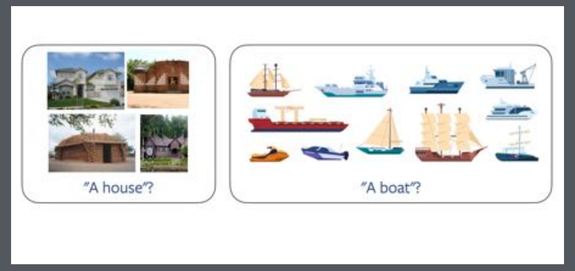
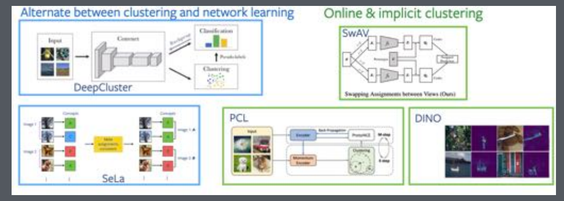
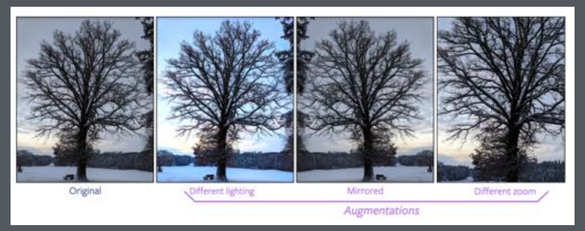
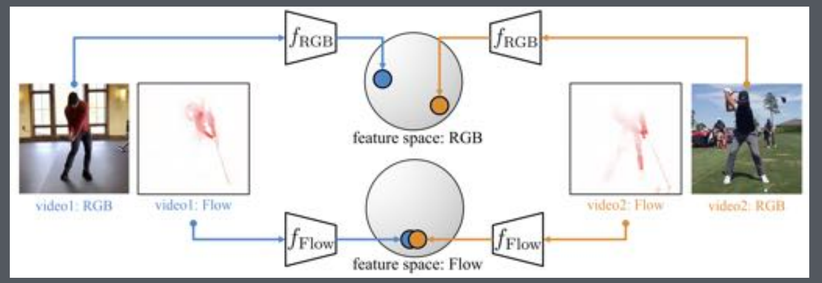
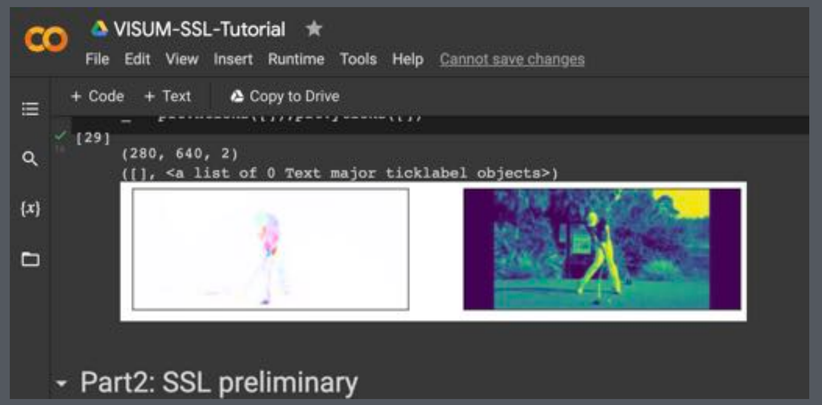
Self-supervised learning for computer vision from images, video and audio. The *why* & *how*.











@ VISUM SUMMERSCHOOL

YUKI M. ASANO, QUVA LAB MANAGER, ASSISTANT PROFESSOR, UNIVERSITY OF AMSTERDAM

TENGDA HAN, VISUAL GEOMETRY GROUP, UNIVERSITY OF OXFORD

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

Why do we want to do *SSL*?

How to do *SSL*?

Deep-dive into one image-*SSL* method

Why multi-modal data for *SSL*?

Brief intro into two video-*SSL* methods

Hi, I'm Yuki

- Assistant Professor with VISLab
  - Self-supervised Learning
  - Multi-modal Learning
  - Privacy and Bias in Computer Vision
  - Other Interdisciplinary Research
- Prior to this:
  - PhD at VGG in Oxford;
    Applied Mathematics/Physics/Economics at Oxford/Munich/Hagen
- Scientific Manager (with Prof. Snoek, Prof. Welling, Prof. Gavves)
  - Qualcomm-UvA (QUvA) Lab

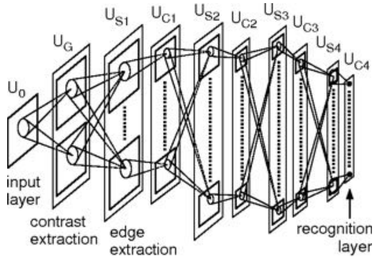
Introduction to self-supervised learning in computer vision

Part: "What"?

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

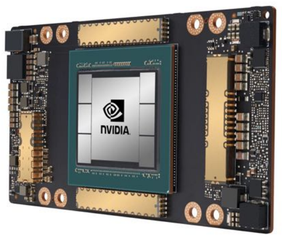
Algorithms

Deep neural networks



Hardware

GPUs



Data

Large scale datasets

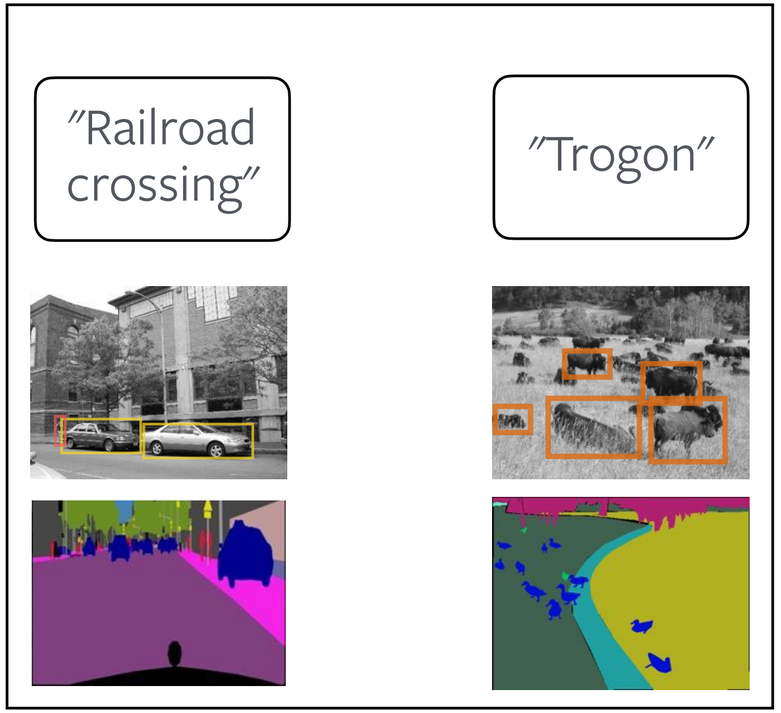
Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Fukushima, K., Biol. Cybernetics 1980
Object Recognition with Gradient-Based Learning. LeCun et al. Shape, Contour and Grouping in Computer Vision 1999
ImageNet: A Large-Scale 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

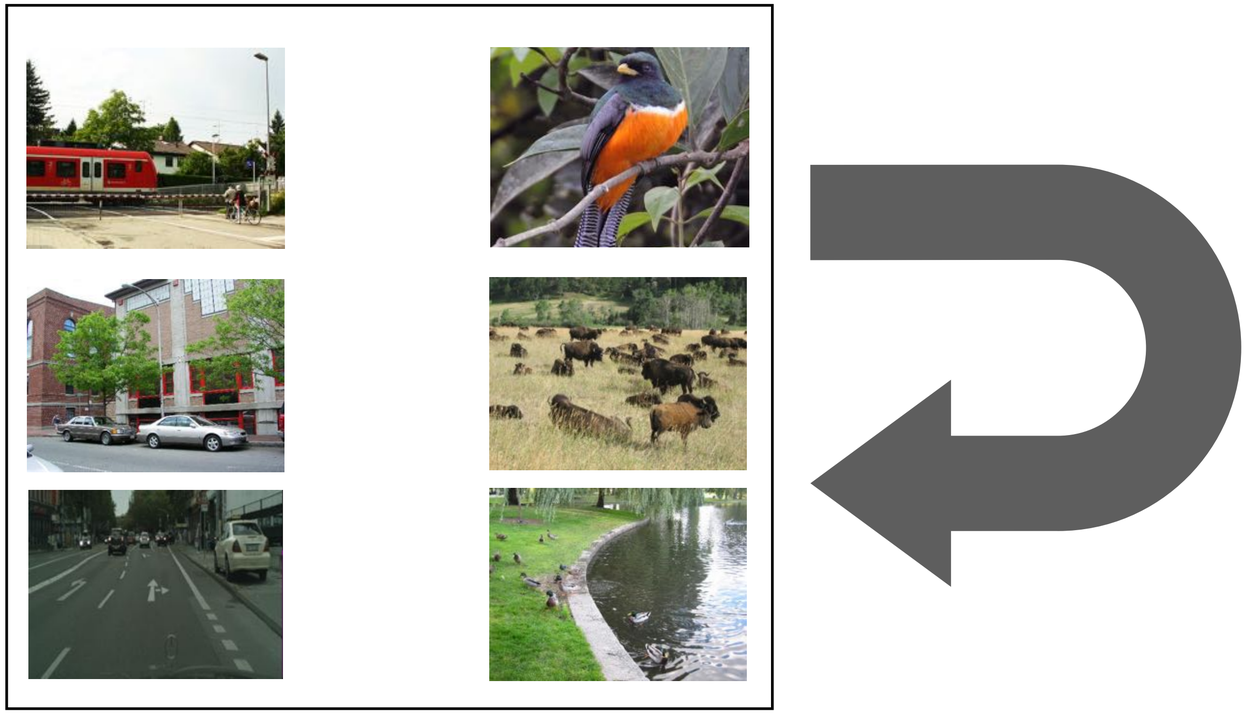
Supervised Learning

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.



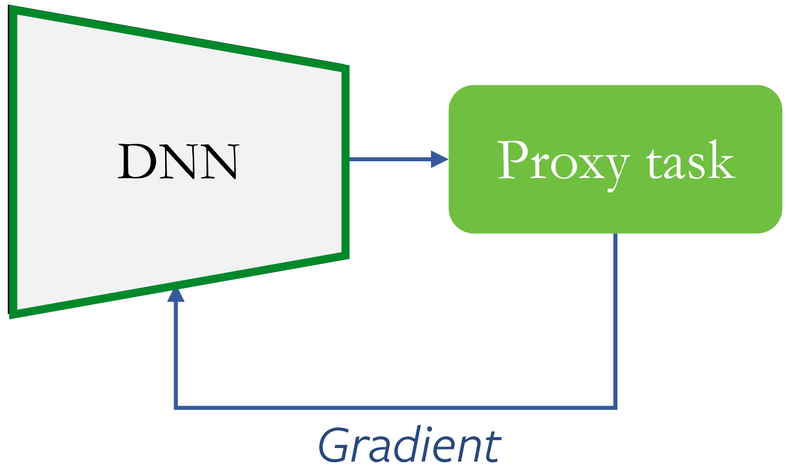
Self-supervision

Extract a supervisory signal from the raw data alone

General procedure of self-supervised learning.

Phase 1: Pretraining

Unlabelled data
+ transformations

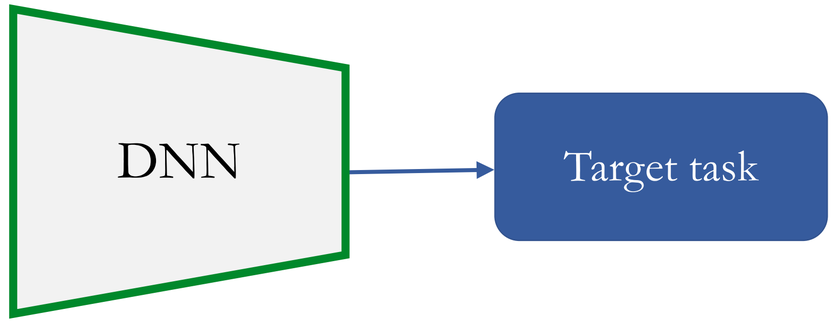


Types:

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

Phase 2: Downstream tasks

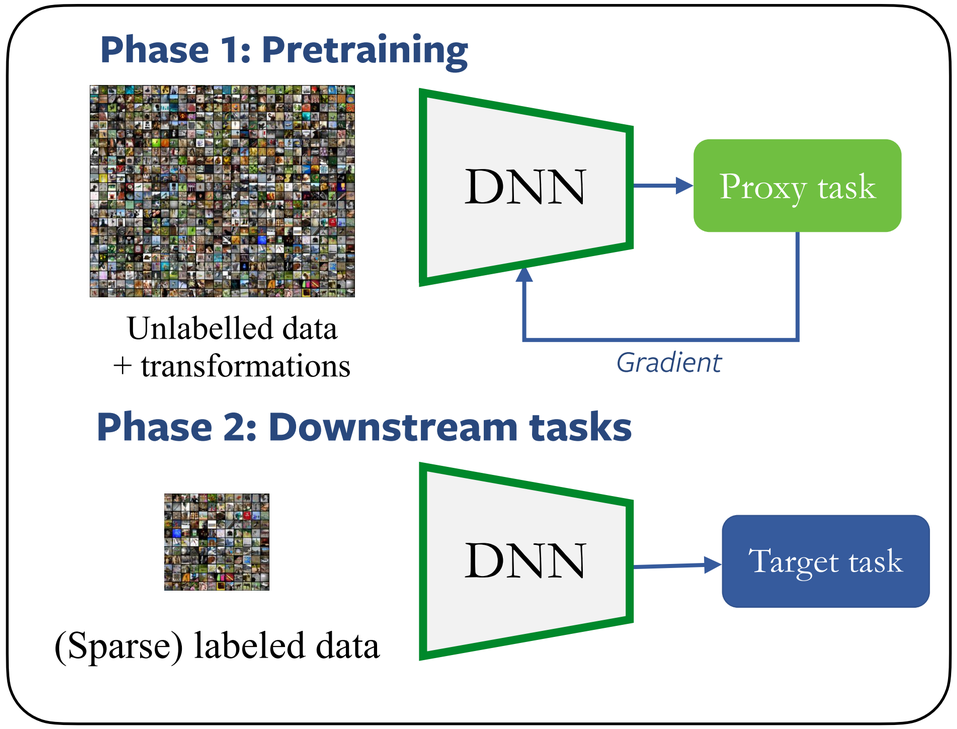
(Sparse) labeled data

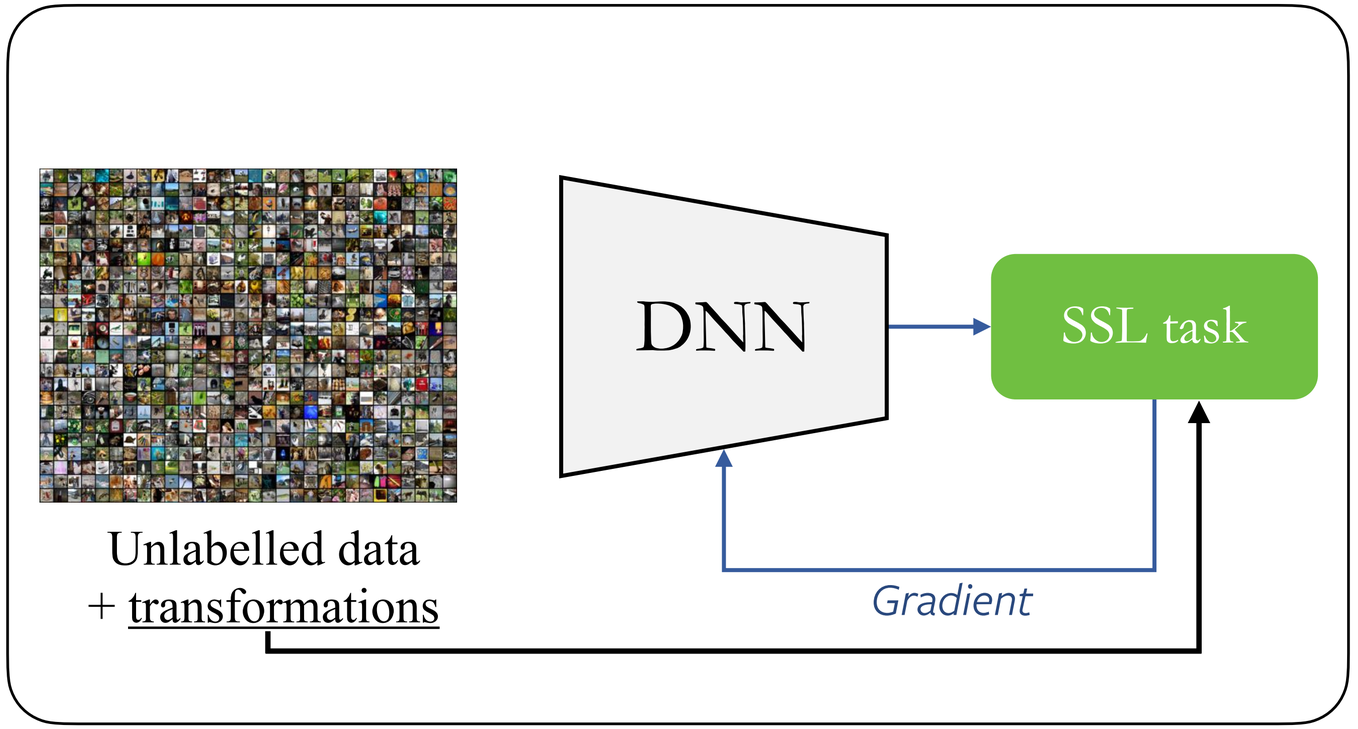


Types:

- 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.

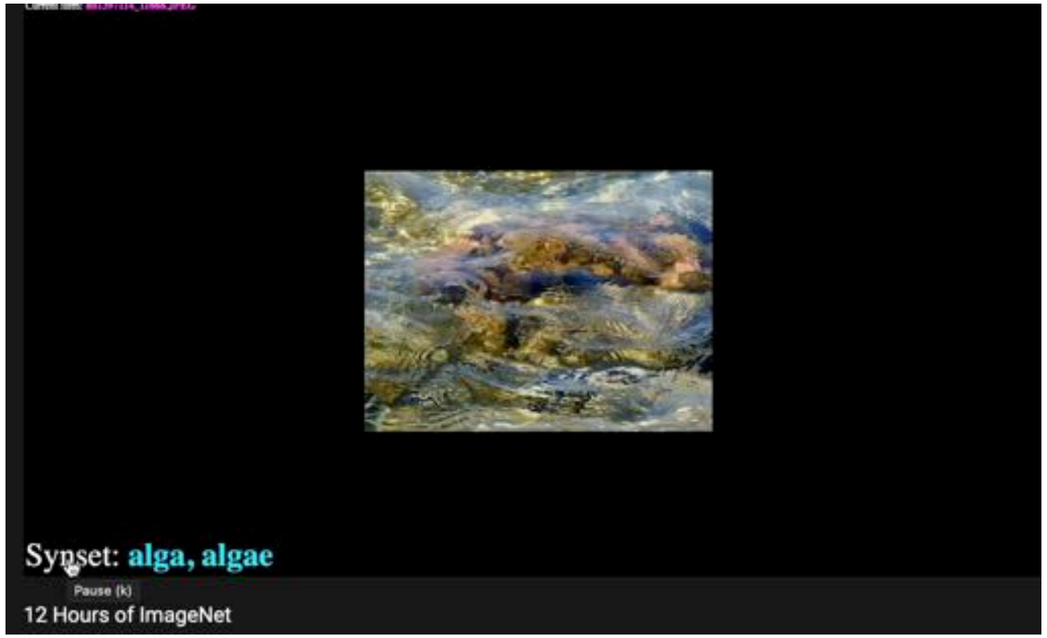
SSL object detection, SSL speaker detection,
SSL dataset labelling etc..

Introduction to self-supervised learning in computer vision

Part: "Why"?

Reason 1: Scalability

(above) x 50 =1.2M images

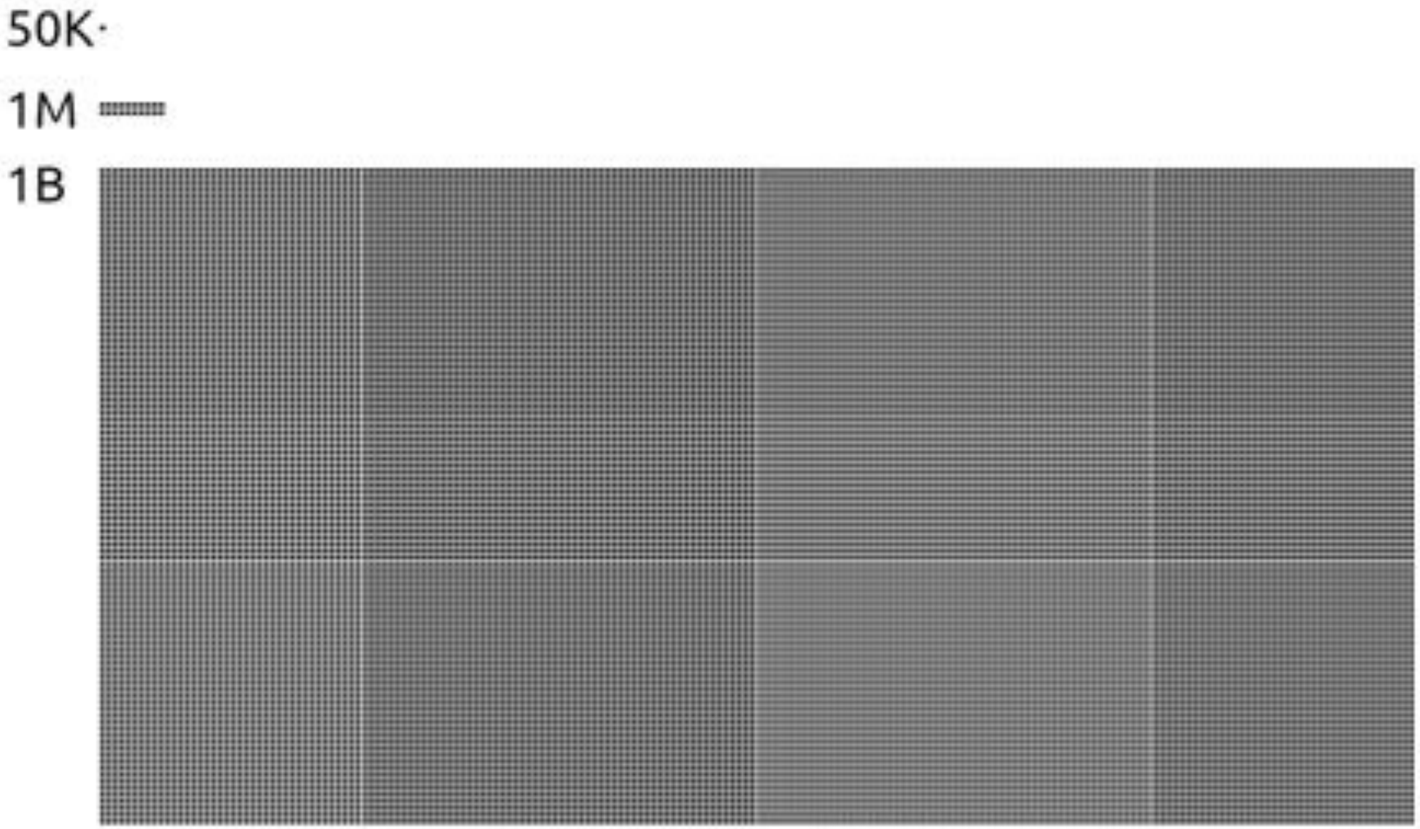


90ms * 1.2M = 30h

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

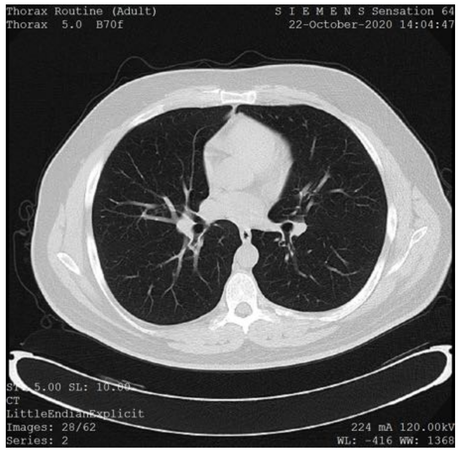


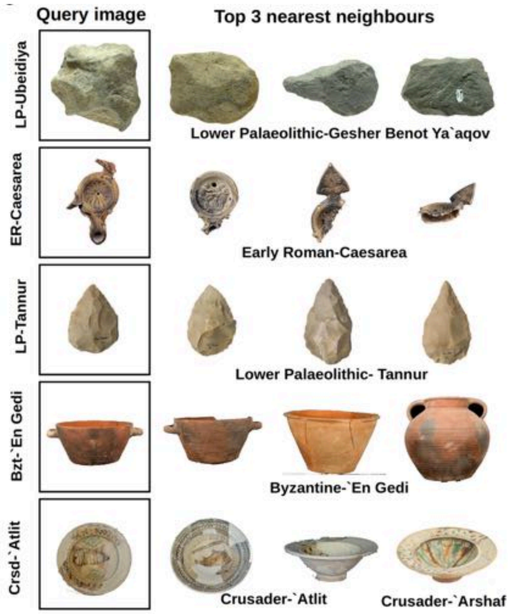
Annotation is expensive, yet datasets keep getting bigger.

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.

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 4: Ambiguity of labels

"A house"?

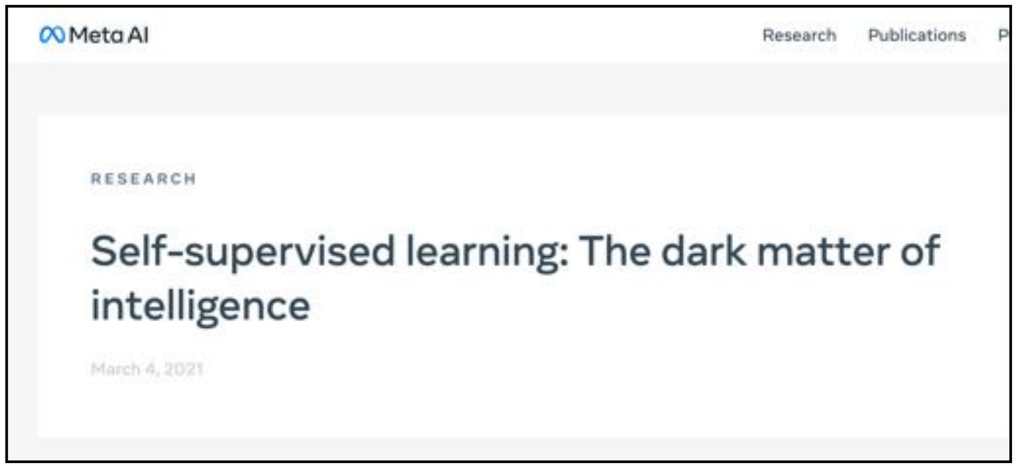
"A boat"?

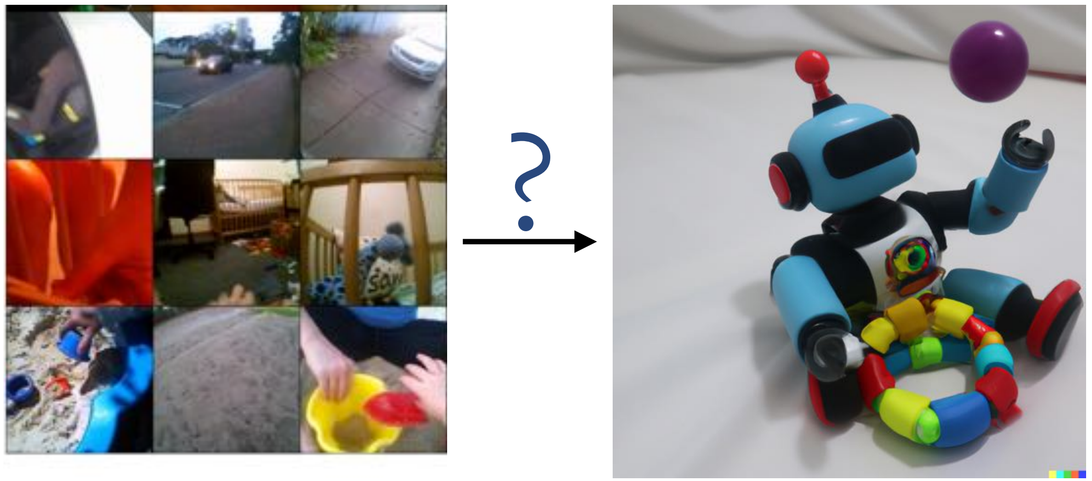
Nonsensical visual labels

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 5: Investigating the fundamentals of visual understanding





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.

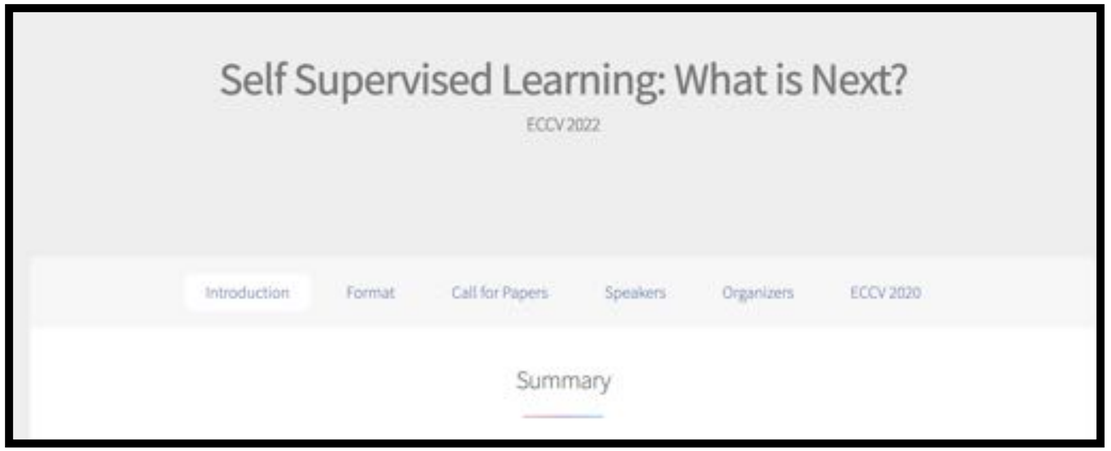
What, if there are, are the limits of learning without labels?

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

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

Here, we will only cover a fraction of the existing literature. Further details and recent developments can be found here:

CVPR'21 Tutorial by Bursuc et al.



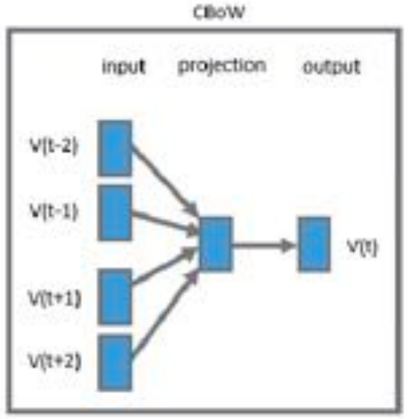
ECCV'20/22 workshop by Asano et al.

This October!

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

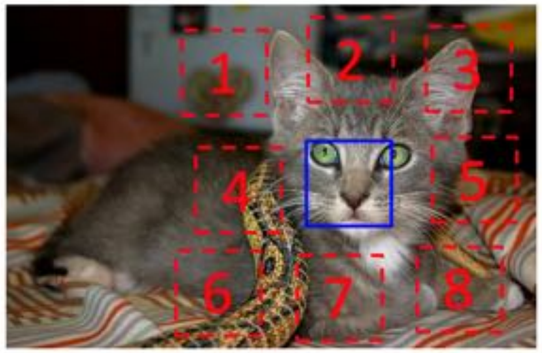
Early methods

Word2Vec

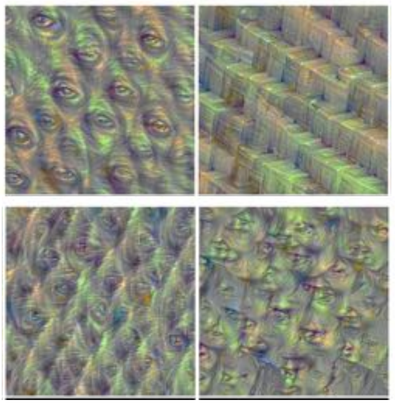


Motivated from NLP

Context Prediction



X = (,); Y = 3

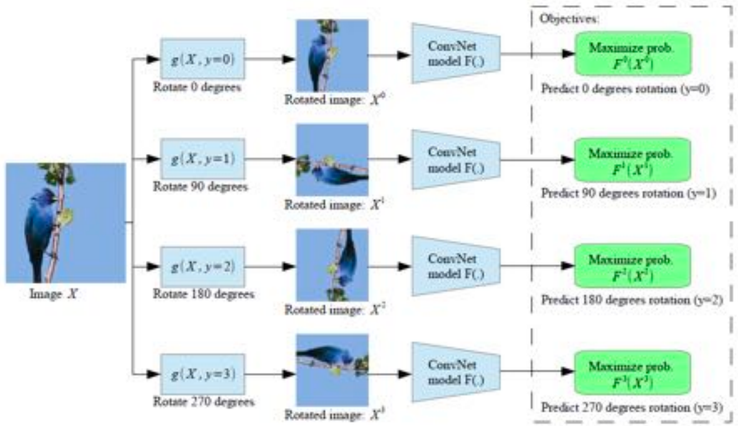


Context Encoders

(a) Input context

(c) Context Encoder
($L2$ loss)

RotNet



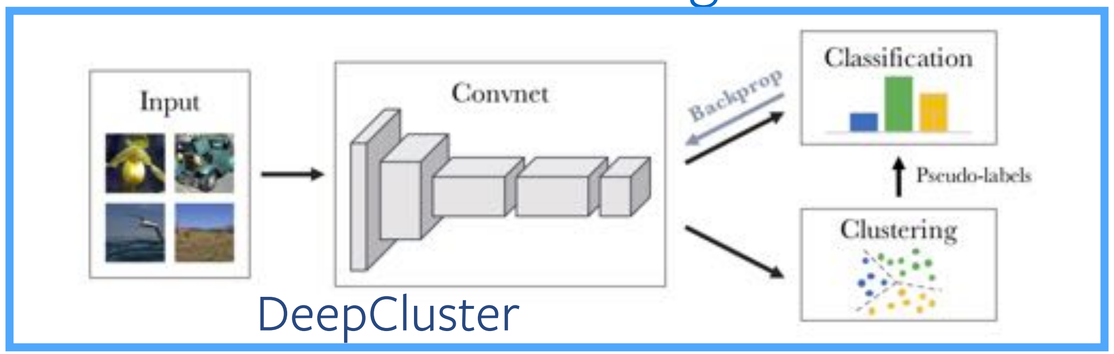
Learning without labels is meaningful and possible.

https://www.researchgate.net/figure/Word2Vec-CBOW-and-Skip-gram-There-are-two-different-methods-in-the-Word2Vec-algorithm_fig2_320829283

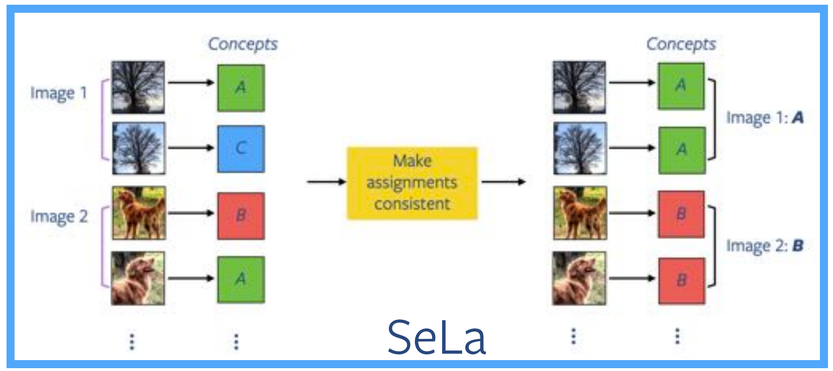
Doersch et al. *Unsupervised Visual Representation Learning by Context Prediction.* ICCV 2015.
Pathak et al. *Context Encoders: Feature Learning by Inpainting.* CVPR 2016.
Gidaris et al. *RotNet: Unsupervised Representation Learning by Predicting Image Rotations.* ICLR 2018

Clustering

Alternate between clustering and network learning

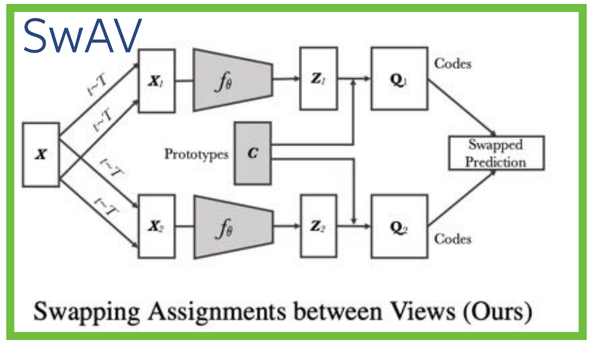


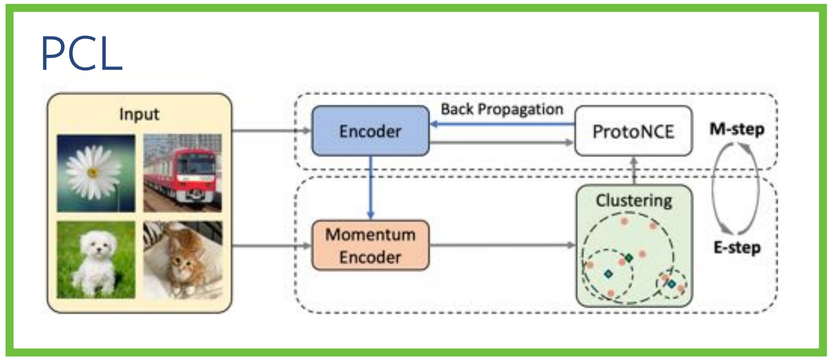
DeepCluster

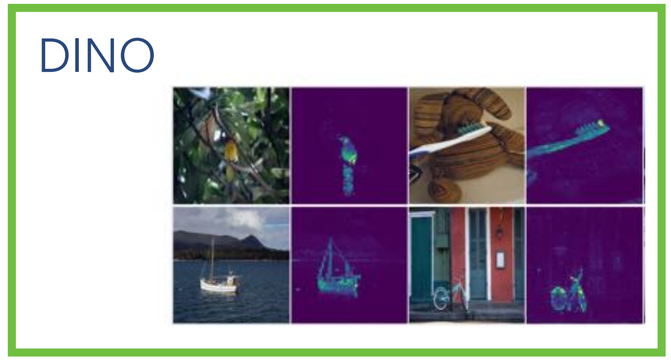


SeLa

Online & implicit clustering



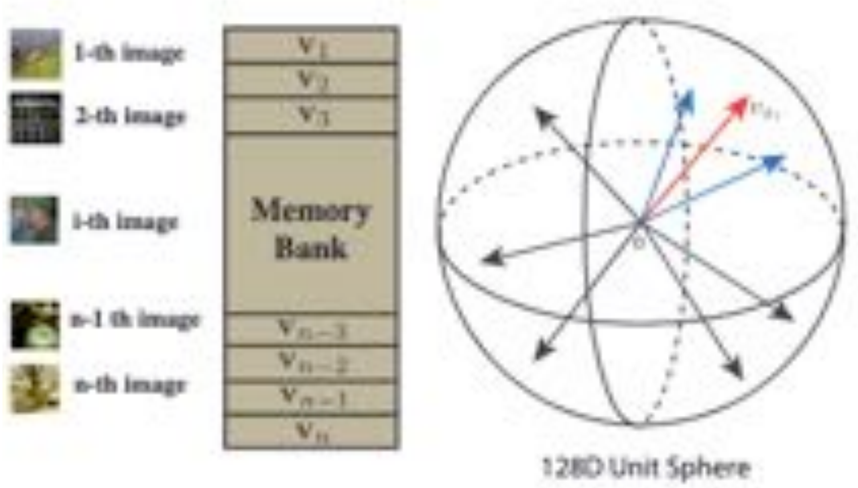




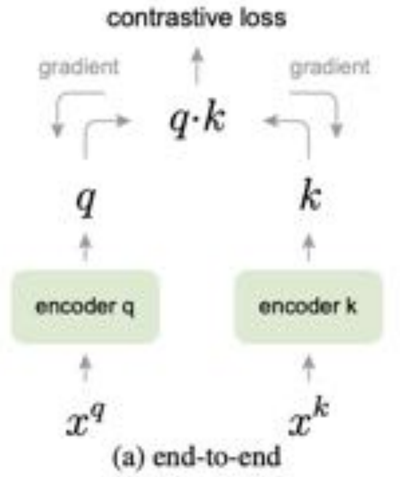
Clustering is a strong pretext task and serves a useful purpose (~labelling/categorizing).

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

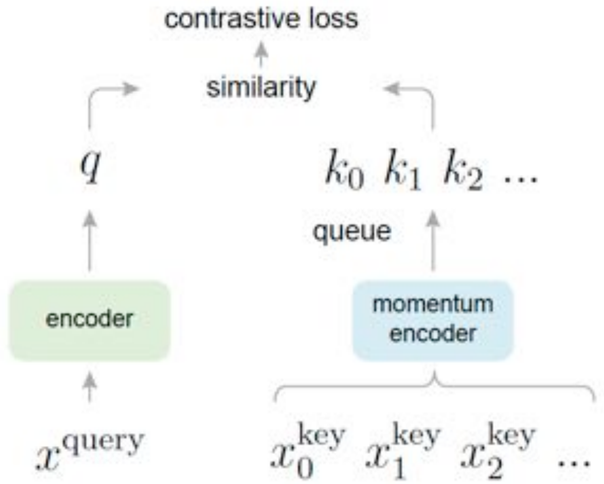
Noise-contrastive self-supervised learning



NPID



SimCLR



MoCo

Momentum encoder:

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

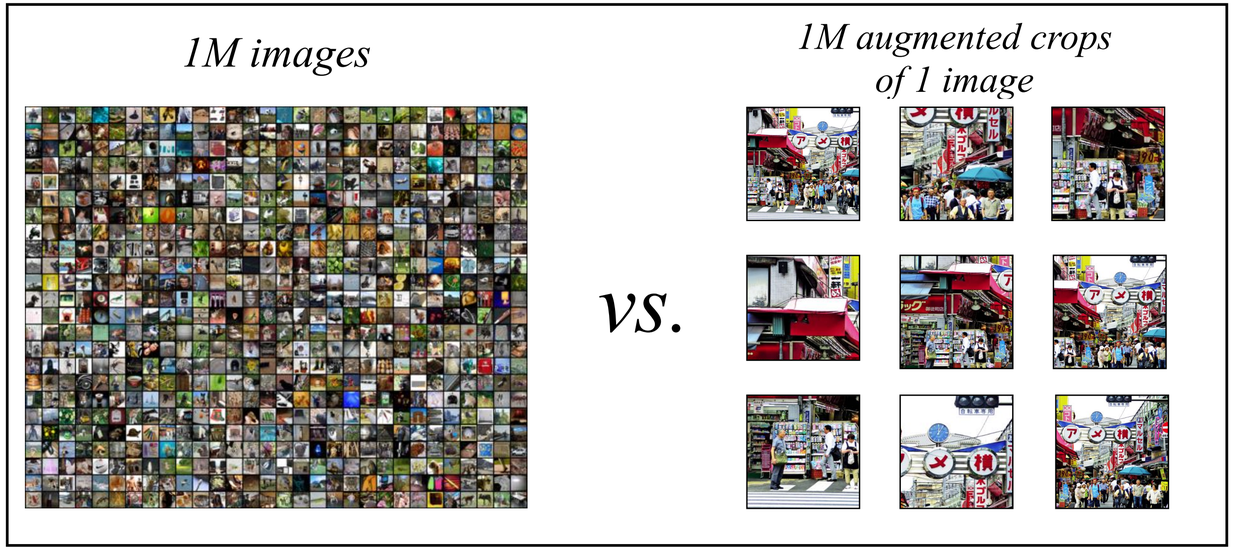
The start of large-scale & industrial self-supervised learning.
These works heavily rely on image augmentations.

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

How does one implement contrastive learning?

Learn it at the tutorial!

About image-augmentations for self-supervised learning.



Downstream task: object detection, 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

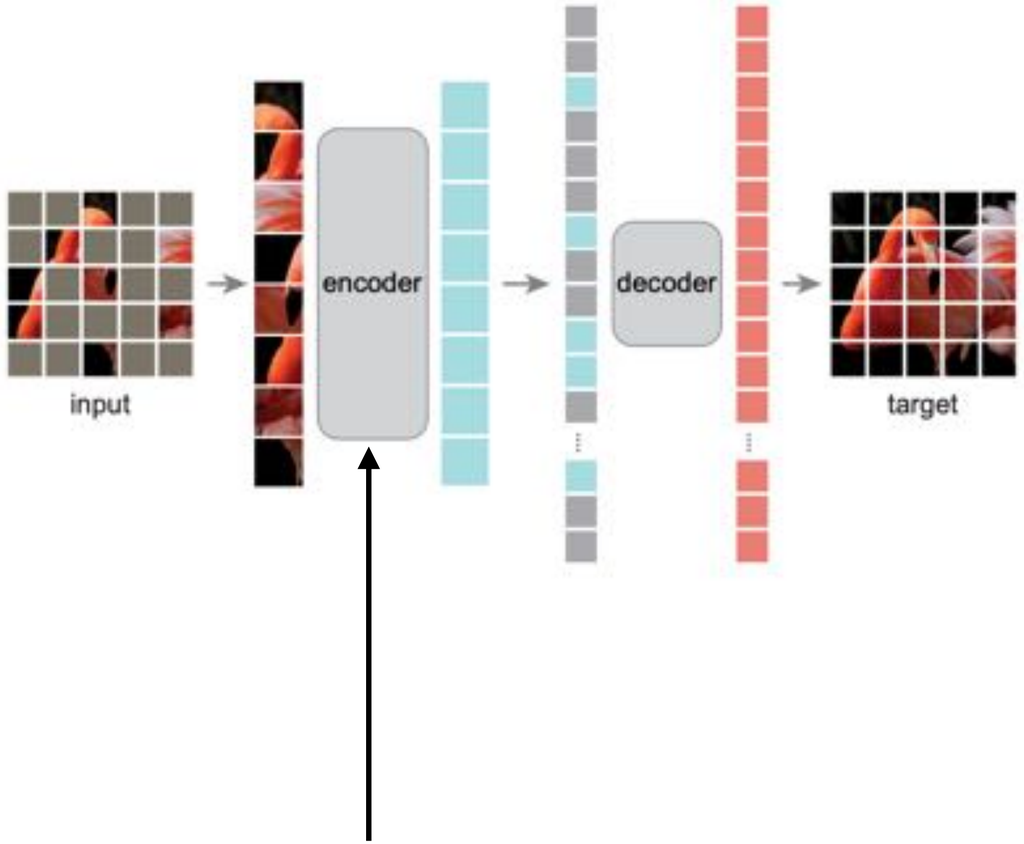
Within 3% of MoCo-v2 on full ImageNet

Method, Image A		
BiGAN	RotNet	DeepCluster
		
20.4	19.9	20.7

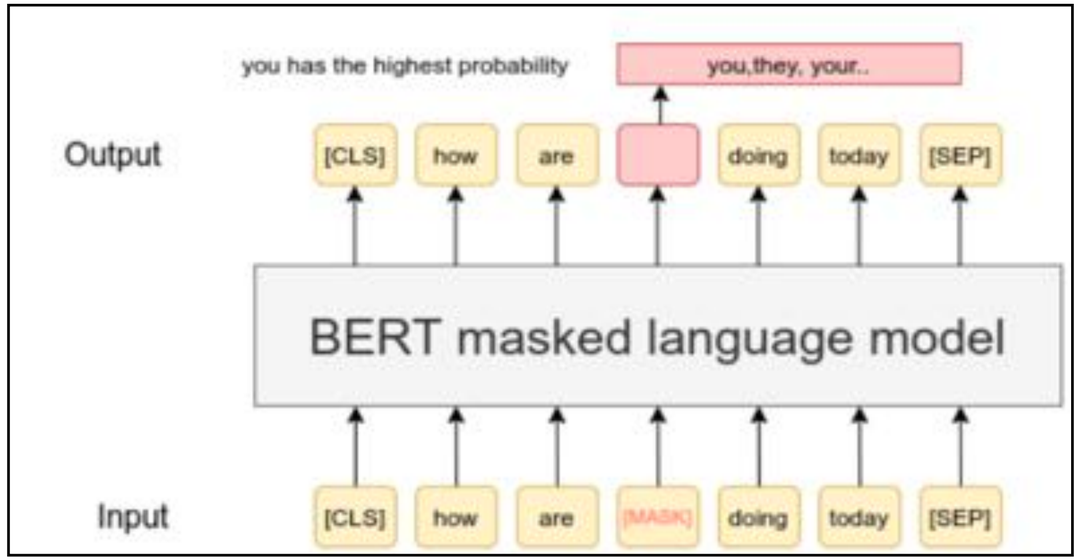
Self-supervised learning's "learning": mostly from augmentations.

Asano et al. *A critical analysis of self-supervision, or what we can learn from a single image.* ICLR'20

Masked Image Modelling



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

Datasets: 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

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

Generalizeable representations and `free labelling' of images

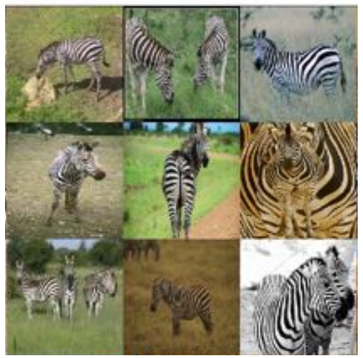
Self-labelling via simultaneous clustering and representation learning (ICLR'20 spotlight)

YUKI M. ASANO, CHRISTIAN RUPPRECHT, ANDREA VEDALDI

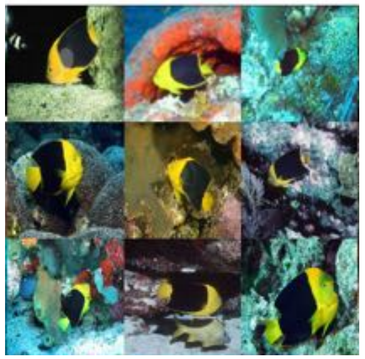
Goal: Discover visual concepts without annotations.

(above) x 50 =1.2M images

?



concept ″A″



concept ″Z″

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

The key to image understanding is separating meaning from appearance.

Original

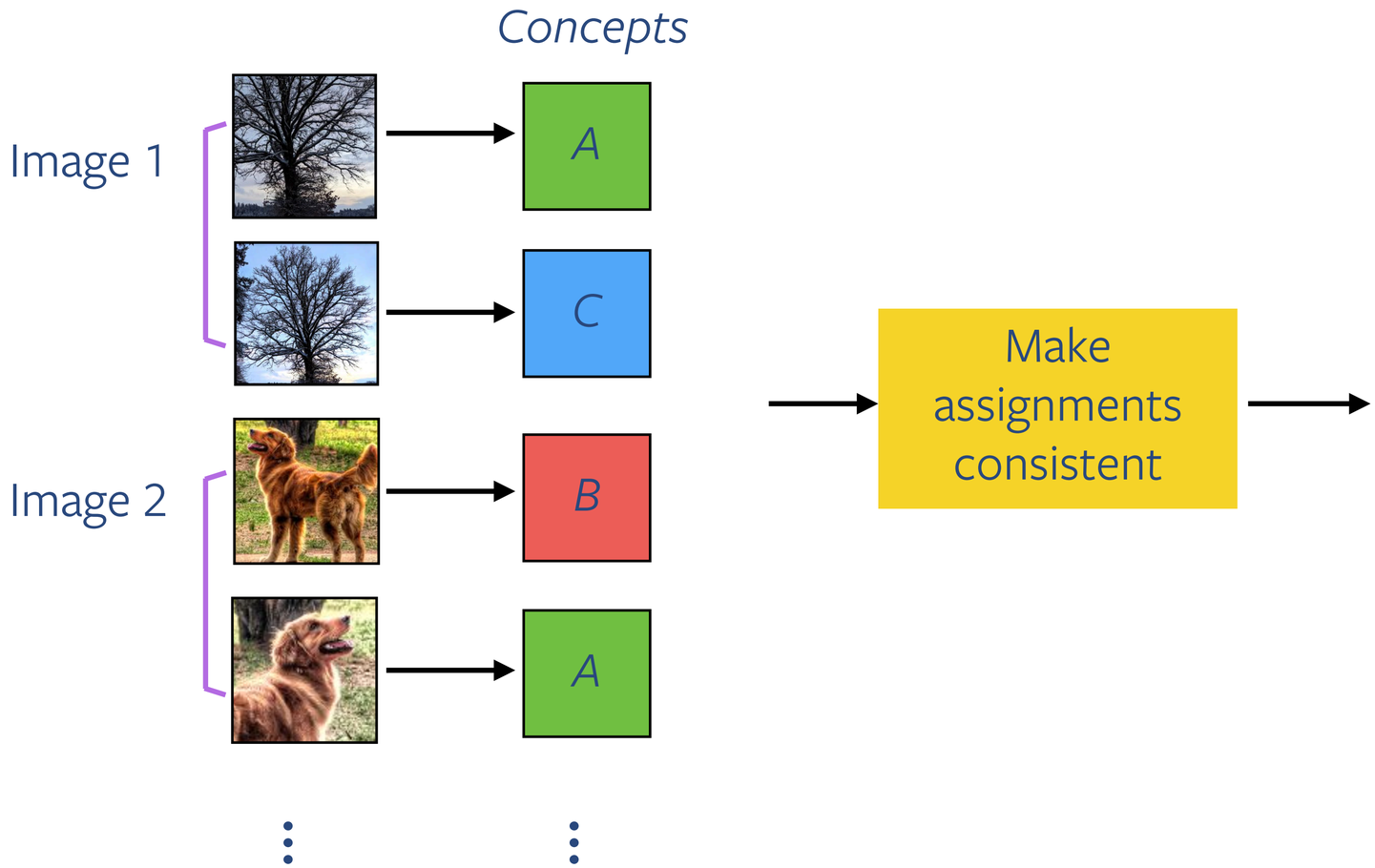
Different lighting

Mirrored

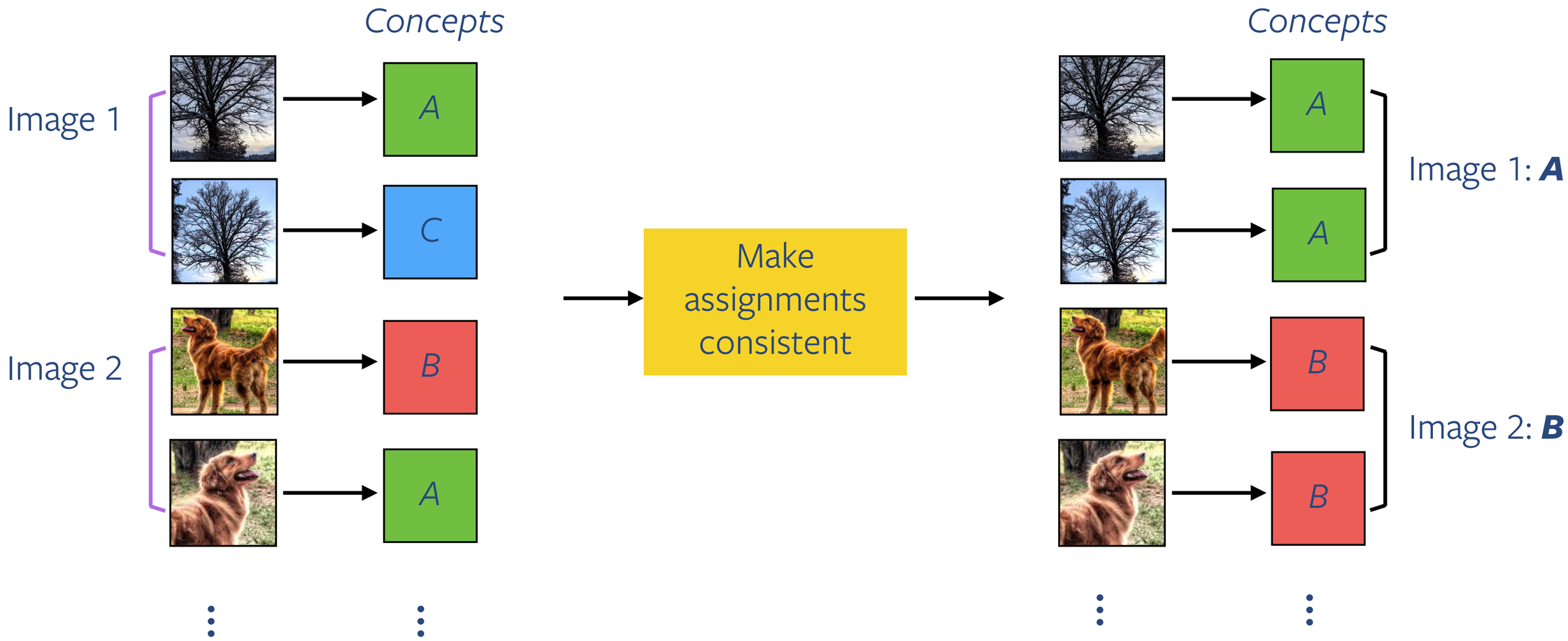
Different zoom

Augmentations

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

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

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

- 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:

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

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

2. Use pseudolabels an equal number of times: 

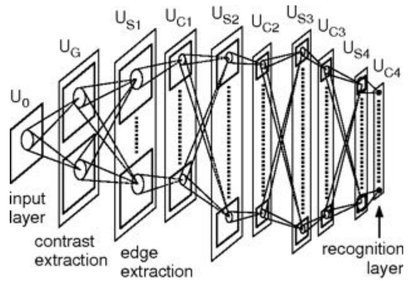
3. Pose as approximate optimal transport:

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

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

Algorithm

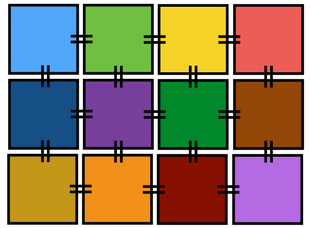
Model Φ



Optimal labelling

Cross entropy training with augmentations

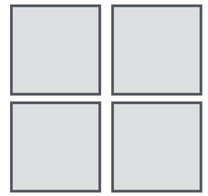
Label assignments q



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

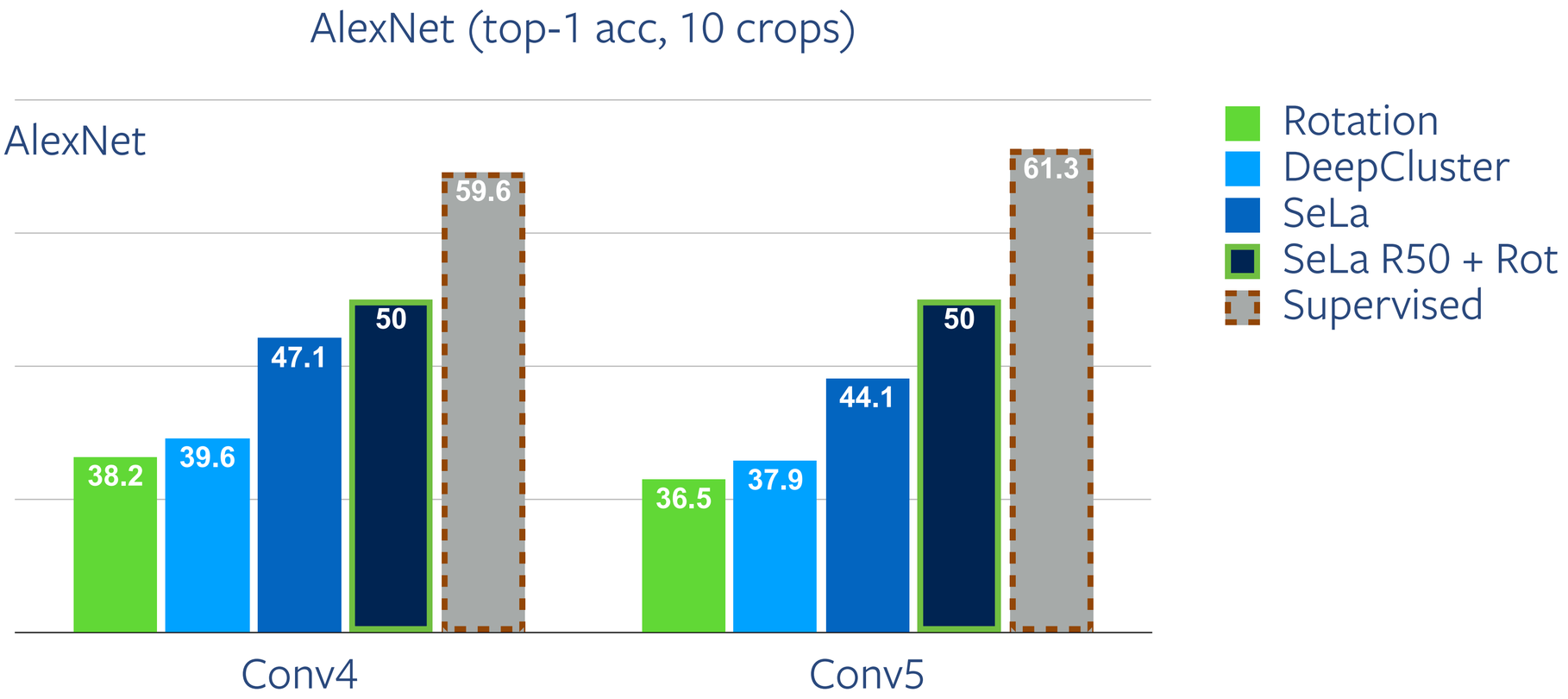
Manually annotated label (>2.5y of work)

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

- 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.

More recently...

Method	Top-1 2x224	Top-1 2x160+4x96	Δ
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

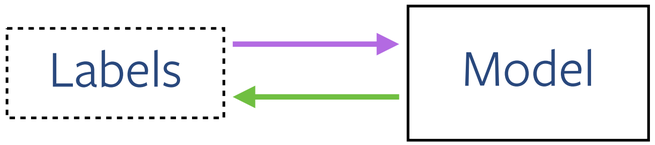
- DINO with SeLa's SK: same performance.

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

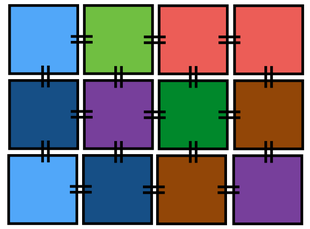
Self-supervised labelling from three core ideas

Invariance to augmentations

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

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

Multi-modal SSL: leveraging video & audio for labelling

The key to image understanding is separating meaning from appearance.

Original

Different lighting

Mirrored

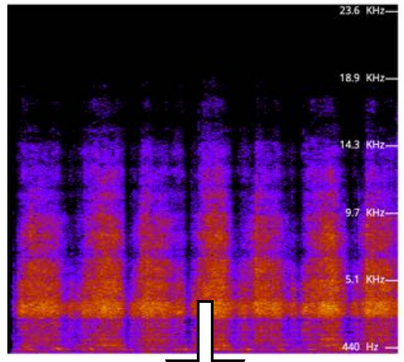
Different zoom

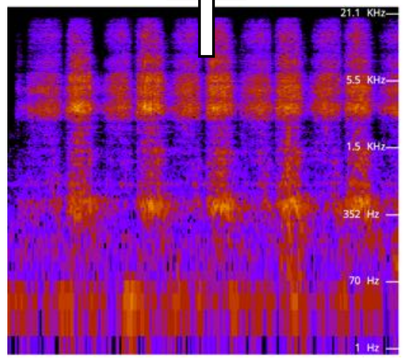
Augmentations

Multiple modalities can yield useful semantic information.

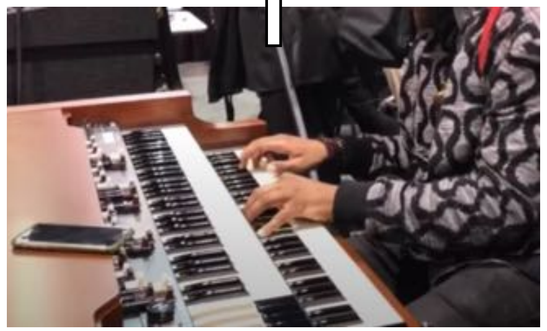
Video

Audio

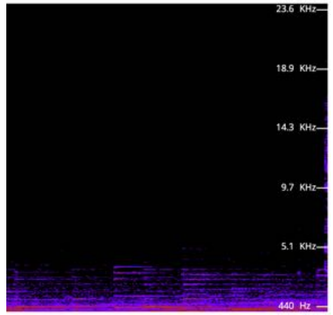


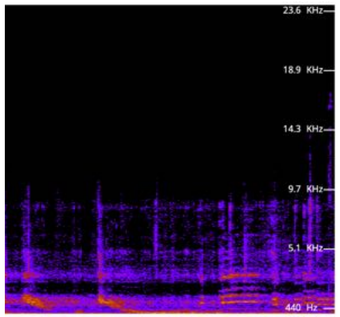


Video



Audio





https://www.youtube.com/watch?v={ZlS8t8_nAHo, tvVYdaykaZM, Kt8_u_i-anQ, J4X6tc1HYts}

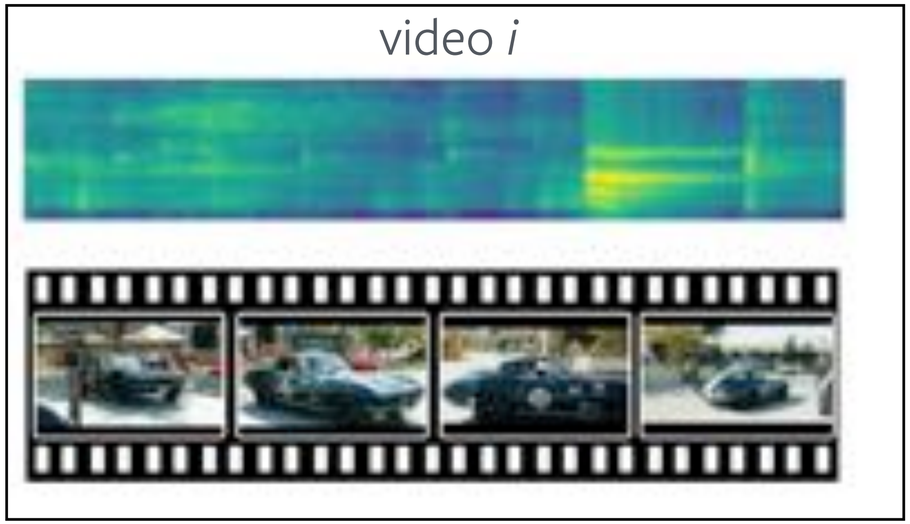
Why clustering for videos?

🤔 Clustering works well for images

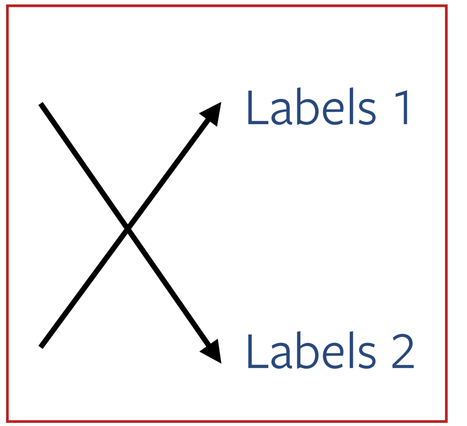
🤑 Videos are expensive to annotate.

📈 Video content is rapidly increasingly.

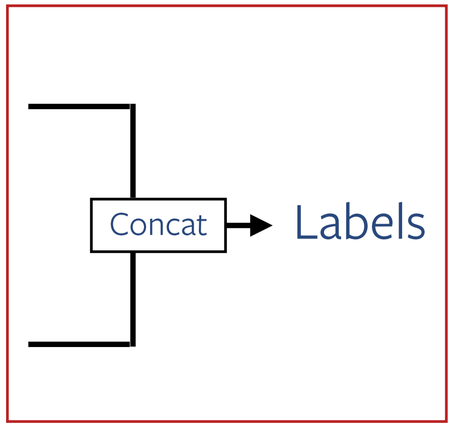
Clustering multi-modal data



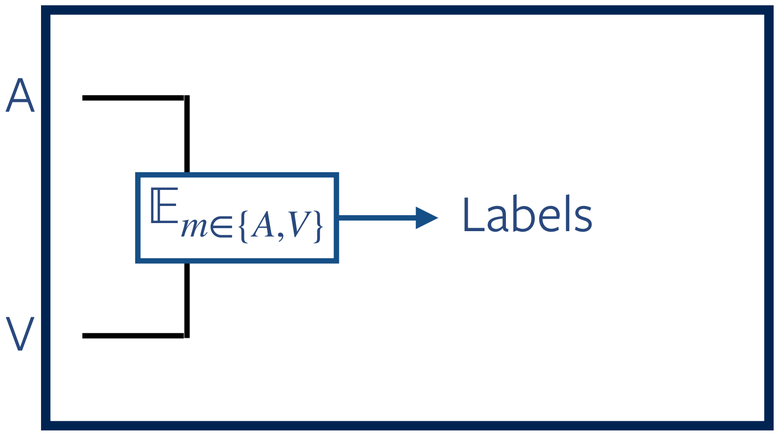
× does not use same-source information
× two different sets of clusters



× two different sets of clusters
× hard to interpret



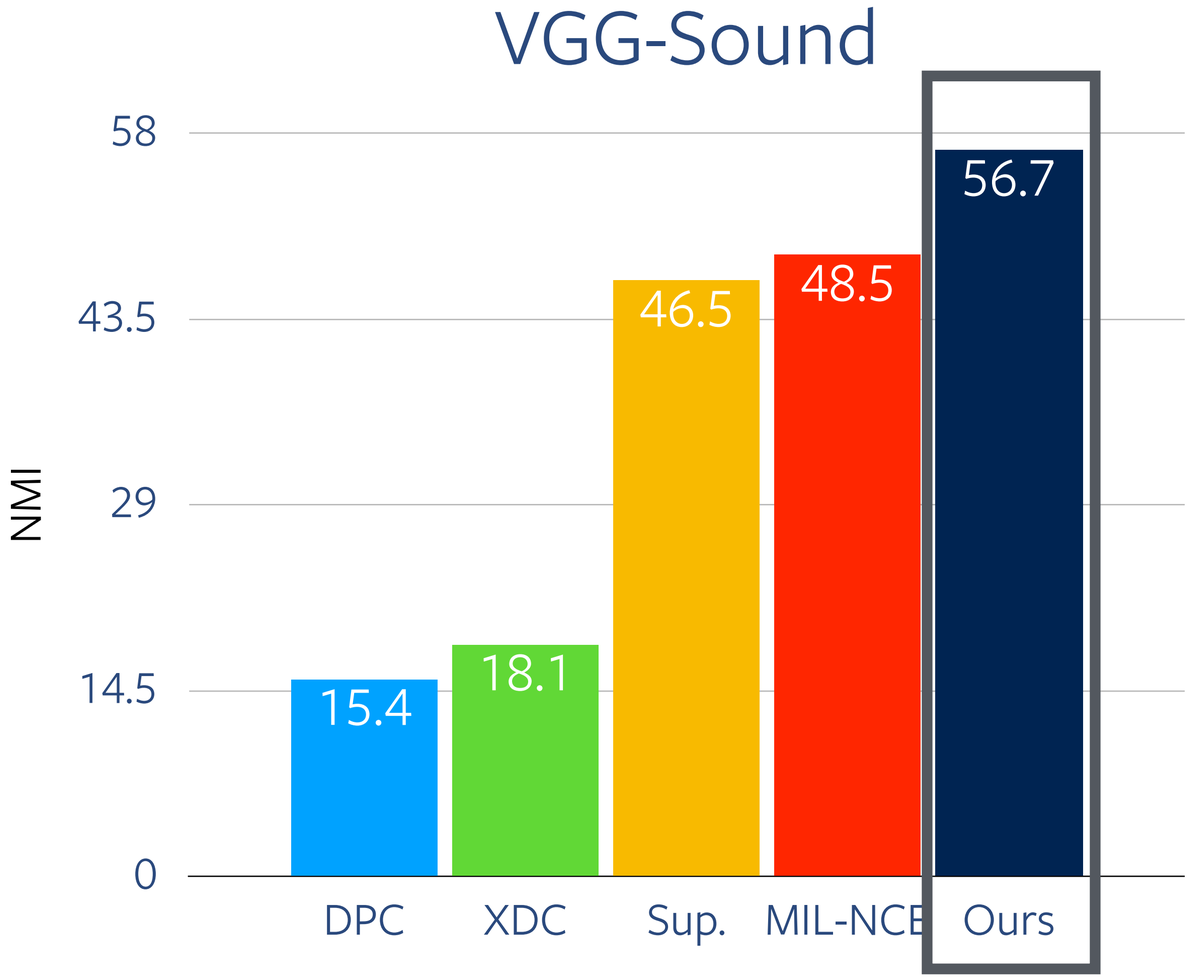
× concatenation can just rely on stronger modality and ignore the other

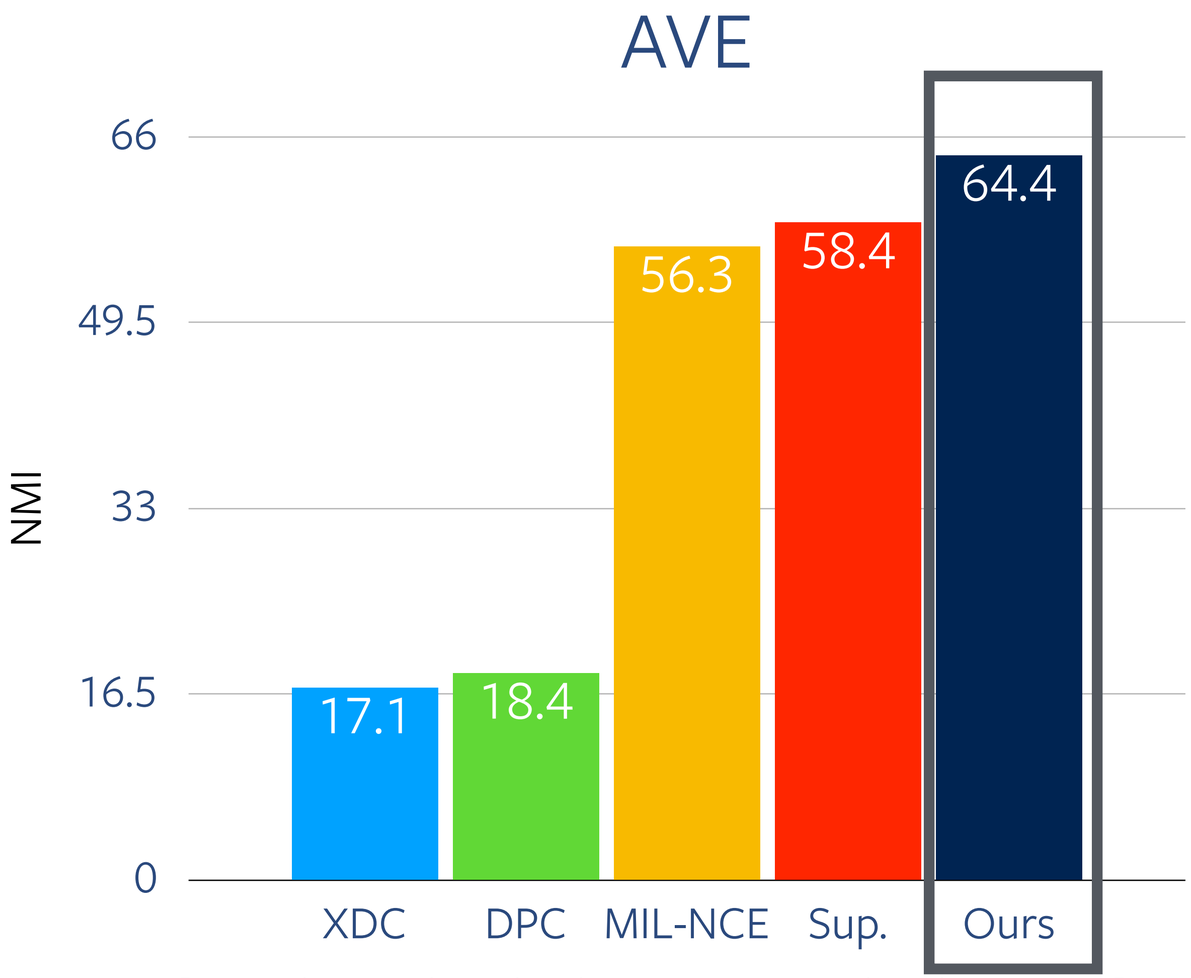


Our idea:
view each modality as an *augmentation*

Asano et al. *Labelling unlabelled videos from scratch with multi-modal self-supervision.* NeurIPS 2020

Good feature representations ↛ good clustering

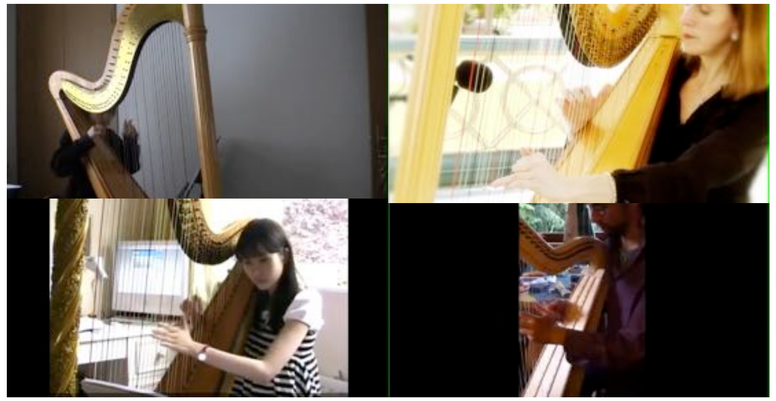




[Han et al., arXiv; Alwassel et al., NeurIPS 2020; Miech et al., CVPR 2020]

Asano et al. *Labelling unlabelled videos from scratch with multi-modal self-supervision.* NeurIPS 2020

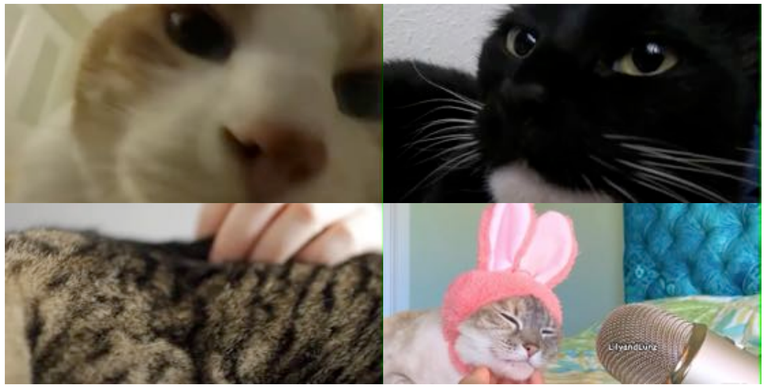
Discovering concepts without manual annotations from 230K videos



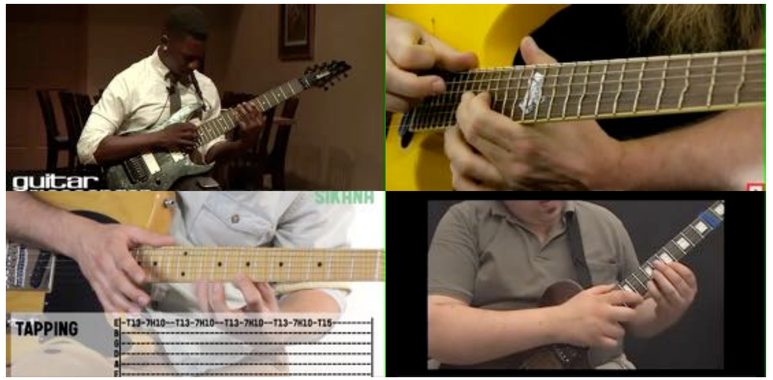
"Playing harp"

"Fireworks"

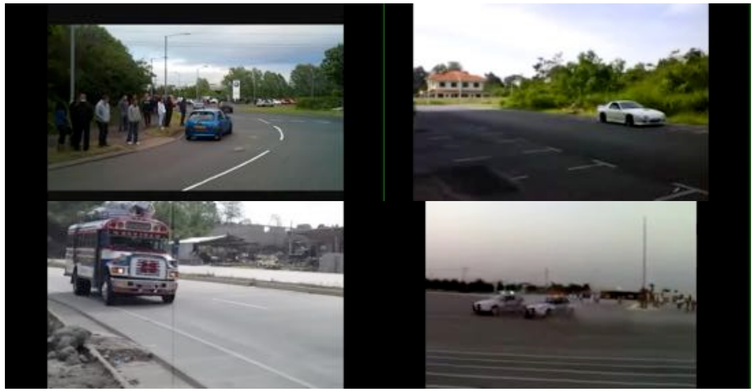
"Hockey game"



"Cat growling"



"Electric guitar"

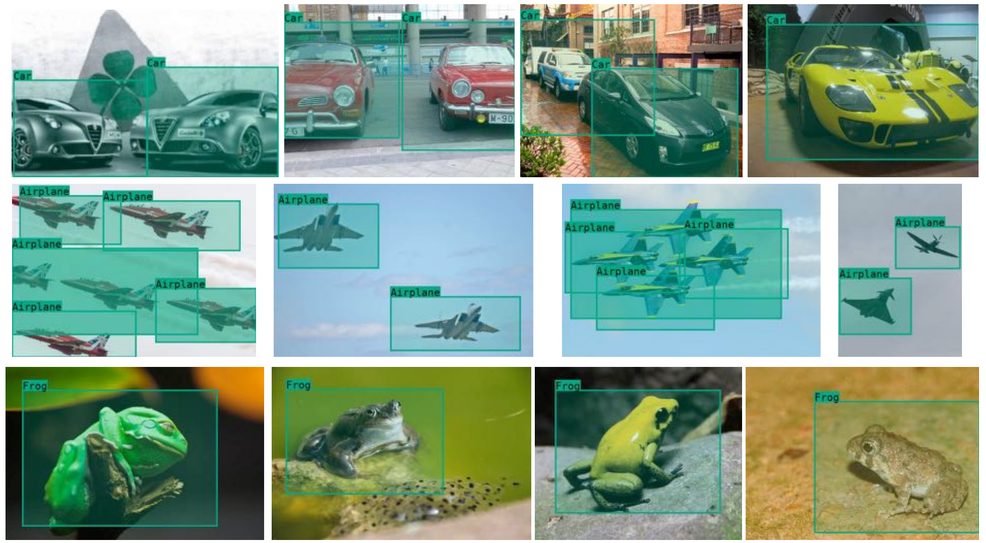


"Vehicle driving"

Asano et al. *Labelling unlabelled videos from scratch with multi-modal self-supervision.* NeurIPS 2020

"Self-Labelling" videos: three main findings

- Clustering framework of SeLa is well-suited
- Generalizeable to model any distribution (Zipf/exponential etc.)
- Good feature representations do not imply good clustering



Follow-up work:
using this + contrastive learning, we can detect objects self-supervisedly.

Asano et al. Labelling unlabelled videos from scratch with multi-modal self-supervision. NeurIPS 2020
Afouras et al. *Self-supervised object detection from audio-visual correspondence.* CVPR'21

On Compositions of Transformations in Contrastive Self-Supervised Learning

Mandela Patrick*, Yuki M. Asano*, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, Andrea Vedaldi
ICCV'21

Invariance *vs* distinctiveness

In contrastive learning, we define positives and negatives.

Should the representations enforce invariance or distinctiveness?

?

=

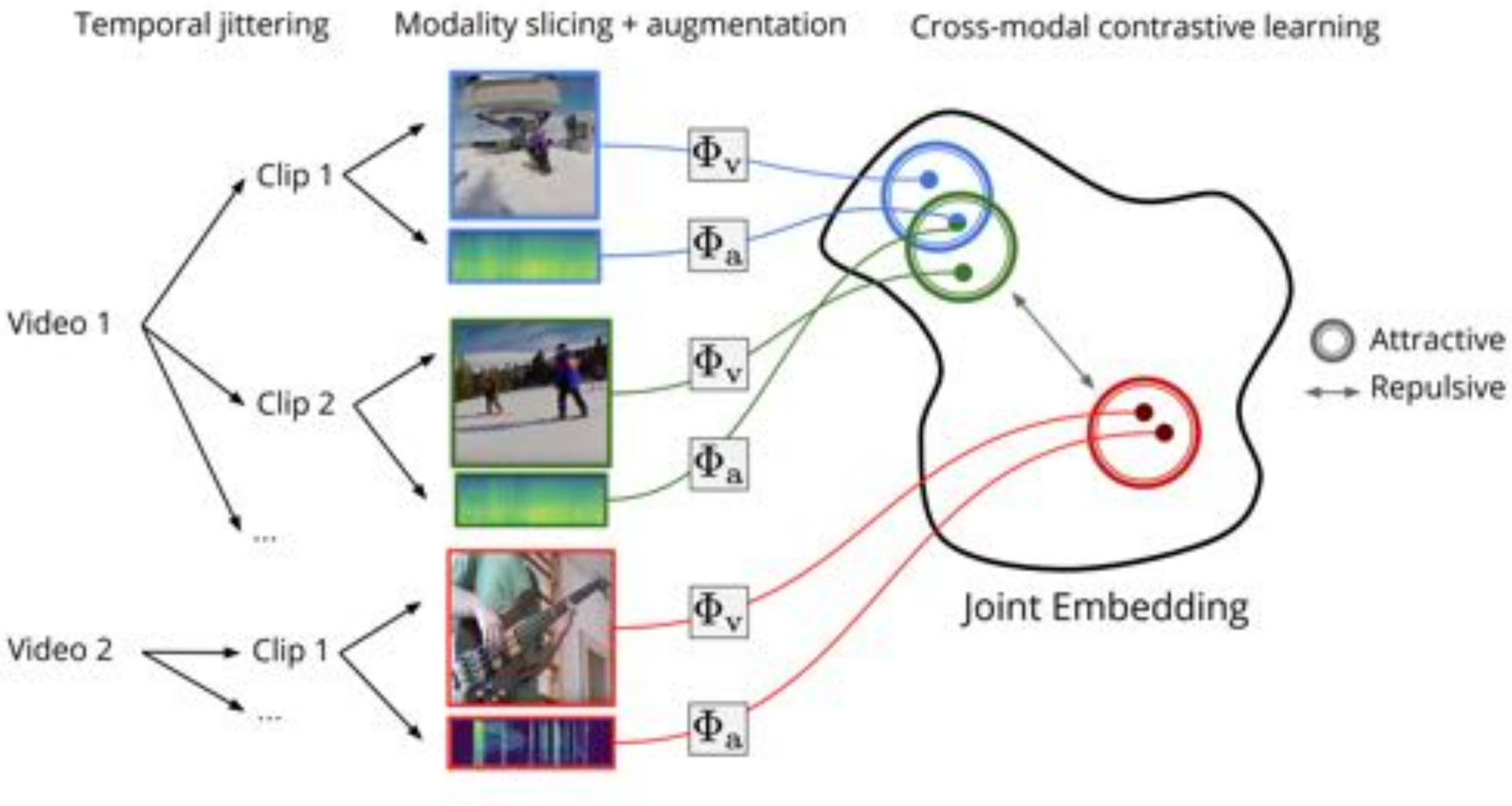
Learning hypotheses we test:

1. Sample Distinctiveness

2. Time Reversal

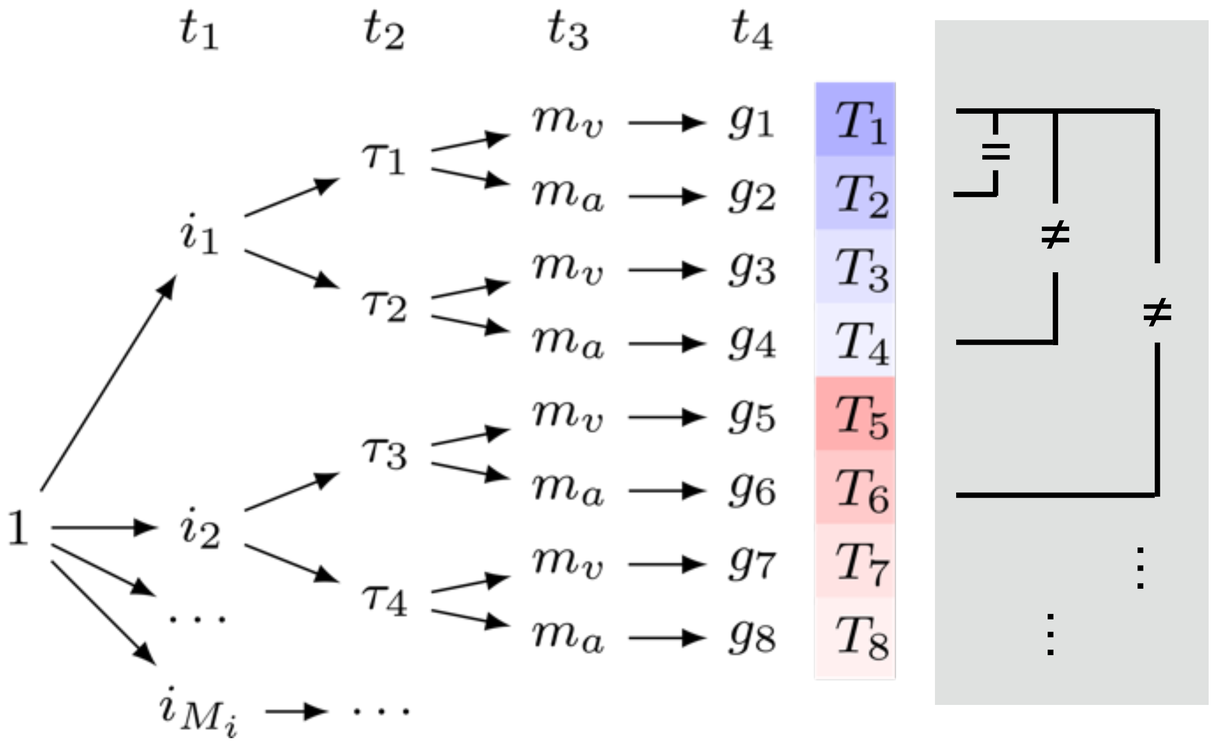
3. Time Shift

Framework



Patrick et al. *On Compositions of Transformations in Contrastive Self-Supervised Learning*. ICCV'21

Example: distinctive to sample & time shift, invariant to modality



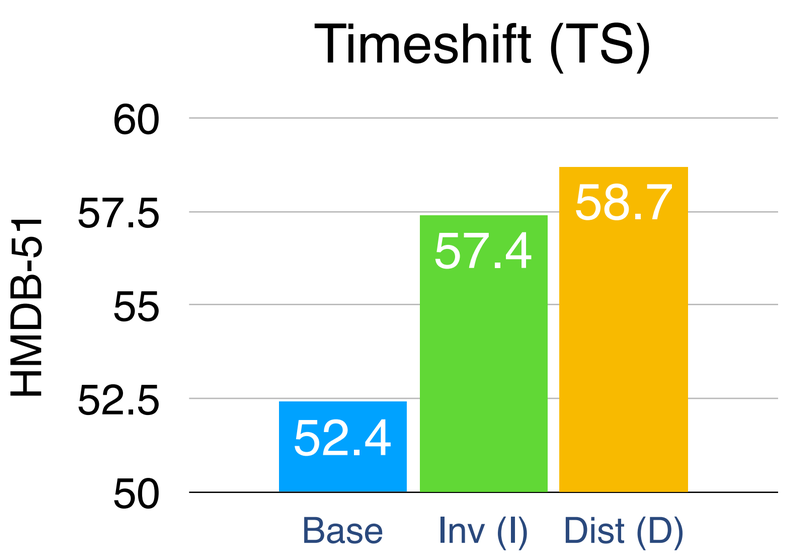
Same embedding for video and audio at *same* time from *same* video

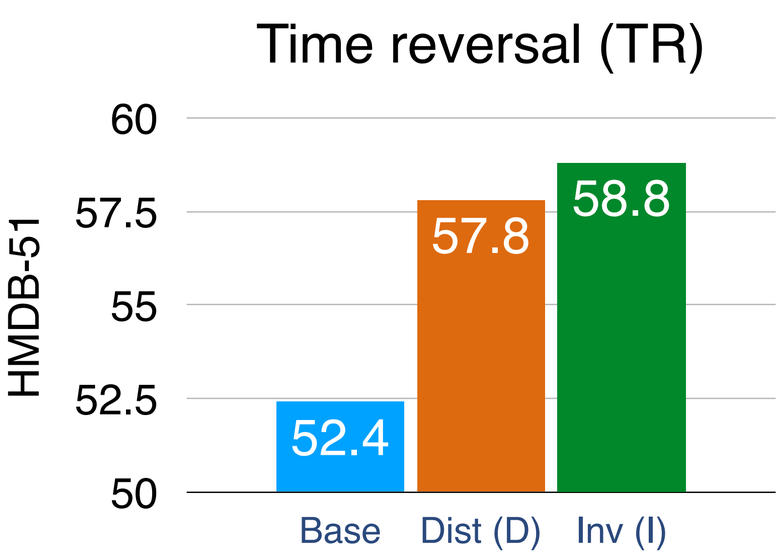
Different embedding for *different* time-indices from the *same* video

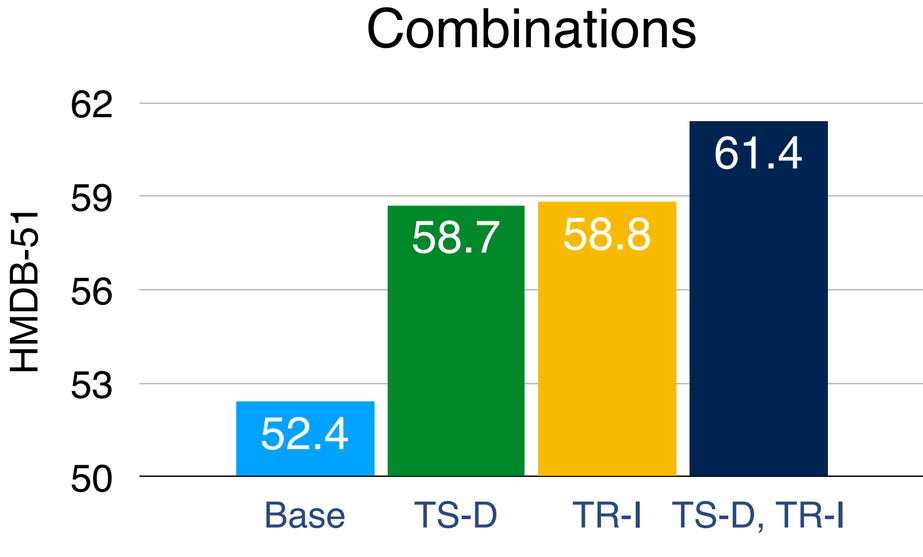
Different embedding for *different* video

Cross-modally

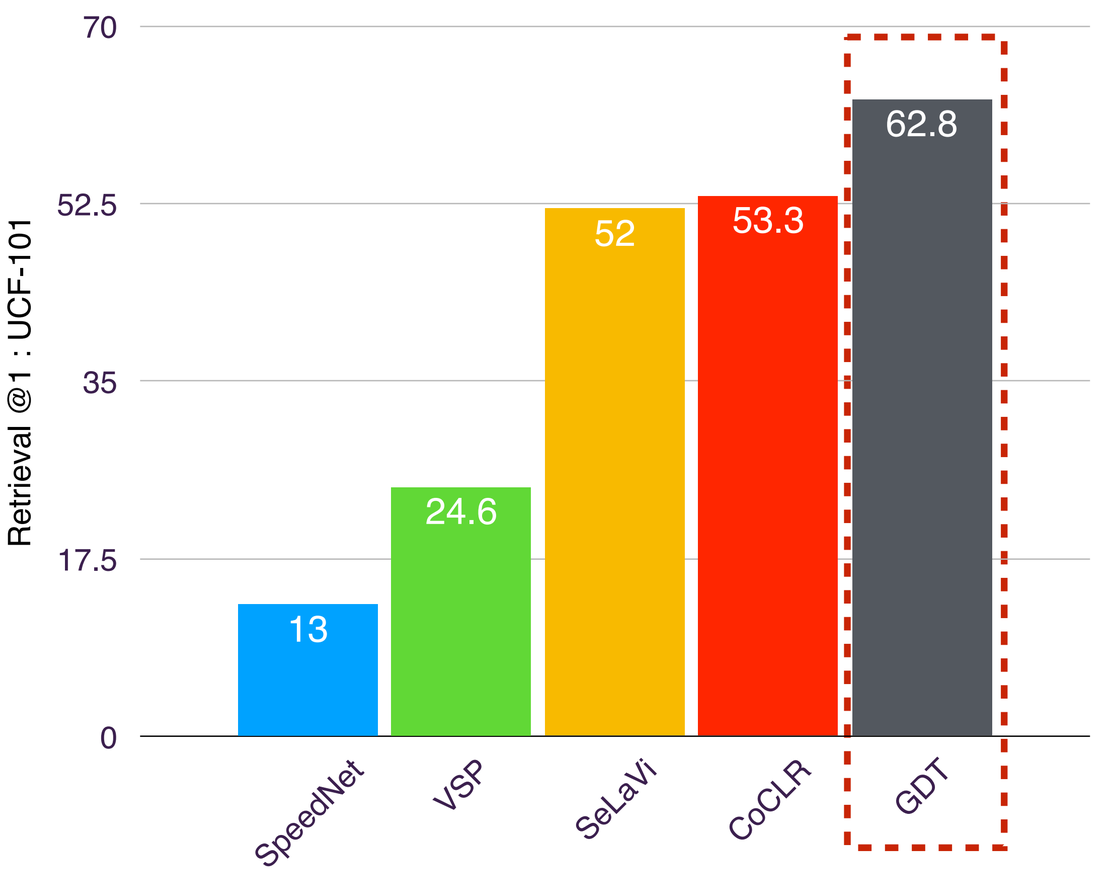
Gains from hypotheses

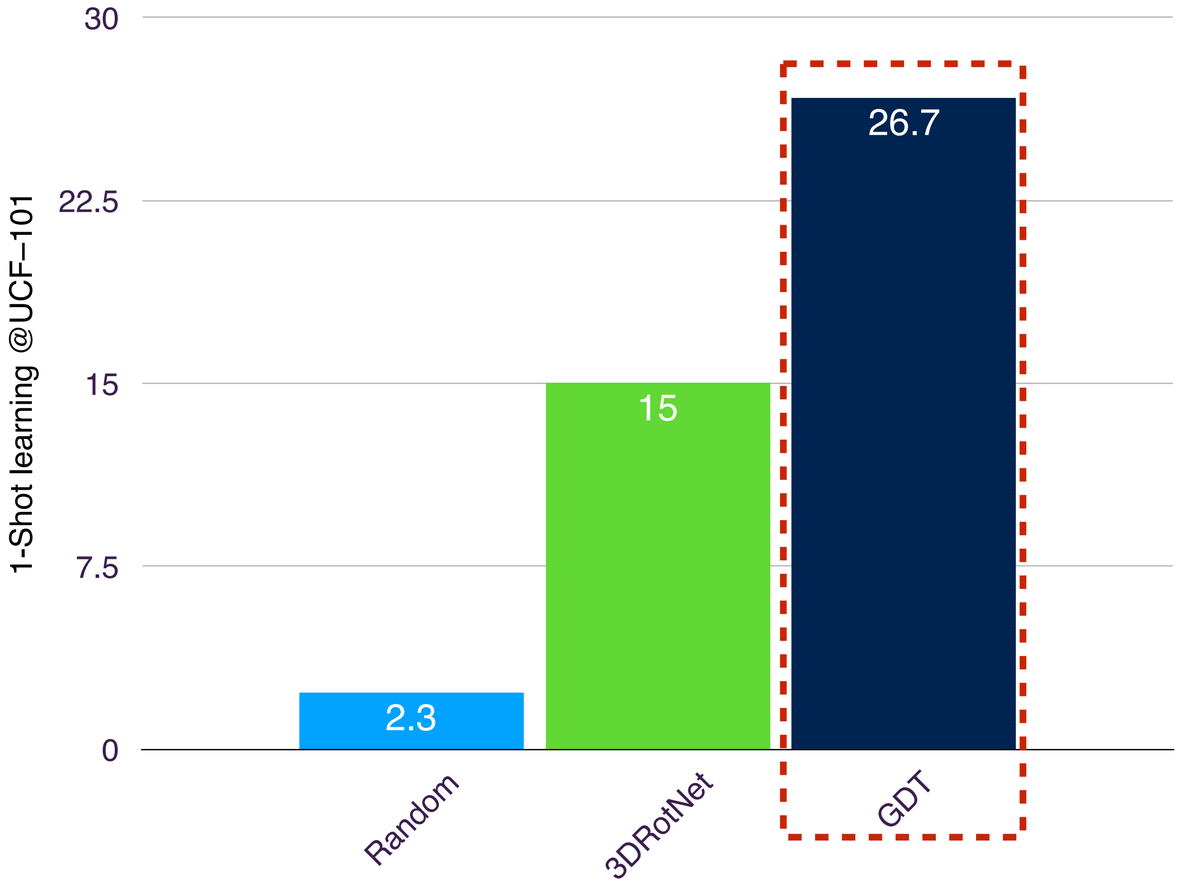






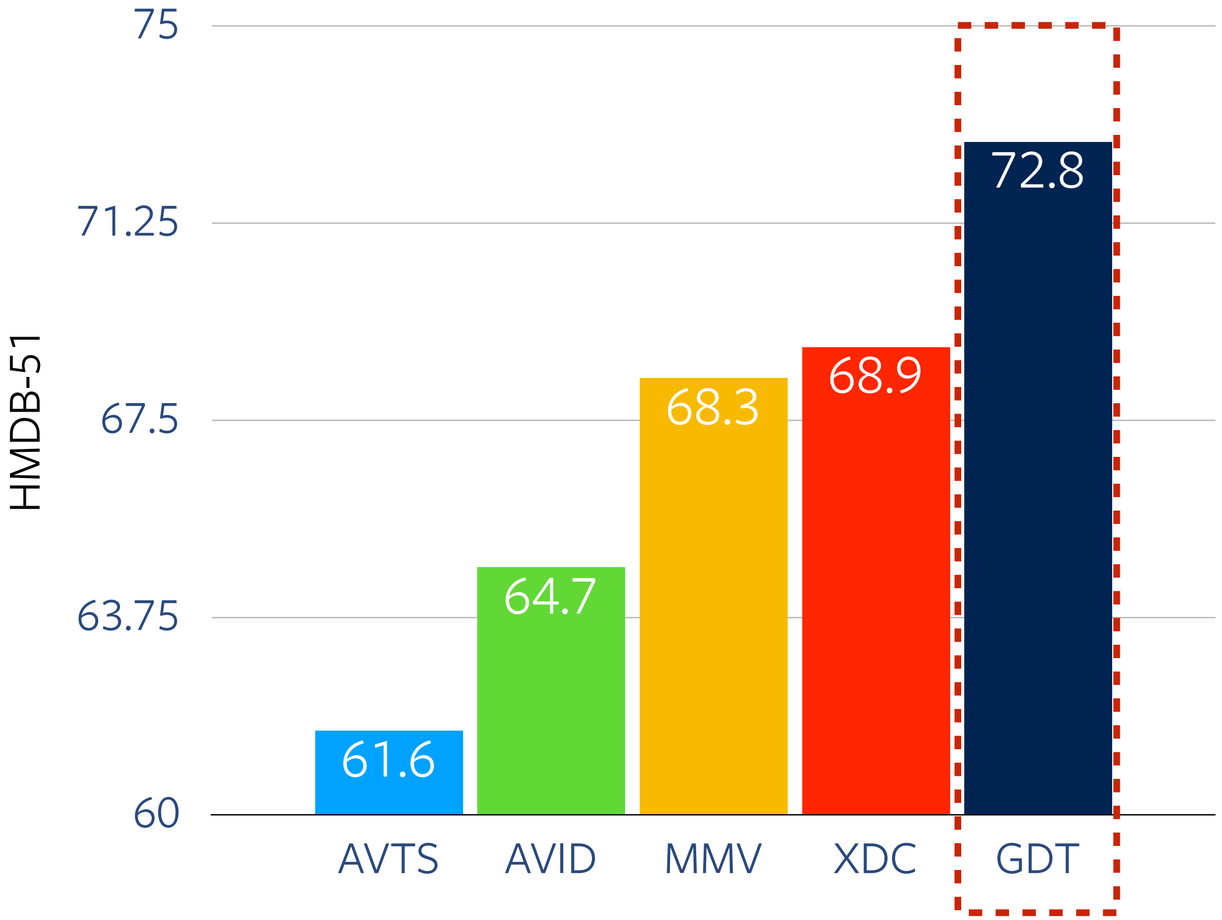
SOTA video action retrieval and few-shot learning results

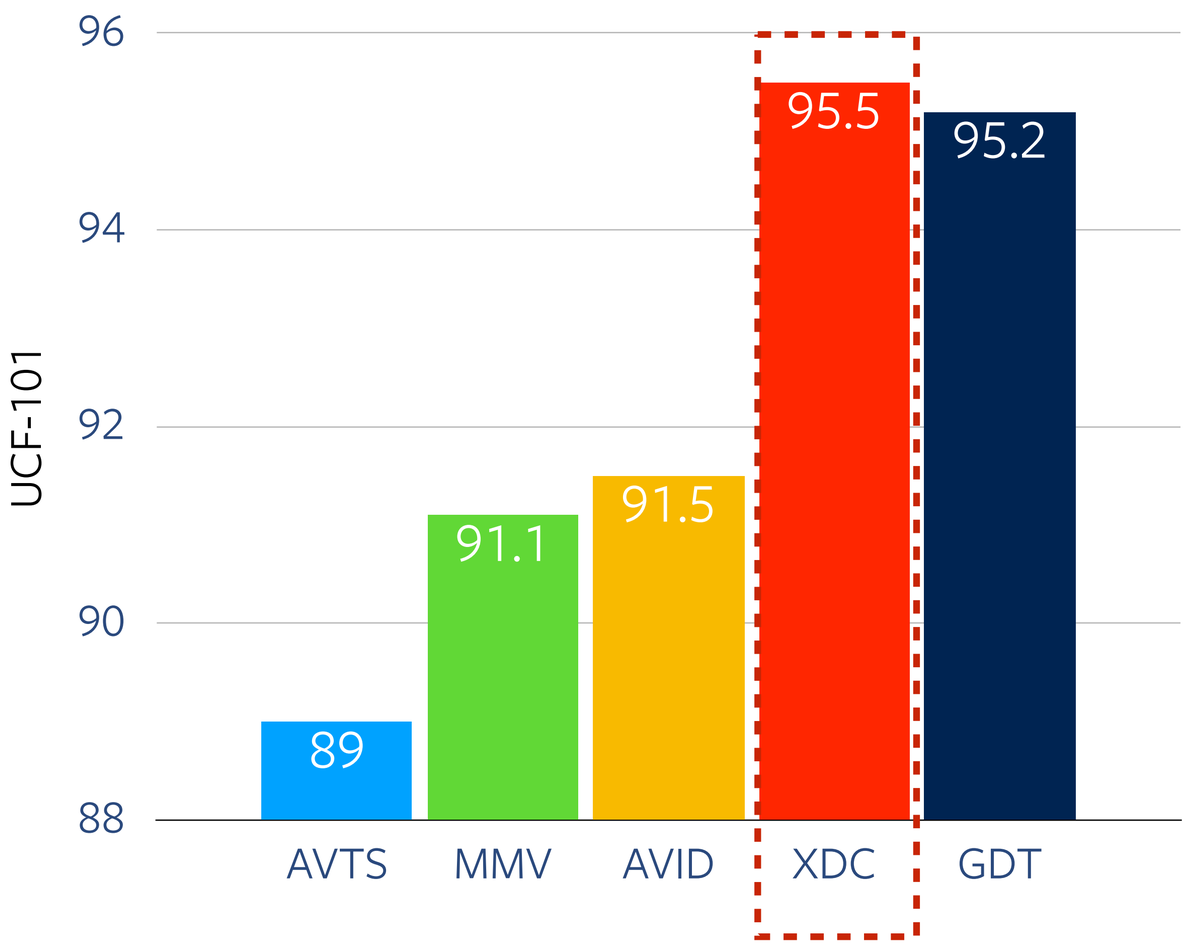




[Benaim et al, CVPR 2020; Cho et al., ArXiv; Asano et al., NeurIPS 2020; Han et al., NeurIPS 2020]
[Jing and Tian, arXiv]

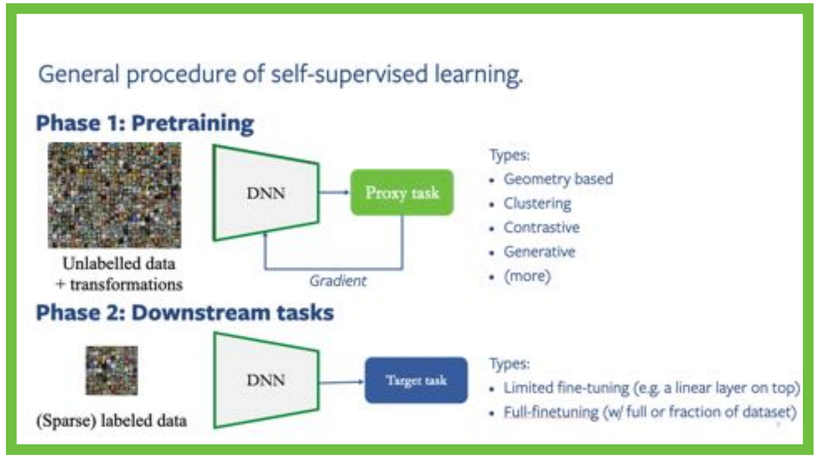
SOTA finetuning video-action recognition results

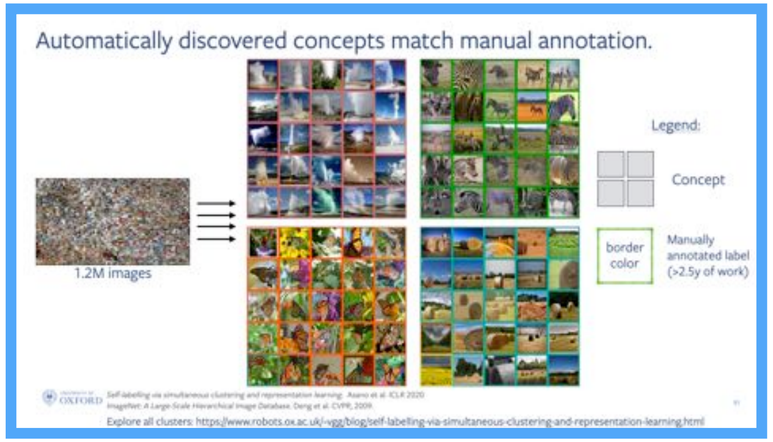


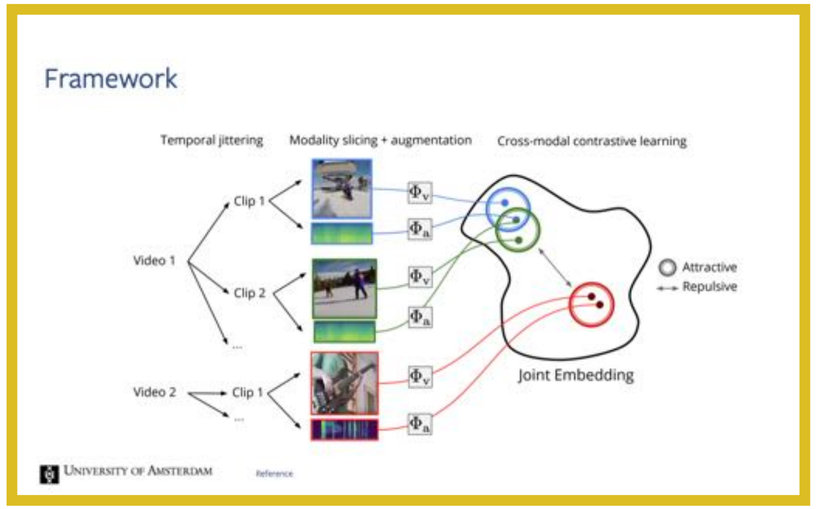


[Bruno et al., NeurIPS 2018; Morgado et al., CVPR 2021; Alayrac el al., NeurIPS 2020; Alwassel et al., NeurIPS 2020]

Visual recap of this talk







Tutorial:

https://urlis.net/5mj36