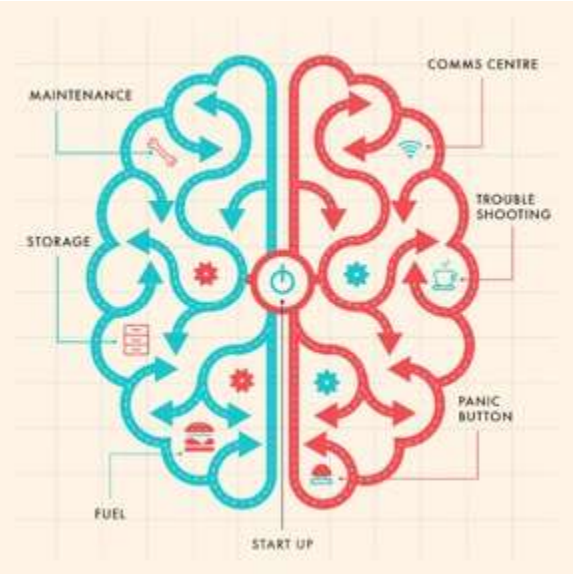
Deep Neural Networks for Object Recognition: Which is the most Brain-like?

Giuseppe di Pellegrino
Department of Psychology, University of Bologna

g.dipellegrino@unibo.it



Cognition and Neuroscience
Second cycle Degree in Artificial Intelligence – 2024/25

The Ventral Visual Stream

Starting with the seminal work of Hubel and Wiesel in the 1960s, research in visual systems neuroscience has shown that the brain actively reformats incoming sensory data to better serve behavioral goals.

In primates, invariant object recognition arises through a hierarchically organized series of cortical areas called the ventral visual stream.

The first 100 ms after a visual stimulus triggers a rapid, cascading wave of neural activity across these areas.

Challenges in Understanding Higher Ventral Areas

Despite progress, explaining neural encoding in higher ventral areas remains an open challenge.

The brain's possible transformations from retinal input to behavior are vast.

One approach to understanding object recognition is to test whether artificial neural networks (ANNs) can serve as functional analogs of biological processes.

Yamins et al., PNAS , 2014

Historical Development of HCNNs

Hierarchical Convolutional Neural Networks (HCNNs) stem from early models (with hand-designing parameters) like Fukushima (1980), with major contributions by LeCun & Bengio (1995), Riesenhuber & Poggio (1999), and Rumelhart.

Though inspired by Hubel & Wiesel, early models could not fully solve core object recognition.

A major shift occurred in 2012 when ANNs, inspired by the ventral stream, began to achieve primate-level object categorization (Krizhevsky et al., 2012)

.

Models of higher ventral areas should

- image-computable and fully runnable
- capable of providing information useful for supporting human-level behavioral performance
- mappable (i.e., layers correspond to distinct regions within the visual system, such as V1, V2, V4, IT)
- neurally predictive (at the single-unit level, and at the neural population level)

Yamins et al., PNAS , 2014

Neuroscience & AI: Key Points of Intersection

an **architecture class** capturing neuroanatomical knowledge;

- AI: CNNs, RNNs, ViTs
- Neuro: Visual cortex, hippocampus, etc.

an **objective function** capturing hypotheses about the signals driving learning;

- AI: Loss functions (e.g., error minimization)
- Neuro: Reward signals, predictive coding

a **training dataset** capturing the environment in which the system learns;

- AI: Curated datasets (e.g., ImageNet)
- Neuro: Sensory input during development

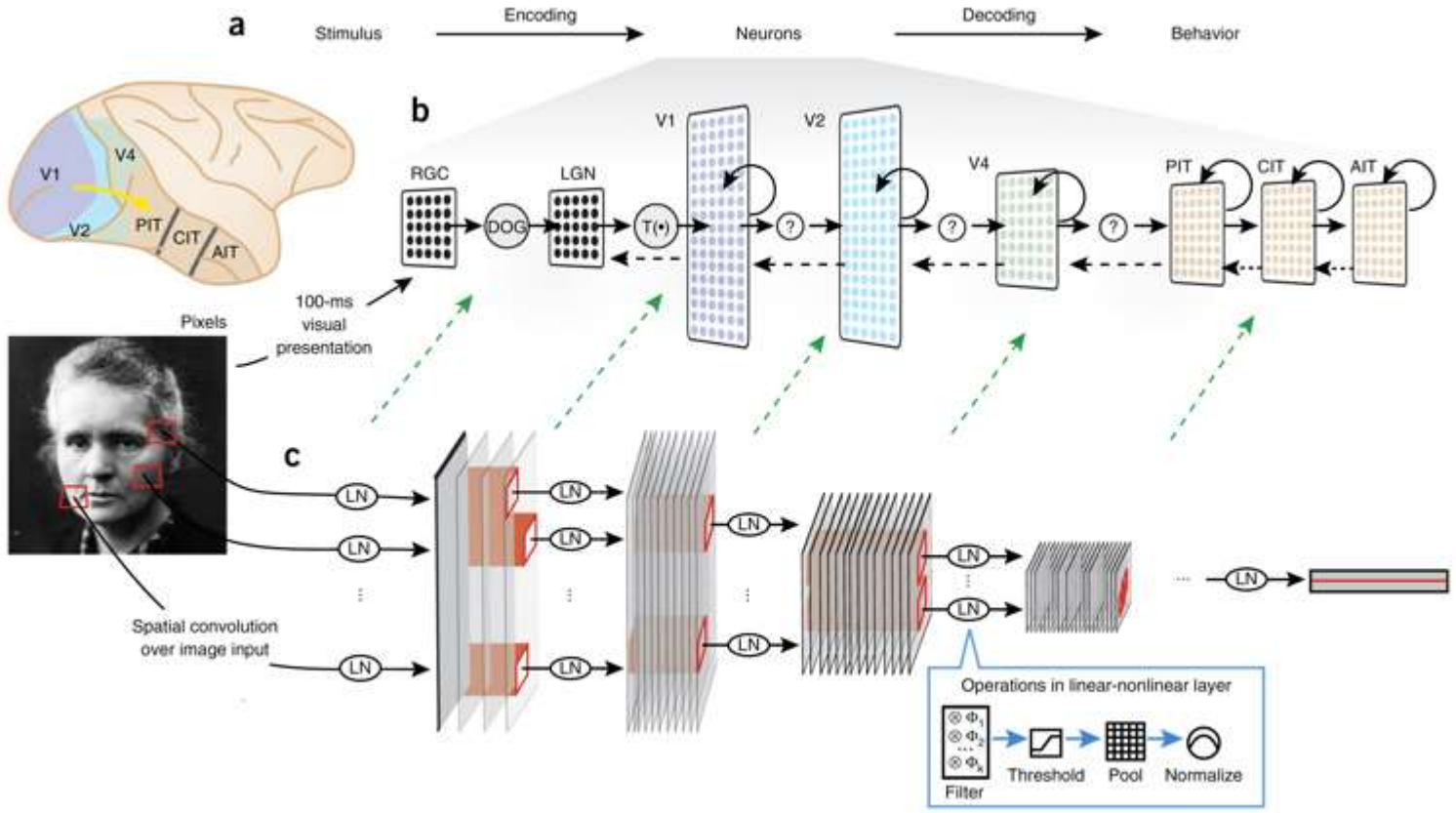
a **learning rule** for actually converting learning signals into system updates, capturing neural plasticity.

- AI: Backpropagation, gradient descent
- Neuro: Hebbian learning, STDP (Spike-Timing Dependent Plasticity)

Hierarchical convolutional neural networks (HCNNs)

HCNNs are good candidates for models of the ventral visual pathway and have achieved near-human-level performance on challenging object categorization tasks.

HCNNs are multilayer neural networks, arranged in series, each of whose layers performing linear-nonlinear (LN) transformation on the input data, analogous to the transformation produced in the ventral stream.



Key operations:

(i) filtering, a linear operation that takes the dot product of local patches in the input stimulus with a set of templates (convolution);

(ii) activation, a nonlinearity—typically either a rectified linear threshold (ReLU) or a sigmoid;

(iii) pooling, a nonlinear aggregation operation—typically the mean or maximum of local values;

(iv) divisive normalization, correcting output values to a standard range;

Hierarchical Depth and Spatial Transformation

Convolutional layers preserve spatial layout, allowing deep stacking.

With depth:
- Receptive fields grow.
- Networks become less retinotopic due to striding/pooling.
- Filter numbers increase, shifting from **wide & shallow to deep & narrow.**

- Eventually, spatial resolution shrinks, requiring fully connected layers for final classification outputs.

Inductive Bias in HCNNs

HCNNs have a built-in **inductive bias** favoring local, adjacent spatial processing (locality).

This reflects the natural structure of images, where nearby pixels are often related.

Early layers focus on local features; deeper layers integrate them into global representations.

This mirrors biological vision, progressing from edge detection to complex object recognition, guided by the hierarchical architecture.

Yamins et al. (2014)

Combining computational and electrophysiology techniques (single-unit recordings), Yamins et al. (2014) explored a wide range of biologically plausible HCNN models and then assessed them against measured IT and V4 neural response data, as well as human performance data.

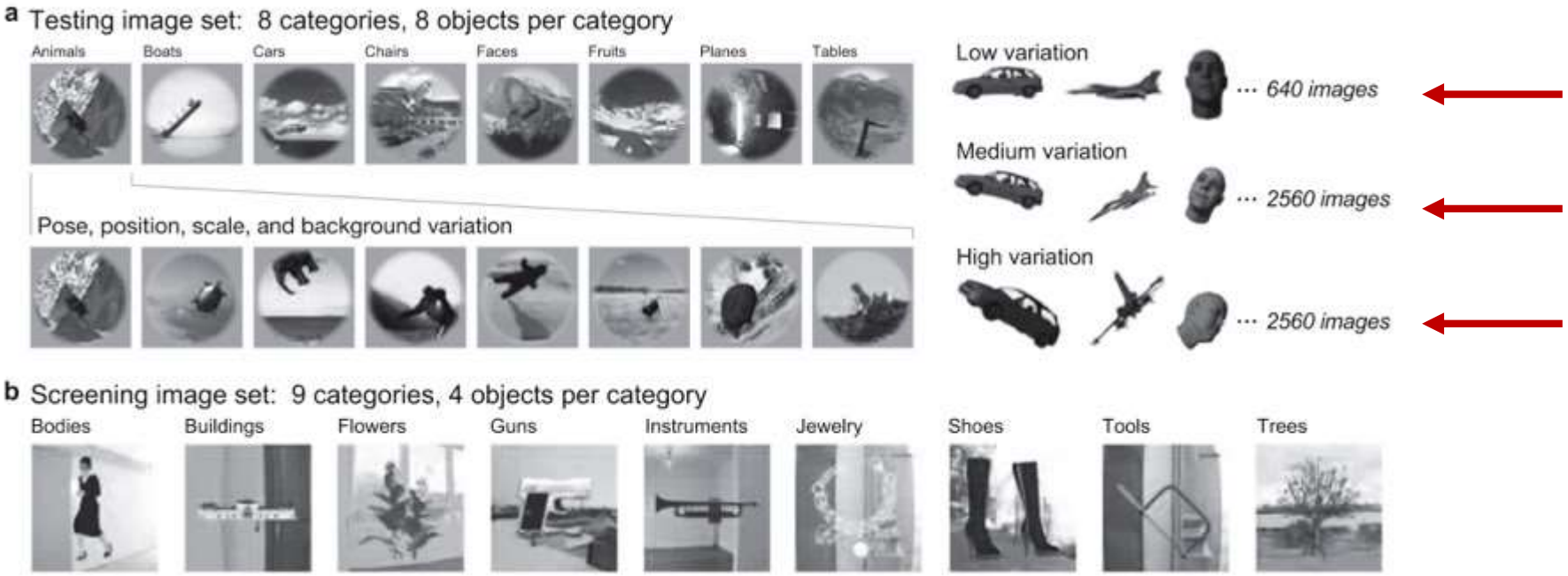
The idea of this approach is to first optimize network parameters for performance on an challenging object recognition task, once network parameters have been fixed, compare networks to neural data.

Data collection

Collected neural data from V4 and IT (128 units in V4, and 168 units in IT) in two macaque monkeys, assessed human behavior, and tested HCNN models on a common image set.

Fixating animals were presented in the center of screen (8° visual angle) with images in pseudorandom order, each for 100 ms, followed by a 100-ms gray blank period with no image.

For each image and electrode, firing rates were by averaging spike counts in the period 70–170 ms after stimulus presentation.
Final neuron output responses were obtained for each image and site by averaging
over image repetitions.



Testing images used to collect neural data and evaluate HCNN models.

Images of 64 distinct objects from 8 categories (animals, boats, cars, chairs, faces, fruits, planes, tables), with 8 specific exemplars of each category, repeated 30 times for a total of 5,760 images, presented with 3 levels of difficulty, low, medium and high variation.

Screening images used to discover the hierarchical modular optimization (HMO) model contained 4,500 images of 36 objects in 9 categories

Models were drawn from a large parameter space of HCNN
Any given HCNN is characterized by the following:

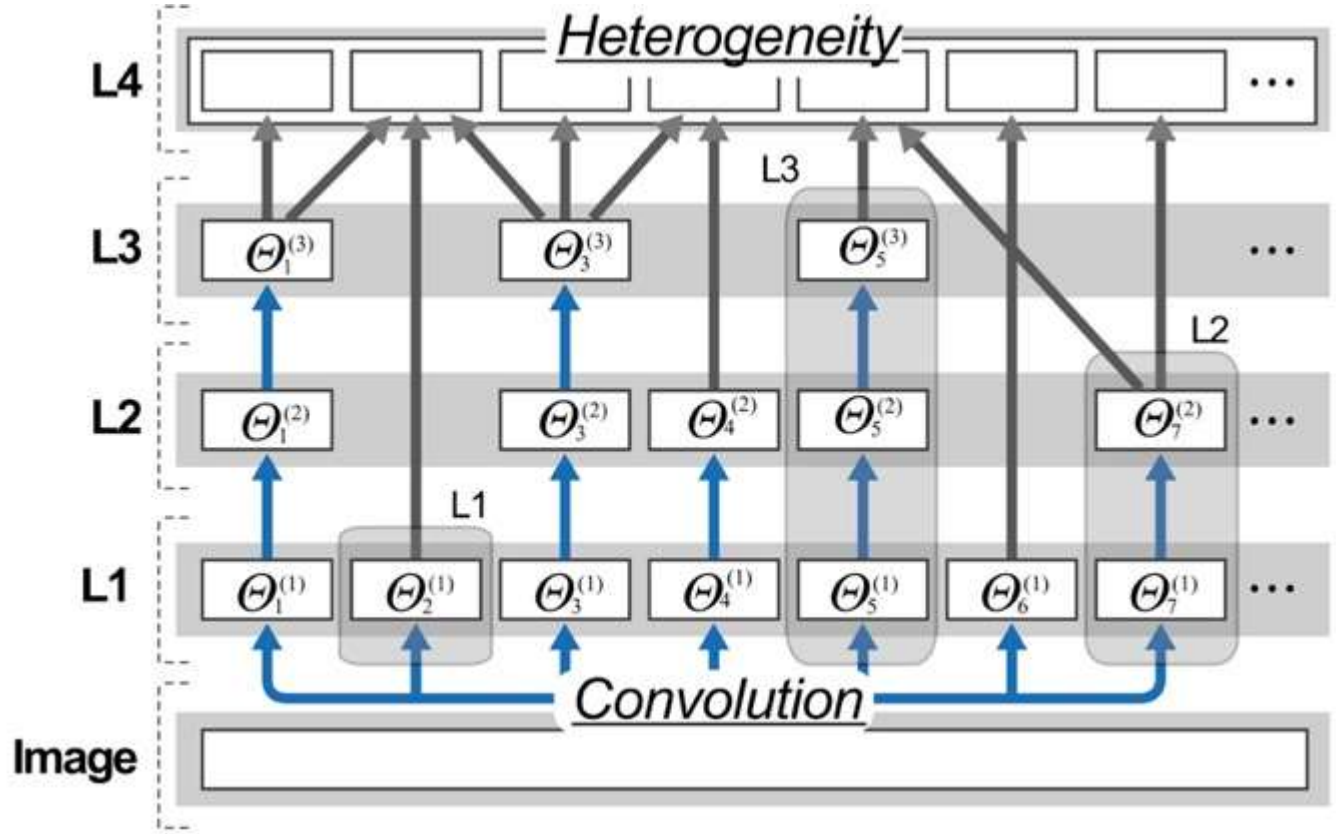
discrete architectural parameters, including the number of layers, for each layer, discrete parameters specifying the number of filters; the local radius of each filtering, pooling and normalization operation; the pooling type; threshold values, and other choices required by the specific HCNN implementation;

continuous filter parameters, specifying the filter weights of;

Let Nk denote the space of stacked networks (N) of depth k, the study used HCNN of depth 3 or less.

Mixture Models – Extending Standard CNNs

- Combine **diverse modules** at the same layer (horizontal composition $\oplus$).
- Modules differ in:
  - Filter sizes (e.g., 3×3, 5×5),
  - Pooling strategies (max, average),
  - Number of filters, depth.



Biological Inspiration & Relevance

Brain's Ventral Stream:

- Not strictly feedforward—features bypass connections (e.g., V1 → V4).
- Heterogeneous neurons in same areas (e.g., V1 processes orientation, motion, color).

Mixture Models Mirror This:

- Parallel pathways process different features at the same stage.
- Fast, shallow modules handle low-level features.
- Slow, complex modules handle detailed representations.

Key Operations:

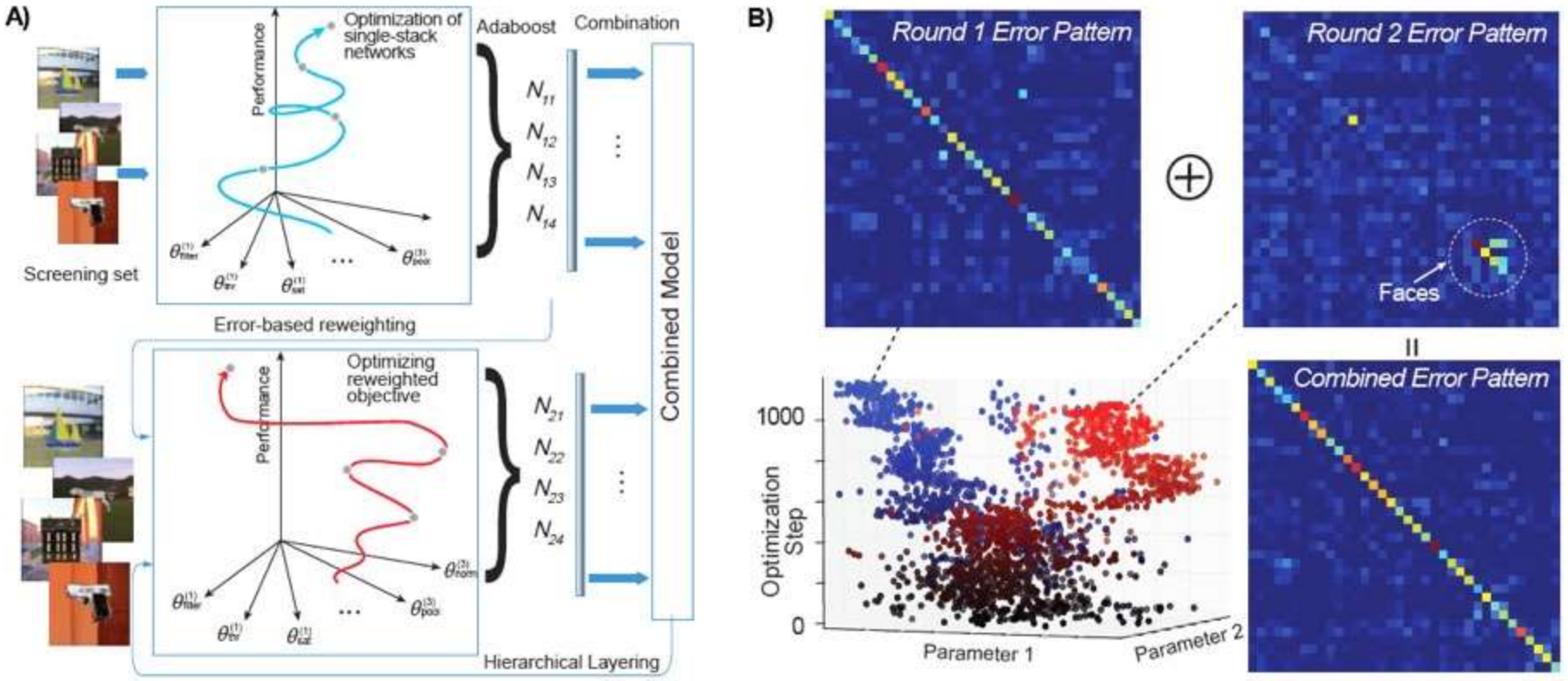
- $\otimes$ (Vertical): Depth, hierarchical complexity (V1 → V2 → V4 → IT).
- $\oplus$ (Horizontal): Breadth, diverse processing within the same layer.

Purpose:

- Enhance neural predictivity by capturing the brain's parallel, diverse processing.
- Achieve efficient, specialized computation like biological systems.

Hierarchical Modular Optimization (HMO)

An **evolutionary algorithm** that optimizes CNN architectures without using backpropagation.
Inspired by biological plausibility: avoids global gradient signals, uses local learning.



Biological Relevance of HMO

Local Learning: Modules are optimized without global error signals.
Greedy, Stepwise Learning: Layers are built progressively, reflecting cortical development.
No Weight Symmetry: Avoids the biologically implausible requirement of symmetric forward/backward weights.
Adaptive Focus: Error-based reweighting mirrors attention and plasticity.

Outcomes:

- Networks predict neural activity (e.g., IT cortex) well, even without backprop.
- Achieve competitive object recognition performance.
- Reflect parallel, heterogeneous processing in the brain's ventral stream.

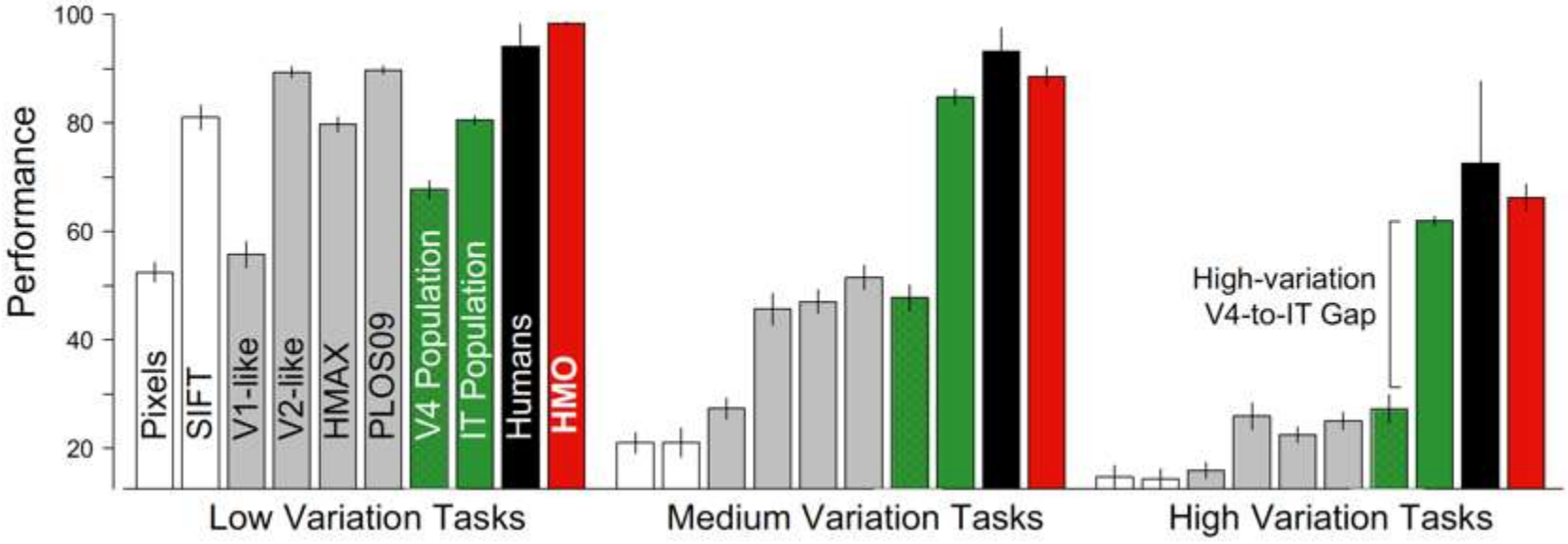
HMO combines Bayesian optimization, boosting, and modular stacking to create brain-like, adaptive networks.

Categorization performance

Object recognition performance was assessed by training linear SVM classifiers with regularization on model and neural output.

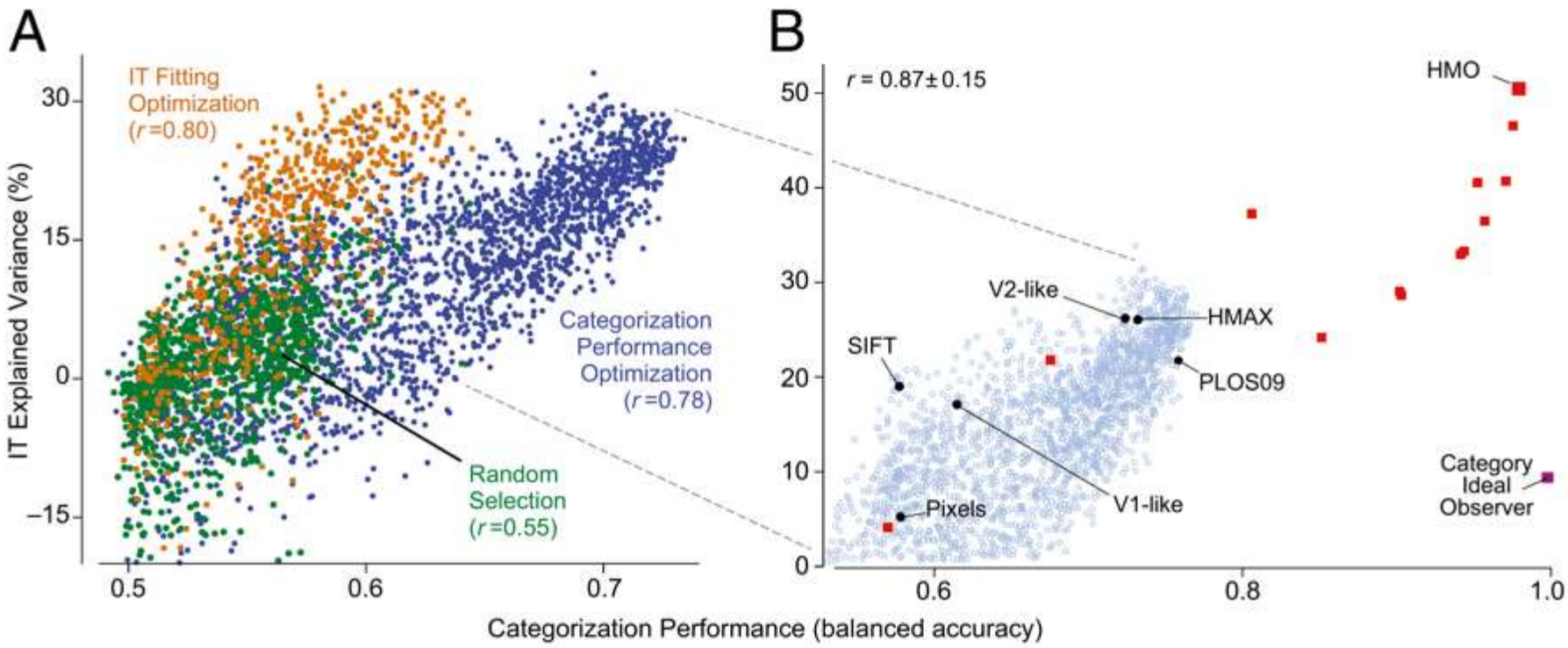
For models, the output features on each stimulus are defined as the set of scalar values for each top-level model unit when evaluated on that stimulus.

For neuronal sites, the output features are defined as the vector of scalar firing rates for each unit, as is typical in neural decoding studies.



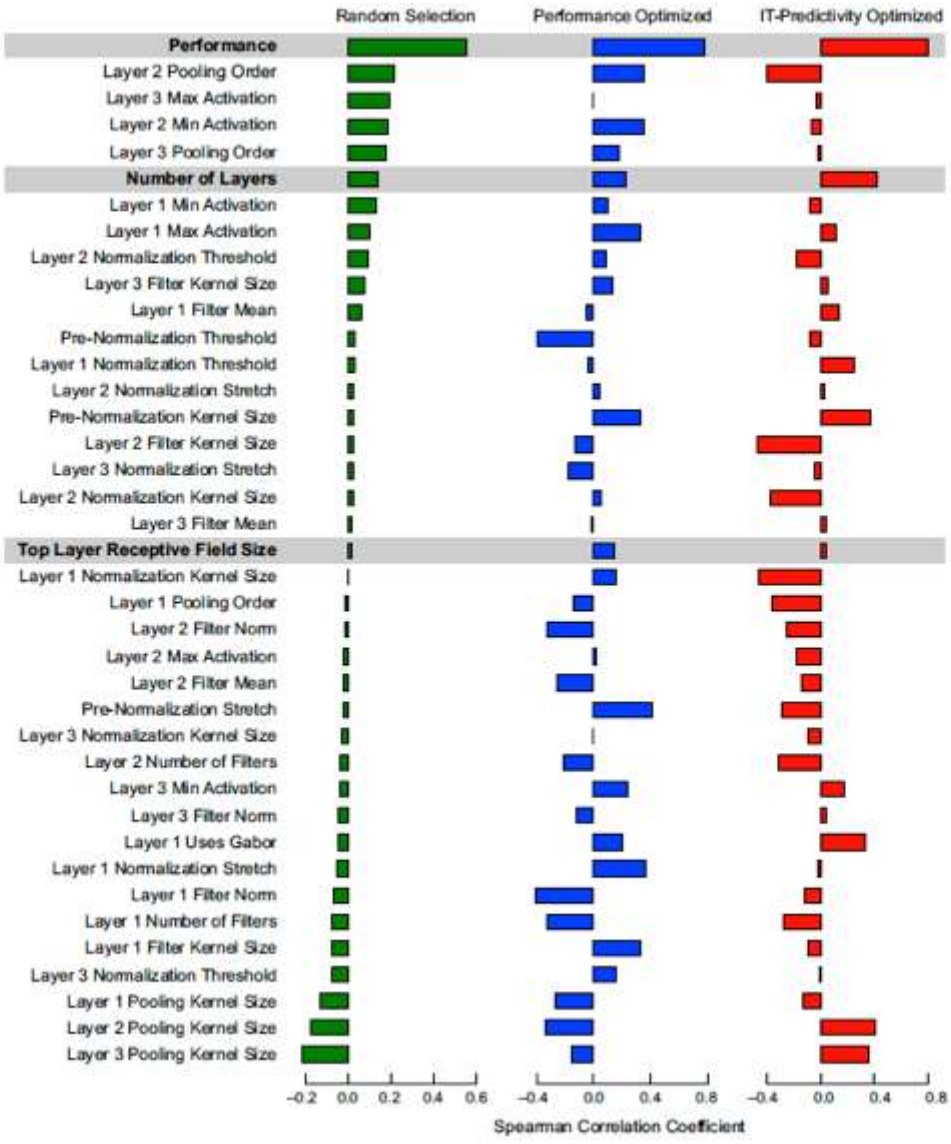
Models were selected for evaluation (measuring categorization performance and IT neural predictivity for each model) by one of three procedures:

- random sampling of the models from the parameter space N3 (n = 2.016, green dots);
- searched models that maximized performance on the high-variation categorization task (n = 2,043, blue dots);
- searched models that maximized for IT neural predictivity (n = 1.876, orange dots).



Performance/IT-predictivity correlation

The x axis shows performance (balanced accuracy, chance is 0.5) of the model output features on a high-variation categorization task; the y axis shows the median single site IT explained variance percentage (n=168 sites) of that model. Each dot corresponds to a distinct model selected from a large family of HCNN.



No individual model parameters correlated nearly as strongly with IT predictivity as performance, indicating that the performance/IT predictivity correlation cannot be explained by simpler mechanistic considerations (e.g., receptive field size of the top layer).

Correlation of model parameters with IT predictivity for the three types of model selection

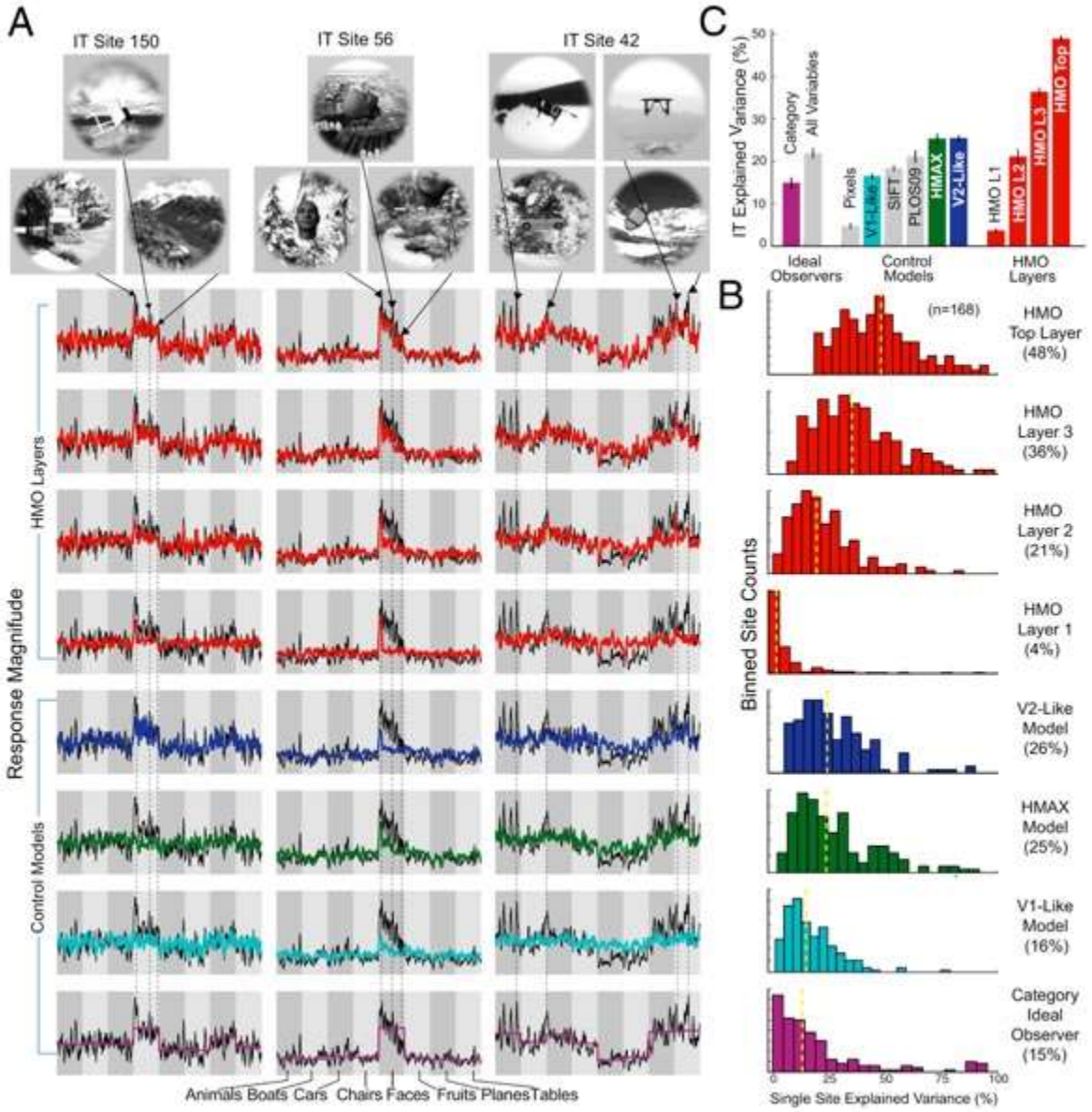
Predicting neural responses in individual IT and V4 neural sites.

To assess model's ability to predict a given neuron output, a standard linear regression methodology was used, in which each neuron site is modeled as linear combination of model outputs.

Briefly, a partial least squares (PLS) regression procedure was used to determine weightings of top-level model outputs which best fit a given neurons' output on a randomly chosen subset of the testing images.

The percentage of explained variance was then computed on a per-site basis using the R2 prediction value for that site

IT neural predictions



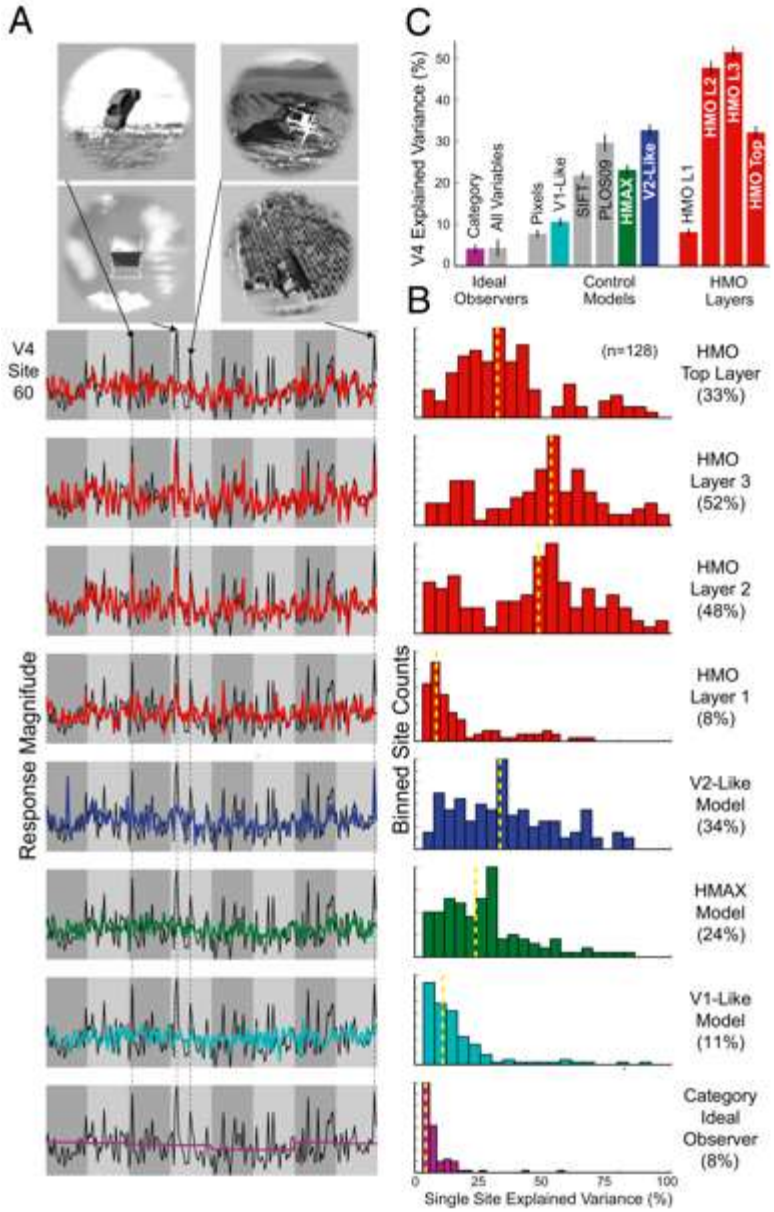
Actual neural response (black trace) vs. model predictions (red trace) for three individual IT neural sites.

Each successive layer predicted IT units increasingly well.

The HMO model layers show that category selectivity and tolerance to more drastic image transformations emerges gradually along the hierarchy.

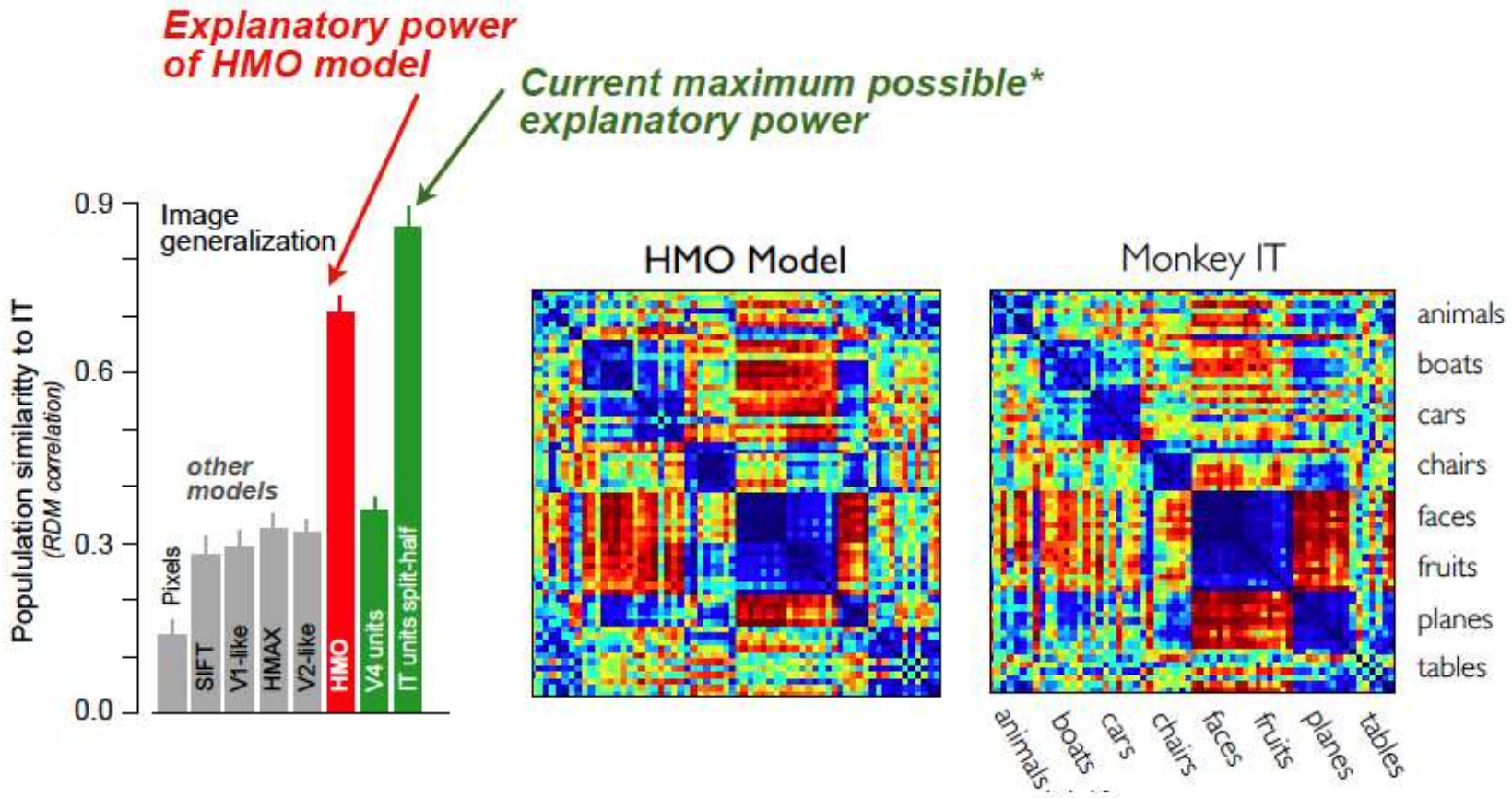
Top layer of the high-performing HMO model achieves high predictivity for individual IT neural sites, predicting 48:5 ±1:3% of the explainable IT neuronal variance.

V4 neural predictions



HMO model's penultimate layer is highly predictive of V4 neural responses (51:7± 2:3% explained V4 variance), providing a significantly better match to V4 than either the model's top or bottom layers.

Representation Dissimilarity Matrix (RDM)



RDM is a tool comparing two representations on a common stimulus set.
Each entry in the RDM corresponds to one stimulus pair, with high/low values indicating that the population as a whole treats the pair stimuli as very different/similar.
Taken over the whole stimulus set, the RDM characterizes the layout of the images in the high dimensional neural population space.

(blue = 0th distance percentile, red = 100th percentile.

Core object recognition

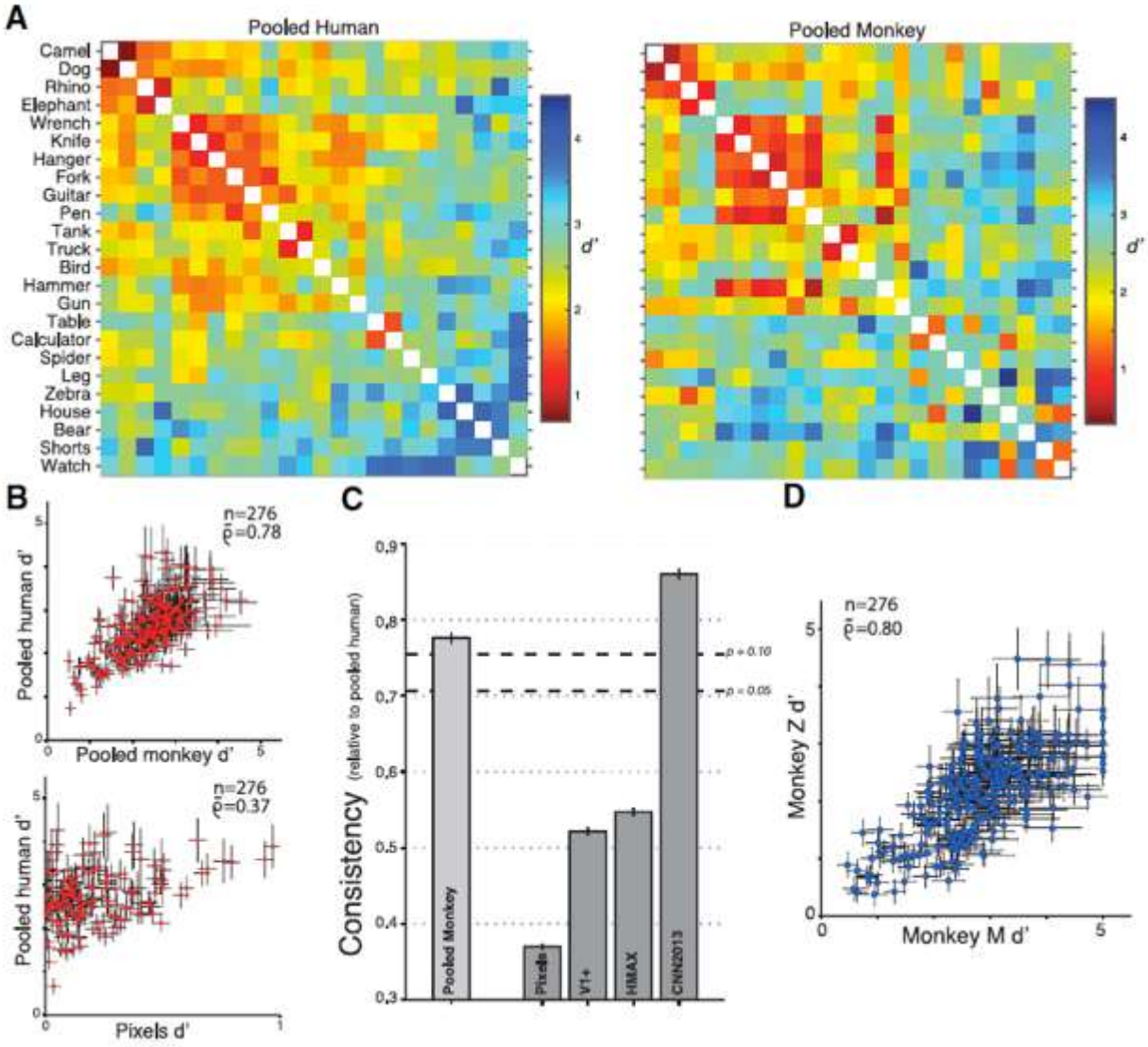
Ability to rapidly identify objects in the central visual field, in a single natural fixation (~200 ms), despite various image transformations (i.e., changes in viewpoint) and background.

Nonhuman and human primates reveal similar invariant visual object recognition when performing the same binary object recognition tasks

Monkey performance shows a pattern of object confusion that is highly correlated (consistency) with human performance confusion pattern (0.78).

Importantly, low-level visual representations (pixels) do not share these confusion patterns (pixels, 0.37).

These results are in line with with the hypothesis that rhesus monkeys and humans share a common neural shape representation that directly supports object recognition.



Rajalingham et al., J. Neurosci, 2015

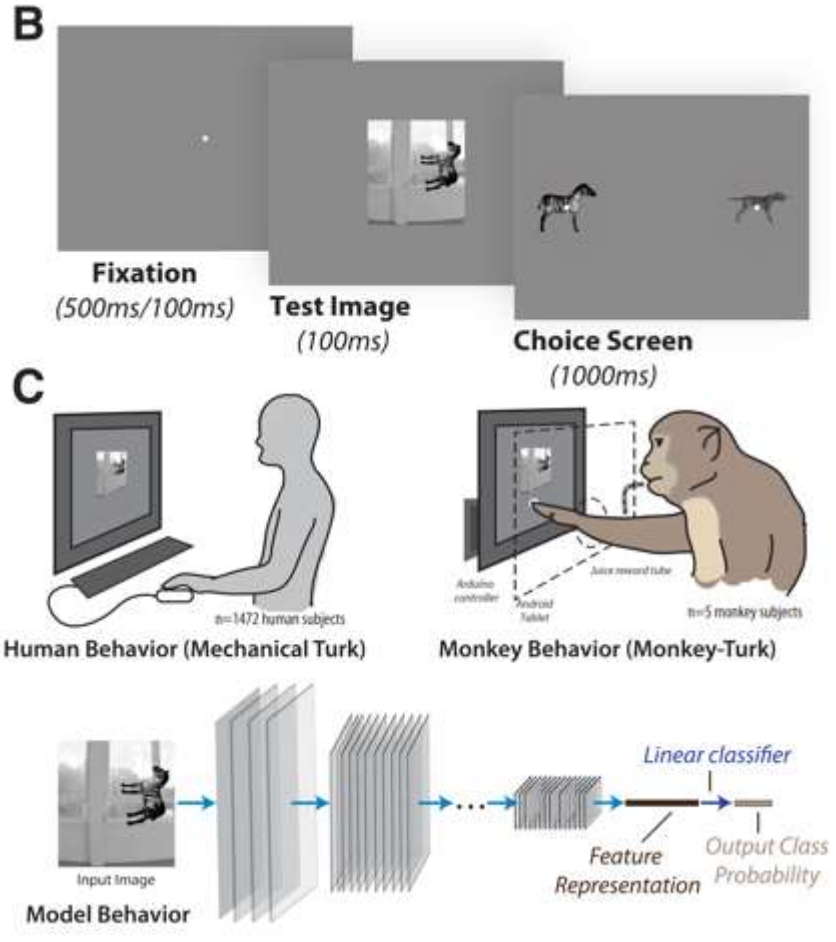
Deep convolutional neural networks (DCNNs), optimized by supervised training on large scale category-labeled image sets (for instance, ImageNet) display internal feature representations similar to neuronal representations along the primate ventral visual stream and they exhibit behavioral patterns similar to the behavioral patterns of pairwise object confusions of primates.

However….

several studies have shown that DCNN models can diverge drastically from humans in object recognition behavior.

Such failures of the current DCNN models would likely not be captured using low-resolution behavioral measures (i.e., object-level) but could be revealed at higher resolution (image level).

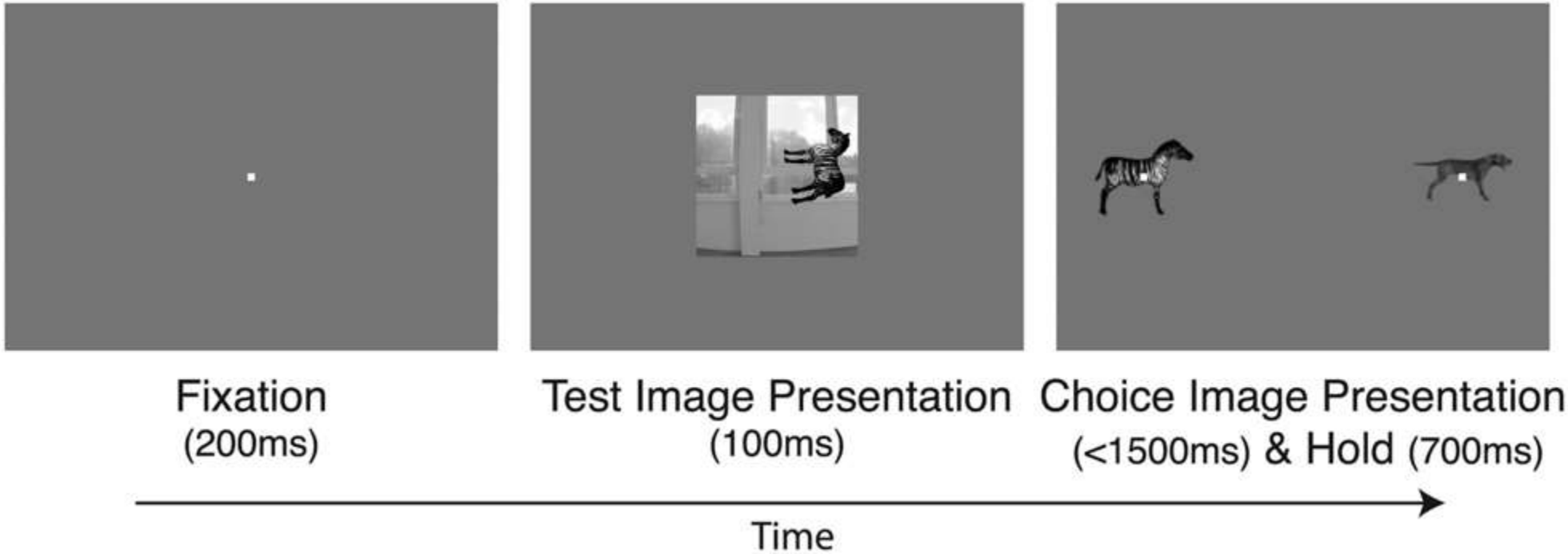
A recent study employed both low- and high-resolution measurements of behavior (over a million behavioral trials) from 1472 anonymous humans and five male macaque monkeys with 2400 images for over 276 binary object discrimination tasks.



Two example images for each of the 24 objects.

To enforce invariant object recognition behavior, each image included one object, with randomly chosen viewing parameters (e.g., position, rotation and size) placed onto a randomly chosen, natural background.

Rajalingham et al., J. Neurosci, 2018



For monkeys, each trial was initiated when they held fixation on a central point for 200 ms, after which a test image (6° of visual angle) appeared at the center for 100 ms. Immediately after extinction of the test image, two choice images, each displaying the canonical view of a single object with no background, were shown to the left and right. The monkey was allowed to freely view the response images for up to 1500ms and respond by holding fixation over the selected image for 700 ms.

Rajalingham et al., J. Neurosci, 2018

Table 1. Definition of behavioral performance metrics

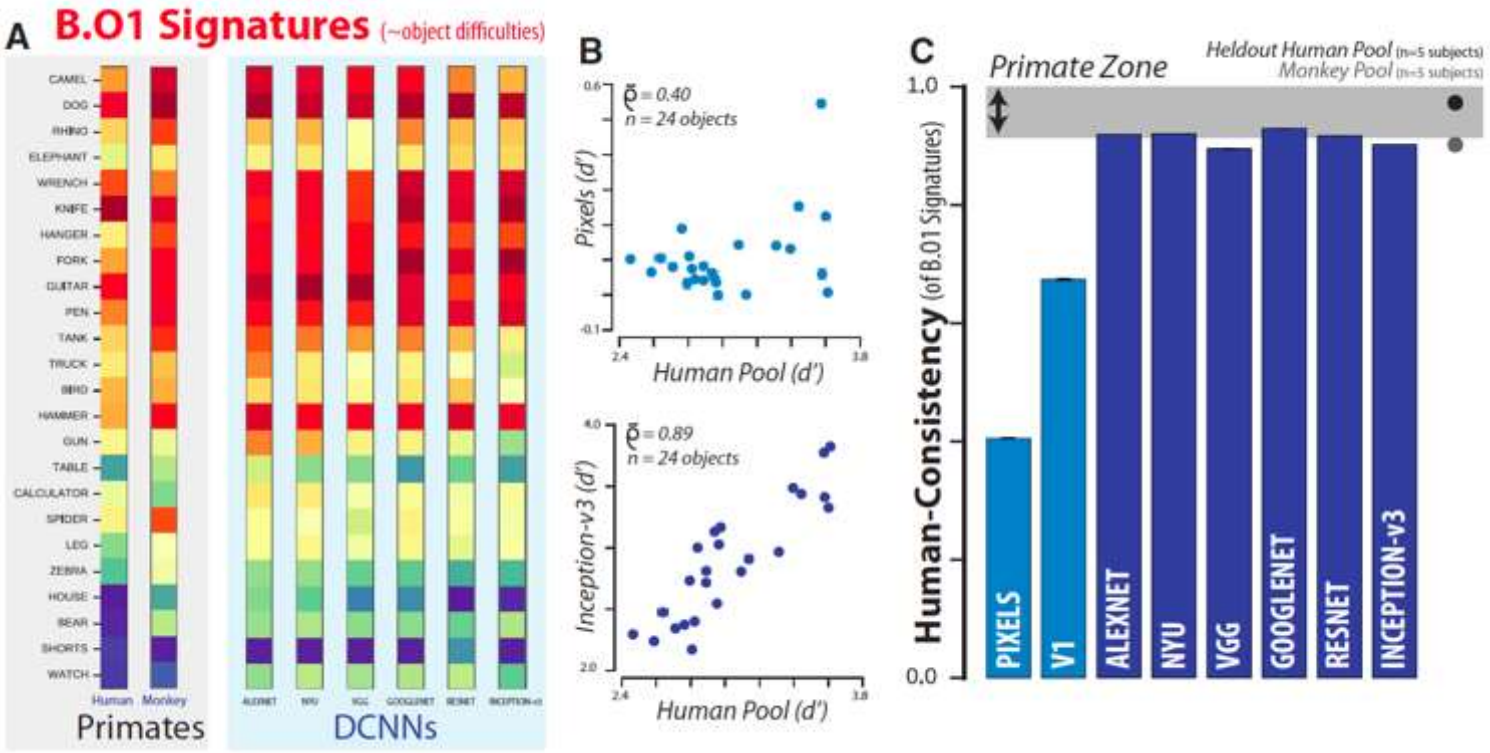
Behavioral metric	HR	FAR
One-versus-all object-level performance (B.01) ($N_{objects} \times 1$) $O_1(i) = Z(HR(i)) - Z(FAR(i))$, $i = 1, 2, \ldots, N_{objects}$	Proportion of trials when images of object *i* were correctly labeled as object *i*	Proportion of trials when any image was incorrectly labeled as object *i*
One-versus-other object-level performance (B.02) ($N_{objects} \times N_{objects}$) $O_2(i, j) = Z(HR(i, j)) - Z(FAR(i, j))$, $i = 1, 2, \ldots, N_{objects}$ $j = 1, 2, \ldots, N_{objects}$	Proportion of trials when images of object *i* were correctly labeled as *i* when presented against distractor object *j*	Proportion of trials when images of object *j* were incorrectly labeled as object *i*
One-versus-all image-level performance B.I1 ($N_{images} \times 1$) $I_1(ii) = Z(HR(ii)) - Z(FAR(ii))$, $ii = 1, 2 \ldots, N_{images}$	Proportion of trials when image *ii* was correctly classified as object *i*	Proportion of trials when any image was incorrectly labeled as object *i*
One-versus-other image-level performance B.I2 ($N_{images} \times N_{objects}$) $I_2(ii, j) = Z(HR(ii, j)) - Z(FAR(ii, j))$, $ii = 1, 2 \ldots, N_{images}$ $j = 1, 2 \ldots, N_{objects}$	Proportion of trials when image *ii* was correctly classified as object *i* when presented against distractor object *j*	Proportion of trials when images of object *j* were incorrectly labeled as object *i*

The first column provides the name, abbreviation, dimensions, and equations for each of the raw performance metrics. The next two columns provide the definitions for computing the hit rate (HR) and false alarm rate (FAR), respectively.

Each behavioral metric computes a sensitivity (discriminability) index: d' = Z(HitRate) - Z(FalseAlarm-Rate), where Z is the standard z score

Rajalingham et al., J. Neurosci, 2018

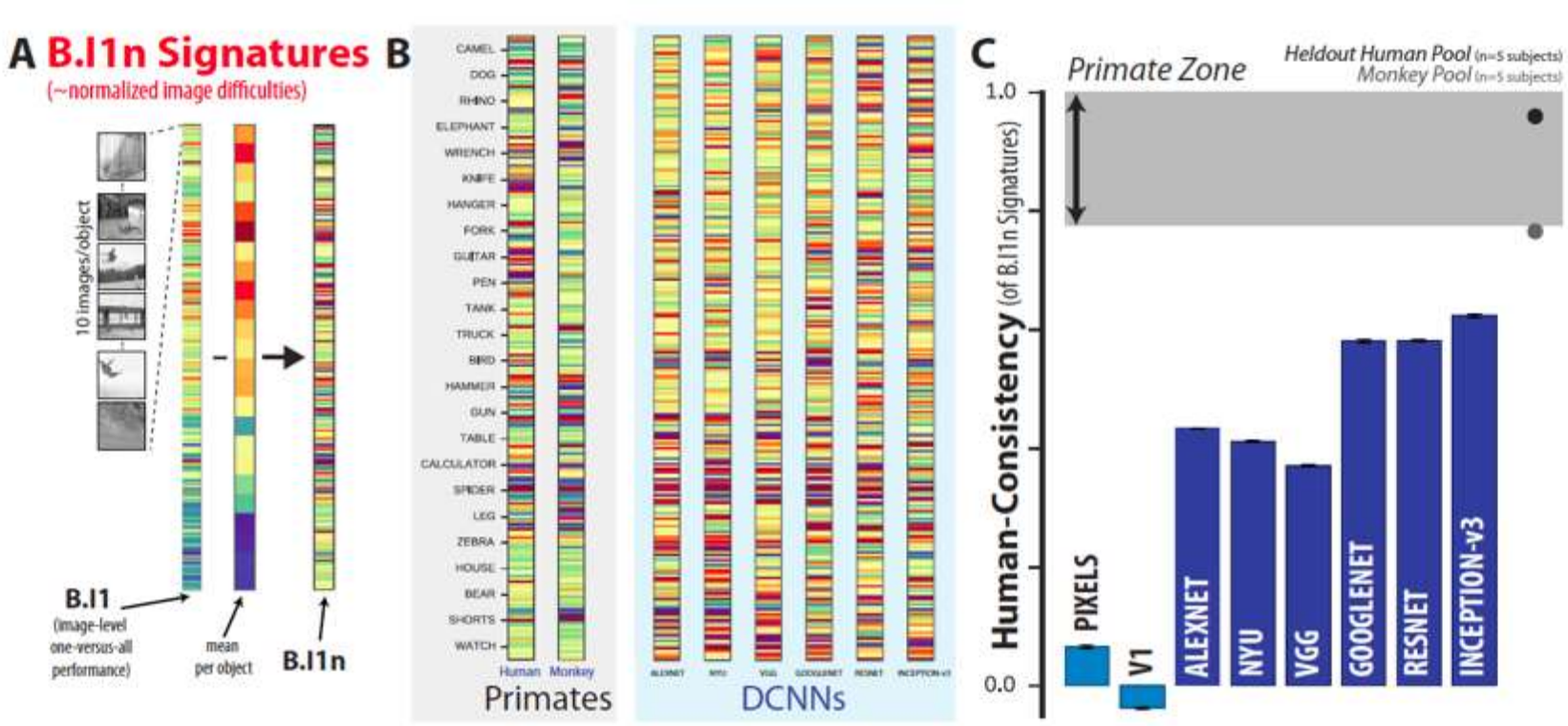
Object-level (across all images and distractors) behavioral comparison



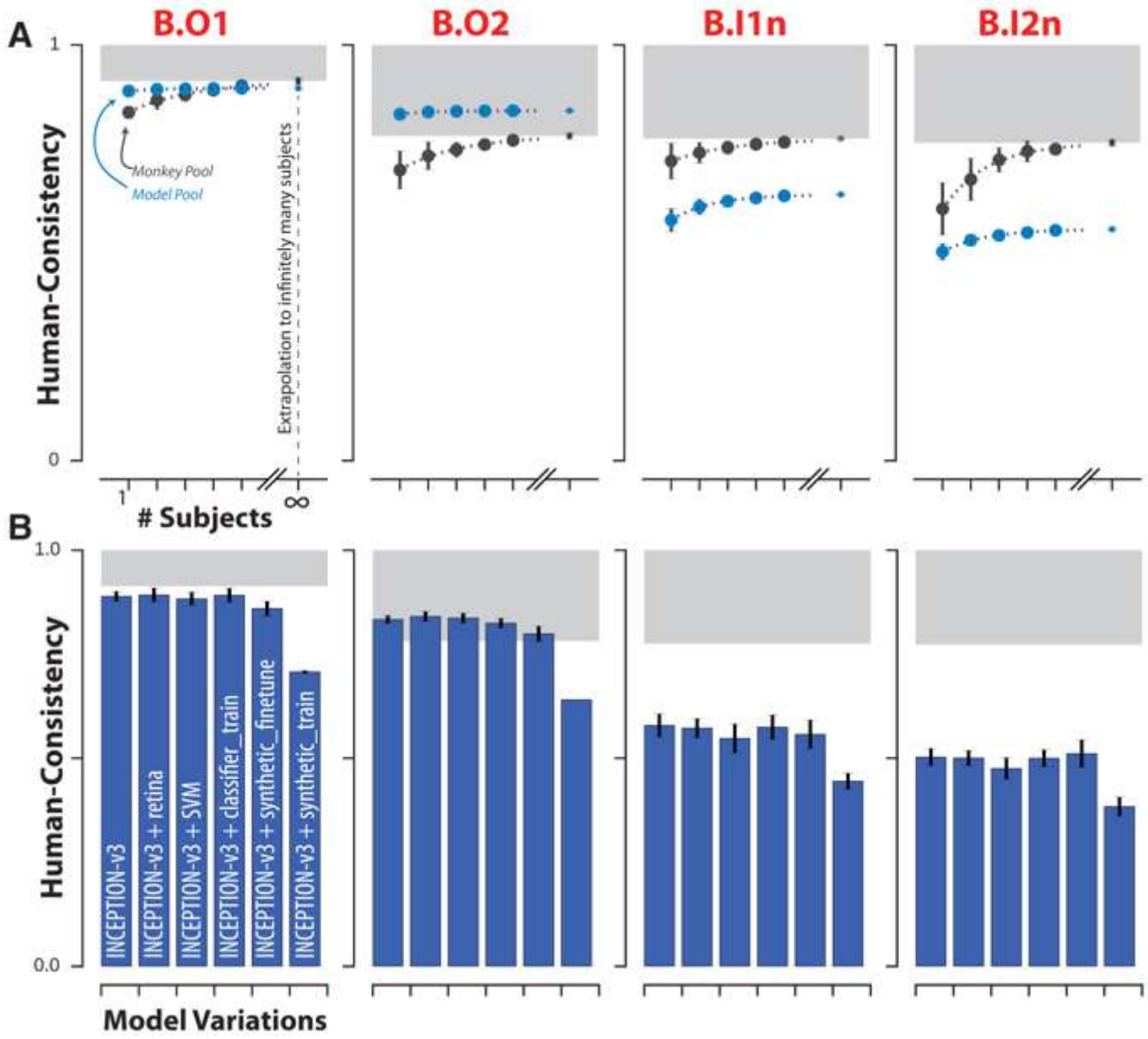
B.O1 signatures (discriminability measures) for the human (n=1472), monkey (n=5), and several DCNN models as 24-dimensional vectors using a color scale (warm colors=lower discriminability). Each element of the vector corresponds to the system's discriminability of one object against all others that were tested (i.e., all other 23 objects).

Human consistency was a used to quantify the similarity between a model visual system and the human visual system with respect to a given behavioral metric (signatures).

Image-level behavioral comparison



The one-versus-all image-level signature (B.I1) is shown as a 240-dimensional vector (a subset of 240 images, 10 images/object) using a color scale, where each colored bin corresponds to the system's discriminability of one image against all distractor objects.


A
B.O1
B.O2
B.I1n
B.I2n
1
0
Human-Consistency
Monkey Pool
Model Pool
Extrapolation to infinitely many subjects
1
∞
Subjects
B
1.0
0.0
Human-Consistency
INCEPTION-v3
INCEPTION-v3 + retina
INCEPTION-v3 + SVM
INCEPTION-v3 + classifier_train
INCEPTION-v3 + synthetic_finetune
INCEPTION-v3 + synthetic_train
Model Variations

Examining behavior at the higher resolution of individual images, all leading DCNN models failed to replicate the image-level behavioral signatures of primates.

Rhesus monkeys are more consistent with the archetypal human than any of the tested DCNN models (at the image level).

Synthetic image-optimized models were no more similar to primates than ANN models optimized only on ImageNet, suggesting that the tested ANN architectures have one or more fundamental flaws that cannot be readily overcome by manipulating the training environment.

DCNN models diverge from primates in their core object recognition behavior.

This suggests that either the model architectural (e.g., convolutional, feedforward) and/or the optimization procedure (including the diet of visual images) that define this model subfamily are fundamentally limiting.

Rajalingham et al., J. Neurosci, 2018

”Fundamental flaws" in ANN models:

1.Feedforward-only architectures:

1. Most DCNNs at the time (like AlexNet, VGG, etc.) are purely feedforward.
2. But the primate visual system uses **recurrent processing**, feedback loops, and dynamic activity over time.
3. Maybe models need recurrence to better match primate vision, especially for hard-to-recognize images.

2.Lack of attention mechanisms:

1. Primates use **selective attention** to focus on important parts of an image.
2. Most early models don’t dynamically shift attention or prioritize regions.

3.Learning mechanisms:

1. Biological systems might not learn purely through supervised classification like ImageNet.
2. **Unsupervised**, **self-supervised**, or **reinforcement learning** might better reflect how primates learn about the world.

4.Energy constraints / sparsity / noise robustness:

1. The brain operates under physical constraints and uses **sparse, noisy signals** efficiently.
2. Current models might be too "brute force" compared to the brain's elegance.

5.Hierarchical representation differences:

1. Even though DCNNs have layers like the visual hierarchy (V1, V2, IT), the **features** they learn at each level might not align with what the brain actually represents.

Known Misalignments Between Brains and HCNNs

Robustness to Noise & Texture Bias (Geirhos et al., 2018):

- HCNNs like VGG-19 and ResNet-152 are less robust than humans to Gaussian noise in object recognition tasks.
- HCNNs tend to rely more on **texture cues**.
- Humans rely more on **shape** for object recognition

Adversarial Vulnerability (Goodfellow et al., 2014):

- Small, imperceptible pixel changes can drastically alter HCNN predictions (e.g., "dog" → "church").
- Humans are largely insensitive to these perturbations.
- Newer models show improved alignment but gaps remain (Gaziv et al., 2023).

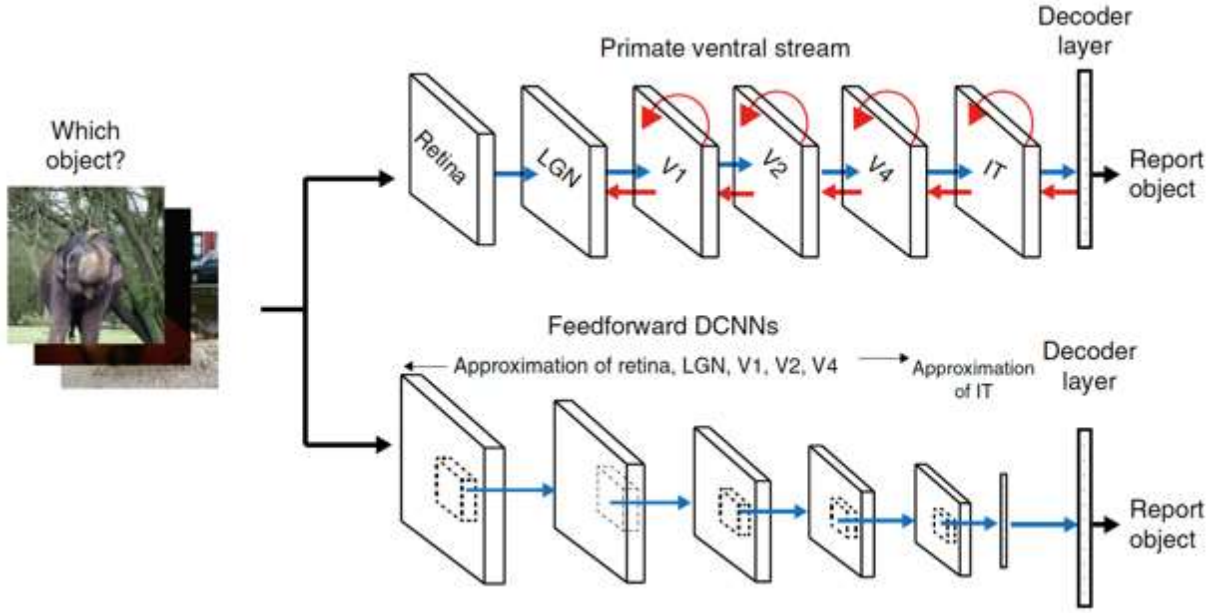
Visual Perception Phenomena

HCNNs struggle with:

- Global shape processing.
- Object part relationships.
- Illusory and uncrowding effects.

Recurrent neural networks

"*Recurrent circuits are the brain's way of keeping the past alive – both to interpret the present, and to shape future learning".*



Deep CNNs trained on object categorization are the best predictors of primate behavioral patterns across multiple core object recognition tasks;

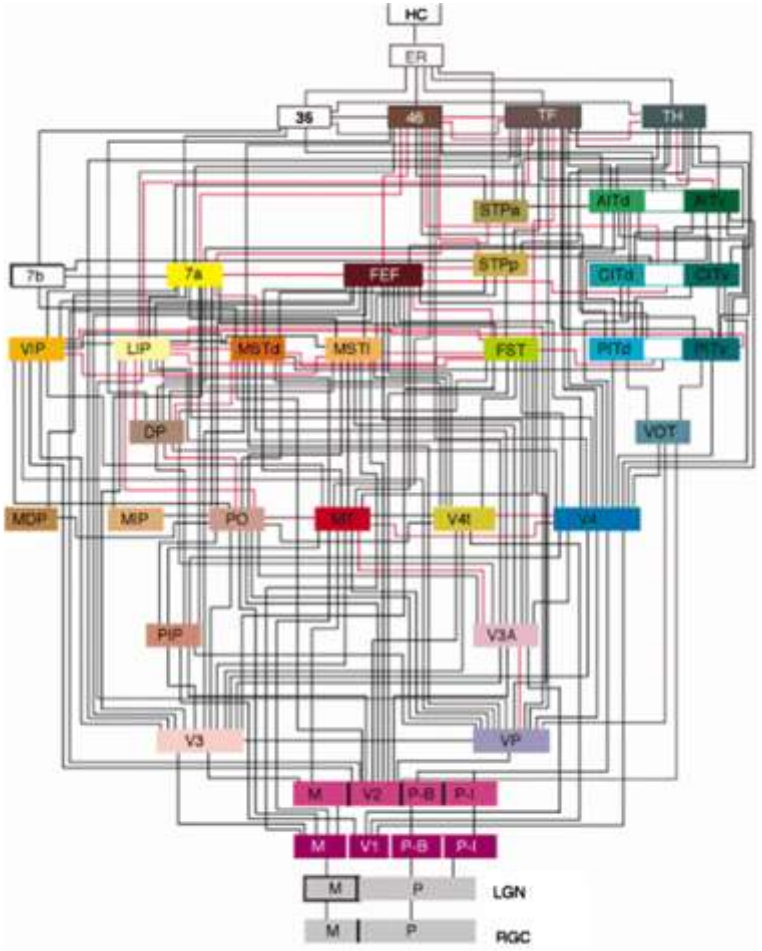
These networks are also the best predictors of individual responses of macaque IT neurons

Unlike the primate ventral stream, these neural networks in this family are almost entirely feedforward and lack cortico-cortical, subcortical, and intra-areal recurrent circuits.

Kar et al., Nat. Neurosci,. 2019

Felleman & Van Essen, (1991)

Distributed Hierarchical Processing in the Primate Cerebral Cortex



Recurrent Circuits for Processing

- **Sustained activity** (e.g., in working memory tasks).
- **Contextual integration** (e.g., figure-ground segregation in vision).
- **Dynamic updating** of sensory inputs based on internal states or attention. Example: **Attractor networks** for stabilizing perceptual states.

Recurrence for Learning

- Recurrent loops help with **error-driven learning**, **credit assignment**, and **synaptic modification**.
- They may support **Hebbian learning** mechanisms over time through **reverberating activity**.
- **Re-entry** and **recurrent activation** can reinforce or refine synaptic weights during **offline replay** or during tasks involving **delayed feedback;** Example: **Sleep-related memory consolidation** involves reactivation of recurrent networks.

It has been argued that recurrent circuits might operate at much slower time scales (Hinton et al., Science, 1995)

The short duration (~200 ms) needed to accomplish accurate object identity inferences in the ventral stream suggests the possibility that recurrent circuit-driven computations are not critical for these inferences.

In addition, it has been argued that recurrent circuits might operate at much slower time scales, being more relevant for processes such as regulating synaptic plasticity (learning vs. processing).

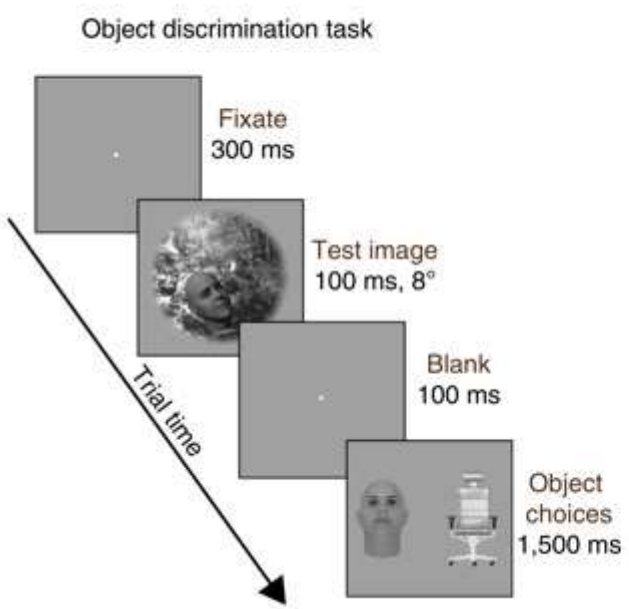
One hypothesis is that core object recognition behavior does not require recurrent processing.
However….

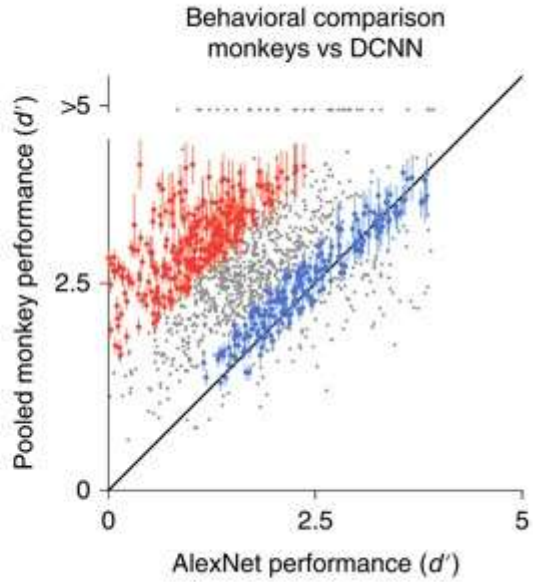
- Feedforward DCNNs fail to accurately predict primate behaviour in many situations.

- Specific images (i.e., blurred, cluttered, occluded) for which the object identities are difficult for DCNNs, but are nevertheless easily solved by primates, might involve recurrent computations.

- The impact of recurrent computations on the ventral stream might be most relevant at later time in the object recognition process.

Kar et al., Nat. Neurosci,. 2019





To compare the behavioral performance of primates (humans and macaques) and current DCNNs image-by-image, a binary object discrimination task was used, with 1,320 images (132 images per object) in which the object belonged to 1 of 10 different categories.

Macaques and humans outperform AlexNet (2012).

There were 266 challenge images (red dots) and 149 control images (blue dots)

Reaction times (RTs) for both humans and macaques for challenge images were significantly higher than for the control images (monkeys: ΔRT = 11.9 ms, humans: ΔRT = 25 ms=, suggesting that additional processing time is required for the challenge images.

Kar et al., Nat. Neurosci,. 2019

Image examples

Challenge images

Dog

Apple

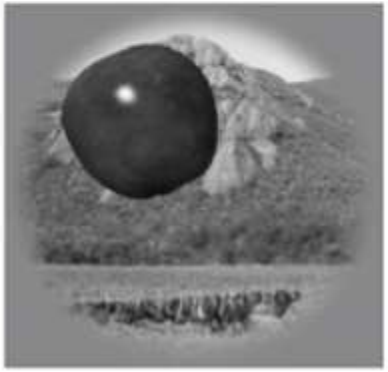
Elephant

Chair

Control images

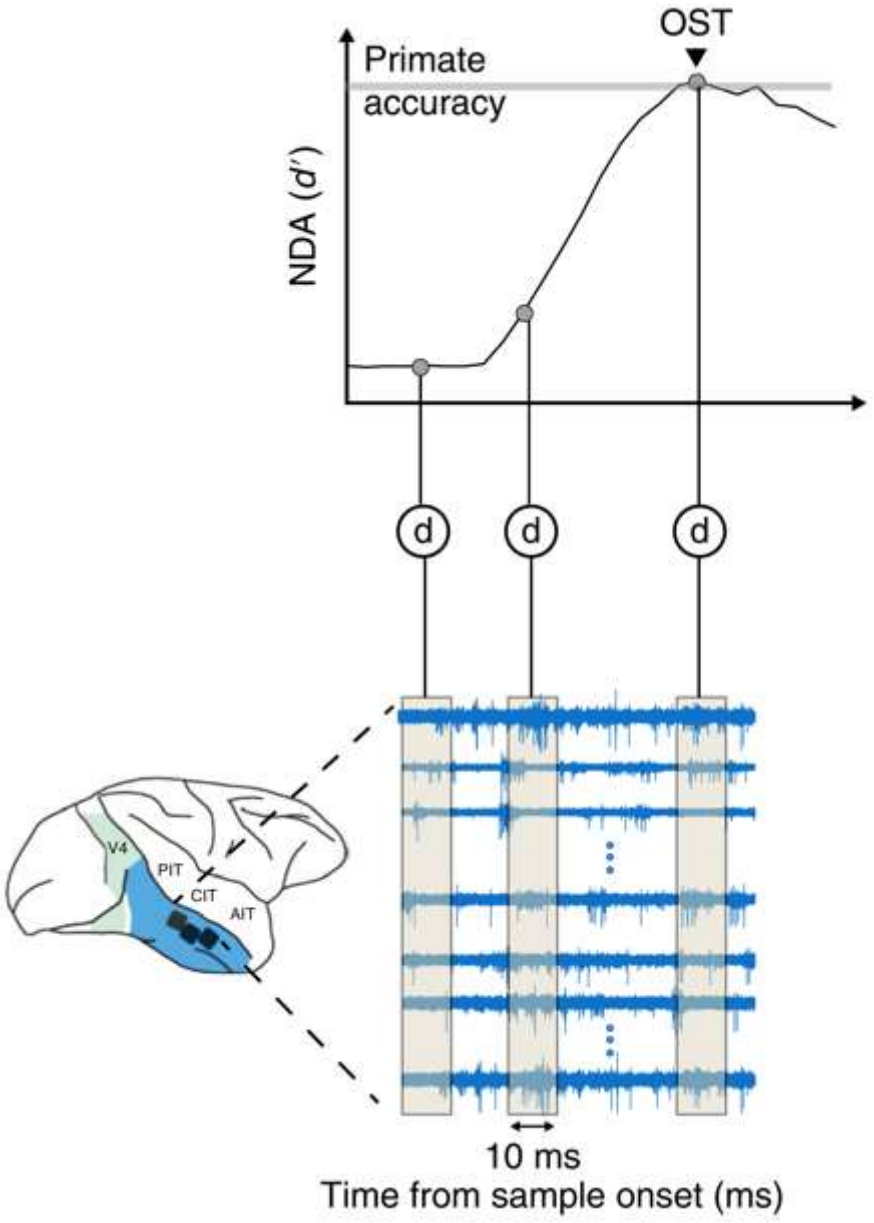
Dog

Apple



Elephant

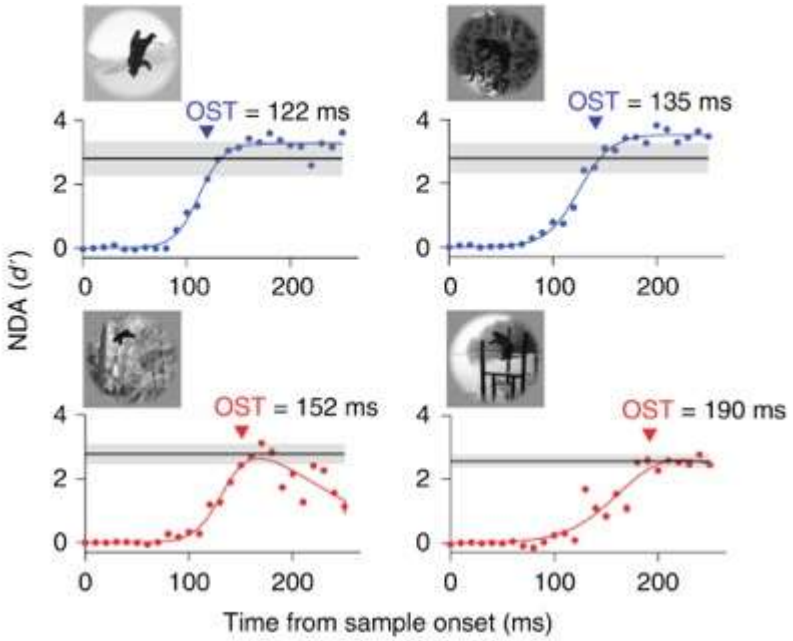
Chair

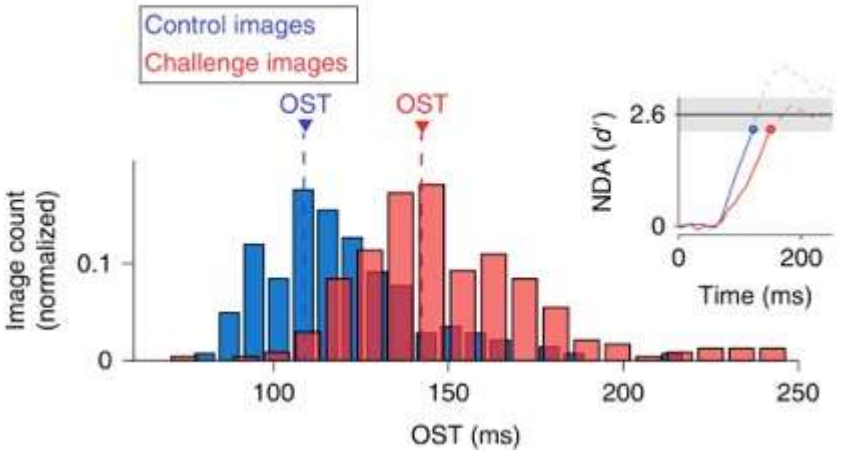


To determine the time at which object identities are formed in the IT cortex, neural decode accuracy (NDAs) was estimated for each image, every 10 ms (from stimulus onset), by training and testing linear classifiers per object independently at each time bin.

The term object solution time (or OST) refers to the time at which the NDA measured for each image reached the level of the behavioral accuracy of each subject (pooled monkey).

Kar et al., Nat. Neurosci,. 2019





First, for both the control and the challenge images, the accuracy of the IT decodes become equal to the behavioral accuracy of the monkeys at some time point after the image onset.

Second, the IT decode solutions for challenge images emerge slightly later than the solutions for the control images (average difference ~30 ms).

The challenge image required an additional time of ~30 ms to achieve full solution compared with the control images regardless of whether the animal was actively performing the task or passively viewing the images.

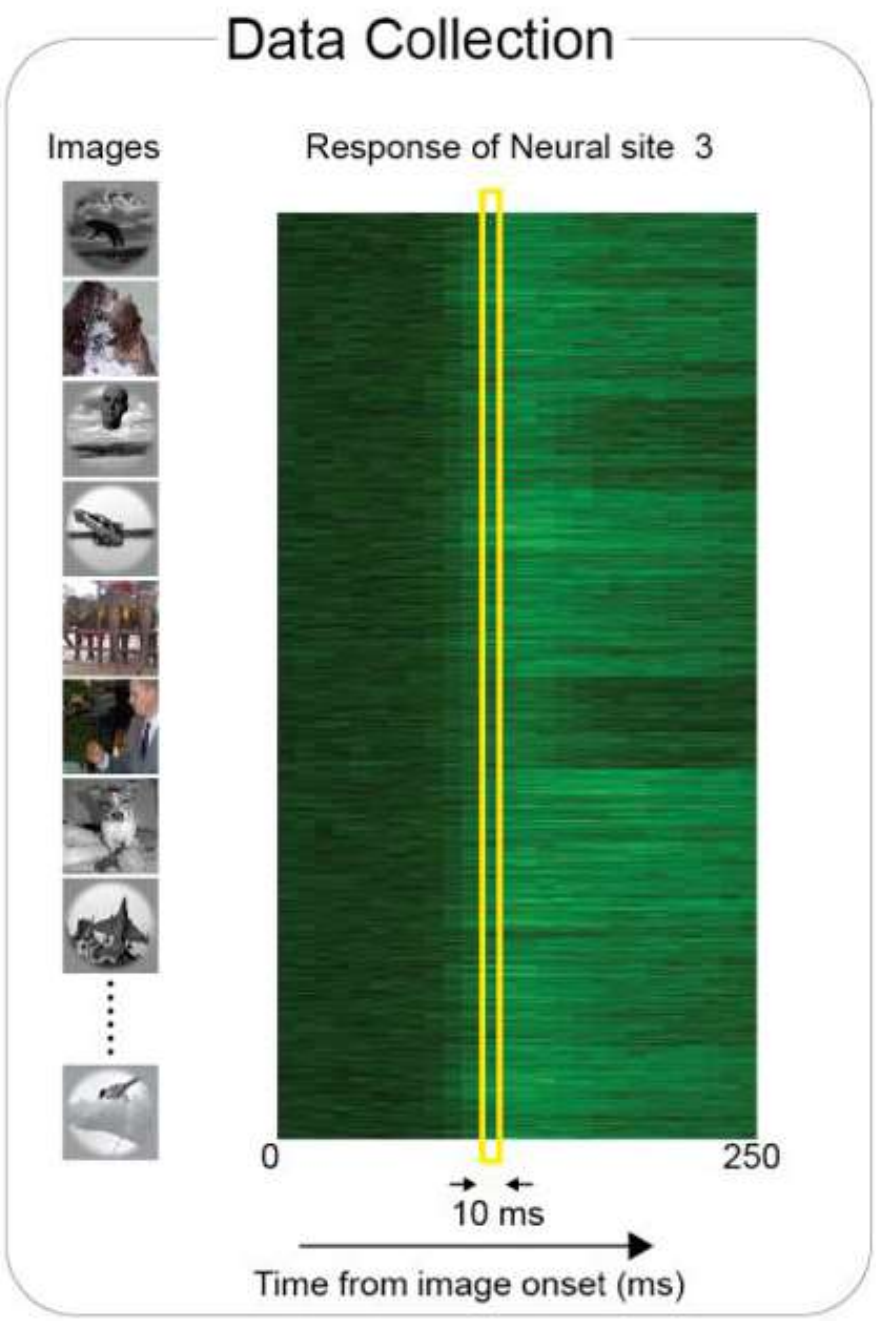
Kar et al., Nat. Neurosci,. 2019

IT predictivity across time from feedforward DCNNs

If the late-emerging IT solutions for challenges images are dependent on recurrent computations, then purely feedforward DCNNs:

- should accurately predict IT neural responses for control images,
- should fail to predict IT neural responses for challenges images

To test this idea, it was investigated how well the DCNN features could predict the time-evolving IT population response using a partial least square analysis.



Predicting IT neural responses with DCNN features

Data collection:

Neural responses are collected for each of the 1320 images (50 repetitions);

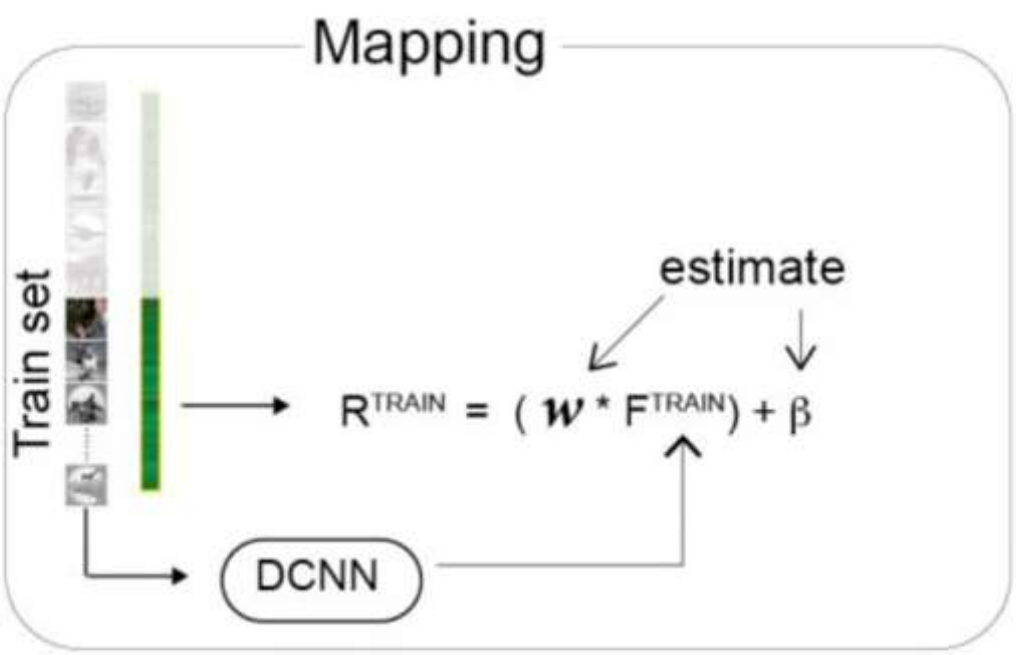
e.g. shown is that of example neural site #3, across 10 ms time bins.

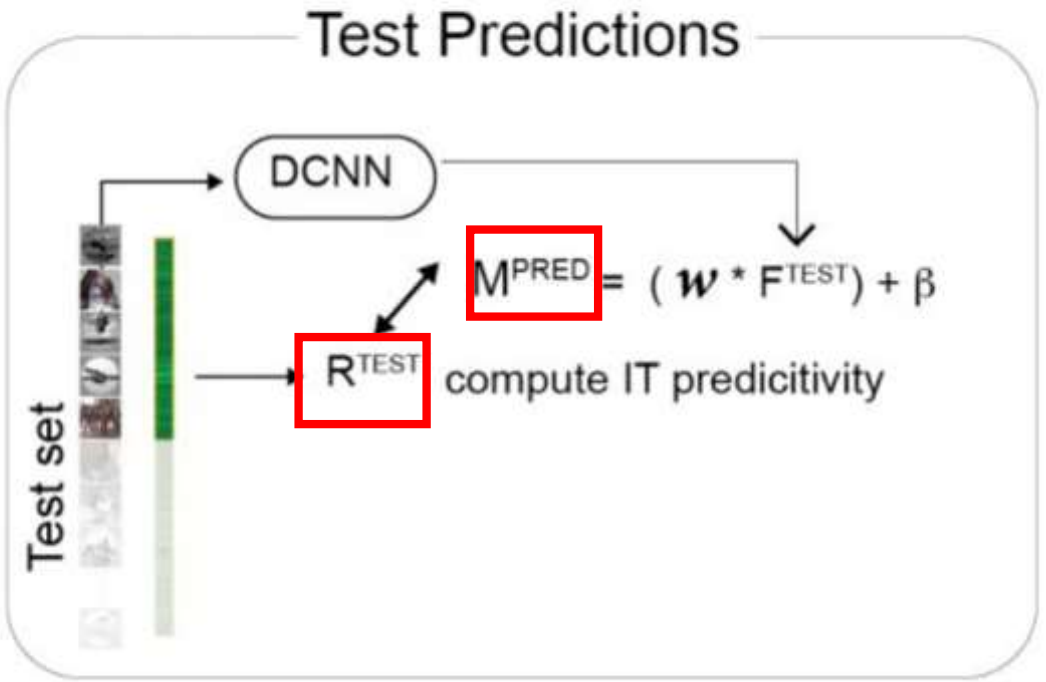
IT predictivity

Data collection: Neural responses are collected for each of the 1320 images (R^{TRAIN}) across 10 ms time-bins for each recorded IT neuron.

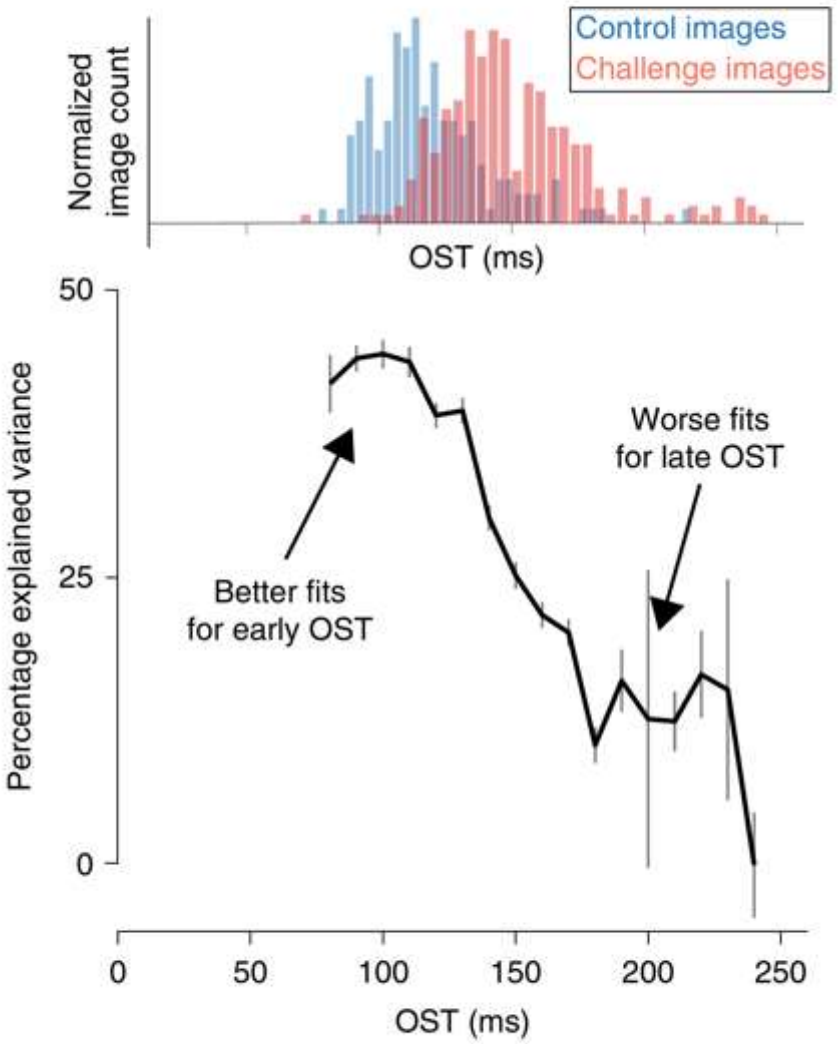
Mapping: For the train images, the image evoked activations (F^{TRAIN}) of the DCNN model from a specific layer was computed. Partial least square regression was used to estimate the set of weights (w) and biases (β) that allows to best predict R^{TRAIN} from F^{TRAIN}.

Test Predictions: Given the best set of weights (w) and biases (β) that linearly map the model features onto the neural responses, the predictions (M^{PRED}) from this synthetic neuron were generated for the test image evoked activations of the model F^{TEST}. These predictions were then compared with the test image evoked neural features (R^{TEST}) to compute the IT predictivity of the model.





IT predictivity across time from feedforward DCNNs



The fc7 layer of AlexNet predicted 44.3 ± 0.7% of the explainable IT neural response variance during the early (putative largely feedforward) response phase (90–110 ms).

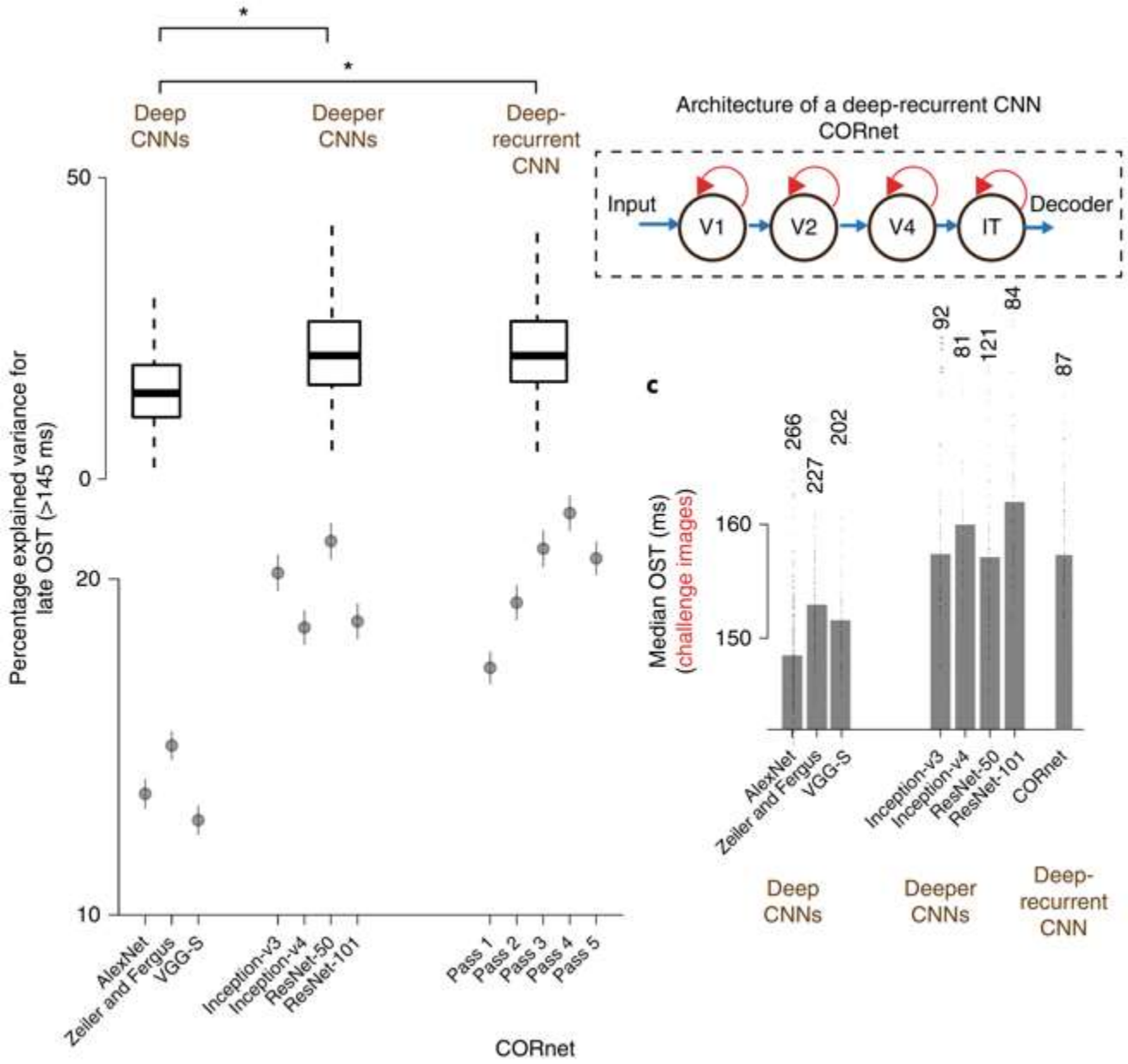
However, the ability of DCNN to predict the IT population pattern significantly worsened (<20% EV) as that response pattern evolved over time.

IT predictivity across time from feedforward DCNNs

This gradual drop in IT predictivity of feedforward DCNNs is consistent with the hypothesis that late-phase IT population responses are modified by the action of recurrent circuits.

Consistent with the hypothesis that challenge images rely more strongly on those recurrent circuits than control images, it was observed that the drop in IT predictivity coincided with the object solution times of the challenge images.

Evaluation of deeper CNNs and recurrent models of ventral visual stream processing



First, top layers of deeper CNNs predicted IT neural responses at the late phases (150–250 ms) significantly more than ‘regular-deep’ models such as AlexNet

This observation suggests that deeper CNNs might indeed be approximating ‘unrolled’ versions of the recurrent circuits of the ventral stream.

Second, it was observed a reduced number of challenge images for deeper CNNs.

Third, it was found that the images that remain unsolved by these deeper CNNs (that is, challenge images for these models) showed even longer OSTs in the IT cortex than the original full set of challenge images.

This suggests that the newer, deeper CNNs have implicitly, but only partially, approximated—in a feedforward network—some of the computations that the ventral stream implements recurrently to solve some of the challenge images.

CORnet (2018), is a four-layered recurrent neural network model

The top layer of CORnet (comparable to IT) has within-area recurrent connections (with shared weights).

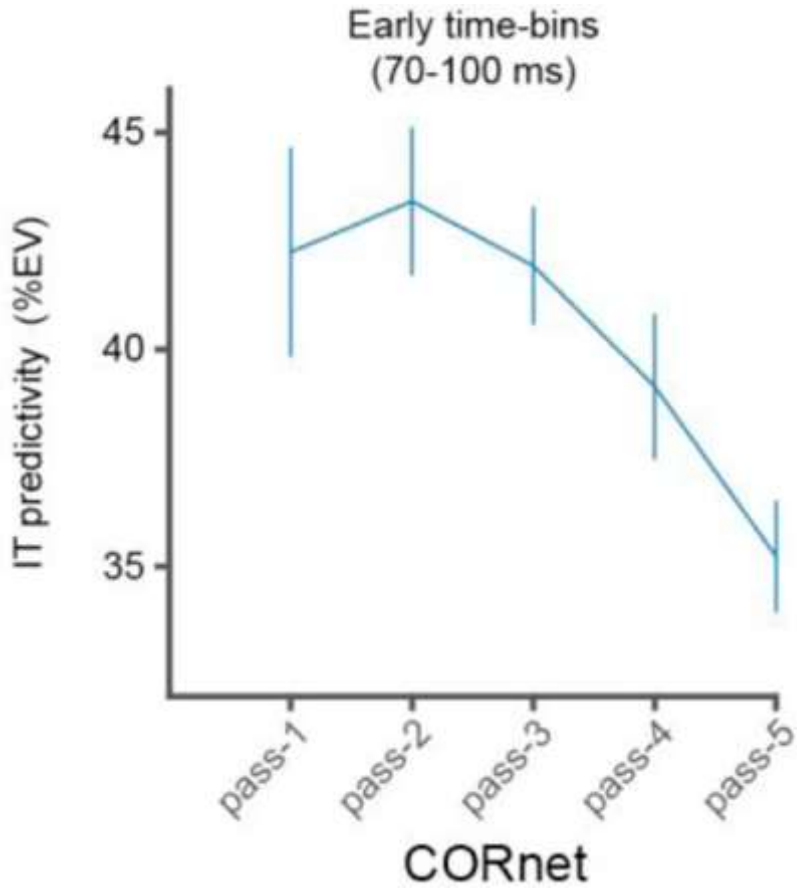
The model implements five time-steps (pass 1 to pass 5).

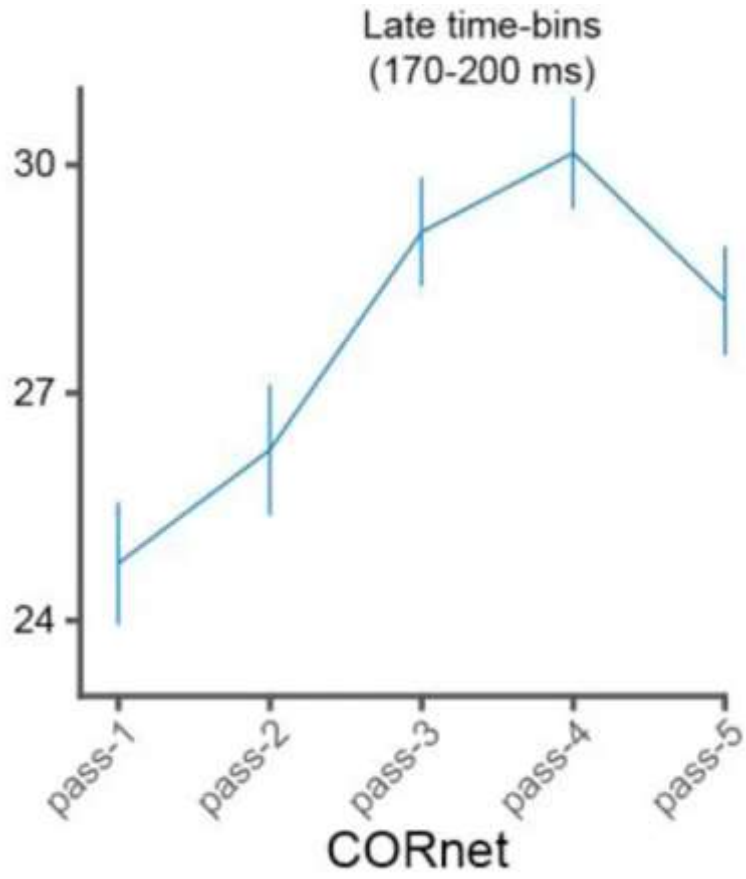
CORnet had higher IT predictivity for the late-phase of IT responses.

Pass 1 and pass 2 of the network are better predictors of the early time bins (relevant for control images).

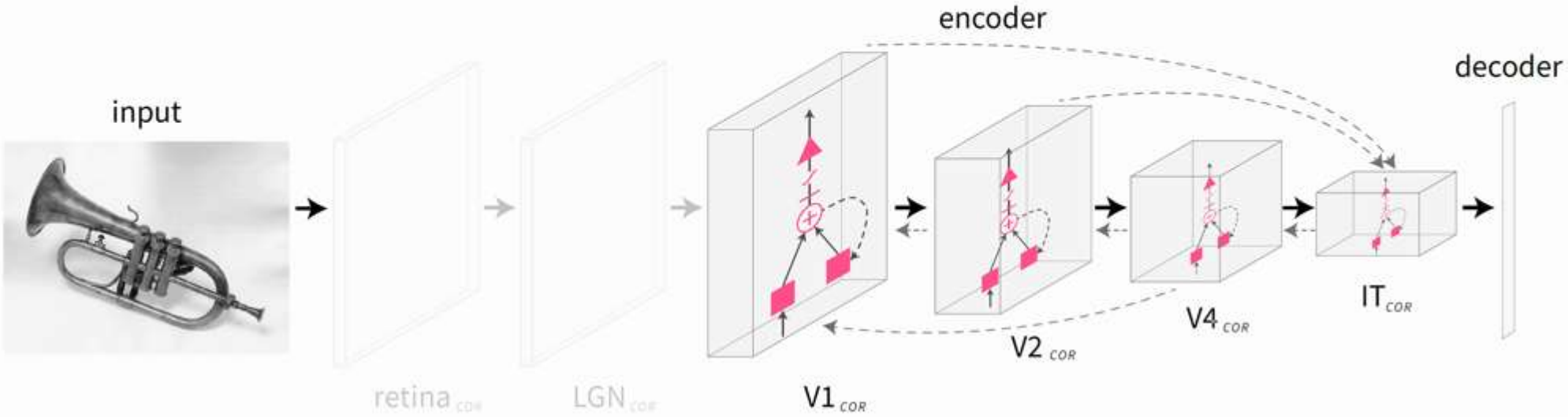
Additionally, late passes (especially pass 4) are better at predicting late (170–200 ms) phases of IT responses (crucial for challenge images).

Taken together, these results further argue for recurrent computations in the ventral stream.


Early time-bins
(70-100 ms)
IT predictivity (%EV)
45
40
35
pass-1
pass-2
pass-3
pass-4
pass-5
CORnet


Late time-bins
(170-200 ms)
30
27
24
pass-1
pass-2
pass-3
pass-4
pass-5
CORnet

CORnet Architecture



Kubilius, J., et al. CORnet: modeling the neural mechanisms of core object recognition. biorXiv (2018)

These data do not yet explain the exact nature of the computational problem solved by recurrent circuits during core object recognition.

Deeper CNNs such as inception-v3, -v4, and ResNet-50, which introduce more nonlinear transformations to the image pixels compared to shallower networks such as AlexNet, are better models of the behaviorally critical late phase of IT responses.

What computer vision has achieved by stacking more layers into the CNNs is a partial approximation of something that is more efficiently built into the primate brain architecture in the form of recurrent circuits.

During core object recognition, recurrent computations act as additional nonlinear transformations of the initial feedforward.

Image completion and RNN

Recurrent computations for visual pattern completion

Animals routinely make inferences from partial data across all cognitive domains.

In perception, pattern completion enables recognition of poorly visible or occluded objects.

Tang et al. (2018) combined psychophysics, physiology, and computational models to test the hypothesis that pattern completion is implemented by recurrent computations.

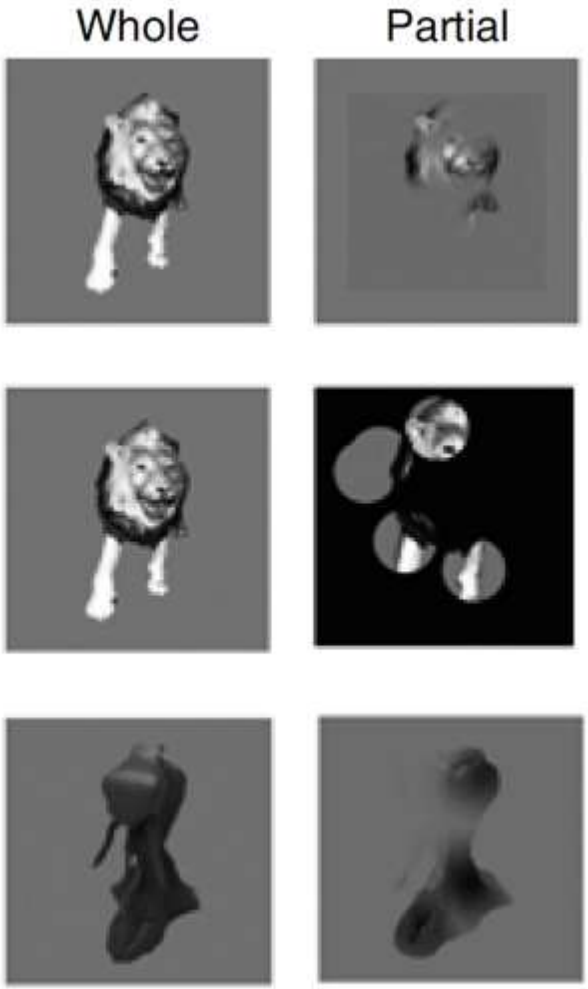
Tang et al., PNAS, 2018

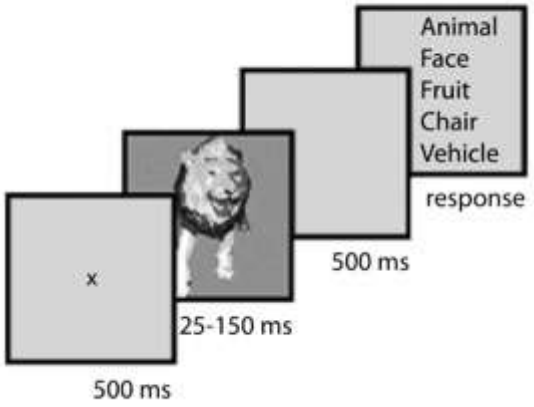
The visual system is capable of making inferences even when only 10–20% of the object is visible.

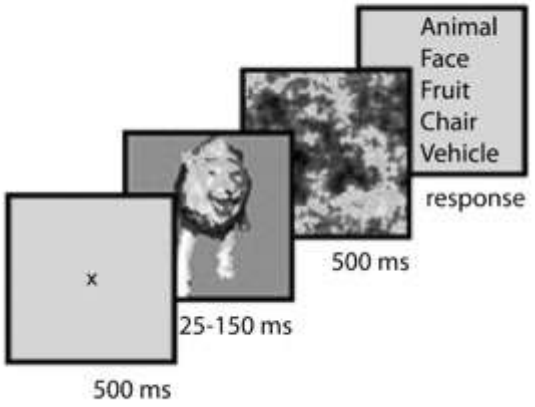
Whole

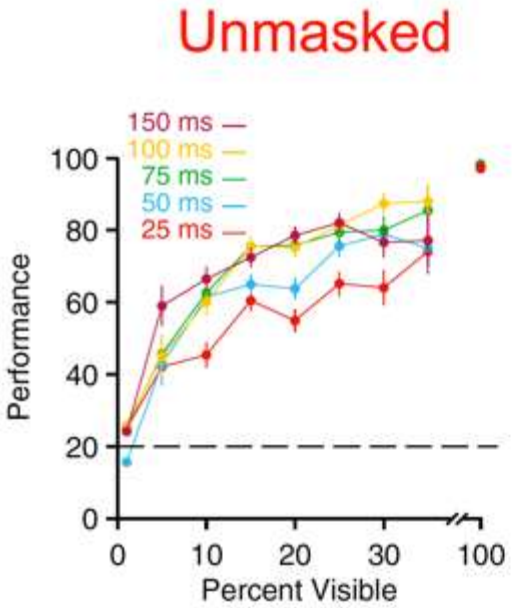
Occluded

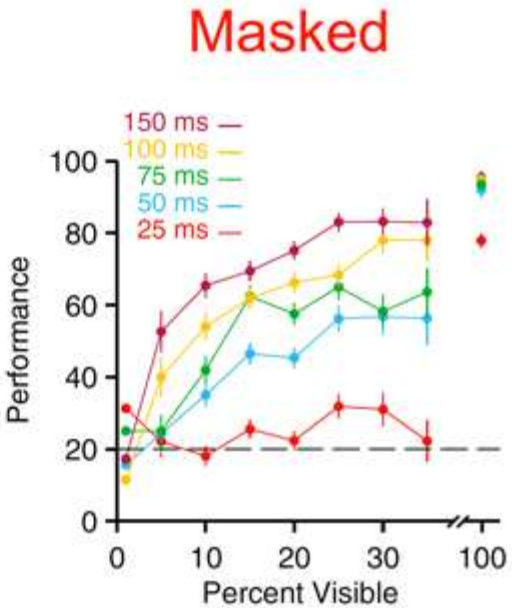
Partial











In backward masking, processing of a visual stimulus is interrupted by the presentation of a second stimulus, the mask.

Tang et al., PNAS, 2018

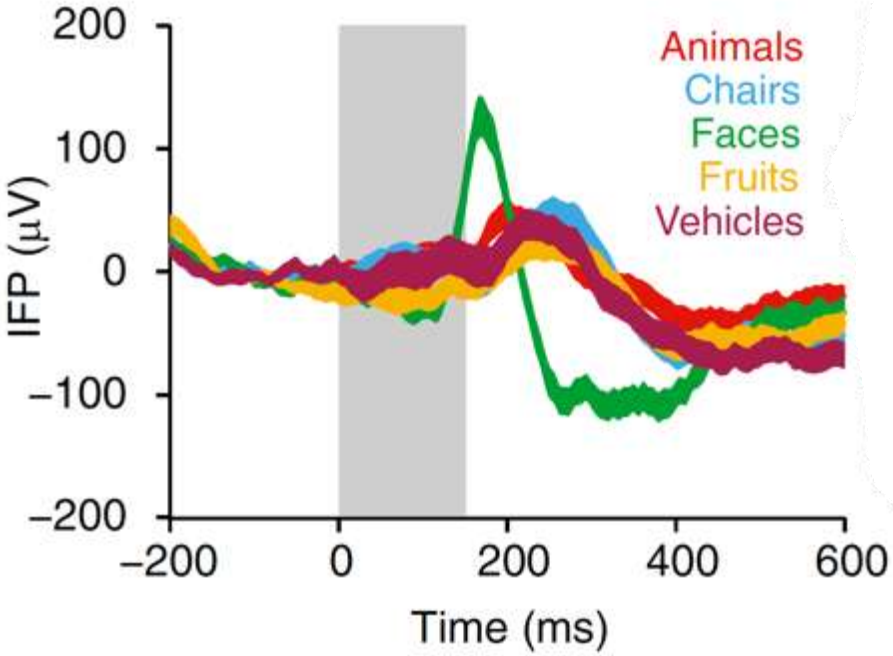
Visual categorization of objects is robust to limited visibility

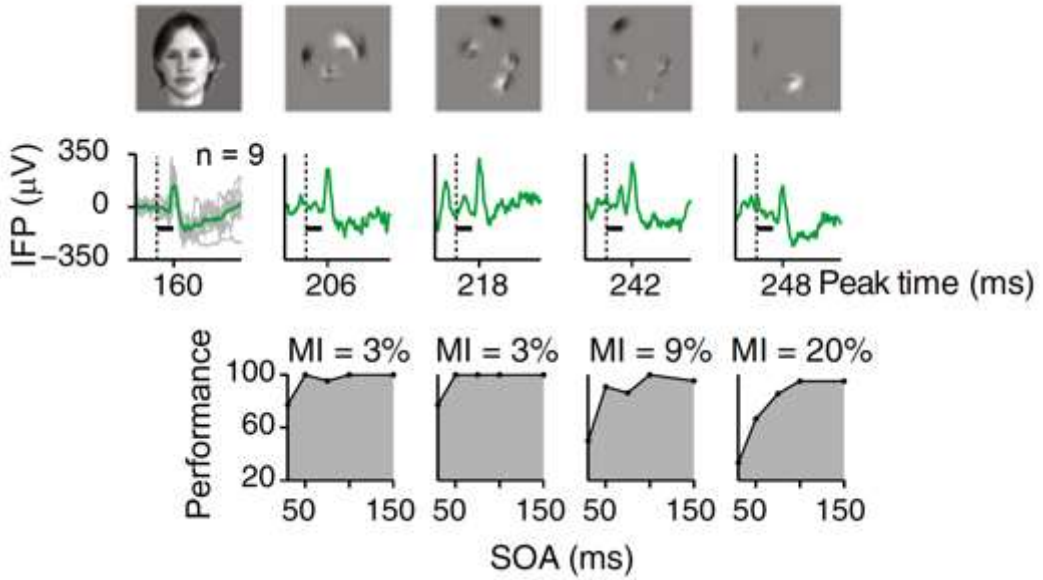
Subjects robustly recognized partial and novel objects across a wide range of visibility levels despite the limited information provided. For whole objects and without a mask, behavioral performance was at 100%

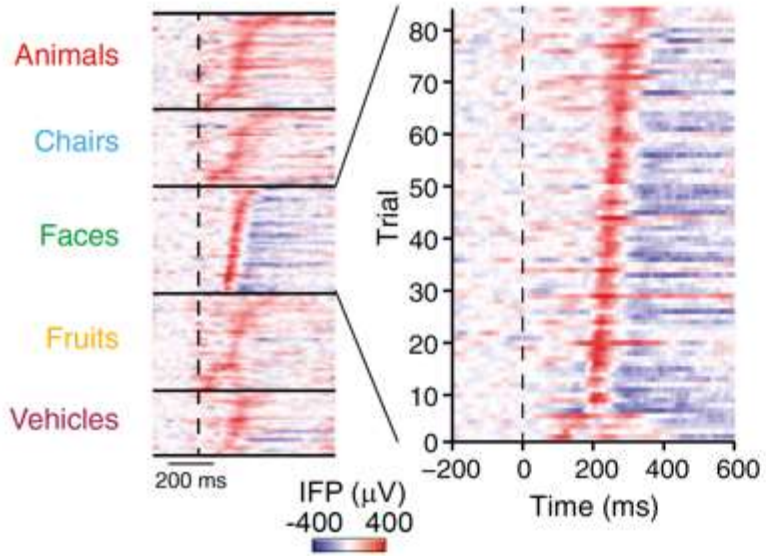
Backward masking disrupts recognition of partially visible objects.

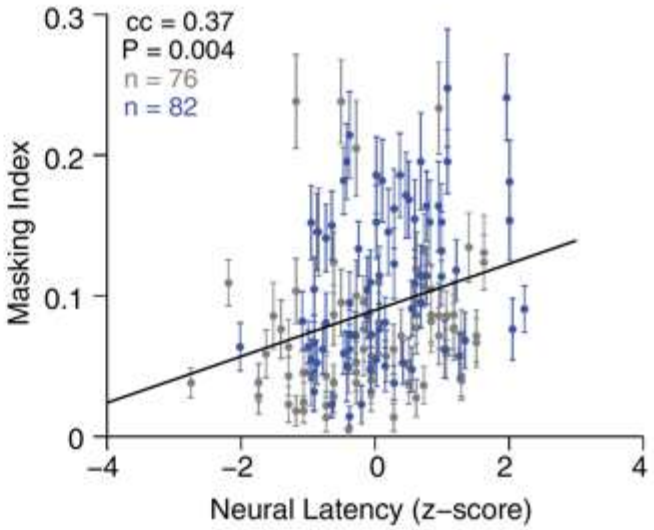
When an image is rapidly followed by a spatially overlapping mask, this interrupts any additional, presumably recurrent, processing of the image.

Tang et al., PNAS, 2018



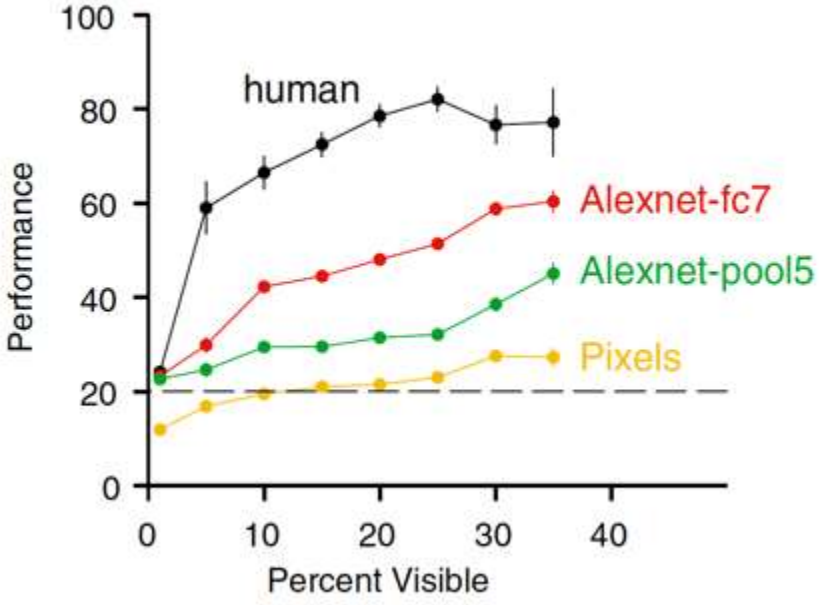


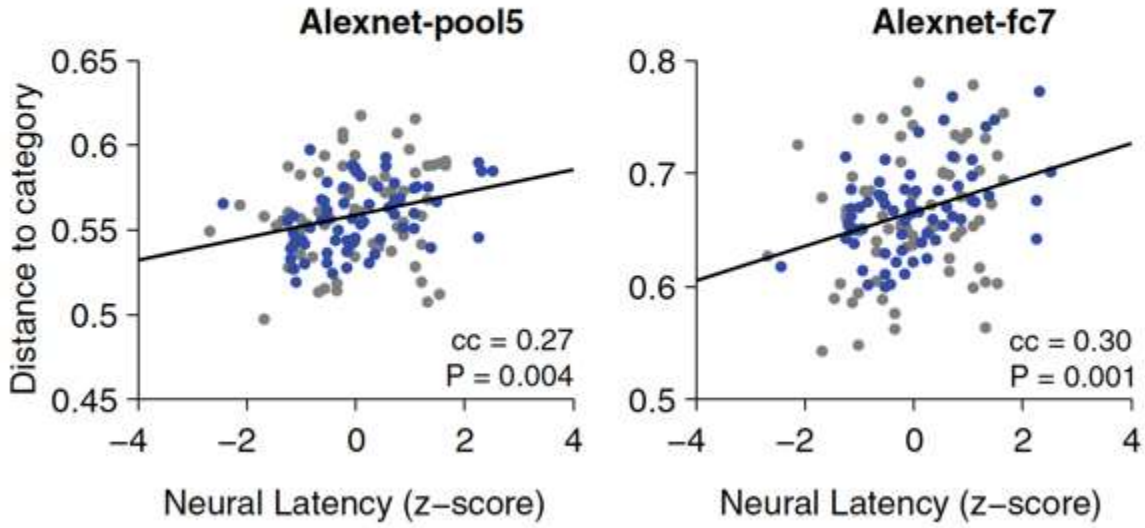




Tang et al., PNAS, 2018

Standard feed-forward models are not robust to occlusion





Performance of feed-forward models (AlexNet an 8-layer CNN trained via back-propagation on ImageNet) was evaluated using the same 325 objects (13,000 trials).

Feed-forward CNN were comparable to humans at full visibility, hover performance declined at limited visibility.

There was a modest but significant correlation at the object-by-object level between the latency of neural response and computational distance of each partial object to its whole object category mean for AlexNet pool5 and fc7 features.

Tang et al., PNAS, 2018

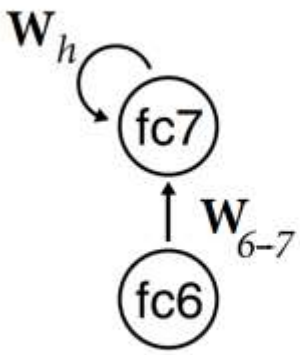
Recurrent Neural Networks improve recognition of partially visible objects

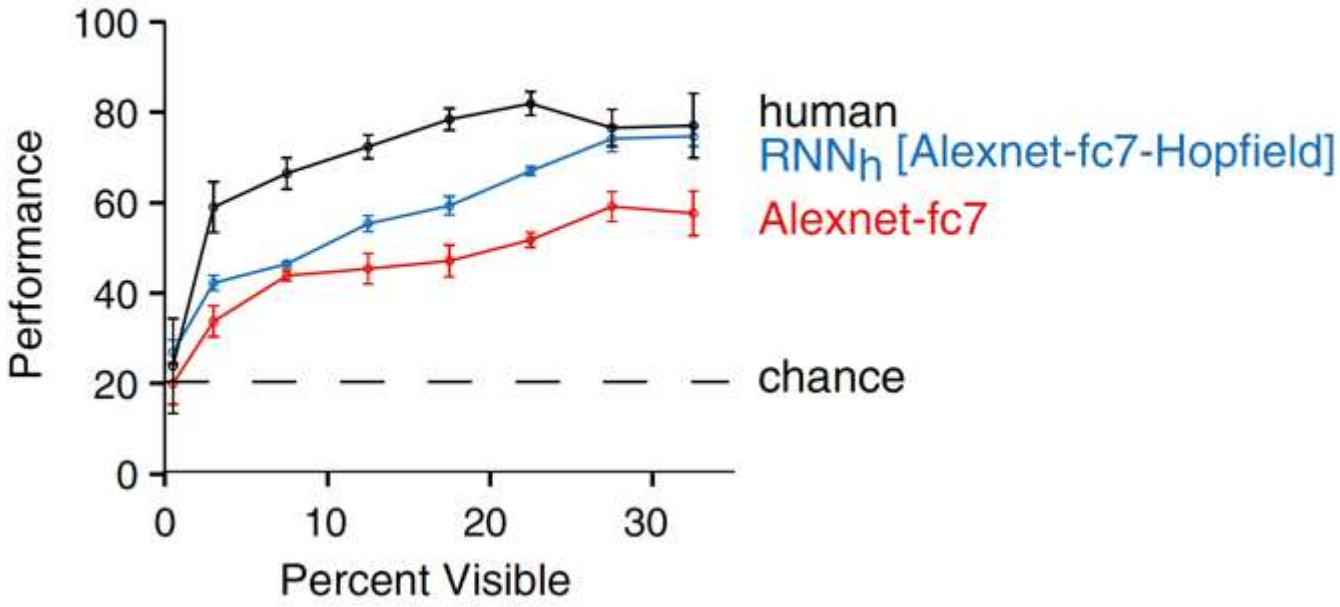
Attractor networks can perform pattern completion.

In the Hopfield network (1982), units are connected in an all-to-all fashion with weights defining fixed attractor points dictated by the whole objects to be represented. Images that are pushed farther away by limited visibility would require more processing time to converge to the appropriate attractor, consistent with the behavioral and physiological observations.

AlexNet architecture was added with recurrent connections to the fc7 layer. with one attractor for each whole object.

The RNNh model demonstrated a significant improvement over the standard AlexNet





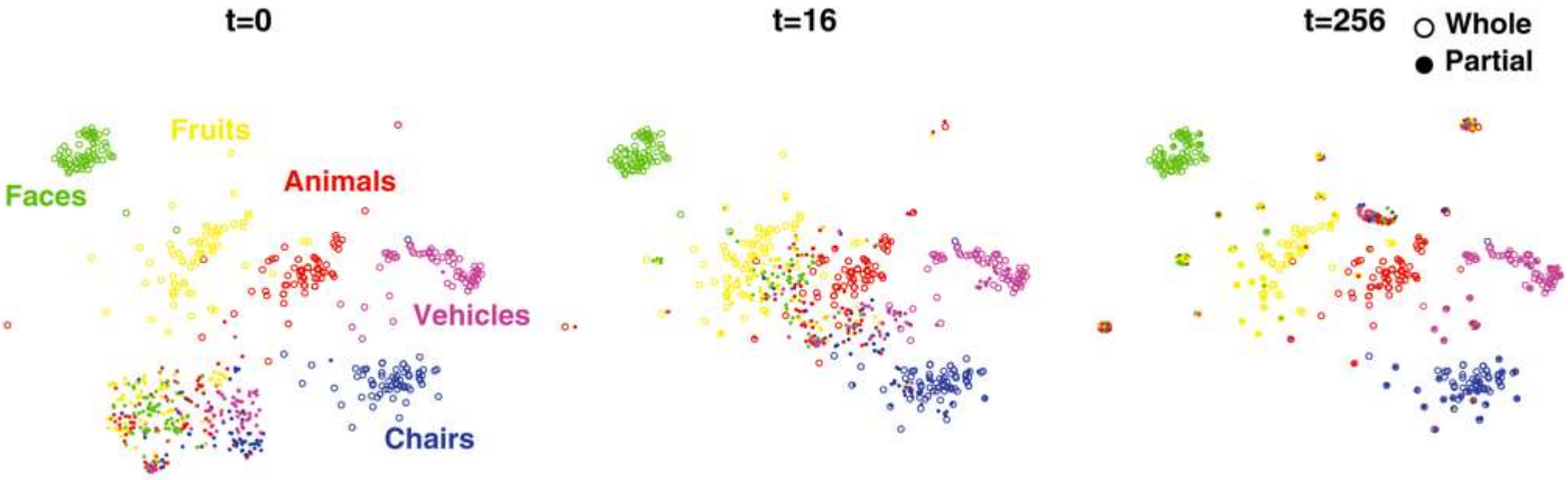
Attractor networks can perform pattern completion.

Tang et al., PNAS, 2018

Temporal evolution of the feature representation for RNN as visualized with stochastic neighborhood embedding.

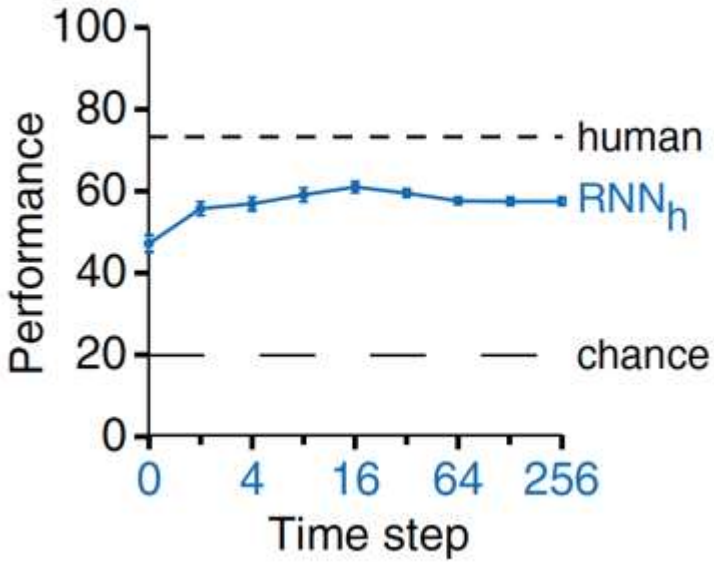
The representation of whole objects (open circles) showed a clear separation among categories, but partial objects from different categories (filled circles) were more similar to each other than to their whole object counterparts.

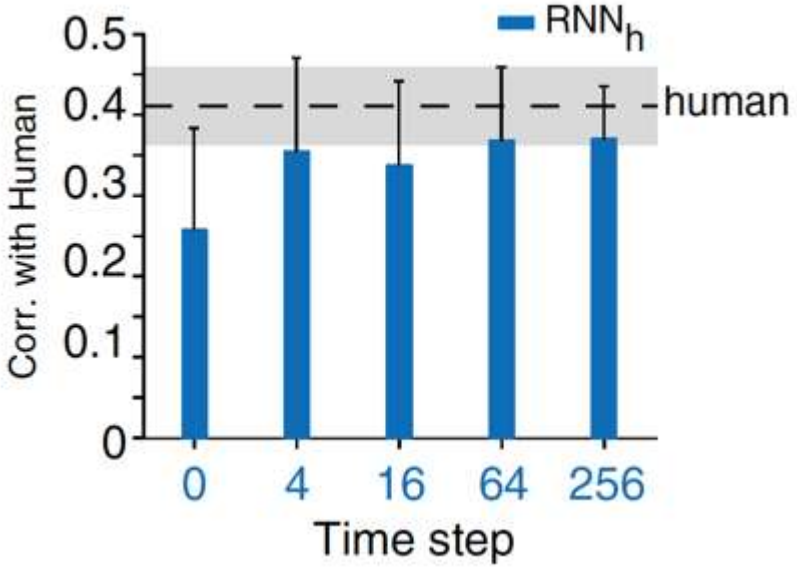
Over time, the representation of partial objects approaches the correct category in the clusters of whole images.



Correlation (Corr.) in the classification of each partially object between the RNNh and humans.

Over time, the recurrent model–human correlation increased toward the human–human upper bound.





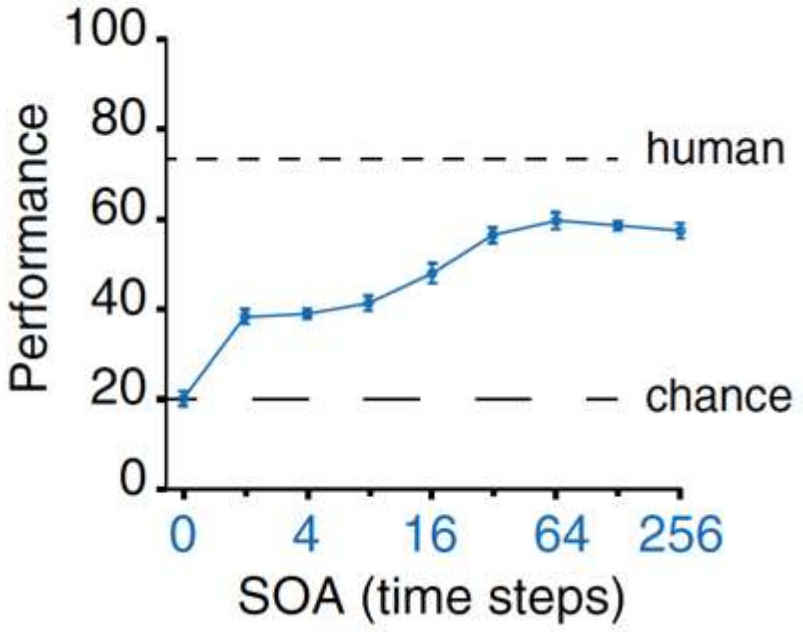
The RNNh model's performance and correlation with humans saturates at around 10–20 time steps. A combination of feed-forward signals and recurrent computations is consistent with the physiological responses to heavily occluded objects arising at around 200 ms.

Tang et al., PNAS, 2018

Backward masking impairs RNN model performance

If backward masking impairs performance by interrupting processing, this should be reproduced in the RNN model.

Presenting the mask reduced RNN performance from 58 ± 2% (SOA = 256 time steps) to 37 ± 2% (SOA = two time steps).



Tang et al., PNAS, 2018

Recognition of **partially visible objects** requires **longer reaction times** and is associated with **delayed neural representations** compared to the recognition of **whole objects.**

These delays strongly suggest that additional, more **complex computations**—beyond the initial feedforward sweep—are necessary for the brain to **interpret incomplete visual information (pattern completion)**.

Critically, when these **additional computations** are **interrupted** by a **backward mask**, recognition performance for partially visible objects is disproportionately **impaired**. This vulnerability highlights the role of **recurrent computations**, which operate over time to **complete** the visual pattern.

The **disruptive effect** of backward masking is thus linked to the **interruption** of these **recurrent loops**, which are essential for pattern completion (Kovacs et al,, PNAS, 1995).

Tang et al., PNAS, 2018

State-of-the-art feed-forward computational architectures were not robust to partial visibility.

However

- **RNNs** show improved recognition of occluded objects as they process inputs over **multiple iterations**, mimicking the **temporal unfolding** of pattern completion in the brain.

- When recurrent processing is **cut short**, similar to backward masking, model performance drops significantly, mirroring human behavioral impairments.

- Finally, recurrent models **closely approximate human performance** on partially visible objects and outperform **feedforward models**, which lack the ability to refine their representations over time.

Recurrent vs. Bottom-Up Architectures in Visual Pattern Completion

Bottom-up (feedforward) models can recognize partially visible objects, especially if unfolded from a recurrent model into many layers (an additional layer for each time step).

However, recurrent architectures offer key advantages:

- Fewer parameters (units and weights).
- Flexible number of computational steps—adapts dynamically to task difficulty.
- Temporal dynamics evolve features over time, refining partial object representations to resemble whole objects.

These dynamics mirror behavioral and neural delays:

RNNs reach human-like performance after ~10–20 time steps .

Consistent with neural responses to occluded objects emerging around 200 ms.

Conclusion: A combination of feedforward and recurrent processing is crucial for effective recognition under occlusion, aligning computational, behavioral, and physiological findings.

Unsupervised neural networks

Unsupervised neural network models of the ventral visual stream

Today's best models of visual cortex are trained on ImageNet, a dataset that contains millions of category-labeled images organized into thousands of categories.

Such supervision is however highly implausible, since human infants and nonhuman primates simply do not receive millions of category labels during development.

Supervised DCNN cannot provide a correct explanation of how such representations are learned in the brain.

Substantial effort has been devoted to unsupervised learning algorithms with the goal of learning representations from natural statistics without high-level labeling.

The developmental trajectory of object recognition robustness: Children are like small adults but unlike big deep neural networks

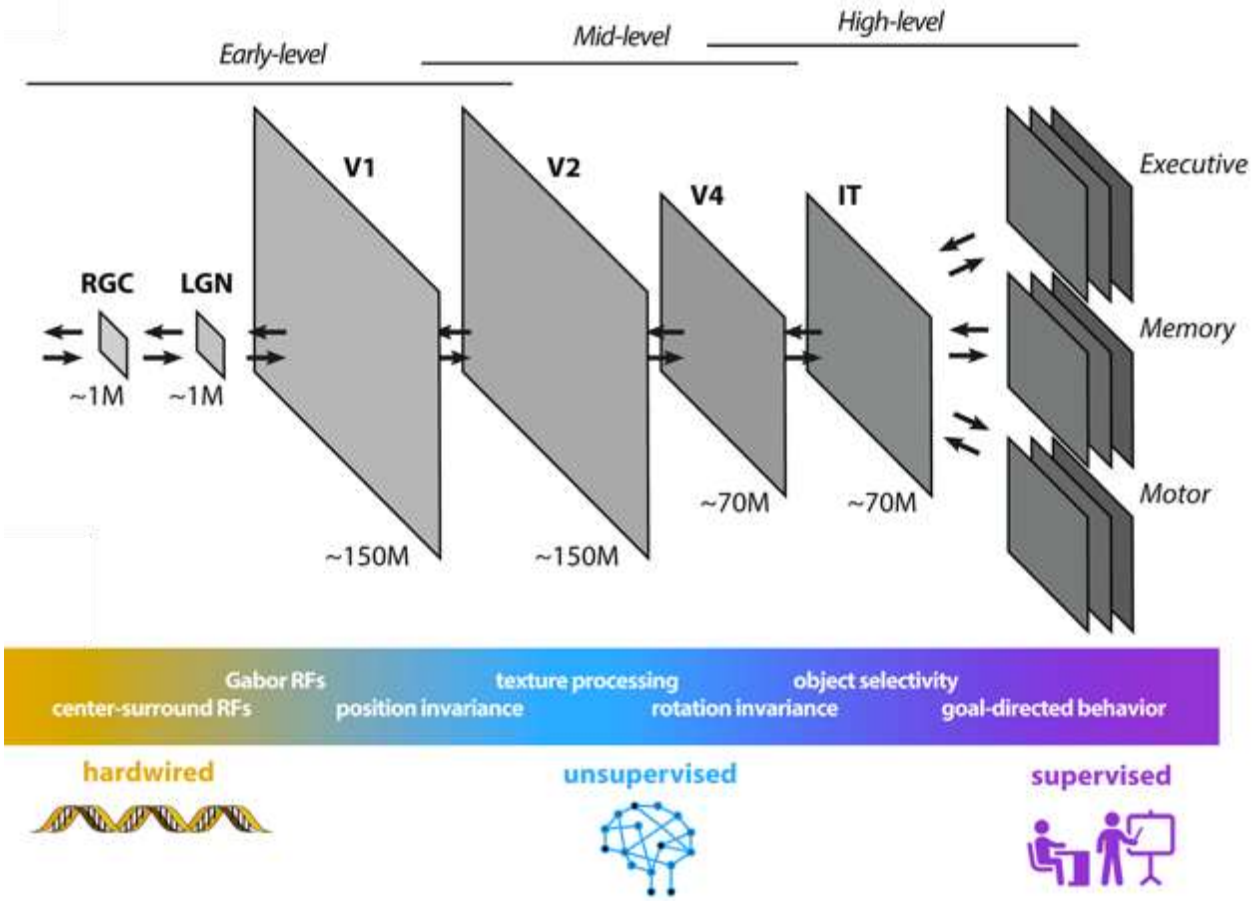
In laboratory object recognition tasks based on undistorted photographs, both adult humans and deep neural networks (DNNs) perform close to ceiling. Unlike adults', whose object recognition performance is robust against a wide range of image distortions, DNNs trained on standard ImageNet (1.3M images) perform poorly on distorted images. However, the last 2 years have seen impressive gains in DNN distortion robustness, predominantly achieved through ever-increasing large-scale datasets—orders of magnitude larger than ImageNet. Although this simple brute-force approach is very effective in achieving human-level robustness in DNNs, it raises the question of whether human robustness, too, is simply due to extensive experience with (distorted) visual input during childhood and beyond. Here we investigate this question by comparing the core object recognition performance of 146 children (aged 4–15 years) against adults and against DNNs. We find, first, that already 4- to 6-year-olds show remarkable robustness to image distortions and outperform DNNs trained on ImageNet. Second, we estimated the number of images children had been exposed to during their lifetime. Compared with various DNNs, children's high robustness requires relatively little data. Third, when recognizing objects, children—like adults but unlike DNNs—rely heavily on shape but not on texture cues. Together our results suggest that the remarkable robustness to distortions emerges early in the developmental trajectory of human object recognition and is unlikely the result of a mere accumulation of experience with distorted visual input. Even though current DNNs match human performance regarding robustness, they seem to rely on different and more data-hungry strategies to do so.

Human data is continuous and egocentric this is not the case for standard image databases; model input is most often unimodal, human input is multimodal

Humans may rely on different inductive biases—that is, constraints or assumptions prior to training (learning)—allowing for more data-efficient learning (i.e., objects obey the laws of physics and behave in a causally predictable way)

Humans may enlarge their initial dataset by using already encountered instances to create new instances during offline states (i.e., imagination, dreaming)

Huber, L. S., Geirhos, R., & Wichmann, F. A. (2023). *Journal of Vision*, *23*(7):4, 1–30,



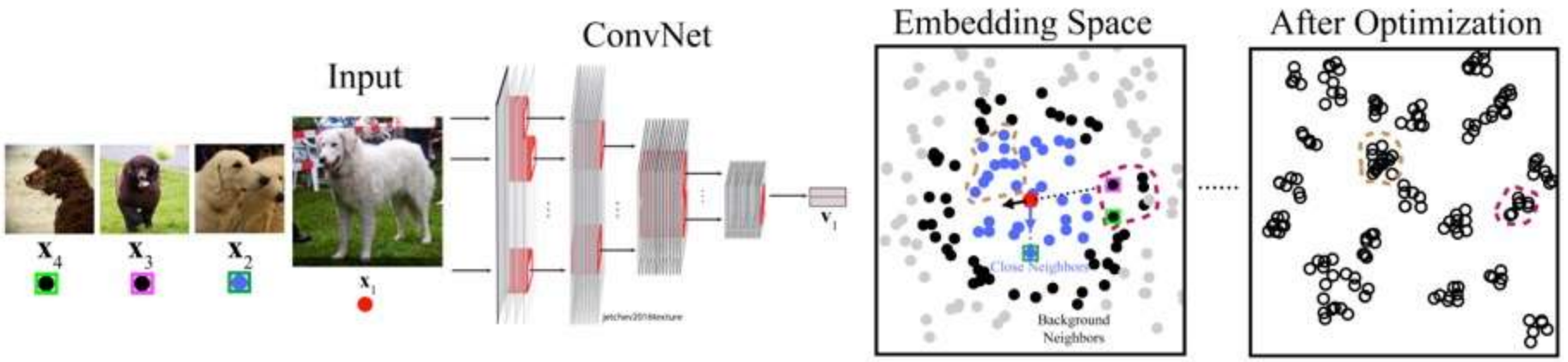
Unsupervised learning could support the continuous adaptation of cortical sensory representations to sensory input statistics, **acting as a bridge between** the largely hard-wired and evolutionary determined processing circuits of low-level areas (e.g. the retina) and the categorial/conceptual representations learned under supervision in higher-order memory/decision centers.

Unsupervised learning algorithms

Table 1. Short descriptions of optimization goals of unsupervised learning tasks

Method	Description
AutoEncoder	First embed the input images to lower-dimension space and then use the embedding to reconstruct the input
PredNet	Predict the next frame as well some of the network responses to the next frame using previous frames
CPC	Predict the embedding of one image crop using the embeddings of its spatial neighbors
Depth prediction	Predict the per-pixel relative depth image from the corresponding RGB image
Relative position	Predict the relative position of two image crops sampled from a 2×2 image grid
Colorization	Predict the down-sampled color information from the grayscale image
Deep cluster	Embed all images into a lower-dimension space and then use unsupervised clustering results on these embeddings as "category" labels to train the networks
CMC	Embed grayscale and color information of one image into two embedding spaces and push together two corresponding embeddings while separating them from all of the other embeddings
Instance recognition	Make the embedding of one image unchanged under data augmentations while separating it from the embeddings of all of the other images
SimCLR	Aggregate the embeddings of two data-augmented crops from one image while separating them from the embeddings of other images in one large batch
Local aggregation	Aggregate the embeddings of one image to its close neighbors in the embedding space while separating them from further neighbors

CPC represents contrastive predictive coding (33).

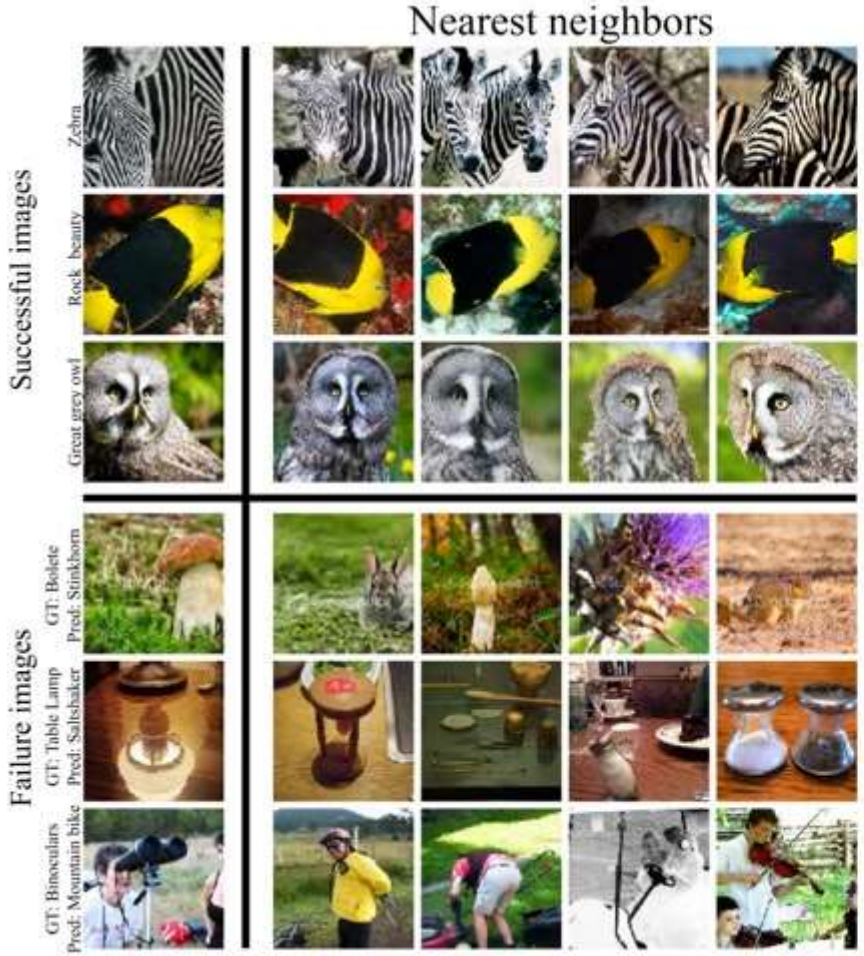


Local Aggregation (LA) method.

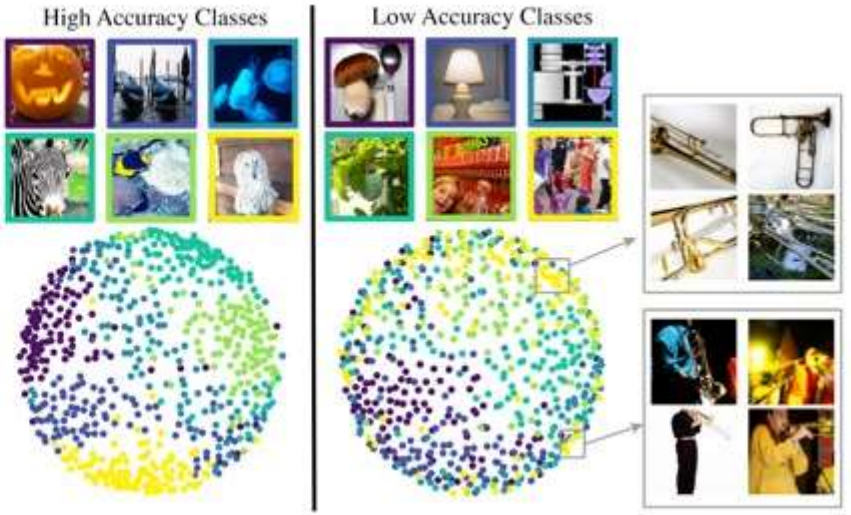
For each input image, a DCNN was used to embed it into a lower dimension space (”Embedding Space”).

Its close neighbors (blue dots) and background neighbors (black dots) were identified.

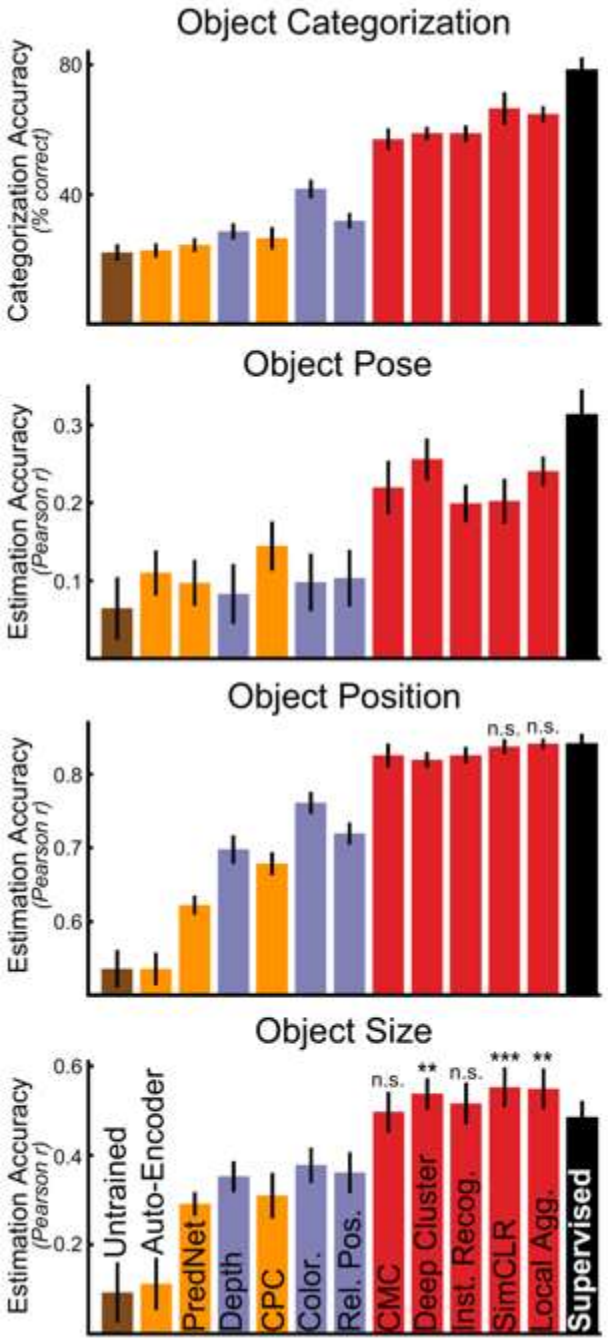
The optimization seeks to push the current embedding vector (red dot) closer to its close neighbors and further from its background neighbors. The blue arrow and black arrow are examples of influences from different neighbors on the current embedding during optimization. ”After Optimization” panel illustrates the typical structure of the final embedding after training.



Top three rows show the images that were successfully classified using a weighted K-nearest-neighbour (KNN) classifier in the embedding space (K top nearest neighbours), while Bottom three rows show unsuccessfully classified examples.



Multi-dimensional scaling (MDS) algorithm used to visualize the embedding space.
Classes with high validation accuracy (left) and classes with low validation accuracy (right). For each class, 100 images of that class were randomly chosen from the training set and apply the MDS algorithm to the resulting 600 images. Dots represent individual images in each color-coded category.



Contrastive embedding methods yield high-performing neural networks

A standard ResNet18 network architecture was used. Training data were drawn from ImageNet, a large-scale database of hand-labeled natural images.
Across all evaluated objective functions, contrastive embedding objectives showed substantially better transfer than other unsupervised methods.

The best of the unsupervised methods (Sim-CLR and local aggregation) equaled or even outperformed the category-supervised model in several tasks, including object position and size estimation.

Unsurprisingly, all unsupervised methods are still somewhat outperformed by the category-supervised model on the object categorization task.

Unsupervised neural network were compared to neural data from macaque V1, V4, and IT cortex.

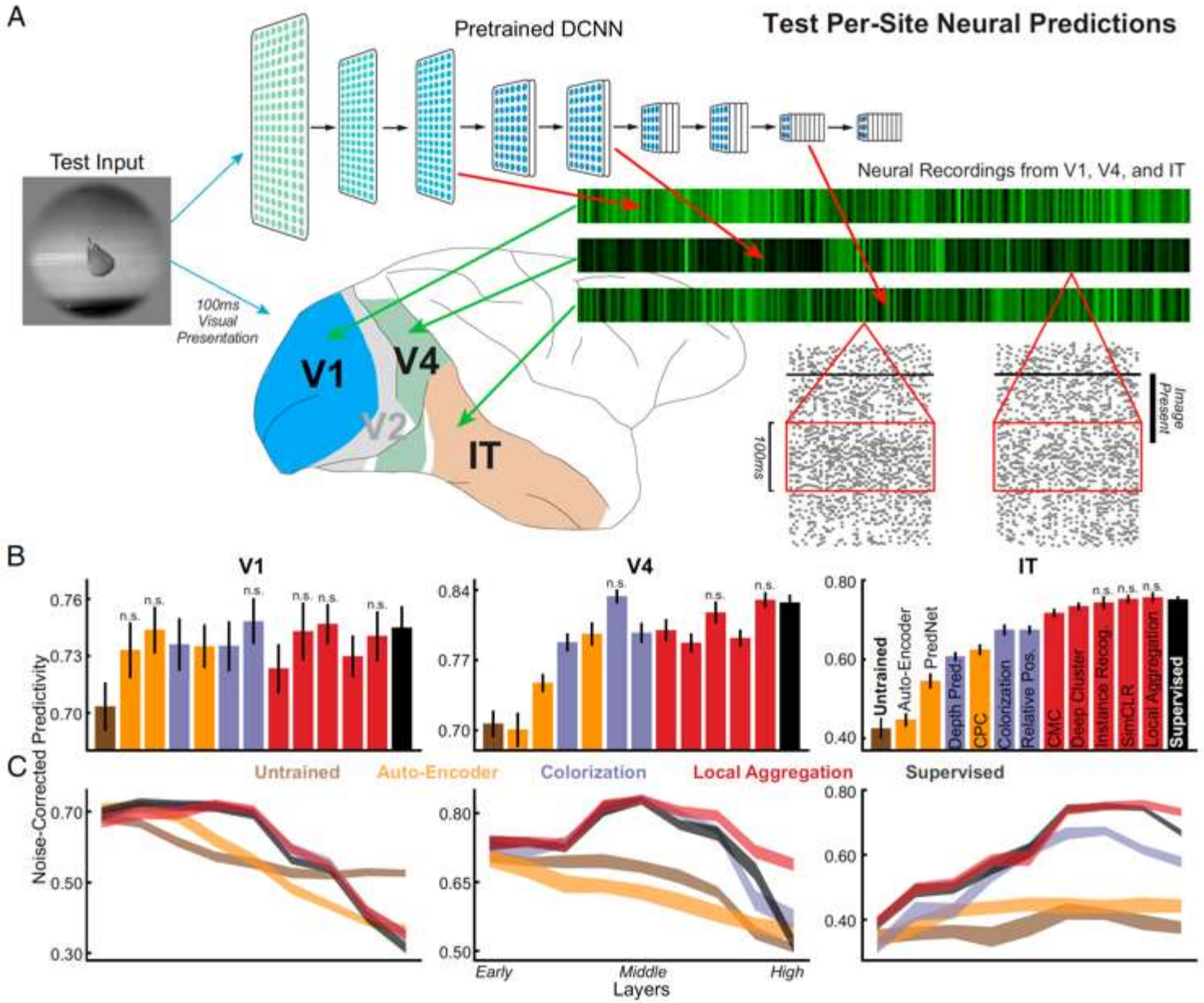
Previously established technique for mapping artificial network responses to real neural response patterns were used.

The figure in the next slide shows the correlation between model and neural responses across held-out images, for the best-predicting layer for each model.

Area V1: all unsupervised methods were significantly better than the untrained baseline at predicting neural responses, although none were statistically better from the category-supervised model on this metric.

Area V4: only a subset of methods achieved parity with the supervised model in predictions of responses.

Area IT: only the best-performing contrastive embedding methods achieved neural prediction parity with supervised models.



Unsupervised neural network were compared to neural data from macaque V1, V4, and IT cortex

Deep Contrastive Learning on first-person video data from children

ImageNet dataset used to train unsupervised networks diverges significantly from real biological data streams

- ImageNet contains single images of a large number of distinct instances of objects in each category, presented cleanly from stereotypical angles;
- Human infants receive images from a much smaller set of object instances, under much noisier continuous;

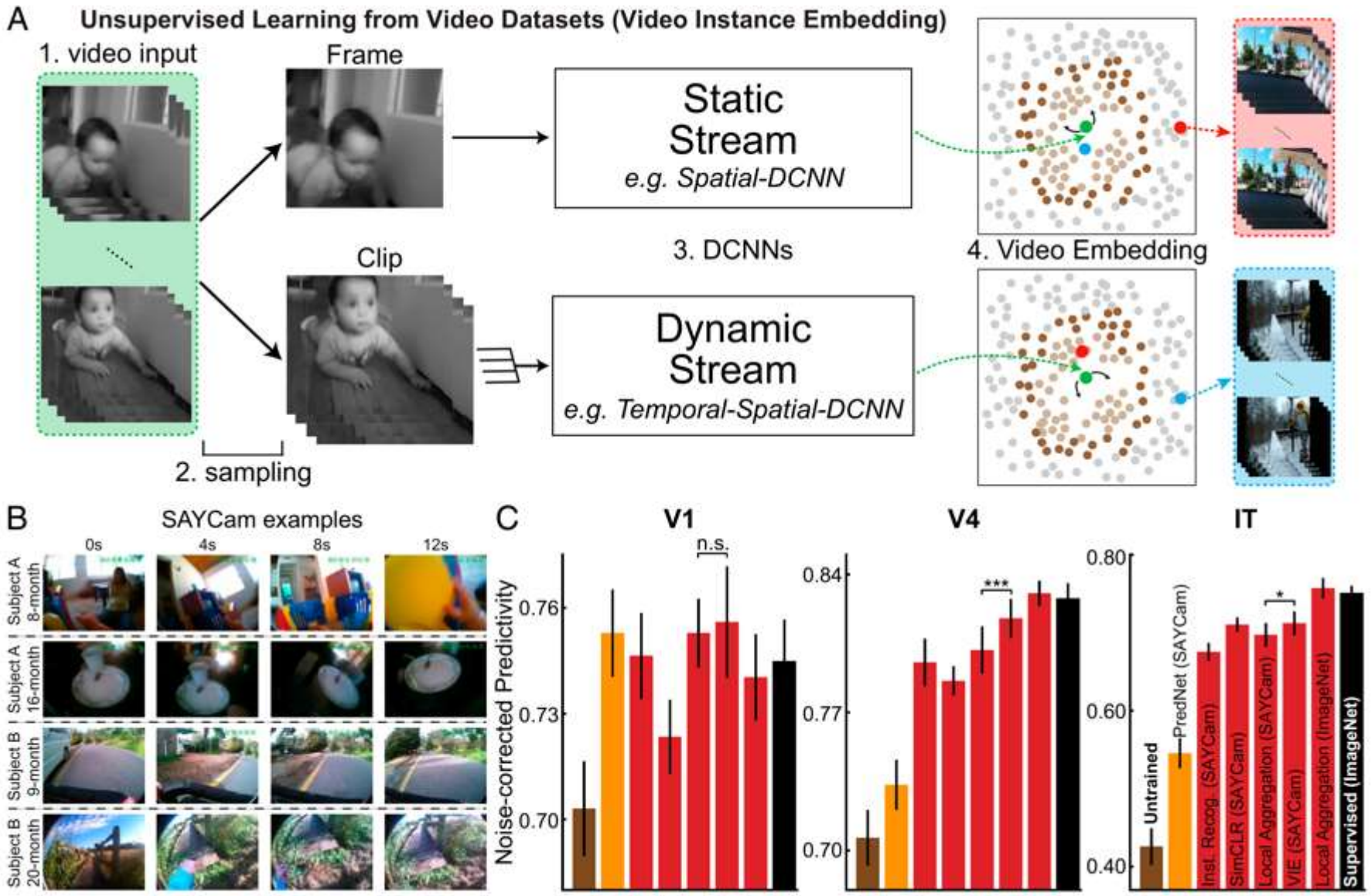
- ImageNet consists of statistically independent static frames;
- Human infants receive streams of temporally correlated inputs;

Is deep contrastive unsupervised learning sufficiently robust to handle real-world developmental video streams such as SAYCam?

A better proxy of the real infant data stream is represented by the recently released SAYCam dataset, which contains head-mounted video camera data from three children (about 2 h/wk spanning ages 6 to 32 mo).

To test whether contrastive unsupervised learning is sufficiently robust to handle real-world developmental video streams such as SAYCam, video instance embedding (VIE) algorithm was used

VIE algorithm is an extension of LA to video, which achieves state-of-the-art results on a variety of dynamic visual task, including action recognition.



Representations learned by VIE are highly robust approaching the neural predictivity of those trained on ImageNet

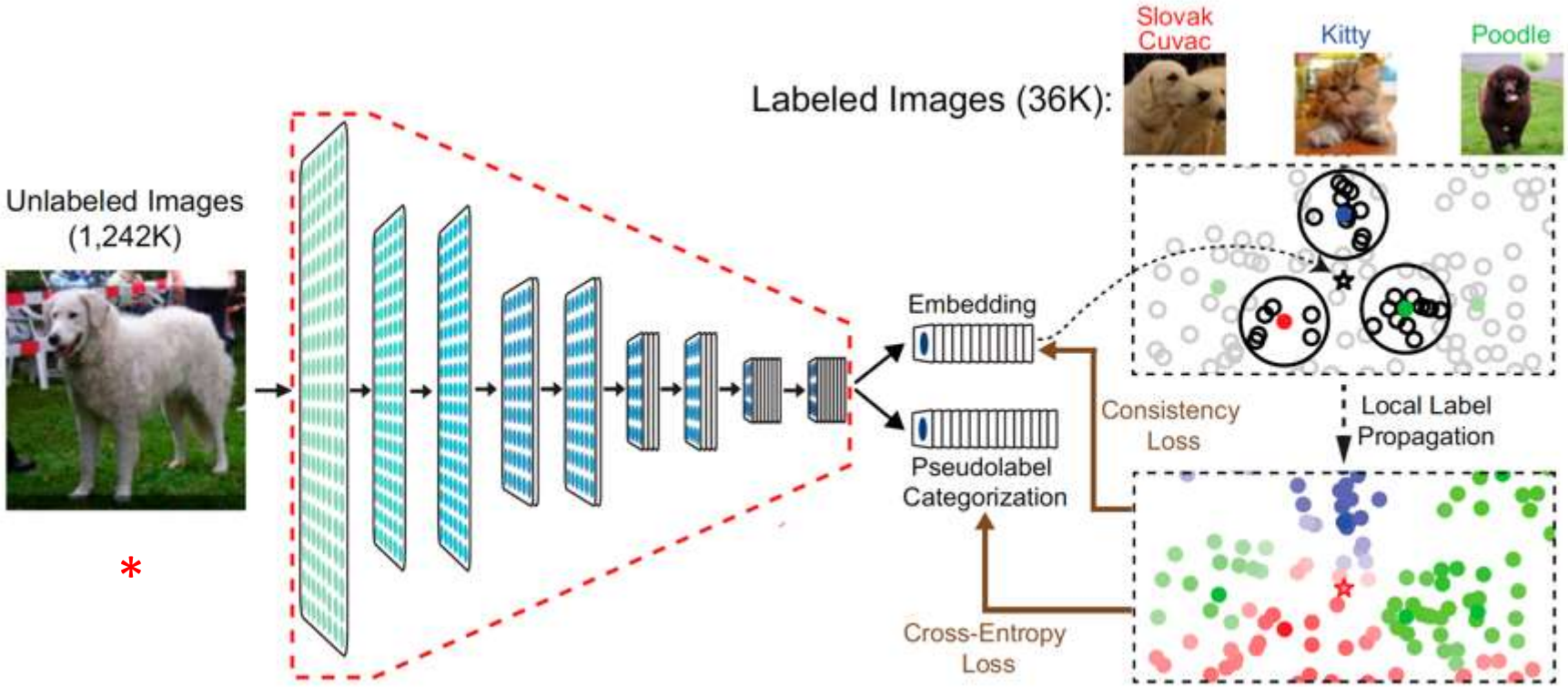
Semisupervised learning seeks to leverage small numbers of labeled datapoints in the context of large amounts of unlabeled data.

A semisupervised learning algorithm, local label propagation (LLP) embeds datapoints into a compact embedding space, but additionally takes into account the embedding properties of sparse labeled data

This algorithm first uses a label propagation method to infer the pseudolabels of unlabeled images from those of nearby labeled images.

The network is then jointly optimized to predict these inferred pseudolabels while maintaining contrastive differentiation between embeddings with different pseudolabels.

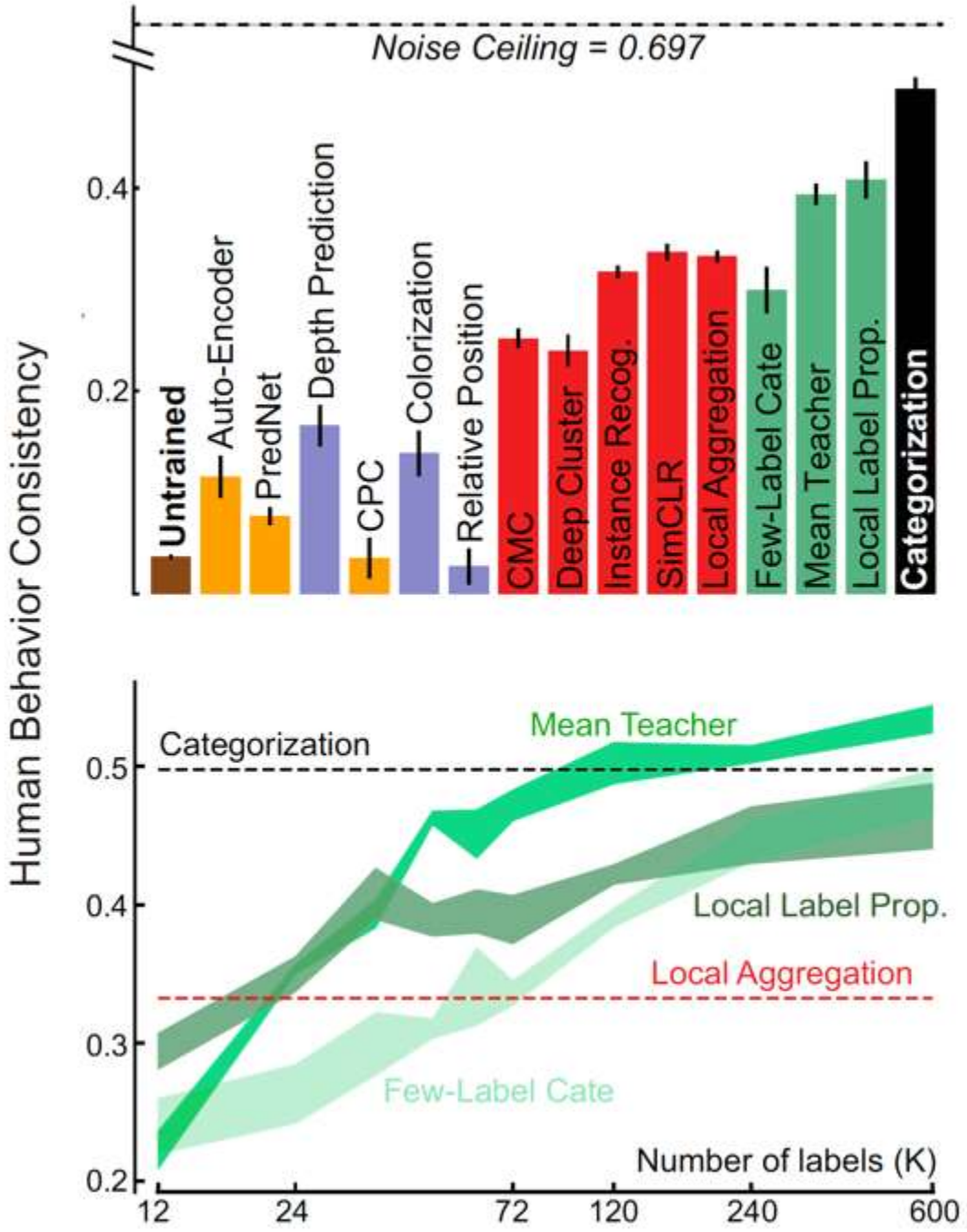
Semi-Supervised Learning: Local Label Propagation



The embedding * of an unlabeled input is used to infer its pseudolabel considering its labeled neighbors with voting weights determined by their distances from * and their local density (the highlighted areas).

DCNNs is jointly optimized to predict these inferred pseudolabels while maintaining contrastive differentiation between embeddings with different pseudolabels

.



Pearson correlations between human and different models' behavior performing the same object recognition task on 2400 images of 24 different objects

Using just 36,000 labels (corresponding to 3% supervision), semisupervised models lead to representations that are substantially more behaviorally consistent than purely unsupervised methods, although a gap to the supervised models remains.

Unsupervised models represent high-performing but biologically plausible visual learning system.

The neural predictivity of the best unsupervised method only slightly surpasses that of supervised categorization models.

Both for neural response pattern and behavioral consistency metrics, results show that there remains a substantial gap between all models (supervised and unsupervised) and the noise ceiling (variance) of the data.

Known Misalignments Between Brains and HCNNs

Robustness to Noise & Texture Bias (Geirhos et al., 2018):

- HCNNs like VGG-19 and ResNet-152 are less robust than humans to Gaussian noise in object recognition tasks.
- HCNNs tend to rely more on texture cues.
- Humans rely more on shape for object recognition

Adversarial Vulnerability (Goodfellow et al., 2014):

- Small, imperceptible pixel changes can drastically alter HCNN predictions (e.g., "dog" → "church").
- Humans are largely insensitive to these perturbations.
- Newer models show improved alignment but gaps remain (Gaziv et al., 2023).

Visual Perception Phenomena

HCNNs struggle with:

- Global shape processing.
- Object part relationships.
- Illusory and uncrowding effects.