# A connectivity-constrained computational account of topographic organization in high-level visual cortex

Nicholas M. Blauch[1,2] Marlene Behrmann[2,3] David C. Plaut[2,3]
[1]Program in Neural Computation [2]Neuroscience Institute [3]Department of Psychology
Carnegie Mellon University
{blauch,behrmann,plaut}@cmu.edu

May 29, 2021

**Abstract**

Inferotemporal cortex (IT) in humans and other primates is topographically organized, with multiple domain-selective areas and other general patterns of functional organization. What factors underlie this organization, and what can this neural arrangement tell us about the mechanisms of high level vision? Here, we present an account of topographic organization involving a computational model with two components: 1) a feature-extracting encoder model of early visual processes, followed by 2) a model of high-level hierarchical visual processing in IT subject to specific biological constraints. In particular, minimizing the wiring cost on spatially organized feedforward and lateral connections within IT, combined with constraining the feedforward processing to be strictly excitatory, results in a hierarchical, topographic organization. This organization replicates a number of key properties of primate IT cortex, including the presence of domain-selective spatial clusters preferentially involved in the representation of faces, objects, and scenes, within-domain topographic organization such as animacy and indoor/outdoor distinctions, and generic spatial organization whereby the response correlation of pairs of units falls off with their distance. The model supports a view in which both domain-specific and domain-general topographic organization arise in the visual system from an optimization process that maximizes behavioral performance while minimizing wiring costs.

## 1 Introduction

Inferotemporal cortex (IT) subserves higher-order visual abilities in primates, including the visual recognition of objects and faces. By adulthood in humans, IT cortex, and ventral temporal cortex more generally, contains substantial functional topographic organization, including the presence of domain-selective spatial clusters in reliable spatial locations, including clusters for faces (Kanwisher et al., 1997; Gauthier et al., 2000b; Grill-Spector et al., 2017), objects (Grill-Spector et al., 2000), buildings and scenes (Aguirre et al., 1998; Epstein and Kanwisher, 1998), and words (McCandliss et al., 2003). Similar domain-level topographic properties have been found in rhesus macaque monkeys, including multiple regions of clustered face selectivity (Tsao et al., 2003, 2006; Freiwald and Tsao, 2010). Intriguingly, this selectivity is encompassed in a larger scale "mosaic" of category-selectivity, in which areas of category-selectivity themselves have further columnar clustering within them (Sato et al., 2013; Tanaka, 1996, 2003), pointing to more general principles of organization beyond the domain level. In line with this idea, human IT cortex also exhibits larger-scale organization for properties such as animacy and real-world size (Konkle and Oliva, 2012; Konkle and Caramazza, 2013), and midlevel features characteristic of these properties and domains have been shown to account well for patterns of high-level visual selectivity (Long et al., 2018). How these domain-level and more general facets of

functional organization arise, how they are related, and whether and in what ways they rely on innate specification and/or experience-based developmental processes remain contentious. Recent work has demonstrated that the neural basis of face recognition depends crucially on experience, given that deprivation of face viewing in juvenile macaque monkeys prevents the emergence of face-selective regions (Arcaro and Livingstone, 2017). Relatedly, the absence of exposure to written forms through reading acquisition precludes the emergence of word-selective regions (Dehaene et al., 2015; Carreiras et al., 2009). That there exists a clustered neural response selectivity for evolutionarily new visual categories such as written words offers further evidence that the topographic development of the human visual system has a critical experience-dependent component (Dehaene and Cohen, 2007; Arcaro et al., 2019). In contrast with a system in which innate mechanisms are specified a priori through natural selection, this experiential plasticity permits the tuning of the visual system based on the most frequent and important visual stimuli that are actually encountered, thereby enabling greater flexibility for ongoing adaptation across the lifespan.

There is considerable computational evidence that experience-dependent neural plasticity can account for the response properties of the visual system at the single neuron level. Classic work demonstrated that the statistics of natural images sufficed for the learning of V1-like localized edge-tuning within a sparse coding framework (Olshausen and Field, 1996, 1997). More recently, the more complex, high-level tuning of IT neurons has been predicted well by deep convolutional neural networks (DCNNs) trained on image classification. Moreover, neural responses at earlier levels of the visual hierarchy (V2, V4) have also been reproduced by these DCNNs, in a hierarchically consistent manner, where deeper layers of the DCNN map onto later layers of the anatomical hierarchy (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014).

Above the single-neuron level, considerable prior work has demonstrated that topographic organization in V1 may emerge from self-organizing, input-driven mechanisms (von der Malsburg, 1973; Linsker, 1986; Miller et al., 1989; Durbin and Mitchison, 1990; Obermayer et al., 1990; Goodhill and Willshaw, 1990; Goodhill, 1993) (for review, see Swindale, 1996). For example, the pinwheel architecture of spatially repeating smooth orientation selectivity overlaid with global retinotopy has been shown to be well-accounted for by Kohonen's Self-Organizing Map (SOM) (Kohonen, 1982; Durbin and Mitchison, 1990; Obermayer et al., 1990). The SOM contains a single map layer that applies Winner-Take-All (i.e., competitive) Hebbian learning with local cooperation, and learns to faithfully represent the input space, with neighboring units learning similar tuning functions and distant units learning more dissimilar tuning functions. As the SOM decreases its local cooperative radius over learning, the topography of unit response properties within the map comes to approximate both the larger and smaller scale patterns of similarity in the input data. The ability of the SOM to account for pinwheel organization has lead to the proposal that pinwheels reflect a dimensionality-reduced mapping of low-level image statistics (i.e., 2D retinotopic location and orientation) (Durbin and Mitchison, 1990; Obermayer et al., 1990).

Some work has applied SOMs to higher-level topographic organization. For example, the stronger topographic clustering in IT for faces compared to other object categories (e.g., chairs, shoes) has been shown to arise in SOMs applied to mid-level feature representations of natural images, due to the greater within-category similarity among faces compared to other categories (Cowell and Cottrell, 2013). However, in general, the vast majority of computational approaches investigating topographic mappings in the visual system have been applied to early visual cortex. One possible reason for this is that the SOM, and other related topographic mapping algorithms, is a single-layer system that re-expresses existing input similarities topographically but does not reorganize or transform these similarities hierarchically, the way needed to untangle invariant visual object representations from the statistics of natural images (DiCarlo et al., 2012). The other key reason for limited application to IT cortex is that—until the recent application

of DNNs—the complex and invariant tuning of the IT feature space had been difficult to capture using more standard approaches (Yamins and DiCarlo, 2016).

Other neural network models are capable of learning hierarchical topographic representations by incorporating distance-dependent constraints on learning into a general neural network model, rather than designing the neural network algorithm around topographic organization *per se* (Kohonen, 1982, 1995). For example, the Laterally Interconnected Synergistically Self-Organizing Map (LISSOM) (Sirosh and Miikkulainen, 1997; Miikkulainen et al., 2006) used Gaussian receptive fields connecting areas in an arbitrarily-deep hierarchy of 2D map layers, and was able to account for a range of topographic properties such as retinotopy, ocular dominance, and orientation hypercolumns. However, the learned representations were less clean than in the SOM, and the model was limited in expressive power by its use of Hebbian learning rather than the error-driven learning employed in the powerful deep learning architectures capable of predicting IT neuron properties. Jacobs and Jordan (1992) developed a similar hierarchical neural network, but rather than imposing fixed Gaussian receptive fields, they utilized a spatial loss function that penalized large weights to a greater degree on longer versus shorter connections. Moreover, rather than Hebbian learning, they used error-driven learning, where weight updates in the network simultaneously sought to maximize task performance and minimize spatial wiring costs. This model was shown to lead to the formation of topographic organization in simple neural networks trained on structured tasks. Plaut (2002); Plaut and Behrmann (2011) used a similar approach in which weight error gradients during learning were reduced as a function of connection length. Plaut and Behrmann (2011) demonstrated that a spatially-constrained model could account for the global organization of domain-specific areas in IT cortex based on input and output demands, with face and word selectivity emerging close to foveally-responsive cortex due to demands on foveal retinotopic inputs, and scene selectivity emerging near peripherally responsive cortex due to demands on peripheral inputs (Levy et al., 2001). Moreover, Plaut and Behrmann (2011) accounted for the lateralization of words to the left-hemisphere owing to a pressure to be functionally coupled with left-lateralized language representations, and the resulting lateralization of faces to the right-hemisphere as a result of competition with words for foveally biased model cortex. In general, however, these models have been limited to small-scale demonstrations, and their topography has been analyzed at only a relatively coarse level. Nonetheless, this early work raises the possibility that the application of distance-dependent constraints in a modern deep neural architecture trained on natural images might provide a more comprehensive account of topographic organization in IT.

A useful first step in modeling the topography of IT cortex using deep neural networks (Lee et al., 2020) successfully accounted for a number of characteristics of face-selective neurons in non-human primates, including their topographic organization, by explicitly encouraging units within a layer of the network to be spatially nearer to units with correlated responses, and farther from units with uncorrelated or anti-correlated responses. Note, however, that this approach *imposes* topographic functional organization on the network rather than *deriving* it from more basic principles of cortical structure and function, such as constraints on connectivity. Ideally, in a model of IT cortex, topography should emerge naturally from such basic network principles, rather than being imposed based on measurements of functional responses. This would provide greater insights into both the functional and structural constraints on development and computation within the visual system.

In the current work, we combined the approaches of task-optimized DCNN modeling (Yamins and DiCarlo, 2016; Lee et al., 2020) with flexible connectivity-constrained architectures (Jacobs and Jordan, 1992; Plaut and Behrmann, 2011) to develop a hierarchical model of topographic organization in IT cortex. We implemented a bias towards local connectivity through minimization of an explicit wiring cost function (Jacobs and Jordan, 1992) alongside a

task performance cost function. We observed that this pressure on local connectivity was, on its own, insufficient to drive topographic organization in our model. This led us to explore two neurobiological constraints on the sign of connectivity—strictly excitatory feedforward connectivity, and the separation of excitation and inhibition—with the result that both, and particularly excitatory feedforward connectivity, provided a powerful further constraint on the development of topographic organization when combined with local connectivity. We explored the functional specialization of the models, revealing a large but graded degree of domain specialization. Last, we uncovered more detailed forms of topographic organization in the model, recapitulating the finding that nearby units exhibit greater correlation in their responses over different images, and that more distant units exhibit less correlation than nearby units (Lee et al., 2020). This distance-dependent organization of responses manifests in other intriguing ways beyond domain organization, evincing sub-domain topographic clustering for animate-inanimate objects, indoors-outdoors and natural-manmade scenes, and face genders. Our computational model thus builds on existing research by proposing that topographic organization arises through task-optimized, low wiring cost solutions, and by incorporating other well-established sign-based constraints on neural computation. The general principles and mechanistic approach captured in the present research holds promise for elucidating other aspects of functional brain organization in both humans and non-human primates, both within and outside of the visual system.

# 2 Results

## 2.1 A biologically-constrained model of ventral temporal cortex produces hierarchical, domain-selective response topography

Our Interactive Topographic Network (ITN) framework for modeling high-level visual cortex consists of an *encoder* that approximates early visual cortex, followed by *interactive topography* (IT) areas that approximate inferotemporal cortex (Figure 1; see Methods for details). We first present the results of simulations of a single architectural variant of the ITN model, in which a ResNet-50 encoder is pre-trained on a large dataset including several categories from the domains of objects, faces, and scenes (each domain matched in total training images). The trained encoder provides input to a 3-area IT with separate posterior (pIT), central (cIT), and anterior (aIT) areas. Each IT area consists of separate banks of excitatory (E) and inhibitory (I) units, and feedforward connectivity between areas is limited to the E units. After training, the model performed well on each domain, reaching a classification accuracy of 86.4% on the face domain, 81.8% on the object domain, and 65.9% on the scene domain (see Supplementary Figure S1). Performance differences across domains are unlikely to be an artifact of the specific architecture as they can be seen across a variety of CNNs, reflecting the intrinsic difficulty of each task given the variability within and between categories of each domain for the given image sets.

The trained model exhibited hierarchical, topographically organized domains, as shown in Figure 2A. We found that domain-level topography was linked across corresponding sectors of each hierarchical layer. This result demonstrates that the distance-dependent constraints on feedforward connectivity pressured units that have minimal between-area distances to learn a similar tuning, which means that each layer is roughly overlapping in their respective (separate) 2D topography. We also found that the topographic organization gets somewhat smoother moving from pIT to cIT in particular, most likely because units in cIT and aIT (but not pIT) have local feedforward receptive fields and thus greater constraint on local cooperation.

We next scrutinized the topography in aIT, where there are very smooth domain-level responses, and where we can directly compare responses with those of the recognition readout mechanism. We computed mean domain

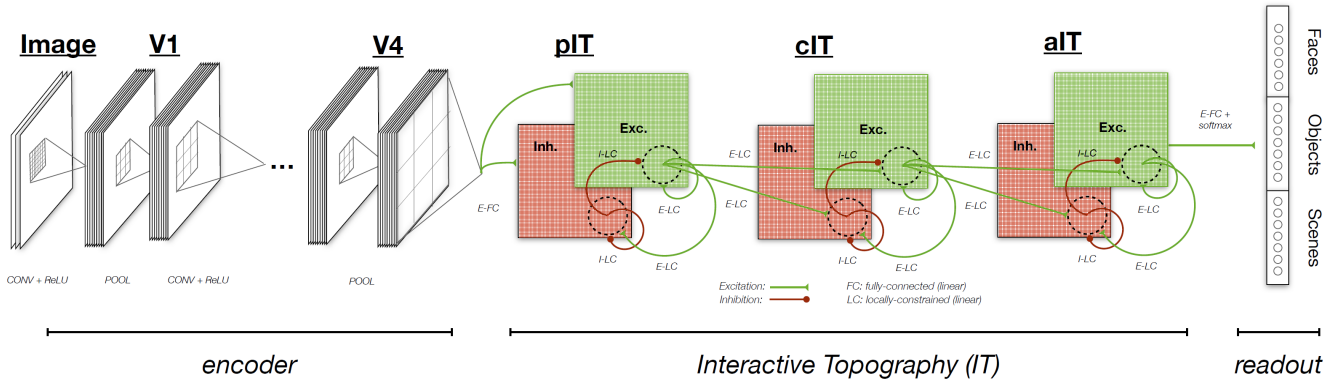



**Figure 1:** Architecture diagram of the Interactive Topographic Network (ITN). An ITN model consists of three components: an *encoder* that approximates early visual processing prior to inferotemporal cortex, the *interactive topography* (IT) areas that approximate inferotemporal cortex, and the *readout* mechanism for tasks such as object, scene, and face recognition. The architecture of each component is flexible. For example, a 4-layer simple convolutional network or a deep 50-layer ResNet can be used as the encoder; whereas the former facilitates end-to-end training along with a temporally-precise IT model, the latter supports better learning of the features that discriminate among trained categories. In this work, topographic organization is restricted to the IT layers. The figure depicts the main version of the ITN containing three constraints: a spatial connectivity cost pressuring local connectivity, separation of neurons with excitatory and inhibitory influences, and the restriction that all between-area connections are sent by the excitatory neurons. The final IT layer projects to the category readout layer containing one localist unit per learned category, here shown organized into three learned domains. (Note that this organization is merely visual and does not indicate any architectural segregation in the model.)

responses, plotted in the first column of Figure 2B, and domain selectivity, plotted in the second column, which demonstrates corresponding topographic organization. We confirmed the functional significance of response topography by conducting a searchlight analysis inspired by multivariate approaches to analyzing functional magnetic resonance imaging (fMRI) data (Kriegeskorte et al., 2008). The searchlight analysis, shown in the third column of Figure 2B, revealed a topographic organization of information for discriminating between categories of each domain—an organization which is strongly correlated with the domain selectivity maps for each domain (all $ps < 0.0001$).

To further confirm the functional significance of the topographic organization, we analyzed the spatial organization of readout weights from aIT to the localist category readout layer. We evaluated whether each domain placed more weight in reading out from the units for which there was greater selectivity, by calculating the mean domain response weight for each unit, averaged over classes in each domain. This produced a map for each domain, shown in the last column of Figure 2B. We find a large positive correlation between the mean readout weight and the mean response for each domain (all $r$s>0.7, all $p$s< 0.0001), further demonstrating the functional significance of the response topography.

## 2.2 Domain-level functional organization indicates strong yet graded specialization

We next performed a series of 'lesion' analyses in the model in order to compare with neuropsychological data on face and object recognition (Farah, 1990; Moscovitch et al., 1997; Geskin and Behrmann, 2017). First, we performed focal lesions, as would be experienced by most patients with acquired brain damage. To simulate the impairment of patients with maximally specific deficits, we located these focal lesions at the center of domain selectivity. Our approach for this involved first smoothing the domain-selective maps by averaging the metric at each unit using the 5% nearest units on the 2D grid, and then finding the unit with the peak metric value. We then designed circular lesions of various sizes centered on this peak unit, removed the designated units from the model, and measured the resulting performance for recognizing images from each domain separately.

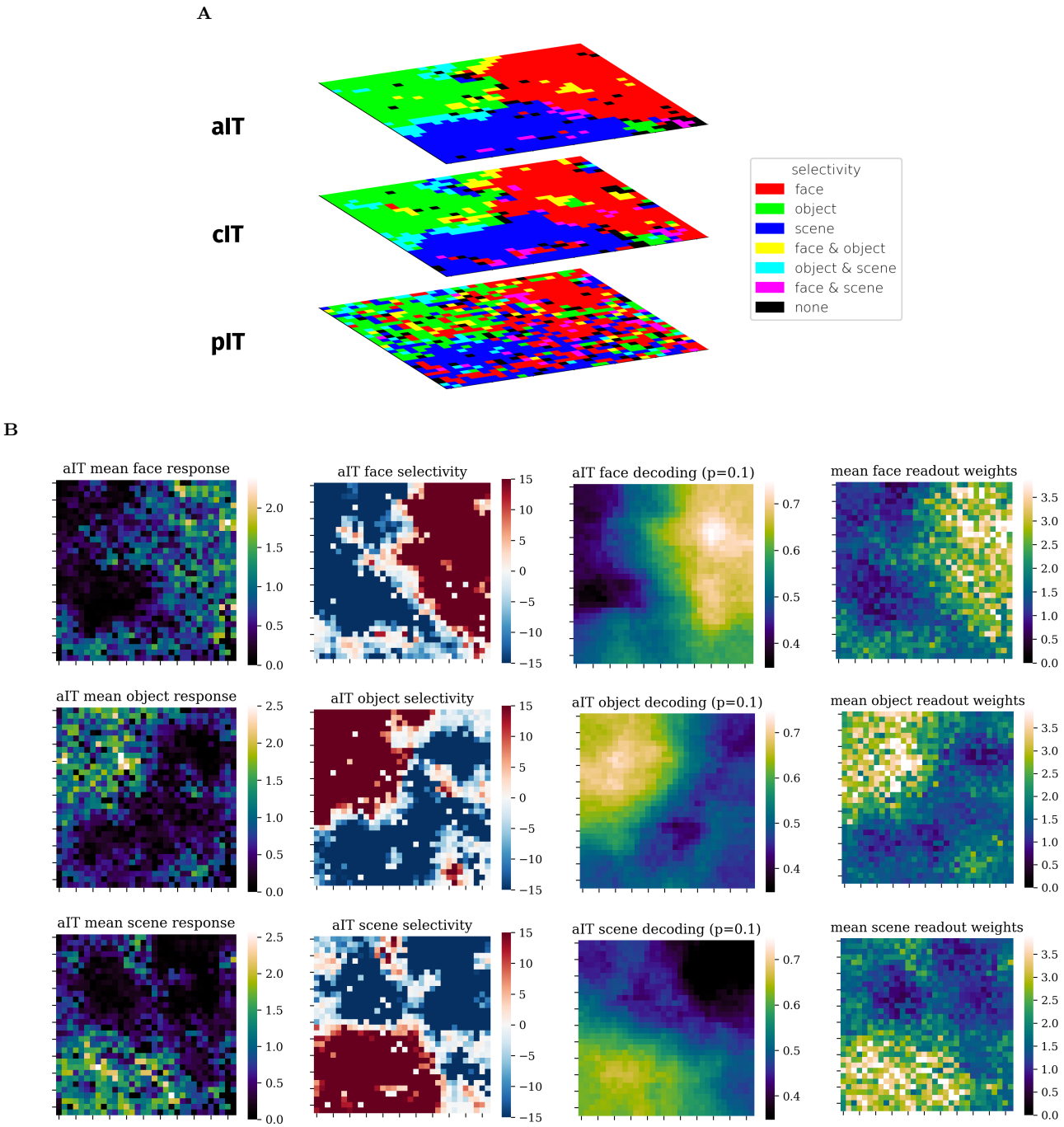



**Figure 2:** Domain-selective topography in the Interactive Topographic Network. **A.** Domain selectivity at each level of the IT hierarchy. Selectivity is computed separately for each domain, and then binarized by including all units corresponding to $p < 0.001$. Each domain is assigned a color channel in order to plot all selectivities simultaneously. Note that a unit can have zero, one, or two selective domains, but not three, as indicated in the color legend. **B.** Detailed investigation of domain-level topography in aIT. Each heatmap plots a metric for each unit in aIT. The first column shows the mean domain response for each domain, the second column row shows domain selectivity, the third column shows the within-domain searchlight decoding accuracy, and the fourth column shows the mean of weights of a given aIT unit into the readout categories of a given domain.

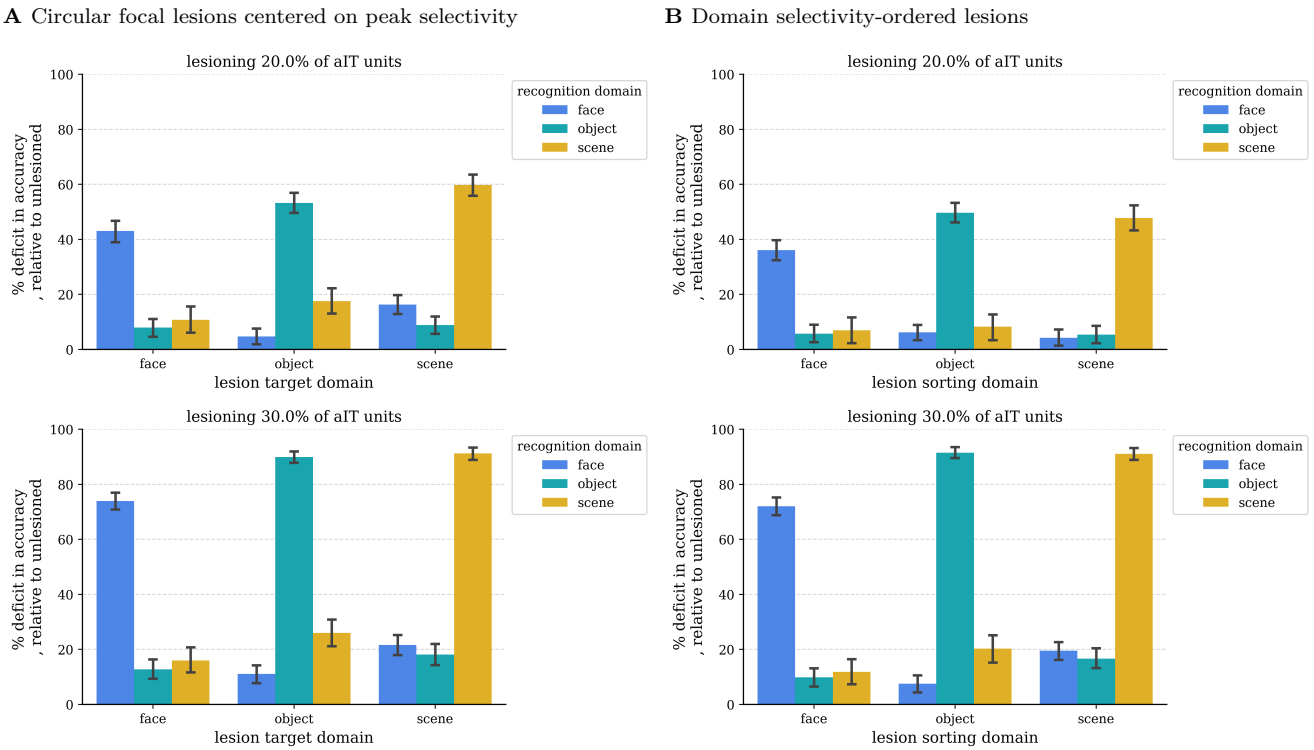



**Figure 3:** Lesion results in the ITN model. Each plot shows the relative effects of a set of same-sized lesions on recognition performance for each domain, relative to the performance on the same domain in the undamaged model. Error bars show bootstrapped 95% confidence intervals over trials; thus, the statistical significance of a given lesion can be assessed by determining whether the confidence interval includes 0. **A.** Damage from various sizes of circular focal lesions centered on the peak of smoothed selectivity for each domain. **B.** Damage from various sizes of selectivity-ordered lesions for each domain.

The results of this lesion analysis for two representative lesion sizes—using 20% and 30% of the aIT units—are shown in Figure 3A. Focal lesions centered on each domain led to an especially severe deficit in recognition for that domain, and milder but significant deficits for the other domains as well (all $p$s<0.05). With a 30% lesion, more substantial domain-general effects were seen, even while performance on the selected domain was not entirely impaired (see Supplementary Figure S2 for an examination of a broader range of lesion sizes, including large lesions that result in profound impairment).

Are these more general effects of circumscribed lesions on non-preferred domains the result of imperfect (patchy) or non-circular topographic organization of an underlying modular organization? To answer this question, we performed selectivity-ordered lesions, in which units were sorted by their selectivity for a given domain, and selected according to their sorting index, shown in Figure 3B. In general, for 20% lesions, we found smaller but still significant deficits for both the preferred and non-preferred domains. This suggests that some but not all of the damage to the non-preferred domain induced by focal lesions may be due to imperfect or non-circular topographic functional organization. Importantly, these more distributed effects of lesions indicate that the functional organization, while highly specialized, is not strictly modular, at least with respect to one influential definition of modularity (Fodor, 1983). Supplementary Figures S2, S3, and S4 provide additional data on the nature of domain specialization in the network.

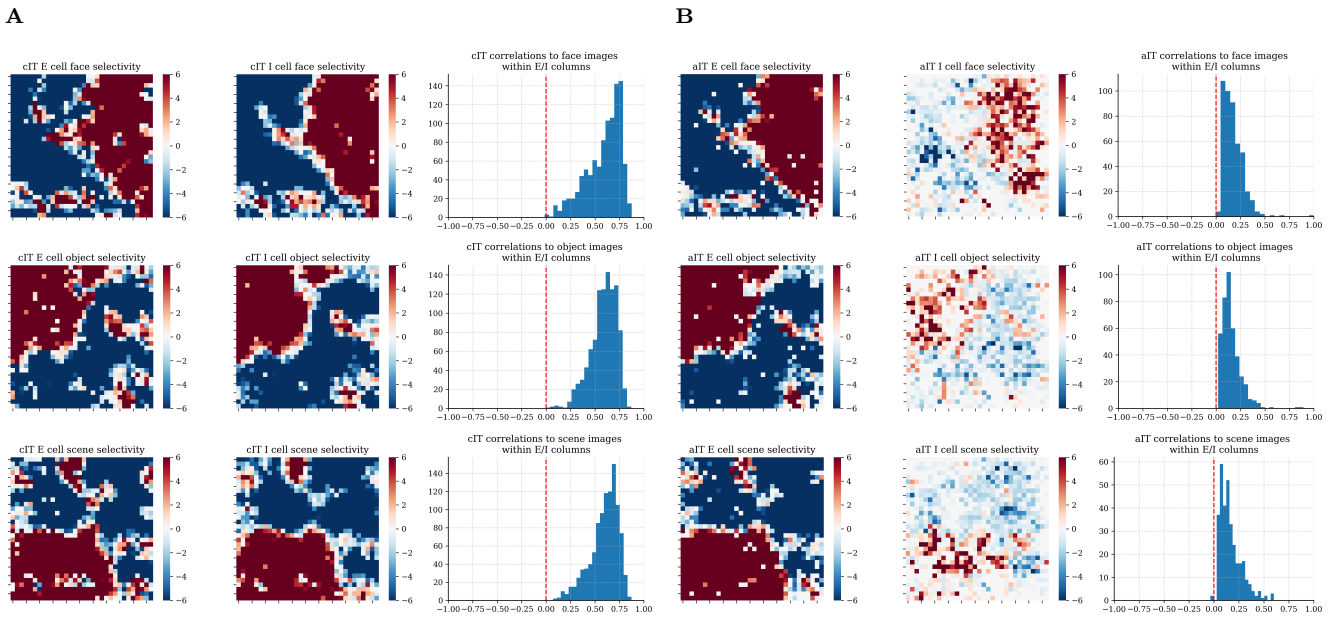



**Figure 4:** E/I columnar response selectivity. The selectivity for faces, objects, and scenes, respectively, is plotted for both E and I cells, and the histogram of correlations between E and I cells over all images of the given domain is computed over all locations on the grid, revealing dominantly positive correlations. **A.** cIT **B.** aIT.

## 2.3 Excitatory and inhibitory units operate as functional columns

Thus far we have focused on the representations in the E cells, both for convenience and clarity, and because it is the E units that exclusively project to downstream areas (including the category readout units). We next assessed whether the I units showed a similar topographic organization, and whether it was linked with the E cells. The selectivity for E and I cells is plotted and correlated in Figure 4. As evident from this figure, the I cells did indeed show similar domain-selective topography, and, moreover, this topography was linked with that of the E cells. Moreover, the activities of E and I units in the same 2D location had highly correlated activities over each domain of images, as well as over all images. If we consider a pair of E and I neurons at a given location on the 2D sheet to correspond to a cortical column, our result is reminiscent of the finding that biological neurons in different layers at the same location on the 2D flattened cortex have similar response properties (Hubel and Wiesel, 1969). Put simply, E and I units appear to act as functional columns. Intriguingly, in aIT, the population of I cells responded more sparsely than the E cells, suggesting that fewer I cells would suffice. This is consistent with the finding that the brain contains far fewer inhibitory than excitatory neurons (see also Alreja et al., 2020).

## 2.4 Domain selectivity exists within a broader organizational scheme similar to that of primate IT cortex

Previous empirical research has demonstrated that the response correlations between pairs of neurons fall off roughly exponentially with distance (data from Majaj et al., 2015, as plotted in Lee et al. (2020)). This finding has been used to develop a class of topographic neural network models that explicitly fits the spatial layout of units to this relationship (Lee et al., 2020). We explored whether this relationship emerged naturally in our network under its constrained connectivity, in line with the emergence of domain-selective topography. We thus computed the correlations among pairs of unit activations across images as a function of the distance between the units, focusing on aIT. As shown in Figure 5, there is, indeed, an exponential decay of response correlations with distance, matching the qualitative trend

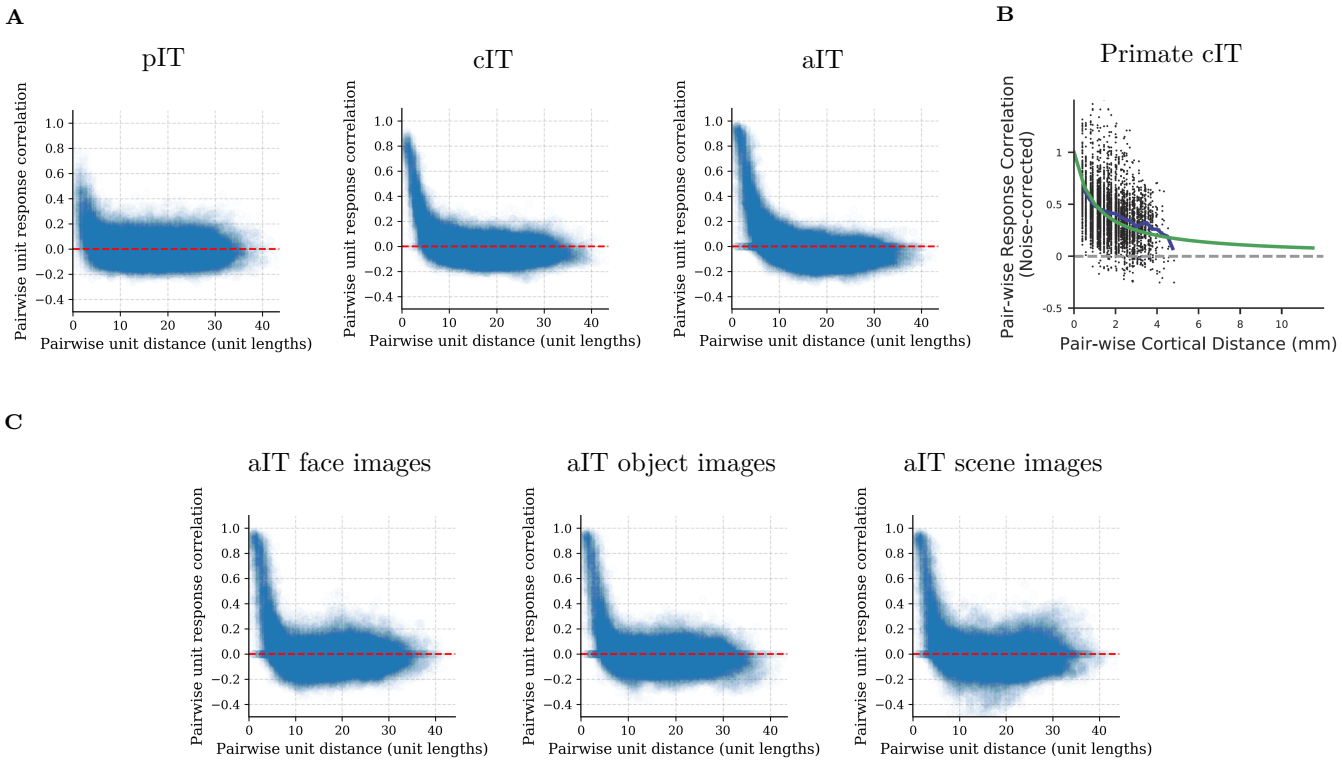



**Figure 5:** Distance-dependent response correlation arises from biologically-plausible constraints on connectivity. **A.** Distance-dependent response correlation in the excitatory cells of each layer, using images from all three domains (objects, faces, scenes). **B.** Distance-dependent response correlation in macaque IT (reproduced from Lee et al., 2020, per CC-BY-NC-ND license). **C.** Distance-dependent response correlation in aIT using images from each domain separately, highlighting within-domain generic functional organization.

in the empirical data (Lee et al., 2020; Majaj et al., 2015). This result is not explained by mere differences between domains, as it also exists when examining responses to images within each domain separately (see the bottom of Figure 5). In line with those of Lee et al. (2020), our results indicate that the domain-level topography is simply a large-scale manifestation of a more general representational topography in which the information represented by neighboring units is more similar than that represented by more distal units and, moreover, that this scheme can arise under explicit wiring length and sign-based constraints on connectivity.

## 2.5 Within-domain clustering of categories in aIT

The generic distance-dependent functional relationship shown in the previous section would suggest that functional organization may be exhibited at finer scales than the domain level. To assess this, we performed a clustering analysis on the readout weights from aIT. We adopted this approach due to the similarity between the readout weights and response topography in aIT (Figure 2). A given category will achieve its maximal output response when the activation pattern in aIT most closely aligns with the readout weights. Thus, the readout weights for a given category act as a sort of category template to match with representations in aIT. Clustering the readout weights directly, rather than interpreting a set of activations to natural images, allows for clustering solutions to be explicitly linked to each category. This allows for a concise clustering solution containing one element for each category: the readout weights projecting from aIT to the identity unit for that category. We thus performed k-means clustering on the readout weights of all categories separately for each domain using $k = 3$ clusters, finding the centroids of these clusters, and

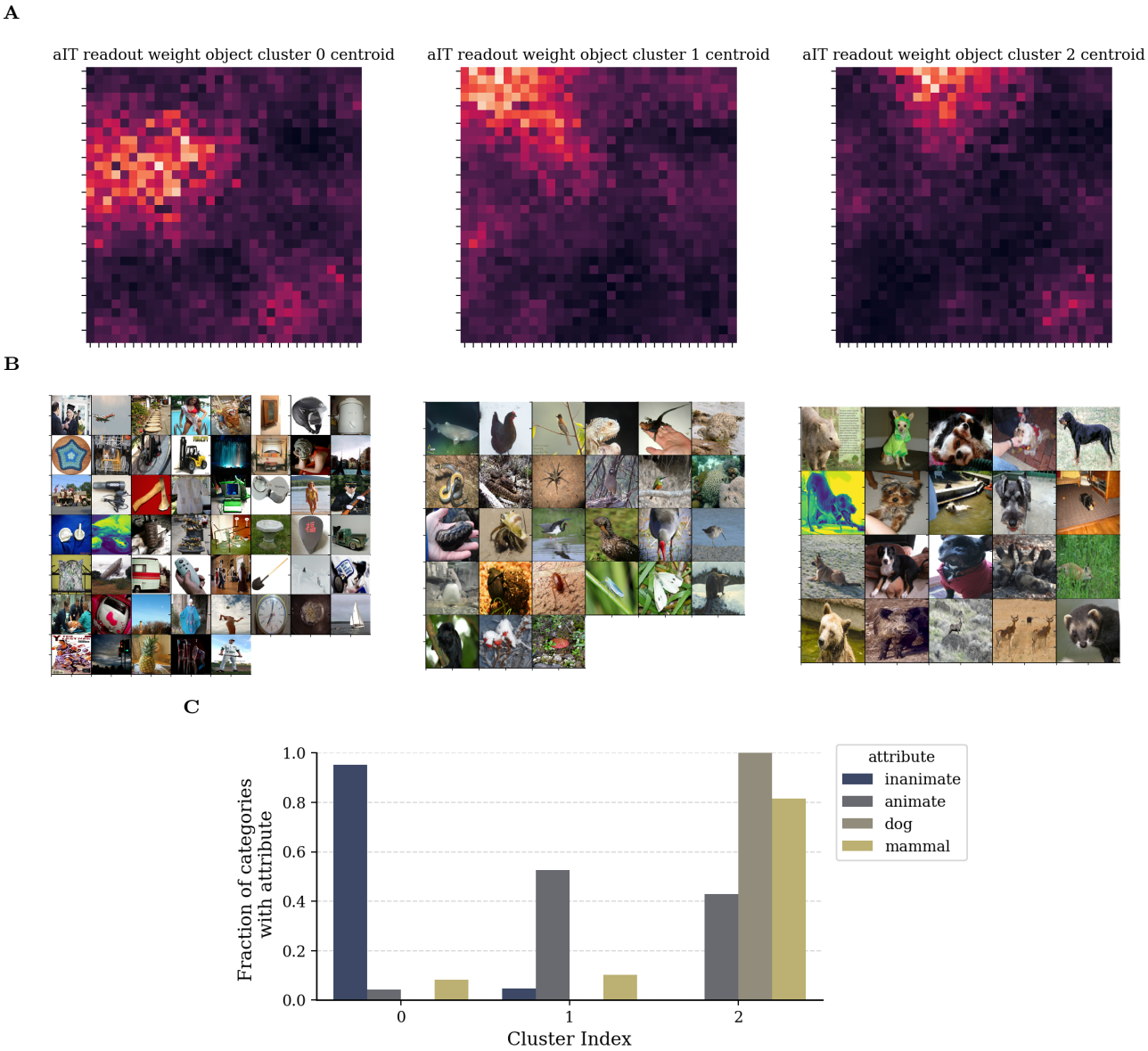



**Figure 6:** Clustering object category readout weights from aIT. $k = 3$ clusters were used in a k-means++ clustering algorithm (Arthur and Vassilvitskii, 2007) that clustered the vectors of aIT readout weights over all 100 object categories. **A.** The centroids of the clusters were then reshaped into the 2D aIT coordinates and visualized as heat maps. Note that there is no topographic constraint on the clustering algorithm itself. **B.** Visualization of one image from each of the categories contained in each cluster. **C.** Quantification of three attributes (animate, dog, mammal) over the members of each cluster, revealing quantitative characteristics of each cluster.

visualizing them in the 2D layout of aIT.

The results for the object stimuli are shown in Figure 6. Visual inspection suggests a striking organization for different classes of object categories, which is confirmed through cluster assignment quantification in Figure 6C. The first two clusters represent the vast majority of animate categories, with the first cluster representing mostly non-mammalian animate categories such as birds and reptiles, and the second cluster representing mostly dogs and other mammals such as bears and raccoons. Last, the third cluster represents the vast majority of inanimate objects such as clocks and various tools.

The same analysis for the scene categories (see Figure 7) revealed a similarly striking organization in which

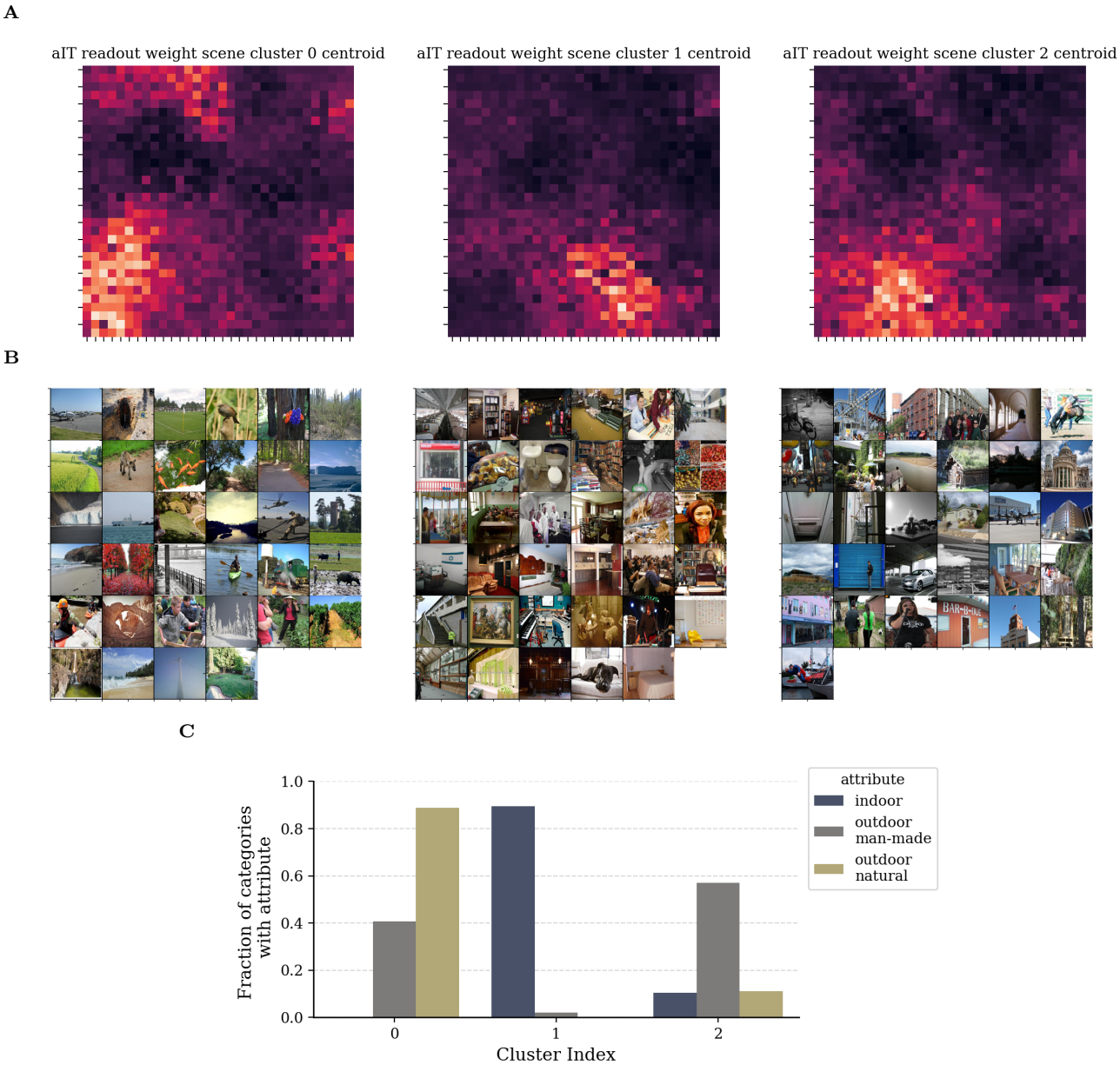



**Figure 7:** Clustering scene category readout weights from aIT. $k = 3$ clusters were used in a K-means++ clustering algorithm that clustered the vectors of aIT readout weights over all 100 scene categories. **A.** The centroids of the clusters were then reshaped into the 2D aIT coordinates and visualized as heat maps. **B.** Visualization of one image from each of the scene categories contained in each cluster. **C.** Quantification of three binary but non-mutually exclusive attributes (indoor, outdoor/man-made, outdoor/natural) over the members of each cluster, revealing quantitative characteristics of each cluster.

scenes were topographically clustered by both indoor-outdoor and natural-manmade dimensions. Particularly, the first cluster contained a preponderance of outdoor scenes and all of the natural outdoor scenes, the second cluster contained entirely indoor scenes, and the third cluster contained a high proportion of outdoor manmade scenes (quantified in Figure 7C). These clusters largely correspond to the right, left, and middle parts of the scene cluster, respectively. Although we do not show the results for face clustering (due to privacy concerns), we note that faces appeared to cluster topographically by gender and hair color (Supplementary Figure S7). Thus, the network exhibited within-domain clustering for each domain, further highlighting the general topographical organization induced by

task-optimized learning under distance-dependent and sign-based connectivity constraints.

## 2.6 Networks can reduce spatial costs and maintain performance by increasing topographic organization

The optimization problem introduced by (Jacobs and Jordan, 1992) and employed in this work (Equation 4) explicitly works to both maximize visual recognition performance through a task-based loss term $\mathcal{L}_t$, and to minimize overall wiring cost through a connection-based lost term $\mathcal{L}_w$ that scales with the square of connection distance. To what extent does minimizing this wiring cost term compromise performance? To answer this question, we tested multiple ITN models with varying wiring cost penalties $\lambda_w$ and measured the resulting wiring cost and task performance. We compute wiring cost in two ways. The first way is by using the $\mathcal{L}_w$ term, which takes into account both the length and strength of connections. The second way is inspired by the wiring cost minimization framework (Koulakov and Chklovskii, 2001), which cares only about the presence—rather than the strength—of connections, along with their distance. To compute such this unweighted wiring cost $\mathcal{L}_{w,u}$, we sparsified the network to contain only the 1% strongest connections, and took the sum of squares of these distances (Chklovskii, 2004); this sparsification introduces minimal performance deficits in the main ITN model (see Equation 6, and Supplementary Figure S10) The results, shown in Figure 8, demonstrate that increasing the wiring cost penalty $\lambda_w$ by an order of magnitude decreased the first spatial cost $\mathcal{L}_w$ by roughly an order of magnitude. Precisely, the log-log plot in Figure 8A revealed a power law relationship of the form $y = Ax^m$, where $m = -1.24$ ($p < 0.001$). Similarly, the unweighted wiring cost on sparsified networks $\mathcal{L}_{w,u}$ (Figure 8B) revealed a power law relationship with a somewhat weaker slope $m = -0.86$ ($p < 0.001$); the weaker slope as a function of $\lambda_w$ is to be expected given that this cost only accounts for the presence, and not the strength, of preserved connections following sparsification. In contrast to the wiring costs, the final classification performance was only marginally affected by $\lambda_w$ (log-log slope $m = -0.0016$, $p < 0.001$, explained variance $r^2 = 0.582$; fit was not significantly better than log-linear regression, $m = -0.0028$, $p < 0.001$, explained variance $r^2 = 0.583$) and the final top5 classification performance was unaffected by $\lambda_w$ ($p > .1$; see Figure 8C,D). Last, increasing the wiring cost penalty gradually resulted in the emergence of domain-selective areas, along with distance-dependent pairwise response correlations (see Figure 8E,F). Thus, models with a large wiring cost penalty perform similarly to models with unconstrained connectivity but achieve very small wiring cost, through the development of topographic functional organization.

## 2.7 Determining which constraints are necessary to develop topographic organization

Having established that the main ITN architecture produces a variety of interesting and empirically grounded topographic organizational phenomena, we next performed a constraint-removal analysis to determine which constraints are necessary for the development of topographic organization. We varied four constraints: whether between-area feedforward connections were excitatory only (EFF), whether the model employed separate E and I unit populations within each area (E/I), whether the model contained lateral (recurrent) connections within each area (RNN), and which connections the local constraint was applied to (all, ff, rec). We removed each of these constraints separately and jointly (see Methods for details) and compared the domain-selective topography for the resulting models with that of the standard model used in the earlier analyses.

As shown in Figure 9, a variety of simplified models lead to emergent topography, which is typically consistent across domain-level and generic measures. An important finding, however, is that local connectivity on its own (local RNN; fourth column) was insufficient to produce clustered domain-level topography, and produced a very weak generic

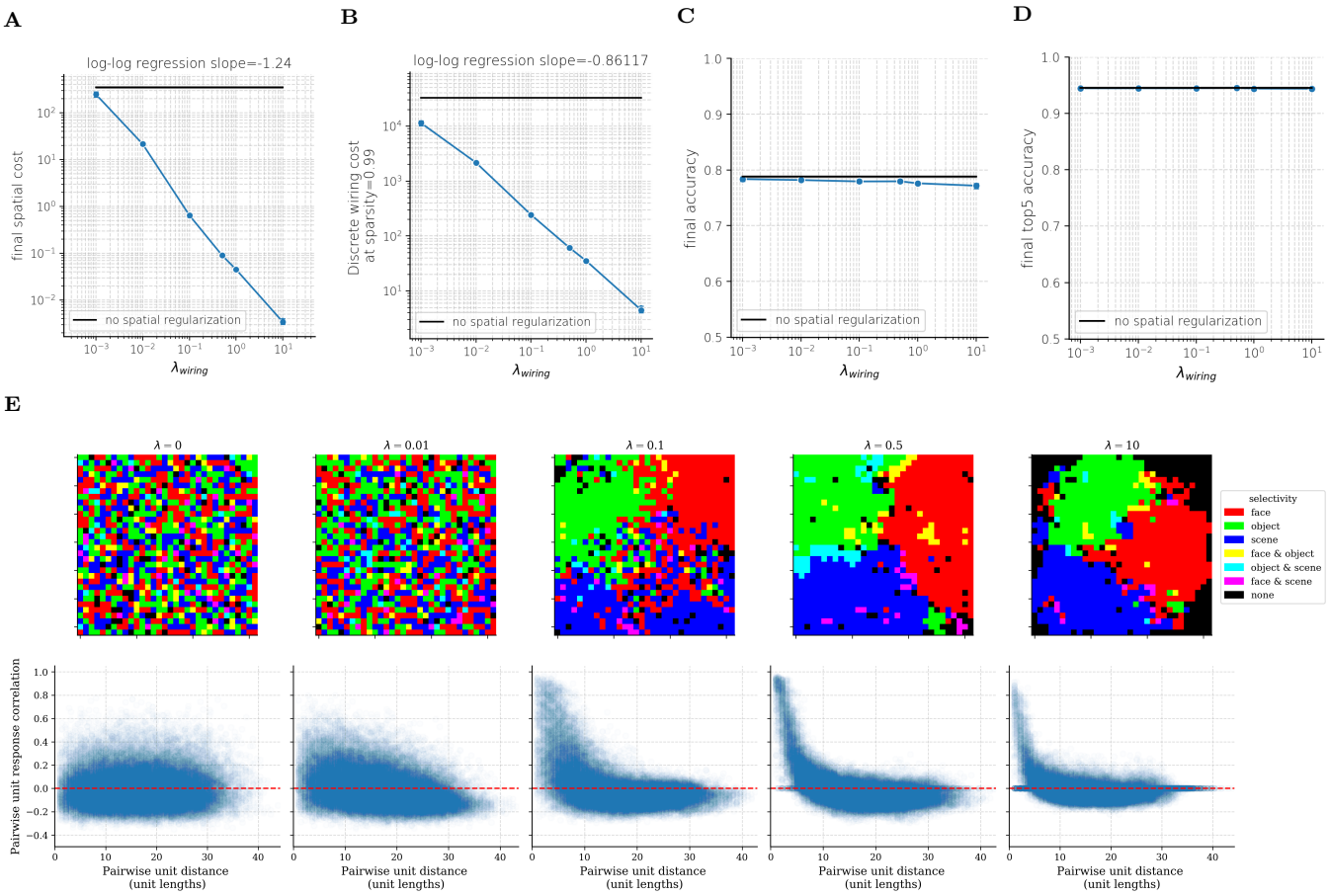



**Figure 8:** Spatial cost, visual performance, and emergent topography as a function of spatial wiring penalty $\lambda_w$. **A-D** use 4 randomly initialized model instances per regularization strength. Error bars around markers shows 95% confidence intervals for the plotted metric at a given spatial regularization strength, and black bands show 95% confidence intervals of the metric for a matched model without a spatial wiring penalty ($\lambda_w = 0$). **A.** Weighted spatial cost ($\mathcal{L}_w$) versus spatial wiring penalty $\lambda_w$, plotted on log-X and log-Y axes. **B.** Unweighted spatial cost following sparsification versus spatial wiring penalty $\lambda_w$, plotted on log-X and log-Y axes. **C.** Mean top1 accuracy on validation images from all domains versus versus spatial wiring penalty $\lambda_w$ on a log X-axis and linear Y-axis. **D.** Mean top5 accuracy on validation images from all domains versus spatial wiring penalty $\lambda_w$ on a linear Y-axis. **E.** Emergent topographic organization in one model instance of each spatial wiring penalty $\lambda_w$. Top row: domain-selective aIT topography. Bottom row: generic aIT topography.

distance-dependent response correlation over units. Each of the other models, which combined local connectivity (either feedforward, lateral, or both) with a sign-based constraint, led to greater topographic organization than the baseline local RNN, indicating that the sign-based constraints were important for the development of topography. In particular, we found that the excitatory constraint on feedforward connectivity led to very clean topographic organization in two simplified models: one removing the separation of excitation and inhibition (EFF-local-RNN; column 3), and another additionally removing learned lateral connectivity entirely (EFF-local-FNN+LN; column 7). This latter, most simplified model contains local feedforward excitatory connectivity and layer normalization (fixed lateral interactions serving to standardize layer pre-activations), and processes inputs for only one time step without learned recurrent connection weights. Since these models appear to generate similar topographic organization to the main model, we opted to evaluate their relative strengths and weaknesses in more detail.

Specifically, we compared the EFF-local-RNN (or *recurrent*) and EFF-local-FNN+LN (or *feedforward*) model variants with the main model, in terms of performance, between-area spatial cost, and functional specialization (for

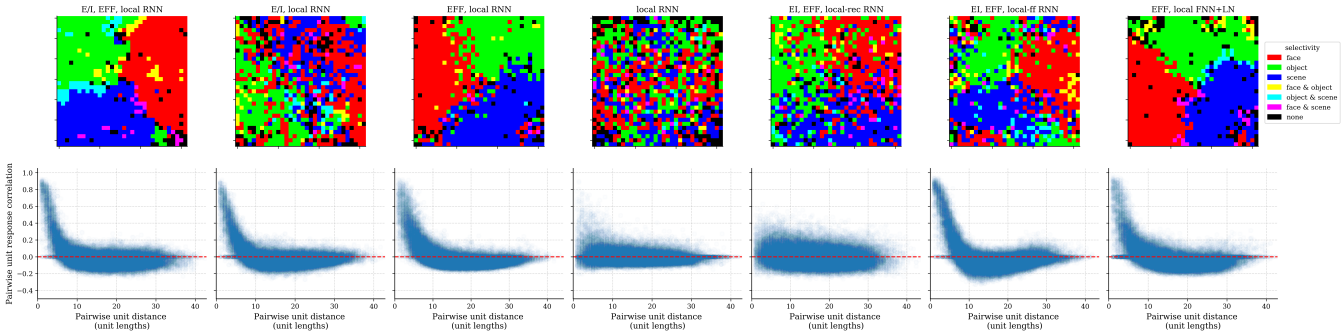



**Figure 9:** Domain-level and generic topography in layer aIT across models with different constraints implemented. From left to right: E/I EFF local RNN is the full model with separate excitation and inhibition, excitatory feedforward connections, and spatial regularization on recurrent and feedforward connections; E/I local RNN has separate excitation and inhibition but both cell types send feedforward connections to the next layer; EFF local RNN does not separate excitation and inhibition (one map of units), but feedforward connections are restricted to be excitatory; local RNN contains no sign-based constraints but has spatial regularization on recurrent and feedforward connections; E/I EFF, local-rec RNN has both sign-based constraints but spatial regularization only on recurrent (lateral) connections; E/I EFF, local-ff RNN has both sign-based constraints but spatial regularization only on feedforward connections; EFF local FNN+LN does not separate excitation and inhibition but limits feedforward connections to be excitatory, does not have recurrent connections besides the layer normalization also implemented in all of the other models, and has spatial regularization on the feedforward connections.

more detailed analysis of each model, see Supplementary Figures S5 and S6). We focused on between-area spatial cost for two reasons. The first, practical reason, is that each model contains such between-area connections, in contrast to lateral connections which are not present in the feedforward variant. The second reason is that between-area spatial costs may be substantially more biologically burdensome than within-area spatial costs, since they incorporate between-area distances in addition to aligned point-to-point distances. For simplicity, and because modeling the complexities of cortical between versus within-area distances was beyond the scope of this work, feedforward spatial costs in these models only take in to account the aligned point-to-point distances.

The results of our analyses, shown in Figure 10, revealed that the simplest, feedforward variant exhibited both significantly reduced performance and significantly greater between-area feedforward spatial costs. Intriguingly, the recurrent variant performed similarly to the main model, but exhibited intermediate between-area spatial costs. Increased spatial costs in the recurrent variant indicate that the lateral connections may be doing less work in this model relative to the feedforward connections. The greater spatial costs and reduced performance in the feedforward variant indicate that the lateral connections play an important functional role in networks with sign-constrained feedforward connectivity, allowing for functional integration of information over a wider range of units, and weighted inhibition.

We next focused on the emergent domain-level specialization in each of these model variants, using the selectivity-ordered lesion approach applied in earlier analyses of the main model. These results are shown in Figure 10C. Both simplified variants evinced greater domain-level specialization than the baseline model. For 20% selectivity-ordered lesions, the recurrent variant exhibited a very large degree of specialization, with no significant damage to the non-preferred domains; 30% selectivity-ordered lesions revealed some degree of representational sharing between domains, with small but significant non-specific effects of damage from lesions to object-selective units on both face and scene tasks, and from scene-selective lesions on the object task. The feedforward variant revealed a very similar pattern of damage to the recurrent one, with one additional significant non-specific damage effect for 20%

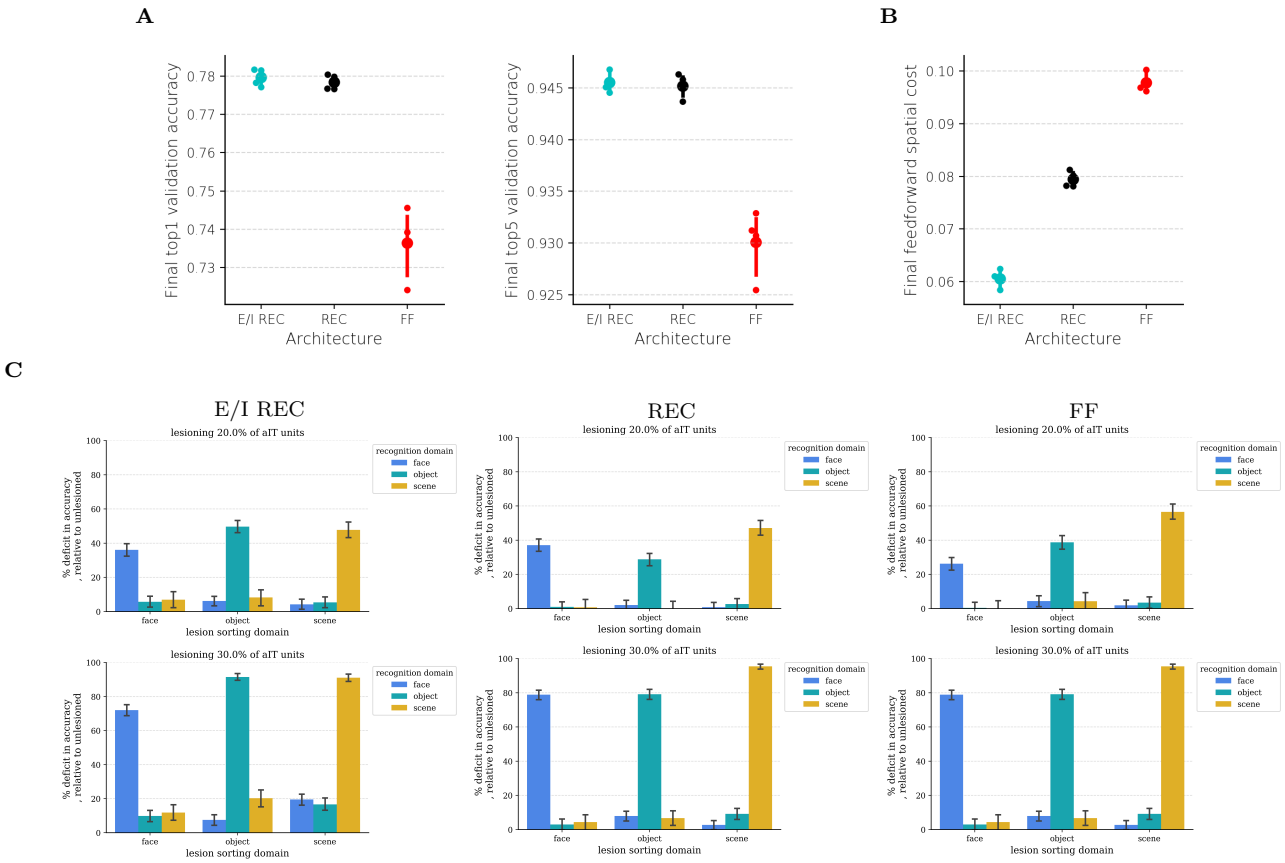



**Figure 10:** Comparing performance, spatial cost, and domain-level specialization of the full E/I recurrent model (E/I-EFF- local-RNN, or simply E/I REC) with two simpler architectures: 1) with excitatory-only feedforward connections but no separation of neurons responsible for excitation and (lateral) inhibition (EFF-local-RNN, or simply REC), and 2) a feedforward architecture employing local excitatory feedforward connections and layer normalization (EFF-local-FNN+LN, or simply FF). **A.** Performance. Left: final classification accuracy of 4 randomly initialized versions of each architecture. Right: final top5 classification accuracy for the same models (trial considered correct if correct prediction is in the top-5 highest output category activities). **B.**. Final feedforward spatial cost for the same models. **C.** Domain-level functional specialization of a single model of each variant measured using selectivity-ordered lesions.

lesions—namely an object-selective lesion on the face task. Thus, the degree of domain-level specialization is strong in each model, and is somewhat dependent on the particular model architecture. However, given that at least some more generalized damage was seen in each case before entirely obliterating performance on a given domain, we argue that these networks are still better described as highly topographically specialized rather than modular *per se*.

# 3 Discussion

Is IT cortex a collection of independent, possibly hard-wired domain-specific modules, or a more general-purpose, interactive, and plastic system? The investigations presented here demonstrate that many of the key findings thought to support a modular view of separable, innately-specified mechanisms for the recognition of different high-level domains (faces, objects, scenes) can be accounted for within a learning-based account operating under generic connectivity constraints (also see Arcaro et al., 2019). By simulating a biologically plausible neural-network model of IT without domain-specific innate structure, we found that we can "let the structure emerge" (McClelland et al., 1986, 2010). Specifically, we observed that our model developed strongly domain-selective spatial clusters which contain

preferential information for each domain, and which, when lesioned, produced largely (but not purely) specific deficits.

Beyond domain-level spatially clustered organization, we found that our model exhibited a more generic form of topographic organization, whereby nearby units had more correlated responses over images compared to more distant units, a relationship which has been demonstrated in macaque IT cortex (Majaj et al., 2015; Lee et al., 2020). Lee et al. (2020) explicitly fit a DCNN to produce the distance-dependent response correlation seen in IT. These authors devised a loss function that directly encouraged units to obey the pairwise response correlation as a function of distance found in macaque IT, using spatially embedded DCNN units, arguing that this is a reasonable proxy for explicit wiring length minimization. They found that the model produced domain-level organization, focusing on the emergence of face-selective clusters despite the model being trained on mostly objects and animals, along with one face class. Our work constitutes a significant advance beyond that of Lee et al. (2020) in that it uses explicit wiring minimization—rather than a proxy loss term serving to place correlated, but unconnected, units nearby—to develop topographic organization. Intriguingly, our model recapitulated the distance-dependent response correlation pattern that was used to generate their loss term, which is both a strength of our model and an indication that their proxy loss term was a reasonable choice. Thus, our work provides mechanistic insights into how this generic functional organization may develop.

An additional difference between our work and that of Lee et al. (2020) is that they argued that face selectivity could emerge solely from learning faces as one of many object categories, whereas we assumed that faces were a substantial domain of recognition requiring individuation over many different face identities. Prior work (Blauch et al., 2017, 2021a) demonstrated that the features learned through object recognition provided a basic but relatively poor foundation from which to perform face individuation. Thus, it is intriguing to note that whereas face selectivity might emerge without substantial experience at the individual level or any specialized mechanisms (see also Cowell and Cottrell, 2013), such extensive experience appears to be necessary to induce a representational space capable of supporting rapid, expert-level individual face recognition (see also Young and Burton, 2018; Sunday and Gauthier, 2018; Blauch et al., 2021b; Yovel and Abudarham, 2021; Young, 2021). The fact that a DCNN model trained without individual-level face experience is able to reproduce many of the hallmarks of face selectivity in primate cortex (Lee et al., 2020), and may provide a superior fit to one trained only with individual-level face experience (Chang et al., 2020), suggests that non-human primates may not be the optimal model system for studying the neural basis of human expert individual face recognition (Rossion and Taubert, 2017; Rossion, 2019). Given that many aspects of the non-human primate visual system make it an excellent model system for studying visual functional organization more generally, perhaps caution is warranted when human and non-human primates do not perform a given task with equivalent proficiency.

Beyond domain-level and generic distance-dependent organization in our model, we discovered interpretable, within-domain organization as well: objects were topographically organized into subclusters of inanimate objects, non-mammalian animals, and mammals; scenes were topographically organized into subclusters of indoor scenes, and natural and manmade outdoor scenes; and faces were topographically organized by gender and hair color. These results are reminiscent of Sato et al. (2013), who demonstrated a mosaic-like organization of domain and within-domain topographic organization in IT cortex. We expect that even finer-grained, more nuanced organization is present in the ITN, as in a Self-Organizing Map (SOM), in which the most salient representational distinctions exhibit global topography and more specific distinctions are repeated within it (Kohonen, 1982). Uncovering these more subtle aspects of organization and relating them more directly to empirical observations in biological systems is an obvious and important direction for future research.

We demonstrated how several key biological details—spatially local recurrent connectivity, the principality of excitation, and the separation of excitation and inhibition—contributed to topographic organization in task-optimized deep neural networks. Notably, in the current task-optimization framework applied to image recognition, spatially local feedforward and recurrent connectivity on its own was insufficient to drive topographic organization. There could be several reasons for this, but the simplest is that the combination of task and architecture did not impose sufficient demands on the lateral connectivity for the local connectivity constraint to drive the functional organization of the system towards local smoothness. Lateral connections in non-convolutional layers do not appear to provide much functional benefit to image recognition (except under constraints such as connectivity sign-restriction, see Figure 10). This is likely because increasing the depth of the hierarchy of fully-connected layers also does not increase performance, and these layers with lateral connections can be seen as temporally unrolled versions of a deeper hierarchy. However, lateral connections have been demonstrated to provide additional functional power to convolutional layers (Nayebi et al., 2018; Kubilius et al., 2018), where increasing depth does typically come with superior performance. Further, other tasks such as segmentation may place greater demands on lateral connections (Linsley et al., 2018, 2020). Thus, further exploration of recurrent architectures and tasks is needed to determine whether local constraints on their own would be sufficient to drive topographic functional organization without the need for sign-based constraints. However, there is also reason to suspect that the sign-based constraints are important.

In particular, local strictly excitatory feedforward connectivity essentially ensures that units in a local neighborhood are positively correlated (note that this is not the case for local feedforward connectivity unconstrained in sign). Since the network is constrained to perform multiple tasks, all units cannot be positively correlated to reach high performance, so competitive and cooperative dynamics through learning leads to local units learning correlated representations and more distant units learning uncorrelated or anti-correlated representations. Additionally, we found that the separation of excitation and inhibition contributed to topographic organization, but less so than the excitatory restriction on feedforward connectivity. We hypothesize that the separation of excitation and inhibition serves to enhance topographic organization by inducing greater pressure on the lateral connections. With separated excitation and inhibition, learned inhibitory responses must be mediated through lateral connections from E to I cells and back to E cells. When this connectivity is constrained to be local, it creates a pressure for local communication to be functionally smooth, so that related aspects of the representational space can communicate with each other. Consistent with this hypothesis, we found that the recurrent model variant had a smaller feedforward cost than the corresponding version without E/I separation, indicating that the former learned to place more functional weight on the lateral connections, thereby being able to reduce the burden of feedforward connections.

Our constraint removal analysis allowed us to discover an intriguingly simple model capable of producing many of the hallmarks of topographic organization in the main model. This feedforward variant contained local excitatory feedforward connections and no learned lateral connectivity, with lateral communication restricted to the layer normalization operation. We reason that this model was capable of producing topographic organization in a way similar to the classical Self-Organizing Map (SOM) (Kohonen, 1982), and to other algorithms applied to early visual cortex topographic organization (von der Malsburg, 1973; Miller et al., 1989). Each of these algorithms implements a form of local cooperation alongside broader competition. Specifically, in the SOM, global competition is implemented by selecting a winning unit on each trial, and suppressing the responses of all other units, and local cooperation is mediated through Hebbian learning scaled by a Gaussian neighborhood around the winning unit. In our feedforward variant, the local excitatory feedforward connections appear to implement the local cooperation, ensuring that neighboring units are positively correlated; the layer normalization operation then implements a global

competition. Specifically, layer normalization attempts to convert the distribution of pre-activations to a standard normal distribution, such that for a bias of 0, approximately half of the units can transmit positive activation through the positive rectification activation function (in practice, however, the distribution of pre-activities is often right-skewed, such that there are more normalized pre-activities with value less than 0, resulting in sparser activity). Moreover, standardization of layer activities ensures that units represent different aspects of the feature space. Thus, layer normalization implements both competition and interactivity that, when combined with the local representational cooperation induced by local excitatory feedforward connections, leads to a smooth topographic organization whereby the unit feature tuning is systematically more similar for nearby units than for farther units.

There are several advantages to the feedforward and other ITN variants, relative to SOMs and other previous topographic mapping algorithms. The first is that the ITN approach is naturally hierarchical, allowing for multiple interacting levels of topographically organized representations, rather than assuming a single feature space to be arranged in a single topographic map. This allows the ITN to explain the presence of multiple domain-selective regions arranged in a stream from earlier to later parts of IT (Kanwisher et al., 1997; Gauthier et al., 2000a; Kriegeskorte et al., 2007; Grill-Spector et al., 2017). Secondly, the ITN is task-optimized, allowing it to be naturally integrated into deep learning architectures without separate Hebbian updates. Of course, this second point could also be seen as a drawback, as Hebbian learning is more plausible than back-propagation of error. There are, however, several promising ongoing efforts to implement scalable biologically-plausible alternatives to back-propagation (Guerguiev et al., 2017; Lillicrap et al., 2020; Richards et al., 2019), so back-propagation might be considered an interim proxy for a different algorithm. Lastly, the ITN framework is extremely flexible, allowing for future research to examine different encoders, different IT architectures and topologies, and different task training environments and readout mechanisms. This makes the ITN an attractive approach for future research examining topographic organization in the visual system.

One limitation of our current work is that it only addresses the topographic organization of high-level representations, since the aforementioned constraints were not applied within the convolutional layers of the encoder network that was used to model early and mid-level vision. In the current deep learning paradigm, convolutions are a crucial aspect of achieving good task performance, whereas local connectivity suffers from a relative lack of inductive bias, thereby requiring more parameters and longer training time to learn the same features (but see Vaswani et al., 2017; Jaegle et al., 2021, for evidence that multiplicative interactions in fully connected visual "Transformer" layers may scale well to visual tasks). Developing a biologically plausible account of functional organization in early visual cortex is a crucial next step which would bring deep neural network models in closer alignment with the primate visual system. In future work, we plan to apply the same constraints implemented here to the feature channels in convolutional layers. We expect that this will induce topographic organization over earlier visual representations, likely encompassing the distance-dependent response correlation (Lee et al., 2020), without necessarily leading to domain-level specialization in each layer. However, a more pertinent line of future research is to jointly model organization over both retinotopic and featural dimensions. Doing so is challenging within the ITN framework, since jointly organizing retinotopic and featural dimensions through constraints on connectivity will require abandoning convolutional architectures. This is because the spatial constraints on connectivity in a system in which both retinotopic location and feature channel are modeled would necessitate that each channel have potentially different connections with other channels across different retinotopic positions. Abandoning convolution is conceptually straightforward but would likely result in extended training times and much poorer performing models, which may limit their ability to recapitulate the task-dependent functional organization of IT. However, convolution is merely an efficient implementation of learning

similar features at different locations in visual cortex and, more importantly, feature tuning is not actually uniform across the visual field (Silson et al., 2015). Thus, relaxing the convolution assumption and undertaking a more detailed connectivity-based topographic modeling study of early and mid-level visual areas is an important line of work that deserves future attention.

The work presented here takes several important steps to model both quantitatively and qualitatively the factors underlying visual cortical development throughout the visual hierarchy. Here, we have focused on constraints local to the IT circuit. However, a currently unexplored question in our framework is why and how regions emerge in consistent locations across members of a given primate species (Behrmann and Plaut, 2013; Kanwisher, 2010; Grill-Spector et al., 2017; Srihasam et al., 2014; Arcaro and Livingstone, 2017). We expect that modeling long-range connectivity-based constraints with regions external to IT (e.g., Plaut, 2002; Plaut and Behrmann, 2011), along with adapting the ITN architecture to contain two hemispheres which communicate via a model corpus callosum, will allow for the location of model cortical areas to be anchored to their connectivity with upstream and downstream areas. In particular, the retinotopic organization of upstream early visual cortical areas is thought to encourage faces towards foveally-biased cortex, and scenes towards peripherally-biased cortex (Levy et al., 2001; Plaut and Behrmann, 2011). Moreover, innate connectivity biases with downstream nonvisual areas is thought to play a further role in shaping the global organization of domain-selective areas in IT (Price and Devlin, 2011; Saygin et al., 2012, 2016; Plaut and Behrmann, 2011; Mahon and Caramazza, 2011; Powell et al., 2018). These biases, such as left-hemispheric language biases, and other more fine-grained patterning of connections with domain-selective downstream areas (i.e., socially-responsive areas for faces, memory areas for scenes, motor areas for manipulable objects) should be explored in future work to better understand IT organization both within and between hemispheres. We fully expect graded connectivity to bias the resulting locations of domain-selective regions successfully. However, based on the results here, we argue that such long-range connectivity is not a necessary condition for topographic domain-selectivity; rather, the pressure for low wiring cost solutions to hierarchical visual computation within IT appears to be sufficient to drive such organization. While such constraints may also not be necessary once domain-preferential long-range biases are modeled, it seems clear that they are present, and they are much simpler to model within visual computational neuroscience than long-range connectivity biases with external areas and circuits. An extension of the Interactive Topographic Network framework to modeling the interaction of inferotemporal cortex with early visual representations and other cognitive capacities should provide interesting and important insights into these issues.

# 4 Materials and Methods

## 4.1 The Interactive Topographic Network

Here we introduce the Interactive Topographic Network (ITN), a framework for computational modeling of high-level visual cortex, under specific biological constraints and in the service of specific task demands. ITN operates according to a set of principles which build upon previous work (Plaut and Behrmann, 2011), and can be divided into two components: an *encoder* that approximates early visual cortex, and *interactive topography* (IT) layers that approximate inferotemporal cortex. The goal of the encoder is to extract general visual features which describe the visual world along dimensions that support a broad range of downstream readout tasks. However, our main modeling focus is on IT, which consists of a series of pairs of recurrent layers that are subject to biological constraints. For computational simplicity, such constraints are not modeled in the encoder, although future work that incorporated similar constraints could be used to model topographic organization throughout the visual hierarchy.

## 4.2 Encoder architecture and training

We used a ResNet-50 (He et al., 2016) encoder to allow the ITN to extract deep and predictive features of the trained inputs. The encoder is pre-trained on equal sized subsets of faces, objects, and scenes from the VGGFace2 (Cao et al., 2018), ImageNet (Deng et al., 2009), and Places365 (Zhou et al., 2018) datasets, respectively, matched in terms of total training images. We reused the same subsets of faces and objects as in (Blauch et al., 2021a), and an additional scene domain was constructed to match the other two domains in total images. An initial learning rate of 0.01 was used, and this learning rate was decayed 5 times by a factor of 10 upon plateau of the validation error; after the 5th learning rate decay, the next validation error plateau determined the end of training. Stochastic gradient descent with momentum ($\rho = 0.9$) and $l2$ weight decay ($\lambda = 0.0001$) was used, with batch size of 256 on a single GPU.

## 4.3 Recurrent neural network formulation of IT

Our model of IT extends the standard discrete-time Recurrent Neural Network (RNN) formulation common in computational neuroscience (e.g., Song et al., 2016). We begin with the continuous-time dynamics of units in an RNN layer, where $\boldsymbol{x}^{(a)}$ is the vector of pre-activation activities in area $a$ of IT, $\boldsymbol{r}^{(a)}$ is the vector of post-activation activities in area $a$, $\boldsymbol{b}^{(a)}$ is the vector of baseline activities in area $a$, $\tau$ is the scalar neuronal time constant, and $W^{(a,b)}$ is the matrix of weights from area $a$ to area $b$:

$$\tau \frac{d\boldsymbol{x}_t^{(a)}}{dt} = -\boldsymbol{x}_t^{(a)} + W^{(a,a)}\boldsymbol{r}_t^{(a)} + W^{(a-1,a)}\boldsymbol{r}_t^{(a-1)} + \boldsymbol{b}^{(a)} \tag{1}$$

where the activation function $\boldsymbol{r}_t^{(a)} = [\boldsymbol{x}_t^{(a)}]_+$ is positive rectification, also called a Rectified Linear Unit (ReLU). Applying the Euler method to integrate this first-order ordinary differential equation, with time step size $\Delta t$, and substituting $\alpha = \frac{\Delta t}{\tau}$, yields the discrete time update:

$$\boldsymbol{x}_t^{(a)} = (1-\alpha)\boldsymbol{x}_{t-1}^{(a)} + \alpha\left(W^{(a,a)}\boldsymbol{r}_{t-1}^{(a)} + W^{(a-1,a)}\boldsymbol{r}_{t-1}^{(a-1)} + \boldsymbol{b}^{(a)}\right) \tag{2}$$

Note that this formulation differs from the standard machine learning implementation of RNNs, which can be derived as a special case where $\Delta t = \tau$ or $\alpha = 1$, in which the time constant is set such that the previous activity of a neuron decays exactly to zero in the time between updates, such that it can be set to 0 in the update equation.

When training models with separate excitatory and inhibitory units, we noted that training could be extremely unstable and required some mechanism for achieving stability. To this end, we adopted layer normalization (Ba et al., 2016), without the trainable scaling parameter that is sometimes used (see Ba et al., 2016, for more details). Where $\mu(\boldsymbol{x})$ is the mean of $\boldsymbol{x}$, and $\sigma(\boldsymbol{x})$ is the standard deviation of $\boldsymbol{x}$, and $\boldsymbol{b}$ is the learned bias term (moved outside of the layer normalization), the layer-normalized activities are given as:

$$\boldsymbol{z}_t = \frac{\boldsymbol{x}_t - \mu(\boldsymbol{x_t})}{\sigma(\boldsymbol{x_t})} + \boldsymbol{b}$$
$$\boldsymbol{r}_t = [\boldsymbol{z}_t]_+$$

Incorporating layer normalization into our update equation yields the final update equation:

$$\boldsymbol{x}_t^{(a)} = (1-\alpha)\boldsymbol{z}_{t-1}^{(a)} + \alpha\left(W^{(a,a)}\boldsymbol{r}_{t-1}^{(a)} + W^{(a-1,a)}\boldsymbol{r}_{t-1}^{(a-1)}\right) \tag{3}$$

## 4.4 Extending the standard RNN framework with biological constraints

Here, we outline the major biological constraints implemented in this work.

#### 4.4.1 Spatial organization

An essential aspect of an ITN model is that each IT layer has a spatial organization. We chose to model layers as square grids, with each layer of the hierarchy of equal size (typically, a grid size length of 32, corresponding to a layer of 1024 units). We normalize the coordinates to lie in the range [0,1]. Each unit thus has a unique $(x, y)$ coordinate which will be used to determine the distance-dependent network topology. In general, the specific choices about map spatial arrangement are not critical to the predictions of the model, but they can potentially be manipulated in certain ways in the service of other theoretical goals.

#### 4.4.2 Spatial connectivity costs

We impose distance-dependent constraints on connectivity through a cost on longer connections throughout training. This basic formulation of the loss was introduced by (Jacobs and Jordan, 1992) as a way to induce spatially organized task specialization, and was shown to do so in a simple neural network model trained on small-scale tasks. To our knowledge, no other research has examined this loss in modern deep learning architectures trained on natural images. We use a simple modification of the original loss function, using the squared Euclidean distance $(\mathcal{D}_{i,j})^2 = ||\boldsymbol{r}_i - \boldsymbol{r_j}||_2^2$ (in place of $(\mathcal{D}_{i,j})^{10} = ||\boldsymbol{r}_i - \boldsymbol{r_j}||_{10}^{10}$ distance used by Jacobs and Jordan (1992)). By using the squared distance, we penalize longer connections disproportionally compared to shorter connections. The spatial loss on connections between areas $a$ and $b$, $\mathcal{L}_w^{(a,b)}$, is given by:

$$\mathcal{L}_w^{(a,b)} = \sum_{i,j} \frac{\left(\mathcal{D}_{ij}^{(a,b)}\right)^2 \left(W_{ij}^{(a,b)}\right)^2}{1 + \left(W_{ij}^{(a,b)}\right)^2} \tag{4}$$

The total spatial loss is the sum of the area-to-area spatial losses $\mathcal{L}_w = \sum_{a,b} \mathcal{L}_w^{(a,b)}$, and is added to the task-based loss as $\mathcal{L} = \mathcal{L}_t + \lambda_w \mathcal{L}_w$, on which gradient descent is performed.

#### 4.4.3 Connection noise

To approximate axon-specific variability in instantaneous firing rate (Cipollini and Cottrell, 2013), We apply multiplicative noise on the individual connections between neurons which is uniform over distance and layers. In practice, we find that connection noise helps to regularize the activations in the network, encouraging a more distributed representation. Noise is sampled independently from a Gaussian distribution $\mathcal{N}$ centered at 0 with variance $\sigma^2$ at each time step of each trial, and is squashed by a sigmoidal function $S(x) = \frac{2}{1+e^{-x}}$, ensuring that the sign of each weight is not changed and each magnitude does not change by more than 100%. Thus, the noisy weight matrix $W_n^{(a,b)}$ from area $a$ to area $b$ on a given trial and time step is:

$$W_n^{(a,b)} = S\left(\mathcal{N}(0, \sigma)\right) * W^{(a,b)} \tag{5}$$

#### 4.4.4 Sign-based restrictions on neuronal connectivity

As has been discussed in prior computational work (Song et al., 2016), standard neural networks gloss over a key detail of neuronal morphology—that single neurons obey Dale's Law, whereby all of their outgoing connections are either excitatory or inhibitory (ignoring modulatory neurons and other special, rare cases). We employ this principle within our framework by replacing the single sheet of unconstrained neurons with parallel sheets of excitatory (E) and inhibitory (I) neurons. We follow the method of (Song et al., 2016) to enforce the sign of connectivity in our

neuronal populations. The second sign-based restriction we implement is that between-area interactions are carried out predominantly by excitatory pyramidal neurons. Thus, we restrict between-area feedforward connectivity to originate from the excitatory neurons only. Both E and I neurons receive feedforward inputs, which in the case of the I neurons is often misleadingly referred to as feedforward inhibition.

## 4.5 Task demands

Extensive work in computational systems and cognitive neuroscience has demonstrated that task-driven computational models learn representations that better account for neural responses in visual cortex than models which are designed by hand (Yamins et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Yamins and DiCarlo, 2016). More recent work has shown that a supervised version of task-driven learning is not essential to superior prediction of neural data, with semi-supervised contrastive learning algorithms performing very close to the supervised state-of-the-art (Zhuang et al., 2020). For simplicity and as a first step, we use supervised learning: Given a set of categories, the network is tasked with classifying images into one of these categories. The cross-entropy loss is used as an objective function, optimized using stochastic gradient descent with weight decay and momentum. However, we fully expect our general findings to hold for any learning algorithm that places comparable demands on high-level representation learning.

Visual systems often have to perform multiple tasks, where the specific organization of tasks is not known ahead of time. Therefore, a pre-specified task segregation (c.f. Dobs et al., 2019) is not possible. By specifying only a domain-general architecture, and optimizing to performance over multiple tasks, ITN models can discover the task-specific organization that maximizes task performance in the context of its task-general architectural constraints. Here, we simulate the requirement of performing three somewhat different visual tasks from common inputs, using common resources. The first task is face identification, for which we use VGGFace2 (Cao et al., 2018). The second task is object and animal recognition (hereafter just referred to as objects), for which we use the ImageNet dataset Deng et al. (2009). The third task is scene recognition, for which we use the Place365 dataset (Zhou et al., 2018). These three tasks constitute three separate "domains," each containing several categories which the network must learn to discriminate between. However, the network has no prior knowledge of the separability of these domains, and they are fully interleaved during training.

## 4.6 IT architecture and training

The main ITN model consists of 3 IT layers (pIT, cIT, aIT) with separate E and I populations, and feedforward connections sent only by E units. To facilitate training many models with fewer computational demands, the model is trained using a fixed pre-trained ResNet-50 encoder on smaller subsets of faces, objects, and scenes. Specifically, we created image subsets equal to the size of the popular CIFAR-100 dataset but at higher image resolution, containing 100 categories with 500 training images and 100 validation images, resized to 112x112 pixels. Thus, the combined dataset contained 300 categories with 150,000 training images and 30,000 validation images. The same learning rate schedule as used for training the encoder was used. Stochastic gradient descent with momentum ($\rho = 0.9$) was used, with batch size of 1024 on a single GPU. We used spatial regularization with $\lambda_w = 0.05$, without an additional weight decay penalty.

## 4.7 IT model variants

To better understand the factors that contribute to the development of topographic organization, we examine a variety of IT model variants containing different subsets of implemented constraints. Some of these models do not use

separate populations of E and I units, but still restrict feedforward connectivity to be excitatory. In this case, we simply restrict the feedforward weights to be positive, despite the same neuron having both positive and negative lateral connections. In another variant, we remove learned lateral connections entirely. This model is trained for a single time step, and the only recurrent computation is that of a single pass of layer normalization. Lastly, we explore a range of spatial regularization strengths.

## 4.8 Analyses of trained models

After training, the responses in IT layers are probed to investigate emergent task specialization and its topographic organization. We use three main approaches to this end.

### 4.8.1 Mass univariate analyses

The first analytic approach is the simple mass-univariate approach, in which each unit is analyzed separately for its mean response to each stimulus domain (objects, faces, scenes), using untrained validation images from the same categories used in training. In addition to computing the mean response to each domain, we compute *selectivity*, a ubiquitous metric used in neuroscience, to analyze how responsive a unit is to one domain compared to all others. We compare the responses of each domain versus the others using a two-tailed $t$-test, and given the test statistic $t$, the significance value $p$ of the test, and the sign of the test statistic $s = sign(t)$, we compute the selectivity as $-s\log(p)$.

### 4.8.2 Searchlight decoding analysis

The second analysis approach we use is the multivariate searchlight analysis commonly used in fMRI (Kriegeskorte et al., 2008), in which a pool of units are selected in a (circular) spatial window around each unit, and the accuracy for discriminating between different categories (e.g., hammer vs. screw-driver) in each domain (e.g., objects) is computed using the activations of only that pool of units; the mean accuracy value is assigned to the center unit, and the process is repeated for all units.

### 4.8.3 Lesion analysis

To assess the causal role of certain units in the performance of specific tasks, we adopt a lesioning approach in which the activities of lesioned units are set to 0 throughout perception. This effectively removes them from processing, allowing the network's dynamics to unfold independently of these units. The effect of a lesion is measured by computing the accuracy following the lesion and relating that to the baseline accuracy.

The first type of lesion we perform is a spatial or *focal* lesion in which a circular neighborhood of size $p * n$ units is selected, where $p$ is the fraction of units selected and $n$ is the total number of units in the area where the lesion is performed. The lesion is centered on a unit $u_{i,j}$ either randomly or according to the peak of a specific metric such as selectivity. In the main analyses, we attempt to lesion spatial neighborhoods corresponding to regions of high domain selectivity. To do so, we take the selectivity map, perform spatial smoothing, and select the unit $u$ of peak smoothed selectivity. We then systematically vary $p$ while keeping $u$ fixed to assess the causal role of increasingly large regions centered on the peak of smoothed selectivity.

The second type of lesion sorts units according to a given *selectivity* metric irrespective of their spatial location. In this analysis, the $p * n$ most selective units are chosen for a given lesion. This is done separately for the selectivity of each domain, as in the focal lesions. When the topography is smooth and the regions approximately circular, the

selectivity-ordered and focal lesions yield similar results. However, to the extent that the topography is not perfectly smooth or circular, the selectivity-ordered lesion may knock-out a more relevant set of units for a given task.

#### 4.8.4 Distance-dependent response correlation

We calculate the correlations of the responses of all pairs of units as a function of distance between them. Response correlation is computed for a given time step over a large number of images, either from all domains, or from each domain separately. We additionally compute a scalar metric of this analysis by taking the Spearman correlation of response correlation and distance. This metric can be easily visualized over many time steps, layers, cell types, models, etc.

#### 4.8.5 Analyzing spatial costs of trained networks

To understand the wiring cost of certain trained models, we analyze the spatial cost of a network, as given by Equation 4, as a function of architectural parameters such as the spatial regularization strength $\lambda_w$. In one analysis, we analyze only the feedforward spatial cost, which simply requires summing spatial costs over pairs of areas $a$ and $b$ where $a! = b$. Similarly, to analyze only the recurrent spatial cost, we can sum spatial cost over pairs of areas $a$ and $b$ where $a = b$.

#### 4.8.6 Unweighted spatial cost of sparsified networks

While wiring cost in an artificial neural network should depend to some extent on the strength of connections—stronger connections may require greater myelination, and strong connections in an artificial neural network may correspond to a larger number of connections in a biological neural network—there is another notion of wiring cost whereby it depends only on whether or not two neurons are connected. This notion of wiring costs has been commonly applied to the study of cortical areal layout and early visual cortical maps (e.g. Durbin and Mitchison, 1990; Mitchison, 1991; Koulakov and Chklovskii, 2001; Chklovskii, 2004). Moreover, the analysis of binary connectivity in thresholded networks is also common in graph-theoretic analysis of brain data (Bullmore and Sporns, 2009). To analyze this notion of wiring costs, we pruned our trained models to a desired connection sparsity level $s$, setting to 0 the $100 * (1 - s)\%$ connections with the smallest magnitude. Sparsity was enforced globally within IT and from IT to readout, rather than individually for each set of connections. We then analyzed an unweighted wiring cost $\mathcal{L}_{w,u}^{(a,b)}$ that computes the sum of squared Euclidean distance values between connected units $i$ and $j$ in areas $a$ and $b$, given that $(a, b)$ are in the set of connected areas $C$:

$$\mathcal{L}_{w,u}^{(a,b)} = \sum_{(a,b)\in C} \sum_{i,j} \left(\mathcal{D}_{ij}^{(a,b)}\right)^2 \left(W_{ij}^{(a,b)} \neq 0\right) \tag{6}$$


# Acknowledgments

We thank Michael J. Tarr, Leila Wehbe, Vlad Ayzenberg, Jacob Prince and the VisCog research group at Carnegie Mellon University for helpful discussions and comments.


# Competing interests

We declare no competing interests.

# References


Aguirre, G. K., Zarahn, E., and D'Esposito, M. (1998). An area within human ventral cortex sensitive to "Building" stimuli: Evidence and implications. *Neuron*, 21(2):373–383.

Alreja, A., Nemenman, I., and Rozell, C. (2020). Constrained brain volume in an efficient coding model explains the fraction of excitatory and inhibitory neurons in sensory cortices. Preprint, Neuroscience.

Arcaro, M. J. and Livingstone, M. S. (2017). A hierarchical , retinotopic proto- organization of the primate visual system at birth. *eLife*, pages 1–24.

Arcaro, M. J., Schade, P. F., and Livingstone, M. S. (2019). Universal Mechanisms and the Development of the Face Network: What You See Is What You Get. *Annual Review of Vision Science*, 5(1):341–372.

Arthur, D. and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. In *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer Normalization. *arXiv:1607.06450 [cs, stat]*.

Behrmann, M. and Plaut, D. C. (2013). Distributed circuits, not circumscribed centers, mediate visual recognition. *Trends in Cognitive Sciences*, 17(5):210–219.

Blauch, N., Aminoff, E., and Tarr, M. J. (2017). Functionally localized representations contain distributed information : Insight from simulations of deep convolutional neural networks. *Proceedings of the Cognitive Science Society*, 39.

Blauch, N. M., Behrmann, M., and Plaut, D. C. (2021a). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition*, 208.

Blauch, N. M., Behrmann, M., and Plaut, D. C. (2021b). Deep learning of shared perceptual representations for familiar and unfamiliar faces: Reply to commentaries. *Cognition*, 208.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

Carreiras, M., Seghier, M. L., Baquero, S., Estévez, A., Lozano, A., Devlin, J. T., and Price, C. J. (2009). An anatomical signature for literacy. *Nature*, 461(7266):983–986.

Chang, L., Egger, B., Vetter, T., and Tsao, D. Y. (2020). What computational model provides the best explanation of face representations in the primate brain? Preprint, Neuroscience.

Chklovskii, D. (2004). Synaptic Connectivity and Neuronal MorphologyTwo Sides of the Same Coin. *Neuron*, 43(5):609–617.

Cipollini, B. and Cottrell, G. (2013). Uniquely human developmental timing may drive cerebral lateralization and interhemispheric collaboration. In *Proceedings of the Cognitive Science Society, 35(35)*, pages 334–339.

Cowell, R. A. and Cottrell, G. W. (2013). What Evidence Supports Special Processing for Faces? A Cautionary Tale for fMRI Interpretation. *Journal of Cognitive Neuroscience*, 25(11):1777–1793.

Dehaene, S. and Cohen, L. (2007). Cultural Recycling of Cortical Maps. *Neuron*, 56(2):384–398.

Dehaene, S., Cohen, L., Morais, J., and Kolinsky, R. (2015). Illiterate to literate: behavioural and cerebral changes induced by reading acquisition. *Nature Reviews Neuroscience*, 16(4):234–244.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pages 248–255. IEEE.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73:415–434.

Dobs, K., Isik, L., Pantazis, D., and Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications*, 10(1).

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

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676):598–601.

Farah, M. J. (1990). *Visual Agnosia: Disorders of Object Recognition and What They Tell Us About Normal Vision*. MIT Press.

Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. The MIT Press, Cambridge, MA.

Freiwald, W. A. and Tsao, D. Y. (2010). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, 330(6005):845–851.

Gauthier, I., Skudlarski, P., Gore, J. C., and Anderson, A. W. (2000a). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*, 3(2):191–197.

Gauthier, I., Tarr, M. J., Moylan, J., Skudlarski, P., Gore, J. C., and Anderson, A. W. (2000b). The fusiform "face area" is part of a network that processes faces at the individual level. *Journal of cognitive neuroscience*, 12(3):495–504.

Geskin, J. and Behrmann, M. (2017). Congenital prosopagnosia without object agnosia? A literature review. *Cognitive Neuropsychology*, 3294:1–51.

Goodhill, G. J. (1993). Topography and ocular dominance: A model exploring positive correlations. *Biological Cybernetics*, 69:109–118.

Goodhill, G. J. and Willshaw, D. J. (1990). Application of the elastic net algorithm to the formation of ocular dominance stripes. *Network: Computation in Neural Systems*, 1(1):41–59.

Grill-Spector, K., Kushnir, T., Hendler, T., and Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature Neuroscience*, 3(8):837–843.

Grill-Spector, K., Weiner, K. S., Kay, K., and Gomez, J. (2017). The Functional Neuroanatomy of Human Face Perception. *Annual Review of Vision Science*, 3(1):163:197.

Guerguiev, J., Lillicrap, T. P., and Richards, B. A. (2017). Towards deep learning with segregated dendrites. *eLife*, 6.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778.

Hubel, D. H. and Wiesel, T. N. (1969). Anatomical Demonstration of Columns in the Monkey Striate Cortex. *Nature*, 221(5182):747–750.

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

Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., and Carreira, J. (2021). Perceiver: General perception with iterative attention.

Kanwisher, N. (2010). Functional specificity in the human brain: A window into the functional architecture of the mind. *Proceedings of the National Academy of Sciences of the United States of America*, 107(25):11163–11170.

Kanwisher, N., McDermott, J., and Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, 17(11):4302–4311.

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep Supervised , but Not Unsupervised , Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11).

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

Kohonen, T. (1995). *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Konkle, T. and Caramazza, A. (2013). Tripartite Organization of the Ventral Stream by Animacy and Object Size. *Journal of Neuroscience*, 33(25):10235–10242.

Konkle, T. and Oliva, A. (2012). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron*, 74(6):1114–1124.

Koulakov, A. A. and Chklovskii, D. B. (2001). Orientation Preference Patterns in Mammalian Visual Cortex: A Wire Length Minimization Approach. *Neuron*, 29:519–527.

Kriegeskorte, N., Formisano, E., Sorger, B., and Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 104(51):20600–5.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2(November):4.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., and Dicarlo, J. J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. *bioRxiv*, pages 1–9.

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

Levy, I., Hasson, U., Avidan, G., Hendler, T., and Malach, R. (2001). Center–periphery organization of human object areas. *Nature neuroscience*, 4(5):533–539.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346.

Linsker, R. (1986). From basic network principles to neural architecture: Emergence of orientation columns. *Proc. Natl. Acad. Sci. USA*, page 5.

Linsley, D., Ashok, A. K., Govindarajan, L. N., Liu, R., and Serre, T. (2020). Stable and expressive recurrent vision models. *arXiv:2005.11362 [cs]*.

Linsley, D., Kim, J., Veerabadran, V., Windolf, C., and Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pages 152–164.

Long, B., Yu, C.-P., and Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024.

Mahon, B. Z. and Caramazza, A. (2011). What drives the organization of object knowledge in the brain? *Trends in Cognitive Sciences*, 15(3):97–103.

Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, 35(39):13402–13418.

McCandliss, B. D., Cohen, L., and Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, 7(7):293–299.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., and Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8):348–356.

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1986). *Parallel Distributed Processing*, volume 2. MIT press Cambridge, MA.

Miikkulainen, R., Bednar, J. A., Choe, Y., and Sirosh, J. (2006). *Computational Maps in the Visual Cortex*.

Miller, K., Keller, J., and Stryker, M. (1989). Ocular dominance column development: Analysis and simulation. *Science*, 245(4918):605–615.

Mitchison, G. (1991). Neuronal branching patterns and the economy of cortical wiring. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 245(1313):151–158.

Moscovitch, M., Winocur, G., and Behrmann, M. (1997). What Is Special about Face Recognition? Nineteen Experiments on a Person with Visual Object Agnosia and Dyslexia but Normal Face Recognition. *Journal of Cognitive Neuroscience*, 9(5):555–604.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., Dicarlo, J. J., and Yamins, D. L. K. (2018). Task-Driven Convolutional Recurrent Models of the Visual System. *arXiv*, pages 1–14.

Obermayer, K., Ritter, H., and Schulten, K. (1990). A principle for the formation of the spatial structure of cortical feature maps. *Proc. Nad. Acad. Sci.*, 87:8345–8349.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609.

Olshausen, B. A. and Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325.

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology*, 19(7):603–639.

Plaut, D. C. and Behrmann, M. (2011). Complementary neural representations for faces and words: A computational exploration. *Cognitive Neuropsychology*, 28(3&4):251–275.

Powell, L. J., Kosakowski, H. L., and Saxe, R. (2018). Social Origins of Cortical Face Areas. *Trends in cognitive sciences*, 22(9):752–763.

Price, C. J. and Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in cognitive sciences*, 15(6):246–253.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

Rossion, B. (2019). What can we learn about human individual face recognition from experimental studies in monkeys? *Vision Research*, 157:142–158.

Rossion, B. and Taubert, J. (2017). Commentary : The Code for Facial Identity in the Primate Brain. *Frontiers in Human Neuroscience*.

Sato, T., Uchida, G., Lescroart, M. D., Kitazono, J., Okada, M., and Tanifuji, M. (2013). Object representation in inferior temporal cortex is organized hierarchically in a mosaic-like structure. *Journal of Neuroscience*, 33(42):16642–16656.

Saygin, Z. M., Osher, D. E., Koldewyn, K., Reynolds, G., Gabrieli, J. D., and Saxe, R. R. (2012). Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. *Nature Neuroscience*, 15(2):321–327.

Saygin, Z. M., Osher, D. E., Norton, E. S., Youssoufian, D. A., Beach, S. D., Feather, J., Gaab, N., Gabrieli, J. D., and Kanwisher, N. (2016). Connectivity precedes function in the development of the visual word form area. *Nature Neuroscience*, 19(9):1250–1255.

Silson, E. H., Chan, A. W.-Y., Reynolds, R. C., Kravitz, D. J., and Baker, C. I. (2015). A retinotopic basis for the division of high-level scene processing between lateral and ventral human occipitotemporal cortex. *Journal of Neuroscience*, 35(34):11921–11935.

Sirosh, J. and Miikkulainen, R. (1997). Topographic Receptive Fields and Patterned Lateral Interaction in a Self-Organizing Model of the Primary Visual Cortex. *Neural Computation*, 9(3):577–594.

Song, H. F., Yang, G. R., and Wang, X.-J. (2016). Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework. *PLOS Computational Biology*, 12(2):e1004792.

Srihasam, K., Vincent, J. L., and Livingstone, M. S. (2014). Novel domain formation reveals proto-architecture in inferotemporal cortex. 17(12):1776–1783.

Sunday, M. A. and Gauthier, I. (2018). Face Expertise for Unfamiliar Faces : A Commentary on Young and Burton's "Are We Face Experts ?". *Journal of Expertise*, x(x):1–7.

Swindale, N. V. (1996). The development of topography in the visual cortex: A review of models. *Network: Computation in Neural Systems*, 7(2):161–247.

Tanaka, K. (1996). Inferotemporal Cortex and Object Vision. *Annual Review of Neuroscience*, 19(1):109–139.

Tanaka, K. (2003). Columns for Complex Visual Object Features in the Inferotemporal Cortex : Clustering of Cells with Similar but Slightly Different Stimulus Selectivities. *Cerebral Cortex*, 13:90–99.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., and Tootell, R. B. H. (2003). Faces and objects in macaque cerebral cortex. *Nature Neuroscience*, 6(9):989–995.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., and Livingstone, M. S. (2006). A Cortical Region Consisting Entirely of Face-Selective Cells. *Science*, 311(5761):670–674.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 5999–6009.

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

Yamins, D. L. K. and DiCarlo, J. J. (2016). Eight open questions in the computational modeling of higher sensory cortex. *Current Opinion in Neurobiology*, 37:114–120.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24.

Young, A. W. (2021). Insights from computational models of face recognition: A reply to Blauch, Behrmann and Plaut. *Cognition*, 208.

Young, A. W. and Burton, A. M. (2018). Are We Face Experts? *Trends in Cognitive Sciences*, 22(2):100–110.

Yovel, G. and Abudarham, N. (2021). From concepts to percepts in human and machine face recognition: A reply to Blauch, Behrmann & Plaut. *Cognition*, 208:104424.

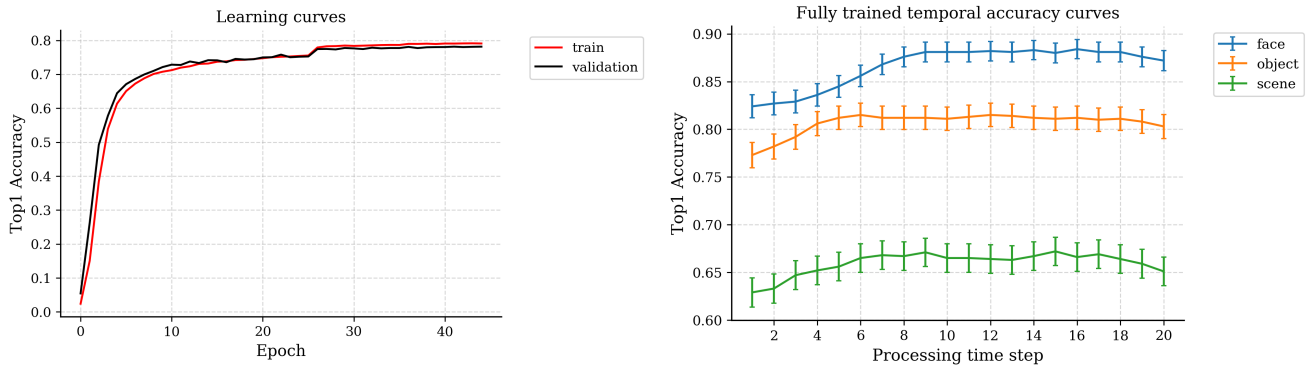
Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

Zhuang, C., She, T., Andonian, A., Mark, M. S., and Yamins, D. (2020). Unsupervised Learning from Video with Deep Neural Embeddings. *arXiv:1905.11954 [cs]*.

# 5 Supplementary Information

## 5.1 Final model performance

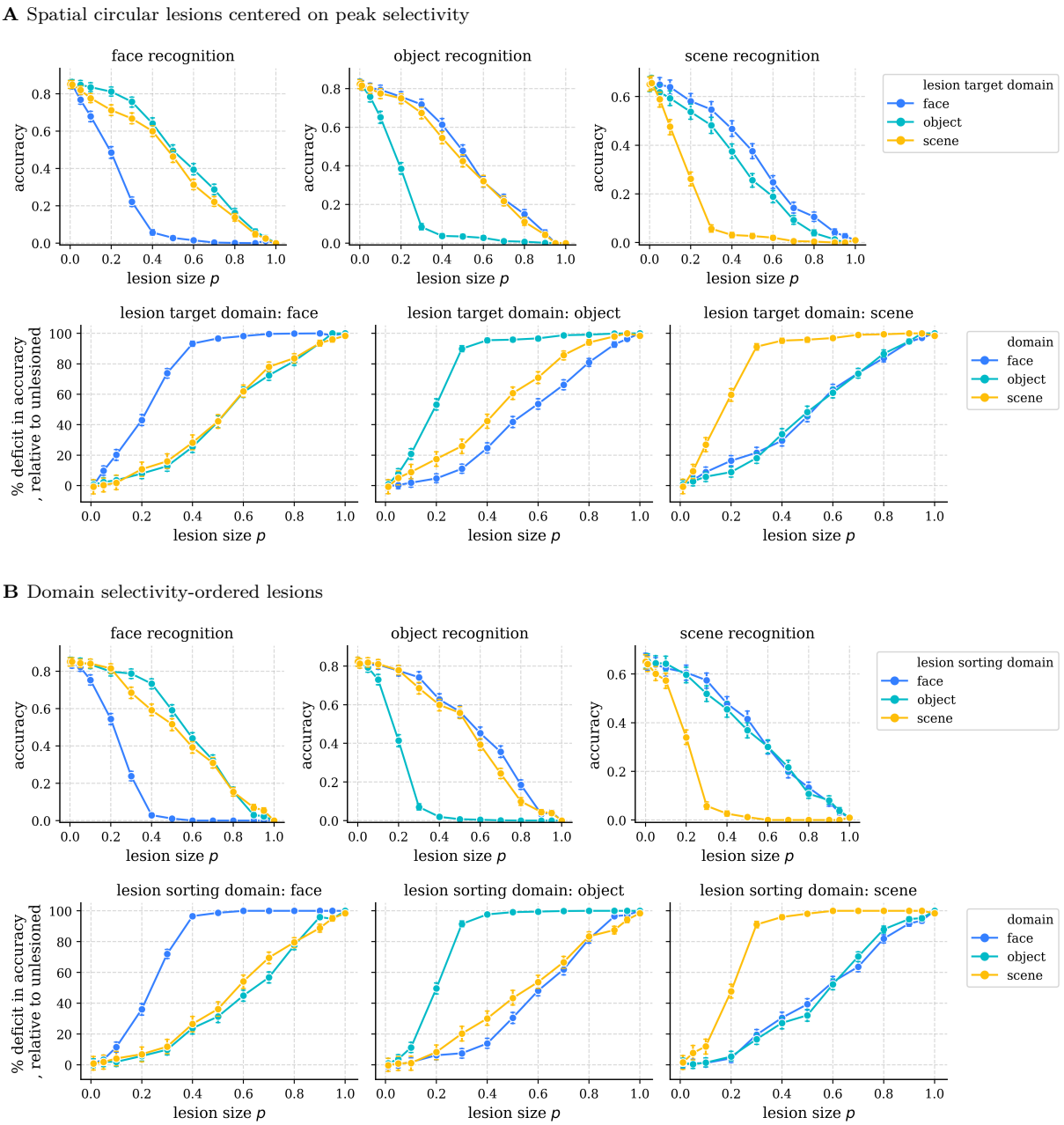
We measured overall recognition performance of the main ITN model throughout training. The results are plotted for training and validation images as a function of training epoch in Supplementary Figure S1. Additionally, we plotted the final accuracy for each domain as a function of processing time step in the model.




**Supplementary Figure S1:** Performance of the main model. **A.** Overall accuracy throughout the training of IT using a pre-trained encoder. **B.** Final accuracy for each domain, plotted over model processing time steps.

## 5.2 A more complete assessment of domain-selective functional organization

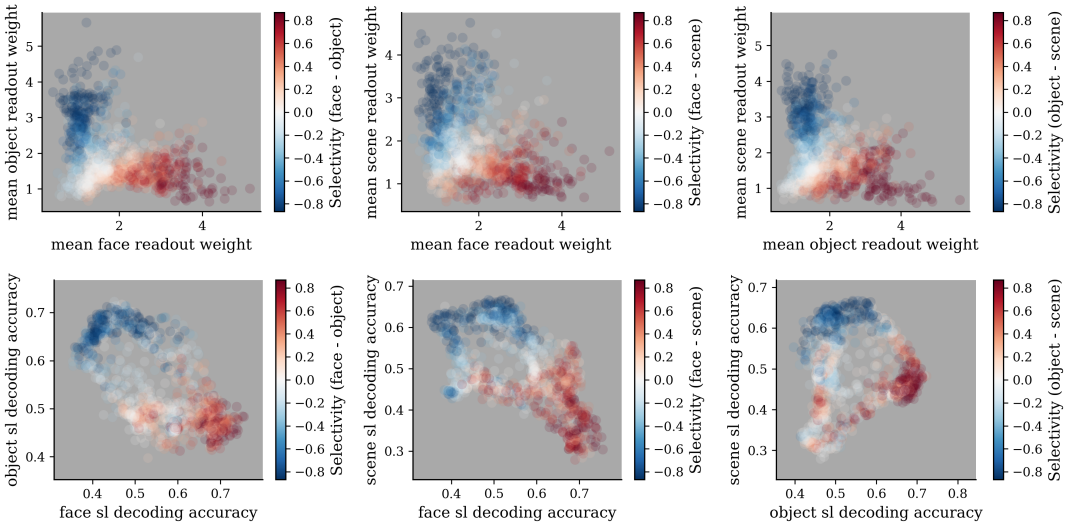
In the main paper, for clarity, we plotted lesion deficits at two specific lesion sizes. Here, for completeness, we plot both raw accuracy values and lesion deficits for a large range of lesions, shown in Supplementary Figure S2.




**Supplementary Figure S2:** Lesion results in the ITN model. **A.** Damage from various sizes of circular topographic lesions centered on the peak of smoothed selectivity for each domain. **B.** Damage from various sizes of non-topographic lesions chosen by sorting units according to their selectivity for each domain. As the selectivity is highly topographic, the lesion masks in the topographic vs. selectivity-ordered lesion plots are very similar; however, for selective regions that are non-circular (such as that seen for objects), the selectivity-ordered lesion provides a more precise way of assaying the selective region.

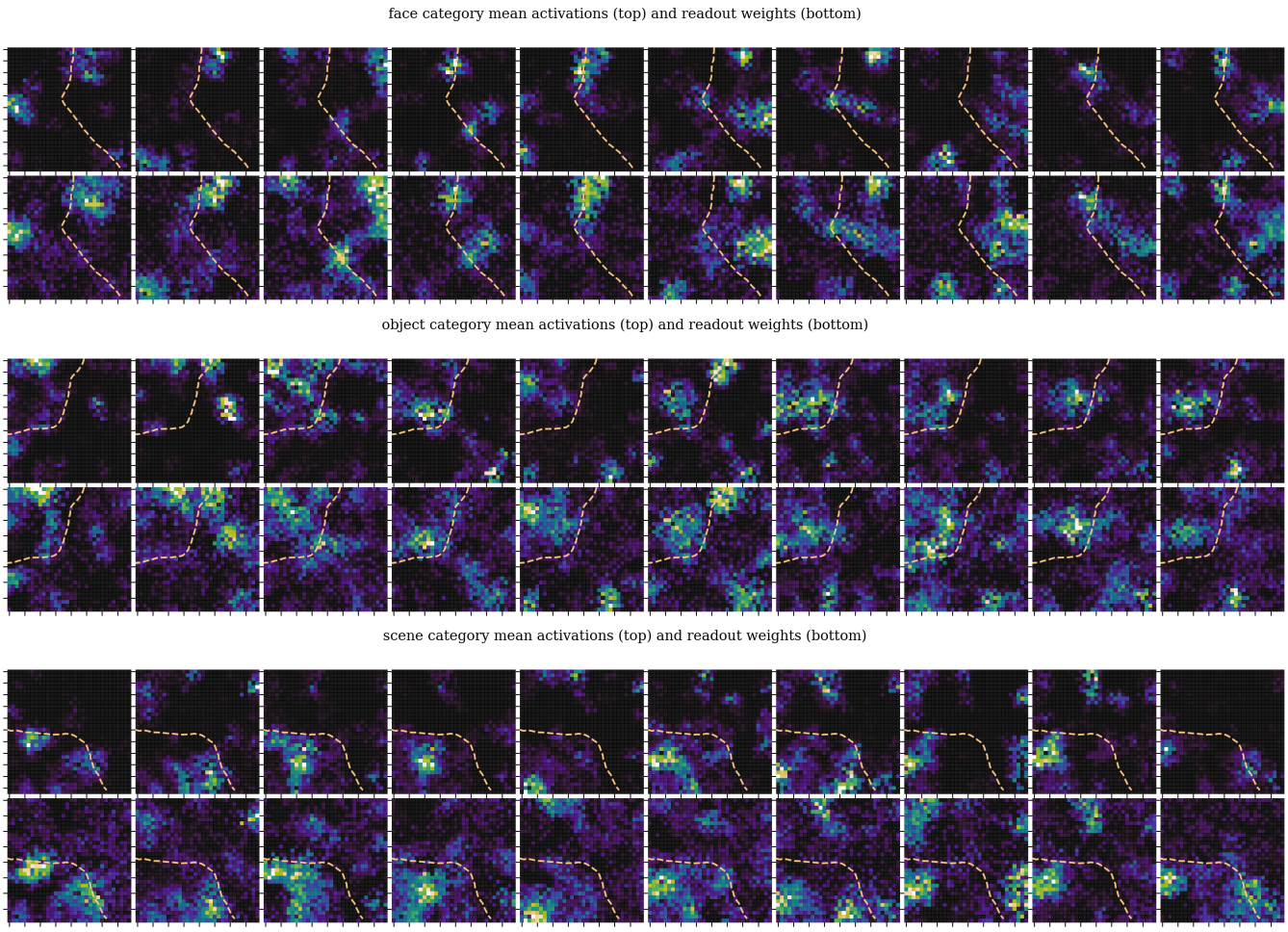
To better understand the degree of representational competition and cooperation, we compare searchlight accuracy and mean readout maps across domains. The results are plotted in Supplementary Figure S3 as scatter plots over units, colored by their selectivity between the two plotted domains, for all pairs of domain. The results corroborate our earlier findings, demonstrating a strong but graded degree of specialization. In particular, we note the large degree of correlation between searchlight accuracies for objects and for scenes. This result is somewhat surprising given the relatively specialized effects of lesions, demonstrating that object and scene information may co-mingle but

still be read out from different largely different sets of units in order to optimize task performance.




**Supplementary Figure S3:** Graded competition and cooperation between domains in aIT. Top: readout weights into each domain, bottom: searchlight accuracy for decoding within each domain.

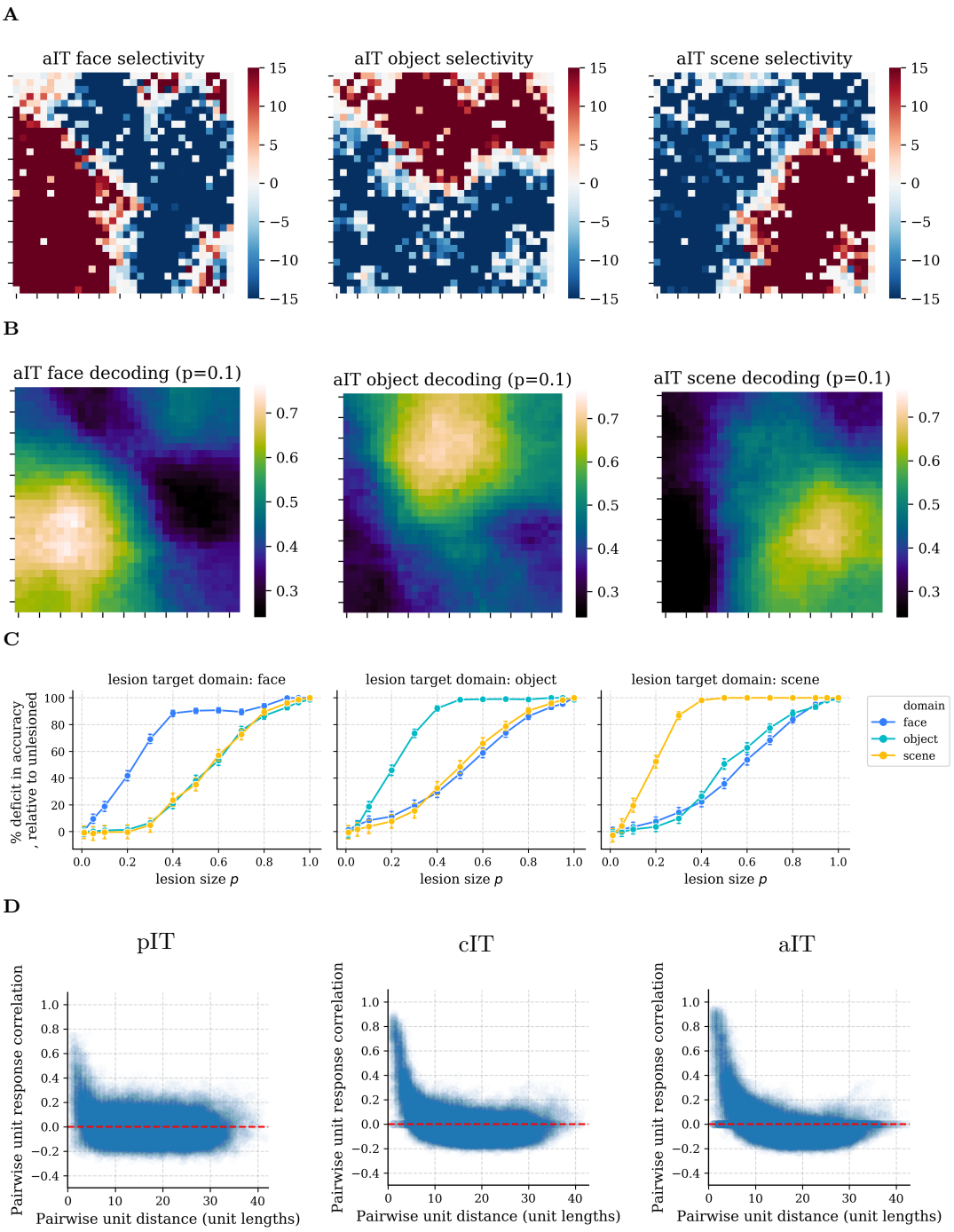
How consistent is the topographic responsiveness of categories within a domain? Within a given domain-selective area, are the weaker responses to non-preferred domains roughly uniform and weak across categories, or do some categories elicit notably strong responses? To answer this question, we plotted the mean aIT response and readout weights of 10 randomly selected categories from each domain, with the contour of significant smoothed domain-selectivity ($p<0.001$; smoothing performed as averaging selectivity over the 5% nearest units). These results are shown in Supplementary Figure S4.




**Supplementary Figure S4:** Category-level mean responses and readout weights for 10 example categories from each domain. For each domain, the top row shows mean category-level responses in aIT, and the bottom row shows the readout weights for the corresponding categories. Dashed yellow lines indicate the contour of statistically significant smoothed domain selectivity for the plotted category's domain ($p<0.001$; smoothing performed as averaging selectivity over the 5% nearest units).

## 5.3 Detailed results for the simplified feedforward model

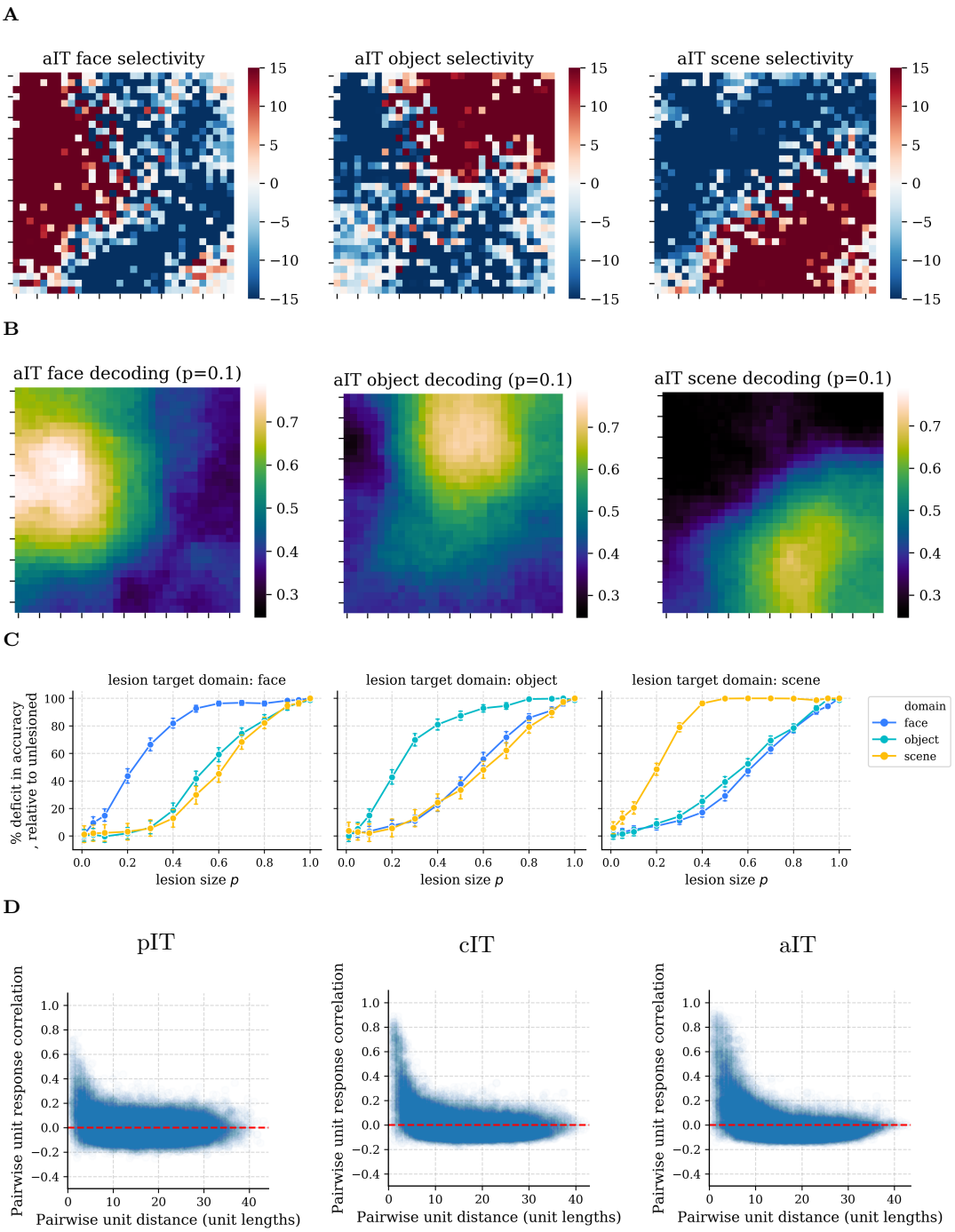
In the main paper, we presented a simplified "feedforward" model that utilizes excitatory feedforward connectivity and layer normalization. In Supplementary Figure S5, we plot a larger range of results for this model, including domain selectivity, within-domain information, lesion deficits, and generic topography.




**Supplementary Figure S5:** A simplified model containing local, excitatory feedforward connectivity and global lateral divisive inhibition (layer normalization) exhibits domain-level and generic topographic organization very similar to the more biologically-detailed model. See Figures 2 and 5 for details on the plots.

## 5.4 Detailed results for an intermediate model, with strictly excitatory feedforward connections but no separation of E and I

In the main paper, we presented a simplified "recurrent" model that utilizes excitatory feedforward connectivity, layer normalization, and recurrent lateral connectivity. In Supplementary Figure S6, we plot a larger range of results for this model, including domain selectivity, within-domain information, lesion deficits, and generic topography.

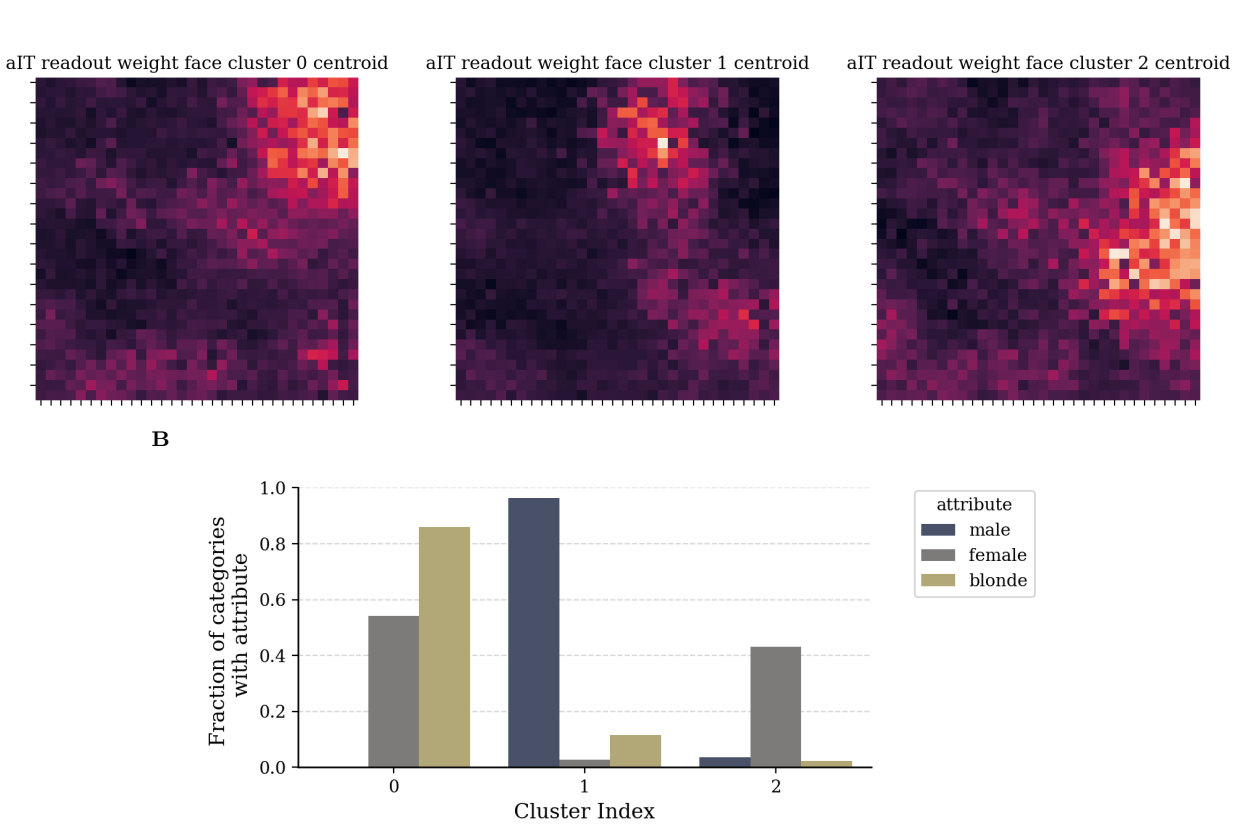



**Supplementary Figure S6:** A simplified model containing local, excitatory feedforward connectivity, local lateral connectivity (one sheet of units without sign constraints), and global lateral divisive inhibition (layer normalization) exhibits domain-level and generic topographic organization very similar to the more biologically-detailed model. See Figures 2 and 5 for details on the plots.

## 5.5 Clustering results for the face domain

We did not include the face domain readout weight clustering analysis in the main paper due to privacy concerns. In Supplementary Figure S7, we show a reduced analysis that preserves privacy, plotting only the cluster centroids and quantification of the attributes of biological sex and whether the modal hair color is blonde (inferred from a sample of 8 images per identity) to illustrate that within-domain clustering is seen for faces along with the other domains

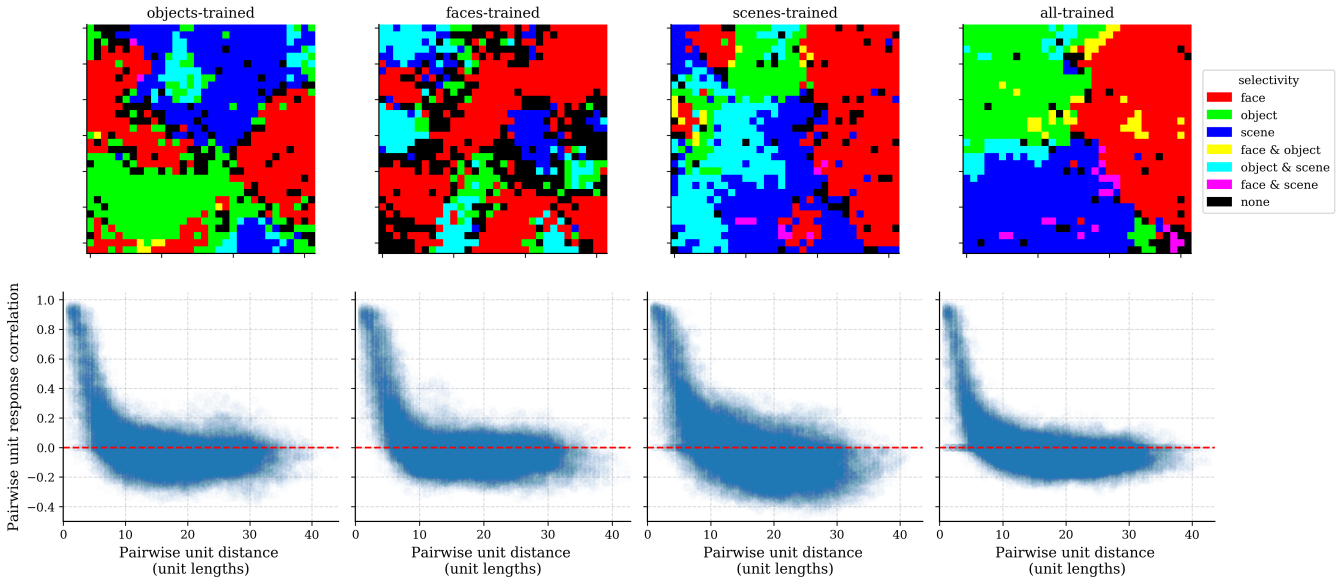
shown in the main paper. As with the other domains, further analysis of other possible attributes (e.g. ethnicity, age) is possible, but is beyond the scope of this work.




**Supplementary Figure S7:** Clustering face category readout weights from aIT. $k = 3$ clusters were used in a K-means++ clustering algorithm that clustered the vectors of aIT readout weights over all 100 face categories. **A.** The centroids of the clusters were then reshaped into the 2D aIT coordinates and visualized as heat maps. **B.** Quantification of biological sex over the members of each cluster, revealing quantitative characteristics of each cluster.

## 5.6 The impact of visual domain experience on emergent domain-level topography

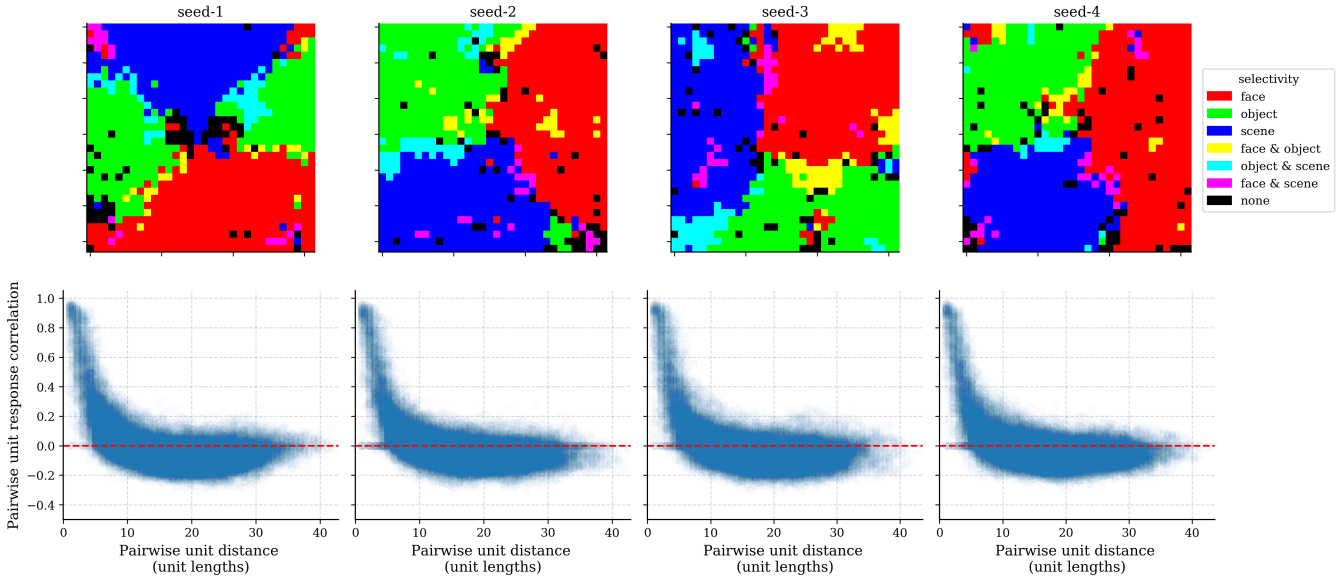
To assess the dependence of emergent topography on visual experience, we trained three additional ITN models with the E/I-EFF-local-RNN architecture, each trained on only one domain (objects, faces, or scenes), with all other parameters identical. The encoder was pre-trained with the same domain images used to train the encoder in main model, removing the images from the other domains. Similarly, IT was trained with the same domain images used to train IT in the main model, removing the images from the other domains. Domain-level topography and generic distance-dependent response correlations are plotted in Supplementary Figure S8. Somewhat smooth domain responses, along with an inverse relationship between pairwise unit distance and response correlation, are seen in each model. However, the model trained on all domains produces the cleanest global topographic organization, followed by object- and scene-trained models. The face-trained model appears to demonstrate the most specific topographic organization, biases towards faces, which makes sense in light of the more specialized data present in cropped (but naturalistic) face pictures compared to naturalistic photographs of objects and scenes.




**Supplementary Figure S8:** Topographic organization under different visual training experiences. The main model trained on objects, faces, and scenes is compared with models trained on each of the domains separately. As in the main model, a pre-trained encoder was used to learn the topography in the IT layers; in each case, the specific training environment was applied to both the encoder and IT components, using a larger dataset for pre-training the encoder.

## 5.7 Robustness of results

To verify the robustness of our results, we ran multiple versions of the main ITN model using different random seeds controlling the random weight initialization and training stimulus presentation order, both of which can in theory effect the emergent functional organization. We plot these results in Supplementary Figure S9.



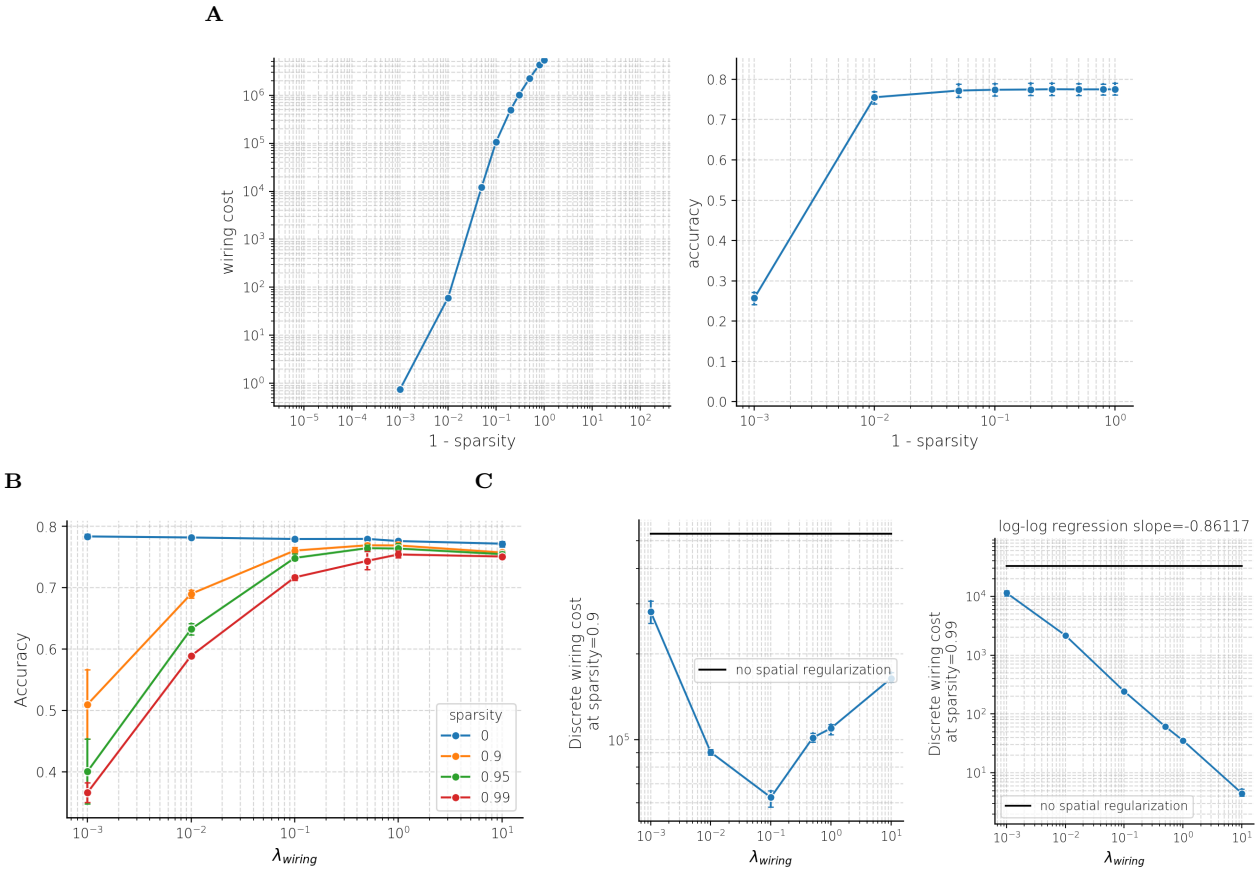

**Supplementary Figure S9:** Robustness of results. Top: domain-selective topography in aIT. Bottom: generic distance-dependent response correlations in aIT.

## 5.8 Further analysis of wiring costs on sparsified networks

In the main paper, we assessed an unweighted wiring cost $\mathcal{L}_{w,u}$ on sparsified ITN models, inspired by the wire minimization framework (Koulakov and Chklovskii, 2001; Chklovskii, 2004), as reproduced here:

$$\mathcal{L}_{w,u}^{(a,b)} = \sum_{(a,b)\in C} \sum_{i,j} \left(\mathcal{D}_{ij}^{(a,b)}\right)^2 \left(W_{ij}^{(a,b)} \neq 0\right)$$

To provide some basic insight into this approach, we first assessed the binary-connection wiring cost in the main ITN model over a range of sparsity values, shown in Supplementary Figure S10A. As expected, we found that increasing the sparsity of networks decreased the wiring cost. We next validated that sparsified models perform well, examining how accuracy relates to the level of induced sparsity, also shown in Supplementary Figure S10A. We found that accuracy was preserved well across a range of sparsity levels, even when only 1% of connections remained, but not when 0.1% of connections remained. We next examined the accuracy of models with a range of wiring cost penalties at a range of sparsity values (see Supplementary Figure S10B). For each non-zero sparsity value, we found that networks with a larger wiring cost penalty performed better. Lastly, we examined the wiring cost of these networks at two sparsity levels. We found that the wiring cost at a sparsity of 0.9 decreased up to a point, but then increased, indicating that increasing the wiring cost penalty did not strictly minimize the distance of the 10% strongest connections. In contrast, analyzing wiring cost at a sparsity of 0.99 indicated that increasing the wiring cost penalty up to a very large level consistently reduced the distances of the 1% strongest connections. These results may be reconciled by considering that networks with a high wiring penalty may learn to put the strongest weights on the shortest connections, but use intermediate weights for intermediate connection lengths, allowing for communication between somewhat distant units along with dominantly local computation.




**Supplementary Figure S10:** Analyzing wiring costs of binary connections. **A.** Wiring cost of binary connections, and accuracy, in sparsified versions of the main ITN model. **B.** Accuracy at a range of sparsity levels across a range of spatial regularization strength $\lambda_w$. **C.** Wiring cost at sparsity=0.9 and sparsity=0.99 for ITN models across a range of spatial regularization strength $\lambda_w$.