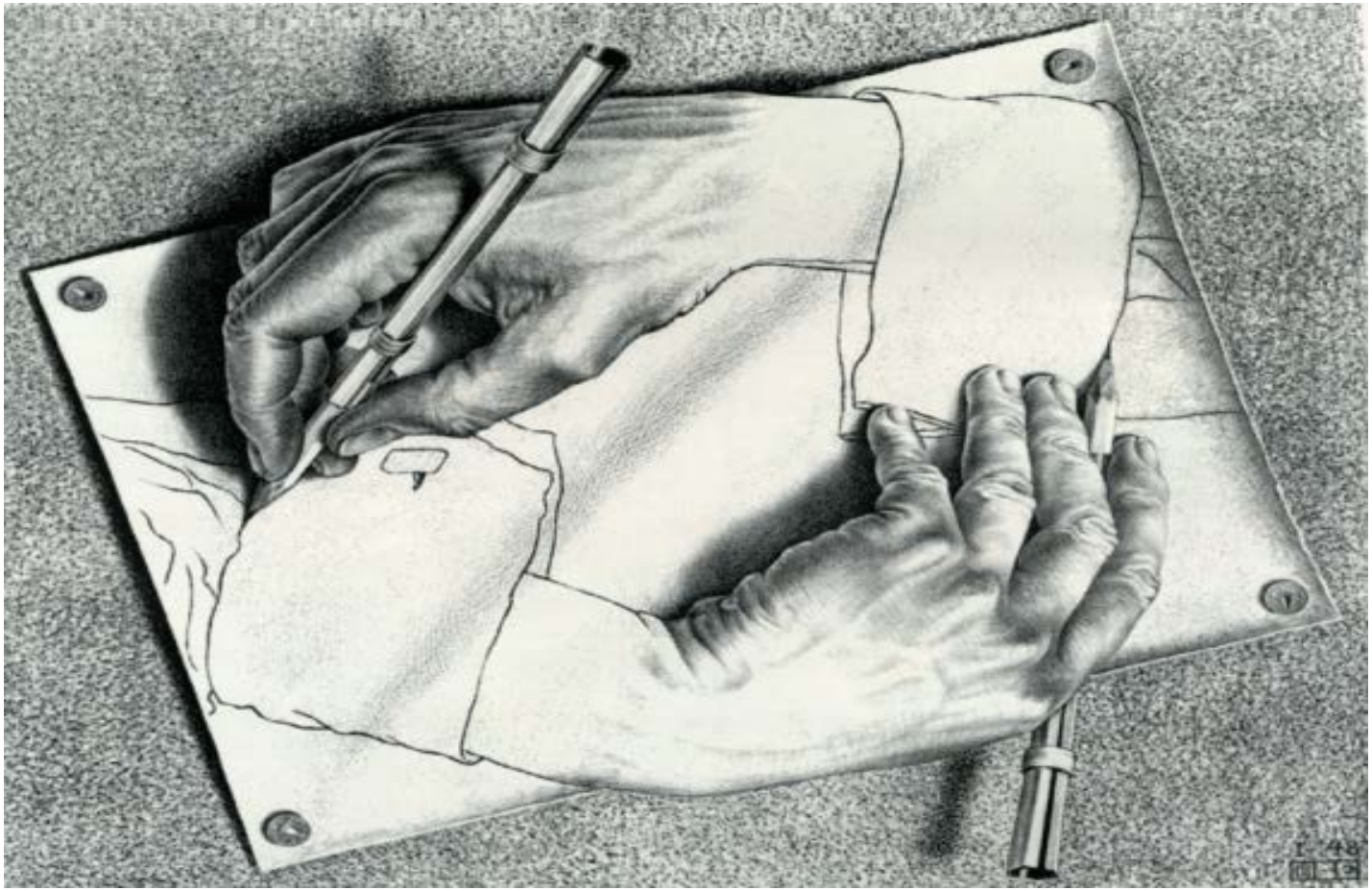
Self-supervised learning



M.C. Escher, *Drawing Hands* (1948) – via A. Efros

Motivation

- Overcoming reliance on *supervised pre-training*

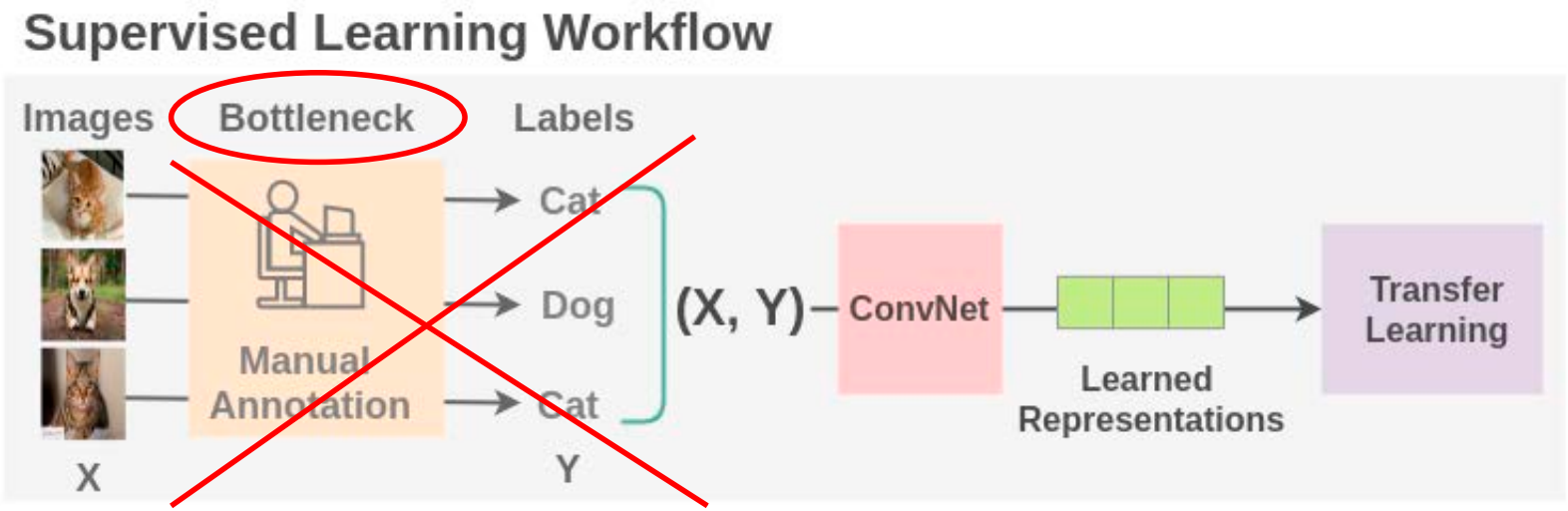
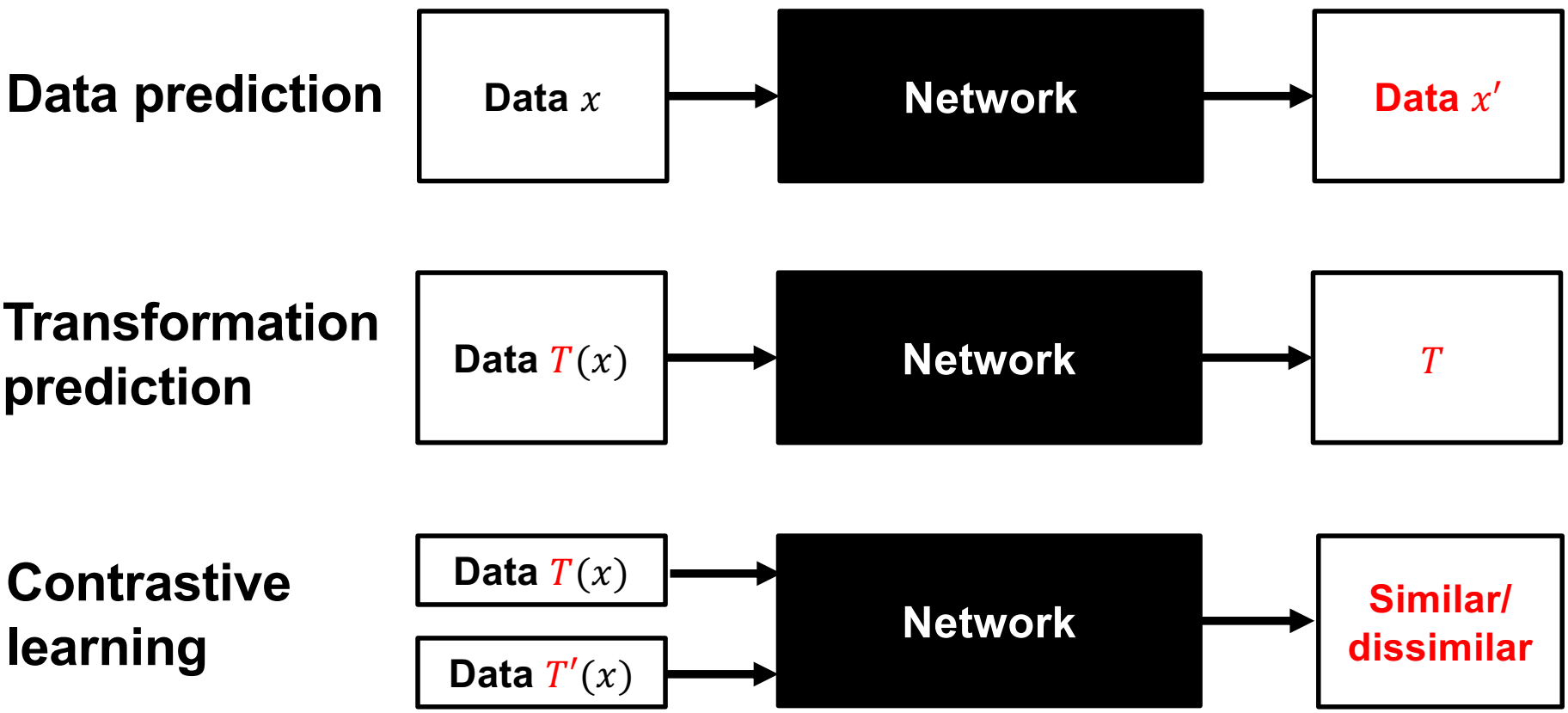


Figure source

Self-supervised vs. *unsupervised* learning

- The terms are sometimes used interchangeably in the literature, but self-supervised learning is a particular kind of unsupervised learning
- **Self-supervised learning**: the learner "makes up" labels from the data and then solves a supervised task
- **Unsupervised learning**: any kind of learning without labels
  - Clustering and quantization
  - Dimensionality reduction, manifold learning
  - Density estimation
  - Learning to sample

Types of self-supervised learning



Source: A. Efros

Self-supervised learning: Outline

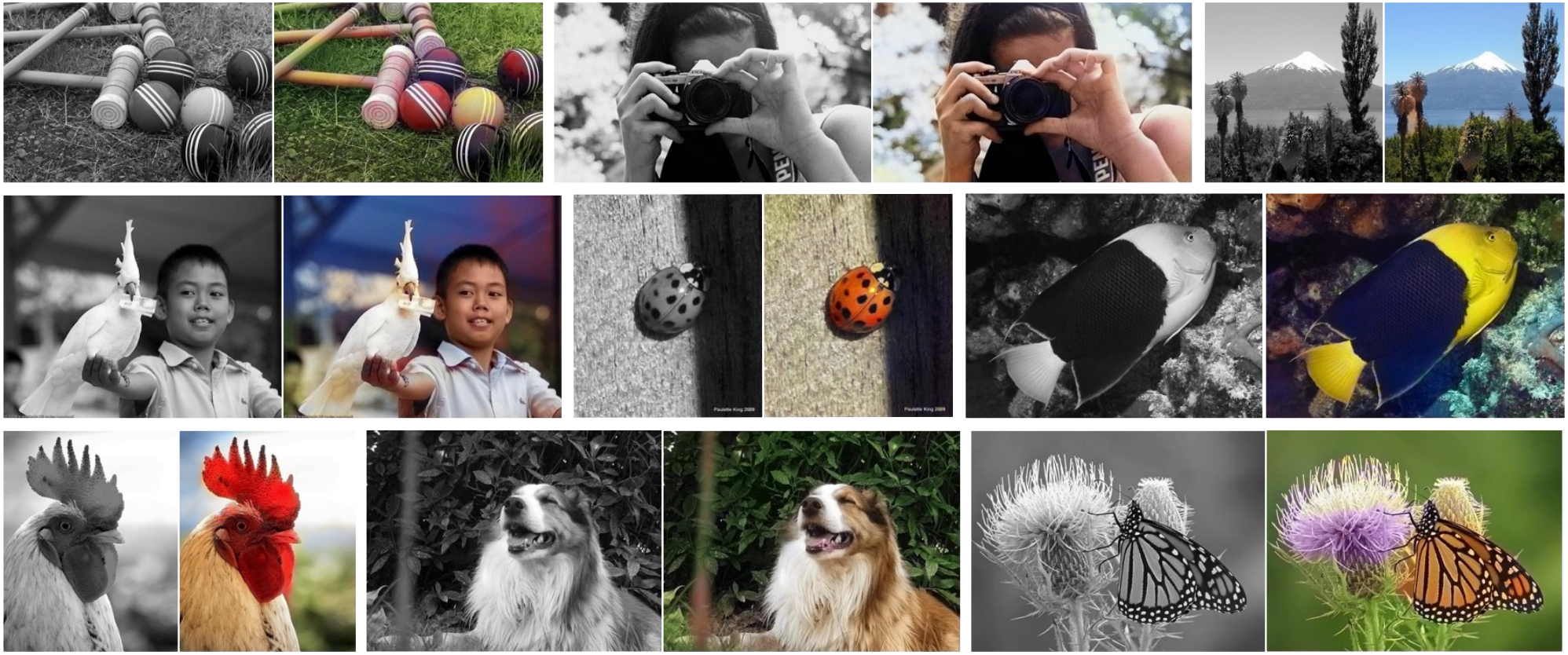
- Data prediction
  - Colorization
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Deep clustering and instance prediction
- Contrastive learning
  - PIRL, MoCo, SimCLR, SWaV
- Self-supervision beyond still images
  - Audio, video, language

Self-Supervision as data prediction

- Colorization
- Inpainting
- Cross-channel encoding
- Future prediction

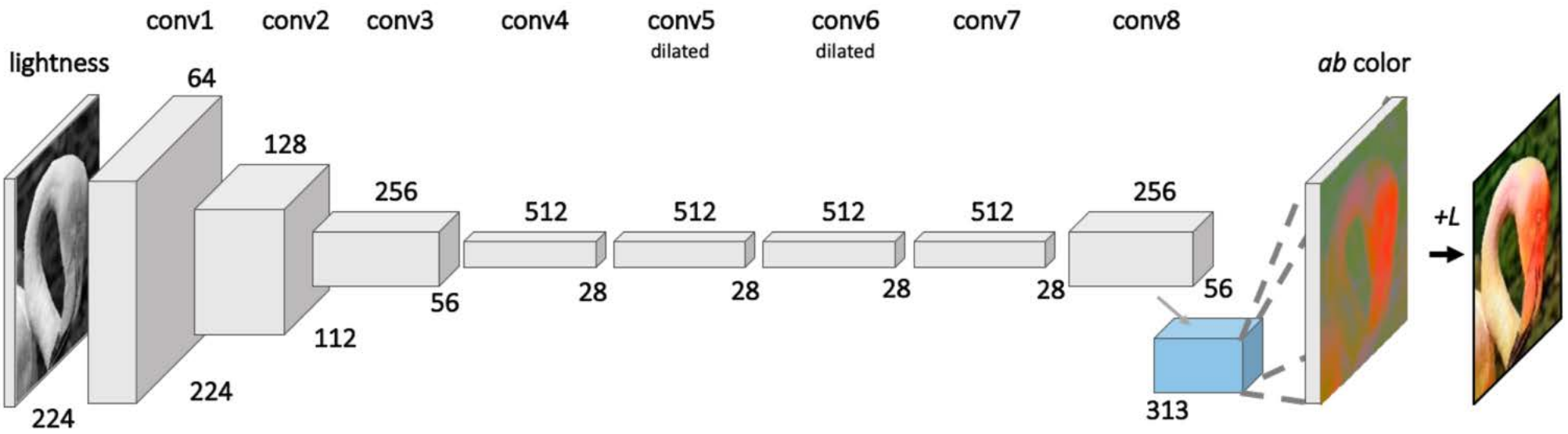
Source: A. Efros

Colorization

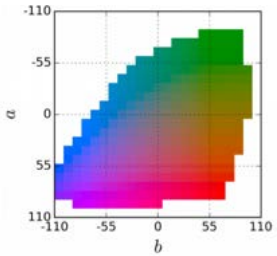


R. Zhang, P. Isola, and A. Efros, Colorful Image Colorization, ECCV 2016

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

Colorization: Results

R. Zhang, P. Isola, and A. Efros, Colorful Image Colorization, ECCV 2016

Failure Cases

Source: A. Efros, R. Zhang

Inherent Ambiguity

Grayscale

Source: A. Efros, R. Zhang

Inherent Ambiguity

Prediction

Ground Truth

Source: A. Efros, R. Zhang

Biases

Source: A. Efros, R. Zhang

Biases

Source: A. Efros, R. Zhang

Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction

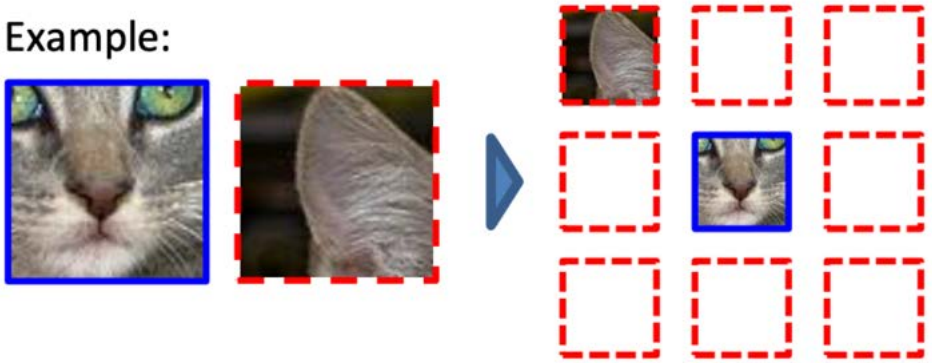
Self-supervision by transformation prediction

- Context prediction
- Jigsaw puzzle solving
- Rotation prediction

Source: A. Efros

Context prediction

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

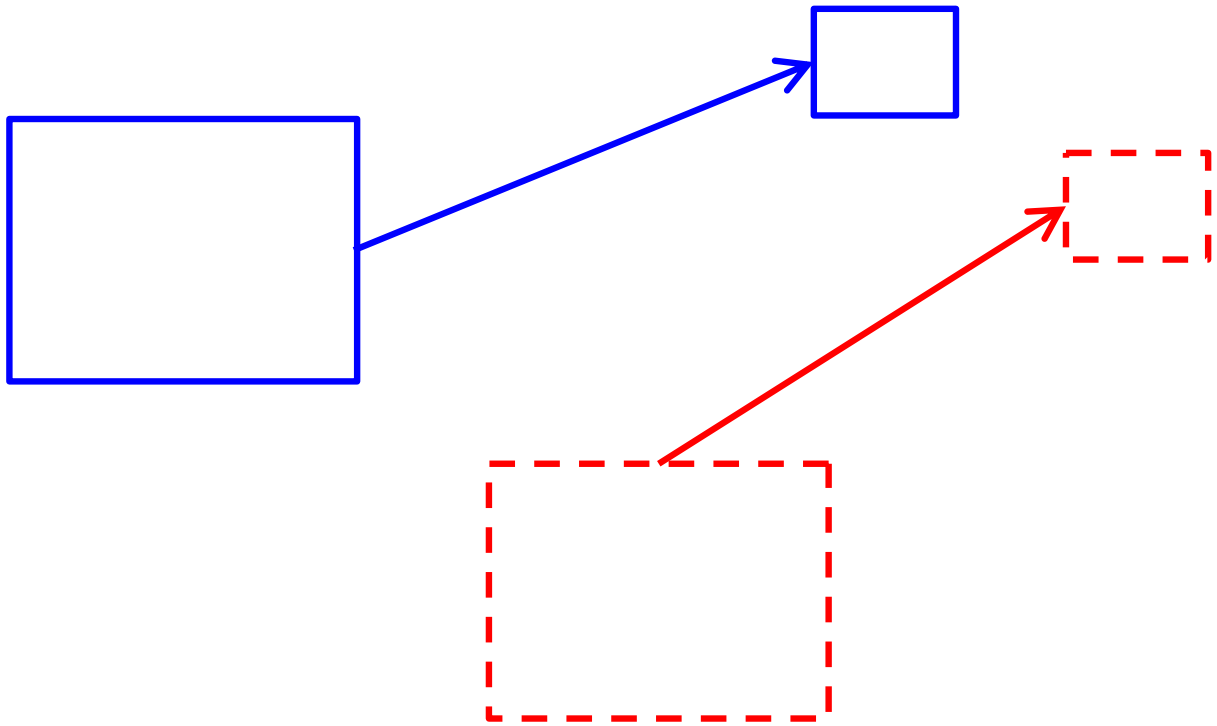


A: Bottom right

A: Top center

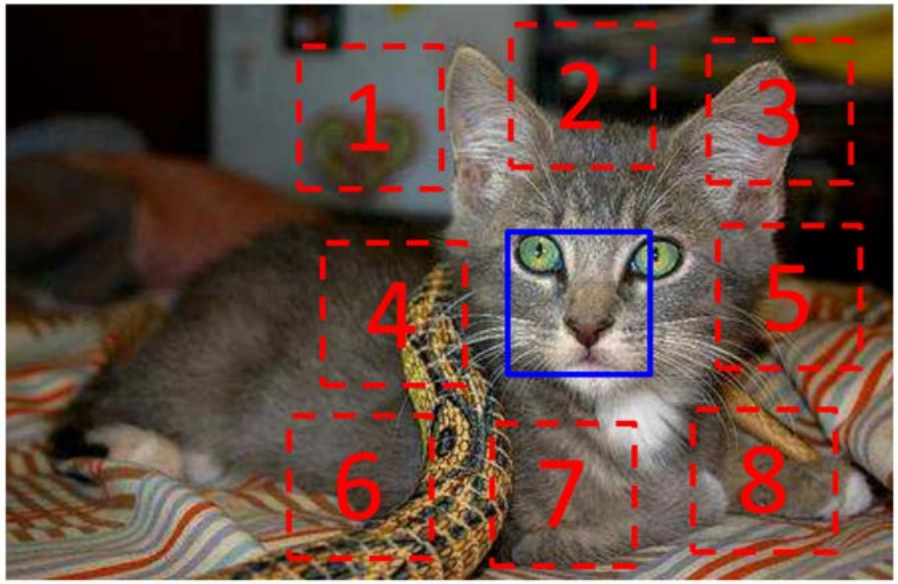
C. Doersch, A. Gupta, A. Efros. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

! "#$%&$' ()%*+,$+"#-' .% / 0#$+,1'2)" / '0'#"#31% / 0#$+,'$014



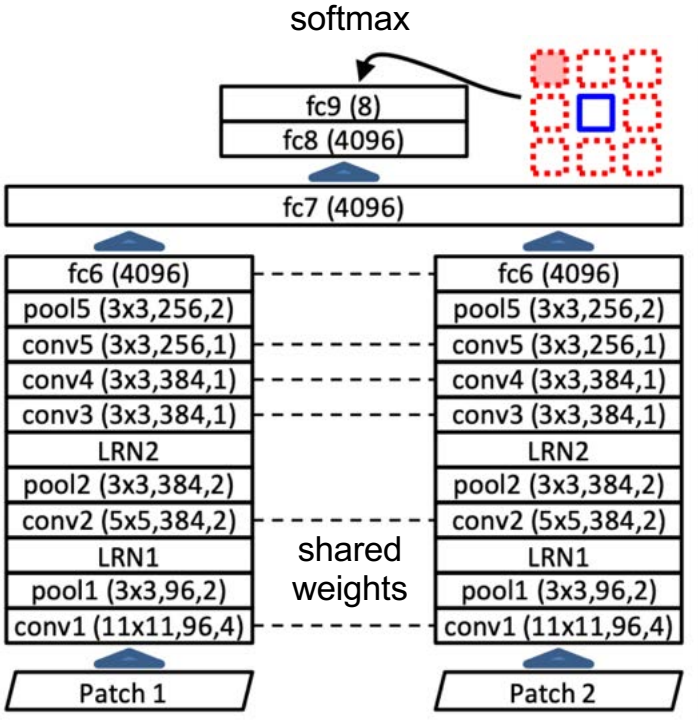
. "5),%-'67'82)"1

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

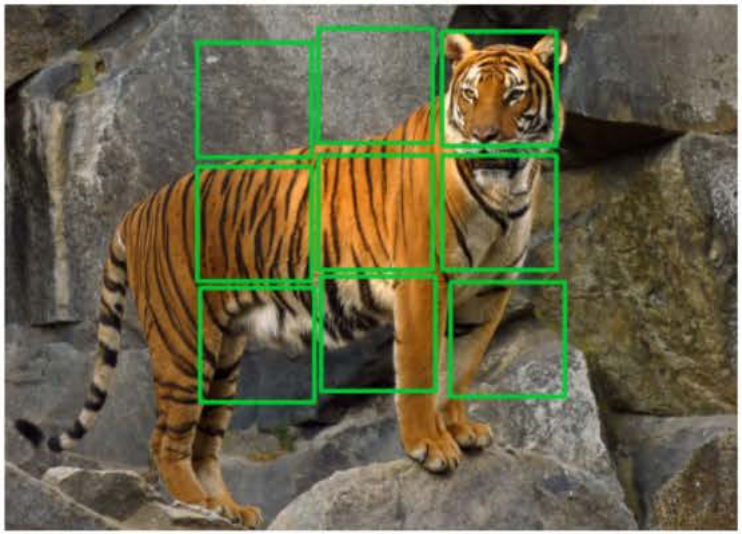
Context prediction: Results

- Use learned weights in R-CNN model to perform detection on PASCAL VOC 2007
- Unsupervised pre-training is 5% mAP better than training from scratch, but still 8% below pre-training with ImageNet label supervision

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

Jigsaw puzzle solving

Crop out tiles



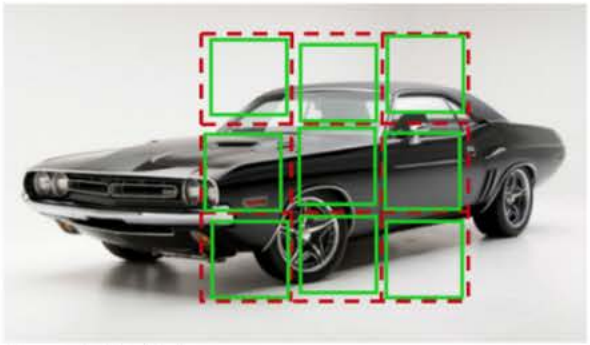
Shuffle

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



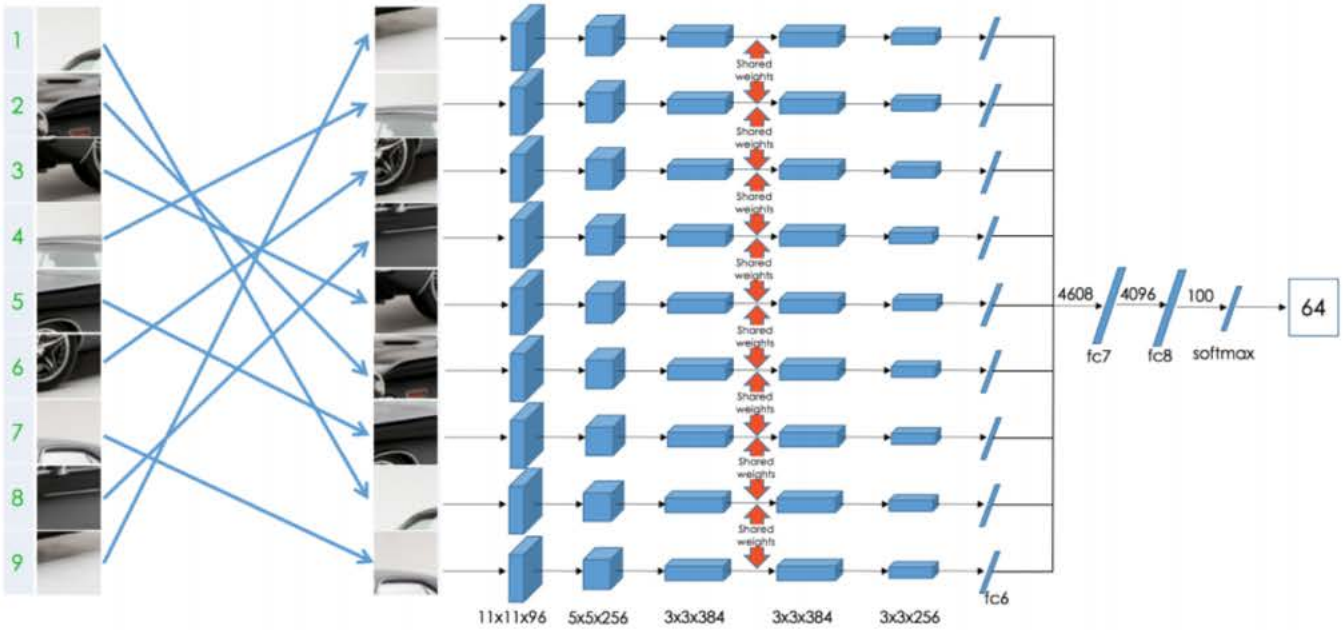
Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

Reorder patches according to the selected permutation

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

Context free network (CFN)



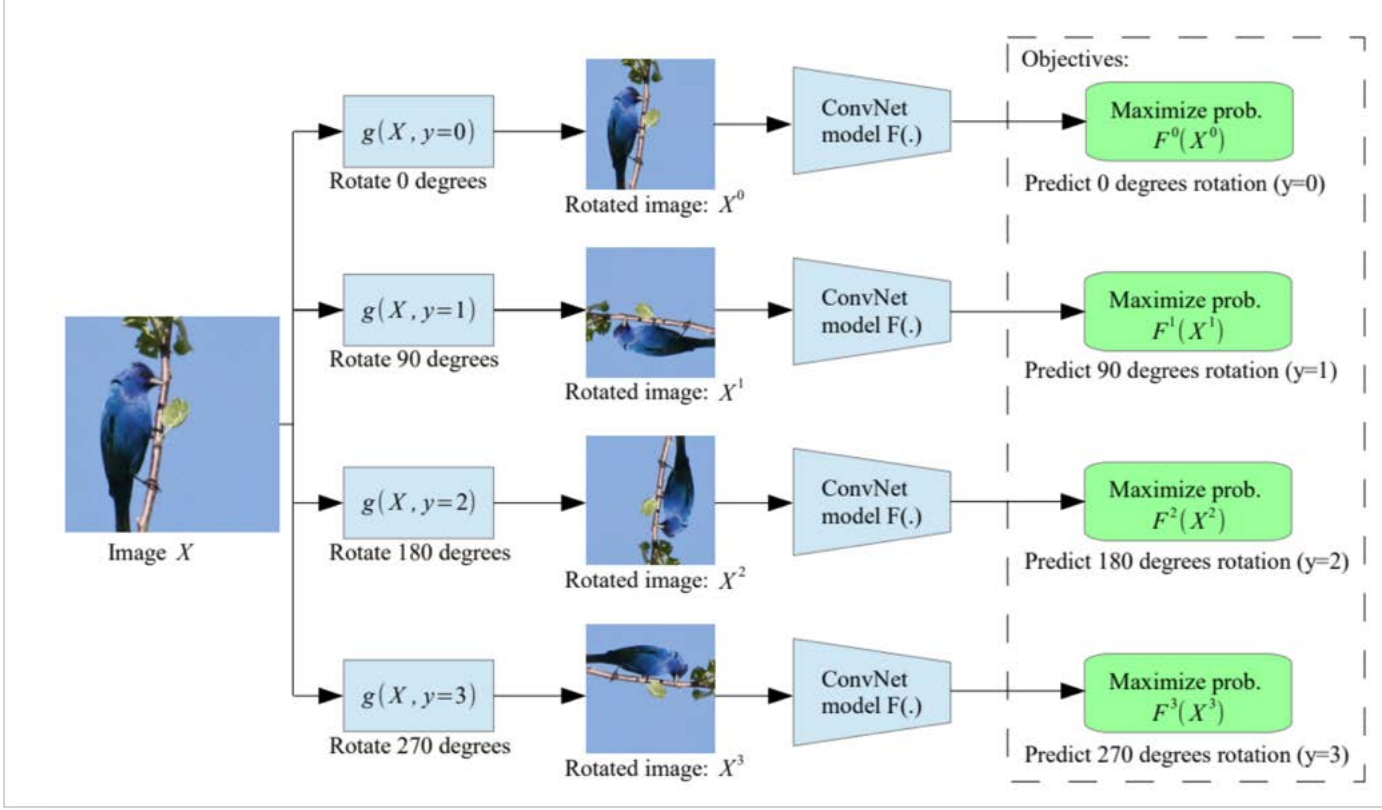
M. Noroozi and P. Favaro. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. ECCV 2016

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

Rotation prediction: PASCAL VOC Transfer results

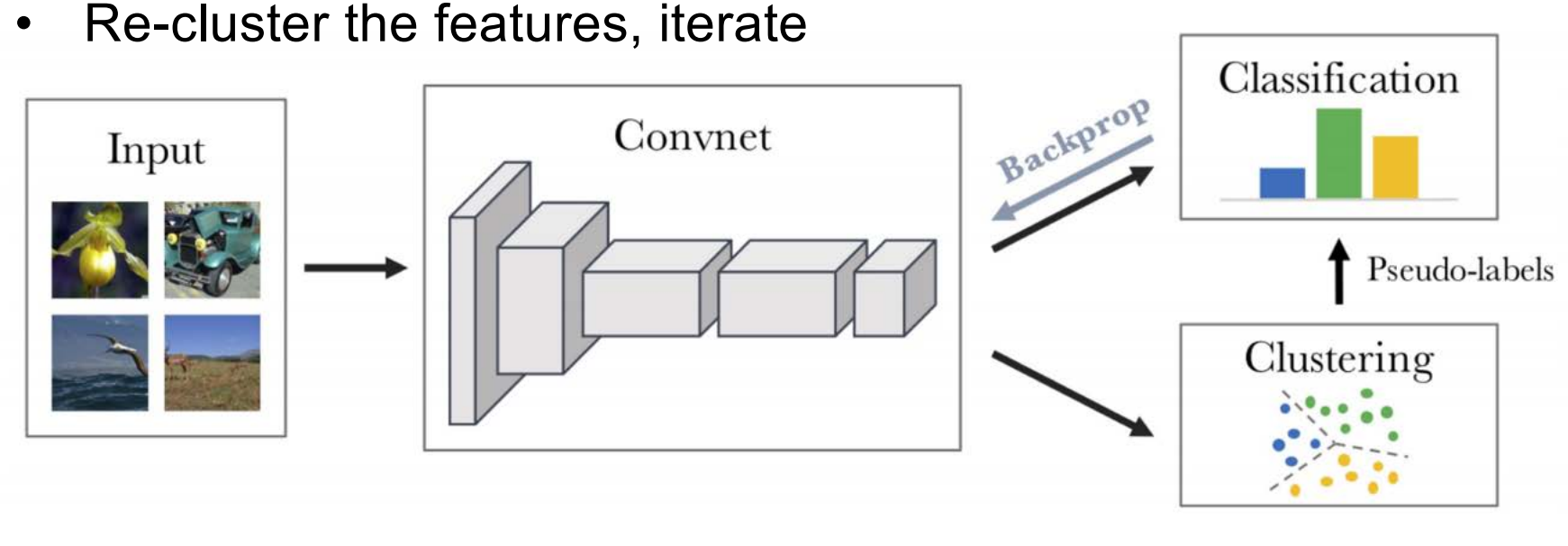
Method	Classification	Detection (mAP)	Segmentation (mIoU)
Supervised (ImageNet)	79.9	56.8	48.0
Colorization	65.6	46.9	35.6
Context	65.3	51.1	
Jigsaw	67.6	53.2	37.6
Rotation	73.0	54.4	39.1

Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Deep clustering and instance prediction

Deep Clustering

- Cluster the features to obtain pseudo-labels
- Use pseudo-label prediction as pretext task to train the network
- Re-cluster the features, iterate



M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. ECCV 2018

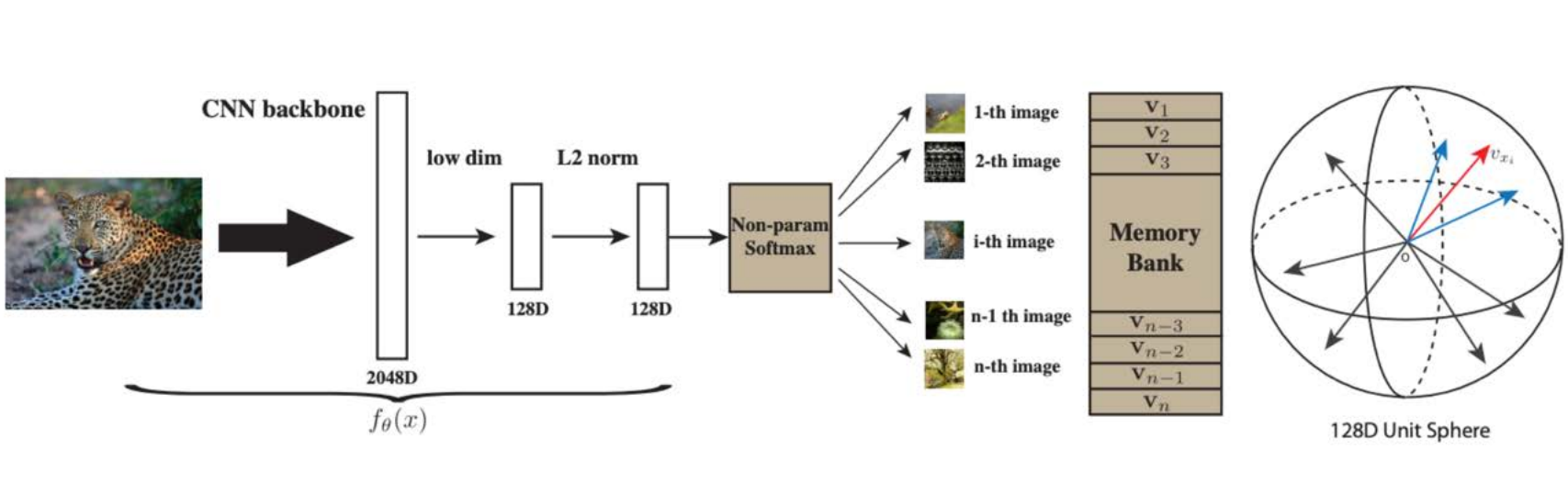
Deep Clustering: PASCAL VOC Transfer results

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
DeepCluster	73.7	55.4	45.1

M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. ECCV 2018

Instance prediction

- Key idea: make each instance into its own class



Z. Wu, Y. Xiong, S. Yu, and D. Lin. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. CVPR 2018

Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Deep clustering and instance prediction
- Contrastive learning
  - PIRL, MoCo, SimCLR, SWaV

Contrastive methods

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

Source: A. Efros

Contrastive methods

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

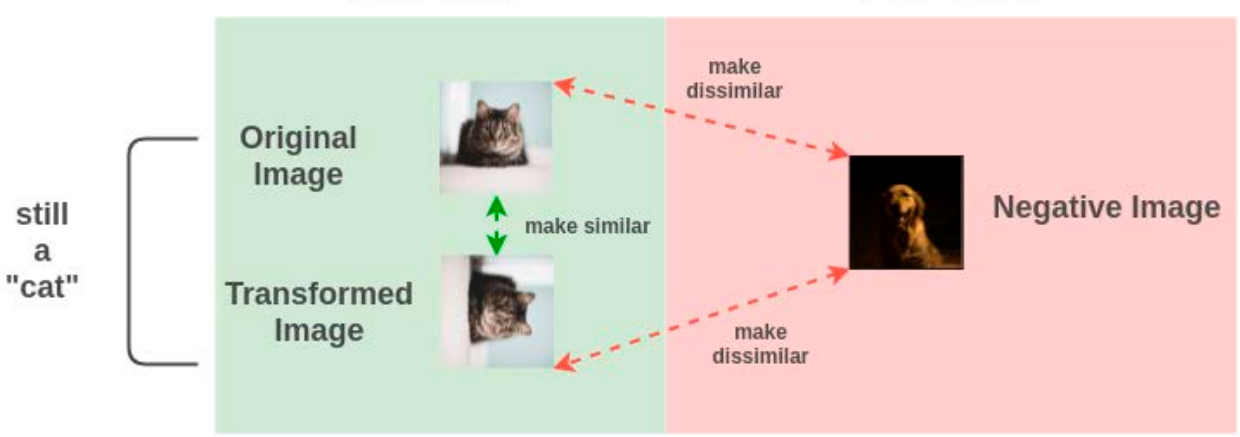


Figure source

Contrastive loss formulation

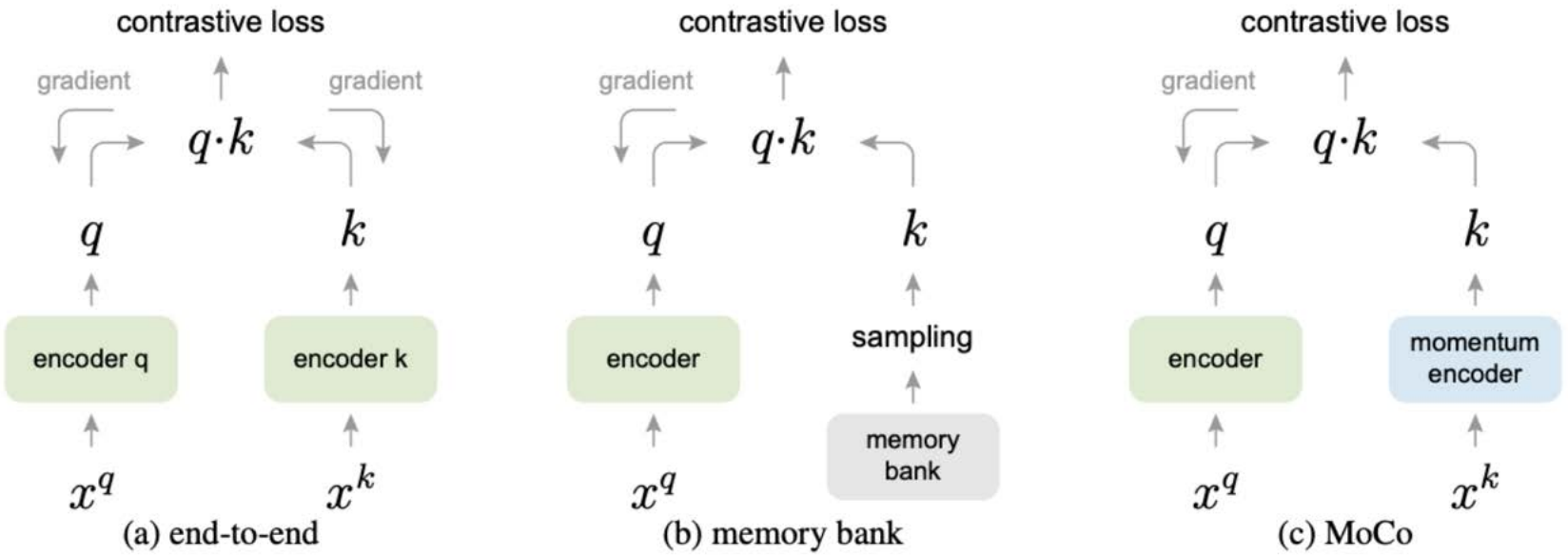
- Given: query point x, positive samples x^+, negative samples x^-
  - Positives are typically transformed versions of x, negatives are random examples from the same mini-batch or *memory bank*
- Key idea: learn representation to make x similar to x^+, dissimilar from x^- (similarity is measured by dot product of normalized features)
- Intuitively, contrastive loss for x, x^+ is the loss of a softmax classifier that tries to classify x as x^+ :

$$l(x, x^+) = -\log \frac{\exp(f(x)^T f(x^+)/\tau)}{\exp(f(x)^T f(x^+)/\tau) + \sum_{j=1}^{N} \exp\left(f(x)^T f\left(x_j^-\right)/\tau\right)}$$

  - τ is the *temperature* hyperparameter (determines how concentrated the softmax is)

Momentum contrast

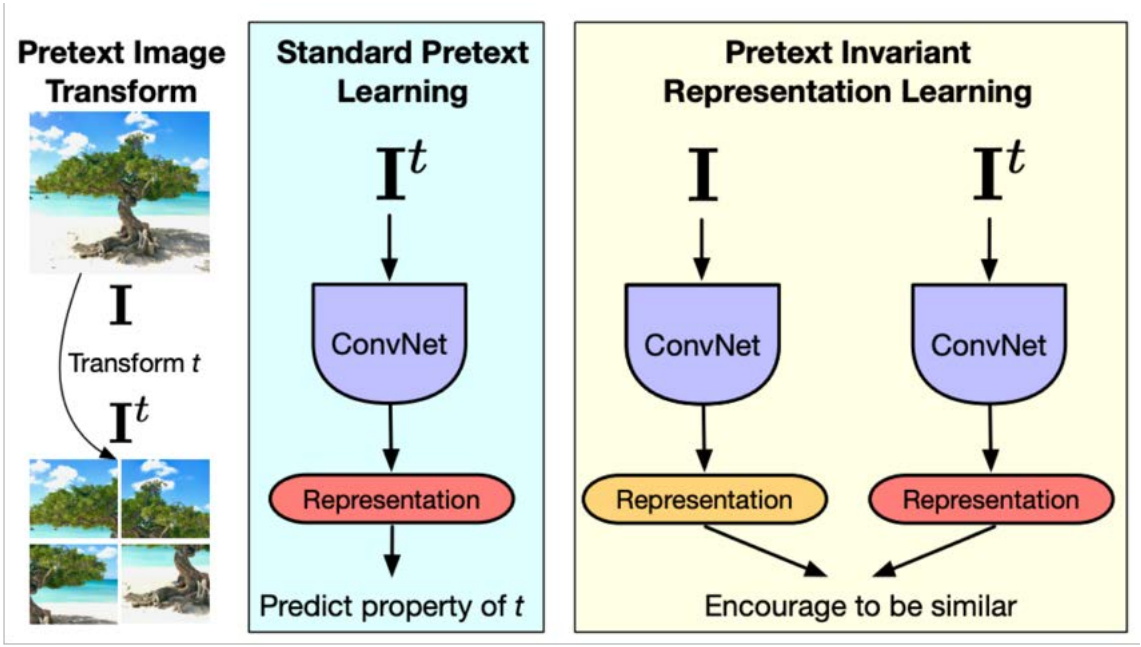
- Use instance discrimination as pretext task, transform query and key by random augmentations, use queue encoded by a momentum encoder instead of memory bank



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

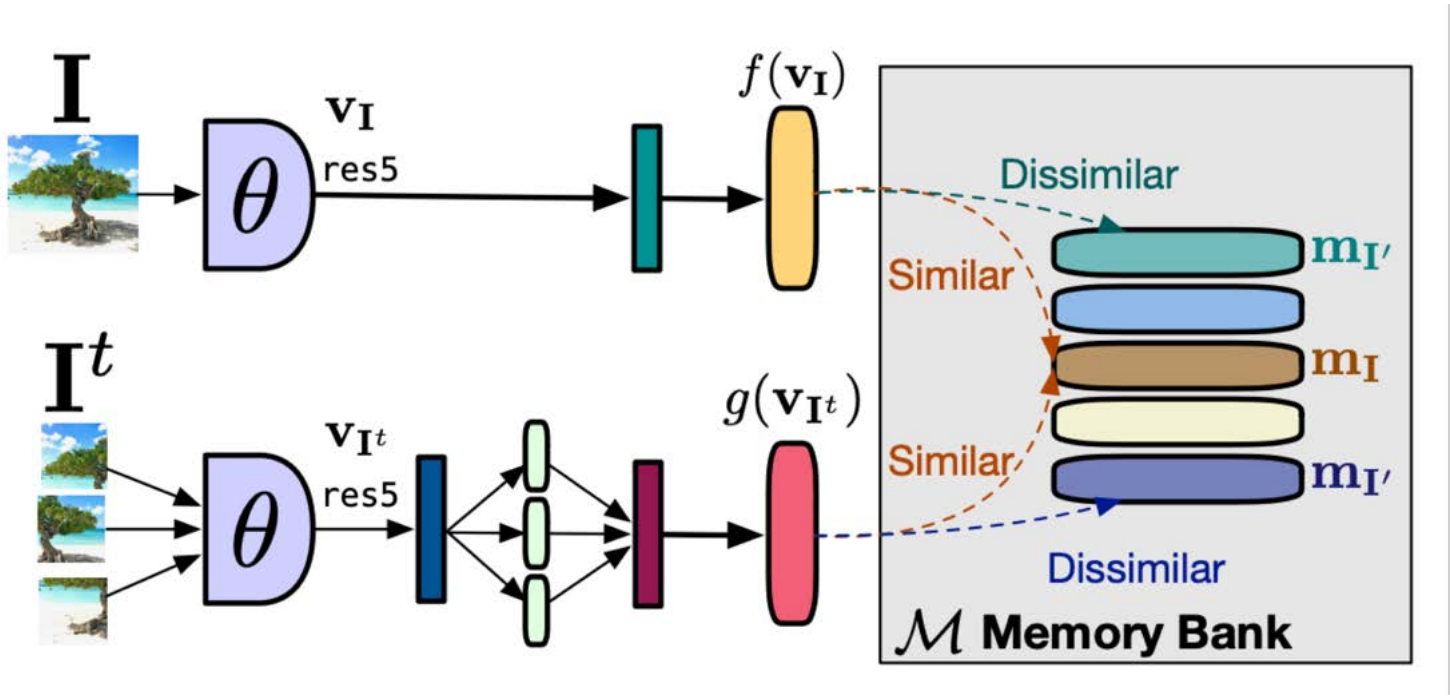
Pretext-invariant representation learning (PIRL)

- Key idea: instead of predicting the transformation of the input, learn a representation *invariant* to the transformation



I. Misra and L. van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. CVPR 2020

Pretext-invariant representation learning (PIRL)



Jigsaw representation

I. Misra and L. van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. CVPR 2020

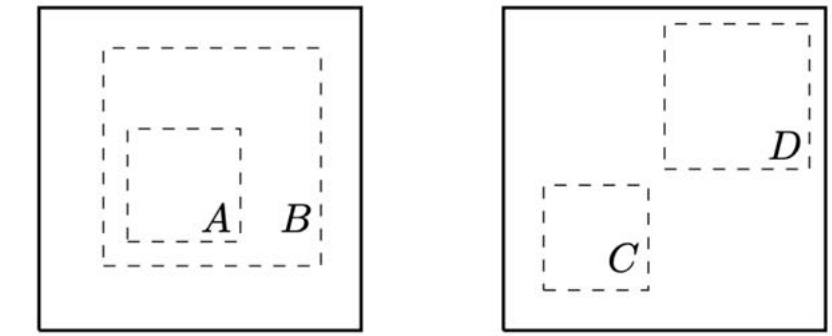
PIRL: Results

Method	Network	AP^{all}	AP^{50}	AP^{75}	ΔAP^{75}
Supervised	R-50	52.6	**81.1**	57.4	=0.0
Jigsaw [19]	R-50	48.9	75.1	52.9	-4.5
Rotation [19]	R-50	46.3	72.5	49.3	-8.1
NPID++ [72]	R-50	52.3	79.1	56.9	-0.5
PIRL **(ours)**	R-50	**54.0**	80.7	**59.7**	**+2.3**
CPC-Big [26]	R-101	–	70.6*	–	
CPC-Huge [26]	R-170	–	72.1*	–	
MoCo [24]	R-50	55.2*†	81.4*†	61.2*†	

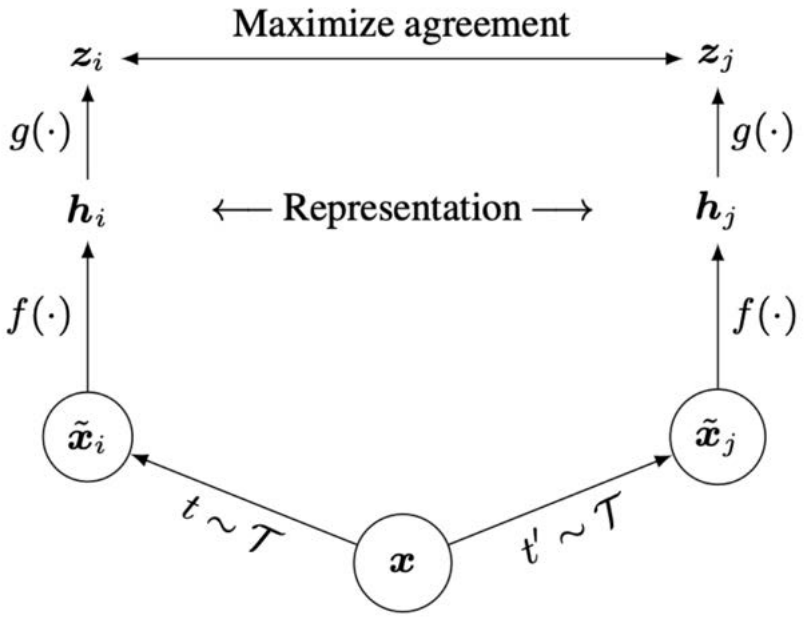
Table 1: Object detection on VOC07+12 using Faster R-CNN. Detection AP on the VOC07 test set after finetuning Faster R-CNN models (keeping BatchNorm fixed) with a ResNet-50 backbone pre-trained using self-supervised learning on ImageNet. Results for supervised ImageNet pre-training are presented for reference. Numbers with * are adopted from the corresponding papers. Method with † finetunes BatchNorm. PIRL significantly outperforms supervised pre-training without extra pre-training data or changes in the network architecture. Additional results in Table 6.

SimCLR

- Form two views of the input by composing data augmentations
  - Cropping and resizing, color distortion, blur



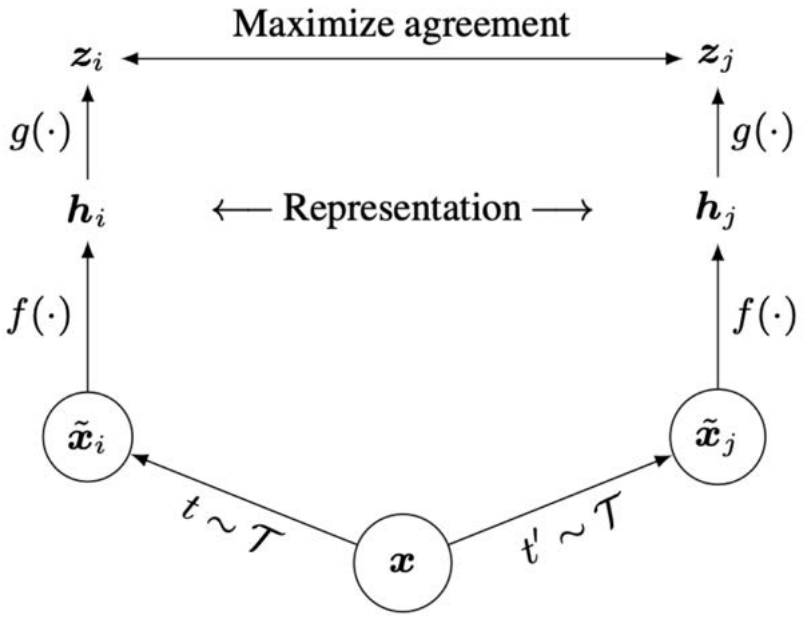
(a) Global and local views. (b) Adjacent views.



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

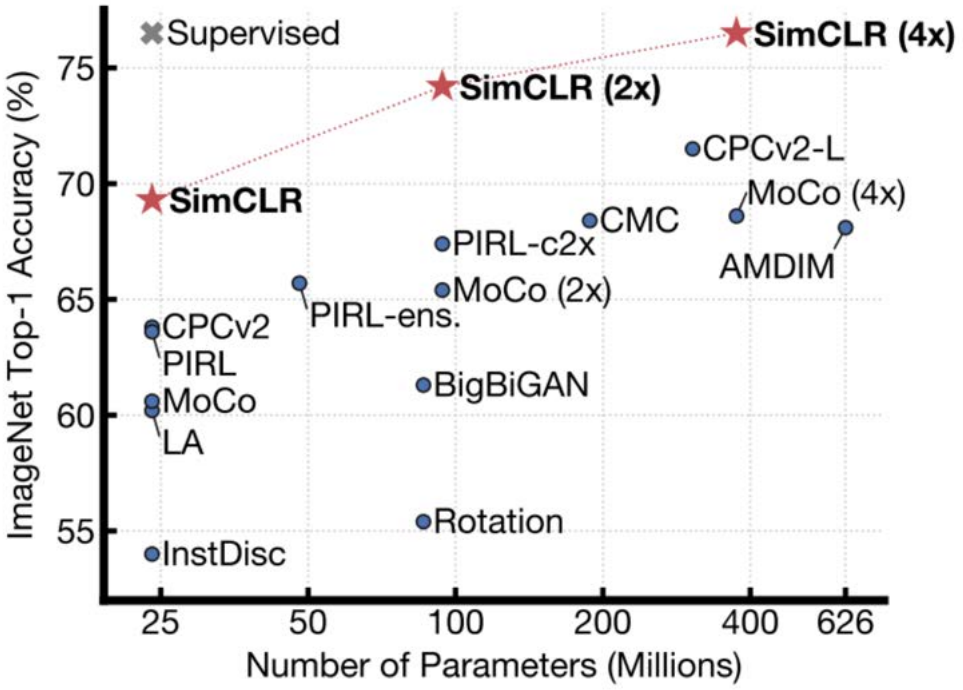
SimCLR

- Form two views of the input by composing data augmentations
  - Cropping and resizing, color distortion, blur
- No memory bank, large mini-batch size (on cloud TPU)
- Introduce nonlinear transformation between representation and contrastive loss (or, use representation a few layers below the contrastive loss)



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

SimCLR: Evaluation

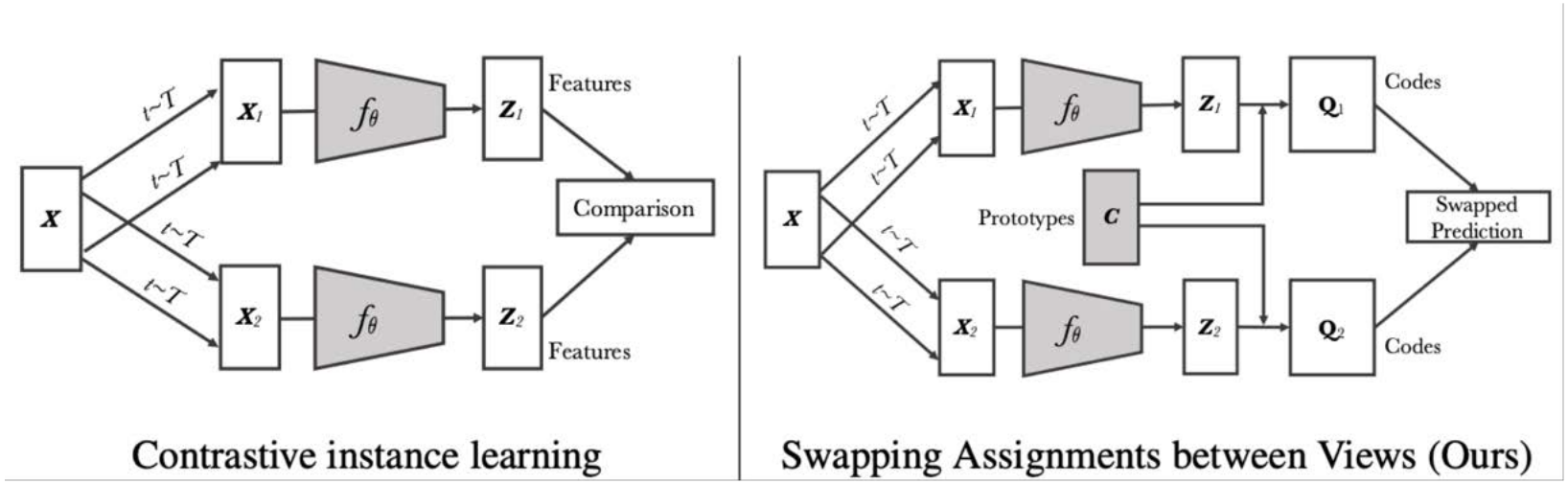


No detection evaluation

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

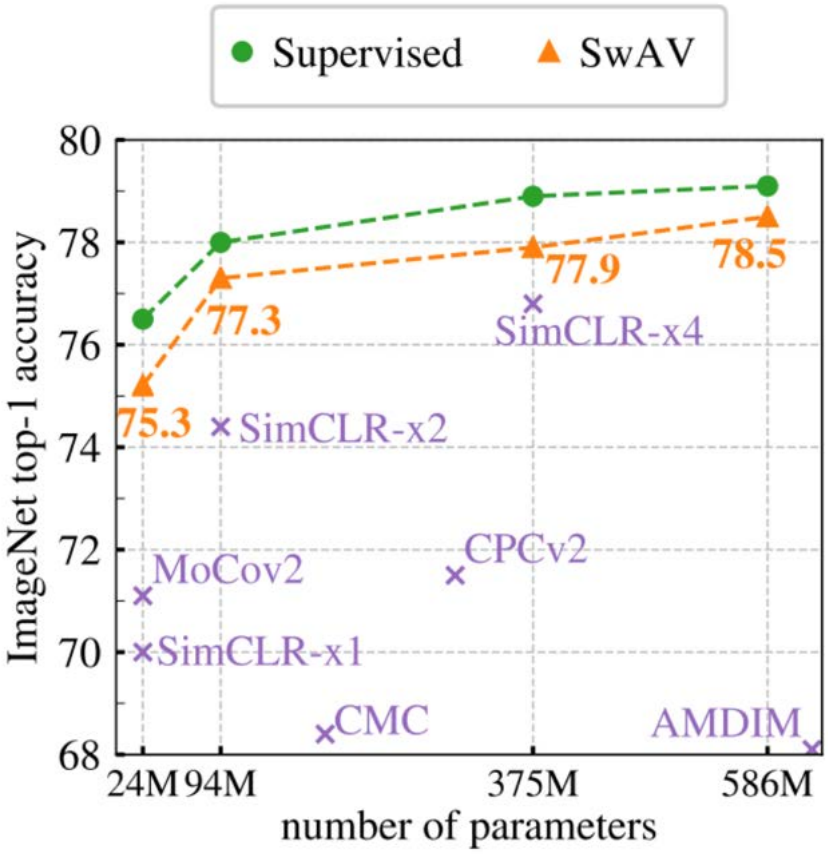
Swapping Assignments Between Views (SWaV)

- Predict cluster assignment of one “view” (transformed version of input image) from representation of another “view”
  - Prototypes or cluster centers are learned online within mini-batch
- Once again, data augmentation strategy matters



M. Caron et al. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. arXiv 2020

SWaV: Results



	Object Detection	
	VOC07+12 (Faster R-CNN)	COCO (DETR)
Supervised	81.3	40.8
SWaV	**82.6**	**42.1**

M. Caron et al. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. arXiv 2020

Why do contrastive methods work?

- L2 normalization of features (before computing dot product to estimate similarity) is important (Wang and Isola, 2020)
- The essential property of the loss is enforcing closeness of positive features while maximizing uniformity of the distribution of features over the hypersphere (Wang and Isola, 2020)
- The choice of data augmentation operations or transformations between two positive “views” is also important and needs further study (Tian et al., 2020)

Provisional conclusion: The Gelato Bet

Made at Berkeley on September 23, 2014 between Alyosha Efros and Jitendra Malik:

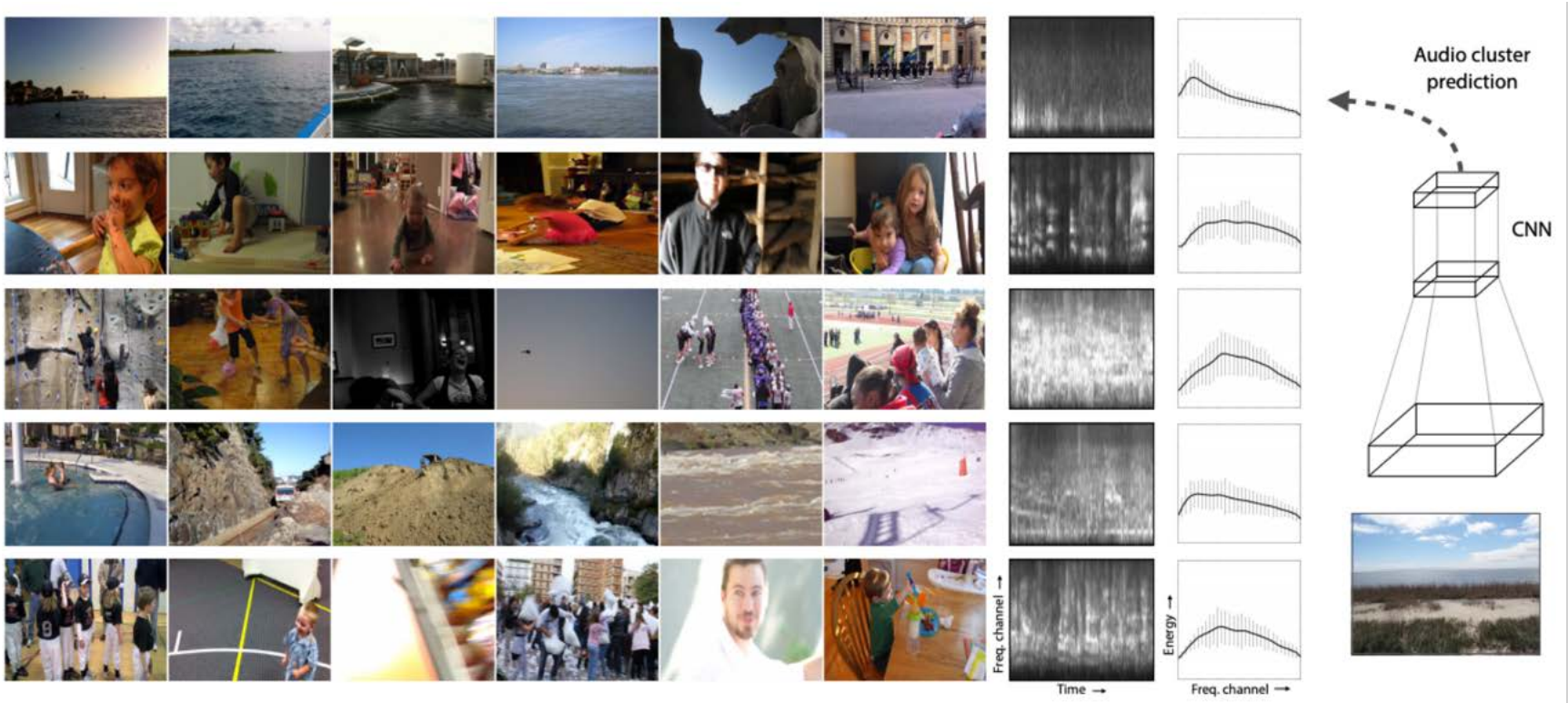
"*If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, Mr. Malik promises to buy Mr. Efros one (1) gelato (2 scoops: one chocolate, one vanilla).*"

https://people.eecs.berkeley.edu/~efros/gelato_bet.html

Self-supervised learning: Outline

- Data prediction
  - Colorization
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Deep clustering and instance prediction
- Contrastive learning
  - PIRL, MoCo, SimCLR, SWaV
- Self-supervision beyond still images
  - Video, audio, language

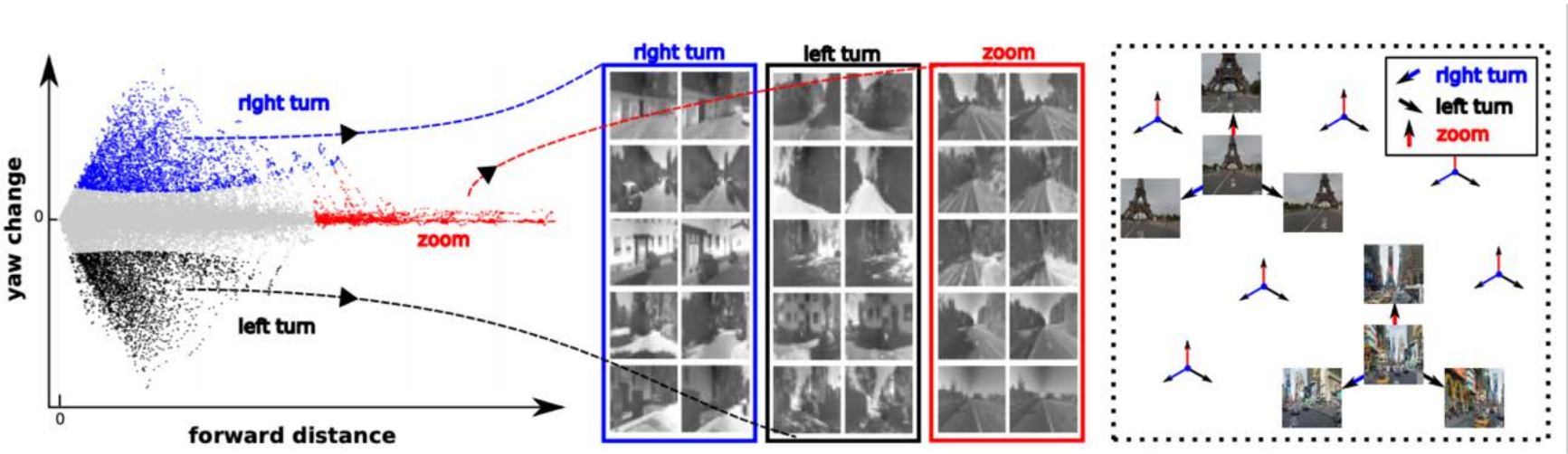
Learning from audio



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

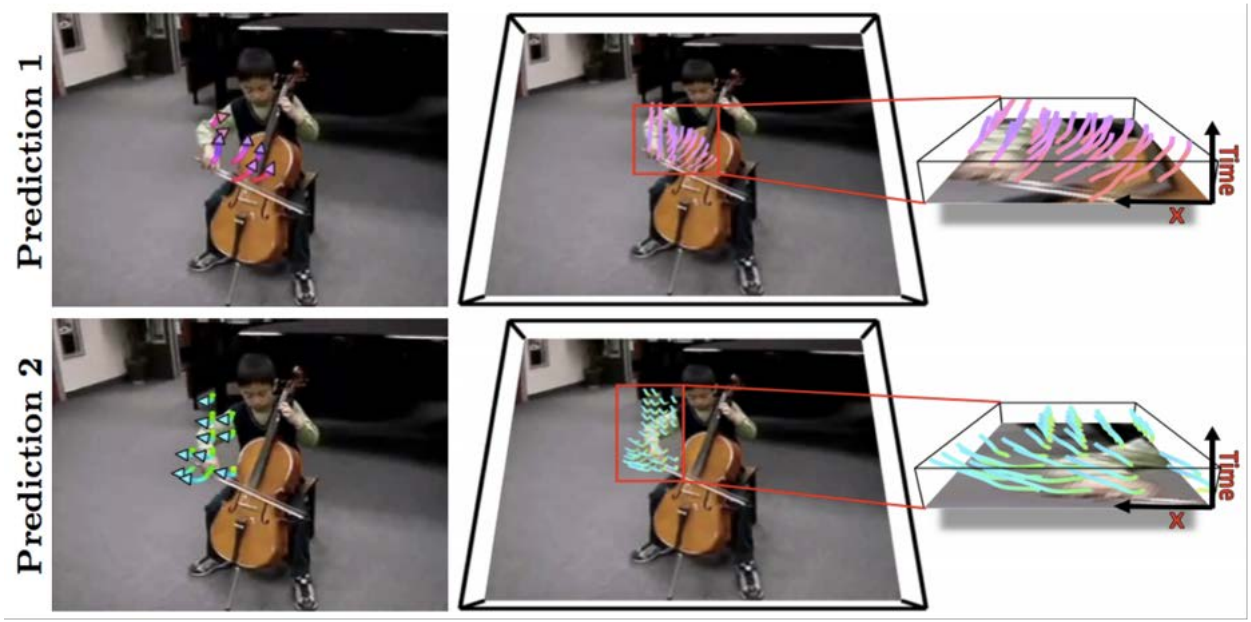
A. Owens et al. Ambient Sound Provides Supervision for Visual Learning. ECCV 2016

Ego-motion features



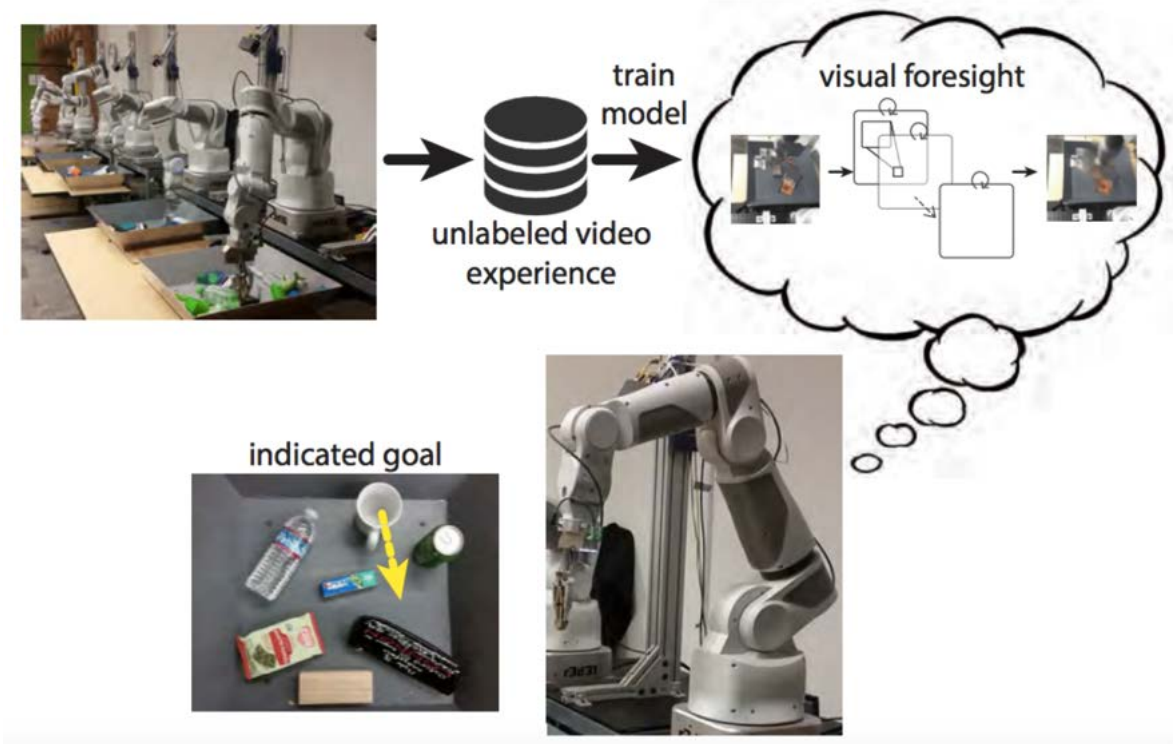
D. Jayaraman and K. Grauman. Learning image representations tied to ego-motion. ICCV 2015

Future prediction



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

Future prediction



C. Finn and S. Levine. Deep Visual Foresight for Planning Robot Motion. ICRA 2017

Self-supervised learning in NLP (coming up)

- word2vec, GloVe, BERT, ELMO, GPT, …

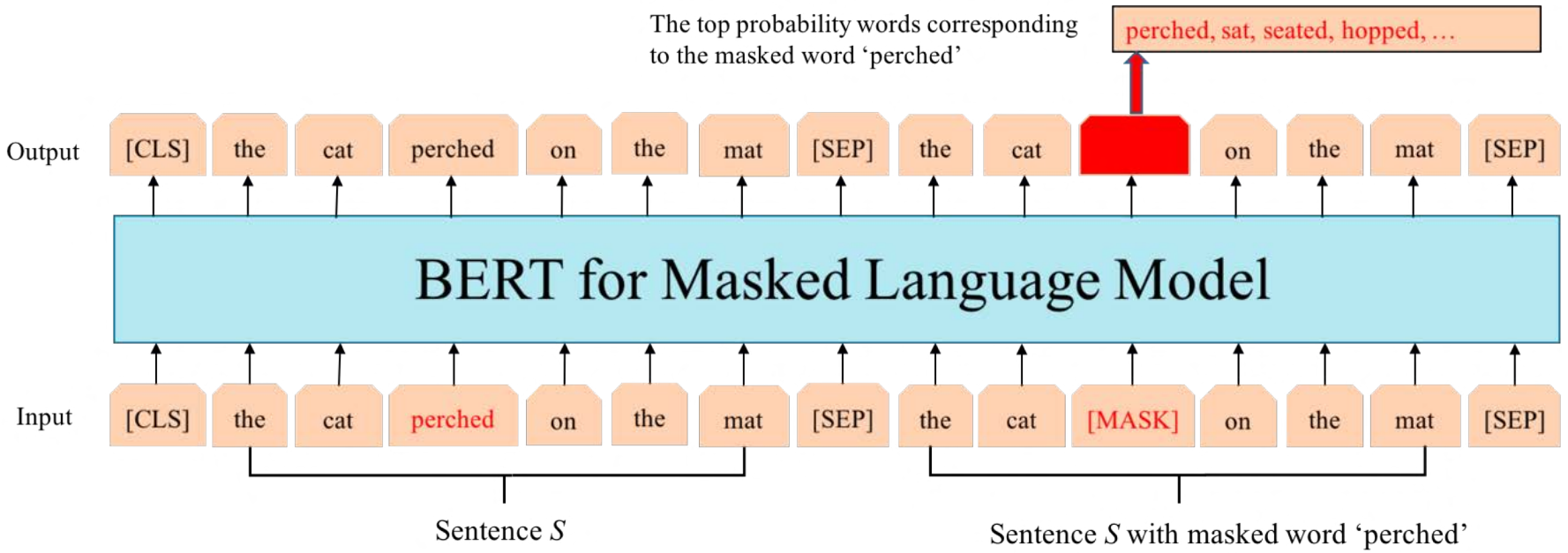


Figure source

For further reading

https://github.com/jason718/awesome-self-supervised-learning