Under review at ICML 2024 AI for Science workshop

## Abstract

Across the primate cortex, neurons that perform similar functions tend to be spatially grouped together. In high-level visual cortex, this widely observed biological rule manifests itself as a modular organization of neuronal clusters, each tuned to a specific object category. The tendency toward short connections is one of the most widely accepted views of why such an organization exists in the brains of many animals. Yet, how such a feat is implemented at the neural level remains unclear. Here, using artificial deep neural networks as test beds, we demonstrate that a topographical organization similar to that in the primary, intermediate, and high-level human visual cortex emerges when units in these models are laterally connected and their weight parameters are tuned by top-down credit assignment. Importantly, the emergence of the modular organization in the absence of explicit topography-inducing learning rules and objectives questions their necessity and suggests that local lateral connectivity alone may be sufficient for the formation of the topographic organization across the cortex.

## 1. Introduction

Functional organization, arrangement of neurons across the cortical sheet according to their functional similarity, stands out as a ubiquitous phenomenon in neuroscience research, manifesting in topographic maps within various brain regions such as the visual system, auditory cortex, parietal cortex, sensorimotor areas, and entorhinal cortex (Hubel & Wiesel, 1962; Harvey et al., 2013; Humphries et al., 2010; Obenhaus et al., 2022; Gu et al., 2018; Wong et al., 1978). These organized patterns have played a critical role in advancing our understanding of both the functionality and potential dysfunctions of the cortex, the roots of which can be traced back to the groundbreaking work of Wilder Penfield in the 1950s and 1960s (Penfield & Rasmussen, 1950). Using electrical stimulation, Dr. Penfield made one of the first attempts in producing a global functional map of the human cortex, thereby unveiling the intricate ways in which the brain governs perception and action.

Building upon this foundational work, subsequent studies, notably those by Hubel and Wiesel (Hubel & Wiesel, 1962), provided a more detailed view of the topographical organization within the visual cortex, particularly emphasizing the structured arrangement of cortical columns based on orientation selectivity and ocular dominance. These observations were later extended to higher stages of the ventral visual cortex with the discovery of neuronal clusters that were selective for faces, scenes, and body parts among others (Margalit et al., 2023).

The ubiquitous cortical topography in primates has highlighted two profound questions in the past decades: “Why are neurons spatially organized in this highly regular fashion?” and, “What are the neural mechanisms underlying such multi-scale topography across the cortex?”. A leading theory, widely accepted as a plausible answer to the why question is the Wiring Cost Minimization (WCM) (Jacobs & Jordan, 1992), which explains topographical organization as the product of an evolutionary process that minimizes the amount of nerve volume connecting neurons across and within cortical areas. Based on this theory, various models have been proposed to answer the ”how” question of cortical topography using a variety of computational approaches for inducing position-dependent covariation between unit responses. Among those, several models use position-aware update rules (Kohonen, 1982), while others use learning objective functions that either enforce cortex-like position dependent pairwise response/weight correlations (Lee et al., 2020; Margalit et al., 2023; Lu et al., 2023), or directly penalize the weight connections between distant units (Lu et al., 2023). Each of these approaches have recapitulated aspects of the topographical organization when adopted to simulate the organization of neurons in the visual cortex. For example, many of these models have replicated category-selective neuronal clusters such as face and place selective areas in the primate inferotemporal cortex. However, the formation of such clusters is not uniquely predicted by any one approach and can similarly be replicated by a number of computational approaches.

Moreover, these prior models face certain limitations. First, several of these models have limited scope in terms of their correspondence with the brain [i.e. which cortical regions, ref to (Blauch et al., 2022; Lee et al., 2020) or the variety of object selective categories that could be explained (Blauch et al., 2022). Second, a number of these models which rely on self-organizing principles factorize learning into separate stages of representation learning and topography induction and assume sequential or independent mechanisms governing the two without explicit descriptions (Durbin & Mitchison, 1990; Doshi & Konkle, 2023). Third, all topographical models exhibit substantially diminished performance on ecologically relevant tasks such as object categorization (Blauch et al., 2022; Durbin & Mitchison, 1990; Doshi & Konkle, 2023), questioning the utility of topographical organization. Finally, most prior models rely on strong assumptions about the underlying biological network that are often not entirely met or remain unspecified.

There is a substantial evidence that certain aspects of the topographical organization begin to shape prior to eye opening (Smith et al., 2018). For example, experiments using different animal models like cats, macaques, and ferrets have demonstrated that ocular dominance columns and long-range correlational structures already exist prior to birth and without any visual experiences (Swindale, 1996). These maps continue to further develop after birth and converge to their mature state. Importantly, sensory deprivation or restrictions could lead to strong degradation in the resulting cortical map (Hubel et al., 1977). Moreover, there is a large body of literature that using tracing techniques have documented the details of the synaptic connectivity across neurons in primary visual cortex of many animal species (Muir et al., 2011; Muir & Douglas, 2011). Three key insights from these work are: 1) neurons in the primary visual cortex are far more likely to establish lateral connections when they have similar functional selectivity; 2) the likelihood of lateral connections often decay with increasing cortical distance, although there exists patches of distant neurons with strong lateral connections; 3) patchy lateral connectivity patterns called (aka Daisy patterns) have been reported in almost all animal species with strong topographical cortices and is virtually non-existent in animals with weak organization such as in rodents (Muir et al., 2011).

These results could be interpreted in one of two ways: 1) that neurons with similar selectivity find and establish lateral connections between themselves or; 2) that distinct patterns of selectivity emerge as a function of innate lateral connectivity. The former possibility requires the neurons to have the capacity to trace other neurons with similar functionality and establishing synapses in between while the latter states that neuronal selectivity across the neural network emerges as a function of learning in a neural network with prespecified constraints in connectivity (e.g. enforced via genetic encoding).

## 2. Locally Laterally Connected Neural Networks (LLCNNs)

Our work introduces Local Laterally Connected Neural Network (LLCNN) (Fig. 1), a topographical neural network model designed to overcome existing limitations. The LLCNN implements the hypothesis that *local lateral connections are sufficient for the emergence of topographic organization in deep neural networks* which challenges the prevalent consensus of the wiring length cost hypothesis. Drawing inspiration from the well-established concept of local lateral connectivity (LLC) (Muir et al., 2011) ubiquitous in species with pronounced topographical organization within the cortex, we seamlessly integrated LLC into the layers of a convolutional neural network. Notably, our approach not only induces a naturally emerging smooth topographic organization but also facilitates the learning of robust representations as an intrinsic byproduct of computation.

In particular, our model achieves the following outcomes: 1) replicates the arrangement of neurons in the early visual cortex, aligning with orientation, spatial frequency, and color (Fig. 2,3); 2) forms object-selective clusters in deep layers analogous to those found in the human inferotemporal cortex (Fig. 4, 5); 3) predicts unit selectivity in inferotemporal regions with previously uncharted object-selectivity (Fig. 6); 4) enhances the trade-off between object recognition performance and cortex-like topography compared to previous models (Fig. 7); and 5) exhibits notable improvement in robustness against adversarial noise (Fig. 7), suggesting a potential functional role for LLC in learning robust representations. Our results underscore the efficacy of incorporating local lateral connectivity in artificial neural network models for reproducing cortex-like topographical organization. Importantly, this achievement is realized without the necessity for specific topography-inducing learning objectives or rules, showcasing the model's inherent capacity to develop such organization.

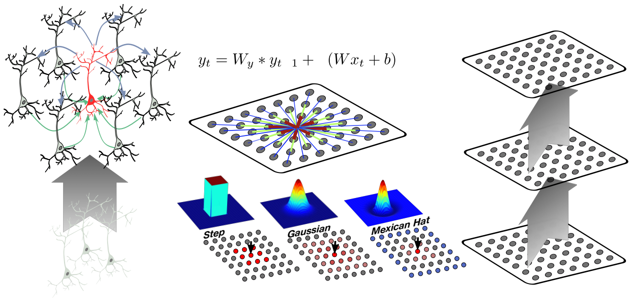

*Figure 1.* **Local Laterally Connected Neural Networks (LLCNNs)** There are not only connections across layers but also lateral connections **within layers** in visual cortex. We incorporated 3 types of lateral connections (i.e. Mean, Gaussian and Mexicanhat) into LLCNNs to induce the modular organization.

# 3. Results

## 3.1. V1 Topography

Some of the earliest and most established investigations of cortical topography have been performed on the primary visual cortex of cats, macaque monkeys and other primates (Hubel & Wiesel, 1962; Blasdel & Salama). Broadly, these studies have described the organization of cortex as a periodic map whose periodicity is governed by a number of sensory-dependent features including ocular dominance, eccentricity, visual angle, orientation selectivity, and spatial frequency.

Given the abundance of prior experimental work in this area, we first examined whether and to what degree the LLCNN model could replicate the same organization in its early layers that best matched the primary visual cortex. We evaluated the selectivity of each unit in each model layer to the stimulus orientation, spatial frequency, and color. For this, we used a stimulus set consisting of gratings with different orientations (0-180 degree), spatial frequencies (1-14Hz), and chromaticity (black/white vs. colored) (Margalit et al., 2023).

We then visualized the selectivity map of units within each layer for each stimulus factor (Fig. 2A). The resulting maps demonstrated smoothly transitioning unit selectivity along the two spatial axes of the simulated cortical sheet for all three stimulus features. We observed that the change in unit selectivity increased to a value of 1 with increasing distance between unit pairs (Fig. 2C), where 1 is the selectivity change expected from random arrangement of units. Likewise, the pairwise response correlation decreased exponentially (Pearson correlation, Fig. 2B). The observed decay in pairwise unit correlation with increasing distance in the early layers of the LLC model suggests that proximal units exhibit a more congruent response to a set of sine grating stimuli than their distant counterparts, akin to prior observations from the macaque primary visual cortex (Muir et al., 2011). Furthermore, the smooth maps extended across the layers of the LLCNN further mimicking the smooth transitioning of these feature selectivity in the brain across cortical areas (e.g. V1 to V2).

It was also reported that 60-75% of laterally connected neurons in the primary visual cortex of tree shrew (Bosking et al., 1997), cat (Schmidt et al., 1997), and macaque monkeys (Malach et al., 1993) have orientation selectivity that falls within ±45° of the preferred orientation of the neuron at the injection site. We investigated the distribution of difference in orientation selectivity between each unit and its neighboring units and found that 60% of neighboring units have selectivity within 45° of the center unit (Fig. 2D).

While there are substantial differences in selectivity maps across individuals, these changes are still obey particular rules. Notably, using two-photon imaging in macaque's primary visual cortex, it was shown that the selectivity map gradients (directions on the cortical surface along which the feature tuning changes) of orientation and spatial frequency were markedly skewed towards orthogonality. We investigated whether the gradients of orientation and spatial frequency selectivity in the LLCNN model followed the same pattern. We computed the gradients of selectivity to each feature and computed the angle between the two gradients at intersection points (Fig. 2E). Our results showed a strong tendency towards crossings at angles close to 90°, echoing the prior findings in the primate primary visual cortex.

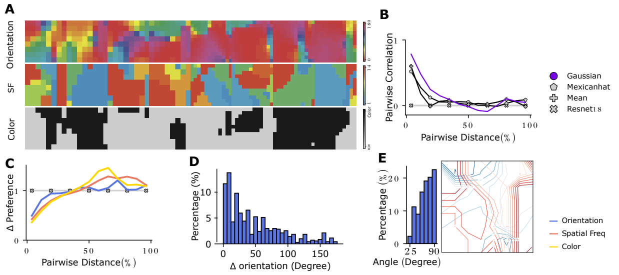

*Figure 2.* **V1 Topography** Early LLC model layers replicate hallmarks of visual processing in the primate primary visual cortex. We first evaluated the topographical similarity of our model with that in the primate V1 by evaluating unit responses to sine grating images of varying orientation, spatial frequency, and color, similar to reference (Margalit et al., 2023). We observed: A) smoothly changing selectivity when considering each of the three factors; B) the similarity decayed exponentially with distances ; C) difference in feature selectivity as a function of distance in an early layer; D) distribution of orientation difference within the laterally connected area. The proportion of orientation difference ±45° is 60% which aligns with the experimental observation from (Muir et al., 2011); E) a tendency towards orthogonal angles between spatial frequency and orientation gradients similar to prior experimental work (Nauhaus et al., 2012)

We also visualize the V1 topography from locally connected LLCNNs which does not have covariance across spatial locations (Fig. 3). The first layer with lateral connectivity size equal to the whole map still exhibit smooth and patchy topography in V1. We can observe the emergence of linear sections, singularities, and pinwheels, similar to those found in the visual cortex (Bosking et al., 1997).

## 3.2. IT Topography

Neurons assume increasingly selectivity to increasingly more complex visual patterns the later in the ventral visual pathway they are with neurons in the human an non-human primate's inferotemporal cortex exhibiting selectiv-

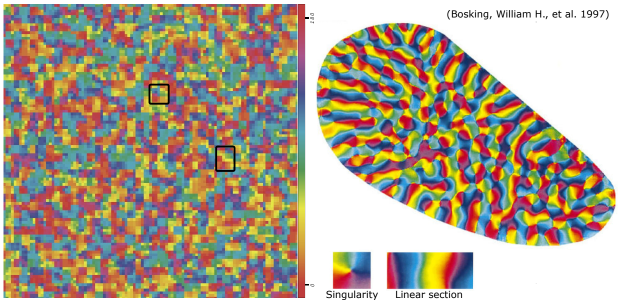

*Figure 3.* **V1 Topography in locally-connected model** The emergence of linear sections, singularities and pinwheels (Bosking et al., 1997) in the early layer of LLCNN with locally connected convolution

ity to specific object categories such as faces, body parts, and scenes. In the high level visual cortex, the principle of cortical topography manifests as distinct cortical patches of category-selective neurons, a phenomenon observed in diverse species, including macaque monkeys and humans. Moreover, areas involved in processing specific object classes are shown to arrange into partly parallel pathways (Bao et al., 2020).

We first investigated whether units in deeper layers of our neural network similarly cluster into category-selective bands with similar category-selectivity as in the human brain. For this, we quantified selectivity of each model unit to each of a number of object categories previously reported in recent literature (Bao et al., 2020; Allen et al., 2022) and visualized them on the simulated cortical sheet of the model (Fig. 4A), concerning six distinct categories of images, namely face, scene, body, characters, objects, no-man's land (Bao et al., 2020) as well as to animacy and size (Konkle & Caramazza, 2013).

We found distinct unit clusters selective to each of the probed categories across the deeper layers of the network. Most category-selective clusters were significantly stretched along the posterior-anterior axis of the simulated cortical sheet and were arranged into semi-continuous pathways expanding across multiple layers of the network (Fig. 4A). The degree of elongation along the posterior-anterior axis changed as a function of the decay rate of the lateral connectivity window size, with faster decay rates leading to less elongation (Fig. 4D). Likewise, the size of category-selective patches co-varied wit the size of the lateral connectivity window (Fig. 4E).

Face- and scene-selective pathways were notably non-overlapping and positioned on the opposite sides of the axis orthogonal to posterior-anterior axis of the simulated cortical sheet (Fig. 4A). Likewise, the model displayed two parallel streams that encoded animacy of objects and their size (Fig. 5), similar to prior observations from human visual cortex (Konkle & Caramazza, 2013).

All model variations, regardless of their lateral connectivity function displayed continuously changing selectivity maps that were significantly smoother than the non-topographical model (Fig. 4C). Similar to our findings in the early model layers, pairwise unit response correlations in these layers also followed an exponentially decaying trend as a function of simulated physical distances between the units (Fig. 4B).

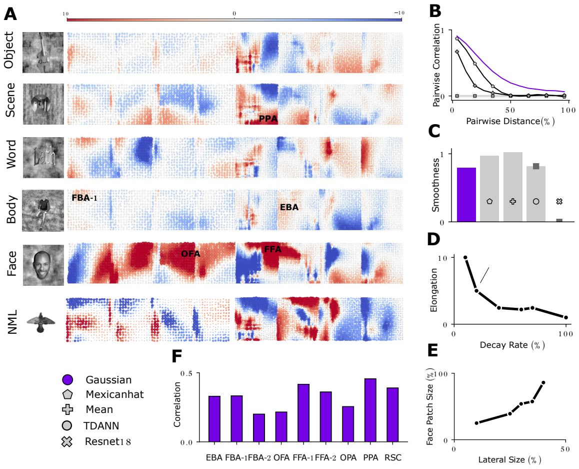

*Figure 4.* **IT Topography** We next investigated the similarity of topographical organization in deeper layers of the network and IT cortex by quantifying unit selectivity using t-value measure (Margalit et al., 2023) Unit responses were assessed concerning six distinct categories of images, namely face, scene, body, characters, objects, no-man's land (Bao et al., 2020). We observed that A,D) continuous and smooth patches selective each of the six categories emerged in the deeper layers of the model (blocks 3-4) that were extended along the shallow-deep axis of the model, similar to typical elongation of category selective patches along the posterior-anterior axis of ventral visual cortex; B) Pairwise unit correlations decayed exponentially as a function of distance; D) the patch elongation was decreased as a function of how fast the lateral connection range was decayed; E) patch sizes were modulated by the range of lateral connectivity.

### 3.3. Using topographical models to reveal selectivity at uncharted cortical landscape

We expected our model to not only reproduce the previously known category selective patches in the brain but also to predict the selectivity in cortical regions with unspecified selectivity. For this, we used the LLCNN model to predict the activity in the parts of the human IT cortex that falls in between two well-known category selective regions, FFA and PPA, using the NSD dataset. We calculated the Pearson

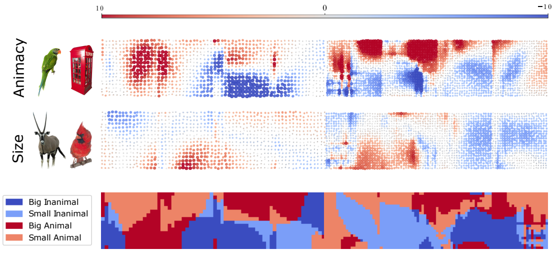

*Figure 5.* **Parallel streams of animacy and size patch** The model displayed two parallel streams that encoded animacy and size of objects, similar to prior observations from human visual cortex (Konkle & Caramazza, 2013)

correlation maps between LLCNNs units (Gaussian) and each of the three patches (FFA1, PPA, and the patch in between them). Interestingly, we found that the intermediate patch was most highly correlated with units in the model that fall in between model's FFA1 and PPA patches (Fig. 6A). Furthermore, we quantified the similarity of the intermediate model patch with that in the brain by measuring the patch correlation between the responses from models and their human counterparts (Fig. 6B) which showed strong similarity in all three model variations and specially in the Gaussian model.

### 3.4. Behavioral performance and Robustness

While the model displayed a significant drop ( 20%) in its object recognition performance compared to its non-topographical counterpart, its accuracy was still substantially higher than the state-of-the-art topographical model (Margalit et al., 2023) (TDANN=43.9%, Gaussian=53%, ResNet18=69.57%; Fig. 7C).

In addition, we also found that LLCNN gives rise to more robust representation. The neural network with lateral connections displayed strong resilience to pixel perturbations (Fig. 7A) compared to the non-topographical model that also increased with larger lateral connection range (Fig. 7B; AutoAttack $\epsilon_{L2} = 1$).

Moreover, the wiring length of LLCNNs was also significantly minimized compared to ResNet18 and baseline topographic models, suggesting that optimization of the locally laterally connected models on the object recognition objective leads to minimization of wiring cost as a byproduct (Fig. 7D), exceeding other topographical models such as TDANN (Margalit et al., 2023). We speculate that this may arise from the continuous padding in LLCNN, which encourages the formation of a continuous category-selective pathways across different layers, thereby significantly reducing the wiring length across layers compared to TDANN.

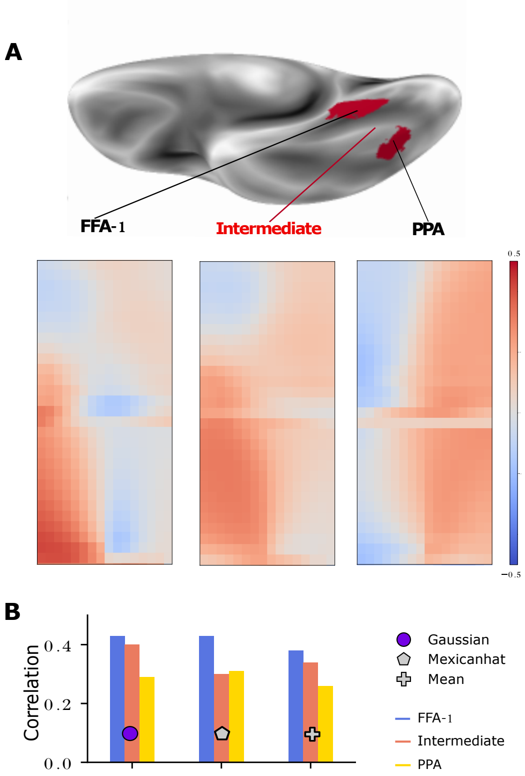

*Figure 6.* **Predicting activity in the intermediate parts of the human IT cortex.** A) identifying an intermediate cortical patch between FFA1 and PPA from NSD dataset. B) correlation maps between LLC units (Gaussian) and each patch.

## 4. Discussion

Local lateral connectivity is ubiquitously found across the cortex and specific patterns of lateral connectivity has previously been proposed as a reliable indicator of cortical topography in adult animals of different species. We incorporated local lateral connections into deep convolutional neural networks and showed that DNNs with these lateral connections can closely simulate the topographical organization of neurons across the hierarchy of human visual cortex.

**Topographical organization without topography-inducing learning rules and objectives.** In the past several years, a number of models have been proposed that aim to incorporate cortex-like topographical organization into deep neural network models. These models make use of auxiliary learning objectives or learning rules that encourage higher covariance across model units in closer proximity (Margalit et al., 2023; Blauch et al., 2022; Lee et al., 2020; Finzi et al., 2023; Lu et al., 2023). One of the oldest and most widely used topographical models is the Kohonen's self-organizing feature maps which proposes a positionally-aware competitive learning rule to encourage developing topographically organized units within a layer of ANNs. SOFMs were shown to successfully replicate the organization of feature selectivity in the early visual cortex and more recently in the high-level visual cortex (Durbin & Mitchison, 1990; Doshi & Konkle, 2023). Motivated by the wiring cost minimization theory of cortical topography Blauch et al.(Blauch et al., 2022) developed a model of cortical topography that operated by tuning the DNN parameters to minimze the total wiring cost in the network. Other recent work considered a more direct approach and defined a similarity based objective functions that encouraged the unit responses or connectivity patterns to follow a brain-like topographical organization (Margalit et al., 2023; Yamins et al., 2013; Lee et al., 2020; Finzi et al., 2023; Lu et al., 2023).

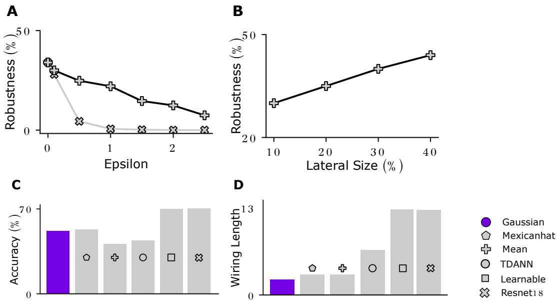

*Figure 7.* **Behavioral performance and Robustness** We compared the performance of the LLCNN with that of baseline models. A,B) The neural networks with lateral connections showed stronger resilience to AutoAttacks with L2 perturbations compared to the non-topographical model, and larger lateral size gives rise to larger robustness. C)LLCNN exhibits a better tradeoff between accuracy and topography compared to other baselines. D) The significant reduce in wiring length of LLCNN.

Our work demonstrates that by including local lateral connections into common DNNs such as convolutional networks, cortex-like topography could emerge without adopting any proprietary learning objectives or learning rules. Compared to previous approaches, LLCNNs are more biologically plausible in that they solely rely on a simple architectural motif that has been widely observed in the animal brains and do not introduce any additional assumptions about the neural circuits or the learning process.

**Top-down influence on shaping the topographical organization.** Much of the literature on studying how topographical organization is shaped in the cortex revolves around the bottom-up and lateral connections. Likewise, most models of cortical topography that have been used to simulate the organization of neurons according to their selectivity have also made use of local learning rules or objectives (Von der Malsburg, 1973; Kohonen, 1982). Yet, our present work as well as other recent modeling work have highlighted the potential role of top-down credit assignment mechanisms on shaping the cortical topographical organization. While the biological plausibility of the backpropagation algorithm used in this work is still under debate, several biologically-aligned variations of this algorithm have been proposed in recent years (Alexandre Payeur et al., 2022).

**Towards globally topographic neural network models.** Earlier models of cortical topography including some of the recent deep neural network implementations had limited scope in their topographical organization. For example, models such as Malsberg's and Swindale's were only designed to replicate the properties of the primary visual cortex (Von der Malsburg, 1973; Willshaw & Von Der Malsburg, 1976) while several recent topographic DNN models (Blauch et al., 2022; Lee et al., 2020) simulated topography in only one layer of the network. More recent work (Margalit et al., 2023; Finzi et al., 2023) has addressed this issue by applying the organization rules in all or most layers of the network leading to network architectures in which units within each layer are topographically organized.

Yet, it is known that in animals with topographically organized cortices, such organization is not limited to each region independent of others and at least some aspects of cortical topography continuously change across the boundaries of classical cortical regions such as V1 and V2. For example the eccentricity and spatial angle respects a continuum across these areas (Arcaro et al., 2009). Building on the ubiquity of local lateral connectivity in the cortex, the present work takes a step further towards developing models with globally shaped topographical organization by positioning units across all layers of the neural network within a global coordinate system where the same connectivity structure is consistently applied across all of them.

**Topographical organization beyond category-selective patches.** There are many computational models of cortical topography available, with some of them dating back to 1970s (Von der Malsburg, 1973; Willshaw & Von Der Malsburg, 1976). Many of these models consisted of single differential equations that worked with abstract inputs that coded for stimulus features such as orientation or patterns of bright/dark dots overlaid on small arrays (Swindale, 1996). Consequently, these models were only validated on their ability to reproduce observed aspects of the cortical topography in early visual areas like V1.

More modern variations of these models, implemented in deep neural networks, were also proposed in the last few years. These models have extended the potential of these models to explain the organization of neurons in higher stages of visual processing (Margalit et al., 2023; Doshi & Konkle, 2023). However, these models are primarily tested on whether they contain previously discovered category-selective patches and are only occasionally used to generate new predictions about the brain.

In this work, we explored a new direction for using these models for generating predictions. Namely, we used LLCNN models to discover selectivity at cortical landscape at intermediate points between known category-selective patches. We believe that existing and upcoming topographical models constitute invaluable tools for making new discoveries about the organization of the cortex in large parts of the visual cortex that are under explored.

**Topography as a consequence of circuit motif that improves robustness**. The question of why cortex is topographically organized has peaked the minds of neuroscientist for decades. Cortical columns and their highly regular arrangements on the cortical surface arguably do not serve any computational roles (Horton & Adams, 2005). Yet, the functional modularity as a function of topographical organization has been proposed to lead to bias towards short connections which is a valuable possible evolutionary goal (Jacobs & Jordan, 1992). Indeed, several computational modeling studies have shown that minimization of cross-unit connections in neural networks leads to the emergence of cortex-like topographical organization, suggesting that the orderly arrangement of the cortex results from evolutionary pressures aimed to optimize the limited space within the skull, constructing a brain that minimizes the volume of nerves needed to connect neurons across different regions (Jacobs & Jordan, 1992; Blauch et al., 2022).

Our results provide an alternative and perhaps complementary view to this theory. They show that cortex-like topography could emerge as a byproduct of top-down learning in neural networks with local lateral connectivity and without any explicit incentive for such arrangement. In turn, this topographical organization yields substantial reduction in the overall connection volume. Moreover, neural networks with lateral connections lead to formation of neural ensembles that are not only clustered together but also that are computationally interconnected. Consequently, robustness to input perturbations in LLCNNs are substantially improved. This view of neural computation based on neural ensembles is closely related to classic approaches in machine learning such as bagging, boosting, and random forests showing that ensembles of weak processing units construct stronger and more robust information processing units.

**Topography as modular learning**. Moreover, the functional organizational framework is analogous to the modular learning in deep learning (Pfeiffer et al., 2023). This connection offers a potential solution by organizing parameters into distinct, independently accountable regions. The profound link between causal representations and modularity is noteworthy. Structural causal models commonly assume the breakdown of knowledge about causal influences into independent mechanisms (Parascandolo et al., 2018) — a concept that highlights the strong connection between cause-and-effect models and modularity.

**Topographical organization improves interpretability**. Additionally, because of their modular structure, LLCNNs have the capability to alleviate the notorious challenges associated with polysemanticity, thereby enhancing the interpretability of deep learning. Decomposing topographic neural networks into smaller modules (Olah et al., 2020) offers a promising avenue for improving interpretability. By understanding the function of each component, and how they interact, we could reason about the behavior of the entire network. Identifying and analyzing the correct components is the initial step in this process, and the feature-selective patches in topographic models facilitate the localization of target components. Furthermore, many neurons exhibit polysemanticity, responding to mixtures of seemingly unrelated inputs. This polysemanticity complicates the understanding of network behavior in terms of individual neuron activity. Topographic neural models present a potential solution by encouraging a significantly higher number of selective units. This feature makes topographic models a promising approach for identifying an interpretable set of features.

**The link between topographical organization and fault tolerant computation**.

A natural comparison can be made between kernel average pooling in neural networks and the majority voting mechanisms in von Neumann's fault-tolerant automata (Von Neumann, 1956). Both methods enhance robustness and fault tolerance through redundancy and aggregation.

In neural networks, particularly those employing kernel average pooling, the local lateral connectivity (LLC) encourages the formation of multiple redundant neurons with similar functions. This redundancy mirrors the self-repair mechanisms found in biological systems (Zlokapa et al., 2022). Just as majority voting in von Neumann's automata ensures reliable computation despite faulty components by relying on the consensus of multiple redundant elements, kernel average pooling mitigates the impact of noisy or corrupted inputs through averaging.

This approach suggests that neural networks with LLC and kernel average pooling can achieve higher resilience and maintain performance under ablation tests better than non-topographical counterparts. The inherent redundancy and fault tolerance in these models allow them to carry out reliable computations even when individual components are compromised.

## 5. Methods

**Architecture** We used the ResNet18 architecture (He et al., 2016) for all model variations in our work. In contrast to the original ResNet18, we arranged the model units (i.e. convolutional kernels) on a 2D plane that simulated the 2D surface of the cortex (i.e. *simulated cortical sheet*). The arrangement was systematic both within and across layers of the network, grounding each unit in a physical space that allowed defining physical distance between each pair of units both within a layer and across (Fig. 1), thereby breaking the symmetry not only among units within kernels but also among kernels themselves. This was inspired by the continuous physical proximity of neurons within cortical areas that are hierarchically close (e.g. V1 and V2).

We considered the computations performed by the local lateral connections to be primarily captured by the following equation:

$$y_{t+1} = W_y \circledast y_t + \sigma(W_x * x_t + b) \tag{1}$$

where $x_t$ and $y_t$ denote the input to and output from a given network layer, $W_x$ the kernel weights of the convolution operation $*$ applied along the spatial dimensions of the input, and $W_y$ the local lateral connection kernel of the Kernel Pooling (KP) operation $\circledast$ (Bashivan et al., 2022) applied along the kernel dimensions of $y_t$. This approach mirrors the principles of spatial pooling operation but *applied along the kernel dimension(s) of the layer activations*. For simplicity, in our simulations, we consider a single-step variation Eq. 1 where the output is computed only once and not iteratively for multiple steps. Therefore, the convolutional layer with this simplified kernel pooling essentially computes the following equation:

$$y = W_y \circledast \big(\sigma(W * x + b)\big) \tag{2}$$

To allow embedding of units from different layers within the same simulated cortical sheet, despite the differences in the number of kernels within each layer, we replaced the traditional zero padding with Continuous Padding which involved appending the size-matched activation from the preceding layer to the current layer's activation before applying the KP operation. Several variations of the KP models were trained including: 1) Kernel Average Pooling (Mean): Computes the average of unit activations within the region of the feature map covered by the filter; 2,3) Kernel Gaussian Pooling (Gaussian), Kernel Mexican-hat Pooling (Mexican-hat): Computes the weighted average of the unit activation within the filter regions based on a Gaussian and Mexican-hat weighting function respectively; 4) Learnable average pooling where the KP parameters $W_y$ are considered as learnable parameters and are optimized along with other network parameters on minimizing the objective function

More details on methodology can be found in the Appendix.

## References

Alexandre Payeur, Jordan Guerguiev, Friedemann Zenke, Blake Richards, and Richard Naud. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature Neuroscience*, 24, 2022. URL https://www.nature.com/articles/s41593-021-00857-x.

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Arcaro, M. J., McMains, S. A., Singer, B. D., and Kastner, S. Retinotopic organization of human ventral visual cortex. *Journal of neuroscience*, 29(34):10638–10652, 2009.

Bao, P., She, L., McGill, M., and Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

Bashivan, P., Ibrahim, A., Dehghani, A., and Ren, Y. Learning robust kernel ensembles with kernel average pooling. *arXiv*, 2022.

Blasdel, G. G. and Salama, G. Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. *Nature*, 321(6070):579–585. ISSN 1476-4687. doi: 10.1038/321579a0. URL https://www.nature.com/articles/321579a0. Number: 6070 Publisher: Nature Publishing Group.

Blauch, N. M., Behrmann, M., and Plaut, D. C. A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, 119 (3):e2112566119, 2022.

Bosking, W. H., Zhang, Y., Schofield, B., and Fitzpatrick, D. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of neuroscience*, 17(6):2112–2127, 1997.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

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

Doshi, F. R. and Konkle, T. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25):eade8187, 2023. doi: 10.1126/sciadv.ade8187. URL https://www.science.org/doi/full/10.1126/sciadv.ade8187. Publisher: American Association for the Advancement of Science.

Durbin, R. and Mitchison, G. A dimension reduction framework for understanding cortical maps. *Nature*, 343(6259): 644–647, 1990.

Finzi, D., Margalit, E., Kay, K., Yamins, D. L. K., and Grill-Spector, K. A single computational objective drives specialization of streams in visual cortex. 2023. doi: 10.1101/2023.12.19.572460. URL http://biorxiv.org/lookup/doi/10.1101/2023.12.19.572460.

Gu, Y., Lewallen, S., Kinkhabwala, A. A., Domnisoru, C., Yoon, K., Gauthier, J. L., Fiete, I. R., and Tank, D. W. A map-like micro-organization of grid cells in the medial entorhinal cortex. *Cell*, 175(3):736–750, 2018.

Harvey, B. M., Klein, B. P., Petridou, N., and Dumoulin, S. O. Topographic representation of numerosity in the human parietal cortex. *Science*, 341(6150):1123–1126, 2013.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Horton, J. C. and Adams, D. L. The cortical column: a structure without a function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456): 837–862, 2005.

Hubel, D. H. and Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

Hubel, D. H., Wiesel, T. N., LeVay, S., Barlow, H. B., and Gaze, R. M. Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 278 (961):377–409, 1977.

Humphries, C., Liebenthal, E., and Binder, J. R. Tonotopic organization of human auditory cortex. *Neuroimage*, 50 (3):1202–1211, 2010.

Jacobs, R. A. and Jordan, M. I. Computational consequences of a bias toward short connections. *Journal of cognitive neuroscience*, 4(4):323–336, 1992.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kohonen, T. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

Konkle, T. and Caramazza, A. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L., and DiCarlo, J. J. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *BioRxiv*, pp. 2020–07, 2020.

Lu, Z., Doerig, A., Bosch, V., Krahmer, B., Kaiser, D., Cichy, R. M., and Kietzmann, T. C. End-to-end topographic networks as models of cortical map formation and human visual behaviour: moving beyond convolutions, 2023. URL http://arxiv.org/abs/2308.09431.

Malach, R., Amir, Y., Harel, M., and Grinvald, A. Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. *Proceedings of the National Academy of Sciences*, 90(22): 10469–10473, 1993.

Margalit, E., Lee, H., Finzi, D., DiCarlo, J. J., Grill-Spector, K., and Yamins, D. L. A unifying principle for the functional organization of visual cortex. *bioRxiv*, 2023.

Muir, D. R. and Douglas, R. J. From neural arbors to daisies. *Cerebral Cortex*, 21(5):1118–1133, 2011. ISSN 1047-3211. doi: 10.1093/cercor/bhq184. URL https://doi.org/10.1093/cercor/bhq184.

Muir, D. R., Da Costa, N. M., Girardin, C. C., Naaman, S., Omer, D. B., Ruesch, E., Grinvald, A., and Douglas, R. J. Embedding of cortical representations by the superficial patch system. *Cerebral Cortex*, 21(10):2244–2260, 2011.

Nauhaus, I., Nielsen, K. J., Disney, A. A., and Callaway, E. M. Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. *Nature neuroscience*, 15(12):1683–1690, 2012.

Obenhaus, H. A., Zong, W., Jacobsen, R. I., Rose, T., Donato, F., Chen, L., Cheng, H., Bonhoeffer, T., Moser, M.-B., and Moser, E. I. Functional network topography of the medial entorhinal cortex. *Proceedings of the National Academy of Sciences*, 119(7):e2121655119, 2022.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Parascandolo, G., Kilbertus, N., Rojas-Carulla, M., and Schölkopf, B. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pp. 4036–4044. PMLR, 2018.

Penfield, W. and Rasmussen, T. The cerebral cortex of man; a clinical study of localization of function. 1950.

Pfeiffer, J., Ruder, S., Vulić, I., and Ponti, E. M. Modular deep learning. *arXiv preprint arXiv:2302.11529*, 2023.

Ringach, D. L., Shapley, R. M., and Hawken, M. J. Orientation selectivity in macaque v1: diversity and laminar dependence. *Journal of neuroscience*, 22(13):5639–5651, 2002.

Schmidt, K. E., Goebel, R., Löwel, S., and Singer, W. The perceptual grouping criterion of colinearity is reflected by anisotropies of connections in the primary visual cortex. *European Journal of Neuroscience*, 9(5):1083–1089, 1997.

Smith, G. B., Hein, B., Whitney, D. E., Fitzpatrick, D., and Kaschube, M. Distributed network interactions and their emergence in developing neocortex. *Nature Neuroscience*, 21(11):1600–1608, 2018. ISSN 1546-1726. doi: 10.1038/s41593-018-0247-5. URL https://www.nature.com/articles/s41593-018-0247-5. Number: 11 Publisher: Nature Publishing Group.

Stigliani, A., Weiner, K. S., and Grill-Spector, K. Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, 35(36):12412–12424, 2015.

Swindale, N. The development of topography in the visual cortex: a review of models. *Network: Computation in neural systems*, 7(2):161–247, 1996.

Von der Malsburg, C. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100, 1973.

Von Neumann, J. Probabilistic logics and the synthesis of reliable organisms from unreliable components. *Automata studies*, 34(34):43–98, 1956.

Walke, E. Y., Sinz, F. H., Froudarakis, E., Fahey, P. G., Muhammad, T., Ecker, A. S., Cobos, E., Reimer, J., Pitkow, X., and Tolias, A. S. Inception in visual cortex: in vivo-silico loops reveal most exciting images. *bioRxiv*, pp. 506956, 2018.

Willshaw, D. J. and Von Der Malsburg, C. How patterned neural connections can be set up by self-organization. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 194(1117):431–445, 1976.

Wong, Y., Kwan, H., MacKay, W., and Murphy, J. Spatial organization of precentral cortex in awake primates. i. somatosensory inputs. *Journal of Neurophysiology*, 41 (5):1107–1119, 1978.

Yamins, D. L., Hong, H., Cadieu, C., and DiCarlo, J. J. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque it and human ventral stream. *Advances in neural information processing systems*, 26, 2013.

Zlokapa, A., Tan, A. K., Martyn, J. M., Fiete, I. R., Tegmark, M., and Chuang, I. L. Biological error correction codes generate fault-tolerant neural networks. *arXiv preprint arXiv:2202.12887*, 2022.

# A. Appendix

**LLC model with locally connected layer** We additionally trained a variation of LLC model where we replaced the first convolutional layer of the network is replaced with a locally-connected layer. Unlike convolution layers, which share weights across the spatial dimensions, locally connected layers employ independent sets of weights at each spatial location, enhancing the expressiveness of the preference map, particularly within V1. However, it's noteworthy that this modification significantly intensifies GPU memory usage. Therefore, in our implementation, we opt to replace only one layer with a locally-connected layer in each model to manage memory constraints. Following prior observations reporting the arrangement of neural selectivity according to eccentricity and polar angle (Arcaro et al., 2009), we arranged the layer weights corresponding to different patches on the simulated cortical sheet following a similar pattern .

**Training** Each neural network model was trained on the Imagenet dataset (Deng et al., 2009) for 100 epochs. We used the Adam optimizer (Kingma & Ba, 2014) for computing the parameter updates from gradients and a scheduler with an initial learning rate of 0.1. We considered training models with fixed-size (fixed sized LC) and exponentially decaying (decayed LC) lateral connection size. In the fixed sized-LC mode, the size of the lateral connection kernel $W_y$ was predetermined and held constant as 0.1 and 0.23 during training while in the decayed-LC mode, we began training using a maximally-sized lateral connection kernel $W_y$ which was equal to the size of the layer map and exponentially decayed the size throughout training. Empirically, we observed that the decayed-LC training leads to the emergence of topographical organization at increasingly finer scales.

**Objective function** We trained the neural network models using supervised or unsupervised learning objectives. The preferred supervised networks were solely trained to minimize the object classification cross-entropy loss on the ImageNet dataset. For the unsupervised training, we considered a non-contrastive approach which has been shown to produce rich visual representations across various architectures (Chen et al., 2020).

## A.1. V1

**Preference map** For analyses presented in Fig. 4, we quantified the orientation, spatial frequency and chromaticity of selectivity of units in layers of each neural network model. We used a stimulus set consisting of grating stimuli of various orientations, spatial frequencies and colors (black/white vs. red/cyan). For each unit, we identified the orientation, spatial frequency, and color that elicited the highest response in that unit and considered that as the preferred stimulus property within each category. To visualize the preferred stimulus property maps, we assigned a color to each unit in the map that indicated the preferred stimulus, which maximizes unit activation strength. The unit size encodes the stimulus selectivity. To quantitatively measure orientation selectivity, we employed the circular variance as in Ringach et al. (Ringach et al., 2002), where lower values signify sharper and more selective tuning, approaching 0.

**Pairwise preference difference** For the analysis presented in Fig. 2 and 4, we quantified the change in unit preferences as a function of pairwise distance between units. For this, we first quantified the preferred stimulus value for each unit (A.1) and then computed all pairwise unit preference differences within all units in each layer. We divide the difference values by the chance value obtained by random resampling of unit pairs. In this fashion, $\Delta$ preference equals to 1 indicating the value expected by random. We then plot the average absolute difference in pairwise stimulus preferences as a function of the pairwise distances on the simulated cortical sheet (Margalit et al., 2023).

**Pairwise correlation** Prior work (Margalit et al., 2023) quantified the rate of pairwise neural correlations as a function of cortical distance between recording sites which shows an exponentially decaying pattern across different brain areas. To replicate this in the model, we quantified the pairwise correlation between units within each unit in each model layer across all stimuli in the NSD dataset. We then compute the average pairwise Pearson correlation between model units within each layer and plot those values against the pairwise distances on the simulated cortical sheet. In contrast to the pairwise preference difference analysis which solely compares the unit activity at their maximum activation for a specific stimulus type, this analysis comprehensively compares entire patterns of unit activity in response to all types of stimuli.

**Distribution of orientation differences** It is known that the orientation selectivity of 60-75% of laterally connected neurons falls within 45° of the selectivity of the source neuron (Muir et al., 2011). To investigate this in our model, we calculated the distribution of delta orientation selectivity between each unit and its neighboring units within each neural network layer.

**Orthogonality** As previously demonstrated (Nauhaus et al., 2012), orientation and spatial frequency gradient directions on the surface of cortex were highly skewed towards orthogonality. To quantify this in our model, we first mapped the contours of orientation and spatial frequency preference maps within each model layer. We then superimposed the contour maps of the two attributes and identified the cross sections and quantified the angles between them using opencv. Finally, we visualized the distribution of angles at the intersections of these maps.

**V4 Best stimulus set** We determined the inputs from MEI (Walke et al., 2018), contour and shape datasets which elicited the maximum units activations, and visualized them in the simulated cortical sheets with units' physical positions.

### A.2. IT

**Selectivity map** We evaluated the model units in terms of their selectivity to faces, body parts, scenes, word forms, no-man's land, and other objects. For this we used the functional localizer stimulus set (fLoc) (Stigliani et al., 2015) to quantify similarity to all categories except no man's land which was quantified on the stimulus set from (Bao et al., 2020). In addition to selectivity to these categories we also we quantified the selectivity to animate and inanimate objects as well as object size using (Konkle & Caramazza, 2013).

Selectivity was measured by computing the t-value measure adopted from (Margalit et al., 2023), where we measure the difference of responses to the target category in contrast to other categories (e.g. face vs. other objects) and normalize the difference by amount variance in each distribution. In this way, the higher the t value is between two distribution of responses the larger and more significant is the difference between the two distributions. Accordingly, higher t values could be interpreted as stronger selectivity of units for a particular class of objects.

$$t = \frac{\mu_{on} - \mu_{off}}{\sqrt{\frac{\sigma_{on}^2}{N_{on}} + \frac{\sigma_{off}^2}{N_{off}}}} \tag{3}$$

where $\mu_{on}$ and $\mu_{off}$ are averages, and $\sigma_{on}$ and $\sigma_{off}$ are standard deviations of on and off categories respectively, N denotes the number of samples within categories

We then visualize the t-values for each neuron overlaid on the map of simulated cortical sheet where, each unit's color and size both denote the selectivity of that unit towards the target category.

**Smoothness** To compute the smoothness of topography, we define the smoothness score by comparing the maximum and minimum correlation.

**Elongation** Most category-selective patches in human cortex are elongated along the posterior-anterior axis of the temporal cortex. We qualitatively observed that the when training LLC networks with progressively decaying lateral connectivity, the corresponding category-selective patches in the model also exhibit elongation along the posterior-anterior axis (shallow to deep layers) of the model. To quantify this phenomenon, we evaluated the elongation of patches at different rates of decay during training. We defined the patch elongation ratio as the length of the patch measured along the posterior-anterior axis divided by the length of patch along the lateral direction which is defined as the axis orthogonal to the posterior-anterior axis.

**Patch size** To quantify the relationship between the patch size and local lateral connectivity size, we evaluate the size of category patches in models with different kernel pooling sizes. Using a t-value threshold of 5, we determine the size of candidate patches in the selectivity map by counting the number of activated units within contiguous regions.

**Patch to patch similarity** After identifying the model patch corresponding to each category-selective brain patch, we measured the representational similarity between each pair using Pearson correlation. For this, we computed the correlation between each unit's average activation in the model patch and the average fMRI response in the corresponding visual cortex patch when both are activated by the same stimuli from the Natural Scene Dataset (NSD) (Allen et al., 2022).

**Identifying the category selective patches corresponding to each brain area** We determined the corresponding visual cortex patches of category patch in models by identifying the maximum patch-to-patch similarity. The visual cortex patches that exhibited the highest similarity to the category patch were considered as corresponding patches in the model.

### A.3. Intermediate Patch Analysis

**Unit correlation map** To create the unit correlation maps, we calculated the Pearson correlation between each model unit's activity and the corresponding fMRI responses within the target patch to the full set of stimuli from the NSD dataset. We then visualized the unit correlations using a heatmap plot where the degree of correlation was displayed by the assigned color to each unit.

**Intermediate patch correspondence** In evaluating topographical models of cortex in terms of their alignment with high level visual cortex, these models are commonly inspected for existence of category-selective unit clusters within their layers. While essential, this type of analysis is restricted to known category-selective patches and does not probe whether the topography of these models matches that of the brain outside of these patches. To expand our assessment, we employ an intermediate patch analysis which aims to demonstrate the model's capability to predict the unknown intermediate patch between two category patches in the visual cortex. To perform this analysis we chose two highly consistent category selective brain regions, namely the Fusiform Face Area (FFA) and Parahippocampal Place Area (PPA) in each subject's brain. We then found the center voxel of each patch in that subject and calculate the straight line that connects the two points across the surface of the cortex. We then identify the middle point on that line and define an area of 10 by 15 voxels around that point as the intermediate patch. We follow a similar procedure in the model to identify the model's intermediate patch and compute the correlation between the two corresponding intermediate patches in the model and the brain in each model.

As a secondary analysis, we also quantify the topographical match between the model and the brain by calculating the degree of curvature a model exhibits when one moves along the trajectory on the model's simulated cortical sheet that best corresponds to the straight line on the cortex. For this, we compute a sequence of 5 intermediate patch centroid fMRI values in the brain and correlate them with the model unit activity. Consequently, we identify a series of corresponding intermediate patches within the models. Our assumption is that if the topography is similar, the transition of these intermediate patches in models should follow a linear trajectory, with a curvature of the shortest path connecting these intermediate patches being 0. To quantify the accuracy of this prediction, we assess the curvature of the shortest path from the face to the place patch within our models.

### A.4. Behavioral performance and Robustness

**Object recognition performance.** We utilized the ImageNet test dataset (Deng et al., 2009), comprising over 5000 labeled images across 1000 classes, for evaluating our model. Images were preprocessed with standard transformations. Evaluation metrics included Top-1 and Top-5 accuracies.

**Wiring length** To quantify wiring length, we followed the approach from Margalit et al. (Margalit et al., 2023) where units with the highest responses in each layer are first identified and subsequently, the length of inter-layer connectivity necessary to link these identified units are calculated according to their physical distances on the simulated cortical sheet. Specifically, for a given stimulus, we pinpoint the top 5% most responsive units in each of the four adjacent sub-layers. The inter-layer connectivity is established using the k-means clustering algorithm, with connections continuously added until the total ”inertia” of the k-means clustering falls below a specified threshold. The total wiring length is then determined as the sum of the lengths of each individual inter-layer connectivity and the intra-layer connectivity which connects the centroids across layers. It is important to note a distinction from a previous study (Margalit et al., 2023), where the average wiring length across all shift directions was reported. In our model, the anterior and posterior directions are explicitly defined, eliminating the need for direction shifting in our analysis.

**Robustness** We evaluated the model's robustness to natural and synthetic input perturbations. For adversarial robustness, we evaluated the model's performance under the AutoAttack (Croce & Hein, 2020) with L2 perturbation epsilon from 0 to 2.5 which is a complex and reliable ensemble attack consisting of several white-box and black-box attacks. To measure the relationship between robustness and laterally connectivity size, we evaluated the model's performance with varying lateral size under AutoAttack.