# Modeling Visual Cortex by Maximizing Layerwise Multiscale Manifold Capacity

**Thomas Edward Yerxa**[1] **SueYeon Chung**[2,3,4] **Eero P. Simoncelli**[1,2]

[1] Center for Neural Science, New York University
[2] Center for Computational Neuroscience, Flatiron Institute
[3] Department of Physics, Harvard University
[4] Kempner Institute, Harvard University

### Abstract

Deep neural networks provide strong predictive models of primate visual cortex, but are typically trained with end-to-end objectives requiring global backpropagation. We introduce ST-MMCR, a layerwise self-supervised learning scheme that trains successive stages of a convolutional hierarchy with local objectives matched to their receptive-field scale. The objective is based on maximum manifold capacity representations: temporally nearby views are pooled into compact, discriminable manifolds using projection, local spatial pooling, temporal pooling, and a nuclear-norm loss. This implements complexity matching through the architecture rather than through separate hand-crafted augmentation streams as was done in previous work. Evaluated on macaque V1, V2, and V4 datasets and human-aligned object-classification benchmarks, ST-MMCR matches or exceeds architecture-matched supervised and self-supervised baselines in neural predictivity, approaches adversarially robust models, and improves out-of-distribution and human-aligned behavior when used as a visual front end.

## Motivation

Task-optimized deep neural networks are among the best current phenomenological models of the primate ventral stream (Willeke et al., 2023; Yamins et al., 2014; Zhuang et al., 2021), but their biological interpretation is limited by their reliance on global backpropagation and by the weak constraints imposed on intermediate representations. Layerwise learning offers a natural alternative: each stage receives its own local training signal, allowing internal representations to be shaped directly. Prior work has shown that such learning is most successful when each layer's task complexity is matched to its representational capacity (Parthasarathy et al., 2024). However, existing implementations often achieve this using different augmented inputs for different layers, whereas biological visual systems learn from a common visual stream. We propose a layerwise objective that instead matches task complexity to representational capacity through spatial and temporal scale.

## Approach

We train a three-stage AlexNet-like convolutional network, with stages corresponding approximately to V1, V2, and V4 (Krizhevsky et al., 2012). Each stage receives a self-supervised loss computed from synthetic videos generated by smoothly transforming ImageNet images over time. Because nearby frames are related by simpler transformations than distant frames, temporal pooling provides a natural axis for controlling invariance; because receptive fields grow with depth, fixed-size pooling on deeper feature maps corresponds to larger effective regions of the input.

Our objective is based on maximum manifold capacity representations (MMCR), which cast self-supervised learning as an approximate optimization of the number of linearly separable transformation manifolds that can be stored in a representation (Chung et al., 2018; Yerxa et al., 2023). For each input video $\mathbf{x}_b$, the $i$th stage produces a spatiotemporal feature map. After the projection head $g_{\phi_i}$ and local spatial pooling, frame-wise responses are projected onto the unit sphere by $\Pi$, then averaged over a local temporal window to form a centroid:

$$\begin{aligned}
\mathbf{c}_b^{(1)} &= \mathrm{LTP}\left(\frac{t_f}{4}\right) \circ \Pi \circ \mathrm{LSP}(s_f) \circ g_{\phi_1} \circ f_{\theta_1}(\mathbf{x}_b), \\
\mathbf{c}_b^{(2)} &= \mathrm{LTP}\left(\frac{t_f}{2}\right) \circ \Pi \circ \mathrm{LSP}(s_f) \circ g_{\phi_2} \circ f_{\theta_2} \circ f_{\theta_1}(\mathbf{x}_b), \\
\mathbf{c}_b^{(3)} &= \mathrm{LTP}(t_f) \circ \Pi \circ \mathrm{LSP}(s_f) \circ g_{\phi_3} \circ f_{\theta_3} \circ f_{\theta_2} \circ f_{\theta_1}(\mathbf{x}_b).
\end{aligned} \tag{1}$$

Here $\mathrm{LSP}(s_f)$ denotes local spatial pooling, $\mathrm{LTP}(t)$ temporal pooling over $t$ frames, and $\Pi$ channel-wise normalization. The same computation is used at each stage, but the effective spatial region and temporal window increase with depth.

For a minibatch of $B$ videos, the centroids from stage $i$ are assembled into a matrix

$$\mathbf{C}^{(i)} = \left[\mathbf{c}_1^{(i)}, \ldots, \mathbf{c}_B^{(i)}\right]. \tag{2}$$

Extended abstract presented at the *9th Conference on Cognitive Computational Neuroscience*, New York, NY, USA, 2026. 

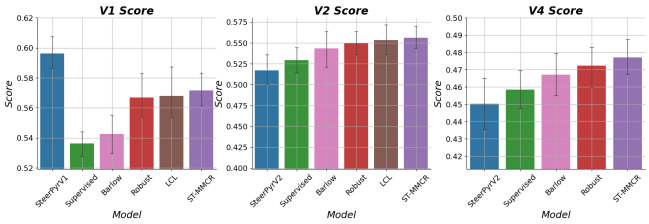


Figure 1: **Neural predictivity across models.** ST-MMCR is evaluated against architecture-matched supervised, self-supervised, robust, random, and handcrafted baselines on neural recordings from macaque V1, V2, and V4. Bars show the best layer score for each model and dataset.

The MMCR loss for each stage is the negative nuclear norm of this centroid matrix,

$$\mathcal{L}^{(i)}_{\text{MMCR}} = -\left\|\mathbf{C}^{(i)}\right\|_*, \tag{3}$$

and the full training objective is

$$\mathcal{L} = \sum_{i=1}^{3} \mathcal{L}^{(i)}_{\text{MMCR}}. \tag{4}$$

This objective aligns temporally related views while spreading centroids from different images apart, combining invariance and discriminability in a single term. Because the centroid matrix has one column per input video rather than one column per frame, the objective does not grow quadratically with the number of temporally pooled samples, making it well suited to learning invariances over extended temporal windows. Gradients from each $\mathcal{L}^{(i)}_{\text{MMCR}}$ update only the corresponding stage and projection head, yielding layerwise learning without global backpropagation across stages. This makes the learning problem increasingly difficult across the hierarchy without changing the input distribution seen by each stage. In this sense, ST-MMCR converts hand-designed layer-specific augmentation schedules into a single architectural principle: deeper stages integrate information over larger regions of space and time.

# Evaluation

We evaluated model-brain alignment using recordings from macaque V1, V2, and V4. For V1/V2, stimuli were synthetic naturalistic textures and spectrally matched noise images (Freeman et al., 2013; Ziemba et al., 2016); for V4, stimuli were natural images and parametrically texturized variants (Lieber et al., 2024). Model activations were mapped to trial-averaged firing rates using partial least-squares regression, and performance was measured by the median cross-validated Pearson correlation across neurons.

ST-MMCR reproduces the key prediction of complexity-matched learning: each stage is most predictive of its corresponding cortical area when the spatial and temporal scale of the learning problem is matched to that stage's receptive-field size. With stage-adapted pooling, a single network trained on one stream of full-size videos achieves near-peak predictivity across V1, V2, and V4. Across baseline comparisons, ST-MMCR outperforms standard supervised AlexNet and end-to-end self-supervised AlexNet in all three areas, while approaching the performance of an adversarially robust AlexNet baseline. Ablations indicate that removing layerwise gradient isolation or intermediate losses modestly reduces neural predictivity.

We also evaluated classifiers trained on frozen ST-MMCR front ends using ImageNet, out-of-distribution image sets, and human choice-alignment metrics (Geirhos et al., 2021). Although ST-MMCR front ends reduce standard ImageNet accuracy relative to full supervision, the V2-stage front end improves out-of-distribution generalization and agreement with human choices and errors, suggesting that biologically motivated self-supervision can improve behavioral alignment without directly optimizing for it.

# Discussion

ST-MMCR provides a biologically motivated alternative to end-to-end task optimization for modeling the ventral stream. Its central contribution is to implement complexity-matched layerwise learning without requiring distinct input distributions or hand-crafted augmentation strengths for different layers. Instead, task complexity is controlled by factors that also organize the biological hierarchy: increasing receptive-field size and temporal integration.

The learning rule uses local losses, applies the same objective form at each stage, and computes that objective using operations related to common cortical motifs, including rectification, normalization, spatial pooling, and temporal integration (Carandini & Heeger, 2012). While the final nuclear-norm computation does not yet have a direct biological implementation, the framework suggests a concrete target for future work on local circuit approximations to capacity-based objectives. More broadly, these results suggest that constraints on credit assignment and task scale can produce representations competitive with task-optimized models in neural predictivity and better aligned with human behavior under distribution shift. This provides a possible bridge between efficient-coding accounts of sensory representation and modern self-supervised objectives for hierarchical vision models.

## Disclosure

The authors acknowledge the use of a Large Language Model (LLM), ChatGPT, to assist with writing the text of the extended abstract.

## References

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature reviews neuroscience*, *13*(1), 51–62.

Chung, S., Lee, D. D., & Sompolinsky, H. (2018). Classification and geometry of general perceptual manifolds. *Physical Review X*, *8*(3), 031003.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, *16*(7), 974–981.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, *34*, 23885–23899.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Lieber, J. D., Oleskiw, T. D., Simoncelli, E. P., & Movshon, J. A. (2024). Responses of neurons in macaque v4 to object and texture images. *BioRxiv*, 2024–02.

Parthasarathy, N., Henaff, O. J., & Simoncelli, E. P. (2024). Layerwise complexity-matched learning yields an improved model of cortical area V2. *Transactions on Machine Learning Research*.

Willeke, K. F., Restivo, K., Franke, K., Nix, A. F., Cadena, S. A., Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Ecker, A. S., et al. (2023). Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. *bioRxiv*, 2023–05.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, *111*(23), 8619–8624.

Yerxa, T., Kuang, Y., Simoncelli, E., & Chung, S. (2023). Learning efficient coding of natural images with maximum manifold capacity representations. *Advances in Neural Information Processing Systems*, *36*, 24103–24128.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118.

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque v2. *Proceedings of the National Academy of Sciences*, *113*(22), E3140–E3149.