A Cognitive Architecture for Probing Hierarchical Processing and Predictive Coding in Deep Vision Models

Brennen Hill*
Department of Computer Science
University of Wisconsin-Madison
Madison, WI 53706
bahill4@wisc.edu

Zhang Xinyu
Department of Computer Science
National University of Singapore
Singapore 119077
zhang_xinyu_@u.nus.edu

Timothy Putra Prasetio
Department of Computer Science
National University of Singapore
Singapore 119077
timothy_prasetio@u.nus.edu

Abstract

Despite their success, understanding the internal cognitive processes of modern deep neural networks remains a critical challenge, situated between high-level behavioral evaluations and low-level mechanistic interpretability. Cognitive science, which seeks to explain cognition in biological systems, offers a rich theoretical foundation for bridging this gap. This paper introduces the Visual Cortex Network (VCNet), a novel neural architecture designed as a computational testbed for prominent cognitive theories of vision. VCNet explicitly operationalizes key neuroscientific principles, including the hierarchical organization of distinct cortical areas, dual-stream segregation of information, and top-down predictive feedback. We evaluate VCNet's emergent behaviors and processing capabilities on two specialized benchmarks chosen to probe its architectural priors: the Spots-10 animal pattern dataset, which tests for evolutionarily relevant feature learning, and the Stanford Light Field dataset, which examines the model's ability to process richer, more naturalistic visual data. Our results show that VCNet achieves state-of-the-art performance, with classification accuracies of 92.08% on Spots-10 and 74.42% on the light field dataset, surpassing contemporary models of comparable size. This work demonstrates how integrating principles of cognitive neuroscience into network design can foster more robust and efficient visual processing, offering a promising direction for building and interpreting more capable artificial vision systems.

1 Introduction: Towards Cognitive Interpretability in Vision Models

The remarkable capabilities of contemporary deep learning models in image recognition are often juxtaposed with fundamental limitations that hinder their broader applicability and trustworthiness. These models frequently demand massive labeled datasets [Krizhevsky et al., 2012], struggle to generalize to out-of-distribution examples [Sagawa et al., 2020], and are notoriously susceptible to

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp: Workshop on Interpreting Cognition in Deep Learning Models.

adversarial attacks and partial occlusions [Liu et al., 2022]. These shortcomings highlight a crucial gap in our understanding: we can measure what these models do, but it is much harder to explain how they do it. This challenge calls for a shift in perspective, moving beyond purely behavioral evaluations towards cognitive interpretability, the systematic interpretation of high-level cognition in deep learning models.

The primate visual system, in stark contrast to current artificial systems, represents a paragon of efficiency and robustness. Humans can learn from few examples [Lake et al., 2015], generalize across novel contexts with ease [Geirhos et al., 2018], robustly perceive occluded objects [Hegdé et al., 2008], and perform these feats with unparalleled energy efficiency [Lennie, 2003]. Cognitive science and neuroscience have attributed these capabilities to specific architectural and computational principles of the visual cortex, most notably its fine-grained hierarchical organization [Felleman and van Essen, 1991, Grill-Spector and Malach, 2004] and its use of predictive processing to build and maintain an internal model of the world [Rao and Ballard, 1999, de Lange et al., 2018]. These principles offer not just inspiration for new architectures, but a theoretical framework for understanding the algorithms that support complex visual cognition.

In this work, we embrace this perspective by developing VCNet, a novel neural network whose macro-architecture is explicitly derived from the primate visual cortex. We approach this not merely as an engineering exercise to boost performance metrics, but as a form of computational cognitive science. We seek to understand how embedding cognitive theories of vision into an architecture influences the learning process and the resulting behaviors. Our contributions are framed through this lens:

- We introduce **VCNet**, a deep neural network architecture that serves as a cognitive model, operationalizing the high-level information flow between major visual cortical areas. This includes implementing computational hypotheses about dual-stream processing, recurrent computation, and, critically, top-down predictive feedback.
- We provide a **behavioral account** of VCNet by evaluating it on the Spots-10 animal pattern benchmark. This task is chosen to probe the model's inductive biases, testing the hypothesis that an architecture shaped by evolutionary pressures for pattern recognition will exhibit superior learning and generalization on such data.
- We present a **processing account** by further evaluating VCNet on the Stanford Light Field dataset. This provides evidence that the model's internal algorithms, inspired by the brain's handling of rich visual streams, are better suited for processing multi-view data that more closely approximates the natural input to the human visual system.

2 Related Work

Our research is situated at the confluence of computational neuroscience and neuro-inspired AI.

Neuro-Inspired Architectures The brain has long been a source of inspiration for artificial intelligence. Models like CorNet [Kubilius et al., 2019] have sought to create architectures that not only perform well but also whose internal activations correlate with neural recordings. These models often focus on replicating the feedforward ventral stream. VCNet differentiates itself by modeling a more comprehensive set of cortical principles: (1) the explicit separation and interaction of the ventral and dorsal streams, (2) the inclusion of recurrent dynamics, and (3) the implementation of a top-down predictive coding loop.

Predictive Coding and Generative Models Predictive coding posits that the brain is fundamentally a generative model of its environment [Rao and Ballard, 1999]. Higher cortical areas generate predictions about lower-level sensory input, and only the residual error is propagated forward. This principle is computationally efficient and has deep connections to Bayesian inference [Friston, 2010]. Our implementation of predictive coding in VCNet serves a similar purpose, encouraging the network to learn an internal model of the visual world.

3 The VCNet Architecture: A Computational Model of the Visual Cortex

While a complete, neuron-for-neuron replication of the visual system remains beyond our reach, our research focuses on emulating the macro-scale organization and information flow within the visual cortex. This architectural scaffolding, informed by decades of neuroscientific research, allows us to investigate how high-level cognitive principles can be instantiated as computational mechanisms. VCNet is a deep neural architecture engineered to systematically operationalize these principles.

3.1 Biologically-Inspired Design as Cognitive Hypotheses

Our model's design is predicated on two foundational theories of primate vision, which we treat as core computational hypotheses: the hierarchical nature of representation and the role of predictive feedback in perception.

Hypothesis 1: Hierarchical Processing Creates Abstract Representations Visual information is known to propagate from the retina through a well-defined hierarchy of cortical areas (V1, V2, V3, V4, V5), with each stage specialized for extracting increasingly complex and abstract features [Huff et al., 2023a]. The primary visual cortex (V1) responds to simple elements like oriented edges. It projects to V2, which processes intermediate conjunctions like contours and texture. V2, in turn, projects to higher-order areas: V4, which is crucial for object form and color perception, and V5 (or MT), which is specialized for motion analysis [Huff et al., 2023b]. This complex connectivity, meticulously mapped using neuronal tracing techniques [Fulton, 2001], is hypothesized to form a highly efficient cascade for constructing invariant object representations [Sheth and Young, 2016].

Hypothesis 2: Predictive Coding Refines Perception The visual cortex is not a simple feedforward pipeline. A prominent theory, predictive coding, posits that it is a generative model constantly trying to predict its sensory inputs. In this framework, higher-level cortical areas send top-down predictions to lower-level areas. The ascending, bottom-up signals carry the actual sensory information, and discrepancies between the top-down predictions and bottom-up inputs generate prediction errors. These error signals are then propagated up the hierarchy to update and refine the brain's internal model, with the goal of minimizing future prediction errors and thus forming an efficient and robust representation of the world [Lowet and Uchida, 2024, Urgen and Miller, 2015].

3.2 Architectural Framework and Its Cognitive Correlates

Departing from monolithic CNN architectures, VCNet is structured as a directed acyclic graph that models the known connectivity between major visual cortical areas. The channel capacity of each module is scaled to approximate the relative neuronal populations in its biological counterpart, reflecting a hypothesis about relative computational load.

The architecture operationalizes the dual-stream hypothesis by separating processing into a ventral (what) stream for object identification and a dorsal (where/how) stream for spatial reasoning, which are interconnected to integrate information. To instantiate its core cognitive hypotheses, VCNet employs several specialized computational blocks, including multi-scale front-ends (V1), recurrent processing (MT/MST), attentional modulation (CBAM), and lateral interaction. Critically, we implement the predictive coding hypothesis via a top-down feedback loop from the highest-level representation (AIT) back to the primary visual module (V1).

4 Experiments: Probing the Cognitive Behaviors of VCNet

We benchmarked VCNet's performance to investigate how its cognitive architecture influences its learning and processing capabilities. The experiments were designed not merely to achieve high scores, but to test specific hypotheses about the consequences of our architectural choices, focusing on data modalities that are particularly relevant to the function of biological vision.

4.1 Behavioral Account: Inductive Biases for Animal Pattern Classification

Motivation and Hypothesis A key evolutionary driver for primate vision was the need to rapidly identify patterns for tasks like finding food and avoiding predators [Kaas, 2012, Fornalé et al.,

2012]. The primate visual cortex is therefore highly optimized for this domain. We chose Spots-10 over standard benchmarks like CIFAR-10 to specifically probe the hypothesis that a neuro-inspired architecture possesses a strong inductive bias for these evolutionarily-relevant visual patterns. This experiment serves as a behavioral test of our model's alignment with these cognitive priors.

Methodology We utilized the Spots-10 dataset, which contains 50,000 grayscale 32x32 pixel images across 10 classes of animal patterns [Atanbori, 2024]. We trained VCNet and compared its performance against a suite of established and highly-optimized distilled models.

Table 1: Test accuracy and model size on the Spots-10 animal pattern benchmark. The results probe the model's inductive bias for a cognitively-relevant task. Baseline models are distilled, as noted. All models, including VCNet Mini, were finetuned for the same number of epochs.

Model	Test Accuracy (%)	Model Size (MB)
VCNet Mini (Ours)	**92.08**	**0.04**
DenseNet121 Distiller	81.84	0.07
ResNet101V2 Distiller	80.29	0.07
ResNet50V2 Distiller	79.03	0.07
MobileNet Distiller	78.26	0.07
MobileNetV3-Small Distiller	78.04	0.07

Results and Interpretation As shown in Table 1, VCNet Mini achieves a test accuracy of 92.08%, substantially outperforming the strongest baseline by 10.24 percentage points. To ensure a fair comparison with the lightweight distilled baselines, we scaled down VCNet's hidden-layer widths to create the *Mini* variant, which uses only 0.04 MB of storage. This superior performance and efficiency provide a strong behavioral account supporting our hypothesis. The results suggest that the architectural priors in VCNet foster internal representations that are exceptionally well-suited to the statistical regularities of natural patterns, a task domain central to primate visual cognition.

4.2 Processing Account: Light Field Classification

Motivation and Hypothesis Standard 2D images are flat, information-poor projections of the 3D world. The human visual system (HVS) processes a much richer input, leveraging binocular vision and eye movements to sample the 7D plenoptic function [Adelson and Bergen, 1991]. This allows it to perceive a robust 3D scene representation using cues from the light field, like parallax and view-dependent reflectance Xia et al. [2014]. Light field cameras, which capture both the intensity and angular direction of light rays, provide data that is a much closer analogue to the HVS's input [Lin et al., 2024]. We therefore hypothesize that an architecture designed to emulate the visual cortex will employ a superior processing algorithm for this richer data modality.

Methodology We evaluated VCNet on the Stanford Light Field dataset [Raj et al., 2016]. We compared its performance against established benchmark models, which were pre-trained on ImageNet and finetuned for the same number of epochs as VCNet.

Table 2: Performance and Size Comparison on Light Field Image Classification. This experiment tests the hypothesis that VCNet's architecture provides a more effective processing algorithm for rich, multi-view data. Baselines are standard ImageNet pre-trained models. All models were finetuned for the same number of epochs.

Model	Test Accuracy (%)	Model Size (MB)
VCNet (Ours)	**74.42**	**3.52**
MobileNetV2	72.09	8.66
ResNet18	65.12	42.69
VGG11_BN	51.16	491.39

Results and Interpretation The results, summarized in Table 2, provide a compelling processing account. VCNet achieved the highest test accuracy (74.42%) while being by far the most compact model (3.52 MB). This outcome suggests that the architectural features of VCNet, such as its dual-stream design and multi-scale front-end, constitute a more effective algorithm for integrating the high-dimensional information present in light field data. This supports our hypothesis that emulating the brain's processing strategies leads to models that are better adapted to handle the complexity of naturalistic visual inputs.

5 Conclusion and Future Work

In this work, we introduced VCNet, an architecture guided by the computational principles of the primate visual cortex and designed to serve as a computational testbed for cognitive theories of vision. By explicitly modeling principles like hierarchical dual-stream processing, recurrence, and predictive coding, we moved beyond standard engineering to develop a cognitive architecture. Our experiments provide both behavioral and processing accounts, demonstrating that these neuro-inspired priors lead to superior performance and parameter efficiency on tasks that probe for evolutionarily relevant biases and the ability to handle rich sensory data. Our work underscores the value of grounding AI development in cognitive science, suggesting that this approach can help bridge the gap between what models can do and how they do it.

This research opens several avenues for future work aimed at deepening our cognitive understanding of these models.

- **Systematic Cognitive Probing**: Future work should involve systematic ablation studies, treating the architectural components as testable hypotheses. For example, by removing the predictive coding loop or disabling the dorsal stream, we can quantify their specific contributions to behavior, robustness, and internal representation, mirroring the lesion studies of classical neuroscience.
- **Richer Processing Accounts**: We plan to investigate more sophisticated and biologically plausible mechanisms. This includes exploring alternative implementations of predictive coding, such as those with explicit precision-weighting of error signals, or incorporating temporal prediction. This would allow for a more detailed processing account of how models build and maintain a dynamic internal model of their environment.
- **Developmental Accounts**: Integrating reinforcement learning could allow the model to learn adaptive visual representations tied to behavioral goals. This would enable a developmental account of how goal-directed behavior shapes the emergence of visual cognition, offering a principled path toward understanding and solving the challenge of out-of-distribution generalization
- **Richer Benchmarks**: We will expand our evaluation to include comparisons against more modern architectures (e.g., Vision Transformers) and provide more comprehensive efficiency metrics, including parameter counts and FLOPs, to ensure fair and complete comparisons.

Author Contributions

Brennen Hill: Project lead, conceptualization, software, engineering, investigation, research, writing
Zhang Xinyu: Software, engineering, investigation, research,writing.
Timothy Putra Prasetio: Software, engineering, investigation, research, writing.

References

Edward H. Adelson and James R. Bergen. The Plenoptic Function and the Elements of Early Vision. In Michael S. Landy and J. Anthony Movshon, editors, *Computational Models of Visual Processing*, pages 3–20. MIT Press, Cambridge, MA, 1991.

John Atanbori. SPOTS-10: Animal Pattern Benchmark Dataset for Machine Learning Algorithms. https://paperswithcode.com/paper/spots-10-animal-pattern-benchmark-dataset-for, 2024. Accessed on November 14, 2024.

Floris P. de Lange, Micha Heilbron, and Peter Kok. How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, 22(9):764–779, 2018. doi: 10.1016/j.tics.2018.06.002.

Daniel J. Felleman and David C. van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991. doi: 10.1093/cercor/1.1.1.

Francesca Fornalé, Stefano Vaglio, Caterina Spiezio, and Emanuela Prato Previde. Red-green color vision in three catarrhine primates. *Communicative & Integrative Biology*, 5(6):583–589, 2012. doi: 10.4161/cib.21414.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138, 2010. doi: 10.1038/nrn2787.

James T. Fulton. *Processes in Biological Vision*. Self-published, 2001. Available from: https://www.researchgate.net/publication/225026362_Processes_in_Biological_Vision.

Robert Geirhos, Carlos R. Medina Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge, and Felix A. Wichmann. Generalisation in Humans and Deep Neural Networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7547–7558. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf.

Kalanit Grill-Spector and Rafael Malach. THE HUMAN VISUAL CORTEX. *Annual Review of Neuroscience*, 27:649–677, 2004. doi: 10.1146/annurev.neuro.27.070203.144220.

Jay Hegdé, Fang Fang, Scott O. Murray, and Daniel Kersten. Preferential responses to occluded objects in the human visual cortex. *Journal of Vision*, 8(4):16, 2008. doi: 10.1167/8.4.16.

Trevor Huff, Navid Mahabadi, and Prasanna Tadi. Neuroanatomy, Visual Cortex. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2023a. Updated 2023 Aug 14. Available from: https://www.ncbi.nlm.nih.gov/books/NBK482504/.

Trevor Huff, Navid Mahabadi, and Prasanna Tadi. *Neuroanatomy, Visual Cortex*. StatPearls Publishing, 2023b. Available from: https://www.ncbi.nlm.nih.gov/books/NBK482504/.

Jon H. Kaas. The Evolution of the Visual System in Primates. In Todd M. Preuss and Jon H. Kaas, editors, *Evolution of the Primate Brain*, pages 441–460. Academic Press, 2012. doi: 10.1016/B978-0-12-398315-7.00021-0.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012. Available from: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, N. Apurva Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-like object recognition with high-performing shallow recurrent anns. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pages 12785–12796. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/2019/file/941366319835c0652341977412193b9d-Paper.pdf.

Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. doi: 10.1126/science.aab3050.

Peter Lennie. The cost of cortical computation. *Current Biology*, 13(6):493–497, 2003.

Bingzhi Lin, Yuan Tian, Yue Zhang, Zhijing Zhu, and Depeng Wang. Deep learning methods for high-resolution microscale light field image reconstruction: a survey. *Frontiers in Bioengineering and Biotechnology*, 12:1500270, 2024. doi: 10.3389/fbioe.2024.1500270.

Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, and Jonghye Woo. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 1:1–48, 2022. Available from: https://arxiv.org/pdf/2208.07422.

Adam S. Lowet and Naoshige Uchida. Predictive coding: A distinction — without a difference. *Current Biology*, 34(20):R926–R929, 2024. doi: 10.1016/j.cub.2024.09.026.

Abhilash Sunder Raj, Michael Lowney, Raj Shah, and Gordon Wetzstein. Stanford Lytro Light Field Archive. http://lightfields.stanford.edu/LF2016.html, 2016. Accessed on November 26, 2024.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1xSgCEFvS.

Bhavin R. Sheth and Ryan Young. Two Visual Pathways in Primates Based on Sampling of Space: Exploitation and Exploration of Visual Information. *Frontiers in Integrative Neuroscience*, 10:37, 2016. doi: 10.3389/fnint.2016.00037.

Burcu A. Urgen and Luke E. Miller. Towards an Empirically Grounded Predictive Coding Account of Action Understanding. *The Journal of Neuroscience*, 35(12):4789–4791, 2015. doi: 10.1523/JNEUROSCI.0144-15.2015.

Ling Xia, Sylvia C. Pont, and Ingrid Heynderickx. The visual light field in real scenes. *i-Perception*, 5(7):613–629, 2014. doi: 10.1068/i0654.

A Technical Appendices and Supplementary Material

This appendix provides the supplementary figure and detailed descriptions of the architectural components of VCNet that were summarized in the main paper.

A.1 Architectural Components and Cognitive Correlates

The following describes the implementation of the core cognitive hypotheses within the VCNet architecture.

The Ventral Stream as an Object Recognition Algorithm: This what pathway models the cognitive process of object identification. It progresses from a V1 module through modules representing V2 (Interstripe, Thin Stripe), V4, and the inferotemporal (PIT, CIT, AIT) cortices. This stream is specialized for extracting features related to an object's form and identity.

The Dorsal Stream as a Spatial Processing Algorithm: This where/how pathway models spatial reasoning and motion analysis. It flows from V1 through V2 (Thick Stripe), the middle temporal (MT) and medial superior temporal (MST) areas, and onward toward parietal regions.

These streams are interconnected, allowing the model to integrate what an object is with where it is, a crucial aspect of holistic scene understanding. The final representation is formed in the AIT module, which receives convergent inputs and feeds into the classification layer

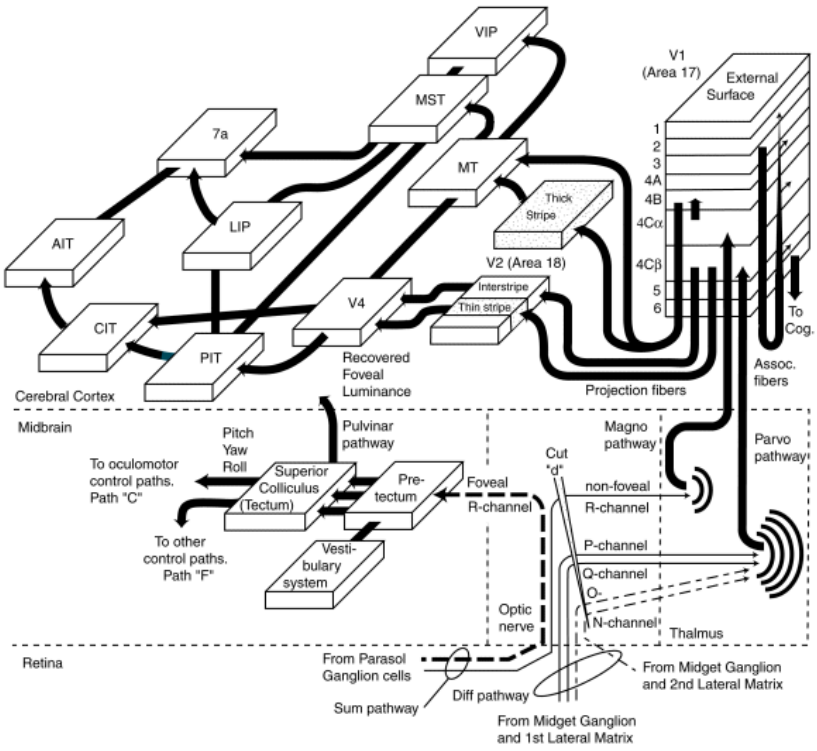

Figure 1: A high-level model of information pathways in the primate visual cortex, illustrating the hierarchical series of feature extraction stages [Fulton, 2001]. This organization, which separates information into specialized processing streams, forms the architectural basis of our cognitive model, VCNet.

A.1.1 Multi-Scale Feature Extraction (V1): Modeling Receptive Field Diversity

To emulate the diverse receptive field sizes in V1, which allows the biological system to perceive features at multiple scales simultaneously, our V1 module processes input through three parallel depthwise separable convolution streams with different kernel sizes (3x3, 5x5, 7x7). The resulting feature maps are concatenated providing a rich, multi-scale initial representation to all subsequent layers, hypothesizing that this is critical for robust feature detection.

A.1.2 Recurrent Processing Blocks (MT/MST): Modeling Iterative Refinement

To model the cognitive process of iterative refinement, where a perceptual hypothesis is updated over time, the MT and MST modules incorporate Recurrent Blocks. These blocks apply a convolutional transformation with shared weights for a fixed number of iterations ($t = 3$), with each iteration receiving the output of the previous one plus a residual connection from the initial input. This models the recurrent processing loops thought to be crucial for motion integration.

A.1.3 Attentional Modulation (CBAM): Modeling Selective Attention

To emulate the brain's ability to focus on salient features, key modules (V1, MT, V4) incorporate a Convolutional Block Attention Module (CBAM). CBAM sequentially infers and applies channel-wise and spatial attention maps, allowing the network to learn to adaptively reweigh features. This is a direct implementation of the cognitive theory of selective attention.

A.1.4 Lateral Interaction Module (V1): Modeling Contextual Modulation

The V1 module includes a Lateral Interaction block, implemented as a convolution followed by channel-wise self-attention within a residual connection. This mechanism simulates the function of

horizontal connections within cortical layers that mediate contextual effects like lateral inhibition, a fundamental process for enhancing edges and contours against their background.

A.1.5 Predictive Coding Loop (AIT to V1): A Direct Implementation of a Cognitive Theory

We implement the predictive coding hypothesis via a top-down connection from the highest level of the ventral stream (AIT) back to the primary visual module (V1). The AIT module, representing the most abstract understanding of the input, generates a prediction of V1 feature activations. This prediction is subtracted from the actual bottom-up V1 activity to compute a prediction error:

$$\epsilon = \text{ReLU}(\text{V1}_{\text{bottom-up}} - \text{AIT}_{\text{top-down}})$$

This error signal ϵ is then used as an auxiliary learning signal. This serves as a potent, cognitively-inspired learning signal, driving the network to learn a better generative model of its visual world.

A.1.6 Neuromodulatory Gating: Modeling Global State Changes

To model the global gain control exerted by neuromodulators (like acetylcholine or dopamine), which can alter the brain's information processing state, we introduce a Neuromodulation block in key modules (V1, MT, V4). This block applies a learnable, channel-wise multiplicative scaling factor to feature maps, allowing the network to dynamically adjust the excitability of different feature pathways based on the input.

A.2 Training Details

All models were trained using the Adam optimizer [Kingma and Ba, 2015] with a learning rate of 10^{-3}. We used a batch size of 16 and applied standard data augmentation techniques, including random horizontal flips and random rotations. All experiments were conducted using Google Colab.