Self-supervised learning

CMPSCI 682: Neural Networks: A Modern Introduction

Chuang Gan

College of
INFORMATION AND
COMPUTER SCIENCES

UMASS
AMHERST

Today’s Class

- Recap
  - Supervised vs Unsupervised Learning
  - Why not always label data?
- Semi-supervised Learning
  - Concepts
  - Example: pseudo-labels / self-training
- Self-supervised Learning
  - Concepts
  - Pretext tasks
  - Contrastive Learning
  - Beyond images

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Today's Class

- Recap
  - Supervised vs Unsupervised Learning
  - Why not always label data?
- Semi-supervised Learning
  - Concepts
  - Example: pseudo-labels / self-training
- Self-supervised Learning
  - Concepts
  - Pretext tasks
  - Contrastive Learning
  - Beyond images

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

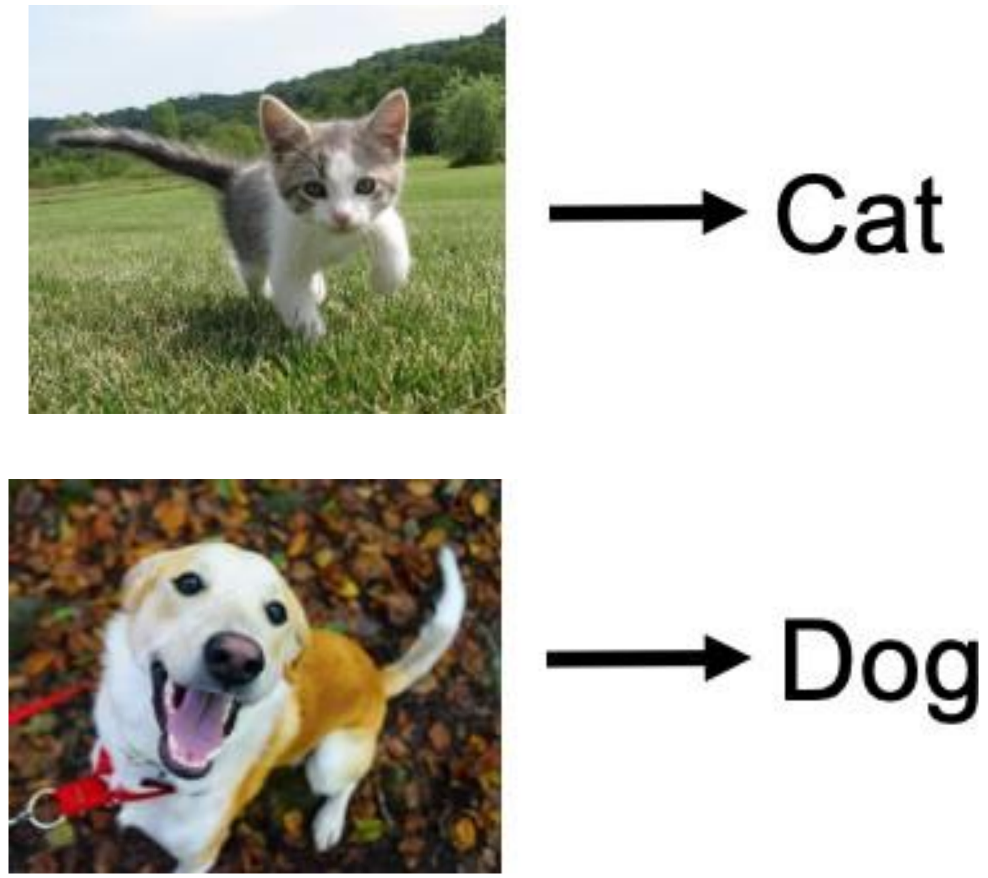
Recap: Supervised vs Unsupervised Learning

Supervised Learning

Data: (X, y)

X = input/feature/image/…

y = label/target



Unsupervised Learning

Data: X

Just X, no labels

Learn about the *structure* of the data, i.e. P(X)

…..

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

So let's always use Supervised Learning?

Supervised Learning

Data: (X, y)

X = input/feature/image/…

y = label/target

 → Cat

 → Dog

"Standard" Supervised Learning:

1. Collect a large set of data (images..) as the "training set"
2. Label each one as cat / dog / monkey / …
3. Train a model mapping image to label

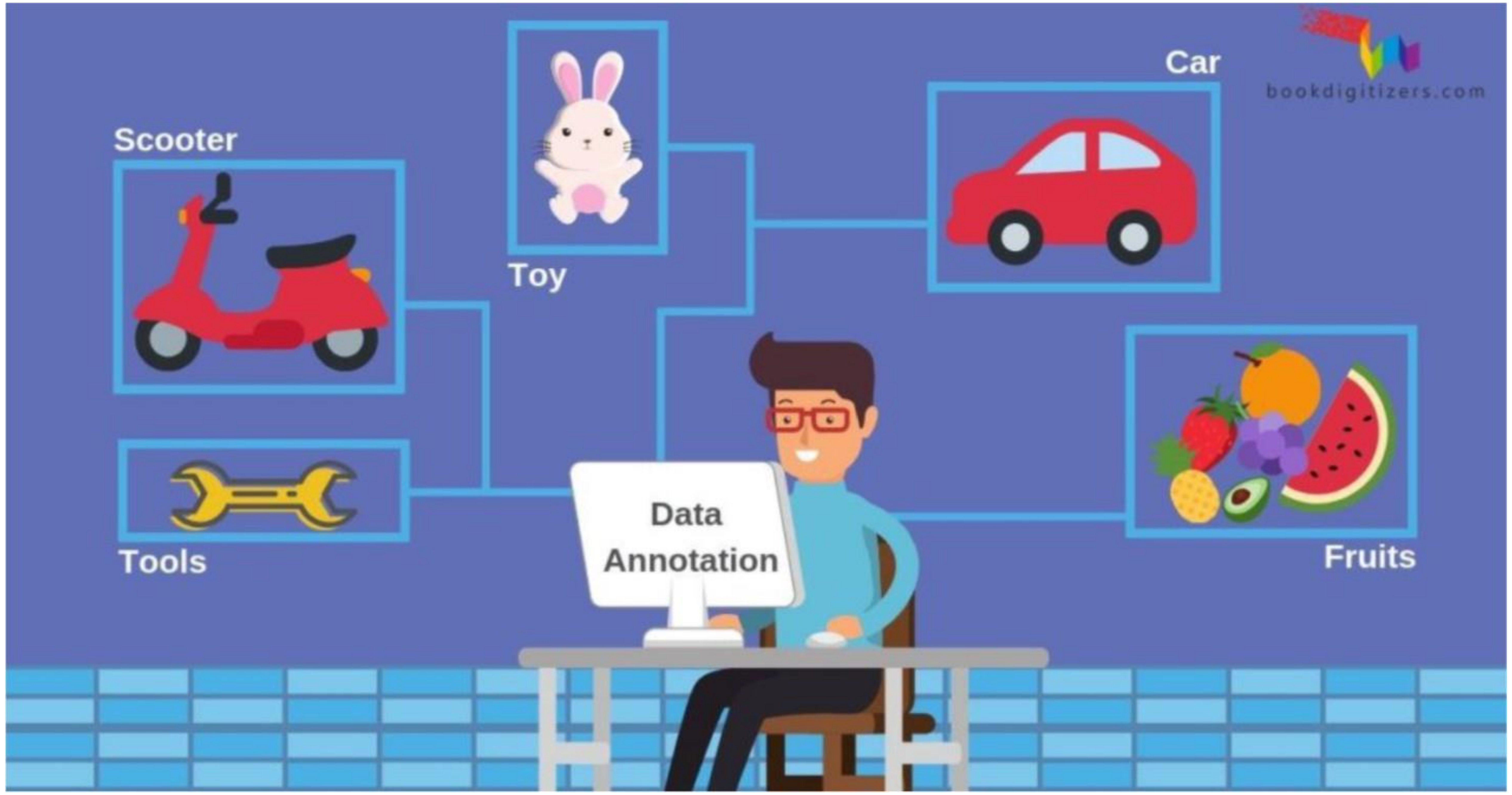
$$f : \mathbf{X} \rightarrow y$$

4. Go forth and classify the world with f !

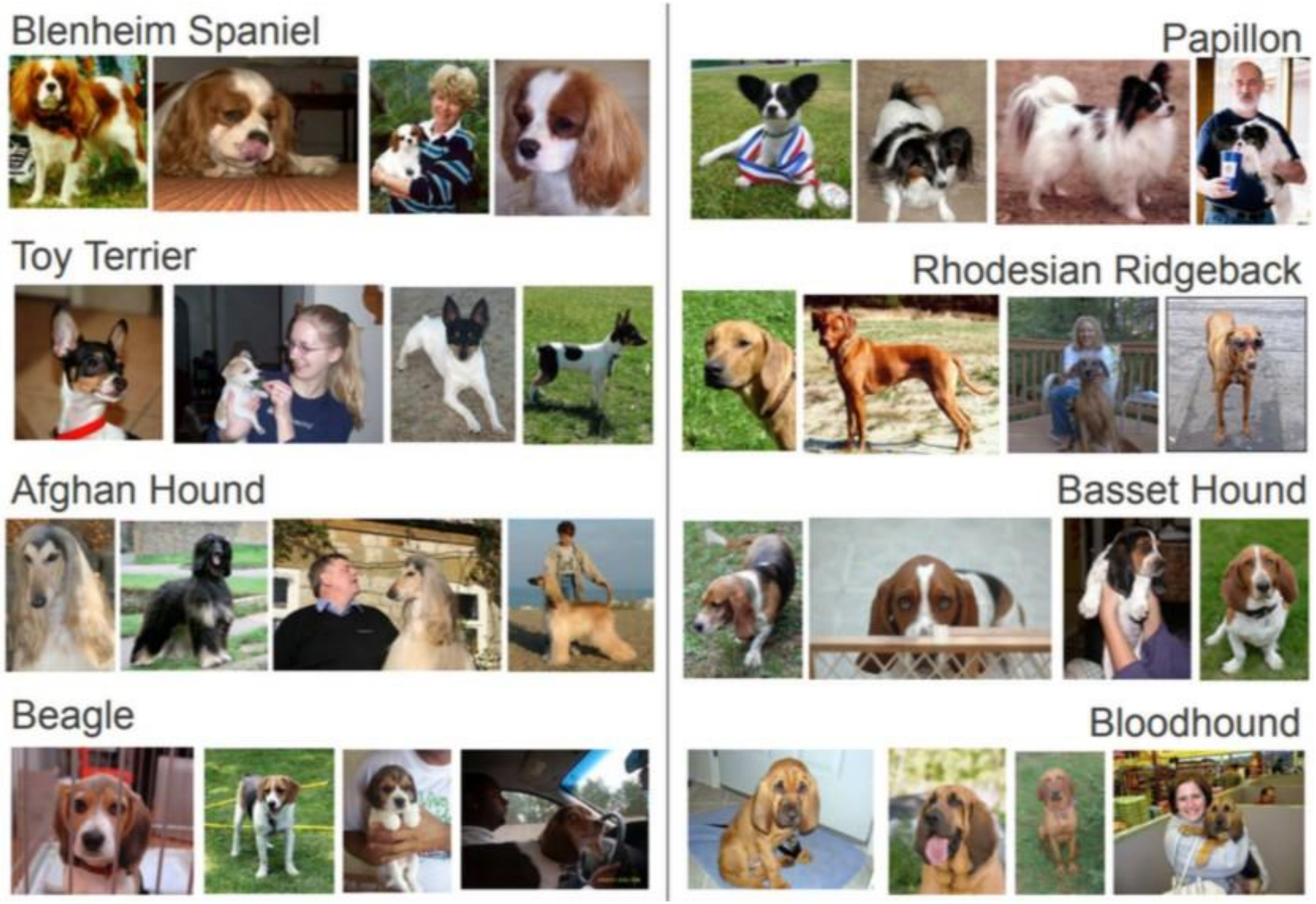
Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Data Annotation

Supervised Learning first requires labeling a very large amount of data



Labeling Image Categories - “Easy” Until ….



- Over **120 dog breeds** in ImageNet dataset for image classification
- Non-expert labelers may not be aware of these **fine-grained** differences, leading to ***labeling errors***
- ***E.g.,*** the Caltech UCSD birds dataset has 4% labeling error (NABirds, Van horn et al. CVPR15)

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

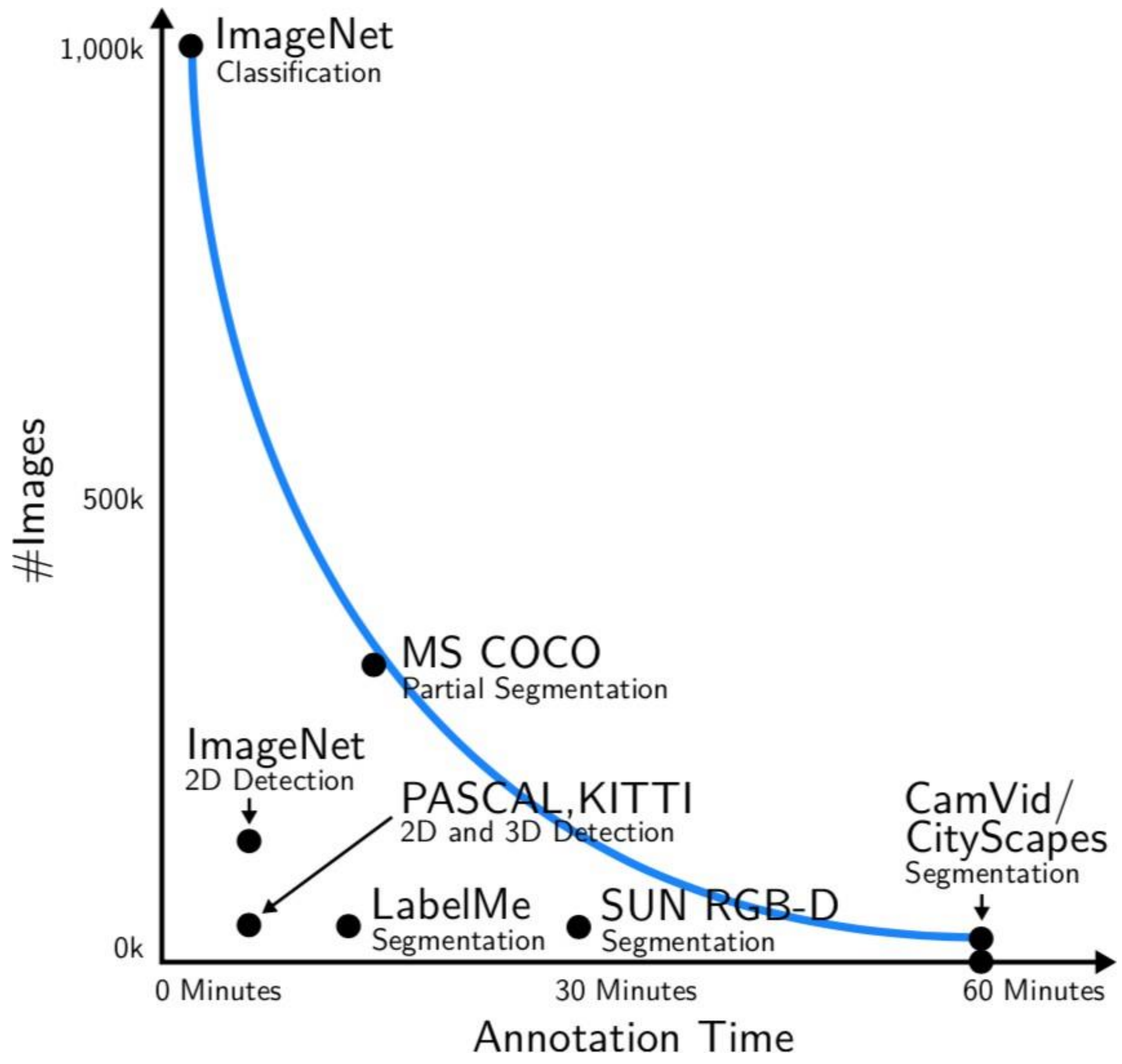
Dense Semantic and Instance Labels

“Cityscape” dataset: Labeling every pixel as person/road/sidewalk …

Annotation time **60-90 minutes per image**

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Annotate Everything — Expensive, doesn't Scale!



Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Motivation - Humans learn with little supervision

Provided with very few “labeled” examples (someone pointing something out to us explicitly), we can generalize quite well.

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Today’s Class

- Recap
  - Supervised vs Unsupervised Learning
  - Why not always label data?
- Semi-supervised Learning
  - Concepts
  - Example: pseudo-labels / self-training
  - Example: Distillation, Student/Teacher
- Self-supervised Learning
  - Concepts
  - Pretext tasks
  - Contrastive Learning

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Semi-supervised Learning

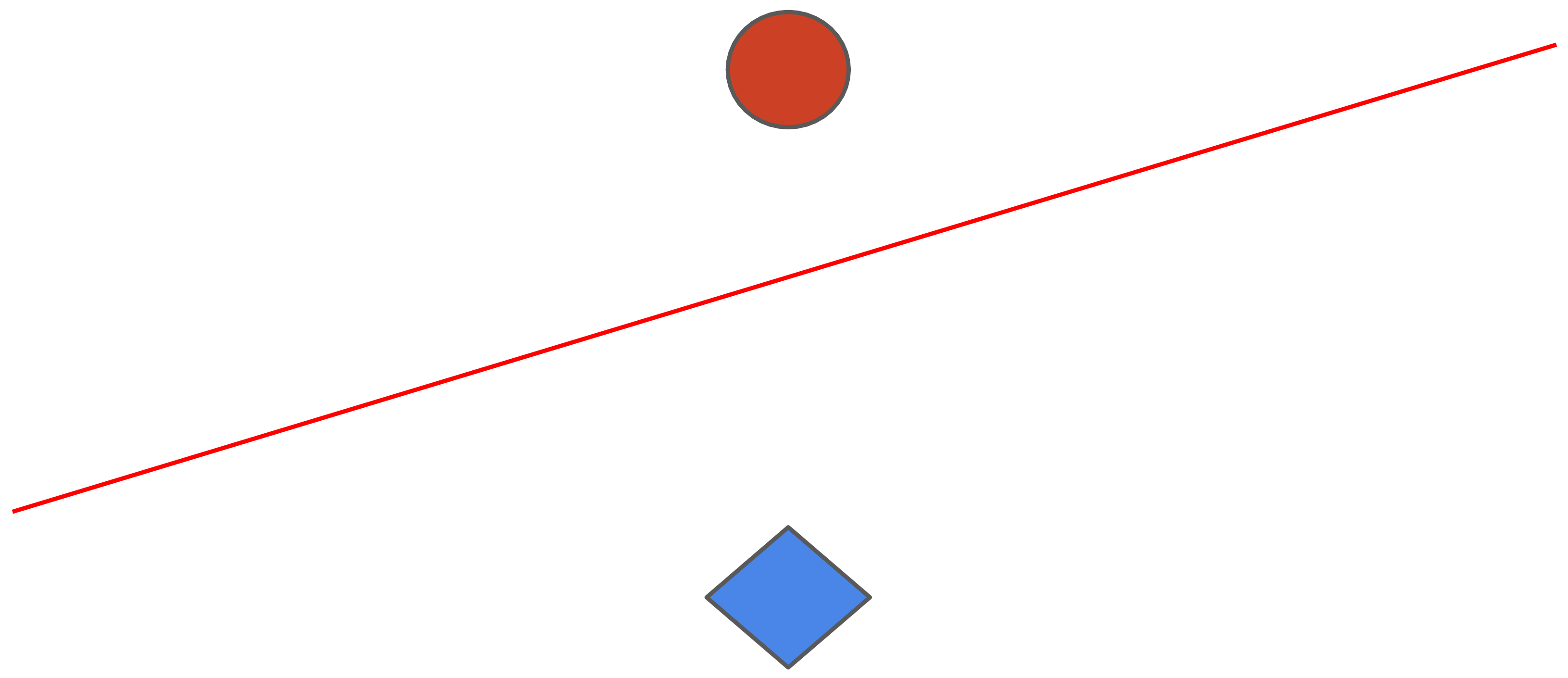
- Given a small amount of ***labeled*** data $\mathcal{X}_L$
- Given (usually) large amount of ***unlabeled*** data $\mathcal{X}_U$
- Can $\mathcal{X}_U$ help us in getting a better model?

What is a good decision boundary for these points?

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Semi-supervised Learning

- Given a small amount of ***labeled*** data $\mathcal{X}_L$
- Given (usually) large amount of ***unlabeled*** data $\mathcal{X}_U$
- Can $\mathcal{X}_U$ help us in getting a better model?



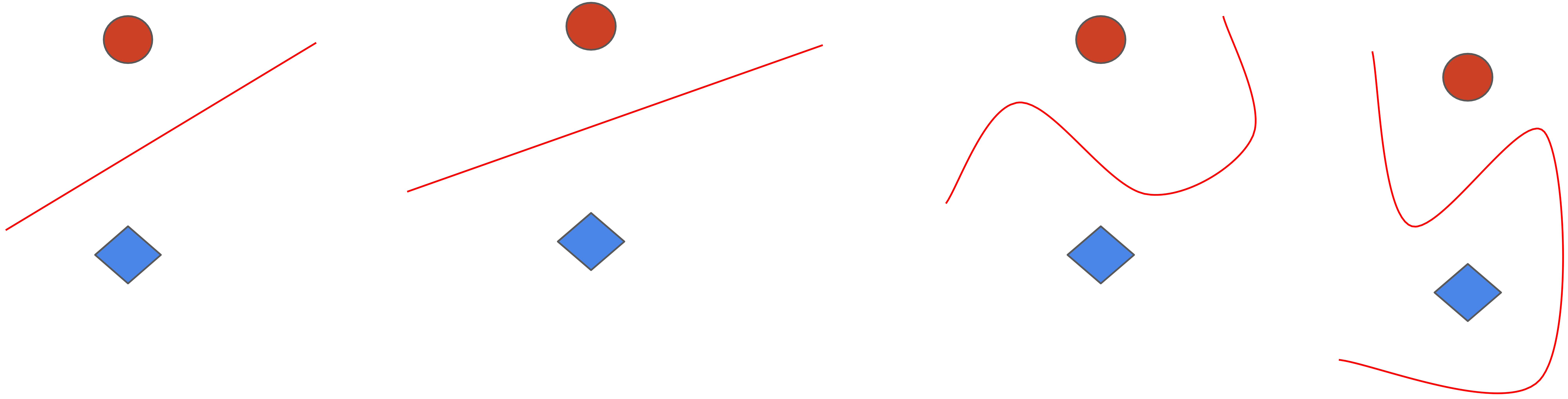
What is a good decision boundary for these points?

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Semi-supervised Learning

- Given a small amount of ***labeled*** data $\mathcal{X}_L$
- Given (usually) large amount of ***unlabeled*** data $\mathcal{X}_U$
- Can $\mathcal{X}_U$ help us in getting a better model?

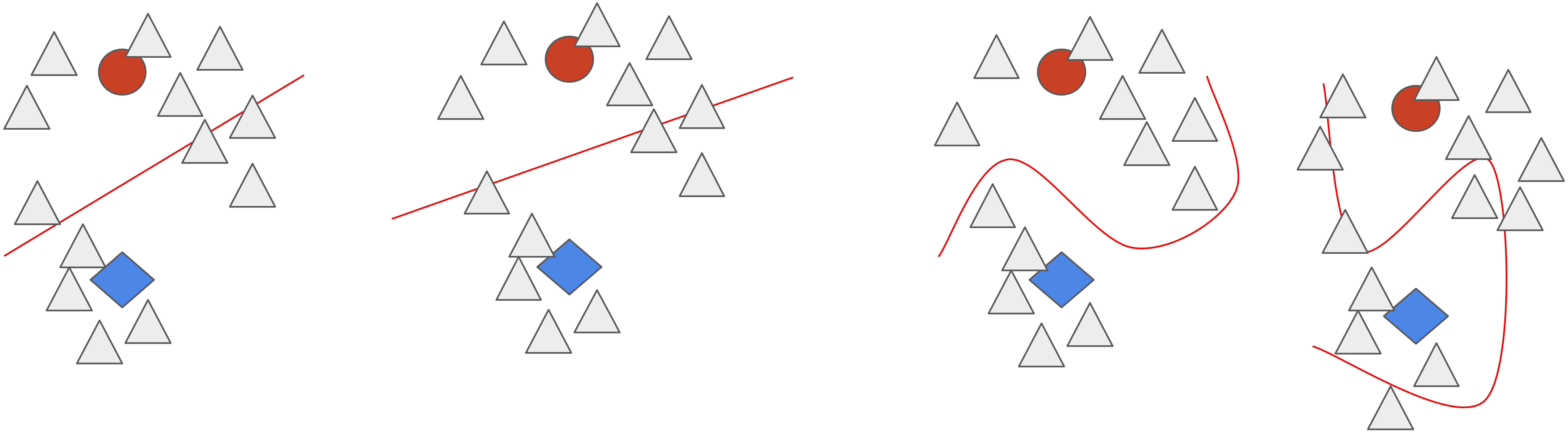
Which one is your favourite?



Semi-supervised Learning

- Given a small amount of ***labeled*** data $\mathcal{X}_L$
- Given (usually) large amount of ***unlabeled*** data $\mathcal{X}_U$
- Can $\mathcal{X}_U$ help us in getting a better model?

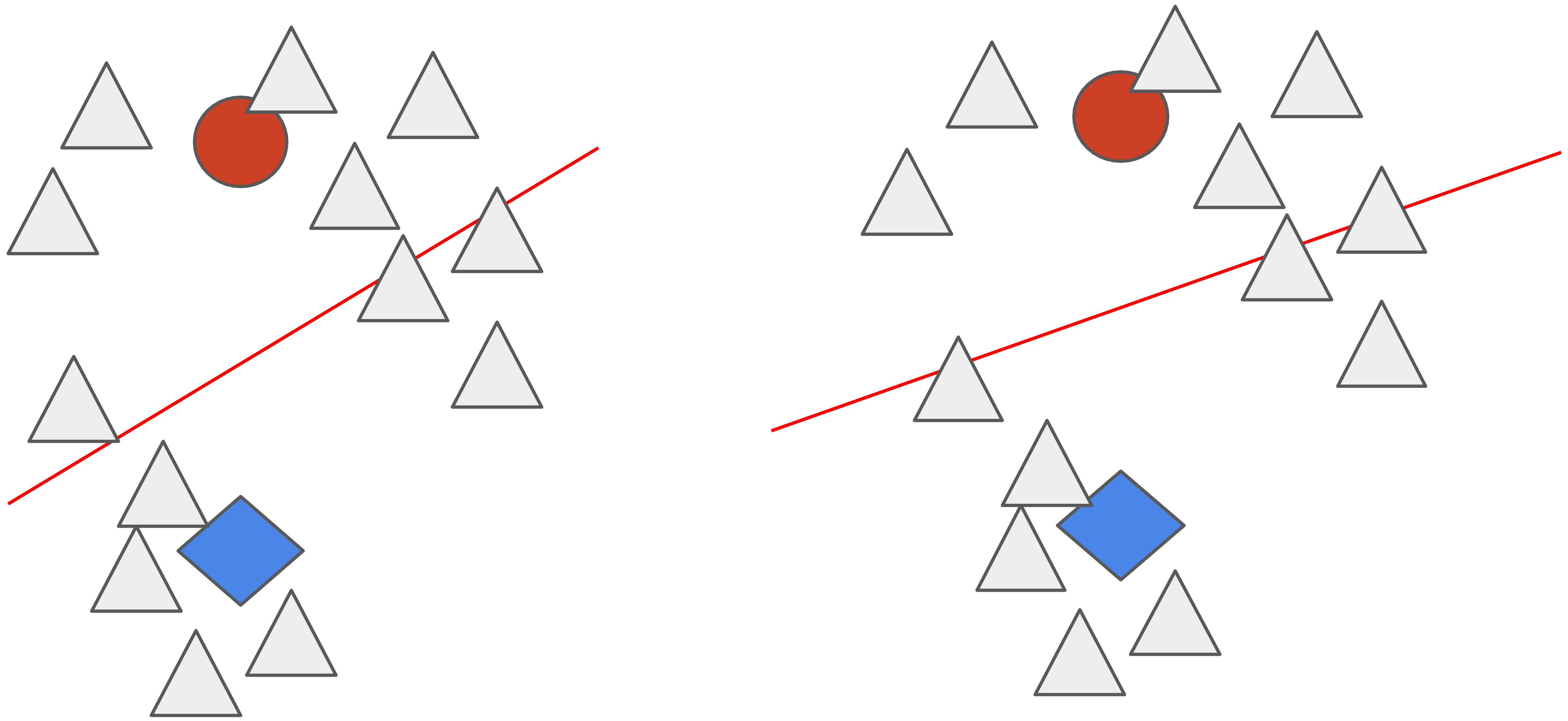
Now we see some unlabeled data points ….

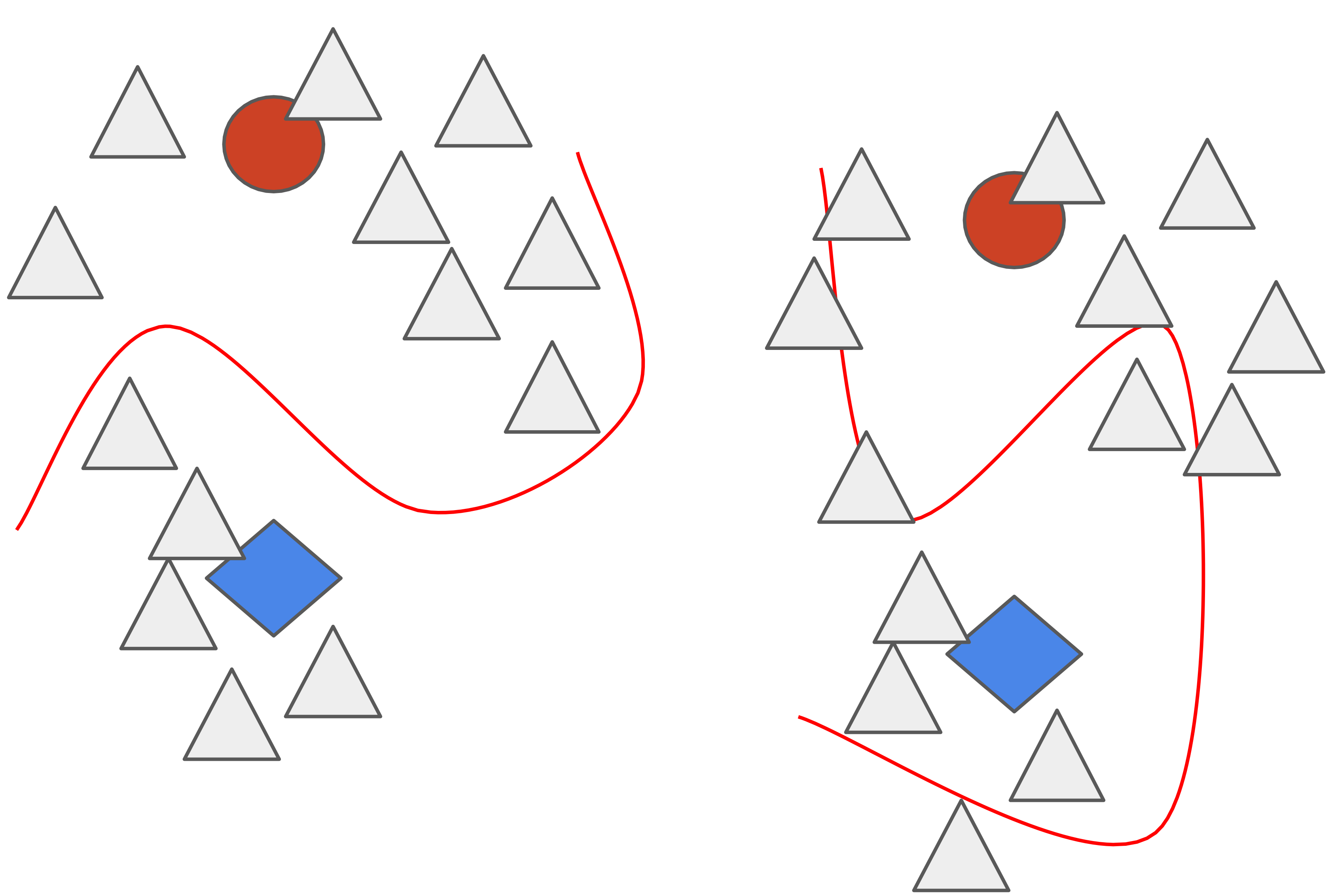


Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Semi-supervised Learning - intuitions

- Unlabeled samples tell us about P(**X**), which is useful in the predictive posterior P(y | **X**)

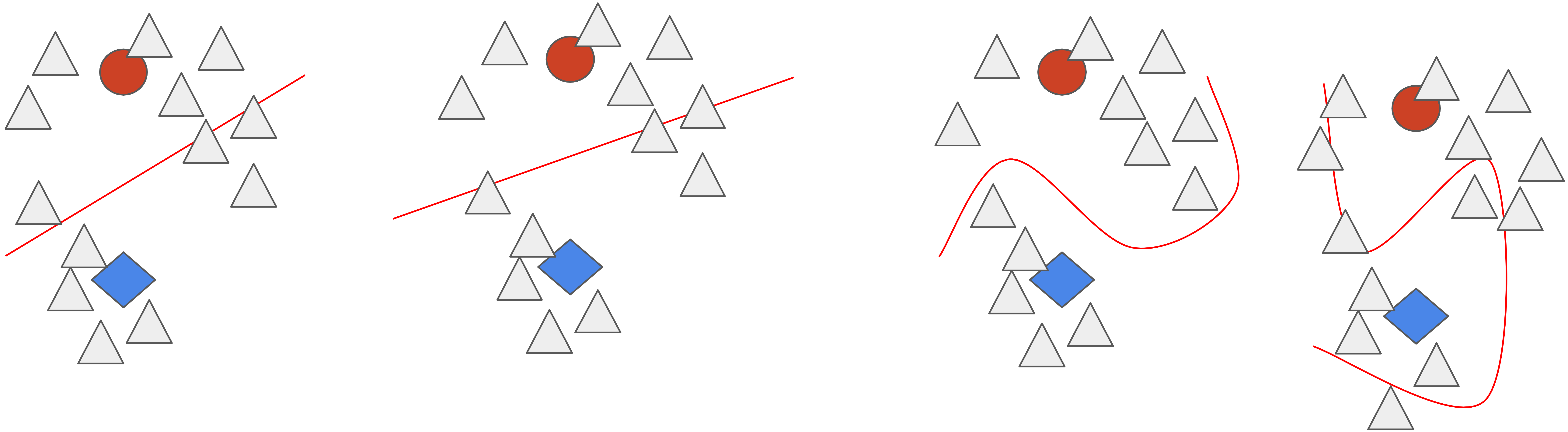




Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Semi-supervised Learning - definitions

- **Smoothness assumption:** if x_1, x_2 are close, labels y_1, y_2 are also “close”
- **Low-density separation:** x_1, x_2 are separated by *low-density region then* labels are not “close”
- **Cluster assumption:** points in same cluster likely to have same label



Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Semi-supervised Learning Approaches

- We will look at *a simple approach* to semi-supervised learning
- **Self-training** or **pseudo-labeling**
  - Age-old method
  - Surprisingly good with modern deep learning methods
  - But many variations …

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Self-training

- Assume: one's own high confidence predictions are correct!

- Train model f on $\mathcal{X}_L := \{x_L, y_L\}$
- Use f to predict "pseudo-labels" on $\mathcal{X}_U := \{x_u\}$
- Add $\{x_u, f(x_u)\}$ to labeled data
- Repeat

Self-training - variations

- Assume: one's own high confidence predictions are correct!

- Train model f on $\mathcal{X}_L := \{x_L, y_L\}$
- Use f to predict "pseudo-labels" on $\mathcal{X}_U := \{x_u\}$
- Add $\{x_u, f(x_u)\}$ to labeled data
- Repeat

1) Add only a few most confident predictions on Xu
2) Add all predictions on Xu
3) Add all predictions, weighted by the confidence of the prediction

….

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Self-training advantages

- The simplest semi-supervised method!
- It’s a “wrapper” - the classifiers or models can be arbitrarily complex, we do not need to delve into those details to apply self-training
- Often quite good in practice, e.g. in natural language tasks
- Also some vision tasks …

Data Distillation: Towards Omni-Supervised Learning

Ilija Radosavovic Piotr Dollár Ross Girshick Georgia Gkioxari Kaiming He
Facebook AI Research (FAIR)

Abstract

We investigate omni-supervised learning, *a special regime of semi-supervised learning in which the learner exploits all available labeled data plus internet-scale sources of unlabeled data. Omni-supervised learning is lower-bounded by performance on existing labeled datasets, offering the potential to surpass state-of-the-art fully supervised methods. To exploit the omni-supervised setting, we propose* data distillation, *a method that ensembles predictions from multiple transformations of unlabeled data, using a single model, to automatically generate new training annotations. We argue that visual recognition models have recently become accurate enough that it is now possible to apply classic ideas about self-training to challenging real-world data. Our experimental results show that in the cases of human keypoint detection and general object detection, state-of-the-art models trained with data distillation surpass the performance of using labeled data from the COCO dataset alone.*

1. Introduction

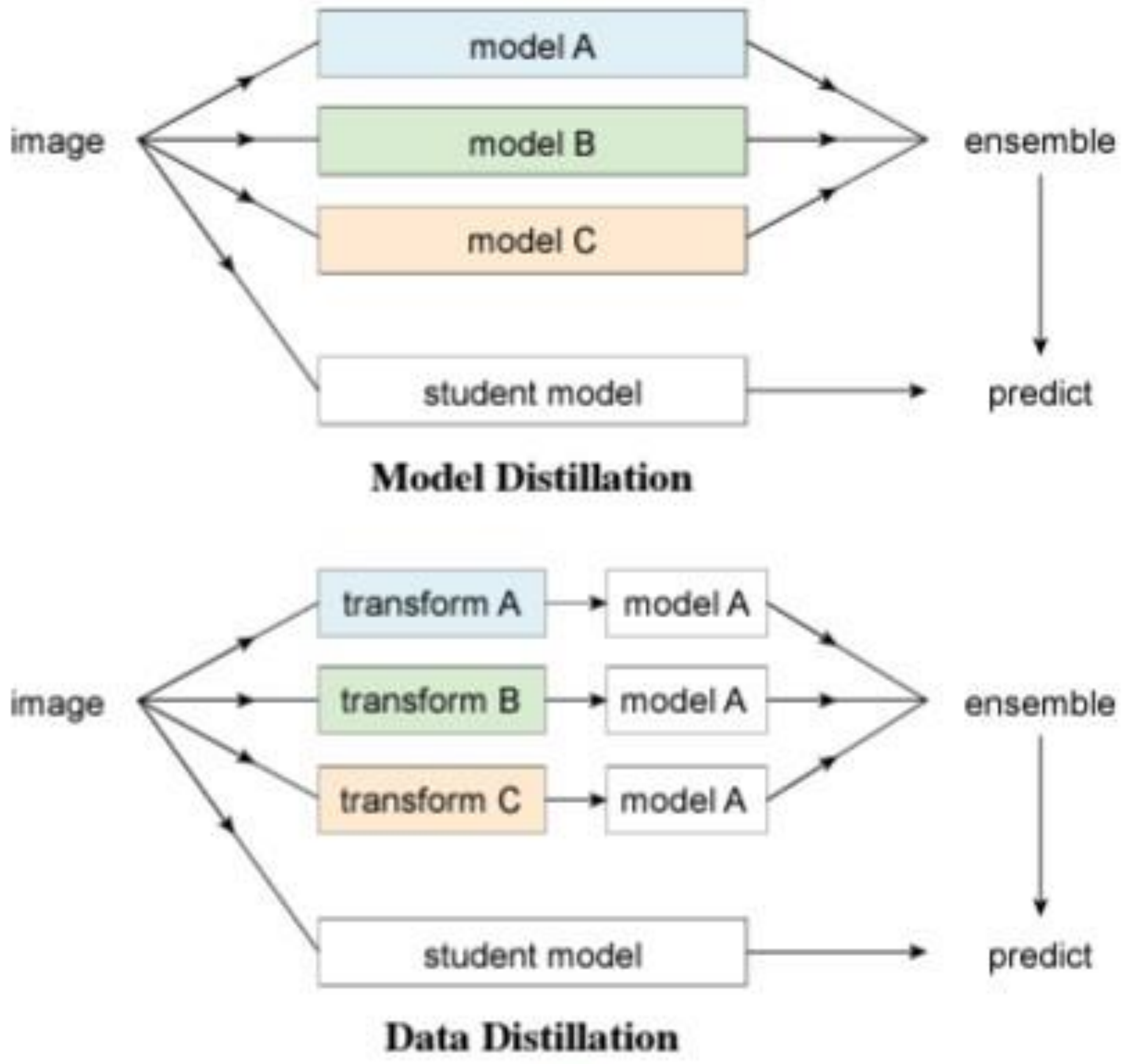


Figure 1. **Model Distillation [18] *vs*. Data Distillation.** In data distillation, ensembled predictions from a single model applied to multiple transformations of an unlabeled image are used as automatically annotated data for training a student model.

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Figure 2. **Ensembling keypoint predictions from multiple data transformations can yield a single superior (automatic) annotation.** For visualization purposes all images and keypoint predictions are transformed back to their original coordinate frame.

backbone	DD	AP	AP_{50}	AP_{75}	AP_M	AP_L
ResNet-50		65.1	86.6	70.9	59.9	73.6
ResNet-50	✓	**66.6**	**87.3**	**72.6**	**61.6**	**75.0**
ResNet-101		66.1	87.7	71.7	60.5	75.0
ResNet-101	✓	**67.5**	**87.9**	**73.9**	**62.4**	**75.9**
ResNeXt-101-32×4		66.8	87.5	73.0	61.6	75.2
ResNeXt-101-32×4	✓	**68.0**	**88.1**	**74.2**	**63.1**	**76.2**
ResNeXt-101-64×4		67.3	88.0	73.3	62.2	75.6
ResNeXt-101-64×4	✓	**68.5**	**88.8**	**74.9**	**63.7**	**76.5**

(c) **Large-scale, dissimilar-distribution data.** Data distillation (DD) is performed on `co-115` with labels and `s1m-180` without labels, comparing with the supervised counterparts trained on `co-115`.

Table 1. Data distillation for COCO keypoint detection. Keypoint AP is reported on COCO `val2017`.

Disadvantages of self-training?

Any guesses?

Disadvantages of self-training?

- Early mistakes can reinforce themselves
  - We have heuristic solutions, like discarding samples if the confidence of prediction falls below some threshold
- Convergence
  - Hard to say if these steps of self-train and repeat will converge

Domain shifts can have a large impact

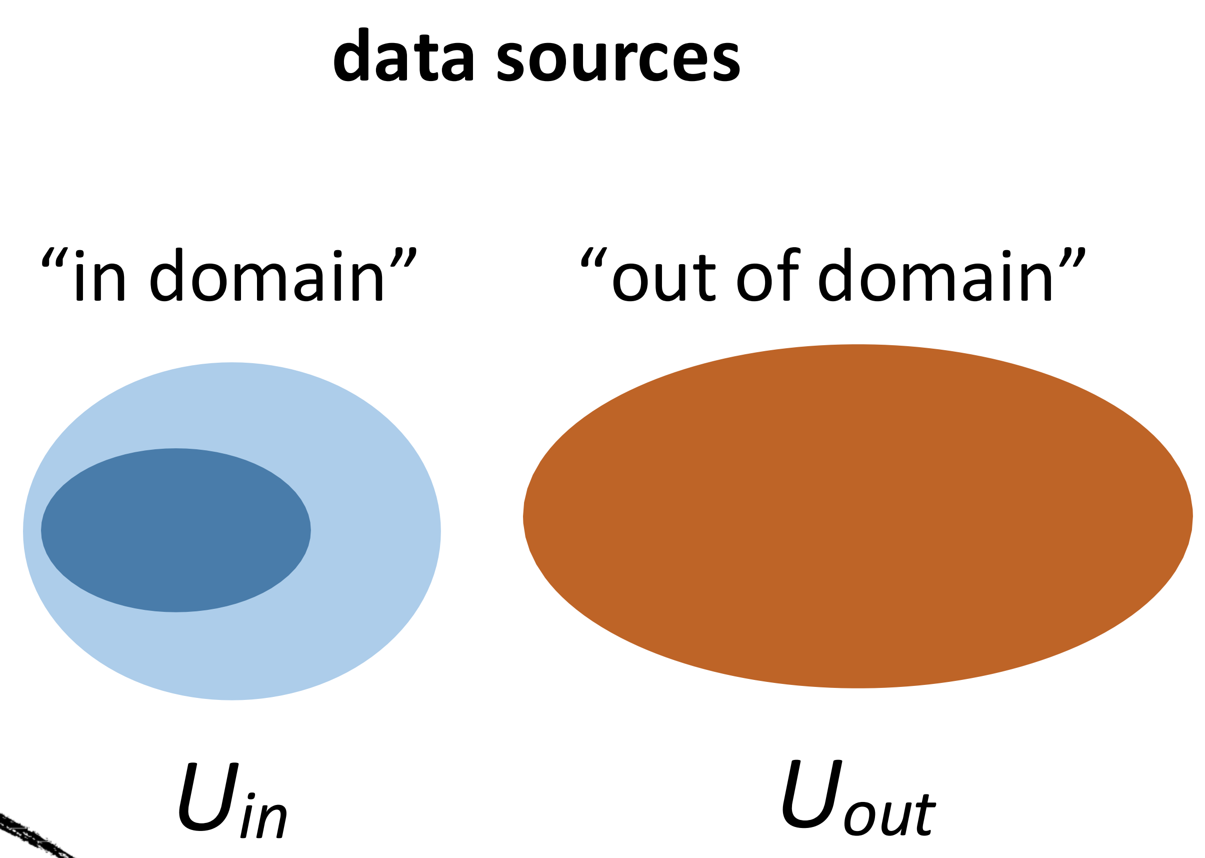
“Small” domain shifts can impact performance

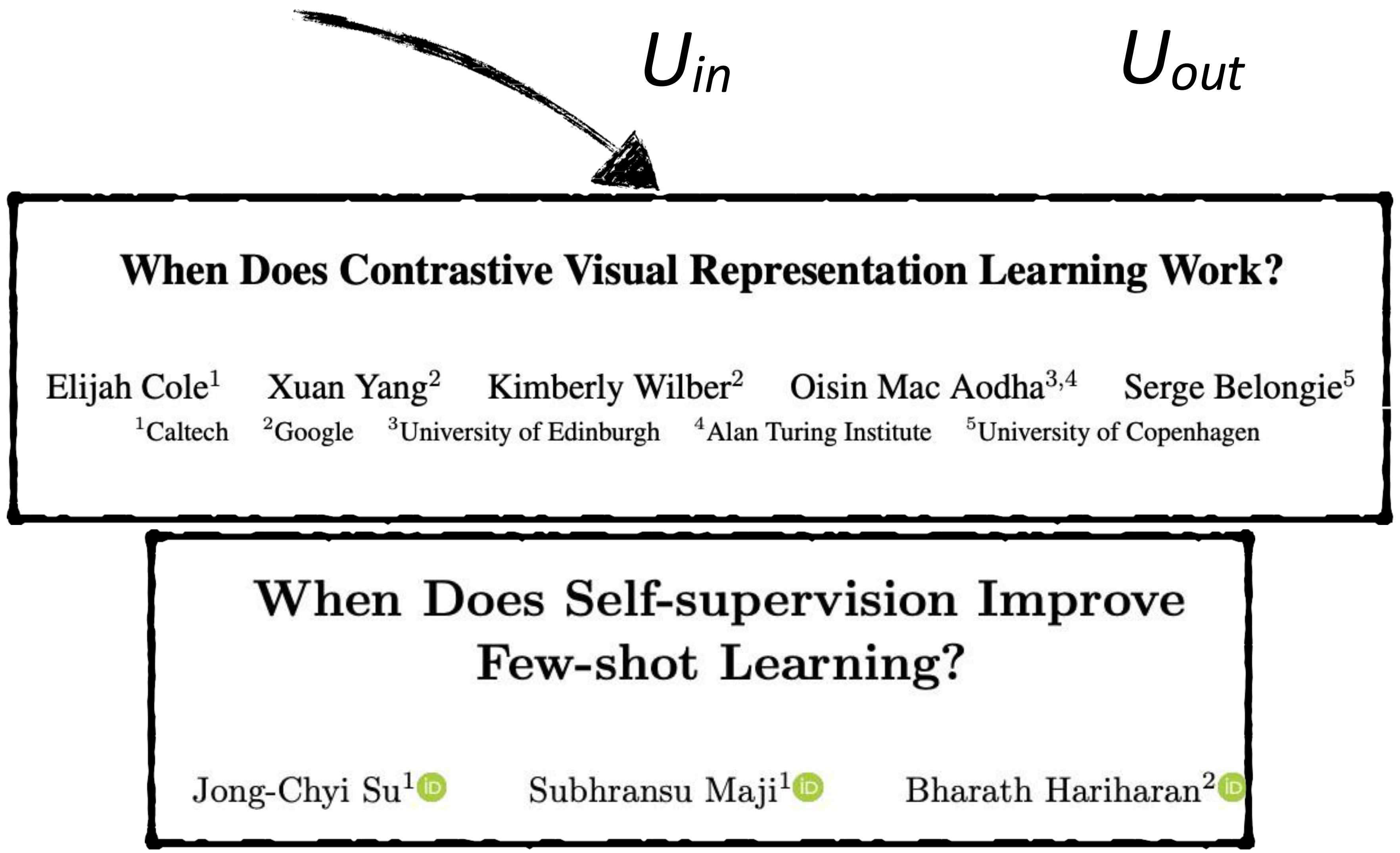
- resolution, size/pose/class, novel classes

Self/semi-supervised learning is brittle in fine-grained domains

- difficult task, long-tailed data

Need “guardrails” against biased data



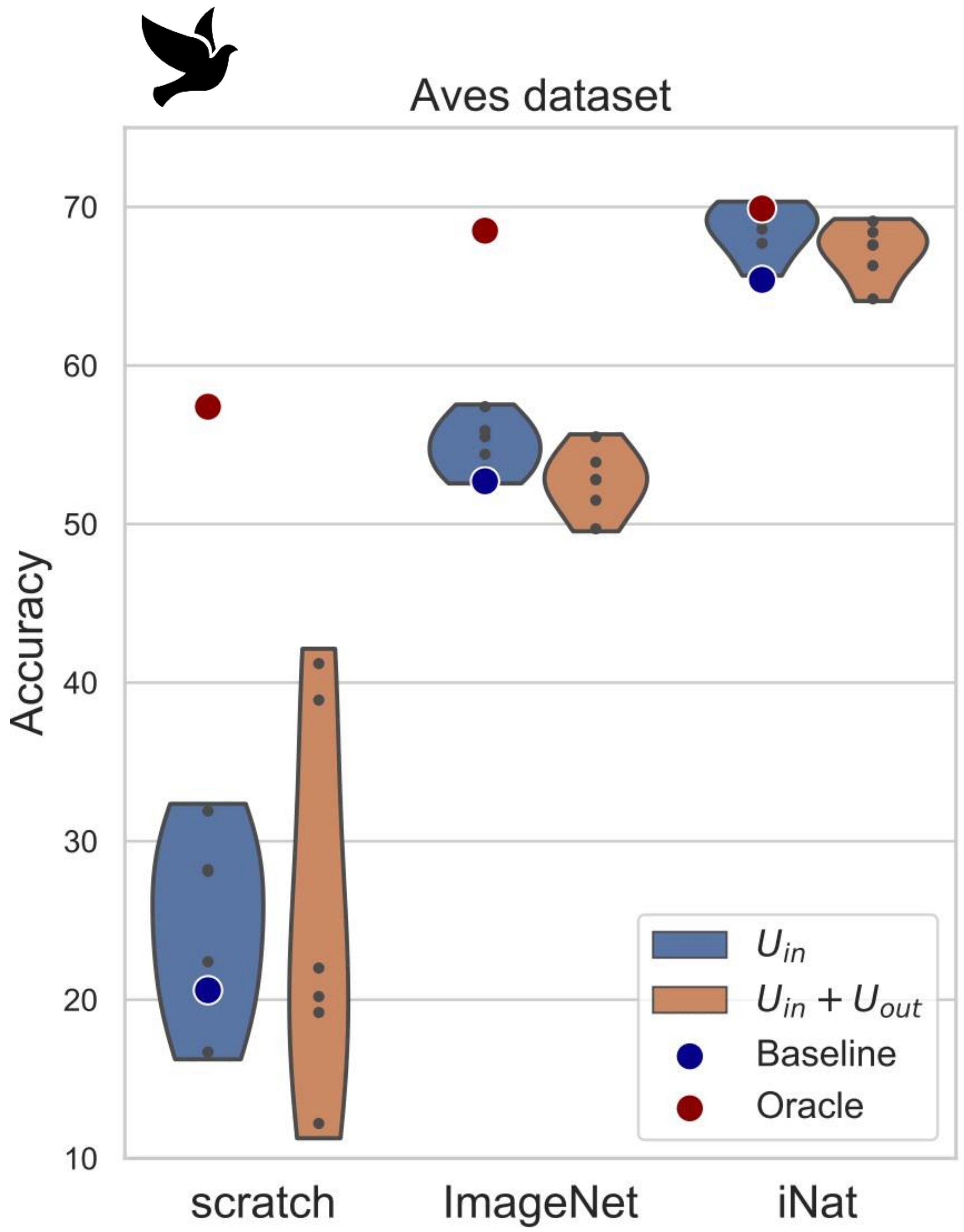

When Does Contrastive Visual Representation Learning Work?

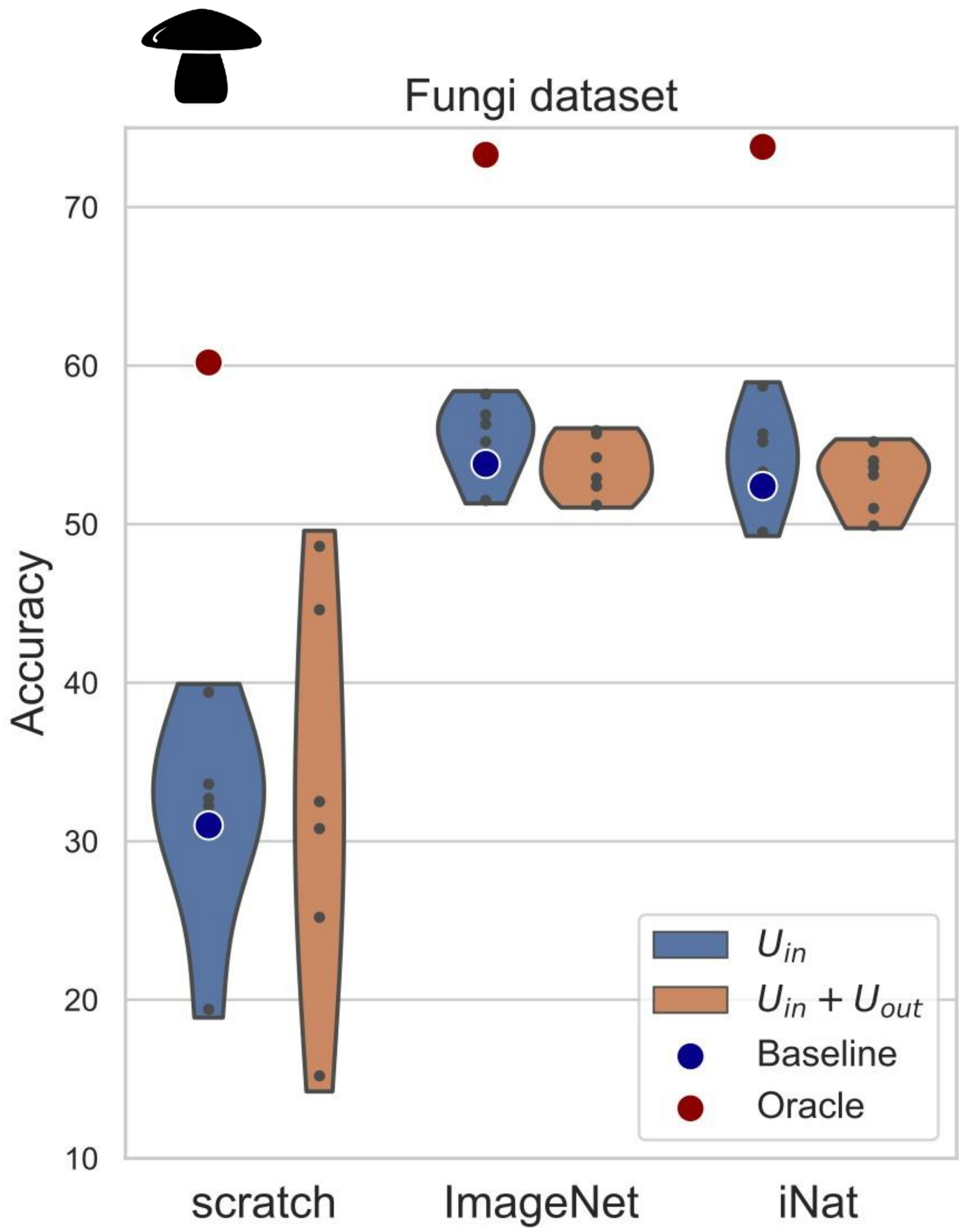
Elijah Cole[1] Xuan Yang[2] Kimberly Wilber[2] Oisin Mac Aodha[3,4] Serge Belongie[5]
[1]Caltech [2]Google [3]University of Edinburgh [4]Alan Turing Institute [5]University of Copenhagen

When Does Self-supervision Improve Few-shot Learning?

Jong-Chyi Su[1] Subhransu Maji[1] Bharath Hariharan[2]

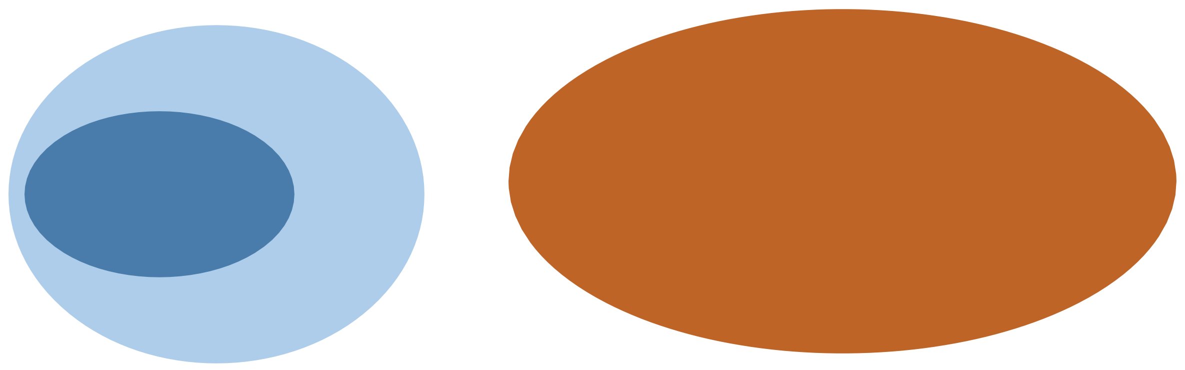
How robust is semi-supervised learning?



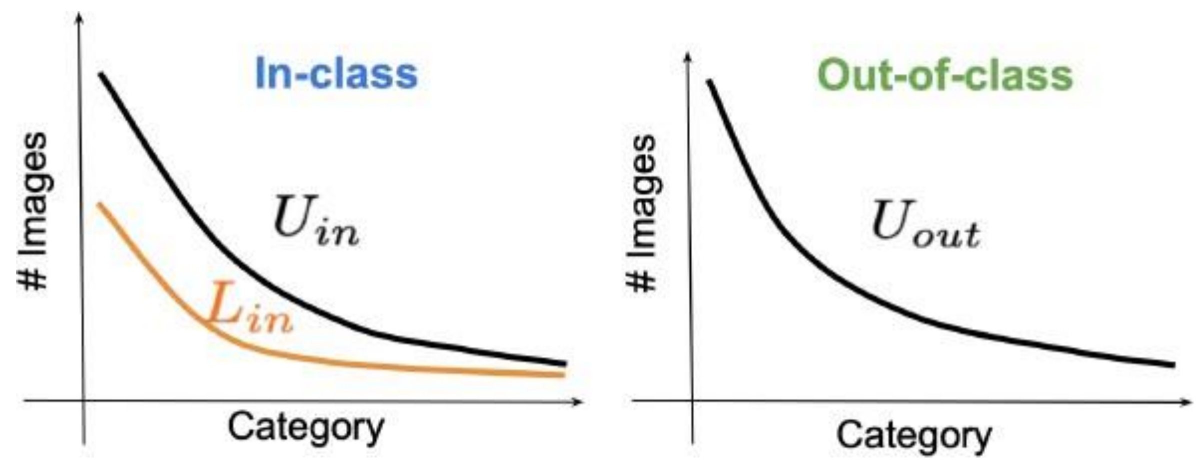


data sources

"in domain" "out of domain"



U_{in} U_{out}



Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

More pointers on semi-supervised learning

- Vast literature both in terms of theory and applications

- Other methods:
  - **Entropy minimization:** adds a loss that encourages the neural network model to make high confidence predictions (minimize “entropy”) on all unlabeled samples
  - Mean Teacher, FixMatch, NoisyStudent, …
  - Combine with methods to detect “out of domain” data

Today’s Class

- Recap
  - Supervised vs Unsupervised Learning
  - Why not always label data?
- Semi-supervised Learning
  - Concepts
  - Example: pseudo-labels / self-training
- Self-supervised Learning
  - Concepts
  - Pretext tasks
  - Contrastive Learning

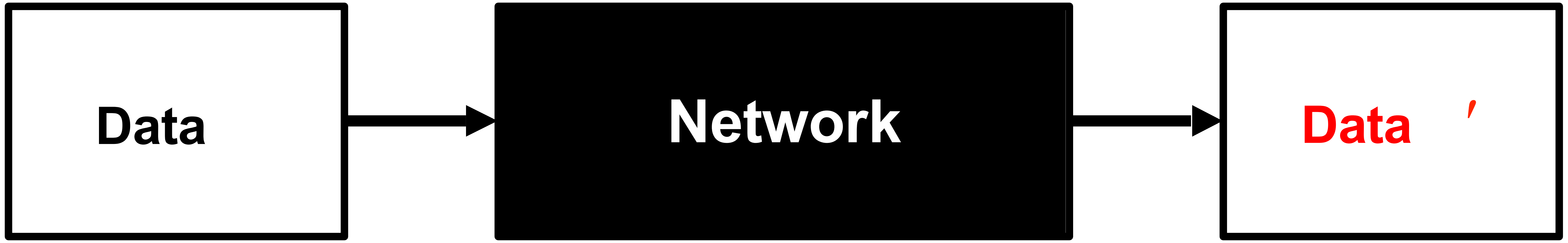
Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
  - “Siamese” methods
    - Contrastive methods
    - Non-contrastive methods
- Self-supervision beyond still images
  - 3D, audio, video, language

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Self-supervision as data prediction



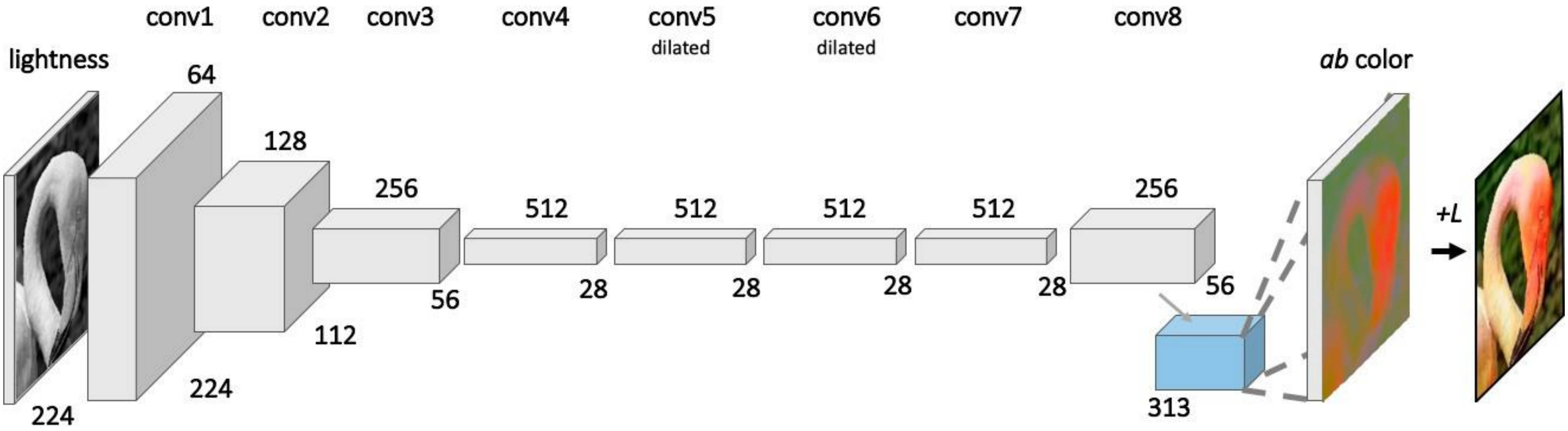
- Colorization
- Inpainting
- Future prediction
- …

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

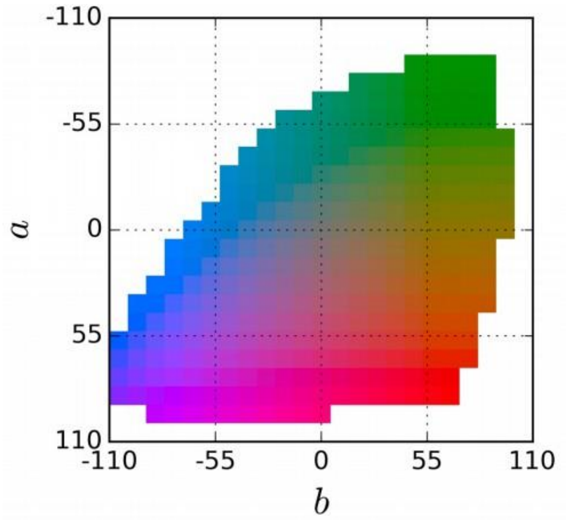
Colorization

R. Zhang, P. Isola, and A. Efros, [Colorful Image Colorization](), ECCV 2016

Colorization: Architecture



At each spatial location, predict probability distribution over 313 quantized (a,b) values

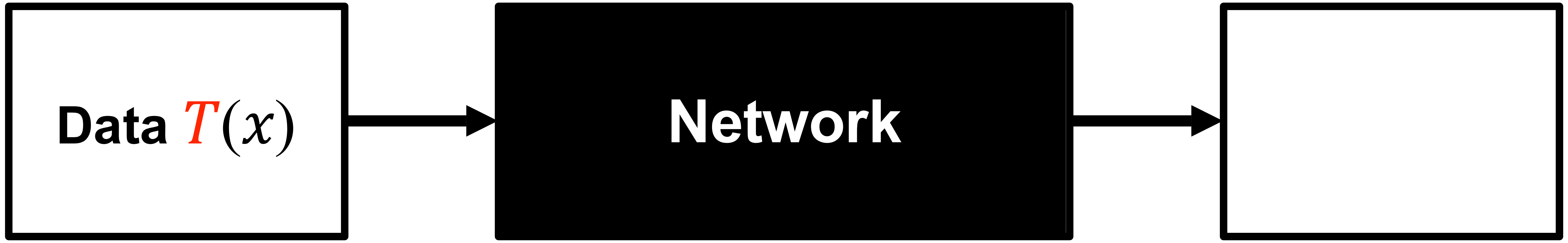


R. Zhang, P. Isola, and A. Efros, [Colorful Image Colorization](), ECCV 2016

Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction

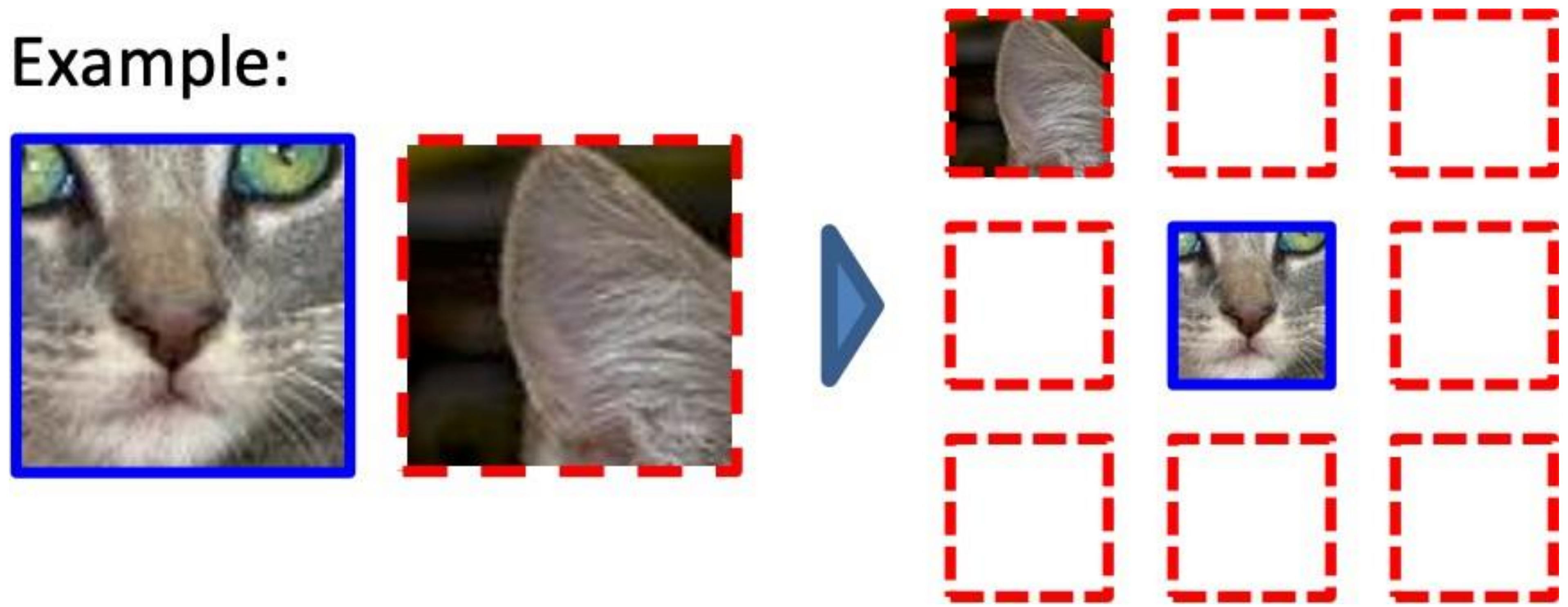
Self-supervision by transformation prediction



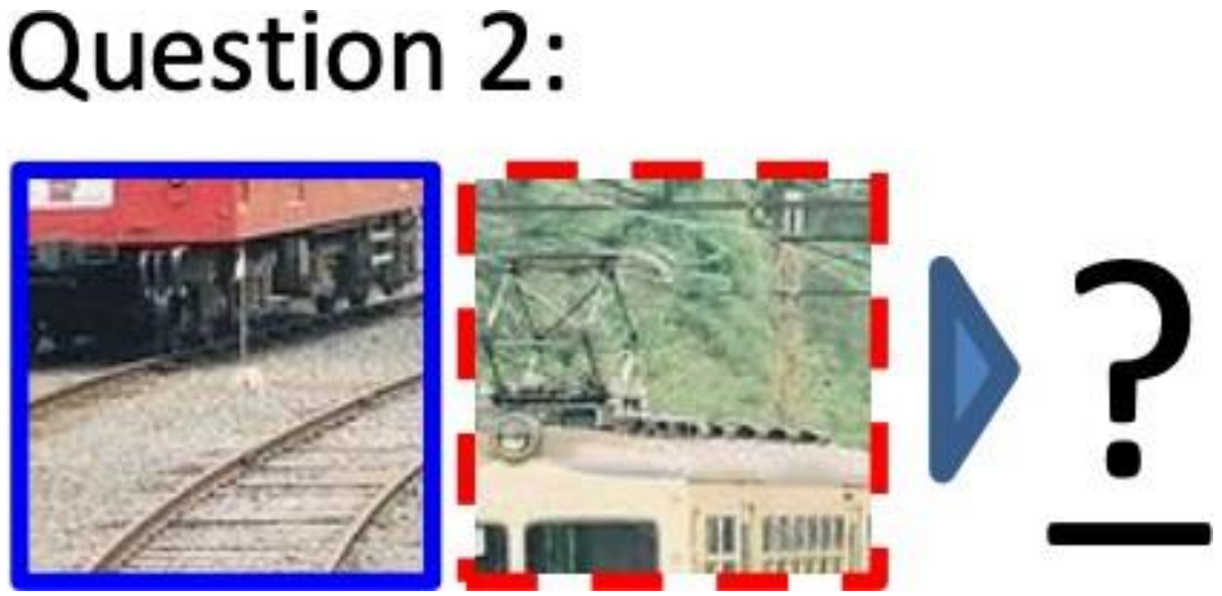
- Context prediction
- Jigsaw puzzle solving
- Rotation prediction

Context prediction

- *Pretext task*: randomly sample a patch and one of 8 neighbors
- Guess the spatial relationship between the patches



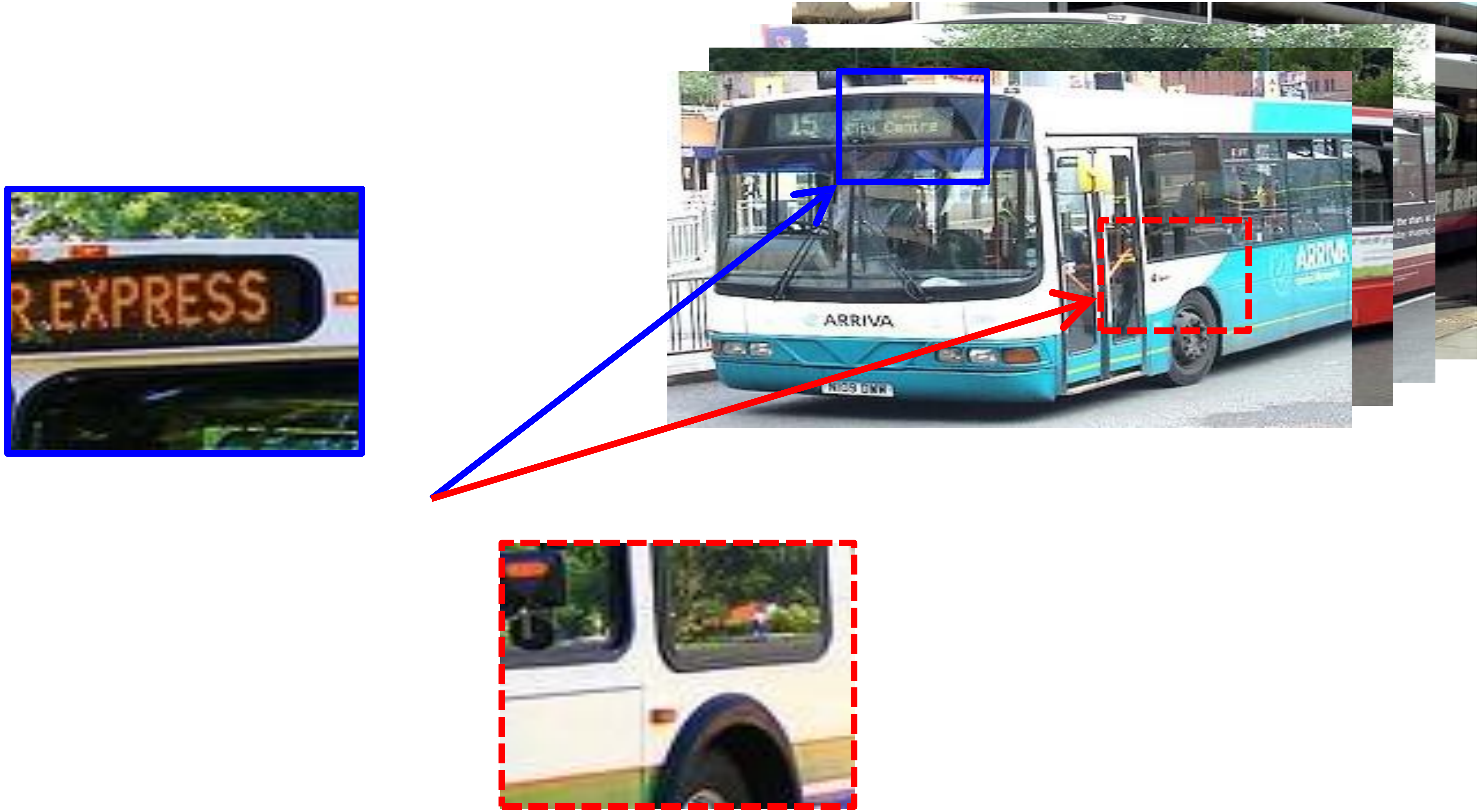
A: Bottom right



A: Top center

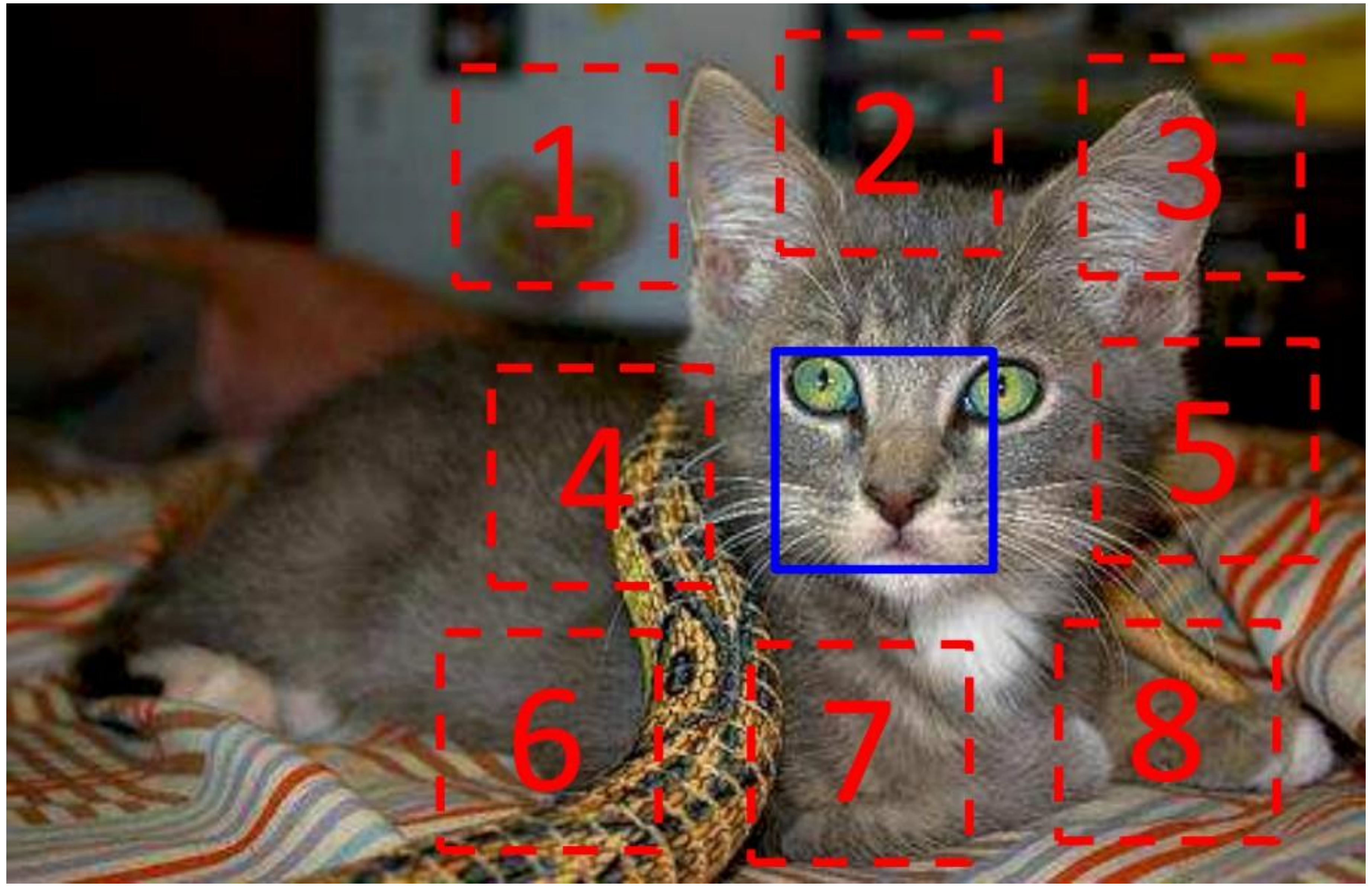
C. Doersch, A. Gupta, A. Efros. [Unsupervised Visual Representation Learning by Context Prediction](). ICCV 2015

Context prediction: Semantics from a non-semantic task



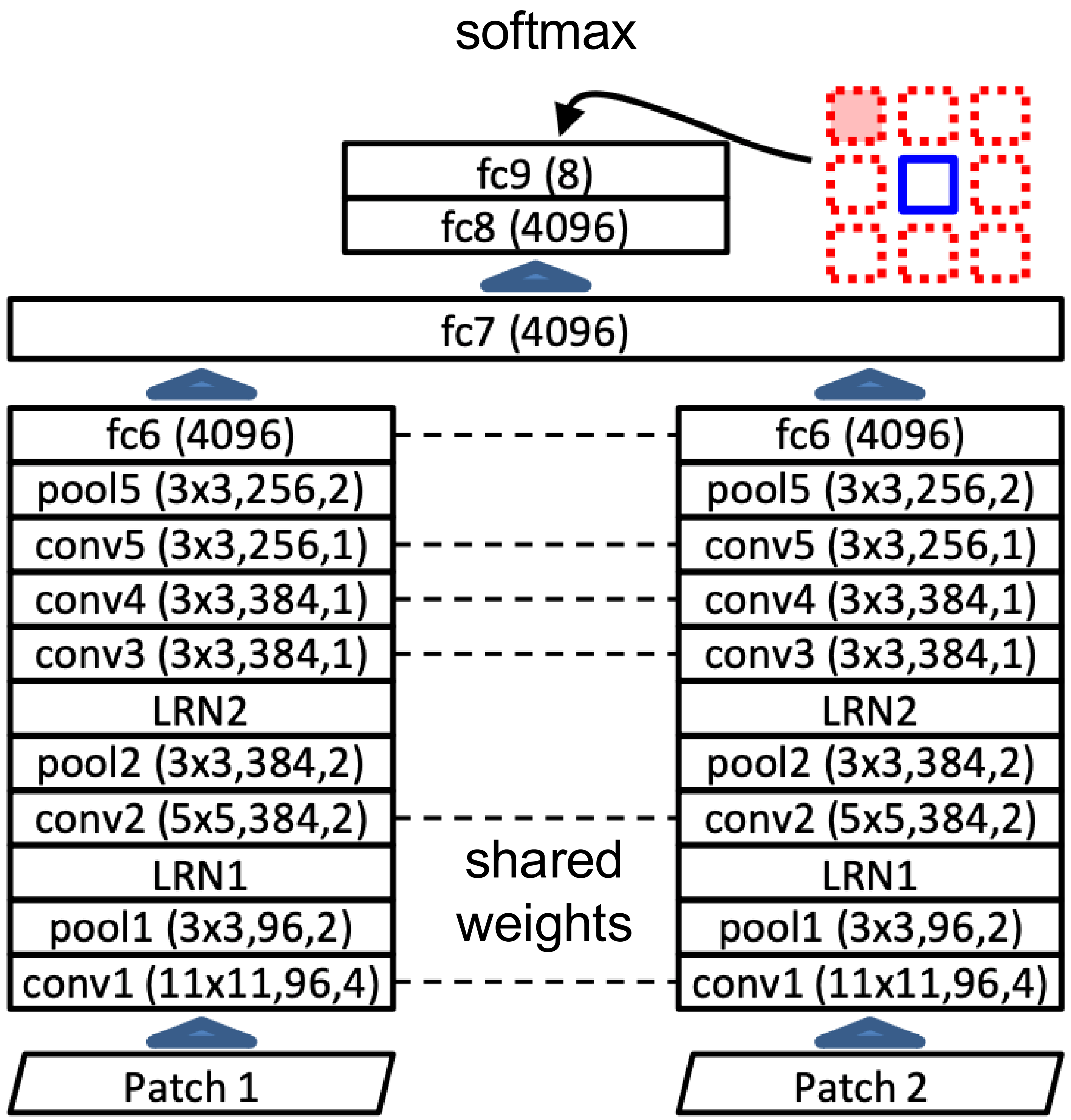
Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Context prediction: Details



Prevent “cheating”: sample patches with gaps, pre-process to overcome chromatic aberration

AlexNet-like architecture



C. Doersch, A. Gupta, A. Efros. [Unsupervised Visual Representation Learning by Context Prediction](). ICCV 2015

Jigsaw puzzle solving

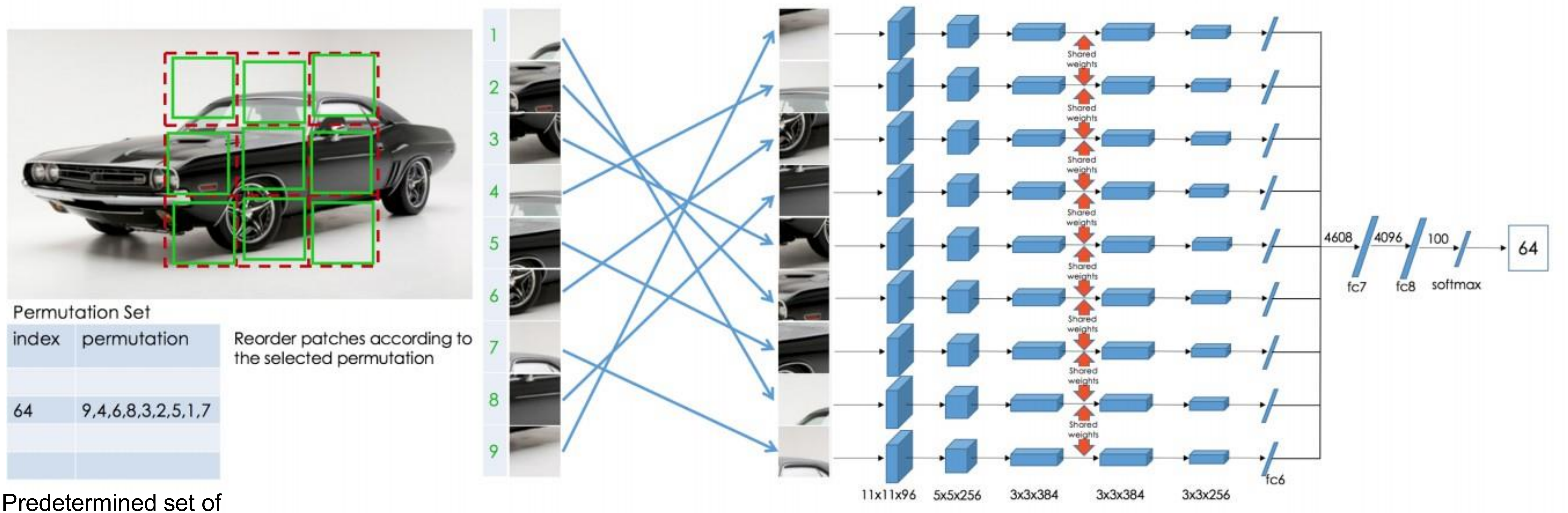
Crop out tiles | Shuffle | Pretext task: reassemble

Claim: jigsaw solving is easier than context prediction, trains faster, transfers better

M. Noroozi and P. Favaro. [Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles](). ECCV 2016

Jigsaw puzzle solving: Details

Context free network (CFN)



Predetermined set of 1000 permutations (out of 362,880 possible)

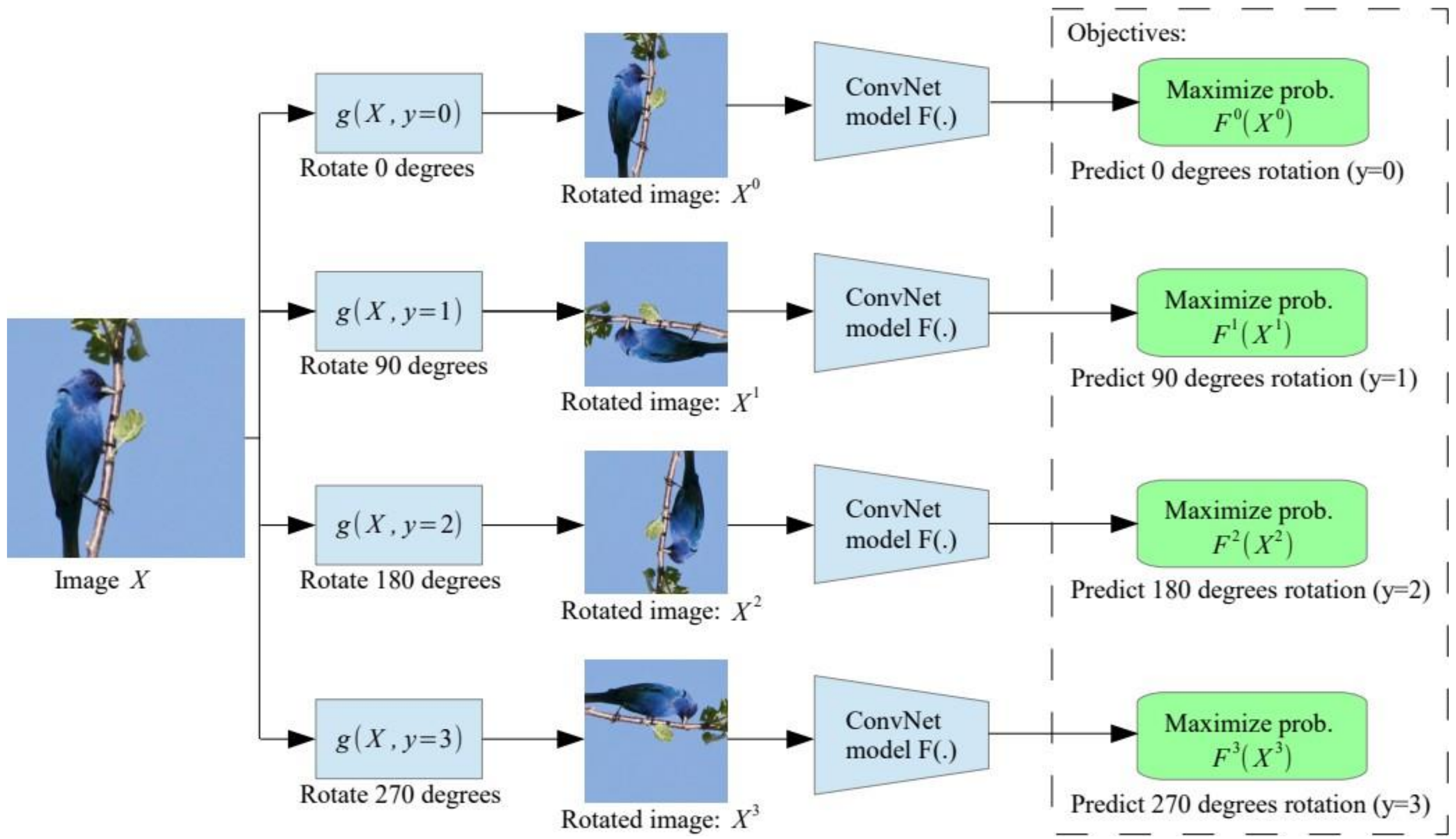
M. Noroozi and P. Favaro. **Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles**. ECCV 2016

Rotation prediction

- Pretext task: recognize image rotation (0, 90, 180, 270 degrees)

S. Gidaris, P. Singh, and N. Komodakis. **Unsupervised representation learning by predicting image rotations**. ICLR 2018

Rotation prediction



During training, feed in all four rotated versions of an image in the same mini-batch

S. Gidaris, P. Singh, and N. Komodakis. **Unsupervised representation learning by predicting image rotations**. ICLR 2018

PASCAL VOC Transfer Results

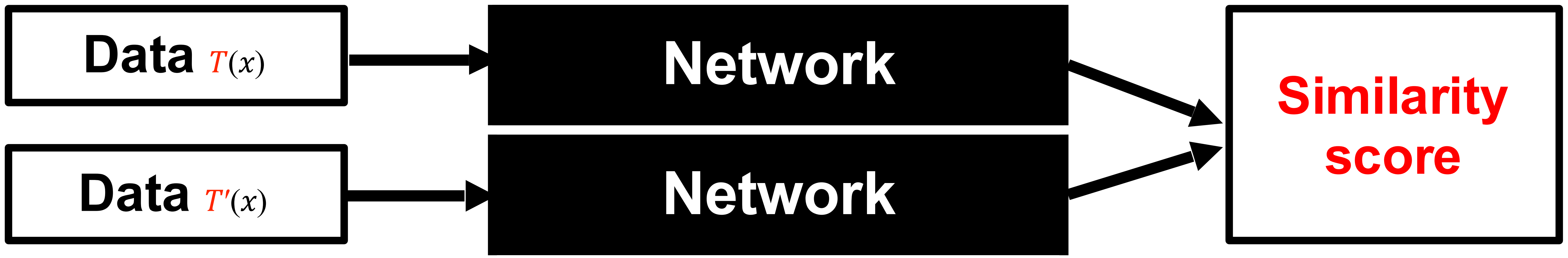
Method	Classification	Detection (mAP)	Segmentation (mIoU)
Supervised (ImageNet)	**79.9**	**56.8**	**48.0**
Colorization	65.6	46.9	35.6
Context	65.3	51.1	
Jigsaw	67.6	53.2	37.6
Rotation	73.0	54.4	39.1

Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- “Siamese” methods

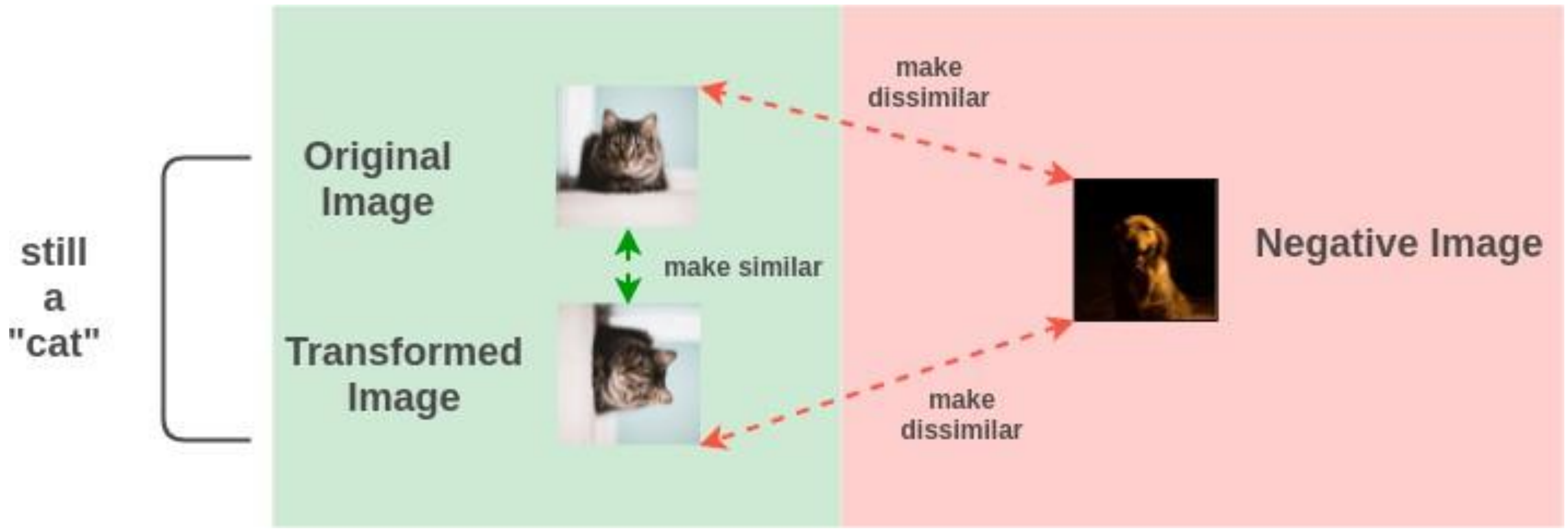
"Siamese" methods

- Extract representations from two transformed versions of a data point, encourage these representations to be similar (or to have other desirable properties)
  - **Contrastive methods**: train using both positive (similar) and negative (dissimilar) pairs
  - **Non-contrastive methods**: train with only positive examples



Contrastive methods

- Encourage representations of transformed versions of the same image to be the same and different images to be different



Contrastive loss formulation

- Given:
  - *Query* point
  - *Positive* sample x^+: version of subjected to a random transformation or augmentation (cropping, rotation, color change, etc.)
  - *Negative* samples

x^+

x^-

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Contrastive loss formulation

- Given: query , positive sample x^+, negative samples
- Measure similarity by dot product of L2-normalized feature representations:

$$\mathrm{sim}(x, y) = \frac{f(x)}{\|f(x)\|_2} \cdot \frac{f(y)}{\|f(y)\|_2}$$

- **Contrastive loss**: make similar to x^+, dissimilar from :

$$l(x, x^+) = -\log \frac{\exp(\mathrm{sim}(x, x^+)/\tau)}{\exp(\mathrm{sim}(x, x^+)/\tau) + \sum_{j=1}^{N} \exp(\mathrm{sim}(x, x_j^-)/\tau)}$$

- Intuitively, this is the loss of a softmax classifier that tries to classify as x^+

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Mechanisms for obtaining negative samples

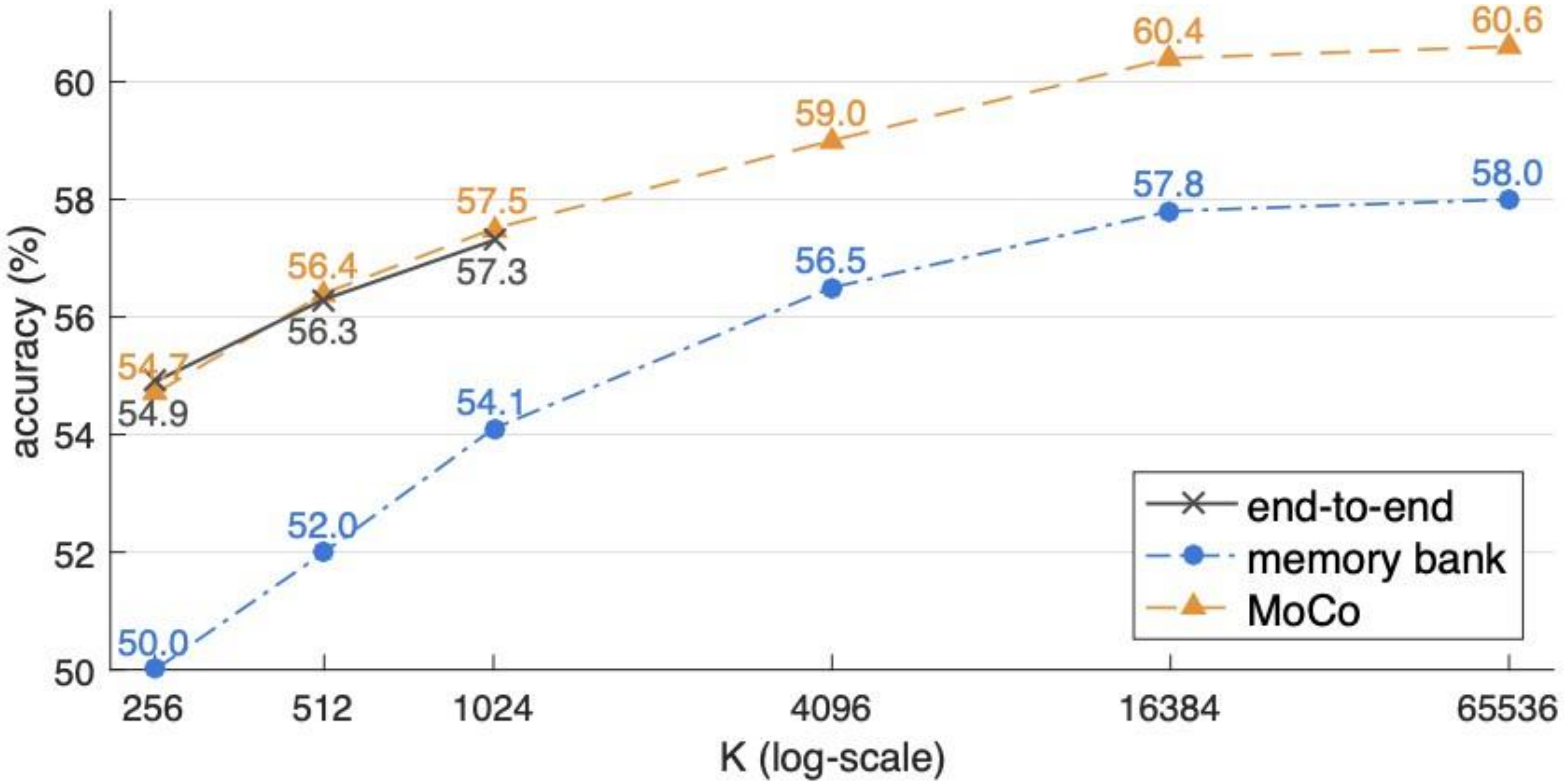
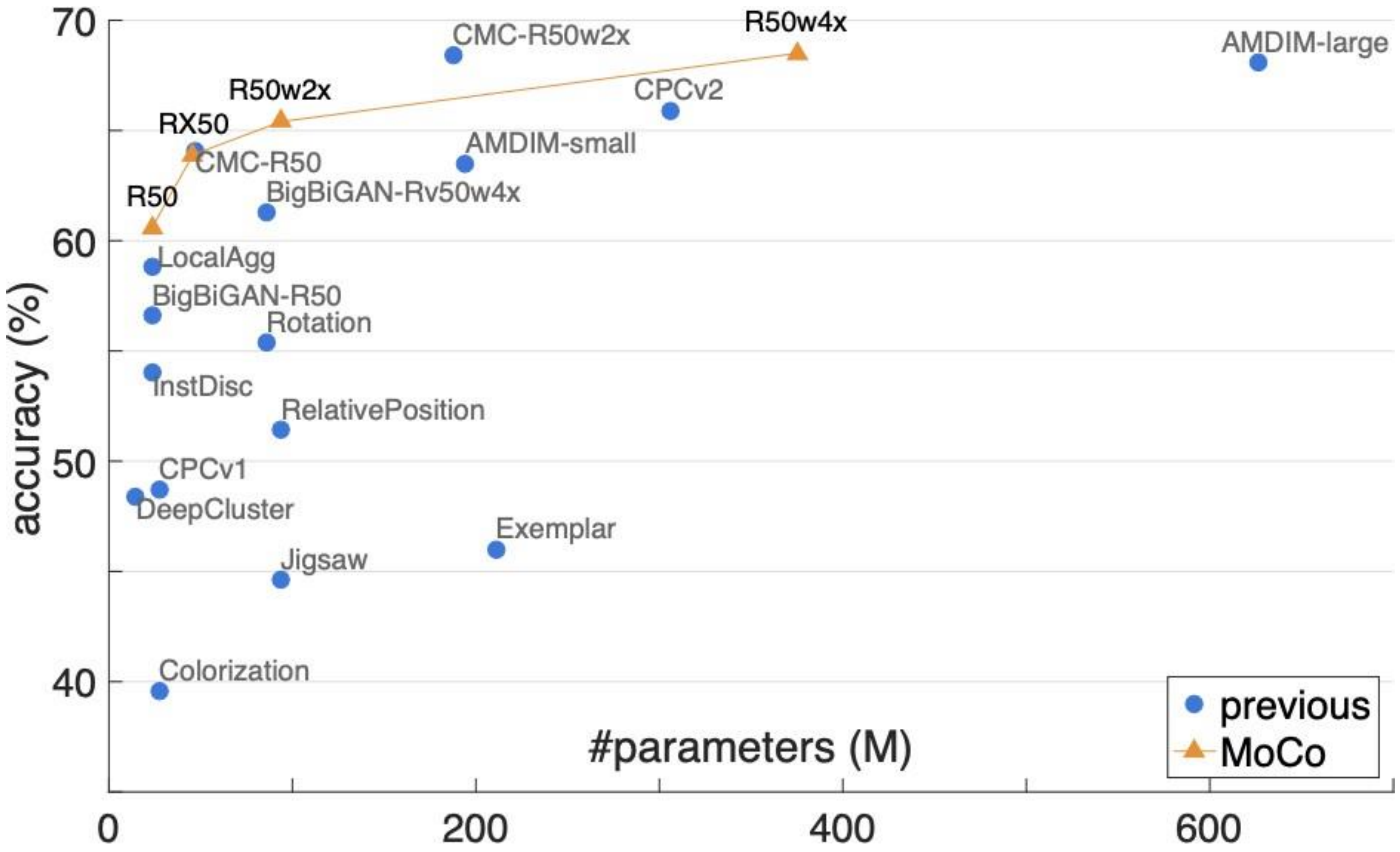


Figure 3. **Comparison of three contrastive loss mechanisms** under the ImageNet linear classification protocol. We adopt the same pretext task (Sec. 3.3) and only vary the contrastive loss mechanism (Figure 2). The number of negatives is K in memory bank and MoCo, and is $K-1$ in end-to-end (offset by one because the positive key is in the same mini-batch). The network is ResNet-50.

K. He et al. [Momentum Contrast for Unsupervised Visual Representation Learning](). CVPR 2020

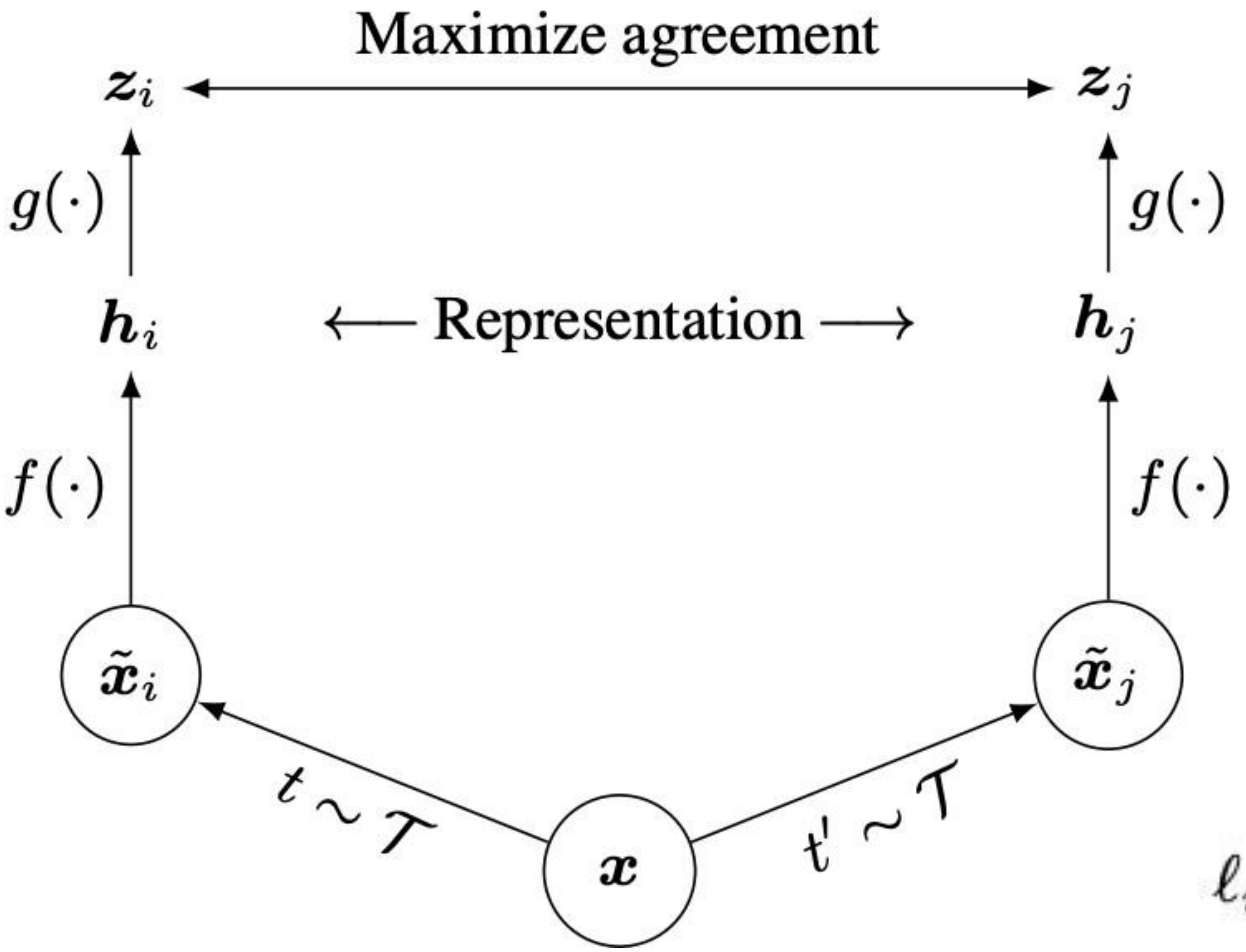
MoCo results

Comparison on linear ImageNet classification
(supervised accuracy above 75%)



K. He et al. **Momentum Contrast for Unsupervised Visual Representation Learning**. CVPR 2020

SimCLR



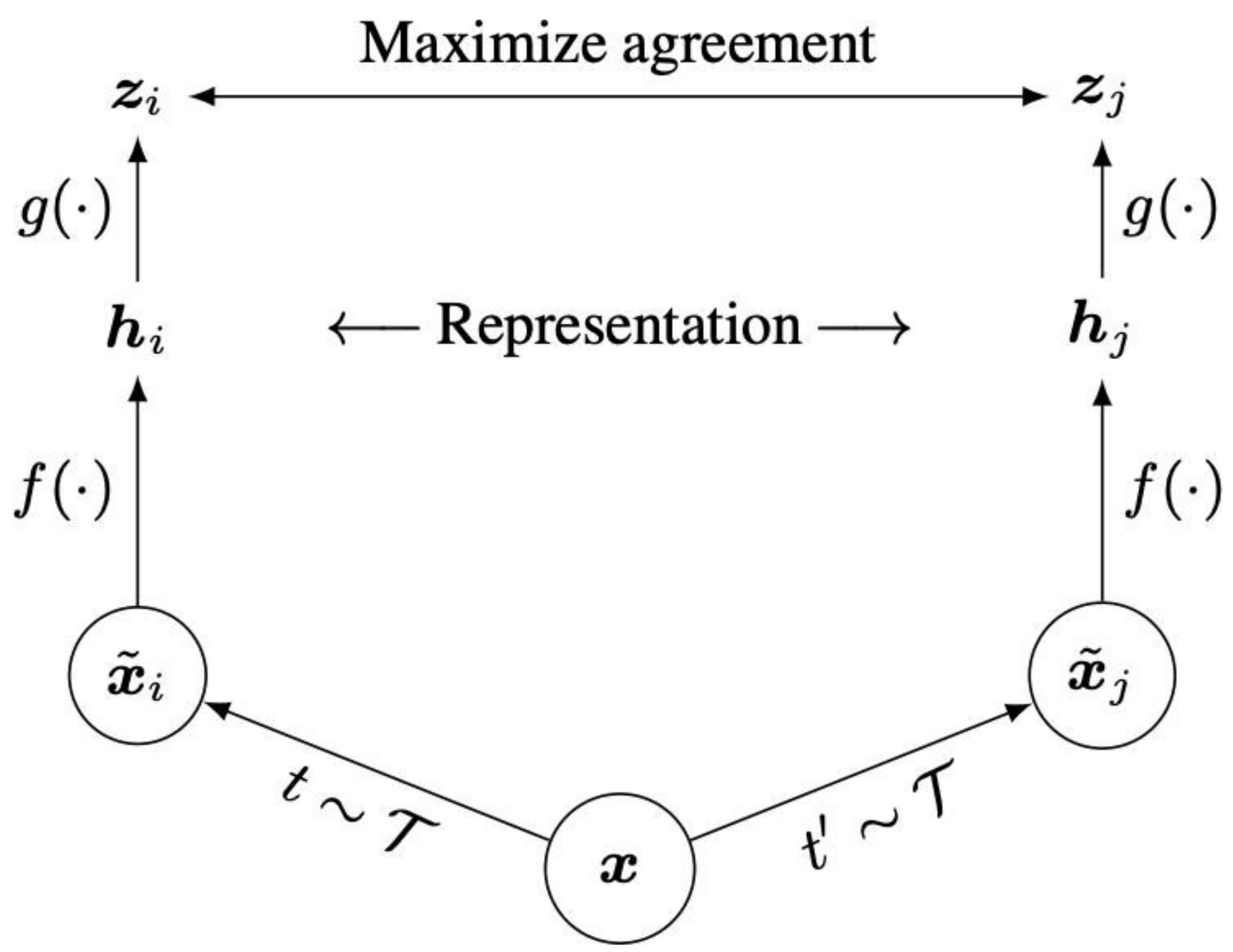
- Instead of memory bank or queue, use large mini-batch size (on cloud TPU)
- Introduce nonlinear *projection* ( ) between representation ( ) and feature used for computing contrastive loss ( )

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$$

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. [A Simple Framework for Contrastive Learning of Visual Representations](). ICML 2020

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

SimCLR



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

Algorithm 1 SimCLR's main learning algorithm.

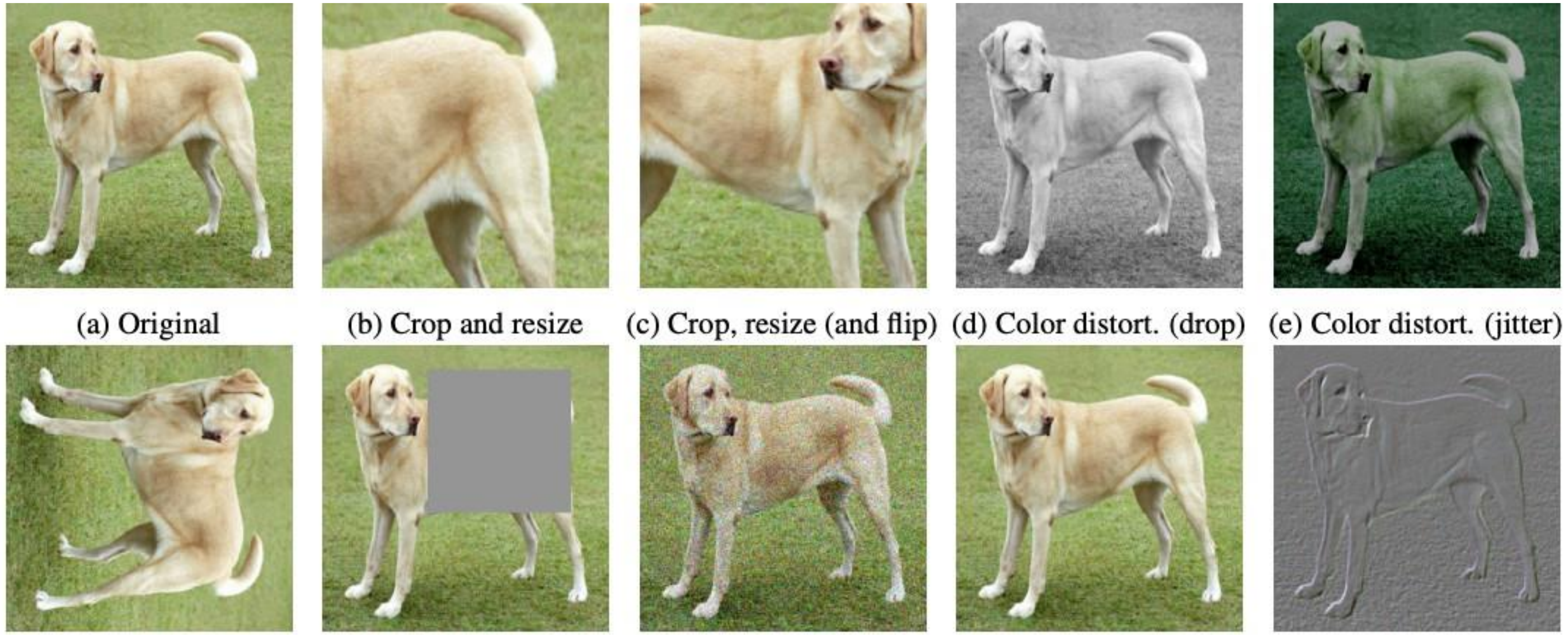
input: batch size N, constant τ, structure of f, g, $\mathcal{T}$.
for sampled minibatch $\{\boldsymbol{x}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    draw two augmentation functions $t \sim \mathcal{T}$, $t' \sim \mathcal{T}$
    # the first augmentation
    $\tilde{\boldsymbol{x}}_{2k-1} = t(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k-1} = f(\tilde{\boldsymbol{x}}_{2k-1})$ # representation
    $\boldsymbol{z}_{2k-1} = g(\boldsymbol{h}_{2k-1})$ # projection
    # the second augmentation
    $\tilde{\boldsymbol{x}}_{2k} = t'(\boldsymbol{x}_k)$
    $\boldsymbol{h}_{2k} = f(\tilde{\boldsymbol{x}}_{2k})$ # representation
    $\boldsymbol{z}_{2k} = g(\boldsymbol{h}_{2k})$ # projection
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^\top \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i,j)$ **as** $\ell(i,j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update networks f and g to minimize $\mathcal{L}$
end for
return encoder network $f(\cdot)$, and throw away $g(\cdot)$

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

SimCLR

- Performed extensive ablation study of data augmentations
- Found that composing multiple augmentations gives the best results



(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

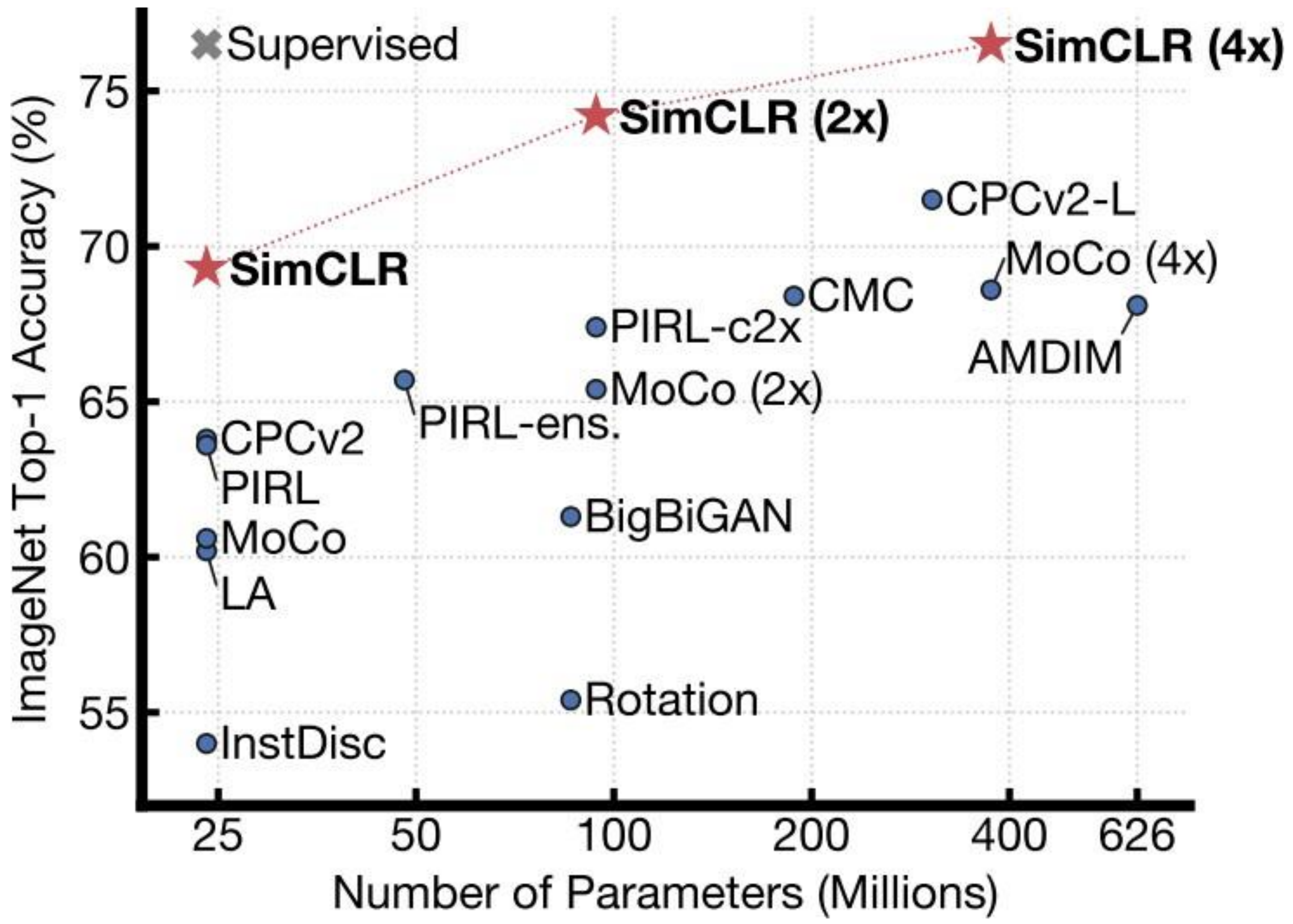
Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

Figure 5. Linear evaluation (ImageNet top-1 accuracy) under individual or composition of data augmentations, applied only to one branch. For all columns but the last, diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations (applied sequentially). The last column reflects the average over the row.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. [A Simple Framework for Contrastive Learning of Visual Representations](). ICML 2020

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

SimCLR: Evaluation



No detection evaluation

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. [A Simple Framework for Contrastive Learning of Visual Representations](). ICML 2020

Improved Baselines with Momentum Contrastive Learning

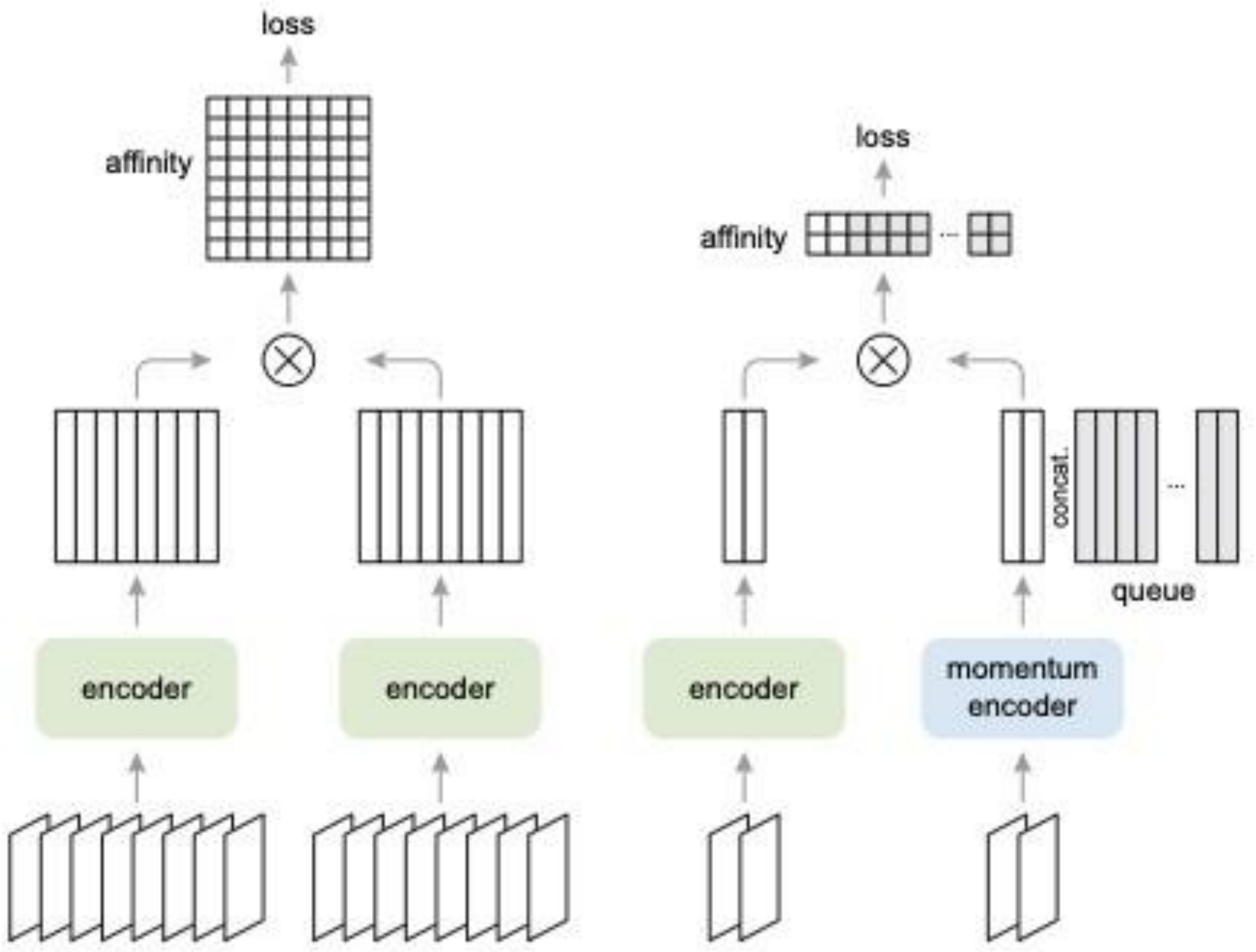
Xinlei Chen   Haoqi Fan   Ross Girshick   Kaiming He
Facebook AI Research (FAIR)

Abstract

Contrastive unsupervised learning has recently shown encouraging progress, e.g., *in Momentum Contrast (MoCo) and SimCLR. In this note, we verify the effectiveness of two of SimCLR's design improvements by implementing them in the MoCo framework. With simple modifications to MoCo—namely, using an MLP projection head and more data augmentation—we establish stronger baselines that outperform SimCLR and do not require large training batches. We hope this will make state-of-the-art unsupervised learning research more accessible. Code will be made public.*

1. Introduction

Recent studies on unsupervised representation learning from images [16, 13, 8, 17, 1, 9, 15, 6, 12, 2] are converging on a central concept known as *contrastive learning* [5]. The



(a) end-to-end   (b) Momentum Contrast

Figure 1. A **batching** perspective of two optimization mechanisms for contrastive learning. Images are encoded into a representation space, in which pairwise affinities are computed.

Ideas from SimCLR improve MoCo too!

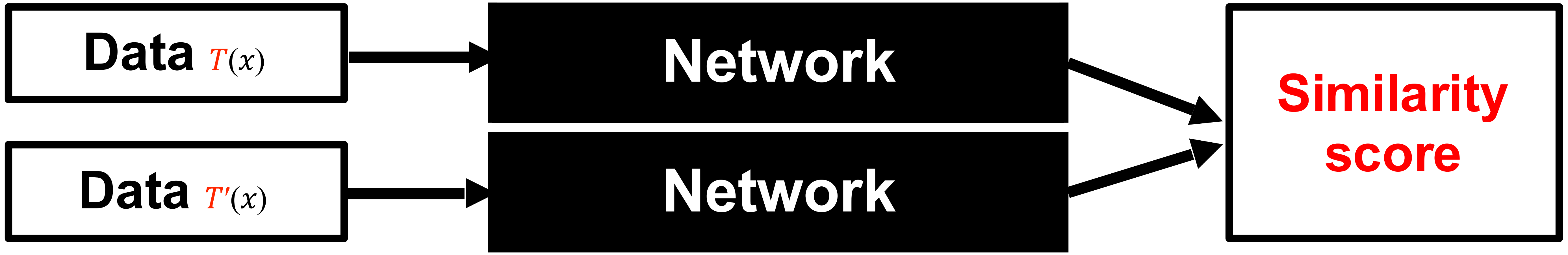
case	MLP	aug+	cos	epochs	batch	ImageNet acc.
	unsup. pre-train					
MoCo v1 [6]				200	256	60.6
SimCLR [2]	✓	✓	✓	200	256	61.9
SimCLR [2]	✓	✓	✓	200	8192	66.6
MoCo v2	✓	✓	✓	200	256	**67.5**
results of ***longer*** *unsupervised training follow:*						
SimCLR [2]	✓	✓	✓	1000	4096	69.3
MoCo v2	✓	✓	✓	800	256	**71.1**

Table 2. **MoCo** ***vs*****. SimCLR**: ImageNet linear classifier accuracy (**ResNet-50, 1-crop 224×224**), trained on features from unsupervised pre-training. "aug+" in SimCLR includes blur and stronger color distortion. SimCLR ablations are from Fig. 9 in [2] (we thank the authors for providing the numerical results).

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Non-contrastive methods

- Extract representations from two transformed versions of a data point, encourage these representations to be similar (or to have other desirable properties)
  - **Contrastive methods**: train using both positive (similar) and negative (dissimilar) pairs
    - Key challenge: sampling of negative pairs
  - **Non-contrastive methods**: train with only positive examples
    - Key challenge: avoiding degenerate solutions (all representations collapsing to constant output value)



Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

BYOL

- Use momentum encoder, but without the queue of negative examples
- Use projection head like SimCLR, add prediction head to online network

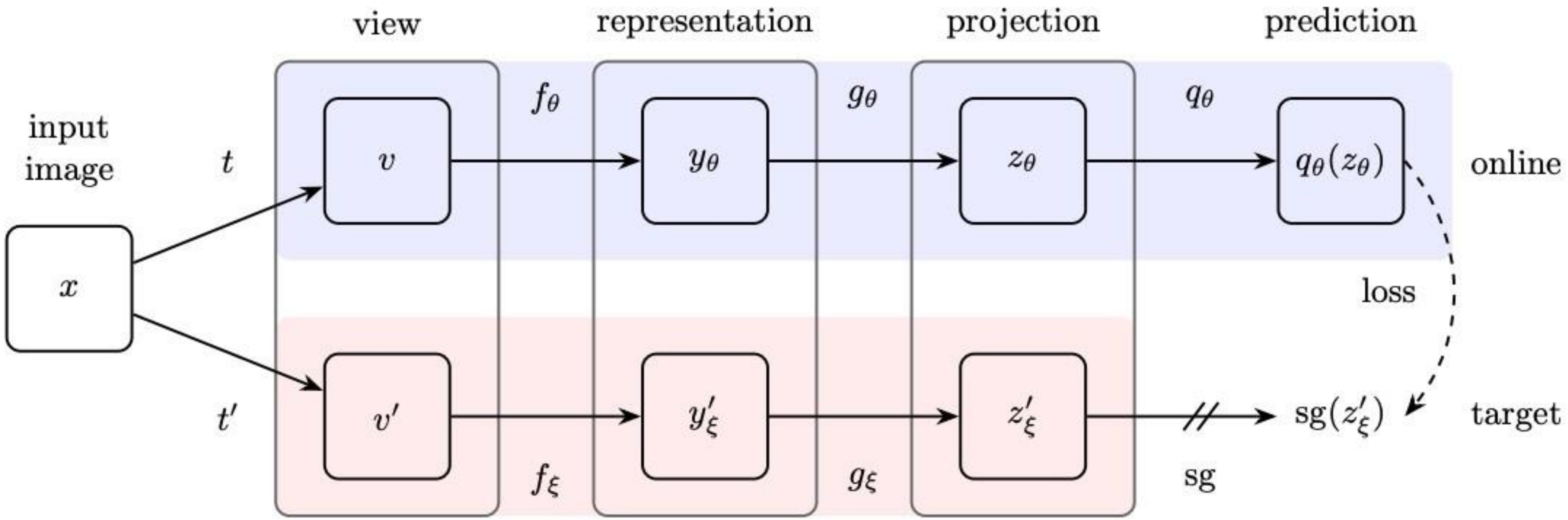


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $\mathrm{sg}(z'_\xi)$, where θ are the trained weights, ξ are an exponential moving average of θ and sg means stop-gradient. At the end of training, everything but f_θ is discarded, and y_θ is used as the image representation.

J.-B. Grill et al. **Bootstrap Your Own Latent A New Approach to Self-Supervised Learning**. NeurIPS 2020

BYOL: Evaluation

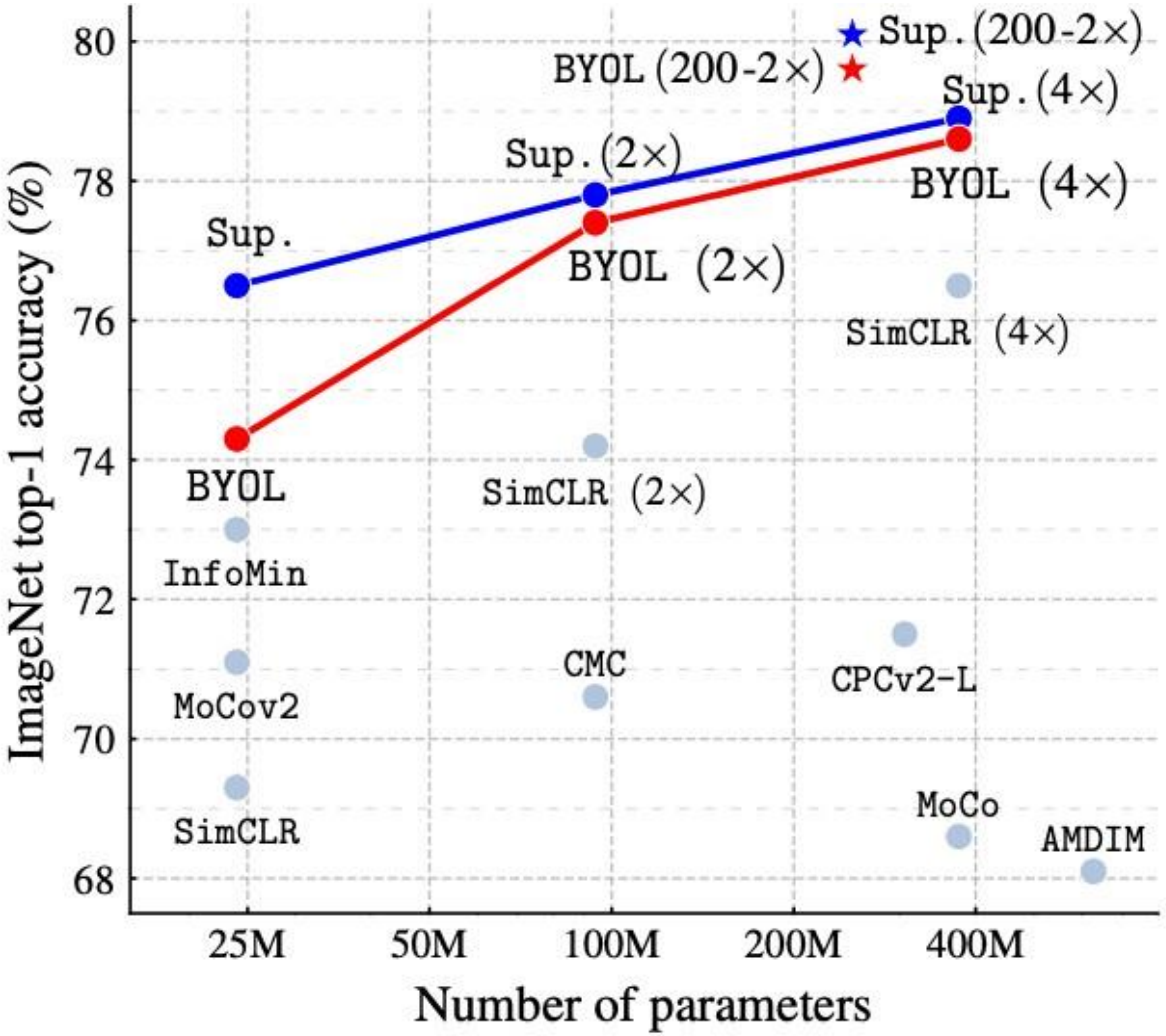


Figure 1: Performance of BYOL on ImageNet (linear evaluation) using ResNet-50 and our best architecture ResNet-200 (2×), compared to other unsupervised and supervised (Sup.) baselines [8].

But remember …

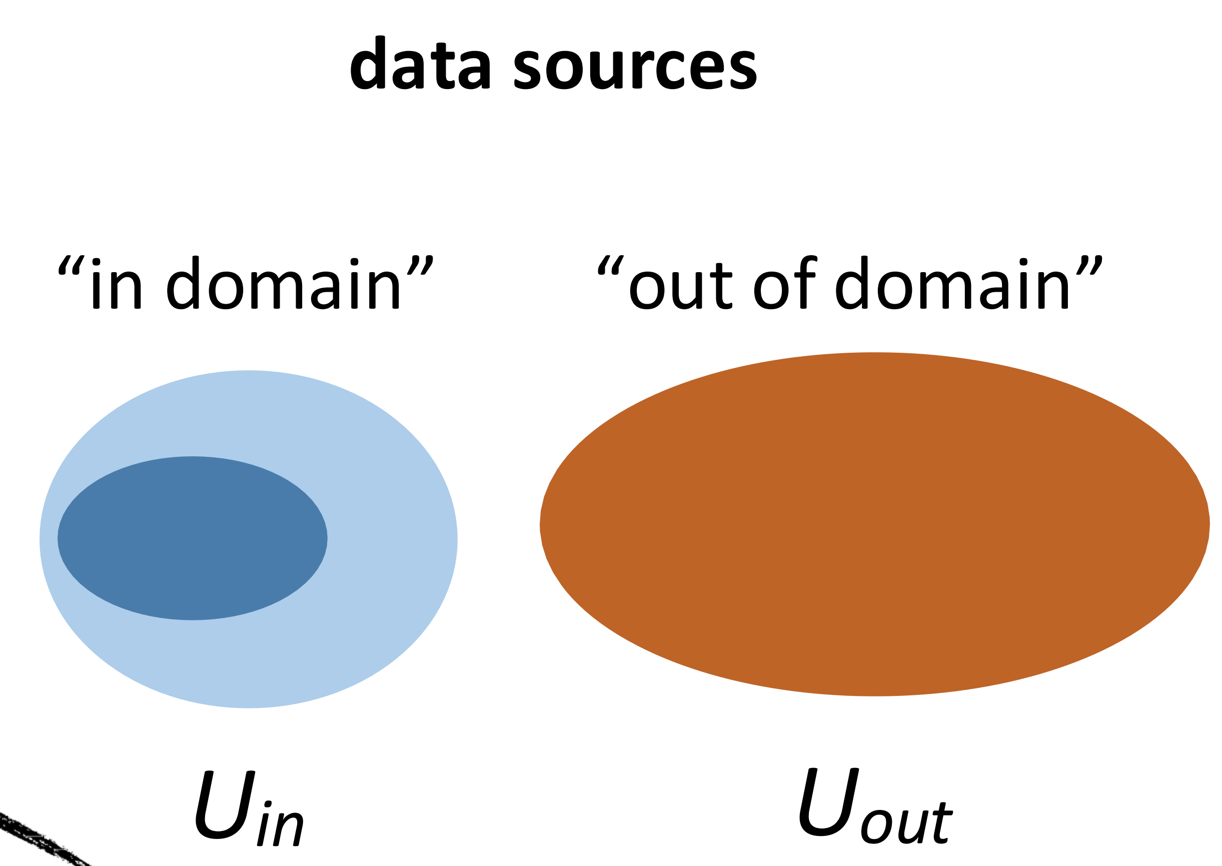
“Small” domain shifts can impact performance

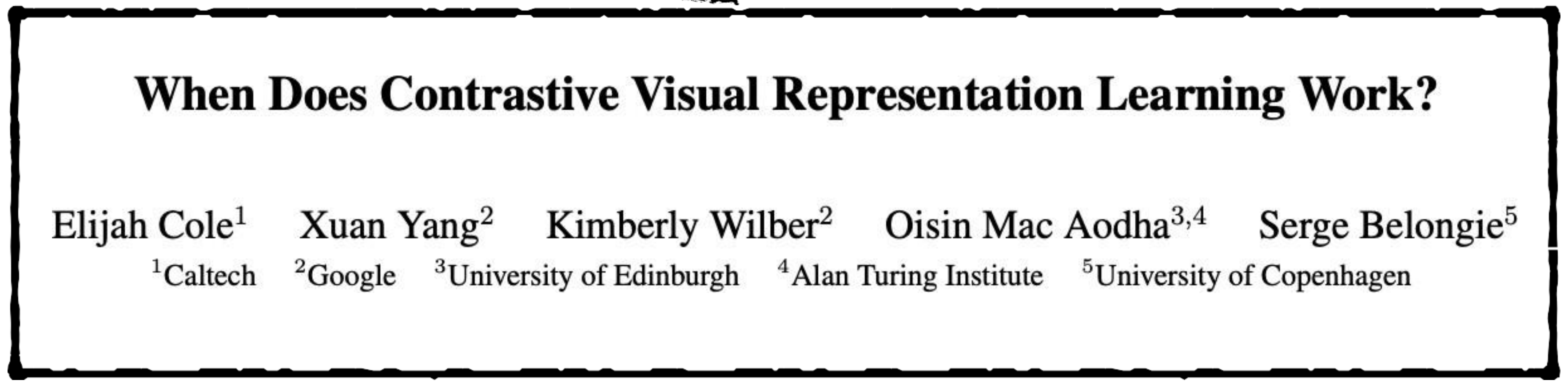
- resolution, size/pose/class, novel classes

Self/semi-supervised learning is brittle in fine-grained domains

- difficult task, long-tailed data

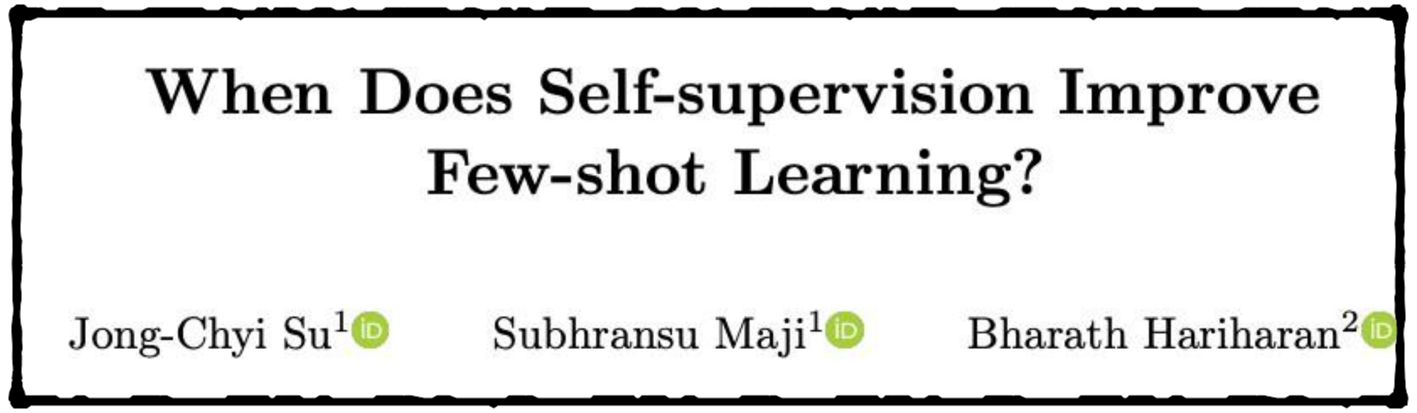
Far from working on non-curated data!





When Does Contrastive Visual Representation Learning Work?

Elijah Cole[1] Xuan Yang[2] Kimberly Wilber[2] Oisin Mac Aodha[3,4] Serge Belongie[5]
[1]Caltech [2]Google [3]University of Edinburgh [4]Alan Turing Institute [5]University of Copenhagen



When Does Self-supervision Improve Few-shot Learning?

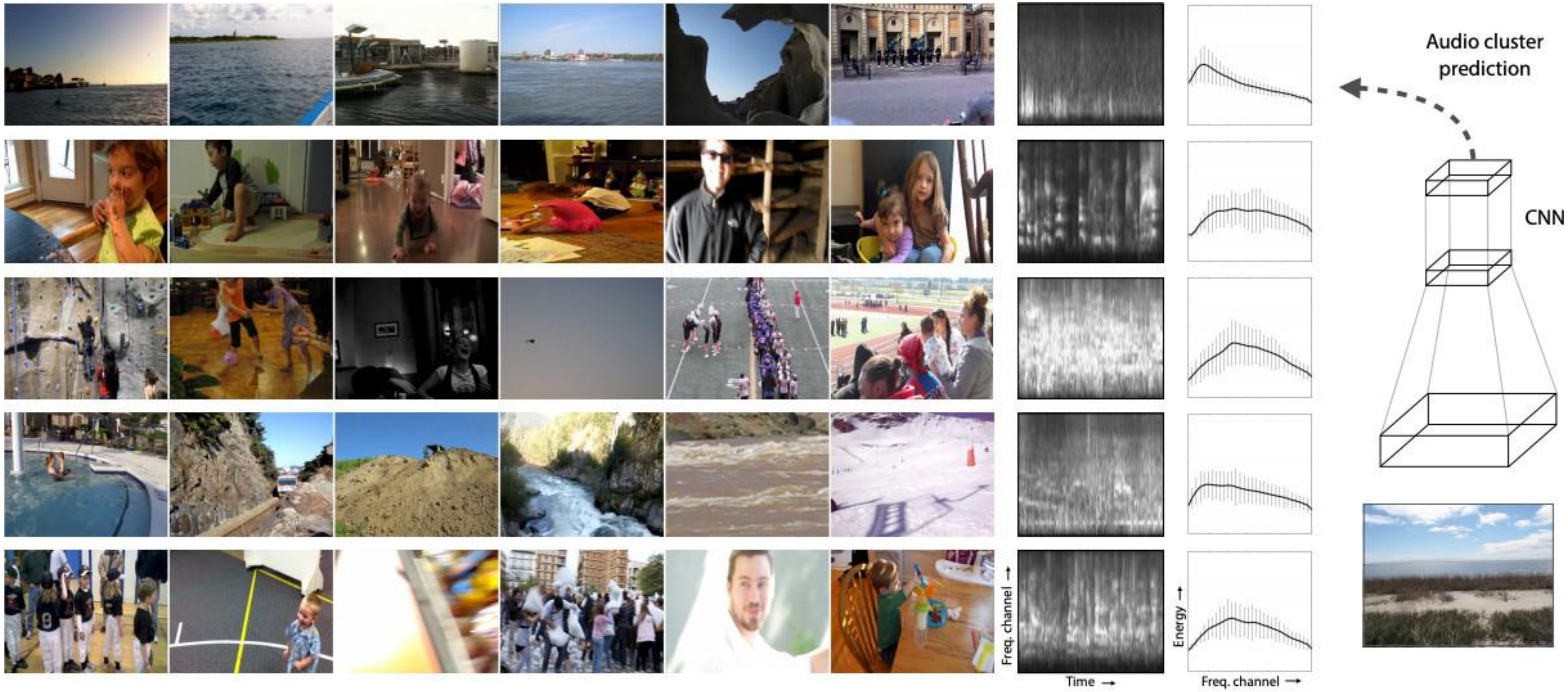
Jong-Chyi Su[1] Subhransu Maji[1] Bharath Hariharan[2]

Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- “Siamese” methods
  - Contrastive methods
  - Non-contrastive methods
- Self-supervision beyond still images
  - Video, audio, language

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

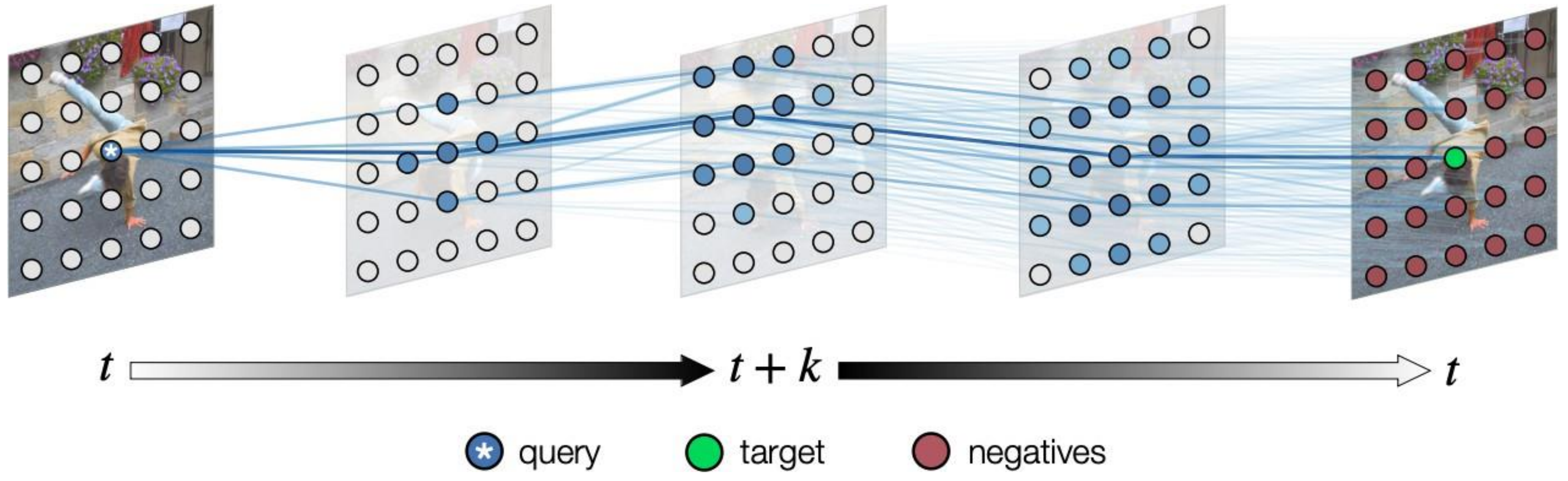
Learning from audio



(a) Images grouped by audio cluster (b) Clustered audio stats. (c) CNN model

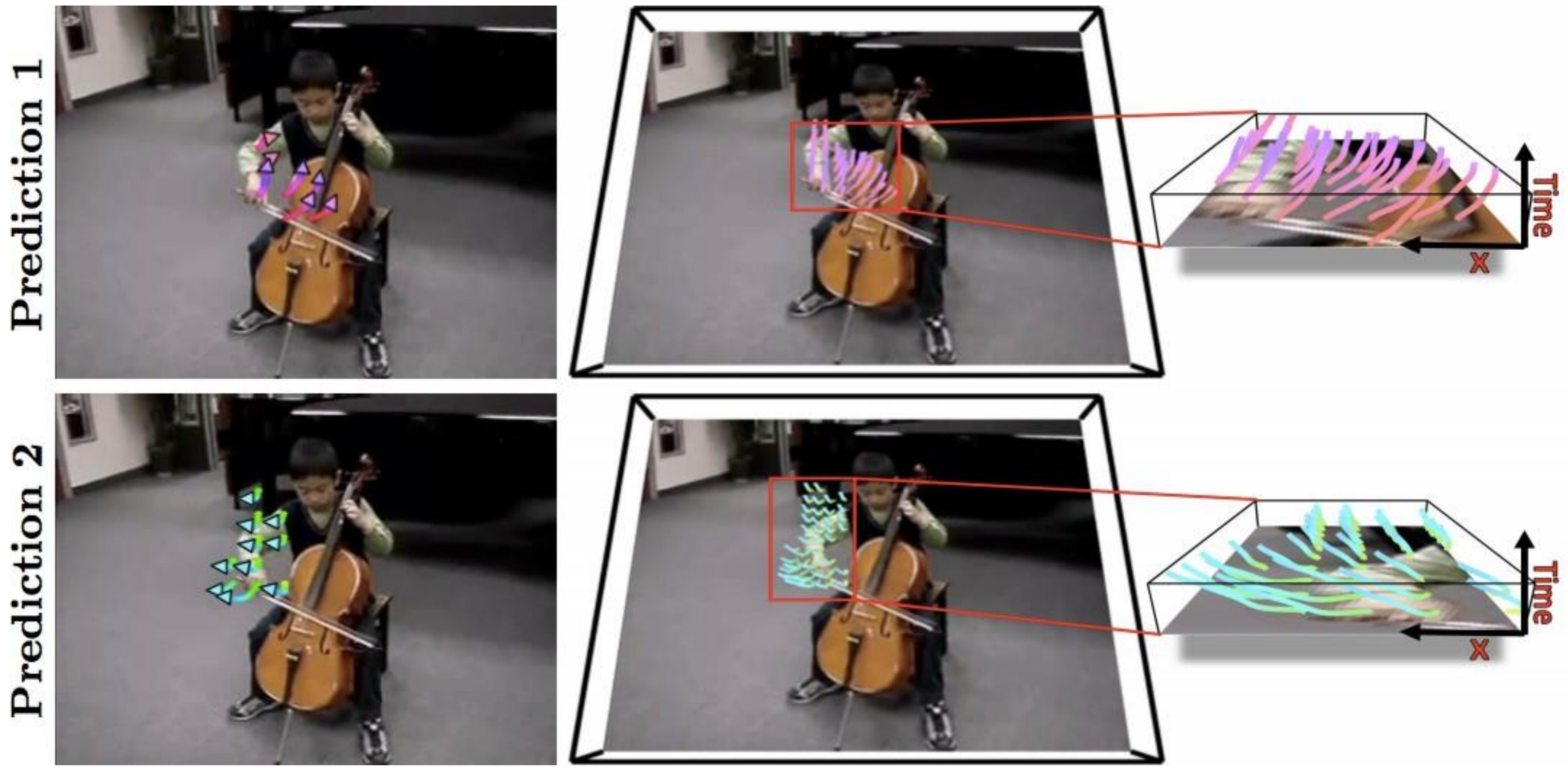
A. Owens et al. [Ambient Sound Provides Supervision for Visual Learning](). ECCV 2016

Video correspondence features



A. Jabri, A. Owens, and A. Efros. [Space-time correspondence as a contrastive random walk](). NeurIPS 2020

Future prediction



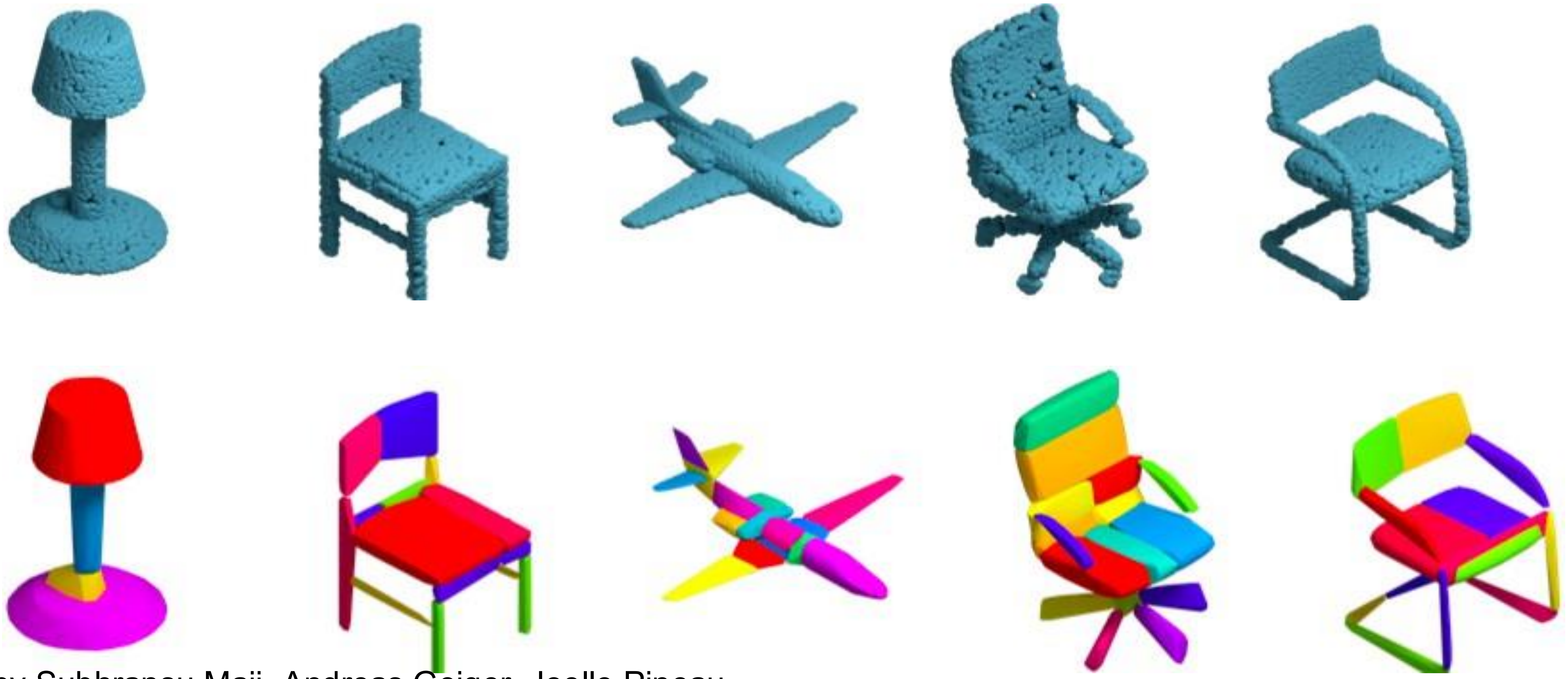
J. Walker et al. [An Uncertain Future: Forecasting from Static Images Using Variational Autoencoders](). ECCV 2016

3D shapes and convexity

- **Final Task:** separate 3D ***objects*** (chairs, tables..) into ***parts*** (legs, back, handles…)

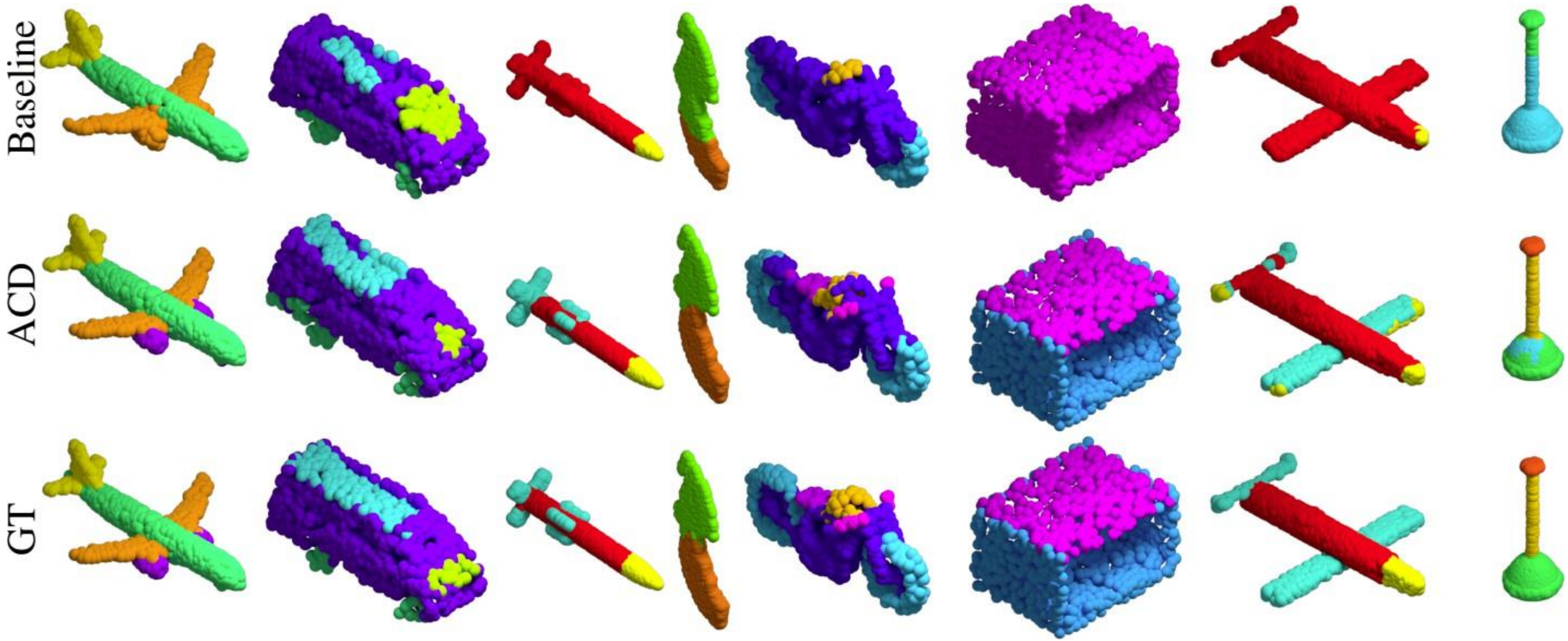
More on the pretext task - (approximate) convexity

- **Pretext Task:** off-the-shelf package for approximate convex decomposition
  - Get a large number of unlabeled 3D shapes
  - Run **off-the-shelf "ACD" software** to get decompositions
  - Train your favorite 3D neural network on this, and then apply on final task



[ECCV 2020]

10-Shot Segmentation Results



Gadelha and RoyChowdhury, et al., ECCV 2020

Large Language Models

pre-train transformers on text

Next-token-prediction

The model is given a sequence of words with the goal of predicting the next word.

Example:
Hannah is a ___

Hannah is a *sister*
Hannah is a *friend*
Hannah is a *marketer*
Hannah is a *comedian*

Masked-language-modeling

The model is given a sequence of words with the goal of predicting a 'masked' word in the middle.

Example
Jacob [mask] reading

Jacob *fears* reading
Jacob *loves* reading
Jacob *enjoys* reading
Jacob *hates* reading

Human examples
Human preferences
RLHF

Finetuning

ChatGPT

https://twitter.com/thealexbanks/status/1624400398114234370

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Masked Autoencoders Are Scalable Vision Learners

Kaiming He*,† Xinlei Chen* Saining Xie Yanghao Li Piotr Dollár Ross Girshick

*equal technical contribution †project lead

Facebook AI Research (FAIR)

Abstract

This paper shows that masked autoencoders (MAE) are scalable self-supervised learners for computer vision. Our MAE approach is simple: we mask random patches of the input image and reconstruct the missing pixels. It is based on two core designs. First, we develop an asymmetric *encoder-decoder architecture, with an encoder that operates only on the visible subset of patches (without mask tokens), along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens. Second, we find that masking a high proportion of the input image,* e.g.*, 75%, yields a nontrivial and meaningful self-supervisory task. Coupling these two designs enables us to train large models efficiently and effectively: we accelerate training (by 3× or more) and improve accuracy. Our scalable approach allows for learning high-capacity models that generalize well:* e.g.*, a vanilla ViT-Huge model achieves the best accuracy (87.8%) among*

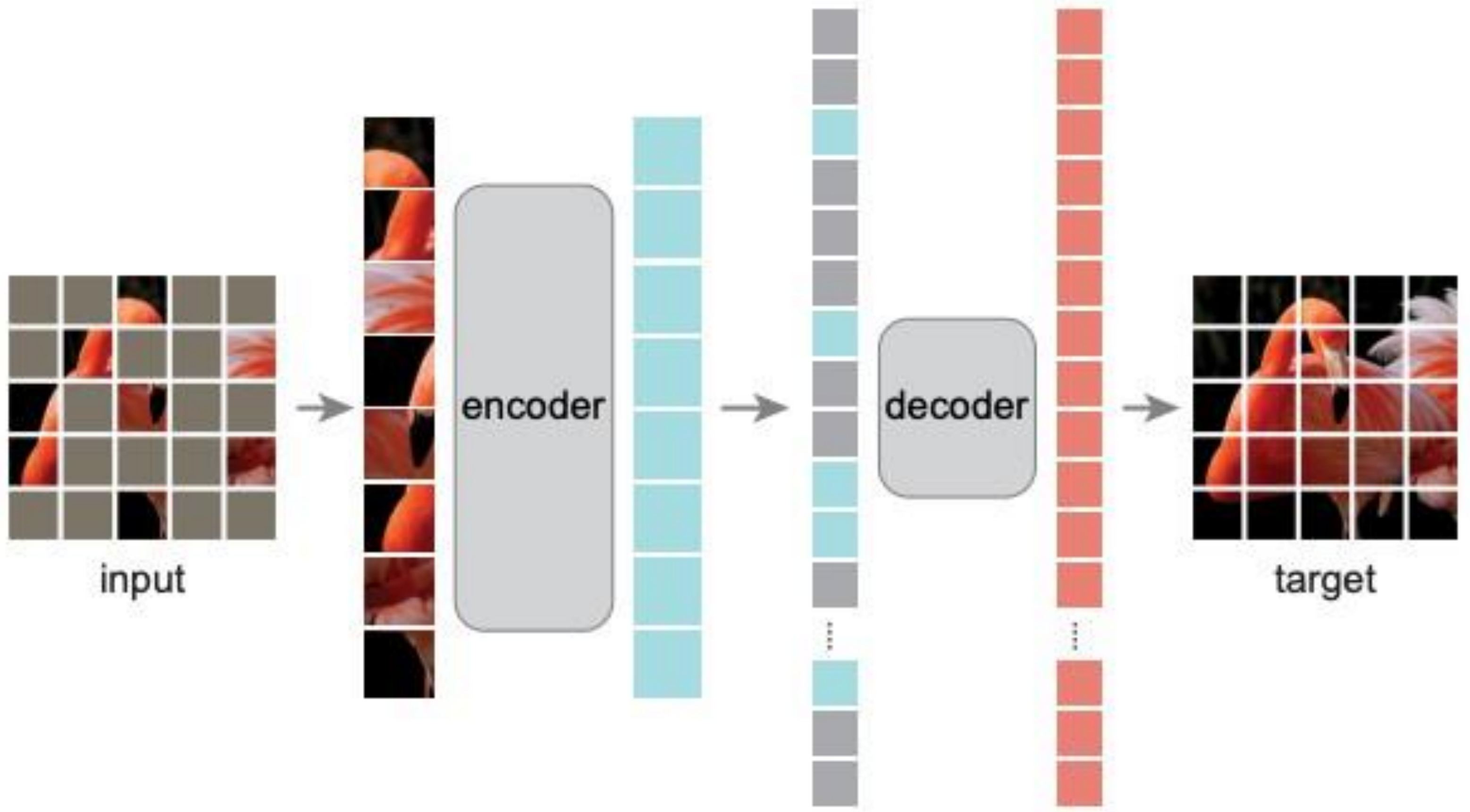


Figure 1. **Our MAE architecture**. During pre-training, a large random subset of image patches (*e.g.*, 75%) is masked out. The encoder is applied to the small subset of *visible patches*. Mask tokens are introduced *after* the encoder, and the full set of encoded patches and mask tokens is processed by a small decoder that reconstructs the original image in pixels. After pre-training, the decoder is discarded and the encoder is applied to uncorrupted images (full sets of patches) for recognition tasks.

Some slides kindly provided by Subhransu Maji, Andreas Geiger, Joelle Pineau

Summary of self-supervision via pretext-tasks

Pretext Tasks:

- Pretext tasks focus on “visual common sense”, e.g., rearrangement, predicting rotations, inpainting, colorization, etc.
- The models are forced learn good features about natural images, e.g., semantic representation of an object category, in order to solve the pretext tasks
- We don’t care about pretext task performance, but rather about the utility of the learned features for downstream tasks (classification, detection, segmentation)

Problems:

- Designing good pretext tasks is tedious and some kind of “art”
- The learned representations may not be general