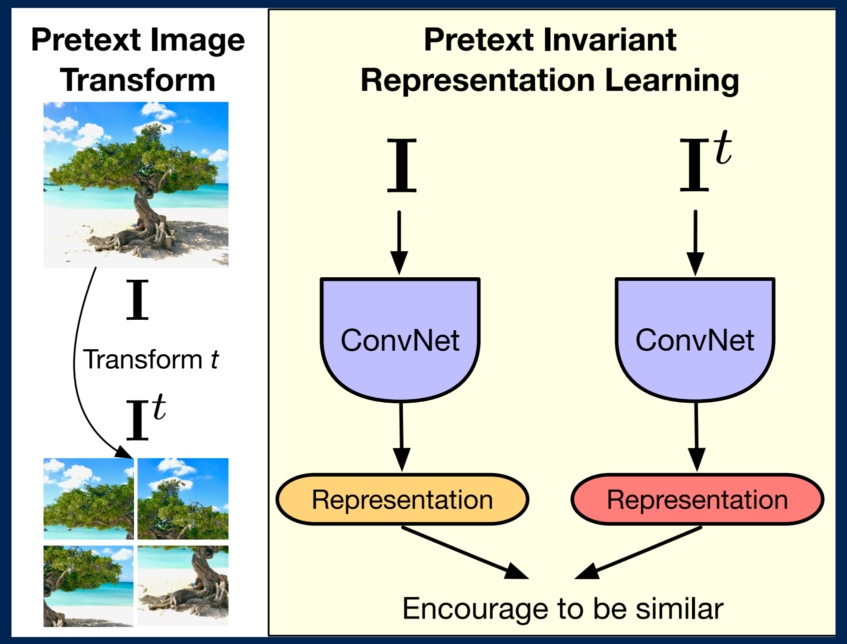




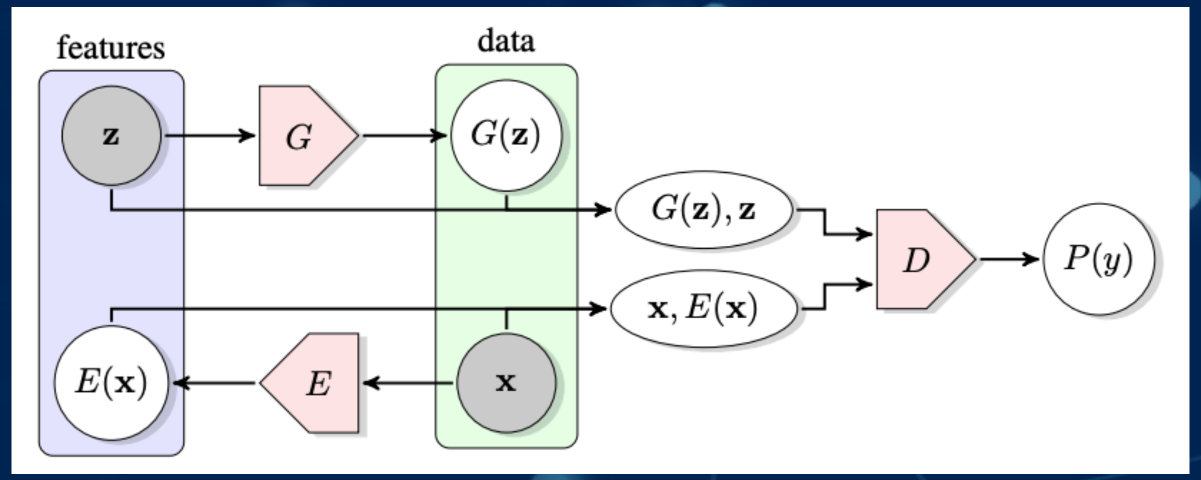
Contrastive/Clustering

Related

Unrelated



Generative



AutoEncoder, VAE, GAN, BiGAN

Predict more information