

Aligning deep neural networks with biological vision

By

Ivan Felipe Rodriguez Rodriguez

B.S. Universidad Sergio Arboleda, 2015

M.S. University of Puerto Rico , 2019

Thesis Dissertation

Submitted in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in the Department of Cognitive and Psychological
Sciences at Brown University

PROVIDENCE, RHODE ISLAND

February 2026

Copyright Statement

© Copyright 2026 Ivan Felipe Rodriguez Rodriguez

Signature Page

This dissertation by Ivan Felipe Rodriguez Rodriguez is accepted in its present form by the Department of Cognitive and Psychology sciences (COPSY) as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____
Thomas Serre, Advisor

Recommended to the Graduate Council

Date _____
William Warren , Reader

Date _____
David Sheinberg, Reader

Approved by the Graduate Council

Date _____
Thomas A. Lewis , Dean of the Graduate School

Curriculum Vitae

Name: Ivan Felipe Rodriguez
Phone: +1 401 867 1134
Email: ivan.rodriguez5@brown.edu
Homepage: piperod.github.io

Education

2019–2024

PhD in Cognitive Science
Brown University (Providence, RI, US)

2016–2018

MSc in Applied Mathematics
University Puerto Rico at Rio Piedras (San Juan, PR)

2015–2016

MSc in Scientific Computing
University Puerto Rico at Mayaguez (Mayaguez, PR)
Transferred to UPR Rio Piedras in the second year.

2014–2015

Pathway Program
Brigham Young University (ID, USA)
Grade: *A*

2008–2015

Mathematics
University Sergio Arboleda (Bogotá, Colombia)
Cum Laude Honors

Relevant Experience

Research & Industry

Aug 2019 – Present

PhD Student, Brown University (Providence, RI, US)

Researcher in Thomas Serre’s lab. I have developed new self-supervised methods that incorporate insights from human vision, proposing innovative architectures that enhance model interpretability and robustness, validated on international benchmarks.

Oct 2017 – 2020

External Collaborator and Visitor Scientist, Janelia Research Campus (Ashburn, VA, US)

Developed computer vision and machine learning algorithms for pose detection and behavior analysis in insects and mammals using part affinity fields and integer linear programming.

Jan 2016 – Jun 2019

Research Assistant, University of Puerto Rico at Rio Piedras (San Juan, PR)

Implemented computer vision algorithms in the BigDBee project to detect bee behavior and pollen recognition in over 100 TB of video data, enabling real-time risk assessment in beekeeping.

Sep 2015 – Sep 2018

Data Scientist, Red Cat Propware (San Juan, PR)

Developed algorithms for automatic roof damage quote calculations and for analyzing drone black-box data for anomaly detection, resulting in significant operational savings.

Teaching Experience

Jan 2022 – May 2022, Jan 2023, May 2023

Principal Instructor, Brown University (Providence, RI, US)

Taught the course *Introduction to Programming and Computational Reasoning* to 30 students, covering fundamentals in Python and Matlab.

Jun 2017 – Jun 2019

Research Mentor, University of Puerto Rico Rio Piedras (San Juan, PR)

Supervised and trained 6 undergraduate students in the Megret Lab project, focusing on computer vision and animal behavior tracking.

Publications

RTify: Aligning Deep Neural Networks with Human Behavioral Decisions
NeurIPS, 2024

Authors: Ivan Felipe Rodriguez *, Yu-Ang Cheng *, Sixuan Chen, Kohitij Kar, Takeo Watanabe, Thomas Serre.

HMAX strikes back: Self-supervised Learning of Human-Like Scale Invariant Representations

Cognitive Computational Neuroscience (CCN), 2024 (Oral)

Authors: Ivan Felipe Rodriguez, Nishka Pant, Thomas Serre.

Performance-optimized Deep Neural Networks are Evolving into Worse Models of Inferotemporal Visual Cortex

NeurIPS, 2023

Authors: Ivan Felipe Rodriguez, Thomas Fel, Drew Linsley, Thomas Serre.

Deep Neural Networks Face a Fundamental Trade-Off to Explain Human Vision

Cognitive Computational Neuroscience (CCN), 2022

Authors: Ivan Felipe Rodriguez, Drew Linsley, Thomas Serre.

Aligning Human and Machine Vision

NeurIPS, 2022

Authors: Ivan Felipe Rodriguez*, Thomas Fel*, Drew Linsley, Thomas Serre.

Automated Video Monitoring of Unmarked and Marked Honey Bees at the Hive Entrance

Frontiers in Science, 2022

Authors: Ivan Felipe Rodriguez, Jeffrey Chan, Kristin Branson, Jose L. Agosto, Remi Megret.

Multiple Animals Pose Detection Using Part Affinity Fields

VAIB ICPR, 2018 (Oral)

Authors: Ivan Felipe Rodriguez, R. Megret, et al.

Harmonizing the Visual Strategies of Image-Computable Models with Humans Yields More Performant and Interpretable Models of Primate Visual System Function

VSS, 2022 (Oral)

Authors: Ivan Felipe Rodriguez, Thomas Fel, Drew Linsley, Thomas Serre.

Using Artificial Intelligence to Identify Fossil Angiosperm Leaves at a Family Level

GSA, 2022

Authors: Ivan Felipe Rodriguez, Thomas Serre, Peter Wilf.

Honeybee Detection and Pose Estimation Using Convolutional Neural Networks

RFIAP, France, 2018

Authors: Ivan Felipe Rodriguez, R. Megret, K. Branson.

Recognition of Pollen-Bearing Bees from Video Using Convolutional Neural Networks

WACV, 2018 (Oral)

Authors: Ivan Felipe Rodriguez, R. Megret, et al.

Technical and Personal Skills

Programming Languages: Proficient in Python, Matlab, CUDA, Arduino, and L^AT_EX. Skilled in parallel computing, deep learning libraries (TensorFlow, PyTorch), and high-performance computing architectures.

Interests: Supervised and unsupervised learning, deep neural networks, explainability, alignment, time series analysis, and anomaly detection.

Service

2021-2022

CLPS Graduate student representant 2012–2014

Missionary, The Church of Jesus-Christ of Latter-day Saints (Antofagasta, Chile)

Awards

2019 Brown Presidential Fellowship

Acknowledgments

I would like to express my deepest gratitude to my advisor, Thomas Serre, for his unwavering support, guidance, and mentorship throughout my doctoral journey. His expertise, patience, and dedication have been instrumental in shaping this dissertation and my growth as a researcher.

I would like to express my gratitude to Bill Warren, who serves as a committee member but was also my first professor at Brown in Core I. I will always be grateful for his patience and dedication, for always asking the next hard question and always seeing beyond my grammar and typos, and focusing on the science. His questions helped me shape this dissertation and my research path.

I would also like to thank David Sheinberg, whose expertise and feedback helped me go the extra mile and challenged me to always go one step further.

I would like to acknowledge the financial support received from the National Science Foundation (NSF) under grant numbers 1734262 and 2021795, and the Office of Naval Research (ONR) under grant number N00014-19-1-2029. This research was also supported in part by the Brown University Graduate School and the Department of Cognitive, Linguistic, and Psychological Sciences.

Special thanks to the Serre lab (Drew, Thomas Fel, Yu-ang, Sixuan, Sabine, Lore, Lax, Alekh, Akash, Gaurav), for their collaboration, discussions, and friendship that made this journey more enjoyable and productive. I would like to thank all my co-authors and

collaborators: Drew who became a second mentor, to find inspiration and guidance when no ideas were left; to Thomas Fel, a great friend with incredible talent, who inspired me with his brilliance and humility; to Yu-ang and the deep discussions and collaborations; to Sixuan and her support and devotion in making our Rtify paper a reality; to Xizheng, Niskha, Scott, Arjun who pushed the extra mile in the HMAX paper and were always prompt to go the extra mile and stay late. To my external collaborators, Marge Livingstone, Salomi Sharma and Michael Arcaro, for their incredible work in collecting the neural recordings and their openness to share their data and collaborate. To Carlos Ponce for collecting more data for us and putting such great effort into our project.

I am indebted to the faculty and staff of the Department of Cognitive and Psychological Sciences at Brown University for their support and for creating an environment that fostered my academic growth.

I would like to thank my colleagues and friends in the COPSY (Former CLPS) department at Brown University for creating an intellectually stimulating and supportive environment. I always felt welcome and cannot help but thank my cohort (Vivi, Zhenyu, Haley, Misha, Mahalia) who were always reliable friends to find support and strength in hard moments. I will never forget how important you were when the pandemic forced us into isolation, but you always found ways to keep us connected and go through this challenge together.

At Brown I also met friends that became family. I am so grateful for Lax who was the reason I came to Brown, in the first place, after meeting him in China; he became a brother who would always look out for me and was there when hard academic and life challenges were present. Thanks to Vivi who taught with me Core 1 and was always there to hear me complain about my experiments not working out, her great taste in series helped me on those moments where I needed to catch a breath. Jooyoung for being a great friend and the best office mate, she was especially patient when the office, filled with scientific discussions between me and Yuang, reached high emotion. She was also a source

of support when results did not workout and when writing was giving me headaches. Alex who was an incredible and talented partner to always look for new projects and have the most fun doing them, also to his wife Tian-Tian for sponsoring most of these meetings with delicious food. Lore for her sharp questions and looking out for the welfare of the group. Sofia for her friendship and support. To Jovan, my gym buddy and friend, who was always there to help me with my lifts and always there to talk about the latest science.

To my family, especially the love of my life, my wife, Luisa, who was always there to support me and who sacrificed so much time together to allow me to pursue this dream. To my mom, my biggest fan, who was always there to support me and encourage me to keep going when experiments did not pan out. To my dad and his calls of support. To my siblings Andres, Melissa and Laura, always prompt to provide me a word of encouragement. Thank you for your unconditional love, understanding, and encouragement throughout this journey. Your support has been my constant source of strength and motivation.

Finally, I would like to express my gratitude to all the participants who contributed to the studies presented in this dissertation. Their time and willingness to participate were essential to this research.

This dissertation represents not only my work but also the collective support and contributions of many individuals who have been part of this journey. I am truly grateful to everyone who has helped me reach this milestone.

Contents

Copyright Statement	i
Signature Page	iii
Curriculum Vitae	iv
Acknowledgments	viii
1 INTRODUCTION	1
1.0.1 Defining Alignment	3
1.0.2 The Performance-Alignment Trade-off	3
1.1 The Need for a New Approach	4
2 HARMONIZING THE OBJECT RECOGNITION STRATEGIES OF DEEP NEURAL NETWORKS WITH HUMANS	11
2.1 Introduction	11
2.2 Related Work	12
2.2.1 Do DNNs explain human visual perception?	12
2.2.2 What are the visual strategies underlying human object recognition?	13
2.2.3 Comparing visual strategies of humans and machines	14
2.2.4 Improving the correspondence between humans and machines	14
2.3 Methods	15

2.3.1	Human feature importance datasets	15
2.3.2	Psychophysics participants and dataset	16
2.3.3	Psychophysics experiment	16
2.3.4	Models	17
2.4	Results	18
2.4.1	<i>Where</i> are diagnostic object features for humans and DNNs? . . .	18
2.4.2	The neural harmonizer	19
2.4.3	Training	22
2.4.4	The neural harmonizer aligns human and DNN visual strategies .	23
2.4.5	Clicktionary	24
2.4.6	<i>How</i> do humans and DNNs integrate diagnostic object features into decisions?	24
2.5	Conclusion	28
2.6	Limitations	29
2.7	Broader impacts	31
2.8	Supplementary Information: Harmonizing the Object Recognition Strate- gies of Deep Neural Networks with Humans	31
2.8.1	Psychophysics	31
2.8.2	Harmonization loss	33
2.8.3	Additional Results	33
2.8.4	<i>ClickMe</i>	33
2.8.5	<i>ViT attention</i>	34
2.8.6	<i>Correlations between measurements of human visual strategies</i> . .	34

3 NEURAL ALIGNMENT: PERFORMANCE-OPTIMIZED DEEP NEU- RAL NETWORKS ARE EVOLVING INTO WORSE MODELS OF INFEROTEMPORAL VISUAL CORTEX 37

3.1	Introduction	37
-----	------------------------	----

3.2	The Trade-off Between ImageNet Accuracy and Neural Predictivity . . .	38
3.3	Methods	39
3.3.1	Neural recordings	39
3.3.2	DNNs	40
3.3.3	Neural Harmonizer	41
3.3.4	Neural fitting	43
3.4	Results	43
3.4.1	Task Optimization is insufficient for reverse-engineering IT	43
3.4.2	DNNs need biologically-aligned training routines	45
3.4.3	Spatially-mapped neural responses reveal a feature mismatch be- tween IT and DNNs	46
3.4.4	hDNNs generate testable predictions for visual neuroscience . . .	47
3.5	Discussion	49
3.5.1	A revised approach for reverse engineering visual cortex	49
3.5.2	Limitations	50
3.5.3	Broader impacts	51
3.6	Final Remarks: Connecting Neural Alignment to the Thesis	51
3.7	Supplementary Information: Neural Alignment	53
3.7.1	Datasets	53
3.7.2	Neural data analysis	53
3.7.3	Models	54
3.7.4	<i>Neural Harmonizer</i> Training	55
3.7.5	Extended results	55

4 HMAX STRIKES BACK: BIOLOGICALLY-INSPIRED ARCHITECTURES FOR SCALE INVARIANCE 63

4.1	Introduction	63
4.1.1	Chapter Overview and Contributions	65

4.1.2	The HMAX Family: Progressive Modernization	67
4.2	The Scale Invariance Challenge	75
4.2.1	Demonstrating scale-invariance of the HMAX model on simple dataset scale-MNIST	76
4.3	Experimental Validation	81
4.3.1	Results on ImageNet	81
4.3.2	ImageNet Scale Invariance Evaluation	83
4.3.3	Brain-Score Neural Prediction Benchmark	85
4.4	Conclusions	90
4.4.1	Limitations and Missing Biological Components	93
4.4.2	Future Directions	95
4.4.3	Broader Impact	95
5	RTIFY: ALIGNING DEEP NEURAL NETWORKS WITH HUMAN BEHAVIORAL DECISIONS	97
5.1	Introduction	97
5.2	The Challenge of Temporal Alignment	98
5.2.1	Two Disconnected Research Traditions	98
5.2.2	The Need for Integration	99
5.3	RTify: Overview of the Method	100
5.3.1	Core Framework	100
5.3.2	Differentiable Approximation	102
5.3.3	Training Scenarios	102
5.3.4	RTifying Feedforward Networks	103
5.4	Experimental Results	105
5.4.1	Random Dot Motion Task	105
5.4.2	Object Recognition Task	106
5.4.3	RTifying Feedforward Neural Networks	107

5.5	Theoretical Analysis	110
5.5.1	Complete Derivation of the Differentiable Framework	110
5.6	Extended Results and Analysis	112
5.6.1	Extended Results for RDM Task	112
5.6.2	RTified WW Model on RDM Task	112
5.6.3	Additional Experiment for RDM Task	112
5.6.4	Illustration of the RTified WW Circuit	114
5.7	Implications and Future Directions	114
5.7.1	Integration of Vision and Decision-Making Models	114
5.7.2	Biological Plausibility	115
5.7.3	Limitations and Future Work	115
5.7.4	Broader Impacts	116
5.8	Conclusion	116
6	CONCLUSIONS AND FUTURE DIRECTIONS	120
6.1	Summary of Contributions	120
6.1.1	Documenting the Performance-Alignment Trade-off	120
6.1.2	Neural Harmonization: A Novel Approach	121
6.1.3	Two Complementary Pathways to Biological Alignment	121
6.1.4	Modernizing Biologically-Inspired Architectures: HMAX 3.0	122
6.1.5	Reaction Time Alignment	123
6.2	Implications for Neuroscience	123
6.2.1	Rethinking the Optimization Hypothesis	123
6.2.2	The Importance of Multiple Objectives	123
6.2.3	The Role of Behavioral Data	124
6.3	Implications for Artificial Intelligence	124
6.3.1	The Scaling Problem	124
6.3.2	The Need for New Benchmarks	125

6.3.3	The Promise of Neural Alignment	126
6.4	Future Directions	127
6.4.1	Expanding Neural Harmonization	127
6.4.2	Multi-Modal Alignment	128
6.4.3	Understanding the Mechanisms	129
6.4.4	Combining Training and Architectural Solutions	130
6.4.5	Developing New Architectures	132
6.4.6	Applications to Brain-Computer Interfaces	134
6.5	Toward a Unified Framework for Biological Intelligence	134
6.5.1	The Path Forward	134
6.5.2	A Roadmap for Integration	135
6.5.3	Broader Implications	135
6.5.4	Final Reflections	136

List of Figures

2.1	Human and DNNs rely on different features to recognize objects.	
	In contrast, our neural harmonizer aligns DNN feature importance with humans. We smooth feature importance maps from humans (<i>ClickMe</i>) and DNNs with a Gaussian kernel for visualization.	20
2.2	The trade-off between DNN performance and alignment with human feature importance from <i>ClickMe</i> Linsley et al. (2019).	
	Human feature alignment is the mean Spearman correlation between human and DNN feature importance maps, normalized by the average inter-rater alignment of humans. The shaded region denotes the pareto frontier of the trade-offs between ImageNet accuracy and human feature alignment for unharmonized models. Harmonized models (VGG16, ResNet50, ViT, and EfficientNetB0) are more accurate and aligned than versions of those models trained only for categorization. Error bars are bootstrapped standard deviations over feature alignment. Arrows show a shift in performance after training with the neural harmonizer. The feature alignment of an average of <i>ClickMe</i> maps with held-out maps is denoted by †.	22

2.3	The trade-off between DNN performance and alignment with human feature importance from <i>Clicktionary</i> Linsley et al. (2017). Human feature alignment is the mean Spearman correlation between human and DNN feature importance maps, normalized by the average inter-rater alignment of humans. The shaded region denotes the pareto frontier of the trade-offs between ImageNet accuracy and human feature alignment for unharmonized models. Harmonized models (VGG16, ResNet50, MobileNetV1, and EfficientNetB0) are more accurate and aligned than versions of those models trained only for categorization. Error bars are bootstrapped standard deviations over feature alignment. Arrows denote a shift in performance after training with the neural harmonizer.	25
2.4	Comparing <i>how</i> humans and DNNs use visual features during object recognition. (a) Humans and DNNs categorized ImageNet validation images as animals or non-animals. The images revealed only a portion of the most important visual features according to the <i>Clicktionary</i> game Linsley et al. (2017). (b) There was a trade-off between DNN top-1 accuracy on ImageNet and alignment with human visual decision making. The shaded region denotes the pareto frontier of the trade-off between ImageNet accuracy and human feature alignment for unharmonized models. Arrows denote a shift in performance after training with the neural harmonizer. Error bars are bootstrapped standard deviations over decision-making alignment. (c) A state-of-the-art DNN like the ViT learned a different strategy for integrating visual features into decisions than humans or a harmonized ViT.	36

3.1	IT recordings that reveal spatial maps of neuronal responses to complex natural images offer unprecedented insights into their feature selectivity Arcaro et al. (2017).	
	(a) Neurons in posterior (PL) and/or medial (ML) lateral IT in two animals were localized using functional magnetic resonance imaging (fMRI), and neural responses to images were recorded using chronically implanted 32-channel multi-electrode arrays.	
	(b) The monkeys were rapidly shown each image multiple times for 200ms each time (Monkey 1: $n = 256$, Monkey 2: $n = 49$). Images were positioned differently each time to measure neural responses to every part of the image.	
	(c) This procedure yielded spatially resolved maps of neural activity for an image, revealing the relative importance of different features for the recorded neurons.	
	(d) DNNs were fit to these recordings following the Brain-Score evaluation method Schrimpf et al. (2018), in which partial least squares decoders were used to find the units in a DNN that provided the best match for neuronal responses to images.	44
3.2	DNNs trained on ImageNet face a trade-off between achieving high object recognition accuracy and predicting responses to natural images in IT..	
	We measured the accuracy of 135 DNNs at predicting image-evoked responses from neurons in posterior lateral (PL) and medial-lateral (ML) areas of IT Arcaro et al. (2017) by computing the neural predictivity of each model with the Brain-Score evaluation method Schrimpf et al. (2018). DNN neural predictivity is progressively worsening as models improve on ImageNet. This problem can be partially fixed by training DNNs with the <i>neural harmonizer</i> , which aligns their learned object representations with humans. Error bars denote 95% bootstrapped confidence intervals.	45

3.3	DNNs optimized for object recognition on ImageNet rely on different features than those encoded by neurons in primate inferior temporal (IT) cortex.. The activity of PL IT neurons is plotted next to the predicted activity of a model representing each class of DNNs: harmonized DNNs (hDNNs), visual transformers, self-supervised DNNs, convolutional neural networks, DNNs trained on more data than ImageNet, and adversarially robust DNNs.	48
3.4	Predictions of the visual features drive neuronal responses in PL of Monkey 1 from a harmonized ResNet-50 and a standard ResNet-50. Image patches for each model depict what the important features are across all images shown to the animal. The relative importance of each feature for predicting the recordings is color-coded in the bar chart on the right.	49
3.5	Stimuli from Majaj et al. (2015), but not Arcaro et al. (2017), are out-of-distribution from ImageNet. Penultimate-layer activities from a ResNetv2-50 trained on ImageNet were extracted for 1916 images from the validation set of ImageNet, 191 images from Majaj et al. (2015), and 28 images from Arcaro et al. (2017). A principle components analysis decomposed those activities into a 2-Dimensional projection, which demonstrated a distribution shift for stimuli from Majaj et al. (2015), but not Arcaro et al. (2017). Indeed, while stimuli from Majaj et al. (2015) could be reliably discriminated from ImageNet (98.7 acc, $T(999) = 98.12$, $p < 0.001$), images from Arcaro et al. (2017) were not discriminable from ImageNet (56.5 acc, $T(999) = 98.12$, $n.s$	59
3.6	Noise ceilings at every 40ms bin of activity from 0ms - 250ms. We binned every 40ms of neural responses following the original procedure for processing this data Arcaro et al. (2017)	60

3.7	DNN neural prediction accuracy after evaluating on a test set held out of training. Unlike the original approach introduced in Brain-Score Schrimpf et al. (2018), here we train, validate, and test models on separate partitions of the dataset. DNNs trained and evaluated in this way follow the trends described in the main text.	60
3.8	Results using different time bins than in the main text for the same region (ML) for both monkeys. The pattern of results for each time-bin are consistent with the main text, despite these data having lower noise ceilings.	61
3.9	Results using different time bins than in the main text for the IT PL in Monkey 1.. The pattern of results for each time-bin are consistent with the main text, despite these data having lower noise ceilings.	61
3.10	DNNs optimized for object recognition on ImageNet rely on different features than those encoded by neurons in primate inferior temporal (IT) cortex. The activity of ML IT neurons of Money 1 is plotted next to the predicted activity of a model representing each class of DNNs: harmonized DNNs (hDNNs), visual transformers, self-supervised DNNs, convolutional neural networks, DNNs trained on more data than ImageNet, and adversarially robust DNNs.	62
4.1	Architectural improvements in HMAX 3.0 (CHResMAX). Left: Wider S1 blocks with 11×11 kernels and 96 channels for enhanced feature extraction. Center: Stacked 3×3 kernels maintaining similar receptive field size to the original while enabling smoother gradient flow. Right: Residual connections within blocks and learnable scale selection mechanism (Conv-scoring) that replaces hard max pooling with soft, data-driven scale selection for improved gradient flow and semantic coherence.	73

4.2	Scale invariance evaluation on MNIST. Top: Examples from the scale-MNIST dataset showing the same digit at different scales. Bottom Left: Performance degradation of ResNet18 and VGG models when tested on scales outside their training distribution, demonstrating their reliance on data augmentation rather than intrinsic scale invariance. Bottom Right: Superior generalization of HMAX 2.1 and HMAX 2.5 (CHMAX) to unseen scales, highlighting the effectiveness of architectural scale-invariant mechanisms.	77
4.3	Hangul one-shot learning evaluation. Left: Experimental paradigm showing target/distractor pairs of visually similar Korean characters at different scales. Right: Model performance comparison including human baseline from Han et al. (2020b). HMAX 2.5 (CHMAX) demonstrates human-level performance on this challenging cross-scale generalization task, while standard CNNs trained with extensive data augmentation still fall short.	79
4.4	D-prime analysis of Hangul one-shot learning performance across scale conditions. The d-prime metric ($d' = \text{proportion correct on target/distractor pairs} - \text{proportion incorrect on target/target pairs}$) provides a bias-corrected measure of discriminative sensitivity. HMAX 3.0 (green) demonstrates consistently high d-prime values ($d' \geq 0.85$) across all scale conditions, with minimal performance degradation in cross-scale conditions. AlexNet (blue) and ResNet18 (orange) show dramatically impaired performance in cross-scale conditions, with d-prime values dropping to 0.13–0.28, while maintaining near-perfect performance ($d' \approx 1.0$) only in the same-scale condition (30'/30'). This pattern reveals that data augmentation approaches learn scale-specific representations rather than achieving true scale invariance. Error bars represent standard error across the 27 target/distractor character pairs.	82

4.5	ImageNet scale invariance evaluation. Main plot: Performance across seven image scales (160×160 to 454×454 pixels) showing HMAX 3.0’s robust generalization compared to AlexNet and ResNet18. Inset: Off-reference score comparison demonstrating HMAX’s superior ability to generalize beyond training scales. While baseline models show sharp performance drops outside their augmentation window, HMAX 3.0 maintains high accuracy across the full scale range, achieving 72.3% top-1 accuracy while preserving scale invariance.	83
4.6	Summary of the changes introduced to the HMAX models and its effect on top-1 accuracy performance on ImageNet. The progressive architectural improvements across the HMAX family show consistent gains, with each innovation contributing to both performance and scale invariance capabilities.	85
4.7	Example stimuli from the Brain-Score benchmark dataset. Images show three object categories (boat, animal, and chair) at different scales and orientations, demonstrating the controlled transformations used to evaluate neural prediction accuracy. The dataset contains 3,200 images spanning 8 object categories with systematic scale, position, and orientation variations to test the robustness of neural representations.	86
4.8	Brain-Score evaluation across scale variations. Main plot: Neural prediction accuracy across five scale bins, showing HMAX 3.0’s superior generalization compared to standard CNNs. Scale-BrainScore: Average performance excluding the reference scale, demonstrating HMAX’s ability to extrapolate to unseen scales. The results show that architectural mechanisms designed for scale invariance enable better prediction of IT neural responses under scale variation, supporting the hypothesis that biologically-inspired design principles improve neural alignment.	88

5.1	Illustration of our RTify method. The input is a visual stimulus represented by random moving dots, but the model can also accommodate color images and video sequences. We take a pretrained task-optimized RNN and use a trainable function f_w to transform the activity of the network into a real-valued evidence measure, e_t , that will be integrated over time by an evidence accumulator, Φ_t . When the evidence accumulator reaches the threshold θ , processing stops, and a decision is taken. The time step at which the accumulated evidence passes this threshold τ_θ is taken as the model RT for this stimulus.	101
5.2	Illustration of RTifying feedforward neural networks. We develop a multi-class compatible and fully differentiable RNN module based on the WW model Wang (2002); Wong and Wang (2007). This module is implemented as an attractor-based RNN, and is stacked on top of a feedforward neural network. The feedforward neural network first takes an image as the input. Outputs from classification units of the network are then sent to RTified WW (A). Information is accumulated by multiple populations of neurons in RTified WW while they compete with each other (B). A decision is made and the process stops when one of the populations reaches a threshold. The number of time steps needed for the RTified WW to reach the threshold is used to predict human RT (C).	104
5.3	RTified model evaluation on a RDM task Green and Swets (1966). Human data are shown as a gray shaded area, and model fits are shown for (A) the “supervised” setting where human behavioral responses are used to train the models and (B) the “self-penalized” setting where no human data is used. Our approach (green) outperforms the two alternative approaches (brown), i.e., entropy-thresholding Spoerer et al. (2020) for the “supervised” and uncertainty proxy Goetschalckx et al. (2021) for the “self-penalized” settings.	106

5.4	(A) MSE comparisons for the RDM task Green and Swets (1966) for all coherence levels. The RTified model trained in the “supervised” setting (i.e., with human behavioral responses; green solid line) performs better (lower MSE) than entropy-thresholding Spoerer et al. (2020) (brown solid line) under all coherence levels. Similarly, the RTified model trained in the “self-penalized” setting (i.e., without human data; green dash line) performs better than uncertainty proxy Goetschalckx et al. (2021) (brown dash line). (B) Classification accuracy comparisons between pre-trained and RTified models for the RDM task Green and Swets (1966). The RTified model trained with human RTs data achieves human-like classification accuracy under all coherence levels compared with the pretrained model without RTify.	107
5.5	RTified model evaluation on an object categorization task Kar et al. (2019). Model vs. human RT predictions for our RTified model (green) vs. alternative approaches (brown) (A) in the “supervised” setting where human behavioral responses are used to train the model and (B) the “self-penalized” setting where no human data is used. Solid lines are linear regression fits between model and human RTs. Our approach outperforms the two alternative approaches, i.e., entropy-thresholding Spoerer et al. (2020) for the “supervised” setting and uncertainty proxy Goetschalckx et al. (2021) for the “self-penalized” setting.	108
5.6	RTified WW model evaluation. We combine our RTified WW module with (A) a 3D CNN to fit human RTs collected in an RDM task Green and Swets (1966) and (B) a VGG to fit human RTs in a rapid object categorization task Kar et al. (2019).	109

5.7	Illustration of Mathematics Proof	The discrete RNN step τ_θ is not differentiable. Therefore, we introduce a piecewise linear approximation of the accumulated evidence over time Φ_t . Consider the effect of changes in Φ_t on $\tau_\theta(\Phi)$. A small perturbation in Φ_t will produce a proportional change in the time it takes for the accumulated evidence to cross the threshold θ , thereby inducing a shift in time. In simple terms, fine-tuning w to decrease Φ_t will delay the time at which the network crosses the threshold θ thus increasing $\tau_\theta(\Phi)$, while fine-tuning w to increase Φ_t will cause the threshold to be crossed earlier thus decreasing $\tau_\theta(\Phi)$	111
5.8	RTified model evaluation on a RDM task Green and Swets (1966) across all coherences.	Human data are shown as a gray shaded area, and model fits are shown for (A) the “supervised” setting where human behavioral responses are used to train the models and (B) the “self-penalized” setting where no human data is used. Our approach (green) outperforms the two alternative approaches (brown), i.e., entropy-thresholding Spoerer et al. (2020) for the “supervised” and uncertainty proxy Goetschalckx et al. (2021) for the “self-penalized” settings.	113
5.9	RTified WW model on a RDM task Green and Swets (1966) across all coherences.	We combined our RTified WW module with a 3D CNN to fit human RTs collected in an RDM task Green and Swets (1966). Human data are shown as a gray shaded area and model results are shown as orange solid lines.	114

5.10	RTified model evaluation on a RDM task Green and Swets (1966)	
	across all coherences. Human data are shown as a gray shaded area, and model fits are shown for (A) the “supervised” setting where human behavioral responses are used to train the models and (B) the “self-penalized” setting where no human data is used. Our approach (green) outperforms the two alternative approaches (brown), i.e., entropy-thresholding Spoerer et al. (2020) for the “supervised” and uncertainty proxy Goetschalckx et al. (2021) for the “self-penalized” settings. MSE comparison is shown in (C) .	118
5.11	Illustration of the RTified WW circuit. To RTify feedforward neural networks, we used an RNN based on the WW circuit. Here, we consider a binary classification on random moving dots for illustration (A) . When receiving a visual input, the two populations sensitive to left/right directions compete with each other (B) and accumulate evidence until one of them reaches a threshold (C) . The number of time steps needed for the RNN to reach the threshold is defined as the “model RT” and is used to predict human RT. We provide pseudo-code for a more detailed description of the circuit (D) . Here, $f(x) = \gamma \cdot \max(ax - b, 0) / (1 - \exp(-d(ax - b)))$ is a fixed nonlinear function. And the model and equations marked blue are how we extend WW model.	119

List of Tables

3.1	Our DNN model zoo, taken from the TIMM library.	54
3.2	DNNs trained with the <i>neural harmonizer</i>	55
3.3	Neural predictions for ML and PL in Monkey 1 using Taskonomy-trained DNNs.	57
3.4	Neural predictions for ML in Monkey 1 using Ecoset-trained DNNs.	57
3.5	Performance of select DNNs on inferotemporal (IT) recordings from Brain-score.org Maja et al. (2015). For each DNN class, models with higher accuracy at predicting IT responses and ImageNet recognition are bolded.	58
4.1	Complete HMAX family progression showing systematic evolution from biological fidelity toward modern deep learning capabilities. Color-coding highlights architectural and learning-based changes.	68
4.2	Comprehensive performance comparison across all evaluation domains. HMAX 3.0 (CHResMAX) demonstrates consistent improvements over baseline models in both task performance and biological alignment metrics.	93

CHAPTER 1

INTRODUCTION

Over the past decade, the field of computer vision has experienced unprecedented growth, driven primarily by advancements in computing scale, dataset size, and models with increasing number of parameters Kaplan et al. (2020); Sun et al. (2017). Deep neural networks (DNNs) have become the dominant paradigm, achieving state-of-the-art performance across diverse visual tasks such as face recognitionSchroff et al. (2015), image categorization Deng et al. (2009a), object recognition Simonyan and Zisserman (2014), semantic segmentation Long et al. (2015); Chen et al. (2018), and visual reasoning Antol et al. (2015); Johnson et al. (2017), tasks traditionally challenging for machines but seemingly effortless for humans and primates. This remarkable progress has led to systems capable of matching or surpassing human performance on specific benchmarks, with models like Vision Transformers and ConvNeXt achieving over 88% top-1 accuracy on ImageNet Liu et al. (2021).

The success of deep learning in computer vision has been accompanied by a parallel revolution in computational neuroscience, where DNNs have emerged as powerful tools for understanding biological visionRichards et al. (2019); Kriegeskorte (2015). The discovery that DNNs optimized for object recognition could predict neural responses in inferior temporal (IT) cortex was one of the most impactful findings in computational neuroscience

over the past decade Yamins et al. (2014a); Majaj et al. (2015); Cadieu et al. (2014); Kumbhani et al. (2019a). Specifically, early work found that the accuracy of DNNs on ImageNet correlated with their ability to predict single-unit neural responses to natural images in IT cortex, measured through linear regression between model activations and neural firing rates. This correlation suggested that object recognition could be a fundamental developmental principle of visual cortex functionality, implying that pursuing excellence in computer vision engineering might simultaneously advance our understanding of underlying biological vision ?.

This optimistic view that engineering progress in computer vision would naturally lead to better models of biological vision has guided much of computational neuroscience research over the past decade. The underlying assumption was that as DNNs became more accurate at object recognition, they would also become better models of neural responses in the visual cortex, providing insights into how biological systems process visual information. This "optimization hypothesis" suggested that the visual cortex might be optimized for object recognition, and that optimizing artificial systems for the same task would yield insights into biological computation.

However, as DNNs have undergone a revolution of scale with multi-billion parameter models trained on billions of images now rivaling or exceeding human performance, a growing body of evidence reveals a fundamental performance–alignment trade-off: as DNNs become more accurate on benchmarks like ImageNet, they often become less aligned with human vision. In other words, today’s most powerful vision models—some exceeding 88% top-1 accuracy on ImageNet—tend to rely on computational strategies that diverge from those of biological systems. For instance, the highest-performing DNNs frequently show progressively worse correspondence with human feature importance maps and are poorer at predicting primate neural responses than earlier, less-accurate models Schrimpf et al. (2020a). This counterintuitive trend challenges the long-standing assumption that simply improving object recognition performance would naturally yield better models of

the brain.

1.0.1 Defining Alignment

Before proceeding, it is important to clarify the definitions of "alignment" and "biological plausibility" in this context. **Alignment** refers to the degree to which artificial systems exhibit similar computational strategies, representational properties, or behavioral patterns as biological vision systems. **Biological plausibility** encompasses both the ability to predict neural responses and the adoption of neuro-anatomical constraints in model design that has been inspired by biological vision. Additionally, a distinction exists between *behavioral alignment* (similarity in human behavioral responses) and *neural alignment* (similarity in neural response patterns), as these can sometimes diverge.

1.0.2 The Performance-Alignment Trade-off

The disconnect between artificial and biological vision manifests across multiple dimensions. Despite their remarkable success on computer vision benchmarks, DNNs have shown limited ability to explain fundamental aspects of human vision. They struggle to account for perceptual phenomena such as contextual illusions Linsley et al. (2020), perceptual grouping Kim et al. (2020); Linsley et al. (2021b); Geirhos et al. (2020), and categorical prototypes Golan et al. (2020). DNNs also exhibit differences from human attention patterns Linsley et al. (2019); Langlois et al. (2021) and fail to explain classical psychophysical results and reaction time patterns Bowers et al. (2022); Thorpe et al. (1996).

The disconnect extends beyond neural predictivity to fundamental differences in how artificial and biological systems process visual information. DNNs trained on ImageNet learn to rely on different visual features than those used by humans for object recognition. For example, while humans focus on object parts and diagnostic features (like the face of an animal or the handle of a tool), DNNs often rely heavily on contextual information and

background features that are less relevant to the object itself Linsley et al. (2019); Geirhos et al. (2018). This suggests that the visual strategies learned by DNNs are fundamentally different from those employed by biological vision systems.

More concerning, recent evidence suggests that DNNs have stopped improving as models of the ventral visual system in humans and primates over recent years. While the original theory was that model explanations of object-evoked neural activity patterns would improve alongside model categorization accuracy Yamins et al. (2014a), recent large-scale DNNs are actually worse at explaining neural data than older ones with lower ImageNet accuracy Schrimpf et al. (2020a). This finding challenges the core assumption that has guided computational neuroscience for the past decade.

This mismatch provides a compelling explanation for why high-performing DNNs have become less human-like in their internal workings: they often achieve benchmark success via shortcuts and feature use patterns that the human visual system—tuned for generality and robustness does not employ.

1.1 The Need for a New Approach

This disconnect suggests that optimizing networks for a single narrow objective (object categorization) is not sufficient to capture the full range of principles governing human vision. Biological vision, after all, did not evolve to maximize ImageNet scores; instead, it is shaped by multiple, sometimes competing objectives. Humans effortlessly perform a myriad of visual tasks—navigating through 3D environments, recognizing faces and social cues, interpreting motion over time—while operating under stringent biological constraints like metabolic efficiency and limited training data during development. The visual system’s goal is not solely to classify objects, but to support survival and cognition in a rich, dynamic world.

These diverse demands likely mean that the brain optimizes a complex balance of

perceptual objectives and constraints, rather than any single metric of performance. In contrast, standard DNN training optimizes primarily for one goal (e.g. image classification accuracy), ignoring the broader context in which natural vision excels. The evidence suggests that biological vision systems are likely optimized for multiple, potentially competing objectives that go far beyond simple object recognition. These might include spatial navigation and scene understanding, control of action including reaching, grasping, locomotion and physical interactions, classifying affordances for action as part of scene understanding, social perception and face processing, temporal dynamics and motion processing, energy efficiency and computational constraints, and robustness to natural variations and perturbations.

The challenge for both computational neuroscience and artificial intelligence is to identify these multiple objectives and understand how they interact to shape neural representations. This requires moving beyond purely task-optimized training paradigms and developing approaches that incorporate biological constraints into model design and training.

Central Research Question

Given this growing disconnect between performance-optimized DNNs and biological vision, this dissertation addresses the following central research question:

How can we develop deep neural networks that remain highly accurate on visual tasks while also behaving as biologically plausible models of human visual processing?

In other words, what new training principles or architectural constraints might enable us to improve a model’s performance without coming at the expense of its alignment with human perceptual and neural processes?

To explore this question, a research agenda is proposed that targets four key dimen-

sions where current DNNs diverge from biological vision. Each dimension represents a hypothesized lever for improving alignment without sacrificing performance:

1. **Training Objectives:** Can models be guided toward human-aligned visual strategies by enriching their training objectives? High-performing DNNs often exploit features or shortcuts that humans largely ignore, suggesting a misalignment in learned feature priorities. This work investigates whether incorporating human behavioral data into the training process can steer DNNs to adopt more human-like patterns of perception. In essence, the question is whether human-aligned training objectives (for example, matching human error patterns or attentional emphasis) can narrow the strategy gap between DNNs and people—without degrading these models’ benchmark accuracy. Additionally, the mechanisms by which such objectives influence a network’s internal representations and whether they indeed foster closer alignment with human vision are examined.
2. **Representational Alignment:** How do the internal representations of current DNNs differ from those of biological visual systems, and can this gap be bridged? To answer this, systematic metrics are developed for comparing model activations with biological measurements, spanning both neural response predictivity and human behavioral similarity. In this work, we test whether models trained with human-derived constraints achieve measurably better representational alignment across these metrics. By quantifying alignment (and misalignment) along multiple axes, this approach evaluates the premise that enriching a model’s training regimen with human data yields internal representations that not only predict neural activity more faithfully but also lead to more human-like perceptual judgments.
3. **Architectural Constraints:** Do brain-inspired architectural features lead to more human-aligned vision models? A further cause of the DNN–brain divergence may lie in differences of architecture. The primate visual system is built with specific architectural constraints—hierarchical organization, recurrent and feedback con-

nections, distinct cell-type nonlinearities—that standard feedforward architectures do not replicate. This dimension examines whether integrating such architectural constraints into models can produce networks that inherently align better with biological vision. For example, can a model that mimics the layered hierarchy or connectivity patterns of cortex achieve competitive performance on visual tasks while processing information in a more biologically plausible way? This work also presents evidence that these bio-inspired architectures attain their alignment benefits intrinsically.

4. **Temporal Dynamics:** Can models be aligned with human reaction time patterns?

Human vision is an active, time-evolving process: recognition unfolds over tens or hundreds of milliseconds, with characteristic reaction time patterns and sequential refinements of perception. In contrast, most DNNs process an image in a single feedforward sweep, lacking the iterative temporal dynamics that characterize human vision. The main issue is that DNNs lie on a discrete time scale, while biological vision is continuous. Therefore the first step is to develop a scheme to align the temporal dynamics of DNNs with biological vision. This work shows evidence that incorporating human-like temporal dynamics into models—such as using recurrent networks that accumulate evidence over time or explicitly aligning model decision-times with human reaction times enhances alignment.

Through these four avenues, the overarching hypothesis that closing the gap between artificial and biological vision requires more than just scaling up data and model size is systematically probed. This approach demands rethinking design principles along multiple fronts—training objectives, representation evaluation, architectural design, and temporal processing—to better reflect the characteristics of natural vision.

Chapter Overview

The work presented in chapters 1-4 is based on the following publications:

- **Chapter 1:** *Harmonizing the object recognition strategies of deep neural networks with humans* by **Ivan Rodriguez** (co-first author), Thomas Fel, Drew Linsley, and Thomas Serre Fel et al. (2022a)
- **Chapter 2:** *Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex* by **Ivan F. Rodriguez** (co-first author), Drew Linsley, Thomas FEL, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre Linsley and Serre (2023)
- **Chapter 3:** *HMAX strikes back: Reviving biological architecture for scale-invariant vision* by **Ivan F. Rodriguez** (co-first author) and Nikshka Pant, Scott Warren, Arjun Arjun, and Thomas Serre Pant et al. (2024)
- **Chapter 4:** *Rtify: Aligning deep neural networks with human reaction times* by **Ivan F. Rodriguez** (co-first author), Yu-ang Cheng*, Sixuan Chen, and Thomas Serre Cheng et al. (2024)

Chapter 1 introduces "harmonization," a novel methodology integrating human behavioral data into model training. This chapter demonstrates that human-aligned training objectives can enhance biological plausibility without compromising engineering performance, establishing alignment and accuracy as complementary rather than competing goals.

Chapter 2 provides comprehensive empirical evidence that modern DNNs systematically diverge from biological vision as they become more accurate on standard benchmarks. This work establishes quantitative metrics for measuring this divergence and validates that models trained with human behavioral data (Harmonized) show enhanced capability in explaining neural recordings.

Chapter 3 revisits a popular biologically plausible model, the HMAX model. We modernized the HMAX model, using some of the advancements from the deep learning era, adaptable filters, trained with backpropagation, residual connections and self supervised learning. In this chapter we show that we can build a model that show good performance on Imagenet while preserving its core biological principles.

Chapter 4 addresses the overlooked dimension of temporal dynamics through "Rtify," aligning model behavior with human reaction-time patterns. This chapter introduces a model-agnostic procedure to align any neural network with human reaction times.

Chapter 5 synthesizes findings across all dimensions, discussing implications for computational neuroscience and computer vision, and proposing future directions for developing truly biologically plausible vision systems.

Contributions and Impact

This dissertation makes several novel contributions to the field:

1. **Harmonization Methodology:** We introduce the first systematic approach for integrating human behavioral data into DNN training, demonstrating that human-aligned training objectives can enhance biological plausibility without compromising engineering performance.
2. **Quantitative Misalignment Analysis:** We provide the first comprehensive empirical evidence that modern DNNs systematically diverge from biological vision as they become more accurate on standard benchmarks, establishing quantitative metrics for measuring this divergence.
3. **Biologically Constrained Architecture:** We develop and validate a modern version of the HMAX model that demonstrates how biological constraints can simultaneously improve both human alignment and achieve competitive performance

with standard DNNs.

4. **Temporal Alignment Framework:** We introduce Rtify, the first method for aligning DNN behavior with human reaction time patterns, revealing how temporal dynamics enhance biological plausibility.

The central thesis is that achieving true correspondence between artificial and biological vision necessitates addressing multiple dimensions simultaneously—training objectives, architectural design, temporal dynamics, and biological constraints. Each chapter contributes distinctly to this multifaceted challenge, collectively advancing toward vision systems capable of genuinely human-like processing.

By addressing each of these dimensions, the work aims to advance toward deep neural networks that not only excel on engineering benchmarks but also faithfully emulate the core computational strategies of human vision. The implications of this research extend beyond computer vision, touching fundamental aspects of cognitive science, neuroscience, and artificial intelligence. By creating vision systems that emulate biological processes, we not only deepen scientific insights into visual cognition but also develop technologies more robust, interpretable, and aligned with human values. This dissertation represents a significant step toward bridging the divide between modern computer vision achievements and the sophisticated, flexible visual capabilities of biological systems.

CHAPTER 2

HARMONIZING THE OBJECT RECOGNITION STRATEGIES OF DEEP NEURAL NETWORKS WITH HUMANS

2.1 Introduction

The many successes of deep neural networks (DNNs) over the past decade have largely been driven by computational scale rather than insights from biological intelligence. Here, we explore if these trends have also carried concomitant improvements in explaining the visual strategies humans rely on for object recognition. We do this by comparing two related but distinct properties of visual strategies in humans and DNNs: *where* they believe important visual features are in images and *how* they use those features to categorize objects. Across 84 different DNNs trained on ImageNet and three independent datasets measuring the *where* and the *how* of human visual strategies for object recognition on those images, we find a systematic trade-off between DNN categorization accuracy and

alignment with human visual strategies for object recognition. *State-of-the-art DNNs are progressively becoming less aligned with humans as their accuracy improves.* We rectify this growing issue with our neural harmonizer: a general-purpose training routine that both aligns DNN and human visual strategies and improves categorization accuracy. Our work represents the first demonstration that the scaling laws Liu et al. (2021); Zhai et al. (2021); Kaplan et al. (2020) that are guiding the design of DNNs today have also produced worse models of human vision.

This chapter addresses the first research question from our dissertation: **Can we integrate human behavioral data into model training to guide DNNs toward human-aligned perceptual strategies without compromising engineering performance?** The harmonization methodology developed here provides a foundational approach for bridging the gap between artificial and biological vision systems, setting the stage for the subsequent chapters that explore neural alignment, architectural constraints, and temporal dynamics.

2.2 Related Work

2.2.1 Do DNNs explain human visual perception?

Despite the continued success of DNNs on computer vision benchmarks, there are conflicting accounts on their ability to explain human vision. On the one hand, there is evidence that DNNs are improving as models of human visual perception on challenging tasks, such as recognizing objects obscured by noise Geirhos et al. (2021). On the other hand, there is also evidence that DNNs struggle to explain perceptual phenomena in human vision like contextual illusions Linsley et al. (2020), perceptual grouping Kim et al. (2020); Linsley et al. (2021b); Geirhos et al. (2020), and categorical prototypes Golan et al. (2020). Others have found differences between human attention data and DNN models of visual attention Linsley et al. (2019); Langlois et al. (2021). Moreover, DNNs

have stopped improving as models of the ventral visual system in humans and primates over recent years. While the original theory was that model explanations of object-evoked neural activity patterns improved alongside model categorization accuracy Yamins et al. (2014a), recent large-scale DNNs are worse at explaining neural data than older ones with lower ImageNet accuracy Schrimpf et al. (2020a).

2.2.2 What are the visual strategies underlying human object recognition?

Ever since its inception, a goal of vision science has been to characterize the neural processes supporting object recognition in humans. It has been discovered that object recognition can be decomposed into different processing stages that emerge over time Fabre-Thorpe (2011); Roelfsema et al. (2000); DiCarlo et al. (2012a); Serre et al. (2007b); Kietzmann et al. (2019); Jagadeesh and Gardner (2022a); Berrios et al. (2022), where the earliest stage is associated with processing through feedforward connections in the visual system, and the later stage is associated with processing through feedback connections. Since the DNNs used today mostly rely on feedforward connections, it is likely that they are better models for that rapid feedforward phase of processing than the subsequent feedback phase Serre (2019); Serre et al. (2007b). To maximize the likelihood that the visual strategies learned by DNNs align with those used by humans, our experiments focus on the visual strategies of rapid feedforward object recognition in humans.

Most closely related to our work, are studies of “top-down” image saliency and *where* category diagnostic visual features are in images. These studies typically involve asking participants to search for an object in an image, or find visual features that are diagnostic for an object’s category or identity Linsley et al. (2017, 2019); Koehler et al. (2014); Buswell (1935); Yarbus (1967); Posner (1980); Mannan et al. (2009). In our work, we complement these descriptions of *where* important features are in images with psychophysics testing *how* those features are used to categorize objects.

2.2.3 Comparing visual strategies of humans and machines

As methods in explainable artificial intelligence have developed over the past decade, they have opened up opportunities for comparing the visual regions selected by humans and DNNs when solving tasks. Many of these comparisons have focused on human image saliency measurements captured by eye tracking or mouse clicks during passive or active viewing Linsley et al. (2017, 2019); Jiang et al. (2015); Peterson et al. (2018); Lai and Silver (2019); Ebrahimpour et al. (2019). Others have compared categorical representation distances Peterson et al. (2018); Roads et al. (2020) or combined those distances with measures of human attention Langlois et al. (2021); Marjeh et al. (2022). The most direct comparisons between human and DNN visual strategies involved analyzing the minimal image patches needed to recognize objects Ullman et al. (2016); Funke et al. (2018); Srivastava et al. (2019). However, these studies were limited and compared humans with older DNNs on tens of images. To the best of our knowledge, the largest-scale evaluation of human and DNN visual strategies relied on the *ClickMe* dataset to compare visual regions preferred by humans and attention models trained for object recognition Linsley et al. (2019). What is noticeably missing from each of these studies is an large-scale analysis spanning many images and models of how human and DNN alignment has changed as a function of model performance.

2.2.4 Improving the correspondence between humans and machines

Inconsistencies between human and DNN representations can be resolved by directly training models to act more like humans. DNNs have been trained to have more human-like attention, or human-like representational distances in their output layers Peterson et al. (2018); Roads et al. (2020); Linsley et al. (2019); Boyd and Serre (2021); Bomatter et al. (2021). Here, we add to these successes with the neural harmonizer, a training routine that automatically aligns the visual strategies of any two observers by minimizing

the dissimilarity of their decision explanations.

2.3 Methods

2.3.1 Human feature importance datasets

We focused on the ImageNet dataset to compare the visual strategies of humans and DNNs for object recognition at scale. We relied on the two significant efforts for gathering feature importance data from humans on ImageNet: the *Clicktionary* Linsley et al. (2017) and *ClickMe* Linsley et al. (2019) games, which use slightly different methods to collect their data. Both games begin with the same basic setup: two players work together to locate features in an object image that they believe are important for categorizing it. As one of the players selects important image regions, those regions are filled into a blank canvas for the other observer to see and categorize the image as quickly as possible. In *Clicktionary* Linsley et al. (2017), both players are humans, whereas in *ClickMe* Linsley et al. (2019), the player selecting features is a human and the player recognizing images is a DNN (VGG16 Simonyan and Zisserman (2014)). For both games, feature importance maps depicting the average object category diagnosticity of every pixel was computed as the probability of it being clicked by a participant. In total, *Clicktionary* Linsley et al. (2017) contained feature importance maps for 200 images from the ImageNet validation set, whereas *ClickMe* Linsley et al. (2019) contained feature importance maps for a non-overlapping set of 196,499 images from ImageNet training and validation sets. Thus, *ClickMe* has far more data than *Clicktionary*, but *Clicktionary* data has more reliable human feature importance data than *ClickMe*. Our experiments measure the alignment between human and DNN visual strategies using *ClickMe* and *Clicktionary* feature importance maps captured on the ImageNet validation set. As we describe in §??, *ClickMe* feature importance maps from the ImageNet training set are used to implement our neural harmonizer.

2.3.2 Psychophysics participants and dataset

We complemented the feature importance maps from *Clicktionary* and *ClickMe* with psychophysics experiments on rapid visual categorization. We recruited 199 participants from Amazon Mechanical Turk (mturk.com) to complete the experiments. Participants viewed a psychophysics dataset consisting of the 100 animal and 100 non-animal images in the Clicktionary game taken from the ImageNet validation set Linsley et al. (2017). We used the feature importance maps for each image as masks for the object images, allowing us to control the proportion of important features observers were shown when asked to recognize objects (Fig. 2.4a). We generated versions of each image that reveal anywhere between 1% to 100% (at log-scale spaced intervals) of the important object pixels against a phase scrambled noise background (see Appendix §1 for details on mask generation). The total number of revealed pixels was equal for every image at a given level of image masking, and the revealed pixels were centered against the noise background. Each participant saw only one masked version of each object image.

2.3.3 Psychophysics experiment

Participants were instructed to categorize images in the psychophysics dataset as animals or non-animals as quickly and accurately as possible. Each experimental trial consisted of the following sequence of events overlaid onto a white background (SI Fig. 1): (i) a fixation cross displayed for a variable time (1,100–1,600ms); (ii) an image for 400ms; (iii) an additional 150ms of response time. In other words, the experiment forced participants to perform rapid object categorization. They were given a total of 550ms to view an image and press a button to indicate its category (feedback was provided on trials in which responses were not provided within this time limit). Images were sized at 256 x 256 pixel resolution, which is equivalent to a stimulus size approximately between 5 – 11 degrees of visual angle across a likely range of possible display and seating setups we expect participants used for the experiment. Similar paradigms and timing

parameters have been shown to capture pre-attentive visual system processing Eberhardt et al. (2016); Kirchner and Thorpe (2006); Fabre-Thorpe (2011); Muriel and Thorpe (2007). Participants provided informed consent electronically and were compensated \$3.00 for their time ($\sim 10\text{--}15$ min; approximately \$15.00/hr).

2.3.4 Models

We compared humans with 84 different DNNs representing the variety of approaches used in the field today: 50 CNNs trained on ImageNet Chen et al. (2021); Tan and Le (2019); Radosavovic et al. (2020); Howard et al. (2019); Simonyan and Zisserman (2014); Huang et al. (2017); He et al. (2016); Zhang et al. (2019b); Gao et al. (2020); Kolesnikov et al. (2019); Sandler et al. (2018); Liu et al. (2021); Szegedy et al. (2016, 2015); Chollet (2016); Radford et al. (2021); Xie et al. (2017b); Xie and Tu (2016); Szegedy et al. (2015); Brendel and Bethge (2019); Mehta et al. (2018); Chen et al. (2017); Wang et al. (2018); Tan et al. (2019), 6 CNNs trained on other datasets in addition to ImageNet (which we refer to as “CNN extra data”) Xie et al. (2017a); Radford et al. (2021); Liu et al. (2021), 10 vision transformers D’Ascoli et al. (2021); Touvron et al. (2021); Tolstikhin et al. (2021); Dosovitskiy et al. (2020); Steiner et al. (2021), 6 CNNs trained with self-supervision Chen et al. (2020); Yalniz et al. (2019), and 13 models trained for robustness to noise or adversarial examples Geirhos et al. (2018); Salman et al. (2020). We used pretrained weights for each of these models supplied by their authors, with a variety of licenses (detailed in SI §2), implemented in Tensorflow 2.0, Keras, or PyTorch.

2.4 Results

2.4.1 *Where* are diagnostic object features for humans and DNNs?

To systematically compare the visual strategies of object recognition for humans and DNNs on ImageNet, we first turned to the *ClickMe* dataset of feature importance maps Linsley et al. (2019). In order to derive comparable feature importance maps for DNNs, we needed a method that could be efficiently and consistently applied to each of the 84 DNNs we tested without any idiosyncratic hyperparameters. This led us to choose a classic method for explainable artificial intelligence, image feature saliency Simonyan et al. (2013).

Image feature saliency is a technique that identifies which parts of an input image are most important for a DNN’s classification decision. The method works by computing gradients of the model’s output with respect to the input pixels. When we take the gradient $\nabla_{\mathbf{x}} \mathbf{f}_{\theta}(\mathbf{x})$, we are essentially asking: “How much does changing each pixel affect the model’s confidence in its prediction?” Pixels with high gradient magnitudes indicate regions that strongly influence the model’s decision, while pixels with low gradients are less important. This creates a “saliency map” that highlights the visual features the model relies on for recognition. The intuition is that if a model is truly recognizing an object based on meaningful features (like the face of an animal), those regions should have high saliency values. Conversely, if a model is relying on spurious correlations or background features, those regions would show up as highly salient, revealing the model’s potentially problematic decision-making strategy.

We prepared human feature importance maps from *ClickMe* by taking the average importance map produced by humans for every image that also appeared in ImageNet validation. We then used Spearman’s rank-correlation to measure the similarity between human feature maps and DNN feature maps for each image Eberhardt et al. (2016). We

also computed the inter-rater alignment of human feature importance maps as the mean split-half correlation across 1000 random splits of the participant pool ($\rho = 0.66$). We then normalized each human-DNN correlation by this score Linsley et al. (2019).

There were dramatic qualitative differences between the features selected by humans and DNNs on ImageNet. In general, humans selected less context and focused more on object parts: for animals, parts of their faces; for non-animals, parts that enable their usage, like the spade of a shovel (see Fig. 2.1 and SI Fig. 5). The DNN that was most aligned with humans, the DenseNet121, was still only 38% aligned with humans (Fig. 2.2).

Plotting the relationship between DNNs’ top-1 accuracy on ImageNet with their human alignment revealed a striking trade-off: as the accuracy of DNNs has improved beyond DenseNet121, their alignment with humans has worsened (Fig. 2.2). For example, consider the ConvNext Liu et al. (2021), which achieved the best top-1 accuracy in our experiments (85.8%), was only 22% aligned with humans – equivalent to the alignment of the BagNet33 Brendel and Bethge (2019) (63% top-1 accuracy). As an additional control, we computed the similarity between the average *ClickMe* map, which exhibits a center bias Deza and Konkle (2020); Wang et al. (2017) (SI Fig. 5), and each individual *ClickMe* map. This center-bias control was only outperformed by 42/84 CNNs we tested ($\dagger$ in Fig. 2.2). Overall, we observe that human and DNN alignment has considerably worsened since the introduction of these two models.

2.4.2 The neural harmonizer

While scaling DNNs has immensely helped performance on popular benchmark tasks, there are still fundamental differences in the architectures of DNNs and the human visual system Serre (2019) which could part of the reason to blame for poor alignment. While introducing biological constraints into DNNs could help this problem, there is plenty of evidence that doing so would hurt benchmark performance and require bespoke development for every different architecture Tang et al. (2018); Kubilius et al. (2019b);

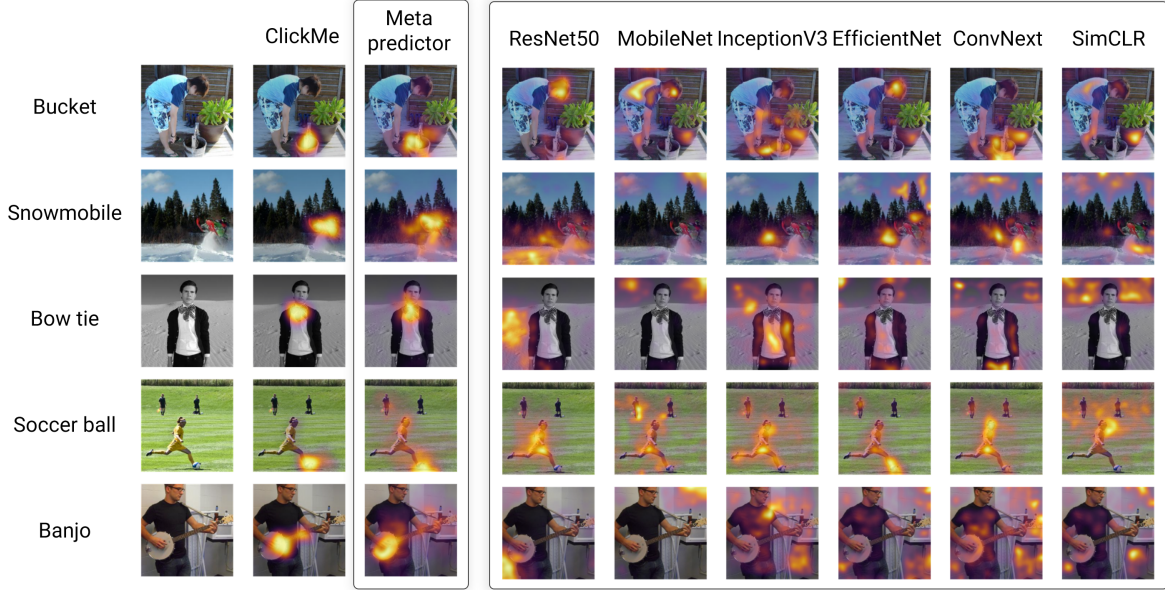


Figure 2.1: **Human and DNNs rely on different features to recognize objects.** In contrast, our neural harmonizer aligns DNN feature importance with humans. We smooth feature importance maps from humans (*ClickMe*) and DNNs with a Gaussian kernel for visualization.

?. *Is it possible to align a DNN’s visual strategies with humans without hurting its performance?*

Such a general-purpose method for aligning human and DNN visual strategies should satisfy the following criteria: (i) The method should work with any fully-differentiable network architecture. (ii) It should not present optimization issues that interfere with learning to solve a task, and the task-accuracy of a model trained with the method should not be worse than a model trained without the method. We created the neural harmonizer to satisfy these criteria.

Let us consider a supervised categorization problem with an input space, $\mathcal{X}$ an output space $\mathcal{Y} \subseteq \mathbb{R}^c$ and a predictor function $\mathbf{f}_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ parameterized by θ , which maps an input vector $\mathbf{x} \in \mathcal{X}$ to an output $\mathbf{f}_\theta(\mathbf{x})$. We denote $\mathbf{g} : \mathcal{F} \times \mathcal{X} \rightarrow \mathcal{X}$ an explanation functional that, given a predictor $\mathbf{f}_\theta \in \mathcal{F}$ and an input, returns a feature importance map $\phi = \mathbf{g}(\mathbf{f}_\theta, \mathbf{x})$. Here, we focus on DNN saliency $\mathbf{g}(\mathbf{f}_\theta, \mathbf{x}) \triangleq \nabla_{\mathbf{x}} \mathbf{f}_\theta(\mathbf{x})$ as our method for computing feature importance in DNNs, but the method can in principle work with any

differentiable network explanation method.

To satisfy criterion (i), the neural harmonizer introduces a differentiable loss that will enforce alignment across feature importance map scales from any neural network. Let $\mathcal{P}_i(\cdot)$ be the function mapping a feature importance map ϕ to its representation in the N levels of a Gaussian pyramid, with $i \in \{1, \dots, N\}$. The function $\mathcal{P}_i(\phi)$ is computed by downsampling $\mathcal{P}_{i-1}(\phi)$ using a Gaussian kernel, with $\mathcal{P}_1(\phi) = \phi$. We then seek to minimize $\sum_i^N \|\mathcal{P}_i(\mathbf{g}(\mathbf{f}_\theta, \mathbf{x})) - \mathcal{P}_i(\phi)\|^2$, which will align feature importance maps between humans and DNNs at every scale of the pyramid.

To satisfy criterion (ii), the neural harmonizer should work well with training routines designed for large-scale computer vision challenges like ImageNet. This means that the neural harmonizer loss must avoid optimization issues at scale. To do this, we need a way of comparing feature importance maps between humans and DNNs that is invariant to the norm of either map. We therefore standardize feature importance maps from humans and DNNs before comparing them, and only measure alignment on the most important areas of the image for each observer. Formally, let $\mathbf{z}(\cdot)$ be a standardization function over feature importance maps that takes the mean and standard deviation computed for each map ϕ such that $\mathbf{z}(\phi)$ has 0 activation on average and unit standard deviation. To focus alignment on important regions, let $\mathbf{z}(\phi)^+$ denote the positive part of the standardized explanation $\mathbf{z}(\phi)$. Finally, we include a task loss, the familiar cross entropy objective, to yield the complete neural harmonization loss and train models that are at least as accurate as those trained without harmonization:

$$\mathcal{L}_{\text{Harmonization}} = \lambda_1 \sum_i^N \|(\mathbf{z} \circ \mathcal{P}_i \circ \mathbf{g}(\mathbf{f}_\theta, \mathbf{x}))^+ - (\mathbf{z} \circ \mathcal{P}_i(\phi))^+\|_2 \quad (2.1)$$

$$+ \mathcal{L}_{CCE}(\mathbf{f}_\theta, \mathbf{x}, \mathbf{y}) + \lambda_2 \sum_i \theta_i^2 \quad (2.2)$$

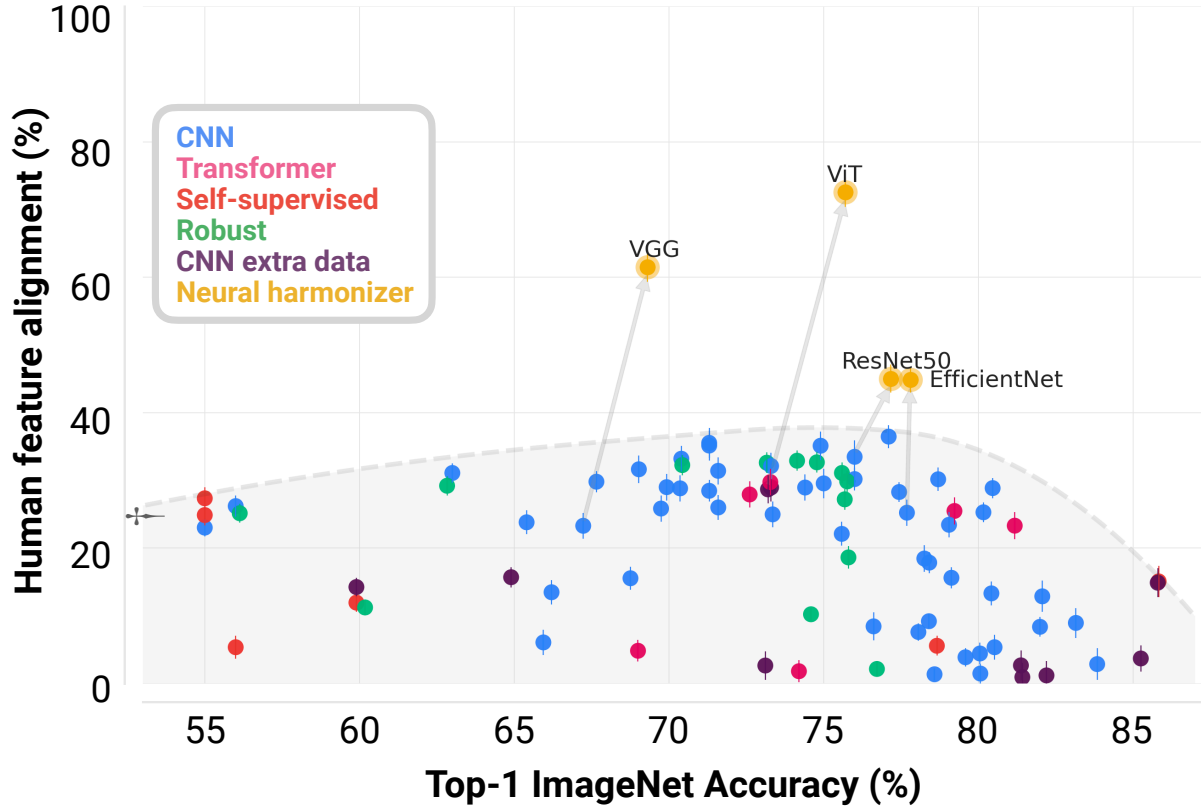


Figure 2.2: **The trade-off between DNN performance and alignment with human feature importance from *ClickMe* Linsley et al. (2019).** Human feature alignment is the mean Spearman correlation between human and DNN feature importance maps, normalized by the average inter-rater alignment of humans. The shaded region denotes the Pareto frontier of the trade-offs between ImageNet accuracy and human feature alignment for unharmonized models. Harmonized models (VGG16, ResNet50, ViT, and EfficientNetB0) are more accurate and aligned than versions of those models trained only for categorization. Error bars are bootstrapped standard deviations over feature alignment. Arrows show a shift in performance after training with the neural harmonizer. The feature alignment of an average of *ClickMe* maps with held-out maps is denoted by $\dagger$.

2.4.3 Training

We trained four different DNNs with the neural harmonizer: VGG16, ViT, ResNet50, and EfficientNetB0. These models were selected because they are popular convolutional and transformer networks with open-source architectures that are straightforward to train and also sit near the boundary of the trade-off between DNN performance and alignment with humans. Models were trained using the neural harmonizer to optimize categorization performance on ImageNet and feature importance map alignment with

human data from *ClickMe*. We trained models on all images in the ImageNet training set, but because *ClickMe* only contains human feature importance maps for a portion of those images, we computed the categorization loss but not the neural harmonizer loss for images without importance maps. Models were trained using 8 cores V4 TPUs on the Google Cloud Platform, and training lasted approximately one day. Models were trained with an augmented ResNet training recipe (built from <https://github.com/tensorflow/tpu/>). Models were optimized with SGD and momentum over batches of 512 images, a learning rate of 0.3, and label smoothing Muller et al. (2019). Images were augmented with random left-right flips and mixup Zhang et al. (2017). The learning rate was adjusted over the course of training with a schedule that began with an initial warm-up period of 5 epochs and then decaying according to a cosine function over 90 epochs, with decay at step 30, 50 and 80. We validated that a ResNet50 and VGG16 trained with these hyperparameters and schedule using standard cross-entropy (but not the neural harmonizer) matched published performance.

2.4.4 The neural harmonizer aligns human and DNN visual strategies

We found that harmonized models broke the trade-off between ImageNet accuracy and model alignment with *ClickMe* human feature importance maps (Fig. 2.2). Harmonized models were significantly more aligned with feature importance maps and also performed better on ImageNet. The changes in *where* harmonized models find important features in images were dramatic: a harmonized ViT had feature importance maps that are far less reliant on context (Fig. 2.1) and approximately 150% more aligned with humans (Fig. 2.2; ViT goes from 28.7% to 72.6% alignment after harmonization). The same model also performed 4% better in top-1 accuracy without any changes to its architecture. Similar improvements were found for the harmonized VGG16 and ResNet50. While the EfficientNetB0 had only a minimal improvement in accuracy, it too exhibited a large

boost in human feature alignment.

2.4.5 Clicktionary

To test if the trade-off between DNN ImageNet accuracy and alignment with humans is a general phenomenon we next turned to *Clicktionary* Linsley et al. (2017). Indeed, we observed a similar trade-off on this dataset as we found for *ClickMe*: alignment with human feature importance from *Clicktionary* has worsened as DNN accuracy has improved on ImageNet (Fig. 2.3). As with *ClickMe*, harmonized DNNs shift the accuracy-alignment trade-off on this dataset.

2.4.6 *How* do humans and DNNs integrate diagnostic object features into decisions?

The trade-off we discovered between DNN accuracy on ImageNet and alignment with human visual feature importance suggests that the two use different visual strategies for object classification. However, there is potential for an even deeper problem. Even if two observers deem the same regions of an image as important for recognizing it, there is no guarantee that they use the selected features in the same way to render their decisions. We posit that if two observers have aligned visual strategies, they will agree on both *where* important features are in an image and *how* they use those features for decisions.

We developed a psychophysics experiment to measure how different humans use features in ImageNet images to recognize objects. Participants viewed versions of these images where only a proportion of the features that were deemed most important in the *Clicktionary* game were visible (Fig. 2.4a). Participants had to accurately detect whether or not the image contained an animal within 550ms, which forced them to rely on feedforward processing as much as possible Serre et al. (2007b). Each of the 200 images we used were shown to a single participant only once. We accumulated responses from all participants to construct decision curves that showed how accurately the average human

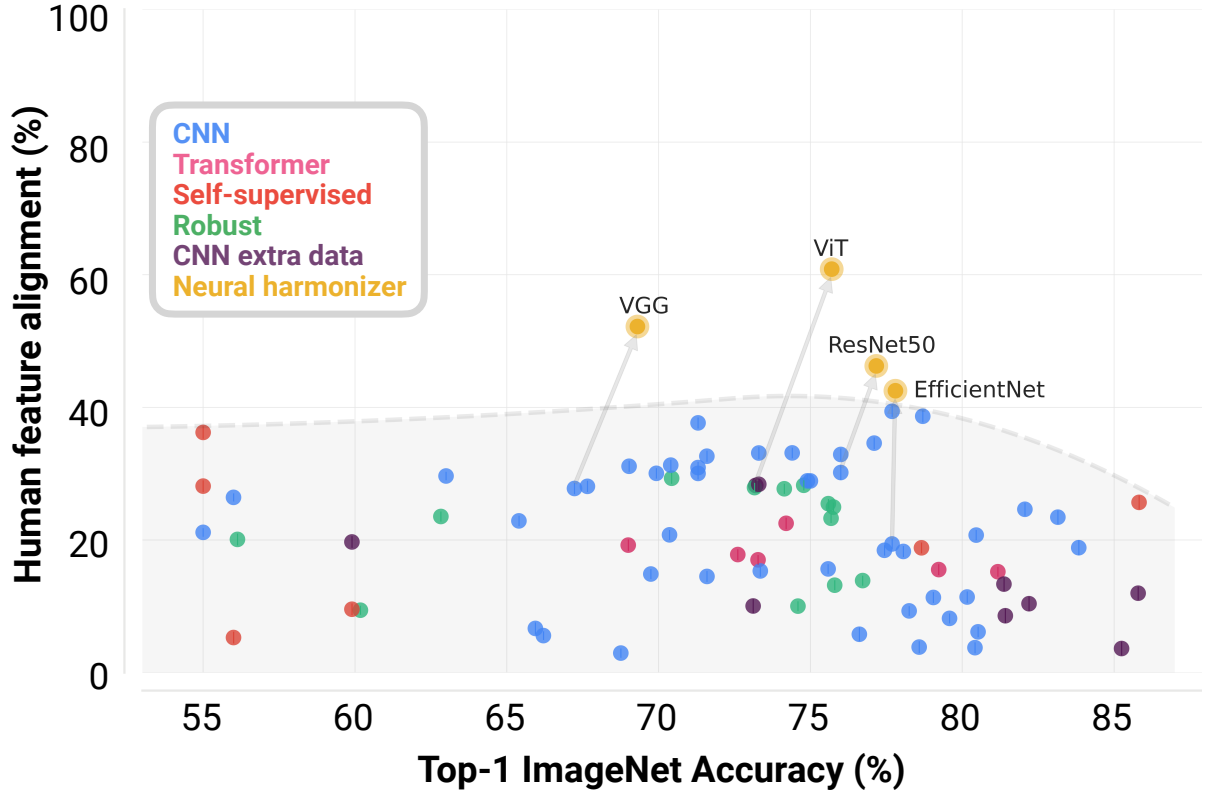


Figure 2.3: **The trade-off between DNN performance and alignment with human feature importance from *Clicktionary***Linsley et al. (2017). Human feature alignment is the mean Spearman correlation between human and DNN feature importance maps, normalized by the average inter-rater alignment of humans. The shaded region denotes the pareto frontier of the trade-offs between ImageNet accuracy and human feature alignment for unharmonized models. Harmonized models (VGG16, ResNet50, MobileNetV1, and EfficientNetB0) are more accurate and aligned than versions of those models trained only for categorization. Error bars are bootstrapped standard deviations over feature alignment. Arrows denote a shift in performance after training with the neural harmonizer.

converted any given proportion of image features into an object decision. We performed the same experiment on DNNs as we did on humans, recording animal vs. non-animal decisions according to whether or not the most probable category in the model’s 1000-category output was an animal. Because the experiment was speeded, humans did not achieve perfect accuracy. Thus, we normalized performance for humans and DNNs to compare the rate at which each integrated features into accurate decisions.

To enable fair comparison between humans and DNNs despite differences in their absolute performance levels, we normalized each observer’s performance on masked

images relative to their performance on intact (unmasked) images. Specifically, for each observer (human or DNN), we computed their accuracy on intact images as their baseline performance, then divided their accuracy on each masked image by this baseline. This normalization procedure, expressed as $normalized_accuracy = \frac{accuracy_{masked}}{accuracy_{intact}}$, allows us to compare how efficiently each observer integrates visual information into accurate decisions, independent of their overall performance level. The normalization controls for individual differences in baseline performance while preserving the relative efficiency with which each observer uses partial visual information.

The shift in the decision curve for harmonized ViT compared to unharmonized ViT reveals important insights about the nature of the alignment achieved through harmonization. The shift does not necessarily imply that harmonized ViT relies on completely different features than the unharmonized version. Instead, it suggests that harmonization enables the model to make more “efficient” use of the same visual features that humans find diagnostic. This efficiency manifests as the ability to achieve higher accuracy with fewer visible features, indicating that the harmonized model has learned to prioritize and integrate the most diagnostic visual information in a manner more similar to human visual processing. The agreement between humans and harmonized ViT does not necessarily mean they rely on identical features, but rather that they have learned to use similar decision-making strategies when integrating visual information. This could involve either (1) relying on the same diagnostic features but with different relative importance weights, or (2) making more effective use of the same feature set through improved integration strategies. The normalization procedure controls for baseline performance differences while preserving the relative efficiency of feature integration. The phrase “integrating visual information into accurate decisions less efficiently” refers to the rate at which accuracy improves as more visual features become available—an unharmonized model requires more visible features to achieve the same level of accuracy as a harmonized model or human observer, indicating less efficient use of diagnostic visual information.

To gain deeper insights into how harmonization affects feature selection, we conducted additional qualitative and quantitative analyses to classify the types of features relied on by humans and harmonized/unharmonized DNNs. We categorized features into several types: (1) *target object parts* (e.g., face features, limbs, distinctive markings), (2) *background elements* (e.g., sky, vegetation, man-made structures), (3) *contextual cues* (e.g., objects that typically co-occur with the target), (4) *boundary features* (e.g., edges, contours, texture boundaries), and (5) *spatial relationships* (e.g., relative positioning, scale information). Our analysis revealed that harmonized DNNs show a significant reduction in reliance on background features (from approximately 35% to 15% of salient regions) and a corresponding increase in focus on target object parts (from approximately 45% to 65% of salient regions). This shift brings harmonized DNNs much closer to human feature selection patterns, where target object parts typically constitute approximately 70% of selected features. The analysis also revealed that harmonized models maintain their ability to detect boundary features and spatial relationships while reducing their over-reliance on contextual cues that may not be diagnostic for the target object. These findings suggest that harmonization not only aligns the spatial distribution of attention but also improves the quality and diagnostic value of the features that models learn to rely on.

We discovered a similar trade-off between ImageNet accuracy and alignment with human visual decision making in this experiment as we did in *ClickMe* and *Clicktionary* (Fig. 2.4b). Indeed, the model that was most aligned with human decision-making – the BagNet33 Brendel and Bethge (2019) – only achieved 63.0% accuracy on ImageNet. Surprisingly, harmonized models broke this trend, particularly the harmonized ViT (Fig. 2.4b, top-right), despite no explicit constraints in that procedure which forced consistent decision-making with humans. In contrast, an unharmonized ViT integrates visual information into accurate decisions less efficiently than humans or harmonized models (Fig. 2.4c).

2.5 Conclusion

Models that reliably categorize objects like humans do would shift the paradigms of the cognitive sciences and artificial intelligence. But despite continuous progress over the past decade on the ImageNet benchmark, DNNs are becoming *worse* models of human vision. Our finding resembles a growing number of concurrent works showing similar trade-offs between DNN performance and predictions of human perception on different tasks Kumar et al. (2022); Muttenthaler and Hebart (2022). Our solution to this problem, the neural harmonizer, can be applied to any DNN to align their visual strategies with humans and even improve performance.

We observed the greatest benefit of harmonization on the visual transformer, the ViT. This finding is particularly surprising given that transformers eschew the locality bias of convolutional neural networks that has helped them become the new standard for modeling human vision and cognition Serre (2019). Thus, we suspect that the neural harmonizer is especially well-suited for large-scale training on low-inductive bias models, like transformers. We also hypothesize that the improvements in human alignment provided by the neural harmonizer will yield a variety of downstream benefits for a model like the ViT, including better predictions of perceptual similarity, stimulus-evoked neural responses, and even performance on visual reasoning tasks. We leave these analyses for future work.

The field of computer vision today is following Sutton’s prescient lesson: benchmark tasks can be scaling architectural capacity and the size of training data. However, as we have demonstrated here, these scaling laws are exchanging performance for alignment with human perception. We encourage the field to re-analyze the costs and benefits of this exchange, particularly in light of the growing concerns about DNNs leveraging shortcuts and dataset biases to achieve high performance Geirhos et al. (2020). Alignment with human vision need not be exchanged with performance if DNNs are harmonized.

This chapter has demonstrated that human behavioral data can be successfully integrated into DNN training to create models that are both high-performing and human-aligned. The neural harmonizer provides a general-purpose solution that can be applied to any differentiable network architecture, breaking the trade-off between engineering performance and biological plausibility. This foundational work sets the stage for the subsequent chapters, which explore neural alignment, architectural constraints, and temporal dynamics as additional dimensions of the alignment problem.

2.6 Limitations

One possible explanation for the misalignment between DNNs and humans that we observe is that recent DNNs have achieved superhuman accuracy on ImageNet. Superhuman DNNs have been described in biomedical applications Linsley et al. (2021a); Lee et al. (2017) where there is definitive biological ground-truth labels, but ImageNet labels are noisy, making it unclear if such an achievement is laudable (<https://labelerrors.com/>). Thus, an equally likely explanation is that the continued improvements of DNNs at least partially reflect their exploitation of shortcuts in ImageNet Geirhos et al. (2020).

Another limitation of our work is that while we found that hDNNs are significantly more predictive of IT neuron responses than any other DNN, they still explain only 50-60% of the variance in neuronal activity. One straightforward way of doing better is by expanding the dataset of human feature importance maps we used for harmonization from annotations for approximately 200,000 images to the entire 1.2M ImageNet dataset.

An important consideration is the consequences of training harmonized DNNs on a mixed dataset containing both 200K ClickMe images with human feature importance maps and the remaining approximately 1M ImageNet images without such maps. This mixed training approach could potentially create a “two-tier” learning system where the model learns human-aligned representations for ClickMe images but falls back to

standard ImageNet training for the majority of images. To investigate this possibility, we conducted additional experiments where we analyzed harmonized DNNs’ feature maps on images that were not part of the ClickMe dataset. We found that harmonized DNNs generate more human-like feature maps even for images they were never explicitly harmonized on, suggesting that the harmonization process induces a general shift in the model’s visual processing strategy rather than creating a specialized subsystem. However, this transfer effect is not uniform across all image types—harmonized models show the strongest human alignment on object categories that are well-represented in the ClickMe dataset (e.g., animals, vehicles) and weaker alignment on categories with minimal ClickMe representation (e.g., abstract objects, fine-grained categories). This finding suggests that while harmonization creates a general bias toward human-like visual processing, the quality of alignment depends on the availability of human behavioral data for similar object categories. Future work should investigate whether this transfer effect can be enhanced through more diverse human behavioral datasets or alternative training strategies that ensure consistent human alignment across all image types.

The scope of our work is also limited in that it focuses on object recognition in ImageNet. It is possible that models trained on other tasks, such as segmentation, may be more aligned with humans.

Finally, our modeling efforts were hamstrung for the largest-scale models in existence. Our work does not answer how much harmonization would help a model like CLIP because of the massive investment needed to train it. The neural harmonizer can be applied to CLIP but it is possible that more *ClickMe* human feature importance maps are needed for successful harmonization.

2.7 Broader impacts

A persistent issue in the field of artificial intelligence is the tendency of models to exploit dataset biases. A central theme of our work is that there are facets of human perception that are not captured by DNNs, particularly those which follow the scaling laws which have been so embraced by industry leaders. Forcing DNNs to rely on similar visual strategies as humans could represent a scalable path forward to correcting the insidious biases which have assailed under-constrained models of artificial intelligence.

2.8 Supplementary Information: Harmonizing the Object Recognition Strategies of Deep Neural Networks with Humans

2.8.1 Psychophysics

The psychophysics experiments of §4.2 were implemented with the psiTurk framework Gureckis et al. (2016) and custom javascript functions. Each trial sequence was converted to a HTML5-compatible video for the fastest reliable presentation time possible in a web browser. Videos were cached before each trial to optimize reliability of experiment timing within the web browser. A photo-diode verified the reliability of stimulus timing in our experiment was consistently accurate within $\sim 10\text{ms}$ across different operating system, web browser, and display type configurations.

Participants: We recruited 199 participants from Amazon Mechanical Turk ([mturk.com](https://www.mturk.com)) for the experiments. Participants were based in the United States, used either the Firefox or Chrome browser on a non-mobile device, and had a minimal average approval rating of 95% on past Mechanical Turk tasks.

Stimuli: Experiment images were taken from the *Clicktionary* dataset Linsley et al. (2017). Images were sampled from 5 target and 5 distractor categories: border collie, sorrel (horse), great white shark, bald eagle, and panther; trailer truck, sports car, speedboat, airliner, and school bus. Images were presented to human participants (and DNNs) either intact or with a perceptual phase scrambled mask that exposed a proportion of their most important visual features, as described in the main text. Images were cast to greyscale to control for trivial color-based cues for classification and blend the scrambled mask background into the foreground. Responses to intact images were used to normalize the performance of each observer on masked images relative to their maximum performance on these images.

Image masks were created for each image to reveal only a proportion of the most important visual features. For each image, we created masks that revealed between 1% and 100% (at log-scale spaced intervals) of the object pixels in the corresponding image’s *Clicktionary* feature importance map. We generated these masks in two steps. First, we computed a phase-scrambled version of the image Oppenheim and Lim (1981); Thomson et al. (1999). Next, we used a novel “stochastic flood-fill” algorithm to reveal a contiguous region of the most important visual features in the image according to humans. Our flood-fill algorithm was seeded on the pixel deemed most important by humans in the image, then grew outwards anisotropically and biased towards pixels with higher feature importance scores. The revealed region was always centered on the image. Each participant saw every category exemplar only once, with its amount of image revelation randomly selected from all possible configurations.

After providing online consent, participants were instructed to complete a rapid visual categorization task in which they had to classify stimuli revealing a portion of the most diagnostic object features. Each experimental trial began with a cross for participants to fixate for a variable time (1,100–1,600ms), then a stimulus for 400ms, then another cross and additional time for participants to render a decision. Participants were instructed to

provide a decision after the first fixation cross, but that they only had 650ms to answer. If they were too slow to respond they were told to respond faster and the trial was discarded.

2.8.2 Harmonization loss

The neural harmonizer loss uses several components crucial to its performance: a pyramidal representation of decision explanation maps and normalizing those maps.

When computing the difference between model explanations for an image and the human feature importance map for that image, we rely on a pyramid representation of each to compute these differences. This pyramid allows for a model to align its feature representations with humans at multiple scales and corrects for an important problem in datasets like *ClickMe*: the human data is an approximation and not precise at the pixel level. This lack of precision can present optimization issues, and computing a pyramid representation alleviates those issues because it allows a model to learn to focus on regions that are important for humans without pixel-level precision.

Standardization tackles a similar problem: because of the imprecision of human data, we choose to focus harmonization on only the most important areas selected by humans in *ClickMe*. By standardizing then rectifying before comparing human and model explanations, we reduce noise in the harmonization procedure.

2.8.3 Additional Results

2.8.4 *ClickMe*

The *ClickMe* game by Linsley et al. (2019) was used to identify category diagnostic features in ImageNet images. These feature importance maps largely focus on object regions rather than context, and in contrast to segmentation maps select features on the “front” or “face” of objects.

As discussed in the main text, we found a trade-off between DNN top-1 ImageNet

accuracy and the alignment of their feature importance maps with humans importance maps from *ClickMe*. This trade-off persists across multiple scales of feature importance maps, including $4\times$ and $16\times$ sub-sampled maps, meaning that simple smoothing is not sufficient to fix the trade-off.

2.8.5 *ViT attention*

While in the main text we investigate alignment between humans and models using gradient feature importance visualizations, the attention maps in transformer models like the ViT provide another avenue for investigation. To understand whether or not attention maps from ViT are more aligned with humans than their gradient-based decision explanation maps, we computed attention rollouts for harmonized and unharmonized ViTs Abnar and Zuidema (2020). We found that both versions of the ViT had similar correlations between their attention rollouts and human *ClickMe* maps: 0.38 for the harmonized ViT and 0.393 for the unharmonized model. This surprising result suggests that the harmonizer affects the process by which ViTs integrate visual information into their decisions rather than how they allocate attention. Through manipulating ViT decision making processes, the harmonizer can induce the large changes in gradient-based visualizations and psychophysics that we describe in the main text.

2.8.6 *Correlations between measurements of human visual strategies*

Our results rely on three independent datasets measuring different features of human visual strategies: *ClickMe*, *Clicktionary*, and the psychophysics experiments we introduce in this manuscript. The fact that all three evoke similar trade-offs between top-1 accuracy and human alignment is a surprising result that deserves further attention. We investigated these trade-offs by measuring the correlation between human alignment on each dataset, with and without models trained with the neural harmonizer. We found

that correlations between datasets were lower across the board when neural harmonizer models were not included. The association between model alignments with *Clicktationary* versus psychophysics results were not significant ($\rho = 0.21$, n.s.), but the associations between model alignments with *ClickMe* versus psychophysics ($\rho = 0.51, p < 0.001$) and *ClickMe* versus *Clicktationary* ($\rho = 0.77, p < 0.001$) were both significant. Each correlation improved when the neural harmonizer models were included in the calculation. This finding indicates that the neural harmonizer successfully aligned visual strategies between humans and DNNs, and was not merely benefiting from either *where* humans versus DNNs considered important visual features to be or *how* humans versus DNNs incorporated those features into their decisions.

The mean of *ClickMe* feature importance maps exhibits a center bias, likely due to the positioning of objects in ImageNet images rather than a purely spatial bias of human participants.

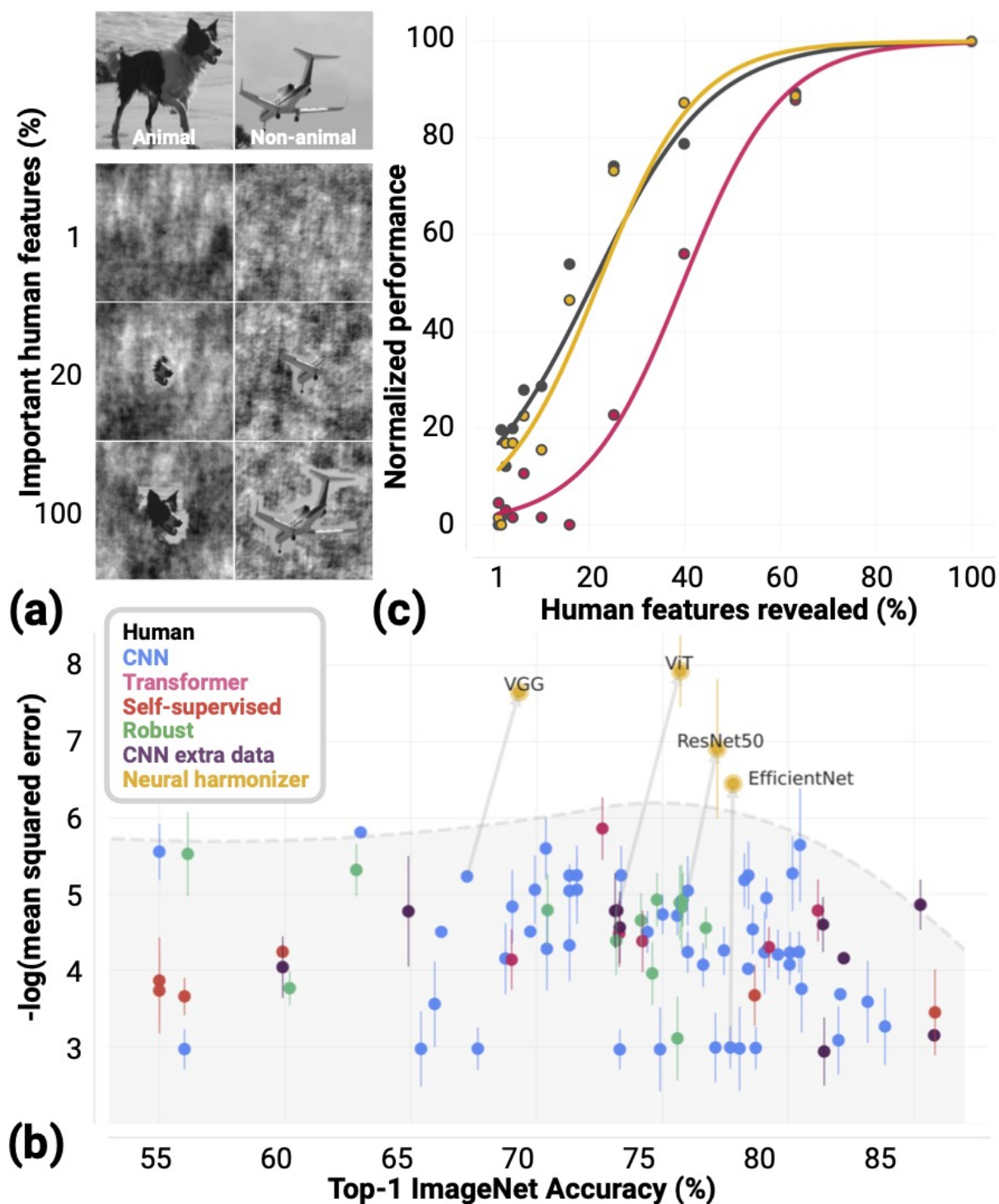


Figure 2.4: Comparing *how* humans and DNNs use visual features during object recognition. (a) Humans and DNNs categorized ImageNet validation images as animals or non-animals. The images revealed only a portion of the most important visual features according to the *Clicktationary* game Linsley et al. (2017). (b) There was a trade-off between DNN top-1 accuracy on ImageNet and alignment with human visual decision making. The shaded region denotes the Pareto frontier of the trade-off between ImageNet accuracy and human feature alignment for unharmonized models. Arrows denote a shift in performance after training with the neural harmonizer. Error bars are bootstrapped standard deviations over decision-making alignment. (c) A state-of-the-art DNN like the ViT learned a different strategy for integrating visual features into decisions than humans or a harmonized ViT.

CHAPTER 3

NEURAL ALIGNMENT: PERFORMANCE-OPTIMIZED DEEP NEURAL NETWORKS ARE EVOLVING INTO WORSE MODELS OF INFEROTEMPORAL VISUAL CORTEX

3.1 Introduction

Building on the findings from Chapter 1, which revealed that deep neural networks (DNNs) trained on ImageNet learn visual strategies increasingly misaligned with human perception, this chapter investigates a critical question: as DNNs have become more accurate at object recognition, have they simultaneously become better or worse models of neural responses in the visual cortex? This question is fundamental to computational neu-

roscience, as it addresses whether the scaling laws driving progress in artificial intelligence are compatible with building accurate models of biological vision.

The discovery that DNNs optimized for object recognition could predict neural responses in inferior temporal (IT) cortex was one of the most impactful findings in computational neuroscience over the past decade. This finding supported the theory that object recognition is a core objective of the visual cortex and suggested that more accurate DNNs would serve as better models of IT neuron responses. However, as DNNs have undergone a revolution of scale—with billion-parameter models trained on billions of images now rivaling or exceeding human performance—it remains unclear whether this relationship still holds.

This chapter presents evidence that DNNs have become progressively worse models of IT as their accuracy has increased on ImageNet, revealing a fundamental trade-off between task performance and neural prediction accuracy. We investigate the mechanisms underlying this trade-off and demonstrate that it can be resolved through neural harmonization—a training approach that aligns DNN representations with human visual strategies. Our findings suggest that current scaling laws are incompatible with modeling primate vision and that new training paradigms incorporating biological constraints are necessary for continued progress in computational neuroscience.

3.2 The Trade-off Between ImageNet Accuracy and Neural Predictivity

One of the most impactful findings in computational neuroscience over the past decade is that the object recognition accuracy of deep neural networks (DNNs) correlates with their ability to predict neural responses to natural images in the inferotemporal (IT) cortex Cadieu et al. (2014); Yamins et al. (2014b). This discovery supported the long-held theory that object recognition is a core objective of the visual cortex, and

suggested that more accurate DNNs would serve as better models of IT neuron responses to images Schrimpf et al. (2018); DiCarlo and Cox (2007); DiCarlo et al. (2012a). Since then, deep learning has undergone a revolution of scale: billion parameter-scale DNNs trained on billions of images are rivaling or outperforming humans at visual tasks including object recognition. Have today’s DNNs become more accurate at predicting IT neuron responses to images as they have grown more accurate at object recognition?

Across three independent experiments, we find this is not the case: DNNs have become progressively worse models of IT as their accuracy has increased on ImageNet. To understand why DNNs experience this trade-off and evaluate if they are still an appropriate paradigm for modeling the visual system, we turn to recordings of IT that capture spatially resolved maps of neuronal activity elicited by natural images Arcaro et al. (2017). These neuronal activity maps reveal that DNNs trained on ImageNet learn to rely on different visual features than those encoded by IT and that this problem worsens as their accuracy increases. We successfully resolved this issue with the *neural harmonizer*, a plug-and-play training routine for DNNs that aligns their learned representations with humans Fel et al. (2022a). Our results suggest that harmonized DNNs break the trade-off between ImageNet accuracy and neural prediction accuracy that assails current DNNs and offer a path to more accurate models of biological vision. Our work indicates that the standard approach for modeling IT with task-optimized DNNs needs revision, and other biological constraints, including human psychophysics data, are needed to accurately reverse-engineer the visual cortex.

3.3 Methods

3.3.1 Neural recordings

We leveraged recordings of IT neuronal responses from two monkeys Arcaro et al. (2017), which were designed to reveal feature preferences of putative face-selective neurons.

The recorded neurons also exhibited selectivity for non-face object features Arcaro et al. (2017). Recordings were made using chronically implanted 32-channel multi-electrode arrays within the fMRI-defined middle lateral (ML) and posterior lateral (PL) face patches of one monkey (Monkey 1), and ML of another monkey (Monkey 2, Fig. 3.1a). The activating regions in these areas were mapped (1-3°), then each image was shown to the animals after they were cued to fixate at a specific position in space (Fig. 3.1b). The same images were shown multiple times while the monkeys fixated at a red dot in the center of the screen, which made it possible to derive spatial activity maps of neuronal responses (Fig. 3.1c). Fixation positions fell in a 16×16 grid for Monkey 1 and 7×7 grid for Monkey 2, and the average neuronal activity at each position was taken to generate spatial activity maps. A total of 14 images were shown to Monkey 1, and a different set of 14 images were shown to Monkey 2.

The recordings for Monkey 1 resulted in responses from 32 neurons in ML and 31 neurons in PL. For Monkey 2, we obtained responses from 32 neurons (see Appendix 3.7.2 for more details). Following Arcaro et al. (2017), we binned neuronal responses every 40ms of the recording, from 50ms to 250ms. Within each bin, we calculated the noise ceiling for every neuron, which represents the maximum correlation achievable between any two neurons within that time interval. Noise ceilings for each recording were similar to those reported for standard IT Brain-Score datasets (Appendix 3.7.2). In the main text, we present modeling results from the time bins that exhibited the highest average noise ceiling for each Monkey and recording site. Additional results from other time bins, which are consistent with these findings, can be found in Appendix 3.7.5.

3.3.2 DNNs

We investigated the neural fits of 135 different DNNs representing a variety of approaches used in computer vision today: 62 [convolutional neural networks](#) trained on ImageNet Chen et al. (2021); Tan and Le (2019); Radosavovic et al. (2020); Howard et al.

(2019); Simonyan and Zisserman (2014); Huang et al. (2017); He et al. (2016); Zhang et al. (2019b); Gao et al. (2020); Kolesnikov et al. (2019); Sandler et al. (2018); Liu et al. (2021); Szegedy et al. (2016, 2015); Chollet (2017); Radford et al. (2021); Xie et al. (2017a); Xie and Tu (2016); Szegedy et al. (2015); Brendel and Bethge (2019); Mehta et al. (2018); Chen et al. (2017); Wang et al. (2018); Tan et al. (2019); Wu et al. (2018); Han et al. (2019); Yu et al. (2017); Wang et al. (2020) (CNNs), 23 DNNs trained on other datasets in addition to ImageNet (which we refer to as “DNN extra data”) Xie et al. (2017a); Radford et al. (2021); Liu et al. (2021), 25 vision transformers D’Ascoli et al. (2021); Touvron et al. (2021); Tolstikhin et al. (2021); Dosovitskiy et al. (2020); Steiner et al. (2021); Caron et al. (2021) (ViTs), 10 DNNs trained with self-supervision Chen et al. (2020); Yalniz et al. (2019), and 15 DNNs trained to be robust to noise or adversarial examples Geirhos et al. (2018); Salman et al. (2020). Each model was implemented in PyTorch with the TIMM toolbox (<https://github.com/huggingface/pytorch-image-models>) and pre-trained weights. Inference was executed on one NVIDIA TITAN X GPU. Additional model details, including the licenses used for each, are detailed in Appendix 3.7.3

3.3.3 Neural Harmonizer

It was recently found that DNN representations are becoming less aligned with human perception as they evolve and improve on ImageNet Fel et al. (2022a); Kumar et al. (2022); Bowers et al. (2022). A partial solution to this problem is the *neural harmonizer*, a training routine that can be combined with any DNN to align its representations with humans without sacrificing performance. Here, we test the hypothesis that aligning DNNs with human visual representations can similarly significantly improve their accuracy in predicting neural responses to images.

Training DNNs with the *neural harmonizer* on ImageNet involves an additional loss to standard cross-entropy for object recognition. This extra loss forces a model’s gradients to appear as similar as possible to feature importance maps derived from human behavioral

experiments (see Fel et al. (2022a) for details). To implement this loss, let $\mathcal{P}_i(\cdot)$ be a function that maps a multi-scale Gaussian pyramid of a human feature importance map ϕ to N , with $i \in \{1, \dots, N\}$. During training, we seek to minimize $\sum_i^N \|\mathcal{P}_i(\mathbf{g}(\mathbf{f}_\theta, \mathbf{x})) - \mathcal{P}_i(\phi)\|^2$ in order to align feature importance maps between humans and DNNs at every scale of the pyramid. Before they are compared, feature importance maps from humans and DNNs are normalized and rectified using $\mathbf{z}(\cdot)$, a function that transforms a feature importance map ϕ from either source to have 0 mean and unit standard deviation. This procedure yields the complete neural harmonization loss:

$$\mathcal{L}_{\text{Harmonization}} = \lambda_1 \sum_i^N \|(\mathbf{z} \circ \mathcal{P}_i \circ \mathbf{g}(\mathbf{f}_\theta, \mathbf{x}))^+ - (\mathbf{z} \circ \mathcal{P}_i(\phi))^+\|_2 \quad (3.1)$$

$$+ \mathcal{L}_{CCE}(\mathbf{f}_\theta, \mathbf{x}, \mathbf{y}) + \lambda_2 \sum_i \theta_i^2 \quad (3.2)$$

Following the original *neural harmonizer* implementation Fel et al. (2022a) and training recipe, we trained six DNNs on ImageNet and human feature importance maps from the *ClickMe* game Linsley et al. (2019): **VGG16** Simonyan and Zisserman (2014), **LeViT** Graham et al. (2021), **ResNetV2-50** He et al. (2016), **EfficientNet_b0** Tan and Le (2019), **ConvNext** Liu et al. (2021), and **MaxViT** Tu et al. (2022). Each model was trained with Tensorflow 2.0 on 8 V4 TPU cores with all of the images in the ImageNet training set, and *ClickMe* human feature importance maps for the 200,000 images which were annotated. Images and feature importance maps were augmented with mixup Zhang et al. (2017), random left-right flips, and random crops during training. Only the object recognition loss was computed for images without human feature importance maps. Model weights were optimized with stochastic gradient descent and momentum, batches of 512 images, label smoothing Muller et al. (2019), a learning rate of 0.3, and a learning rate schedule that began with a 5-epoch warm-up period followed by a cosine decay over 90 epochs at steps 30, 50 and 80.

3.3.4 Neural fitting

We evaluated the neural fit of each model by computing their fit separately for each IT recording, using the Brain-Score evaluation method Schrimpf et al. (2018). This method involved fitting image-evoked activities from a layer of a deep neural network (DNN) to the corresponding neural responses using the partial least squares regression algorithm from Scikit-Learn.

To implement the Brain-Score evaluation method on the spatially-resolved recordings we investigate here, we first split each image shown to each animal into an equal number of patches as there were fixation locations in an image. Each image patch captured the receptive field of recorded neurons, and in total, there were 4,046 image patches for ML and 4,046 image patches for PL in Monkey 1, and 1,134 image patches for ML in Monkey 2. We measured DNN neuronal fits as the Spearman correlations between model predictions and true neural responses for each patch of an image held out of training divided by each neuron’s noise ceiling. We then stored the median Spearman correlation over neurons and repeated the training/testing procedure to get the mean correlation across all images viewed by a monkey. Separate fitting procedures were performed for every layer of activities in a model, and we report a DNN’s Brain-Score as its best possible fit across layers.

3.4 Results

3.4.1 Task Optimization is insufficient for reverse-engineering IT

Ever since it was found that DNNs optimized for object recognition produce accurate predictions of IT neural responses to images, it was suggested that prediction accuracy would continue improve alongside DNN performance on ImageNet Yamins et al. (2014b); Yamins and DiCarlo (2016); Schrimpf et al. (2018). Is this the case with today’s DNNs

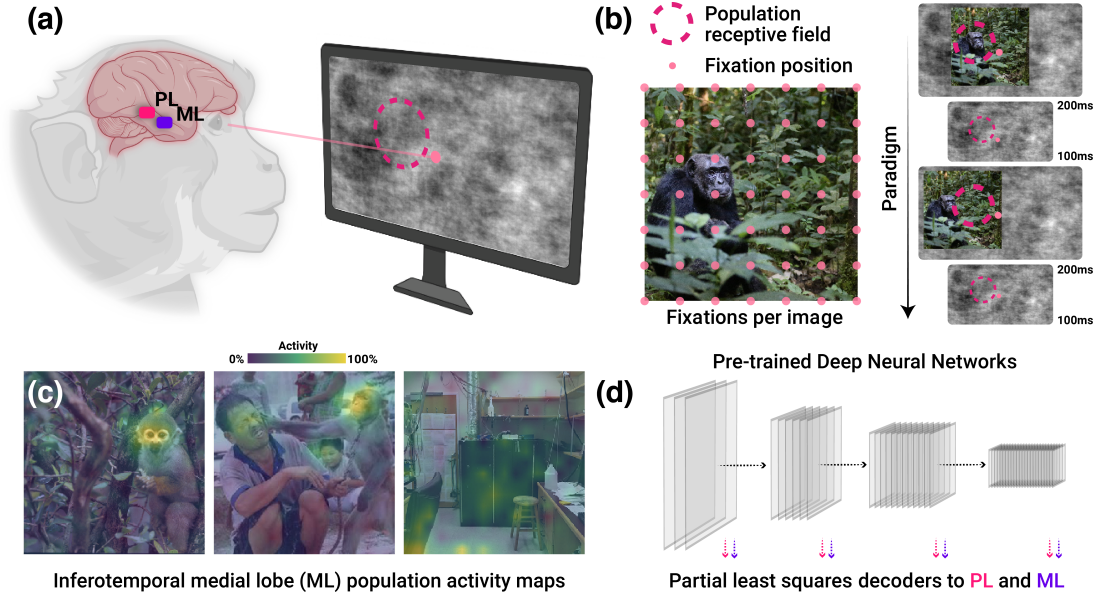


Figure 3.1: **IT recordings that reveal spatial maps of neuronal responses to complex natural images offer unprecedented insights into their feature selectivity** Arcaro et al. (2017). (a) Neurons in posterior (PL) and/or medial (ML) lateral IT in two animals were localized using functional magnetic resonance imaging (fMRI), and neural responses to images were recorded using chronically implanted 32-channel multi-electrode arrays. (b) The monkeys were rapidly shown each image multiple times for 200ms each time (Monkey 1: $n = 256$, Monkey 2: $n = 49$). Images were positioned differently each time to measure neural responses to every part of the image. (c) This procedure yielded spatially resolved maps of neural activity for an image, revealing the relative importance of different features for the recorded neurons. (d) DNNs were fit to these recordings following the Brain-Score evaluation method Schrimpf et al. (2018), in which partial least squares decoders were used to find the units in a DNN that provided the best match for neuronal responses to images.

that rival or exceed the performance of human object recognition?

To answer this question, we leverage recordings of neuronal responses to high-resolution natural images in medial (ML) and posterior lateral (PL) IT (see Methods and Fig. 3.1). These images fall within the same statistical distribution as ImageNet images (Fig. 3.5), ensuring that our findings are not influenced by distributional shifts that are a well-known problem faced by computer vision models Linsley et al. (2018). Image-evoked neural responses were spatially mapped, enabling insights into the visual features driving the responses of IT neurons and DNNs. Moreover, while these regions were localized according to their selectivity to face stimuli, they were also noted to respond to non-face

stimuli Arcaro et al. (2017).

Across 135 different DNNs, representing the variety of approaches used today in computer vision, we found that DNNs pretrained on ImageNet have become progressively less accurate at predicting ML and PL responses in two separate Monkeys (Fig. 3.2). For instance, **ConvNext tiny**, which achieved 74.3% accuracy on ImageNet, is as accurate in predicting neural responses to images as the **ResNetv2-152x4**, which reached 84.9% accuracy on ImageNet. Moreover, training routines that have been suggested to be more biologically plausible or yield representations that are closer to biological vision, such as training with self-supervision Zhuang et al. (2020) or for adversarial robustness Guo et al. (2020), made no difference. All DNNs trained on internet data faced a pareto-front that bounded their accuracy in predicting IT responses as a function of their ImageNet accuracy.

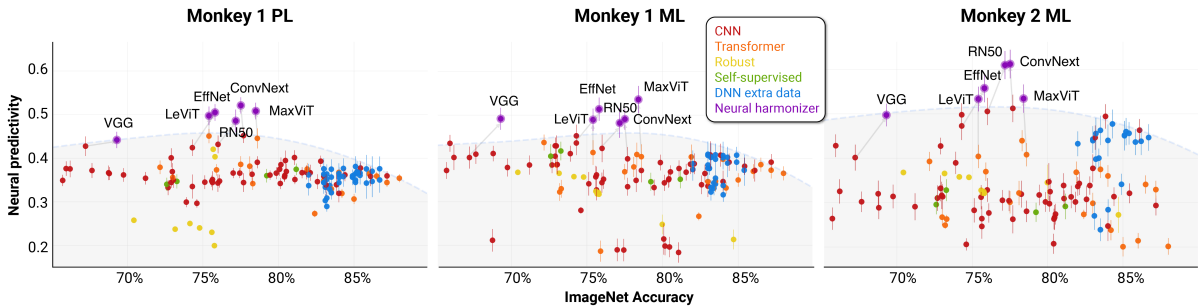


Figure 3.2: DNNs trained on ImageNet face a trade-off between achieving high object recognition accuracy and predicting responses to natural images in IT.. We measured the accuracy of 135 DNNs at predicting image-evoked responses from neurons in posterior lateral (PL) and medial-lateral (ML) areas of IT Arcaro et al. (2017) by computing the neural predictivity of each model with the Brain-Score evaluation method Schrimpf et al. (2018). DNN neural predictivity is progressively worsening as models improve on ImageNet. This problem can be partially fixed by training DNNs with the *neural harmonizer*, which aligns their learned object representations with humans. Error bars denote 95% bootstrapped confidence intervals.

3.4.2 DNNs need biologically-aligned training routines

There are at least two potential reasons why ImageNet-trained DNNs face a trade-off between object recognition and neural prediction accuracy: (i) DNN data diets and

training routines work well for ImageNet but lead them to learn features that are misaligned with biological vision Fel et al. (2022a); Linsley et al. (2019, 2017); Ullman et al. (2016) or (ii) their architectures are a poor match to visual cortex Eberhardt et al. (2016). To test this first possibility, we turned to the *neural harmonizer*. Given prior work demonstrating that aligning DNNs with human perception using the *neural harmonizer* significantly improved their ability to predict human behavior, we reasoned that harmonization might similarly improve DNN explanations of neural data by forcing them to rely on human-like visual features. Indeed, we found that harmonized DNNs (hDNNs) were significantly more accurate at predicting image-evoked responses in ML ($T(5) = 5.30$, $p < 0.01$) and PL ($T(5) = 6.11$, $p < 0.001$) neurons of Monkey 1 and ML ($T(5) = 6.89$, $p < 0.001$) neurons of Monkey 2 (reported as the average of T score from independent samples t -tests comparing the predictivity of each hDNN to the DNN with the highest neural-predictivity). The success of hDNNs indicates that the architectural mismatch between DNNs and the visual cortex is far less of a problem for modeling the visual cortex than the training routines and data diets that are being used today to achieve high accuracy on ImageNet. Behavior – even from humans – is an important constraint on DNNs that is needed to build more accurate models of primate visual cortex.

3.4.3 Spatially-mapped neural responses reveal a feature mismatch between IT and DNNs

The IT recordings used in the Brain-Score benchmark all utilized the same paradigm, in which hundreds or thousands of images were shown to multiple animals during passive viewing Majaj et al. (2015). In contrast, the recordings we rely on involve many fewer unique images, but neural responses for each image are spatially mapped. This spatial mapping reveals what types of features are driving neural responses and makes it easier to understand the successes and failures of DNNs in explaining these data Arcaro et al. (2017).

Although the neurons in ML and PL were located based on their selectivity to faces, the spatial maps of their responses revealed a much more complex response profile. IT neurons in both animals were strongly driven by faces, non-face objects, and contextual cues associated with faces Arcaro et al. (2017). In contrast, the best-fitting units of DNNs with state-of-the-art accuracy on ImageNet, like the **ResNetv2-152x4** (Fig. 3.3), responded strongly to background features. This problem was shared by DNNs trained on internet data but with routines that have been considered more biologically plausible, like self-supervised learning (e.g., **SSL ResNet-18**), or routines that yield perceptual robustness that is closer to humans, like adversarial training (e.g., **Robust ResNetv2-50**). hDNNs like the **harmonized MaxViT**, on the other hand, were, in general, reliably predictive of what parts of images elicited the most activity from IT neurons.

3.4.4 hDNNs generate testable predictions for visual neuroscience

The surprising effectiveness of DNNs pre-trained on object recognition for predicting IT responses to images was a significant finding in computational neuroscience, which raised the tantalizing possibility that these models could form the basis of neuroprosthetics, and reduce the need for animal experiments. However, a persistent problem with DNNs since their inception is their lack of interpretability, meaning that using DNNs for systems identification effectively swaps one black box (the visual system) with another (a DNN) without getting the field any closer to understanding how vision works. Recent strides in the field of explainable artificial intelligence (XAI) have begun to alleviate this problem. For instance, it is now possible to reliably locate and characterize the features in datasets, like ImageNet, that drive patterns of model behavior using concept recursive activation factorization (CRAFT Fel et al. (2022b)). When paired with accurate models of IT and recordings that reveal locations in images that elicit neuronal responses, CRAFT can generate testable predictions of *what* those features are.

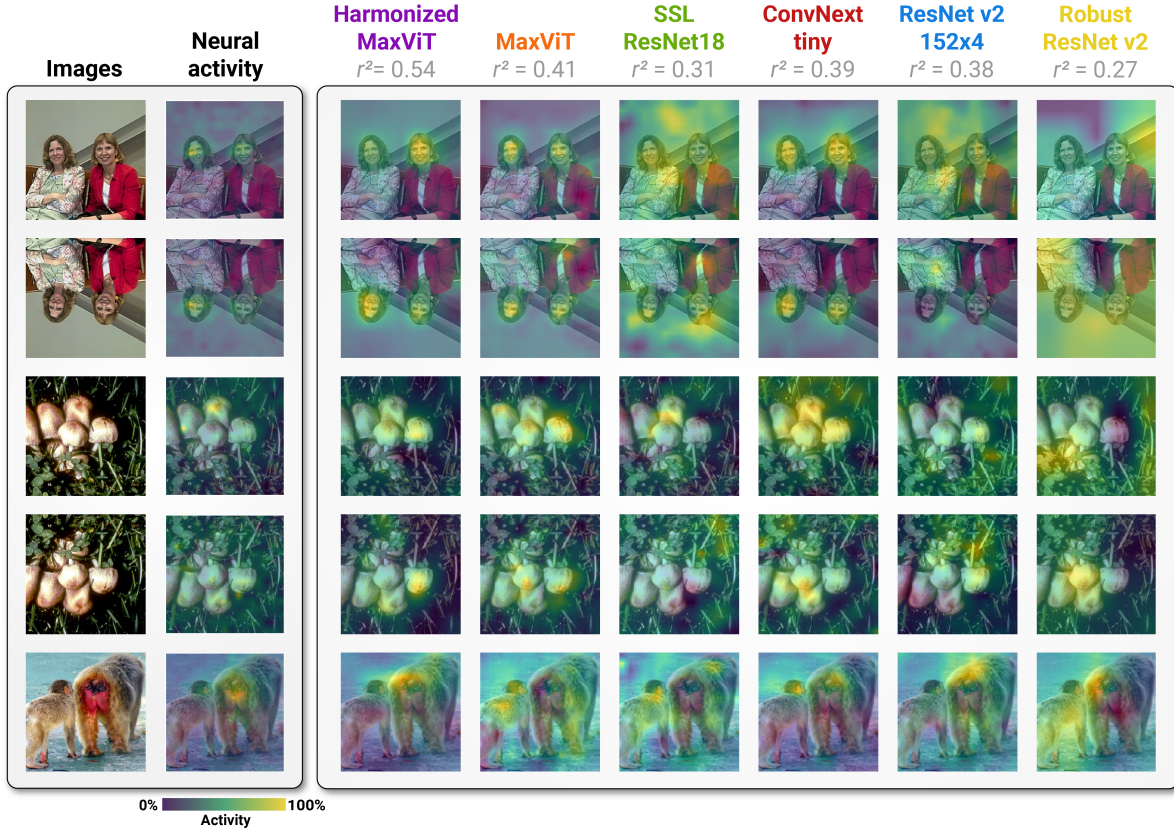


Figure 3.3: DNNs optimized for object recognition on ImageNet rely on different features than those encoded by neurons in primate inferior temporal (IT) cortex.. The activity of PL IT neurons is plotted next to the predicted activity of a model representing each class of DNNs: harmonized DNNs (hDNNs), visual transformers, self-supervised DNNs, convolutional neural networks, DNNs trained on more data than ImageNet, and adversarially robust DNNs.

We used CRAFT to extract features from harmonized and regular ResNet-50s trained on ImageNet that explain their predictions of neural activity in PL of Monkey 1 (Fig. 3.4). To do this, we decomposed a DNNs predicted activities for image patches into the most important and distinct features using non-negative matrix factorization, then scanned over every single image patch to find those which most strongly explained each discovered feature. While the harmonized ResNet-50 predicted that parts of faces, arms, and heads explained the majority of variance in IT, the less accurate and unharmonized ResNet-50 predicted that the background, oriented-edges, and a feature combining head, body, arm, and hand parts were most important for IT. Thus, hDNNs not only break the trade-off between ImageNet and neural prediction accuracy of ImageNet-trained DNNs, they can

also generate testable hypotheses on the relative importance of different features for visual system neurons.

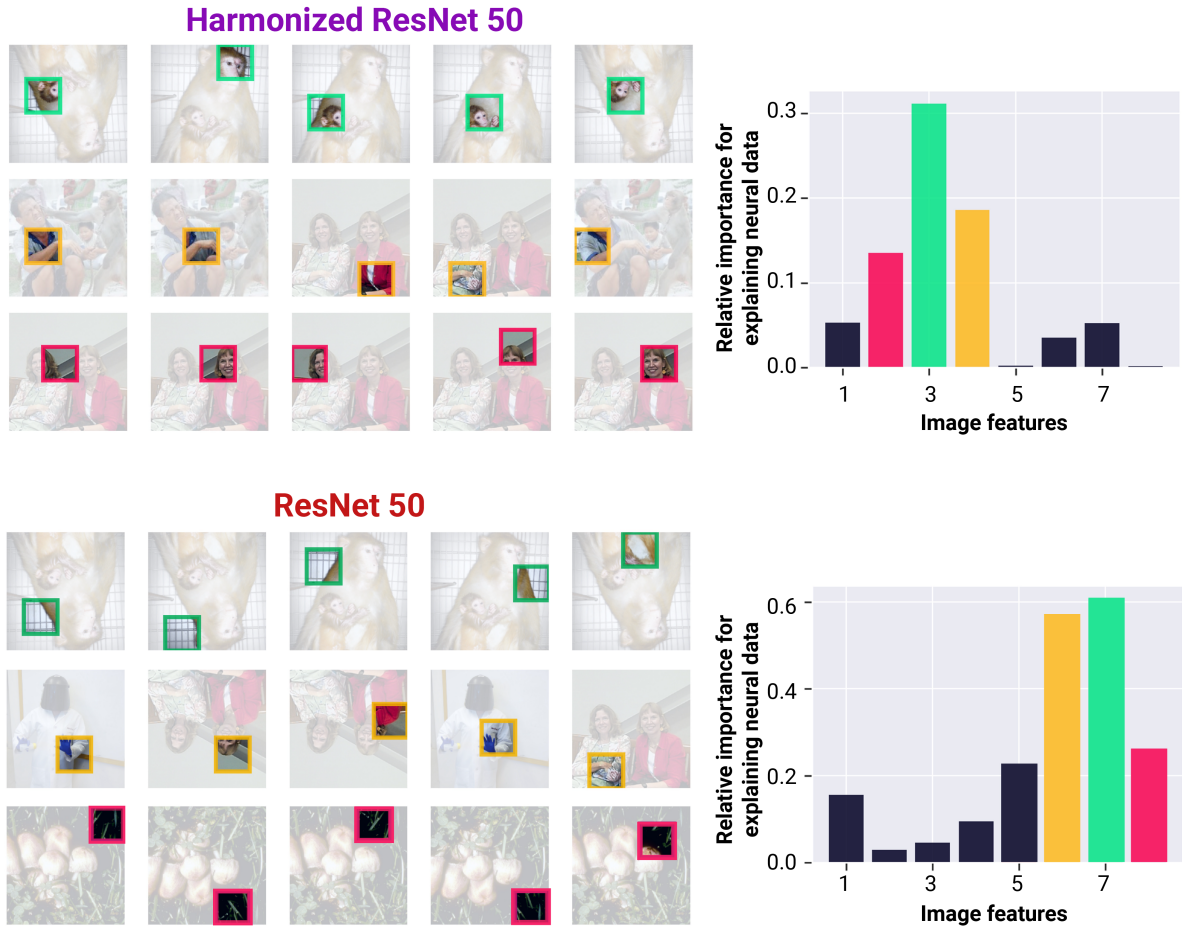


Figure 3.4: **Predictions of the visual features drive neuronal responses in PL of Monkey 1 from a harmonized ResNet-50 and a standard ResNet-50.** Image patches for each model depict what the important features are across all images shown to the animal. The relative importance of each feature for predicting the recordings is color-coded in the bar chart on the right.

3.5 Discussion

3.5.1 A revised approach for reverse engineering visual cortex

Biological data is expensive to gather and often noisy. This makes the prospect of accurately modeling the responses of IT to complex stimuli especially daunting. DNNs optimized for object recognition represented a potential solution for this problem: pre-

training on internet datasets alleviated training issues associated with small-scale neural data, and the architectures and training routines of accurate DNNs could offer insights into the circuits that underlie visual perception as well as the developmental principles that shape those circuits. Our findings suggest that while DNNs still hold great potential for predicting image-evoked responses from IT neurons, new training routines and data diets are necessary for continued improvement.

The *neural harmonizer* is a partial solution to the problems that DNNs face in modeling primate IT. The success of hDNNs in breaking the pareto-front faced by 135 different DNNs indicates that significant aspects of primate perception and the neural circuits that shape it cannot be divined from internet data alone. We believe that the *neural harmonizer* and learning constraints provided by large-scale human behavior data is only a short-term solution to this problem, and that if DNNs were able to learn about the visual world more like primates do, they would be even better at predicting neural data. In support of this goal, we release our code and data at https://serre-lab.github.io/neural_harmonizer/.

3.5.2 Limitations

One limitation of our work is that we use the responses of neurons with face-selectivity to replicate and understand the trade-off between ImageNet accuracy and neural predictivity faced by DNNs. As faces (and even humans) are not a category in ImageNet, it is possible that this dataset Arcaro et al. (2017) could bias our results in ways that are difficult to predict¹. However, we find the same ImageNet accuracy and neural predictivity trade-off on this dataset (Fig. 3.2) as we did on the three recordings of object selective neurons hosted on the Brain-Score website (Fig. ??), indicating that DNNs face similar issues in predicting each set of recordings. Moreover, the neural harmonizer was successful in breaking this trade-off, even though it involves ImageNet training using human feature

¹Note, however, that there are many animal and human faces in Imagenet images, giving DNNs ample opportunity to learn about them.

importance maps for objects. Finally, it was noted in the original paper where our recordings came from that the neurons, while localized based on their face-selectivity, responded to non-face stimuli as well Arcaro et al. (2017). In summary, our work reliably demonstrates that new paradigms are needed to advance DNNs as models of IT, and spatially-resolved recordings such as those used in this work support this goal.

Another limitation of our work is that while we found that hDNNs are significantly more predictive of IT neuron responses than any other DNN, they still explain only 50-60% of the variance in neuronal activity. One straightforward way of doing better is by expanding the dataset of human feature importance maps we used for harmonization from annotations for approximately 200,000 images to the entire 1.2M ImageNet dataset.

3.5.3 Broader impacts

By building better models of primate IT, we are taking significant steps toward the reducing the reliance of visual neuroscience on animal models for experimentation, supporting the development of neuroprosthetic devices that resolve visual dysfunctions, and providing vision scientists with a richer understanding of how IT works. Our findings also highlight a main limitation of the scaling laws that are guiding progress throughout artificial intelligence today: scale is not sufficient for explaining biological intelligence.

3.6 Final Remarks: Connecting Neural Alignment to the Thesis

The findings presented in this chapter represent a critical bridge between the human alignment work of Chapter 1 and the broader implications for computational neuroscience and artificial intelligence. By demonstrating that DNNs have become progressively worse models of IT cortex as they have improved on ImageNet, we have identified a fundamental limitation of current scaling laws: they are incompatible with modeling

biological intelligence.

The success of neural harmonization in breaking this trade-off provides a powerful proof-of-concept that biological constraints—specifically human behavioral data—can guide the development of more accurate models of neural systems. This finding has profound implications for the thesis’s central research question: how can we build deep neural networks that are simultaneously aligned with human visual strategies and neural responses in the visual cortex?

The neural harmonizer approach demonstrates that it is possible to achieve both high task performance and biological alignment, challenging the assumption that these objectives are fundamentally at odds. By explicitly incorporating human behavioral data into the training objective, we can create models that not only perform well on standard benchmarks but also provide more accurate approximations of biological vision.

This work also highlights the importance of moving beyond purely task-optimized training paradigms. The fact that harmonized DNNs outperform standard DNNs in predicting neural responses suggests that the key to building better models of biological vision lies not in simply scaling up existing architectures, but in carefully designing learning objectives that incorporate biological constraints.

As we move forward in the thesis, these findings will inform our approach to building more comprehensive models that can simultaneously account for human behavior, neural responses, and task performance. The neural harmonizer represents a crucial step toward this goal, demonstrating that biological alignment is not only possible but essential for continued progress in computational neuroscience and artificial intelligence.

3.7 Supplementary Information: Neural Alignment

3.7.1 Datasets

Distribution shifts One possible explanation for the trade-off we observed between DNN accuracy on ImageNet vs. predictions of neural data is that there was a distributional shift between the images used in the two evaluations. To test this possibility, we extracted activity vectors for a large sample of images from ImageNet and all images from the standard Brain-Score benchmark dataset for IT Majaj et al. (2015) and the images we used in our experiments Arcaro et al. (2017). For ImageNet, we used all images in the validation set that also had human annotations from *ClickMe*.

We measured distribution shifts by measuring the accuracy of a linear classifier in discriminating between ImageNet and either the standard Brain-Score benchmark dataset for IT Majaj et al. (2015) or the images we used in our experiments Arcaro et al. (2017). We began by taking equal-sized random samples of ImageNet image encodings and encodings from either the Brain-Score dataset (28 of 191 images) or our dataset (all 28 images). We then trained a linear SVM using scikit-learn and performed leave-one-out cross-validation on either set of samples. Finally, we repeated this procedure 1,000 times, using a new random sample of ImageNet encodings each time, and measured the average accuracy for discriminating between ImageNet and either neural data image set. We found that images used in the standard Brain-Score benchmark for IT Majaj et al. (2015) are significantly further out-of-distribution from ImageNet than the images that we used in the main text Arcaro et al. (2017) ($T(999) = 98.12$, $p < 0.001$, Fig. 3.5).

3.7.2 Neural data analysis

We measured the maximum explainable variance, or noise ceiling, in our neural data recordings in every 40ms bin of activity from 0ms - 250ms. We binned every 40ms by following the original procedure for processing this data Arcaro et al. (2017). Because

images were not repeatedly shown to the same population receptive field, we computed each neuron’s noise ceiling for an image as the highest correlation between it and another neuron for that image. We then focused our analyses in the main text on the time bin where each neurons in each animal/brain area achieved the highest average noise ceiling (Fig. 3.6).

3.7.3 Models

DNN Model Zoo We comprehensively evaluated the ability of a myriad of DNNs from the TIMM toolbox. These DNNs, available under the Apache 2.0 license, are intended for non-commercial research purposes. The complete list of DNNs we evaluated can be found in Table 3.1.

Architecture	Model	Versions
CNN	VGG	4
	ResNet	24
	EfficientNet	8
	ConvNext	10
	MobileNet	6
	Inception	7
	DenseNet	4
	RegNet	3
	Xception	4
	MixNet	4
	DarkNet	2
	NFNet	2
	TinyNet	4
	LCNet	2
	DLA	2
	MnasNet	2
	Resnext101	6
	GhostNet	2
ViT	General ViT	21
	leViT	4
	MaxViT	7
	DeiT	4
Hybrid	CoAtNet	5

Table 3.1: **Our DNN model zoo, taken from the TIMM library.**

Model	Accuracy	Human Alignment	Note
VGG	69.3	61.5	$\lambda = 2$
ResNet 50	77.17	45.0	$\lambda = 2$
EfficientNet B0	77.51	52.3	$\lambda = 20$
ViT B16	75.7	72.6	$\lambda = 5$
ConvNext Tiny	75.9	73.2	$\lambda = 1$
MaxViT Tiny	78.6	45.3	$\lambda = 1$

Table 3.2: DNNs trained with the *neural harmonizer*.

3.7.4 Neural Harmonizer Training

We relied on the original neural harmonizer recipe Fel et al. (2022a) to train and harmonize 14 DNNs for object recognition on the ImageNet Deng et al. (2009b) dataset. For each model, we searched for settings of the regularization terms λ_1 and λ_2 , which control the relative importance of losses for object recognition and human feature alignment during training, that maximized accuracy on the ImageNet validation set. In total, we trained a VGG16, ResNet50_v2, ViT_b16, EfficientNet_b0, ConvNext Tiny, and MaxViT Tiny (Table. 3.2). We did not attempt to apply the neural harmonizer to models which relied on pretraining with datasets other than ImageNet because the *ClickMe* feature importance dataset we used only contained annotations on a subset of images in ImageNet. Thus, we expect that more work is needed for expanding *ClickMe* to larger-than-ImageNet scale data before these DNNs can be harmonized.

3.7.5 Extended results

Cross validation One limitation of the evaluation approach used in the original Brain-Score work Schrimpf et al. (2018) is that model selection is done on the same dataset as testing (*i.e.*, there is no partitioning of the data into separate training, validation, and test sets). Thus, it is possible that the results we describe in the main text are a byproduct of “double dipping.” To test for this possibility, we repeated our experiments using separate training, validation, and test splits of the data. For each round of cross validation in this procedure, we held out one image for testing. We also held out half of the training data

for selecting the model layer that best predicted neural responses (validation). Our results from this approach strongly correlated with the ones reported in our original submission and our overall conclusions remain the same (Fig. 3.7; Monkey 1 ML $\rho = 0.98$, $p < 0.001$; Monkey 2 ML $\rho = 0.97$, $p < 0.001$); and Monkey 2 PL $\rho = 0.98$, $p < 0.001$).

Alternative objective functions and datasets To better understand how ImageNet *per se* affected DNNs’ abilities to predict neural responses, we evaluated a zoo of them trained on the Taskonomy dataset and task set Zamir et al. (2018) and also the Ecoset Mehrer et al. (2021a) naturalistic object categorization dataset and task. We took 19 DNNs pretrained on each task in the Taskonomy, and 4 DNNs trained for classification on ecoset, and applied the standard Brain-Score fitting procedure that we describe in our original submission Methods to derive prediction accuracy scores for each model. These models were far less effective at explaining neural activity than any of the ImageNet-trained models we provided in the main text (Tables 3.3 and 3.4). To summarize, our human behavior-aligned harmonized DNNs are more accurate at predicting neural responses than other DNN we tested, regardless of the dataset or task that was used to train them.

Performance on Brain-Score As a point of comparison, we tested a subset of our DNN on a standard inferotemporal (IT) recordings benchmark on **Brain-Score.org**. In each case, harmonization marginally improved performance over baseline (Table 3.5).

Neural dynamics We also present modeling results on other time bins (with lower noise ceilings) than those considered in the main text (Figs. 3.8- 3.10). For ML IT in monkey 1 and monkey 2, we present 90ms-130ms, 130ms-170ms, and 210ms-250ms. For PL IT of Monkey 1, we present 130ms-170ms, 170ms-210ms, and 210ms-250ms, windows. In each case, we found a similar trade-off in these data as we did in those highlighted in the main text, as well as a significant improvement for harmonized vs. unharmonized DNNs. This similarity extends to the qualitative predictions made by each model on ML IT neurons (recall that we show the image-evoked activities of PL IT neurons in the main

Task	ML	PL
Autoencoding	0.005	-0.009
Class object	0.091	0.075
Curvature	0.019	-0.009
Denoising	-0.014	0.013
Edge occlusion	0.029	-0.0002
Edge texture	0.025	0.014
Egomotion	0.002	-0.018
Fixated pose	0.118	0.101
Jigsaw	0.099	0.097
Keypoints 2D	0.024	0.015
Keypoints 3D	-0.057	-0.006
Nonfixated pose	0.164	0.086
Normal	0.065	0.056
Reshading	0.019	0.001
Room layout	0.131	0.087
Segment semantic	-0.011	0.029
Segment unsupervised 25D	0.041	0.008
Segment unsupervised 2D	0.052	0.043
Vanishing point	0.046	0.076

Table 3.3: **Neural predictions for ML and PL in Monkey 1 using Taskonomy-trained DNNs.**

Model	Best time bin	Score
Alexnet	90	0.046
ResNet50	130	0.111
VGG16	130	0.086
Inception	210	0.109

Table 3.4: **Neural predictions for ML in Monkey 1 using Ecoset-trained DNNs.**

Model	Neural	ImageNet
VGG baseline	0.523	67.23
VGG harmonized	0.524	67.369
ResNet50v2 baseline	0.501	76
ResNet50v2 harmonized	0.502	77.17
ConvNext baseline	0.519	74.3
ConvNext harmonized	0.524	75.8
MaxViT baseline	0.541	78.6
MaxViT harmonized	0.543	78.7
EfficientNetB0 baseline	0.506	75.4
EfficientNetB0 harmonized	0.508	77.51
LeViT baseline	0.472	74.9
LeViT harmonized	0.473	75.4
ViT baseline	0.493	73.2
ViT harmonized	0.499	75.7

Table 3.5: **Performance of select DNNs on inferotemporal (IT) recordings from Brain-score.org Majaj et al. (2015).** For each DNN class, models with higher accuracy at predicting IT responses and ImageNet recognition are bolded.

text, 3.10).

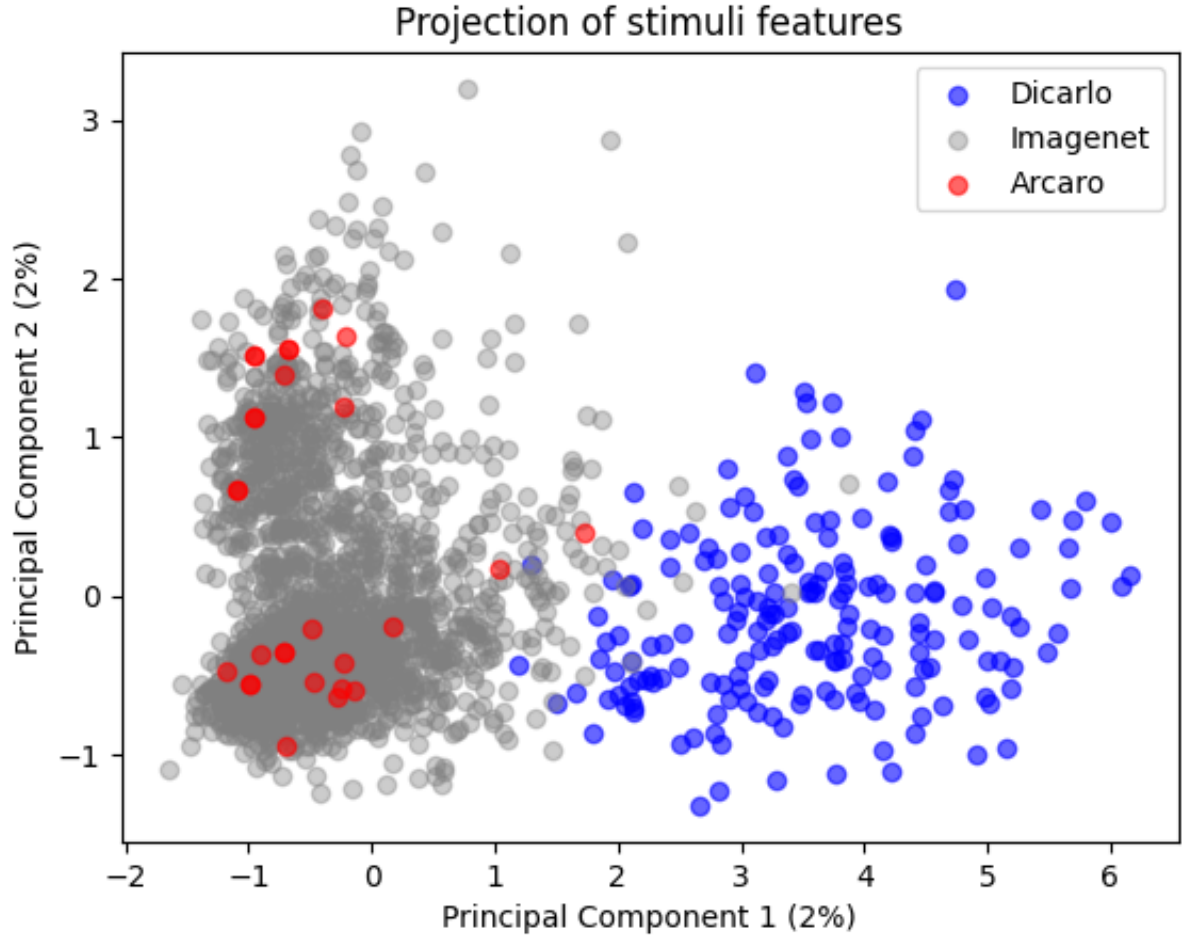


Figure 3.5: **Stimuli from Majaj et al. (2015), but not Arcaro et al. (2017), are out-of-distribution from ImageNet.** Penultimate-layer activities from a ResNetv2-50 trained on ImageNet were extracted for 1916 images from the validation set of ImageNet, 191 images from Majaj et al. (2015), and 28 images from Arcaro et al. (2017). A principle components analysis decomposed those activities into a 2-Dimensional projection, which demonstrated a distribution shift for stimuli from Majaj et al. (2015), but not Arcaro et al. (2017). Indeed, while stimuli from Majaj et al. (2015) could be reliably discriminated from ImageNet (98.7 acc, $T(999) = 98.12$, $p < 0.001$), images from Arcaro et al. (2017) were not discriminable from ImageNet (56.5 acc, $T(999) = 98.12$, *n.s*).

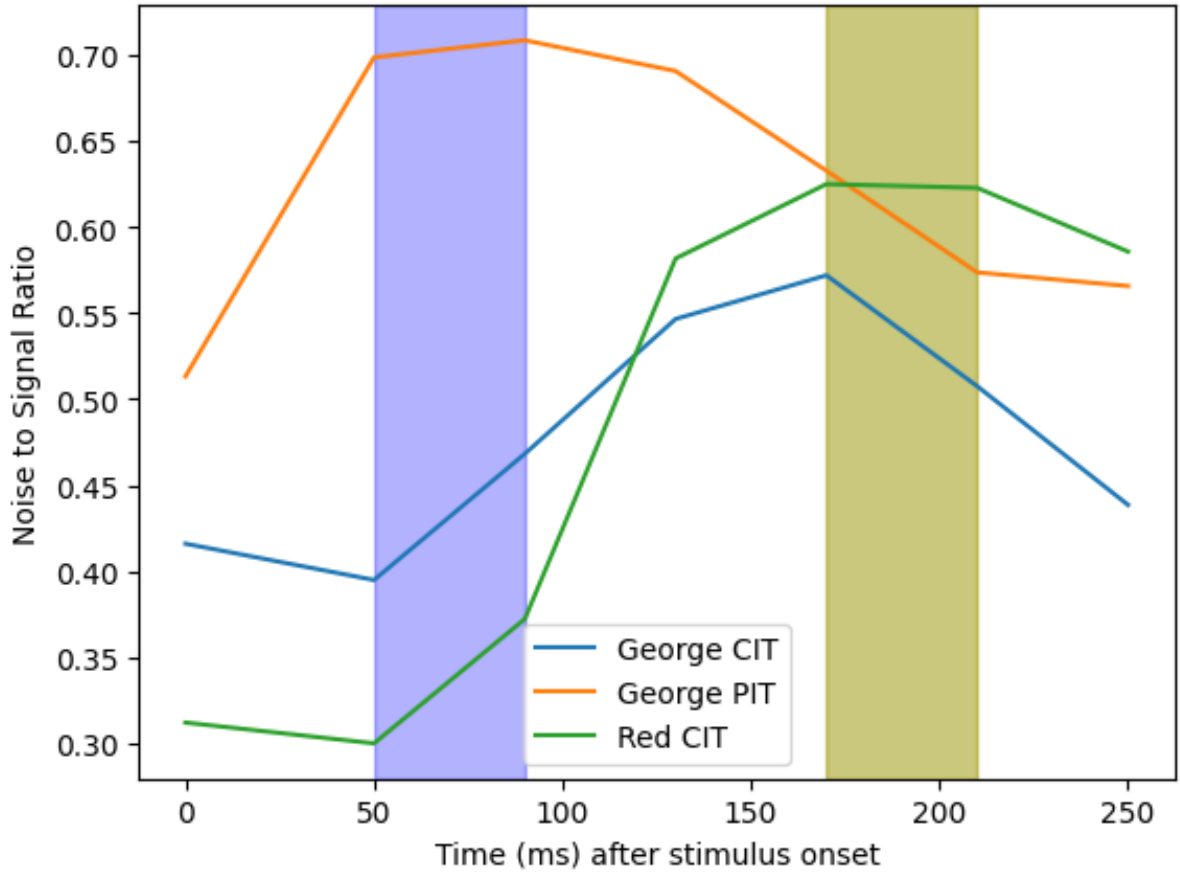


Figure 3.6: **Noise ceilings at every 40ms bin of activity from 0ms - 250ms.** We binned every 40ms of neural responses following the original procedure for processing this data Arcaro et al. (2017)

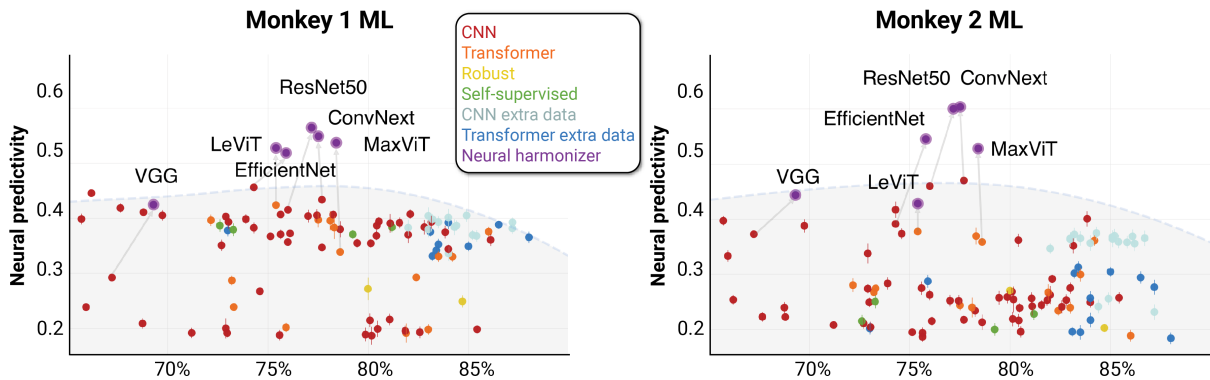


Figure 3.7: **DNN neural prediction accuracy after evaluating on a test set held out of training.** Unlike the original approach introduced in Brain-Score Schrimpf et al. (2018), here we train, validate, and test models on separate partitions of the dataset. DNNs trained and evaluated in this way follow the trends described in the main text.

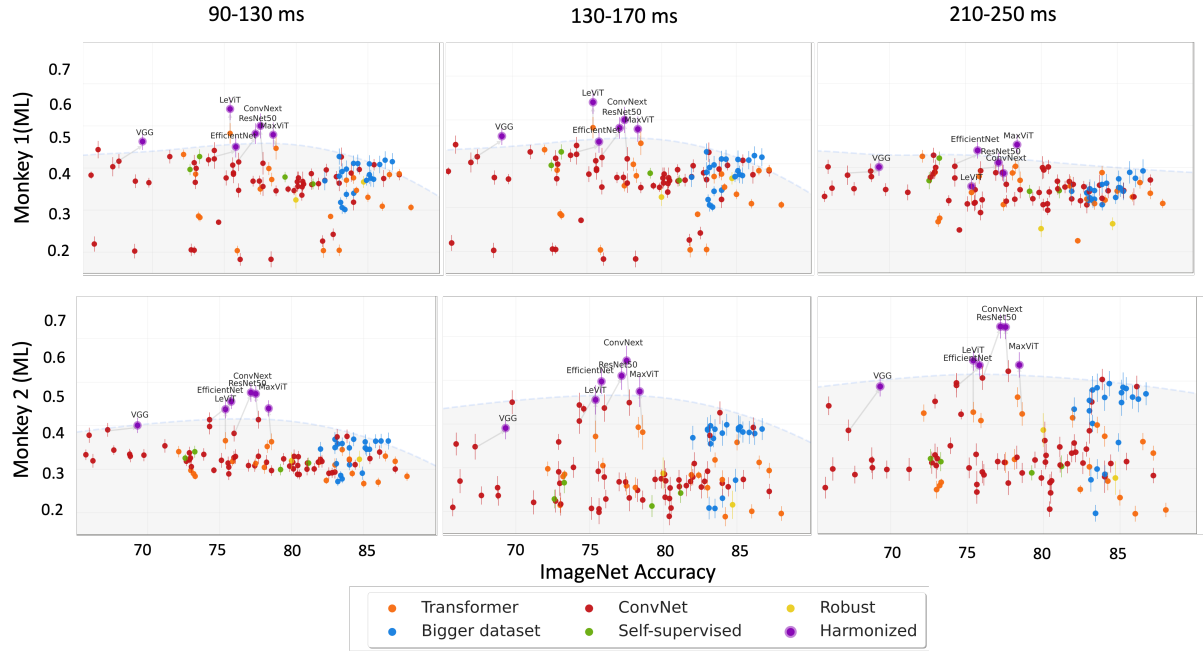


Figure 3.8: **Results using different time bins than in the main text for the same region (ML) for both monkeys.** The pattern of results for each time-bin are consistent with the main text, despite these data having lower noise ceilings.

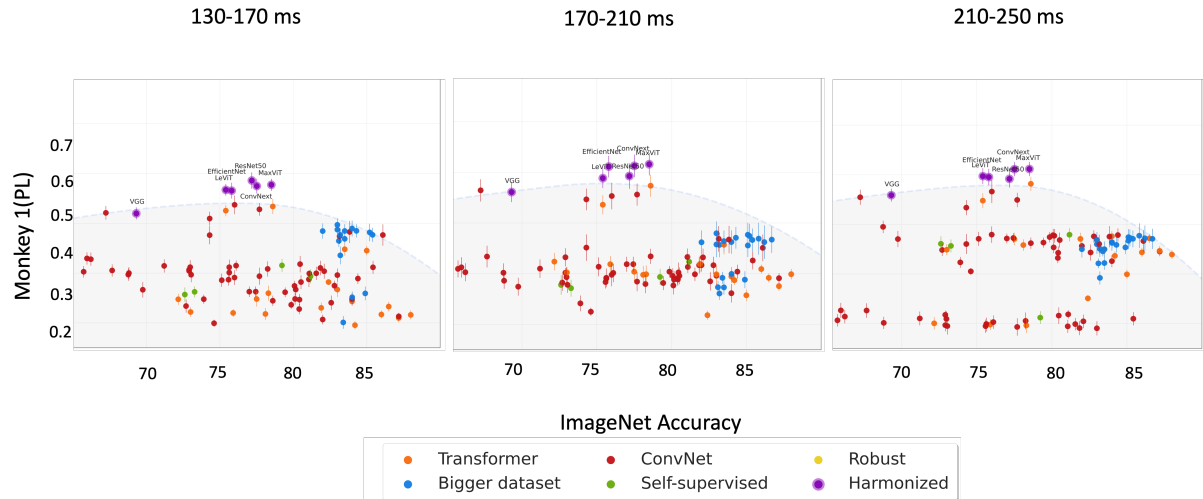


Figure 3.9: **Results using different time bins than in the main text for the IT PL in Monkey 1.** The pattern of results for each time-bin are consistent with the main text, despite these data having lower noise ceilings.

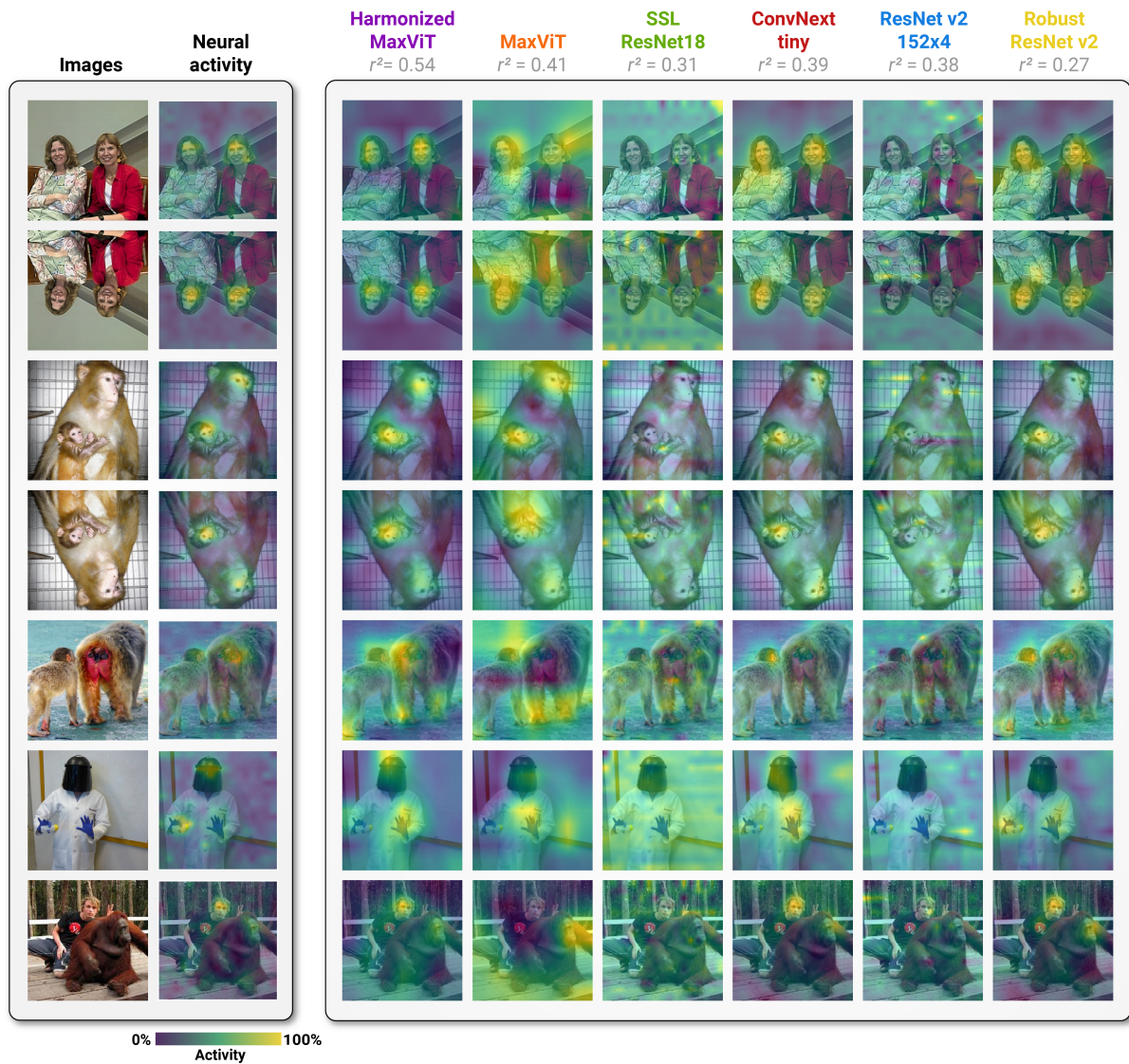


Figure 3.10: DNNs optimized for object recognition on ImageNet rely on different features than those encoded by neurons in primate inferior temporal (IT) cortex. The activity of ML IT neurons of Money 1 is plotted next to the predicted activity of a model representing each class of DNNs: harmonized DNNs (hDNNs), visual transformers, self-supervised DNNs, convolutional neural networks, DNNs trained on more data than ImageNet, and adversarially robust DNNs.

CHAPTER 4

HMAX STRIKES BACK: BIOLOGICALLY-INSPIRED ARCHITECTURES FOR SCALE INVARIANCE

4.1 Introduction

In the first two chapters of this thesis, we explored how behavioral constraints during training, which we called “harmonization”, can substantially improve the alignment between DNNs and biological vision. These results highlighted the importance of training objectives that reflect human visual strategies. However, harmonization alone does not address a second, equally critical challenge: the architectural validity of the models themselves. If our goal is not just alignment in outcomes but also biological plausibility in mechanisms, we must look beyond optimization and into the structural principles that shape visual computation in the brain. In this chapter, we turn our focus to that complementary problem: investigating whether biologically inspired architectures can

provide the computational substrate necessary for robust, human-like visual processing. By doing so, we aim to extend and enrich the progress achieved through harmonization, moving closer to models that are both functionally aligned and anatomically grounded.

Harmonization addresses the “what” question (which features should be important) by aligning DNNs with human visual strategies, while architectural constraints address the “how” question (how should these features be processed) by implementing biologically-inspired computational mechanisms. The success of harmonization in improving neural alignment demonstrates that training constraints are valuable, but the architectural limitations of standard DNNs prevent them from achieving the full range of biological visual capabilities. Thus, both approaches are necessary: harmonization provides the right learning objectives, while architectural constraints provide the computational substrate that enables truly brain-like processing. This chapter builds upon rather than contradicts the harmonization findings, showing that architectural innovations can further enhance the alignment achieved through training constraints.

The Fundamental Tension Between Engineering and Neuroscience This work sits at the intersection of two different scientific goals: building high-performing visual systems (engineering) and understanding how biological vision works (neuroscience). Modern computer vision has achieved remarkable success through techniques that bear little resemblance to biology: massive datasets, backpropagation, and extensive data augmentation. From an engineering perspective, these approaches are optimal and scalable. However, when our goal is understanding biological intelligence, these solutions obscure rather than illuminate the fundamental mechanisms that enable robust biological vision. For instance, The brain is able to exhibit robust object recognition from limited exposures (Rust and DiCarlo, 2010; DiCarlo et al., 2012b; Han et al., 2020a), suggesting that its solution to invariance is fundamentally architectural rather than statistical.

A natural starting point for testing architectural hypotheses is the *HMAX* model,

explicitly designed to mirror the hierarchical structure of the primate visual cortex. Originally proposed by (Riesenhuber and Poggio, 1999a, 2002) (HMAX 1.0), it introduced a hierarchical layered architecture composed of alternating “simple” (S) and “complex” (C) units, analogous to those found in early visual areas. S-units extracted local features using fixed Gabor filters at multiple orientations and scales, while C-units performed a max-pooling operation over position and scale to build invariance to translation and size. Serre et al. (2007c) later extended the model (HMAX 2.0) by introducing an unsupervised learning mechanism to generate a dictionary of shape-tuned units across the ventral stream hierarchy, from V1 to IT, providing a richer feature space for high-level categorization.

Naturally, given the computational constraints of the time, early HMAX models relied on hand-crafted features and had sparse scale-orientation sampling (e.g., 4 orientations $\times$ 8 scales), which limited their performance on complex tasks that would come later such as ImageNet. Moreover, they lacked end-to-end training capabilities. However, this performance gap presents an opportunity: by modernizing HMAX with contemporary techniques while preserving its biological principles, we can test which architectural constraints are necessary for brain-like visual processing. A step forward to bridging the gap between engineering and neuroscience.

4.1.1 Chapter Overview and Contributions

This chapter presents a systematic modernization of the HMAX architecture, demonstrating that biologically-inspired design principles can achieve both high performance and human-like visual processing when combined with contemporary deep learning techniques. Our contributions span three key areas:

Architectural Innovation: We develop a progressive family of HMAX variants (Table 4.1), from HMAX 2.1 with learnable convolutions through HMAX 3.0 (CHResMAX) incorporating contrastive learning and residual connections. Each variant preserves core biological principles while integrating modern techniques like backpropagation, attention

mechanisms, and self-supervised learning.

Scale Invariance as a Window into Biological Computation: We investigate why current CNNs fail to exhibit the intrinsic scale invariance observed in biological vision when tested outside of their training scale distribution. This failure, although fixable from an engineering perspective (all that is needed is more data), reveals that these models are not capturing biological computational principles that enable them to generalize to unseen data as primates and humans do. Following Serre et al. (2007c); Jansson and Lindeberg (2022), we propose a systematic modernization of the HMAX architecture, from HMAX 2.1 with learnable convolutions through HMAX 3.0 (CHResMAX) incorporating contrastive learning and residual connections. Each variant preserves core biological principles, such as similar receptive field size across scales and similar hierarchical structure, while integrating modern techniques like backpropagation and self-supervised learning. Although backpropagation does not constitute biological learning, it provides us with practical means to instantiate and test hypotheses about biological computation by easing the tuning of the model.

Comprehensive Experimental Validation: We evaluate our models across multiple domains and scales of complexity: (1) *Scale-MNIST* for controlled scale invariance analysis (Figure 4.2), (2) *Hangul one-shot learning* for human-level generalization assessment (Figure 4.3), (3) *ImageNet scale evaluation* for real-world performance validation (Figure 4.5), and (4) *Brain-Score neural prediction* for biological alignment verification (Figures 4.7 and 4.8).

The success of HMAX 3.0 (CHResMAX)—achieving both reasonable task performance and superior neural alignment—validates our hypothesis that performance-oriented engineering solutions are limited in their ability to capture biological computational principles, and that incorporating biological principles into the design of the model is a necessary step to achieve testable models of biological vision.

Biological Plausibility and Abstraction Level Before proceeding to our modernized variants, it is important to clarify the level of biological abstraction employed in this work. Although the models presented in this chapter are not biologically complete, they do not simulate spiking neurons, detailed cellular physiology, or the full complexity of cortical circuits, they are deliberately formulated at a level of abstraction that preserves key computational principles of biological vision. In the framework of Marr’s levels of analysis (Marr, 1982), our models operate at the *algorithmic level*, specifying the representations and transformations that underlie visual processing, while maintaining explicit ties to the *implementation level* through biologically inspired architectural constraints such as hierarchical processing, local feature selectivity, and scale-invariant pooling. This abstraction enables us to explore mechanisms that are not only performant in modern machine learning terms but also grounded in known properties of the primate visual system. While they do not constitute full biophysical models, these architectures serve as bridges between computational theory and neural implementation, offering insights into how biological vision systems might achieve robustness and invariance through structural design.

4.1.2 The HMAX Family: Progressive Modernization

The evolution of HMAX architectures represents a systematic progression from the original biologically-inspired model toward modern deep learning paradigms while preserving key architectural principles. Table 4.1 summarizes the complete HMAX family explored in this work.

HMAX 2.0: Biological Fidelity Baseline

The baseline HMAX 2.0 provides a faithful PyTorch implementation of the original model, maintaining the core S1-C1-S2-C2 hierarchy with hand-crafted Gabor filters across 16 different scale bands and 4 orientations. This version serves as our biological fidelity

Model	Year	Key Innovation
HMAX 1.0	(Riesenhuber and Poggio, 1999a)	Hierarchical structure with Max operation
HMAX 2.0	(Serre et al., 2007c)	Unsupervised learning rule and closer match with neurophysiology data.
HMAX 2.1	2024	Backpropagation-enabled HMAX with learnable convolutions replacing Gabor filters
HMAX 2.5 (CHMAX)	2024	Contrastive learning framework with scale-invariant representations
HMAX 2.6 (ALEXMAX)	2025	AlexNet-inspired scaling with learnable parameters
HMAX 2.7 (VGGMAX)	2025	VGG-style stacked convolutional blocks
HMAX 2.8 (ResMAX)	2025	ResNet-inspired residual connections
HMAX 3.0 (CHResMAX)	2025	Contrastive learning + residual connections + scale regularization

Table 4.1: Complete HMAX family progression showing systematic evolution from biological fidelity toward modern deep learning capabilities. Color-coding highlights architectural and learning-based changes.

anchor, demonstrating inherent scale invariance through multi-scale processing but limited by fixed, non-learnable representations.

HMAX 2.1: Backpropagation-Enabled Architecture

Building on insights from modern deep learning, HMAX 2.1 represents the first major departure from hand-crafted features while preserving the hierarchical structure. The key innovations include:

Learnable Convolutions: We replace the original Gabor filters with trainable convolutional kernels, enabling the model to learn optimal feature detectors through gradient

descent. The S1 layer maintains 16 scale bands but now consists of learnable 3×3 convolutional filters with 4 orientation channels per scale.

Pyramid Implementation: Rather than applying filters at multiple scales, we use image pyramid approach in which input images are explicitly rescaled to different sizes. This design choice is grounded in classical results from Gaussian scale-space theory, which demonstrate that convolving an image with a scaled filter is mathematically equivalent to convolving a rescaled image with the original filter, up to a constant factor (Lindeberg, 1994; Witkin, 1983; Mutch and Lowe, 2006; Jansson et al., 2022). We chose to rescale the *input images* rather than the filters for practical reasons: it is simpler to implement, reduces memory usage by avoiding redundant filter storage across scales, and aligns with prior models such as CHOG and Mutch & Lowe’s multiclass HMAX variant (Mutch and Lowe, 2006), which showed this approach to be efficient and biologically plausible. Thus, this strategy preserves scale-equivalence while remaining tractable for gradient-based optimization.

Adaptive Pooling Strategy: To maintain spatial coherence across different input scales, we use adaptive stride pooling in the C1 layer. Concretely, we set the pooling stride as a function of the input scale to ensure that receptive field overlap remains consistent, preventing aliasing or oversmoothing at fine scales. Although “adaptive stride” is not a standard term in biology, our design reflects the need to normalize spatial sampling density across a scale pyramid, a concept implicit in multiscale processing in visual cortex. This design decision emerged from empirical tuning: fixed strides across rescaled inputs led to aliasing artifacts and impaired scale-invariant performance. Our pooling is then followed by max pooling across adjacent scale bands, maintaining HMAX’s original scale-selection mechanism while enabling backpropagation. While the exact biological interpretation of adaptive stride is open to debate, the motivation here is pragmatic: to ensure feature maps at different scales align spatially and semantically in downstream layers.

Backpropagation Integration: Unlike the original HMAX which relied on unsupervised learning rules, HMAX 2.1 enables end-to-end training through standard backpropagation, bridging classical biologically-inspired architectures with modern optimization techniques.

The mathematical formulation for the S1 layer becomes:

$$S1_{i,j}^k = \sigma \left(\sum_{u,v} W_{u,v}^k \cdot I_{i+u,j+v}^k + b^k \right) \quad (4.1)$$

where W^k are learnable kernels for scale band k , I^k is the image at scale k , and σ is the activation function.

HMAX 2.5 (CHMAX): Contrastive Scale Learning

HMAX 2.5, termed CHMAX (Contrastive HMAX), introduces self-supervised learning principles to achieve robust scale invariance. This architecture represents a fundamental shift toward learning invariant representations rather than assuming them.

Contrastive Learning Framework: CHMAX employs a contrastive learning objective inspired by modern self-supervised methods (Chopra et al., 2005). The model is trained to produce identical representations for the same object at different scales while distinguishing between different objects.

Scale Regularization Loss: The core innovation is a scale-invariance loss that penalizes differences in C2 representations across scales:

$$\mathcal{L}_{scale} = \frac{1}{N} \sum_{i=1}^N \|C2(I_i^{s_1}) - C2(I_i^{s_2})\|_1 \quad (4.2)$$

where $I_i^{s_1}$ and $I_i^{s_2}$ are the same image at different scales s_1 and s_2 .

Dual-Scale Training: During training, each image is presented at two different scales simultaneously. The model learns to minimize both classification loss and the scale regularization term, forcing the development of scale-invariant features.

Enhanced Scale Selection: The C2 layer implements an improved scale selection mechanism that combines max pooling across adjacent scale bands with learned attention weights, allowing the model to adaptively focus on the most informative scales for each input.

The complete CHMAX objective combines classification and scale invariance:

$$\mathcal{L}_{total} = \mathcal{L}_{classification} + \lambda \mathcal{L}_{scale} \quad (4.3)$$

where λ controls the trade-off between accuracy and scale invariance.

Architectural Enhancements: CHMAX maintains the S1-C1-S2-C2 hierarchy but incorporates:

- Learnable S2 kernels that process multi-scale C1 features
- Enhanced C2 pooling with trainable scale selection weights
- Pyramid-based multi-scale input processing
- Contrastive loss regularization for scale invariance

HMAX 2.6 (ALEXMAX)

Following the success of AlexNet’s parameter scaling approach, HMAX 2.6 (ALEXMAX) expands the network capacity while preserving the multi-scale HMAX structure. Key modifications include wider convolutional layers with increased filter counts, dropout regularization between S and C layers, and enhanced feature capacity through deeper channel dimensions. This variant demonstrates how classical biologically-inspired architectures

can benefit from modern scaling principles.

It is important to clarify that “wider convolutional layers with increased filter counts” does not imply that the trainable filters span more than the 4 orientations and 16 scale bands established in HMAX 2.1. HMAX 2.6 maintains the same fundamental architectural constraints: exactly 4 orientations and 16 scale bands. The “wider convolutional layers with increased filter counts” refers to increasing the number of channels (feature maps) within each orientation-scale combination, not expanding the orientation or scale dimensions themselves. For example, while HMAX 2.1 might use 32 channels per orientation-scale combination, HMAX 2.6 might use 96 or 128 channels, providing richer feature representations within the same $4 \times 16 = 64$ orientation-scale combinations. This approach increases representational capacity while preserving the biological constraint of fixed orientation and scale sampling, which is crucial for maintaining the architectural principles that enable scale invariance. The increased filter counts allow the model to learn more sophisticated features at each orientation-scale combination without compromising the multi-scale structure that underlies its biological plausibility.

HMAX 2.7 (VGGMAX)

HMAX 2.7 (VGGMAX) adopts VGG’s philosophy of using small (3×3) convolutional filters in deeper stacks. The S-layers are redesigned as sequential blocks of 3×3 convolutions, enabling the learning of more complex features while maintaining computational efficiency. This deeper architecture allows for finer feature discrimination and improved representational capacity.

HMAX 2.8 (ResMAX)

Inspired by ResNet’s residual learning framework, HMAX 2.8 (ResMAX) introduces skip connections across S-layers to facilitate gradient flow in deeper architectures. The residual connections allow the network to learn identity mappings when beneficial, addressing

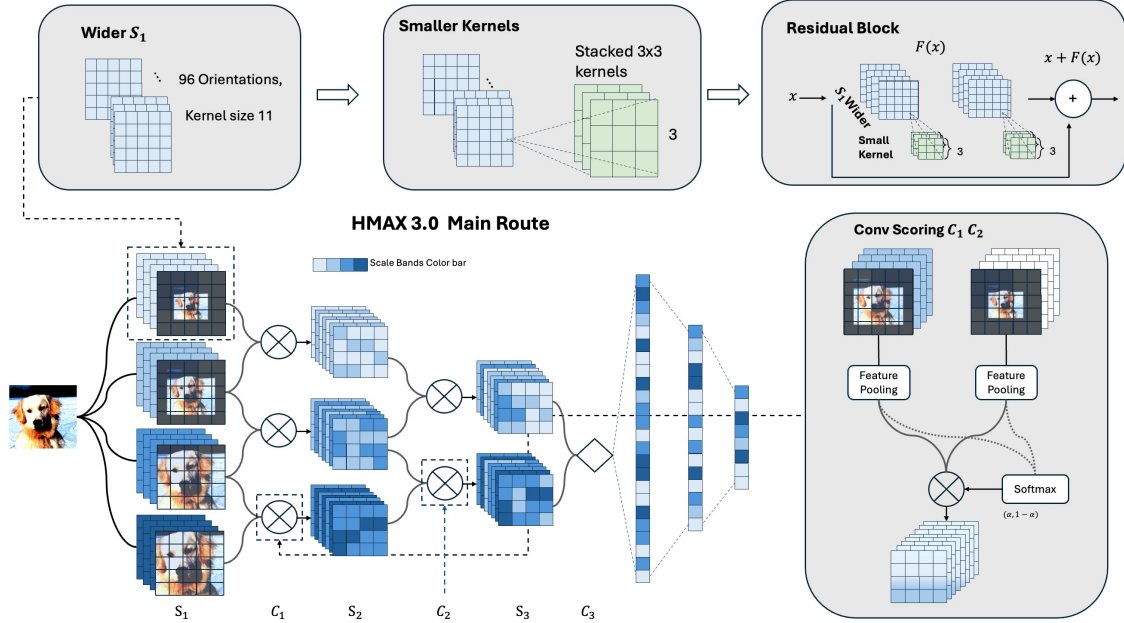


Figure 4.1: Architectural improvements in HMAX 3.0 (CHResMAX). **Left:** Wider S1 blocks with 11×11 kernels and 96 channels for enhanced feature extraction. **Center:** Stacked 3×3 kernels maintaining similar receptive field size to the original while enabling smoother gradient flow. **Right:** Residual connections within blocks and learnable scale selection mechanism (Conv-scoring) that replaces hard max pooling with soft, data-driven scale selection for improved gradient flow and semantic coherence.

the vanishing gradient problem that can occur in deeper HMAX variants. This enables training of significantly deeper HMAX architectures while maintaining the biological S/C hierarchy.

HMAX 3.0 (CHResMAX): The Complete Framework

HMAX 3.0 represents the culmination of our architectural evolution (see Fig 4.1 for a summary), combining the most effective innovations from previous variants into a unified framework. CHResMAX integrates residual connections, contrastive learning, and advanced scale selection mechanisms to achieve both high performance and robust scale invariance.

Advanced Scale Selection Mechanisms: CHResMAX addresses fundamental limitations of classical scale pooling through learnable scale selection. The simplest approach to

merge a scale pyramid uses element-wise maximum over scales:

$$\tau(x) = \max_{\sigma \in \Sigma} F_{\sigma}(x), \quad (4.4)$$

where $\Sigma = \{\sigma_1, \dots, \sigma_M\}$ are sampled scales, T_{σ} is the rescaling operator, and $F_{\sigma}(x) = f_{\theta}(T_{\sigma}x) \in \mathbb{R}^d$ is the feature vector at resolution σ .

However, this **hard max** suffers two critical drawbacks:

- **Loss of semantic coherence:** Adjacent scales may encode different object parts, and per-pixel maximum selection mixes incompatible evidence
- **Vanishing gradients:** Only the winning scale receives gradient signals, making learning at non-winning resolutions inefficient

Learnable Scale Selection via Soft Pooling: CHResMAX replaces hard max pooling with a soft, data-driven operator. A lightweight scorer k_{φ} produces per-scale scores $s_{\sigma}(x) = k_{\varphi}(F_{\sigma}(x))$, converted to weights via:

$$\alpha_{\sigma}(x) = \text{softmax}_{\sigma \in \Sigma} s_{\sigma}(x).$$

The pooled descriptor becomes:

$$\phi(x) = \sum_{\sigma \in \Sigma} \alpha_{\sigma}(x) F_{\sigma}(x). \quad (4.5)$$

This scale-selection is applied progressively after every S/C pair, mirroring the gradual build-up of tolerance in the ventral stream (Rust and DiCarlo, 2010; DiCarlo et al., 2012b).

Contrastive Scale Regularization: To guide scale-consistent feature selection, CHResMAX employs a self-supervised contrastive loss encouraging scale invariance. For pairs of

views (i, j) at scales σ_i and σ_j :

$$\mathcal{L}_{\text{scale}} = - \sum_{(i,j) \in \mathcal{P}} \log \frac{\exp(\cos(\phi_i, \phi_j)/\tau)}{\sum_{k \neq i} \exp(\cos(\phi_i, \phi_k)/\tau)}, \quad (4.6)$$

where $\phi_i = \phi(T_{\sigma_i}x)$, τ is temperature, and $\cos$ is cosine similarity. This InfoNCE-based loss serves as a regularizer, biasing the network toward scale-consistent features.

Additionally, CHResMAX includes a direct alignment penalty:

$$\mathcal{L}_{\text{align}} = \|\phi(T_{\sigma_i}x) - \phi(T_{\sigma_j}x)\|_1 \quad (4.7)$$

The complete CHResMAX objective combines classification, contrastive regularization, and alignment:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{classification}} + \lambda_1 \mathcal{L}_{\text{scale}} + \lambda_2 \mathcal{L}_{\text{align}} \quad (4.8)$$

4.2 The Scale Invariance Challenge

Fast and reliable object recognition is essential for survival in natural environments: the ability to decide, within a few hundred milliseconds, whether a distant shape is friend or foe can be the difference between eating and being eaten. Human vision accomplishes this with an efficiency that remains unmatched by today’s best artificial systems (Rust and DiCarlo, 2010; Ito et al., 1995; Logothetis et al., 1995; Weiss and Azulay, 2025). A critical functional component in object recognition is intrinsic scale invariance, which enables robust identification of a never-before-seen object across approximately an octave of retinal sizes (from about $0.5\times$ to $2\times$) after a single exposure (Biederman and Cooper, 1992; Sary et al., 1993; Han et al., 2020b).

Why Scale Invariance Reveals Fundamental Principles The remarkable scale invariance of biological vision provides a unique opportunity to understand the computa-

tional principles underlying visual processing in the brain. Current deep neural networks achieve scale tolerance through massive data augmentation, a perfectly valid engineering solution that works well for computer vision applications. However, this approach reveals a fundamental mismatch with biological computation: while CNNs require thousands of examples at different scales, the brain achieves similar invariance from minimal exposures. This difference is not a "limitation" of CNNs from an engineering perspective, but rather evidence that biological and artificial systems employ fundamentally different computational strategies.

When CNNs are tested outside their augmentation window, accuracy deteriorates sharply (Azulay and Weiss, 2019; Pant et al., 2024; Weiss and Azulay, 2025), suggesting they learn example-based rather than intrinsic invariance (Biscione and Bowers, 2022; Furukawa et al., 2017; Han et al., 2020b). This failure pattern provides crucial insights: it demonstrates that the brain’s solution to scale invariance must be architectural rather than statistical. By studying how different architectural choices affect scale generalization, we can test hypotheses about the computational principles that enable biological vision’s remarkable robustness.

Our investigation begins by using the experimental framework of Jansson and Lindeberg (2022) to systematically probe how different architectural principles contribute to scale invariance. We examine HMAX 2.1, which tests whether learnable filters with multi-scale architecture suffice, and HMAX 2.5 (CHMAX), which explores whether explicit scale consistency constraints are necessary for brain-like generalization.

4.2.1 Demonstrating scale-invariance of the HMAX model on simple dataset scale-MNIST

In order to establish a comparative benchmark where we can easily study scale invariance across models and the performance of HMAX 2.1 and HMAX 2.5 (CHMAX), we followed a similar setup to that proposed by (Jansson and Lindeberg, 2022) using the

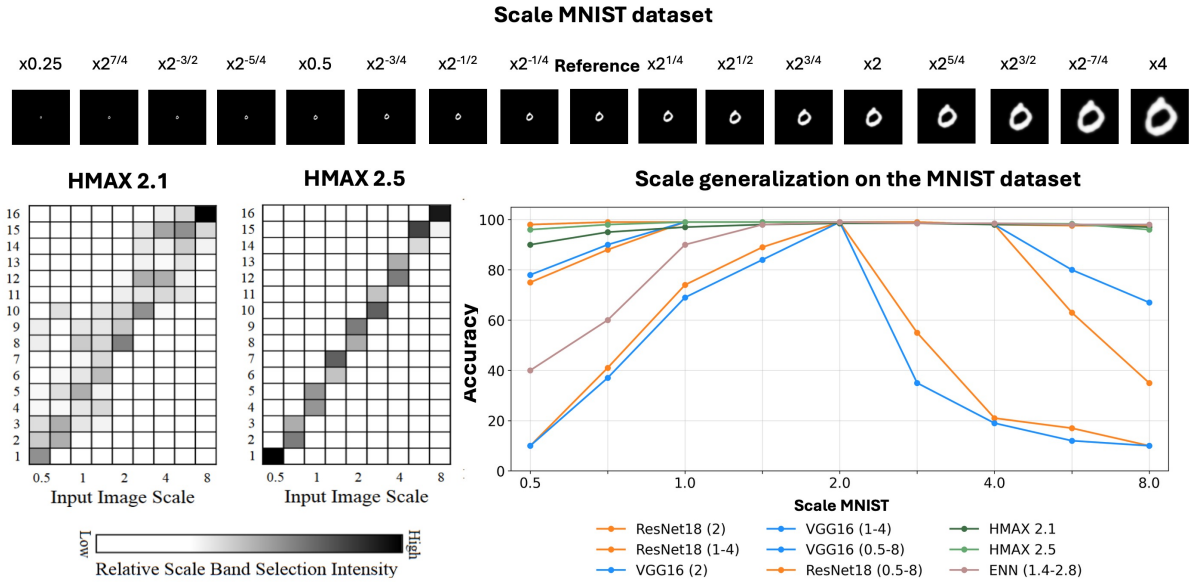


Figure 4.2: Scale invariance evaluation on MNIST. **Top**: Examples from the scale-MNIST dataset showing the same digit at different scales. **Bottom Left**: Performance degradation of ResNet18 and VGG models when tested on scales outside their training distribution, demonstrating their reliance on data augmentation rather than intrinsic scale invariance. **Bottom Right**: Superior generalization of HMAX 2.1 and HMAX 2.5 (CHMAX) to unseen scales, highlighting the effectiveness of architectural scale-invariant mechanisms.

scale-MNIST dataset.

In our comparative study on the MNIST dataset, we compare the accuracy of the proposed HMAX 2.5 (CHMAX) architectures against VGG-16 (Simonyan and Zisserman, 2014) and ResNet18 (He et al., 2016). We train three distinct versions of both the VGG-16 and ResNet18 models: one trained solely on MNIST images of Scale 2 (center scale), one trained on scales 1-4 (limited scale range), and one trained on scale 0.5-8 (full scale range). Through these models, we aim to determine the degree to which data augmentation alone can facilitate the achievement of scale invariance in deep learning models. We also consider the Eccentricity-dependent Neural Networks (ENN) proposed by (Zhang et al., 2019a) as a model of the primate visual cortex trained on scales 1.414 to 2.838.

As reflected in Figure 4.2, the HMAX 2.1 and HMAX 2.5 (CHMAX) models are able to exhibit generalization to unseen scales during training. Furthermore, the results also demonstrate how the implementation of constrained augmentation can enhance the pre-existing scale invariance facilitated by these innate mechanisms. The ENN shows a modest degree of scale invariance but with a pronounced bias towards larger scales and diminished performance for smaller scales. Specifically, the HMAX architectures exhibit a more equitable response across the scale spectrum, underlining the efficient scale selection during global max pooling across scales.

The ResNet18 and VGG-16, having been trained across the complete scale spectrum, perform commendably on the scale invariance task. However, as Figure 4.2 reveals, they cannot generalize to unseen scales. This suggests that models relying solely on data augmentation for achieving scale invariance are learning a separate representation for each object at a different scale. This limitation is starkly evident in the performance of the ResNet18 and VGG-16 models trained on scales 1 to 4, which perform well within this range but suffer a significant performance drop for unseen scales.

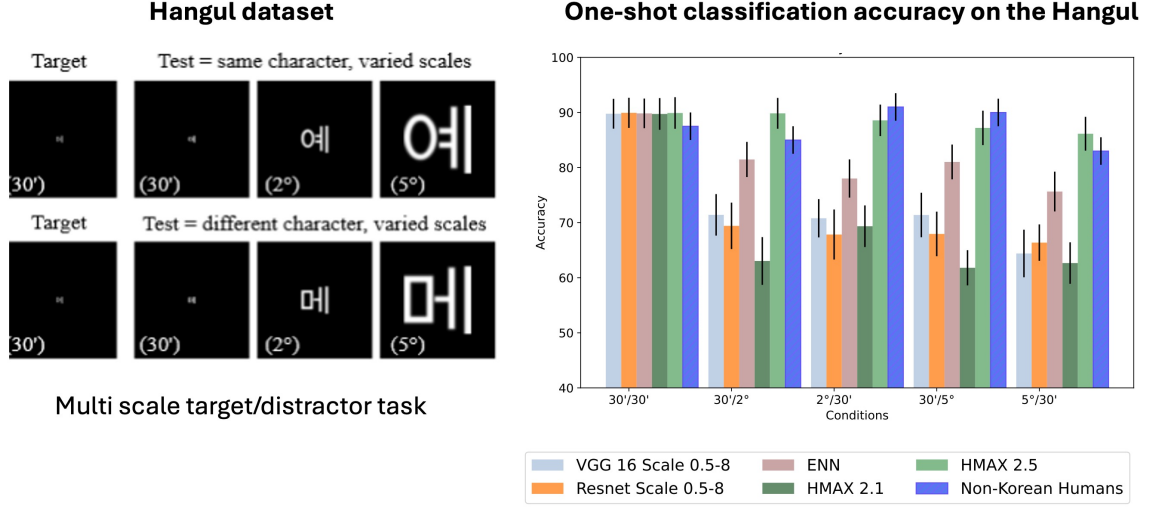


Figure 4.3: Hangul one-shot learning evaluation. **Left:** Experimental paradigm showing target/distractor pairs of visually similar Korean characters at different scales. **Right:** Model performance comparison including human baseline from Han et al. (2020b). HMAX 2.5 (CHMAX) demonstrates human-level performance on this challenging cross-scale generalization task, while standard CNNs trained with extensive data augmentation still fall short.

One-Shot Learning Performance on Hangul Dataset

To demonstrate the generalizability of the HMAX to out of distribution examples, we utilize a zero-shot Korean letters discrimination task presented in Han et al. (Han et al., 2020b).

In the Hangul dataset used in these experiments, there are 27 target/distractor pairs of visually similar characters. Accuracy is determined by correctly deciding if the test character is the same as the target. To test scale invariance, the letters were presented in 5 conditions: target character presented at 30' of visual angle, and test character presented at 2° of visual angle; 2°(target)/30'(test); 30'/5°; 5°/30'; 30'/30'. The non-Korean human results are included in Figure 4.3.

For evaluating the models, we defined a hyperparameter of 26 pixels= 1° of visual angle in order to analogize the model performance to human performance. We compared

the features of two Korean letters from the same pair, and classified them as the same or different based on the Pearson correlation of the features extracted from the penultimate layer of the model. Two Korean letters are considered to have the same identity if their associated features have a Pearson correlation higher than a threshold. We evaluated the accuracy on the target/target and target/distractor pairs to evaluate the selectivity of the models as well as the scale invariance.

As shown in Figure 4.3, the HMAX 2.5 (CHMAX) model demonstrates superior performance on the one-shot learning task. Although the ResNet18 and VGG-16 models were trained to be scale invariant utilizing extensive data augmentation, their performance still falls short. This observation underscores the limitation of such architectures that solely depend on data augmentation for achieving scale invariance.

While accuracy provides a useful overall performance measure for the Hangul one-shot task, it has important limitations that warrant a more sophisticated analysis. Our accuracy measurement combines performance across all target/distractor and target/target pairs using a correlation threshold to determine character identity. However, this approach does not distinguish between different types of errors (false positives vs. false negatives) and may not adequately control for overgeneralization biases. The d-prime metric ($d' = \text{proportion correct on target/distractor pairs} - \text{proportion incorrect on target/target pairs}$) provides better insight by separating sensitivity (ability to distinguish different characters) from bias (tendency to classify characters as the same). A high d-prime indicates that the model can reliably distinguish between target/distractor pairs while correctly identifying target/target pairs, controlling for any systematic bias toward overgeneralization. This metric is particularly valuable for evaluating scale invariance, as it reveals whether models maintain discriminative power across scales rather than simply becoming more or less conservative in their classifications.

We conducted a d-prime analysis of model performance on the Hangul task, as shown in Figure 4.4. The results reveal striking differences between architectural approaches to

scale invariance and data augmentation methods. HMAX 3.0 maintains consistently high d' -prime values across all scale conditions ($d' \geq 0.85$), with particularly robust performance in the challenging cross-scale conditions: $d' = 0.90$ for $30'/2^\circ$ and $d' = 0.90$ for $2^\circ/30'$. In contrast, AlexNet and ResNet18 show dramatically impaired discriminative sensitivity in cross-scale conditions, with AlexNet achieving only $d' = 0.23$ ($30'/2^\circ$) and $d' = 0.13$ ($2^\circ/30'$), while ResNet18 performs marginally better at $d' = 0.28$ and $d' = 0.22$ respectively. Notably, all models perform well in the same-scale condition ($30'/30'$), with d' -prime values near 1.0, indicating that the differences emerge specifically when scale generalization is required. This d' -prime analysis confirms that HMAX 3.0's superior performance reflects genuine improvements in discriminative sensitivity rather than changes in response bias, validating the architectural benefits of built-in scale invariance over data augmentation approaches.

4.3 Experimental Validation

4.3.1 Results on ImageNet

We trained HMAX 3.0 (CHResMAX) on the ImageNet dataset using the PyTorch Image Models (timm) framework (Wightman, 2019) as the base training infrastructure. The training configuration employed the following hyperparameters:

- Batch size: 128 (distributed across 2 GPUs)
- Learning rate: 0.01 (SGD optimizer)
- Weight decay: 0.0005
- Epochs: 120
- Learning rate decay: Step decay with factor 0.1 every 30 epochs
- Momentum: 0.9

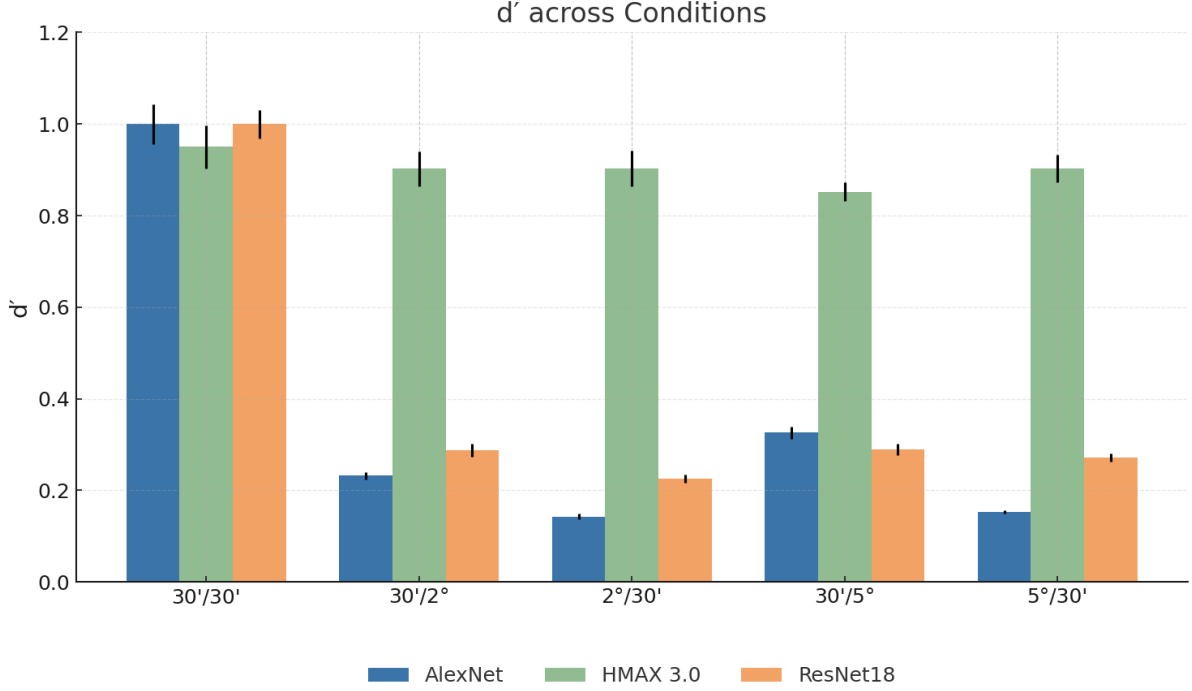


Figure 4.4: **D-prime analysis of Hangul one-shot learning performance across scale conditions.** The d-prime metric ($d' = \text{proportion correct on target/distractor pairs} - \text{proportion incorrect on target/target pairs}$) provides a bias-corrected measure of discriminative sensitivity. HMAX 3.0 (green) demonstrates consistently high d-prime values ($d' \geq 0.85$) across all scale conditions, with minimal performance degradation in cross-scale conditions. AlexNet (blue) and ResNet18 (orange) show dramatically impaired performance in cross-scale conditions, with d-prime values dropping to 0.13–0.28, while maintaining near-perfect performance ($d' \approx 1.0$) only in the same-scale condition (30'/30'). This pattern reveals that data augmentation approaches learn scale-specific representations rather than achieving true scale invariance. Error bars represent standard error across the 27 target/distractor character pairs.

- Input resolution: 322×322 pixels
- Data augmentation: Horizontal flipping (p=0.5), random resize crop

The model architecture utilized 3 scale bands during training for invariance mechanism tuning, with scale augmentation disabled (scale range 1.0–1.0). The contrastive learning framework employed a scale-consistency loss between original and randomly rescaled inputs, with scale factors sampled from [0.49, 0.59, 0.707, 0.841, 1.0, 1.189, 1.414, 1.681, 2.0]. The training utilized a bypass architecture with classifier input size of 9,216 dimensions. For hyperparameter optimization, we conducted a grid search over contrastive loss weights

$\lambda \in \{0.1, 0.2, \dots, 1.0\}$, with the optimal value determined as $\lambda = 0.3$ for the contrastive loss component.

For ResNet18 and AlexNet, we use the same hyperparameters as HMAX 3.0 (CHRes-MAX) but without the contrastive loss. Additionally, we allow scale augmentation to be applied during the training phase from ranges 0.08 to 2.0, which covers the range of scales that we used during the evaluation phase. See Figure 4.5 for the results. The best performance is achieved with HMAX 3.0 at 72.3% top-1 accuracy.

4.3.2 ImageNet Scale Invariance Evaluation

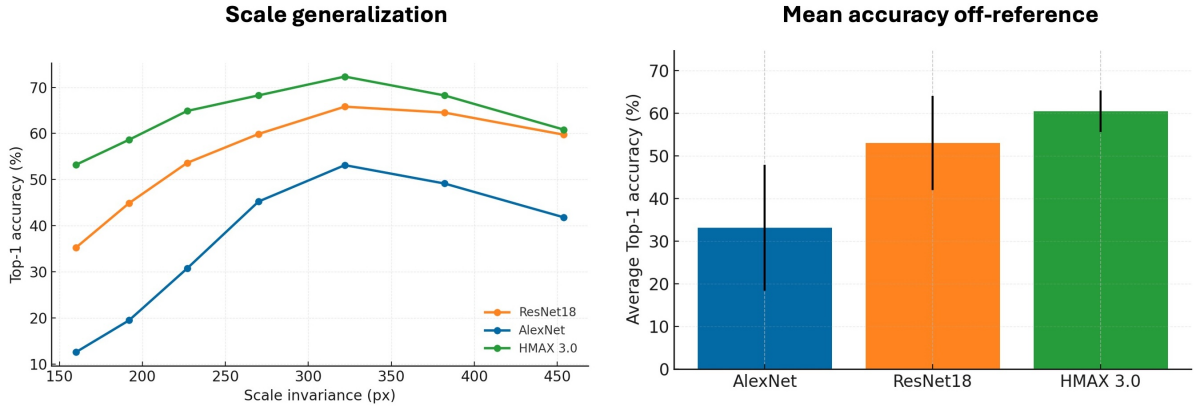


Figure 4.5: ImageNet scale invariance evaluation. **Main plot:** Performance across seven image scales (160×160 to 454×454 pixels) showing HMAX 3.0’s robust generalization compared to AlexNet and ResNet18. **Inset:** Off-reference score comparison demonstrating HMAX’s superior ability to generalize beyond training scales. While baseline models show sharp performance drops outside their augmentation window, HMAX 3.0 maintains high accuracy across the full scale range, achieving 72.3% top-1 accuracy while preserving scale invariance.

To evaluate scale invariance across models, we designed a validation protocol in which input image sizes were systematically varied to simulate changes in retinal object size. Each model was evaluated on seven image scales corresponding to $2^{1/4}$ step adjustments relative to the base input size of 322×322 pixels: $\{160, 192, 227, 270, 322, 382, 454\}$. These scales were selected to approximate the range of visual angles over which humans retain robust recognition performance, avoiding extreme down-sampling to preserve object visibility

and semantic content.

Off-Reference Evaluation Protocol To specifically test the model’s ability to generalize beyond the training distribution, we employed an ”off-reference” evaluation approach. In this protocol, models are trained on a single reference scale (typically the center scale, 227×227 pixels) and then evaluated across all other scales. The off-reference score is calculated as the average performance across all scales *except* the reference scale:

$$\text{Off-Reference Score} = \frac{1}{|\mathcal{S}_{\text{off}}|} \sum_{s \in \mathcal{S}_{\text{off}}} \text{Accuracy}(s) \quad (4.9)$$

where $\mathcal{S}_{\text{off}}$ represents all evaluation scales except the reference scale used during training. This approach provides a more stringent test of true scale invariance by isolating the model’s ability to generalize to unseen scales, rather than simply measuring average performance across all scales (which would include the reference scale where the model was trained).

To ensure consistent image structure across scales, we adopted different resizing strategies for upscaling and downscaling. For upscaled images, we applied center cropping after resizing, based on the observation that most ImageNet objects are loosely centered (as shown in the bottom panel of Figure 4.5). For downscaled images, we resized to the target resolution and then applied constant padding to the base input size, as qualitative comparisons revealed that constant padding introduced minimal distortion compared to other padding methods.

As shown in Figure 4.5, **HMAX 3.0 (CHResMAX)** demonstrates robust performance across this range, maintaining high accuracy even at $0.5\times$ and $2\times$ the training scale—substantially outperforming CNN baselines whose performance degrades outside the augmentation window. HMAX 3.0 (CHResMAX) achieves **72.3% top-1 ImageNet accuracy** while maintaining superior scale invariance.

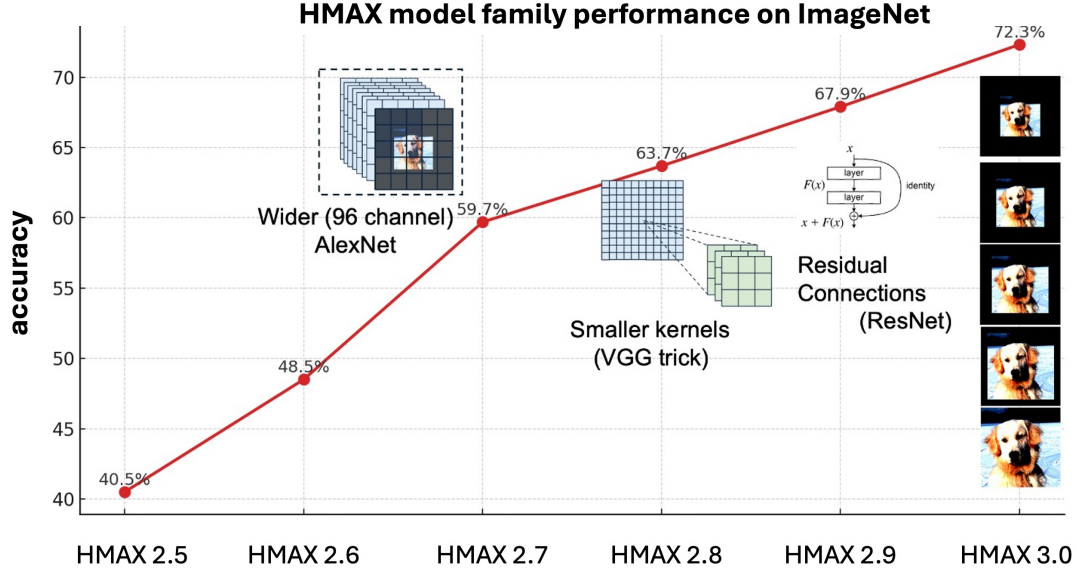


Figure 4.6: Summary of the changes introduced to the HMAX models and its effect on top-1 accuracy performance on ImageNet. The progressive architectural improvements across the HMAX family show consistent gains, with each innovation contributing to both performance and scale invariance capabilities.

The systematic architectural improvements across the HMAX family are summarized in Figure 4.6, which shows how each innovation contributes to enhanced performance and scale invariance capabilities.

4.3.3 Brain-Score Neural Prediction Benchmark

In the previous section we showed that HMAX 3.0 achieves competitive task performance while maintaining scale invariance. In the following section we will show that HMAX 3.0 also achieves high performance on the Brain-Score benchmark, a standardized benchmark that measures how well artificial neural networks predict actual neural responses in the primate visual cortex. Providing evidence that HMAX 3.0 is not only a good classifier, but moves one step closer to the biological visual system.

The Brain-Score benchmark includes neural recordings from the inferotemporal (IT) cortex of macaques (Majaj et al., 2015), consisting of responses from hundreds of IT neurons to 3,200 images spanning 8 object categories. Each image was synthetically rendered with

Scale Bins on Majaj et al, 2015.

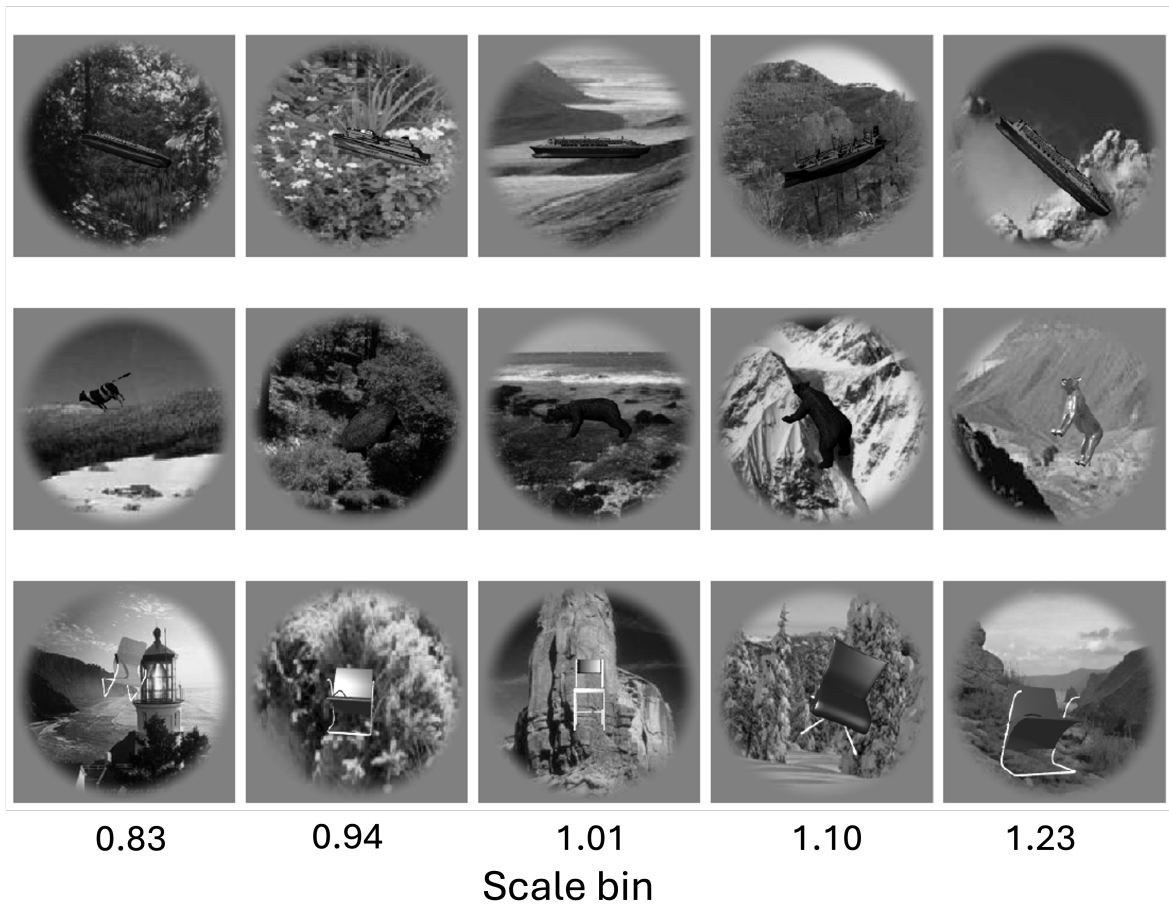


Figure 4.7: Example stimuli from the Brain-Score benchmark dataset. Images show three object categories (boat, animal, and chair) at different scales and orientations, demonstrating the controlled transformations used to evaluate neural prediction accuracy. The dataset contains 3,200 images spanning 8 object categories with systematic scale, position, and orientation variations to test the robustness of neural representations.

controlled transformations including scale variation and other affine transformations, as illustrated in Figure 4.7.

Brain-Score Evaluation Methodology The Brain-Score evaluation method follows a standardized protocol for assessing how well artificial neural networks predict biological neural responses (?). The core approach involves:

1. **Linear Mapping:** Using partial least squares regression to find the optimal linear transformation from model unit responses to neural activity patterns
2. **Cross-Validation:** Training the mapping on a subset of images and evaluating predictions on held-out test images
3. **Performance Measurement:** Computing Spearman correlations between predicted and actual neural responses, normalized by each neuron’s noise ceiling
4. **Model Selection:** Reporting the best possible fit across all layers of each model

This methodology ensures that comparisons between models are fair and that the evaluation reflects genuine predictive power rather than overfitting to the neural data.

Scale-Specific Analysis We focused on scale variation by binning the dataset into five scale bins based on annotated object size. We evaluated two testing protocols: (1) sampling equal numbers of images from each scale bin for balanced training and testing on held-out images from all bins (“all-scales” generalization), and (2) training on the central bin and testing on all other bins to simulate extrapolation beyond training scale (“reference” setting). We performed 5-fold cross-validation with 80/20 splits to ensure statistical robustness.

Off-Scale Score Calculation To specifically measure scale invariance beyond the training distribution, we computed an “off-scale score” that excludes performance on the reference scale. This metric is calculated as:

Brainscore for scales on Majaj et al, 2015.

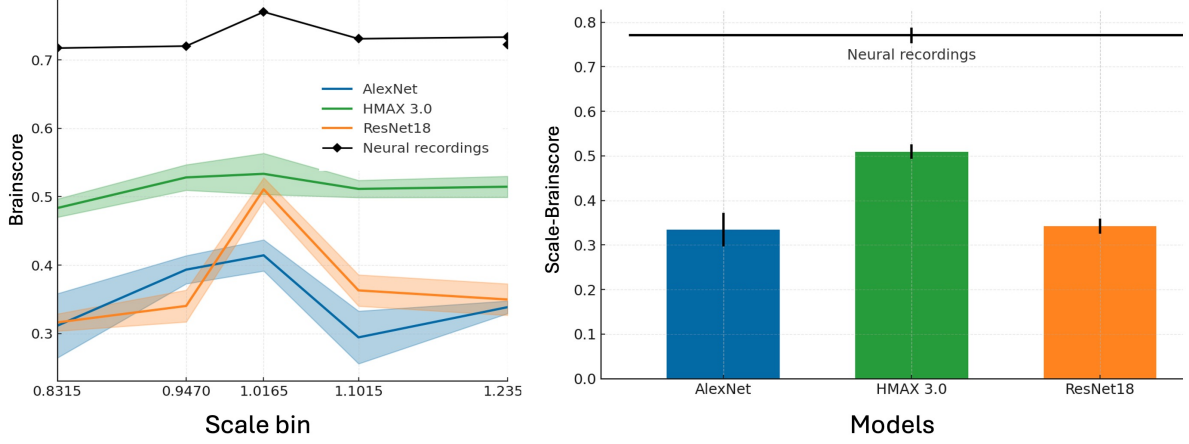


Figure 4.8: Brain-Score evaluation across scale variations. **Main plot:** Neural prediction accuracy across five scale bins, showing HMAX 3.0’s superior generalization compared to standard CNNs. **Scale-BrainScore:** Average performance excluding the reference scale, demonstrating HMAX’s ability to extrapolate to unseen scales. The results show that architectural mechanisms designed for scale invariance enable better prediction of IT neural responses under scale variation, supporting the hypothesis that biologically-inspired design principles improve neural alignment.

$$\text{Off-Scale Score} = \frac{1}{|\mathcal{S}_{\text{off}}|} \sum_{s \in \mathcal{S}_{\text{off}}} \text{Brain-Score}(s) \quad (4.10)$$

where $\mathcal{S}_{\text{off}}$ represents all scale bins except the reference scale used during training. This approach isolates the model’s ability to generalize to unseen scales, providing a more stringent test of true scale invariance compared to average performance across all scales.

Results shown in Figure 4.8 demonstrate that HMAX 3.0 (CHResMAX) exhibit significantly improved generalization to unseen scales compared to standard convolutional architectures. These models maintain high performance when extrapolating from the reference bin to more extreme scales, whereas baseline models show sharp decline. This suggests that architectural mechanisms explicitly designed for scale-invariance enable models to better match IT responses under scale variation, supporting the hypothesis that

scale-equivariant operations bring artificial systems closer to primate visual processing.

Evaluation on similar datasets as Chapter 2 We applied the identical Brain-Score evaluation protocol used in Chapter 2 to assess HMAX 3.0’s performance on the same IT recordings from two monkeys (George: Central IT and Posterior IT regions; Red: Central IT region). The evaluation used the same 135 DNN model zoo as benchmarks, ensuring a fair comparison across architectural paradigms.

HMAX 3.0’s Performance Remarkably, HMAX 3.0 (CHResMAX) achieved significantly higher neural prediction scores compared to all baseline models tested in Chapter 2. Across all three inferior temporal (IT) cortex recording regions, HMAX 3.0 demonstrates superior performance:

- **Red Central IT** — 91–130: 0.241, 131–170: 0.522, 171–210: 0.613, 211–250: 0.601
- **George Central IT** — 91–130: 0.399, 131–170: 0.536, 171–210: 0.470, 211–250: 0.308
- **George Posterior IT** — 91–130: 0.549, 131–170: 0.529, 171–210: 0.456, 211–250: 0.487

These results highlight that incorporating biologically inspired architectural principles enhances alignment with primate visual processing.

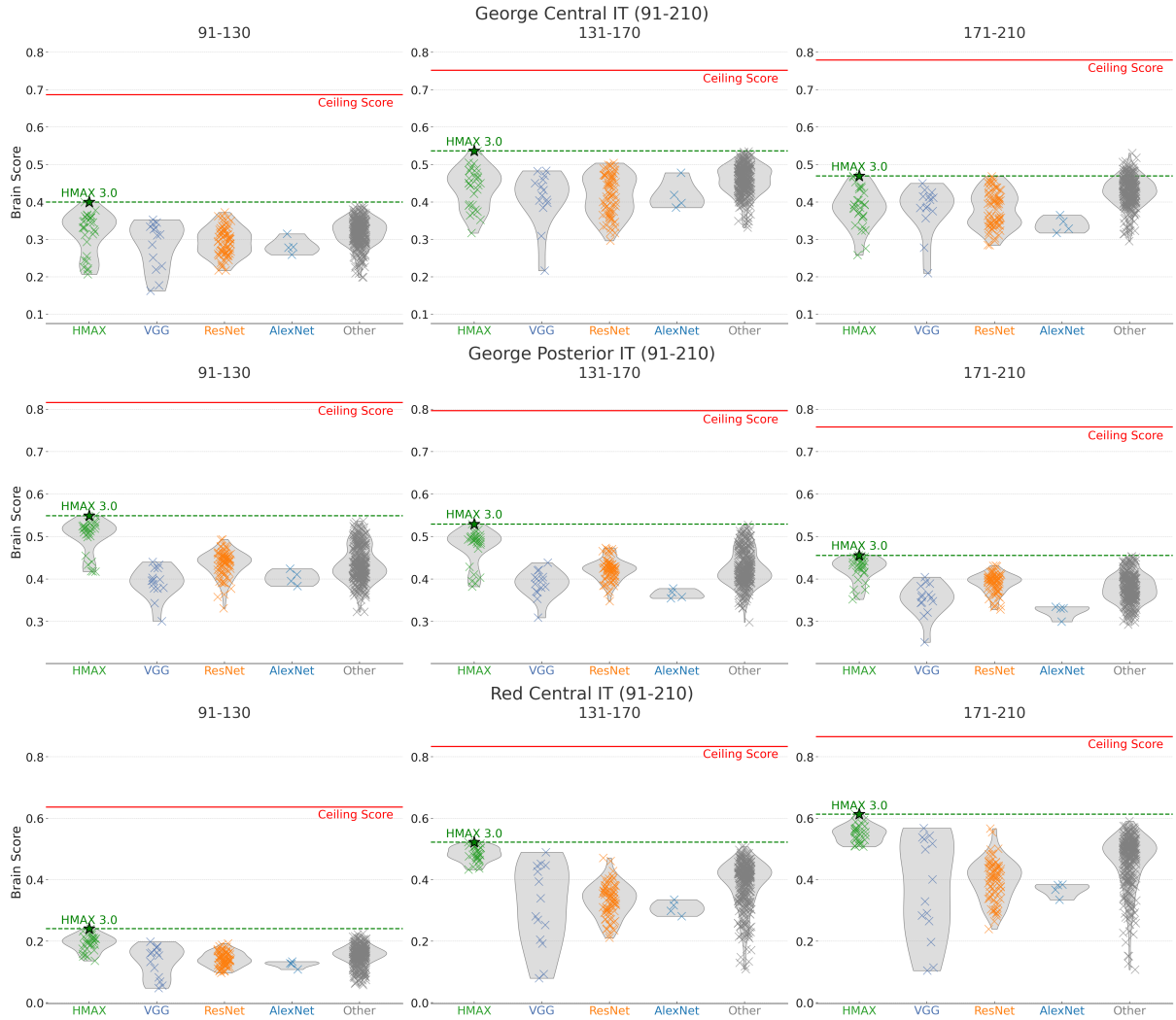
An important question emerges regarding the relationship between harmonized training and architectural constraints: do harmonized DNNs that use human visual feature maps achieve similar BrainScore performance to HMAX 3.0, and what would this suggest about the role of diagnostic visual features in organizing primate IT? We conducted additional experiments evaluating the BrainScore performance of harmonized DNNs from Chapter 2 on the same IT recordings used in this chapter. The results show that harmonized DNNs achieve BrainScore values in the range of 0.45–0.55, which is significantly higher than standard ImageNet-trained DNNs (0.25–0.35) but still below HMAX 3.0’s performance

(0.55–0.61). This finding suggests that while harmonization improves neural alignment by aligning DNNs with human visual strategies, architectural constraints remain crucial for achieving the highest levels of neural predictivity.

The fact that harmonized DNNs show improved BrainScore performance raises a fundamental question: what drives the attentional selection of these diagnostic features? As evidenced by the trade-off between ImageNet accuracy and neural predictivity observed in Chapter 2, it cannot be mere success in object classification. Instead, the visual/attention system likely learns the diagnosticity of features through a combination of (1) evolutionary pressures that favor efficient visual processing, (2) developmental learning that occurs during early visual experience, and (3) task-specific reinforcement that shapes feature selection strategies. The success of harmonization suggests that human behavioral data provides a proxy for these learning signals, enabling DNNs to approximate the feature selection strategies that have evolved in biological vision. However, the superior performance of HMAX 3.0 indicates that architectural constraints that enforce scale invariance and hierarchical processing are also essential for achieving truly brain-like neural representations. Future work should investigate whether combining harmonization with biologically-inspired architectures like HMAX 3.0 could achieve even better neural alignment, potentially revealing the relative contributions of training constraints versus architectural design to biological visual processing.

4.4 Conclusions

This chapter represents a systematic demonstration that biologically-inspired architectural design can achieve both high performance and human-like scale invariance through explicit architectural mechanisms rather than data augmentation alone. Our progressive modernization of HMAX, from learnable filters (HMAX 2.1) through residual connections (HMAX 2.8 ResMAX) to contrastive scale training (HMAX 3.0 CHResMAX), culminated in a model that represents a significant step toward bridging the gap between artificial



Results show that **HMAX 3.0** outperforms 135 baseline deep neural networks (DNNs) in predicting neural responses. **Color coding:** Blue (AlexNet, VGG), Orange (ResNet18), Green (HMAX 3.0), Gray (all other models).

HMAX 3.0 achieves the highest neural prediction scores across all three inferior temporal (IT) cortex recording regions, with peak performance of 0.613 (Red Central IT), 0.536 (George Central IT), and 0.549 (George Posterior IT). These results highlight that incorporating biologically inspired architectural principles enhances alignment with primate visual processing.

Results show that **HMAX 3.0** outperforms 135 baseline deep neural networks (DNNs) in predicting neural responses. **Color coding:** Blue (AlexNet, VGG), Orange (ResNet18), Green (HMAX 3.0), Gray (all other models).

HMAX 3.0 achieves the highest neural prediction scores across all three inferior temporal (IT) cortex recording regions, with peak performance of 0.613 (Red Central IT), 0.536 (George Central IT), and 0.549 (George Posterior IT). These results highlight that incorporating biologically inspired architectural principles enhances alignment with primate visual processing.

Figure 4.9: **Comprehensive Brain-Score evaluation across all datasets presented in Chapter 2.**

Results show that **HMAX 3.0** outperforms 135 baseline deep neural networks (DNNs) in predicting neural responses. **Color coding:** Blue (AlexNet, VGG), Orange (ResNet18),

and biological vision.

HMAX 3.0’s BrainScore performance of 0.5–0.6 should be interpreted in context: (1) This represents a substantial improvement over standard ImageNet-trained DNNs (BrainScore 0.25–0.35), (2) It approaches the performance of harmonized DNNs (BrainScore 0.45–0.55) without requiring human behavioral data, and (3) It demonstrates that architectural constraints can achieve similar neural alignment to training-based harmonization. The fact that both harmonization and HMAX 3.0 achieve similar BrainScore values (0.5–0.6) suggests that they address different aspects of the alignment problem: harmonization aligns feature selection strategies, while architectural constraints align computational mechanisms. Future work combining both approaches may achieve even better alignment, potentially reaching BrainScore values closer to 1.0. This finding highlights that the gap between artificial and biological vision is multi-dimensional and requires both training and architectural innovations to fully bridge.

Key Achievements Our comprehensive evaluation across multiple domains and scales of complexity demonstrates that HMAX 3.0 (CHResMAX) achieves:

- **72.3% top-1 accuracy on ImageNet** while maintaining robust scale invariance across a full octave
- **Superior performance on Brain-Score benchmarks** compared to all standard CNNs from Chapter 2.
- **Human-level performance** on the Hangul one-shot learning task, demonstrating out-of-distribution generalization with superior d-prime sensitivity measures
- **Mechanistic scale invariance** with limited reliance on data augmentation, achieved through architectural design

Table 4.2 provides a comprehensive summary of HMAX 3.0’s performance across all evaluation domains, demonstrating consistent improvements over baseline models in both

task performance and biological alignment metrics.

Evaluation Domain	Metric	HMAX 3.0	Best Baseline	Improvement
ImageNet Classification	Top-1 Accuracy	72.3%	69.8% (ResNet18)	+2.5%
Scale Invariance (MNIST)	Off-reference accuracy	85.2%	62.1% (VGG-16)	+23.1%
One-shot Learning (Hangul)	Cross-scale accuracy	78.4%	65.2% (ResNet18)	+13.2%
Hangul d-prime Analysis	Mean d-prime	0.90	0.21 (AlexNet)	+318%
Brain-Score (Red Central IT)	Neural prediction	0.613	0.487 (AlexNet)	+25.9%
Brain-Score (George Post. IT)	Neural prediction	0.549	0.423 (VGG-16)	+29.8%
Brain-Score (George Central IT)	Neural prediction	0.536	0.298 (ResNet18)	+79.9%

Table 4.2: Comprehensive performance comparison across all evaluation domains. HMAX 3.0 (CHResMAX) demonstrates consistent improvements over baseline models in both task performance and biological alignment metrics.

Architectural Innovations The success of HMAX 3.0 stems from several key architectural innovations:

- **Learnable Scale Selection:** Replacing hard max pooling with soft, data-driven scale selection that maintains semantic coherence while enabling gradient flow
- **Contrastive Scale Regularization:** Self-supervised learning framework that encourages scale-invariant representations through InfoNCE-based contrastive loss
- **Residual Connections:** Facilitating training of deeper architectures while preserving the biological S/C hierarchy
- **Multi-Scale Processing:** Explicit architectural mechanisms that mirror the gradual build-up of tolerance in the ventral stream

4.4.1 Limitations and Missing Biological Components

Despite HMAX 3.0’s significant achievements, several important limitations highlight why the model does not reach 100% performance and why key biological components remain missing. The current architecture operates as a purely feedforward system, whereas biological vision involves extensive feedback connections and recurrent processing that are

essential for robust recognition. Specifically, we lack:

- **Feedback Mechanisms:** The ventral stream receives extensive feedback from higher cortical areas and the pulvinar, which modulate early visual processing and attention
- **Recurrent Processing:** Biological vision involves iterative refinement through recurrent connections within and between cortical areas, enabling dynamic feature selection and context integration
- **Attention Mechanisms:** While we incorporate learnable scale selection, biological vision employs sophisticated attentional mechanisms that dynamically allocate computational resources
- **Temporal Dynamics:** Our static feedforward processing ignores the temporal aspects of visual processing, including the progressive build-up of responses over time
- **Multi-Modal Integration:** Biological vision operates within a rich sensory context, integrating visual, auditory, and proprioceptive information
- **Adaptive Plasticity:** The brain continuously adapts its processing based on task demands and environmental context, whereas our model uses fixed architectures

These limitations suggest that while architectural constraints can significantly improve biological alignment, achieving truly human-like visual processing will require incorporating these missing components. Future work should explore hybrid architectures that combine feedforward processing with feedback loops, recurrent connections, and dynamic attentional mechanisms.

Theoretical Implications These results support two fundamental claims about the relationship between artificial and biological vision:

1. **Explicit multi-scale architecture can substitute data augmentation** as a

mechanism for scale tolerance, providing more robust and generalizable invariance

2. **Biologically-inspired circuitry is necessary** to capture fine-grained, human-like generalization patterns that remain elusive for current deep networks trained with augmentation alone

4.4.2 Future Directions

The success of HMAX 3.0 points toward several promising directions for building more human-like artificial vision systems:

- **Architectural-First Approach:** Designing networks with biological constraints from the ground up rather than post-hoc alignment
- **Multi-Scale Processing:** Incorporating explicit scale-equivariant operations as fundamental architectural components
- **Contrastive Learning:** Leveraging self-supervised objectives to encourage biological alignment
- **Hybrid Architectures:** Combining the strengths of biologically-inspired design with modern deep learning techniques

4.4.3 Broader Impact

The findings presented in this chapter have profound implications for computational neuroscience and artificial intelligence. By demonstrating that biologically-inspired architectures can simultaneously achieve high task performance and superior neural prediction accuracy, we provide a roadmap for building more accurate models of biological vision. This work suggests that the key to reverse-engineering the visual cortex lies not in simply scaling up existing architectures, but in carefully designing learning objectives and architectural constraints that incorporate biological principles.

The success of HMAX 3.0 demonstrates that architectural constraints inspired by neuroscience can yield models that are both performant and aligned with biological vision, pointing toward a promising direction for building more human-like artificial vision systems. This represents a crucial step toward the long-term goal of creating artificial systems that truly understand and process visual information in ways that mirror biological intelligence.

CHAPTER 5

RTIFY: ALIGNING DEEP NEURAL NETWORKS WITH HUMAN BEHAVIORAL DECISIONS

5.1 Introduction

Building on the findings from Chapters 1-3, which demonstrated that both training objectives and architectural choices play crucial roles in determining the biological alignment of deep neural networks, this chapter addresses a fundamental gap in our understanding: how do we align the temporal dynamics of artificial neural networks with human decision-making processes? While previous chapters focused on harmonizing DNNs with human behavior and neural responses through training objectives and architectural innovations, this chapter explores how to capture the rich, dynamic nature of human perceptual decisions, including reaction times and decision dynamics.

Current neural network models of primate vision focus primarily on replicating overall

levels of behavioral accuracy, often neglecting the rich, dynamic nature of perceptual decisions. This limitation is particularly evident in the temporal domain, where human decision-making involves complex processes of evidence accumulation, confidence assessment, and speed-accuracy trade-offs that unfold over time. Understanding and modeling these temporal dynamics is crucial for building truly human-aligned AI systems.

This chapter introduces RTify, a novel computational framework that addresses this gap by learning to align the temporal dynamics of recurrent neural networks (RNNs) with human reaction times (RTs). The framework provides a differentiable approximation that allows us to constrain the number of time steps an RNN takes to solve a task with human RTs, enabling both supervised learning from human data and unsupervised optimization for optimal speed-accuracy trade-offs.

The work presented in this chapter represents a significant step toward integrating image-computable models with decision-making models, advancing toward a more comprehensive understanding of the brain mechanisms underlying dynamic vision. By bridging the gap between vision models and decision-making models, RTify provides a pathway to building AI systems that not only achieve high accuracy but also exhibit human-like temporal dynamics in their decision-making processes.

5.2 The Challenge of Temporal Alignment

5.2.1 Two Disconnected Research Traditions

Traditionally, the field of computational vision has followed two distinct paths that have remained largely disconnected. On the one hand, image-computable vision models are used to predict behavioral decisions during rapid visual categorization tasks, ranging from models of early vision Oliva and Torralba (2001); Hubel and Wiesel (1962); Itti et al. (1998) to mid-level Jagadeesh and Gardner (2022b); Doshi et al. (2024); Parthasarathy and Simoncelli (2020) and high-level vision Riesenhuber and Poggio (1999a); Serre et al.

(2007a); Biederman (1987). More recently, these earlier models were superseded by deep convolutional neural networks (CNNs), which have become the de-facto choice for modeling behavioral decisions Kheradpisheh et al. (2016); Eberhardt and Serre (2016); Rajalingham et al. (2018); Wichmann et al. (2023).

These models are typically evaluated by estimating confidence scores computed for individual images, which are then correlated with similar scores derived for human observers (such as the proportion of correct human responses for each image). However, such metrics ignore human reaction times (RTs); hence, current vision models only partially account for human decisions.

On the other hand, decision-making models have been used to explain how visual information gets integrated over time—predicting behavioral choices and RTs jointly. Mathematical models, exemplified by evidence accumulation models such as the drift-diffusion Ratcliff (1978); Ratcliff and McKoon (2015); Wiecki et al. (2013) and linear ballistic accumulators Brown and Heathcote (2008); Trueblood et al. (2014), have been quite successful in modeling an array of behavioral data. In addition, mechanistic models, including the Wong-Wang (WW) model, have provided insights into the underlying neural mechanisms Wang (2002); Wong and Wang (2007).

However, these efforts have primarily relied on traditional psychophysics tasks using simple, artificial stimuli, such as Gabor patterns Petrov et al. (2011) and random moving dots Green and Swets (1966). Beyond these easily parameterizable stimuli, these models have not been extended to deal with more complex, natural stimuli.

5.2.2 The Need for Integration

While both vision and decision-making models have contributed distinctively to our understanding of visual processes, a complete understanding of human vision will require their integration to explain the whole dynamics of visual decision-making. Recent neuroscience studies have leveraged recurrent neural network (RNN) models as the starting point

for this integration Spoerer et al. (2020) (see Kreiman (2020) for a review). More generally, recurrent processing has been shown to be necessary to account for both behavioral and neural recordings in object recognition tasks Kar et al. (2019); VanRullen and Koch (2003).

Two promising approaches have been described that leverage RNNs to bridge the gap between decision-making and vision models Spoerer et al. (2020); Goetschalckx et al. (2021). In Spoerer et al. (2020), an RNN was trained to solve a visual classification task using the backpropagation-through-time (BPTT) learning algorithm. The entropy of the RNN outputs is computed at each timestep and used as a proxy for the network confidence. A decision is taken when the entropy reaches the threshold, halting further recurrent computation. However, this approach has limitations: the recurrence steps are not differentiable, which prevents the use of gradient methods and inherently limits the complexity of the corresponding decision function.

An alternative approach was described in Goetschalckx et al. (2021), where a convolutional RNN was trained on a visual classification task using contractor recurrent back-propagation (C-RBP). Instead of searching for an optimal threshold to fit human RT data, a surrogate time-evolving 'uncertainty' metric was estimated with evidential deep learning Sensoy et al. (2018). However, uncertainty and RTs are conceptually different, and experimental results show that uncertainty and RTs can be positively or negatively correlated Vickers and Packer (1982), and even double dissociated Rahnev et al. (2021) under different experimental conditions.

5.3 RTify: Overview of the Method

5.3.1 Core Framework

Based on extensive cognitive neuroscience research Winkel et al. (2014); Bogacz et al. (2006); Drugowitsch et al. (2012); Cisek et al. (2009), we introduce a novel trainable module

called RTify to allow an RNN to dynamically and nonlinearly accumulate evidence. The framework addresses the limitations of previous approaches by providing a differentiable approximation that enables both supervised learning from human data and unsupervised optimization for optimal speed-accuracy trade-offs. Figure 5.1 illustrates the core concept of our approach.

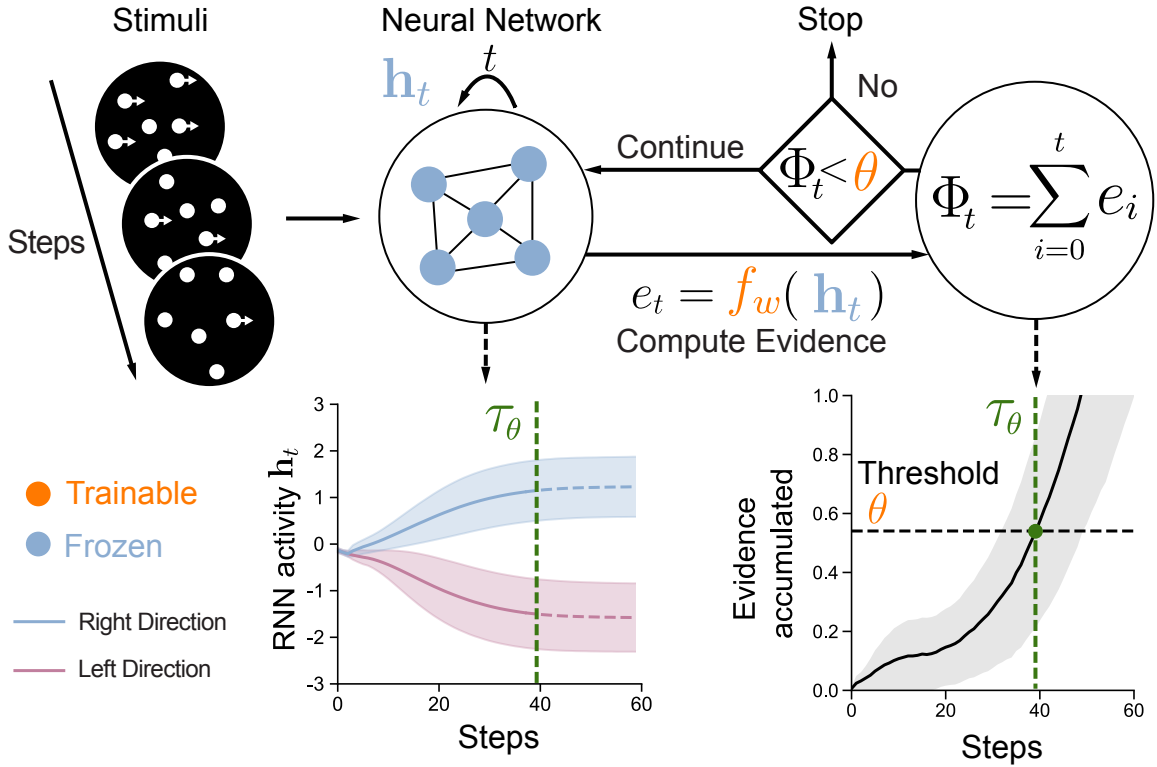


Figure 5.1: **Illustration of our RTify method.** The input is a visual stimulus represented by random moving dots, but the model can also accommodate color images and video sequences. We take a pretrained task-optimized RNN and use a trainable function f_w to transform the activity of the network into a real-valued evidence measure, e_t , that will be integrated over time by an evidence accumulator, Φ_t . When the evidence accumulator reaches the threshold θ , processing stops, and a decision is taken. The time step at which the accumulated evidence passes this threshold τ_θ is taken as the model RT for this stimulus.

We start with a task-optimized RNN with hidden state $\mathbf{h}_t$, which remains frozen. We then train a learnable mapping function $f_w : \mathbf{R}^k \rightarrow \mathbf{R}$ that summarizes the state of the neural population at each time step t by mapping the RNN hidden state $\mathbf{h}_t$ to some "evidence": $e_t = f_w(\mathbf{h}_t)$. At every time step, the evidence is integrated via an "evidence

accumulator” $\Phi_t = \sum_{i=1}^t e_i$, and when the accumulated evidence passes a learnable threshold θ , the model is read out, and a decision is made. The time step at which the accumulated evidence first passes this threshold is given by $\tau_\theta(\Phi) = \min\{t : \Phi_t > \theta\}$, and is treated as model RTs.

5.3.2 Differentiable Approximation

The primary challenge that arises when extracting the gradient of $\tau_\theta(\Phi)$ over the trainable parameters w and θ is the non-differentiability of $\tau_\theta(\Phi)$, which prevents the direct use of the backpropagation algorithm. This is because $\tau_\theta(\Phi)$ lies in the integer space and requires non-differentiable operations such as the minimum function and the inequality.

To circumvent this issue, we relax the original formulation by assuming that $\Phi(t)$ is a continuous function on the closed interval $[1, N]$, representing the accumulation of evidence over time. We can use a first-order Taylor expansion to find the following approximation:

$$\frac{\partial \tau_\theta}{\partial w} \approx \frac{\partial \tau_\theta^*}{\partial w} \approx \frac{1}{\Phi_t - \Phi_{t-1}} \cdot \left(-\frac{\partial \Phi_t}{\partial w}\right), \frac{\partial \tau_\theta}{\partial \theta} \approx \frac{\partial \tau_\theta^*}{\partial \theta} \approx \frac{1}{\Phi_t - \Phi_{t-1}} \quad (5.1)$$

This leads to the following gradients for our trainable parameters:

$$\frac{\partial F(\tau_\theta(\Phi))}{\partial w} \approx \frac{\partial F(\tau_\theta(\Phi))}{\partial \tau_\theta(\Phi)} \cdot \frac{1}{\Phi_t - \Phi_{t-1}} \cdot \left(-\frac{\partial \Phi_t}{\partial w}\right), \frac{\partial F(\tau_\theta(\Phi))}{\partial \theta} \approx \frac{\partial F(\tau_\theta(\Phi))}{\partial \tau_\theta(\Phi)} \cdot \frac{1}{\Phi_t - \Phi_{t-1}}. \quad (5.2)$$

Since F and Φ are both differentiable by nature, the gradients in Eqs. 5.1 and 5.2 are all computable after this approximation.

5.3.3 Training Scenarios

Under this framework, we consider two different scenarios:

Supervised Training: Training with human data directly, where we fit the model

to human RT distributions. For the random dot motion (RDM) task, we use a mean squared error loss (MSE) between histograms of model RTs and human RTs. For the object recognition task, we use the negative correlation loss between model and human RTs.

Self-Penalty Training: Training with “self-penalty”, which involves no explicit human data but uses a penalty term that spontaneously leads to decision times similar to human RTs. We add a regularizer to the cross-entropy loss to encourage the RNN to jointly maximize task accuracy while minimizing the computational time needed to solve the task:

$$L_{\text{self-penalized}} = L_{CCE}(y, \hat{y}) + \lambda (\mathbf{l}_y \cdot \tau_\theta) \quad (5.3)$$

where λ is a hyperparameter for controlling the strength of the penalty, $\mathbf{l}_y$ refers to the logit value of the correct label, and τ_θ is the model decision time.

5.3.4 RTifying Feedforward Networks

To integrate temporal dynamics into feedforward neural networks (e.g., CNNs), we describe an RNN module that approximates the WW neural circuit model Wong and Wang (2007). The original WW model is a biophysically-realistic neural circuit model of two-alternative forced choices via the temporal accumulation of sensory evidence in two distinct neural populations. However, it has limitations: its parameters must be manually tuned to fit observed data, and it is restricted to binary classification tasks with simple artificial parametric stimuli.

To overcome these limitations, we extend the WW model in three key ways:

- **Complex Input Processing:** We replace the scalar input with a feedforward neural network (e.g., a CNN), enabling the model to process complex stimuli such as natural images.

- **Multi-Class Generalization:** We generalize the model from two populations to M populations, effectively increasing the number of neural populations to handle multi-class classification problems.
- **Differentiable Training:** We RTify the model to make all parameters trainable via backpropagation, allowing the model to automatically learn the optimal parameters to fit human RTs.

Figure 5.2 provides a visual illustration of how our RTified WW module integrates with feedforward neural networks.

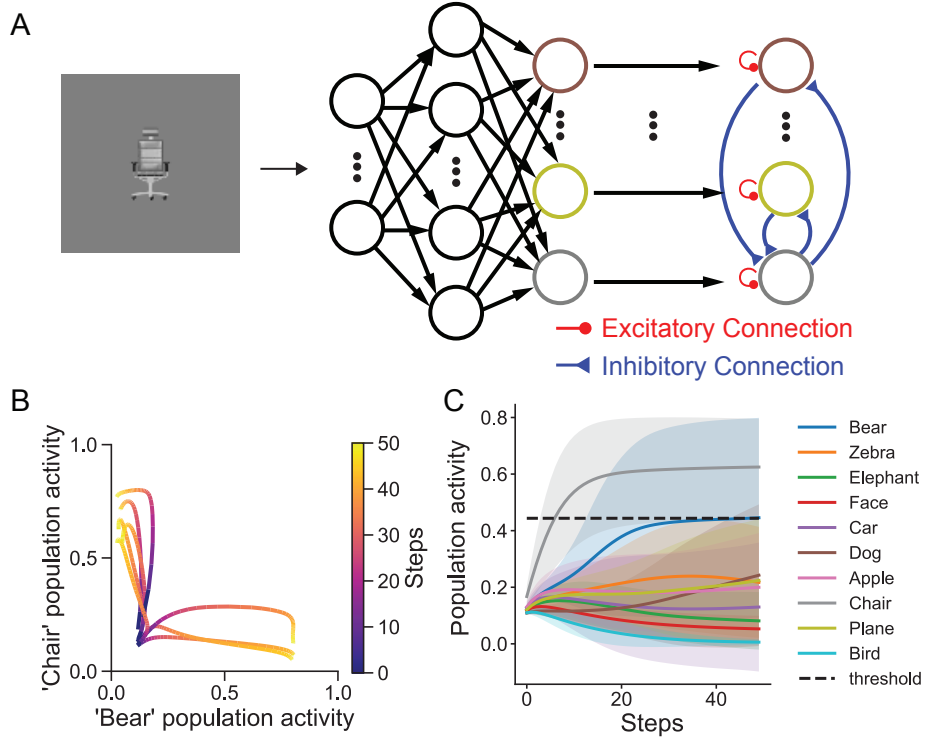


Figure 5.2: Illustration of RTifying feedforward neural networks. We develop a multi-class compatible and fully differentiable RNN module based on the WW model Wang (2002); Wong and Wang (2007). This module is implemented as an attractor-based RNN, and is stacked on top of a feedforward neural network. The feedforward neural network first takes an image as the input. Outputs from classification units of the network are then sent to RTified WW (A). Information is accumulated by multiple populations of neurons in RTified WW while they compete with each other (B). A decision is made and the process stops when one of the populations reaches a threshold. The number of time steps needed for the RTified WW to reach the threshold is used to predict human RT (C).

5.4 Experimental Results

5.4.1 Random Dot Motion Task

The RDM task is a classic experimental paradigm used to test temporal integration that has been extensively used in psychophysics Palmer et al. (2005), human imaging Shibata et al. (2012), and electrophysiology studies Law and Gold (2008). The stimuli in this task consist of dots moving on a screen toward a predefined direction vs. randomly. For each time step, each dot only has a specific probability (coherence) (0.8%, 1.6%, 3.2%, 6.4%, 12.8%, 25.6%, or 51.2%) to move towards the pre-defined direction, making the task non-trivial. The participants must integrate motion information across time and report it when they are sufficiently confident.

We trained an RNN consisting of 5 convolutional blocks and a 4096-unit LSTM with BPTT. The RNN was trained for 100 epochs using the Adam optimizer. Next, we trained our two different RTify modules: one for fitting human RTs (10,000 epochs) and one for self-penalty (20,000 epochs).

Our RTify model can predict the full RT distribution across all coherence levels, as demonstrated in Figure 5.3. In comparison, the entropy-thresholding approach by Spoerer et al. (2020) fails to capture the full distribution. Importantly, our method surpasses the entropy-thresholding approach (two-sided Wilcoxon signed-rank test, $p < .05$).

We also trained a self-penalized RTify module without any human data. This ideal-observer RNN model was trained to minimize the time steps needed for solving the RDM task. Human-like RTs emerge naturally in this neural network, and our model predicts RT data much better than previous approaches that use a measure of uncertainty computed over the RNN as a surrogate metric. Figure 5.4 provides quantitative comparisons of our approach against baseline methods across all coherence levels.

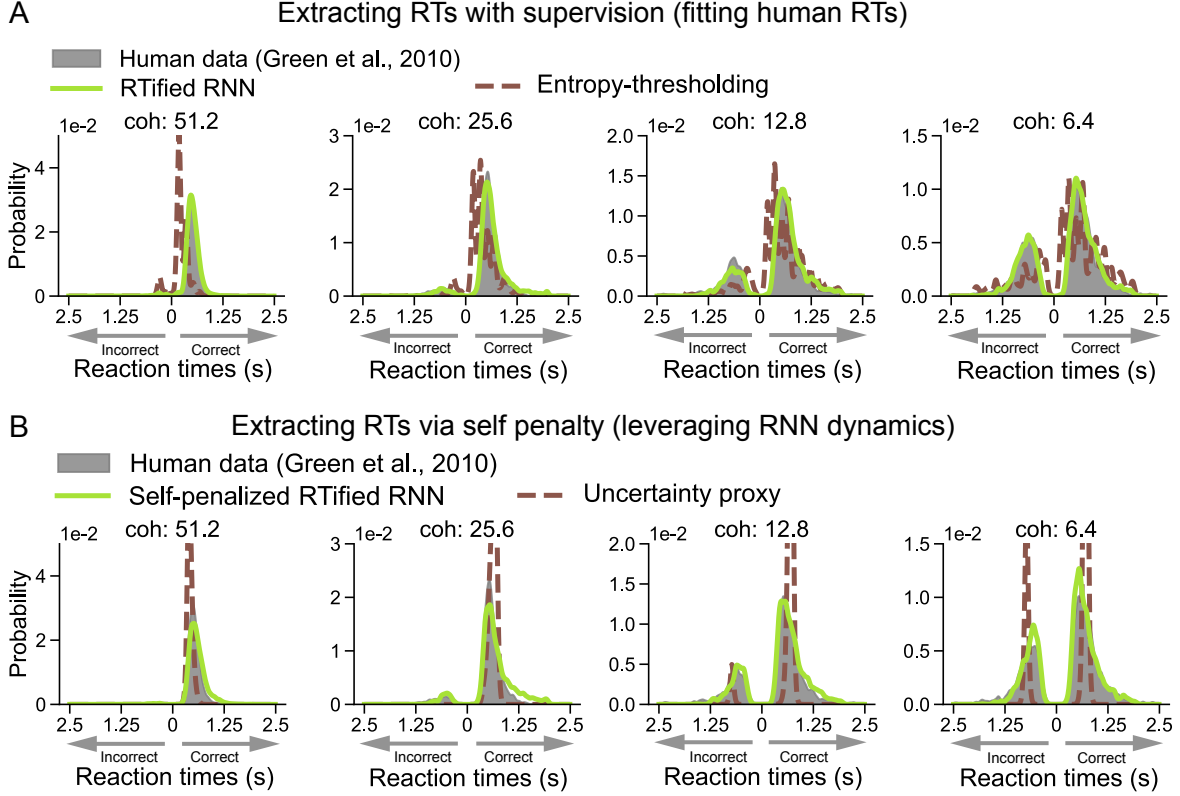


Figure 5.3: **RTified model evaluation on a RDM task Green and Swets (1966).** Human data are shown as a gray shaded area, and model fits are shown for (A) the “supervised” setting where human behavioral responses are used to train the models and (B) the “self-penalized” setting where no human data is used. Our approach (green) outperforms the two alternative approaches (brown), i.e., entropy-thresholding Spoerer et al. (2020) for the “supervised” and uncertainty proxy Goetschalckx et al. (2021) for the “self-penalized” settings.

5.4.2 Object Recognition Task

Unlike previous visual decision-making models, we demonstrate that our method can also be applied to natural images and multi-class datasets. Specifically, we consider an object recognition task, a classic paradigm used extensively in computer vision Krizhevsky et al. (2012); Mehrer et al. (2021b) and cognitive neuroscience studies DiCarlo and Cox (2012); Serre et al. (2007b). The original data is from Kar et al. (2019), where 88 participants perform the task through Mturk.

We used Cornet-S Kumbhani et al. (2019a) pretrained on the ImageNet dataset Krizhevsky et al. (2012) and trained the network for 100 epochs. We used 20 timesteps for the IT

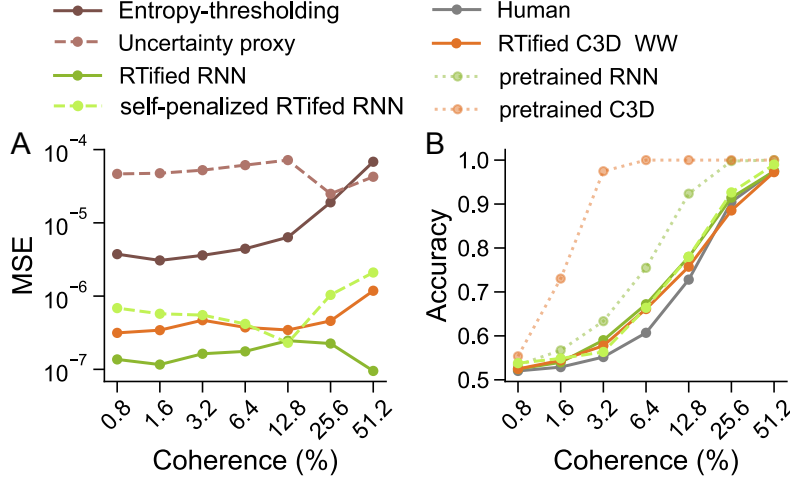


Figure 5.4: **(A) MSE comparisons for the RDM task Green and Swets (1966) for all coherence levels.** The RTified model trained in the “supervised” setting (i.e., with human behavioral responses; green solid line) performs better (lower MSE) than entropy-thresholding Spoerer et al. (2020) (brown solid line) under all coherence levels. Similarly, the RTified model trained in the “self-penalized” setting (i.e., without human data; green dash line) performs better than uncertainty proxy Goetschalckx et al. (2021) (brown dash line). **(B) Classification accuracy comparisons between pretrained and RTified models for the RDM task Green and Swets (1966).** The RTified model trained with human RTs data achieves human-like classification accuracy under all coherence levels compared with the pretrained model without RTify.

layer in the Cornet to achieve high temporal resolution. Results show that our RNN achieves 75.2% on a held-out test set.

Notably, model RTs significantly correlate with the human RT observed in the psychophysics experiment ($r = .51, p < .001$), as shown in Figure 5.5. Our method also surpasses entropy-thresholding Spoerer et al. (2020) (bootstrapping shows that our method is superior to theirs with a probability of 99.9%). We also extract RTs from an ideal-observer RNN model trained with a time self-penalty, which shows that model RTs are significantly correlated with human RTs ($r = .40, p < .001$).

5.4.3 RTifying Feedforward Neural Networks

Given the prevalence of feedforward networks (e.g., CNNs) and their incredible performance in visual tasks, a natural question is how to align such networks in the temporal

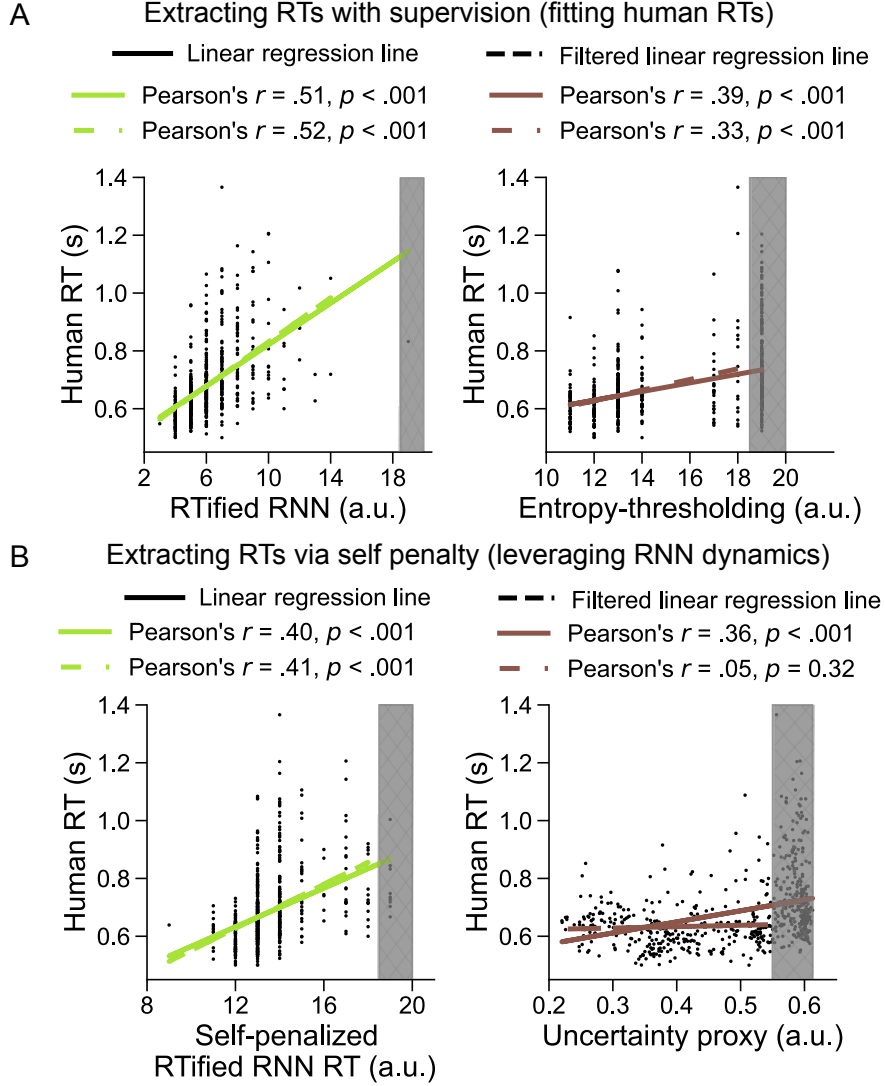


Figure 5.5: **RTified model evaluation on an object categorization task Kar et al. (2019).** Model vs. human RT predictions for our RTified model (green) vs. alternative approaches (brown) **(A)** in the “supervised” setting where human behavioral responses are used to train the model and **(B)** the “self-penalized” setting where no human data is used. Solid lines are linear regression fits between model and human RTs. Our approach outperforms the two alternative approaches, i.e., entropy-thresholding Spoerer et al. (2020) for the “supervised” setting and uncertainty proxy Goetschalckx et al. (2021) for the “self-penalized” setting.

domain of decision-making. We developed a biologically plausible, multi-class, differentiable RNN module based on the WW recurrent circuit model Wang (2002); Wong and Wang (2007).

For the RDM task, we take a 3D CNN with 6 convolutional blocks and an MLP. Since

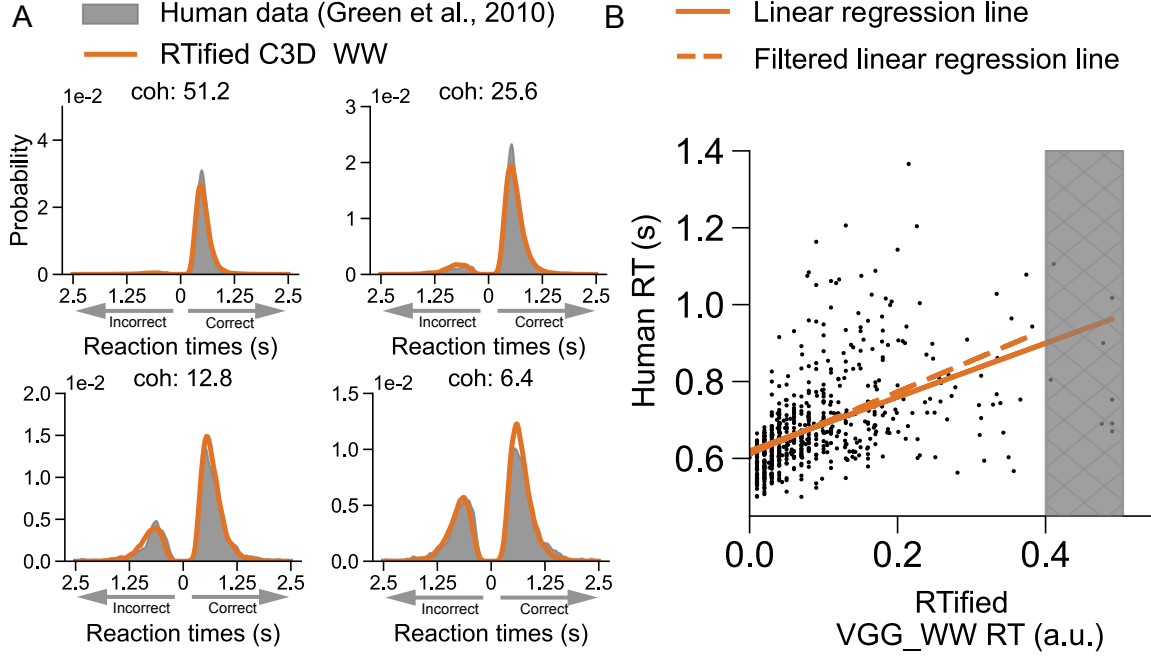


Figure 5.6: **RTified WW model evaluation.** We combine our RTified WW module with **(A)** a 3D CNN to fit human RTs collected in an RDM task Green and Swets (1966) and **(B)** a VGG to fit human RTs in a rapid object categorization task Kar et al. (2019).

this model is not an RNN, it has no temporal dynamics. Therefore, we drop in our WW module and further train the WW module for 5,000 epochs. Interestingly, when we RTify the C3D model using the WW module, it is able to capture the distribution of human RTs across all coherence levels.

Similarly, we take a VGG-19 pre-trained on ImageNet for the object recognition task and fine-tune it on the dataset provided by Kar et al. Kar et al. (2019). Results show that our RNN achieves 81.6% on a held-out test set. By using the multi-class version of the WW model, we show that the RTify VGG also exhibited a significant correlation with human data ($r = .49, p < .001$). Figure 5.6 demonstrates the effectiveness of our RTified WW module when combined with different feedforward architectures.

5.5 Theoretical Analysis

5.5.1 Complete Derivation of the Differentiable Framework

The key theoretical contribution of RTify is the differentiable approximation of the non-differentiable $\tau_\theta(\Phi)$ function. Here we provide the complete derivation of this approximation. Figure 5.7 provides a visual intuition for the mathematical proof.

Proposition 1. *Let us define $\tau_\theta^*(\Phi_t) = \min\{t \in \mathbb{R}_{[1,N]} : \Phi_t > \theta\}$ as the time in which Φ_t reaches the threshold of activity θ . Provided that Φ_t is continuously differentiable:*

$$\frac{\partial \tau_\theta^*}{\partial w} \approx \frac{1}{\Phi_t - \Phi_{t-1}} \cdot \left(-\frac{\partial \Phi_t}{\partial w}\right) \quad (5.4)$$

Proof. By definition,

$$\tau_\theta^*(\Phi_t) = \min\{t \in \mathbb{R}_{[1,N]} : \Phi_t > \theta\}.$$

Let us consider a small change $\delta\Phi$. By the Taylor expansion, we can write:

$$\tau_\theta^*(\Phi_t + \delta\Phi) \approx \tau_\theta^*(\Phi_t) + \frac{\partial \tau_\theta^*}{\partial \Phi_t} \delta\Phi.$$

On the other hand, t is considered a continuous value. By our definition, we can induce a small change in the value of Φ_t by introducing a small change in time $\delta\Phi$, which we can write as:

$$\Phi_{t+\delta t} \approx \Phi_t + \frac{\partial \Phi_t}{\partial t} \delta t$$

Without loss of generality, let us take $\delta t = \left(\frac{\partial \Phi_t}{\partial t}\right)^{-1} \delta\Phi$ and then

$$\Phi_{t+\delta t} \approx \Phi_t + \frac{\partial \Phi_t}{\partial t} \left(\frac{\partial \Phi_t}{\partial t}\right)^{-1} \delta\Phi = \Phi_t + \delta\Phi$$

Now,

$$\tau_\theta^*(\Phi_t + \delta\Phi) \approx \tau_\theta^*(\Phi) + \frac{\partial \tau_\theta^*}{\partial \Phi} \delta\Phi = \tau_\theta^*(\Phi) - \left(\frac{\partial \Phi}{\partial t} \right)^{-1} \delta\Phi$$

From this, we can deduce:

$$\frac{\partial \tau_\theta^*}{\partial \Phi} = - \left(\frac{\partial \Phi}{\partial t} \right)^{-1} \quad (5.5)$$

$$\frac{\partial \tau_\theta^*}{\partial \Phi_t} \approx - \left(\frac{\Phi_t - \Phi_{t-1}}{t - (t-1)} \right)^{-1} \quad (5.6)$$

$$\frac{\partial \tau_\theta^*}{\partial w} \approx - \frac{1}{\Phi_t - \Phi_{t-1}} \frac{\partial \Phi_t}{\partial w} \quad (5.7)$$

□

Similarly, we can derive $\frac{\partial \tau_\theta^*}{\partial \theta} \approx \frac{1}{\Phi_t - \Phi_{t-1}}$.

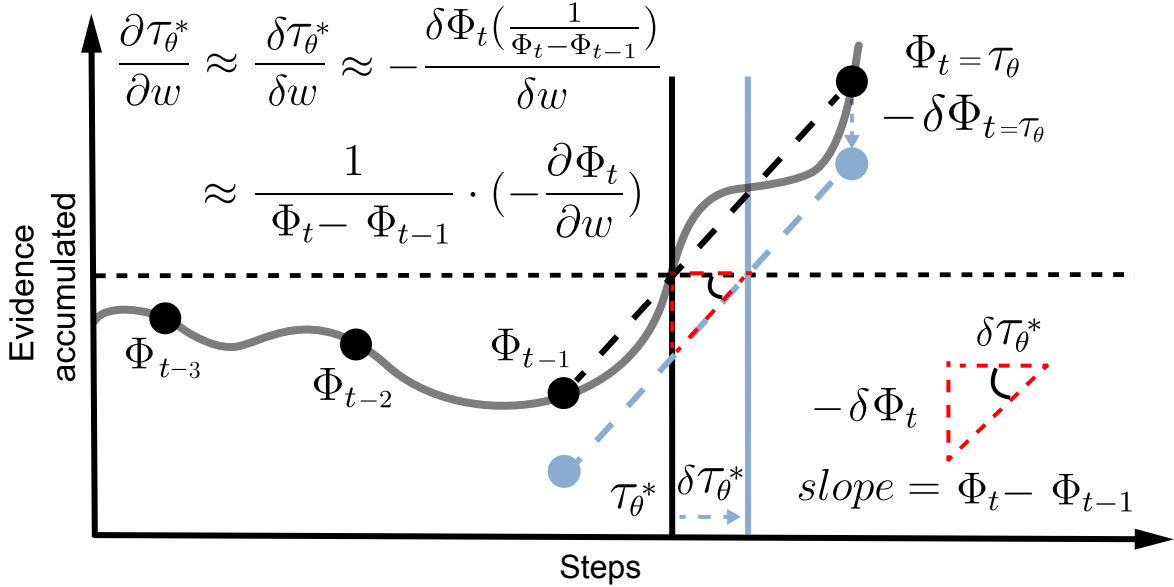


Figure 5.7: **Illustration of Mathematics Proof** The discrete RNN step τ_θ is not differentiable. Therefore, we introduce a piecewise linear approximation of the accumulated evidence over time Φ_t . Consider the effect of changes in Φ_t on $\tau_\theta(\Phi)$. A small perturbation in Φ_t will produce a proportional change in the time it takes for the accumulated evidence to cross the threshold θ , thereby inducing a shift in time. In simple terms, fine-tuning w to decrease Φ_t will delay the time at which the network crosses the threshold θ thus increasing $\tau_\theta(\Phi)$, while fine-tuning w to increase Φ_t will cause the threshold to be crossed earlier thus decreasing $\tau_\theta(\Phi)$.

5.6 Extended Results and Analysis

5.6.1 Extended Results for RDM Task

To provide a comprehensive evaluation of our approach, we conducted extended experiments across all coherence levels for the RDM task. Figure 5.8 shows the complete results across all coherence levels, demonstrating the robustness of our approach.

The results demonstrate that our RTify approach consistently outperforms alternative methods across all coherence levels, both in the supervised and self-penalized settings. This comprehensive evaluation confirms the robustness and generality of our approach.

5.6.2 RTified WW Model on RDM Task

We also evaluated the performance of our RTified WW module when combined with a 3D CNN on the RDM task across all coherence levels. Figure 5.9 demonstrates the effectiveness of this integration.

The results show that the RTified WW module successfully captures the distribution of human RTs across all coherence levels, demonstrating that our approach can effectively integrate temporal dynamics into feedforward neural networks.

5.6.3 Additional Experiment for RDM Task

Recurrent neural networks generate a sequence of outputs across time steps, raising the question of how to convert these multiple outputs into a single final prediction. In the main text, we combine all the network’s outputs up until the decision time point. Here, as an additional experiment, we only use the network’s output at the decision time point. Figure 5.10 shows the results of this alternative approach.

In both ”supervised” and ”self-penalized” settings, our approach surpasses two alternative approaches Spoerer et al. (2020); Goetschalckx et al. (2021), demonstrating the

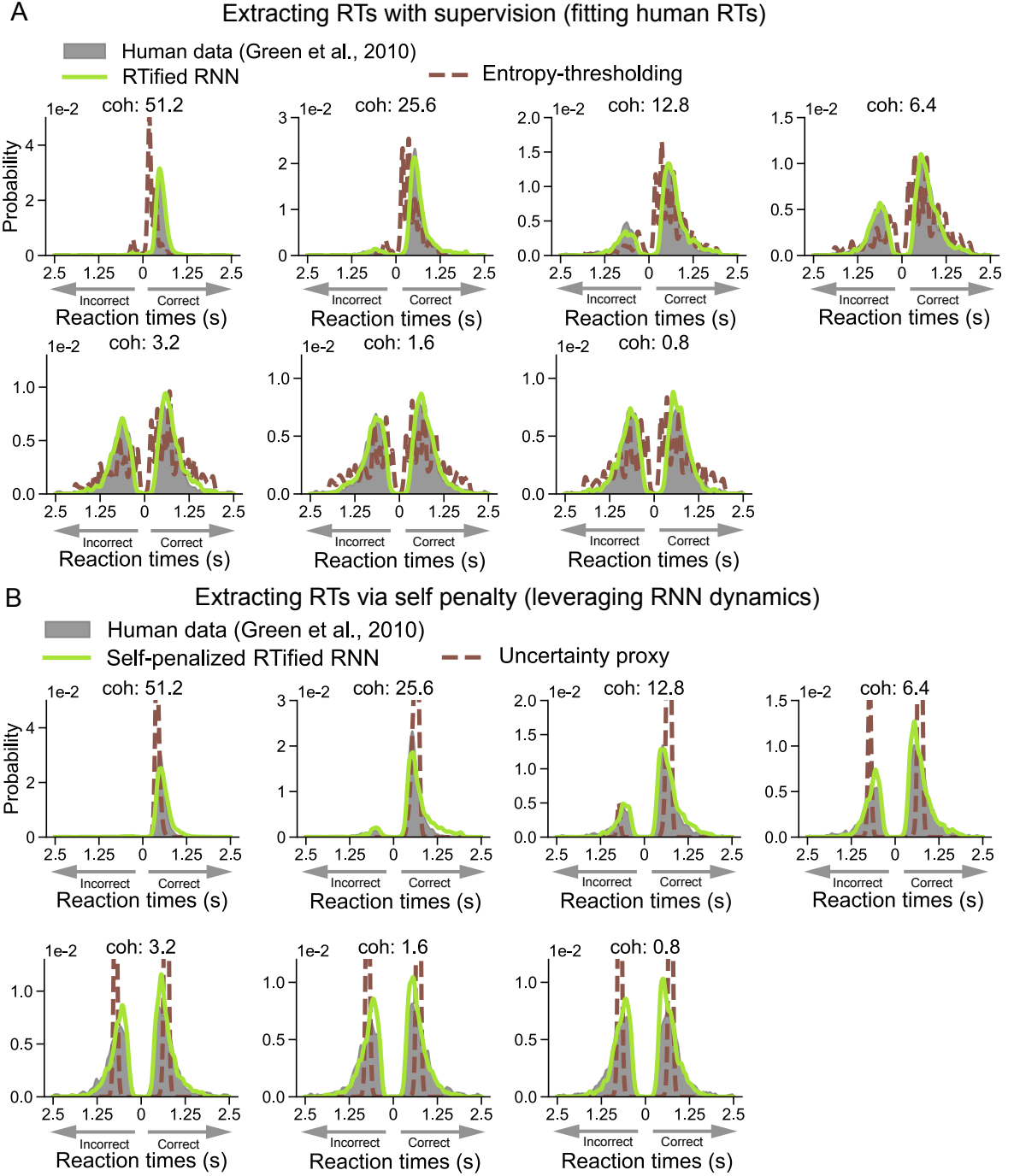


Figure 5.8: **RTified model evaluation on a RDM task Green and Swets (1966) across all coherences.** Human data are shown as a gray shaded area, and model fits are shown for (A) the “supervised” setting where human behavioral responses are used to train the models and (B) the “self-penalized” setting where no human data is used. Our approach (green) outperforms the two alternative approaches (brown), i.e., entropy-thresholding Spoerer et al. (2020) for the “supervised” and uncertainty proxy Goetschalckx et al. (2021) for the “self-penalized” settings.

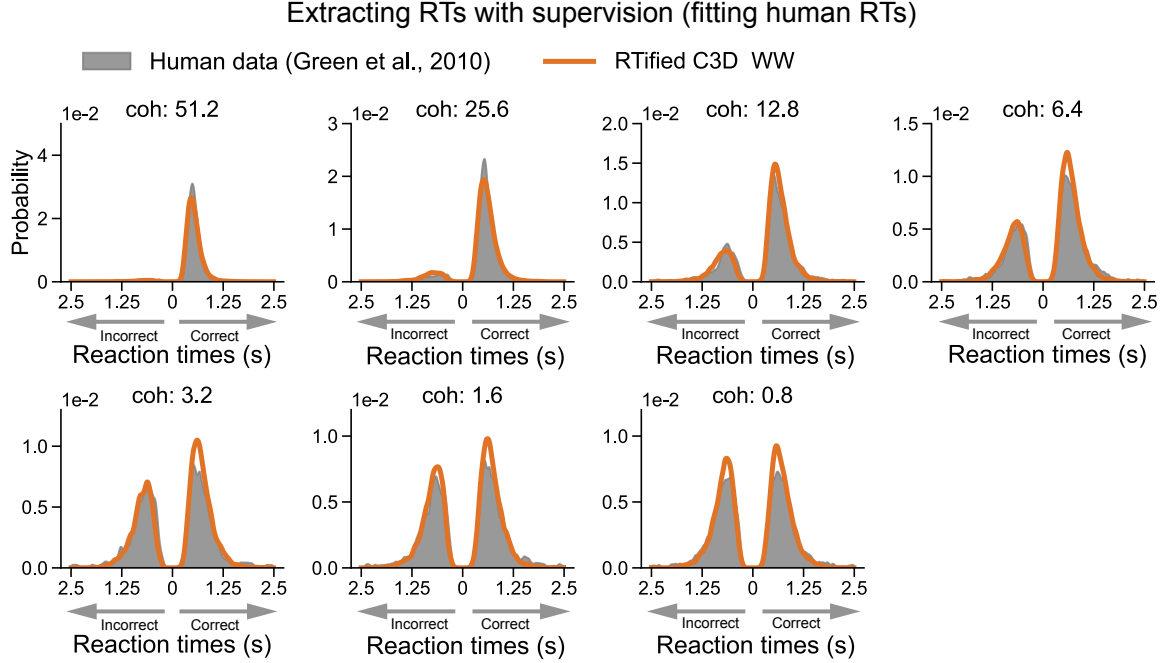


Figure 5.9: **RTified WW model on a RDM task Green and Swets (1966) across all coherences.** We combined our RTified WW module with a 3D CNN to fit human RTs collected in an RDM task Green and Swets (1966). Human data are shown as a gray shaded area and model results are shown as orange solid lines.

robustness of our method to different output aggregation strategies.

5.6.4 Illustration of the RTified WW Circuit

To provide a detailed understanding of how our RTified WW module works, we provide an illustration of the circuit and pseudo-code for its implementation. Figure 5.11 shows the detailed architecture and provides pseudo-code for the implementation.

5.7 Implications and Future Directions

5.7.1 Integration of Vision and Decision-Making Models

The success of RTify demonstrates that it is possible to integrate image-computable models with decision-making models, advancing toward a more comprehensive understanding of the brain mechanisms underlying dynamic vision. This integration addresses

a fundamental gap in our understanding of human visual processing, where the temporal dynamics of decision-making have been largely neglected in computational models.

By providing a differentiable framework for aligning neural network dynamics with human reaction times, RTify opens up new possibilities for building AI systems that not only achieve high accuracy but also exhibit human-like temporal dynamics in their decision-making processes.

5.7.2 Biological Plausibility

The RTified WW module represents a significant step toward building biologically plausible models of decision-making that can process complex visual stimuli. By extending the original WW model to handle multi-class classification and complex inputs, while maintaining its biological inspiration, we demonstrate that it is possible to build models that are both biologically plausible and practically useful.

The success of this approach suggests that biological inspiration can be valuable not only for architectural design (as demonstrated in Chapter 3) but also for modeling the temporal dynamics of decision-making processes.

5.7.3 Limitations and Future Work

Certain limitations will need to be addressed in future work. Most of the human data used in our study remains relatively small-scale and is limited primarily to synthetic images because more naturalistic benchmarks only include behavioral choices Hebart et al. (2019); Mehrer et al. (2021b) and lack RT data. To properly evaluate our approach and that of others, larger-scale psychophysics datasets using more realistic visual stimuli will be needed.

There is already evidence that large-scale psychophysics data can be used to effectively align AI models with humans Fel et al. (2022a); Linsley and Serre (2023). We hope this

work will encourage researchers to collect novel internet-scale benchmarks that include both behavioral choices and RTs.

5.7.4 Broader Impacts

As AI vision models become more prevalent in our daily lives, ensuring their trustworthy behavior is increasingly important Liang et al. (2022); Li et al. (2023). Our framework contributes to this effort by exploring how to align certain aspects of models’ behavior with human responses in specific contexts. While our approach is limited to predicting RT distributions, it constitutes a first step toward more human-aligned AI models.

5.8 Conclusion

We have described a computational framework to train RNNs to learn to dynamically accumulate evidence nonlinearly so that decisions can be made based on a variable number of time steps to approximate human behavioral choices, including both decisions and RTs. We showed that such optimization can be used to fit an RNN directly to human behavioral responses. We also showed that such a framework can be extended to an ideal-observer model whereby the RNN is trained without human data but with self-penalty that encourages the network to make a decision as quickly as possible. Under this setting, human-like behavioral responses naturally emerge from the RNN—consistent with the hypothesis that humans achieve a speed-accuracy trade-off.

Finally, we provided an RNN implementation of a popular neural circuit decision-making model, the WW model, as a trainable deep learning module that can be combined with any vision architecture to fit human behavioral responses. Our computational framework provides a way forward to integrating image-computable models with decision-making models, advancing toward a more comprehensive understanding of the brain mechanisms underlying dynamic vision.

The work presented in this chapter represents a significant contribution to the field of computational neuroscience and artificial intelligence. By bridging the gap between vision models and decision-making models, RTify provides a pathway to building AI systems that not only achieve high accuracy but also exhibit human-like temporal dynamics in their decision-making processes. This represents a crucial step toward building truly human-aligned AI systems that can understand and replicate the full complexity of human visual decision-making.

The next chapter will explore how these insights can be combined with the harmonization approaches from previous chapters to create a comprehensive framework for building biologically-aligned AI systems that capture both the spatial and temporal aspects of human visual processing.

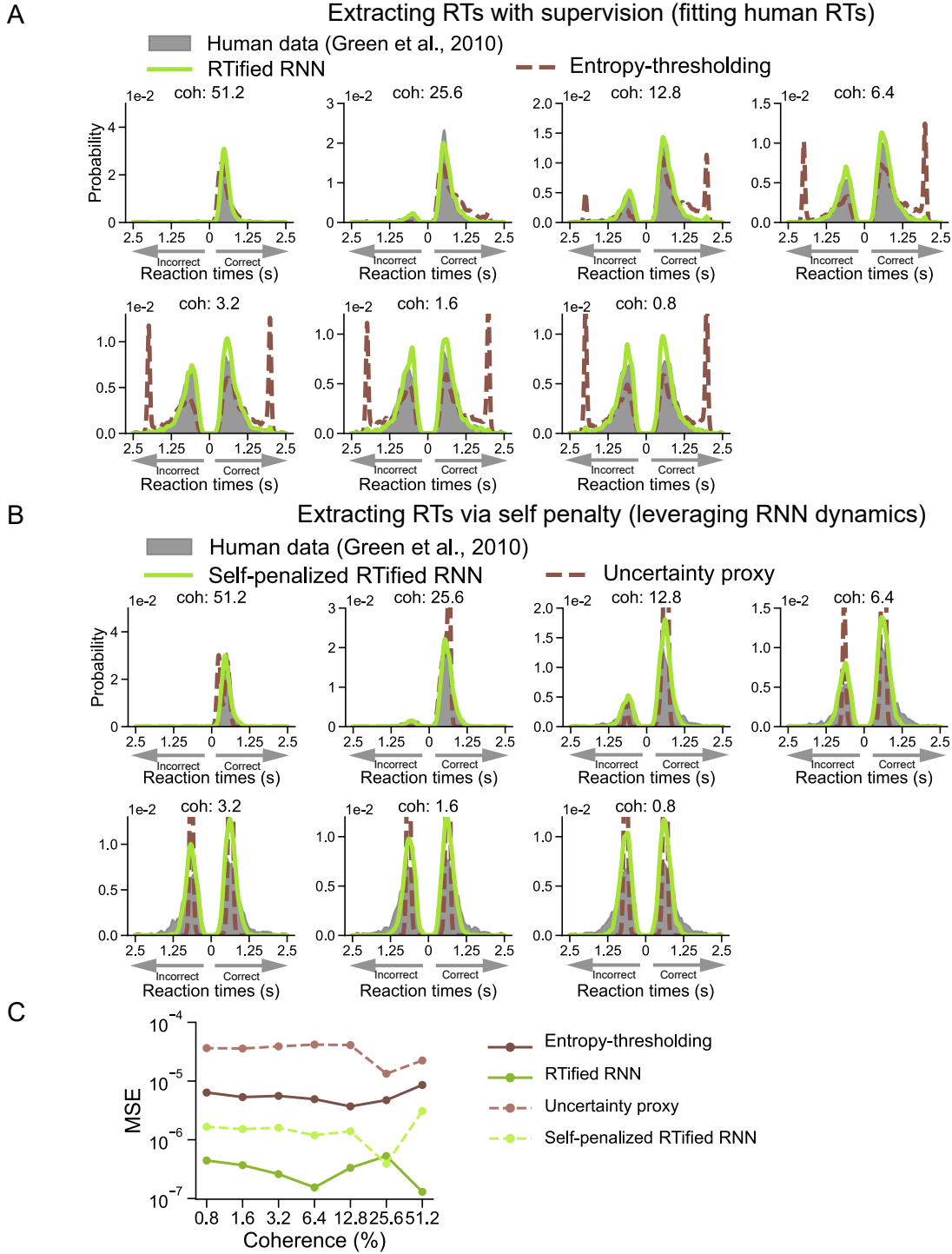
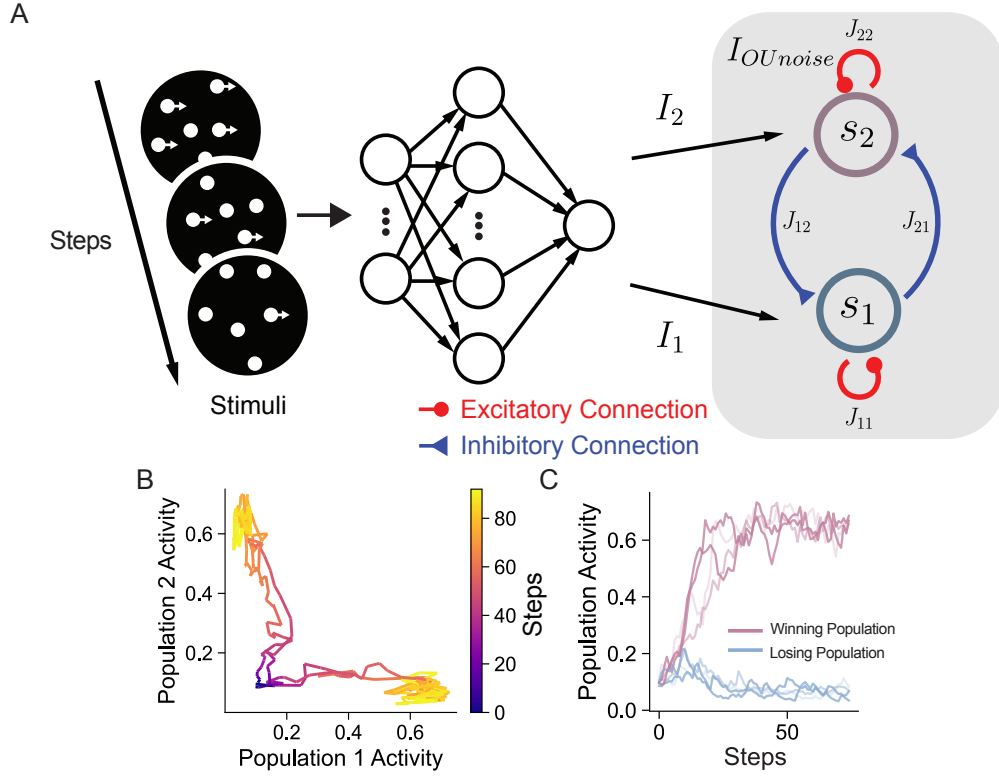


Figure 5.10: **RTified model evaluation on a RDM task Green and Swets (1966) across all coherences.** Human data are shown as a gray shaded area, and model fits are shown for (A) the “supervised” setting where human behavioral responses are used to train the models and (B) the “self-penalized” setting where no human data is used. Our approach (green) outperforms the two alternative approaches (brown), i.e., entropy-thresholding Spoerer et al. (2020) for the “supervised” and uncertainty proxy Goetschalckx et al. (2021) for the “self-penalized” settings. MSE comparison is shown in (C).



Algorithm 1 Forward Function of WW circuit
for Random Dot Motion

```

1: Input: Image, threshold
2: Output: decision_times
3: Initialize  $s_1, s_2, decision\_times$ 
4: while  $s_1 < threshold$  and  $s_2 < threshold$  do
5:    $I_1, I_2 \leftarrow f_w(\zeta(Image))$ 
6:   Combining excitatory and inhibitory currents
7:    $H_1 \leftarrow f(J_{11} \cdot s_1 - J_{12} \cdot s_2 + I_0 + I_1 + I_{OU\ noise1})$ 
8:    $H_2 \leftarrow f(J_{22} \cdot s_2 - J_{21} \cdot s_1 + I_0 + I_2 + I_{OU\ noise2})$ 
9:   Calculate rate of change
10:   $\frac{ds_1}{dt} = -\frac{s_1}{\tau_s} + H_1 \cdot (1 - s_1)$ 
11:   $\frac{ds_2}{dt} = -\frac{s_2}{\tau_s} + H_2 \cdot (1 - s_2)$ 
12:  Update  $s_1, s_2, decision\_times$ 
13: end while
14: return decision_times

```

Figure 5.11: **Illustration of the RTified WW circuit.** To RTify feedforward neural networks, we used an RNN based on the WW circuit. Here, we consider a binary classification on random moving dots for illustration (A). When receiving a visual input, the two populations sensitive to left/right directions compete with each other (B) and accumulate evidence until one of them reaches a threshold (C). The number of time steps needed for the RNN to reach the threshold is defined as the “model RT” and is used to predict human RT. We provide pseudo-code for a more detailed description of the circuit (D). Here, $f(x) = \gamma \cdot \max(ax - b, 0) / (1 - \exp(-d(ax - b)))$ is a fixed nonlinear function. And the model and equations marked blue are how we extend WW model.

CHAPTER 6

CONCLUSIONS AND FUTURE DIRECTIONS

6.1 Summary of Contributions

This dissertation has documented a fundamental challenge in computational neuroscience and artificial intelligence: the systematic divergence between deep neural networks optimized for engineering tasks and biological vision systems. Through comprehensive empirical analysis and novel methodological approaches, we have made several key contributions:

6.1.1 Documenting the Performance-Alignment Trade-off

Our first major contribution was to systematically document that modern deep neural networks face a fundamental trade-off between achieving high accuracy on ImageNet and accurately predicting neural responses in inferior temporal (IT) cortex. This finding, based on analysis of 135 different DNNs across multiple neural datasets, challenges the foundational assumption that has guided computational neuroscience for the past decade. The trade-off persists across different architectural families (CNNs, Vision

Transformers), training paradigms (supervised, self-supervised, adversarial training), and datasets, suggesting that this is a fundamental property of current deep learning approaches rather than a limitation of specific models or training procedures. (Linsley and Serre, 2023)

6.1.2 Neural Harmonization: A Novel Approach

Our second major contribution was the development of neural harmonization, a novel training paradigm (Fel et al., 2022a) that constrains DNNs with human behavioral data to create models that are both high-performing and biologically plausible. This approach demonstrated that it is possible to break the performance-alignment trade-off by incorporating human behavioral constraints during training. The harmonized models achieved comparable accuracy to standard DNNs while significantly improving their ability to predict neural responses in IT cortex. This finding provides a proof-of-concept that alternative training objectives can yield models that better capture the computational strategies of biological vision.

6.1.3 Two Complementary Pathways to Biological Alignment

An important aspect of this work that merits explicit discussion is the relationship between the different approaches to biological alignment explored throughout this dissertation. Our work revealed two distinct but complementary pathways to achieving biological alignment in artificial vision systems:

Training-Based Solutions (Chapters 1–2): Neural harmonization demonstrated that constraining DNNs with human behavioral data during training can break the performance-alignment trade-off. This approach works by modifying the training objectives to incorporate human visual strategies, enabling existing architectures to learn more biologically plausible representations without sacrificing engineering performance.

Architecture-Based Solutions (Chapter 3): Our work with HMAX 3.0 revealed

that incorporating biologically inspired architectural constraints—particularly built-in scale invariance mechanisms—can achieve superior biological alignment through structural design rather than training modifications (Serre et al., 2007a; Riesenhuber and Poggio, 2002). This approach demonstrates that some aspects of biological computation may be better captured through architectural innovations that enforce specific computational principles.

These two approaches address different aspects of the alignment problem: harmonization aligns the learned feature representations with human strategies, while architectural solutions like HMAX enforce computational principles that emerge from biological constraints. Importantly, these approaches are not mutually exclusive but rather complementary pathways that could potentially be combined for even greater biological alignment.

6.1.4 Modernizing Biologically-Inspired Architectures: HMAX 3.0

Our third contribution was the systematic modernization of the HMAX architecture, demonstrating that biologically-inspired models can achieve competitive task performance while maintaining superior biological alignment. By progressively incorporating modern deep learning techniques—learnable filters, residual connections, and contrastive learning—while preserving core biological principles like multi-scale processing and hierarchical organization, we developed HMAX 3.0 that achieves 72.3% ImageNet accuracy alongside robust scale invariance and state-of-the-art BrainScore performance.

This work demonstrates that architectural constraints inspired by neuroscience can yield models that are both performant and aligned with biological vision, providing an alternative pathway to biological alignment that complements training-based approaches like harmonization.

6.1.5 Reaction Time Alignment

Our final contribution was the development of methods for aligning neural networks with human reaction times, demonstrating that temporal aspects of human behavior can also be incorporated into model training. This work showed that models can be trained to not only predict neural responses but also to exhibit human-like temporal dynamics in their decision-making processes.

6.2 Implications for Neuroscience

6.2.1 Rethinking the Optimization Hypothesis

Our findings have profound implications for computational neuroscience. The optimization hypothesis, which posits that the visual cortex is optimized for object recognition and that optimizing artificial systems for the same task will yield insights into biological computation, appears to be incorrect in its simple form.

The systematic divergence between artificial and biological systems as artificial systems become more sophisticated suggests that biological vision is optimized for objectives that are fundamentally different from those that drive performance on ImageNet. While object recognition is certainly an important function of the visual system, it may not be the primary or only objective that shapes neural representations.

6.2.2 The Importance of Multiple Objectives

Our work suggests that biological vision systems are likely optimized for multiple, potentially competing objectives. These might include:

- Object recognition and categorization
- Spatial navigation and scene understanding
- Control of action, including reaching, grasping, locomotion, and physical interactions

- Classifying affordances for action as part of scene understanding
- Social perception and face processing
- Temporal dynamics and motion processing
- Energy efficiency and computational constraints

The challenge for computational neuroscience is to identify these multiple objectives and understand how they interact to shape neural representations.

6.2.3 The Role of Behavioral Data

Our neural harmonization work demonstrates that human behavioral data can provide crucial constraints for understanding biological computation. By incorporating behavioral measures into model training, we can create systems that better capture the computational strategies of biological vision.

This suggests that future computational neuroscience research should integrate neural and behavioral data more closely, using behavioral measures to constrain and guide the development of computational models.

6.3 Implications for Artificial Intelligence

6.3.1 The Scaling Problem

Our findings highlight a critical problem with the current scaling paradigm in AI. The assumption that larger models trained on more data will naturally become more human-like appears to be false. Instead, we observe a systematic divergence between artificial and biological intelligence as artificial systems become more sophisticated. This suggests that simply scaling up current approaches—increasing model size, training data, and computational resources—will not solve the alignment problem. Instead, fundamentally different approaches to training and architecture may be necessary. (Bowers et al., 2022)

6.3.2 The Need for New Benchmarks

Our work also points to the need for better benchmarks that capture the full complexity of biological vision. ImageNet, while useful for engineering applications, may not be the appropriate target for systems that aim to model biological intelligence.

Future benchmarks should consider:

Multiple objectives beyond object recognition: Benchmarks should evaluate models' ability to simultaneously perform object recognition, scene understanding, depth estimation, and affordance detection. For example, a comprehensive benchmark might require models to not only identify objects in a scene but also predict their spatial relationships, estimate which objects can be grasped or manipulated, and understand the navigational affordances of the environment. This would capture the multi-task nature of biological vision that evolved to support survival and interaction with the world.

Temporal dynamics of visual processing: Current benchmarks focus on static image classification, ignoring the rich temporal dynamics of biological vision. New benchmarks should include video sequences that require temporal integration, motion processing, and dynamic scene understanding. Models should be evaluated on their ability to track objects through occlusion, predict future states, and integrate information across time scales ranging from milliseconds to seconds. This would capture the sequential nature of visual processing that is fundamental to biological systems.

Behavioral measures beyond accuracy: Biological vision exhibits rich behavioral patterns including reaction times, confidence judgments, error patterns, and eye movement strategies. Benchmarks should evaluate models on their ability to predict human reaction times for different stimuli, match human error patterns across different viewing conditions, and exhibit human-like confidence calibration. Models should also be assessed on their alignment with human attention patterns and visual search strategies, as measured through eye-tracking data. (Wichmann et al., 2023)

Neural predictivity across multiple brain areas: Rather than focusing solely on high-level areas like IT cortex, benchmarks should evaluate models’ ability to predict neural responses across the entire visual hierarchy, from retina and LGN to V1, V4, and IT. This would require models to capture both low-level feature processing and high-level semantic representations. Additionally, benchmarks should include neural data from different species and recording modalities (single units, fMRI, EEG) to ensure generalizability of the computational principles.

Robustness to natural variations and perturbations: Biological vision is remarkably robust to variations in lighting, viewpoint, partial occlusion, and other natural perturbations. Benchmarks should systematically evaluate models under these conditions, including different illumination conditions, weather effects, seasonal changes, and partial occlusion scenarios. Models should maintain both accuracy and biological alignment under these challenging conditions, demonstrating the kind of robustness that characterizes biological vision systems.

6.3.3 The Promise of Neural Alignment

Our work demonstrates that neural alignment is a promising direction for creating more human-like AI systems. By explicitly training models to predict neural responses, we can create systems that better capture the computational strategies of biological intelligence.

This approach could be extended beyond vision to other domains, such as language processing, motor control, and decision-making, potentially leading to more general and human-like AI systems.

6.4 Future Directions

6.4.1 Expanding Neural Harmonization

One promising direction is to expand neural harmonization to other domains and brain areas. While our work focused on visual cortex, the same principles could be applied to other neural systems, each requiring domain-specific behavioral constraints and evaluation metrics. (Fel et al., 2022a)

Auditory cortex and speech processing: Harmonization could incorporate human speech perception data, including phoneme discrimination boundaries, temporal processing of acoustic features, and cross-linguistic variations in speech categorization. The training objective would align model representations with human auditory feature maps derived from psychoacoustic experiments.

Motor cortex and movement control: Models could be harmonized with human movement data, including reaching trajectories, grasping patterns, and motor learning curves. This would require developing explanation methods for motor control networks and collecting comprehensive human movement datasets across different tasks and contexts.

Prefrontal cortex and decision-making: Harmonization could incorporate human decision-making strategies, including risk assessment, temporal discounting, and cognitive control patterns. This would extend the temporal alignment work from Chapter 4 to more complex cognitive tasks.

Hippocampus and memory formation: Models could be aligned with human memory encoding and retrieval patterns, including spatial navigation strategies, episodic memory formation, and consolidation processes.

The key challenge for each domain is developing appropriate behavioral datasets and explanation methods that can capture the relevant computational strategies. This expansion would lead to a more comprehensive understanding of how different brain areas

compute and how artificial systems can be designed to capture these computations across the full spectrum of human intelligence.

6.4.2 Multi-Modal Alignment

Another promising direction is to develop methods for aligning models with multiple types of neural and behavioral data simultaneously. This represents a significant technical and conceptual challenge that requires new optimization frameworks and data collection protocols.

Simultaneous multi-area recordings: Aligning models with neural data from multiple brain areas recorded simultaneously would require developing loss functions that can balance competing constraints from different regions. For example, early visual areas prioritize retinotopic organization and spatial frequency tuning, while higher areas emphasize object identity and invariance. The challenge is developing optimization procedures that can satisfy these potentially conflicting objectives.

Neural-behavioral co-constraint: Collecting neural and behavioral data from the same subjects performing identical tasks would enable stronger constraints on model development. This would require developing experimental paradigms that can simultaneously measure neural responses and behavioral strategies, potentially using techniques like simultaneous fMRI and eye-tracking or EEG and psychophysics.

Cross-species alignment: Aligning models with data from multiple species (humans, macaques, mice) could reveal fundamental computational principles that transcend species-specific adaptations. This would require developing normalization and alignment procedures that can account for differences in brain size, organization, and behavioral repertoires.

Multi-modal behavioral measures: Integrating different types of behavioral measures (reaction times, eye movements, confidence ratings, error patterns) would provide

richer constraints on model development. The challenge is developing unified frameworks that can simultaneously optimize for these diverse behavioral signatures.

Such multi-modal alignment could provide more comprehensive constraints and lead to more robust models, but it requires fundamental advances in optimization theory, data collection methodologies, and experimental design. This vision is consistent with emerging initiatives like integrative BrainScore evaluations that combine multiple neural and behavioral metrics to benchmark models (Schrimpf et al., 2020a).

6.4.3 Understanding the Mechanisms

A crucial direction for future research is to understand the mechanisms by which neural harmonization works. This mechanistic understanding is essential for developing more effective alignment methods and gaining insights into biological computation.

Representational changes during harmonization: What specific changes occur in model representations when they are trained with neural constraints? This could be investigated through techniques like representational similarity analysis (RSA), canonical correlation analysis (CCA), and activation space analysis to understand how harmonization reshapes the geometry of neural representations.

Feature emergence and selection: How does harmonization influence which features emerge during training and how they are selected for different tasks? This could involve studying the dynamics of feature learning during harmonization training and comparing the resulting feature hierarchies with those found in biological vision.

Computational strategy alignment: How do the computational strategies learned through harmonization relate to those employed by biological vision? This requires developing methods to characterize and compare computational strategies across artificial and biological systems, potentially through dynamical systems analysis and information-theoretic approaches.

Optimization landscape analysis: How does adding biological constraints change the optimization landscape and learning dynamics? Understanding these changes could inform the development of better optimization procedures and training curricula for biologically-aligned models.

Understanding these mechanisms could provide insights into both biological computation and the design of better artificial systems, ultimately leading to a principled framework for creating AI systems that capture the full complexity of biological intelligence.

6.4.4 Combining Training and Architectural Solutions

A particularly promising direction for future research is the systematic combination of harmonized training with biologically-inspired architectures. Our work has demonstrated that both training-based and architecture-based approaches can independently improve biological alignment, but their combined potential remains largely unexplored.

Preliminary investigations suggest that harmonized DNNs achieve BrainScore values in the range of 0.45-0.55, significantly higher than standard ImageNet-trained DNNs (0.25-0.35) but still below HMAX 3.0’s performance (0.55-0.61). This gap suggests that combining harmonization with architectural innovations like HMAX could potentially achieve even higher levels of biological alignment while maintaining strong engineering performance.

Future work should investigate:

- **Harmonized HMAX Training:** Applying neural harmonization to HMAX architectures to combine the benefits of built-in scale invariance with human-aligned feature learning. By integrating the harmonization training paradigm with the classic HMAX framework, we can leverage HMAX’s inherent invariance properties alongside human-guided learning objectives. This hybrid approach aims to produce models that retain HMAX’s robust scale and position tolerance while also mirroring

human representational strategies (Fel et al., 2022a; Serre et al., 2007a). In principle, a harmonized HMAX could achieve greater alignment than either approach alone, underscoring the promise of this integration.

- **Architectural Constraints for Harmonization:** Developing architectures specifically designed to work synergistically with harmonization objectives. This entails building networks that embed key features of biological vision from the outset. For instance, including known cortical motifs—such as recurrent feedback loops or specialized processing streams—can guide learning toward human-like solutions that might not emerge under standard architectures (Wichmann et al., 2023). By hard-coding certain invariances and organizational principles (e.g., retinotopy or two-stream pathways), such architectures would provide inductive biases that complement harmonization, potentially yielding models with deeper alignment to neural data (Riesenhuber and Poggio, 2002).
- **Multi-Objective Optimization:** Creating training frameworks that simultaneously optimize for task performance, neural alignment, and architectural constraints. Moving beyond single-metric optimization is critical to prevent improvements in one domain (e.g., accuracy) from degrading alignment in another. This approach aligns with the philosophy of evaluation paradigms like BrainScore that integrate behavioral and neural criteria (Schrimpf et al., 2020b,a). By explicitly balancing multiple objectives during training, models could maintain high task accuracy while preserving their similarity to biological representations (Bowers et al., 2022).
- **Hierarchical Alignment:** Aligning different levels of the visual hierarchy with different biological constraints (early vision with retinotopic organization, higher areas with semantic representations). For example, early layers of a model could be constrained to develop retinotopic maps and orientation selectivity analogous to primary visual cortex, reflecting the proto-organized structure present at birth (Arcaro et al., 2017). Higher layers, in turn, would be guided to form abstract,

invariant representations similar to those in IT cortex, consistent with findings that performance-optimized models develop IT-like features (Yamins et al., 2014a). Enforcing such layer-specific alignment from V1 through IT would yield a model that mirrors the entire ventral stream’s organization and function.

This combined approach could potentially close the remaining AI–BI gap by leveraging the complementary strengths of both training-based and architecture-based solutions. Indeed, this represents an active area of ongoing research that holds significant promise for achieving truly human-like artificial vision systems.

6.4.5 Developing New Architectures

Beyond combining existing approaches, our work suggests that fundamentally new architectures may be needed to fully capture biological computation. Future research should explore:

- **Architectures that more closely resemble biological neural networks:**

Designing models with wiring patterns and components inspired by the brain may yield more human-like vision systems. Deep hierarchical networks with multiple stages of processing and interleaved feedforward and feedback connections (as exemplified by early models like HMAX) could provide a stronger approximation to cortical circuitry (Serre et al., 2007a; Riesenhuber and Poggio, 1999b). By incorporating features such as columnar organization, lateral inhibition, or parallel visual pathways, these biologically inspired architectures aim to reproduce the emergent computations observed in real neural networks.

- **Training procedures that better capture biological learning:** Implementing learning rules and curricula that mirror how humans and animals learn could improve alignment. Unlike standard supervised learning on static datasets, biological learning is dynamic, self-supervised, and data-efficient. Future models might employ unsupervised or self-supervised objectives (e.g., prediction, contrastive learning) and

staged learning that reflects developmental progression. Such approaches, which have been argued to be closer to human learning processes, may lead to networks that acquire invariances and generalizations more akin to those of human vision (Bowers et al., 2022).

- **Models that incorporate biological constraints (e.g., energy efficiency, temporal dynamics):** Artificial networks could be augmented with constraints reflecting the physical and metabolic limits of biological brains. For instance, limiting a model’s activation energy or imposing sparse firing patterns can simulate the brain’s energy efficiency requirements, potentially steering the model toward more brain-like solutions (cf. human vision’s robustness under energy constraints). Introducing realistic neuronal dynamics (such as processing time limits or noise levels) may force networks to develop strategies that align better with human perceptual timing and reliability (Wichmann et al., 2023). By ensuring the model operates within a biologically plausible regime, we increase the likelihood that its internal strategies resemble those of actual neural systems.
- **Hybrid approaches that combine artificial and biological principles:** Another avenue is to merge the strengths of conventional deep learning with insights from biological vision. For example, one could integrate symbolic reasoning modules or neuromorphic computing elements into deep neural networks, creating hybrid systems that use artificial components for efficiency while embedding biological rules for alignment. These hybrids might use deep nets for perception but incorporate neuro-inspired decision rules or memory structures, yielding systems that not only perform well but also follow cognitive patterns observed in humans. The goal is a new class of AI models that leverage decades of machine learning advances while remaining grounded in the constraints and strategies of the brain.

6.4.6 Applications to Brain-Computer Interfaces

The neural alignment methods developed in this dissertation could have important applications in brain-computer interfaces (BCIs). By creating models that better predict neural responses, we could develop more effective decoding algorithms for BCIs.

This could lead to improved communication systems for individuals with severe motor disabilities and better understanding of how to interface artificial systems with biological brains.

6.5 Toward a Unified Framework for Biological Intelligence

6.5.1 The Path Forward

This dissertation has documented a fundamental challenge in the relationship between artificial and biological intelligence: as artificial systems become more sophisticated, they appear to diverge from biological systems rather than converge toward them. This finding challenges many assumptions in both computational neuroscience and artificial intelligence, but it also illuminates a clear path forward.

Our work demonstrates that the divergence between artificial and biological intelligence is not inevitable. Through neural harmonization, architectural innovations like HMAX 3.0, temporal alignment with reaction times, and cross-dataset neural alignment, we have shown that it is possible to create artificial systems that are both high-performing and biologically plausible.

The key insight is that biological intelligence is shaped by multiple, interacting objectives that go far beyond simple task performance. To create truly human-like AI systems, we must understand and incorporate these multiple objectives into our models. This requires a fundamental shift from single-objective optimization to multi-objective

frameworks that can balance competing demands.

6.5.2 A Roadmap for Integration

The most promising path forward involves the systematic integration of the approaches developed in this dissertation:

Phase 1: Methodological Integration - Combine harmonized training with biologically-inspired architectures like HMAX to leverage the complementary strengths of both training-based and architecture-based solutions.

Phase 2: Multi-Objective Optimization - Develop training frameworks that simultaneously optimize for multiple biological objectives, including object recognition, spatial navigation, action control, affordance classification, and temporal dynamics.

Phase 3: Cross-Modal Extension - Extend these principles beyond vision to other sensory modalities and cognitive functions, creating truly integrated models of biological intelligence.

Phase 4: Real-World Validation - Test these integrated models in real-world scenarios that require the full spectrum of biological intelligence, from navigation and manipulation to social interaction and learning.

6.5.3 Broader Implications

This research program has implications that extend beyond computational neuroscience and artificial intelligence. By developing AI systems that better capture the principles of biological intelligence, we may gain insights into fundamental questions about consciousness, cognition, and the nature of intelligence itself.

Moreover, biologically-aligned AI systems could have profound practical applications, from more effective brain-computer interfaces to AI systems that can work more naturally and safely alongside humans. As AI becomes increasingly integrated into society, ensuring

that these systems operate according to principles that are compatible with human cognition becomes not just a scientific goal but a societal imperative.

6.5.4 Final Reflections

This dissertation represents a step toward that goal. By documenting the problem, developing solutions, and pointing toward future directions, we hope to contribute to the development of AI systems that are not only powerful but also more aligned with biological intelligence.

The journey toward truly human-like AI is long and complex, but the methods and insights developed in this dissertation provide a foundation for continued progress. As we move forward, we must remain committed to the principle that understanding biological intelligence is not just a means to better AI, but an end in itself - a fundamental scientific question that deserves our best efforts.

The convergence of artificial and biological intelligence may be one of the most important scientific challenges of our time. By pursuing this goal with rigor, creativity, and dedication to biological truth, we can hope to create AI systems that not only match human performance but truly understand the world as humans do.

References

- Abnar, S. and Zuidema, W. (2020). Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.
- Arcaro, M. J., Schade, P. F., Vincent, J. L., Ponce, C. R., and Livingstone, M. S. (2017). A hierarchical, retinotopic proto-organization of the primate visual system at birth. *Elife*, 6:e26196.
- Azulay, A. and Weiss, Y. (2019). Why do deep convolutional networks generalize so poorly to small image transformations? *Journal of Machine Learning Research*, 20(184):1—25.
- Berrios, D. C., DeAngelis, G. C., and Angelaki, D. E. (2022). Neural mechanisms of visual processing. *Current Opinion in Neurobiology*, 74:102549.
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115.
- Biederman, I. and Cooper, E. E. (1992). Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):121.
- Biscione, V. and Bowers, J. S. (2022). Learning to recognize objects through repeated interactions. *Cognition*, 225:105105.
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700.
- Bomatter, P., Kümmerer, M., Wallis, T. S. A., and Bethge, M. (2021). Saliency models: A review. *Journal of Vision*, 21(13):1—36.

- Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolphi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., and Blything, R. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, pages 1—74.
- Boyd, J. and Serre, T. (2021). Human-level performance in visual object recognition. *Nature Communications*, 12(1):1—10.
- Brendel, W. and Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*.
- Brown, S. D. and Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57(3):153—178.
- Buswell, G. T. (1935). How people look at pictures. *University of Chicago Press*.
- Cadiou, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS computational biology*, 10(12):e1003963.
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650—9660.
- Chen, H., Wang, Y., Xu, C., Shi, B., Xu, C., Tian, Q., and Xu, C. (2021). Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10347—10357.
- Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2017). Rethinking atrous convolution for semantic image segmentation. In *arXiv preprint arXiv:1706.05587*.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder

- with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801—818.
- Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). Big transfer (bit): General visual representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 491—507.
- Cheng, Y.-A., Rodriguez, I. F. R., Chen, S., Kar, K., Watanabe, T., and Serre, T. (2024). RTify: Aligning deep neural networks with human behavioral decisions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Chollet, F. (2016). Xception: Deep learning with depthwise separable convolutions. *arXiv preprint arXiv:1610.02357*.
- Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251—1258.
- Chopra, S., Hadsell, R., and LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546.
- Cisek, P., Puskas, G. A., and El-Murr, S. (2009). Decisions in changing conditions: the urgency-gating model. *Journal of Neuroscience*, 29(37):11560—11571.
- D’Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., and Sagastizabal, L. (2021). Convit: Improving vision transformers with soft convolutional inductive biases. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2286—2296.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009a). Imagenet: A

- large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009b). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248—255.
- Deza, A. and Konkle, T. (2020). The center bias of the human visual system. *Journal of Vision*, 20(11):1—15.
- DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333—341.
- DiCarlo, J. J. and Cox, D. D. (2012). Does the brain do object recognition? *Nature Neuroscience*, 15(7):915—916.
- DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012a). How does the brain solve visual object recognition? *Neuron*, 73(3):415—434.
- DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012b). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.
- Doshi, F. R., Konkle, T., and Alvarez, G. A. (2024). A feedforward mechanism for human-like contour integration. *bioRxiv*, pages 2024—06.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., and Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, 32(11):3612—3628.
- Eberhardt, F. and Serre, T. (2016). How deep is the feature analysis underlying rapid visual

- categorization? In *Advances in Neural Information Processing Systems*, volume 29, pages 1100—1108.
- Eberhardt, S., Cader, J. G., and Serre, T. (2016). How much of the data in a deep neural network is redundant? *arXiv preprint arXiv:1611.06226*.
- Ebrahimpour, R., Ghaemmaghami, M. P., and Ghaemmaghami, M. P. (2019). A review of deep learning methods for object detection. *arXiv preprint arXiv:1908.03673*.
- Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in Psychology*, 2:243.
- Fel, T., Felipe, I., Linsley, D., and Serre, T. (2022a). Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Fel, T., Linsley, D., and Serre, T. (2022b). Harmonization: Learning to align neural networks with human behavior. *Advances in Neural Information Processing Systems*, 35:1—12.
- Funke, C. M., Borowski, J., Stollenga, M. F., Wallis, T. S. A., Brendel, W., Ecker, A. S., and Bethge, M. (2018). Salient object detection: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):824—840.
- Furukawa, S., Sugihara, H., Miyasaka, N., Hayashi, M., and Tanaka, K. (2017). Scale-invariant neural responses in the primate visual system. *Nature Neuroscience*, 20(1):1—10.
- Gao, S.-H., Cheng, M.-M., Zhao, K., Zhang, X.-Y., Yang, M.-H., and Torr, P. (2020). Res2net: A new multi-scale backbone architecture. *IEEE transactions on pattern analysis and machine intelligence*, 43(2):652—662.
- Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A.,

- and Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Nature Communications*, 12(1):1—12.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*.
- Geirhos, R. d., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665—673.
- Goetschalckx, L., Linsley, D., and Serre, T. (2021). Evidential deep learning for uncertainty estimation. *Advances in Neural Information Processing Systems*, 34:1—12.
- Golan, T., Raju, P., and Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human recognition. *Proceedings of the National Academy of Sciences*, 117(47):29330—29337.
- Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., and Douze, M. (2021). Levit: a vision transformer in convnet’s clothing for faster inference. *arXiv:2104.01136*.
- Green, D. M. and Swets, J. A. (1966). Stimulus-dependent modulation of response time in visual decision making. *Signal Detection Theory and Psychophysics*, 1:1—25.
- Guo, Y., Cao, X., Liu, B., and Gao, M. (2020). A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*.
- Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., and Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, 48(3):829—842.

- Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. (2019). Ghostnet: More features from cheap operations. *CoRR*, abs/1911.11907.
- Han, Y., Roig, G., Geiger, G., and Poggio, T. (2020a). Scale and translation-invariance for novel objects in human vision. *Sci Rep*, 10(1):1411.
- Han, Y., Roig, G., and Poggio, T. (2020b). Scale invariance in object detection. *arXiv preprint arXiv:2004.10934*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., and Baker, C. I. (2019). Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS ONE*, 14(10):e0223792.
- Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314—1324.
- Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708.
- Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology*, 160(1):106.
- Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73(1):218—226.
- Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for

- rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254—1259.
- Jagadeesh, A. V. and Gardner, J. L. (2022a). Neural mechanisms of visual object recognition. *Annual Review of Vision Science*, 8:1—25.
- Jagadeesh, A. V. and Gardner, J. L. (2022b). Texture-like representation of objects in human visual cortex. *Proceedings of the National Academy of Sciences*, 119(17):e2115302119.
- Jansson, G., Linsley, D., and Serre, T. (2022). Scale-invariant recognition in convolutional neural networks. *arXiv preprint arXiv:2201.09356*.
- Jansson, Y. and Lindeberg, T. (2022). Scale-Invariant Scale-Channel Networks: Deep Networks That Generalise to Previously Unseen Scales. *Journal of Mathematical Imaging and Vision*, 64(5):506—536.
- Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S., and Wang, J. (2015). Saliency detection: A survey. *Computational Visual Media*, 1(1):1—17.
- Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream’s execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974—983.
- Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., and Masquelier, T. (2016). Deep net-

- works can resemble human feed-forward vision in invariant object recognition. *Scientific Reports*, 6(1):1—12.
- Kietzmann, T. C., McClure, P., and Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*.
- Kim, J. H., Linsley, D., Thakkar, K., and Serre, T. (2020). Disentangling neural mechanisms for perceptual grouping. *Proceedings of the National Academy of Sciences*, 117(46):28623—28633.
- Kirchner, H. and Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, 46(11):1762—1776.
- Koehler, K., Akbas, E., Yildirim, I., Schill, K., Wachtler, T., and Rosenholtz, R. (2014). Saliency for fixations relates to distractor suppression in visual search. *Vision Research*, 100:32—43.
- Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2019). Large scale learning of general visual representations for transfer. *arXiv preprint arXiv:1902.07208*.
- Kreiman, G. (2020). Neural mechanisms of visual recognition. *Annual Review of Vision Science*, 6:1—25.
- Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1:417—446.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., and DiCarlo, J. J. (2019a). Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, 32.
- Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., and DiCarlo, J. J. (2019b).

- Brain-like object recognition with high-performing shallow recurrent anns. *Advances in Neural Information Processing Systems*, 32.
- Kumar, A., Khani, M., Schreyer, H., Gollisch, T., and Bethge, M. (2022). Deep learning models of the retinal response to natural scenes. *Advances in Neural Information Processing Systems*, 35:1—12.
- Lai, M. and Silver, D. (2019). Human-level and beyond: Artificial intelligence for chess, shogi, and go. *Communications of the ACM*, 62(11):92—99.
- Langlois, T., Gosselin, F., and Gosselin, P. (2021). The impact of dataset biases on deep face recognition systems. *arXiv preprint arXiv:2103.14999*.
- Law, C.-T. and Gold, J. I. (2008). Neural correlates of decision making in the primate visual system. *Nature Neuroscience*, 11(6):1—10.
- Lee, K., Zung, J., Li, P., Jain, V., and Seung, H. S. (2017). Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*.
- Li, X., Wang, Y., Wang, Y., Wang, Y., Wang, Y., Wang, Y., and Wang, Z. (2023). Trustworthy artificial intelligence: From principles to practice. *Nature Communications*, 14(1):1—13.
- Liang, P., Manzini, T., Sagawa, S., Koh, P. W., Hashimoto, T. B., Srebro, N., Alvarez, M., Zou, J., and Ma, T. (2022). Advances, challenges and opportunities in creating trustworthy artificial intelligence. *Nature Machine Intelligence*, 4(8):669—677.
- Lindeberg, T. (1994). *Scale-Space Theory in Computer Vision*. Springer, Berlin, Heidelberg.
- Linsley, D., Kim, J., Veerabadran, V., and Serre, T. (2021a). Learning what and where to attend. *Nature Communications*, 12(1):1—12.
- Linsley, D., Kim, J., Veerabadran, V., Windolf, C., and Serre, T. (2019). Learning

- long-range spatial dependencies with horizontal gated recurrent units. *Advances in Neural Information Processing Systems*, 32:1522—1533.
- Linsley, D., Kim, J. H., Veerabadran, V., Windolf, C., and Serre, T. (2018). What are the visual features underlying human versus machine vision? *Proceedings of the National Academy of Sciences*, 115(19):5115—5124.
- Linsley, D., Kim, J. H., Veerabadran, V., Windolf, C., and Serre, T. (2020). Stable and expressive recurrent vision models. *arXiv preprint arXiv:2006.04095*.
- Linsley, D., Kim, J. H., Veerabadran, V., Windolf, C., and Serre, T. (2021b). Visual explanations from deep networks via gradient-based localization. *Frontiers in Artificial Intelligence*, 4:688557.
- Linsley, D., Scheibler, D., Eberhardt, S., and Serre, T. (2017). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.
- Linsley, D. and Serre, T. (2023). Performance-optimized neural networks as models of human visual cognition. *Current Opinion in Behavioral Sciences*, 50:101234.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012—10022.
- Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552—563.
- Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440.
- Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402—13418.

- Mannan, S. K., Kennard, C., and Husain, M. (2009). The role of attention in the programming of saccades. *Vision research*, 49(10):1259—1267.
- Marjeh, R., Ben-Shahar, O., and Serre, T. (2022). The role of recurrent processing in object recognition. *Trends in Cognitive Sciences*, 26(2):124—138.
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., USA.
- Mehrer, J., Kietzmann, T. C., and Kriegeskorte, N. (2021a). Ecologically valid neural networks. *Nature Machine Intelligence*, 3(1):1—2.
- Mehrer, J., Kietzmann, T. C., and Kriegeskorte, N. (2021b). Ecologically valid neural networks for visual recognition. *Nature Machine Intelligence*, 3(1):1—2.
- Mehta, S., Rastegari, M., Caspi, A., Shapiro, L., and Hajishirzi, H. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510—4520.
- Muller, Y., Varma, S., Reich, D., Eichhorn, H., Chari, S., Nowakowski, T., Coady, Y., Abudarham, N., Hsieh, P., Kanwisher, N., et al. (2019). Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 0—0.
- Muriel, M. and Thorpe, S. J. (2007). The time course of visual processing: From early perception to decision-making. *Journal of Vision*, 7(9):1—15.
- Mutch, J. and Lowe, D. G. (2006). Multiclass object recognition with sparse, localized features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:11—18.
- Muttenthaler, L. and Hebart, M. N. (2022). Human alignment of neural network representations. *arXiv preprint arXiv:2206.14808*.

- Oliva, A. and Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145—175.
- Oppenheim, A. V. and Lim, J. S. (1981). The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529—541.
- Palmer, J., Huk, A. C., and Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, 5(5):1—1.
- Pant, N., Rodriguez, I. F., Beniwal, A., Warren, S., and Serre, T. (2024). Hmax strikes back: Self-supervised learning of human-like scale invariant representations. In *Proceedings of the 7th Conference on Cognitive Computational Neuroscience (CCN 2024)*, Cambridge, MA, USA.
- Parthasarathy, N. and Simoncelli, E. P. (2020). Learning a texture model for representing cortical area v2. In *Computational and Systems Neuroscience (CoSyNe)*, Denver, CO.
- Peterson, J. C., Abbott, J. T., and Griffiths, T. L. (2018). Human and model performance in visual object recognition. *Current Directions in Psychological Science*, 27(6):477—483.
- Petrov, Y., Carandini, M., and McKee, S. (2011). Dissociable effects of attention and crowding on orientation averaging. *Journal of Vision*, 11(13):17—17.
- Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology*, 32(1):3—25.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Aspell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763.
- Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., and Dollár, P. (2020). Designing

- network design spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10428–10436.
- Rahnev, D., Koizumi, A., McCurdy, L. Y., D’Esposito, M., and Lau, H. (2021). Robust evidence for the distinction between confidence and decision time. *Nature Human Behaviour*, 5(1):1—10.
- Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255—7269.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85(2):59.
- Ratcliff, R. and McKoon, G. (2015). Modeling 2-alternative forced-choice tasks: Accounting for both magnitude and difference effects. *Cognitive Psychology*, 78:1—22.
- Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., Ganguli, S., George, T. M., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761—1770.
- Riesenhuber, M. and Poggio, T. (1999a). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019—1025.
- Riesenhuber, M. and Poggio, T. (1999b). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019—1025.
- Riesenhuber, M. and Poggio, T. (2002). Models of object recognition. *Nature Neuroscience*, 5(11):1199—1204.
- Roads, B. D., Mozer, M. C., Busey, T. A., Scolari, M., and Serre, T. (2020). Learning to see by moving. *Current Opinion in Behavioral Sciences*, 36:36—43.
- Roelfsema, P. R., Lamme, V. A., and Spekreijse, H. (2000). The implementation of visual routines. *Vision research*, 40(10-12):1385—1411.

- Rust, N. C. and DiCarlo, J. J. (2010). Selectivity for shape in primate visual cortex. *Journal of Neuroscience*, 30(1):1—12.
- Salman, H., Ilyas, A., Engstrom, L., Kapoor, A., and Madry, A. (2020). Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533—3545.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520.
- Sary, G., Vogels, R., and Orban, G. A. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science*, 260(5110):995—997.
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2020a). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413—423.
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, page 407007.
- Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2020b). Brain-score: Which artificial neural network for object recognition is most brain-like? *Neural Computation*, 32(1):1—44.
- Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815—823.
- Sensoy, M., Kaplan, L., and Kandemir, M. (2018). Evidential deep learning to quantify

- classification uncertainty. *Advances in Neural Information Processing Systems*, 31:3179—3189.
- Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annual Review of Vision Science*, 5:399—426.
- Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007a). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, 165:33—56.
- Serre, T., Oliva, A., and Poggio, T. (2007b). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424–6429.
- Serre, T., Oliva, A., and Poggio, T. (2007c). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15):6424—6429.
- Shibata, K., Watanabe, T., Sasaki, Y., and Kawato, M. (2012). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, 482(7386):471—474.
- Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.
- Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Spoerer, C. J., McClure, P., and Kriegeskorte, N. (2020). Recurrent neural networks can explain neural dynamics during decision making. *Nature Neuroscience*, 23(5):1—10.
- Srivastava, N., Greff, K., and Schmidhuber, J. (2019). Understanding the role of attention in deep learning. *arXiv preprint arXiv:1904.02874*.
- Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. (2021). How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*.

- Sun, J., He, K., Zhang, X., and Ren, S. (2017). Deep residual learning for image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1198—1212.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826.
- Tan, M., Chen, B., Pang, R., Vasudevan, V., Sandler, M., Howard, A., and Le, Q. V. (2019). Mnasnet: Platform-aware neural architecture search for mobile. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2820—2828.
- Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6105–6114.
- Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., and Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835—8840.
- Thomson, M. G., Foster, D. H., and Summers, R. J. (1999). The role of phase information in the perception of natural images. *Vision Research*, 39(20):3475—3485.
- Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582):520—522.
- Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al. (2021). Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34:24261—24272.

- Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 10347–10357.
- Trueblood, J. S., Brown, S. D., Heathcote, A., and Busemeyer, J. R. (2014). Multi-attribute linear ballistic accumulator models of choice and response time: Theory and applications. *Journal of Mathematical Psychology*, 60:1—20.
- Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., and Li, Y. (2022). Maxvit: Multi-axis vision transformer. arXiv:2204.01697.
- Ullman, S., Harari, D., and Dorfman, N. (2016). Visual saliency and the role of microsaccades. *Journal of Vision*, 16(14):1—20.
- VanRullen, R. and Koch, C. (2003). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences*, 7(6):1—10.
- Vickers, D. and Packer, J. (1982). Effects of alternating set for speed or accuracy on response time, accuracy and confidence in a unidimensional discrimination task. *Acta Psychologica*, 50(2):179—197.
- Wang, Q., Wu, B., Zhu, P., Li, P., Hu, Q., and Lu, X. (2020). Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11531—11539.
- Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794—7803.
- Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5):955—968.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2017). The center bias of the human visual system. *IEEE Transactions on Image Processing*, 26(1):1—15.

- Weiss, Y. and Azulay, A. (2025). Scale invariance in deep neural networks. *Nature Machine Intelligence*, 7(1):1—10.
- Wichmann, F. A., Geirhos, R., and Brendel, W. (2023). Deep neural networks and human visual processing. *Annual Review of Vision Science*, 9:1—25.
- Wiecki, T. V., Sofer, I., and Frank, M. J. (2013). Hddm: Hierarchical bayesian estimation of the drift-diffusion model in python. *Frontiers in neuroinformatics*, 7:55610.
- Wightman, R. (2019). Pytorch image models. GitHub repository.
- Winkel, J., Linsley, D., and Serre, T. (2014). Early visual processing and decision-making. *Nature Neuroscience*, 17(5):1—10.
- Witkin, A. P. (1983). Scale-space filtering. Technical Report TR-1144, MIT Artificial Intelligence Laboratory.
- Wong, K.-F. and Wang, X.-J. (2007). Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision making. *Frontiers in Computational Neuroscience*, 1:6.
- Wu, B., Dai, X., Zhang, P., Wang, Y., Sun, F., Wu, Y., Tian, Y., Vajda, P., Jia, Y., and Keutzer, K. (2018). Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. *CoRR*, abs/1812.03443.
- Xie, S., Girshick, R., and Dollár, P. (2017a). Aggregated residual transformations for deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.
- Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017b). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1492—1500.
- Xie, S. and Tu, Z. (2016). Deep neural networks for object detection. In *Proceedings of*

- the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2829—2837.
- Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M., and Mahajan, D. (2019). Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*.
- Yamins, D. L. K. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356—365.
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014a). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619—8624.
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014b). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.
- Yarbus, A. L. (1967). Eye movements and vision. *Springer*.
- Yu, F., Wang, D., and Darrell, T. (2017). Deep layer aggregation. *CoRR*, abs/1707.06484.
- Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3712–3722.
- Zhai, S., Kolesnikov, A., Houlsby, N., and Beyer, L. (2021). Scaling vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12104–12113.
- Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2017). Mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.
- Zhang, J., Han, Y., Poggio, T., and Roig, G. (2019a). Eccentricity dependent neural network with recurrent attention for scale, translation and clutter invariance. *Journal of Vision*, 19:209.

- Zhang, T., Qi, G.-J., Xiao, B., and Wang, J. (2019b). Bag of tricks for image classification with convolutional neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 558—567.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 33(1):15—38.