UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

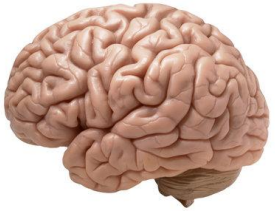
> Reinventa el mundo <

Deep Learning

Week 7 : NeuroAI Origin, Basic Principles, and Introduction to Psychophysics in Humans & Machines

TRANSFORMATEC

Neuroscience



Neuroscience is the scientific discipline that consists on the study and discovery of the mechanisms of processing and representation in the brain and nervous systems of biological systems.

What the Frog's Eye Tells the Frog's Brain*

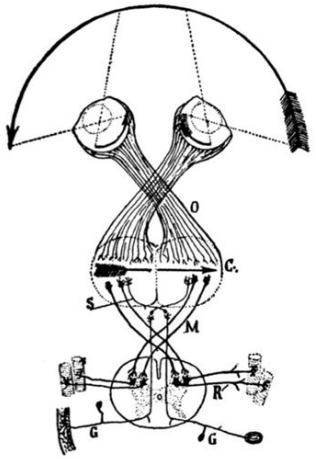
J. Y. LETTVIN†, H. R. MATURANA‡, W. S. McCULLOCH‖, SENIOR MEMBER, IRE, AND W. H. PITTS‖

Summary—In this paper, we analyze the activity of single fibers in the optic nerve of a frog. Our method is to find what sort of stimulus causes the largest activity in one nerve fiber and then what is the exciting aspect of that stimulus such that variations in everything else cause little change in the response. It has been known for the past 20 years that each fiber is connected not to a few rods and cones in the retina but to very many over a fair area. Our results show that for the most part within that area, it is not the light intensity itself but rather the pattern of local variation of intensity that is the exciting factor. There are four types of fibers, each type concerned with a different sort of pattern. Each type is uniformly distributed over the whole retina of the frog. Thus, there are four distinct parallel distributed channels whereby the frog's eye informs his brain about the visual image in terms of local pattern independent of average illumination. We describe the patterns and show the functional and anatomical separation of the channels. This work has been done on the frog, and our interpretation applies only to the frog.

INTRODUCTION

Behavior of a Frog

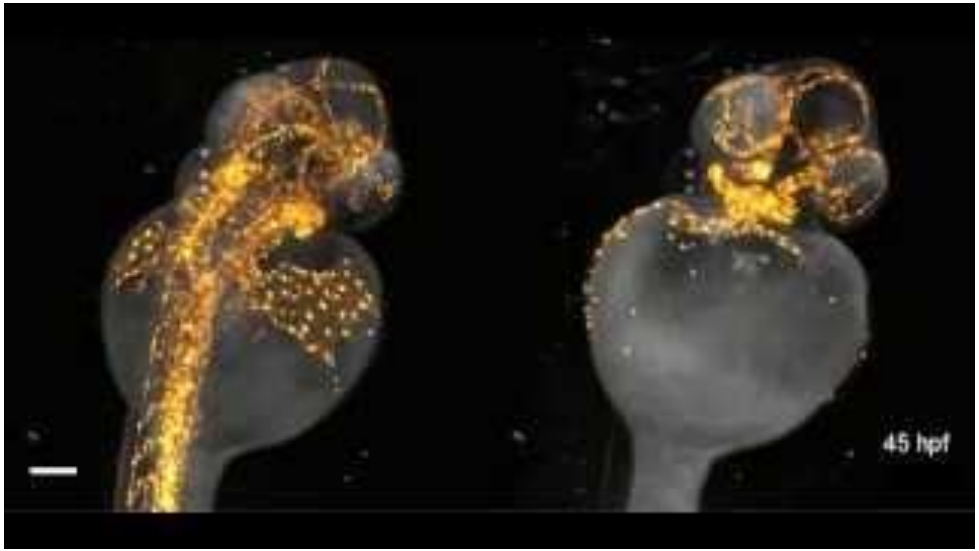
A FROG hunts on land by vision. He escapes enemies mainly by seeing them. His eyes do not move, as do ours, to follow prey, attend suspicious events, or search for things of interest. If his body changes its position with respect to gravity or the whole visual world is rotated about him, then he shows compensatory eye movements. These movements enter his hunting and evading habits only, *e.g.*, as he sits on a



InVivo / Experimental of Frog

Neuroscience

Neuroscience is the scientific discipline that consists on the study and discovery of the mechanisms of processing and representation in the brain and nervous systems of biological systems.



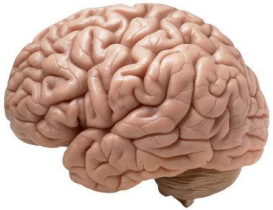
InVivo / Experimental of ZebraFish

Neuroscience

Neuroscience is the scientific discipline that consists on the study and discovery of the mechanisms of processing and representation in the brain and nervous systems of biological systems.

Behaviour in Goats

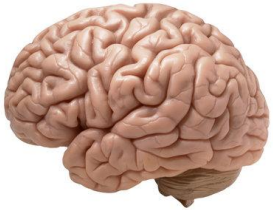
Neuroscience



Neuroscience is the scientific discipline that consists on the study and discovery of the mechanisms of processing and representation in the brain and nervous systems of biological systems.

Anomalous Behaviour in Humans (Prosopagnosia)

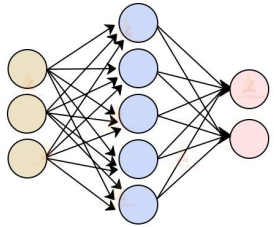
Neuroscience



Neuroscience is the scientific discipline that consists on the study and discovery of the mechanisms of processing and representation in the brain and nervous systems of biological systems.

Anomalous Behaviour in Humans (Prosopagnosia)

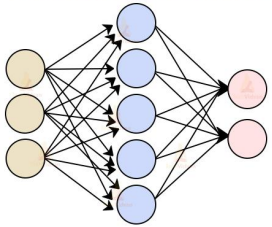
Artificial Intelligence



Artificial Intelligence consists of the engineering discipline that has the ultimate goal of
creating and understanding intelligent machines.

AI for Learning + Control

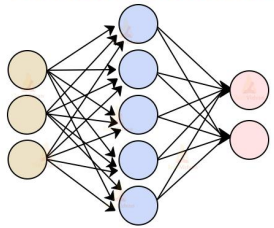
Artificial Intelligence



Artificial Intelligence consists of the engineering discipline that has the ultimate goal of
creating and understanding intelligent machines.

Computer Vision

Artificial Intelligence

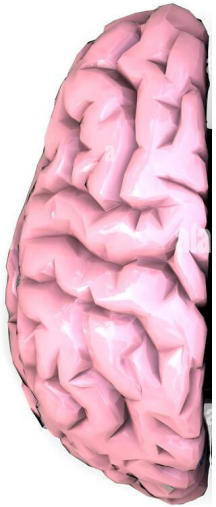


Artificial Intelligence consists of the engineering discipline that has the ultimate goal of
creating and understanding intelligent machines.

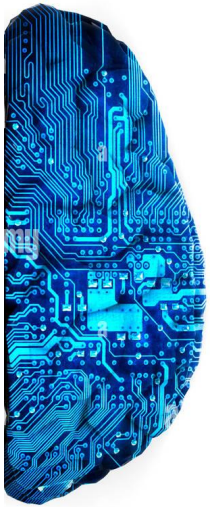
Natural Language Processing &
Multi-Modal Understanding

NeuroAI

How do we use Neuroscience to invent better systems of Artificial Intelligence (AI)



NeuroAI

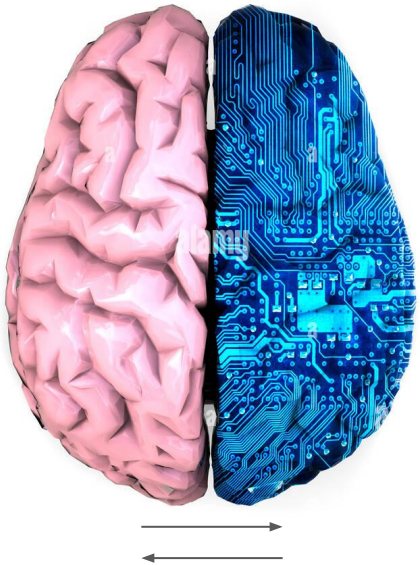


How do we use Artificial Intelligence to discover principles in Neuroscience?

NeuroAI

Engineering

How do we use Neuroscience to invent better systems of Artificial Intelligence (AI)

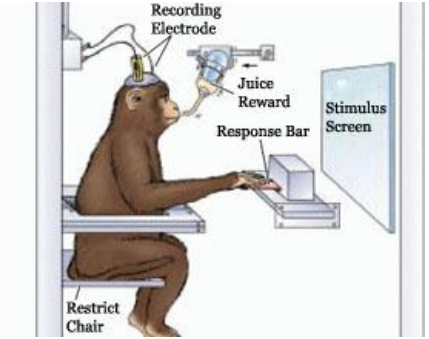


Science

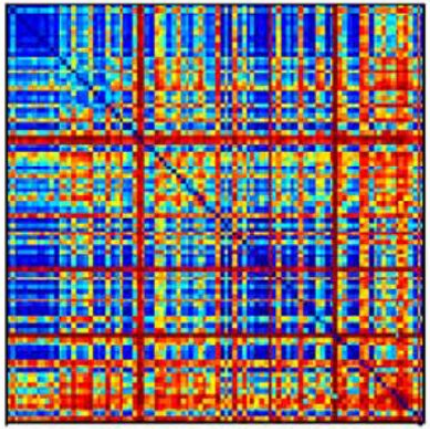
How do we use Artificial Intelligence to discover principles in Neuroscience?

Is a symbiotic new discipline between Artificial Intelligence and Neuroscience that mixes both aspects of science and engineering

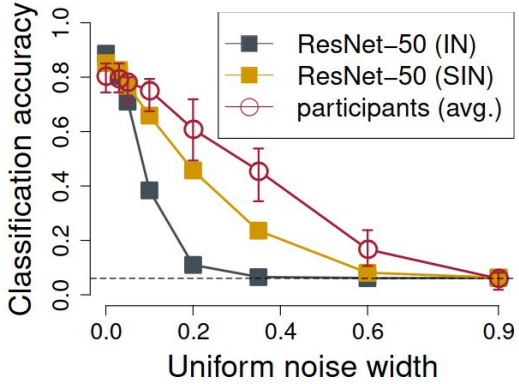
Feature Representation & Data Collection



Representational Similarity Analysis (RSA)



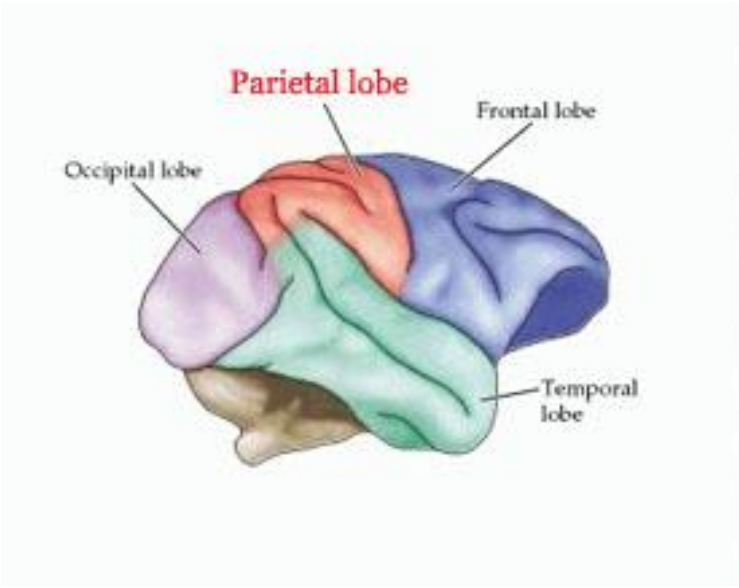
PsychoPhysics



Assessment of Internal Representation

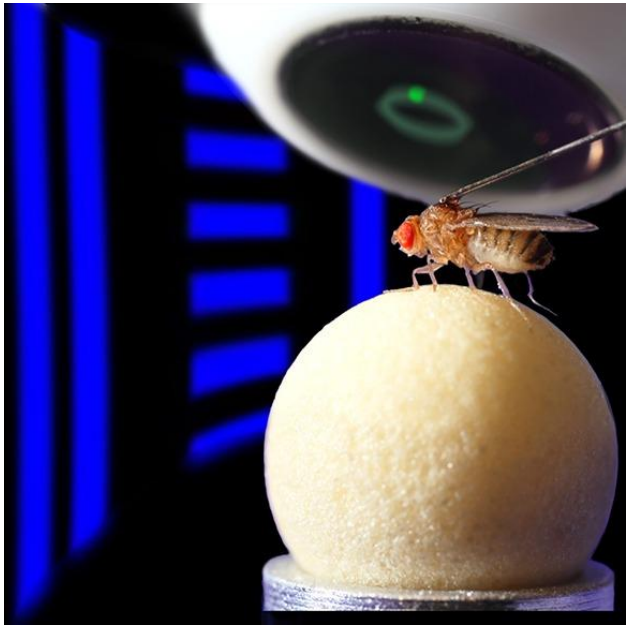
Assessment of Behaviour/Output

Invasive Data Collection (Biology / Neuro)



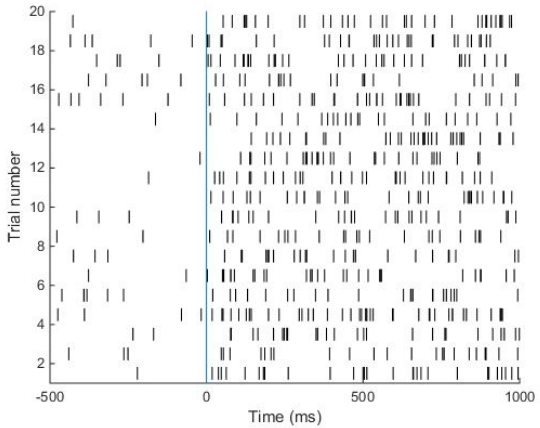
Non-Invasive Data Collection (Biology / Neuro for Monkeys)

Semi-Invasive Data Collection (Biology / Neuro for Flies)

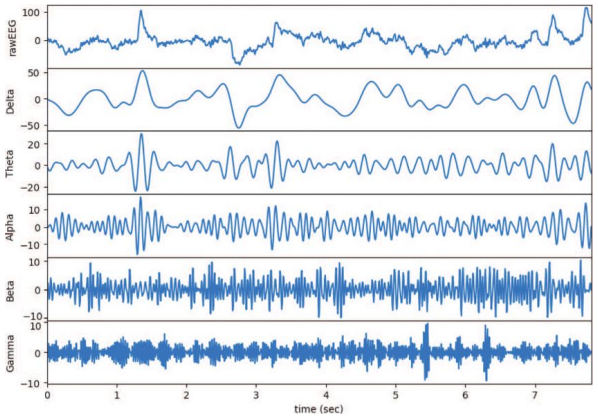


Some example of how Bio/Neuro Data collection looks like in Humans

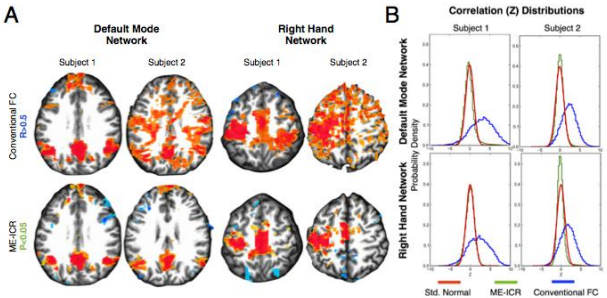
Neurophysiology



EEG

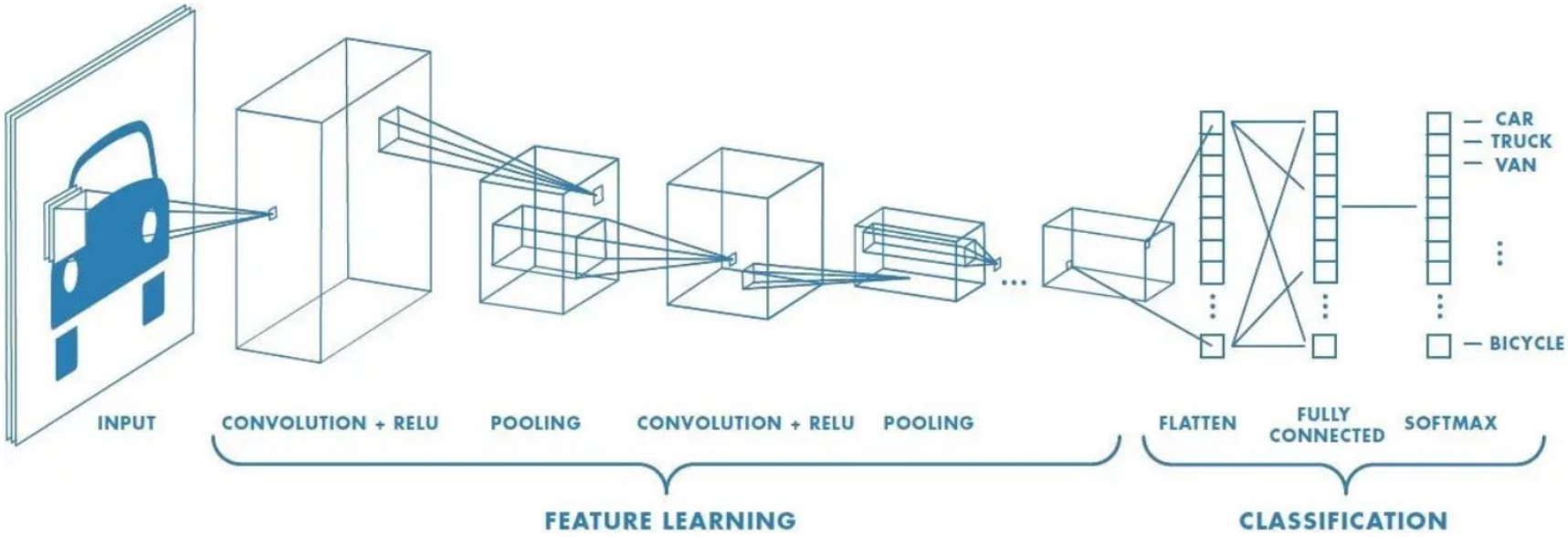


fMRI

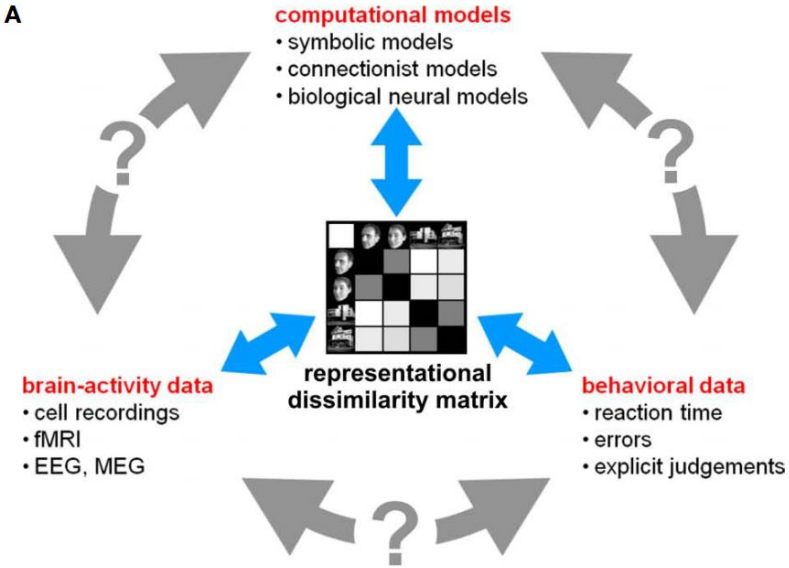


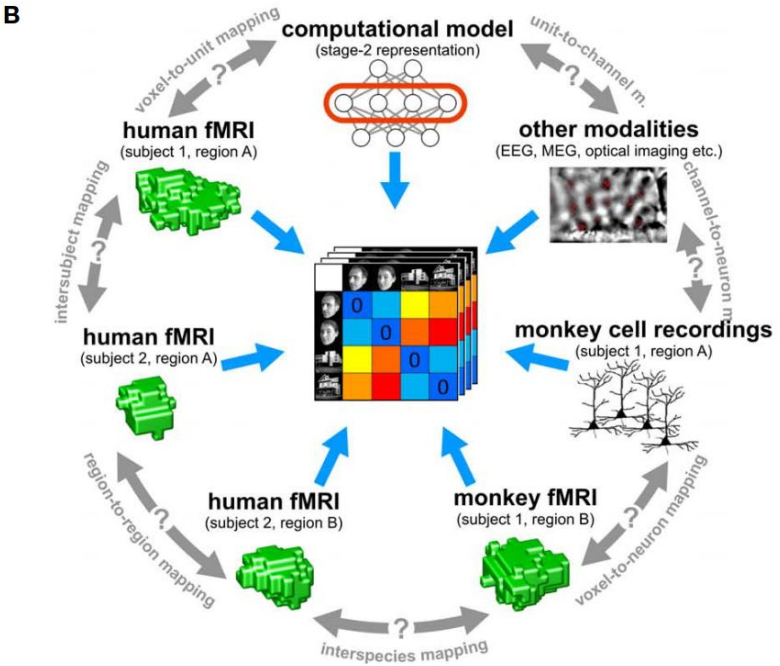
WhiteBoard Example of Data Collection

Data Collection in Machines



Representational Similarity Analysis (RSA)

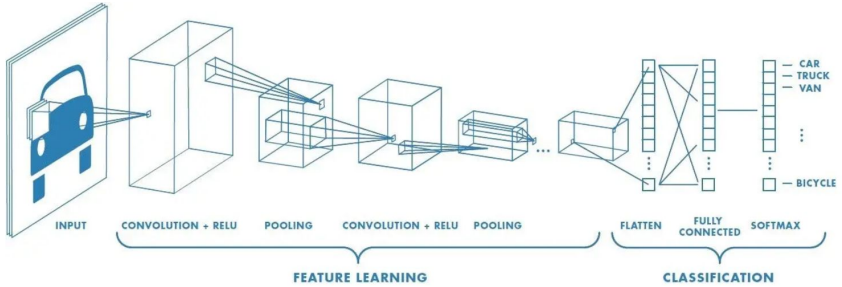




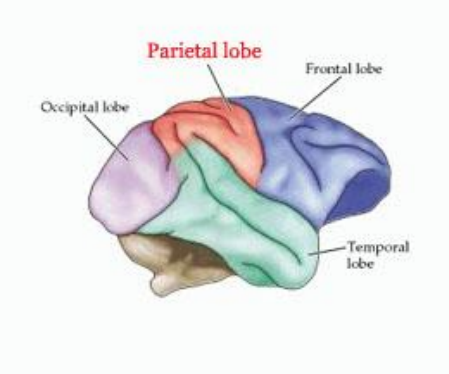
Representational similarity analysis – connecting the branches of systems neuroscience

Nikolaus Kriegeskorte[1,], Marieke Mur[1,2] and Peter Bandettini[1]*

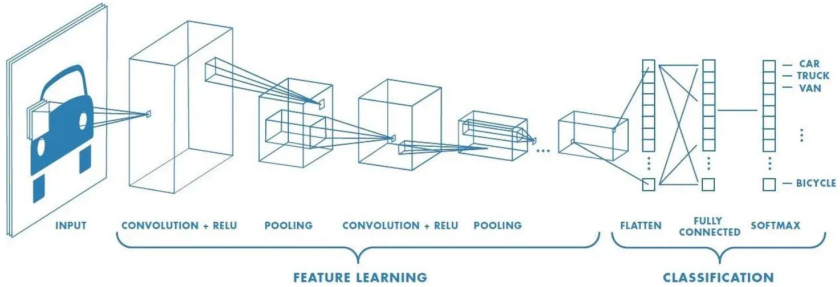
RSA allows us to compare both these systems:



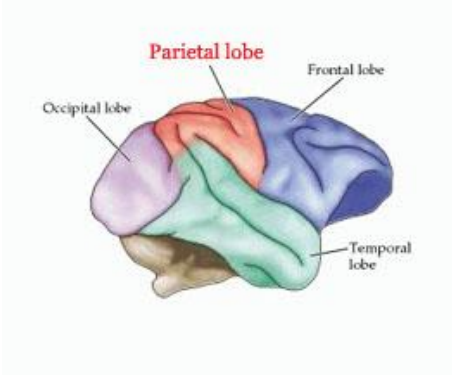
vs



RSA allows us to compare both these systems:



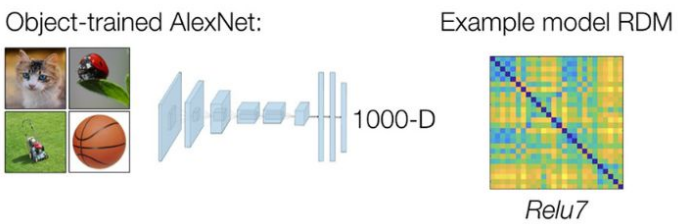
vs

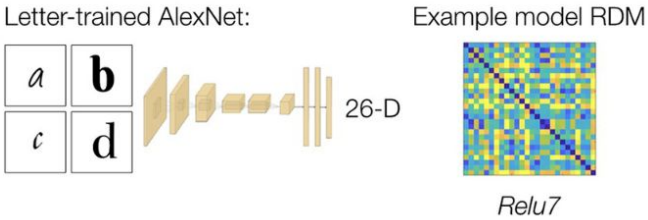


Examples of uses of RSA

PLOS COMPUTATIONAL BIOLOGY

A. Neural Network Model Feature Spaces





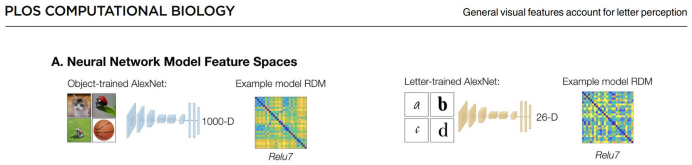
In this paper authors explored the question of how different neural networks trained on different datasets can interpret similarity of visual letter stimuli

General object-based features account for letter perception

Daniel Janini[1]*, Chris Hamblin[1], Arturo Deza[1,2], Talia Konkle[1]

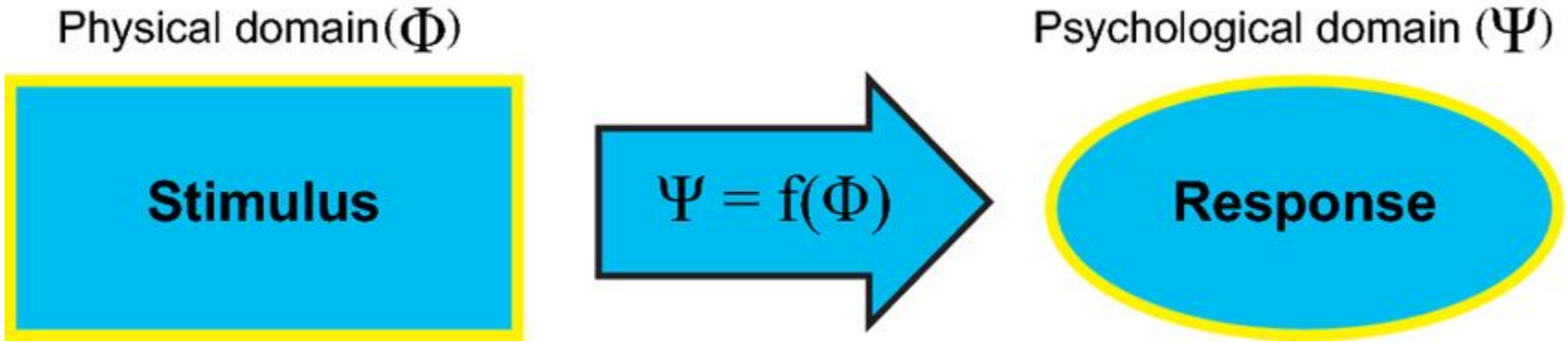
1 Department of Psychology, Harvard University, Cambridge, Massachusetts, United States of America,
2 Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Examples of uses of RSA (Walk-Through)

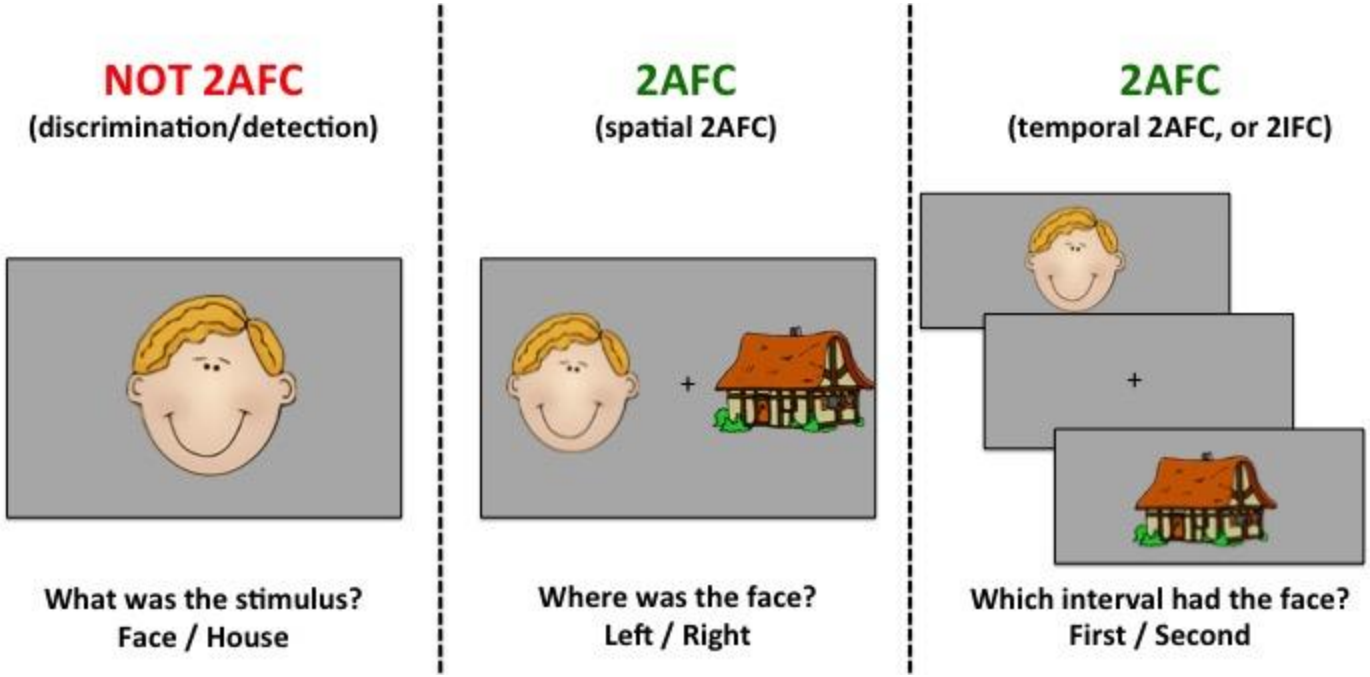


In this paper authors explored the question of
how different neural networks trained on different datasets can interpret similarity of visual letter stimuli

Psychophysics (Behaviour)



(The word psychophysics literally comes from a fusion of using methods of physics to find an underlying mechanism, and applying it not to physical laws, but to behaviour and perception)

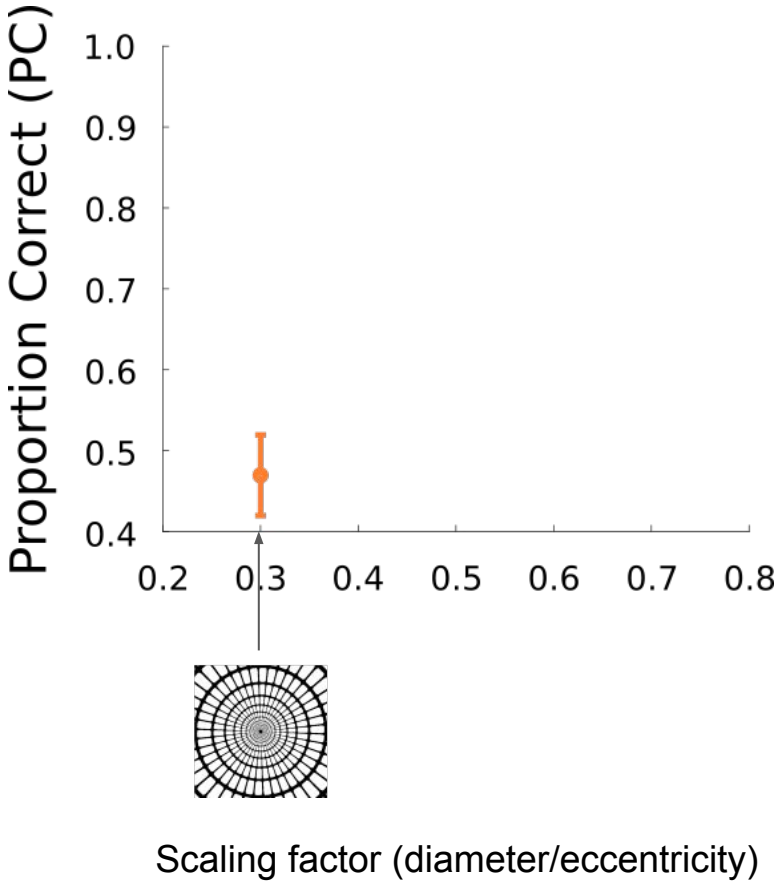


2AFC stands for: Two Alternative Forced Choice

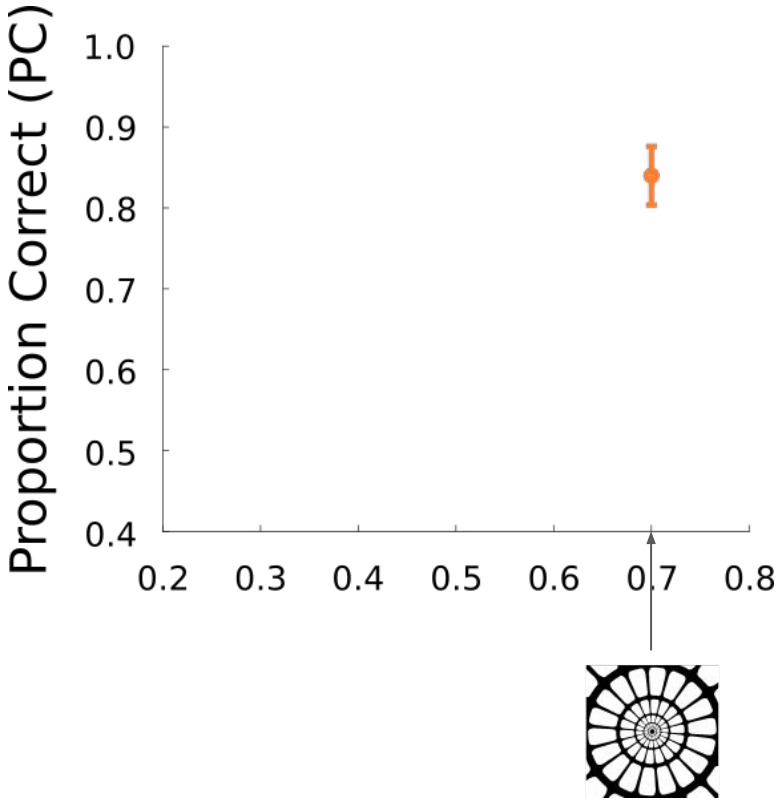
Psychophysics Example 1

Fitting a distortion variable to human limits of perception

How do we test this?

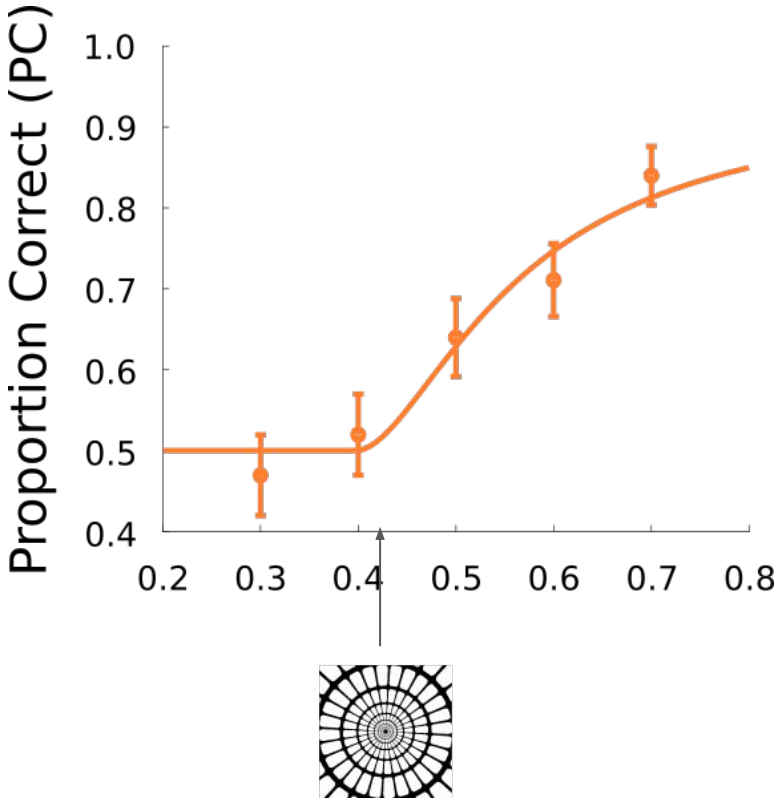


How do we test this?



Scaling factor (diameter/eccentricity)

How do we test this?



Scaling factor (diameter/eccentricity)

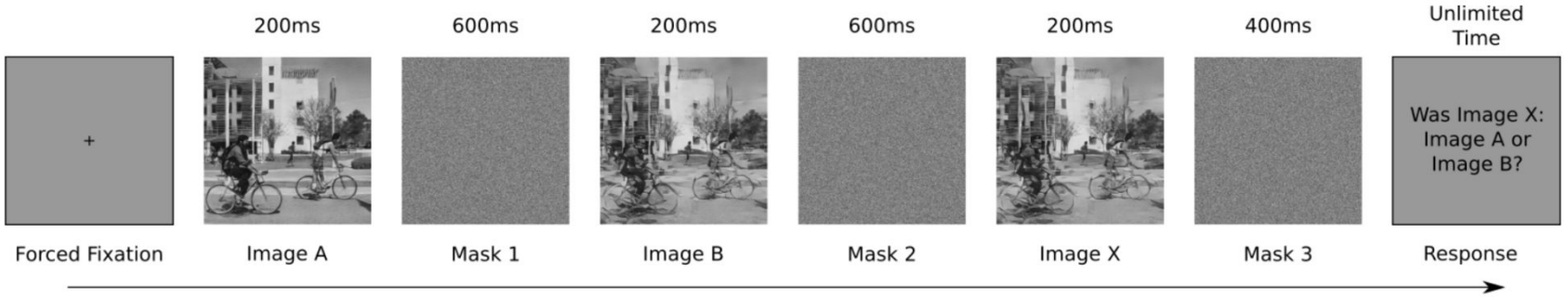
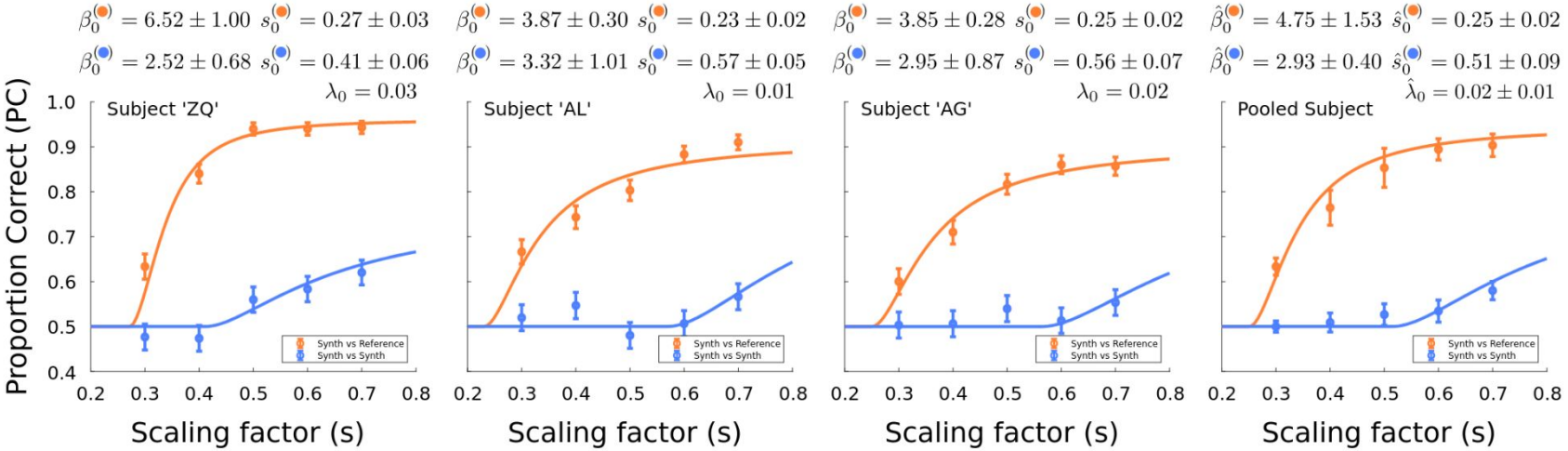


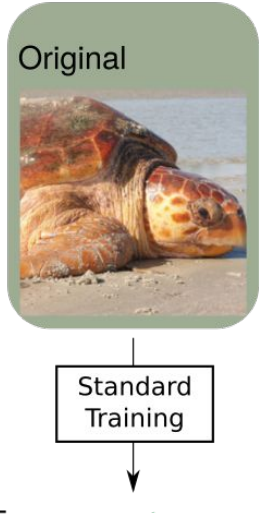
Figure 11: Experiment 2 shows the ABX metamer discrimination task done by the observers. Humans must fixate at the center of the image (no eye-movements) throughout the trial for it to be valid.



Psychophysics Example 2

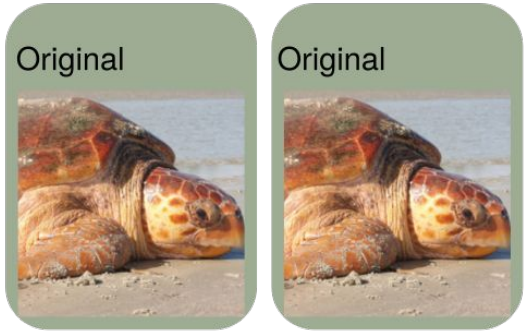
Testing Limits of Perception in Humans given renderings of images in Machines

How are Convolutional Neural Networks trained?



CNN		Standard Accuracy:
		Adversarially Robust:
Human		Biologically-Plausible: (Set of Transformations)

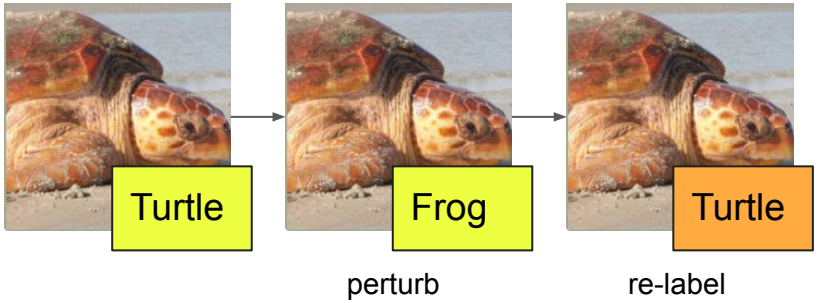
How are **robust** Convolutional Neural Networks trained?

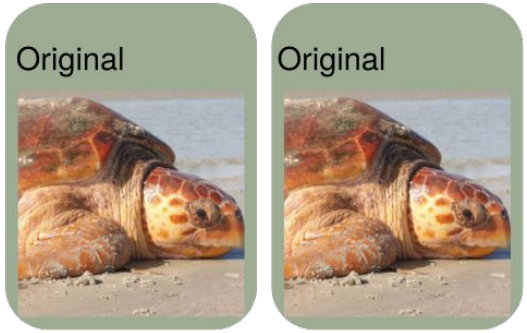


		Standard Training	Adversarial Training
CNN	Standard Accuracy:	✓	✓
	Adversarially Robust:	✗	✓
Human	Biologically-Plausible: (Set of Transformations)	-	✗

Madry et al. ICLR 2018.

How are **robust** Convolutional Neural Networks trained?

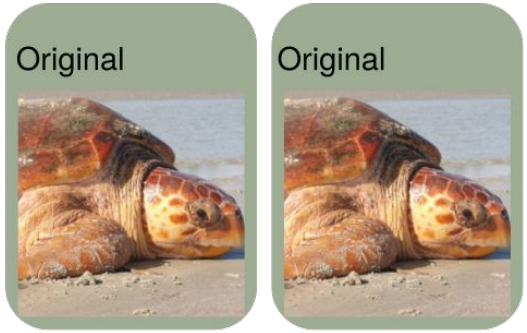




CNN	Standard Accuracy:	✓ ✓
	Adversarially Robust:	✗ ✓
Human	Biologically-Plausible: (Set of Transformations)	– ✗

Madry et al. ICLR 2018.

How are **robust** Convolutional Neural Networks trained?



		Standard Training	Adversarial Training
CNN	Standard Accuracy:	✓	✓
	Adversarially Robust:	✗	✓
Human	Biologically-Plausible: (Set of Transformations)	-	✗

Madry et al. ICLR 2018.

Training image

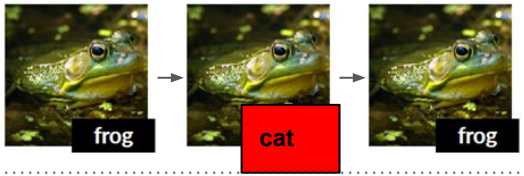
$\mathcal{D}$

"Adversarial Examples are not Bugs, they are Features"

Ilyas et al., *NeurIPS, 2019.*

Adversarial Training as in Tsipras et al. *ICLR, 2019.*

[Example: Adversarially Perturb Frog to Cat, Re-Train as Frog]

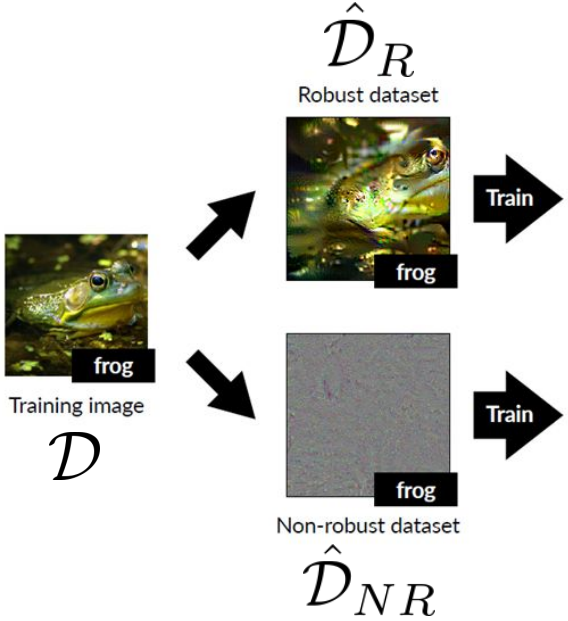


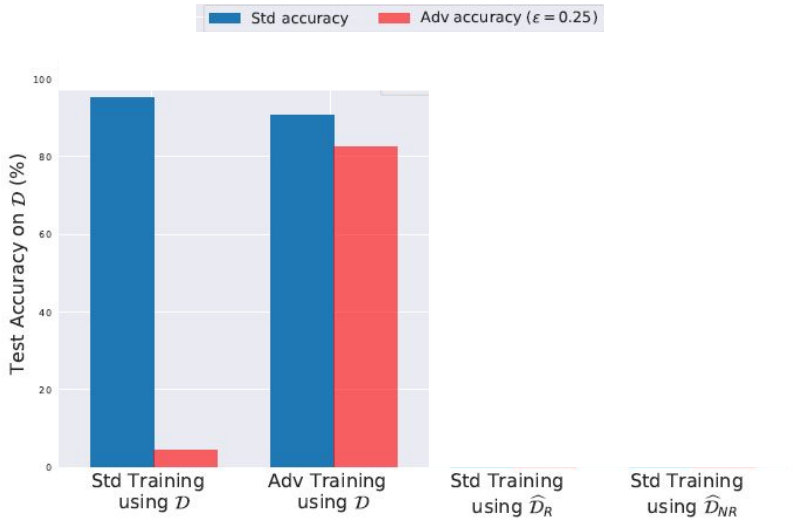
Training image

$\mathcal{D}$

"Adversarial Examples are not Bugs, they are Features"

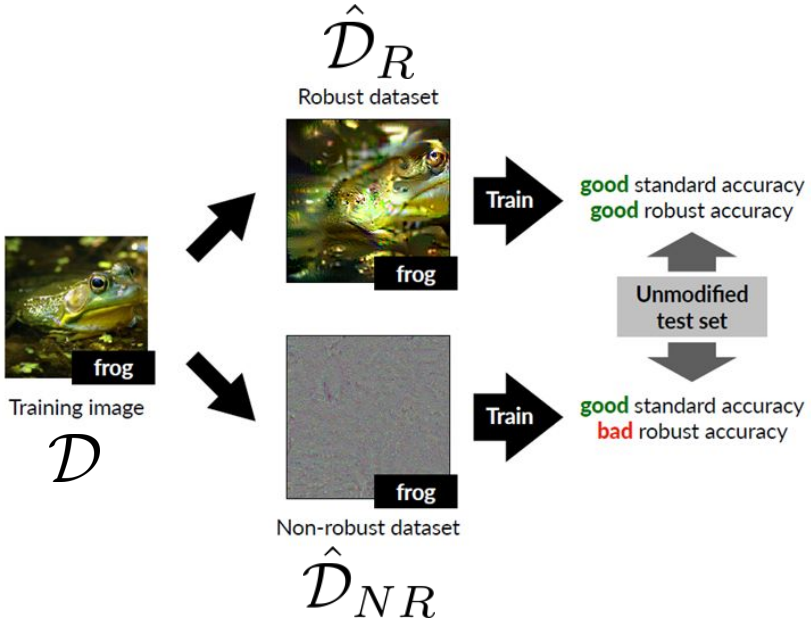
Ilyas et al., *NeurIPS, 2019.*





"Adversarial Examples are not Bugs, they are Features"

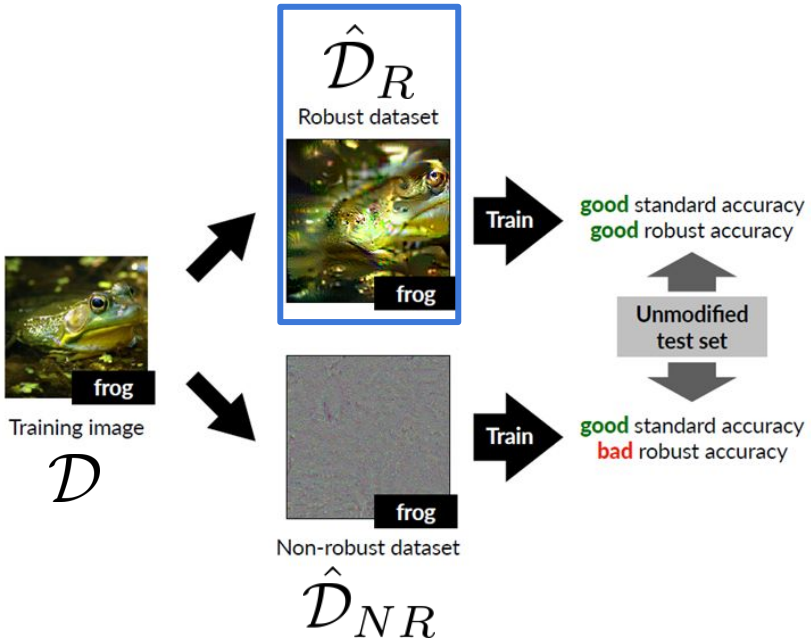
Ilyas et al., *NeurIPS, 2019.*

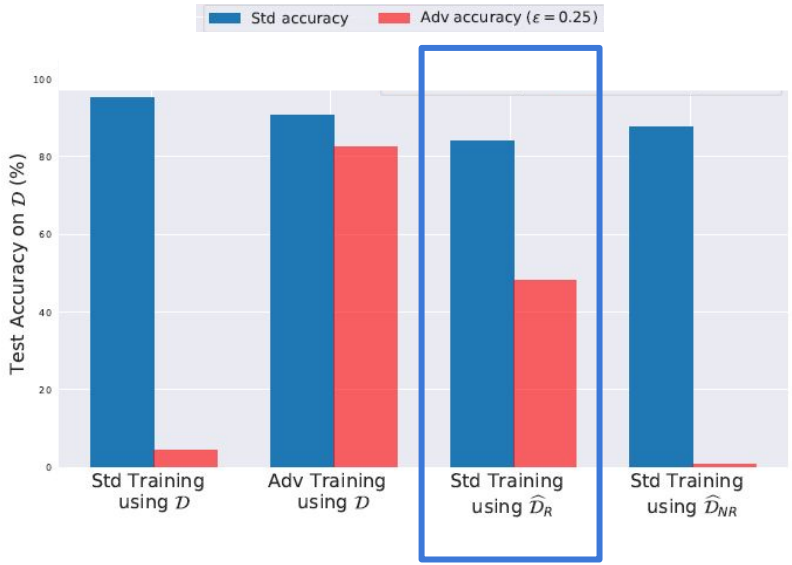


"Adversarial Examples are not Bugs, they are Features"

Ilyas et al., *NeurIPS, 2019.*

But is this transformation biologically plausible?
[Computer Vision argues it is not!]

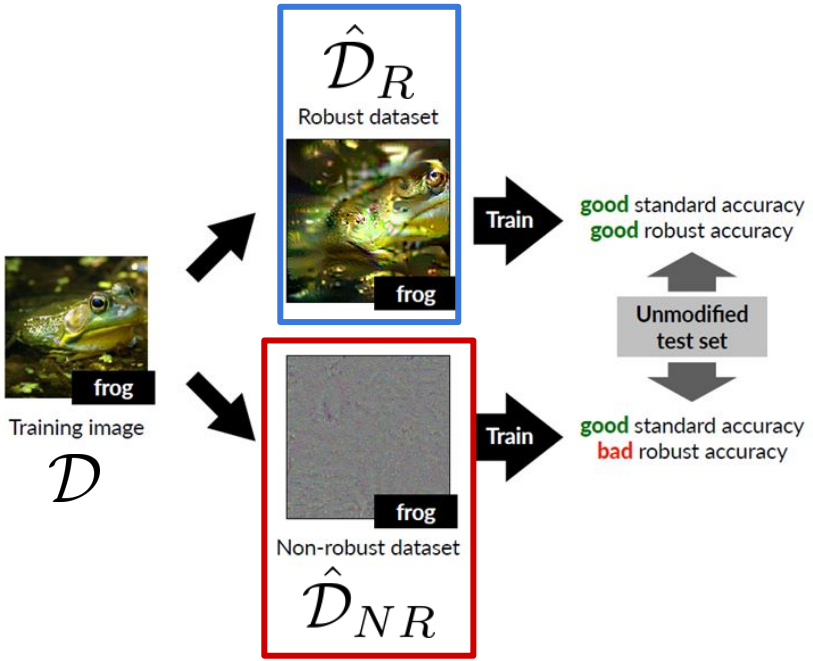


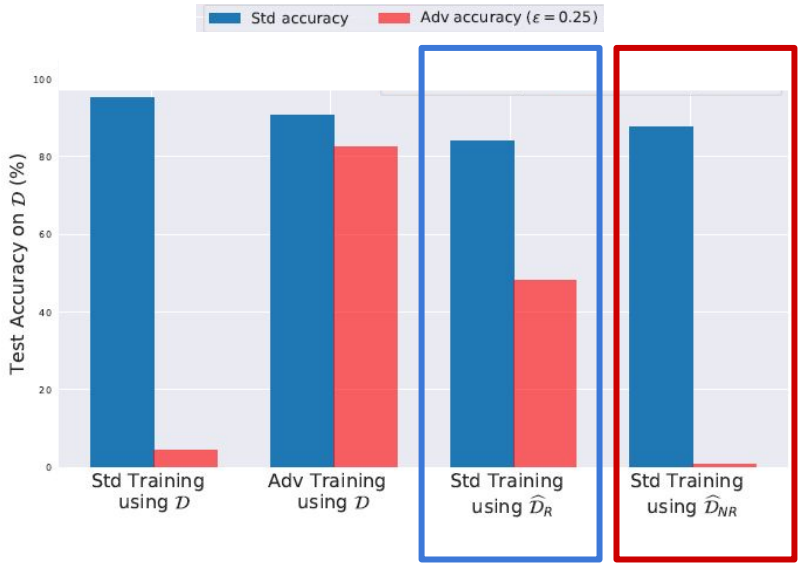


"Adversarial Examples are not Bugs, they are Features"

Ilyas et al., *NeurIPS, 2019.*

But is this transformation biologically plausible?
[Computer Vision argues it is not!]



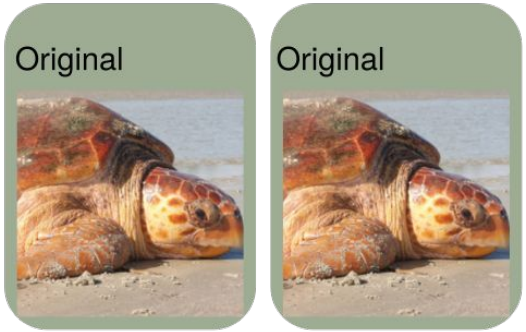


"Adversarial Examples are not Bugs, they are Features"

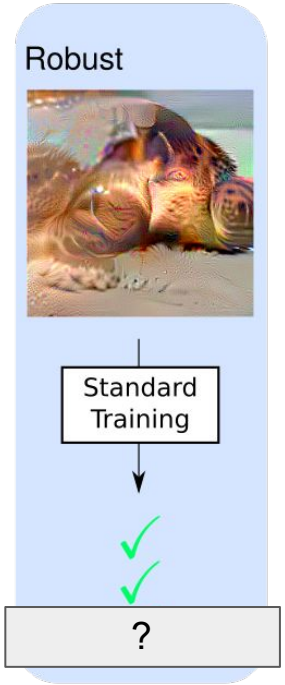
Ilyas et al., *NeurIPS, 2019.*

How are **robust** Convolutional Neural Networks trained?

CNN	 Standard Accuracy:	✓
	Adversarially Robust:	✗
Human	Biologically-Plausible: (Set of Transformations)	-



Stimuli Synthesized from:



Harrington & Deza. *(Accepted at ICLR 2022 as a Spotlight). [Top 1%]*

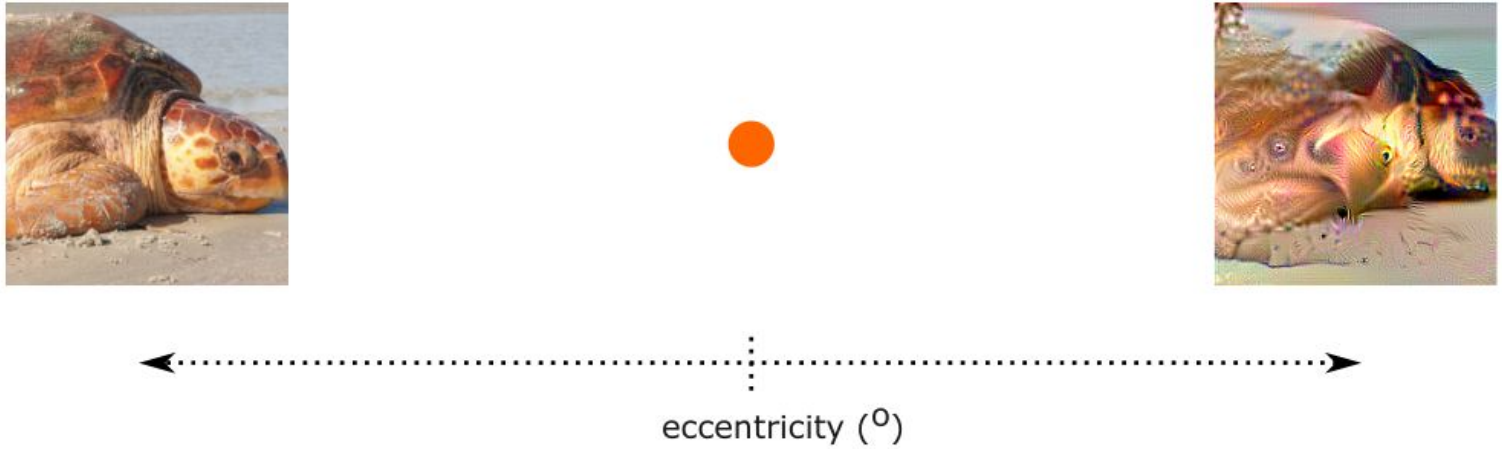
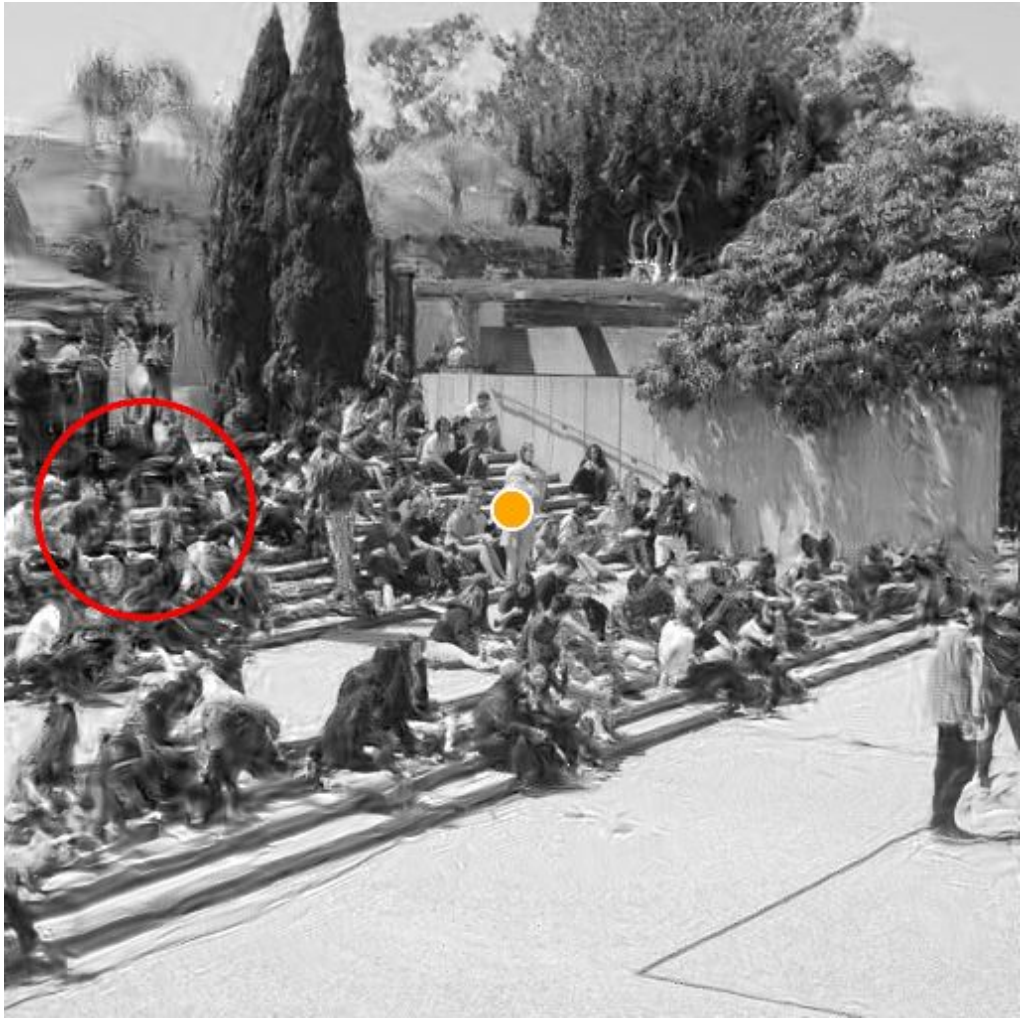


Figure 1: A sample un-perturbed (left) and synthesized (right) image are shown peripherally. When a human observer fixates at the orange dot (center), both images – now placed away from the fovea – are perceptually indistinguishable to each other (i.e. *metameric*). In this paper, we psychophysically test this phenomena over a variety of images as we manipulate retinal eccentricity, showing biological plausibility of adversarially robust features via peripheral computation on 12 human subjects.

Harrington & Deza. *(Accepted at ICLR 2022 as a Spotlight). [Top 1%]*

Rosenholtz (2011)

Freeman & Simoncelli (2011)



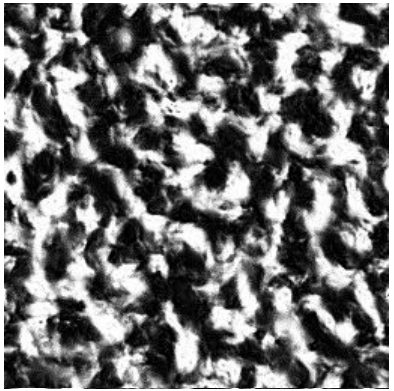
Rosenholtz (2011)

Freeman & Simoncelli (2011)

Balas, Nakano & Rosenholtz. *Journal of Vision, 2009.*
Portilla & Simoncelli. *International Journal of Computer Vision, 2000.*

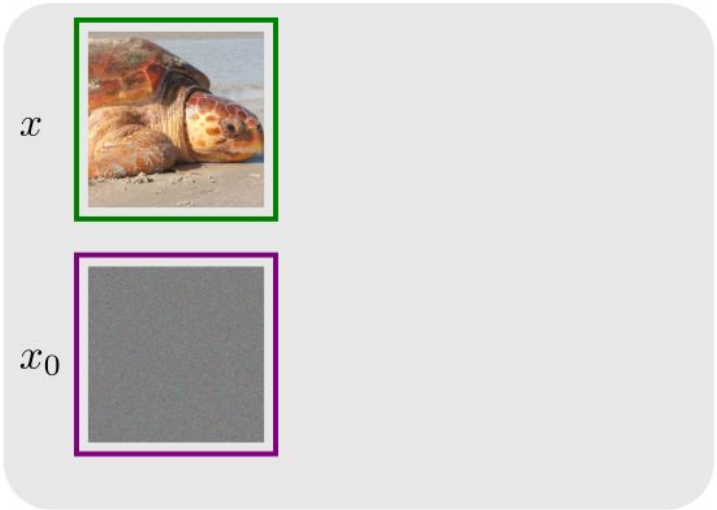
Texture Synthesis used to model Peripheral Computation + Visual Crowding

$$\xrightarrow{f(\cdot)} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} \hat{x_1} \\ \hat{x_2} \\ \vdots \\ \hat{x_N} \end{bmatrix} \xleftarrow{f(\cdot)}$$



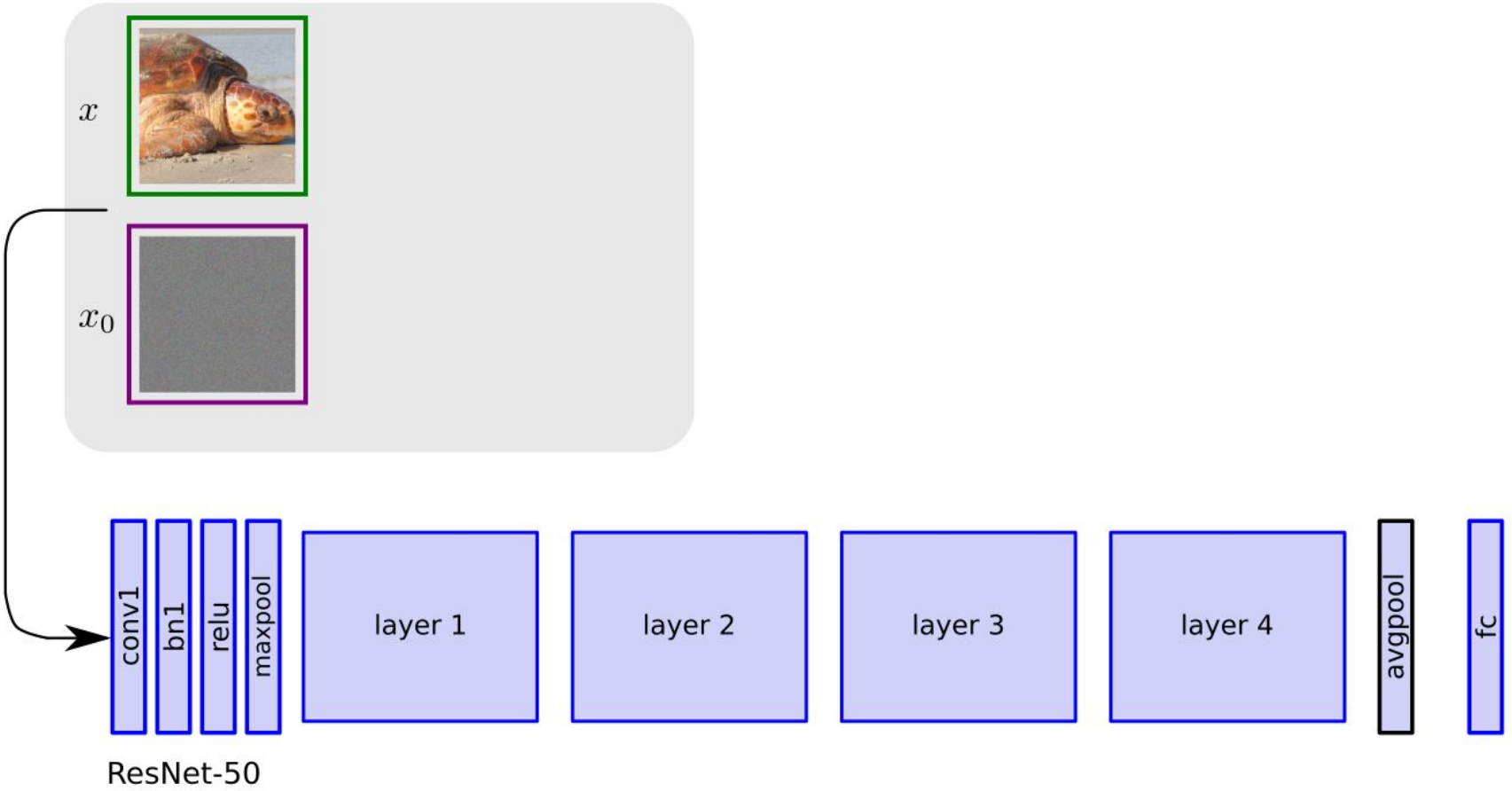
Balas, Nakano & Rosenholtz. *Journal of Vision, 2009.*
Portilla & Simoncelli. *International Journal of Computer Vision, 2000.*

We can do Analysis-by-Synthesis with a DeepNet as well!



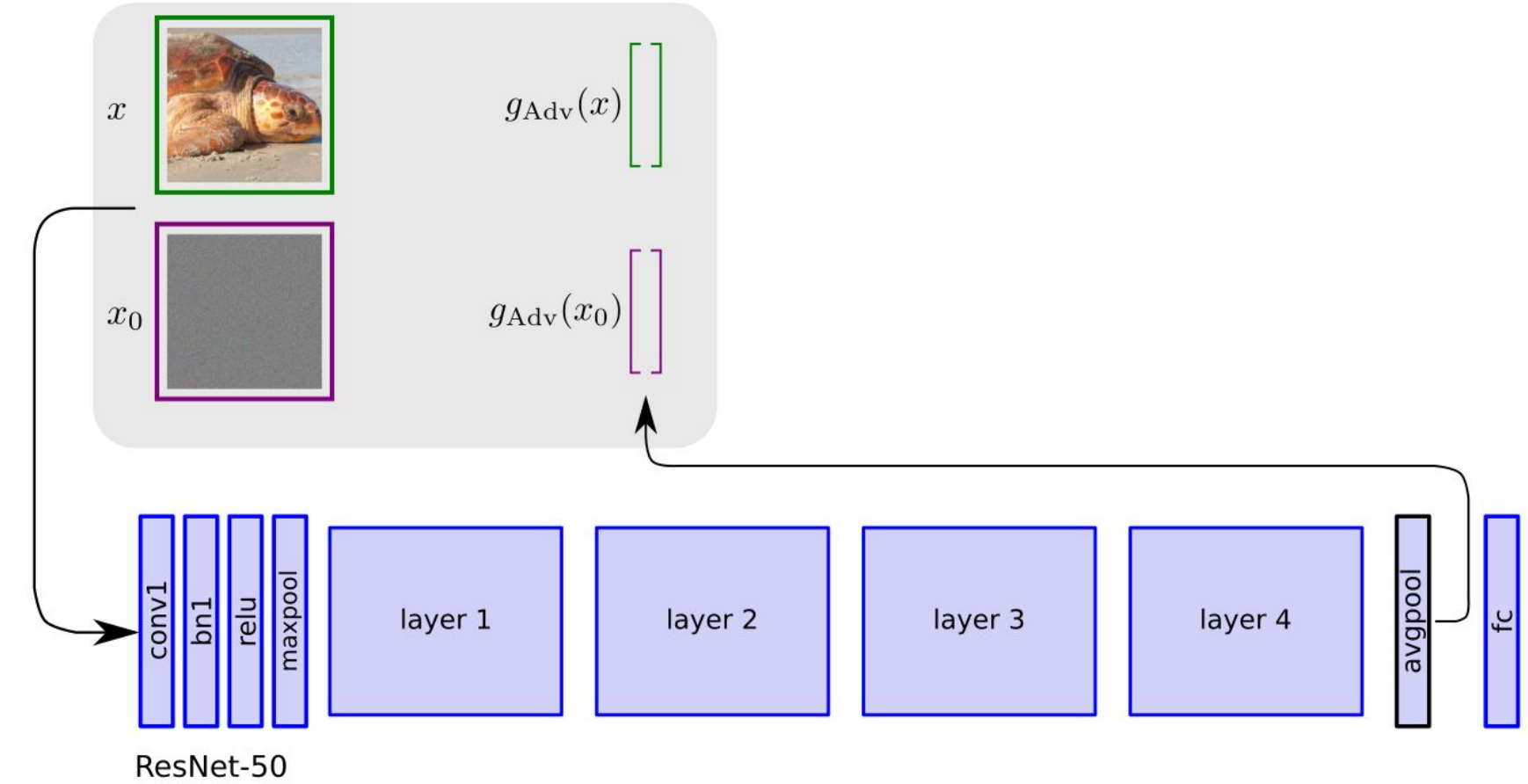
Santurkar et al. & Engstrom et al., *NeurIPS, 2019.*

We can do Analysis-by-Synthesis with a DeepNet as well!



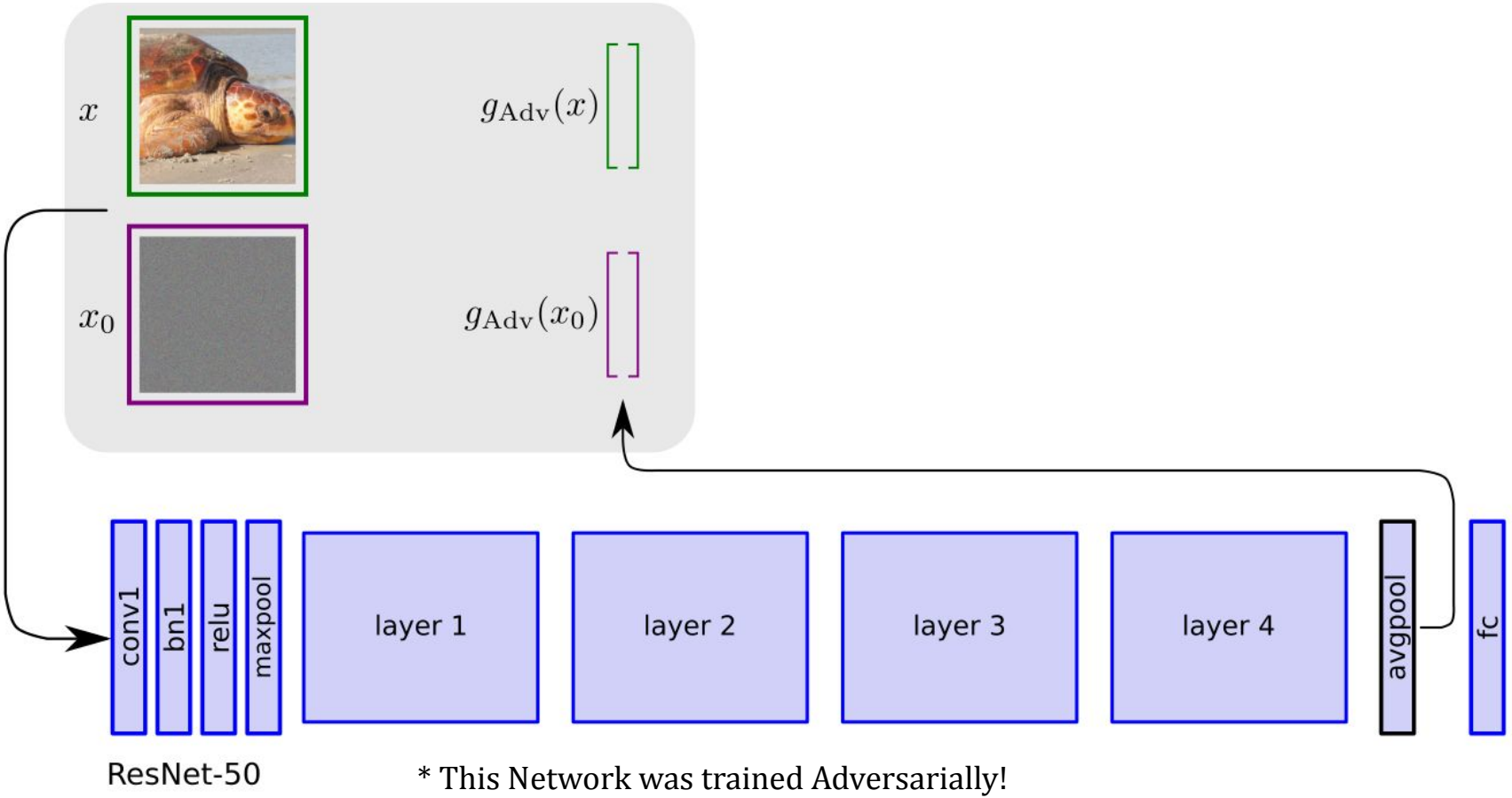
Santurkar et al. & Engstrom et al., *NeurIPS, 2019.*

We can do Analysis-by-Synthesis with a DeepNet as well!



Santurkar et al. & Engstrom et al., *NeurIPS, 2019.*

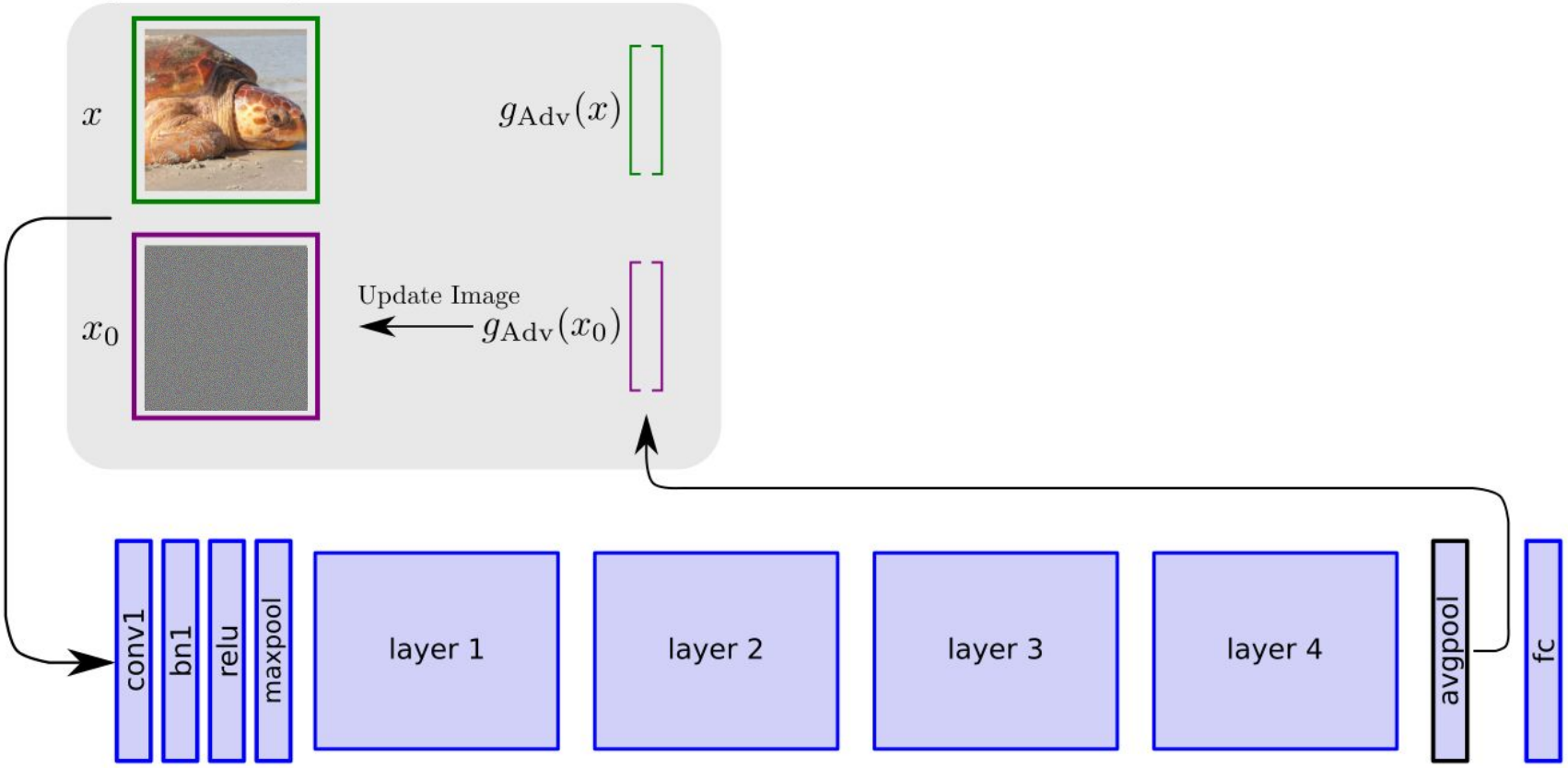
We can do Analysis-by-Synthesis with a DeepNet as well!



Santurkar et al. & Engstrom et al., *NeurIPS, 2019.*

We can do Analysis-by-Synthesis with a DeepNet as well!

$- g_{\text{Adv}}(x_0)||_2$



* This Network was trained Adversarially!

Santurkar et al. & Engstrom et al., *NeurIPS, 2019.*

Adversarially Trained Networks

~

Human Peripheral Computation

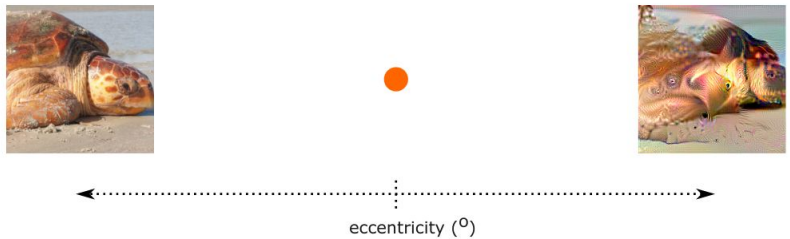
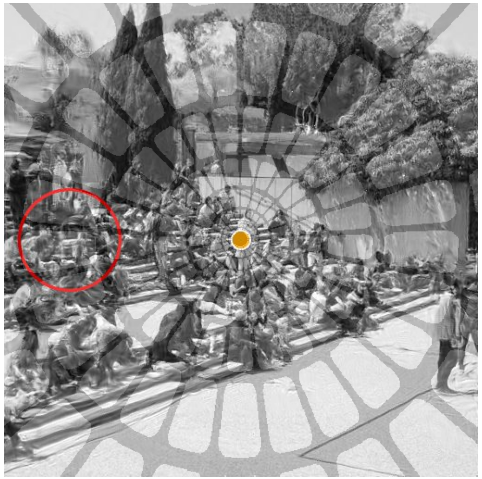
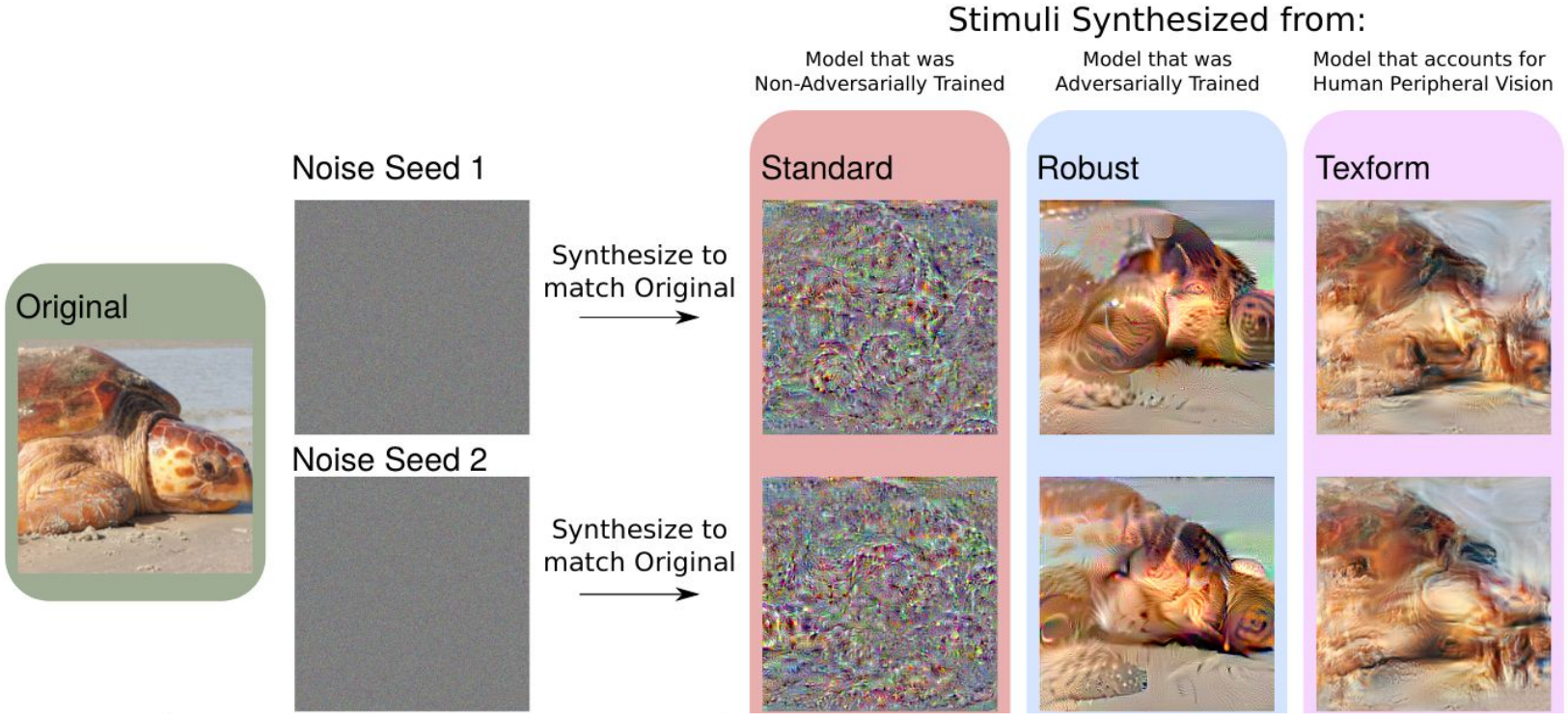


Figure 1: A sample un-perturbed (left) and synthesized (right) image are shown peripherally. When a human observer fixates at the orange dot (center), both images – now placed away from the fovea – are perceptually indistinguishable to each other (i.e. *metameric*). In this paper, we psychophysically test this phenomena over a variety of images as we manipulate retinal eccentricity, showing biological plausibility of adversarially robust features via peripheral computation on 12 human subjects.



"Finding Biological Plausibility for Adversarially Robust Features via Metameric Tasks"

Harrington & Deza. ICLR 2022.


Stimuli Synthesized from:
Model that was Non-Adversarially Trained
Model that was Adversarially Trained
Model that accounts for Human Peripheral Vision
Original
Noise Seed 1
Noise Seed 2
Synthesize to match Original
Synthesize to match Original
Standard
Robust
Texform

Human Psychophysics!

"Finding Biological Plausibility for Adversarially Robust Features via Metameric Tasks"

Harrington & Deza. ICLR 2022.

In Real Experiment: Stimuli is flashed for 100 ms

[Peripheral-to-Peripheral] Oddity task

Harrington & Deza. ICLR 2022 (Accepted as Spotlight)

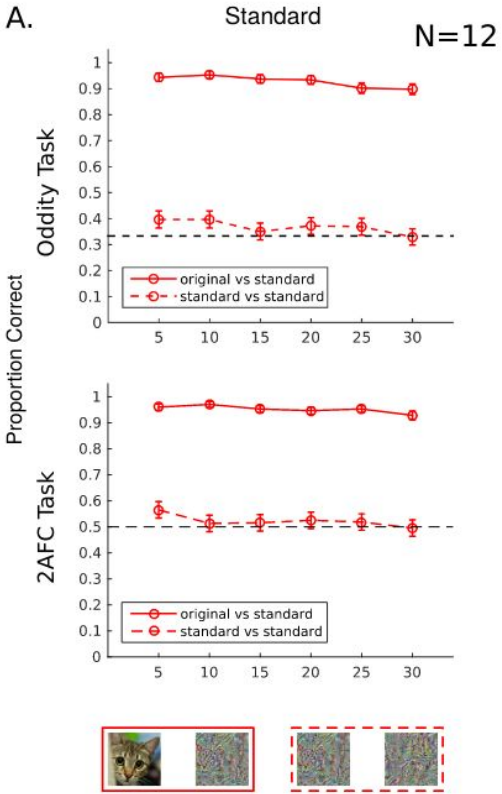
In Real Experiment: Stimuli is flashed for 100 ms

[Foveal-to-Peripheral] 2AFC Matching task

+

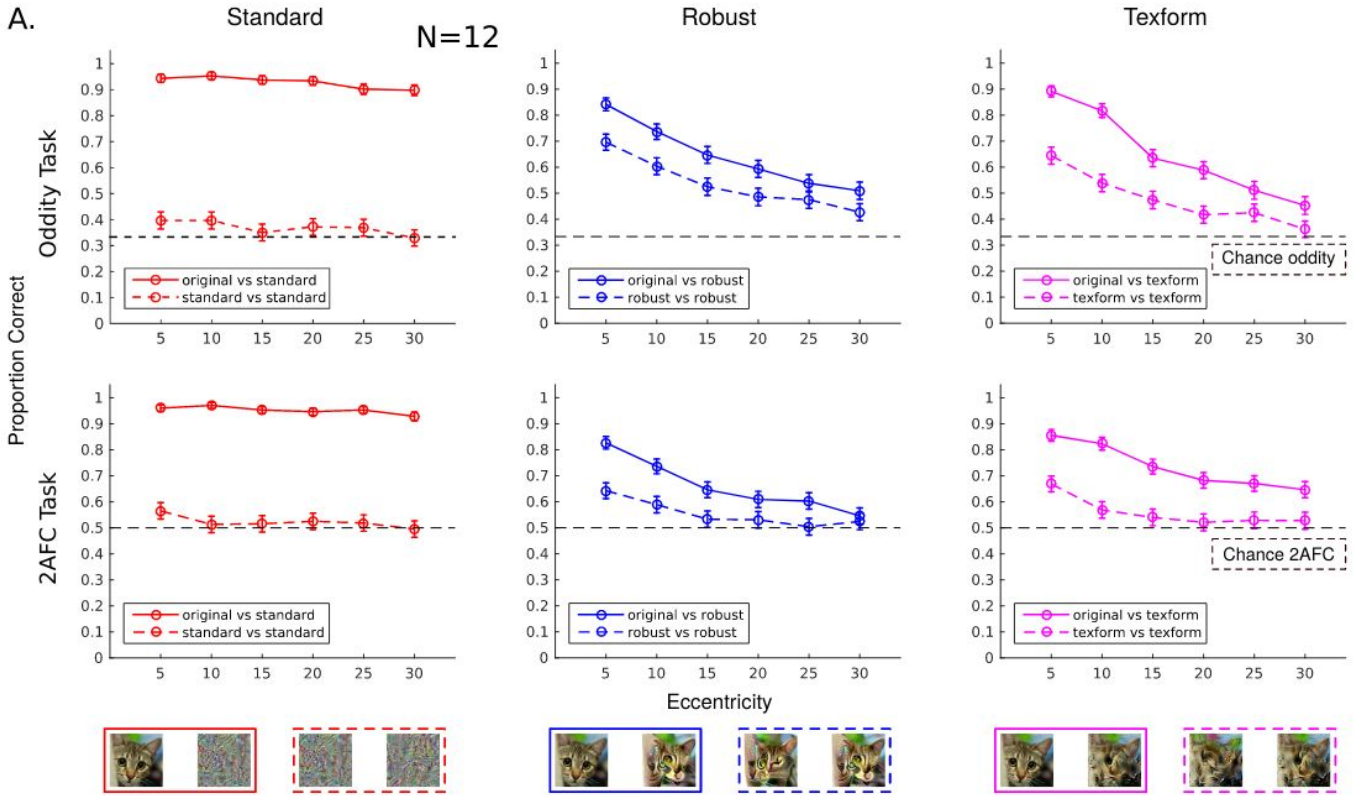
Harrington & Deza. ICLR 2022 (Accepted as Spotlight)

Results



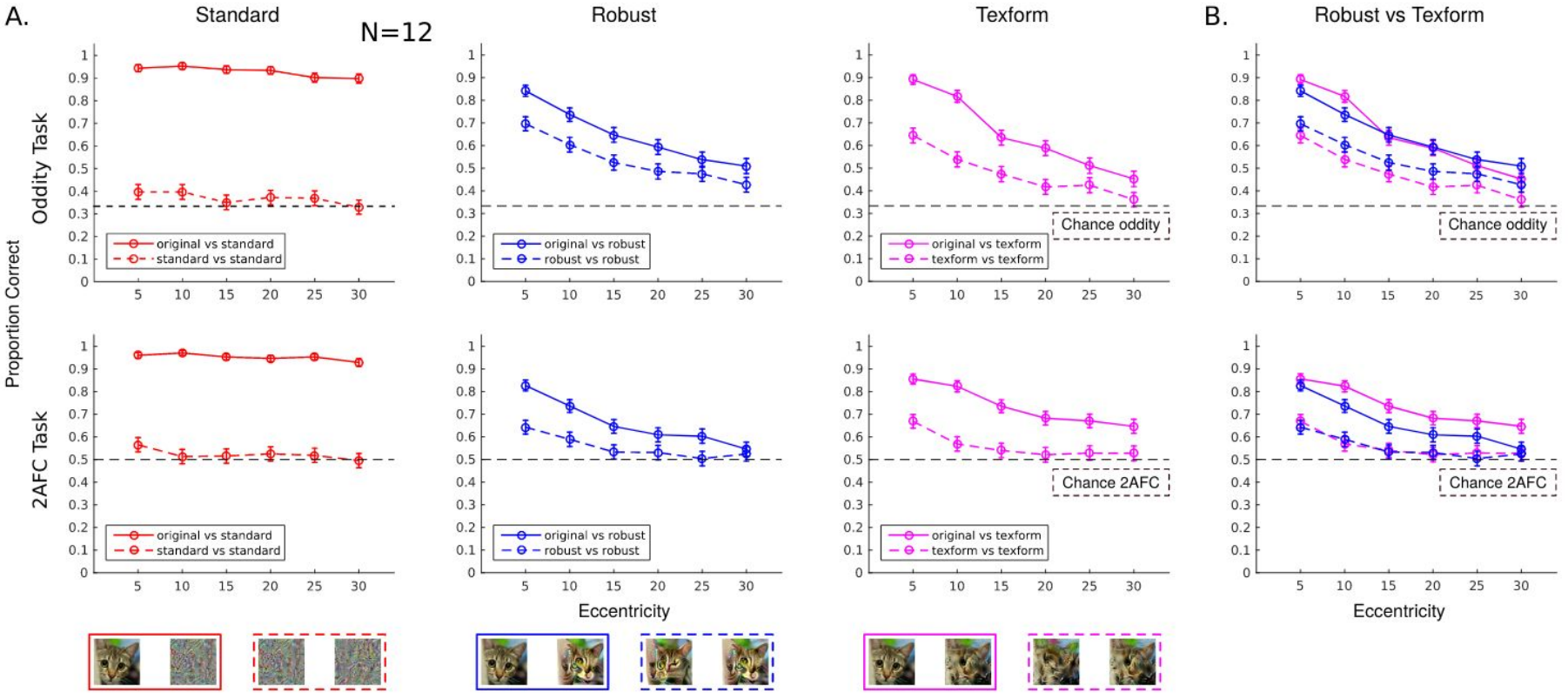
Harrington & Deza. ICLR 2022 (Accepted as Spotlight)

Results



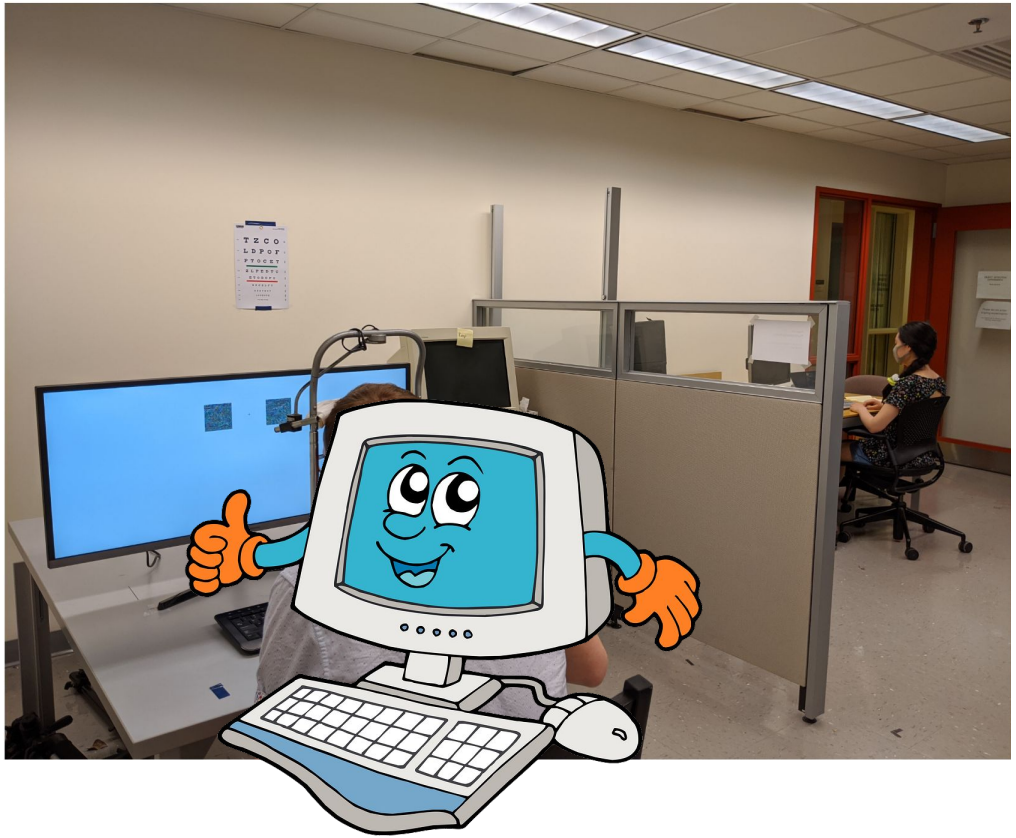
Harrington & Deza. ICLR 2022 (Accepted as Spotlight)

Results



Harrington & Deza. ICLR 2022 (Accepted as Spotlight)

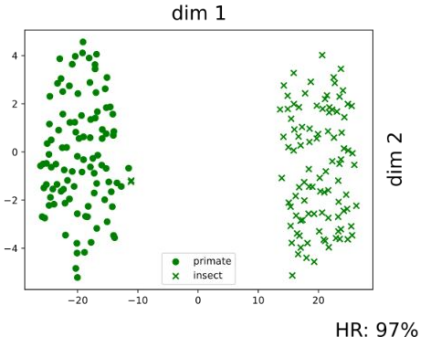
“Machine” Psychophysics!



Testing Stimuli

ResNet50 trained on Original Images

Standard Training



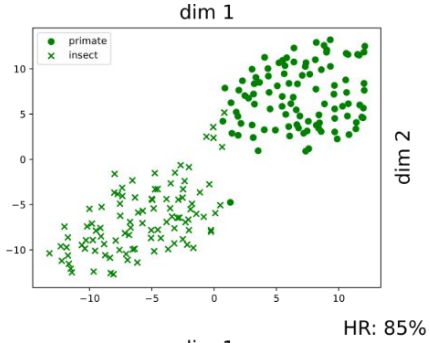
HR: 97%

Original Stimuli

Testing Stimuli

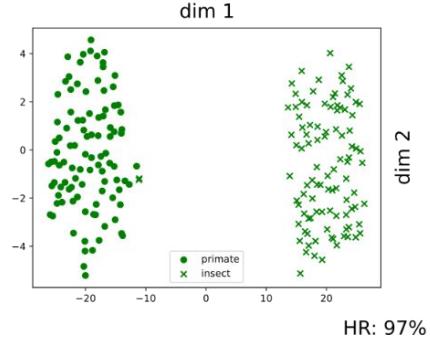
ResNet50 trained on Original Images

Adversarial Training



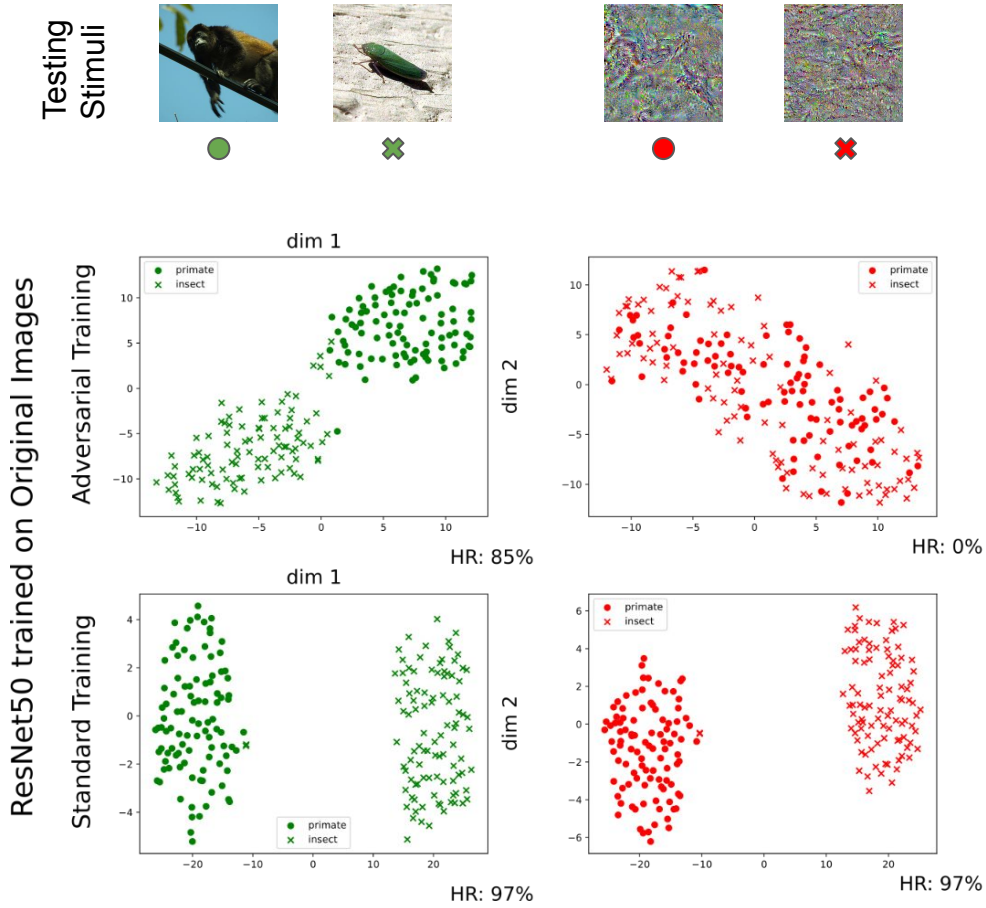
HR: 85%

Standard Training



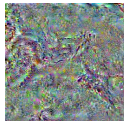
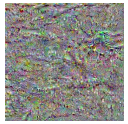
HR: 97%

Original Stimuli


Testing Stimuli
ResNet50 trained on Original Images
Adversarial Training
Standard Training
dim 1
dim 2
primate
insect
HR: 85%
HR: 0%
HR: 97%
HR: 97%
Original Stimuli
Stimuli derived from a Standard Trained Network

Testing Stimuli

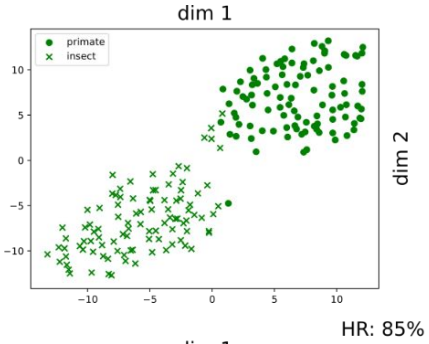


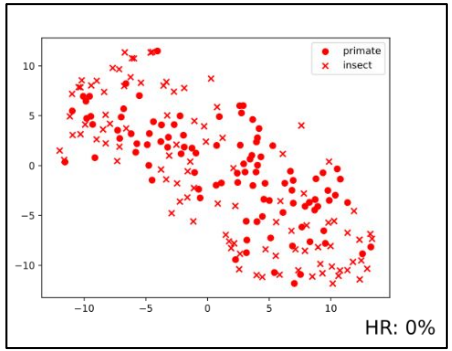



A Human would do poorly as well!

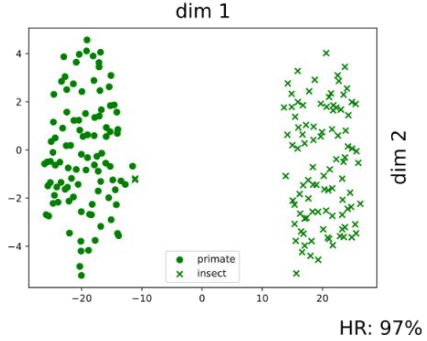
ResNet50 trained on Original Images

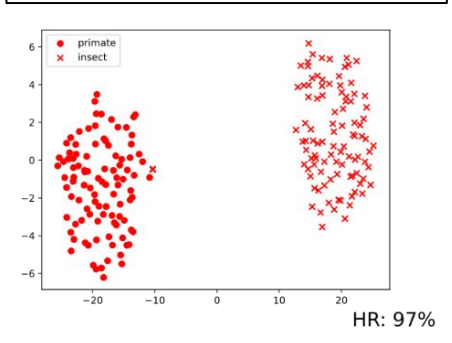
Adversarial Training





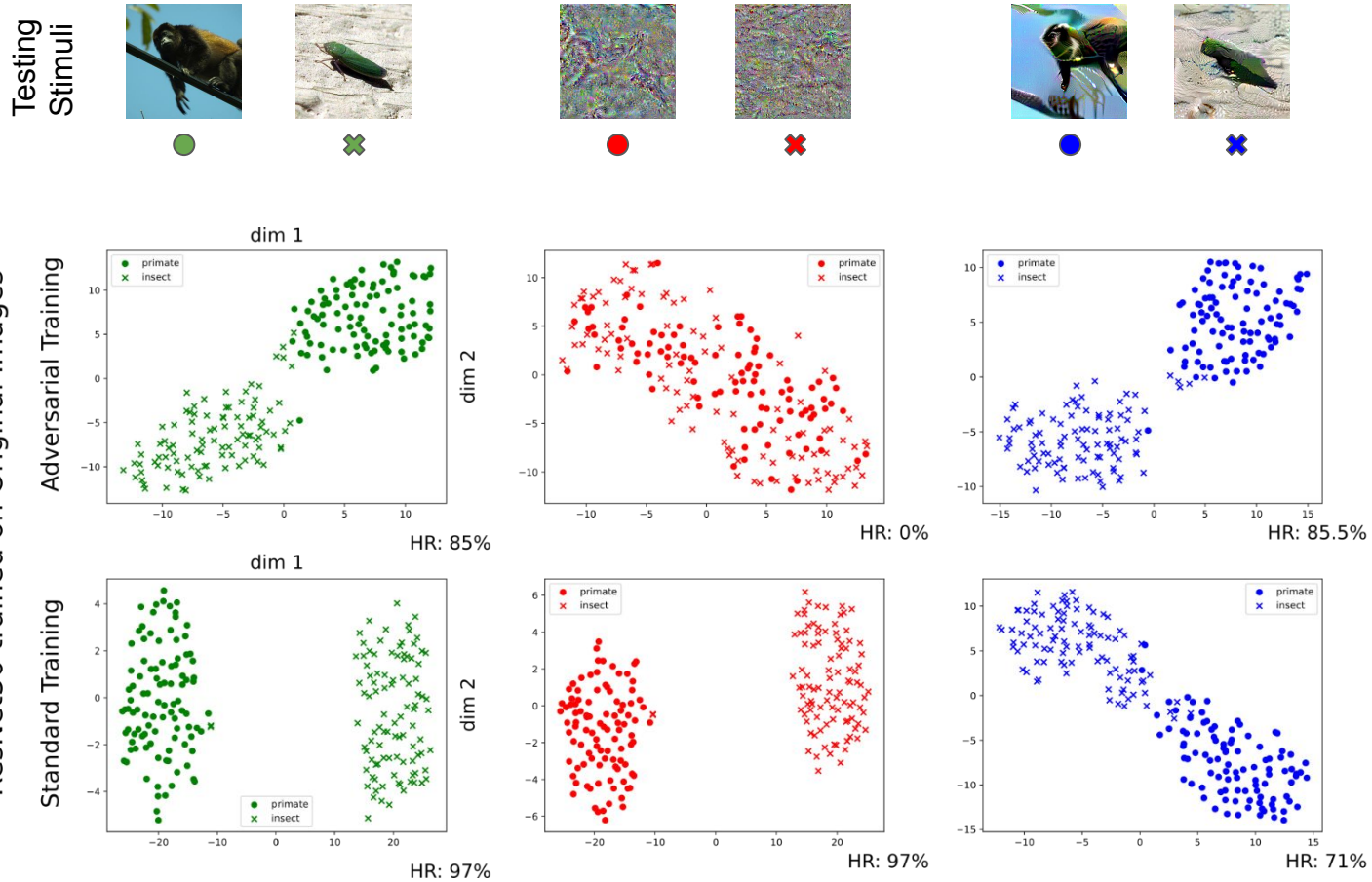
Standard Training

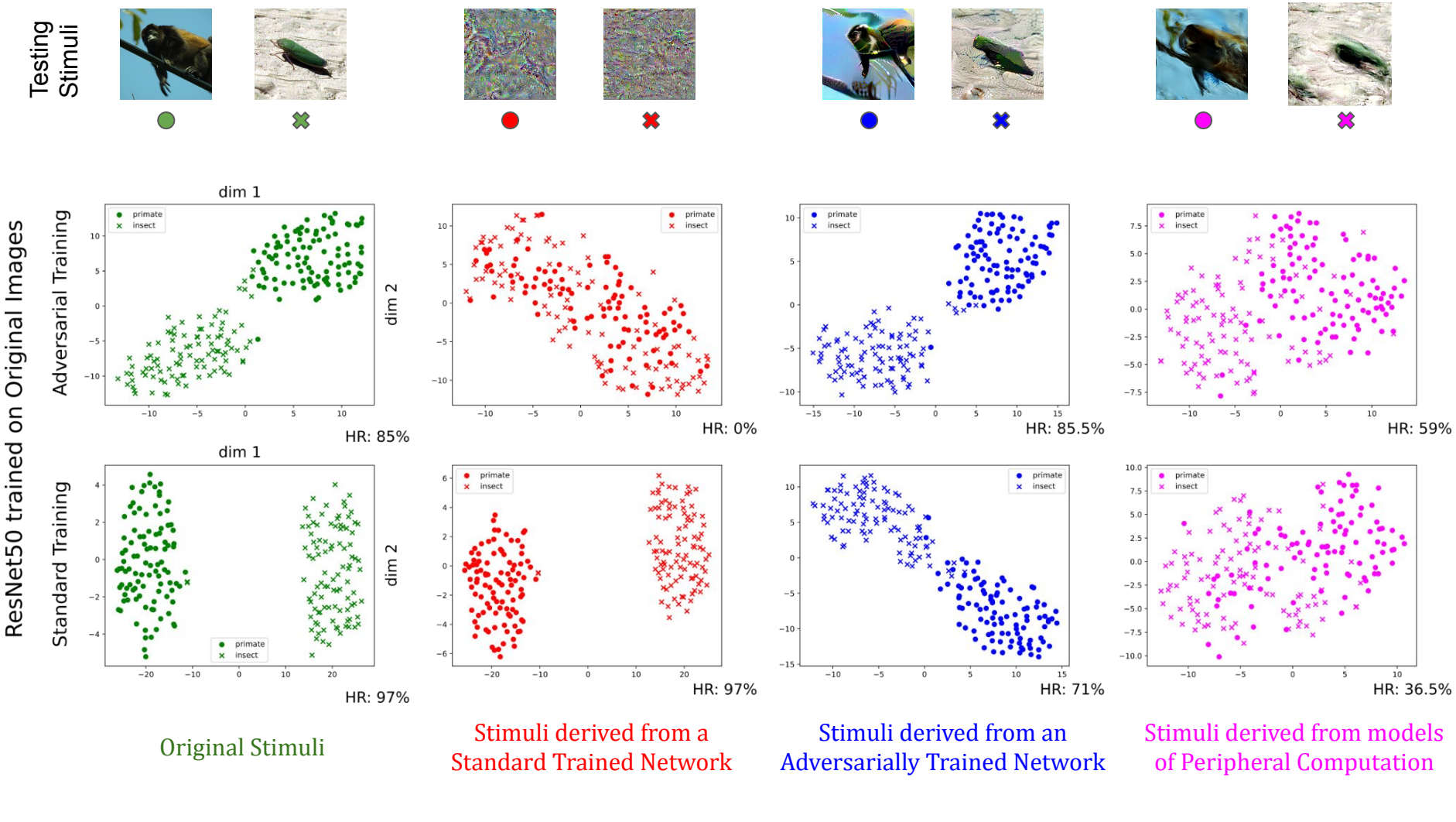


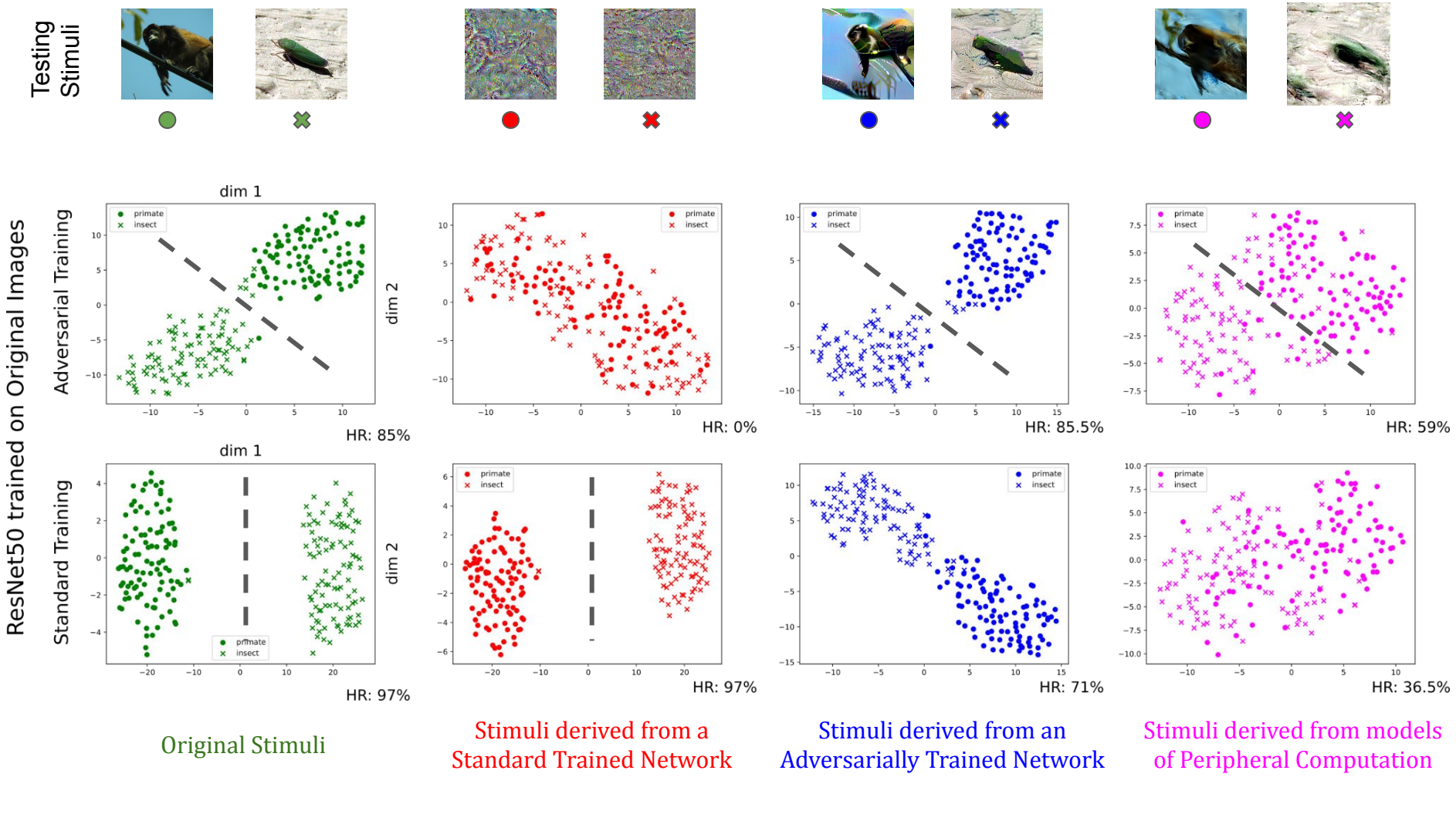


Original Stimuli

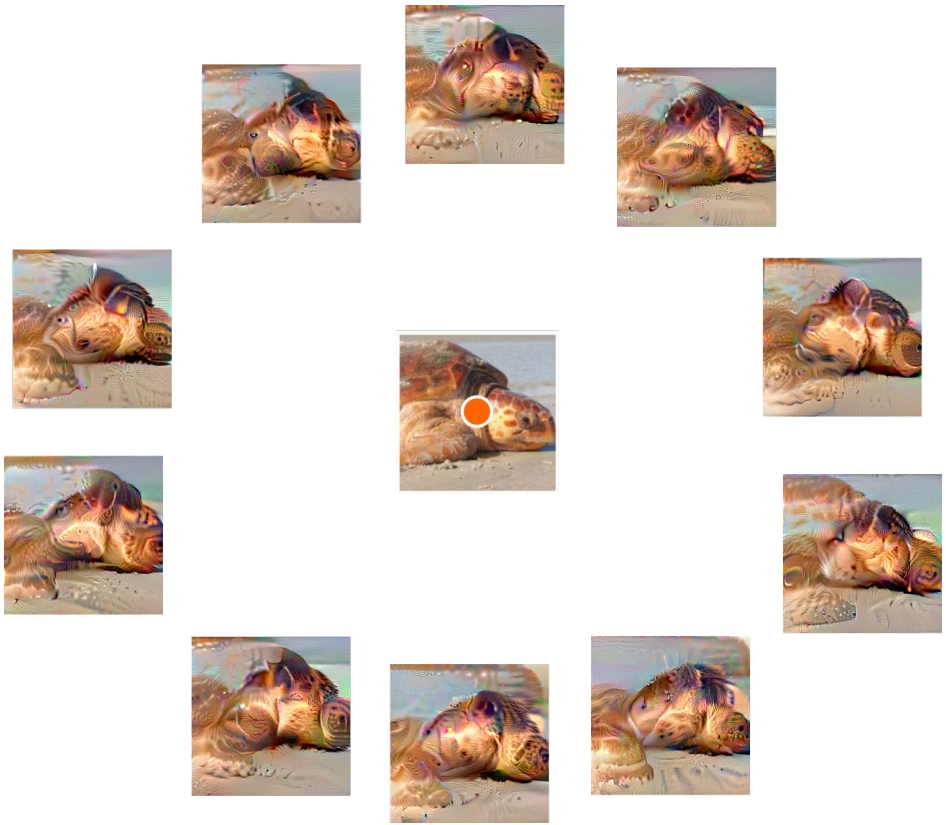
Stimuli derived from a Standard Trained Network


Testing Stimuli
ResNet50 trained on Original Images
Adversarial Training
Standard Training
dim 1
dim 2
primate
insect
HR: 85%
HR: 0%
HR: 85.5%
HR: 97%
HR: 97%
HR: 71%
Original Stimuli
Stimuli derived from a Standard Trained Network
Stimuli derived from an Adversarially Trained Network


Testing Stimuli
ResNet50 trained on Original Images
Adversarial Training
Standard Training
dim 1
dim 2
primate
insect
HR: 85%
HR: 0%
HR: 85.5%
HR: 59%
HR: 97%
HR: 97%
HR: 71%
HR: 36.5%
Original Stimuli
Stimuli derived from a Standard Trained Network
Stimuli derived from an Adversarially Trained Network
Stimuli derived from models of Peripheral Computation


Testing Stimuli
ResNet50 trained on Original Images
Adversarial Training
Standard Training
dim 1
dim 2
primate
insect
HR: 85%
HR: 0%
HR: 85.5%
HR: 59%
HR: 97%
HR: 97%
HR: 71%
HR: 36.5%
Original Stimuli
Stimuli derived from a Standard Trained Network
Stimuli derived from an Adversarially Trained Network
Stimuli derived from models of Peripheral Computation

Peripheral Computation could have a **representational goal** in humans !



[...] as it seems like peripheral vision encodes a similar set of representations as an adversarially trained network

Application of RSA + Psychophysics: Are Transformers good models of the Ventral Stream?

(Open Question)

Are Transformers good models of the Ventral Stream?

William Berrios

← **Tweet**

Simon Kornblith @skornblith · Apr 1, 2021

Has anyone looked at how well the representations of Vision Transformers match brain data? @martin_schrimpf @qbilius @GeigerFranziska

5 5 38

Martin Schrimpf
@martin_schrimpf

Replying to @skornblith and @qbilius

The transformers we have tested do not match brain data well. We have ViT and DeiT on brain-score.org and they score much worse than other models (even though their ImageNet performance is better)

10:14 AM · Apr 1, 2021 · Twitter Web App

11 Retweets **4** Quote Tweets **78** Likes

Grace Lindsay @neurograce · Apr 1, 2021

I remember when the results were first coming out that trained convolutional neural networks are good predictors of activity in the visual system some people had the attitude of "that's not interesting because obviously anything that does vision well will look like the brain"

> **Martin Schrimpf** @martin_schrimpf · Apr 1, 2021
> Replying to @skornblith and @qbilius
> The transformers we have tested do not match brain data well. We have ViT and DeiT on brain-score.org and they score much worse than other models (even though their ImageNet performance is better)

8 13 82

Dileep George @dileeplearning · Apr 1, 2021

so if transformers actually turn out to be how the brain works, then what would your argument be?

Or are you saying that you know transformers are not how the visual cortex works? How do you know that?

Doesn't this point to a core problem?

2 4

Grace Lindsay
@neurograce

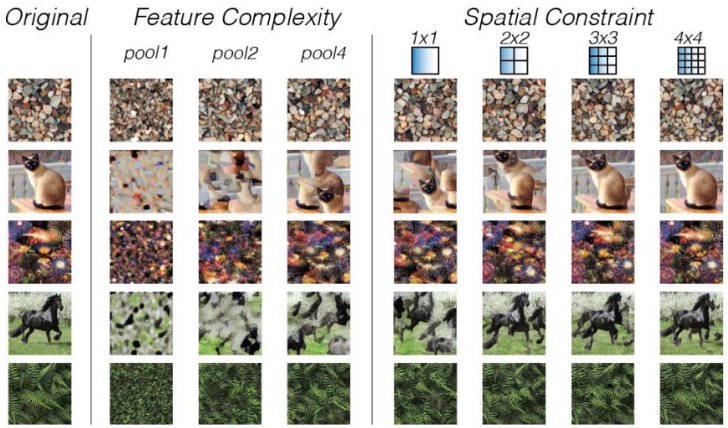
Replying to @dileeplearning

Transformers score worse on Brainscore; that is, they can't predict neural activity as well (despite performing well on Imagenet). That suggests they are worse models.

2:16 PM · Apr 1, 2021 · Twitter Web App

4 Likes

Multi-Resolution + Local Texture Computation

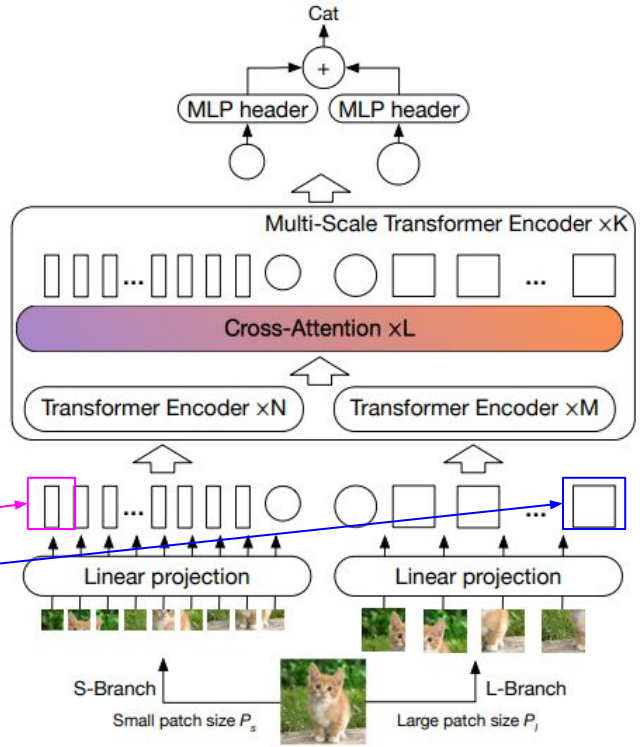


Gatys et al., 2016
Jagadeesh & Gardner, 2021

$$T_M = \begin{bmatrix} \phi_1(x)\phi_1(x) & \phi_1(x)\phi_2(x) & \dots & \phi_1(x)\phi_n(x) \\ \phi_2(x)\phi_1(x) & \phi_2(x)\phi_2(x) & \dots & \phi_2(x)\phi_n(x) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_n(x)\phi_1(x) & \phi_n(x)\phi_2(x) & \dots & \phi_n(x)\phi_n(x) \end{bmatrix}$$

Deza et al., 2020

- Dual-branch vision transformer to extract multi-scale feature representations + Gramian-like local texture computation

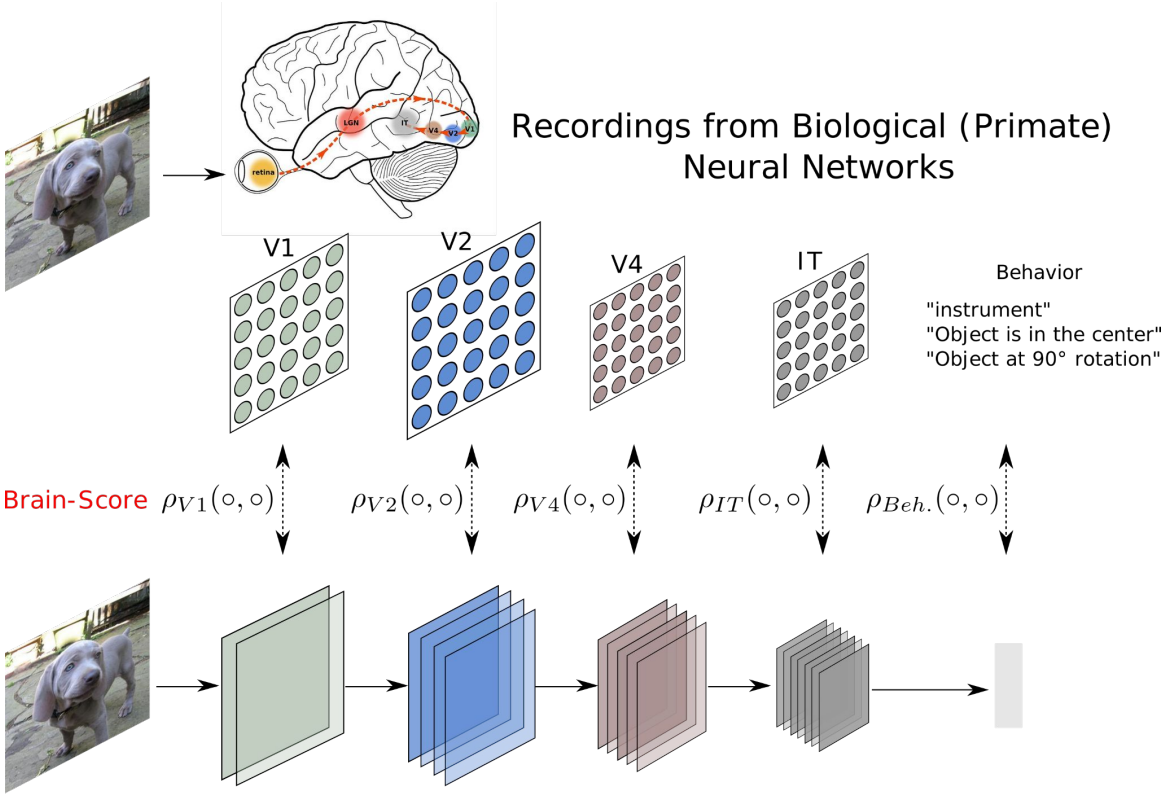


CrossViT

Chen et. al.,2021

Brain-Score

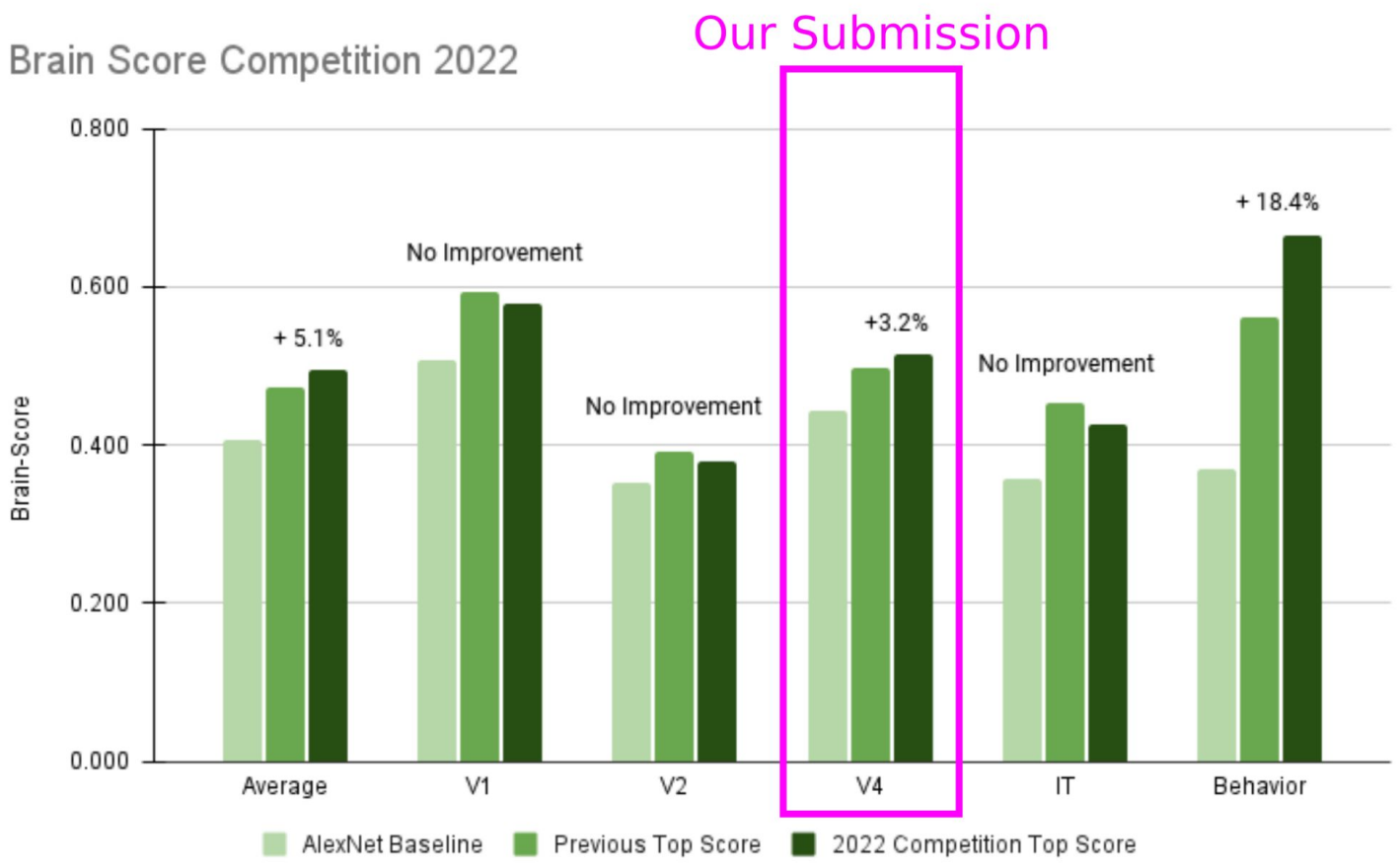
www.brain-score.org



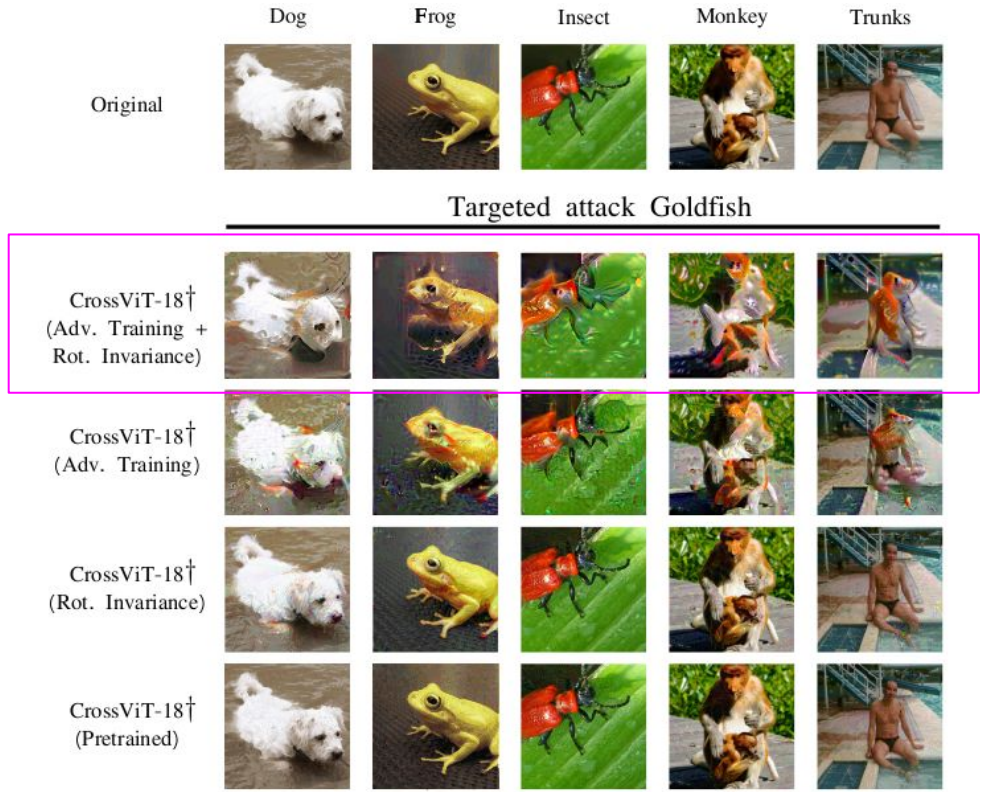
Our Model beyond the Competition

Rank	Model (submitted by)	average	V1 (23 benchmarks)	V2 (1 benchmark)	V4 (4 benchmarks)	IT (5 benchmarks)	behavior (1 benchmark)	engine
1	effnetb1_cutmixpatch_augmix_robust32_avge4e7_manylayers_324x288 — Alexander Riedel	.495	.568	.360	.481	.412	.652	.297
2	voneresnet-50-robust — Tiago Marques	.471	.531	.391	.471	.417	.545	X
3	custom_model_cv_18_dagger_408 — William Berrios	.467	.493	.342	.514	.425	.562	.473
4	resnet50_finetune_cutmix_e3_robust_linf8255_e0_247x234 — Alexander Riedel	.466	.584	.362	.472	.364	.549	
5	effnetb1_cutmix_augmix_sam_e1_5avg_424x377 — Alexander Riedel	.463	.482	.291	.499	.381	.664	.033
6	resnet50_finetune_cutmix_AVGe2e3_robust_linf8255_e0_247x234 — Alexander Riedel	.462	.584	.360	.464	.368	.536	.285
7	voneresnet-50-non_stochastic — Tiago Marques	.461	.569	.326	.484	.398	.530	.552

	Violet Xiang							
200	deit_tiny_patch16_224_id *Violet Xiang*	.151	.091	.251	.120	.081	.211	
201	ode_net_mar15 *Agrim Sharma*	.150	.423	.152	.120	.058		
202	resnet-18-LC_untrained *Roman Pogodin*	.137	.347	.056	.221	.115	-0.055	
203	dcn_full_mar15 *Agrim Sharma*	.134	.391	.128	.103	.047		
204	dcgan *Brain-Score Team*	.112	.158	.226	.108	.043	.023	
205	dcn_ode *Agrim Sharma*	.108	.144	.208	.129	.056		
206	unet_entire *Mike Ferguson*	.100	.135	.204	.119	.048	-0.006	
207	pixels-baseline *Brain-Score Team*	.051	.158	.003	.048	.028	.020	
208	resnet-50x4_untrained *Roman Pogodin*	X	X	X	X	X	X	

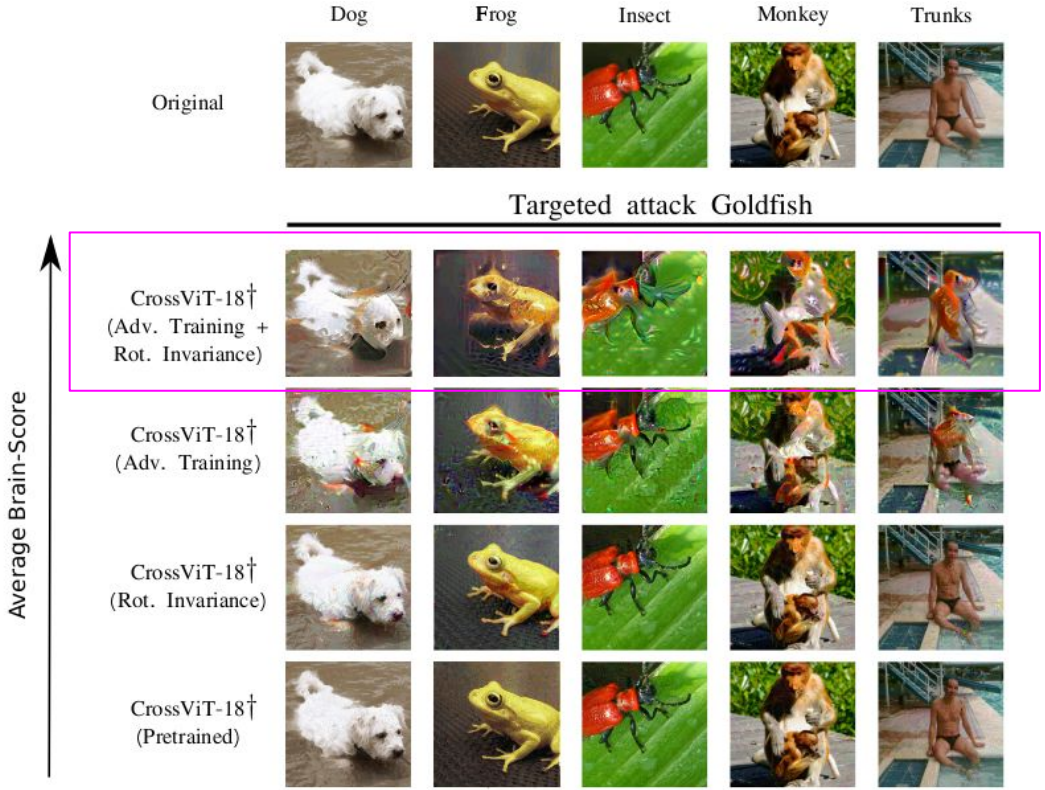

Our Submission
Brain Score Competition 2022
Brain-Score
0.800
0.600
0.400
0.200
0.000
+ 5.1%
No Improvement
No Improvement
+3.2%
No Improvement
+ 18.4%
Average
V1
V2
V4
IT
Behavior
AlexNet Baseline
Previous Top Score
2022 Competition Top Score

Human-machine perceptual alignment of the CrossViT via the effects of adversarial perturbations



Berrios & Deza. BSW 2022

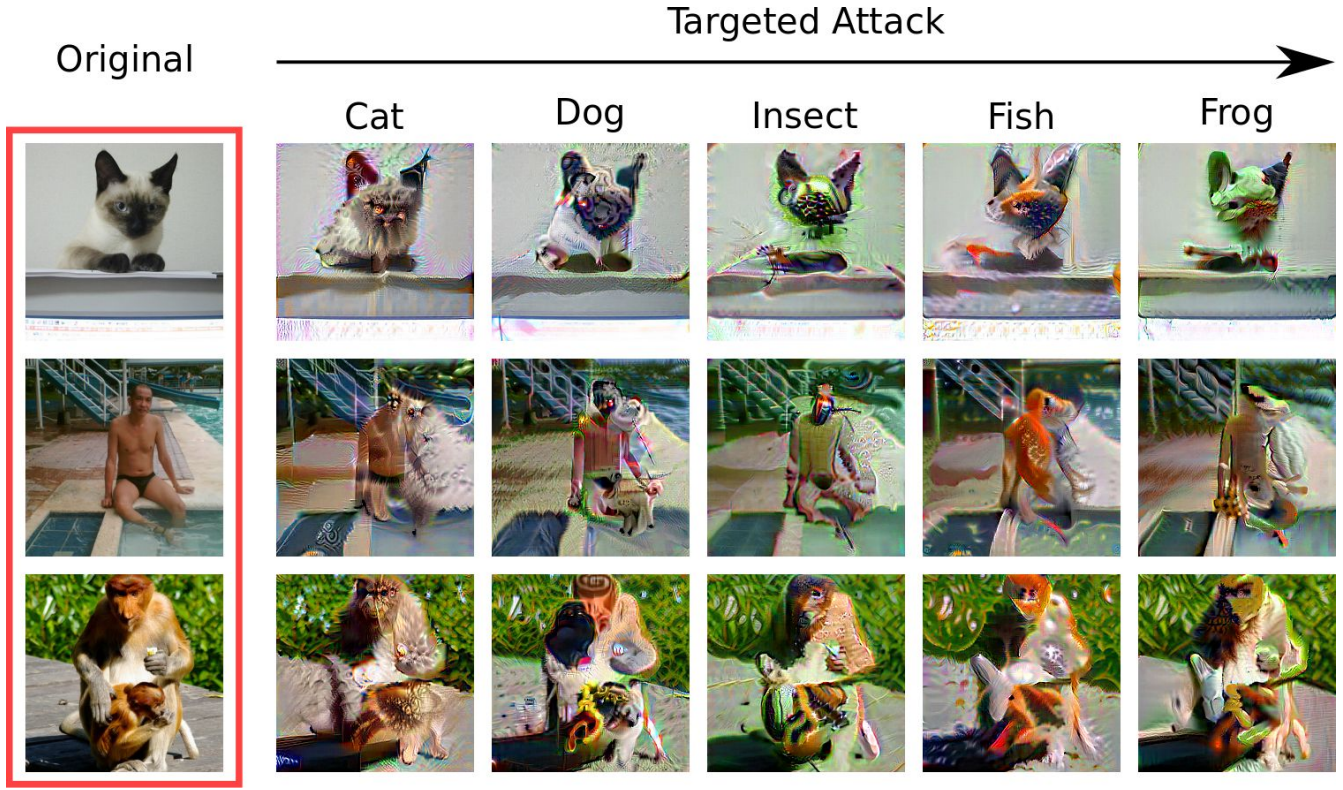
Human-machine perceptual alignment of the CrossViT via the effects of adversarial perturbations



"As the average Brain-Score increases in our system, the distortions seem to fool a human as well"

Berrios & Deza. BSW 2022

Human-machine perceptual alignment of the CrossViT via the effects of adversarial perturbations



Berrios & Deza. BSW 2022

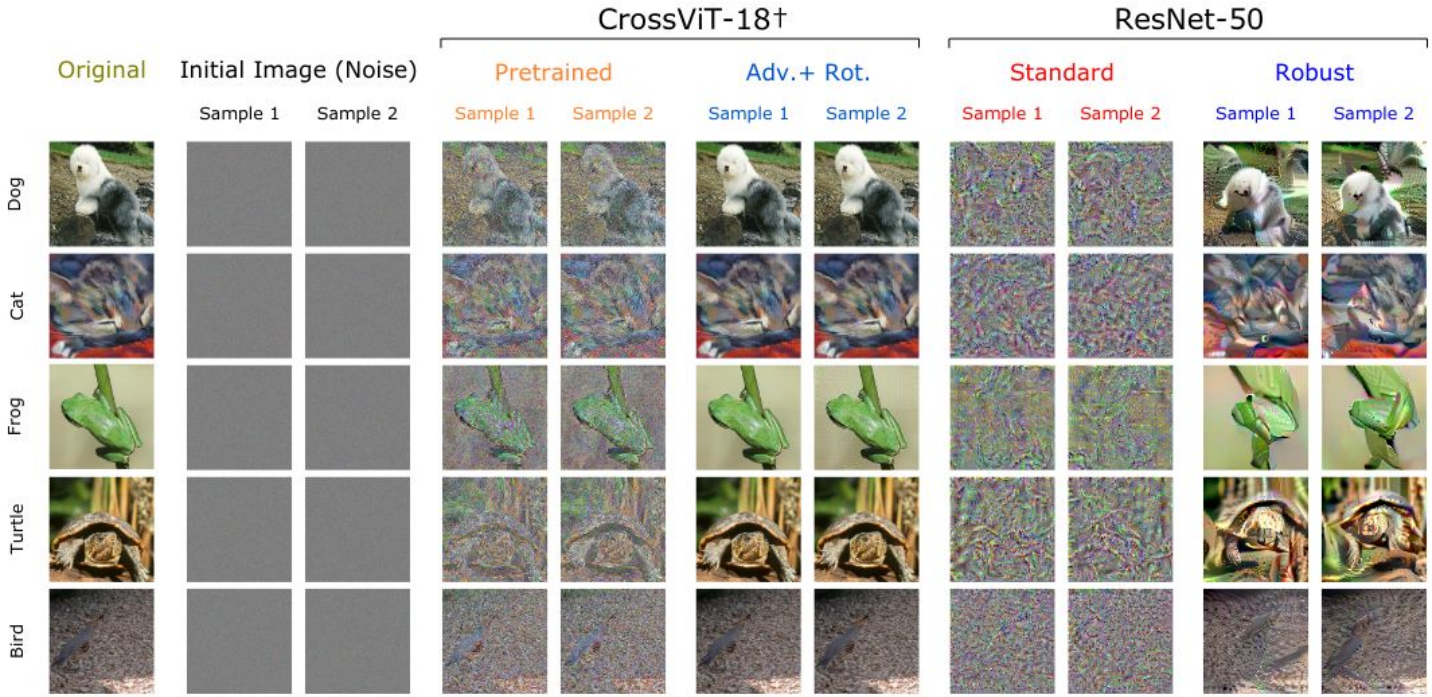


Figure 6: A summary of Feature Inversion models when applied on two different randomly samples noise images from a subset of the stimuli used in Harrington & Deza (2022). Standard and Pretrained models poorly invert the original stimuli leaving high spatial frequency artifacts. Adversarial training improves image inversion models, and this is even more evident for Transformer models. Notice that Transformer models independent of their optimization seem to preserve a higher shape bias as they recover the global structure of the original images. Extended figure can be viewed in the Appendix.

On-going work in NeuroAI

Strong and Precise Modulation of Human Percepts via Robustified ANNs

Guy Gaziv* **Michael J. Lee*** **James J. DiCarlo**
McGovern Institute for Brain Research, Dept. of Brain and Cognitive Sciences
Massachusetts Institute of Technology

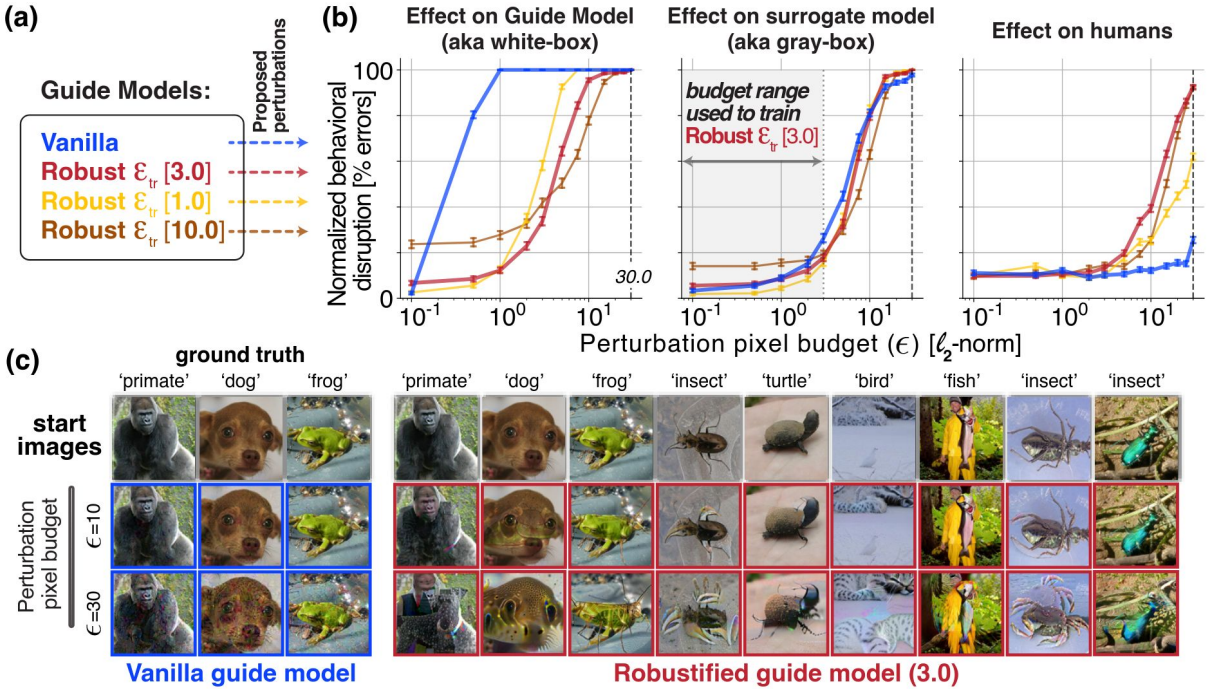
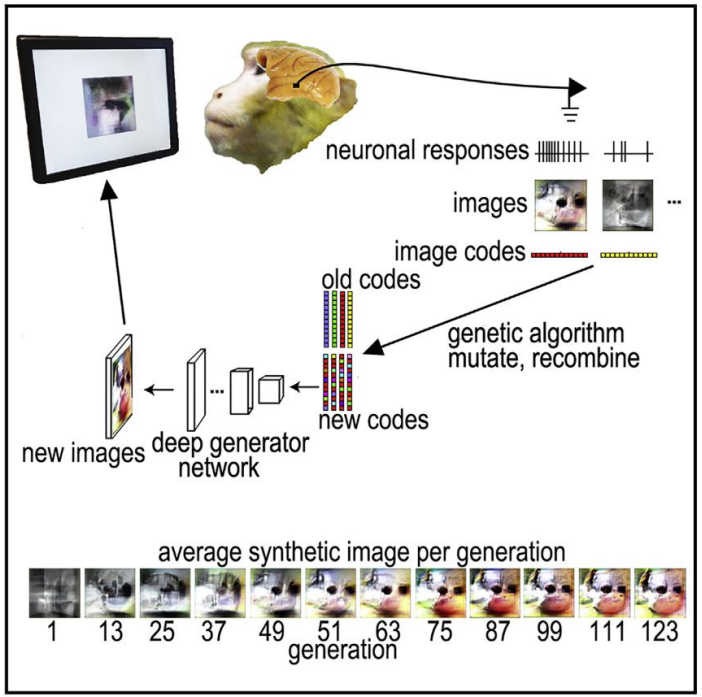


Figure 2: **Low-norm image perturbations discovered by robustified models strongly disrupt human category judgements.** ***(a)*** *The Guide Models used for Disruption Modulation (DM) image generation.*

Cell

Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences

Graphical Abstract



Authors

Carlos R. Ponce, Will Xiao,
Peter F. Schade, Till S. Hartmann,
Gabriel Kreiman, Margaret S. Livingstone

Correspondence

crponce@wustl.edu (C.R.P.),
mlivingstone@hms.harvard.edu (M.S.L.)

In Brief

Neurons guided the evolution of their own best stimuli with a generative deep neural network.

High-performance Evolutionary Algorithms for Online Neuron Control

Binxu Wang*
Harvard Medical School
Boston, MA, USA
binxu_wang@hms.harvard.edu

Carlos R. Ponce
Harvard Medical School
Boston, MA, USA
carlos@hms.harvard.edu

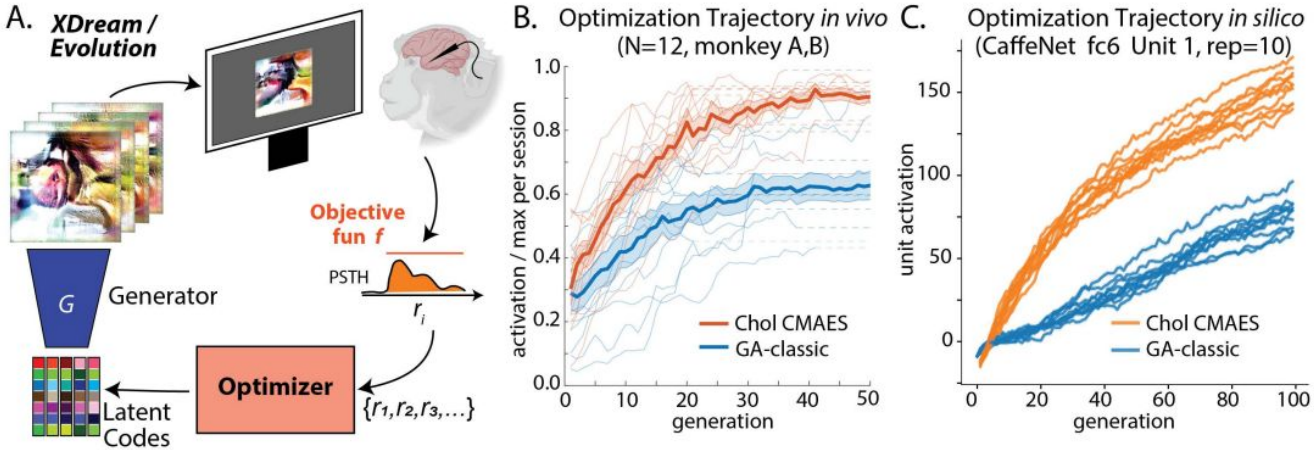


Figure 1: Cholsky-CMAES Excelled in Activation Maximization both *in silico* and *in vivo*. A. Schematics of the XDream Evolution experiment. **B.** *in vivo* optimization trajectory from 12 paired Evolution experiments in two monkeys. Thin curves show the trajectories for individual experiments; shaded thick curves show the mean and standard error (SEM) of trajectories across experiments. Experiments that terminated earlier were extrapolated by constant (dashed line) to match the generation number for mean and SEM calculation. **C.** *in silico* optimization trajectory comparison for unit 1 in fc6 layer of CaffeNet, mean activation per generation is plotted.

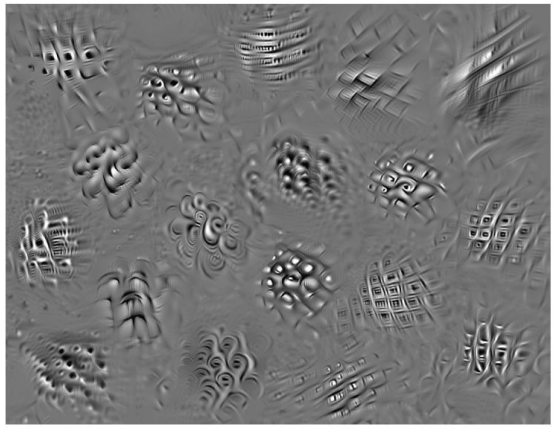
NEUROSCIENCE

Neural population control via deep image synthesis

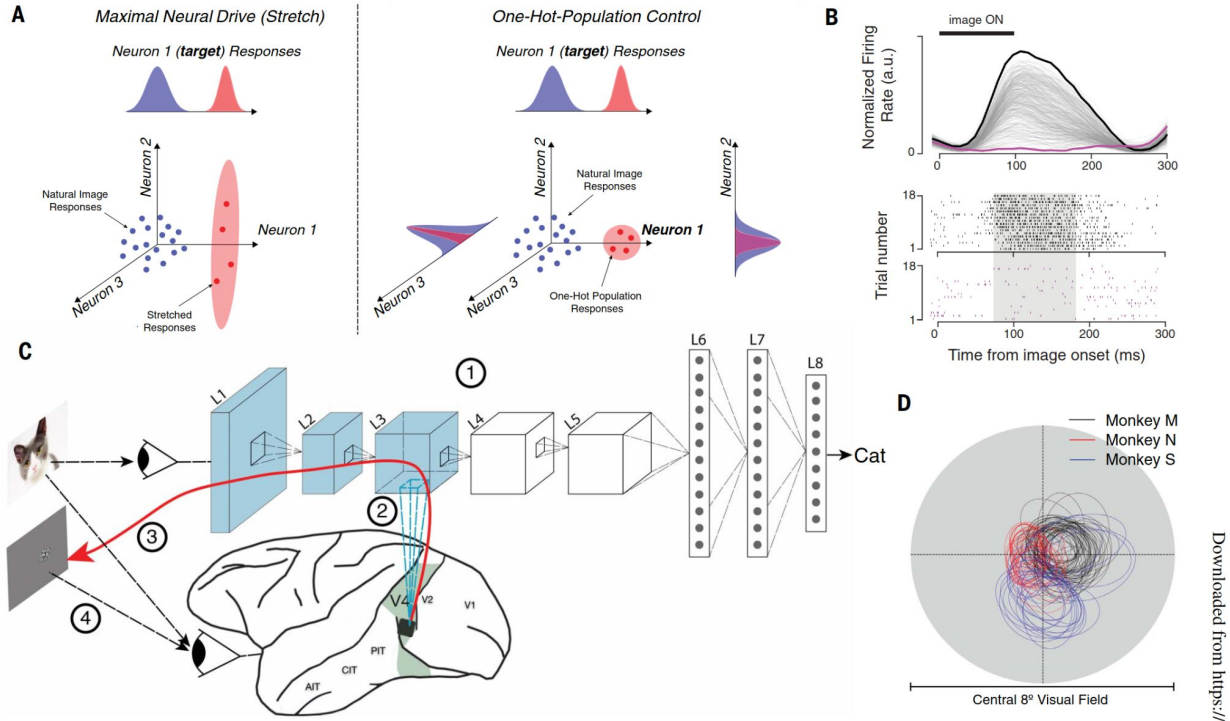
Pouya Bashivan*, Kohitij Kar*, James J. DiCarlo†

INTRODUCTION: The pattern of light that strikes the eyes is processed and re-represented via patterns of neural activity in a "deep" series of six interconnected cortical brain areas called the ventral visual stream. Visual neuroscience research has revealed that these patterns of neural activity underlie our ability to recognize objects and their relationships in the world. Recent advances have enabled neuroscientists to build ever more precise models of this complex visual processing. Currently, the best such models are particular deep artificial neural network (ANN) models in which each brain area has a corresponding model layer and each brain neuron has a corresponding model neuron. Such models are quite good at predicting the responses of brain neurons, but their contribution to an understanding of primate visual processing remains controversial.

RATIONALE: These ANN models have at least two potential limitations. First, because they aim to be high-fidelity computerized copies of the brain, the total set of computations performed by these models is difficult for humans to comprehend in detail. In that sense, each model seems like a "black box," and it is unclear what form of understanding has been achieved. Second, the generalization ability of these models has been questioned because they have only been tested on visual stimuli that are similar to those used to "teach" the models. Our goal was to assess both of these potential limitations through nonhuman primate neurophysiology experiments in a mid-level visual brain area. We sought to answer two questions: (i) Despite these ANN models' opacity to simple "understanding," is the knowledge embedded in them already useful for a



Collection of images synthesized by a deep neural network model to control the activity of neural populations in primate cortical area V4. We used a deep artificial neural network to control the activity pattern of a population of neurons in cortical area V4 of macaque monkeys by synthesizing visual stimuli that, when applied to the subject's retinae, successfully induced the experimenter-desired neural response patterns.



What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines?

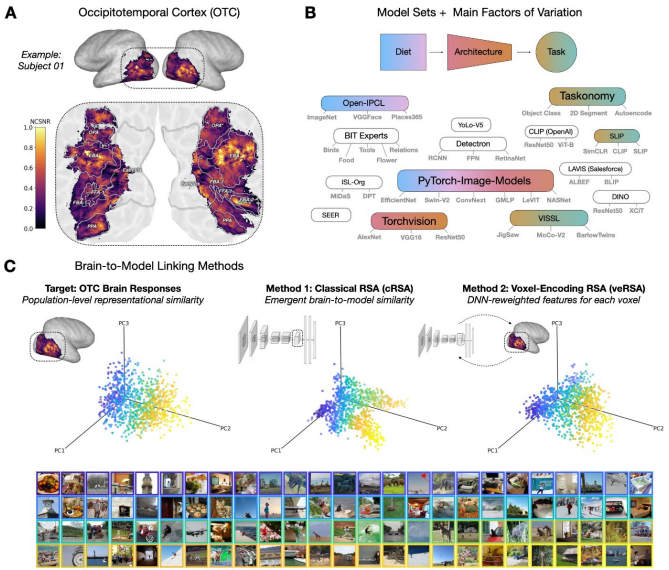
Colin Conwell[1]*, Jacob S. Prince[1], Kendrick N. Kay[2], George A. Alvarez[1], Talia Konkle[1,3,4]

[1]Department of Psychology, Harvard University

[2]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota

[3]Center for Brain Science, Harvard University

[4]Kempner Institute for Natural and Artificial Intelligence, Harvard University



Abstract

The rapid development and open-source release of highly performant computer vision models offers new potential for examining how different inductive biases impact representation learning and emergent alignment with the high-level human ventral visual system. Here, we assess a diverse set of 224 models, curated to enable controlled comparison of different model properties, testing their brain predictivity using large-scale functional magnetic resonance imaging data. We find that models with qualitatively different architectures (e.g. CNNs versus Transformers) and markedly different task objectives (e.g. purely visual contrastive learning versus vision-language alignment) achieve near equivalent degrees of brain predictivity, when other factors are held constant. Instead, variation across model visual training diets yields the largest, most consistent effect on emergent brain predictivity. Overarching model properties commonly suspected to increase brain predictivity (e.g. greater effective dimensionality; learnable parameter count) were not robust indicators across this more extensive survey. We highlight that standard model-to-brain linear re-weighting methods may be too flexible, as most performant models have very similar brain-predictivity scores, despite significant variation in their underlying representations. Broadly, our findings point to the importance of visual diet, challenge common assumptions about the methods used to link models to brains, and more concretely outline future directions for leveraging the full diversity of existing open-source models as tools to probe the common computational principles underlying biological and artificial visual systems.

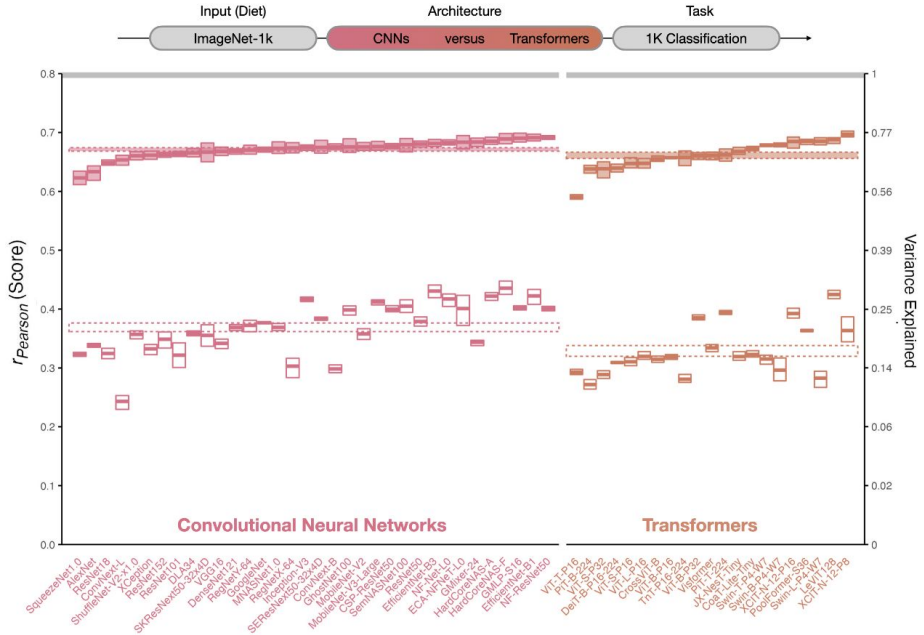


Figure 2: **Architecture Variation**. *Degree of brain predictivity ($r_{Pearson}$) is plotted for the controlled set of convolutional neural networks (CNNs) and transformer models in our survey. Each small box corresponds an individual model. The horizontal midline of each box indicates the model's mean score across the 4 subjects, with the height of the box indicating the grand-mean-centered 95% bootstrapped confidence intervals (CIs) (Morey et al., 2008) of the model's score across subjects. The cRSA score is plotted in open boxes, and the veRSA score is plotted in filled boxes. For each class of model architecture (convolutional, transformer) the class mean is plotted as a striped horizontal ribbon. The width of this ribbon reflects the 95% grand-mean-centered bootstrapped 95% CIs over the mean score for all models in a given set. The aggregate noise ceiling of the occipitotemporal brain data is plotted in the gray horizontal ribbon at the top of the plot, and reflects the mean and 95% CIs of the noise ceilings computed for each individual subject. The secondary y-axis shows explainable variance explained (the squared model score, divided by the squared noise ceiling).*

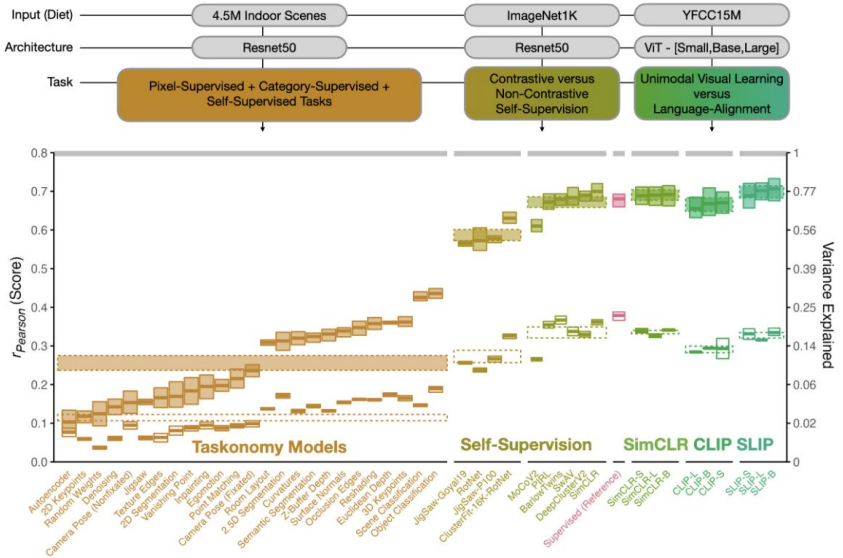


Figure 3: **Task Variation**. *Degree of brain predictivity ($r_{Pearson}$) is plotted for the sets of models with controlled variation in task. The first set of models shows scores across the ResNet50 encoders from Taskonomy, trained on a custom dataset of 4.5 million indoor scenes. The second set of models shows the difference between contrastive and non-contrastive self-supervised learning ResNet50 models (with a category-supervised ResNet50 for reference), trained on ImageNet1K. The third set of models shows the scores across the vision-only and vision-language contrastive learning ViT-[Small,Base,Large] models from FaceBook's SLIP Project, trained on the images (or image-text pairs) of YFCC15M. Each small box corresponds an individual model. The horizontal midline of each box indicates the model's mean score across the 4 subjects, with the height of the box indicating the grand-mean-centered 95% bootstrapped confidence intervals (CIs) of the model's score across subjects. The cRSA score is plotted in open boxes, and the veRSA score is plotted in filled boxes. The class mean for each distinct set of models is plotted in striped horizontal ribbons across the individual models. The width of this ribbon reflects the 95% grand-mean-centered bootstrapped 95% CIs over the mean score for all models in this set. The aggregate noise ceiling of the occipitotemporal brain data is plotted in the gray horizontal ribbon at the top of the plot, and reflects the mean and 95% CIs of the noise ceilings computed for each individual subject. The secondary y-axis shows explainable variance explained (the squared model score, divided by the squared noise ceiling).*