Self-Supervision for Learning from the Bottom Up

© Quint Buchholz

Alexei A. Efros

UC Berkeley

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

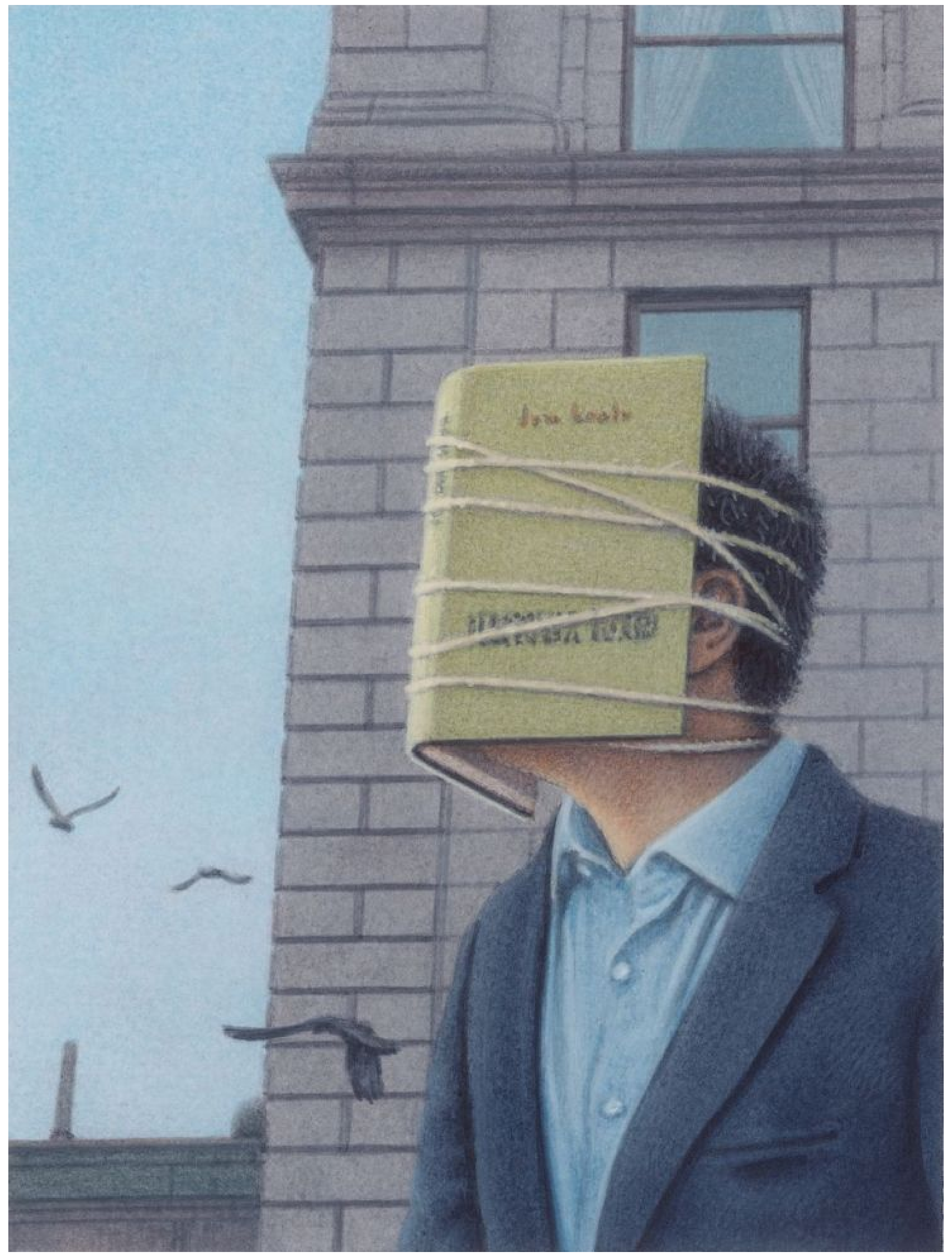
Why Self-Supervision?

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

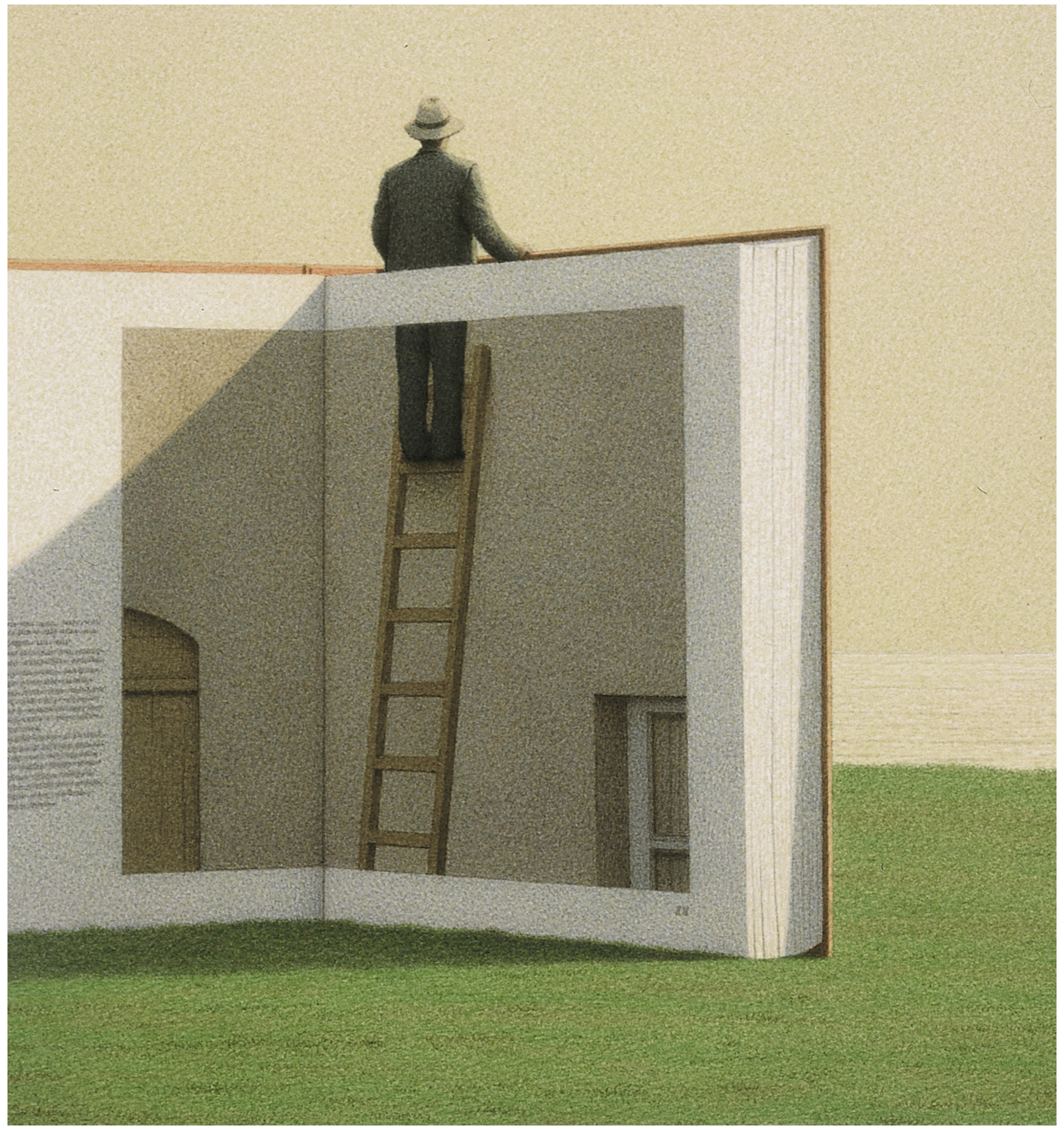
Earth Population: 7.8 billion

Why Self-Supervision?

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



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

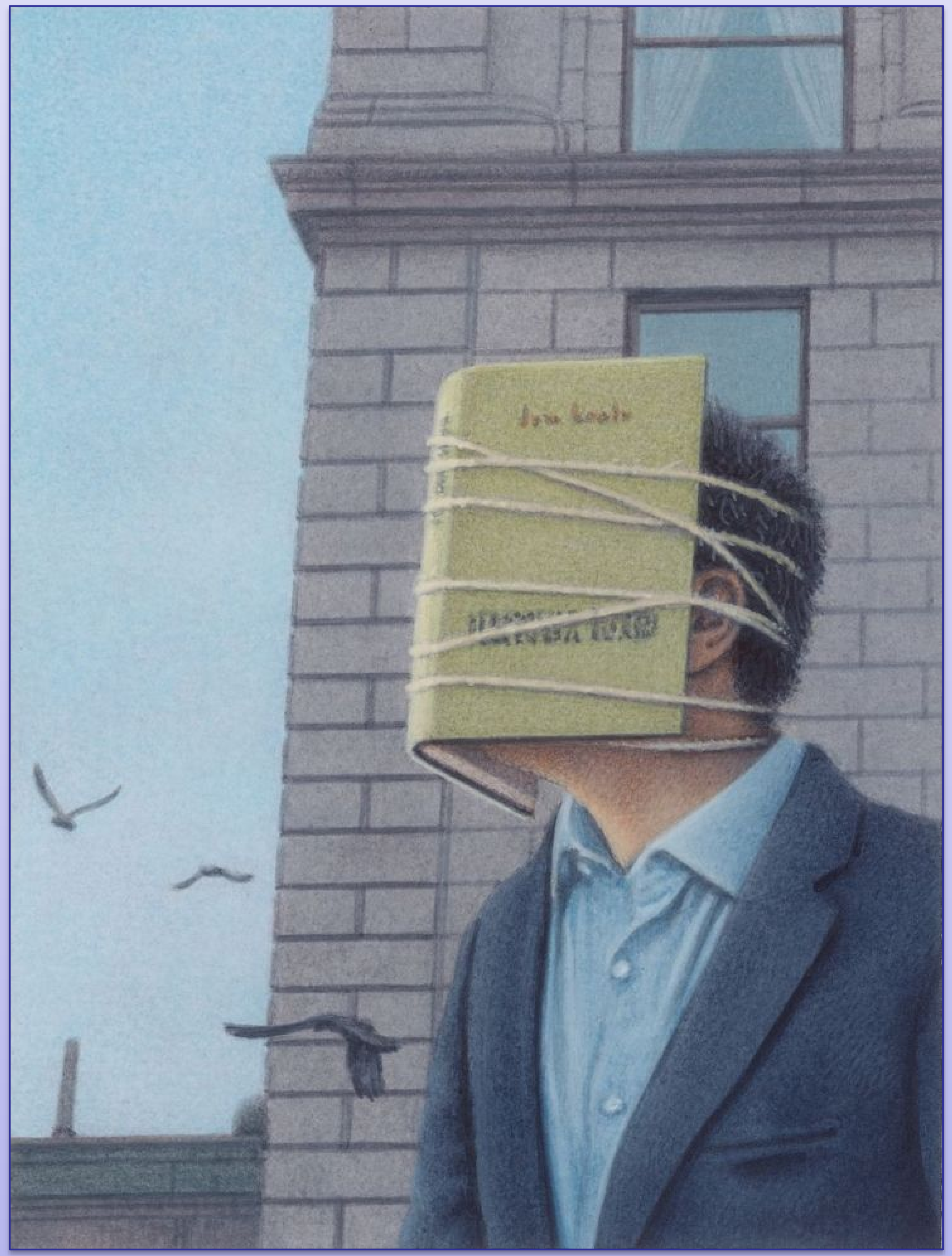


3. To get away from <u>fixed objectives</u>

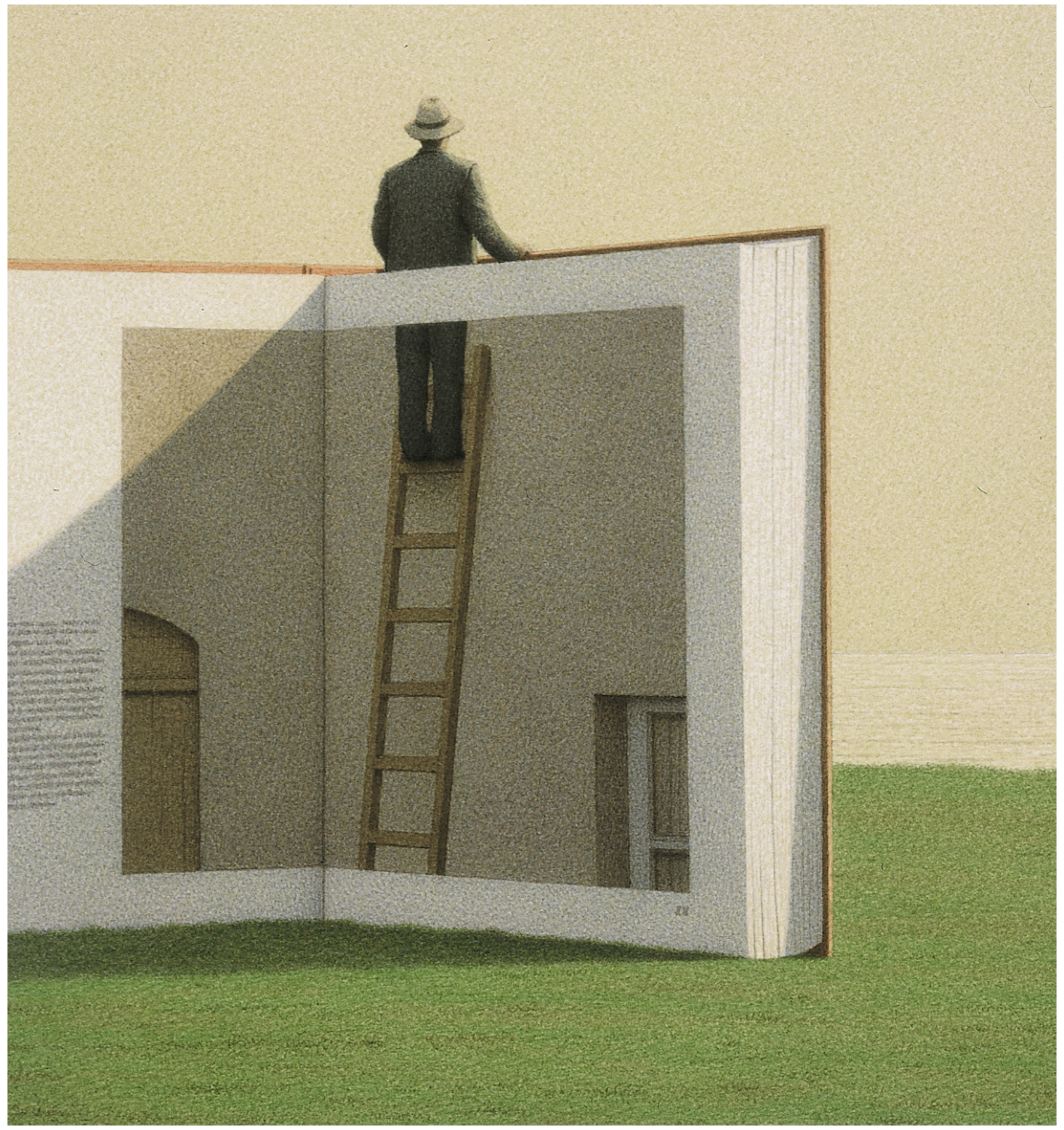
Artwork by *Quint Buchholz*

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>

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)

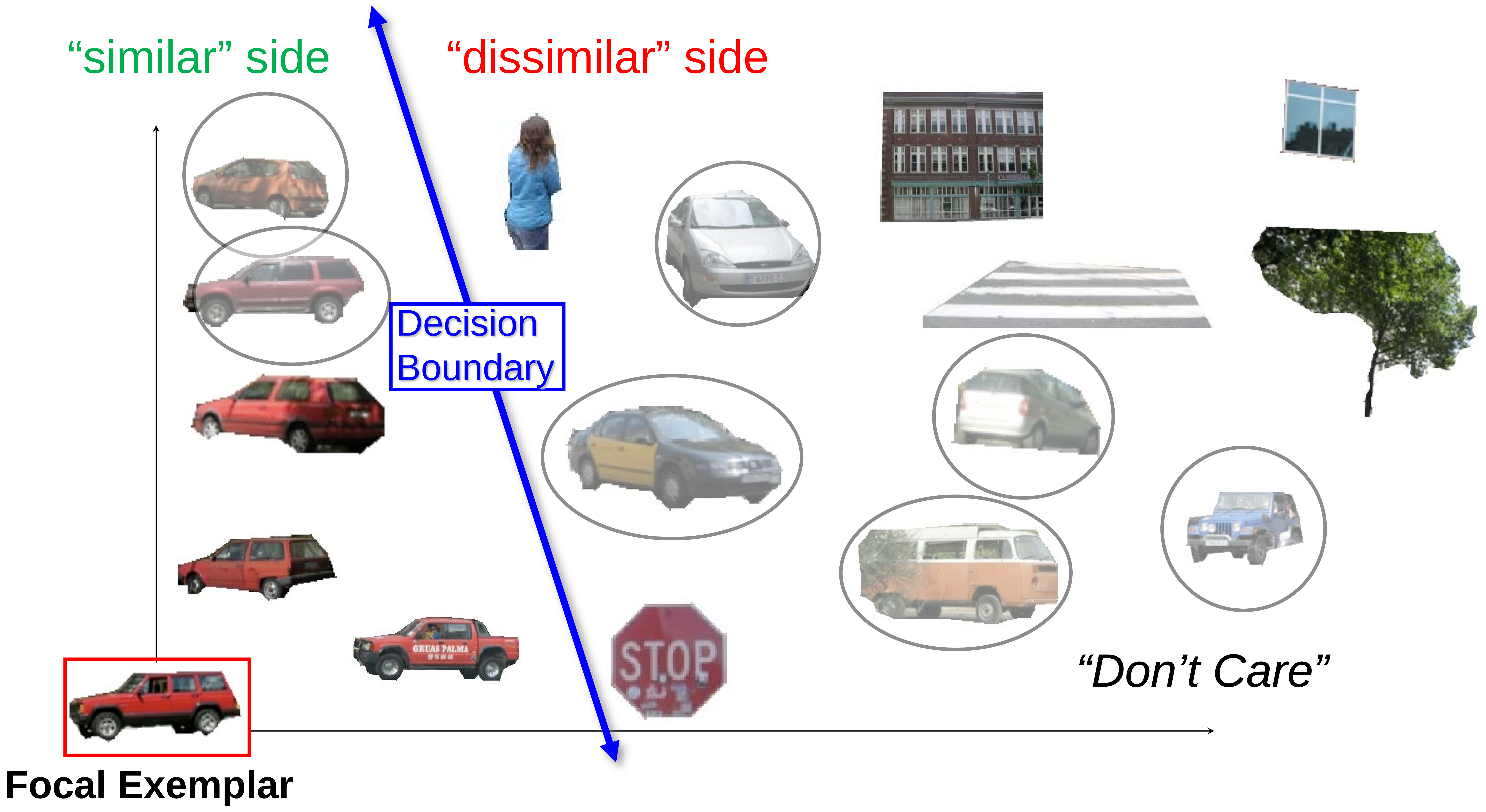
Ask not *"What is it?"*
Ask "What is it <u>like</u>?"

-- Moshe Bar, 2008

Learning Per-Exemplar Distances (CVPR 2008)

[Malisiewicz & Efros. *Recognition by Association via Learning Per-exemplar Distances.* CVPR'08]()

Learning Per-Exemplar Distances (CVPR 2008)



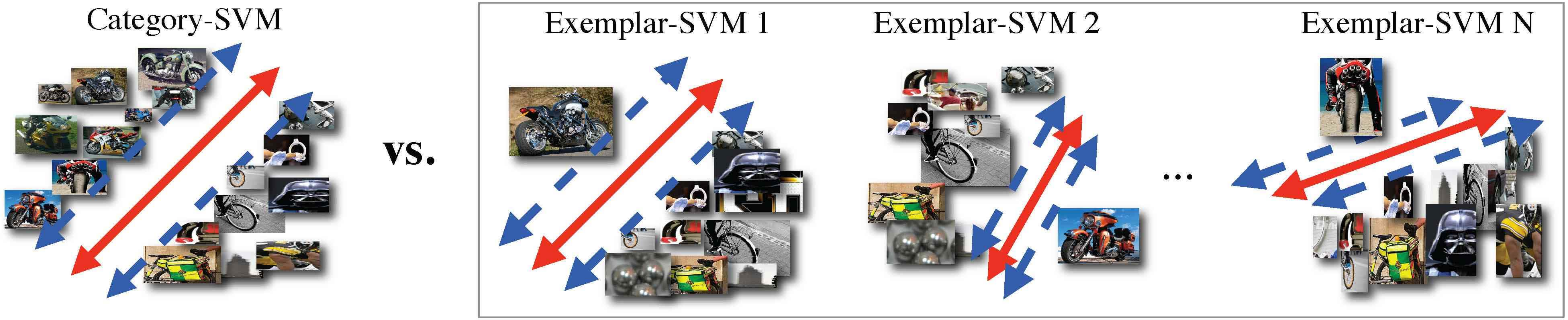
[Malisiewicz & Efros. *Recognition by Association via Learning Per-exemplar Distances.* CVPR'08]()

Learning Per-Exemplar Distances (CVPR 2008)

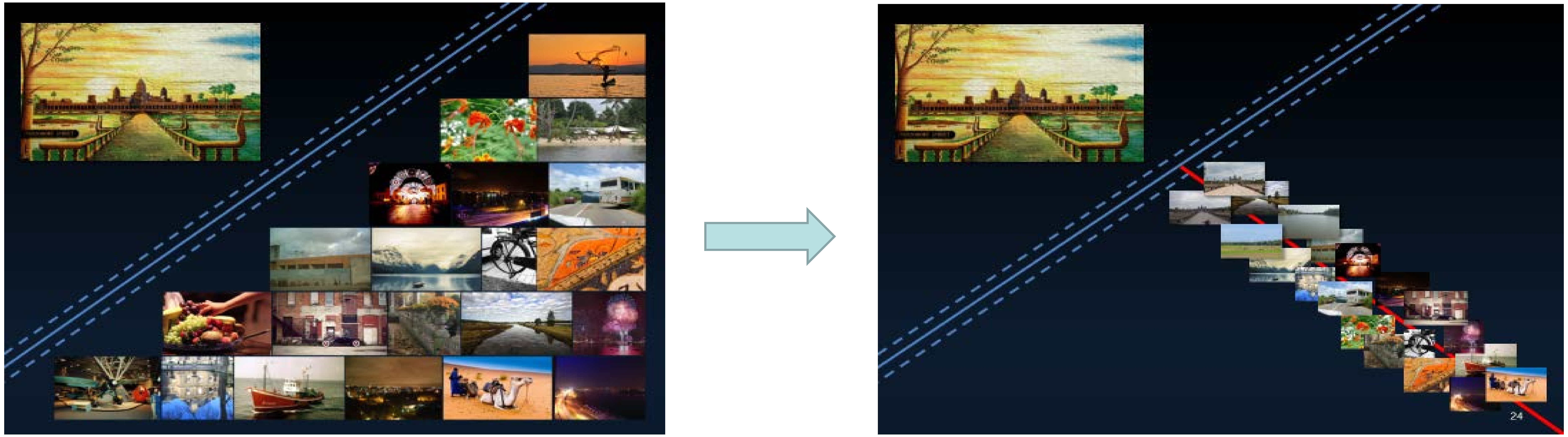
Malisiewicz & Efros. *Recognition by Association via Learning Per-exemplar Distances.* CVPR'08

Exemplar-SVM: *defining yourself by what you are <u>not</u>*

[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

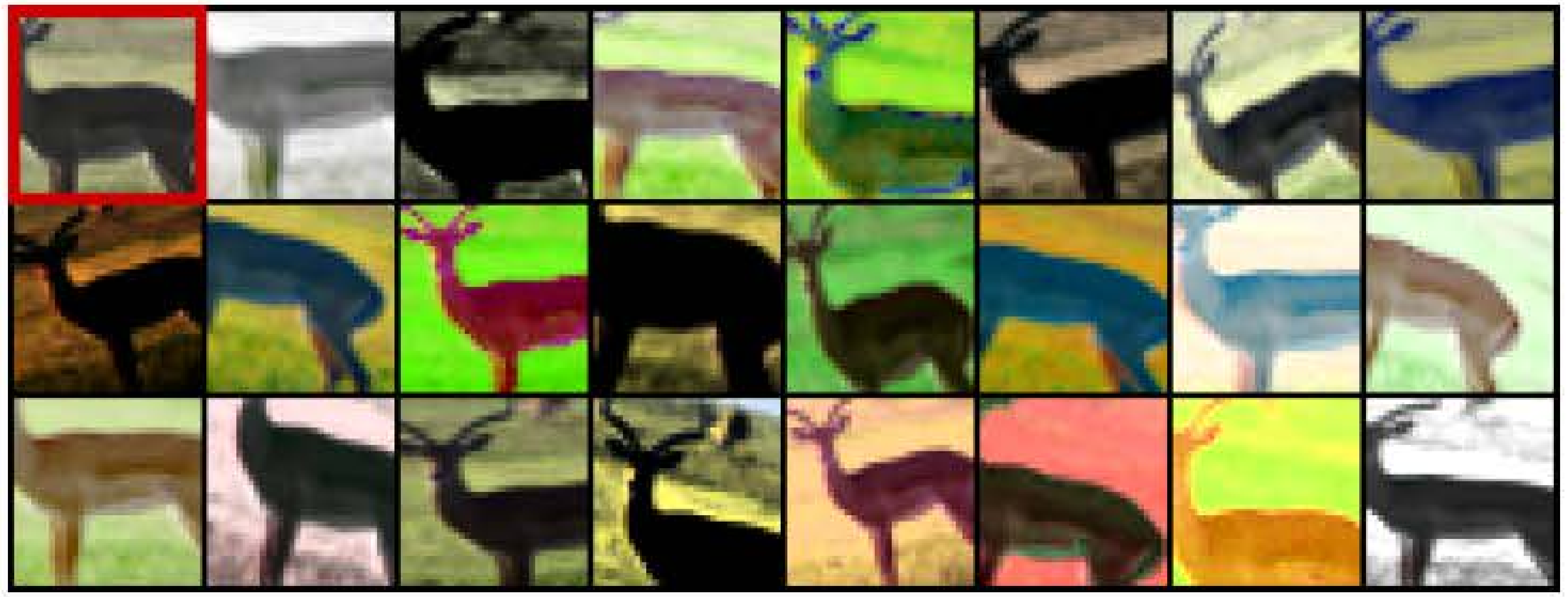
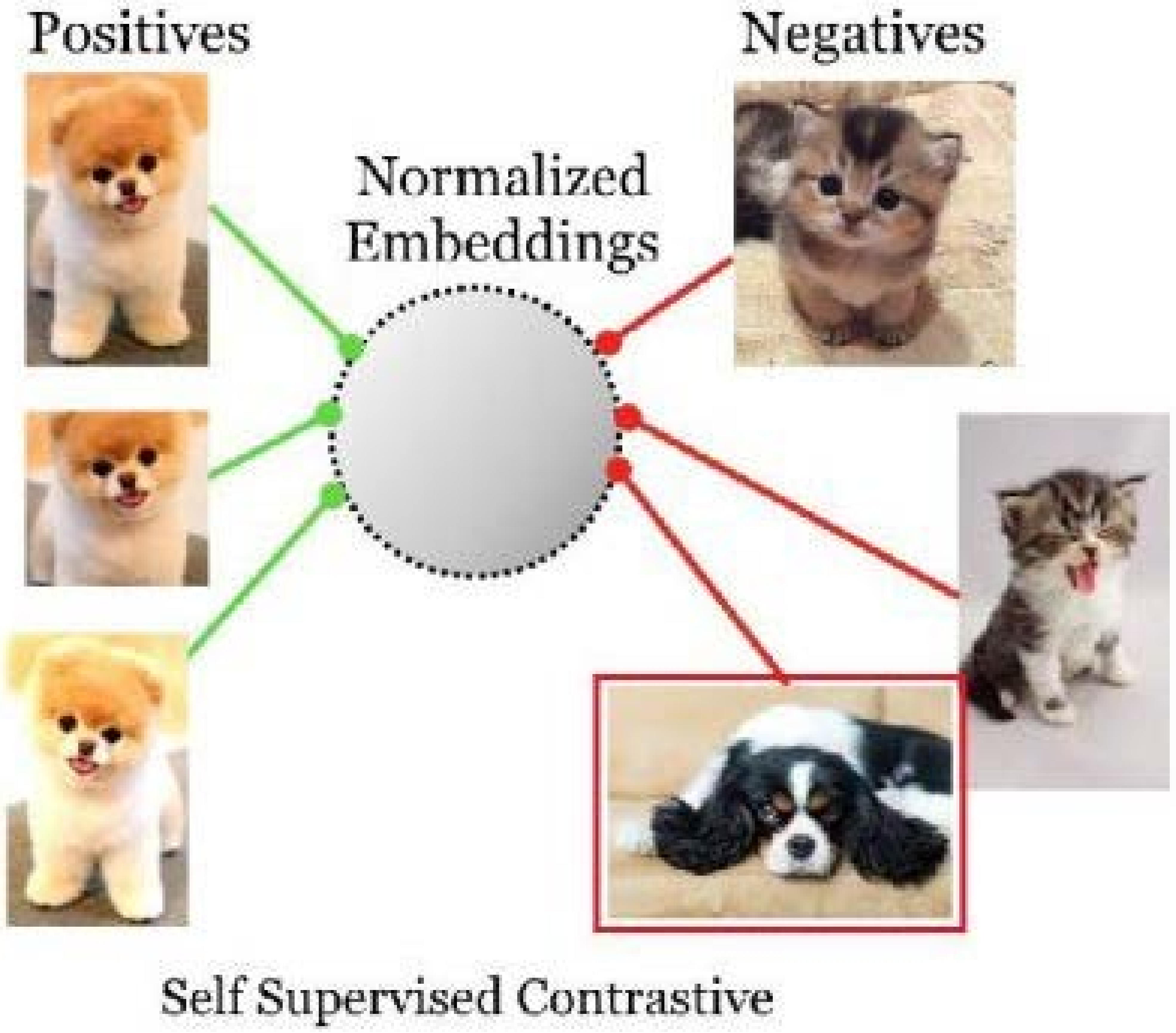


Fig. 2. Several random transformations applied to one of the patches extracted from the STL unlabeled dataset. The original ('seed') patch is in the top left corner.

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

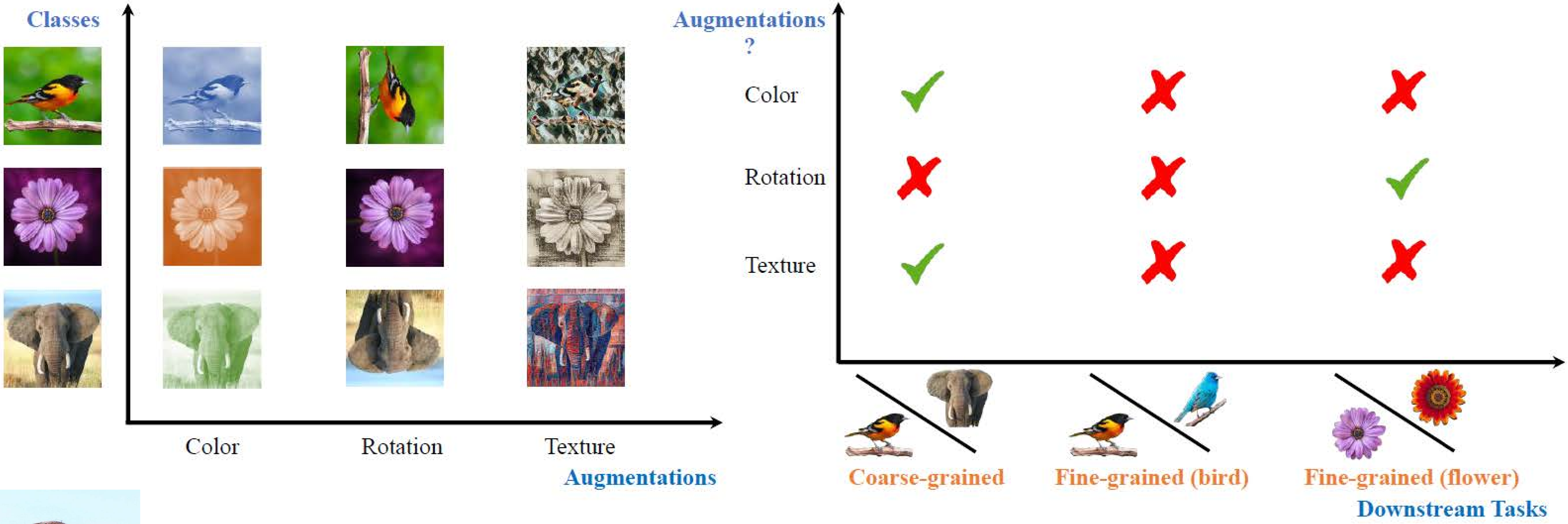
input

color
crop
flip
blur

Views

SimCLR augmentations (Chen et al, 2020)

The choice of data augmentation is itself supervision



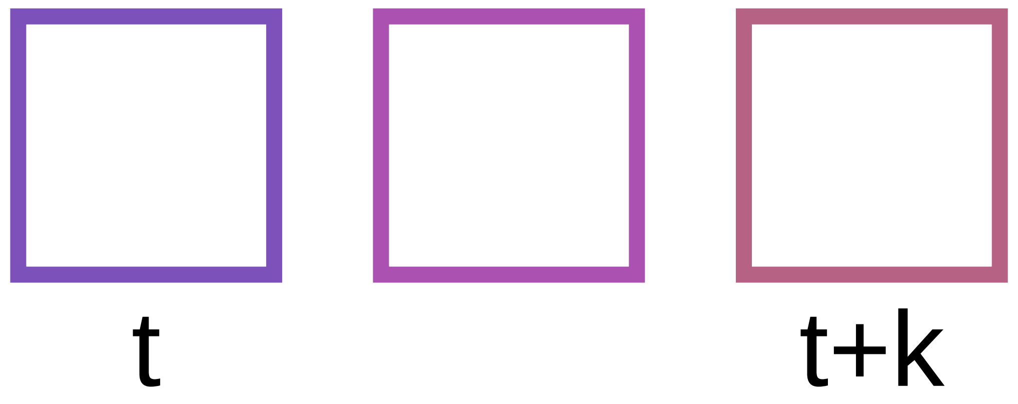
What should not be contrastive in contrastive learning
T Xiao, X Wang, AA Efros, T Darrell - ICLR 2021

Video as Data Augmentation

Wehrtheimer (1938)

Learning Representations from Video

Temporal Coherence across Frames



Slow Features, Sparse Coding, ICA

Foldiak (1989)
Wiskott & Sejnowski (2001)
Olshausen et al (2003)
Hurri & Hyvarinen (2003)

Auto-encoders

Hinton (1989)
Zou et al (2011)
Goroshin et al (2013)

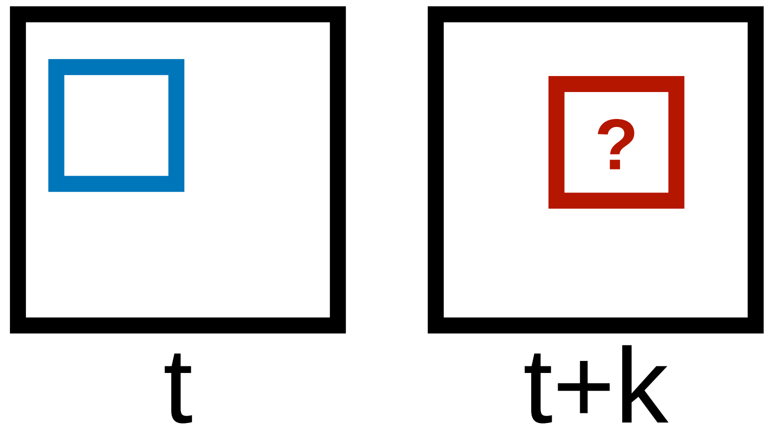
Contrastive Learning

Hyvarinen et al (2016)
Sermanet et al (2018)
Gordon et al (2020)

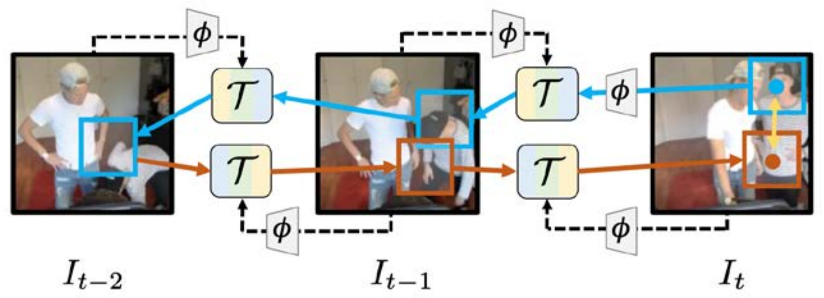
Pretext Tasks

Ranzato et al (2014)
Agrawal et al (2015)
Jayaraman et al (2016)
Mishra et al (2017)
Wei et al (2018)
Vondrick et al (2019)

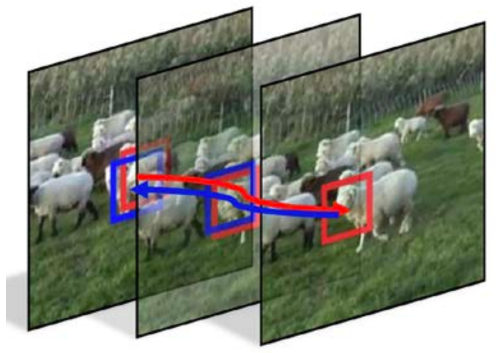
Mining Correspondence



Tracking



Wang & Jabri et al, 2019



Ning et al, 2019

Detection

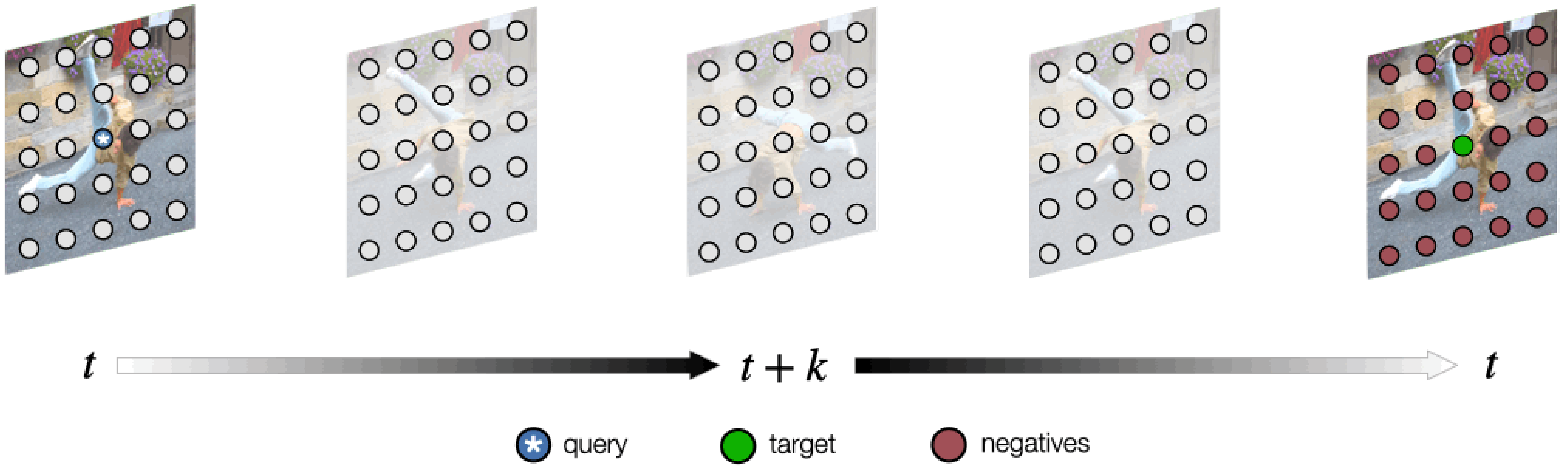
Pirk et al, 2019

Romijnders et al, 2020

Among others, e.g.:
Wang et al (2015), Lee et al (2015), Pathak et al (2016)

Space-Time Correspondence as a Contrastive Random Walk

NeurIPS 2020





Allan A. Jabri
UC Berkeley

Andrew Owens
U. Michigan

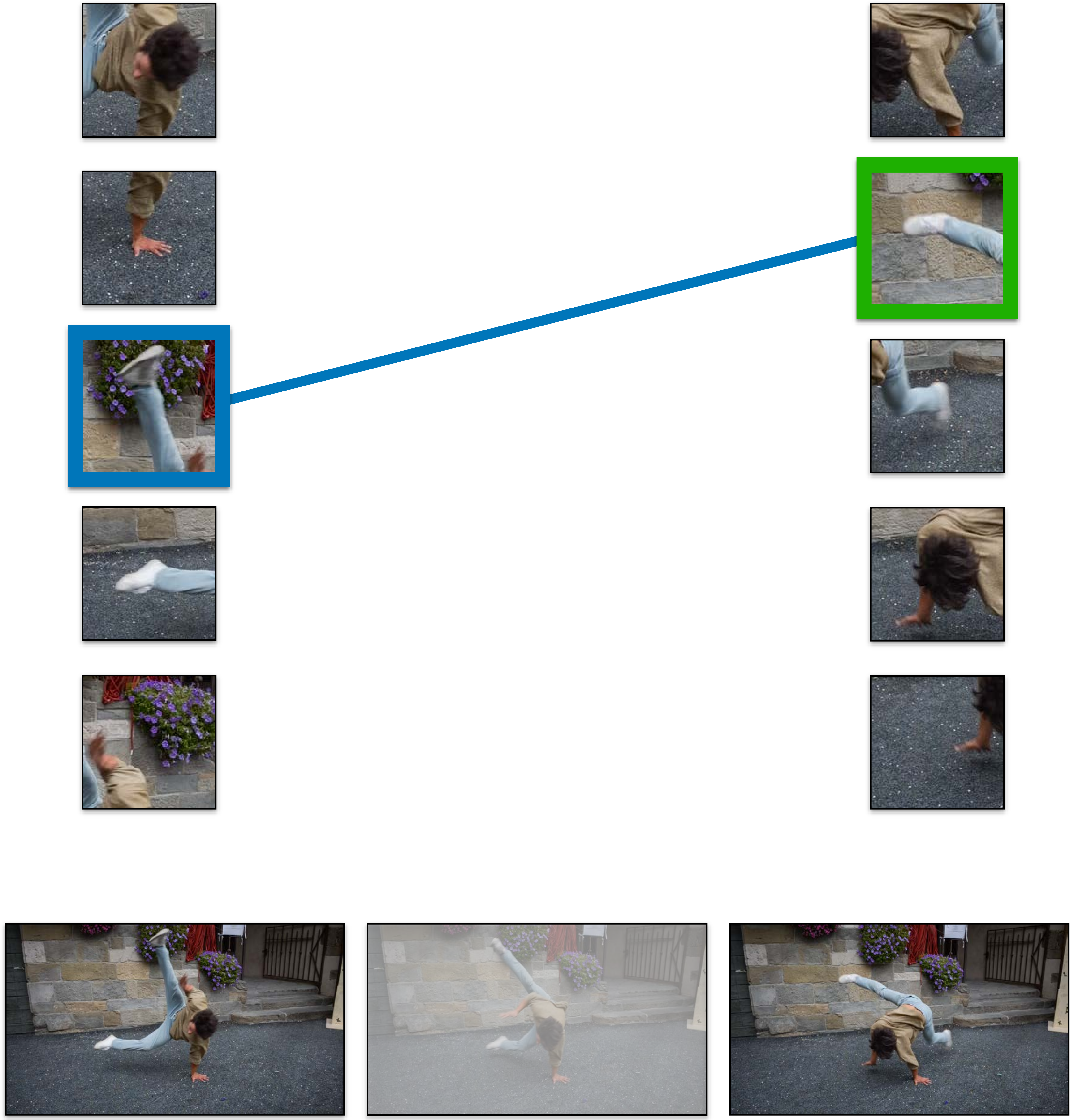
Alexei A. Efros
UC Berkeley

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

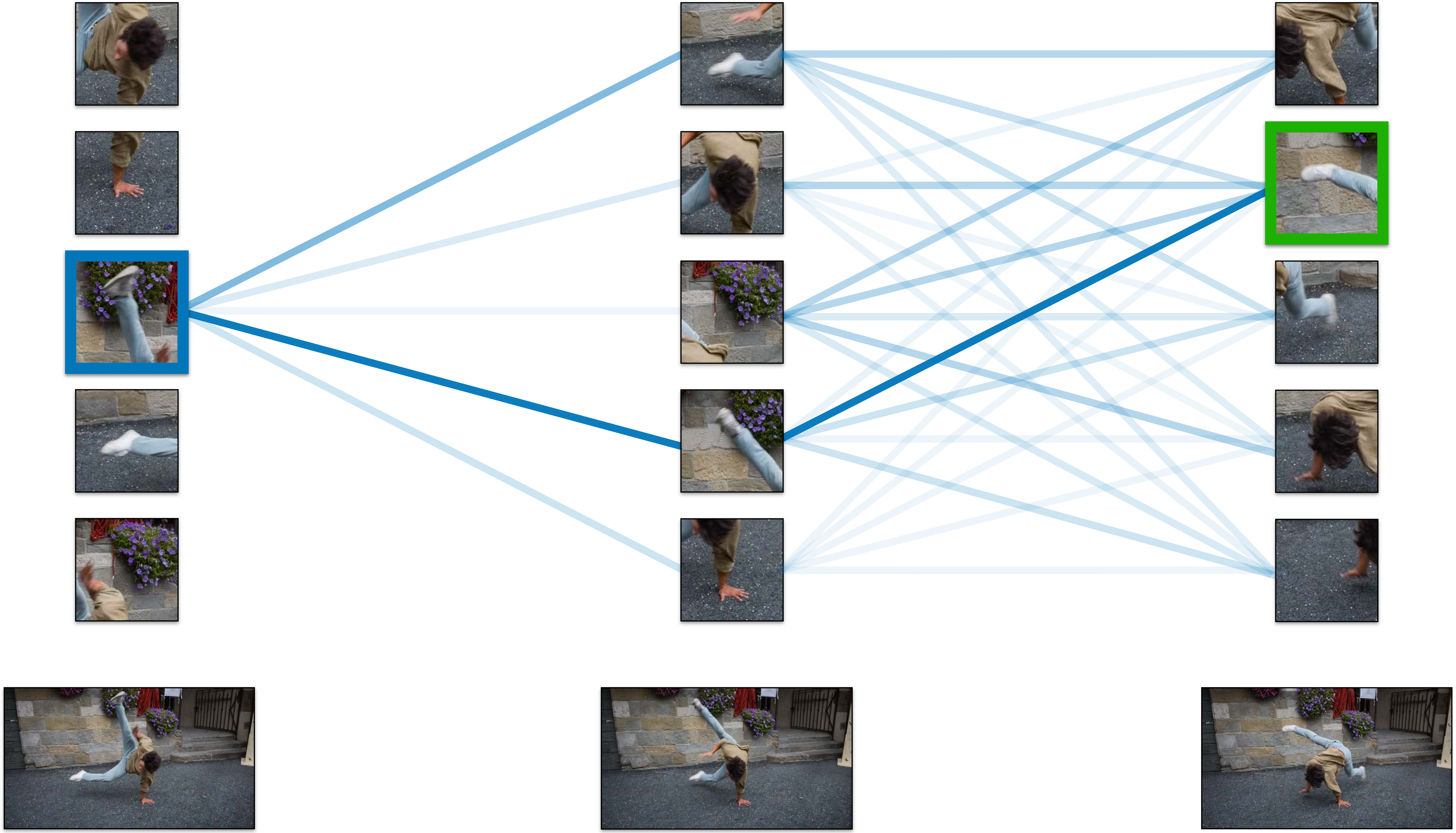
NEURAL INFORMATION PROCESSING SYSTEMS

Warm up: the supervised case

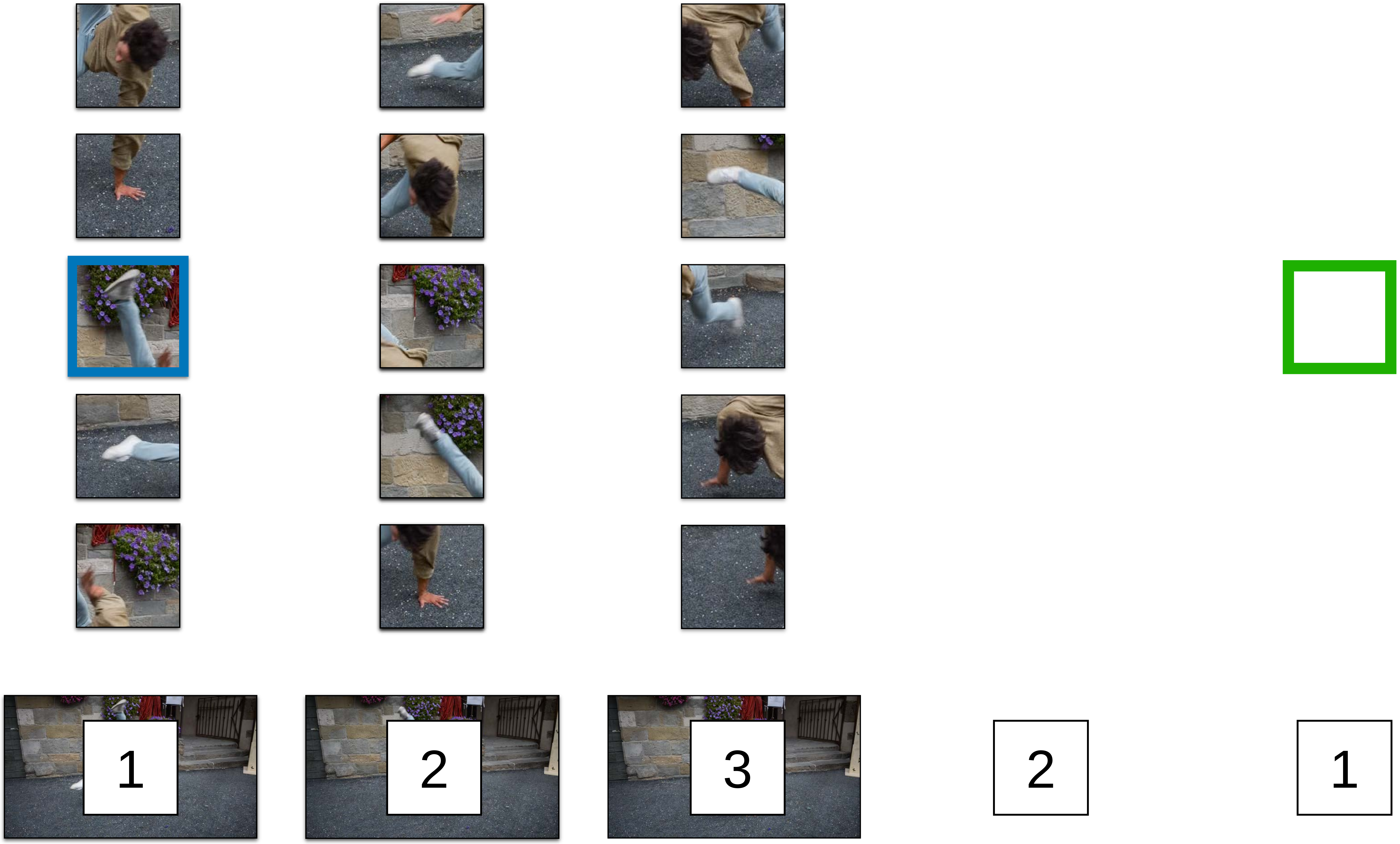
Supervised Distance Learning



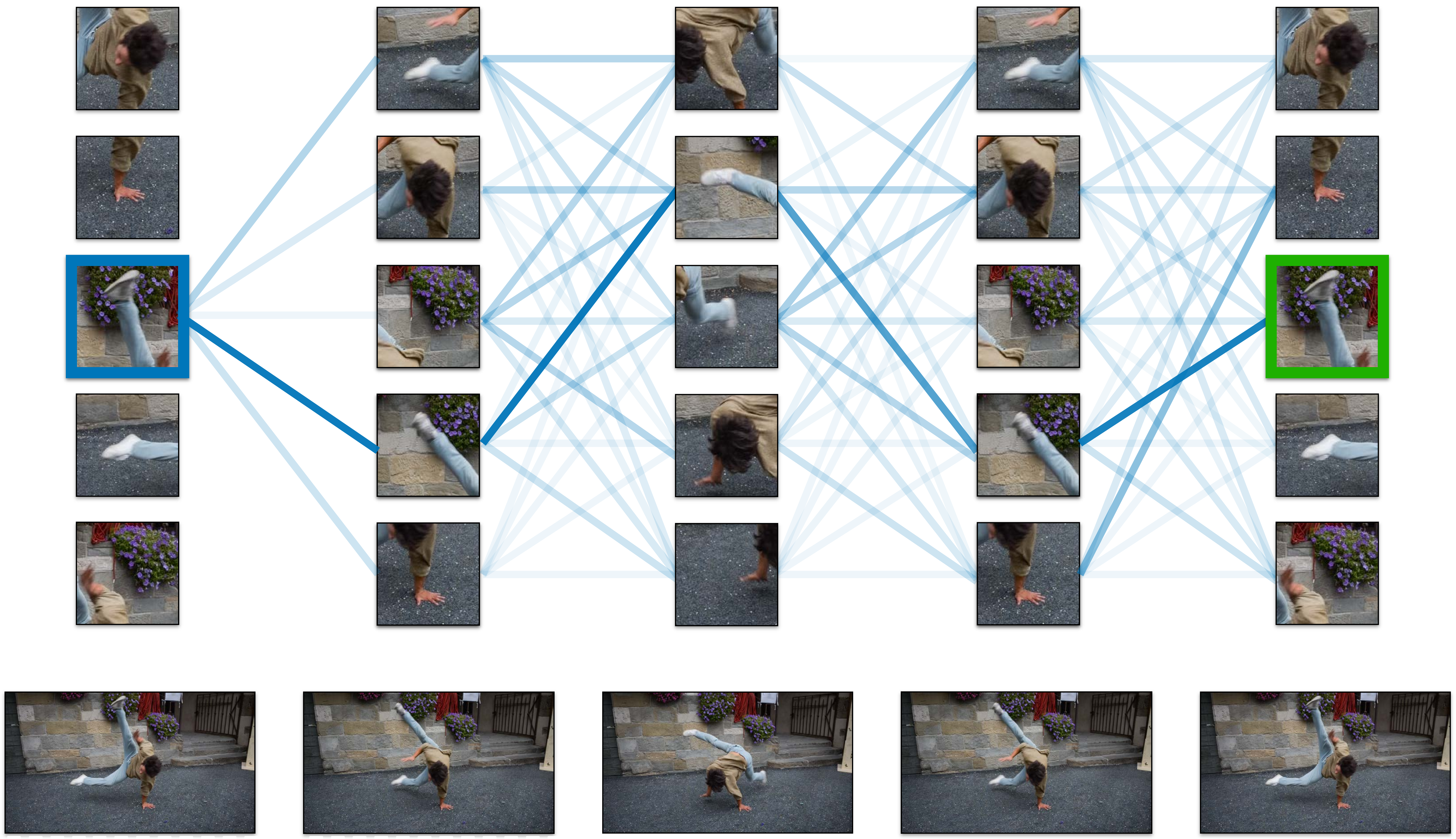
Implicit “data augmentation” with latent views

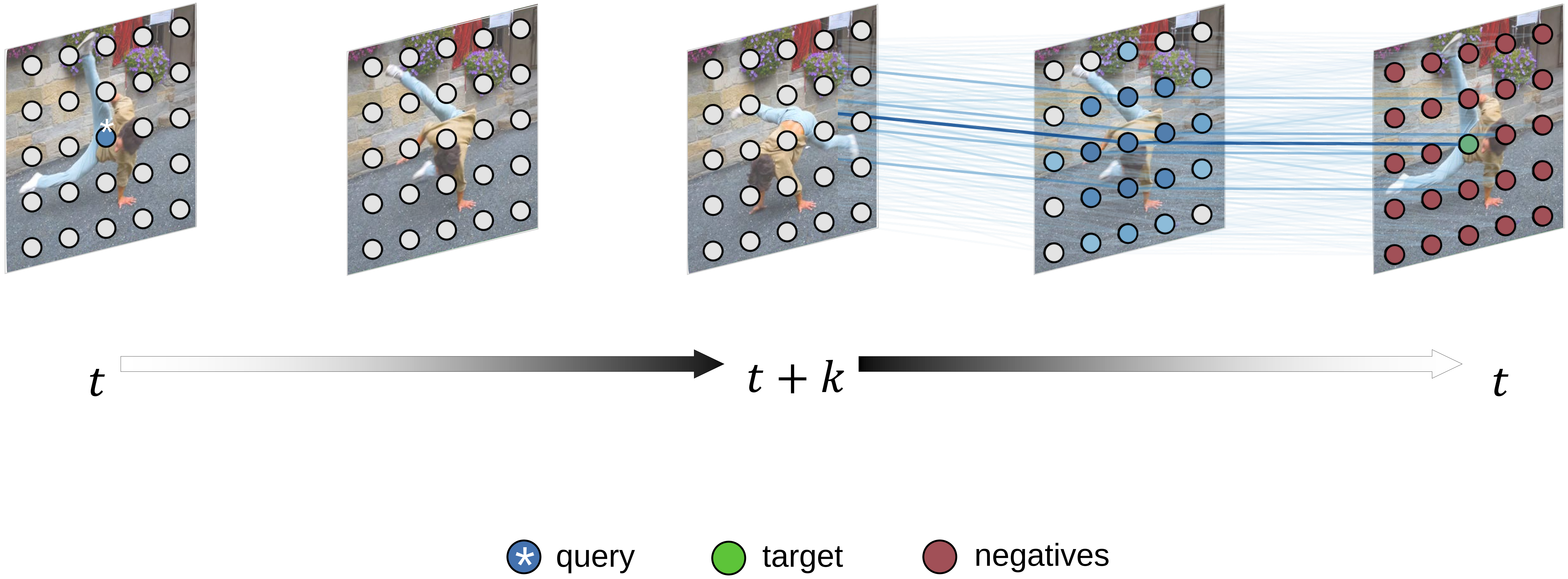


Palindromes (cycles in time)



Self-supervised Learning





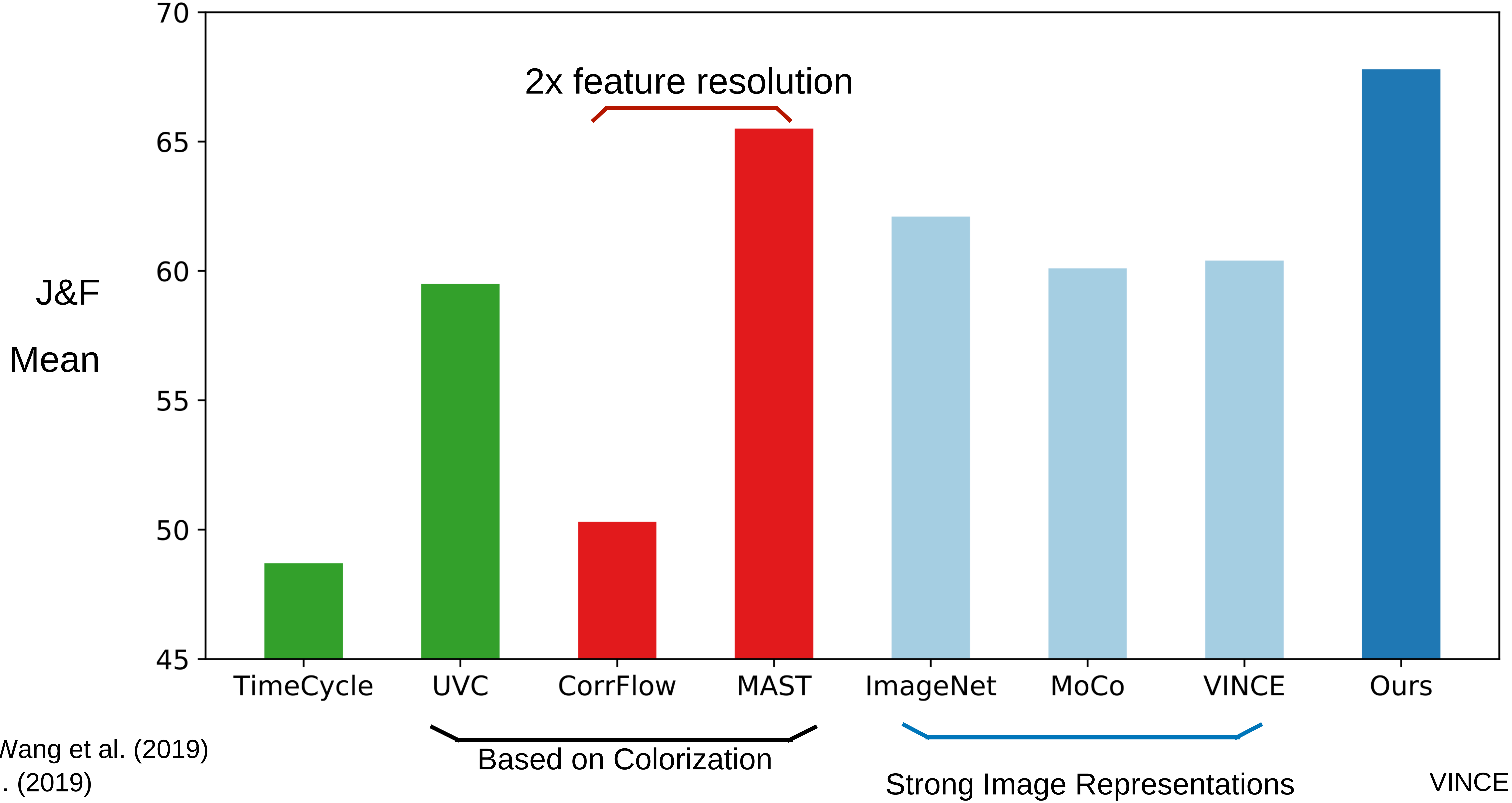
t
$t + k$
t
query
target
negatives

Qualitative Results: Video Object Propagation (DAVIS)

UVC
Li et al. (2019)

Ours

DAVIS Video Object Segmentation

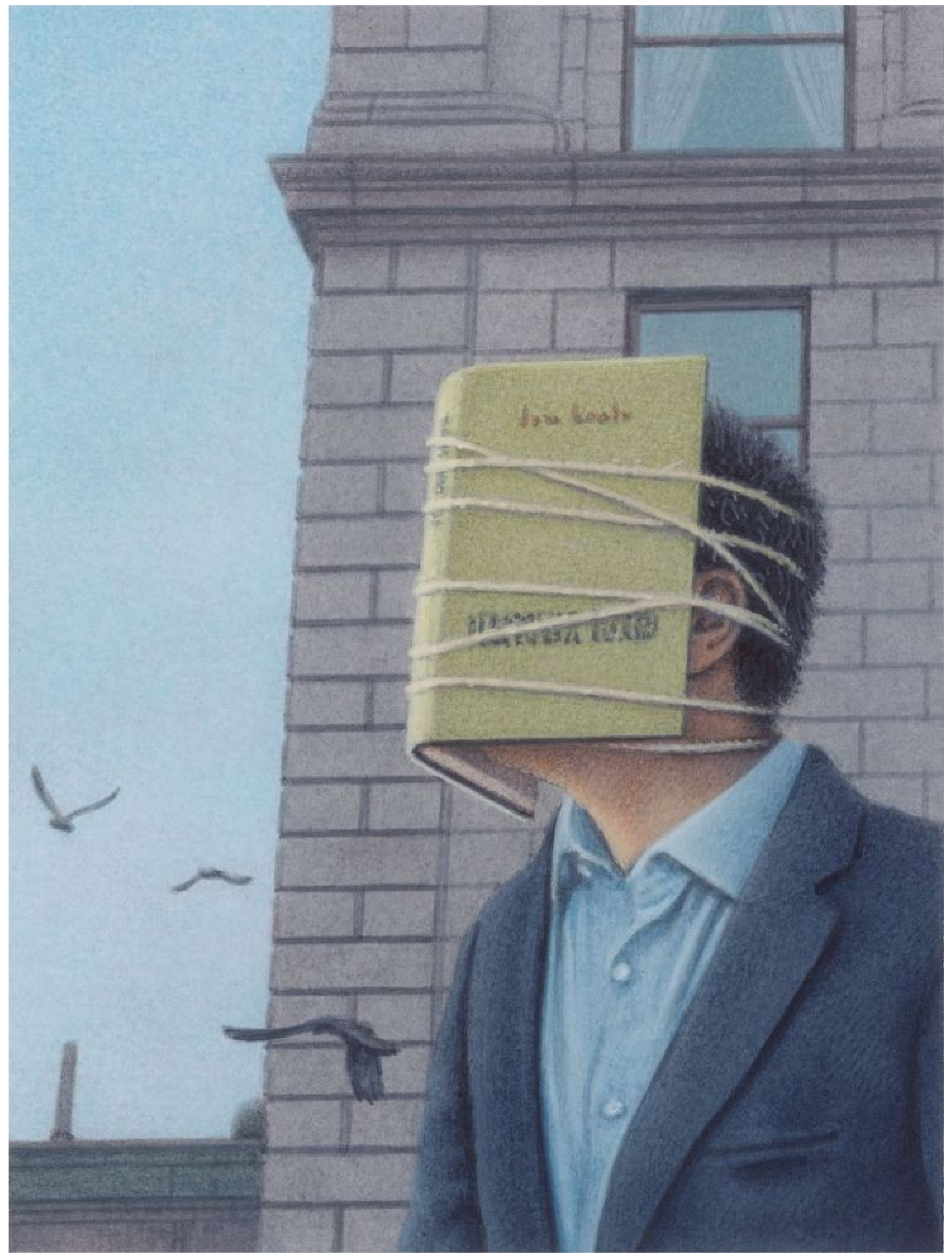


TimeCycle: Wang et al. (2019)
UVC: Li et al. (2019)
CorrFlow: Lai et al. (2019)
MAST: Lai et al. (2019)

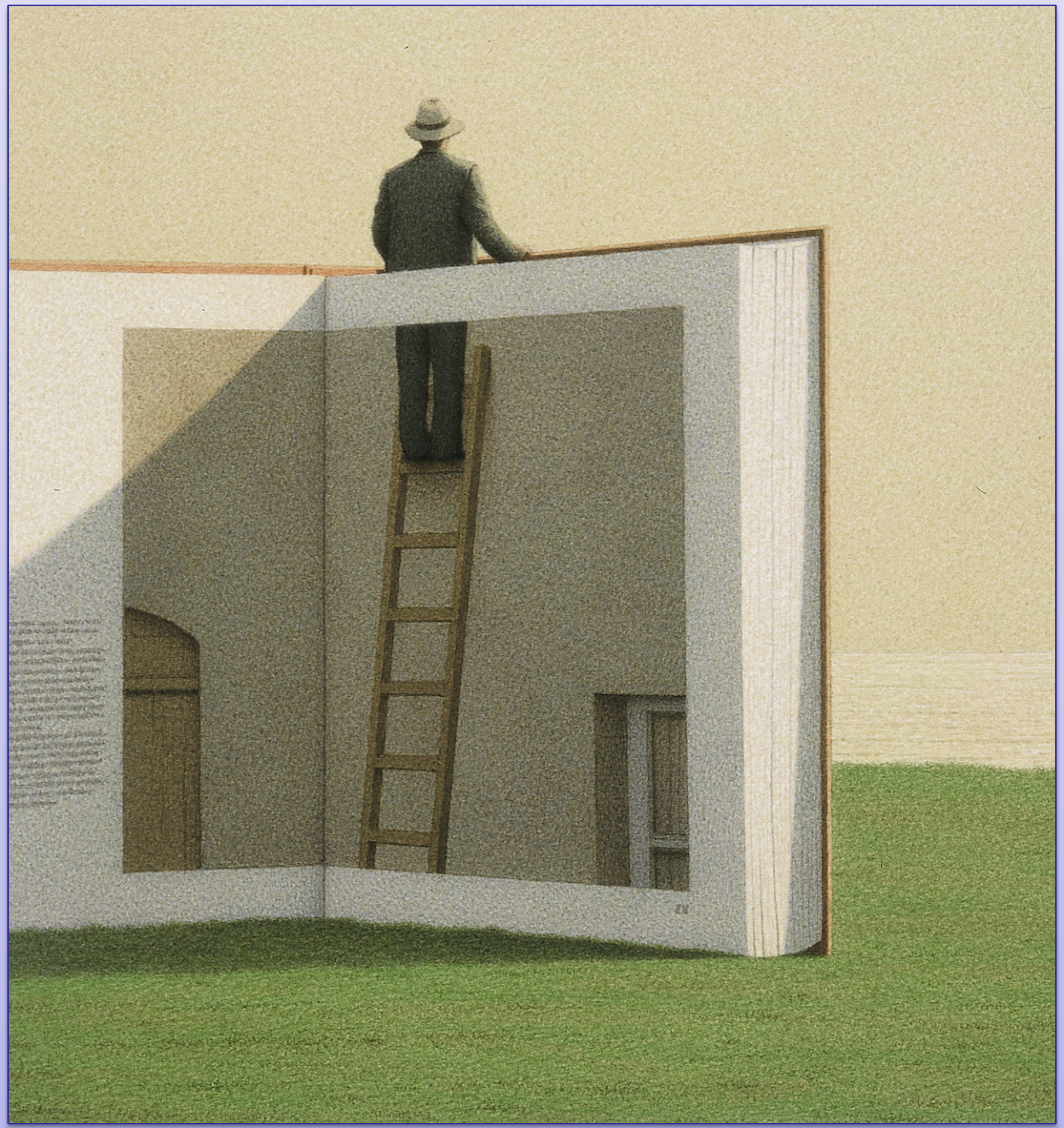
VINCE: Gordon et al. (2020)
MoCo: He et al. (2019)

Why Self-Supervision?

1. To get away from semantic categories



2. To get away from fixed datasets

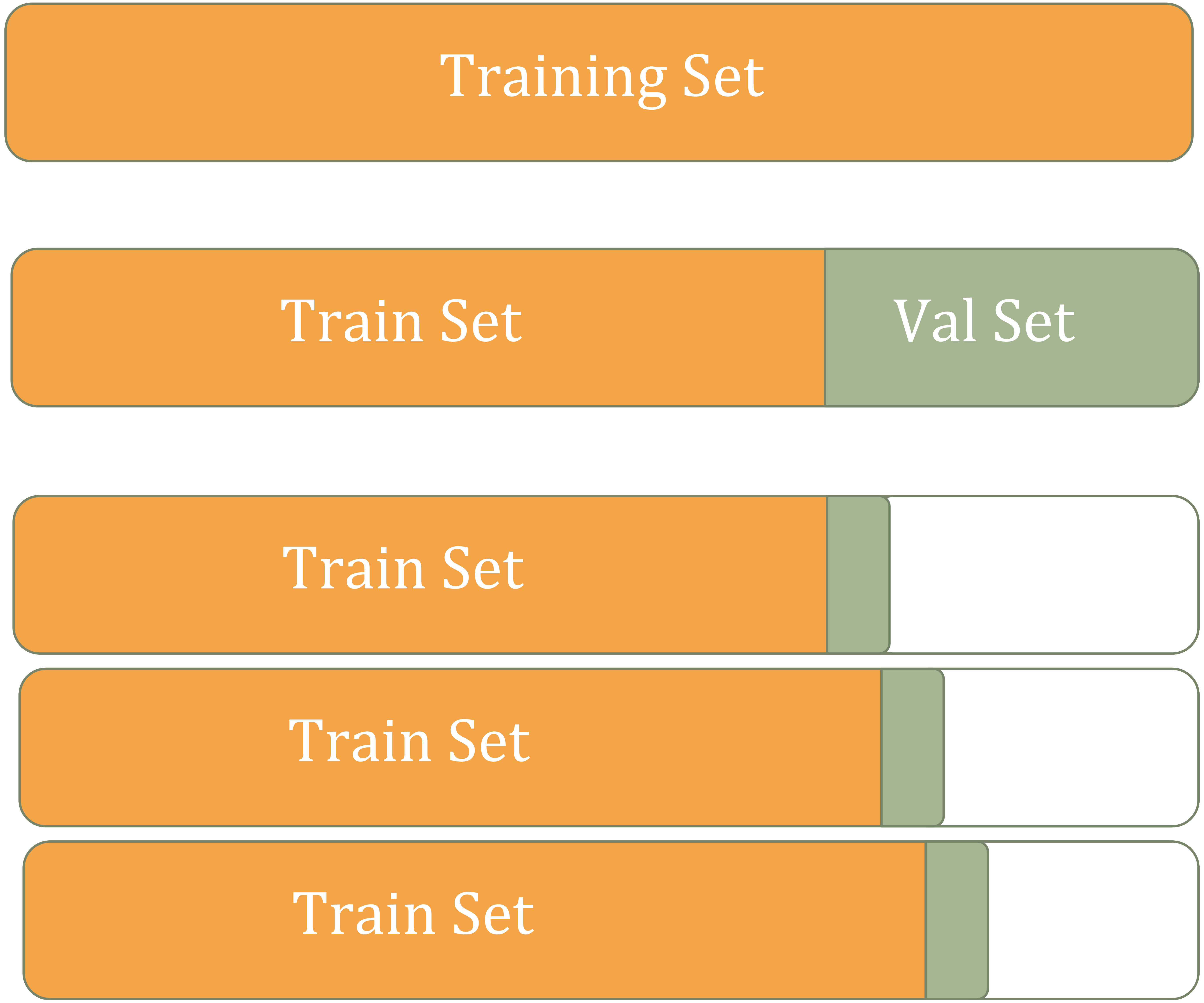


3. To get away from fixed objectives

What’s wrong with Fixed Datasets?

- 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
  - Might explain why data augmentation works so well
- The only reason for fixed datasets is annotation expanse
- **But with self-supervision, there is no excuse for reusing data / multiple epochs**

Online Continual Learning



Train / val split

Continual learning

as continual Validation/testing

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)

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

For a test distribution Q

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

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)

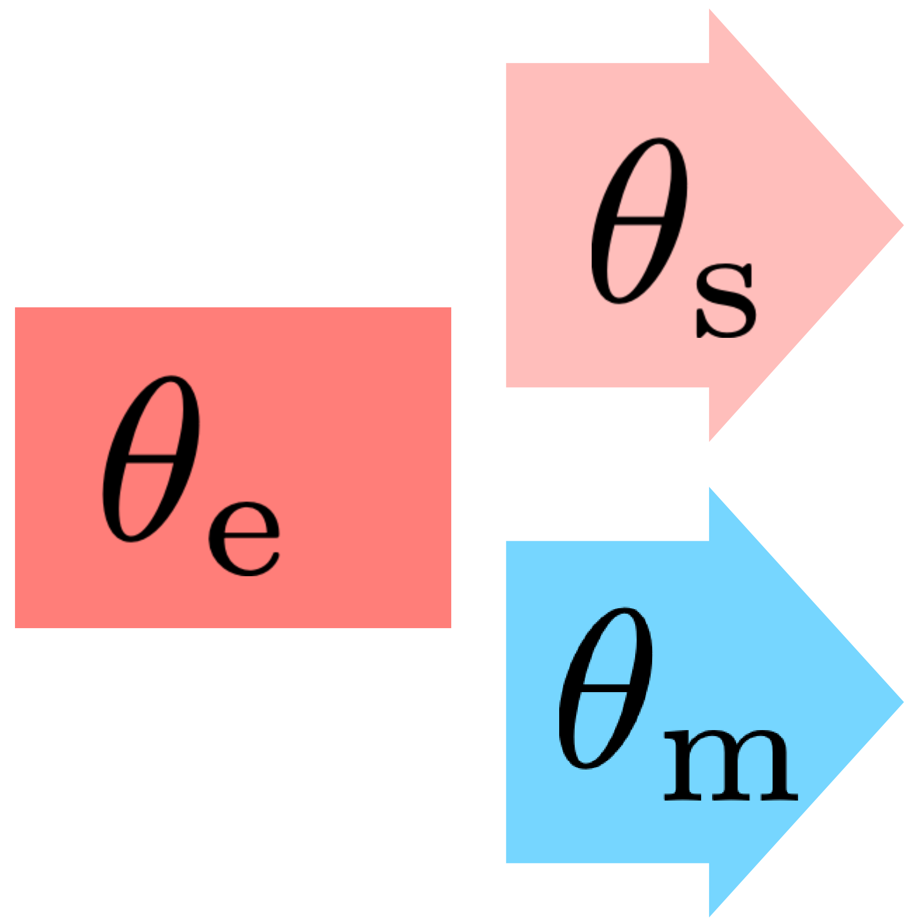
$$\text{standard test error} = \mathbb{E}_Q[\ell(x, y);\ \theta]$$

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

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!**

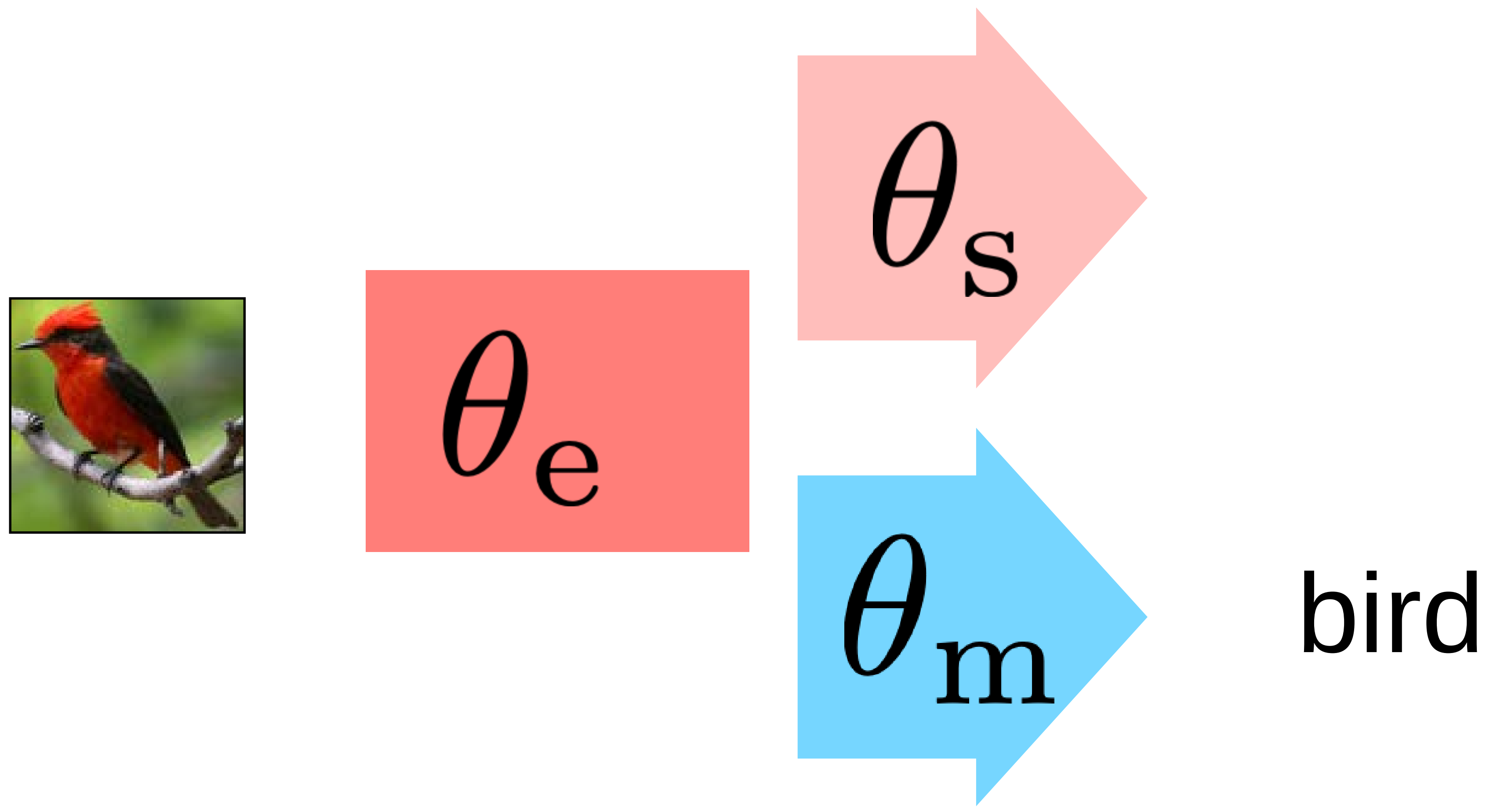
Algorithm for TTT



network
architecture

Algorithm for TTT

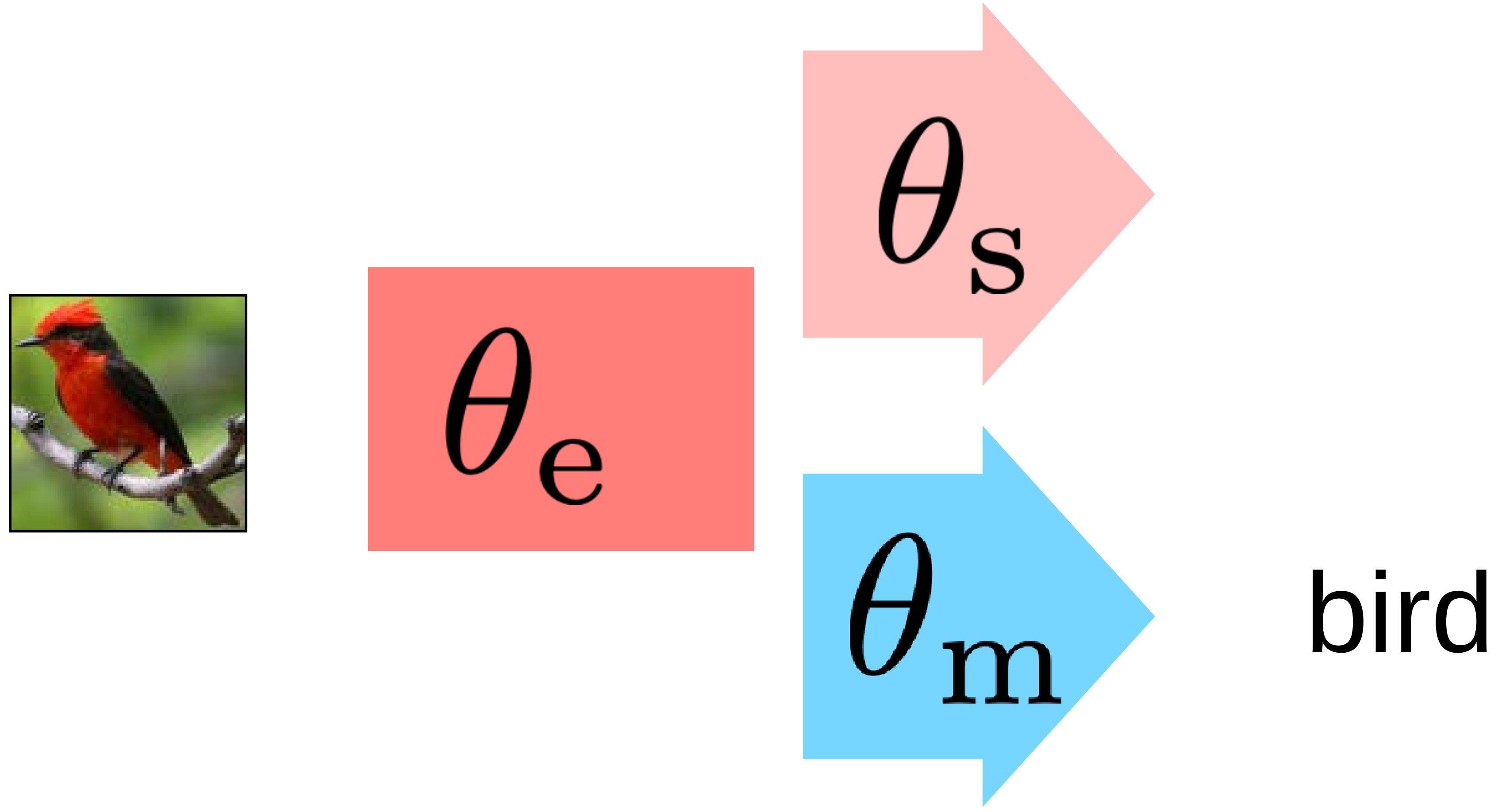
training



Algorithm for TTT

training

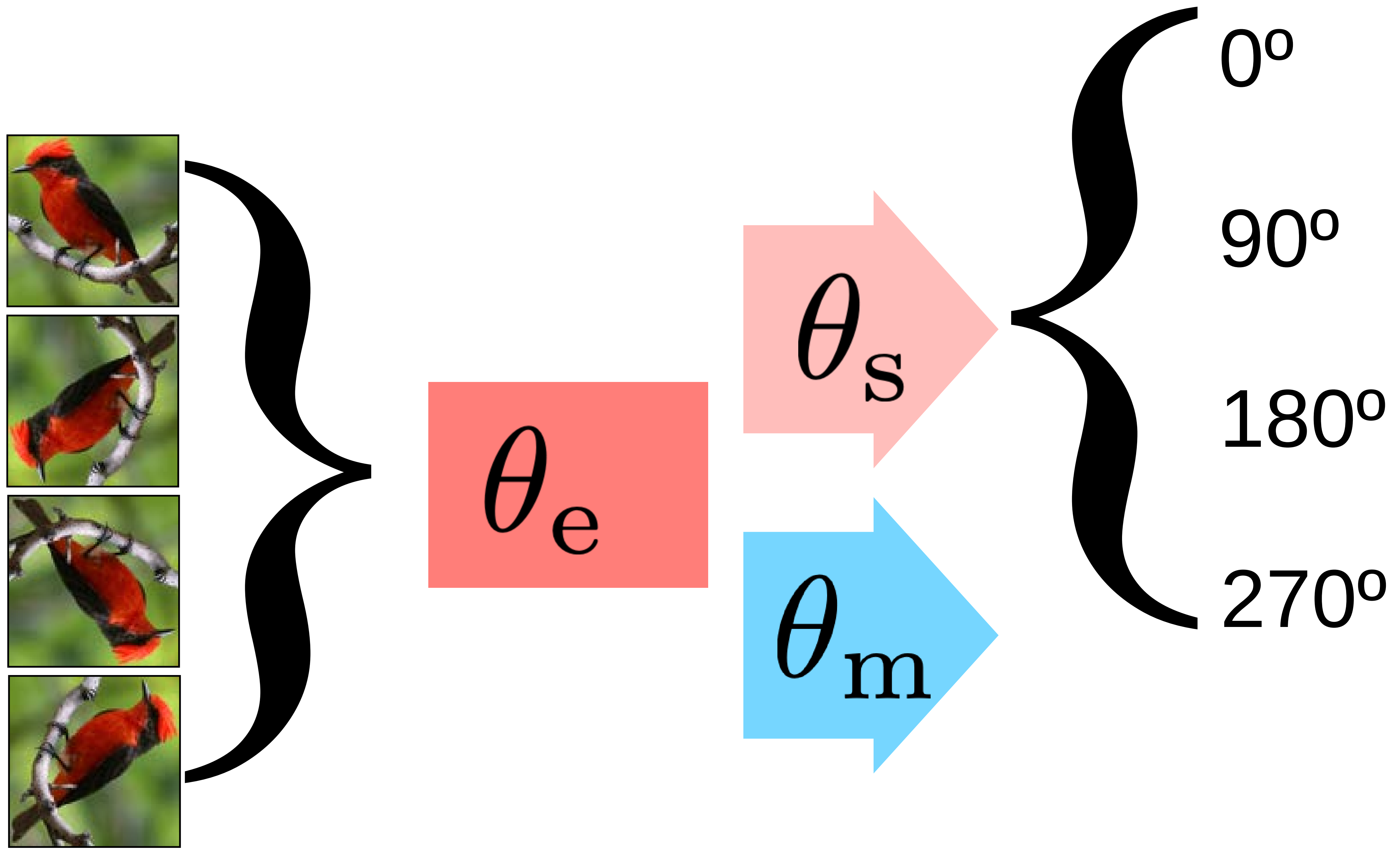
$\ell_{\mathrm{m}}(x, y; \theta_{\mathrm{e}}, \theta_{\mathrm{m}})$



Algorithm for TTT

training

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

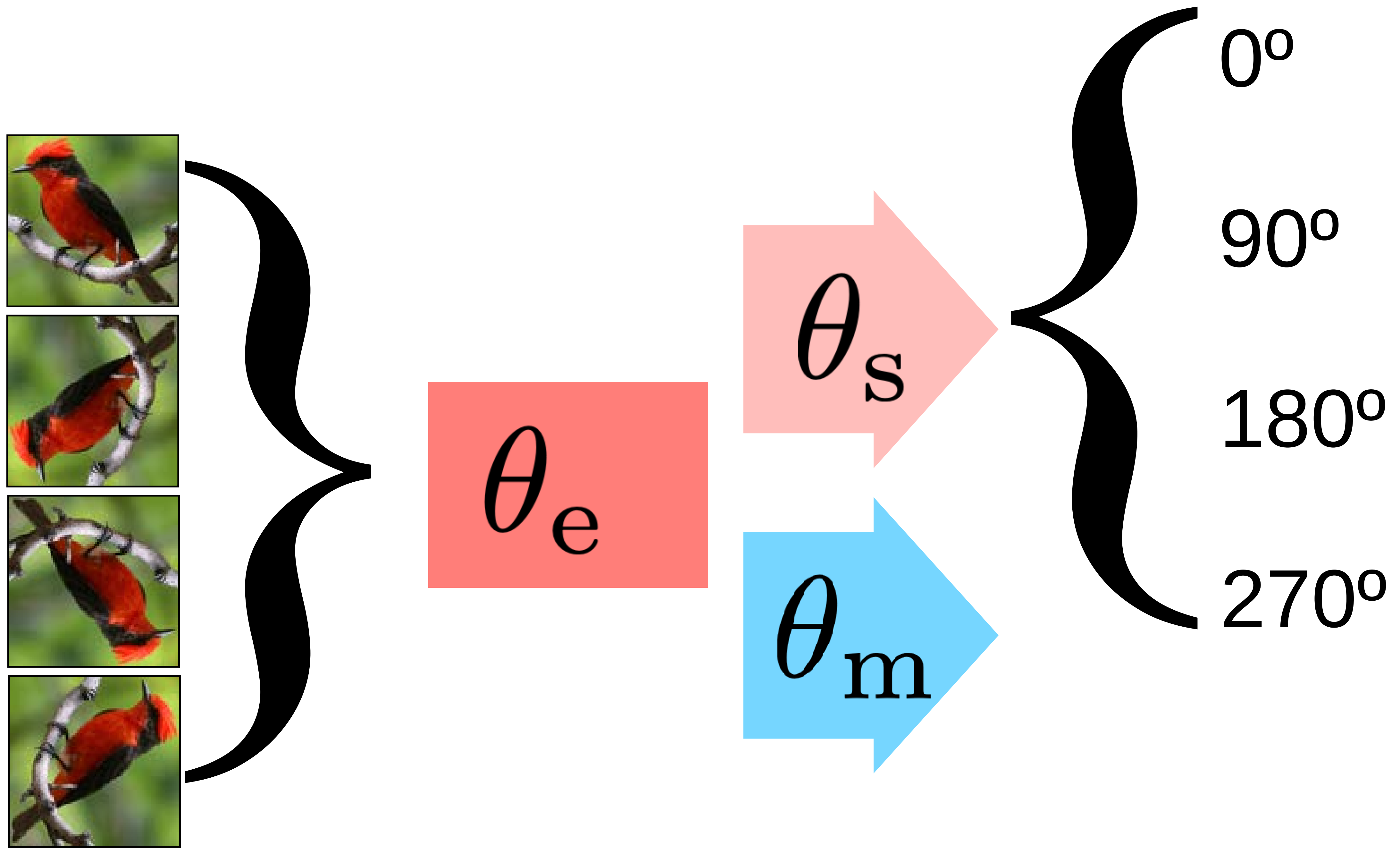


Algorithm for TTT

training

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

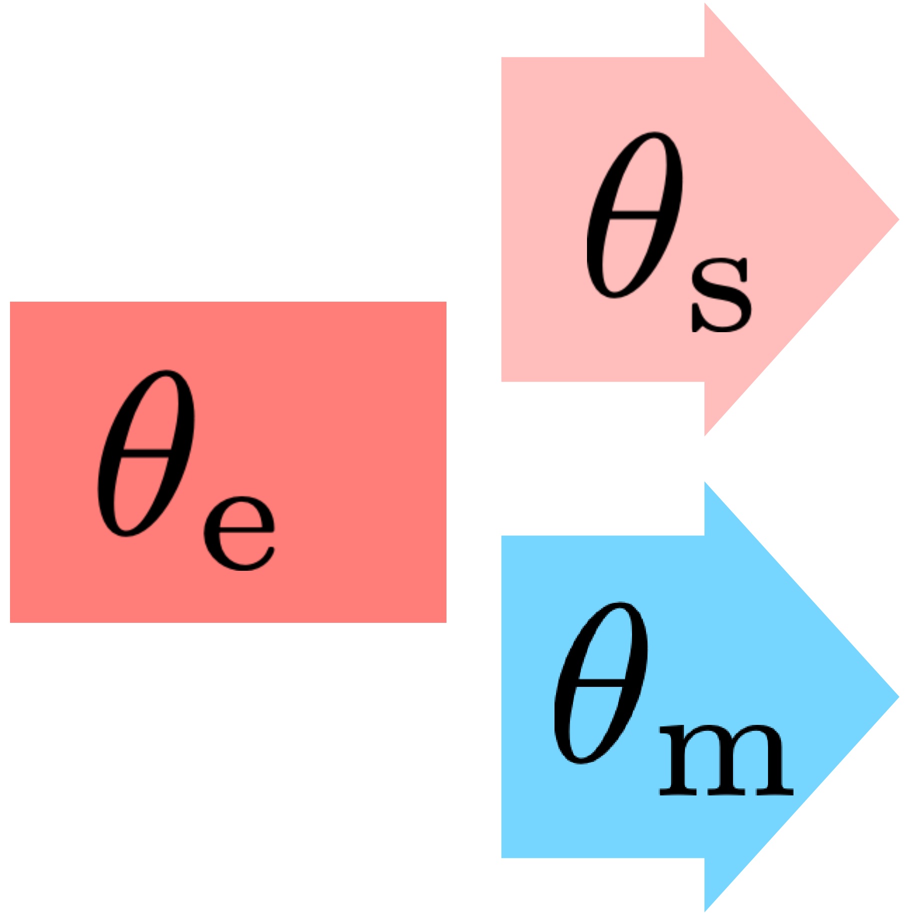
$+\, \ell_s(x, y_{\mathrm{s}}; \theta_e, \theta_s)$



Algorithm for TTT

training

$$\min_{\theta_\mathrm{e},\theta_\mathrm{s},\theta_\mathrm{m}} \mathbb{E}_P \begin{bmatrix} \ell_\mathrm{m}(x, y; \theta_\mathrm{e}, \theta_\mathrm{m}) \\ + \, \ell_s(x, y_\mathrm{s}; \theta_e, \theta_s) \end{bmatrix}$$

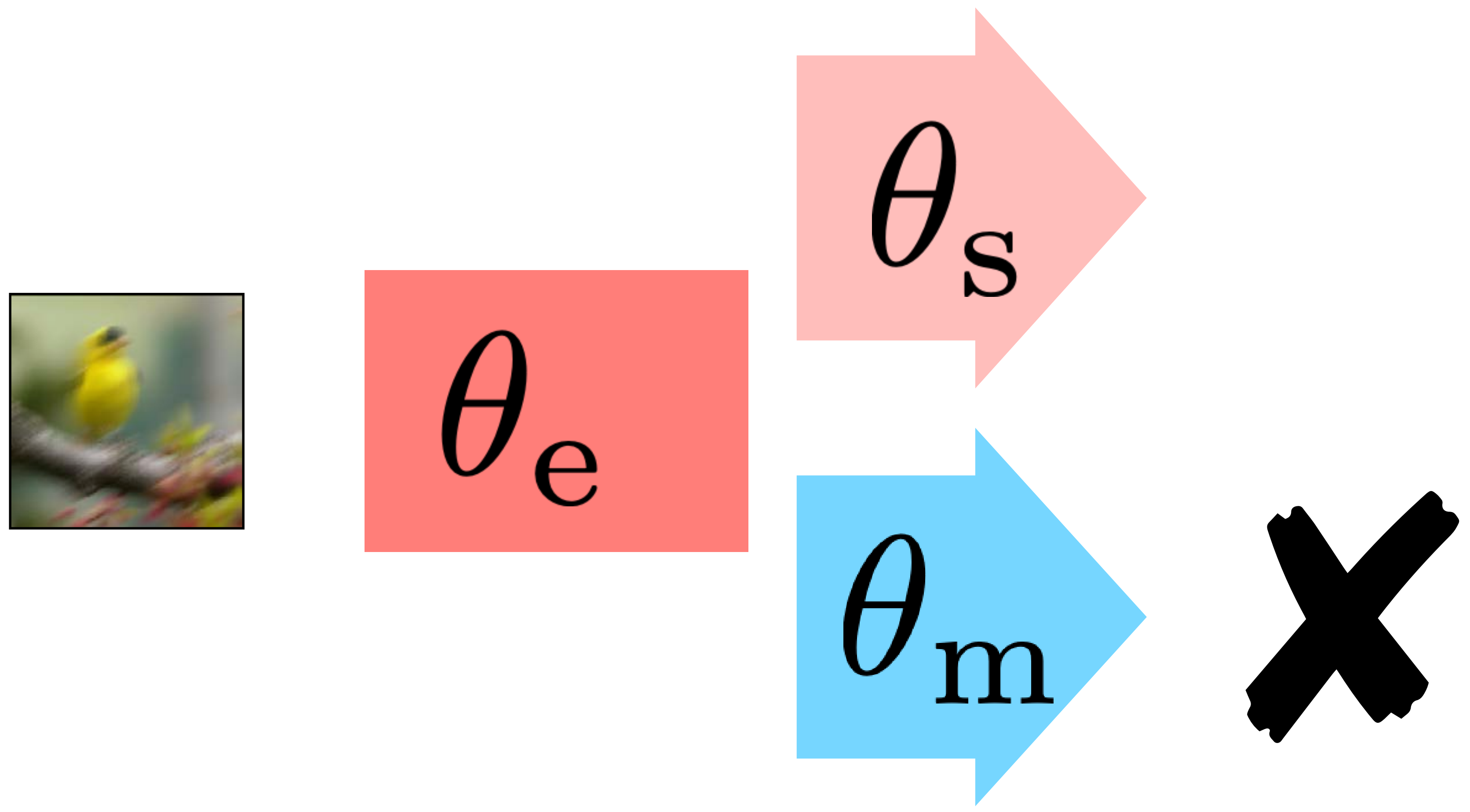


Algorithm for TTT

training

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

testing

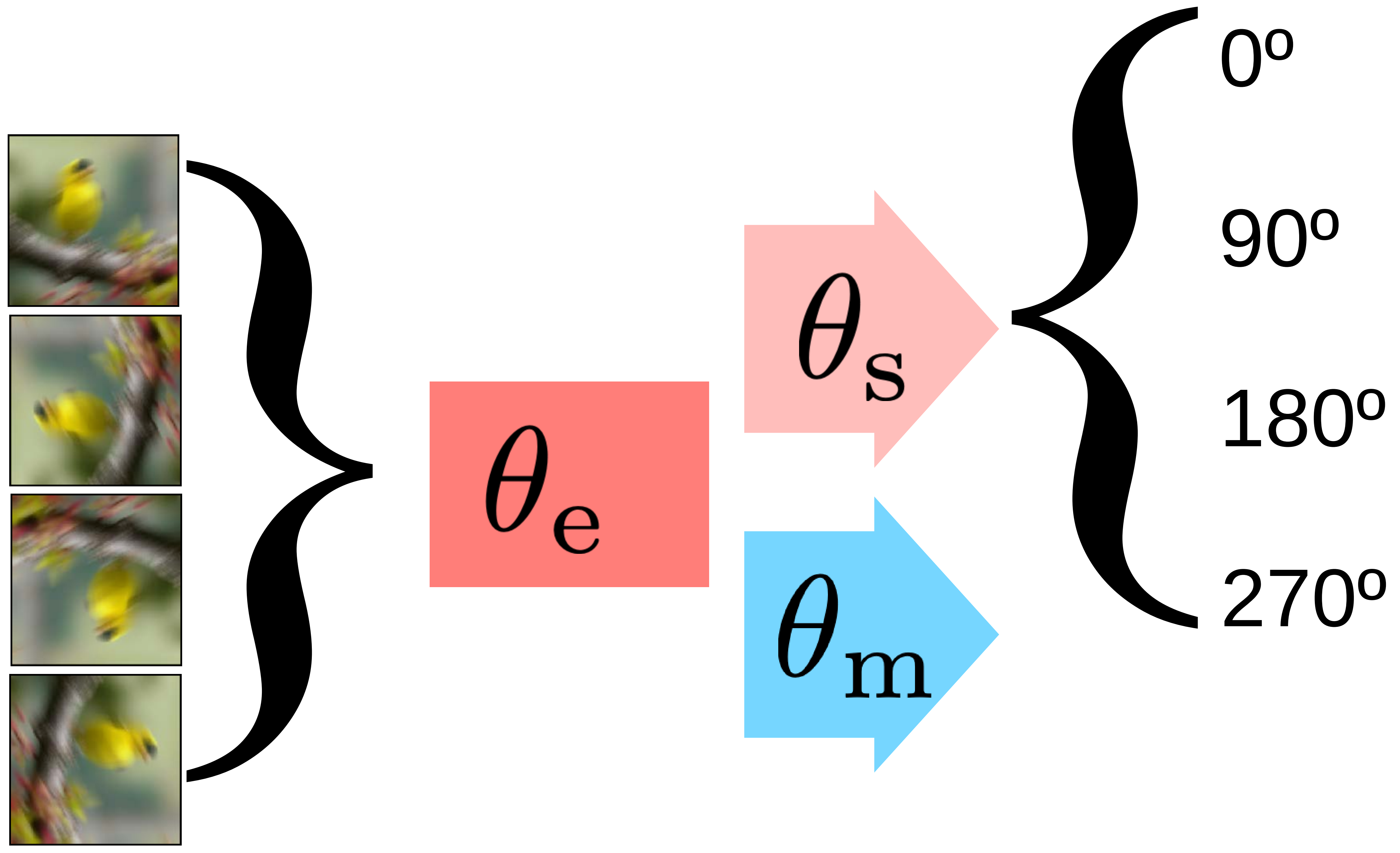


Algorithm for TTT

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



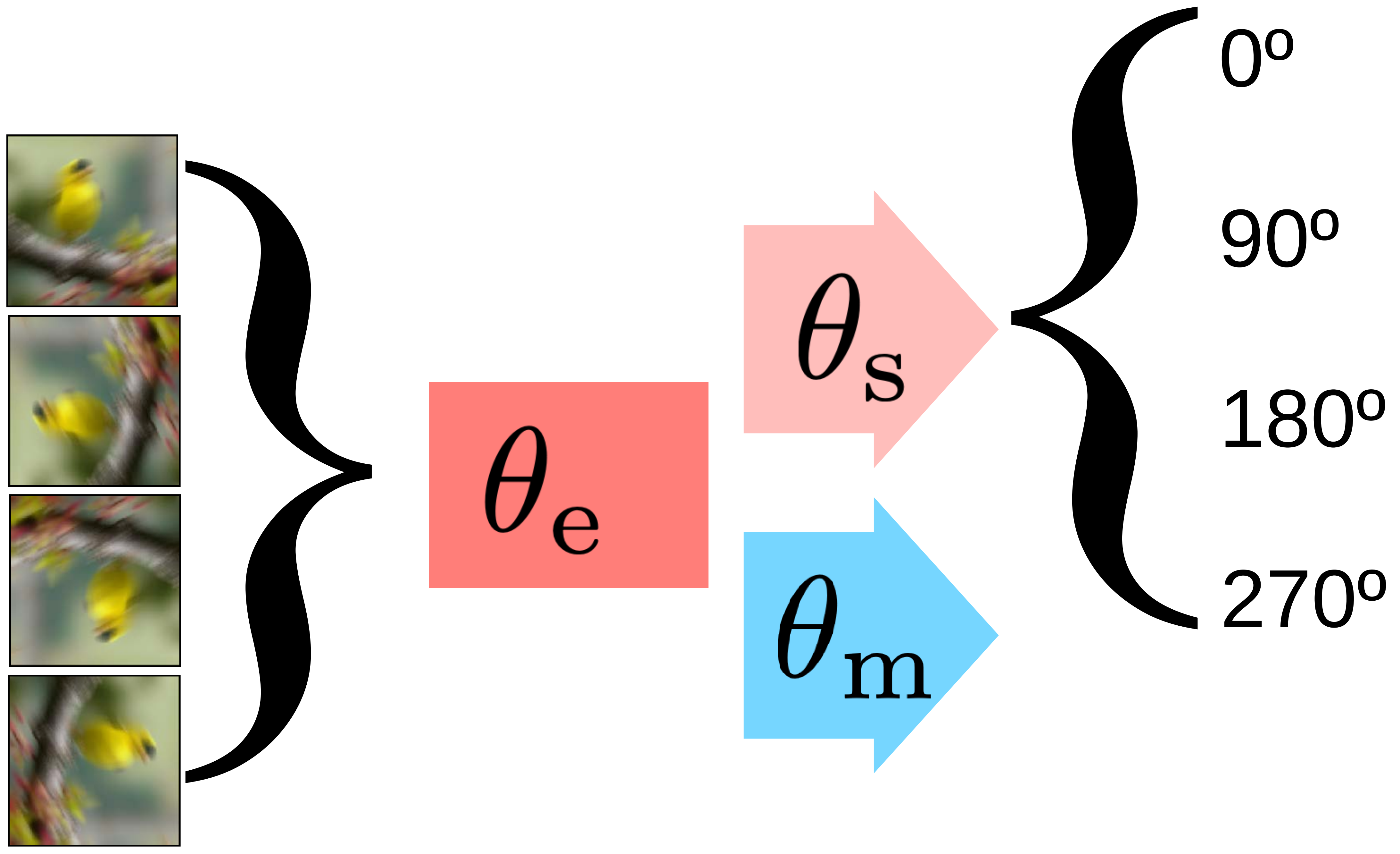
Algorithm for TTT

training

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

testing

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



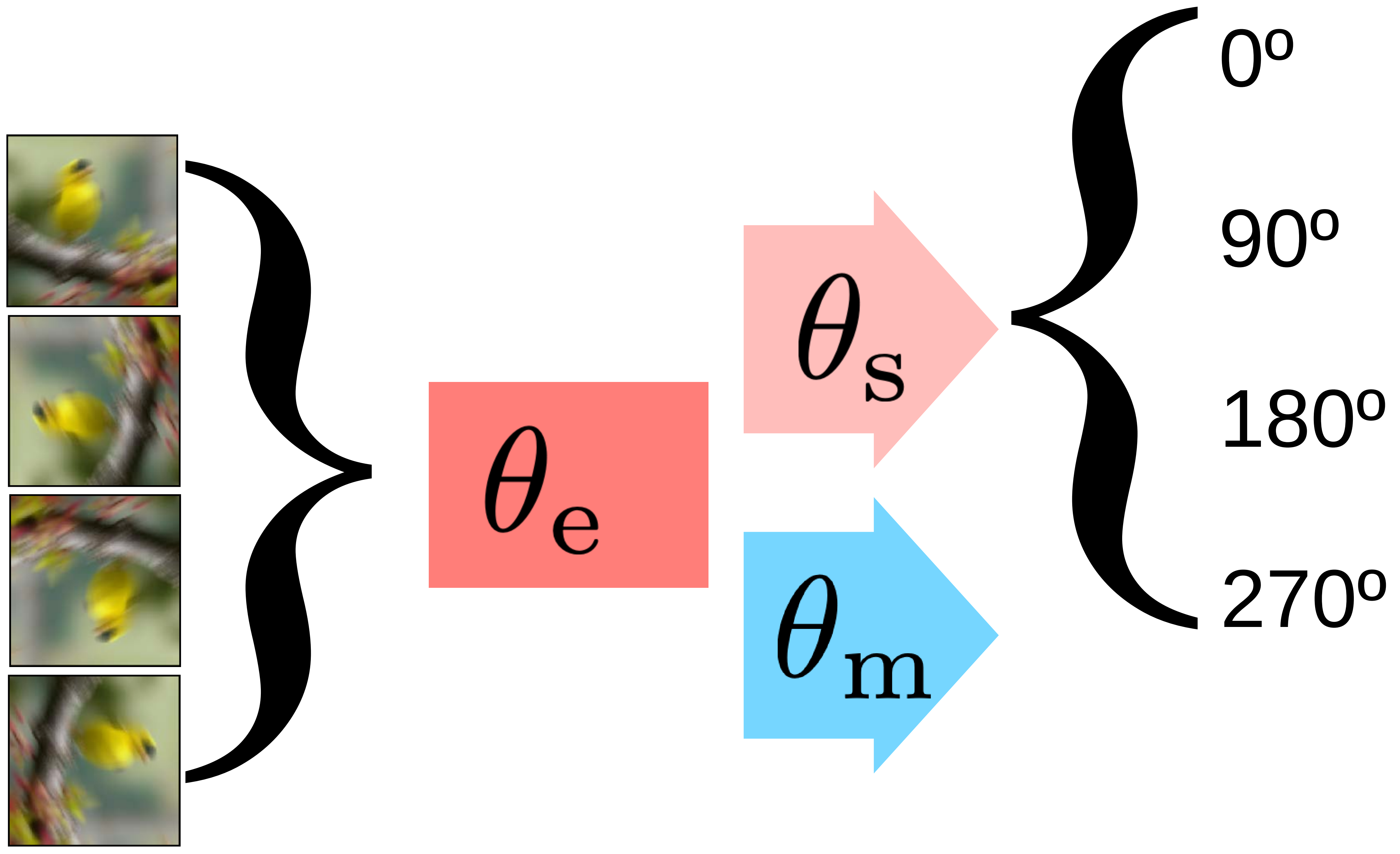
Algorithm for TTT

training

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

testing

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



Algorithm for TTT

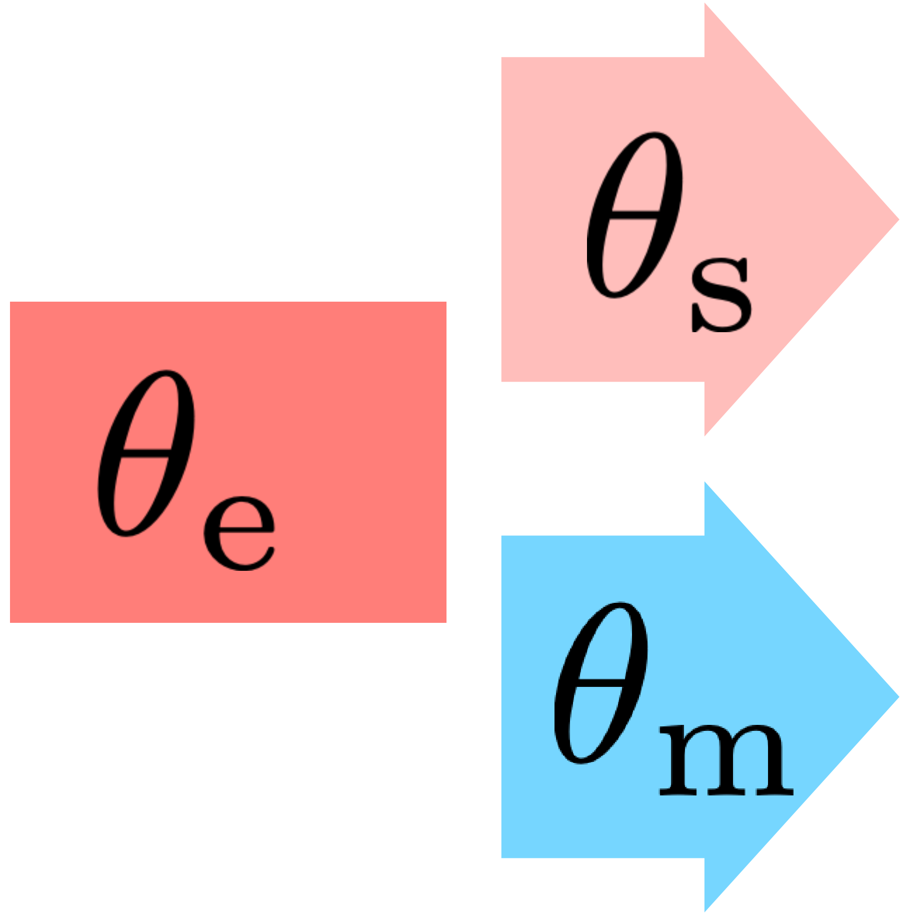
training

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

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

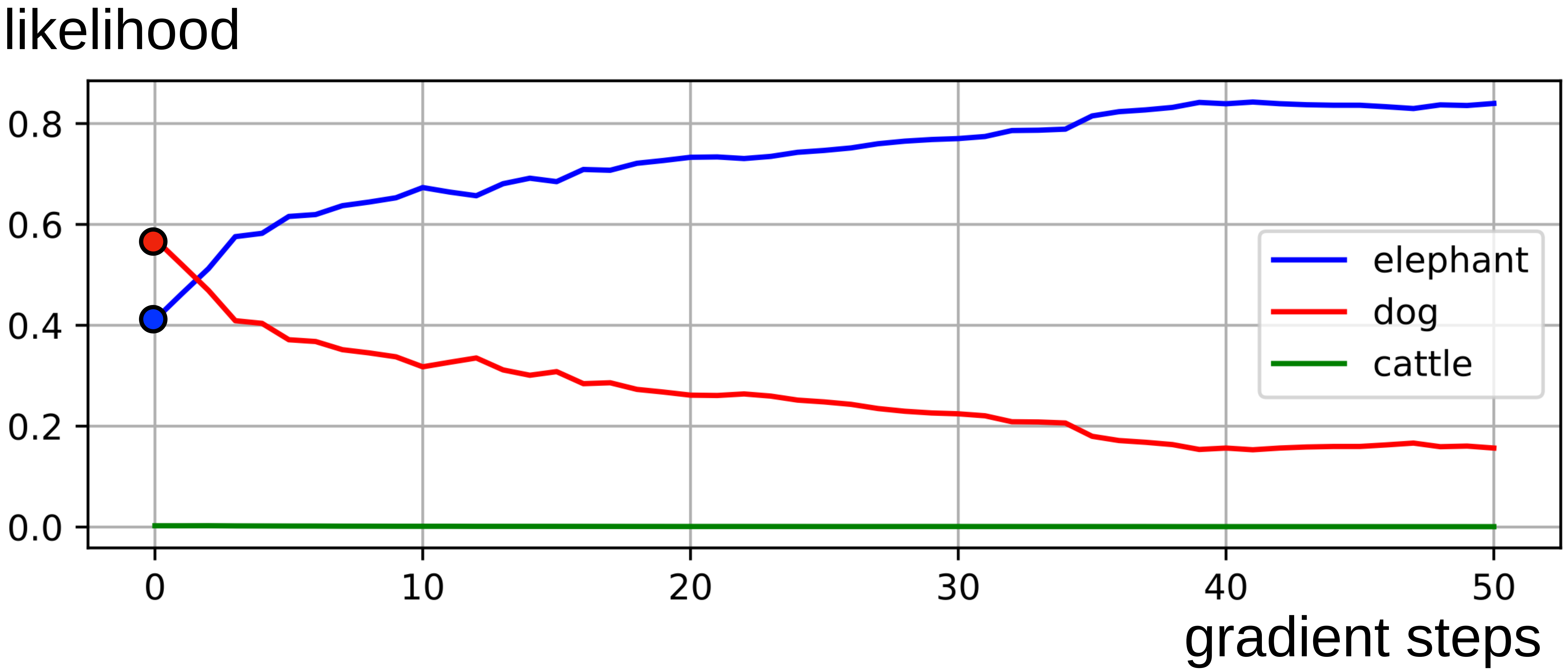
training

$$\min_{\theta_\mathrm{e},\theta_\mathrm{s},\theta_\mathrm{m}} \mathbb{E}_P \begin{bmatrix} \ell_\mathrm{m}(x, y; \theta_\mathrm{e}, \theta_\mathrm{m}) \\ + \, \ell_s(x, y_\mathrm{s}; \theta_e, \theta_s) \end{bmatrix}$$

testing

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

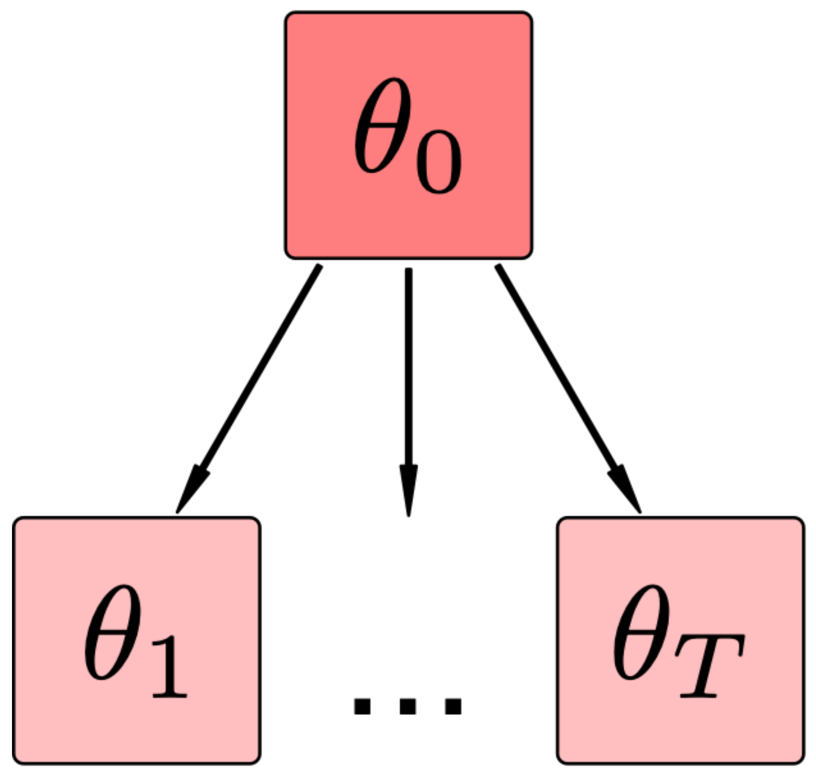
$\rightarrow \theta(x)$: make prediction on x



Algorithm for TTT

multiple test samples $x_1, \ldots, x_T$

θ_0: parameters after joint training



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

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

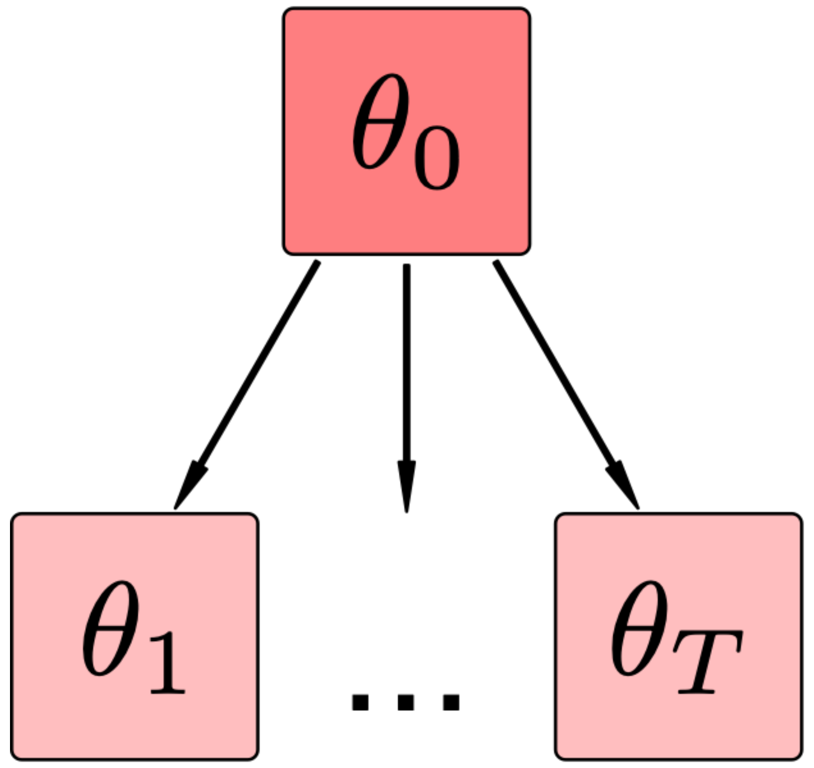
$\rightarrow \theta(x)$: make prediction on x

multiple test samples $x_1, \dots, x_T$

θ_0: parameters after joint training

standard version

no assumption on the test samples



Algorithm for TTT

training

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

testing

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

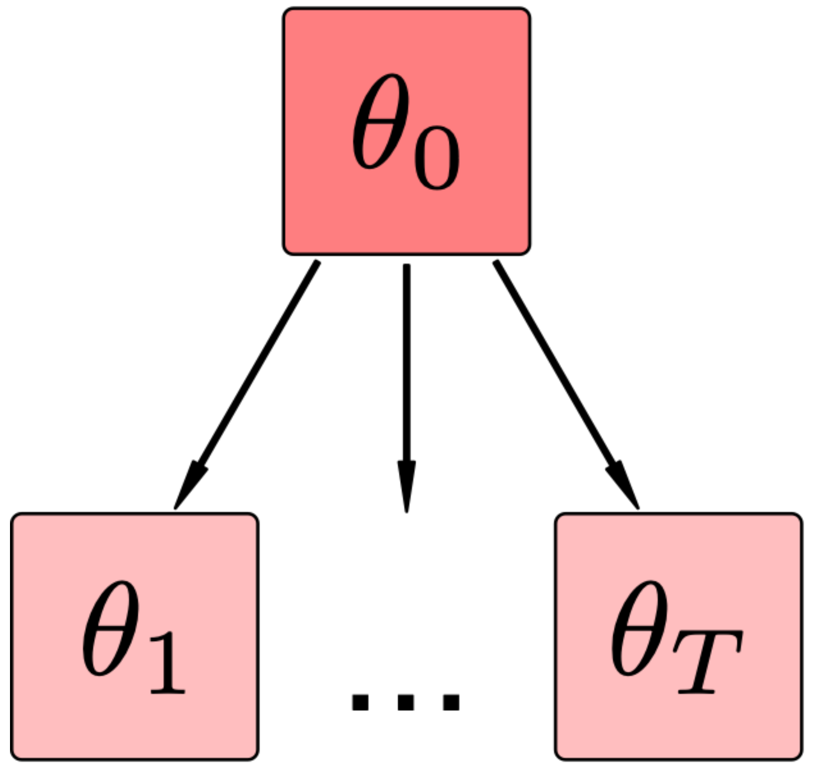
$\rightarrow \theta(x)$: make prediction on x

multiple test samples $x_1, \dots, x_T$

θ_0: parameters after joint training

standard version

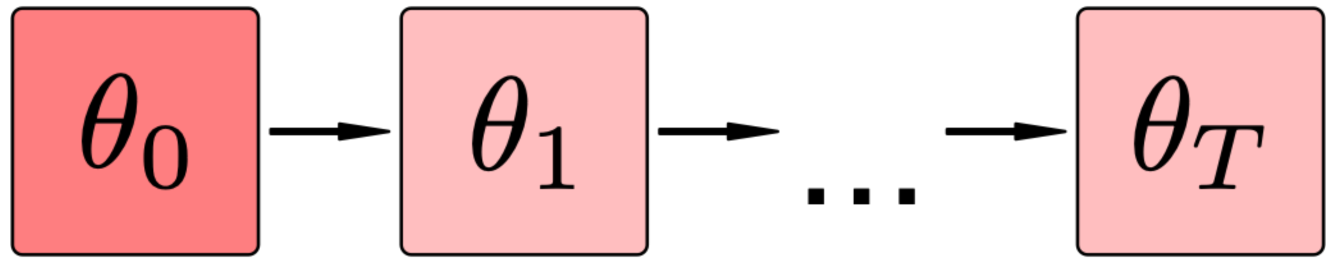
no assumption on the test samples



online version

$x_1, \dots, x_T$ come from the same Q

or <u>smoothly</u> changing $Q_1, \dots, Q_T$



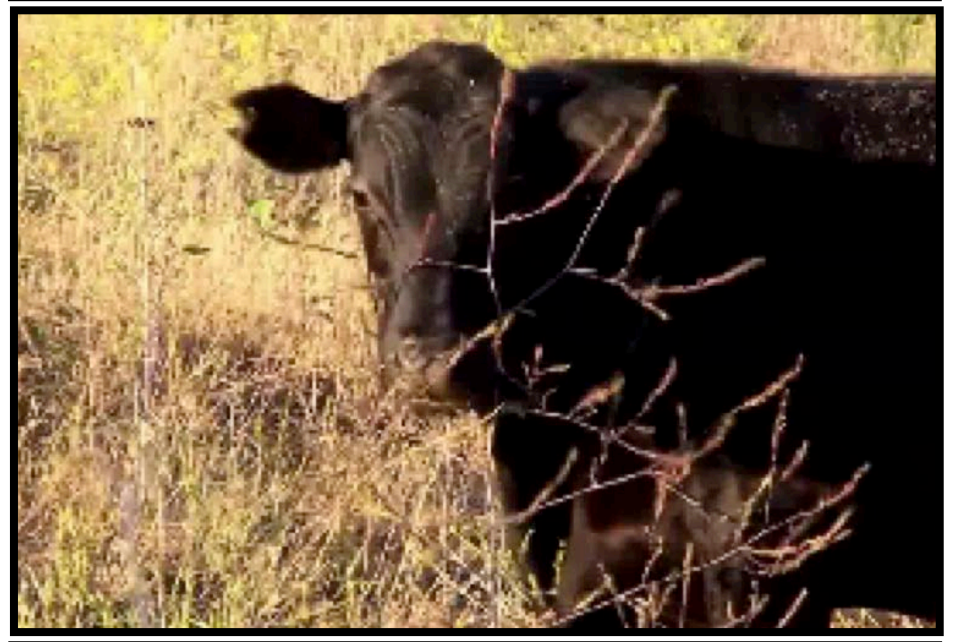
Results

Method	CIFAR-10 accuracy (%)	ImageNet accuracy (%)
Object recognition task only	41.4	62.7
Joint training (Hendrycks et al. 2019)	42.4	63.5
TTT standard	45.2	63.8
TTT online	45.4	64.3

examples

Join training: dog

TTT: elephant



Join training: dog

TTT: cattle

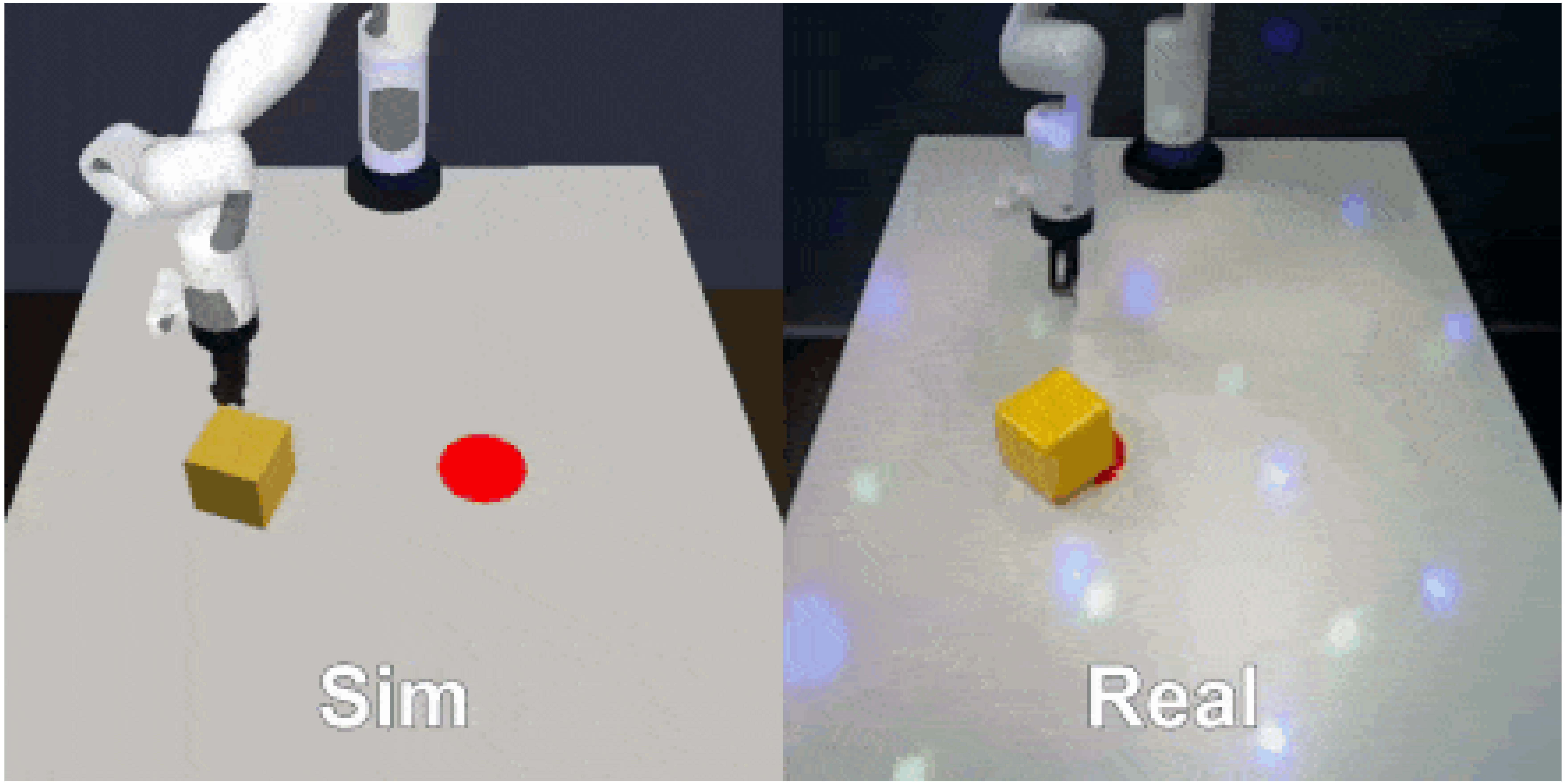
Join training: car

TTT: bus

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

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

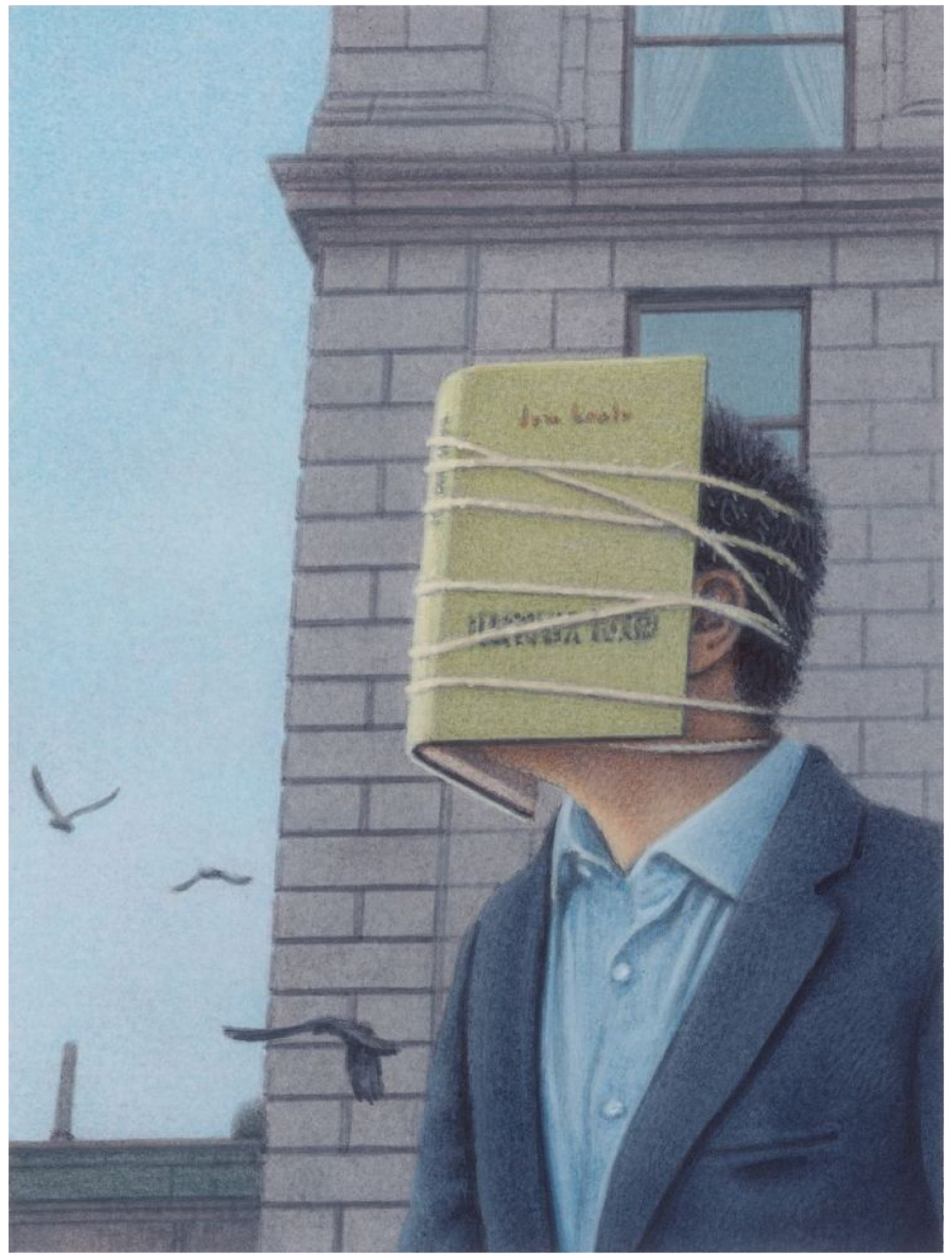
Self-Supervised Policy Adaptation during Deployment [ICLR’21]



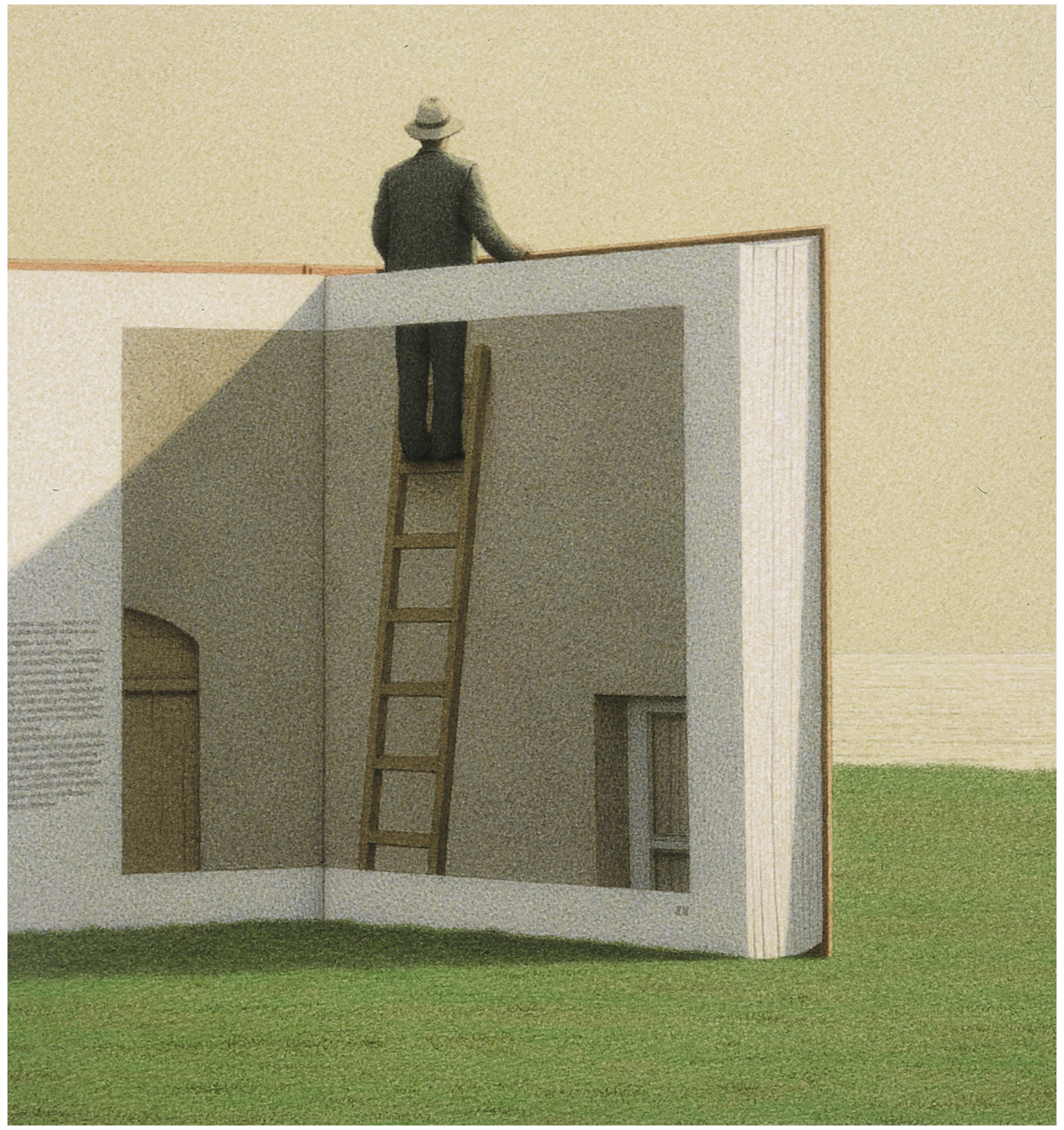
[Nicklas Hansen, Yu Sun, Pieter Abbeel, Alexei A. Efros, Lerrel Pinto, Xiaolong Wang, ICLR 2021]

Why Self-Supervision?

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



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



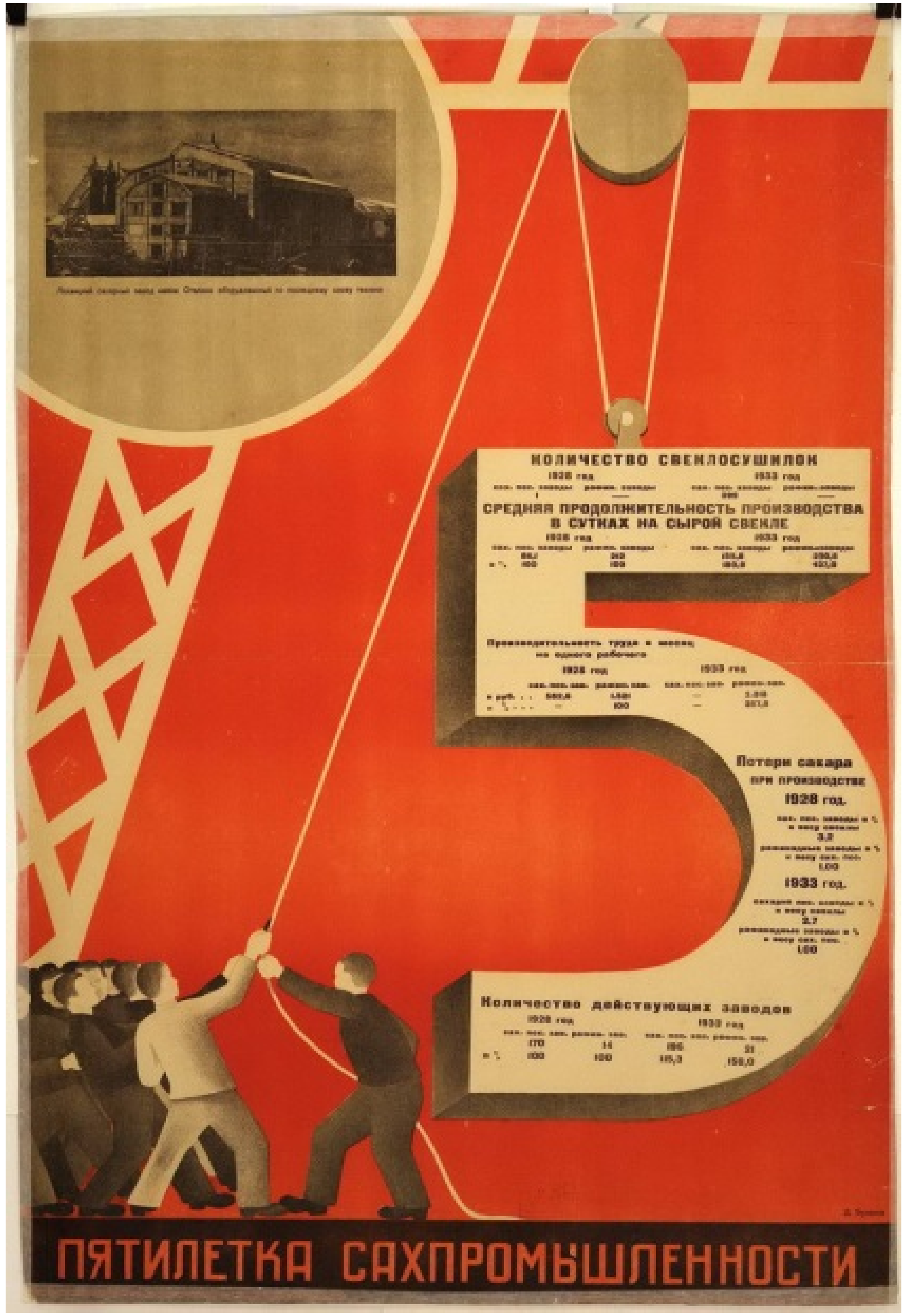
3. To get away from <u>fixed objectives</u>

What’s wrong with fixed objectives?

- Genetic Algorithms
  - Just another way to optimize some objective function
- The magic of evolution:
  - It doesn’t optimize any objective
  - Objectives emerge on their own
    - Even death is an emergent property!
- How we can get emergent objectives?
  - Biologists’ and economists’ answer: **arms races**


ПАРИКМАХЕРСКАЯ

The Five-Year Plan



The Five-Year Plan

HP-80 (1973)

Elektronika DD (1973)

Arms races create emergent objectives

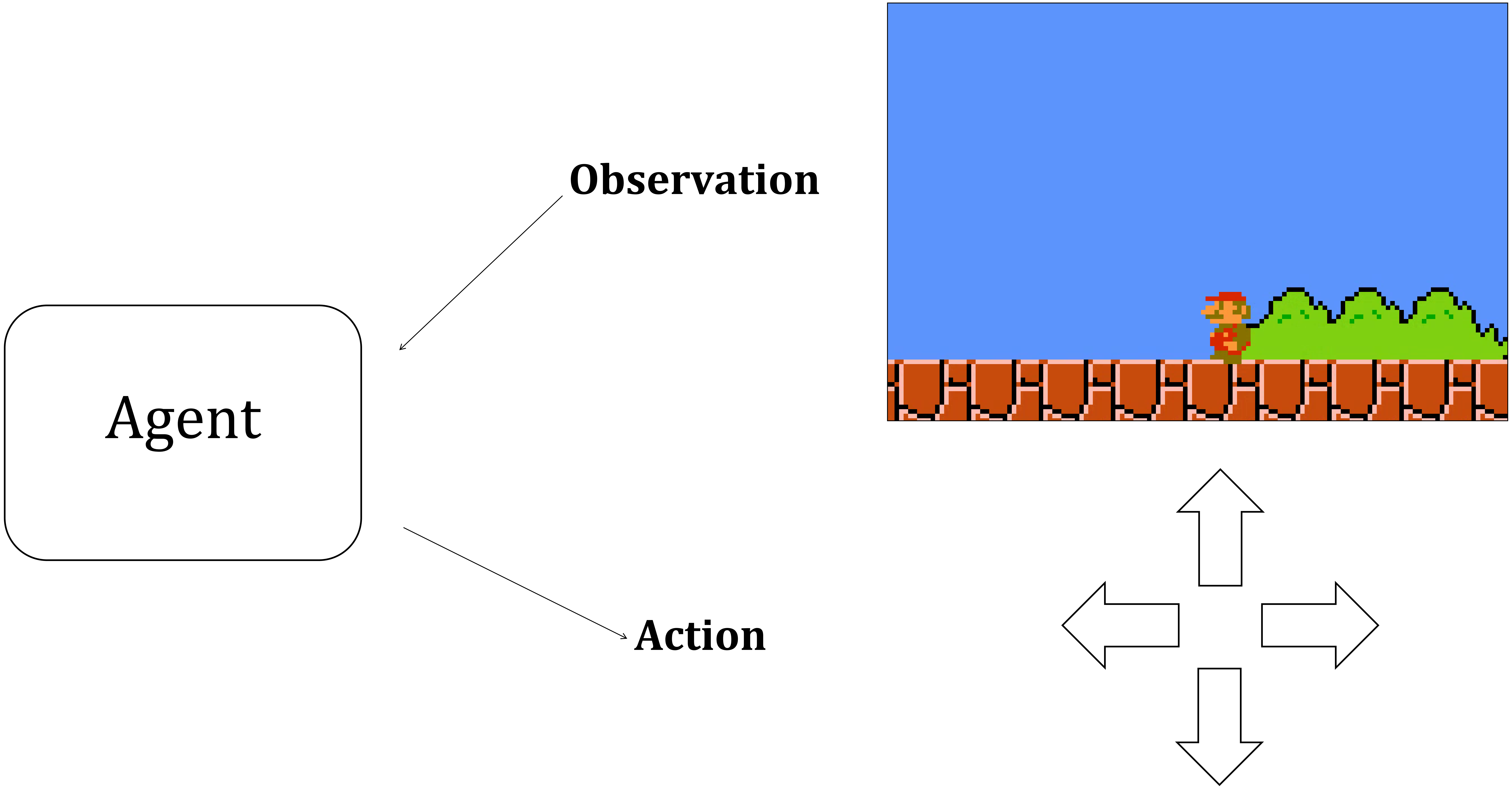
Setting up arms races

- Self-play
  – symmetric: agent vs itself
  – in practice, still have to specify final objective
- GANs
  – asymmetric: Generator vs. Discriminator
  – in practice, discriminator ends up dominating
- Prediction as meta-objective
  – The world as adversary
  – asymmetric: predicting agent vs. complexity of the world
  – open-ended: in complex world, can always predict further

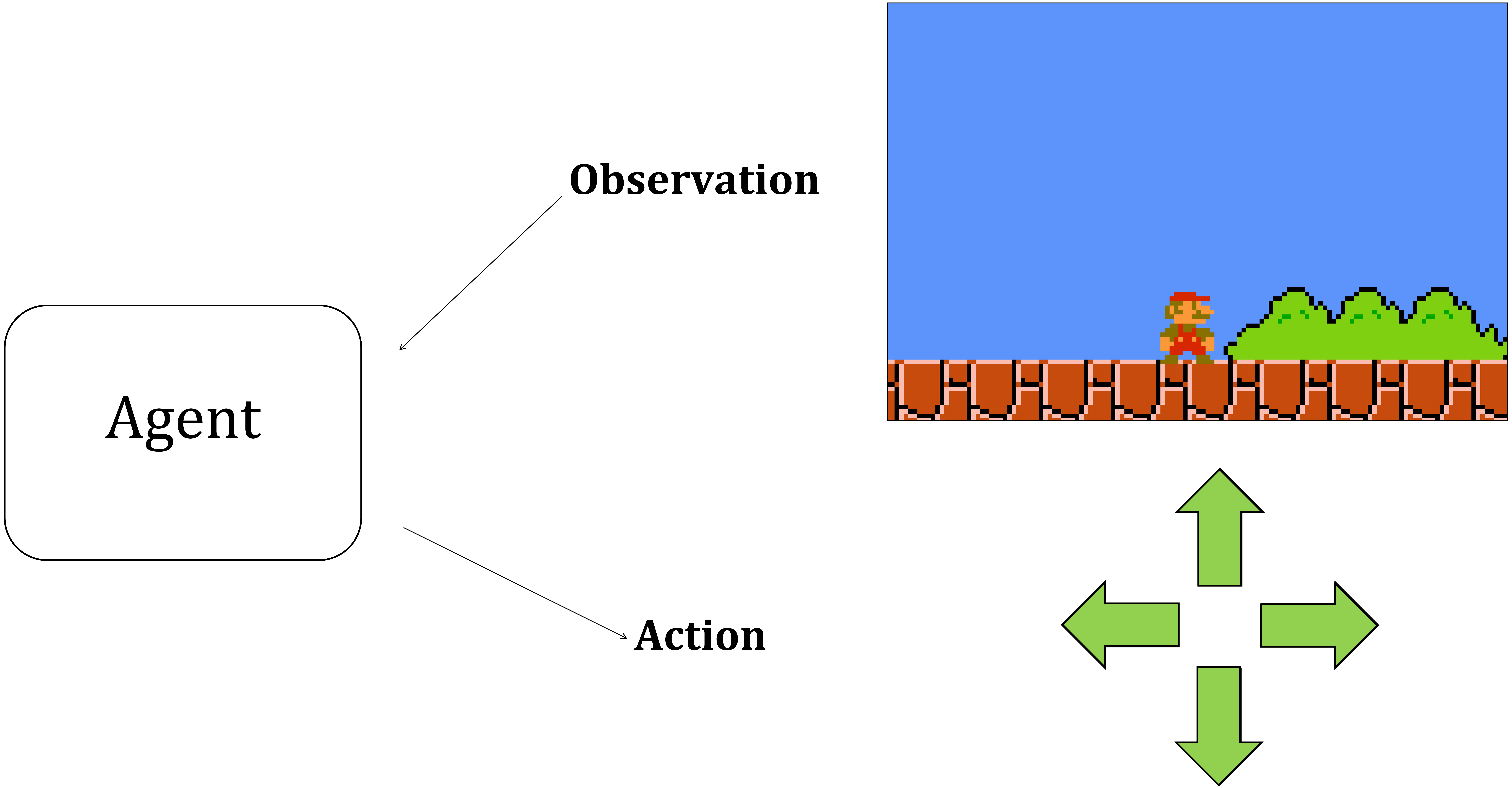
Curiosity-driven Exploration by Self-Supervised Prediction

Deepak Pathak, Pulkit Agrawal, Alexei Efros, Trevor Darrell

ICML 2017



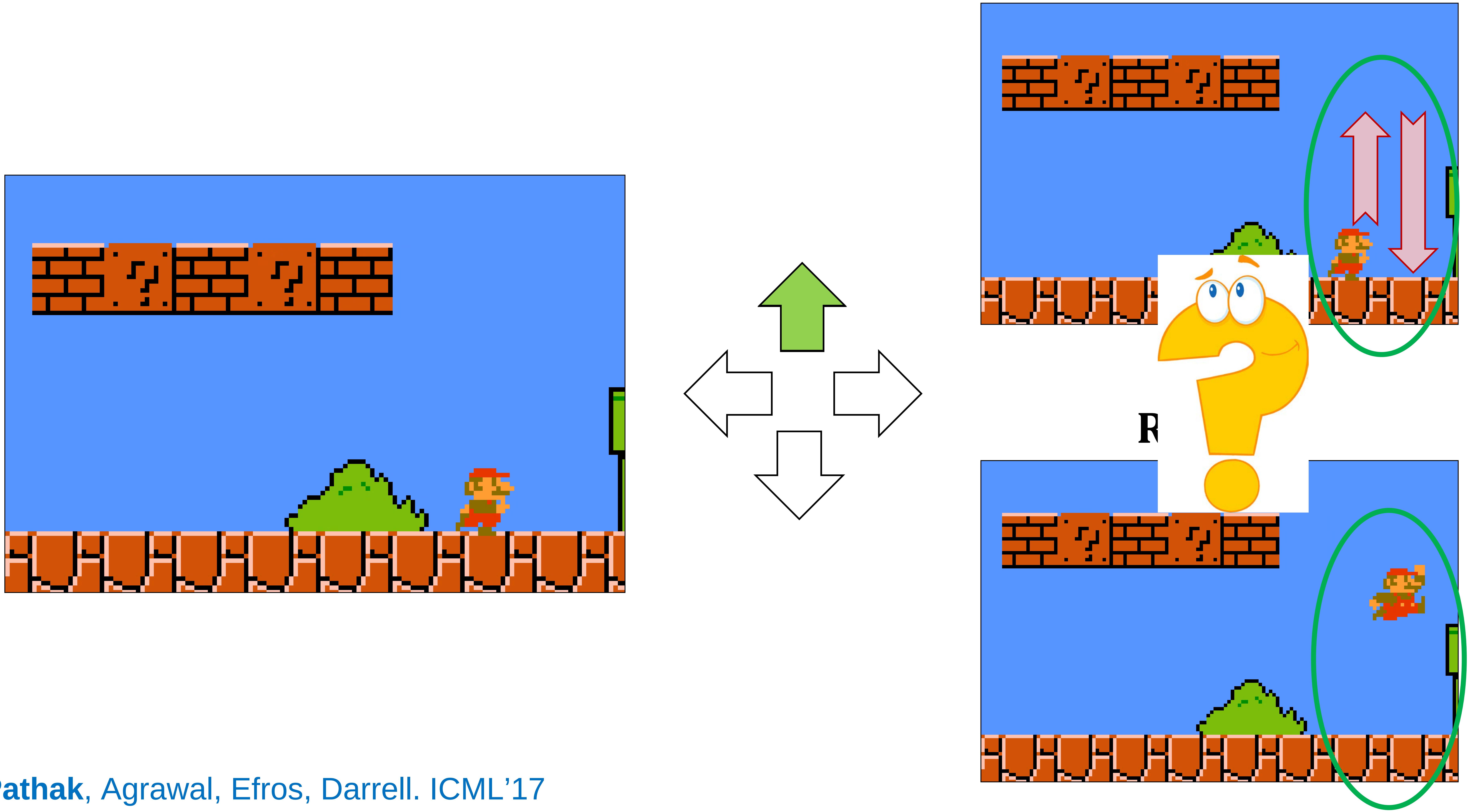
[**Pathak**, Agrawal, Efros, Darrell. ICML'17]()



“Down” has no effect

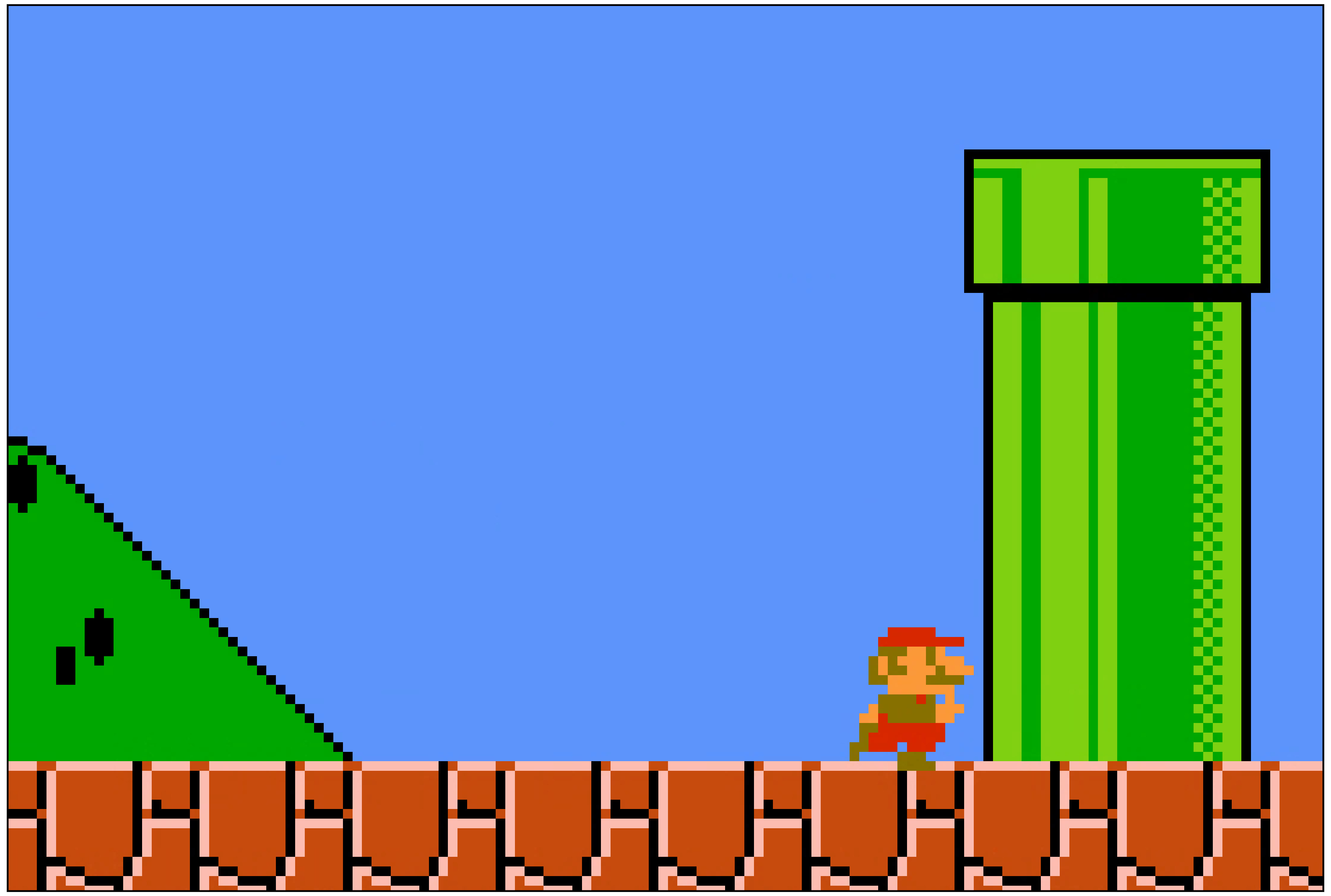
Pathak, Agrawal, Efros, Darrell. ICML’17

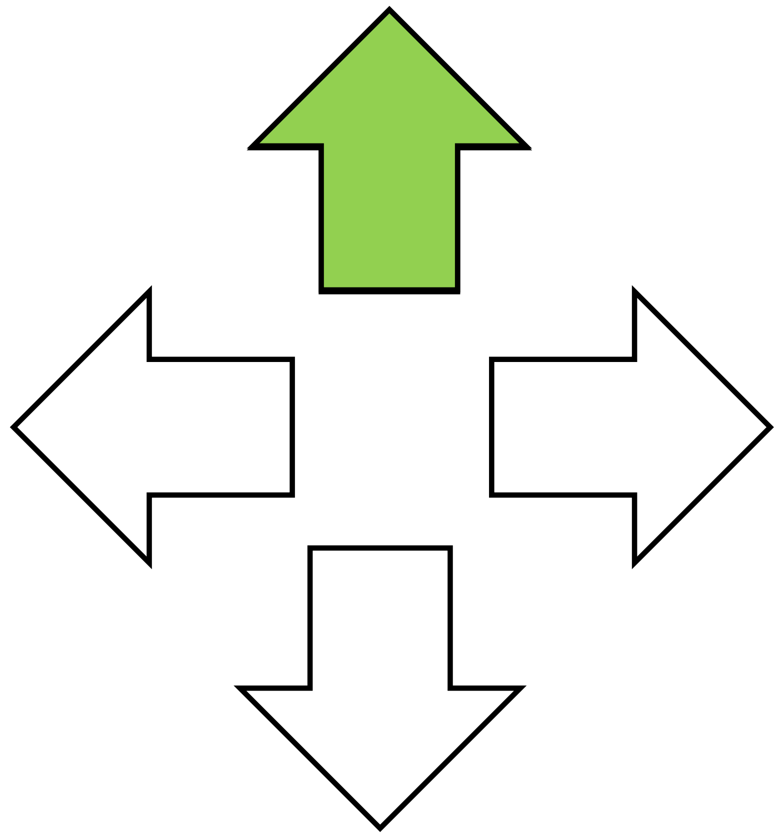
Prediction



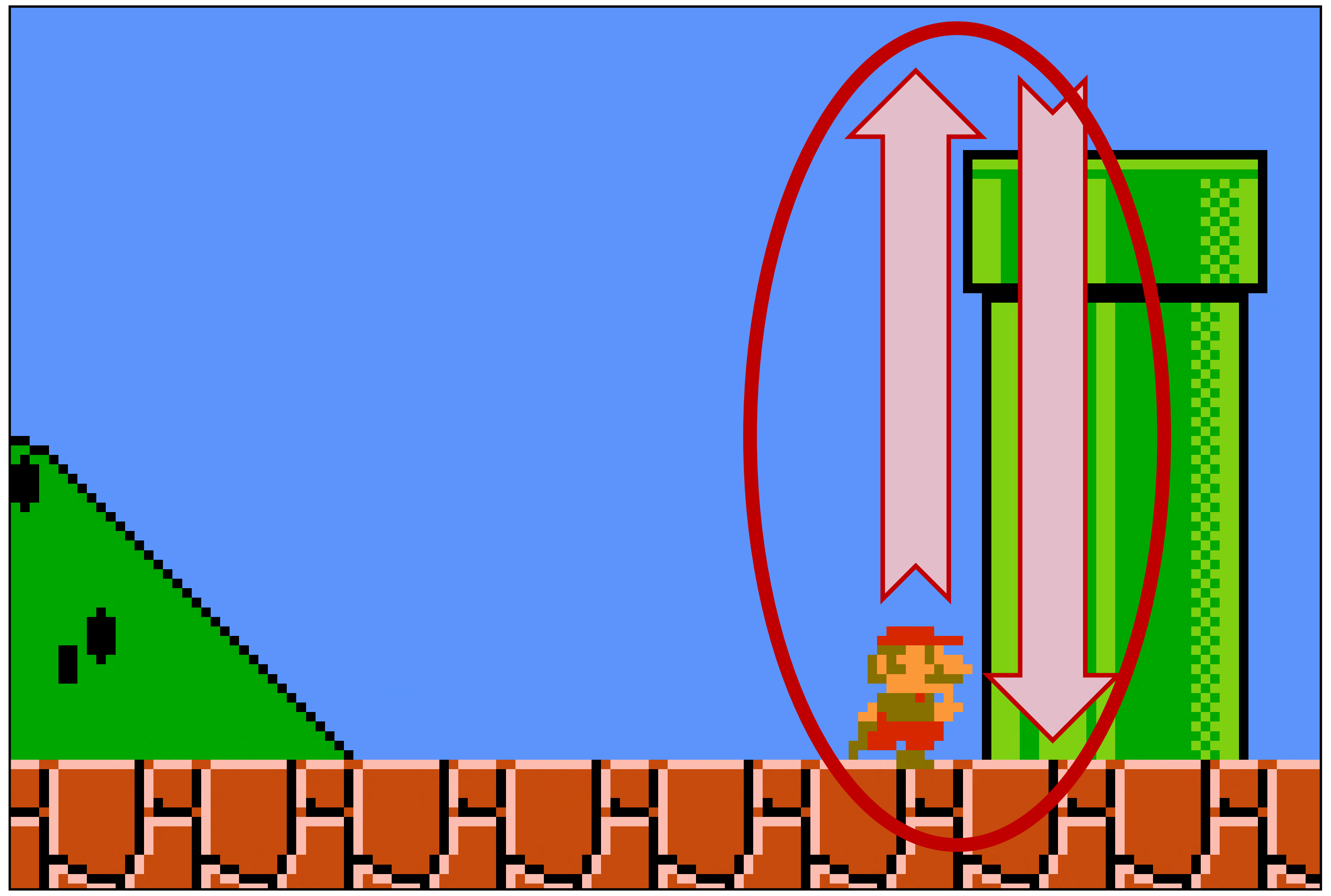
Pathak, Agrawal, Efros, Darrell. ICML’17

Curiosity $\triangleq$ Prediction Error

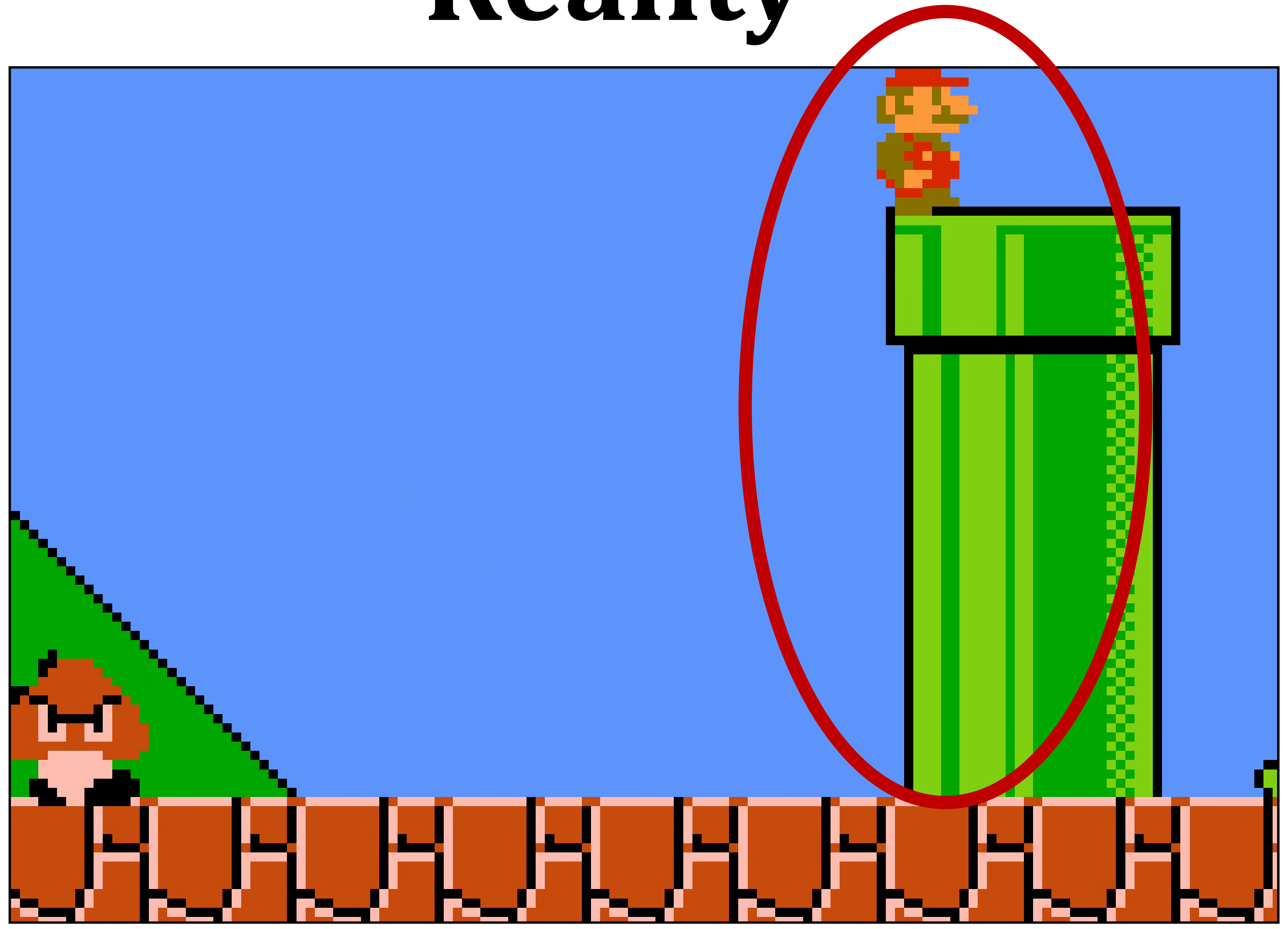




Prediction



Reality



Pathak, Agrawal, Efros, Darrell. ICML’17

Train a model

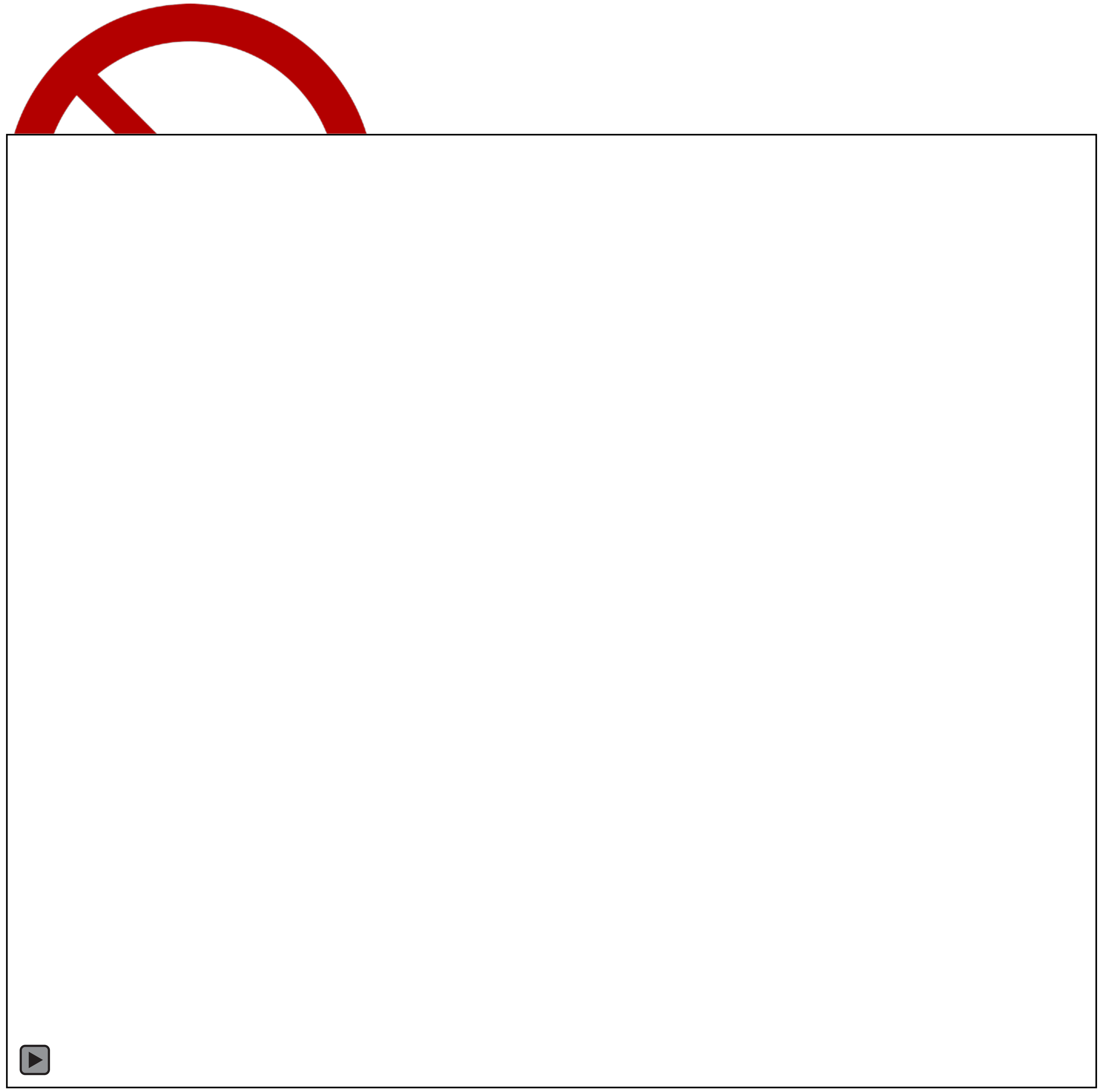
Predict consequences of the action

Bad prediction → **High** Curiosity

Ongoing battle between predictor vs. curiosity

Pathak, Agrawal, Efros, Darrell. ICML’17

No external reward, only curiosity



After curiosity-driven training

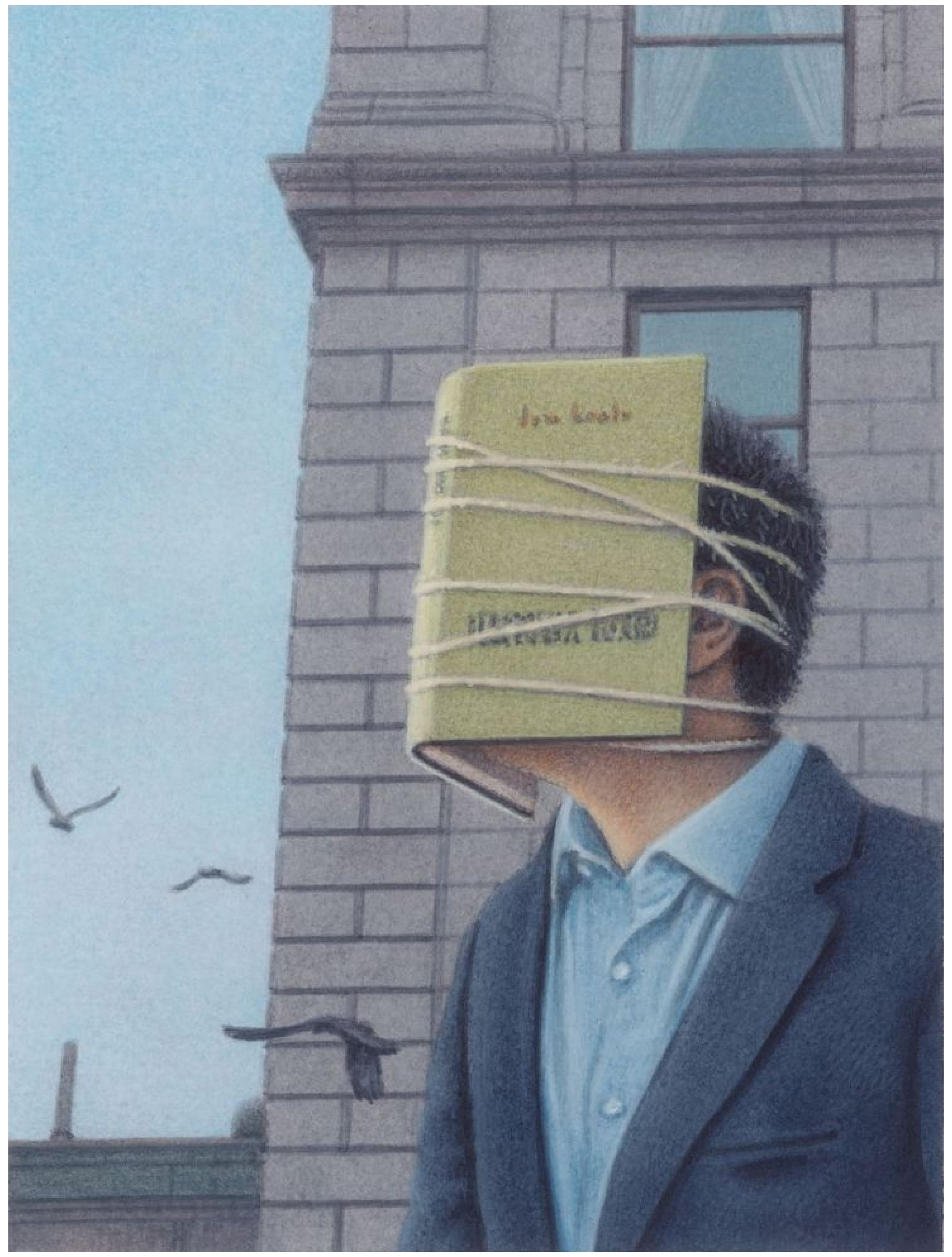
At the start of training

Curiosity on both sides… makes a rally

Environment: Multi-player Pong

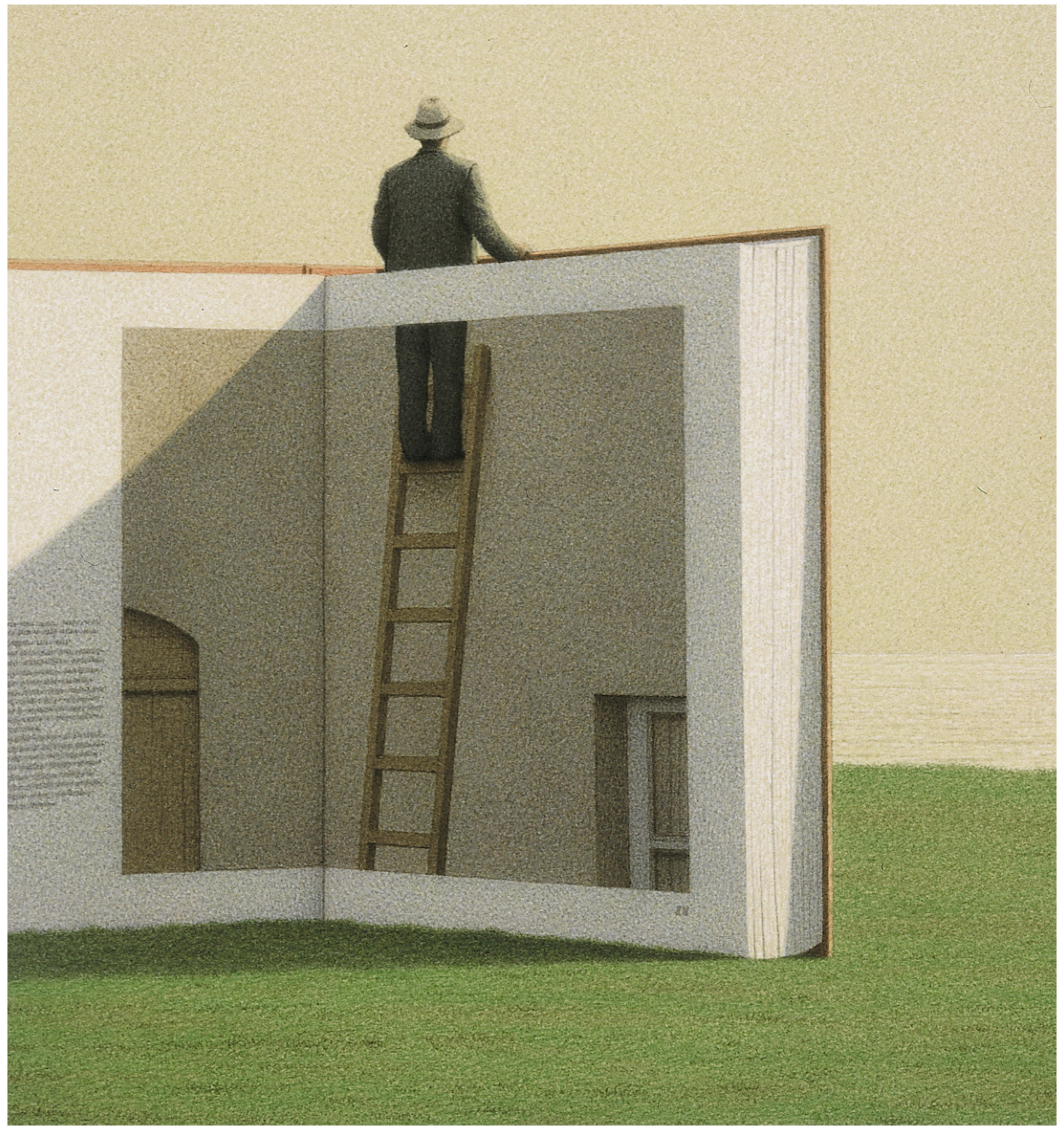
Summary: Why Self-Supervision?

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



⇨ data-driven association

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

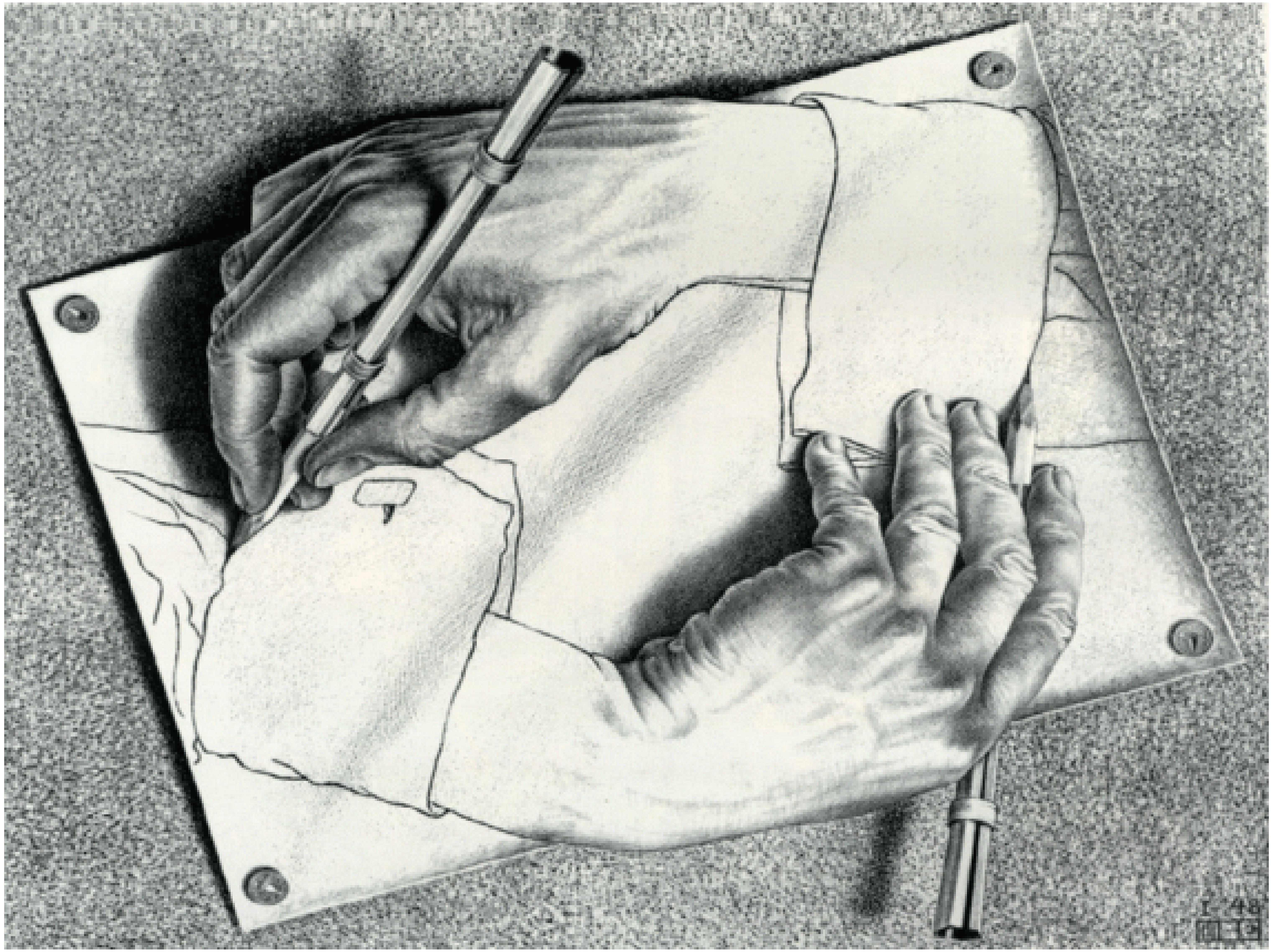


⇨ continuous, life-long learning

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

⇨ emergent objectives

Artwork by *Quint Buchholz*



Thank You