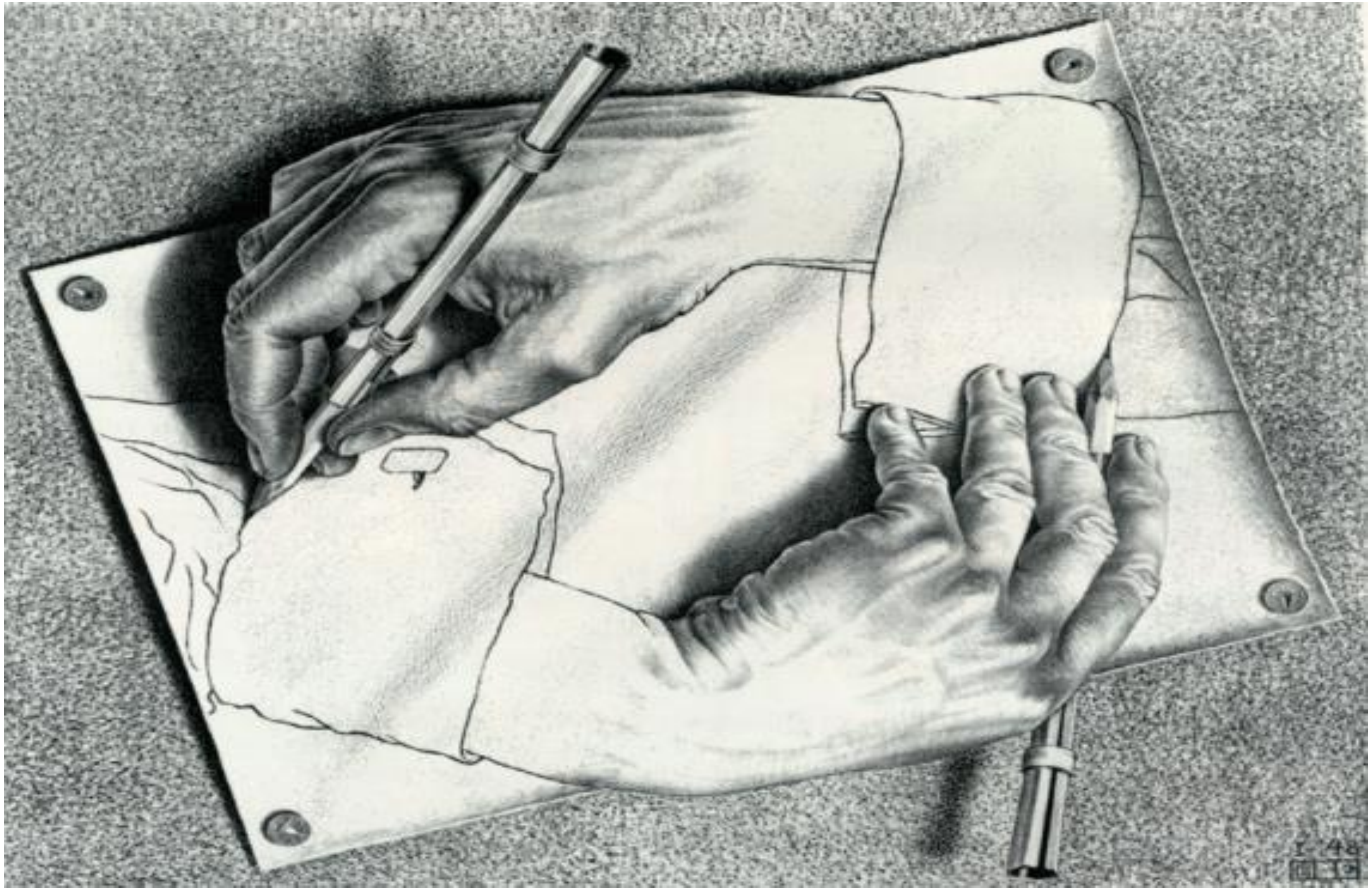
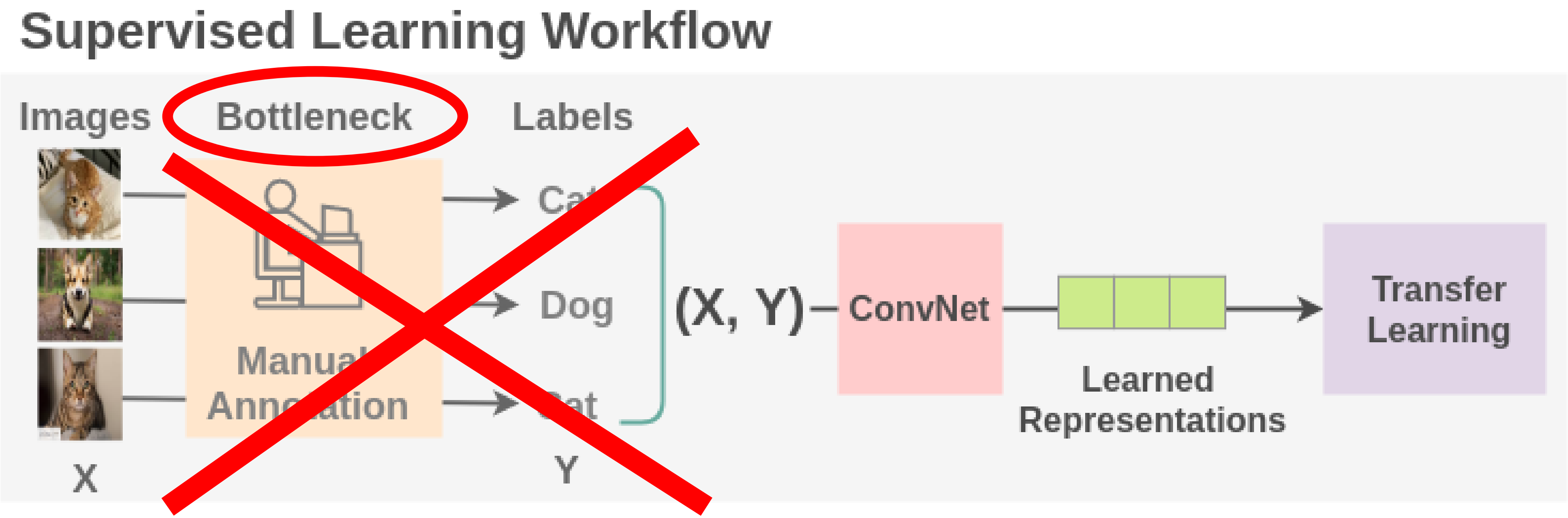
Self-supervised learning



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

Motivation

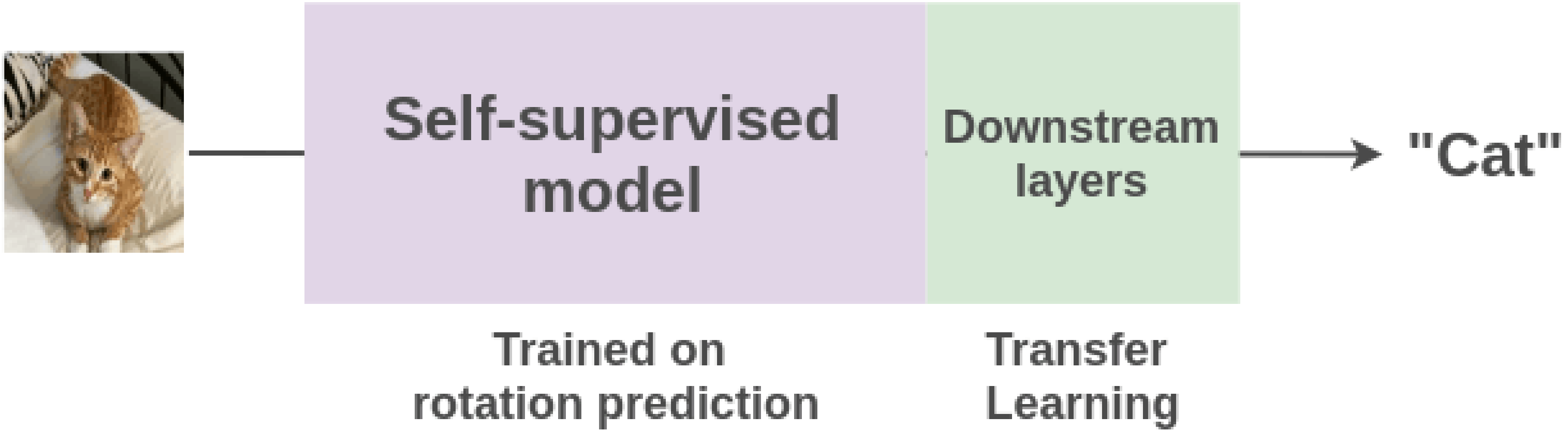
- Overcoming reliance on *supervised pre-training*



Can we design the task in such a way that we can generate virtually unlimited labels from our existing images and use that to learn the representations?

Figure source

Motivation

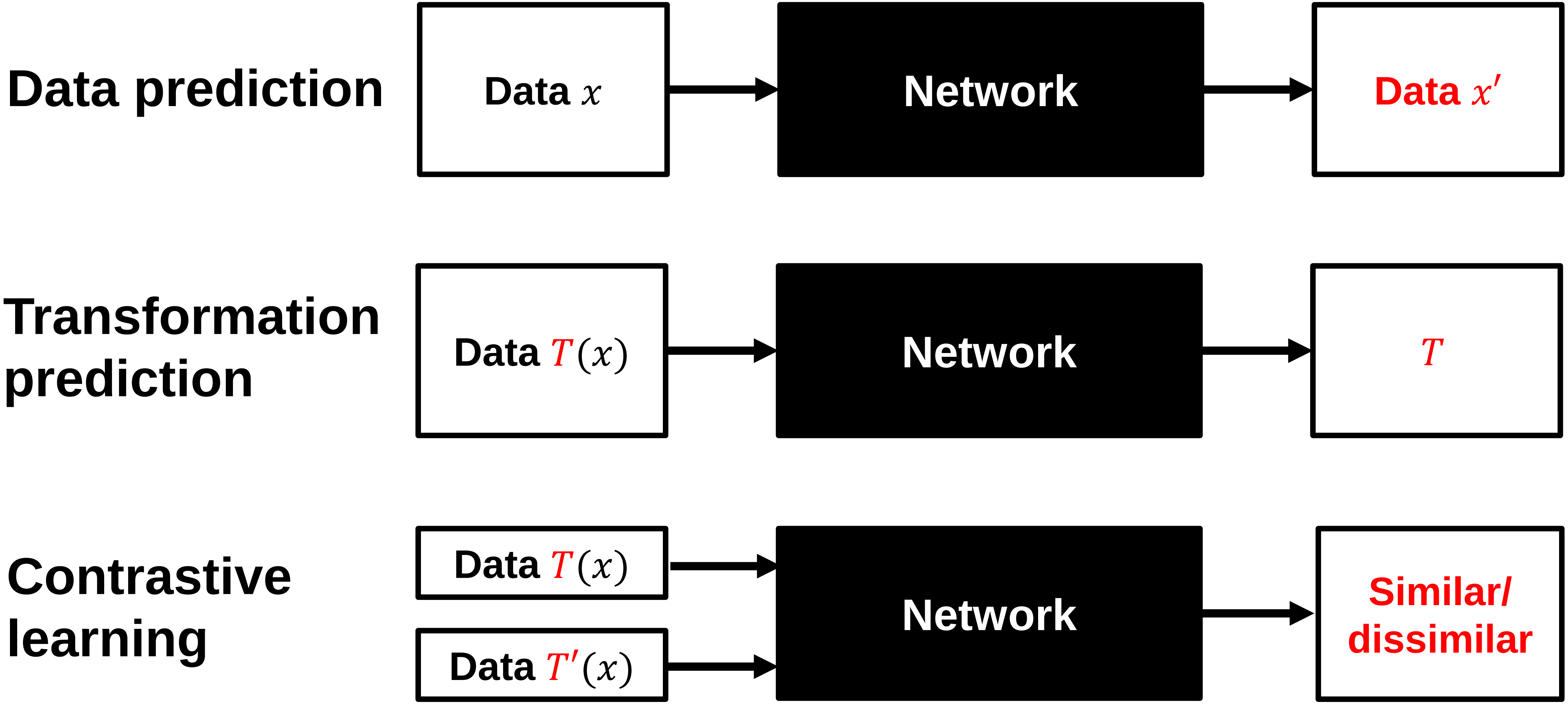


Once we learn representations from these millions of images, we can use transfer learning to fine-tune it on some supervised task like image classification of cats vs dogs with very few examples.

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

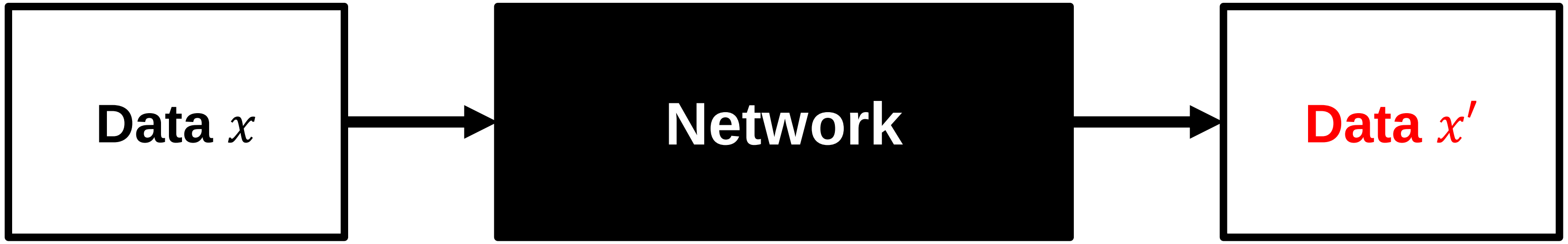
Self-supervised learning: Outline

- Data prediction
  - Colorization, Superresolution, Inpainting, Cross channel encoding
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Automatic Label Generation
  - Image clustering, Synthetic imagery
- Contrastive learning
  - PIRL, MoCo, SimCLR, SWaV
- Self-supervision beyond still images
  - Audio, video, language

Self-supervised learning: Outline

- Data prediction
    - Colorization, Superresolution, Inpainting, Cross channel encoding

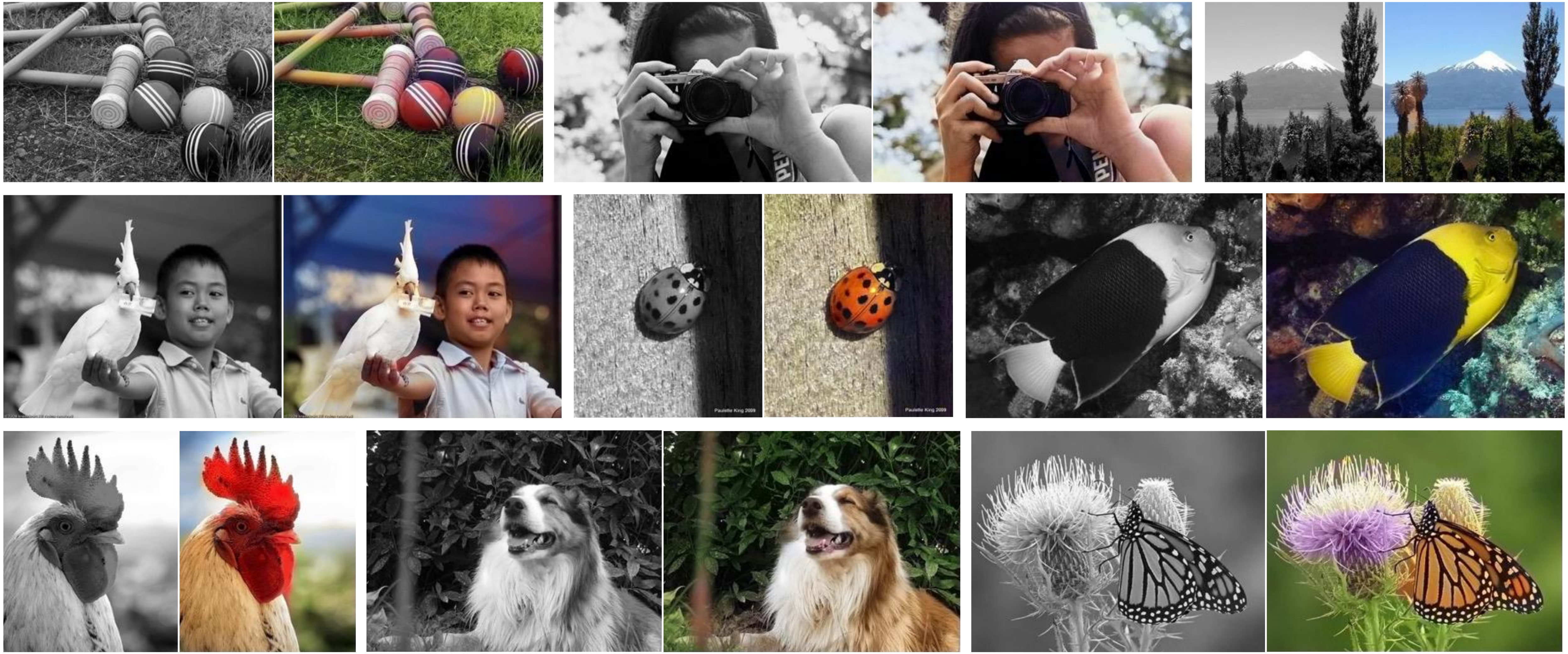
Self-Supervision as data prediction



- Colorization
- Superresolution
- Inpainting
- Cross-channel encoding

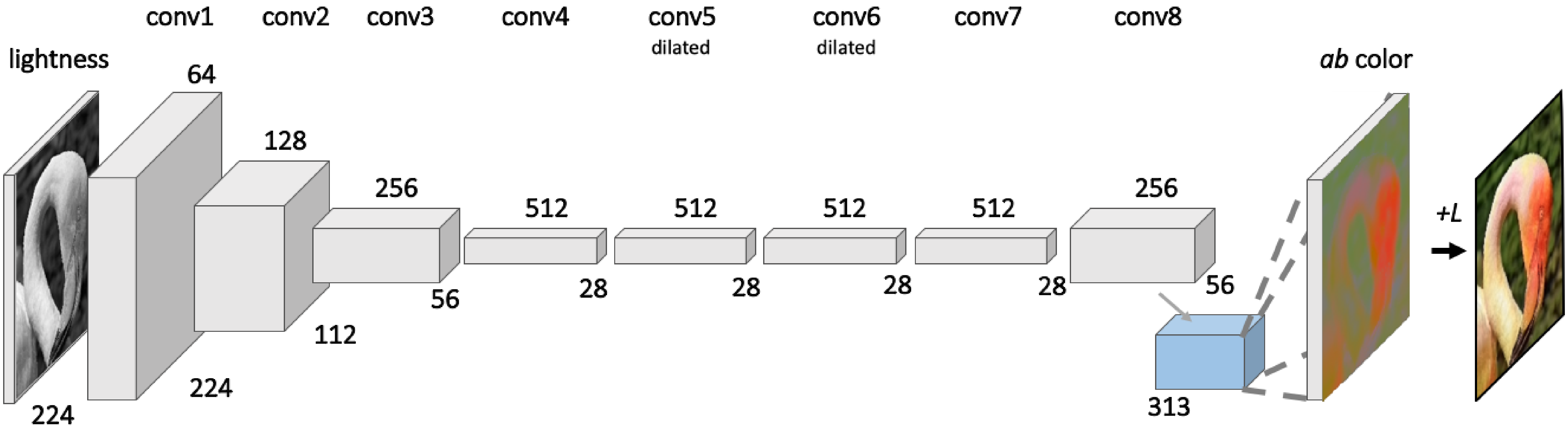
Source: A. Efros

Colorization

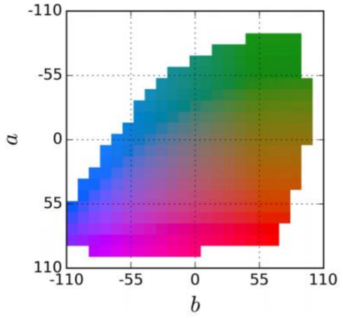


R. Zhang, P. Isola, and A. Efros, [Colorful Image Colorization](), ECCV 2016

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

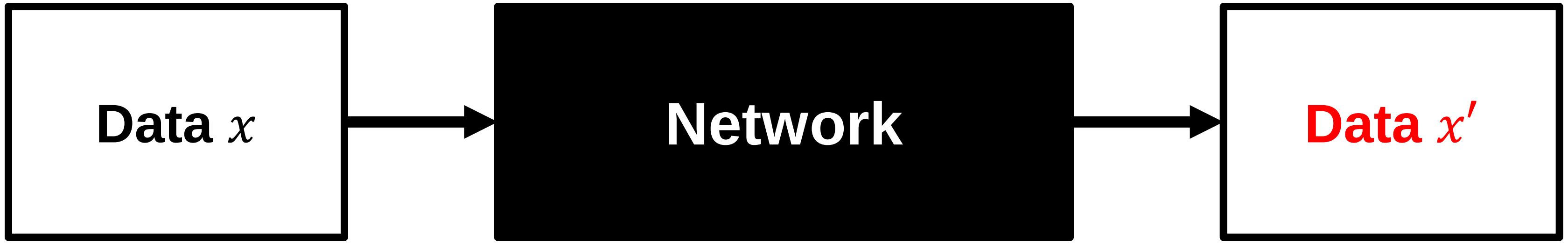
Inherent Ambiguity

Prediction

Ground Truth

Source: A. Efros, R. Zhang

Self-Supervision as data prediction



- Colorization
- Superresolution
- Inpainting
- Cross-channel encoding

Source: A. Efros

Image Superresolution

What if we prepared training pairs of (small, upscaled) images by downsampling millions of images we have freely available?

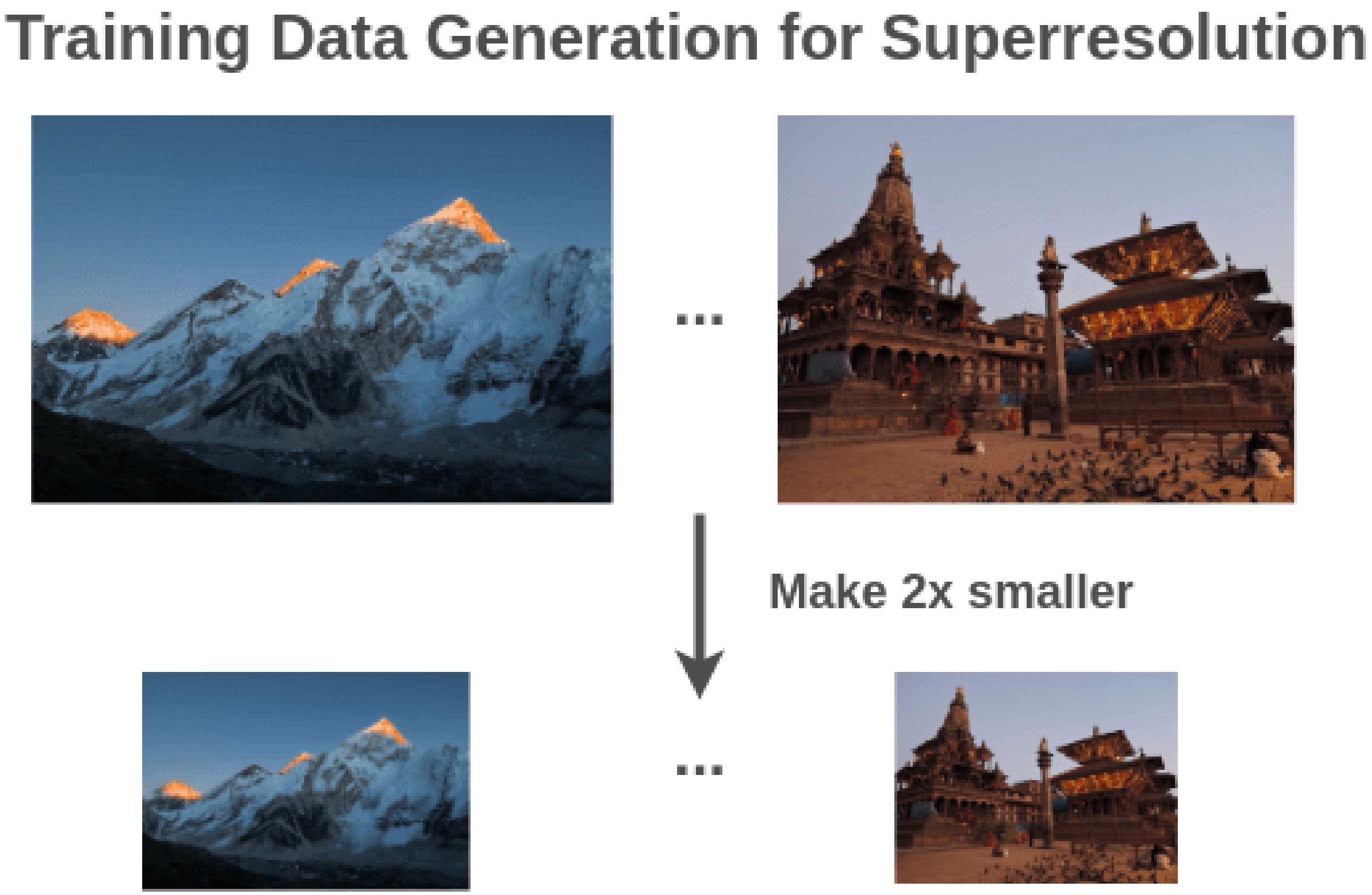


Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network

Image Superresolution

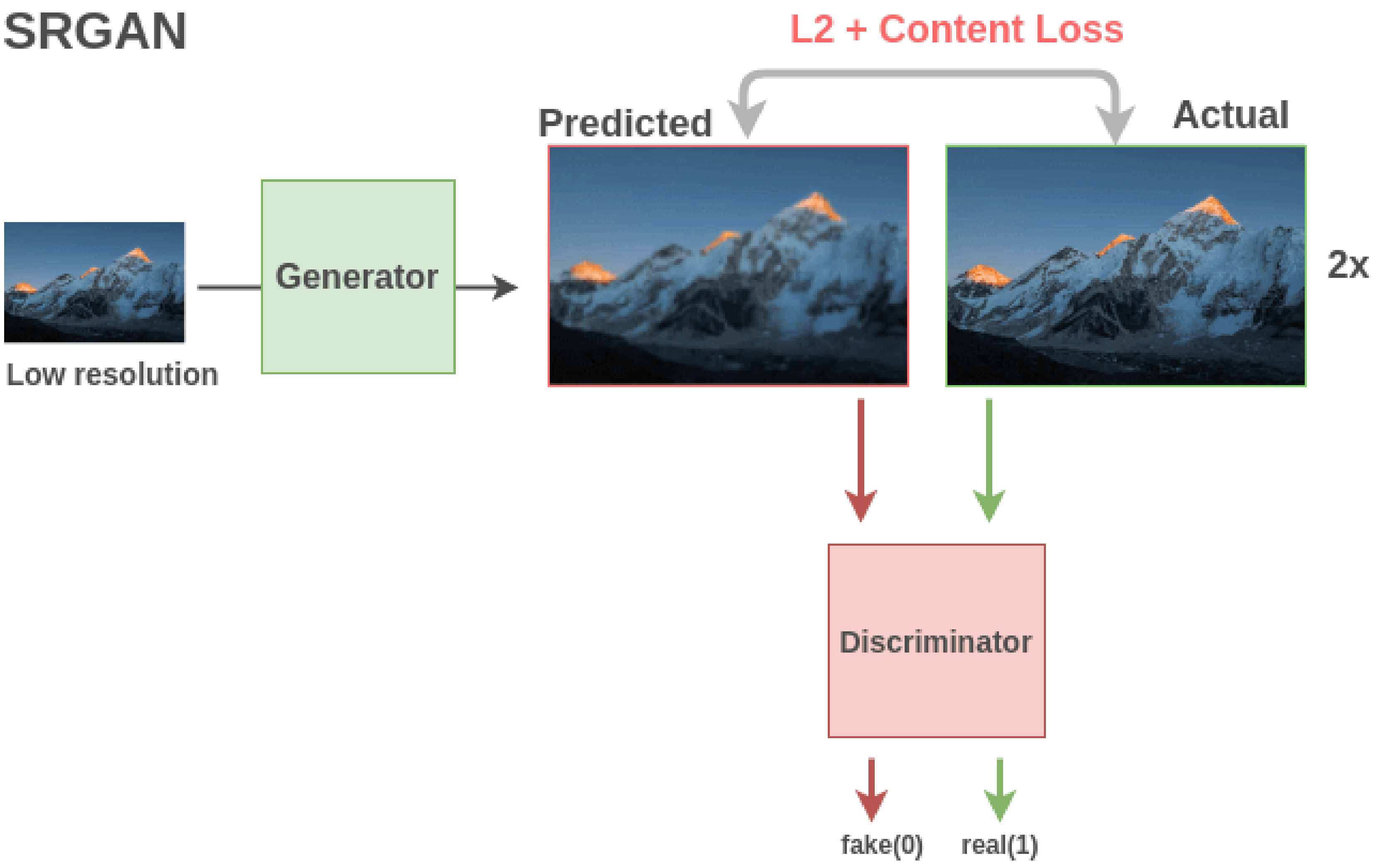
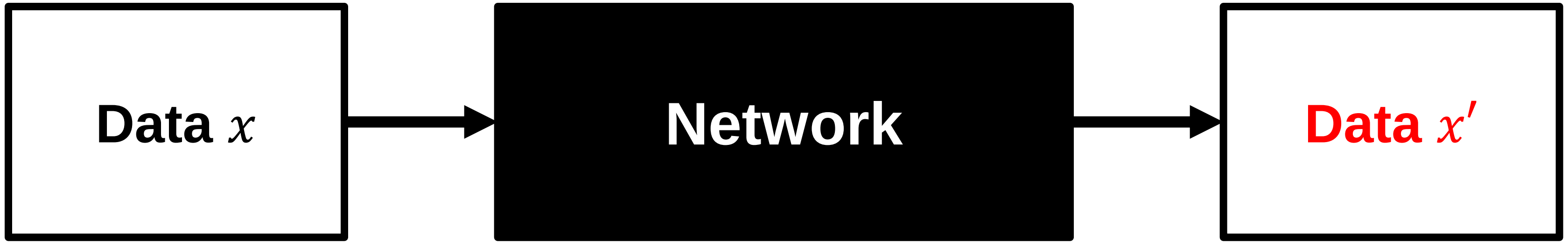


Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network

Self-Supervision as data prediction

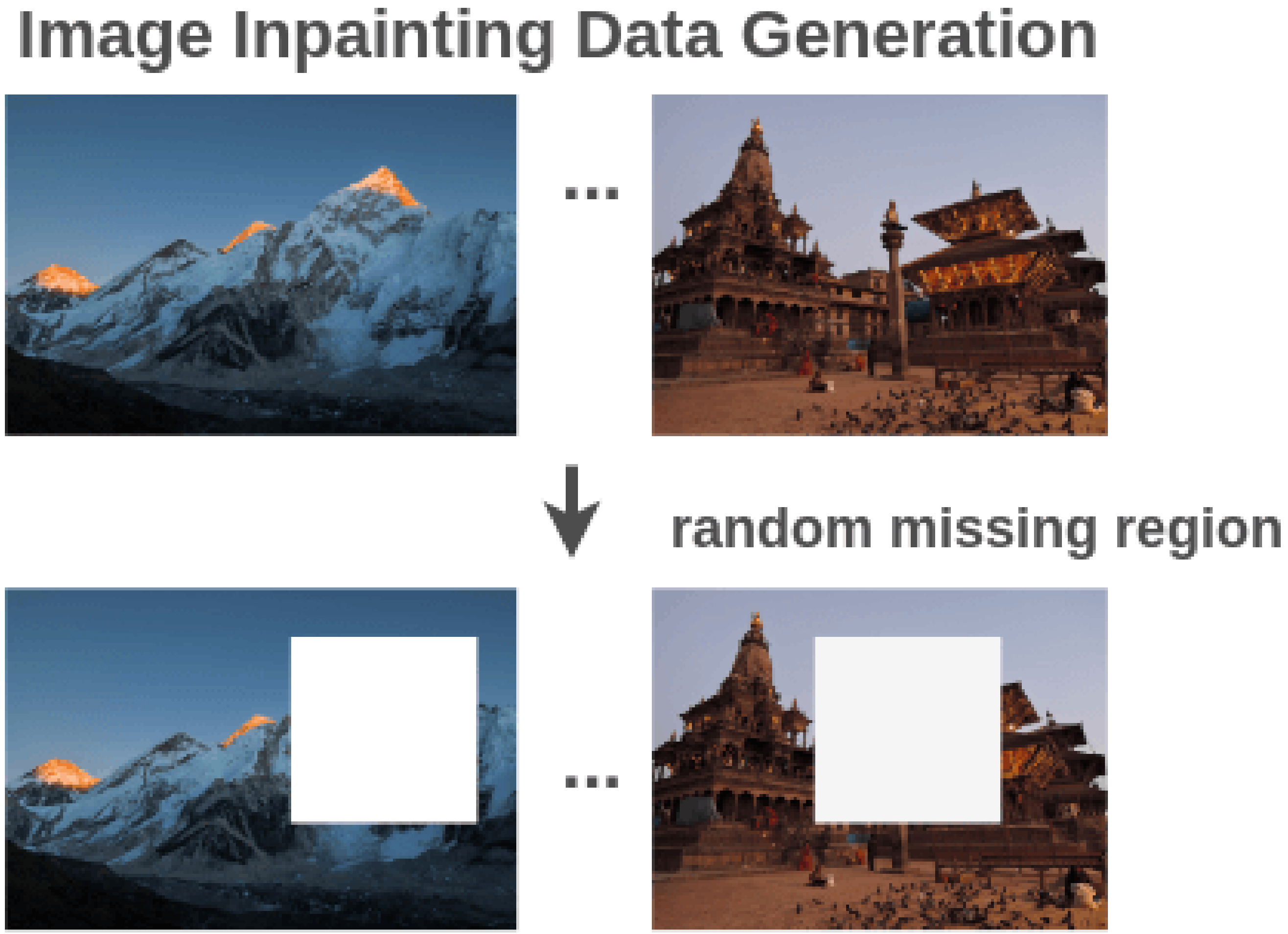


- Colorization
- Superresolution
- Inpainting
- Cross-channel encoding

Source: A. Efros

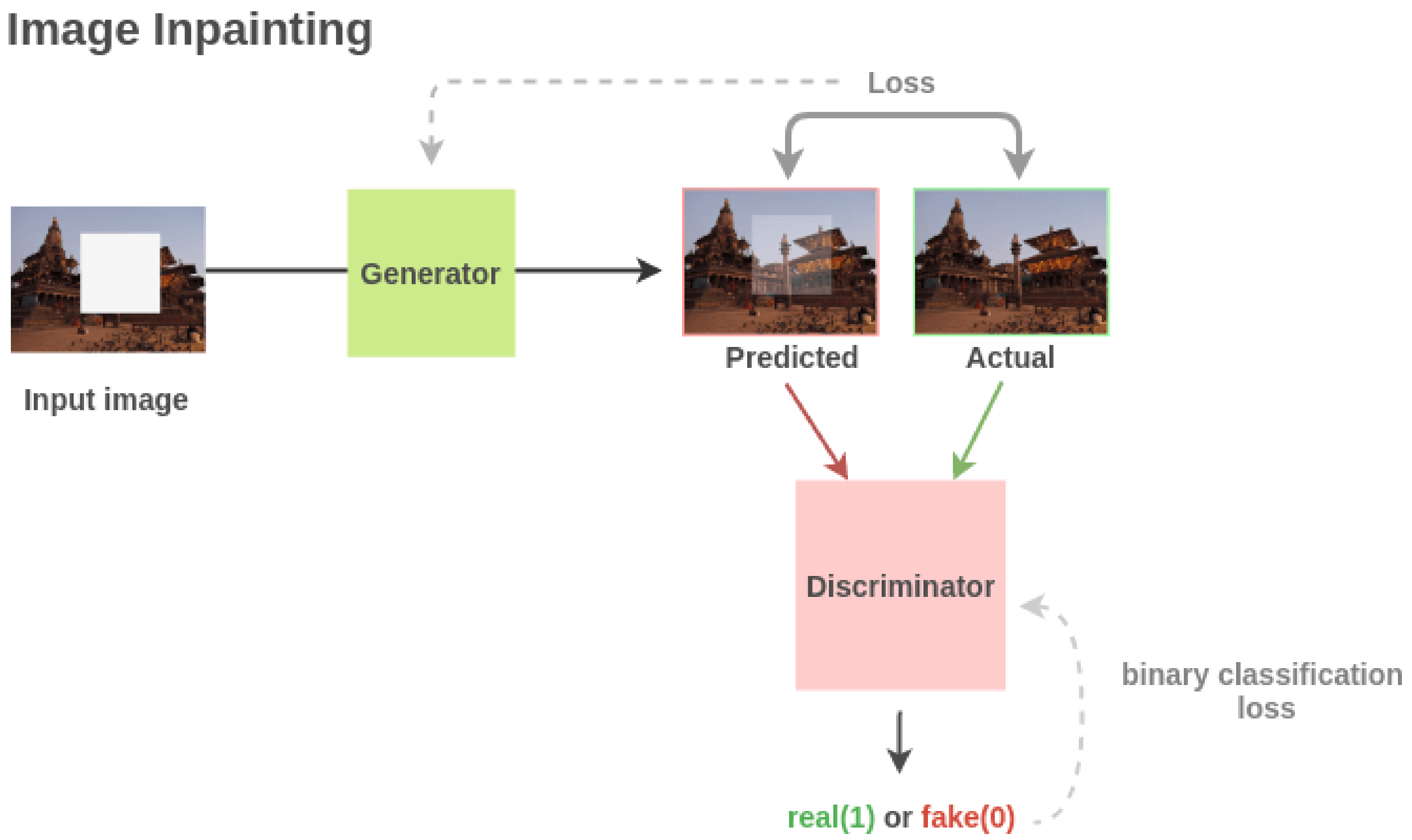
Image Inpainting

What if we prepared training pairs of (corrupted, fixed) images by randomly removing part of images?



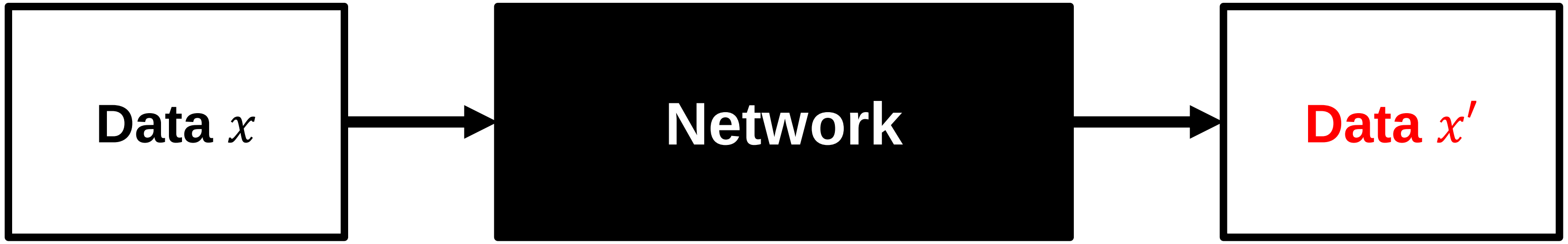
[Context encoders: Feature learning by inpainting]

Image Inpainting



[Context encoders: Feature learning by inpainting]

Self-Supervision as data prediction

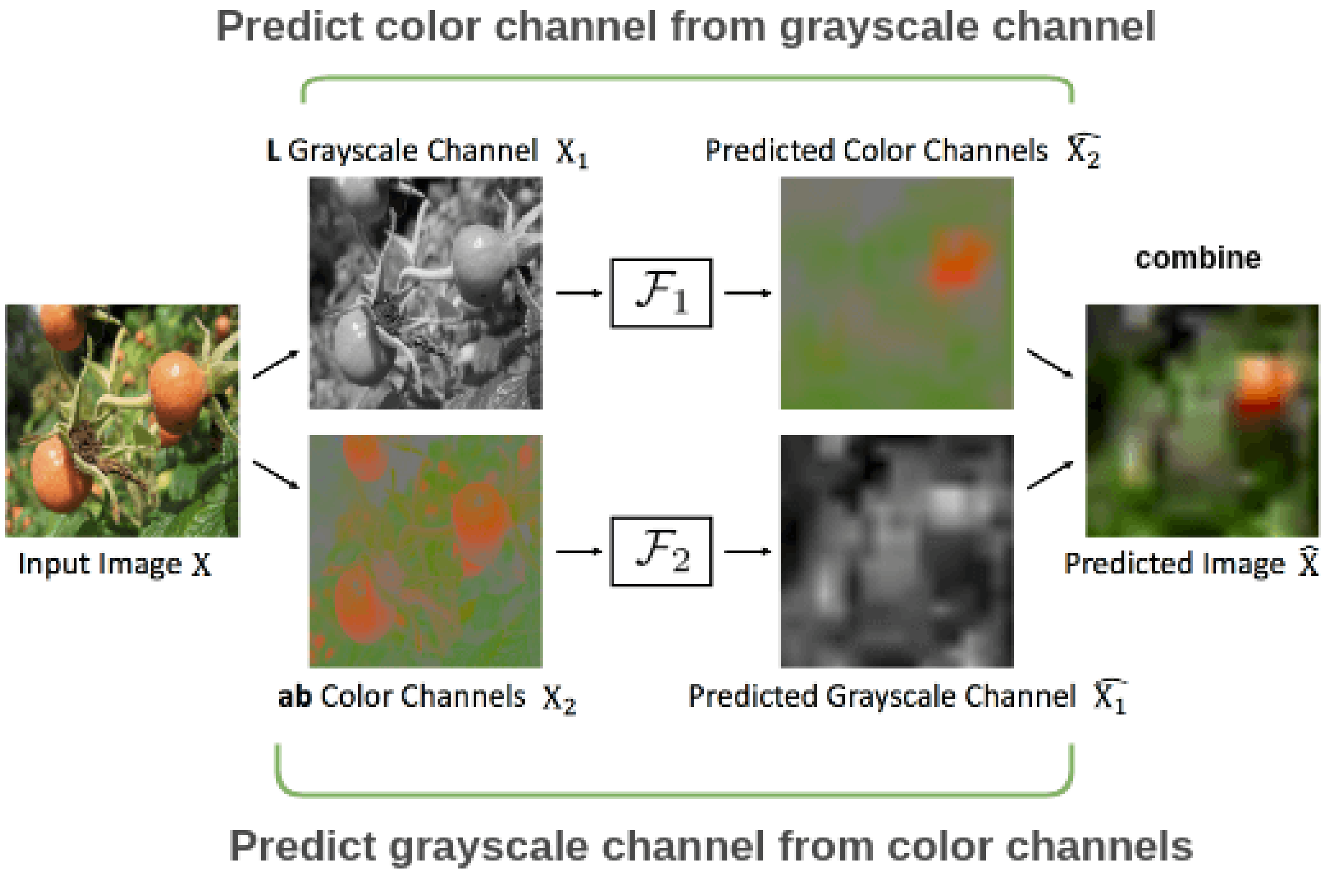


- Colorization
- Superresolution
- Inpainting
- Cross-channel encoding

Source: A. Efros

Cross-channel encoding

What if we predict one channel of the image from the other channel and combine them to reconstruct the original image?

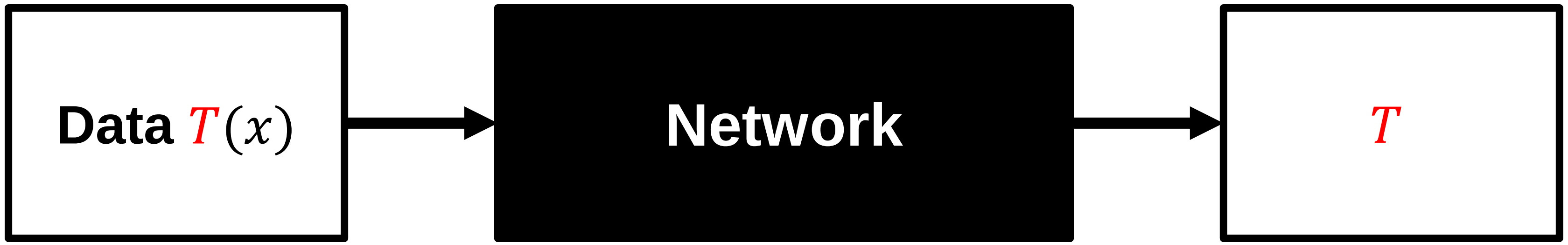


Example adapted from “Split-Brain Autoencoder”

Self-supervised learning: Outline

- Data prediction
  - Colorization, Superresolution, Inpainting, Cross channel encoding
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction

Self-supervision by transformation prediction

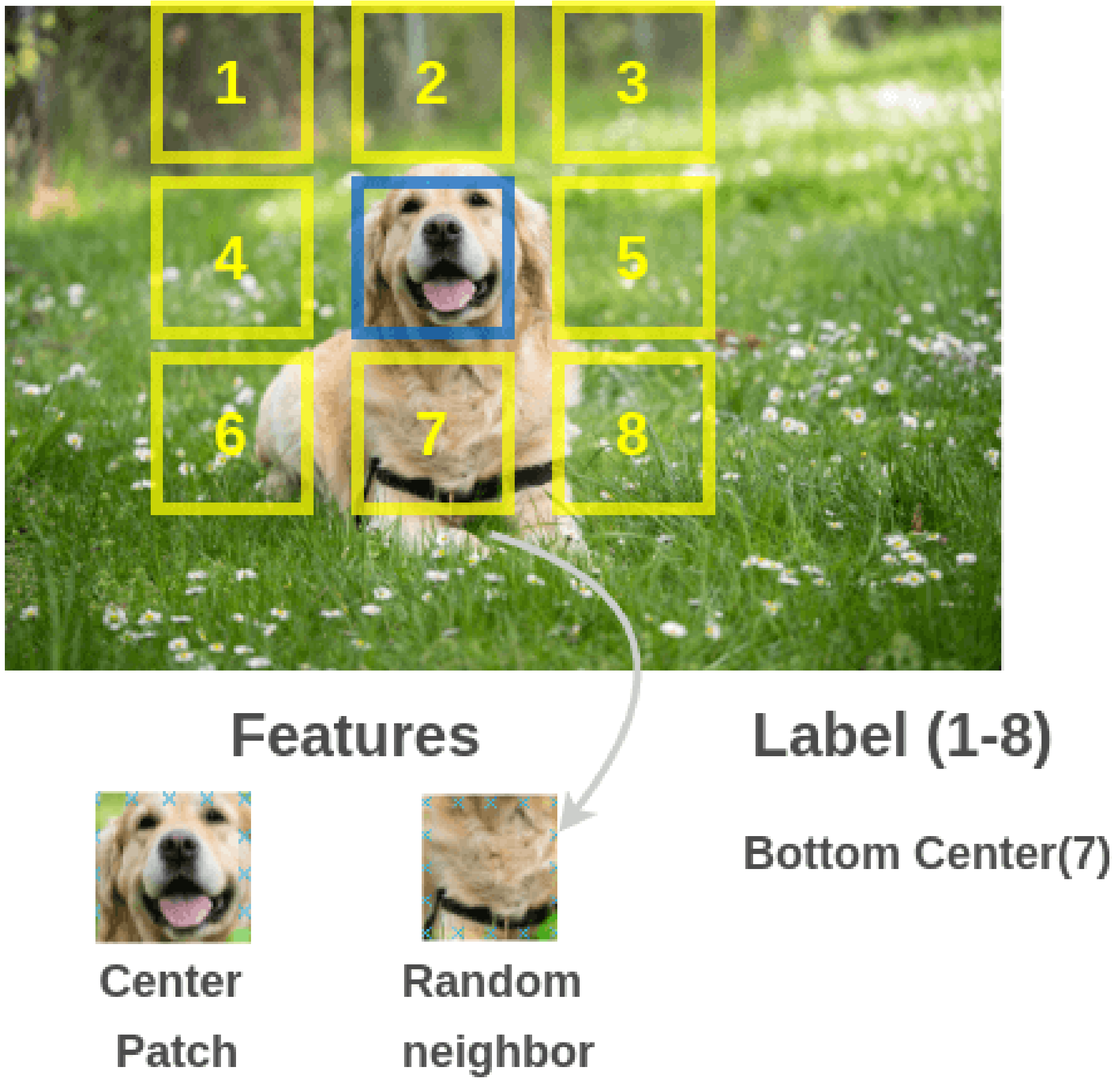


- Context prediction
- Jigsaw puzzle solving
- Rotation prediction

Source: A. Efros

Context prediction

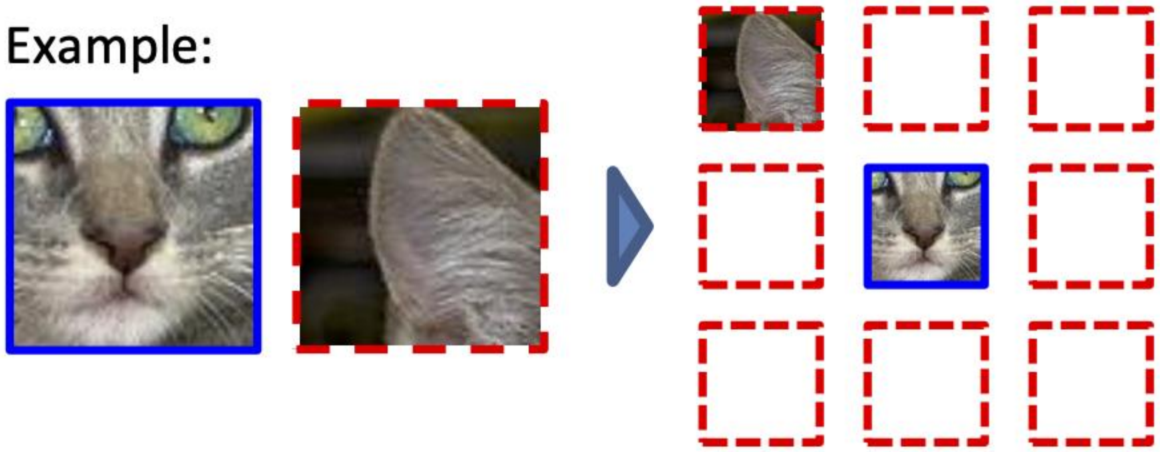
What if we prepared training pairs of (image-patch, neighbor) by randomly taking an image patch and one of its neighbors around it from large, unlabeled image collection?



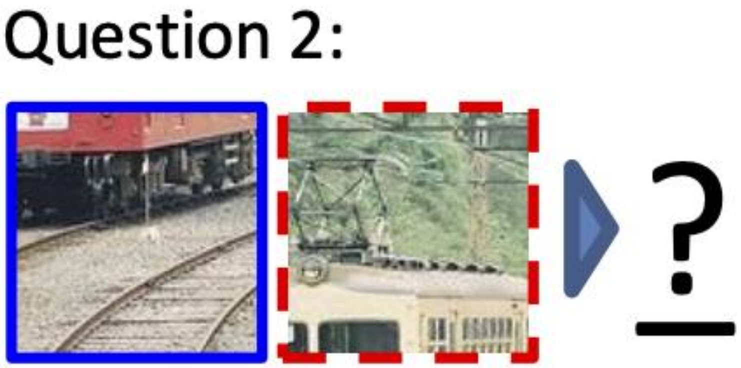
C. Doersch, A. Gupta, A. Efros. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

Context prediction

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



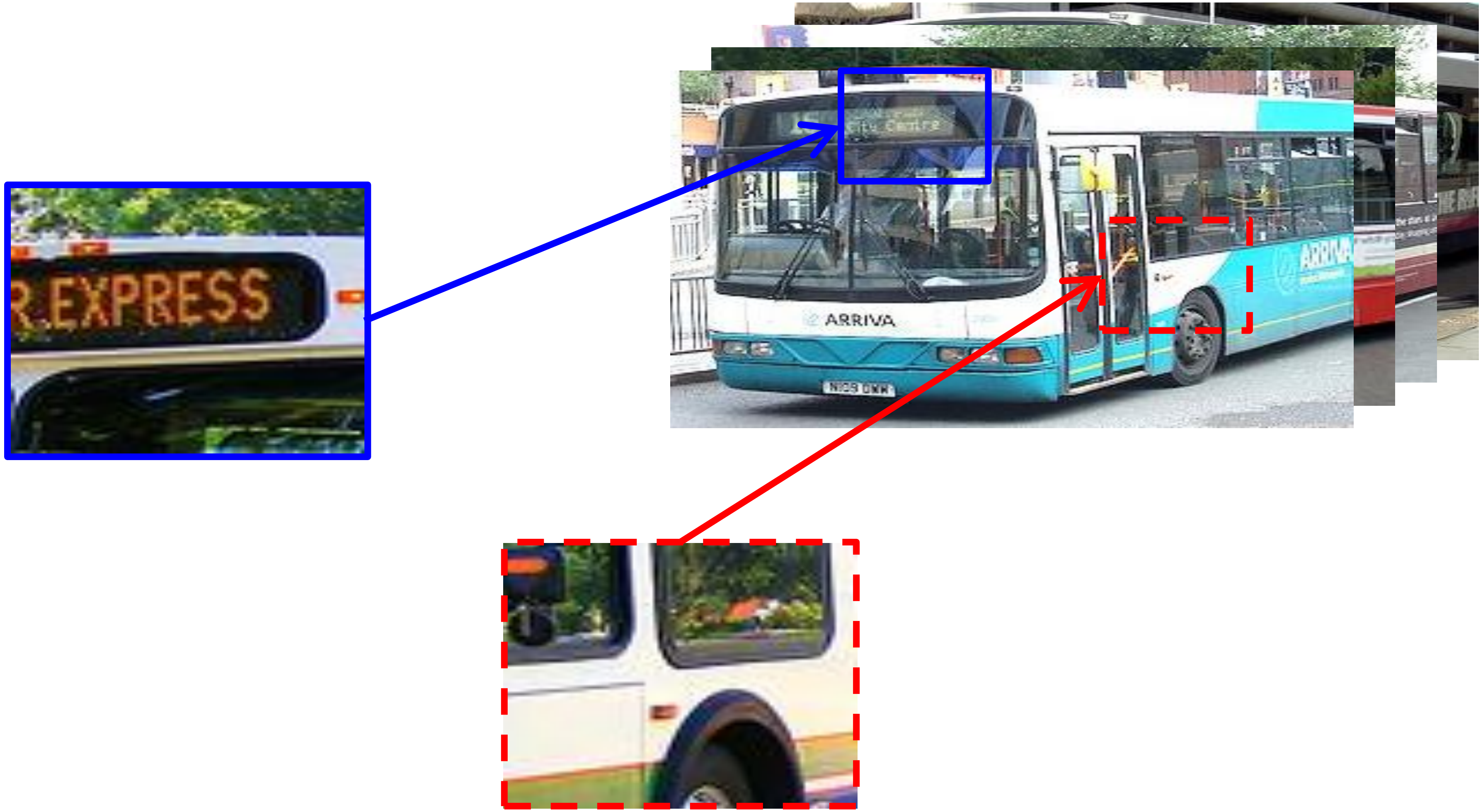
A: Bottom right



A: Top center

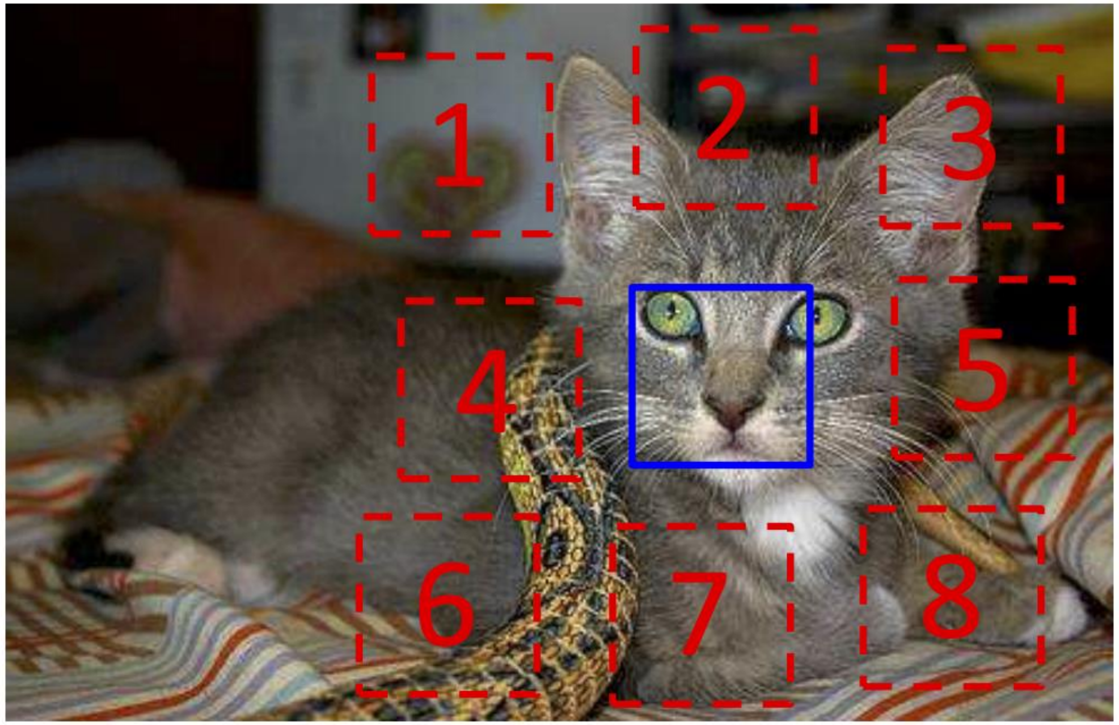
C. Doersch, A. Gupta, A. Efros. Unsupervised Visual Representation Learning by Context Prediction. ICCV 2015

Context prediction: Semantics from a non-semantic task



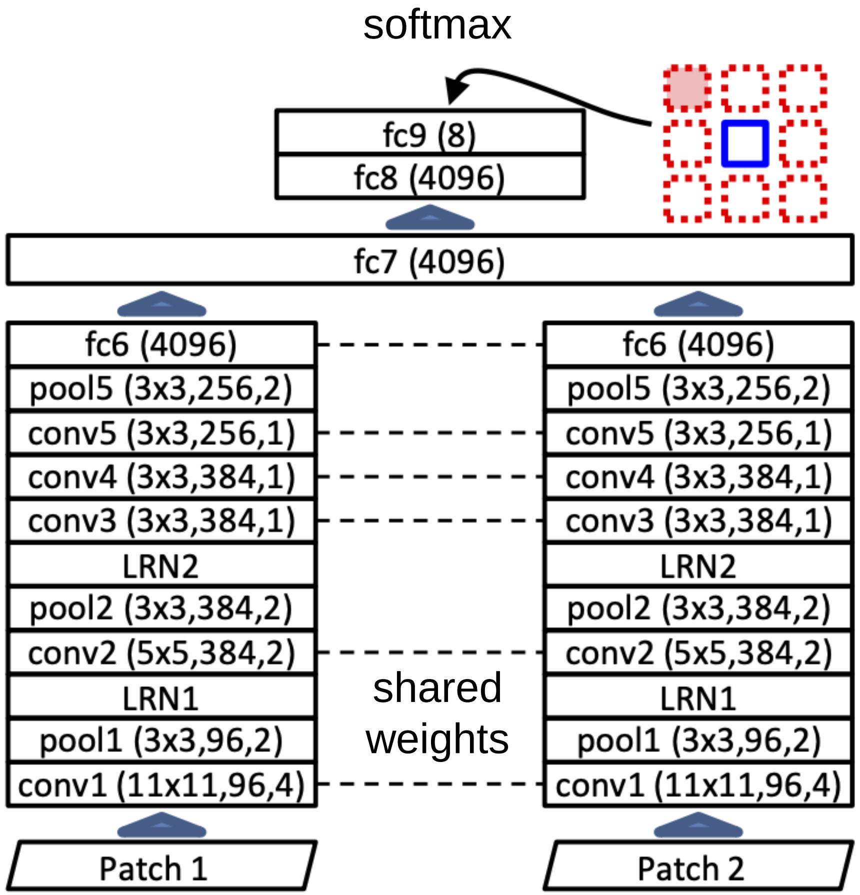
Source: A. Efros

Context prediction: Details



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

AlexNet-like architecture



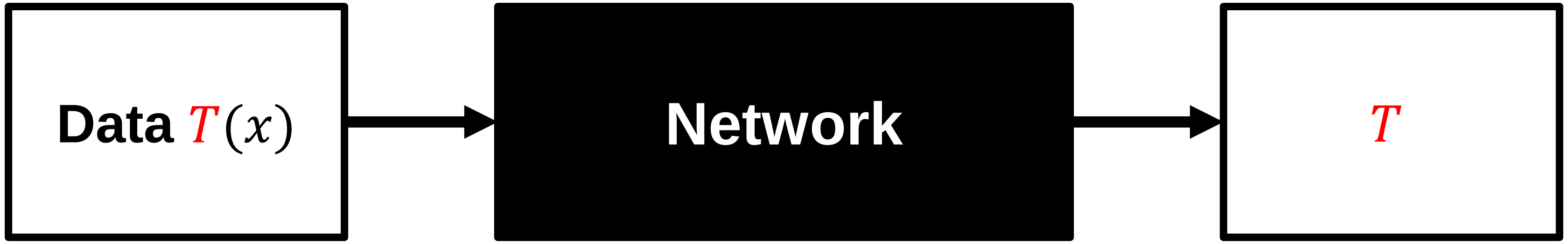
C. Doersch, A. Gupta, A. Efros. [Unsupervised Visual Representation Learning by Context Prediction](). ICCV 2015

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

Self-supervision by transformation prediction



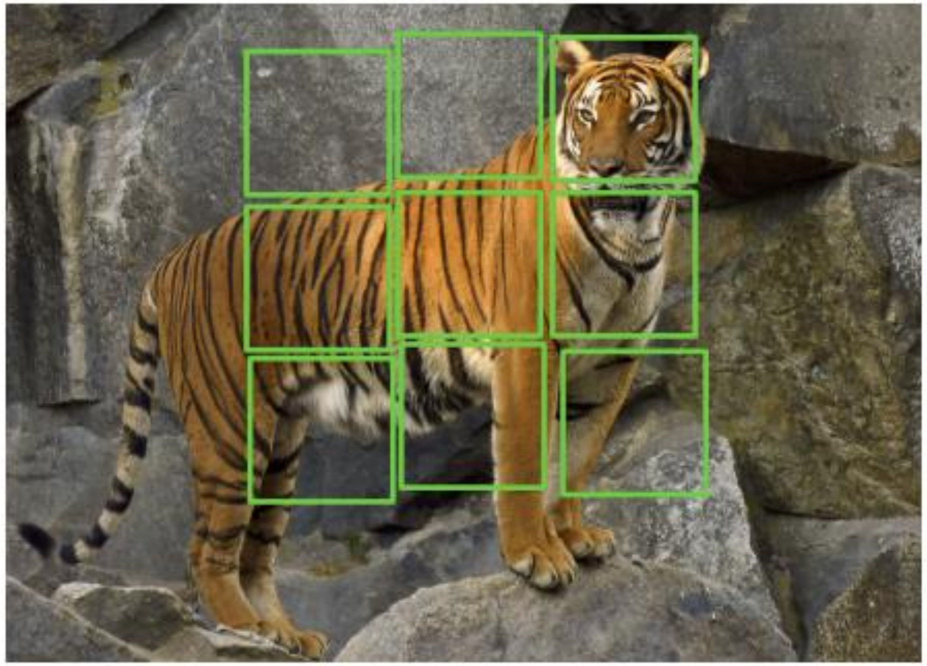
- Context prediction
- Jigsaw puzzle solving
- Rotation prediction

Source: A. Efros

Jigsaw puzzle solving

What if we prepared training pairs of (shuffled, ordered) puzzles by randomly shuffling patches of images?

Crop out tiles



Shuffle

Pretext task: reassemble

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

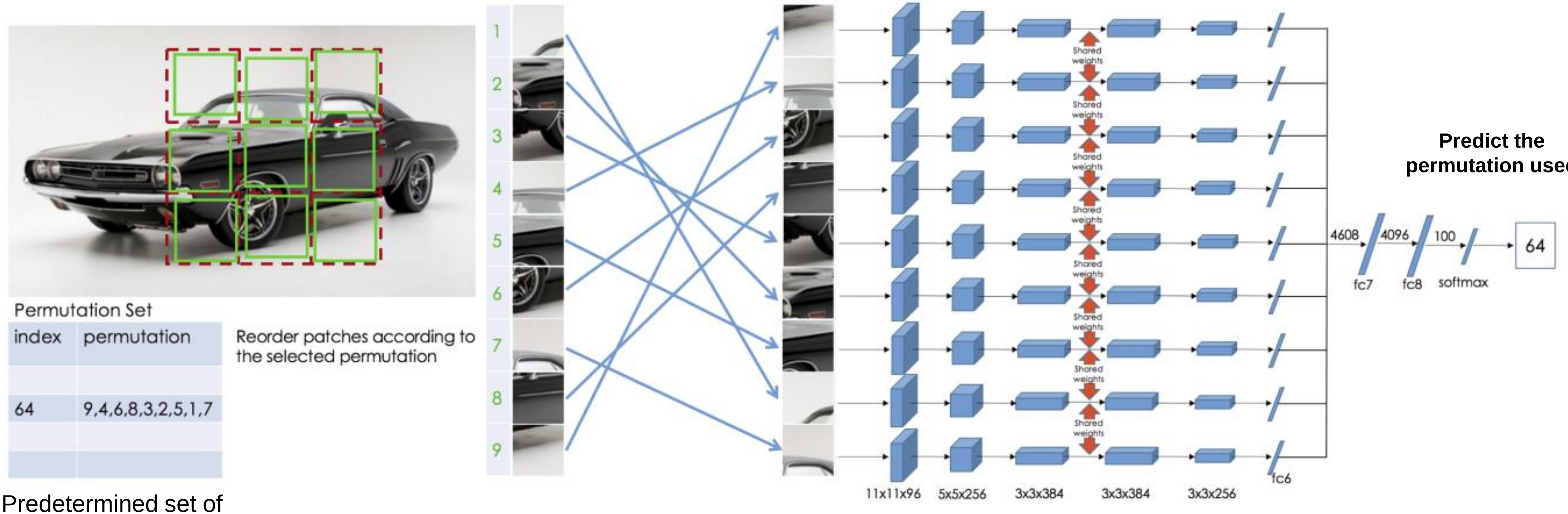
M. Noroozi and P. Favaro. [Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles](). ECCV 2016

Jigsaw puzzle solving: Details

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

Jigsaw puzzle solving: Details

Context free network (CFN)

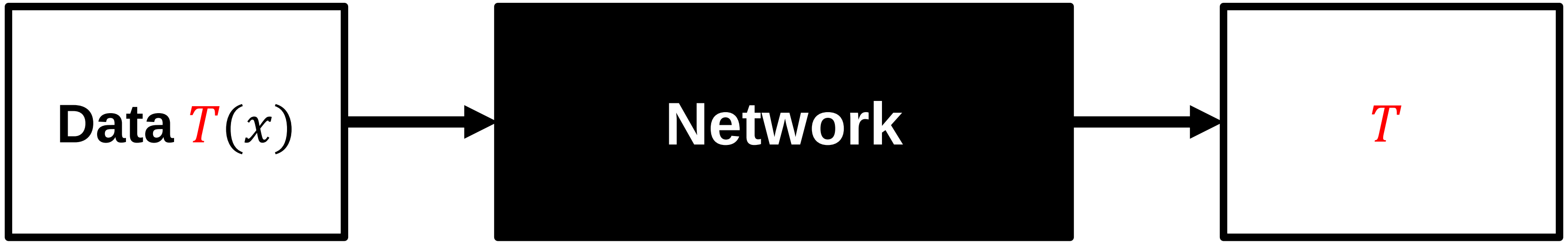


Predict the permutation used

Predetermined set of 100 permutations (out of 362,880 possible)

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

Self-supervision by transformation prediction



- Context prediction
- Jigsaw puzzle solving
- Rotation prediction

Source: A. Efros

Rotation prediction

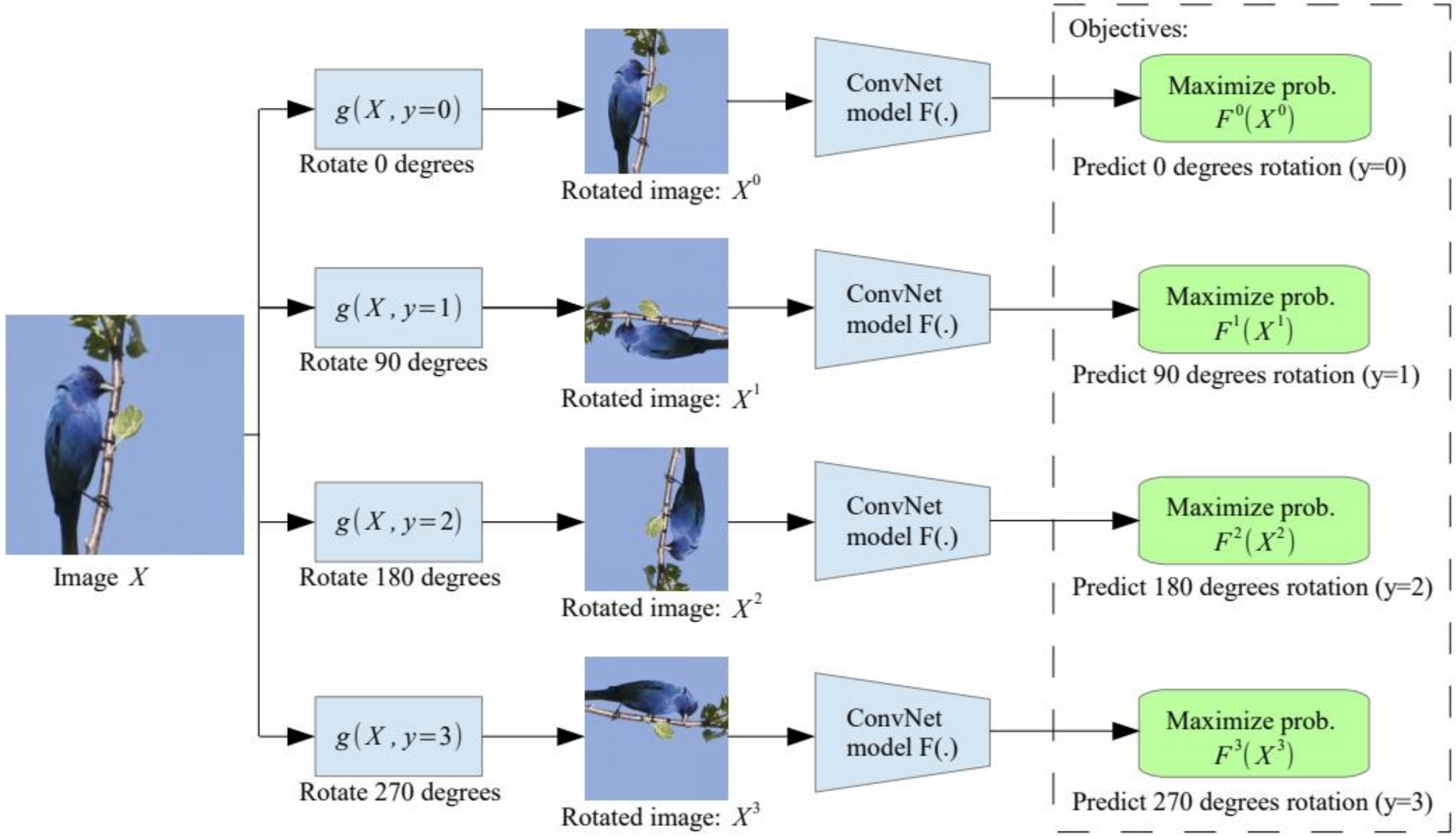
What if we prepared training pairs of (rotated-image, rotation-angle) by randomly rotating images by (0, 90, 180, 270) from large, unlabeled image collection?

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

90° rotation | 270° rotation | 180° rotation | 0° rotation | 270° rotation

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

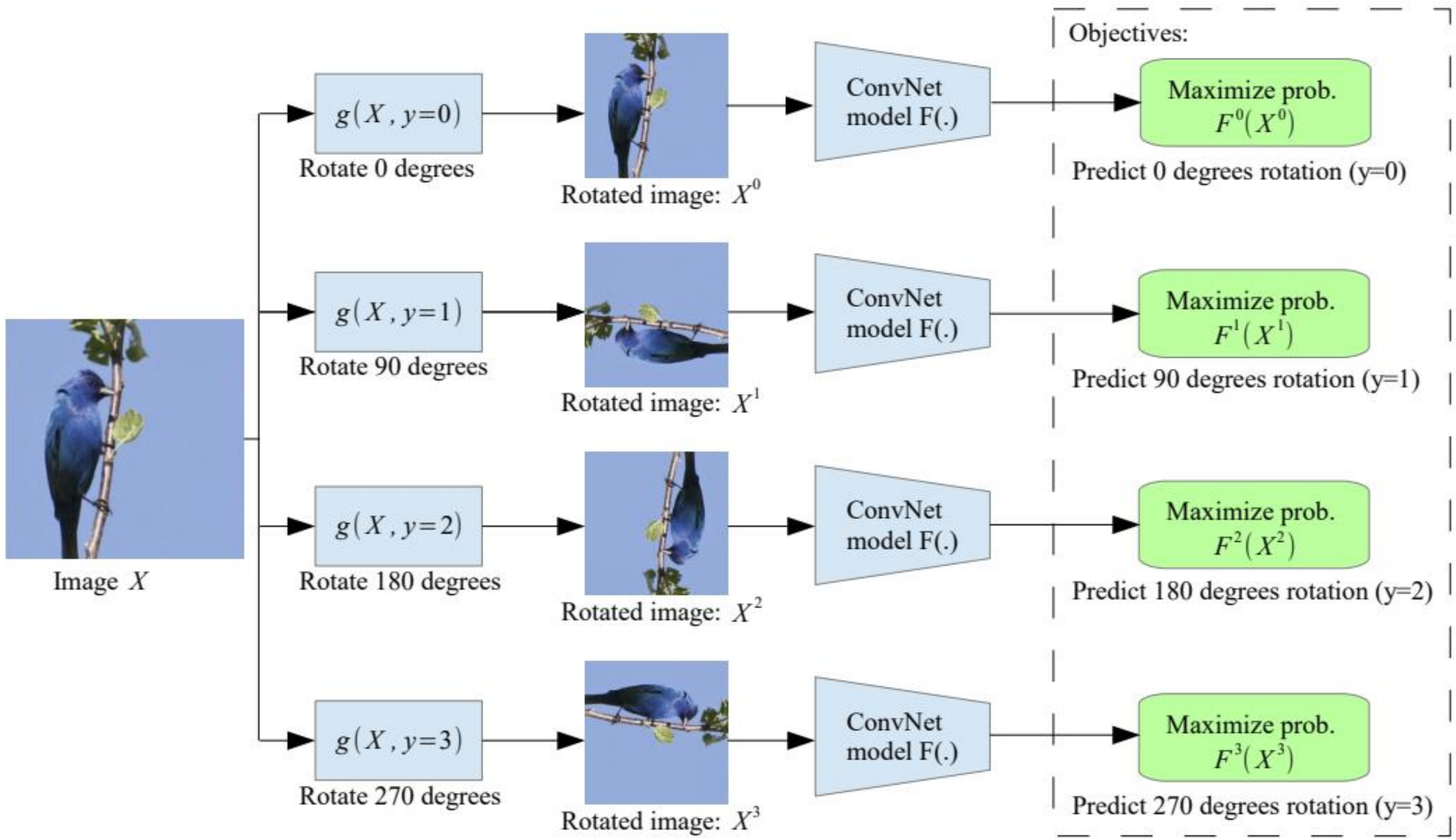
Rotation prediction



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



Though a very simple idea, the model has to understand location, types and pose of objects in an image to solve this task and as such, the representations learned are useful for downstream tasks.

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

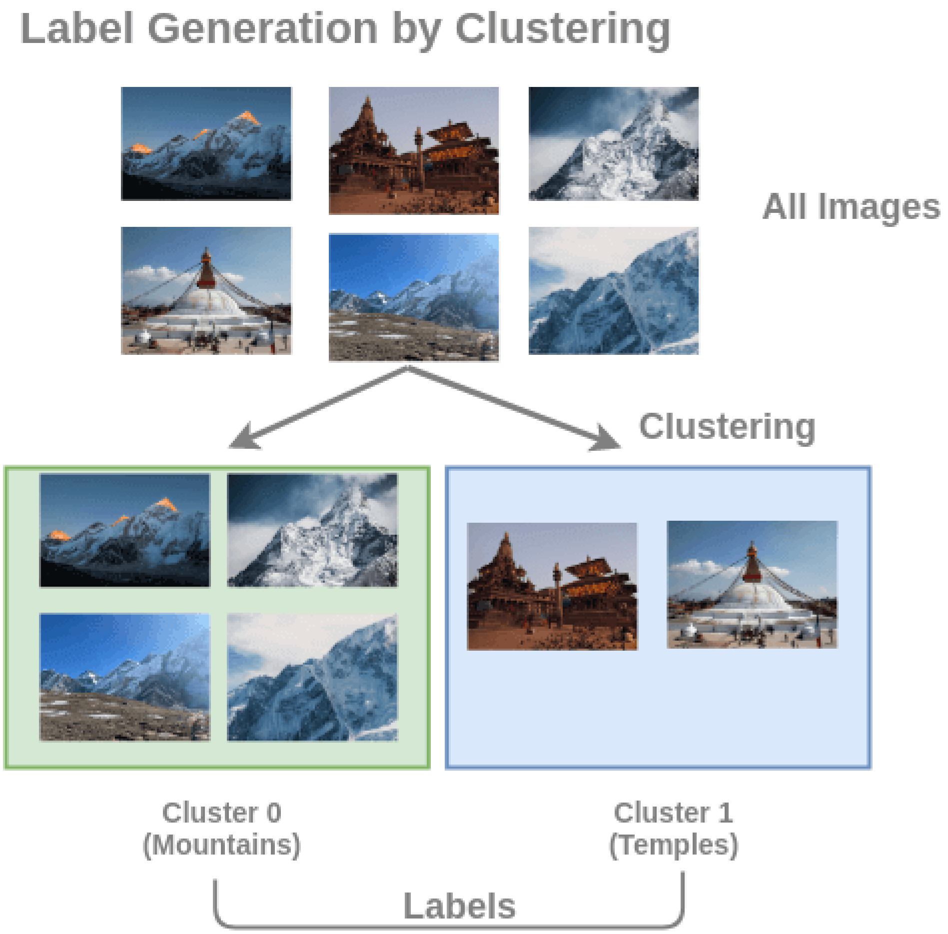
Method	Classification	Detection (mAP)	Segmentation (mIoU)
Supervised (ImageNet)	79.9	56.8	48.0
Colorization	65.6	46.9	35.6
Context	65.3	51.1	
Jigsaw	67.6	53.2	37.6
Rotation	73.0	54.4	39.1

Self-supervised learning: Outline

- Data prediction
  - Colorization, Superresolution, Inpainting, Cross channel encoding
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Automatic Label Generation
  - Image clustering, Synthetic imagery

Image Clustering

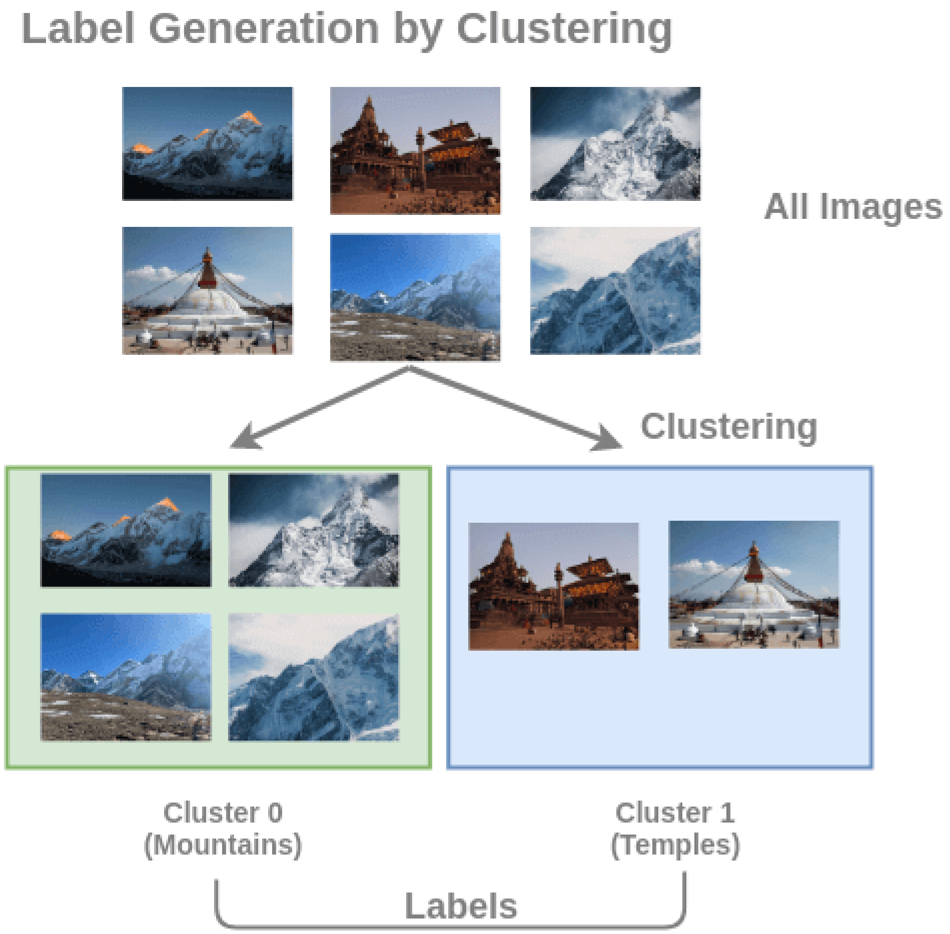
What if we prepared training pairs of (image, cluster-number) by performing clustering on large, unlabeled image collection?

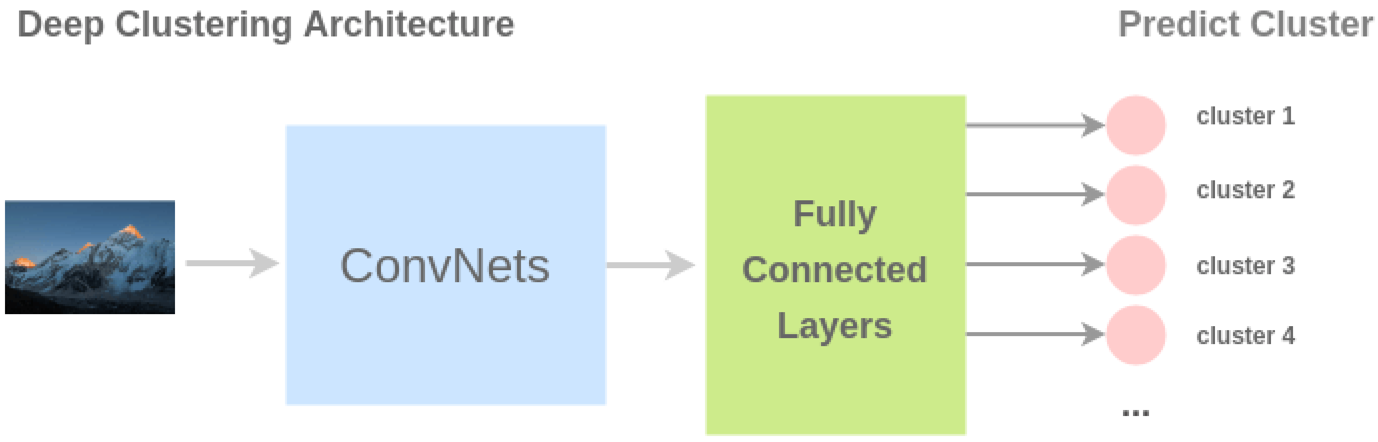


[Deep clustering for unsupervised learning of visual features]()
[Self-labelling via simultaneous clustering and representation learning]()
[CliqueCNN: Deep Unsupervised Exemplar Learning]()

Image Clustering

What if we prepared training pairs of (image, cluster-number) by performing clustering on large, unlabeled image collection?

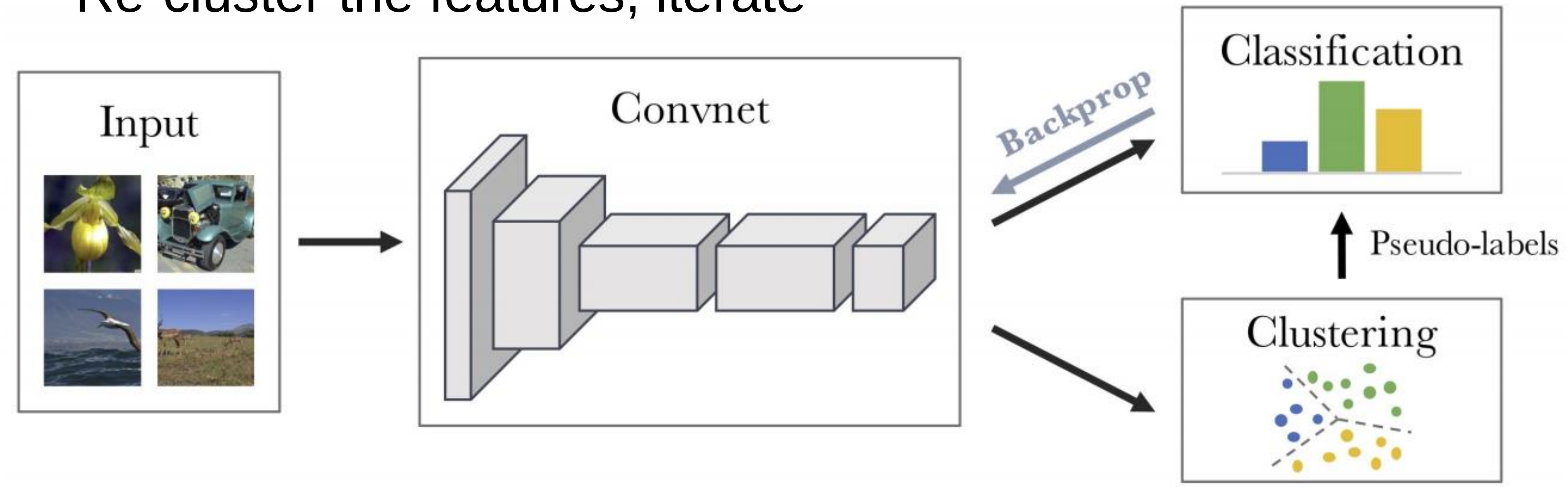




[Deep clustering for unsupervised learning of visual features]
[Self-labelling via simultaneous clustering and representation learning]
[CliqueCNN: Deep Unsupervised Exemplar Learning]

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

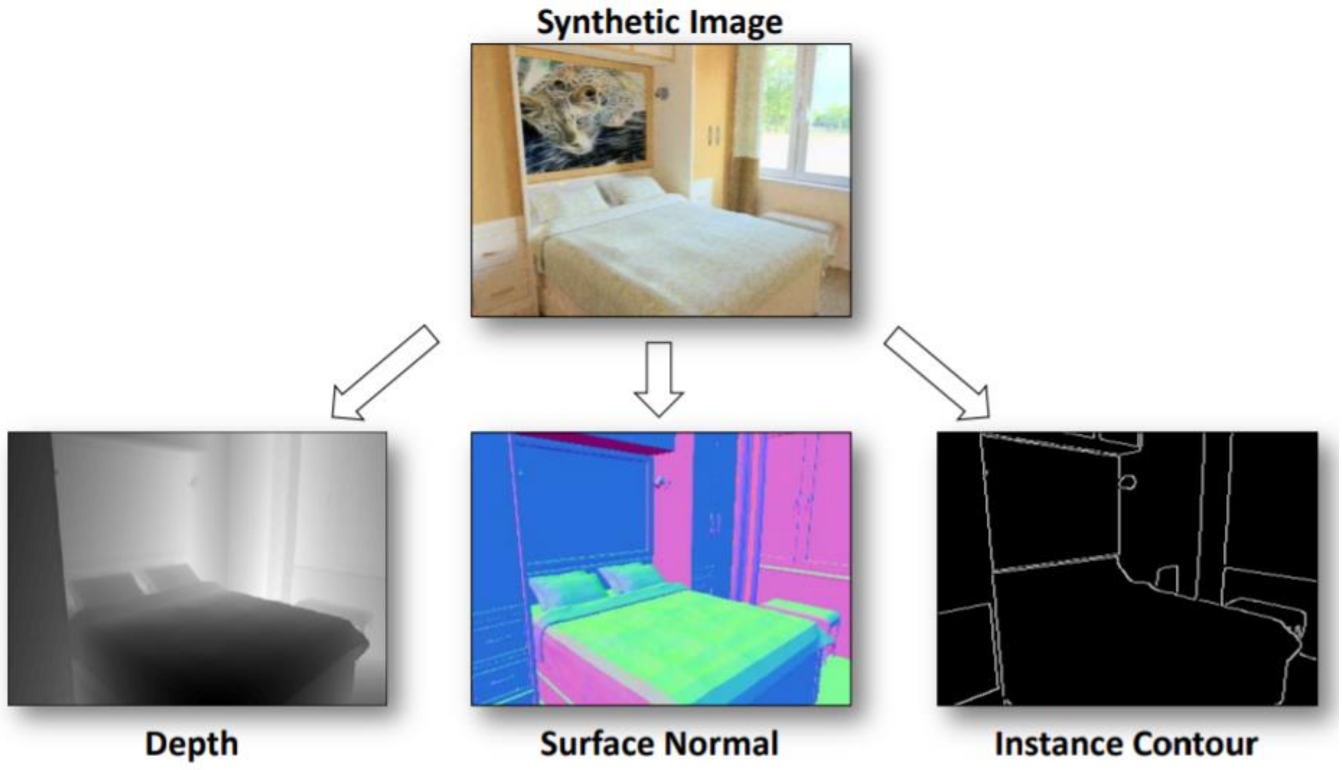
Deep Clustering: PASCAL VOC Transfer results

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

Synthetic Imagery

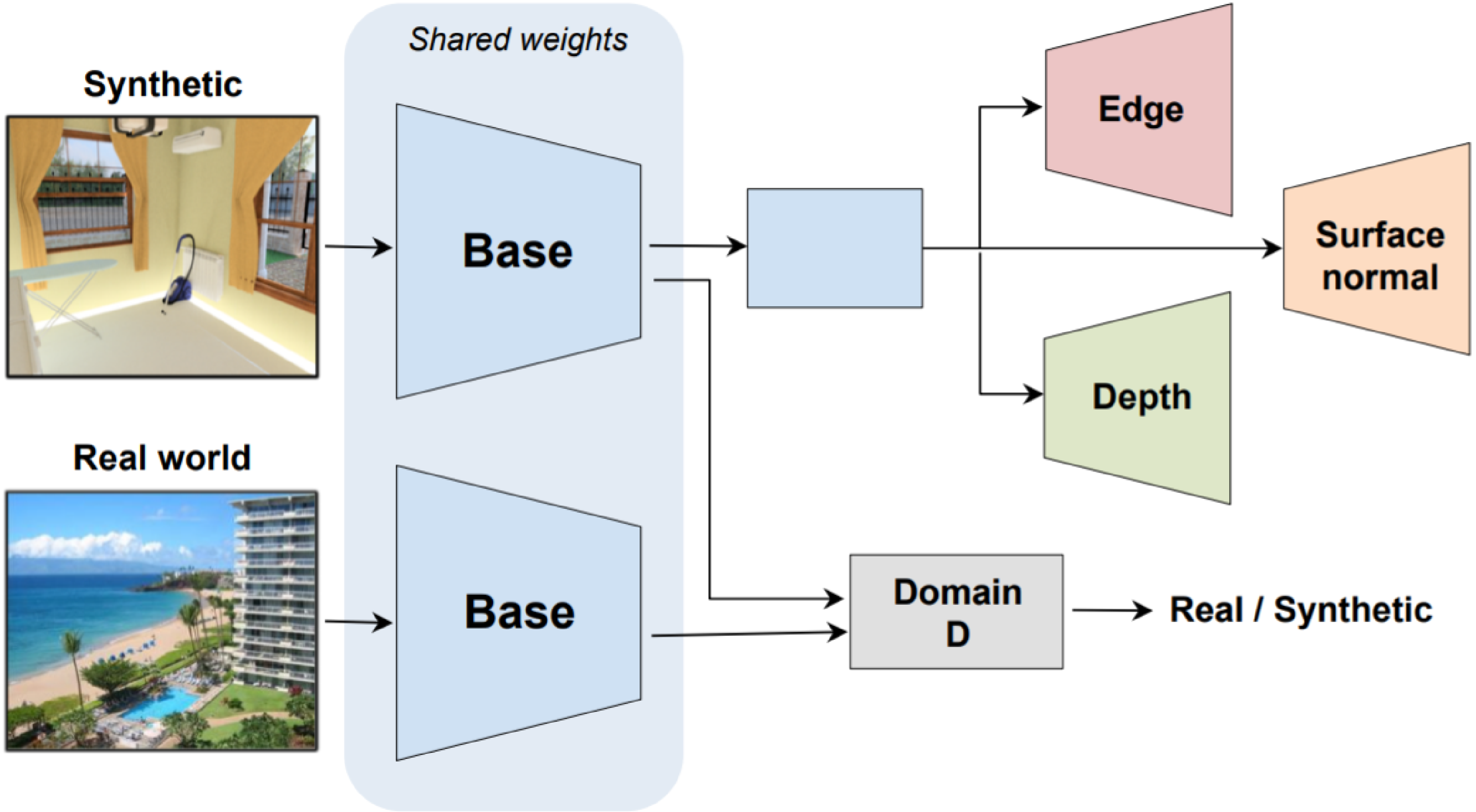
What if we prepared training pairs of (image, properties) by generating synthetic images using graphics engines and adapting it to real images?



Cross-Domain Self-supervised Multi-task Feature Learning using Synthetic Imagery

Synthetic Imagery

The upper net takes a synthetic image and predicts its depth, surface normal, and instance contour (edge) map. The bottom net extracts features from a real-world image. The domain discriminator D tries to differentiate real and synthetic features. The learned blue modules are used for transfer learning on real-world tasks.



Cross-Domain Self-supervised Multi-task Feature Learning using Synthetic Imagery

Self-supervised learning: Outline

- Data prediction
  - Colorization, Superresolution, Inpainting, Cross channel encoding
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Automatic Label Generation
  - Image clustering, Synthetic imagery
- Contrastive learning
  - PIRL, MoCo, SimCLR, SWaV

Problems with earlier approaches

- As such, the image representations learned can overfit to the transformation and not generalize well on downstream tasks.
- The representations will be **covariant** with the transformation.
- It will only encode essential information to predict the transformation and could discard useful semantic information.

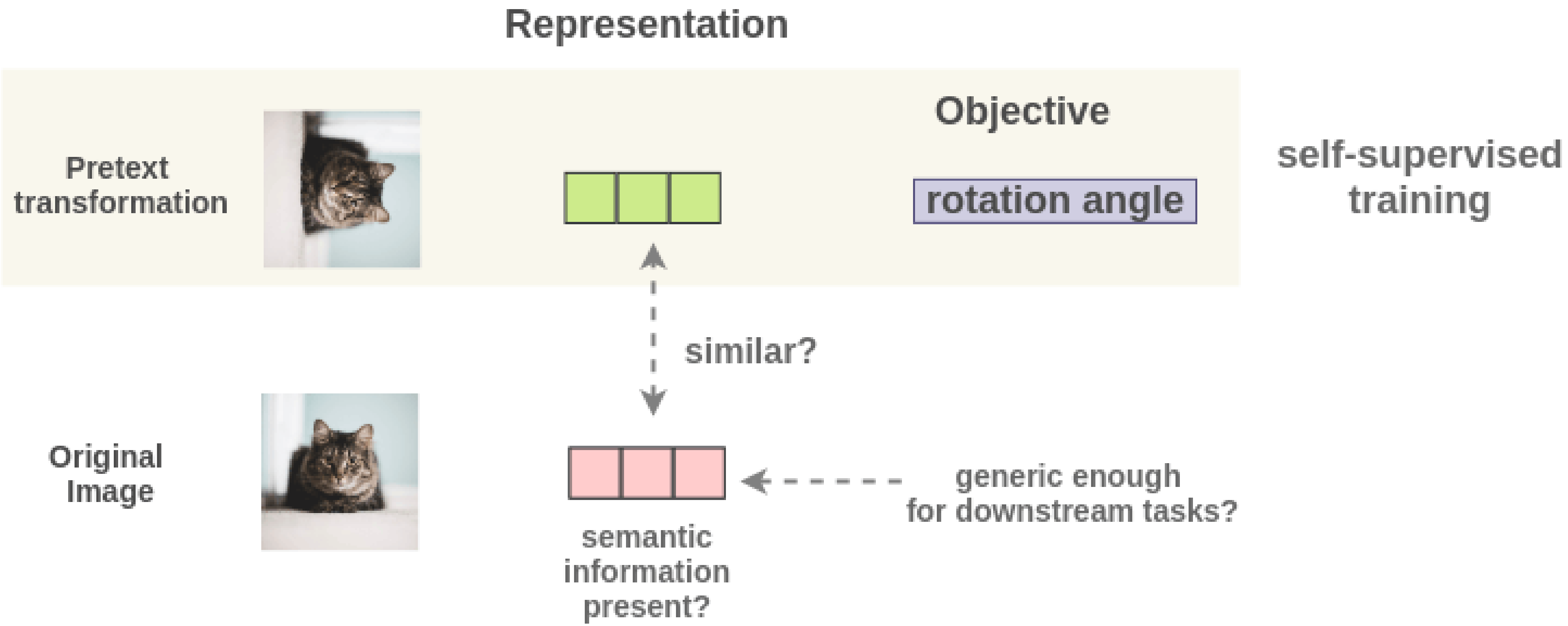


Figure source

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

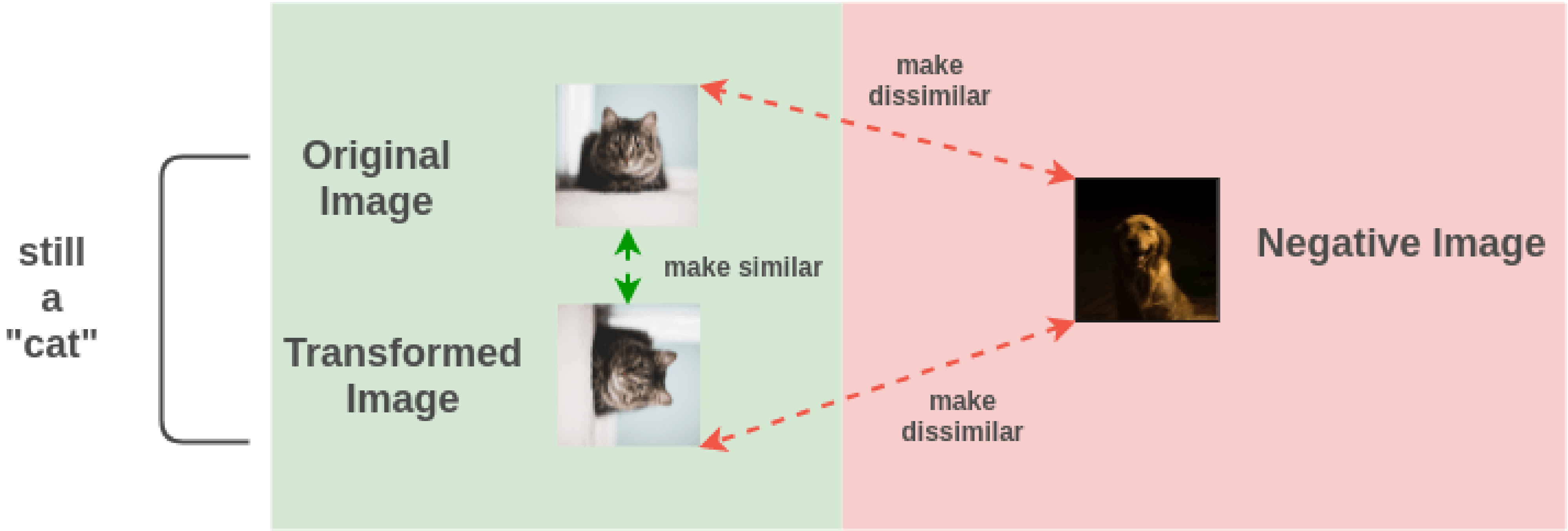


Figure source

Contrastive loss formulation

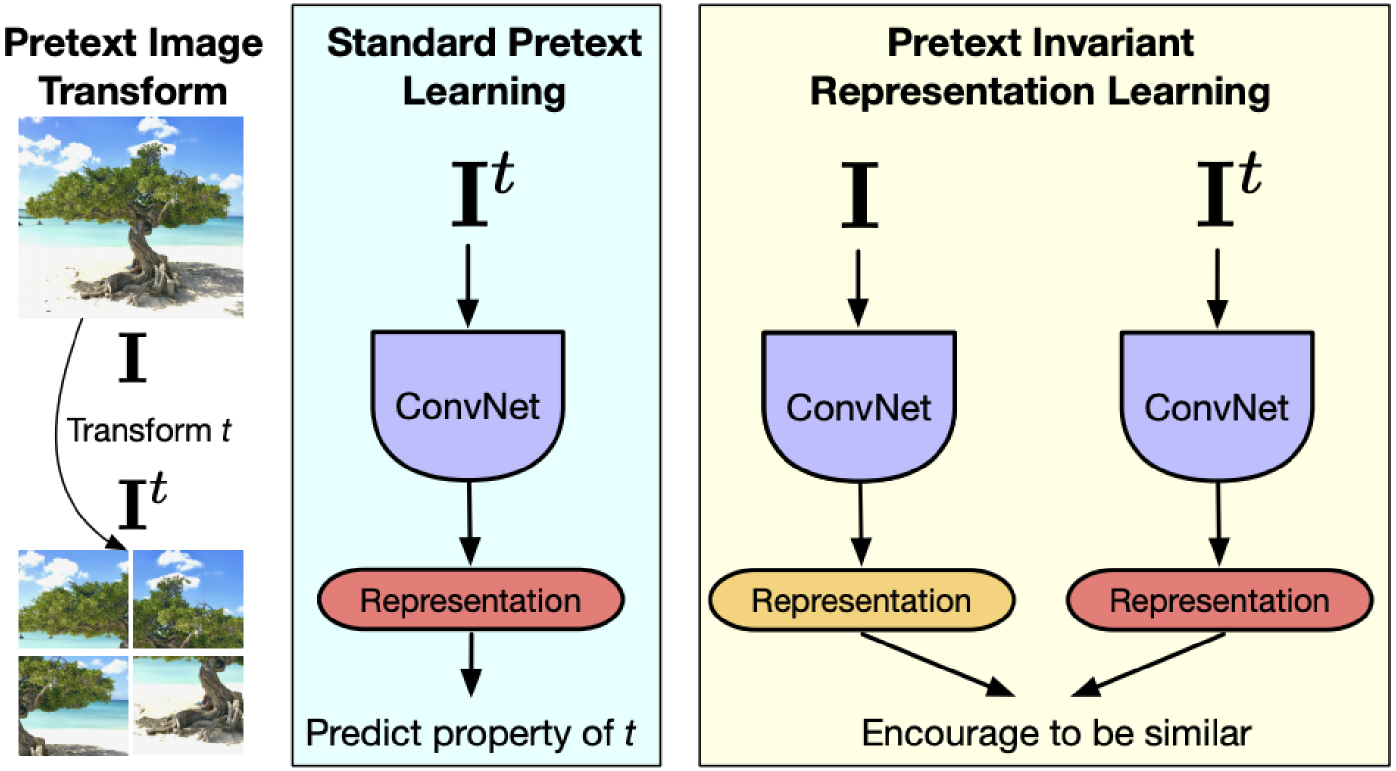
- Given: query point x, positive samples x^+, negative samples x^-
  - Positives are typically transformed versions of x, negatives are random examples from the same mini-batch or *memory bank*
- Key idea: learn representation to make x similar to x^+, dissimilar from x^- (similarity is measured by dot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

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

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

Pretext-invariant representation learning (PIRL)

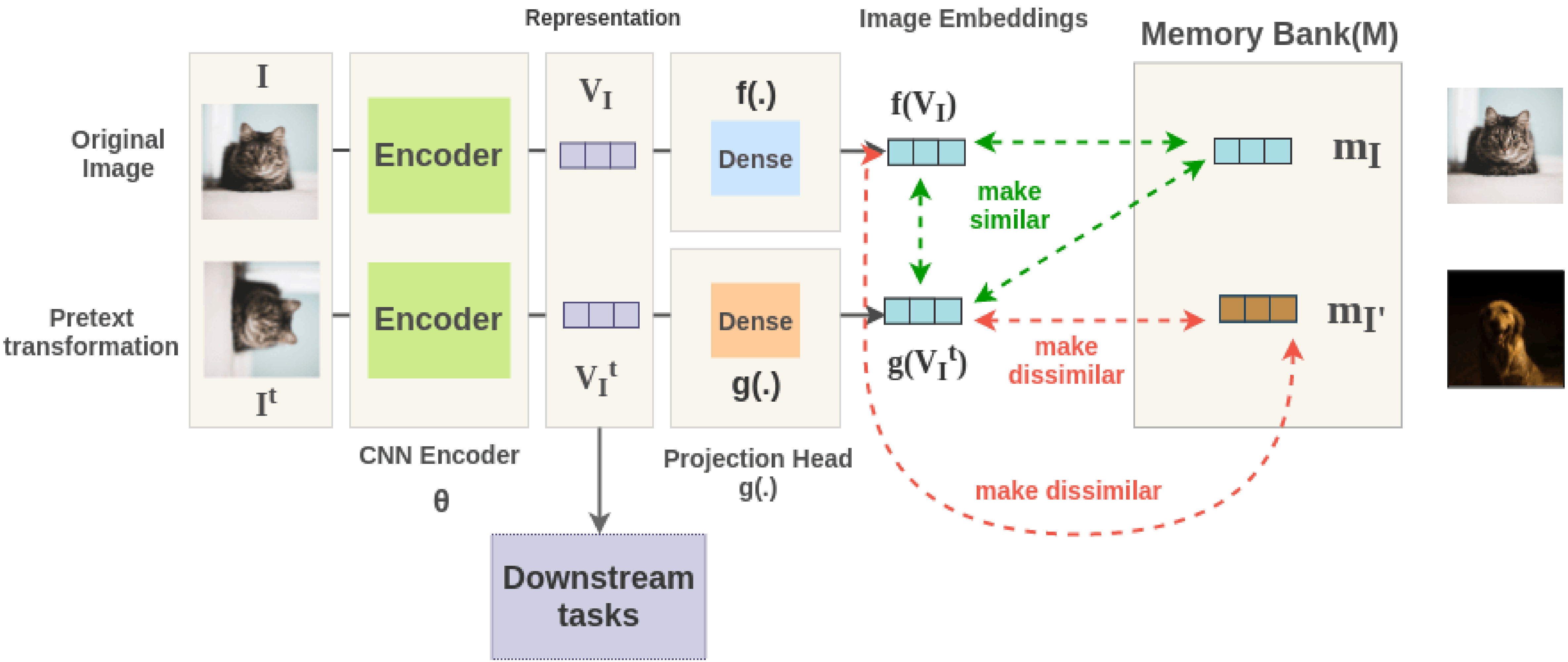
- Key idea: instead of predicting the transformation of the input, learn a representation *invariant* to the transformation



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

Pretext-invariant representation learning (PIRL)

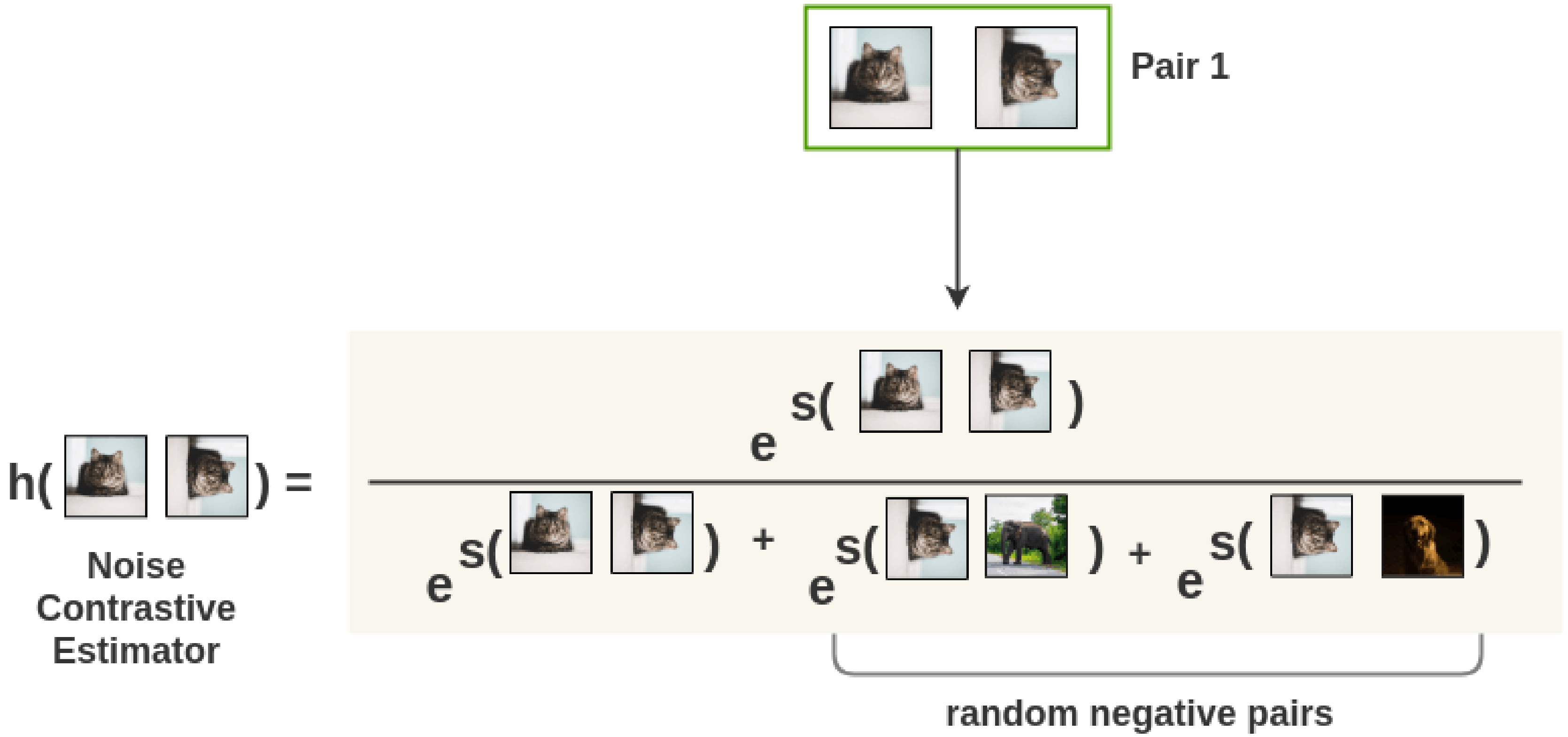
PIRL Generic Framework



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

Pretext-invariant representation learning (PIRL)

PIRL uses Noise Contrastive Estimator (NCE)



$$L_{NCE}(I, I^t) = -log[h(f(V_I), g(V_{I^t}))] - \sum_{I' \in D_N} log[1 - h(g(V_{I^t}), f(V_{I'}))]$$

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

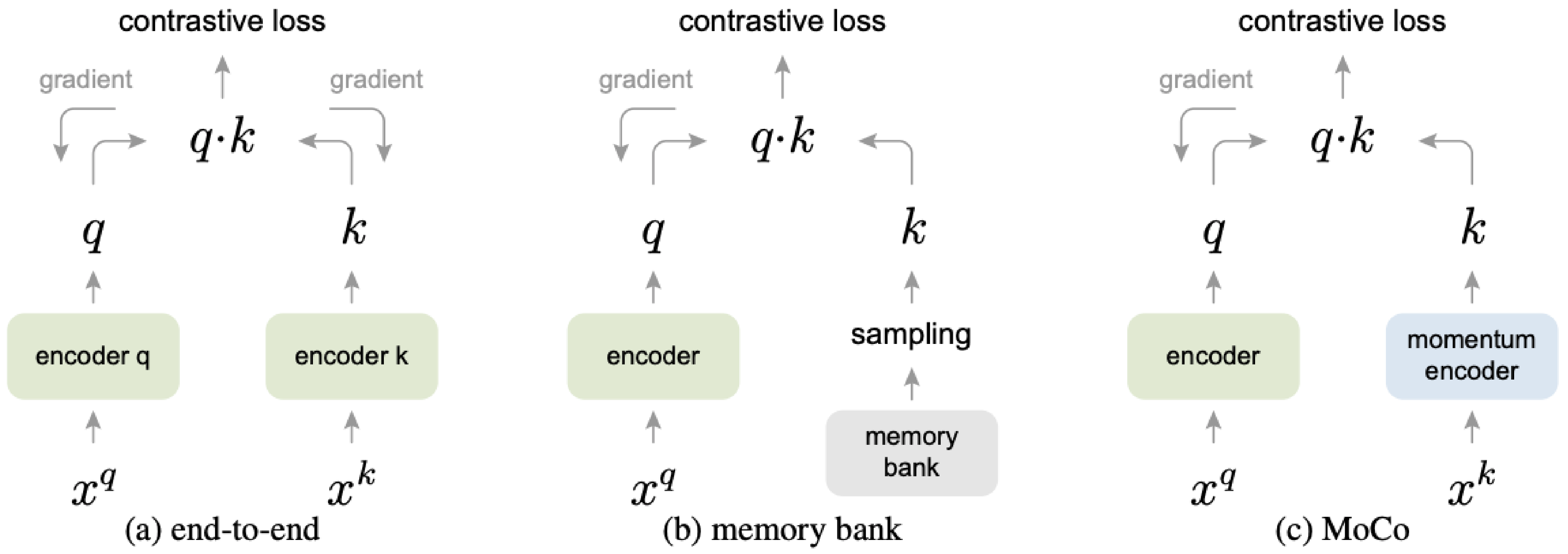
PIRL: Results

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	$55.2^{*\dagger}$	$81.4^{*\dagger}$	$61.2^{*\dagger}$	

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.

Momentum contrast

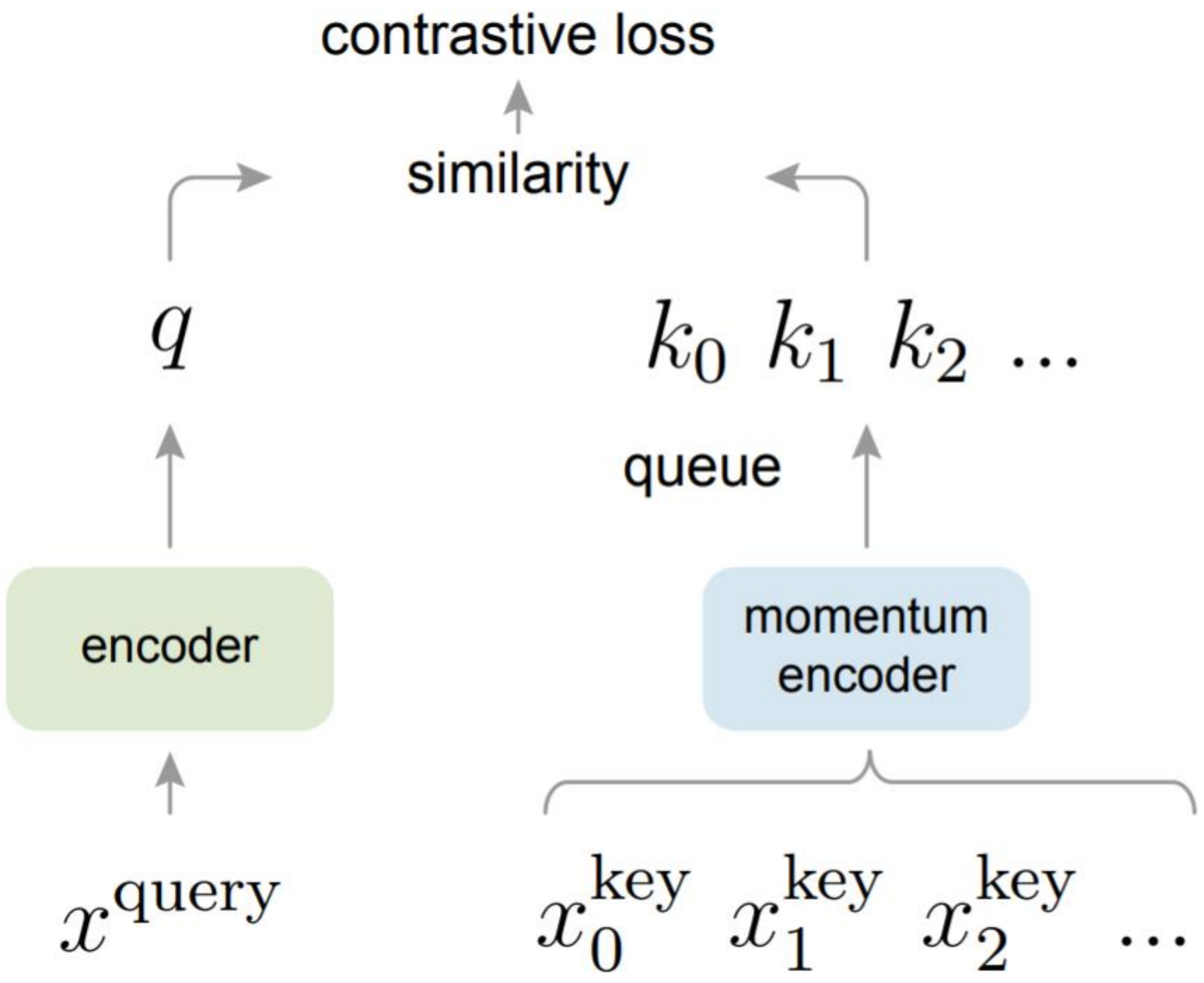
- Use instance discrimination as pretext task, transform query and key by random augmentations, use queue encoded by a momentum encoder instead of memory bank



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

Momentum contrast

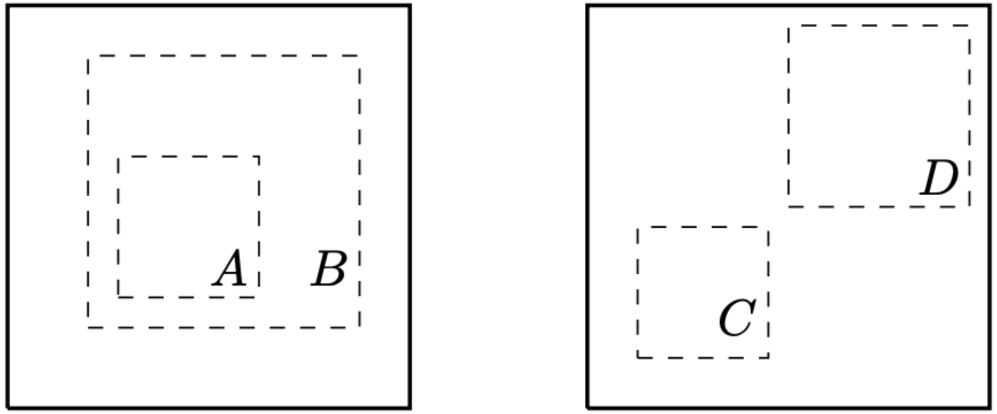
- Momentum Contrast (MoCo) trains a visual representation encoder by matching an encoded query q to a dictionary of encoded keys using a contrastive loss.
- The dictionary keys {k0, k1, k2, ...} are defined on-the-fly by a set of data samples.
- The dictionary is built as a queue, with the current mini-batch enqueued and the oldest mini-batch dequeued, decoupling it from the mini-batch size.
- The keys are encoded by a slowly progressing encoder, driven by a momentum update with the query encoder.
- This method enables a large and consistent dictionary for learning visual representations.



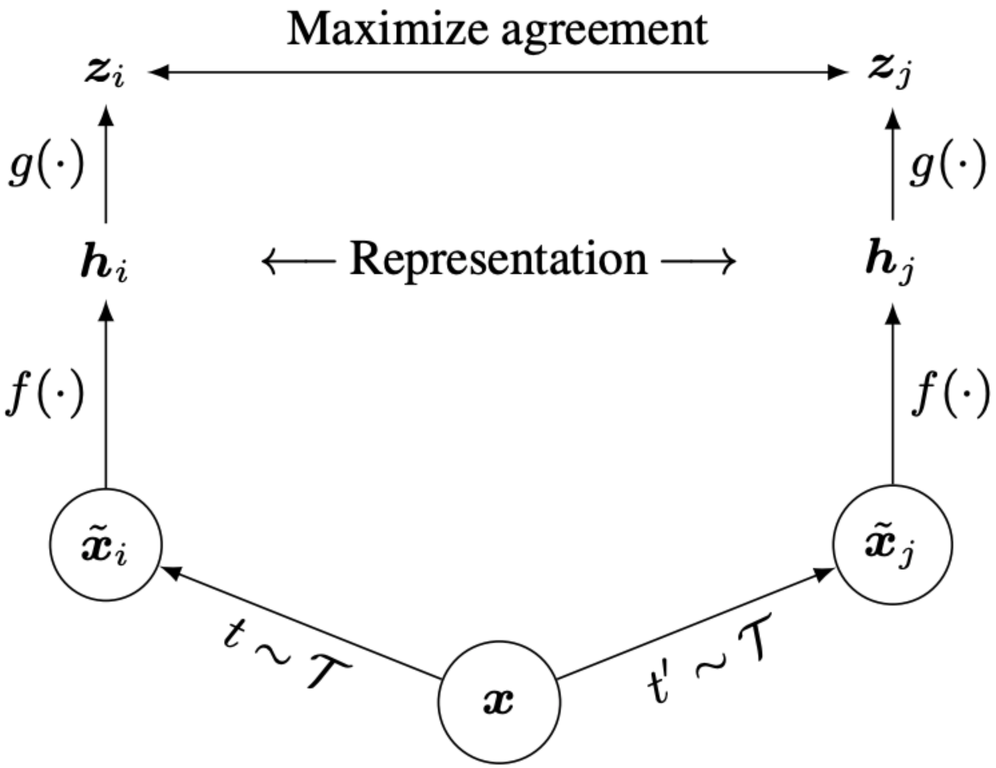
K. He et al. Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020

SimCLR

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



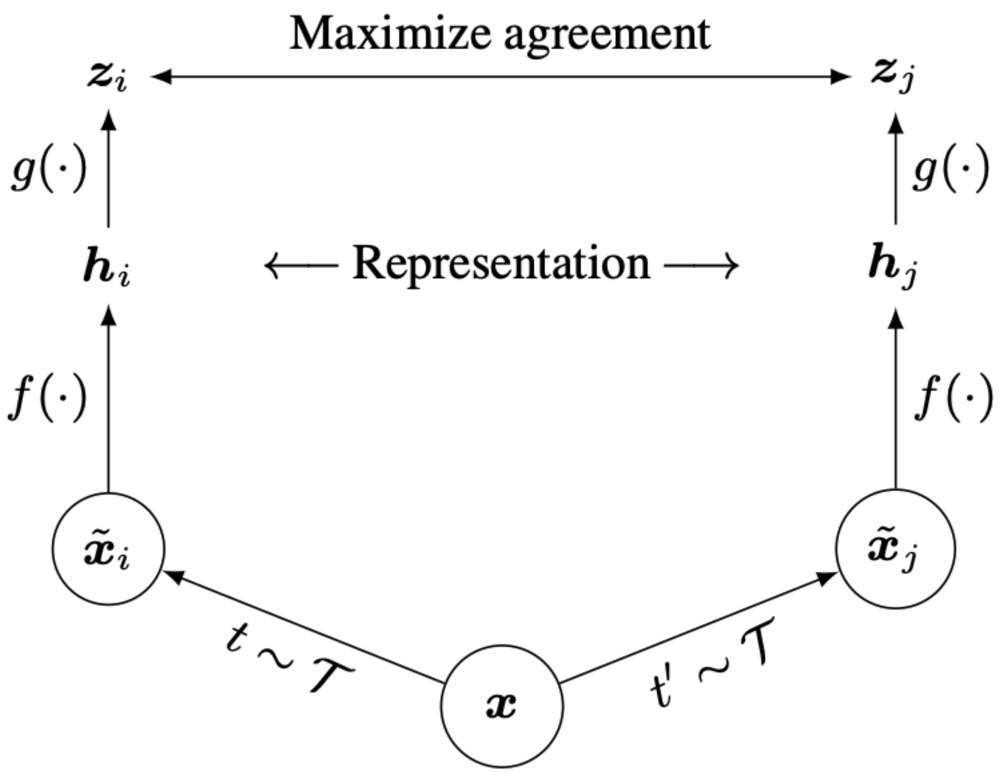
(a) Global and local views. (b) Adjacent views.



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

SimCLR

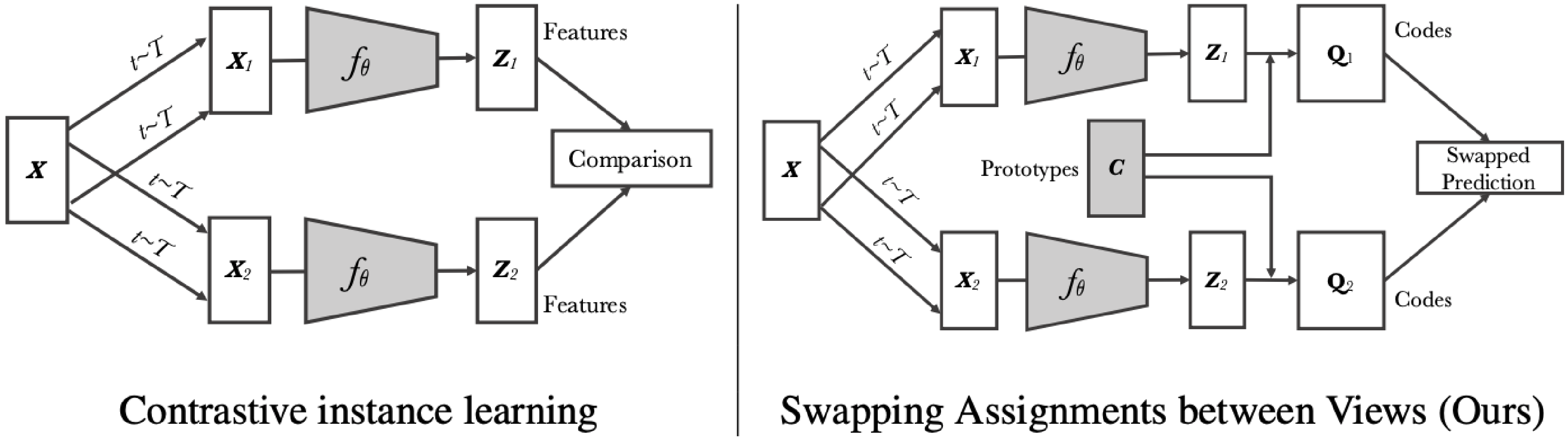
- Form two views of the input by composing data augmentations
  - Cropping and resizing, color distortion, blur
- No memory bank, large mini-batch size (on cloud TPU)
- Introduce nonlinear transformation between representation and contrastive loss (or, use representation a few layers below the contrastive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



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

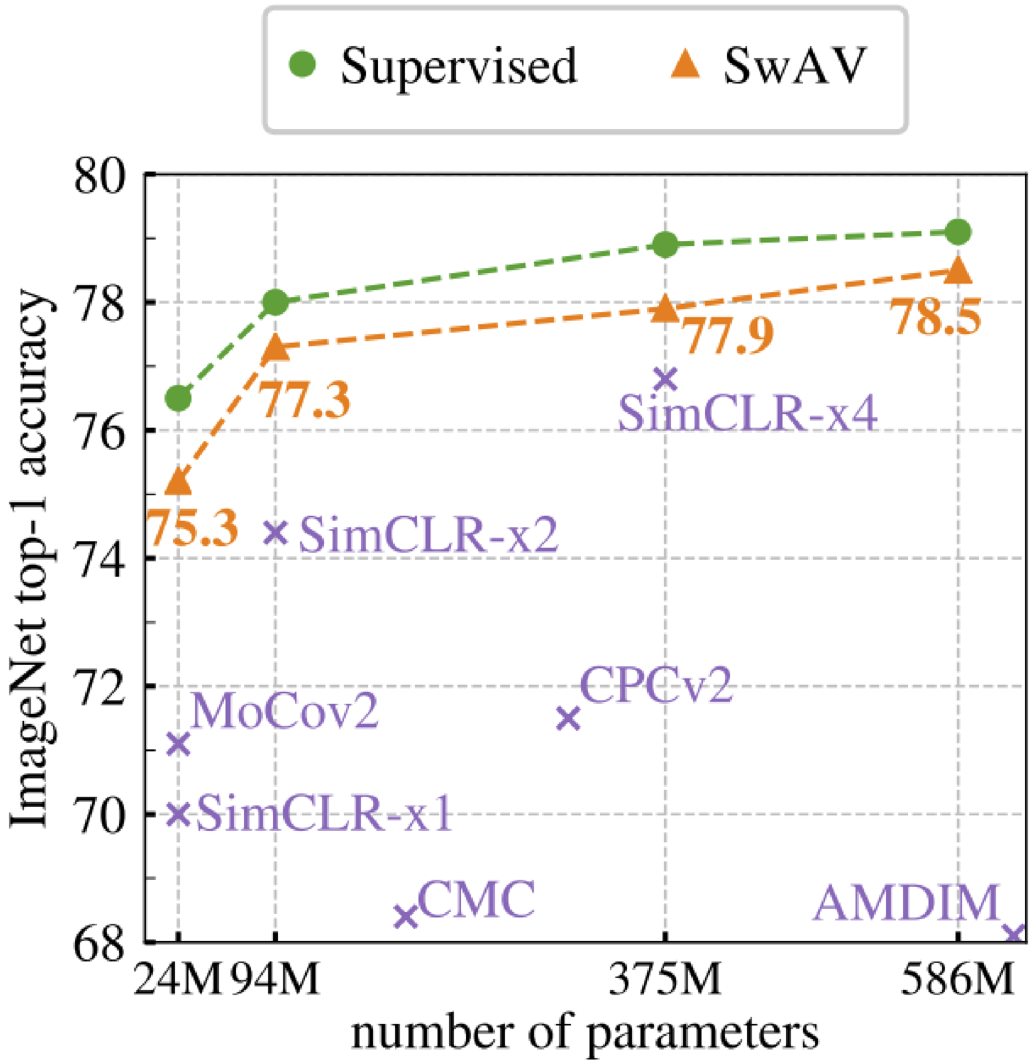
Swapping Assignments Between Views (SWaV)

- 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
- Simply put, it uses a swapped prediction mechanism where it predicts the code of a view from the representation of another view.
- Once again, data augmentation strategy matters



M. Caron et al. [Unsupervised Learning of Visual Features by Contrasting Cluster Assignments](). NeurIPS 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. NeurIPS 2020

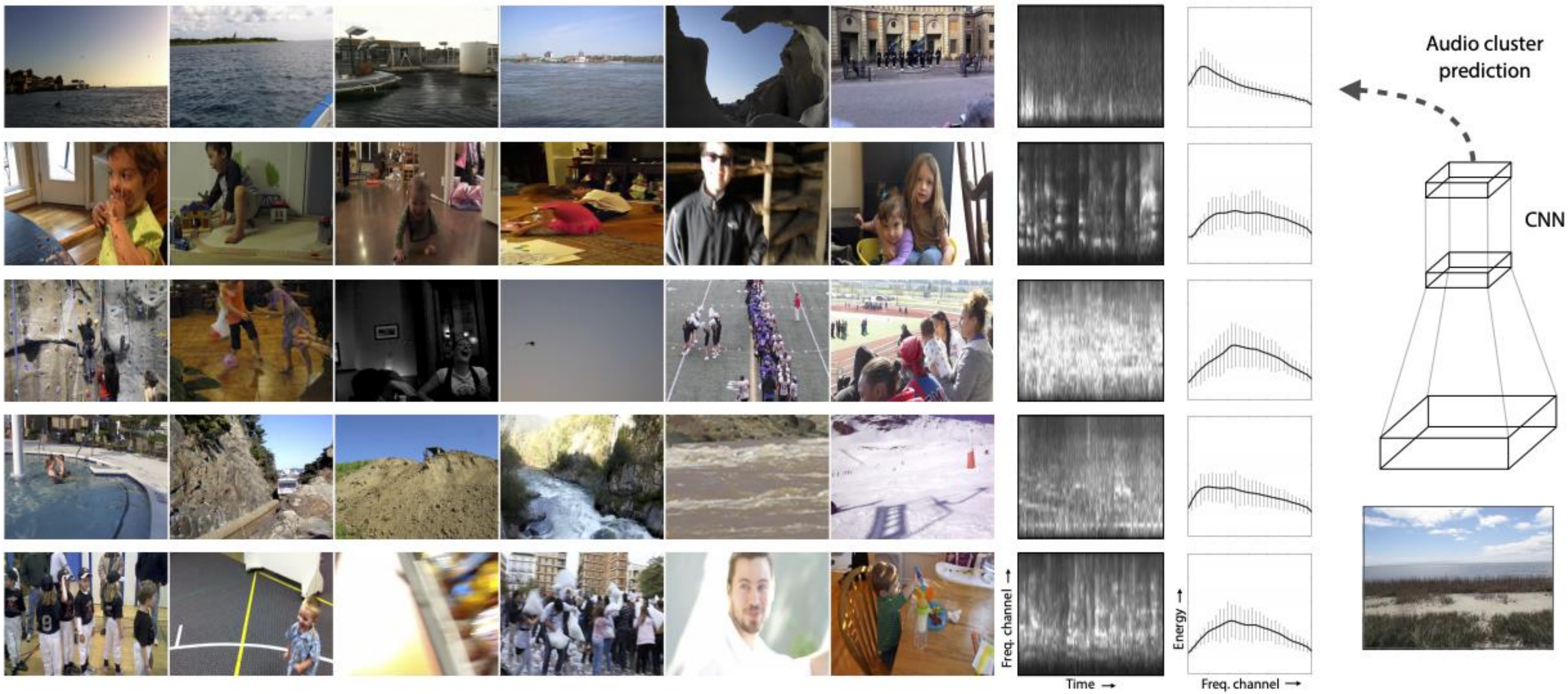
Why do contrastive methods work?

- 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)

Self-supervised learning: Outline

- Data prediction
  - Colorization, Superresolution, Inpainting, Cross channel encoding
- Transformation prediction
  - Context prediction, jigsaw puzzle solving, rotation prediction
- Automatic Label Generation
  - Image clustering, Synthetic imagery
- Contrastive learning
  - PIRL, MoCo, SimCLR, SWaV
- Self-supervision beyond still images
  - Audio, video, 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

Self-Supervised Learning From Video

Frame Order Verification

What if we prepared training pairs of (video frames, correct/ incorrect order) by shuffling frames from videos of objects in motion?

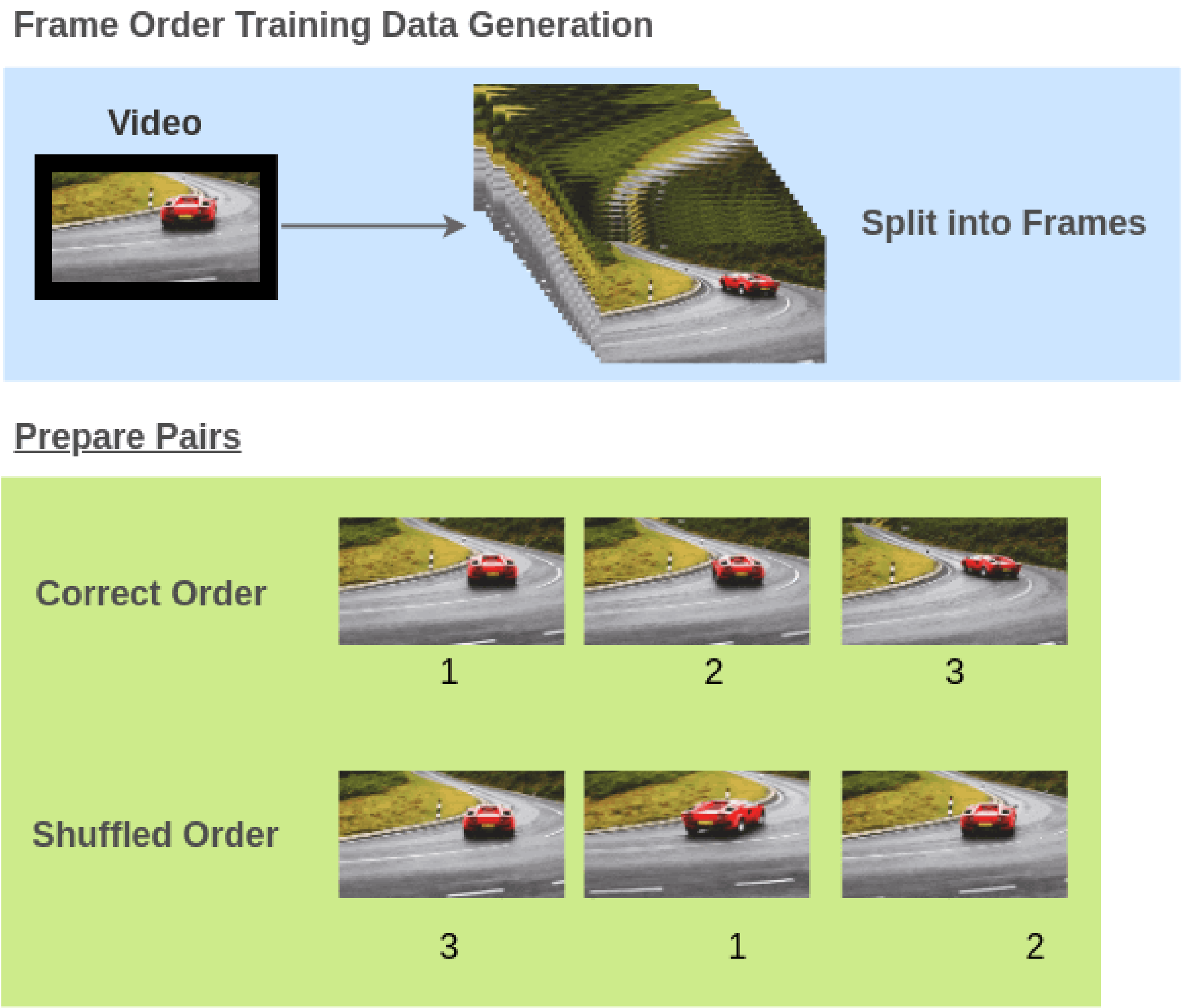
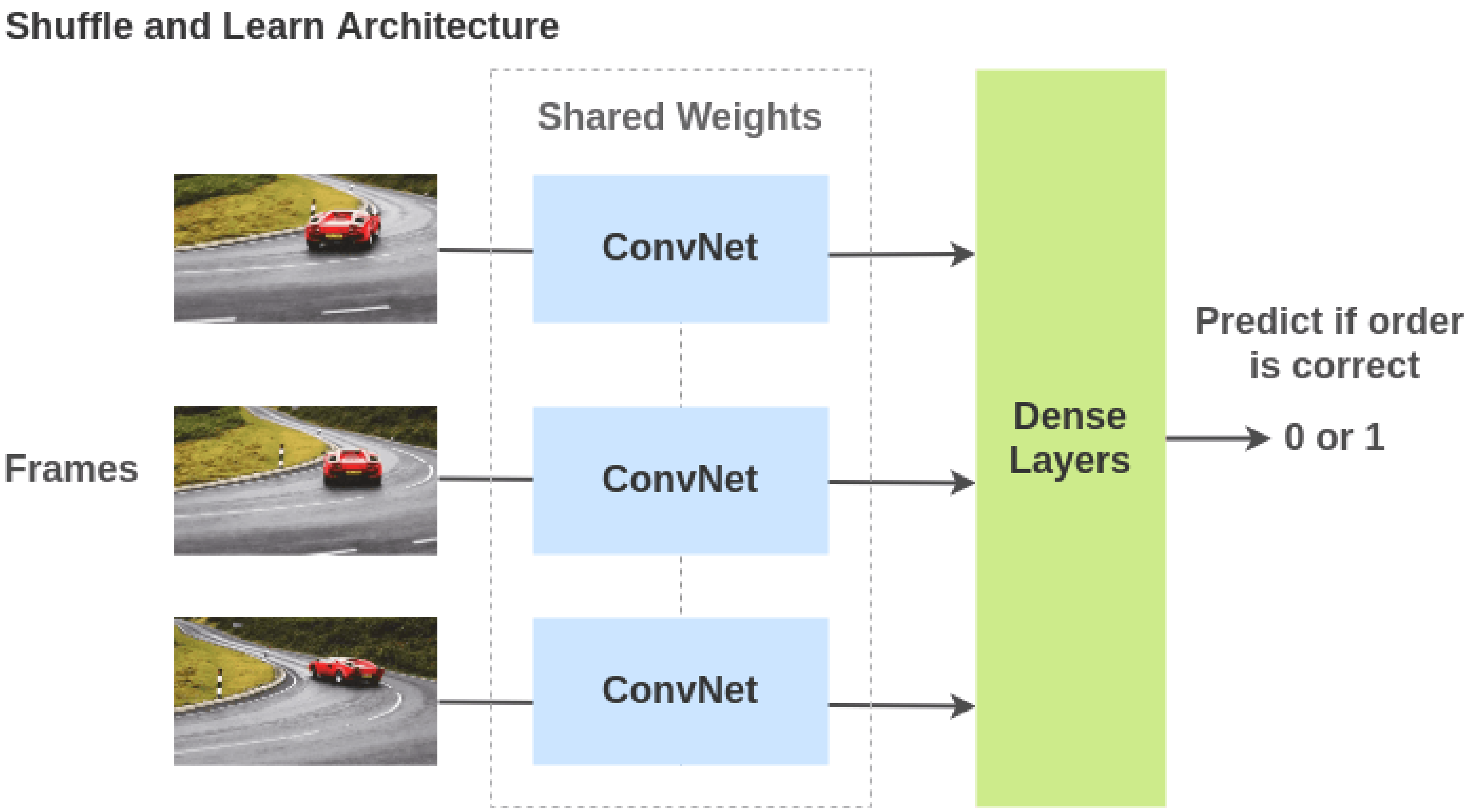


Figure source

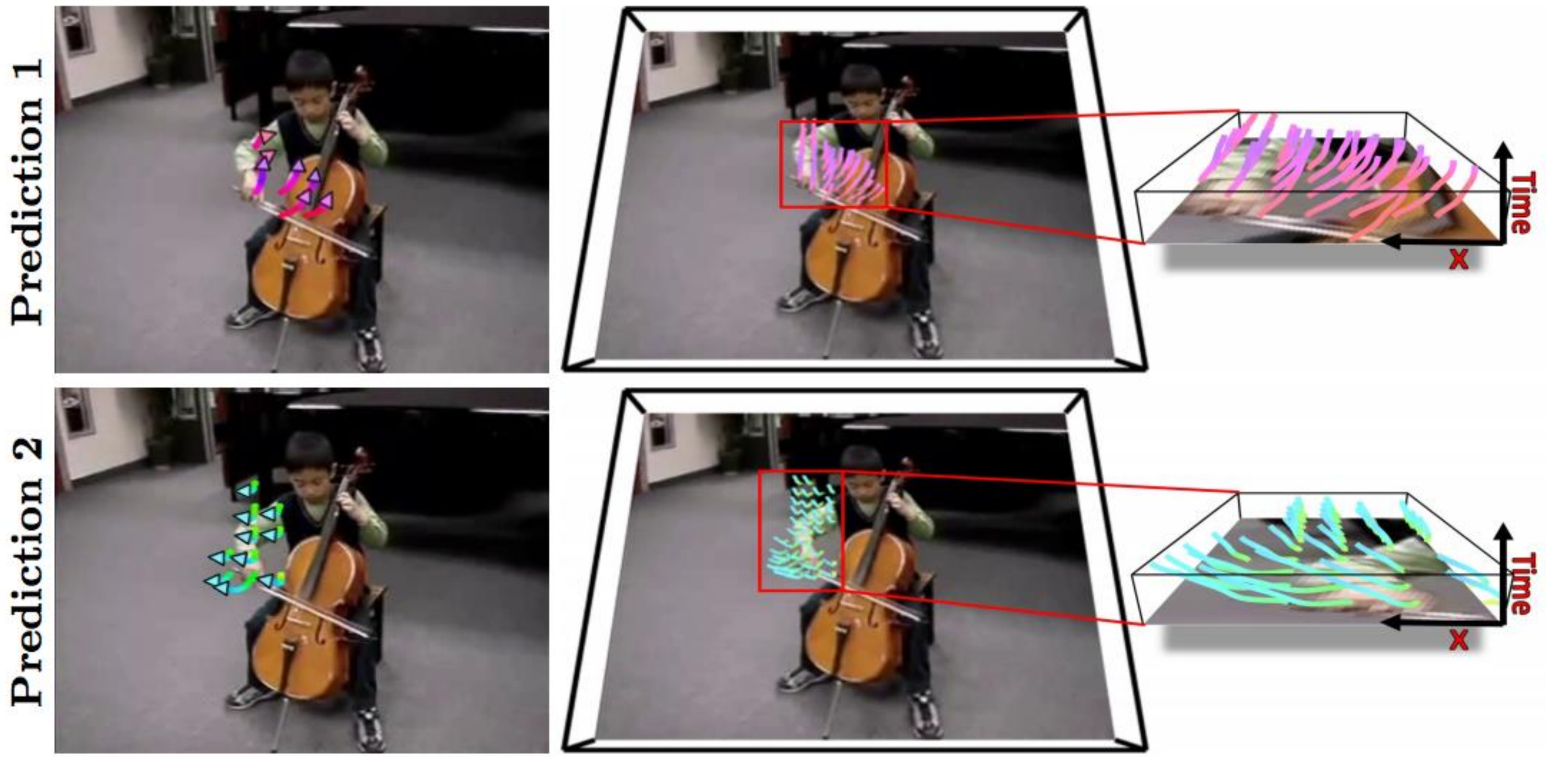
Self-Supervised Learning From Video



Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

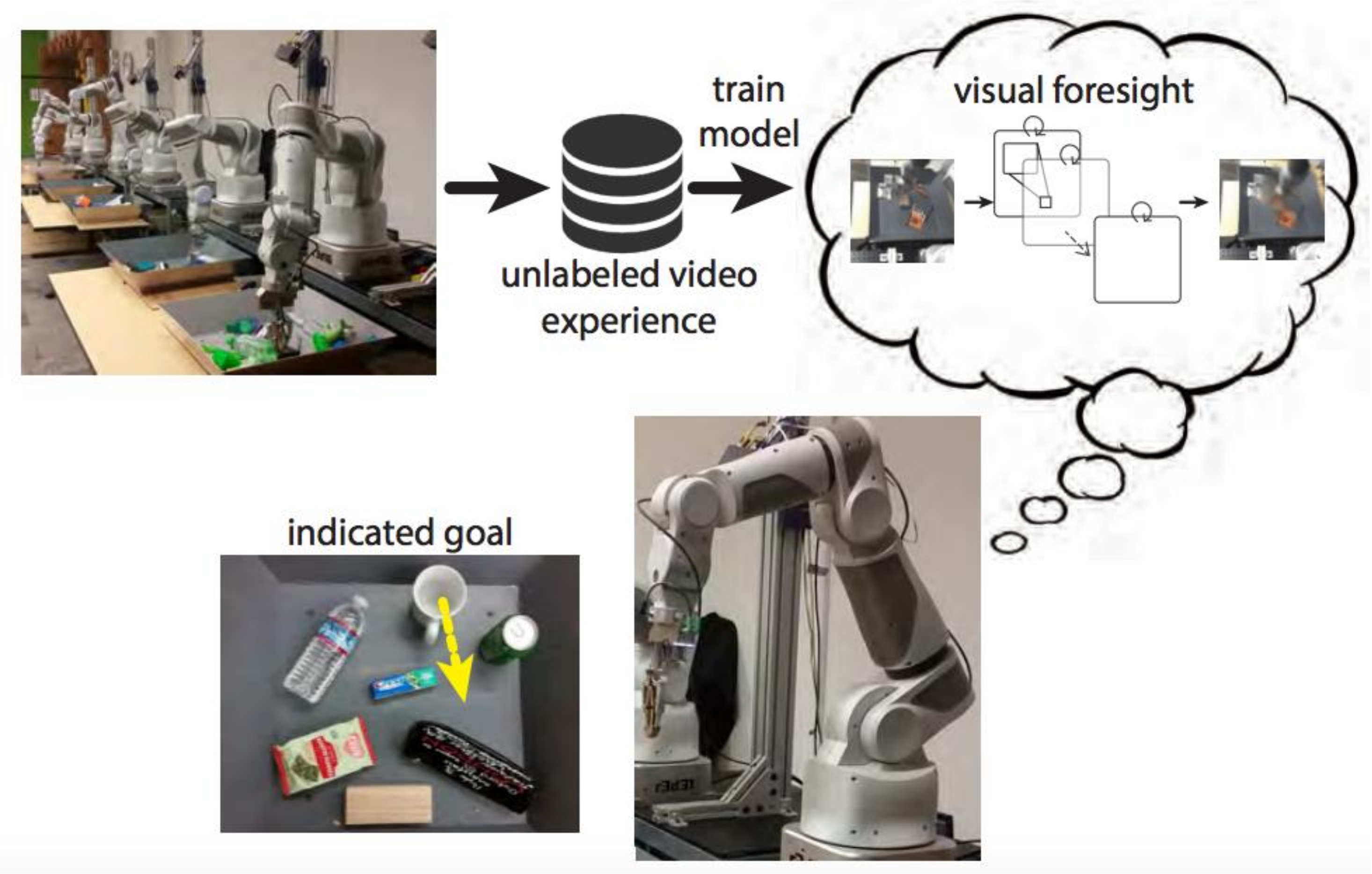
Figure source

Future prediction



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

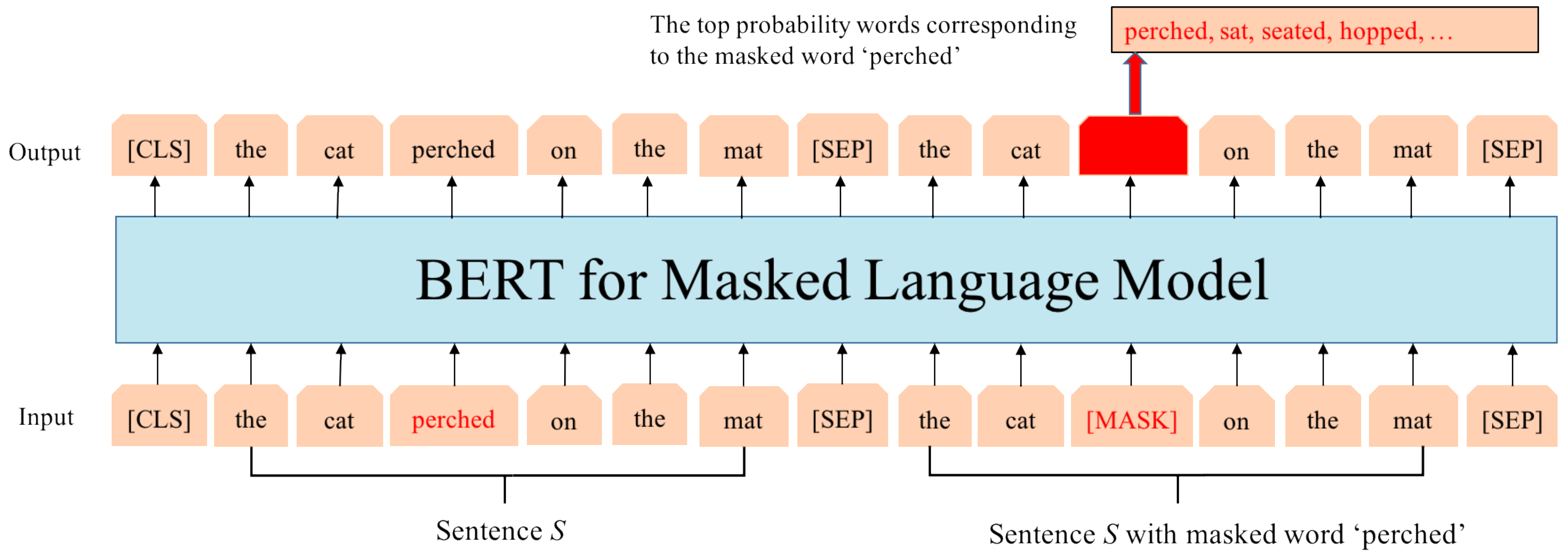
Future prediction



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

Acknowledgement

Thanks to the following courses and corresponding researchers for making their teaching/research material online

- Deep Learning, Stanford University
- Introduction to Deep Learning, University of Illinois at Urbana-Champaign
- Introduction to Deep Learning, Carnegie Mellon University
- Convolutional Neural Networks for Visual Recognition, Stanford University
- Natural Language Processing with Deep Learning, Stanford University
- And Many More ......