Self-Labelling via simultaneous clustering and representation learning

YUKI M. ASANO* MANDELA PATRICK* CHRISTIAN RUPPRECHT ANDREA VEDALDI

ICLR EML 2021; PAPER ORIGINALLY PUBLISHED IN NEURIPS'20

yuki@robots.ox.ac.uk, @y_m_asano

UNIVERSITY OF OXFORD

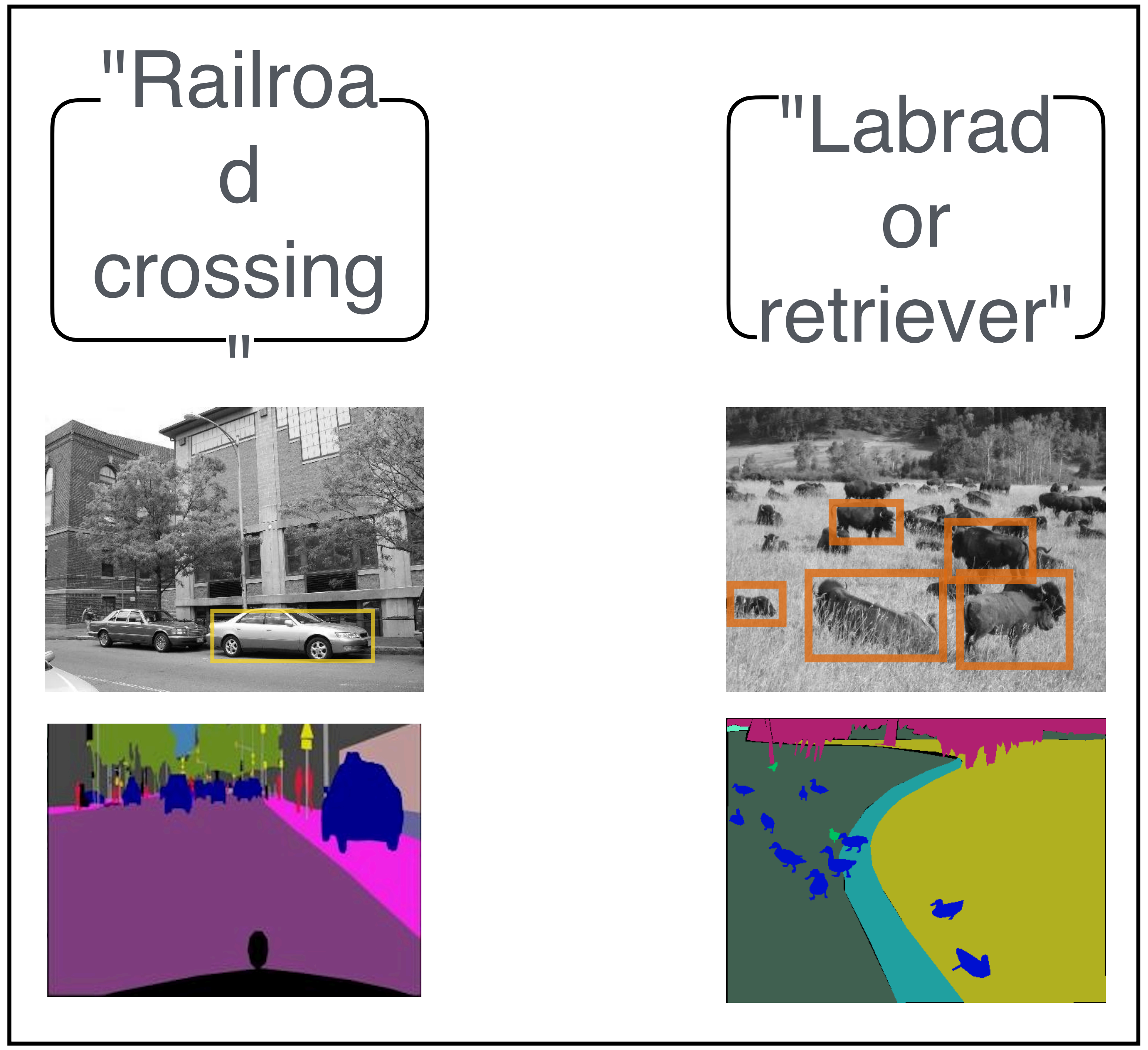
VGG
UNIVERSITY OF OXFORD

FACEBOOK

Manual annotations for the data are limiting.

Data is often cheap

But manual annotations are expensive



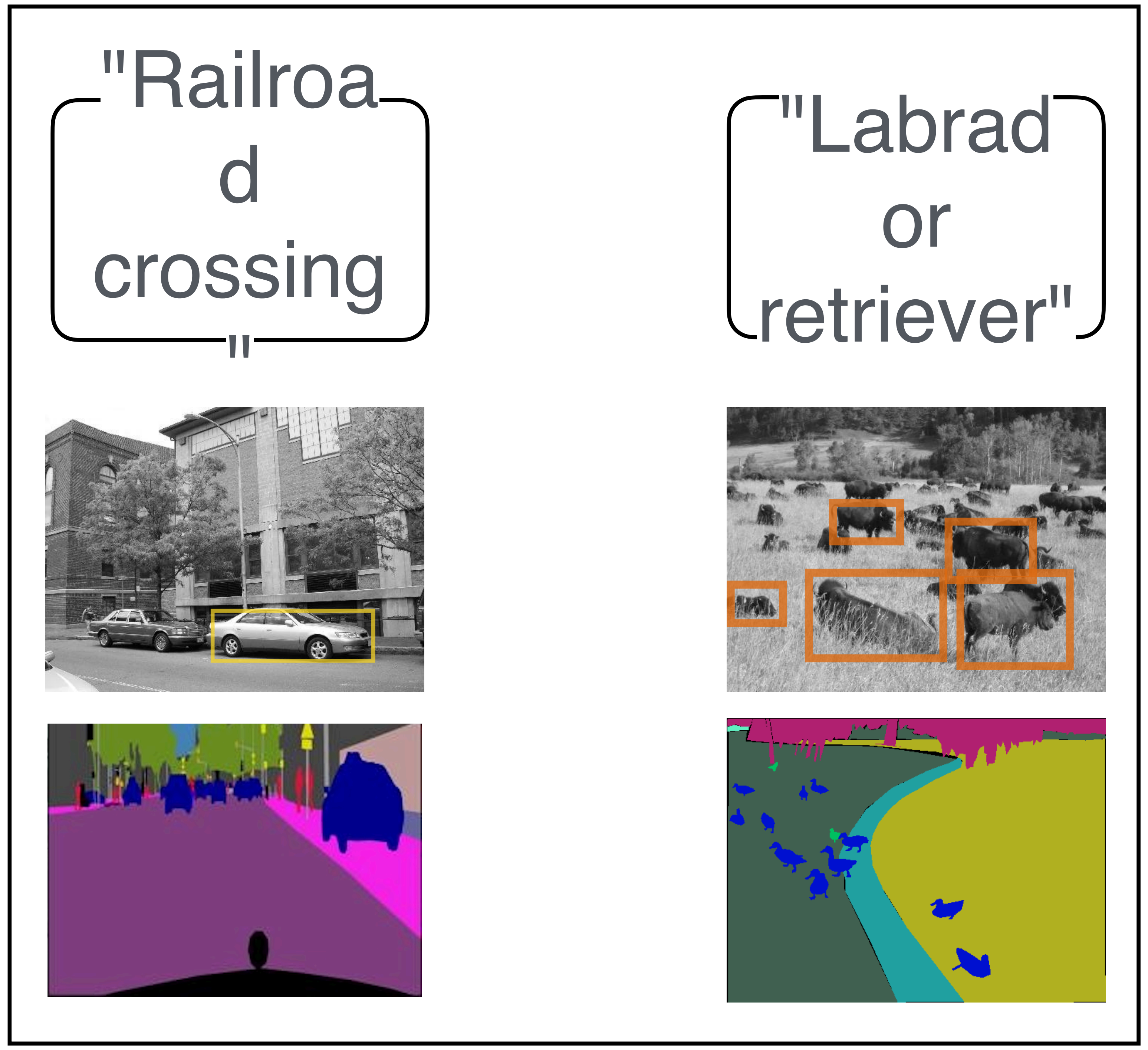
UNIVERSITY OF OXFORD *Scene parsing through ADE20K dataset.* Zhou et al. CVPR 2017.

Replacing manual annotations by self-supervised learning.

Clustering

Detection

Segmentation



Images

CliqueCNN (Bautista, NeurIPS'16)
DeepCluster (Caron, ECCV'18)
IIC (Ji, ICCV'19)
SeLa (Asano, ICLR'19)
SCAN (Gansbeke, ECCV'20)
and more

SSOD (Afouras arxiv'21)

MaskContrast (Gansbeke arxiv'21)

Video

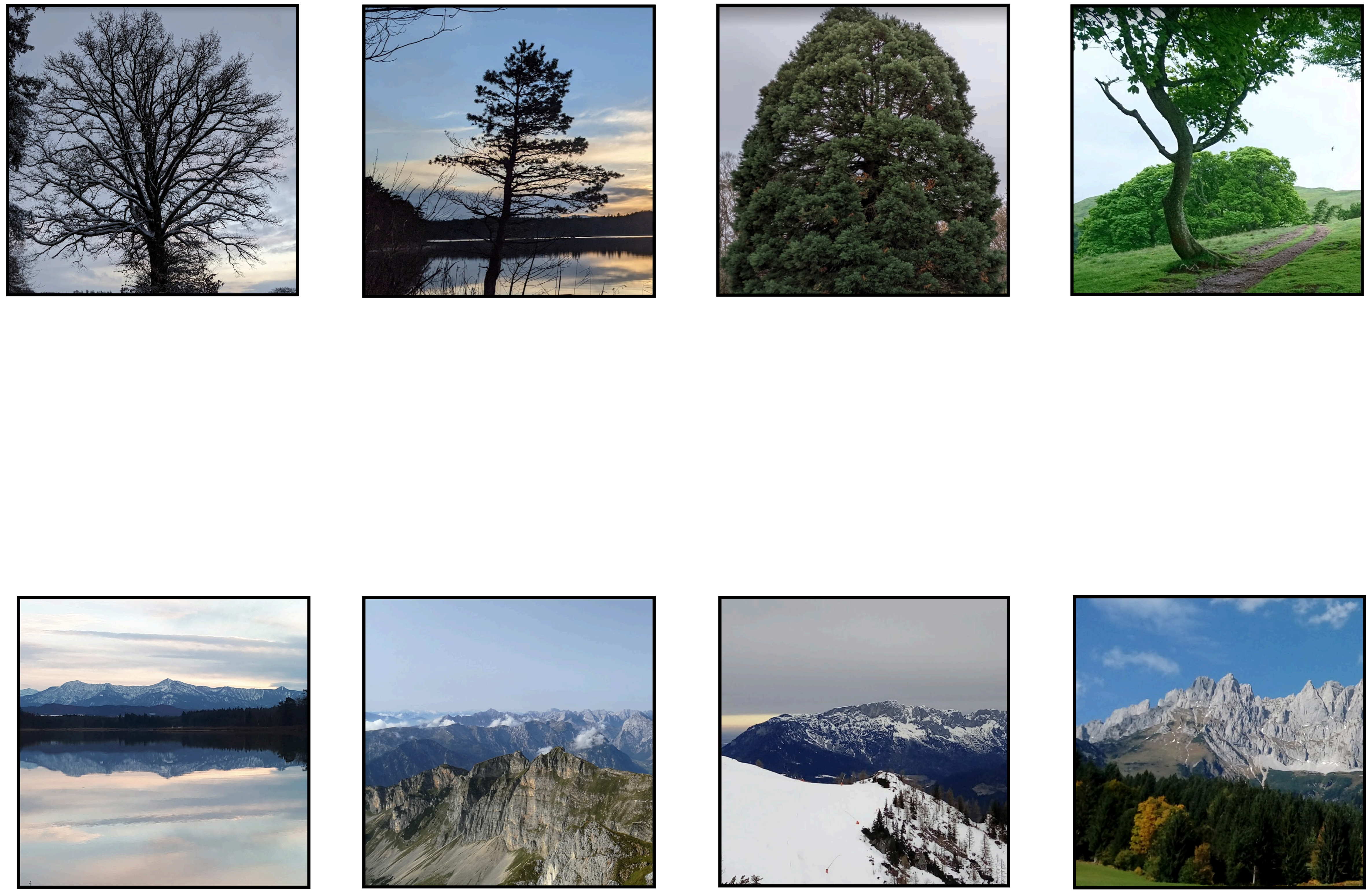
[Sight from Sound (Owens ECCV'16)]
XDC (Alwassel NeurIPS'20)
SeLaVi (Asano NeurIPS'20)

Boxes:
SSOD (Afouras arxiv'21)

[Heatmaps]:
Objects that sound (Arandjelović ECCV'18)
DMC (Hu CVPR'19)
DSOL (Hu NeurIPS'20)

UNIVERSITY OF OXFORD

Can we label the dataset *without* humans?

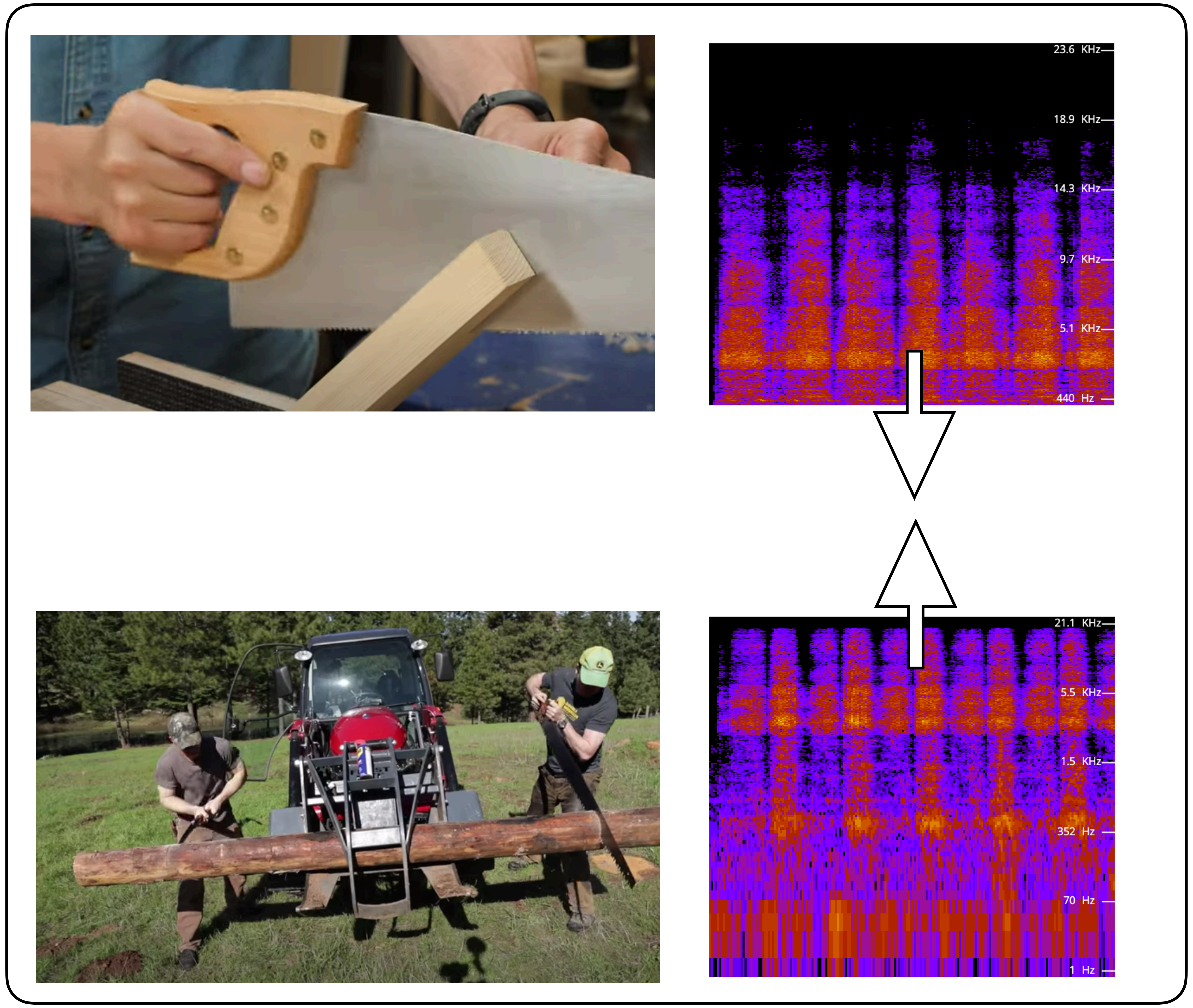


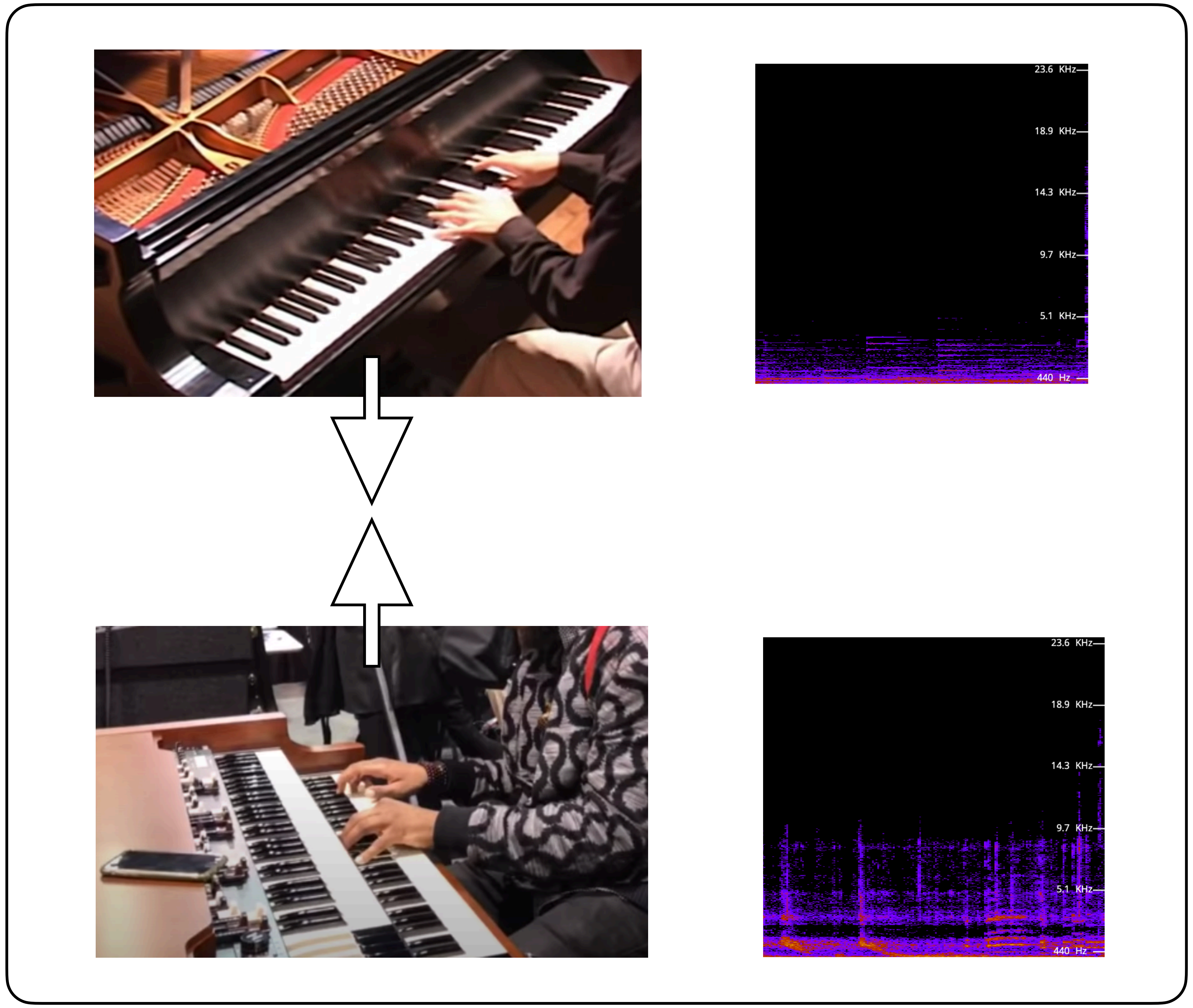
?

Label A

Label B

Multiple modalities can help us infer semantic similarity.



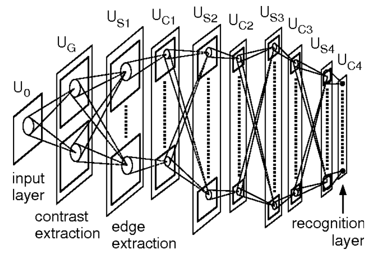


https://www.youtube.com/watch?v={ZlS8t8_nAHo, tvVYdaykaZM, Kt8_u_i-anQ, J4X6tc1HYts}

Learning with labels.

$\mathbf{x}$

Model Φ



$p(y\,|\,\mathbf{x})$

y_{gt}

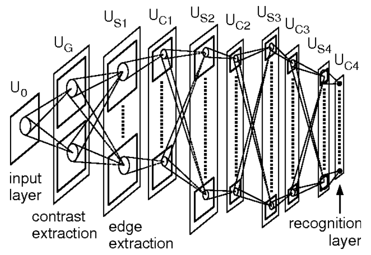
What if we don't have labels?

Minimise the cross-entropy loss w.r.t to **labels**

Learning without labels.

$\mathbf{x}$

Model Φ



$p(y \mid \mathbf{x})$

y_{gt}

(Minimise the cross-entropy loss w.r.t to **labels**)

+

(**optimize pseudolabels**)

How can we *optimize* labels?

If we had ground-truth labels:

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

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

Idea: Representing the labels as an assignment table q:

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

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

Solution: Force all labels to be used an *fixed* number of times and pose as optimal transport.

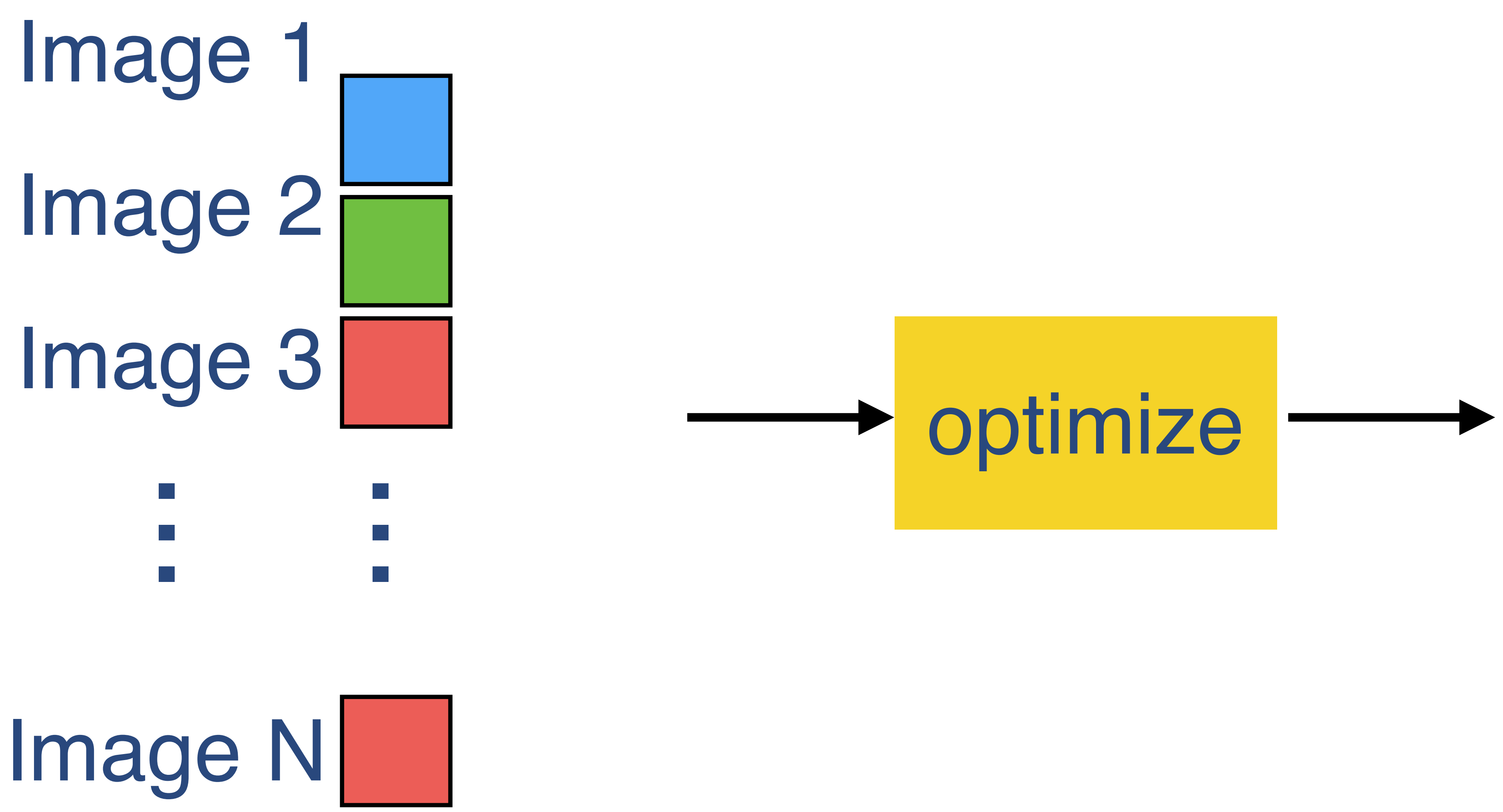
$$\min_{q, \Phi} L(q, \Phi) \quad \text{s.t.} \quad \sum_{i=1}^{N} q(y \,|\, \mathbf{x}_i) = \frac{N}{K},$$

with the iterative solution $Q_{ij} = u_i p^{\lambda} v_j$

UNIVERSITY OF OXFORD

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

Solution: “Fixed marginal” label optimization

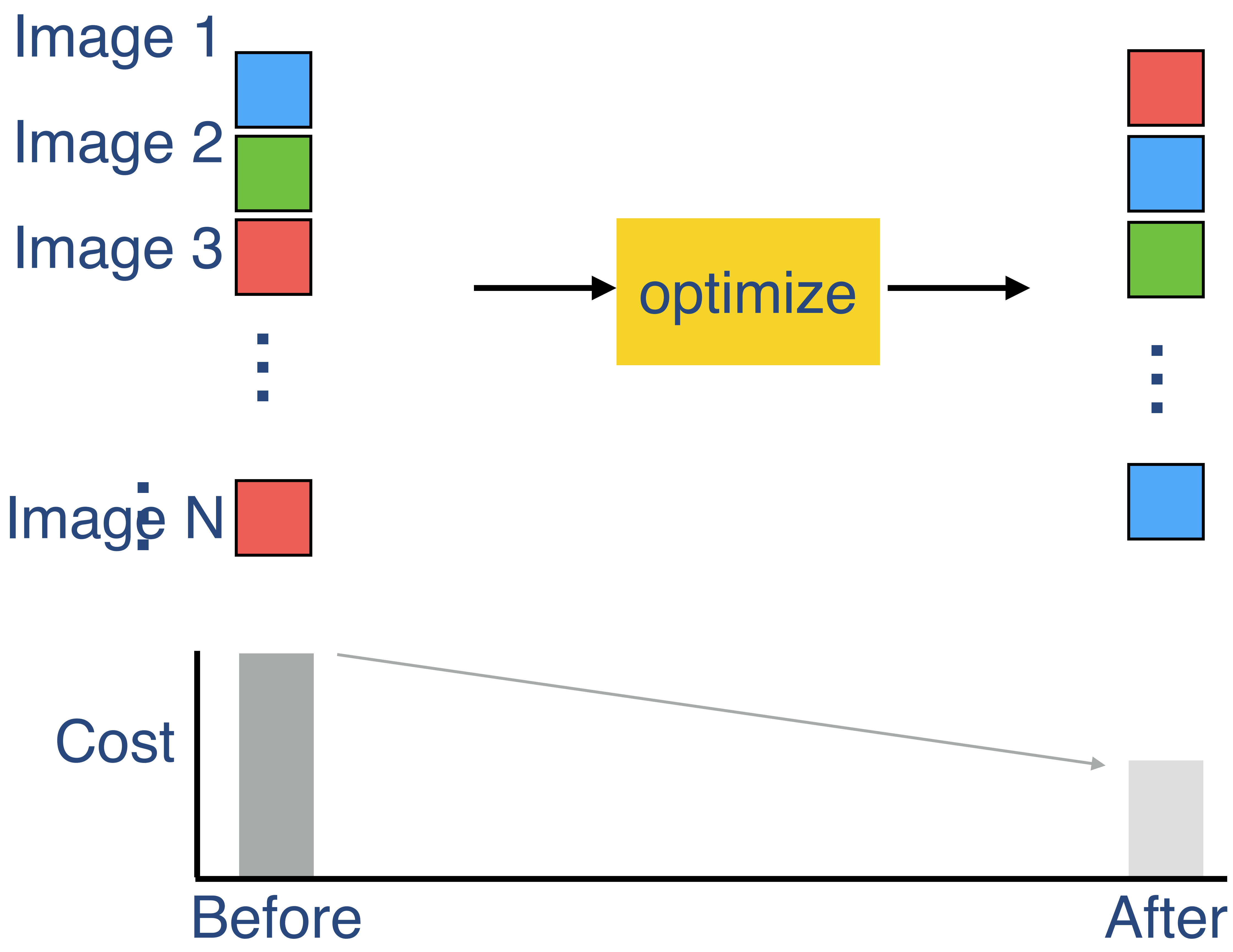


UNIVERSITY OF OXFORD

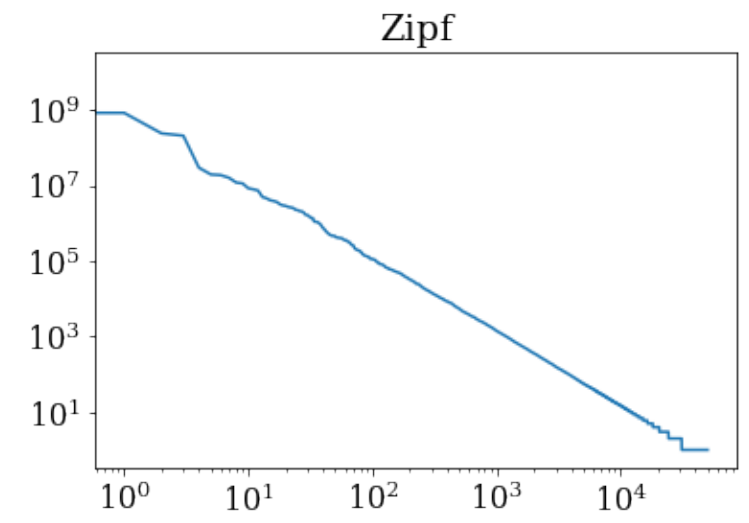
Self-labelling via simultaneous clustering and representation learning.
Asano et al., ICLR 2020

Solution: “Fixed marginal” label optimization (Sinkhorn-Knopp)

Intuition: Shuffle fixed set of labels around s.t. it best fits current model



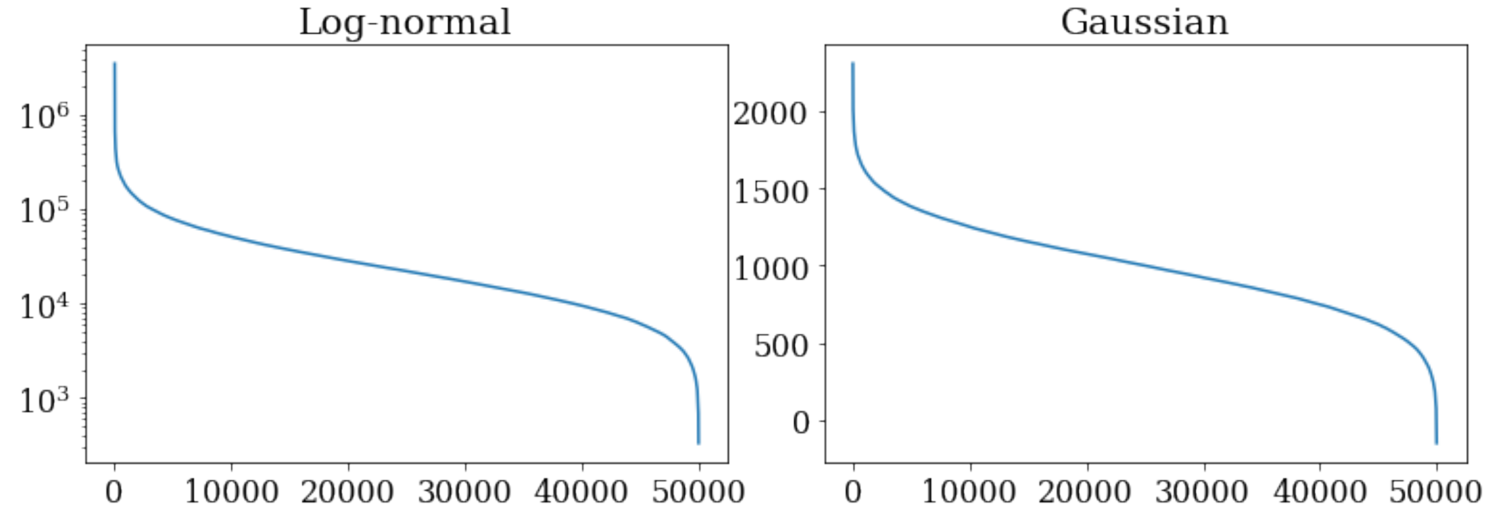
Flexibly use any marginals:



$$\min_{Q\in U(r,c)} \langle Q, -\log P\rangle + \frac{1}{\lambda}\mathrm{KL}(Q\|rc^\top) \rightarrow$$

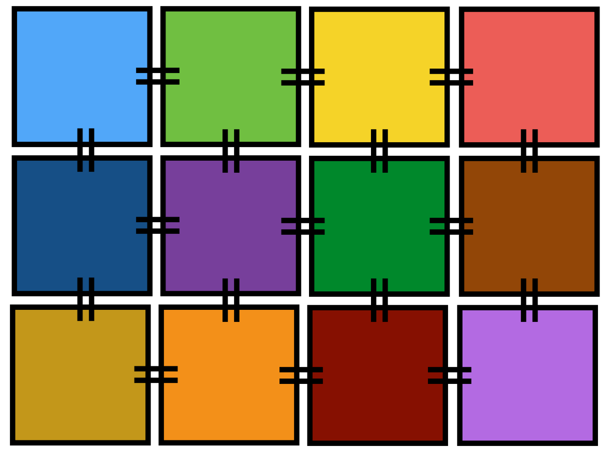
$$\langle Q, -\log P\rangle + \frac{1}{\lambda}\mathrm{KL}(Q\|Rrc^\top)$$

$$= \text{const} + \sum_y -q(y)\,[R\log r]_y\,.$$



Algorithm

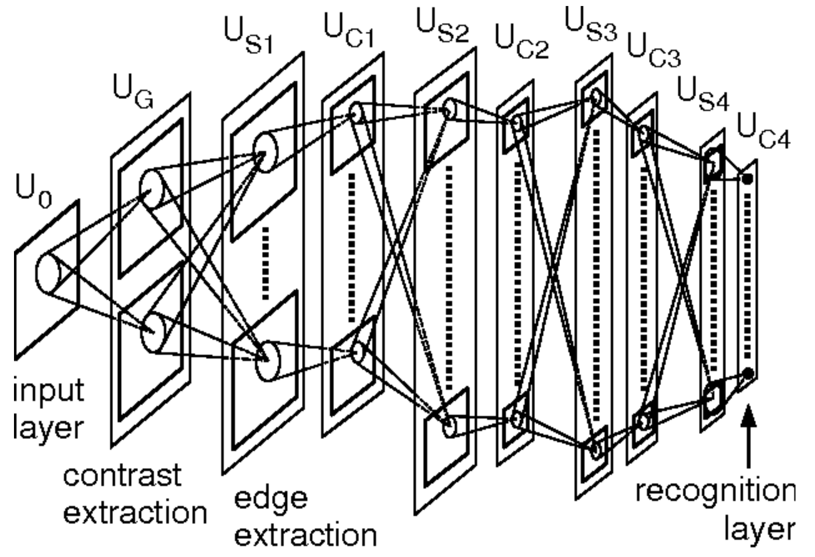
Label assignments q



Optimal labelling

Cross entropy training
with augmentations

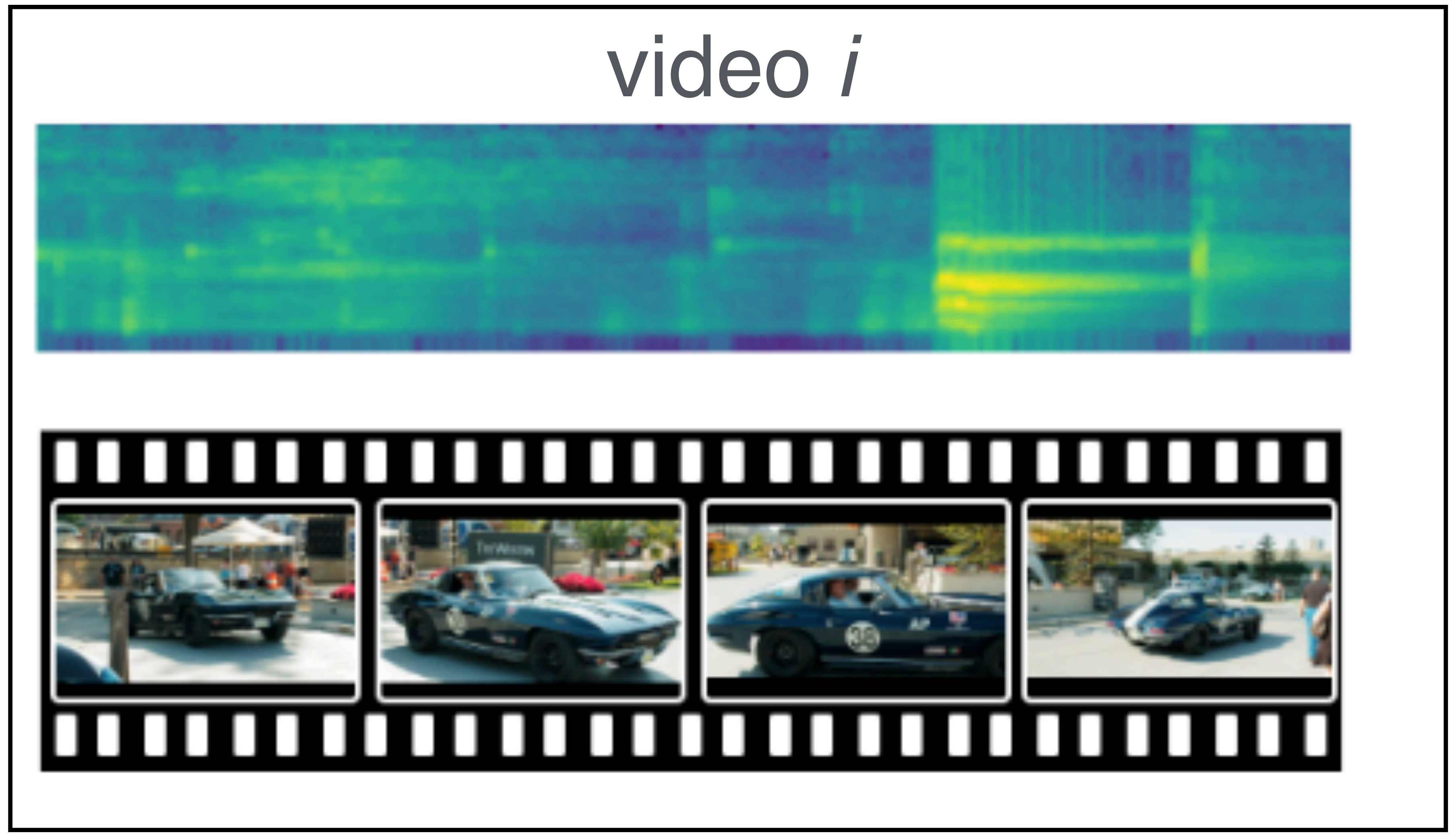
Model Φ



UNIVERSITY OF OXFORD

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

Clustering multi-modal data



A → Labels 1

V → Labels 2

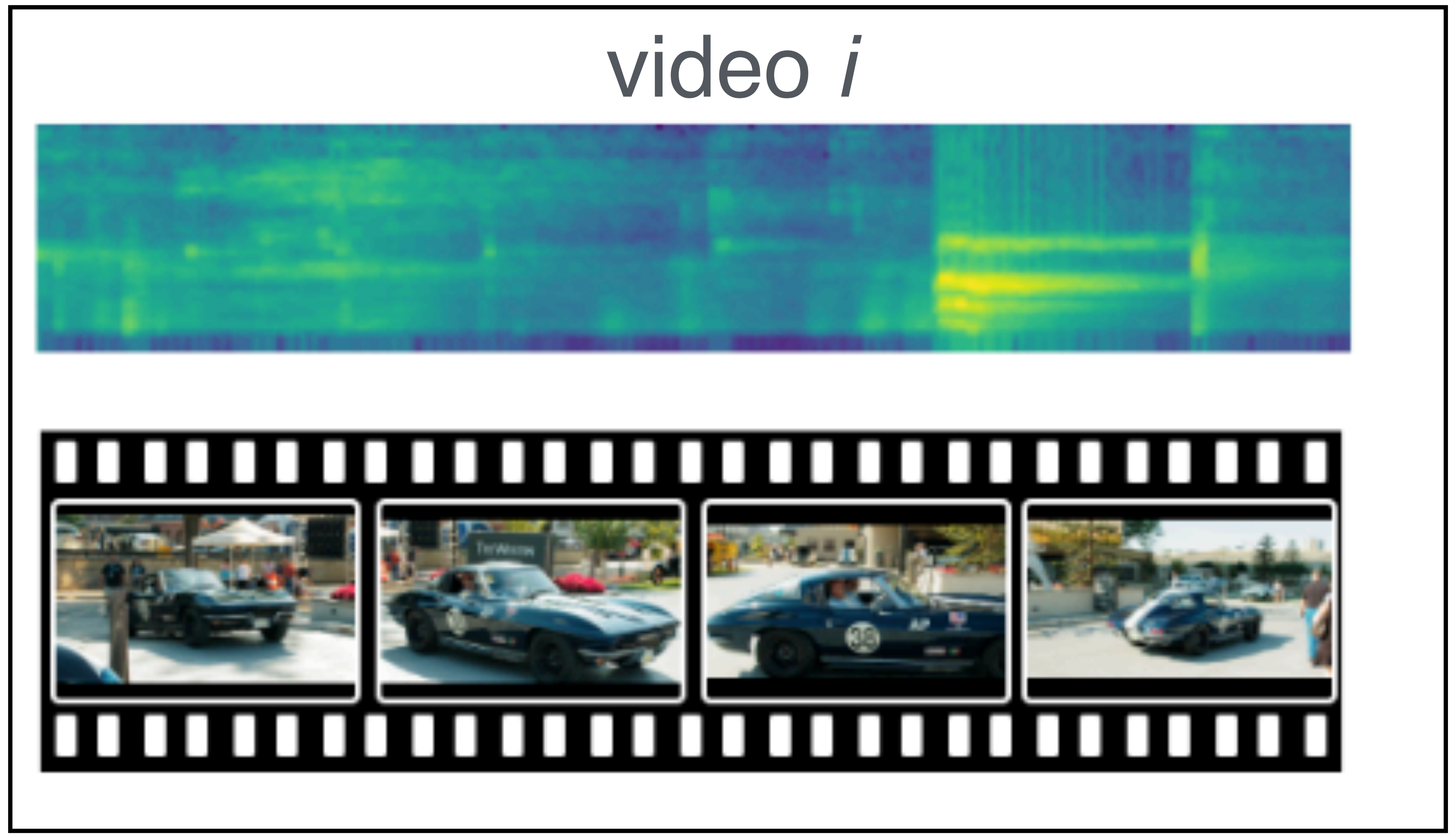
- × does not use same-source information
- × two different sets of clusters

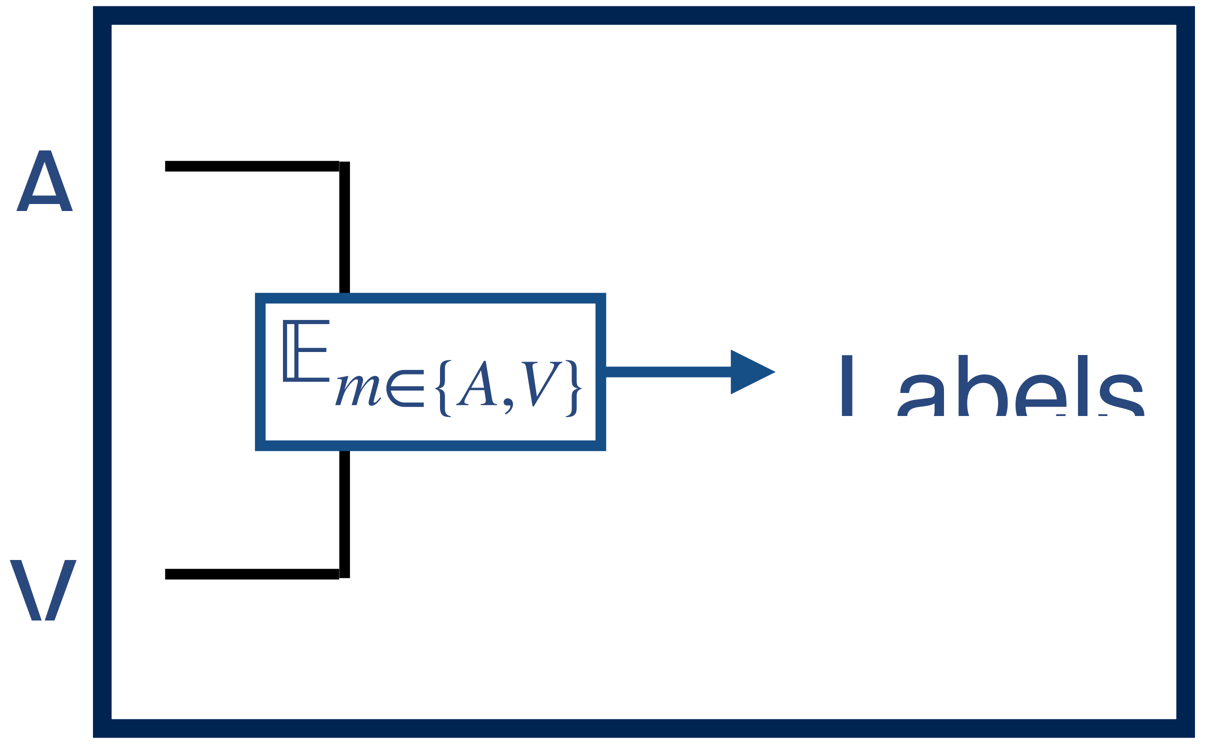
Concat → Labels

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

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

Our idea: view each modality as an *augmentation.*

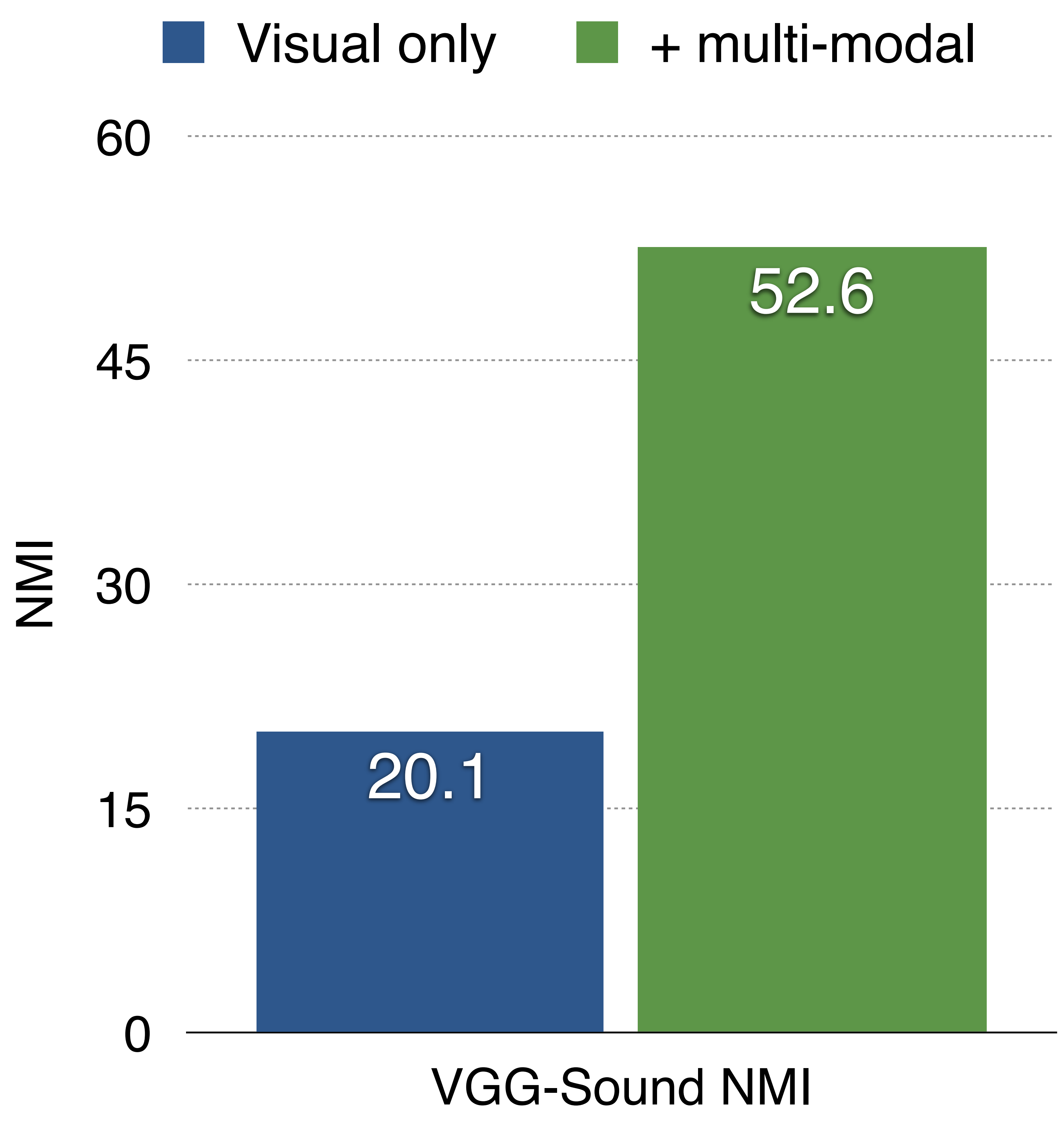




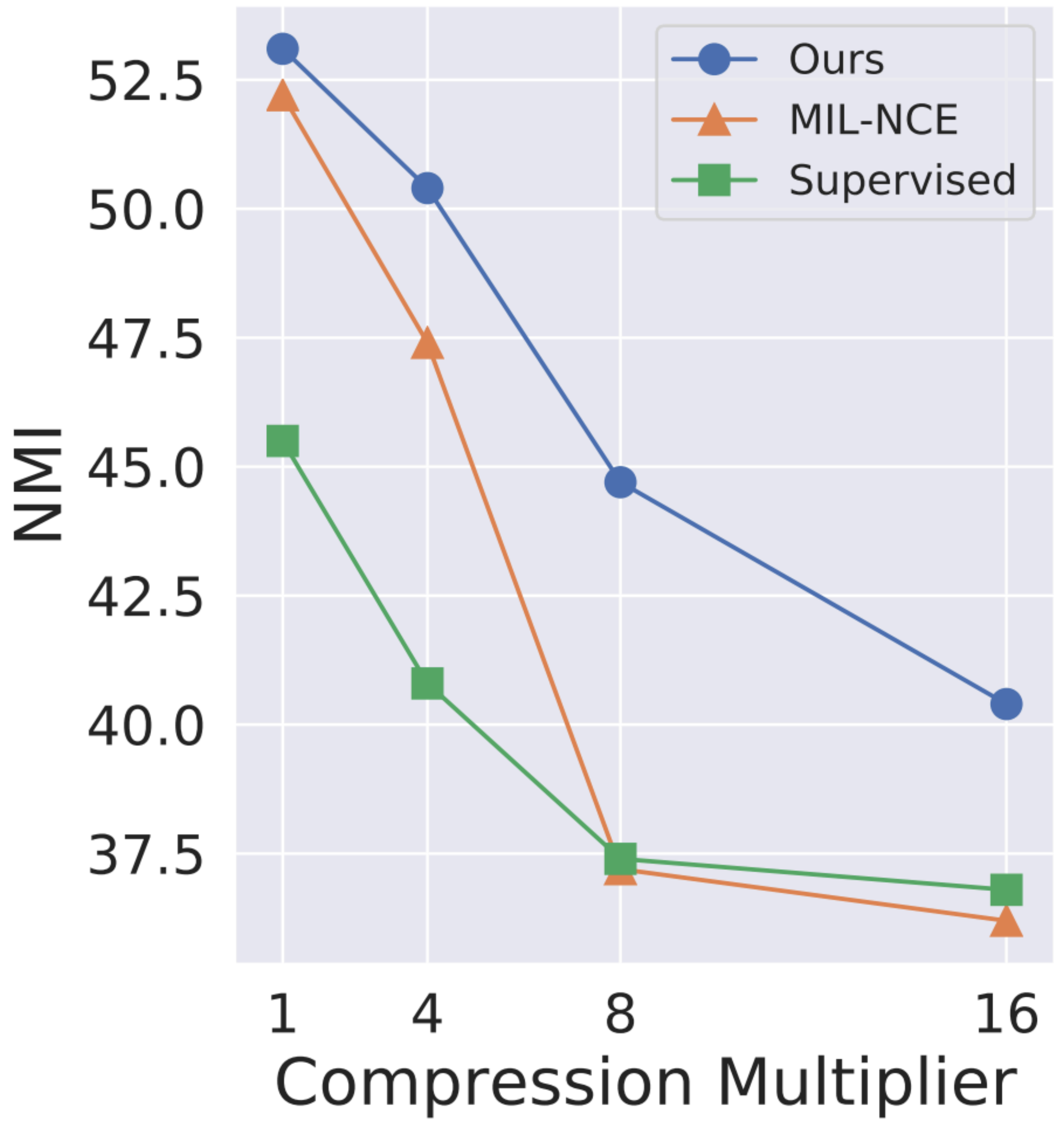
The *same* clusters are produced from *either* modality

$$E(\Phi, q) \propto \sum_{i,c,m} q(c\,|\,i)\left[\log \operatorname*{sftmx}_{c} \Phi_a(\text{audio}(\mathbf{x}_i)) + \log \operatorname*{sftmx}_{c} \Phi_v(\text{video}(\mathbf{x}_i))\right]$$

Multi-modality clustering is key.



Clustering works much better when also using the audio.



Our clustering formulation degrades less quickly thanks to treating audio equally.

Simultaneous clustering and representation learning is better.

Ours (train VGG-Sound)

vs

pre-train + K-means:

DPC (train Kinetics-400)

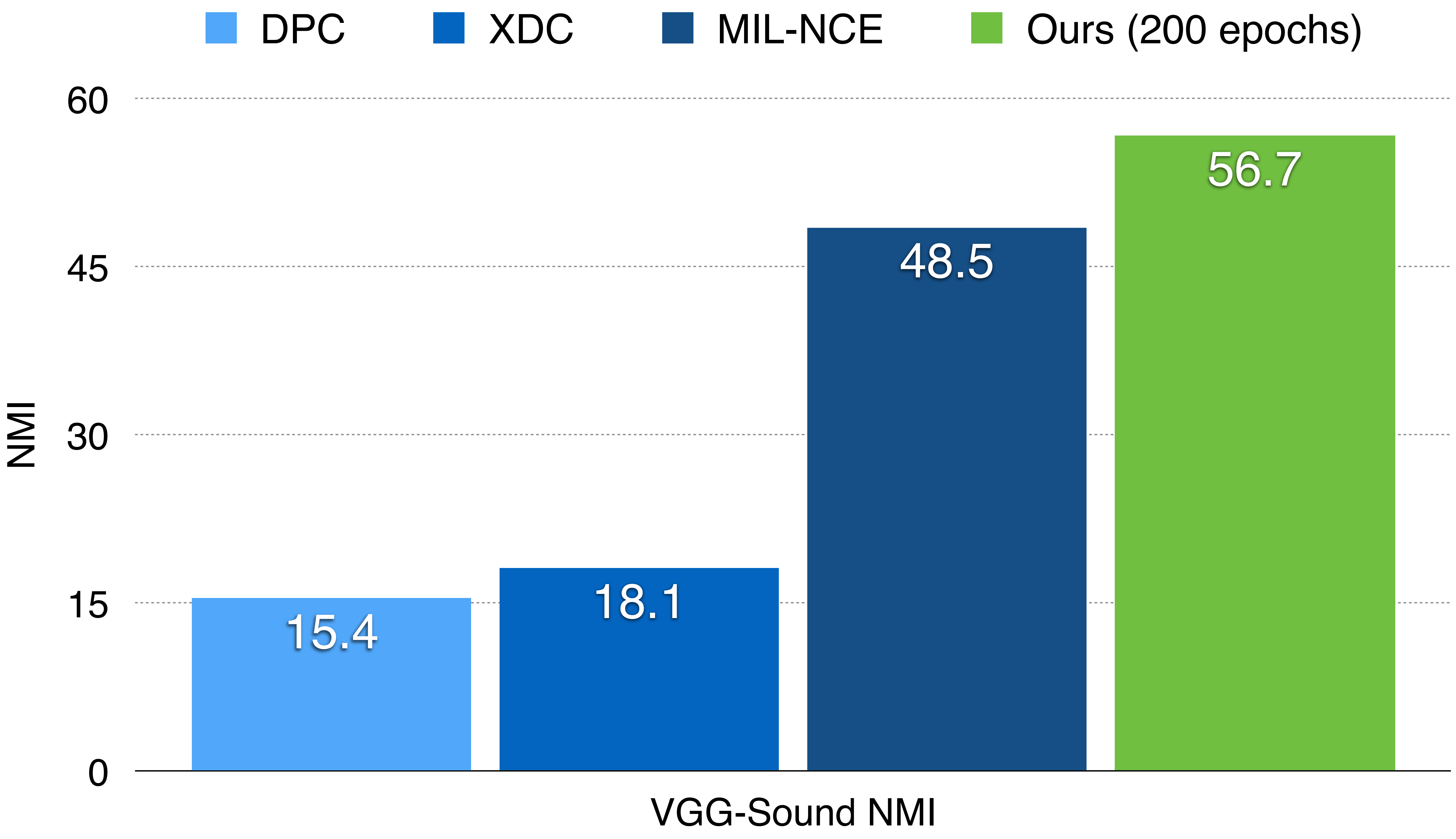
Video representation learning by dense predictive coding, Han, Xie, and Zisserman, ICCV, 2019

XDC (train Kinetics-400)

Self-supervised learning by cross-modal audio-video clustering, Alwassel, Mahajan, Torresani, Ghanem, and Tran, arXiv, u

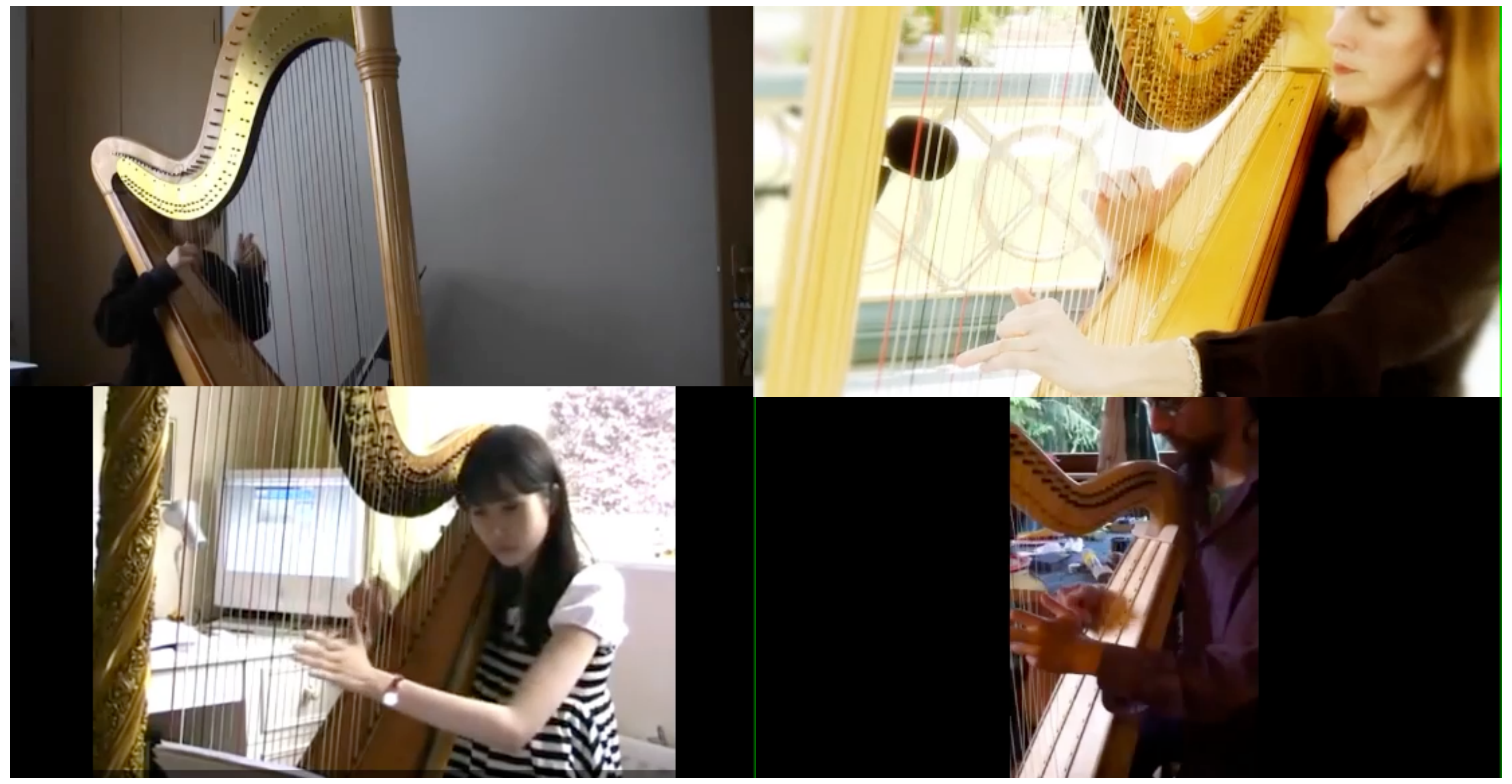
MIL-NCE (train on HowTo100M)

End-to-end learning of visual representations from uncurated instructional videos, Miech, Alayrac, Smaira, Laptev, Sivic, and Zisserman, arXiv, 2019



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

Clusters are highly consistent thanks to utilising both modalities.



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

View all clusters here: https://www.robots.ox.ac.uk/~vgg/research/selavi/#den