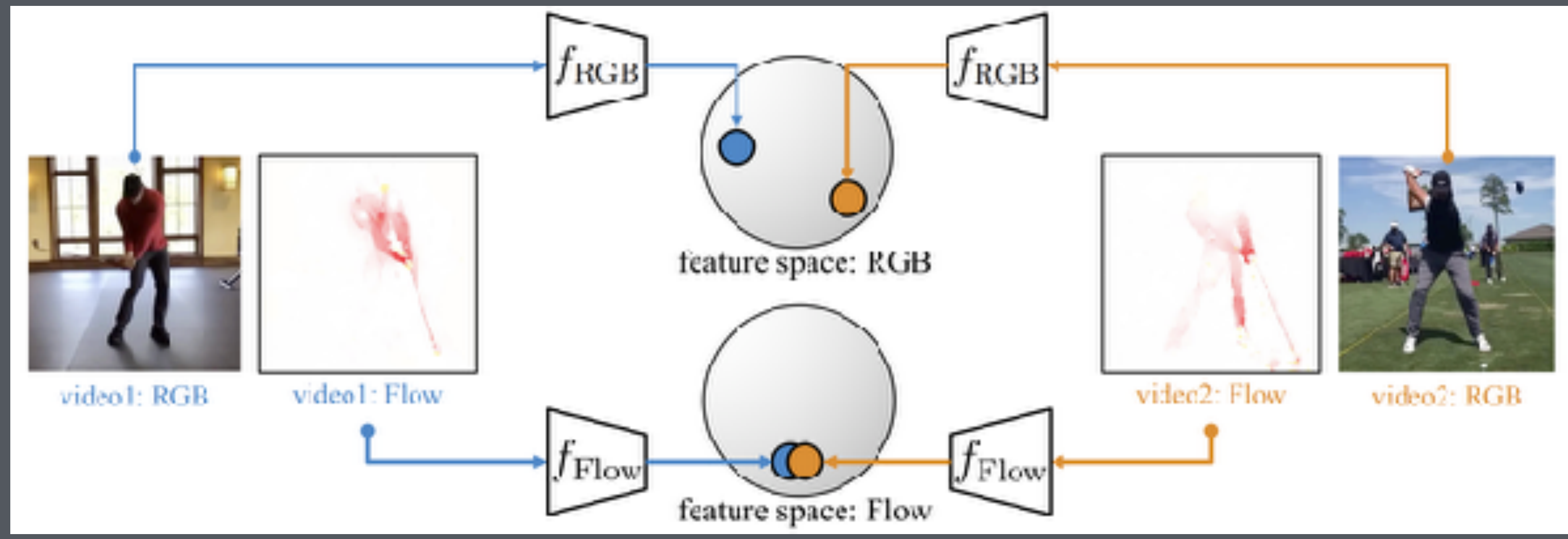
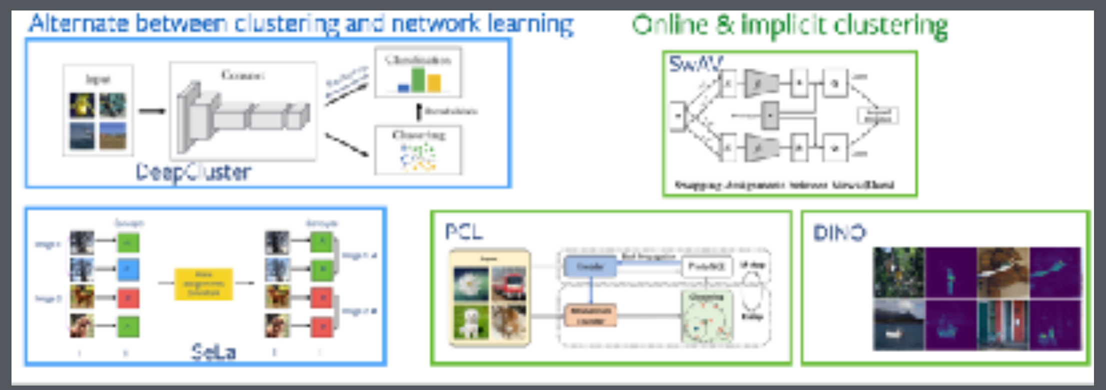
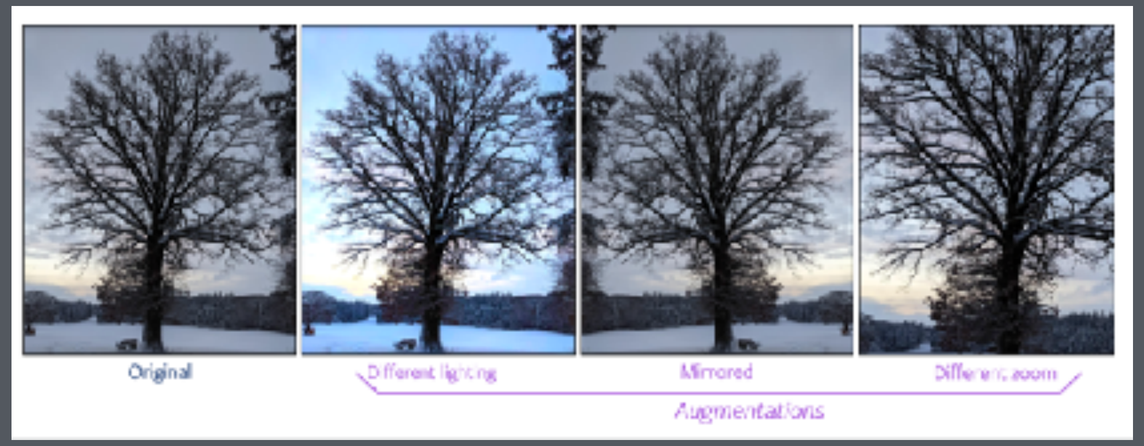
Self-supervised learning for computer vision from images, video and audio. Part 1





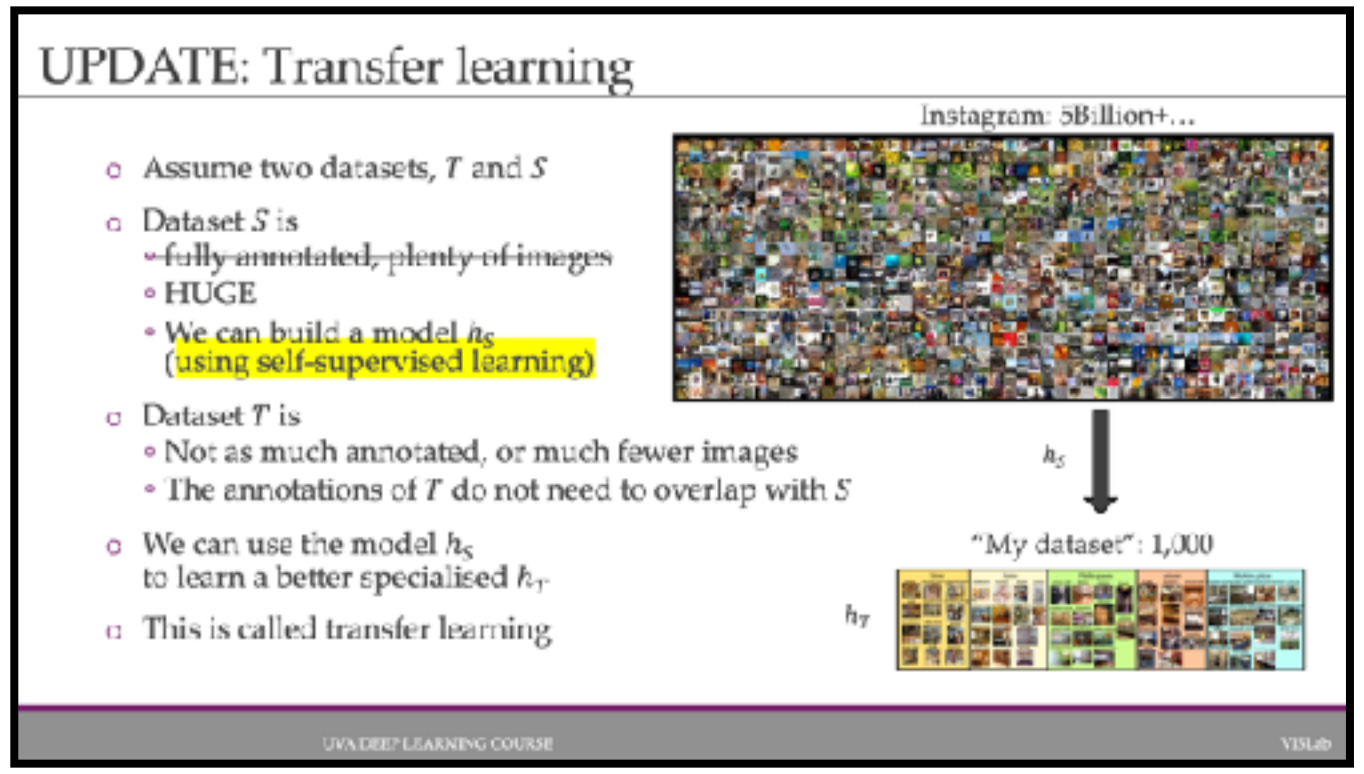


@ DEEP LEARNING 1

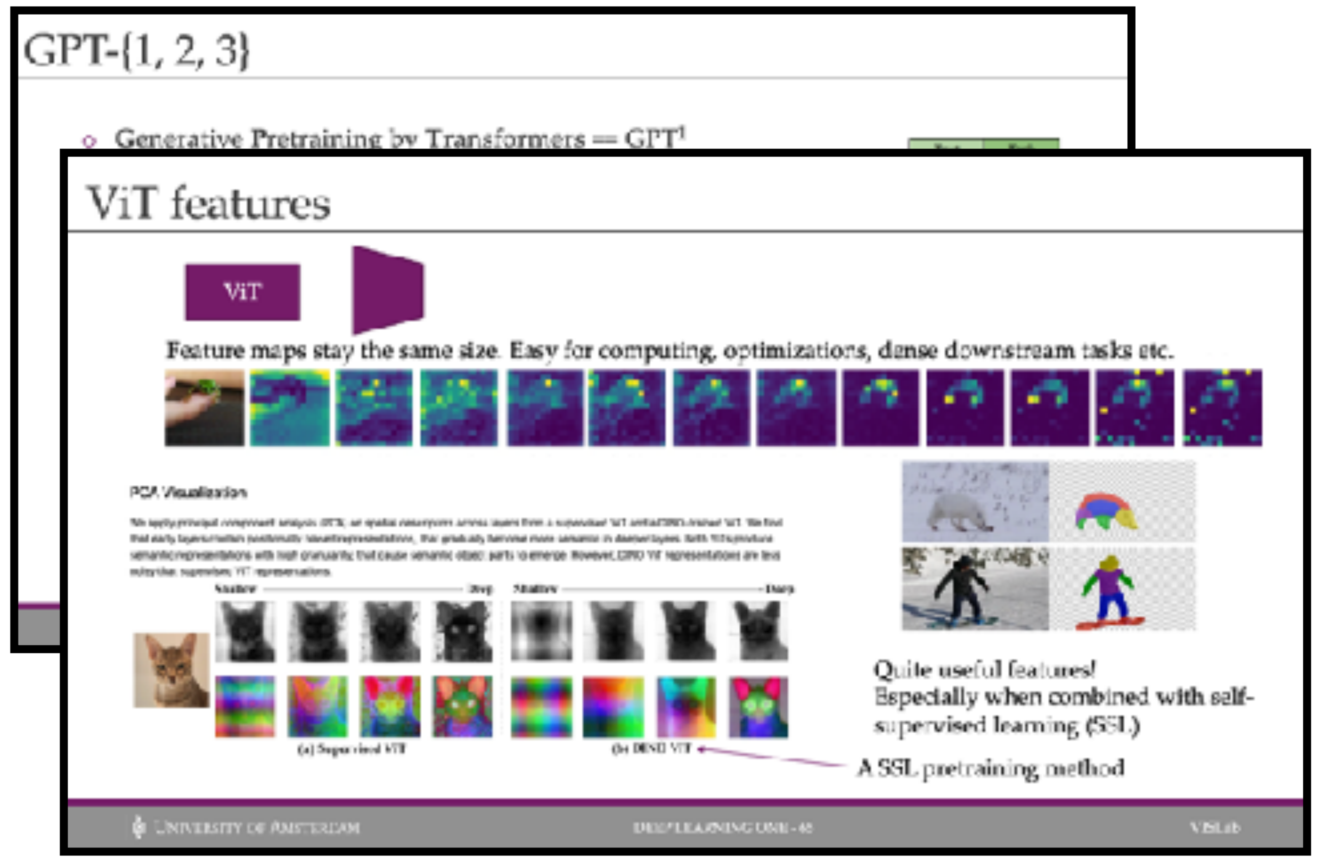
YUKI M. ASANO

LECTURE 13, 13TH DEC 2022

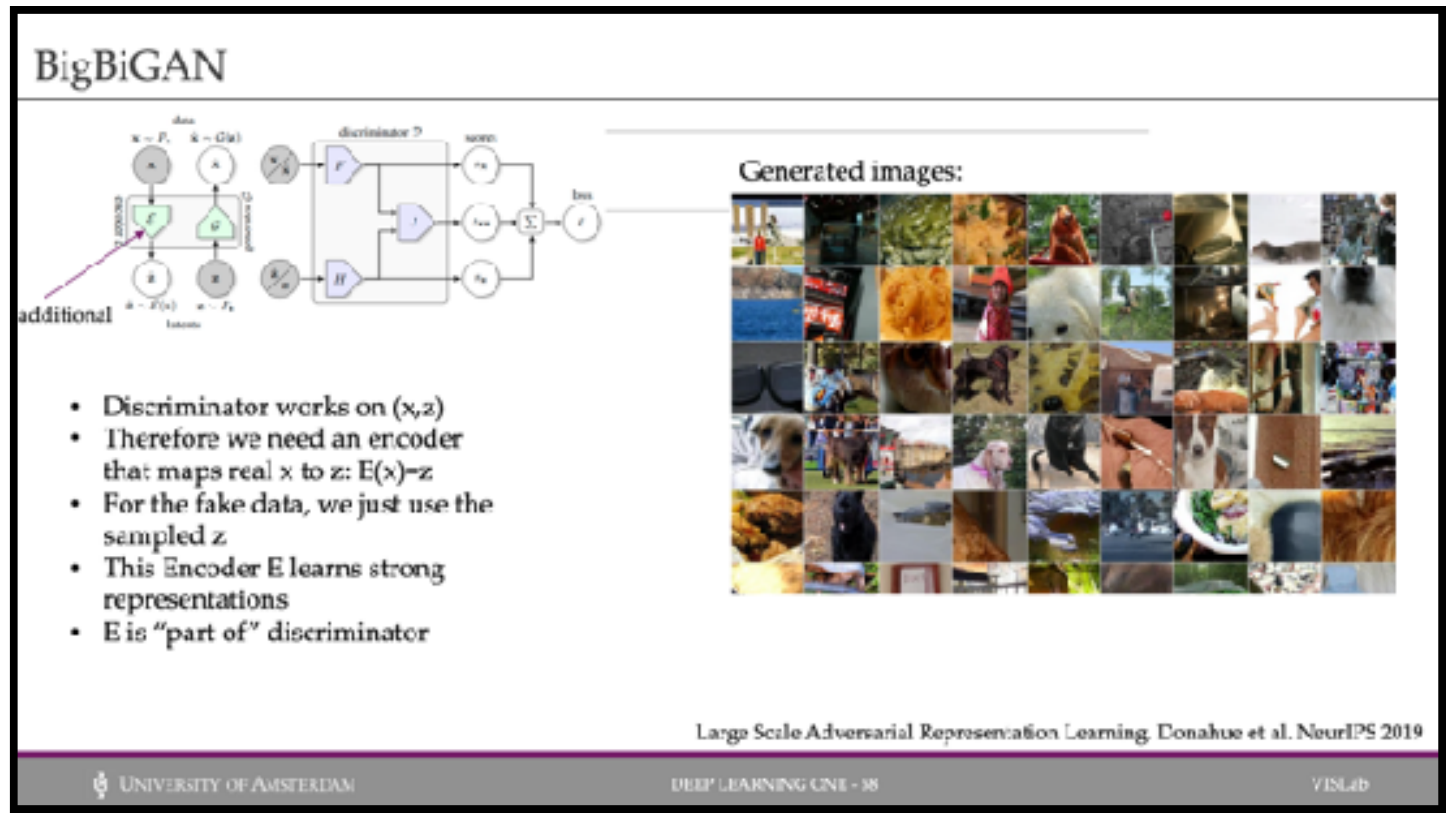
Self-supervised learning came up in multiple previous lectures.



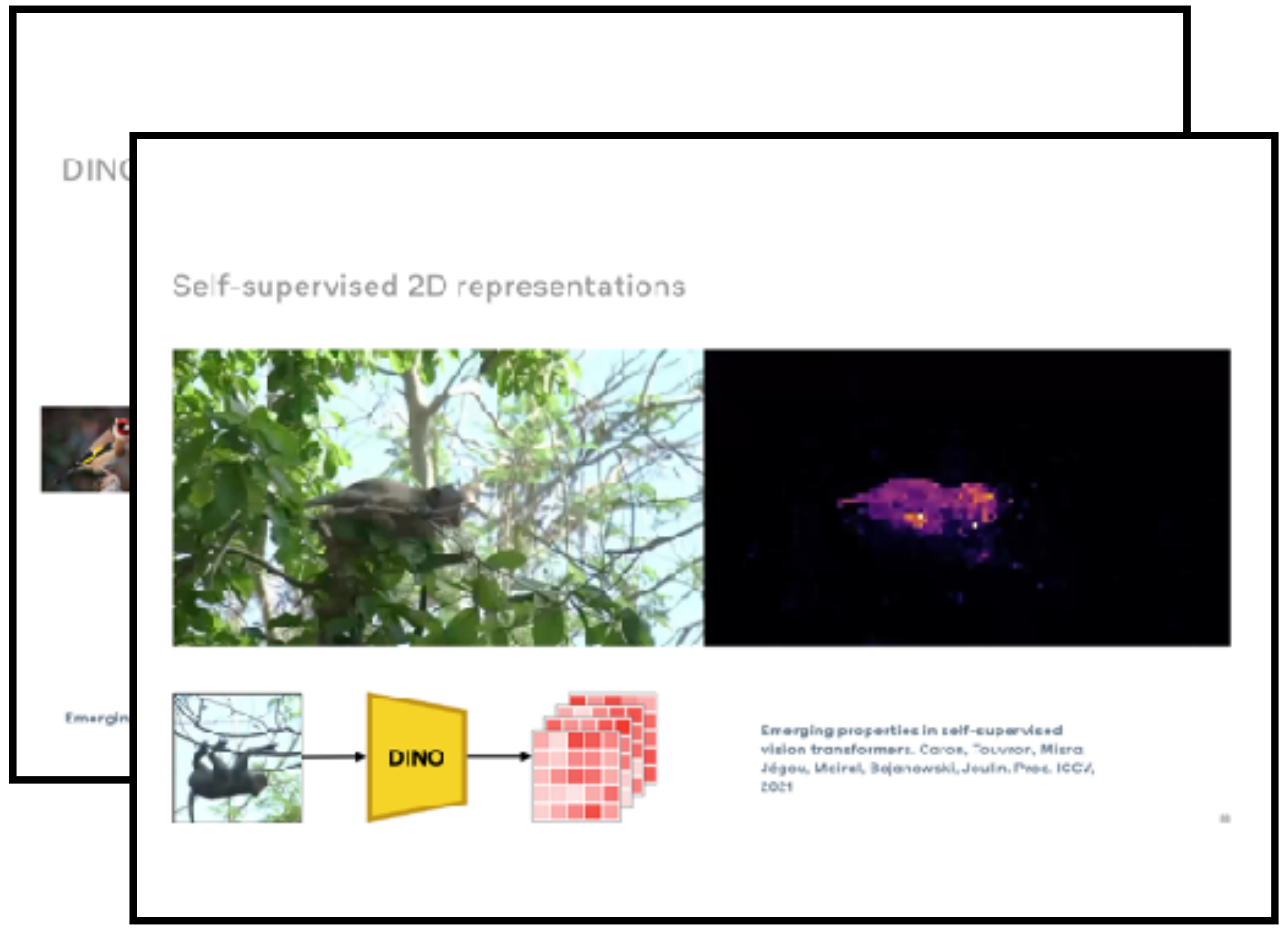
Lecture 5



Lecture 7



Lecture 10



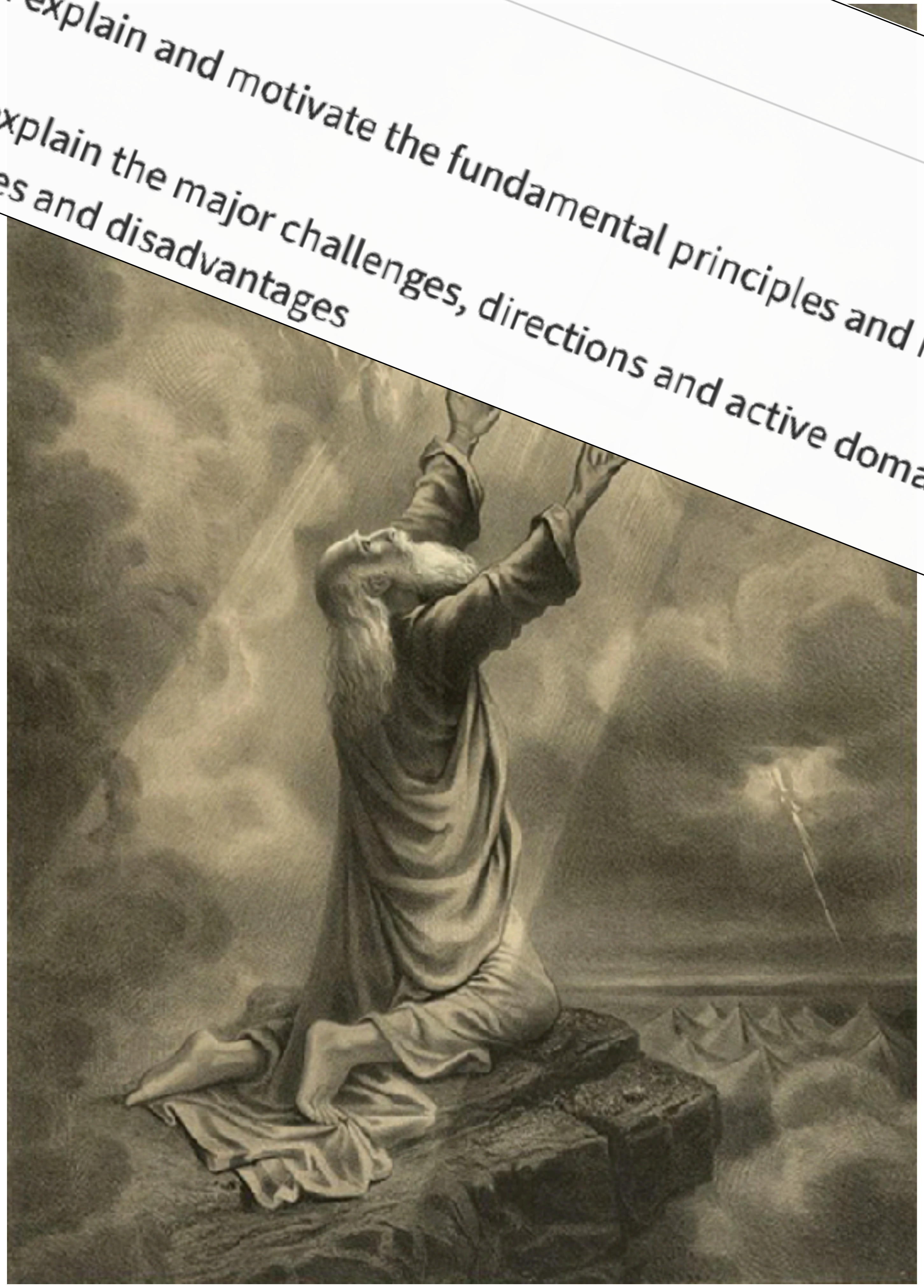
(Guest) Lecture 11

SSL is key for these two Objectives.

Course manual 2022/2023

Objectives

- The students can explain and motivate the fundamental principles and mechanisms behind Deep Learning's past present and future
- The students can explain the major challenges, directions and active domains of research in the field of deep learning along with their advantages and disadvantages



Today:

What is *Self-supervised learning* (SSL)?

Why do we want to do *SSL*?

How to do *SSL*?

* The data

* The augmentations

* The methods

Note: SSL is an active research field with many new weekly discoveries.
Things change and there's no good textbook yet, so we will cover some research papers today.

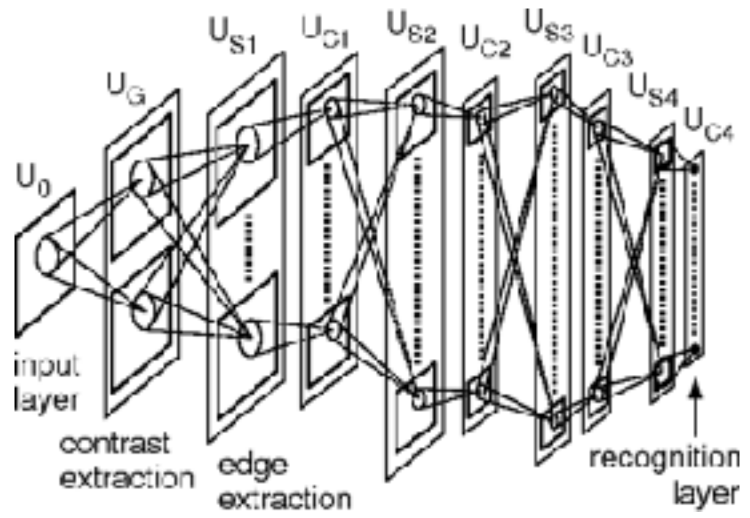
Introduction to self-supervised learning in computer vision

Part: "What"?

The field of AI has made rapid progress, the crucial fuel is data

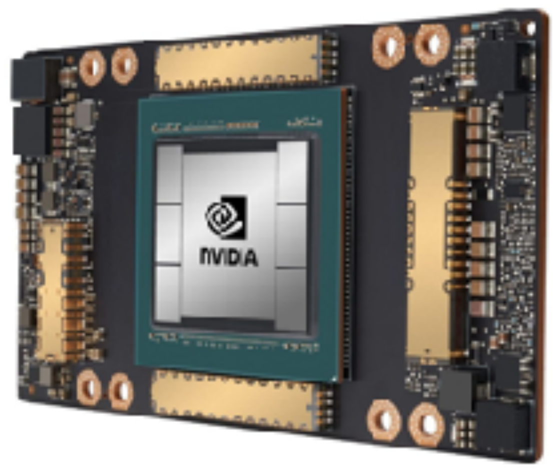
Algorithms

Deep neural networks



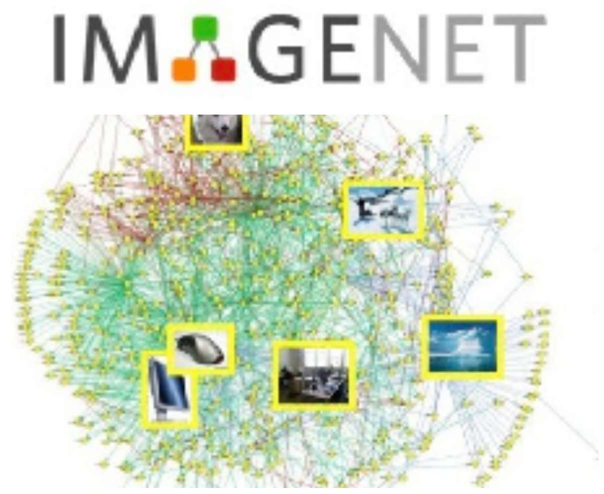
Hardware

GPUs



Data

Large scale datasets

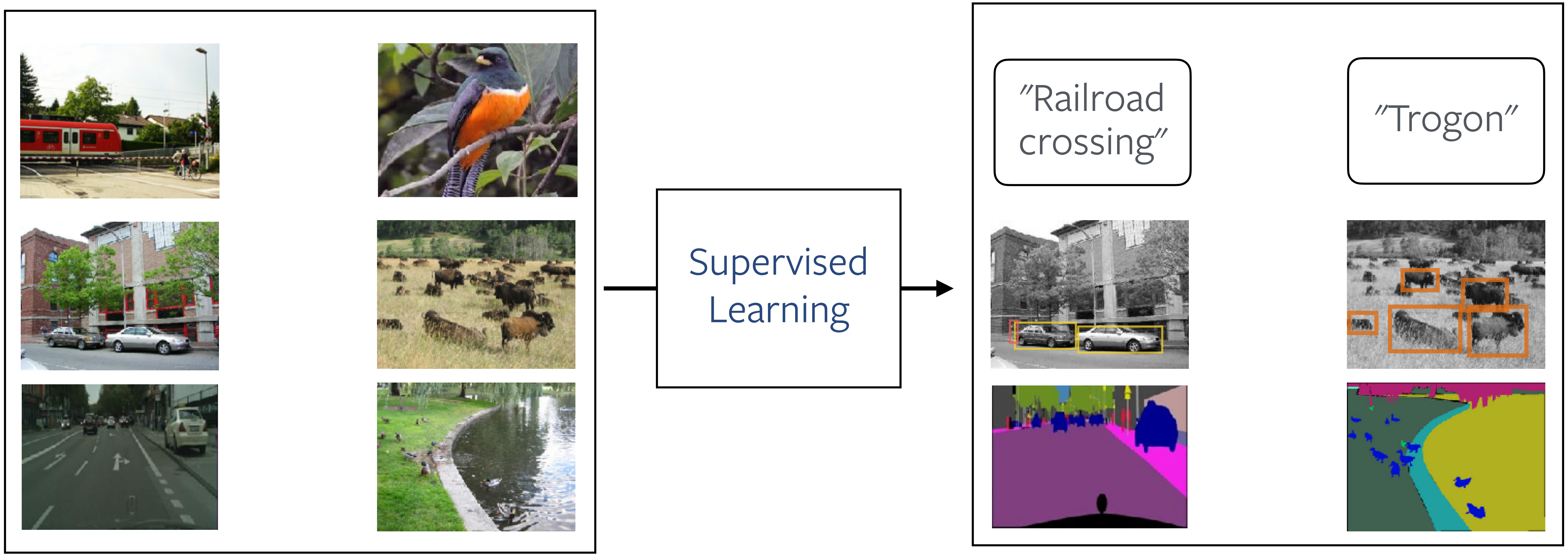


Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Fukushima, K., Biol. Cybernetics 1980
Object Recognition with Gradient-Based Learning. LeCun et al. Shape, Contour and Grouping in Computer Vision 1999
ImageNet: A Large-Scale Hierarchical Image Database. Deng, et al. CVPR, 2009.
ImageNet Classification with Deep Convolutional Neural Networks., Krizhevsky et al., NeurIPS 2012

Manual annotations for the data are limiting.

Images are often cheap

But manual annotations are expensive:
e.g. 30min per image / requiring experts



ImageNet: A Large-Scale Hierarchical Image Database. Dong et al. CVPR 2009
The Cityscapes Dataset for Semantic Urban Scene Understanding. Cordts et al. CVPR 2016
Scene parsing through ADE20K dataset. Zhou et al. CVPR 2017.

Solving the problem of expensive annotations: self-supervision.

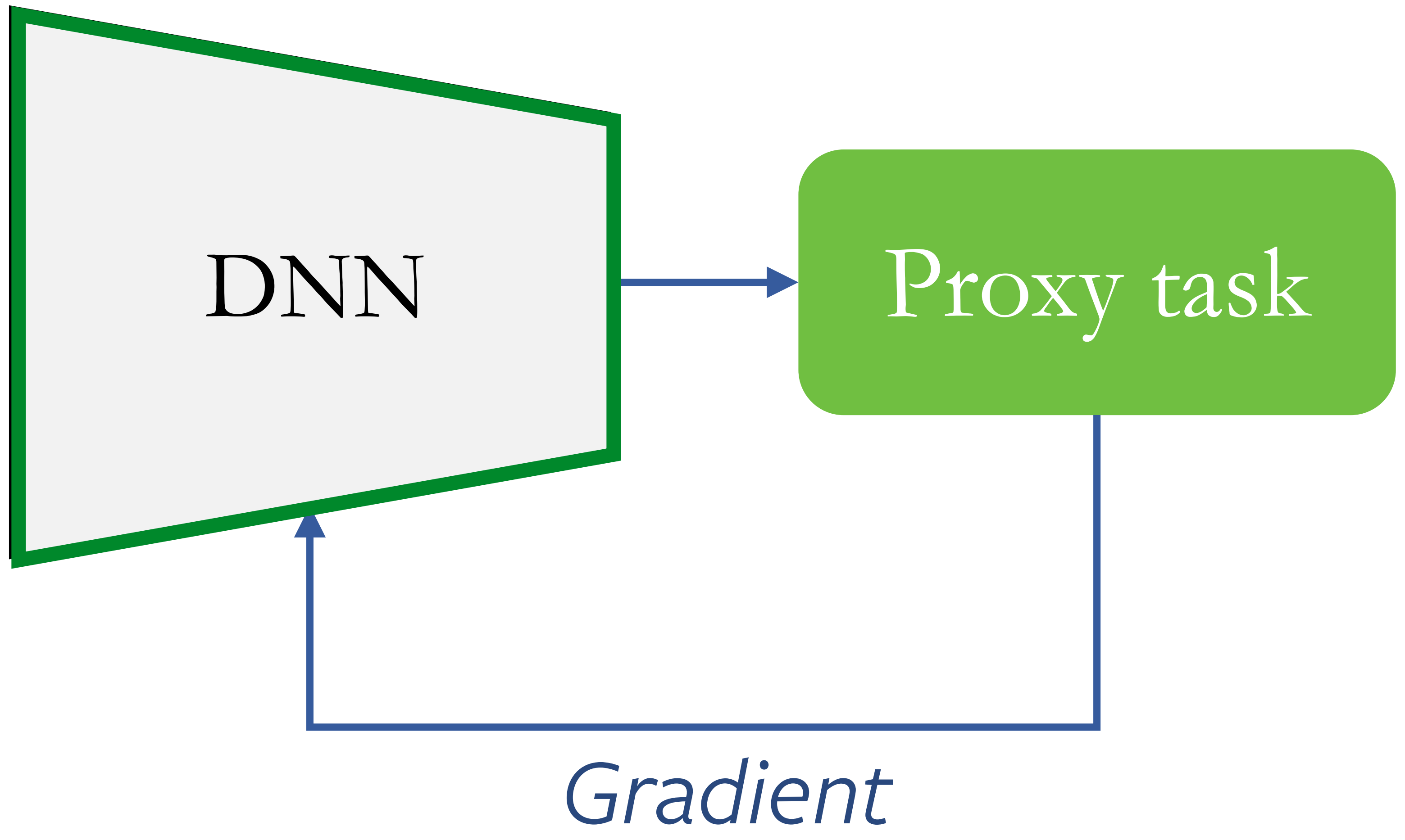
Self-supervision

Extract a supervisory signal from the raw data alone

General procedure of self-supervised learning.

Phase 1: Pretraining

Unlabelled data
+ transformations

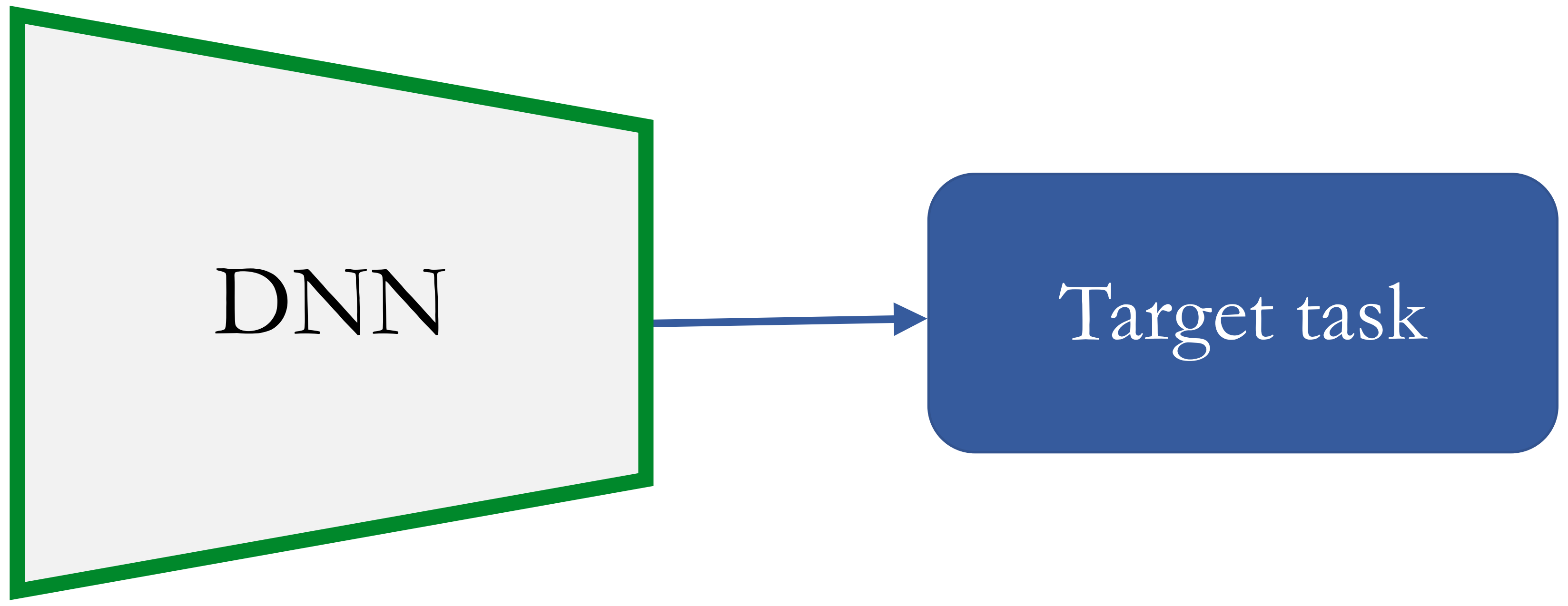


Types:

- Geometry based
- Clustering
- Contrastive
- Generative (partial/full)
- (more)

Phase 2: Downstream tasks

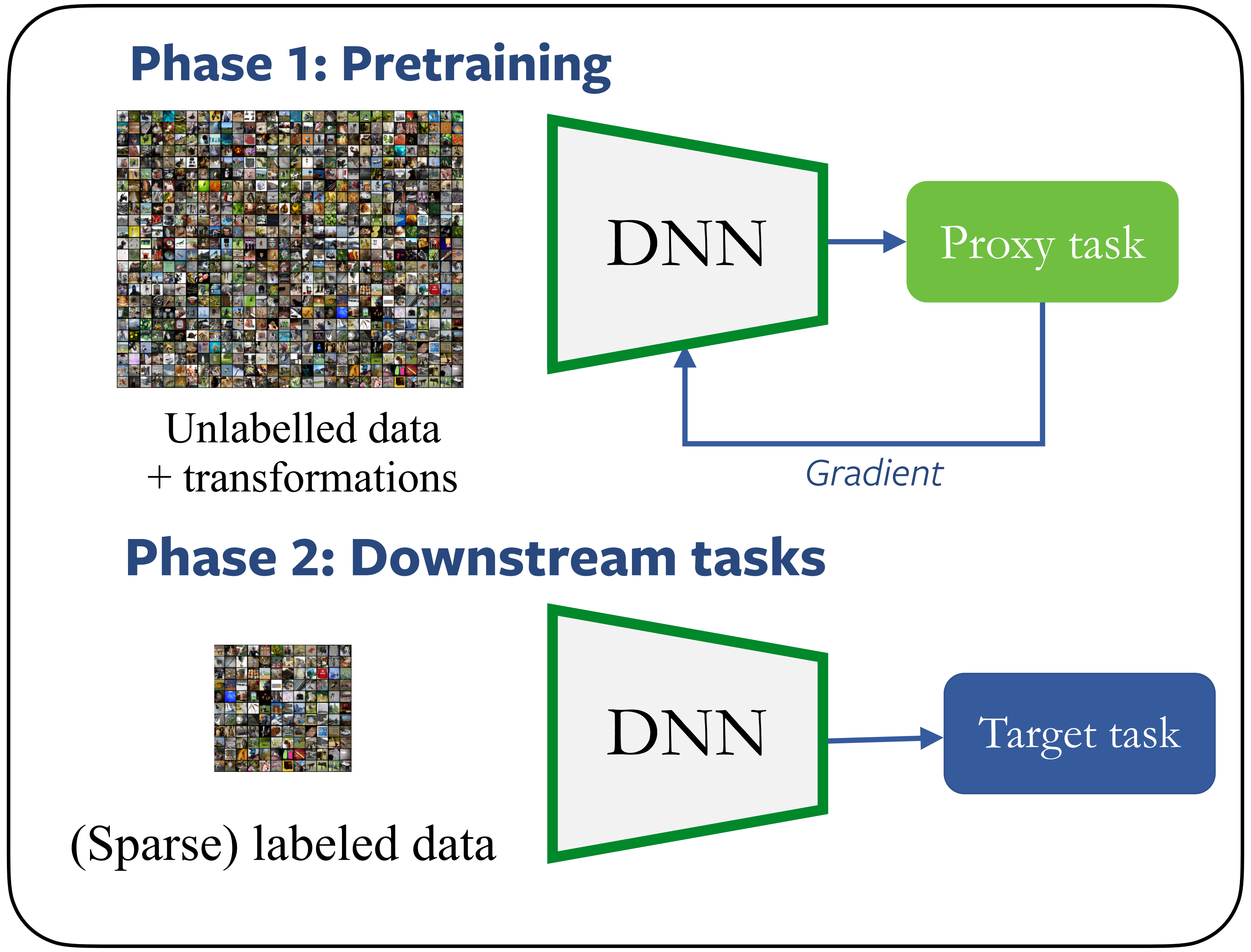
(Sparse) labeled data

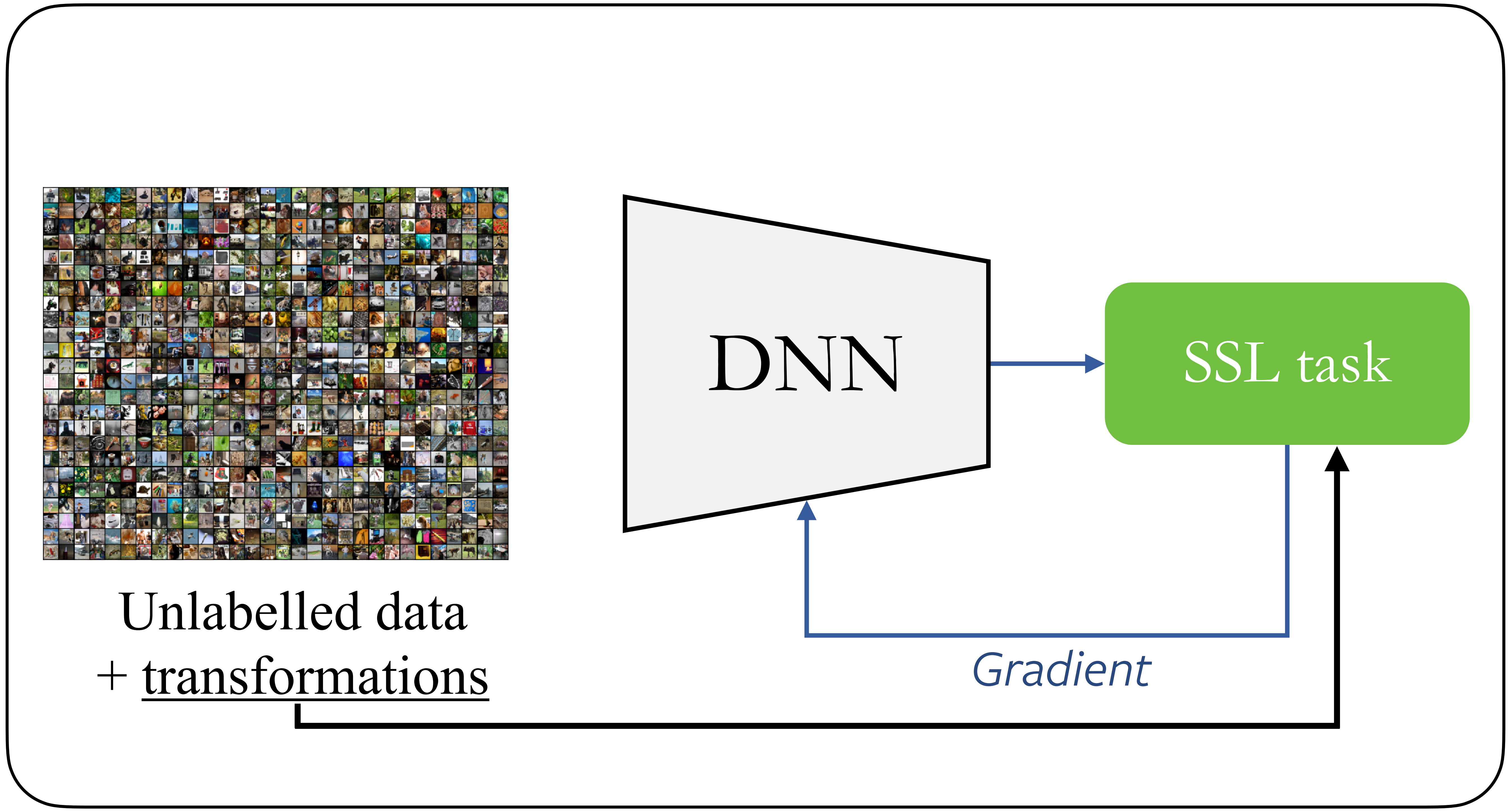


Types:

- Limited fine-tuning (e.g. linear layer)
- Finetuning (w/ full or fraction of dataset)

General procedure of self-supervised learning.





Representation Learning

Useful Self-supervised Learning, e.g.

SSL object detection & segmentation

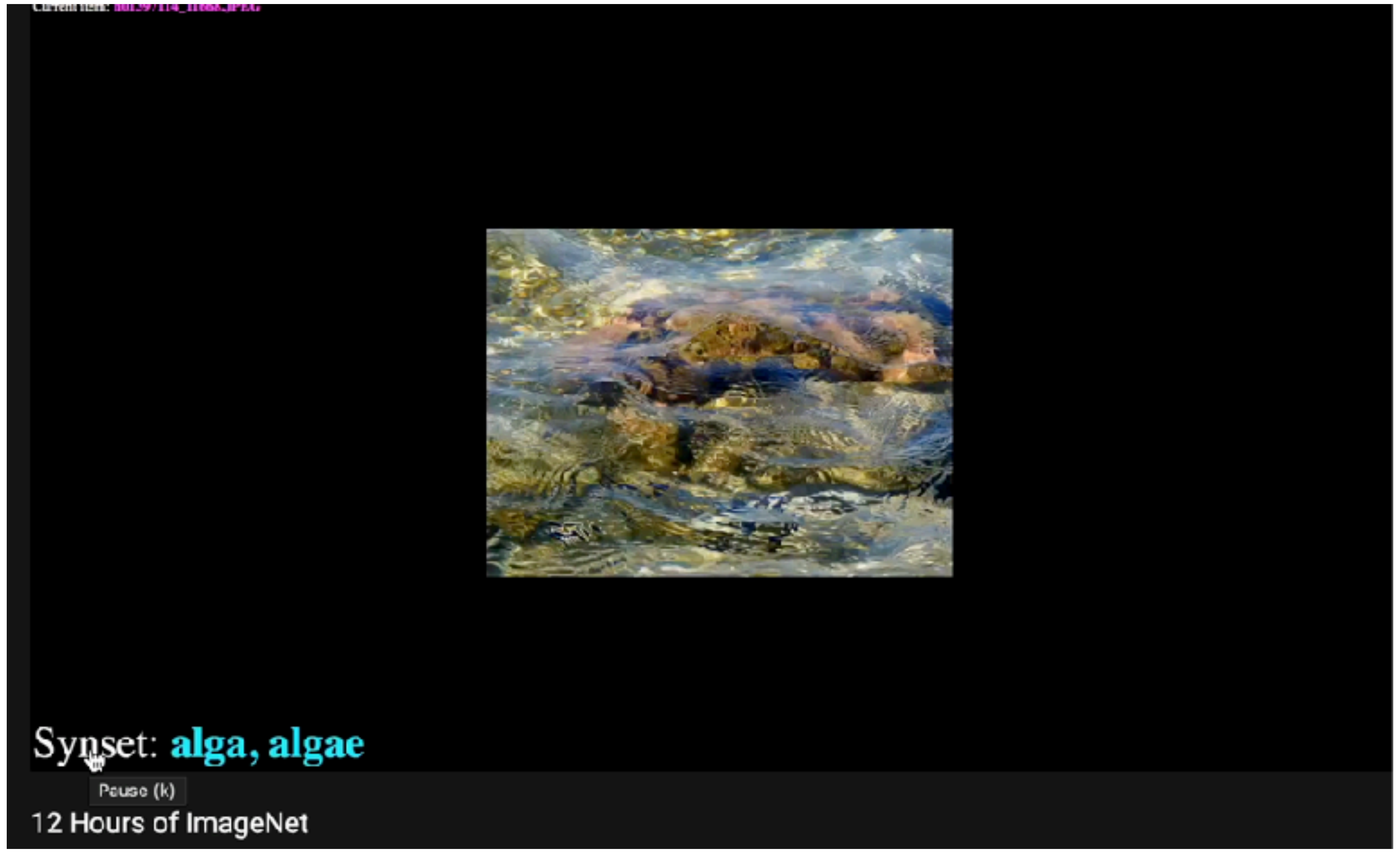
SSL speaker detection, SSL dataset labelling etc..

Introduction to self-supervised learning in computer vision

Part: "Why"?

Reason 1: Scalability

(above) x 50 =1.2M images

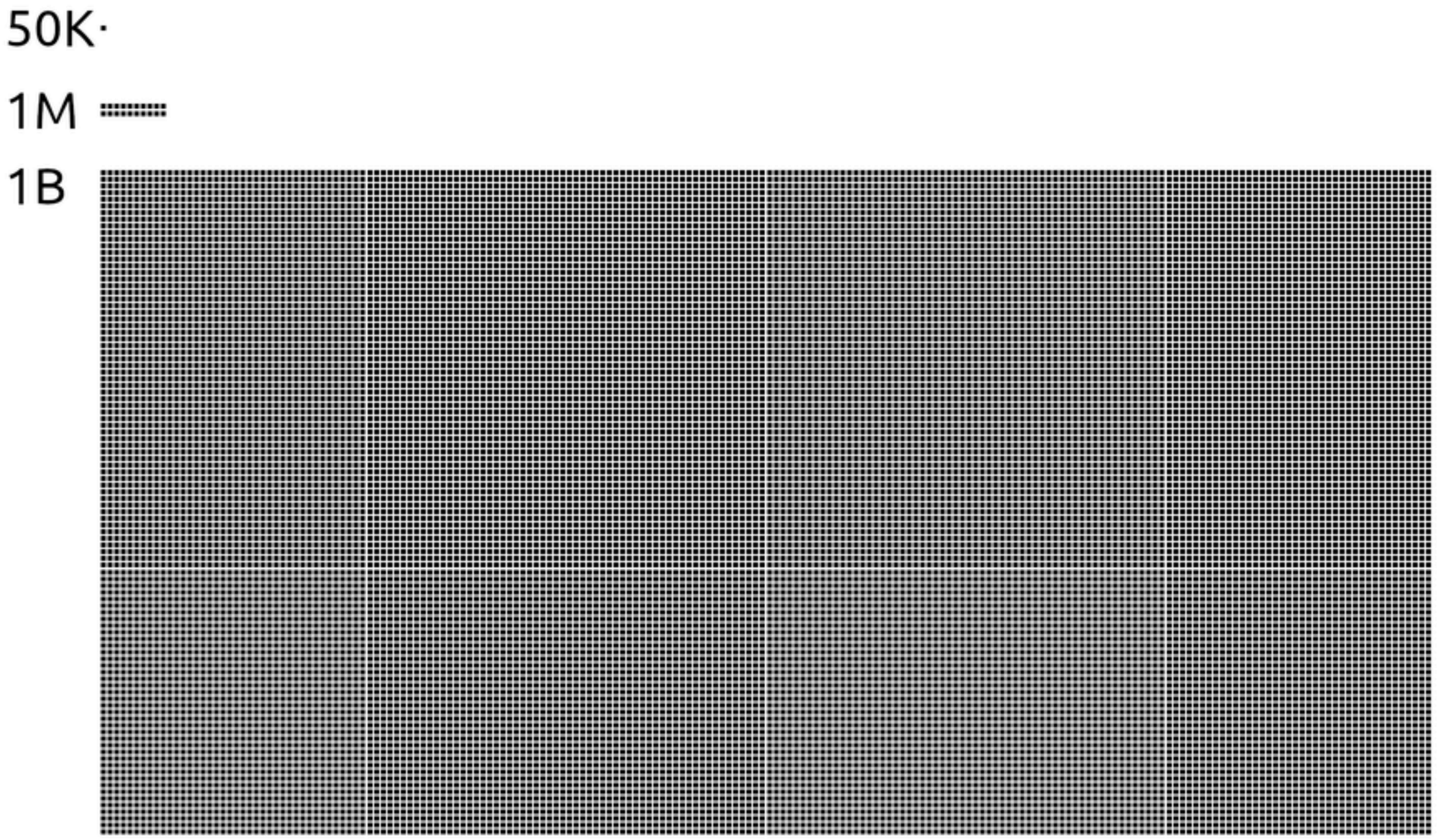


90ms * 1.2M = 30h

ImageNet: A Large-Scale Hierarchical Image Database. Deng, et al. CVPR, 2009.
12 hours of ImageNet: https://www.youtube.com/watch?v=PC60JL-lMzA

Reason 1: Scalability

Instagram: >50B images

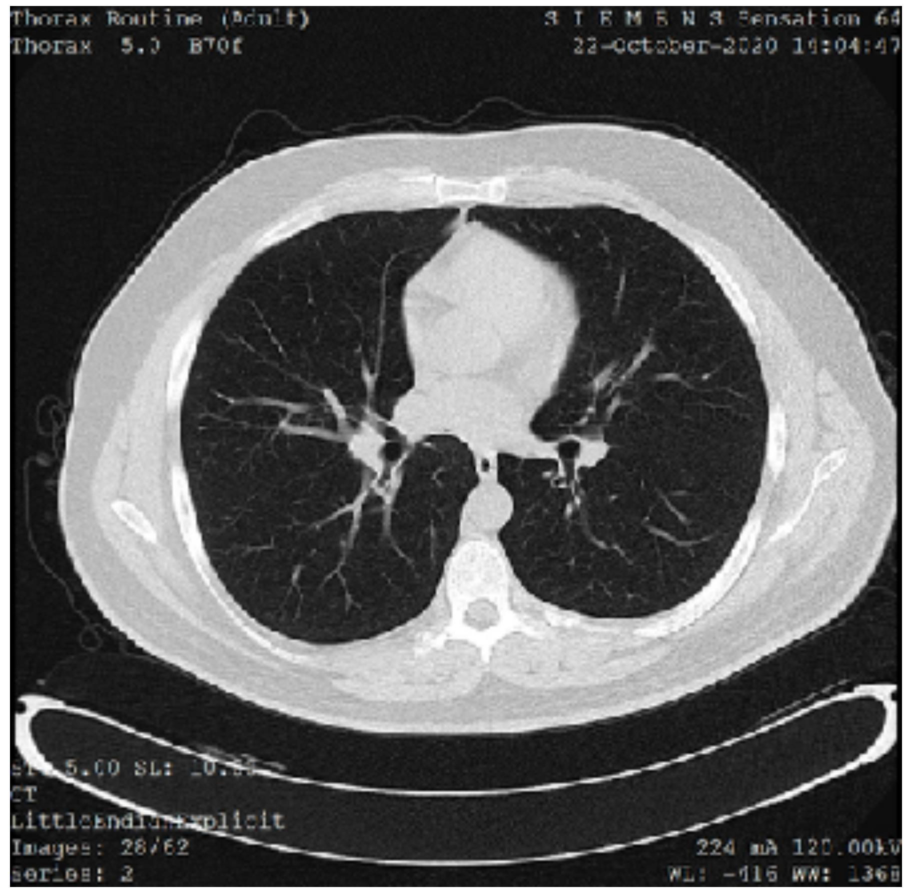


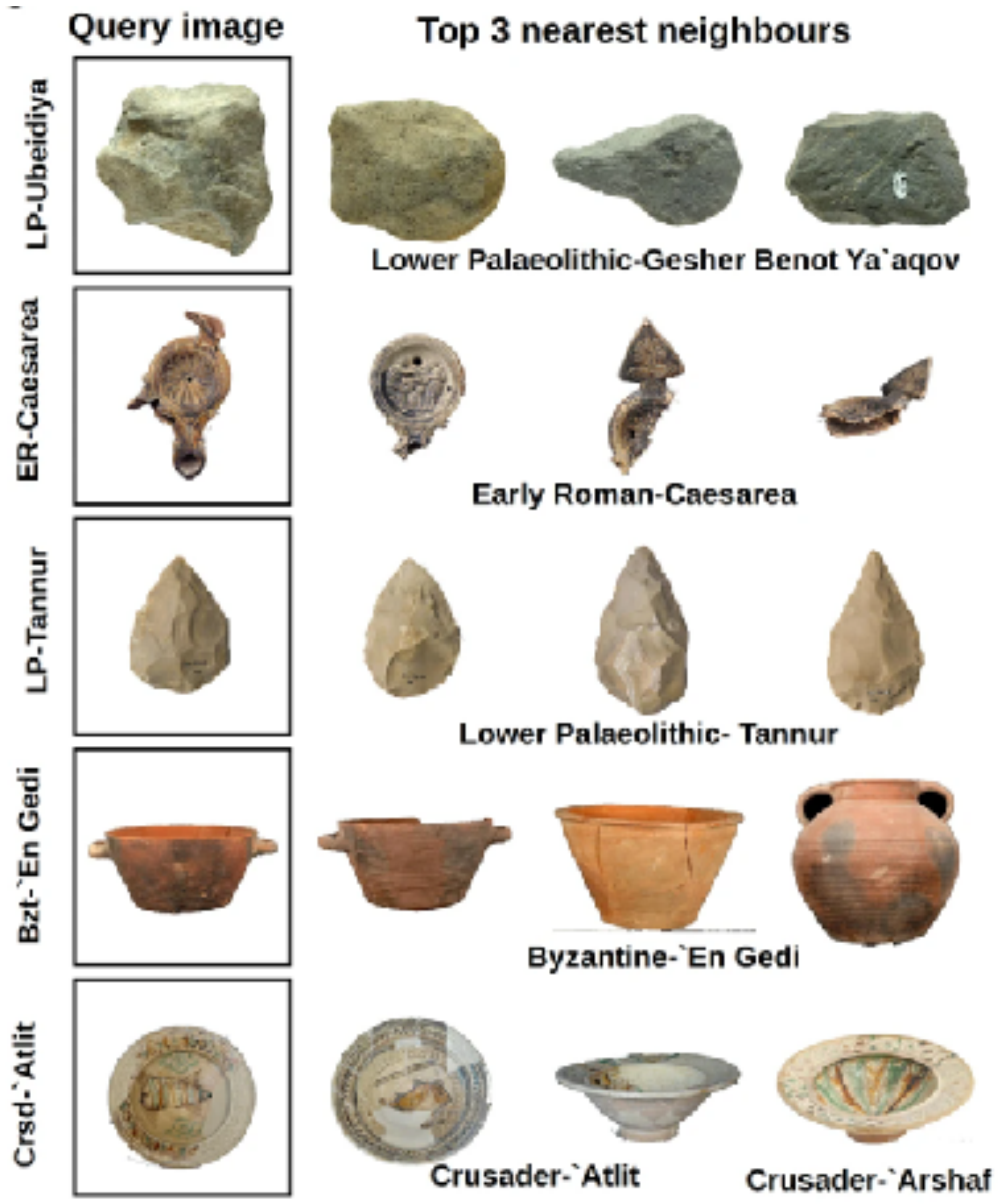
Annotation is expensive, yet datasets keep getting bigger.

Reason 2: Constantly changing domains

Unclear when & what to relabel. Again, large costs just to "keep up".

Reason 2: Accessibility & generalisability





Pretrained models are very useful for a variety of tasks.

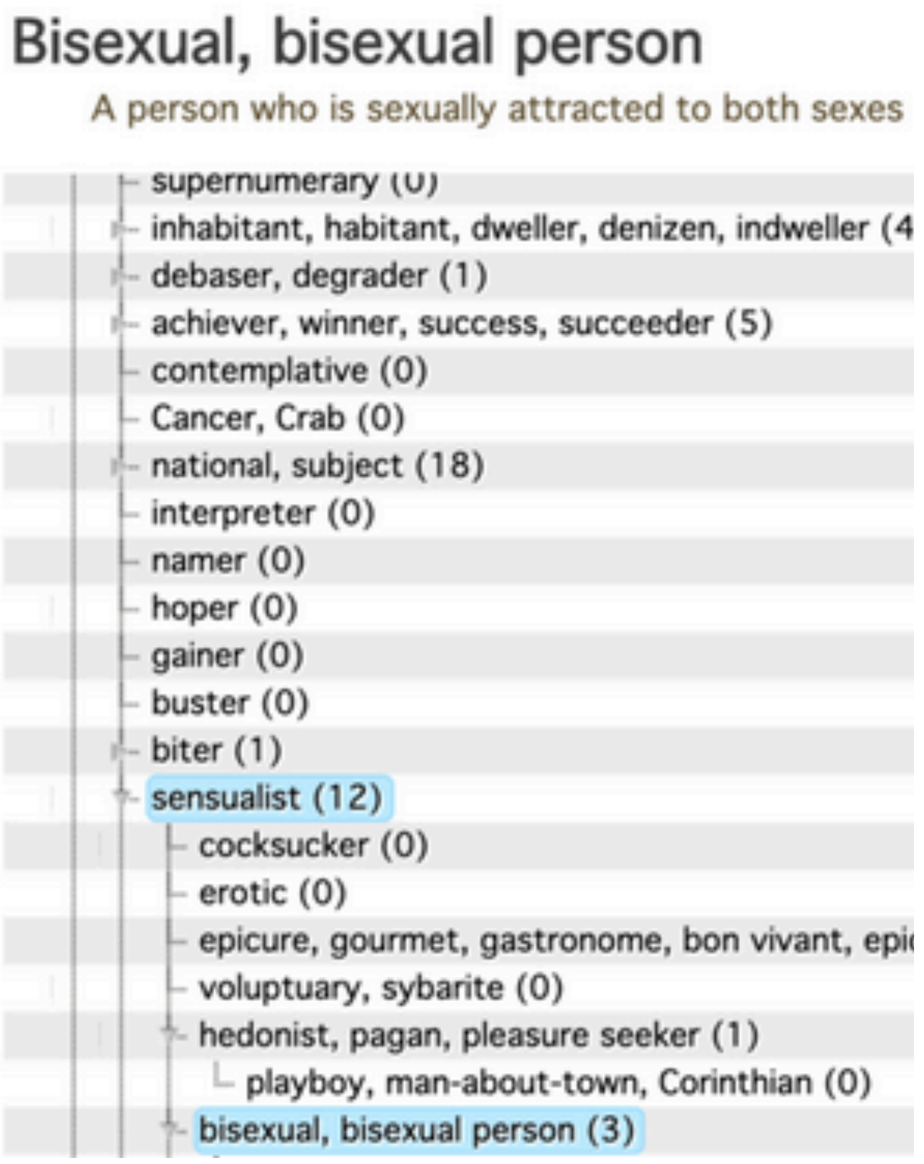
UNIVERSITY OF AMSTERDAM

https://www.kaggle.com/c/herbarium-2019-fgvc6, https://en.wikipedia.org/wiki/Medical_imaging#/media/File:CT_Scan_General_Illustration.jpg
Schaefer et al. Deep convolutional neural networks as strong gravitational lens detectors. Astronomy & Astrophysics.
Resler et al. A deep-learning model for predictive archaeology and archaeological community detection. Nature Humanities & Social Sciences Communications.

Reason 3: Ambiguity of labels

"A house"?

"A boat"?

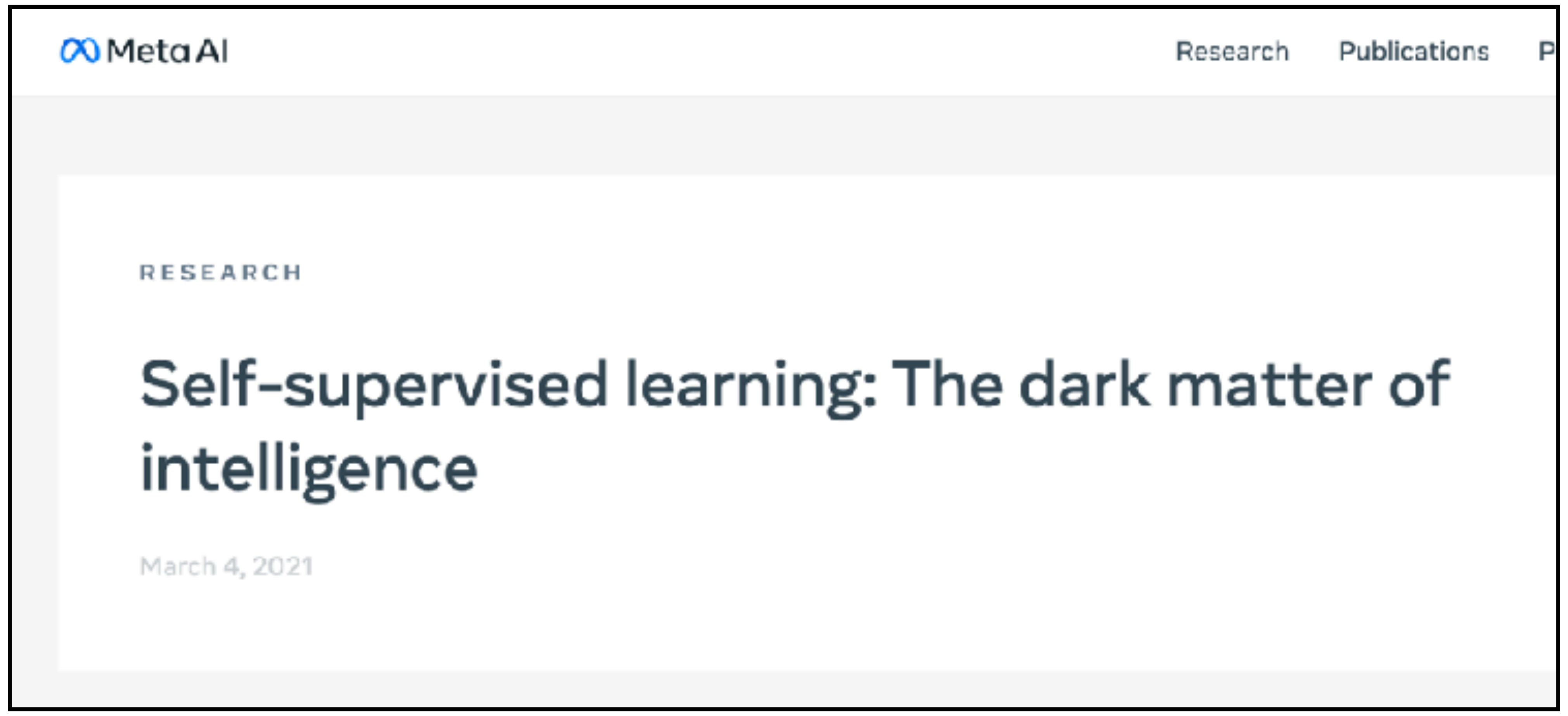


Nonsensical visual labels

**Labels are ambiguous at best, discriminating and bias-propagating at worst.
Do we really wish to provide our models with these priors?**

https://en.wikipedia.org/wiki/List_of_house_styles
https://www.shutterstock.com/image-illustration/flat-ships-sailing-yachts-marine-sailboats-1903407259
https://excavating.ai/ Crawford & Paglen

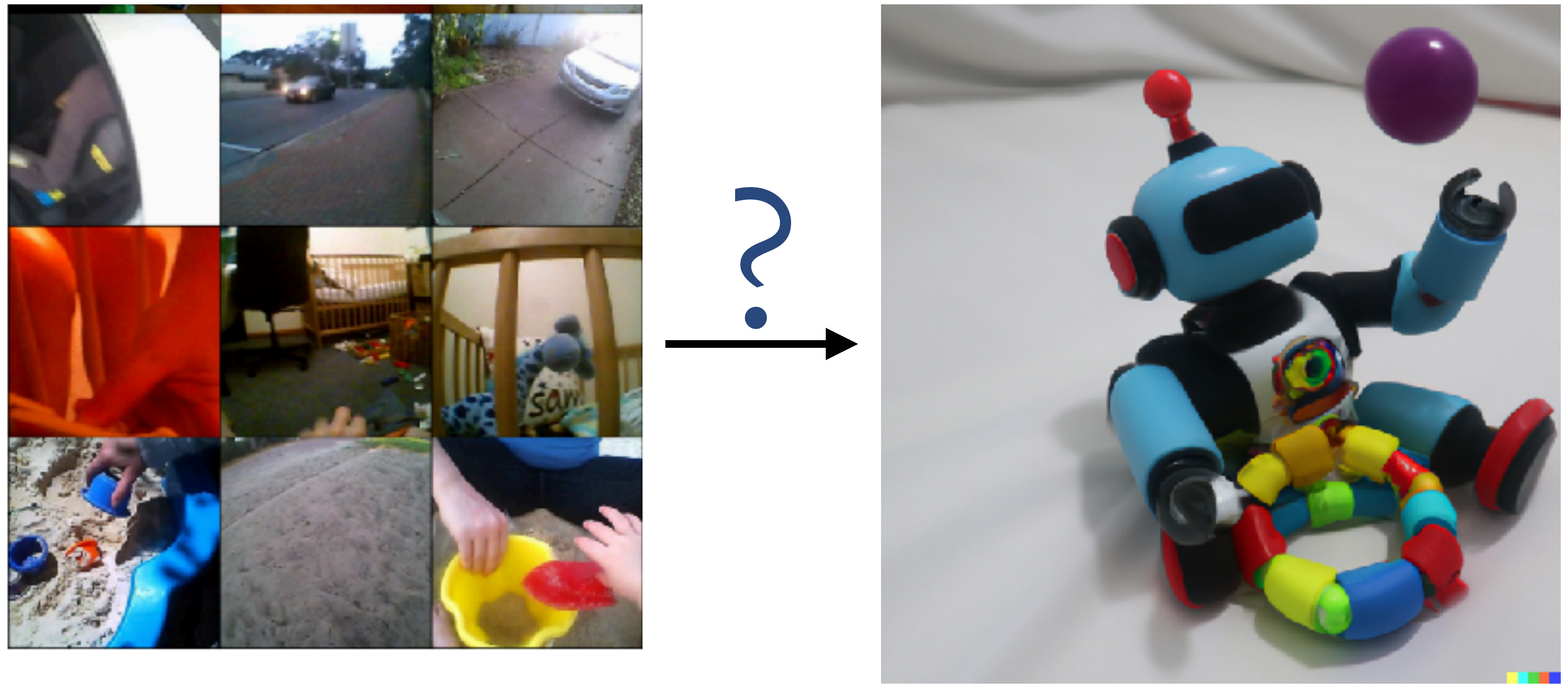
Reason 4: Investigating the fundamentals of visual understanding


Meta AI Research Publications P

RESEARCH

Self-supervised learning: The dark matter of intelligence

March 4, 2021



As babies, we learn how the world works largely by observation. We form generalized predictive models about objects in the world by learning concepts such as object permanence and gravity. Later in life, we observe the world, act on it, observe again, and build hypotheses to explain how our actions change our environment by trial and error.

What, if there are, are the limits of learning without labels?

https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/
Orhan et al. Self-supervised learning through the eyes of a child. NeurIPS 2020

Quiz: turn to your neighbour and briefly explain the core idea behind self-supervised learning.

Food for thought:

What are the core principles and ideas?

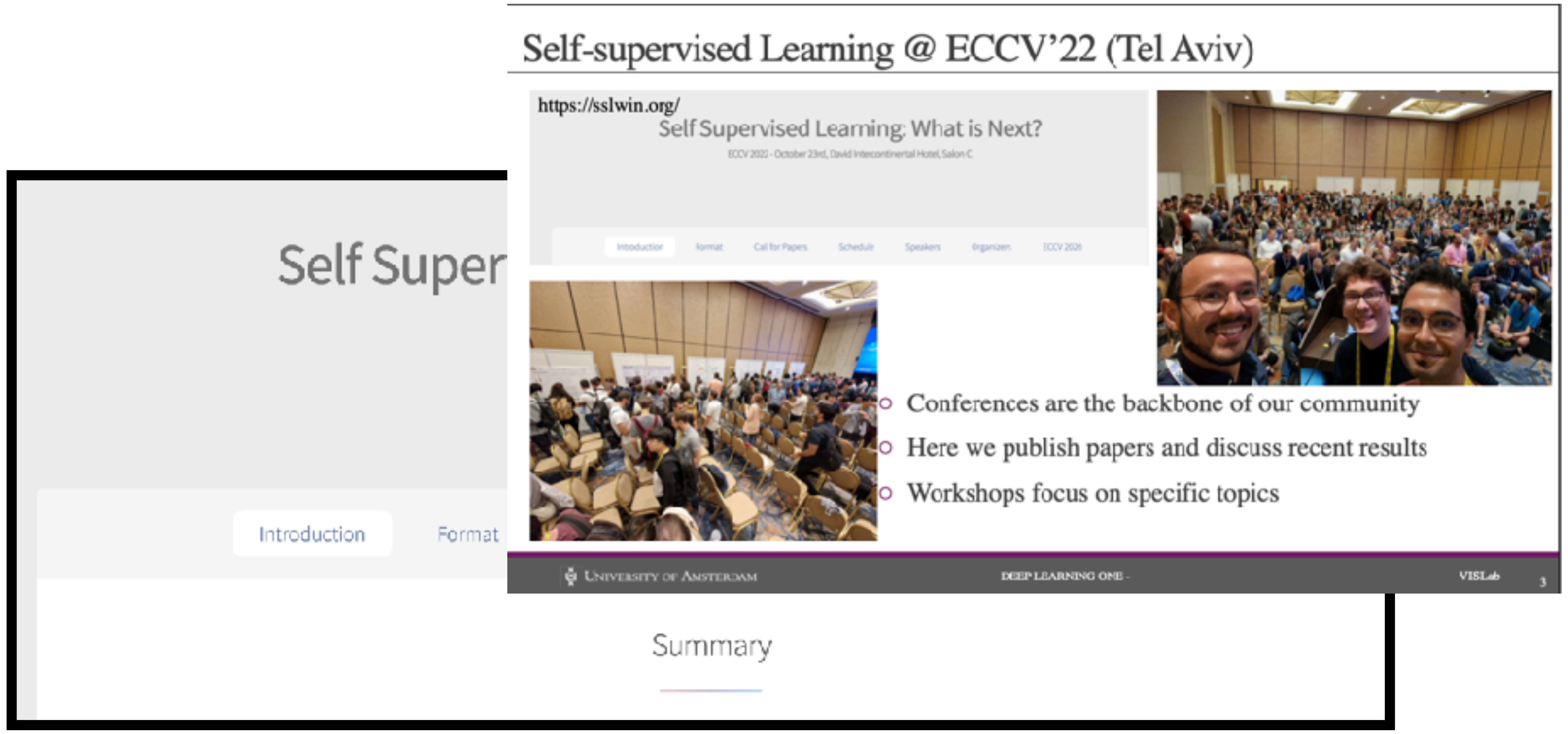
What is intuitive? What is (so far) unclear?

Our human learning experience vs the ML perspective?

Overview of self-supervised learning methods (the ”how”)

Here, we will only cover the most important works.
Further details and recent developments can be found here:

CVPR'21 Tutorial by Bursuc et al.



ECCV'20/22 workshop by Asano et al.

Was this October!

https://www.youtube.com/watch?v=MdD4UMshl1Q
https://sslwin.org/

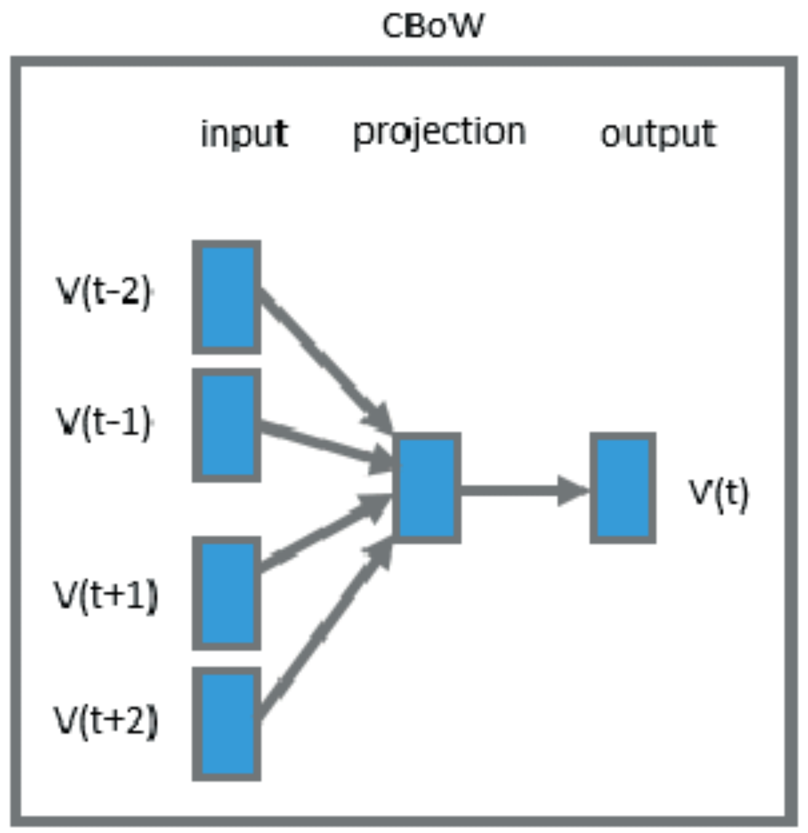
How does one learn without labels?

Need to generate a loss that provides gradients. Types of signals that we can leverage include:

- Reconstruction (full image or some within-image patch(es))
- Geometry
- Augmentation invariance
- Image uniqueness
- Assumed structure (clustering)
-

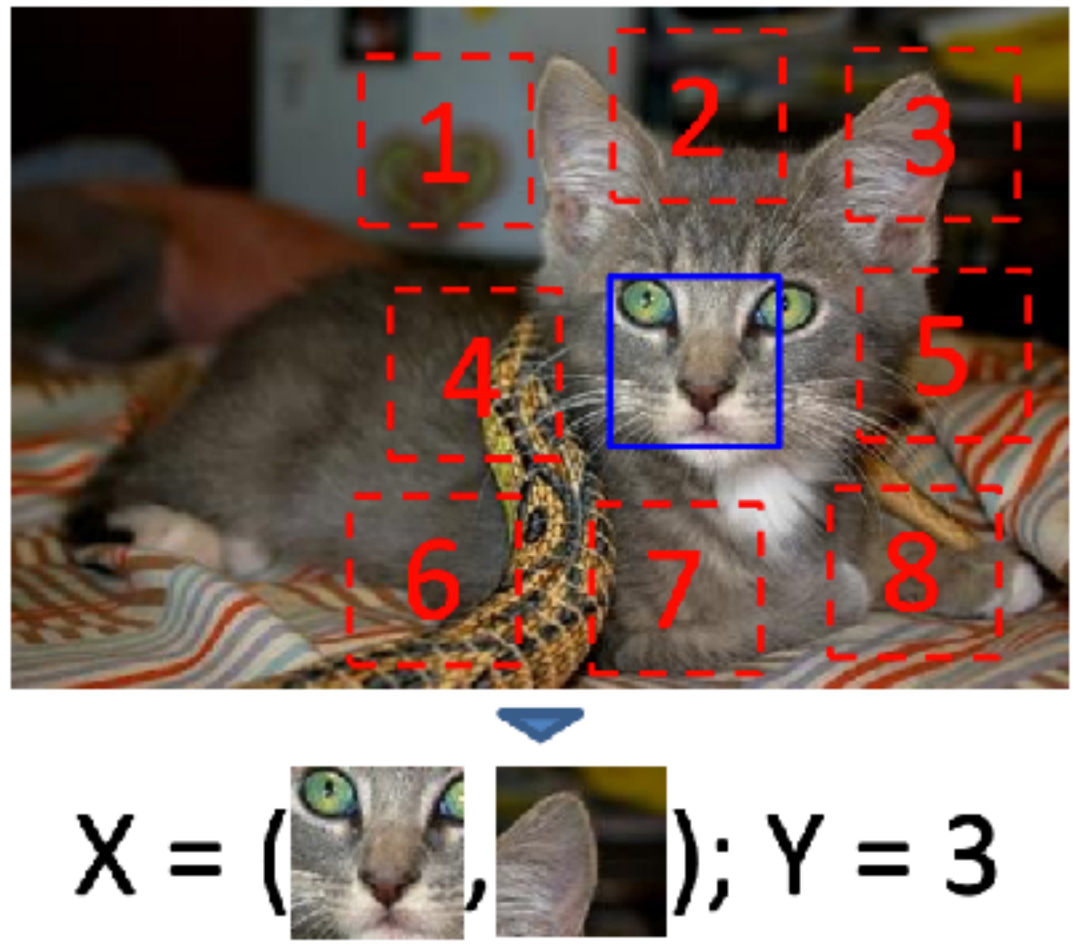
Early methods: Context prediction

Word2Vec

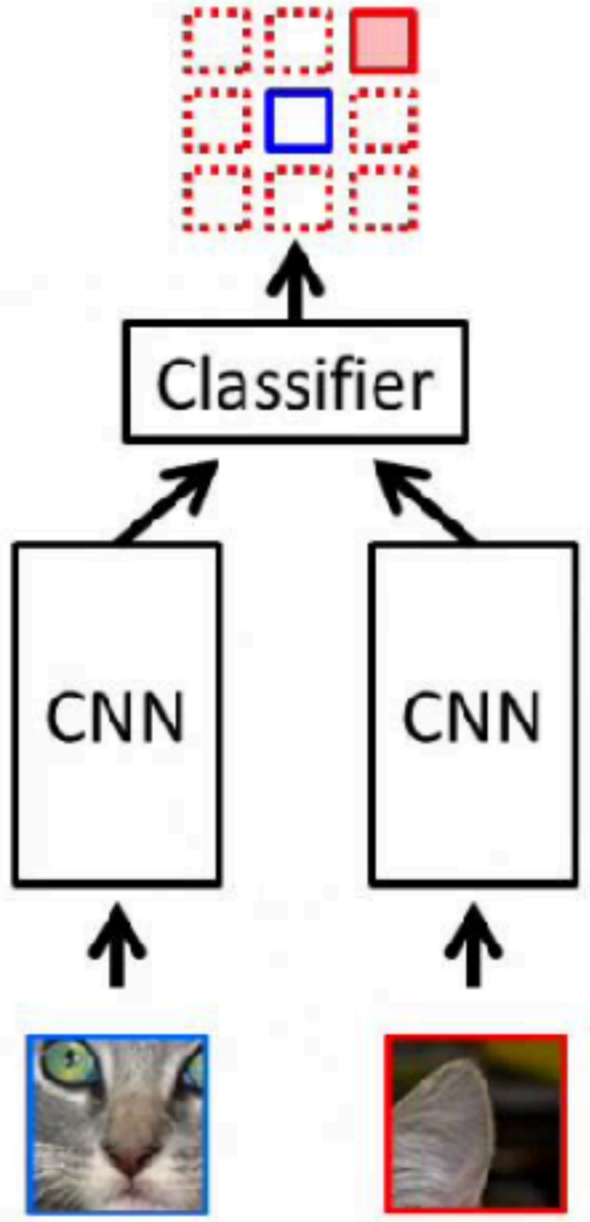


Motivated from NLP

Take some 3x3 patches



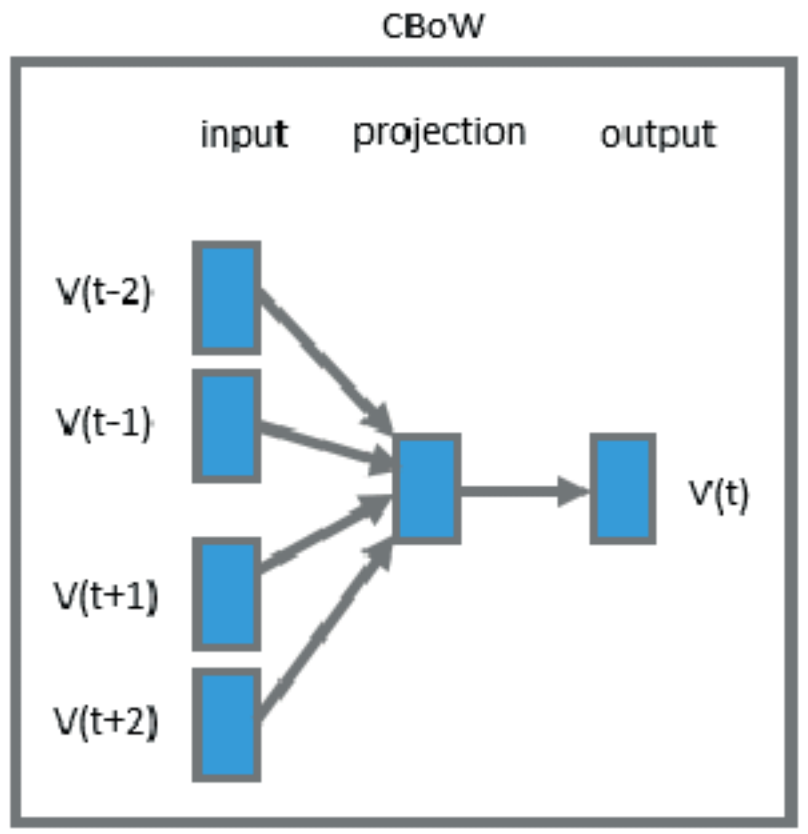
Predict where right patch comes from



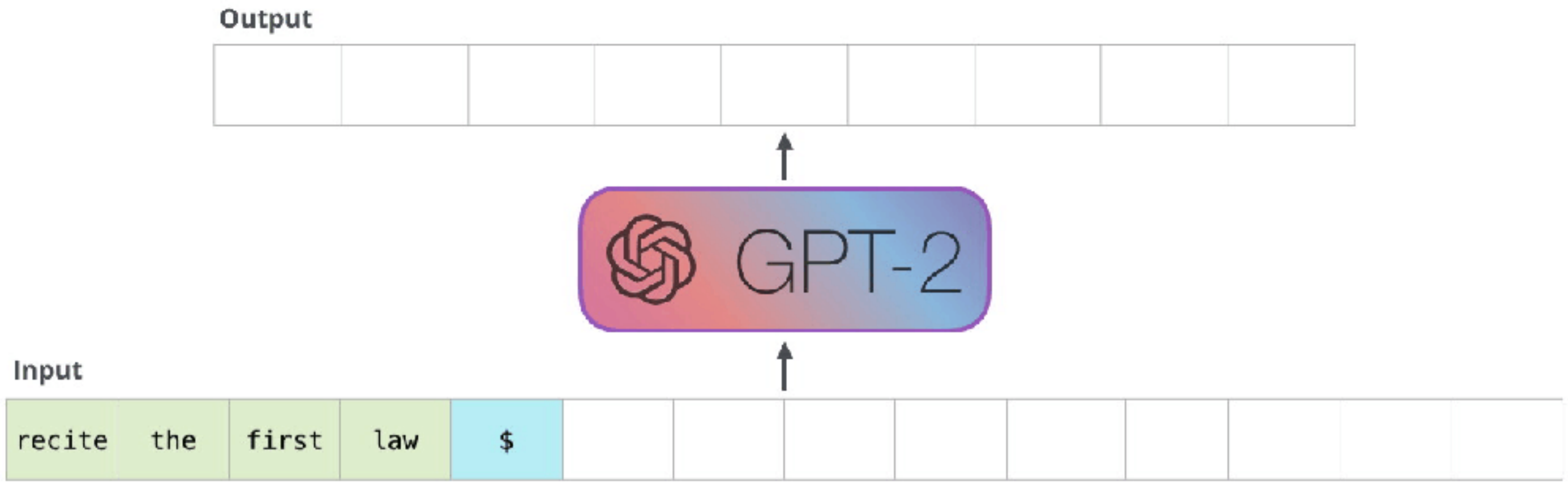
https://www.researchgate.net/figure/Word2Vec-CBOW-and-Skip-gram-There-are-two-different-methods-in-the-Word2Vec-algorithm_fig2_320829283
Doersch et al. *Unsupervised Visual Representation Learning by Context Prediction.* ICCV 2015.

Note: this is how GPT and pretty much all LLMs have been trained

Word2Vec



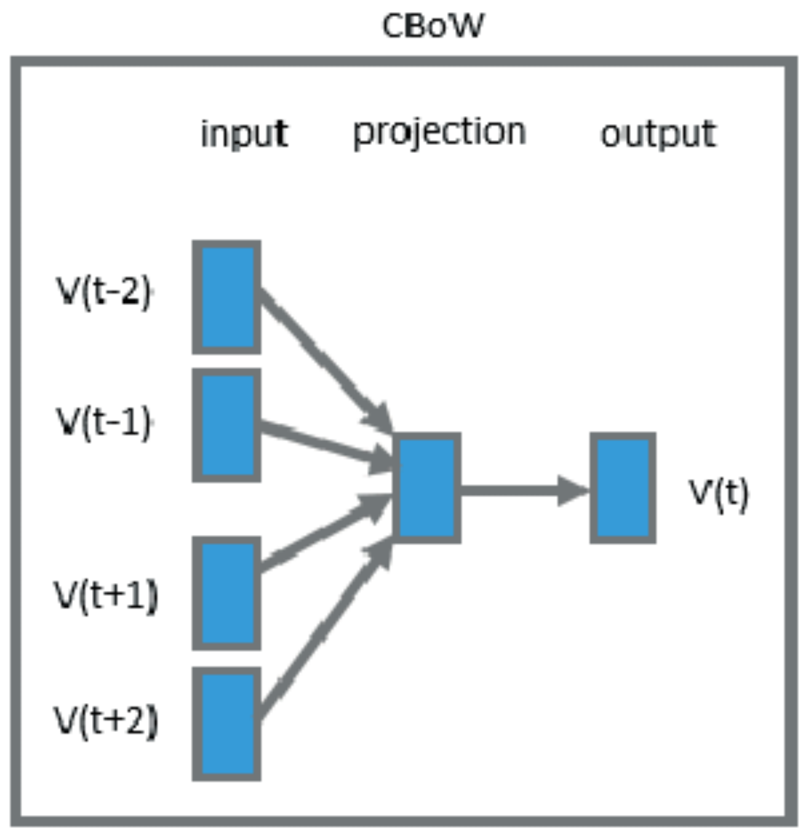
Motivated from NLP



https://www.researchgate.net/figure/Word2Vec-CBOW-and-Skip-gram-There-are-two-different-methods-in-the-Word2Vec-algorithm_fig2_320829283
Doersch et al. *Unsupervised Visual Representation Learning by Context Prediction.* ICCV 2015.

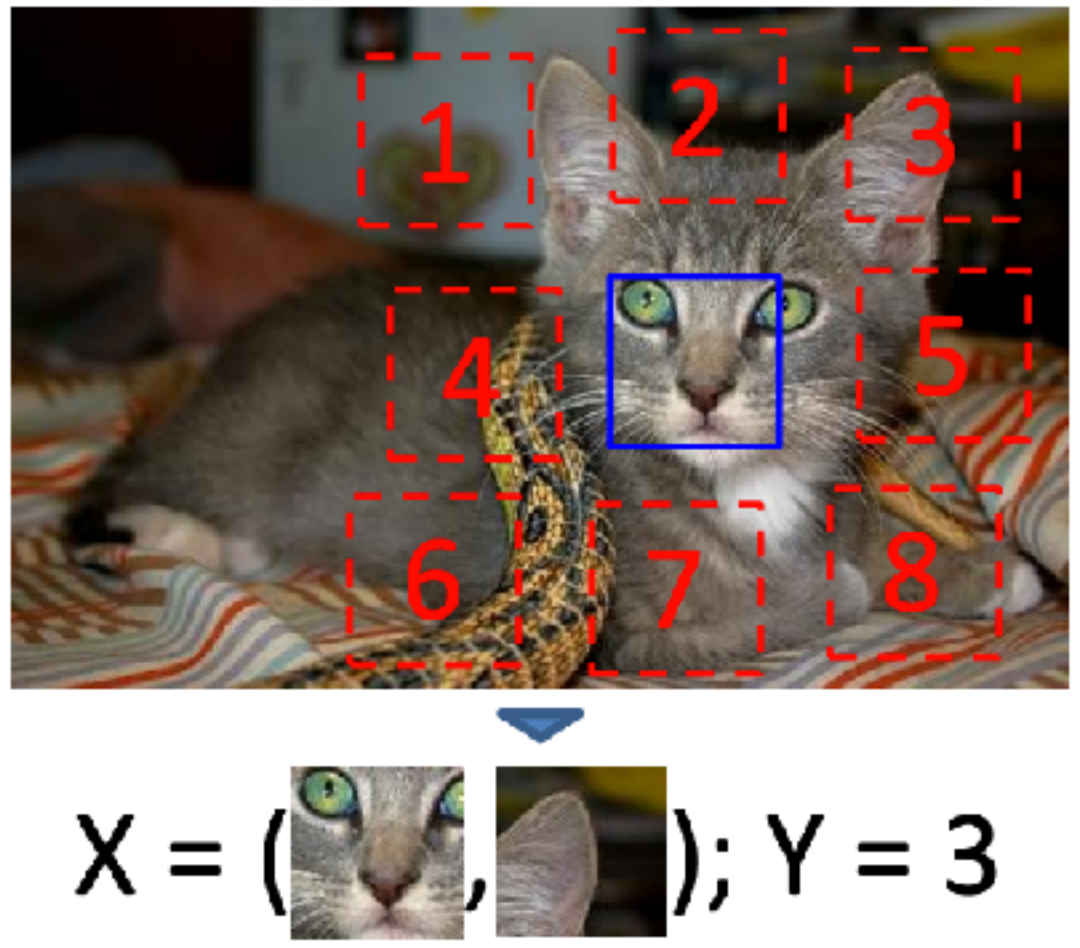
Early methods

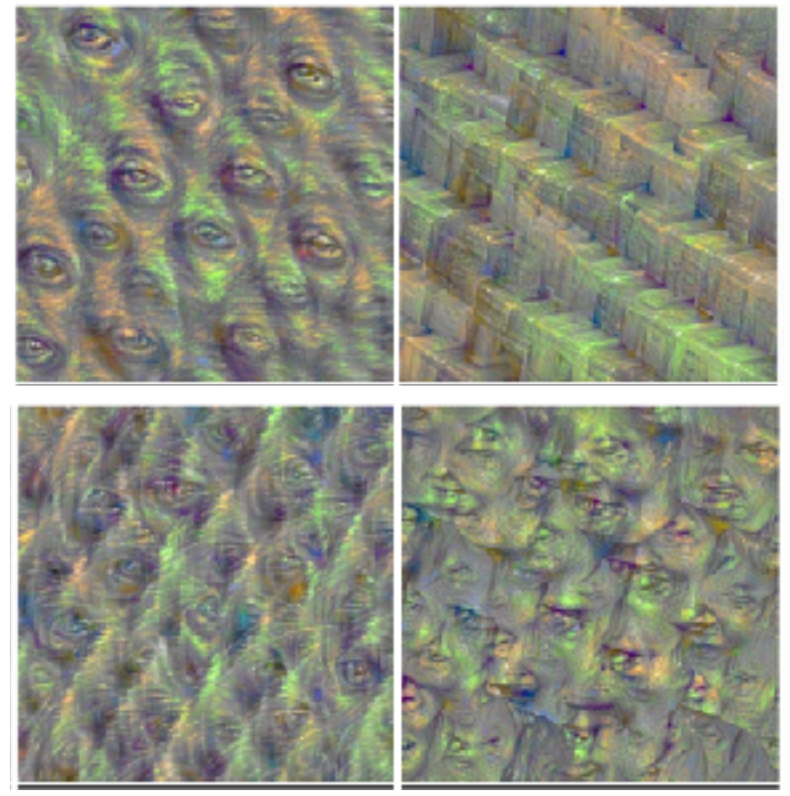
Word2Vec



Motivated from NLP

Context Prediction



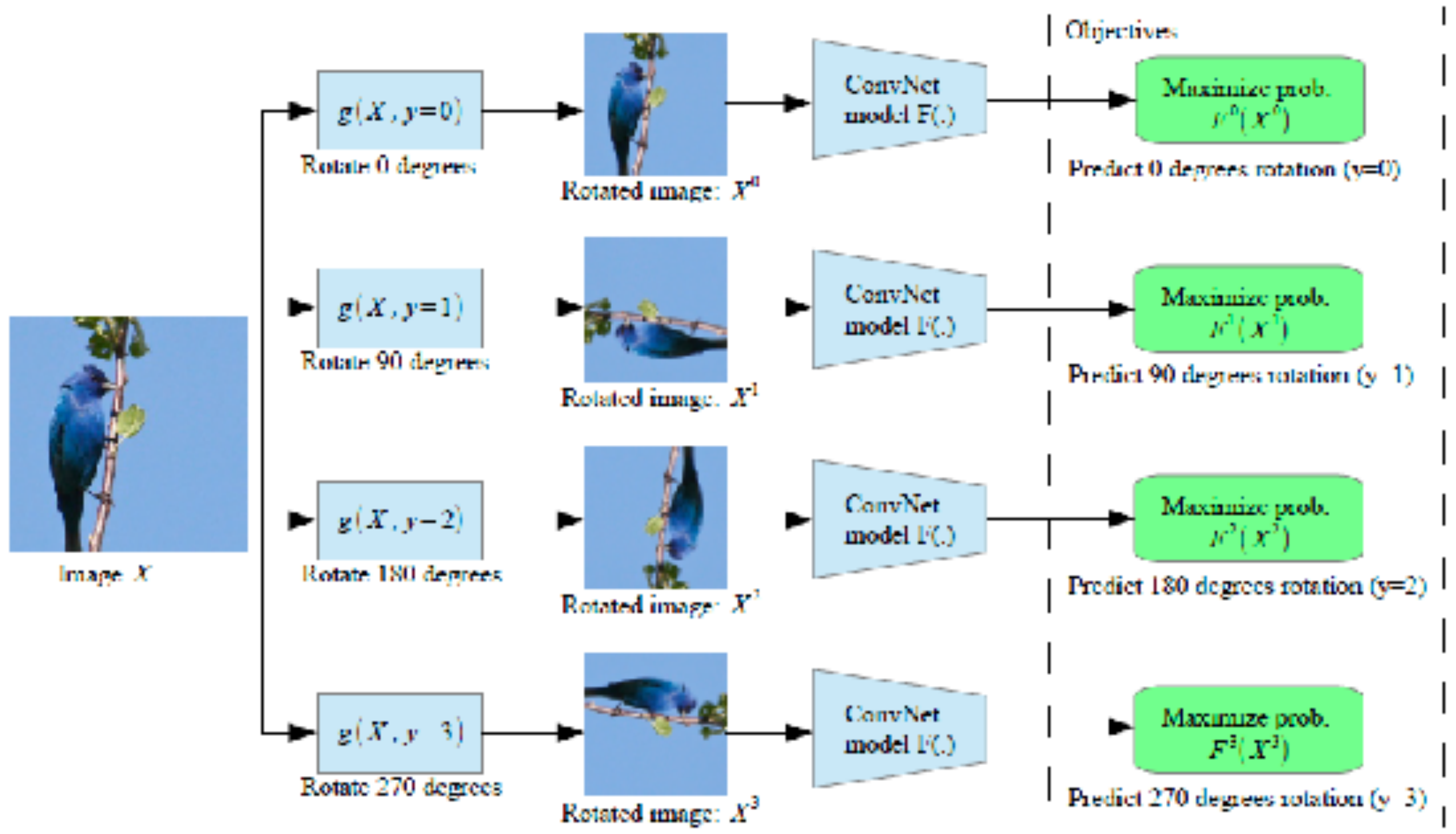


Context Encoders

(a) Input context

(c) Context Encoder
($L2$ loss)

RotNet



Learning without labels is meaningful and possible.

https://www.researchgate.net/figure/Word2Vec-CBOW-and-Skip-gram-There-are-two-different-methods-in-the-Word2Vec-algorithm_fig2_320829283
Doersch et al. *Unsupervised Visual Representation Learning by Context Prediction*. ICCV 2015.
Pathak et al. *Context Encoders: Feature Learning by Inpainting*. CVPR 2016.
Gidaris et al. *RotNet: Unsupervised Representation Learning by Predicting Image Rotations*. ICLR 2018

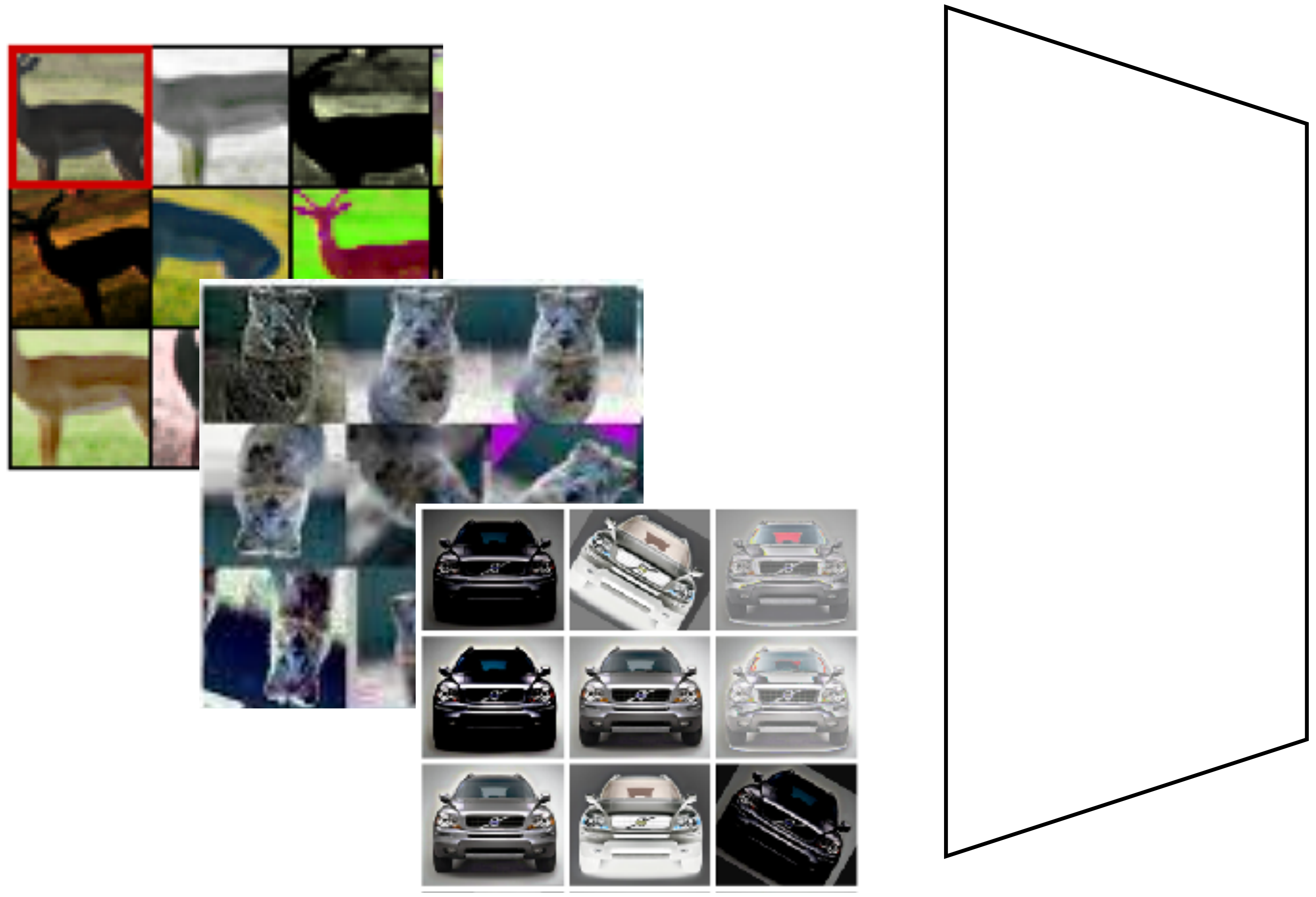
Geometry: RotNet: learn features by predicting “which way is up”.

Train model to predict rotation degree

But:

Unsupervised Representation Learning by Predicting Image Rotations. Gidaris et al., ICLR 2018

Image-uniqueness: Exemplar CNN, precursor to contrastive learning



→ Class 1

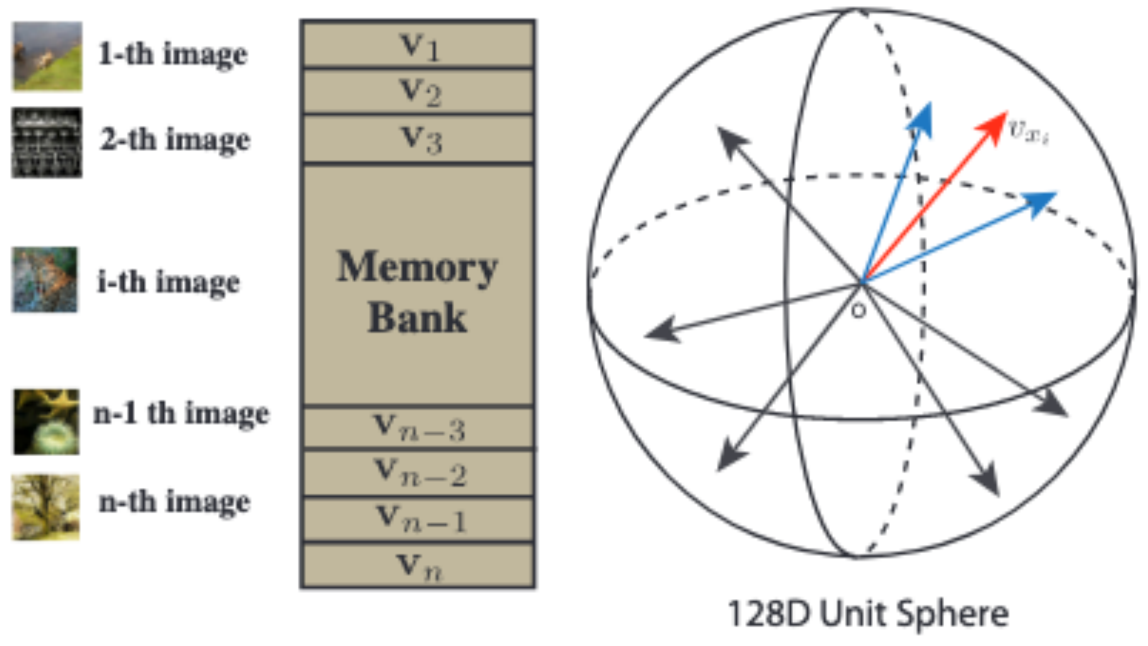
→ Class k

→ Class n

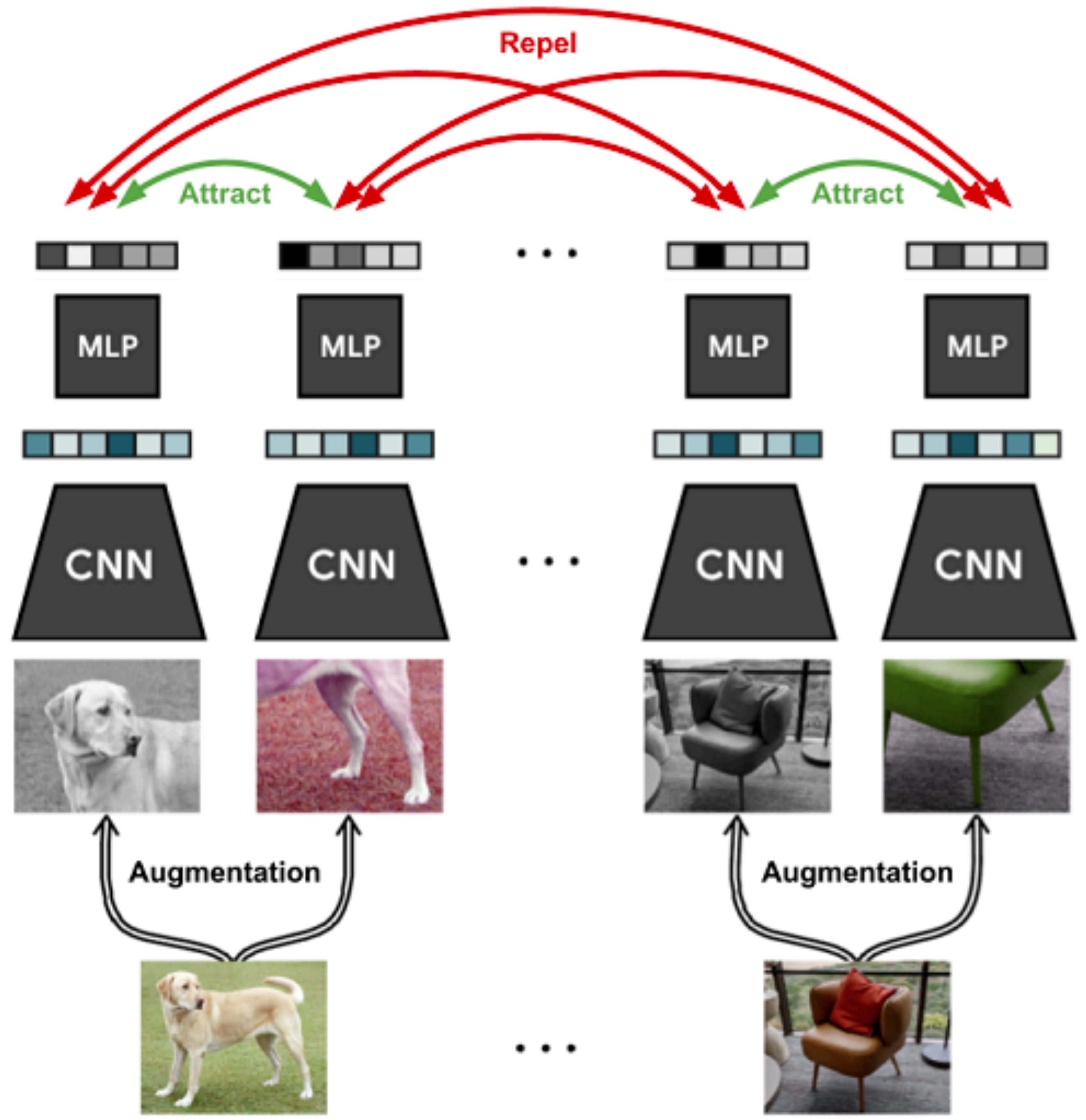
Uses image-uniqueness and *enforces* augmentation-invariance

Dosovitskiy et al. Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. NeurIPS 2014.

Modern Noise-contrastive self-supervised learning



NPID



SimCLR

The contrastive loss for positive pairs i,j:

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

with z_i, z_j embeddings for images i and j,

τ a temperature, sim() *is the dot-product*

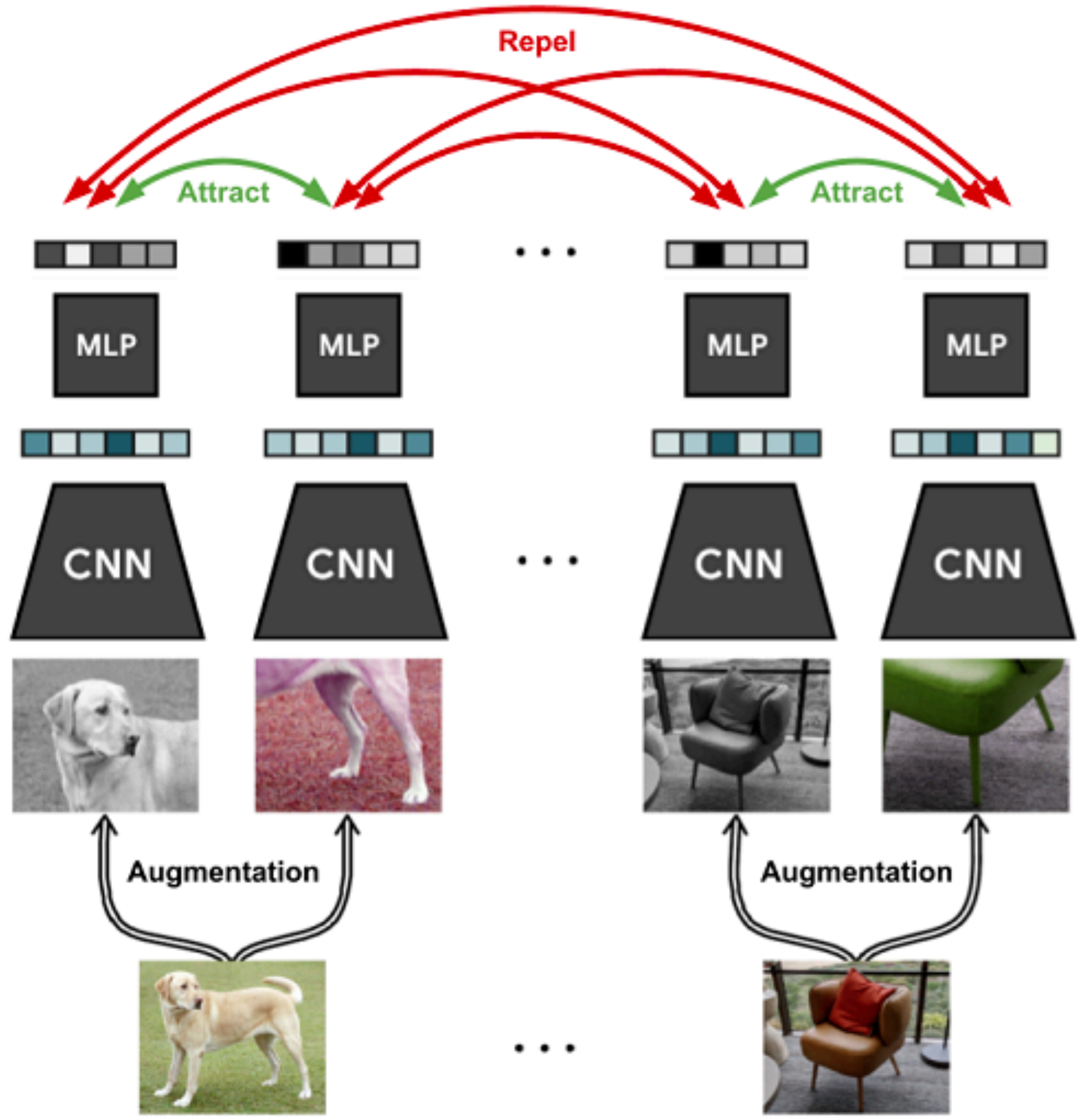
"non-parametric" softmax

Enforces image-uniqueness and

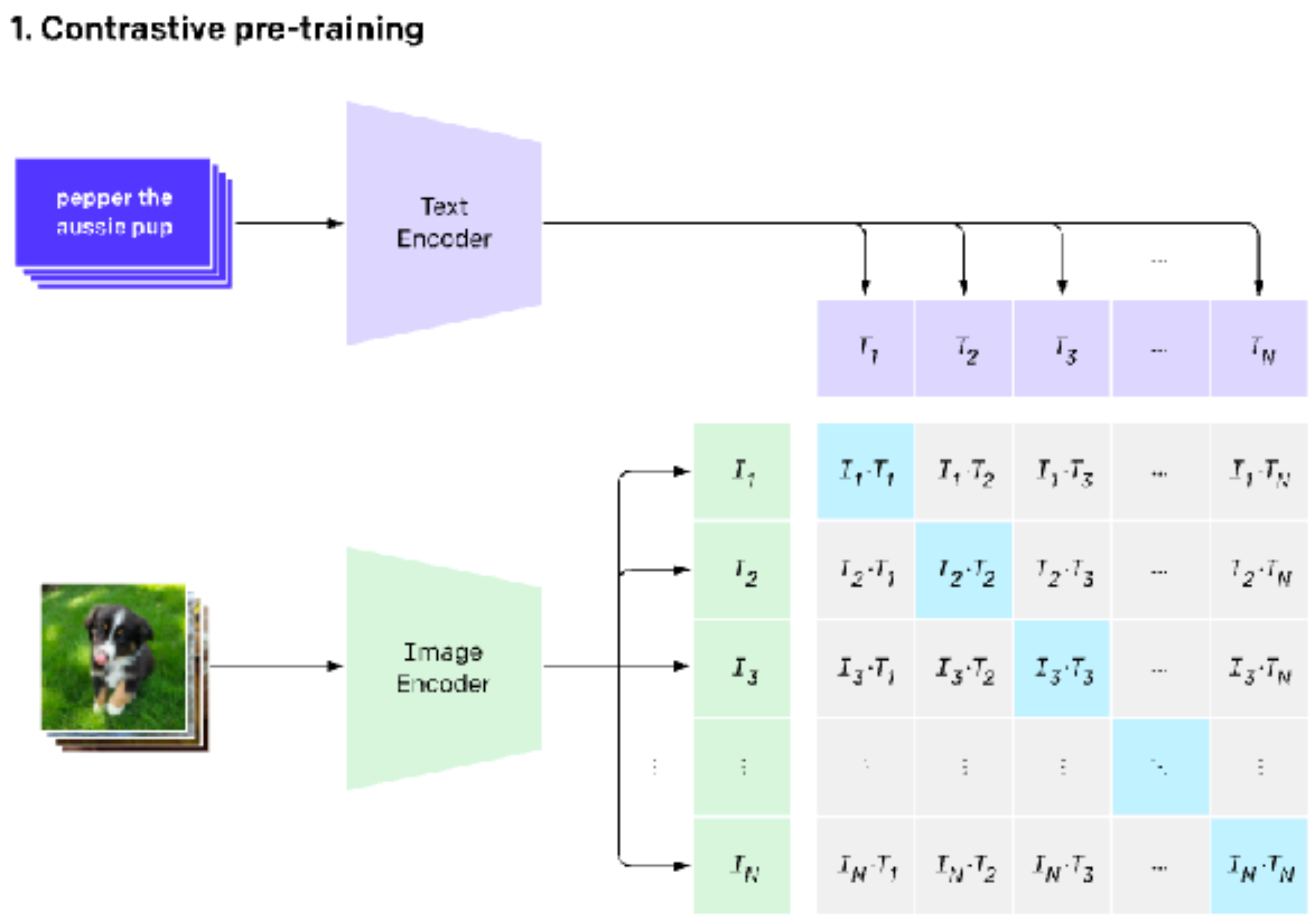
enforces augmentation-invariance (more on that later)

Wu et al. *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination.* CVPR 2018
Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations.* ICML 2020

CLIP from Lect 9 and assigment 2 simply applies SimCLR across modalities

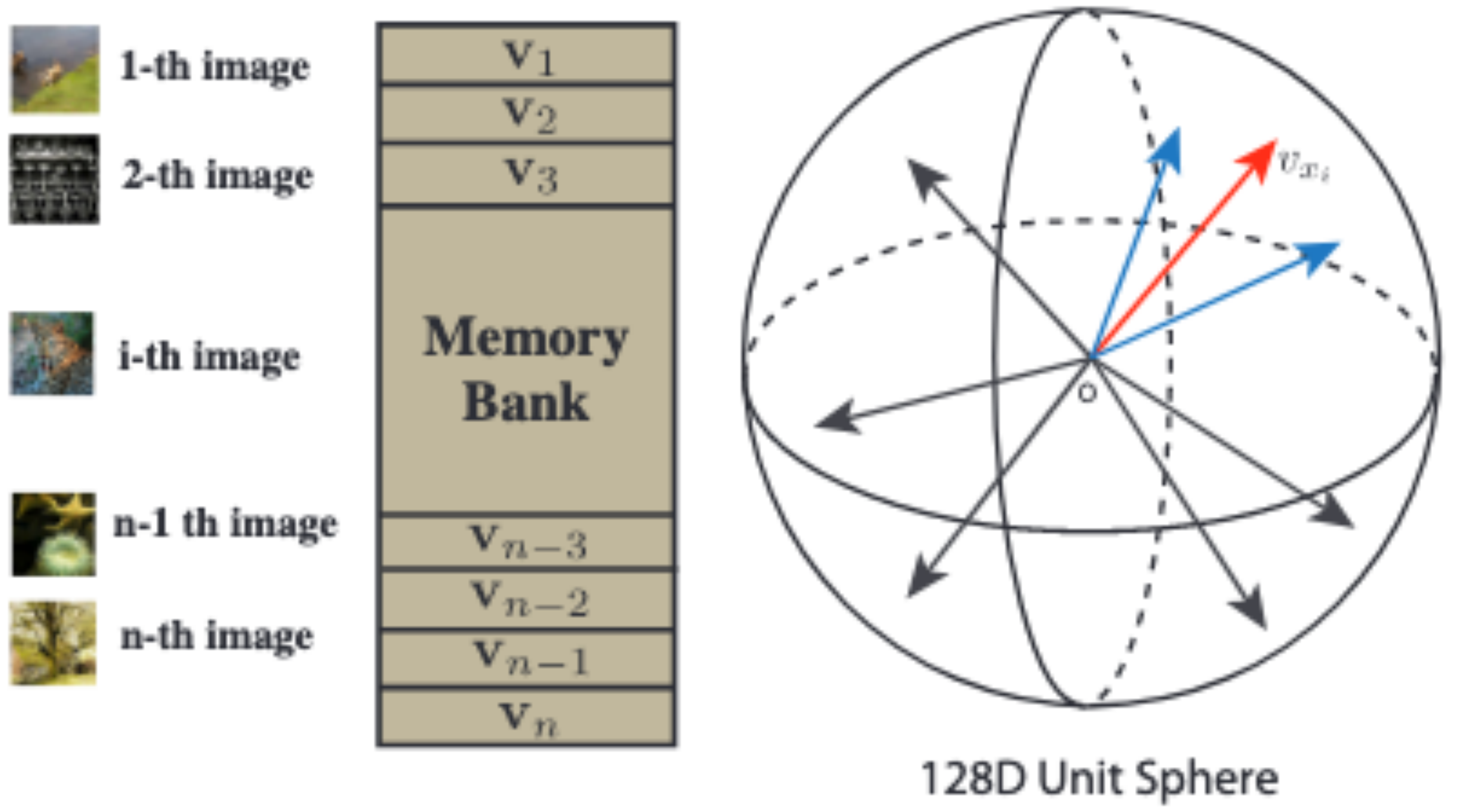


SimCLR

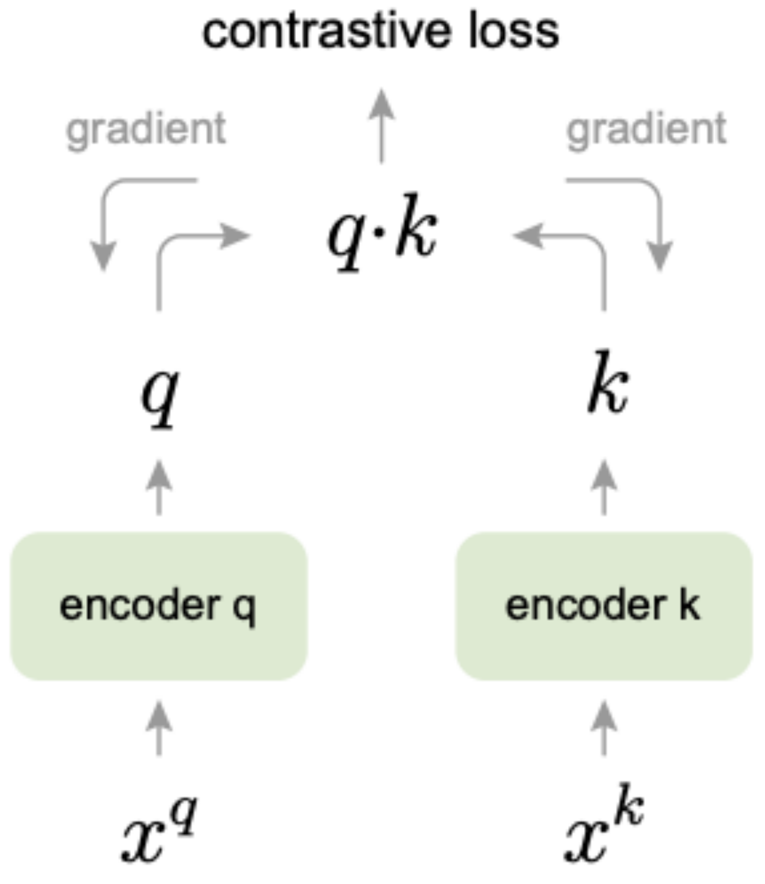


CLIP: instead of augmentation, uses an image caption

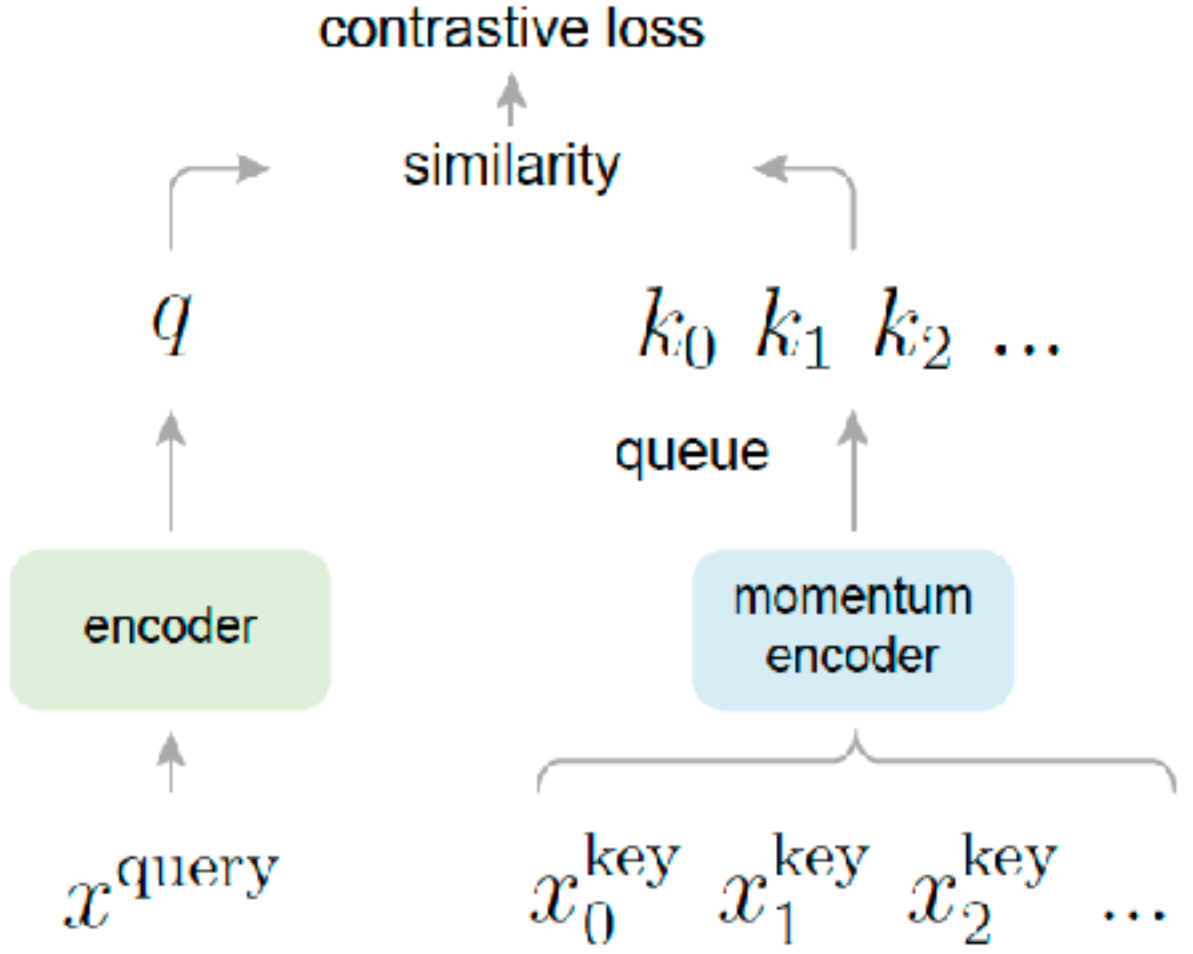
Modern Noise-contrastive self-supervised learning



NPID



SimCLR



MoCo

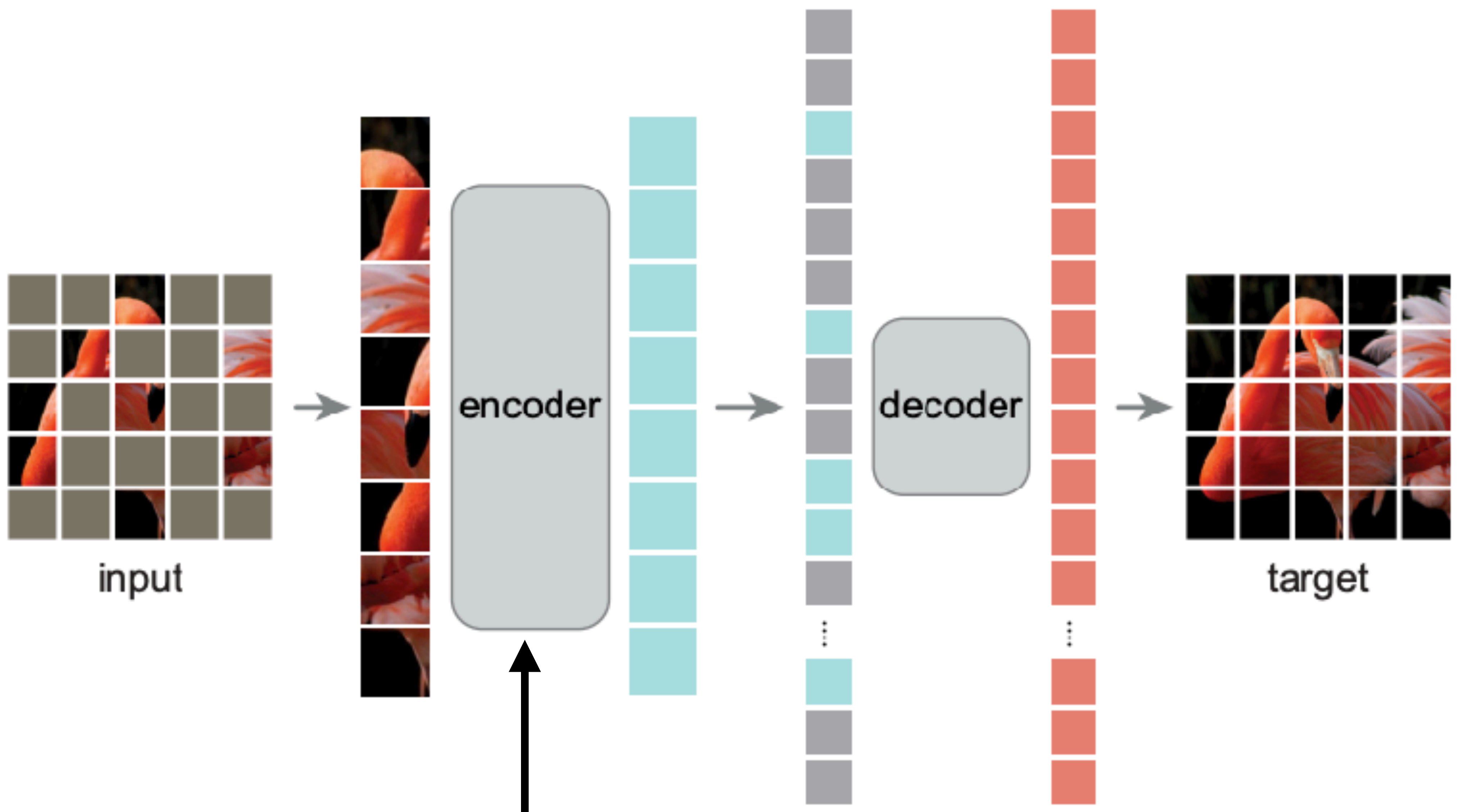
Momentum encoder:

```
# momentum update: key network
f_k.params = m*f_k.params+(1-m)*f_q.params
```

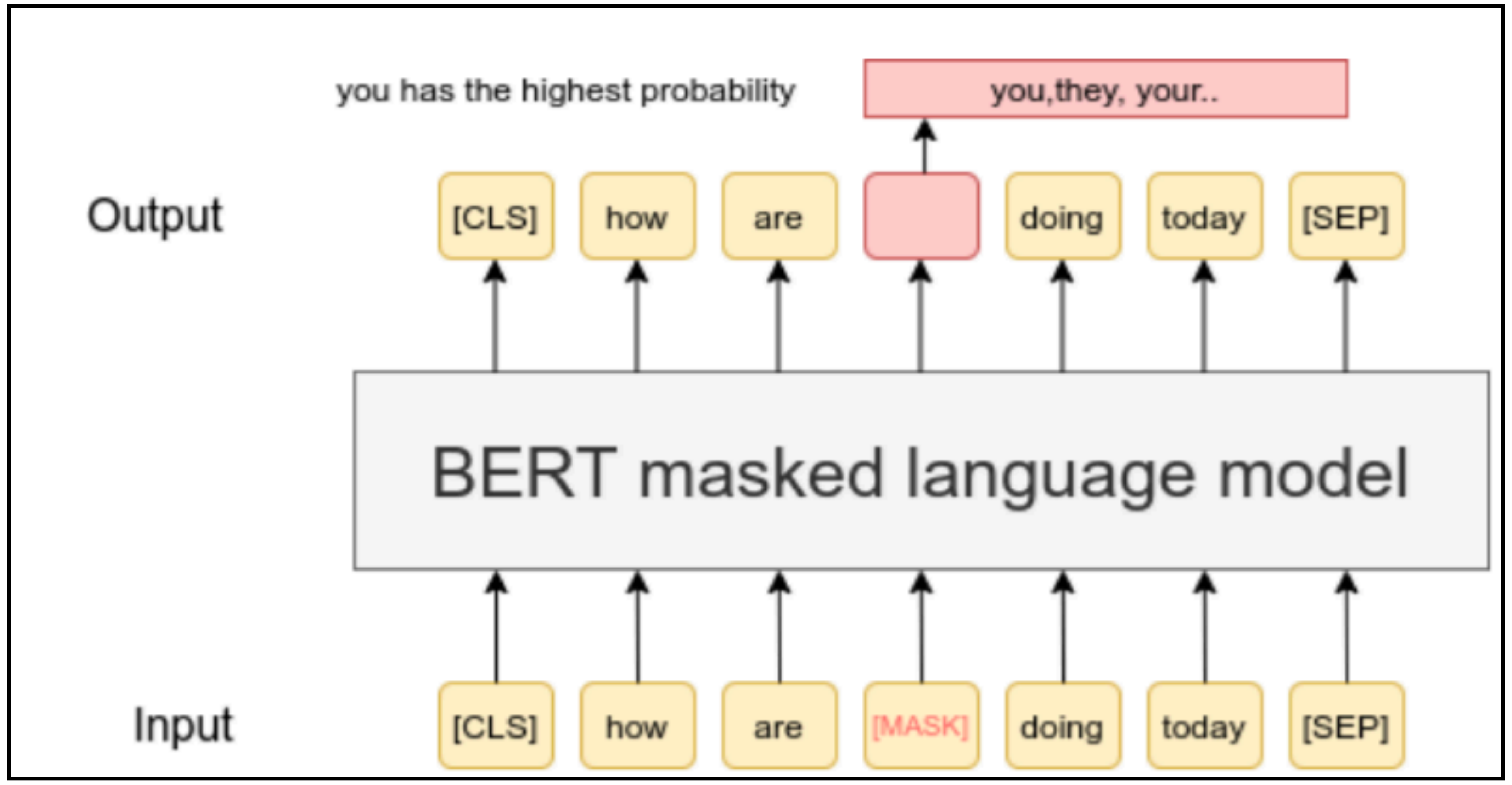
The start of large-scale & industrial self-supervised learning.
These works heavily rely on image augmentations.

Wu et al. *Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination.* CVPR 2018
Chen et al. *A Simple Framework for Contrastive Learning of Visual Representations.* ICML 2020
He et al. *Momentum Contrast for Unsupervised Visual Representation Learning.* CVPR 2020

Masked Image Modelling (recent development)



Vision Transformer

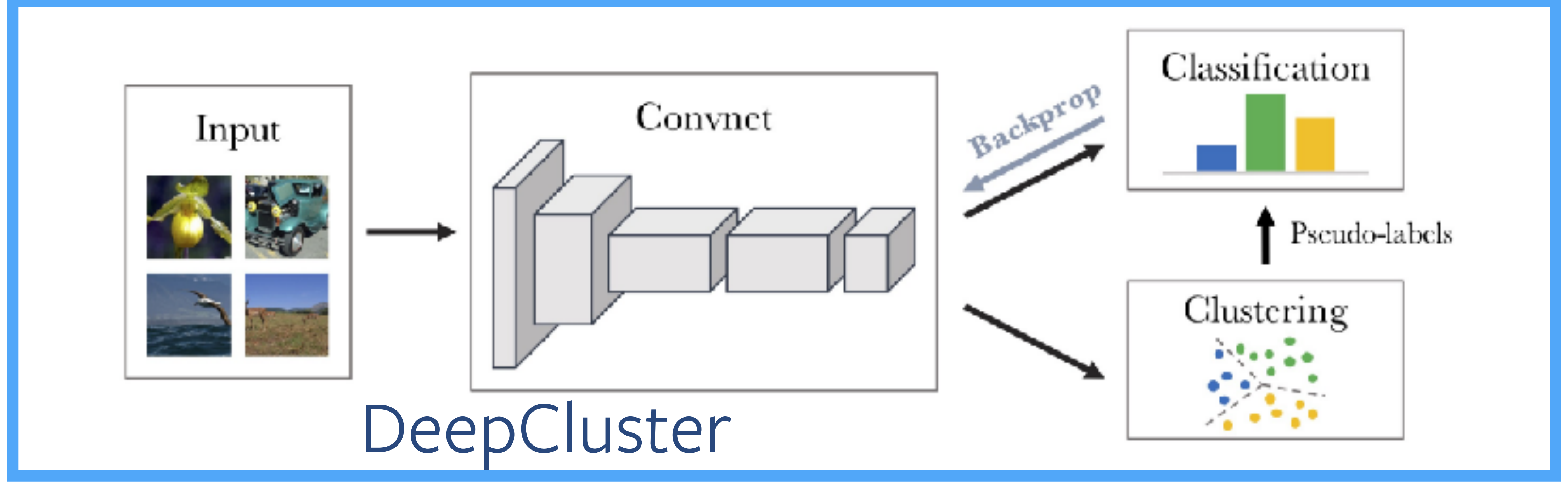


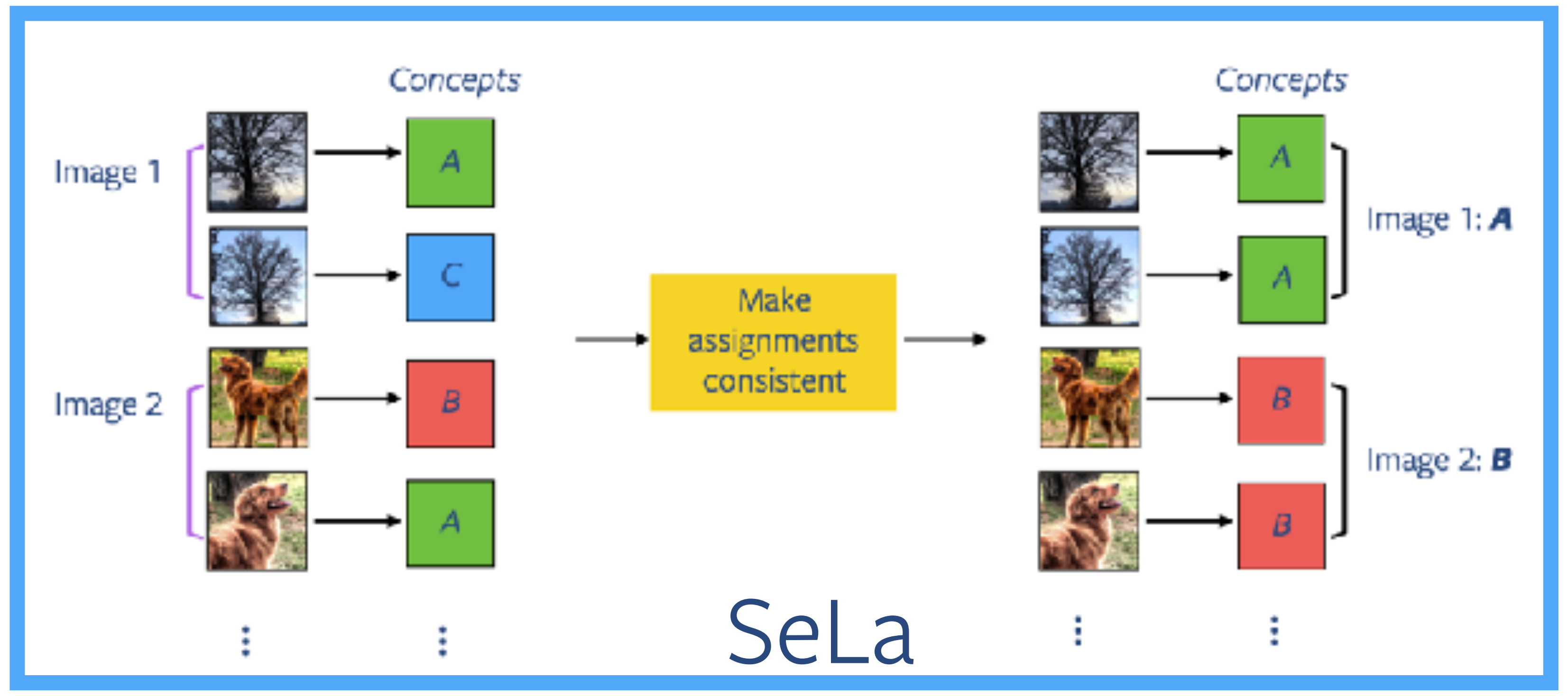
Back to NLP

He et al. *Masked Autoencoders Are Scalable Vision Learners*. CVPR'21
Xie et al. *SimMIM: A Simple Framework for Masked Image Modeling. ArXiv*
Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. ICLR'21
https://www.sbert.net/examples/unsupervised_learning/MLM

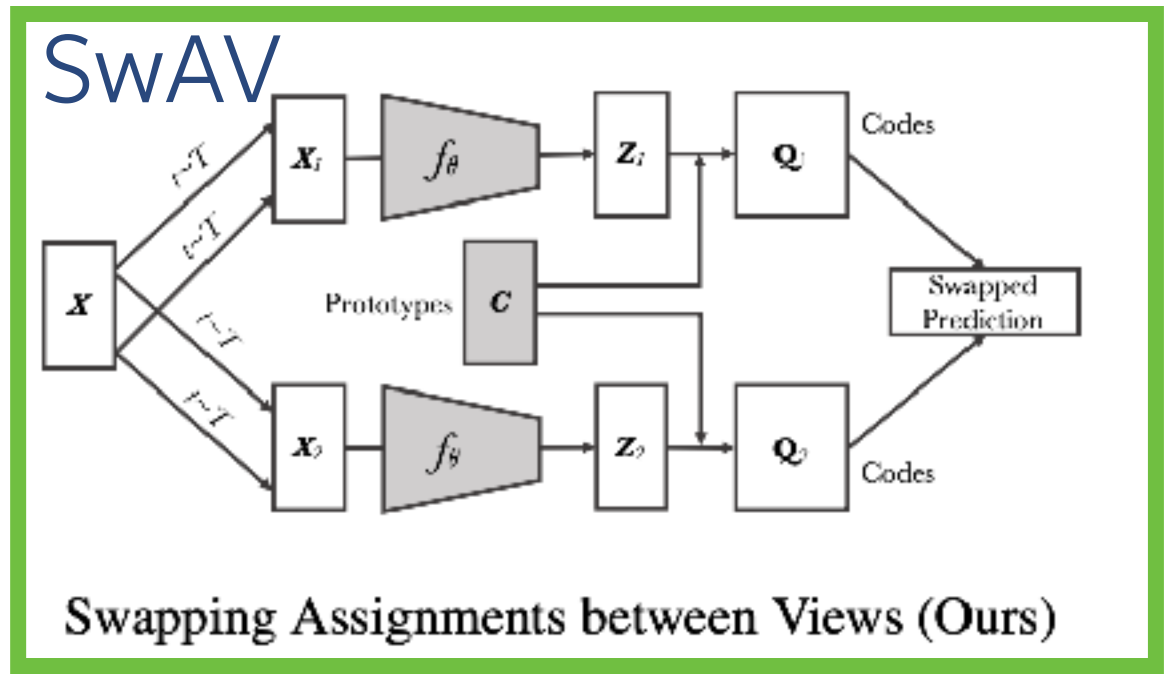
Clustering

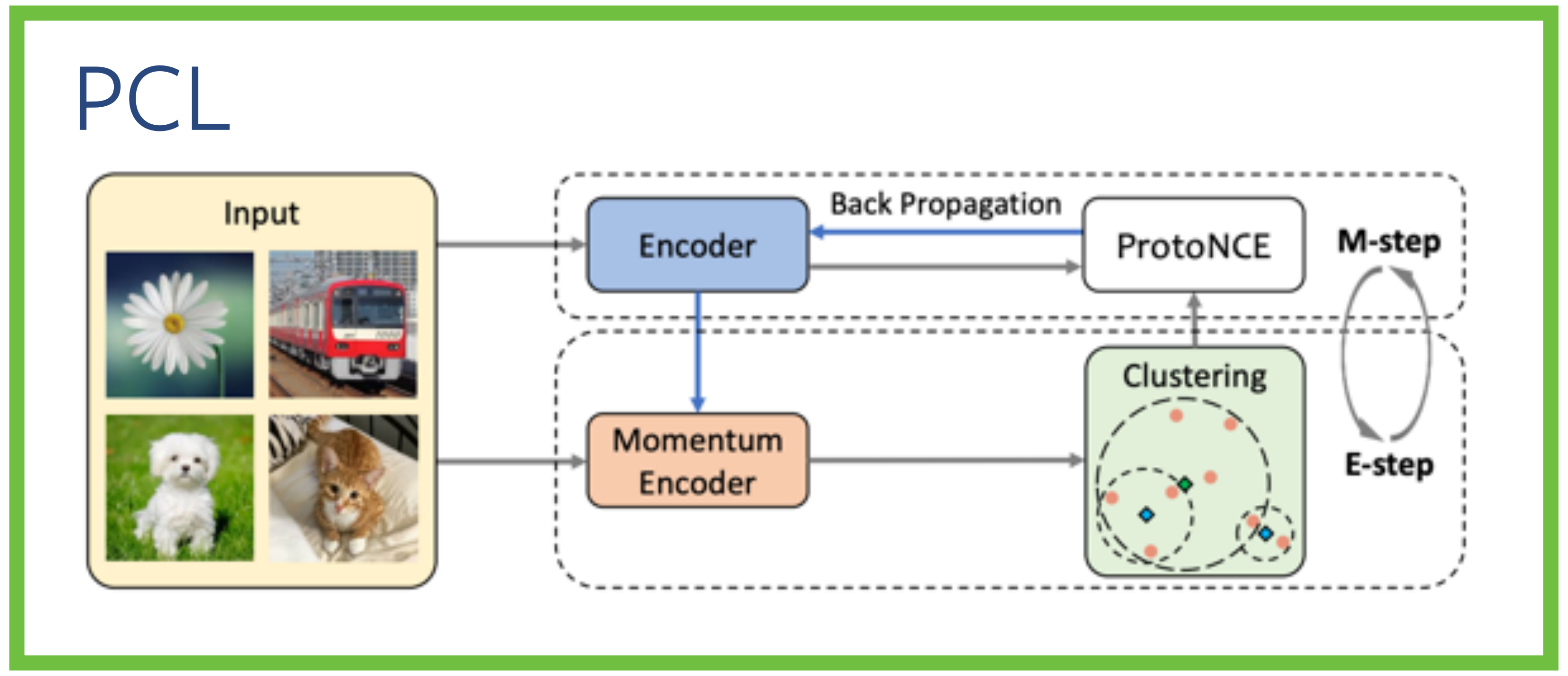
Alternate between clustering and network learning

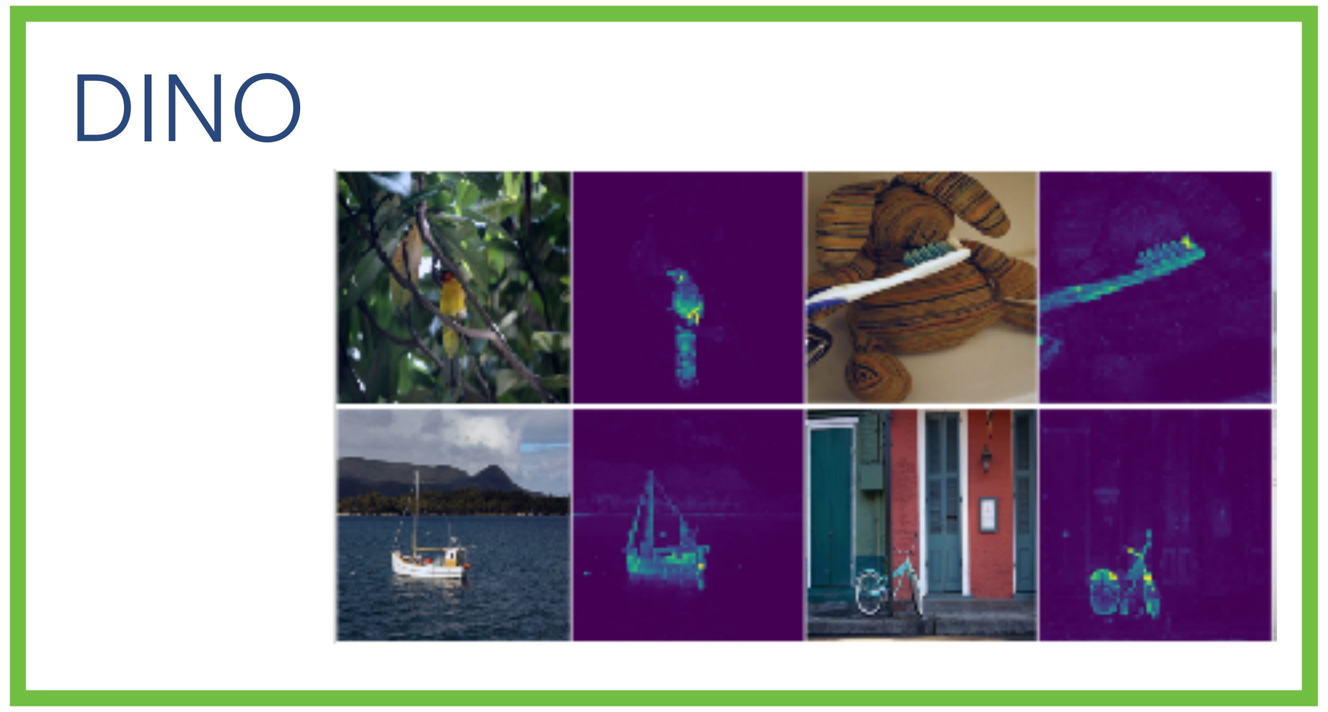




Online & implicit clustering







Clustering is a strong pretext task and serves a useful purpose (~labelling/categorizing).

Caron et al. *Deep Clustering for Unsupervised Learning of Visual Features.* ECCV'18
Asano et al. *Self-labelling via simultaneous clustering and representation learning.* ICLR'19
Caron et al. *Unsupervised Learning of Visual Features by Contrasting Cluster Assignments.* NeurIPS'20
Li et al. *Prototypical Contrastive Learning of Unsupervised Representations.* ICLR'21
Caron et al. *Emerging Properties in Self-Supervised Vision Transformers.* ICCV'21

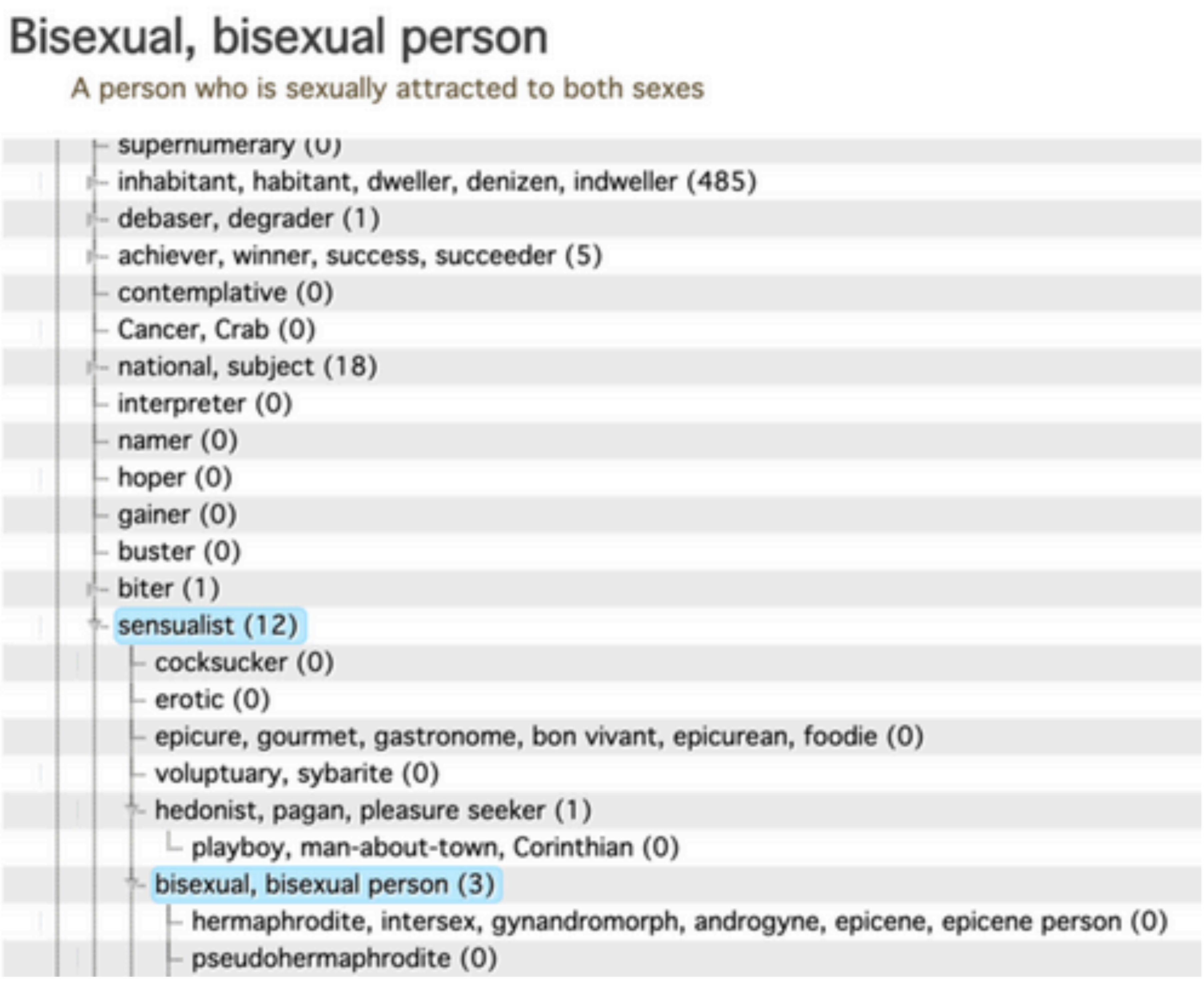
On which datasets are self-supervised methods trained and evaluated?

Datasets for images: Pretraining and downstream

- Class-balanced dataset, via search engine
- Unclear image licences
- Particular choice of classes, e.g. 120 classes of dogs
- Object-centric, stereotypical images
- Many problematic images (see Prabhu & Birhane)

Deng, et al. *ImageNet: A Large-Scale Hierarchical Image Database*. CVPR, 2009.
Prabhu & Birhance. *Large image datasets: A pyrrhic win for computer vision?* FaCCT 2020
Asano et al. *PASS: Pictures without humAns for Self-Supervised Pretraining*. NeurIPS-Data'21

Recent surge in research on problematic images in ImageNet



class_number	label	mean_gender_audit	mean_age_audit	mean_nsfw_train
445	bikini, two-piece	0.18	24.89	0.859
638	maillot	0.18	25.91	0.802
639	maillot, tank suit	0.18	26.67	0.769
655	miniskirt, mini	0.19	29.95	0.62
459	brassiere, bra, bandeau	0.16	25.03	0.61

Table 5: Table of the 5 classes for further investigation that emerged from the NSFW analysis

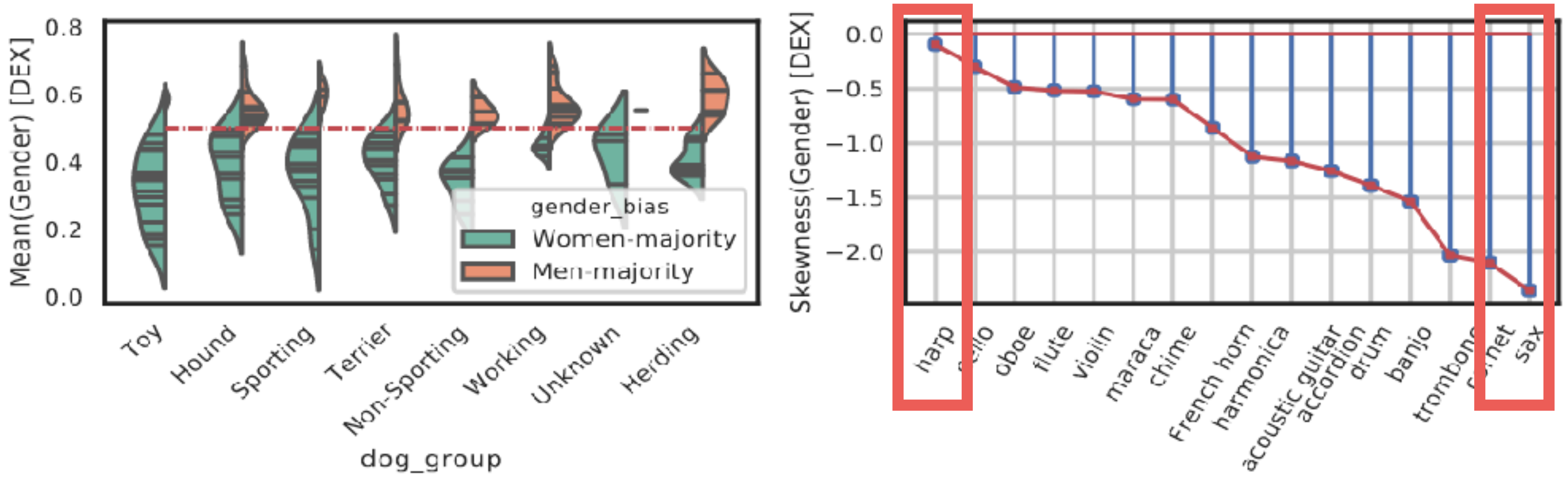


Figure 4: Known *human co-occurrence* based gender-bias analysis

https://excavating.ai/, Crawford and Paglen
Large image datasets: A pyrrhic win for computer vision?, Prabhu & Birhance, FaCCT 2020

Remove all the humans!

Data generation pipeline

YFCC-100M (Thomee et al. 2015) → 17M → 10M → 6M → PASS

Creative Commons 'CC-BY' license only

Download and removal of broken/ single color files

Automated removal of images with faces and humans

Balance image contributions by user

PASS

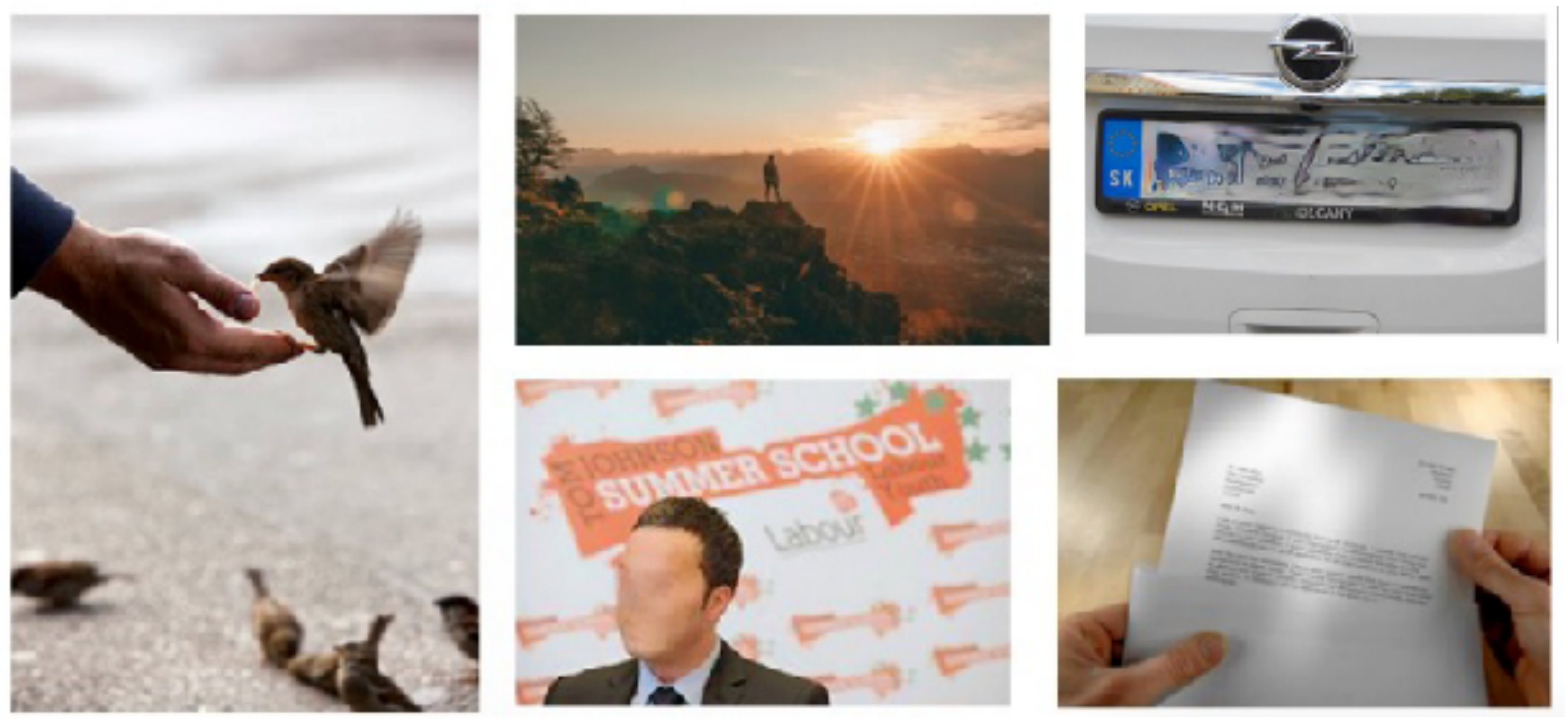
- **1.4M images**
- <80 images per user
- Full-licence files

Human verification: no humans

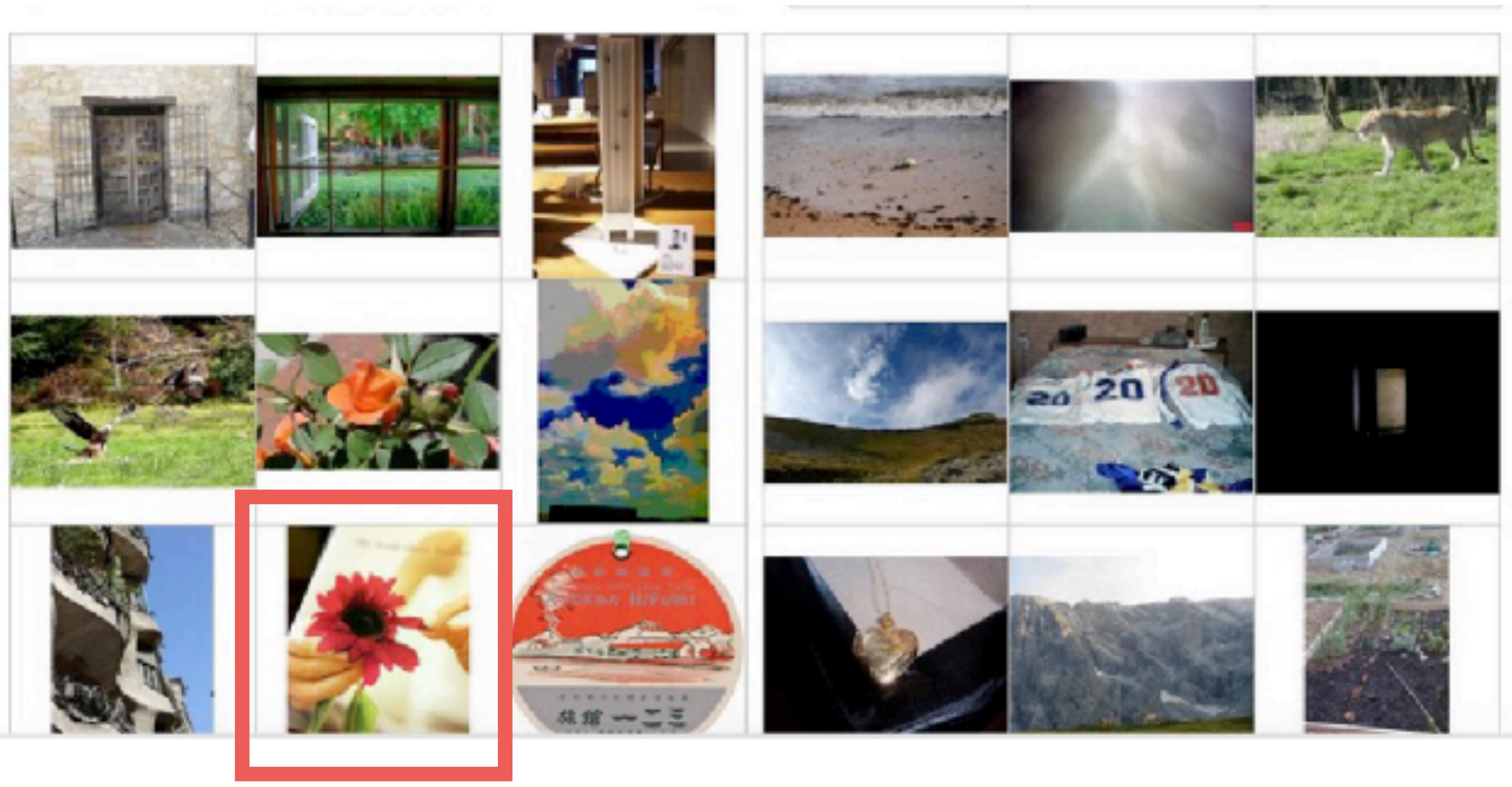
Human verification

Flag all images that contain: people, body parts and personal information (ID, licence plate, names etc.)

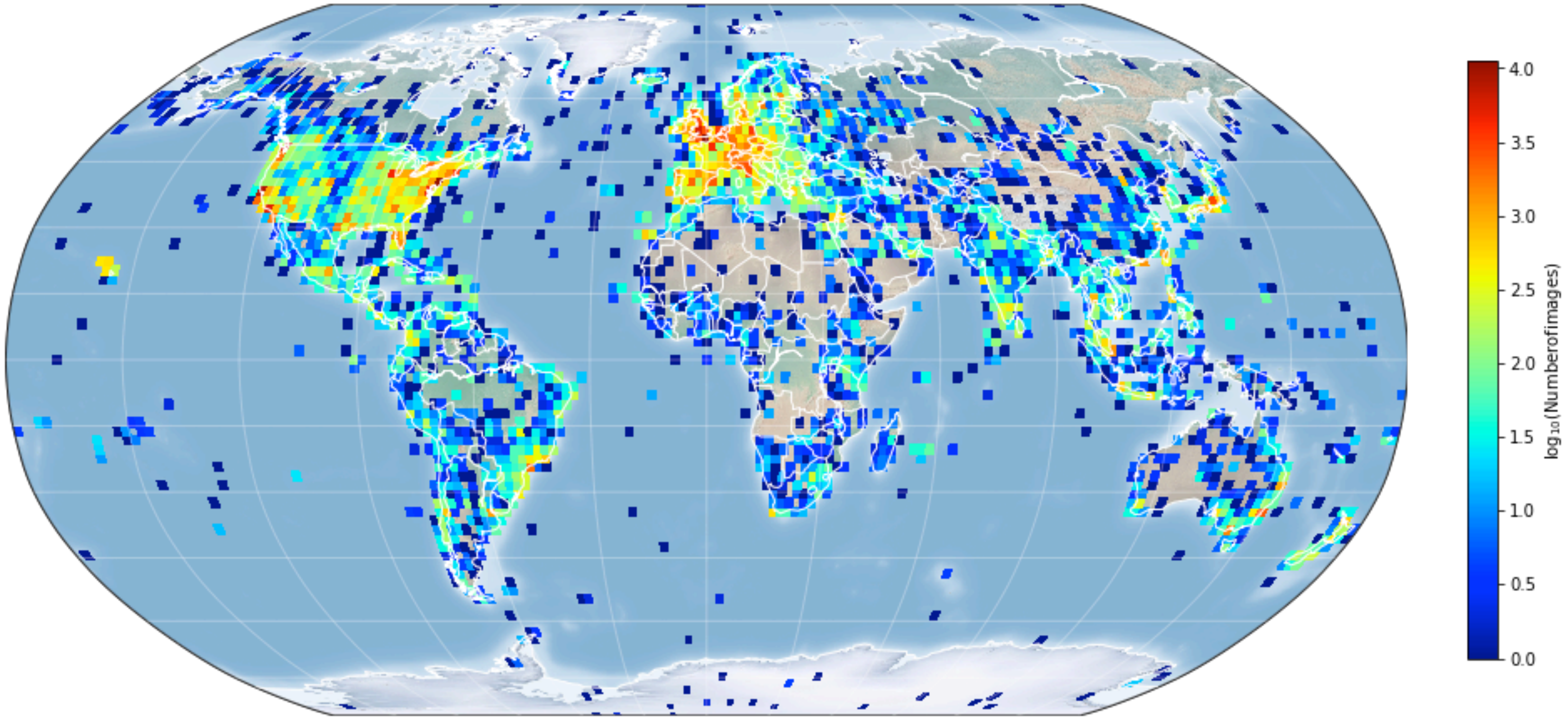
Examples:



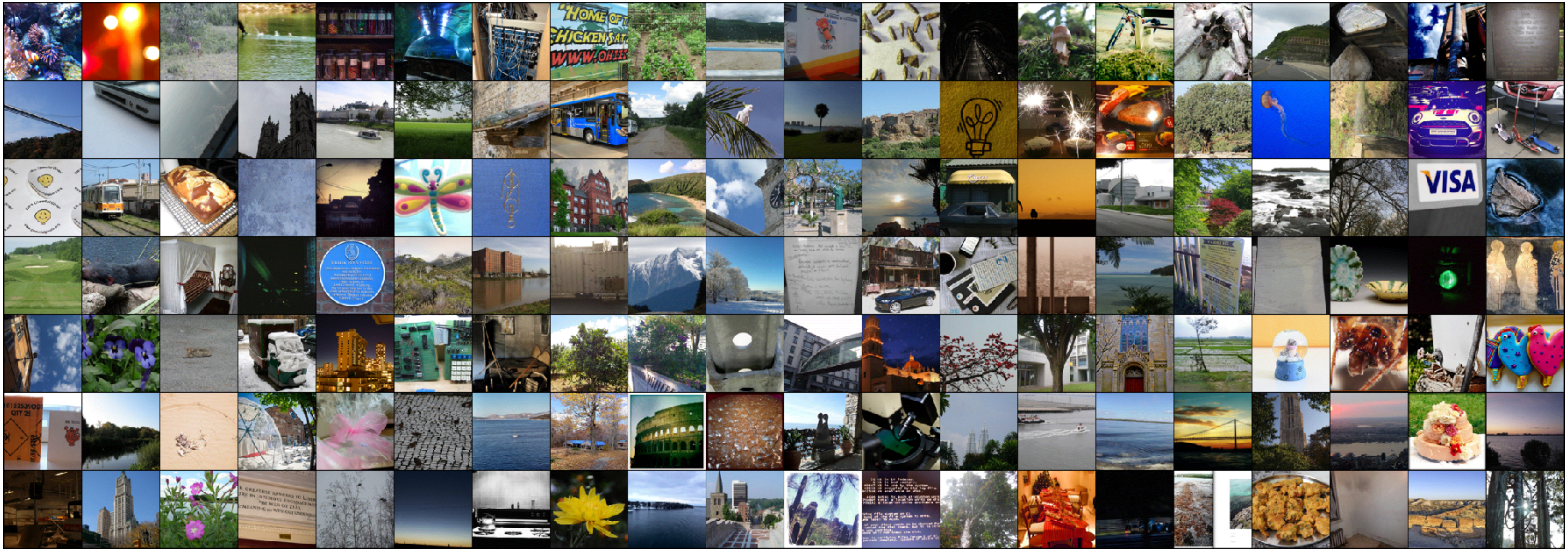
Real data:



The dataset: 30% of images contain location meta-data.



The dataset: diverse, containing nature and buildings.



Datasets for images: Pretraining and downstream

- Class-balanced dataset, via search engine
- Unclear image licences
- Particular choice of classes, e.g. 120 classes of dogs
- Object-centric, stereotypical images
- Many problematic images (see Prabhu & Birhane)

- Random images from YFCC-100M
- All images with complete CC-BY licences
- No people, nor identifiable information
- Natural images as humans take them
- Likely a better indicator for billions-level pretraining

Clustering (❄️)

IN-1k
ObjectNet
Places205
Flowers

SVM low-shot (❄️+🔥)

Places205
Pascal VOC
Herbarium-19

Linear probing (❄️+🔥)

IN-1k
Places205
CIFAR-100
Flowers
…

Finetuning (🔥)

MS-COCO:
detection, segmentation, key point detection, dense pose estimation

Pascal VOC:
detection

LVIS v1.0:
detection

University of Amsterdam

Deng, et al. *ImageNet: A Large-Scale Hierarchical Image Database*. CVPR, 2009.
Prabhu & Birhance. *Large image datasets: A pyrrhic win for computer vision?* FaCCT 2020
Asano et al. *PASS: Pictures without humAns for Self-Supervised Pretraining*. NeurIPS-Data'21

Downstream semi-supervised tasks: Self-supervised Learning helps

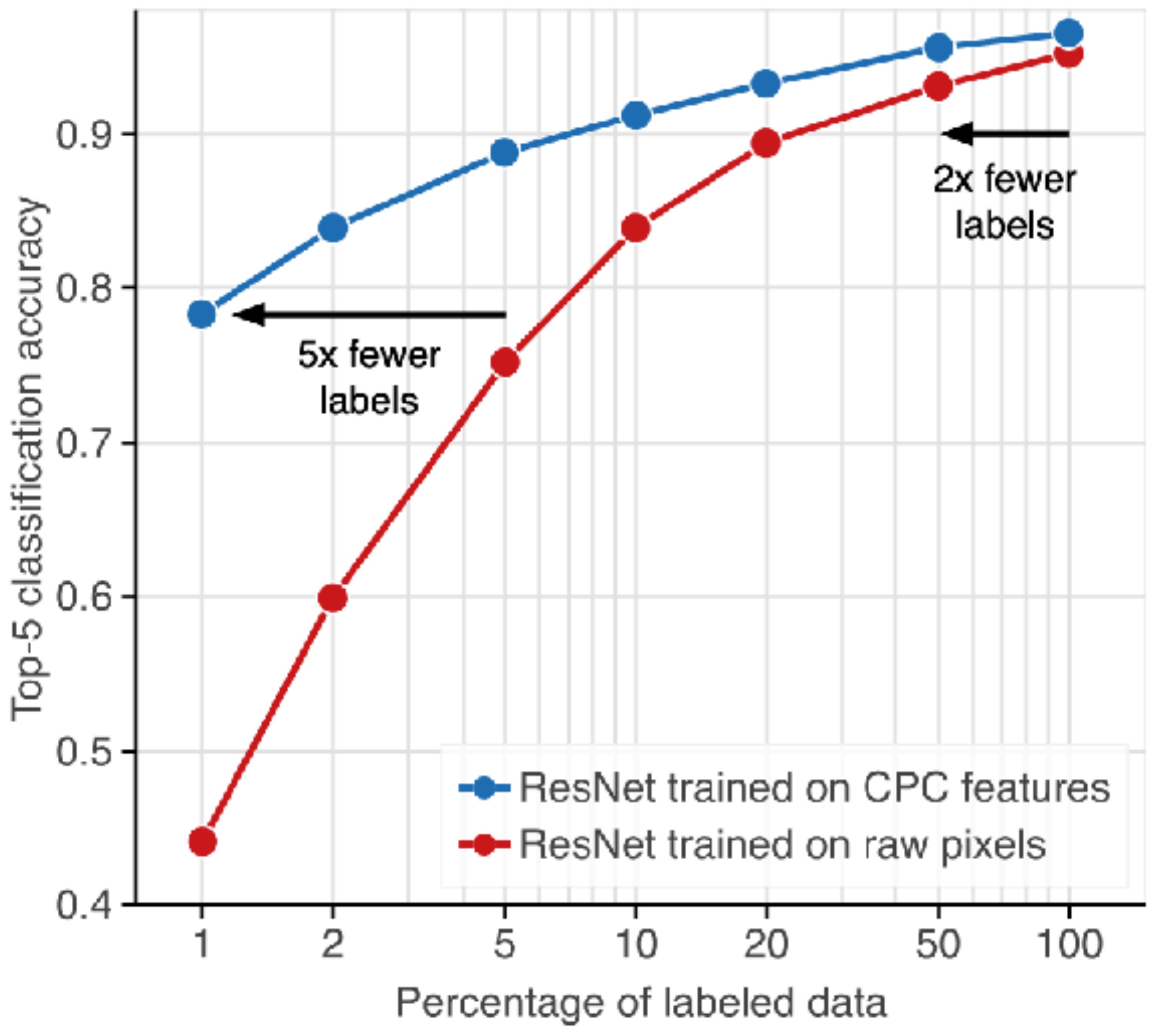


Figure 1. Data-efficient image recognition with Contrastive Predictive Coding. With decreasing amounts of labeled data, supervised networks trained on pixels fail to generalize (red). When trained on unsupervised representations learned with CPC, these networks retain a much higher accuracy in this low-data regime (blue). Equivalently, the accuracy of supervised networks can be matched with significantly fewer labels (horizontal arrows).

Once pretrained, self-supervised networks good for quick transfer learning even with few labels

Achieves much better performance for low number of annotated data

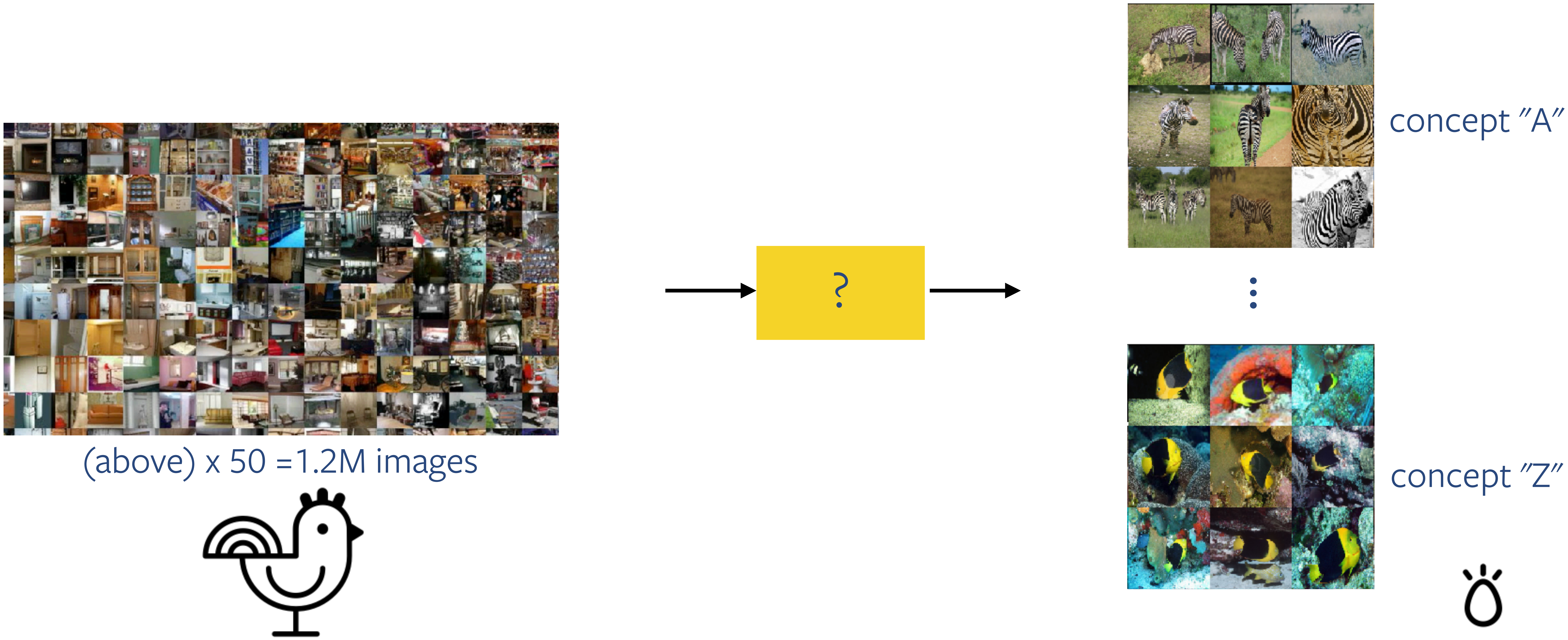
This is the case if you were to found a startup and tackle a new problem (annotation=expensive)

Self-supervised learning using optimal-transport based clustering

Self-labelling via simultaneous clustering and representation learning (ICLR'20 spotlight)

YUKI M. ASANO, CHRISTIAN RUPPRECHT, ANDREA VEDALDI

Goal: Discover visual concepts without annotations.



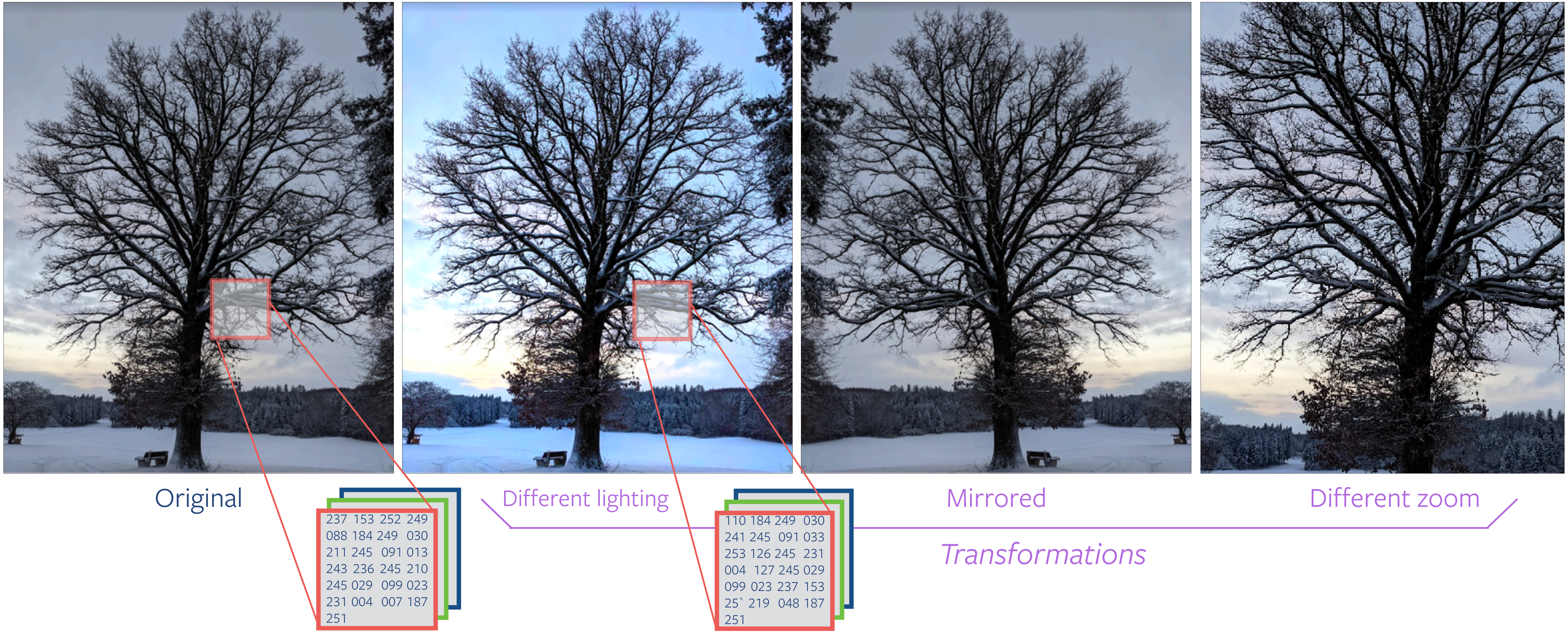
ImageNet: A Large-Scale Hierarchical Image Database. Deng, et al. CVPR, 2009.

How can we solve this chicken and egg problem?

Augmentations!

(I have no idea where this gif is from)

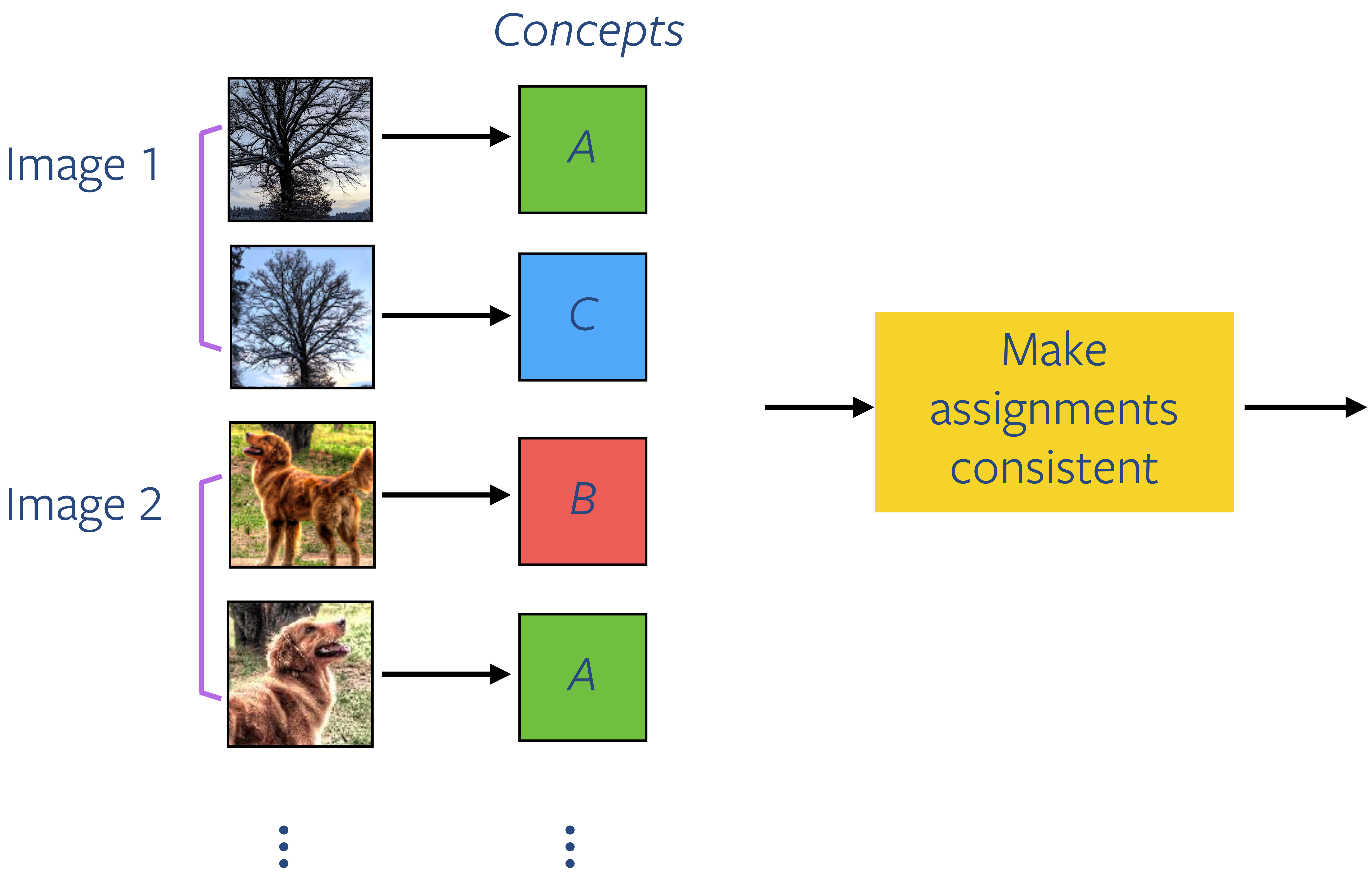
The key to image understanding is separating meaning from appearance.



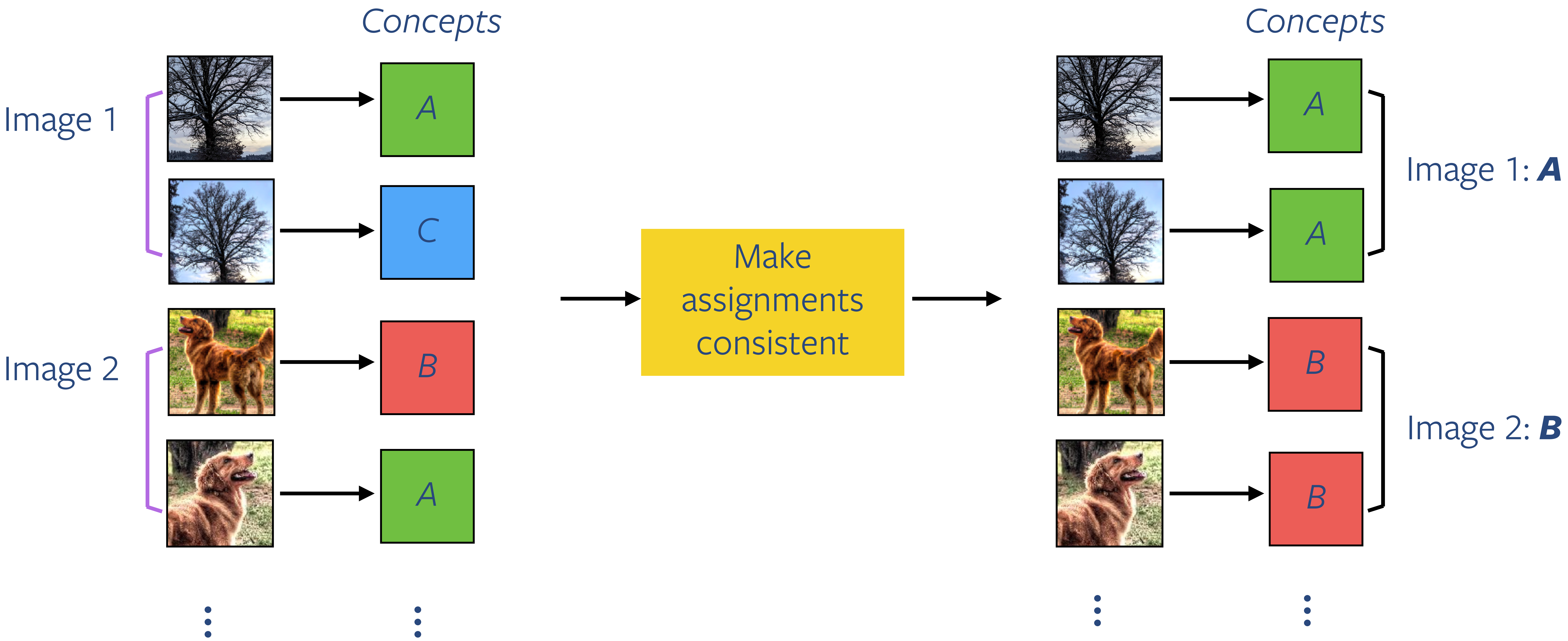
Quiz: What other ways of incorporating prior knowledge have we already learned about? (MC)

1) Choosing the right learning rate
2) Setting the network architecture
3) Picking the optimizer
4) Choosing the number of epochs
5) Choosing the loss

Our work applies the idea of augmentation invariance to assign concepts.



Our work applies the idea of transformation invariance to assign concepts.



How can we optimize the labels and make assignments consistent?

If we had ground-truth labels

$$\min_{\Phi} L(y, \Phi),$$

where $L(y, \Phi) = \frac{1}{N}\sum_{i=1}^{N} \log p(y_i \mid \mathbf{x}_i, \Phi)$

- L is the loss (cost) function
- Φ is the deep neural network model
- y are the labels

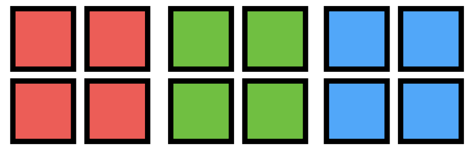
Our novel contribution *without* ground-truth

Solution sketch:

1. Represent via an assignment table q and optimize:

$$L(q, \Phi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{y} q(y \mid \mathbf{x}_i) \log p(y \mid \mathbf{x}_i, \Phi)$$

But: The trivial solution for q is to set all labels to be the same

2. Use pseudolabels an equal number of times: 

3. Pose as approximate optimal transport:

$$\min_{q,\Phi} L(q, \Phi) \quad \text{s.t.} \quad \sum_{i=1}^{N} q(y \mid \mathbf{x}_i) = \frac{N}{K},$$

Self-labelling via simultaneous clustering and representation learning. Asano et al. ICLR 2020
Sinkhorn distances: Lightspeed computation of optimal transport. Cuturi. NeurIPS 2013

SK optimisation (not needed for exam)

$$\min_{P\in U} F(P) = \min_{P\in U}\Big[\langle Q, -\log P\rangle - \lambda h(P)\Big]$$

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}p_{ij}}F \\
&= \frac{\mathrm{d}}{\mathrm{d}p_{ij}}\Big[\sum_{ij} Q_{ij}P_{ij} + \lambda P_{ij}\log(P_{ij}) + \sum_i \alpha_i(\sum_{ij} P_{ij} - 1) + \sum_j \beta_j(\sum_{ij} P_{ij} - 1)\Big] \\
&= Q_{ij} + \lambda \log(P_{ij}) + \lambda + \alpha_i + \beta_j
\end{aligned}$$

Hence

$$\begin{aligned}
P_{ij} &= \exp\big(-\lambda^{-1}\alpha_i - \lambda^{-1}Q_{ij} - 1 - \lambda^{-1}\beta_j\big) \\
&= u_i e^{-\lambda^{-1}Q_{ij}} v_j \quad = u_i e^{\lambda^{-1}log(q)} v_j
\end{aligned}$$

SK optimisation of assignments Q (not needed for exam)

$$\min_{Q\in U} L = \min_{Q\in U}\left[\langle Q, \underbrace{-\log P}_{C\,\geq\,0,\quad \text{costs}}\rangle - \frac{1}{\lambda}h(Q)\right]$$

using $\quad H(Q) = H(r) + H(c) - D_{KL}(Q\|rc^{\mathrm{T}}) = \log(NK) - D_{KL}(Q\|rc^{\mathrm{T}})$

$$\min_{Q\in U} L = \min_{Q\in U}\left[\langle Q, C\rangle + \frac{1}{\lambda}D_{KL}(Q\|rc^{\mathrm{T}})\right] + \text{const}\,.$$

Find minimum:

$$0 = \frac{\mathrm{d}}{\mathrm{d}q_{ij}}F \quad = \frac{\mathrm{d}}{\mathrm{d}q_{ij}}\left[\sum_{ij} Q_{ij}C_{ij} + \frac{1}{\lambda}Q_{ij}\log(Q_{ij}) + \sum_i \alpha_i(\sum_{ij} Q_{ij} - 1) + \sum_j \beta_j(\sum_{ij} Q_{ij} - 1)\right]$$

$$= C_{ij} + \frac{1}{\lambda}\log(Q_{ij}) + \lambda + \alpha_i + \beta_j$$

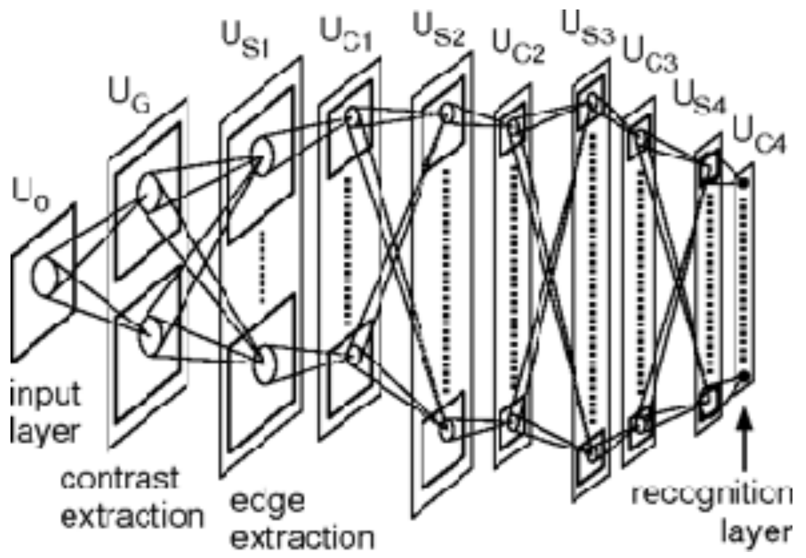
Hence:

$$Q_{ij} = \exp\left(-\lambda\alpha_i - \lambda C_{ij} - 1 - \lambda\beta_j\right)$$

$$= u_i e^{-\lambda C_{ij}} v_j = u_i e^{\lambda \log(p)} v_j = u_i p^{\lambda} v_j$$

Algorithm

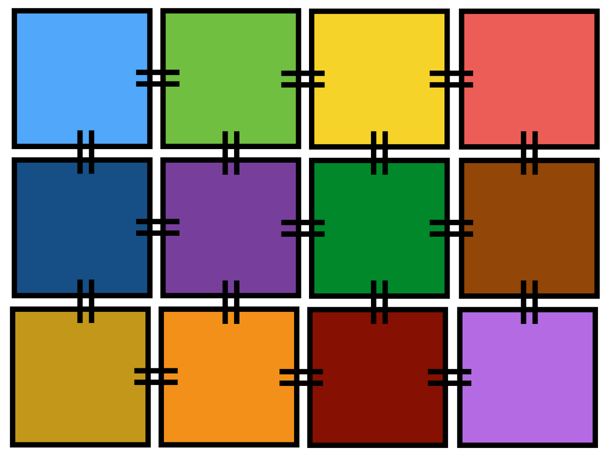
Model Φ



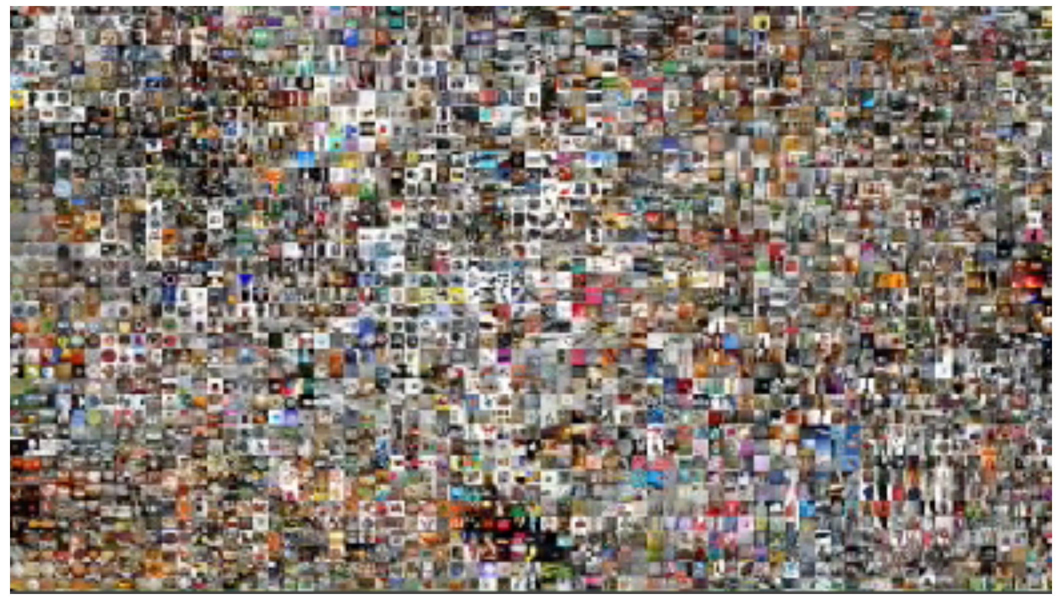
Optimal labelling

Cross entropy training
with augmentations

Label assignments q

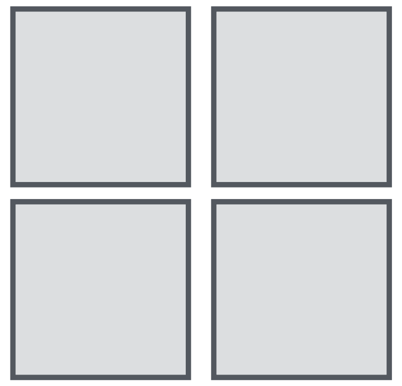


Our method applied on 1.2 million images: Examples



1.2M images

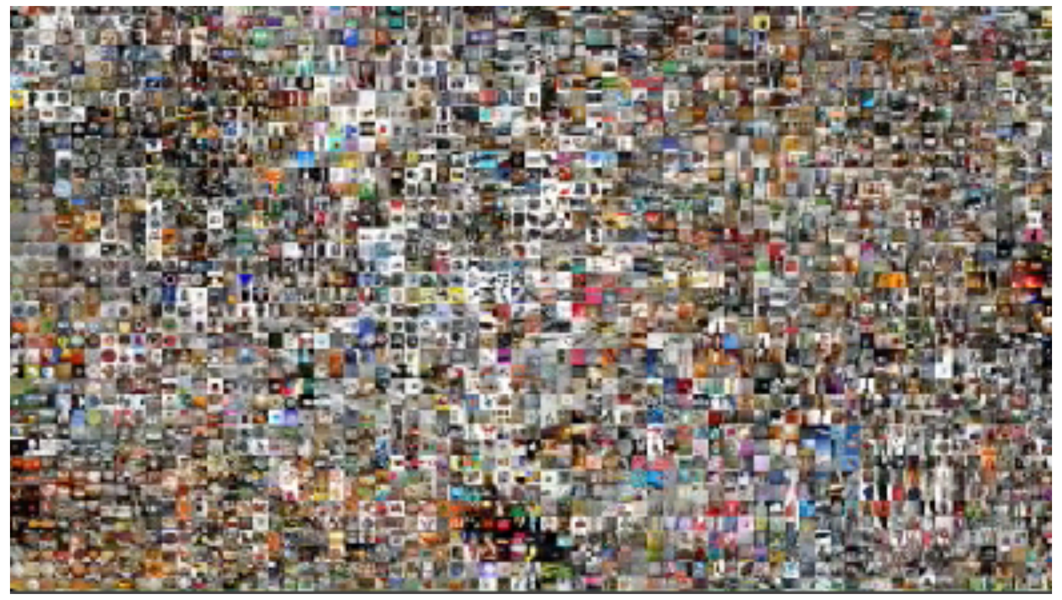
Legend:



Concept

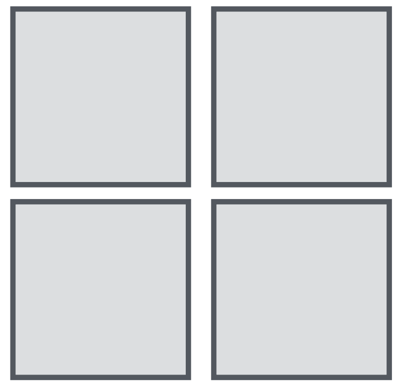
Self-labelling via simultaneous clustering and representation learning. Asano et al. ICLR 2020

Automatically discovered concepts match manual annotation.



1.2M images

Legend:



Concept

border color

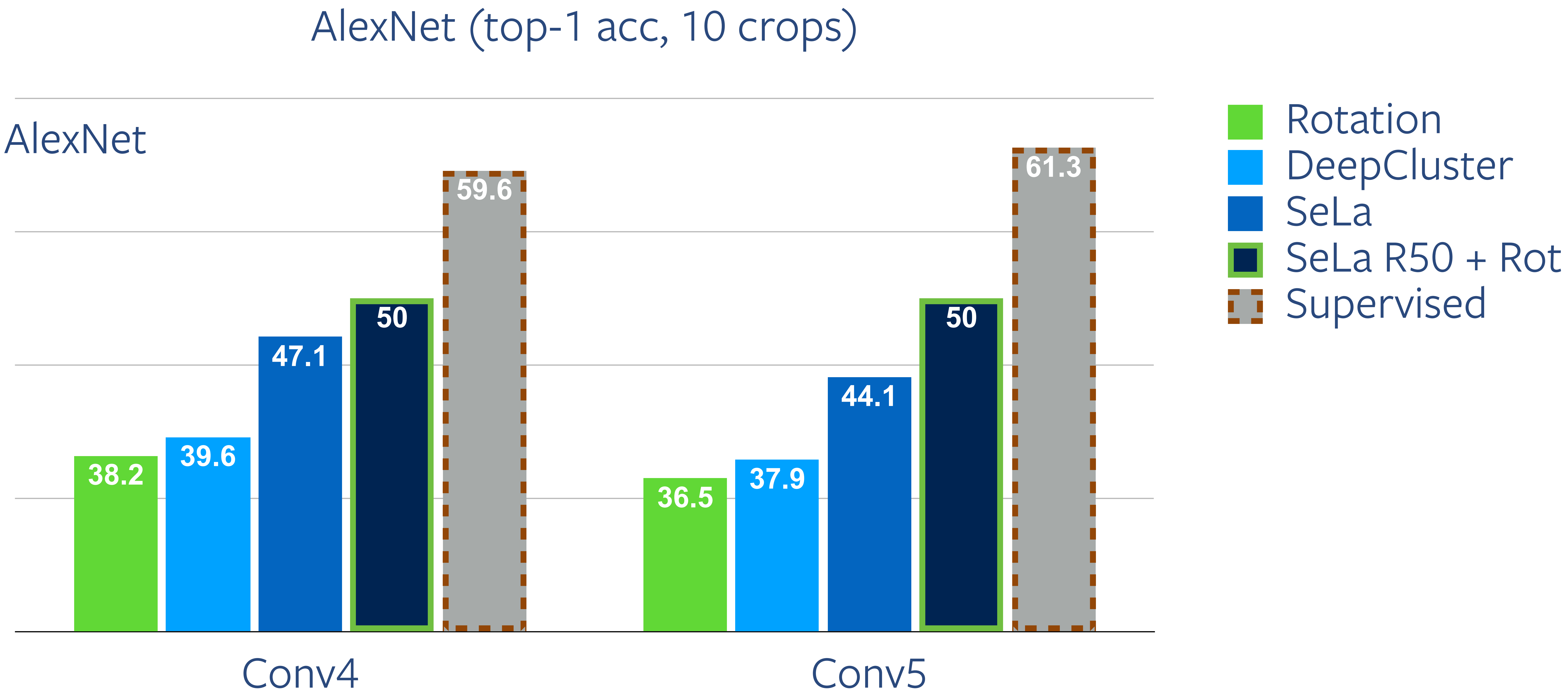
Manually annotated label (>2.5y of work)

Explore all clusters:

Self-labelling via simultaneous clustering and representation learning. Asano et al. ICLR 2020
ImageNet: A Large-Scale Hierarchical Image Database. Deng et al. CVPR, 2009.

AlexNet, ImageNet linear probes (remember Lecture 5)

- Big jump on DeepCluster
- SoTA or close to SoTA for AlexNet



Unsupervised representation learning by predicting image rotations. Gidaris et al. ICLR,2018.
Self-supervised representation learning by rotation feature decoupling. Feng et al. CVPR 2019.

Self-supervised labelling from three core ideas

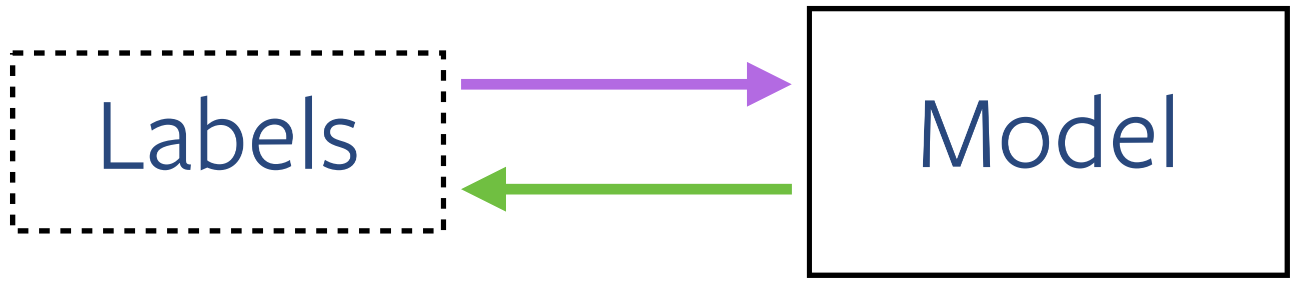
Invariance to augmentations

Original

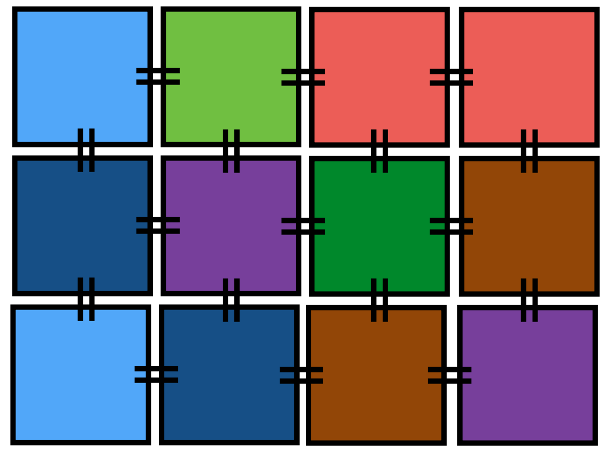
Different lighting

Different zoom

Virtuous cycle of labelling and representation learning



Balanced labelling



(1) Transformations

- Data augmentations “infuse knowledge”

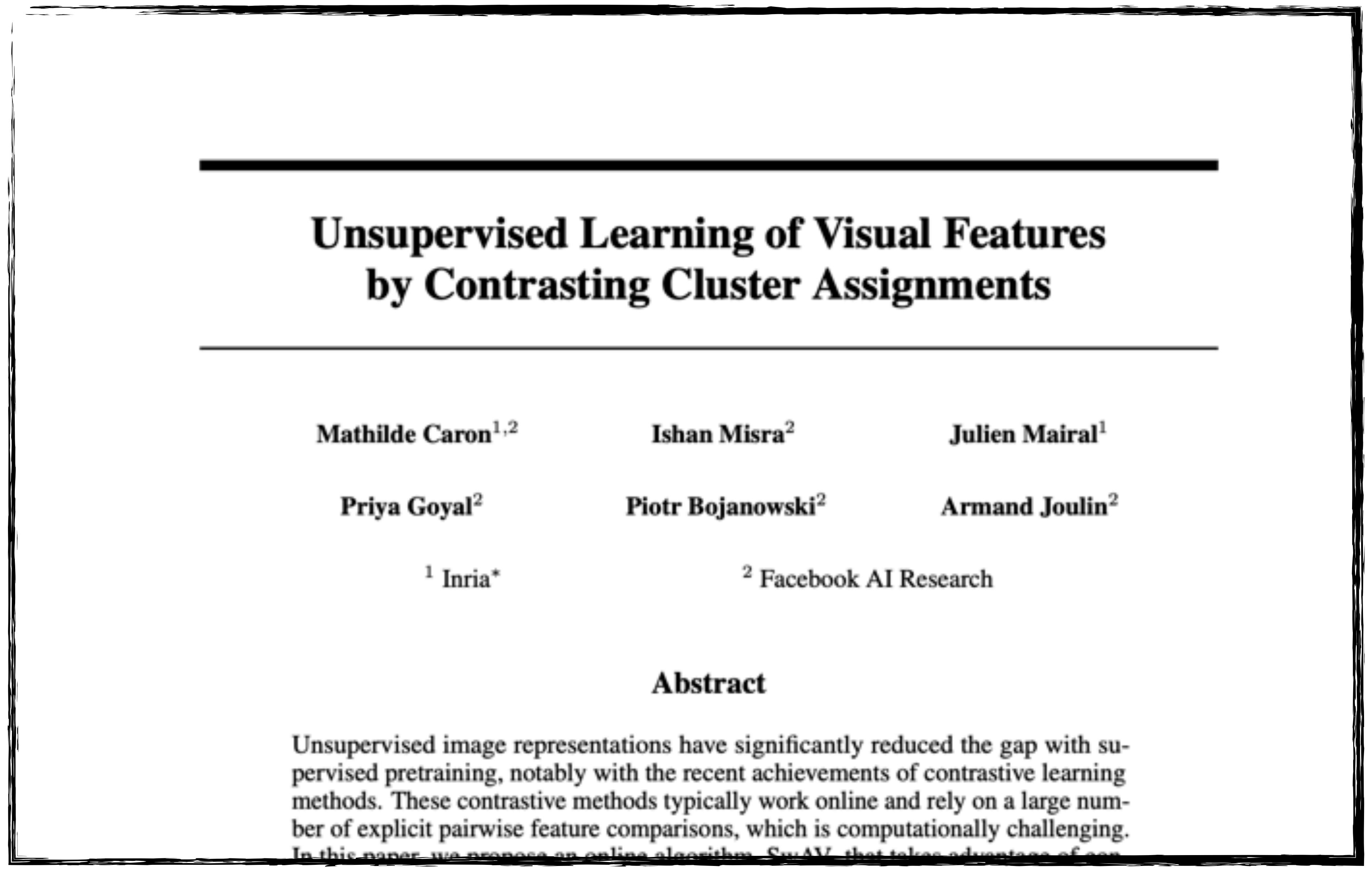
(2) Useful labels

- Labels discovered are similar to ground-truth
- Can be used to analyze how the network “sees” the data

(3) Balanced pseudo-labelling

- Well defined, fast objective
- No trivial solutions

More recently...



Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

Mathilde Caron[1,2] Ishan Misra[2] Julien Mairal[1]

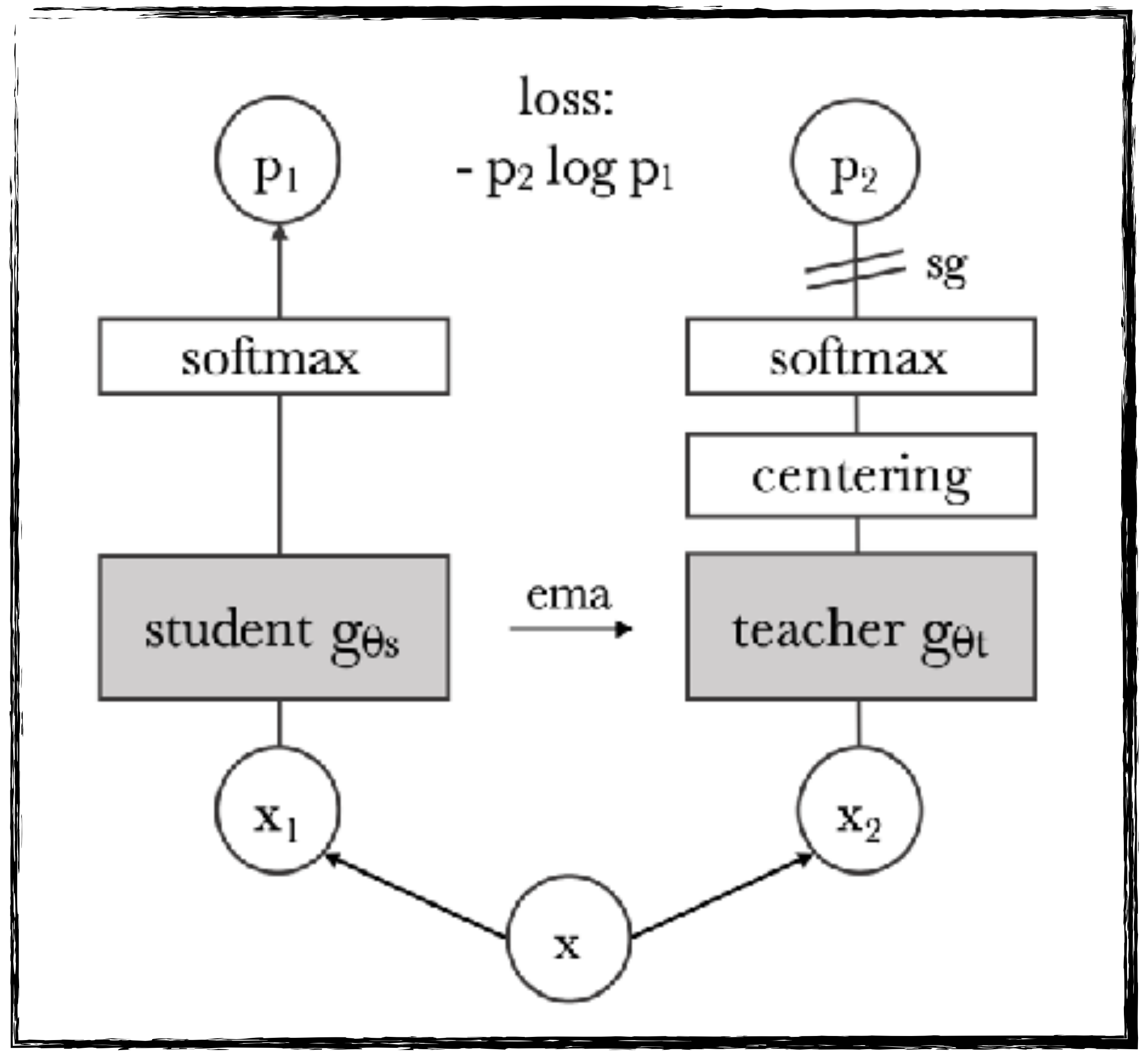
Priya Goyal[2] Piotr Bojanowski[2] Armand Joulin[2]

[1] Inria* [2] Facebook AI Research

Abstract

Unsupervised image representations have significantly reduced the gap with supervised pretraining, notably with the recent achievements of contrastive learning methods. These contrastive methods typically work online and rely on a large number of explicit pairwise feature comparisons, which is computationally challenging.

SwAV: generalises SeLa to cluster online



DINO: uses momentum ViT encoder, replaces online SK with centering and softmax

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. Caron et al. NeurIPS 2020.
Emerging Properties in Self-Supervised Vision Transformers. Caron et al. ICCV 2021.

More recently…

	Top-1		Δ
Method	2x224	2x160+4x96	
Supervised	76.5	76.0	−0.5
Contrastive-instance approaches			
SimCLR	68.2	70.6	+2.4
Clustering-based approaches			
SeLa-v2	67.2	71.8	+4.6
DeepCluster-v2	70.2	74.3	+4.1
SwAV	70.1	74.1	+4.0

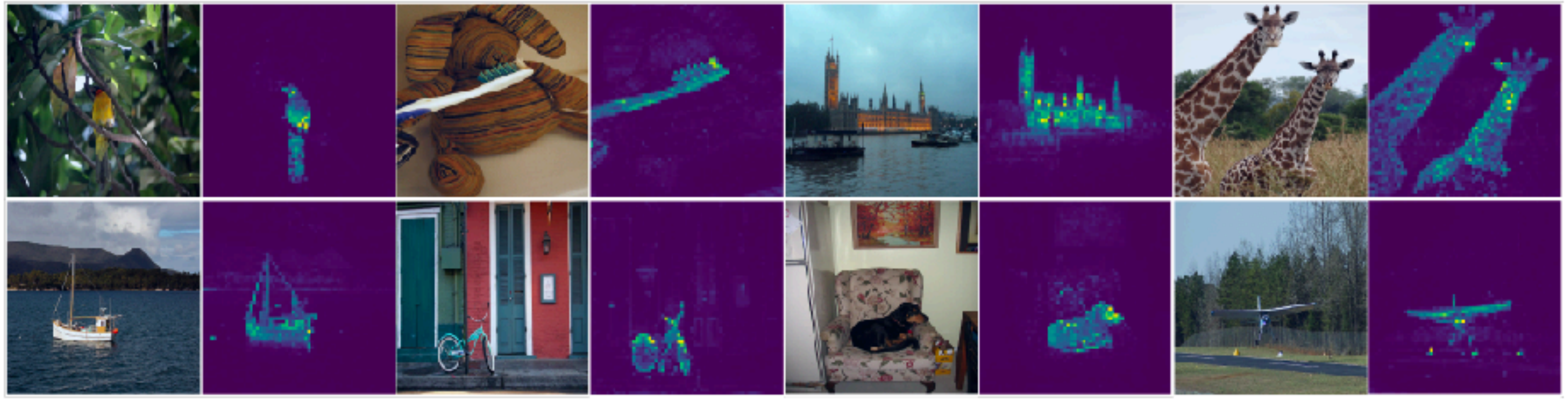
- SwAV uses SeLa's SK algo
- SeLa-v2 better than SimCLR

	Method	Momentum	Operation	Top-1
1	DINO	✓	Centering	76.1
2	–	✓	Softmax(batch)	75.8
3	–	✓	Sinkhorn-Knopp	76.0
4	–		Centering	0.1
5	–		Softmax(batch)	72.2
6	SwAV		Sinkhorn-Knopp	71.8

- DINO with SeLa's SK: same performance.

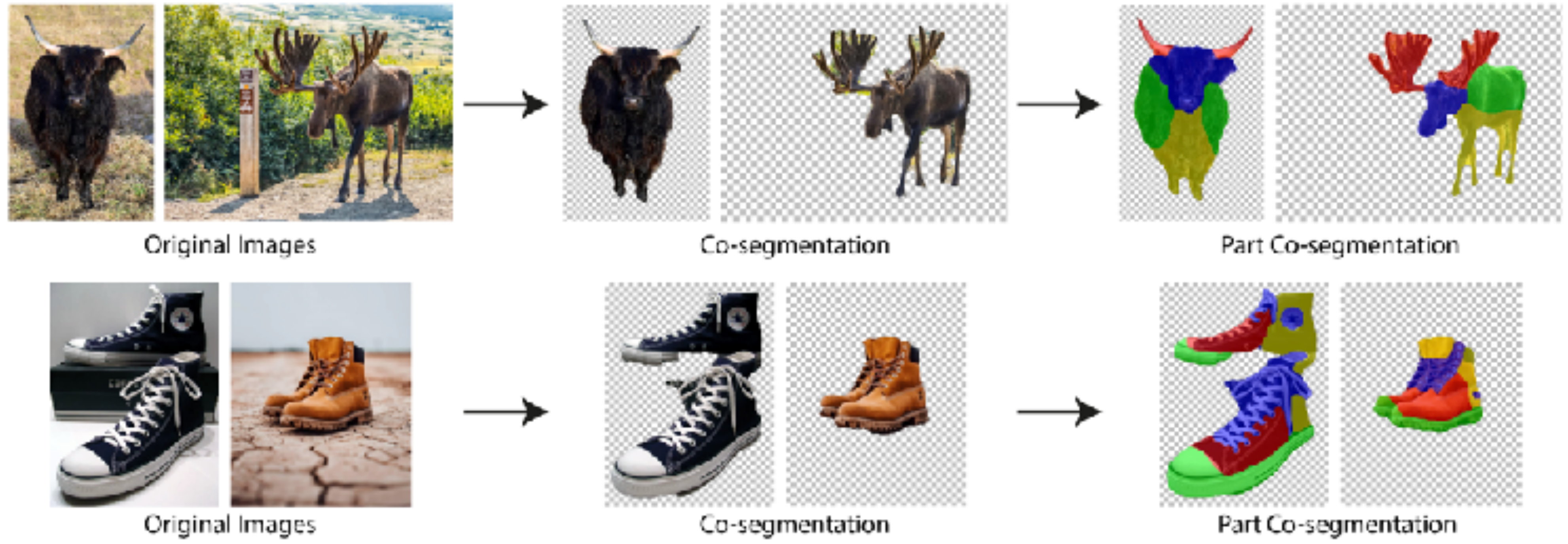
Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. Caron et al. NeurIPS 2020.
Emerging Properties in Self-Supervised Vision Transformers. Caron et al. ICCV 2021.

DINO has remarkable properties



The attention matrix of the [CLS] token with the spatial patches highlights the salient objects..

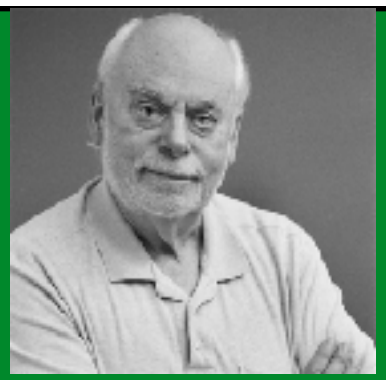
.. which is not the case for supervised learning



Spatial features even capture semantics across classes

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. Caron et al. NeurIPS 2020.
Emerging Properties in Self-Supervised Vision Transformers. Caron et al. ICCV 2021.
Deep ViT Features as Dense Visual Descriptorsg. Amir et al. ECCV'SSLWIN 2022.

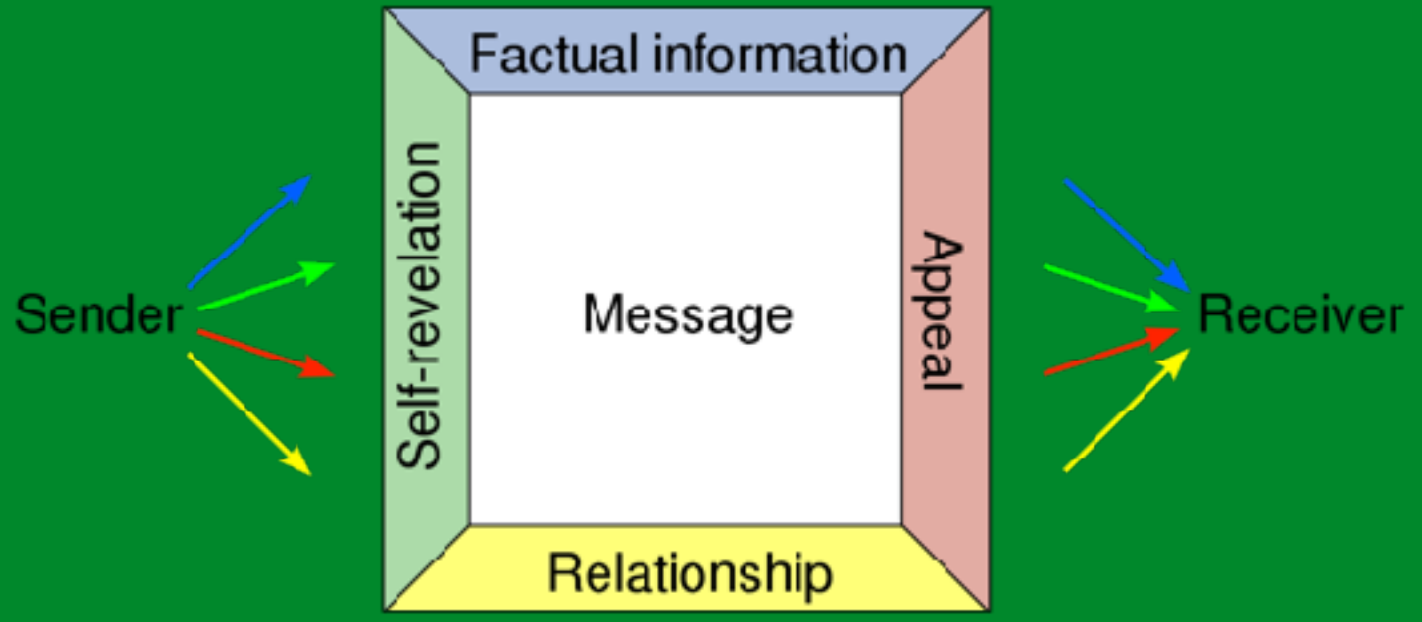
How research gets done: part 9



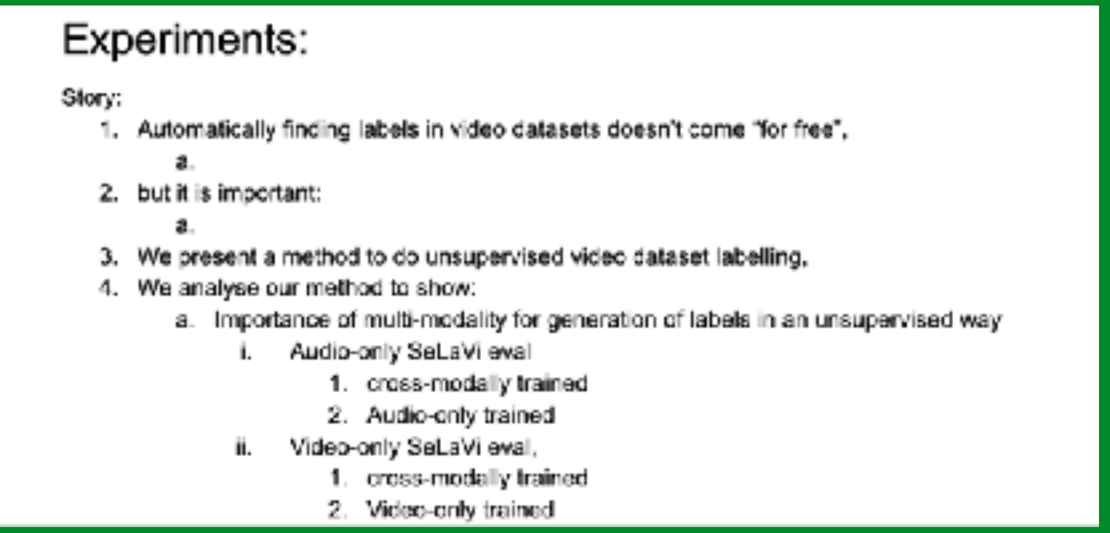
Fraser Stoddart:
"You've got to break the rules"

Previous parts:
[fundamental understanding/read papers, how-to-read-papers, implement & tinker with code, realise and seek *funny* moments, MVP/principles/benchmarks/baselines, when to (not) give up/impact-vs-work, importance of Ablations]

- Ideally before or latest when all previous steps are (more or less) completed we develop the *storyline*
- Why a story? Aren't we writing a hard, cold, scientific paper?
  - Yes, but: (science) communication not as easy:
  - So we need to put in a lot of work
  - What's the *rode draad*/overarching motif?



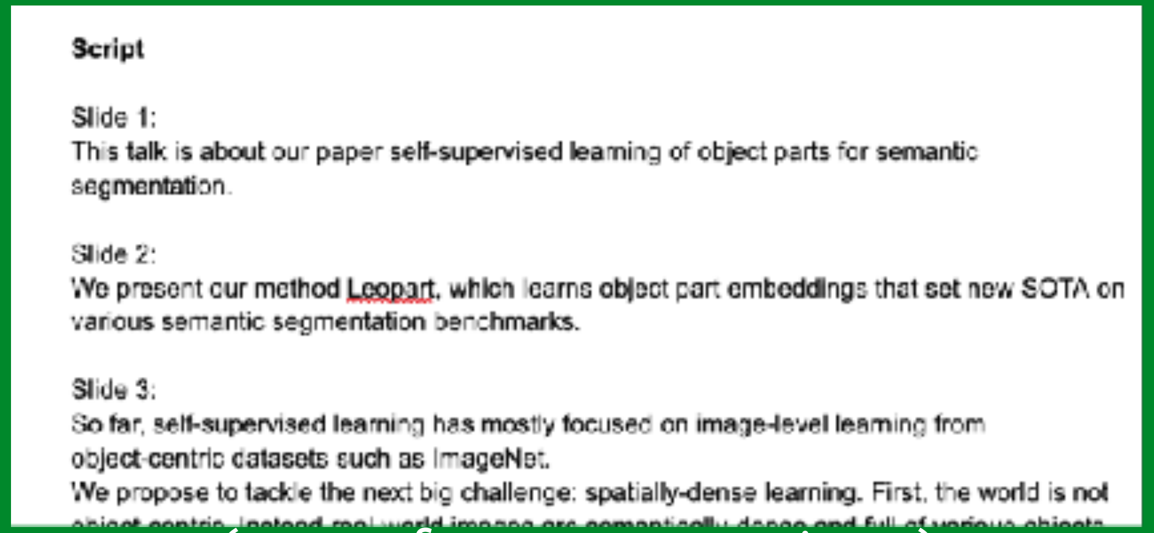
- Use google docs, don't make it super nice, just re-iterate from scratch multiple times.



Experiments:

Story:

1. Automatically finding labels in video datasets doesn't come "for free",
   a.
2. but it is important:
   a.
3. We present a method to do unsupervised video dataset labelling.
4. We analyse our method to show:
   a. Importance of multi-modality for generation of labels in an unsupervised way
      i. Audio-only SeLaVi eval
         1. cross-modally trained
         2. Audio-only trained
      ii. Video-only SeLaVi eval,
         1. cross-modally trained
         2. Video-only trained



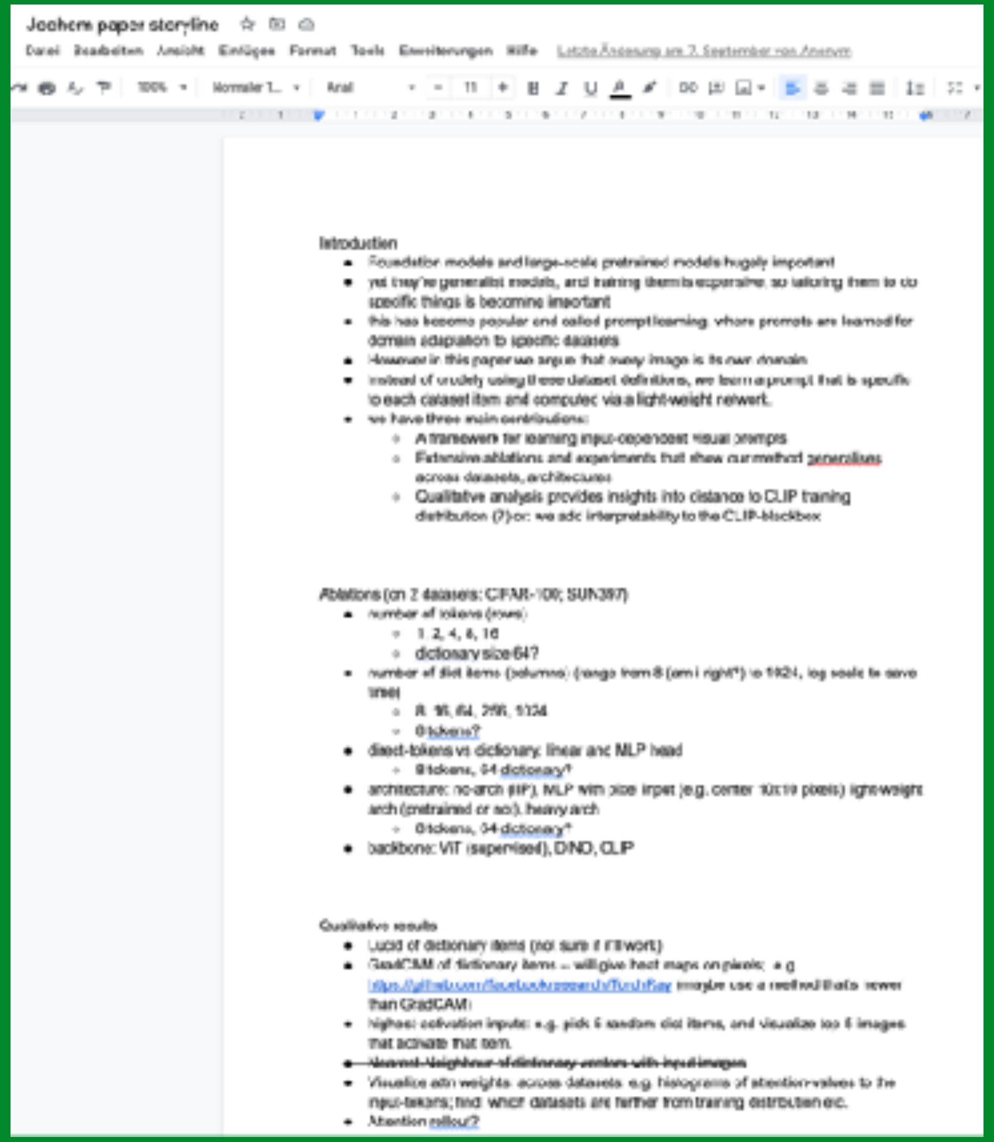
Script

Slide 1:
This talk is about our paper self-supervised learning of object parts for semantic segmentation.

Slide 2:
We present our method Leopart, which learns object part embeddings that set new SOTA on various semantic segmentation benchmarks.

Slide 3:
So far, self-supervised learning has mostly focused on image-level learning from object-centric datasets such as ImageNet.
We propose to tackle the next big challenge: spatially-dense learning. First, the world is not

(same for presentations)



Introduction

- Foundation models and large-scale pretrained models hugely important
- Qualitative analysis provides insights into distance to CLIP training distribution

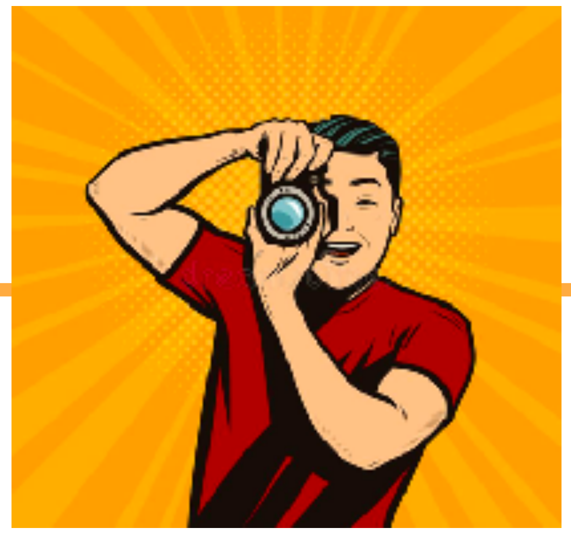
Ablations (on 2 datasets: CIFAR-100; SUN397)

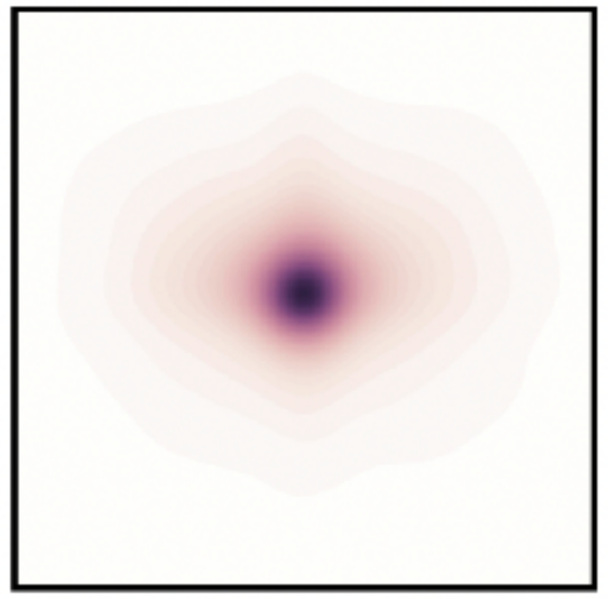
- 1, 2, 4, 8, 16
- dictionary size 64?
- backbone: ViT (supervised), DINO, CLIP

Qualitative results

- Best thing: you'll discover important missing experiments & have the introduction part of paper almost done

However: The world is not object-centric.

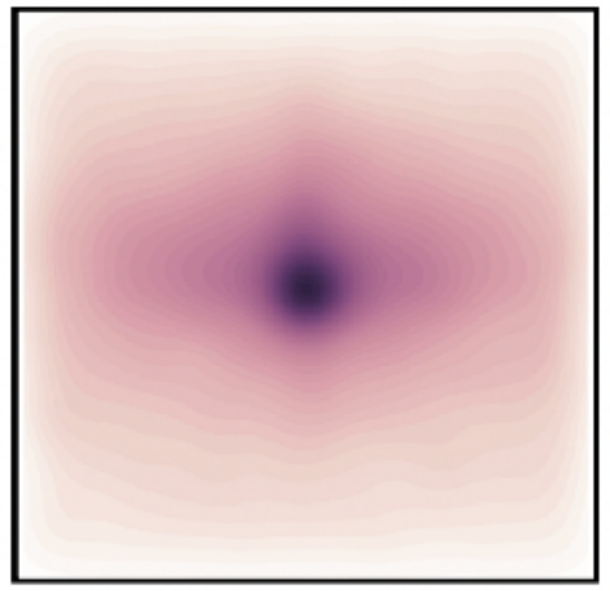




ILSVRC-DET 2014 (train+val)
0.35M images

Object-centric image

Dense real-world image



COCO 2017 (train+val)
0.12M images

Self-Supervised Learning of Object Parts for Semantic Segmentation
CVPR 2022

Adrian Ziegler*, Yuki M Asano

Technical University of Munich, University of Amsterdam
* work done as MSc thesis

Self-Supervised **Le**arning of **O**bject **Part**s for Semantic Segmentation 🐆

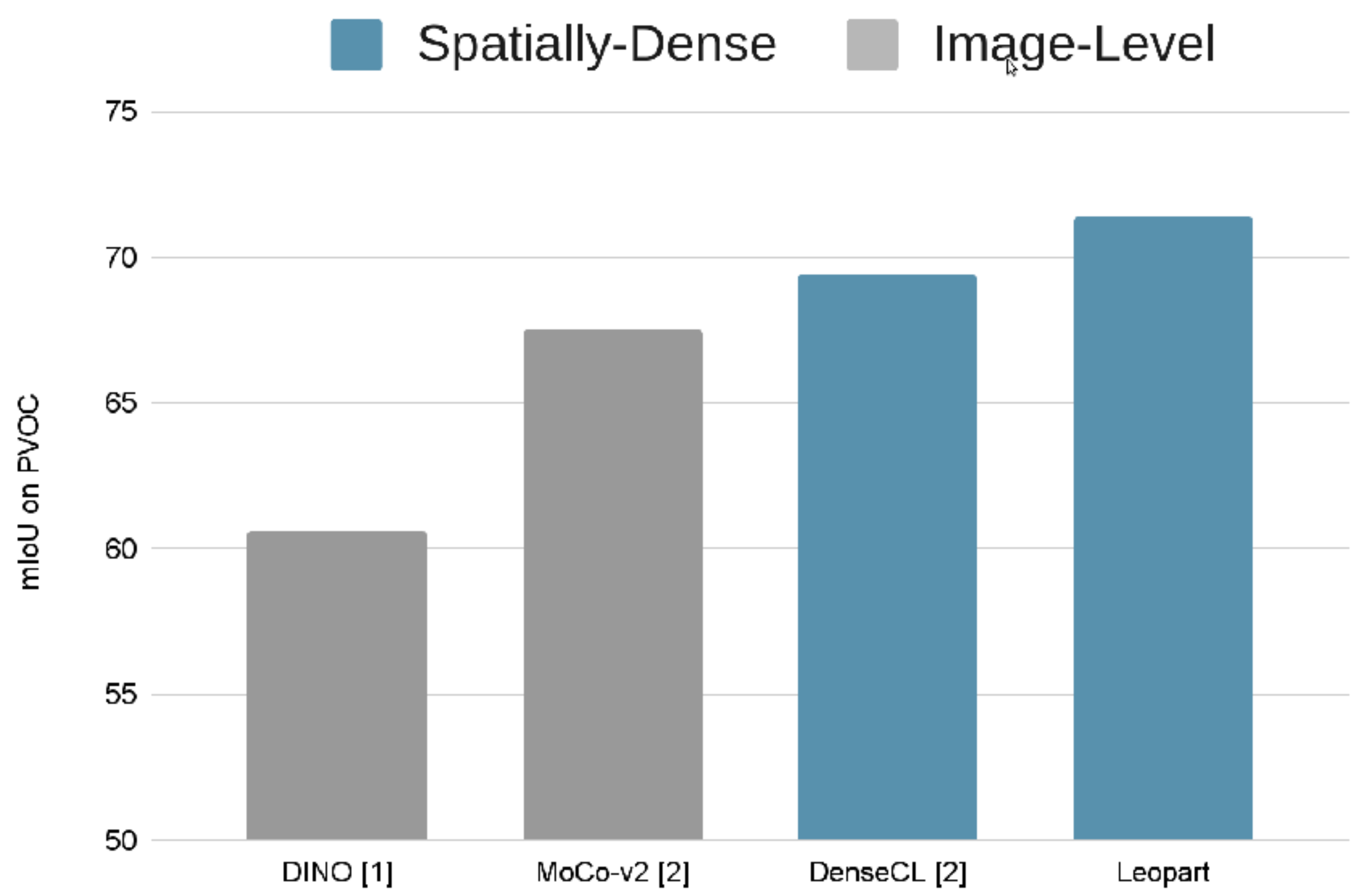
CVPR 2022

Adrian Ziegler*, Yuki M Asano

Technical University of Munich, University of Amsterdam
* work done as MSc thesis

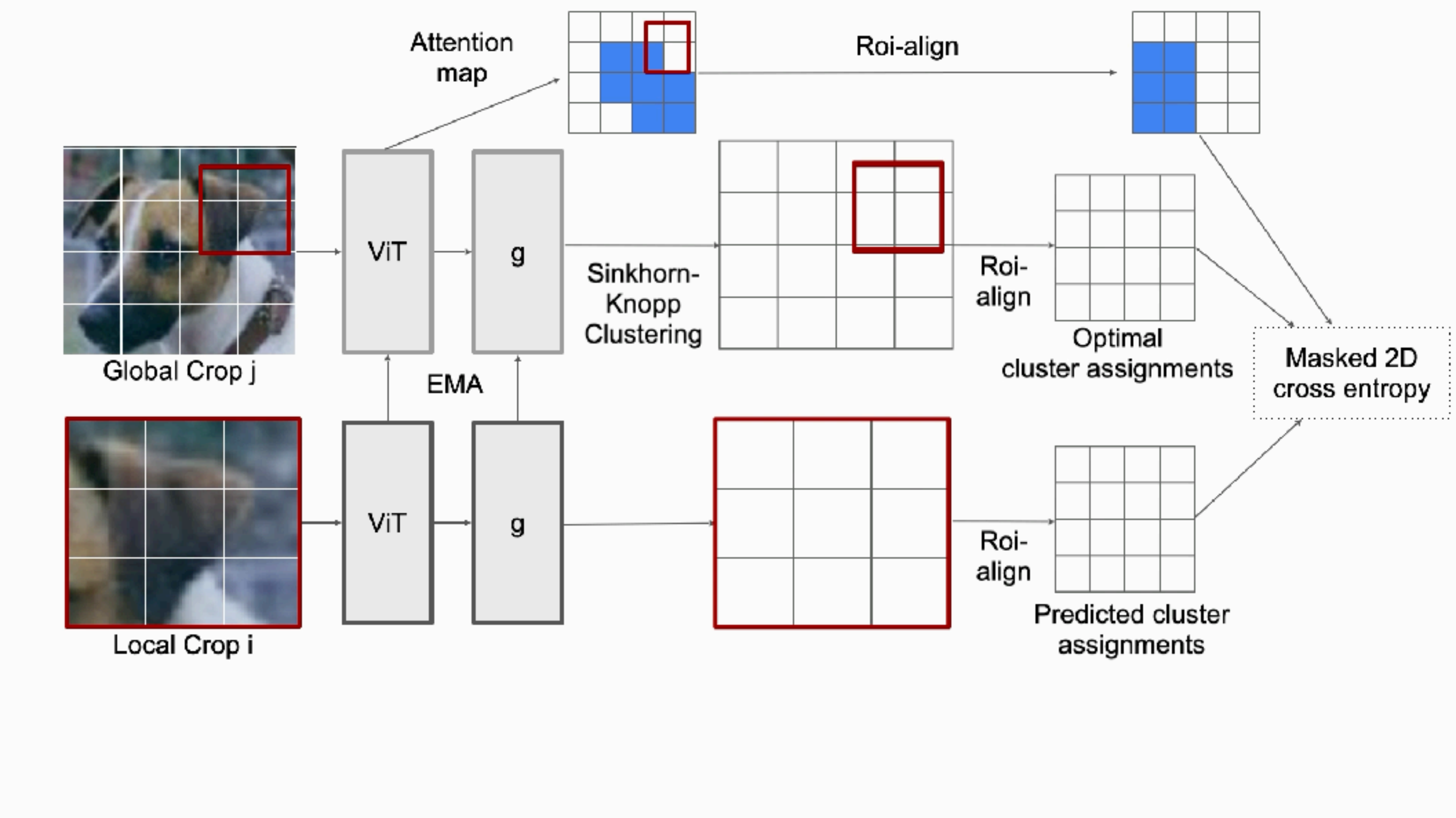
Self-Supervised Learning has to move from image-level to spatially-dense learning

1) The world is not object-centric
2) Spatially-dense learning scales better
3) Spatially-dense learning improves performance on dense prediction tasks



[1] Caron et. al, ICCV 2021 [2] He et. al, CVPR 2020 [3] Wang et. al, CVPR 2021

We propose a dense clustering pretext task to learn object parts



Quiz: Why did we use RoI-Align and not RoI-Pool?

1) RoI-Align is faster to compute
2) RoI-Align can take care of non-rectangular selections
3) RoI-Align works for non-integer locations
4) RoI-Pool would have worked as well

DINO

+ Pretext Task

Additional Innovation 1: Cluster-Based Foreground Extraction (CBFE)

- Use attention masks as ″noisy foreground″
- Assign clusters that have large IoU with to this ″foreground″.
- Improves foreground extraction by >10% in comparison to DINO's attention map.

CBFE

DINO Attention Masks

Leopart Cluster Masks

DINO

+ Pretext Task

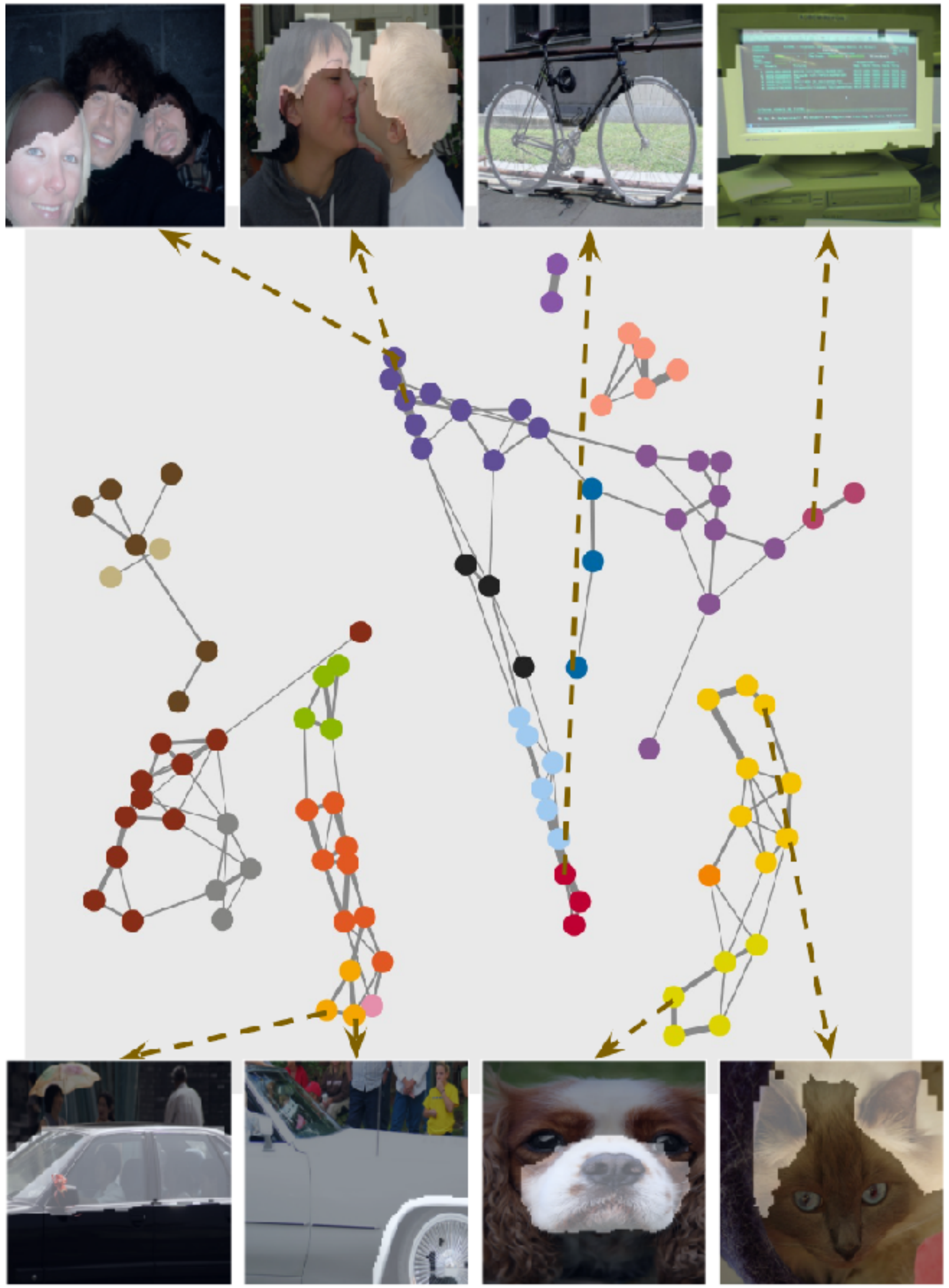
+ CBFE

Additional Innovation 2: Overclustering with Community Detection (CD)

- Interpret objects as co-occurring object parts
- Construct undirected weighted graph
  - Each node corresponds to a cluster
  - Edges weights by co-occurrence probability

Overclustering with Community Detection

- Run community detection algorithm on graph to merge to objects
- The network shows
  - Semantically close object parts are in the same community.
  - Object parts are learned that do not latch on low-level features.



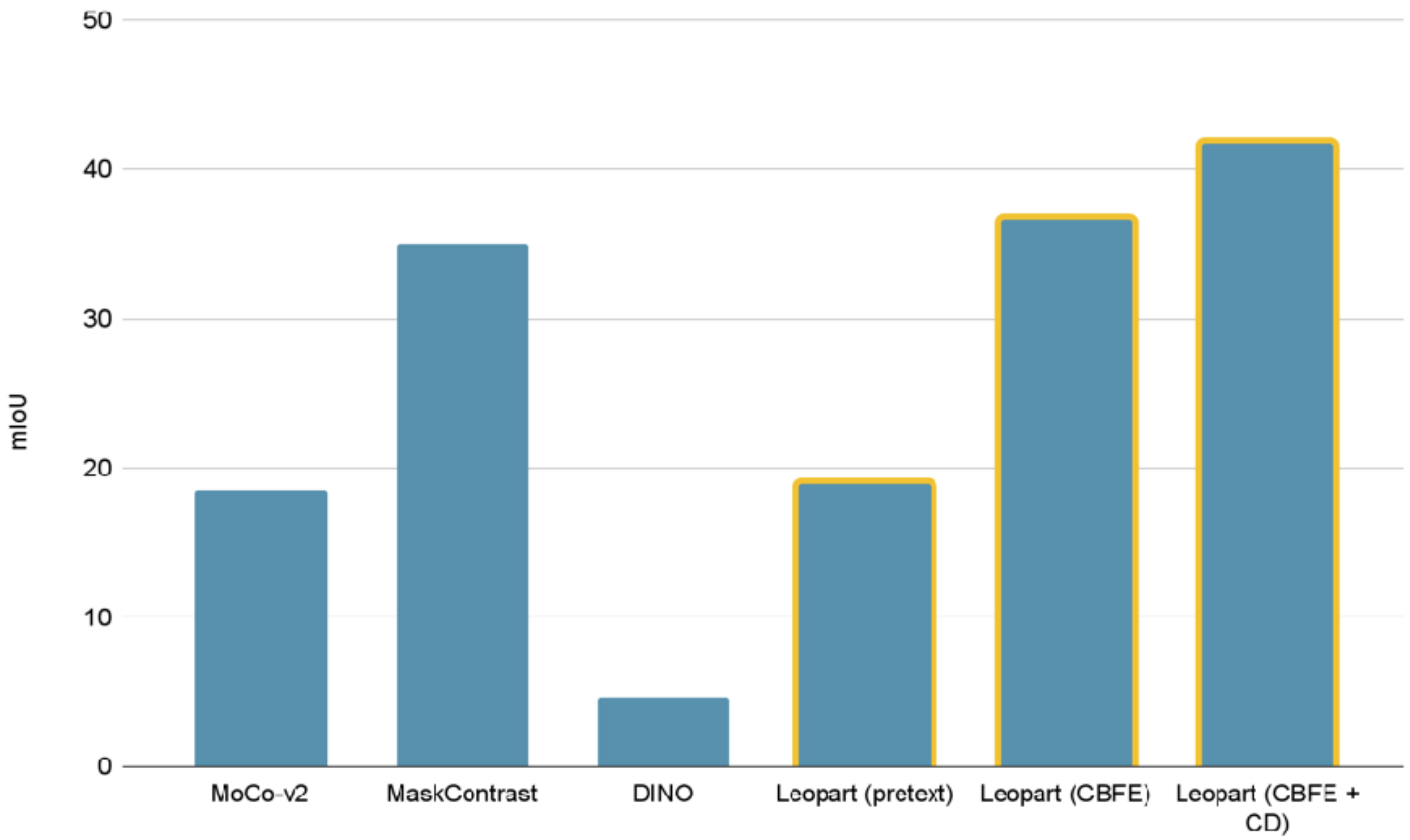
DINO

+ Pretext Task

+ CBFE

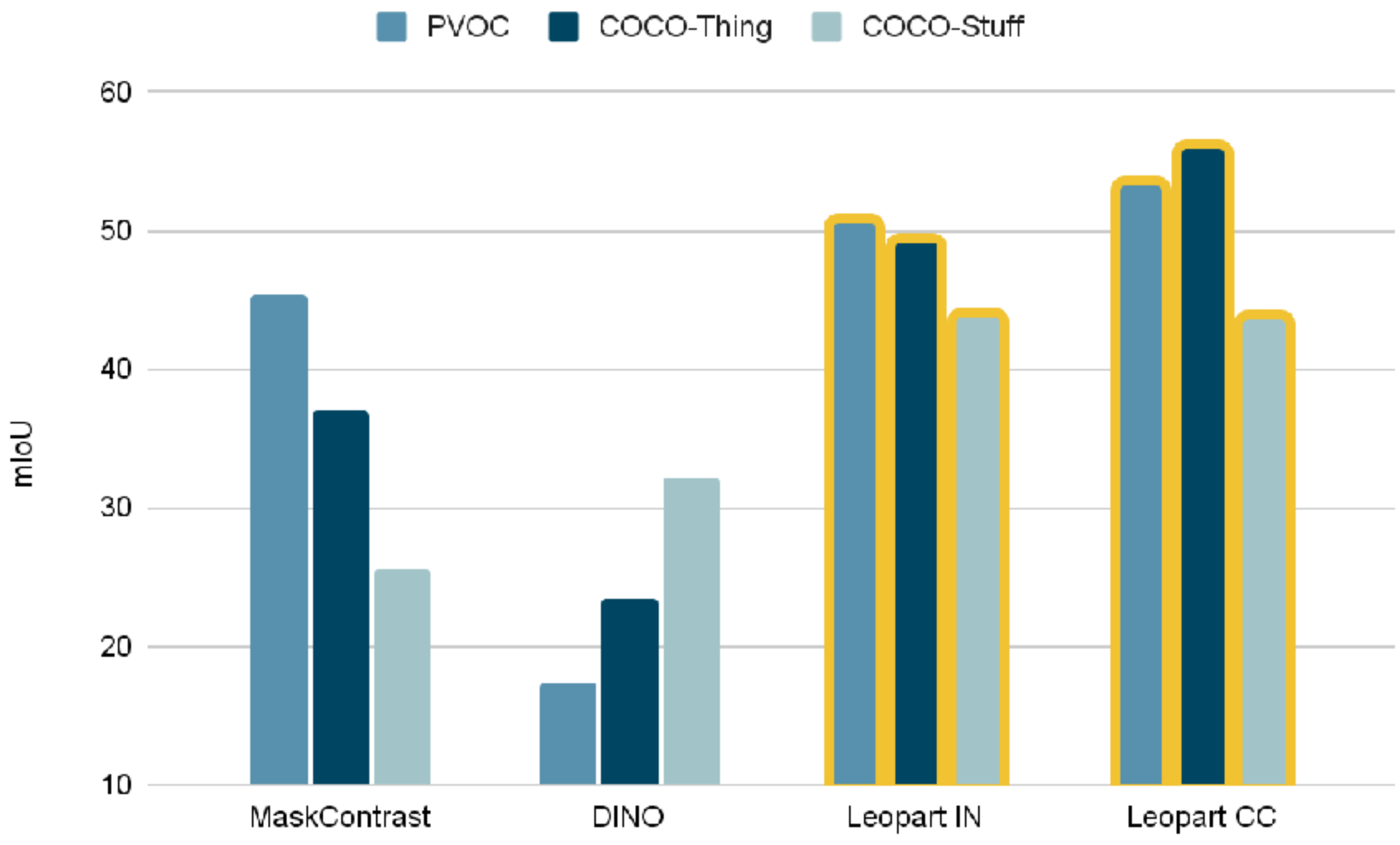
+CD

Leopart improves fully unsupervised SOTA by >6%

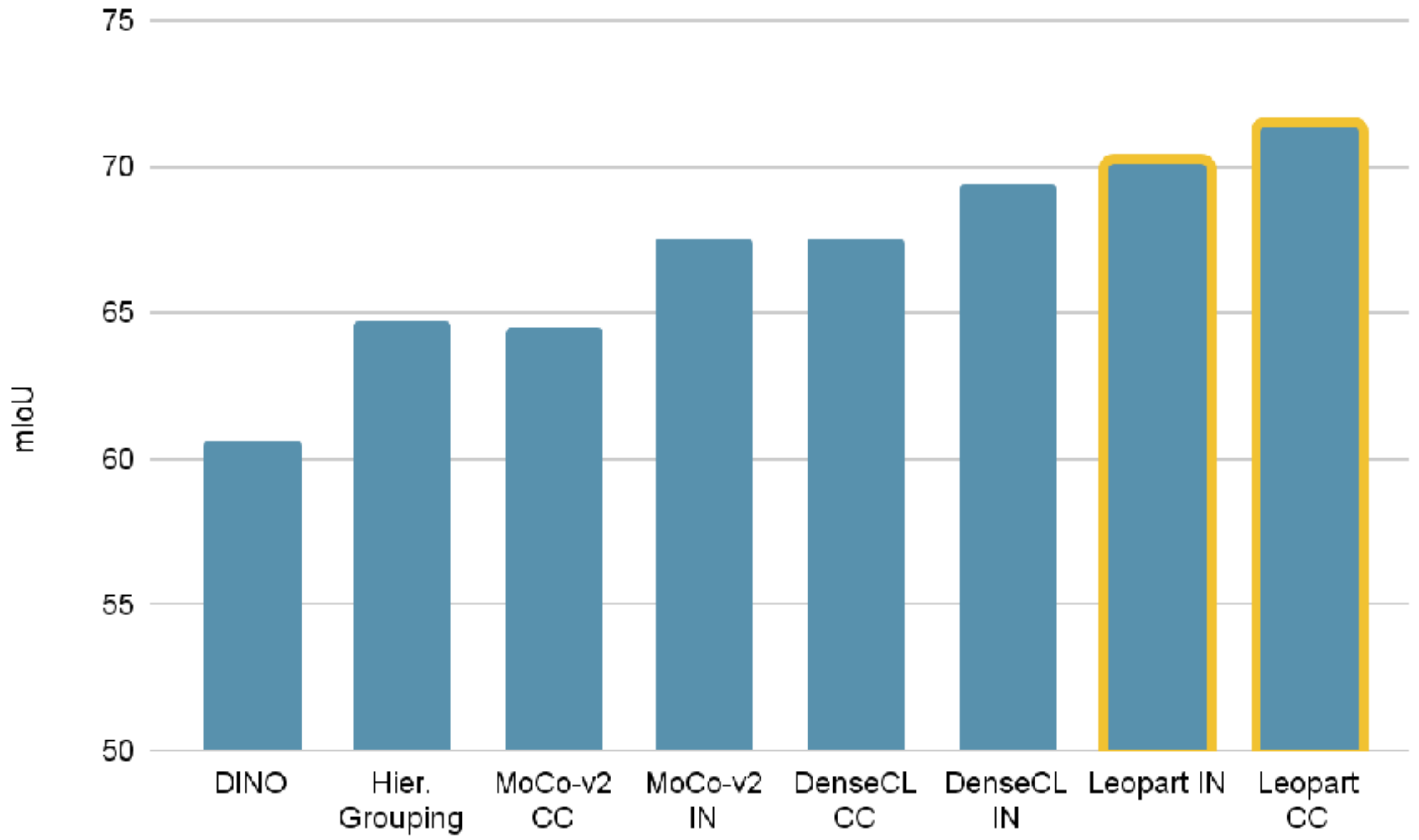


[2] Van Gansbeke et. al, ICCV 2021

Leopart achieves transfer SOTA on three datasets simultaneously



Semantic Segmentation Results with K=500
IN = ImageNet, CC= COCO



Semantic Segmentation Results with FCN Head on PVOC

Augmentations were key for both SeLa and Leopart.

Note how they were used
a) to solve the chicken-and-egg problem
b) not only as invariances in Leopart (the crop's location was essential)

Next we will investigate these augmentations a bit more in detail.

How can we isolate the effect of augmentations?

By learning from a single image

Data
1M images

VS.

Data
1M crops of 1 image

How do we go about this?

Data

1M crops of 1 image

Process

- generate a dataset of 1.2M transformations of the same image
- train using an off-the-shelf SSL method
- compare to using 1.2M different images

What do we learn?

1) How much a single image can take us from a random initialization
2) Whether self-supervised learning can extract more information than that

Tested images

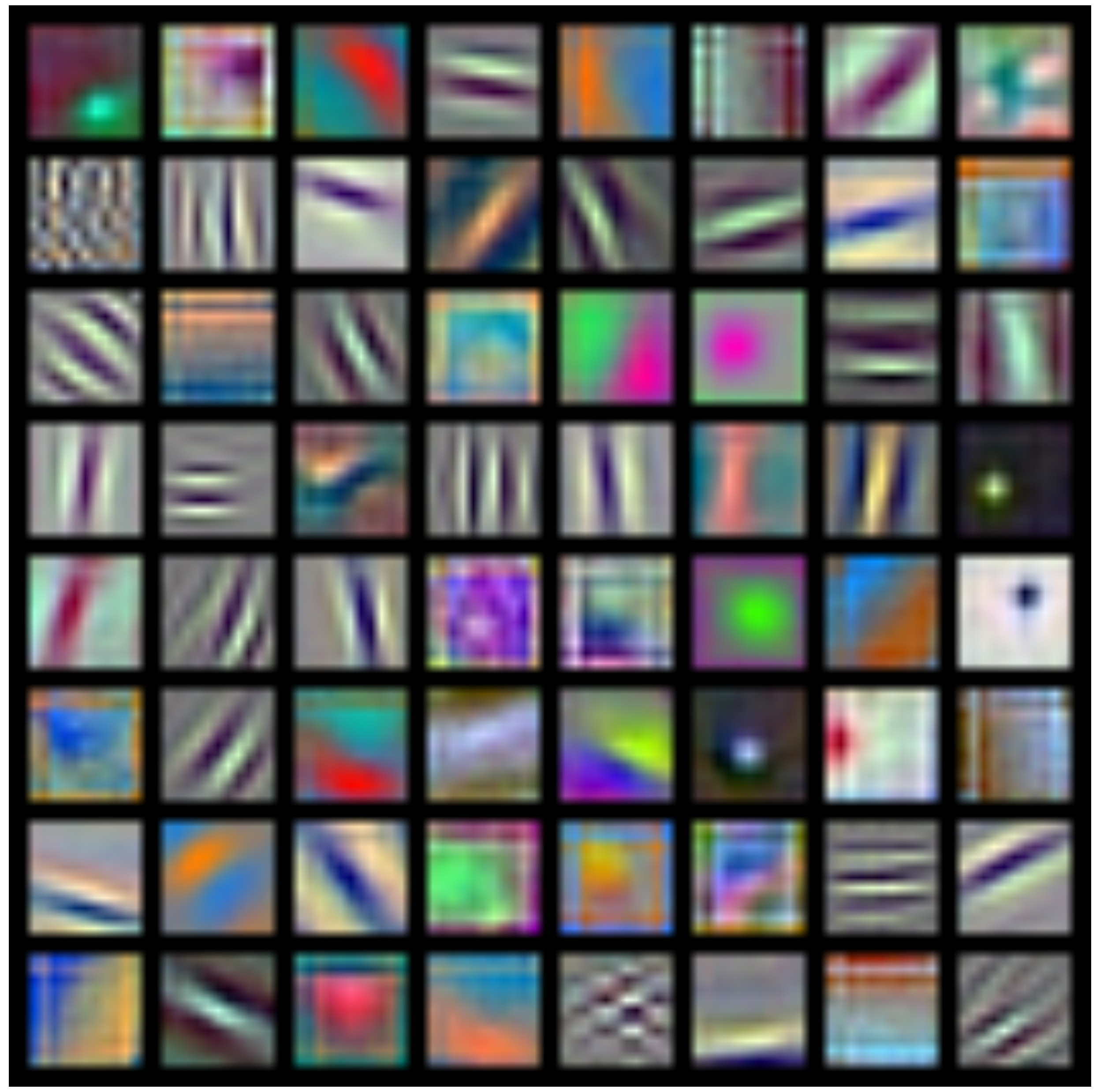
Image A

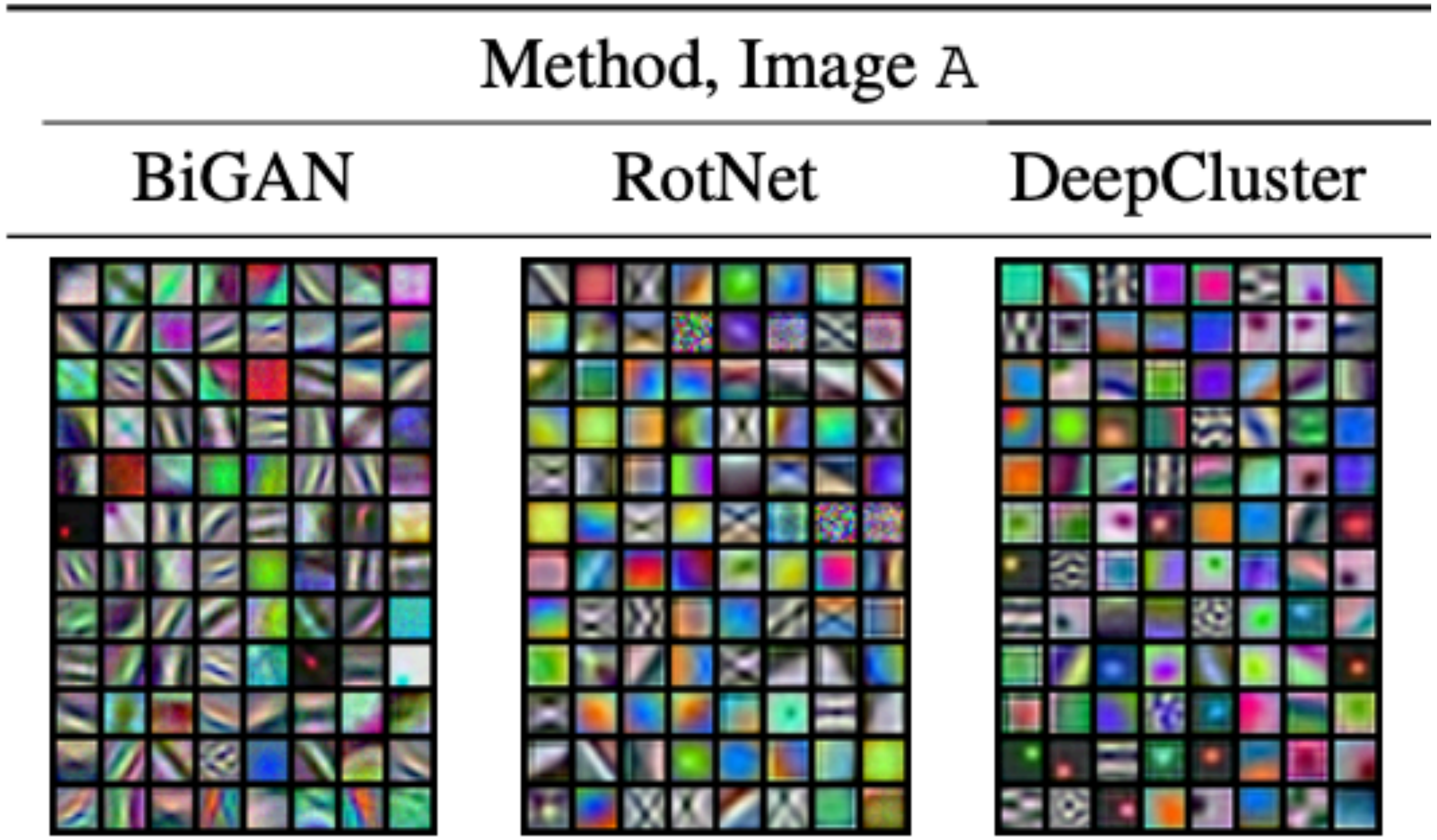
Image B

Image C

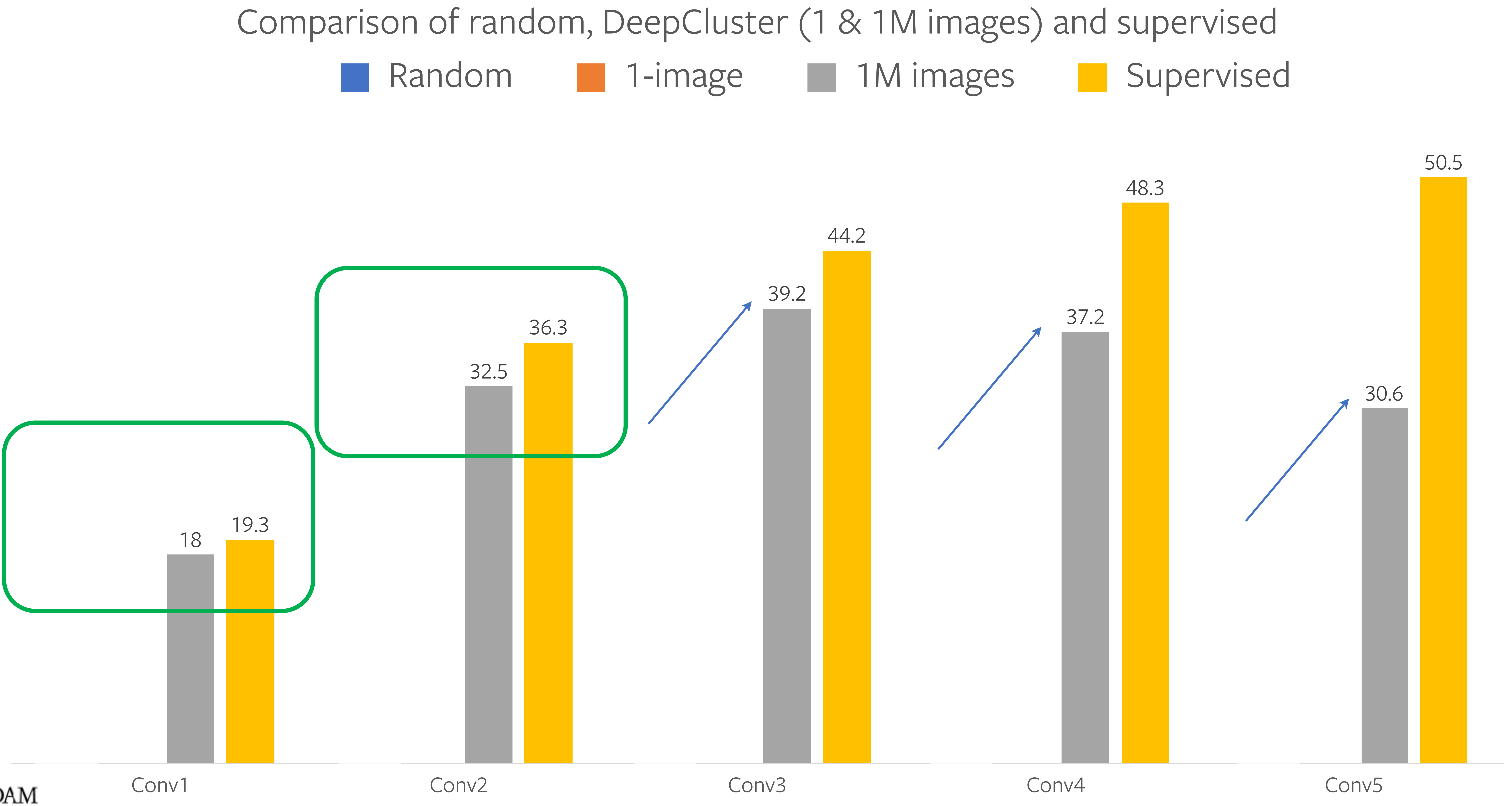
Self-supervised learning from one image: First convolutional layer

1.2M images, supervised

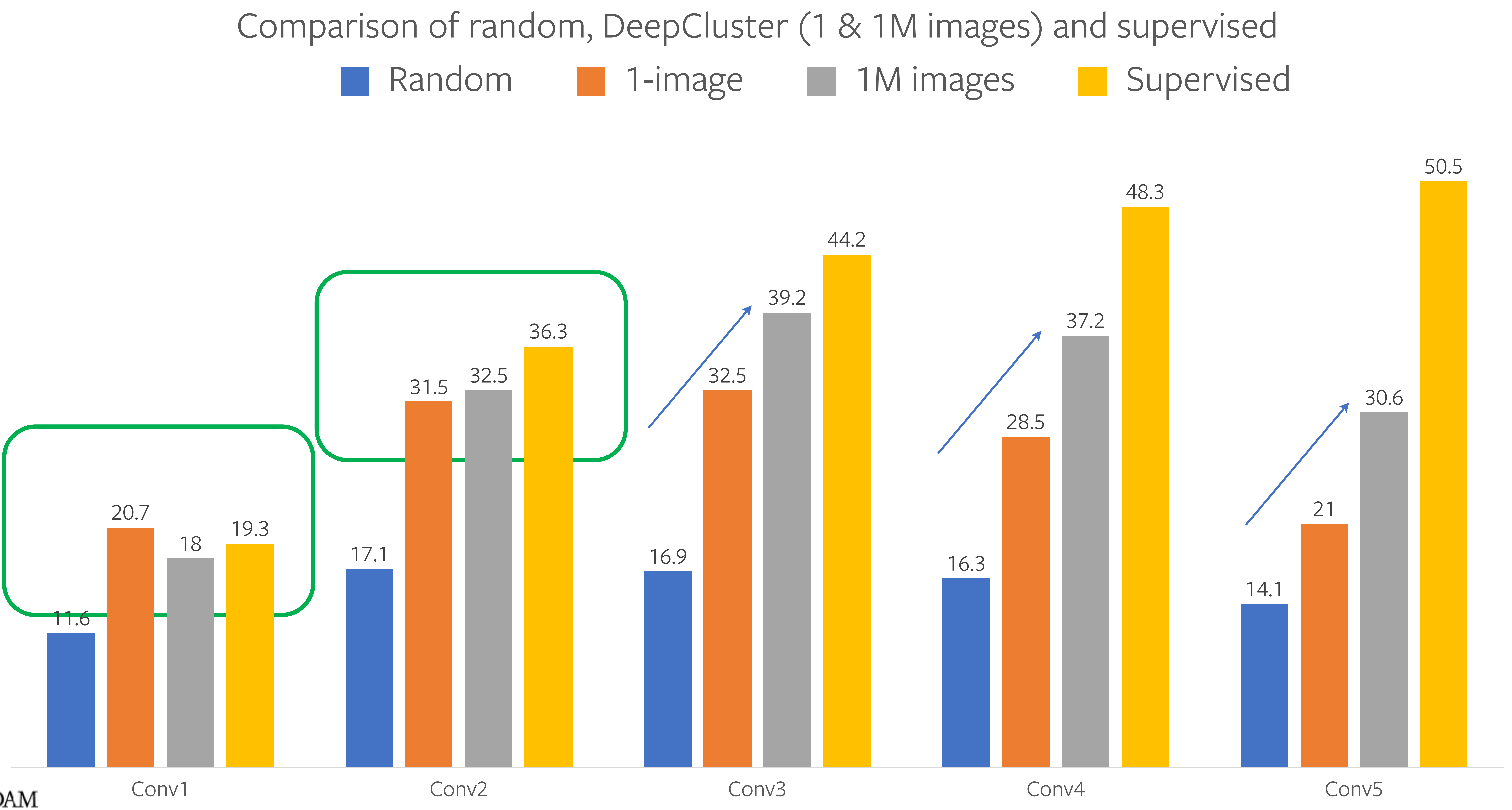




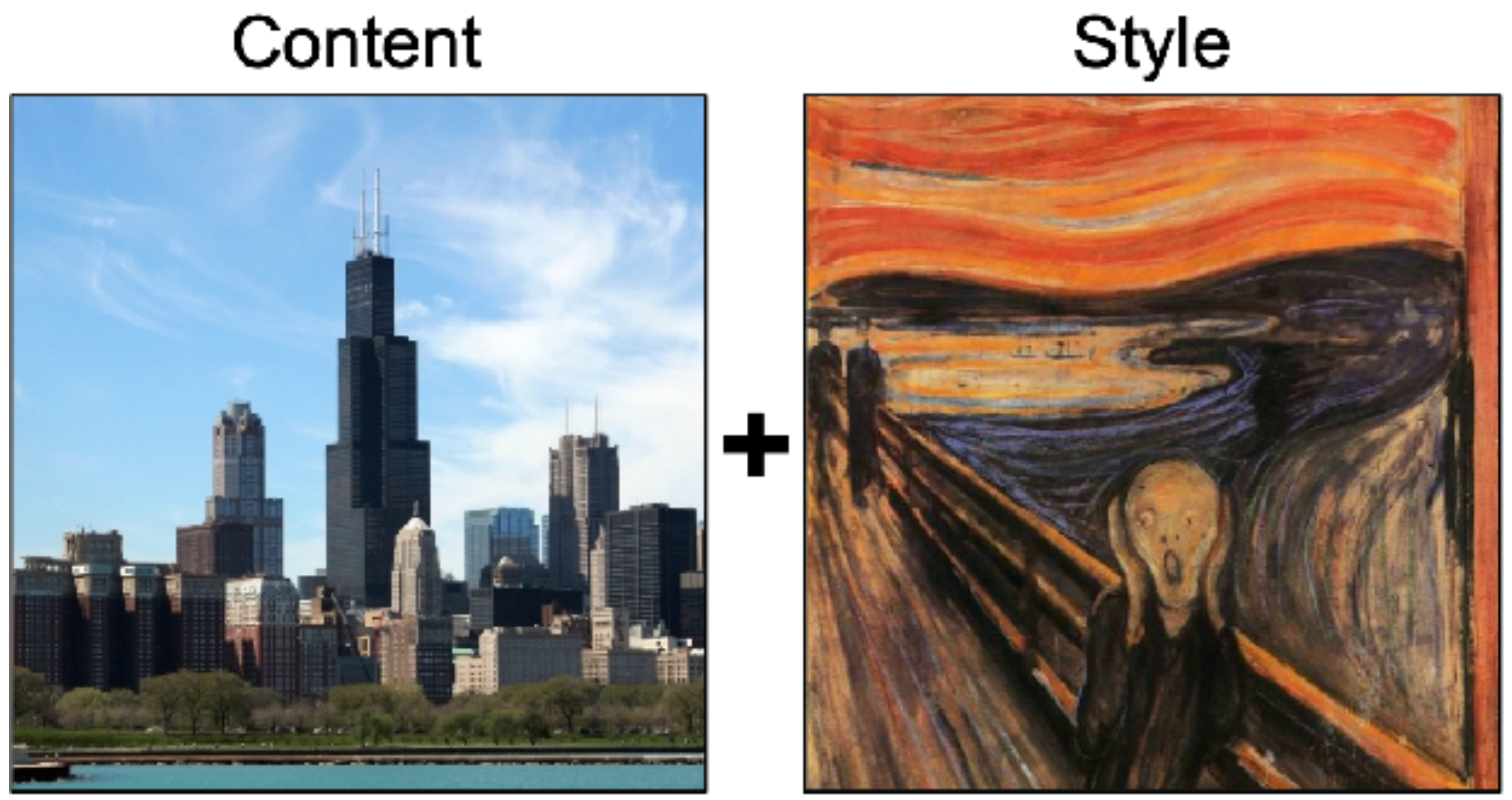
Self-supervised learning from one image: Quality (ImageNet linear probes)



Self-supervised learning from one image: Quality (ImageNet linear probes)



Style transfer with a 1-image trained CNN



UNIVERSITY OF OXFORD

[Update Feb 2021] Using a ResNet-50 and MoCo loss, we get even closer for fine-tuning tasks.

COCO R50-C4 finetuning, 1x

pre-train	Bounding-box			Segmentation		
	AP^{bb}	AP^{bb}_{50}	AP_{75}	AP^{mk}	AP^{mk}_{50}	AP^{mk}_{75}
Random	26.4	44.0	27.8	29.3	46.9	30.8
Supervised	38.2	58.2	41.2	33.3	54.7	35.2
ours 1-image A	36.5	55.2	39.2	32.1	52.2	34.0
MoCo-v1	38.5	58.3	41.6	33.6	54.8	35.6
MoCo-v2	39.0	58.6	41.9	34.2	55.4	36.2

+10% mAP from a single image and augmentations

Within 3% of MoCo-v2 on full ImageNet

Surface normal estimation on NYUv2

	Angle Distance ↓		Within $t°$ ↑		
Initialization	Mean	Median	11.25	22.5	30
Random	26.3	16.1	37.9	60.6	69.0
ImageNet supervised	26.4	17.1	36.1	59.2	68.5
1-image	**24.3**	**15.0**	**40.9**	**62.4**	**70.6**
Jigsaw ImageNet	24.2	14.5	41.2	64.2	72.5

(Unpublished)

Update 2:

https://single-image-distill.github.io/

.. using knowledge distillation.

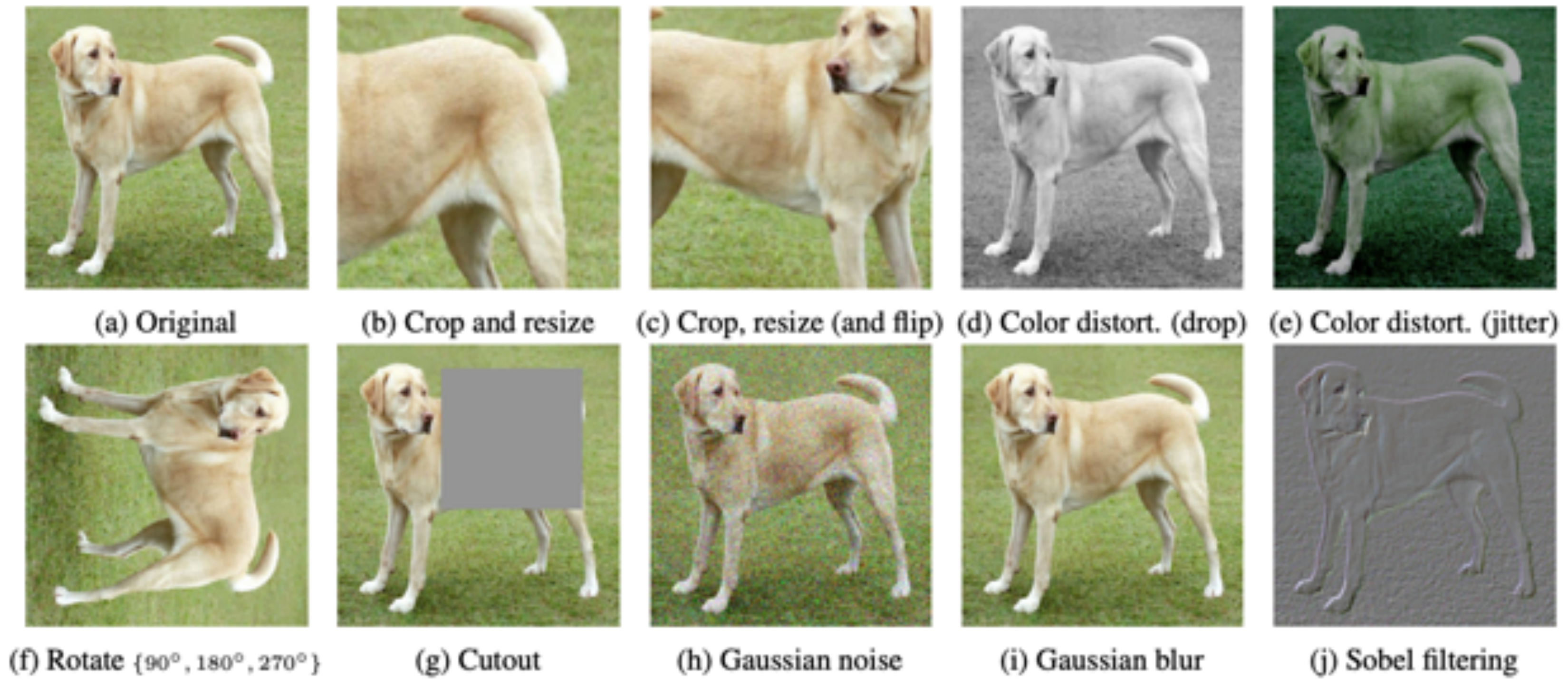
Conclusion

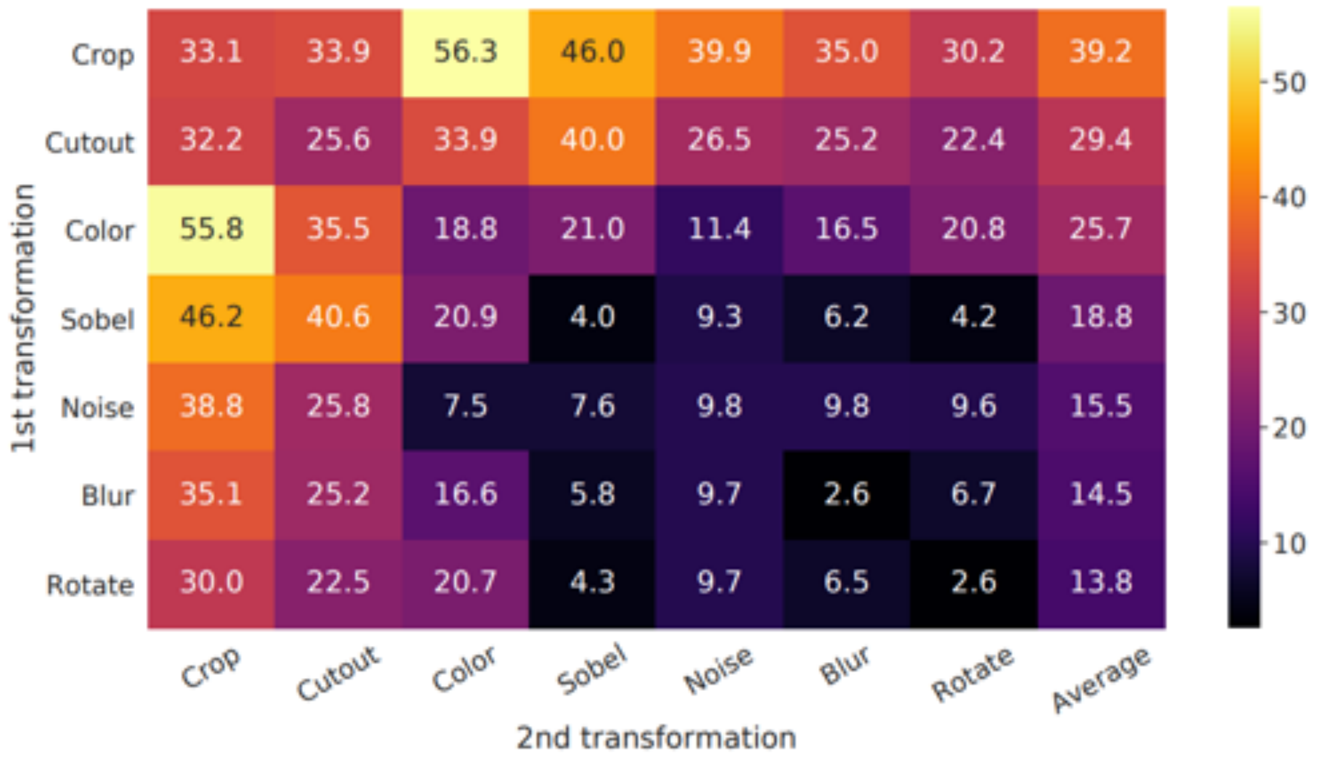
Self-supervised learning works (to a very large extent) thanks to augmentations.

Augmentations revisited

case	MLP	aug+	cos	epochs	ImageNet acc.	VOC detection AP_{50}	AP	AP_{75}
supervised					76.5	81.3	53.5	58.8
MoCo v1				200	60.6	81.5	55.9	62.6
(a)	✓			200	66.2	82.0	56.4	62.6
(b)		✓		200	63.4	82.2	56.8	63.2
(c)	✓	✓		200	67.3	**82.5**	57.2	63.9
(d)	✓	✓	✓	200	67.5	82.4	57.0	63.6
(e)	✓	✓	✓	**800**	**71.1**	**82.5**	**57.4**	**64.0**

augmentations (besides longer training and MLP head and better learning rate schedule) have huge impact





intuition:

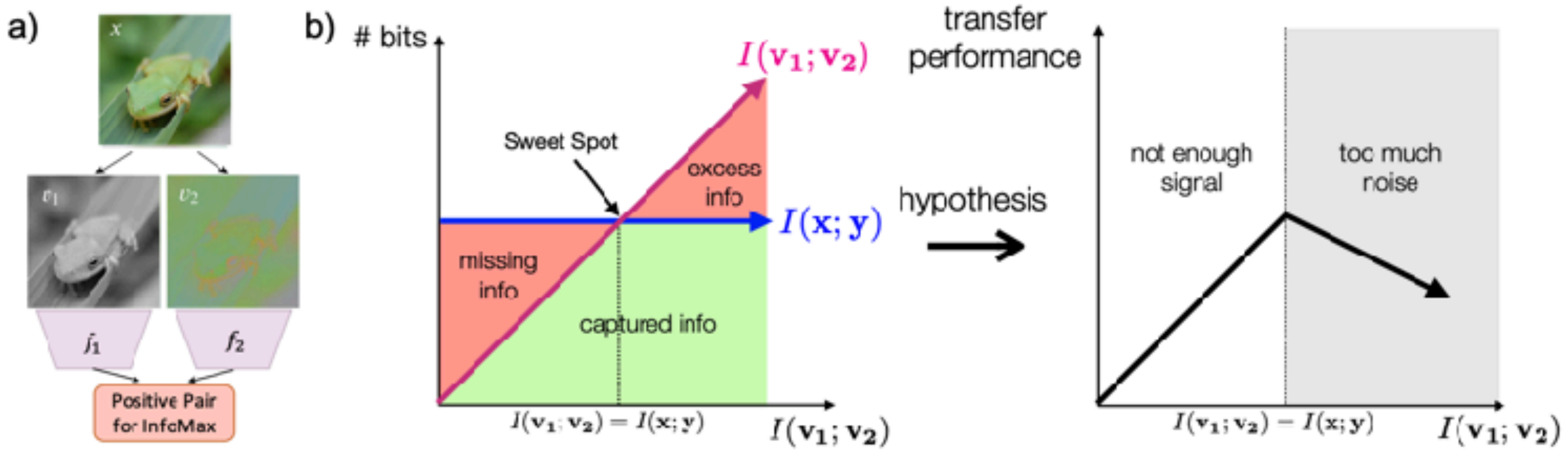
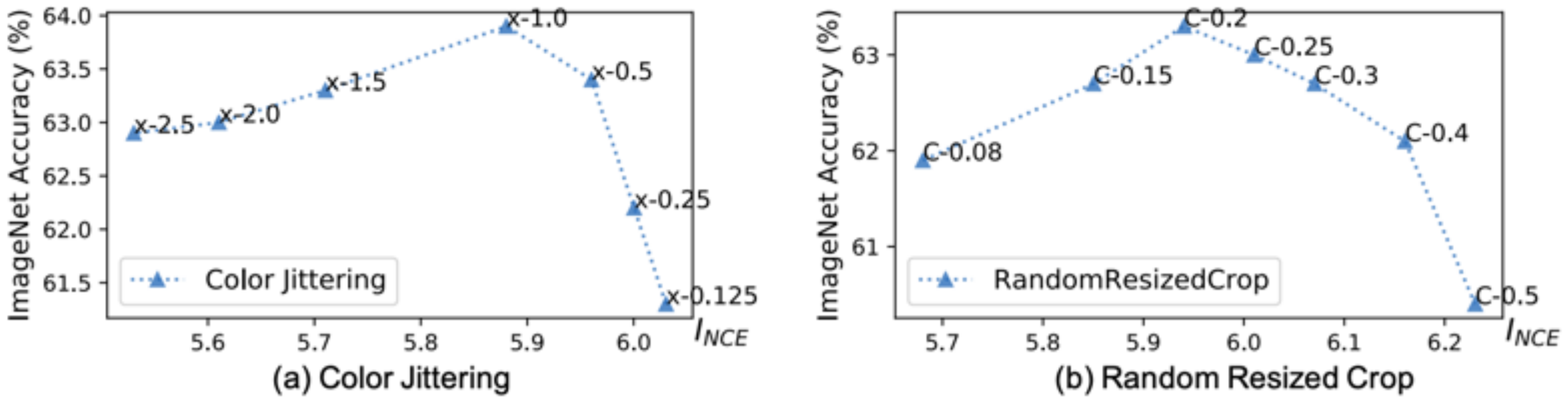


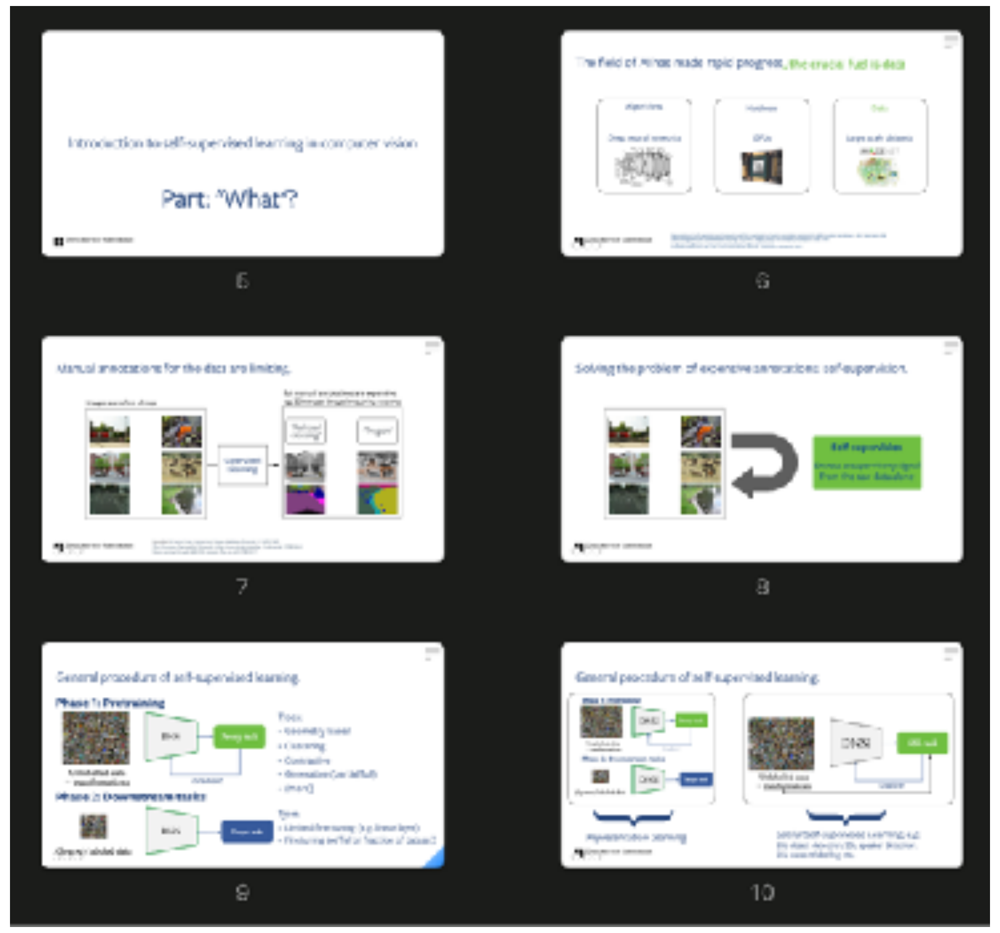
Figure 1: **(a)** Schematic of multiview contrastive representation learning, where an image is split into two views, and passed through two encoders to learn an embedding where the views are close relative to views from other images. **(b)** When we have views that maximize $I(\mathbf{v_1}; \mathbf{y})$ and $I(\mathbf{v_2}; \mathbf{y})$ (how much task-relevant information is contained) while minimizing $I(\mathbf{v_1}; \mathbf{v_2})$ (information shared between views, including both task-relevant and irrelevant information), there are three regimes: *missing information* which leads to degraded performance due to $I(\mathbf{v_1}; \mathbf{v_2}) < I(\mathbf{x}; \mathbf{y})$; *excess noise* which worsens generalization due to additional noise; *sweet spot* where the only information shared between $\mathbf{v_1}$ and $\mathbf{v_2}$ is task-relevant and such information is complete.

evidence:

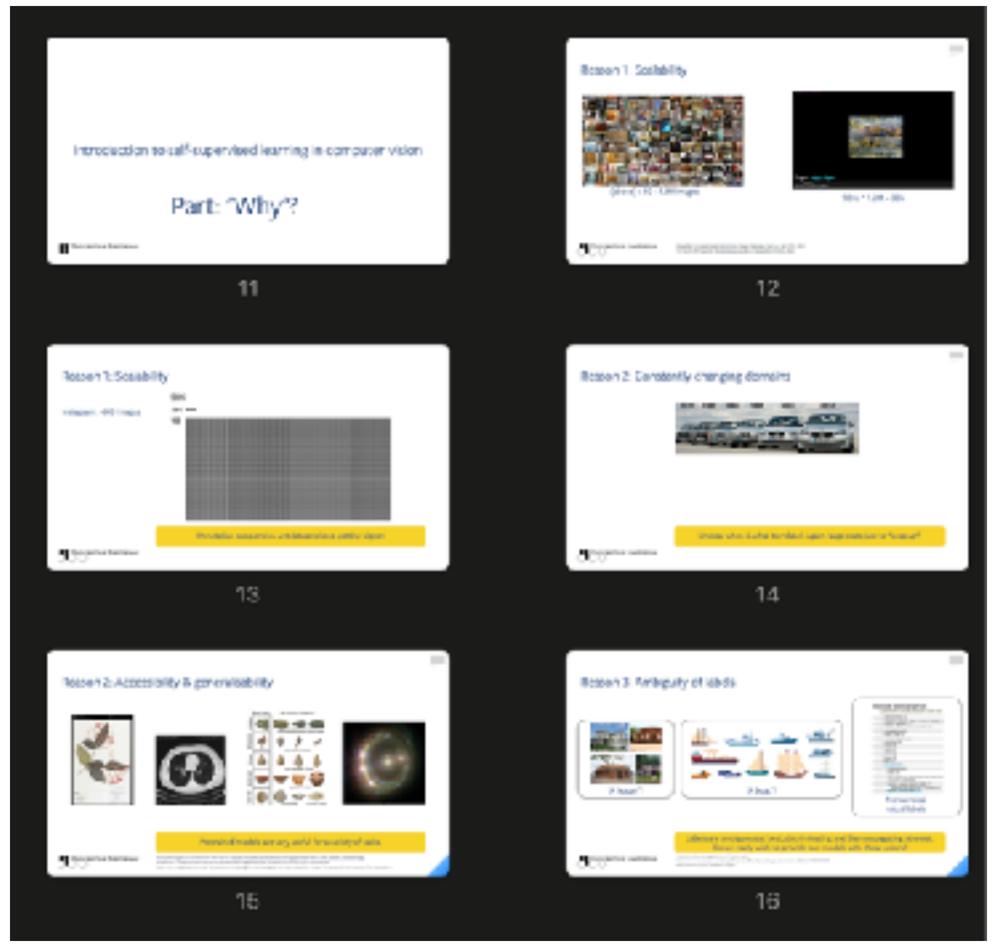


A Simple Framework for Contrastive Learning of Visual Representations. Chen et al. ICML 2020
Improved Baselines with Momentum Contrastive Learning. Chen et al. 2020
What Makes for Good Views for Contrastive Learning? Tian et al. NeurIPS 2020

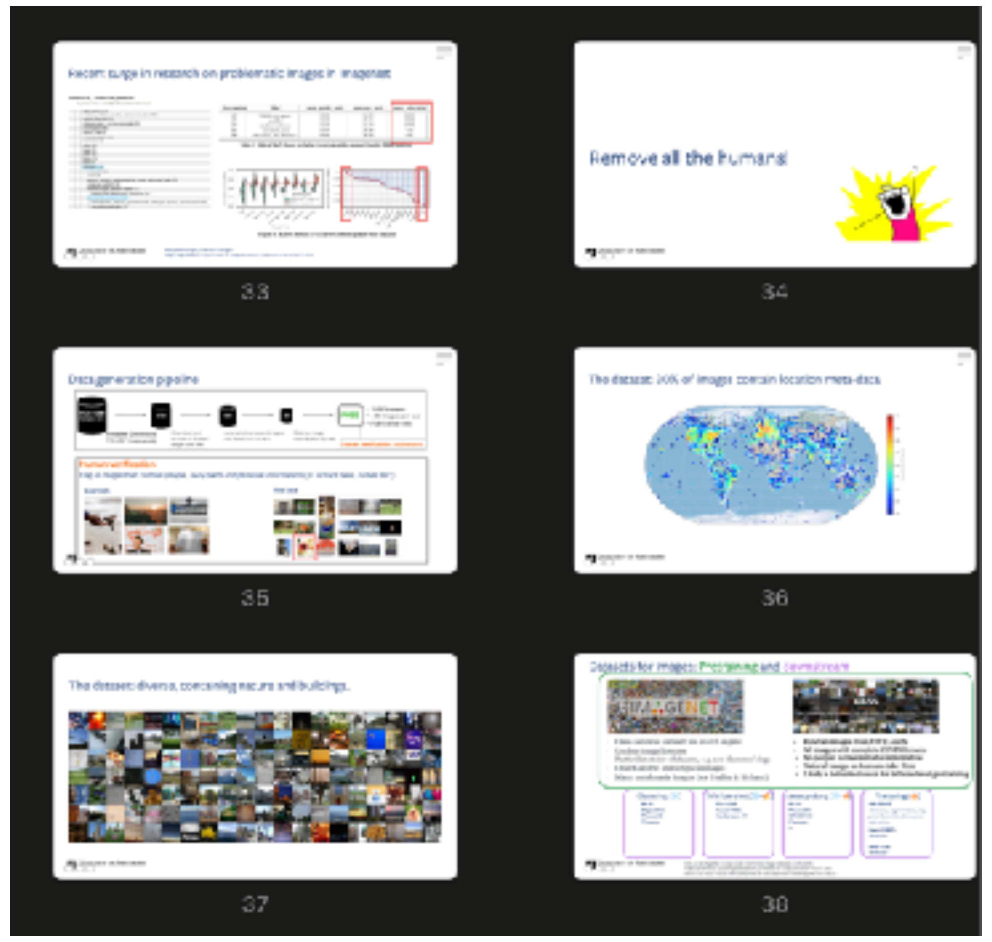
Summary



Why SSL?

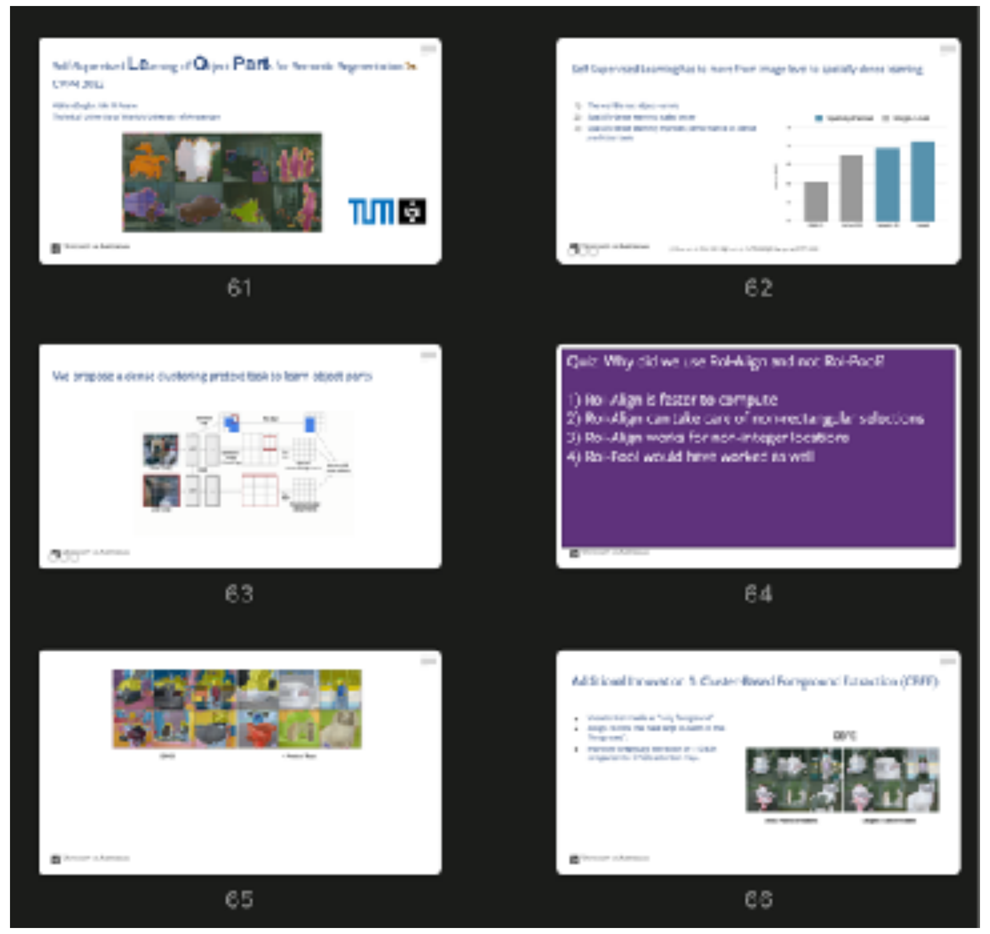


What is SSL?

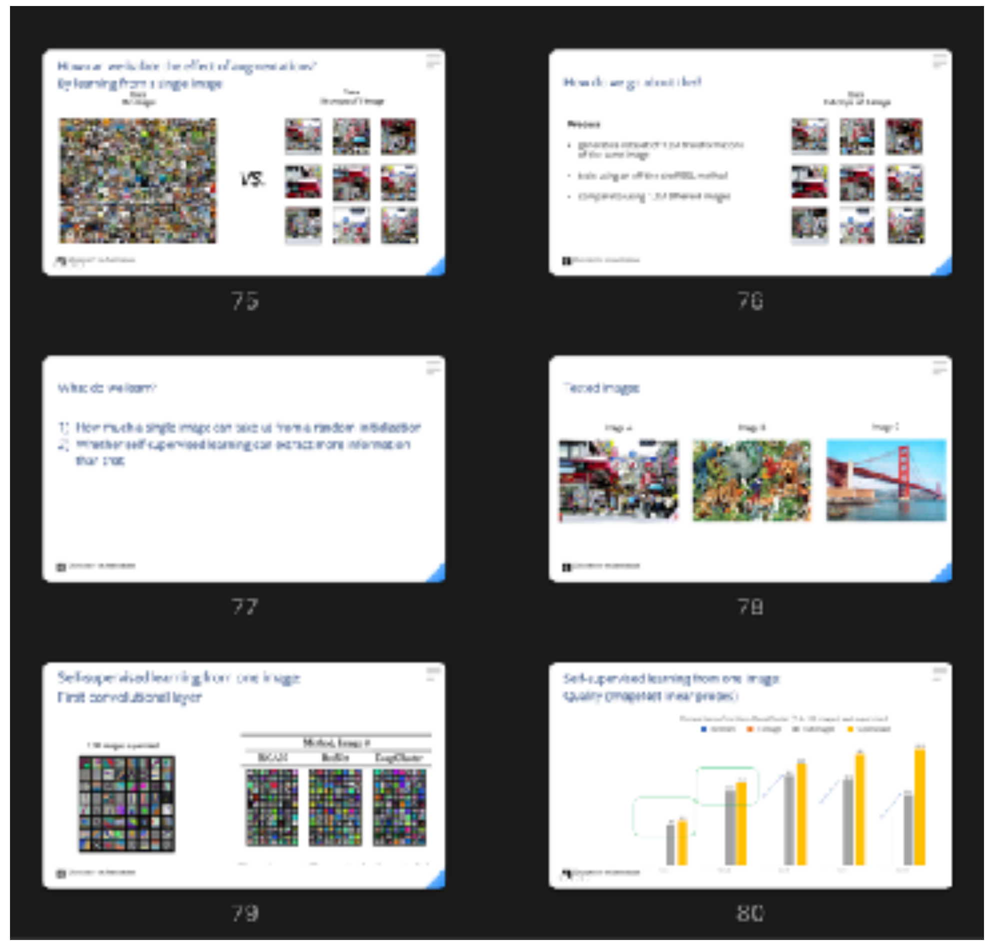


What kind of data? [1]

How SSL? (e.g. clustering [2])



SSL for segmentation [3]



Role of augmentations [4]

UNIVERSITY OF AMSTERDAM

[1] PASS: Pictures without humAns for Self-Supervised Pretraining. Asano et al. NeurIPS-Data'21.
[2] Self-labelling via simultaneous clustering and representation learning. Asano et al. ICLR 2020.
[3] Self-Supervised Learning of Object Parts for Semantic Segmentation. Ziegler & Asano. CVPR 2022
[4] A critical analysis of self-supervision, or what we can learn from a single image. Asano et al. ICLR 2020.

Please provide feedback for this lecture 🙏

https://evasys.uva.nl/evasys/public/online/index/index?online_php=&p=P5FET