Local lateral connectivity is sufficient for replicating cortex-like topographical organization in deep neural networks

Xinyu Qian[1], Amir Ozhan Dehghani[2], Asa Borzabadi Farahani[1], Pouya Bashivan[1,2,3]
[1] Department of Computer Science, McGill University
[2] Department of Physiology, McGill University
[3] Mila, Université de Montréal

August 6, 2024

Abstract

Across the primate cortex, neurons that perform similar functions tend to be spatially grouped together. In the high-level visual cortex, this biological principle manifests itself as a modular organization of neuronal clusters, each tuned to a specific object category. The tendency toward short connections is widely believed to explain the existence of such an organization in the brains of many animals. However, the neural mechanisms underlying this phenomenon remain unclear. Here, we use artificial deep neural networks as test beds to demonstrate that a topographical organization akin to that in the primary, intermediate, and high-level human visual cortex emerges when units in these models are locally laterally connected and their weight parameters are tuned by top-down credit assignment (i.e. backpropagation of error). The emergence of modular organization without explicit topography-inducing learning rules or objective functions challenges their necessity and suggests that local lateral connectivity alone may suffice for the formation of topographic organization across the cortex. Furthermore, the incorporation of lateral connections in deep convolutional networks enhances their robustness to small input perturbations, indicating an additional role for these connections in learning robust representations.

1 Introduction

Functional organization, characterized by the spatial arrangement of neurons across the cortical sheet according to their functional similarity, stands out as a ubiquitous phenomenon in neuroscience research. This multiscale organization is evident in widely observed topographic maps throughout the brain [HW62, HKPD13, HLB10, OZJ+22, GLK+18, WKMM78] and has facilitated investigations through lesion experiments, stimulation techniques, and advanced neuroimaging methods such as functional MRI, providing deeper insight into brain functionality.

One of the earliest and most significant contributions to mapping the functional organization of the cortex was made by Wilder Penfield. Using electrical stimulation, Penfield produced a global functional map of the human cortex, revealing the intricate ways in which the brain governs perception and action [PR50]. Building upon this foundational work, subsequent studies, notably those by Hubel and Wiesel [HW62], provided a more detailed examination of 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 [KMC97, DJSK01, EK98].

The ubiquitous topographical organization of neurons in primate cortices raises two critical questions: "Why are neurons organized topographically?" and, "What neural mechanisms enable such multiscale topography across the cortex?". A prominent theory addressing the "why" question is the Wiring Cost Minimization (WCM) [JJ92], which posits that topographical organization is an evolutionary adaptation aimed at minimizing the volume of nerve connections across and within cortical areas.

To elucidate the "how" of cortical topography, various computational models have been developed

based on the WCM theory. These models implement different strategies to induce position-dependent covariation between unit responses. Some models use position-aware update rules [Koh82, VdM73, WVDM76], while others adopt learning objective functions that enforce cortex-like position-dependent pairwise response/weight correlations [LMJ+20, MLF+24, LDB+], or directly penalize weight connections between distant units [BBP22]. Each approach has successfully mimicked aspects of cortical topography, such as the formation of category-selective clusters observed in the primate inferotemporal cortex, including areas selective for faces and places. However, the specific formation of these clusters is not exclusively predicted by any single approach and can be achieved through various computational methods.

Moreover, these prior models face certain limitations. First, several of these models have limited scope regarding their correspondence with the brain, often simulating only early or high-level visual cortex [VdM73, Lin86, BBP22, LMJ+20], or a narrow variety of object-selective categories [BBP22]. Second, several models that rely on self-organizing principles factorize the problem of representation learning and topography and assume sequential or independent mechanisms governing the two which is in contrast to biology[DM90, DK22]. Third, all topographical models exhibit substantially diminished performance on ecologically relevant tasks such as object categorization [BBP22, DM90, DK22], questioning the utility of topographical organization. Finally, most prior models rely on strong assumptions about the biological network that are often unmet or remain unspecified.

Substantial evidence suggests that certain aspects of the topographical organization begin to form prior to eye opening. 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 [HH96, LSS78] (see [BS] for a review). These maps continue to develop postnatally, converging to their mature state [SHW+18]. Importantly, sensory deprivation or restrictions could lead to strong degradation in the resulting cortical map [HW97]. Moreover, there is a large body of literature that documenting the details of the synaptic connectivity across neurons in primary visual cortex of many animal species [RL82, RL83, CK90, MDCG+11]. These observations suggests that the inherent structure of the cortical neural network may play a key role in shaping its topographical organization.

Three key insights from these studies 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 decays with increasing cortical distance, although there exist patches of distant neurons with strong lateral connections; 3) Patchy lateral connectivity patterns (i.e. daisy patterns) have been reported in almost all animal species with strong topographical cortices and are virtually nonexistent in animals with weak organization, such as rodents. These insights can in turn be interpreted in two ways: 1) Neurons with similar selectivity profiles actively build lateral connections among themselves; 2) Inherent lateral connectivity patterns enforce structure on the spatial distribution of neuronal selectivity profiles on cortex. The first scenario suggests that neurons are capable of identifying and synapsing with others that share functional similarities, while the second implies that neuronal selectivity across the network emerges due to learning with pre-specified constraints in connectivity, which is potentially encoded genetically.

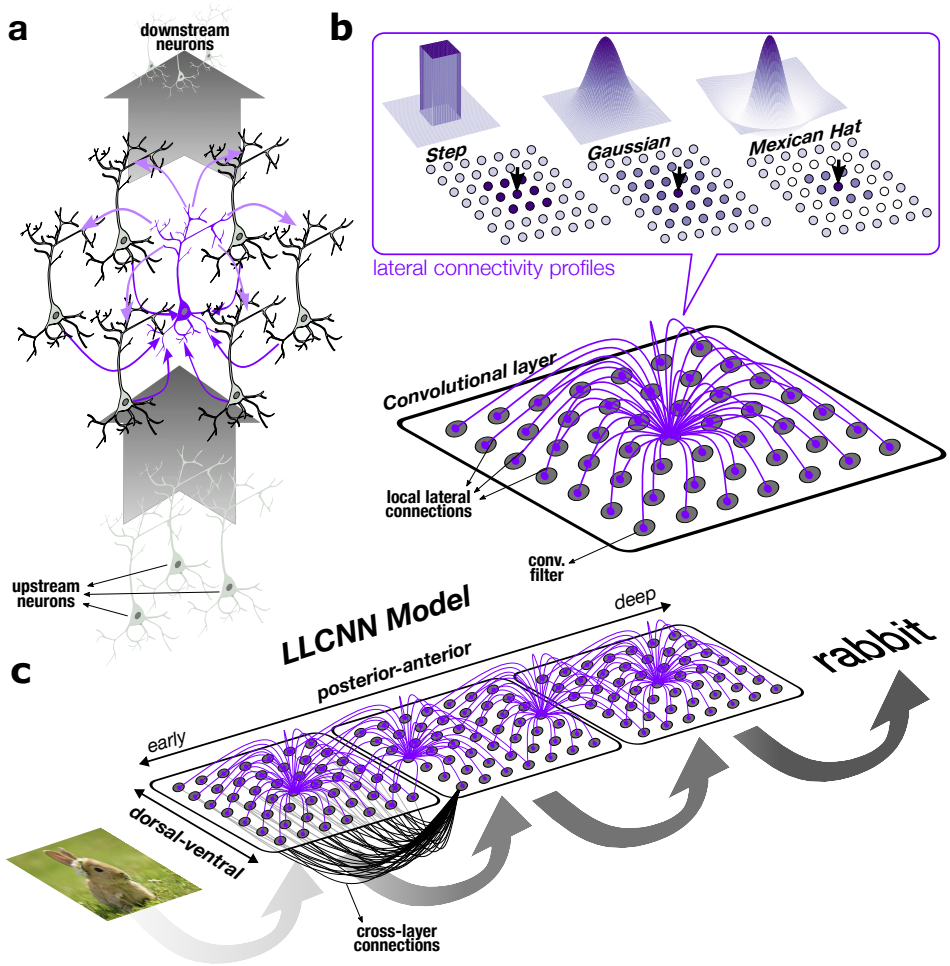

Figure 1: **Locally laterally connected neural networks (LLCNNs).** a) A schematic of the main circuitary with two types of connections, namely, **local lateral connections** (in purple) and **bottom-up connections** from neurons in an upstream area (in black). b) Convolutional filters are arranged on a two-dimensional "simulated cortical sheet" that defines their spatial configuration. Three variations of distance-dependent connectivity patterns used. Each unit's output consists of its bottom-up computation (i.e. convolution) and local lateral connectivity whose parameters are fixed and pre-specified as one of three patterns: Step, Gaussian, and Mexican hat. c) Model layers are arranged on a 2D layout with two distinct axes akin to the posterior-anterior and dorsal-ventral axes in the human ventral visual cortex, breaking the permutation symmetry of neurons in neural networks. Inputs are fed from one end and outputs are generated on the other end.

In this work, we present Locally-Laterally Connected Neural Networks (LLCNNs), a topographical neural network model that integrates lateral connections into deep convolutional neural networks (Fig. 1). This LLCNN model architecture embodies the hypothesis that local lateral connections alone are sufficient to establish the topographic organization of neurons according to feature selectivity. LLCNN consists of a series of convolutional layers with local lateral connections within each layer. These connections represent the reciprocal lateral connections between pyramidal cells in layers 2 and 3, also known as the superficial patch system [LAB03]. Units within each layer are arranged on a 2-dimensional sheet simulating a simplified view of the arrangement of neurons in the cortex (Fig. 1b). Moreover, all layers were also arranged on a 2D layout with two distinct axes akin to the posterior-anterior and dorsal-ventral axes in the human ventral visual cortex. We explored three types of local lateral connection profiles 1) Step-like function, assigning uniform strength to every connection within a fixed range (LLCNN-S); 2) Gaussian, where connection strengths decay exponentially with distance between units (LLCNN-G); and 3) Mexican-hat, which where connection strengths follows the Mexican-hat pattern with distance between units (LLCNN-MH, Fig. 1b). While local lateral

connections could in principle be added to any neural network architecture, here we used a commonly used feedforward convolutional neural network model called ResNet18 [HZRS16] that consists of 18 convolutional layers. Besides the addition of local lateral connections connections, all other aspects of LLCNN networks, including training data and procedures, closely matched with conventional methods (see Methods).

The locally-laterally connected neural network demonstrates several significant outcomes: 1) It replicates the arrangement of neurons in the early visual cortex according to the orientation, spatial frequency, and color selectivity (Fig. 2); 2) It forms a multi-scale topographical organization in its deep layers, similar to those found in the human inferotemporal cortex (Fig. 3, 4); 3) It reproduces the organization in mid-level visual area V4 (Fig. 5), and predicts unit selectivity in inferotemporal regions with previously uncharted object-selectivity (Fig. 6); 4) It improves the trade-off between object recognition performance and cortex-like topography compared to prior models (Fig. 7); 5) It shows notable improvement in robustness against adversarial noise, suggesting a previously unknown functional role for cortical lateral connections in learning robust representations (Fig. 7).

These results reveal the multifaceted role of cortical local lateral connections, not only in optimizing intercortical nerve volume but also in enhancing the robustness of neural representations. Crucially, our findings demonstrate that the emergence of complex multi-scale topographical maps can occur without specific topography-inducing learning rules, challenging longstanding theoretical frameworks. This paradigm-shifting insight has profound implications for both neuroscience research and the development of advanced neural network algorithms, marking a significant step forward in our quest to decode and replicate the intricate processes of the human brain.

2 Results

2.1 Topographical organization of units in early model layers mimics that of V1

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 [HW62, BS]. 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 extent the LLCNN model could replicate the same organization in its early layers (i.e. layer 1 and 2 of ResNet architecture) that corresponded with 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) [MLF+24]. 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 and for all three variations of LLCNN model (Figs. 2a, **??**. 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 the 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 [NNDC12]. Furthermore, the smooth maps extended across the layers of the LLCNN models, 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 [BZSF97], cat [SGLS97], and macaque monkeys [MAHG93] have orientation selectivity that falls within $\pm 45^\circ$ 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

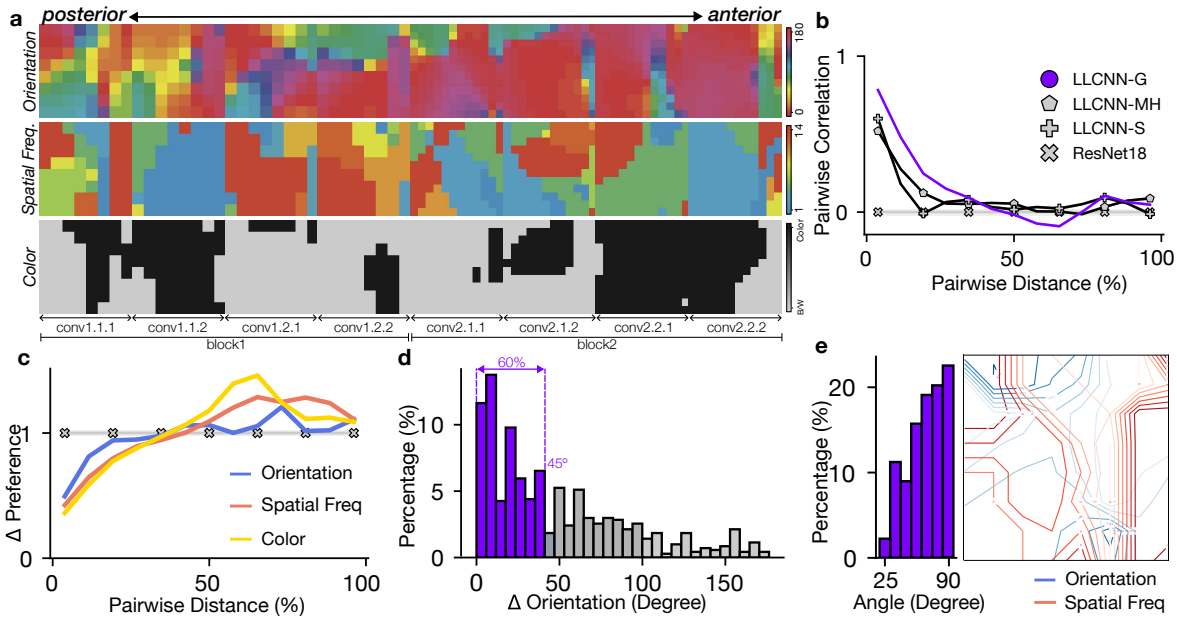

Figure 2: **Early LLCNN-G model layers replicate hallmarks of visual processing in the primate primary visual cortex.** We first evaluated the topographical similarity of our model with that in the primate V1 by evaluating unit responses to sine grating images of varying orientation, spatial frequency, and color, similar to reference [MLF+24]. a) Selectivity maps to orientation, spatial frequency and color smoothly change across the spatial dimensions of the simulated cortical sheet, both within and across layers. b) Pairwise correlation between units decayed exponentially as a function of unit distance on the simulated cortical sheet. c) Difference in feature selectivity increases and plateaus as a function of distance in an early model layers. d) Distribution of difference in orientation selectivity across immediate neighboring units (i.e. strongly laterally connected units). The proportion of orientation difference $\pm 45°$ is $\sim$60% which aligns with the experimental observation from [MDCG+11]; e) A tendency towards orthogonal angles between spatial frequency and orientation gradients similar to prior experimental work [NNDC12]. Orientation (blue) and spatial frequency (red) contours are depicted on the simulated cortical sheet of the first model layer.

found that 60% of neighboring units have selectivity within 45° of the center unit (Fig. 2d).

While there are substantial differences in selectivity maps between individuals, these changes still follow 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 [NNDC12]. 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.

2.2 Unit arrangement in deep LLCNN layers replicates multi-scale topography in high-level human visual cortex

Neurons demonstrate selectivity for increasingly complex visual patterns as one progresses along the ventral visual pathway. In the high level visual cortex, the principle of cortical topography manifests as distinct cortical patches of category-selective neurons, a phenomenon observed across various species, including macaque monkeys and humans. Notable among these are the face-selective fusiform face area (FFA), the body-selective extrastriate body area (EBA) and the scene-selective parahippocampal place area (PPA). Moreover, regions involved in processing specific object classes are organized into partially parallel bands, predominantly running along the posterior-anterior axis of the temporal cortex

[BSMT20].

Based on prior literature, we expected the later layers of our model to better correspond to higher visual areas. Consistent with our findings in the early model layers and congruent with prior experimental work [LMJ+20], pairwise unit response correlations in these layers also exhibited an exponentially decaying trend as a function of simulated physical distances between units (Fig. 3b). All model variants, regardless of their lateral connectivity function, displayed continuously changing selectivity maps that were significantly smoother than those of the nontopographical model (Fig. 3c).

We then examined whether units in deeper layers of our neural network similarly clustered according to their category-selectivity as in the human brain [KMC97, EK98, DJSK01]. We quantified selectivity of each model unit to six distinct image categories–face, scene, body, characters, objects, no-man's land [ASYW+22, BSMT20]–and visualized them on the model's simulated cortical sheet (Fig. 3a). We identified distinct unit clusters for each category in the deeper layers of the network (layer 3 and 4 of ResNet18 architecture). Pairwise response correlations decayed exponentially as a function of distances between units in all model variations (Fig. 3b) and the unit selectivity changed smoothly along the spatial dimensions of the simulated cortical sheet (Fig. 3c), comparable to previous model of cortical topography TDANN [MLF+24]).

In addition, most category-selective clusters were significantly elongated along the posterior-anterior axis of the simulated cortical sheet, forming semi-continuous bands that expanded over multiple layers of the network (Fig. 3a). This feature closely mirrors the relative elongation of similar patches in the human and monkey visual cortex [NLD+11, BSMT20]. Additionally, the extent of this elongation along the posterior-anterior axis varied with the decay rate of the lateral connectivity window size during training, with faster decay rates leading to less elongation (Fig. 3d). Training the LLCNN model with a fixed lateral connection window size resulted in blob-like category-selective patches with milder elongation effect along the anterior-posterior axis (Fig. S7). Similarly, the size of category-selective patches co-varied with the size of the lateral connectivity window (Fig. 3f). Similar to the human cortex, face- and scene-selective were strongly anti-correlated and units with selectivity to each category formed notably non-overlapping pathways positioned on the opposite sides of the dorsal-medial axis of the simulated cortical sheet (Fig. 3a) [**?**]. The highest model patch similarity to OFA appeared in layer 3, preceding the FFA, which emerged in layer 4 (similar patterns were observed for FBA-1 and EBA regions). For other object categories, while multiple selective patches appeared across the model layers, they did not form clear bands like those for face and scene processing. Interestingly, apart from face and scenes, selectivity to NML category was among the highest and most widespread among the six categories. Finally, the activity in LLCNN-G category-selective patches were highly correlated with those measured from different category selective human cortex, improving upon TDANN model in several cortical areas including OFA, FFA1, and OPA (Fig. 3e). We observed that training the LLCNN models using an unsupervised learning objective [ZJM+21] resulted in qualitatively similar results (Fig. S6). However, we didn't perform quantitative comparisons between supervised and unsupervised trained models.

Having investigated the topographical organization of the model at mesoscale, we next explored whether it also replicated the macroscale organization observed in the human visual cortex. We analyzed unit arrangements according to their selectivity for object animacy and real-world size, extensively investigated in human studies [KO12, KC13]. Consistent with human observations, model units clustered according to preferences for animate versus inanimate objects and for small versus large real-world objects. Within each map (animacy and size), units with similar selectivity formed parallel streams along the posterior-anterior axis of the model (Fig. 4a). This finding differed significantly from those previously reported in self-organized topographical neural network models [DK] but closely resembled the extended arrangement of these pathways along the posterior-anterior axis in the human ventral visual cortex [KC13].

Moreover, it has been shown that the similarity between small and big animals is significantly greater than that between small and big objects [KC13]. We investigated whether these differences in selectivity also existed in the LLCNN model units. For this, we calculated each unit preference for big or small objects using the T-statistic and contrasted the distribution of selectivity for size as a function of unit's preference for animate or inanimate objects (Fig. 4b). The result showed that LLCNN-G units that are selective to animals were significantly less sensitive to size, compared to those that are selective for

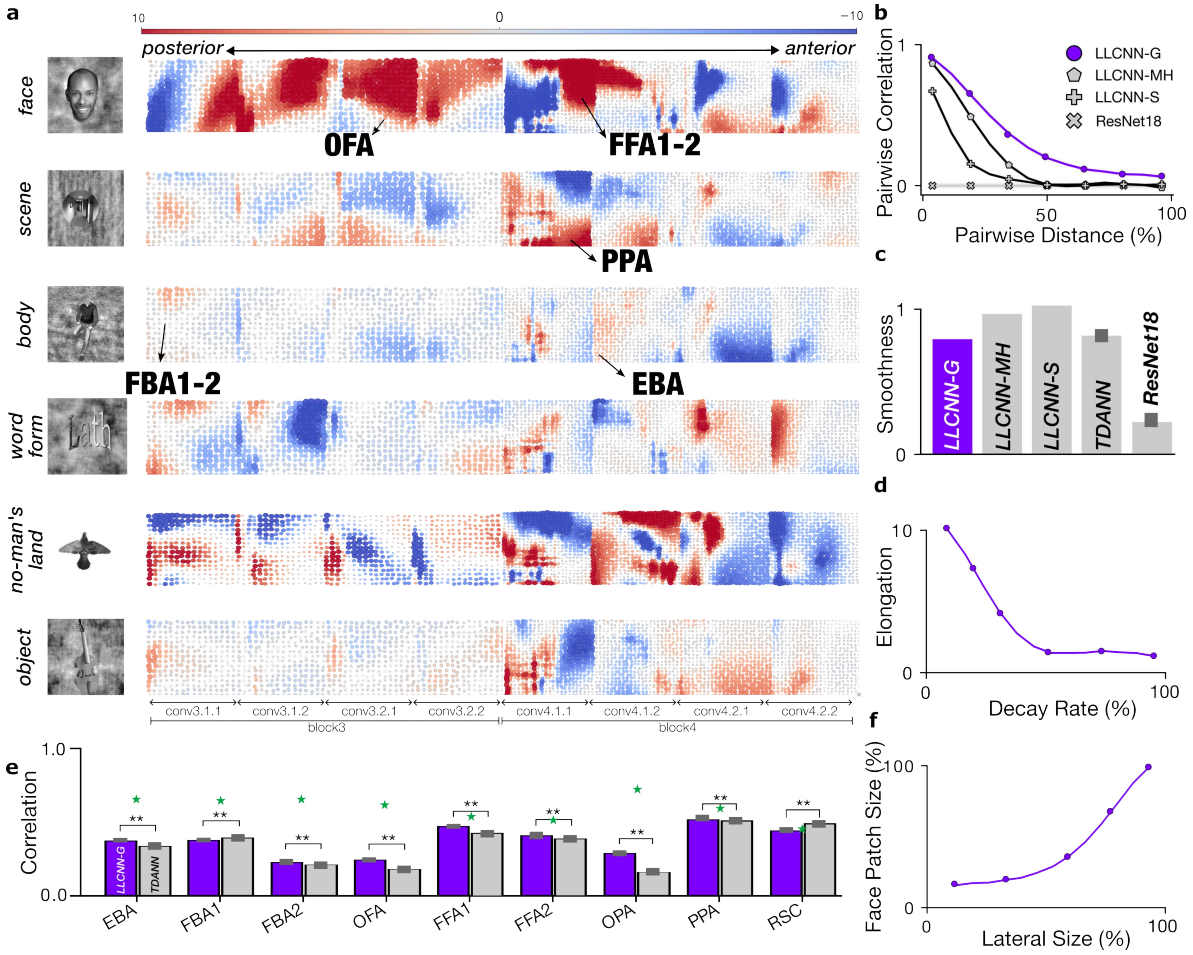

Figure 3: **Topograhical arrangement according to object categories in deep layers of LLCNN-G.** a) Unit responses were assessed concerning their selectivity to six distinct categories of images, namely face, scene, body, characters, objects, no-man's land [BSMT20]. Continuous and smooth patches selective for each of the six categories emerged in the deeper layers of the model (layer 3-4) that were significantly elongated along the posterior-anterior 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; c) LLCNN-G yielded a smoother topography compared to other variants and other topographical neural network model with notably smaller variance; d) The patch elongation decreases as a function of how fast the lateral connection range was decayed during training; e) Activity in category-selective LLCNN-G patches closely resembled those in previously studied category-selective cortical patches. Comparisons were made using the NSD dataset [ASYW+22] and error bars indicate the standard deviation across subjects. Green stars indicate the internal consistency of the brain activity signal in each region; f) Patch sizes were positively modulated by the size of lateral connections.

inanimate objects. In addition, we found that the 70% of units in the scene-selective PPA patch of the LLCNN-G model were selective for big objects, while 83% of units in model's FFA were selective for small objects (Fig. 4c), further replicating prior observations of similar overlaps in the human cortex [KC13].

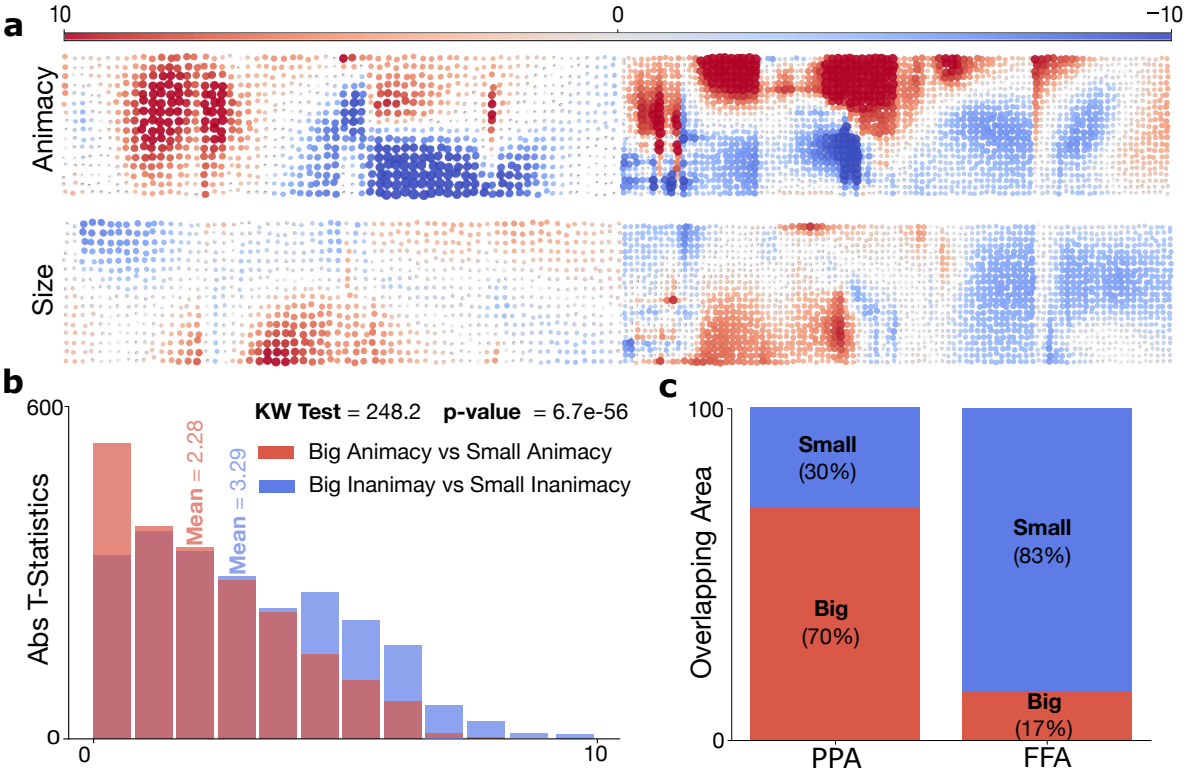

Figure 4: **Macro-scale organization based on animacy and object size emerges across LLCNN-G layers.** a) Unit selectivity to animate vs. inanimate objects, as well as, large vs. small objects emerge along continuous bands. The model displayed two parallel streams that encoded animacy and size of objects, similar to prior observations from human visual cortex. Plot shows the 8 convolutional layers in Blocks3 and 4 of the model. b) The t-statistics for big animacy vs. small animacy and big inanimacy vs. small inanimacy show a statistically significant difference between the means of these two sets of t-values. c) The overlapping area between size and PPA/FFA reveals that the PPA overlaps mostly with patches that are selective to big size (70%), while the FFA mostly overlaps with patches selective to small size(83%). [KC13]

2.3 Topographical organization in uncharted and non-categorical cortical landscape

We further investigated whether the representations in intermediate layers of the LLCNN model also correspond to the intermediate visual areas in the primate brain. Neurons in macaque monkeys area V4 were previously shown to form clusters according to their selectivity to corner or curvature stimuli [JALT21]. We examined the selectivity of units in LLCNN model to corners and curvatures using the same stimulus set used in prior work. We found that the units in block 3 of the LLCNN model were similarly clustered according to their preference for corner versus curvature stimuli (Fig. 5a). Furthermore, similar to our previous observations in early and deep layers of the model, the pairwise unit correlations exponentially decayed as a function of distance between units (Fig. 5b). The selectivity for curve or corner patterns also followed a similar pattern to experimental data from macaques (Fig. 5c; [JALT21]). Visual examination of the best stimulus maps for stimuli from alternative stimulus sets [WRF+23, PC02] further confirmed the existence of robust unit clusters and smooth transition in selectivity across the model units (Fig. S8).

We next asked whether the topographical organization of our model also aligns with that in the human

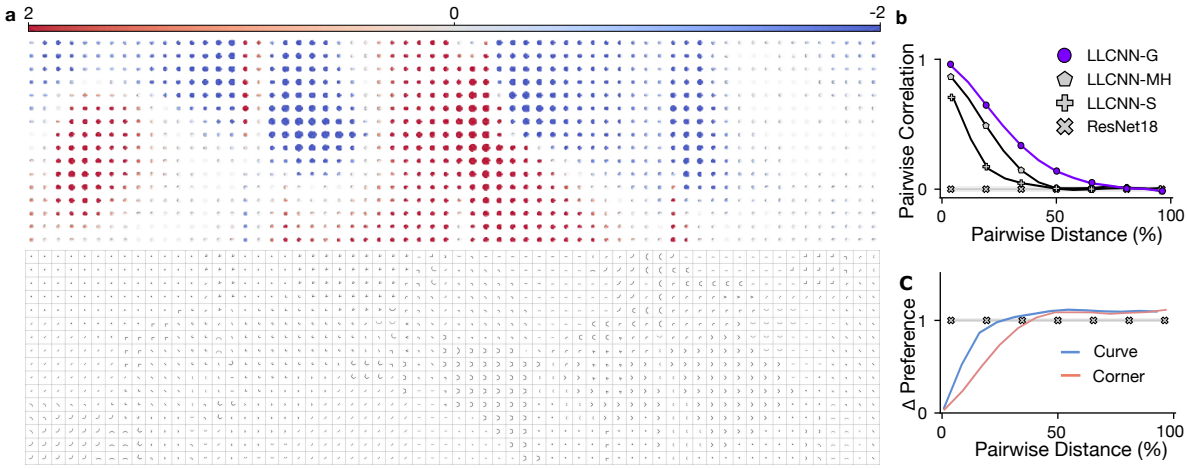

Figure 5: **Topographical organization of units in intermediate layers of the LLCNN-G model according to selectivity to corners and curves.** a) We investigated the topography of area V4 by identifying the stimuli from curves and corners that elicit maximum activations from the units and the t-values. Plot shows the 4 layers in block 3 of the model. b) The pairwise correlation of unit activations responding to shape dataset decayed exponentially as a function of pairwise distance. c) The units' selectivity to corners and curves increased exponentially as a function of pairwise physical distance.

cortex outside of the category-selective regions. To perform this analysis, we first identified the center point of FFA-1 and PPA regions that are most consistent across individuals. We next designated the intermediate point between the FFA-1 and PPA center points in each subject as the *intermediate* patch (Fig. 6a). Our hypothesis was that if the model's organization is a good match to that in the human visual cortex, the activity in this intermediate patch should be well correlated with the activity in units positioned at the intermediate point between the model's respective patches. We calculated the Pearson correlation between all LLCNN-G units and each patch in the brain. The resulting correlation maps showed that indeed the units that were most highly correlated with the intermediate brain patch were located at spatial positions in between the units highest correlated with FFA-1 and PPA (Fig. 6b). All LLCNN model variations could predict the brain activity at the intermediate patch relatively well while the Gaussian variation was significantly better than the other two variations (Fig. 6c). This result suggested that the unit arrangement in these models is not only aligned with the brain in terms of replicating the category-selective patches but also generalizes to uncharted cortical landscape in between.

2.4 Impact of local lateral connections on learned behavior and internal representations

Performance on visual categorization. How does local lateral connectivity affect behavioral performance? Prior work on topographical neural network models had reported a conflict between topography-inducing objective functions and task learning (e.g. visual categorization) [MLF+24]. In this perspective, the parameter updates necessary for learning better representations in practice clash with those necessary for organizing units in a topographical organization, leading the training dynamics to reach an equilibrium that reflects the trade-off between the two objectives. The magnitude of this impact on visual categorization is often large and undesired (∼25.7% reduction on Imagenet accuracy in TDANN [MLF+24]). We evaluated the categorization performance of different LLCNN variants and compared them with nontopograhpical ResNet18 and a recent topographical neural network model [MLF+24]. Although LLCNN models displayed a significant drop (16.5%) in object recognition performance compared to their non-topographical counterpart, their accuracy was still substantially higher than the state-of-the-art topographical model [MLF+24] (TDANN=43.9%, LLCNN-G=53%, RN18=69.57%; Fig. 7c).

Connection lengths. According to the WCM theory, the primary role of the modular organization

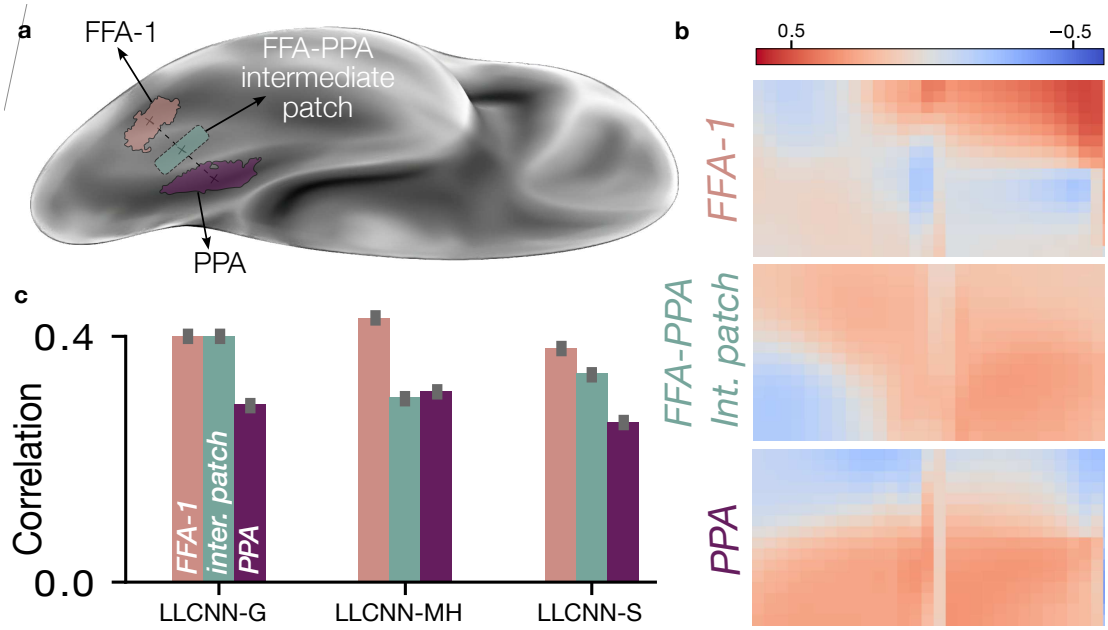

Figure 6: **Predicting activity in the intermediate parts of the human IT cortex.** a) Identifying an intermediate cortical patch between FFA-1 and PPA from NSD dataset (subject 1 is shown). b) Correlation maps between LLCNN-G units and each patch in the brain. c) Correlation between intermediate LLCNN-G unit and the intermediate cortical patch for different model variations.

in cortex is to minimize the volume of interconnecting nerves. We next investigated whether inclusion of local lateral connections results in the reduction of connection lengths between layers and to what degree. We used a previously used metric[MLF+24] to estimate the global connection weights in the network from the connection weights. Using this metric, we found that the wiring length of LLCNNs is significantly lower than the nontopographical ResNet18 model as well as another recent topographic model that was explicitly optimized for this objective [MLF+24]. In fact, the wiring length in LLCNN-G was 16.7% of the non-topographical one and about half of that in TDANN (LLCNN-G=2, TDANN=6, ResNet18=12; Fig. 7d). This suggests that top-down learning in neural networks with local lateral connectivity on ecologically-relevant tasks such as object recognition minimizes the wiring cost as a byproduct. It surprisingly does so even more effectively than alternative topographical models that induce topography by enforcing brain-like distance-dependent correlation structure between model units [MLF+24].

Robust representations. We posited that the learned unit clusters in LLCNNs may be implementing a form of ensemble model where each feature value is computed by finding the average estimate of multiple units with similar selectivity. Machine learning ensemble approaches such as Bagging and Boosting were previously shown to be robust to input perturbations [Fre95, Bre96, Bre01, HTFF09]. We tested the LLCNN models to verify if such robustness to input perturbations is also present in the models by subjecting each model to input images that were adversarially perturbed using a strong adversarial attack called AutoAttack [CH20]. The neural network with lateral connections displayed strong resilience to pixel perturbations (Fig. 7a) compared to the non-topographical model. While the performance of both models eventually reached zero with larger perturbations (i.e. epsilon denoting the magnitude of the perturbation measured with L_2 norm), the LLCNN model with step-like local lateral connectivity showed consistently higher robustness compared to the topographical one. Furthermore, we observed that robustness increases with larger lateral connection range in the model (Fig. 7b; AutoAttack $\epsilon_{\mathrm{L2}} = 1$). This suggested a possible auxiliary role for lateral connections in the cortex in learning robust representations.

Effective dimensionality. Lower effective dimensionality has previously been reported in topographic models of visual cortex and was proposed as an inherent characteristic of these models that leads to better alignment with the neural responses in the brain [MLF+24]. We found that while the effective dimensionality of all topographical models were substantially lower than the original ResNet18 architecture, this difference becomes insignificant for most models, when the categorization performance of these models is matched (Fig. 7e). Without controlling the accuracy of ResNet-18, there is a positive correlation between complexity and accuracy before over fitting [GBC16] that may lead to higher dimensionality in ResNet-18 with maximum accuracy. This result suggests that the significantly lower-dimensional representations in topographical models is a byproduct their lower categorization accuracy rather than an inherent feature.

To gain further insight into the underlying representation in the LLCNN-G model, we examined the unit activity profiles [DK17] in a category-independent manner. For this, we computed the principal components of the transposed activation matrix (in which units and stimuli are arranged along the rows and columns respectively) in response to the NSD stimulus set. In this representation of activations, each sample represents the specific neuron's activity profile to the complete set of stimulus set, making the PCA space a proxy for the function space learned by neurons in LLCNNs.

We found that units in nontopographical ResNet18 span a sphere-like **function manifold** showcasing diverse and uniform response profiles across the units, capturing 49% of the variance within its top 10 PCs. In contrast, the function manifold in the LLCNN-G model is much lower dimensional, showing distinct topological properties (Fig. 8a) in which 97% of the variance in the activations are explained by the top 10 PCs. By coloring the neurons based on their physical positions in LLCNN-G model, we observed that the manifold preserves the pairwise relative distances in the model, indicating that neurons in close spatial proximity tend to produce more similar responses. Similar structure existed in block3 of the model as well (Fig. S9). Similarly, repeating this analysis on brain activity using the NSD dataset, we observed a similar smooth and low dimensional manifold in the human IT cortex (Fig. 8B). However, we could not identify any significant semantic meaning associated with these directions. Most human IT cortex PCs were significantly more aligned with PCs from the LLCNN-G model compared to those from nontopographical ResNet18.

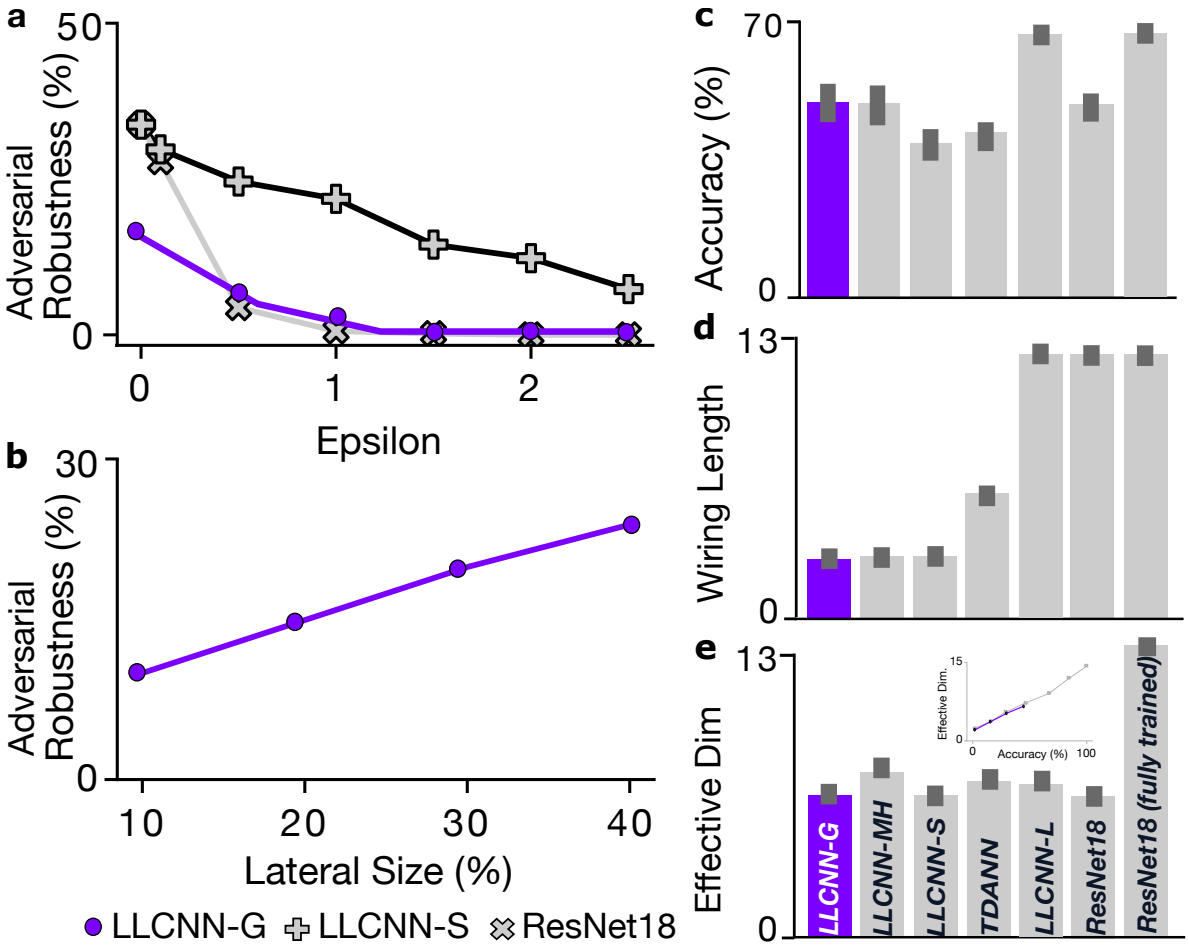

Figure 7: **Behavioral performance and Robustness** We compared the performance of the LLCNN with that of baseline models. a) LLCNNs show stronger resilience to adversarial perturbations (AutoAttack-L2) compared to the non-topographical model ResNet18. Epsilon denotes the strength of the allowed image perturbation measured with L_2 norm; b) Larger lateral size yields larger robustness to adversarial perturbations. c) LLCNN-G and LLCNN-MH exhibit better trade-off between accuracy and topography compared to other baselines. d) Training LLCNNs with gradient descenet leads to minimization of the total wiring length in the network. e) The effective dimensionality does not significantly differ between topographical and non-topographical models with equalized accuracy (33%). Inset shows the relation between effective dimensionality and ImageNet accuracy for intermediate checkpoints of ResNet18 (gray) and LLCNN-G (purple) during training. Effective dimensionality linearly increases with improved accuracy.

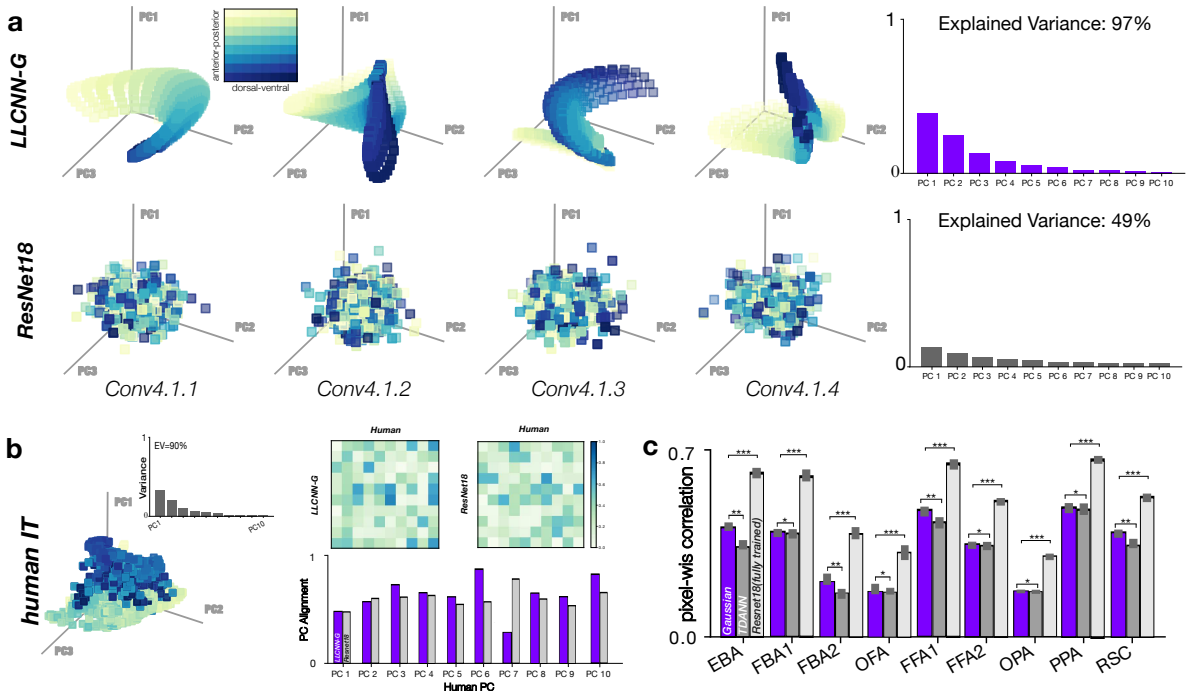

Figure 8: **Low dimensional manifold of activity profiles in LLCNN-G model block4 and human ventral visual cortex.** a) Units' activity profiles are visualized in the top 3 PC space. Each dot corresponds to a unit in the model and units are color coded according to their position on the simulated cortical sheet. LLCNN-G units live on a low-dimensional manifold with maintained physical positions. In contrast ResNet18 units fall on a higher dimensional manifold with no specific position-dependent structure. b) (left) Activity profile manifold similar to **(a)** but computed over IT voxels from NSD dataset also exhibit topographical clustering and a low-dimensional manifold. (right-top) cosine similarity between each pair of model and human brain PCs computed using the NSD dataset. (right-bottom) Most top 10 PCs from the human ventral visual cortex are better aligned with the LLCNN-G model compared to ResNet18; c) The unit-voxel correlation between fMRI and LLCNN-G model activations.

Finally, comparing individual unit activities in each model with voxel activations in the human subjects revealed a higher correspondence (Pearson correlation) between units in the LLCNN model and brain in several category-selectivity areas including EBA and FFA1 compared to TDANN [MLF+24]. However, the correlations were significantly higher for the non-topographical ResNet18 model (Fig. 8c).

3 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. The inclusion of lateral connections also yielded an unexpected behavioral outcome: robustness to input perturbations which is one of the most remarkable characteristics of sensory processing in humans and other animal species.

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 [MLF+24, BBP22, LMJ+20, FMK+, LDB+]. As mentioned in the outset, most recent models rely on strong assumptions about structure and function in biological circuits which challenges

their biological plausibility. For instance, approaches that use topography-inducing objective functions [LMJ+20, MLF+24, BBP22, LDB+] implicitly assume that information about synaptic efficacies and axonal lengths are centrally available at the level of each cortical region for the objective function to be rendered computable. Other approaches that rely on local learning rules have also been shown to lack the capacity to learn rich hierarchical representations that are similar to those in the human brain [DQBB24]. In contrast, in our work, we showed that the topographical organization of the cortex can be closely replicated by incorporating lateral connections into deep neural networks that are trained with top-down credit assignment algorithms. This substantially simplifies the additional assumptions required on neural circuitary necessary for reproducing the cortical topographical organization as such lateral connections has been widely observed in the cortex of many animal species.

Top-down influence on shaping the topographical organization. Much of the classic literature on studying how topographical organization is shaped in the cortex revolves around the bottom-up and lateral connections [VdM73, WVDM76, Lin86]. 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 [Koh82, VdM73]. Yet, our present work as well as other recent modeling work have highlighted the potential role of top-down credit assignment mechanisms (i.e. backpropagation of error) on shaping the cortical topographical organization [DQBB24]. Collectively, these works demonstrate that top-down signal that contain information about each unit's role on a downstream task leads to reducing the widely reported conflict between parameter updates necessary for topographic organization and representation learning. Thereby, these models can often achieve more cortex-like topography while maintaining high task performance, and richer representations.

We used the backpropagation algorithm in our work because of its efficacy and computational efficiency. While the biological plausibility of the backpropagation algorithm is still under debate, several biologically-aligned variations of this algorithm have been proposed in recent years [AJF+22, LSM+20].

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 were even argued to not carry any computational significance [HA05]. 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. Indeed, recent computational modeling studies have shown that minimization of cross-unit connections across layers of neural network models leads to the emergence of cortex-like topographical organization [BBP22], suggesting that the orderly arrangement of the cortex may have resulted 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.

Our results provide an alternative and perhaps complementary view to this theory. They show that learning in locally laterally connected neural networks leads to the minimization of overall cross-layer connection volume, despite the lack of any explicit penalty that encourages shorter inter-neuron connections. More importantly, neural networks with lateral connections leads to formation of neural ensembles that are not only clustered together but also that are computationally interconnected that in turn substantially improve the robustness of such networks to input perturbations. This view of neural computation based on neural ensembles is closely related to classic approaches in machine learning such as bagging [Bre96], boosting [Fre95], and random forests [Bre01] showing that ensembles of weak processing units construct stronger and more robust information processing units.

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 [VdM73, WVDM76] while several recent topographic DNN models [BBP22, LMJ+20] simulated topography in only one layer of the network. More recent work [MLF+24, FMK+] 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 con-

tinuously 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 [AMSK09]. Building on the ubiquity of local lateral connectivity in the cortex, the present work takes a step further towards developing models with global 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 units across layers.

Topographical organization beyond category-selective patches. There are many computational models of cortical topography available, with some of them dating back to 1970s [VdM73, WVDM76]. 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 [Swi96]. 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 [MLF+24, DK]. 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 two avenues for using these models for generating predictions: 1) to discover selectivity at cortical landscape at intermediate points between known category-selective patches; 2) to map out the organization of cortical selectivity at intermediate cortical patches that are less known. We believe that existing and future topographical models constitute invaluable tools for making new discoveries about the organization of the cortex in large parts of the visual cortex that are currently under-explored.

Topography as modular learning. Moreover, the functional organizational framework is analogous to the modular learning within the context of machine learning. This connection offers a potential solution by organizing parameters into distinct, independently accountable regions.

Topographical organization improves interpretability. Decomposing topographic neural networks into smaller modules [OCS+20] 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.

4 Methods

4.1 Laterally-connected topographic neural networks (LLCNNs)

Architecture. We used the ResNet18 architecture [HZRS16] for all model variants 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 permutation 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:

$$\boldsymbol{y}_t^i = \sigma(\boldsymbol{W}_x^i * \boldsymbol{x}_t + b^i) \tag{1}$$

$$\boldsymbol{Y}_{t+1} = \boldsymbol{W}_y \circledast \boldsymbol{Y}_t, \quad \boldsymbol{Y}_t = [\boldsymbol{y}_0, \boldsymbol{y}_1, \dots \boldsymbol{y}_n] \tag{2}$$

where $\boldsymbol{x}_t$ and $\boldsymbol{y}_t^i$ denote the input to a given network layer, and output of the ith kernel from that layer respectively. $\boldsymbol{W}_x$ denotes the kernel weights of the convolution operation $*$ applied along the spatial dimensions of the input, and $\boldsymbol{W}_y$ the local lateral connection kernel of the Kernel Pooling (KP) operation $\circledast$ applied along the kernel dimensions of $\boldsymbol{Y}_t$ [BIDR22]. This approach mirrors the principles of spatial pooling operation but *applied along the channel 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.

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 (CP) 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 (KAP): Computes the average of unit activations within the region of the feature map covered by the filter (LLCNN-S); 2,3) Kernel Gaussian Pooling (KGP), Kernel Mexican-hat Pooling (KMP): Computes the weighted average of the unit activation within the filter regions based on a Gaussian and Mexican-hat weighting function respectively (LLCNN-G and LLCNN-MH); 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 (LLCNN-L).

Training. Each neural network model was trained on the Imagenet dataset [DDS+09] for 100 epochs. We used the Adam optimizer [KB14] 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 $\boldsymbol{W}_y$ 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 $\boldsymbol{W}_y$ 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 contrastive approach, Barlow Twins, which has been shown to produce rich visual representations across various architectures [ZJM+21]. We do not believe that different types of supervision make a significant difference in the eventual topography as our results from supervised and unsupervised learning yield similar topographic organization.

4.2 V1 analyses

Preference map. For analyses presented in Fig. 2, 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) [MLF+24]. 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. [RSH02].

Pairwise preference difference. For the analysis presented in Fig. 2 and 3, we quantified the change in unit preferences as a function of pairwise distance between units following previous work [MLF+24]. For this, we first quantified the preferred stimulus value for each unit (4.2) and then calculated all pairwise unit preference differences between all units in each layer. We divide the difference values by

the chance value obtained by random resampling of unit pairs. In this fashion, Δ preference equals 1 indicating the value expected by chance. We then plot the average absolute difference in pairwise stimulus preferences as a function of the pairwise distances on the simulated cortical sheet.

Pairwise correlation. Prior work 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 [LMJ+20]. To replicate this in the model, we quantified the pairwise correlation between units within 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 [MDCG+11]. 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. We then plotted a histogram of those values and counted the percentage of units with difference in orientation selectivity $< 45°$.

Orthogonality. As previously demonstrated [NNDC12], orientation and spatial frequency gradient directions on the surface of cortex are 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 Python opencv pacakge. Finally, we visualized the distribution of angles at the intersections of these maps.

V4 analyses Best stimulus set. We determined the inputs from MEI [WSF+18], contour and shape datasets[] which elicited the maximum units activations, and visualized them in the simulated cortical sheets with units' physical positions.

4.3 IT analyses

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) [SWGS15] to quantify similarity to all categories except no man's land which was quantified on the stimulus set from [BSMT20] (shared privately by the authors). In addition to selectivity to these categories, we also quantified the selectivity to animate and inanimate objects as well as object real-world size using [KC13].

For each analysis, selectivity was measured by computing the t-statistic [MLF+24], 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-statistic 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 a stronger selectivity of units for a particular class of objects.

$$t = \frac{\mu_c - \mu_o}{\sqrt{\frac{\sigma_c^2}{N_c} + \frac{\sigma_o^2}{N_o}}} \tag{3}$$

where μ, σ, N denote the average, standard deviation, and the number of samples respectively within the target category c and other categories o. We then visualize the t-values for all units 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 the difference between the maximum and minimum correlation values in the pairwise correlation plot.

$$Smoothness = \max_{correlation} - \min_{correlation} \tag{4}$$

Elongation analysis. Most category-selective patches in human cortex are elongated along the posterior-anterior axis of the temporal cortex [NLD+11, BSMT20]. We qualitatively observed that when training LLC networks with progressively decaying lateral connectivity (decayed-LC), 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 analysis. To quantify the relationship between the patch size and local lateral connectivity size, we evaluated the size of category patches in models with different kernel pooling (KP) 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. We then plotted the estimated patch size as a function of KP size.

Model-brain patch similarity analysis. We first identified all category-selective patches in the model To do this, we computed the category-selectivity t-statistic for each model unit in each layer. We classified each unit with a t-value greater than 5 as category-selective and considered any continuous cluster consisting of more than 20 units as a patch. We then computed the cross-validated Pearson correlation between each unit's average activation across all units within the model patch and the average fMRI response across voxels in the corresponding cortical patch when both were activated by the same stimuli from the Natural Scene Dataset (NSD) [ASYW+22].

4.4 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 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 mapped the straight line that connects the two points across the surface of the cortex. We then identified the middle point on that line between the boundaries of the two category-selective regions and defined an area of 10 by 10 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.

4.5 Behavioral performance, robustness, and other measures

Object recognition performance. We evaluated our model's object recognition performance (top-1 accuracy) using the ImageNet validation dataset [DDS+09], which contains 50,000 labeled images distributed across 1,000 classes. The preprocessing routine for these images involved standard transformations: resizing each image to 256 $\times$ 256 pixels and then extracting a center crop of 224 $\times$ 224 pixels.

Quantifying wiring length. To quantify wiring length, we adopted the approach from [MLF+24] 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 [MLF+24], 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 tested the model's robustness to input perturbations by evaluating the model's performance under the AutoAttack [CH20] adversarial attack with L2 perturbation epsilon ranging from 0 to 2.5. AutoAttack is one of the most difficult adversarial attacks which consists of a suite of 4 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. Because of the high computational cost of this attack, we chose to evaluate the performance of each model using 500 randomly selected samples from Imagenet validation set.

Function manifold. We applied Principal Component Analysis (PCA) to the transposed activations matrix from layer 4 responding to NSD dataset and projected the activation of each neuron onto the PCA space. Subsequently, we projected the manifold onto the first three principal components (PC1, PC2, and PC3) to determine the activation values of each neuron in response to the stimulus along these directions.

Acknowledgement

We thank Dr. Nancy Kanwisher for her helpful discussions and feedback on the manuscript. We also thank Drs. Pinglei Bao and Doris Tsao for sharing the stimulus set from *Bao et al. 2020 Nature* and Dr. Andreas Tolias for sharing the stimulus set from *Willeke et al. 2023 BioRxiv*. This research was supported by the Healthy-Brains-Healthy-Lives startup supplement grant, the NSERC Discovery grant RGPIN-2021-03035, and CIHR Project Grant PJT-191957. P.B. was supported by FRQ-S Research Scholars Junior 1 grant 310924, and the William Dawson Scholar award. All analyses were executed using resources provided by the Digital Research Alliance of Canada (Compute Canada) and funding from Canada Foundation for Innovation project number 42730.

References

[AJF+22] 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.

[AMSK09] Michael J Arcaro, Stephanie A McMains, Benjamin D Singer, and Sabine Kastner. Retinotopic organization of human ventral visual cortex. *Journal of neuroscience*, 29(34):10638–10652, 2009.

[ASYW+22] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[BBP22] Nicholas M Blauch, Marlene Behrmann, and David C Plaut. 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.

[BIDR22] Pouya Bashivan, Adam Ibrahim, Amirozhan Dehghani, and Yifei Ren. Learning robust kernel ensembles with kernel average pooling. 2022.

[Bre96] Leo Breiman. Bagging predictors. 24:123–140, 1996.

[Bre01] Leo Breiman. Random forests. 2001. ISBN: 9781441993267.

[BS] Gary G. Blasdel and Guy Salama. Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. 321(6070):579–585. Number: 6070 Publisher: Nature Publishing Group.

[BSMT20] Pinglei Bao, Liang She, Mason McGill, and Doris Y Tsao. A map of object space in primate inferotemporal cortex. *Nature*, 583(7814):103–108, 2020.

[BZSF97] William H Bosking, Ying Zhang, Brett Schofield, and David Fitzpatrick. Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. *Journal of neuroscience*, 17(6):2112–2127, 1997.

[CH20] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[CK90] Edward M Callaway and Lawrence C Katz. Emergence and refinement of clustered horizontal connections in cat striate cortex. *Journal of Neuroscience*, 10(4):1134–1153, 1990.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[DJSK01] Paul E Downing, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

[DK] Fenil R. Doshi and Talia Konkle. Cortical topographic motifs emerge in a self-organized map of object space. 9(25):eade8187. Publisher: American Association for the Advancement of Science.

[DK17] Jörn Diedrichsen and Nikolaus Kriegeskorte. *Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis*, volume 13. 2017. Publication Title: PLoS Computational Biology Issue: 4 ISSN: 15537358.

[DK22] Fenil R Doshi and Talia Konkle. Visual object topographic motifs emerge from self-organization of a unified representational space. *BioRxiv*, pages 2022–09, 2022.

[DM90] Richard Durbin and Graeme Mitchison. A dimension reduction framework for understanding cortical maps. *Nature*, 343(6259):644–647, 1990.

[DQBB24] Amir Ozhan Dehghani, Xinyu Qian, Asa Borzabadi, and Pouya Bashivan. Credit-based self organization for training high-performing topographic neural networks. *arXiv*, 2024.

[EK98] Russell Epstein and Nancy Kanwisher. A cortical representation of the local visual environment. *Nature*, 392(6676):598–601, 1998.

[FMK+] Dawn Finzi, Eshed Margalit, Kendrick Kay, Daniel L. K. Yamins, and Kalanit Grill-Spector. A single computational objective drives specialization of streams in visual cortex.

[Fre95] Yoav Freund. Boosting a weak learning algorithm by majority. 121(2):256–285, 1995.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[GLK+18] Yi Gu, Sam Lewallen, Amina A Kinkhabwala, Cristina Domnisoru, Kijung Yoon, Jeffrey L Gauthier, Ila R Fiete, and David W Tank. A map-like micro-organization of grid cells in the medial entorhinal cortex. *Cell*, 175(3):736–750, 2018.

[HA05] Jonathan C Horton and Daniel L Adams. The cortical column: a structure without a function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):837–862, 2005.

[HH96] J. C. Horton and D. R. Hocking. An adult-like pattern of ocular dominance columns in striate cortex of newborn monkeys prior to visual experience. *Journal of Neuroscience*, 16(5):1791–1807, 1996. Publisher: Society for Neuroscience Section: Articles.

[HKPD13] Ben M Harvey, Barrie P Klein, Natalia Petridou, and Serge O Dumoulin. Topographic representation of numerosity in the human parietal cortex. *Science*, 341(6150):1123–1126, 2013.

[HLB10] Colin Humphries, Einat Liebenthal, and Jeffrey R Binder. Tonotopic organization of human auditory cortex. *Neuroimage*, 50(3):1202–1211, 2010.

[HTFF09] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[HW62] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106, 1962.

[HW97] David Hunter Hubel and Torsten Nils Wiesel. Ferrier lecture - functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 198(1130):1–59, 1997. Publisher: Royal Society.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[JALT21] Rundong Jiang, Ian Max Andolina, Ming Li, and Shiming Tang. Clustered functional domains for curves and corners in cortical area v4. 10:e63798, 2021. Publisher: eLife Sciences Publications, Ltd.

[JJ92] Robert A Jacobs and Michael I Jordan. Computational consequences of a bias toward short connections. *Journal of cognitive neuroscience*, 4(4):323–336, 1992.

[KB14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[KC13] Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

[KMC97] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, 1997.

[KO12] Talia Konkle and Aude Oliva. A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, 74(6):1114–1124, 2012. Number: 6 Publisher: Elsevier.

[Koh82] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[LAB03] Jennifer S. Lund, Alessandra Angelucci, and Paul C. Bressloff. Anatomical substrates for functional columns in macaque monkey primary visual cortex. 13(1):15–24, 2003.

[LDB+] Zejin Lu, Adrien Doerig, Victoria Bosch, Bas Krahmer, Daniel Kaiser, Radoslaw M. Cichy, and Tim C. Kietzmann. End-to-end topographic networks as models of cortical map formation and human visual behaviour: moving beyond convolutions.

[Lin86] R Linsker. From basic network principles to neural architecture: emergence of orientation columns. 83(22):8779–8783, 1986. Publisher: Proceedings of the National Academy of Sciences.

[LMJ+20] Hyodong Lee, Eshed Margalit, Kamila M Jozwik, Michael A Cohen, Nancy Kanwisher, Daniel LK Yamins, and James J DiCarlo. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *BioRxiv*, pages 2020–07, 2020.

[LSM+20] Timothy P. Lillicrap, Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020. Number: 6 Publisher: Springer US.

[LSS78] Simon Levay, Michael P. Stryker, and Carla J. Shatz. Ocular dominance columns and their development in layer IV of the cat's visual cortex: A quantitative study. *Journal of Comparative Neurology*, 179(1):223–244, 1978. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cne.901790113.

[MAHG93] Rafi Malach, Yonah Amir, Michal Harel, and Amiram Grinvald. 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.

[MDCG+11] Dylan Richard Muir, Nuno MA Da Costa, Cyrille C Girardin, Shmuel Naaman, David B Omer, Elisha Ruesch, Amiram Grinvald, and Rodney J Douglas. Embedding of cortical representations by the superficial patch system. *Cerebral Cortex*, 21(10):2244–2260, 2011.

[MLF+24] Eshed Margalit, Hyodong Lee, Dawn Finzi, James J DiCarlo, Kalanit Grill-Spector, and Daniel LK Yamins. A unifying framework for functional organization in early and higher ventral visual cortex. *Neuron*, 2024.

[NLD+11] Shahin Nasr, Ning Liu, Kathryn J. Devaney, Xiaomin Yue, Reza Rajimehr, Leslie G. Ungerleider, and Roger B. H. Tootell. Scene-selective cortical regions in human and nonhuman primates. *The Journal of Neuroscience*, 31(39):13771–13785, 2011.

[NNDC12] Ian Nauhaus, Kristina J Nielsen, Anita A Disney, and Edward M Callaway. Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. *Nature neuroscience*, 15(12):1683–1690, 2012.

[OCS+20] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

[OZJ+22] Horst A Obenhaus, Weijian Zong, R Irene Jacobsen, Tobias Rose, Flavio Donato, Liangyi Chen, Heping Cheng, Tobias Bonhoeffer, May-Britt Moser, and Edvard I Moser. Functional network topography of the medial entorhinal cortex. *Proceedings of the National Academy of Sciences*, 119(7):e2121655119, 2022.

[PC02] Anitha Pasupathy and Charles E. Connor. Population coding of shape in area v4. 5(12):1332–1338, 2002. Number: 12 Publisher: Nature Publishing Group.

[PR50] Wilder Penfield and Theodore Rasmussen. The cerebral cortex of man; a clinical study of localization of function. 1950.

[RL82] Kathleen S Rockland and Jennifer S Lund. Widespread periodic intrinsic connections in the tree shrew visual cortex. *Science*, 215(4539):1532–1534, 1982.

[RL83] Kathleen S Rockland and Jennifer S Lund. Intrinsic laminar lattice connections in primate visual cortex. *Journal of Comparative Neurology*, 216(3):303–318, 1983.

[RSH02] Dario L Ringach, Robert M Shapley, and Michael J Hawken. Orientation selectivity in macaque v1: diversity and laminar dependence. *Journal of neuroscience*, 22(13):5639–5651, 2002.

[SGLS97] Kerstin E Schmidt, Rainer Goebel, Siegrid Löwel, and Wolf Singer. 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.

[SHW+18] Gordon B. Smith, Bettina Hein, David E. Whitney, David Fitzpatrick, and Matthias Kaschube. Distributed network interactions and their emergence in developing neocortex. *Nature Neuroscience*, 21(11):1600–1608, 2018. Number: 11 Publisher: Nature Publishing Group.

[SWGS15] Anthony Stigliani, Kevin S Weiner, and Kalanit Grill-Spector. Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, 35(36):12412–12424, 2015.

[Swi96] NV Swindale. The development of topography in the visual cortex: a review of models. *Network: Computation in neural systems*, 7(2):161–247, 1996.

[VdM73] Chr Von der Malsburg. Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, 14(2):85–100, 1973.

[WKMM78] YC Wong, HC Kwan, WA MacKay, and JT Murphy. Spatial organization of precentral cortex in awake primates. i. somatosensory inputs. *Journal of Neurophysiology*, 41(5):1107–1119, 1978.

[WRF+23] Konstantin F. Willeke, Kelli Restivo, Katrin Franke, Arne F. Nix, Santiago A. Cadena, Tori Shinn, Cate Nealley, Gabrielle Rodriguez, Saumil Patel, Alexander S. Ecker, Fabian H. Sinz, and Andreas S. Tolias. Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization, 2023. Pages: 2023.05.12.540591 Section: New Results.

[WSF+18] Edgar Y Walke, Fabian H Sinz, Emmanouil Froudarakis, Paul G Fahey, Taliah Muhammad, Alexander S Ecker, Erick Cobos, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception in visual cortex: in vivo-silico loops reveal most exciting images. *bioRxiv*, page 506956, 2018.

[WVDM76] David J Willshaw and Christoph Von Der Malsburg. 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.

[ZJM+21] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

A Supplementary Information

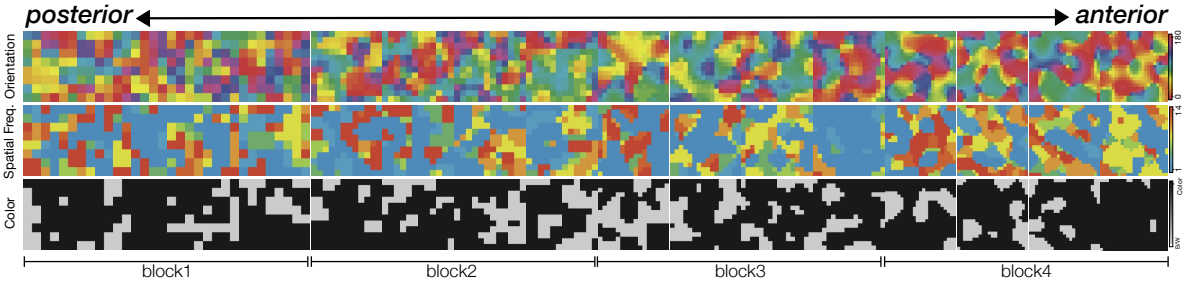

Figure S1: **V1 Topography in LLCNN-S model.** Selectivity maps to orientation, spatial frequency and color smoothly change across the spatial dimensions of the simulated cortical sheet, both within and across layers. Selectivity maps is shown for all units in all layers of the model. Early to late layers of the model are shown from left to right (posterior to anterior).

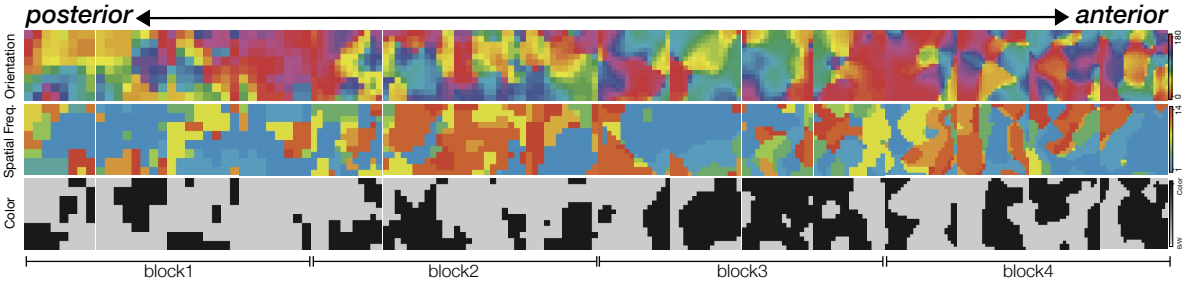

Figure S2: **V1 Topography in LLCNN-MH.** Selectivity maps to orientation, spatial frequency and color smoothly change across the spatial dimensions of the simulated cortical sheet, both within and across layers. Selectivity maps is shown for all units in all layers of the model. Early to late layers of the model are shown from left to right (posterior to anterior).

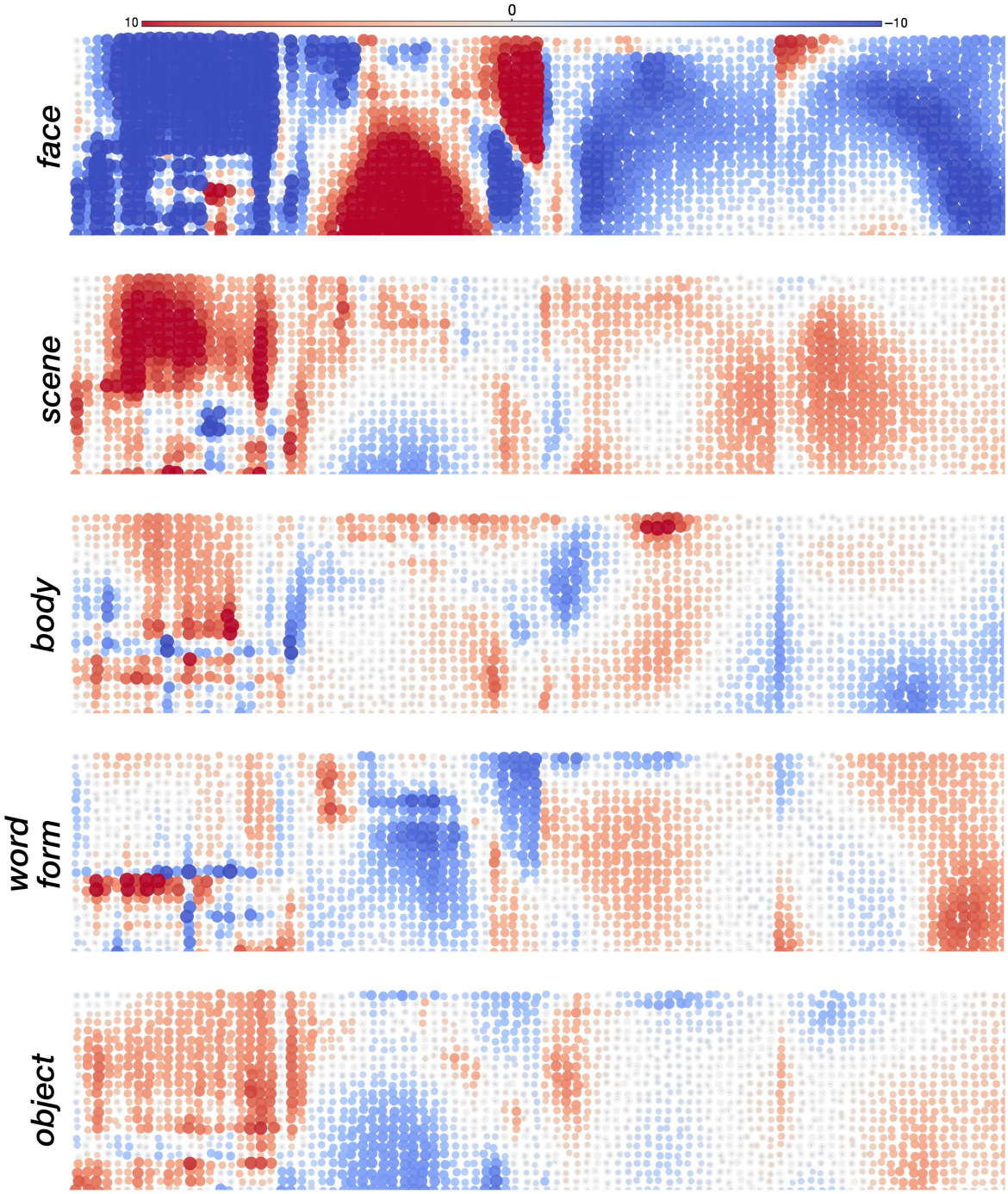

Figure S3: **Topograhical arrangement according to object categories in deep layers of LLCNN-S model.** Unit responses were assessed concerning their selectivity to five distinct categories of images, namely face, scene, body, characters, and objects. Selectivities are shown for the 4 convolutional layers in block 4 of the ResNet18 architecture. Color denotes the selectivity of each unit to each category.

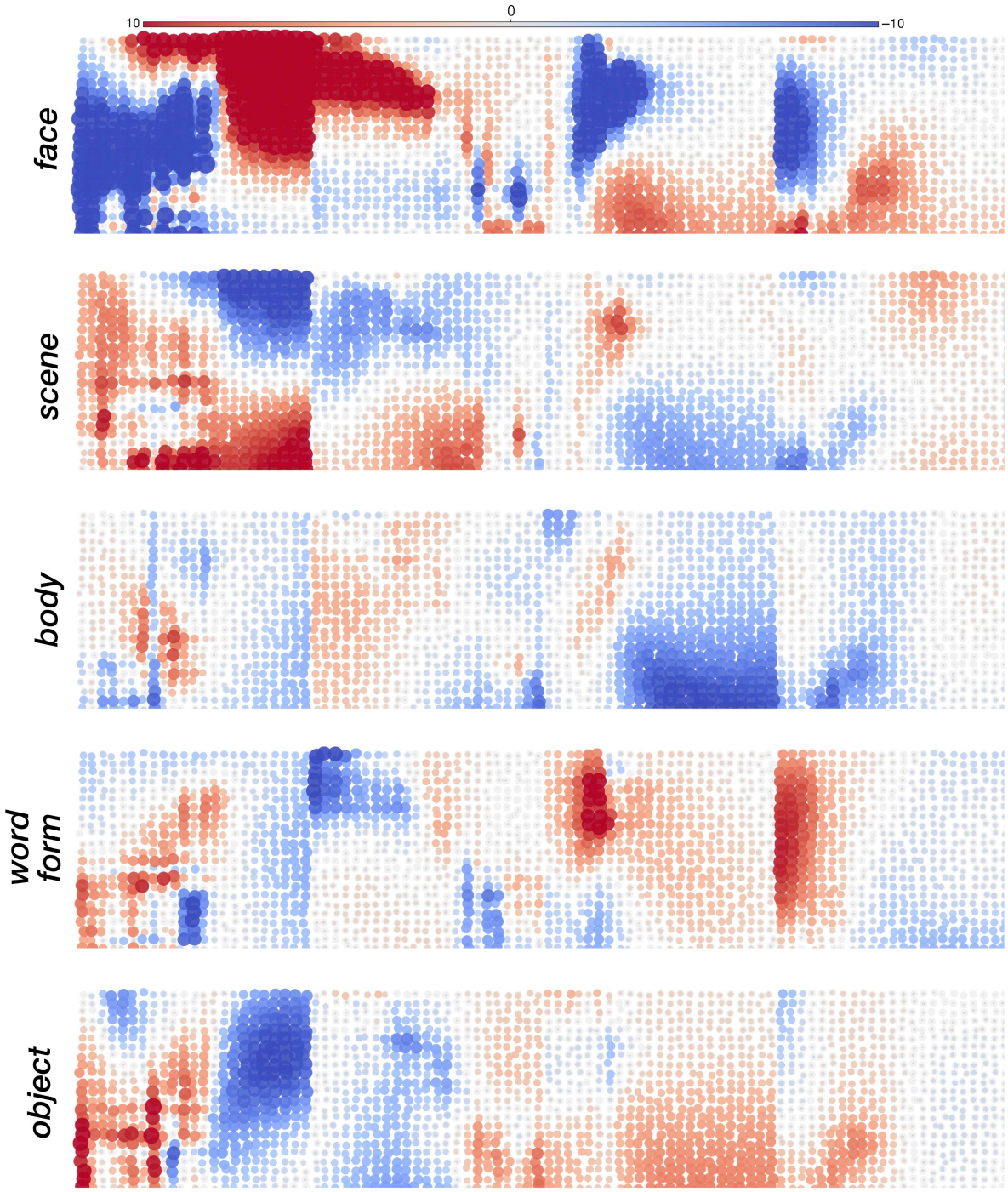

Figure S4: **Topograhical arrangement according to object categories in deep layers of LLCNN-MH model.** Unit responses were assessed concerning their selectivity to five distinct categories of images, namely face, scene, body, characters, and objects. Selectivities are shown for the 4 convolutional layers in block 4 of the ResNet18 architecture. Color denotes the selectivity of each unit to each category.

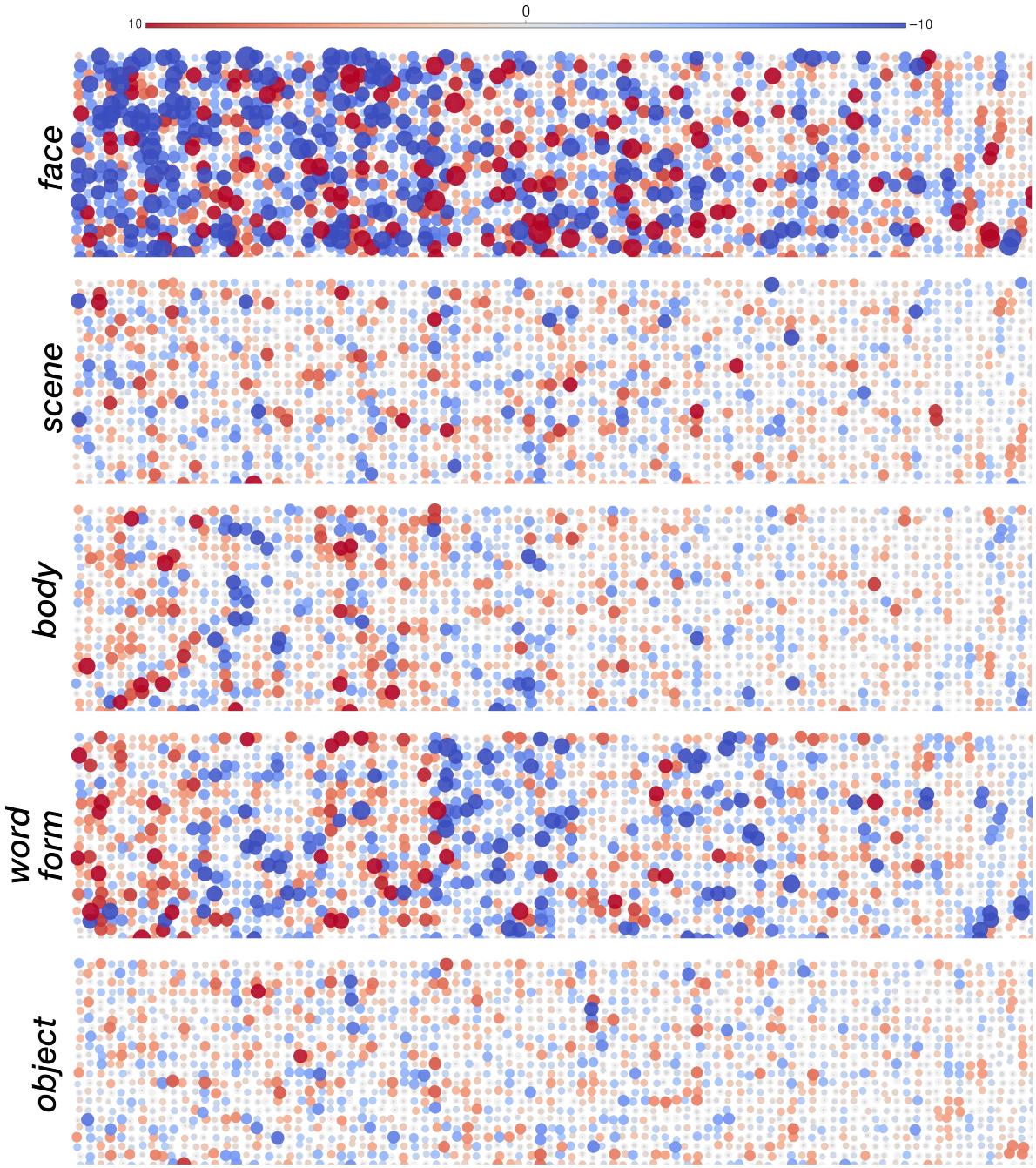

Figure S5: **Topograhical arrangement according to object categories in deep layers of LLCNN-L model.** Unit responses were assessed concerning their selectivity to five distinct categories of images, namely face, scene, body, characters, and objects. Selectivities are shown for the 4 convolutional layers in block 4 of the ResNet18 architecture. Color denotes the selectivity of each unit to each category.

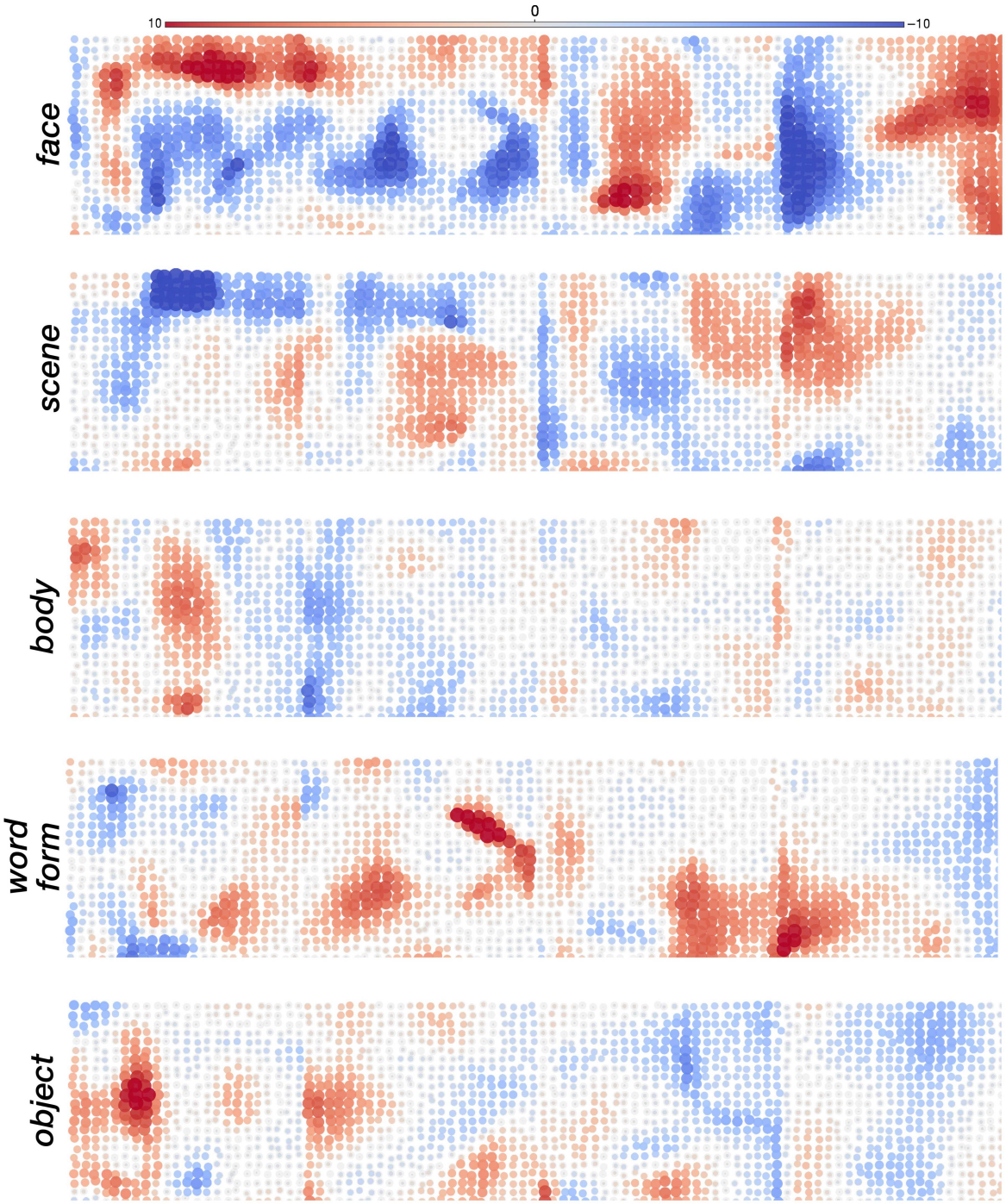

Figure S6: **IT Topography in LLCNN model trained unsupervised via Barlow's Twins [ZJM+21].** Unit responses were assessed concerning their selectivity to five distinct categories of images, namely face, scene, body, characters, and objects. Selectivities are shown for the 4 convolutional layers in block 4 of the ResNet18 architecture. Color denotes the selectivity of each unit to each category.

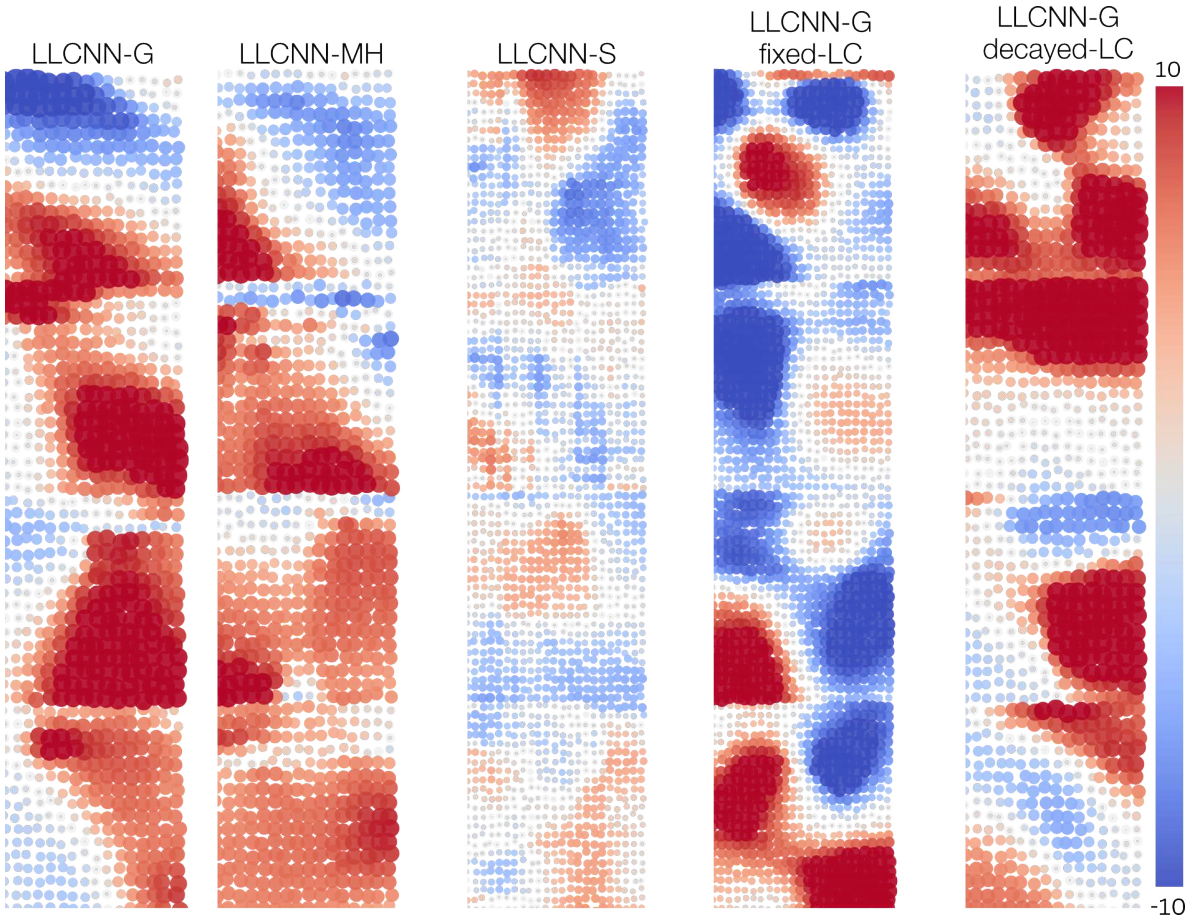

Figure S7: **Face selectivity in LLCNN model across different model variations and training regimes.** Continuous face-selective patches emerged in all variations of the LLCNN model. Face selectivity index was visibly lower in the LLCNN-S model, possibly because of its lower performance on Imagenet categorization task. Training with fixed lateral connections (fixed-LC) led to more blob-like category selective patches and more discontinuities in category-selectivity across layers compared to the version with exponentially decaying lateral connectivity size (decayed-LC).

Figure S8: **A map of best stimuli for units 4 consecutive layers of block3 in LLCNN-G model.** The three rows correspond to stimuli from [WRF+23], [JALT21], and [PC02] respectively.

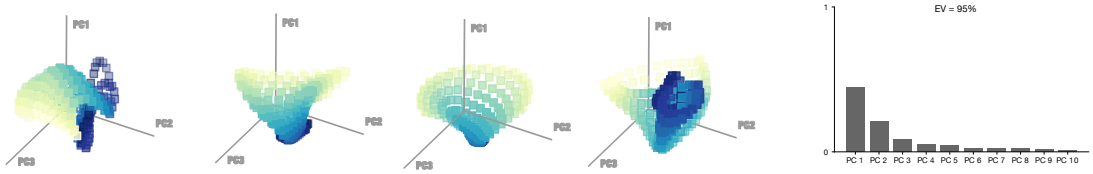

Figure S9: **Low dimensional manifold of activity profiles for block 3 of LLCNN-G model.** Plots are similar to those in Fig. 8a but for block 4 of the LLCNN-G model. The activity profiles lie on a low dimensional manifold that preserves the spatial arrangement of the untis on the simulated cortical sheet as well.