Self-Supervised Learning

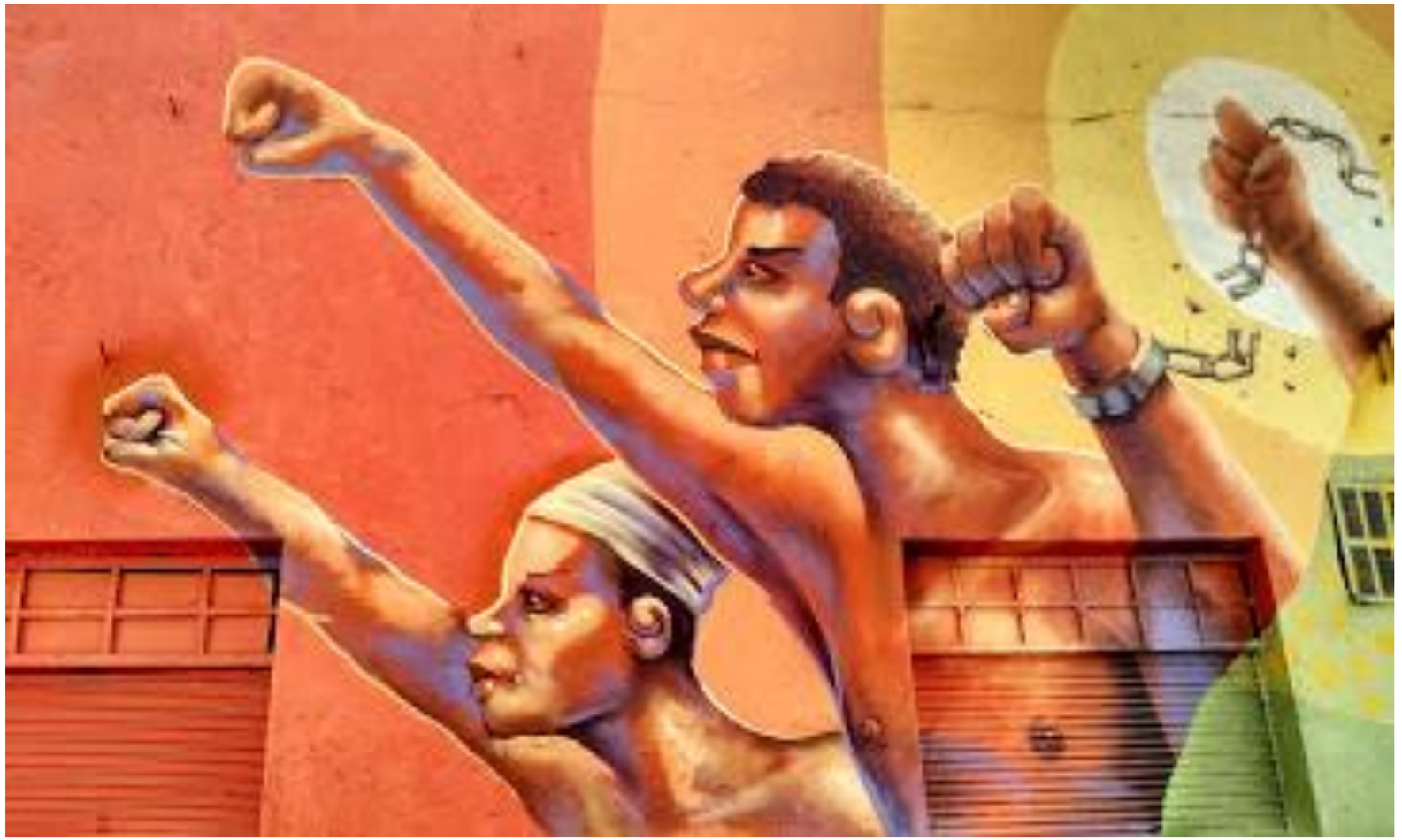


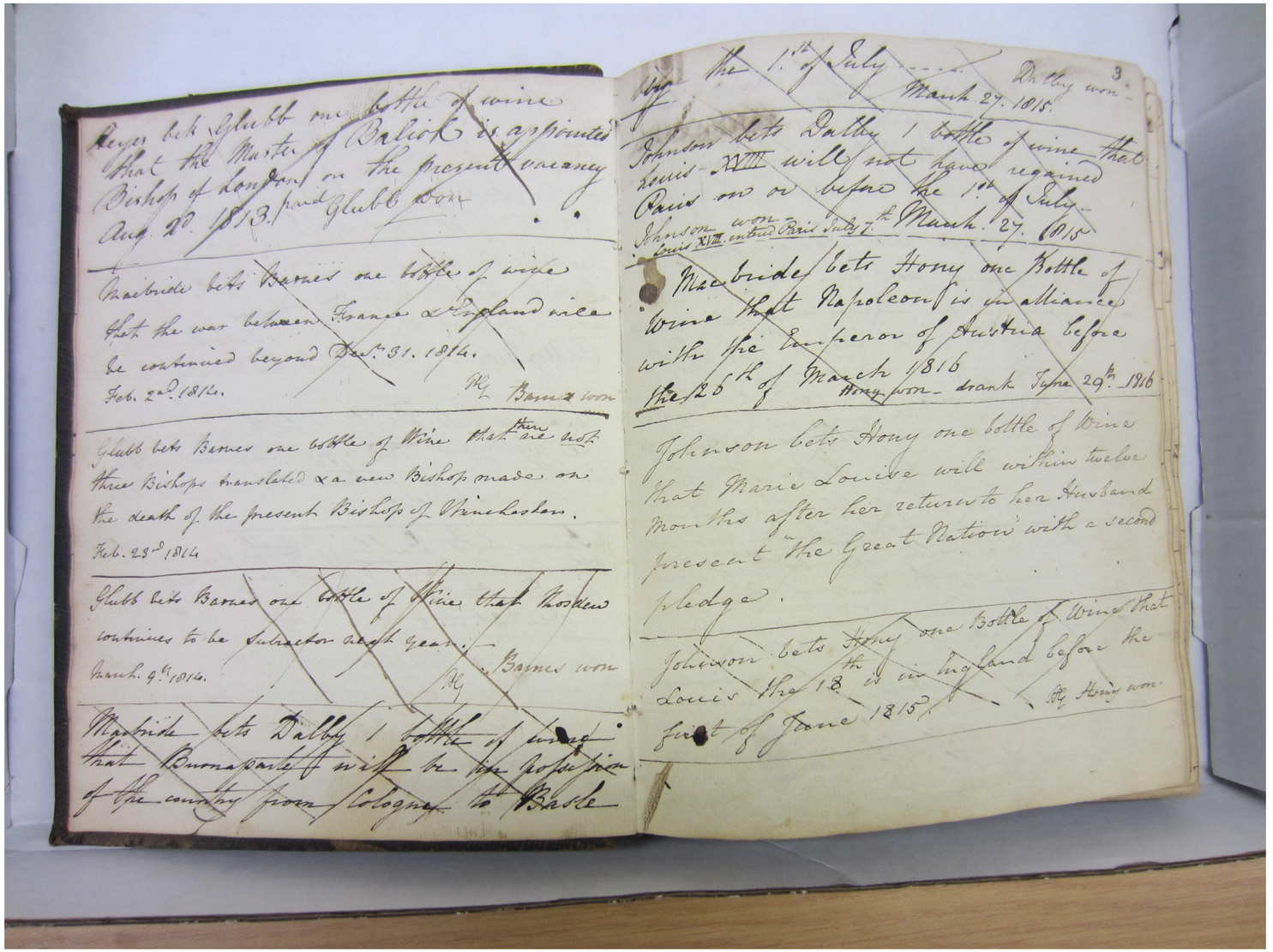
Photo from Santiago, ICCV 2015

“The Revolution will not be supervised!”

Alexei Efros, CS280, Spring 2024

It all started with a bet…

Long Tradition of Scientific Bets



Betting Book
Exeter College
Oxford
1815

The Gelato Bet

Sept 23, 2014

- R-CNN just came out
- It was surprising (to me), that ImageNet pretraining helped in PASCAL detection

The Gelato Bet

Sept 23, 2014

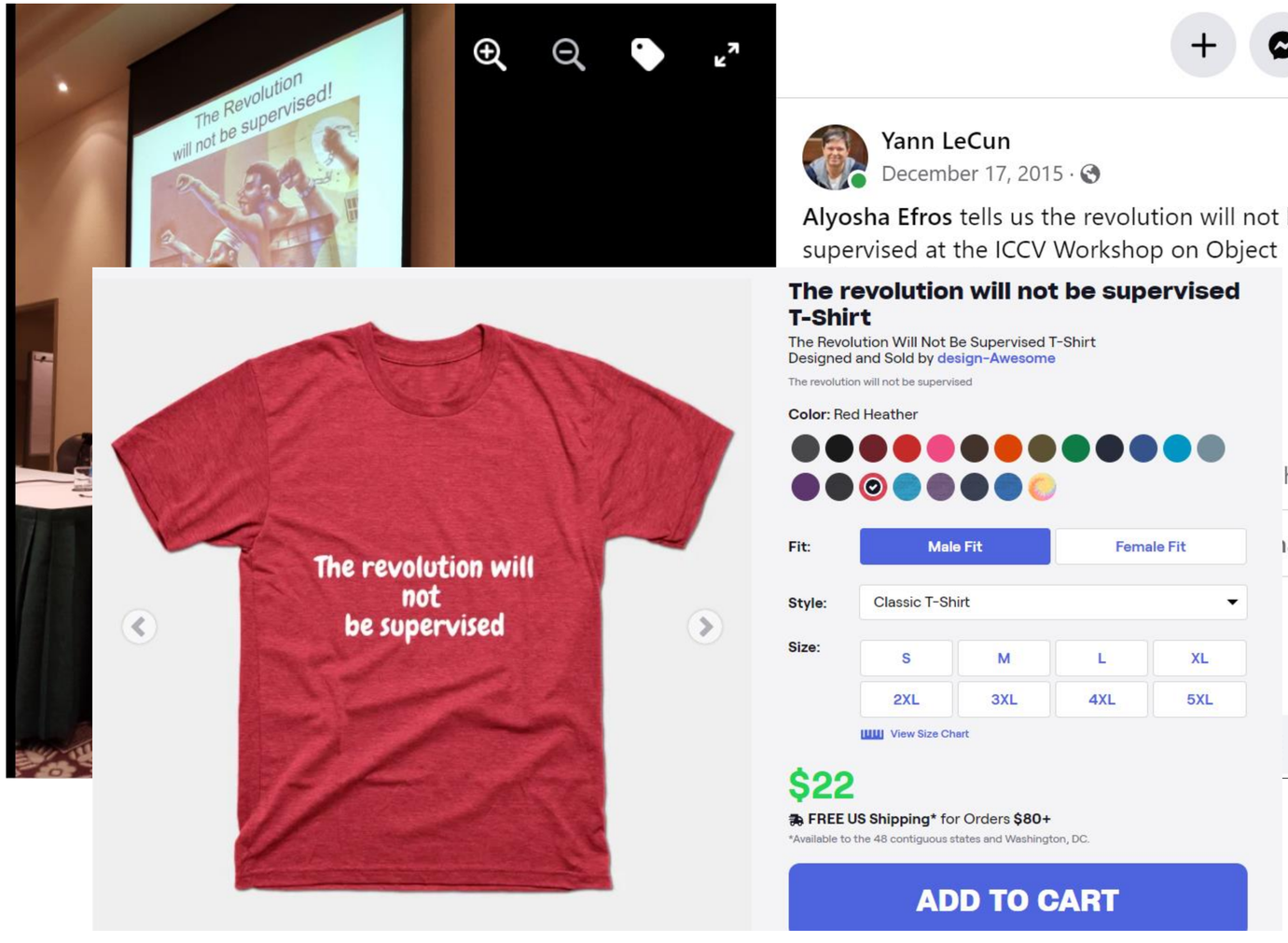
"If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, ***without the use of any extra, human annotations*** *(e.g. ImageNet labels) as pre-training, Mr. Malik promises to buy Mr. Efros one (1) gelato (two scoops: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

One year later…

Of course, I lost!

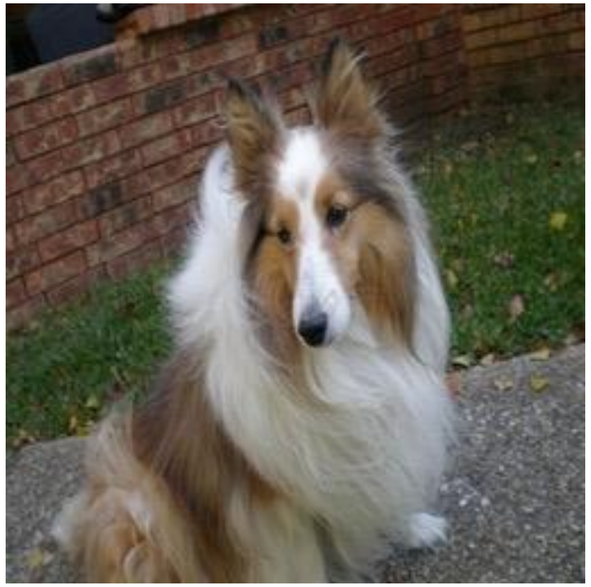
One year later…

- …but:
  - First 4 self-supervised papers presented at **ICCV 2015**
    - Doersch et al, Agrawal et al, Wang et al, and Jayaraman et al.
  - Yann LeCun liked my talk and posted on FB
  - Rest is history…



Problems with top-down (human-supplied) labels

classifiers love to cheat



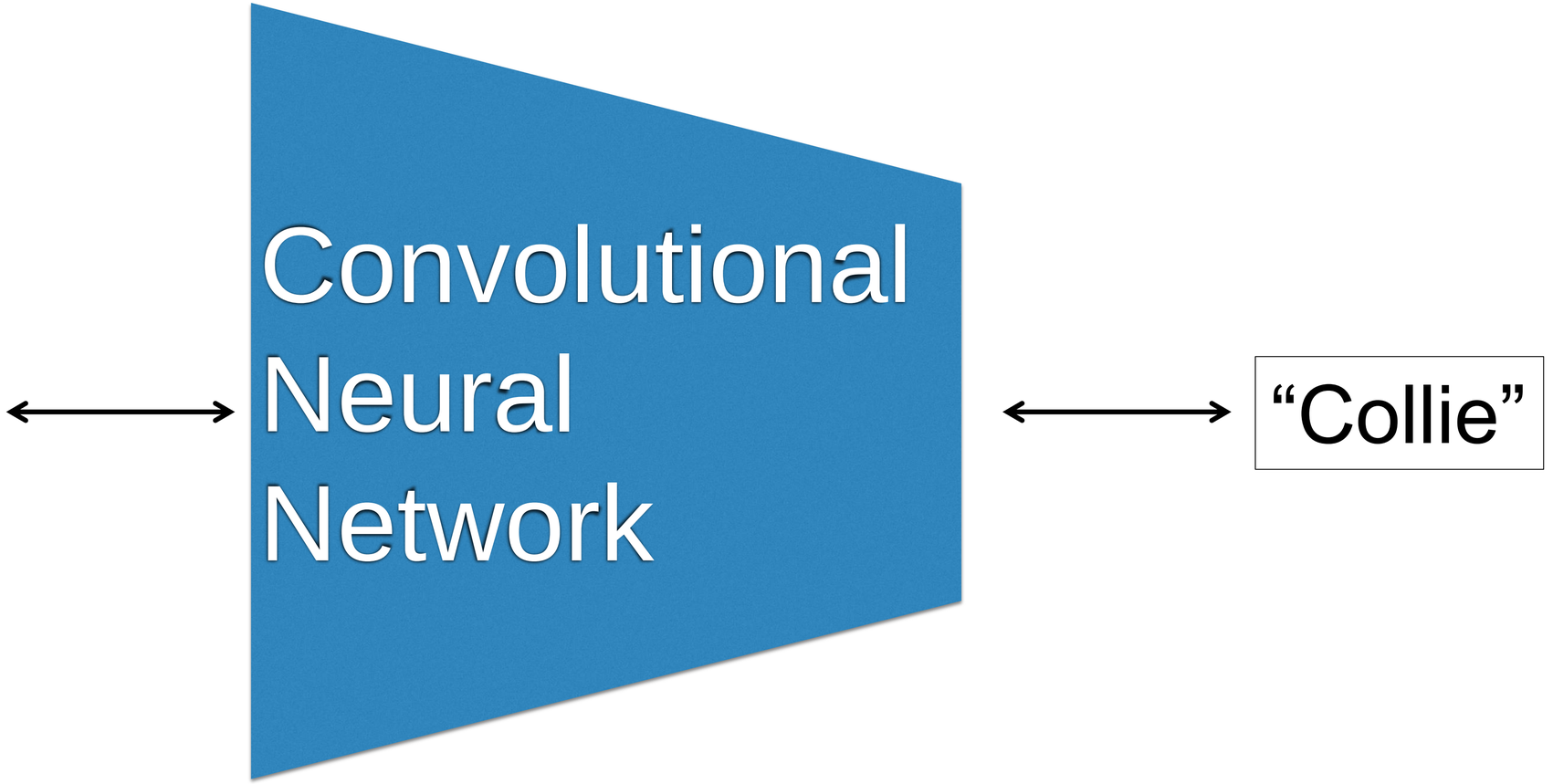


image *X*

label *Y*

classifiers love to cheat

Convolutional Neural Network

“Collie”

image *X*

label *Y*

Gatys et al, 2017

example: action recognition in video

– input video:




– output: class label

- Picking up cup
- Slicing bread
- **Opening fridge**
- etc

example by David Fouhey

semantic classification ~= memorization

We are raising a generation of algorithms who can only “cram for the test” (set)

Why do we have vision?

- “To see what is where by looking”
  - Aristotle, Marr, etc
- .
- .
- .
- .
- “To make babies who make babies, etc”
  - Darwin, Dawkins, etc.

Why do we have vision?

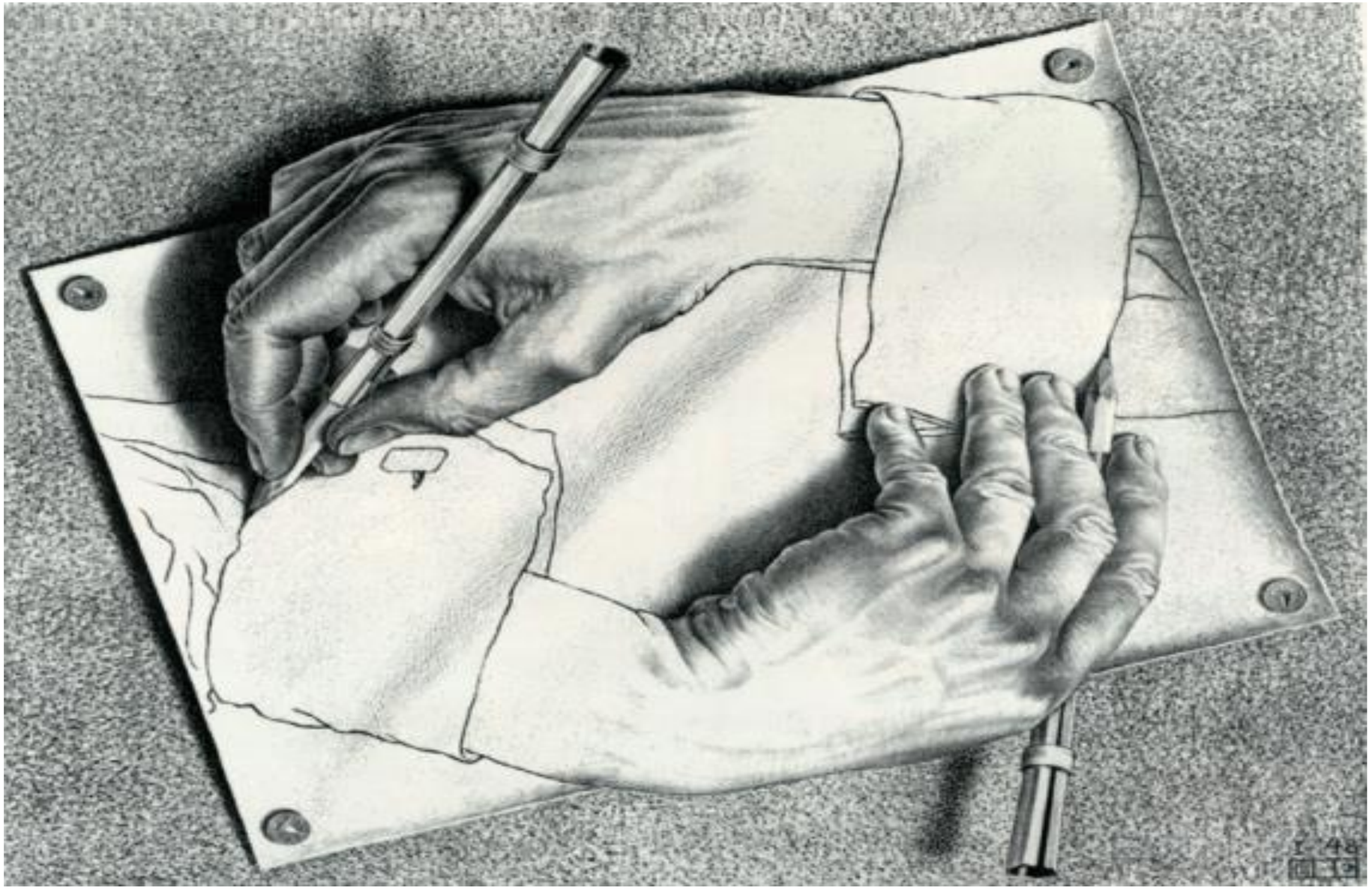
- “To see what is where by looking”
  - Aristotle, Marr, etc.
- .
- “To predict the world”
  - Jakob Uexküll, Jan Koenderink, Moshe Bar, etc.
- .
- “To make babies who make babies, etc”
  - Darwin, Dawkins, etc.

The world as supervision

Try to predict some aspect of the world that we interact with / have effect on:

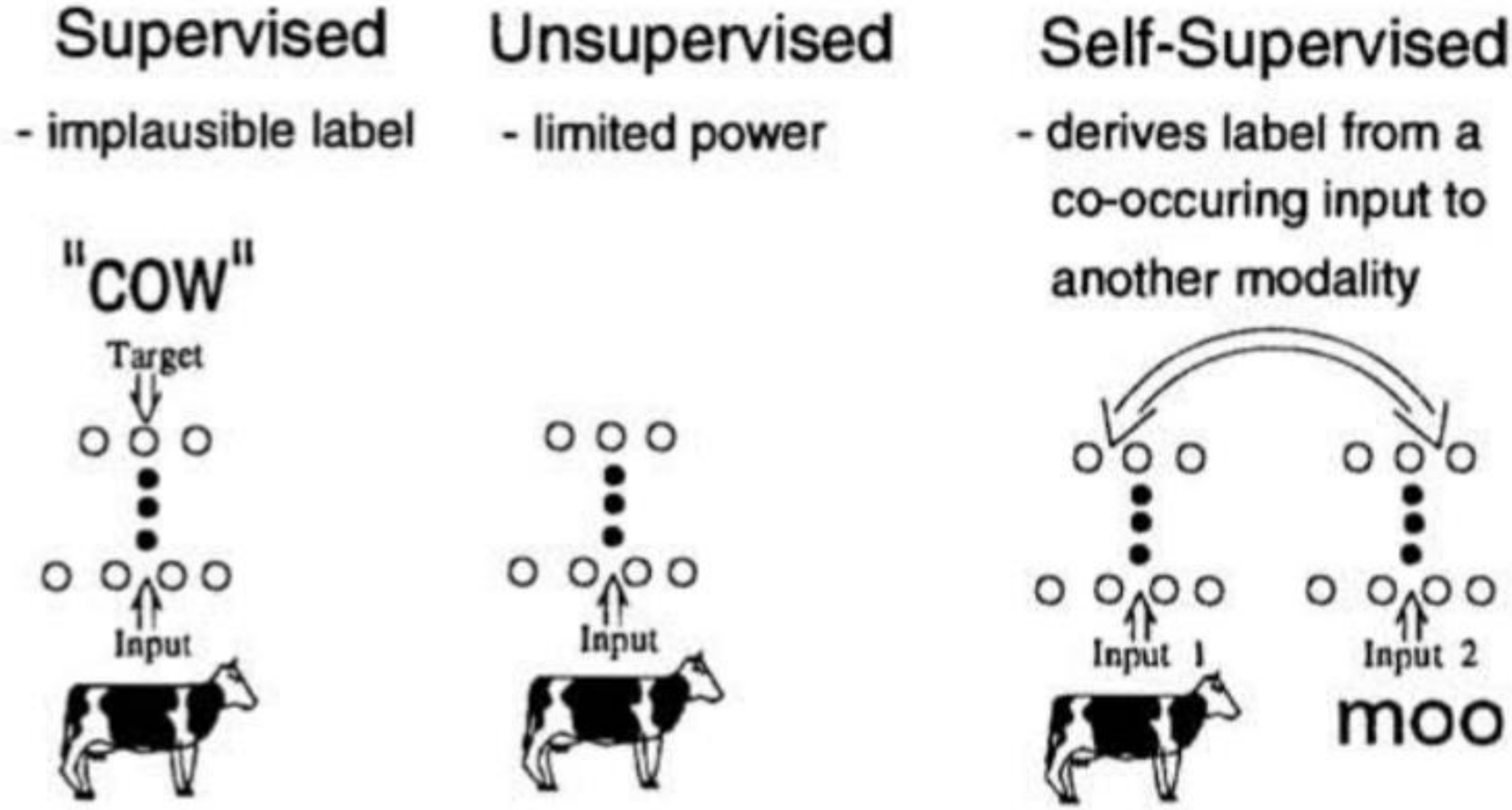
- What’s gonna happen next?
- What’s to my left?
- What can I touch?
- What will make a sound?
- Etc.

Self-Supervision



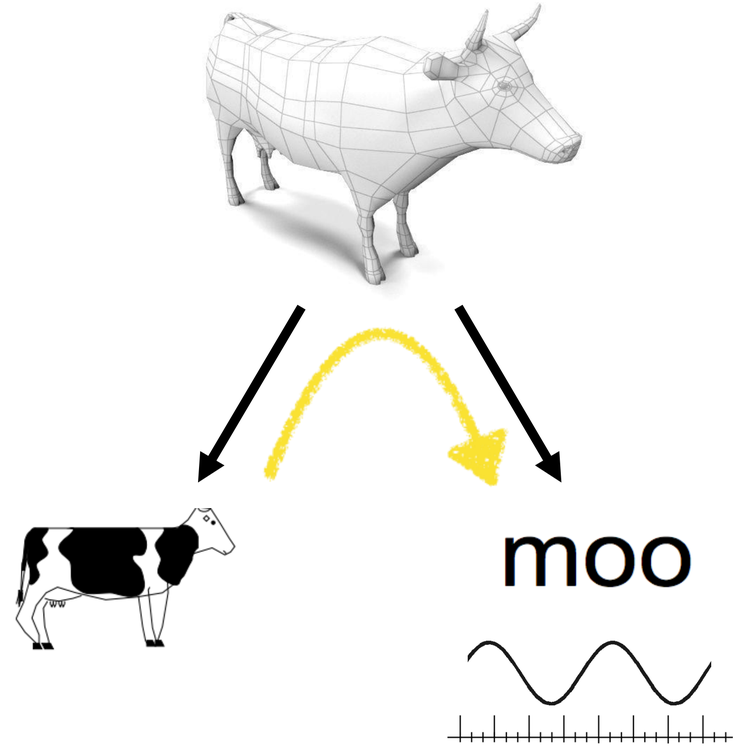
Drawing Hands, M.C. Escher, 1948

Self-Supervision in Multisensory Learning



Virginia De Sa, “Learning classification from unlabelled data”, NIPS 1994

The allegory of the cave



word2vec

I parked the car in a nearby street. It is a red car with two doors, …

I parked the vehicle in a nearby street…

T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013

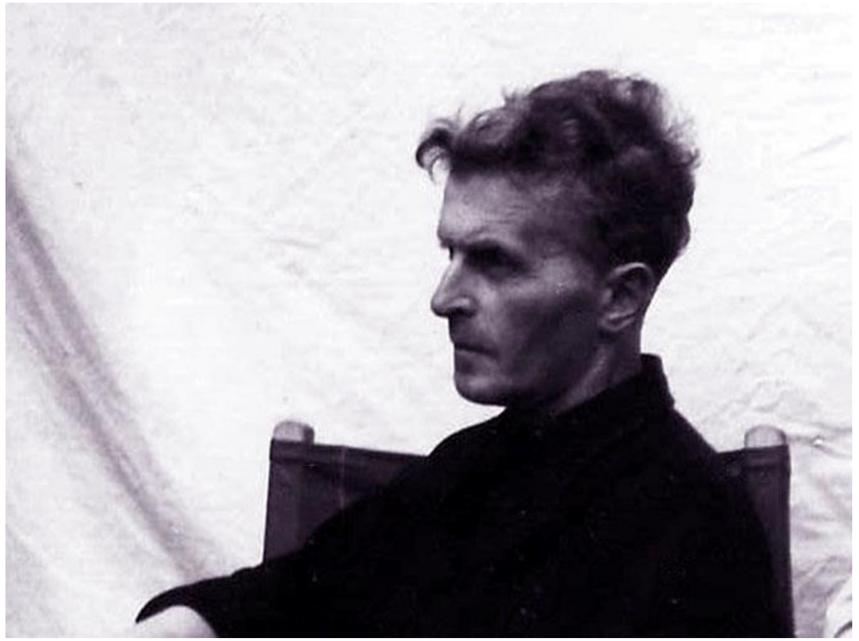
word2vec

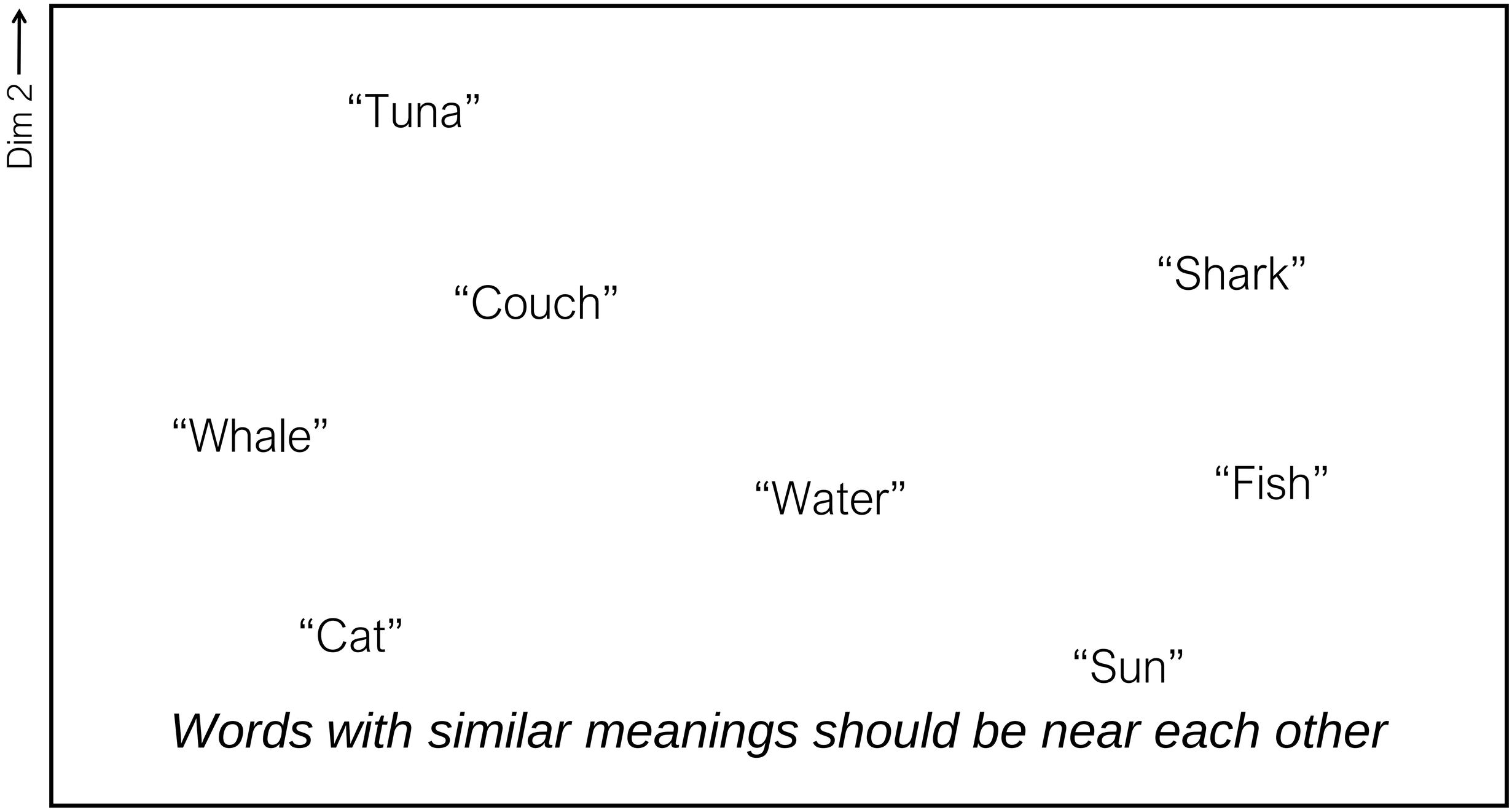
Words with similar meanings should be near each other

Proxy: words that are used in the same context tend to have similar meanings

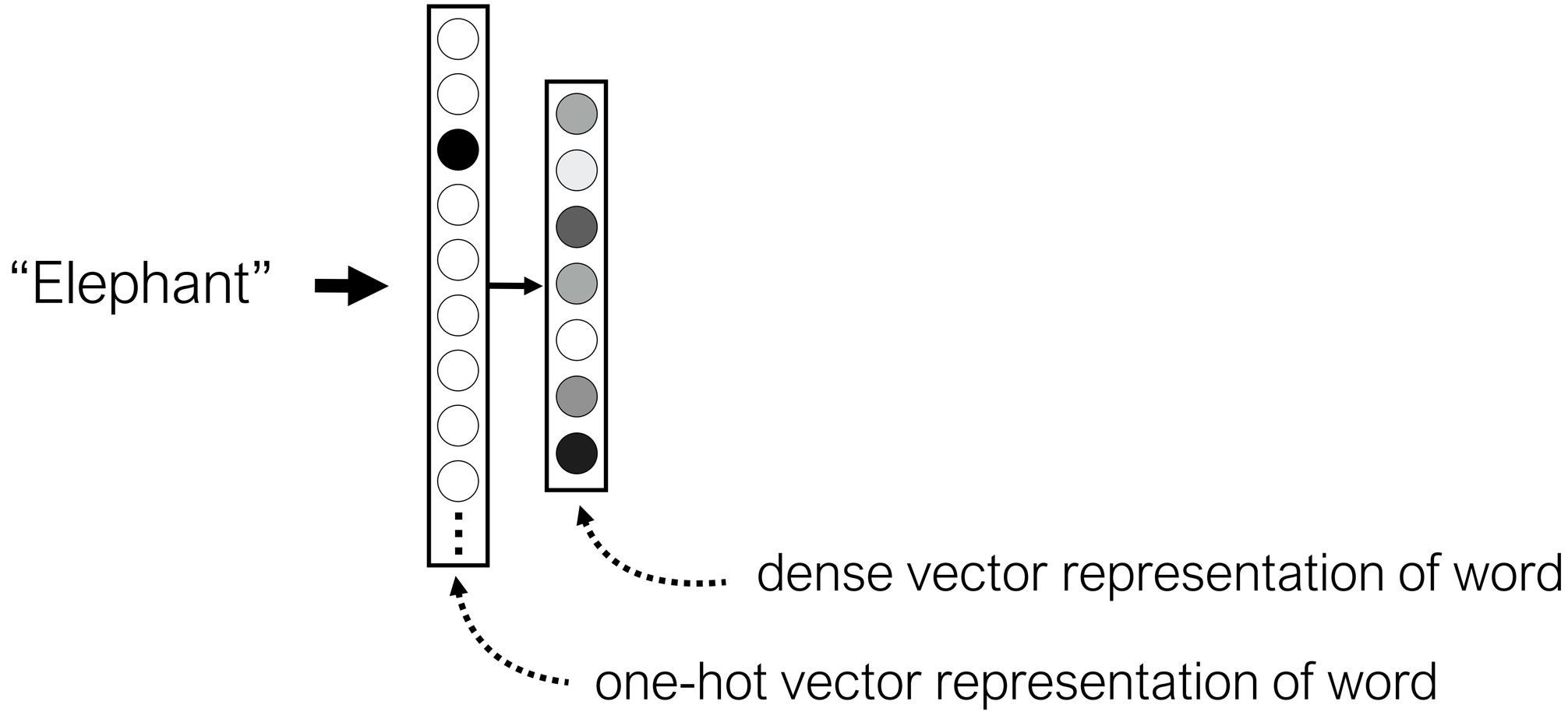
words with similar contexts should be near each other

“Meaning is use” — Wittgenstein




Dim 2
“Tuna”
“Couch”
“Shark”
“Whale”
“Water”
“Fish”
“Cat”
“Sun”
Words with similar meanings should be near each other
Dim 1

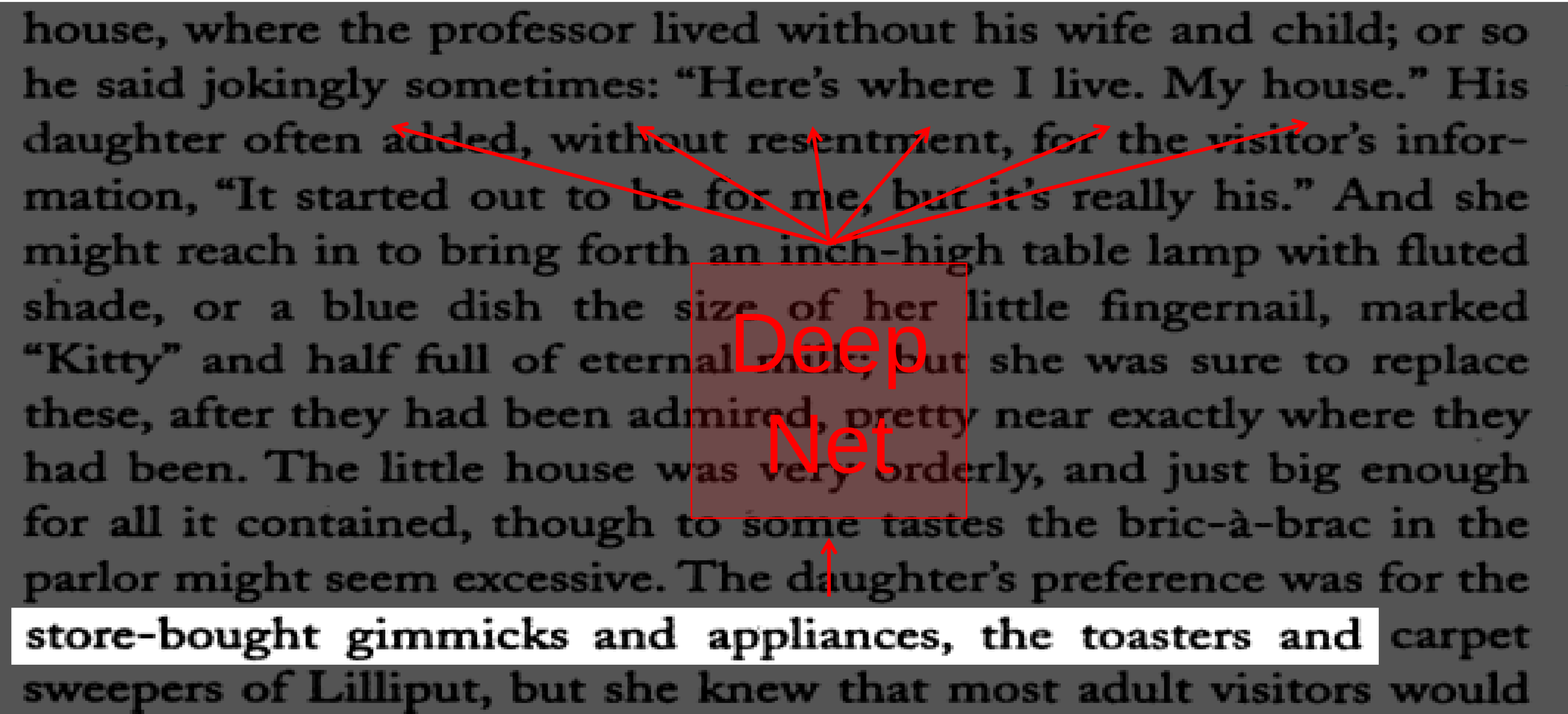
word2vec



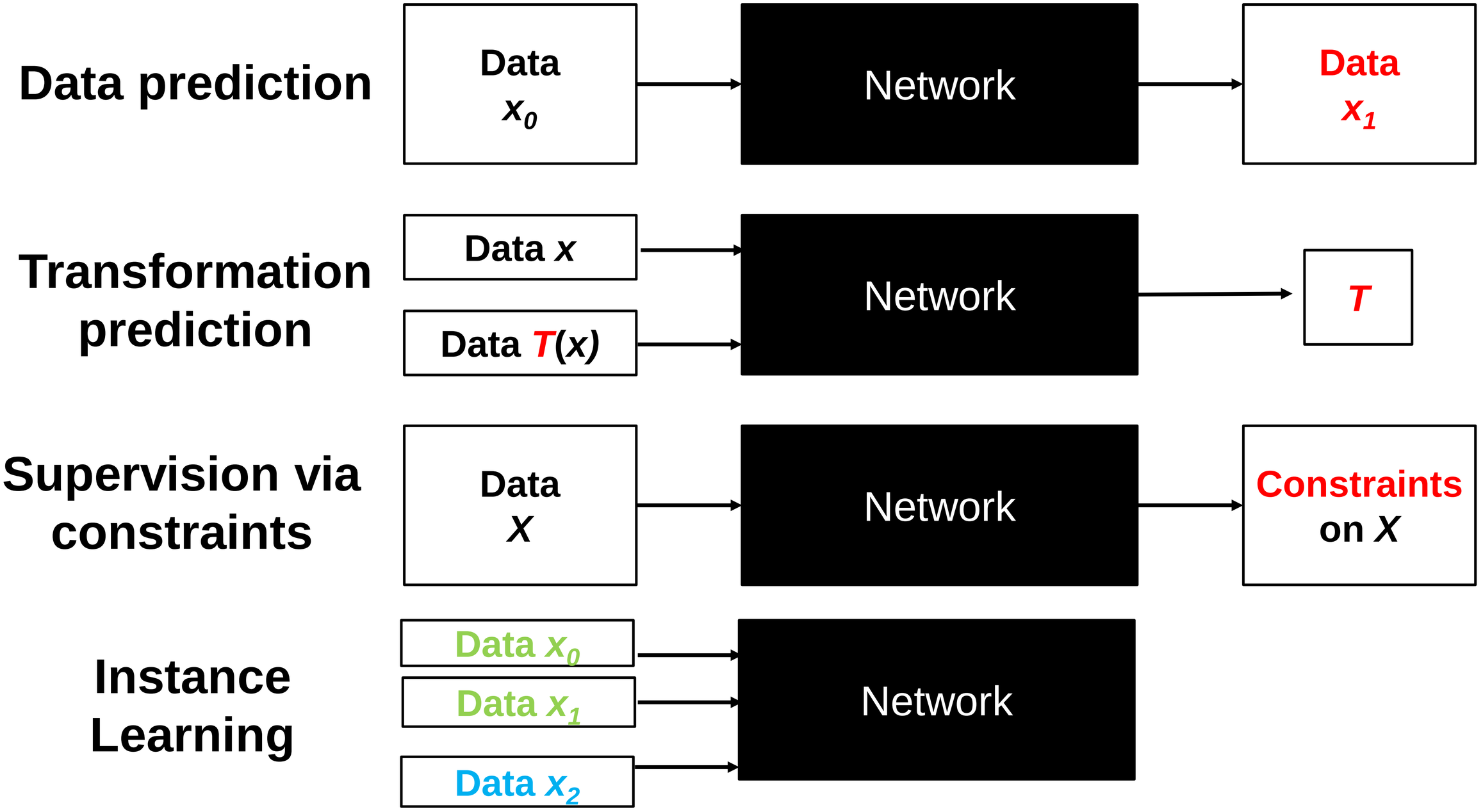
X2vec methods are also called embeddings of X, e.g., a **word embedding**

Context as Supervision

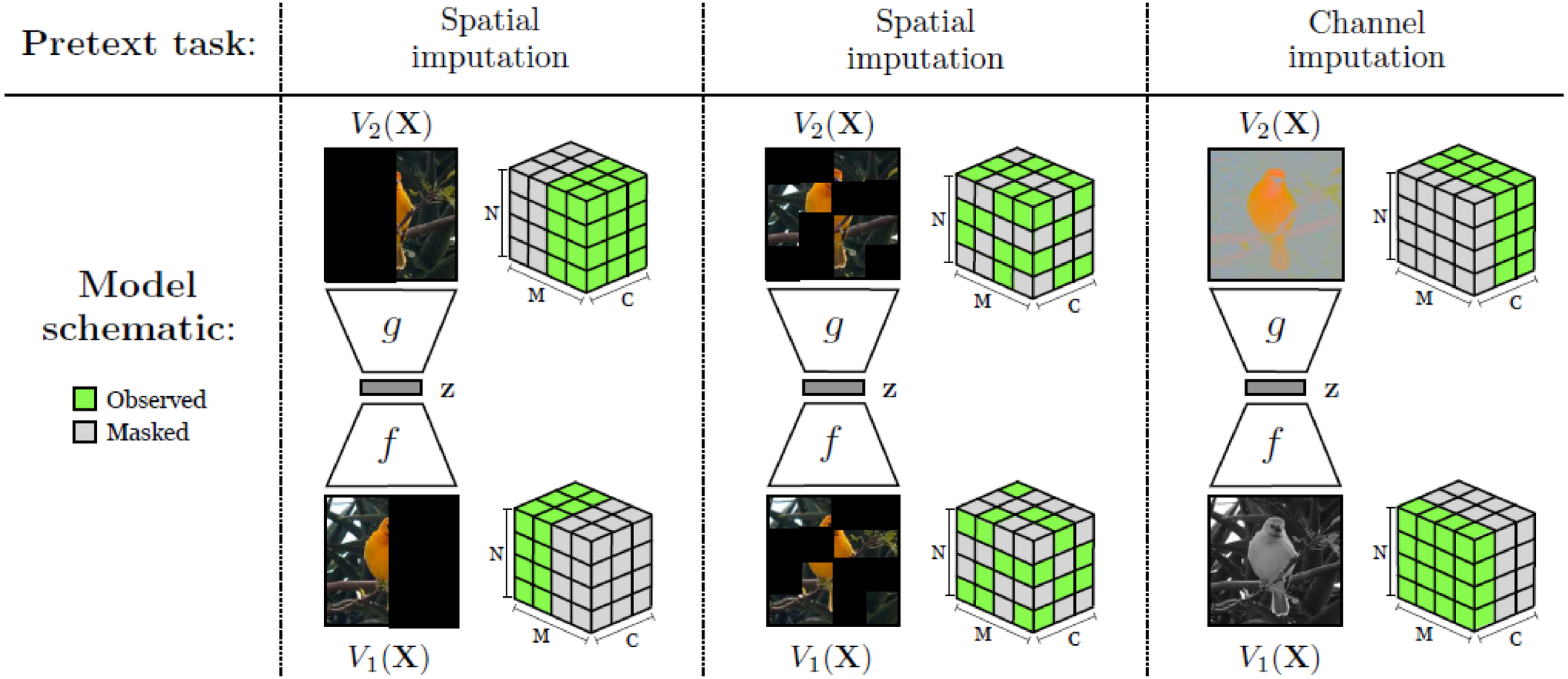
[Collobert & Weston 2008; Mikolov et al. 2013]



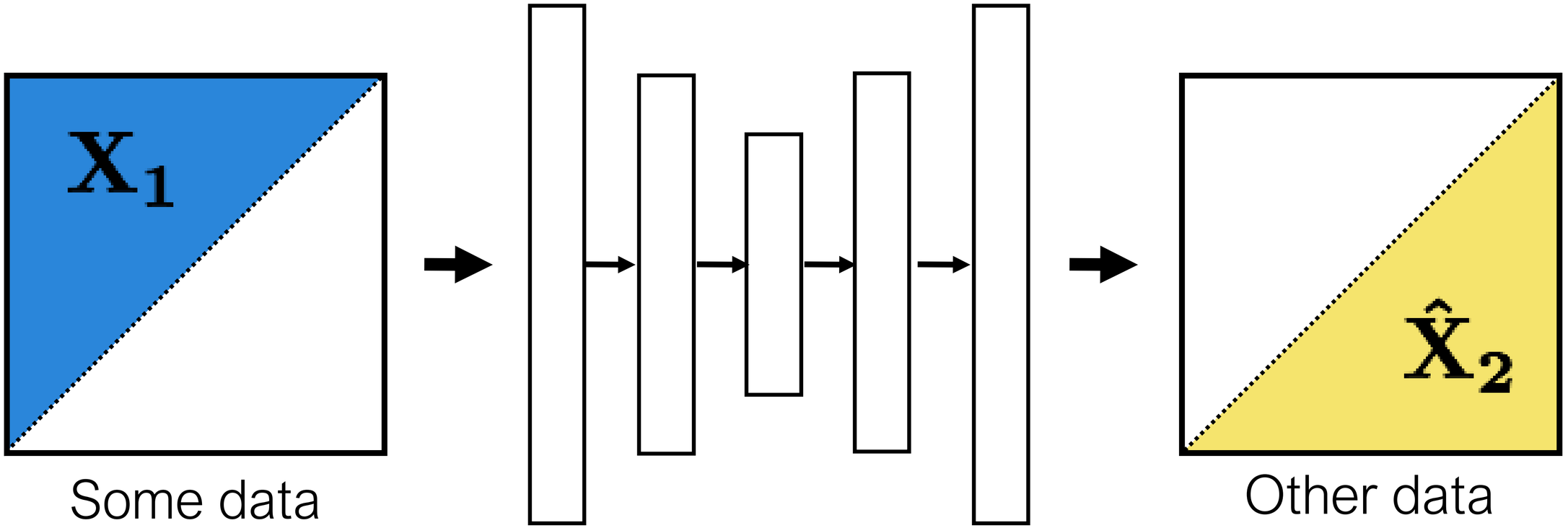
(Partial) Taxonomy of Self-Supervision

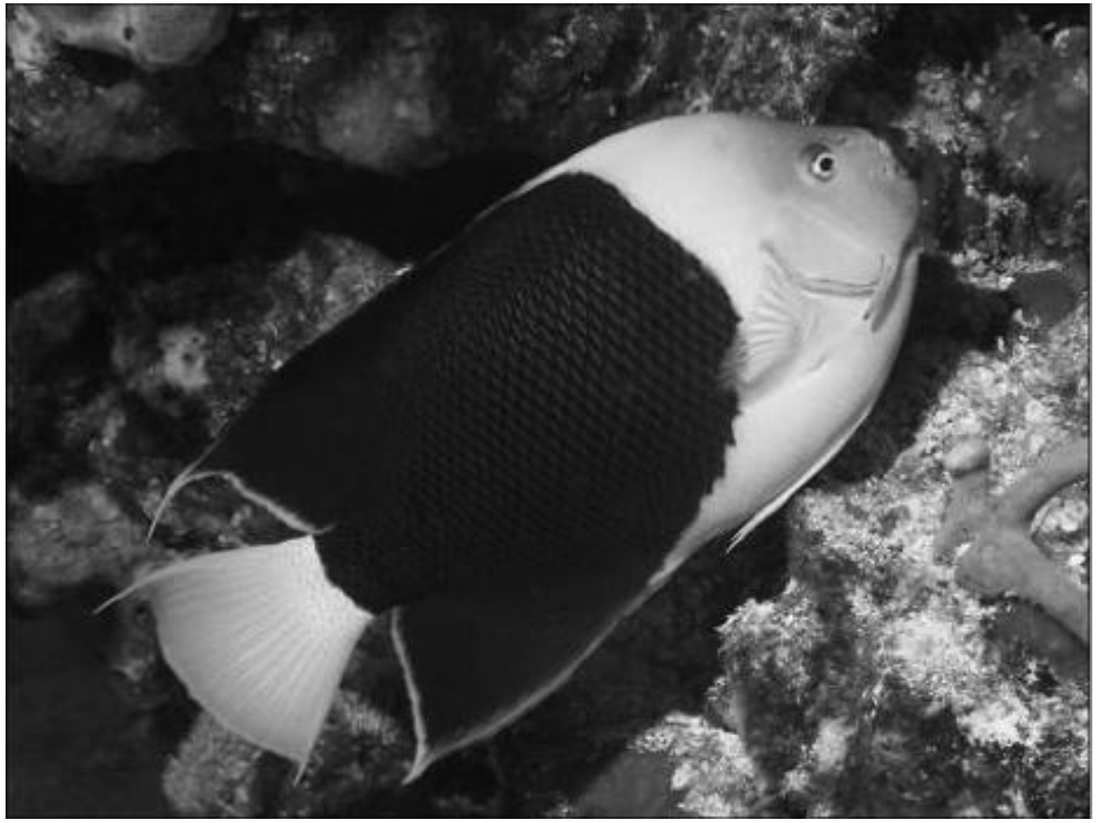


Data prediction



Self-supervision as data prediction





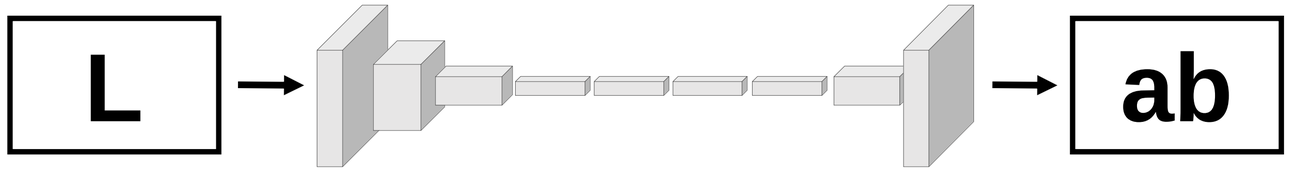
$\mathcal{F}$

Grayscale image: L channel

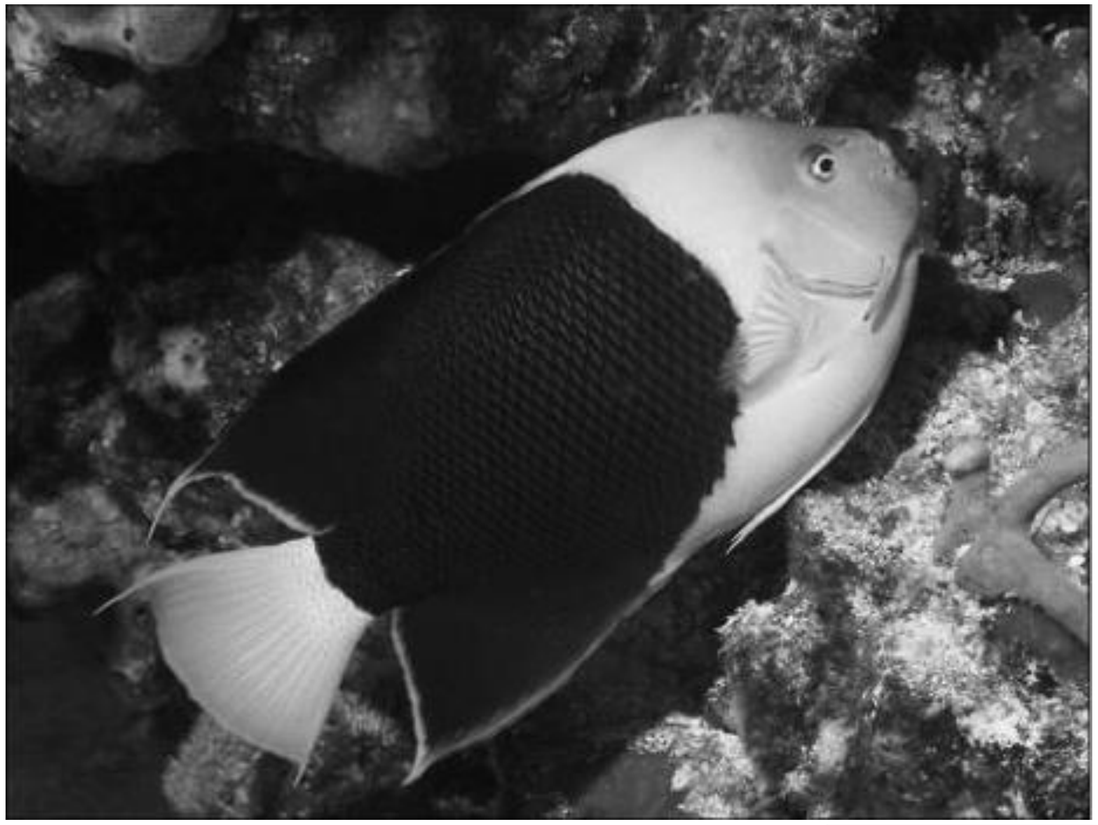
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Color information: ab channels

$\hat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$



[Zhang, Isola, Efros, ECCV 2016]

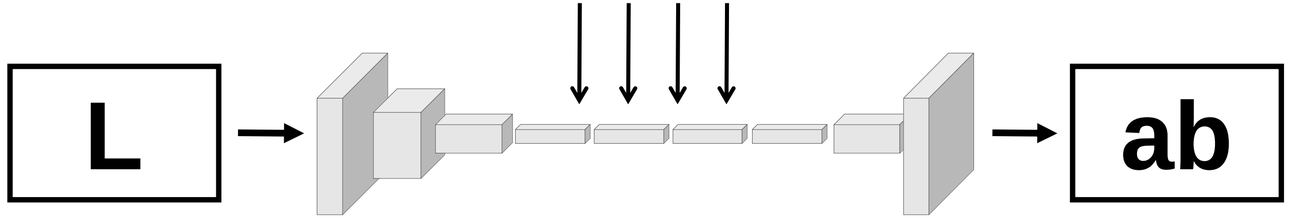


$\mathcal{F}$

Grayscale image: L chan
$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Semantics? Higher-level abstraction?

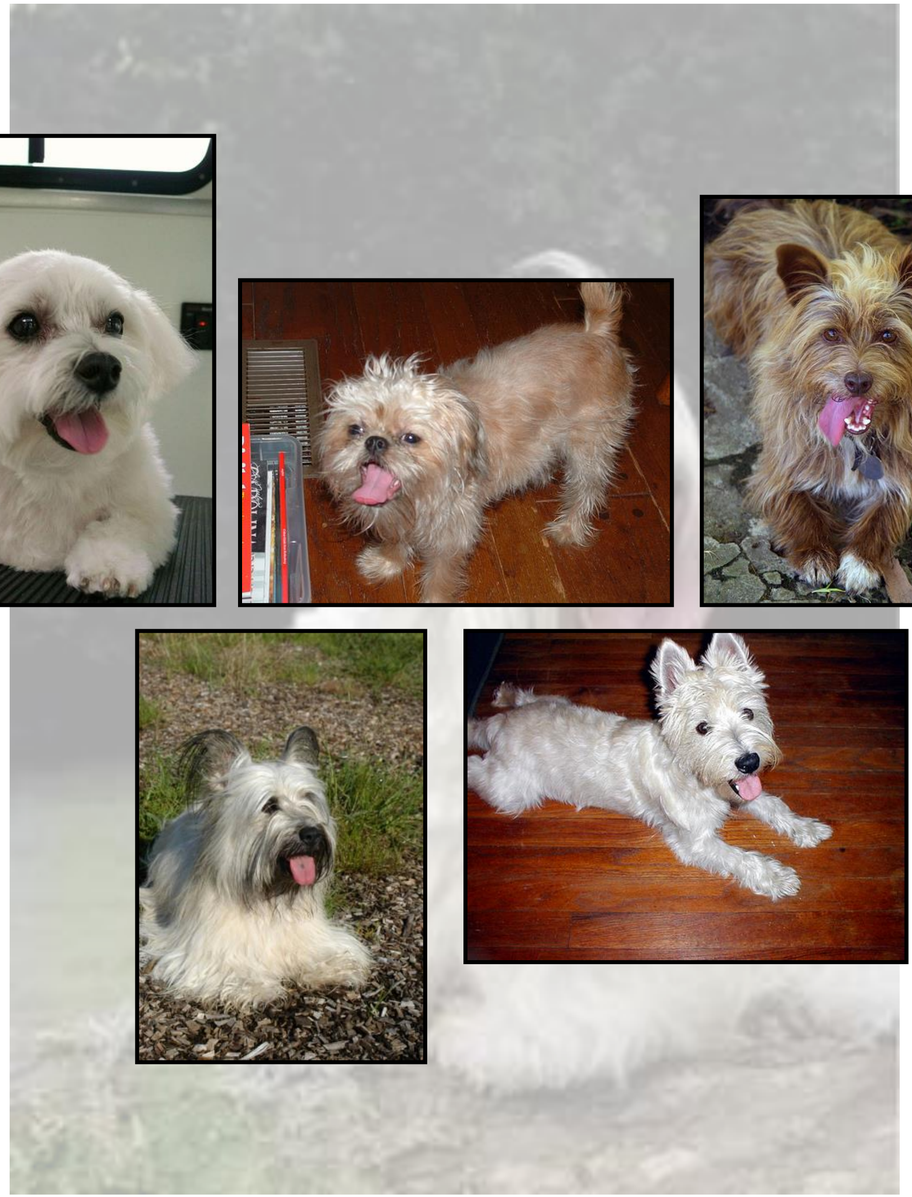
information: ab channels
$\hat{\mathbf{Y}} \in \mathbb{R}^{H \times W \times 2}$

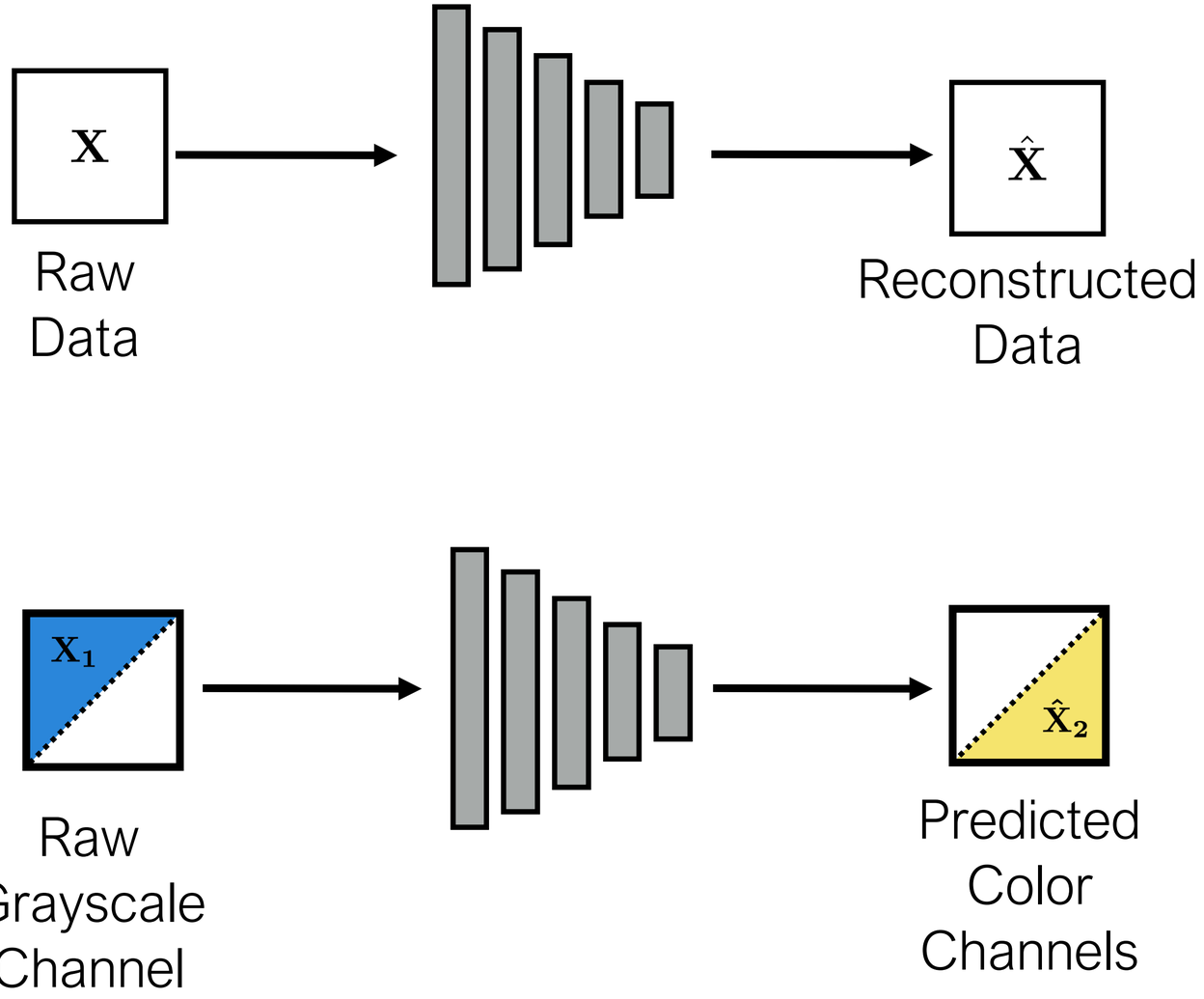


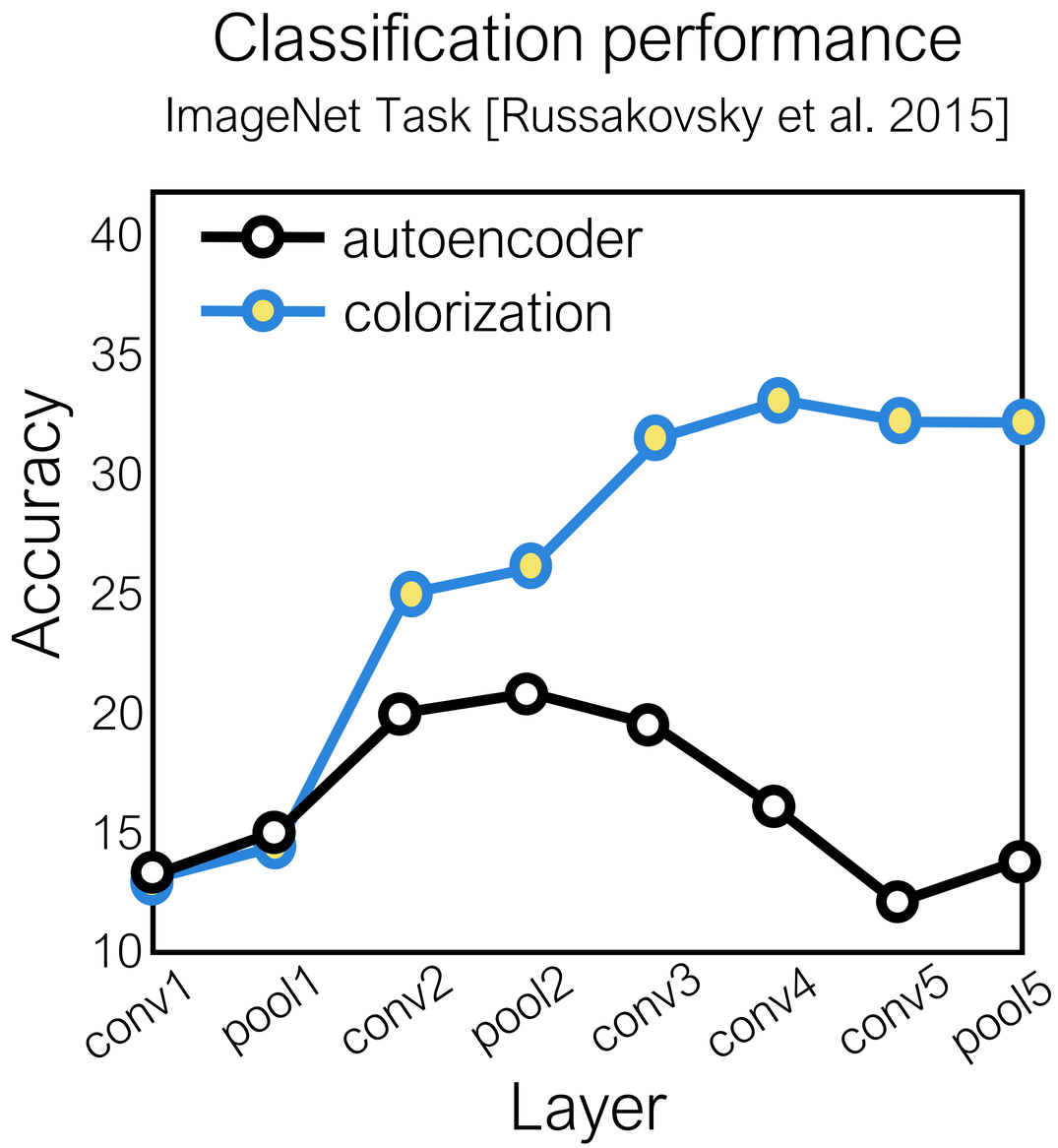
[Zhang, Isola, Efros, ECCV 2016]

Instructive failure

Instructive failure




x
Raw Data
x̂
Reconstructed Data
x1
Raw Grayscale Channel
x̂2
Predicted Color Channels


Classification performance
ImageNet Task [Russakovsky et al. 2015]
autoencoder
colorization
Accuracy
40
35
30
25
20
15
10
conv1
pool1
conv2
pool2
conv3
conv4
conv5
pool5
Layer

ImageNet Linear “Probing” – big mistake!

Dataset and Task Generalization on PASCAL [37]								
		Class. (%mAP)			Det. (%mAP)		Seg. (%mIU)	
fine-tune layers	**[Ref]**	**fc8**	**fc6-8**	**all**	**[Ref]**	**all**	**[Ref]**	**all**
ImageNet [38]	-	76.8	78.9	79.9	[36]	56.8	[42]	48.0
Gaussian	[10]	–	–	53.3	[10]	43.4	[10]	19.8
Autoencoder	[16]	24.8	16.0	53.8	[10]	41.9	[10]	25.2
k-means [36]	[16]	32.0	39.2	56.6	[36]	45.6	[16]	32.6
Agrawal et al. [8]	[16]	31.2	31.0	54.2	[36]	43.9	–	–
Wang & Gupta [15]	–	28.1	52.2	58.7	[36]	47.4	–	–
*Doersch et al. [14]	[16]	44.7	55.1	**65.3**	[36]	**51.1**	–	–
*Pathak et al. [10]	[10]	–	–	56.5	[10]	44.5	[10]	29.7
*Donahue et al. [16]	–	38.2	50.2	58.6	[16]	46.2	[16]	34.9
Ours (gray)	–	**52.4**	**61.5**	**65.9**	–	46.1	–	35.0
Ours (color)	–	**52.4**	**61.5**	**65.6**	–	46.9	–	**35.6**

ion **Table 2.** PASCAL Tests

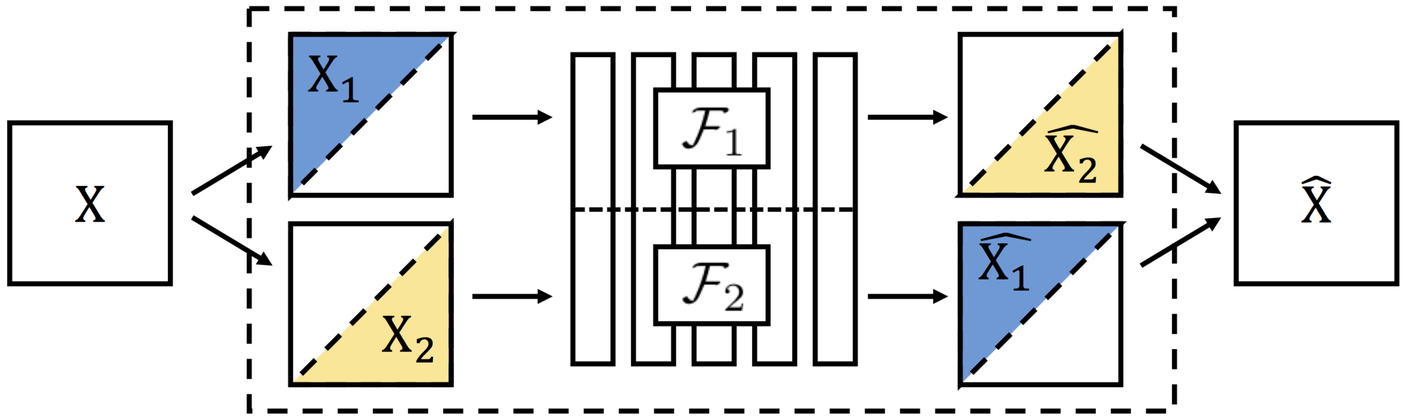
The “probe” has 2000 * 1000 = 2,000,000 parameters!

[Zhang, Isola, Efros, ECCV 2016]

Self-supervision as data prediction

Context Encoder
Pathak et al. CVPR 2016

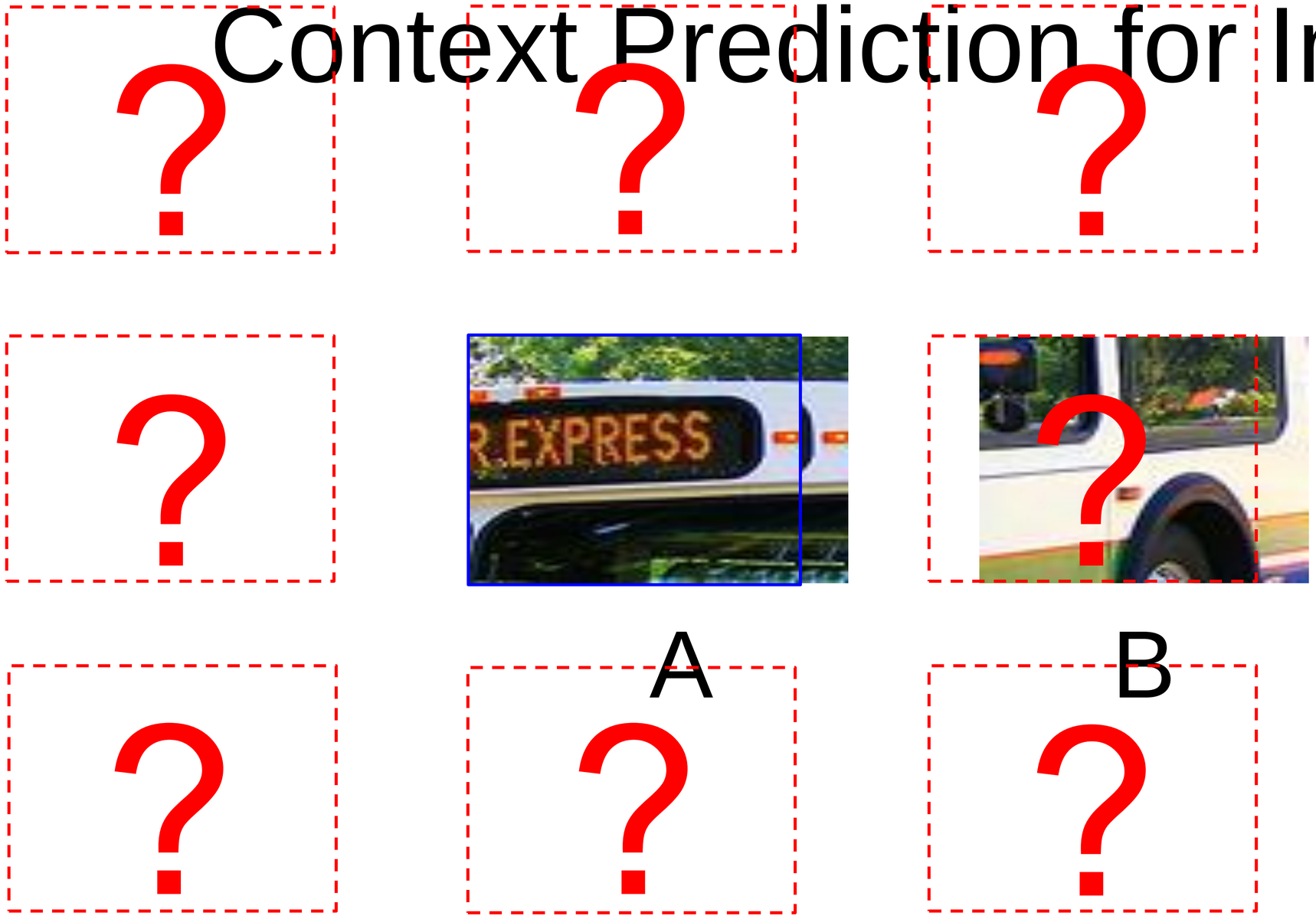
Colorization
Zhang et al, ECCV 2016



Split-brain Autoencoder, Zhang et al, CVPR 2017

Self-supervision as transformation predication

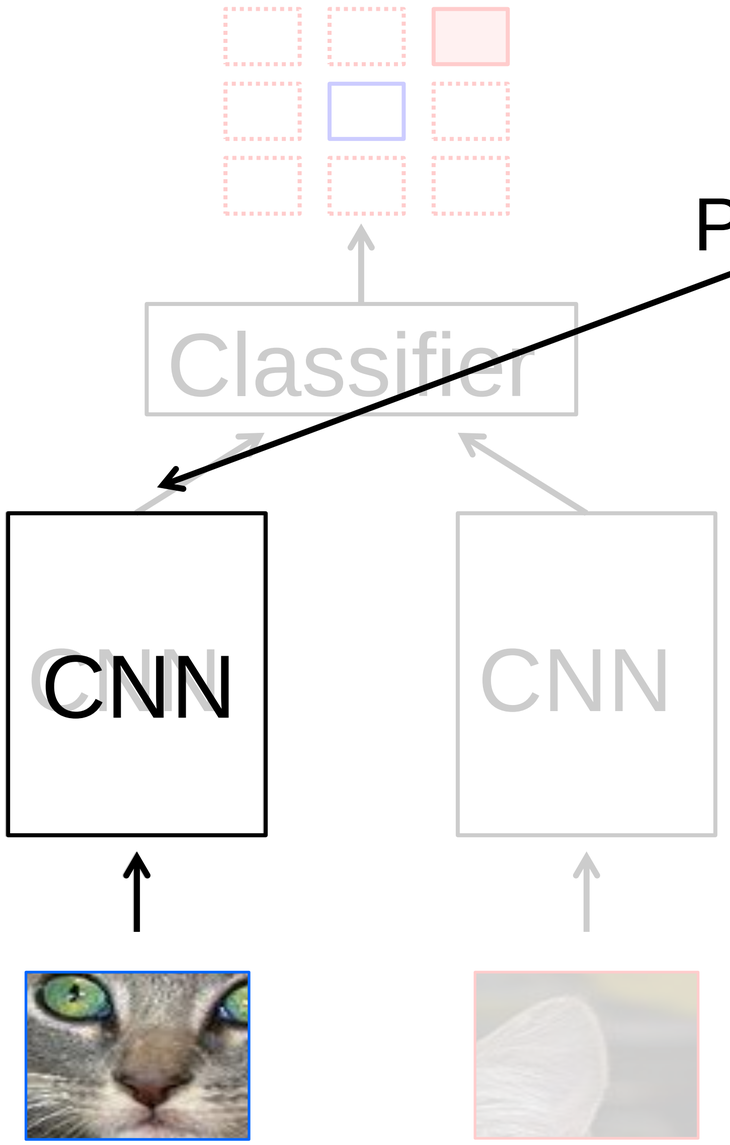
Context Prediction for Images



Deorsch, Gupta, Efros ICCV 2015

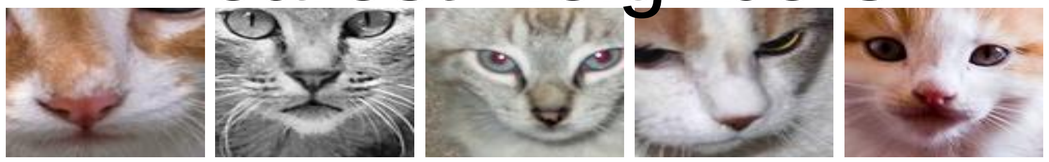
Semantics from a non-semantic task

Relative Position Task



Input

Nearest Neighbors



Correspondence ***across*** instances!

Audio-Visual Scene Analysis with Self-Supervised Multisensory Features

ECCV 2018

Andrew Owens Alexei A. Efros

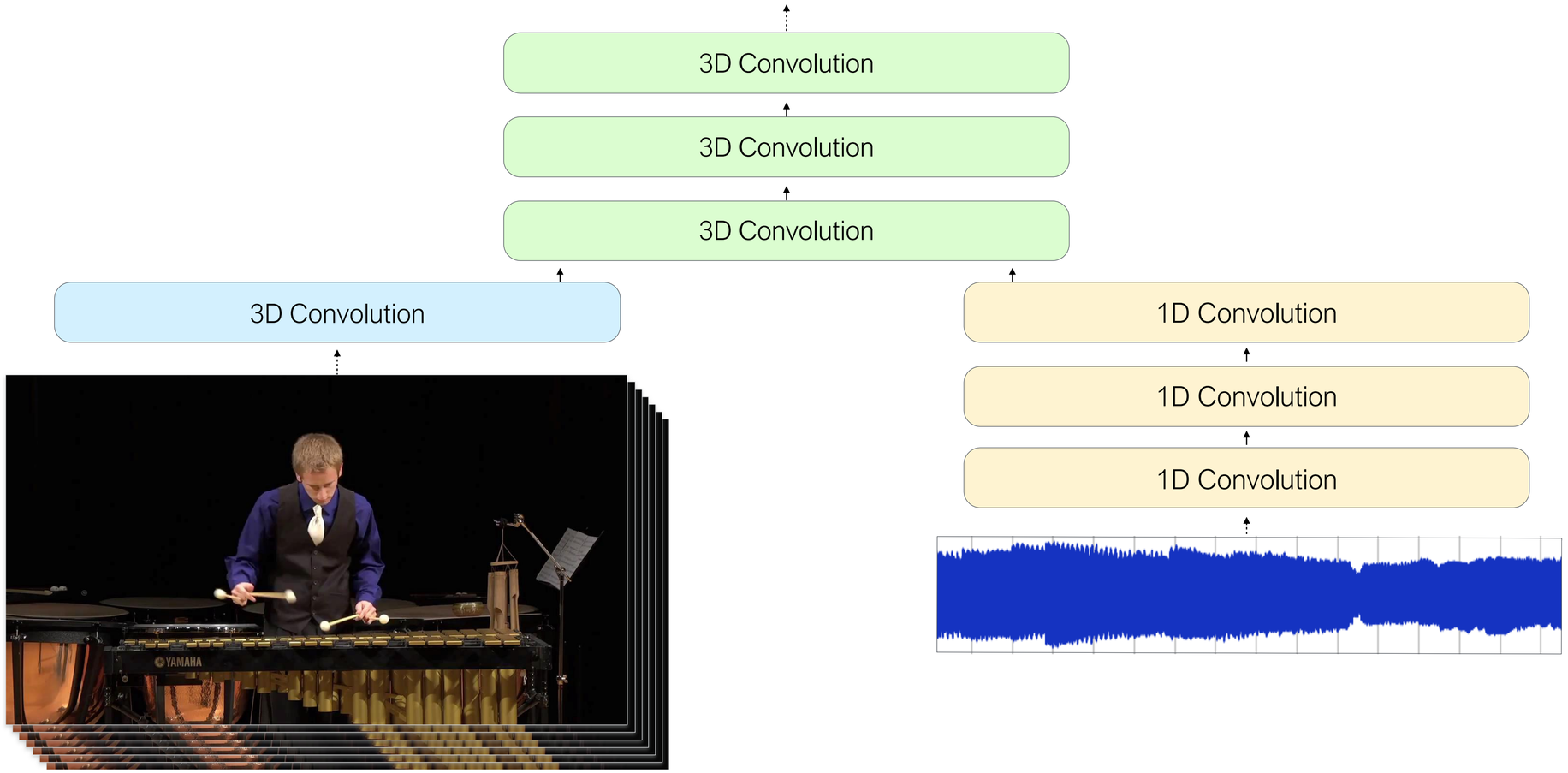
UC Berkeley

(McGurk 1976)

Same audio, different video!

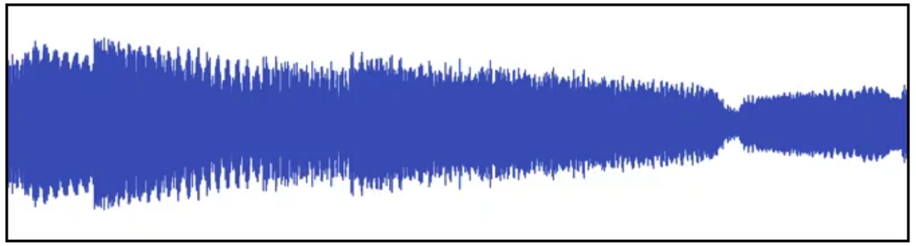
(McGurk 1976)

Multisensory representation



Learning audio-visual correspondences

(

, )

→ **real** or fake?

Related work: L^3-net (Arandjelović & Zisserman 2017), AVTS (Korbar et al. 2018)
Noise-contrastive estimation (Gutmann & Hyvarinen 2010)

Learning audio-visual correspondences

(

,)

→ real or fake?

Related work: L^3-net (Arandjelović & Zisserman 2017), AVTS (Korbar et al. 2018)
Noise-contrastive estimation (Gutmann & Hyvarinen 2010)

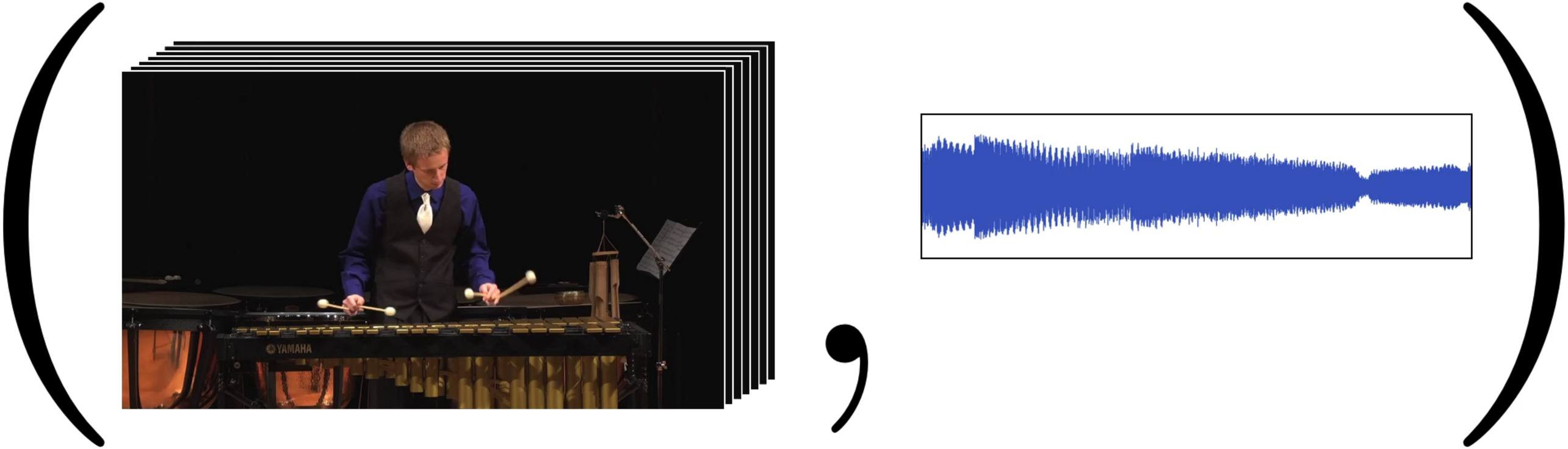
Idea #1: random pairs

,

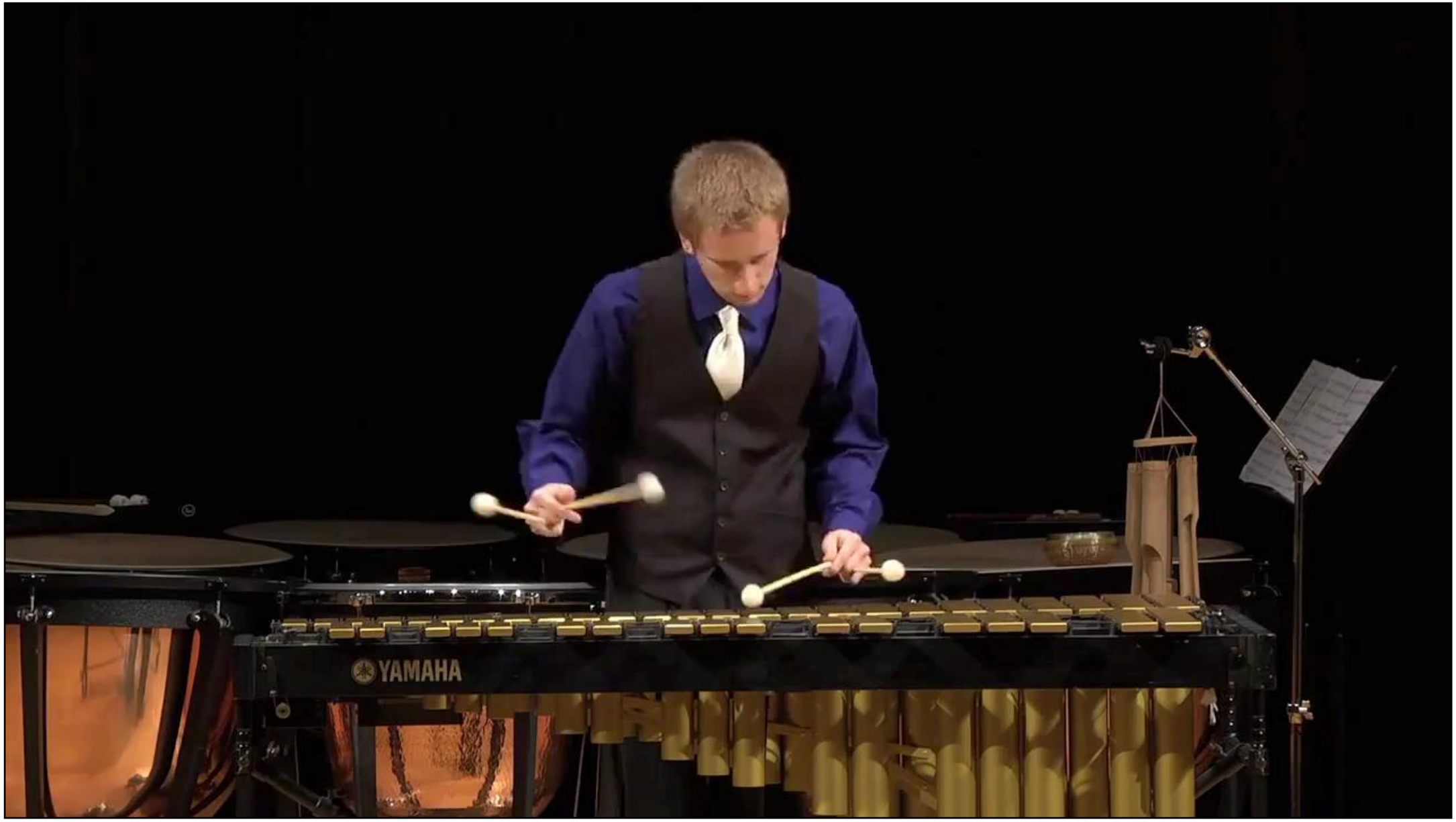
(Arandjelović & Zisserman 2017)

Idea #1: random pairs

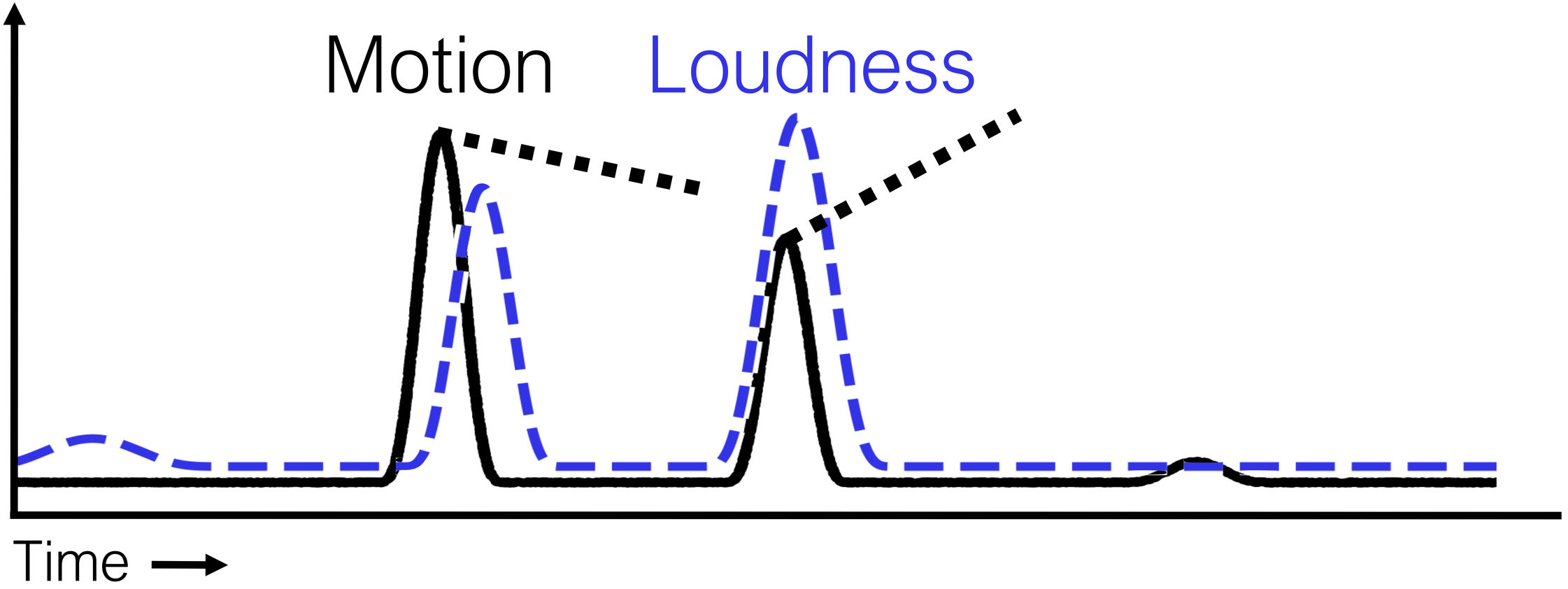
Idea #2: time-shifted pairs



Idea #2: time-shifted pairs

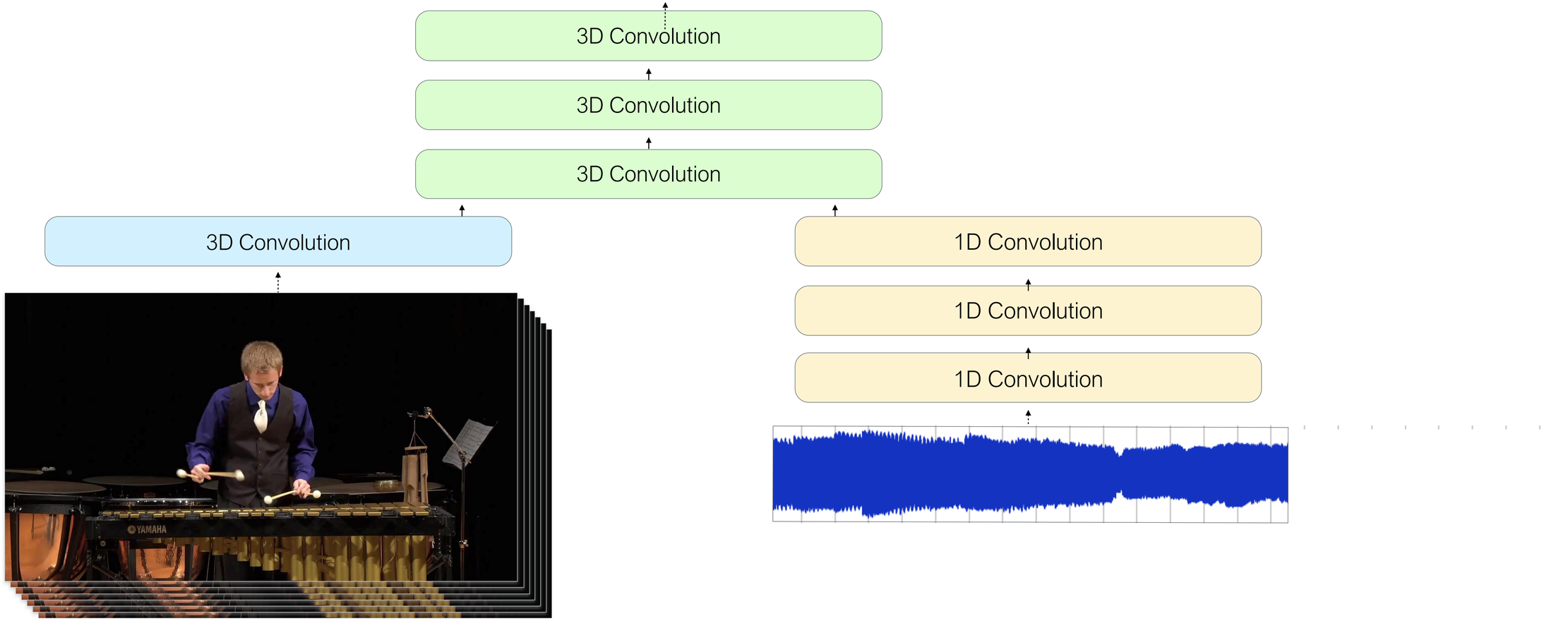


Idea #2: time-shifted pairs

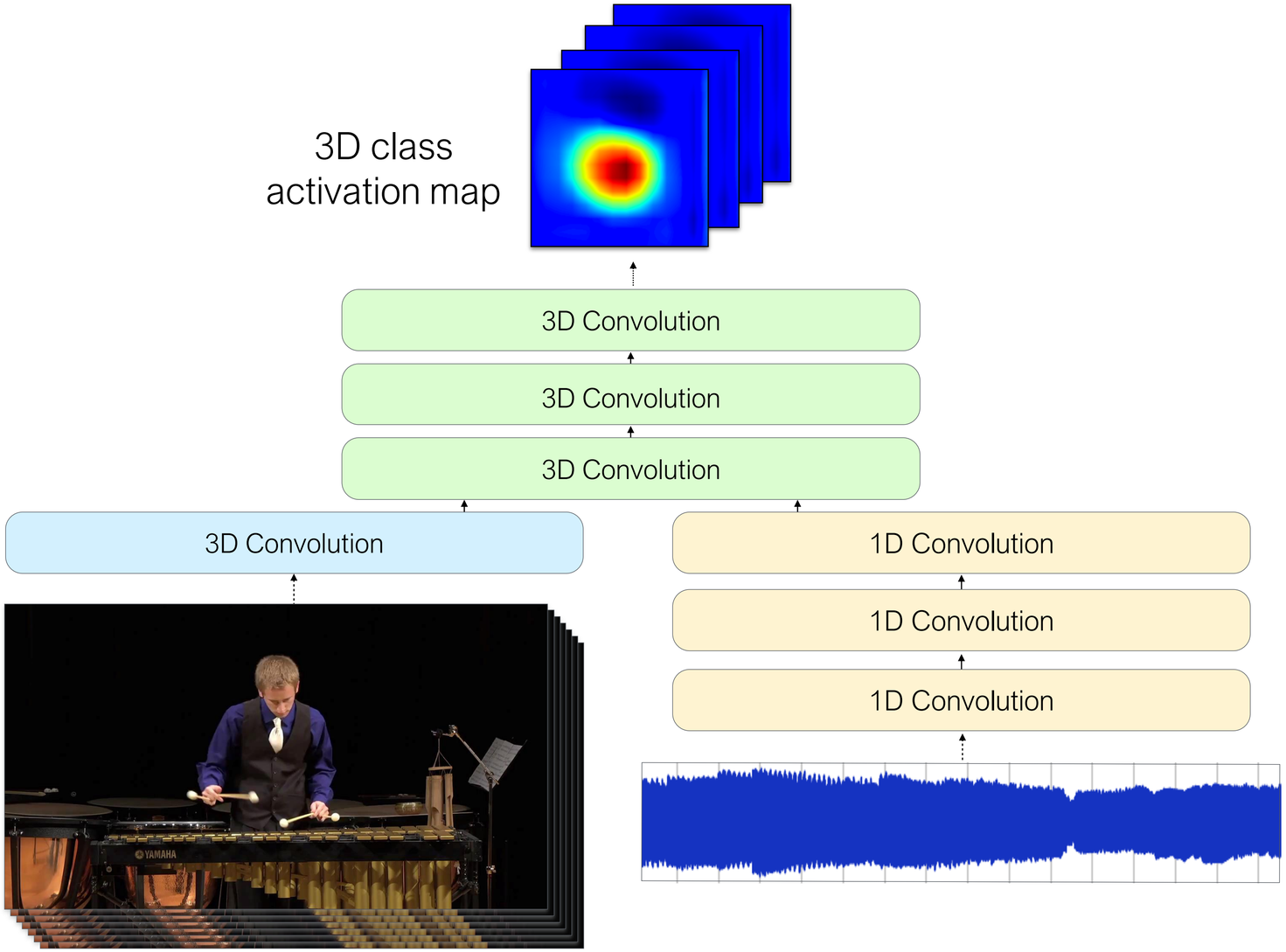


Self-supervised Training

aligned vs. not-aligned



Visualizing the location of sound sources




Top responses per category
(speech examples omitted)
Dribbling basketball

On/off-screen source separation

On/off-screen source separation


Input video
CNN


On-screen prediction
CNN


Off-screen prediction

Supervision via Constraints

Supervision via constraints

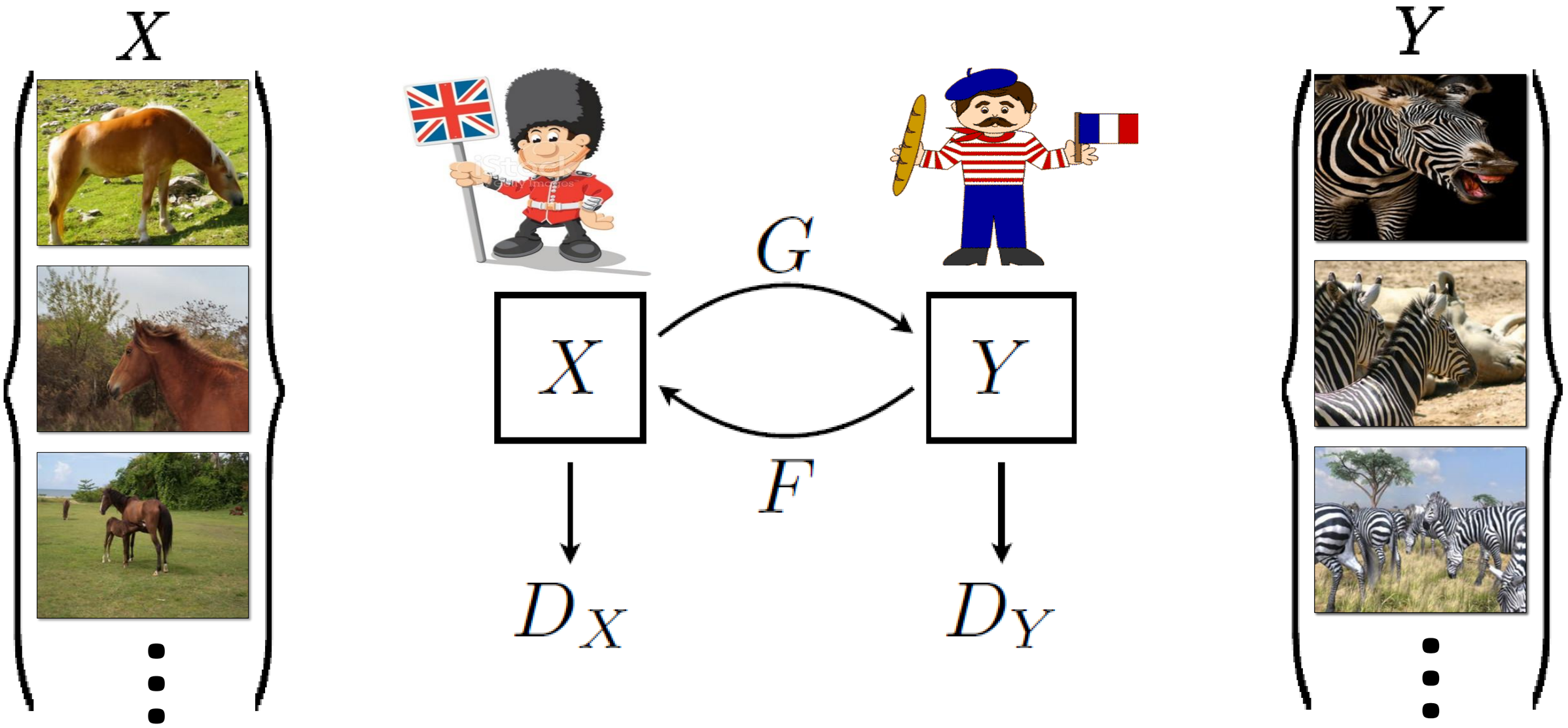
$F(x) = y$
- direct supervision

$F(x) \in Y$
- GANs

$G(F(x)) = x$
- cycle-consistency

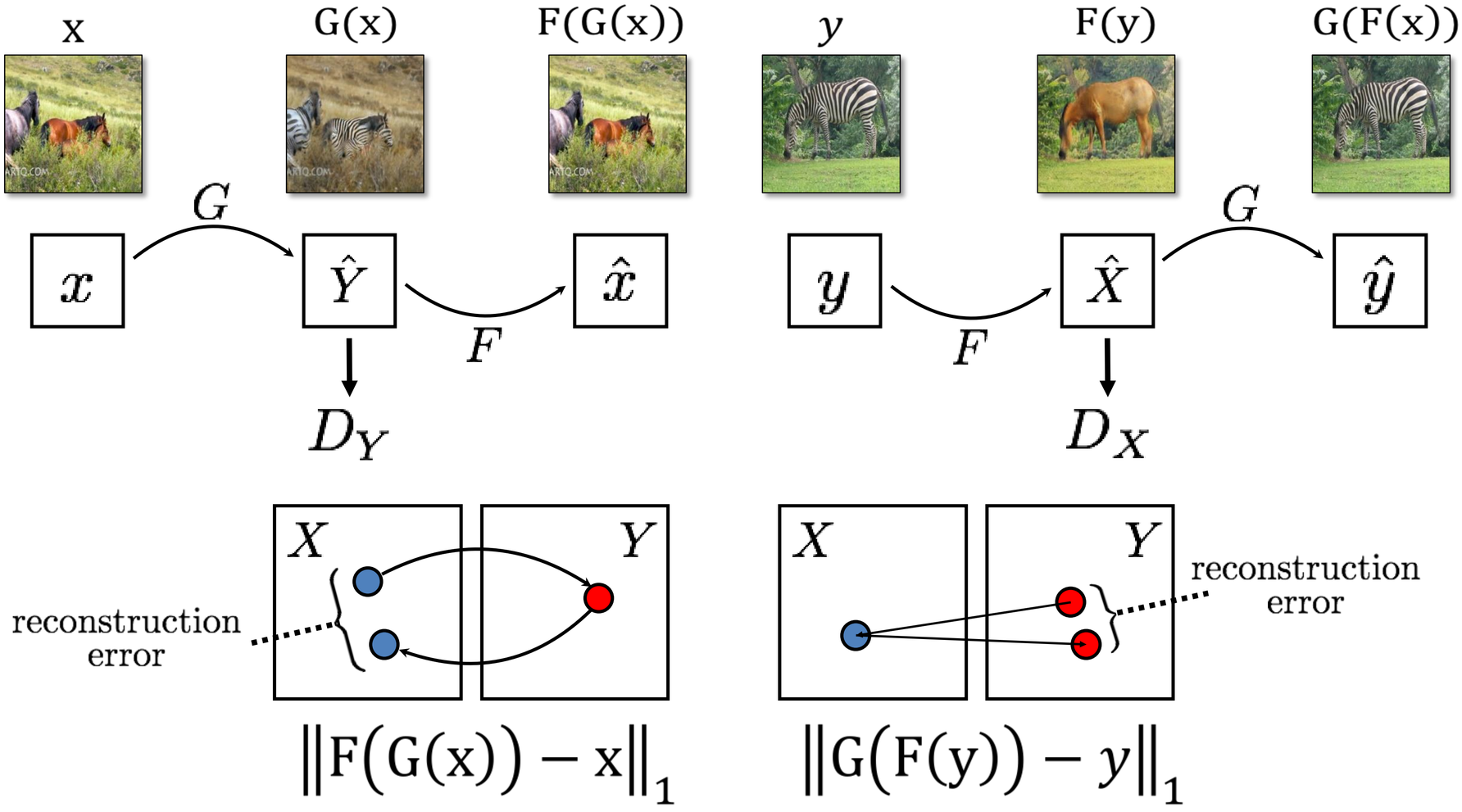
- …

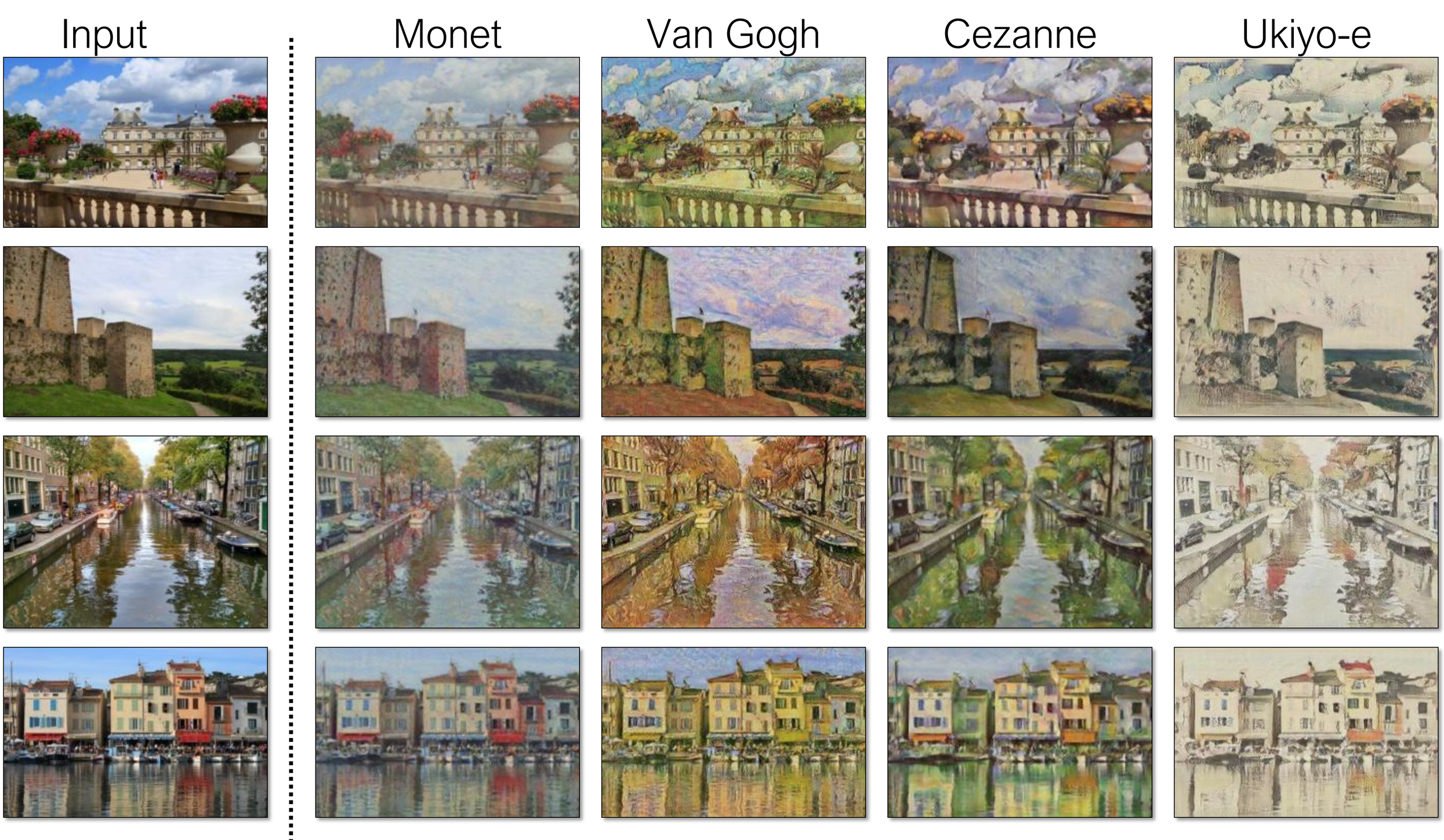
CycleGAN, or “there and back aGAN”



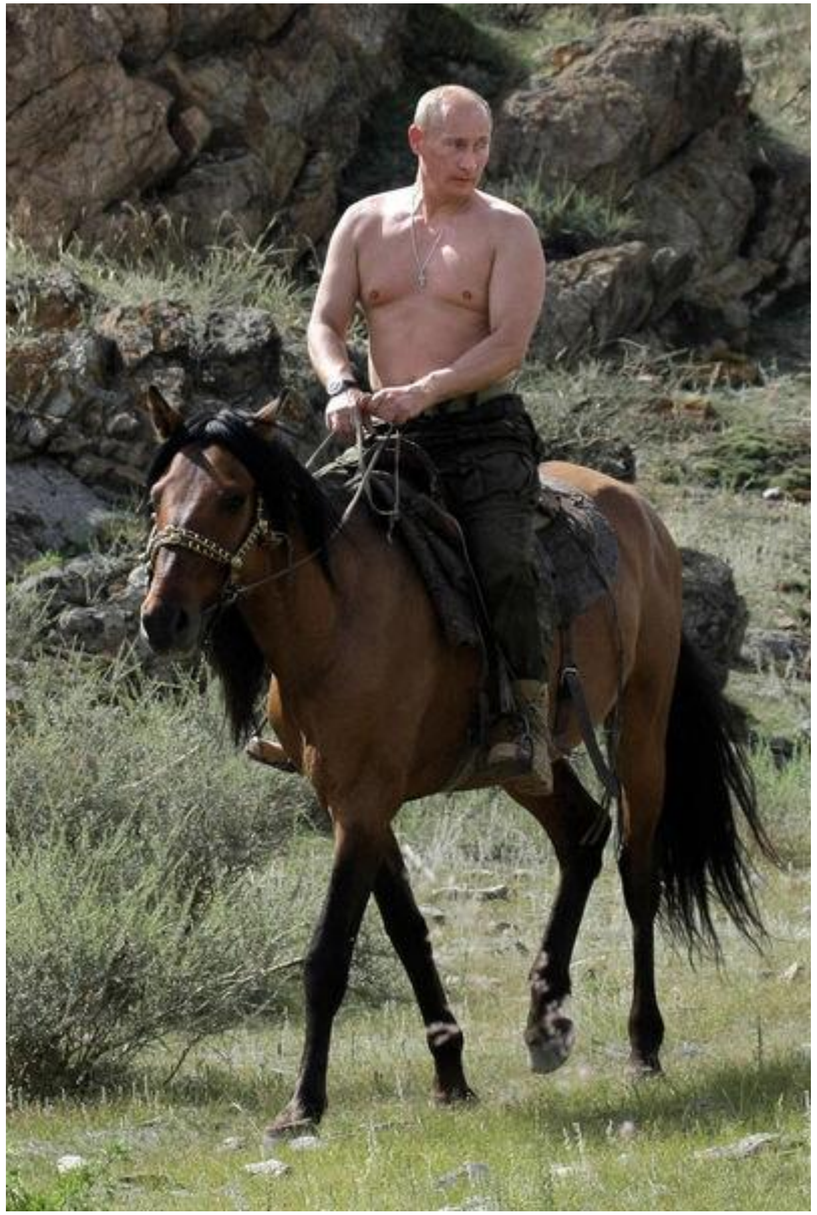
[Zhu*, Park*, Isola, Efros. ICCV 2017]

Cycle Consistency Loss



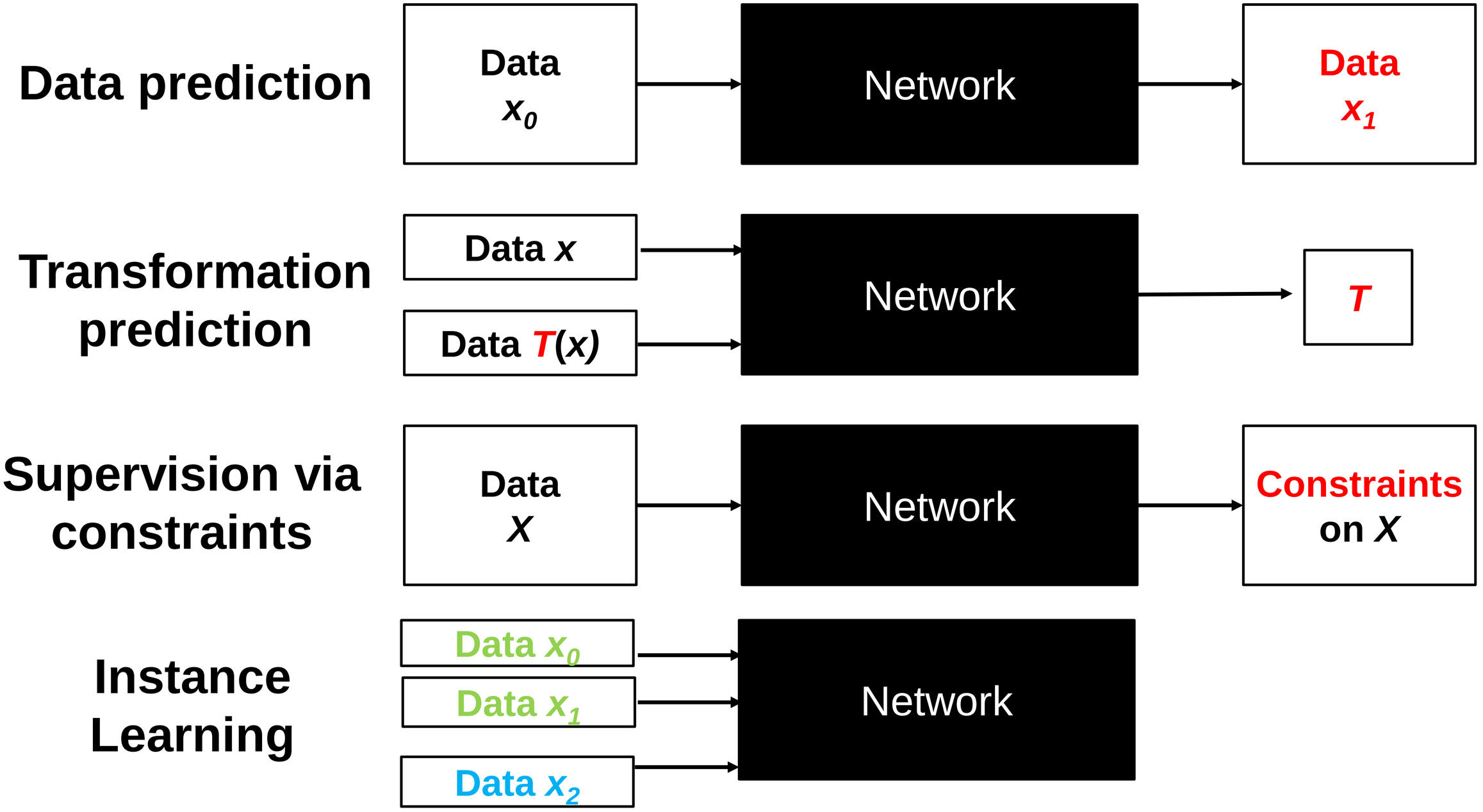

Input
Monet
Van Gogh
Cezanne
Ukiyo-e

Failure case



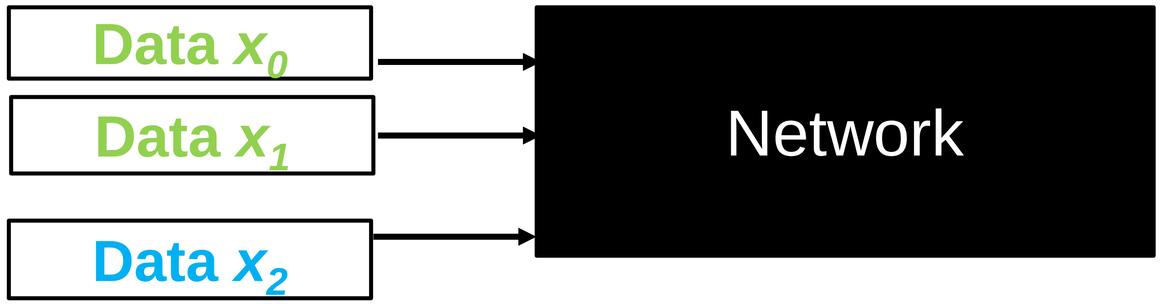
Video

(Partial) Taxonomy of Self-Supervision



Instance Learning

Instance Learning



Problem with semantic categories

- **Chair**




- **City**


With labels like these, we are setting our systems up to fail

Classical View of Categories

- Dates back to Plato & Aristotle
  1. Categories defined by a **list of properties** shared by all elements in a category
  2. Category membership is **binary**
  3. Every member in the category is **equal**

Problems with Classical View

- Humans don’t do this!
  - People don’t rely on abstract definitions / lists of shared properties (Wittgenstein 1953, Rosch 1973)
    - e.g. define the properties shared by all “games”
    - Typicality
    - e.g. Chicken -> bird, but bird -> eagle, sparrow, etc.
  - Language-dependent
    - e.g. In Russian, there is no single word for “chair”: стул, кресло, табуретка

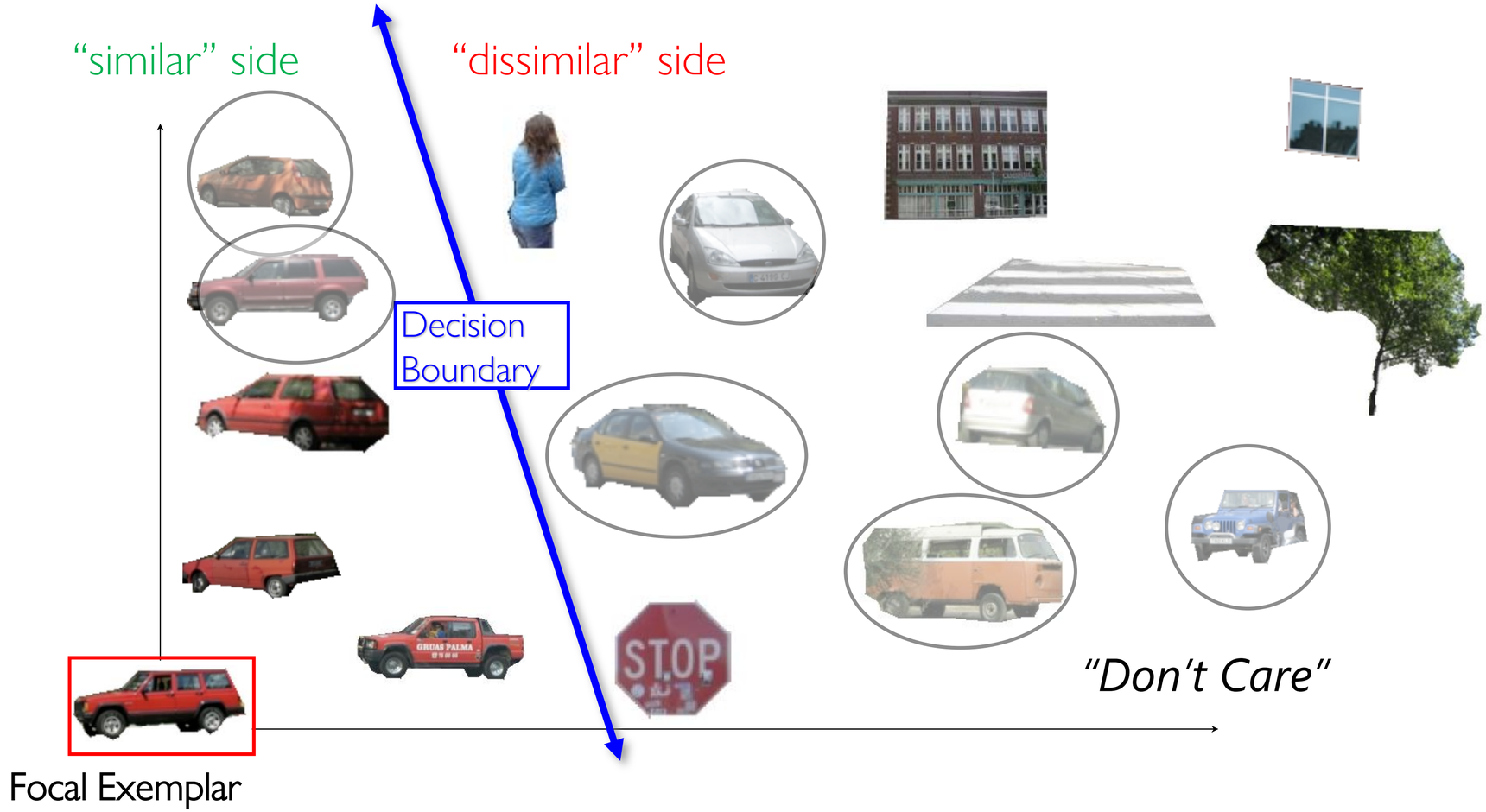
Bottom-up Association
instead of **Top-down Categorization**

- Porotype theory (Rosch, 1973)
- Exemplar Theory (Medin & Schaffer 1978, Nosofsky 1986, Krushke 1992)

Ask not *“What is it?”*
Ask “What is it <u>like</u>?”

-- Moshe Bar, 2008

Learning Per-Exemplar Distances (CVPR 2008)

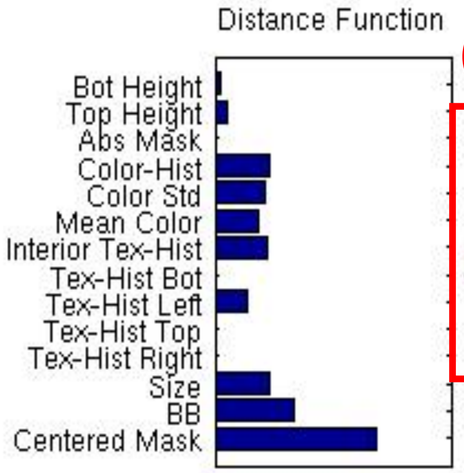


Malisiewicz & Efros. *Recognition by Association via Learning Per-exemplar Distances.* CVPR'0

Learning Per-Exemplar Distances (CVPR 2008)

Query

Baseline Top Nearest Neighbors



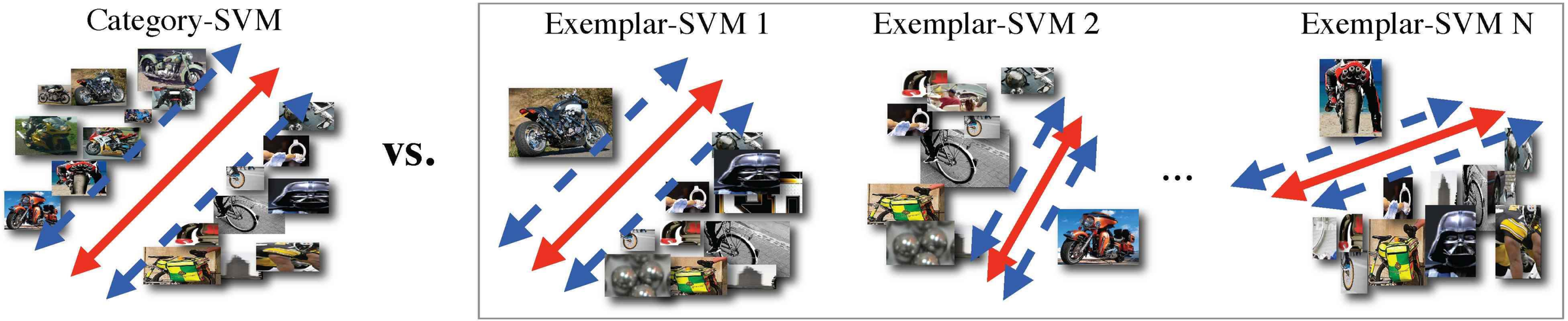
Query

Top Nearest Neighbors with Learned Dist

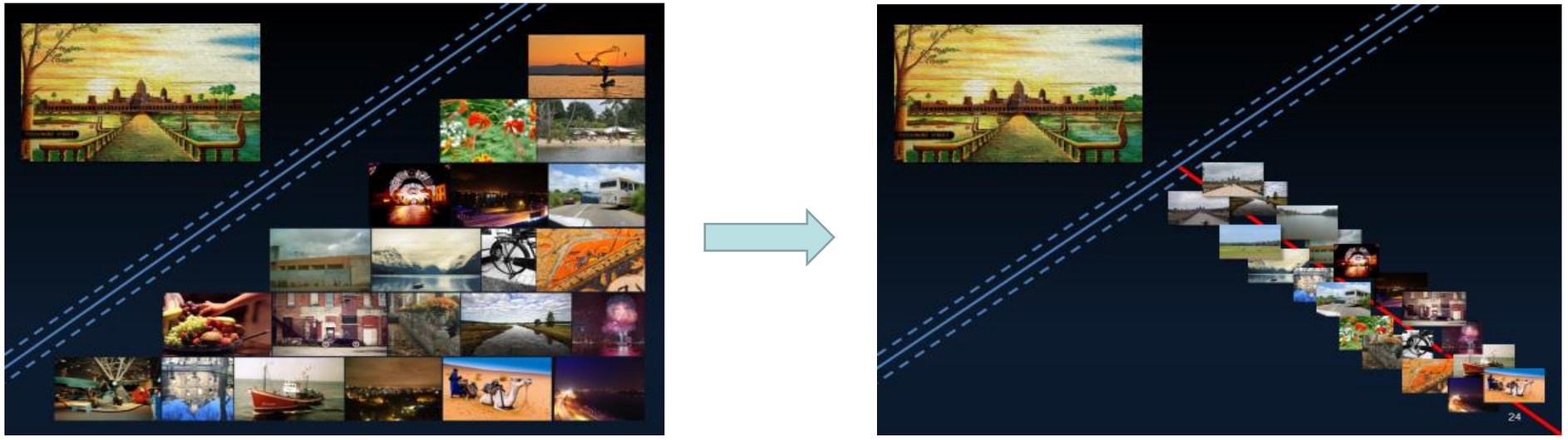
Malisiewicz & Efros. *Recognition by Association via Learning Per-exemplar Distances.* CVPR'0

Exemplar-SVM: *defining yourself by what you are not*

[Malisiewicz, Gupta, Efros, ICCV'11]



One-against-all learning for image retrieval [Srivastava et al, SIGGRAPH'11]



Exemplar-CNN [Dosovitskiy et al, NIPS'14]

- single parametric representation (CNN)
- Data augmentation

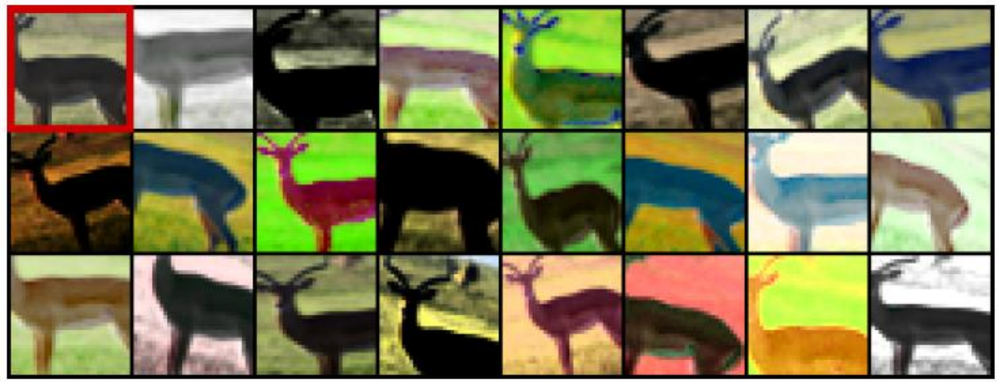
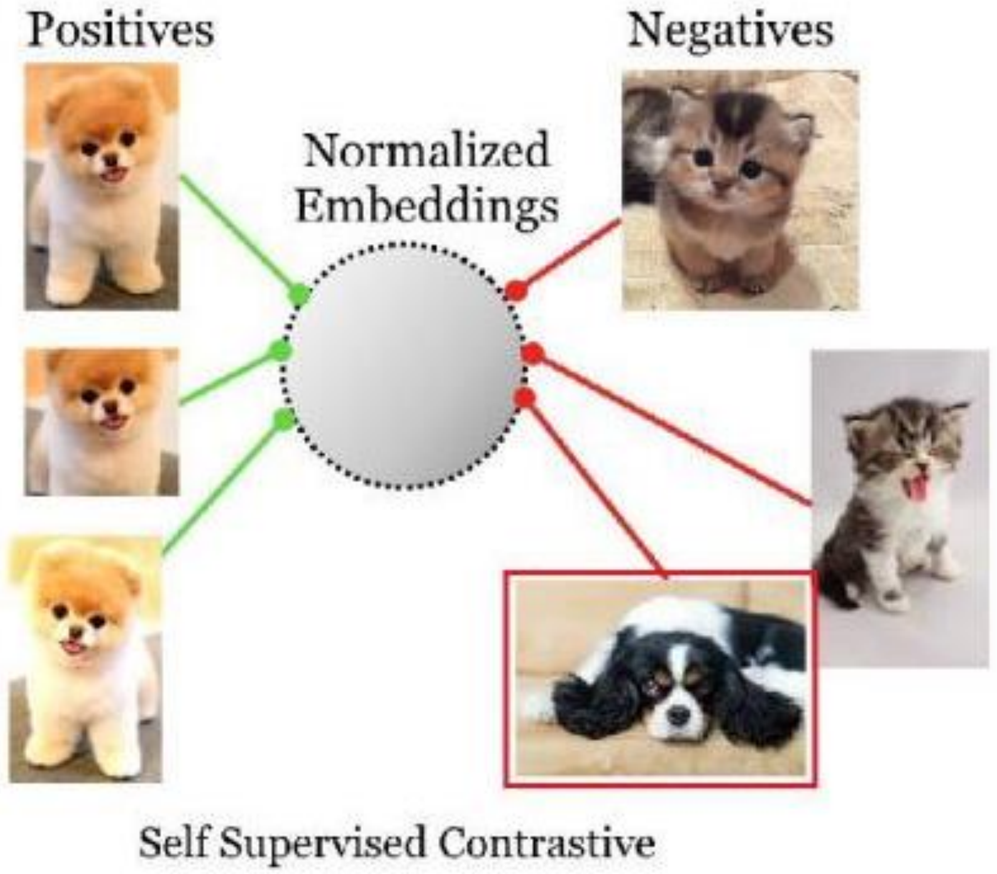


Fig. 2. Several random transformations applied to one of the patches extracted from the STL unlabeled dataset. The original ('seed') patch is in the top left corner.

Modern Day: representations via Similarity Learning

- Metric Learning
  - Siamese Nets
- Contrastive Learning
  - etc

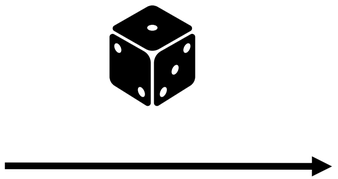


Becker et al (1992)
de Sa (1993)
Bromley et al (1994)
Chopra et al (2005)
Dosovitsky et al (2014)
Bojanowski et al (2017)
Wu et al (2018)
van den Oord et al (2019)
Tian et al (2019)
He et al (2019)
Chen et al (2020)

1. Improvements in representation learning (e.g. Contrastive)
2. **Improved Data Augmentations (e.g. cropping)**

Data Augmentation

input



color

crop

flip

blur

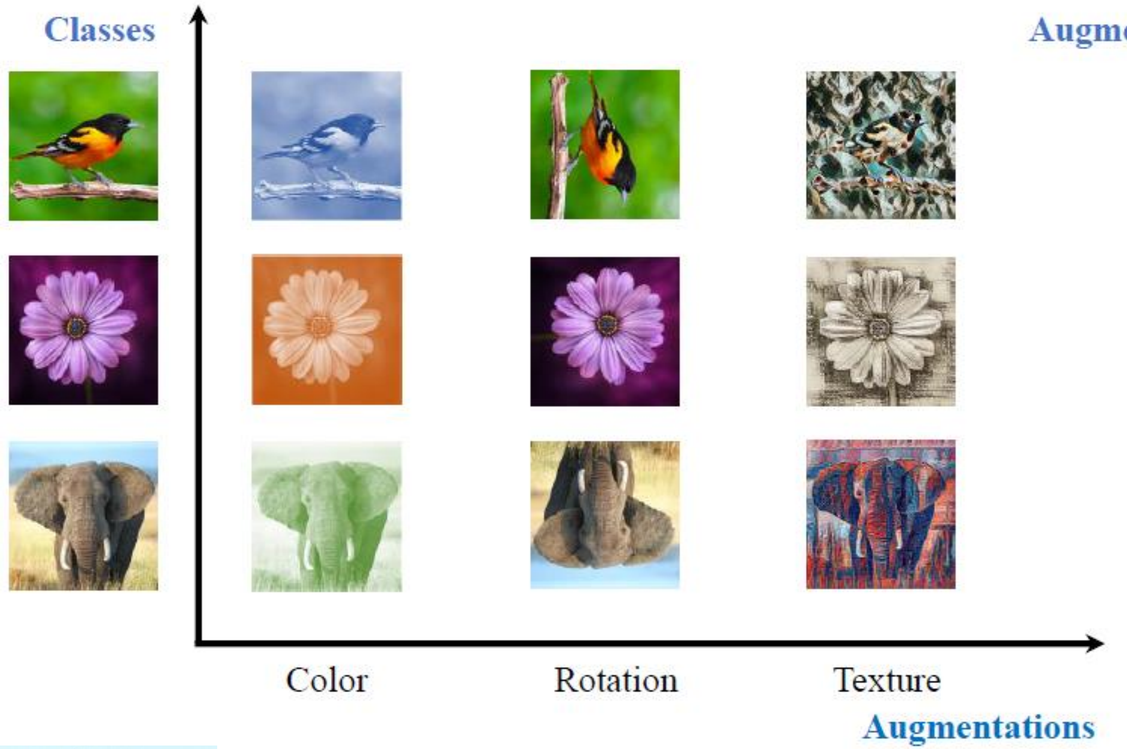
Views

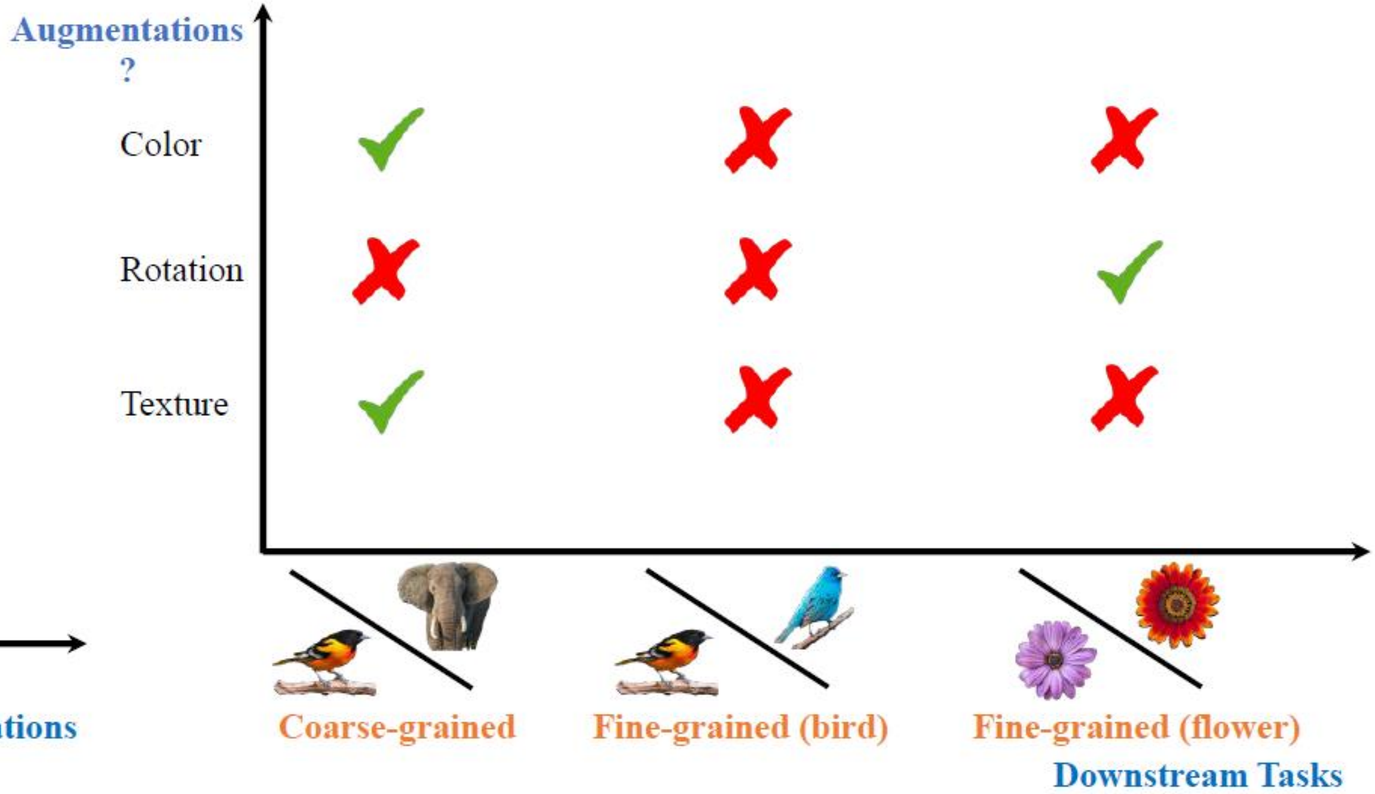
SimCLR augmentations (Chen et al, 2020)

SimCLR

[Chen, Kornblith, Norouzi, Hinton, ICML 2020]

The choice of data augmentation is itself supervision



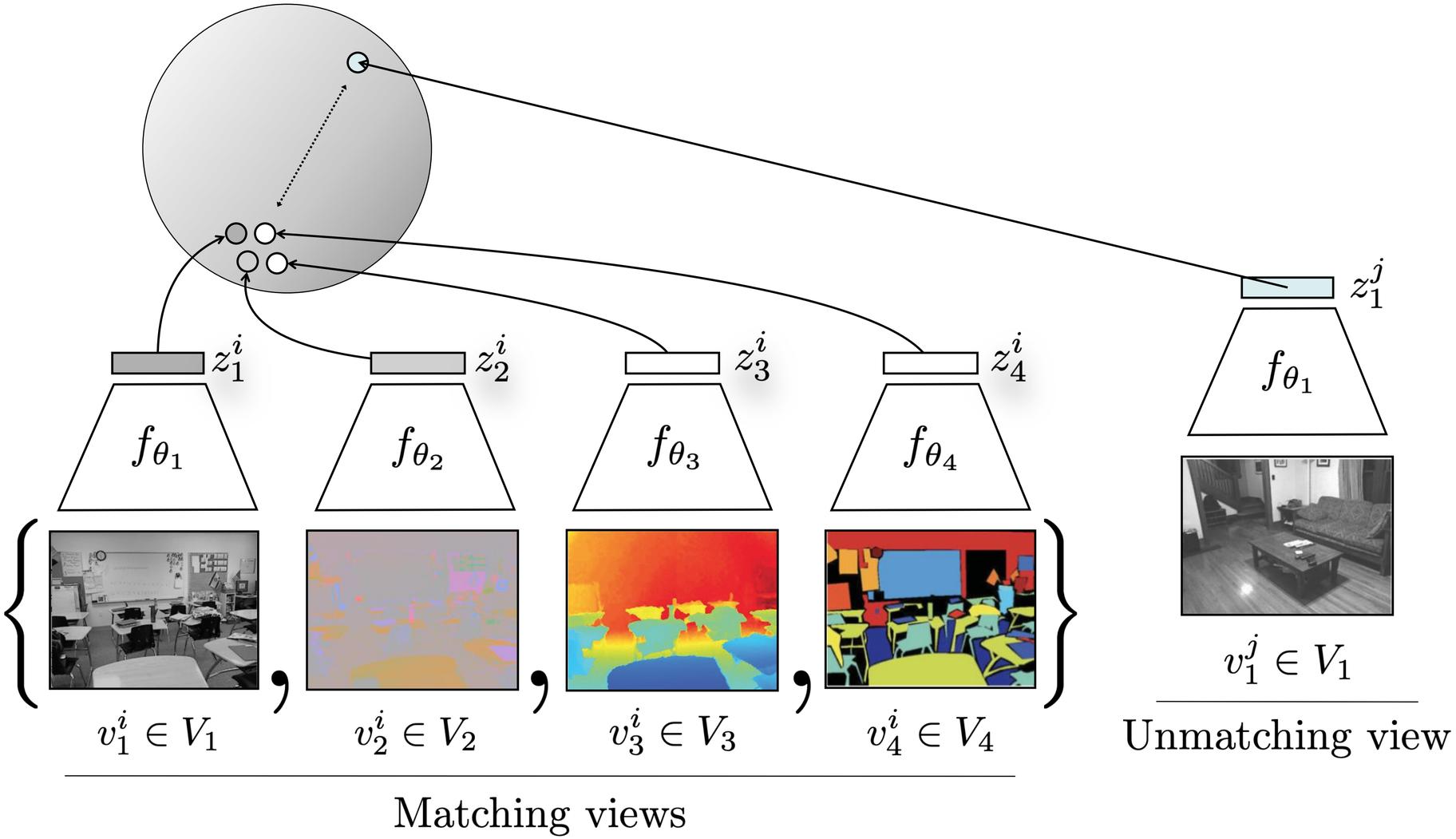


What should not be contrastive in contrastive learning

T Xiao, X Wang, AA Efros, T Darrell - ICLR 2021

Contrastive Multiview Coding

[Tian, Krishnan, Isola, ECCV 2020]



- goal: contrastive learning ***without*** data augmentation

- Approach: make the “views” latent

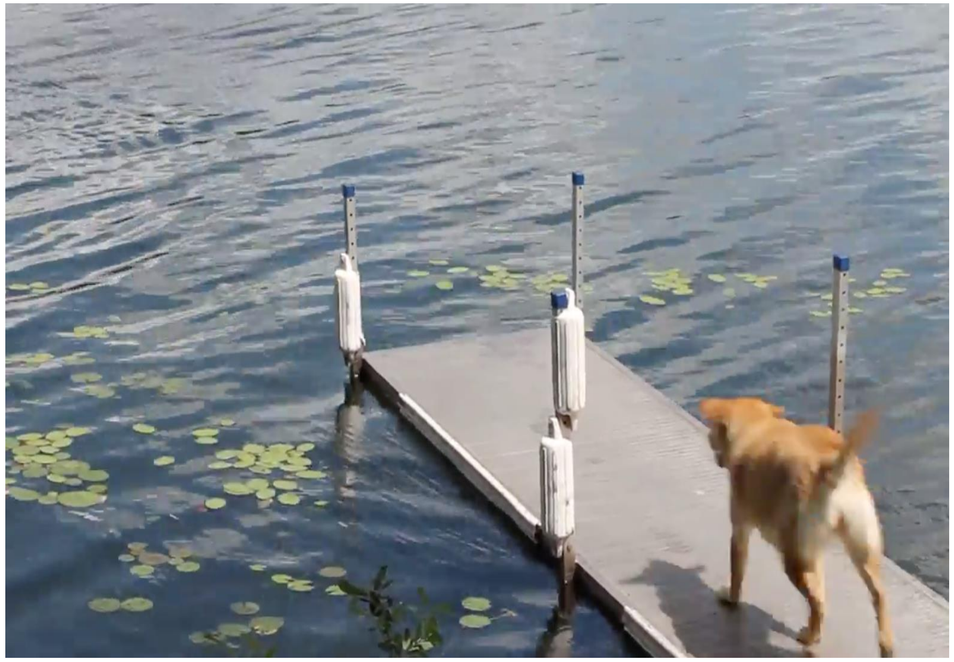
- Key question: where to get supervisory signal?

Time as Supervisory Signal

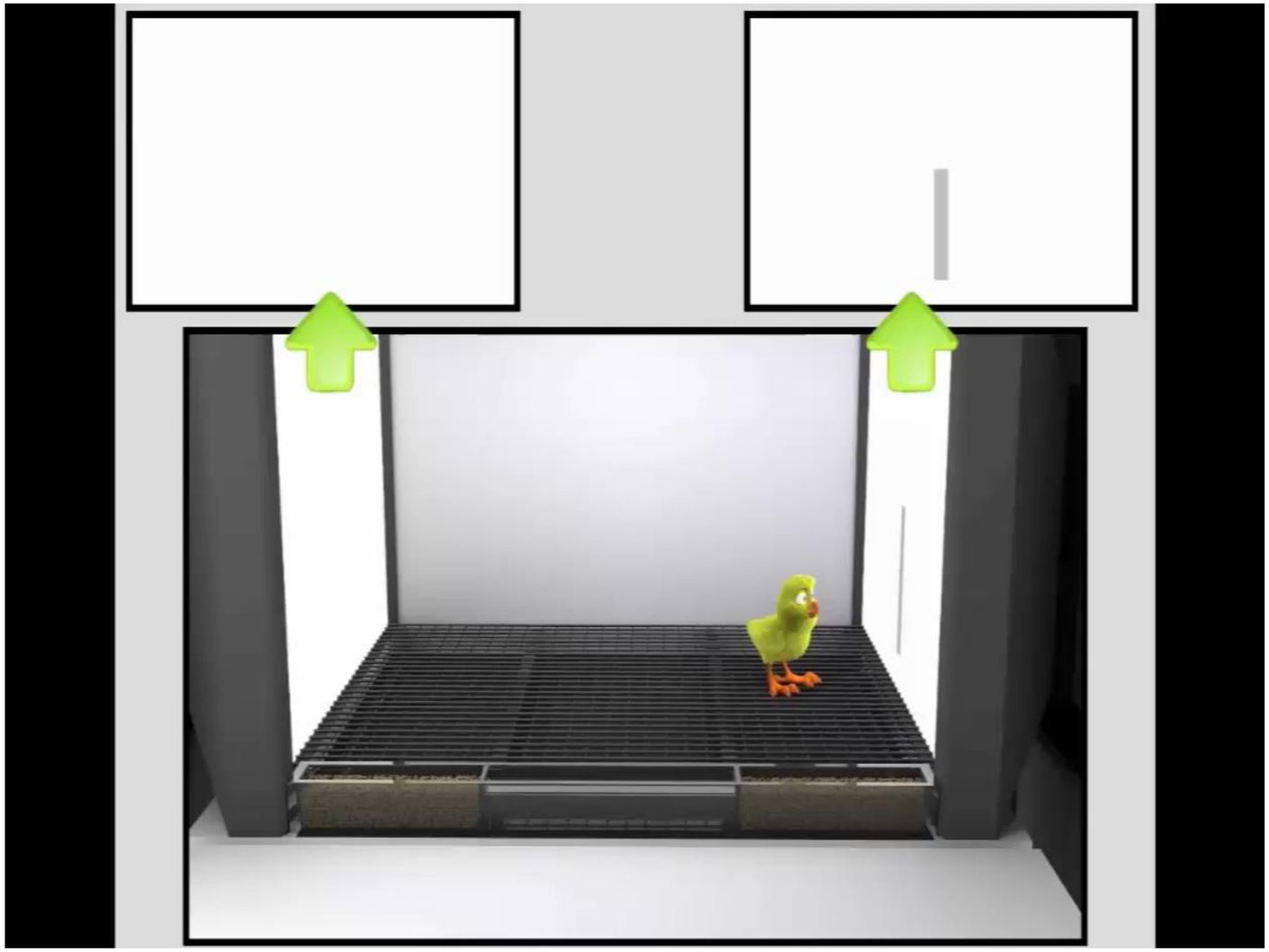
“It irritated him that the “dog” of 3:14 in the afternoon, seen in profile, should be indicated by the same noun as the dog of 3:15, seen frontally…”

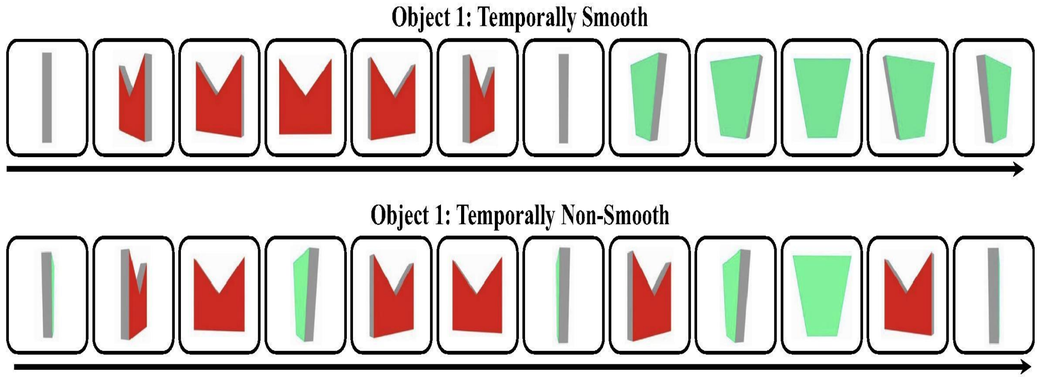
-- from *Funes the Memorious*

Jorge Luis Borges



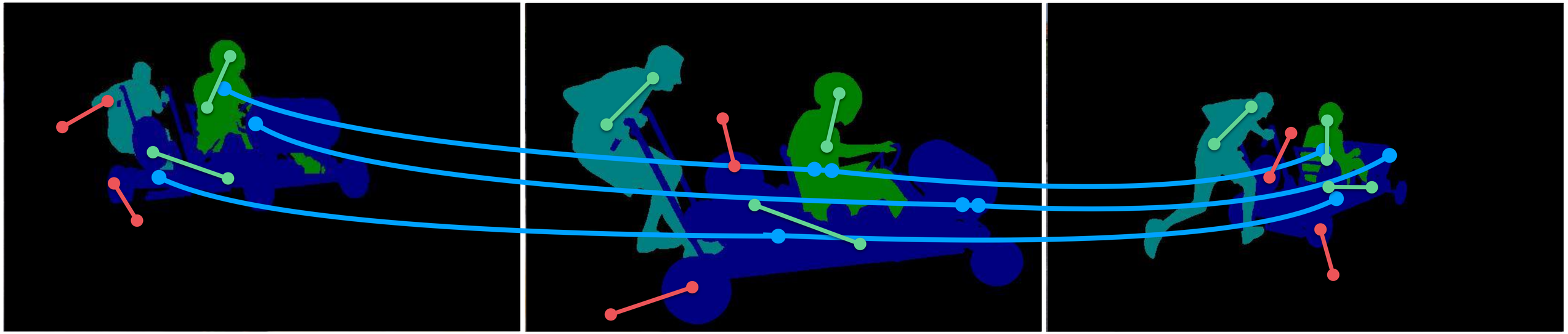
Temporal Continuity crucial for visual development





Wood 2013, 2016, 2018

Video as Data Augmentation



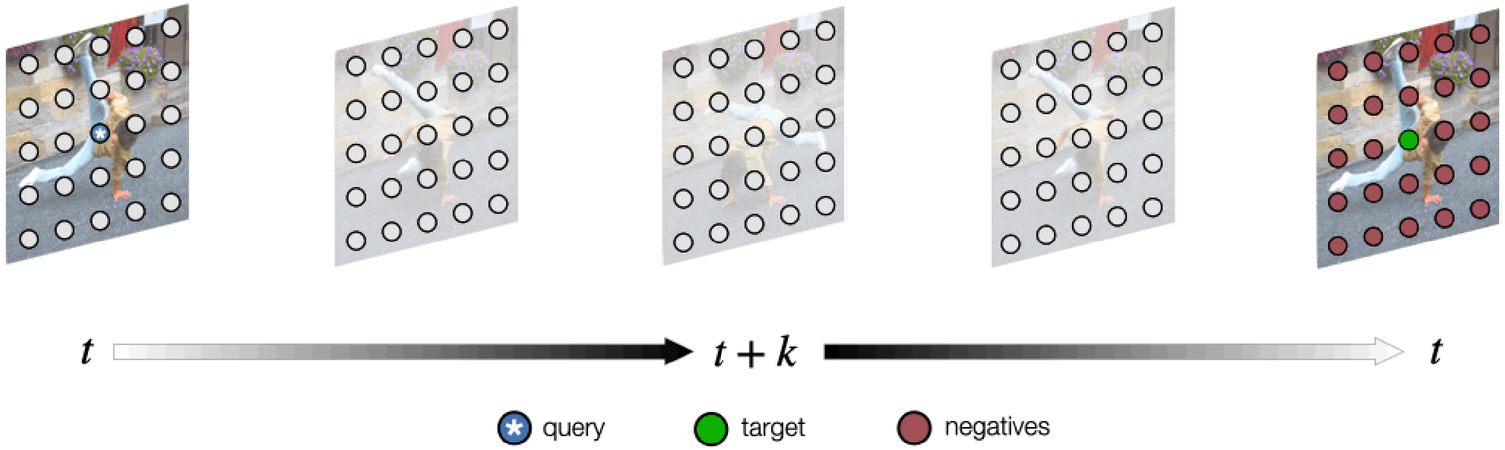
Correspondence

"Common Fate"

Wehrtheimer (1938)

Space-Time Correspondence as a Contrastive Random Walk

NeurIPS 2020



Allan A. Jabri
UC Berkeley

Andrew Owens
U. Michigan

Alexei A. Efros
UC Berkeley

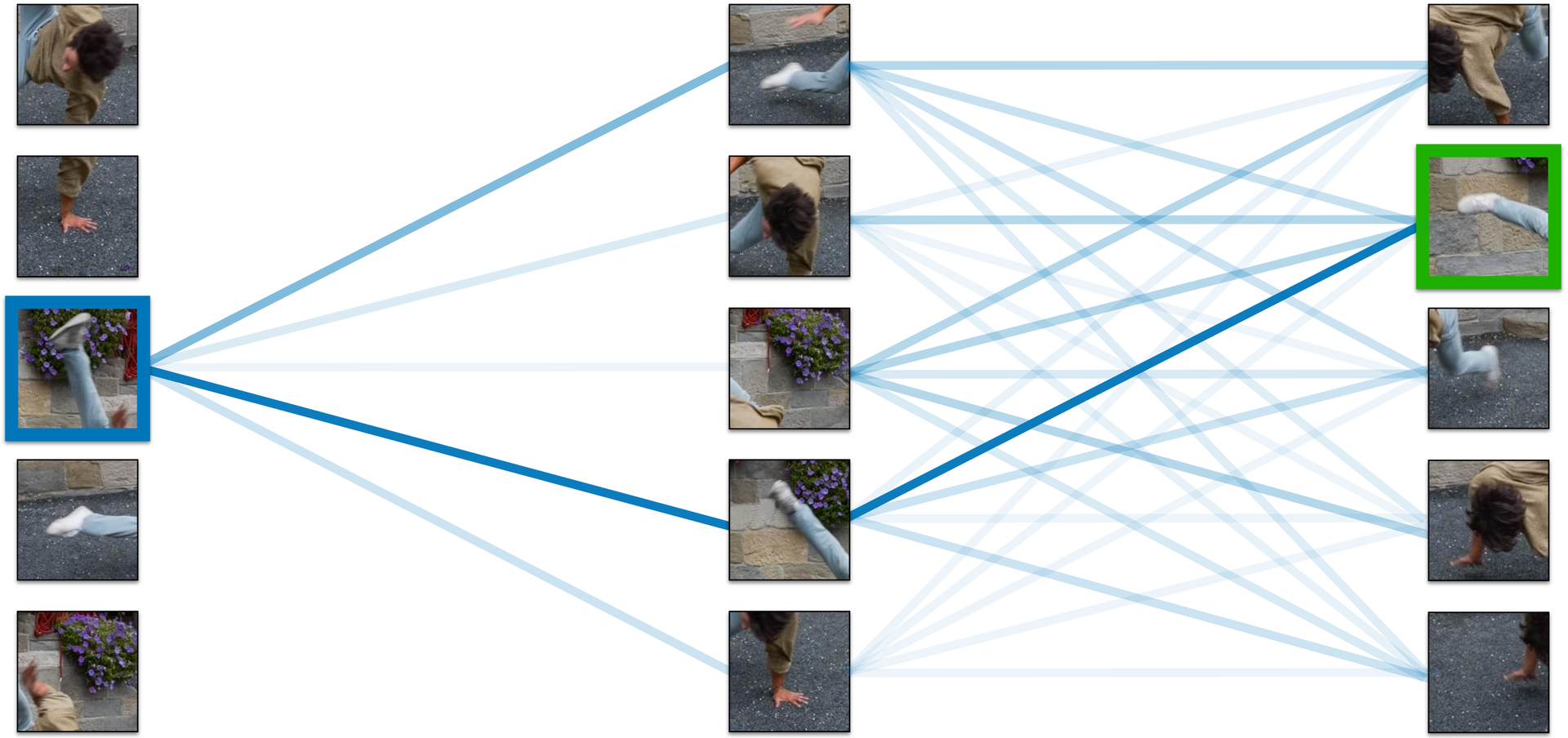
BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

NEURAL INFORMATION PROCESSING SYSTEMS

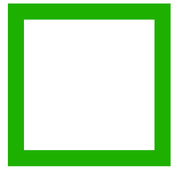
Warm up: the supervised case

Supervised Distance Learning

Implicit “data augmentation” with latent views



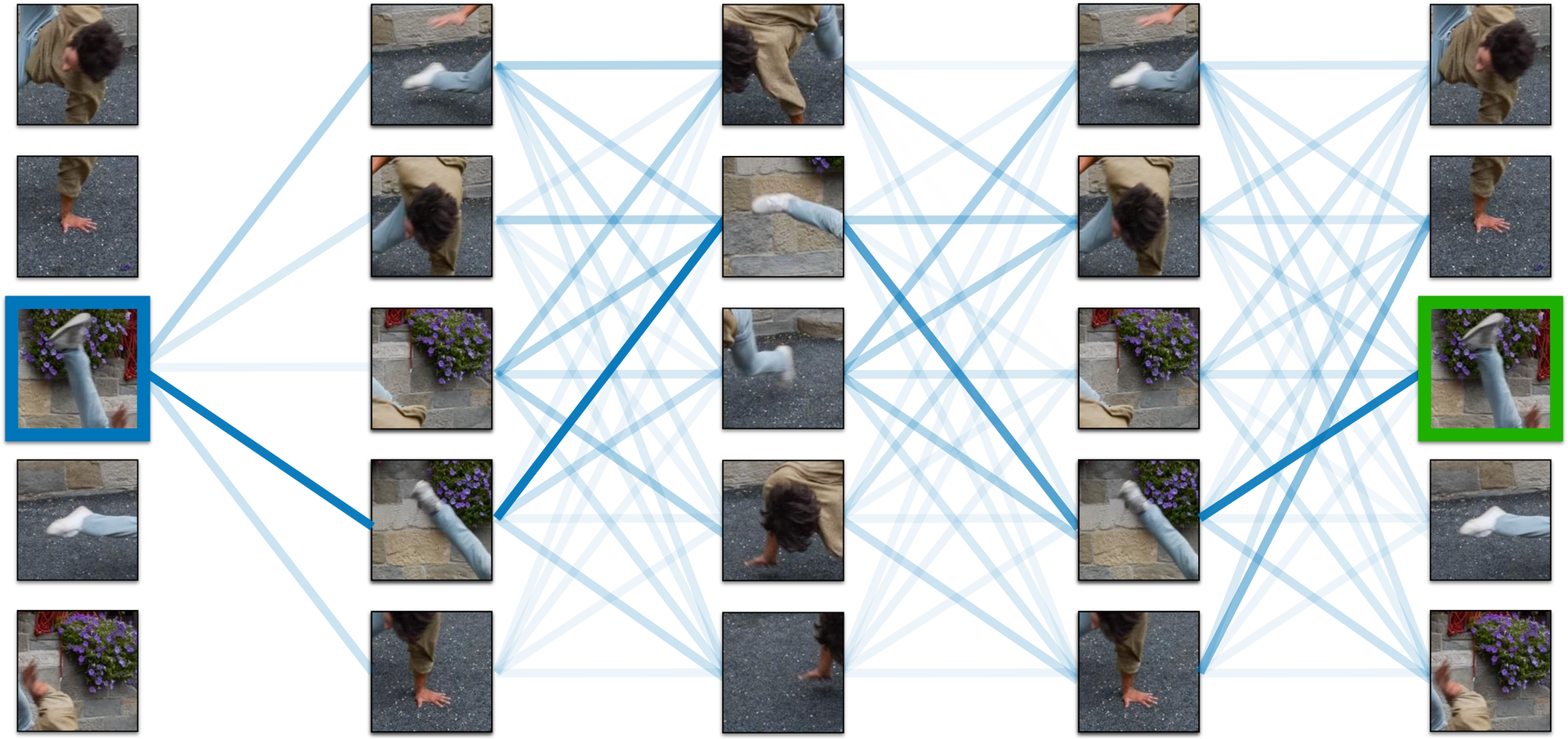
Palindromes (cycles in time)

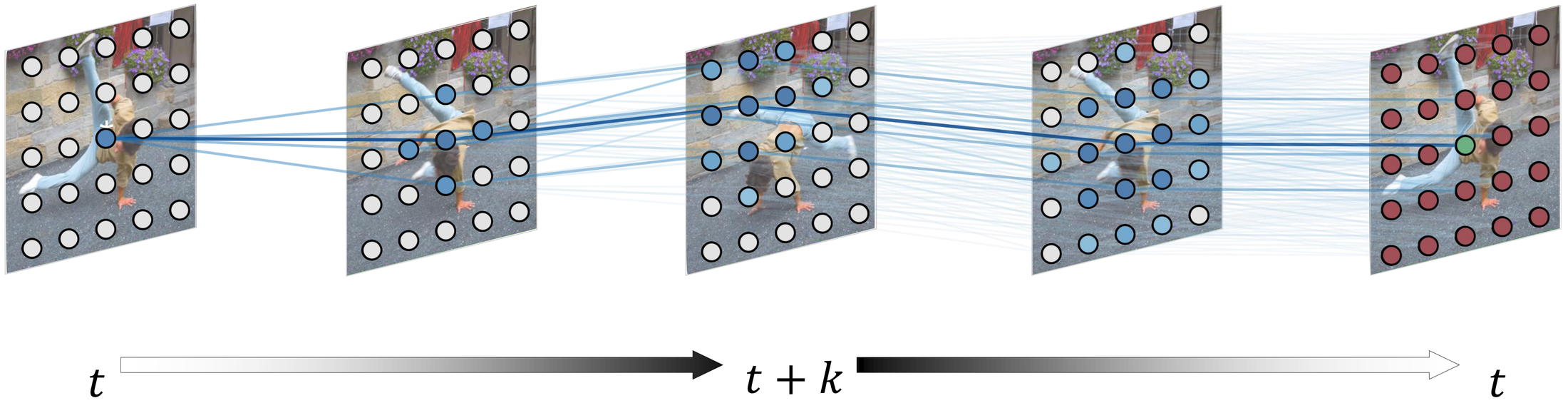


2

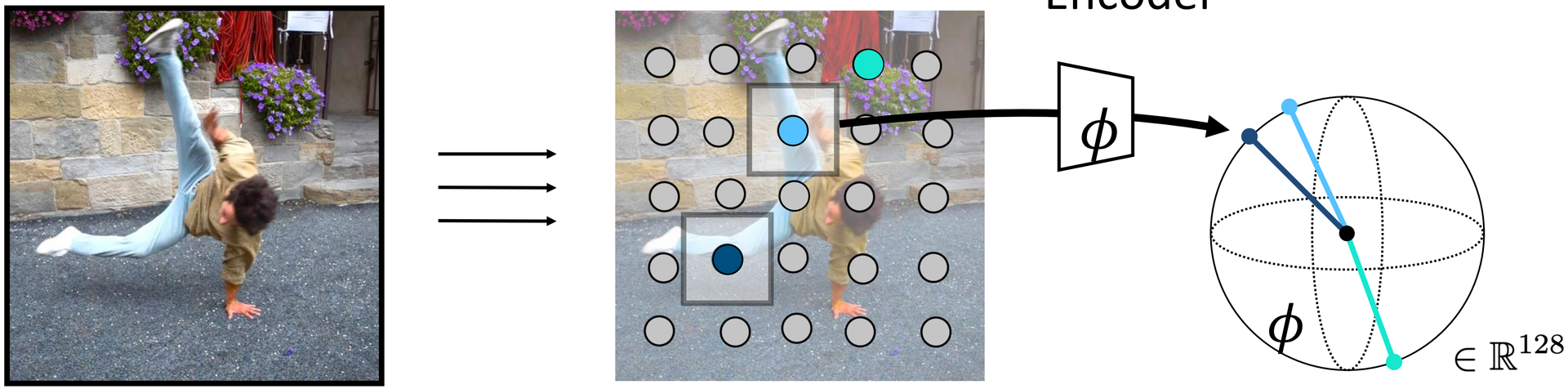
1

Self-supervised Learning




t
$t+k$
t
query
target
negatives

Video as a Graph



Pixels

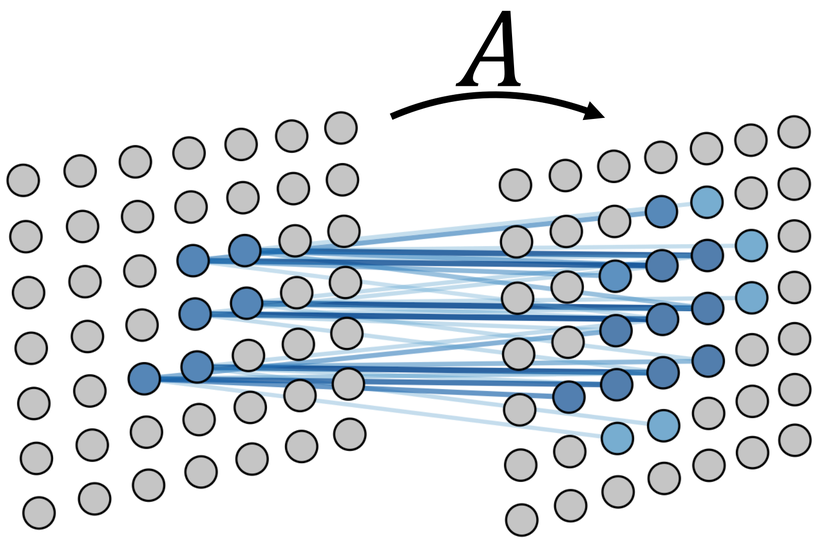
I_t

Nodes

$\mathbf{q}_t$

Representation

Video as a Graph



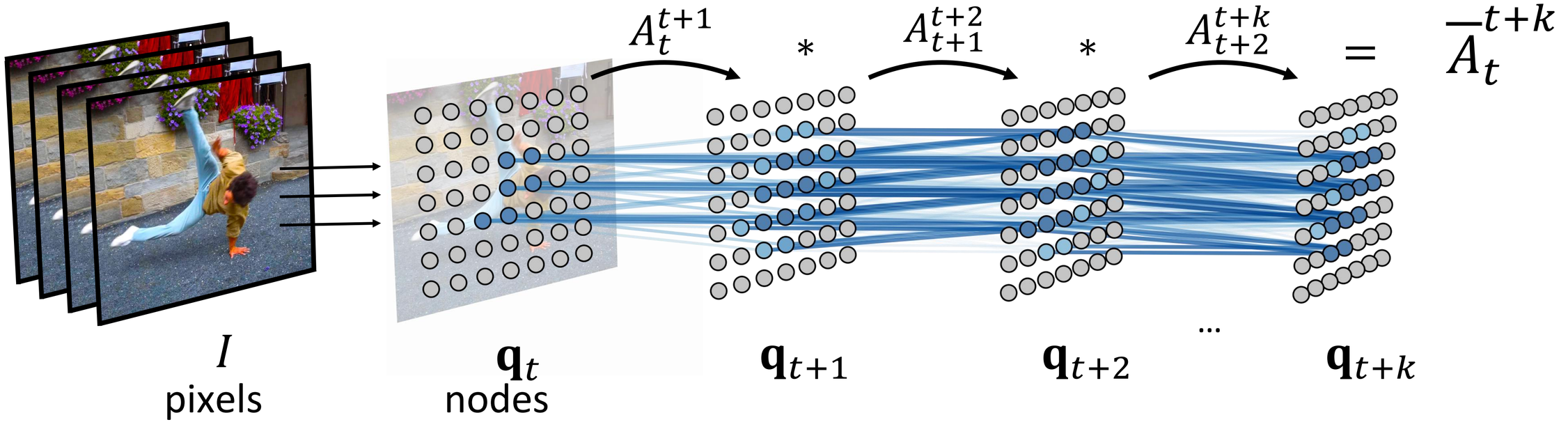
$$A_{ij} = \frac{e^{d_\phi(q_t^i,\, q_{t+1}^j)/\tau}}{\sum_l e^{d_\phi(q_t^i,\, q_{t+1}^l)/\tau}}$$

$$= P(X_{t+1} = j | X_t = i)$$

where $d_\phi(x, y) = \phi(x)^\top \phi(y)$

X_t is the position of walker at time t

Correspondence as a Random Walk



Learn representation ϕ = Fit transition probabilities $\overline{A}_t^{t+k}$

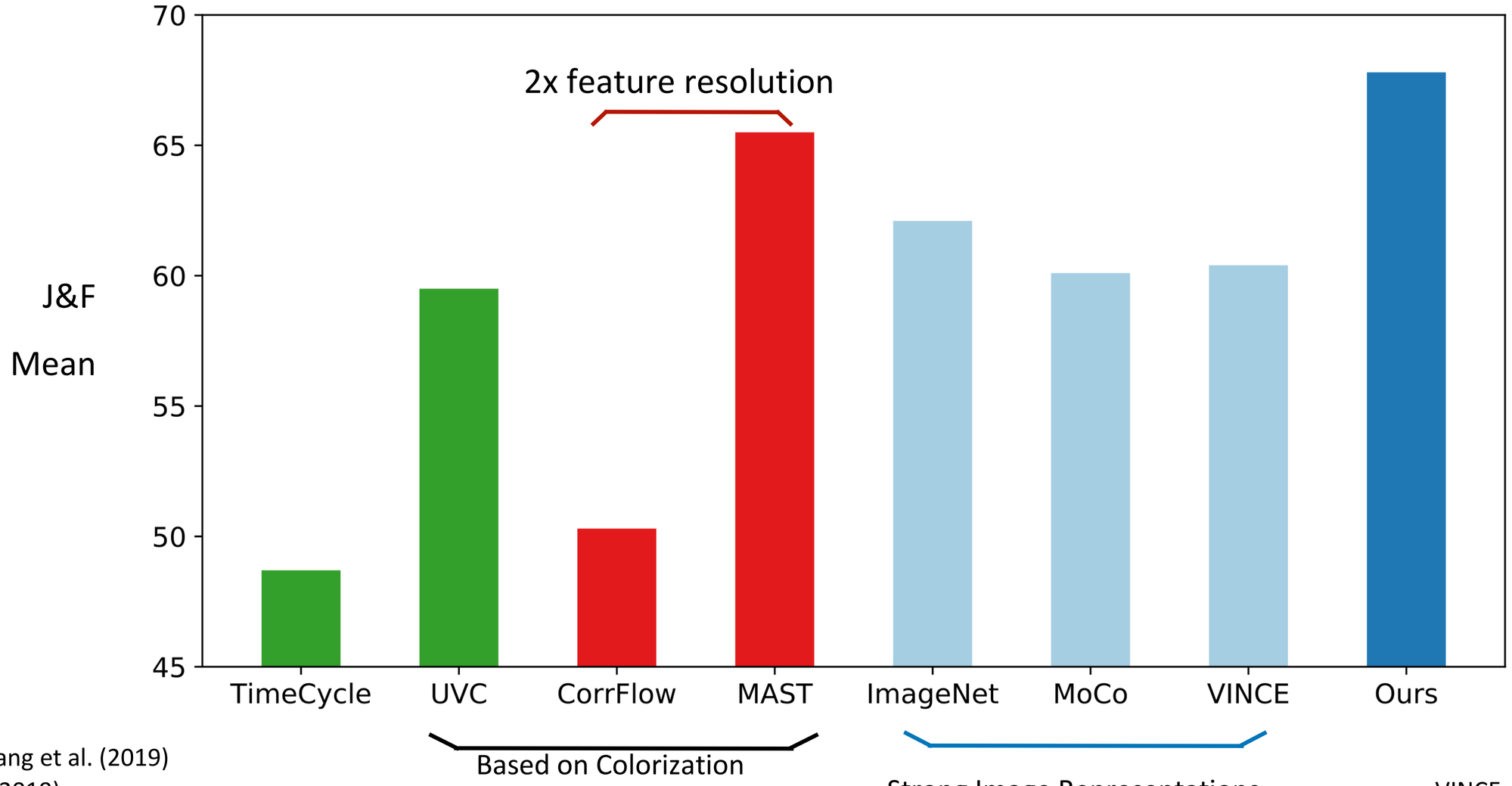
Meila and Shi, *Learning Segmentation by Random Walks*. NIPS 2001

Qualitative Results: Video Object Propagation (DAVIS)

UVC
Li et al. (2019)

Ours

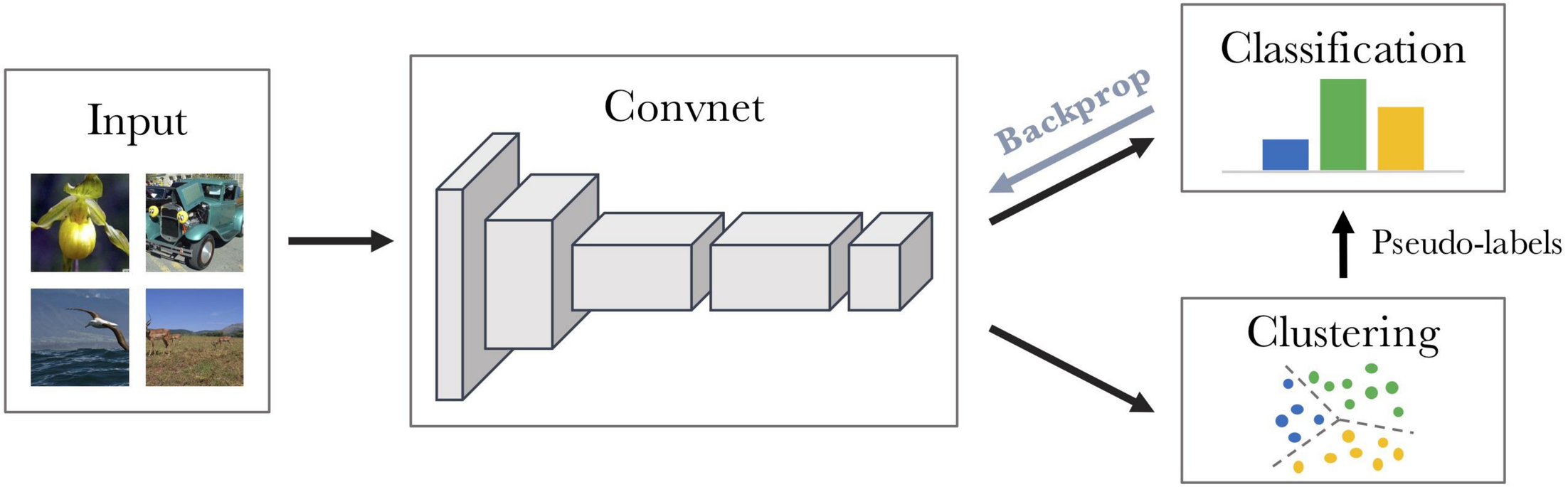
DAVIS Video Object Segmentation



TimeCycle: Wang et al. (2019)
UVC: Li et al. (2019)
CorrFlow: Lai et al. (2019)
MAST: Lai et al. (2019)

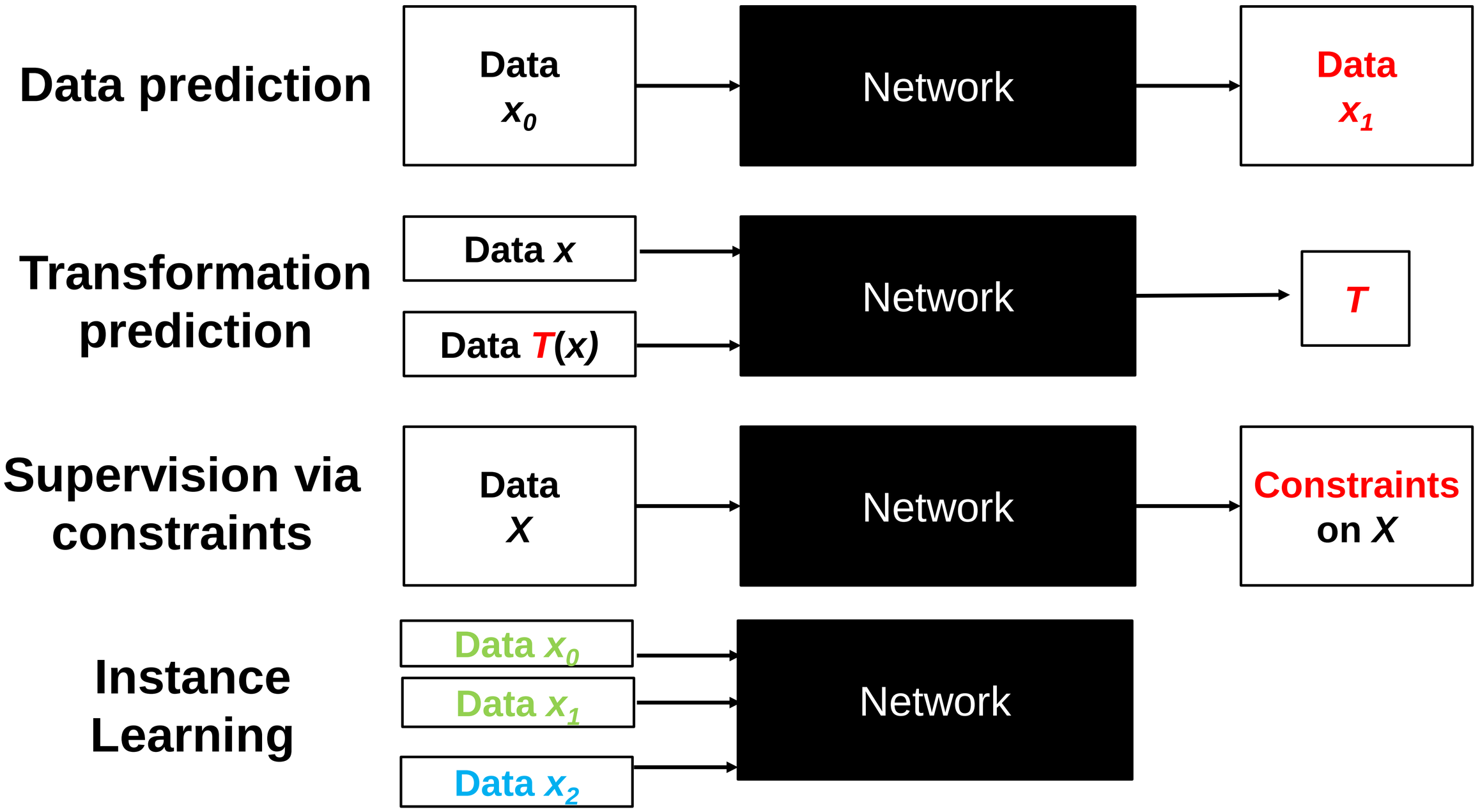
VINCE: Gordon et al. (2020)
MoCo: He et al. (2019)

DeepCluster, Caron et al, ECCV 2018



So, where are we now?

(Partial) Taxonomy of Self-Supervision



And the current winner is…

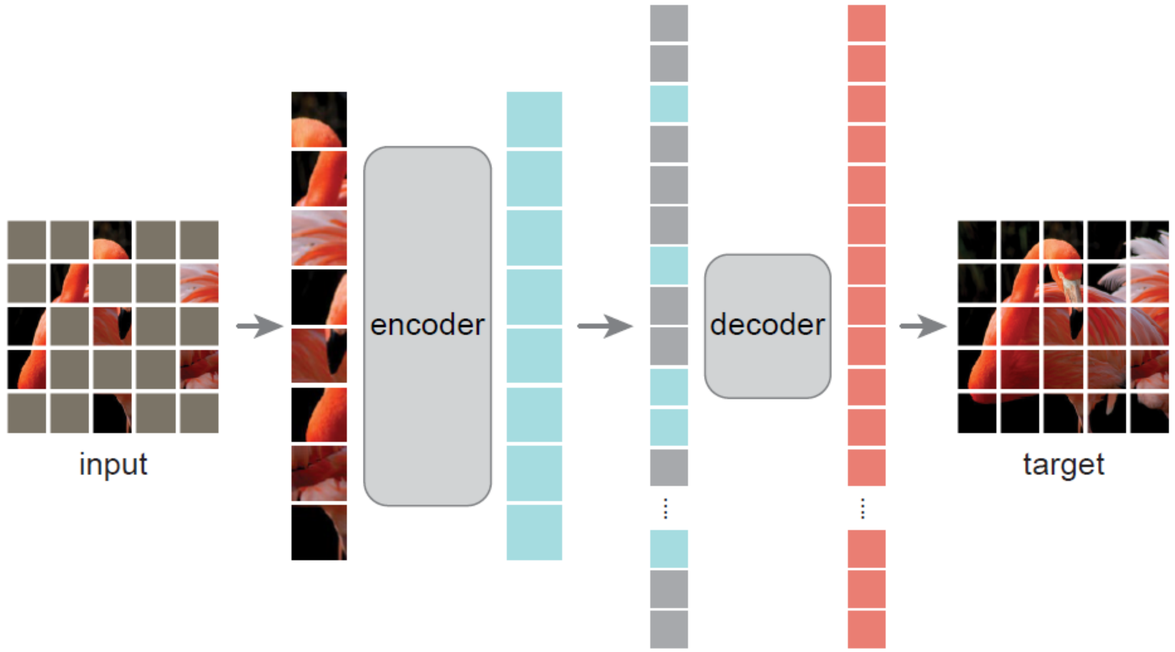
Data prediction

Context Encoder
Pathak et al. CVPR 2016

Masked Autoencoder
He et al. 2021

And the current winner is…

Data prediction



Masked Autoencoder
He et al. 2021

Why Do Self-Supervision?

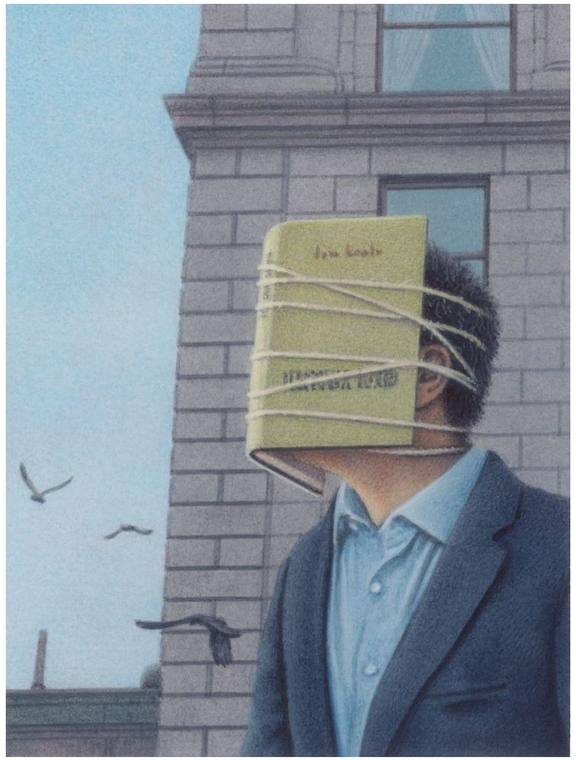
- A common answer:
  - *“Because labels are expensive”*

Supervised vs. Self-Supervised Learning

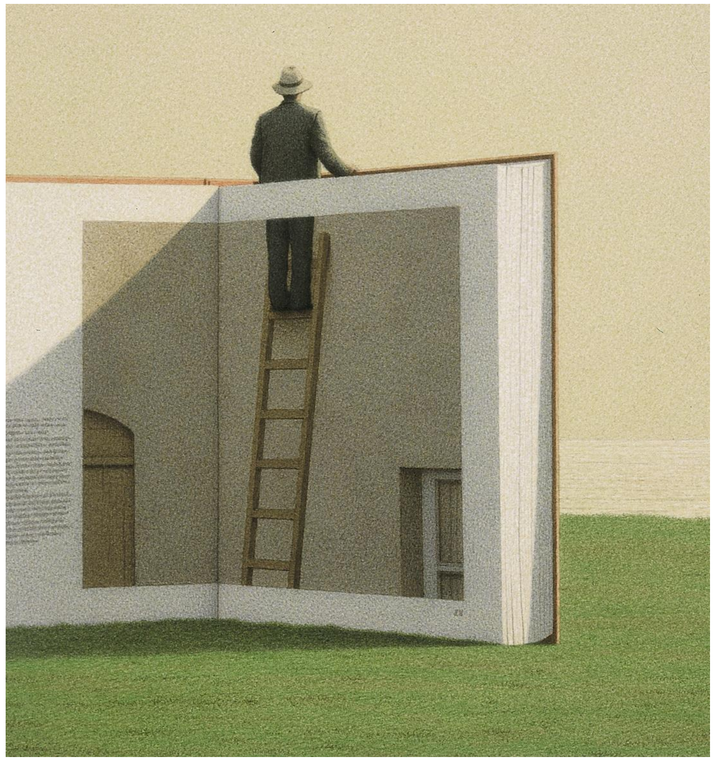
- If **training task / distribution == test task / distribution**, supervised learning is almost always preferable
  - Self-driving cars
    - 7.8B people / 64M km of roads = **122 people / km**

Why Self-Supervision?

1. To get away from <u>semantic categories</u>



2. To get away from <u>fixed datasets</u>

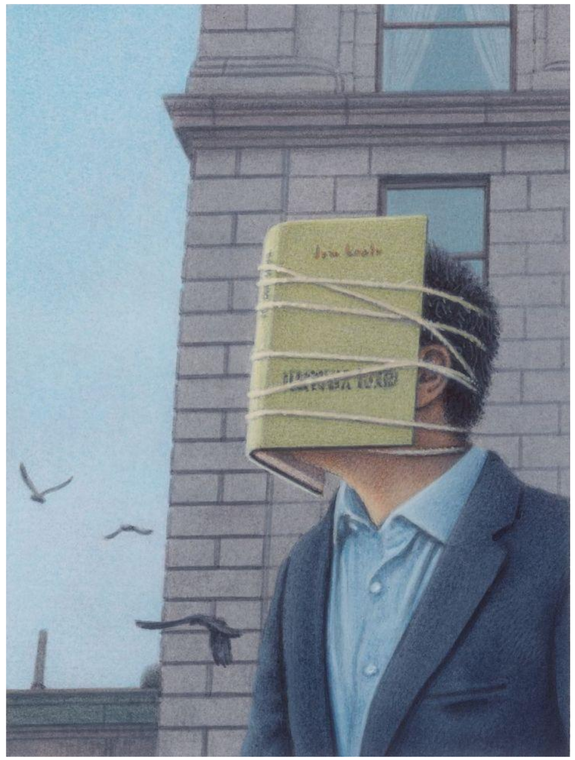


3. To get away from <u>fixed objectives</u>

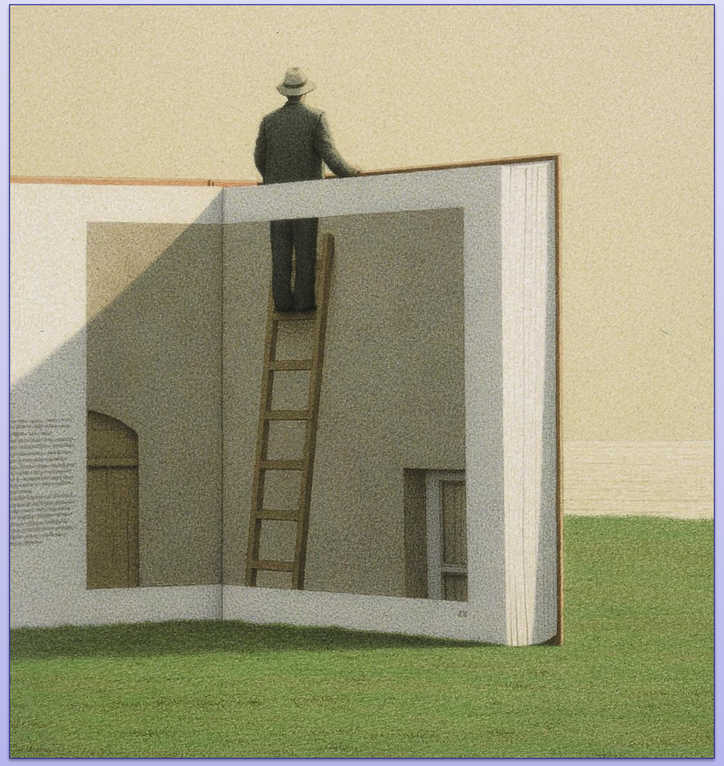
Artwork by *Quint Buchholz*

Why Self-Supervision?

1. To get away from <u>semantic categories</u>

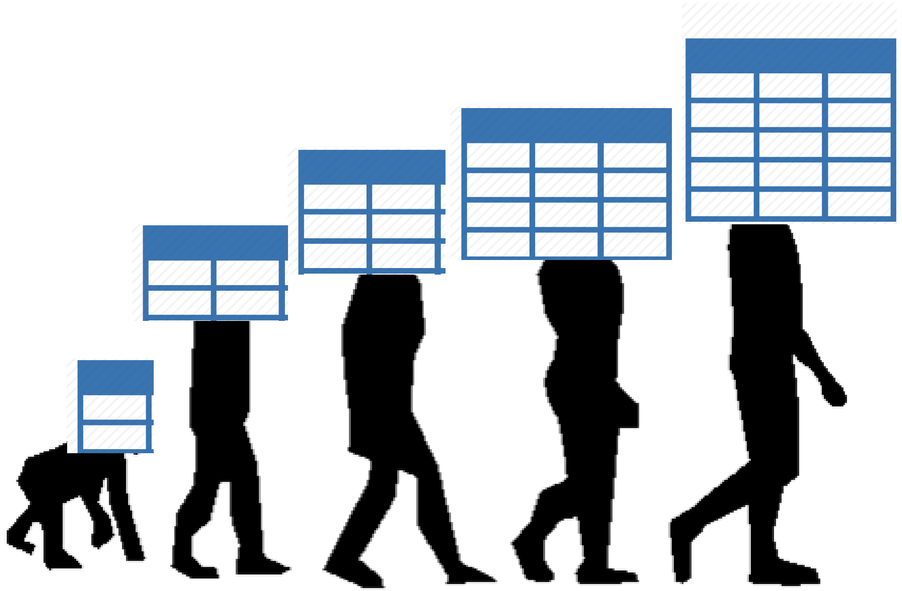


2. To get away from <u>fixed datasets</u>

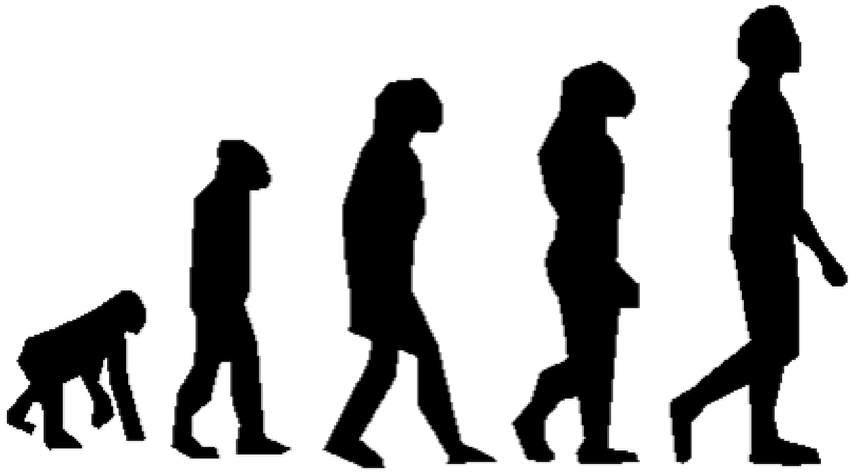


3. To get away from <u>fixed objectives</u>

Static vs. Changing World



- Lookup table / memory
- kNN converges to Bayes risk! [Cover & Hart, 1967]
- “Specialist”



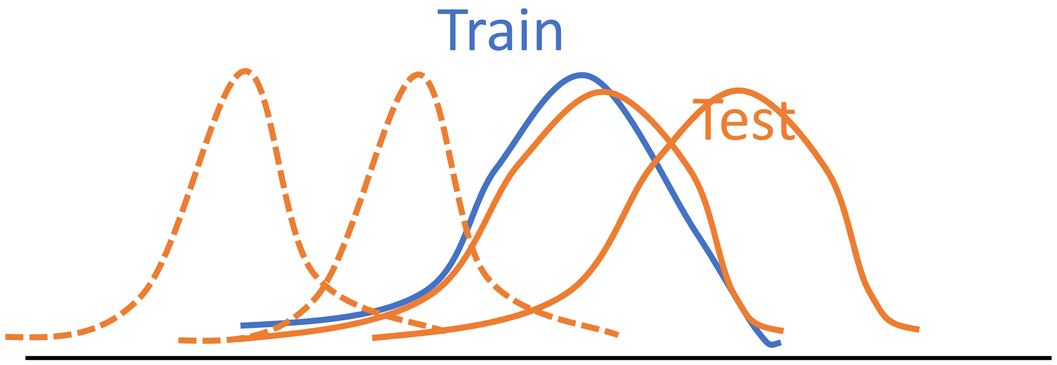
- Brain
- Worse than kNN
- “Generalist”

Fixed Datasets can't represent a Changing World

- Real-world motivation
  - Biological agents never see the same data twice!
  - Every new piece of data is first "test", then "train"
- Repeating the same sample might encourage memorization / discourage generalization
- The only reason for fixed datasets is annotation expanse
- **But with self-supervision, there is no excuse for reusing data / multiple epochs**

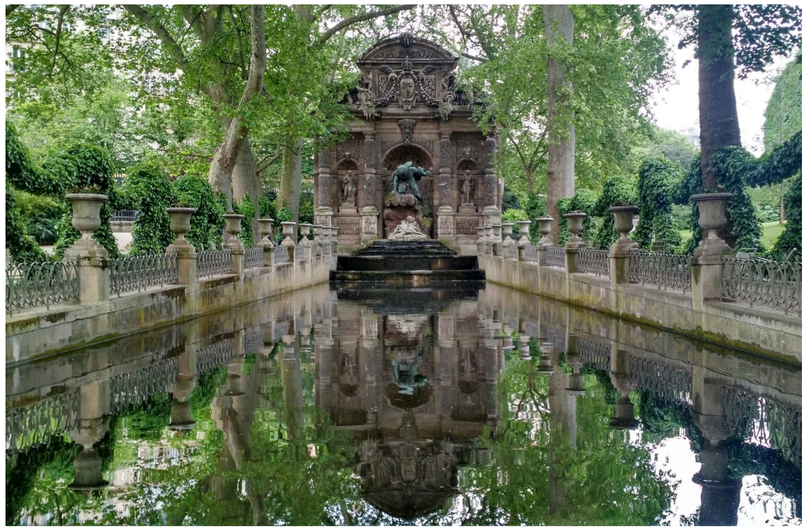
Two Approaches to Generalization

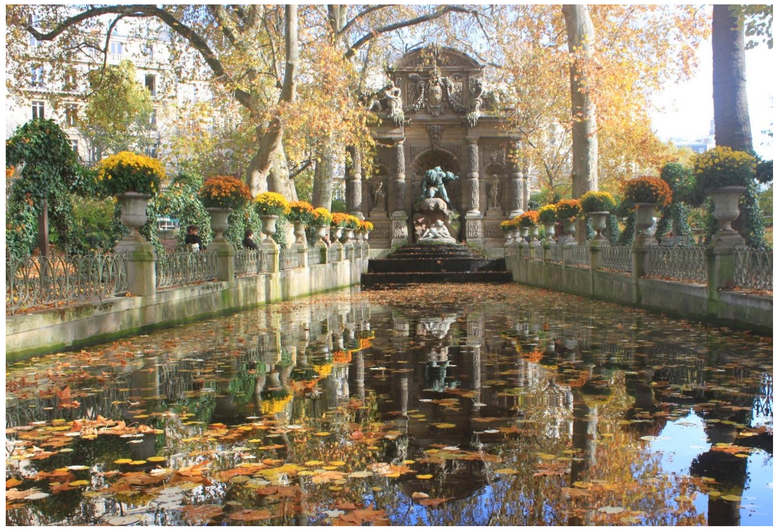
In theory: same distribution for train and test



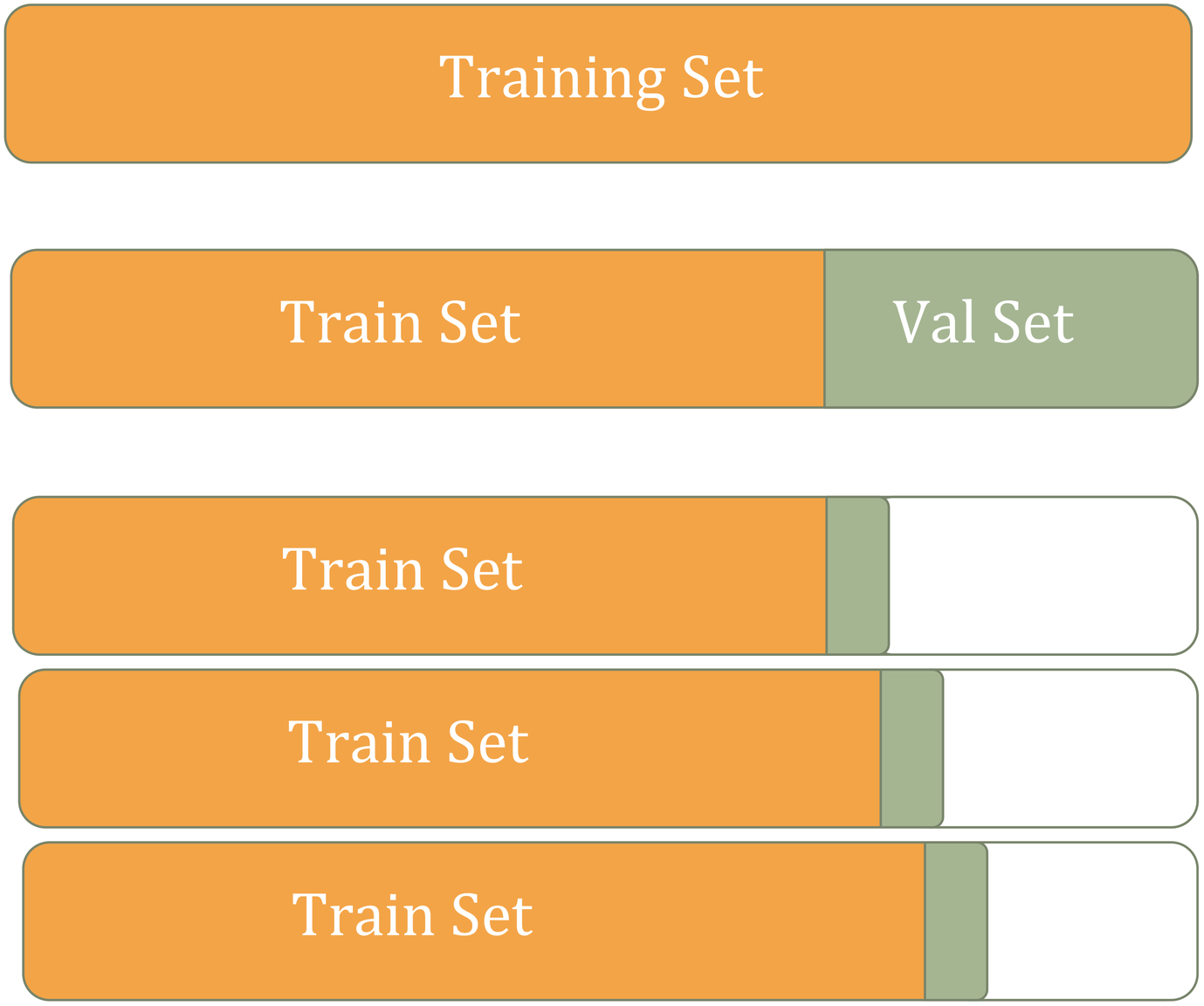
In the real world: this is rarely the case!

Adapt to the future once it arrives





Online Learning

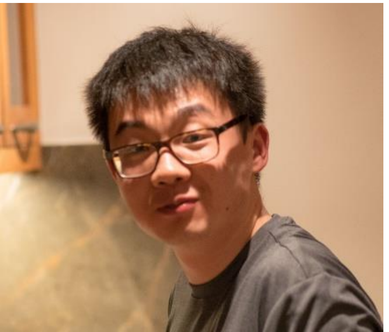


Test-Time Training

- Our attempt to operationalize continual learning on an infinite, changing data stream
- Approach: use **self-supervision** to adapt to new test data
  - Already applied this to RL with our curiosity work e.g. [Pathak, ICML'17]
  - What about vision tasks?

Test-Time Training with Self-Supervision for Generalization under Distribution Shifts

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, Moritz Hardt
UC Berkeley



ICML 2020

Test-Time Training (TTT)

$$\text{standard test error} = \mathbb{E}_Q[\ell(x, y);\ \theta]$$

For a test distribution Q

Test-Time Training (TTT)

$$\text{standard test error} = \mathbb{E}_Q[\ell(x, y);\ \theta]$$

For a test distribution Q

- The test sample x gives us a hint about Q

Test-Time Training (TTT)

$$\text{standard test error} = \mathbb{E}_Q[\ell(x, y);\ \theta]$$

$$\text{continual test error} = \mathbb{E}_Q[\ell(x, y);\ \theta(x)]$$

For a test distribution Q

- The test sample x gives us a hint about Q
- No fixed model, but adapt at test time

Test-Time Training (TTT)

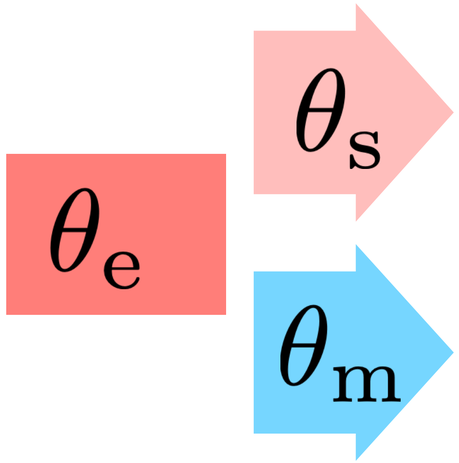
$$\text{standard test error} = \mathbb{E}_Q[\ell(x, y);\ \theta]$$

$$\text{continual test error} = \mathbb{E}_Q[\ell(x, y);\ \theta(x)]$$

For a test distribution Q

- The test sample x gives us a hint about Q
- No fixed model, but adapt at test time
- One sample learning problem
- No label? Self-supervision!

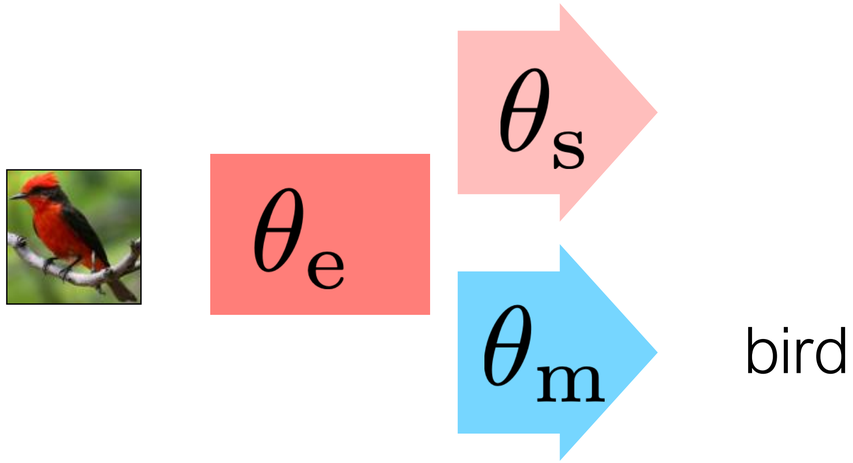
Algorithm for TTT



network
architecture

Algorithm for TTT

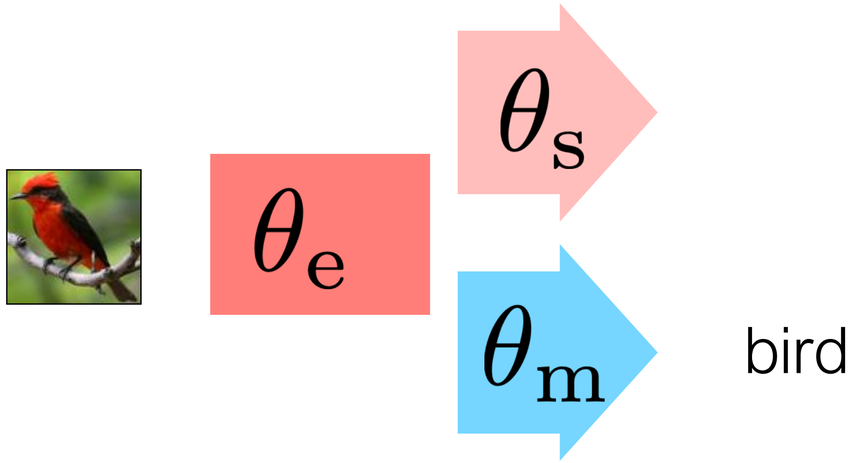
training



Algorithm for TTT

training

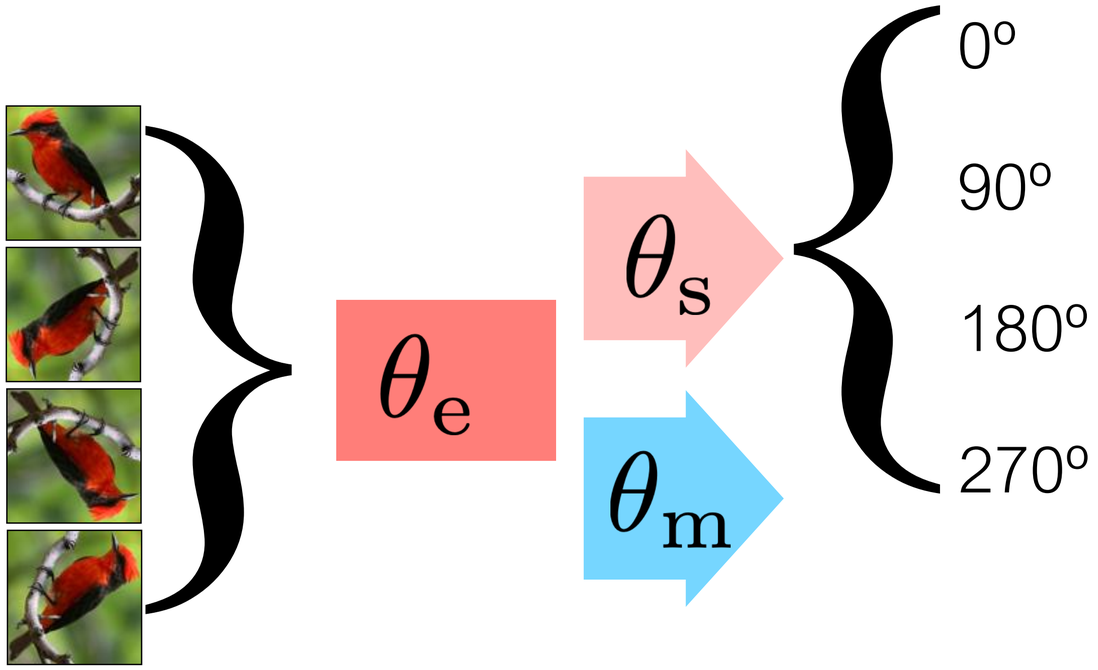
$\ell_{\mathrm{m}}(x, y; \theta_{\mathrm{e}}, \theta_{\mathrm{m}})$



Algorithm for TTT

training

$\ell_{\mathrm{m}}(x, y; \theta_{\mathrm{e}}, \theta_{\mathrm{m}})$

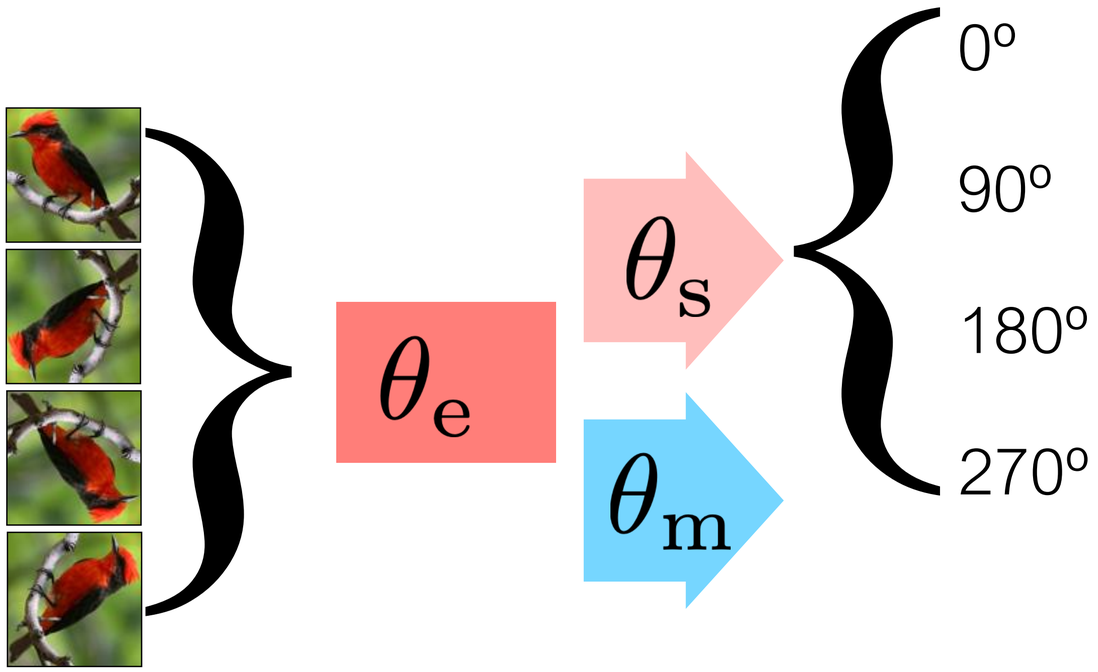


Algorithm for TTT

training

$\ell_{\mathrm{m}}(x, y; \theta_{\mathrm{e}}, \theta_{\mathrm{m}})$

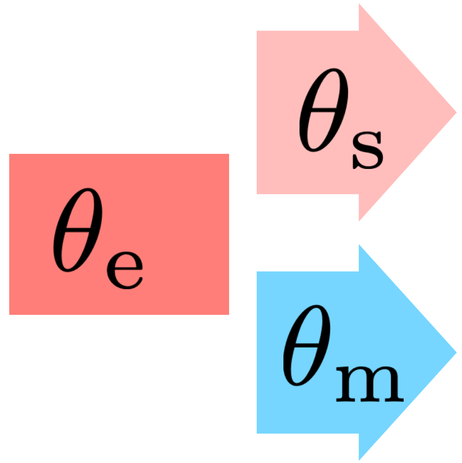
$+ \ell_s(x, y_{\mathrm{s}}; \theta_e, \theta_s)$



Algorithm for TTT

training

$$\min_{\theta_{\mathrm{e}},\theta_{\mathrm{s}},\theta_{\mathrm{m}}} \mathbb{E}_P \left[\begin{array}{l} \ell_{\mathrm{m}}(x, y; \theta_{\mathrm{e}}, \theta_{\mathrm{m}}) \\ + \ell_s(x, y_{\mathrm{s}}; \theta_e, \theta_s) \end{array} \right]$$



Algorithm for TTT

training

$$\min_{\theta_{\mathrm{e}},\theta_{\mathrm{s}},\theta_{\mathrm{m}}} \mathbb{E}_P \left[\begin{array}{l} \ell_{\mathrm{m}}(x, y; \theta_{\mathrm{e}}, \theta_{\mathrm{m}}) \\ + \ell_s(x, y_{\mathrm{s}}; \theta_e, \theta_s) \end{array} \right]$$

testing

θ_{e}

θ_{s}

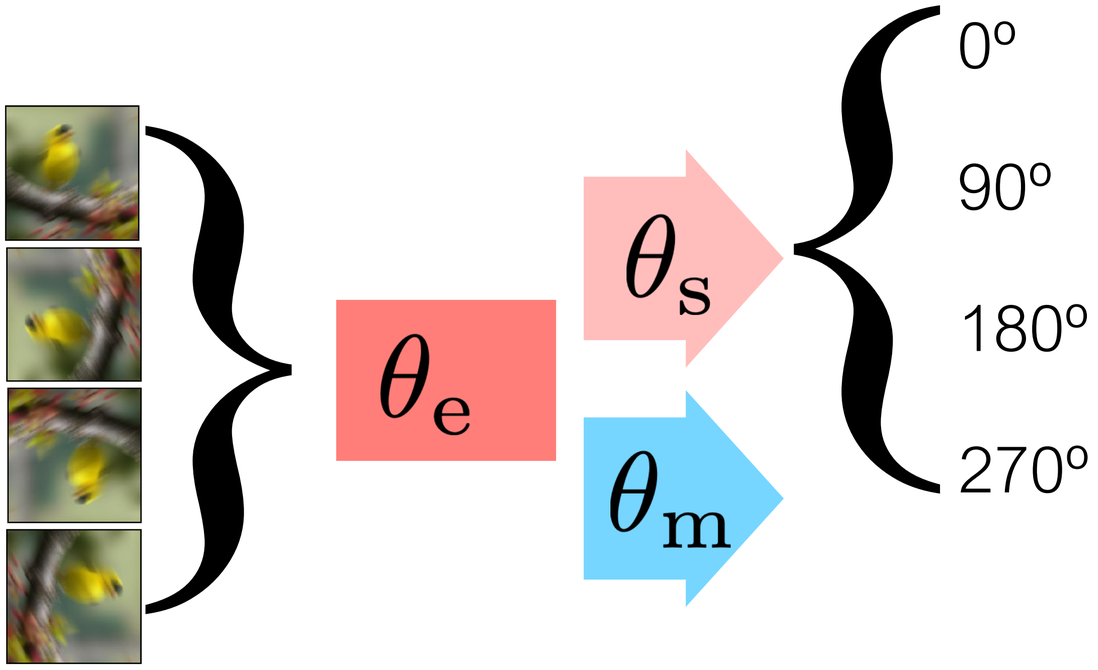
θ_{m}

Algorithm for TTT

training

$$\min_{\theta_{\mathrm{e}},\theta_{\mathrm{s}},\theta_{\mathrm{m}}} \mathbb{E}_P \left[\begin{array}{l} \ell_{\mathrm{m}}(x, y; \theta_{\mathrm{e}}, \theta_{\mathrm{m}}) \\ + \ell_s(x, y_{\mathrm{s}}; \theta_e, \theta_s) \end{array} \right]$$

testing



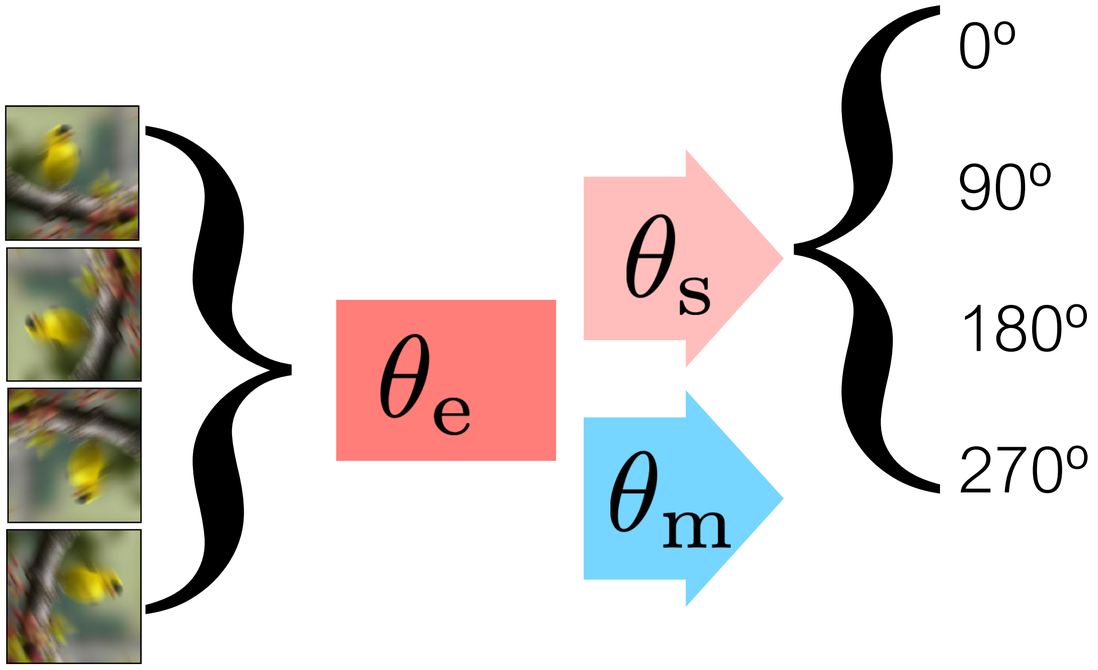
Algorithm for TTT

training

$$\min_{\theta_e,\theta_s,\theta_m} \mathbb{E}_P \left[\begin{array}{l} \ell_m(x, y; \theta_e, \theta_m) \\ + \ell_s(x, y_s; \theta_e, \theta_s) \end{array} \right]$$

testing

$$\min_{\theta_e,\theta_s} \left[\ell_s(x, y_s; \theta_e, \theta_s) \right]$$



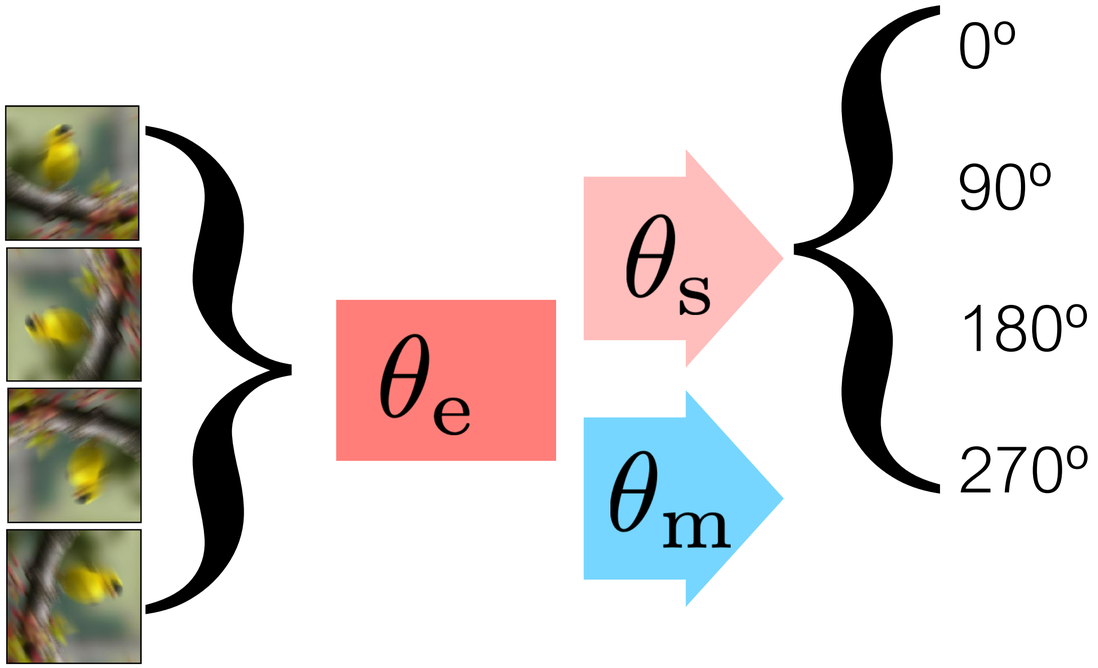
Algorithm for TTT

training

$$\min_{\theta_e,\theta_s,\theta_m} \mathbb{E}_P \left[\begin{array}{l} \ell_m(x, y; \theta_e, \theta_m) \\ + \ell_s(x, y_s; \theta_e, \theta_s) \end{array} \right]$$

testing

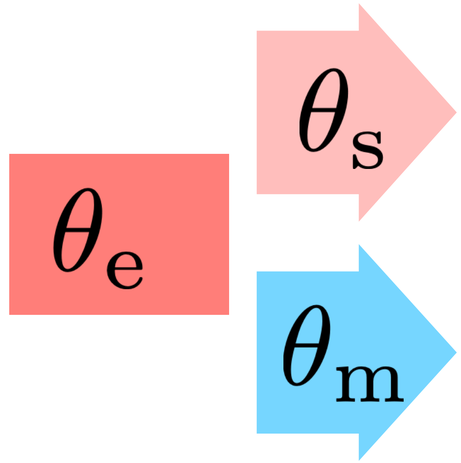
$$\min_{\theta_e,\theta_s} \mathbb{E}_Q \left[\ell_s(x, y_s; \theta_e, \theta_s) \right]$$



Algorithm for TTT

training

$$\min_{\theta_{\mathrm{e}},\theta_{\mathrm{s}},\theta_{\mathrm{m}}} \mathbb{E}_P \left[\begin{array}{l} \ell_{\mathrm{m}}(x, y; \theta_{\mathrm{e}}, \theta_{\mathrm{m}}) \\ + \ell_s(x, y_{\mathrm{s}}; \theta_e, \theta_s) \end{array} \right]$$



testing

$$\min_{\theta_{\mathrm{e}},\theta_{\mathrm{s}}} \mathbb{E}_Q \left[\ell_s(x, y_{\mathrm{s}}; \theta_e, \theta_s) \right]$$

$\rightarrow \theta(x)$: make prediction on x

Algorithm for TTT

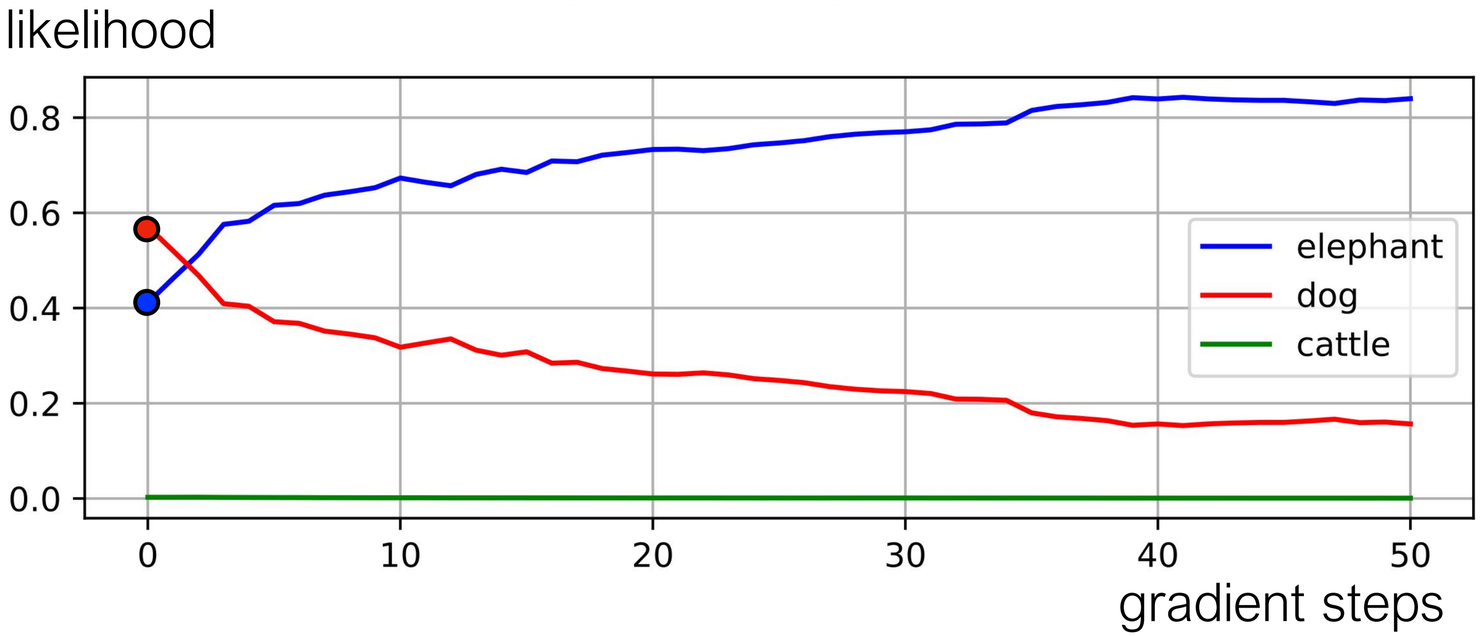
training

$$\min_{\theta_e,\theta_s,\theta_m} \mathbb{E}_P \left[\begin{array}{l} \ell_m(x, y; \theta_e, \theta_m) \\ + \ell_s(x, y_s; \theta_e, \theta_s) \end{array} \right]$$

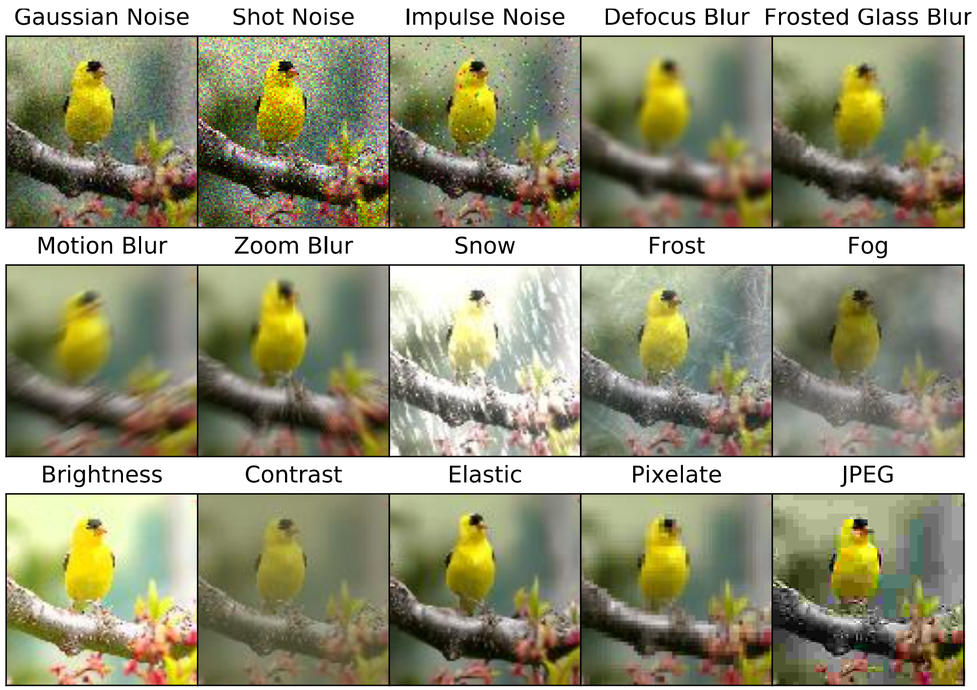
testing

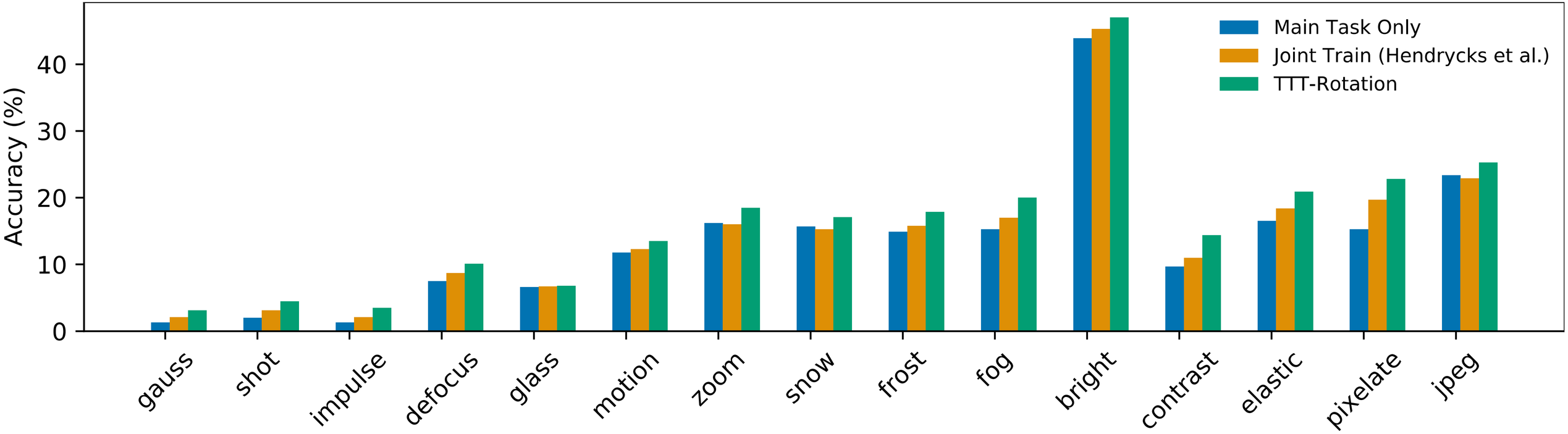
$$\min_{\theta_e,\theta_s} \mathbb{E}_Q \left[\ell_s(x, y_s; \theta_e, \theta_s) \right]$$

$\rightarrow \theta(x)$: make prediction on x

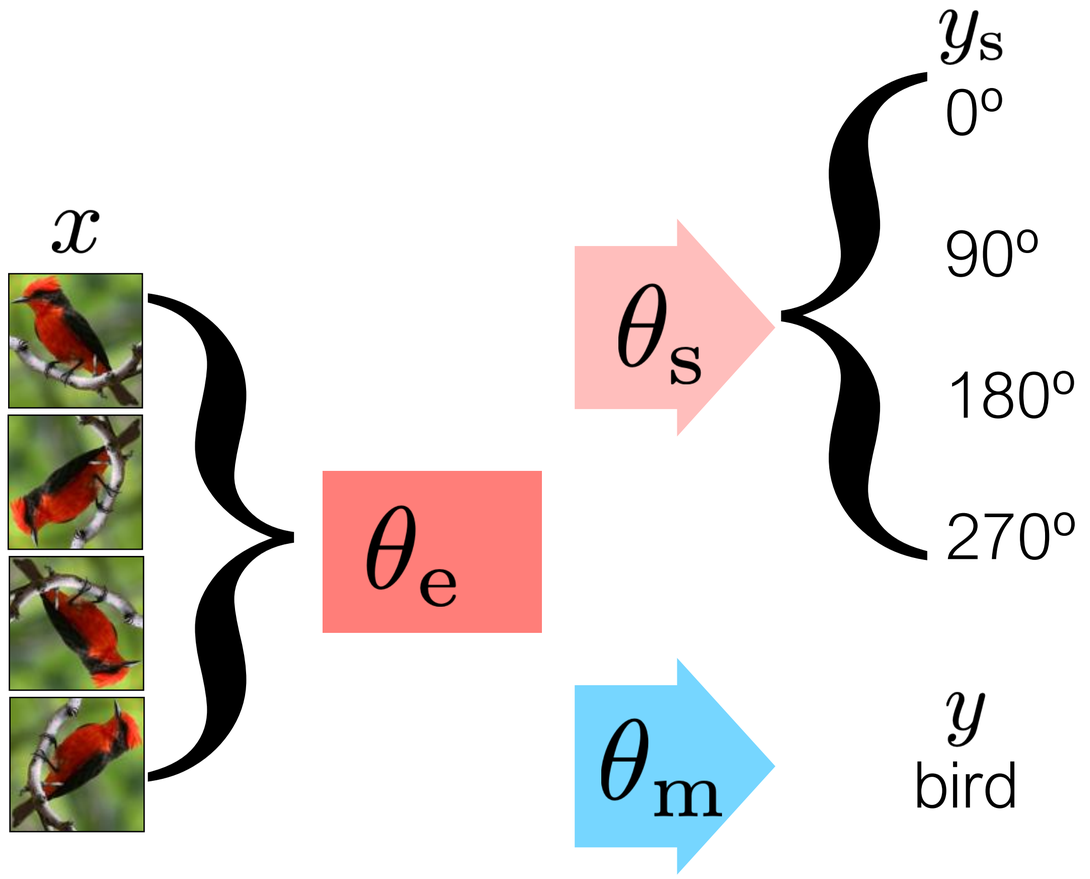


TTT-Rotation on ImageNet-C





Hidden Assumption: task alignment



We assume that self-supervised task is aligned with my main task

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Meta AI

Test-Time Training with Masked Autoencoders

Yossi Gandelsman* — UC Berkeley

Yu Sun* — UC Berkeley

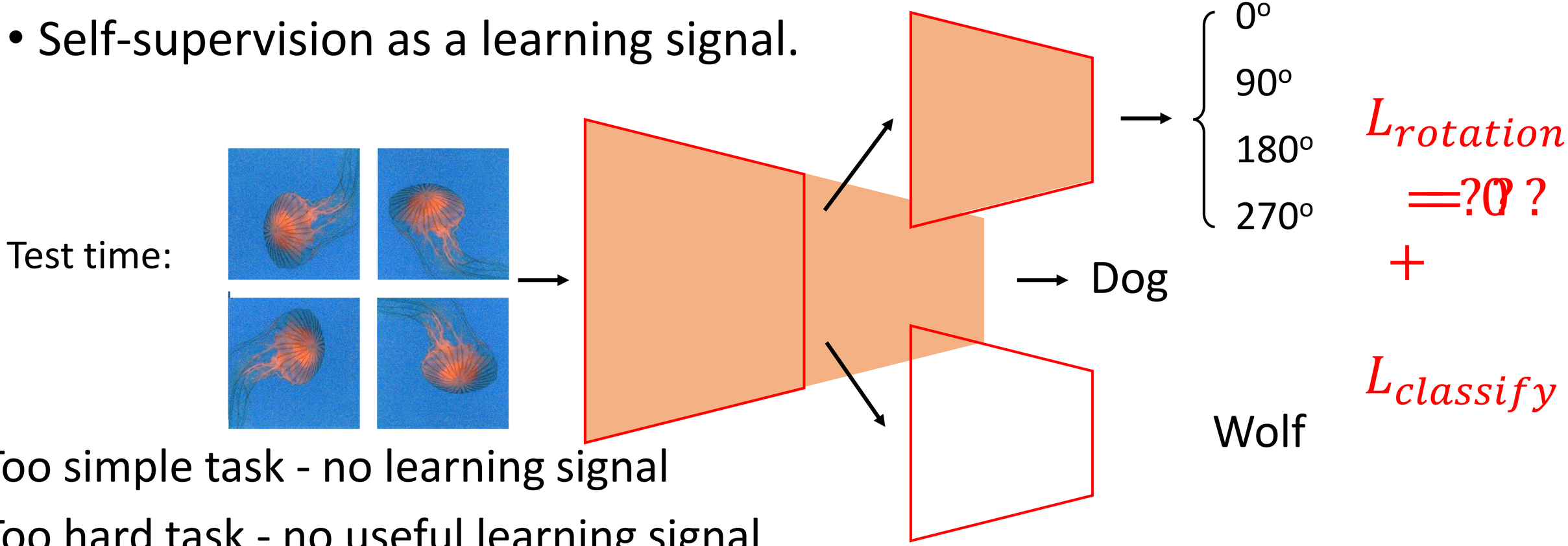
Xinlei Chen — Meta AI

Alexei Efros — UC Berkeley

*Equal contribution

Test-Time Training (TTT)

- The test input gives a hint about the test distribution.
- Train on the test input before making a prediction.
- Self-supervision as a learning signal.

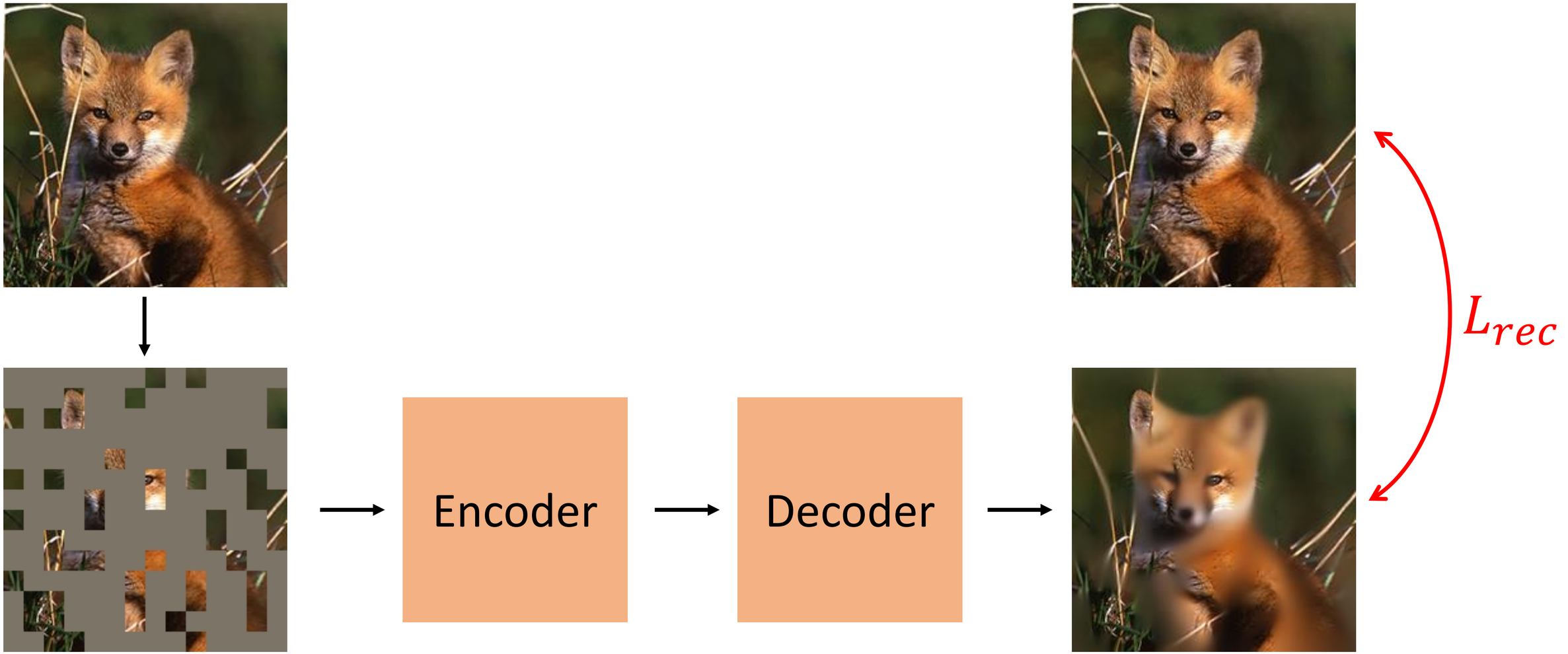


Too simple task - no learning signal

Too hard task - no <u>useful</u> learning signal

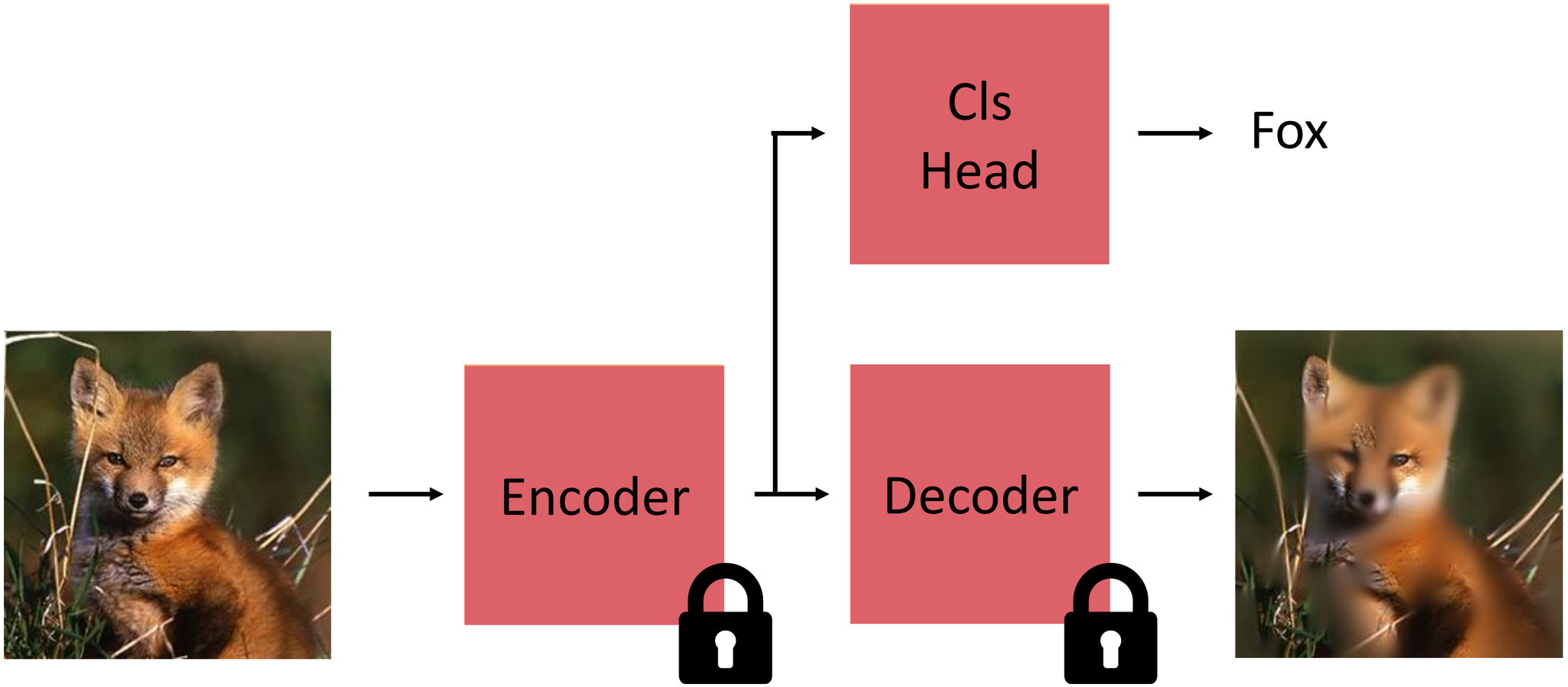
Sun et al., “Test-Time Training with Self-Supervision for Generalization under Distribution Shifts”, ICML 2020

Masked Autoencoders



He et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022

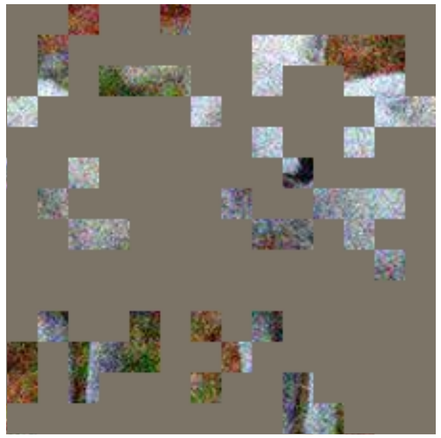
Masked Autoencoders



He et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022

Masked Autoencoders

Test input

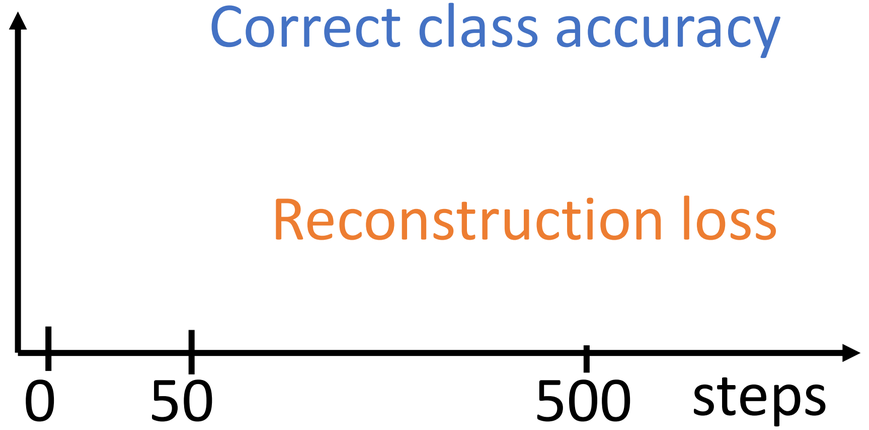


Encoder

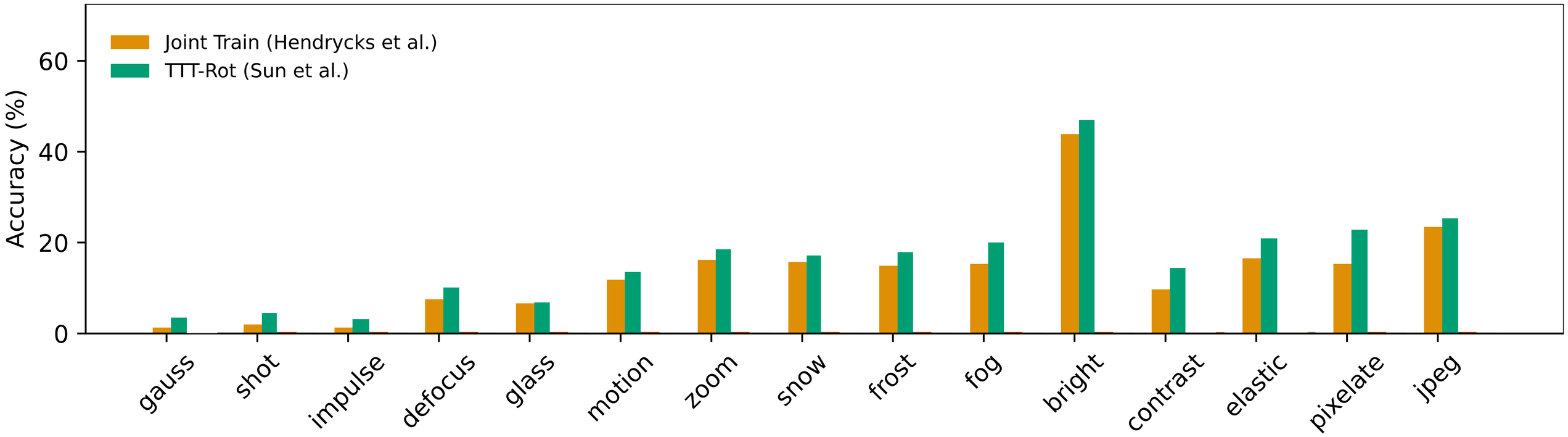
Cls Token

Decoder

Wolf



Test-Time Training on ImageNet-C



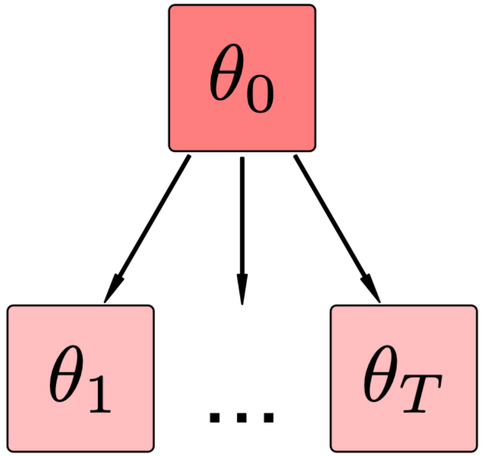
Hendrycks et al., “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”, ICLR 2019

Multiple test inputs $x_1, \ldots, x_T$

Reset to θ_0

θ_0: parameters after joint training

no assumption on test inputs

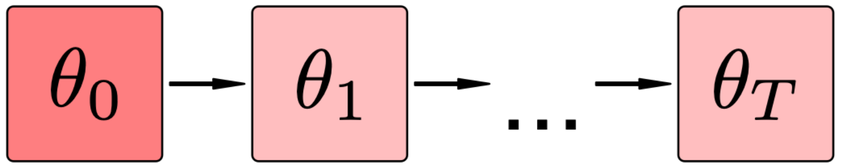


No reset (online)

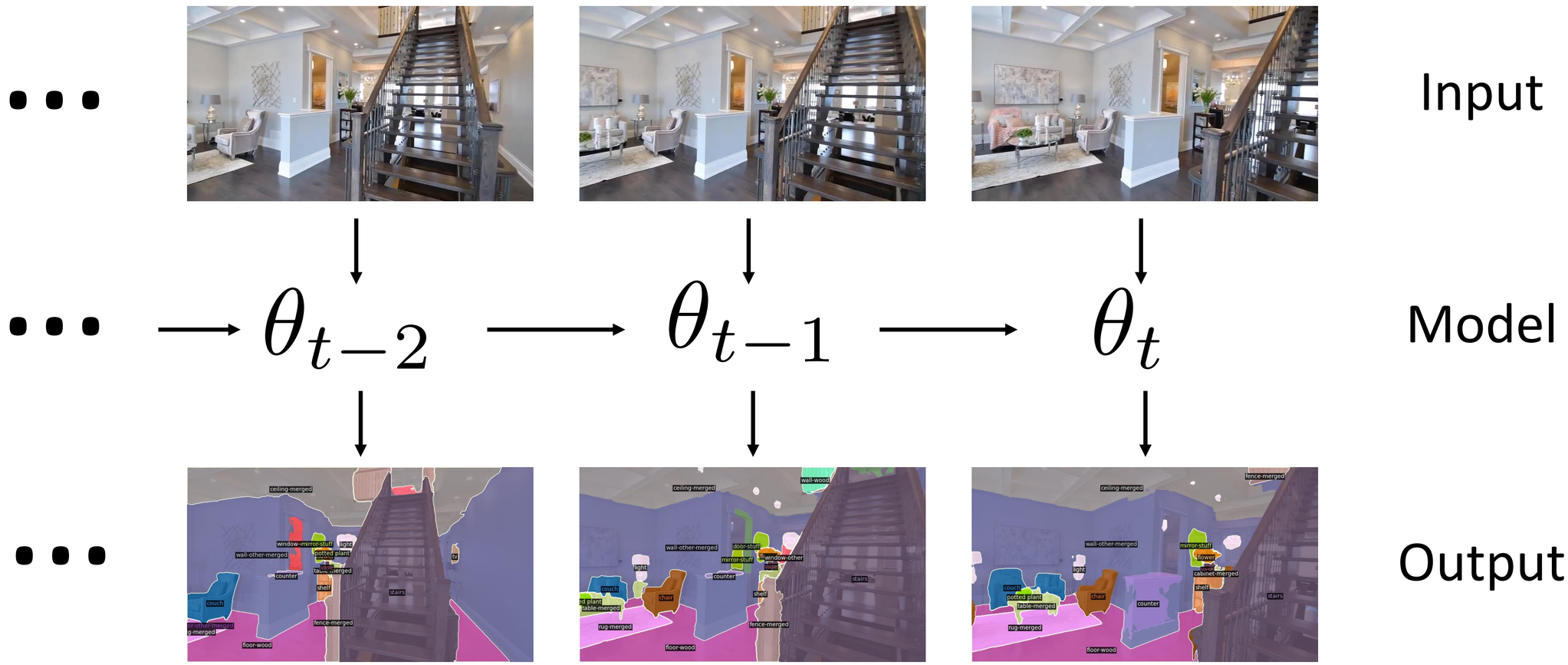
$x_1, \ldots, x_T$ come from

smoothly changing data stream

e.g. video



Test-Time Training on Video Streams



Test-Time Training on Video Streams (Coming soon on arXiv)

Renhao Wang*, Yu Sun*, Alexei Efros, Xiaolong Wang

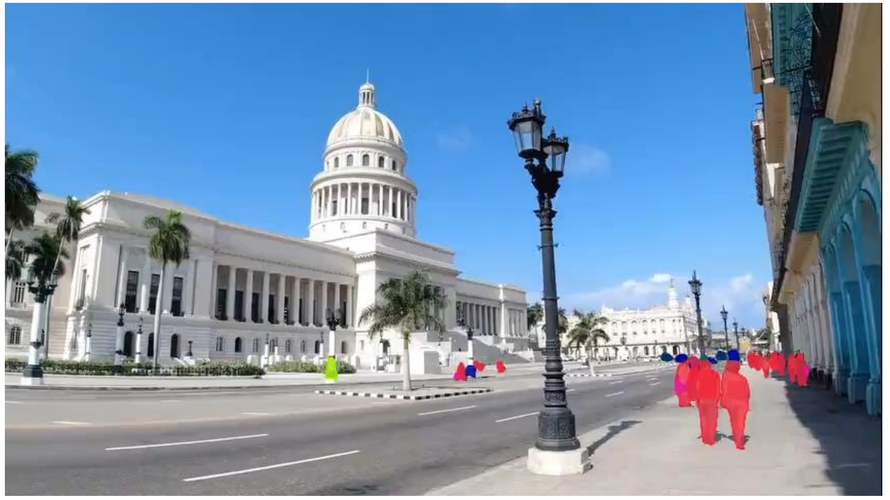
Online version - videos

“Test-time training on video streams”, Wang et al. 2022

Results – instance segmentation

Input

Baseline



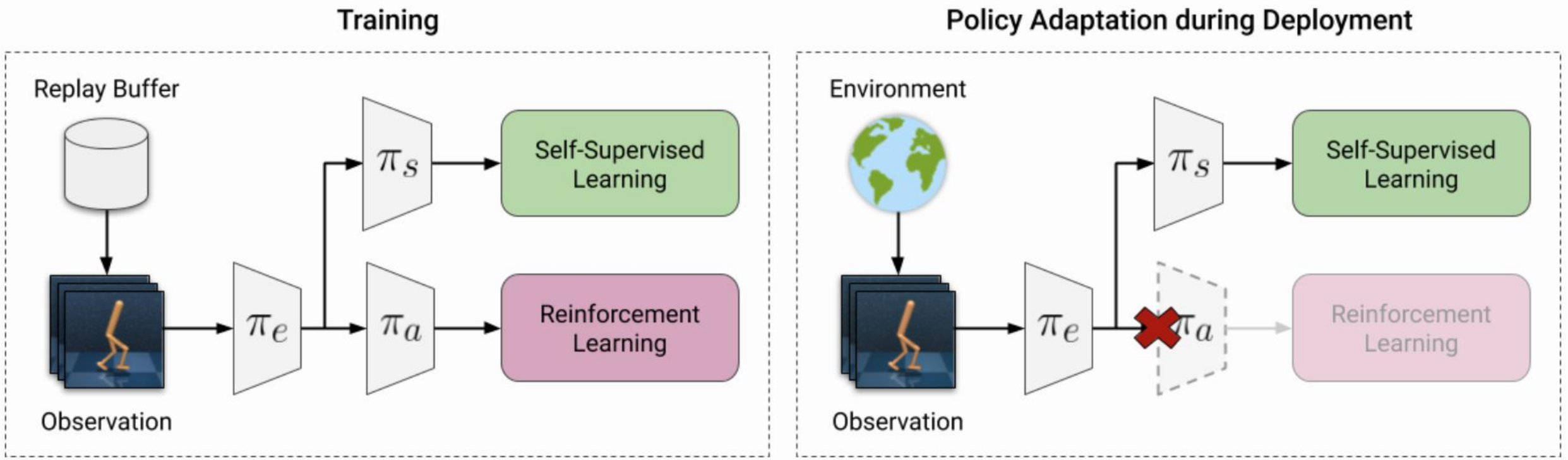
TTT

Results - colorization

Baseline

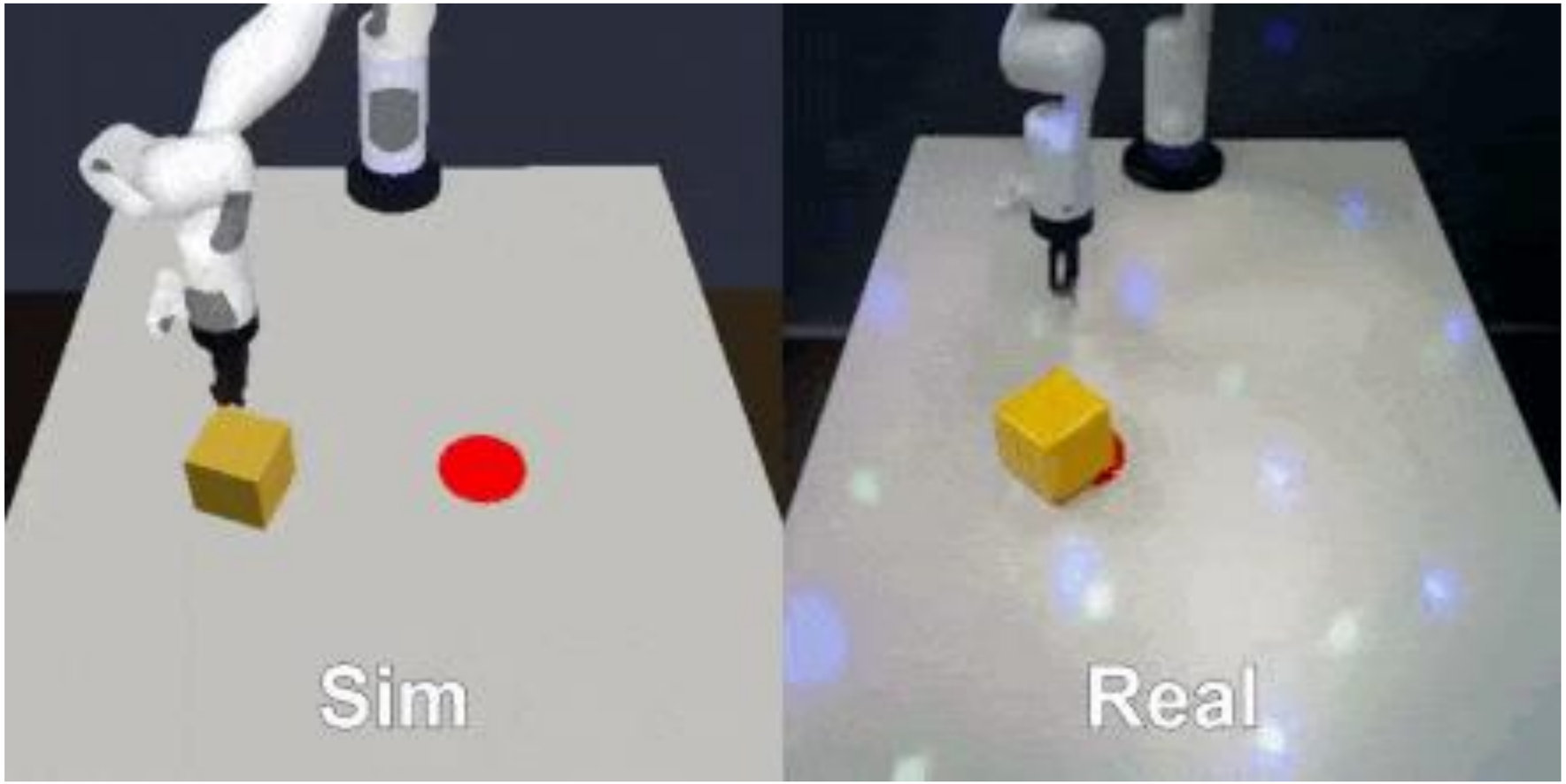
TTT

Self-Supervised Policy Adaptation during Deployment [ICLR'21]



Nicklas Hansen, Yu Sun, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, Xiaolong Wang, ICLR 2021

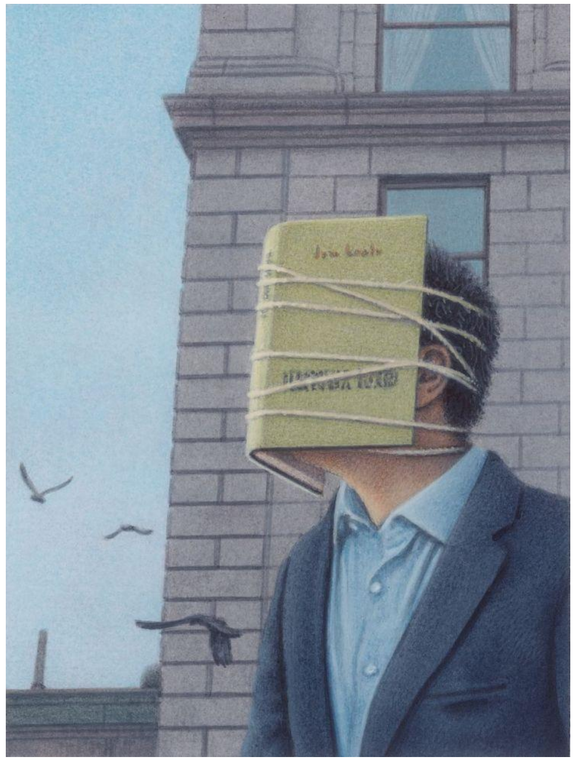
Self-Supervised Policy Adaptation during Deployment [ICLR'21]



Nicklas Hansen, Yu Sun, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, Xiaolong Wang, ICLR 2021

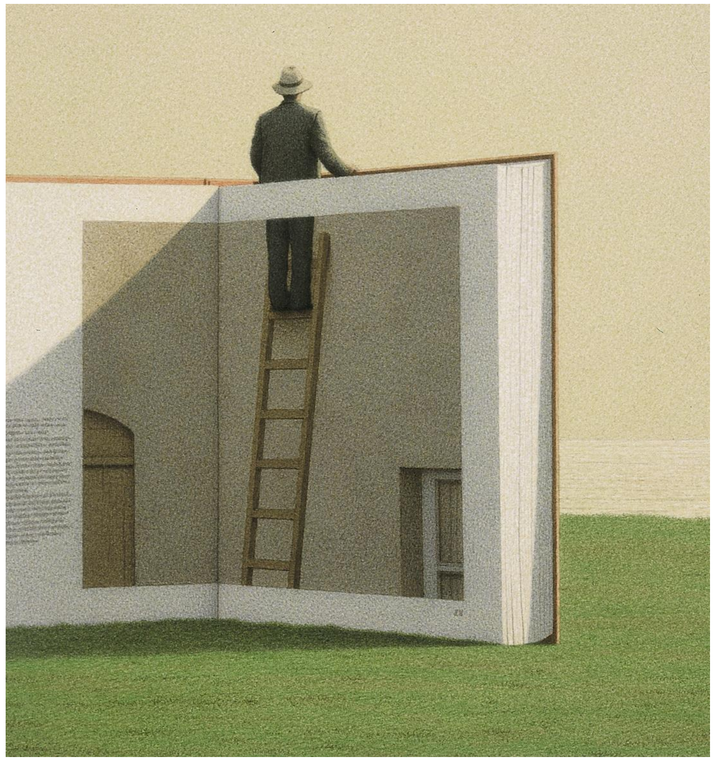
Why Self-Supervision?

1. To get away from <u>semantic categories</u>



⇨ data-driven association

2. To get away from <u>fixed datasets</u>



⇨ continuous, life-long learning

3. To get away from <u>fixed objectives</u>

⇨ emergent objectives

Artwork by *Quint Buchholz*