CS375 / Psych 249:
Large-Scale Neural Network Models for Neuroscience

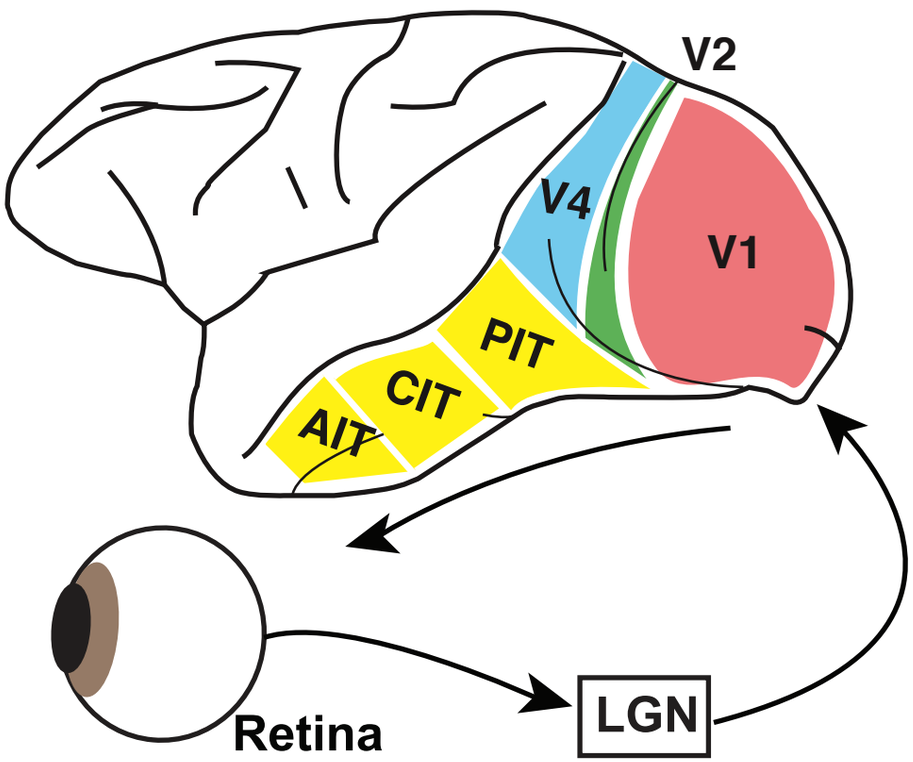
Lecture 6: Unsupervised Models of the Visual System

2025.01.26

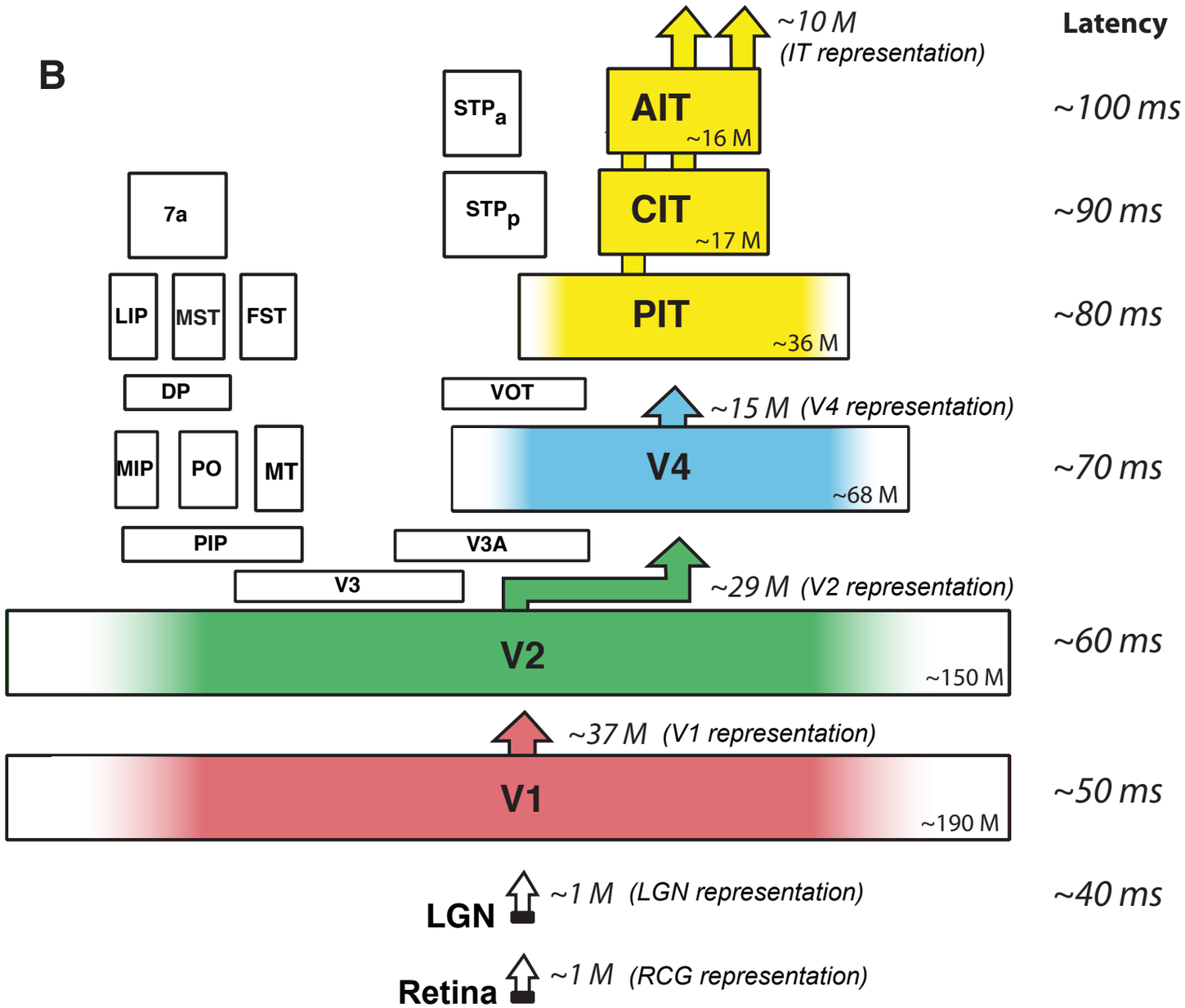
Daniel Yamins

Departments of Computer Science and of Psychology
Stanford Neuroscience and Artificial Intelligence Laboratory
Wu Tsai Neurosciences Institute
Stanford University

A



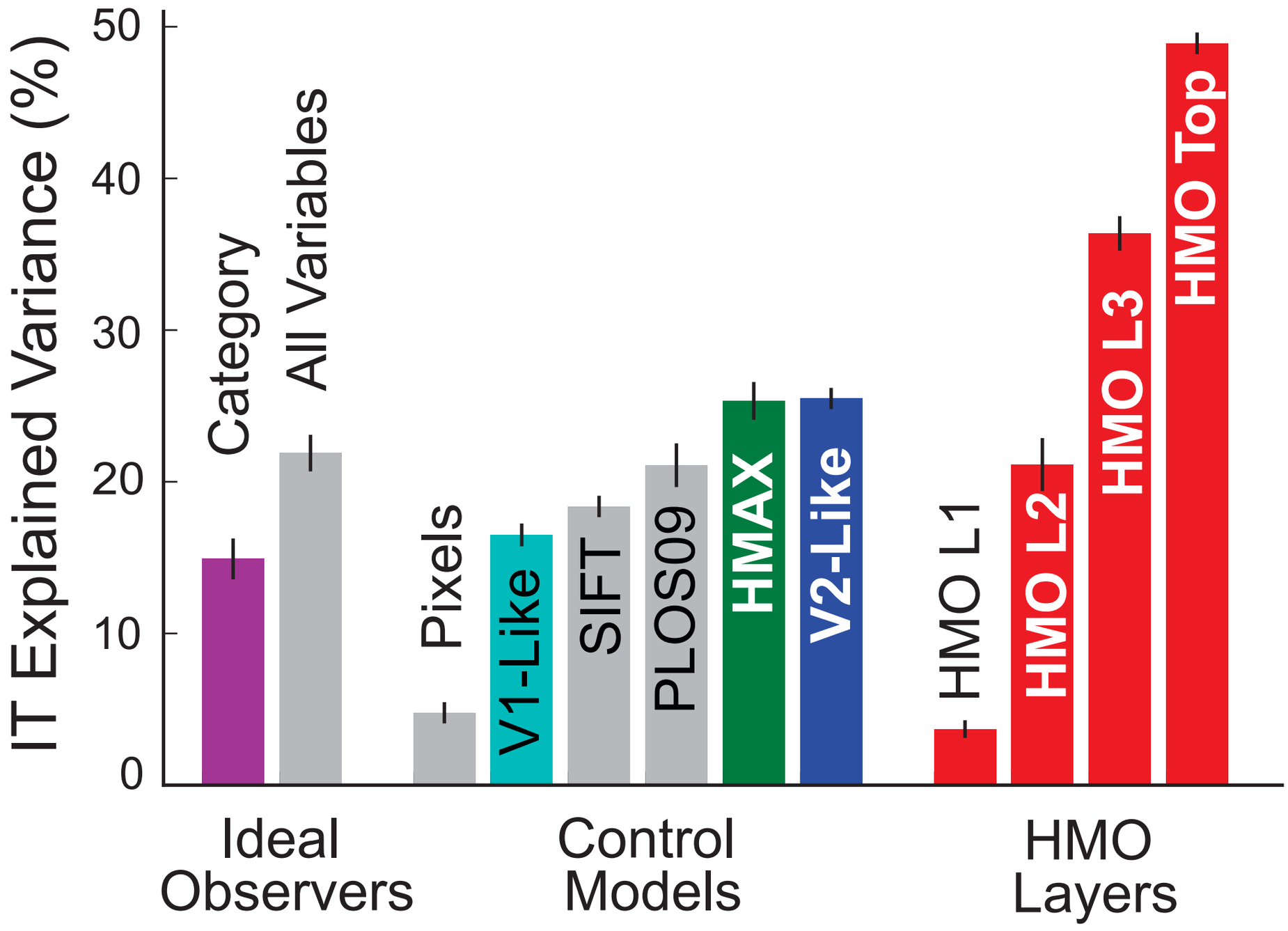
B



Adapted from DiCarlo et al. 2012

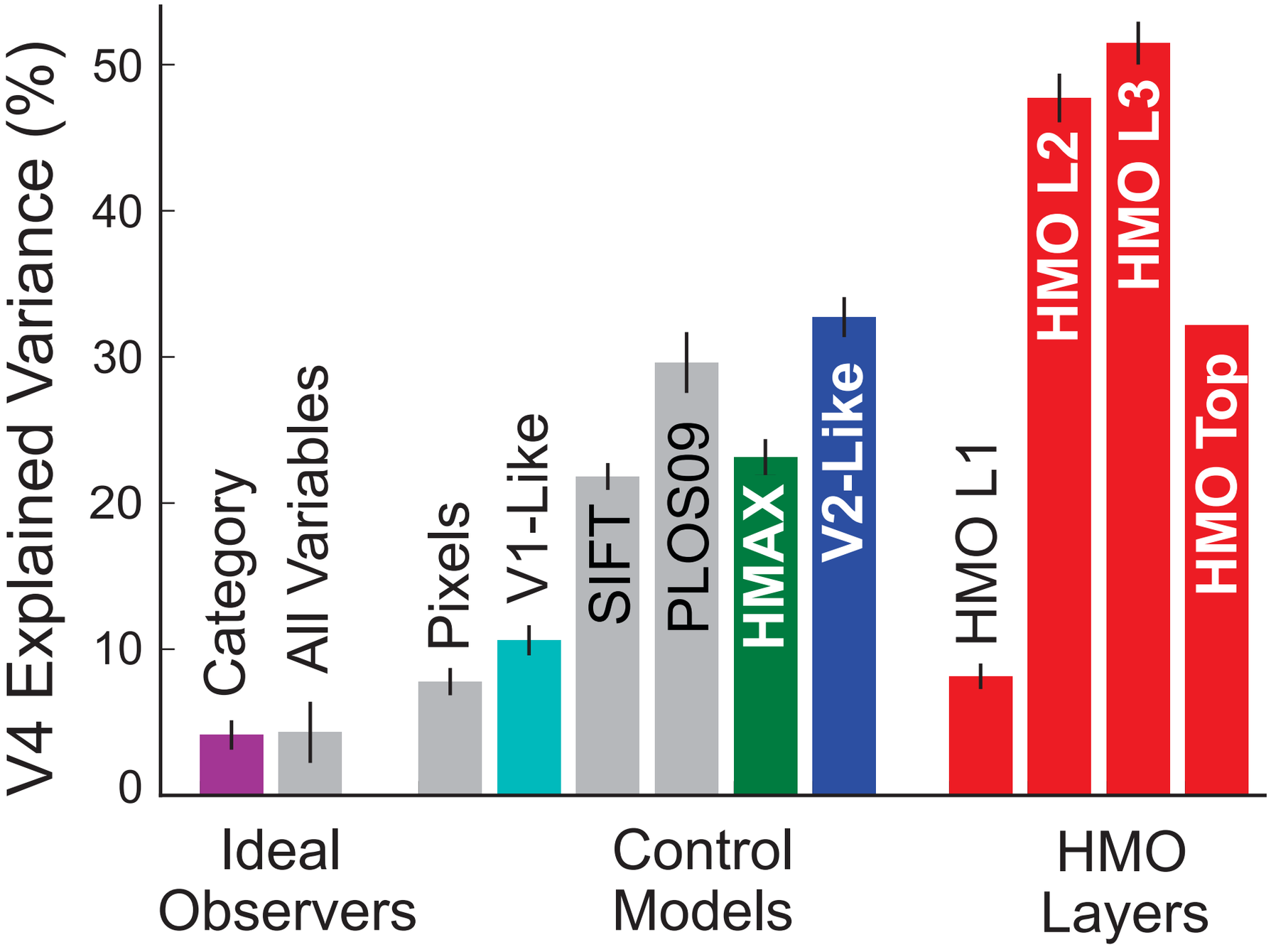
Predicting IT Neural Responses

Yamins and Hong* et. al.* ***PNAS*** *(2014)*



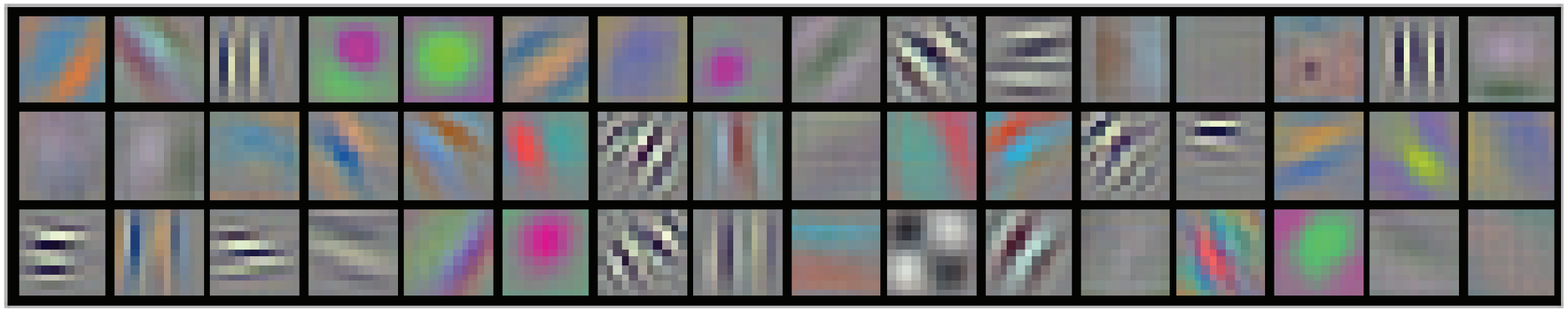
Predicting V4 Neural Responses

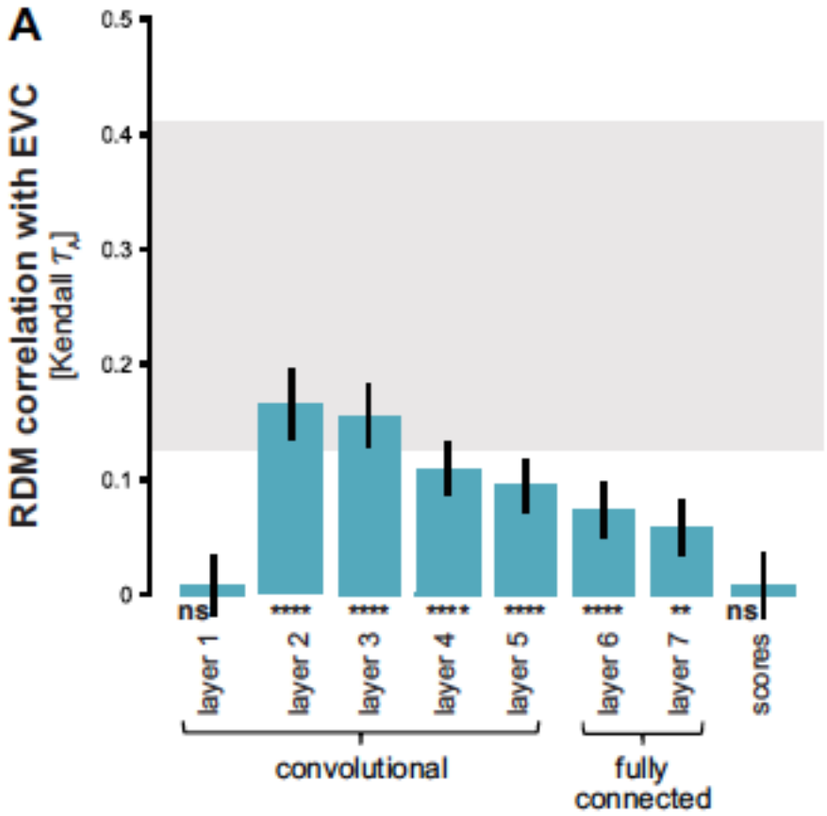
Yamins and Hong* et. al.* ***PNAS*** *(2014)*



Layer-area correspondence

Emergently, AlexNet filters at lowest layer resemble Gabor wavelets:





Model early layers are best explanation of fMRI data in V1. (*with Darren Seibert and Justin Gardner*)

Kaligh-Razavi and Kriegeskorte (2014)

Similar result: Guclu & Van Gerven (2015)

Four Principles of Goal-Driven Modeling

1.

A = *architecture class*

2.

T = *task/objective*

3.

D = *dataset*

4.

L = *learning rule*

Four Principles of Goal-Driven Modeling

1.

A = *architecture class*

2.

T = *task/objective*

3.

D = *dataset*

4.

L = *learning rule*

Best proxies thus far for ventral stream:

A = *ConvNets of reasonable depth*

T = *multi-way object categorization*

D = *ImageNet images*

L = *evolutionary architecture search + filter learning through gradient descent*

Four Principles of Goal-Driven Modeling

1.

A = *architecture class* **= *circuit neuro-anatomy***

2.

T = *task/objective* **= *ecological niche***

3.

D = *dataset* **= *environment***

4.

L = *learning rule* **= *natural selection + synaptic plasticity***

<u>Best proxies</u> thus far for ventral stream:

A = *ConvNets of reasonable depth*

T = *multi-way object categorization*

D = *ImageNet images*

L = *evolutionary architecture search + filter learning through gradient descent*

Four Principles of Goal-Driven Modeling

1.

A = *architecture class* **= circuit neuro-anatomy**

solving

2.

T = *task/objective* **= ecological niche**

situated in

3.

D = *dataset* **= environment**

updating according to

4.

L = *learning rule* **= natural selection + synaptic plasticity**

<u>Best proxies</u> thus far for ventral stream:

A = *ConvNets of reasonable depth*

T = *multi-way object categorization*

D = *ImageNet images*

L = *evolutionary architecture search + filter learning through gradient descent*

Big Problems in Each Area

***bad** = obviously deeply wrong as model of the brain or behavior

PROBLEM

1. ***Xbad***

A = *architecture class*

e.g. **CNNs**

2.

T = *task/objective*

e.g. **Object Categorization**

3.

D = *dataset*

e.g. **ImageNet**

4.

L = *learning rule*

e.g. **Arch. Srch.** + **Grad. Desc.**

Big Problems in Each Area

***bad** = obviously deeply wrong as model of the brain or behavior

PROBLEM

RECURRENCE and FEEDBACK!!?

1. ***bad***

A = *architecture class*

e.g. **CNNs**

2.

T = *task/objective*

e.g. **Object Categorization**

3.

D = *dataset*

e.g. **ImageNet**

4.

L = *learning rule*

e.g. **Arch. Srch.** + **Grad. Desc.**

Big Problems in Each Area

***bad** = obviously deeply wrong as model of the brain or behavior

PROBLEM

1. ✗***bad***

A = *architecture class*

e.g. **CNNs**

RECURRENCE and FEEDBACK!!?

2. ✗***bad***

T = *task/objective*

e.g. **Object Categorization**

TOO MUCH LABELLED DATA REQUIRED!!?

3.

D = *dataset*

e.g. **ImageNet**

4.

L = *learning rule*

e.g. **Arch. Srch.** + **Grad. Desc.**

Big Problems in Each Area

***bad** = obviously deeply wrong as model of the brain or behavior

PROBLEM

1. ✗***bad***

A = *architecture class*

e.g. **CNNs**

RECURRENCE and FEEDBACK!!?

2. ✗***bad***

T = *task/objective*

e.g. **Object Categorization**

TOO MUCH LABELLED DATA REQUIRED!!?

3. ✗***bad***

D = *dataset*

e.g. **ImageNet**

REAL NOISY VIDEO DATASTREAMS vs STEREOTYPED CLEAN STILL IMAGES

4.

L = *learning rule*

e.g. **Arch. Srch.** + **Grad. Desc.**

Big Problems in Each Area

***bad** = obviously deeply wrong as model of the brain or behavior

PROBLEM

1. ✗*bad*

A = *architecture class*

e.g. **CNNs**

RECURRENCE and FEEDBACK!!?

2. ✗*bad*

T = *task/objective*

e.g. **Object Categorization**

TOO MUCH LABELLED DATA REQUIRED!!?

3. ✗*bad*

D = *dataset*

e.g. **ImageNet**

REAL NOISY VIDEO DATASTREAMS vs STEREOTYPED CLEAN STILL IMAGES

4. ✗*bad*

L = *learning rule*

e.g. **Arch. Srch.** + **Grad. Desc.**

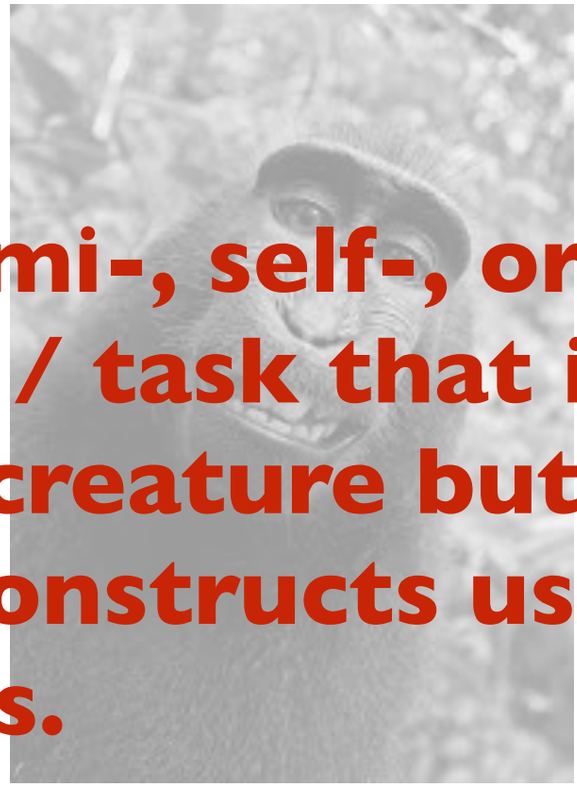
BACKPROP AND ITS DISCONTENTS

The Supervision Problem

There's just no way that these creatures receive millions of high-level semantic labels during learning.

Effective proxy, but just obviously deeply wrong.

The Problem

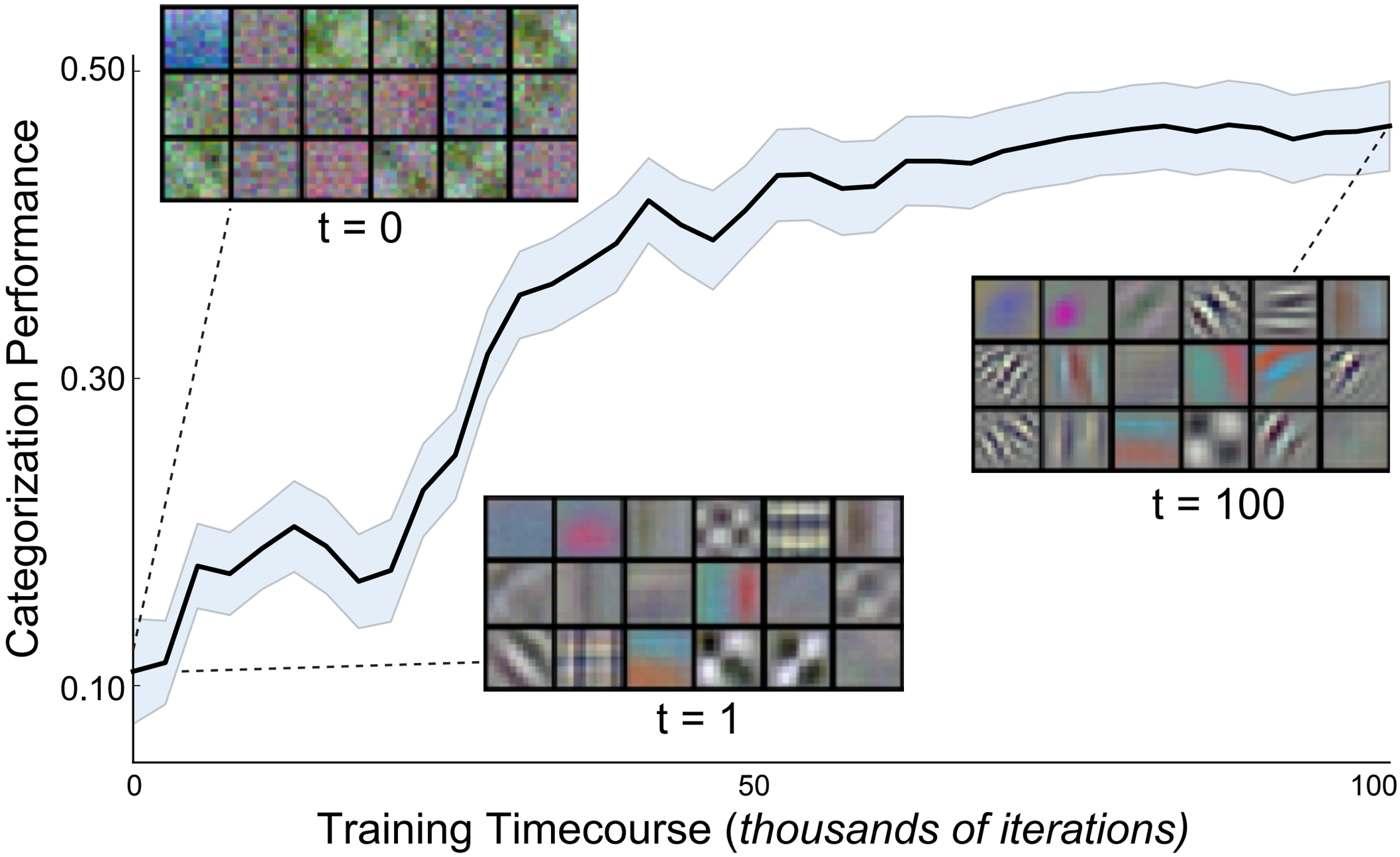


Must find some sort of semi-, self-, or unsupervised loss function / task that is "realistically costly" to the creature but is sufficiently powerful that it constructs useful representations.

There's just no way that these creatures receive millions of high-level semantic labels during learning.

Effective proxy, but just obviously deeply wrong.

Goal: Developmental Model



Survey of Unsupervised Methods

Generic:

- Clustering
- Mixtures
- Factorization
- Manifold learning

Semi-generic:

- Auto encoders
- BiGANs

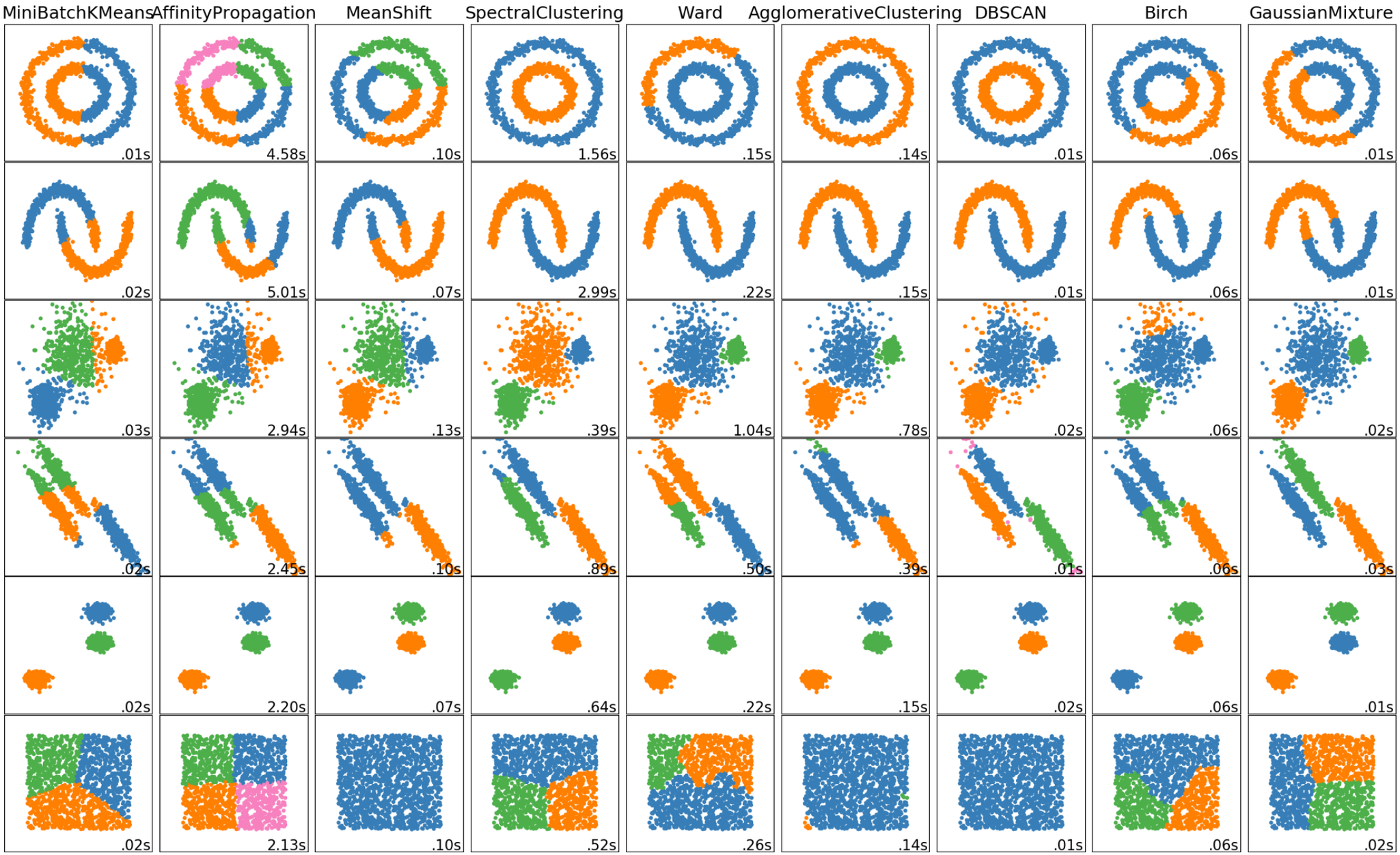
Not generic:

- Other problem-domain specific stuff

Survey of Unsupervised Methods: **Clustering**

Clustering: assign datapoint to natural groups based on the way the data is laid out.

Some of the many methods of clustering



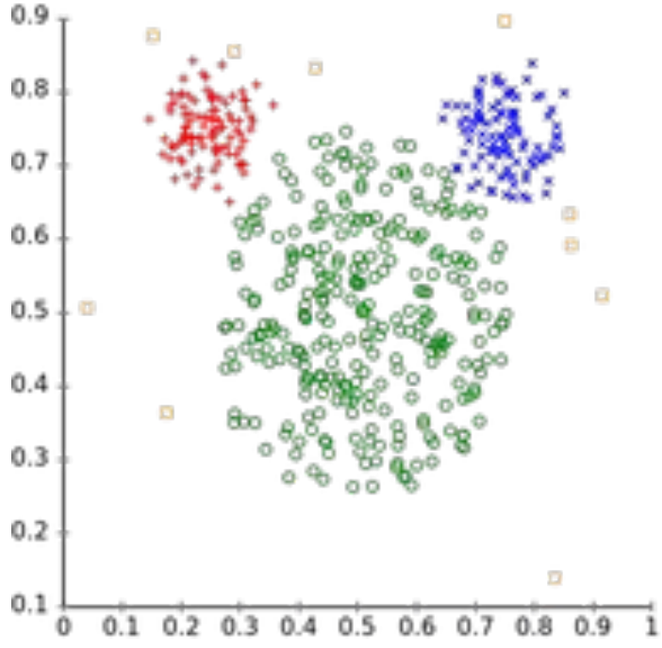
Survey of Unsupervised Methods: **Clustering**

K-means (Lloyd's algorithm)

k = number of clusters

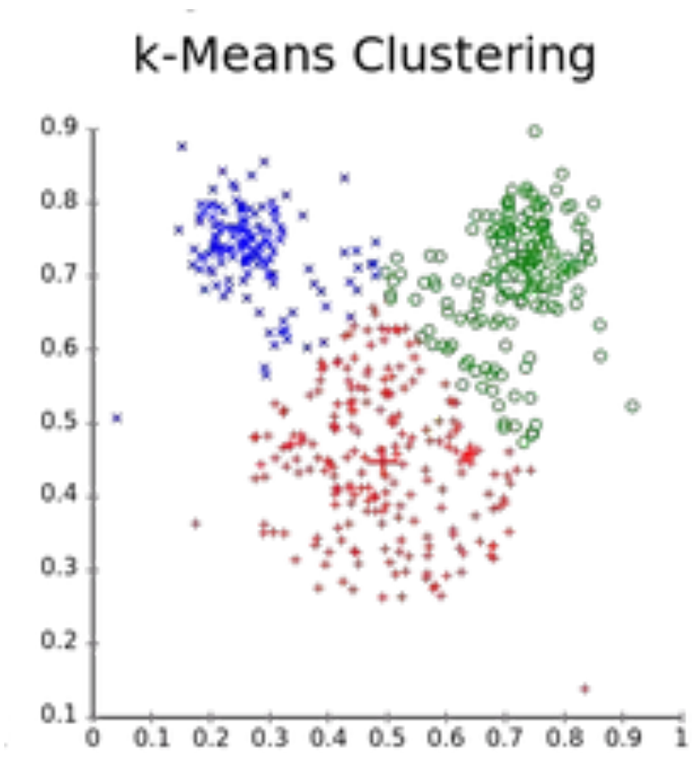
C = partition into clusters

μ_i = mean of i-th cluster



C assignment chosen ('learned') to minimize:

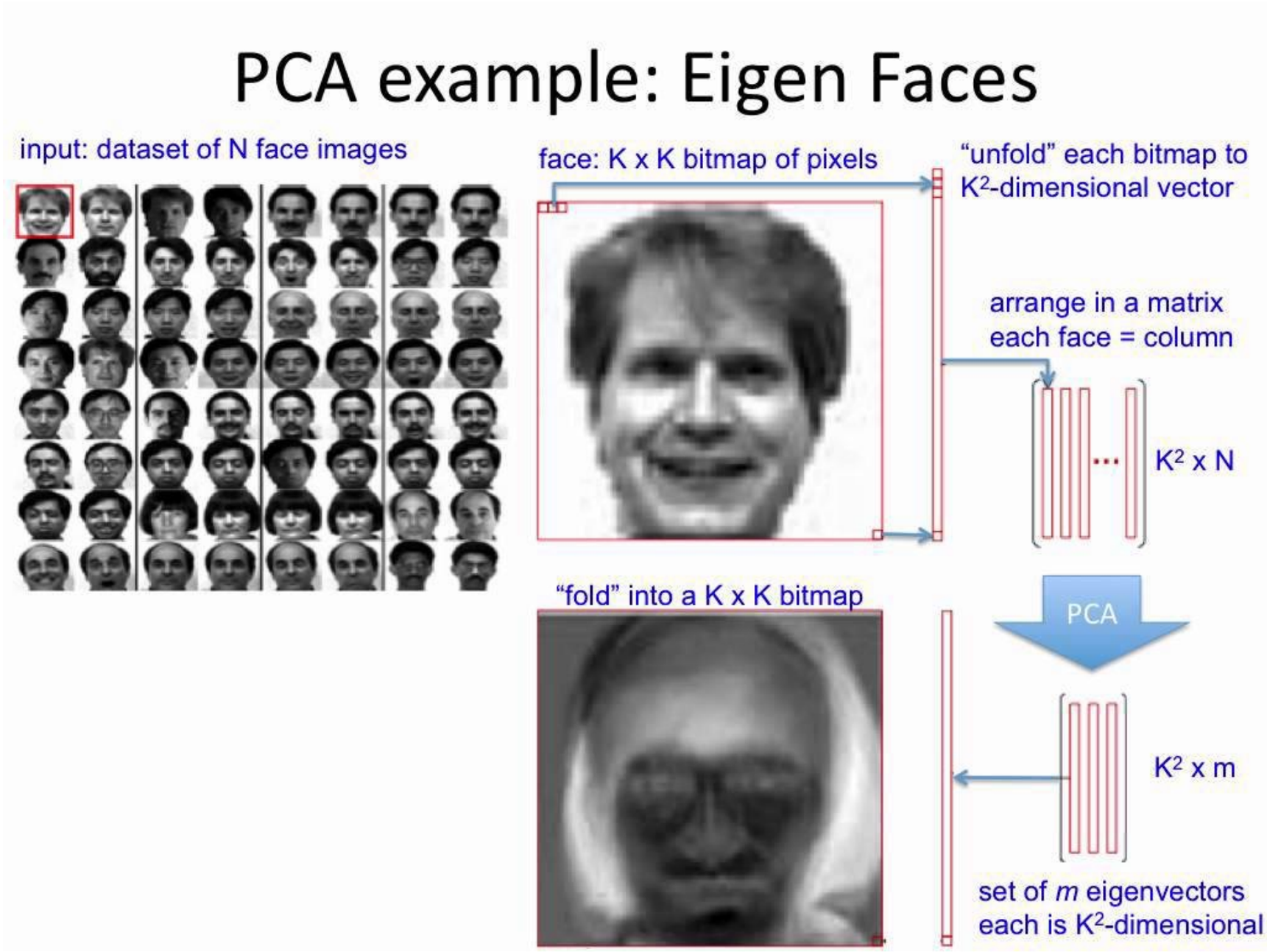
$$\sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 = \sum_{i=1}^{k} |C_i| \cdot Var(C_i)$$



1 parameter, *k*, learned via supervision

Survey of Unsupervised Methods: **(Linear) Factorization**

PCA

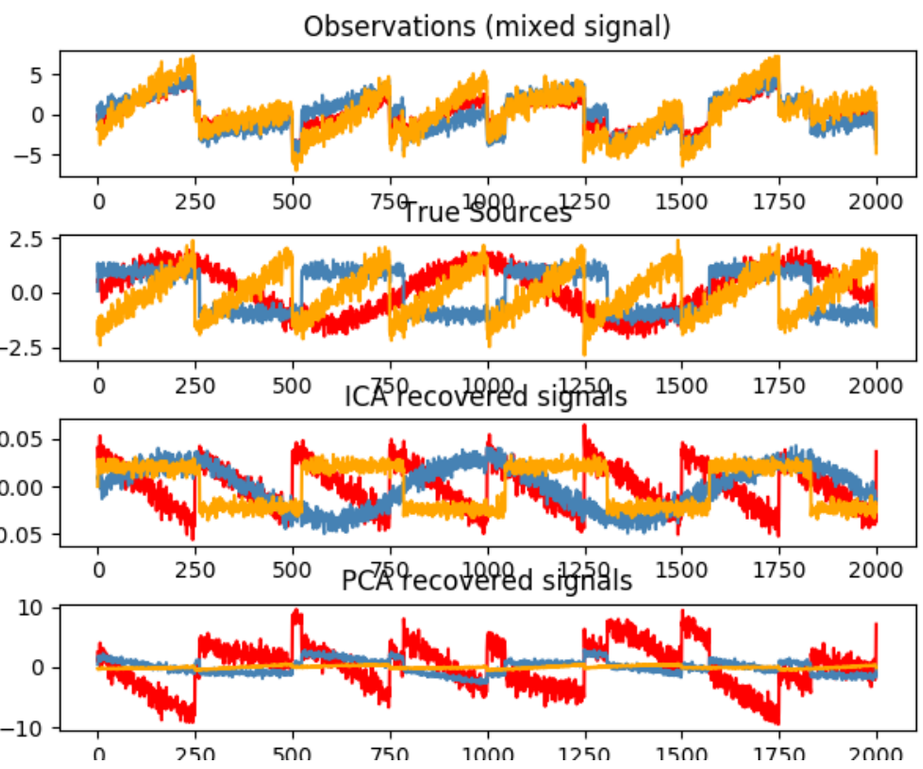


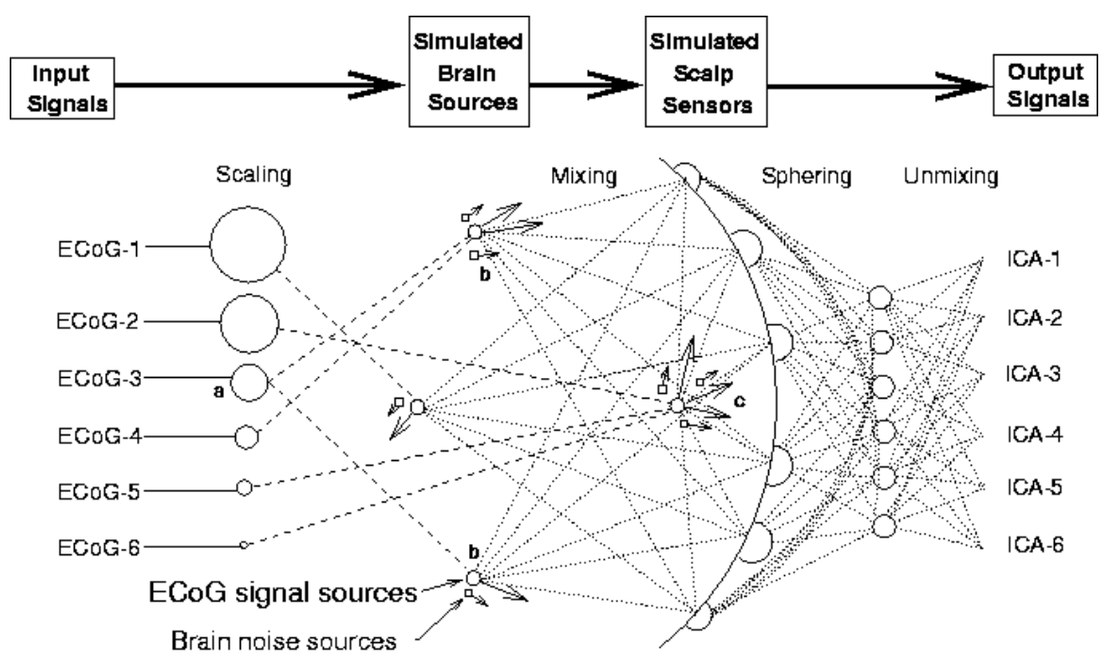
Survey of Unsupervised Methods: **Factorization**

Independent Component Analysis (ICA)

$$x = A \cdot s$$

where the s_i are statistically independent signals





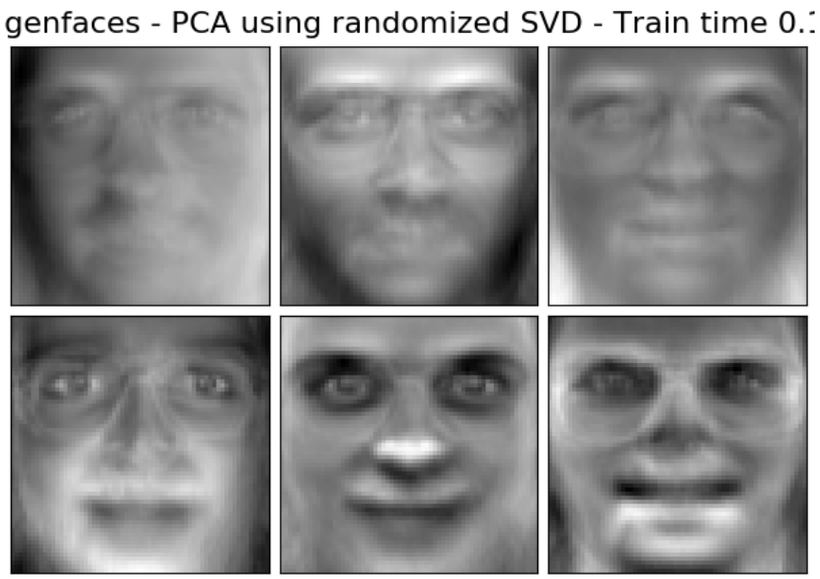
ERP Simulation using ECoG Data

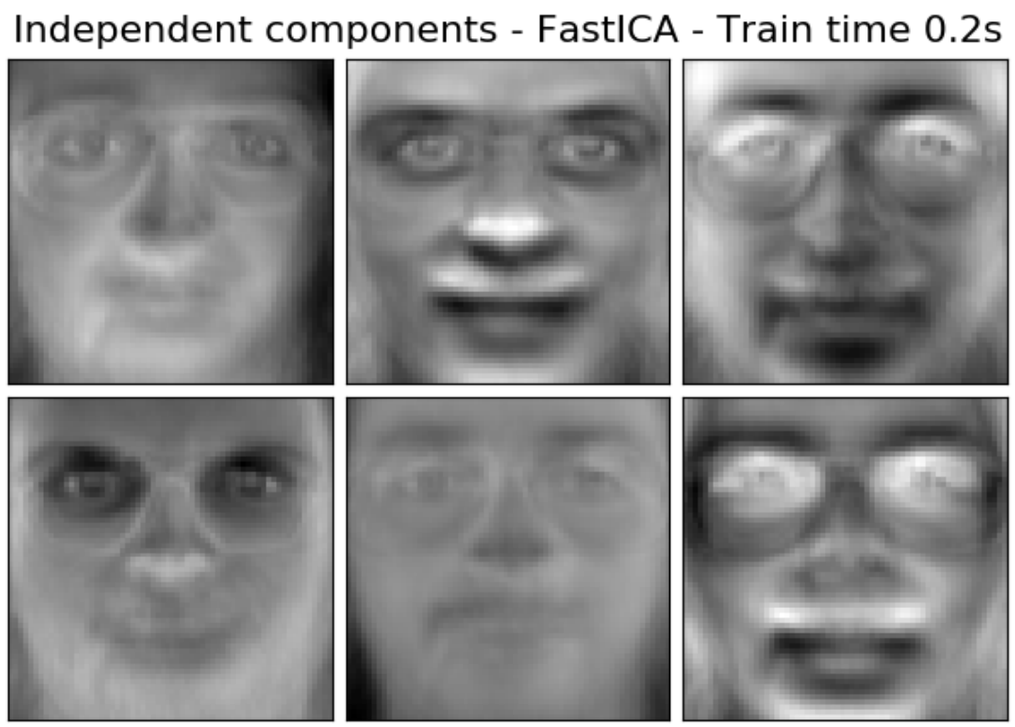
Survey of Unsupervised Methods: **Factorization**

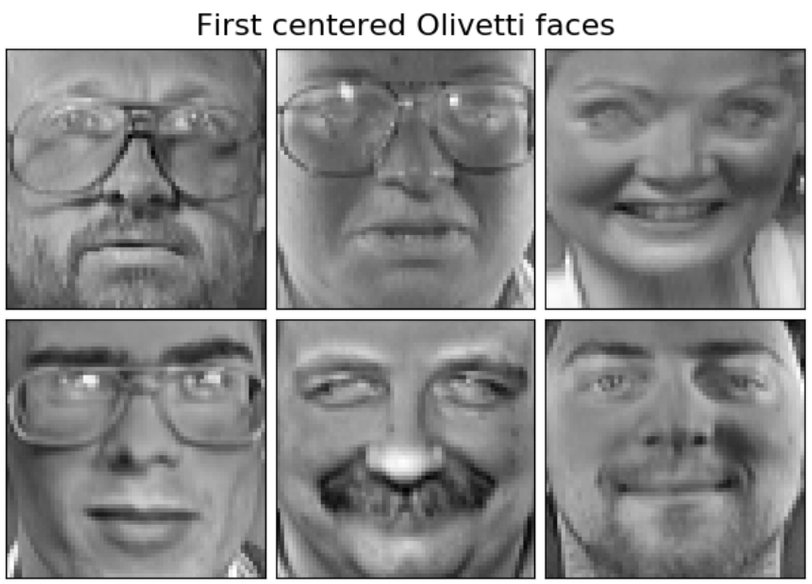
Independent Component Analysis (ICA)

$$x = A \cdot s$$

where the s_i are statistically independent signals







Survey of Unsupervised Methods: **Factorization**

fMRI response data collected* on 165 commonly heard natural sound stimuli.

Man speaking
Flushing toilet
Pouring liquid
Tooth-brushing
Woman speaking
Car accelerating
Biting and chewing
Laughing
Typing
Car engine starting
Running water
Breathing
Keys jangling
Dishes clanking
Ringtone
Microwave
Dog barking

Road traffic
Zipper
Cellphone vibrating
Water dripping
Scratching
Car windows
Telephone ringing
Chopping food
Telephone dialing
Girl speaking
Car horn
Writing
Computer startup sound
Background speech
Songbird
Pouring water
Pop song
Water boiling

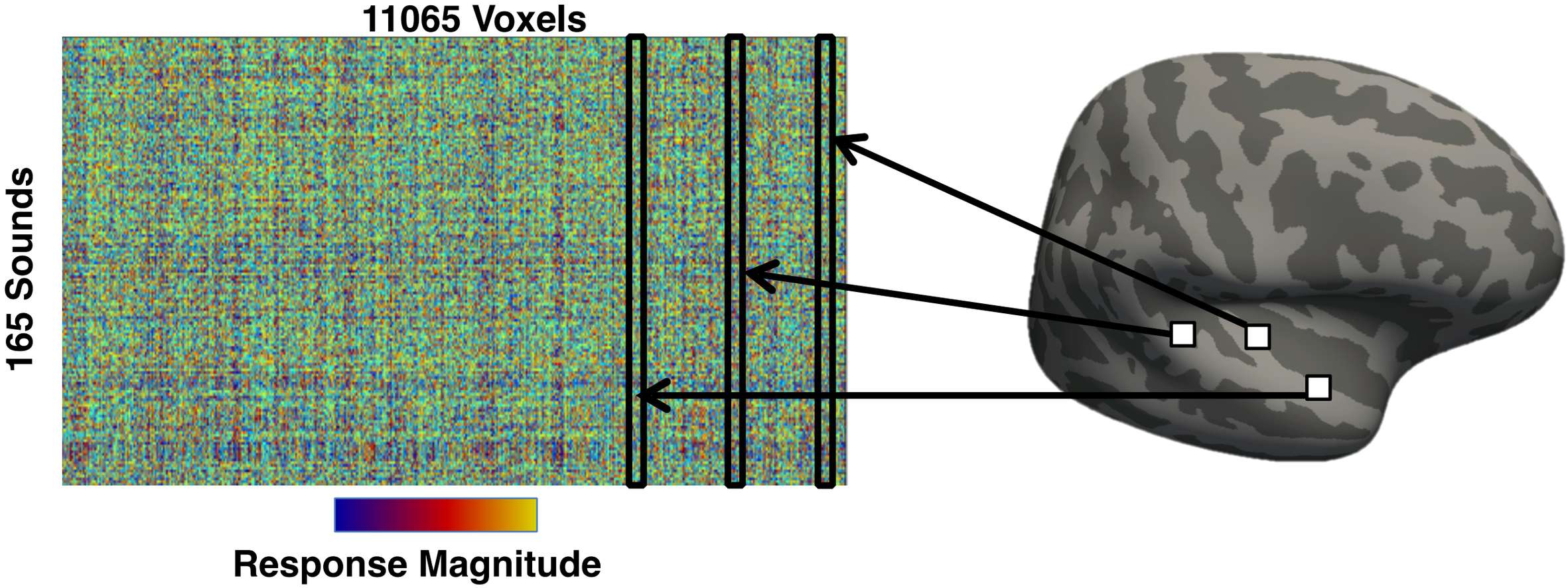
Guitar
Coughing
Crumpling paper
Siren
Splashing water
Computer speech
Alarm clock
Walking with heels
Vacuum
Wind
Boy speaking
Chair rolling
Rock song
Door knocking

•
•
•

*Sam Norman-Haignere, Nancy Kanwisher, and Josh McDermott

Survey of Unsupervised Methods: **Factorization**

For each voxel, measured average response to each sound:



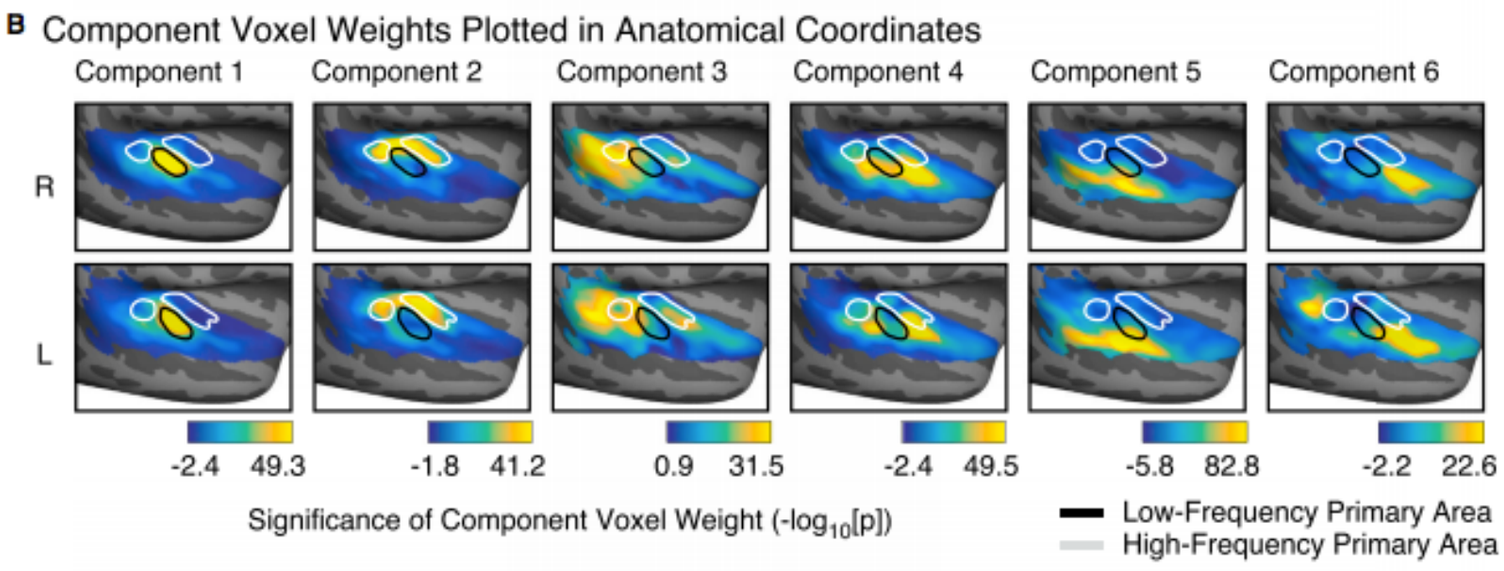
Data matrix: voxels X sounds.

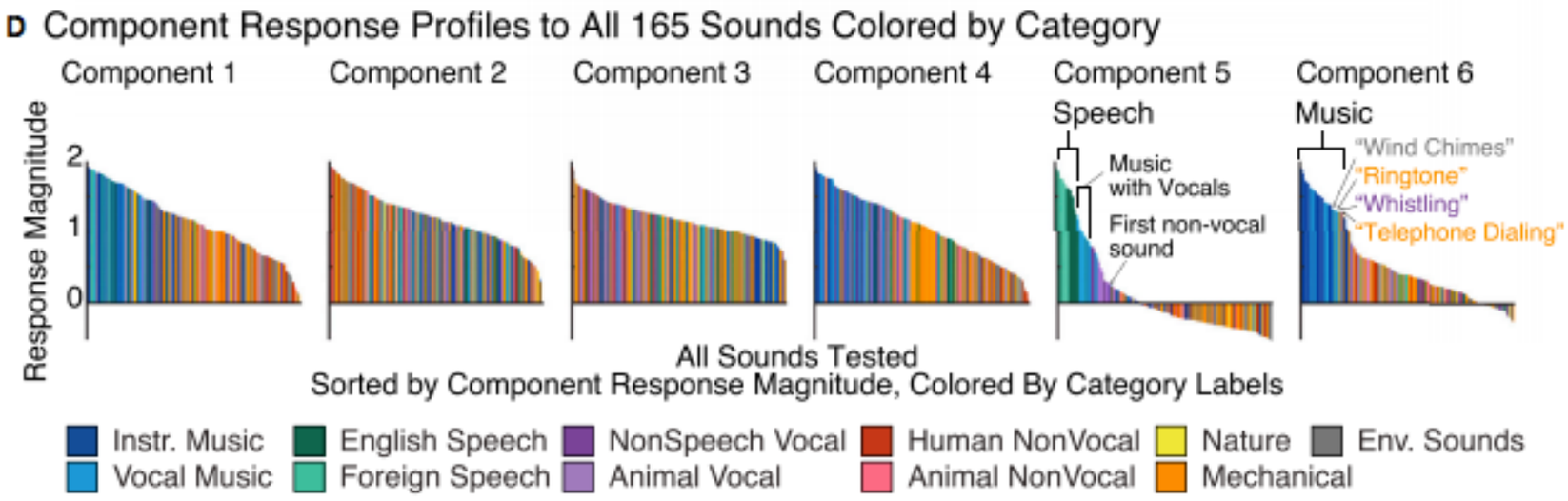
Survey of Unsupervised Methods: **Factorization**

Independent Component Analysis (ICA)

$$x = A \cdot s$$

where the s_i are statistically independent signals





Norman-Haigniere, et. al. 2015

Survey of Unsupervised Methods: **Factorization**

Non-negative Matrix Factorization (NMF)

Sebastian Seung

$$X \sim W \cdot H \qquad\qquad W, H \geq 0$$

minimize: $$\frac{1}{2} \sum_{ij} (X_{ij} - (W \cdot H)_{ij})^2$$

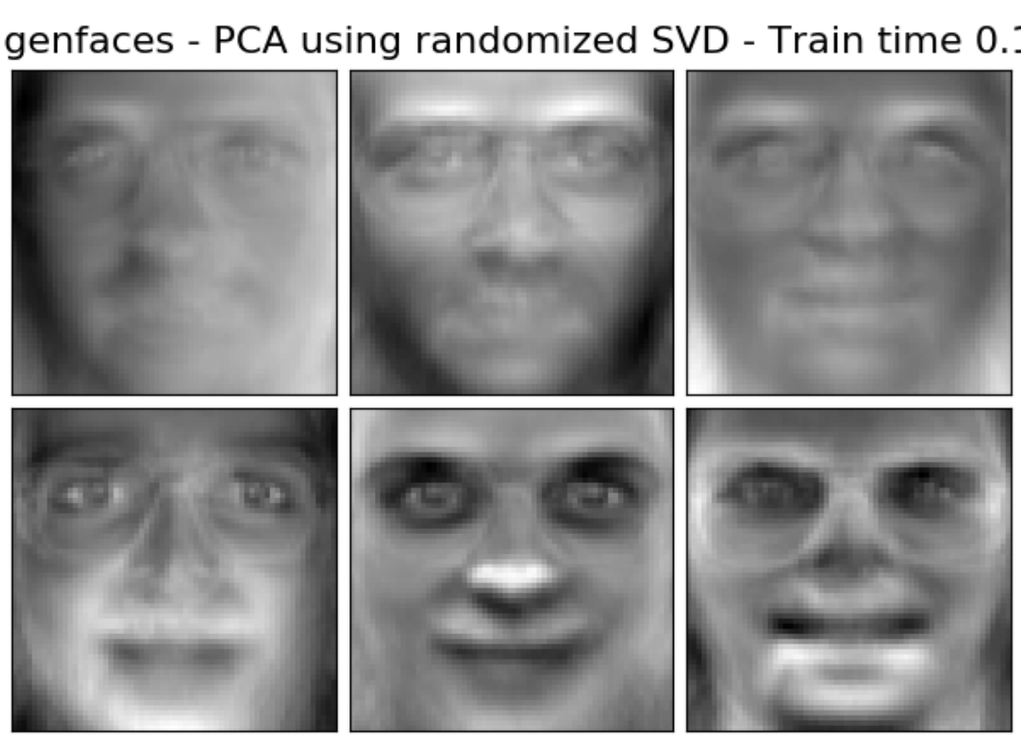
Survey of Unsupervised Methods: **Factorization**

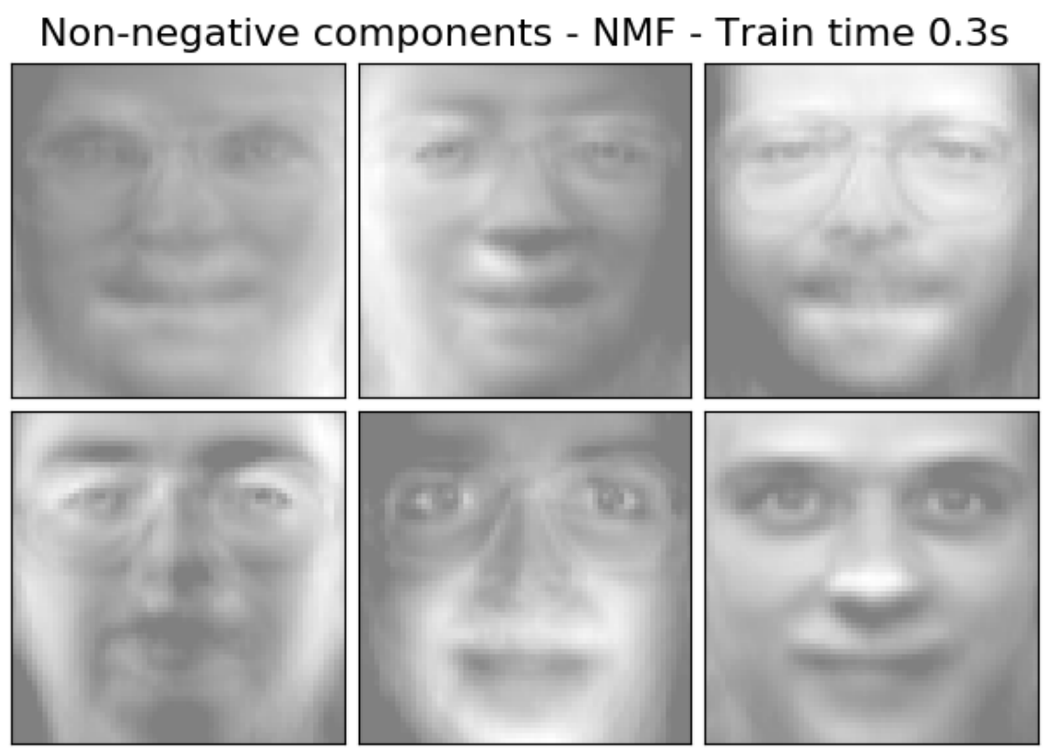
Non-negative Matrix Factorization (NMF)

Sebastian Seung

$$X \sim W \cdot H \qquad\qquad W, H \geq 0$$

minimize: $$\frac{1}{2} \sum_{ij} (X_{ij} - (W \cdot H)_{ij})^2$$





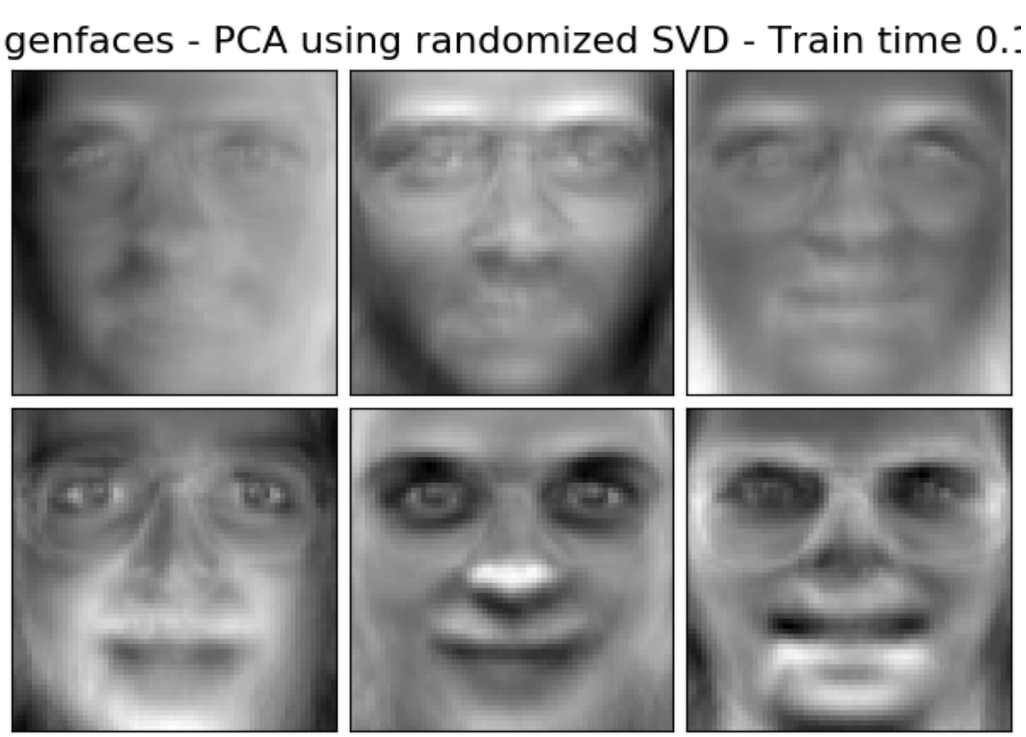
Survey of Unsupervised Methods: **Factorization**

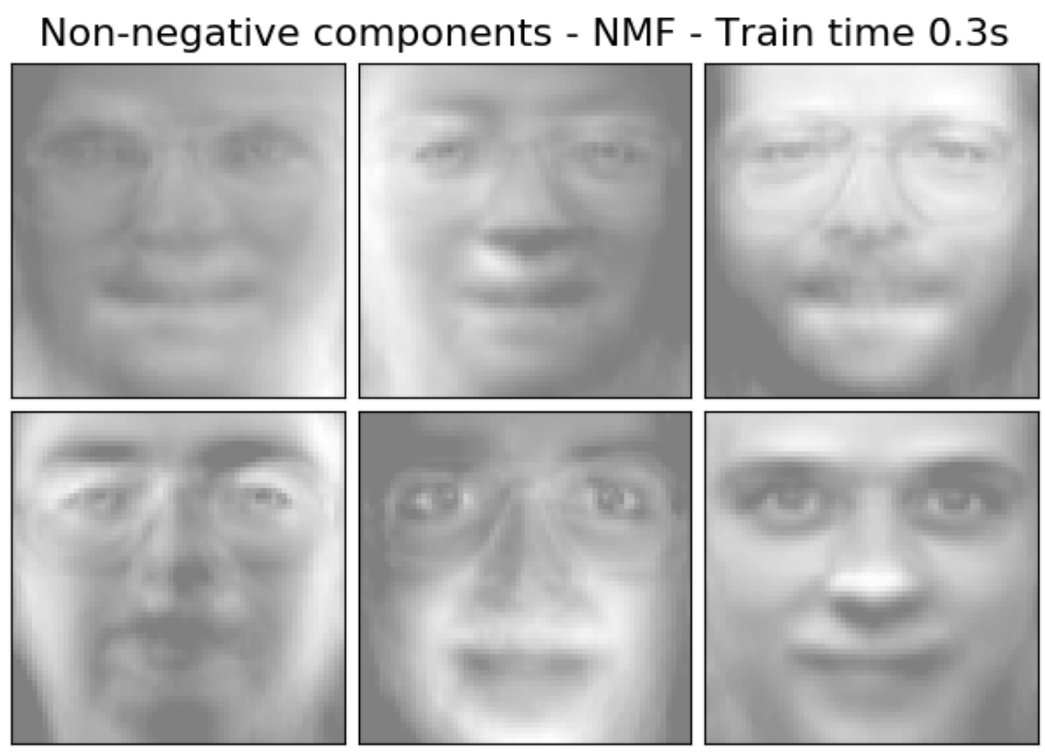
Non-negative Matrix Factorization (NMF)

Sebastian Seung

$$X \sim W \cdot H \qquad\qquad W, H \geq 0$$

minimize: $$\frac{1}{2}\sum_{ij}(X_{ij} - (W \cdot H)_{ij})^2$$





regularization: $$\frac{1}{2}\sum_{ij}(X_{ij} - (W \cdot H)_{ij})^2 + \alpha(||W||_1 + ||H||_1)$$

Survey of Unsupervised Methods: **Factorization**

Non-negative Matrix Factorization (NMF)

NEURORESOURCE

Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data

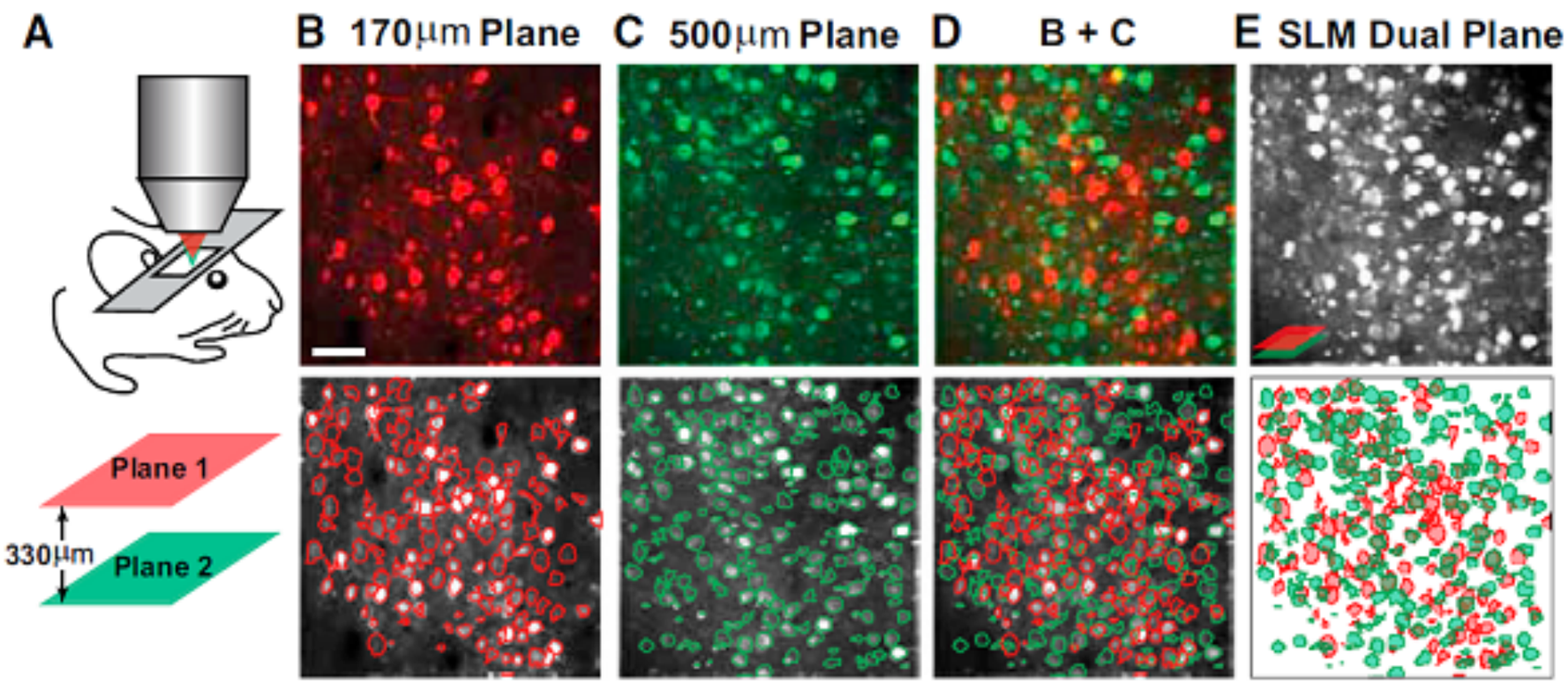
Eftychios A. Pnevmatikakis, Daniel Soudry, Yuanjun Gao, Timothy A. Machado, Josh Merel, David Pfau, Thomas Reardon, Yu Mu, Clay Lacefield, Weijian Yang, Misha Ahrens, Randy Bruno, Thomas M. Jessell, Darcy S. Peterka, Rafael Yuste, Liam Paninski

Published Online: January 07, 2016

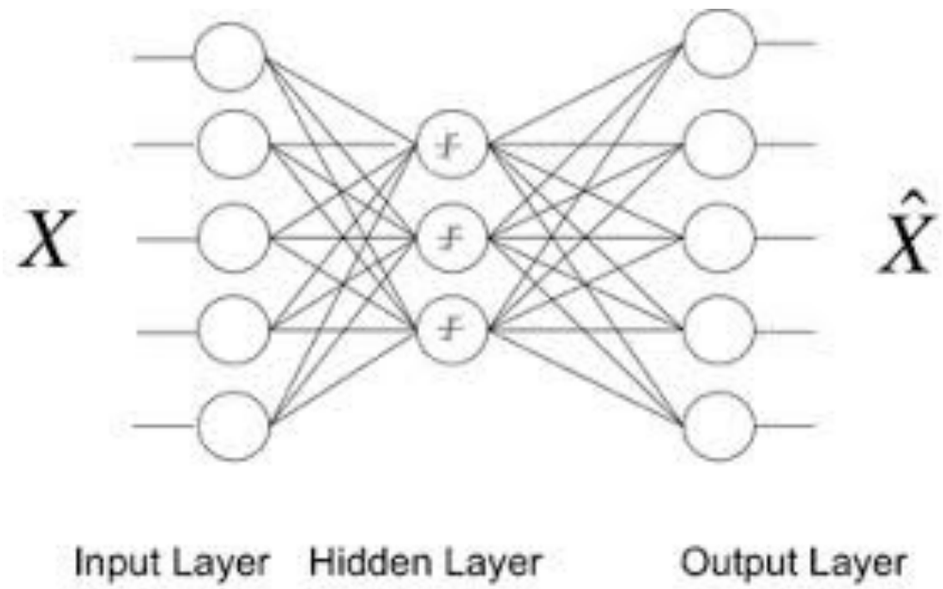
Open Archive | PlumX Metrics

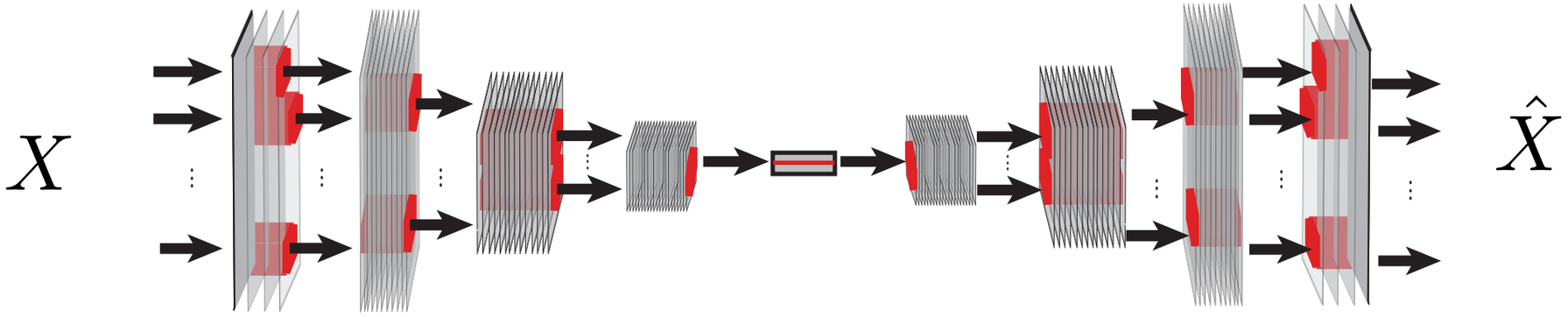
DOI: http://dx.doi.org/10.1016/j.neuron.2015.11.037 | CrossMark

Article Info



Survey of Unsupervised Methods: **Autoencoders**



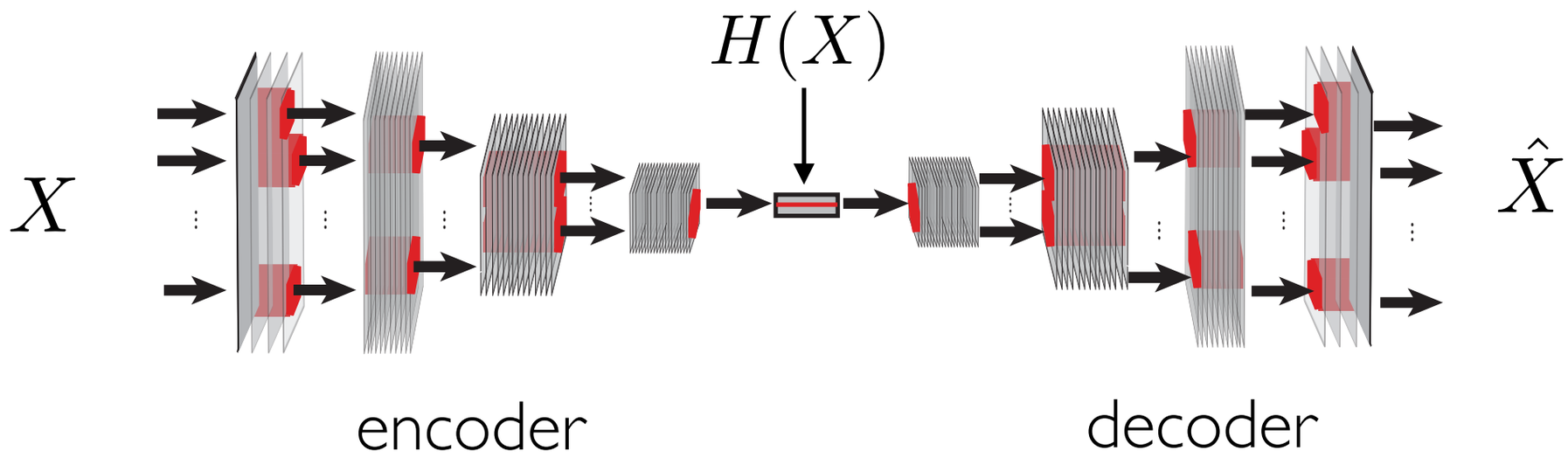


encoder decoder

$$\text{Loss} = ||X - \hat{X}||^2 + Penalty(H(X))$$

reconstruction complexity metric

Survey of Unsupervised Methods: **Autoencoders**



$$\text{Loss} = \quad ||X - \hat{X}||^2 + Penalty(H(X))$$

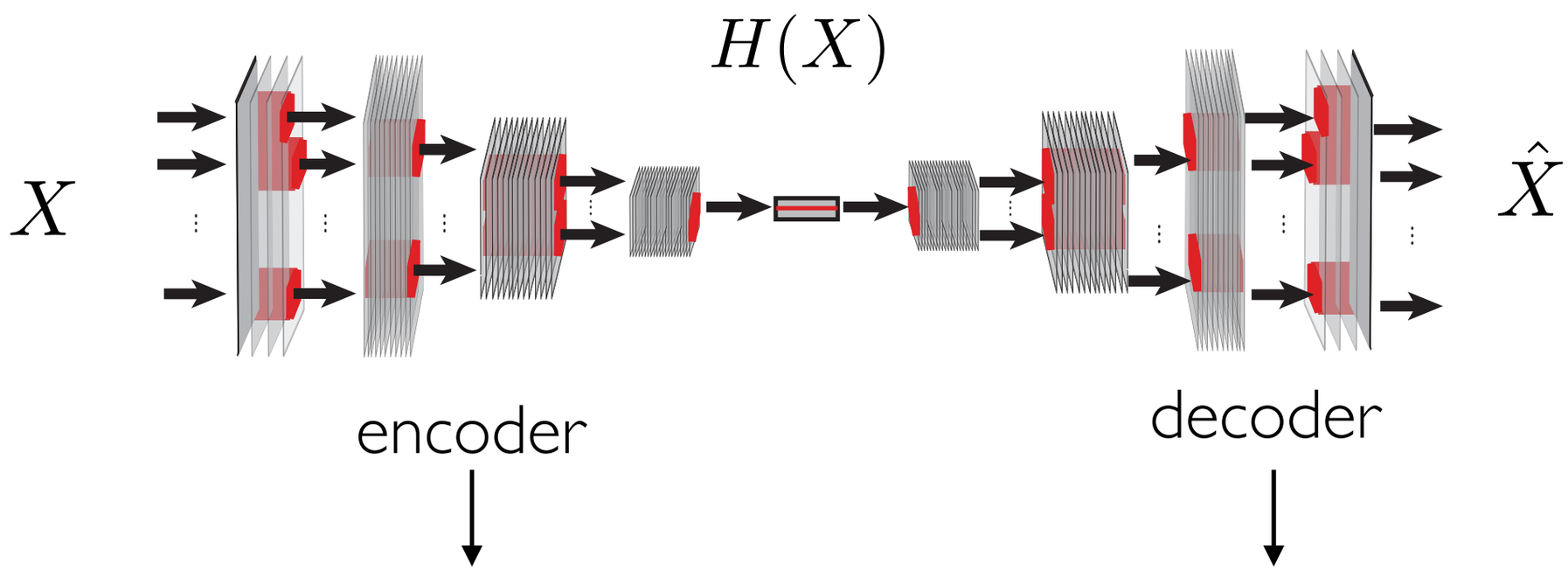
reconstruction complexity metric

Various penalties:

- low dimensionality of H(X) e.g. compression
- Penalty(X) = |X|, e.g. activation sparseness
- Penalty(X) = KL divergence to some simple distribution

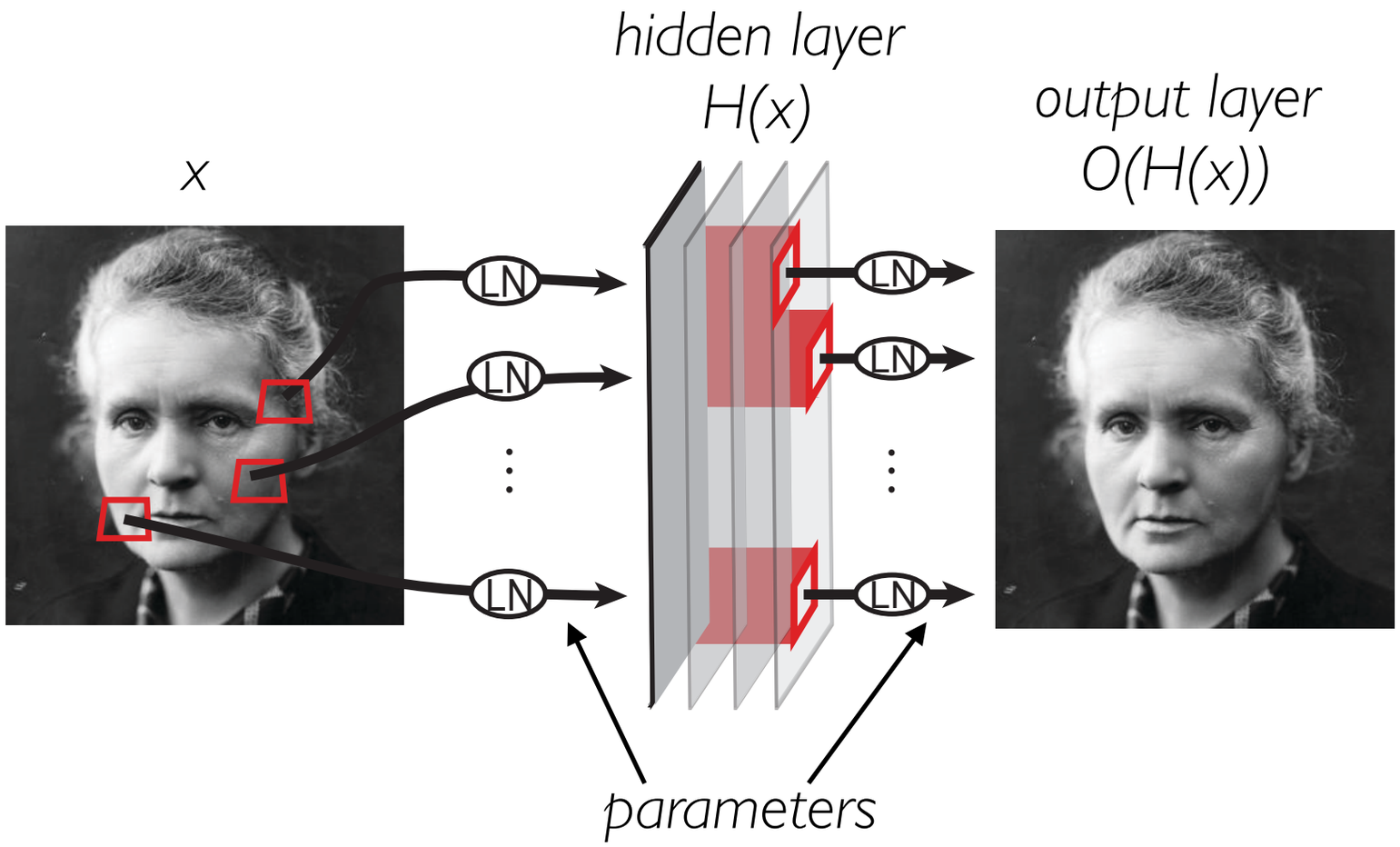
Parameters: whatever the parameters of the encoder & decoder are.

Survey of Unsupervised Methods: **Autoencoders**



original method:	"dictionary" learned offline by e.g. backdrop	reconstruction weights estimated "online" in an inner loop (no params)
modern method:	"dictionary" learned offline by e.g. backprop	also learned via backprop parametrizing FF neural network

Survey of Unsupervised Methods: **Sparse Autoencoders**

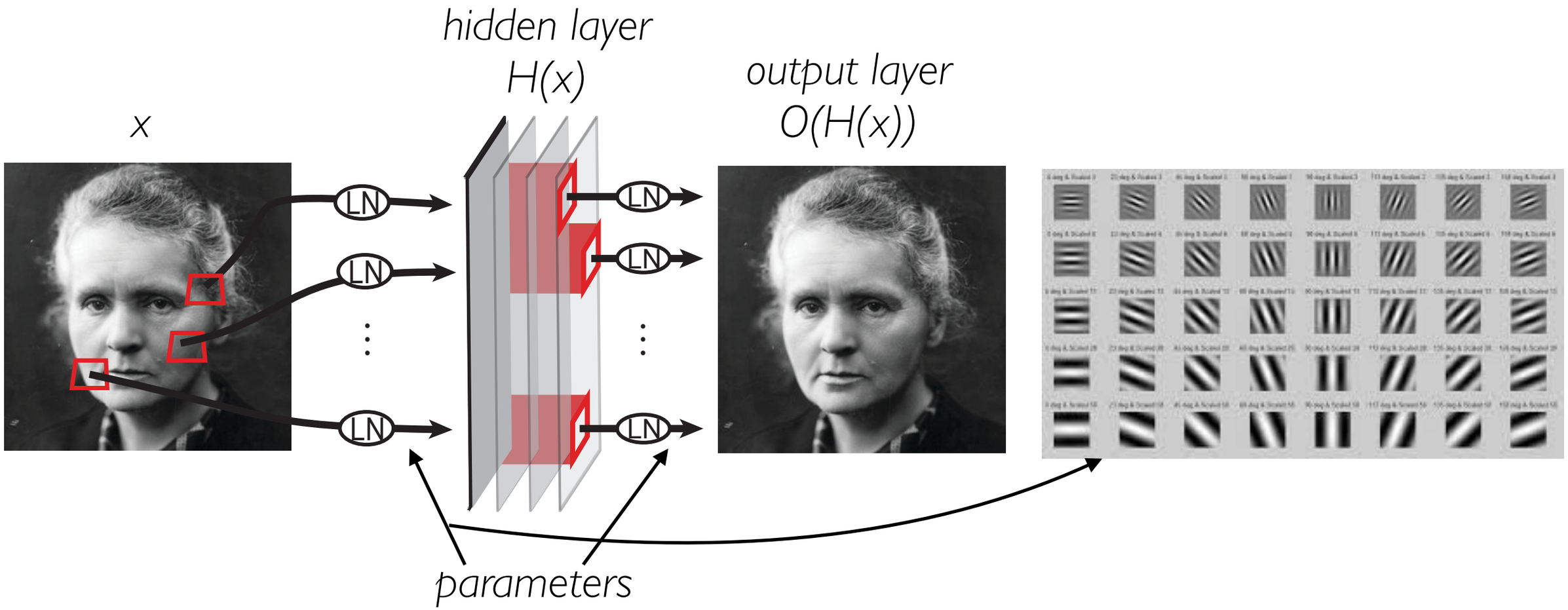


$$L(x) = |x - O(H(x))|^2 + \lambda \cdot |H(x)|$$

Sparse Coding Foldiak, Olshausen, mid 1990s

→neurons have to represent their environment, as efficiently as possible

Survey of Unsupervised Methods: **Sparse Autoencoders**

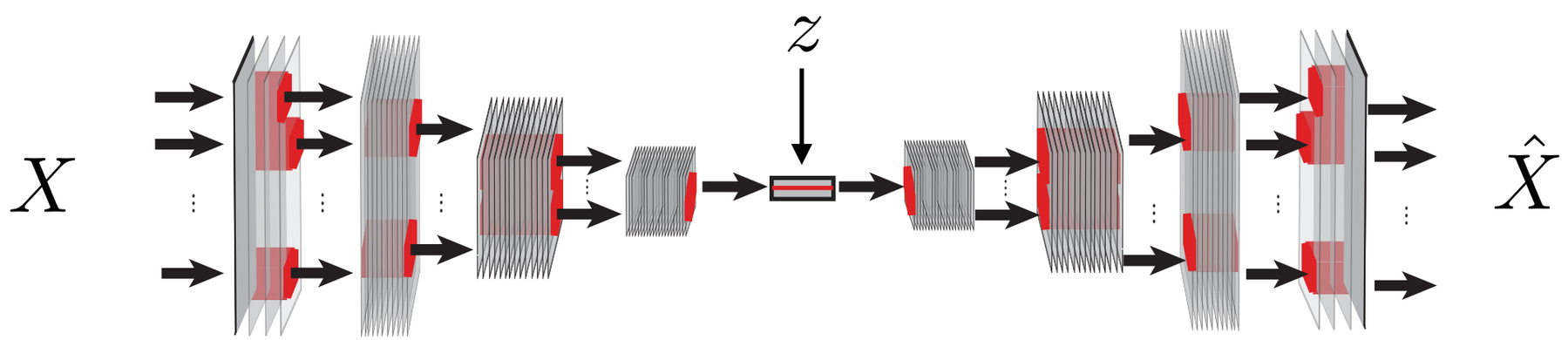


$$L(x) = |x - O(H(x))|^2 + \lambda \cdot |H(x)|$$

Sparse Coding Foldiak, Olshausen, mid 1990s

→neurons have to represent their environment, as efficiently as possible

Survey of Unsupervised Methods: **Variational Autoencoders**



encoder decoder

want to minimize

$$-log(p(\hat{x})) \leq -\sum_z q(z|x) \log \frac{p(z,x)}{q(z|x)}$$

Bayes rule

$$= -\sum_z q(z|x) \log \frac{p(x|z)p(z)}{q(z|x)}$$

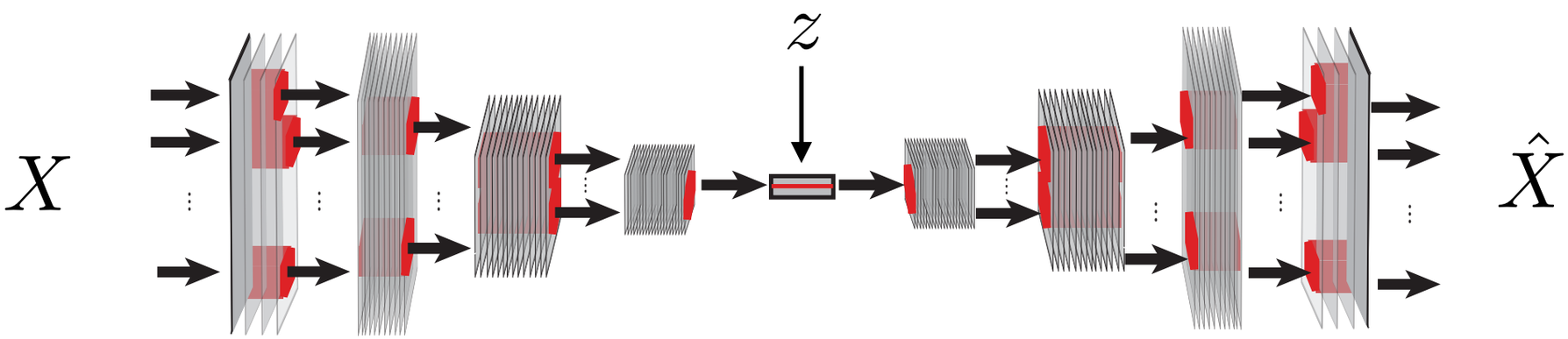
property of logarithms

$$= -\sum_z q(z|x) \log \frac{p(z)}{q(z|x)} - \sum_z q(z|x) \log(p(x|z))$$

definition of "expectation" and KL divergence

$$= -E_z[\log p(x|z)] + KL(q(z|x)||p(z))$$

Survey of Unsupervised Methods: **Variational Autoencoders**



z = ("identity", "gender", "age", "expression")

FaceApp

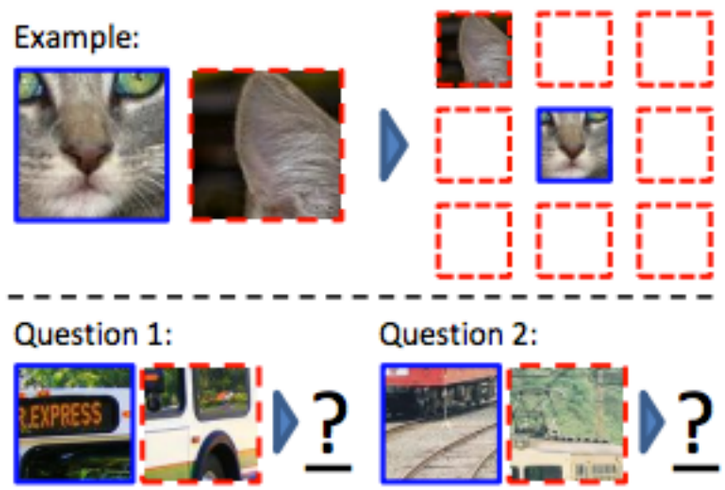
just need dataset varying with the four variables, and the decision to use one uniform and and three gaussian knobs … automatically discovers them

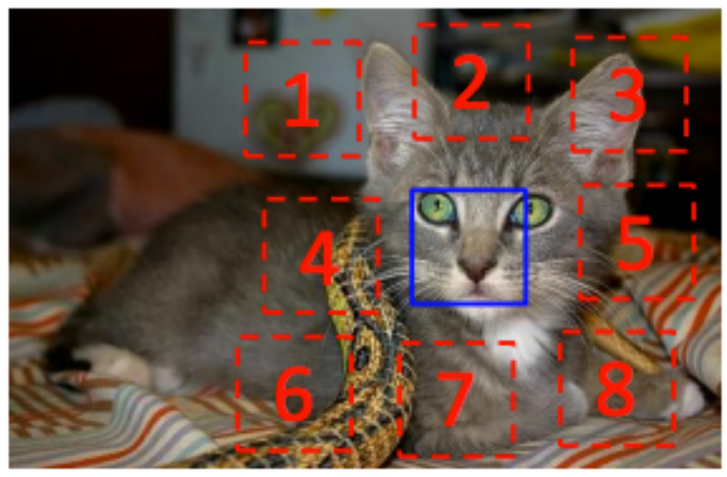
Survey of Unsupervised Methods: Context Prediction

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Abhinav Gupta, Alexei A. Efros

(Submitted on 19 May 2015 (v1), last revised 16 Jan 2016 (this version, v3))





X = (,); Y = 3

This is a discrete classification task.

VOC-2007 Test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
DPM-v5[17]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
[8] w/o context	52.6	52.6	19.2	25.4	18.7	47.3	56.9	42.1	16.6	41.4	41.9	27.7	47.9	51.5	29.9	20.0	41.1	36.4	48.6	53.2	38.5
Regionlets[58]	54.2	52.0	20.3	24.0	20.1	55.5	68.7	42.6	19.2	44.2	49.1	26.6	57.0	54.5	43.4	16.4	36.6	37.7	59.4	52.3	41.7
Scratch-R-CNN[2]	49.9	60.6	24.7	23.7	20.3	52.5	64.8	32.9	20.4	43.5	34.2	29.9	49.0	60.4	47.5	28.0	42.3	28.6	51.2	50.0	40.7
Scratch-Ours	52.6	60.5	23.8	24.3	18.1	50.6	65.9	29.2	19.5	43.5	35.2	27.6	46.5	59.4	46.5	25.6	42.4	23.5	50.0	50.6	39.8
Ours-projection	58.4	62.8	33.5	27.7	24.4	58.5	68.5	41.2	26.3	49.5	42.6	37.3	55.7	62.5	49.4	29.0	47.5	28.4	54.7	56.8	45.7
Ours-color-dropping	60.5	66.5	29.6	28.5	26.3	56.1	70.4	44.8	24.6	45.5	45.4	35.1	52.2	60.2	50.0	28.1	46.7	42.6	54.8	58.6	46.3
Ours-Yahoo100m	56.2	63.9	29.8	27.8	23.9	57.4	69.8	35.6	23.7	47.4	43.0	29.5	52.9	62.0	48.7	28.4	45.1	33.6	49.0	55.5	44.2
ImageNet-R-CNN[21]	64.2	69.7	50	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
K-means-rescale [31]	55.7	60.9	27.9	30.9	12.0	59.1	63.7	47.0	21.4	45.2	55.8	40.3	67.5	61.2	48.3	21.9	32.8	46.9	61.6	51.7	45.6
Ours-rescale [31]	61.9	63.3	35.8	32.6	17.2	68.0	67.9	54.8	29.6	52.4	62.9	51.3	67.1	64.3	50.5	24.4	43.7	54.9	67.1	52.7	51.1
ImageNet-rescale [31]	64.0	69.6	53.2	44.4	24.9	65.7	69.6	69.2	28.9	63.6	62.8	63.9	73.3	64.6	55.8	25.7	50.5	55.4	69.3	56.4	56.5
VGG-K-means-rescale	56.1	58.6	23.3	25.7	12.8	57.8	61.2	45.2	21.4	47.1	39.5	35.6	60.1	61.4	44.9	17.3	37.7	33.2	57.9	51.2	42.4
VGG-Ours-rescale	71.1	72.4	54.1	48.2	29.9	75.2	78.0	71.9	38.3	60.5	62.3	68.1	74.3	74.2	64.8	32.6	56.5	66.4	74.0	60.3	61.7
VGG-ImageNet-rescale	76.6	79.6	68.5	57.4	40.8	79.9	78.4	85.4	41.7	77.0	69.3	80.1	78.6	74.6	70.1	37.5	66.0	67.5	77.4	64.9	68.6

Table 1. Mean Average Precision on VOC-2007.

Survey of Unsupervised Methods: Context Prediction

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Abhinav Gupta, Alexei A. Efros

(Submitted on 19 May 2015 (v1), last revised 16 Jan 2016 (this version, v3))

Context Encoders: Feature Learning by Inpainting

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros

(Submitted on 25 Apr 2016 (v1), last revised 21 Nov 2016 (this version, v2))

(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

Survey of Unsupervised Methods: Context Prediction

Unsupervised Visual Representation Learning by Context Prediction

Carl Doersch, Abhinav Gupta, Alexei A. Efros

(Submitted on 19 May 2015 (v1), last revised 16 Jan 2016 (this version, v3))

Context Encoders: Feature Learning by Inpainting

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, Alexei A. Efros

(Submitted on 25 Apr 2016 (v1), last revised 21 Nov 2016 (this version, v2))

Learning Features by Watching Objects Move

Deepak Pathak[1,2,*], Ross Girshick[1], Piotr Dollár[1], Trevor Darrell[2], and Bharath Hariharan[1]

[1]Facebook AI Research (FAIR)
[2]University of California, Berkeley

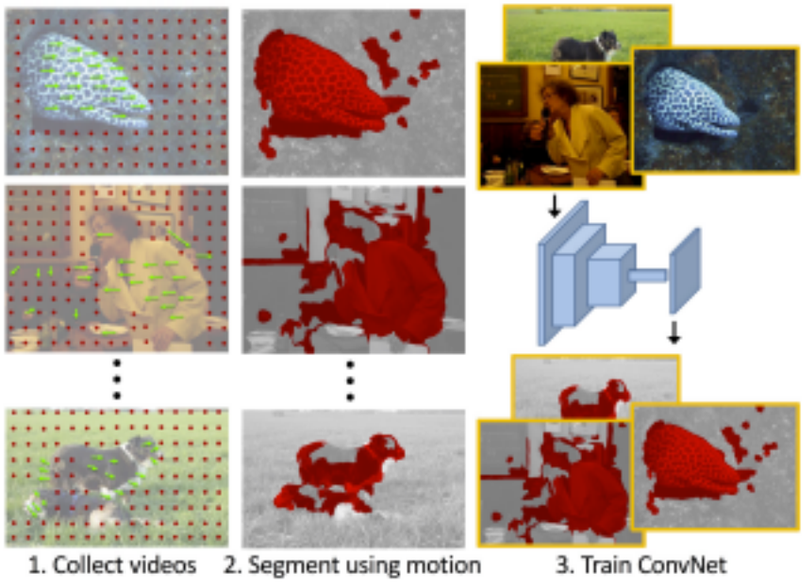


Figure 2. Overview of our approach. We use motion cues to segment objects in videos *without any supervision*. We then train a ConvNet to predict these segmentations from *static frames*, *i.e.* without any motion cues. We then transfer the learned representation to other recognition tasks.

Survey of Unsupervised Methods: Colorization

Colorful Image Colorization

Richard Zhang, Phillip Isola, Alexei A. Efros
{rich.zhang,isola,efros}@eecs.berkeley.edu

University of California, Berkeley

Fig. 1. Example input grayscale photos and output colorizations from our algorithm. These examples are cases where our model works especially well. Please visit `http://richzhang.github.io/colorization/` to see the full range of results and to try our model and code. Best viewed in color (obviously).

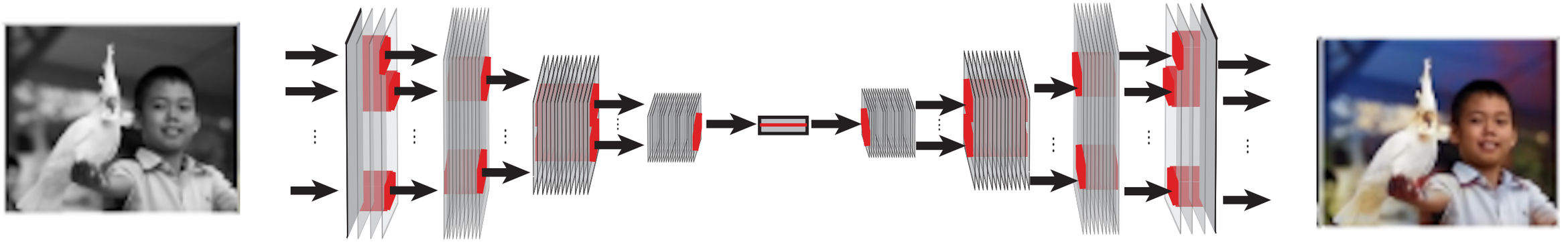
Survey of Unsupervised Methods: Colorization

Colorful Image Colorization

Richard Zhang, Phillip Isola, Alexei A. Efros
{rich.zhang,isola,efros}@eecs.berkeley.edu

University of California, Berkeley

Fig. 1. Example input grayscale photos and output colorizations from our algorithm. These examples are cases where our model works especially well. Please visit `http://richzhang.github.io/colorization/` to see the full range of results and to try our model and code. Best viewed in color (obviously).



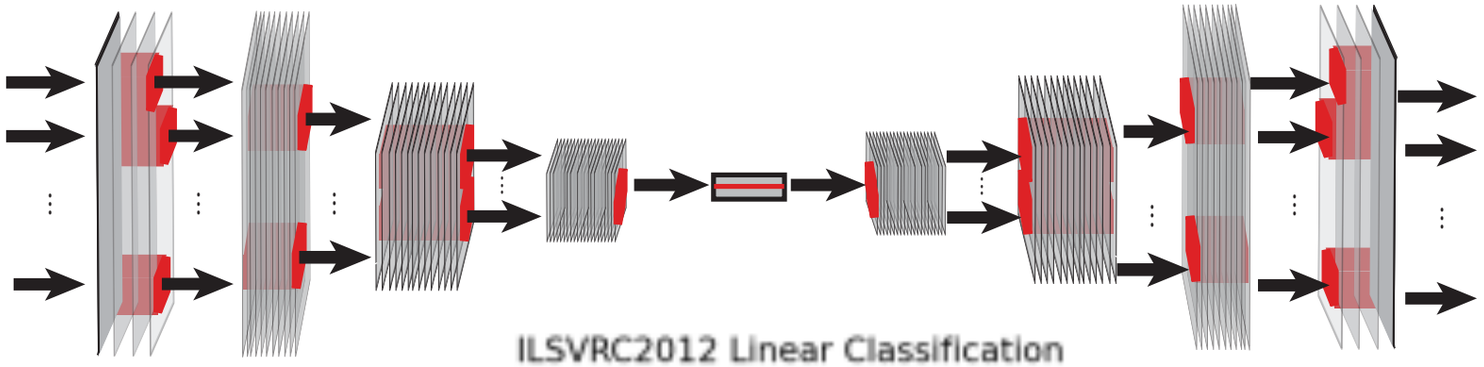
Survey of Unsupervised Methods: Colorization

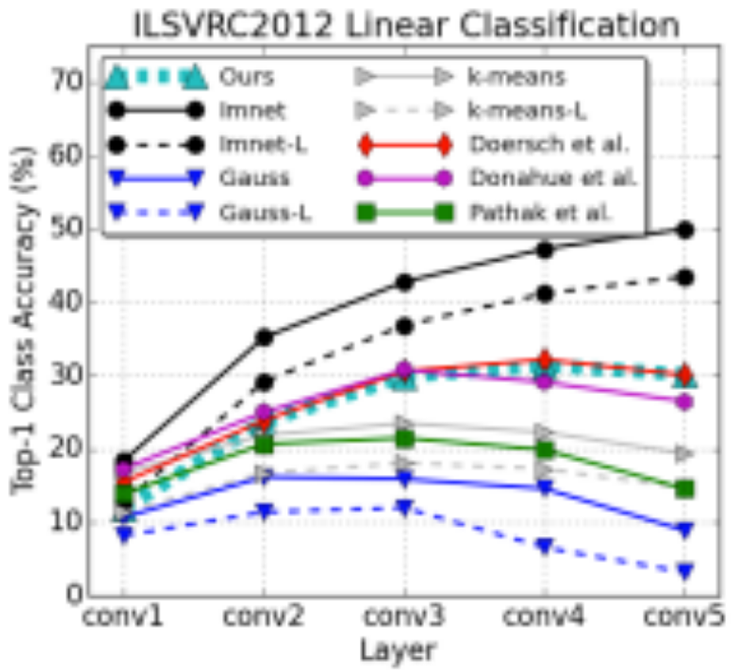
Colorful Image Colorization

Richard Zhang, Phillip Isola, Alexei A. Efros
{rich.zhang,isola,efros}@eecs.berkeley.edu

University of California, Berkeley

Fig. 1. Example input grayscale photos and output colorizations from our algorithm. These examples are cases where our model works especially well. Please visit `http://richzhang.github.io/colorization/` to see the full range of results and to try our model and code. Best viewed in color (obviously).





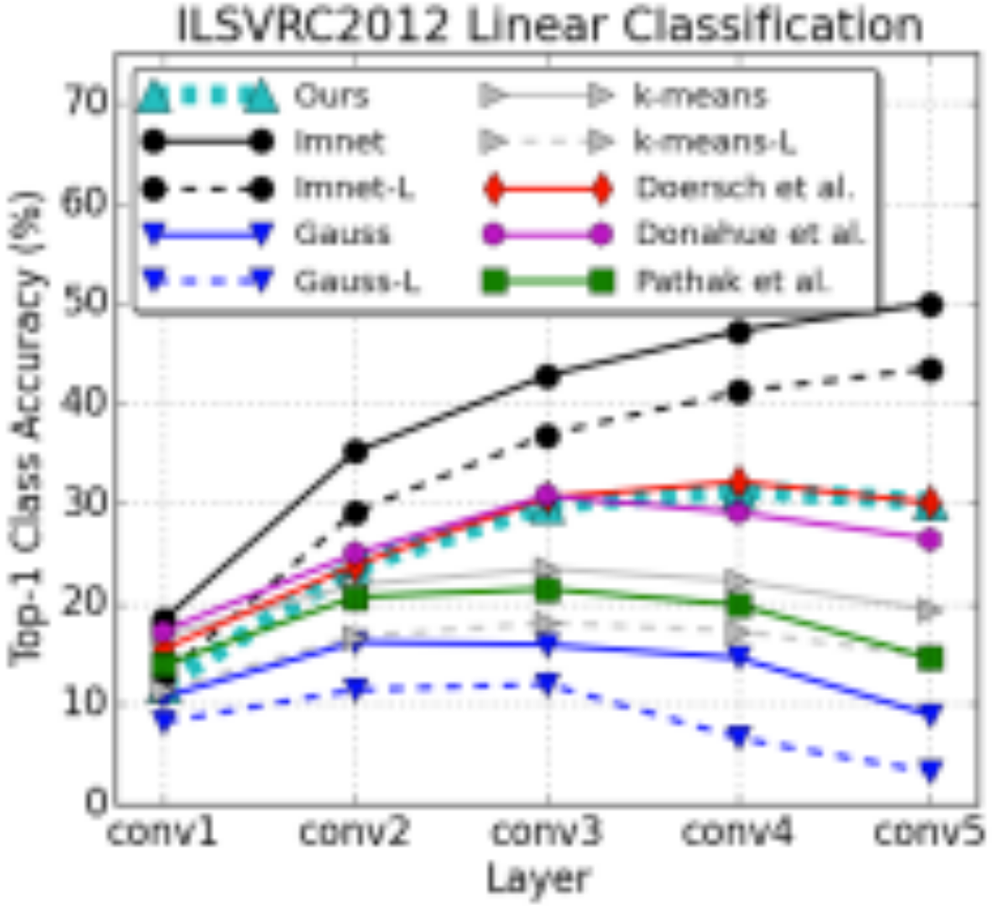
Survey of Unsupervised Methods: Colorization

Colorful Image Colorization

Richard Zhang, Phillip Isola, Alexei A. Efros
{rich.zhang,isola,efros}@eecs.berkeley.edu

University of California, Berkeley

Fig. 1. Example input grayscale photos and output colorizations from our algorithm. These examples are cases where our model works especially well. Please visit `http://richzhang.github.io/colorization/` to see the full range of results and to try our model and code. Best viewed in color (obviously).



Survey of Unsupervised Methods: Rotation

UNSUPERVISED REPRESENTATION LEARNING BY PREDICTING IMAGE ROTATIONS

Anonymous authors
Paper under double-blind review

Figure 1: Images rotated by random multiples of 90 degrees (e.g., 0, 90, 180, or 270 degrees). The core intuition of our self-supervised feature learning approach is that if someone is not aware of the concepts of the objects depicted in the images, he cannot recognize the rotation that was applied to them.

Survey of Unsupervised Methods: Rotation

UNSUPERVISED REPRESENTATION LEARNING BY PREDICTING IMAGE ROTATIONS

Anonymous authors
Paper under double-blind review

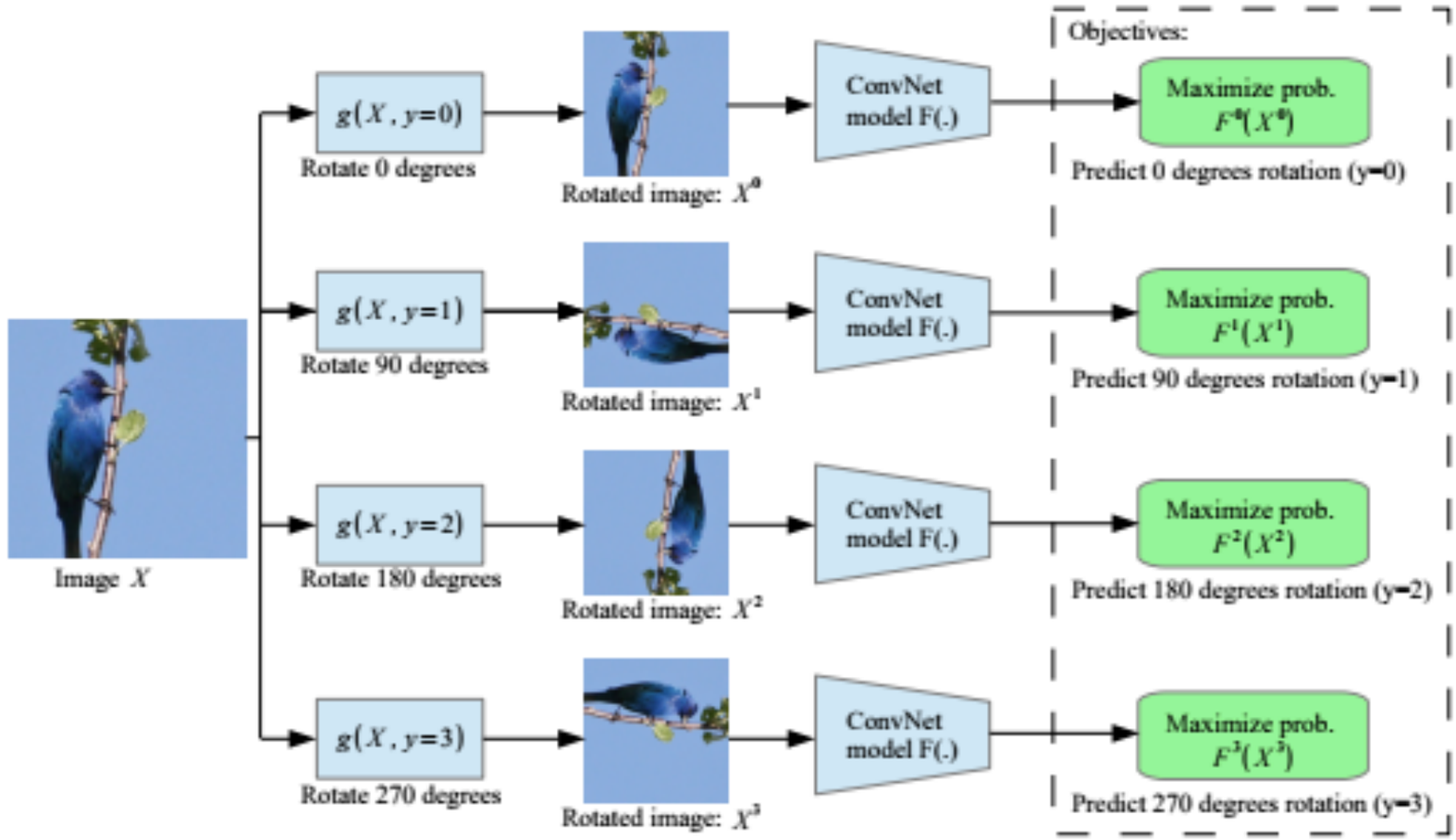


Figure 2: Illustration of the self-supervised task that we propose for semantic feature learning. Given four possible geometric transformations, the 0, 90, 180, and 270 degrees rotations, we train a ConvNet model $F(.)$ to recognize the rotation that is applied to the image that it gets as input. $F^y(X^{y^*})$ is the probability of rotation transformation y predicted by model $F(.)$ when it gets as input an image that has been transformed by the rotation transformation y^*.

Survey of Unsupervised Methods: Rotation

Colorization, jigsaw, rotation, &c approaches are of this form

$$X \mapsto f_\theta(X)$$

Goal: from $f_\theta(X)$, predict θ

ex: $f_\theta(X)$ = rotation by θ

ex: $f_\theta(X)$ = masking (e.g. jigsaw) at some location(s)

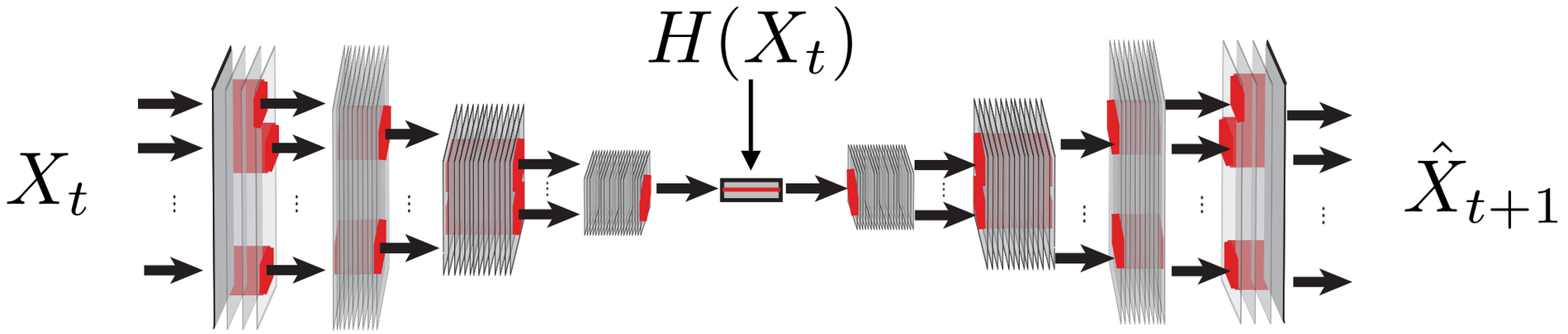
ex: $f_\theta(X)$ = grayscaling (no dependence on theta)

Key common feature of colorization, jigsaw, rotation, &c approaches: no dependence on $\boldsymbol{X}$ is allowed. Only $f_\theta(X)$ is given as input for figuring out θ.

. . . unlike auto-encoders. Giving $\boldsymbol{X}$ *makes the problem too easy.*

Auto-Encoding like methods: Predictive Coding

X_t X_{t+1}



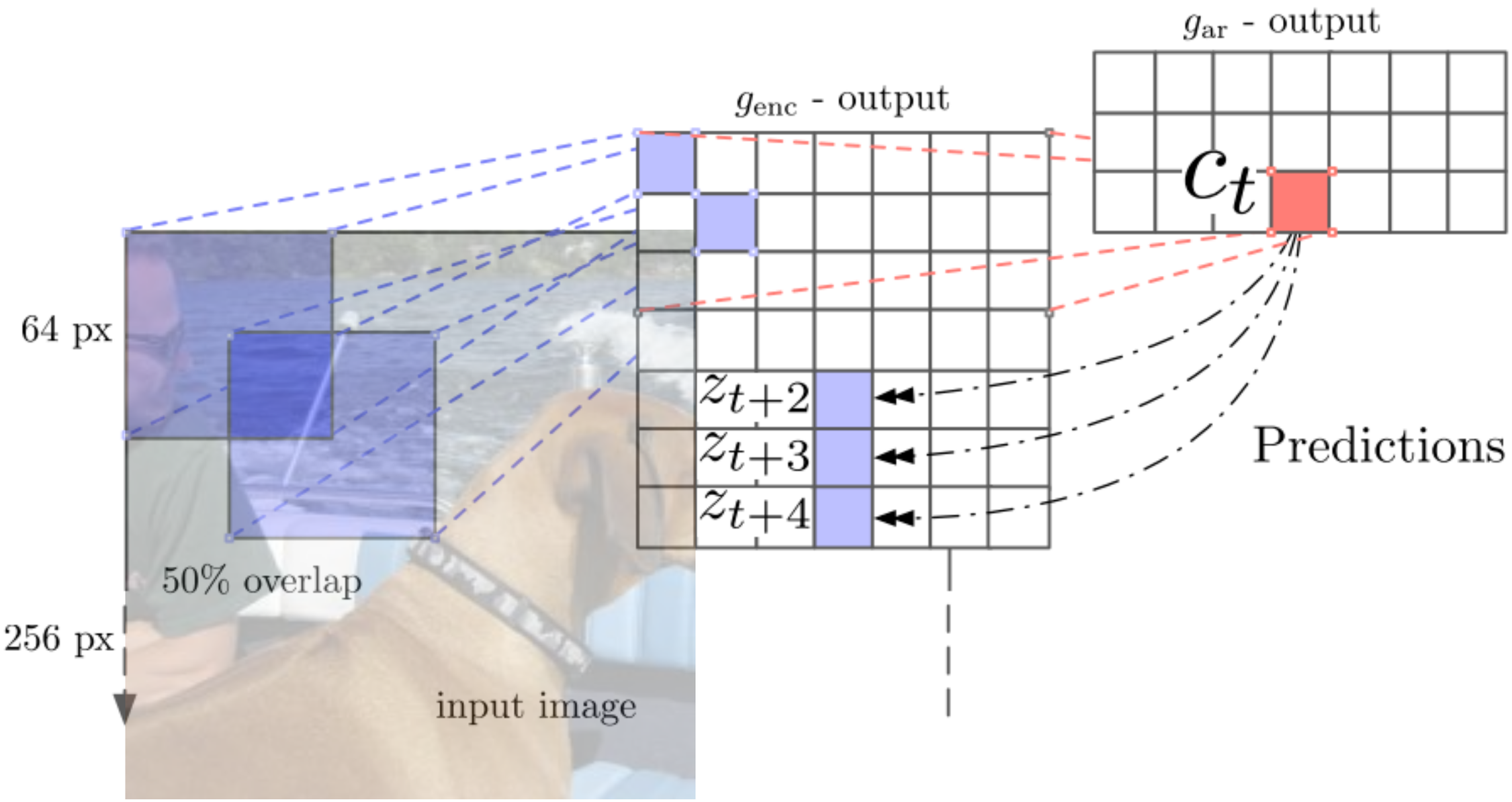
encoder decoder

$$\text{Loss} = \|X_{t+1} - \hat{X}_{t+1}\| + Penalty(H(X_t))$$

reconstruction complexity metric

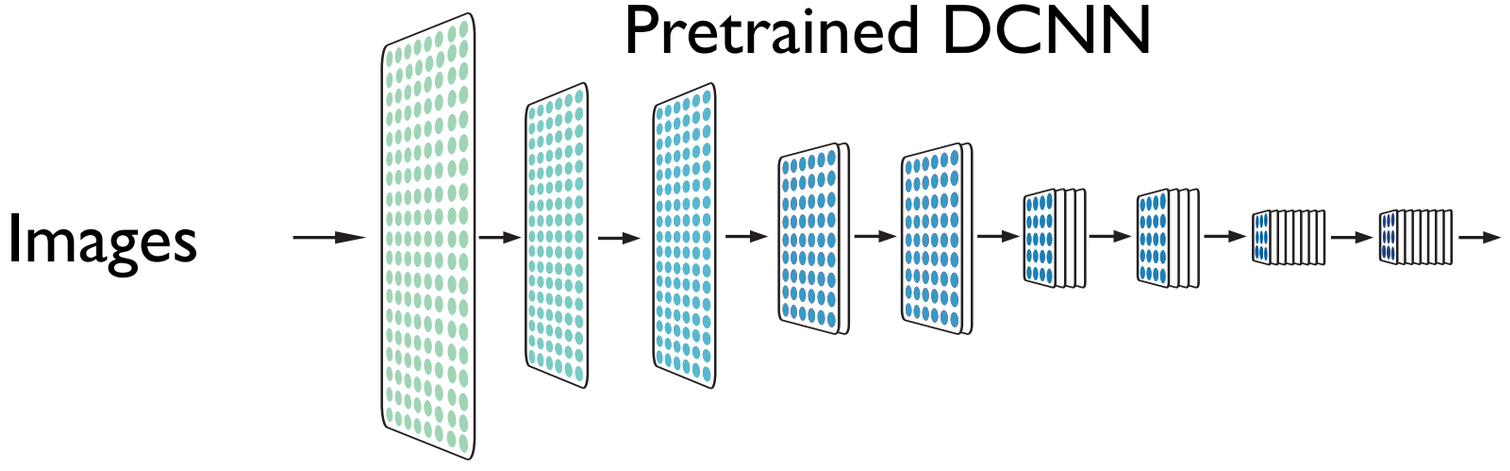
Lotter et al. 2017

Auto-Encoding like methods:
Contrastive Predictive Coding

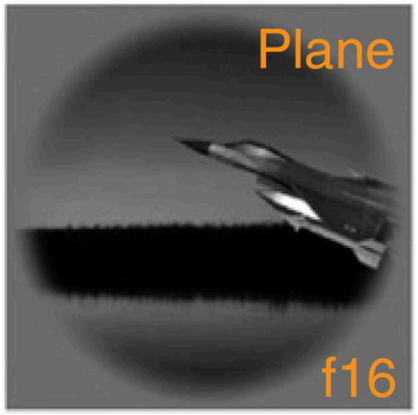


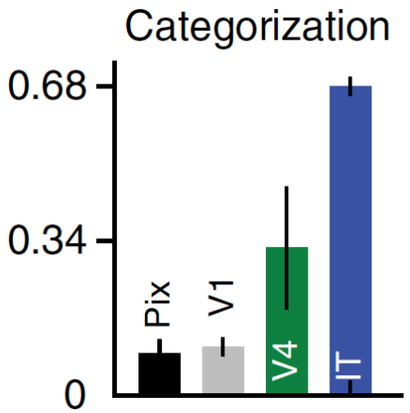
van den Oord et al. 2018

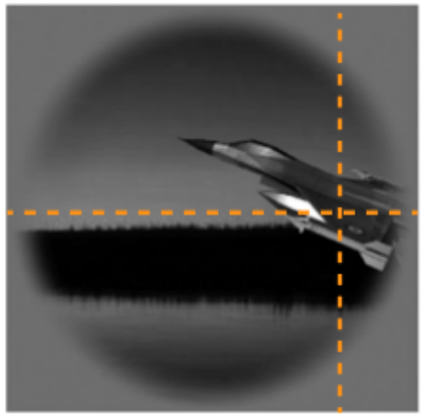
Downstream Task Performance

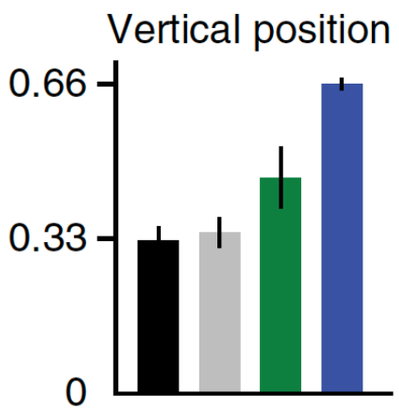


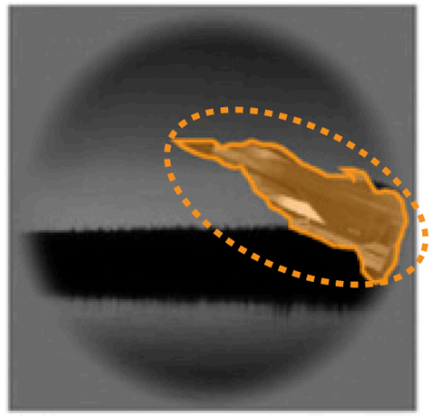
Hong et al. 2016

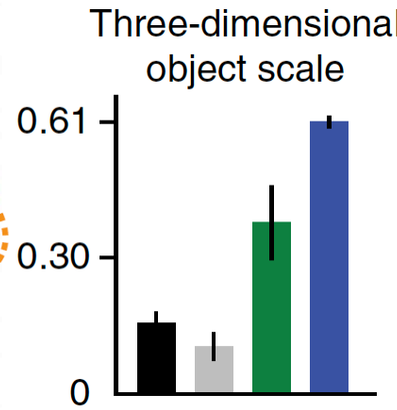






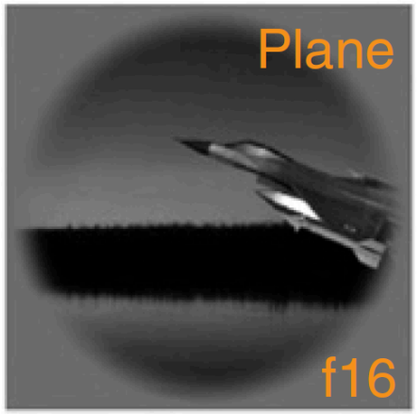


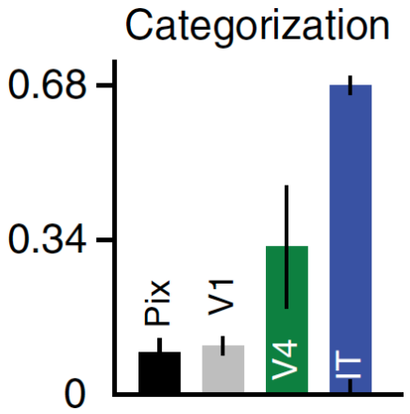


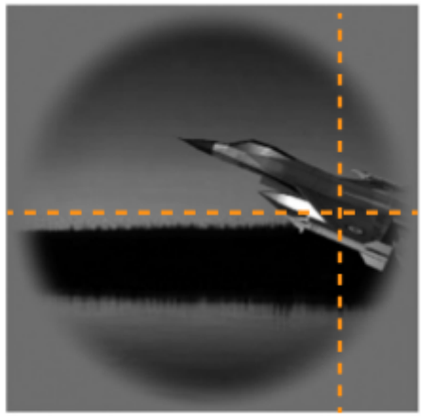


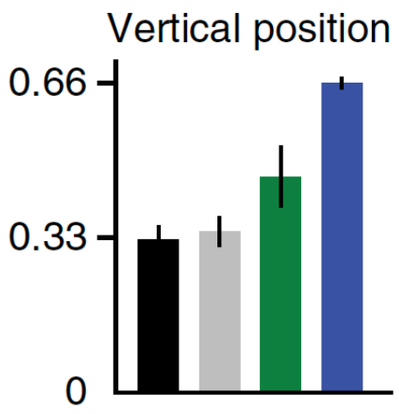
AutoEncoder, PredNet, and CPC show relatively poor downstream performance

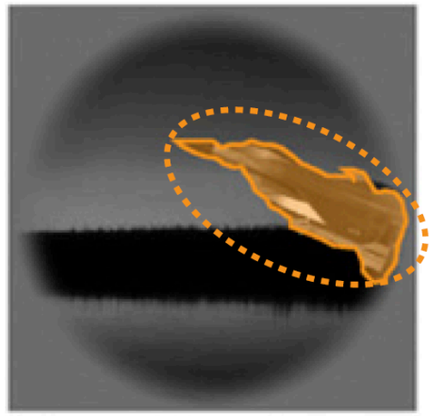
Hong et al. 2016

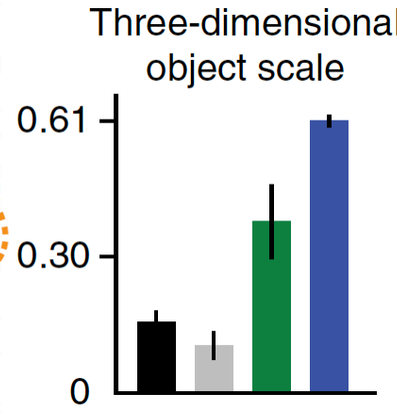


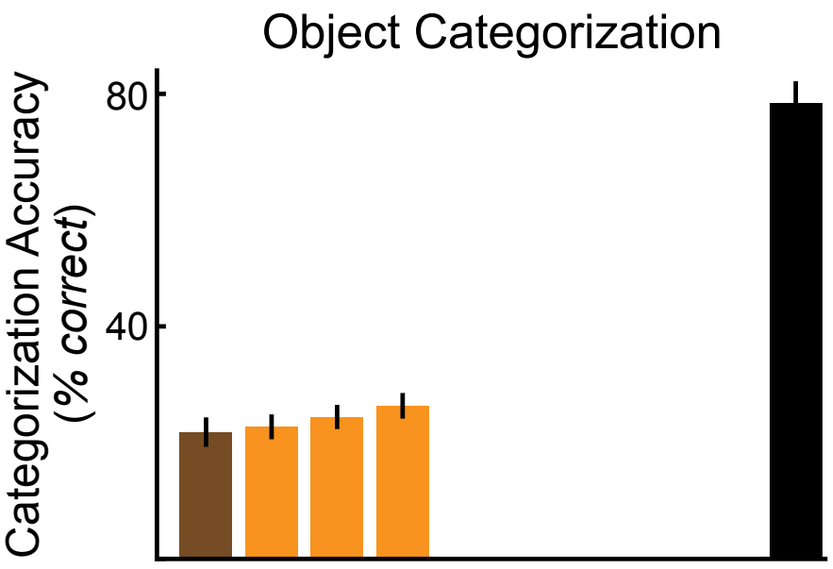


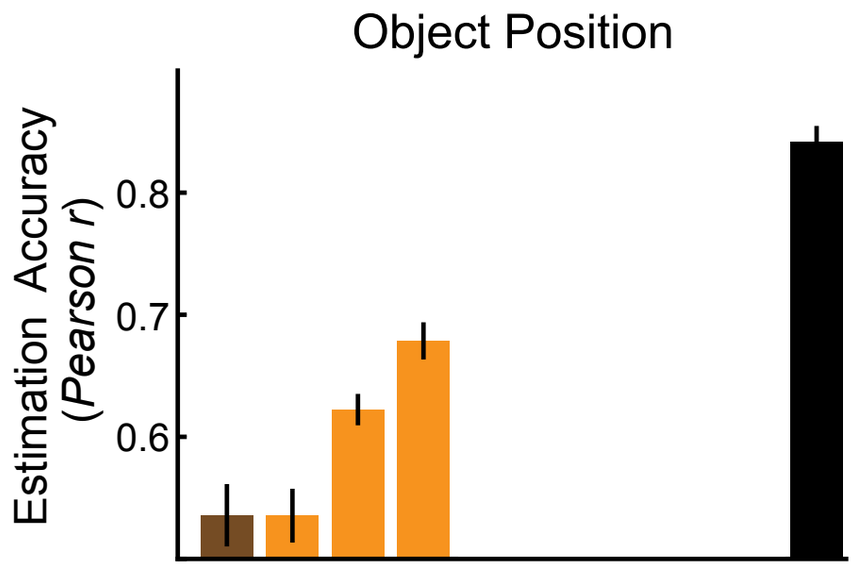


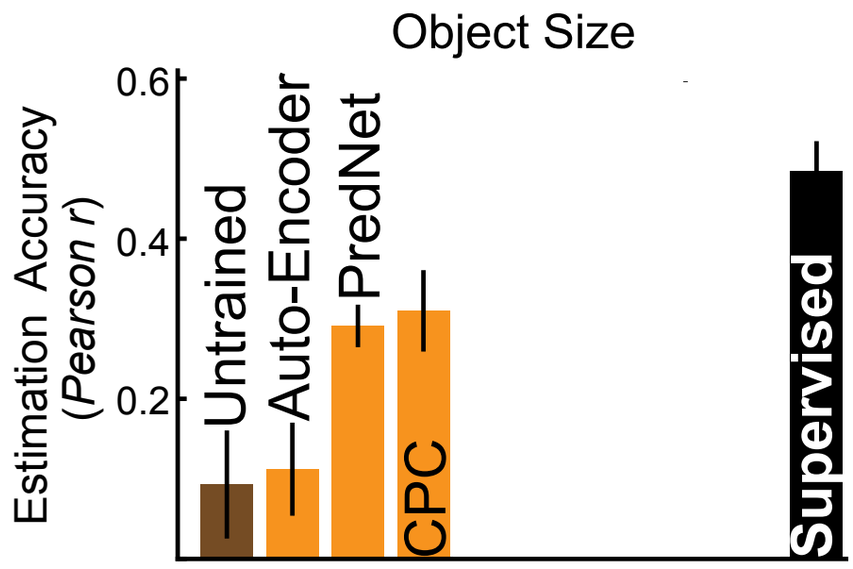




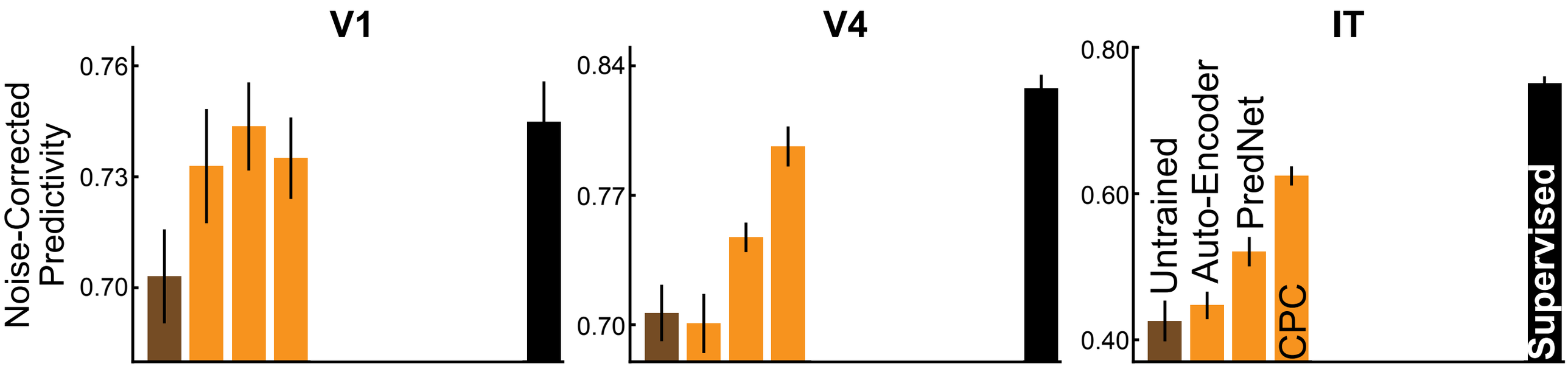








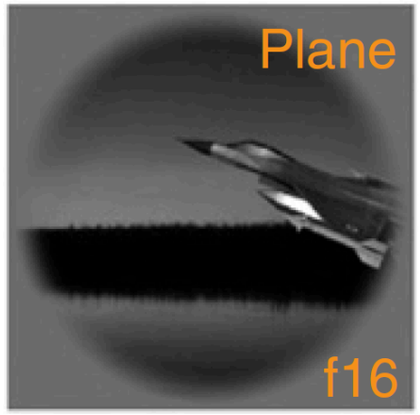
AutoEncoder is only good for V1
CPC is good for V1, not bad in V4 and IT

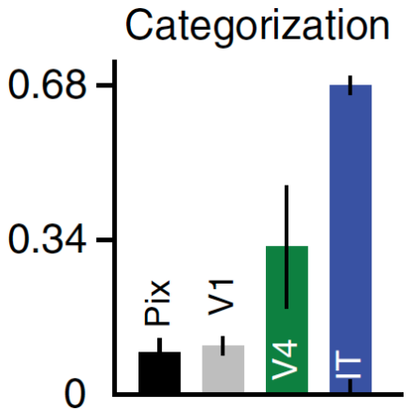


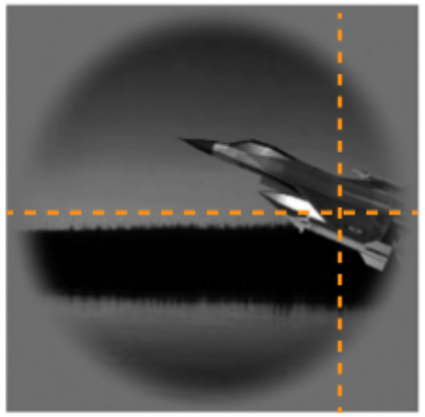
Best Pearson correlations across all layers are reported

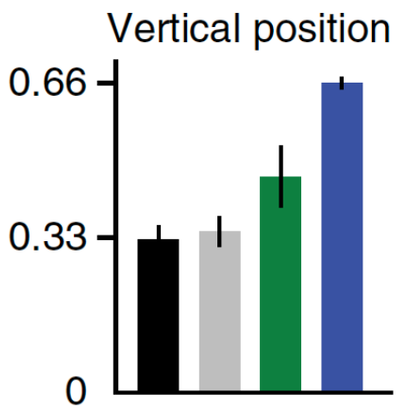
Self-supervised tasks show slightly better downstream performance

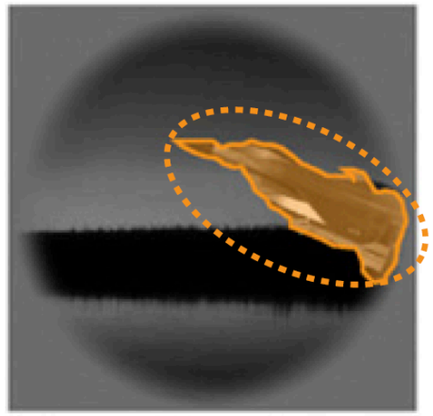
Hong et al. 2016

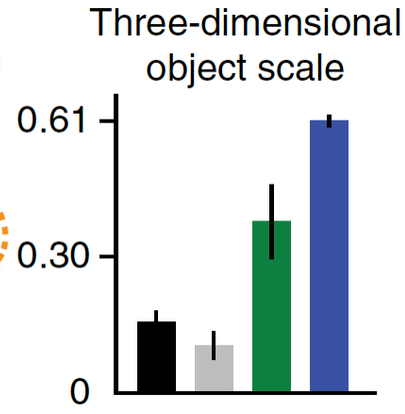


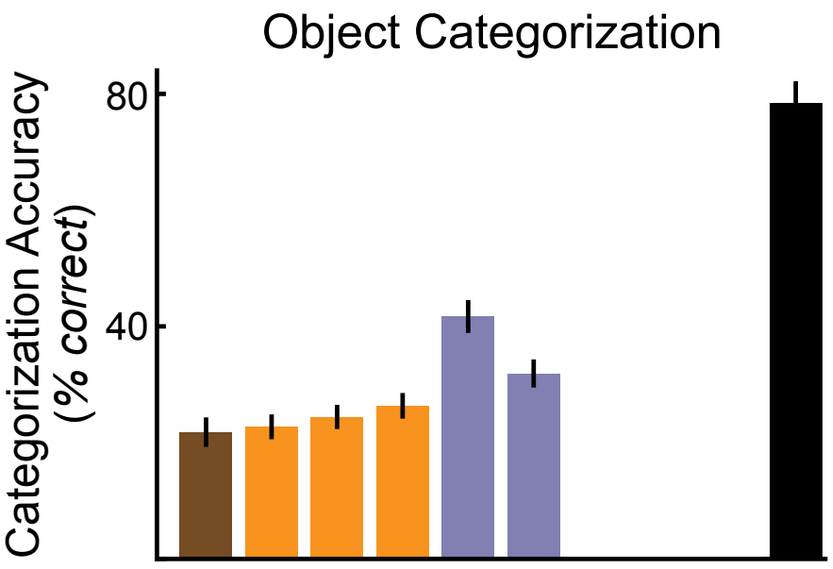


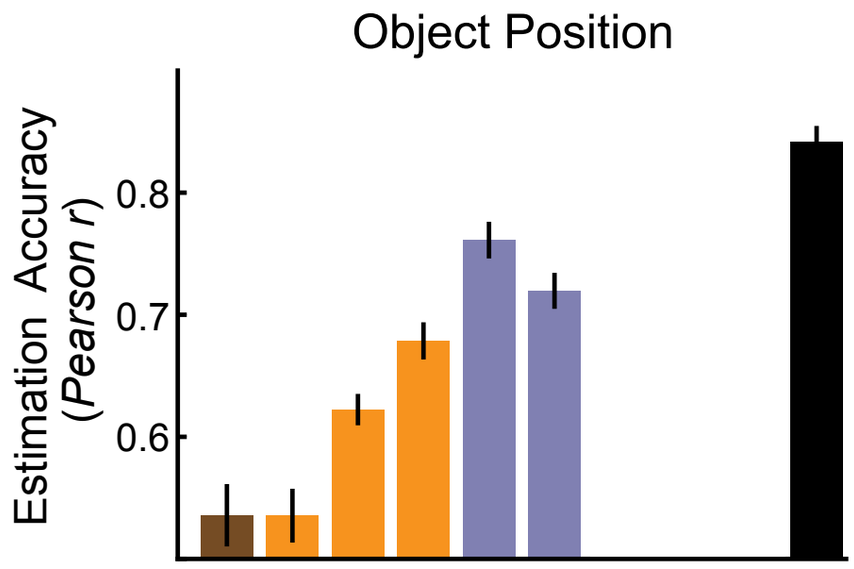


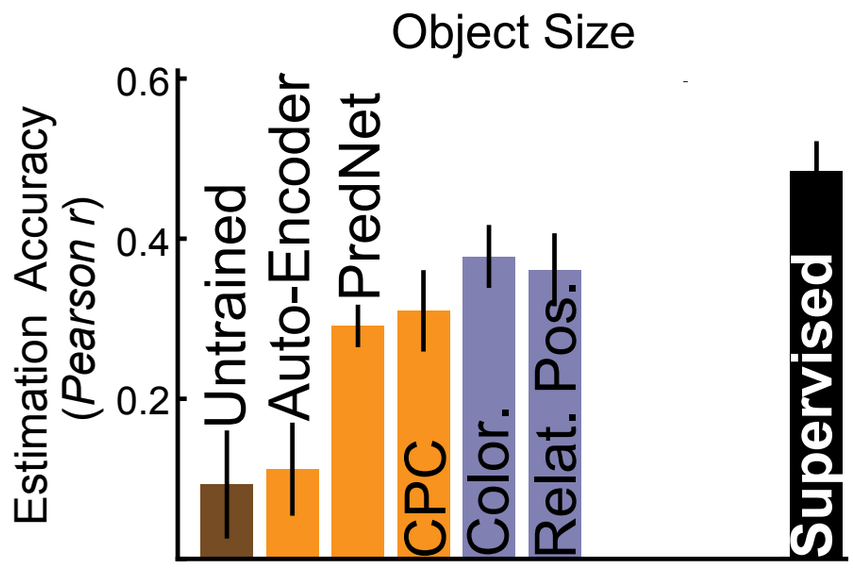




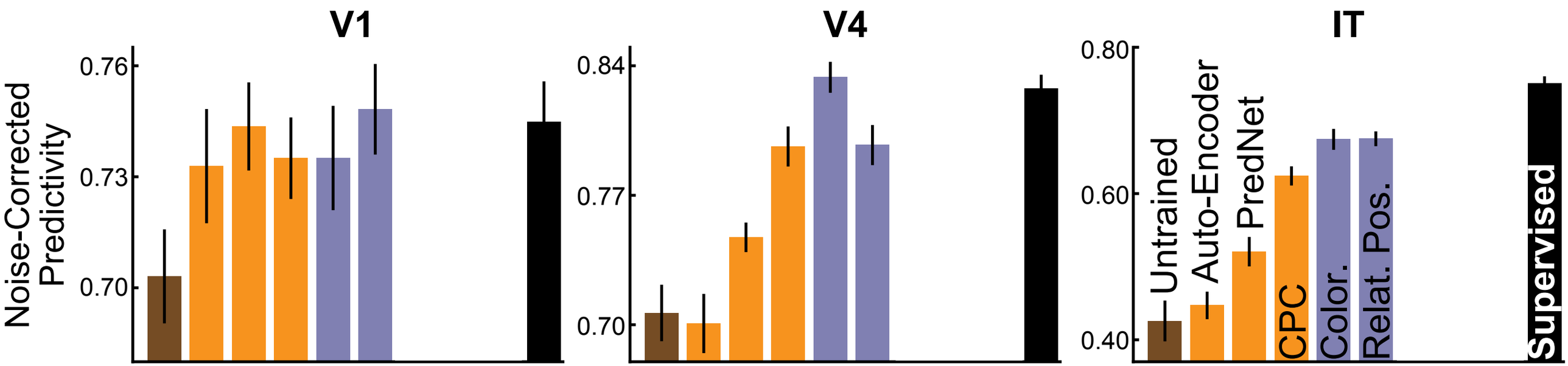






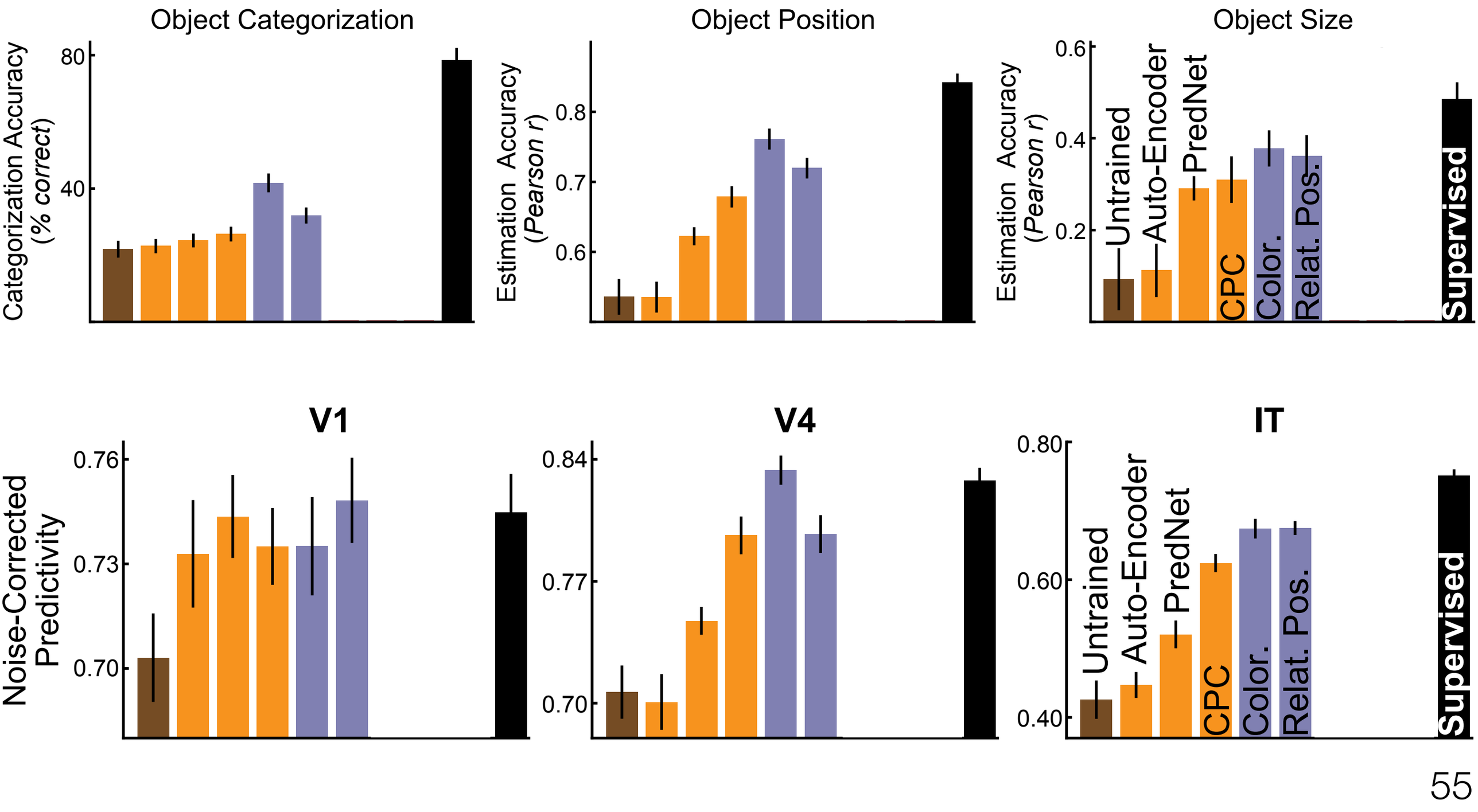


Self-supervised tasks show better V4 and IT neural predictivity



Best Pearson correlations across all layers are reported

Still, none of the algorithms show good task performance and IT predictivity.



Contrastive learning tasks

High-level idea of these methods: make the representations non-trivially robust to data augmentations

Wu et al. 2018

Contrastive learning tasks: Instance Recognition

Raw Input

Augmented Input

Wu et al. 2018

Contrastive learning tasks: Instance Recognition

Raw Input

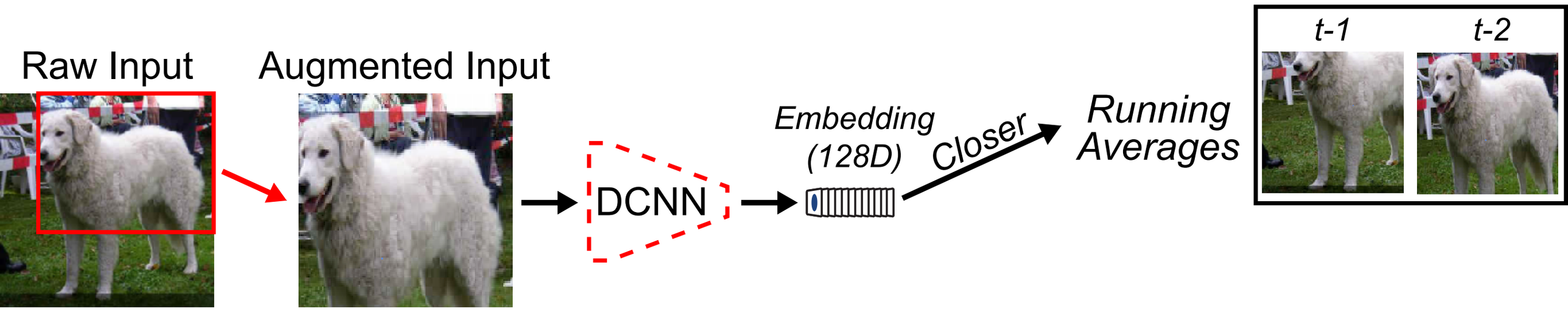
Augmented Input

DCNN

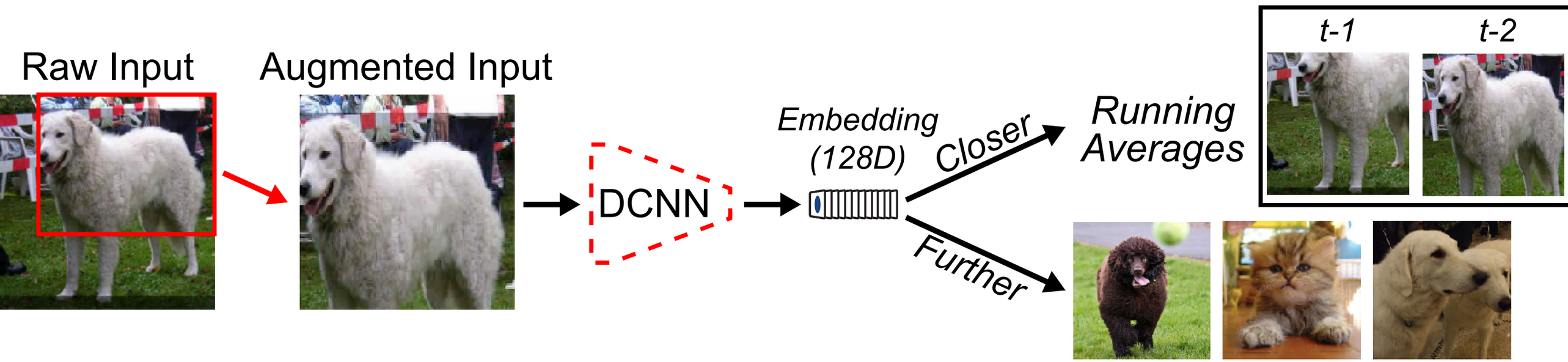
Embedding (128D)

Contrastive learning tasks: Instance Recognition

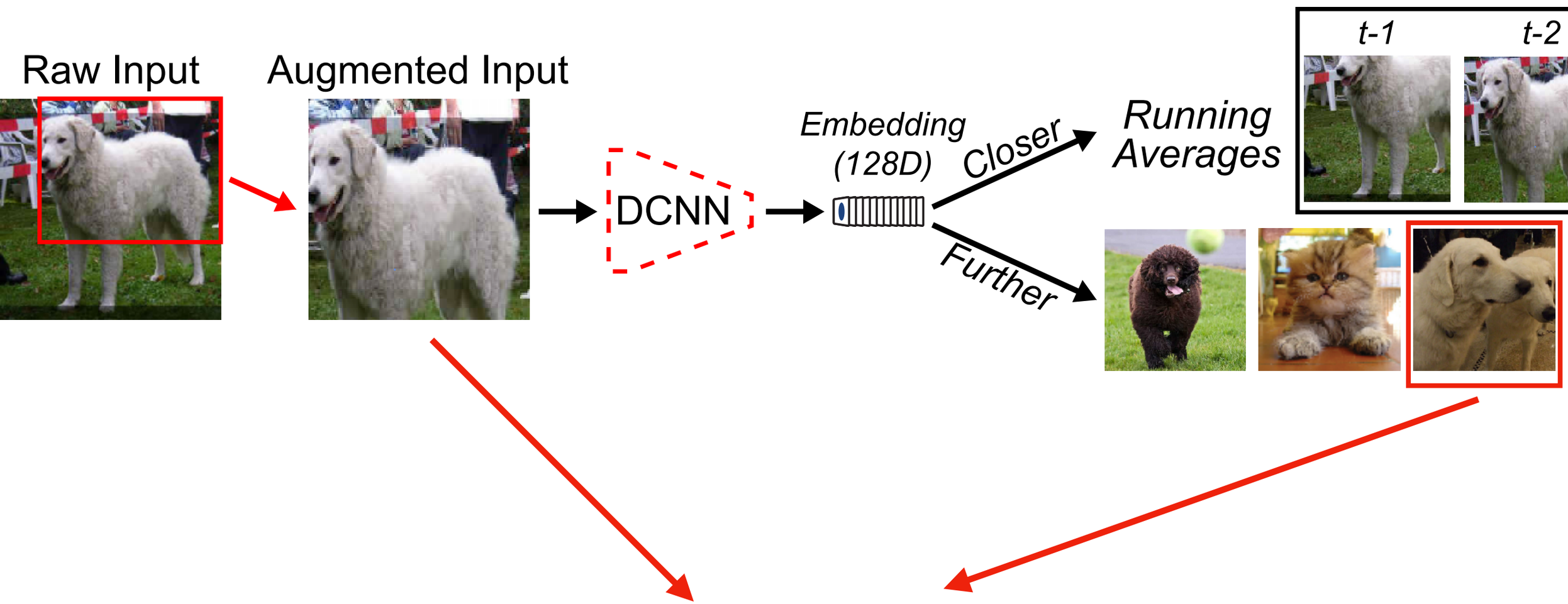
Robust Recognition to Data Augmentations



Contrastive learning tasks: Instance Recognition



Avoid Collapsing through Spreading across the Space



Why separating everything given that there are naturally examples within the same category?

Constrastive Embedding Models

Chengxu Zhuang

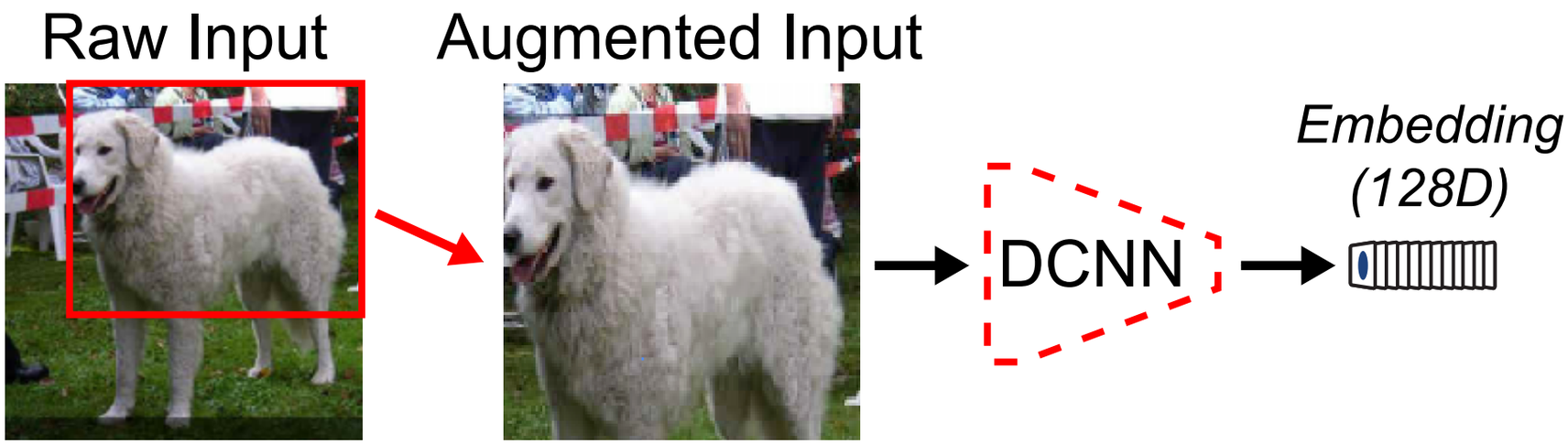
Zhuang et al. **Local Aggregation for Unsupervised Learning of Visual Embeddings.** *(ICCV 2019)*

Zhuang et al. **Local Label Propagation for Large-Scale Semi-Supervised Learning.** https://arxiv.org/abs/1905.11581

Zhuang et al. **Unsupervised Learning from Video with Deep Neural Embeddings.** *(CVPR 2020)*
https://arxiv.org/abs/1905.11954

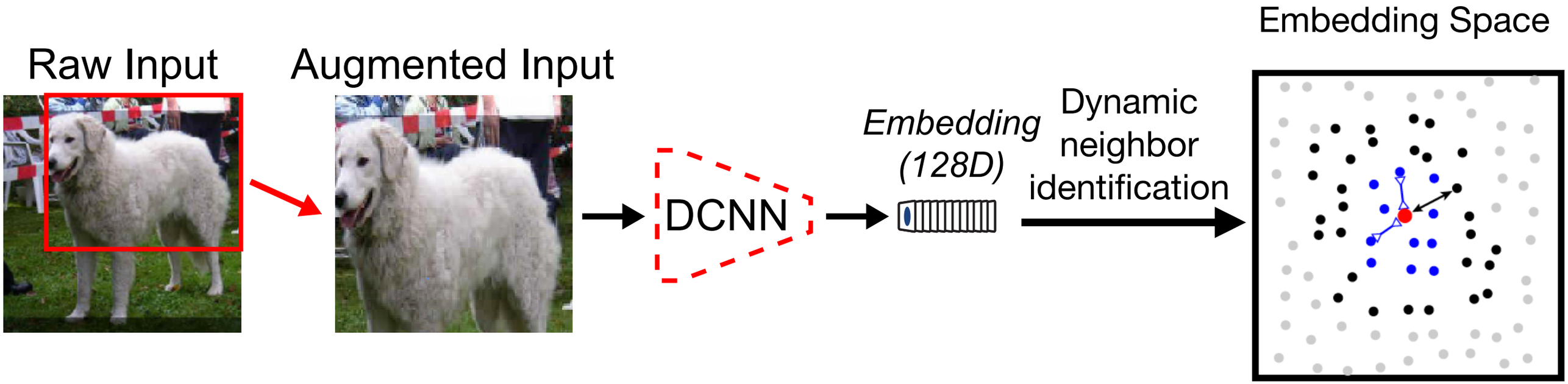
Zhuang et al. **Unsupervised Neural Network Models of the Ventral Visual Stream.** *(PNAS, 2021)*

Local Aggregation



Zhuang et al. 2019a

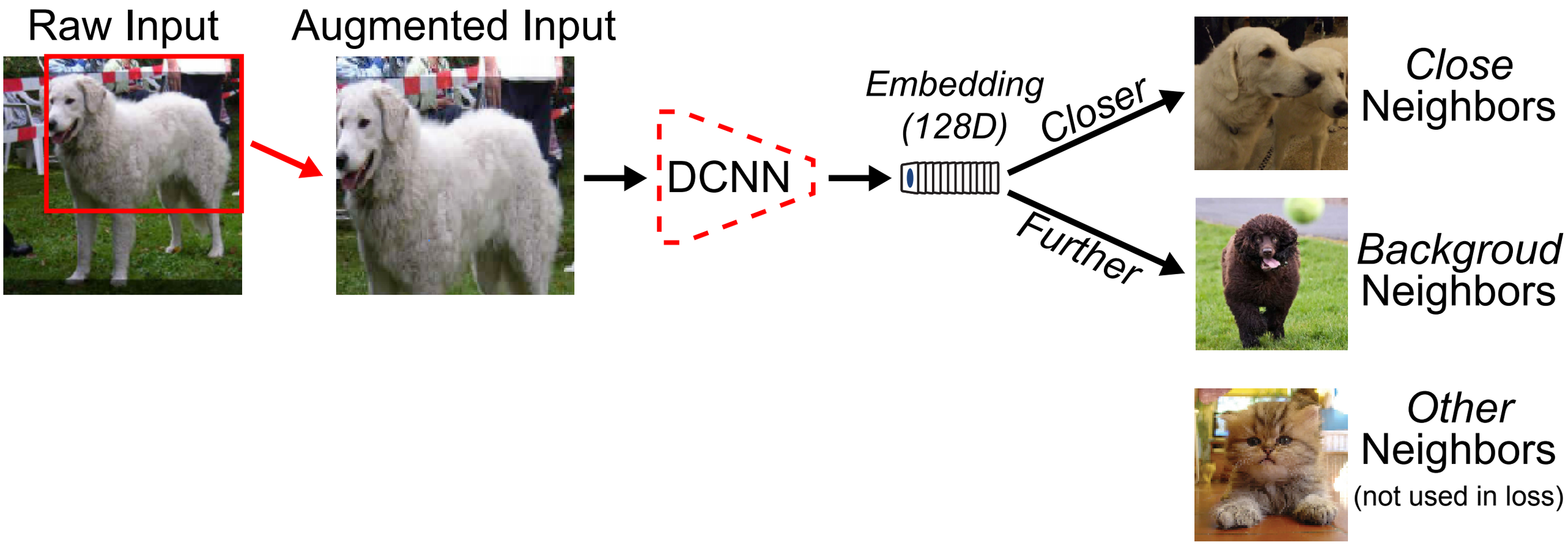
Local Aggregation



Dynamic neighbor identification in the embedding space for each image.

Zhuang et al. 2019a

Local Aggregation



Locally aggregate the close neighbors and the current image.

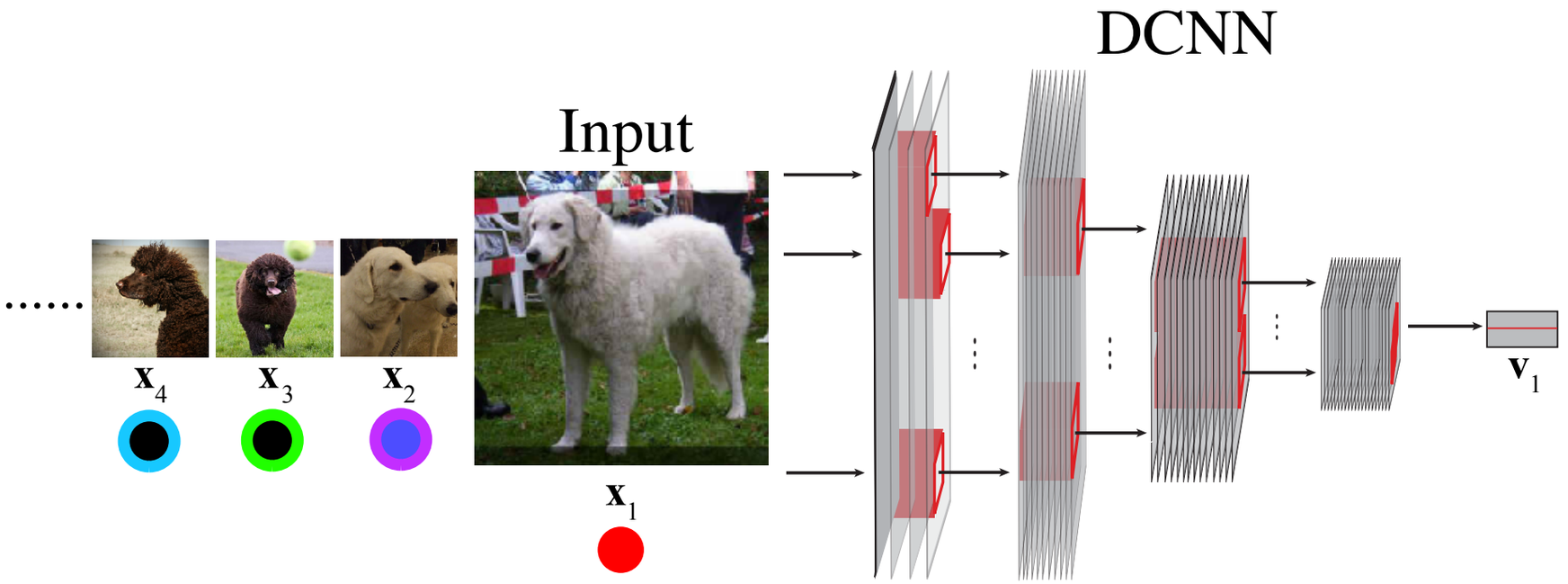
Zhuang et al. 2019a

Constrastive Embedding Models

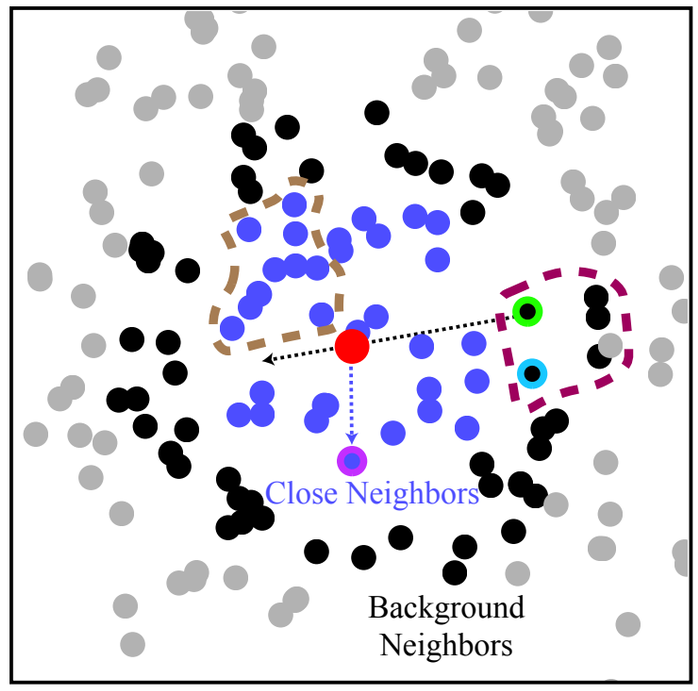
Family of new methods from unsupervised learning called

deep contrastive embeddings.

Chengxu Zhuang



Before training…



"neural code" space

Progressively modify network synapses to minimize:

$$L(\mathbf{C},\mathbf{B}) = -\log \frac{P(\mathbf{C}\cap\mathbf{B})}{P(\mathbf{B})}$$

eg. increase probability of being clustered together, if close in neural code

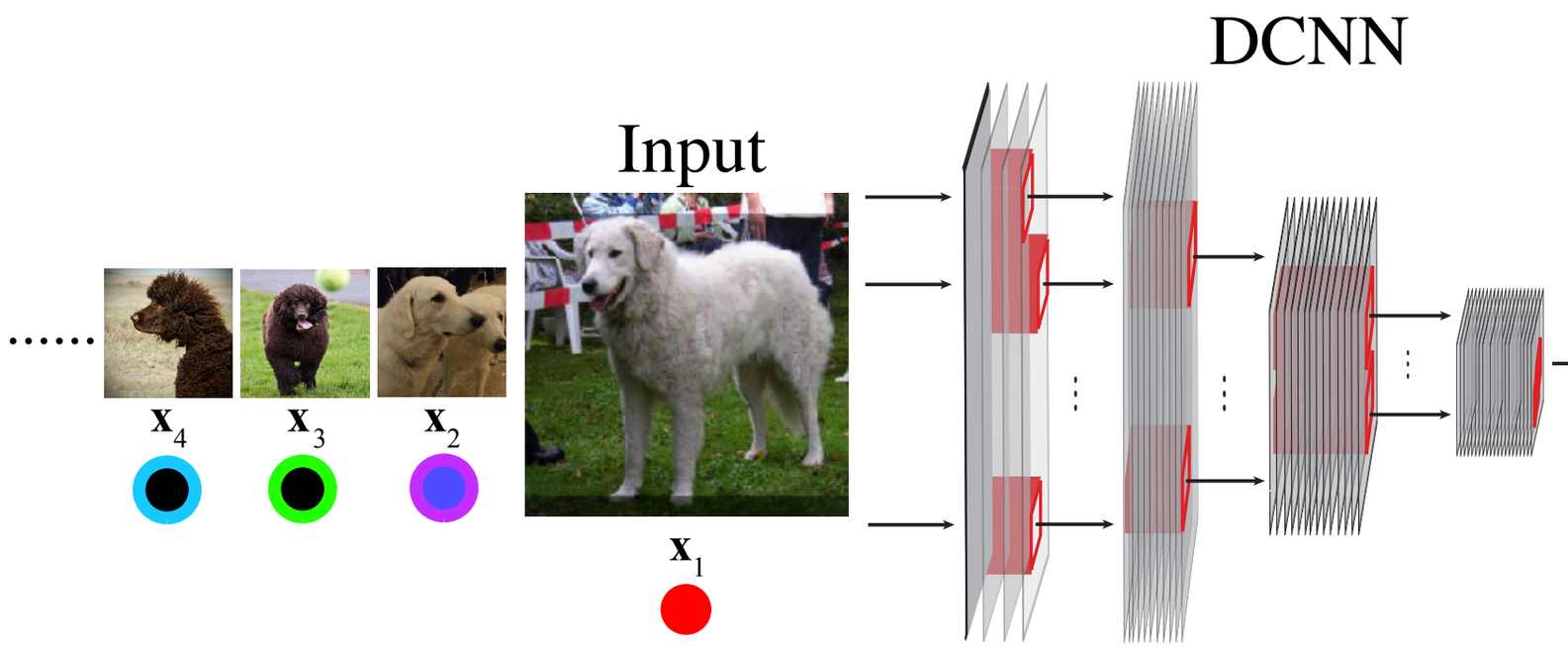
Zhuang et al. **Local Aggregation for Unsupervised Learning of Visual Embeddings.** *(ICCV 2019)*

Constrastive Embedding Models

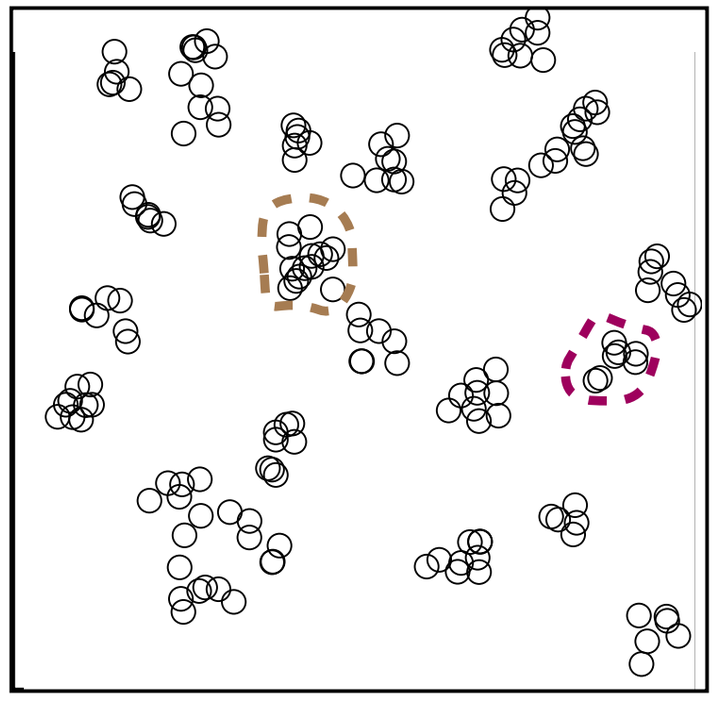
Family of new methods from unsupervised learning called

deep contrastive embeddings.

Chengxu Zhuang



… *after training*



"neural code" space

Progressively modify network synapses to minimize:

$$L(\mathbf{C}, \mathbf{B}) = -\log \frac{P(\mathbf{C} \cap \mathbf{B})}{P(\mathbf{B})}$$

eg. increase probability of being clustered together, if close in neural code

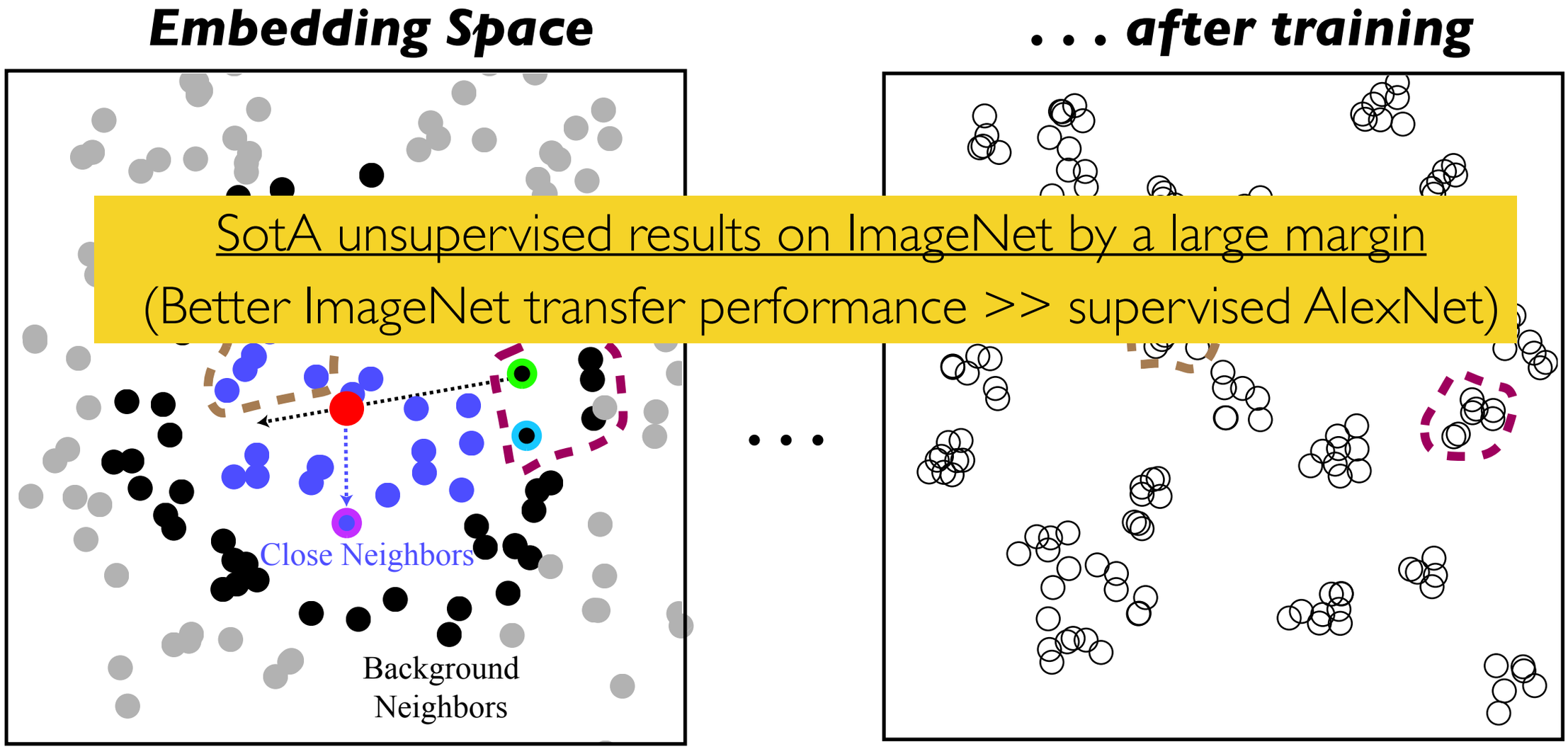
Zhuang et al. **Local Aggregation for Unsupervised Learning of Visual Embeddings.** *(ICCV 2019)*

New Unsupervised Method: **Local Aggregation**

We have achieved substantial boost above previous state-of-the-art using a method we call **Local Aggregation**.

Chengxu Zhuang

Zhuang et al. **Local Aggregation for Unsupervised Learning of Visual Embeddings.** *(ICCV 2019)*

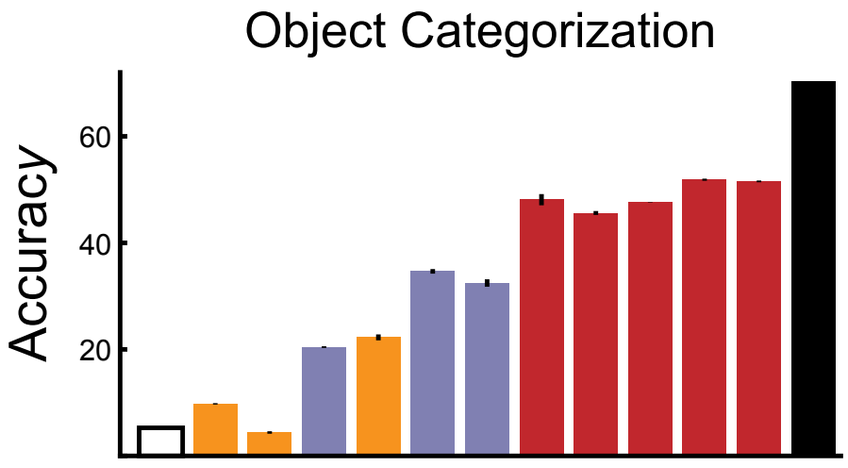


Allows similar points to move closer while pushing dissimilar points further away

New Unsupervised Method: **Local Aggregation**

Performance increases not just on object categorization but also many other visual tasks … suggesting general improvement in representation.

Chengxu Zhuang



Autoencoders

Missing-Data Tasks

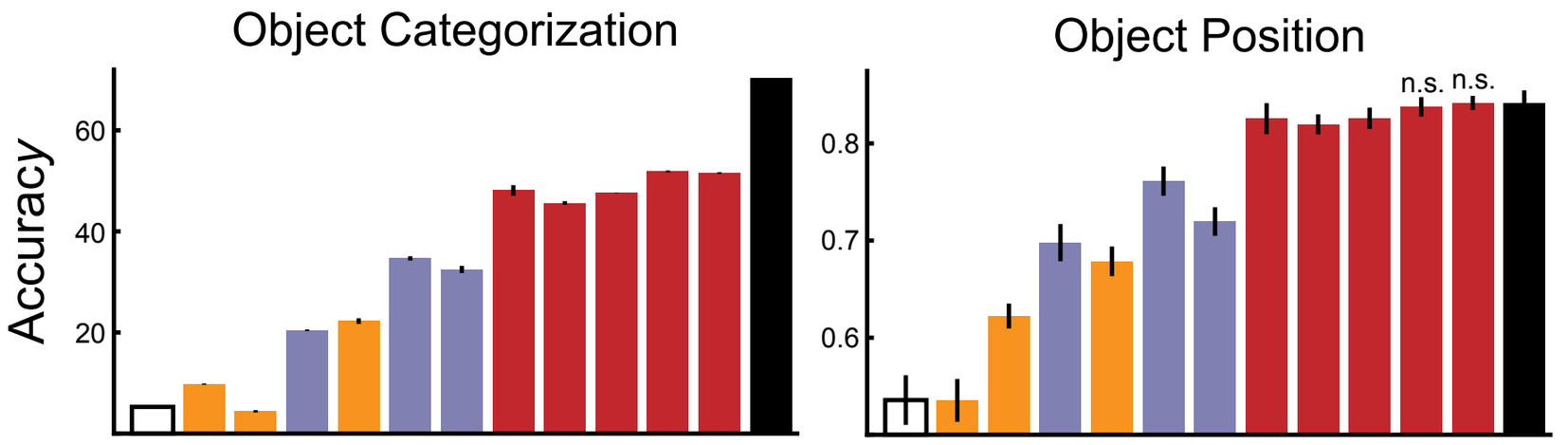
Deep Contrastive Embeddings

Zhuang et al. **Local Aggregation for Unsupervised Learning of Visual Embeddings.** *(ICCV 2019)*

New Unsupervised Method: **Local Aggregation**

Chengxu Zhuang

Performance increases not just on object categorization but also many other visual tasks … suggesting general improvement in representation.



Autoencoders

Missing-Data Tasks

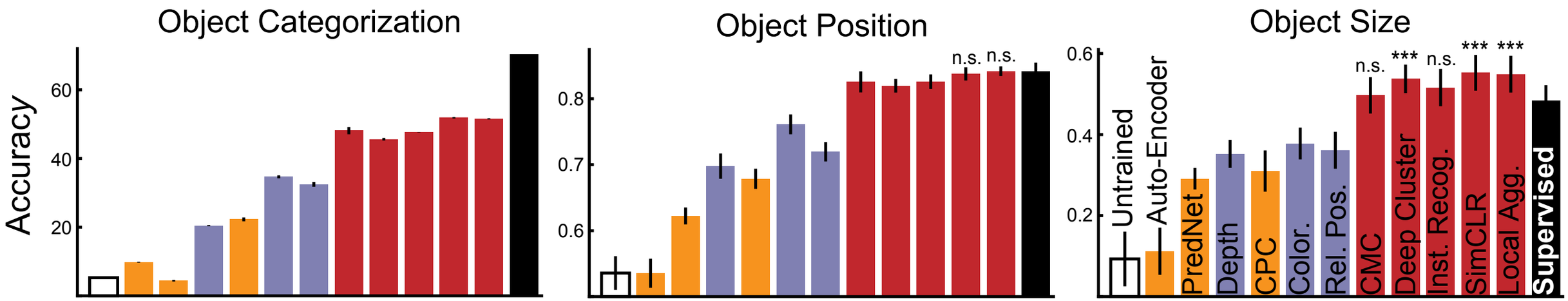
Deep Contrastive Embeddings

Zhuang et al. **Local Aggregation for Unsupervised Learning of Visual Embeddings.** *(ICCV 2019)*

New Unsupervised Method: **Local Aggregation**

Performance increases not just on object categorization but also many other visual tasks … suggesting general improvement in representation.

Chengxu Zhuang



Autoencoders

Missing-Data Tasks

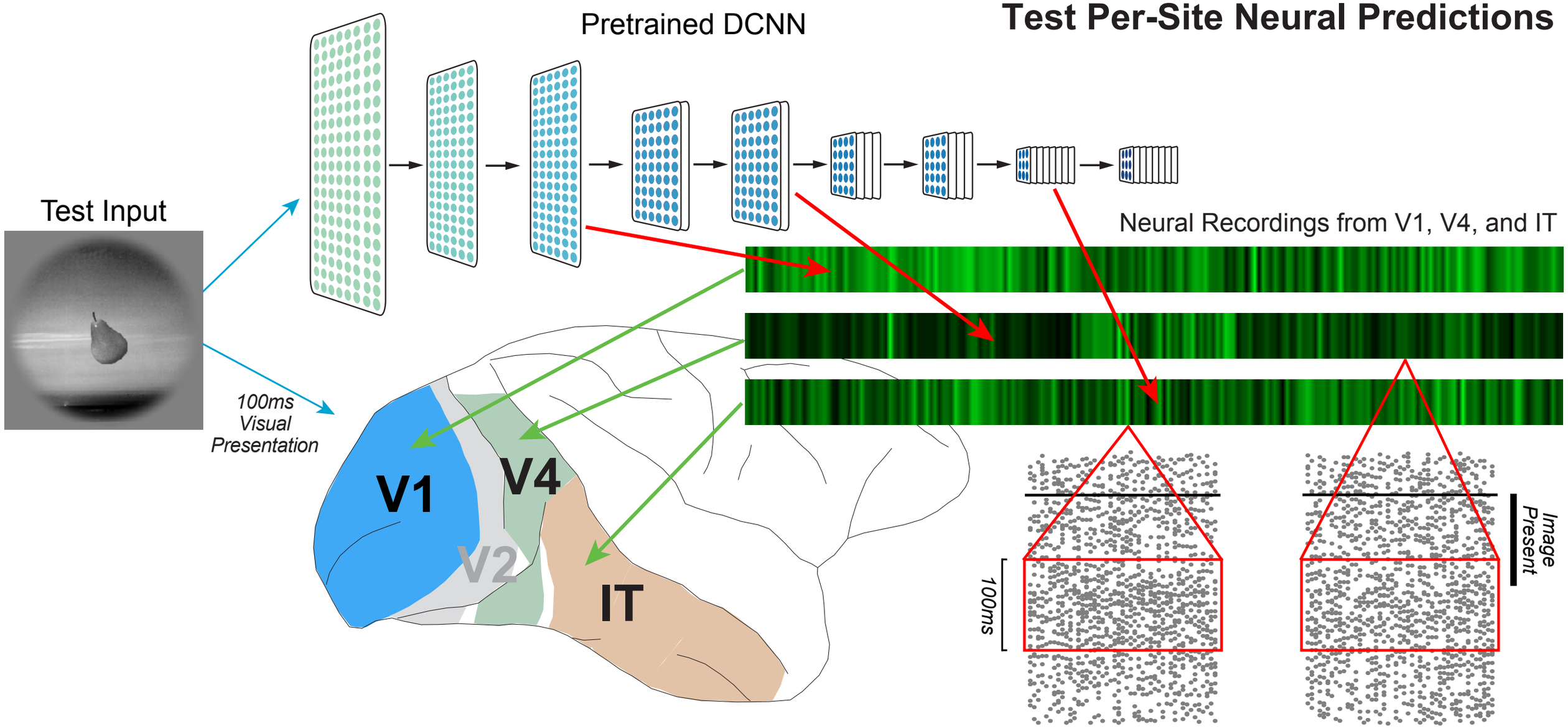
Deep Contrastive Embeddings

Zhuang et al. **Local Aggregation for Unsupervised Learning of Visual Embeddings.** *(ICCV 2019)*

Comparison to Neural Data

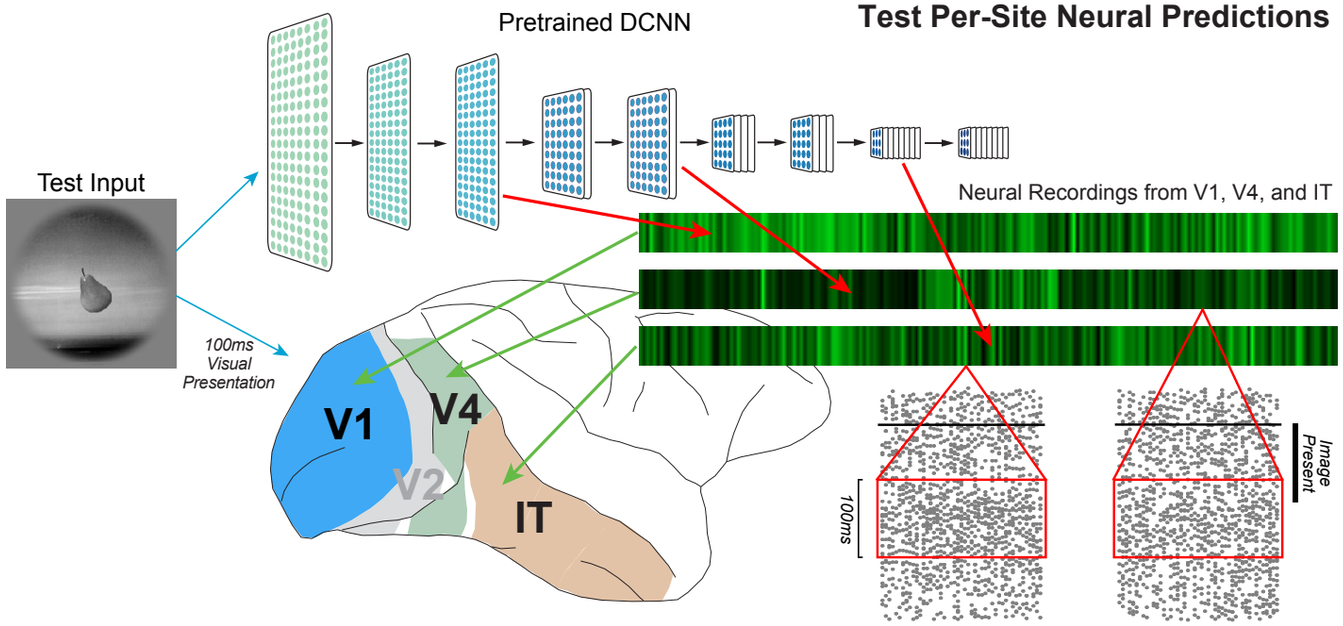
How well does it match neural data?

Chengxu Zhuang

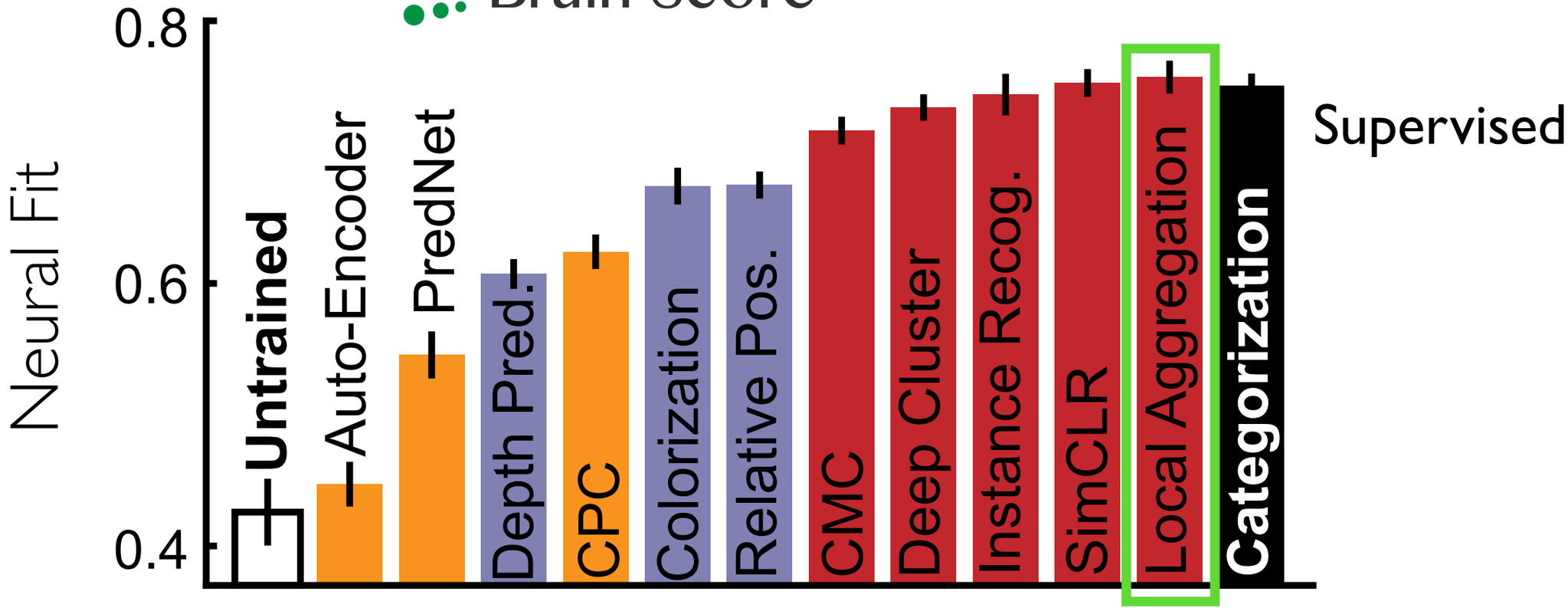


V1 data from Cadena et al. Deep convolutional models improve predictions of macaque **V1** responses to natural images *PLoS Comp. Bio.,* (2019)

V4 & IT data from Majaj et al. Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance *J. Neurosci.* (2015)



Brain-Score



Quantitatively accurate unsupervised model of a higher brain area.

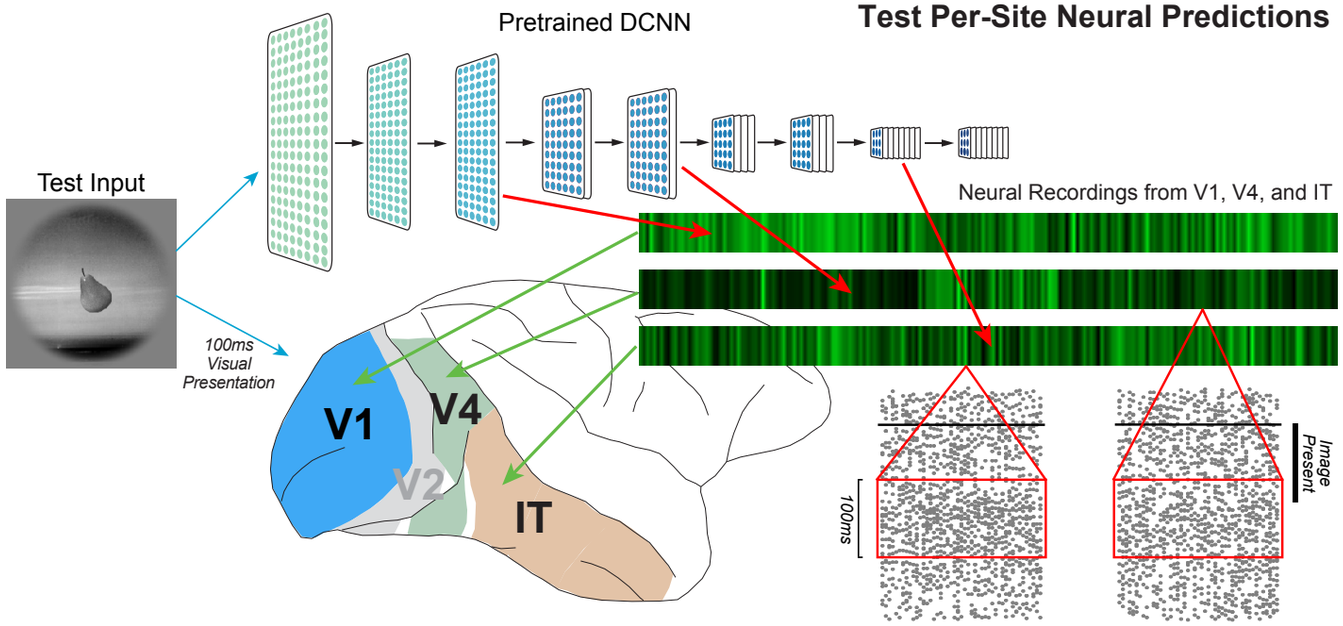
Autoencoders

Missing-Data Tasks

Deep Contrastive Embeddings

Chengxu Zhuang

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M, DiCarlo JJ, & Yamins D (2021). Unsupervised Neural Network Models of the Ventral Visual Stream. *(PNAS)*

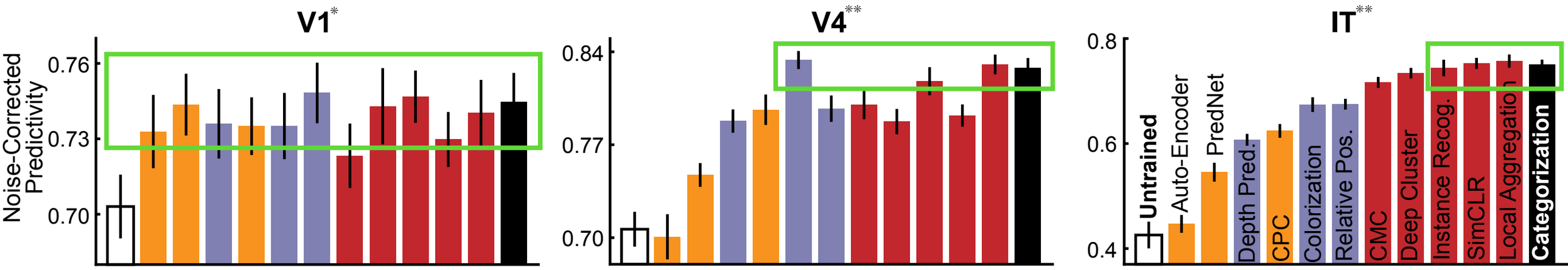


Autoencoders

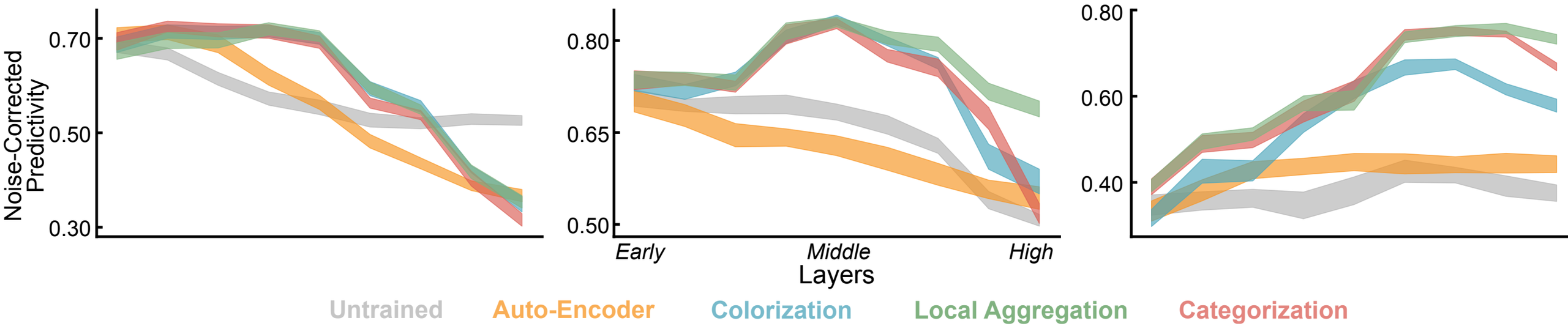
Missing-Data Tasks

Deep Contrastive Embeddings

Chengxu Zhuang



Can also measure "anatomical mapping consistency":



The Supervision Problem

2. *e.g.* **Object Categorization** ✗*bad*

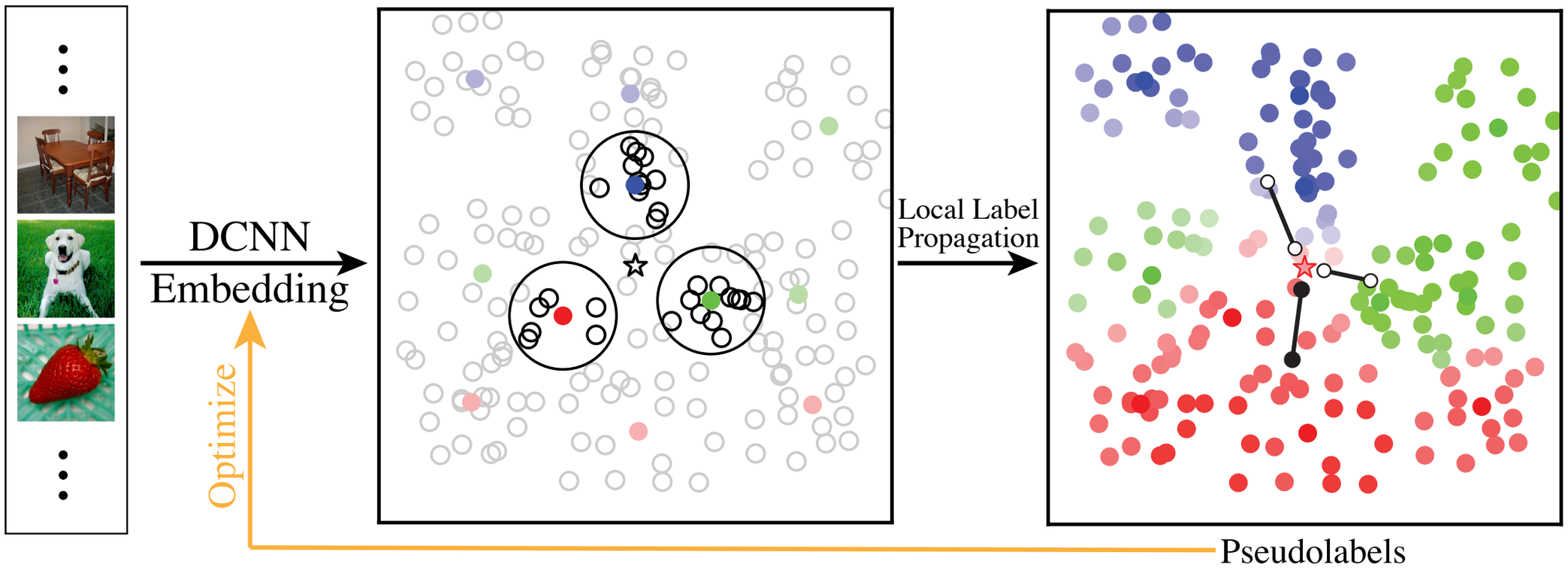
T = *task/objective*

Actually they do get **SOME** labels ….

New Semi-supervised Method: **Local Label Propagation**

Chengxu Zhuang

Local Label Propagation for Large-Scale Semi-Supervised Learning. https://arxiv.org/abs/1905.11581

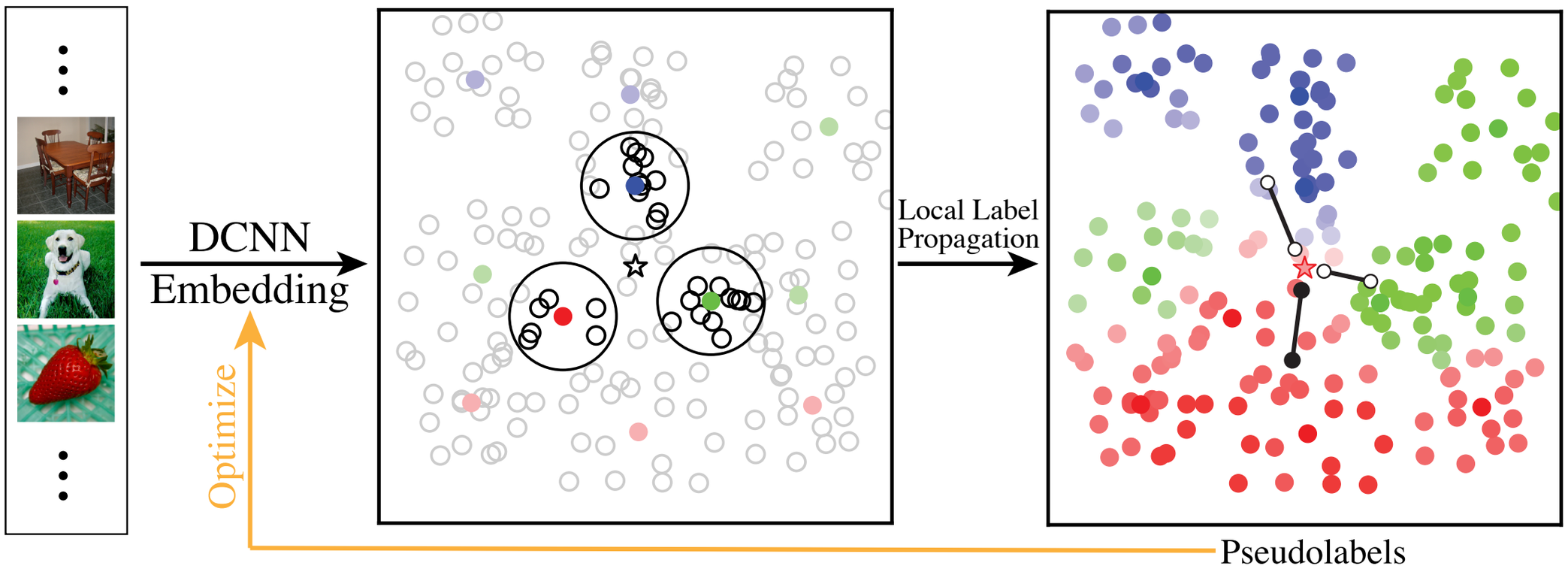


recall: $\mathcal{L}_{\mathrm{LA}}(\mathbf{C}, \mathbf{B}|\theta) = L(\mathbf{C}, \mathbf{B}|\theta) + \lambda||\theta||_2^2$

New Semi-supervised Method: **Local Label Propagation**

Local Label Propagation for Large-Scale Semi-Supervised Learning. https://arxiv.org/abs/1905.11581

Chengxu Zhuang



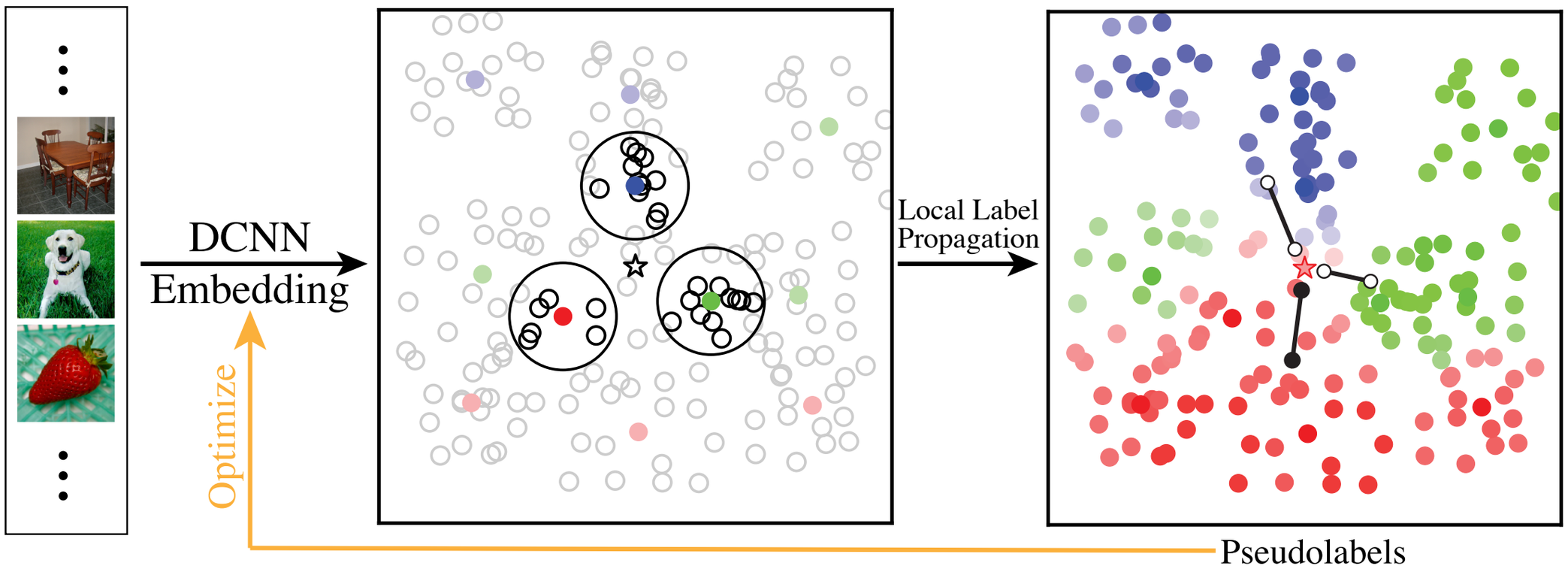
recall: $\mathcal{L}_{\mathrm{LA}}(\mathbf{C}, \mathbf{B}|\theta) = L(\mathbf{C}, \mathbf{B}|\theta) + \lambda||\theta||_2^2$

$$\mathcal{L}_{\mathrm{semi}}(x|\theta) \sim L_{\mathrm{LA}}(x|\theta) + L_{\mathrm{Cross\text{-}Ent}}(y, y_{\mathrm{pseudo}})$$

New Semi-supervised Method: **Local Label Propagation**

Local Label Propagation for Large-Scale Semi-Supervised Learning. https://arxiv.org/abs/1905.11581

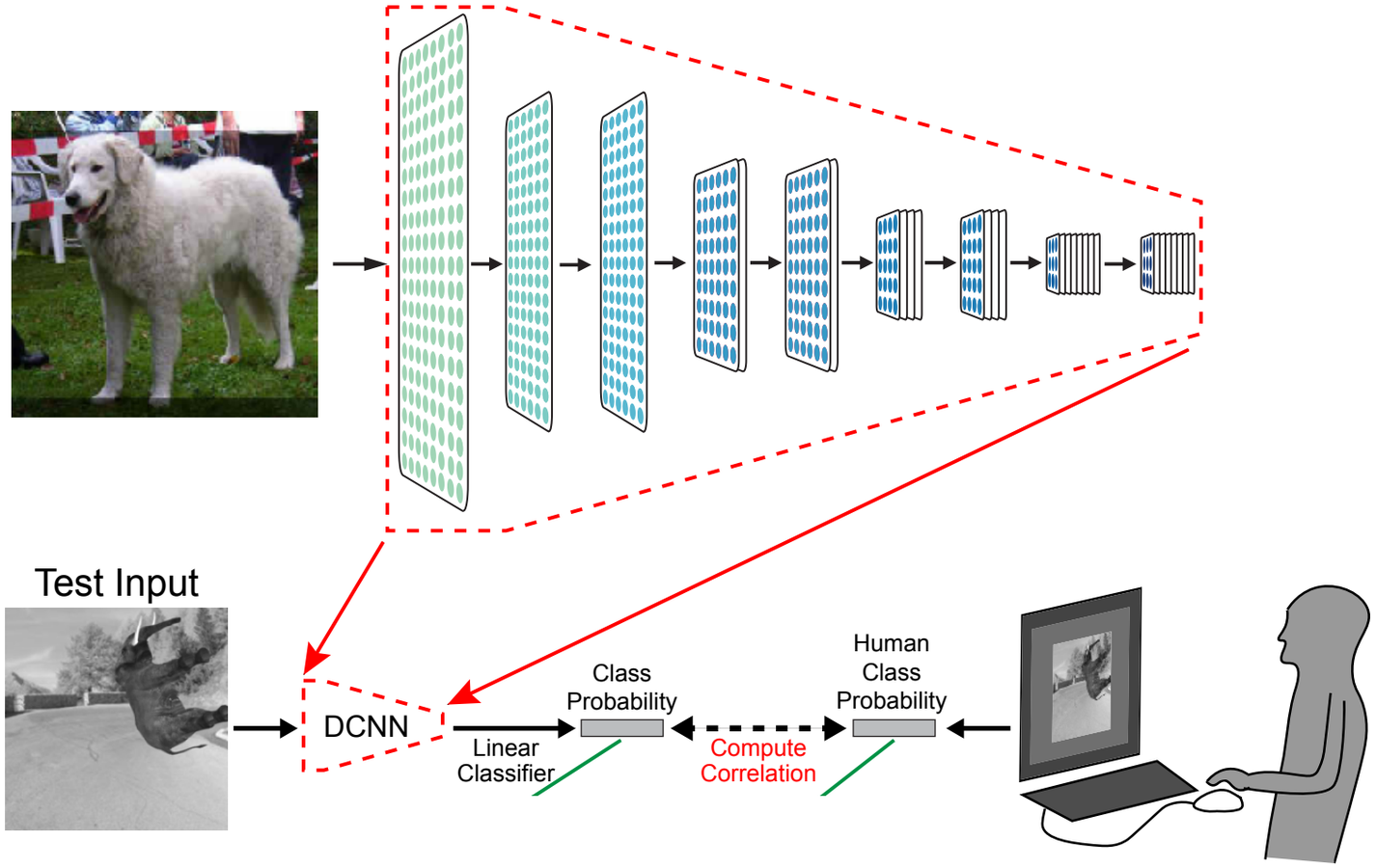
Chengxu Zhuang

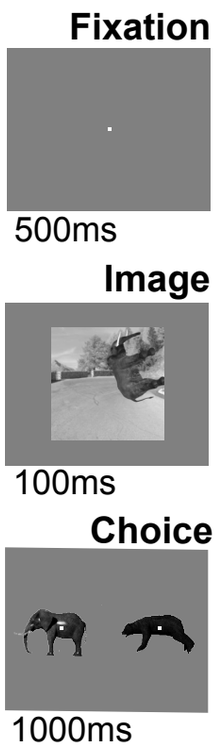


recall: $\mathcal{L}_{\mathrm{LA}}(\mathbf{C}, \mathbf{B}|\theta) = L(\mathbf{C}, \mathbf{B}|\theta) + \lambda||\theta||_2^2$

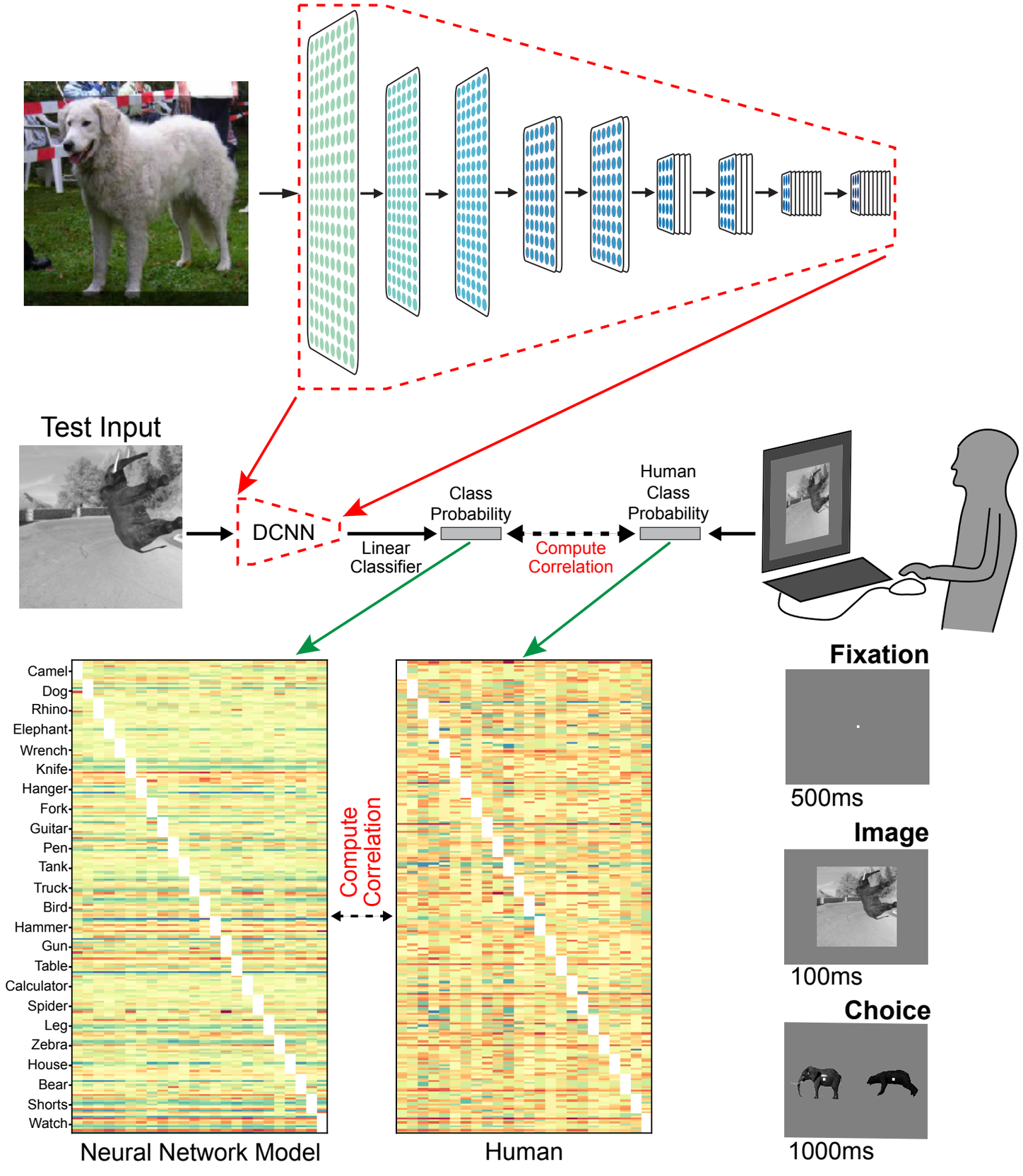
$$\mathcal{L}_{\mathrm{semi}}(x|\theta) \sim \mathrm{confidence}(y_{\mathrm{pseudo}}) \cdot [L_{\mathrm{LA}}(x|\theta) + L_{\mathrm{Cross\text{-}Ent}}(y, y_{\mathrm{pseudo}})]$$

Rajalingham, et al. ***Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks.*** Journal of Neuroscience 38.33 (2018): 7255-7269.

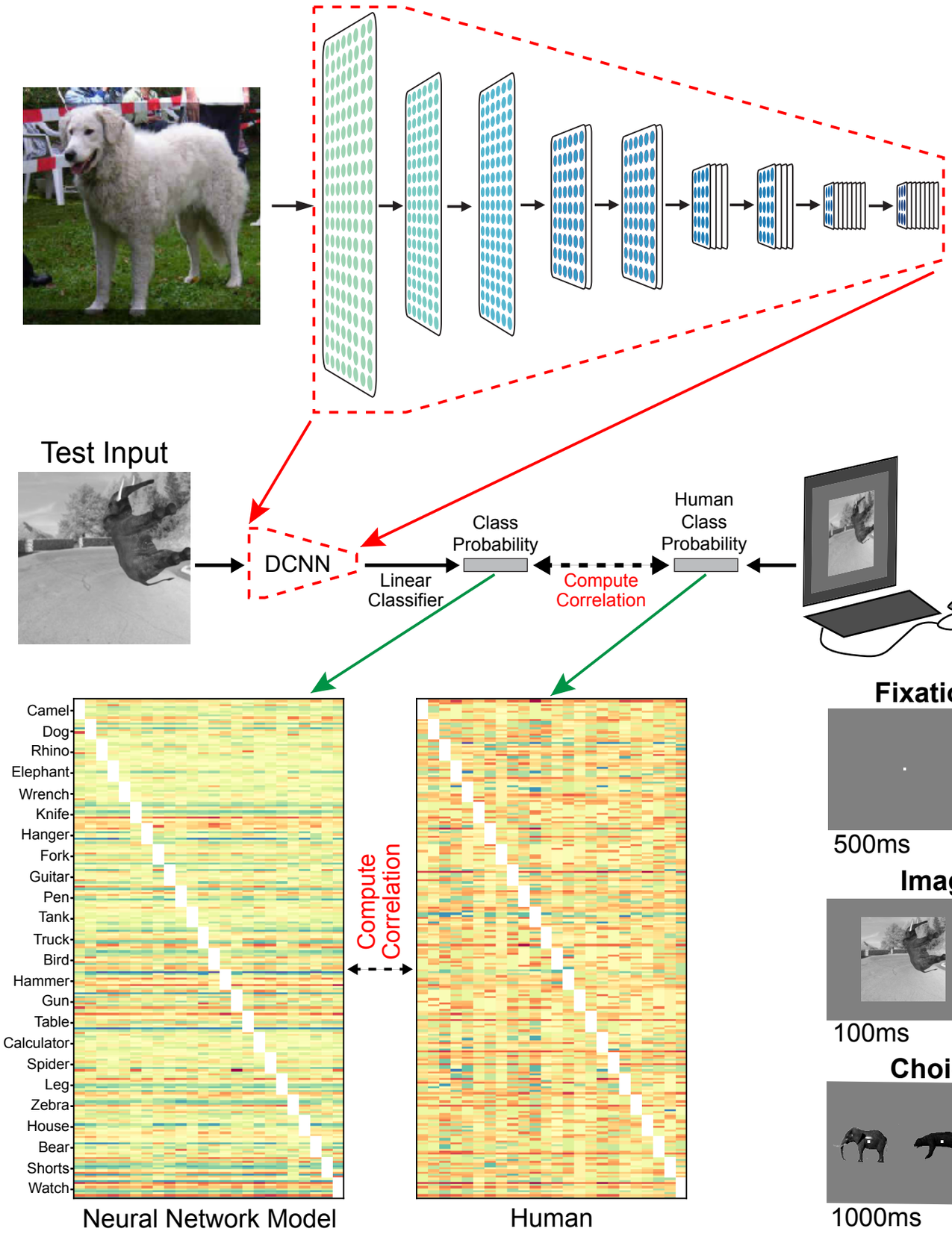


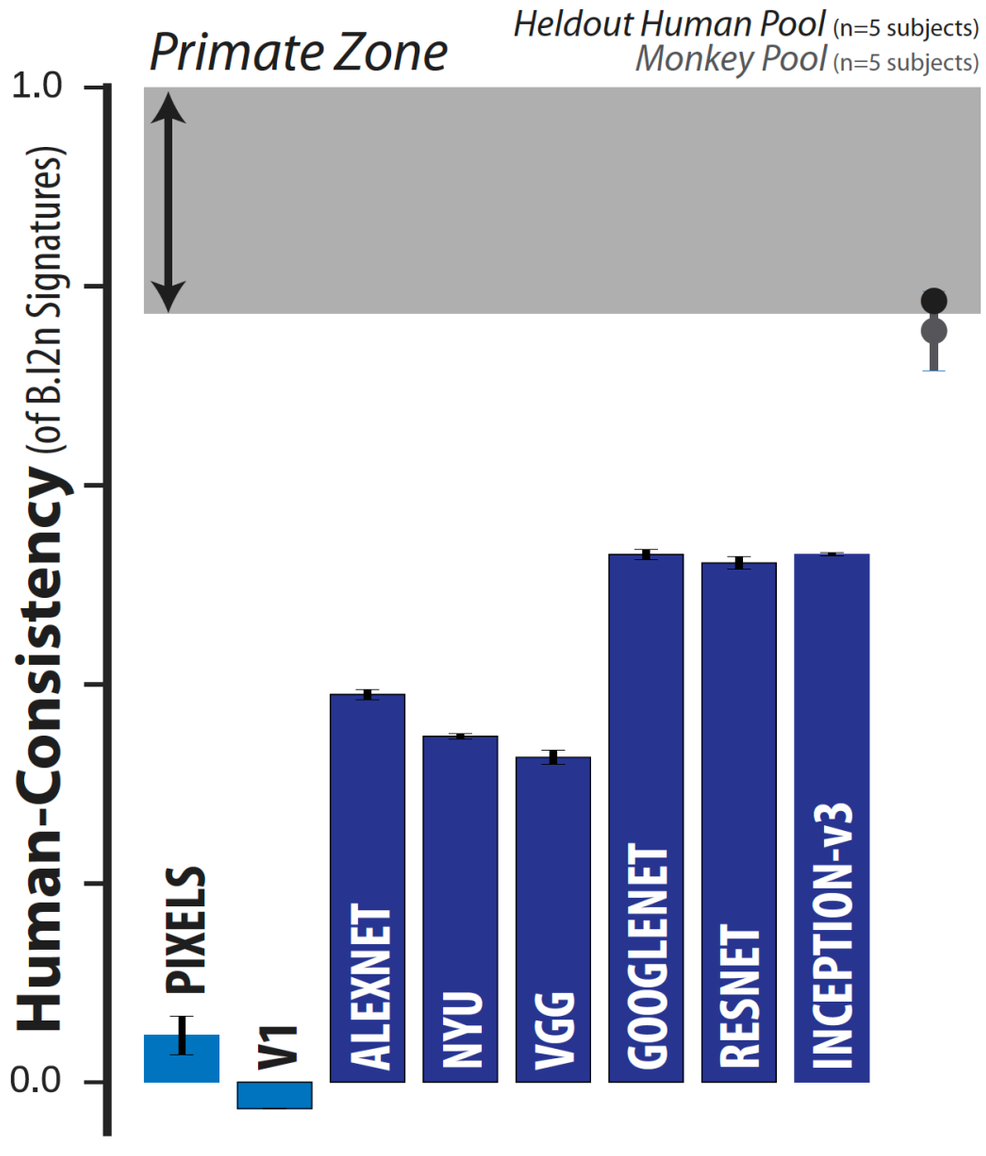


Rajalingham, et al. ***Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks.*** Journal of Neuroscience 38.33 (2018): 7255-7269.



Rajalingham, et al. ***Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks.*** Journal of Neuroscience 38.33 (2018): 7255-7269.





Behavioral Comparison

~3%
Labeled Images (36K):
~35k labels

Slovak Cuvac
Kitty
Poodle

Unlabeled Images
(1,242K)

Embedding
Pseudolabel Categorization
Consistency Loss
Local Label Propagation
Cross-Entropy Loss

Test Input

DCNN
Linear Classifier
Class Probability
Compute Correlation
Human Class Probability

Fixation
Image
Choice
500ms
100ms
1000ms

Camel
Dog
Rhino
Elephant
Wrench
Knife
Hanger
Fork
Guitar
Pen
Tank
Truck
Bird
Hammer
Gun
Table
Calculator
Spider
Leg
Zebra
House
Bear
Shorts
Watch

Compute Correlation

Neural Network Model
Human

Behavioral Comparison

~3%
Labeled Images (36K):
~35k labels

Slovak Cuvac
Kitty
Poodle

Unlabeled Images
(1,242K)

Embedding
Pseudolabel
Categorization
Consistency Loss
Local Label Propagation
Cross-Entropy Loss

Test Input

DCNN
Linear Classifier
Class Probability
Compute Correlation
Human Class Probability

Fixation
Image
Choice
500ms
100ms
1000ms

Camel
Dog
Rhino
Elephant
Wrench
Knife
Hanger
Fork
Guitar
Pen
Tank
Truck
Bird
Hammer
Gun
Table
Calculator
Spider
Leg
Zebra
House
Bear
Shorts
Watch

Compute Correlation

Neural Network Model
Human

Noise Ceiling = 0.697
Human Behavior Consistency
0.4
0.2
Untrained
Auto-Encoder
PredNet
Depth Prediction
CPC
Colorization
Relative Position
CMC
Deep Cluster
Instance Recog.
SimCLR
Local Aggregation
Few-Label Cate.
Mean-Teacher
Local Label Propagation
Categorization
Self-supervised
Semi-supervised

Behavioral Comparison

~3%
Labeled Images (36K):
~35k labels

Slovak Cuvac
Kitty
Poodle

Unlabeled Images
(1,242K)

Embedding
Pseudolabel
Categorization
Consistency Loss
Local Label Propagation
Cross-Entropy Loss

Test Input

DCNN
Linear Classifier
Class Probability
Compute Correlation
Human Class Probability

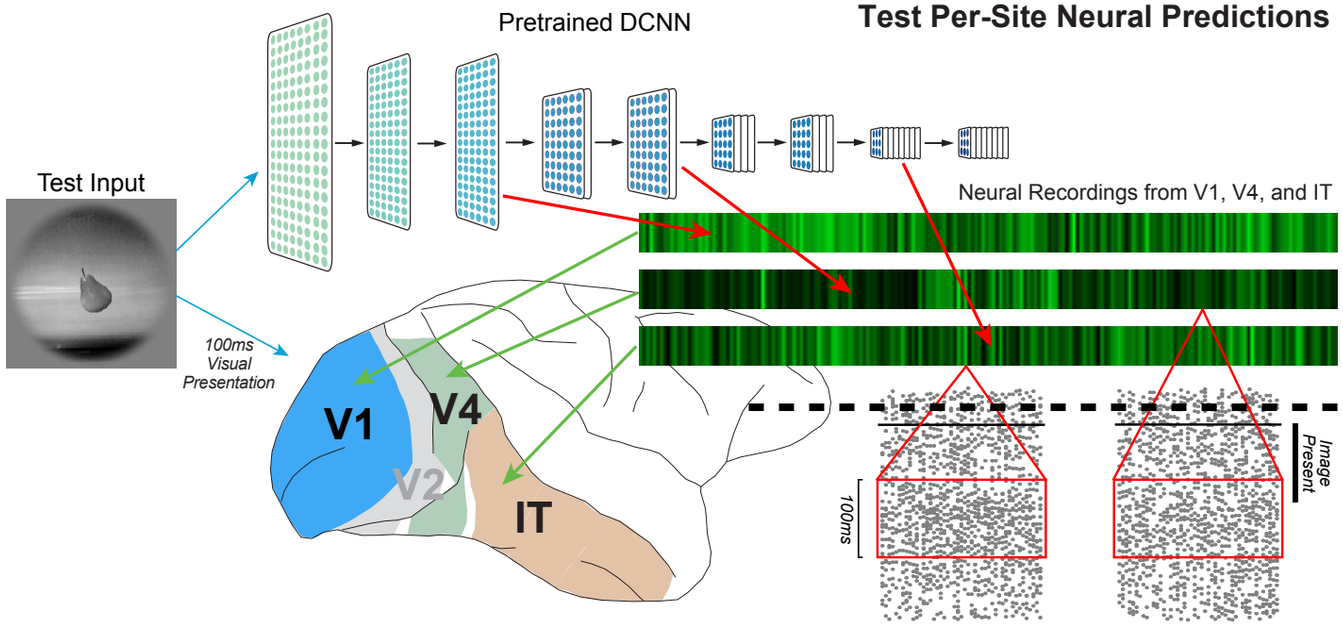
Fixation
Image
Choice
500ms
100ms
1000ms

Camel
Dog
Rhino
Elephant
Wrench
Knife
Hanger
Fork
Guitar
Pen
Tank
Truck
Bird
Hammer
Gun
Table
Calculator
Spider
Leg
Zebra
House
Bear
Shorts
Watch

Compute Correlation

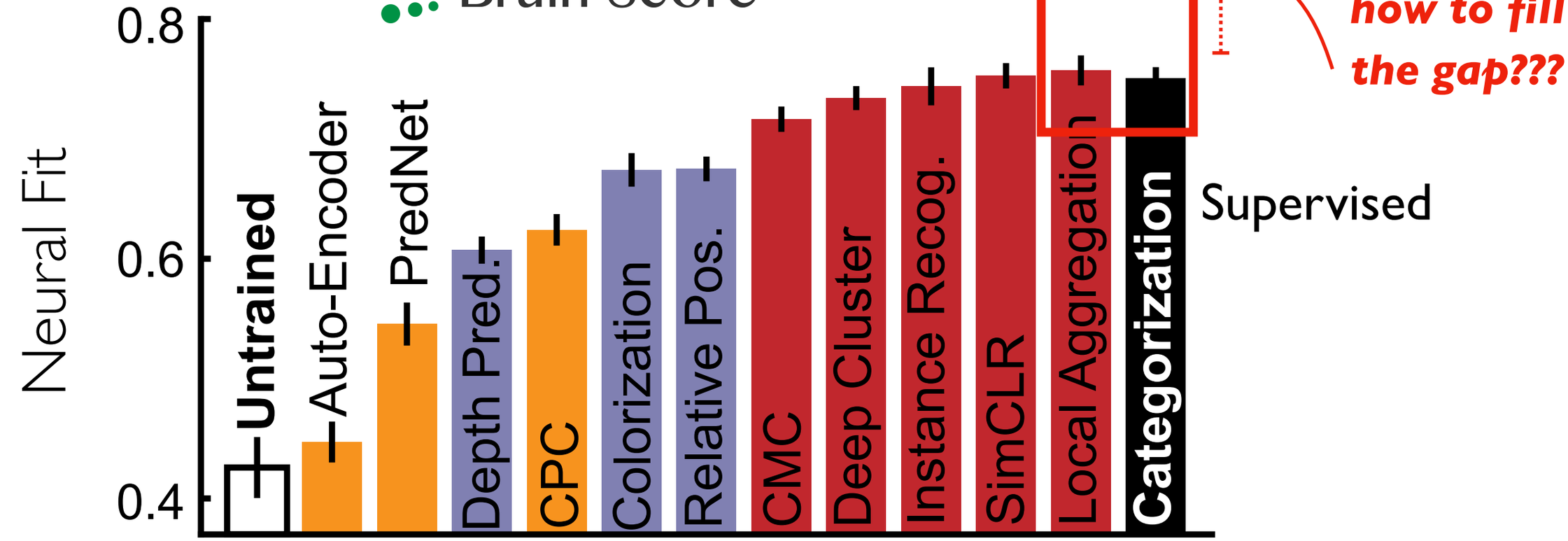
Neural Network Model
Human

Noise Ceiling = 0.697
Human Behavior Consistency
0.4
0.2
Untrained
Auto-Encoder
PredNet
Depth Prediction
CPC
Colorization
Relative Position
CMC
Deep Cluster
Instance Recog.
SimCLR
Local Aggregation
Few-Label Cate.
Mean-Teacher
Local Label Propagation
Categorization
Self-supervised
Semi-supervised



Chengxu Zhuang

Brain-Score



not really better!

how to fill the gap???

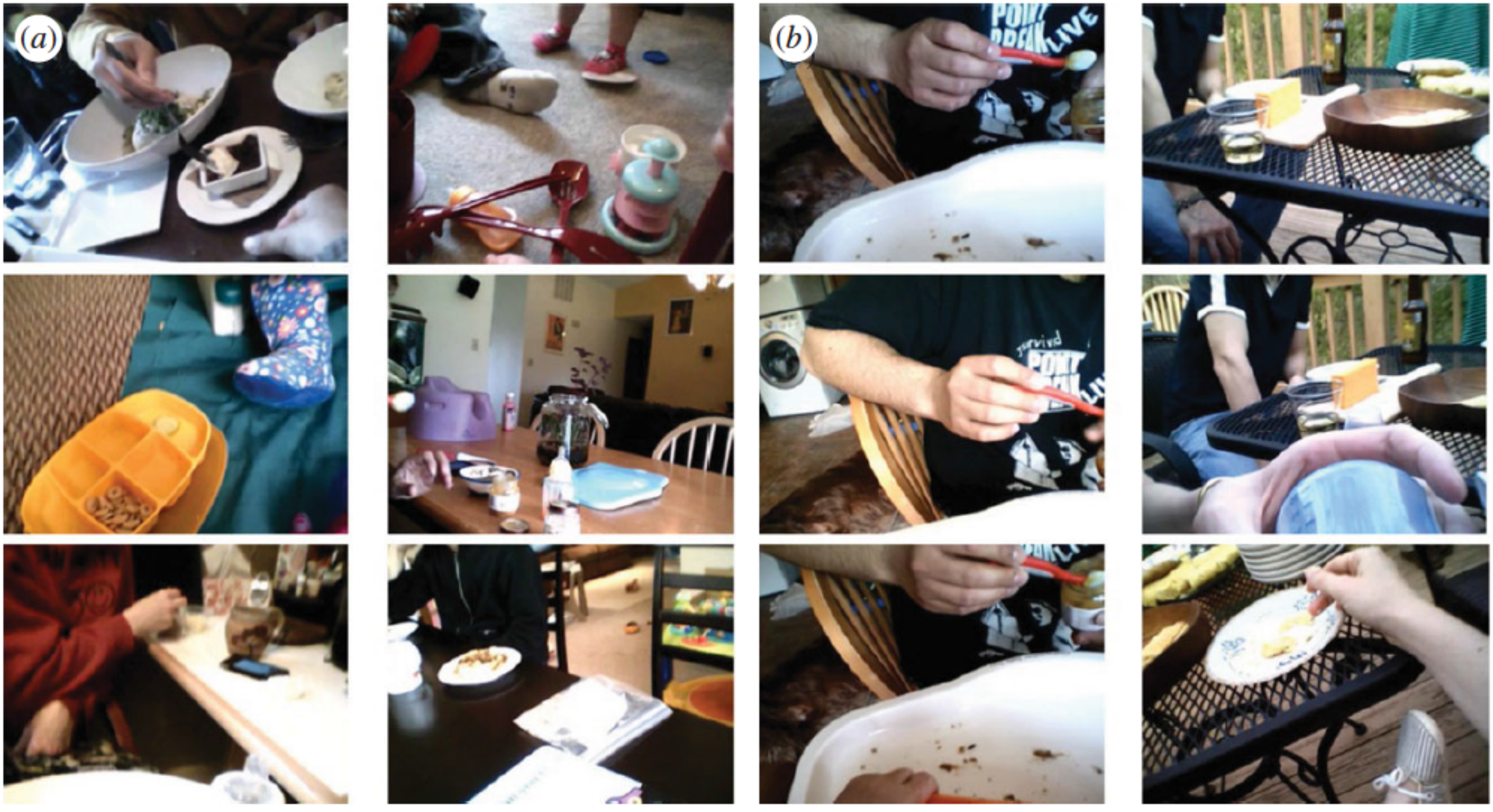
Quantitatively accurate unsupervised model of a higher brain area.

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M, DiCarlo JJ, & Yamins D (2021). Unsupervised Neural Network Models of the Ventral Visual Stream. *(PNAS)*

Take-aways:

Contrastive unsupervised approaches finally have largely made up the "supervision gap" in performance & neural fits.

This is more like what real visual experience looks like:



Clerkin, Hart, Rehg, Yu, & Smith (2017)

Contrastive Embeddings in the Wild

SAYCam Dataset:

Three infants aged 6-32 months

Mike Frank

Contrastive Embeddings in the Wild

SAYCam Dataset:

Three infants aged 6-32 months

Head-mounted camera

Mike Frank

Contrastive Embeddings in the Wild

SAYCam Dataset:

Three infants aged 6-32 months

Head-mounted camera

Mono video and audio channels

~2 hours per week

Mike Frank

Contrastive Embeddings in the Wild

SAYCam Dataset:

Three infants aged 6-32 months

Head-mounted camera

Mono video and audio channels

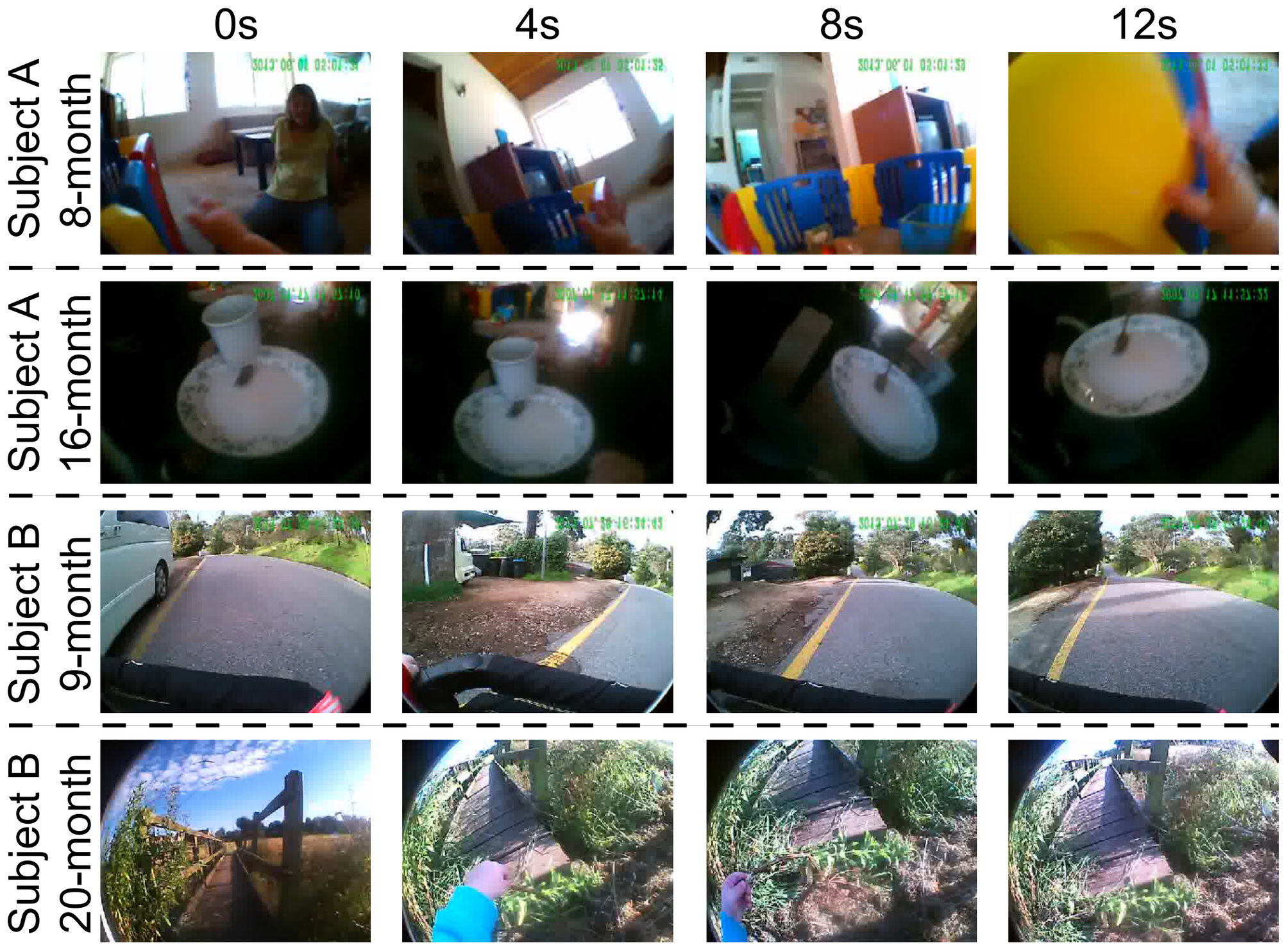
~2 hours per week

Mike Frank

Q: How would you use this dataset to learn a representation?

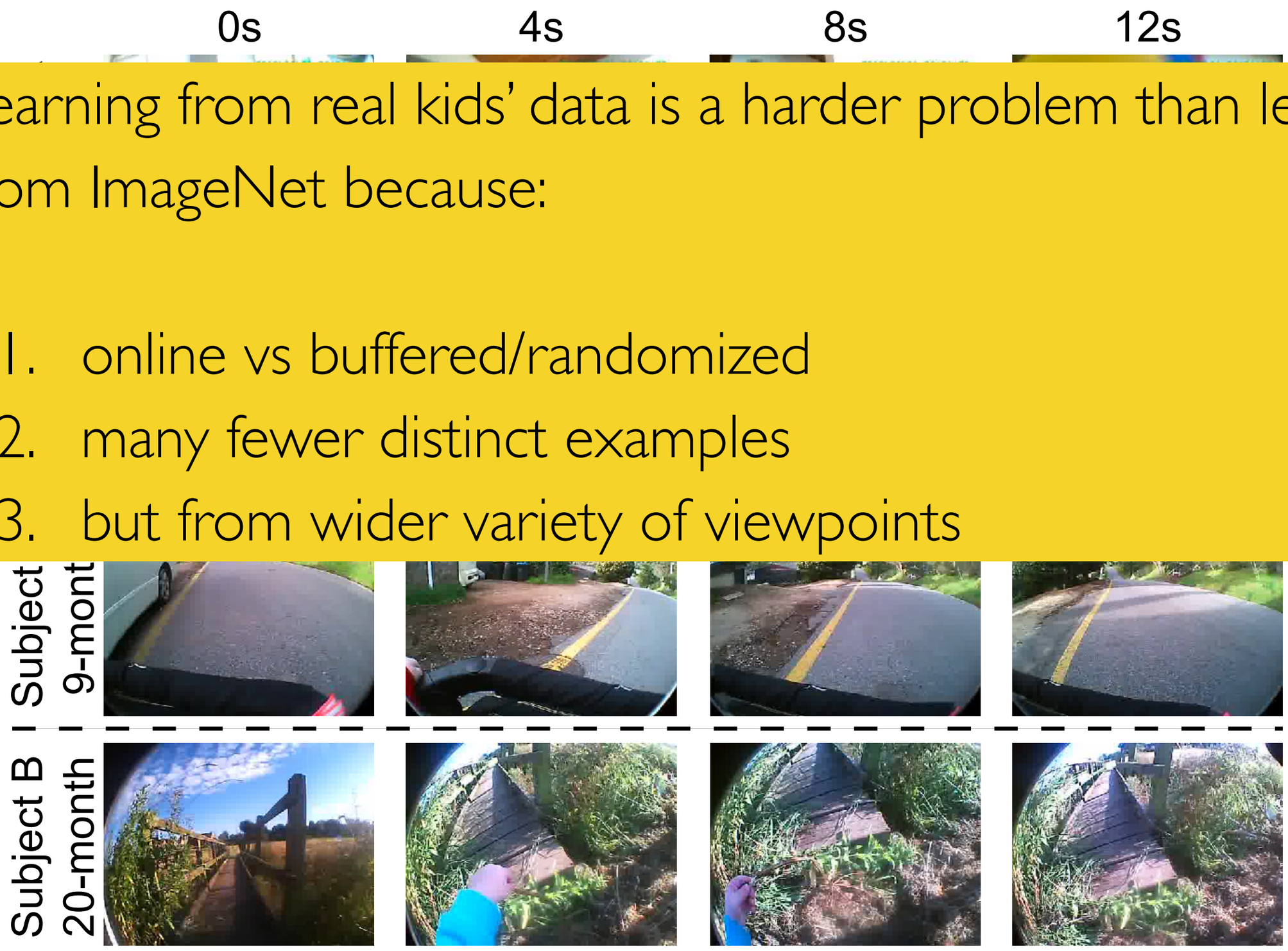
Learning from real datastreams

SAY-Cam examples



Learning from real datastreams

SAY-Cam examples



Learning from real kids' data is a harder problem than learning from ImageNet because:

1. online vs buffered/randomized
2. many fewer distinct examples
3. but from wider variety of viewpoints

Learning from real datastreams

SAYCam Dataset:

Three infants aged 6-32 months

Head-mounted camera

Mono video and audio channels

~2 hours per week

Mike Frank

Q: How would you use this dataset to learn a representation?

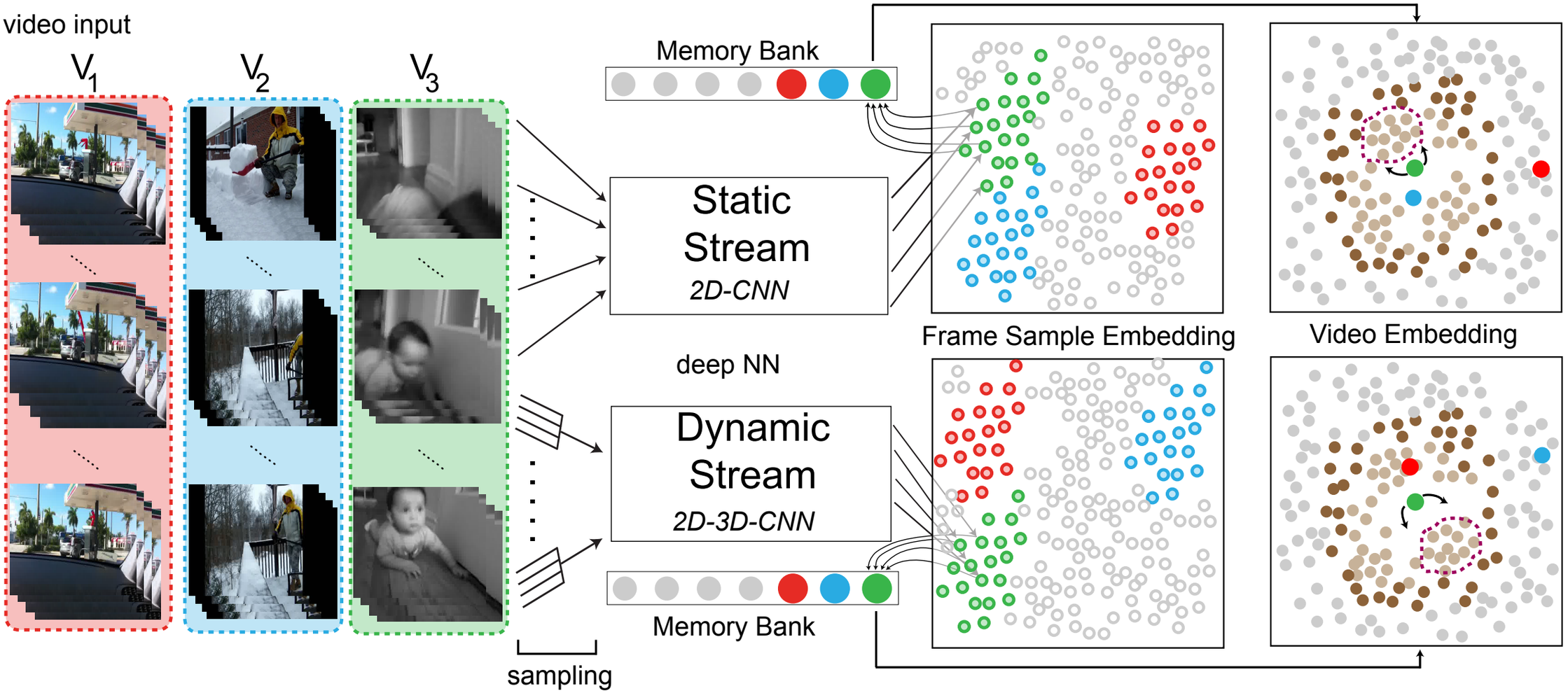
A: Extend deep embedding approach to videos?

Learning from real datastreams

Chengxu Zhuang

Unsupervised Learning from Video with Deep Neural Embeddings.

(CVPR 2020) https://arxiv.org/abs/1905.11954



Video Instance Embedding (VIE) → $\mathbf{e} = \frac{\mathbb{E}_\rho[\phi_\theta(\mathbf{f})]}{||\mathbb{E}_\rho[\phi_\theta(\mathbf{f})]||_2}$

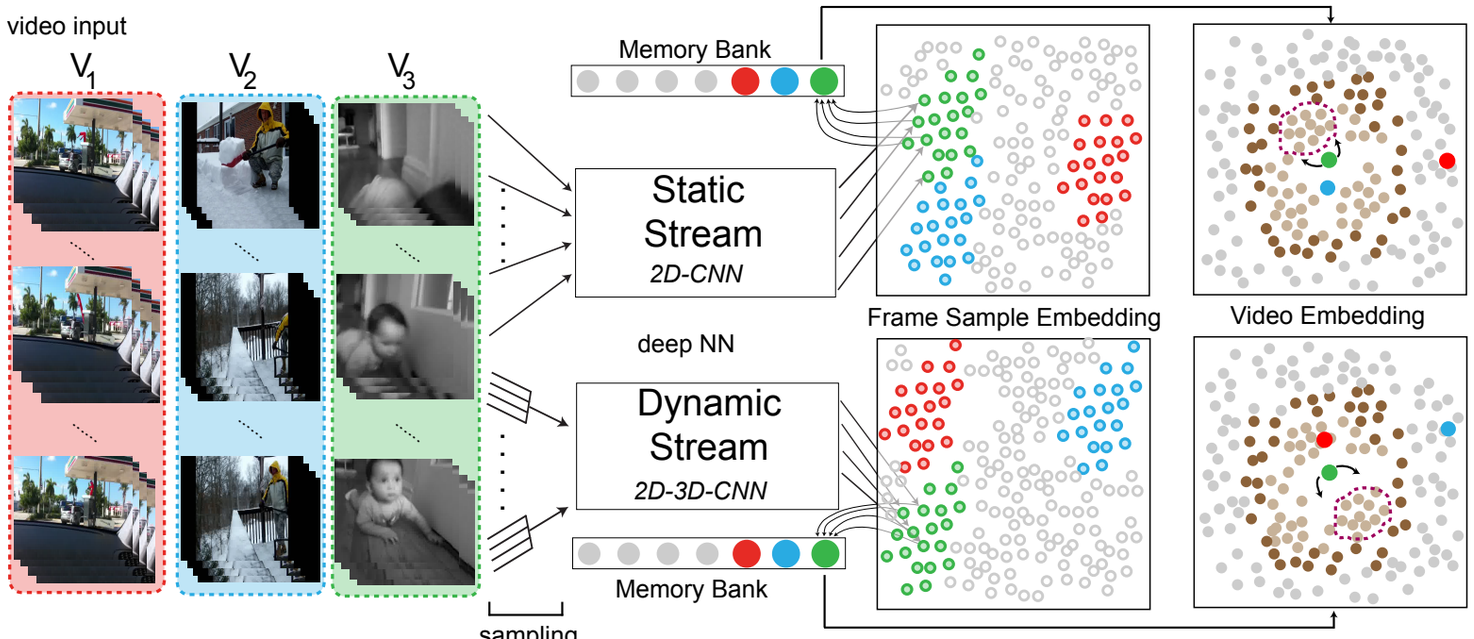
Contrastive Embeddings in the Wild

SAYCam Dataset:

Head-mounted camera, 3 infants aged 6-32 months

Mike Frank

Using contrastive learning with head-cam video

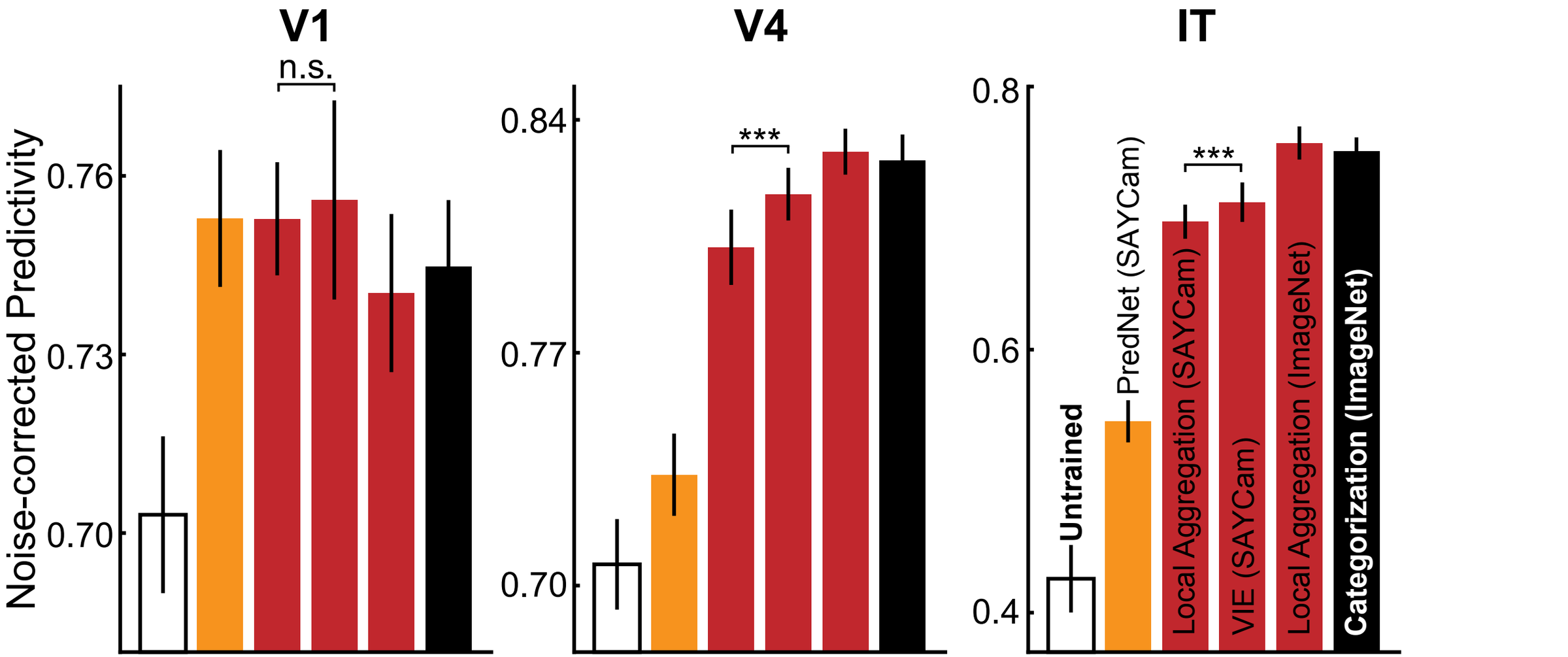


Unsupervised Learning from Video with Deep Neural Embeddings.

(CVPR 2020) https://arxiv.org/abs/1905.11954

Learning from real datastreams

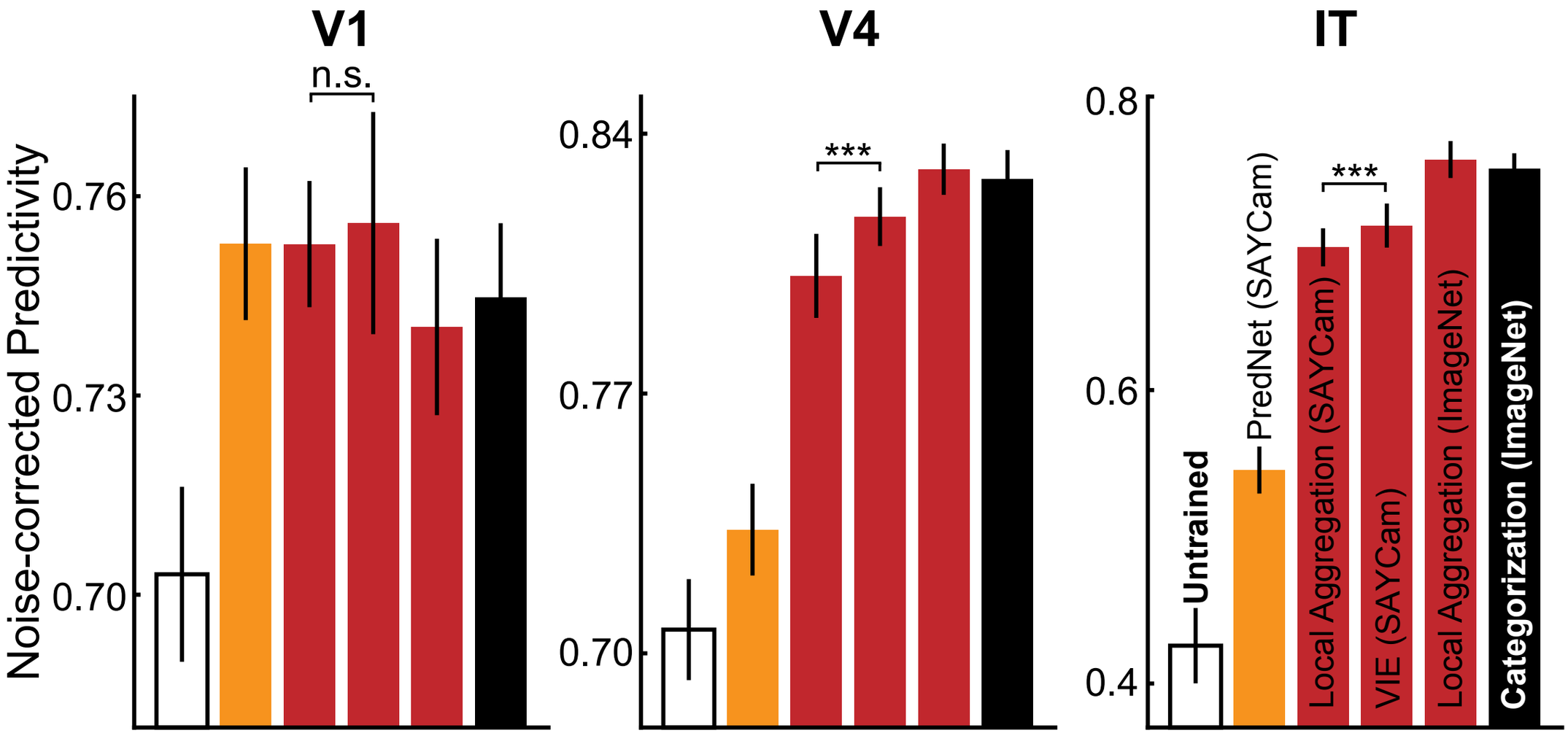
Video learning from SAY-Cam with deep contrastive embeddings predicts neurons substantially than stronger alternatives (predictive coding)



Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M, DiCarlo JJ, & Yamins D (2021). Unsupervised Neural Network Models of the Ventral Visual Stream. *(PNAS)*

Learning from real datastreams

Video learning from SAY-Cam with deep contrastive embeddings predicts neurons substantially than stronger alternatives (predictive coding)

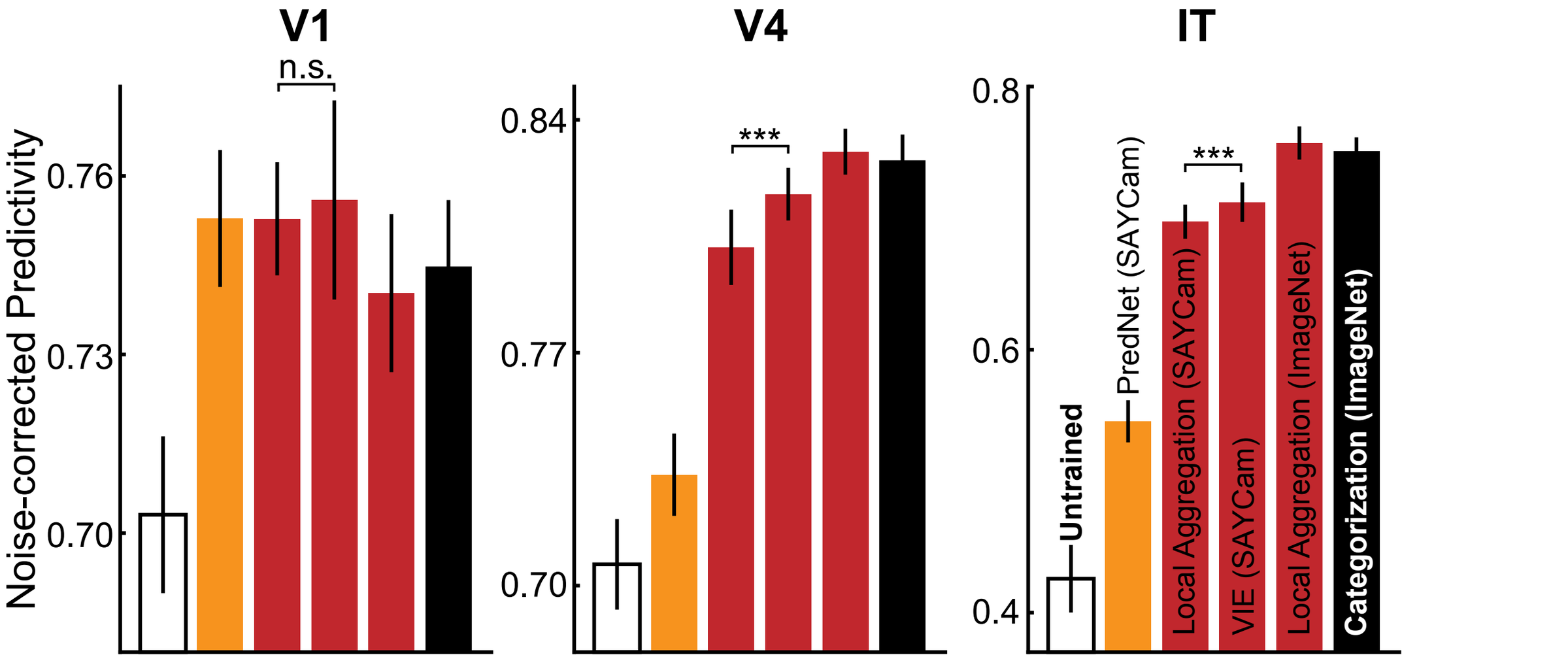


Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M, DiCarlo JJ, & Yamins D (2021). Unsupervised Neural Network Models of the Ventral Visual Stream. *(PNAS)*

Advantage to video compared to still-frames

Learning from real datastreams

Video learning from SAY-Cam with deep contrastive embeddings predicts neurons substantially than stronger alternatives (predictive coding)



Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M, DiCarlo JJ, & Yamins D (2021). Unsupervised Neural Network Models of the Ventral Visual Stream. *(PNAS)*

Advantage to video compared to still-frames

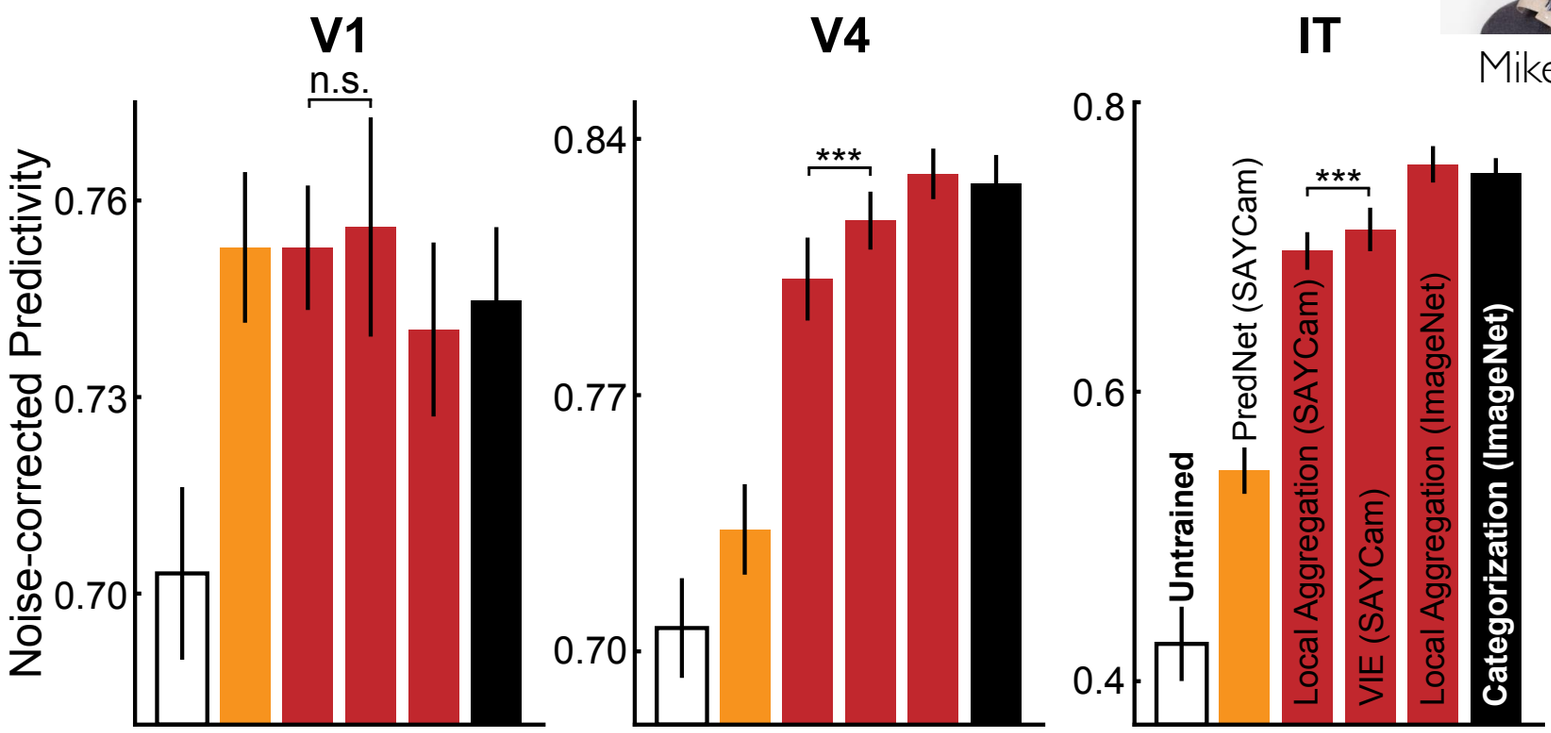
But, still some gap between training on ImageNet and training on SAY-Cam

Contrastive Embeddings in the Wild

SAYCam Dataset:

Head-mounted camera, 3 infants aged 6-32 months

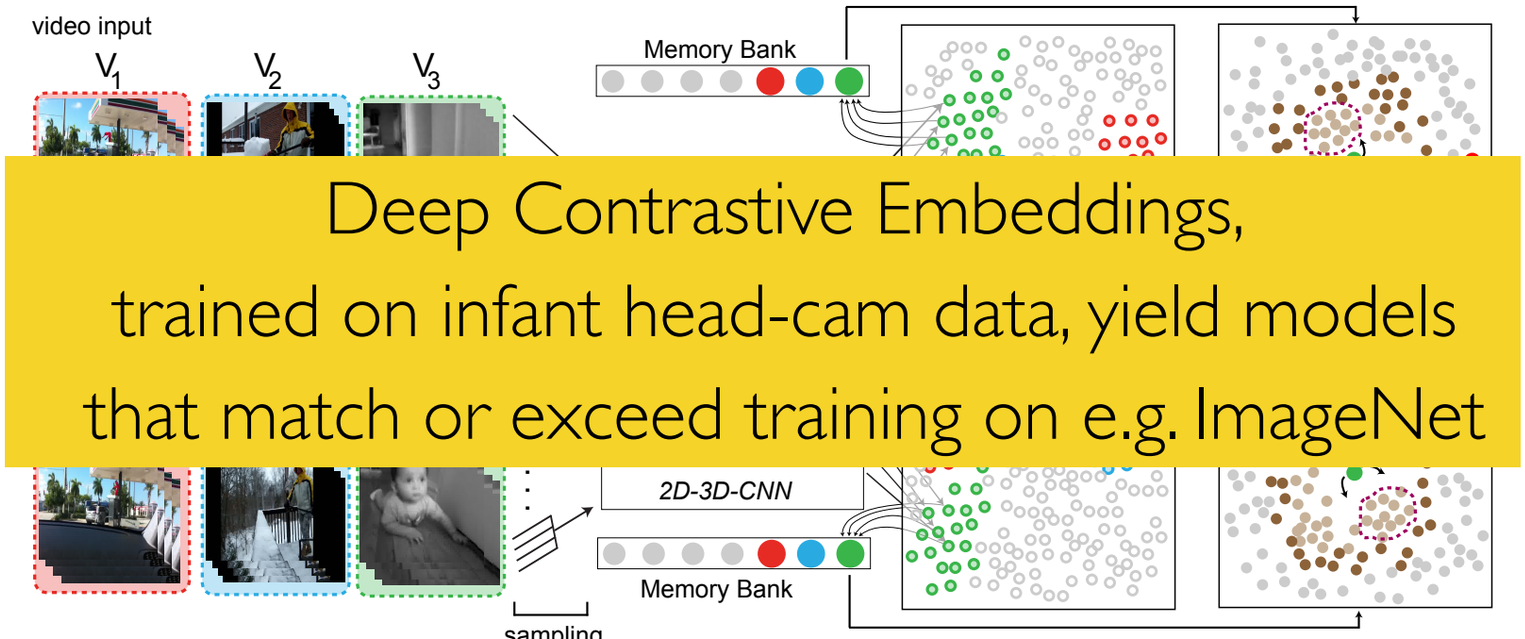
Mike Frank



Brain-Score

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M, DiCarlo JJ, & Yamins D (2021). Unsupervised Neural Network Models of the Ventral Visual Stream. *(PNAS)*

Using contrastive learning with head-cam video



Deep Contrastive Embeddings,
trained on infant head-cam data, yield models
that match or exceed training on e.g. ImageNet

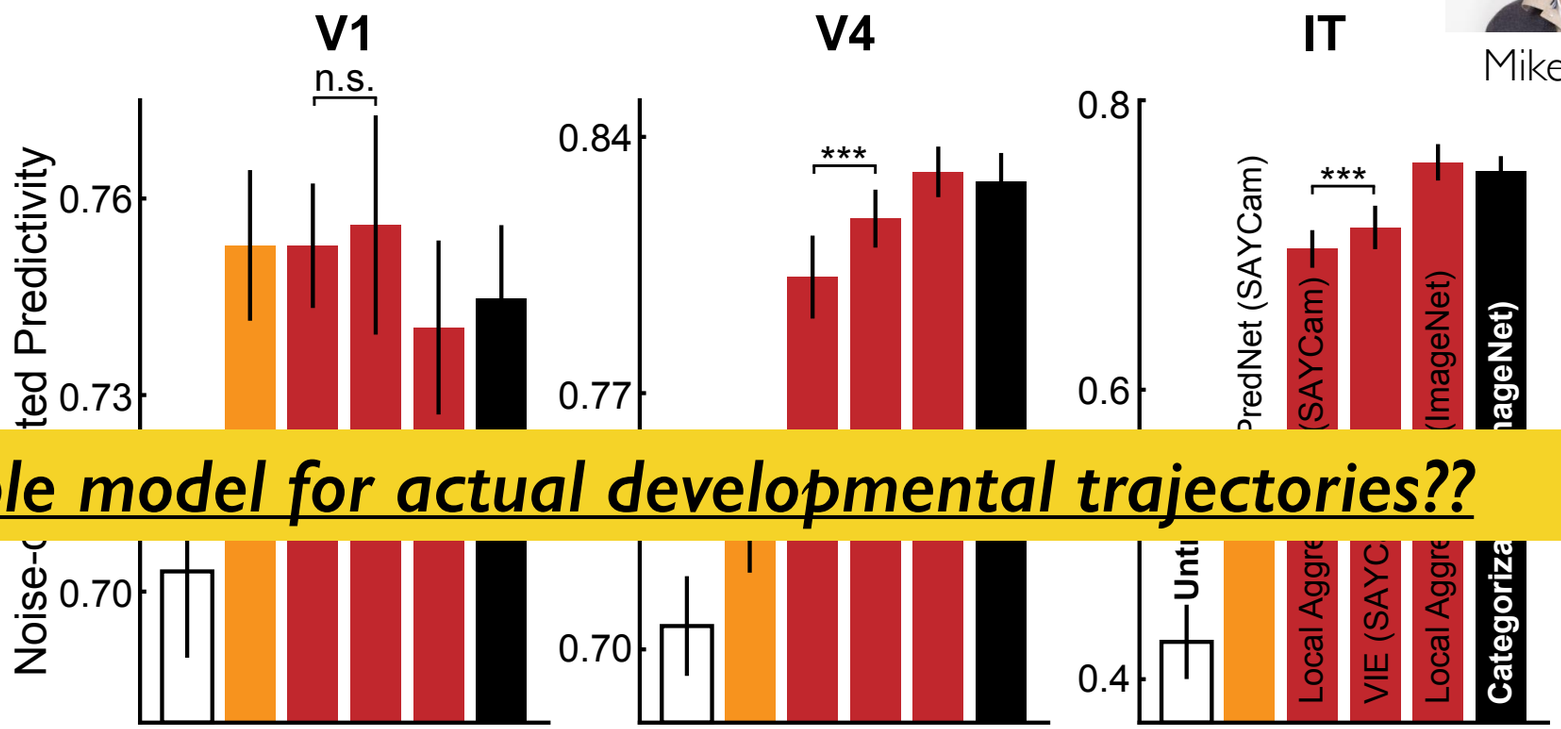
Unsupervised Learning from Video with Deep Neural Embeddings.
(CVPR 2020) https://arxiv.org/abs/1905.11954

Contrastive Embeddings in the Wild

SAYCam Dataset:

Head-mounted camera, 3 infants aged 6-32 months

Mike Frank

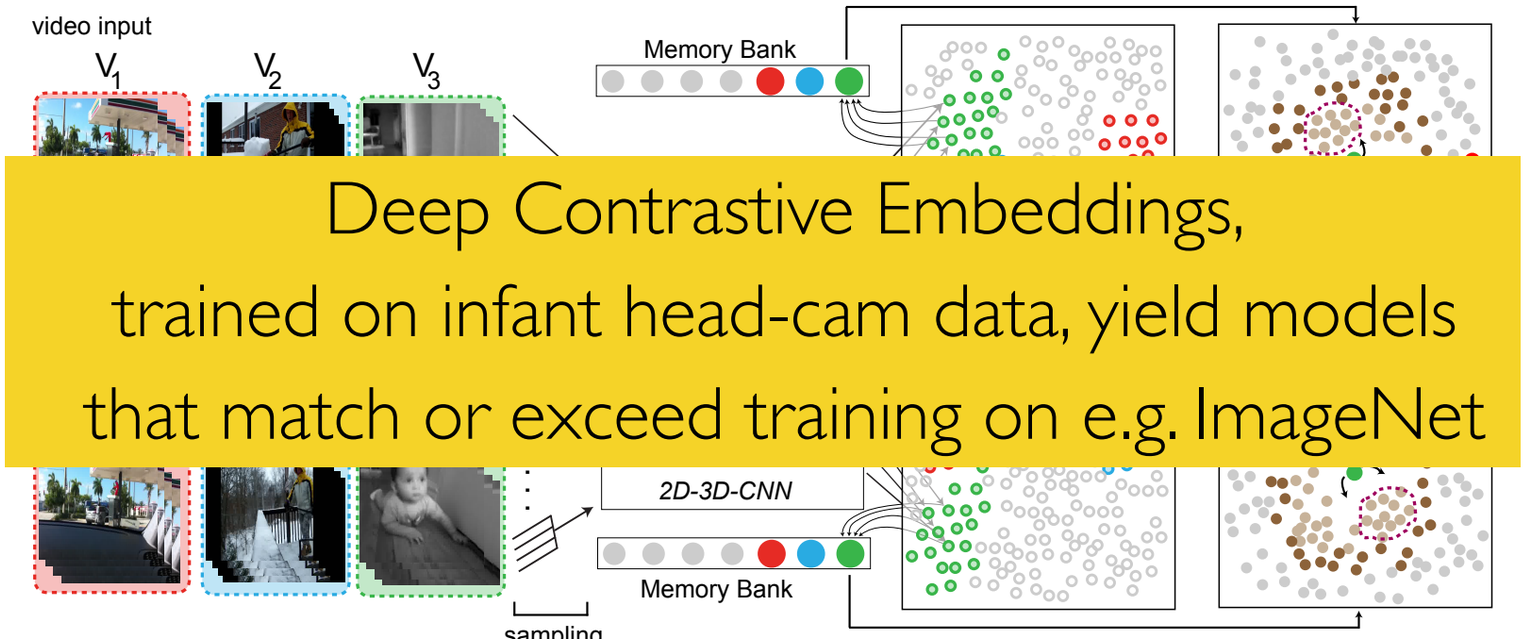


Brain-Score

Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank M, DiCarlo JJ, & Yamins D (2021). Unsupervised Neural Network Models of the Ventral Visual Stream. *(PNAS)*

Possible reasonable model for actual developmental trajectories??

Using contrastive learning with head-cam video



Deep Contrastive Embeddings, trained on infant head-cam data, yield models that match or exceed training on e.g. ImageNet

Unsupervised Learning from Video with Deep Neural Embeddings.
(CVPR 2020) https://arxiv.org/abs/1905.11954

Big Problems in Each Area

***✓ok** = we've really nailed it
***✓ok-ish** = **harder to reject out of hand**
***bad** = obviously deeply wrong

PROBLEM

1. ***✓ok-ish**

A = *architecture class*

e.g. **CNNs**

NO TOPOGRAPHICAL STRUCTURE

2. ***✓ok-ish**

T = *task/objective*

e.g. **Object Categorization**

TOO MUCH LABELLED DATA REQUIRED!!?

3. ***✓ok-ish**

D = *dataset*

e.g. **ImageNet**

REAL NOISY VIDEO DATASTREAMS vs STEREOTYPED CLEAN STILL IMAGES

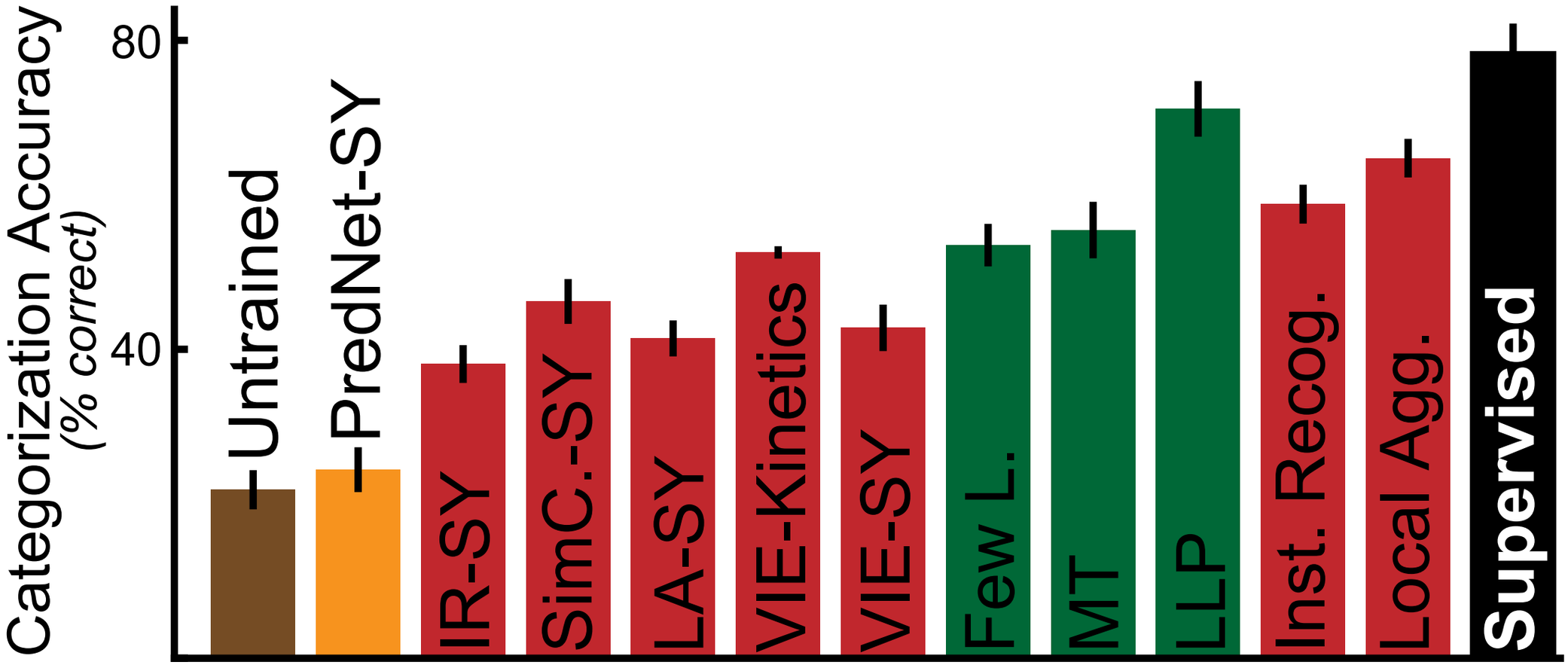
4. ***Xbad**

L = *learning rule*

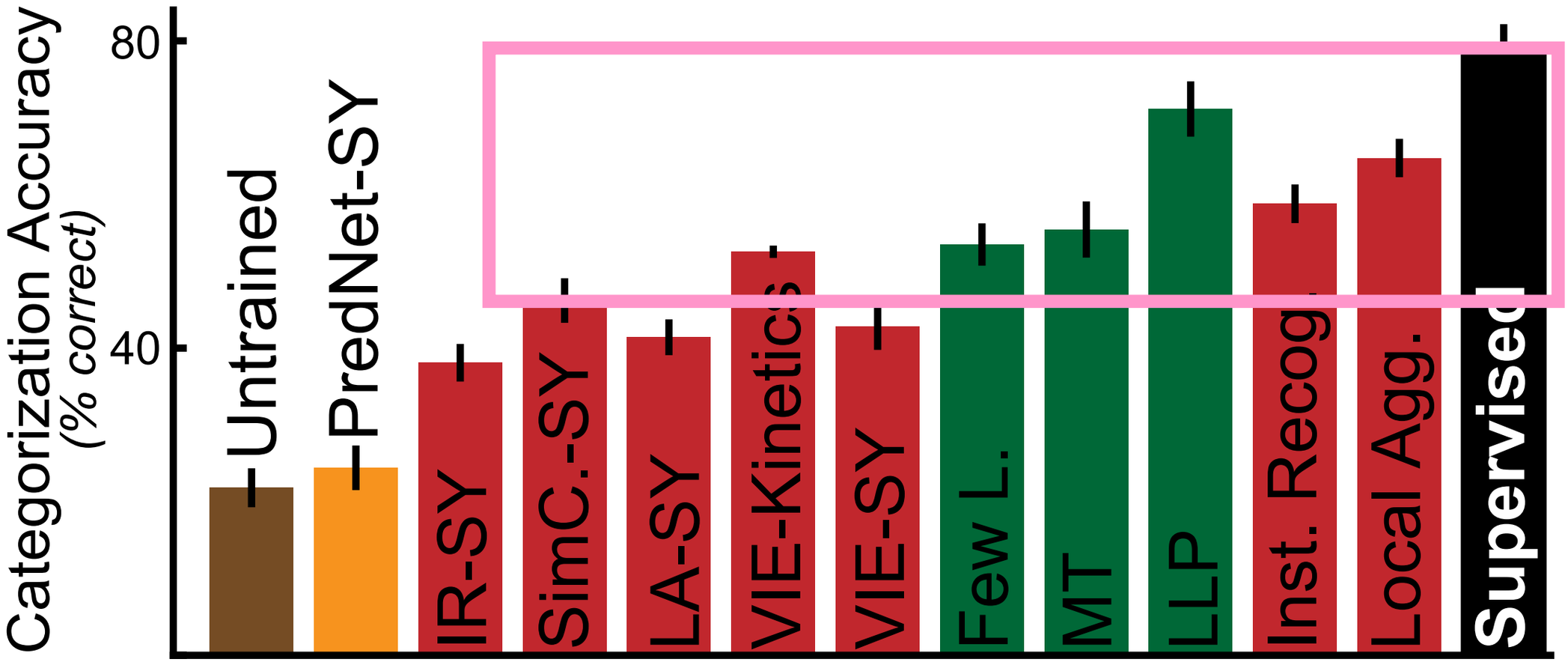
e.g. **Arch. Srch.** + **Grad. Desc.**

BACKPROP AND ITS DISCONTENTS

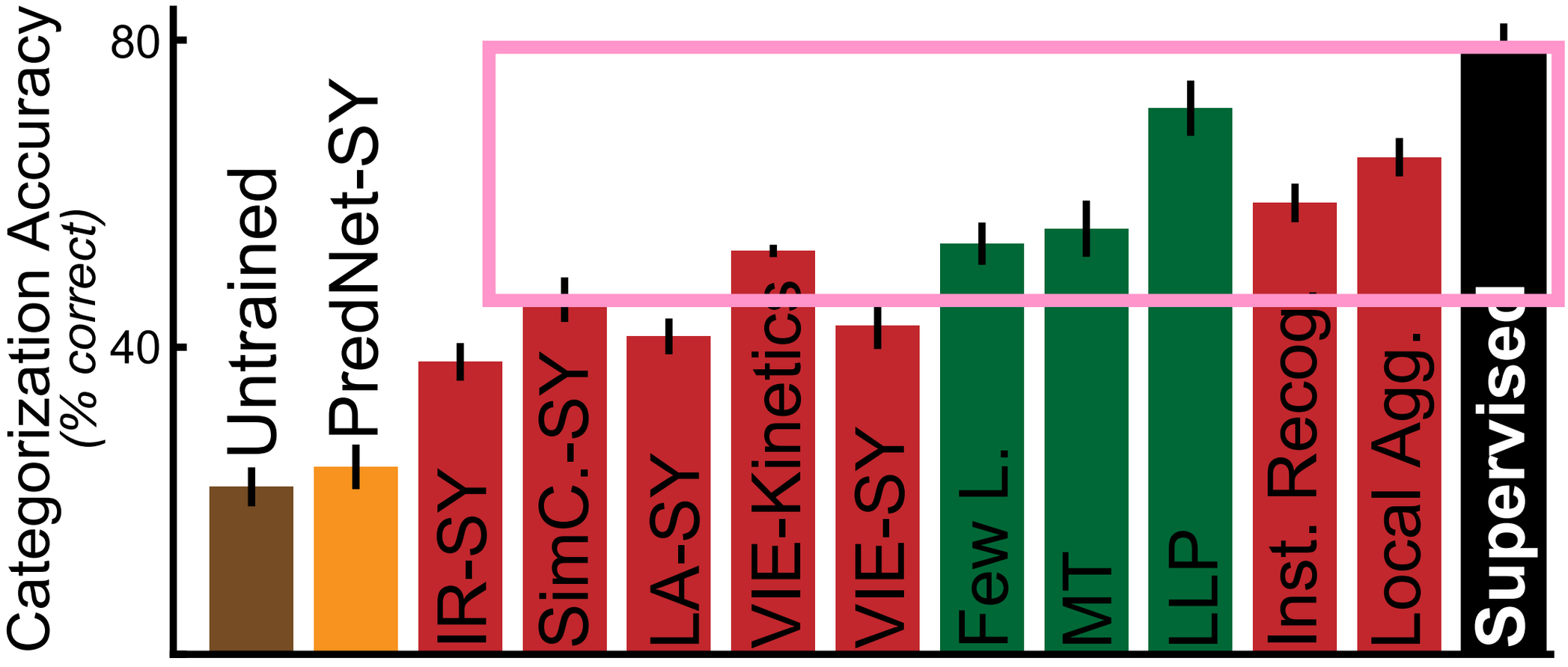
But still quite imperfect. . .



But still quite imperfect. . .



But still quite imperfect. . .



Problem: Current algorithms trained on **existing developmentally-appropriate datasets** don't learn very strong representations.

Chen et al. 2020

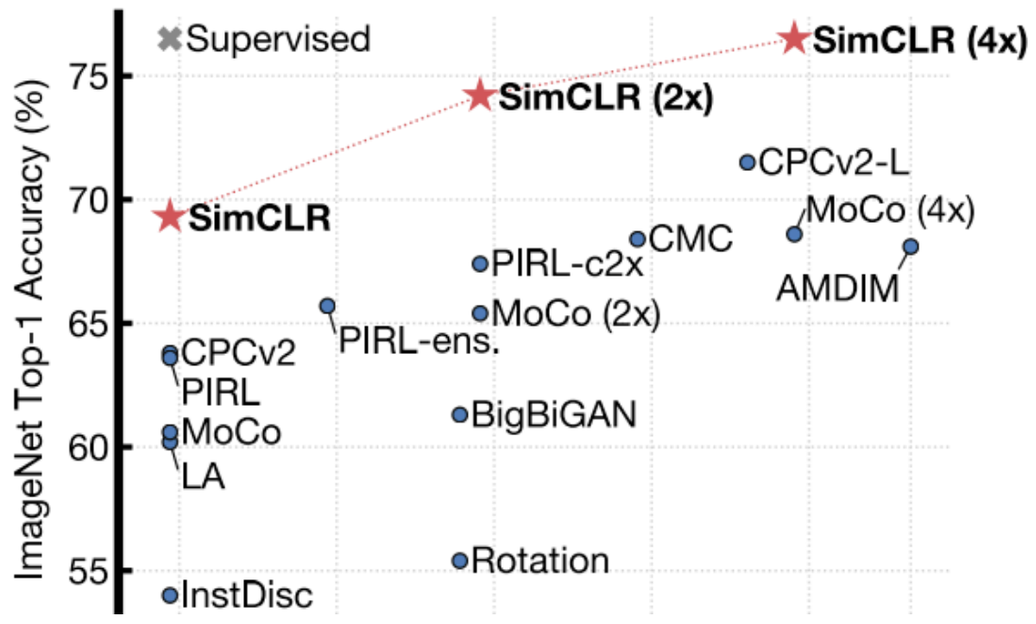
Since then, many more algorithms have been proposed.

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen [1] Simon Kornblith [1] Mohammad Norouzi [1] Geoffrey Hinton [1]

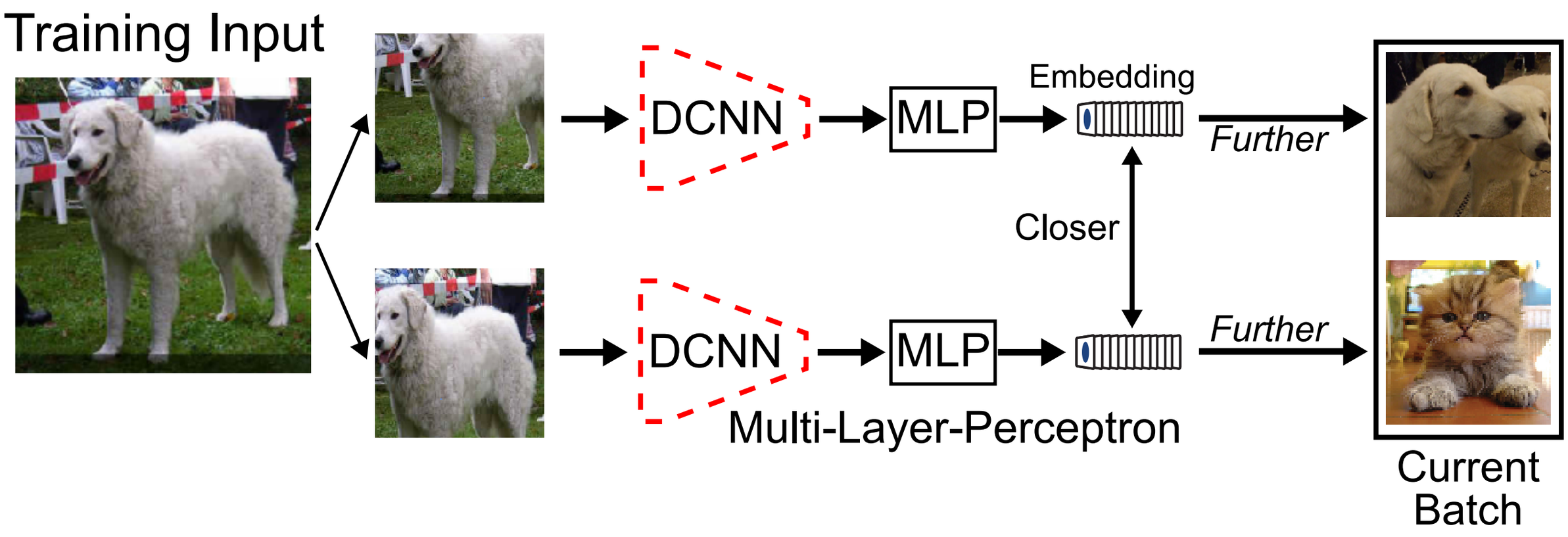
Abstract

This paper presents *SimCLR*: a simple framework for contrastive learning of visual representations. We simplify recently proposed contrastive self-supervised learning algorithms without requiring specialized architectures or a memory bank. In order to understand what enables the contrastive prediction tasks to learn useful representations, we systematically study the major components of our framework. We show that (1) composition of data augmentations plays a critical role in defining



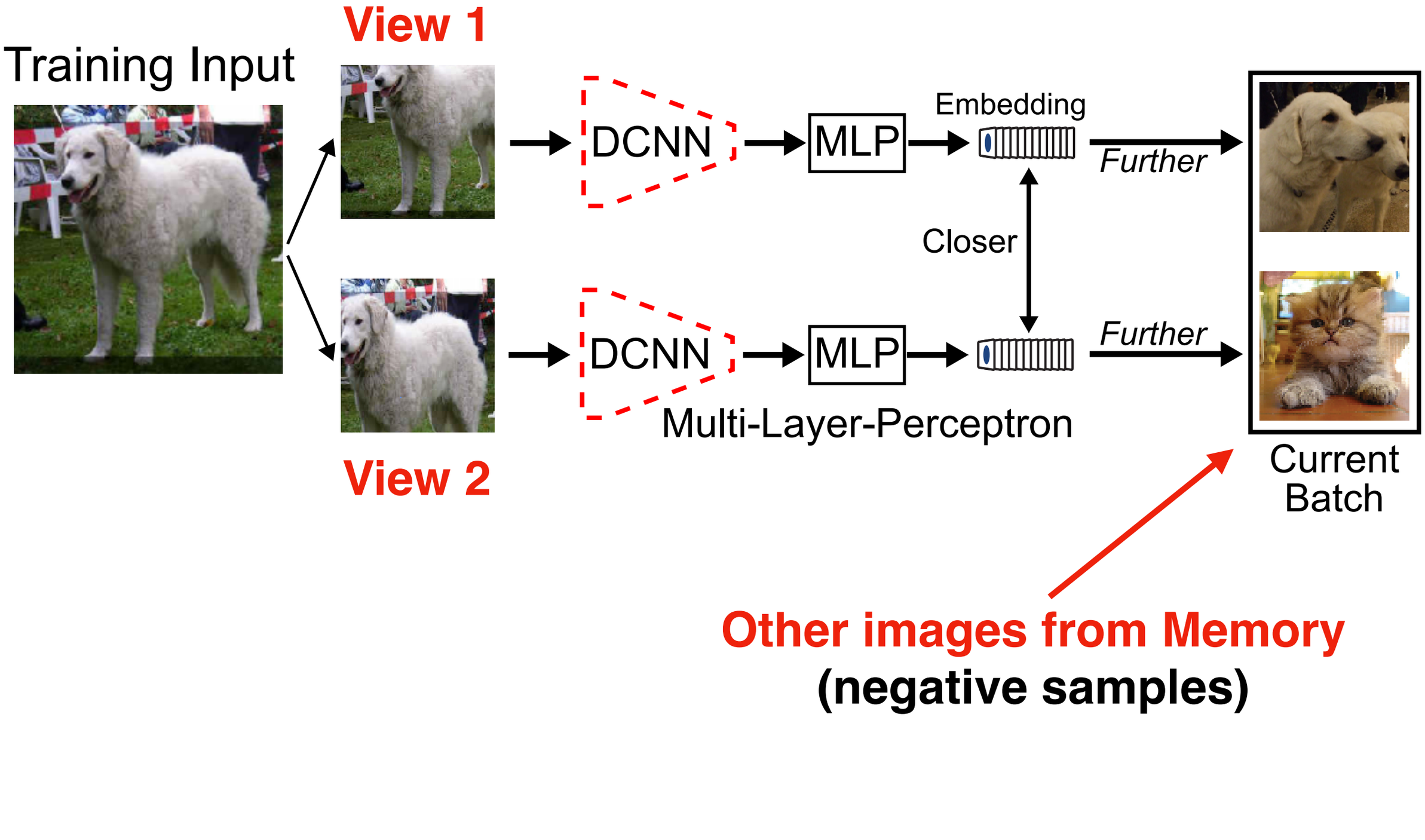
Chen et al. 2020

SimCLR



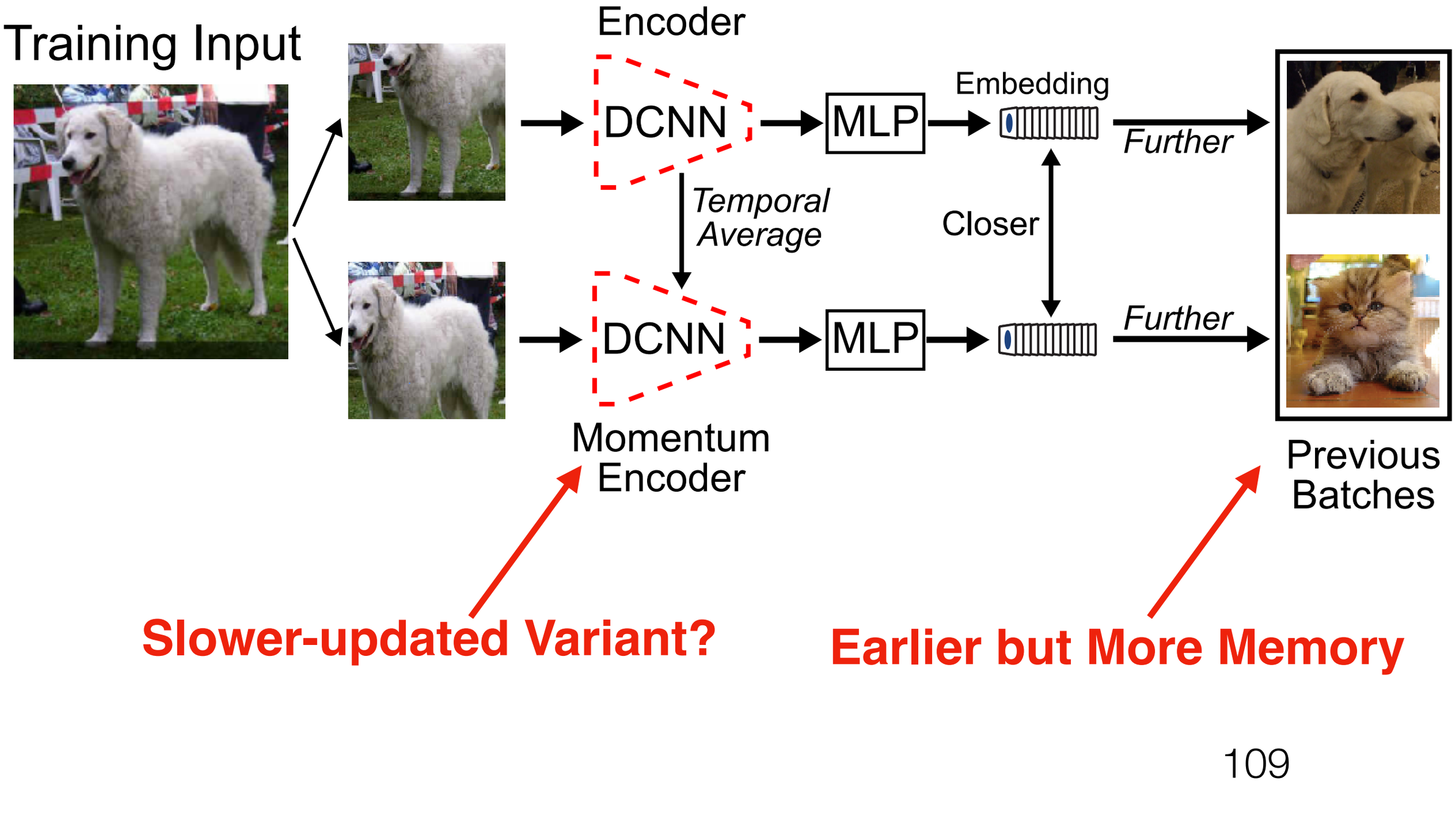
Chen et al. 2020

SimCLR

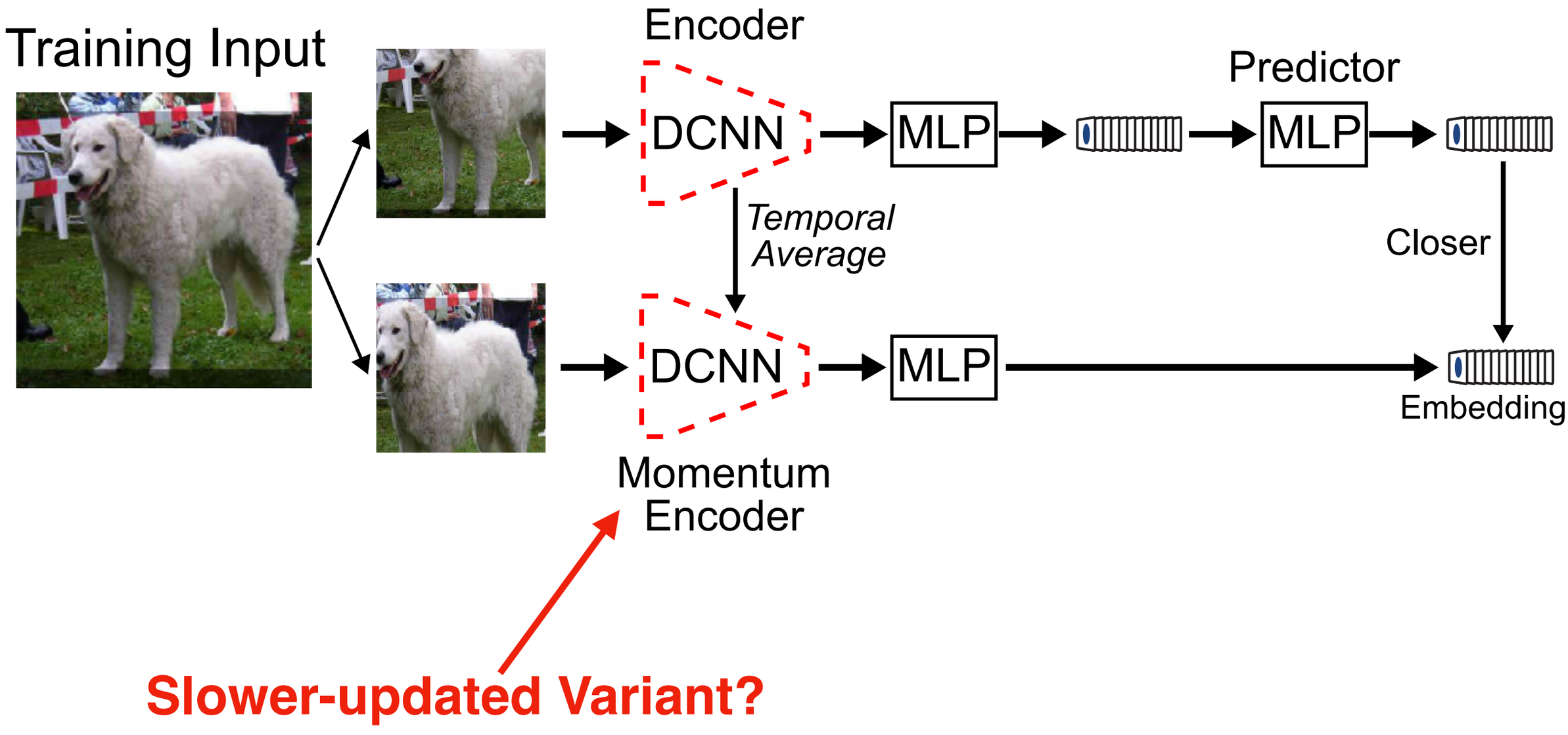


X. Chen et al. 2020

MoCo v2 (Momentum Contrast)



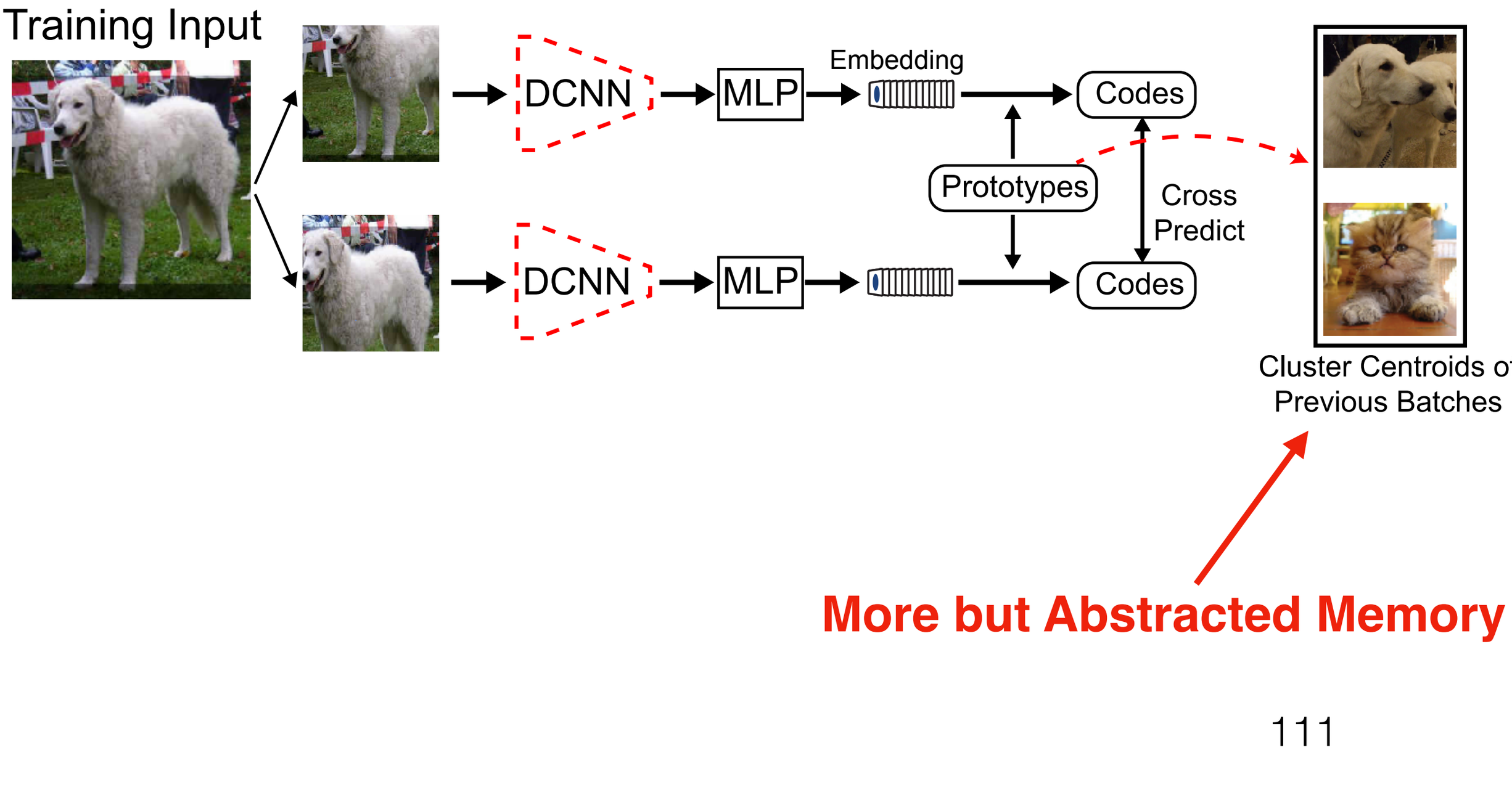
BYOL (Bootstrap Your Own Latent)



Slower-updated Variant?

Get rid of "memory" component (negative samples) due to implementation/hardware concerns

SwAV (Swapping Assignments between Views)



Masked Autoencoders (MAEs)

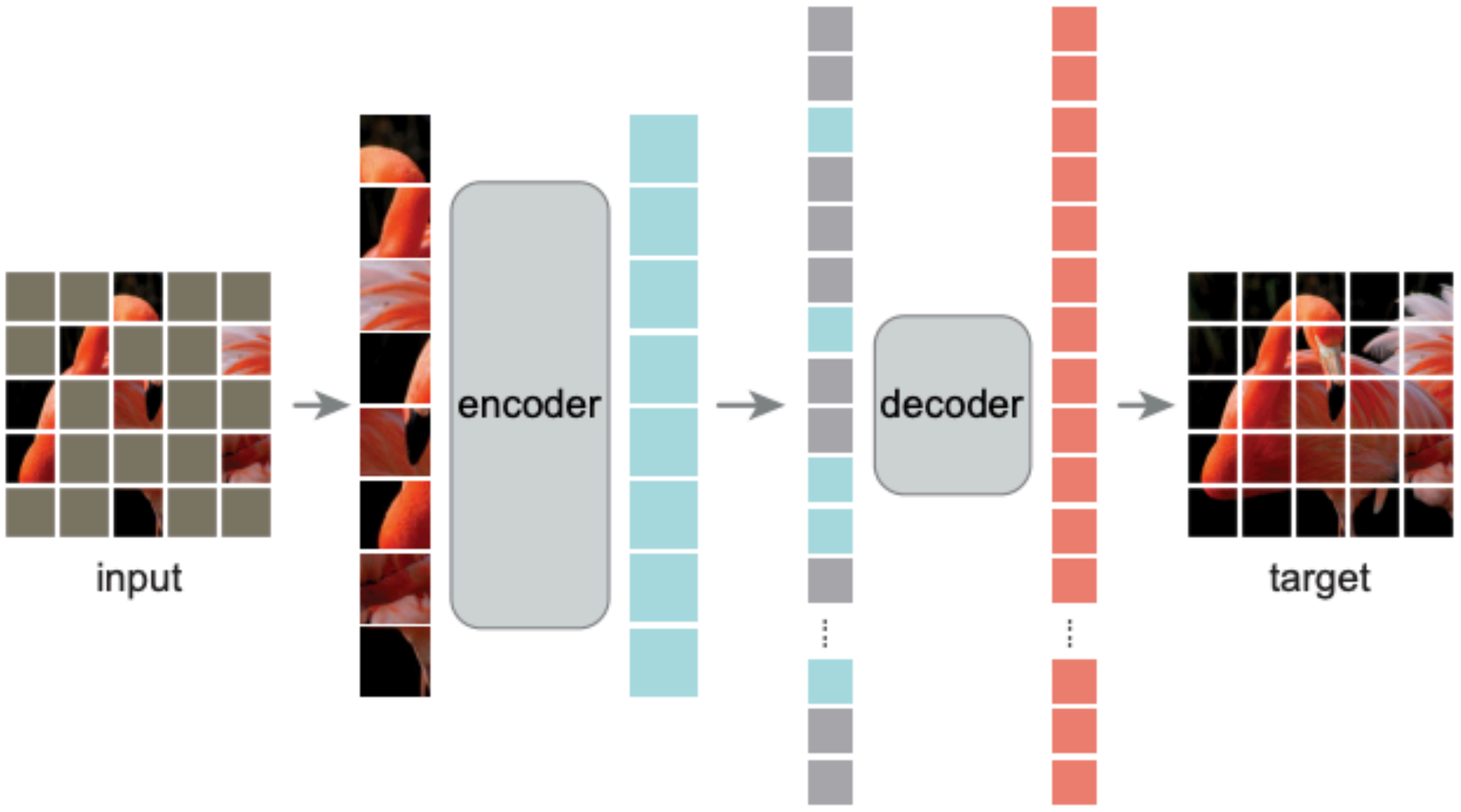
[Submitted on 11 Nov 2021 (v1), last revised 19 Dec 2021 (this version, v3)]

Masked Autoencoders Are Scalable Vision Learners

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

This paper shows that masked autoencoders (MAE) are scalable self-supervised learners for computer vision. Our MAE approach is simple: we mask random patches of the input image and reconstruct the missing pixels. It is based on two core designs. First, we develop an asymmetric encoder-decoder architecture, with an encoder that operates only on the visible subset of patches (without mask tokens), along with a lightweight decoder that reconstructs the original image from the latent representation and mask tokens. Second, we find that masking a high proportion of the input image, e.g., 75%, yields a nontrivial and meaningful self-supervisory task. Coupling these two designs enables us to train large models efficiently and effectively: we accelerate training (by 3x or more) and improve accuracy. Our scalable approach allows for learning high-capacity models that generalize well: e.g., a vanilla ViT-Huge model achieves the best accuracy (87.8%) among methods that use only ImageNet-1K data. Transfer performance in downstream tasks outperforms supervised pre-training and shows promising scaling behavior.

F(img-25%) → img-all



F(img-25%) → img-all

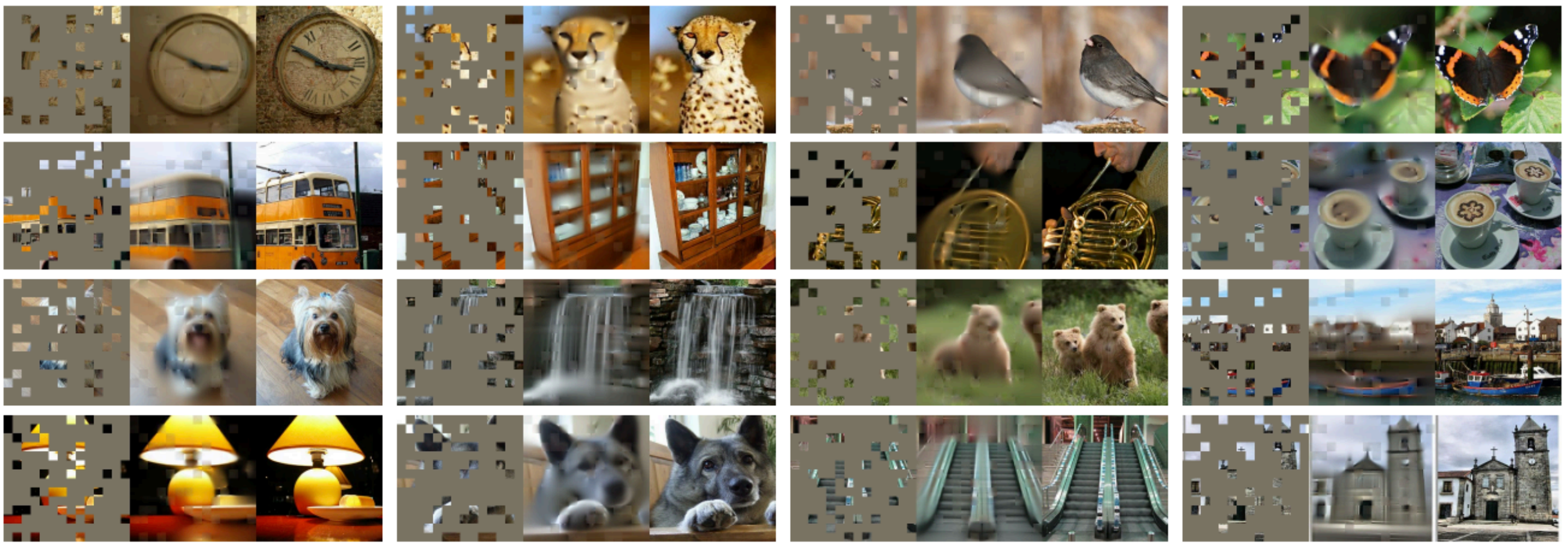


Figure 2. Example results on ImageNet *validation* images. For each triplet, we show the masked image (left), our MAE reconstruction† (middle), and the ground-truth (right). The masking ratio is 80%, leaving only 39 out of 196 patches. More examples are in the appendix.
†*As no loss is computed on visible patches, the model output on visible patches is qualitatively worse. One can simply overlay the output with the visible patches to improve visual quality. We intentionally opt not to do this, so we can more comprehensively demonstrate the method's behavior.*

Figure 3. Example results on COCO validation images, using an MAE trained on ImageNet (the same model weights as in Figure 2). Observe the reconstructions on the two right-most examples, which, although different from the ground truth, are semantically plausible.

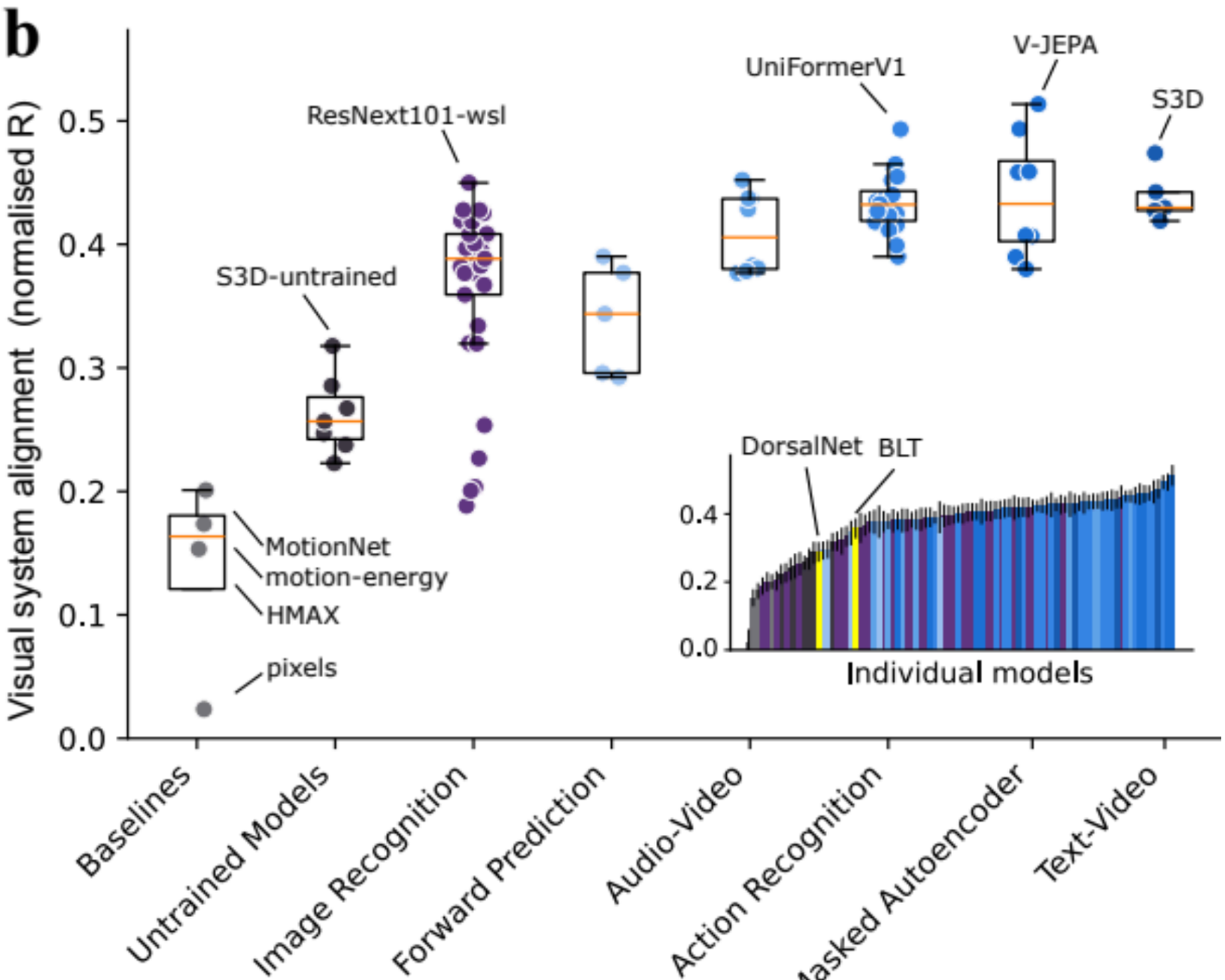
Trends:

Remove Memory

Remove contrasting

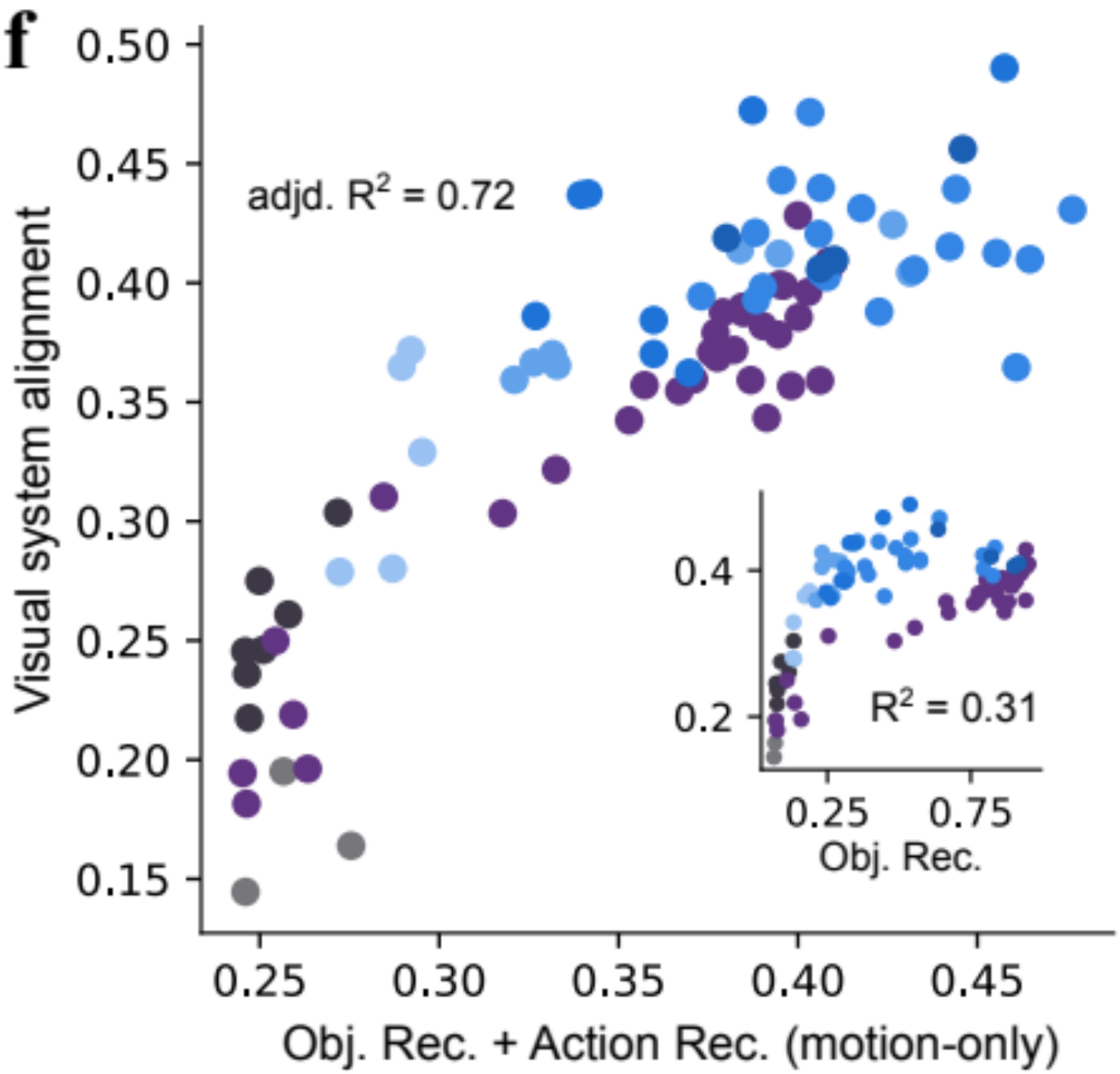
Diverse Perceptual Representations Across Visual Pathways Emerge from A Single Objective

Yingtian Tang[1], Abdulkadir Gokce[1], Khaled Jedoui Al-Karkari[2], Daniel Yamins[2], and Martin Schrimpf[1]



Diverse Perceptual Representations Across Visual Pathways Emerge from A Single Objective

Yingtian Tang[1], Abdulkadir Gokce[1], Khaled Jedoui Al-Karkari[2], Daniel Yamins[2], and Martin Schrimpf[1]



Problem: Current algorithms trained on
existing developmentally-appropriate datasets
don't learn very strong representations.

Learning from real kids' data is a harder problem than learning from ImageNet because:

1. online vs buffered/randomized
2. many fewer distinct examples
3. but from wider variety of viewpoints

Problem: Current algorithms trained on **existing developmentally-appropriate datasets** don't learn very strong representations.

Two Main Hypotheses:

Problem: Current algorithms trained on
existing developmentally-appropriate datasets
don't learn very strong representations.

Two Main Hypotheses:

The **algorithms** are insufficient

vs.

Problem: Current algorithms trained on **existing developmentally-appropriate datasets** don't learn very strong representations.

Two Main Hypotheses:

The **algorithms** are insufficient

vs.

The **data** are insufficient

Problem: Current algorithms trained on
existing developmentally-appropriate datasets
don't learn very strong representations.

Two Main Hypotheses:

The **algorithms** are insufficient

vs.

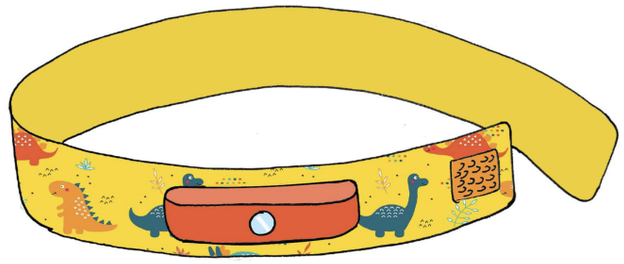
The **data** are insufficient

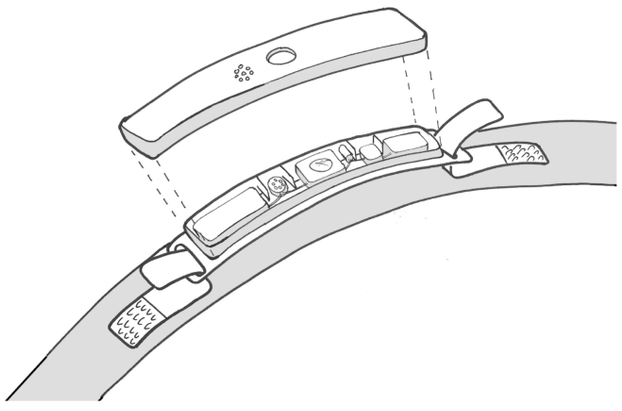
Until very recently, not enough data to know.

Our strategy: Get more data!

BabyCam++

40 Bay Area families
6 months - 3 years
Recording ~5 hours/week
Custom high resolution babycam video+accelerometer
Unprecedented resource for studying development



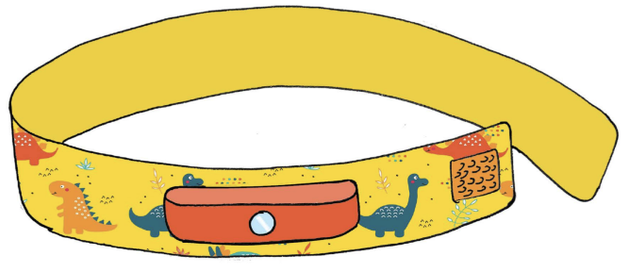


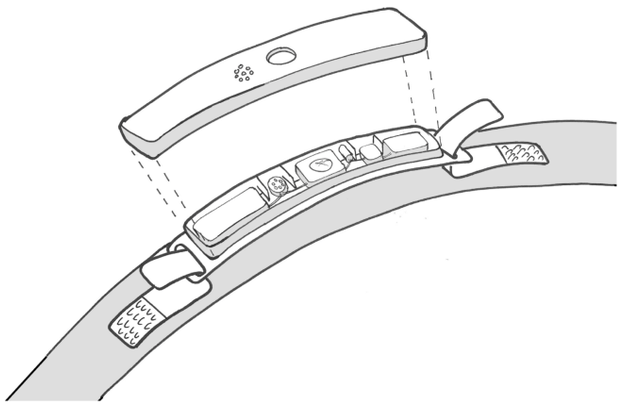
SAYCam (~0.1 child-years)

Our strategy: Get more data!

BabyCam++

40 Bay Area families
6 months - 3 years
Recording ~5 hours/week
Custom high resolution
babycam video+accelerometer
Unprecedented resource for studying development





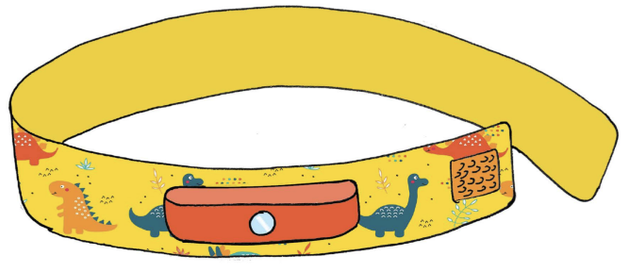
SAYCam (~0.1 child-years)

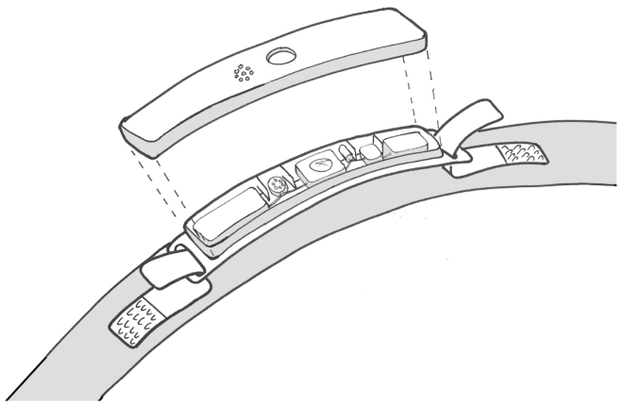
1 child-year

Our strategy: Get more data!

BabyCam++

40 Bay Area families
6 months - 3 years
Recording ~5 hours/week
Custom high resolution babycam video+accelerometer
Unprecedented resource for studying development





The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Bria Long

now starting her lab at UCSD!

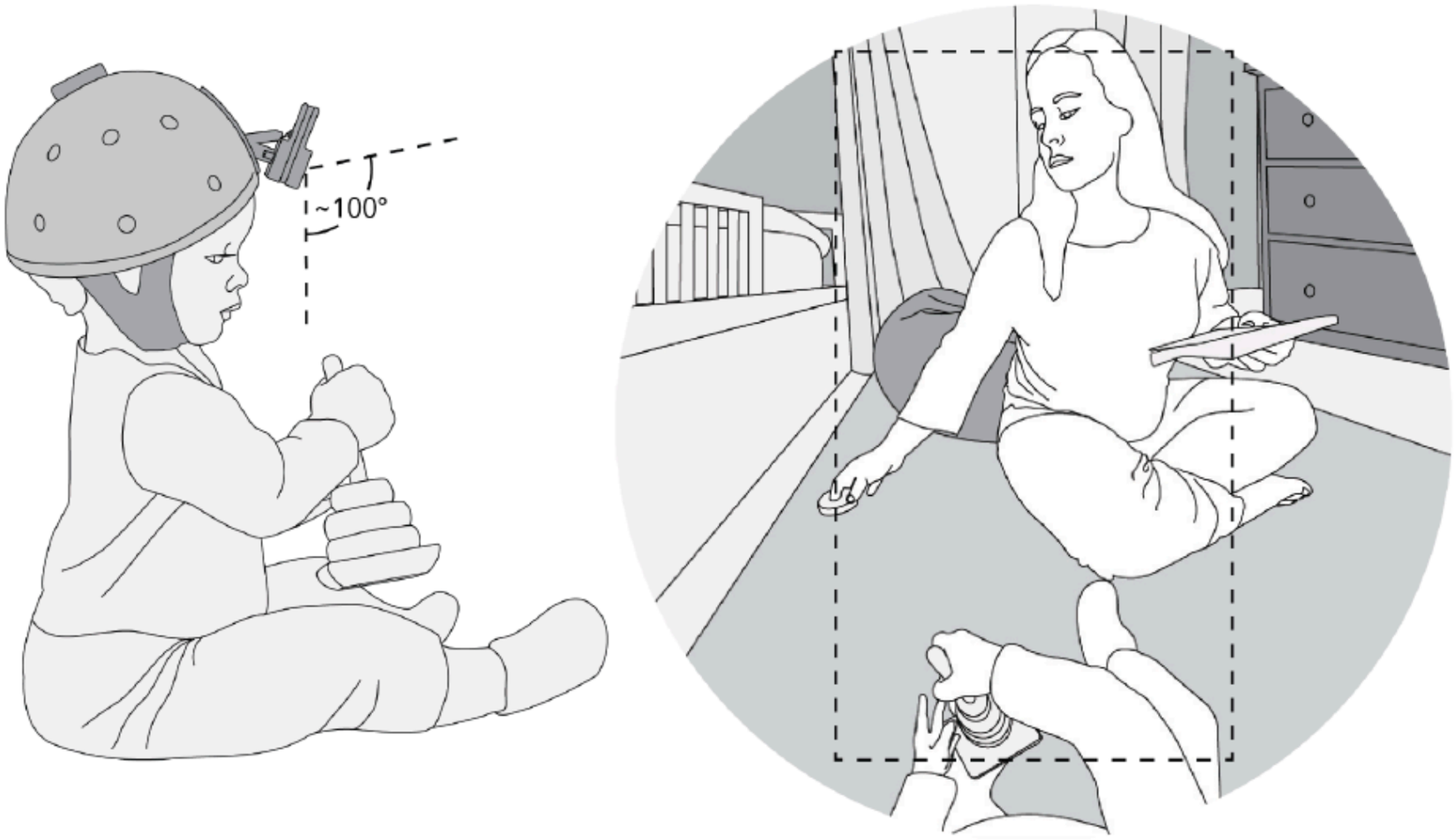


Figure 1. Schematic illustration of the BabyView camera's orientation (left) and field of view (right; dotted line), highlighting that this camera angle captures both the objects that children are interacting with as well as the social information in the child's view. See Figures 3 and 5 for example images.

Figure 2. Overview of the BabyView Camera design process, showing (a) the assembled device, (b) the original camera, (c) babysafe helmet, and (d) and 3D printed mounting equipment.

Figure 5. Example images and off-the-shelf Mask-R CNN segmentations (confidence > .3) on frames from the BabyView camera. These higher-resolution egocentric images provide better data for segmentation than previous cameras, yet are still quite challenging for state-of-the-art models.

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

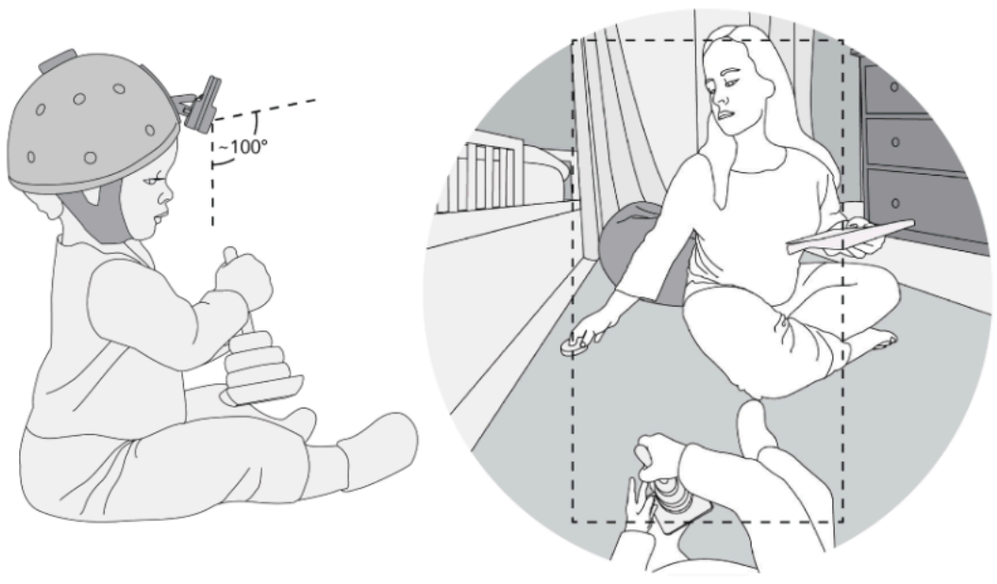


Figure 1. Schematic illustration of the BabyView camera's orientation (left) and field of view (right; dotted line), highlighting that this camera angle captures both the objects that children are interacting with as well as the social information in the child's view. See Figures 3 and 5 for example images.

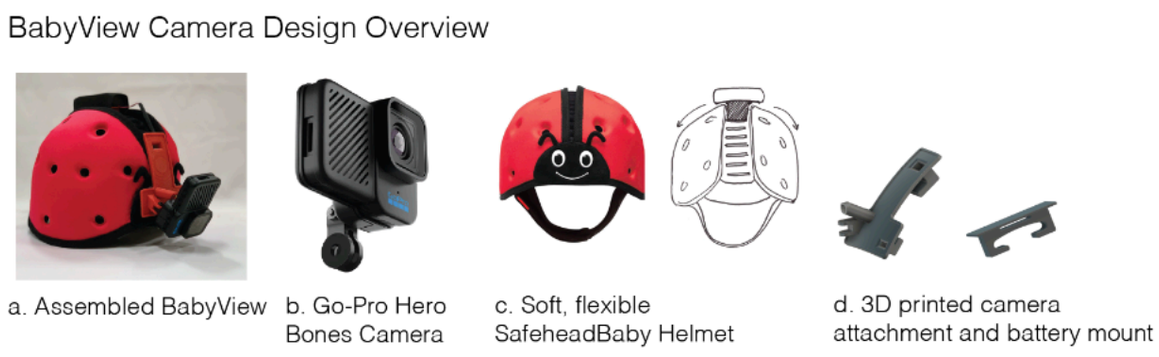


Figure 2. Overview of the BabyView Camera design process, showing (a) the assembled device, (b) the original camera, (c) babysafe helmet, and (d) and 3D printed mounting equipment.

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

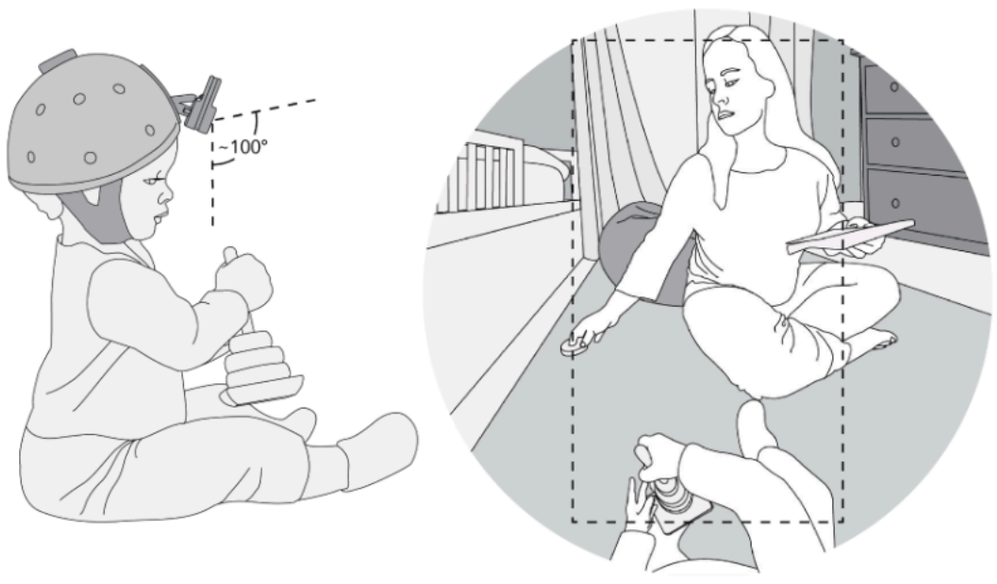


Figure 1. Schematic illustration of the BabyView camera's orientation (left) and field of view (right; dotted line), highlighting that this camera angle captures both the objects that children are interacting with as well as the social information in the child's view. See Figures 3 and 5 for example images.

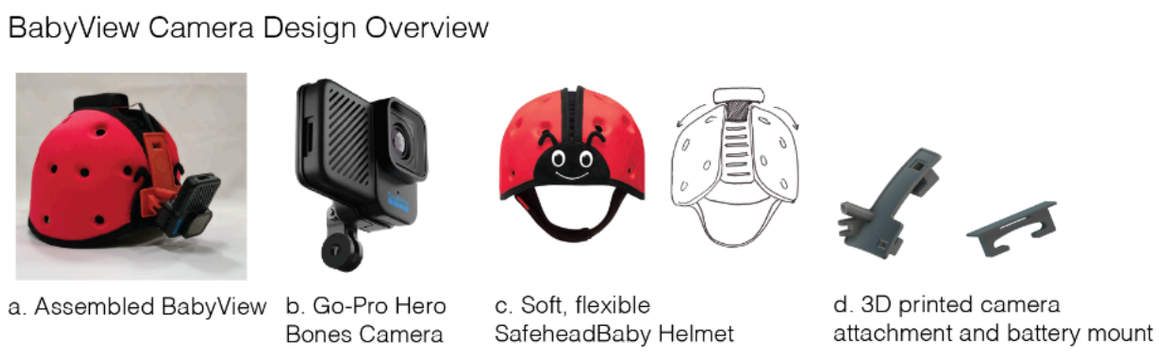


Figure 2. Overview of the BabyView Camera design process, showing (a) the assembled device, (b) the original camera, (c) babysafe helmet, and (d) and 3D printed mounting equipment.

1000+ hours
Audio/Video/Gyroscope
Text Transcript

`BabyView v1.0` data to be released Sep. 2024

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

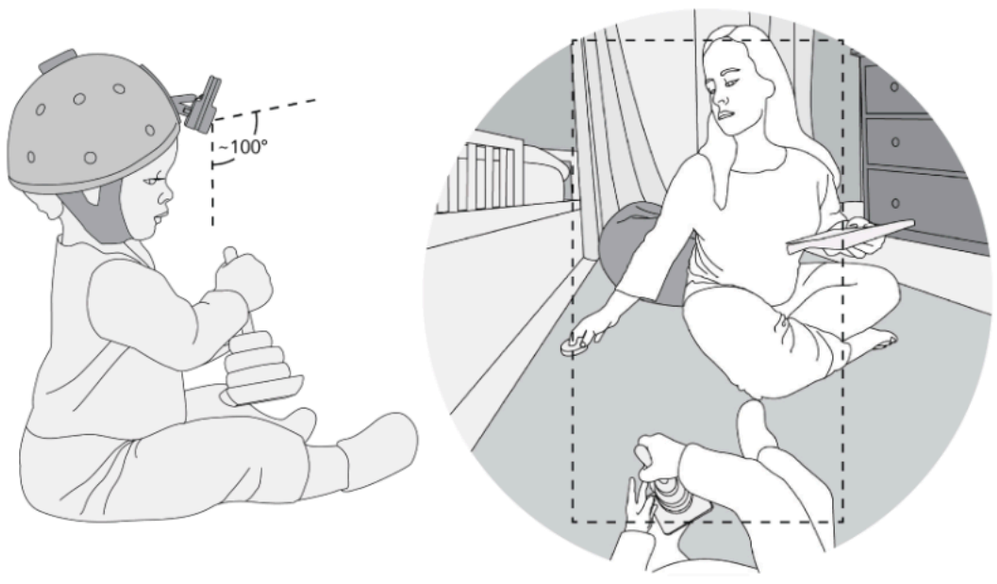


Figure 1. Schematic illustration of the BabyView camera's orientation (left) and field of view (right; dotted line), highlighting that this camera angle captures both the objects that children are interacting with as well as the social information in the child's view. See Figures 3 and 5 for example images.

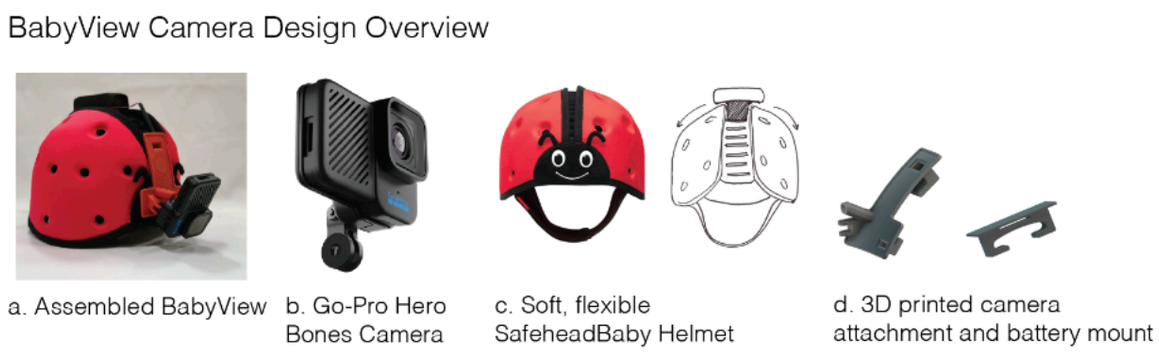


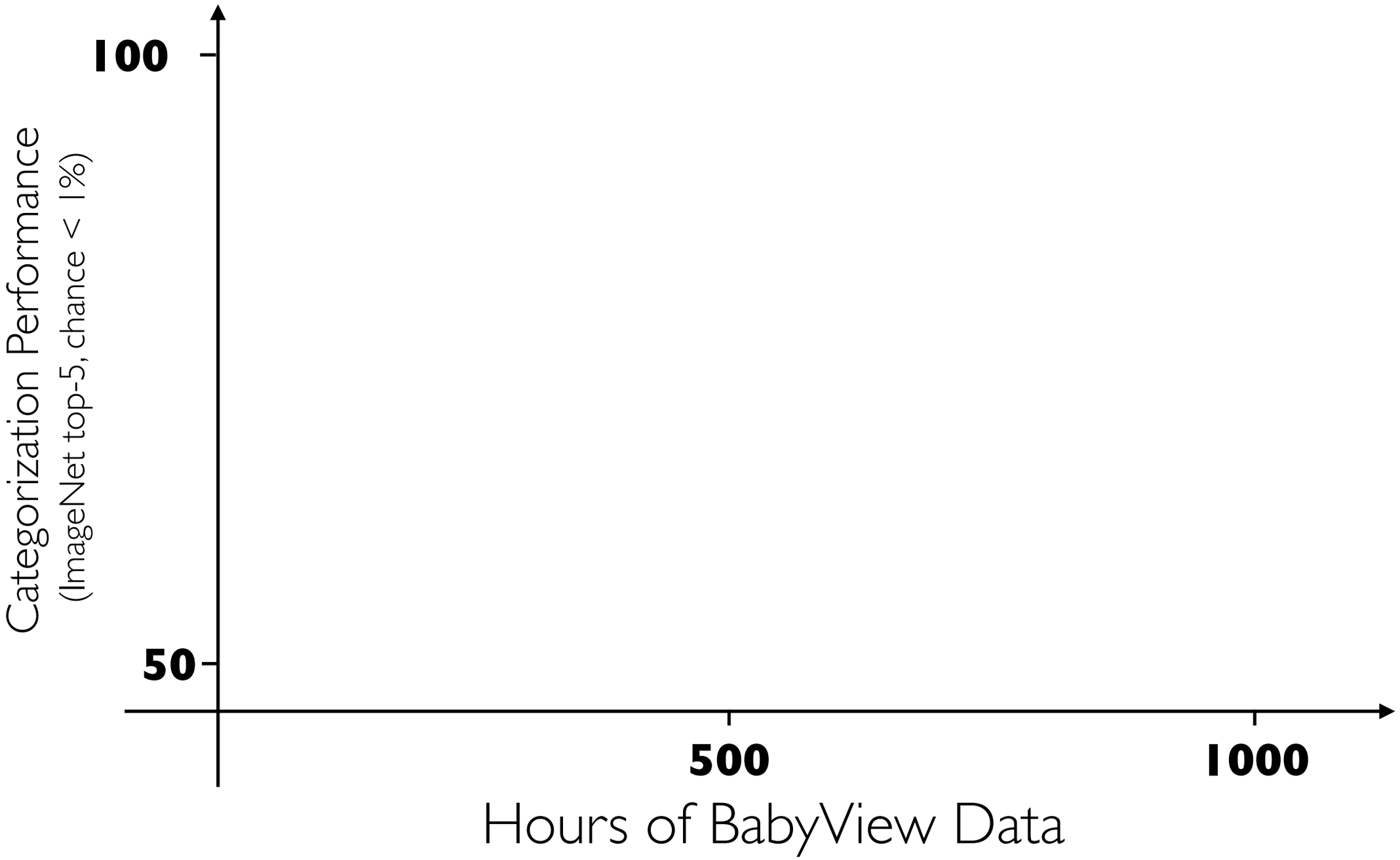
Figure 2. Overview of the BabyView Camera design process, showing (a) the assembled device, (b) the original camera, (c) babysafe helmet, and (d) and 3D printed mounting equipment.

4000+ hours
Audio/Video/Gyroscope
Text Transcript

`BabyView v2.0` data to be released Sep. 2025

Stefan Stojanov

Initial ('hot off the press') results:



PsyArXiv Preprints

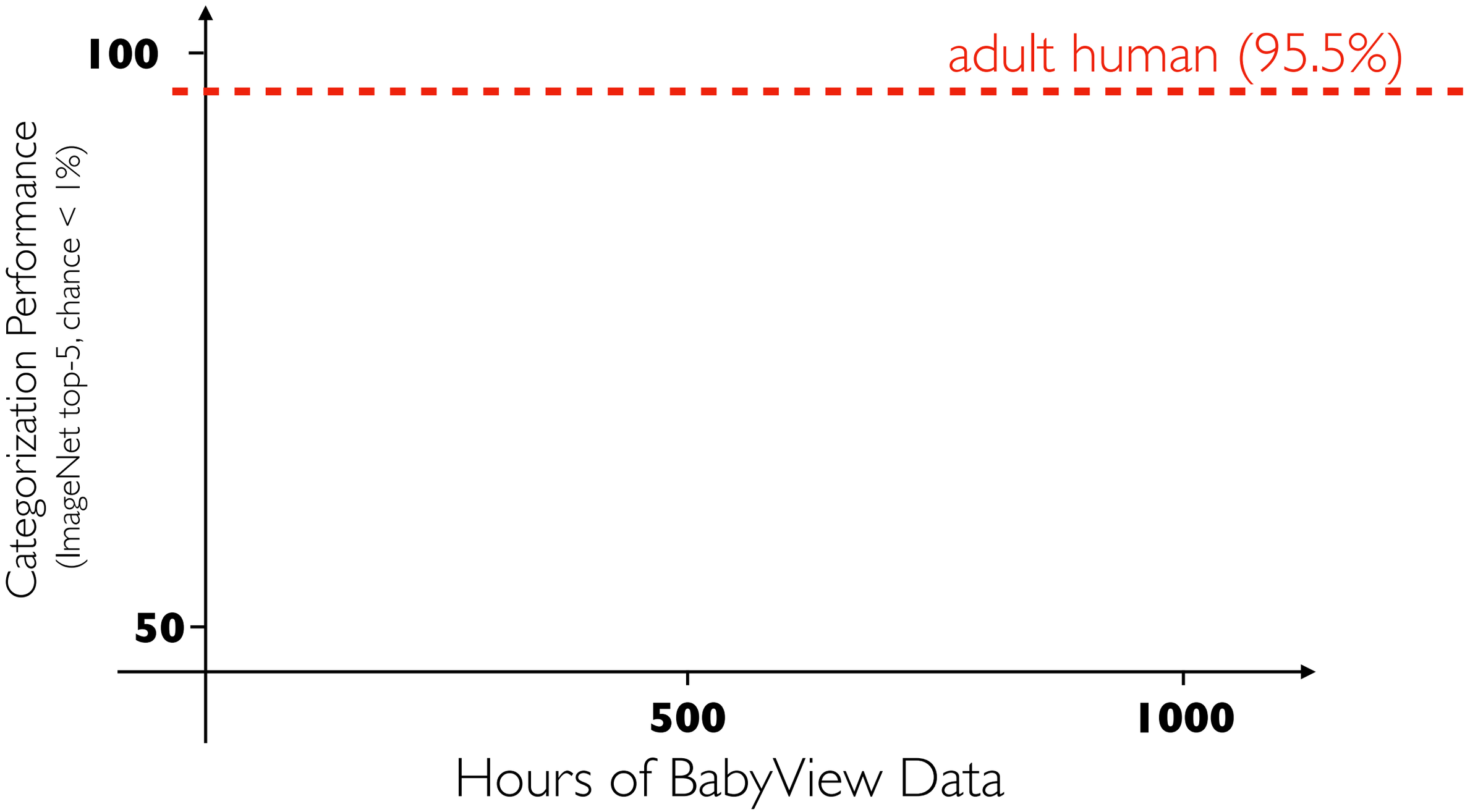
My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov

Initial ('hot off the press') results:



PsyArXiv Preprints

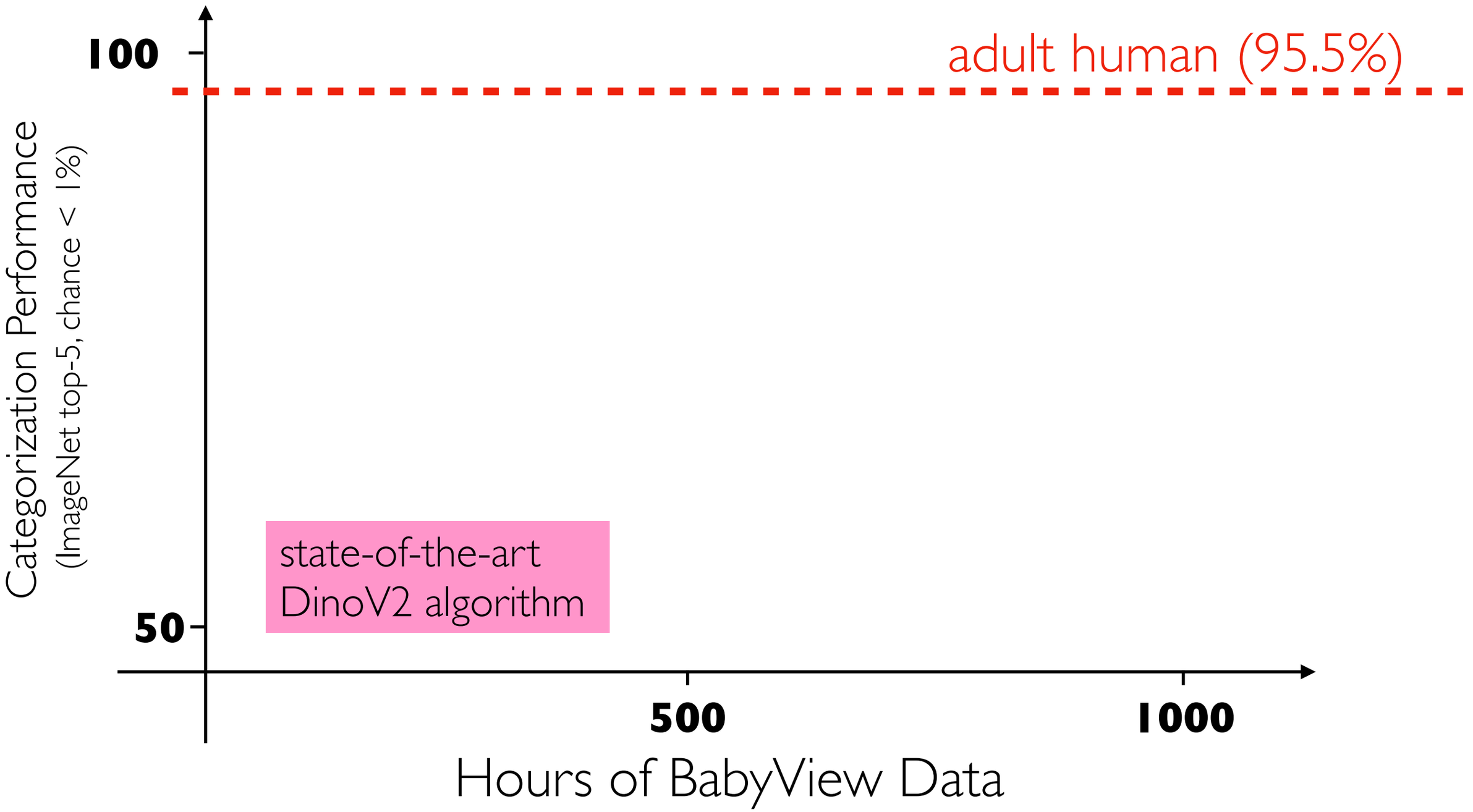
My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov

Initial ('hot off the press') results:



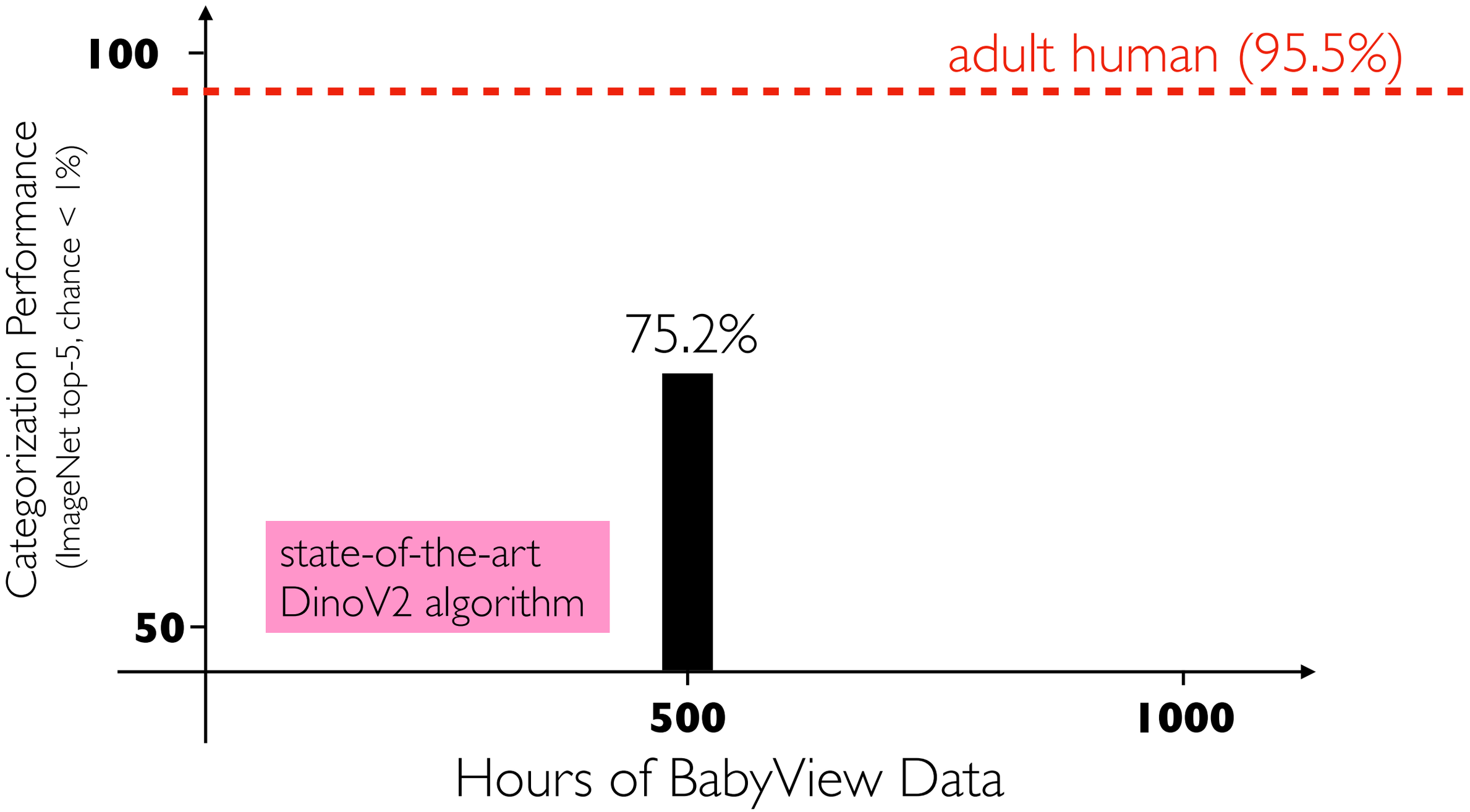
PsyArXiv Preprints My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov

Initial ('hot off the press') results:



PsyArXiv Preprints

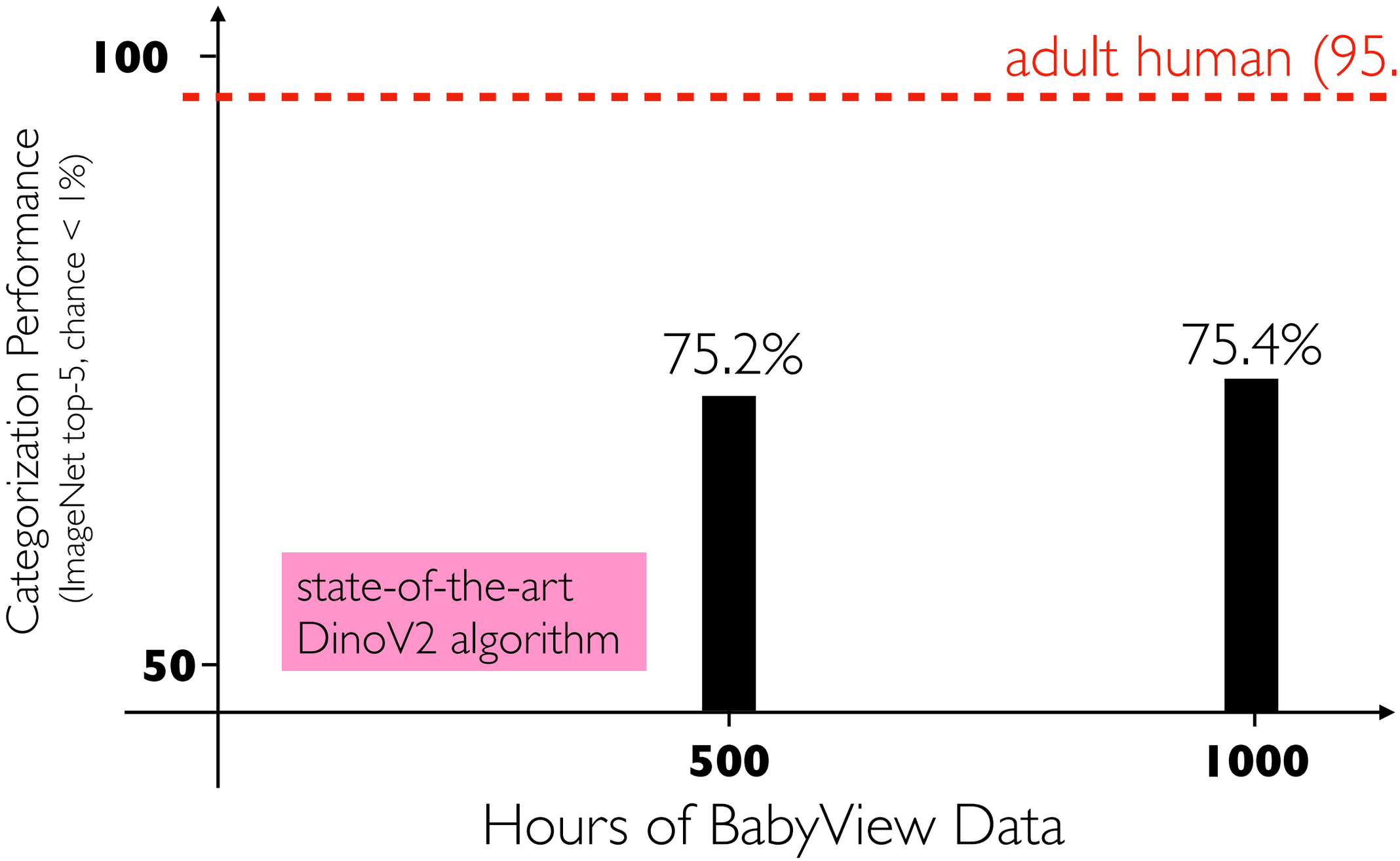
My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov

Initial ('hot off the press') results:



PsyArXiv Preprints

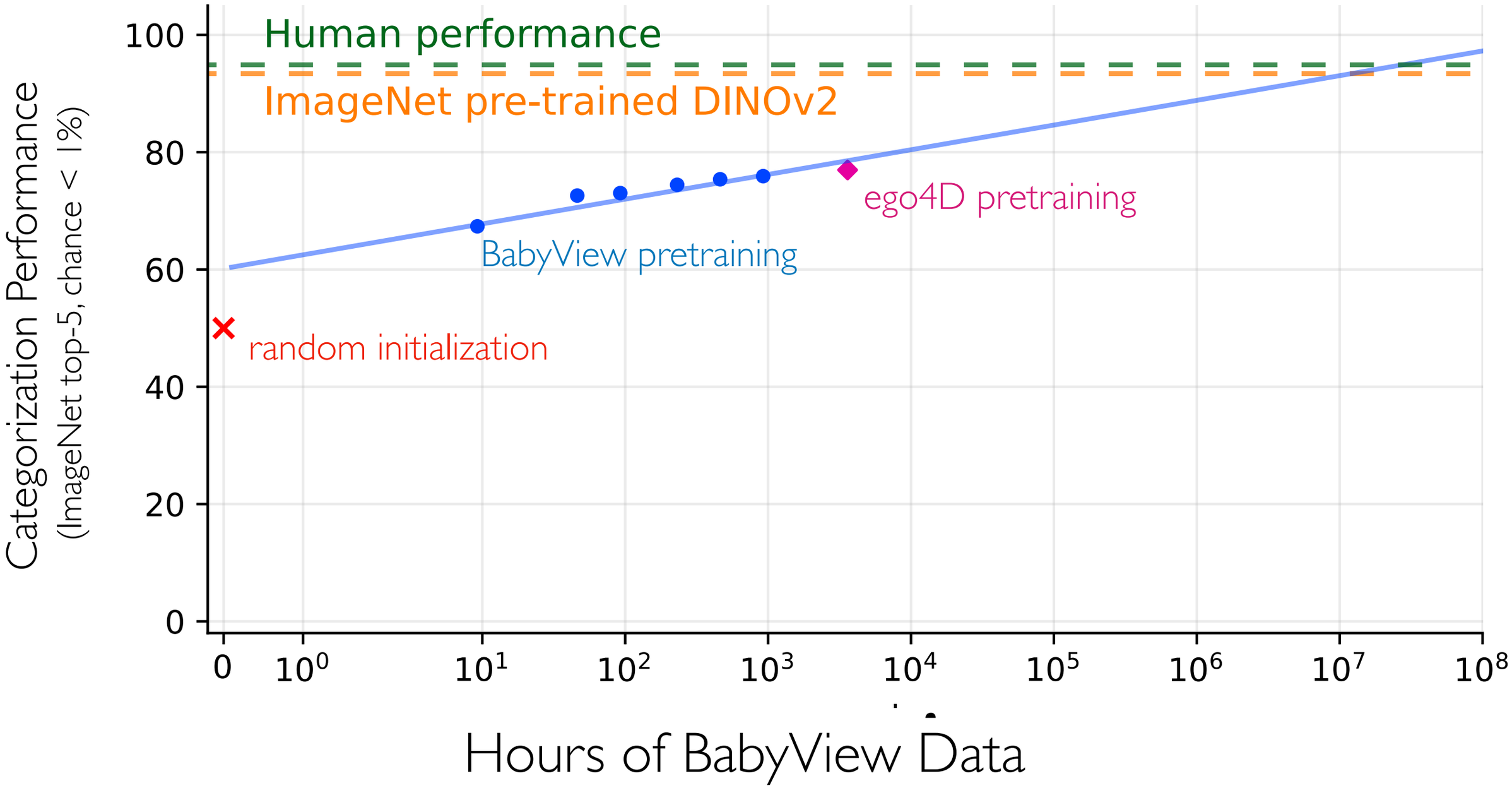
My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov

Categorization



PsyArXiv Preprints

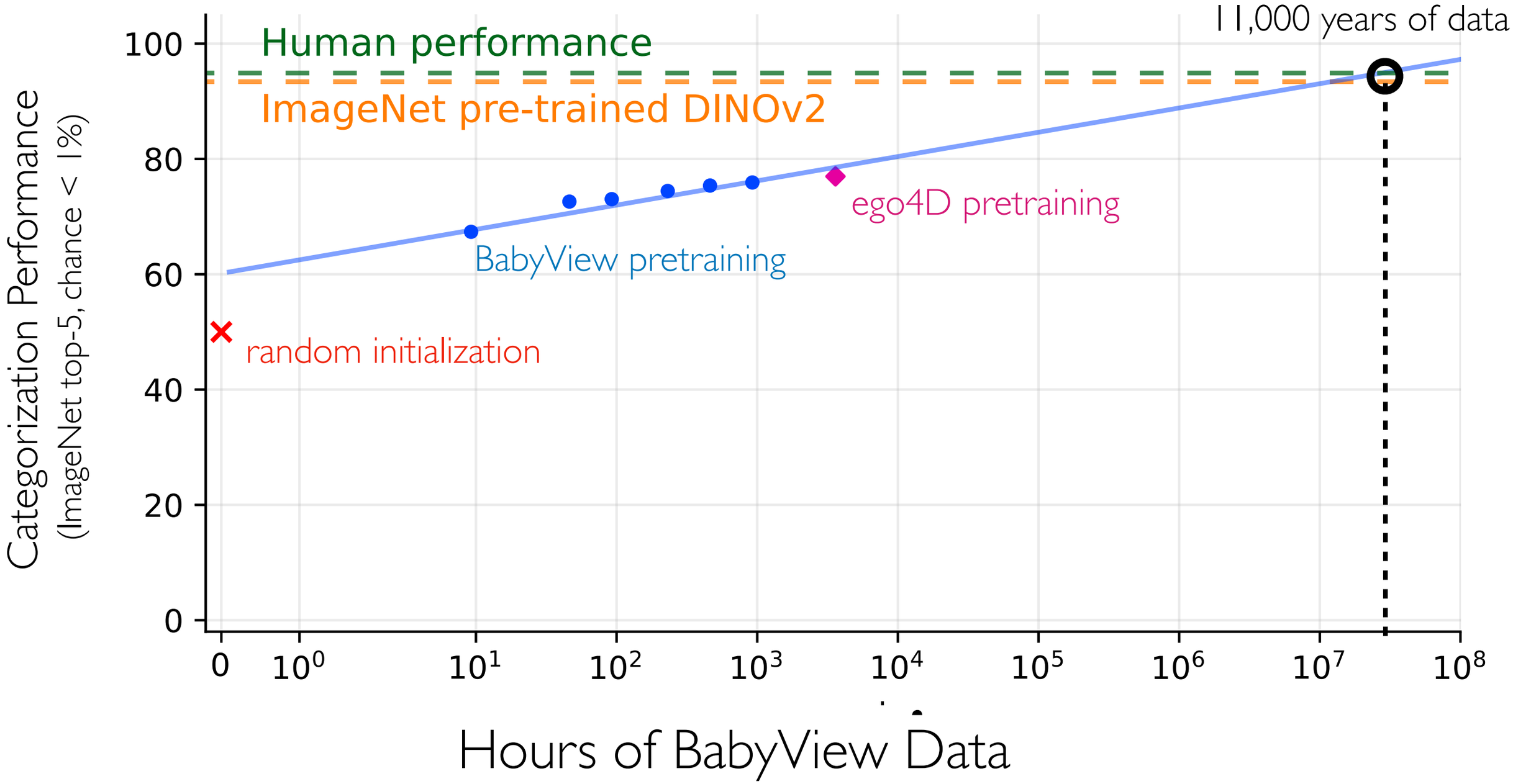
My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov

Categorization



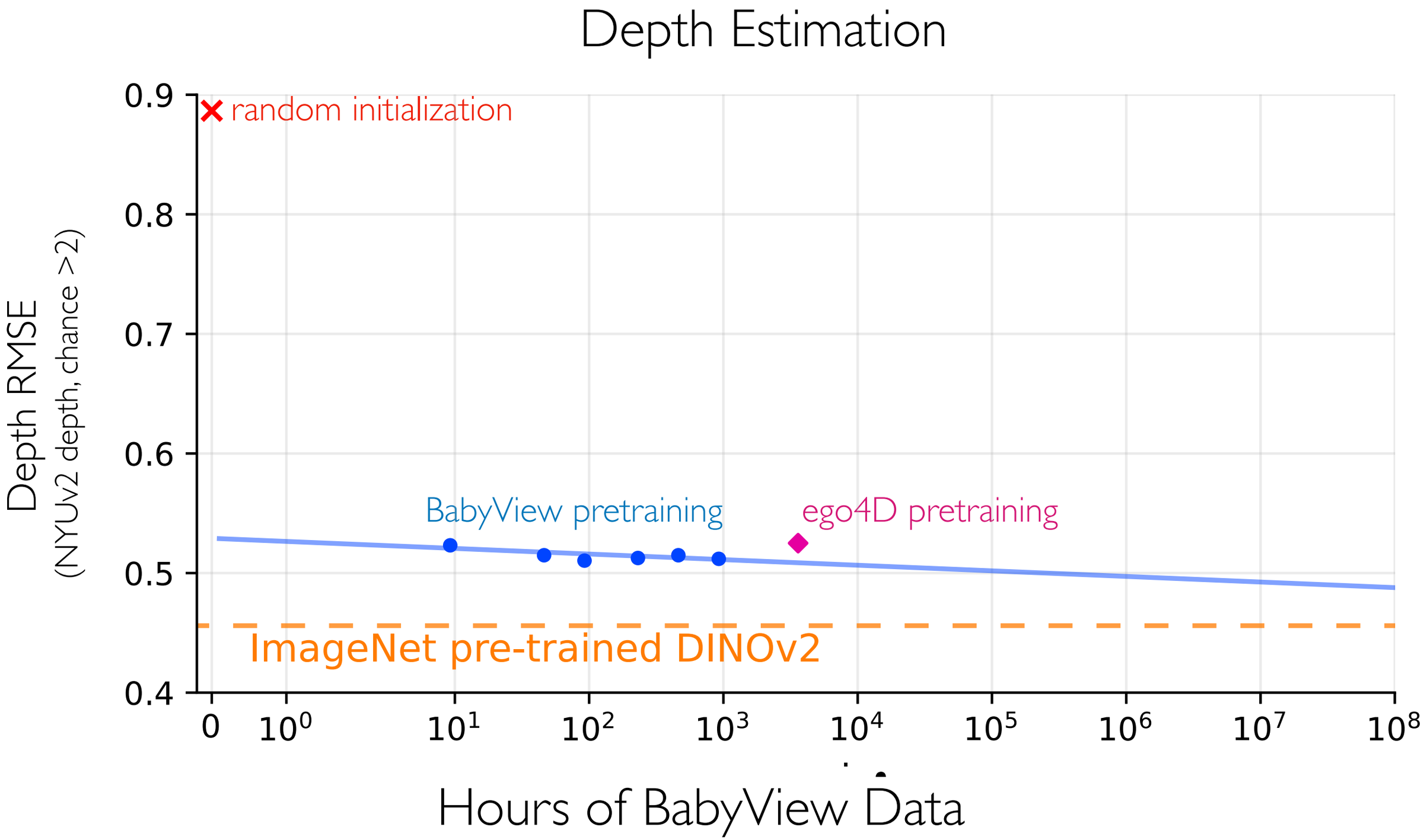
PsyArXiv Preprints

My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov



PsyArXiv Preprints

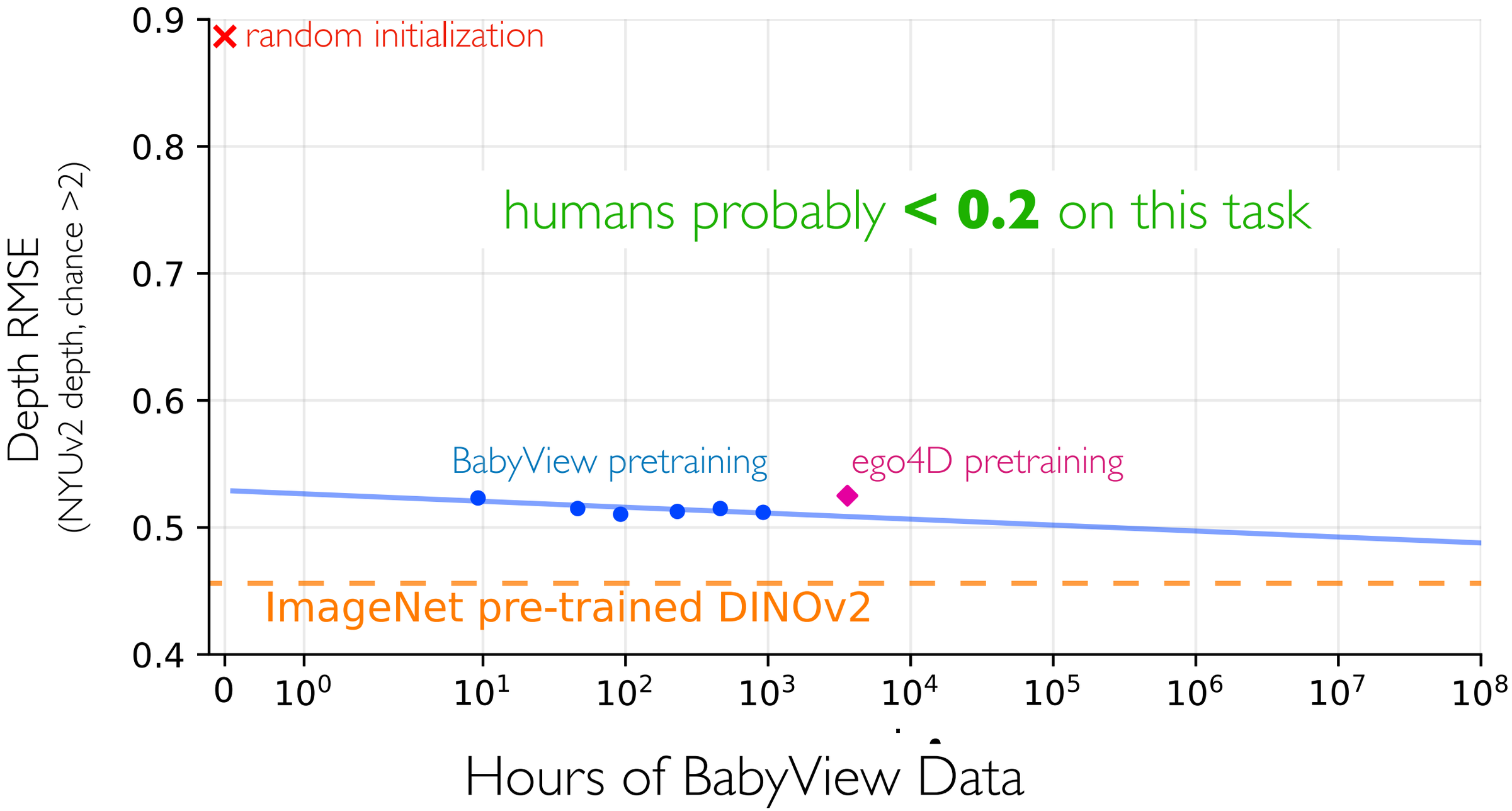
My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov

Depth Estimation



PsyArXiv Preprints

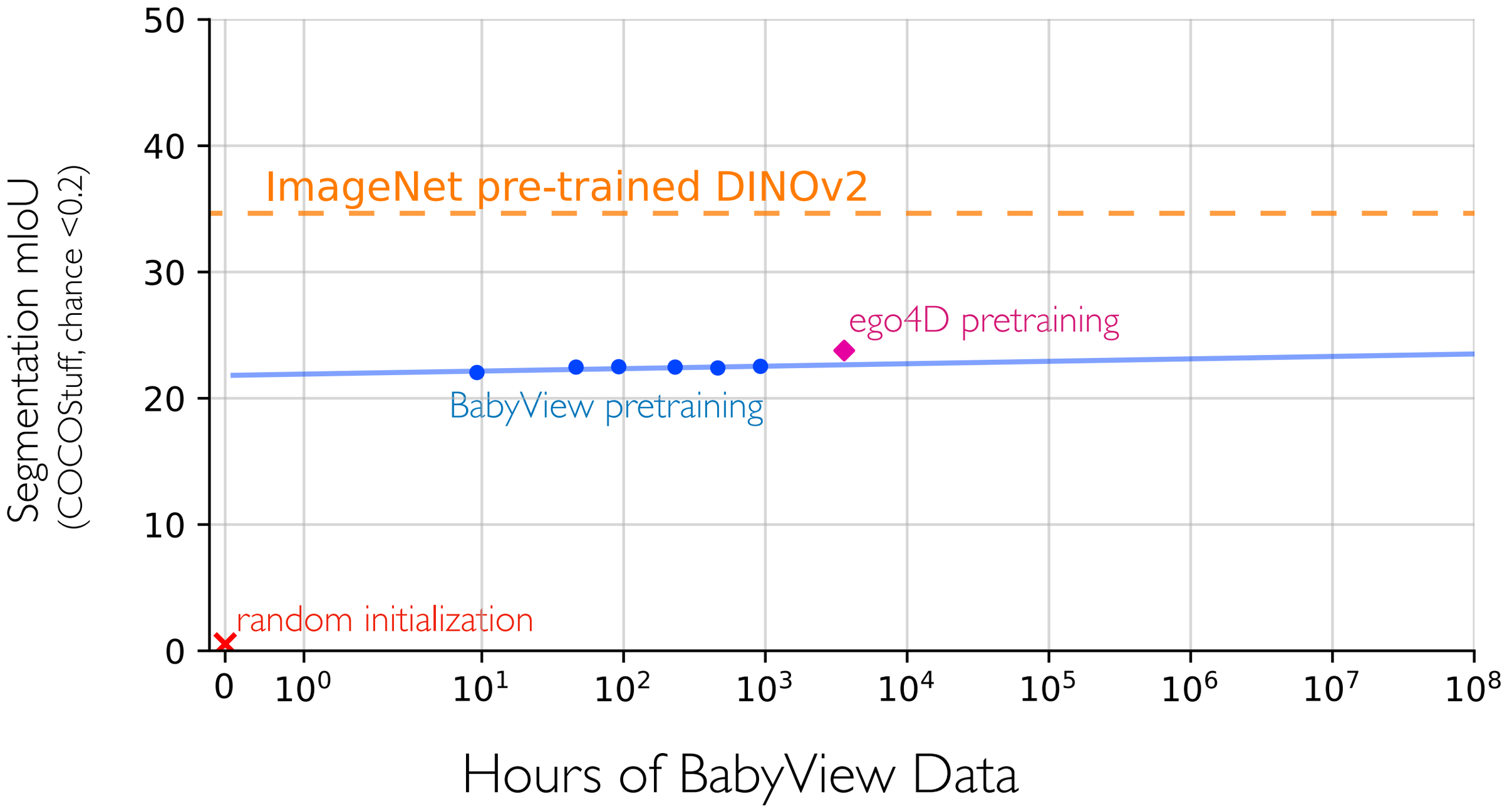
My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov

Basic Semantic Segmentation

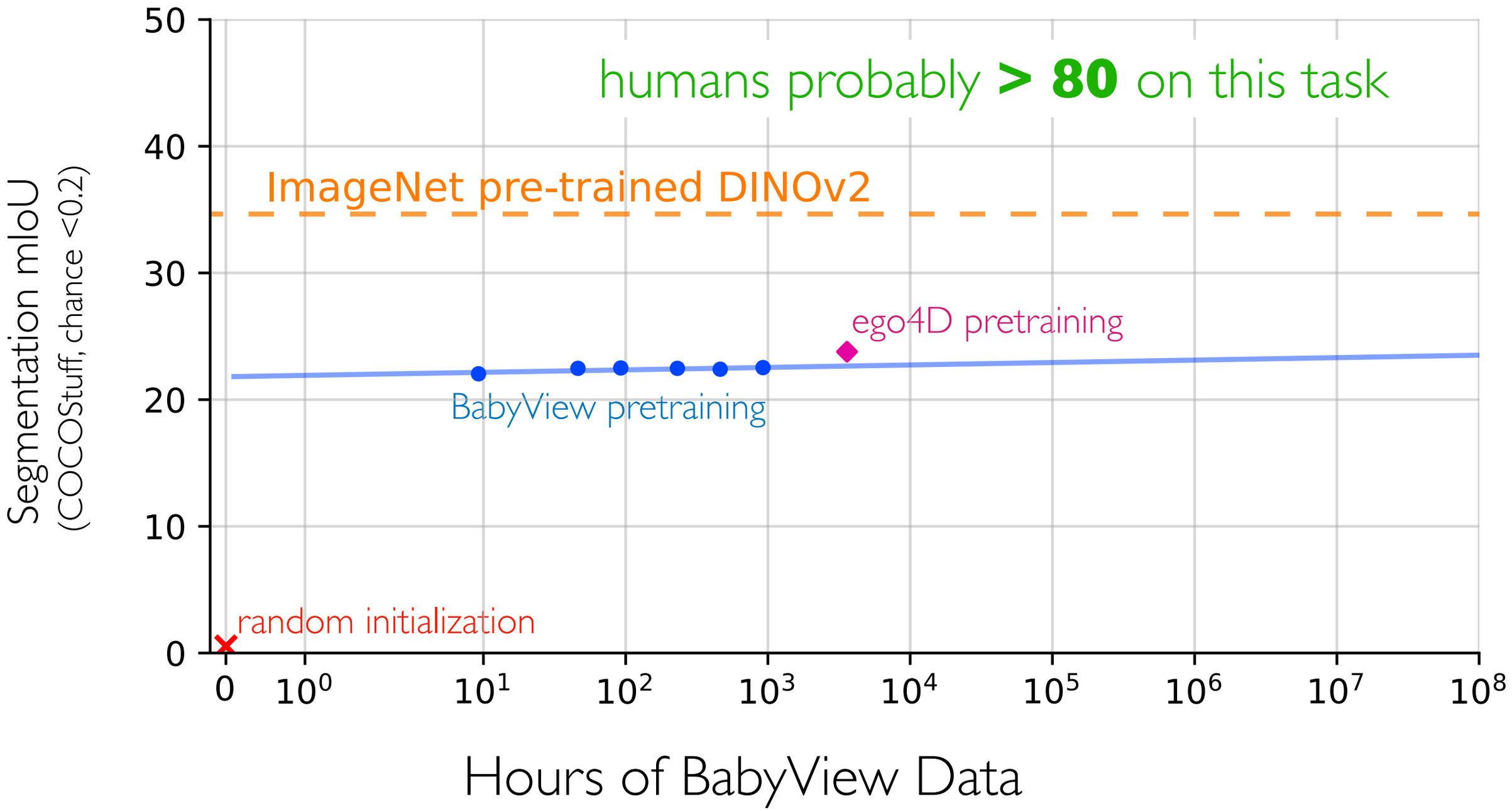


PsyArXiv Preprints My Preprints Add a preprint Donate Sign Up Sign In

The BabyView Camera: Designing a New Head-mounted Camera to Capture Children's Early Social and Visual Environment

AUTHORS
Bria Long, Sarah Goodin, George Kachergis, Virginia A. Marchman, Samaher Radwan, Robert Z. Sparks, Violet Xiang, Chengxu Zhuang, Oliver Hsu, Brett Newman, Daniel Yamins, and Michael C. Frank

Stefan Stojanov



A **new hypothesis** is *probably* needed. Like what??

0. You just need *more* (developmental, egocentric) data.

A **new hypothesis** is *probably* needed. Like what??

~~0. You just need *more* (developmental, egocentric) data.~~

A **new hypothesis** is *probably* needed. Like what??

~~0. You just need *more* (developmental, egocentric) data.~~

1. You need other *kinds* of data:

A **new hypothesis** is *probably* needed. Like what??

~~0. You just need *more* (developmental, egocentric) data.~~

1. You need other *kinds* of data:

    a. other modalities

A **new hypothesis** is *probably* needed. Like what??

~~0. You just need *more* (developmental, egocentric) data.~~

1. You need other *kinds* of data:

    a. other modalities

        — audio-visual? (but blind/deaf people learn fine?)

A **new hypothesis** is *probably* needed. Like what??

~~0. You just need *more* (developmental, egocentric) data.~~

1. You need other *kinds* of data:

    a. other modalities

        — audio-visual? (but blind/deaf people learn fine?)

        — language? (but monkeys learn vision fine?)

A **new hypothesis** is *probably* needed. Like what??

~~0. You just need *more* (developmental, egocentric) data.~~

1. You need other *kinds* of data:

    a. other modalities

        — audio-visual? (but blind/deaf people learn fine?)

        — language? (but monkeys learn vision fine?)

    b. embodiment e.g. action streams/joint policy learning (not clear evidence for this)

A **new hypothesis** is *probably* needed. Like what??

~~0. You just need *more* (developmental, egocentric) data.~~

1. You need other *kinds* of data:

    a. other modalities

        — audio-visual? (but blind/deaf people learn fine?)

        — language? (but monkeys learn vision fine?)

    b. embodiment e.g. action streams/joint policy learning (not clear evidence for this)

2. You need other kinds of algorithms.