A single computational objective drives specialization of streams in visual cortex

Dawn Finzi[1,2,✉], Eshed Margalit[3], Kendrick Kay[4*], Daniel L. K. Yamins[1,2,5,*], and Kalanit Grill-Spector[1,5,*]

[1] Department of Psychology, Stanford University, Stanford, CA 94305

[2] Department of Computer Science, Stanford University, Stanford, CA 94305

[3] Neurosciences Graduate Program, Stanford University, Stanford, CA 94305

[4] Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN 55455

[5] Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA 94305

[*] co-senior author

Human visual cortex is organized into dorsal, lateral, and ventral streams. A long-standing hypothesis is that the functional organization into streams emerged to support distinct visual behaviors. Here, we use a neural network-based computational model and a massive fMRI dataset to test how visual streams emerge. We find that models trained for stream-specific visual behaviors poorly capture neural responses and organization. Instead, a self-supervised Topographic Deep Artificial Neural Network, which encourages nearby units to respond similarly, successfully predicts neural responses, spatial segregation, and functional differentiation across streams. These findings challenge the prevailing view that streams evolved to separately support different behaviors, and suggest instead that functional organization arises from a single principle: balancing general representation learning with local spatial constraints.

processing streams | topography | vision | neural network

Correspondence: *dfinzi@stanford.edu*

Confronted by the blooming, buzzing confusion of the world around us, we perform a diverse range of computations on our visual inputs, including rapidly and accurately identifying objects in our surroundings, their locations, and their actions. The human brain is accordingly thought to be divided into three processing streams, beginning in early visual cortex and ascending through multiple areas to form different processing pathways: (1) a "what" Ventral stream ascending from early visual cortex to the inferior aspects of occipital and temporal cortices [1, 2], (2) a "where" or "visually-guided grasping" Dorsal stream extending superiorly along occipito-parietal cortex [1, 2], and (3) a Lateral stream, extending through lateral occipitotemporal cortex to the superior temporal sulcus (STS), thought to be involved in dynamic perception [3], particularly of actions [4], and social information [5, 6].

The prevailing hypothesis (*multiple behavioral demands hypothesis*) suggests that the organization into streams is an outcome of evolutionary optimization for independent visual behaviors that can be done in parallel with dedicated neural machinery, yielding a fast and efficient visual system [1, 7, 8, 9, 10, 11, 12]). An alternative hypothesis suggests that a set of physical constraints, such as wiring length, could produce the functional organization of the brain into streams (*spatial constraints hypothesis*). According to Nelson and Bower [13], "if the brain's estimated 10^{11} neurons were placed on the surface of a sphere and fully interconnected by individual axons 0.1 µm in radius, the sphere would have to have a diameter of more than 20 km to accommodate the connections." As a result of physical constraints and a need for fast processing, there is a known bias in the brain toward short-range connections [14, 15]. One way to minimize wiring length locally is to encourage nearby neurons to respond similarly [16]. Indeed, locally correlated responses are evident in the many cortical topographic maps, such as maps of the visual field [17, 18]. From an information theoretic standpoint, positioning neurons frequently involved in processing related information close together also makes neural processing faster and more efficient [16, 19]. Thus, we ask: is the functional organization of visual cortex into streams due to optimization for multiple distinct behaviors, or due to balancing spatial and information constraints?

To test these hypotheses, we use a Deep Artificial Neural Network (DANN) approach. We instantiate the multiple behavioral demands hypothesis by training three different models, each using a state-of-the-art DANN that is trained using supervision on the stream-specific visual behavior: Dorsal: object detection [22], Lateral: action recognition [23], Ventral: object categorization [24] (Fig. 1a). Each model is trained separately to encourage maximum differentiation of the learned representations. We instantiate the spatial constraints hypothesis using a topographic DANN (TDANN, [20, 25]), in which model units in each layer are assigned a position in a 2D simulated cortical sheet, and during training a spatial constraint is balanced together with contrastive self-supervised learning (SimCLR [26], Fig. 1b). The spatial constraint encourages nearby units to have more correlated responses than distant units, and SimCLR encourages two snapshots of the same image (differing in incidental properties such as color or field of view) to have similar representations that are distinct from others [26]. We choose SimCLR because it is one of the best performing self-supervised approaches, and because it generates broadly useful representations that are beneficial for a range of visual tasks [27].

To evaluate whether the multi-behavior models or the TDANNs better predict brain responses to visual stimuli, as well as the cortical organization into different streams, we leverage the Natural Scenes Dataset (NSD) [28]. NSD is a massive, high-resolution, fMRI dataset that measured responses to tens of thousands of natural images across eight individuals. The same images are given as input to each of the candidate models. By comparing model and brain responses on these images, we can test which model best predicts cortical responses and spatial segregation into visual streams.

A major challenge in comparing computational models to the brain lies in establishing a mapping between model representations and brain representations. We develop a new algorithm that estimates an optimal 1-to-1 mapping between model units and voxels. The algorithm matches each model unit to a voxel by finding pairings that have the highest response correlation using an iterative version of the Kuhn-Munkres algorithm [21] with an additional spatial prior (Fig. 1c, Alg. 1) and has several appealing features. First, this 1-to-1 mapping allows us to evaluate topographic organization, unlike typical approaches that either match a linear combination of model units to a brain voxel (linear regression, [11, 29]) or examine the distributed representational similarity structure across units/voxels [30], thus obscuring the topographic organization. Second, a unit in a neural network model abstracts neural computations; as such, it may be a good model for the aggregated neural response of a voxel (i.e., in the Goldilocks zone of computational abstraction [31]). Third, a 1-to-1 mapping provides a more stringent test of models [32, 33], allowing a more rigorous evaluation of equivalences between DANNs and brains [34].

The multi-behavior models fail to capture the functional organization of cortex into streams, while the TDANN provides a much better match

To evaluate whether models match the functional organization into streams, we quantify the spatial and functional correspondence between candidate models and the NSD across subjects, hemispheres, and model seeds,

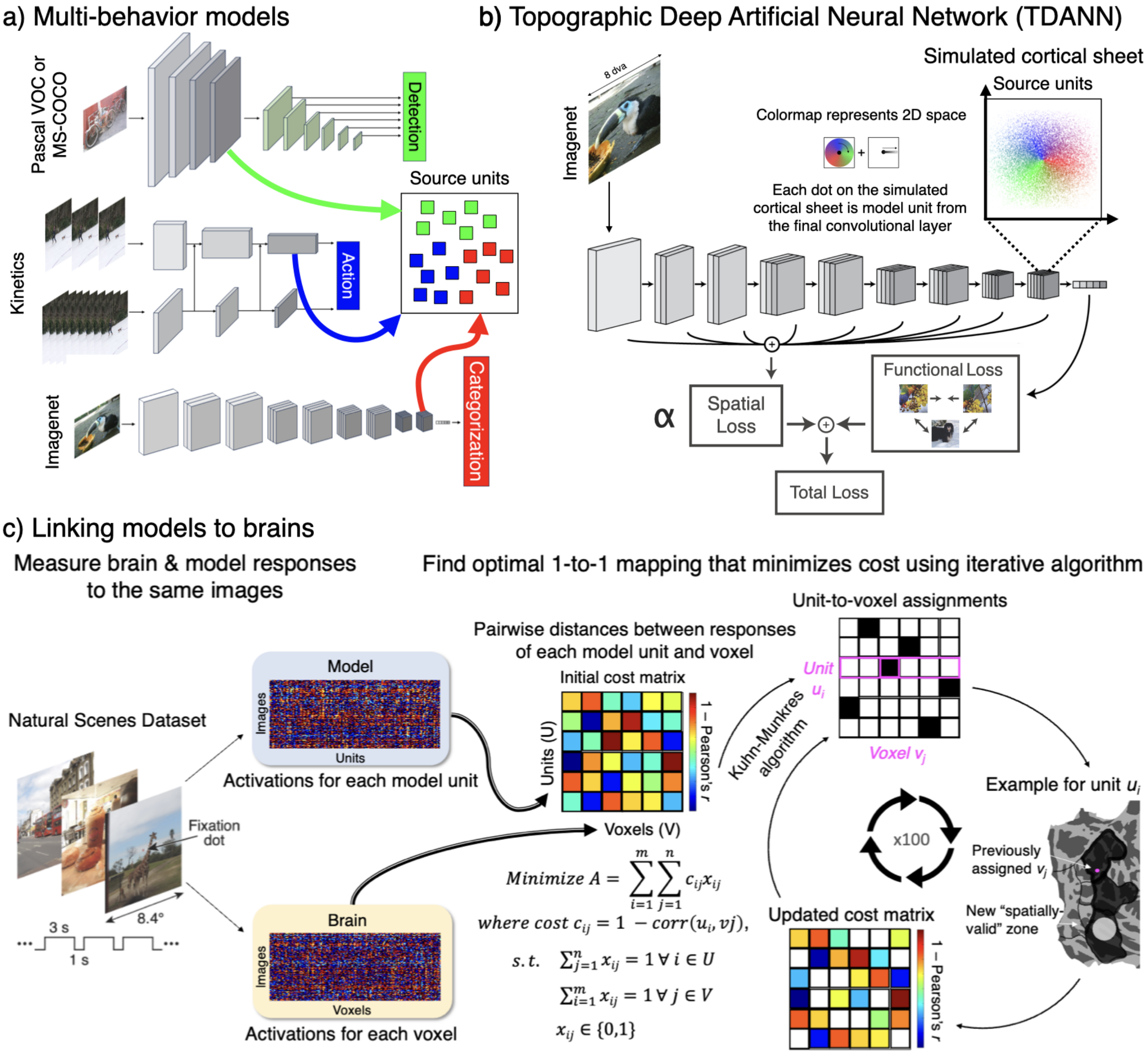
a) Multi-behavior models
Pascal VOC or MS-COCO
Detection
Source units
Kinetics
Action
Imagenet
Categorization
b) Topographic Deep Artificial Neural Network (TDANN)
Imagenet
8 dva
Simulated cortical sheet
Source units
Colormap represents 2D space
Each dot on the simulated cortical sheet is model unit from the final convolutional layer
α
Spatial Loss
Functional Loss
Total Loss
c) Linking models to brains
Measure brain & model responses to the same images
Find optimal 1-to-1 mapping that minimizes cost using iterative algorithm
Natural Scenes Dataset
Fixation dot
8.4°
3 s
1 s
Model
Images
Units
Activations for each model unit
Brain
Images
Voxels
Activations for each voxel
Pairwise distances between responses of each model unit and voxel
Initial cost matrix
Units (U)
Voxels (V)
1 – Pearson's r
Kuhn-Munkres algorithm
Unit-to-voxel assignments
Unit u_i
Voxel v_j
x100
Updated cost matrix
1 – Pearson's r
Example for unit u_i
Previously assigned v_j
New "spatially-valid" zone
$Minimize\ A = \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}x_{ij}$
$where\ cost\ c_{ij} = 1 - corr(u_i, vj),$
$s.t.\ \sum_{j=1}^{n} x_{ij} = 1\ \forall\ i \in U$
$\sum_{i=1}^{m} x_{ij} = 1\ \forall\ j \in V$
$x_{ij} \in \{0,1\}$

Figure 1. To test competing theories, we use two candidate model classes: **(a) Multi-behavior models: object detection, action recognition, & object categorization.** Units are sampled from the final convolutional layer of the backbone architecture for all models and pooled to create candidate source units that are mapped to the brain. **(b) Topographic Deep Artificial Neural Networks (TDANNs)** developed in [20] contain units that are assigned positions on a simulated cortical sheet prior to training and are trained to minimize the sum of a functional loss and a spatial loss, controlled by a free parameter α. Units from the final convolutional layer are used as candidate source units for mapping to the brain. **(c) Overview of 1-to-1 mapping approach for linking model to brain.** 10,000s of images of natural scenes were presented to 8 individuals and candidate models. Responses are extracted from model units and brain voxels, and correlations between each unit-voxel pair computed. Correlations are transformed into an initial cost matrix $(1 - \mathrm{correlation})$. Using the Kuhn-Munkres optimization algorithm [21] we determine an initial assignment such that each unit is assigned to a unique voxel and the average cost across all unit-voxel pairings is minimized (black: assignment). To promote general smoothness in the mapping, neighboring units' assigned voxels are used to calculate "spatially-valid zones", such that any voxels outside a unit's zone are set to have a prohibitive cost (indicated in white, see Alg. 1 for details). This updated cost matrix is used to redetermine assignments, repeated for 100 iterations.

evaluating 1216 model-to-brain mappings in total. Using the unique images seen by each participant, we first find the model-to-brain mapping between units in the model's convolutional end-layer and voxels in anatomical regions of interest (ROIs) corresponding to the ends of each stream. We assess spatial correspondence qualitatively on the cortical sheet (Fig. 2a) and quantitatively by computing how many units are mapped into the corresponding stream (e.g. object categorization and Ventral, see Methods: Evaluating spatial correspondence for details; Fig. 2b). We assess functional correspondence by evaluating the average correlation between unit responses and voxel responses on an independent set of 515 left-out images seen by all subjects (Fig. 2c). We hypothesize that the best model for the human brain is another human brain. That is, the between-subject spatial and functional correspondence can serve as a benchmark to evaluate shared organizational principles [35]. Thus, to estimate a noise ceiling reference point, we use the same 1-to-1 mapping algorithm to map from one subject's brain to another subject's brain (brain-to-brain noise ceiling; gray bars).

Our analyses reveal that the multi-behavior model does not explain the organization of cortex into streams. We find

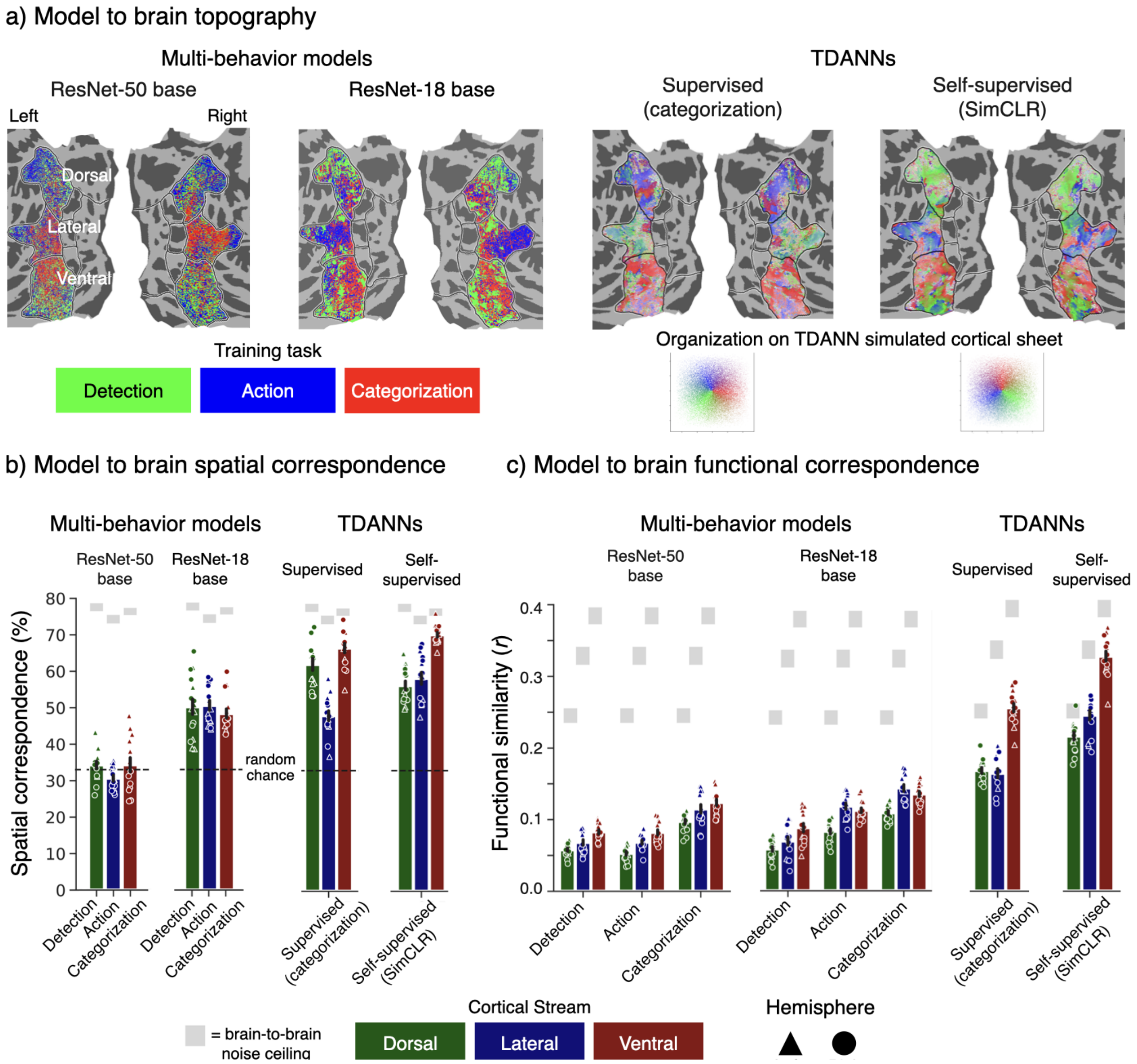

Figure 2. The TDANN better matches spatial and functional organization of human visual system into streams (a) Unit-to-voxel mapping on an example subject flattened cortical surface. Left: multi-behavior models, each voxel is colored based on the training task for its assigned unit. Right: TDANN models, voxels are colored by location on the simulated cortical sheet of the last convolutional layer; polar angle (red -> green -> blue) and eccentricity (opacity, which decreases away from center). **(b)** Quantification of spatial correspondence. For multi-behavior models, we show the percentage of voxels for each stream that are assigned to units from the model trained on their hypothesized task. For TDANNs, we show the percentage of voxels for each stream are assigned to the corresponding third of the simulated cortical sheet (max correspondence, up to rotation and reflection). Each symbol represents a model-to-subject mapping. **(c)** Quantification of functional correspondence. Unit-to-voxel correlations on the test set (left-out 515 images). For (b) and (c) error bars: mean across subjects and hemispheres ± SE. Self-supervised and supervised TDANNs shown for an optimal weighting of the spatial loss ($\alpha = 0.25$ for self-supervised, $\alpha = 2.5$ for supervised).

that neither highly performant models based on ResNet-50 [24, 23, 22], nor models based on ResNet-18 matching the TDANN architecture [36, 20, 25], recapitulate stream organization (Fig. 2). While we expected that a higher proportion of units from a model implementing the visual behavior associated with a stream would be assigned to the corresponding stream than other streams (e.g., units trained on object categorization would be primarily assigned to the Ventral stream), we instead find that unit-to-voxel assignments for the ResNet-50 models are noisy (Fig. 2a) with spatial correspondence not different from chance across all three streams ($ps > .2$), except Lateral which is significantly lower than chance ($t(7) = -4.9$, $p = 0.002$; Fig. 2b). Functionally, unit-voxel correspondences (Fig. 2c) are also poor: below $r = 0.13$ for all streams and tasks. Surprisingly, there is no dissociation of stream by visual behavior, with units trained on object categorization yielding the highest correlations to brain responses across all three streams. While a multi-behavior model based on a shallower ResNet-18 architecture provides significantly better spatial correspondence across all three streams and significantly better functional correspondence in Dorsal and Lateral than the ResNet-50 model (all $ps \leq .018$, Supplemental Tables 1 and 2), its correspondence remains lacking. Spatially, only about 50% of the units are assigned to their correct stream. And functionally, correspondence remains low (at or below $r = 0.15$ for all streams and tasks), with again no dissociation of stream by visual behavior, as units trained on object categorization best match brain responses across all three streams.

In contrast to the multi-behavior models, the self-supervised TDANN captures both stream function and topography. Three distinct clusters of TDANN units map to the three different streams, qualitatively more smoothly on the cortical sheet than multi-behavior models (Fig. 2a). Quantitatively, the self-supervised TDANN achieves significantly higher spatial correspondence than the multi-behavior models across all three streams (all $ps \leq .012$, Supplemental Table 1, Fig. 2b). Additionally, functional correspondence of model units from the self-supervised TDANN to cortex is significantly higher (Fig. 2c, Supplemental Table 2), almost doubling from the best multi-behavior ResNet-18 models across streams (improvement, Dorsal: mean 91% increase, Lateral: 65%, Ventral: 133%). Notably, the functional correspondence approaches the brain-to-brain noise ceiling (Fig. 2c-gray bars) in both the Dorsal and Ventral streams, as does the spatial correspondence in the Ventral stream (Fig. 2b-gray bars).

As the categorization task yields the best spatial and functional match among multi-behavior models, we also implement a TDANN trained on object categorization. This TDANN achieves significantly better spatial correspondence in the Dorsal and Ventral streams and better functional correspondence across all three streams than the corresponding ResNet-18 trained on categorization (all $ps \leq .044$, Supplemental Tables 1 and 2). This difference between the object categorization TDANN and the standard object categorization model is particularly striking for the Ventral stream, suggesting that not only does the spatial constraint in the TDANN change the layout of units, but it also changes their response properties to be more “brain-like". Nonetheless, across models tested, the self-supervised TDANN provides the best functional and spatial match to the brain. The success of the TDANN in matching both functional and spatial brain organization, above and beyond the multi-behavior models, suggests a new explanation of why the brain is organized into visual processing streams.

Both contrastive self-supervision and the spatial constraint during training are critical for functional organization into streams

We next ask what factors contribute to the emergence of streams in TDANNs. The TDANN has two key components that may affect its performance: the training task and the relative strength of the spatial constraint. Thus, we test TDANN models (5 seeds each), trained with either supervised categorization or self-supervised SimCLR, across a range of spatial weightings (α) from $\alpha = 0$, where the model is essentially a standard ResNet-18 minimizing only the task loss, to $\alpha = 25$, at which point the task is dwarfed by the spatial constraint. Models are evaluated on both spatial (Fig. 3B-top panel) and functional (Fig. 3B-bottom panel) correspondence to the brain. As TDANNs contain simulated cortical sheets, we evaluate the model-to-brain spatial correspondence using a distance similarity metric that quantifies the similarity between the spatial topography of the model and that of the brain.

Across all three streams, self-supervised TDANNs with a spatial weight $0.25 \leq \alpha \leq 0.5$ provide the best spatial and functional match to the brain. The clearest stream structure is evident for a self-supervised TDANN with $\alpha = 0.25$, with each stream largely mapping to a distinct contiguous third of the simulated cortical sheet (Fig. 3a); this structure is also visible using a continuous spatial gradient without pre-assigning voxels into streams (Supplemental Fig. S4). Notably, there are significant and large gaps between self-supervised (purple) and supervised (gold) TDANNs in their spatial and functional correspondence to the brain. Additionally, correspondence significantly varies with the level of the spatial constraint and there is a significant interaction between the spatial constraint and training task (Supplemental Tables 3 to 8). Self-supervised TDANNs achieve peak functional and spatial correspondence at $0.25 \leq \alpha \leq 0.5$ (Fig. 3b). While a commonly-used, less strict mapping of model-units-to-brain functional correspondence using linear regression estimates a higher functional correspondence, it critically masks the effects of training task and spatial constraint (Fig. 3d). In fact, the improved functional correspondence to the brain between TDANNs trained with biologically-plausible self-supervised training and models trained on supervised object categorization nearly vanishes when models are evaluated using linear regression.

Another characteristic of visual cortex as a computational system is that the dimensionality of its representational space, that is, the Effective Dimensionality (ED) of encoded information, is relatively low [37, 38, 39]. This characteristic is thought to allow the brain to be robust to noise and well-generalize to new input distributions. We postulate that if the TDANN is a good model of the brain, it should also exhibit this property. Thus, we hypothesize that TDANNs that are more functionally similar to the brain may also be more similar to the brain in ED. Fig. 3c-gray bars shows the functional similarity (horizontal bar) vs. the ED (vertical bar) of each stream. Comparing TDANNs to this brain data reveals that self-supervised TDANNs (Fig. 3c-purple) have lower ED than supervised categorization TDANNs (Fig. 3c-yellow), and increasing the spatial weighting in self-supervised TDANNs further decreases the ED. Strikingly, in all streams, TDANNs that produce the most brain-like functional correspondence also have the most brain-like ED.

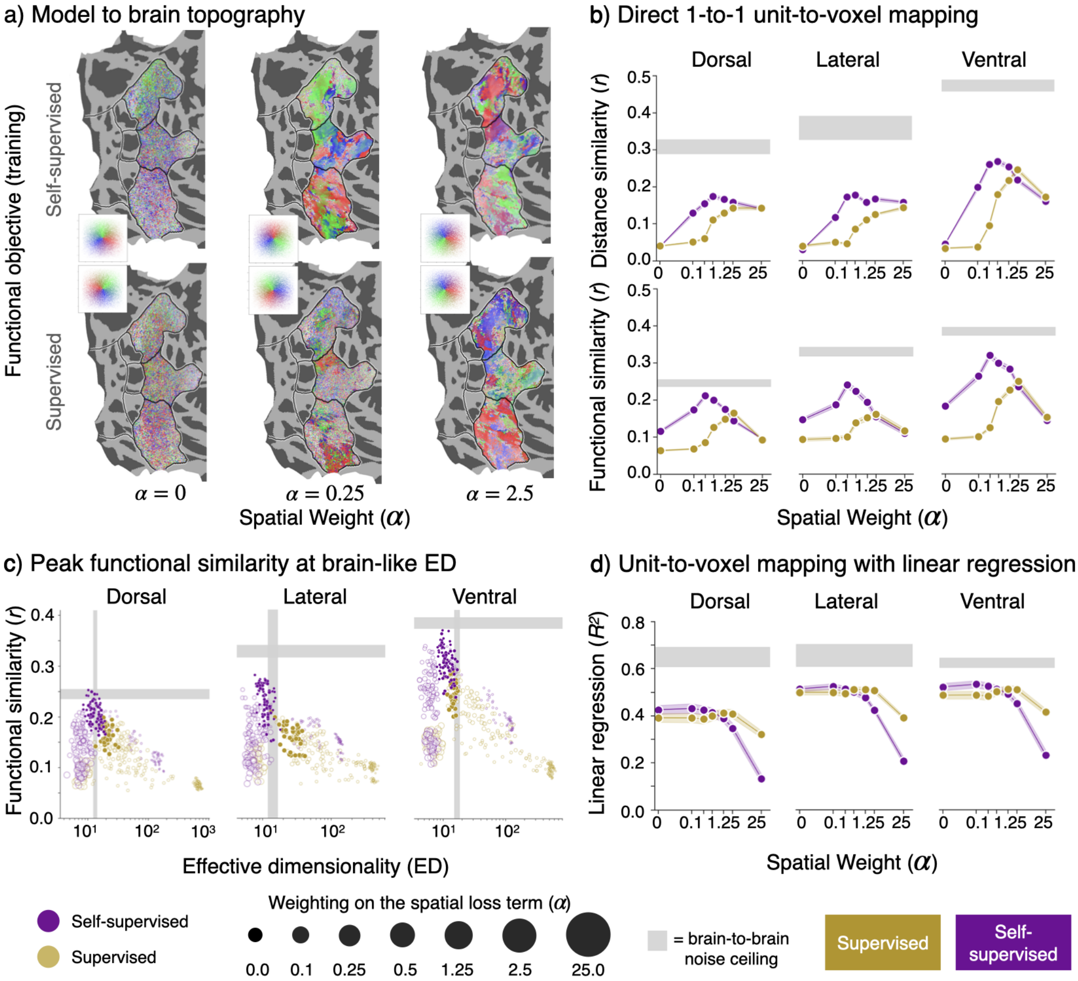

Figure 3. Both self-supervised training and a mid-weight spatial constraint are key to predicting spatial and functional organization into streams. (a) Mapping between TDANN and the brain for an example flattened right hemisphere. Voxels are colored by location on the simulated cortical sheet (see inset). Left: TDANNs trained with no spatial weighting ($\alpha = 0$); Middle: TDANNs trained with $\alpha = 0.25$, best functional similarity for self-supervised TDANNs; Right: TDANNs trained with $\alpha = 2.5$, best functional similarity for TDANNs trained on categorization. **(b)** Across all three streams, the topographic and functional match to brain is best for self-supervised TDANNs trained with $0.25 \leq \alpha \leq 0.5$. Top row: Distance similarity between unit-voxel pairings per stream. Data averaged over unit-voxel pairing for lowest third of distances. Bottom row: average functional similarity (correlation) between unit-voxel pairings per stream. **(c)** TDANNs with most brain-like functional similarity also have most brain-like effective dimensionality (ED). Horizontal shaded gray bar: brain-to-brain functional similarity, averaged across subjects and hemispheres ($\pm$SE across target subjects). Vertical shaded gray bar: ED of brain responses by stream, averaged across subjects and hemispheres ($\pm$SE across subjects). Each dot: one model seed and subject combination, averaged across hemispheres. Opaque points: models with the highest correspondence (b) for each training task ($0.25 \leq \alpha \leq 0.5$: self-supervised; $\alpha = 2.5$: categorization). **(d)** Linear (ridge) regression mapping between model units and voxels for each stream. In (b) and (d): Values are averaged across model seeds and hemispheres, error bars: SE across subjects.

Functional segregation emerges from the TDANN model

Our findings suggest that visual processing streams can emerge in a network that learns via a single, self-supervised task, under a spatial constraint to minimize wiring. Nonetheless, empirical findings imply that there are functional differences across visual processing streams [1, 8, 5] such as differences in population receptive fields (pRFs [41, 42]) and differences in task performance [1, 2]. Is the emergence of streams from a single training task at odds with functional differentiation across streams? To gain initial insights into this question, we test the extent to which TDANN model units assigned to different streams exhibit stream-relevant functional properties.

As pRFs in face-selective regions in the Ventral stream are more central than those of face-selective regions in the Lateral stream [41, 42], we evaluate the mean eccentricity of receptive fields (overlapping an $8°$ x $8°$ stimulus) of TDANN face-selective units assigned to the Ventral and Lateral streams, respectively. Results show a qualitative model-to-brain correspondence: Ventral face-selective model units are significantly more foveal than Lateral ones (mean $\pm$ SE: Ventral = 2.82 ± 0.01; Lateral = 2.94 ± 0.007; $t(15) = -8.2, p = 7.7\text{x}10^{-5}$). Next, we test whether TDANN units assigned to the Dorsal and Ventral streams contribute to stream-specific hypothesized behaviors: (1) determining the object's position [40], associated with the Dorsal stream's role in determining where an object is, and (2) determining the object's category (a 1000-way Imagenet categorization task [43, 26]), associated with the Ventral stream role in determining what the object is. While position [40, 41] and category [44, 45] can be decoded from both Dorsal and Ventral streams, we hypothesize that units assigned to the Dorsal stream will outperform Ventral units for

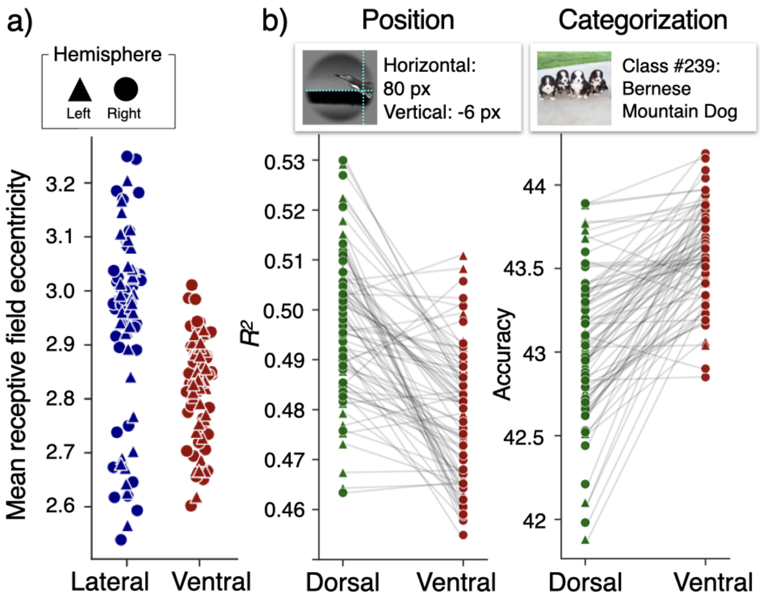

Figure 4. Functional segregation, in alignment with known stream properties, emerges from the TDANN model. Data shown is for self-supervised TDANNs trained with $0.25 \leq \alpha \leq 0.5$. **(a)** Average receptive field eccentricity for face-selective TDANN units is further from the center for those units assigned to the Lateral vs. the Ventral stream. **(b)** Task transfer performance for TDANN units assigned to either the Dorsal or Ventral stream. Left: R^2 on an object position prediction task [40] (location in pixels of the object's center). Right: 1000-way object categorization accuracy on the Imagenet validation set. Green: units mapped to Dorsal stream; Blue: units mapped to Lateral stream; Red: units mapped to Ventral stream. Each dot: a model seed, subject, and hemisphere combination. Triangles: left hemisphere; circles: right.

position prediction and units assigned to the Ventral stream [46, 45], will out-perform Dorsal stream units on object categorization. We find that this is indeed the case: Dorsal units achieve significantly higher performance than Ventral units on predicting object position (mean ± SE: Ventral = $48 \pm 0.06\%$; Dorsal = $50\% \pm 0.10\%$; $t(7) = -16.4, p = 7.7\text{x}10^{-7}$), whereas Ventral units achieve significantly higher accuracy than Dorsal units on object categorization (mean ± SE: Ventral = $43.6\% \pm 0.04\%$; Dorsal = $43.0\% \pm 0.10\%$; $t(7) = 4.21, p = .004$; Supplemental Fig. S5, extended results including Lateral).

Discussion and conclusions

We find that a single, biologically plausible, computational principle - self-supervised learning of the statistics of visual inputs under a spatial constraint that encourages nearby units to have correlated responses - better explains the functional and spatial organization of the human visual system into processing streams than a system trained to perform different visual behaviors in parallel. These data necessitate a rethinking of an inherent idea in philosophy [47], psychology [48, 49], computational theory [50, 12, 51], and neuroscience [52], that different portions of our visual system have explicitly evolved to support a collection of distinct visual behaviors. Instead, our results suggest an intriguing new idea that evolution may lead to the emergence of a flexible visual system that can learn a task-general representation in an self-supervised manner, while being constrained by the physical size and layout of cortical tissue. In this conception, the visual system can learn from visual input alone without necessitating human-unique inputs such as language [53]. Moreover, it still develops distinct streams with functional properties suitable for different visual behaviors [1, 2, 3, 4, 5], and it is information efficient [16, 19, 38].

This understanding would not have been possible without conceptual, empirical, and methodological innovations, including a full end-to-end TDANN that is both trainable and simulates the topographic arrangement of units on the cortical sheet [20, 25], a massive fMRI dataset [28] that enables comparing DANNs to the human brain, and a 1-to-1 mapping algorithm between model units to brain voxels. Recent success of DANNs in explaining neural responses in the visual system [11, 30, 54, 29] has elicited excitement that this class of models has the potential to explain why the brain is organized the way it is [55, 52, 31, 51, 56, 57]. At the same time, there is considerable theoretical debate as how to evaluate if a model accurately explains the brain [35, 32, 31, 58] as commonly used metrics, such as linear regression between model units and brain responses, do not distinguish between models [29, 33]. Here we show that a more stringent criterion - a 1-to-1 mapping between model units and brain voxels - is able to distinguish between models of the brain, including providing evidence for a definitive advantage of a more biologically-plausible self-supervised training over the best-to-date supervised categorization training [27, 11, 30, 29] in explaining brain responses in visual cortex. These findings, together with computational-theoretical advancements in developing metrics to compare systems that preserve neural tuning [33, 32], not just representational space

[11, 12, 59, 29, 51, 60], underscore the necessity of using stricter metrics not only to adjudicate between putative models of the brain, but also to glean new understanding of biological systems.

The success of the TDANN underscores the necessity of modeling not only brain functional responses, as is the prevalent approach [11, 30, 27, 52, 31, 57, 56], but also brain topography. This follows insights from several recent studies that have investigated the emergence of regional topographic maps in the ventral stream [61, 62, 25, 20] suggesting that wiring [61], smoothness [62], and a balancing of spatial and functional constraints [20] can produce topographic organization. The key insight from the present work is that a local spatial constraint that allows fast processing [16] and may contribute to minimizing local wiring length [63, 15] can percolate up to create broad-scale stream structure. As parallel processing streams exist in other species [64], cortical systems [65, 66, 67, 68], and spatial scales [69, 70], future research can test if the same principles trained on other sensory and multimodal inputs lead to the emergence of parallel processing streams across the brain. Overall, this study suggests a paradigm shift: any end-to-end computational model of the brain that learns from the sensory input needs to include physical constraints, and not just behavioral goals, in order to accurately predict brain function.

References

1. Leslie G Ungerleider and Mortimer Mishkin. Two cortical visual systems. analysis of visual behavior. *Analysis of Visual Behavior*, pages 549–586, 1982.
2. Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992.
3. Kevin S Weiner and Kalanit Grill-Spector. Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. *Psychological Research*, 77(1):74–97, 2013.
4. Moritz Wurm and Alfonso Caramazza. Action and object representation in the ventral" what" stream. *PsyArXiv*, 2021.
5. David Pitcher and Leslie G Ungerleider. Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, 2020.
6. Leyla Isik, Kami Koldewyn, David Beeler, and Nancy Kanwisher. Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 114(43):E9145–E9152, 2017.
7. Leslie G Ungerleider and James V Haxby. ‘what’and ‘where’in the human brain. *Current Opinion in Neurobiology*, 4(2):157–165, 1994.
8. Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, 6:414–417, 1983.
9. Patrick Mineault, Shahab Bakhtiari, Blake Richards, and Christopher Pack. Your head is there to move you around: Goal-driven models of the primate dorsal pathway. *Advances in Neural Information Processing Systems*, 34, 2021.
10. H Steven Scholte, Max M Losch, Kandan Ramakrishnan, Edward HF de Haan, and Sander M Bohte. Visual pathways from the perspective of cost functions and multi-task deep neural networks. *cortex*, 98:249–261, 2018.
11. Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.
12. Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.
13. Mark E Nelson and James M Bower. Brain maps and parallel computers. *Trends in neurosciences*, 13(10): 403–408, 1990.
14. David C van Essen. A tension-based theory of morphogenesis and compact wiring in the central nervous system. *Nature*, 385(6614):313–318, 1997.
15. Robert A Jacobs and Michael I Jordan. Computational consequences of a bias toward short connections. *Journal of Cognitive Neuroscience*, 4(4):323–336, 1992.
16. Semir Zeki and Stewart Shipp. The functional logic of cortical connections. *Nature*, 335(6188):311–317, 1988.
17. David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology*, 160(1):106, 1962.
18. Brian A Wandell and Jonathan Winawer. Imaging retinotopic maps in the human brain. *Vision research*, 51(7): 718–737, 2011.
19. David C Van Essen, Charles H Anderson, and Daniel J Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–423, 1992.
20. Eshed Margalit, Hyodong Lee, Dawn Finzi, James J DiCarlo, Kalanit Grill-Spector, and Daniel LK Yamins. A unifying principle for the functional organization of visual cortex. *bioRxiv*, pages 2023–05, 2023.
21. James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

22. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

23. Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

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

25. Hyodong Lee, Eshed Margalit, Kamila M Jozwik, Michael A Cohen, Nancy Kanwisher, Daniel LK Yamins, and James J DiCarlo. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 2020.

26. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. 2020. doi: 10.48550/ARXIV.2002.05709. URL https://arxiv.org/abs/2002.05709.

27. Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

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

29. Colin Conwell, Jacob S Prince, Kendrick N Kay, George A Alvarez, and Talia Konkle. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, pages 2022–03, 2022.

30. Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11):e1003915, 2014.

31. Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, pages 1–20, 2023.

32. Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.

33. Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. *arXiv preprint arXiv:2311.09466*, 2023.

34. Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67, 2021.

35. Rosa Cao and Daniel Yamins. Explanatory models in neuroscience: Part 2–constraint-based intelligibility. *arXiv preprint arXiv:2104.01489*, 2021.

36. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

37. Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

38. Nathan CL Kong, Eshed Margalit, Justin L Gardner, and Anthony M Norcia. Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, 18(1):e1009739, 2022.

39. Eric Elmoznino and Michael F Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *bioRxiv*, pages 2022–07, 2022.

40. Ha Hong, Daniel LK Yamins, Najib J Majaj, and James J DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613–622, 2016.

41. Dawn Finzi, Jesse Gomez, Marisa Nordt, Alex A Rezai, Sonia Poltoratski, and Kalanit Grill-Spector. Differential spatial computations in ventral and lateral face-selective regions are scaffolded by structural connections. *Nature communications*, 12(1):2278, 2021.

42. Magdalena W Sliwinska, Caitlin Bearpark, Julia Corkhill, Aimee McPhillips, and David Pitcher. Dissociable pathways for moving and static face perception begin in early visual cortex: Evidence from an acquired prosopagnosic. *Cortex*, 130:327–339, 2020.

43. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

44. Lior Bugatus, Kevin S Weiner, and Kalanit Grill-Spector. Task alters category representations in prefrontal but not high-level visual cortex. *Neuroimage*, 155:437–449, 2017.

45. Maryam Vaziri-Pashkam and Yaoda Xu. Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. *Journal of Neuroscience*, 37(36):8767–8782, 2017.

46. James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293 (5539):2425–2430, 2001.

47. Jerry A Fodor. *The modularity of mind*. MIT press, 1983.

48. Rainer Goebel. A connectionist approach to high-level cognitive modeling. In *12th Annual Conference. CSS Pod*, pages 852–859. Psychology Press, 2022.

49. Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Montero, Christian Tsvetkov, Guillermo Puebla, Federico G Adolfi, John Hummel, Rachel Flood Heaton, Benjamin Evans, et al. Disagreement and confusion over the status of dnns as models of vision. 2023.

50. David Marr. Vision: A computational investigation into the human representation and processing of visual information, 1982.

51. Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, page 407007, 2020.

52. Nancy Kanwisher, Meenakshi Khosla, and Katharina Dobs. Using artificial neural networks to ask 'why'questions of minds and brains. *Trends in Neurosciences*, 2023.

53. Aria Y Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Better models of human high-level visual cortex emerge from natural language supervision with a large and diverse dataset. *Nature Machine Intelligence*, pages 1–12, 2023.

54. Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

55. Radoslaw M Cichy and Daniel Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317, 2019.

56. Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

57. Grace W Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, 33(10):2017–2031, 2021.

58. Anna A Ivanova, Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. Is it that simple? linear mapping models in cognitive neuroscience. *bioRxiv*, page 438248, 2021.

59. Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, 34:25164–25178, 2021.

60. Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

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

62. Fenil R Doshi and Talia Konkle. Cortical topographic motifs emerge in a self-organized map of object space. *Science Advances*, 9(25):eade8187, 2023.

63. Dmitri B Chklovskii, Thomas Schikorski, and Charles F Stevens. Wiring optimization in cortical circuits. *Neuron*, 34(3):341–347, 2002.

64. David C Van Essen and Jack L Gallant. Neural mechanisms of form and motion processing in the primate visual system. *Neuron*, 13(1):1–10, 1994.

65. Josef P Rauschecker and Biao Tian. Mechanisms and streams for processing of “what” and “where” in auditory cortex. *Proceedings of the National Academy of Sciences*, 97(22):11800–11806, 2000.

66. Takahiro Osada, Akitoshi Ogawa, Akimitsu Suda, Koji Nakajima, Masaki Tanaka, Satoshi Oka, Koji Kamagata, Shigeki Aoki, Yasushi Oshima, Sakae Tanaka, et al. Parallel cognitive processing streams in human prefrontal cortex: Parsing areal-level brain network for response inhibition. *Cell Reports*, 36(12), 2021.

67. Judith Tanné-Gariépy, Eric M Rouiller, and Driss Boussaoud. Parietal inputs to dorsal versus ventral premotor areas in the macaque monkey: evidence for largely segregated visuomotor pathways. *Experimental Brain Research*, 145(1):91–103, 2002.

68. Simone Vossel, Joy J Geng, and Gereon R Fink. Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *The Neuroscientist*, 20(2):150–159, 2014.

69. BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.

70. Richard F Betzel and Danielle S Bassett. Multi-scale brain networks. *Neuroimage*, 160:73–83, 2017.

Supplementary Information

Materials and methods

Code and data availability

Original code for this study is available at https://github.com/dawnfinzi/spacestream. The neural data analyzed in this study comes from the Natural Scenes Dataset (NSD) [1] available at http://naturalscenesdataset.org/.

"Training phase": Neural network architectures and training

Multi-behavior models. To test the *multiple behavioral demands hypothesis*, we used three models, each trained on a different task: object categorization, action recognition, and object detection. We chose these tasks as these are the computer vision equivalents of the proposed behaviors each stream is thought to support (Ventral: what is it, Lateral: what is it doing, Dorsal: where is it).

We used two versions of multi-behavior models: one version with a ResNet-50 base architecture, for optimal task performance, and one version with a ResNet-18 base architecture as a control to more closely match the TDANN architecture, and the number of visual areas in primate cortex [2, 3]. For the object categorization model, we used the base model (ResNet-50 or ResNet-18) [4] trained on object categorization on ILSVRC-2012 (ImageNet Large-Scale Visual Recognition Challenge [5]). For the action recognition model, we used the SlowFast model architecture [6], which is a dual-pathway network with a 3D ResNet backbone trained on the Kinetics-400 video dataset [7]. Finally, for the ResNet-50 object detection model, we used a Faster R-CNN [8] trained on MS-COCO [9]. For the ResNet-18 object detection model, we used a single-stage object detection network, SSD [10], for greater correspondence with the other multi-behavior models, and trained on Pascal VOC (2007 and 2012) [11], in order to avoid any confounds with both training and testing on MS-COCO (as the NSD images are also from MS-COCO).

We randomly subsampled an equal number of units from layer 4.1 or its equivalent from each network (ignoring any "visually non-responsive" units that did not respond to any of the images) so that the total number of units was equal to the total number of voxels. In order to allow for the most direct comparison against TDANN models, the units sampled from the task-trained models were assigned random positions on a two-dimensional simulated cortical sheet. We then followed the same pre-optimization procedure as in Initialization of model unit position: *Stage 2* in order to be able to fairly apply the same mapping algorithm.

Topographic Deep Artificial Neural Network (TDANN).

Model architecture and training The TDANN model class, which we used to evaluate the *spatial constraints hypothesis*, is based on ResNet-18 architecture, with two key differences: (1) model units are assigned positions on a 2D simulated cortical sheet and (2) the model is trained to jointly minimize a spatial and a task loss. All TDANN models were built using the ResNet-18 [4] base architecture (from the *torchvision* implementation) and trained using modifications to the VISSL framework [12]. ResNet-18 was chosen because it has been shown to achieve strong task performance, accurately predict neuronal responses across the visual system [13], and has roughly comparable number of layers to stages (areas) in the primate visual system [14, 3]. Models were trained for 200 epochs using the ILSVRC-2012 [5] training set, with each model being trained from five different random initial seeds. We optimized the network parameters using stochastic gradient descent with momentum (with γ set to 0.9), a batch size of 512, and a learning rate initialized to 0.6, which then decayed according to a cosine learning schedule [15] Models were trained using a self-supervised contrastive objective "SimCLR" [16] or a supervised 1000-way object categorization task.

Initialization of model unit position Prior to training, model units in each layer were assigned fixed positions in a two-dimensional simulated cortical sheet specific to that layer. The size of the cortical sheet in each layer, and the size of the "cortical zone" used during training (computation of the spatial loss is restricted to units within the same cortical zone), was determined by the presumed correspondence, based on previous work comparing convolutional neural networks (CNNs) and the primate visual system [17, 18, 19], between model layers and human visual areas (see [2] for further details). Positions were then assigned in a two-stage process.

Stage 1: Retinoptopic initialization As each layer convolves over the outputs of the previous layer, the resulting responses are structured into spatial grids. To maintain this inherent organization, we assigned each model unit to a specific area of the simulated cortical sheet that aligns with its spatial receptive field.

Stage 2: Pre-optimization of positions In CNNs, filter weights are shared between units at different locations, which means that local updates to one unit affect all units with the same filter weights. This global coordination constraint makes it challenging to achieve local smoothness when the units are arbitrarily positioned. To address this, a pre-optimization of unit positions was necessary to identify a set of positions that enables learning smooth cortical maps. We spatially shuffled the units of a pre-trained CNN on the cortical sheet, so that nearby units had correlated responses to a set of sine grating images. The use of sine gratings is based on studies that show that propagating retinal waves drive development of the visual system in the womb in primates and other mammals [20, 21, 22, 23]

The spatial shuffling works as follows: 1) Randomly select a cortical zone. 2) Compute pairwise response correlations for all units in the zone 3) Select 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 trained the TDANN models using a weighted sum of two types of loss functions: a task loss, which served to encourage learning of visual representations and a spatial loss, which encourages local correlations in responses to visual inputs. Optimization on the total loss function leads to both successful visual representation learning and minimization of inter-layer wiring length [2].

Spatial loss The spatial loss function encourages nearby model units on the simulated cortical sheet to be correlated in their responses to the training stimuli. Specifically, SL_l is the spatial correlation loss computed for the l-th layer and SL_l is computed on a given batch by randomly sampling a local cortical zone and calculating for pairs of units, (1) correlation (Pearson's r) between the response profiles, ($\vec{R}$), and (2) the the stabilized reciprocal Euclidean distances ($\vec{D}$):

$$\vec{D} = \frac{1}{(1+\vec{d})} \tag{S1}$$

where $\vec{d}$ is the vector of pairwise cortical distances. These two terms are then related as follows:

$$\mathrm{SL}_l = 1 - \mathrm{Corr}\left(\vec{R}, \vec{D}\right) \tag{S2}$$

such that SL_l is minimized when nearby units have correlated responses to the training stimuli.

Task loss The task loss (TL) is computed from the output of the final model layer. We tested two candidate TLs: supervised object categorization cross-entropy loss [24] and the self-supervised SimCLR objective [16]. The SimCLR objective is a contrastive loss function which works by creating two "views", or augmentations, of each image in a batch, using random cropping, horizontal flips, color distortion, and Gaussian blur. These views are passed to the network and the final layer outputs are passed through a 2-layer multi-layer perceptron (MLP), producing a low-dimensional representation of each view which serves as the input to the loss function. The SimCLR loss function then attempts to maximize the similarity of representations for two views of the same source image, while pushing that representation away from all other images in the batch.

Overview of training In sum, the TDANN is trained on Imagenet [5] to minimize this total loss, which is the sum of the weighted spatial loss for each layer and the task loss as follows:

$$\text{Total Loss} = \mathrm{TL} + \alpha \sum_{l \in \text{layers}} \mathrm{SL}_l \tag{S3}$$

where α is the weight of the spatial loss component (fixed across all layers), and SL_l is the spatial correlation loss computed for the l-th layer.

The total model training process consists of 6 steps:

1. The ResNet-18 model is trained using the task loss only.
2. Positions are initialized to preserve coarse retinotopy in each layer (Stage 1).
3. Positions are pre-optimized in an iterative process that preserves retinotopy while bringing together units with correlated responses to sine gratings (Stage 2).
4. After pre-optimization, positions are permanently frozen.
5. All network weights are randomly re-initialized.
6. The network is trained to minimize the total loss.

"Mapping phase": Linking model to brain

Neural data. As our neural comparison, we used the Natural Scenes Dataset (NSD) [1], a high-resolution fMRI dataset that densely sampled responses to up to 10,000 natural images in each of eight individuals over the course of 32-40 scan sessions. Full details on data collection and processing can be found in Allen et al. [1]. Briefly, scanning was conducted at 7T using gradient-echo EPI at 1.8-mm isotropic resolution with whole-brain coverage. Images were taken from Microsoft's COCO image database [9], square cropped to 425 pixels x 425 pixels, and presented at fixation at a size of 8.4° x 8.4° for 3 seconds with 1 second gaps in between images. Data were preprocessed using one temporal resampling (to correct for slice timing differences) and one spatial resampling (to correct for head motion, EPI distortions and gradient non-linearities), resulting in upsampled 1.0mm resolution (temporal resolution 1.0 s). Single-trial beta weights were estimated using a general linear model approach designed to optimize the quality of single trial betas (GLMsingle [25]). The single-trial responses were then z-scored across images for each voxel and session and then averaged across 3 trial repeats. We used cortex-based alignment to align all data to the fsaverage surface. Throughout reporting of the results, we refer to the brain units of measurement as "voxels", for interpretability, though they are more technically "vertices" as we are using the fsaverage preparation.

ROIs We defined regions of interest (ROIs) for early, intermediate, and high-level visual cortex for each of the three streams based on a combination of anatomical landmarks, noise ceiling estimates (Supplemental Fig. S1), and a constraint to roughly match the number of voxels per stream. We focus only on the high-level visual ROIs for the purposes of this paper and compare the end point of the models to the end points of the processing stream in each brain. However, the full details for drawing all seven (one early, three intermediate and three higher-level) ROIs are included here for completeness and because the high-level ROIs share boundaries with the intermediate ROIs. ROIs were drawn on the fsaverage surface as follows:

Early visual cortex ROI: The early visual cortex ROI was drawn as the union of the V1v, V1d, V2v, V2d, V3v and V3d ROIs from the Wang retinotopic atlas [26]. Additionally, V2v and V2d, as well as V3v and V3d, were connected such that the part of the occipital pole typically containing foveal representations was included in the ROIs.

Intermediate ROIs: Three intermediate ROIs were drawn corresponding to each of the three streams: Ventral, Lateral and Dorsal. All three ROIs border the early visual cortex ROI on the posterior side. The intermediate Ventral ROI was drawn to reflect the inferior boundary of hV4 from the Wang atlas [26] and includes the inferior occipital gyrus (IOG), with the anterior border of the ROI drawn based on the anterior edge of the inferior occipital sulcus (IOS). The intermediate Lateral ROI was drawn directly superior to the intermediate ventral ROI, with the superior and anterior borders determined as the LO1 and LO2 boundaries from the Wang atlas [26]. The intermediate Dorsal ROI was drawn directly superior to that to include V3A and V3B from the Wang atlas.

Higher-level ROIs: Three higher-level ROIs were drawn for each of the Ventral, Lateral and Dorsal streams, bordering their respective intermediate ROIs on their posterior edges. The ventral ROI was drawn to follow the anterior lingual sulcus (ALS), including the anterior lingual gyrus (ALG) on its inferior border and to follow the inferior lip of the inferior temporal sulcus (ITS) on its superior border. The anterior border was drawn based on the midpoint of the occipital temporal sulcus (OTS). The lateral ROI was drawn such that the higher-level ventral ROI was its inferior border and the superior lip of the superior temporal sulcus (STS) was used to mark the anterior/superior boundary. The rest of the superior boundary traced the edge of angular gyrus, up to the tip of the posterior STS (pSTS). The dorsal ROI was drawn to reflect the boundary of the lateral ROI on its inferior edge and to otherwise trace the borders of and include the union of IPS0, IPS1, IPS2, IPS3, IPS4, IPS5 and SPL1 from the Wang retinotopic atlas.

The three higher-level ROI were then trimmed using the prepared noise ceiling maps for beta version b3 and the fsaverage surface [1]. The noise ceiling estimates represent the amount of variance contributed by the signal expressed as a percentage of the total amount of variance in the data, for the average of responses across three trial presentations. An approximate cutoff of 10% was used to guide trimming of the higher-level ROIs, such that we were left with reduced ROIs where all voxels had a noise ceiling $\geq 10\%$ theoretically predictable variance. These ROIs were contiguous and roughly matched in size (number of voxels per ROI right hemisphere: Dorsal = 6688, Lateral = 6839, Ventral = 5638; left hemisphere: Dorsal = 6182, Lateral = 5849, Ventral = 6126).

Model data. For each of our neural network models, we extracted features in response to the same set of 73,000 total NSD images seen by participants in the scanner. Features were extracted from layer 4.1 in ResNet18 and ResNet50 or equivalent. This layer was chosen as past work shows that it has the best functional correspondence to higher-level visual areas, including on the NSD [3, 2, 27]. This results in a matrix of features of the form *number of images* x *number of units*, where the number of units is 7 x 7 x 512 = 25,088, i.e. the total number of units in layer 4.1, for the TDANN models, and 19,164 total subsampled units for the multi-behavior models (6,388 units per model), which matches the maximum number of voxels per hemisphere. Thus, for all models, the number of source units was of the same order of magnitude as the number of target voxels [28, 29].

1-to-1 mapping. Development of the mapping algorithm was done using the case of the brain-to-brain mapping, with the reasoning being that if we failed to recover an element of the functional organization in the model-to-brain case, we would not be able to arbitrate between a failure in the model and a failure in the mapping, unless we had already shown that such functional organization could be captured in the brain-to-brain case using an identical mapping. We found that incorporation of the spatial prior (see below) into the mapping significantly improves accuracy in assignment (Supplemental Fig. S2). Once satisfied with the mapping algorithm in the brain-to-brain case, we used that algorithm with as few modifications as possible (neighborhood radius and stimuli used, detailed below) to map model-to-brain.

For each pair of subjects (source subject mapped to target subject), we used the single-trial z-scored betas for each source voxel and each target voxel in response to the 515 images shared across all subjects (80% used for assignment, 20% used for evaluation) and computed the correlations between each pair of voxels. This correlation matrix was then transformed into a cost matrix (1 – correlation) and assignment was first attempted purely on the basis of this cost matrix ("functional only") using the Kuhn-Munkres algorithm [30], a combinatorial optimization algorithm which solves the assignment problem in polynomial time. The "functional only" algorithm performed above chance in assigning voxels to the correct streams but did not fully recover the spatial organization (Supplemental Fig. S2). We thus added a minimal smoothness constraint to the optimization procedure; the smoothness constraint encourages neighboring voxels in the target space to “pick” neighboring voxels in the source space (given a small local radius of 5 mm; full algorithm provided below:1). As this recovered more of the known spatial organization in the voxel-to-voxel case (Supplemental Fig. S2), this was the mapping algorithm we chose to then apply in the unit-to-voxel case. To convert the radius of 5 mm used in the voxel-to-voxel case to the model space, we calculated what percentage of the max voxel-to-voxel distance (237 mm) the brain radius cutoff was and then multiplied that percentage by the max model distance (12.9). This yielded a model radius cutoff of approximately 0.27, resulting in the following two radii used in the unit-to-voxel case, 5 mm for the brain distances and 0.27 for the model distances. Additionally, in the unit-to-voxel case, we leveraged the full set of unique images (up to 9485 per individual) to link models to individual brains. The 515 shared images were then reserved for evaluating theories ("Test phase").

Algorithm 1 1-to-1 mapping with spatial prior (Fig. 1c)

```
Require: Cost matrix $C$ of dimension $N_s \times N_t$, where $N_s$ is the number of source units and $N_t$ is the number of target units, and each entry in $C$ is $1-$ the pairwise correlation of the response vectors. Source distance matrix, $D_s$, of dimension $N_s \times N_s$, with the pairwise distances between all units in the source space. Target distance matrix, $D_t$, of dimension $N_t \times N_t$, with the pairwise distances between all units in the target space. Radius $r$ to use as neighborhood size.
  procedure MAPPING ALGORITHM($C, D_s, D_t, r$)
    Assignments $A \leftarrow \texttt{Kuhn-Munkres}(C)$        ▷ Initialized based on "functional only" mapping
    while Mean movement of assignments from iteration to iteration has not converged do
      $C_{\text{temp}} \leftarrow C\texttt{.copy()}$
      for each target unit, $v_t$ do
        Find all neighboring target units, $V_{tn}$ within distance $r$
        Initialize candidate matrix, $V_{sn}$
        for each unit in $V_{tn}$ do
          Find their assigned source unit, $v_s$, from $A$
          $v_{sn} \leftarrow$ all units in source space within distance $r$ from $v_s$
          Append $v_s$ and $v_{sn}$ to $V_{sn}$
        end for
        Fit a 2D Gaussian, $G$, to point cloud, $V_{sn}$ in source space
        $M_{sn} \leftarrow \texttt{mahalanobis}(V_{sn}, G)$
        for each unit, $u$, in $V_{sn}$ do
          if $M_{sn}[u] > 2$ then
            Remove $u$ from $V_{sn}$
          end if
        end for
        $C_{\text{temp}}[v_t, \neg V_{sn}] = 1000$        ▷ all source units not in $V_{sn}$ are set to have a prohibitive cost
      end for
      $A \leftarrow \texttt{Kuhn-Munkres}(C_{\text{temp}})$
    end while
    return $A$
  end procedure
```

Total models tested. We evaluated 5 instances initialized with different random seeds per each of the TDANNs across 2 training tasks (SimCLR and categorization) and 7 levels of spatial weightings (α), as well as 2 base architectures

(Resnet-18 and ResNet-50) for each of the 3 multi-behavior models (detection/action/categorization). Each model was then mapped to the 2 hemispheres for each of the 8 participants, totaling 1216 model-to-brain mappings tested.

"Test phase": Evaluating theories

Evaluating spatial correspondence. We evaluated the spatial correspondence in two ways. First, to compare multi-behavior models and TDANNs directly we calculated a percentage spatial correspondence metric. In the case of multi-behavior models, this is calculated as the percentage of voxels for each stream that were assigned to the multi-behavior model trained using that stream's corresponding task (i.e. Ventral and object categorization). In the case of TDANNs, we divided the simulated cortical sheet into three sections (candidate stream partition scheme), assuming a log-polar transform, and calculated the highest percentage of voxels that match the partition scheme across candidate partitions (reflection and $5°$ rotations).

Second, when comparing across TDANNs that have simulated cortical sheets, we additionally calculated a distance similarity metric. The distance similarity metric measures, for each unit-to-voxel pairing, the correlation between the normalized distances on the model cortical sheet (from that unit to other units) and normalized distances on the actual cortical surface (for the assigned voxel to the other units' assigned voxels), averaged across-unit-to-voxel pairings. For each unit, this metric is calculated across only the closest 33% of units, to simulate stream boundaries, which has the additional benefit of discounting high distances where there are few pairs. The same calculation was performed in the brain-to-brain case to determine the actual cortical level of distance similarity. We report the average distance similarity across unit-to-voxel pairings.

Evaluating functional correspondence. To evaluate functional correspondence between candidate models and the brain, we report the 1-to-1 correlations calculated on the left-out set of 515 shared images, using the unit-to-voxel assignments determined by the mapping procedure. Each unit-to-voxel correlation was normalized by the individual voxel noise ceiling (r) of that assigned voxel (see [1] for information on the calculation of the intra-individual voxel noise ceilings in NSD). 1-to-1 correlations were calculated on an individual subject and hemisphere basis for each of the candidate models. The voxel-to-voxel assignments were used to calculate the overall inter-individual i.e. brain-to-brain noise ceiling (correlations evaluated on test set of 20% of the shared images, averaged across 5 splits for each source and target subject combination).

Linear regression. To compare the 1-to-1 mapping results to the commonly used mapping method of linear regression between model units and the brain [18, 31, 13], we also calculated TDANN model to brain correspondence by regressing model responses from the final convolution layer onto individual voxel responses using ridge regression. As in [13], to decrease computational costs without sacrificing performance, we first projected unit activations into a lower dimensional space using a subsample of the ImageNet validation images and retained the first 1000 PCs. Performance was evaluated on a left-out test set (8/9 train, 1/9 test, shared images excluded, 10 splits) for each subject separately. Test R^2 for each voxel is normalized by the individual voxel noise ceiling (R^2). To evaluate the upper-bound model performance given the shared variance across subjects, we again calculated a brain-to-brain noise ceiling for each stream by repeating the same ridge regression procedure as for model-to-brain but instead using all other subjects' responses to predict the left-out subject (80/20 train-test split using the set of 515 shared images, 10 splits).

Effective dimensionality

To calculate effective dimensionality, i.e. the dimensionality of the space of how information is represented by a system (also referred to as the latent dimensionality), we considered the responses of the subset of units assigned to each stream by subject. Using the unit activations (in the case of the models) or z-scored betas (in the case of the human subjects), we computed the eigenspectrum of these responses to the MSCOCO images used in NSD. Following [32] and [33], we computed effective dimensionality from the eigenvalues (λ) as:

$$ED = \frac{\left(\sum_{i=1}^{N} \lambda_i\right)^2}{\sum_{i=1}^{N} \lambda_i^2} \qquad \textbf{(S4)}$$

where N is the number of eigenvectors. Intuitively, if the eigenspectrum decays slowly, that means there are many informative dimensions, and the ED, which in words is simply the squared sum of the eigenvalues over the sum of squares of the eigenvalues, will be high. On the other hand, if the eigenspectrum decays rapidly, meaning that information is largely encoded in only a few dimensions, then the ED will be low.

Model unit selectivity and receptive field properties

Previous studies reported that voxels in face-selective regions in the Lateral stream are more peripheral than those in face-selective regions of the Ventral stream [34]. To test if this functional feature was also present in the TDANN, we identified face-selective units in each stream and then estimated the eccentricity of their receptive fields. To identify face-selective units in the TDANN, we used the functional localizer (fLoc) stimulus set [35]. fLoc contains stimuli from 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). These stimuli have been previously used to localize and describe category-selective responses in human higher visual cortex in fMRI studies [35, 34, 36]. 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.

$$t = \frac{\mu_{\text{on}} - \mu_{\text{off}}}{\sqrt{\frac{\sigma^2_{\text{on}}}{N_{\text{on}}} + \frac{\sigma^2_{\text{off}}}{N_{\text{off}}}}}, \qquad \textbf{(S5)}$$

where μ_{on} and μ_{off} are the mean responses to the "on" categories (adult and child faces) and "off" categories (all non-face categories), respectively, σ^2 are the associated variances of responses to exemplars from those categories, and N is the number of exemplars being averaged over. As in fMRI experiments, units with $t > 3$ were classified as face-selective. For each unit, eccentricity was then calculated based on the unit's (x,y) position from the center of the $7x7$ filter, converted to degrees of visual angle (by multiplying by $8°$ input stimulus / 7 grid size). From there, we divided these units based on which stream they were assigned to and report the mean eccentricity across face-selective units for each model seed x subject x hemisphere combination.

Task transfer

We tested performance of self-supervised TDANNs units mapped to the Dorsal and Ventral stream, on new tasks, position prediction and object categorization, respectively, associated with each stream (results for the Lateral stream and additional tasks in Supplementary Fig. S5). We refer to this as task transfer performance as the TDANN model was not trained on any of these tasks and model weights were frozen. Performance on the transfer task was tested for the self-supervised TDANNs which best match the brain ($0.25 \leq \alpha \leq 0.5$ in Fig. 4, full results across spatial weightings in Supplementary Fig. S5) and each hemisphere of each subject.

Position task. We evaluated the performance of TDANN units assigned to each stream on predicting the vertical and the horizontal locations in pixels of an object center in an image, using the stimulus set from Hong et al., which has also been used in the evaluation of neural network models of the mouse [38] and primate [31, 39] visual systems. This stimulus set consists of 5760 gray-scale images of 64 distinct objects chosen from one of eight categories (animals, boats, cars, chairs, faces, fruits, planes, tables) placed on randomly chosen, realistic background scene images. Object position, pose and size in this stimulus set varied at different levels from no variation, to medium variation and high variation levels.

For each TDANN model and individual subject, smaller "stream models" were created for each of the three streams by selecting the 5000 units assigned to that stream with the highest correlations. We extracted activations from these units and reduced the dimensionality of the activations to 1000 dimensions using principal components analysis (PCA). We used Ridge regression, with the regularization parameter, α, cross-validated from

$$\alpha \in [0.01, 0.1, 1, 10] \qquad \textbf{(S6)}$$

to predict the vertical and the horizontal locations in pixels of the object center in the image. We performed five-fold cross-validation on the training split of the no- and medium-variation image subsets, consisting of 3200 images, and computed performance on the test split of the high-variation set consisting of 1280 images. Ten different category-balanced train-test splits were randomly selected. We report R^2 on the high-variation test set, averaged across the 10 splits.

Categorization task. To evaluate the performance of TDANN units assigned to each stream on a downstream categorization task, we used the 1000-way ImageNet object categorization task [5]. For each TDANN model and individual subject, smaller "stream models" were created for each of the three streams as in the position prediction task. A single linear layer was then trained directly from the outputs of those units. The linear layer was trained for 28 epochs of the ILSVRC-12 ImageNet training set (1,281,167 images) with a batch size of 1,024 and a learning rate which was initialized to 0.04 and decreased by a factor of 10 every eight epochs. We report the top-1 performance on the held-out validation set (50,000 images).

Supplementary tables

Table 1. Linear mixed-effects model to test effect of candidate model type on spatial correspondence for each of the three streams. To test if there were differences between candidate models on spatial correspondence with the brain, we used linear mixed-effects models, with fixed effects for candidate model type (intercept denotes multi-behavior candidate model with ResNet-18 base, i.e. MB ResNet-18) and a random intercept for each subject. Model specification was as follows: spatial correspondence $\sim$ candidate model type + 1 | subject. A separate model was run for each of the three streams. Positive values indicate better spatial correspondence than the MB ResNet-18 (first row) and negative values indicate worse spatial correspondence. For example, the β coefficient of -15.87 for MB ResNet-50 in Dorsal indicates that there is an average decrease of 15.87% in spatial correspondence for the MB ResNet-50 model relative to the MB ResNet-18 model, while the β coefficient of 11.91 for the supervised TDANN indicates an average increase of 11.91%, again relative to the MB ResNet-18 value of 50.2%. Corrected p-values indicate p-values Bonferroni-corrected for multiple comparisons between candidate model types. Significant predictors ($p < .05$) are shown in bold. MB = multi-behavior. These statistics are related to Fig 2b.

		Coefficients±SE	z-value	p-value	corrected p-value
Dorsal	**Intercept (MB ResNet-18)**	**50.20±1.54**	**32.61**	**2.6x10^{-233}**	
	MB ResNet-50	**-15.87±2.12**	**-7.48**	**7.5x10^{-14}**	**2.2x10^{-13}**
	Self-supervised TDANN	**6.13±2.12**	**2.89**	**.004**	**.012**
	Supervised TDANN	**11.91±2.12**	**5.61**	**2.0x10^{-8}**	**6.1x10^{-8}**
Lateral	**Intercept (MB ResNet-18)**	**50.58±1.43**	**35.28**	**1.3x10^{-272}**	
	MB ResNet-50	**-19.89±1.65**	**-12.06**	**1.8x10^{-33}**	**5.3x10^{-33}**
	Self-supervised TDANN	**7.71±1.65**	**4.67**	**3.0x10^{-6}**	**8.9x10^{-6}**
	Supervised TDANN	-2.51±1.65	-1.52	0.13	0.38
Ventral	**Intercept (MB ResNet-18)**	**48.35±1.34**	**36.14**	**5.9x10^{-286}**	
	MB ResNet-50	**-13.99±1.89**	**-7.39**	**1.4x10^{-13}**	**4.3x10^{-13}**
	Self-supervised TDANN	**21.85±1.89**	**11.55**	**7.5x10^{-31}**	**2.2x10^{-30}**
	Supervised TDANN	**18.22±1.89**	**9.63**	**6.1x10^{-22}**	**1.8x10^{-21}**

Table 2. Linear mixed-effects model to test effect of candidate model type on functional correspondence for each of the three streams. To test if there were differences between candidate models on functional correspondence with the brain, we used linear mixed-effects models, with fixed effects for candidate model type (intercept denotes multi-behavior candidate model with ResNet-18 base, i.e. MB ResNet-18) and a random intercept for each subject. Model specification was as follows: functional correspondence $\sim$ candidate model type + 1 | subject. A separate model was run for each of the three streams. Positive values indicate better functional correspondence than the MB ResNet-18 (first row) and negative values indicate worse functional correspondence. For example, the β coefficient of -0.01 for MB ResNet-50 in Ventral indicates that there is an average decrease in the correlation with brain responses of 0.01 for the MB ResNet-50 model relative to the MB ResNet-18 model, while the β coefficient of 0.18 for the self-supervised TDANN indicates a massive average increase in correlation of 0.18, again relative to the MB ResNet-18's value of 0.14, meaning that the self-supervised TDANN had an average correlation across subjects of 0.32 to Ventral brain responses. Corrected p-values indicate p-values Bonferroni-corrected for multiple comparisons between candidate model types. Significant predictors ($p < .05$) are shown in bold. MB = multi-behavior. These statistics are related to Fig 2c.

		Coefficients±SE	z-value	p-value	corrected p-value
Dorsal	**Intercept (MB ResNet-18)**	**0.11±0.00**	**24.86**	**1.8x10^{-136}**	
	MB ResNet-50	**-0.01±0.01**	**-2.74**	**.01**	**.018**
	Self-supervised TDANN	**0.10±0.01**	**21.26**	**2.7x10^{-100}**	**8.2x10^{-100}**
	Supervised TDANN	**0.05±0.01**	**11.39**	**4.7x10^{-30}**	**1.4x10^{-29}**
Lateral	**Intercept (MB ResNet-18)**	**0.15±0.01**	**19.82**	**2.1x10^{-87}**	
	MB ResNet-50	**-0.03±0.01**	**-4.98**	**6.5x10^{-7}**	**1.9x10^{-6}**
	Self-supervised TDANN	**0.09±0.01**	**15.30**	**7.9x10^{-53}**	**2.4x10^{-52}**
	Supervised TDANN	**0.02±0.01**	**2.44**	**0.015**	**0.044**
Ventral	**Intercept (MB ResNet-18)**	**0.14±0.01**	**22.42**	**2.6x10^{-111}**	
	MB ResNet-50	-0.01±0.01	-2.29	.022	.067
	Self-supervised TDANN	**0.18±0.01**	**31.74**	**4.0x10^{-221}**	**1.2x10^{-220}**
	Supervised TDANN	**0.11±0.01**	**19.56**	**3.4x10^{-85}**	**1.0x10^{-84}**

To test for the effects of TDANN spatial weightings ($\alpha \in [0.0, 0.1, 0.25, 0.5, 1.25, 2.5, 25]$) and training task (self-supervised simCLR vs. supervised object categorization) on (1) model-to-brain distance similarity (Fig. 3b-top), (2) model-to-brain functional similarity (Fig. 3b-bottom), and (3) linear regression brain predictivity, we ran repeated-measures ANOVAs separately for each stream with the factors spatial weighting and training task. Results are reported in Tables 3-11. Num DF indicates numerator degrees of freedom and Den DF indicates denoinator degrees of freedom.

Table 3. Distance similarity (r) for Dorsal.

	F	Num DF	Den DF	p-value
Spatial weighting	71.83	6.0	42.0	1.7x10^{-20}
Training task	72.21	1.0	7.0	6.2x10^{-5}
Spatial weighting:Training task	35.24	6.0	42.0	7.2x10^{-15}

Table 4. Distance similarity (r) for Lateral.

	F	Num DF	Den DF	p-value
Spatial weighting	90.60	6.0	42.0	2.0x10^{-22}
Training task	91.54	1.0	7.0	2.9x10^{-5}
Spatial weighting:Training task	45.86	6.0	42.0	7.0x10^{-17}

Table 5. Distance similarity (r) for Ventral.

	F	Num DF	Den DF	p-value
Spatial weighting	152.15	6.0	42.0	7.5x10^{-27}
Training task	258.05	1.0	7.0	8.8x10^{-7}
Spatial weighting:Training task	138.17	6.0	42.0	5.1x10^{-26}

Table 6. Functional similarity (r) for Dorsal.

	F	Num DF	Den DF	p-value
Spatial weighting	210.92	6.0	42.0	1.0x10^{-29}
Training task	520.02	1.0	7.0	7.9x10^{-8}
Spatial weighting:Training task	370.10	6.0	42.0	1.1x10^{-34}

Table 7. Functional similarity (r) for Lateral.

	F	Num DF	Den DF	p-value
Spatial weighting	189.58	6.0	42.0	9.0x10^{-29}
Training task	548.94	1.0	7.0	6.6x10^{-8}
Spatial weighting:Training task	523.86	6.0	42.0	8.2x10^{-38}

Table 8. Functional similarity (r) for Ventral.

	F	Num DF	Den DF	p-value
Spatial weighting	487.47	6.0	42.0	3.6x10^{-37}
Training task	890.56	1.0	7.0	1.2x10^{-8}
Spatial weighting:Training task	694.57	6.0	42.0	2.4x10^{-40}

Table 9. Linear regression predictivity (R^2) for Dorsal.

	F	Num DF	Den DF	p-value
Spatial weighting	264.61	6.0	42.0	1.0x10^{-31}
Training task	17.10	1.0	7.0	.004
Spatial weighting:Training task	188.09	6.0	42.0	1.1x10^{-28}

Table 10. Linear regression predictivity (R^2) for Lateral.

	F	Num DF	Den DF	p-value
Spatial weighting	797.86	6.0	42.0	1.3x10^{-41}
Training task	148.19	1.0	7.0	5.8x10^{-6}
Spatial weighting:Training task	272.05	6.0	42.0	5.9x10^{-32}

Table 11. Linear regression predictivity (R^2) for Ventral.

	F	Num DF	Den DF	p-value
Spatial weighting	747.19	6.0	42.0	5.1x10^{-41}
Training task	47.92	1.0	7.0	.0002
Spatial weighting:Training task	290.19	6.0	42.0	1.6x10^{-32}

653 Supplementary figures

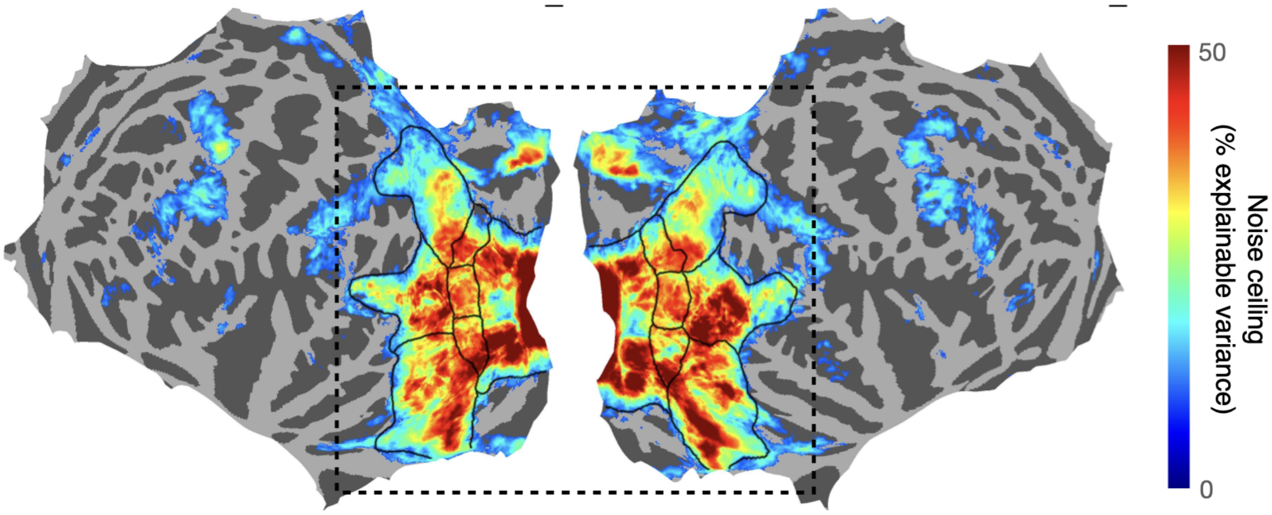

Figure S1. Voxel-wise noise ceiling estimates and ROI boundaries. Noise ceiling estimates (% explainable variance) across all image repeats per subject and then averaged across subjects, visualized on the fsaverage surface. Values are thresholded at 10% explainable variance, the cutoff used to guide drawing of the higher-level ROI boundaries. Figure illustrates that, (1) by design, much of the reliable signal is included in the ROIs drawn, and (2) the noise ceiling is high (minimum 10% explainable variance and numerous voxels above 50% explainable variance) across Ventral, Lateral, and Dorsal. Dashed black line: region shown in main text figures.

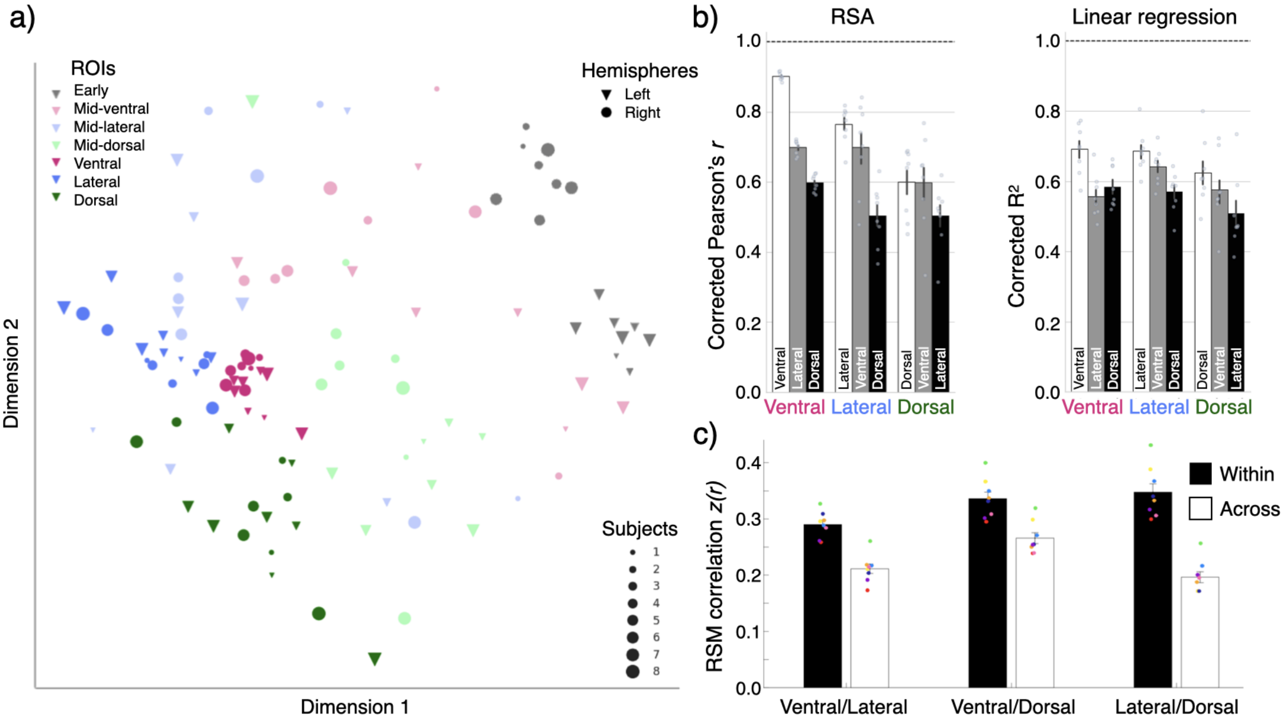

Figure S2. Representations differ across streams in NSD. (a) Multidimensional scaling of representation structure. For each individual and ROI, we computed the similarity (Pearson's *r*) between distributed responses across the ROI to all pairs of shared images, resulting in a representational similarity matrix (RSM) from which we extract the flattened lower triangle as a representation vector. Representation vectors were correlated across all subject and ROI combinations (corrected by the trial-to-trial reliability) to generate a 2nd-order RSM, which characterizes the similarity of representations across subjects and ROIs. To visualize the structure, we computed a multidimensional scaling (MDS) of this matrix. We find a rough hierarchical progression from early visual cortex (EVC) ROIs in the top-right (gray), to mid-level ROIs (light colors, middle), to high-level ROIs in the lower-left. Additionally, there is a large-scale separation by stream for high-level ROIs, rather than subject or hemisphere, with lateral high-level ROIs (blue) separated and more superior from a tight ventral cluster (magenta), which is in turn, largely distinct from the dorsal ROIs (green, though these show greater between subject variability). (b) Comparison of ROIs as models of each other, using representational similarity analysis (RSA) and linear regression (Ridge regression). Pearson's r and R^2 values are normalized by the respective noise ceilings (NC). Each dot represents a subject. White: within-ROI (i.e. subject-to-subject noise ceiling); Gray and Black bars: ROI X's prediction of ROI Y's responses. (c) To further test whether each stream showed a distinct representational structure, we parcellated cortex into 1000 equally spaced ROIs and then calculated the correlation between each pair of parcels. Each comparison was grouped based on whether both parcels were located within the same stream (black) or whether they were located in two different streams (white), revealing significantly higher correlations within than across streams for this three-stream organization (main effect of within vs. across: p=4.19x10^{-7}). The difference in parcel correlations within vs. across streams did not simply reflect anatomical proximity, as the neighboring lateral and dorsal streams showed the greatest differentiation.

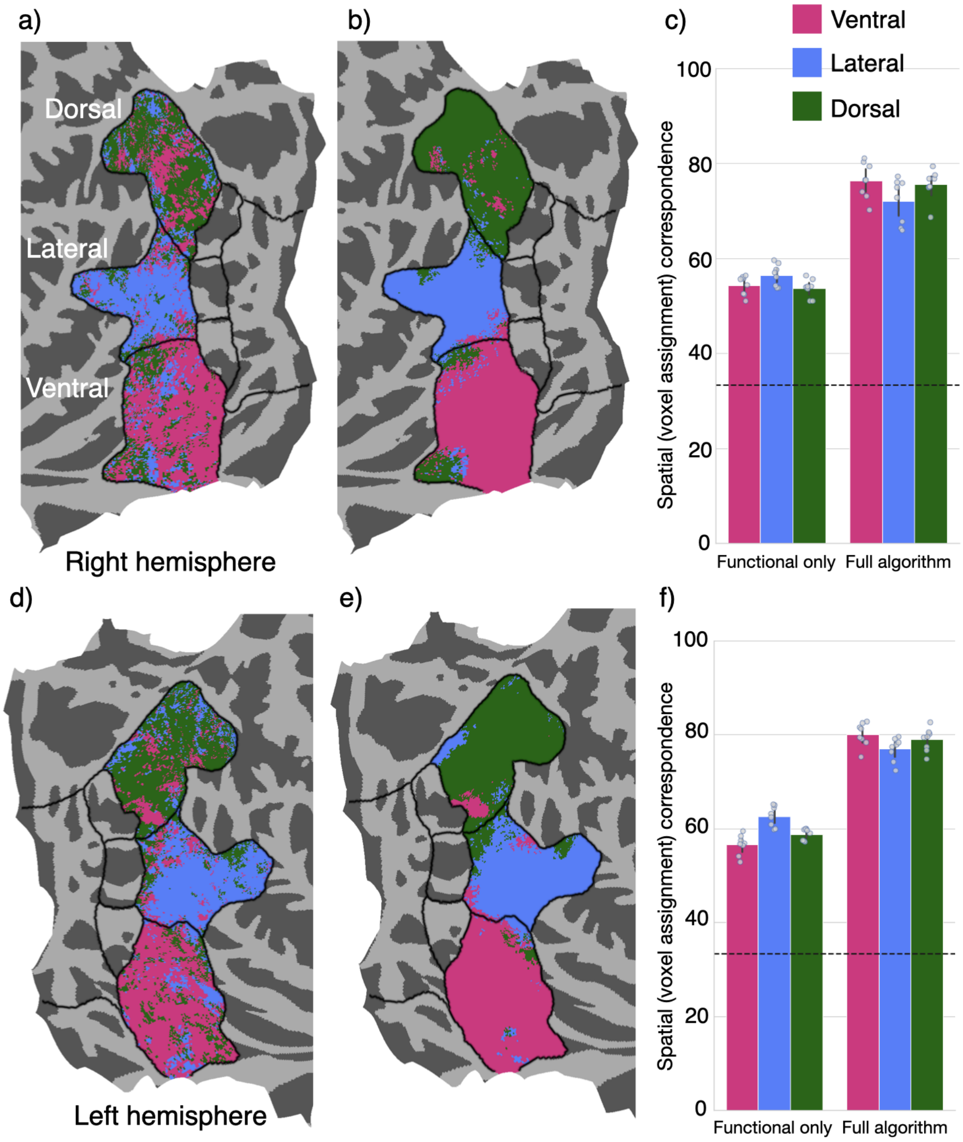

Figure S3. Validating the 1-to-1 mapping algorithm by testing how well it maps one brain onto another brain. (a) Right hemisphere voxels in Target (Subj. 2) brain colored by their assignment to streams in Source (Subj. 1) right hemisphere voxels using the algorithm [30] that matches each unit to a voxel solely based on functional similarity. Essentially, all spatial information was removed and voxels in the Target brain were assigned to voxels from the Source brain purely on the basis of how correlated they were in their responses to the shared stimuli across subjects. We refer to this mapping as functional only. (b) Right hemisphere voxels in Target (Subj. 2) brain colored by their assignment to Source (Subj. 1) right hemisphere voxels using the full algorithm that additionally incorporates a gentle smoothness constraint, which encourages neighboring voxels in the target space to "pick" neighboring voxels in the source space (algorithm used in main text). (c) Spatial correspondence between brains using the functional only or full algorithm. Bars: Comparison of right hemisphere voxel-to-voxel correspondence. Data show substantially higher correspondences when using the full algorithm in both the right (Ventral: $t(7) = 41.3, p = 1.3\text{x}10^{-9}$, Lateral: $t(7) = 15.6, p = 1.1\text{x}10^{-6}$, Dorsal: $t(7) = 27.0, p = 2.5\text{x}10^{-8}$) and left hemisphere (panel f; Ventral: $t(7) = 18.2, p = 3.7\text{x}10^{-7}$, Lateral: $t(7) = 18.1, p = 4.0\text{x}10^{-7}$, Dorsal: $t(7) = 21.8, p = 1.1\text{x}10^{-7}$). Data plotted for high-level ROIs across three streams, averaged across source subjects for each target subject. Color represents stream (pink: Ventral, blue: Lateral, and green: Dorsal). Dotted line: chance level (33%). Error bar: 95% CI, each dot is a subject. Data show that the full algorithm achieves a better brain-to-brain mapping. (d), (e), and (f), same for left hemisphere.

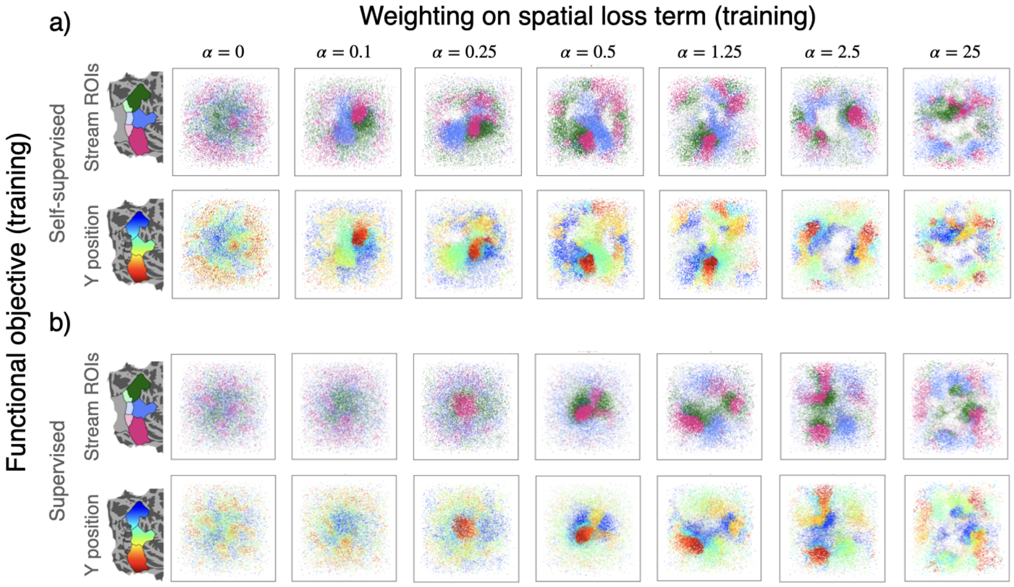

Figure S4. Model-to-brain 1-to-1 mapping visualized on the simulated cortical sheet of the last convolutional layer of the TDANN model. TDANN model-to-brain mapping visualized on the simulated cortical sheet, for TDANNs trained using the self-supervised, SimCLR task (a) or the supervised, object categorization task (b). This visualization is the reverse of the visualization in Fig. 2a and Fig. 3a of the main text. Here, each square panel shows model units (each dot is a model unit) on the simulated cortical sheet. Units are colored based on the spatial location of their assigned voxel in an example target brain. Opacity of the units reflects the strength of the model-to-brain correlation between responses to NSD images and units are colored by stream (top) or superior-to-inferior spatial gradient (y-position in flat map, bottom). This second color scheme is "stream-agnostic" in that it does not presuppose the existence of three streams, yet stream clustering emerges at self-supervised $0.25 \leq \alpha \leq 0.5$. Each column displays the model-to-brain mapping for an example subject and model seed for a particular spatial weight.

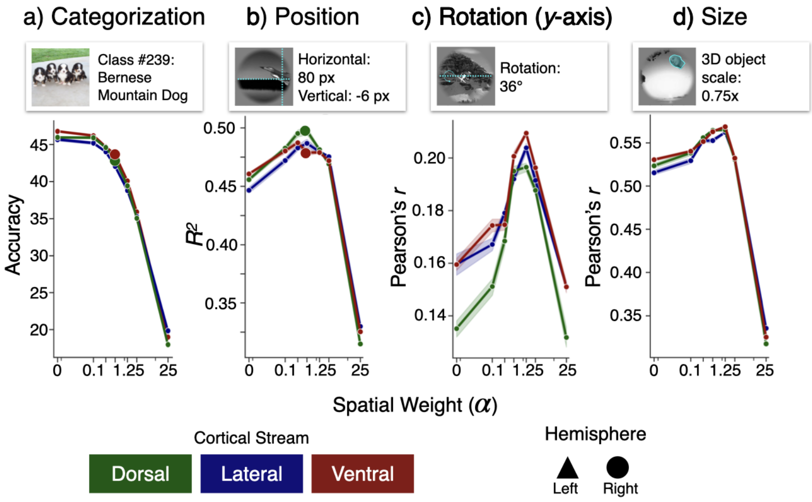

Figure S5. Adding the spatial loss component to the TDANN during training can improve later transfer performance for some tasks. Transfer performance for units from the self-supervised TDANNs across a range of weightings (α) on the spatial loss function for four tasks: object categorization, object position estimation, object pose estimation (y-axis rotation), and object size estimation. For three of the four tasks: position, pose, and size estimation, the addition of a spatial loss term improves performance, with peak performance in the same range of weighting on the spatial loss term that leads to best correspondence with the brain $(0.25 \leq \alpha \leq 1.25)$. Units were divided by their assigned stream and then assessed on transfer performance using supervised linear readouts. Results are averaged across model seeds (5), subjects (8) and hemispheres (2) totaling 80 models per point, with the exception of (a) where results represent only one model seed, due to compute constraints. Shaded error bar: ± SE. Larger circles indicate model by stream combinations evaluated main text Figure 4.

References

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

2. Eshed Margalit, Hyodong Lee, Dawn Finzi, James J DiCarlo, Kalanit Grill-Spector, and Daniel LK Yamins. A unifying principle for the functional organization of visual cortex. *bioRxiv*, pages 2023–05, 2023.

3. Jonas Kubilius, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel LK Yamins, and James J DiCarlo. Cornet: modeling the neural mechanisms of core object recognition. *bioRxiv*, 2018.

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

5. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

6. Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

7. Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

8. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28, 2015.

9. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

10. Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

11. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012.

12. Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, et al. Vissl, 2021.

13. Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, page 407007, 2020.

14. Daniel J Felleman and David C Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.

15. Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

16. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. 2020. doi: 10.48550/ARXIV.2002.05709. URL https://arxiv.org/abs/2002.05709.

17. Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10(11):e1003915, 2014.

18. Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

19. Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib Majaj, Elias Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, et al. Brain-like object recognition with high-performing shallow recurrent anns. *Advances in Neural Information Processing Systems*, 32, 2019.

20. Markus Meister, Rachel OL Wong, Denis A Baylor, and Carla J Shatz. Synchronous bursts of action potentials in ganglion cells of the developing mammalian retina. *Science*, 252(5008):939–943, 1991.

21. Jinwoo Kim, Min Song, Jaeson Jang, and Se-Bum Paik. Spontaneous retinal waves can generate long-range horizontal connectivity in visual cortex. *Journal of Neuroscience*, 40(34):6584–6599, 2020.

22. Todd McLaughlin, Christine L Torborg, Marla B Feller, and Dennis DM O'Leary. Retinotopic map refinement requires spontaneous retinal waves during a brief critical period of development. *Neuron*, 40(6):1147–1160, 2003.

23. Xinxin Ge, Kathy Zhang, Alexandra Gribizis, Ali S Hamodi, Aude Martinez Sabino, and Michael C Crair. Retinal waves prime visual motion detection by simulating future optic flow. *Science*, 373(6553):eabd0830, 2021.

24. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

25. Jacob S Prince, Ian Charest, Jan W Kurzawski, John A Pyles, Michael J Tarr, and Kendrick N Kay. Improving the accuracy of single-trial fmri response estimates using glmsingle. *Elife*, 11:e77599, 2022.

26. Liang Wang, Ryan EB Mruczek, Michael J Arcaro, and Sabine Kastner. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, 25(10):3911–3931, 2015.

27. Dawn Finzi, Daniel LK Yamins, Kendrick Kay, and Kalanit Grill-Spector. Do deep convolutional neural networks accurately model representations beyond the ventral stream? In *2022 Conference on Cognitive Computational Neuroscience*, 2022.

28. Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. *Advances in Neural Information Processing Systems*, 34:4738–4750, 2021.

29. Meenakshi Khosla and Alex H Williams. Soft matching distance: A metric on neural representations that captures single-neuron tuning. *arXiv preprint arXiv:2311.09466*, 2023.

30. James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

31. Chengxu Zhuang, Siming Yan, Aran Nayebi, Martin Schrimpf, Michael C Frank, James J DiCarlo, and Daniel LK Yamins. Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3):e2014196118, 2021.

32. Marco Del Giudice. Effective dimensionality: A tutorial. *Multivariate behavioral research*, 56(3):527–542, 2021.

33. Eric Elmoznino and Michael F Bonner. High-performing neural network models of visual cortex benefit from high latent dimensionality. *bioRxiv*, pages 2022–07, 2022.

34. Dawn Finzi, Jesse Gomez, Marisa Nordt, Alex A Rezai, Sonia Poltoratski, and Kalanit Grill-Spector. Differential spatial computations in ventral and lateral face-selective regions are scaffolded by structural connections. *Nature communications*, 12(1):2278, 2021.

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

36. Marisa Nordt, Jesse Gomez, Vaidehi S Natu, Alex A Rezai, Dawn Finzi, Holly Kular, and Kalanit Grill-Spector. Cortical recycling in high-level visual cortex during childhood development. *Nature human behaviour*, 5(12): 1686–1697, 2021.

37. Ha Hong, Daniel LK Yamins, Najib J Majaj, and James J DiCarlo. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, 19(4):613–622, 2016.

38. Aran Nayebi, NC Kong, C Zhuang, JL Gardner, AM Norcia, and DL Yamins. Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *BioRxiv*, pages 1–37, 2022.

39. Rishi Rajalingham, Elias B Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.