Self-supervised learning for detecting objects and the single image prior

ASCI COMPUTER VISION BY LEARNING

YUKI M. ASANO,

QUVA SCIENTIFIC LAB MANAGER, ASSISTANT PROFESSOR, UNIVERSITY OF AMSTERDAM

UNIVERSITY
OF AMSTERDAM

Overview

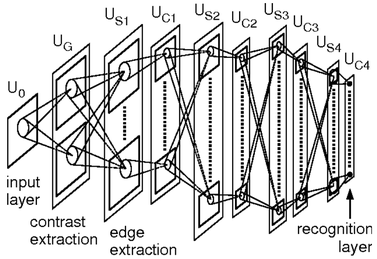
- Brief Intro to Self-supervised learning
- Self-supervised object detection from audio visual correspondence
- Extrapolating from a Single Image to a Thousand Classes using Distillation

Introduction to self-supervised learning in computer vision

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

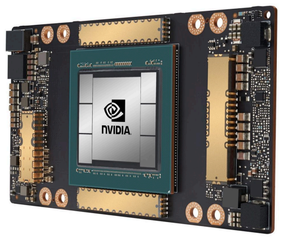
Algorithms

Deep neural networks



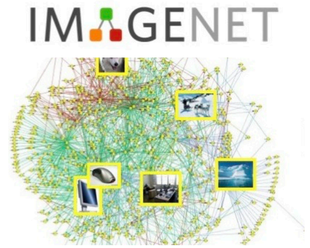
Hardware

GPUs



Data

Large scale datasets

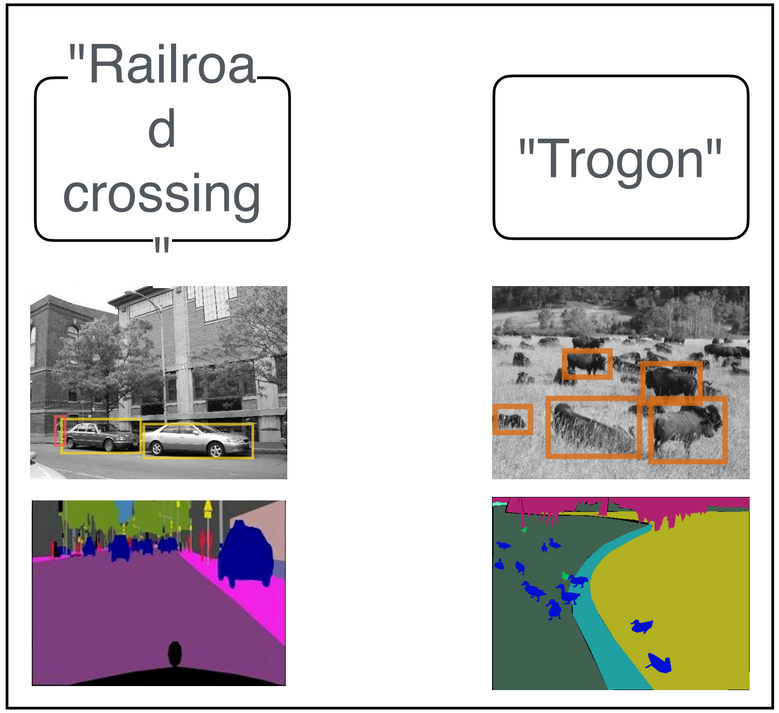


Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Fukushima, K., Biol. Cybernetics 1980
Object Recognition with Gradient-Based Learning. LeCun et al. Shape, Contour and Grouping in Computer Vision 1999
ImageNet: A Large-Scale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

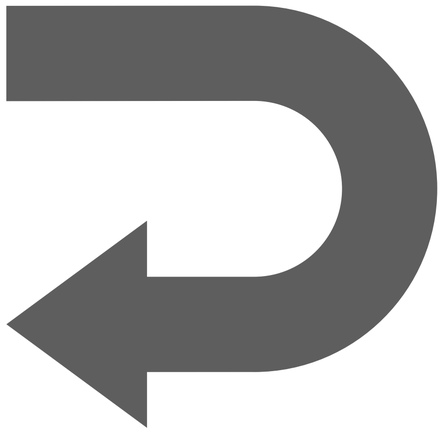
Images are often cheap

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.

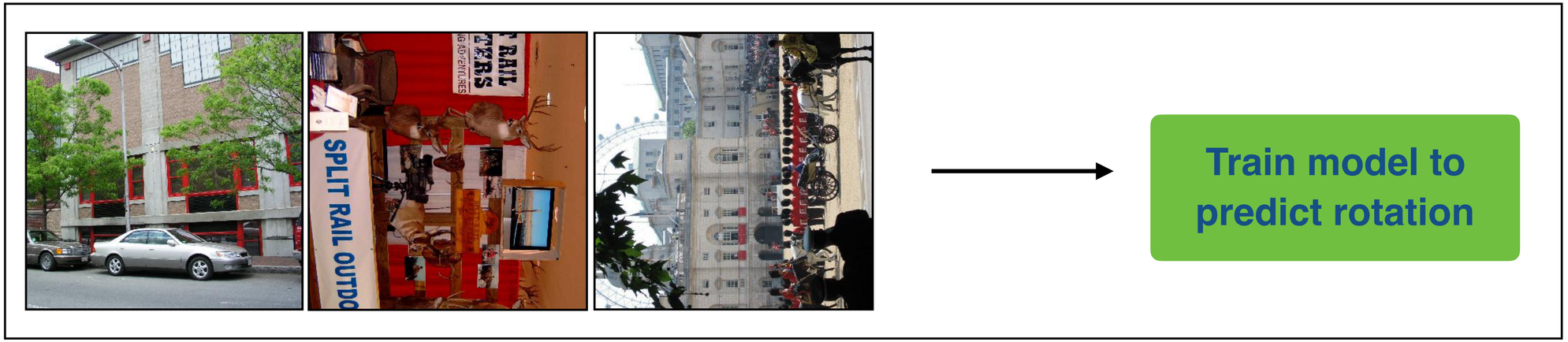
Solving the problem of expensive annotations: self-



Self-supervision

Extract a supervisory signal from the raw data alone

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



But:

Self-supervised representation learning

Goal:

pretrain a representation, e.g. a CNN **without labels**

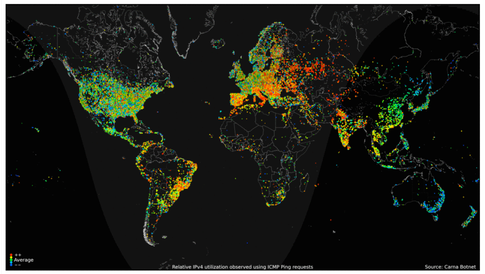
Application:

transfer the representation to new "downstream" tasks

or

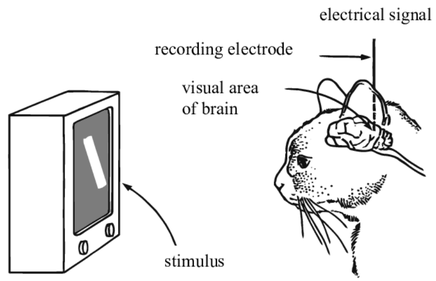
do something useful without labels

Why study self-supervised representation learning?



Scalability

Accessibility



Fundamentals

Ambiguity of labels

Images:
https://www.kaspersky.co.uk/blog/amazing-internet-maps/6420/
https://www.kaggle.com/c/herbarium-2019-fgvc6

https://www.researchgate.net/figure/Fig-In-the-classic-neuroscience-experiment-Hubel-and-Wiesel-discovered-a-cats-visual_fig1_335707980
https://en.wikipedia.org/wiki/List_of_house_styles

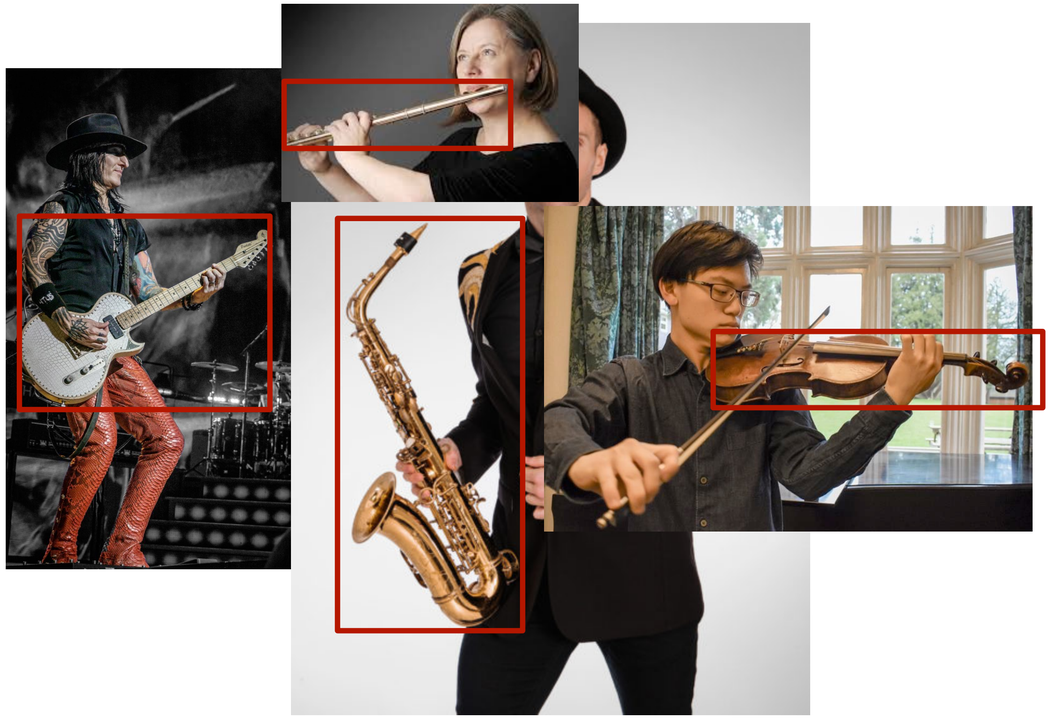
Self-supervised object detection from audio visual correspondence CVPR'22

TRIANTAFYLLOS AFOURAS*, YUKI M. ASANO*,
FRANCOIS FAGAN, ANDREA VEDALDI, FLORIAN METZE

Object detection - supervised training

Detector

Supervised training

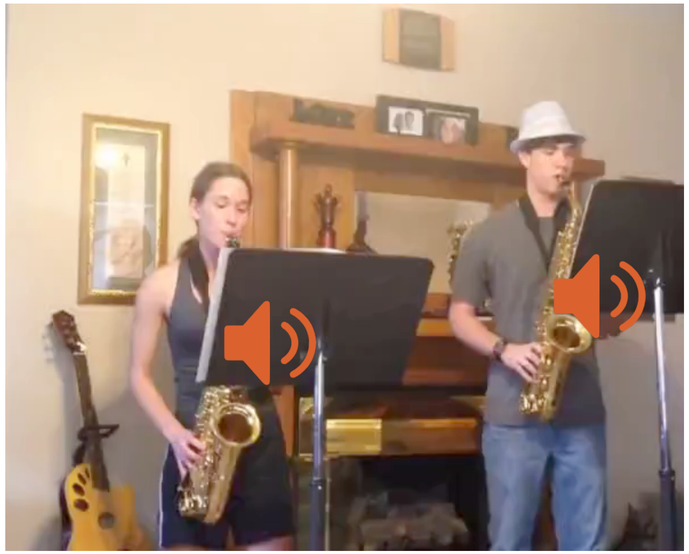


Lots of annotated samples

Data annotation expensive

Process hard to generalise

What we propose instead:

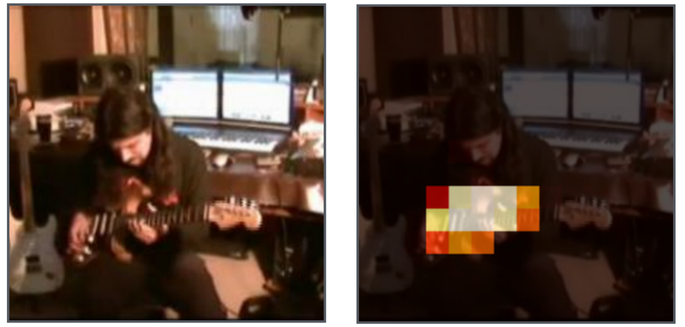


Use sound as supervision for detection

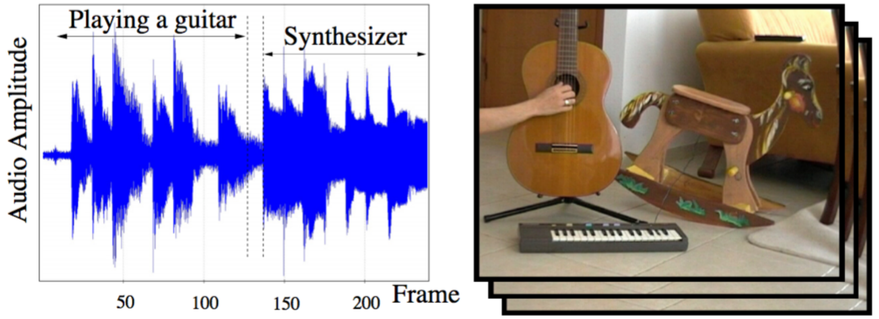
e.g. 1 second window around frame

Related work: sound source localisation

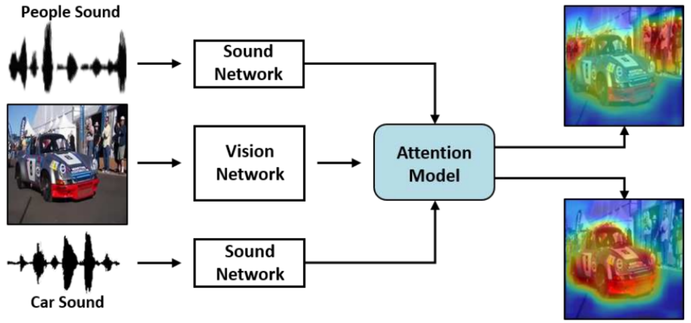
Audio vision: Using audio-visual synchrony to locate sounds. Hershey and Movellan, NeurIPS 2000.



Objects that Sound. Arandjelović and Zisserman, ECCV 2018.

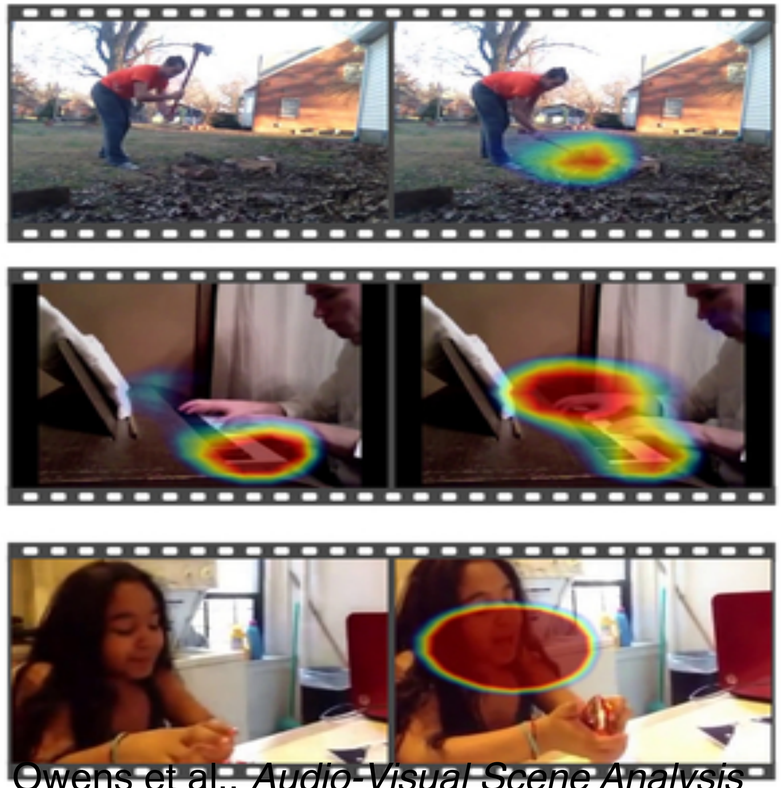


Pixels that Sound. Kidron et al., CVPR 2005.



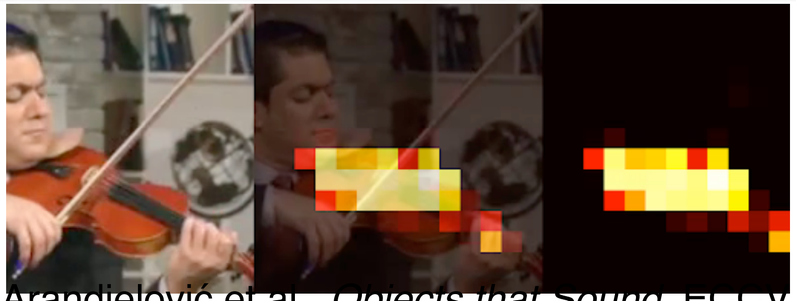
Learning to Localize Sound Source in Visual Scenes. Senocak et al., CVPR 2018.

Related work: Limitations



Owens et al., *Audio-Visual Scene Analysis with Self-Supervised Multisensory Features*, ECCV 2018

Outputs are heatmaps



Arandjelović et al., *Objects that Sound,* ECCV 2018.

No class labels

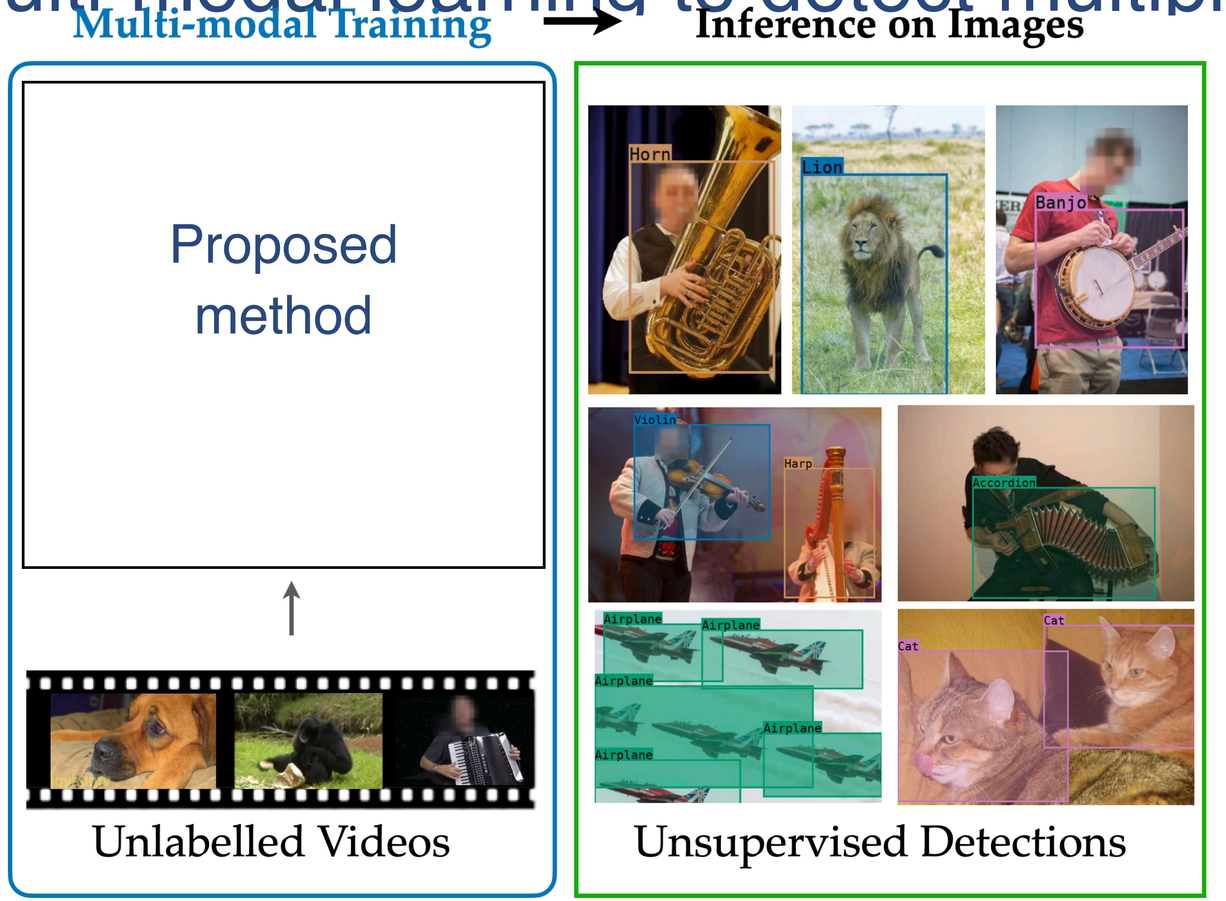
Inference requires audio

Goal of this paper:

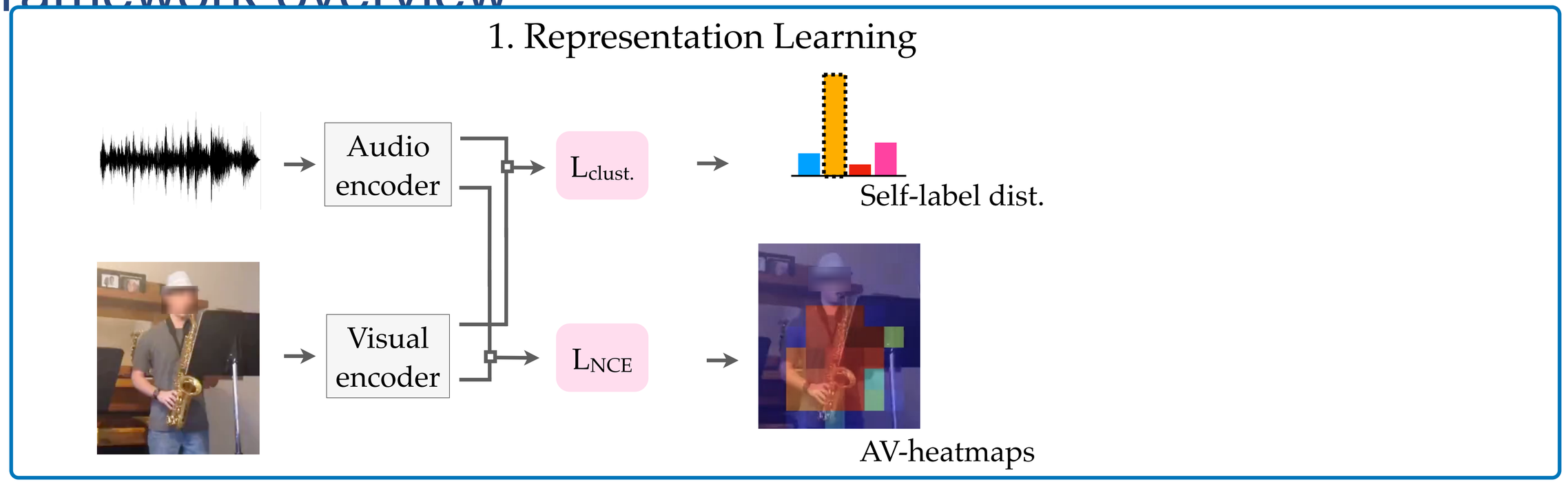
Combining self-labelling and multi-modal learning to detect multiple

Training object detectors without labels:

- ✓ Use only free audio as “supervision”
- ✓ Output bounding boxes & class labels
  - — not just heatmaps
- ✓ No audio required during inference
- ✓ Retrieve all visible instances, not just the actively sounding ones

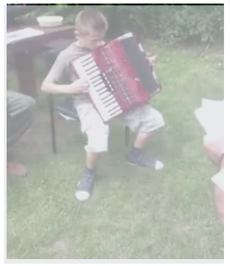


Framework overview

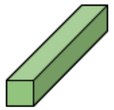


Ingredient 1: training heat maps

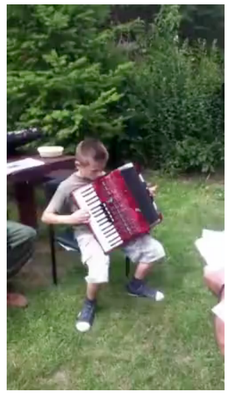
A_p



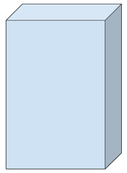
Audio | Pool



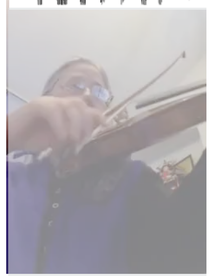
V



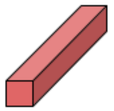
Visual



A_n



Audio | Pool



H_p

Contrastive loss:

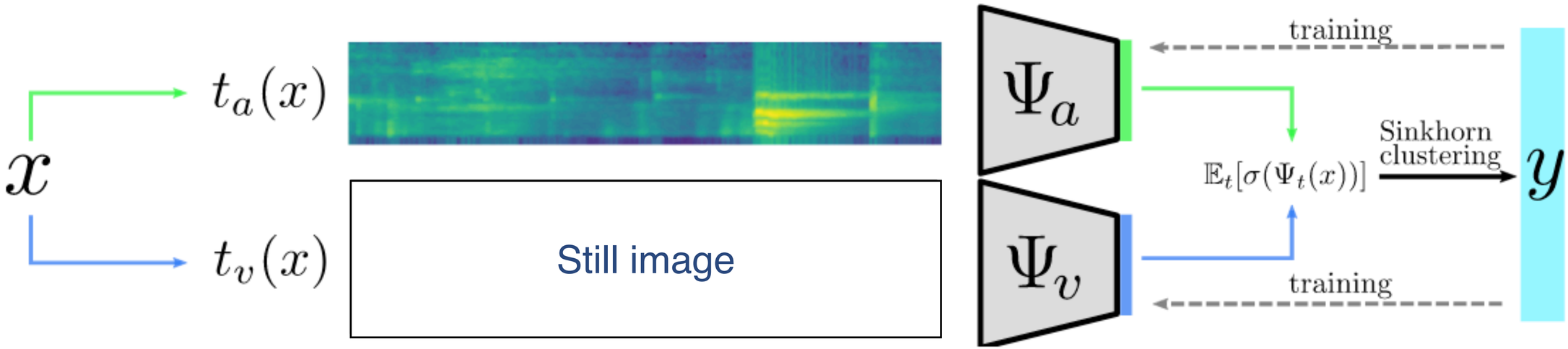
Positive pairs: V & A_p from same clip

Negative pairs: V & A_n from different clip

H_n

$$\mathcal{L}_{NCE} = \log\left(\frac{\max\limits_{h_i \in \mathcal{H}_p} e^{h_i}}{\max\limits_{h_i \in \mathcal{H}_p} e^{h_i} + \max\limits_{h'_i \in \mathcal{H}_n} e^{h'_i}}\right)$$

Ingredient 2: Self-labels training ($L_{clust.}$)



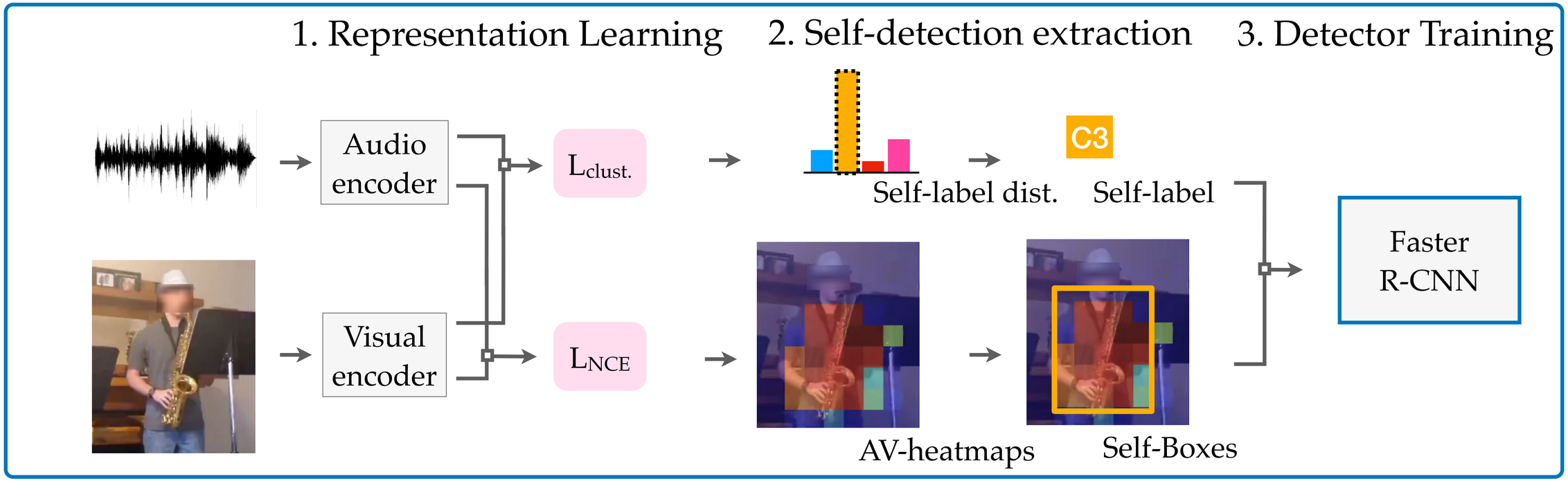
$$\mathcal{L}_v(\mathcal{B}|y) = -\frac{1}{|\mathcal{B}|}\sum_{(v,a)\in\mathcal{B}} \log \text{softmax}(y(v,a)\,|\Psi_v(v))$$

$$\mathcal{L}_a(\mathcal{B}|y) = -\frac{1}{|\mathcal{B}|}\sum_{(v,a)\in\mathcal{B}} \log \text{softmax}(y(v,a)\,|\Psi_a(a))$$

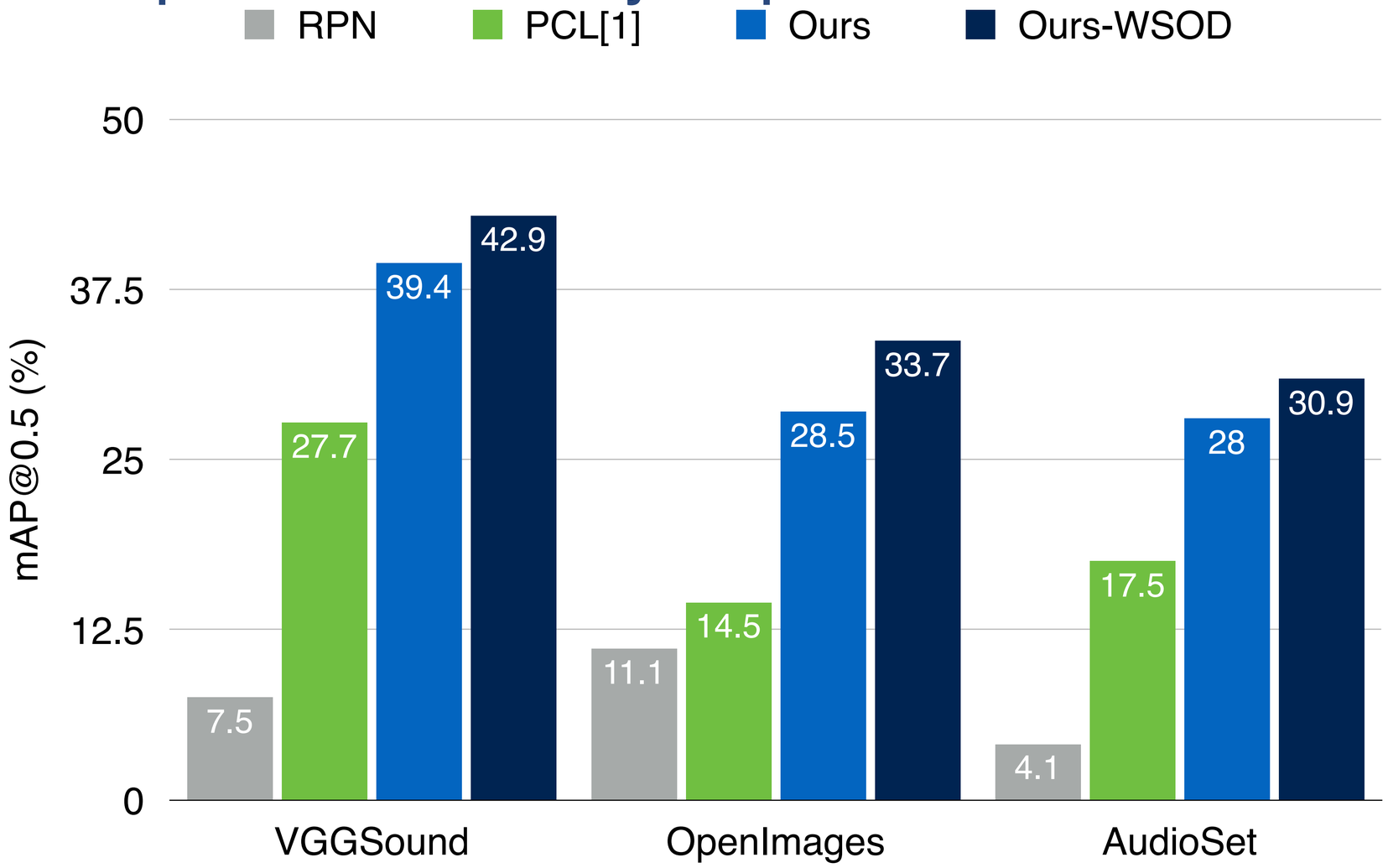
$$\mathcal{L}_{\text{clust}}(\mathcal{B}|y) = (\mathcal{L}_v(\mathcal{B}|y) + \mathcal{L}_a(\mathcal{B}|y))/2$$

See Asano et al., NeurIPS 2020

Framework overview



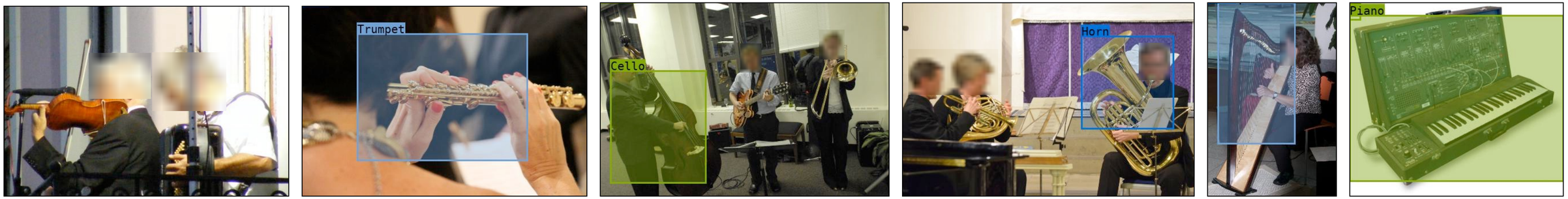
Results compared to weakly-supervised baseline and region



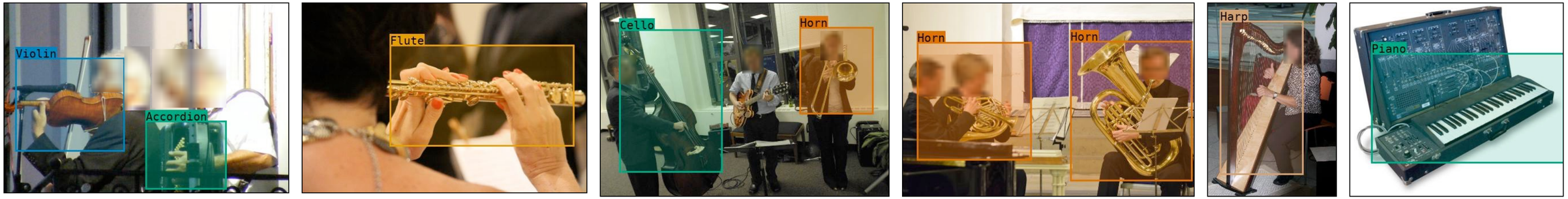
[1] PCL: Proposal Cluster Learning for Weakly Supervised Object Detection. Tang et al., TPAMI 2018

Qualitative results compared to weakly-supervised

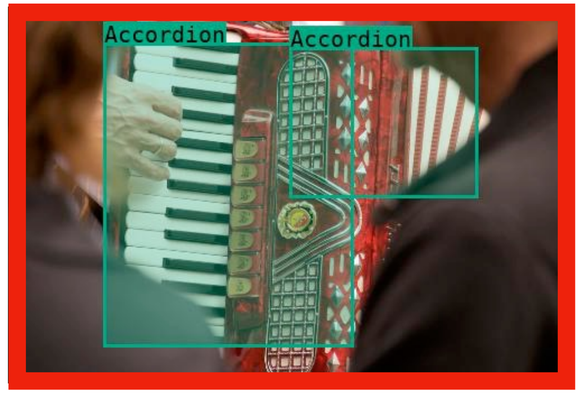
PCL

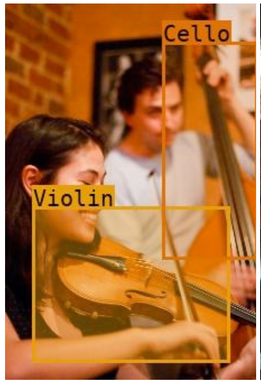


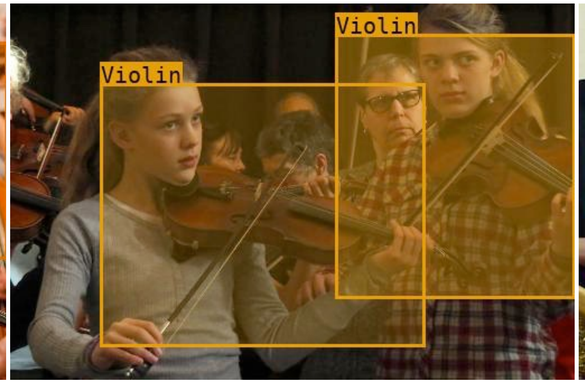
Ours

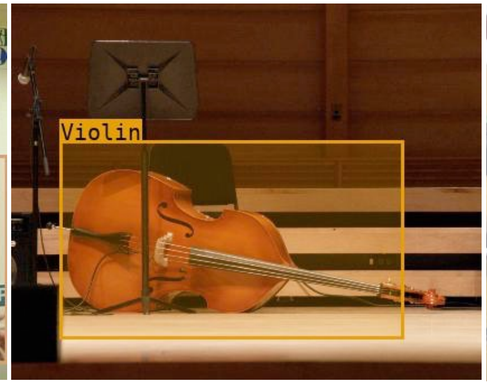


Detection examples and failure cases







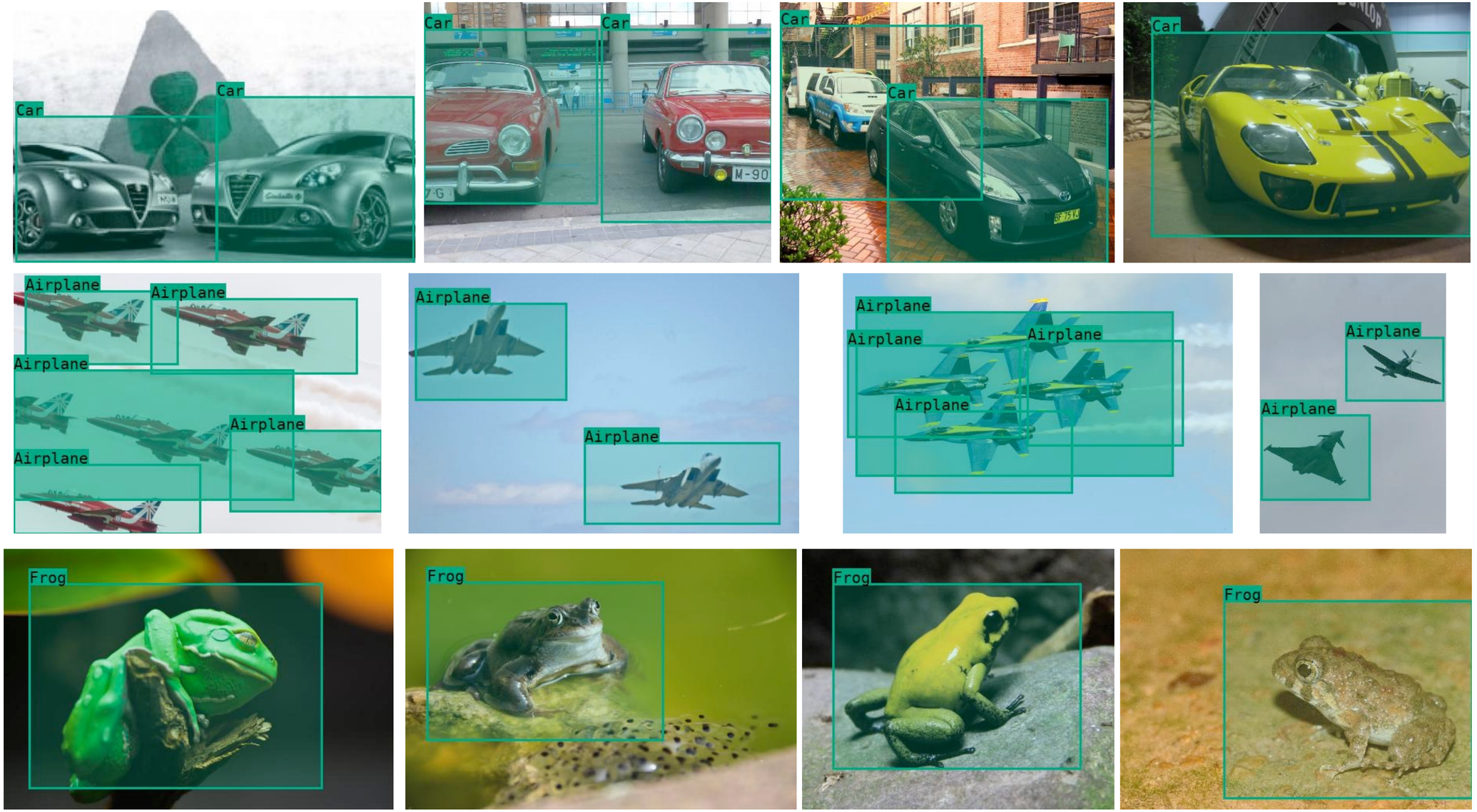


What about more general objects, beyond instruments?

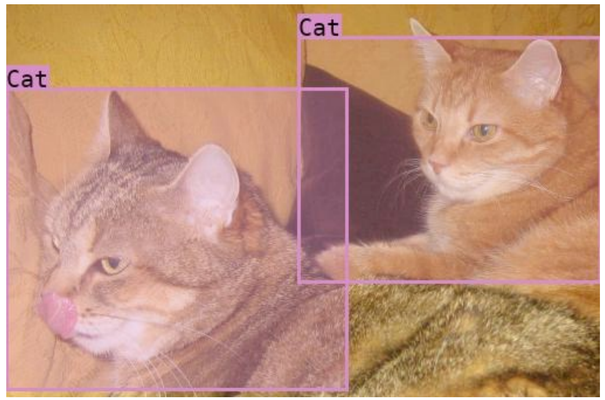
We train on all ~300 VGGSound classes, learning 300 clusters.

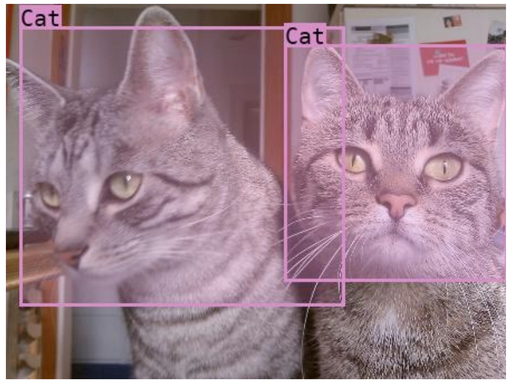
- Same method — no changes at all
- Only match class labels *after* the detector is trained
- Match class labels with as few as 1 sample per cluster (ie 300 “labels”)

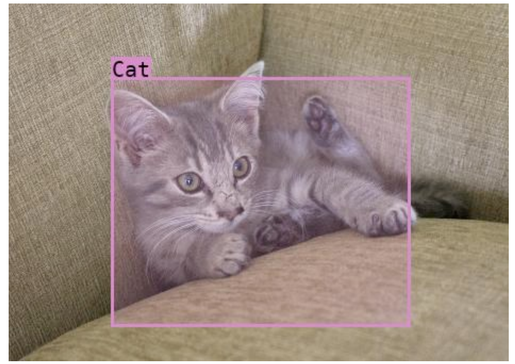
VGGSound object detections

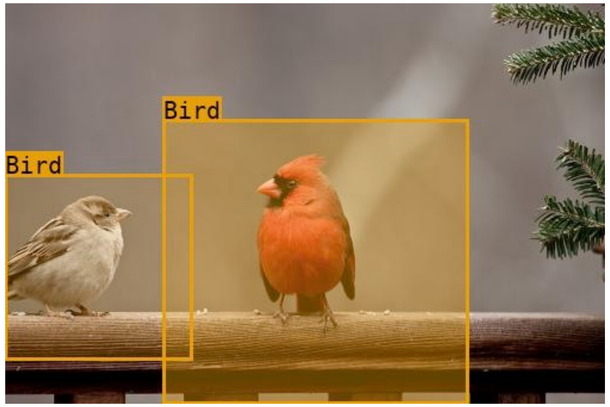


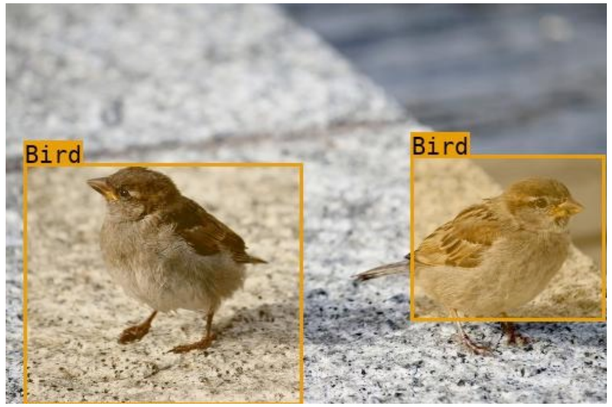
VGGSound object detections

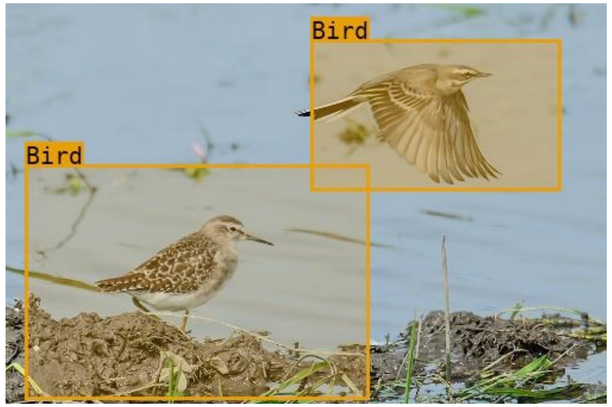


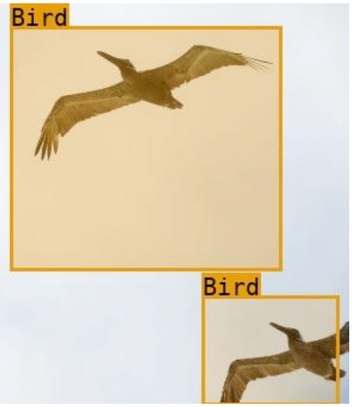


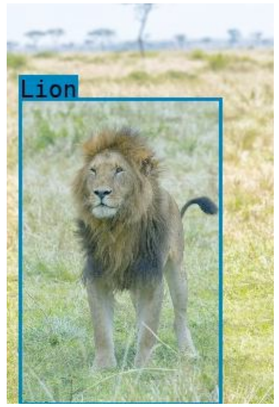




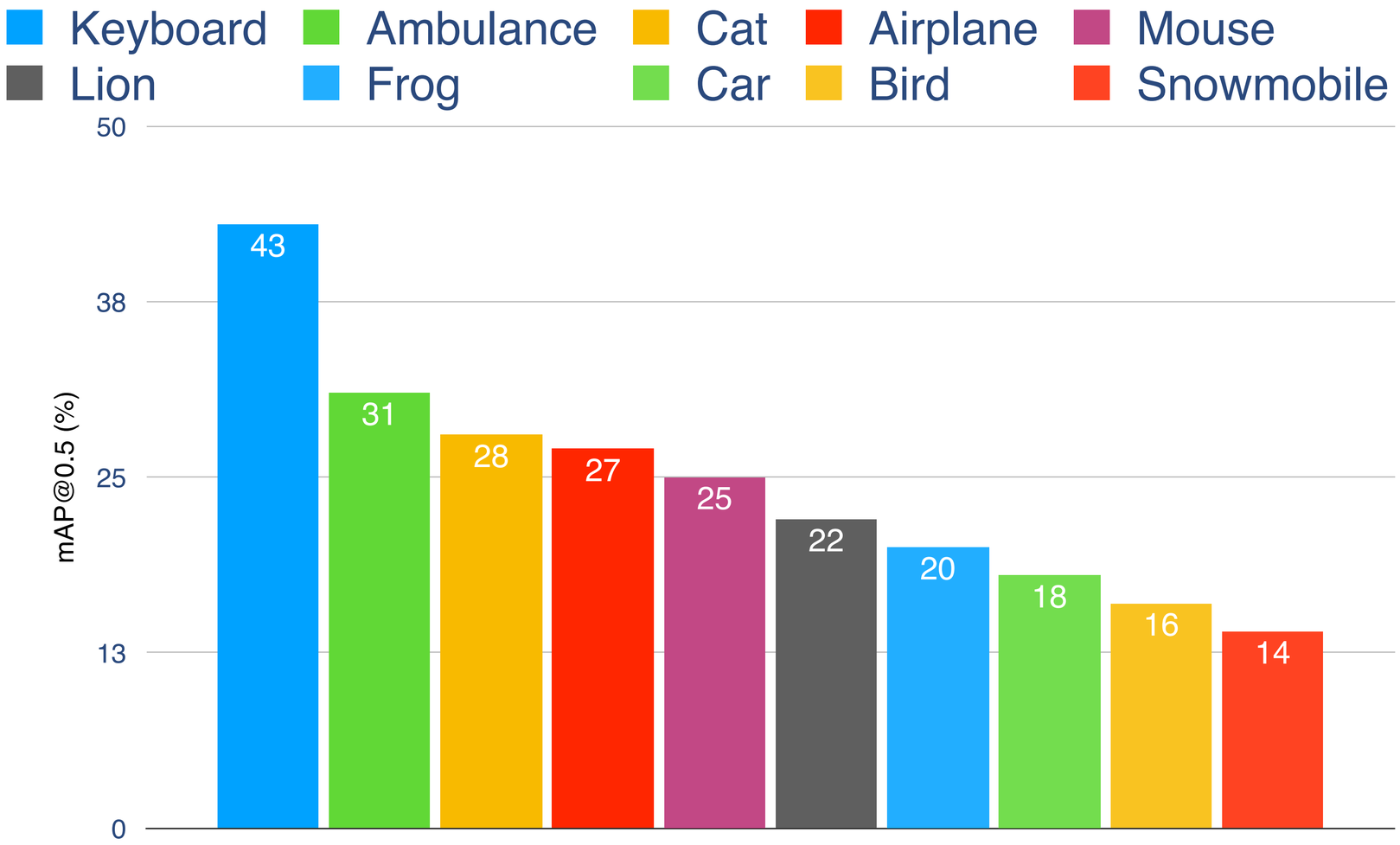








Per class performances



- Still long way to go for most objects
- mAP @0.5 is a hard measure

Extrapolating from a Single Image to a Thousand Classes using Distillation

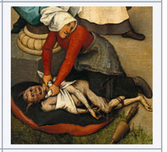
YUKI M. ASANO* & AAQIB SAEED*

Netherlandish Proverbs by Pieter Bruegel the Elder, 1559, Public Domain

Dutch folklore

List of proverbs and idioms featured in the painting [edit]

Expressions featured in the painting[10][11]

	Proverb/idiom	Meaning	Area	Image
001	To be able to tie even the devil to a pillow (fr) (nl)	Obstinacy overcomes everything	Lower left	
002	To be a pillar-biter (fr) (nl)	To be a religious hypocrite	Lower left	
003	Never believe someone who carries fire in one hand and water in the other (fr) (nl)	To be two-faced and to stir up trouble	Lower left	

≈

Some computer vision

Deep Learning requires labels
❌ *Self-supervised learning.*

Object detection requires annotations
❌ *Afouras CVPR'22, Simeoni BMVC'21 etc*

Augmentations ≈ enlarging dataset
❌ *Asano ICLR'20, He CVPR'21 etc.*

Deep Learning requires a lot of data
❌ *This work, let's specify this.*

https://en.wikipedia.org/wiki/Netherlandish_Proverbs;
Self-supervised object detection from audio visual correspondence. Afouras et al. CVPR'22
Localizing Objects with Self-Supervised Transformers and no Labels. Simeoni et al. BMVC'21
A critical analysis of self-supervision, or what we can learn from a single image. Asano et al. ICLR'20
Momentum Contrast for Unsupervised Visual Representation Learning. He et al. CVPR'20

Problem setting

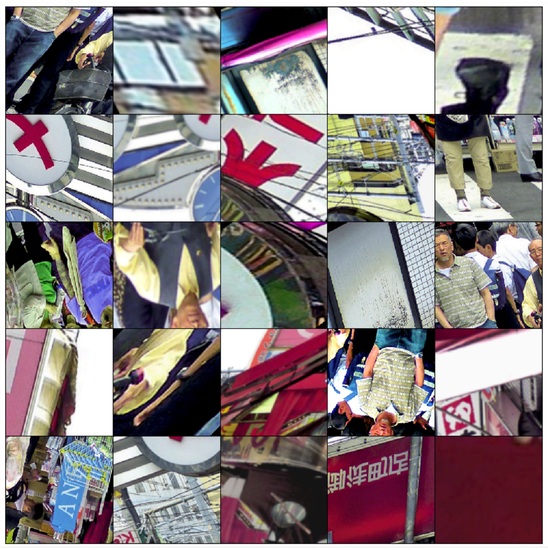
To bang one's head against a brick wall -- To try to achieve the impossible

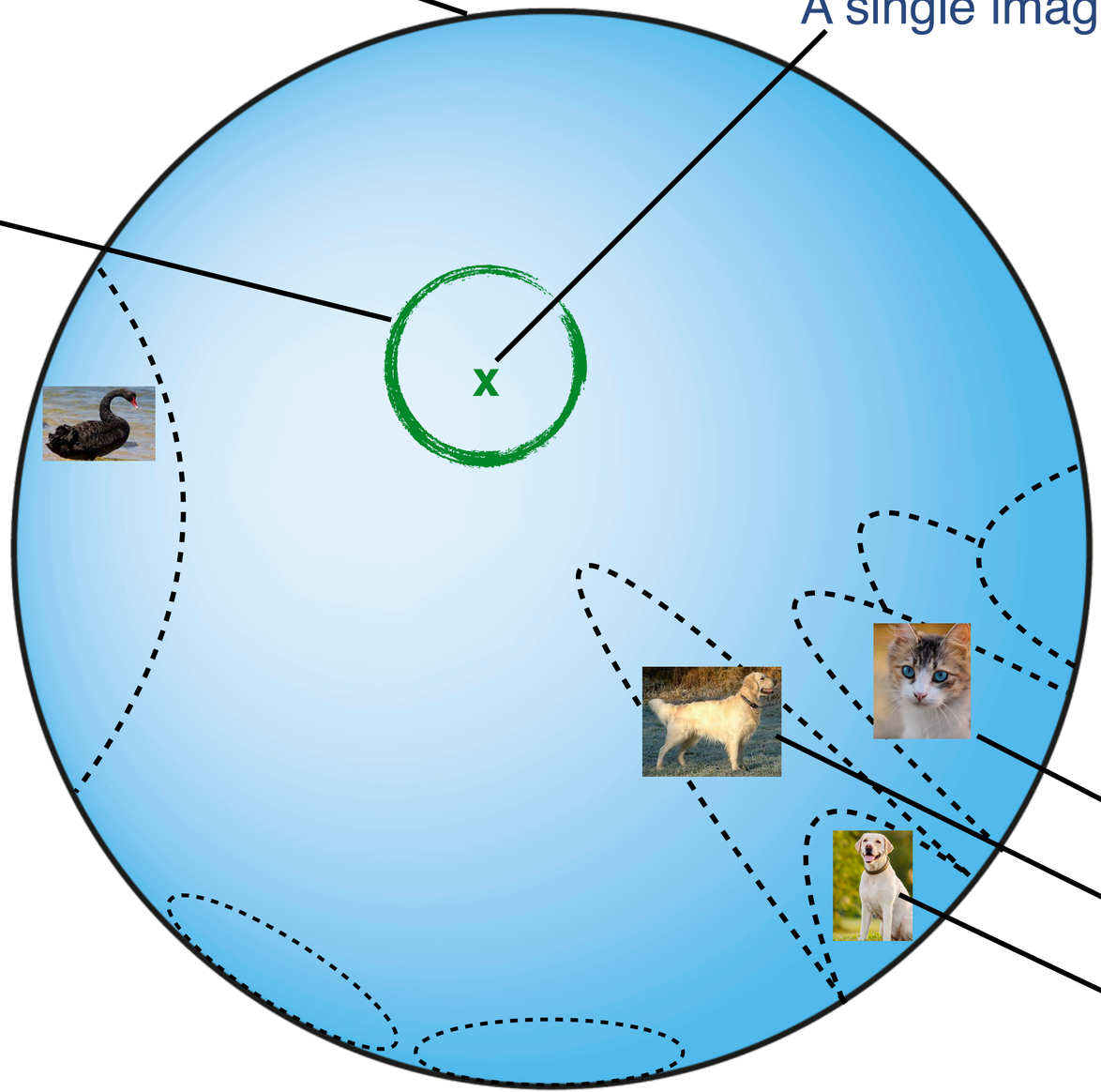
Extrapolating from a single image to semantic categories

All natural images

A single image I

Augmentations $\mathcal{A}(I)$





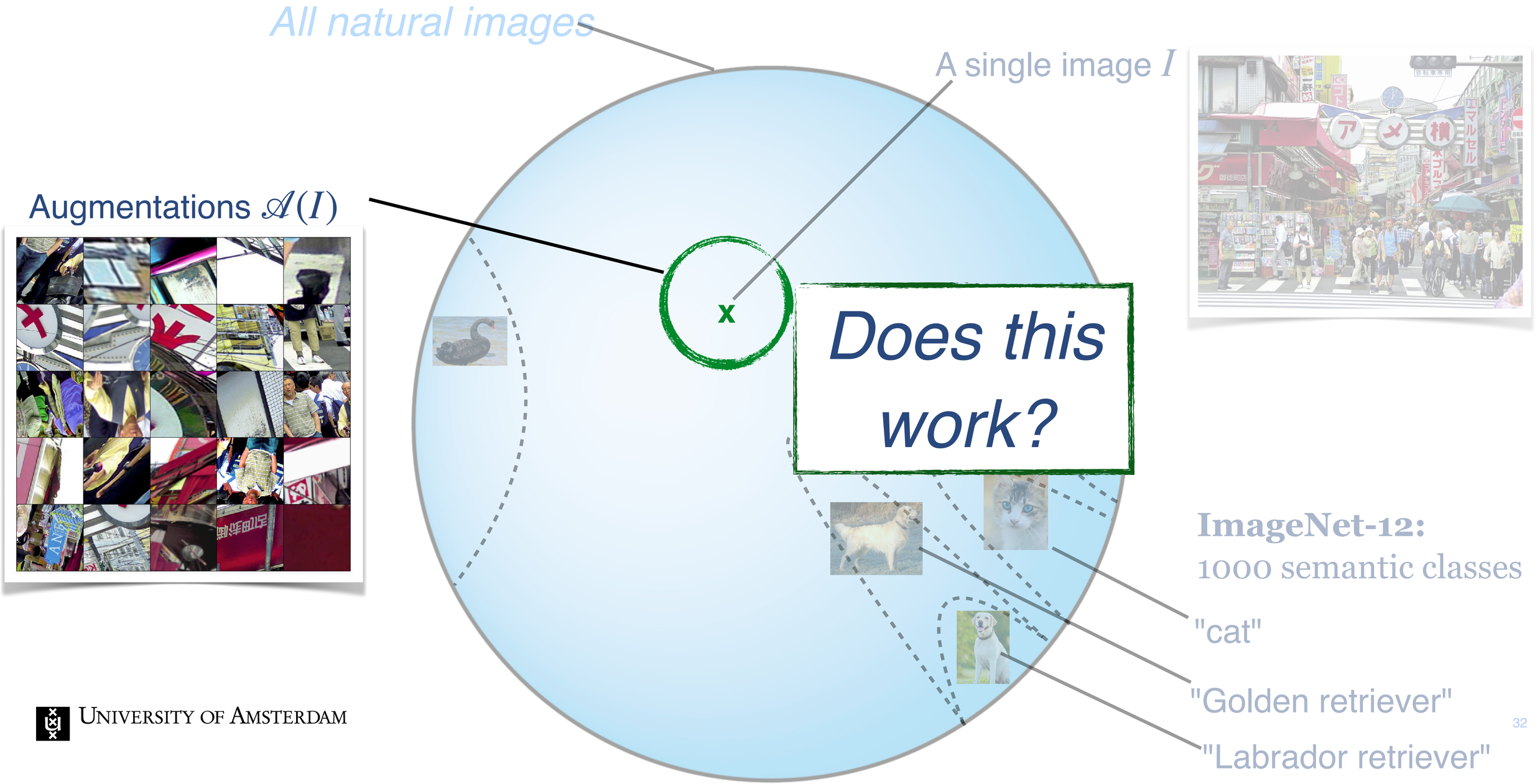
ImageNet-12:
1000 semantic classes

"cat"

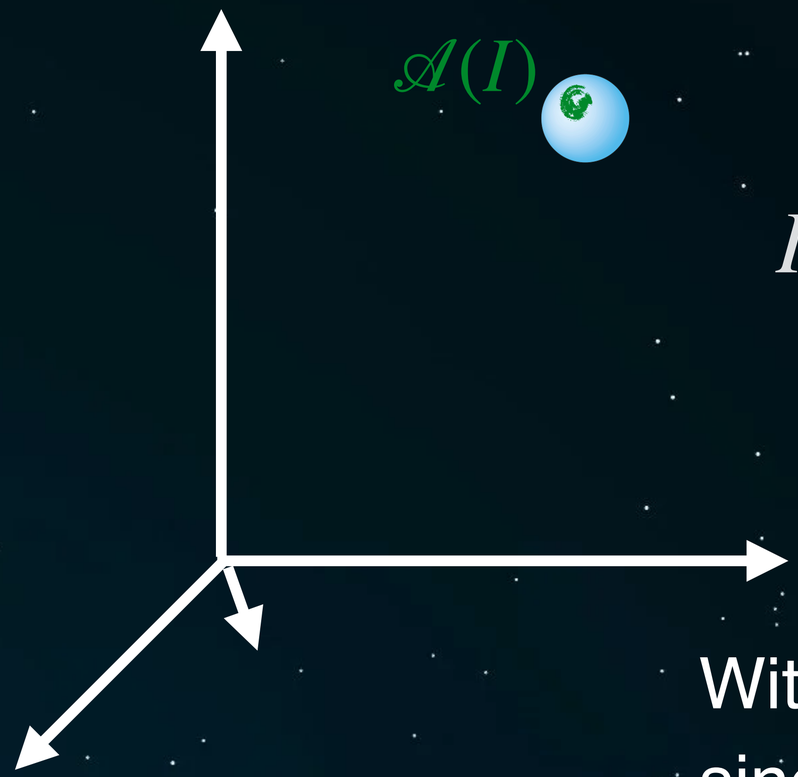
"Golden retriever"

"Labrador retriever"

Extrapolating from a single image to semantic categories



Why it might work: *The single image prior*



$$I \in \mathcal{I}, \quad \mathcal{I} = \{0,..,255\}^{3\times224\times224}$$

Within the space of all possible images $\mathcal{I}$, a single real image I and its augmentations $\mathcal{A}(I)$ provide a very informative prior about all real images

Method

To be suspended between heaven and earth -- To be in an awkward situation

Requirements

- Generic method
- No optimization over the choice of the single image
- Need to infuse the model with semantic categories
- Model needs to be trained from scratch
- Model is only allowed to be trained on image + augmentations

→ Knowledge distillation from a teacher model

Method overview

Results

To sit on hot coals -- To be impatient

Comparison of *datasets (number of pixels)*

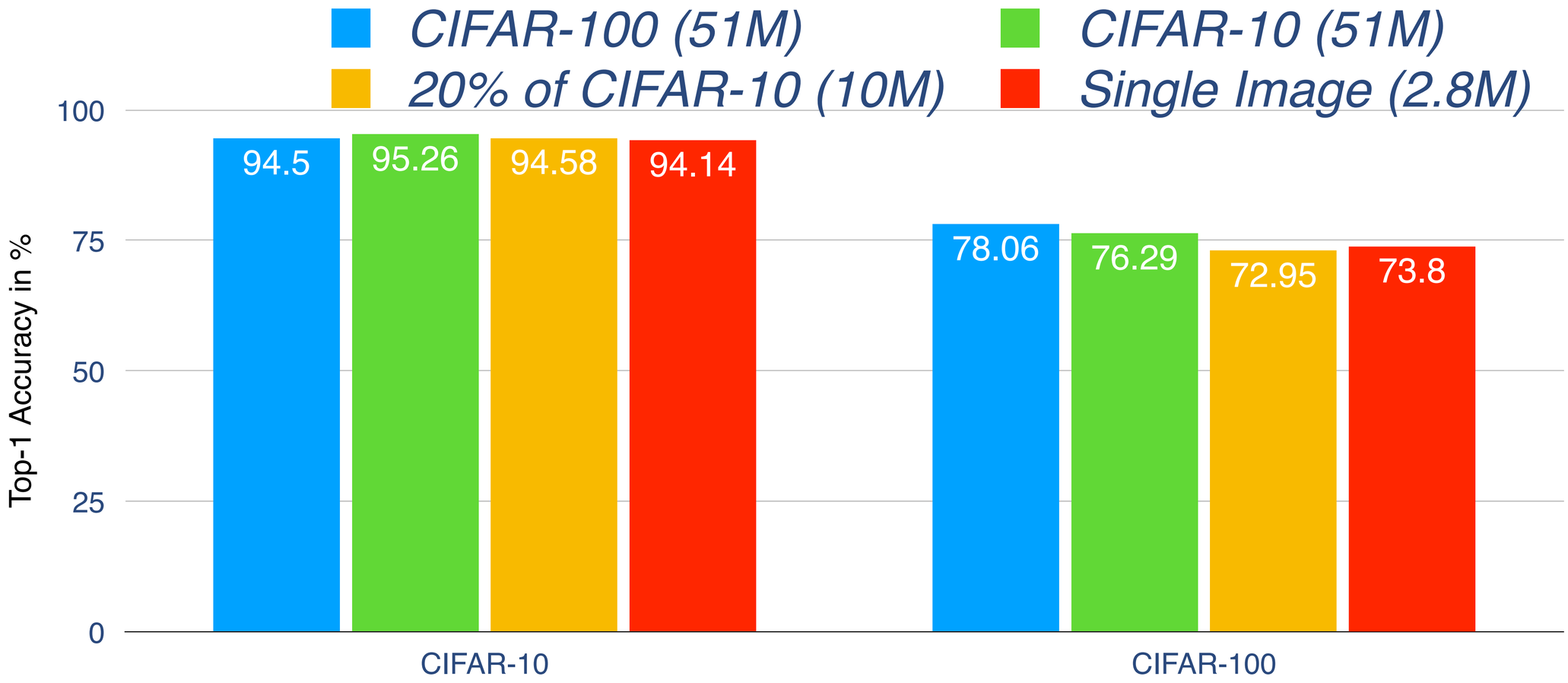
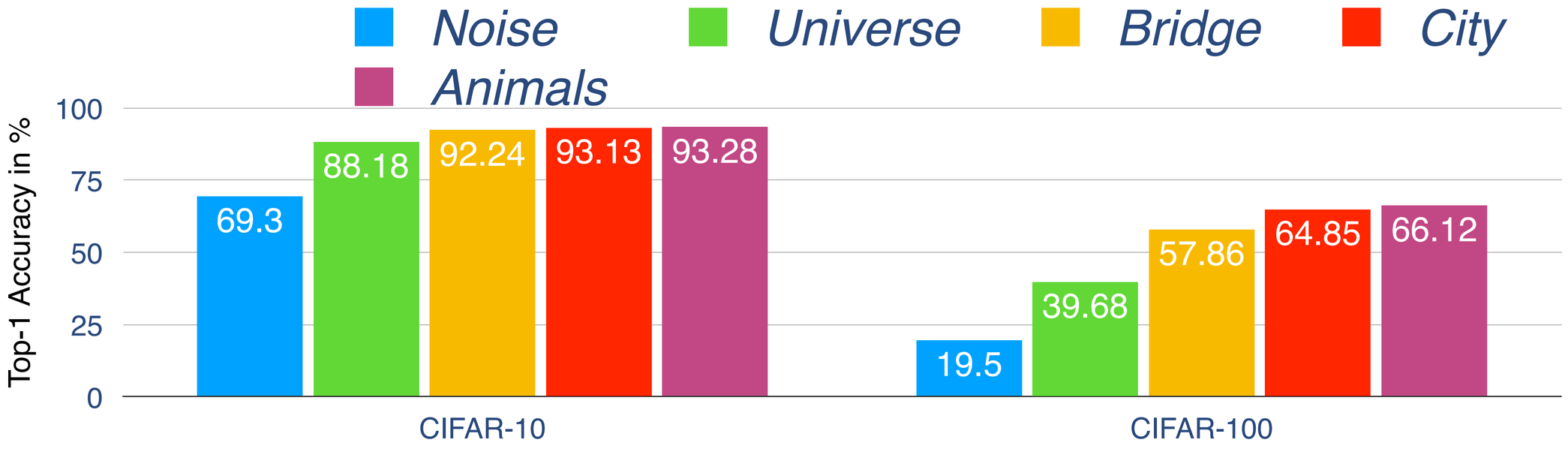
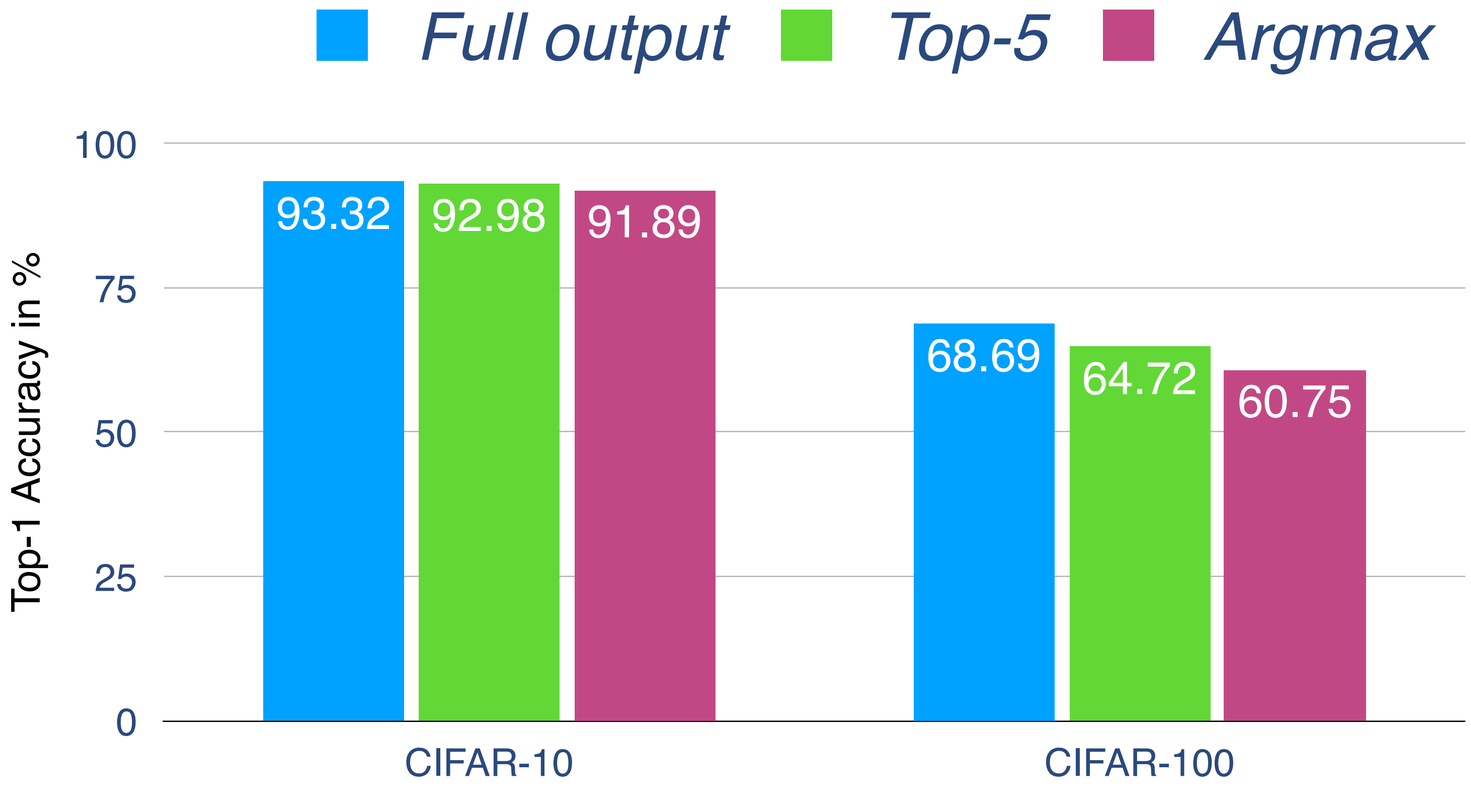


Image choice matters.



Learning signal: even top-5 or argmax works well.



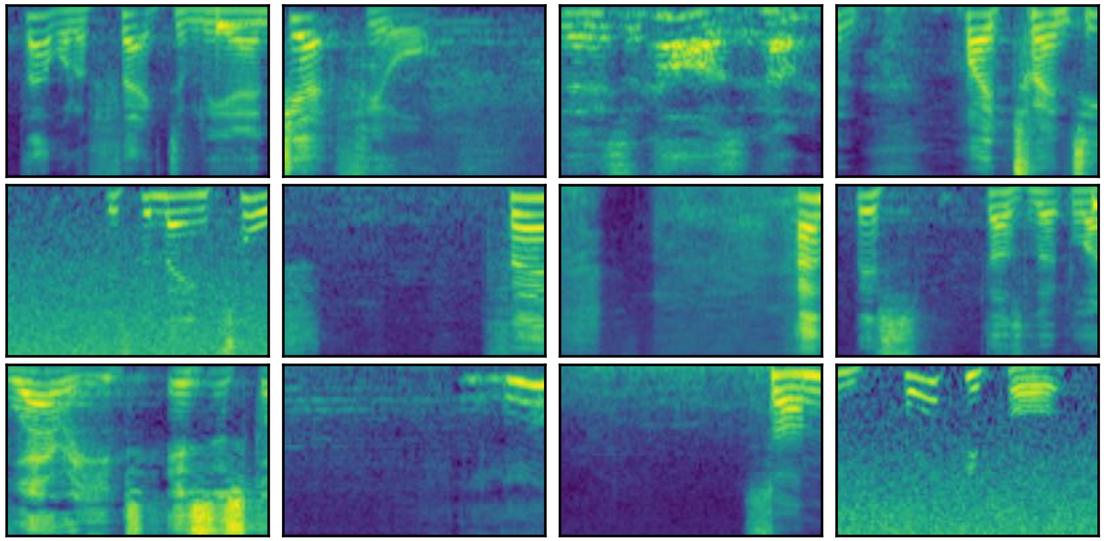
Even with only top-5 predictions (and confidence) or hard distillation, performance only slightly degrades.

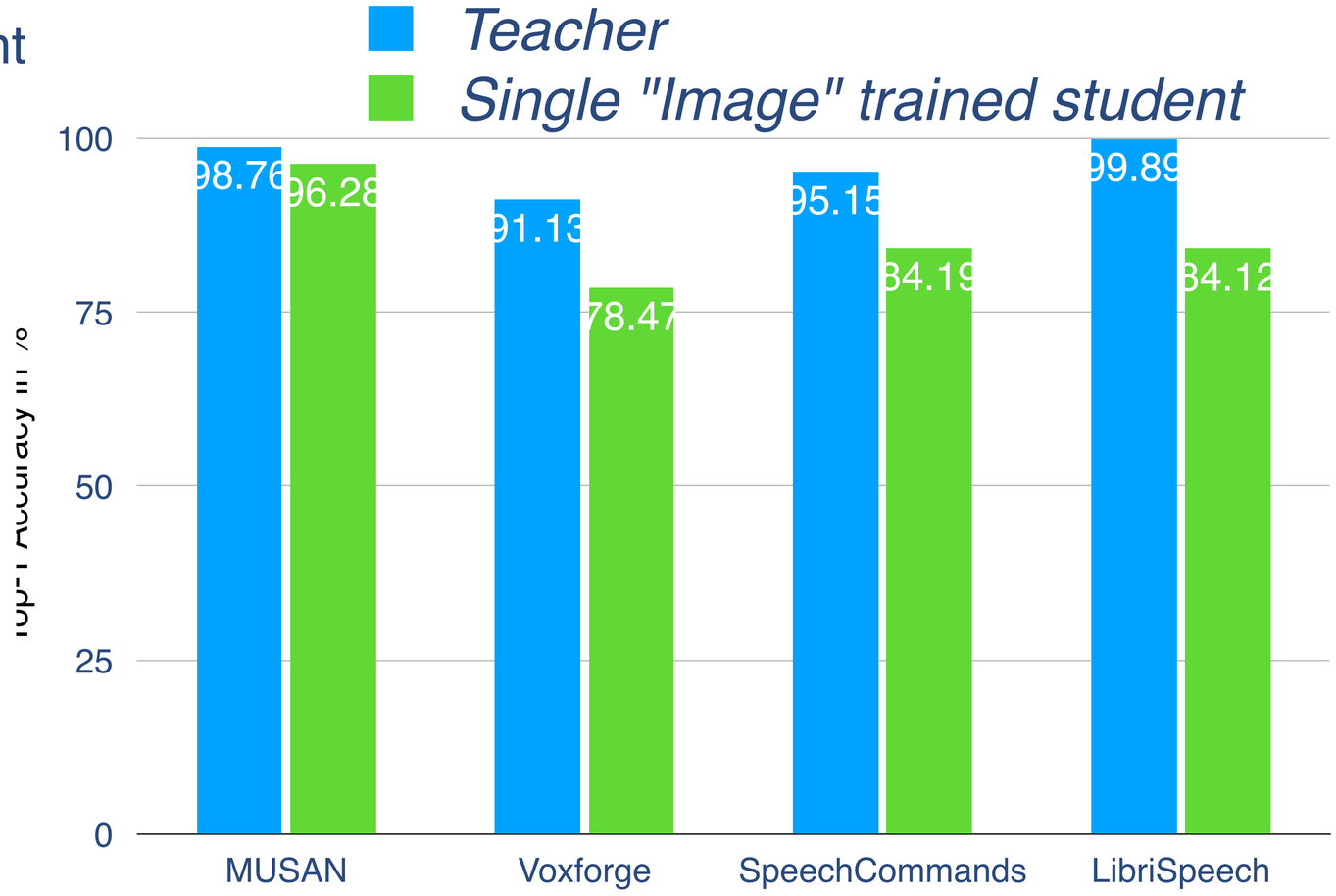
⚠️ API providers! (c.f. Orekondy et al.)

Orekondy et al. Knockoff nets: Stealing functionality of black-box models. CVPR 2019

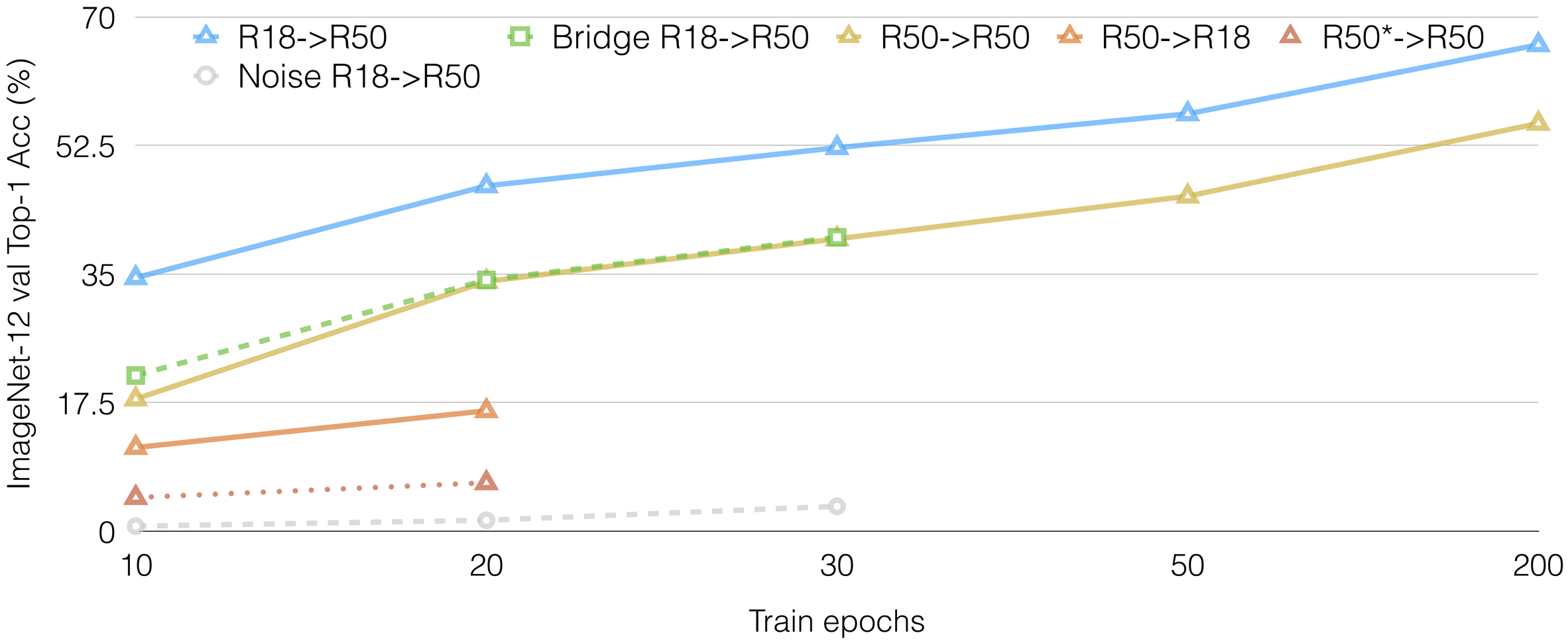
Generalizing to audio.

Extract audio & augment

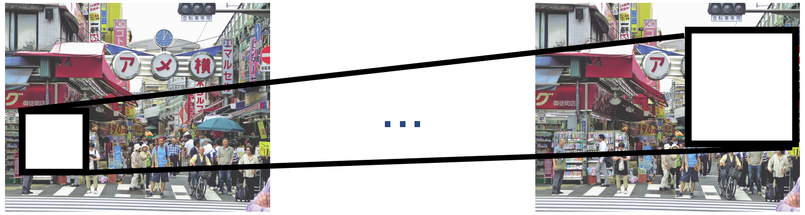


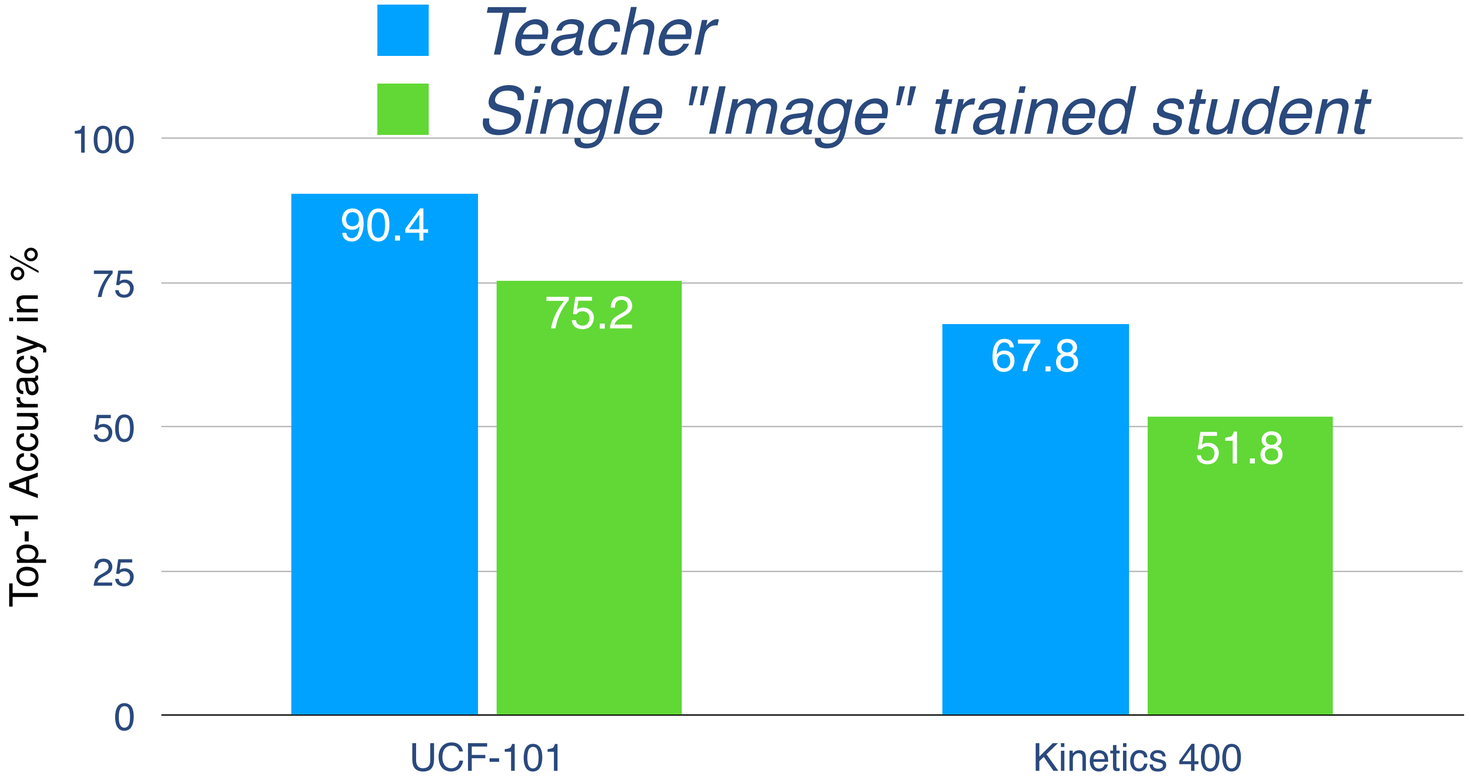


Teacher-student performance comparison

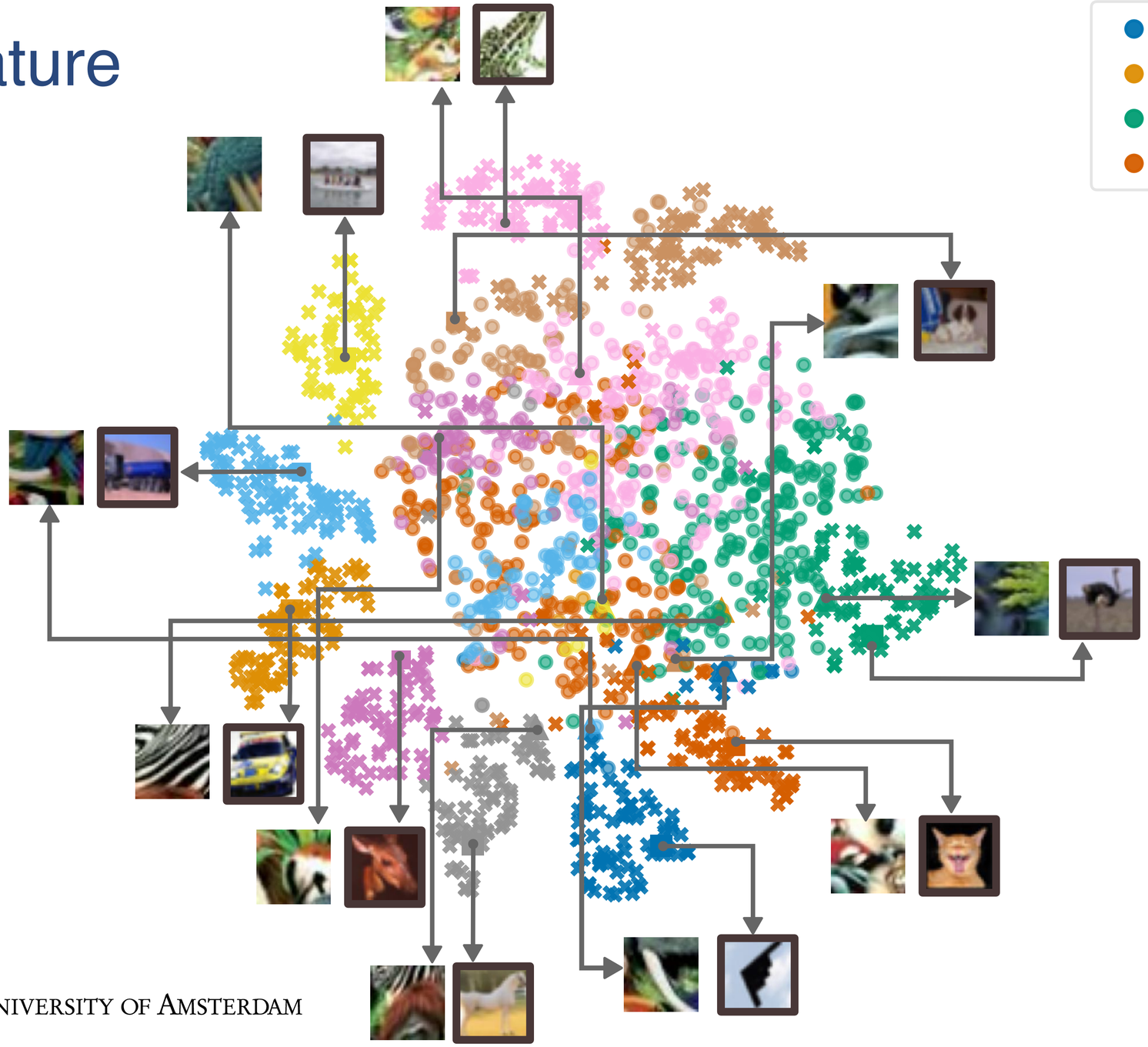


Scaling to video.



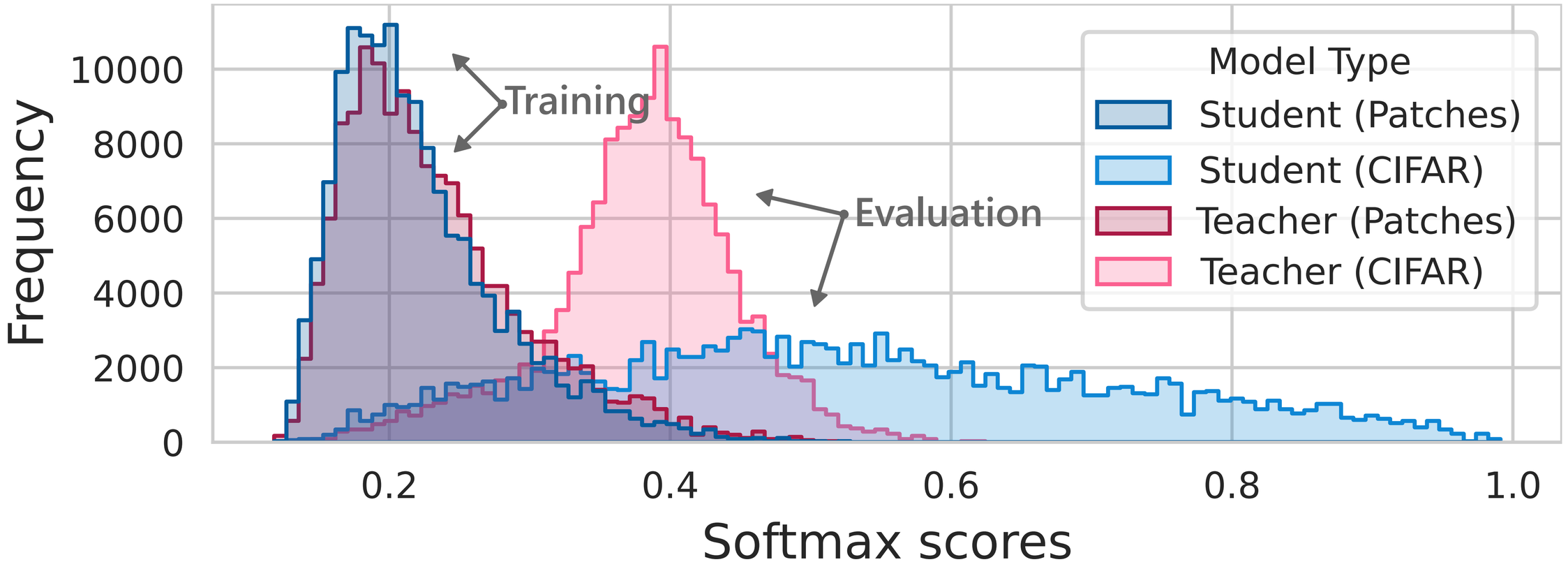


Feature

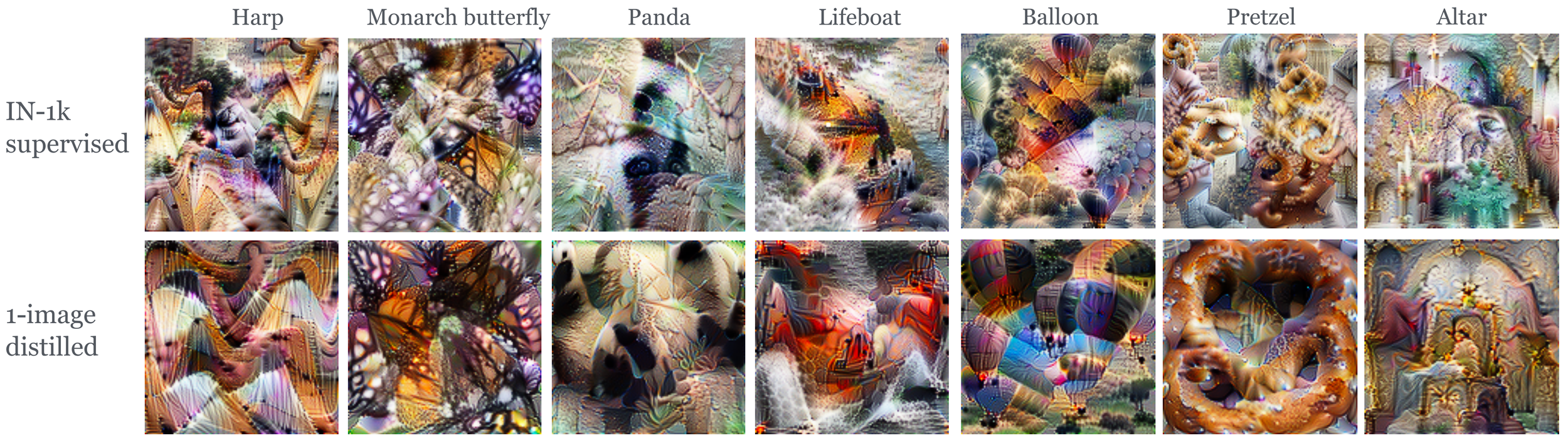


- Semantic extrapolation works
- patches (·) are "inside", real images (×) "outside"

Student learns something different.



Neuron activation maximisation



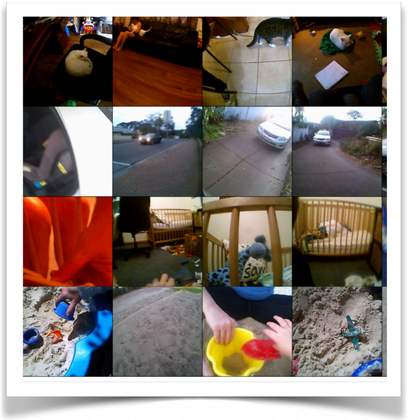
For intermediate layers, see paper.

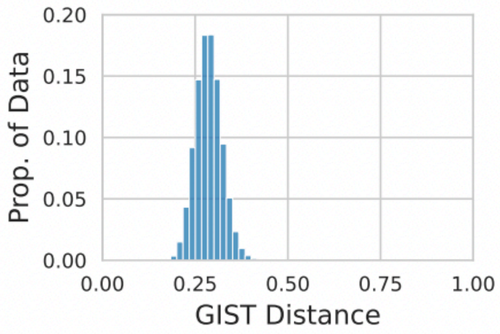
Conclusion

Two dogs over one bone seldom agree -- To argue over a single point

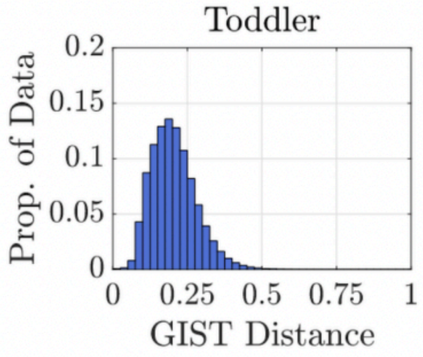
Fundamental research -- yes. But also potential applications.

Why does it work?





(a) **Our 1-image dataset.**



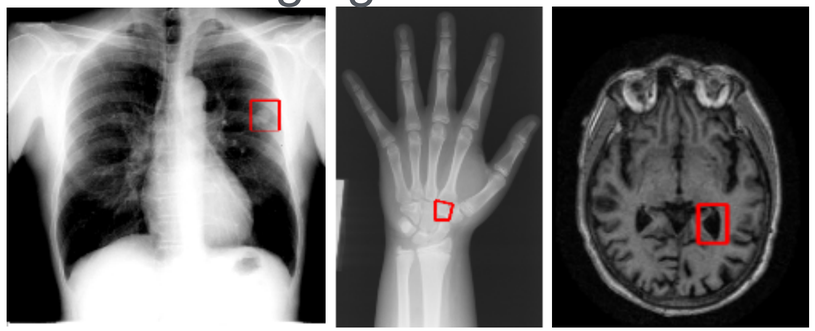
(b) **Toddler data.** Figure from [3].

Pruning & Quantization (Tab 16 in Appx)

Dataset	Model	Standard	Quantization *(source)*	Ours - Quantization *(single-image)*	Pruning *(source)*	Ours - Pruning *(single-image)*
CIFAR-10	ResNet56	93.77	93.64	93.45	93.38	93.18
	VGG11	91.57	91.22	90.97	91.14	90.86
	VGG19	93.28	92.94	93.00	92.84	92.96
	WideRNet16-4	94.81	94.40	94.59	94.76	94.43
	WideRNet40-4	95.42	95.09	94.90	95.29	94.82
CIFAR-100	ResNet56	70.99	70.89	69.85	70.74	70.42
	VGG11	69.65	69.92	68.86	69.77	69.13
	VGG19	70.79	70.60	70.21	70.89	70.56
	WideRNet16-4	75.81	75.45	74.71	75.68	75.51
	WideRNet40-4	78.14	78.27	77.60	77.94	77.78

Table 16. **Self-distillation with a single-image for efficient deep models**. We perform model compression via self-distillation using source data and 50k random patches generated from 'Animals' image for 50% sparsity (in case of pruning) and 8-bits quantization without any noticeable loss in performance.

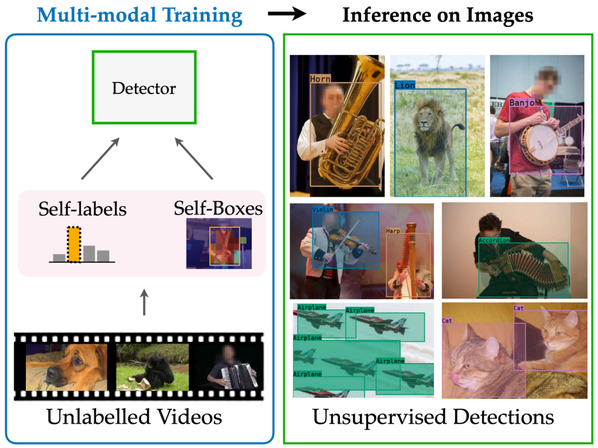
Medical imaging



Visual recap of this talk

Combining self-labelling and multi-modal learning to detect multiple objects

Training object detectors without labels:
- ✓ Use only free audio as “supervision”
- ✓ Output bounding boxes & class labels — not just heatmaps
- ✓ No audio required during inference
- ✓ Retrieve all visible instances, not just the actively sounding ones



Extrapolating from a single image to semantic categories