Self-supervised Learning in Vision

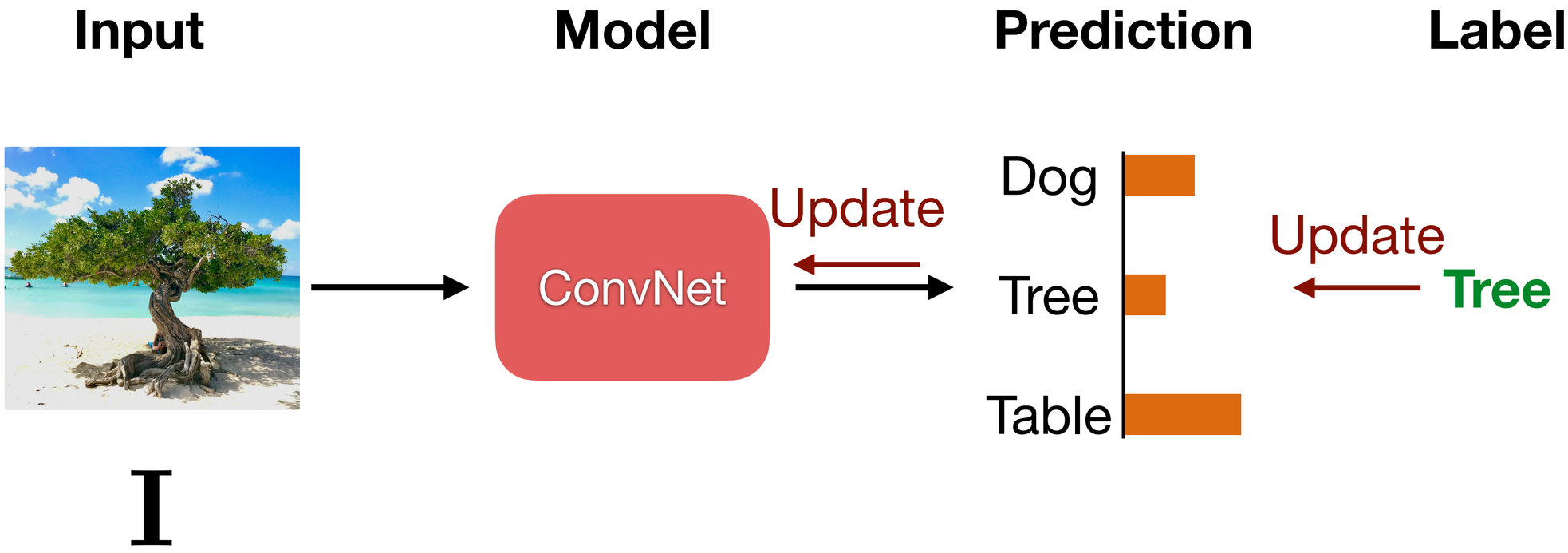
Oxford Machine Learning Summer School, 2022

Ishan Misra

FAIR, Meta

Twitter: @imisra_
https://imisra.github.io/

Supervised Learning

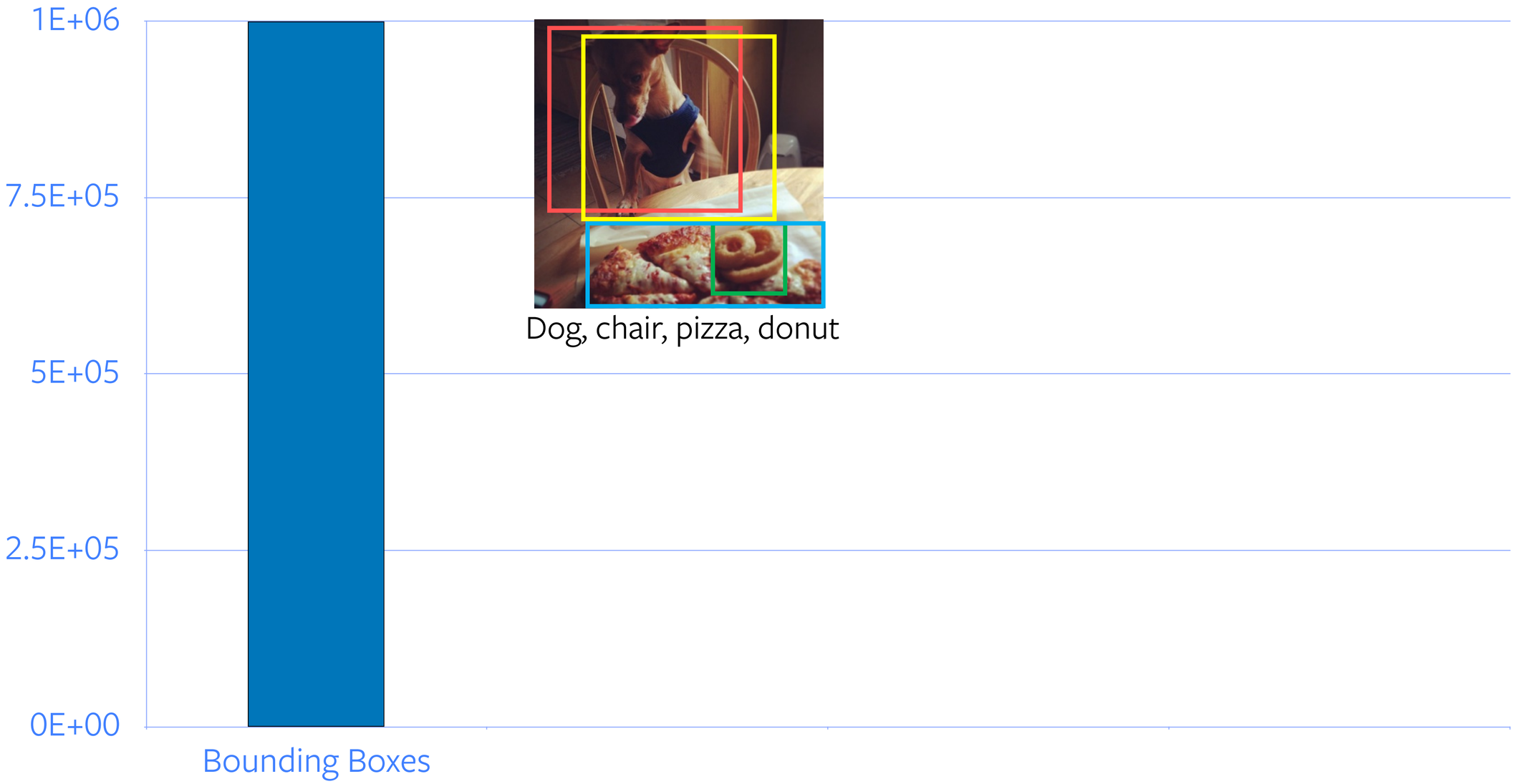


So what is the bottleneck ?

- Supervision!!
- Getting "real" labels is difficult

Can we get labels for all data?

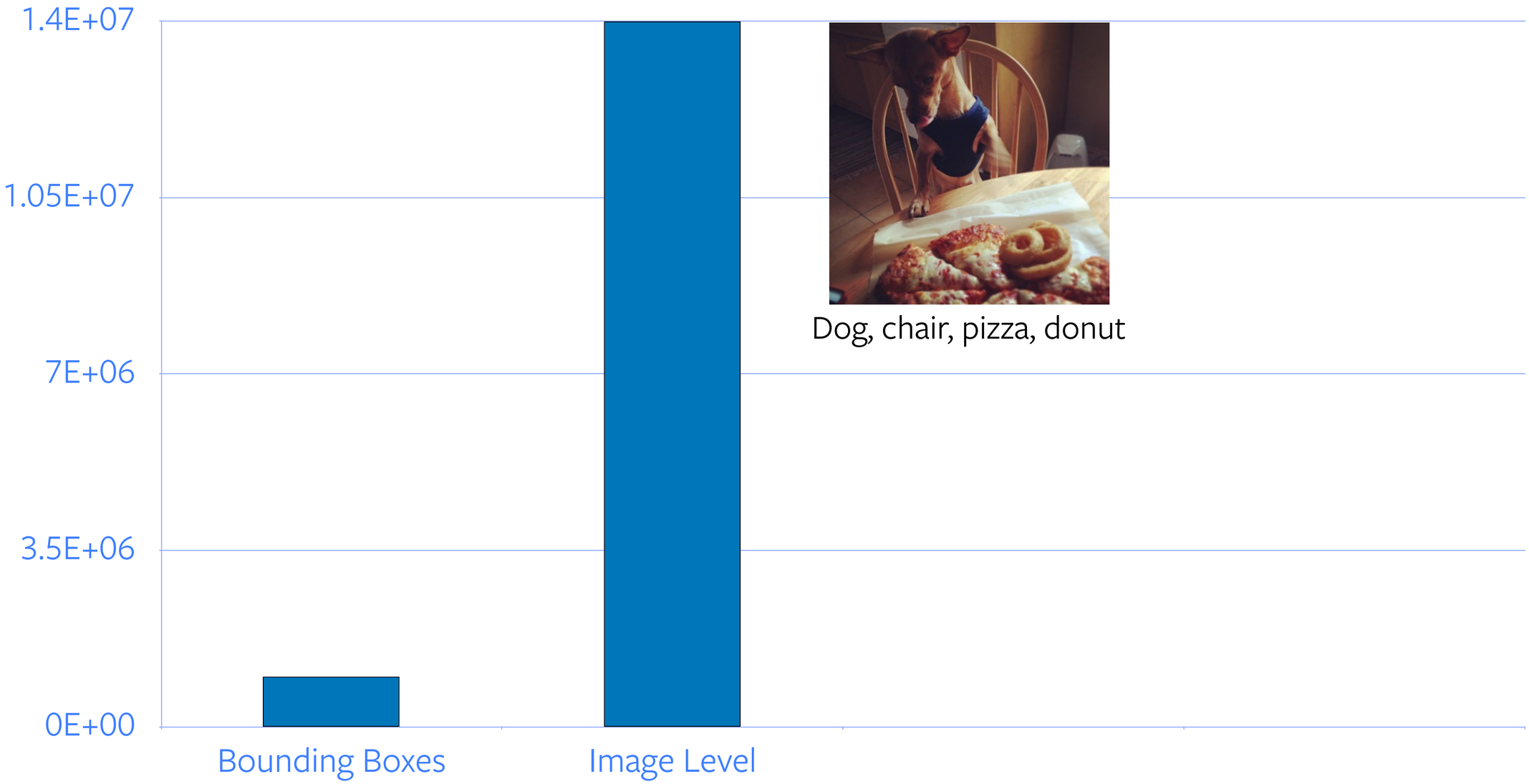
Can we get labels for all data?



Dog, chair, pizza, donut

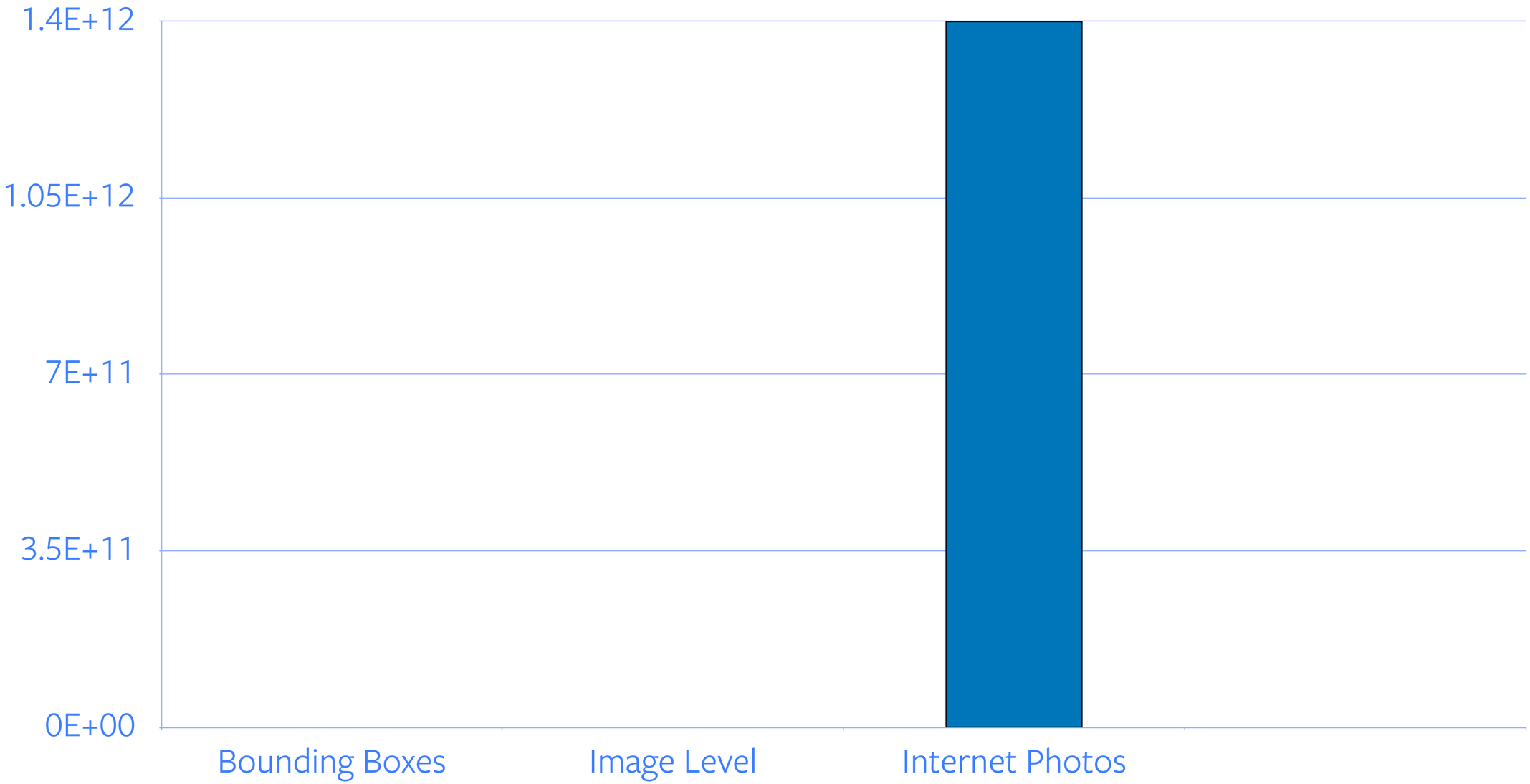
Stats from Pawan Kumar at Oxford

Can we get labels for all data?



Dog, chair, pizza, donut

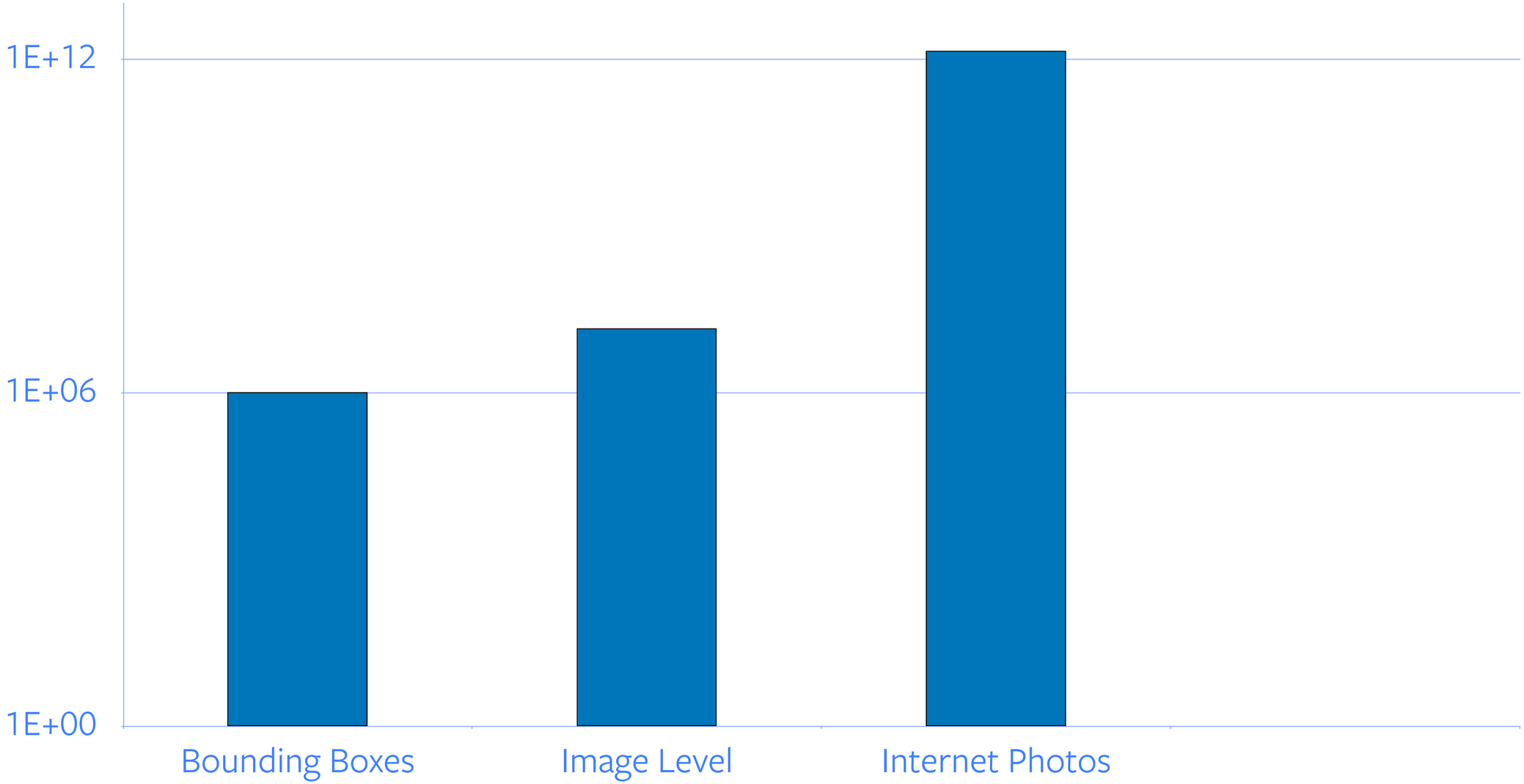
Can we get labels for all data?



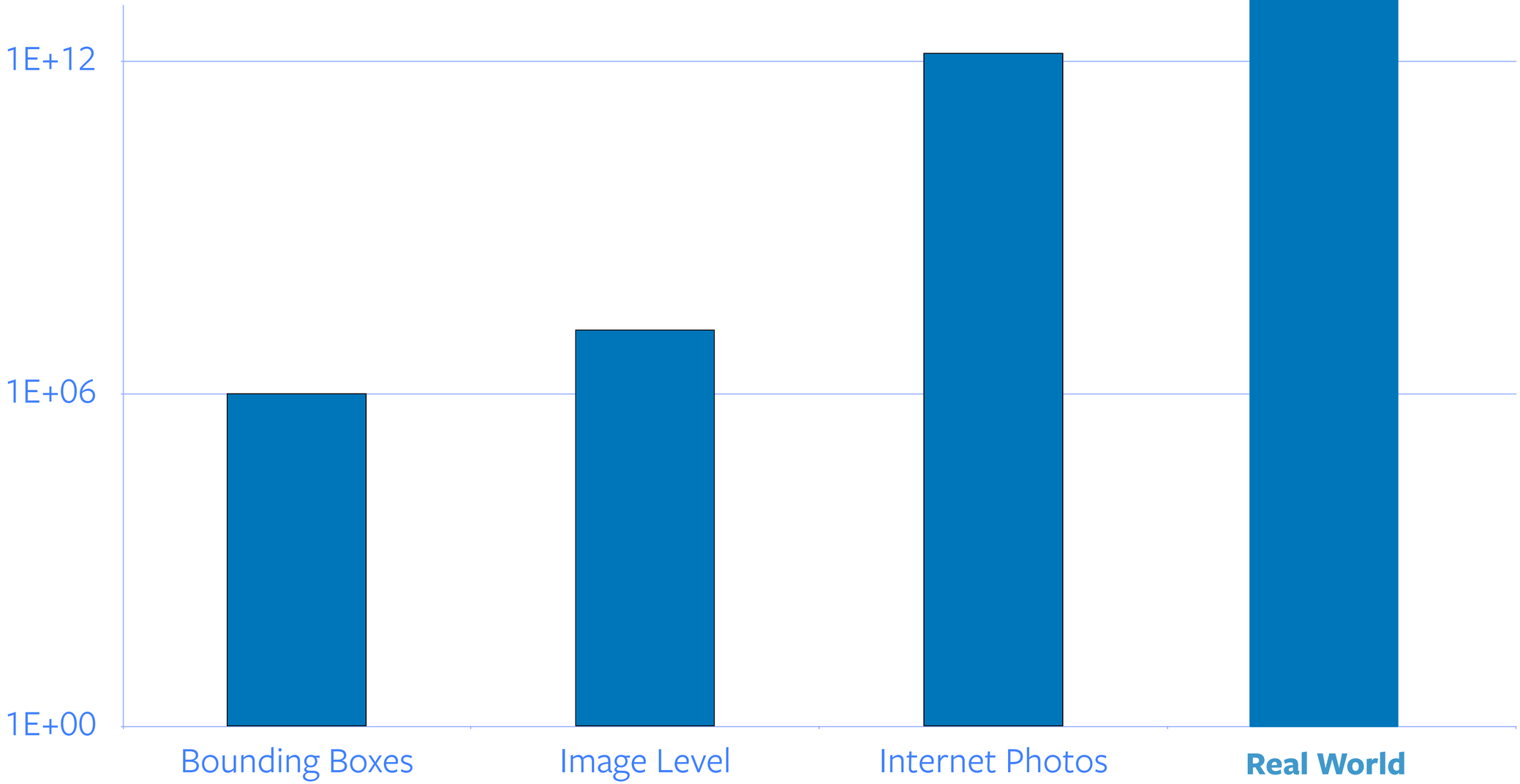
forbes.com

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?

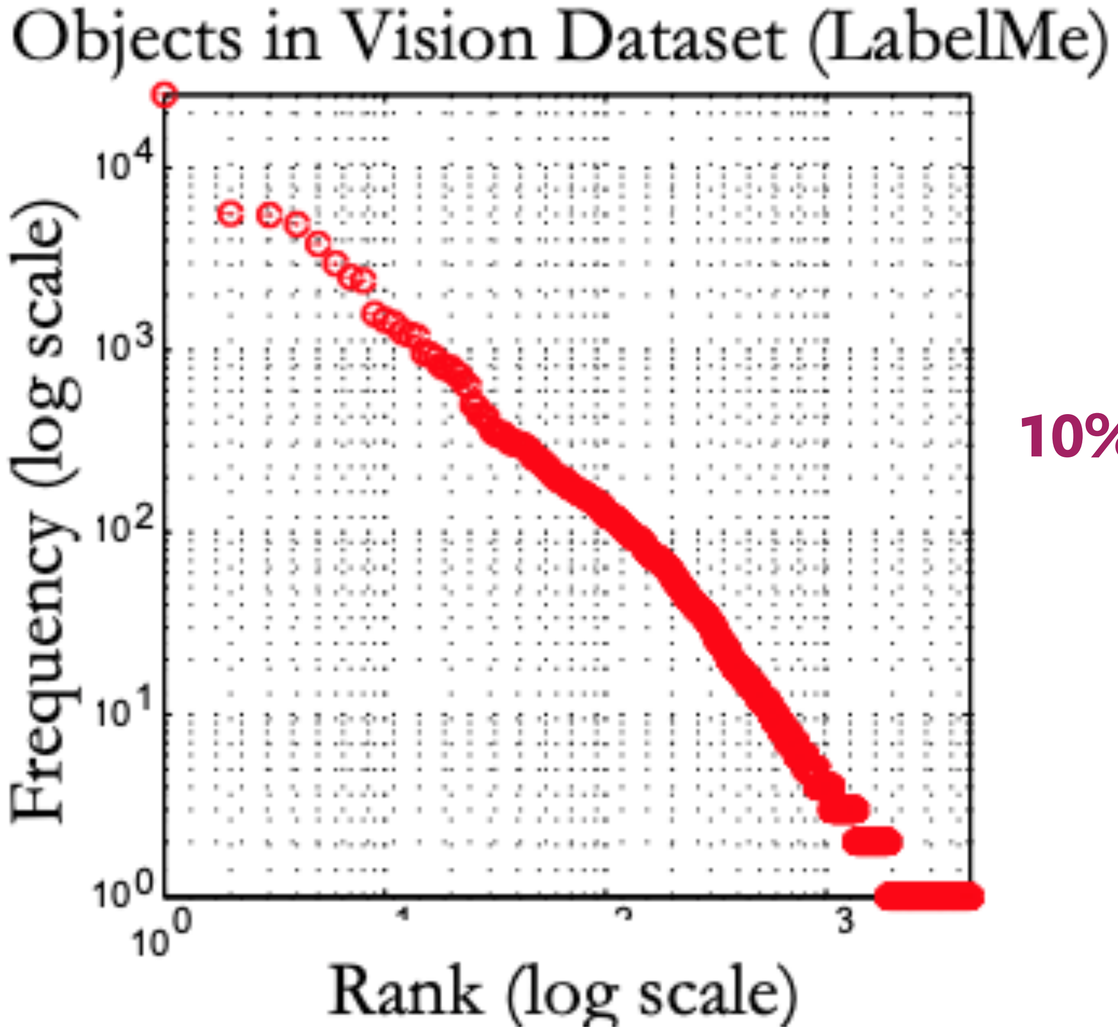


ImageNet (14 million images) needed 22 human years to label

Can we get labels for all data?

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

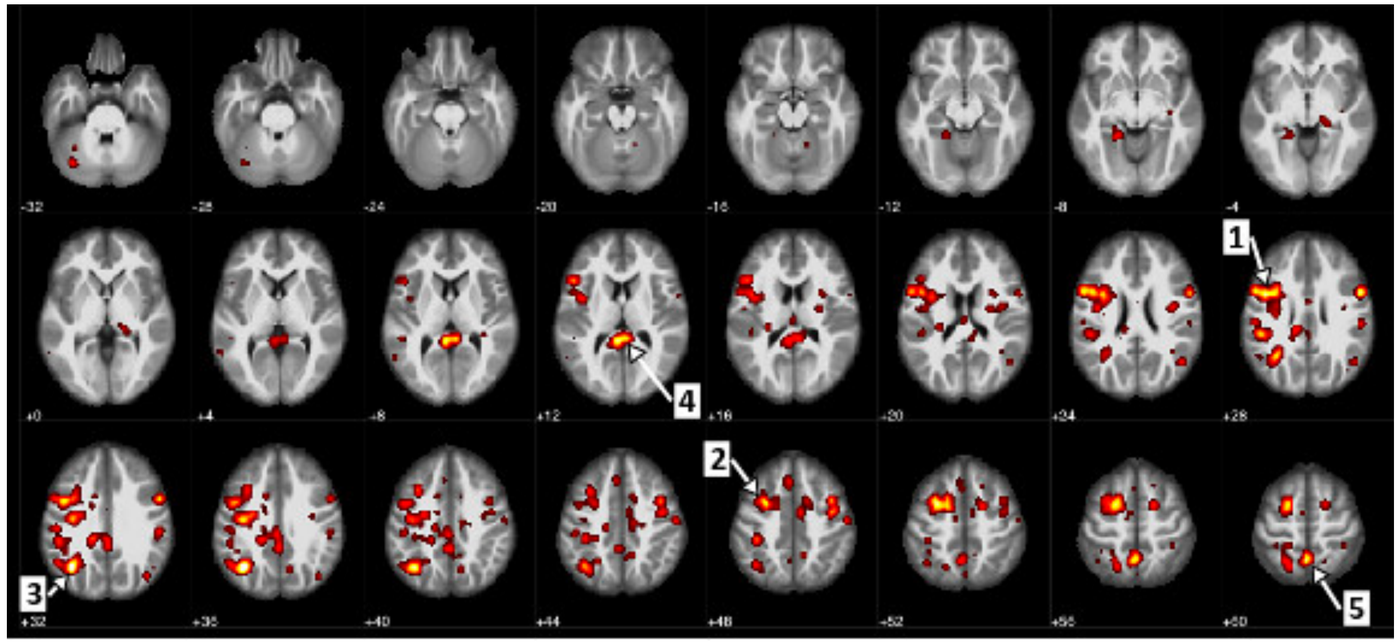
Rare concepts?



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

Slide credit: Rob Fergus

Different Domains?



Labeled data can be hard to obtain

Other Limitations of Supervised Learning

Commercial supervised AI models

Soap

Country of Origin: Nepal
Prediction: Food

Spices

Country of Origin: Philippines
Prediction: Beer

Toothpaste

Country of Origin: Burundi
Prediction: Wood

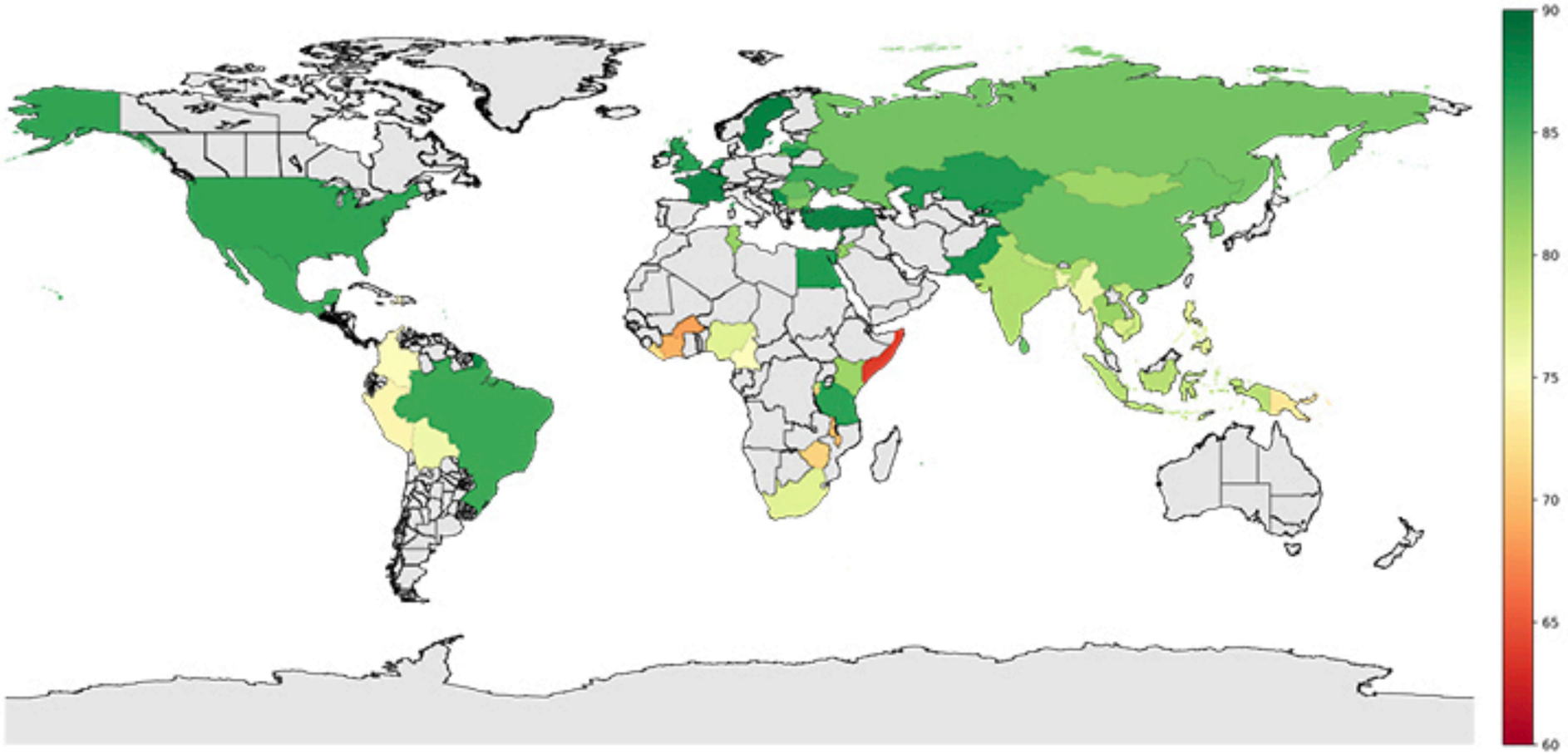
Country of Origin: UK
Prediction: Toiletry

Country of Origin: USA
Prediction: Spice

Country of Origin: USA
Prediction: Toothpaste

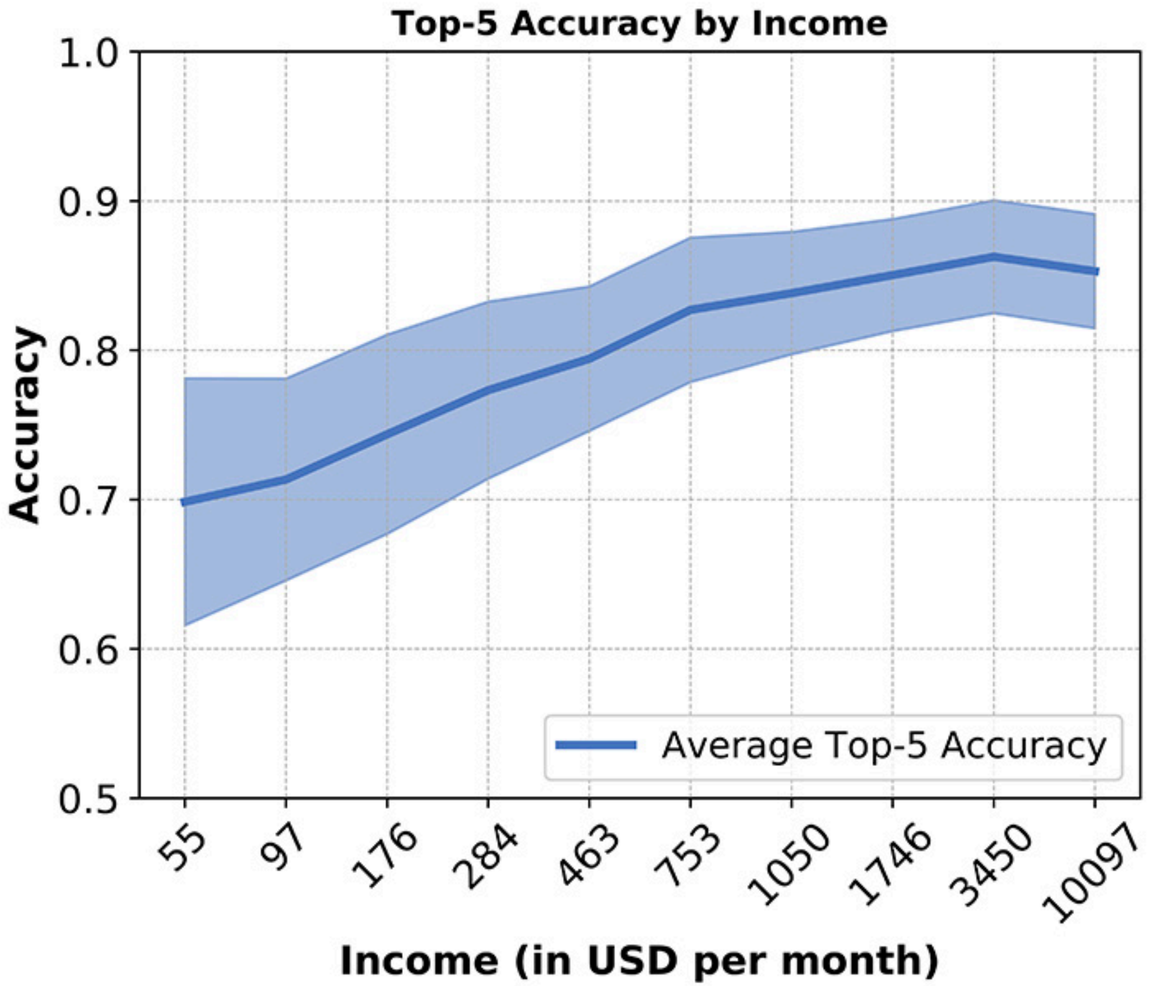
Does Object Recognition Work For Everyone? - DeVries, Misra, Wang, van der Maaten - 2019

Commercial supervised AI models



Does Object Recognition Work For Everyone? - DeVries, Misra, Wang, van der Maaten - 2019

Commercial supervised AI models



Does Object Recognition Work For Everyone? - DeVries, Misra, Wang, van der Maaten - 2019

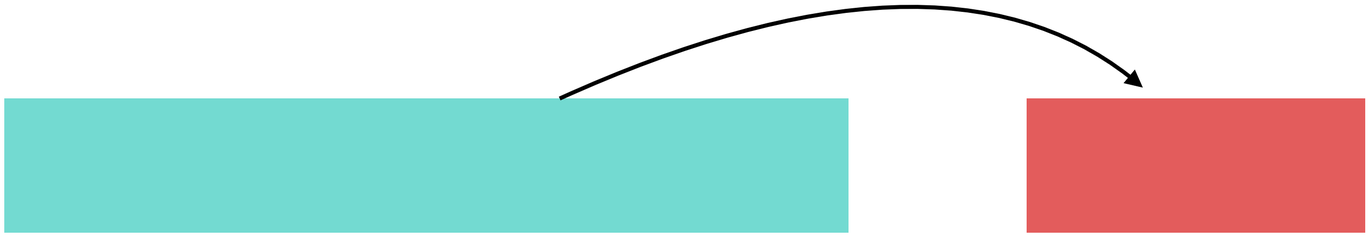
The promise of "alternative" supervision

- 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

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

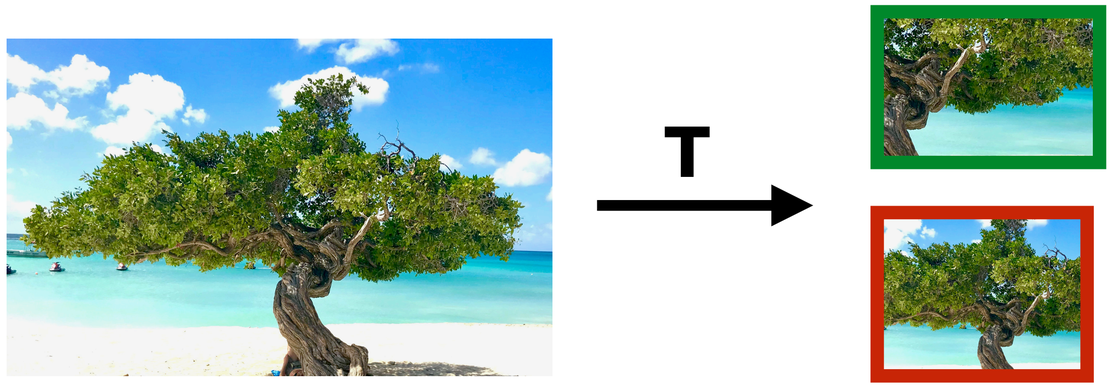
Hidden property of the data

In the context of Computer Vision

Two paradigms

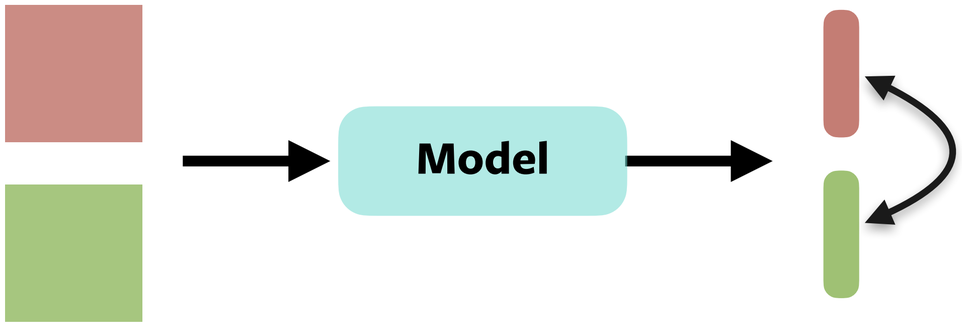
Invariance

Predicting Properties of Data (Pretext)

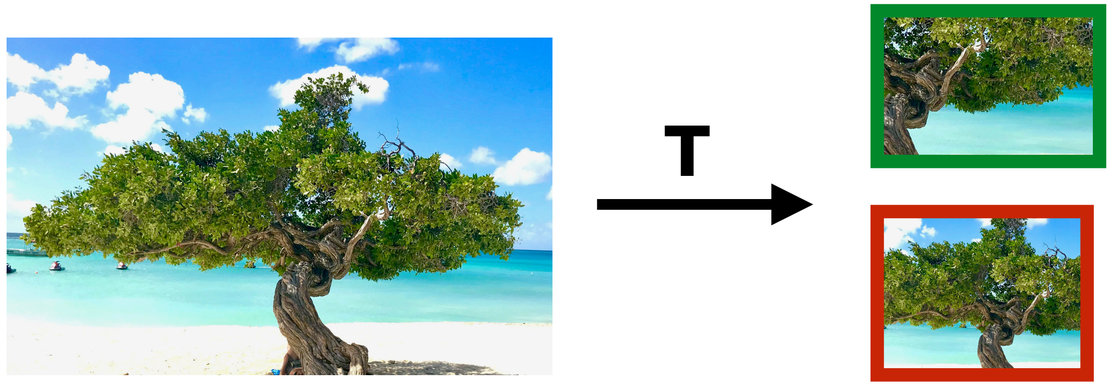


Two paradigms

Invariance

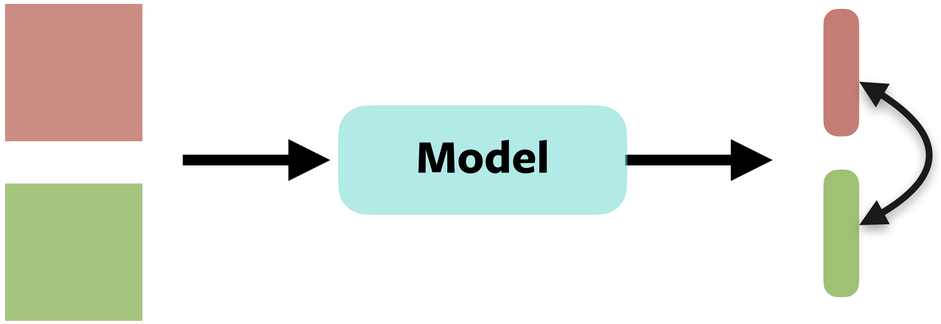


Predicting Properties of Data (Pretext)

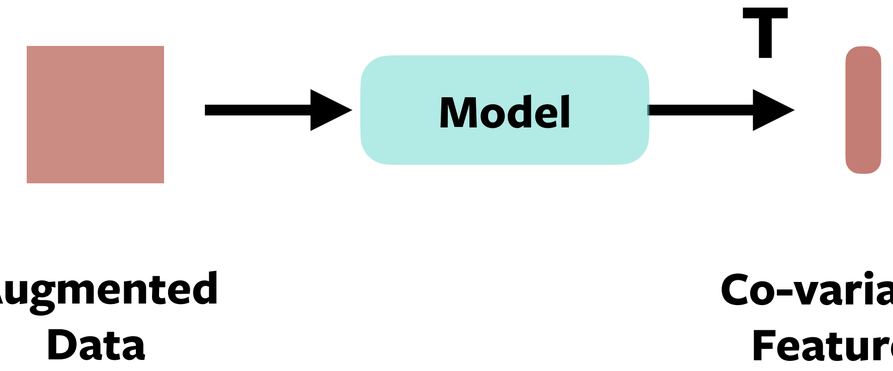


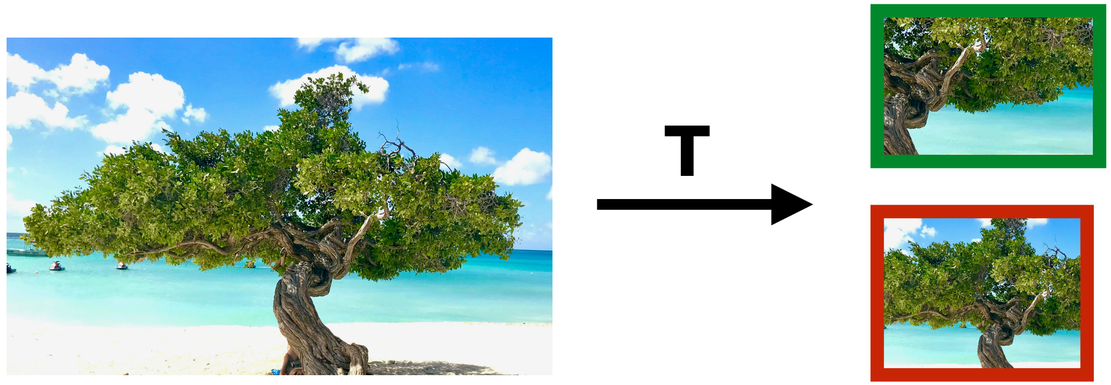
Two paradigms

Invariance



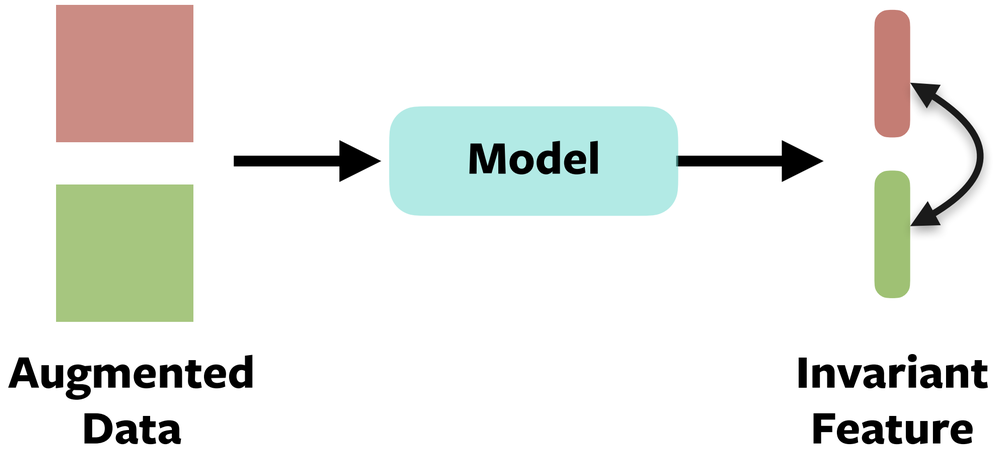
Predicting Properties of Data (Pretext)





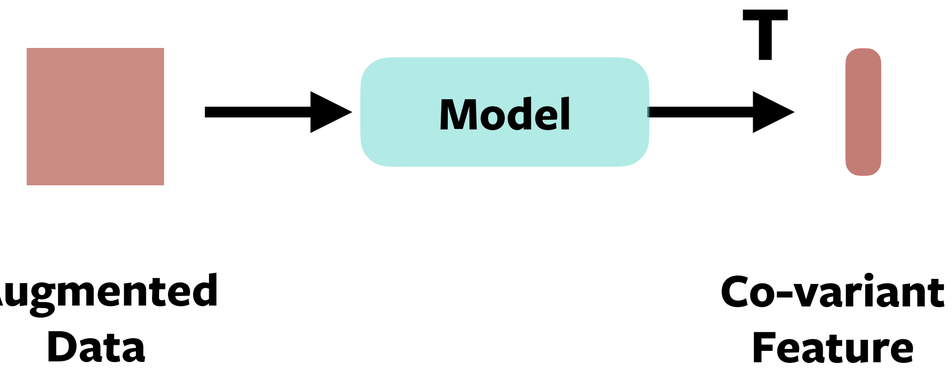
Two paradigms

Invariance



Contrastive Learning
Clustering

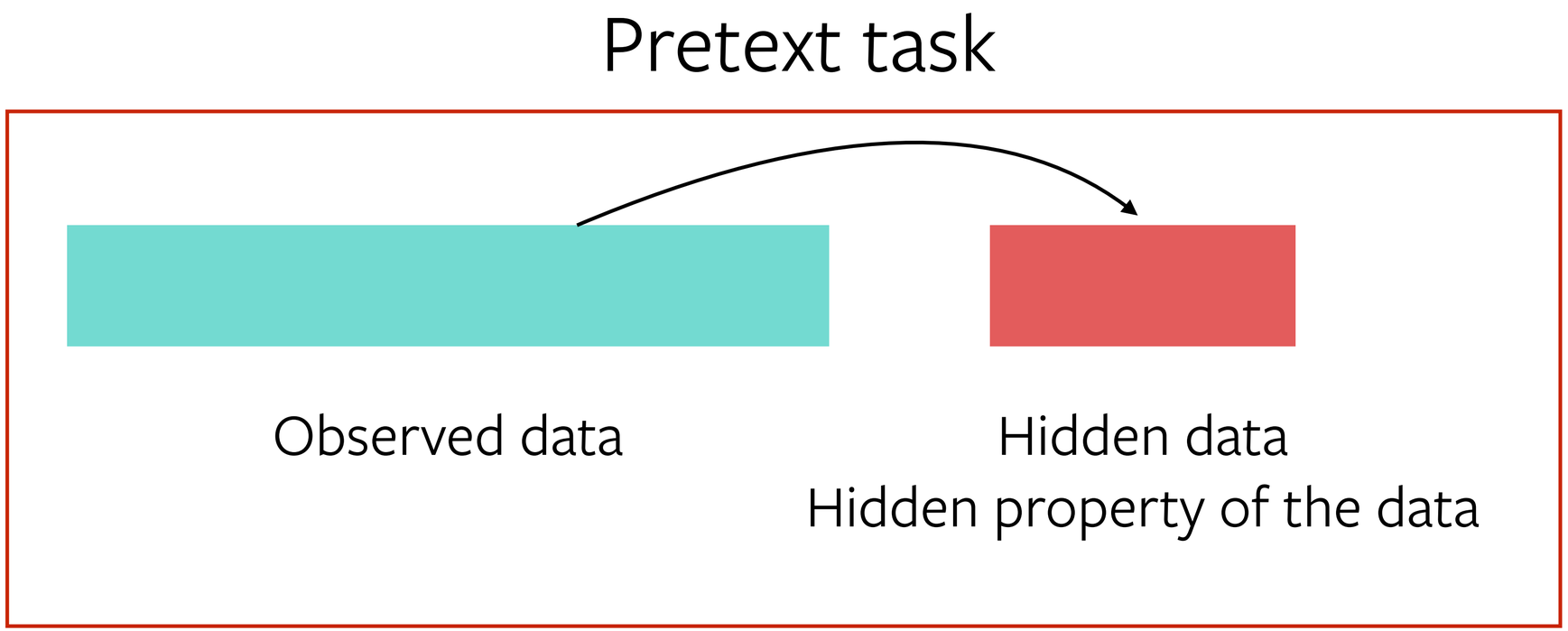
Predicting Properties of Data (Pretext)



Autoencoding
Pretext Tasks

Pretext task

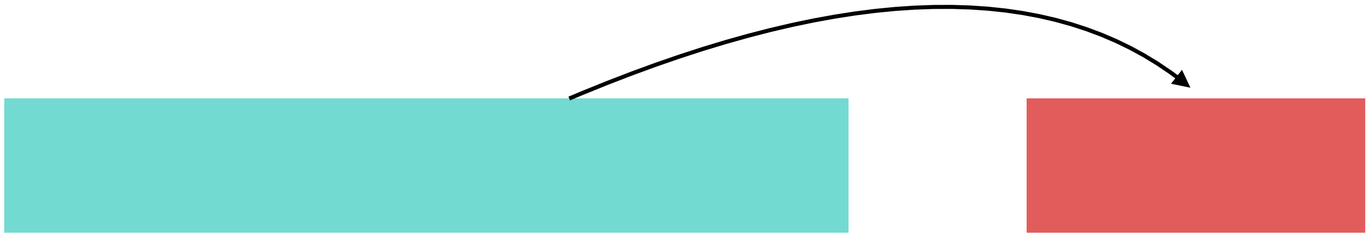
- Self-supervised task used for learning representations
- Often, not the "real" task (like image classification) we care about



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

Pretext task

- Using images
- Using video
- Using video and sound

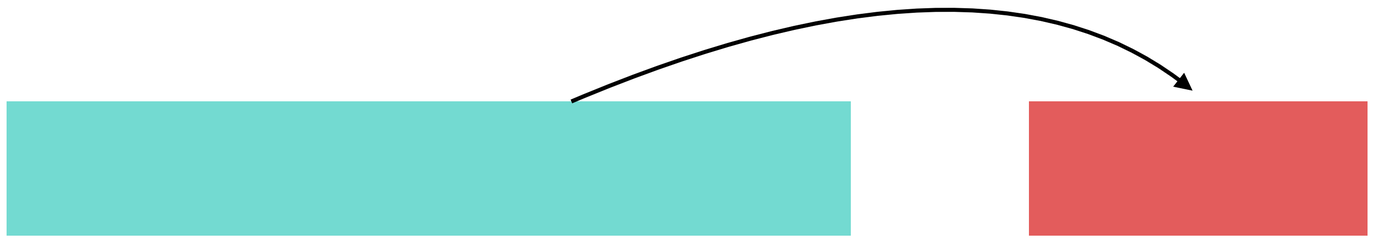


Observed data

Hidden data

Hidden property of the data

Pretext task



Observed data

Hidden **data**

Hidden **property** of the data

Level of “abstraction”
in the hidden targets matter

Pretext task

- Predict pixel values or missing features

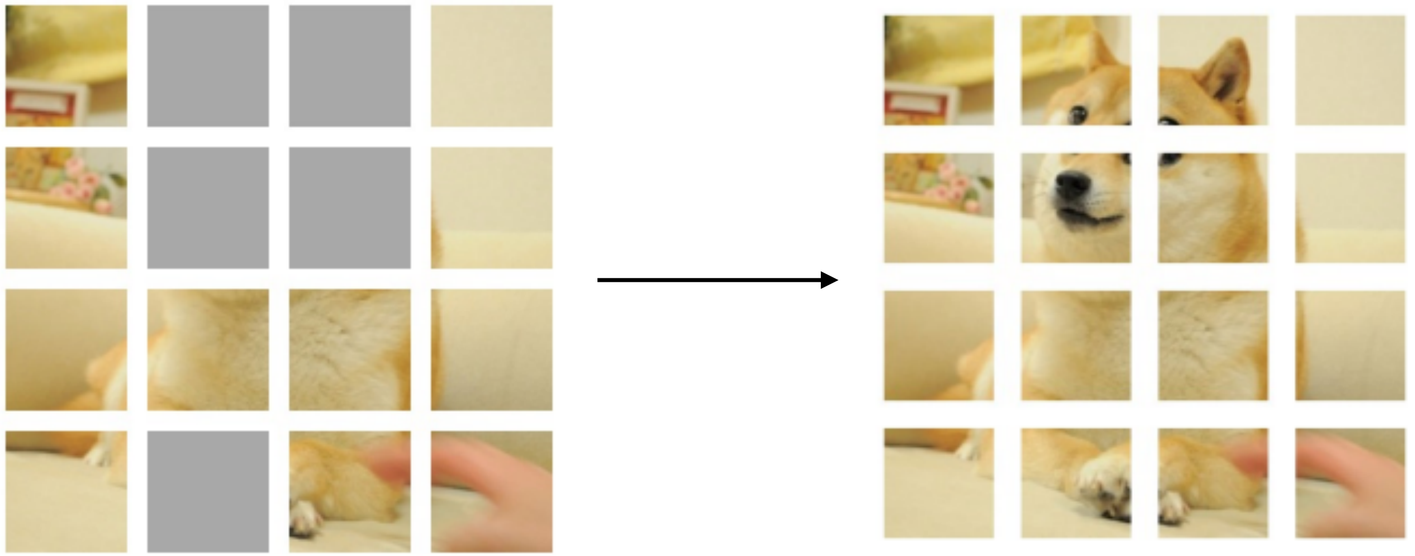
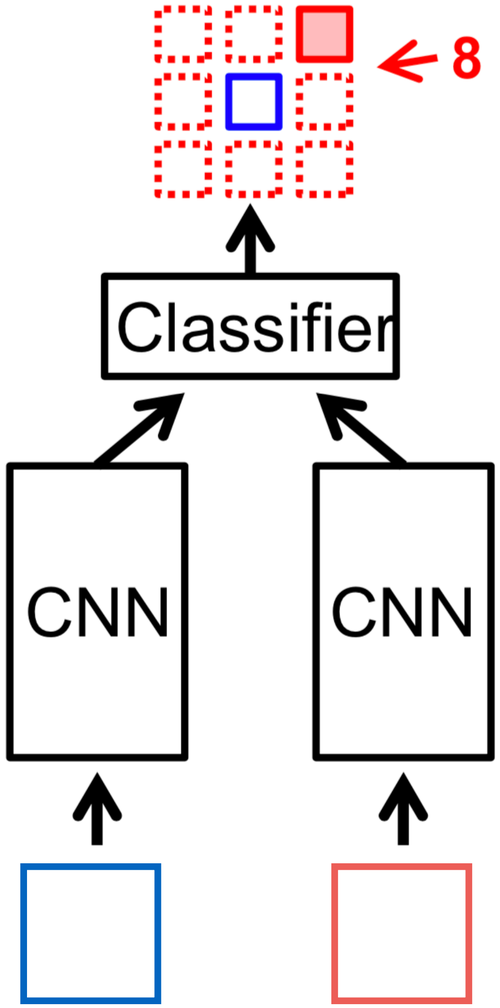
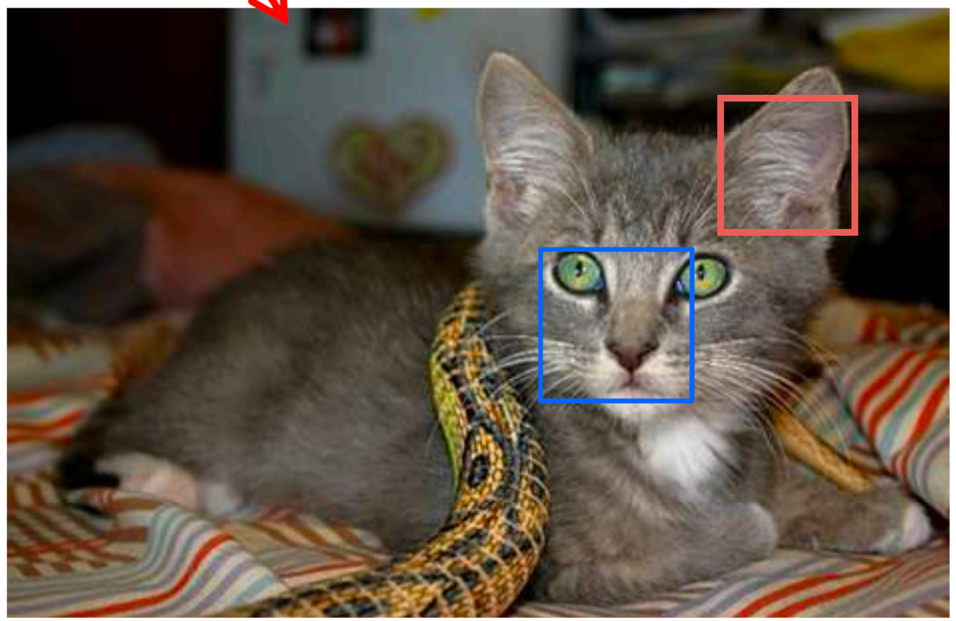


Image from: BEIT: BERT Pre-Training of Image Transformers - Bao et al., 2021

Relative Position of patches



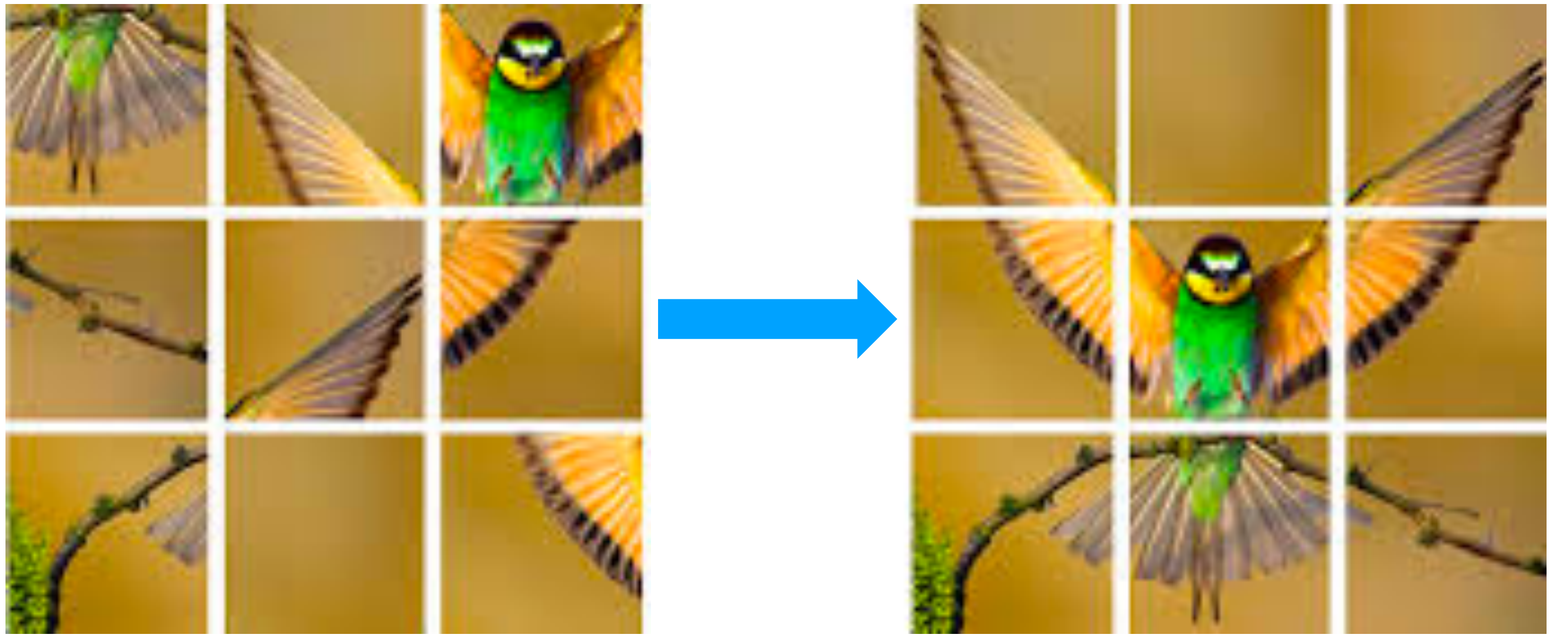


Input: Two patches

Output: 8-way classification

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

Jigsaw



Jigsaw puzzles
(Noorozi & Favaro, 2016)

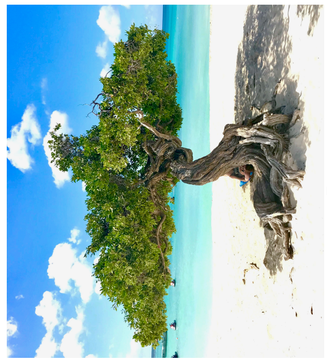
Input: nine patches
Permute using one of N permutations

Output: N-way classification

Set N << 9!

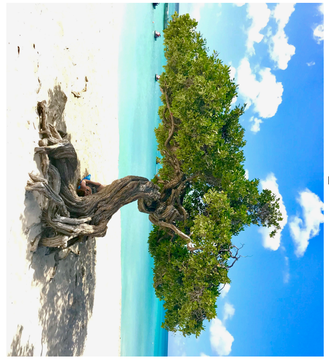
Predicting Rotations

0^0



90^0

180^0



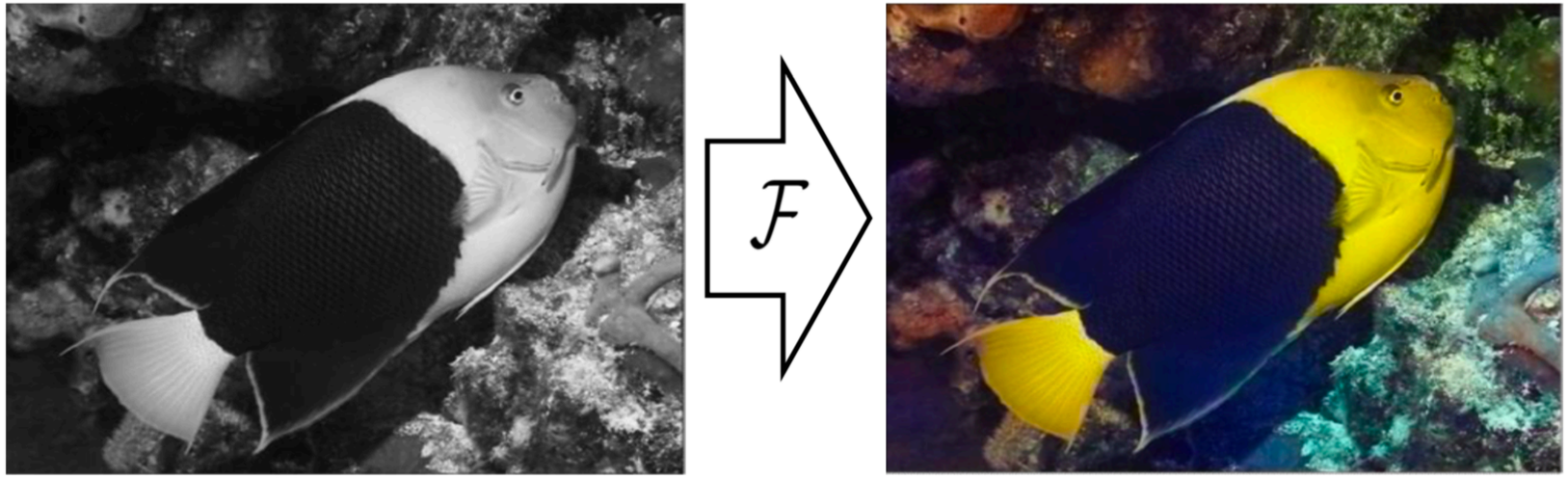
270^0

Input: image rotated by [0, 90, 180, 270]

Output: 4-way classification

Gidaris et al., 2018, Predicting Image Rotations

Colorization

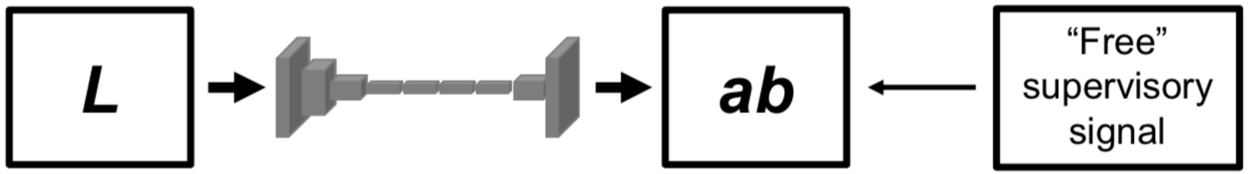


Grayscale image: *L* channel

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

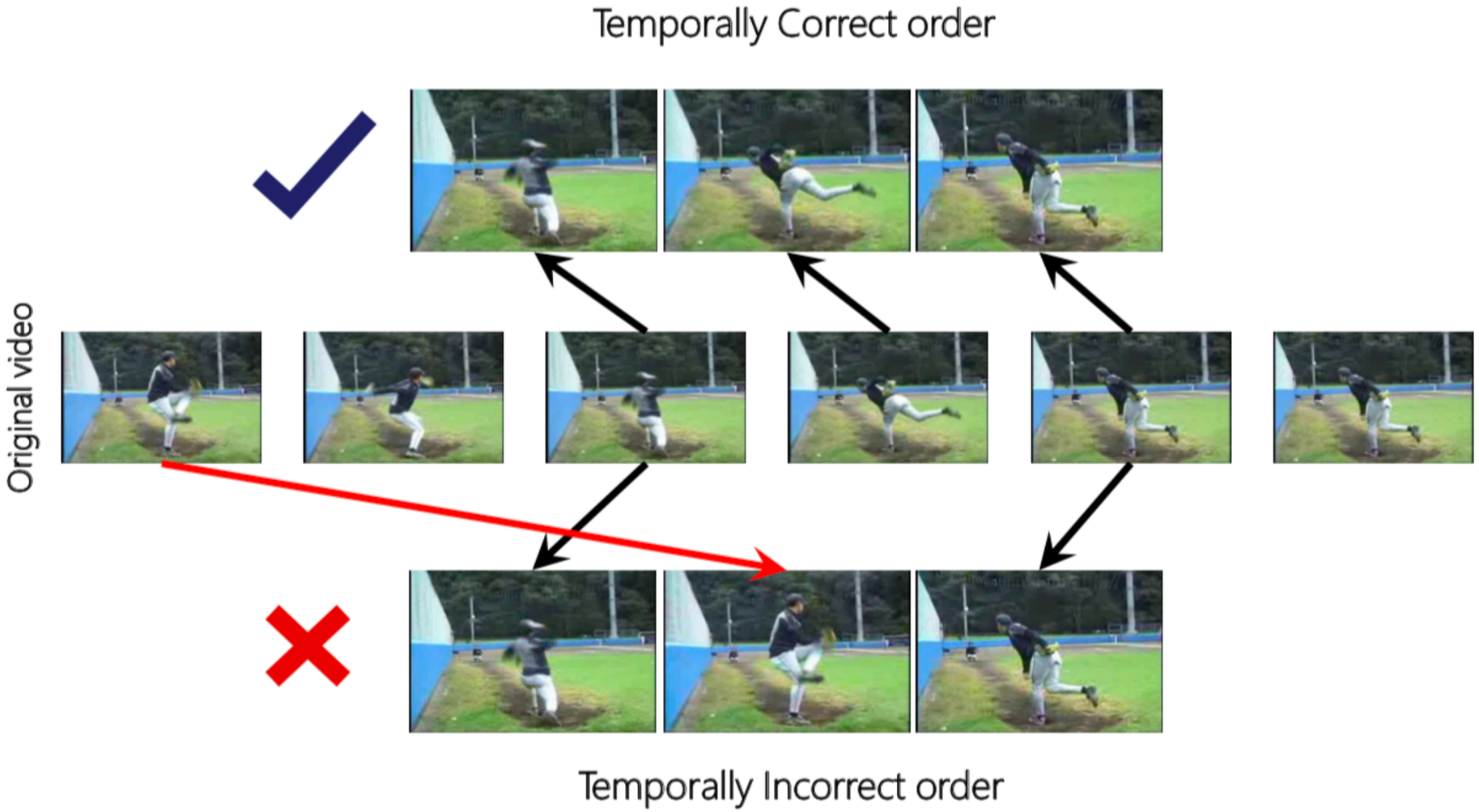
Concatenate (L,ab)

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



Zhang and Efros, 2016, Colorful image colorization

Shuffle & Learn

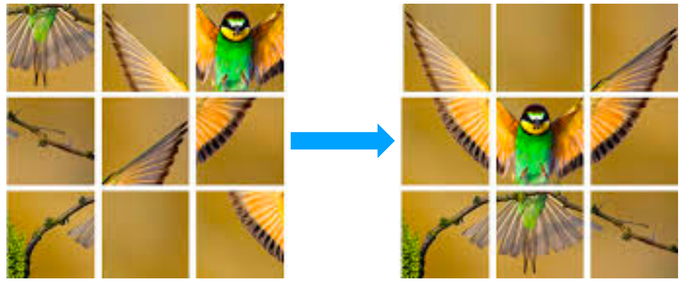


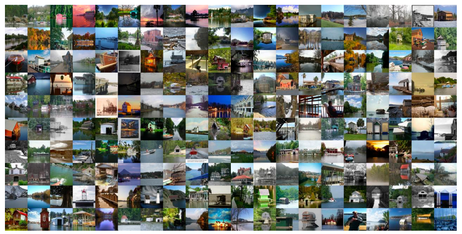
Misra et al., 2016, Shuffle and Learn

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

The hope of generalization

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



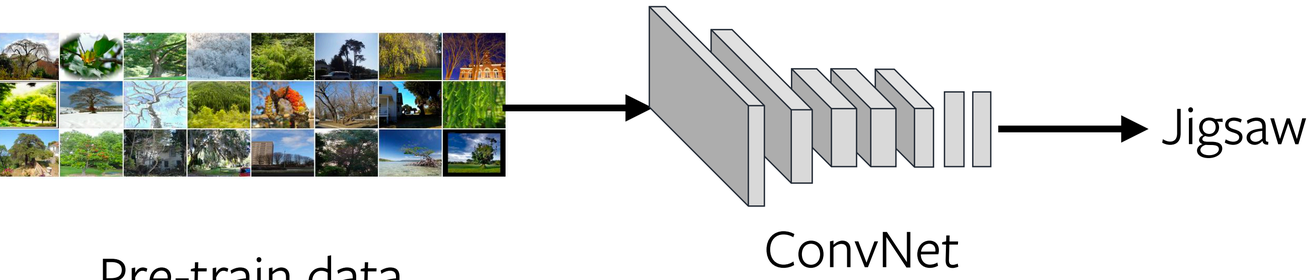


Pre-training

Self-supervised

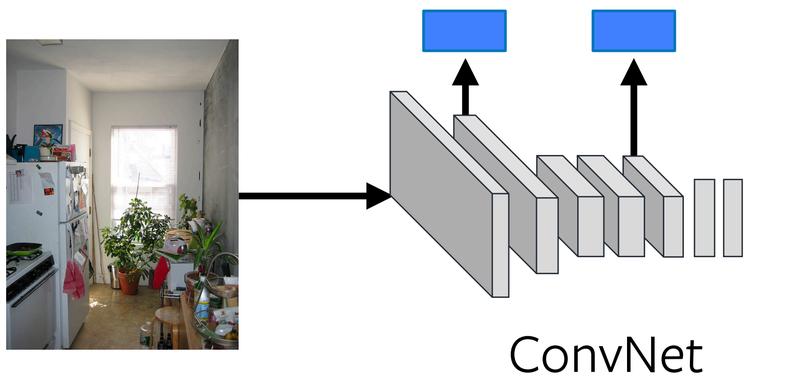
Transfer Tasks

The hope of generalization ... ?



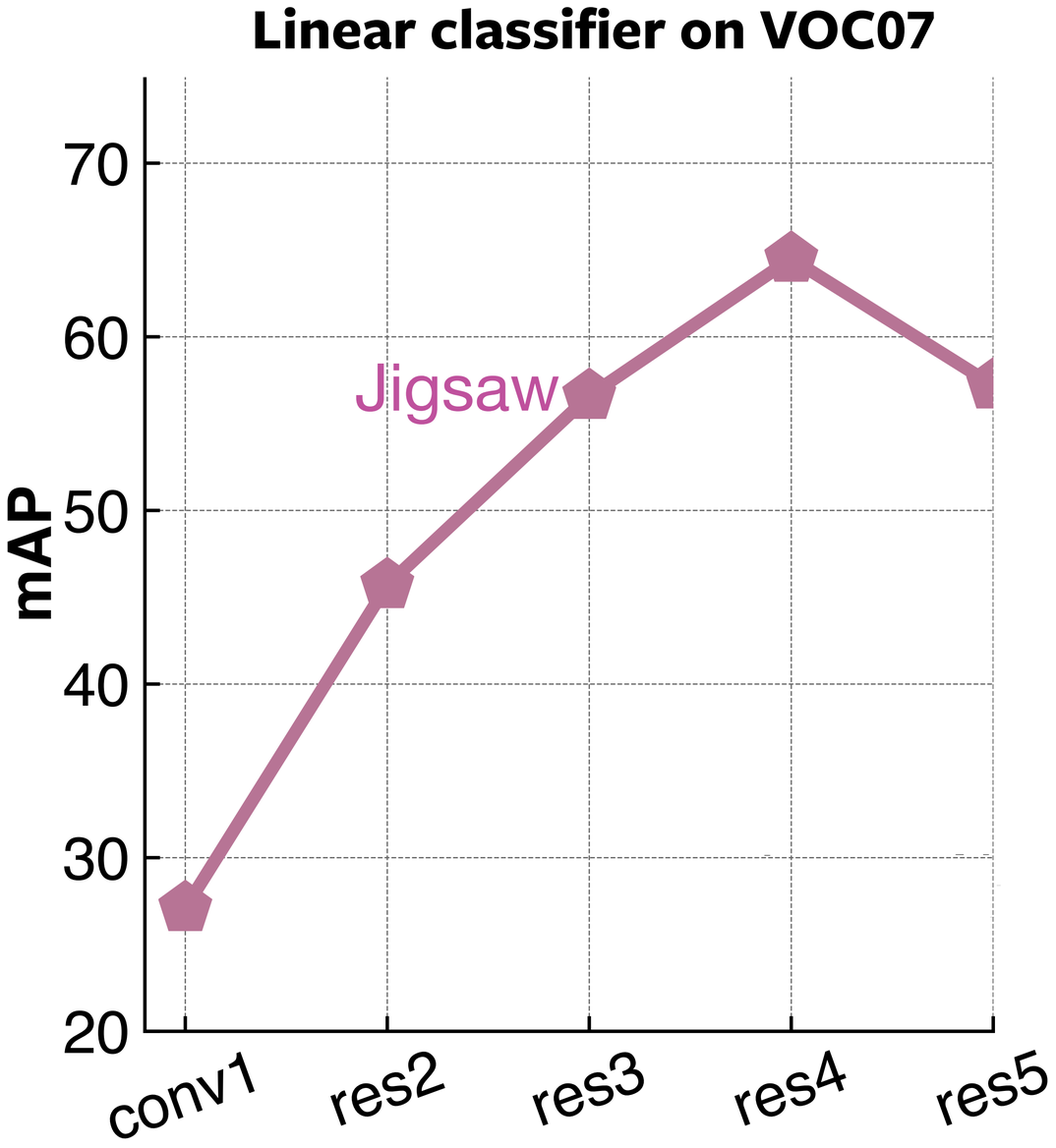
Pre-training

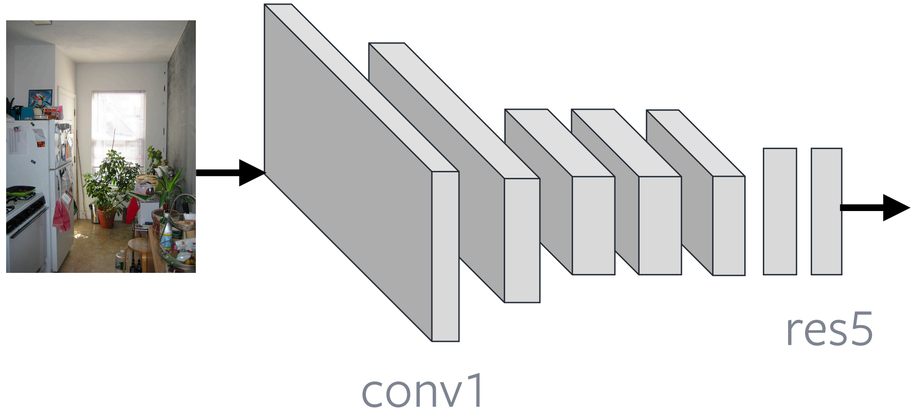
Weak or self-supervised



Transfer

Higher layers do not generalize ...





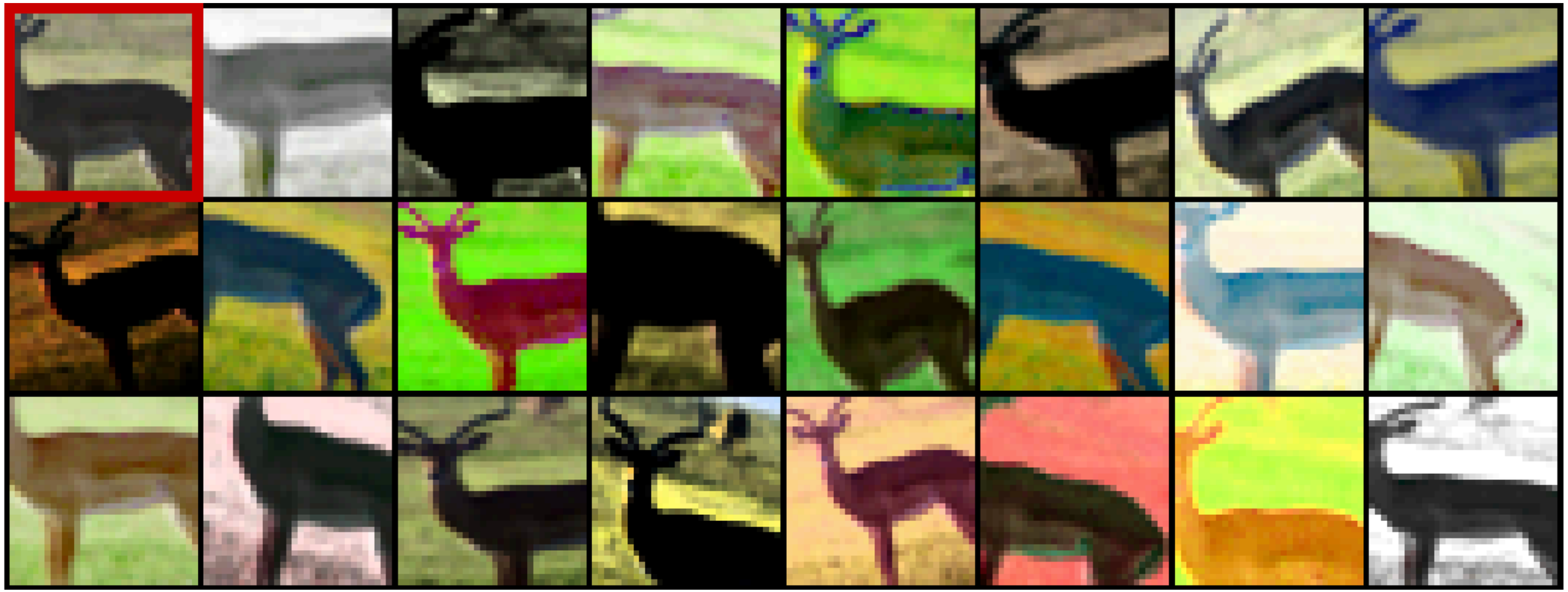
mAP = mean Average Precision
(Higher is better)

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

Invariant feature learning paradigm

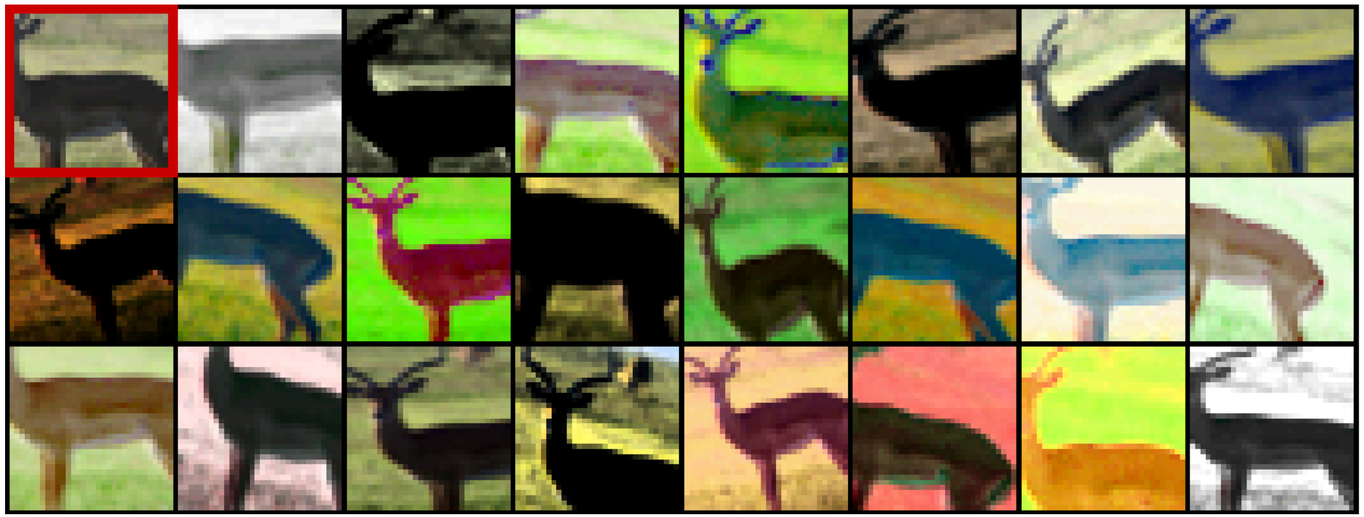


Learn features such that:

$$f_\theta(I) = f_\theta(\text{augment}(I))$$

Figure from Dosovitskiy et al., 2014

Why is it useful?



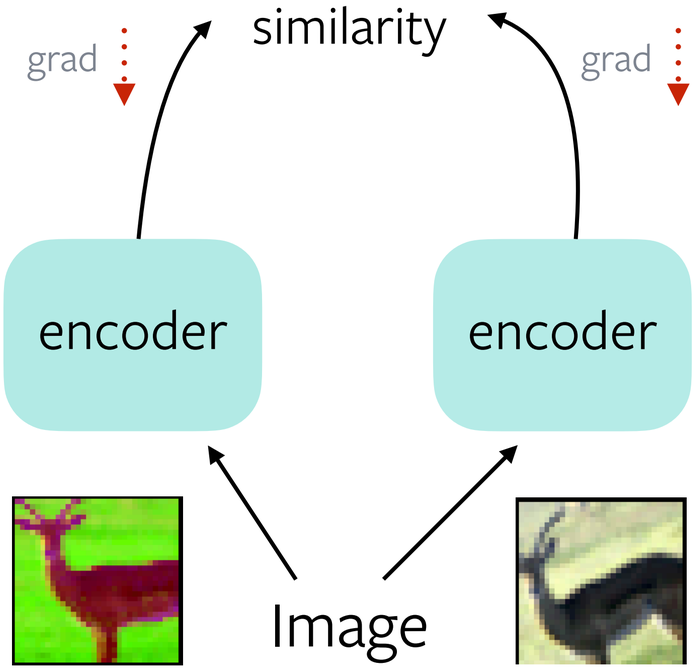
Learn features such that:

$$f_\theta(I) = f_\theta(\text{augment}(I))$$

Learned features are invariant to "nuisance factors"
or data augmentation

Can it work?

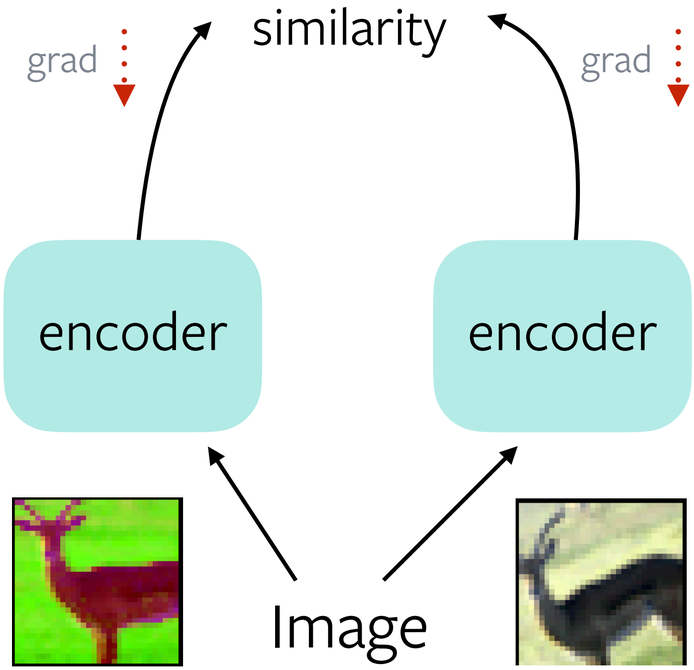
$$f_\theta(I) = f_\theta(\text{augment}(I))$$



Trivial Solutions

$$f_\theta(I) = f_\theta(\text{augment}(I))$$

$$f_\theta(I) = \text{constant}$$



Satisfies the invariance property, but not useful

Categorization of recent self-supervised methods

Invariant feature learning: ways to avoid trivial solutions

Similarity Maximization Objective

- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam, DINO

Redundancy Reduction Objective

- Redundancy Reduction
  - Barlow Twins, VICReg

Many ways to avoid trivial solutions

Similarity Maximization Objective

- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam

Redundancy Reduction Objective

- Redundancy Reduction
  - Barlow Twins

Pretraining

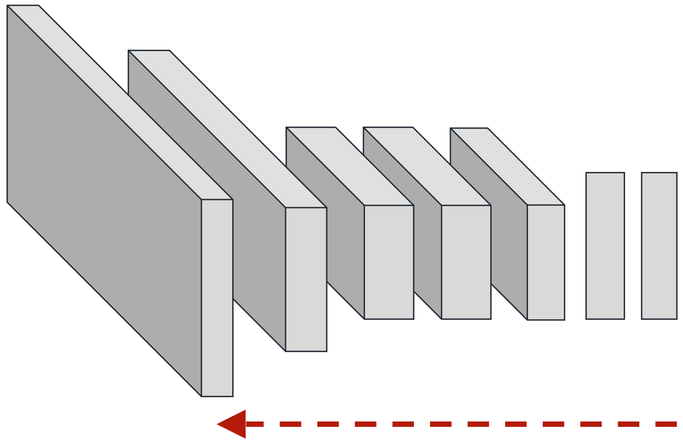
- ImageNet without labels (1.3M images)
- ResNet-50 initialized randomly

Evaluation using Transfer Learning

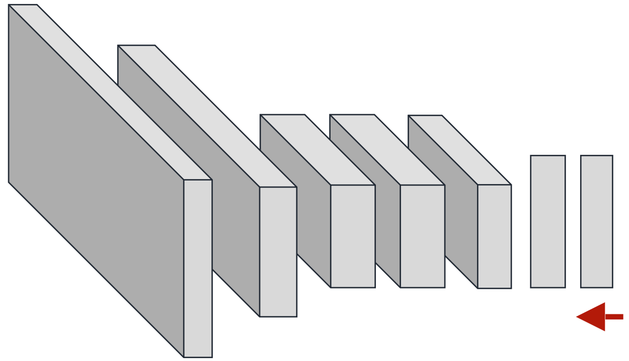
Transfer to downstream task

- Train a linear classifier on frozen features
- Full finetuning of the network

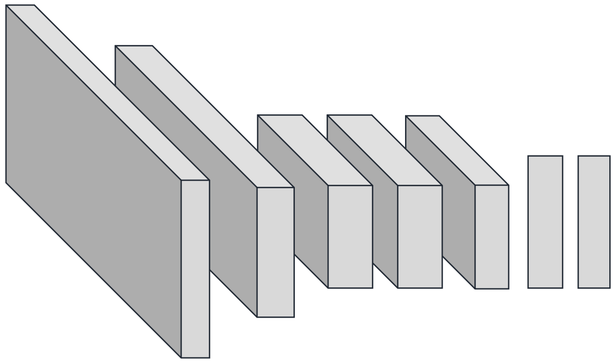
Evaluation – fine-tuning vs. linear classifier vs. kNN



Fine-tune all layers



Linear classifier



kNN

Is this representation learning
OR
learning a good initialization?

The great spiral of research

Pre 2015 - Sparse encoding, RBMs, contrastive

2015 - Pretext

2018/19 - Invariance using Contrastive

2020 - Invariance using non-contrastive

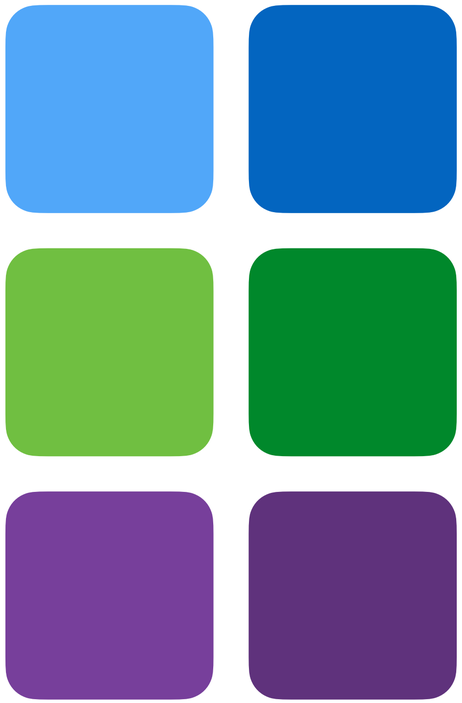
2021/22 - Pretext tasks are cool again

Pretext-Invariant Representation Learning (PIRL)

Ishan Misra, Laurens van der Maaten

Contrastive Learning

Groups of
Related and Unrelated
Images

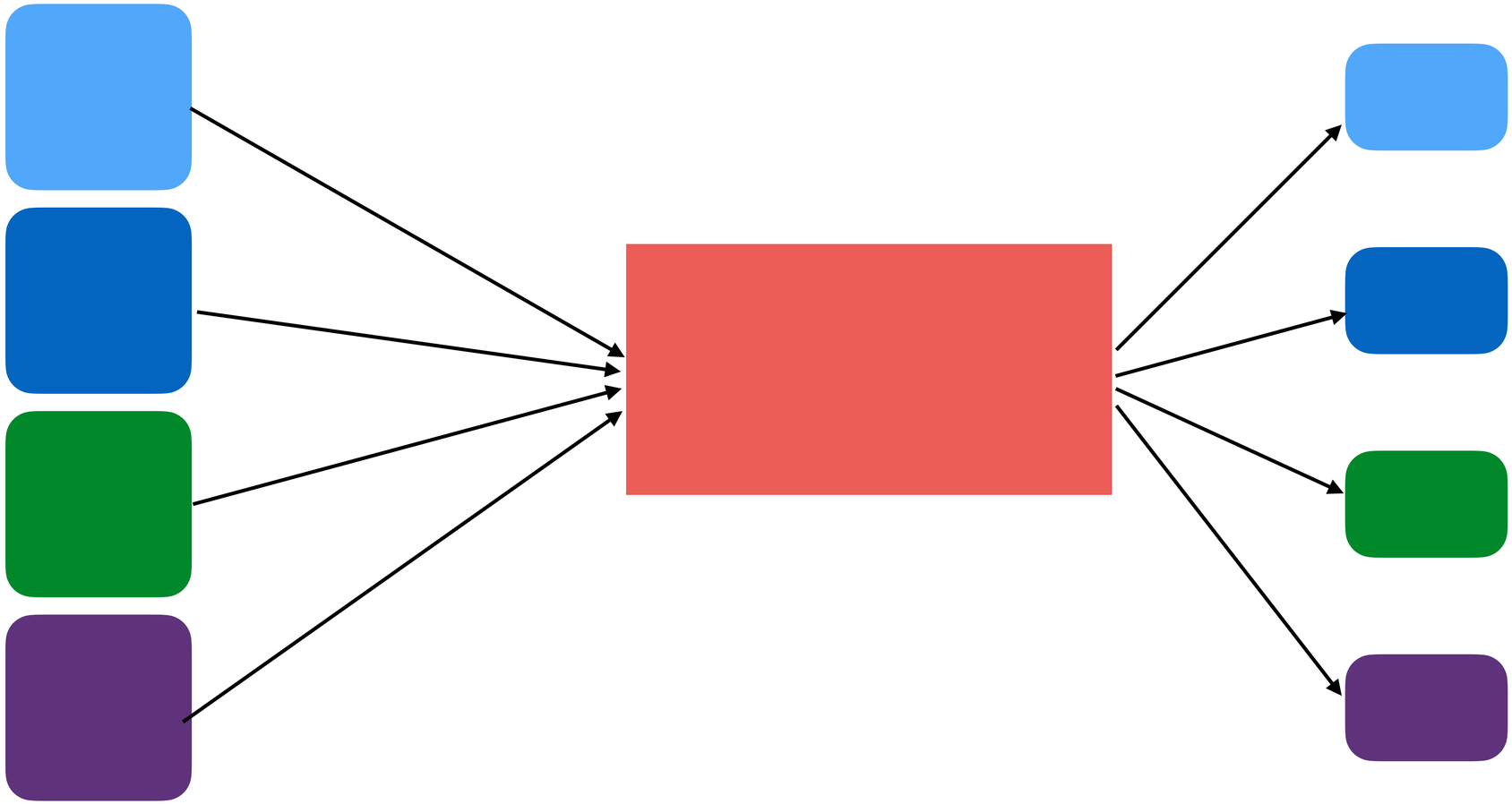


Contrastive Learning

Groups of Related and Unrelated Images

Shared network (Siamese Net)

Image Features (Embeddings)

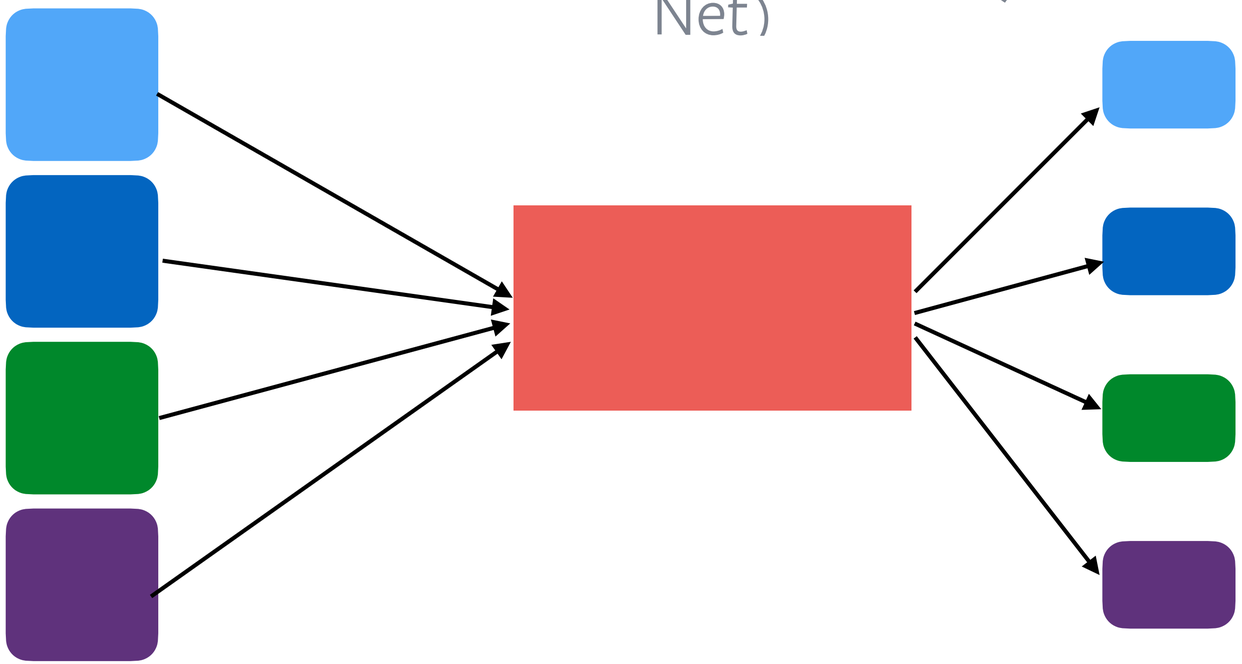


Contrastive Learning

Related and Unrelated Images

Shared network (Siamese Net)

Image Features (Embeddings)

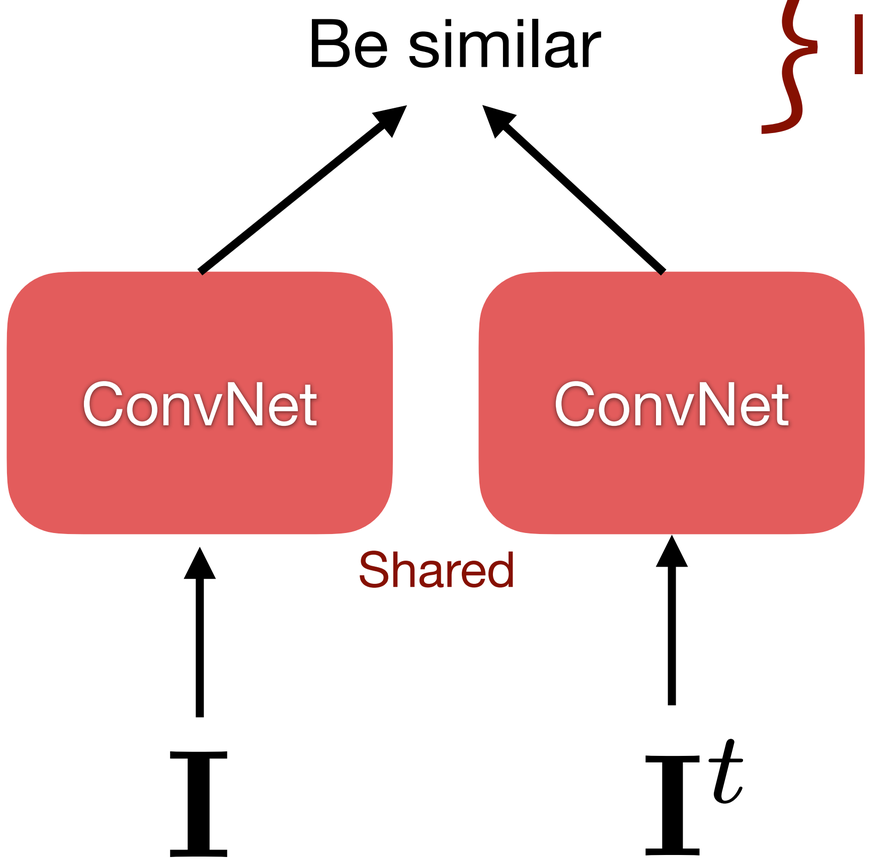


Loss Function

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

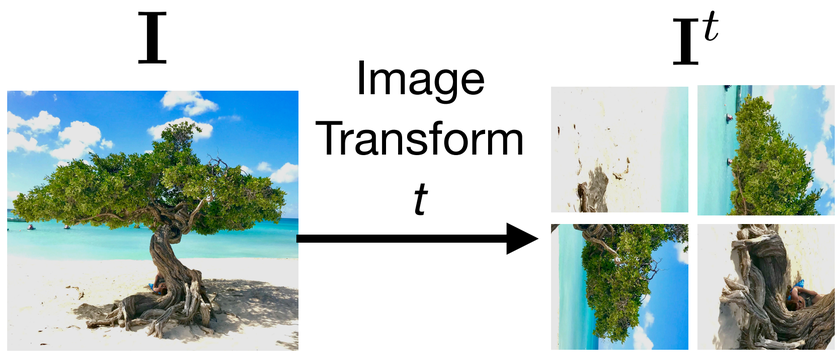
Hadsell et al., 2005, DrLim

PIRL



Invariant to Pretext transform

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

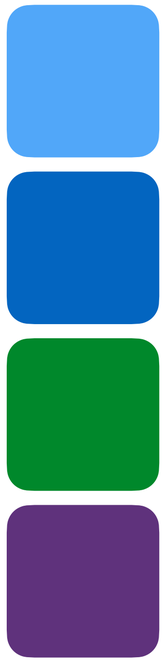


Invariance to

- Data Augmentations
- Multiple views created by pretext task (Jigsaw/Rotation)

Contrastive Learning in PIRL

Dataset

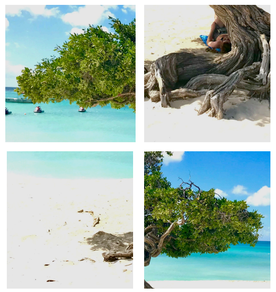


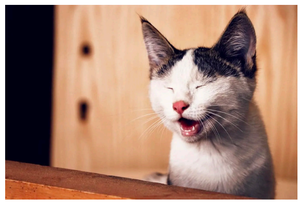
Loss Function

Image Feature & Patch Features

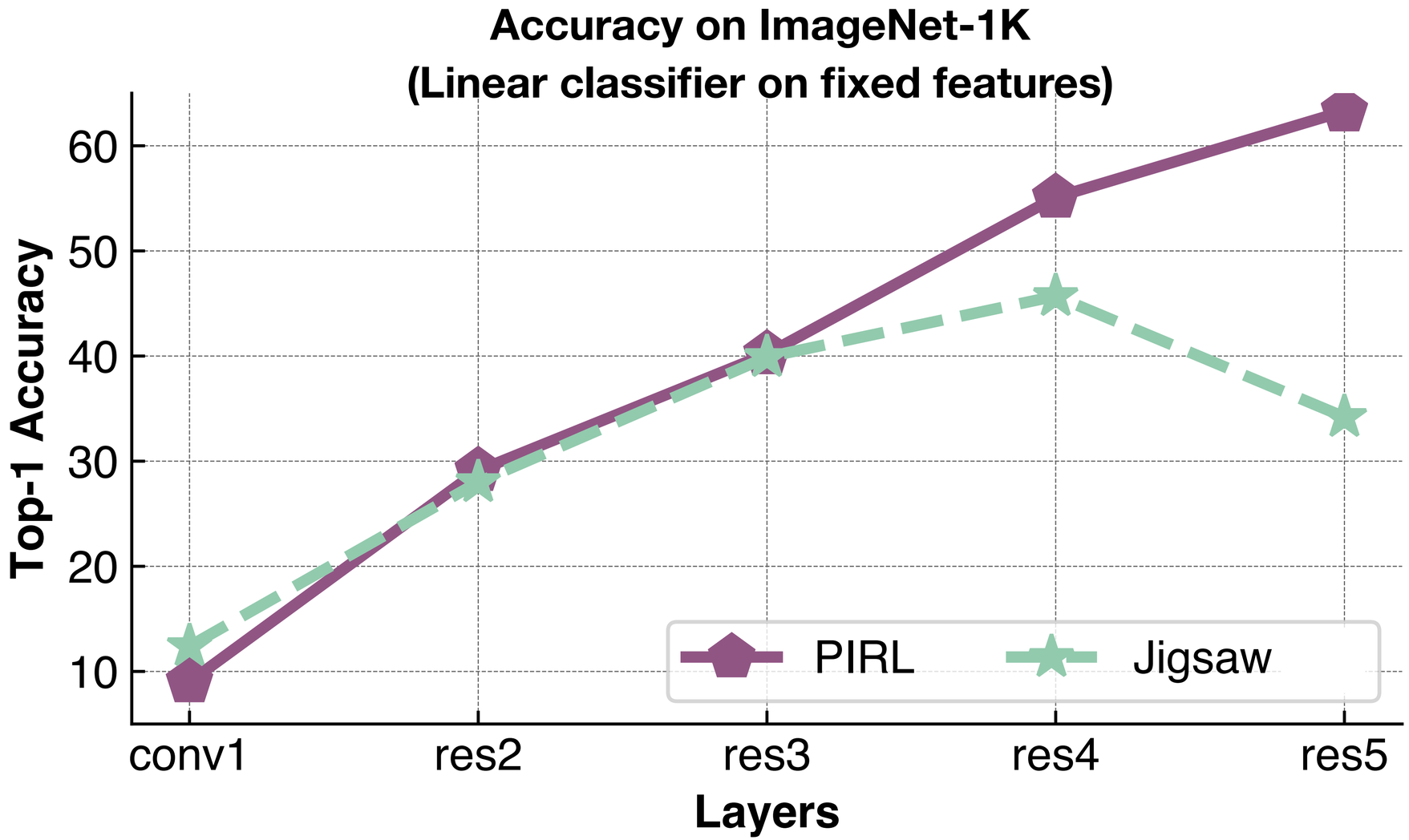
Random Images

$\mathbf{I}$ $\mathbf{I}^t$

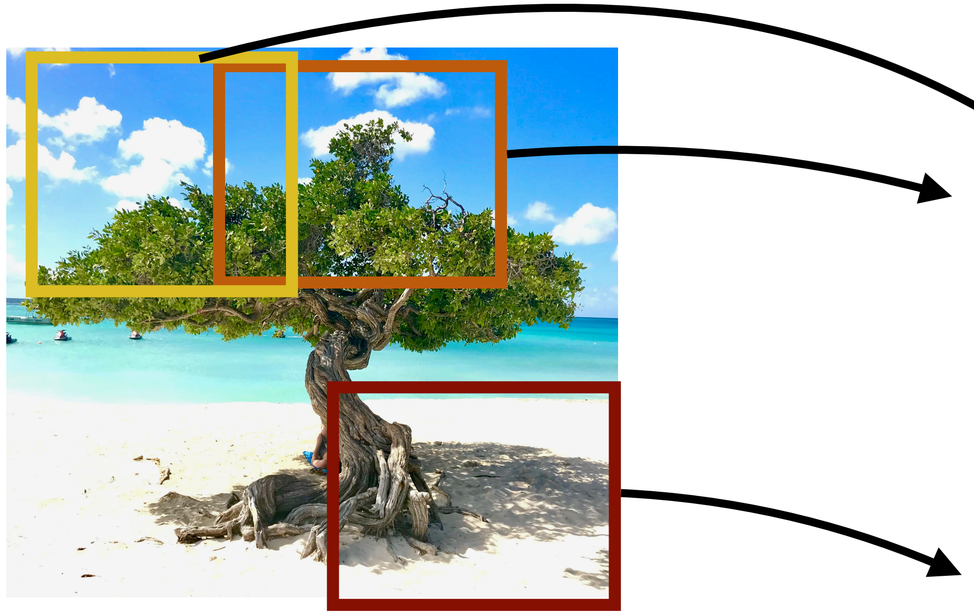




Semantic Features?



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

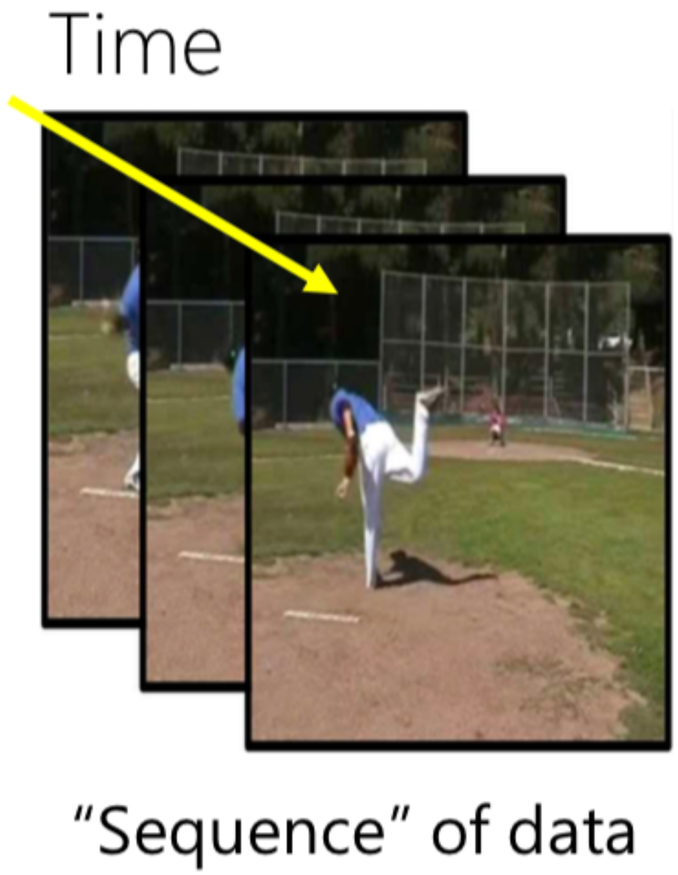
Patches of an image vs. patches of other images

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

Frames of a video

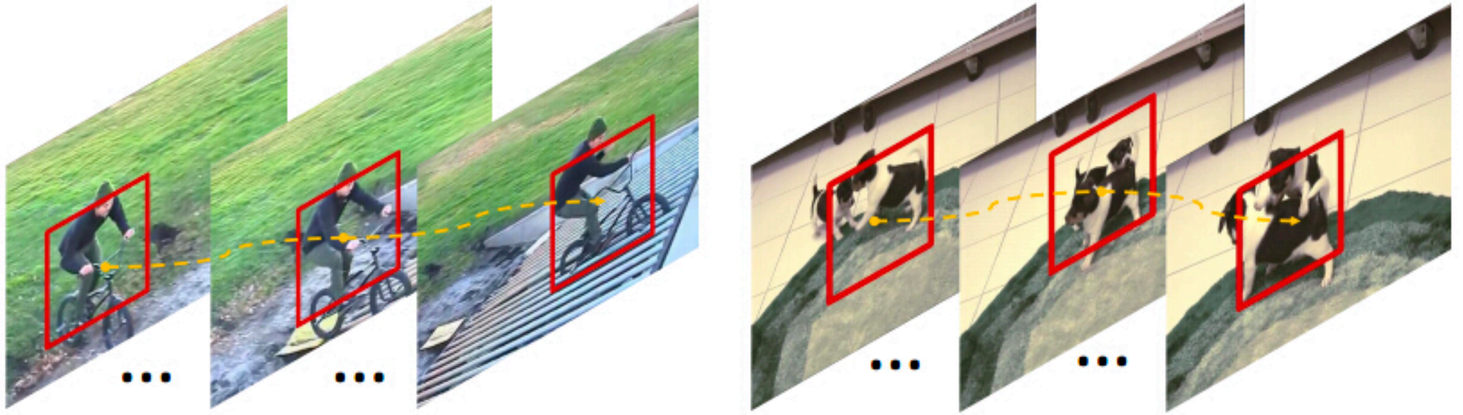


Hadsell et al., 2005, DrLim
van der Oord et al., 2018, CPC

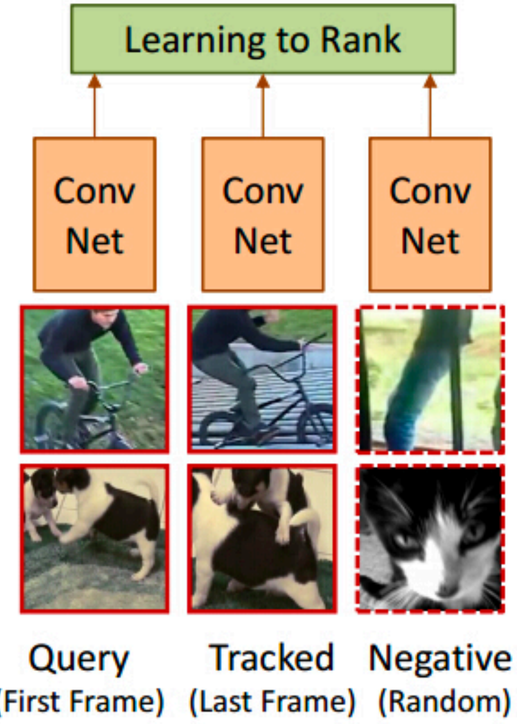
Video & Audio

AVID+CMA - Morgado et al., 2020
GDT - Patrick et al., 2020

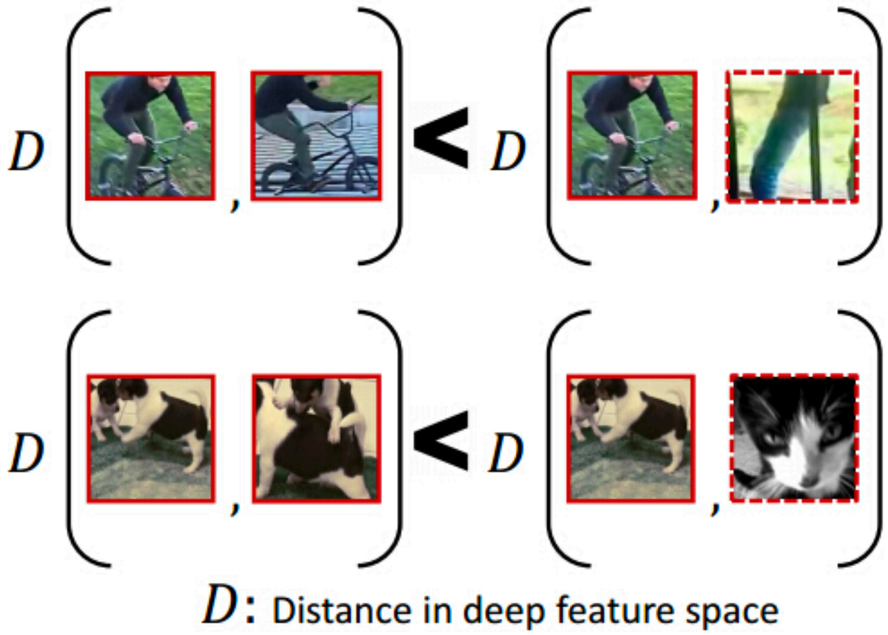
Tracking Objects



(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

3D Point Clouds

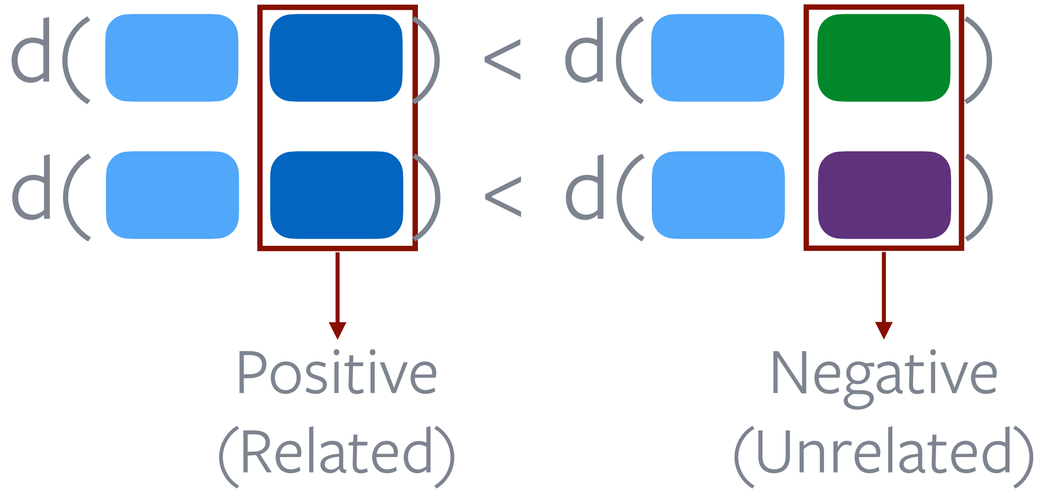
Augmentations

DepthContrast - Zhang et al., ICCV 2021
PointContrast Xie et al., CVPR 2020

Good negatives are necessary

Loss Function

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

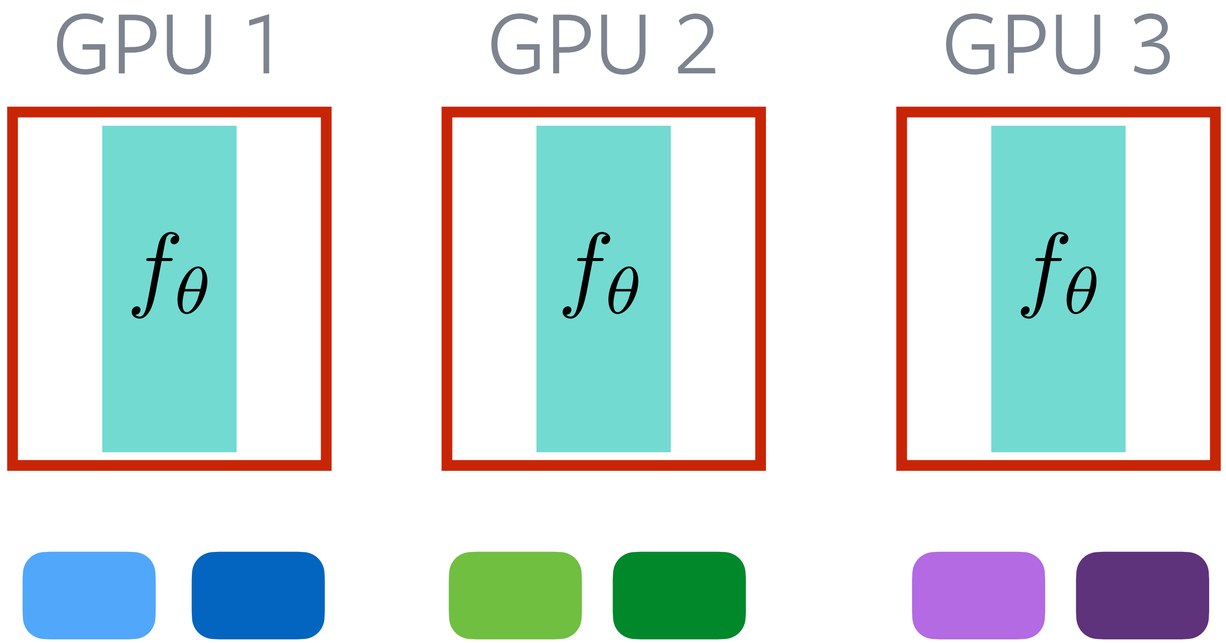


Good negatives are *very* important in contrastive learning

Hadsell et al., 2005, DrLim

SimCLR

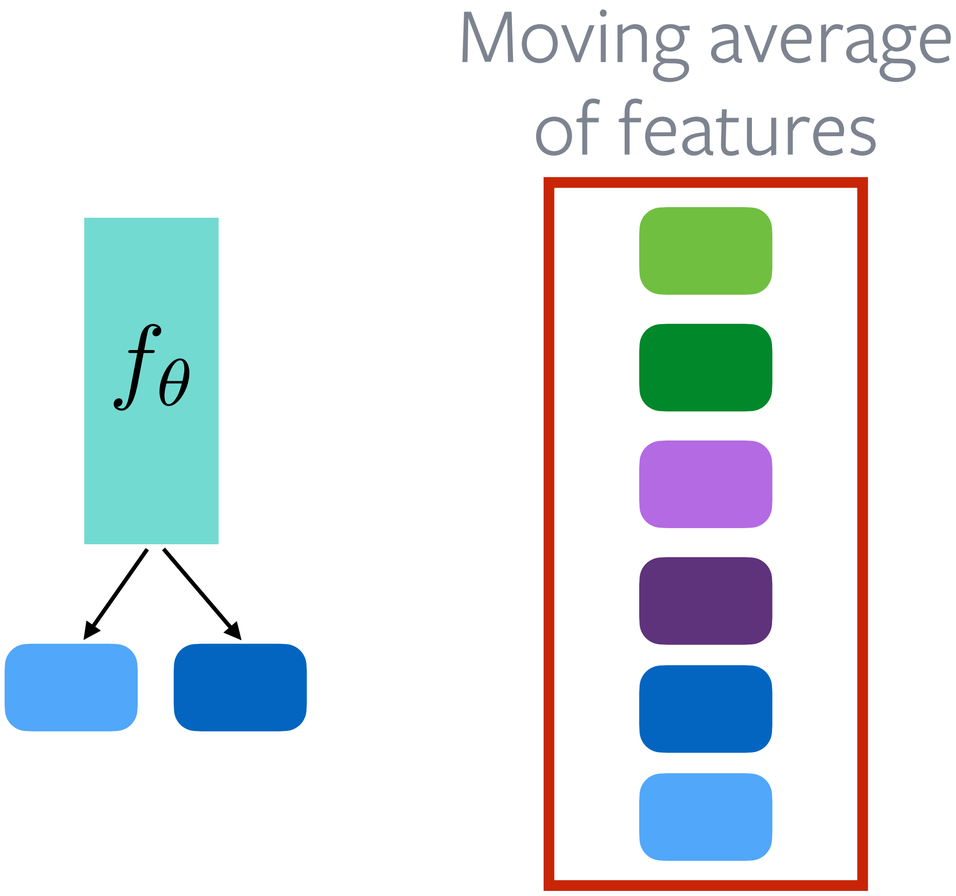
- Large batch size - e.g. in SimCLR
- Pros - Simple to implement
- Cons - Large batch size



A simple framework for contrastive learning - Chen et al., 2020

Memory Bank

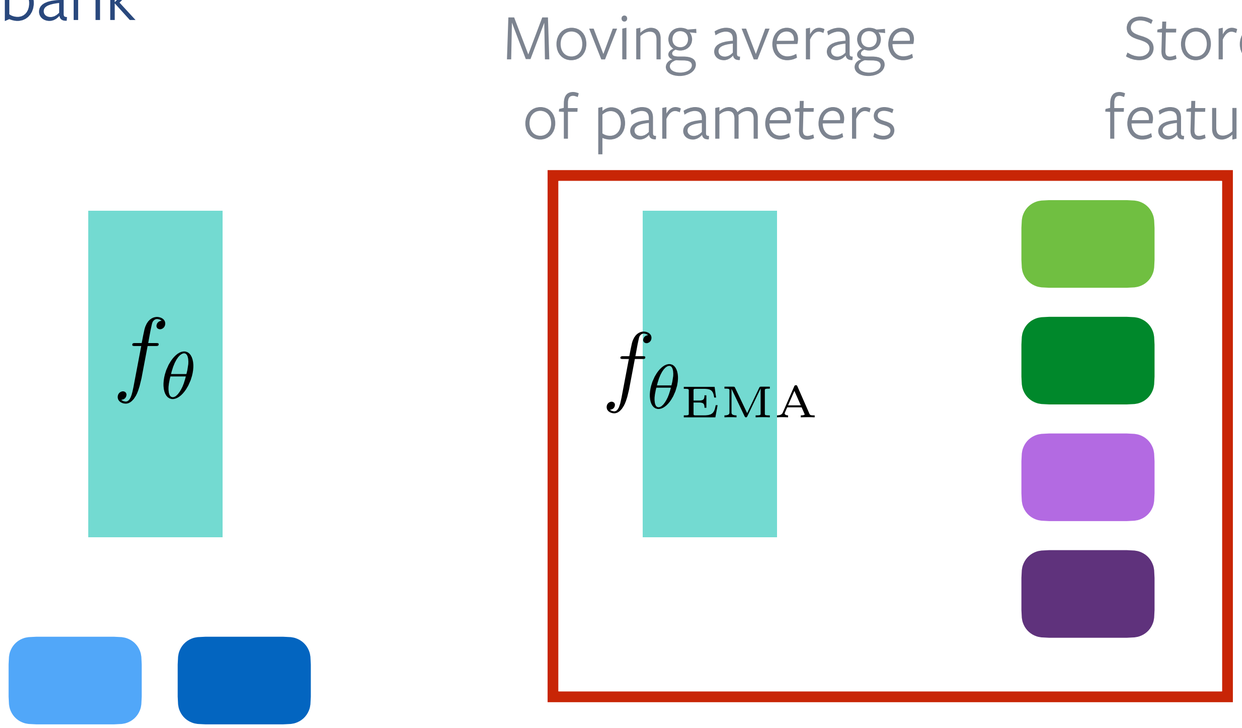
- Maintain a "memory bank" -- momentum of activations
- Pros - compute efficient
- Cons - Needs large memory, not "online"



Non parametric Instance Discrimination - Wu et al., 2018

MoCo

- Maintain "momentum" network - MoCo
- Pros - online
- Cons - extra memory for parameters/stored features, extra fwd pass compared to memory bank



Momentum Contrast - He et al., 2019

Many ways to avoid trivial solutions

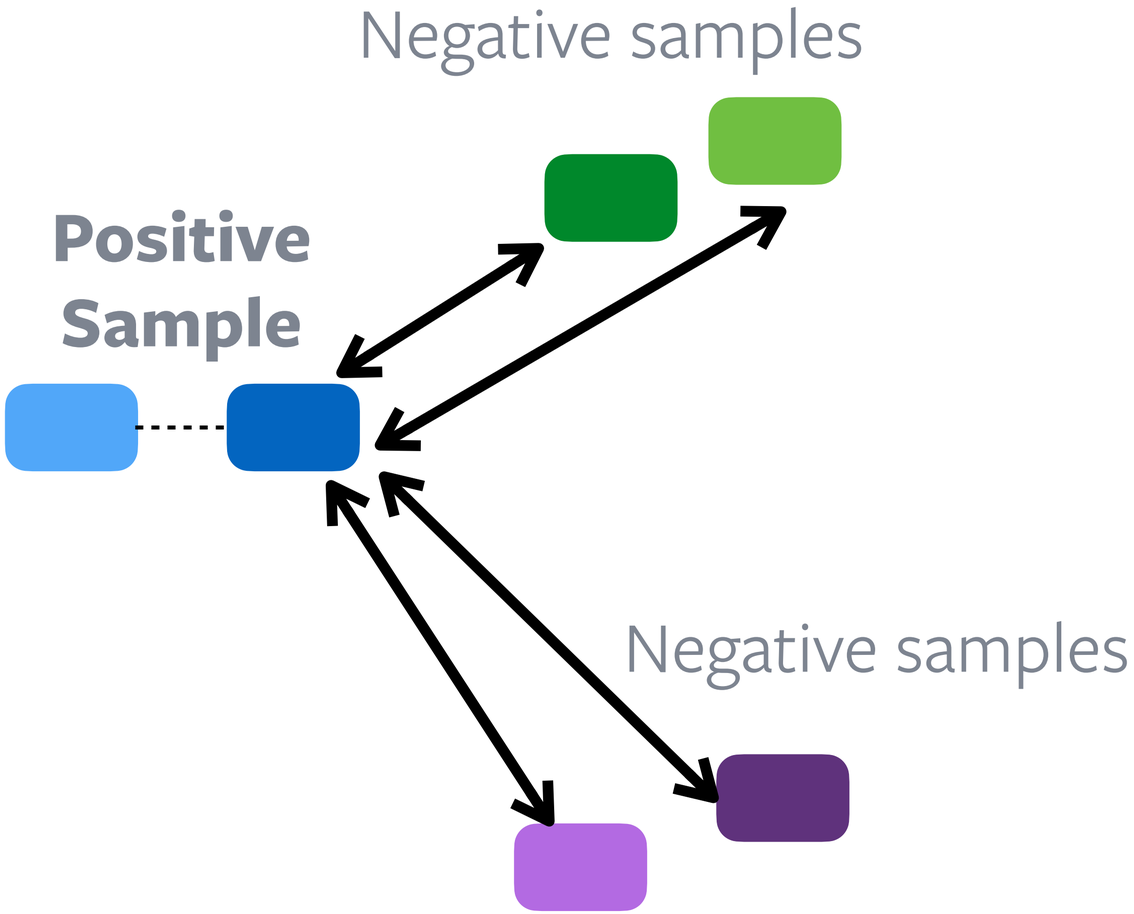
Similarity Maximization Objective

- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam

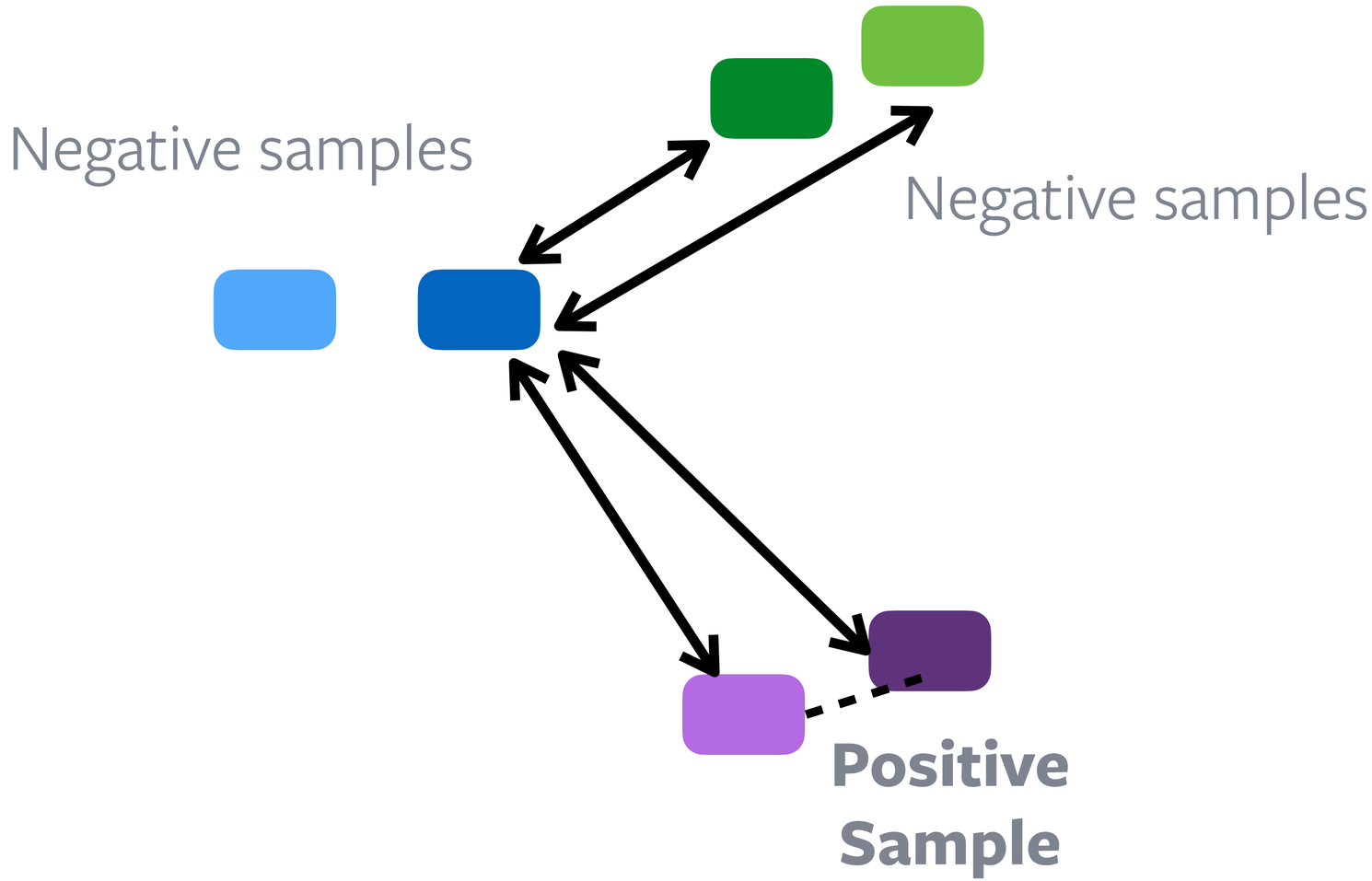
Redundancy Reduction Objective

- Redundancy Reduction
  - Barlow Twins

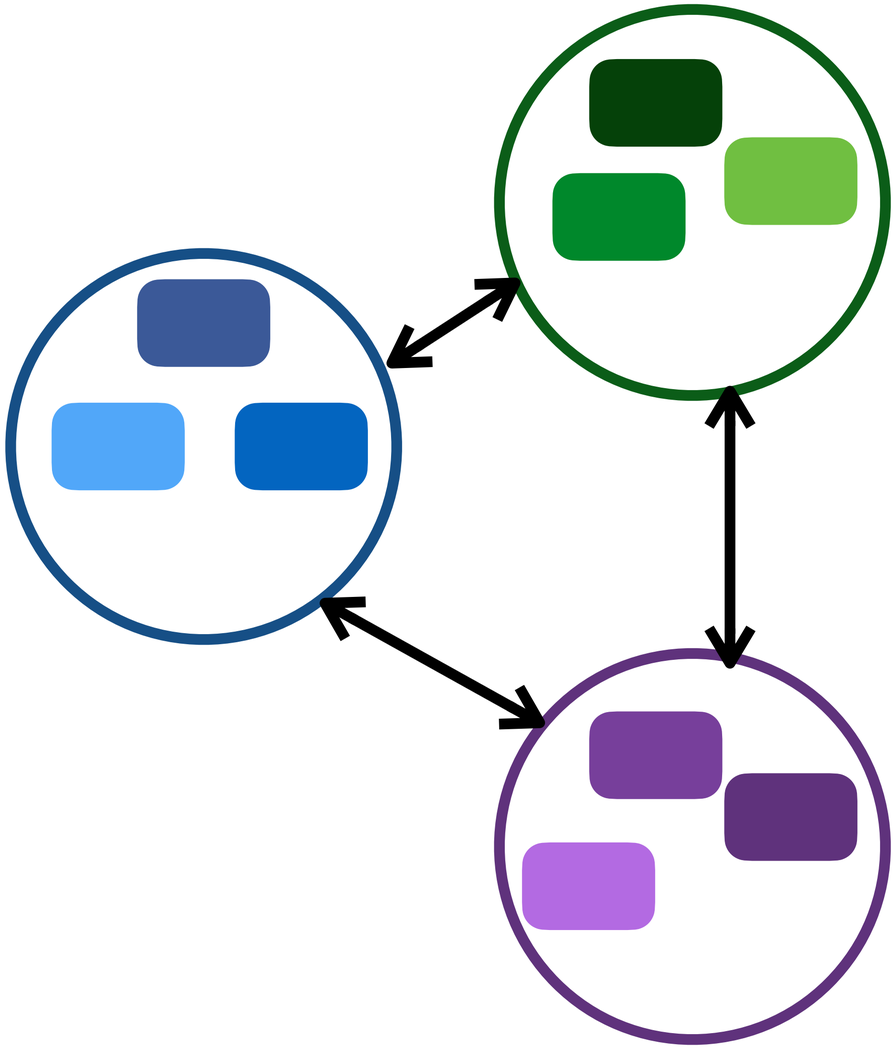
Contrastive learning -- what does it do?



Contrastive learning -- what does it do?

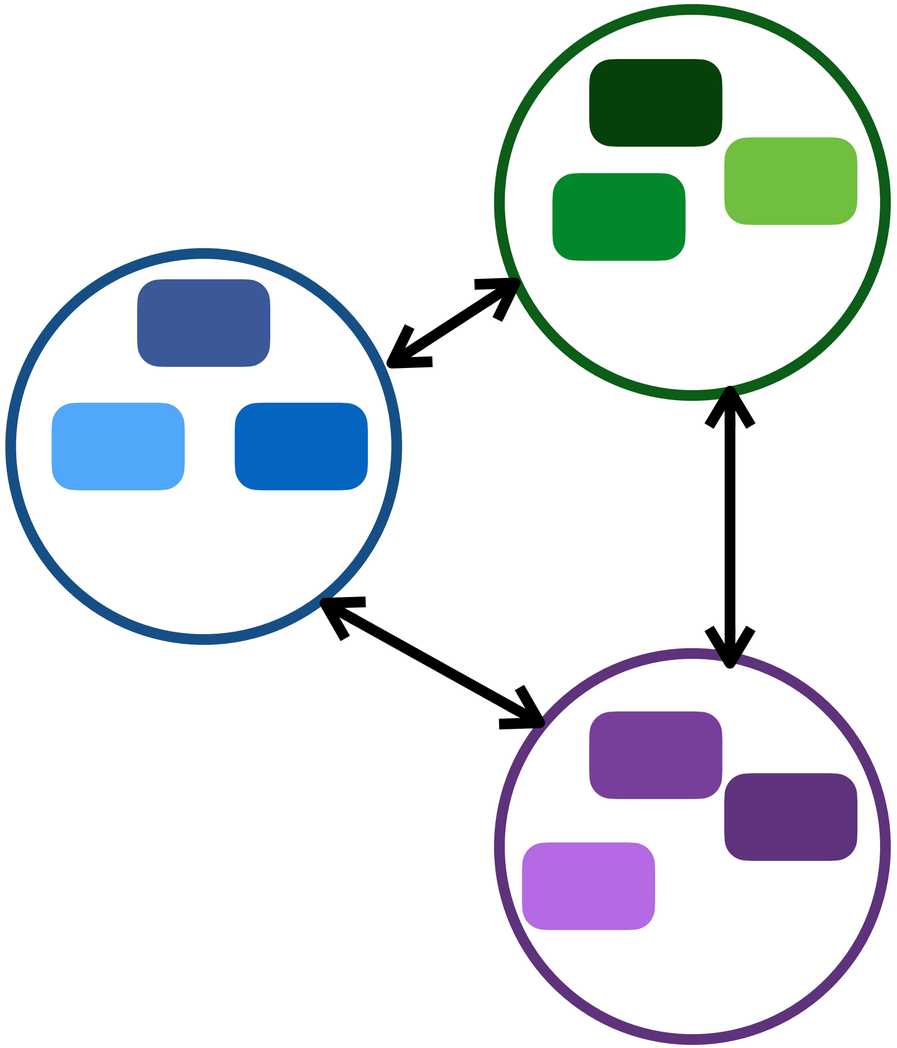


Contrastive Learning => Groups in feature space



Creates groups
in the feature space

Clustering creates groups too



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

Key Idea

- What we want $f_\theta(I) = f_\theta(\text{augment}(I))$

- How we do it - I and augment(I) must belong to the same "group" or cluster

- Prevent trivial solutions by controlling the clustering process

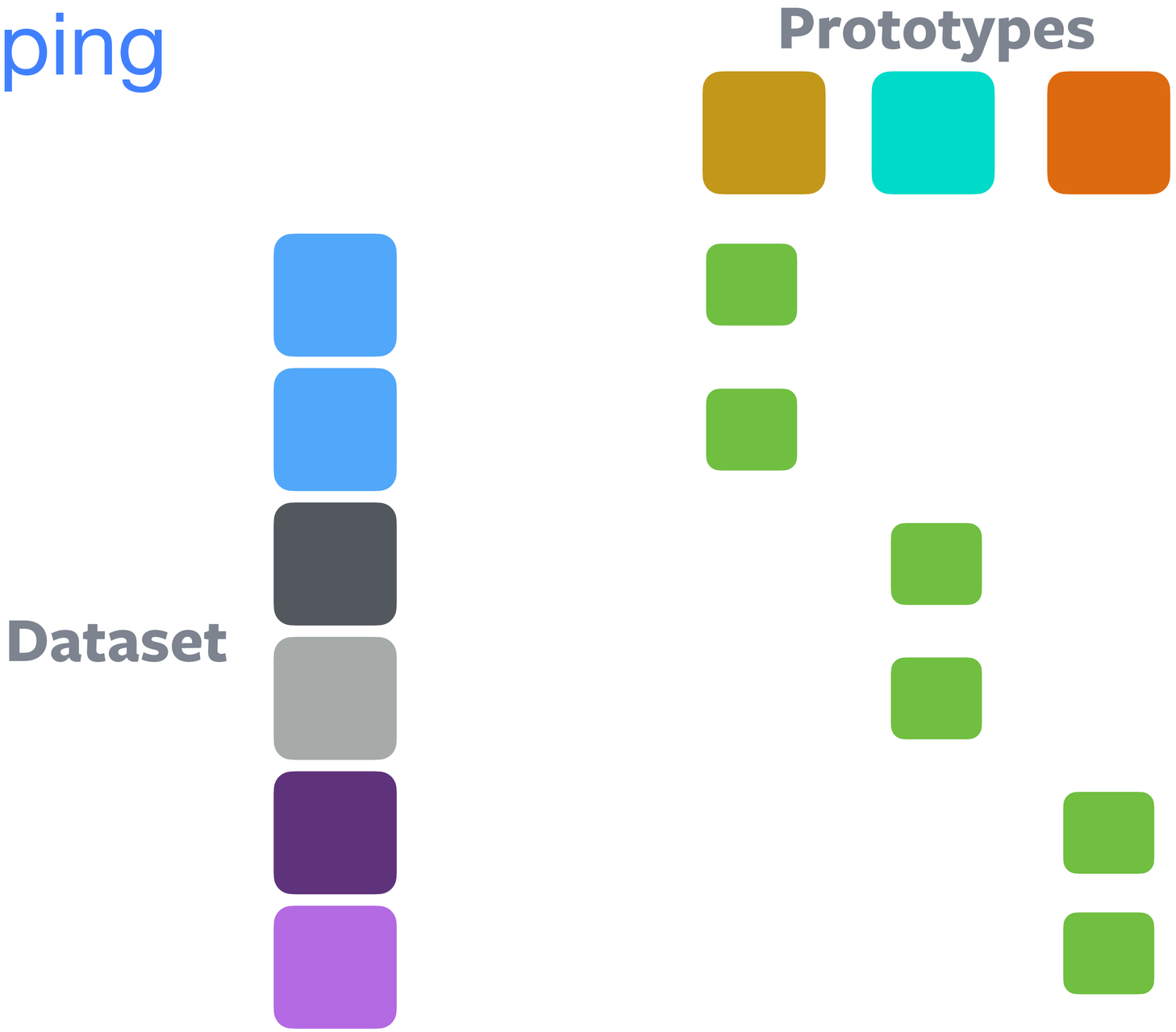
Grouping

Similarity of
dataset sample & prototypes

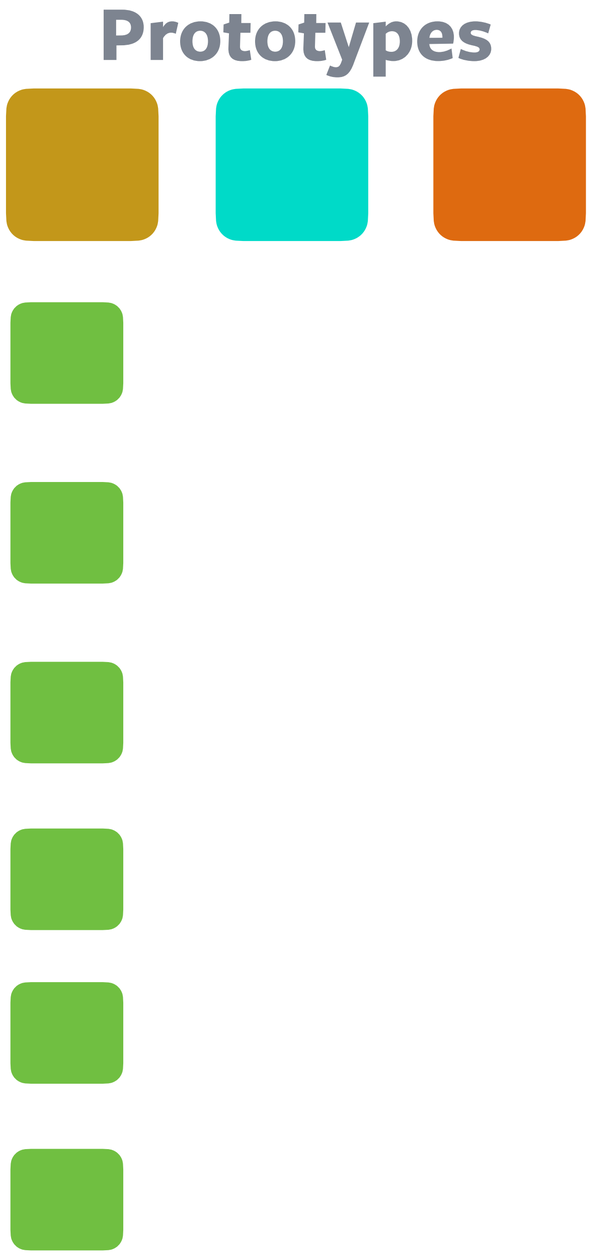
(which cluster does a sample belong to?)

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

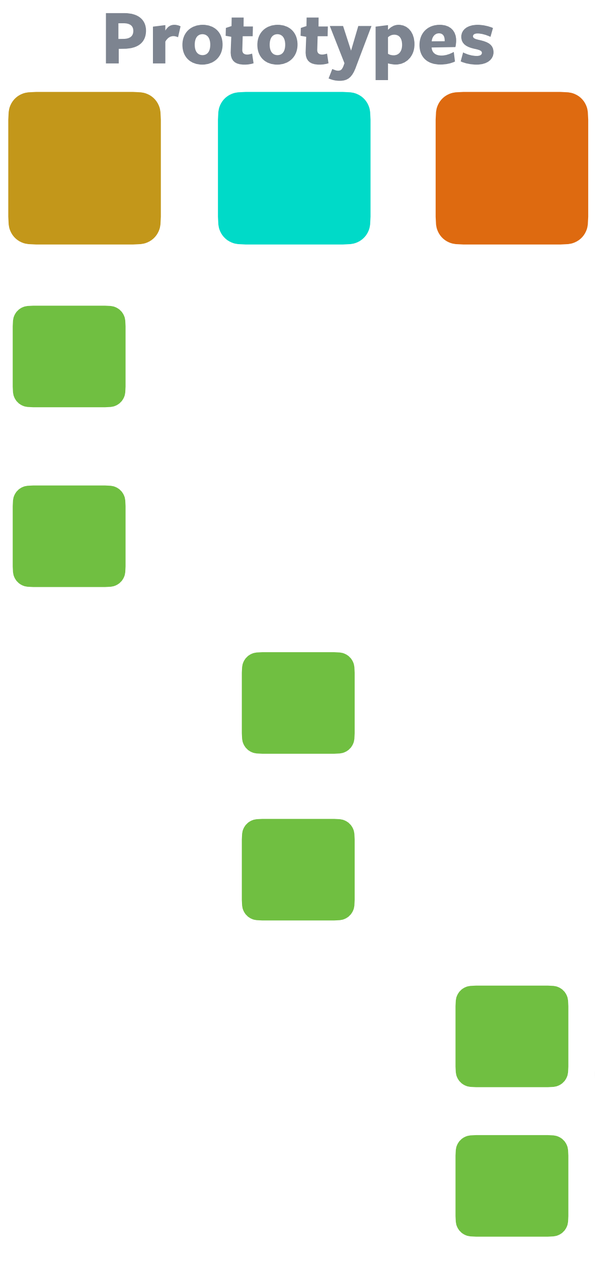
Grouping



Trivial Solutions?



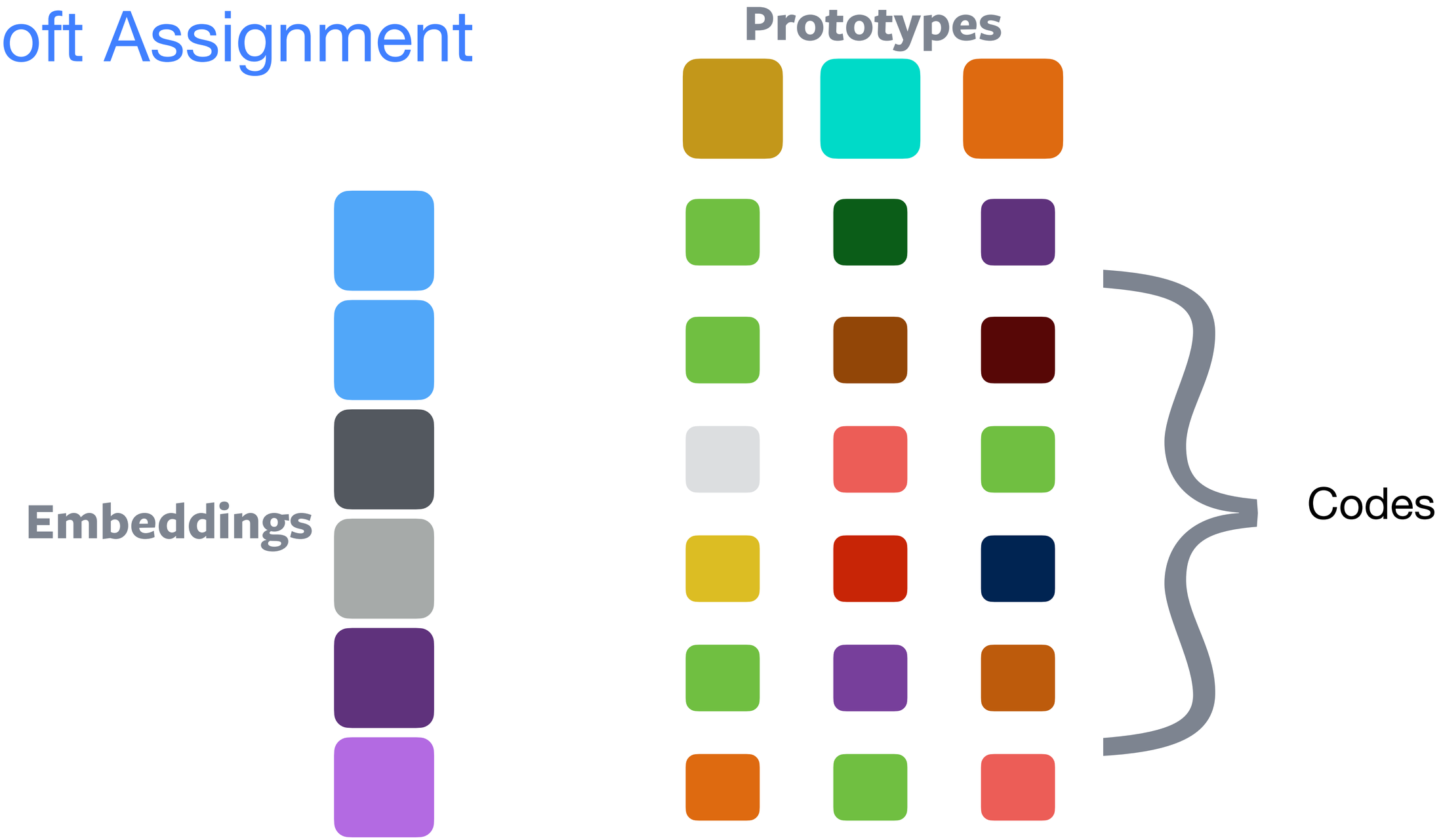
Grouping

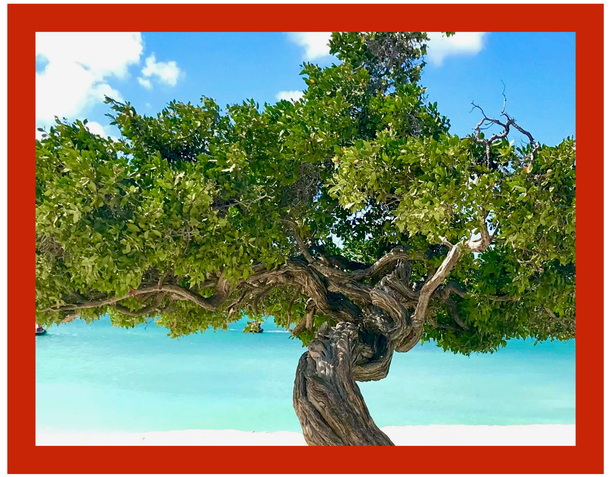


Equipartition constraint --
Given N samples and K prototypes,
each prototype is most similar to N/K samples

Implemented using
Optimal Transport (Sinkhorn-Knopp)

Soft Assignment





Prototypes

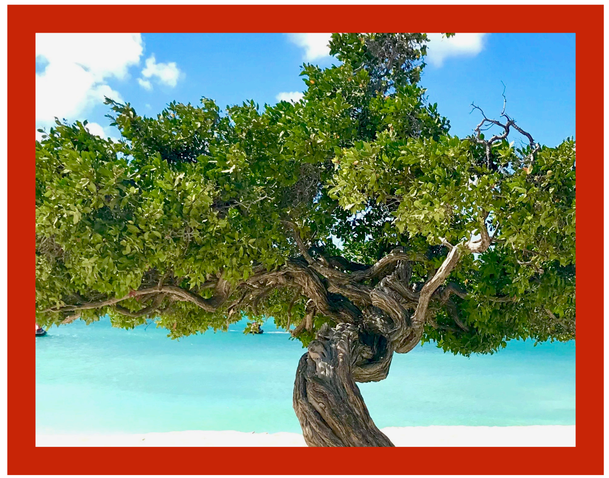
f_θ

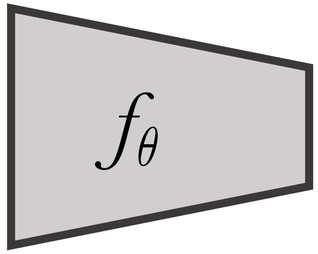
f_θ

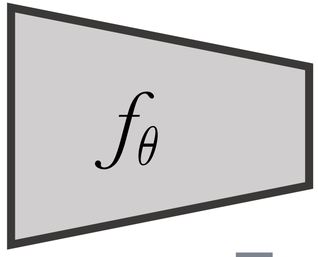
Code 1

Code 2

Embeddings







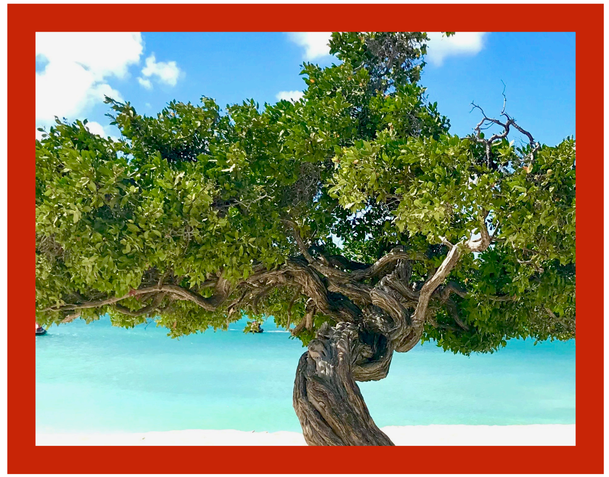
Prototypes

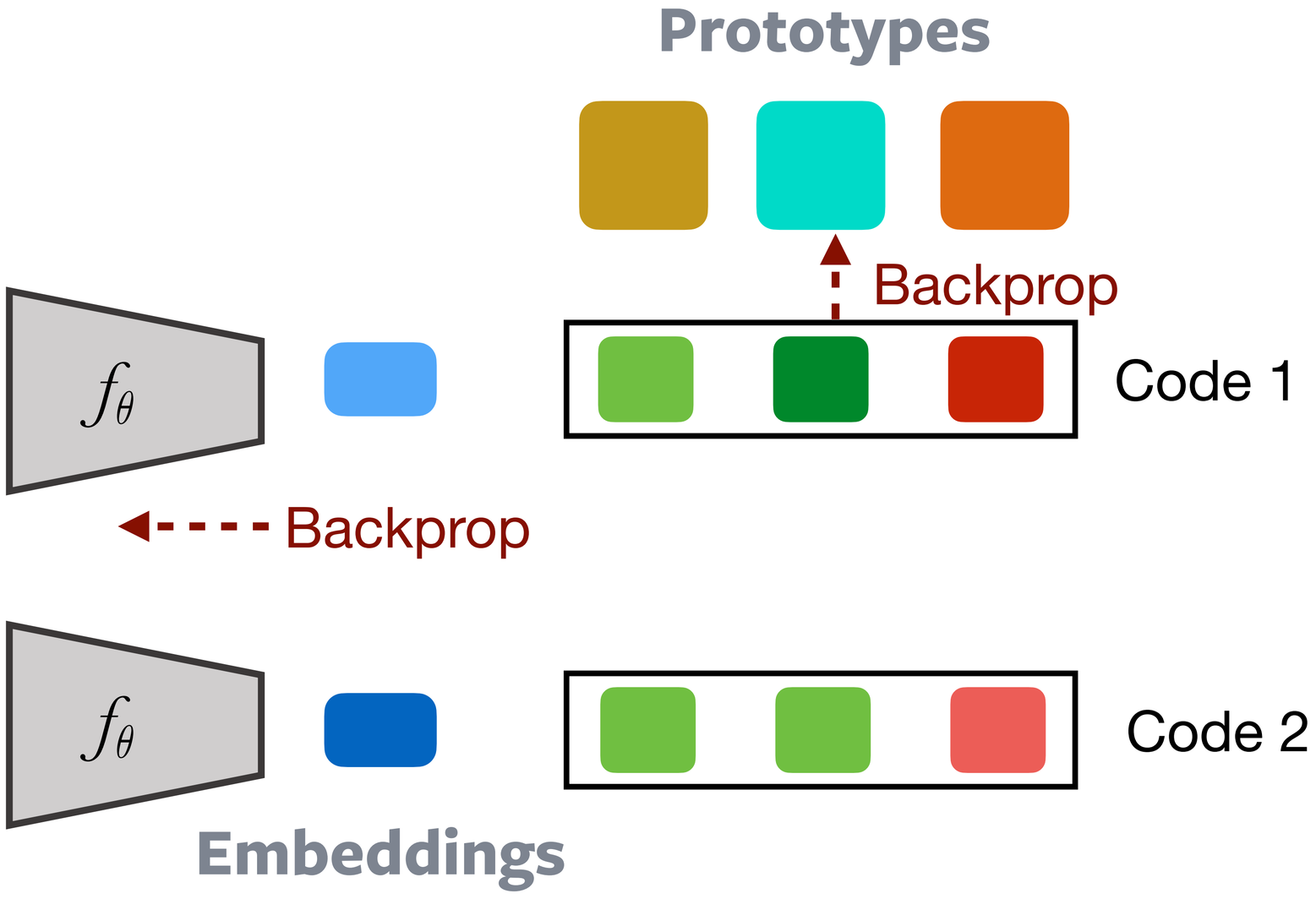
Code 1

Code 2

Predict

Embeddings





Not contrastive!

Key Results

	Linear Classifier (Fixed Features)			Detection (Fine-tuned)	
	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**

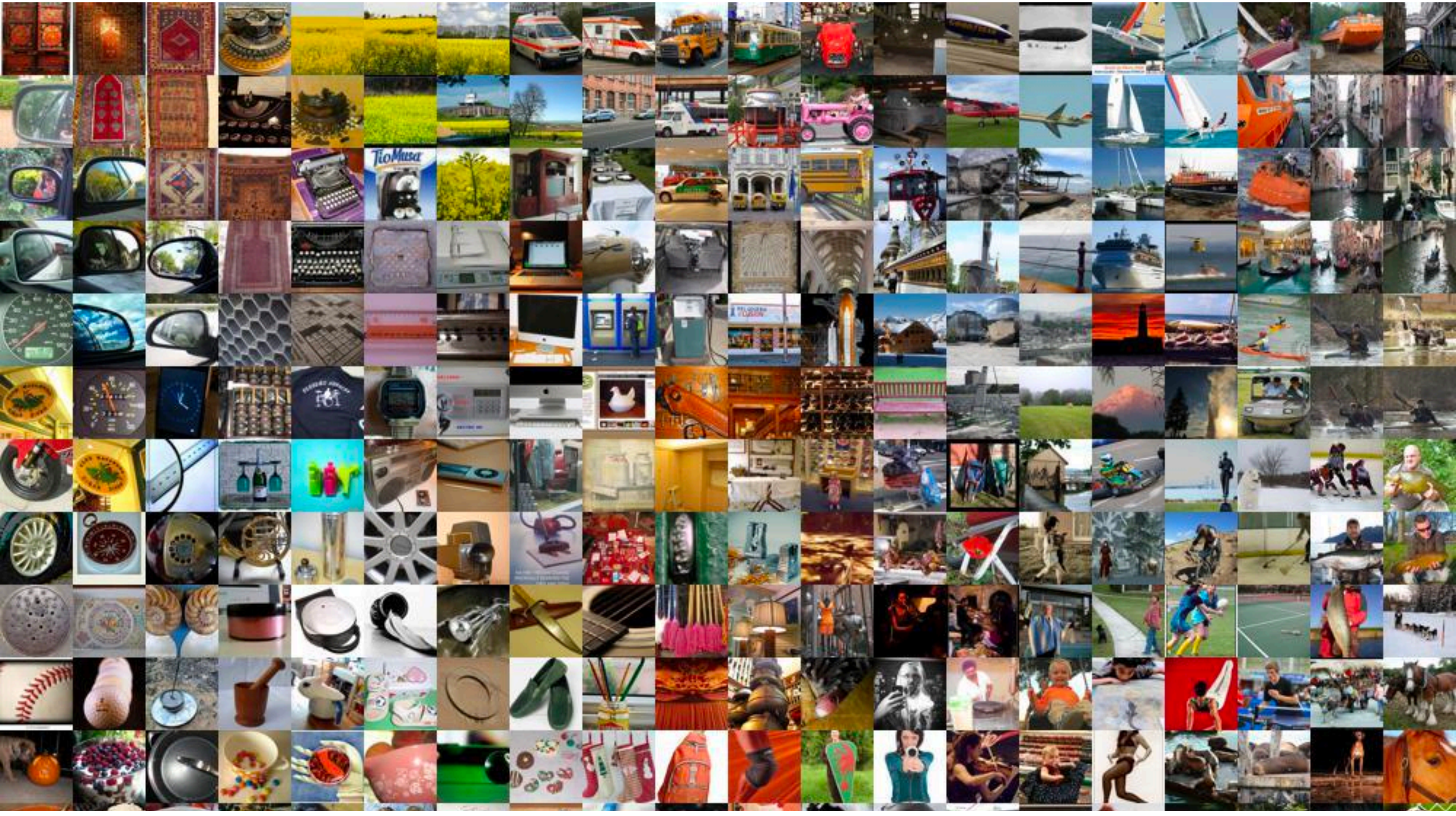
Advantages of SwAV (Clustering-based)

- 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

Pretrained on ImageNet without labels


TioMusa

Curated pretraining data

ImageNet data is "curated"

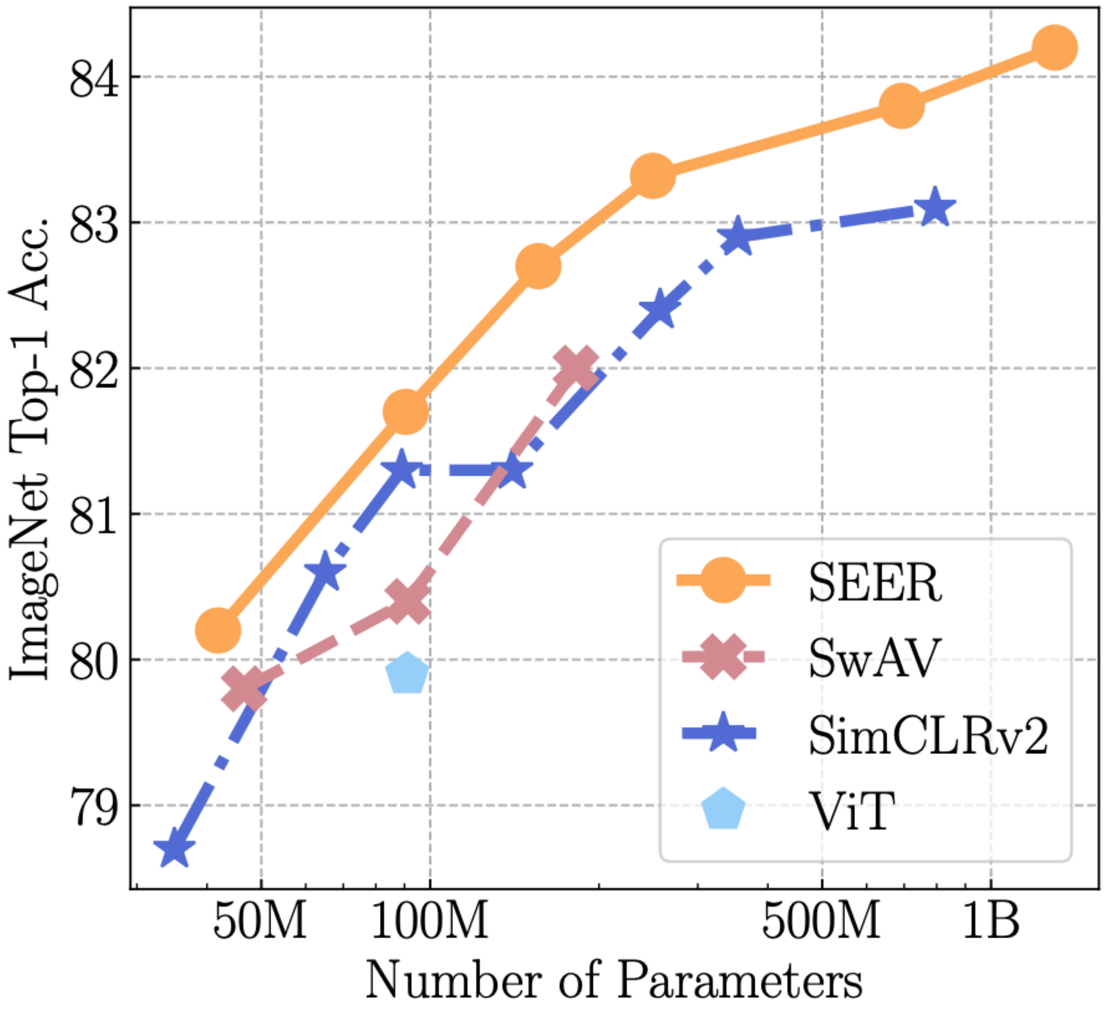
- All images belong to 1000 classes
- All images contain a prominent object
- Very limited clutter

Real world data

Images have

- different distributions (cartoon images, memes)
- no single prominent object

SEER: Learning from uncurated images



Train on 1.3 billion random images

Images are NOT filtered in any way

RegNet trained on 1.3B random internet images

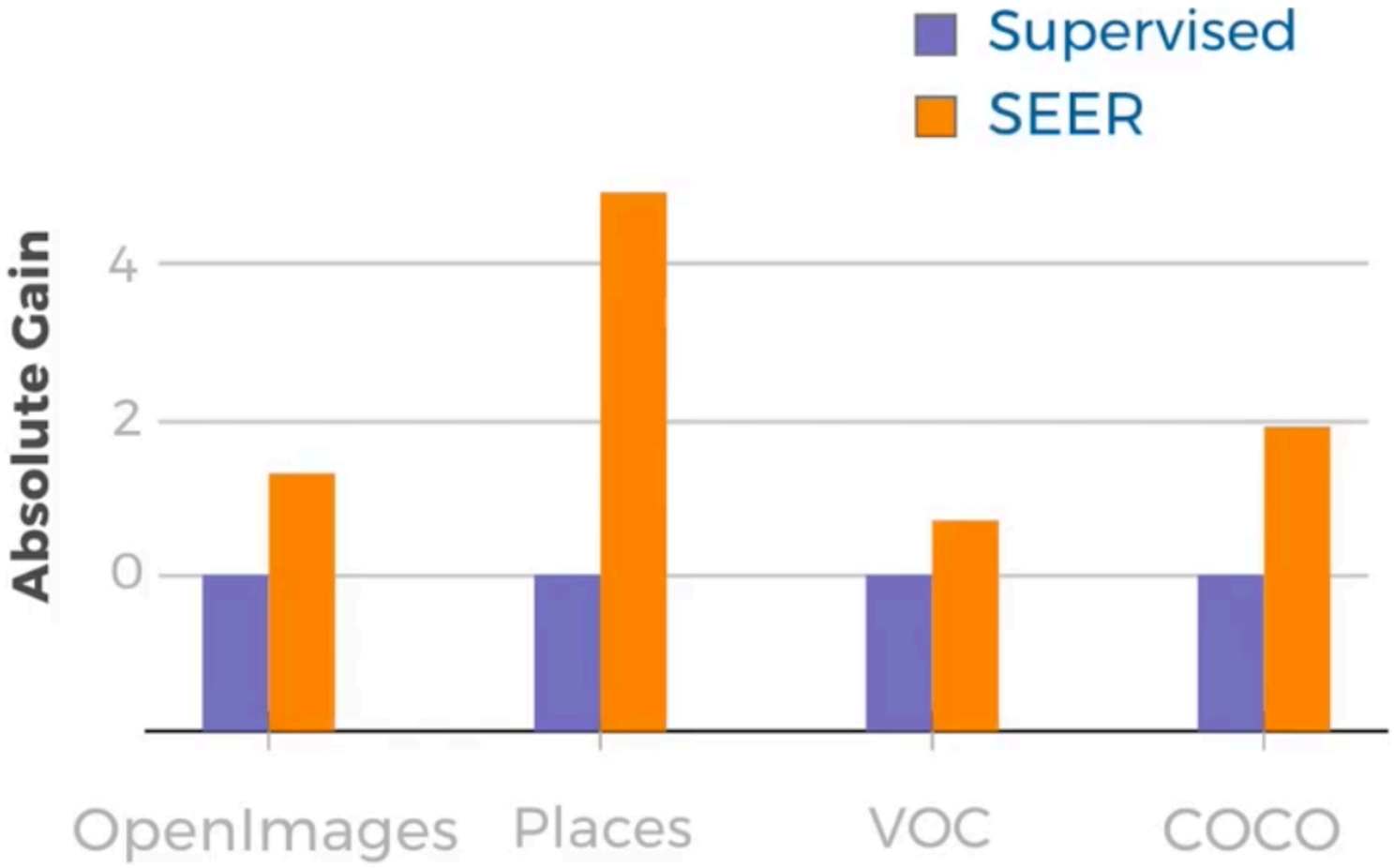
ResNet trained on ImageNet

ResNet (modified) trained on ImageNet

Vision Transformer trained on ImageNet

SEER - Goyal et al., 2021

SEER: Improves performance on benchmarks



SEER - Self-supervised vision model on
1 billion random internet images. **No Labels/metadata.**

SEER - Goyal et al., 2021

SEER: AI that works for everyone

Spices (Nepal)

Supervised - cleaning equipment, kitchen sink, shower

SEER - spices, medication, bowls

Stove (China)

Supervised - lock on front door, power switches, cooking utensils

SEER - cooking utensils, stove

SEER - Goyal et al., 2021

SEER: Learning from uncurated images

Method	Pretrain images	Curated	Arch.	Params	ImageNet top-1
Hashtag prediction	1B	Yes	X101-32x8d	91M	82.6
SEER	1B	No	X101-32x8d	91M	81.6

SEER uses no labels and works on random images

Comparable performance to networks trained on curated data with weak supervision

Architecture: ResNeXt-101-32x8d

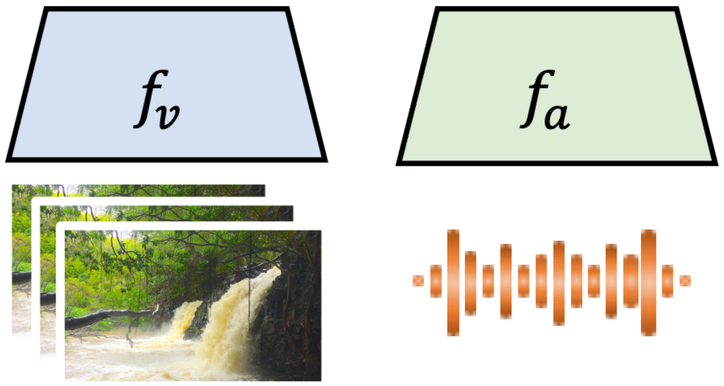
Audio Visual Instance Discrimination with Cross Modal Agreement (AVID + CMA)

Pedro Morgado, Nuno Vasconcelos, Ishan Misra



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

Contrastive (Audio Video Instance Discrimination)



Positives **Negatives**

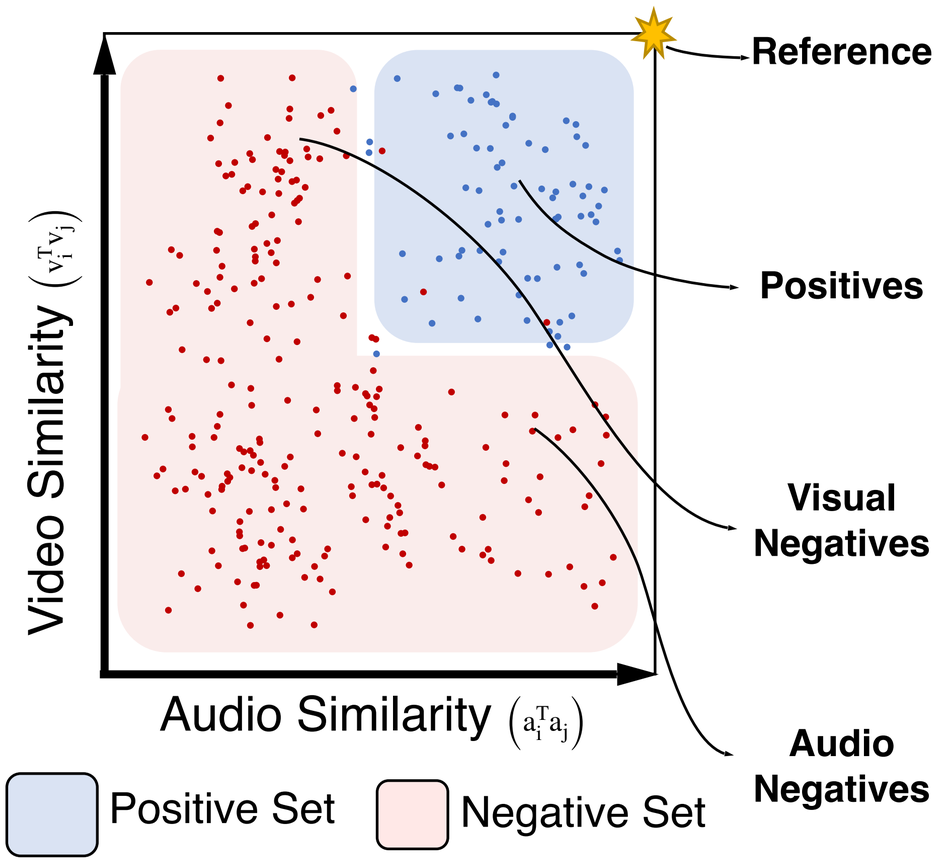
d( ) < d( )

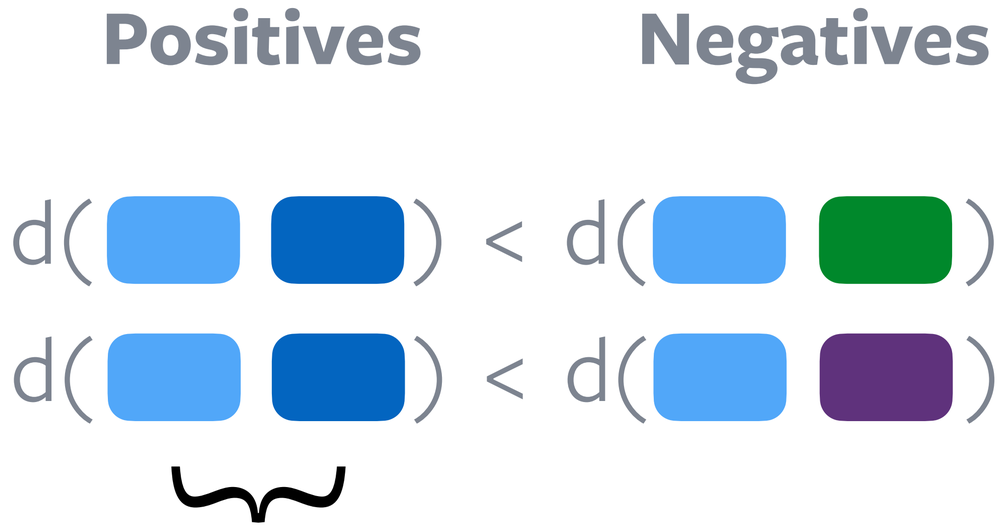
d( ) < d( )

Audio & Video
(same sample)

Relate to other video/audio
using negatives

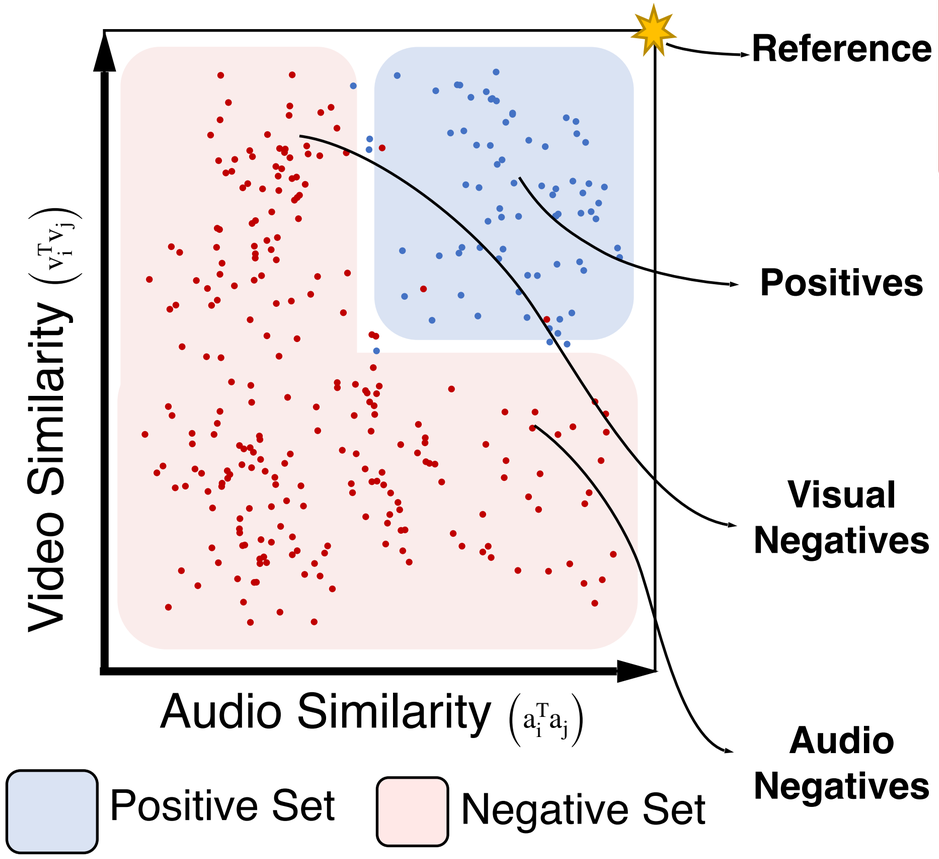
Grouping using Audio-visual Agreements (CMA)

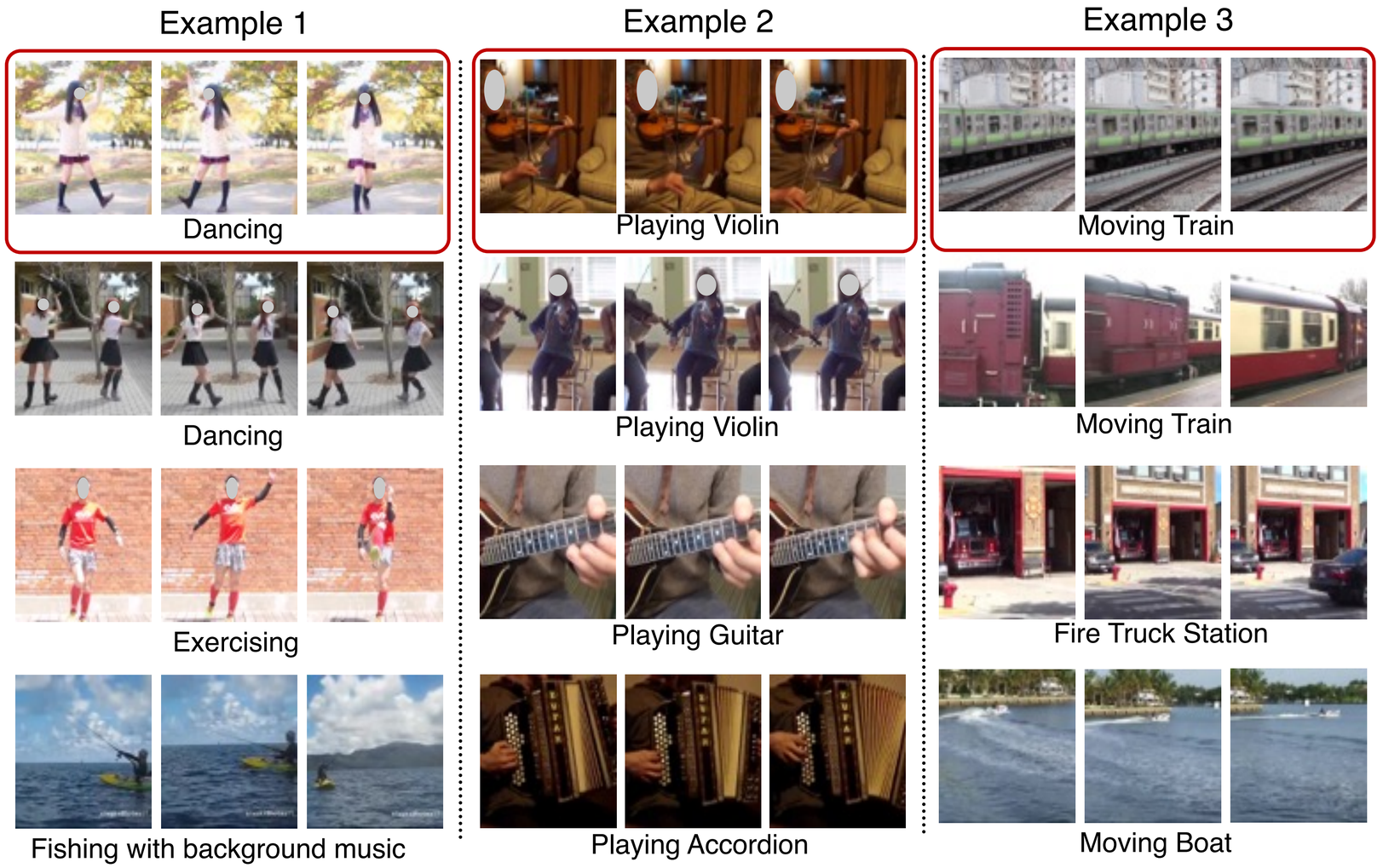




Videos that are similar in audio & video features

Grouping using Audio-visual Agreements (CMA)





Many ways to avoid trivial solutions

Similarity Maximization Objective

- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam, DINO

Redundancy Reduction Objective

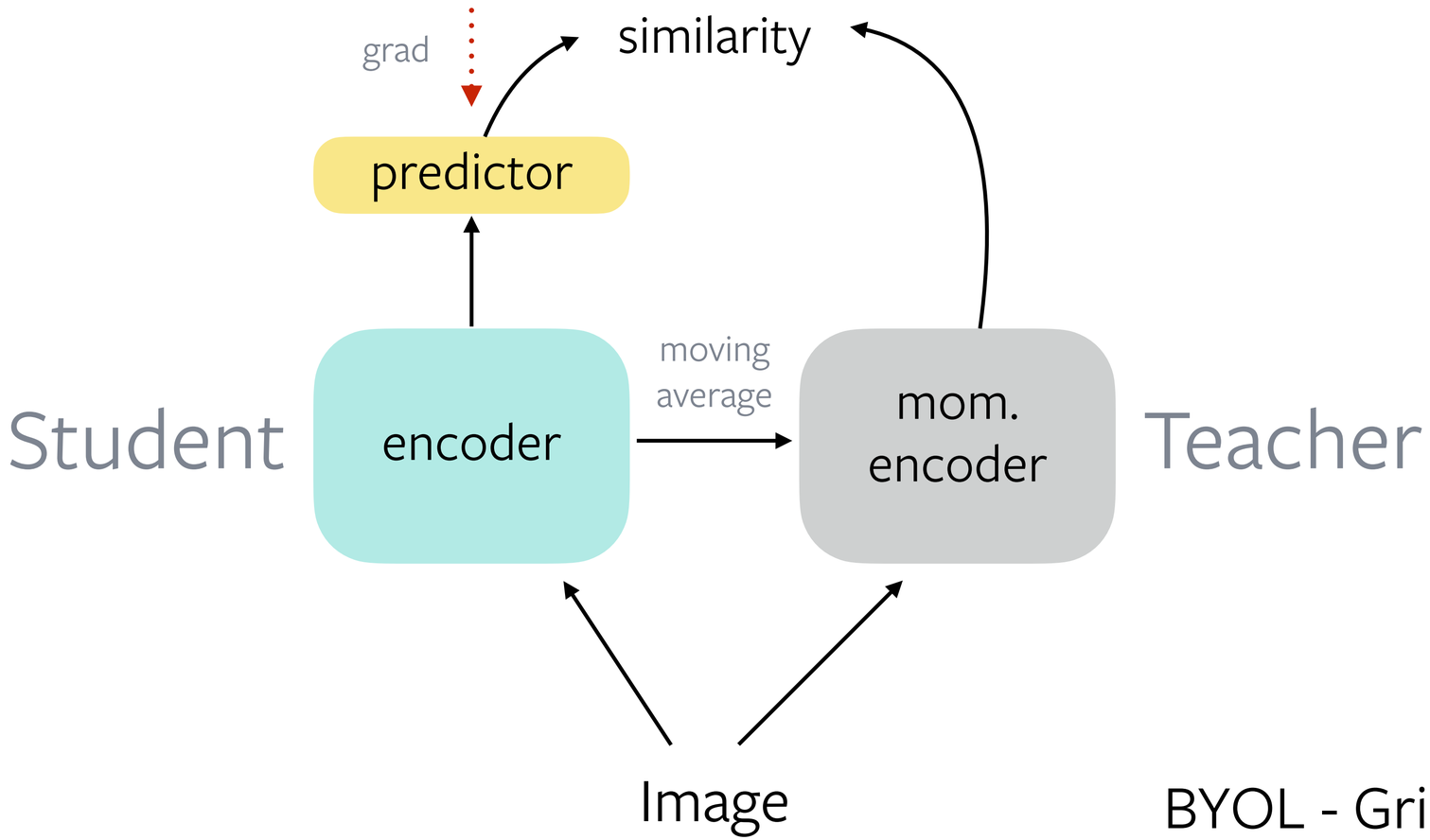
- Redundancy Reduction
  - Barlow Twins

Distillation

- What we want $f_\theta(I) = f_\theta(\text{augment}(I))$
- How we do it $f_\theta^{\text{student}}(I) = f_\theta^{\text{teacher}}(\text{augment}(I))$
- Prevent trivial solutions by asymmetry
  - Asymmetric **learning rule** between student teacher
  - Asymmetric **architecture** between student teacher

BYOL

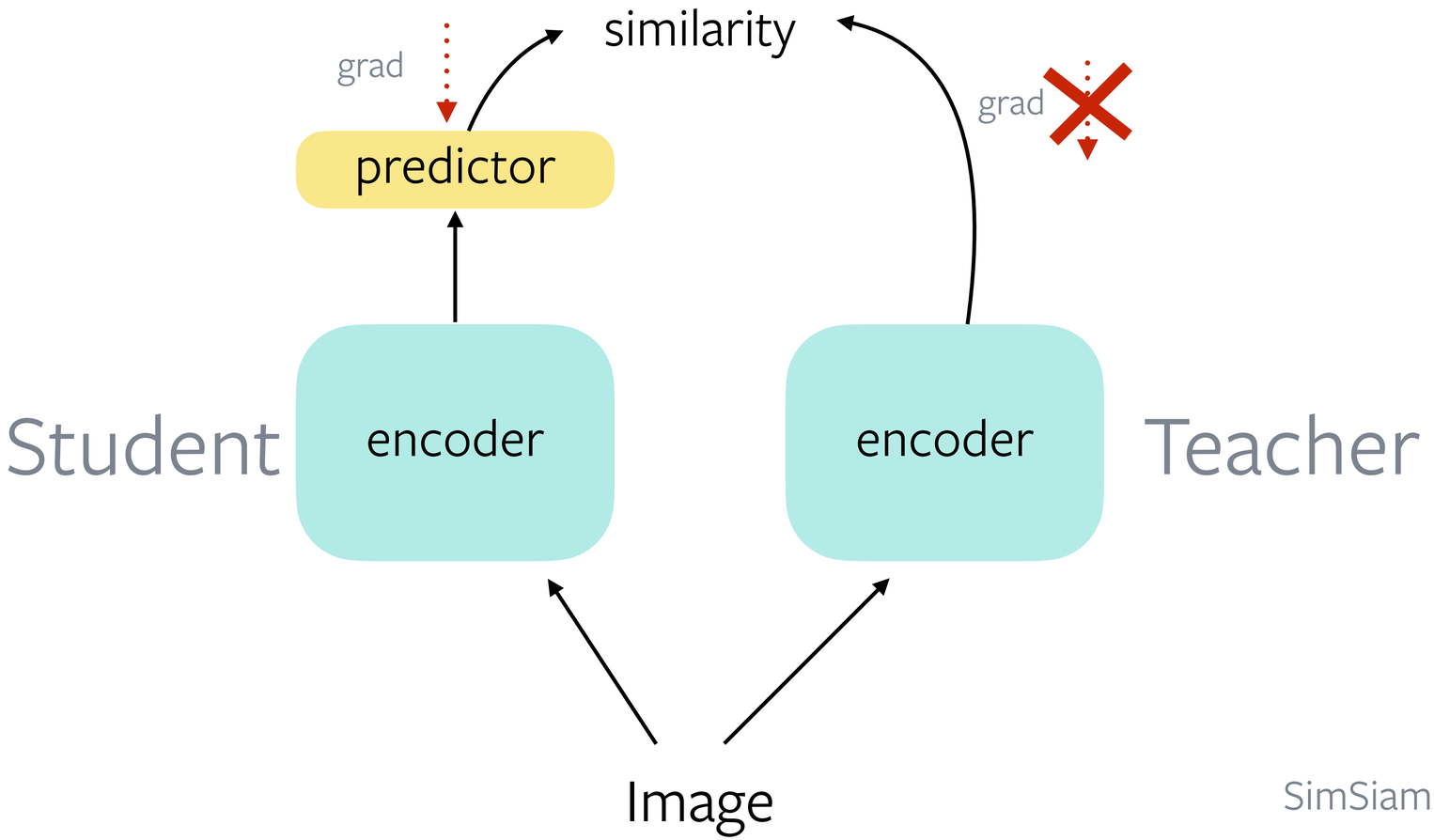
- What we want $f_\theta(I) = f_\theta(\text{augment}(I))$
- How we do it $f_\theta^{\text{student}}(I) = f_\theta^{\text{teacher}}(\text{augment}(I))$



BYOL - Grill et al., 2020

SimSiam

- What we want $f_\theta(I) = f_\theta(\text{augment}(I))$



SimSiam - Chen & He 2020

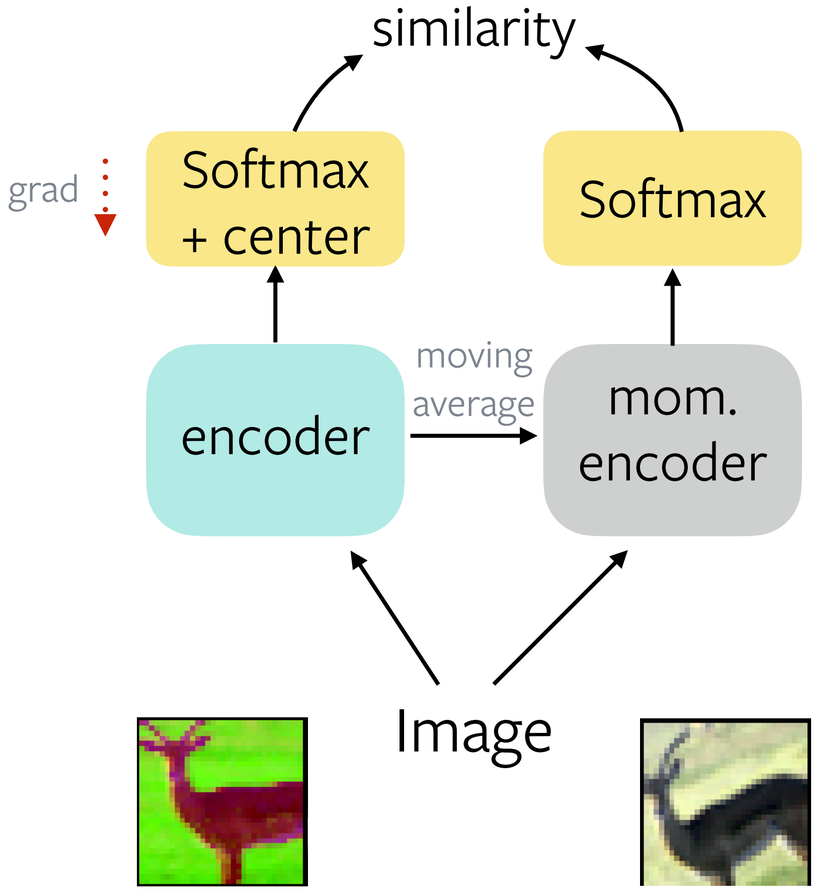
DINO - Distillation with No Labels

Mathilde Caron, Hugo Touvron, Ishan Misra,
Herve Jegou, Julien Marial, Piotr Bojanowski, Armand Joulin

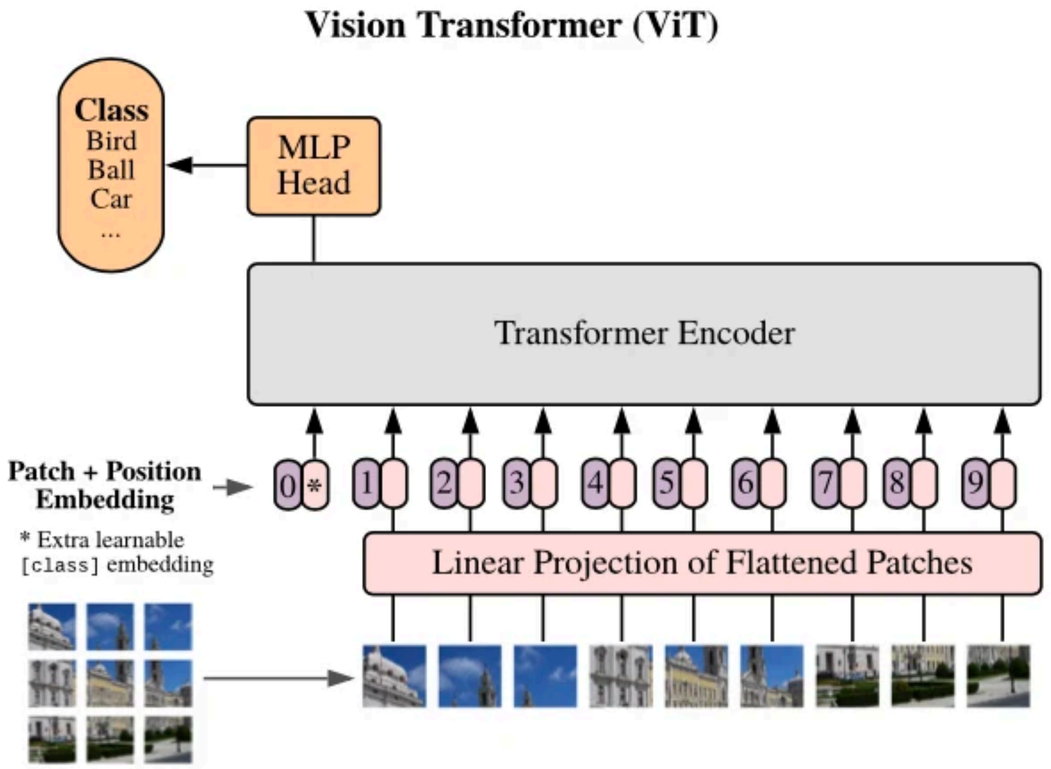
https://github.com/facebookresearch/dino

ICCV 2021

DINO - Main idea



Type of encoder - Vision Transformer



Vision Transformer - Dosovitskiy et al., 2020

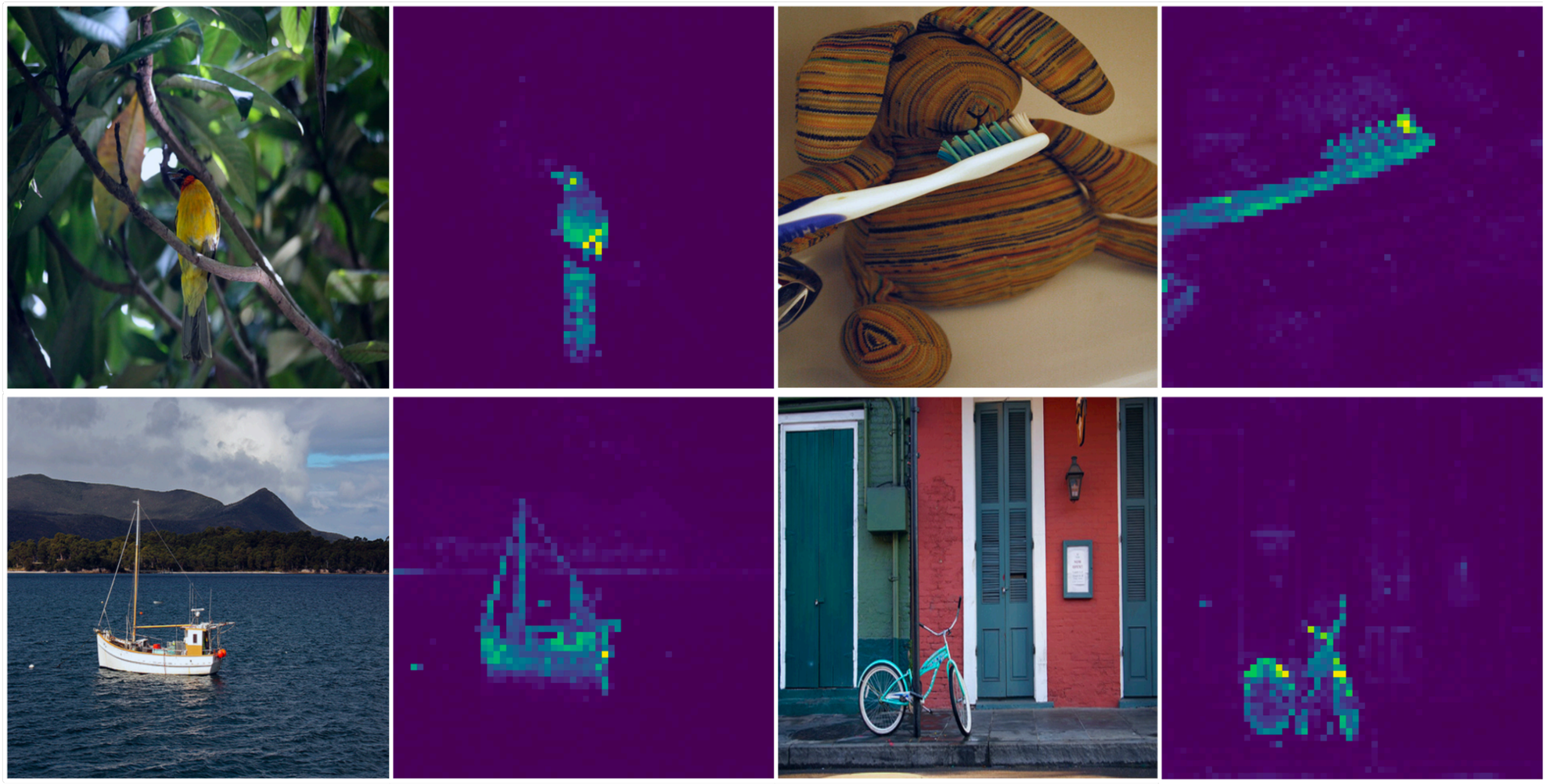
No pooling!

Feature maps in CNN

Feature maps in ViT

Source: https://towardsdatascience.com/vision-transformers-or-convolutional-neural-networks-both-de1a2c3c62e4

Segmentation emerges!

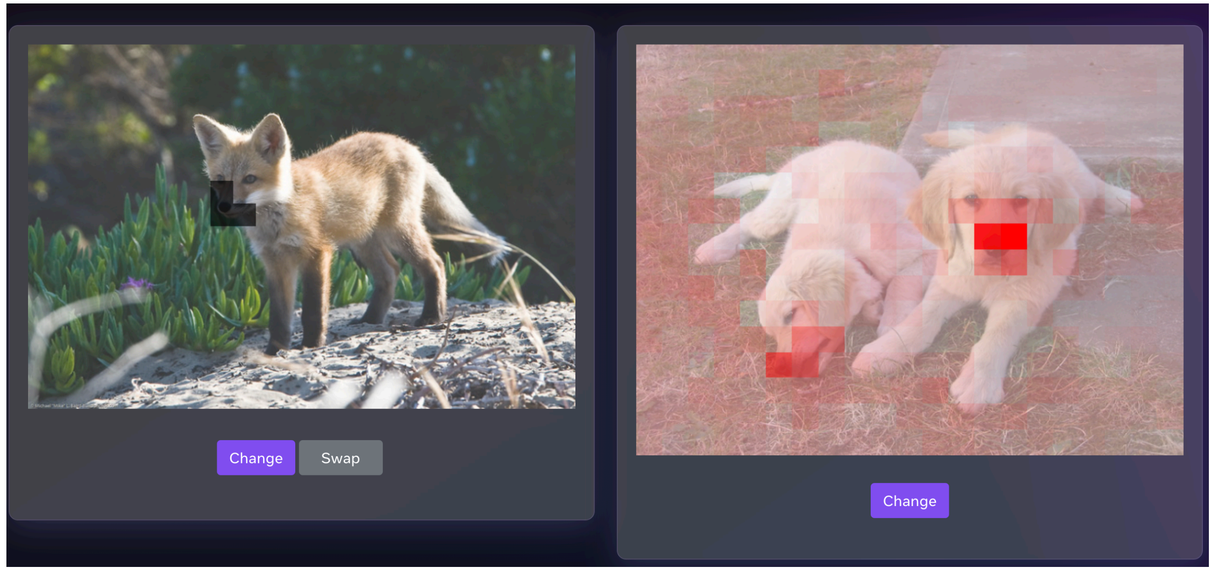


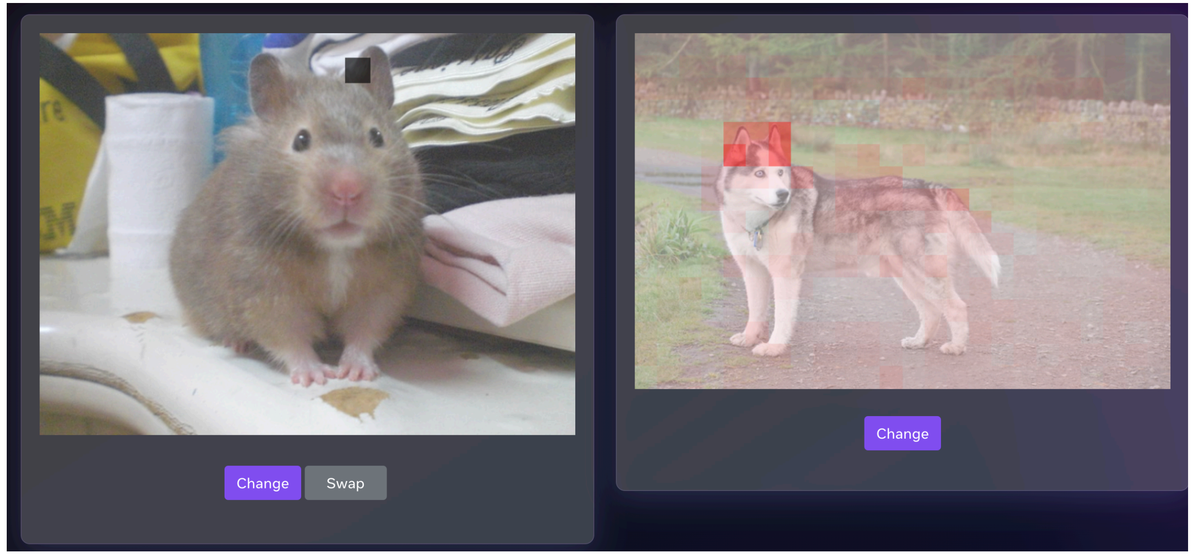
Visualize the “CLS” token attention.

Note that the CLS token or the network are not supervised

Segmentation across different heads

DINO features can match across concepts/ images





https://ai.facebook.com/blog/dino-self-supervised-learning-demo/

Results using Linear & kNN

Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**
Comparison across architectures					
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

Many ways to avoid trivial solutions

Similarity Maximization Objective

- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam

Redundancy Reduction Objective

- Redundancy Reduction
  - Barlow Twins, VICReg

Barlow Twins: Self-supervised Learning via Redundancy Reduction

Jure Zbontar*, Li Jing*, Ishan Misra, Yann LeCun, Stéphane Deny

https://github.com/facebookresearch/barlowtwins

Horace Barlow's Efficient Coding Hypothesis

- Inspired by Information Theory
- Neurons communicate via ″spiking codes″
- Spiking codes aim to reduce redundancy between neurons

Redundancy Reduction

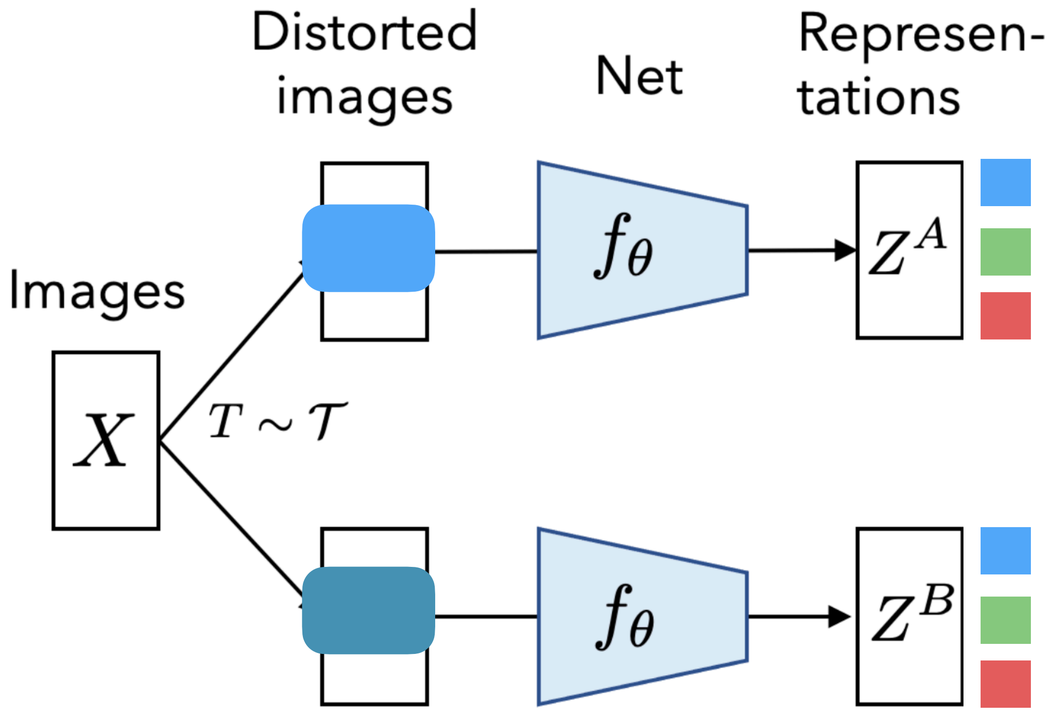
- N neurons produce a representation: N dimensional feature
- Each neuron should satisfy
  - Invariance -- be invariant under different data augmentation
  - Independent of other neurons -- reduce redundancy

- VERY roughly speaking

$$f_\theta(I)[i] = f_\theta(\text{augment}(I))[i]$$

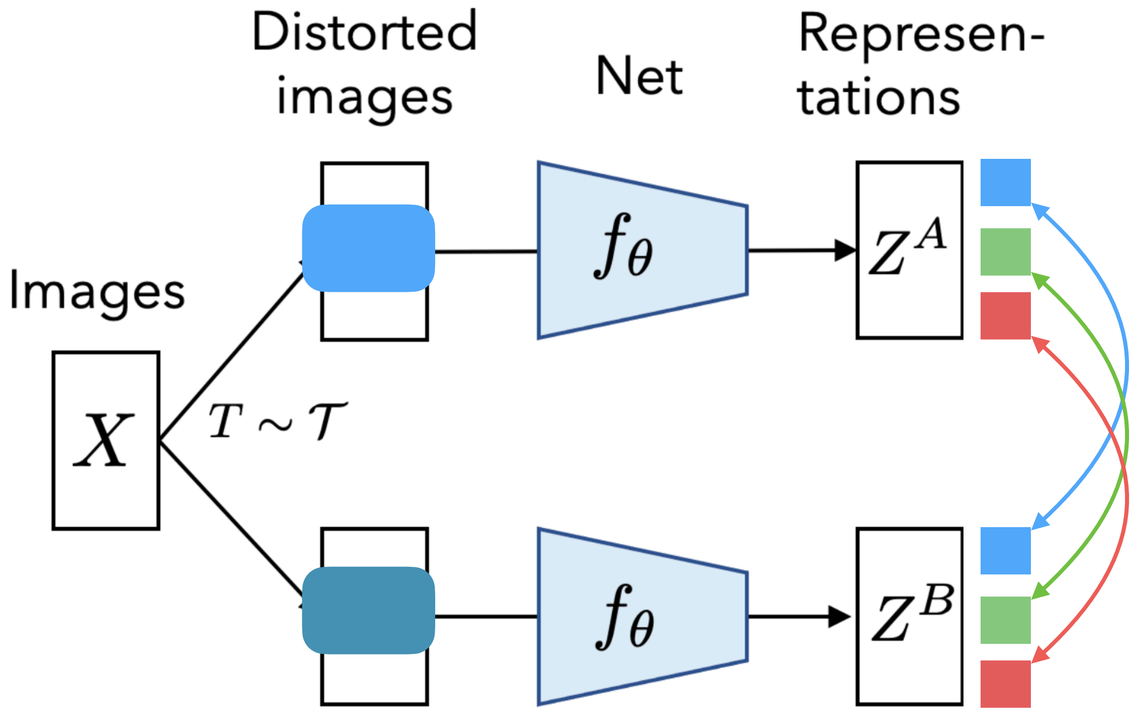
$$f_\theta(I)[i] \neq f_\theta(\text{augment}(I))[j]$$

Barlow Twins



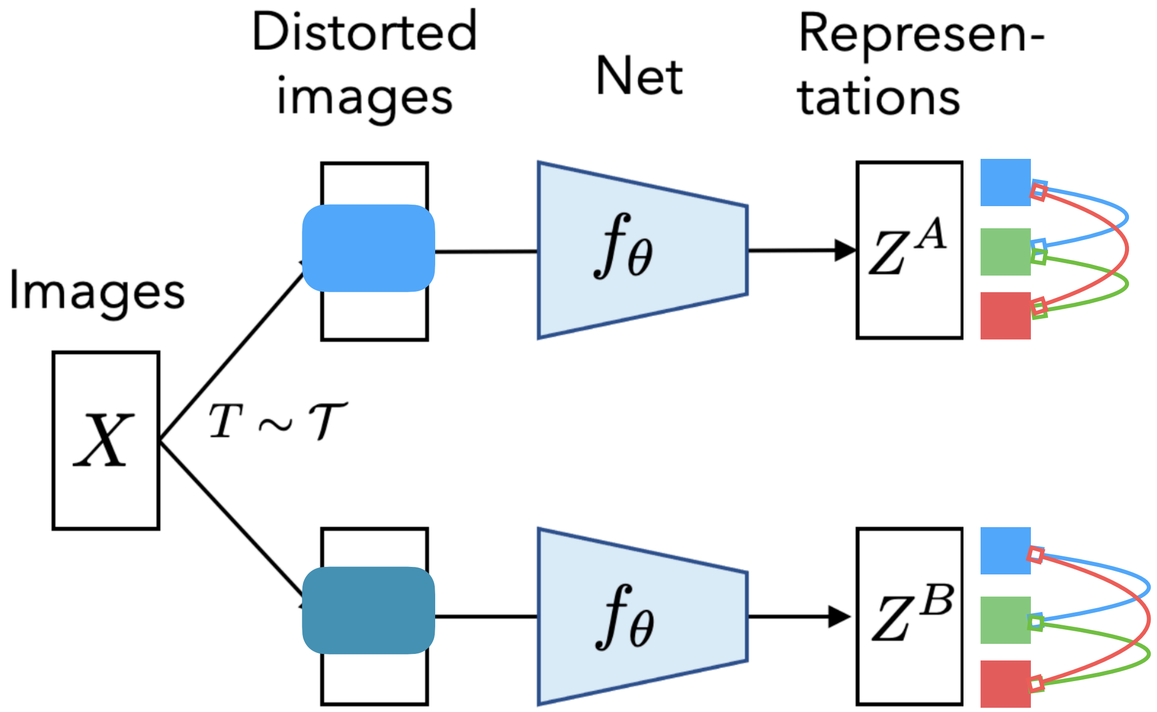
Barlow Twins - Zbontar et al., 2020

Barlow Twins - Invariance



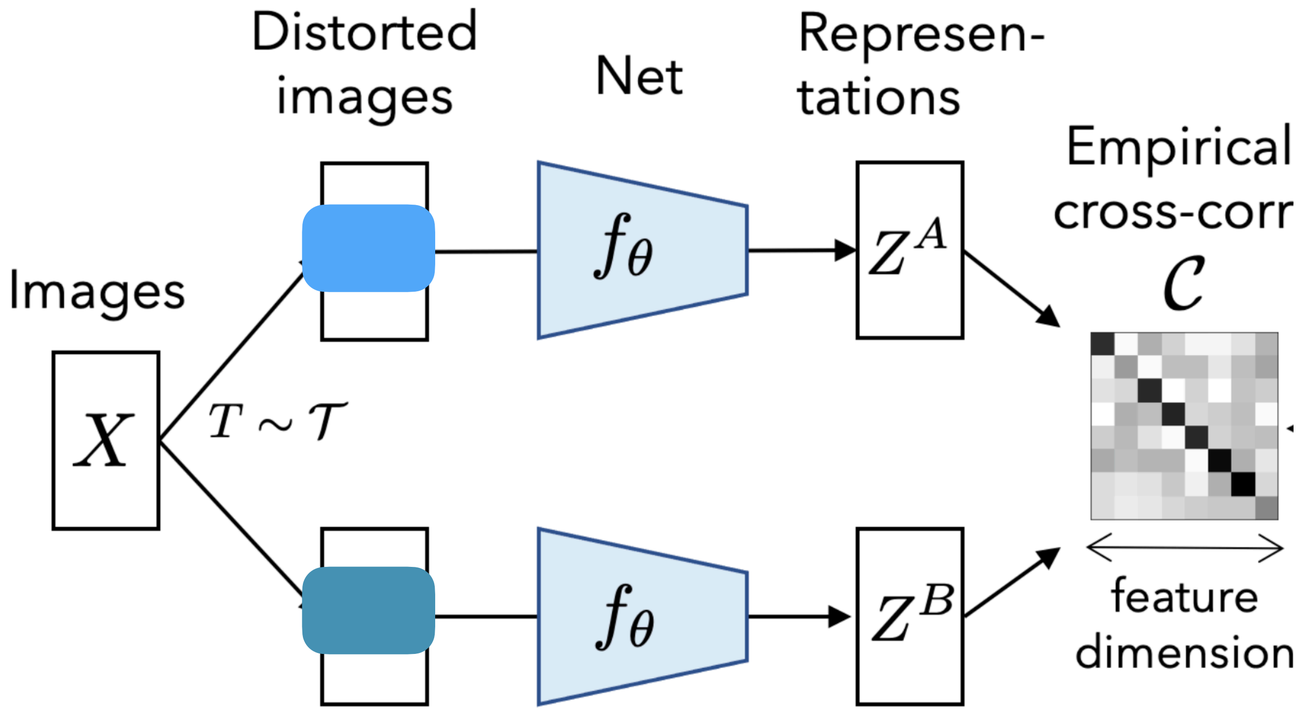
Barlow Twins - Zbontar et al., 2020

Barlow Twins - Redundancy Reduction



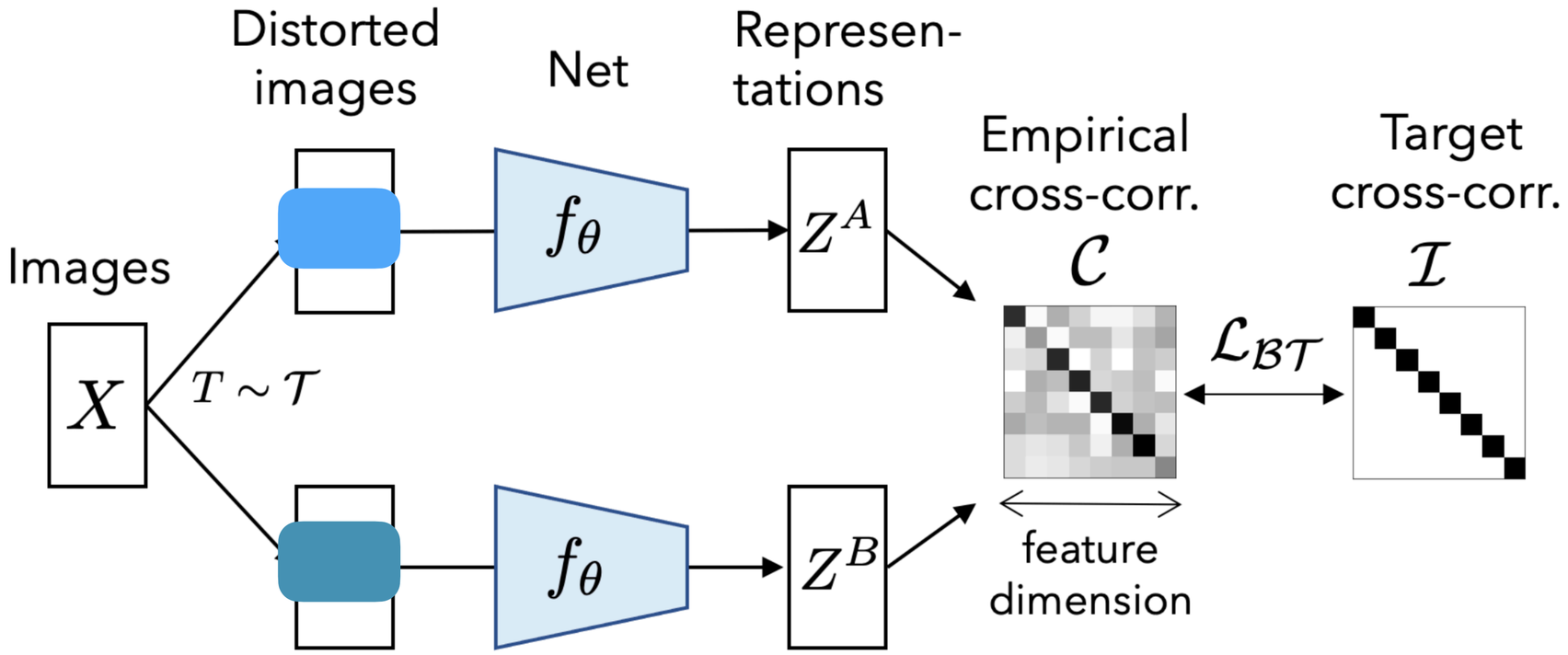
Barlow Twins - Zbontar et al., 2020

Barlow Twins



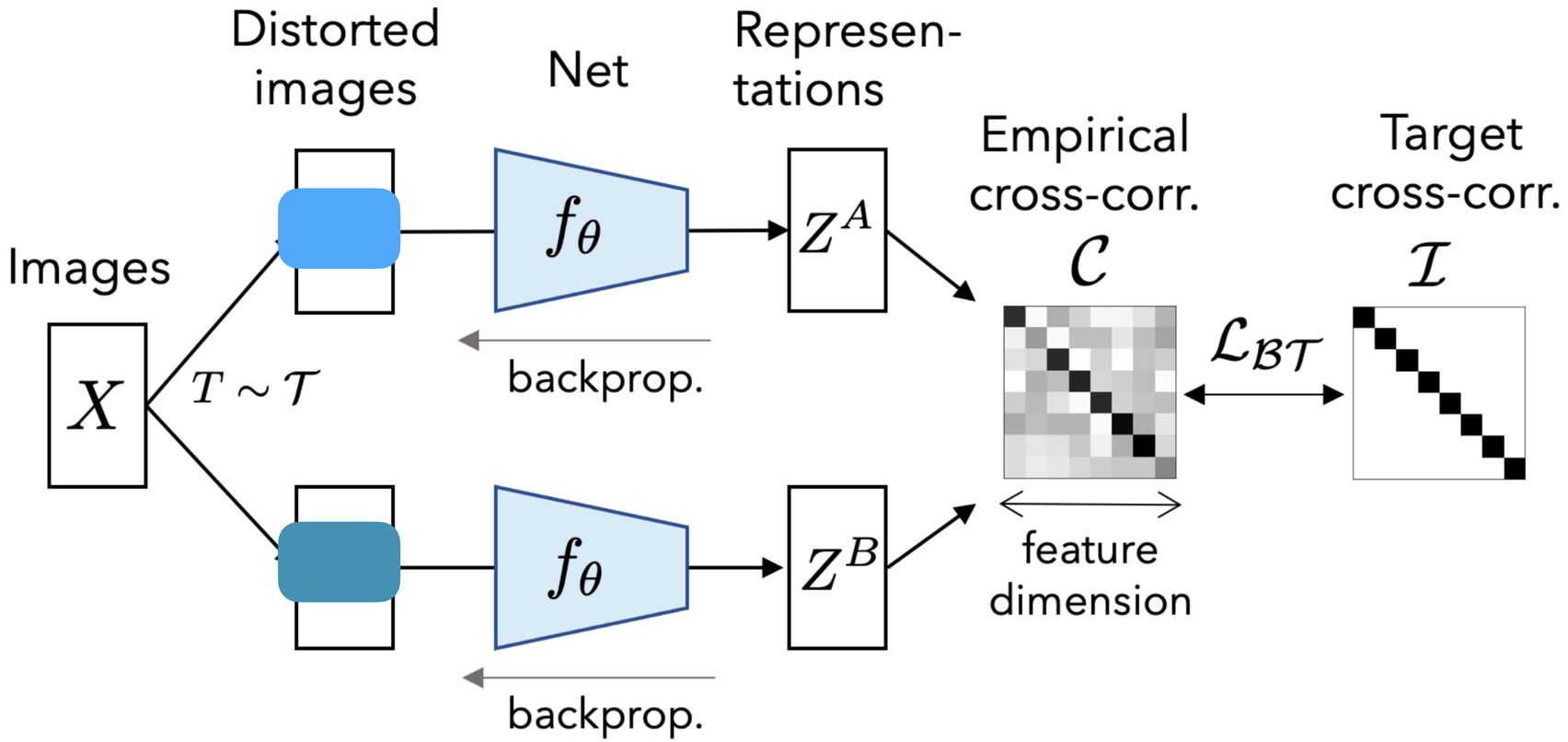
Barlow Twins - Zbontar et al., 2020

Barlow Twins - Loss



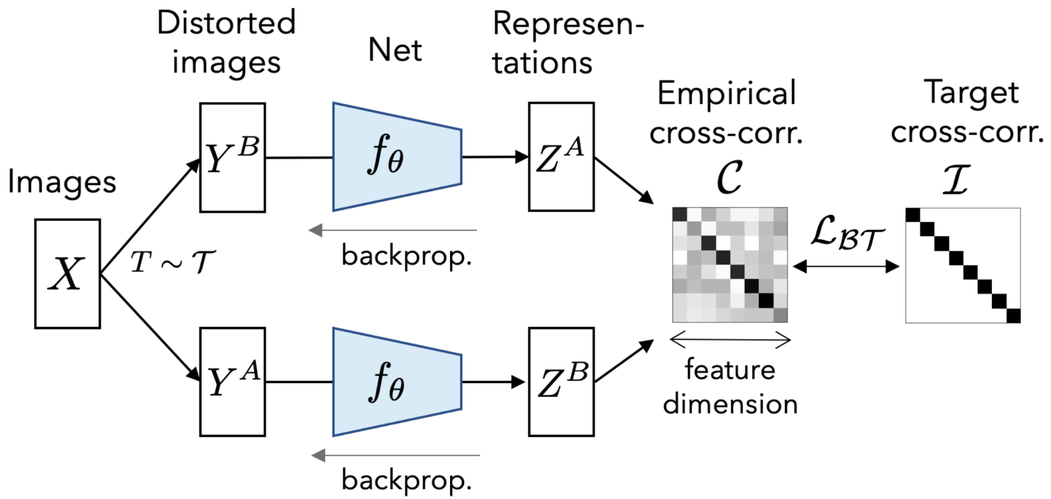
Barlow Twins - Zbontar et al., 2020

Barlow Twins - Loss



Barlow Twins - Zbontar et al., 2020

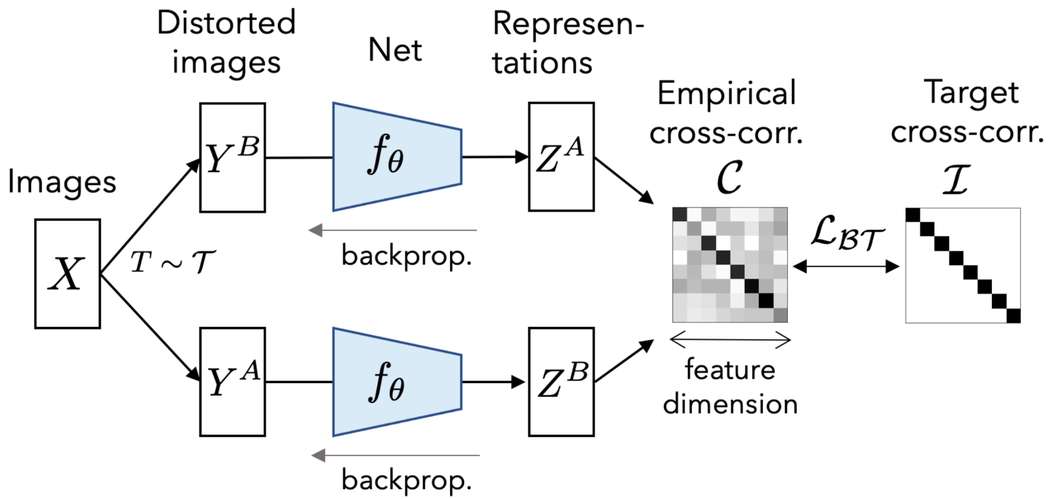
Barlow Twins Objective Function



$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \sqrt{\sum_b \left(z^B_{b,j}\right)^2}}$$

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}{}^2}_{\text{redundancy reduction term}}$$

Trivial Solutions?



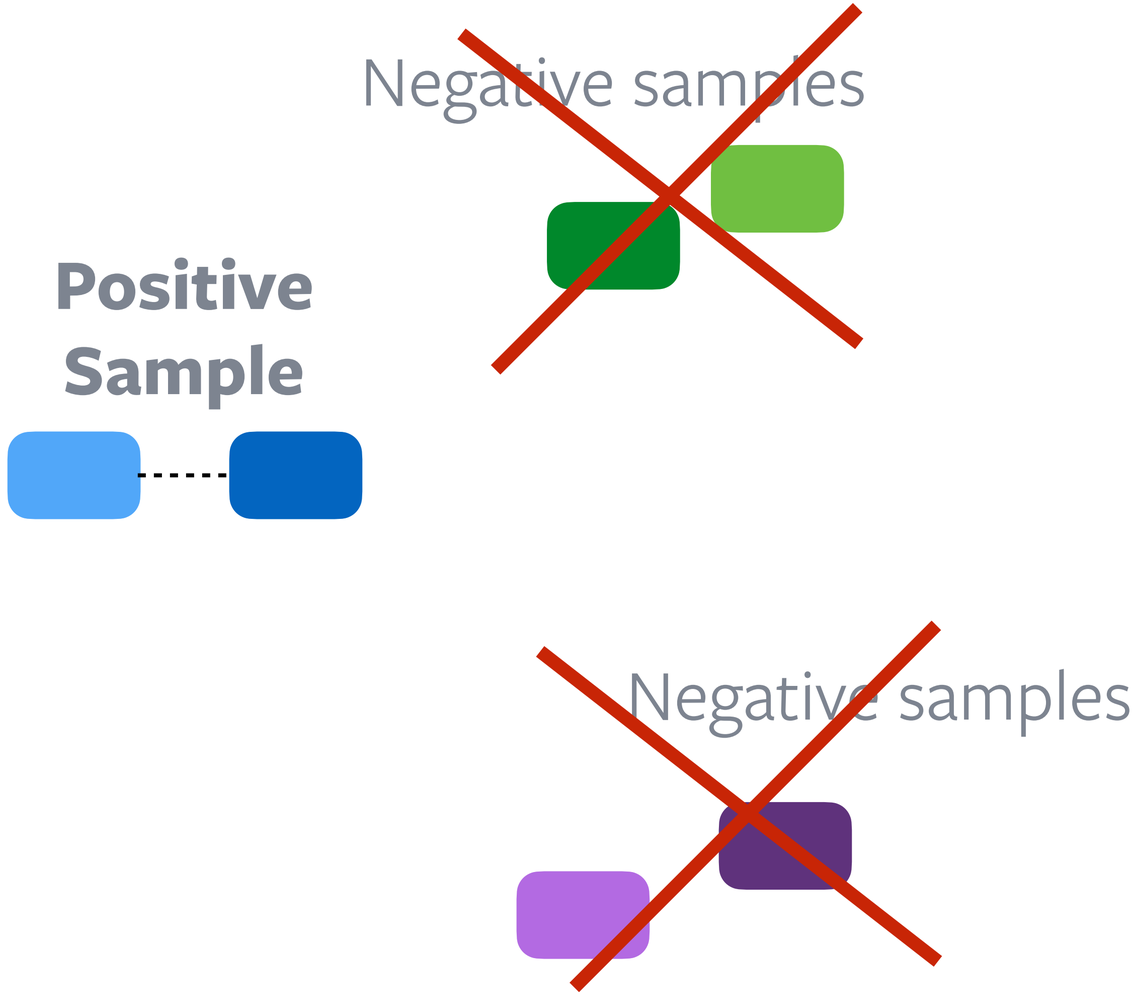
$$\mathcal{C}_{ij} \triangleq \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b \left(z^A_{b,i}\right)^2} \sqrt{\sum_b \left(z^B_{b,j}\right)^2}}$$

$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}{}^2}_{\text{redundancy reduction term}}$$

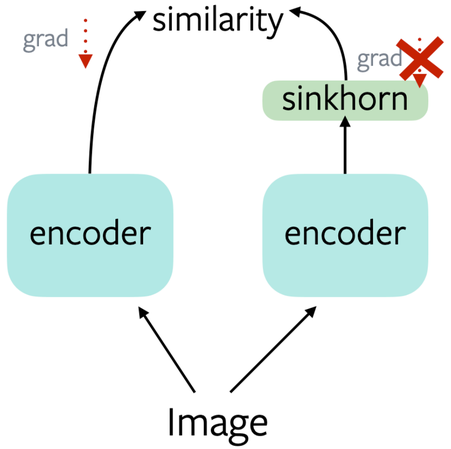
Center Z^A and Z^B before computing cross-correlation

Prevents trivial solutions without

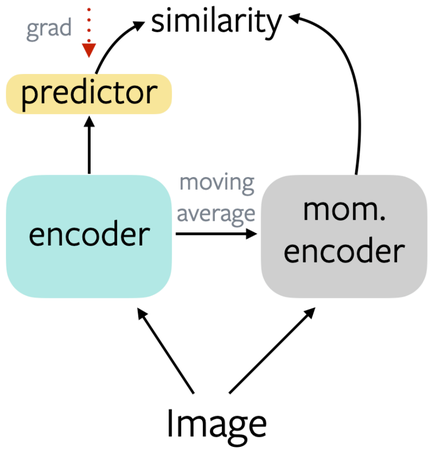
- ″Negatives″ like in contrastive learning



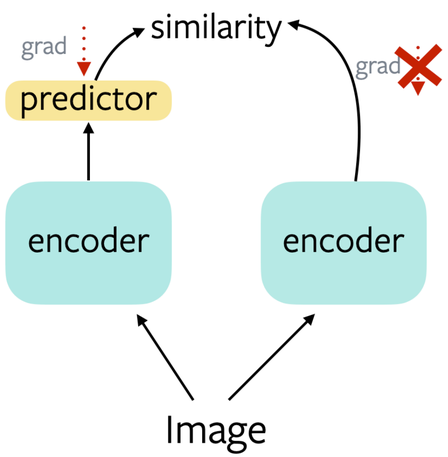
Prevents trivial solutions without Asymmetric Learning



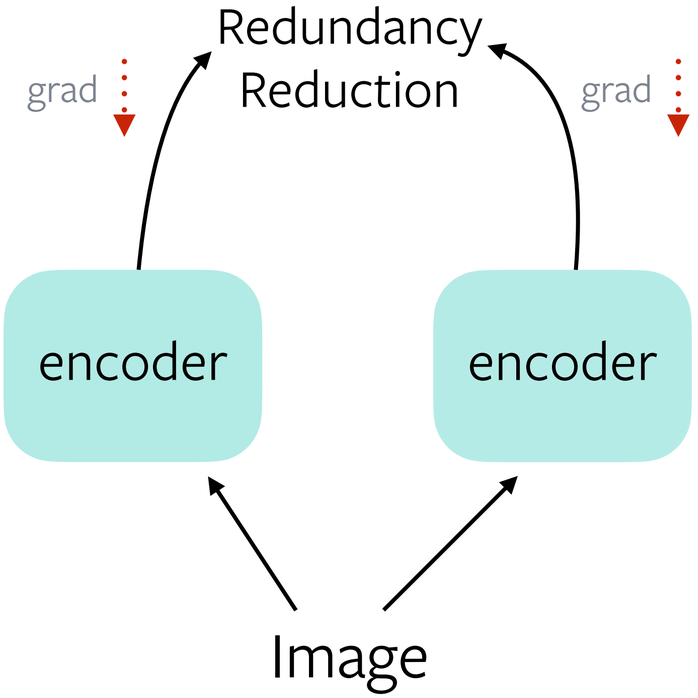
SwAV - Caron et al., 2020



BYOL - Grill et al., 2020



SimSiam - Chen & He, 2020



Barlow Twins

The great spiral of research

Pre 2015 - Sparse encoding, RBMs, contrastive

2015 - Pretext

2018/19 - Invariance using Contrastive

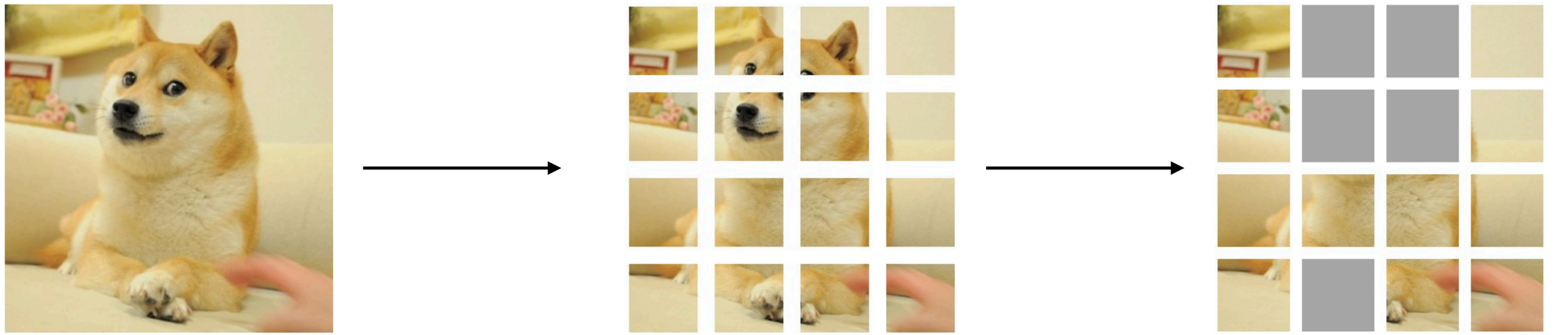
2020 - Invariance using non-contrastive

2021 - Pretext tasks are cool again

BeIT: BERT Pre-Training of Image Transformers

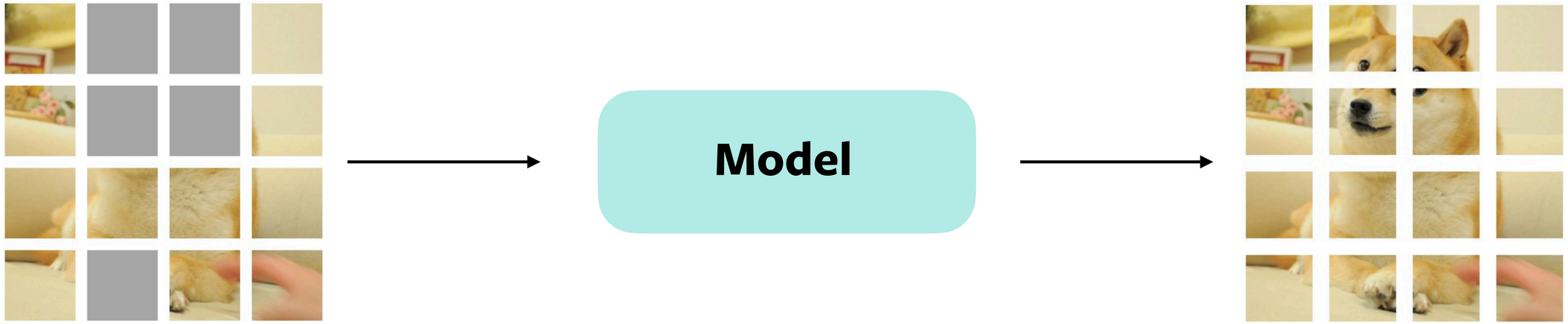
Hangbo Bao, Li Dong, Furu Wei

BeIT

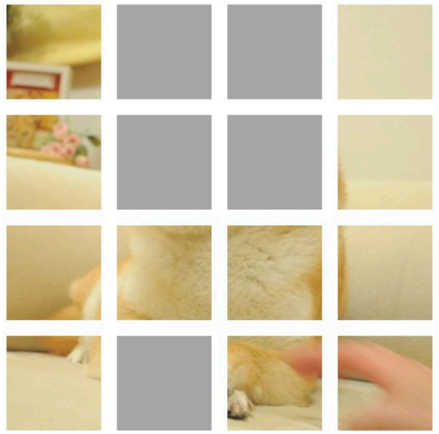


Original Image **Patches** **Masked Patches**

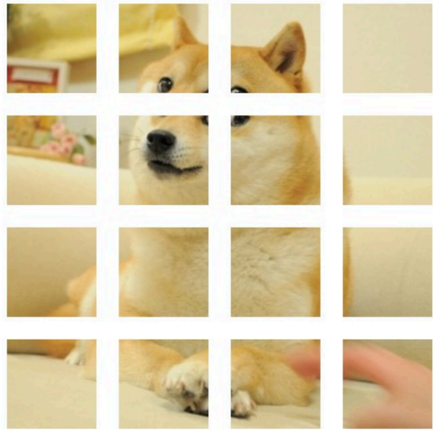
BeIT: Masked Prediction Problem



Masked Prediction: Vision & NLP



Model



Masked Patches

Patches

A _ day

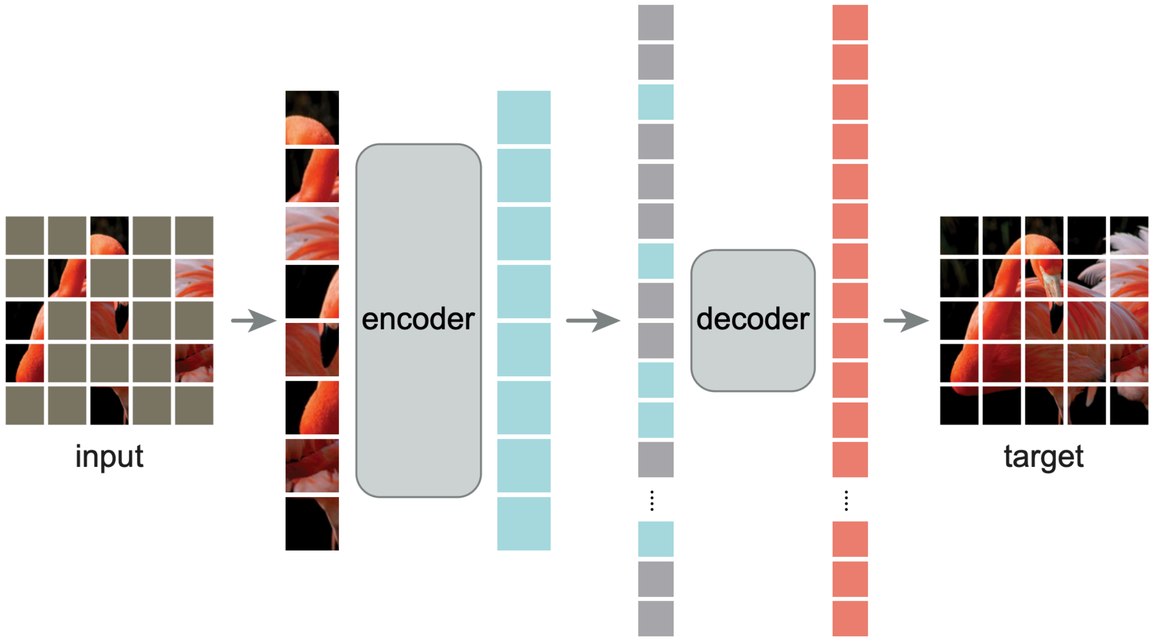
Model

A sunny day

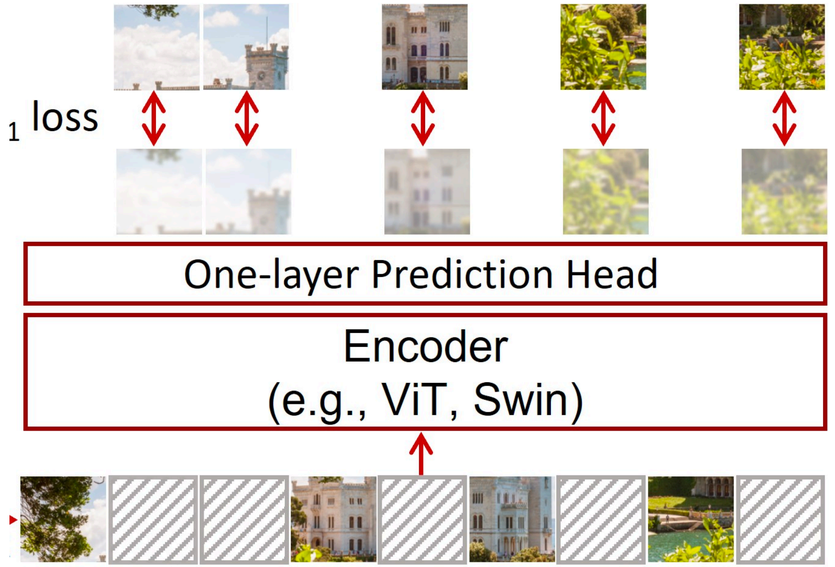
Masked Sentence

Sentence

MAE; SimMIM

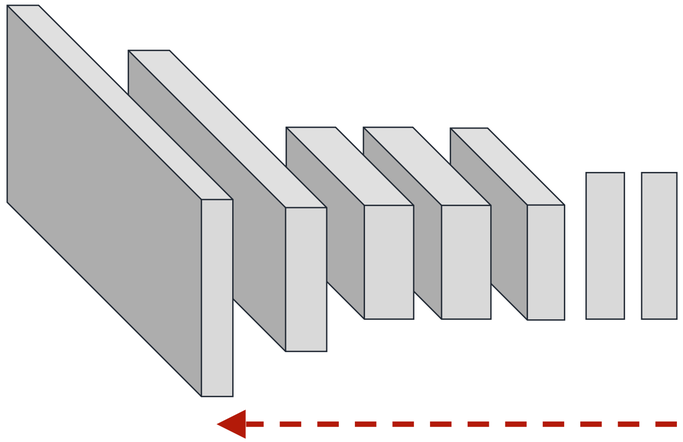


He et al.,
Masked Autoencoders Are Scalable Vision Learners
arXiv, 2021.

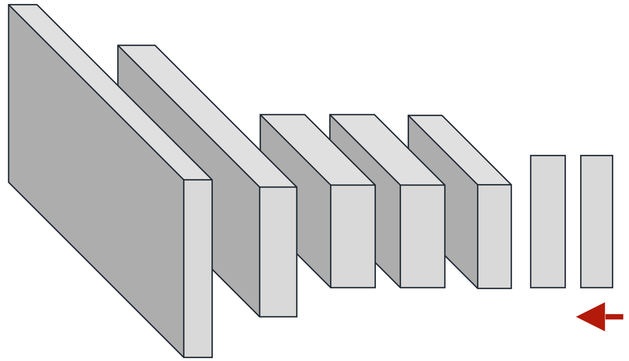


Xie et al.,
A Simple Framework for Masked Image Modeling
arXiv, 2021.

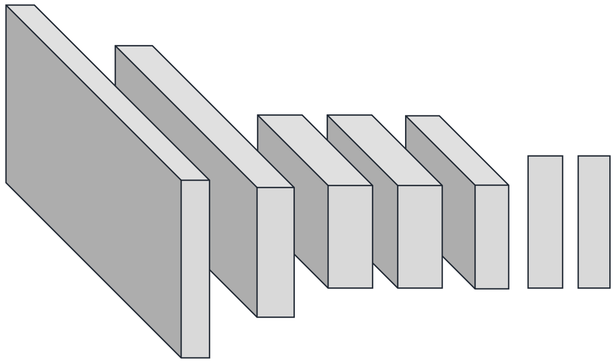
How to evaluate?



Fine-tune all layers



Linear classifier



kNN

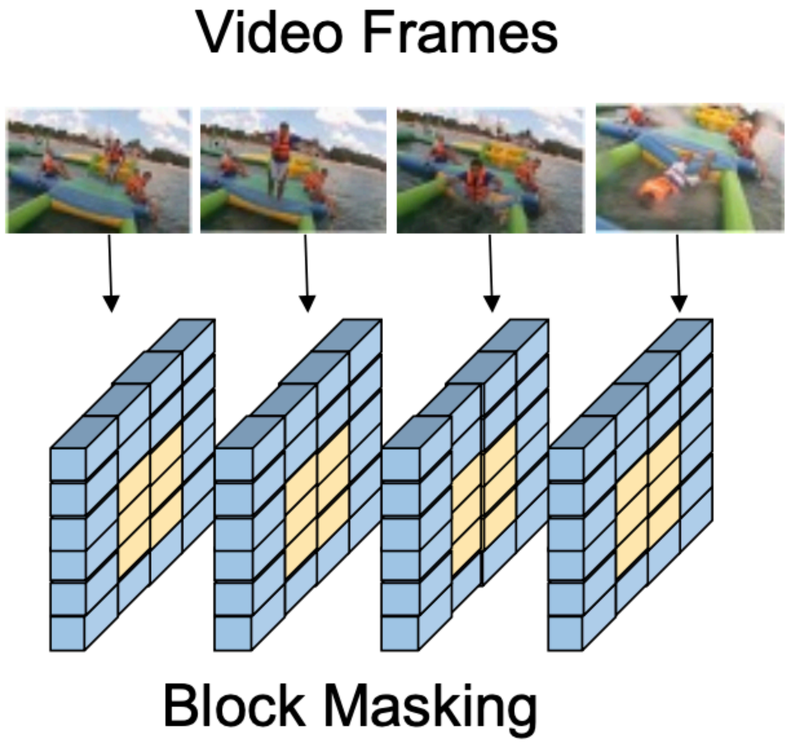
Full finetuning

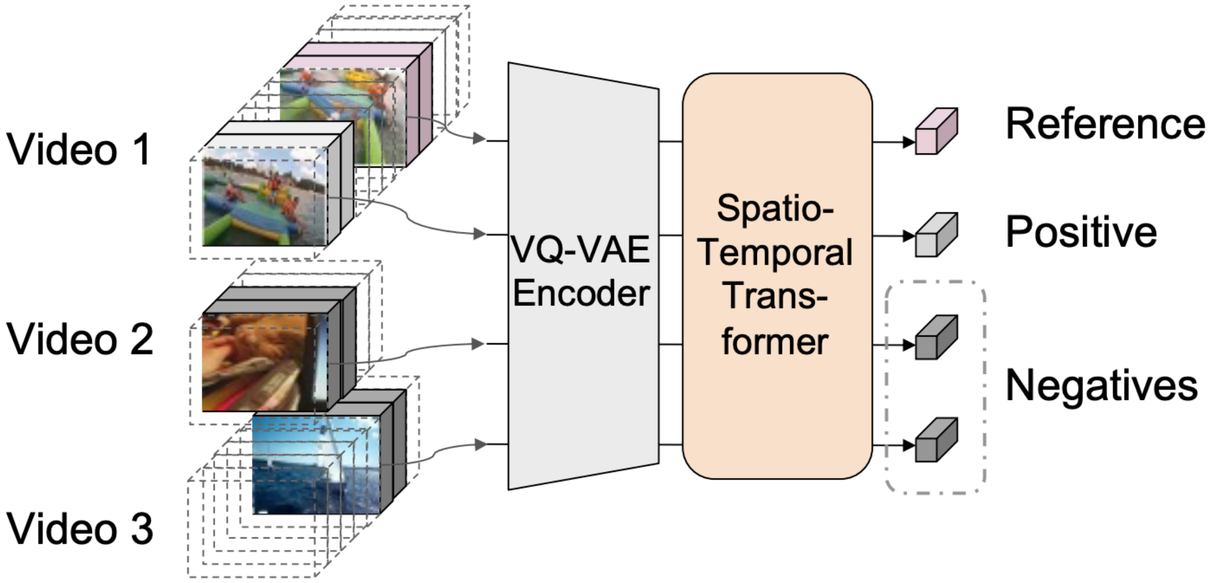
method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**
SimMIM	IN1K			87.1	

Full finetuning - Segmentation

Models	mIoU
Supervised Pre-Training on ImageNet	45.3
DINO (Caron et al., 2021)	44.1
BEIT (ours)	45.6
BEIT + Intermediate Fine-Tuning (ours)	**47.7**

Videos - VIMPAC





VIMPAC: Video Pre-Training via Masked Token Prediction and Contrastive Learning - Tan et al., 2021

Full finetuning

Method	Modality	Pre-Train Dataset	Temporally-Heavy	
			SSV2 [26]	Diving48 [41]
Previous SotA	V	-	65.4 [2]	81.0 [6]
w/o Temporal	V	-	36.6 [6]	-
Self-supervised Pre-Training				
K400 Self-Sup.	V	Kinetics-400 [9]	55.8 [21]	-
MIL-NCE [44]	V+T	HTM [45]	-	-
MMV [1]	V+A+T	AS [23]+HTM [45]	-	-
MoCo [21]	V	IG-Uncurated [24]	53.2	-
VIMPAC	V	HTM [45]	68.1	85.5

Linear classifier or kNN?

Features don’t seem to work well for linear-classification

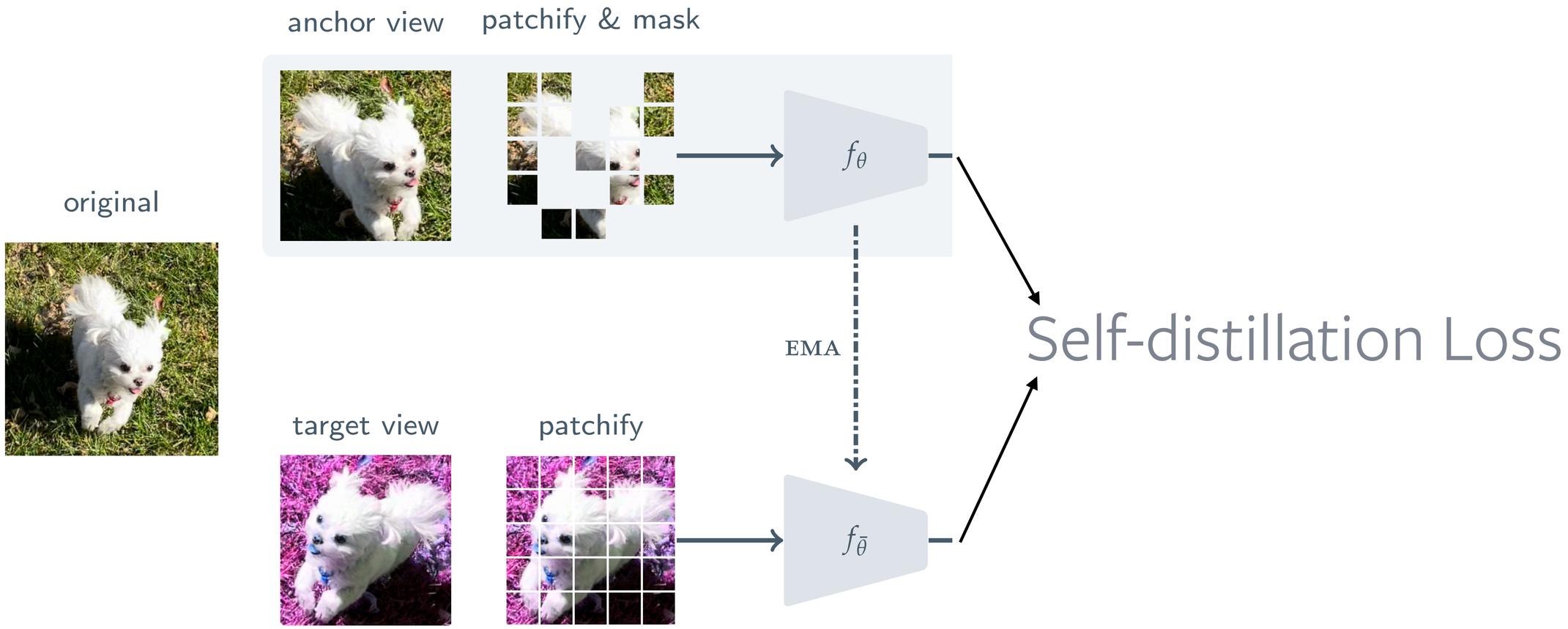
Is this representation learning
OR
learning a good initialization?

Masked Siamese Networks for Label-Efficient Learning

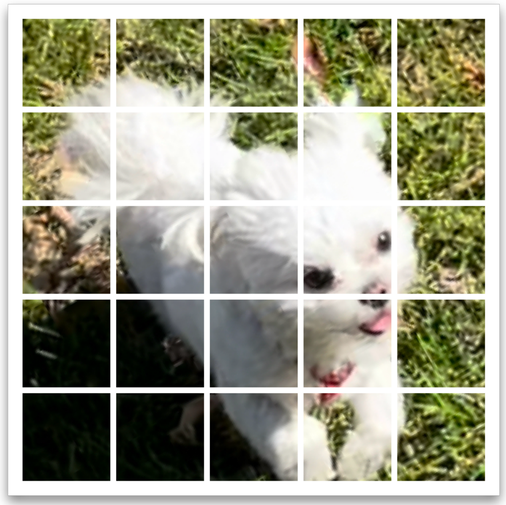
Mido Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski
Florian Bardes, Pascal Vincent, Armand Joulin, Mike Rabbat, Nicolas Ballas

ECCV 2022

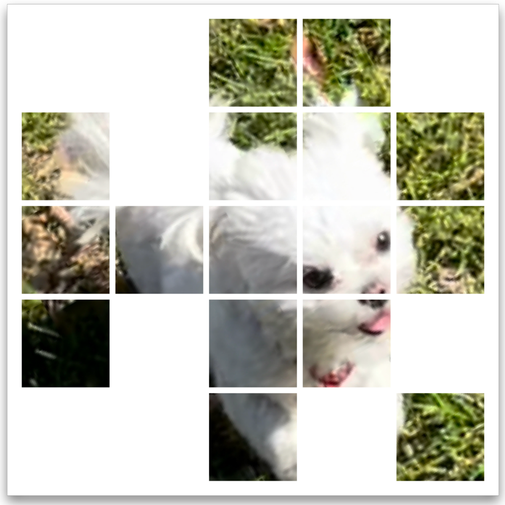
Masked Siamese Networks



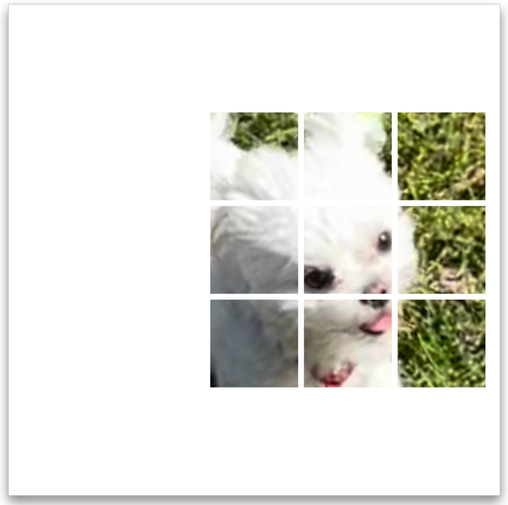
Masked Siamese Networks



Original

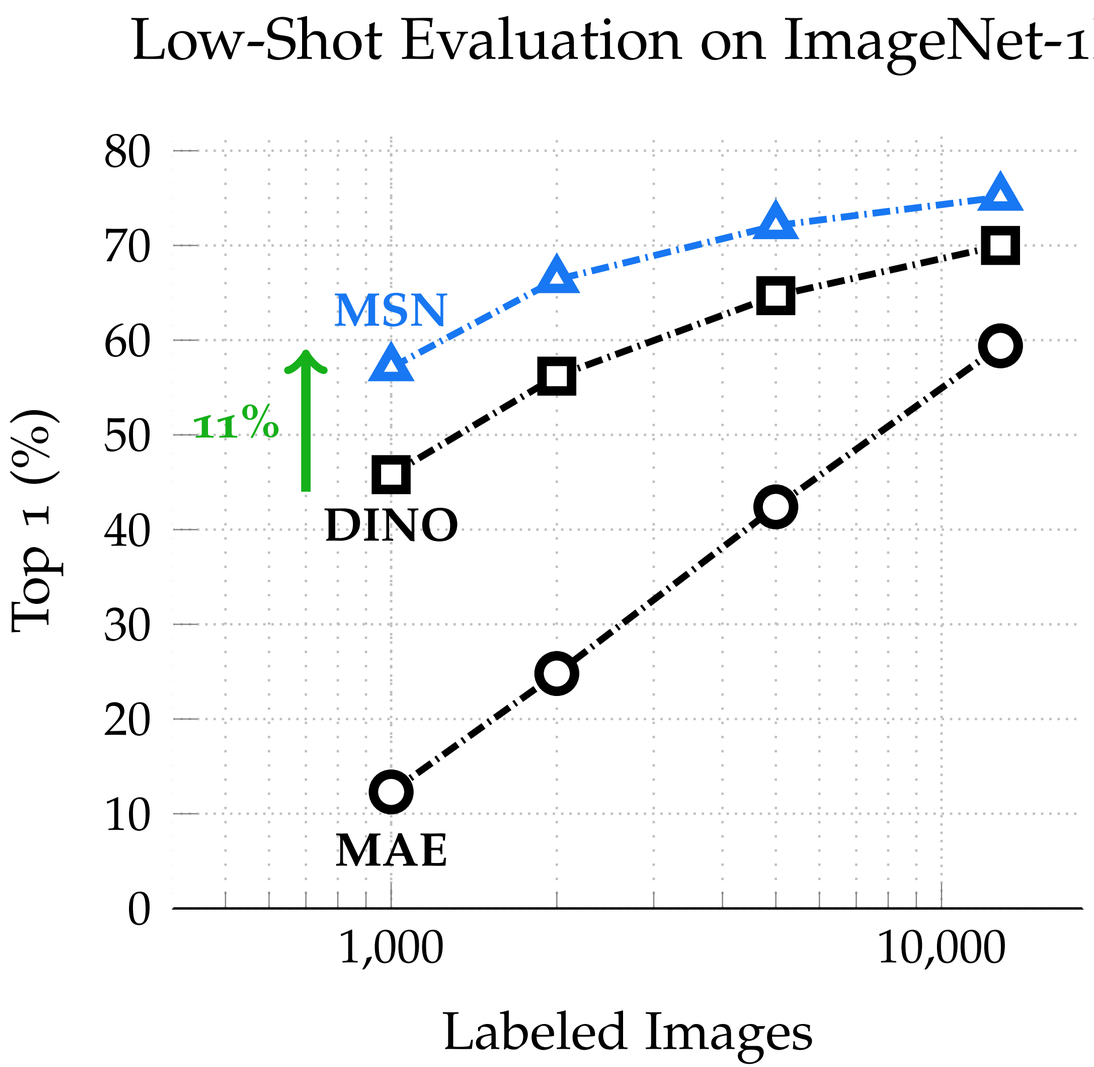


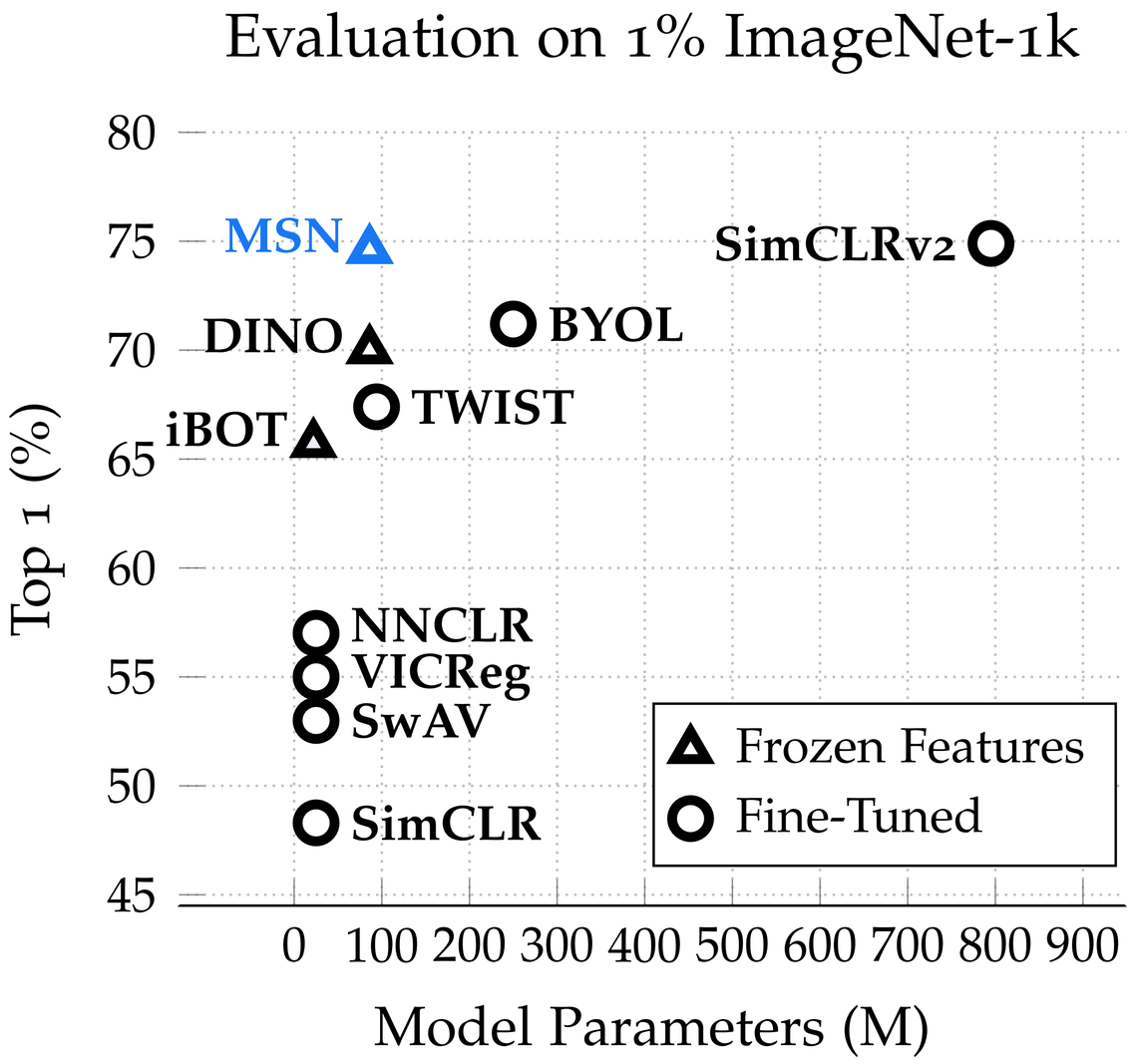
Random Mask



Focal Mask

Label-efficient learning

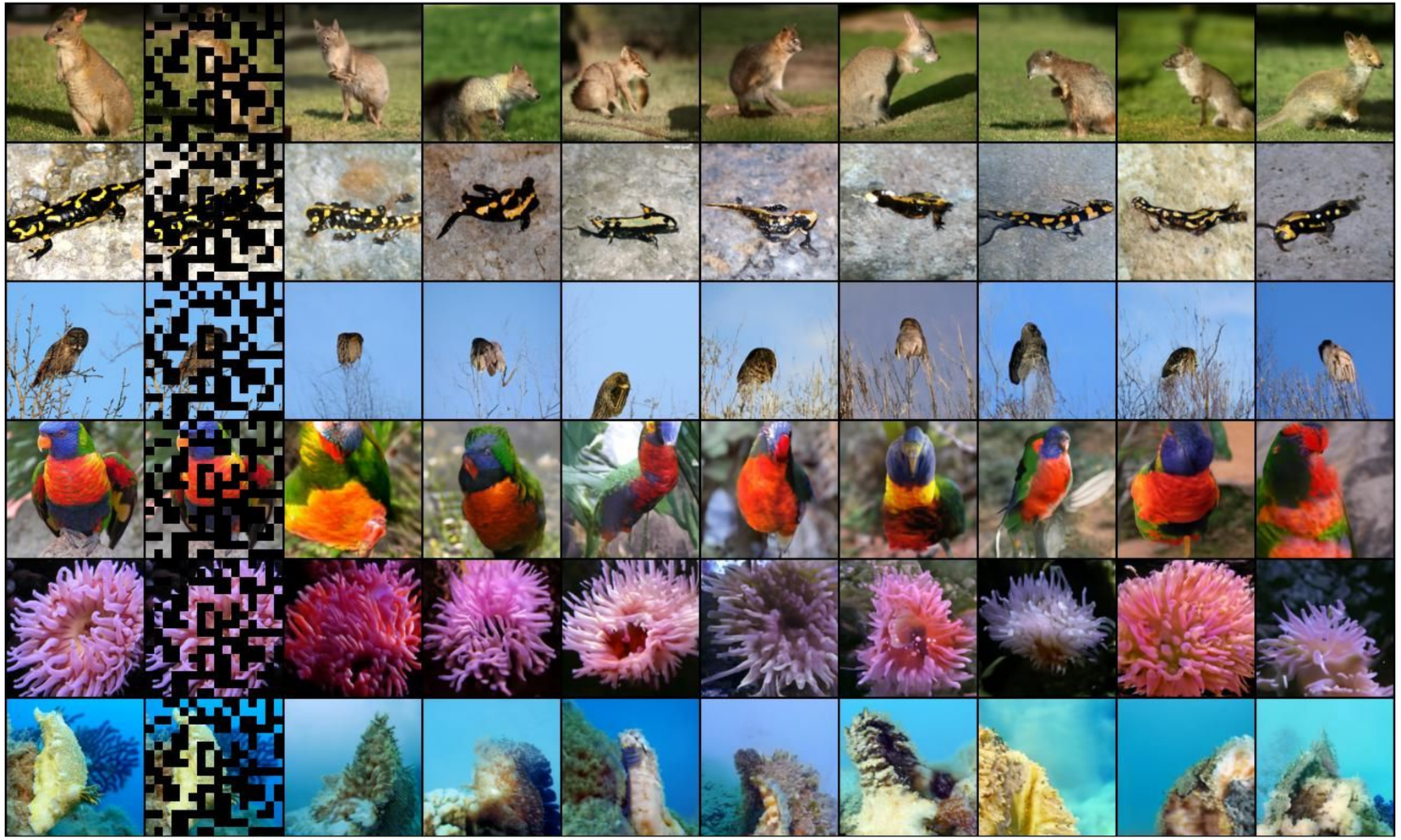




Robust representations

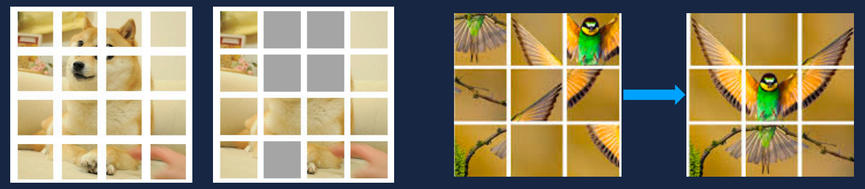
	IN-A (top-1 ↑)	**IN-R** (top-1 ↑)	**IN-Sketch** (top-1 ↑)	**IN-C** (mCE ↓)
Supervised ResNet50	0.04	36.11	24.2	76.7
MAE ViT-B/16 [22]	35.9	48.3	34.5	51.7
MSN ViT-B/16	**37.5**	**50.0**	**36.3**	**46.6**

Reconstructing images

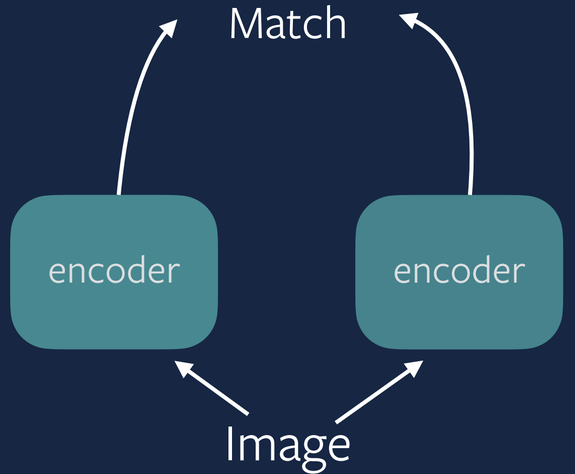


Bordes et al.,
High Fidelity Visualization of What Your Self-Supervised Representation Knows About
arXiv, 2022.

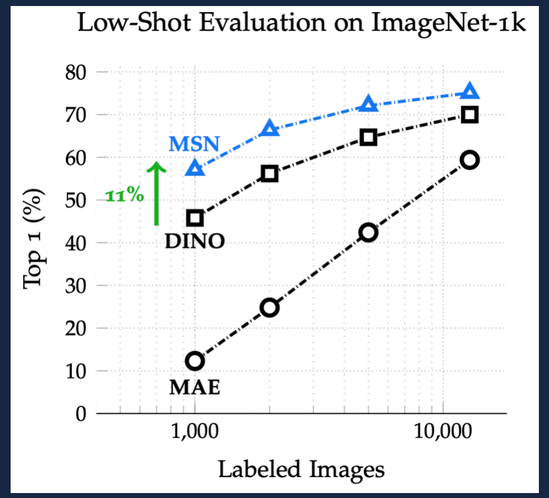
Thanks!



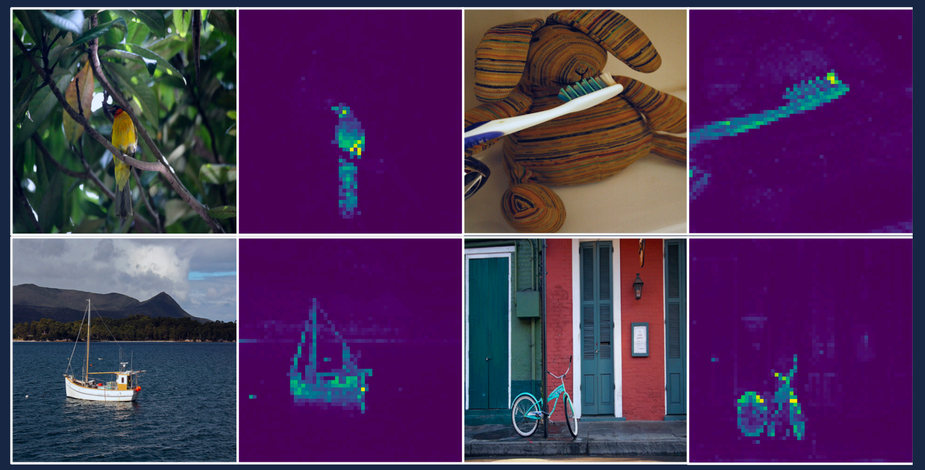
Pretext tasks



Invariance to data augmentation



Label-efficient Learning



Emergent Properties