Do Latent Representations of Deep Visual Architectures Follow the Population Manifold Hypothesis of the Primate AIT Cortex?

Ankit Sharma
Indian Institute of Technology Hyderabad
Hyderabad, India
Email: ai25resch11004@iith.ac.in

Sumohana S. Channappayya
Indian Institute of Technology Hyderabad
Hyderabad, India
Email: sumohana@iith.ac.in

***Abstract*—In computational neuroscience, the statistical nature of primate visual responses has long served as a benchmark for efficient coding. Specifically, previous works demonstrated that in the primate Anterior Inferotemporal (AIT) cortex, population sparseness (S_p) significantly exceeds single-neuron sparseness (S_l), demonstrating that while individual neurons respond to relatively simple features, the total available feature space is vast. In this work, we establish a comparative experimental framework to bridge the gap between biological neural responses and the internal representations of Vision Transformers (ViTs) and ResNet-50. We analyze layer-wise kurtosis dynamics across ViT-B/16, ViT-L/16, ViT-B/32, and ResNet-50 using 3000 images of ImageNet-1k validation and Caltech-101 datasets. Our results reveal a "Semantic Snap" that is not a fluke and that it is a robust architectural phenomenon of visual models. Notably, while ResNet-50 inverts the biological signature ($S_l > S_p$), high-capacity transformer-based models like ViT-L/16 achieve a Lehky ratio, mirroring the distributed manifold coding of the AIT cortex. We further identify a significant magnitude gap in representational bandwidth between attention-based and convolutional architectures. These findings, validated by Pareto tail analysis, robust t-statistics, and Lehky ratio calculations, provide a computational link between transformer scaling laws and the Population Manifold Hypothesis in biological vision.**

***Index Terms*—efficient coding, population sparseness, single-neuron selectivity, Lehky ratio, representational bandwidth.**

I. Introduction

The study of the Vision Transformer (ViT) [1] has marked a paradigm shift in computer vision, challenging the long-standing dominance of Convolutional Neural Networks (CNNs). Unlike the local receptive fields of CNNs, ViTs leverage global self-attention mechanisms to process visual information. Recent benchmarks propose these architectures exhibit a high degree of alignment with the primate ventral stream, specifically the Inferior Temporal (IT) cortex, which is responsible for invariant object recognition [2], [3]. However, the internal "representational bandwidth", which measures how much of the network's total feature space is actively used to understand a single image, remains largely opaque.

Classic studies in computational neuroscience proposed the Sparse Coding Hypothesis, which states that the brain represents complex stimuli using a small, active subset of neurons [4], [5]. The two critical metrics widely used in neuroscience to define this efficiency of neural responses are Single-neuron selectivity (S_l) and Population sparseness (S_p) [6], [7], [8]. While S_l measures an individual neuron's selectivity across the dataset, S_p measures the collective efficiency of the latent representation in encoding a specific input instance. While scaling laws for downstream accuracy are well-documented, how model capacity and patch size govern the internal information bottleneck to modulate these metrics and recover the biological anterior inferotemporal (AIT) signature ($S_p > S_l$) remains completely unexplored.

Crucially, no prior work systematically analyzes how ViT hyperparameters, specifically capacity and patch size, modulate kurtosis dynamics to describe AIT's signature $S_p > S_l$ ratio [7]. In this paper, we investigate the layer-wise Kurtosis dynamics of ViT-B/16, ViT-L/16, ViT-B/32, and ResNet-50 to demonstrate that the intermediate layer transition (Layers $5-8$) represents a fundamental semantic bottleneck where the model discards low-level information in favor of learning underlying semantic information of objects, a process grounded in Information Bottleneck Theory [9].

Our primary contribution is the discovery of a capacity-driven shift in coding strategies. We show that while the moderate-capacity ViT-B/16 mirrors the dual-spike (population and selectivity) behavior and evolves localized sparse coding, the larger ViT-L/16 evolves toward a purely distributed population code of primate AIT. This suggests that, statistically, larger transformers distribute semantic information across a broader population of feature units, analogous to the distributed population coding observed in primate AIT. Finally, we show that this "semantic snap", a non-linear phase transition, consistent with the compression phase described in the Information Bottleneck theory [9], is resolution-dependent. This fails to emerge in coarse-patched architectures like ViT-B/32. Our findings are validated by Lehky ratios$(S_p/S_l) \gg 1$, Pareto tail analysis, and t-statistics ($p < 0.001$), providing a robust computational link between transformer scaling and biological population geometry.

II. Related Work

The principle of sparsity has long been recognized as a fundamental constraint in neural computation. Early work by Olshausen and Field [4] demonstrated that simple-cell receptive fields in the primary visual cortex (V1) emerge naturally from the objective of maximizing representational sparsity. Beyond V1, the nature of sparse coding evolves significantly. Lehky et al. [7] provided a comprehensive statistical analysis of the anterior inferotemporal (AIT) cortex, revealing that S_p consistently exceeds S_l. This discovery indicate that the primate brain utilizes a high-dimensional, distributed code to achieve invariant object recognition, rather than relying on a small set of specialized "expert" neurons.

Deep networks serve as powerful computational proxies for the primate ventral stream; late-stage CNN features correlate well with IT cortex firing rates [2], while ViTs display superior alignment with biological representations and human behavior [3], [10]. Mechanistically, Central Kernel Alignment (CKA) reveals that ViTs retain global representations early on, unlike the strict local-to-global hierarchy of CNNs [11]. Moreover, model scaling fosters heavy-tailed activation distributions [12], yet the precise depth-dependent bottleneck where semantic concentration occurs remains poorly characterized.

Information Bottleneck (IB) theory posits that deep representations optimize via a compression phase that discards task-irrelevant entropy through non-linear phase transitions [9].

III. Methodology

A. Dataset and Model Configurations

We evaluate pretrained **ViT-B/16**, **ViT-L/16**, **ViT-B/32**, and **ResNet-50** across 3 independent trials, sampling $3,000$ images from the ImageNet-1K validation set [13] and Caltech-101 datasets at 224×224 resolution. Activations are extracted post-layer norm for ViTs (discarding the CLS token to preserve the spatial patch sequence) and from the unpooled, flattened residual bottleneck maps for ResNet-50. Final statistics are averaged across all 3 trials.

B. Measuring Representational Sparsity

We adapt the kurtosis-based framework from computational neuroscience [7] to evaluate token-level efficiency. Let $X \in \mathbb{R}^{N\times D}$ denote the activation tensor at a given layer, where D is the hidden dimension, and N is the total number of spatial tokens across the dataset ($N =$ images $\times$ patches per image). The scalar $x_{i,j}$ represents the activation of the j^{th} feature unit for the i^{th} token.

1) Token-Level Population Sparseness (S_p): While standard biological formulations typically compute sparseness per holistic stimulus instance, we define a global S_p over the joint token-feature space to obtain a layer-level summary statistics, pooling both intra-image token variance and cross-image distributional shape. We compute this by flattening the activation tensor X and calculating the excess kurtosis over the entire token distribution:

$$S_p = \frac{\frac{1}{ND}\sum_{i=1}^{N}\sum_{j=1}^{D}(x_{i,j}-\mu)^4}{\sigma^4} - 3, \tag{1}$$

where μ and σ are the global mean and standard deviation of activations across all tokens and dimensions at a given layer.

2) Token-Level Single-Unit Selectivity (S_l): Symmetrically, S_l captures per-unit selectivity across the token manifold as the mean excess kurtosis over feature dimensions:

$$S_l = \frac{1}{D}\sum_{j=1}^{D}\left(\frac{\frac{1}{N}\sum_{i=1}^{N}(x_{i,j}-\mu_j)^4}{\sigma_j^4} - 3\right), \tag{2}$$

where μ_j and σ_j are the mean and standard deviation of the j^{th} neuron's response across the token manifold.

C. The Lehky Ratio and Pareto Analysis

To distinguish between local expert and distributed coding, we define the Lehky Ratio as $R_L = S_p/S_l$ [7]. An exploding ratio ($R_L \gg 1$) indicates information concentrated in rare, high-magnitude population events rather than individual neuronal specialization.

The Pareto tail index k was estimated via the Peaks-Over-Threshold (POT) method. For each layer, the threshold was set to the 95^{th} percentile of the activation magnitude distribution; exceedances above this threshold were shifted to the origin and fitted to a Generalized Pareto Distribution (GPD) using MLE. The theoretical basis for this choice is the Pickands–Balkema–de Haan theorem [14], which guarantees that threshold exceedances converge in distribution to a GPD for any heavy-tailed parent distribution, making the GPD the canonical asymptotic model for this regime.

We use k as a *relative* index of tail heaviness across layers and architectures rather than as an absolute distributional claim; cross-layer and cross-dataset consistency of k values: $k > 0$ indicates a heavy power-law tail, $k = 0$ is exponential, and $k < 0$ is finite. Comparing S_p with k determines if the "Semantic Snap" results from a uniform distribution shift or the genuine emergence of highly selective, heavy-tailed latent representations.

D. Statistical Significance and Robustness

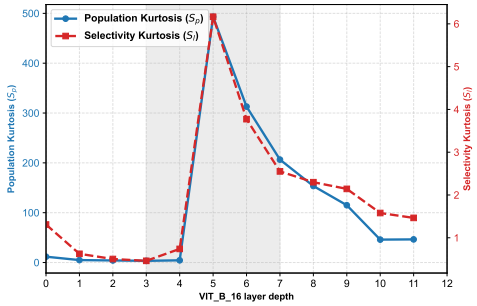
To validate robustness, independent paired t-tests (ttest_rel) compare the mean image-level population kurtosis of bottleneck layers (Layers $5-8$) against early layers (Layers $1-4$). All reported statistics are averaged across 3 independent trials; the resulting tight standard deviations observed in S_p and S_l across these trials confirm the high inter-trial reliability of the measured latent geometries.

IV. Results

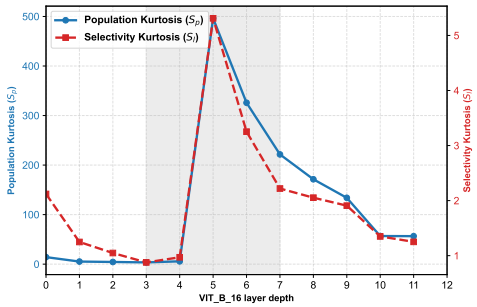
To investigate how vision models compress pixels into semantic concepts, we quantified layer-wise S_p and S_l across different visual models using ImageNet-1k and Caltech-101 datasets.

A. The Canonical Bottleneck and Localized Expert Coding

Early layers produce dense representations with low excess kurtosis, matching primate V1 feature extraction [7]. However, the simultaneous peak at the Layer 6 bottleneck in ViT-B/16 (Fig. 1) stems from its limited capacity (86M

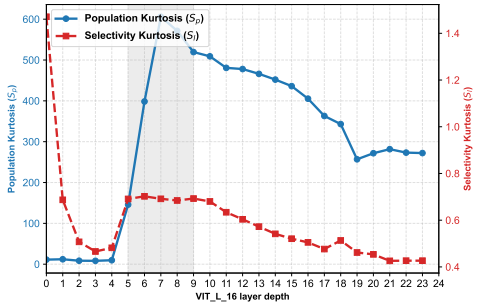


(a) ImageNet-1K: Dual kurtosis plot (S_p Peak L6: **493.21**, S_l Peak L6: **6.16**)

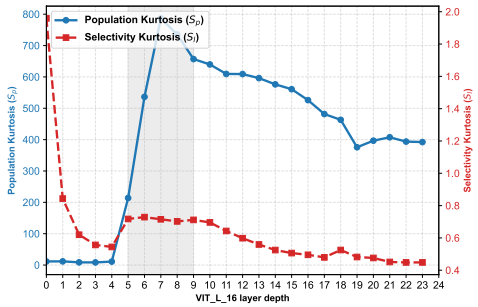


(b) Caltech-101: Dual kurtosis plot (S_p Peak L6: **496.26**, S_l Peak L6: **5.31**)

Fig. 1: ViT-B/16 exhibits a consistent "Semantic Snap" at Layer 6 ($S_p \approx 490$, $S_l \approx 6$, $p < 0.001$). This extreme population suppression indicates localized expert coding, radically diverging from the distributed coding of biological AIT [7].



(a) ImageNet-1K: Dual kurtosis plot (S_p Peak L8: **605.97**, S_l: no peak)



(b) Caltech-101: Dual kurtosis plot (S_p Peak L8: **786.85**, S_l: no peak)

Fig. 2: ViT-L/16 observes "Semantic Snap" at $L8$, reaching extreme ($S_p \approx 600-786$) with no corresponding S_l peak. This suppression drives an enormous Lehky ratio, maximizing representational bandwidth via distributed manifold coding [7].

params, $D = 768$) [1]. Rather than distributing information, it routes features to a few highly selective 'expert' units (high S_l) [15]. Consequently, for any given image, only a small subset of expert neurons are active, resulting in high S_p. This regime resembles a "localized expert regime" [15] in which a few highly selective neurons dominate the representation, efficiently suppressing irrelevant background information but relying on brittle, localized representations and contrasts biological distributed coding.

B. Scaling Leads to Distributed Population Coding

The most significant architectural transition in our study emerges when scaling model capacity to ViT-L/16 ($D = 1024$). In contrast to base-capacity models, ViT-L/16 exhibits a statistical decoupling between S_p and S_l. While S_p retains a massive spike at the semantic bottleneck ($L8$), S_l remains suppressed i.e., no peak is observed (Fig. 2).

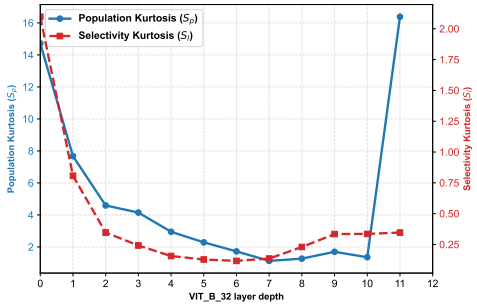
This pattern suggests that the model no longer relies on localized expert coding, but instead adopts a Distributed Population Coding strategy [16]. In this regime, semantic representations are not encoded by isolated units but are distributed across a population of broadly tuned neurons that collectively represent the semantic manifold.

Interestingly, this decoupled signature (high S_p together with low S_l), closely matches the statistical properties reported for the primate AIT cortex by Lehky et al. [7].

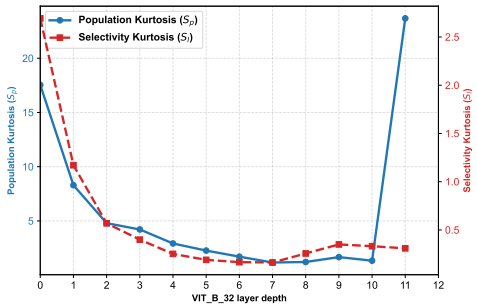
C. The Spatial Resolution Constraint on Semantic Emergence

Crucially, our experiments reveal that the 'Semantic Snap' is not an automatic consequence of depth, but is strictly bound by spatial resolution constraints. When the size of the input patch is increased in the ViT-B/32 architecture, the canonical kurtosis spike collapses entirely (Fig. 3).

In ViT-B/32, the architectural constraint of utilizing 32×32 patches forces the model to heavily compress early-stage visual information. Because each spatial token aggregates a massive grid of pixels, it effectively acts as a structural low-pass filter, likely discarding high-frequency spatial detail that is necessary for fine-grained semantic discrimination. This early-stage information bottleneck prevents the emergence of the semantic snap seen in ViT-B/16 and leads to the observed collapse in representational sparsity (S_p). This mirrors findings in biological systems where low-pass-filtered visual inputs lead to broadly tuned neural responses and a failure of the AIT cortex to achieve high-level representational sparsity [17]. Consequently, we argue that the semantic transformation is highly resolution-dependent; it is not merely a product of model depth, but is fundamentally conditional on fine-grained spatial sampling of the visual manifold.

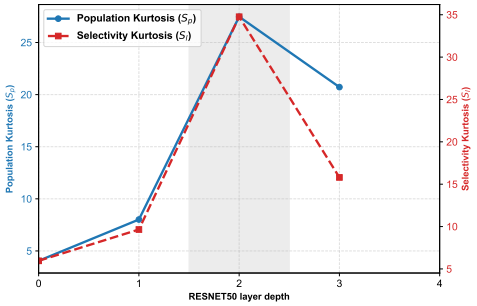


(a) ImageNet-1K: Dual kurtosis plot (S_p: no peak, S_l: no peak)

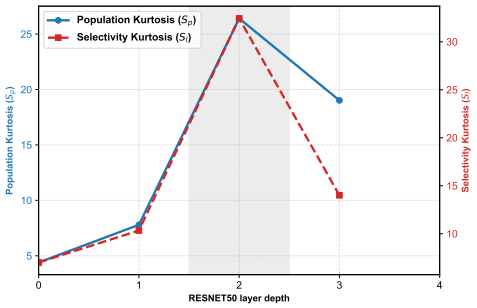


(b) Caltech-101: Dual kurtosis plot (S_p: no peak, S_l: no peak)

Fig. 3: ViT-B/32 Resolution Failure: The absence of a sharp sparsity peak indicates that coarse 32×32 patches cause Semantic Blurring, preventing the high-sparsity seen in higher-resolution models.



(a) ImageNet-1k: Dual kurtosis plot (S_p Peak L3: $\mathbf{27.47}$, S_l Peak S3: $\mathbf{34.76}$)



(b) Caltech-101: Dual kurtosis plot (S_p Peak L3: $\mathbf{26.41}$, S_l Peak S3: $\mathbf{32.44}$)

Fig. 4: ResNet-50 Lehky Violation: While a bottleneck is achieved at Stage 3, the inverted ratio ($S_l > S_p$) reveals highly localized expert coding, where a minimal subset of hyper-specialized channels carries the semantic load rather than a distributed population.

D. ResNet-50: The "Expert Overfit" Signature

ResNet-50 exhibits a coding signature that fundamentally inverts biological norms. At the Stage 3 bottleneck, we observe selectivity dominance ($S_l > S_p$) (Fig. 4), directly violating the $S_p \gg S_l$ signature characteristic of the primate AIT cortex [7].

- The Magnitude Gap: Beyond the distribution, a massive $17\times$ gap in "representational bandwidth" exists between architectures. While ResNet-50 peaks to $(S_p \approx 26-28)$, ViT-B/16 achieves extreme sparseness ($S_p \approx 493-495$). This implies that local convolutional kernels struggle to suppress task irrelevant spatial entropy compared to the global suppression power of self-attention.
- Localized Expert vs. Distributed Population Coding: ResNet-50 relies on highly localized expert coding, where hyper-specialized individual units ($S_l > S_p$) carry the semantic load. This rigid strategy lacks the robust, distributed population coding manifold exhibited by larger Transformers and biological systems.

E. Statistical Reliability and Significance

To validate our experiments:

- Hypothesis Testing: We evaluate the Null Hypothesis (H_0) of no significant difference in token-level representational sparsity between early and mid-stage layers. Across all semantic snaps, $p < 0.001$, rejecting H_0 ($t = 101.0$ for ViT-B/16; $t = 116.2$ for ViT-L/16). These large t-statistics arise from our token-level formulation ($n \approx 588,000$ spatial tokens). However, since tokens from the same image are correlated, the effective number of independent observations is approximately $N_{\text{img}} = 3{,}000$ (one per image). Under this correction, the effect remains highly significant ($t \approx 7-10$, $p \ll 0.001$), reflecting a genuine and large effect, underscoring the structural magnitude of the snap.
- Trial Stability: Low standard deviations across three random trials (Table II) confirm cross-distribution stability, with S_l variance remaining exceptionally low ($\sigma < 0.05$) across all variants.

This high degree of convergence across random subsets and rejection of H_0 confirms that the semantic snap is a robust architectural phenomenon rather than an artifact of specific data distribution.

V. Conclusion

We have demonstrated that the "Semantic Snap", a non-linear sparsity phase transition mirroring the primate AIT cortex, emerges systematically across ViT architectures (Figs. 1, 2). Our results show that ViT-L/16 achieves the strongest statistical alignment with the AIT criterion (Table I: $S_p \approx 606-769$, $S_l \approx 0.7-0.8$, Lehky up to $961\times$). Similarly,

Model	Bottleneck	S_p (Kurt)	S_l (Kurt)	Lehky Ratio	Pareto k	t-stat	Coding Regime	Verdict
High-Resolution Performers								
ViT-B/16	L6	493.2/496.2	6.16/5.31	80.1/93.5	0.27/0.37	101.0/–	Localized Expert	Biological (Nascent)
ViT-L/16	L8	**605.9/768.8**	0.7/0.8	**864/961**	0.36/0.38	116.2/–	Distributed (AIT)	Biological (Superior)
ResNet-50	S3	27.4/26.4	34.8/32.4	0.79/0.81	0.28/0.06	–	Expert Overfit	Lehky Violation (Inverted)
Resolution Failure								
ViT-B/32	Diffuse	no peak	no peak	46.5/45.8	0.23/0.22	–	Resolution Limited	Semantic Blurring (Weak)

TABLE I: **Cross-Dataset Summary (ImageNet-1K / Caltech-101)**:(1) **Sparsity Convergence**: High-resolution Transformers achieve extreme S_p. (2) **Statistical Alignment**: ViT-L/16 recovers the biological AIT signature ($S_p \gg S_l$). (3) **Architectural Inversion**: CNNs (ResNet-50) rely on selectivity-dominant expert coding ($S_l > S_p$). (4) **Resolution Constraint**: Coarse spatial tokenization (ViT-B/32) prevents high-intensity semantic emergence.

Model	Layer	Dataset	S_p (σ)	S_l (σ)
ViT-B/16	L6	ImageNet	4.37	0.02
		Caltech	2.67	0.03
ViT-L/16	L8	ImageNet	8.63	0.004
		Caltech	4.61	0.001
ResNet-50	S3	Both	0.95 / 0.22	0.54 / 0.19
ViT-B/32	L24	Both	0.17 / 0.65	0.006 / 0.003

TABLE II: Statistical Reliability Across 3 Randomized Trials.

ViT-B/16 also exhibits a strong biological signature ($S_p \approx 493-496$, Lehky up to $93.5\times$) within a localized expert regime. Crucially, our analysis of ViT-B/32 reveals that while coarse resolution maintains a distributed population ratio (Lehky $\approx 46\times$), it fails to generate a semantic bottleneck. This implies that fine-grained spatial tokenization determines the absolute magnitude of entropy suppression, while global attention drives the AIT-like population dominance.

In contrast, ResNet-50 exhibits a clear semantic bottleneck ($S_p \approx 27$) yet fails biological convergence ($S_l \approx 34 > S_p$, Lehky $\approx 0.8\times$). This "Lehky Violation" reveals that local convolutional kernels force the model into highly localized channel selectivity, where a few hyper-specialized channels dominate the response, contrasting with the distributed manifold coding observed in biological visual processing. Our comparative framework establishes a clear scaling law: high-resolution global attention acts as a critical mechanism for recovering the statistical signature of the primate visual cortex, highlighting how architectural inductive biases govern representational geometry.

Limitations and Future Directions: While this study establishes a robust foundational baseline for token-level representational sparsity across standard ViT and CNN archetypes, fully characterizing these latent spaces requires broader investigation. Future work will expand this framework by incorporating representational similarity metrics (e.g., CKA, RSA) and information-theoretic measures to directly quantify manifold dimensionality. Furthermore, extending this analysis to hybrid architectures (e.g., ConvNeXt) across larger-scale datasets, and evaluating how these geometric phase transitions correlate with downstream functional performance (e.g., zero-shot transfer, robustness), remain important next steps to definitively link statistical sparsity with model generalization.

REFERENCES

[1] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[2] M. Schrimpf, J. Kubilius, H. Hong, N. J. Majaj, R. Rajalingham, E. B. Issa, K. Kar, P. Bashivan, J. Prescott-Roy, F. Geiger, *et al.*, "Brain-score: Which artificial neural network for object recognition is most brain-like?," *BioRxiv*, p. 407007, 2018.

[3] S. Tuli, I. Dasgupta, E. Grant, and T. L. Griffiths, "Are convolutional neural networks or transformers more like human vision?," p. 1844 – 1850, 2021. Cited by: 34.

[4] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[5] B. A. Olshausen and D. J. Field, "Natural image statistics and efficient coding," *Network: computation in neural systems*, vol. 7, no. 2, p. 333, 1996.

[6] S. R. Lehky and A. B. Sereno, "Comparison of shape encoding in primate dorsal and ventral visual pathways," *Journal of neurophysiology*, vol. 97, no. 1, pp. 307–319, 2007.

[7] S. R. Lehky, R. Kiani, H. Esteky, and K. Tanaka, "Statistics of visual responses in primate inferotemporal cortex to object stimuli," *Journal of Neurophysiology*, vol. 106, no. 3, pp. 1097–1117, 2011.

[8] L. Franco, E. T. Rolls, N. C. Aggelopoulos, and J. M. Jerez, "Neuronal selectivity, population sparseness, and ergodicity in the inferior temporal visual cortex," *Biological cybernetics*, vol. 96, no. 6, pp. 547–560, 2007.

[9] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[10] C. Conwell, D. Mayo, A. Barbu, M. Buice, G. Alvarez, and B. Katz, "Neural regression, representational similarity, model zoology & neural taskonomy at scale in rodent visual cortex," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5590–5607, 2021.

[11] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," *Advances in neural information processing systems*, vol. 34, pp. 12116–12128, 2021.

[12] U. Sharma and J. Kaplan, "Scaling laws from the data manifold dimension," *Journal of Machine Learning Research*, vol. 23, no. 9, pp. 1–34, 2022.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[14] J. Pickands III, "Statistical inference using extreme order statistics," *the Annals of Statistics*, pp. 119–131, 1975.

[15] J. S. Bowers, "On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience.," *Psychological review*, vol. 116, no. 1, p. 220, 2009.

[16] A. Pouget, P. Dayan, and R. Zemel, "Information processing with population codes," *Nature Reviews Neuroscience*, vol. 1, no. 2, pp. 125–132, 2000.

[17] D. J. Field, "What is the goal of sensory coding?," *Neural computation*, vol. 6, no. 4, pp. 559–601, 1994.