HHS Public Access

Author manuscript

*Neuron*. Author manuscript; available in PMC 2025 July 17.

Published in final edited form as:

*Neuron*. 2024 July 17; 112(14): 2435–2451.e7. doi:10.1016/j.neuron.2024.04.018.

# A Unifying Framework for Functional Organization in Early and Higher Ventral Visual Cortex

**Eshed Margalit**[1,*], **Hyodong Lee**[2], **Dawn Finzi**[3,4], **James J. DiCarlo**[2,5,6], **Kalanit Grill-Spector**[3,7,+], **Daniel L. K. Yamins**[3,4,7,+]

[1]Neurosciences Graduate Program, Stanford University, Stanford, CA 94305 USA

[2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

[3]Department of Psychology, Stanford University, Stanford, CA 94305 USA

[4]Department of Computer Science, Stanford University, Stanford, CA 94305 USA

[5]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

[6]Center for Brains Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

[7]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA 94305 USA

## Summary

A key feature of cortical systems is functional organization: the arrangement of functionally distinct neurons in characteristic spatial patterns. However, the principles underlying the emergence of functional organization in cortex are poorly understood. Here we develop the Topographic Deep Artificial Neural Network (TDANN), the first model to predict several aspects of the functional organization of multiple cortical areas in the primate visual system. We analyze the factors driving the TDANN's success and find that it balances two objectives: learning a task-general sensory representation and maximizing the spatial smoothness of responses according to a metric that scales with cortical surface area. In turn, the representations learned by the TDANN are more brain-like than in spatially-unconstrained models. Finally, we provide evidence that the TDANN's functional organization balances performance with between-area connection

[*]Correspondence: eshed.margalit@gmail.com.

[+]co-senior author

Lead contact: Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Eshed Margalit (eshed.margalit@gmail.com).

Author Contributions

E.M. and D.F. performed analyses. E.M., K.G.-S., and D.L.K.Y. wrote the paper. H.L., J.J.D., and D.L.K.Y. originally conceived the approach.

Declaration of Interests

The authors declare no competing interests.

length. Our results offer a unified principle for understanding the functional organization of the primate ventral visual system.

## eTOC Blurb

Margalit et al. develop a topographic artificial neural network that predicts both functional responses and spatial organization of multiple cortical areas of the primate visual system. In turn, the model minimizes between-area wiring length and produces more brain-like responses to visual stimuli than spatially unconstrained alternative models.

## Introduction

Sensory cortical systems can be measured in two ways: by the response patterns of neurons as a function of stimulus input, and by the spatial arrangement of those neurons across the cortical surface. The confluence of these observations is referred to as *functional organization*, the reproducible spatial arrangement of neurons within a cortical area according to their response properties. Functional organization is among the most ubiquitous of neuroscience findings, appearing in the topographic maps of the visual system[50], and in auditory[52], parietal[45], sensorimotor[117], and entorhinal areas[91,42]. These organized structures anchor our understanding of cortical development, function, and dysfunction. Yet, it remains a mystery what processes govern their emergence, and what computational function they serve.

Any theory of functional organization must explain both neuronal response properties and the physical arrangement of neurons. Furthermore, a complete *unified* theory should account for functional organization in all cortical areas. Prior computational models of the organization within single cortical areas have been developed[5,60,92,28,107,123,26,72,84,83,13,54,65,9,53,3], but these approaches do not generalize to multiple areas. Moreover, many prior models utilize hand-crafted features, and thus cannot explain how neuronal response properties are learned from realistic sensory inputs. Deep artificial neural networks (DANNs) trained with large naturalistic datasets are increasingly being used to model neuronal responses in visual, auditory, and language regions[19,57,118,120,58,44,12,68,103,102]. However, standard DANNs impose no spatial arrangement among model units, and thus cannot explain the organization of neurons across the cortical sheet.

Here, we introduce the Topographic Deep Artificial Neural Network (TDANN), a model that takes a step toward unification by predicting many features of functional organization in multiple cortical areas from a single learning framework. The TDANN implements the hypothesis that neural systems are optimized to address two key goals: supporting ecologically-relevant behaviors by producing useful neural representations[76], and doing so in a biophysically efficient manner. A critical component of biophysical efficiency is the minimization of neuronal wiring length, which is theorized to result in the smooth topographic organization observed in many cortical areas[16,65,54]. The TDANN embeds each layer's units in a two-dimensional simulated cortical sheet, then optimizes a *composite objective function* with two components: a functional objective that drives the learning

of useful representations, and a spatial constraint that encourages efficiency with smooth response patterns across the simulated cortical sheet. We test this framework in the primate ventral visual stream, a cortical system in which functional organization has been extensively documented.

The ventral stream is a hierarchical series of cortical areas that support visual recognition, beginning with primary visual cortex (V1) and ascending through intermediate areas to high-level regions: inferotemporal (IT) cortex in macaques and ventral temporal cortex (VTC) in humans. Well-known neuronal response properties in V1 include tuning to edge orientation[50,100,21], spatial frequency[20], and color[122,73]. These response properties are coupled with topography: orientation preferences form a smooth cortical map with pinwheel-like discontinuities[8,40,10,51,88]; spatial frequency is organized in a quasi-periodic map[51,88,43]; and color-preferring neurons cluster in punctate blobs[73] across V1. Higher-level regions such as primate IT[24,41,96,109] and the analogous human VTC contain neurons with stronger responses for items of specific categories (e.g., faces vs non-faces), a property known as *category selectivity*. A core characteristic of functional organization in IT[109,95] and VTC[55,30,27,78,94,115,39] is that neurons selective for certain ecologically-relevant categories – including faces, places, limbs, and visual wordforms – cluster into spatial patches, with characteristic patch sizes, counts, and inter-patch distances. The location of category-selective regions in human VTC has been related to eccentricity biases[71,46,35], spatial frequency and curvature preferences[87,4], chromatic preference[69], and real-world size[64]. Functional organization in V1 has been related to endogenous activity patterns prior to birth[1] and efficient encoding of visual inputs[93]. Here, we apply the TDANN to test if core phenomenology in multiple cortical areas can be predicted by a single computational model.

We find that the TDANN reproduces several key aspects of the functional organization of multiple regions in the ventral stream, including smooth orientation maps with pinwheels in an earlier model layer, and category-selective patches in a later layer that match the number and size of selective regions in human VTC. We then test which specific functional and spatial constraints of the TDANN are critical to its success by instantiating alternative models and measuring their capacity to predict neuronal data. We find that the combination of task and spatial objectives that best matches the functional organization of the ventral stream also makes learned representations more brain-like by constraining their intrinsic dimensionality. We also find that the TDANN learns these representations while indirectly minimizing between-area wiring length, providing further evidence that brain-like functional organization effectively balances performance with metabolic costs.

Finally, because the the TDANN accurately predicts key aspects of the functional organization of the ventral stream, it provides an exciting new platform for simulating experiments that are challenging to implement empirically. As a proof of principle, we perform *in silico* experiments simulating the effect of cortical microstimulation devices that vary in their spatial precision and cortical coverage (Box 1). Taken together, our experiments suggest that the TDANN provides a framework for understanding the emergence of functional organization in multiple cortical areas of the ventral visual stream.

# Results

## Instantiating models that balance task performance with spatial smoothness

Building on optimization-based approaches in computational neuroscience[98,119], we seek a model architecture and an objective function that generate a neural network which matches the neuronal responses and topography of the primate ventral visual stream.

Because standard DANNs have no within-area spatial structure beyond retinotopy, we must augment their architecture to model spatial topography. Specifically, we take the ResNet-18 architecture[48], a DANN with strong object recognition performance and accurate prediction of neuronal responses throughout the ventral visual stream[103], and augment it by embedding the units of each convolutional layer into a two-dimensional simulated cortical sheet (Figure 1a). Given that neurons in visual cortex are organized retinotopically at birth[4], we assign model unit positions retinotopically, such that units responding to similar regions of input images are nearby in the simulated cortical sheet. The size of the simulated cortical sheet in each layer is anchored by estimates of cortical surface area in the human ventral visual stream (Figure 1a). We refer to the resulting model as the *Topographic DANN (TDANN)*.

Given this architecture, the core of the TDANN approach is to train on a composite objective function that sums two components: a task objective encouraging the learning of behaviorally-useful functional representations, and a spatial objective driving the emergence of topographic properties. Recent work has illustrated the training of (non-topographic) DANNs with *constrastive self-supervised* objectives as models of the ventral pathway[124,63]. Contrastive self-supervised networks learn representations that achieve equally strong neural predictivity as category-supervised networks, but without the need for biologicially-implausible category supervision labels. Here we use SimCLR[17], a simple but especially effective contrastive self-supervised objective, as the task component of the TDANN loss function.

For the spatial loss (SL), we introduce an objective that encourages nearby pairs of units to have more correlated responses than distant pairs of units (Figure 1b, see Methods). The SL is computed separately in each convolutional layer, then summed across layers:

$$\text{TDANN Loss} = L_{\text{task}} + \sum_{k \in \text{layers}} \alpha_k \text{SL}_k \tag{1}$$

where $\alpha_k$ is the weight of the spatial loss in the $k$th layer, set to $\alpha_k = 0.25$ for all layers. The value of $\alpha$ is a free parameter that was selected based on quantitative benchmarks comparing model predictions to neuronal functional organization (Figure 4). Other parameters that impact the spatial loss – including the size of each cortical sheet and the maximum distance across which different units can participate in the spatial loss computation – are fixed based on empirical measurements (see Methods).

Training the TDANN on ImageNet[23] successfully minimized both task and spatial losses (Supplemental Figure S1a,b). We tested if adding the spatial loss interferes with

representation learning by measuring the model's object categorization performance with a linear readout. Categorization accuracy was only slightly lower for the TDANN (median across initialization seeds = 43.9%) than "Task Only" models with no spatial loss ($\alpha = 0$, median = 48.5%; Mann-Whitney $U = 25$, $p = .008$). Moreover, adding the spatial loss term had the intended effect of increasing spatial smoothness (Supplemental Figure S1c,d).

To determine if this learned correlation structure corresponds to brain-like topographic maps, we constructed a battery of quantitative benchmarks comparing model predictions with neural data in two stages of the ventral pathway: V1 and VTC, (Figure 1c). Initial results for intermediate ventral visual areas are presented in Supplemental Figure S2, but we do not include them in our core benchmarks due to a relative lack of empirical data to compare against. As in prior work[12,120], we find that earlier model layers best predict V1 responses and later layers best predict responses in higher visual cortex (Supplemental Figure S2d). Accordingly, we designate the fourth and ninth convolutional layers as the "V1-like" and "VTC-like" layers, respectively, and restrict our analyses to these layers when evaluating benchmarks of functional organization.

## The TDANN predicts the functional organization of primary visual cortex

Neurons in primate V1 are organized into maps of preferred stimulus orientation, spatial frequency, and color[73,88,15]. Because data at the resolution necessary to visualize these maps is not available for human V1, we compare the TDANN to macaque V1 using scale-invariant metrics. We tested if the V1-like TDANN layer captures the functional organization of macaque V1 with three quantitative benchmarks. First, we evaluate functional correspondence by asking if model units in the TDANN V1-like layer have similar preferred orientations and orientation tuning strengths as neurons in macaque V1. Second, we assay cortical map structure by measuring pairwise tuning similarity as a function of cortical distance. Third, we measure the density of pinwheel-like discontinuities in the orientation preference map. In addition to the TDANN, we evaluate four control models on these benchmarks. To test the impact of model training and pre-optimization unit shuffling, we used an *Unoptimized* TDANN, in which model weights and unit positions are left randomly initialized. To determine the effect of the spatial constraint in the loss function, we trained a *Task Only* variant with $\alpha = 0$. The other two controls are self-organizing maps (SOMs), which have been proposed as models of V1 functional organization[107,28]: a traditional SOM in which feature dimensions are manually predetermined (as in[107]), referred to here as the "Hand-Crafted SOM", and the DNN-SOM, a novel SOM that organizes the output of a deep neural network (AlexNet) V1-like layer (inspired by[26,123]).

**The TDANN matches orientation tuning in V1—**We measured orientation tuning strength by presenting a set of oriented sine grating images to the model (Figure 2a), computing a tuning curve for each unit, and calculating the circular variance (CV; lower values for sharper tuning) of each tuning curve. We find that the TDANN V1-like layer has a significantly greater proportion of selective units (CV < 0.6, range across model seeds: [20%, 31%]) than Unoptimized models ([1%, 3%]; Mann-Whitney $U = 25$; $p = .008$, Figure 2b), but fewer than Task Only models ([35%, 50%]; $U = 25$; $p = .008$) or macaque V1 (45%; Supplemental Figure S3c). In contrast, neither the Hand-Crafted SOM nor the DNN-SOM

exhibited any units with sharp orientation tuning. We also find that TDANN and Task Only models (but not SOMs or Unoptimized models) show an over-representation of cardinal orientations (0 and 90 degrees) as in macaque V1[21] (Supplemental Figure S3b; see also Henderson & Serences[49]).

**The TDANN predicts the arrangement of orientation-selective V1 neurons—**To evaluate whether the TDANN V1-like layer captures the topographic properties of macaque V1, we consider the spatial distribution of orientation-selective units – the orientation preference map (OPM) – and find a smooth progression of preferred orientations that resembles macaque V1 (Figure 2c, d). Following prior work[14,31,99], we quantify this structure by measuring the absolute pairwise difference in preferred orientation as a function of cortical distance. In both the TDANN and macaque V1 (data from[88]), we find that nearby units have smaller differences in orientation preference than distant pairs (Figure 2e). In contrast, orientation preference similarity does not vary with cortical distance in Task Only or Unoptimized models, and both the Hand-Crafted and DNN-SOMs exhibit OPMs with abnormally high orientation tuning similarity (Figure 2e, Supplemental Figure S3a). We summarize these profiles by computing a *smoothness score* that measures the increase in tuning similarity for nearby unit pairs compared to distant unit pairs. Smoothness of TDANN OPMs ([min, max] across random initialization: [.64, .83]) was consistent with macaque V1 (.68); however, OPMs in the Hand-Crafted SOM ([.92, .92]) and DNN-SOMs ([.81, .86]) were smoother than in macaque V1. In turn, macaque V1 OPMs were smoother than Unoptimized ([.03, .04]) and Task Only ([.28, .39]) models. Jointly comparing each model to macaque V1 orientation tuning strength and OPM smoothness highlights that the TDANN is the only model class that satisfies both criteria (Figure 2j).

As a more stringent test of OPM structure, we computed the density of pinwheel-like discontinuities in the OPM[10] and compared to the expected value of ~3.1 pinwheels / $mm^2$ in macaque V1[56]. Multiple pinwheels are apparent in both the TDANN and the Hand-Crafted SOM (Figure 2k). We find that the TDANN has lower pinwheel density (range across seeds = [2.0, 2.3] pinwheels / column spacing$^2$) than macaque V1, but significantly higher than either the Task Only ([0.2, 0.8]; Mann-Whitney $U = 25$, $p = .008$) or Unoptimized models (0 pinwheels; Figure 2k). The Hand-Crafted SOM has higher pinwheel density ([3.7, 4.5]) than the TDANN, but the DNN-SOM has no detectable pinwheels. We note that absolute pinwheel density can depend on model architecture (Supplemental Figure S3).

**The TDANN predicts maps of spatial frequency and color preference in V1—**While OPMs are the best-studied feature of V1 functional organization, the cortical sheet simultaneously accommodates organized maps of spatial frequency[88] and chromatic tuning[33,73]. An accurate model of V1 should also predict these maps. We compared spatial frequency preference maps in macaque V1 (data from[88]) and in the TDANN V1-like layer and found a smooth progression of preferred spatial frequency in both (Figure 2f). The TDANN map of spatial frequency preference across random initializations = [.38, .54] ([min, max]) is as smooth as the map in macaque V1 (0.53; Figure 2g), whereas maps from Task Only ([.23, .36]) and Unoptimized models ([.02, .03]) are less smooth than

macaque V1, and both the Hand-Crafted SOM ([.79, .81]) and the DNN-SOM ([.83, .86]) are far smoother than the neuronal data. We observe similar results for maps of chromatic preference (Figure 2h, i), where comparisons are made to imaging of cytochrome oxidase uptake that is prevalent in color-tuned neurons (data from Livingstone & Hubel[73]). In the TDANN chromatic map, the fraction of units with opposite color-tuning increases with cortical distance, again exhibiting comparable smoothness to macaque V1 (TDANN smoothness: [.38, .54], macaque: .53). Together, our analyses demonstrate that the TDANN predicts the multifaceted functional organization of macaque V1, providing a stronger match to neuronal data than existing models.

## The TDANN reproduces many features of higher visual cortex functional organization

Because benchmarks measuring the topographic similarity between models and higher visual cortex, i.e. primate inferior temporal (IT) and human ventral temporal cortex (VTC), are underdeveloped, we introduce five quantitative benchmarks that compare both responses and topography. Response properties are compared by measuring the similarity of category selectivity patterns with representational similarity analysis (RSA, Kriegeskorte et al.[66]), as in Margalit et al.[77], Haxby et al.[47]. Topographic properties are then compared against four complementary benchmarks: 1) the smoothness of category selectivity maps, 2) the number of category selective patches, 3) the area of these patches, and 4) the spatial overlap among units selective for different categories. We compute these metrics for the TDANN's VTC-like layer and for VTC data from eight human subjects in the Natural Scenes Dataset (NSD)[2] (Supplemental Figure S4e).

We also evaluate two alternative models of VTC topography: an SOM trained on the outputs of a categorization-pretrained AlexNet (DNN-SOM, cf[26,123]) and an Interactive Topographic Network (ITN) that is trained on the same dataset (ImageNet) we used[9]. Human subjects and models were presented a common set of 1,440 grayscale images from five categories[105]: faces, bodies, written characters, places, and objects.

**The TDANN predicts patterns of category selectivity—**We characterize neuronal responses in VTC by computing a representational similarity matrix (RSM): the similarity among distributed selectivity patterns for each of the five object categories. The average RSM from human VTC indicates high similarity between distributed selectivity patterns for faces and bodies, and low similarity between distributed selectivity for faces and places (Figure 3a). RSMs from different subjects and hemispheres were very similar, with the 95% CI of Kendall's $\tau = [.72, .75]$. We then compute RSMs for each model and find that some models provide a closer match to human VTC than others (ANOVA $F(4, 331) = 630$; $p < 10^{-152}$). TDANN RSMs closely mirror those in human VTC ($\tau = [.69, .73]$), significantly better than DNN-SOM ($\tau = [.31, .35]$; post-hoc Tukey's HSD $p < 10^{-13}$), ITN ($\tau = [.46, .56]$; $p < 10^{-13}$), Task Only ($\tau = [.65, .68]$; $p = .001$) and Unoptimized ($\tau = [.11, .14]$; $p < 10^{-13}$) models (Figure 3b).

**The TDANN predicts category-selectivity maps—**To compare models against topographic benchmarks, we generate selectivity maps for each of the five object categories (Figure 3c), then quantify their structure by measuring the difference in selectivity as a

function of cortical distance between pairs of units (Figure 3d). We find that many models have similar selectivity profiles, with nearby units having more similar selectivity than distant pairs of units (Figure 3d). Summarizing the curves with the same smoothness metric used in V1 (Figure 3e), we find no significant differences between smoothness in human VTC and the TDANN VTC-like layer (permutation test: $p = 0.30$). The ITN also exhibits VTC-like smoothness ($p = 0.10$), although the maps from the Task Only and Unoptimized models were less smooth than human VTC ($ps < 0.001$), and maps from the DNN-SOM were smoother than human VTC ($p < .001$).

For the remaining topographic benchmarks, we follow the literature by thresholding selectivity maps to find strongly selective units (Supplemental Figure S4a-d). Clusters of selective units are identifiable in human VTC, TDANN, the SOM, and ITN models, but not in Task Only or Unoptimized models. We computationally identify large contiguous clusters of selective units as "patches" (Figure 3f). We find similar sets of patches in VTC and the TDANN: both contain a few patches selective for each category. There are two notable exceptions: object-selective patches are present in the TDANN but not VTC, and the TDANN exhibits a large central place-selective patch flanked by face-selective patches, an arrangement not found in VTC. Quantitative comparison supports the similarity of human VTC and TDANN: there is no significant difference in patch count ($p = 0.99$, Figure 3g) or patch area ($p = 0.67$; Figure 3h). In contrast, we find that the ITN has more than twice as many patches as VTC ($p = 1.2 \times 10^{-5}$), although the patches are as large on average as those in VTC ($p = 0.99$). The DNN-SOM fails to match VTC in the other extreme: while the number of patches in the DNN-SOM is similar to that in VTC ($p = 0.15$), the patches are too large ($p < 10^{-10}$). Joint comparison of models and humans on both patch count and size (Figure 3i) highlights the strong correspondence between TDANN and human VTC.

A hallmark of higher visual cortex functional organization is the reproducible spatial arrangement of units selective for different categories, including the close proximity of face-selective and body-selective regions[95,114] and the separation between face- and place-selective regions. Here we measured the co-occurrence of face-selective and body-selective units (and face-selective and place-selective units) with an overlap score that ranges between 1 (face-selectivity perfectly predicts body-selectivity) to 0.5 (no relationship), to 0 (face- and body-selectivity perfectly anti-correlated). As expected, Face-Body overlap scores are high in human VTC (95% CI across subjects and hemispheres: [.66, .72]), whereas Face-Place overlap was significantly lower (95% CI: [.40, .45], Wilcoxon signed-rank test against one-sided alternative $W = 136$; $p = 1.5 \times 10^{-5}$; Figure 3j). The same pattern is apparent in the TDANN: Face-Body Overlap ([.63, .71]) is significantly higher than Face-Place Overlap ([.14, .26]; $W = 15$; $p = .03$). In the ITN, the Face-Body overlap score was lower than in human VTC (.52), but still higher than the Face-Place overlap score (.36). Neither the the DNN-SOM nor the Task Only models had higher Face-Body overlap than Face-Place overlap (Figure 3j; $ps > 0.5$).

To further gain intuition for the tuning profiles of model units, we synthesized images that optimally drive each region of the VTC-like layer. We make the subjective observation that optimal stimuli vary smoothly across the cortical surface, and that optimal stimuli in

face-selective regions often include round objects resembling eyes (Supplemental Figure S4f,g). We also find that training the TDANN on natural images (either ImageNet[23] or Ecoset[81]) produces accurate V1-like and VTC-like maps, whereas training on noise or simpler hand-crafted stimuli fails to provide a unified account of ventral stream topography and predicts only V1-like functional organization (Supplemental Figure S6a-c).

## Multiple signatures of functional organization emerge at the same spatial constraint strength

While most of the parameters in the the TDANN framework are set according to empirical data, the weight of the spatial loss in the training objective, $\alpha$ is a critical free parameter that cannot be assigned ahead of time. Here, we validate our setting of $\alpha = 0.25$ for the results above by demonstrating that many benchmarks of neuronal similarity are simultaneously satisfied at this value.

Comparison of OPMs in the V1-like layer and category-selectivity maps in the VTC-like layer in models trained at different levels of $\alpha$ shows that functional organization is absent when $\alpha = 0$, is structured at intermediate values of $\alpha$, and deteriorates at high values (Figure 4a). We quantify the dependence of functional organization on $\alpha$ with three kinds of benchmarks: functional similarity (Figure 4b), map smoothness (Figure 4c), and presence of topographic phenomena (i.e., pinwheels and patches; Figure 4d). Considering functional similarity, we find that the fraction of V1-like layer units that are orientation selective is closest to macaque V1 when $\alpha$ is low, and representational similarity between the VTC-like layer and human VTC is maximized at $\alpha = 0.25$ (Figure 4b). Considering topography, smoothness of OPMs in the V1-like layer is most brain-like at $\alpha = 0.1$ and smoothness of category-selectivity in the VTC-like layer is most brain-like at $\alpha = 0.25$ (Figure 4c). Finally, the density of pinwheels in the V1-like layer and category-selectivity patches in the VTC-like layer are most similar to measurements in macaque V1 and human VTC, respectively, at $\alpha = 0.25$ (Figure 4d).

A specific range of $\alpha$ values ($0.1 \leq \alpha \leq 0.25$) thus produces experimentally-observed outcomes across a variety of functional and topographic benchmarks in multiple brain areas.

## Two factors underlying functional organization: self-supervised learning and a scalable spatial constraint

To understand the constraints that shape the ventral stream's functional organization, we construct variants of the TDANN with alternative functional and spatial objectives, then evaluate how these factors affect the accuracy of the resulting models' functional organization.

Most studies comparing neural networks to the brain use models trained for supervised object categorization ([58,120,70]; Figure 5a-bottom left). The TDANN, however, uses contrastive self-supervision[124,17]. We thus considered a variant topographic model using standard visual object categorization as the "task component" of its objective function.

We also investigate how the spatial objective function affects emergent functional organization.[70] recently introduced a spatial loss function that subtracts the inverse of pairwise cortical distances from the magnitude of pairwise response correlations (Figure 5a-bottom right). We refer to this as Absolute Spatial Loss ($SL_{\text{Abs}}$), because minimizing it requires an absolute match between response correlations and (inverse) cortical distances. While training models with $SL_{\text{Abs}}$ produces clustering of category-selective units in a late model layer[70], we found that in layers with shorter cortical distances, $SL_{\text{Abs}}$ can only be minimized if response correlations are pathologically high. The TDANN instead uses a more flexible spatial loss function that we term Relative Spatial Loss ($SL_{\text{Rel}}$); Figure 5a-top right). $SL_{\text{Rel}}$ requires only that inverse cortical distances be correlated with response similarity, *regardless of total cortical surface area*. Interestingly, we find that switching from $SL_{\text{Abs}}$ to $SL_{\text{Rel}}$ slightly increased performance in linear readouts of object category (Supplemental Figure S7b).

We compare the full TDANN model (characterized by having both self-supervised task loss and Relative spatial loss) to these variants both in terms of (1) the smoothness of OPMs and face-selectivity maps in the V1-like and VTC-like layers, respectively, and (2) the number of pinwheels and category-selective patches in those layers.

In the V1-like layer, the Categorization-supervised variant was slightly but significantly less smooth than the TDANN (mean smoothness = 0.56, $U = 25$, $p = 0.008$), with an equal pinwheel density (2.07 pinwheels / column spacing $^2$; $U = 10$, $p = 0.69$). Absolute SL models resemble the TDANN qualitatively (Figure 5b), but with significantly lower smoothness (TDANN mean: 0.71, Absolute SL: 0.40; $U = 25$, $p = 0.008$; Figure 5d) and lower pinwheel density (TDANN: 2.14 pinwheels / column spacing $^2$, Absolute SL: 0.89; $U = 21$, $p = 0.09$; Figure 5e).

In the VTC-like layer, category-selectivity maps were much less organized in the Categorization-supervised variant than in the self-supervised TDANN. At the same spatial weight of $\alpha = 0.25$, clusters of category-selective units are observed in self-supervised but not categorization-supervised models (Figure 5c). The Absolute SL models also fail to form organized category-selectivity maps at this level of $\alpha$. Quantitative comparison reveals smoother category selectivity maps in the TDANN (mean smoothness of face-selectivity maps = 0.44) than in either categorization-trained models (0.09; Mann-Whitney $U = 25$, $p = 0.008$; Figure 5f) or in Absolute SL models (0.13). The TDANN also has a significantly higher number of category selective patches (mean = 1.2) than either categorization-trained (mean = 0) or Absolute SL alternatives (mean = 0.08; $U = 25$, $p = 0.008$; Figure 5g). Thus, both the specific form of the task objective (self-supervised learning) and the spatial loss (relative rather than absolute) are critical for producing brain-like functional organization.

## Spatial constraints make learned representations more brain-like, reducing intrinsic dimensionality

A natural question is whether training with spatial objectives also affects the *non-topographic* aspects of learned representations.

One way to test this is to measure how well model unit responses can predict neuronal responses to a large set of naturalistic stimuli[103,12,120,44]. When fitting neuronal responses as a linear combination of model unit responses[120,124,68,103,19], we find the TDANN has similar neuronal predictivity to non-spatial models, and that there is no effect of model training objective type (Figure 6a). This result is conceptually consistent with prior work[124,19], but somewhat at odds with the dramatic differences between models observed on topographic benchmarks in the preceding sections. One possible explanation for this discrepancy is that linear regression may be too permissive of a mapping: even if a model lacks individual units that resemble recorded neurons, a combination of units might still allow for accurate predictions. We thus performed a more stringent one-to-one mapping[32], in which individual VTC-like layer model units are assigned to individual VTC voxels (Supplemental Figure S8a,b). This one-to-one assignment separate models much more effectively, with the TDANN model exhibiting substantially higher NSD voxel correlation[2] than alternative models (Figure 6b). Correlation peaks at $\alpha = 0.25$, the same value identified by topographic benchmarks (Figure 4), providing further evidence that topographic constraints affect functional representations.

Why are the TDANN features more brain-like? To understand this, we next considered the concept of *intrinsic dimensionality*,[106] a measure of the uniqueness of activation patterns across neurons. Intrinsic dimensionality is low when neurons have similar responses to one another, and high when neurons respond independently. Recent work has demonstrated that standard ANN models have higher intrinsic dimensionality than real macaque V1, and that models with lower dimensionality better predict neuronal responses[61] (but cf.[101]). Because the TDANN's spatial constraint encourages units to respond more similarly to one another, we hypothesized that their intrinsic dimensionality might be reduced.

We computed intrinsic dimensionality with a measure called Effective Dimensionality (ED)[29,22] (see Methods). We find that the addition of the spatial constraint decreases ED in the VTC-like layer regardless of the training objective (Figure 6c, Supplemental Figure S8). Non-spatial models ($\alpha = 0$) have higher ED than human VTC (mean across subjects = 16.7), while ED in the VTC-like layer of categorization-trained models (76.8) is much higher than in self-supervised models (27.8). At the spatial weight where the TDANN best matches neural data ($\alpha = 0.25$), the model's VTC-like layer approaches the ED of human VTC (TDANN mean = 13.2). The ED of models trained with $SL_{\text{Abs}}$ decreases too quickly (mean = 6.5), while the ED of categorization-trained models remains higher than human VTC (mean = 42.7). Similar results are observed when summarizing the response eigenspectrum with power law fits, as in[106,61] (Supplemental Figure S8g). Intriguingly, we find that the ED of the TDANN converges to a common value of approximately 15 at $\alpha = 0.25$ across model layers (Figure 6d), raising the possibility that a similar dimension stabilization phenomenon may occur in the brain. These results provide new evidence that the computational constraints generating cortical topography also make non-topographic features more brain-like.

### The TDANN indirectly minimizes between-area (feed-forward) fiber-tract wiring length

Identifying the optimization paradigm that is most consistent with neuronal data provides insight into the constraints underlying neural development, but prompts a deeper question: why would these constraints be favored by evolutionary selection in the first place? One hypothesis is that functionally organized cortical systems also minimize wiring length, consequently reducing brain size, weight, and power consumption[18,54]. Though the TDANN objective does not directly minimize wiring length, we test this hypothesis by asking whether models that best predict functional organization also reduce a measure of between-area wiring length. In feed-forward models that lack within-area connectivity, such as the TDANN, any potential gains in wiring efficiency must be between areas. To test if such gains occur, we estimate the feed-forward wiring length needed to connect populations of co-activated model units in adjacent areas modeled in the the TDANN. For each pair of adjacent layers, we construct virtual fibers that originate in the upstream "source" area and terminate in the downstream "target" area, adding between-area fibers until the total distance between each activated unit and its nearest fiber is below a specified threshold (see Methods, Figure 7a).

In principle, co-activated units could be distributed uniformly throughout the cortical sheet, but we find that presenting the TDANN with natural images leads to localized clusters of responses in the VTC-like layer of all models trained with $\alpha > 0$, with multiple clusters apparent at higher levels of $\alpha$ (Supplemental Figure S9). Critically, we find that this increase in clustering within areas also results in shorter wiring length between areas at higher levels of $\alpha$ (Figure 7b). However, we also find that object categorization performance decreases as wiring efficiency improves (Figure 7c), indicating that models at low-to-intermediate levels of $\alpha$ optimally balance performance with between-area wiring efficiency. This coincidence of optimal $\alpha$ values suggests that the functional organization of the ventral visual stream balances inter-area wiring costs with performance. Finally, we also find more efficient between-area wiring for optimization objectives that yield the most brain-like functional organization: fiber length is lower in the TDANN than categorization-trained models and the Absolute SL-trained models (Figure 7d).

Having identified a model that reproduces many aspects of ventral stream functional organization, it is interesting to consider new opportunities that the TDANN unlocks. In Box 1 we give examples of using the TDANN to simulate microstimulation of neuronal populations and a proof-of-principle for the design of cortical prostheses.

## Discussion

In this work, we use artificial neural networks models to elucidate principles of functional organization in the primate ventral visual stream. We found that training a topographic deep neural network for a specific combination of objectives results in a model, the TDANN, that captures several key functional and spatial properties of ventral stream responses, from the pinwheels of V1 to the category-selective patches of higher visual cortex.

We identified two specific factors critical to brain-like functional organization. First, we find that *self-supervised* learning of *task-general* representations yields more neurally

correct organization than the more commonly-deployed objective of supervised object categorization. Self-supervised objectives are *a priori* compelling because they can be implemented by the organism without the need for unrealistic supervision labels. While previous work on self-supervised visual system models[124,62] has largely shown parity between self-supervised and category-supervised objectives in their ability to explain neural data, our results show how more biologically-plausible self-supervision leads to quantitatively improved models of the visual system. Moreover, while other work has suggested that functional specialization to categories in the ventral stream can arise under joint training for two different supervised recognition tasks, one for faces and one for objects[25], our results demonstrate that functional specialization can be unified under a single unsupervised learning objective on a single training set.

Second, we find that the spatial constraint in our model should compare response similarity and physical similarity according to a metric that scales with the size of each cortical area. This finding suggests a new idea: that circuits shaping the structure of local response correlations should scale with the surface area of cortical regions. Our identification of these two critical factors demonstrates that comparing optimization objectives can yield concrete insights into principles underlying cortical systems.

An intriguing possibility is that these mechanisms might extend to predict the abundant, yet largely unexplained, functional organization non-visual sensory cortex. We hypothesize that the functional organization of auditory[57,90], somatosensory[117], entorhinal[42,91] and parietal cortices[45] may also be explained by contrastive self-supervised learning under spatial smoothness constraints. Under this hypothesis, it is the structure of the input data (e.g., auditory experience, somatosensory input) that changes, but the mechanisms for learning and organization remain universal across cortical systems. Future work can directly test this hypothesis by training TDANN variants to learn spatially-organized representations specific to each system.

The TDANN is the first unified model to predict key functional organization signatures in multiple cortical areas by learning features and topography, from scratch, in an end-to-end optimization framework trained directly on image inputs. As the TDANN is trained end-to-end, it provides the opportunity for modeling the interaction between learned representations and functional organization during development. Preliminary analyses suggest that trajectories of TDANN functional architecture throughout training roughly match the faster development of earlier vs higher cortical regions (Supplemental Figure S6d,e), but more work is needed to develop these ideas. There are several limitations that can be addressed in future work. First, we benchmarked functional organization in only two cortical regions: V1 and VTC, as these regions have the most empirical data to compare against. Future work can test functional organization in all visual areas (Supplemental Figure S2c includes qualitative results in V4) and include other aspects of the ventral stream such as eccentricity bias[46,35]. Second, because the architecture used here is feed-forward, there are no within-layer connections between units, so wiring-length inferences can only be made between layers. A more complex architecture could include both within-layer recurrence and long-range feedback connections[89], although our results demonstrate that explicitly modeling these recurrent connections is not necessary to produce accurate

topographic maps (see also[9] for consistent results), raising the possibility that minimization of the length of long-range fibers may be the key determinant of the functional organization of visual cortex. Third, the TDANN uses a separate, square-shaped cortical sheet for each cortical area. An improved model would model all areas with a single cortical sheet, integrating neuroanatomically-accurate details of the folding and three-dimensional structure of the cortical surface[116,113,36]. Finally, like all convolutional neural networks, the TDANN uses the same filter weights across the entire visual field (termed "weight-sharing") to make large-scale network training feasible. However, weight-sharing is biologically implausible and potentially interferes with topographic map formation, since changing input weights to a unit in one part of the cortical sheet will also change the weights of many distant units. In this work we pre-optimized unit positions in a way that allows the learning of locally-smooth topographic maps even with weight-sharing.

Weight-sharing also requires our approach to wiring-length optimization to be indirect: instead of explicitly minimizing wiring length and then checking for within-layer feature smoothness, we optimize for within-layer smoothness and then test how this affects the length of between-area virtual fibers. This indirect result is interesting, because it shows that wiring length minimization can emerge without having to explicitly build it in, and suggests a simple mechanism by which *between-area* wiring length minimization can emerge purely from a local *within-area* spatial constraint. Future work can reconcile direct optimization of wiring length with the restrictions of weight-sharing (see e.g.[75]). Beyond issues of computational efficiency, our results raise an intriguing question: to the extent that wiring length is minimized during brain development, does the biophysical mechanism that implements this optimization involve direct measurement and control of between-area fiber length, or is it more akin to our within-area local smoothness constraint, which then indirectly minimizes feedforward wiring length?

Finally, an exciting application of the TDANN is the simulation of experiments with spatial manipulations and readouts (Box 1). Indeed, experiments that involve perturbation of local neuron populations (e.g.,[97,104]) that are difficult to do in humans could use the TDANN to predict the downstream behavioral impact of those manipulations. In sum, a unified model of functional organization, the TDANN, now allows a rich comparison between models and cortex.

## STAR Methods

### Resource Availability

**Lead Contact.—**Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Eshed Margalit (eshed.margalit@gmail.com).

**Materials Availability.—**This study did not generate new materials apart from data and code.

### Data and Code Availability.

1. Original stimuli, data, and model checkpoints produced in this paper have been deposited at OSF at https://osf.io/64qv3/. Accession numbers for original and publicly available datasets are listed in the key resource table.
2. All original code has been deposited at https://github.com/neuroailab/TDANN and is publicly available as of the date of publication. DOIs are listed in the key resources table.
3. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

## Method Details

### Neural network architecture and training.

**Model training:** We build off of the *torchvision* implementation of ResNet-18[48] and train models with modifications to the VISSL framework[37]. All models were trained for 200 epochs of the ILSVRC-2012 (ImageNet Large-Scale Visual Recognition Challenge;[23]) training set. Unless otherwise indicated, models were each trained from five different random initial seeds. Network parameters were optimized with stochastic gradient descent with momentum ($\gamma = 0.9$), a batch size of 512, and a learning rate initialized to 0.6 then decaying according to a cosine learning schedule[74]. Models were trained either for supervised 1000-way object categorization or on the self-supervised contrastive objective "SimCLR"[17]. Following training, categorization accuracy for self-supervised models was assessed by freezing the parameters of the model and training a linear readout from the outputs of the final layer. The linear readout is trained for 28 epochs with a batch size of 1,024 and a learning rate initialized to 0.04 and decreasing by a factor of 10 every eight epochs.

**Initialization of model unit positions:** Prior to training, model units in each layer are assigned fixed positions in a two-dimensional cortical sheet that is specific to that layer. For efficiency, we do not embed the units of the very first convolutional layer. The size of the cortical sheet in each layer depends on a mapping between model layers and regions in the human ventral visual pathway, as well as a commitment to the extent of the visual field being modeled. For example, because we map model Layer 4 to human V1, the surface area of the cortical sheet in that layer is set to $13 cm^2$: the mean value reported by[7] for the surface area of the section of human V1 that is sensitive to the central 7 degrees of visual angle. Another critical parameter in our framework is the size of a "cortical neighborhood": during training, computation of the spatial loss is restricted to units within the same cortical neighborhood. We set the neighborhood width to match measurements made of the spatial extent of lateral connections in different cortical areas of the macaque (from[121]), then scale up to achieve estimates that might match the human ventral visual pathway. Table 1 details the sizes of simulated cortical sheets and cortical neighborhoods in all layers.

Positions are assigned in a two-stage process:

Because each layer performs a convolution over the previous layer's outputs, responses are organized into spatial grids. We preserve this intrinsic organization by assigning each model unit to a region of the simulated cortical sheet that corresponds to its spatial receptive field.

Convolutional networks share filter weights between units at different locations; thus, local updates to a single unit entail updates to all units with the same filter weights. It is highly unlikely that an arbitrary configuration of unit positions will permit local smoothness under this global coordination constraint. Thus, we perform pre-optimization of unit positions to identify a set of unit positions for which learning smooth cortical maps is possible. Specifically, we spatially shuffle the units of a pre-trained DCNN on the cortical sheet such that nearby units have correlated responses to a set of sine grating images. The choice of sine gratings here is inspired by observations that edge-like propagating retinal waves drive experience-independent organization of the visual system in primates and other mammals[82,59,80,34].

The spatial shuffling works as follows: 1) Select a cortical neighborhood at random. 2) Compute the pairwise response correlations of all units in the neighborhood. 3) Choose a random pair of units, and swap their locations in the cortical sheet. 4) If swapping positions decreases local correlations (measured as an increase in the Spatial Loss function described below), undo the swap. 5) Repeat steps 3-4 500 times. 6) Repeat steps 1-5 10,000 times.

**Loss functions:** We use two kinds of loss functions: spatial losses that encourage topographic structure, and task losses that encourage the learning of visual representations. We detail each in turn below:

The spatial loss (SL) function encourages nearby pairs of units to have response profiles that are more correlated with one another than those of distant of units. Consider a neighborhood with $N$ units. The vector of pairwise Pearson's response correlations, $\vec{r}$, has length $M = \binom{N}{2}$, the number of unique pairs. Let the corresponding vector of pairwise Euclidean cortical distances be denoted $\vec{d}$.

We define two SL variants:

$$SL_{\text{Abs}} = \frac{1}{M}\sum_{i=1}^{M} | r_i - D_i | , \tag{2}$$

$$SL_{\text{Rel}} = 1 - Corr(\vec{r}, \vec{D}), \tag{3}$$

where $Corr$ is the Pearson's correlation function and $\vec{D}$ is the inverse distance:

$$D_i = \frac{1}{d_i + 1} \tag{4}$$

The task loss is computed from the output of the final model layer. We use two task losses: the object categorization cross-entropy loss used in supervised object recognition (e.g.[67]) and the self-supervised SimCLR objective[17].

On each batch, model weights are updated to minimize a weighted sum of the task loss and the spatial loss contributed by each layer:

$$\text{TDANN Loss} = L_{\text{task}} + \sum_{k \in \text{layers}} \alpha_k \text{SL}_k \tag{5}$$

where $\alpha$ is the weight of the spatial loss.

In summary, models are trained in 6 steps:

1. ResNet-18 is trained on the task loss only.
2. Positions in each layer are initialized to preserve coarse retinotopy (Stage 1).
3. Positions are further pre-optimized in an iterative process that preserves retinotopy while bringing together units with correlated responses to sine gratings images (Stage 2).
4. Positions are frozen and never again modified.
5. All network weights are randomly re-initialized.
6. The network is trained to minimize a weighted combination of the spatial and task loss components.

### Benchmarks comparing macaque V1 to model V1-like layers.

**Stimuli and Tuning Curves:** Tuning to low-level image properties such as orientation, spatial frequency, and chromaticity was assessed by constructing 224 × 224 pixel sine grating images that span 8 orientations evenly spaced between 0 and 180 degrees, 8 spatial frequencies between 0.5 and 12 cycles per degree, 5 spatial phases, and two chromaticities: black/white gratings and red/cyan gratings.

We evaluated tuning for orientations and spatial frequencies by constructing tuning curves for each unit. Color-responsiveness is assessed by comparing the mean response to all black and white gratings to the mean response to all red/cyan gratings. The distribution of model unit activations for a given layer was rescaled to match the minimum and maximum firing rates reported in[100]. We quantify the orientation tuning strength of model units using circular variance (CV), where values closer to 0 correspond to sharper tuning. As in[100], CV is defined as:

$$CV = 1 - \left| \frac{\sum_k r_k e^{i2\theta_k}}{\sum_k r_k} \right| \quad (6)$$

Where $\theta_k$ is the $k$th orientation, in radians, and $r_k$ is the scaled response to that orientation. Orientation tuning curves are additionally fit with a von Mises function whose peak is taken as the preferred orientation.

**Models:** Our hand-crafted self-organizing map (SOM) implementation uses the *MiniSom* library[112], with parameters adapted from[107]. We instantiate the SOM as a 128 x 128 grid of model units.

10,000 training samples were randomly constructed by selecting a random (x, y) location, orientation ([0, $\pi$], spatial frequency ([0, 1]), and chromaticity (black/white, colorful).

As in[107], SOM weights were initialized retinotopically with randomly-selected initial preferred orientations. The SOM is trained by presenting training examples for a total of 700,000 updates. After each example, the "winning" unit (i.e. the one with the highest response) is updated with a learning rate of $\epsilon = 0.02$ to be more strongly aligned with the input stimulus, and its neighbors are updated in proportion to their proximity to the winner, as determined by a Gaussian neighborhood function parameterized by $\sigma = 2.5$.

Following training, each sine grating in the set of probe stimuli is presented to the SOM by projecting it into the six-dimensional space of SOM unit tuning and computing the response of each SOM unit to the stimulus. Once responses to each stimulus are obtained, tuning curves are constructed as usual.

The DNN-SOM is identical to the hand-crafted SOM, except that 1) the inputs are derived from the outputs of the first layer of an AlexNet model pretrained for ImageNet object categorization and 2) the learning rate is increased, which we found helps convergence. Following the approach of[123], we take the responses of the first AlexNet layer to all 50,000 natural images in the ImageNet dataset, reduce their dimensionality with principal components analysis, and train the SOM on those examples.

**Response Benchmarks:** Model responses are compared to macaque V1 by considering preferred orientations and orientation tuning strength. Orientation tuning strength is computed as circular variance (CV) and compared between the population of model units and the empirical distribution provided by[100] with the Kolmogorov-Smirnov distance. To filter out noisy units, we compute CV for model units with a mean response magnitude of at least 1.0. The distribution of preferred orientations is also compared to empirical data collected by[21] by counting the number of units preferring each of four orientations: 0, 45, 90, and 135 degrees. In Supplemental Figure S3b we compute a "Cardinality Index": the fraction of preferred orientations that include, 0, 90, and 180 degrees.

**Topographic Benchmarks:** Orientation preference maps (OPMs) are compared to empirical measurements in two ways: counting pinwheels and quantifying map smoothness.

We interpolate the OPM onto a two-dimensional grid by computing the circular mean of the preferred orientation of units near a given location. If the population of model units near a grid location has high heterogeneity in preferred orientation, we disqualify that pixel for having an unreliable estimate of preferred orientation. Each grid location is assigned a "winding number"[13], computed by considering the preferred orientations of the eight pixels directly bordering the pixel under consideration. Moving clockwise around the bordering eight pixels, the change in preferred orientation from pixel to pixel is summed. A high winding number indicates a clockwise pinwheel, and a low winding number indicates a counterclockwise pinwheel, where the thresholds for "high" and "low" are selected to be consistent with manual annotation of clear pinwheels. When computing pinwheel density, we report the number of identified pinwheels per "column-spacing", i.e. we normalize to the distance between iso-orientation areas. We note that the orientation column spacing in The TDANN (~ 3.5mm width) does not match macaque V1 (~ 1mm). This mismatch, caused in part by our commitment of the TDANN as a model of human visual cortex and not macaque visual cortex, can also be overcome by increasing the number of units in the network at the expense of increased computational cost (see Supplemental Figure S3d-f).

We compute the smoothness of orientation preference maps by constructing a curve relating pairwise difference in preferred orientation to pairwise cortical distance. First, we restrict the population of model units to those with the highest 25% peak-to-peak tuning curve magnitudes. This filtering step removes units with weak responses or responses that would be indistinguishable from a "cocktail blank" background activity level, and we consider it equivalent to neuron selection in electrophysiological and optical imaging studies[100,88]. As in similar approaches to quantifying OPM structure (e.g.[14]), pairs of units are binned according to their distance, and the average absolute different in preferred orientation is plotted for each distance bin. Because there can be hundreds of thousands of units in a given layer, we restrict this analysis to randomly-selected neighborhoods of a fixed width, then sample many neighborhoods from each map. Finally we divide the pairwise difference by the chance value obtained by random resampling of unit pairs, such that a values $< 1$ indicate more similar tuning than would be expected by chance.

The OPM curves are compared to reconstructed macaque V1 data from[88].

We adopt an identical approach for the construction of a neuronal spatial frequency preference map, where data are also provided for the same imaging window in[88]. A similar strategy was used to recover data on cytochrome oxidase (CO) uptake from[73].

We define a smoothness score for a given map by comparing the tuning similarity for the nearest model unit pairs to the tuning similarity of the least similar pairs. Concretely, given a vector $x$ of pairwise tuning similarity values, sorted in order of increasing cortical distance:

$$S(x) = \frac{\max(x) - x_0}{x_0}$$

(7)

### Benchmarks comparing human VTC to model VTC-like layers.

**Stimuli:** We evaluate the selectivity of neurons and model units to visual object categories using the "fLoc" functional localizer stimulus set[105]. fLoc contains five categories, each with two subcategories consisting of 144 images each. The categories are faces (adult and child faces), bodies (headless bodies and limbs), written characters (pseudowords and numbers), places (houses and corridors), and objects (string instruments and cars). Selectivity was assessed by computing the $t$-statistic over the set of functional localizer stimuli and defining a threshold above which units were considered selective.

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

(8)

where $\mu_{\text{on}}$ and $\mu_{\text{off}}$ are the mean responses to the "on" categories (e.g., adult and child faces) and "off" categories (e.g., all non-face categories), respectively, $\sigma^2$ are the associated variances of responses to exemplars from those categories, and $N$ is the number of exemplars being averaged over.

**Human Data:** We compare models to human data from the Natural Scenes Dataset (NSD)[2], a high-resolution fMRI dataset of responses to 10,000 natural images in each of eight individuals (see Allen et al. for details). Models are compared to two aspects of this dataset: single-trial responses to the main set of natural images per participant (see "One-to-one mapping") and selectivity in response to the "fLoc" stimuli. Single-trial responses were $z$-scored across images for each voxel and session and then averaged across three trial repeats. Selectivity was computed on the "fLoc" experiment as described in the previous section, generating $t$-maps for each of the five categories for each individual subject. While some category-selective regions are pre-defined in the NSD dataset, those regions include regions with very weak selectivity. To better align with the literature on category selectivity, we recompute selectivity and patch boundaries in the human data using the same contrasts and thresholds as the models we compare to. The VTC region of interest (ROI) was drawn based on anatomical landmarks to follow the convention in the literature[11] and is provided in the NSD data release as the "Ventral" ROI in the "streams" parcellation.

**Models:** We reconstruct maps from a variant of the ITN in[9] that was trained and evaluated on the same images as the remaining models.

Two related approaches for building SOM models of higher visual cortex have recently been published[123,26]. Because neither paper evaluates the resulting topographic maps with the fLoc stimuli, we approximately reimplement the approach of[123] as follows. We extract the responses of each unit in the final layer of a pretrained AlexNet to all 50,000 images in the ImageNet validation set. The responses are then reduced to the first four principal components. The SOM is initialized as a 200 x 200 grid of model units with a Gaussian

neighborhood function set to $\sigma = 6.2$. The learning rate is set to 1.0 and the SOM is trained for 200,000 total iterations. The fLoc images are presented to the pretrained AlexNet model and projected into the space spanned by the four principal components computed previously. The response of each model unit to each fLoc image is computed by taking the dot product of the unit weight matrix with the projected fLoc images. The SOM is then treated identically to the VTC-like layer of the TDANN. We note that the DNN-SOM implemented here does not perfectly replicate the approach of either[26] or[123], but uses the same DNN architecture, weights, and input images. It remains possible that exact replication of these approaches would yield different results; however, we were unable to achieve stronger results in our exploration of alternative implementations. With respect to[123], we do not include a step to warp the simulated cortical tissue to the morphology of human VTC.

**Response Benchmarks:** We compare functional properties of human VTC and models with representational similarity analysis (RSA)[66]. For any given model or human hemisphere, we compute a representational similarity matrix (RSM) as the pairwise Pearson's correlation between patterns of selectivity for each of the five fLoc categories. The diagonal of the RSM is trivially 1.0 and is ignored in further analysis. The similarity of two RSMs is computed as Kendall's $\tau$.

**Topographic Benchmarks:** We measure pairwise difference in VTC-like layer unit tuning as a function of cortical distance. We draw 25 randoms samples of 500 units each. Each sample is filtered to include only units with a mean response of at least 0.5 a.u.. For each fLoc category, the absolute pairwise difference in selectivity is computed for pairs of units separated by different cortical distances. Curves are normalized by the chance value obtained by randomly shuffling unit positions. Smoothness of maps is computed from these curves, same as in our analysis of V1. To compare a model to a human hemisphere, we compute the mean category-by-category difference in smoothness, e.g. comparing model face map smoothness to human face map smoothness, model body map smoothness to human body map smoothness, etc. Permutation tests randomly assigning category-by-category smoothness profiles to either "model" or "human" were used to assess the statistical significance of the mean difference in smoothness.

Patches are automatically detected in maps of category selectivity by identifying contiguous regions of highly-selective units (or voxels, for human VTC). Patch identification has a small number of parameters that can be adjusted for maps of different sizes and with different dynamic ranges of selectivity values. The first step in identifying patches is to smooth and interpolate discrete selectivity maps. The selectivity map is then thresholded, and contiguous islands surviving the threshold are retained as candidate patches. Each candidate patch is further filtered for reasonable size: patches must be at least $100 mm^2$ and no larger than $45 cm^2$. Finally, the 2D geometry of the patch is constructed by fitting the concave hull of the points within the patch.

The following table identifies the relevant parameters for patch identification in human VTC and for each candidate model class.

A measure of proximity between face- and body-selective regions was previously introduced in[70]. We determine if units (or voxels, for human VTC) that are selective for a pair of categories overlap with one another as follows. First, we bin the cortical sheet into discrete square neighborhoods of width 10mm. In each neighborhood, the fraction of units selective for Category X and Category Y are recorded. We consider two populations as overlapping if there is a strong correlation between the proportions recorded across neighborhoods, i.e., if the frequency of Category 1 selectivity is predictive of Category Y selectivity and vice-a-versa. The X-Y Overlap score is computed as

$$\text{Overlap} = \frac{1 - RankCorr(X, Y)}{2}, \tag{9}$$

where RankCorr is the Spearman's rank correlation coefficient and $\overrightarrow{X}$ is the proportion of units selective for Category X in each cortical neighborhood. The category selectivity threshold was set at $t > 4$. One might consider other measures of inter-patch geometry; for example, face patches are lateral to place-selective patches in human VTC, but this is not apparent in the TDANN. We have not quantified this kind of inter-patch geometry given the complexity in registering coordinates between the simulated cortical sheet and human VTC.

**Linear regression:** Neural predictivity is computed against a given dataset as the mean variance explained across neurons and splits of the data. In practice we follow the parameters and design decisions made by the BrainScore team[103]; they are repeated here for completeness. We use partial least squares (PLS) regression to predict the activity of a given neuron as a linear weighted sum of model units in a given layer. Model activations are preprocessed by first projecting unit responses to ImageNet images onto the first 1000 principal components, i.e. each component is a linear mixture of model units. This projection is used when fitting on the stimuli that were shown to the animal. When fitting IT, we use data from Majaj, Hong, et al., 2015[76], which consists of multi-electrode array data in responses to quasi-naturalistic scenes with a variety of objects on a variety of backgrounds. Variance explained is corrected by dividing raw predictivity by the internal noise ceiling, a measure of the consistency of each recorded neuron.

***One-to-one mapping of visual cortical responses.:*** A direct, one-to-one mapping between units and voxels is computed by assigning each unit in a layer of the network to a single voxel based on responses to a given dataset. In practice, we correlate individual model unit activations to the natural images from the Natural Scenes Dataset[2] with responses to these same images on the single voxel level for a given subject. Unit-to-voxel assignments are determined using a polynomial-time optimal assignment algorithm[86] which maximizes the overall average correlation between unit and voxel pairs, on a given training set. The 515 shared images that all eight subjects viewed three times were held out as a test set and all reported one-to-one correlations are calculated on this test set, using the unit-to-voxel assignments determined from training. Each unit-to-voxel correlation is normalized by the individual voxel noise ceiling of that assigned voxel (see Allen et al. for information on the calculation of the intra-individual voxel noise ceilings in NSD). One-to-one correlations

were calculated on an individual subject basis for each of the self-supervised and supervised models trained at each level of the spatial weight α. The inter-individual, or subject-to-subject, noise ceiling, was calculated in the same manner, this time assigning voxels from one subject to voxels from another subject based on how correlated responses to the shared 515 images were for each potential voxel pair. For the subject-to-subject assignment, we used an 80/20 train/test split and averaged results for each subject combination across 5 splits. A similar analysis appears in[32].

***Wiring Length.:*** We measure the functional wiring length between two adjacent layers, the "source" layer and the "target" layer by first identifying the units with the highest responses in each layer, then computing the length of between-area fibers that would be required to connect them. First, for a given natural image input, we identify the top $p$% most responsive units in each of two adjacent layers. We set $p$ to 5% in the V1-like layers and 1% in the VTC-like layers. We note that for computational tractability, we restrict our analysis to small neighborhoods in the V1-like layers and average results across many random neighborhood selections.

Next, between-area fibers are added one by one, until all activated units in the earlier "source" layer are sufficiently close to the location at which a fiber originates. In practice, we find the optimal fiber origination sites using the $k$-means clustering algorithm, and continue adding fibers until the total "inertia" of the $k$-means clustering falls below a specified threshold, $k_{\text{thresh}}$. Inertia is computed as the sum of the squared distances between each activated unit and its nearest fiber, and $k_{\text{thresh}}$ is set such that the mean distance from each unit to its nearest fiber is not greater than $d_{\text{thresh}}$. $d_{\text{thresh}}$ is set to 10.0mm in the VTC-like layer pairs, and is reduced to 0.9mm in the V1-like layer pairs to reflect the smaller cortical neighborhood. Having established the number of between-area fibers required and their origination sites in the "source" layer, we identify optimal termination sites for those fibers in the "target" layer as follows. The set of target layer termination sites is identified as the centroids from $k$-means clustering, with $k$ set to the number of fibers. Finally, fibers are assigned between origination sites and termination sites with the linear sum assignment algorithm, and the total wiring length is computed as the sum of the lengths of each individual between-area fiber.

A critical decision when measuring wiring length in this way is how to situate units from two layers in a common physical space. By design, each TDANN layer occupies a unique two-dimensional sheet, leaving the spatial relationships between units in different cortical sheets undefined. Here, we assume that the "source" cortical sheet and "target" cortical sheet lie in the same 2D plane, joined at one edge. Concretely, we can position the "target" sheet to the left, right, above, or below the "source" layer. Without reason to choose one of these strategies, we compute the optimal wiring length for each of the four options and report the average across all shift directions.

***Dimensionality.:*** In our analyses of dimensionality, we consider the responses of the full population of model units in each layer to a set of 10,112 natural images from the NSD[2]. Following[29], we perform spatial max-pooling on the convolutional feature maps, then

compute the eigenspectrum of these responses. We summarize the dimensionality of the responses by their effective dimensionality (ED;[22]):

$$ED = \frac{\left(\sum_{i=1}^{N} \lambda_i\right)^2}{\sum_{i=1}^{N} \lambda_i^2}, \tag{10}$$

where $\lambda_i$ is the $i$th eigenvalue, and $N$ is the number of eigenvectors.

***Microstimulation of model units on the simulated cortical sheet.:*** We simulate the microstimulation of local populations of model units to 1) gain insight into the functional properties of local populations, and 2) measure effective connectivity between groups of units in adjacent layers. In all analyses, stimulation is performed by fixing the activity of units to values determined by a 2D Gaussian function. Units near the center of the Gaussian have their activity set to the maximal value, and activity falls off with distance from the center. We consider the top 5% of units, ranked by activity level, as being responsive in either the "Source" layer, where activity is set according to the 2D Gaussian, or in the following "Target" layer, where unit activity is determined by the network architecture and learned weights.

In VTC-like layers, we measure functional alignment between layers by comparing the category selectivity of activated units in the Source layer (Layer 8) with the selectivity of responsive units in the Target layer (Layer 9). For each stimulation site, we compute the mean selectivity ($t$-statistic) of the top 5% most activated units for each of the following categories: faces, bodies, characters, cars, and places. This five-element "selectivity profile" can then be compared to the profile of the top 5% most strongly responding units in the Target layer by computing $\chi^2$ distance between selectivity profiles. Similarity is then taken as the negative log distance and compared to a shuffle-control in which a random subset of units is compared instead of the top 5% most active units.

***Simulation of a Visual Cortical Prosthesis.:*** In Box 1, we demonstrate a proof of concept for using topographic DCNNs to prototype visual cortical prosthetic devices. This proof of concept consists of two distinct stages: 1) generating device-achievable stimulation patterns with a Stimulation Simulator, and 2) generating the estimated percept (Percept Synthesizer) that would result by stimulating cortical areas with those patterns. To generate stimulation patterns, we feed a target image into TDANN and record the precise activation magnitude of each model unit in each layer. If an infinitely high-precision stimulation device with absolute coverage of the cortical sheet in all cortical areas were available, we would stimulate cortex with this set of precise activation patterns. However, real stimulation devices are limited in many ways, including limits to their spatial precision and the set of cortical areas they can access. Thus, we use TDANN to produce *device-achievable* stimulation patterns, i.e., those that are consistent with the limitations of cortical stimulation devices. Here we take a simple approach by considering degradation of high-precision patterns into device-achievable patterns by Gaussian blurring. In each layer, we first interpolate the precise activity patterns onto a high-resolution grid (2500 × 2500 px), then blur the resulting

pattern with a 2D Gaussian kernel whose $\sigma$ parameter is set according to the desired blur level. Because different layers have different cortical sheet sizes (e.g. 70mm on an edge in the VTC-like layer and 37mm on an edge in the V1-like layer), the width of the Gaussian in *pixels* is variable, even though the width of the Gaussian in *mm* is constant. Finally, we perform a nearest-neighbor lookup such that each model unit adopts the activity level of the pixel closest to its location. This set of activity patterns is the final "device-achievable" pattern. The Stimulation Simulator also allows any specific subset of layers to be included; e.g. the first two layers only, or all eight layers. We consider this restriction comparable to the limited access a neuronal stimulation device might be restricted to.

Given a set of device-achievable activity patterns, we seek to determine the estimated percept that would be evoked if that pattern were written into cortex, i.e., the visual input that is most consistent with those patterns. To this end, we follow the example of[38] and use gradient-ascent image optimization methods to synthesize an image such that the activity pattern produced by presenting that image is as close as possible to the device-achievable target pattern. We use the *lucent* Python package to iteratively optimize an image to minimize the total mean squared error, summed across layers, between the target activity patterns and the current evoked patterns at that iteration. We optimize the image for 3000 steps at a learning rate of 0.05; further optimization has little effect on reducing the mean squared error. The optimized result is the predicted percept for a given input image and theoretical cortical stimulation device.

### Quantification and Statistical Analysis

Statistical analyses were performed in Python using the pandas[108,79] and pingouin[110] libraries. The statistical tests used, values of n, and measures of spread are indicated either in the text of the Results section or in figure captions. Where applicable, significance was defined as a $p$-value below 0.05.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

This work was supported by a National Science Foundation Graduate Research Fellowship awarded to E.M., a National Institutes of Health grant (RO1 EY 023915) awarded to K.G.-S., a Simons Foundation grant (543061) awarded to D.L.K.Y., a National Science Foundation CAREER grant (1844724) awarded to D.L.K.Y., and an Office of Naval Research grant (S5122) awarded to D.L.K.Y. We also thank the NVIDIA corporation and the Google TPU Research Cloud group for hardware grants. We are grateful to Ben Sorscher for helpful discussions.

## References

1. Ackman JB, Burbridge TJ, and Crair MC (2012). Retinal waves coordinate patterned activity throughout the developing visual system. Nature 490, 219–225. 10.1038/nature11529. [PubMed: 23060192]
2. Allen EJ, St-Yves G, Wu Y, Breedlove JL, Prince JS, Dowdle LT, Nau M, Caron B, Pestilli F, Charest I, Hutchinson JB, Naselaris T, and Kay K (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. Nat. Neurosci 25, 116–126. 10.1038/s41593-021-00962-x. [PubMed: 34916659]

3. Anderson Keller T, Gao Q, and Welling M (2021). Modeling Category-Selective Cortical Regions with Topographic Variational Autoencoders. arXiv.
4. Arcaro MJ and Livingstone MS (2017). A hierarchical, retinotopic proto-organization of the primate visual system at birth. Elife 6. 10.7554/eLife.26196.
5. Barrow HG, Bray AJ, and Budd JML (1996). A Self-Organizing Model of "Color Blob" Formation. Neural Comput. 8, 1427–1448. 10.1162/neco.1996.8.7.1427. [PubMed: 8823941]
6. Beauchamp MS, Oswalt D, Sun P, Foster BL, Magnotti JF, Niketeghad S, Pouratian N, Bosking WH, and Yoshor D (2020). Dynamic Stimulation of Visual Cortex Produces Form Vision in Sighted and Blind Humans. Cell 181, 774–783.e5. 10.1016/j.cell.2020.04.033. [PubMed: 32413298]
7. Benson NC, Yoon JMD, Forenzo D, Engel SA, Kay KN, and Winawer J (2022). Variability of the Surface Area of the V1, V2, and V3 Maps in a Large Sample of Human Observers. J. Neurosci 42, 8629–8646. 10.1523/JNEUROSCI.0690-21.2022. [PubMed: 36180226]
8. Blasdel GG and Salama G (1986). Voltage-sensitive dyes reveal a modular organization in monkey striate cortex. Nature 321, 579–585. 10.1038/321579a0. [PubMed: 3713842]
9. Blauch NM, Behrmann M, and Plaut DC (2022). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. Proc. Natl. Acad. Sci. U. S. A 119. 10.1073/pnas.2112566119.
10. Bonhoeffer T and Grinvald A (1991). Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns. Nature 353, 429–431. 10.1038/353429a0. [PubMed: 1896085]
11. Bugatus L, Weiner KS, and Grill-Spector K (2017). Task alters category representations in prefrontal but not high-level visual cortex. Neuroimage 155, 437–449. 10.1016/j.neuroimage.2017.03.062. [PubMed: 28389381]
12. Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, and Ecker AS (2019). Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comput. Biol 15, e1006897. 10.1371/journal.pcbi.1006897. [PubMed: 31013278]
13. Carreira-Perpiñán MA, Lister RJ, and Goodhill GJ (2005). A computational model for the development of multiple maps in primary visual cortex. Cereb. Cortex 15, 1222–1233. 10.1093/cercor/bhi004. [PubMed: 15616135]
14. Chang JT, Whitney D, and Fitzpatrick D (2020). Experience-Dependent Reorganization Drives Development of a Binocularly Unified Cortical Representation of Orientation. Neuron 107, 338–350.e5. 10.1016/j.neuron.2020.04.022. [PubMed: 32428433]
15. Chatterjee S, Ohki K, and Reid RC (2021). Chromatic micromaps in primary visual cortex. Nat. Commun 12, 2315. 10.1038/s41467-021-22488-3. [PubMed: 33875667]
16. Chen BL, Hall DH, and Chklovskii DB (2006). Wiring optimization can relate neuronal structure and function. Proc. Natl. Acad. Sci. U. S. A 103, 4723–4728. 10.1073/pnas.0506806103. [PubMed: 16537428]
17. Chen T, Kornblith S, Norouzi M, and Hinton G (2020). A Simple Framework for Contrastive Learning of Visual Representations. arXiv. 10.48550/arXiv.2002.05709.
18. Chklovskii DB, Schikorski T, and Stevens CF (2002). Wiring optimization in cortical circuits. Neuron 34, 341–347. 10.1016/S0896-6273(02)00679-7. [PubMed: 11988166]
19. Conwell C, Prince JS, Kay KN, Alvarez GA, and Konkle T (2023). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? Preprint at bioRxiv. 10.1101/2022.03.28.485868.
20. De Valois RL, Albrecht DG, and Thorell LG (1982). Spatial frequency selectivity of cells in macaque visual cortex. Vision Res. 22, 545–559. 10.1016/0042-6989(82)90113-4. [PubMed: 7112954]
21. De Valois RL, Yund EW, and Hepler N (1982). The orientation and direction selectivity of cells in macaque visual cortex. Vision Res. 22, 531–544. 10.1016/0042-6989(82)90112-2. [PubMed: 7112953]
22. Del Giudice M. (2021). Effective Dimensionality: A Tutorial. Multivariate Behav. Res 56, 527–542. 10.1080/00273171.2020.1743631. [PubMed: 32223436]
23. Deng J, Dong W, Socher R, Li LJ, Li K, and Fei-Fei L (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition pp. 248–255. 10.1109/CVPR.2009.5206848.

24. Desimone R, Albright TD, Gross CG, and Bruce C (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. J. Neurosci 4, 2051–2062. [PubMed: 6470767]
25. Dobs K, Martinez J, Kell AJE, and Kanwisher N (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. Science Advances 8, eabl8913. 10.1126/sciadv.abl8913. [PubMed: 35294241]
26. Doshi FR and Konkle T (2023). Cortical topographic motifs emerge in a self-organized map of object space. Sci Adv 9, eade8187. 10.1126/sciadv.ade8187. [PubMed: 37343093]
27. Downing PE, Jiang Y, Shuman M, and Kanwisher N (2001). A cortical area selective for visual processing of the human body. Science 293, 2470–2473. 10.1126/science.1063414. [PubMed: 11577239]
28. Durbin R and Mitchison G (1990). A dimension reduction framework for understanding cortical maps. Nature 343, 644–647. 10.1038/343644A0. [PubMed: 2304536]
29. Elmoznino E and Bonner MF (2023). High-performing neural network models of visual cortex benefit from high latent dimensionality. Preprint at bioRxiv. 10.1101/2022.07.13.499969.
30. Epstein R and Kanwisher N (1998). A cortical representation of the local visual environment. Nature 392, 598–601. 10.1038/33402. [PubMed: 9560155]
31. Ferreiro DN, Conde-Ocazionez SA, Patriota JHN, Souza LC, Oliveira MF, Wolf F, and Schmidt KE (2021). Spatial clustering of orientation preference in primary visual cortex of the large rodent agouti. iScience 24, 101882. 10.1016/j.isci.2020.101882. [PubMed: 33354663]
32. Finzi D, Margalit E, Kay K, Yamins DLK, and Grill-Spector K (2023). A single computational objective drives specialization of streams in visual cortex. Preprint at bioRxiv. 10.1101/2023.12.19.572460.
33. Garg AK, Li P, Rashid MS, and Callaway EM (2019). Color and orientation are jointly coded and spatially organized in primate primary visual cortex. Science 364, 1275–1279. 10.1126/science.aaw5868. [PubMed: 31249057]
34. Ge X, Zhang K, Gribizis A, Hamodi AS, Sabino AM, and Crair MC (2021). Retinal waves prime visual motion detection by simulating future optic flow. Science 373. 10.1126/science.abd0830.
35. Gomez J, Barnett M, and Grill-Spector K (2019). Extensive childhood experience with Pokémon suggests eccentricity drives organization of visual cortex. Nat Hum Behav 3, 611–624. 10.1038/s41562-019-0592-8. [PubMed: 31061489]
36. Gomez J, Pestilli F, Witthoft N, Golarai G, Liberman A, Poltoratski S, Yoon J, and Grill-Spector K (2015). Functionally defined white matter reveals segregated pathways in human ventral temporal cortex associated with category-specific processing. Neuron 85, 216–227. 10.1016/j.neuron.2014.12.027. [PubMed: 25569351]
37. Goyal P, Duval Q, Reizenstein J, Leavitt M, Xu M, Lefaudeux B, Singh M, Reis V, Caron M, Bojanowski P, and Others (2021). Vissl.
38. Granley J, Riedel A, and Beyeler M (2022). Adapting Brain-Like Neural Networks for Modeling Cortical Visual Prostheses. arXiv. 10.48550/arXiv.2209.13561.
39. Grill-Spector K and Weiner KS (2014). The functional architecture of the ventral temporal cortex and its role in categorization. Nat. Rev. Neurosci 15, 536–548. 10.1038/nrn3747. [PubMed: 24962370]
40. Grinvald A, Lieke E, Frostig RD, Gilbert CD, and Wiesel TN (1986). Functional architecture of cortex revealed by optical imaging of intrinsic signals. Nature 324, 361–364. 10.1038/324361a0. [PubMed: 3785405]
41. Gross CG, Rocha-Miranda CE, and Bender DB (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. J. Neurophysiol 35, 96–111. 10.1152/jn.1972.35.1.96. [PubMed: 4621506]
42. Gu Y, Lewallen S, Kinkhabwala AA, Domnisoru C, Yoon K, Gauthier JL, Fiete IR, and Tank DW (2018). A Map-like Micro-Organization of Grid Cells in the Medial Entorhinal Cortex. Cell 175, 736–750.e30. 10.1016/j.cell.2018.08.066. [PubMed: 30270041]
43. Guan S-C, Ju N-S, Tao L, Tang S-M, and Yu C (2021). Functional organization of spatial frequency tuning in macaque V1 revealed with two-photon calcium imaging. Prog. Neurobiol 205, 102120. 10.1016/j.pneurobio.2021.102120. [PubMed: 34252470]

44. Güçlü U and van Gerven MAJ (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. J. Neurosci 35, 10005–10014. 10.1523/JNEUROSCI.5023-14.2015. [PubMed: 26157000]
45. Harvey BM, Klein BP, Petridou N, and Dumoulin SO (2013). Topographic representation of numerosity in the human parietal cortex. Science 341, 1123–1126. 10.1126/science.1239052. [PubMed: 24009396]
46. Hasson U, Levy I, Behrmann M, Hendler T, and Malach R (2002). Eccentricity bias as an organizing principle for human high-order object areas. Neuron 34, 479–490. [PubMed: 11988177]
47. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, and Pietrini P (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430. 10.1126/science.1063736. [PubMed: 11577229]
48. He K, Zhang X, Ren S, and Sun J (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 770–778.
49. Henderson M and Serences JT (2021). Biased orientation representations can be explained by experience with nonuniform training set statistics. J. Vis 21, 10. 10.1167/jov.21.8.10.
50. Hubel DH and Wiesel TN (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol 160, 106–154. [PubMed: 14449617]
51. Hübener M, Shoham D, Grinvald A, and Bonhoeffer T (1997). Spatial relationships among three columnar systems in cat area 17. J. Neurosci 17, 9270–9284. [PubMed: 9364073]
52. Humphries C, Liebenthal E, and Binder JR (2010). Tonotopic organization of human auditory cortex. Neuroimage 50, 1202–1211. 10.1016/j.neuroimage.2010.01.046. [PubMed: 20096790]
53. Hyvärinen A, Hoyer PO, and Inki M (2001). Topographic independent component analysis. Neural Comput. 13, 1527–1558. 10.1162/089976601750264992. [PubMed: 11440596]
54. Jacobs RA and Jordan MI (1992). Computational Consequences of a Bias toward Short Connections. J. Cogn. Neurosci 4, 323–336. 10.1162/jocn.1992.4.4.323. [PubMed: 23968126]
55. Kanwisher N, McDermott J, and Chun MM (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci 17, 4302–4311. [PubMed: 9151747]
56. Kaschube M, Schnabel M, Löwel S, Coppola DM, White LE, and Wolf F (2010). Universality in the evolution of orientation columns in the visual cortex. Science 330, 1113–1116. 10.1126/science.1194869. [PubMed: 21051599]
57. Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, and McDermott JH (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron 98, 630–644.e16. 10.1016/j.neuron.2018.03.044. [PubMed: 29681533]
58. Khaligh-Razavi S-M and Kriegeskorte N (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol 10, e1003915. 10.1371/journal.pcbi.1003915. [PubMed: 25375136]
59. Kim J, Song M, Jang J, and Paik S-B (2020). Spontaneous Retinal Waves Can Generate Long-Range Horizontal Connectivity in Visual Cortex. J. Neurosci 40, 6584–6599. 10.1523/JNEUROSCI.0649-20.2020. [PubMed: 32680939]
60. Kohonen T. (1982). Self-organized formation of topologically correct feature maps. Biol. Cybern 43, 59–69. 10.1007/BF00337288.
61. Kong NCL, Margalit E, Gardner JL, and Norcia AM (2022). Increasing neural network robustness improves match to macaque V1 eigenspectrum, spatial frequency preference and predictivity. PLoS Comput. Biol 18, e1009739. 10.1371/journal.pcbi.1009739. [PubMed: 34995280]
62. Konkle T. (2021). Emergent organization of multiple visuotopic maps without a feature hierarchy. Preprint at bioRxiv. 10.1101/2021.01.05.425426.
63. Konkle T and Alvarez GA (2022). A self-supervised domain-general learning framework for human ventral stream representation. Nat. Commun 13, 491. 10.1038/s41467-022-28091-4. [PubMed: 35078981]
64. Konkle T and Oliva A (2012). A real-world size organization of object responses in occipitotemporal cortex. Neuron 74, 1114–1124. 10.1016/j.neuron.2012.04.036. [PubMed: 22726840]

65. Koulakov AA and Chklovskii DB (2001). Orientation preference patterns in mammalian visual cortex: a wire length minimization approach. Neuron 29, 519–527. [PubMed: 11239440]
66. Kriegeskorte N, Mur M, and Bandettini P (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci 2, 4. 10.3389/neuro.06.004.2008. [PubMed: 19104670]
67. Krizhevsky A, Sutskever I, and Hinton GE (2012). ImageNet classification with deep convolutional neural networks. Commun. ACM 60, 84–90. 10.1145/3065386.
68. Kubilius J, Schrimpf M, Kar K, Rajalingham R, Hong H, Majaj N, Issa E, Bashivan P, Prescott-Roy J, Schmidt K, Nayebi A, Bear D, Yamins DL, and DiCarlo JJ (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In Advances in Neural Information Processing Systems 32, Wallach H, Larochelle H, g A, Alche-Buc F, Fox E, and Garnett R, eds. (Curran Associates, Inc.) pp. 12805–12816.
69. Lafer-Sousa R and Conway BR (2013). Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. Nat. Neurosci 16, 1870–1878. 10.1038/nn.3555. [PubMed: 24141314]
70. Lee H, Margalit E, Jozwik KM, Cohen MA, Kanwisher N, Yamins DLK, and DiCarlo JJ (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. Preprint at bioRxiv. 10.1101/2020.07.09.185116.
71. Levy I, Hasson U, Avidan G, Hendler T, and Malach R (2001). Center–periphery organization of human object areas. Nat. Neurosci 4, 533–539. 10.1038/87490. [PubMed: 11319563]
72. Linsker R. (1986). From basic network principles to neural architecture: emergence of orientation columns. Proc. Natl. Acad. Sci. U. S. A 83, 8779–8783. 10.1073/pnas.83.22.8779. [PubMed: 3464981]
73. Livingstone MS and Hubel DH (1984). Anatomy and physiology of a color system in the primate visual cortex. J. Neurosci 4, 309–356. [PubMed: 6198495]
74. Loshchilov I and Hutter F (2016). SGDR: Stochastic Gradient Descent with Warm Restarts. arXiv. 10.48550/arXiv.1608.03983.
75. Lu Z, Doerig A, Bosch V, Krahmer B, Kaiser D, Cichy RM, and Kietzmann TC (2023). End-to-end topographic networks as models of cortical map formation and human visual behaviour: moving beyond convolutions. arXiv. 10.48550/arXiv.2308.09431.
76. Majaj NJ, Hong H, Solomon EA, and DiCarlo JJ (2015). Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. Journal of Neuroscience 35, 13402–13418. 10.1523/JNEUROSCI.5181-14.2015. [PubMed: 26424887]
77. Margalit E, Jamison KW, Weiner KS, Vizioli L, Zhang R-Y, Kay KN, and Grill-Spector K (2020). Ultra-high-resolution fMRI of Human Ventral Temporal Cortex Reveals Differential Representation of Categories and Domains.
78. McCandliss BD, Cohen L, and Dehaene S (2003). The visual word form area: expertise for reading in the fusiform gyrus. Trends Cogn. Sci 7, 293–299. 10.1016/s1364-6613(03)00134-7. [PubMed: 12860187]
79. McKinney W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference, van der Walt S and Millman J, eds. (SciPy). 10.25080/majora-92bf1922-00a.
80. McLaughlin T, Torborg CL, Feller MB, and O'Leary DDM (2003). Retinotopic map refinement requires spontaneous retinal waves during a brief critical period of development. Neuron 40, 1147–1160. 10.1016/s0896-6273(03)00790-6. [PubMed: 14687549]
81. Mehrer J, Spoerer CJ, Jones EC, Kriegeskorte N, and Kietzmann TC (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. Proc. Natl. Acad. Sci. U. S. A 118. 10.1073/pnas.2011417118.
82. Meister M, Wong RO, Baylor DA, and Shatz CJ (1991). Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. Science 252, 939–943. [PubMed: 2035024]
83. Miller KD (1994). A model for the development of simple cell receptive fields and the ordered arrangement of orientation columns through activity-dependent competition between ON- and OFF-center inputs. J. Neurosci 14, 409–441. [PubMed: 8283248]

84. Miller KD, Keller JB, and Stryker MP (1989). Ocular dominance column development: analysis and simulation. Science 245, 605–615. 10.1126/science.2762813. [PubMed: 2762813]

85. Moeller S, Freiwald WA, and Tsao DY (2008). Patches with links: a unified system for processing faces in the macaque temporal lobe. Science 320, 1355–1359. 10.1126/science.1157436. [PubMed: 18535247]

86. Munkres J. (1957). Algorithms for the Assignment and Transportation Problems. Journal of the Society for Industrial and Applied Mathematics 5, 32–38. 10.1137/0105003.

87. Nasr S, Echavarria CE, and Tootell RBH (2014). Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. J. Neurosci 34, 6721–6735. 10.1523/JNEUROSCI.4802-13.2014. [PubMed: 24828628]

88. Nauhaus I, Nielsen KJ, Disney AA, and Callaway EM (2012). Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex. Nat. Neurosci 15, 1683–1690. 10.1038/nn.3255. [PubMed: 23143516]

89. Nayebi A, Sagastuy-Brena J, Bear DM, Kar K, Kubilius J, Ganguli S, Sussillo D, DiCarlo JJ, and Yamins DLK (2022). Recurrent Connections in the Primate Ventral Visual Stream Mediate a Tradeoff Between Task Performance and Network Size During Core Object Recognition. Neural Computation 34, 1652–1675. [PubMed: 35798321]

90. Norman-Haignere SV, Feather J, Boebinger D, Brunner P, Ritaccio A, McDermott JH, Schalk G, and Kanwisher N (2022). A neural population selective for song in human auditory cortex. Curr. Biol 32, 1470–1484.e12. 10.1016/j.cub.2022.01.069. [PubMed: 35196507]

91. Obenhaus HA, Zong W, Jacobsen RI, Rose T, Donato F, Chen L, Cheng H, Bonhoeffer T, Moser M-B, and Moser EI (2022). Functional network topography of the medial entorhinal cortex. Proc. Natl. Acad. Sci. U. S. A 119. 10.1073/pnas.2121655119.

92. Obermayer K, Ritter H, and Schulten K (1990). A principle for the formation of the spatial structure of cortical feature maps. Proc. Natl. Acad. Sci. U. S. A 87, 8345–8349. [PubMed: 2236045]

93. Olshausen BA and Field DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609. 10.1038/381607a0. [PubMed: 8637596]

94. Orlov T, Makin TR, and Zohary E (2010). Topographic representation of the human body in the occipitotemporal cortex. Neuron 68, 586–600. 10.1016/j.neuron.2010.09.032. [PubMed: 21040856]

95. Pinsk MA, Arcaro M, Weiner KS, Kalkus JF, Inati SJ, Gross CG, and Kastner S (2009). Neural representations of faces and body parts in macaque and human cortex: a comparative FMRI study. J. Neurophysiol 101, 2581–2600. 10.1152/jn.91198.2008. [PubMed: 19225169]

96. Pinsk MA, DeSimone K, Moore T, Gross CG, and Kastner S (2005). Representations of faces and body parts in macaque temporal cortex: a functional MRI study. Proc. Natl. Acad. Sci. U. S. A 102, 6996–7001. 10.1073/pnas.0502605102. [PubMed: 15860578]

97. Rajalingham R and DiCarlo JJ (2019). Reversible Inactivation of Different Millimeter-Scale Regions of Primate IT Results in Different Patterns of Core Object Recognition Deficits. Neuron 102, 493–505.e5. 10.1016/j.neuron.2019.02.001. [PubMed: 30878289]

98. Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, Christensen A, Clopath C, Costa RP, de Berker A, Ganguli S, Gillon CJ, Hafner D, Kepecs A, Kriegeskorte N, Latham P, Lindsay GW, Miller KD, Naud R, Pack CC, Poirazi P, Roelfsema P, Sacramento J, Saxe A, Scellier B, Schapiro AC, Senn W, Wayne G, Yamins D, Zenke F, Zylberberg J, Therien D, and Kording KP (2019). A deep learning framework for neuroscience. Nat. Neurosci 22, 1761–1770. 10.1038/s41593-019-0520-2. [PubMed: 31659335]

99. Ringach DL, Mineault PJ, Tring E, Olivas ND, Garcia-Junco-Clemente P and Trachtenberg JT (2016). Spatial clustering of tuning in mouse primary visual cortex. Nat. Commun 7, 12270. 10.1038/ncomms12270. [PubMed: 27481398]

100. Ringach DL, Shapley RM, and Hawken MJ (2002). Orientation selectivity in macaque V1: diversity and laminar dependence. J. Neurosci 22, 5639–5651. 20026567. [PubMed: 12097515]

101. Schaeffer R, Khona M, and Fiete I (2022). No free lunch from deep learning in neuroscience: A case study through models of the entorhinal-hippocampal circuit. Adv. Neural Inf. Process. Syst 35, 16052–16067.

102. Schrimpf M, Blank IA, Tuckute G, Kauf C, Hosseini EA, Kanwisher N, Tenenbaum JB, and Fedorenko E (2021). The neural architecture of language: Integrative modeling converges on predictive processing. Proc. Natl. Acad. Sci. U. S. A 118. 10.1073/pnas.2105646118.
103. Schrimpf M, Kubilius J, Lee MJ, Ratan Murty NA, Ajemian R, and DiCarlo JJ (2020). Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. Neuron 108, 413–423. 10.1016/j.neuron.2020.07.040. [PubMed: 32918861]
104. Shahbazi E, Ma T, Pernus M, Scheirer WJ, and Afraz A (2023). The causal role of the inferior temporal cortex in visual perception. Preprint at bioRxiv. 10.1101/2022.10.24.513337.
105. Stigliani A, Weiner KS, and Grill-Spector K (2015). Temporal Processing Capacity in High-Level Visual Cortex Is Domain Specific. J. Neurosci 35, 12412–12424. 10.1523/JNEUROSCI.4822-14.2015. [PubMed: 26354910]
106. Stringer C, Pachitariu M, Steinmetz N, Carandini M, and Harris KD (2019). High-dimensional geometry of population responses in visual cortex. Nature 571, 361–365. 10.1038/s41586-019-1346-5. [PubMed: 31243367]
107. Swindale NV and Bauer H (1998). Application of Kohonen's self–organizing feature map algorithm to cortical maps of orientation and direction preference. Proceedings of the Royal Society of London B: Biological Sciences 265, 827–838. 10.1098/rspb.1998.0367.
108. The pandas development team (2023). pandas-dev/pandas: Pandas.
109. Tsao DY, Freiwald WA, Tootell RBH, and Livingstone MS (2006). A Cortical Region Consisting Entirely of Face-Selective Cells. Science 311, 670–674. 10.1126/science.1119983. [PubMed: 16456083]
110. Vallat R. (2018). Pingouin: statistics in Python. J. Open Source Softw 3, 1026. 10.21105/joss.01026.
111. van der Grinten M, van Steveninck J. d. R., Lozano A, Pijnacker L, Rückauer B, Roelfsema P, van Gerven M, van Wezel R, Güçlü U, and Güçlütürk Y (2022). Biologically plausible phosphene simulation for the differentiable optimization of visual cortical prostheses. Preprint at bioRxiv. 10.1101/2022.12.23.521749.
112. Vettigli G. (2018). MiniSom: minimalistic and NumPy-based implementation of the Self Organizing Map.
113. Weiner KS, Golarai G, Caspers J, Chuapoco MR, Mohlberg H, Zilles K, Amunts K, and Grill-Spector K (2014). The mid-fusiform sulcus: a landmark identifying both cytoarchitectonic and functional divisions of human ventral temporal cortex. Neuroimage 84, 453–465. 10.1016/j.neuroimage.2013.08.068. [PubMed: 24021838]
114. Weiner KS and Grill-Spector K (2010). Sparsely-distributed organization of face and limb activations in human ventral temporal cortex. Neuroimage 52, 1559–1573. 10.1016/j.neuroimage.2010.04.262. [PubMed: 20457261]
115. Weiner KS and Grill-Spector K (2011). Not one extrastriate body area: using anatomical landmarks, hMT+, and visual field maps to parcellate limb-selective activations in human lateral occipitotemporal cortex. Neuroimage 56, 2183–2199. 10.1016/j.neuroimage.2011.03.041. [PubMed: 21439386]
116. Witthoft N, Nguyen M, Golarai G, LaRocque KF, Liberman A, Smith ME, and Grill-Spector K (2014). Where is human V4? Predicting the location of hV4 and VO1 from cortical folding. Cereb. Cortex 24, 2401–2408. 10.1093/cercor/bht092. [PubMed: 23592823]
117. Wong YC, Kwan HC, MacKay WA, and Murphy JT (1978). Spatial organization of precentral cortex in awake primates. I. Somatosensory inputs. J. Neurophysiol 41, 1107–1119. 10.1152/jn.1978.41.5.1107. [PubMed: 100583]
118. Yamins DL, Hong H, Cadieu C, and DiCarlo JJ (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. Adv. Neural Inf. Process. Syst 26.
119. Yamins DLK and DiCarlo JJ (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci 19, 356–365. 10.1038/nn.4244. [PubMed: 26906502]
120. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, and DiCarlo JJ (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U. S. A 111, 8619–8624. 10.1073/pnas.1403112111. [PubMed: 24812127]

121. Yoshioka T, Levitt JB, and Lund JS (1992). Intrinsic lattice connections of macaque monkey visual cortical area V4. J. Neurosci 12, 2785–2802. [PubMed: 1377236]

122. Zeki S. (1983). Colour coding in the cerebral cortex: the reaction of cells in monkey visual cortex to wavelengths and colours. Neuroscience 9, 741–765. 10.1016/0306-4522(83)90265-8. [PubMed: 6621877]

123. Zhang Y, Zhou K, Bao P, and Liu J (2021). Principles governing the topological organization of object selectivities in ventral temporal cortex. Preprint at bioRxiv. 10.1101/2021.09.15.460220.

124. Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, DiCarlo JJ, and Yamins DLK (2021). Unsupervised neural network models of the ventral visual stream. Proc. Natl. Acad. Sci. U. S. A 118. 10.1073/pnas.2014196118.

## Box 1

A unique advantage of a unified topographic model such as the TDANN is that it can be used to predict the effects of simultaneous spatially-localized stimulation across multiple cortical areas.

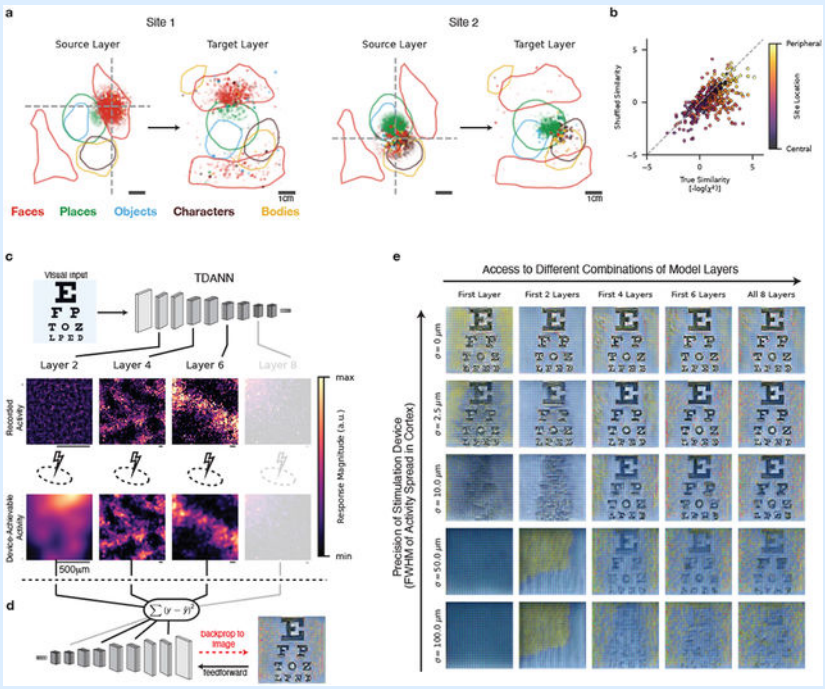
We test this in two scenarios: Electrical stimulation and prototyping a hypothetical multi-region cortical stimulation device. Mirroring results in macaque IT[85], we find that stimulating units in a TDANN face patch drives localized activity in a face patch in the subsequent layer (panels a, b).

Based on recent advances in model-driven prostheses[6,111,38], we simulate a device with two components: 1) a *Stimulation Simulator* that transforms desired activity patterns on the cortical sheet into *device-achievable* patterns, and 2) a *Percept Synthesizer* that visualizes the percept evoked by the stimulation.

Given an image input, the Stimulation Simulator uses the TDANN to predict a spatial pattern of responses in each layer, and then constrains that pattern into one that is physically achievable by a specific hypothetical device (panel c). As a proof-of-principle, we model two such constraints here: spatial precision – the resolution at which the device can create activity patterns, and regional access – the subset of cortical areas that are accessible.

The Percept Synthesizer then visualizes an input image which generates the target activity pattern[38,104] (panel d). Panel e illustrates predicted percepts for cortical stimulation devices with variable precision and access. Unsurprisingly, a device with infinitely high spatial stimulation precision yields sharp percepts even when only early cortical areas are stimulated (panel e, top left). However, percepts quickly deteriorate as the spatial precision of the device decreases (panel e, lower left). Our simulation suggests that at lower precision, the quality of percepts can be improved by adding stimulation of higher cortical areas (panel e, middle rows).

While we have neglected many critical details here, including spatiotemporal processing, cortical magnification, and the need to behaviorally validate percepts, this proof of principle motivates the use of the TDANN to make testable predictions about percepts elicited by cortical stimulation devices.



## Highlights

1. Single model predicts function and spatial structure in early and late visual cortex
2. Best models use self-supervised learning and a scalable spatial constraint
3. More brain-like responses in spatially-accurate than spatially-unconstrained models
4. The local spatial constraint results in lower between-area wiring length

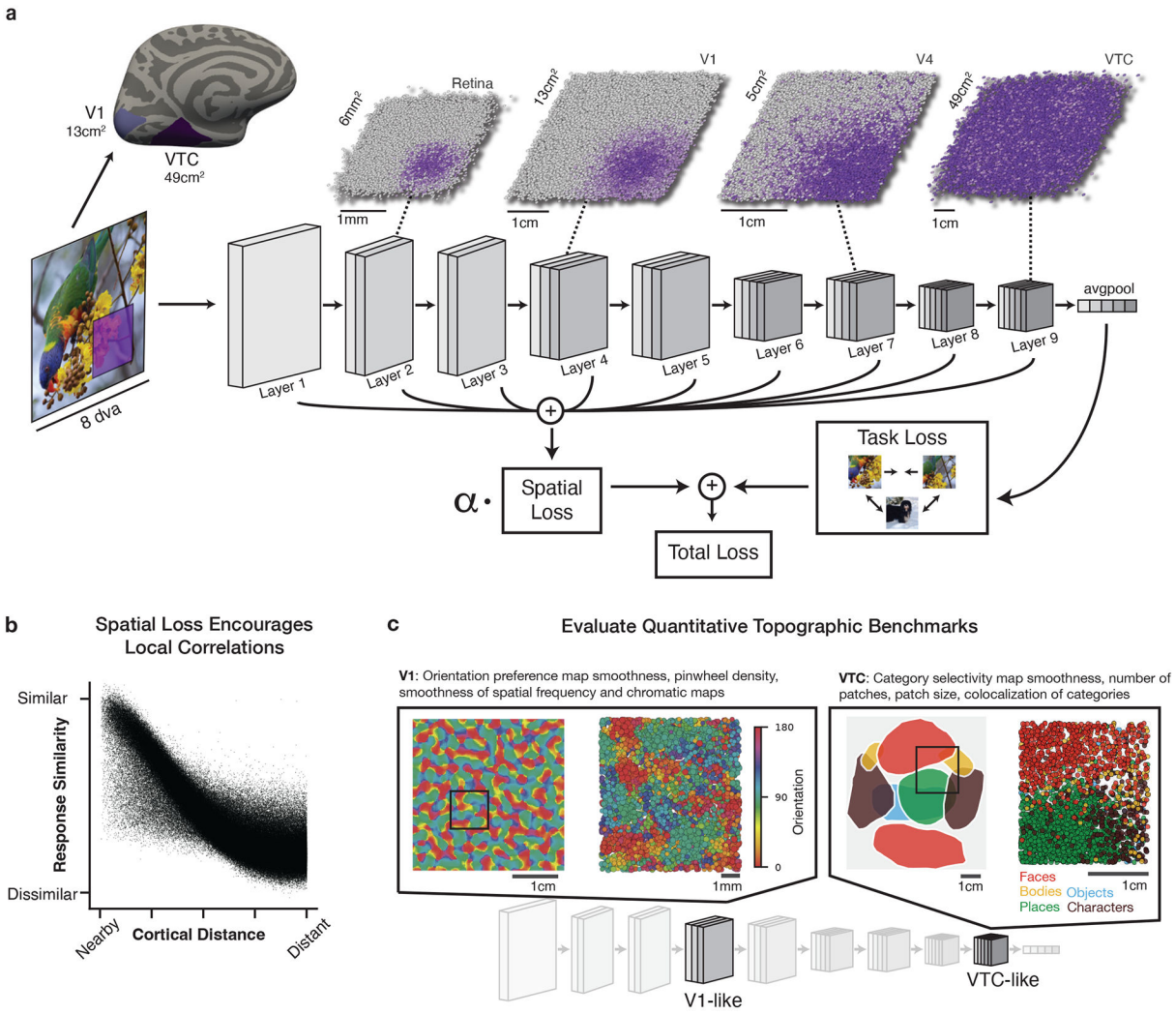



**Figure 1: Constructing a unified model of the functional and spatial constraints of ventral visual cortex.**

**(a)** TDANNs are artificial neural networks whose units are assigned positions in a two-dimensional simulated cortical sheet in each layer. Position assignments are retinotopic, such that location in the cortical sheet corresponds to visual field position. Each dot is one model unit; purple indicates overlap between a unit's receptive field and the purple square on the input image. The TDANN is trained to minimize the sum of a task loss and a spatial loss (SL). $\alpha$ is a free parameter controlling the relative weight of the SL. **(b)** The SL encourages nearby units to develop strong response correlations. Each dot represents the pairwise similarity of responses (y-axis) and cortical distance (x-axis) for a pair of units. **(c)** The TDANN is evaluated on quantitative benchmarks that measure correspondence to topographic features. *Left*: orientation preference map in the V1-like TDANN layer (Figure 2 for details). *Right*: category selectivity map in the VTC-like layer (Figure 3 for details).

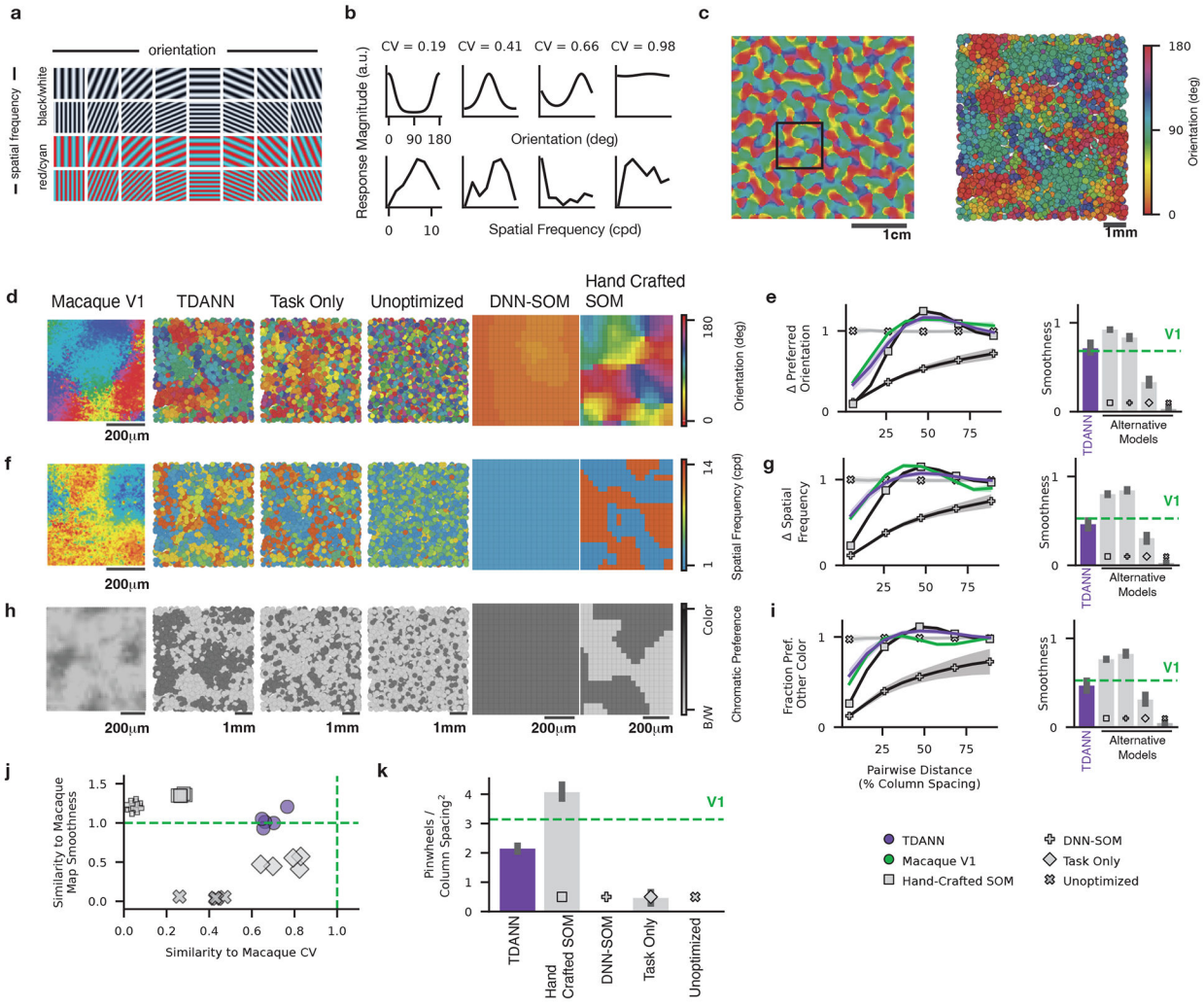



**Figure 2: The TDANN prediction of V1 topography.**

**(a)** Example grating stimuli used to assess tuning for orientation, spatial frequency, and color. **(b)** Tuning curves for orientation (top) and spatial frequency (bottom) for example units in the V1-like layer. **(c)** Smoothed orientation preference map (OPM) in the V1-like layer of the TDANN. Box corresponds to right panel showing individual units labeled by preferred orientation. Results for additional model seeds shown in Supplemental Figure S5. **(d)** OPMs for Macaque V1 (data adapted with permission from Nauhaus et al.[88]), TDANN, and control models: Task Only and Unoptimized neural networks, the DNN-SOM, and Hand Crafted SOM. **(e)** *Left*: Pairwise difference in preferred orientations as a function of pairwise cortical distance, normalized to random-sampling chance level. *Right*: Map smoothness for OPMs in macaque V1 (dashed green line, data from Nauhaus et al.[88]) and four candidate models: the TDANN (purple), the Hand-Crafted self-organizing map (SOM, squares), deep neural network SOM (DNN-SOM, pluses), and Task Only (diamonds). *Error bars:* 95% CI across model seeds and cortical neighborhoods. See Supplemental Figure S3g,h for results from alternative feature spaces. **(f)** Spatial frequency preference, shown for the same region of the TDANN V1-like layer and macaque V1 (data adapted with permission from Nauhaus et al.[88]) as in panel (d). **(g)** Change in preferred spatial

frequency as a function of cortical distance, normalized to chance, for macaque V1 and each model. **(h)** Preference for chromatic stimuli for the same region of the TDANN V1-like layer. *Dark dots:* stronger responses to chromatic than achromatic gratings. *Macaque data:* reconstruction of cytochrome oxidase staining data adapted from Livingstone & Hubel[73] (Copyright 1984 Society for Neuroscience). **(i)** Fraction of units differing in their chromatic preference as a function of cortical distance, normalized to chance. **(j)** Similarity to the smoothness of macaque OPMs (data from Nauhaus et al.[88]) vs. similarity to the distribution of orientation tuning strengths in macaque V1 (data from Ringach et al.[100]) . Duplicate markers indicate different initial model seeds. *Dashed green:* perfect correspondence. **(k)** Density of pinwheels detected in TDANNs, Hand-Crafted SOMs, Task Only models, and Unoptimized models. *Error bars:* CI across model seeds. *Green:* estimated macaque V1 pinwheel density.

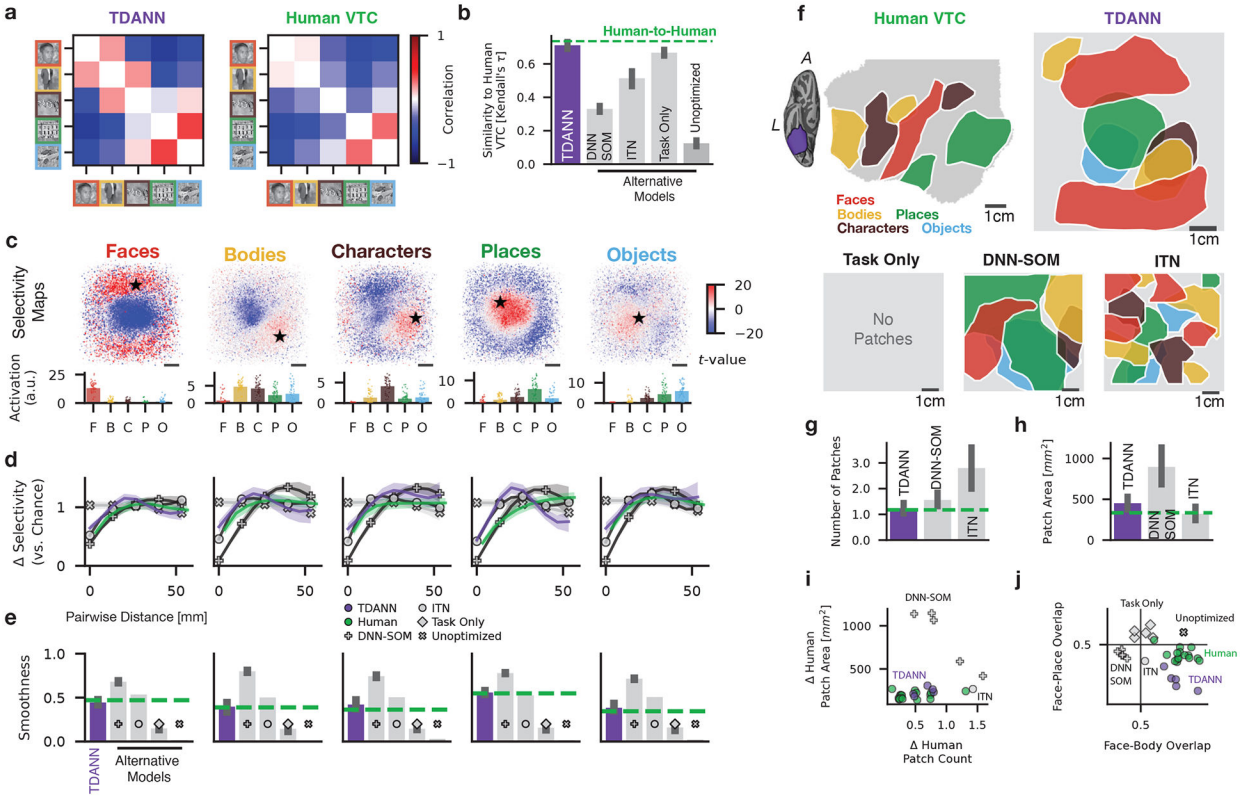



**Figure 3: The TDANN prediction of higher visual cortex topography.**

**(a)** Representational similarity matrices (RSMs) for the TDANN and human VTC, computed across selectivity to five object categories. **(b)** Functional similarity between the TDANN, human VTC, and alternative models, measured as the similarity of RSMs. *Green:* mean of pairwise human-to-human similarity values. **(c)** Selectivity (*t*-value), for each category plotted on the simulated cortical sheet of the VTC-like layer. Responses for an individual unit, marked by black star, plotted below (individual dots: single images, bar height: mean across images). *Scale bar*: 1cm. **(d)** Difference in selectivity as a function of cortical distance for pairs of units in each of five candidate models: the TDANN (purple), deep neural network self-organizing map (DNN-SOM; plus markers), interactive topographic network ("ITN", Blauch et al.[9]; circles), Unoptimized ("x" markers), and Task Only (diamonds). Curves normalized to random-sampling chance. *Green*: Subject-average human data. *Shaded regions*: 95% CI across different unit subsets from models trained with different initial seeds. **(e)** Smoothness of selectivity maps for each category and model. *Dashed green*: human mean. **(f)** Category-selective patches for an example hemisphere in human ventral temporal cortex (VTC; see left inset for location, A: Anterior, L: Lateral), the TDANN, Task Only, DNN-SOM, and "ITN" models. ITN maps adapted with permission from Blauch et al.[9]. Examples from different seeds in Supplemental Figure S5. **(g)** Average number of category-selective patches for each model and human data (dashed green). ANOVA for patch count difference: $F(5, 179) = 32.7$, $p < 10^{-22}$; Significant difference between VTC and ITN ($p = 1.2 \times 10^{-5}$, Post-hoc Tukey's test). **(h)** Surface area of category-selective patches. Same plotting conventions as in (g). ANOVA for patch area

difference: $F(5, 187) = 15.4$, $p < 10^{-11}$; Significant difference between VTC and DNN-SOM ($p < 10^{-10}$, Post-hoc Tukey's test). **(i)** Each human subject and model instance compared to the mean patch area and patch number in the human data. **(j)** Overlap between face- and body-selectivity vs. overlap between face- and place-selectivity, for each human hemisphere (green dots), each TDANN (purple dots), the ITN (gray dot), each DNN-SOM (gray plus signs), and each Task Only model (gray diamonds).

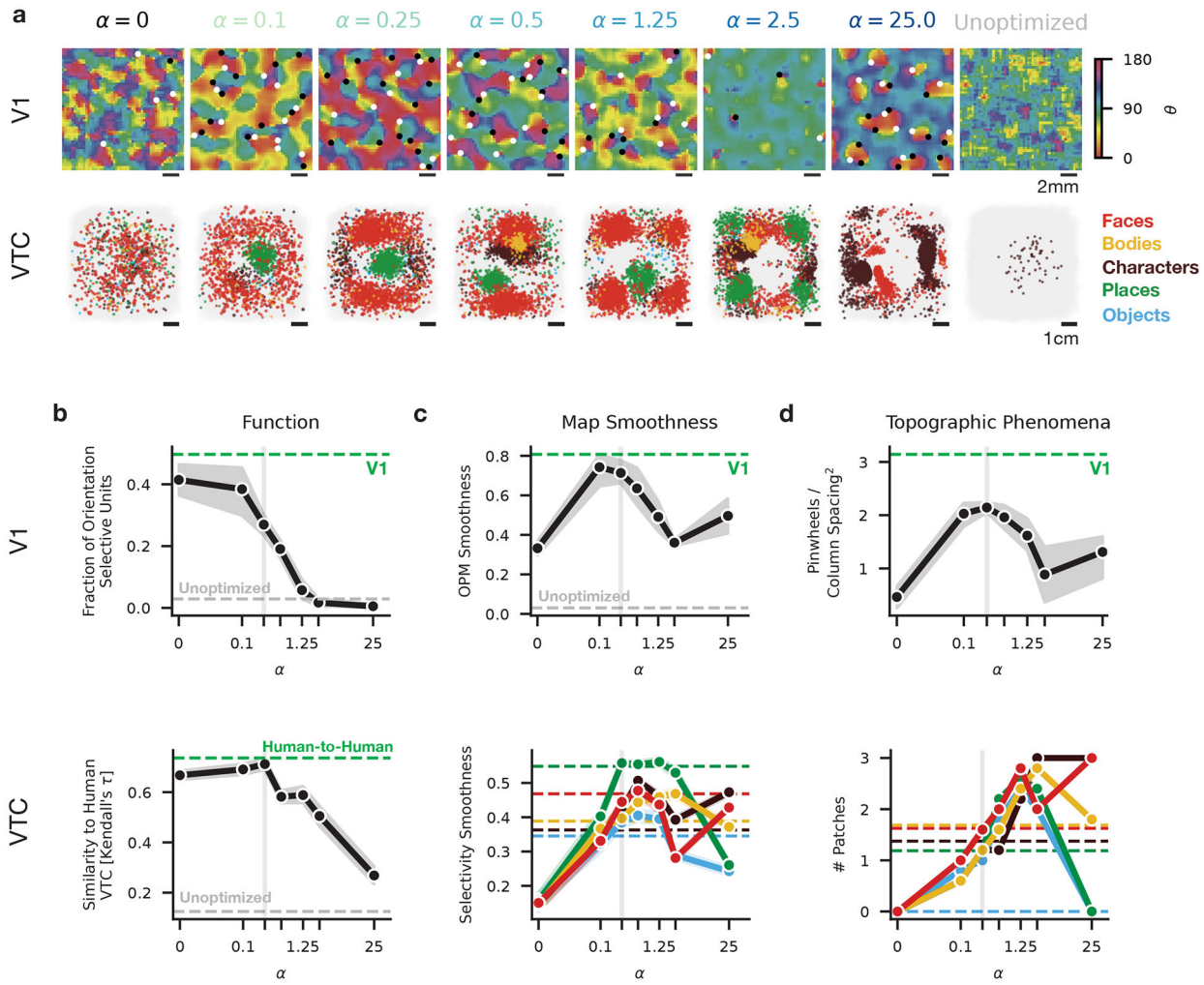



**Figure 4: Convergence of benchmarks indicates balance between functional and spatial constraints.**

**(a)** Topographic maps in the V1-like (top) and VTC-like layer (bottom) of TDANNs trained at different levels of the spatial weight $\alpha$. *Top*: Orientation map structure and pinwheels apparent at $0.1 < \alpha < = 1.25$. *Dots:* estimated pinwheel locations; black: clockwise, white: counterclockwise. *Bottom:* Category selective units ($t > 12$) colored by preferred category. **(b)** Functional correspondence to neural data as a function of $\alpha$. *Top:* Fraction of strongly orientation-selective units (circular variance $\leq 0.6$) in the V1-like layer. *Dashed green:* macaque V1 (from Ringach et al.[100]). *Dashed gray*: mean for Unoptimized models. *Shaded regions:* 95% CI across initial seeds. *Bottom:* Representational similarity between the VTC-like layer and human VTC (as in Figure 3). *Shaded region:* 95% CI across model seeds and human hemispheres. *Vertical line* ($\alpha = 0.25$): value used in prior figures. **(c)** Topographic map smoothness as a function of $\alpha$. *Top:* OPM smoothness in the V1-like layer. *Dashed green:* value in macaque V1. *Dashed gray*: smoothness in an Unoptimized model. *Bottom:* Category selectivity map smoothness in the VTC-like layer. *Dashed lines:* means across human hemispheres from the NSD for each category. **(d)** Density of topographic phenomena

as a function of $\alpha$. *Top:* Pinwheel density in OPMs from the V1-like layer. *Bottom:* Number of category selective patches in the VTC-like layer. *Dashed lines:* Human data.

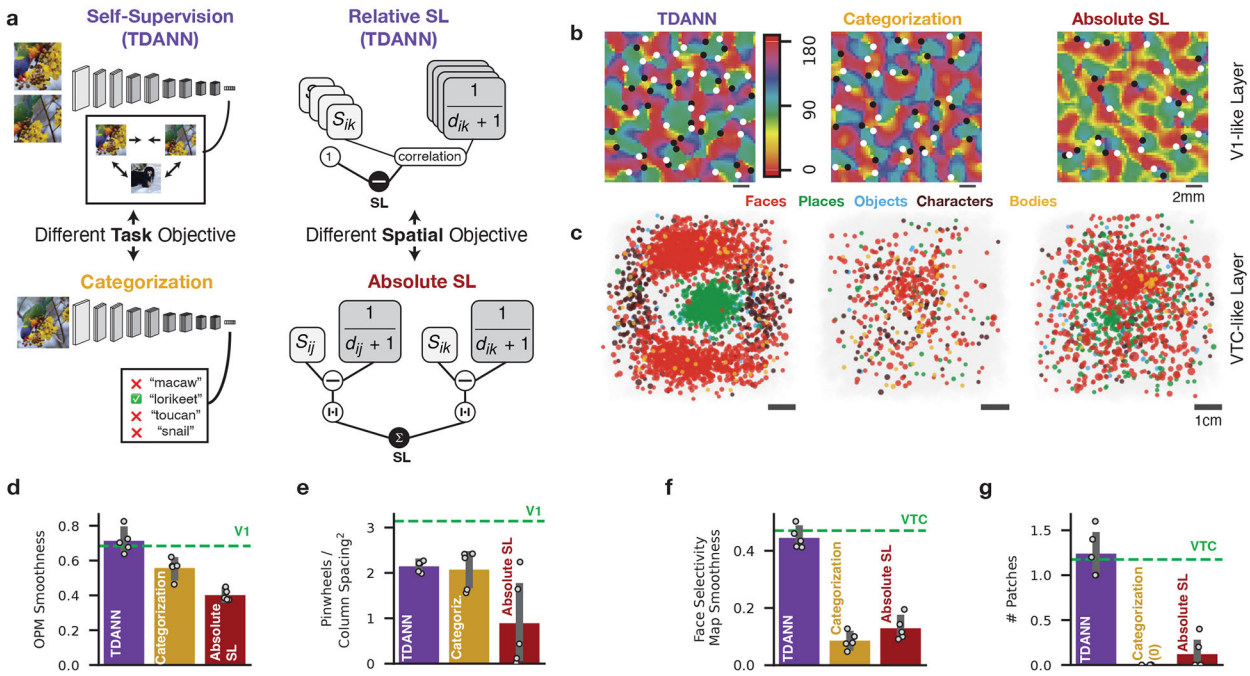



**Figure 5: Self-supervision and scalable spatial constraints underlie the emergence of functional organization.**

In all panels: *purple:* TDANN, *gold*: Categorization-trained, *red*: Absolute SL, and *green:* neural data. **(a)** *Left*: comparison of task objectives. Contrastive self-supervision (top) encourages high similarity for representations of two views of the same image, and low similarity for two views of different images.. Categorization (bottom) compares predicted class probabilities to the human-labeled correct class. *Right*: comparison of spatial objectives. $S_{ij}$: response similarity of units $i$ and $j$. $d_{ij}$: cortical distance between units $i$ and $j$. The TDANN uses the Relative SL (top), which correlates the population of response similarities and pairwise inverse distances across pairs of units. Prior work[70] used the Absolute SL (bottom), which directly subtracts inverse cortical distance from response similarity magnitude. **(b)** Smoothed orientation preference maps (OPMs) in the V1-like layer of the TDANN (left), a Categorization trained model (middle), and a model trained with the Absolute SL (right). *Dots:* pinwheels. $\alpha = 0.25$ for each model. **(c)** Category selective units in the VTC-like layer of each model. **(d)** *Right*: Smoothness of OPMS in the V1-like layer of each model (bars) and macaque V1 (dashed line). **(e)** Density of detected pinwheels in models (bars) and macaque V1 (line). **(f)** *Right*: Smoothness of face selectivity maps in the VTC-like layer of each model (bars) and human VTC (line). **(g)** Number of category-selective patches in the VTC-like layer of each model (bars) and human VTC (line).

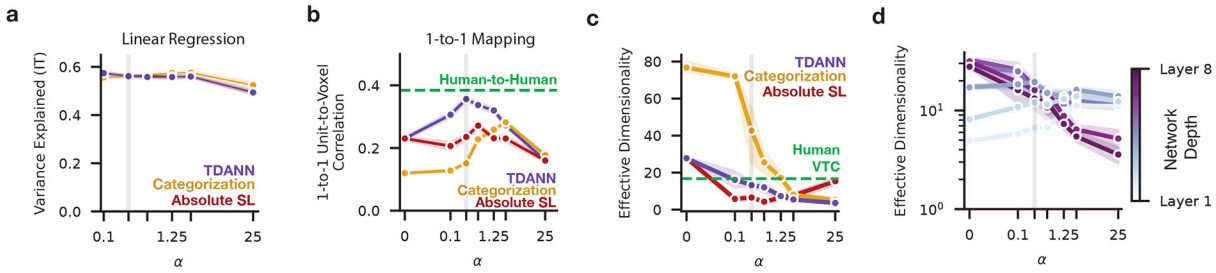



**Figure 6: Spatial constraints make learned representations more brain-like and reduce intrinsic dimensionality**

**(a-c)** metrics as a function of spatial loss weight $\alpha$ and training objective. **(a)** Variance explained under a linear regression mapping between model units and macaque IT neurons. All fits in Supplemental Figure S8d. **(b)** Mean correlation between model units and VTC voxels under a one-to-one mapping. *Green:* mean human-to-human correlation under the same one-to-one mapping. **(c)** Estimated effective dimensionality (cf. Elmoznino & Bonner[29], Del Giudice[22]) of the population response in the VTC-like layer of each model. *Green:* mean ED in human VTC estimated from the NSD dataset. **(d)** Effective dimensionality in the TDANN across all layers and levels of $\alpha$. *Shaded vertical bars:* $\alpha = 0.25$, demonstrated in prior analyses to best match topographic phenomena.

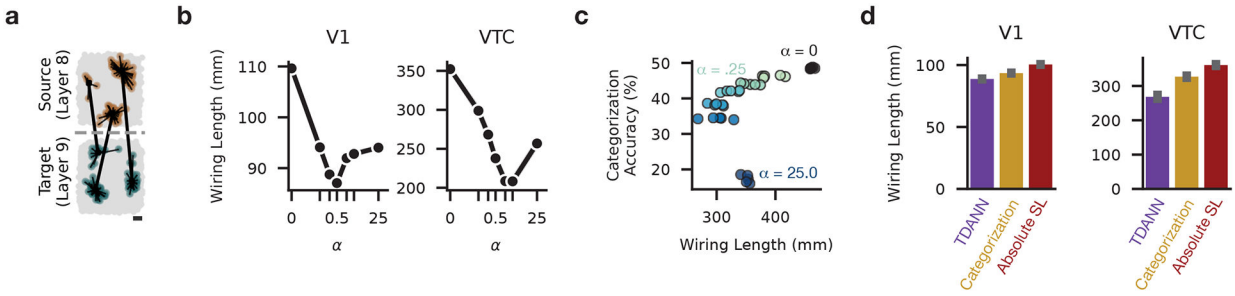



**Figure 7: Minimization of between-area wiring length in models with brain-like functional organization.**

**(a)** Example wiring length computation between adjacent layers. *Brown dots:* top 5% most active units in the Source layer for an arbitrarily-selected natural image. *Green dots:* top 5% most active units in the Target layer. *Black dots:* termination points of virtual fibers that would be required to connect active populations of units. **(b)** Wiring length between layers 4 and 5 (V1-like; left), and layer 8 and 9 (VTC-like, right) as a function of $\alpha$. *Shaded regions:* 95% CI of measurements from different cortical neighborhoods, model seeds, and input images. **(c)** Accuracy on object categorization vs. wiring length; each dot, different $\alpha$. **(d)** Wiring length of models trained with different tasks and spatial objectives ($\alpha = 0.25$ for all). *Error bar*: 95% CI over images and model seeds.

**Table 1.**

Parameters for layer positions. *the value of 1.6mm used in the V1-like layer is known to be inaccurate, but matching the proper value yields too few units in each cortical neighborhood to compute pairwise distances. See Supplemental Figure S3d-f for a discussion and solution to this problem.

| Layer | # Units | Size of Cortical sheet | Neighborhood Size | Region |
|---|---|---|---|---|
| Layer 2 | 200704 | $5.7 mm^2$ | $47 \mu m$ | Retina |
| Layer 3 | 200704 | $5.7 mm^2$ | $47 \mu m$ | Retina |
| Layer 4 | 100352 | $13.5 cm^2$ | $1.6 mm$* | V1 |
| Layer 5 | 100352 | $13.5 cm^2$ | $1.6 mm$* | V1 |
| Layer 6 | 50176 | $12 cm^2$ | $4 mm$ | V2 |
| Layer 7 | 50176 | $5 cm^2$ | $2.5 mm$ | V4 |
| Layer 8 | 25088 | $49 cm^2$ | $31 mm$ | VTC |
| Layer 9 | 25088 | $49 cm^2$ | $31 mm$ | VTC |

**Table 2.**

Patch detection parameters for human VTC and each model.

| Model | Selectivity Threshold | Smoothing $\sigma$ | Minimum Size square mm | Maximum Size square mm |
|---|---|---|---|---|
| Human VTC | 4 | None | 100 | None |
| TDANN | 2 | 2.4 | 100 | 4500 |
| ITN | 8 | 0.7 | 100 | 4500 |
| DNN-SOM | 10 | 2.4 | 100 | 4500 |

## Key Resources Table

| Resource | Source | Identifier |
|---|---|---|
| **Deposited Data** | | |
| Model Weights | This paper | https://osf.io/64qv3/ |
| Sine Grating Images | This paper | https://osf.io/64qv3/ |
| **Software and Algorithms** | | |
| Model Training and Evaluation Code | This paper | https://github.com/neuroailab/TDANN (DOI: 10.5281/zenodo.10294203) |
| **Other** | | |
| VISSL | Goyal et al. | https://github.com/facebookresearch/vissl |
| fLoc Images | Stigliani et al., 2015 | https://github.com/VPNL/fLoc |
| NSD Data and Stimuli | Allen et al., 2022 | https://naturalscenesdataset.org |