University of Trento
Center for Mind/Brain Sciences (CIMeC)
PhD Program in Cognitive and Brain Sciences

~ · ~

37th Cycle (2021–2026)

Modeling Cognition by Pruning and Topography-Learning in Deep Neural Networks

PhD Student
Nhut Truong

Professor
Uri Hasson

Final examination date: 24 April 2026

Thesis Summary

Deep artificial neural networks (DNNs) now match or exceed human accuracy on many benchmarks, yet high performance alone does not imply human-like representational structure. This motivates developing algorithms that not only improve benchmarks but also produce representations that are better aligned with human brain or behavior, so that models can be considered as mechanistic accounts for "in silico" experiments in neurosciences. This thesis contributes to the intersection of cognitive neuroscience and AI by studying how biologically inspired algorithms can produce representations that better align with human cognition, and how they shape the internal representations of models beyond task performance.

Specifically, we focus on two complementary parts that potentially bring cognitive neuroscience and AI closer together: (i) explaining and improving representational alignment between pretrained models and human cognition, using behavioral similarity judgments as a proxy for mental representational geometry; and (ii) enforcing brain-like correlations in topographic networks, to assess their capabilities of modeling high-level visual cortex, and to understand how these correlations impact the network's performance and representations. Across both parts, the common aim is to test whether models can align with specific aspects of human representational structure, and to characterize how these constraints reshape model representations.

In Part I, aiming to improve and explain the alignment between model representations and human semantic knowledge, we focus on representational alignment with human similarity judgments. Rather than relying on all information in the full embeddings extracted from the deep networks, as is common in the literature, we identify the relevant information of the model's representational space that best match human similarity judgments of a certain semantic category. Technically, we implement this using structured pruning over learned feature maps or units in convolutional neural networks to select a subset that can improve the alignment with human judgments. This part consists of three studies. In Study 1.1, we introduce a statistic quantifying how much each feature map contributes to alignment with human similarity judgments called Alignment Importance Score (AIS), which is not only used for improving alignment, but also for explainable AI analyses. By structured pruning low-AIS feature maps, we improve out-of-sample prediction of human judgments while reducing the number of feature maps. Moreover, AIS-pruning can select feature that produces image-space heatmaps highlighting the visual information most relevant for explaining human comparisons among objects, supporting mechanistic interpretability. In Study 1.2, we investigate whether

alignment is driven by a small set of specialized units or by population-level geometry. Using numerosity as a controlled domain, we find that the units critical for capturing similarity judgments, identified via the same pruning method, do not overlap with the units identified via a traditional statistical test (ANOVA). This suggests that human-aligned representation in models is an emergent property of population-level geometry rather than the result of isolated, specialized units. Study 1.3 aims to identified the core representational geometry of an existing model, by extending pruning beyond layer-wise and explicitly supervised targets in previous studies. We introduce Correlation Retaining Iterative Structural Pruning, a geometry-guided procedure that removes redundant feature maps or units, aiming to approximate a target representational geometry. This task-agnostic algorithm formalizes the pruning logic from the earlier studies into a more general framework that can be used to compress models while either retaining a model's own geometry or aligning geometry to external targets such as similarity judgments.

Part II focuses on correlation-based, end-to-end topographic models, where units are arranged on a physical space, then correlated activities are enforced among nearby units via training, thus their activations can be visualized on smooth, brain-like spatial maps. Specifically, we study the capability of capturing cortical organization, and the computational properties of topographic regularizers that encourage correlated representations. This part includes two studies. In Study 2.1, we evaluate whether a current leading state-of-the-art topographic model can capture the fine-grained organization of the human occipitotemporal cortex. Focusing on the action dimension - the degree to which an object is associated with physical manipulation - our results show that while the model successfully captures broad divisions like animacy, it fails to produce an action-related gradient. This finding suggests that generic spatial constraints may be insufficient, and additional requirements are needed to account for the specialized organization of human high-level visual cortex. In the final study, Study 2.2, we investigate the computational advantages of correlated constraints in topographic models and how they shape the network's internal representations, beyond topographic map visualization - the primary goal in most of previous work. We systematically compare two commonly used local constraints in end-to-end convolutional networks: Activation Similarity, which encourages nearby units to have similar activations, and Weight Similarity, which encourages nearby units to develop similar afferent weight vectors. Our analysis shows that the two constraints can produce robustness not only to input perturbations but also to parameter noise. Moreover, the two constraints produce qualitatively different computational properties at the representation levels.

Overall, this thesis investigates cognitively inspired approaches, implemented through structured pruning and topographic constraints, as methods for aligning human-DNN representations, and for shaping DNNs' internal representations. Practically, our studies support an alternative approach to improving human–model alignment: instead of using full embeddings, we improve alignment by selecting the most relevant information within the model. Under this view, AIS and CRISP provide structured pruning tools that improve or preserve alignment-relevant geometry while compressing networks and enabling interpretability, including heatmaps for explaining human similarity comparisons. For topographic modeling, we demonstrate the engineering benefit of locally regularizing correlations, and show that either weight-based or activation-based constraints can be a preferred choice to handle certain types of noise. Theoretically, we show that human-like alignment is better characterized as an emergent property of population-level geometry rather than isolated expert units, and we identify that current topographic models likely require additional constraints to capture the fine-grained organization of the human visual cortex. Collectively, these findings contribute toward more transparent, robust, and cognitively aligned models for both practical applications and in silico cognitive science research.

Contents

Introduction

The deep learning advances in the last decade has brought AI and cognitive neuroscience into closer contact. Deep neural networks (DNNs) can now perform many perceptual and cognitive tasks at or beyond human-level accuracy, making them useful not only as engineered solutions but also as computational objects for modeling cognition. This creates an interaction in which researchers use DNNs to test and formalize hypotheses about human cognition, and conversely use insights from human behavior and the brain to design more human-like systems. Two prominent directions in this field are (i) evaluating and improving alignment between models and human behavior or neural measurements (Schrimpf, Kubilius, H. Hong, Najib J Majaj, et al. 2018a; Sucholutsky et al. 2023; Demircan et al. 2024; Oota et al. 2023; M. W. Mathis and A. Mathis 2025), and (ii) injecting human-inspired inductive biases (built-in assumptions that guide learning) into models to shape their internal representations and behavior (Kriegeskorte 2015; Lake, Ullman, et al. 2017; Goyal and Bengio 2022). This thesis comprises a set of experimental papers that address both directions. In Part I, we aim to model and explain human similarity judgments as a task within the larger human-model alignment problem. In Part II, we study topographic constraints as a biological-inspired inductive bias, testing their ability to capture organization in the visual cortex and characterizing their computational benefits and representational consequences.

0.1 Modeling human similarity judgments via pruning

As AI models have continued to improve, they now match or exceed human performance on many benchmarks (K. He et al. 2015; Achiam et al. 2023; Maslej et al. 2025). However, strong task performance is not necessarily based on human-like internal computations. Models can achieve high performance while relying on different features, exploiting dataset shortcuts, or making errors that humans do not (Geirhos, Rubisch, et al. 2018; Schrimpf, Kubilius, H.

Hong, Najib J Majaj, et al. 2018a; Geirhos, Jacobsen, et al. 2020; Wichmann and Geirhos 2023). This raises an important question: when a model performs well, does its success base on human-like representations? Ensuring human-like internal mechanism has many practical benefits, for example, interpretability, transparency, and effective collaboration with humans. Moreover, mechanistic alignment between models and behavioral or brain data is necessary for building "in silico" models in computational cognitive neuroscience.

The alignment problem spans many levels, from alignment of goals and values, to performance, to alignment of cognitive and neural representations (Shen et al. 2024; Ji et al. 2025; Sucholutsky et al. 2023; Oota et al. 2023; M. W. Mathis and A. Mathis 2025). In our work, we focus on representational alignment: the degree to which the internal representational structure of a model corresponds to that of humans. In deep learning models, representations are typically taken as activation vectors from specific layers containing useful features for the tasks. In humans, representations can be computed from behavioral measures or from neural recordings via fMRI, MEG, EEG, etc (Martin N Hebart, Contier, et al. 2023; Sucholutsky et al. 2023; Oota et al. 2023; M. W. Mathis and A. Mathis 2025; López-Cardona et al. 2025). Work on representational alignment includes measuring the mismatch, bridging the two sides by mapping them into a shared representational space, and improving the match across diverse modalities (images, videos, text, audio, odorants, etc). The results vary depending on the task, dataset, and the specific alignment goal (Sucholutsky et al. 2023). One common way to assess alignment is through representational geometry, rather than matching representations unit-by-unit (Kriegeskorte 2015; Sucholutsky et al. 2023). Representational geometry is the pattern of similarities or dissimilarities among stimuli, which can be viewed as the arrangement of stimuli relative to each other in a representational space. This can be operationalized as a representational similarity or dissimilarity matrix (RSM/RDM).

This motivates us to narrow our focus to the alignment of similarity judgments. For example, in a normal context, if people have to pick the least similar object among a dog, a cat, and a car, they are most likely to pick the car ("odd-one-out"). Perceiving how similar objects are to each other reveals the high-level internal representations of individual human and models . In fact, similarity is an important concept in cognitive science: it helps humans organize complex raw information from sensory input, shapes how humans form categories, and supports downstream functions such as memory and reasoning (Roads and Love 2024). It is considered a more general task than object classification (Roads and Love 2021). Behavioral similarity judgments provide a direct way to estimate this structure, for example, through pairwise

similarity ratings (Peterson et al. 2018) or odd-one-out task (Roads and Love 2024; Martin N Hebart, Contier, et al. 2023). Similarity can also be estimated from brain activity by applying representational similarity measures to neural responses (see datasets for this task in Roads and Love 2024), although this is more noisy than behavioral data (Mur et al. 2013). In both cases, these data can then be compared with the outputs of neural network models, or with internal representations from intermediate layers of the models.

Next, we provide a brief review of models for capturing human similarity judgments, pruning methods (which we apply for this task), and the thesis contributions.

0.1.1 Modeling human similarity judgments

Common pre-deep learning methods for modeling HSJs include geometric, set-theoretic, and graph-based approaches (Roads and Love 2024). Geometric models map behavioral data into a low-dimensional space (e.g., Multidimensional scaling; Shepard 1962; Joseph B Kruskal 1964) and then apply a metric over this representational space to infer the similarity between objects. Set-theoretic models represent an item as a set of features (e.g., a cat is an animal, has fur, is carnivorous), and define similarity in terms of shared versus distinctive features (Tversky 1977). Graph models treat concepts as nodes and relationships as links, such that two concepts are judged more similar when they are separated by only a few steps (Carroll 1976; Cunningham 1978). Together, these models can capture and explain latent mental structure that are reflected in HSJs. However, a major limitation is that they typically rely on hand-crafted, human-defined representations (e.g., feature lists) and well-controlled, simple, non-ecological stimuli (see typical examples in Battleday et al. 2021). They lack the feature-extraction capabilities of deep learning architectures, making them difficult to scale to the complexity of real-world environments.

Deep neural networks can address this problem thanks to the power of high-level feature extraction from natural stimuli. Kamila M Jozwik et al. (2017) showed that the later layers of AlexNet and VGG-16, two common convolutional neural networks (CNNs), outperform human label feature-based models (e.g., parts, colors, textures) on judging similarity in natural images, although the models are not trained explicitly to do this task. Similarly, Kubilius, Bracci, et al. (2016) demonstrated that higher-layer activations of CNNs show high correlation with human perceptual shape judgments from natural objects. Moving beyond just using raw activations as representation, odPeterson et al. (2018) learn reweighting factors for each feature, significantly boosting the performance of predicting HSJs in many models. Attarian et al. (2020)

learn a linear transformation of the penultimate layer of VGG-16, reducing the dimension to half but improving the prediction of HSJs. Muttenthaler, Linhardt, et al. (2023) extend the linear transformation to align the global structure of similarity, such as the relationship between high-level concepts (food and drinks), while preserving the similarity structure within concepts (types of foods), showing benefits towards downstream tasks.

Martin N Hebart, Zheng, et al. (2020) showed that object embeddings can be learned end-to-end in a shallow neural network to predict odd-one-out judgments. Given a large amount of training data, this results in only 49 highly interpretable dimensions, opening up the possibility to interpret the semantic meaning in each dimension. These dimensions employ semantic properties (e.g., "food related" and "fire related"), contrasting with the dominant visual strategy (e.g., color, shape, materials) in models (Mahner et al. 2025). Jha et al. (2023) added a bottleneck layer after the penultimate layer and trained it end-to-end to predict HSJs, and found that only a small number of bottleneck dimensions (around 10-60 compared to 4096 original dimensions) is sufficient to achieve equal prediction of HSJs. Muttenthaler, Dippel, et al. (2022) and Roads and Love (2021) showed that improvements in classification accuracy do not entail better modeling HSJs. Moreover, training objective, such as supervised or self-supervised, and the size and diversity of training datasets drive the performance of modeling HSJs rather than the network architectures (Muttenthaler, Dippel, et al. 2022).

Reweighting or transforming the activations (e.g. Peterson et al. 2018; Kaniuth and Martin N Hebart 2022a; Attarian et al. 2020; Jha et al. 2023) can be viewed as that the models can learn the required features for modeling HSJs, but they need to be post-processed to boost the performance. This can be done by reweighting the original features (Peterson et al. 2018; Kaniuth and Martin N Hebart 2022a), or transformed the original dimensions into latent dimensions containing useful features for this task (Attarian et al. 2020; Jha et al. 2023). An alternative view is that the models already learns sufficient features for the task; they just need selection of specific features for computing the similarity . This is particularly relevant for modeling the representation of specific categories where only a small number of dimensions is relevant. For example, it is improbable that humans consider all features as in models to judge similarity between two dogs, as maybe animal-related features are useful enough. Based on this assumption, Tarigopula et al. (2023) applied a feature selection process to prune irrelevant features from the activations to model HSJs, improving out-of-sample prediction over reweighting (Peterson et al. 2018). Bavaresco, Truong, et al. (2025) applied pruning to show that the prediction of HSJs for different semantic categories employs different subspaces

of the feature set, and even within the same category the relevant subspace can vary across comparisons (e.g., highly-similar vs weakly-similar objects).

0.1.2 Pruning in machine learning research

Pruning is a subfield of research in machine learning. Pruning aims to compress models while maintaining performance, or accept some loss in performance for smaller models, to speed up running in real-world applications, especially on edge devices (see review H. Cheng et al. 2024; F. Chen et al. 2024; Menghani 2023; Y. He and Xiao 2023; Marinó et al. 2023; Lê et al. 2023). Common vision models are used in computational cognitive neuroscience like AlexNet, VGG, or ResNet (trained on CIFAR or ImageNet-1K) can be pruned 90% of the weights, with only less than 10% percentage point drop in accuracy (Y. He and Xiao 2023; H. Cheng et al. 2024), suggesting that these models are heavily over-parametrized.

Pruning techniques can be divided as unstructured or structured. Unstructured pruning means freely removing any weights in the model, while structured pruning removes an entire set of weights associated with a unit in a fully connected layer, or entire filters in a convolutional layer. Structured pruning can speed up model inference without specialized hardware support because it reduces tensor dimensions, unlike unstructured pruning. Criteria for pruning include weight, activation, or gradient magnitude, loss change, and saliency. Pruning can be applied to pretrained models, while training models, or even before training (H. Cheng et al. 2024). A pruning hypothesis, called the lottery ticket hypothesis, states that given a randomly initialized network there exists a subnetwork that can be trained to match the performance of the entire network (Frankle and Carbin 2018; E. Malach et al. 2020).

Here, to model HSJs, we focus on structured pruning: besides achieving speed-ups in running models, more importantly, structured pruning can improve interpretability of existing models, e.g., we can analyze kept/pruned units or feature maps and decoding the semantic meaning of those units. This gives an advantage over reweighting or transforming the activations to model HSJs (e.g. Peterson et al. 2018; Kaniuth and Martin N Hebart 2022a; Attarian et al. 2020; Jha et al. 2023), as they change the profile of units and therefore less straightforward to interpret at unit level compared to pruning. Moreover, to study the utility of common foundation networks as models of vision, we apply pruning to common pretrained models, rather than training new models from scratch for explicitly learning the task (e.g. Martin N Hebart, Zheng, et al. 2020; Mahner et al. 2025).

0.1.3 Thesis contributions

In the first part of the thesis, moving beyond correlations between models and humans, we try to develop a mechanistic understanding of what internal structures support representational alignment via similarity judgments. We propose structured pruning as both a predictive, explanatory, and a diagnostic tool to study human-model alignment.

Study 1.1: Pruning for predicting similarity judgments and explaining human comparisons

In Tarigopula et al. (2023), pruning is shown to be an effective tool to predict HSJs, but the underlined explanation on what high-level semantic features drive the match is unexplored. Here, we use pruning both as a predictive and an explainability tool for the alignment problem. We demonstrate that HSJs within a category (e.g. animals) can be predicted by specific latent dimensions in many CNNs (VGG, ResNet, DenseNet, Inception, etc) that can be isolated and visualized.

In Study 1.1, we introduce the Alignment Importance Score (AIS), a metric derived by pruning feature maps in convolutional layers to quantify their unique contribution to the alignment of model- human similarity judgments. First, we show that selectively pruning low-AIS features significantly improves the model's ability to predict HSJs of natural objects (data is from Peterson et al. 2018), while on average reduces the number of feature maps in convolutional layers to almost half. More importantly, by projecting these scores back into image space, we provide a method for generating heatmaps that explain the visual information underlying human comparisons, showing a dissociation between cognitively relevant features and standard visual saliency (J. Lou et al. 2022).

Study 1.2: Pruning for reassessing number-detector units in CNNs

If study 1.1 provides a tool to identify features in concrete objects, study 1.2 asks whether this alignment holds for abstract concepts. More importantly, here pruning is used as a diagnostic tool to study representational geometry: is alignment supported by highly specialized "expert" units, or by the population-level representational geometry?

To test this, we use numerosity as a case study. Number sense, the ability to perceive the number of items in a set without counting, is found in both human infants and adults, as well as some animals e.g. monkey, crow, chick (Viswanathan and Nieder 2013a; Wagener, Loconsole, Helen M Ditz, et al. 2018a; Kobylkov et al. 2022a). Here we also consider representational

geometry among non-symbolic numbers, represented by images of sets (i.e., small dots on a clean background). Numerosity offers a controlled and scalable similarity structure, providing more data for representational geometry than that of naturalistic images . Therefore, it is easy to simulate behavioral judgments, providing a simple setting for testing alignment. Previous research suggested that models can distinguish numbers thanks to "number-detector" units (Nasr et al. 2019a; G. Kim et al. 2021a), identified via statistical tests such as ANOVA, which are specialized units tuned to specific quantities (analogous to "number neurons" in the brain; Kutter et al. 2018; Dijk et al. 2022).

We use pruning as a diagnostic tool to test whether number-detector units are necessary for population-level numerosity representations in CNNs. First, following Nasr et al. (2019a) and G. Kim et al. (2021a), we apply ANOVA on the activations of many layers in multiple CNN architectures to obtain the set of number-detector units. Second, we prune the same activations using a simulated human number similarity judgments as the alignment target. The results show that these two sets generally do not overlap. This suggests that although number-detector units can emerge in CNNs, they are not critical for capturing population-level numerosity structure aligned with human behavior, therefore the alignment in CNNs is better described as an emergent property of representational geometry rather than the result of isolated, specialized units.

Study 1.3: Sparsity-guided pruning to preserve representational geometry and model human similarity judgments

Instead of treating accuracy as the objective to maintain during pruning, in Study 1.3 we treat a model's representational geometry as the objective to retain. Technically, we aim to identify a smaller subnetwork that preserves the similarity structure of the original model. Achieving a smaller model with similar representational geometry provides benefits similar to knowledge distillation - smaller, faster models, and better accuracy than training from scratch (Hinton et al. 2015; Gou et al. 2021) - but here we obtain a compressed version of the same model instead of training another model as in distillation. This makes pruning a task-agnostic method for isolating the subspaces that carry representational structure, enabling mechanistic analyses of what information is retained or removed in the original models.

Implementing this idea requires moving beyond the layer-wise, supervised pruning used in Studies 1.1 and 1.2. Layer-by-layer pruning can propagate changes to later activations, and supervision with pairwise HSJs requires large dataset collection. To address both issues, we develop an unsupervised, network-level procedure that compares the representational geometry of subnetworks to that of the original unpruned model. The same framework can also use ex-

ternal targets such as HSJs, making it a predictive tool as in previous studies, but using HSJs only at evaluation.

To do this, we develop a method called Correlation Retaining Iterative Structural Pruning (CRISP), which preserves representational geometry while removing redundant information. Following the motivation of structured pruning, we target sparse activations and remove entire neurons or feature maps in CNNs (Huan Hu et al. 2016), both to speed up model running and to provide a potential basis for explainability. CRISP introduces a geometry-guided structural pruning framework that isolates the subspaces carrying representational structure, either preserving a model's own geometry or aligning it to HSJs, and opens up the possibility of interpretation on what semantic feature is kept and what is removed. In this way, CRISP formalizes the pruning logic used in the previous studies into a task-agnostic algorithm. The preliminary results show that CRISP can prune half of the number of units in fully connected layers with less than 10% decreasing the match with the geometry from original, unpruned model.

Overall, the first part of the thesis addresses a question within the broader human–machine alignment problem, which is modeling similarity judgments, using structured pruning as a methodological tool. We show that pruning is not only a predictive tool that can boost or maintain task performance with smaller models, but also an explanatory tool for human visual comparisons, and a diagnostic tool for studying the representational geometry of models.

0.2 Topographic models

Across both machine learning and computational cognitive science, biologically-inspired inductive biases (i.e. built-in assumptions that guide learning) a re often developed with two objectives: boosting the performance of artificial systems and increasing alignment with human behavior and neural data, thereby improving interpretability and supporting mechanistic explanations. In machine learning practice, inductive biases are often evaluated mainly on task performance, such as improvements in classification accuracy, rather than their faithfulness to underlying neurobiological mechanisms (Hassabis et al. 2017). In contrast, computational cognitive science treats biological constraints as central modeling assumptions, aiming to replicate some aspects of behavioral and neural mechanisms. Rather than optimizing performance alone, this approach seeks mechanistic explanations and architectures, and may accept engineering trade-offs to achieve these goals (O'Reilly 1998; D. L. Yamins and DiCarlo 2016; Pulvermüller et al. 2021; Momennejad 2023; Cohen et al. 2022; Jeon and T. Kim 2023; Ororbia et al. 2024). Topographic models mainly be-

long to the latter category, because their focus is to model the physical cortex rather than to maximize task performance.

Specifically, topographic models are computational models inducing smoothness among neighboring units, producing a topography that can be visualized in a physical space. In these models, each neuron is assigned a position in a physical space, such as a 2D grid or even a 3D volume, so that two neurons can be close together or far apart. This adds an extra design compared to simply connecting or disconnecting neurons in standard models. In addition to spatial layout, neurons in these models are usually positioned so that nearby neurons exhibit similar activation patterns, or nearby neurons connect with each other while distant neurons connect rarely or not at all.

These constraints are often motivated by Hebbian locality ("neurons that fire together, wire together", Hebb 1949) and minimization of wiring cost (Chklovskii and Koulakov 2004), which has been observed across many areas of the mammalian cortex (R. B. Levy and Reyes 2012; X. Jiang et al. 2015; Ringach et al. 2016; Ding et al. 2025). Consequently, when we visualize the functions of a group of nearby neurons, we can find that they share similar functions, creating continuous or discontinuous patches of activation. These observations have been documented in many regions of the mammalian brain, most notably in the visual cortex, but also in the auditory cortex (Humphries et al. 2010; Saenz and Langers 2014), the somatosensory cortex (Penfield and Boldrey 1937; Wong et al. 1978), and in the numerosity representation of parietal cortex (Harvey et al. 2013). In primary visual cortex, topography can be seen in several feature maps: orientation preferences appear as a collection of pinwheels (Bonhoeffer and Grinvald 1991; Ohki et al. 2006), spatial-frequency preferences appear as discrete islands (Shoham et al. 1997; Nauhaus et al. 2012), and color preferences appear as blobs (M. S. Livingstone and Hubel 1984; H. D. Lu and Roe 2008). In higher visual cortex, such as ventral temporal cortex in humans, there are category-selective areas for certain categories, such as scenes, faces, bodies, hands, tools, and written words (Kanwisher 2010; Cortinovis, Peelen, et al. 2025). Since topographic maps occupy much of the cortex, provide metabolic benefits, and are central to many theories of visual information processing, they are considered central to sensory processing (Kaas 1997; Silver and Kastner 2009; Patel et al. 2014). Topographic models aim to capture some or many of these organizations, explain the mechanisms behind map formation, and further generate hypotheses about how the cortex is organized.

In the next section, a brief history of this type of model will be sketched, followed by the rationale of the second part of the thesis.

0.2.1 A brief history of topographic models

The first topographic models were developed nearly half a century ago, long before the rise of deep learning–based topographic modeling. At that time, models relied largely on simple algorithms, simple tasks, and simple stimuli, yet they were able to demonstrate that such mechanisms could capture important characteristics of early visual cortex. One of the earliest examples is the work of Von der Malsburg (1973), who proposed a model consisting of a two-dimensional grid of neurons which was exposed to oriented bars. Using an associative learning rule, where connections between afferent inputs and cortical neurons are strengthened when both are active , following the Hebbian principle, each neuron became selective to a specific orientation. Neurons with similar orientation preferences formed clusters on the two-dimensional sheet, resembling orientation columns in primary visual cortex. Willshaw and Von Der Malsburg (1976) extended this self-organization principle to model retinotopic maps, demonstrating ordered, topographic wiring between two neural sheets rather than within a single sheet. Durbin and Mitchison (1990) framed cortical maps as a dimension-reduction problem, in which a high-dimensional feature space is mapped onto the two-dimensional cortical surface while preserving local neighborhood relationships. Their elastic-net model explains the emergence of pinwheel-like orientation maps when retinal position and orientation selectivity are mapped jointly. Around the same time, Obermayer et al. (1990) adopted self-organizing maps (SOMs; Kohonen 1982) and obtained similar results, but with important technical differences: the mapping between cortical sheets was optimized continuously in an iterative process similar to modern deep learning , and the model operated directly on high-dimensional stimuli rather than on preprocessed feature vectors. Beyond retinotopy and orientation maps, related models have been proposed to account for other forms of cortical organization, including color blobs (Barrow et al. 1996), and ocular dominance columns in V1 (see reviews by Erwin et al. 1995 and Swindale 1996).

Most of these models (e.g., SOMs and elastic nets) are based on competition and Hebbian learning, meaning that neurons which respond together strengthen their connections while competing with other neighboring neurons, and topography can emerge through learning. There are also other approaches beyond Hebbian learning, such as correlation-based models, in which cortical cells are learned from the statistical correlations of retinal activity, and spectral models, which explain orientation and ocular dominance maps by directly analyzing the statistical structure of visual input using mathematical filters (Erwin et al. 1995). The stimulus sets used in most of these models are simple and limited: either symbolic inputs or points in low-dimensional spaces,

elongated or circular patches (Swindale 1996). These models cannot capture organization in higher visual cortex, because they cannot automatically learn to represent complex objects from raw pixels. However, they still provide important insights into cortical map formation, showing that topography can emerge from general constraints such as cortical geometry, wiring cost, and self-organization. Moreover, these models can reproduce species-specific differences in map organization (e.g., macaque versus cat) and generate concrete, testable predictions about the dynamics of map development (Goodhill 2007).

Jacobs and Jordan (1992) pioneered an early form of end-to-end learning, giving computational support to the hypothesis that the cortical organization is a result of a biological bias toward minimizing connection length, due to the high biological cost of long-range wiring . They trained shallow, fully connected three-layer networks using backpropagation, optimizing a loss function that combined task performance with a regularization term penalizing connections between distant neurons. Through simulations on simple visual tasks with low-dimensional stimuli, they showed that enforcing short-range connectivity leads hidden and output layers to subdivide into specialized modules. This study is consistent with the idea that physical wiring constraints are a fundamental driver of modular organization in the brain, beyond purely computational or functional considerations, and inspired later end-to-end models that explicitly penalize long-range connections, e.g. Blauch et al. (2022) and X.-J. Zhang et al. (2025).

Recent advances in deep learning bring two major developments. First, end-to-end models (e.g. CNNs) can automatically learn hierarchical representations from low-level to high-level features, mimicking hierarchical processing in the brain. Second, modern deep learning models can process naturalistic stimuli, reducing reliance on handcrafted features by learning representations directly from data, opening up the possibility of modeling multi-unit responses beyond primary visual cortex. Recent models no longer replicate only properties of early visual cortex, but can also capture organization in higher visual areas, e.g. ventral temporal cortex (VTC). Inspired by earlier wiring-length constraints proposed by Jacobs and Jordan, Blauch et al. (2022) showed that imposing a similar wiring penalty in deep networks leads to the emergence of clustered neurons selective for faces, scenes, and objects in deeper layers. Z. Lu et al. (2025) optimized correlations among neighboring weights and, interestingly, captured not only V1- and VTC-like organization but also center–periphery behavioral biases in visual processing. Qian et al. (2026) introduced lateral connections between neighboring neurons, replicating characteristic features of both early and higher-level visual cortex without explicitly enforcing correlation constraints or minimizing wiring length.

An important work is Margalit et al. (2024), in which they optimized correlations between neighboring neurons. Their model can capture the organization of not only single areas but of the entire sequence of ventral visual cortex, from V1, including selectivity for orientation, spatial frequency, and chromatic gratings, to VTC category-selectivity maps, as well as the number, area, and spatial overlap of category-selective patches. Importantly, this work aside from developing a single model, proposed a framework that can be applied to replicate the topographic organization across the entire cortex, and showed good quantitative agreement with human and macaque brain data.

Going beyond vision, topographic modeling has also been applied to other modalities, such as language and audition (Rathi et al. 2024; Binhuraib et al. 2025; Al-Tahan et al. 2025). More recently, analyses on topographic models has extended beyond reproducing cortical maps toward more engineering-oriented goals. For example, Poli et al. (2023) showed that introducing topographic structure can improve robustness to pruning while maintaining classification performance. Deb et al. (2025) focused on making topographic constraints easy to plug into existing models to improve task performance. Qian et al. (2026), D. Zhou et al. (2025), and Bashivan et al. (2025) demonstrated that topography can increase robustness to adversarial noise. In addition, there is also effort to push topographic models toward applications in discovery and intervention. For instance, Kamila Maria Jozwik et al. (2023) provided a simulation demonstrating how such models can be used to discover new category-selective areas, while Mehrer, Lonnqvist, et al. (2025) used topographic models to guide cortical stimulation in macaques to generate visual prosthetics.

0.2.2 Thesis contributions

Study 2.1: Testing general topographic principles with a new organization dimension in high visual cortex

Since end-to-end topographic networks are relatively new models, an important question is how well they capture the organization of the ventral stream (occipitotemporal cortex; OTC). Recent models have shown that such networks can capture several major dimensions of ventral OTC, including animacy and real-world size, and can develop category-selective clusters. In particular, Margalit et al. (2024) proposed a set of organizing principles that, in theory, should be able to capture any kind of topographic organization within and outside visual cortex. Therefore, here we put this claim to test, and we adopt this model to try to replicate a specific type of spatial organization that goes beyond the dimensions traditionally emphasized in ventral OTC.

In this study, we propose action as a new dimension shaping object organi-

zation in OTC, and we use this dimension to test the model. In lateral OTC, we found that action dimension appears as a graded, topographically organized map, with partially overlapping selectivity for body parts and objects that differ in their action-related properties, which is a pattern that is not reducible to animacy, shape, or real-world size (Konkle and Oliva 2012; Konkle and Caramazza 2013; Bracci and H. O. d. Beeck 2016; P. Bao et al. 2020; Yue et al. 2020). We therefore tested whether the topographic models of Margalit et al. (2024) show a corresponding action-based organization. We found that this model did not exhibit an action-related gradient, despite capturing the animacy division, highlighting a limitation of current topographic modeling approaches, and suggesting that additional constraints beyond generic spatial organization may be required to account for the fine-grained topographic structure of high-level visual cortex.

Study 2.2: Comparing topographic constraint implementations and their computational consequences

Given the emerging landscape of end-to-end topographic models, we now have many different implementations of topographic constraints. Some models directly enforce similarity in unit activations across nearby neurons (Lee et al. 2020; Margalit et al. 2024; Rathi et al. 2024; Poli et al. 2023), while others impose similarity on connectivity (Z. Lu et al. 2025). These constraints have different mathematical consequences on the internal representations, which in turn shape the resulting topographic organization and map visualizations. At the same time, topographic constraints come with computational disadvantages, because they are not primarily designed to boost task performance: accuracy often stays similar, and can even be detrimental (Margalit et al. 2024; Z. Lu et al. 2025; Rathi et al. 2024). This limits the use case of topographic models if we only evaluate them by task accuracy or the quality of topographic maps. More broadly, beyond producing brain-like topography, the computational consequences of topographic constraints, and how different implementations shape learned representations, remain unclear.

The second paper in this thesis addresses this gap by systematically comparing two commonly used local topographic constraints in end-to-end trained convolutional networks: Activation Similarity (AS), which directly encourages neighboring units to have similar activations, and Weight Similarity (WS), which encourages neighboring units to develop similar afferent weight vectors so that correlated activity can emerge. We show that the correlations induced by topographic constraints can produce clear computational advantages in terms of robustness to noise, both input noise and parameter noise. WS generally produces clearer benefits than AS under parameter noise, while AS

is generally better under input noise. Moreover, the representations shaped by these two constraints are qualitatively distinct, affecting functional localization and feature tuning. Together, these results highlight that the choice of topographic constraint is not a minor technical detail: different ways of injecting topography impose different consequences for both robustness and representational organization, beyond producing visually smooth maps. With this effort, we aim to bring topographic modeling beyond computational neuroscience to a broader machine learning community, since topography is not just an inductive bias for modeling purposes only, but is also a useful bias to gain engineering advantages.

Topographic modeling is an old line of investigation that has been revived by the recent developments in deep learning. Modern end-to-end topographic networks now can learn from naturalistic stimuli and reproduce several forms of cortical-like organization. However, as a still-emerging model family, these models need to be tested and characterized more carefully, both in terms of what kinds of cortical topography they can (and cannot) capture, and what computational consequences their constraints introduce. Together, the studies presented in the second part of the thesis help characterize a family of models that has the potential to serve as a foundation for cognitive computational neuroscience.

Chapter 1

Pruning for Explaining Human Comparisons Using Alignment-Importance Heatmaps

Abstract We present a computational explainability approach for human comparison tasks, using Alignment Importance Score (AIS) heatmaps derived from deep-vision models. The AIS reflects a feature-map's unique contribution to the alignment between Deep Neural Network's (DNN) representational geometry and that of humans. We first validate the AIS by showing that prediction of out-of-sample human similarity judgments is improved when constructing representations using only higher-scoring AIS feature maps identified from a training set. We then compute image-specific heatmaps that visually indicate the areas that correspond to feature-maps with higher AIS scores. These maps provide an intuitive explanation of which image areas are more important when it is compared to other images in a cohort. We observe a correspondence between these heatmaps and saliency maps produced by a gaze-prediction model. However, in some cases, meaningful differences emerge, as the dimensions relevant for comparison are not necessarily the most visually salient. To conclude, Alignment Importance improves prediction of human similarity judgments from DNN embeddings, and provides interpretable insights into the relevant information in image space.

Code `https://github.com/tlmnhut/ais\_heatmap`

Publication status This chapter is published in the paper: **Truong, N.**, Pesenti, D., & Hasson, U. (2025). Explaining human comparisons using alignment-importance heatmaps. *Computational Brain & Behavior*, 1-21. We also include a preceding analysis that closely related to this paper, in which the author of the thesis contributed towards developing the analysis. The additional analysis is published in the paper: Tarigopula, P., Fairhall, S. L., Bavaresco, A., **Truong, N.**, & Hasson, U. (2023). Improved prediction of behavioral and neural similarity spaces using pruned DNNs. *Neural Networks*, 168, 89-104.

1.1 Introduction

1.1.1 The question: Explaining human comparisons

Work in recent years has shown that DNNs learn feature spaces whose geometry has some similarity to that of humans. This is convincingly shown by the fact that human similarity judgments (HSJs) for pairs of words or images are often quite well predicted by the distances between image-pairs or word-pairs in vision-DNNs or language models (for reviews, see Battleday et al. 2021; Roads and Love 2024; Sucholutsky et al. 2023). These models therefore naturally extract features relevant for modeling HSJs when trained on standard tasks such as image classification or word prediction. While the object-embeddings of such pretrained machine learning models approximate HSJs quite well, it has been further shown that these predictions can be considerably improved using down-stream operations.

One such operation is to learn a reweighting of the products of feature values, which improves prediction of HSJs for both images e.g., Peterson et al. 2018; Kaniuth and Martin N Hebart 2022a and words e.g., Richie and Bhatia 2021. Another approach is to use supervised pruning to assess features' importance in the context of estimating a set of similarity judgments (Tarigopula et al. 2023; Flechas Manrique et al. 2023b). Pruning does not alter the activation weights of the retained features, but instead removes a subset of features from the embedding matrix. Pruning has also been used to identify sub-spaces in language models that optimize particular classification tasks e.g., Cao et al. 2021.

While prior work has shown that pruning of nodes in a DNN's penultimate layer can improve prediction of similarity judgments, here we are interested in its potential to explain what parts of an image matter for the judgment itself. Understanding which information is used as a basis for comparison is a fundamental question in cognitive science. Since the work of Tversky (1977), many studies have shown that comparisons between objects are a function

of those elements that are shared or distinct between them. However, for naturalistic stimuli, it is difficult to know which properties are important when an image is compared to a target set of images. Here we suggest that this question is tractable via a computational solution in which latent dimensions that are related to the comparison process are identified and projected onto the image space as a heatmap.

1.1.2 Logic of the current study

We present the logic here, with a complete formal presentation provided in Section 1.2.1. Our approach relies on evaluating how pruning changes the alignment between human and computer-model representational spaces. Both spaces are operationalized using pairwise distances between images. One set of distances is derived from human behavior (HB_{dist}), the other is computed from a computer model ($Model_{dist}$). We define the baseline isomorphism between the two spaces as the correlation between these two vectors.

In the next step, a perturbation is introduced to the feature representations of an image. Specifically, a feature map in the last convolutional layer is masked. Therefore, the information from that feature map is not encoded in the model, and not propagated onwards to the fully connected layer from which we obtain image embeddings. Subsequently, $Model_{dist}$ is recomputed, as is the isomorphism between the representations. Note that only the target image is affected, and not the other images. Furthermore, HB_{dist} remains unchanged. There are two possible outcomes: *i*) if the encoded information from the feature map is cognitively irrelevant or even confounding, its removal could alter $Model_{dist}$ in a way that improves the isomorphism with human similarity judgments. Conversely, *ii*) if the encoded information from the feature map is cognitively-relevant (e.g., masking a feature map representing an animal's face in context of similarity judgments between animals), its removal will alter $Model_{dist}$ in a way that decreases the isomorphism with human judgments. This occurs because the way that images stand in relation to each other in the DNN representation is now lacking information that underlies human judgments. By iteratively masking all feature maps in the last convolutional layer, each feature map is linked with a perturbation score indicating its importance.

Similar logic was presented in the previous works, but masking was applied on the image space rather than the latent feature space. For instance, Palazzo et al. (2020) masked image patches to evaluate how masking impacted the compatibility between vision-DNN embeddings and EEG data. In other work, Tarigopula et al. (2023) used this approach with human neuroimaging data to explain which parts of an image contain information relevant to the representational space of various brain regions.

Since the author of the thesis contributed this analysis in Tarigopula et al. (2023), we also present it here in the results section as a preceding analysis that motivates the main study. These additional results (not appear in the current published paper of the chapter) serve as a proof of concept that pruning can produce interpretable heatmaps that explain the contribution of image parts to the alignment between model and human data; however, this analysis lacked quantitative evaluation. Below, we list the limitations of this approach, which motivate improving the method for the current study and opening up new utilities, which is explaining human comparisons.

1.1.3 Current aims and contribution

The current study's aims advances over prior studies in three respects: it directly studies human comparison processes, it introduces an advantageous masking procedure, and it evaluates the results against typical saliency maps. The aforementioned studies operationalized representational spaces from multivariate fMRI and EEG recordings but have not studied human comparison processes. Furthermore, the technique they use, namely, mask-sweep over an image, presents several major limitations: 1) the mask size is arbitrary, requiring the use of multiple sizes; 2) an arbitrary decision is required regarding how to combine information from different mask sizes; 3) the process is computationally costly, as masks are ideally applied with each pixel being in the mask center; 4) a theoretical weakness is that the mask is not informed by prior information contained in the model.

Departing from these prior studies, here we directly model human comparison judgments, and use a different, more efficient approach to masking images, which uses information already present in the DNNs own feature space. Specifically, we focus on the feature maps in a deep convolutional layer, and use them to define the masks. Our approach is inspired by Score-weighted Class Activation Maps (Score-CAM) which is an explanatory method that generates heatmaps indicating which sections of a target image are relevant for its classification (H. Wang et al. 2020). Score-CAM takes the information in each feature map, upscales it to the original input resolution, uses it as an information selector for the original input image, and computes the activation for correct class (pre-softmax confidence) when using that feature map alone. After repeating this process for all feature maps, the confidence scores are used as weights to generate a heatmap highlighting image areas important for classification. Using a similar logic, we show that information at the feature-map level is also highly useful for identifying which feature maps are important for the alignment between the DNN and human representational spaces, and that these can be visualized in a similar manner.

Beyond our main explainability objective, we have two other important aims. First, we evaluate whether it is possible to identify feature maps that are particularly important for predicting human representational spaces; using only these feature maps should improve out of sample prediction accuracy for human similarity judgments as compared to using all feature maps. Second, we evaluate the relationship between heatmaps produced using this method, and traditional saliency maps. While the latter operationalizes saliency using information latent in the image itself, the heatmaps we produce highlight information pertinent to image comparisons within a given set.

1.2 Methods

1.2.1 Preliminaries

• Architecture and datasets: In the main analysis, We use VGG-16, a deep neural network (Simonyan and Zisserman 2014), pre-trained on ImageNet (Deng, Dong, Socher, L.-J. Li, K. Li, et al. 2009) and another trained on Ecoset[1] (Mehrer, Spoerer, Jones, et al. 2021). VGG-16 was used because Ecoset was trained on that model. It is also a common architecture used for predicting human similarity judgements (Peterson et al. 2018; Kaniuth and Martin N Hebart 2022a) and has been shown to be a good candidate for behavior or brain alignment (Schrimpf, Kubilius, H. Hong, Najib J Majaj, et al. 2018a). As images we used a dataset provided by Peterson et al. (2018), which consists of 720 images divided into six categories of 120 images. The categories were: Animals, Fruits, Furniture, Various, Vegetables and Automobiles (the latter effectively including any means of transportation including horses, sleds, cranes; Transportation henceforth). Images had a native resolution of 500×500 which was downscaled to 224×224 to fit the model.
• Human Similarity Judgments: Let H be a matrix representing the similarity judgments provided by human assessors for n objects. Each entry $H_{i,j}$ in the matrix corresponds to the similarity judgment between objects i and j. We use the upper triangle of matrix H, denoted as H_u.
• Object distances in feature space: Let C be a matrix representing the embeddings of n images onto d features of the penultimate layer of the pre-trained computer vision model, denoted as $C \in \mathbb{R}^{n \times d}$. Specifically, we use VGG-16 with $d = 4096$, and the number of images in each Peterson's category is $n = 120$. Matrix C is obtained by considering all parameters of the pre-trained model, and specifically all 512 feature maps of the deepest convolutional layer.

[1] Available at https://osf.io/kzxfg/

Z_u is the upper triangle of image-pair similarity matrix Z, computed from the Spearman correlation for each row pair in C.

• Subspaces in matrix C: We produce two variants of C (all with dimension $n \times d$). The first variant ("remove 1"), denoted as $C^{(\neg k)}$, is constructed by excluding feature map k where $k \in \{1, 2, \ldots, 512\}$. The second variant is produced when using only a subset S of feature maps in the model. Let $S \subseteq \{1, 2, \ldots, 512\}$ be a set of selected feature-map indices, and let $C^{(S)}$ be the matrix representing the embedding of n images onto d nodes in the penultimate layer, but when using the subset of feature-maps corresponding to S. Note that in all cases, the (one or more) feature-map activations are propagated to the penultimate layer using the pre-trained weights.

• From the variants of C we derive matching similarity matrices. The first, $Z^{(\neg k)}$, is obtained by computing the cosine similarity for each pair of rows in $C^{(\neg k)}$. The second, $Z^{(S)}$ is formed using the selected feature indices in $C^{(S)}$.

• As indicated, Z_u and H_u denote the vectorized upper triangles of matrices Z and H respectively. The Spearman correlation coefficient between the two is denoted as $\rho(Z_u, H_u)$. We refer to this value as a Baseline Second-Order-Isomorphism (2OI) between the two domains. Analogously, in some cases we compute $\rho(Z_u^{(\neg k)}, H_u)$ and $\rho(Z_u^{(S)}, H_u)$.

1.2.2 Aim 1: Identifying a subset of feature maps that optimizes prediction of human similarity judgments

We define the Alignment Importance Score (AIS) of each feature map in terms of its predictive capacity for the human representation H_u. Intuitively, we aim to determine how the removal of each feature map $k \in \{1, 2, \ldots, 512\}$ affects the baseline isomorphism, $\rho(Z_u, H_u)$. The removal of each feature map produces a modified 2OI score, $\rho(Z_u^{(\neg k)}, H_u)$. Finally, The AIS of feature map k is defined in Equation 1.1, with positive values indicating a relatively important feature map, and negative values a less important one. After computing AIS for all feature-maps, we rank-order them based on their AIS.

$$\text{AIS}_k = \rho(Z_u, H_u) - \rho(Z_u^{(\neg k)}, H_u) \tag{1.1}$$

We then identify an optimal subset of feature maps for predicting H_u. In each iteration, one feature map is added to the subset S in descending order of AIS rank, and we recompute the 2OI, $\rho(Z_u^{(S)}, H_u)$ using that subset of feature maps alone. After these 512 iterations, subset S^* ultimately selected is the one that maximizes 2OI.

To validate AIS, we use an 80:20 cross-validation framework where 80% of the entries in H_u are assigned to a training set, and the remaining 20%

constitute the test set. The optimal subset of feature map indices, S^*, is determined from the training set using sequential features selection as described above. For statistical significance testing, we repeat the entire cross-validation process eight times with different dataset shuffling. This produces 40 Full vs. Retained value-pairs for each relevant comparison. To evaluate generalization, we use only this S^* set of feature maps to predict HSJs on the test set. Prediction performance is compared against a baseline where all 512 features are used for predicting HSJs in the test set. Statistical significance testing, per dataset, is based on the 40 value-pairs produced via cross-validation, which are analyzed using paired two-tailed T-tests (12 tests in all, non-corrected for multiple comparisons). Success of Aim 1 is determined if $\rho(Z_u^{(S^*)}, H_u)$ surpasses $\rho(Z_u, H_u)$, indicating superior prediction compared to the baseline using a subset of feature maps.

As an additional baseline, we used Learned Perceptual Image Patch Similarity (LPIPS), which is a method for obtaining a cognitively-relevant similarity metric between image pairs (R. Zhang et al. 2018). LPIPS fine-tunes a computer vision CNN so that the image distances in the network, calculated as differences between embedding vectors, align with human similarity judgments. LPIPS is fine-tuned using human decision data regarding which of two slightly altered images are closer to an original image, and is based on reweighting all layers of the network. LPIPS has shown to closely match human behavior in 2-Alternative Forced Choice tasks involving minor image distortions and a reference image. To evaluate whether LPIPS is at all viable for our materials and similarity judgments, we applied LPIPS to all images in each dataset to compute pairwise distances between images, and computed the Pearson correlation between the LPIPS distance matrix and the human similarity judgments. Note that the LPIPS method does not allow integration with pruning, as its reweighting function achieves a parallel goal. We use the pre-trained LPIPS weights provided by the original authors as these have been trained on a large set of human judgments and have been argued to predict human behavior in multiple domains.

1.2.3 Aim 2: Explaining human similarity judgments

Our goal is to identify which image patches, in image space, are relevant to comparisons between a target image t and other images in the set. This is visualized by creating a heatmap for t identifying those image sections, as follows. We begin by defining a baseline 2OI for t as the Spearman correlation between the $n-1$ similarity judgments associated with t, as quantified from the model, and the corresponding set of human similarity judgments. As in

Aim 1, we define the AIS of feature map k by computing a value that reflects the departure from baseline, as indicated in Equation 1.1.

We iterate over all 512 feature maps, producing 512 AIS values that indicate the relative importance of each feature map for the alignment between DNN-derived distances and human similarity judgments for target image t. This produces an $n \times k$ matrix (120 [AIS] x 512 [feature map]) for each dataset containing 120 images. We then compare these distributions between the ImageNet and Ecoset-trained models to understand if and how the training regime impacts the distribution of AIS. Histograms are computed for the mean AIS value by feature, and the Mean Absolute Deviation, computed by feature (column) and by image (row).

Image-level heatmaps are then computed as follows. We first convert negative AIS values to zero because they indicate features that encode information less relevant to modeling the human data (see Eq. 1.1). The remaining scores are sum normalized. Subsequently, feature maps for an image are weighted-averaged according to their corresponding AIS to create a heatmap. In the heatmaps, warmer colors indicate image areas associated with the more important features.

To quantify the similarity between the heatmaps generated by Ecoset and Imagenet, we defined a Match score for each image as the Pearson correlation between the heatmap generated by the Ecoset model and the one generated by the Imagenet model. Anticipating the results, in certain instances, the Match score was low. We therefore examined if this occurred for images that did not correspond to classes on which the models were trained. For each image, we computed the entropy of the post-softmax probability distributions, independently for the Ecoset and ImageNet trained models. The higher of these two entropy values was retained and designated as maxEntropy. Subsequently, considering all images in a dataset, we computed the correlation between the Match score and maxEntropy.

1.2.4 Aim 3: Cross-referencing heatmaps against saliency maps

We compare the heatmaps produced by our method to those produced by TranSalNet (J. Lou et al. 2022), which is a state-of-the-art DNN that identifies salient image sections and accurately predicts human gaze patterns (see Figure 14 in Appendix). We cross-reference TranSalNet against our method (AIS) using two approaches: Precision-Recall curves and Subset analyses.

Precision-Recall curves

First, we evaluate how well a pixel's salience predicts its inclusion in an AIS heatmap. When the salience and AIS maps are thresholded at a specified level to form binarized maps, the relationship between them can be understood in terms of precision and recall. The binarized AIS map is treated as the target variable, and the binarized saliency map is the predicting variable. In this case, we have:

$$\text{Precision} = \frac{|TranSalNet \cap AIS|}{|TranSalNet|}$$

and

$$\text{Recall} = \frac{|TranSalNet \cap AIS|}{|AIS|}.$$

We describe this relationship using a Precision-Recall curve. The curve is generated by thresholding the AIS map at a fixed level and then plotting precision versus recall as the saliency map is thresholded across a range of levels.

The following steps were performed for each image: first, we created a heatmap as described in Aim 2 and generated a corresponding saliency map using TranSalNet. We kept the same aspect ratio of the images input to both VGG-16 and TranSalNet for compatibility in later comparisons. We conducted four separate analyses, where we created a binary mask for the AIS map at each of the following percentiles: $P = \{60, 70, 80, 90\}$. In each analysis we thresholded the saliency maps at all percentiles between 1 and 99, with a step size of 2. Percentiles were calculated separately for each image.

Conditional probability analysis

In this analysis we aim to identify whether an image section (specifically, a pixel) identified as salient (Sal) is more likely to also be identified as comparison-relevant (CR; that is, warm-colored in our analysis). To do this we threshold both maps to select the top 5% of Salient and CR pixels, producing Sal, $\neg Sal$, CR and $\neg CR$ partitions of the image pixels. We then compute the Relative Risk (RR) ratio as in Equation 1.2.

$$\text{RR} = P(CR|Sal)\boldsymbol{\nabla} \cdot P(CR|\neg Sal) \tag{1.2}$$

The relative risk as computed here measures the likelihood of Sal pixels being CR pixels compared to $\neg Sal$ pixels. An RR value greater than 1 indicates that salient pixels are more likely to be CR than non-salient ones, while an RR less than 1 indicates the opposite. A main difference between this analysis and the precision-recall one is that it also quantifies joint distributions within the

non-salient pixel-set. We repeat this analyses when thresholding both maps at 10% and 15% top *Sal* and *CR* pixels.

We note that there is no requirement that the two methods identify the same image features. The saliency map is driven by image features (including higher level semantics captured by the DNNs), whereas the heatmap we produce from AIS values is a function of how a certain object stands in relation to other objects in the set. As we will see, this produces cases of very high overlap, but also important distinctions.

1.2.5 Aim 4: Generalization to other architectures and training objectives

In Aims 1, 2 and 3 the image embeddings used were obtained from VGG-16. VGG-16, and a later variant VGG-19, are somewhat unique in that after the deepest convolutional layer, they also include two very large fully connected layers. These layers perform non-linear, abstract interactions over the information in the deepest feature map layer, and are essential for linking this information to the classification task.

Many other computer-vision architectures do not include such layers, and instead use the deepest feature maps, relatively directly, for classification. This is done by implementing global average pooling, which reduces each of these feature maps into a single value, followed by learning a linear combination of these values for classification. Thus, in these architectures, the final layer before classification receives an input corresponding to the number of feature maps (after global pooling), and produces an output corresponding to the number of classes to be learned.

To evaluate the applicability of the AIS-based analysis to other architectures, we applied the analysis developed for Aim 1, with several modifications, to the following models: Inception-V3 (Szegedy et al. 2015), ResNet-152 (K. He et al. 2016), DenseNet-161 (G. Huang et al. 2016), EfficientNet-B3 (Tan and Le 2019), RegNetY-400MF (Radosavovic et al. 2020), and ResNeXt-50-32x4d (Xie et al. 2017). The deepest layers of these architectures contain varying numbers of feature maps: Inception-V3, ResNet-152, and ResNeXt-50-32x4d each have 2,048 feature maps, DenseNet-161 has 2,208, EfficientNet-B3 has 1,536, and RegNetY-400MF has 440.

We note that all these architectures learn features in the context of supervised classification tasks. To evaluate feature maps produced by non-supervised learning, we used a ResNet-50 architecture trained with the Barlow Twins self-supervised learning framework (Zbontar et al. 2021). In this approach, the objective of the the model is learn representations by maximizing

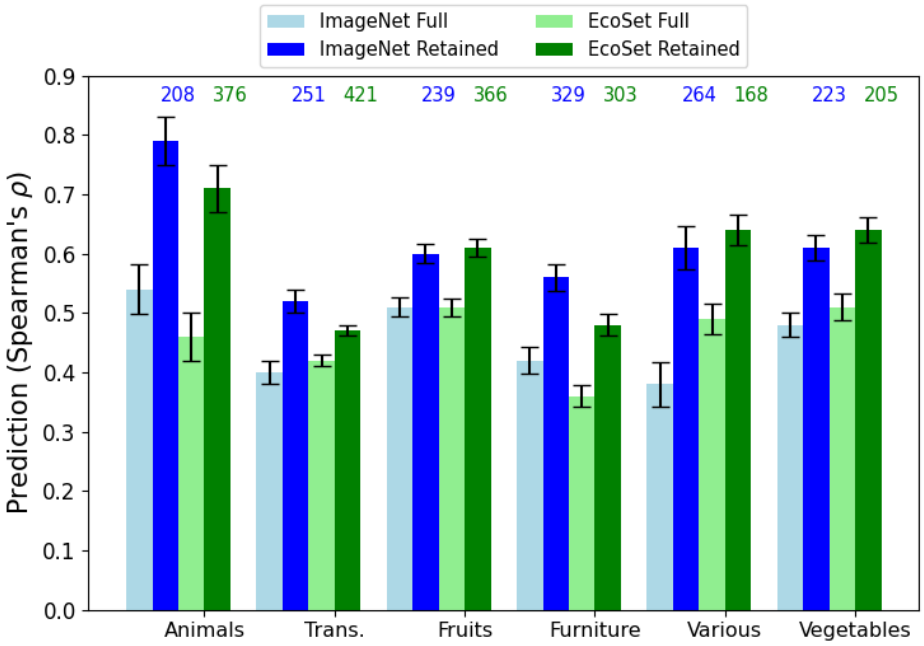


Figure 1.1: Out-of-sample predictions of human similarity judgments using image embeddings. Full: using all 512 feature maps. Retained: using feature maps identified from an independent training set. The numbers above the second and fourth columns in each group represent averages of feature-map set sizes across 40 folds. Error bars indicate standard errors adjusted for paired-comparisons (Loftus and Masson 1994).

the similarity between two augmented versions of the same image. In this way, training extracts general visual features, ignoring small visual distortions.

For each of these architectures we performed five-fold Cross validation, as detailed for Aim1. For all architectures except VGG-16 and VGG-19, object embeddings were generated by applying global pooling to the feature maps from the deepest convolutional layer. For VGG-16 and VGG-19, embeddings were constructed from the penultimate, fully connected layer.

1.3 Results

1.3.1 Aim 1: Identifying a subset of feature maps that optimizes prediction of human similarity judgments

As shown in Figure 1.1, by computing AIS it was possible to identify a subset of 512 feature maps for each dataset, which produced improved out-of-sample predictions compared to a baseline condition where all feature maps were used. This was consistent for models trained on Ecoset or ImageNet, with less than 50% of the 512 feature maps being used in 5/12 cases. Paired T-tests indicated that in all 12 cases, predictions from Full features were less accurate than those from features learned via pruning (p-values < 0.01). The performance metrics of ImageNet and Ecoset were quite similar.

Speaking to category-specific information, AIS values for each feature-map differed across datasets. That is, feature maps important for aligning one category were not necessarily important for another category. To evaluate this issue, we computed pair-wise Pearson correlations between the AIS values of the 512 feature-maps for each pair of datasets (e.g., Fruits vs. Vegetables). For both Ecoset and ImageNet, the strongest correlation was between Fruits and Vegetables (Ecoset $R = 0.48$; ImageNet $R = 0.67$). For Ecoset, the second highest correlation was between Transportation and Furniture ($R = 0.38$), whereas for ImageNet it was between Various and Animals ($R = 0.26$). Most of other correlations, in both analyses, ranged from -0.2 to 0.2.

Finally, we evaluated the LPIPS method for human similarity modeling (see *Methods*). LPIPS image-distances indeed tracked human similarity judgments for all categories, in that higher LPIPS distances were associated with lower similarity. However, these correlations were quite low. Spearman rho values were: Animals 0.15, Automobiles 0.19, Fruits 0.15, Furniture 0.07, Vegetables 0.40, and Various 0.19. Thus, alignment with LPIPS did not approach the levels seen in Figure 1.1, even for the non-pruned cases.

1.3.2 Aim 2: Explaining human similarity judgments

Figure 1.2 shows examples of heatmaps produced by alignment importance scoring. Given that each dataset contained 120 images, we selected 4 images from each dataset according to the principle that two of the images produced apparently sensible results, and the two others were less sensible. It can be seen that the method can identify image-sections that are relevant for inter-category comparisons, such as the faces of animals, central parts of fruits and vegetables, and discriminating elements of artifacts and man made objects. As we will see later, these are not necessarily the most salient aspects of images.

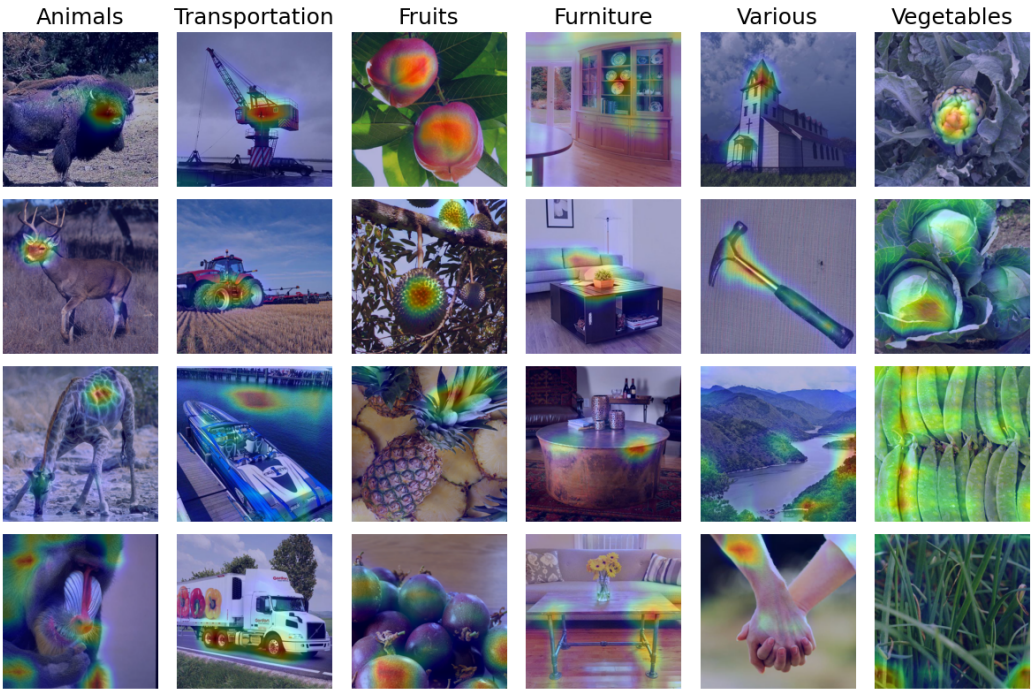


Figure 1.2: Heatmaps generated using Alignment Importance Scores of feature maps trained with Ecoset. For each dataset, two images with subjectively higher interpretability (top two rows) and lower interpretability (bottom two rows) were selected.

To assess the similarity of heatmaps produced by Ecoset and ImageNet, for each image we calculated the correlation between the heatmaps produced by the two methods. The median correlation values were as follows: 0.80 ± 0.16 for Animals, 0.64 ± 0.19 for Transportation, 0.73 ± 0.22 for Fruits, 0.64 ± 0.22 for Furniture, 0.64 ± 0.27 for Various, and 0.56 ± 0.25 for Vegetables. In all datasets the maximum correlation values approached 1.0, while the minimum values often approached zero (see the histogram in Appendix Figure 13). As Appendix Figure 13 shows, for all categories (apart from Animals), around 10% of images showed a low correlation of less than 0.2. Considering a correlation of 0.8 as an (arbitrary) reference point for strong correspondence between heatmaps, we find that for Animals more than 40% of the images showed correlations that exceeded this value, whereas for Transportation and Furniture the value was below 20%.

This means that although agreement was often good, training models on Ecoset or ImageNet often produces different heatmaps. These findings are consistent with those of Aim 1, which showed that the VGG-16 models trained

on the two datasets capture and learn human similarity judgments in slightly different ways.

As detailed section 1.2.3, we evaluated if images that presented a lower Match between Ecoset and ImageNet heatmaps were associated with higher entropy of post-softmax values in either of the two sets (maxEntropy), which would produce a negative correlation between the two quantities. We found that this was indeed the case, for Animals ($R = -0.31$), Fruits ($R = -0.34$), Various ($R = -0.24$), and Vegetables ($R = -0.21$). Weaker, yet sill negative correlations were found for Transportation and -0.11, Furniture, $Rs = -0.11, -0.04$ respectively. Thus, images that do not present information sufficient for classification produce disagreement between the two models. These might be out of distribution images or bad examples of trained categories.

Ultimately, in those cases where heatmaps differ, the results of Aim 1 may be used as a guide to inform whether Ecoset or ImageNet is more plausible with respect to the human representation of a given category. For instance, given the low agreement in heatmaps produced for Transportation and Furniture, one may select to use the ImageNet produced feature maps as these provide better out-of-sample prediction of human behavior.

We also statistically quantified the relation between AIS values obtained for feature maps when produced from models trained on Ecoset or ImageNet. Figure 1.3 shows, for each dataset, histograms computing the Average AIS associated with each feature (log10 scaled); and Figure 1.4, the Mean Absolute Deviation computed per feature (column) and per image (row). The histogram shows that the average AIS rarely exceeded 0.001 for any feature (Figure 1.3). Two-sided Kolmogorov-Smirnov (KS) tests (Hodges Jr 1958) were conducted to verify if the histograms associated with the two training regimes (ImageNet, Ecoset) came from the same distribution. Overall, KS test confirmed significant differences for all six categories ($p < .05$).

With respect to Mean Absolute Deviation (MAD), when computed per feature (Figure 1.4a) we find that the values varied around one order of magnitude, with a few features showing relatively higher values meaning they were much more important for some images than others. The MAD histograms computed from per-image data indicated that ImageNet's AIS distribution was consistently left shifted with respect to Ecoset's (Figure 1.4b). This means that the AIS produced by Ecoset-trained model are more strongly distributed, suggesting a more meaningful separation between those features relevant for alignment and those that are not. Two-sided Kolmogorov-Smirnov tests on Mean Absolute Deviation verify significant differences between the two models in all cases (all six datasets, KS tests, $p < .05$).

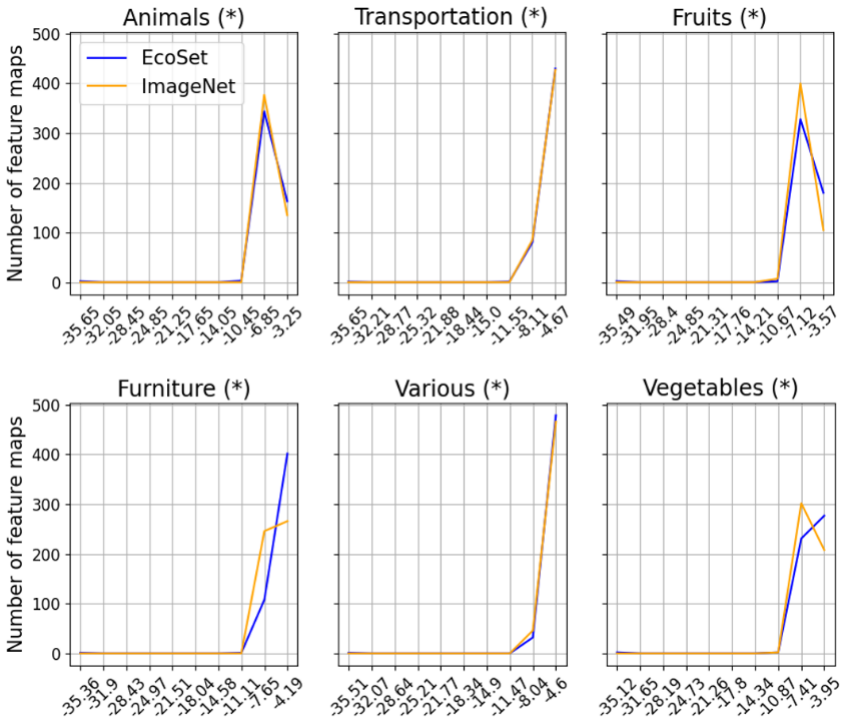


Figure 1.3: Log average of AIS values per feature for models trained on Ecoset or ImageNet. A star symbol (*) indicates a significant difference between the two distributions as determined by a KS test.

1.3.3 Aim 3: Cross-referencing heatmaps against saliency maps

Precision-Recall curves

For each image, we thresholded the AIS-produced heatmap at a given threshold to form a binary prediction target with AIS-related image sections (after thresholding) constituting the positive class. We then evaluated the extent to which these could be predicted by the saliency maps, using a Precision-Recall curve. In this analysis, the target variable is thresholded at a fixed level (e.g., 90th percentile), while the predicting variable is thresholded across a range of levels, with precision and recall computed for each threshold.

Figure 1.5a shows the results when predicting AIS heatmaps produced from ImageNet-trained feature maps, and with the AIS heatmaps thresholded at the 60th percentile. We observe that when saliency maps are thresholded at stringent levels (leftward points on the curve), precision is high for the Animals category, and somewhat lower for other categories, with values ranging from

0.6 to 0.8.

Lowering the threshold increased recall, but also gradually lowered precision as expected, While saliency and AIS maps were clearly related, with the exception of Animals, predicting AIS from saliency appeared limited, even when AIS heatmaps were thresholded at a relatively low value of 60th percentile. The other panels in Figure 1.5 show the same analysis with AIS heatmaps thresholded at the 70th, 80th, and 90th percentiles. In the latter analysis, AIS-relevant pixels are defined as the top 10%, and as shown in Figure 1.5d), saliency predicted membership in this class poorly, with the exception of the Animals category. Another observation is that for the Fruits category, thresholding the saliency map at the most strict level (left-most point) did not produce the highest precision, which was instead achieved at lower thresholds. This suggests that the most salient points were not always the most precise predictors of AIS heatmaps..

In summary, we found that, with the exception of the Animals category, saliency heatmaps could not predict AIS heatmaps with good precision and recall, particularly when AIS heatmaps were thresholded at higher levels. A very similar pattern was found for AIS heatmaps produced from Ecoset feature maps (see Appendix Figure 15).

Conditional probability analysis

We observed that areas identified as comparison-relevant by AIS heatmaps were much more likely to be associated with salient image sections than with non-salient image sections, as indicated by Relative Risk values strongly exceeding 1.0 (Table 1.1). This was found regardless of whether pixels in both heatmaps were thresholded at top 5%, top 10% or top 15%. As the table shows, the RR values often exceeded 5, reaching as high as 30 for Animals. The data were quite similar for ImageNet and Ecoset overall. Furthermore, the Relative Risk values varied significantly across categories, being highest for Animals, and lowest for Vegetables. This suggests that for Animals, elements salient in images are also important for comparison, whereas this is less so for Vegetables. This is numerically consistent with the Precision-Recall analysis where we found that thresholding saliency maps at high percentiles produced good prediction-precision of AIS data.

Figure 1.6 presents images on which we plotted contours reflecting TranSalNet's salience (orange) and alignment score heatmaps (blue) to visualize their overlap. For the two images on the left (bison and crane), the salience and alignment maps consistently show strong agreement across all three thresholding levels. For the two right images, there is no overlap. Specifically, the monkey's facial features are highly salient, but are not identified as important

Table 1.1: Relative Risk values comparing heatmaps computed from Alignment Importance Scores to those generated by TranSalNet, a saliency model that predicts human gaze. Chance values are $RR = 1$.

Category	Ecoset			ImageNet		
	5% vs. 5%	10% vs. 10%	15% vs. 15%	5% vs. 5%	10% vs. 10%	15% vs. 15%
Animals	30.8 ± 32.1	17.0 ± 18.3	12.7 ± 11.0	28.2 ± 34.2	14.9 ± 11.5	11.4 ± 8.2
Transportation	7.8 ± 11.5	5.8 ± 7.0	5.2 ± 6.5	6.4 ± 7.8	5.6 ± 5.8	5.3 ± 5.2
Fruits	9.9 ± 18.5	7.4 ± 10.9	6.2 ± 9.2	9.9 ± 21.4	6.6 ± 11.4	5.4 ± 8.2
Furniture	6.1 ± 10.3	5.1 ± 6.2	4.5 ± 4.8	6.5 ± 12.0	5.2 ± 6.5	4.6 ± 4.5
Various	17.3 ± 27.4	10.2 ± 11.0	8.7 ± 8.9	14.4 ± 31.4	8.2 ± 9.9	6.7 ± 7.2
Vegetables	6.4 ± 10.7	4.9 ± 6.8	4.1 ± 4.2	7.1 ± 14.7	5.0 ± 6.8	4.1 ± 4.2
All datasets	13.0 ± 22.1	8.4 ± 11.7	6.9 ± 8.4	12.1 ± 23.8	7.6 ± 9.6	6.2 ± 6.9

for alignment. In the case of the truck image, the banner area depicting colorful peppers is identified as salient, but the wheel area is identified as important for alignment. This is reasonable, as means of transportation in the set are effectively compared by observing the lower section of the vehicle, which differentiates trucks, cars, buses, motorcycles, trains and so on. Indeed we find these elements are often highly salient in the produced heatmaps. More results with appropriate level of detail are shown in the Appendix section below.

1.3.4 Aim 4: Generalization to other architectures and training objectives

We find that quantifying alignment importance improved out-of-sample prediction of human similarity judgments across all architectures and all six categories tested (see Figure 1.7).

Based on the results, we make the following observations. First, the two VGG-based architectures tended to perform the best overall, ranking first in three of the six image categories and second in all six categories. Second, baseline performance (test-set prediction using all features) tended to be diagnostic of which architecture would perform best using the learned pruned test set: for four of the six categories, the best performing model when using the full feature sets was also the best-performing when using the pruned sets.

However, in three cases, none-VGG models predicted human judgments best. EfficientNet-B3 ranked highest for Fruits and Furniture. The compound scaling used in this architecture, which optimally balances width, depth and resolution has been argued to produce a better representation of relevant image details (see Tan and Le 2019 their Figure 7). Furthermore, as indicated in the Methods section, the fact that this model uses linear combinations of feature-map information for classification (after global pooling) makes it potentially

more interpretable than VGG-16 and VGG-19, which use fully connected layers to learn complex combinations of feature-map information. Finally, the Barlow Twins architecture which is self-supervised and is not guided by a classification objective performed the best on the Various category.

These findings suggest that the VGG architectures show considerable strength overall. However, the impact of removing single feature maps in these architectures is effectively evaluated via the changes in activations in the fully connected layers, which learn interactions between feature maps. Depending on the aims of the analysis, other architectures may be used if such interaction effects are of no interest. Practically, the findings of Aim 4 suggest that when using AIS-based heatmaps as explanations for human comparisons, it is is sensible to use an architecture that best predicts these judgments.

1.3.5 Heatmaps in Tarigopula et al. (2023) that show the impact of brain-supervised pruning on representational space

Here we present the preceding analysis in Tarigopula et al. (2023) that motivates our current study. A similar pruning logic was proposed there, aiming to compute the importance of each part of an image for the alignment between models and human data. An image part is important if removing that part decreases alignment; otherwise alignment remains the same or increases. The main differences are: (1) we computed RDMs from regions of interests (ROIs) extracted from fMRI data, where participants observed a set of natural object images from different categories (King et al. 2019), unlike the homogeneous category setting in Peterson et al. (2018). (2) We removed image parts using a predefined window size and swept the window across the image, similar to convolutional filters, and we varied the window size and then aggregated results across sizes. (3) There was no step of inserting the high-scored kept parts to maximize alignment scores, as in AIS. Below we present the methods and results of this analysis. Since this analysis served as a preliminary proof of concept, there was no evaluation.

Because pruning fleshes out shared dimensions between a brain ROI and a pruned DNN, it is possible to identify, for a given image, the contribution of each image section to those shared dimensions. The principle is based on evaluating the impact of masking a part of a single image on the 2OI between the DNN and Brain RDMs. In brief (see *Appendix* Figure 16 for methods details), we consider as input a set of N images presented for viewing in an fMRI scanner. One image is selected for analysis and for this target image we compute an RDM capturing the correlation between the target image and each

other image. Correlations not involving the target image are not considered. One RDM is computed from Brain data and another from DNN embeddings. The second-order-isomorphism value for these two RDMs is taken as *baseline* 2OI, $2OI_{base}$. A portion of the target image is then masked, and the DNN RDM is recomputed, whereas the brain RDM remains unaltered. This produces a modified second-order isomorphism $2OI_{mask}$.

If the masked area changes the DNN RDM in a way that reduces its 2OI with the brain RDM, i.e., $2OI_{mask} < 2OI_{base}$, this means that the masked area contains information that loads on a latent dimension that contributes to 2OI. In contrast, if masking does not reduce $2OI_{base}$, or even improves on it, the information within it is less relevant to shared dimensions. The contribution of an image patch is therefore simply $Contrib = 2OI_{base} - 2OI_{mask}$ with higher values indicating greater importance of the masked area.

To make this concrete, consider a set of ten images where images 1-5 include a face and images 6-10 do not. Assume that a certain brain area only codes for the presence of a face. This brain area's RDM will separately cluster images 1-5 and images 6-10. Assume also that a DNN has been pruned by this brain area, and therefore produces a similar RDM. The relation between the two RDMs is quantified via $2OI_{base}$. Image 1 is chosen as the target image, and the face depicted in that image is masked. The DNN RDM for correlations with Image 1 changes: now, images [2-5] are strongly clustered but image 1 clusters with images [6-10]. Because the brain RDM remains unaltered. the result is a reduction in $2OI_{base}$ because the masked region was related to a dimension that organized both RDMs. Contrarily, if a non-important part of Image 1 were masked, the DNN RDM would not change, and so $2OI_{base}$ would remain unaltered. Implementation details can be found in the Appendix.

To apply this method we used brain RDMs from ventral temporal cortex (vTC) and parahippocampal place area (PPA), and the two DNNs pruned by these RDMs. For any given target image, the image was masked by sweeping a mask over the entire image, and assigning a Contribution score to a 4×4 pixel area in the center of the mask. Following prior work Palazzo et al. 2020 masks at different scales were applied, and we selected the *Contrib* value that departed most strongly from zero as the value assigned to the center of the mask. As an internal control, the analysis was also repeated by computing DNN RDMs from a non-pruned version of VGG-19. This control identifies shared dimensions between the 'vanilla' non-pruned network and a given brain RDM. The code for this can be found on Github `https://github.com/tlmnhut/Visualize\_PrunedDNN\_by\_HumanSim`.

A sample result is shown in Figure 1.8 (see *Appendix* Figure 17 for more results). As shown, the method is highly useful for identifying types of infor-

mation that may be important for a given brain area. In the outdoors image, for vTC, masking of sky-areas strongly perturbed $2OI_{base}$, but this was found for the pruned DNN only. For PPA, in contrast, the unpruned DNN identified the face as important, but the pruned DNN notably excluded face information. We note that these effects were mediated by the global rather than local structure of the image: applying the method to target images rotated by 180-degrees (e.g., sky is below) produced substantially different heatmaps.

Given these results as a proof of concept that supervised pruning by brain data can produce interpretable heatmaps highlighting image parts that are important for human–model alignment, we improve this method in the current paper to make it both a predictive and explanatory tool for studying human object comparisons, with concrete quantitative evaluation.

1.4 Discussion

Understanding what information is used in human comparisons is important not only for a better understanding of the comparison process itself, but also for comprehending how people form memories and make decisions (Roads and Love 2024). We introduced and validated a feature-map's Alignment Importance as a meaningful parameter relevant to such explanations. We first showed that AIS values generalize to improve prediction of human similarity judgments. This complements current approaches that achieve improvements by using reweighting or pruning of nodes in a DNN's penultimate layer (e.g. Peterson et al. 2018; Attarian et al. 2020; Kaniuth and Martin N Hebart 2022a; Jha et al. 2023; Tarigopula et al. 2023).

We then used AIS to produce explanations for those judgments via heatmaps. These heatmaps offered some correspondence to state-of-the-art saliency maps, in that when saliency maps were thresholded at high percentiles, the resulting representation could sometimes predict (binarized) AIS heatmaps quite well, especially for Animals. However, instances where saliency and AIS-reduced maps diverged are of major theoretical importance as they show it is possible to dissociate visually salient image elements from those that are important for comparison.

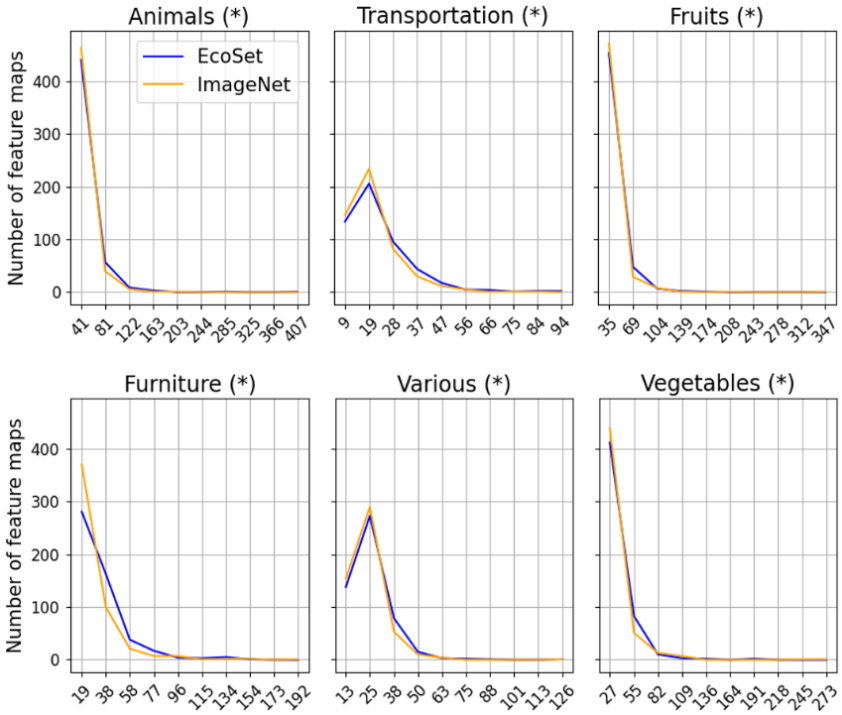
Because the method we present is based on mapping, or aligning a DNN's representational space to a human one via pruning, the feature space of the pretrained-DNN is of fundamental importance. For this reason, in Aim 1 we studied DNNs trained on both ImageNet and Ecoset datasets. We found that AIS scores improved out-of-sample prediction for models trained on either of the training datasets. Thus, both models learn feature maps particularly relevant for accounting for the representational space of specific categories. For

both Ecoset and ImageNet, category-specificity was shown in the fact that the relative ranking of AIS scores varied greatly across categories. Interestingly, Ecoset appears to distribute the AIS scores slightly more uniformly across feature-maps than ImageNet, which is a topic that requires further investigation.

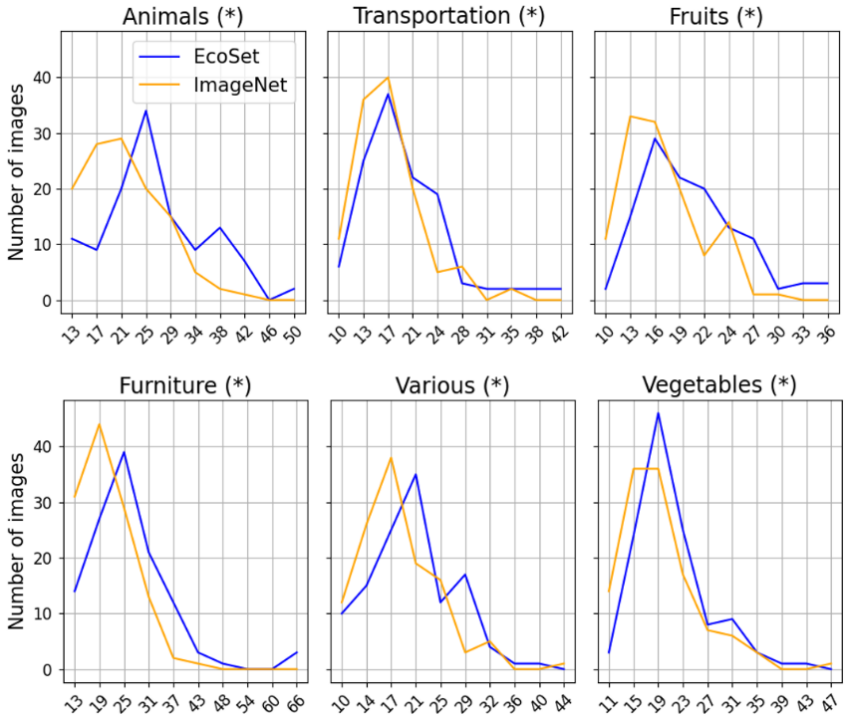
Further speaking to generalization across both training sets, the heatmaps were, for the most part, quite similar when created from Ecoset or ImageNet AIS scores, with average correlations between the heatmaps exceeding 0.75 for the Animals category. However, some images showed low correlations, and these tended to be associated with more uniform post-softmax distributions in the DNN's categorization layer. This means that divergence in heatmaps produced by the two models were more prevalent for images that one of the models found difficult to classify. In practice, we recommend using both Ecoset and ImageNet trained models to create heatmaps and carefully evaluating images with inconsistent results.

The strongest demonstration of generalization of the AIS based approach was provided in Aim 4, where we showed that the method improves out-of-sample prediction of human similarity judgments across eight different architectures. From the perspective of construct validity, the choice of architecture is fundamental for the effective use of the proposed method. An architecture that provides poor out-of-sample predictions of human similarity judgments will offer less meaningful explanations of human behavior compared to one that provides strong predictions. Examining this issue we find that there was no architecture that provided the best prediction across all six image categories. Thus, when explaining human comparisons for a stimulus set, it would be generally important to select an architecture with the best predictive capacity.

However, we also note that predictive capacity should be considered conjointly with the complexity of the architecture. In the current study, we used the VGG architecture in Aims 1, 2 and 3, as it was the reference architecture in prior work on prediction of human similarity judgments from image embeddings (Attarian et al. 2020; Peterson et al. 2018; Tarigopula et al. 2023). As mentioned in the Methods and Results, the two VGG architectures, while providing good predictions, produce embeddings that naturally reflect interactions between feature map information, and so the removal of a feature map is assessed by the impact of its removal on these interaction values. Other architectures that do not use fully connected layer after the deepest convolutions may produce simpler explanations. This is a topic that needs to be explored in future work.

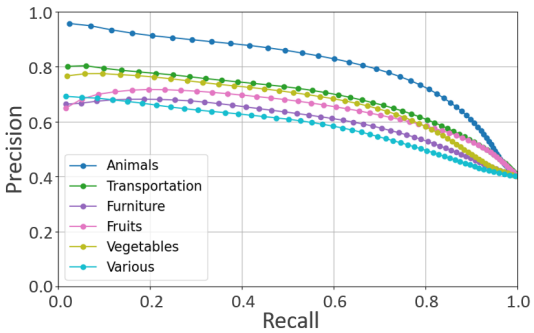


(a) Mean absolute deviation of each feature's AIS values, computed over 120 images.

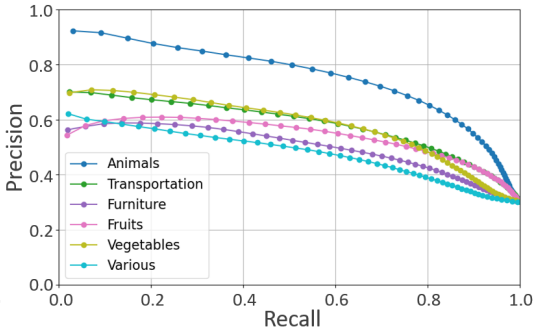


(b) Mean absolute deviation of each image's AIS values, computed over 512 feature maps.

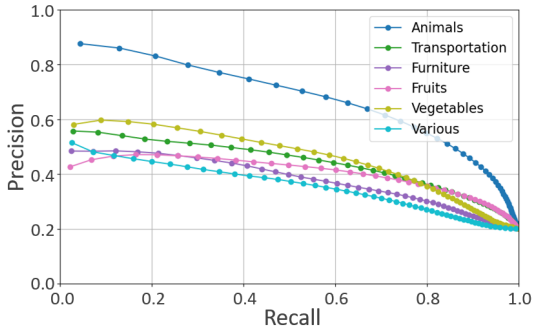
Figure 1.4: Histograms describing statistics of Alignment Importance Score distributions for models trained on Ecoset or ImageNet. The x-axis are displayed in e-4 format. A star symbol (*) indicates a significant difference between the two distributions as determined by a KS test.



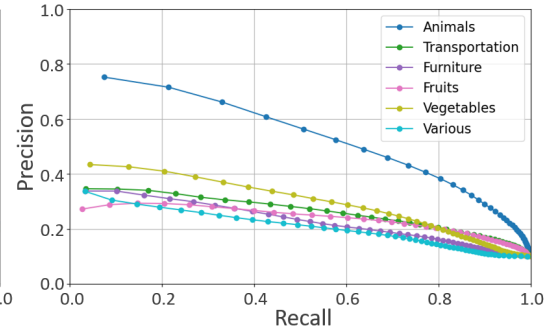
(a) AIS heatmaps thresholded at 60th percentile



(b) AIS heatmaps thresholded at 70th percentile



(c) AIS heatmaps thresholded at 80th percentile



(d) AIS heatmaps thresholded at 90th percentile

Figure 1.5: Precision-Recall Curves when predicting AIS heatmap values from saliency, for different thresholds of AIS heatmaps. The target variable was heatmap values produced from AIS scores computed from ImageNet training. The predicting variable were saliency map values obtained from TranSalNet.

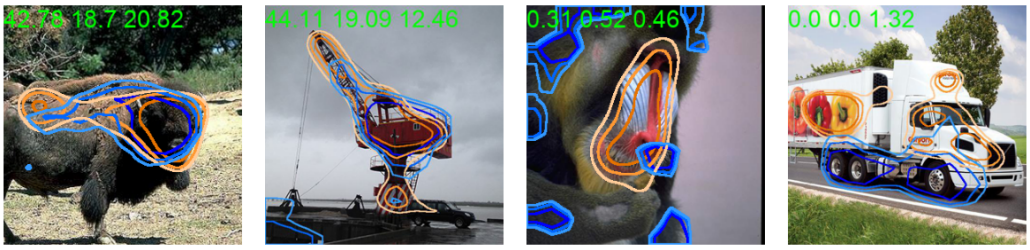


Figure 1.6: Overlap between the heatmaps created by Alignment Importance Scores (blue contours) and the saliency maps from TranSalNet (orange contours). The contours indicate the 5%, 10%, and 15% most important pixels, with increasing color intensity respectively. Relative Risk values computed from top 5%, 10% and 15% pixels in each map are printed on the top of each images. The two left images are examples of cases where AIS and saliency identified similar areas, whereas the two right images present extreme cases of non-overlap.

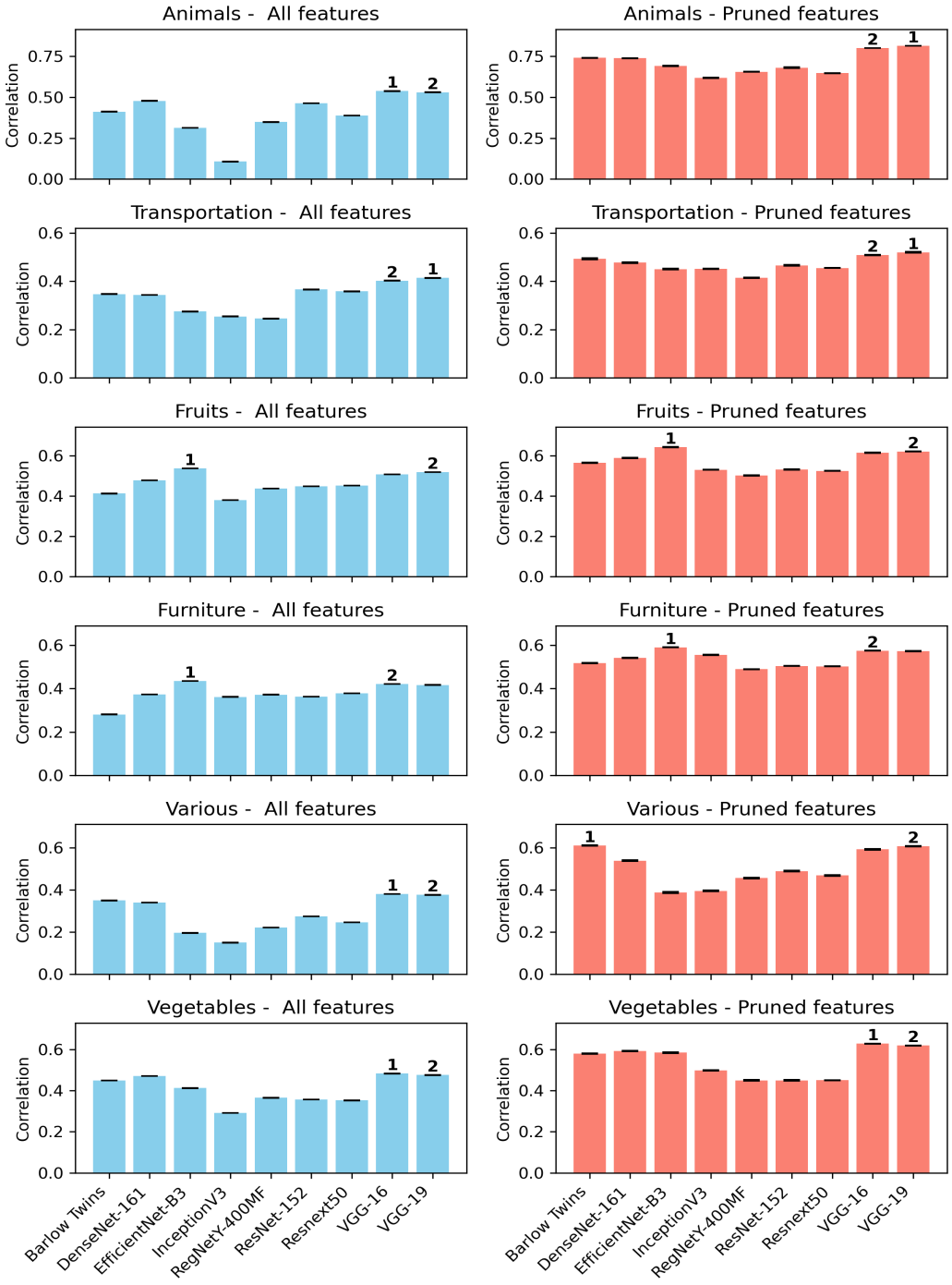


Figure 1.7: Cross-validation performance for models typically used as feature extractors. The numbers '1' and '2' refer to the two best performing models on the test set when using all features or only the features retained from the training set ('Pruned features').

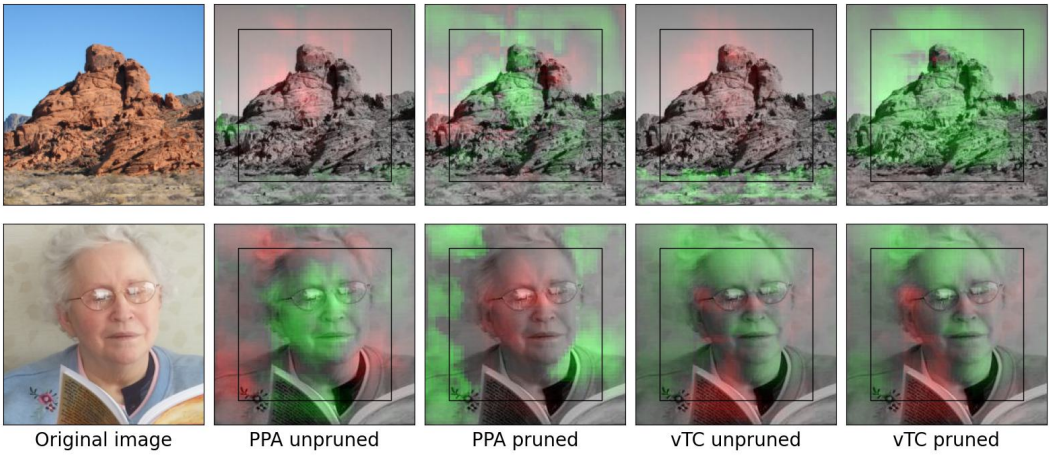


Figure 1.8: Heatmap in Tarigopula et al. (2023) showing the contribution of each image section to second order isomorphism between a DNN RDM and a Brain RDM. 'pruned' and 'unpruned' refer to whether or not the brain region supervised the pruning of the DNN. Green colors indicate image areas whose features contribute to shared DNN/Brain dimensions. The area within the inner black square was captured by masks at all scale-sizes; areas outside the black square also included padded data.

Chapter 2

Pruning for Reassessing Number-Detector Units in Convolutional Neural Networks

Abstract Convolutional neural networks (CNNs) have become essential models for predicting neural activity and behavior in visual tasks. However, their ability to capture higher-level cognitive functions, such as numerosity discrimination, remains debated. Numerosity, the ability to perceive and estimate the number of items in a visual scene, is often proposed to rely on specialized number-detector units within CNNs, analogous to number-selective neurons observed in the brain. In this study, we use CORnet, a biologically inspired CNN architecture inspired by the organization of the primate visual system. To address a limitation of classical Representational Similarity Analysis (RSA)—its assumption that all units contribute equally—we apply pruning, a feature selection approach that identifies the units most relevant for explaining behavioral similarity structure. Our results show that number-detector units are not critical for population-level representations of numerosity, challenging their proposed role in previous studies.

Code `https://github.com/alireza-kr/CORnetOnNumber`

Publication status This chapter is taken from the workshop paper: **Truong, N.**, Noei, S., & Karami, A. (2024, December). Reassessing Number-Detector Units in Convolutional Neural Networks. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*. `https://openreview.net/forum?id=n5cV83qUrs`

2.1 Introduction

Early breakthroughs in the study of biological vision served as the foundation for convolutional neural networks (CNNs; Lindsay 2021). Like the brain, these hierarchical models consist of several feedforward layers, with each layer comprising numerous artificial units that mimic neurons. Since then, CNNs have evolved into state-of-the-art models for predicting neural activity and behavior in visual tasks (Khaligh-Razavi and Kriegeskorte 2014; Daniel L. K. Yamins et al. 2014; Daniel L K Yamins and DiCarlo 2016; Cichy et al. 2016). For example, it has been demonstrated that CNNs trained on an object classification task can account for the brain responses of both humans' and monkeys' inferior temporal cortex (IT), a key region for object recognition (Khaligh-Razavi and Kriegeskorte 2014).

But what happens when the images contain multiple objects? Perceiving and representing the number of items in a set—known as numerosity—without counting is considered a core and ancient cognitive ability shared by humans and many animal species, often referred to as 'number sense' (Dehaene 2011). Specialized neurons, or 'number neurons', that are tuned to the number of items in a visual display have been identified in numerically naive monkeys (Viswanathan and Nieder 2013b), crows (Wagener, Loconsole, Helen M. Ditz, et al. 2018b), and untrained 10-day-old domestic chicks (Kobylkov et al. 2022b), suggesting that numerosity is automatically represented in the brain. Brain imaging studies have also pinpointed regions in the parietal cortex that are responsible for representing numerosity in both adults (Piazza et al. 2004; Castaldi et al. 2019; Karami, Castaldi, Eger, and Piazza 2025) and preverbal infants (Izard et al. 2008; Hyde et al. 2010; Edwards et al. 2015), demonstrating this ability at the population level. In addition, Karami, Castaldi, Eger, M. Hebart, et al. 2025 combining magnetoencephalography (MEG) with fMRI has shown that numerosity representations emerge rapidly after stimulus onset and evolve over time along the visual hierarchy, from early visual cortex to higher-level associative areas. Additionally, fMRI decoding in the parietal regions of adults has been linked to behavioral number discrimination acuity (Lasne et al. 2019). Collectively, these findings highlight the critical role of parietal brain activity in human number discrimination.

Recently, it has been shown that number-detector units, analogous to number neurons recorded in the prefrontal and parietal cortices of monkeys, can emerge in the final layers of CNNs trained for visual object recognition (Nasr et al. 2019b) and even in completely untrained CNNs (G. Kim et al. 2021b). However, Karami, Truong, et al. (2025), using RSA (Kriegeskorte 2008), demonstrated that CNNs fall short of explaining the variance in numerosity representation observed in fMRI data from the human parietal cortex. Further analysis

using multidimensional scaling (MDS; J. B. Kruskal 1964) revealed significant differences in the geometric structure of numerosity representations between human parietal regions and CNNs (Karami, Truong, et al. 2025). In the classical RSA framework used by Karami (2024), all features contribute equally to the final dissimilarity estimate. However, this 'equal weights' assumption conflicts with the notion that, when comparing representational dissimilarity matrices (RDMs), certain features may carry more informative content than others. As a result, this approach can underestimate the true correspondence between the model and a specific brain region or behavior (Kaniuth and Martin N. Hebart 2022b; Tarigopula et al. 2023). Moreover, the classical RSA approach may overemphasize non-relevant units by treating them as equally important as units that carry behaviorally relevant information, such as number-detector units in our case.

To assess the relevance of number-detector units in representing numerosity at the population level within CNNs, we employed a pruning approach. Pruning is a feature selection technique used to identify and retain the most relevant parts of a model, such as specific weights or activations, that best align with the behavior data and improve predictions (Flechas Manrique et al. 2023a; W. Bao and Hasson 2024; Truong et al. 2024). This approach is based on the observation that pretrained models often contain redundant information (Y. Cheng et al. 2015; Frankle and Carbin 2018). Therefore, using the entire model may not be necessary for a specific task, such as numerosity discrimination in our case. Specifically, we pruned different CNN architectures based on the number RDM, which captures the behavioral signature of numerosity perception in humans. This matrix serves as a benchmark for comparing the alignment of CNN representations with human numerosity processing. Our results revealed that number-detector units are not critical for representing numerosity at the population level within these networks. This finding suggests that, while number-detector units may emerge in specific layers of CNNs, they do not play a significant role in capturing the broader, population-level representation of numerosity, as reflected in human behavioral data.

2.2 Methods and experimental setup

2.2.1 Stimuli and Training the CNN

To investigate number-detector units in CNNs we used CORnet-Z and CORnet-S, models with four anatomically mapped areas (V1, V2, V4, and IT) followed by a decoder layer. CORnet-Z is the simplest network in the CORnet family and a lightweight alternative to AlexNet. CORnet-S also has recurrent connec-

tivity and is designed to maximize Brain-Score (Schrimpf, Kubilius, H. Hong, Najib J. Majaj, et al. 2018b). Each anatomically mapped area in the CORnet consists of a single convolution, followed by a ReLU nonlinearity, max pooling and the decoder is a 1000-way linear classifier (Kubilius, Schrimpf, Nayebi, et al. 2018; Kubilius, Schrimpf, Kar, et al. 2019).

We chose CORnet-Z and CORnet-S because it balanced the resemblance to the architectures used by previous studies on numerosity and because it well fit the visual system (Schrimpf, Kubilius, H. Hong, Najib J. Majaj, et al. 2018b). We used three versions of the CORnet:

1. the completely untrained version with randomly initialized weights to reveal the effect of architecture alone (Cichy et al. 2016),

2. a version trained on object recognition using the ImageNet dataset (Deng, Dong, Socher, L.-J. Li, N. K. Li, et al. 2009), which contained 1.2 million images of objects over 1000 categories, as it has been used in a previous study by Nasr et al. (2019b),

3. and a version of the network was specifically trained to discriminate between ten numerosity values: 6, 7, 9, 10, 12, 14, 17, 20, 24, and 29. We specifically trained the networks to discriminate between numbers because Mistry et al. (2023) demonstrated that training a CNN for numerosity discrimination significantly reorganizes the number-detector units. To avoid flawed stimulus design (Park 2022), where low-level visual features like size or dot density correlate with the number of dots, we used the method introduced by DeWind et al. (2015) to generate the dot sets. A sample of the generated stimuli used for training the network is shown in Figure 2.1A. Following the approach of Mistry et al. (2023), we first initialized the network with weights pre-trained on ImageNet, then trained it on the numerosity task for 100 epochs using the Stochastic Gradient Descent (SGD) optimizer with default PyTorch parameters.

After training, the three versions of the model were tested with visual dot sets, where both the number of dots and low-level visual features (average item area and total field area) varied across 32 different conditions: 4 numerosities (6, 10, 17, 29), 4 average item areas, and 2 total field areas (Figure 2.1B). Each input image was 500 × 500 pixels. We selected four layers of the network-analogous to visual brain areas (V1, V2, V4, and IT) and extracted the activations of all nodes in each layer. The results from 100 instances of each condition were averaged to produce a single activity vector for each condition from the output of each layer.

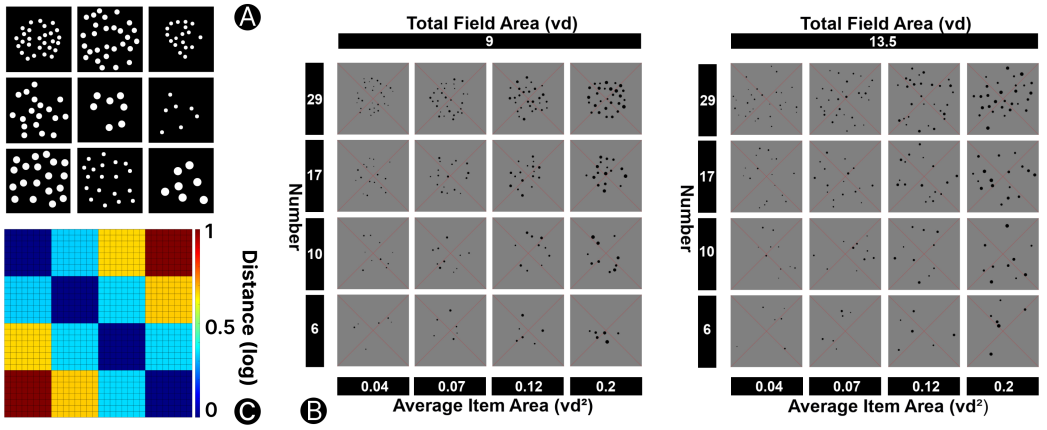


Figure 2.1: (A) Sample stimuli used for training the networks. (B) Sample stimuli used for testing the networks. (C) Number RDM, reflecting human behavioral data, used to prune the network layers.

2.2.2 Identification of Number-Detector Units

Following Nasr et al. (2019b), we used ANOVA to find number-detector units. Specifically, a three-way analysis of variance (ANOVA) with three factors - numerosity, total field area, and average item area - was applied to select number-detector units. The goal was to identify units that had a significant change in response across different numerosities while maintaining an invariant response across variations in total field area and average item area, as well as across the three interactions among pairs of factors, and the interaction among all of the three. A unit is marked as a number-detector if it shows a significant change for numerosity ($p < 0.01$) but no significant change for the other two factors or any of the interactions. Units that did not meet these criteria were classified as non-selective. This method of selecting number-detector units is analogous to the method that has been used to detect numerosity-sensitive neurons in monkeys and humans.

2.2.3 Representational Similarity Analysis between Number RDM and Network RDM

To create the CNNs' RDMs, we used the 32 activity vectors obtained by averaging the 100 instances per condition. We chose 100 instances to address concerns about the limited number of sample images used in previous studies, such as Nasr et al. (2019b), which were criticized for this limitation (X. Zhang and X. Wu 2020). The CNNs' RDMs were constructed using 1 – Pearson correlation between the activations of each layer for each pair of conditions.

The number RDM (Figure 2.1C) was based on the logarithmic distance between the pairs of conditions in terms of numerosity. We then compute the correlation between each CNN's RDM and the number RDM.

Pruning the Layers of Models

The pruning algorithm, which is adapted from Tarigopula et al. (2023), involves three steps. First, the importance of each unit is assessed by removing it from the full set of units. Each time a unit is removed, a new RDM is computed, and its score is compared to the number RDM. A significant drop in the score compared to the full set RDM indicates that the unit is important for matching the number RDM, while a smaller drop or an increase in score suggests the unit is unimportant or possibly encoding noise. Second, all units are ranked based on their importance scores, from highest to lowest. Third, starting with an empty activation vector, units are sequentially added back in the order of their ranking. After each addition, the fit between the RDM derived from the new embedding and the number RDM is re-evaluated. We truncate and select the set of neurons when the highest RSA score is achieved, and refer these units as the 'retained units' after pruning.

2.3 Results

2.3.1 Retained Units After Pruning and Number-Detector Units Often Do Not Overlap

Table 2.1: Number of retained units after pruning, and of number-detector units identified by ANOVA. The full set of units in V1, V2, V4, and IT layer in both models are 262144, 131072, 65536, 32768 respectively. The numbers in parentheses denote the percentage of units compared to the full set.

CORnet	**Layer**	**After Pruning**			**ANOVA**		
		ImageNet	DeWind	Untrained	ImageNet	DeWind	Untrained
Z	**V1**	101511 (39)	100081 (38)	72429 (28)	68 (0.03)	83 (0.03)	77 (0.03)
	V2	45421 (35)	45584 (35)	40176 (31)	44 (0.03)	24 (0.02)	1 (0)
	V4	32264 (49)	31979 (49)	20358 (31)	13 (0.02)	27 (0.04)	13 (0.02)
	IT	6074 (19)	1971 (6)	4139 (13)	32 (0.1)	10 (0.03)	3 (0.01)
S	**V1**	58310 (22)	117169 (45)	100637 (38)	803 (0.31)	939 (0.36)	657 (0.25)
	V2	6886 (5)	39881 (30)	12344 (9)	414 (0.32)	121 (0.09)	454 (0.35)
	V4	423 (1)	16444 (25)	2595 (4)	101 (0.15)	62 (0.09)	257 (0.39)
	IT	66 (0.2)	37 (0.1)	424 (1.3)	126 (0.38)	0 (0)	125 (0.38)

Table 2.2: The overlap scores between the units selected by the two methods. The scores are defined as the number of units presented in both two sets divides by the number of units in the smaller set, ranging from 0 (in case of no overlap) to 1 (in case of the bigger set contains the entire smaller set). The value for CORnet-S trained on DeWind was undefined as there was no unit detected via ANOVA.

CORnet	Layer	ImageNet	DeWind	Untrained
Z	**V1**	0.40	0.57	0
	V2	0.59	0.71	1
	V4	1	0.85	1
	IT	0.09	0	0
S	**V1**	0.31	0.27	0.65
	V2	0.01	0.21	0.01
	V4	0	0	0
	IT	0	-	0

Table 2.1 presents the number of units selected by two methods: pruning and ANOVA. In both models, the number of retained units after pruning is significantly higher than the number of number-detector units identified by ANOVA. ANOVA results in significant more units in CORnet-S compared to Z. Moreover, pruning removed the largest proportion of units in the IT layer, while the ANOVA method showed no significant differences across layers. There is no noticeable difference between the ImageNet-trained and DeWind-trained models. Interestingly, both retained units and number-detector units are found in untrained networks, consistent with the findings of G. Kim et al. (2021b).

The overlap between the units selected by the two methods is defined as the number of units shared by both sets, divided by the number of units in the smaller set: $|\text{Pruning} \cap \text{ANOVA}| \,/\, \min(|\text{Pruning}|, |\text{ANOVA}|)$. The overlap scores, as shown in table 2.2, vary considerably across layers and models. Little to no overlap was observed in the IT layer of both models and in the V4 layer of CORnet-S, while significant overlap was found in the V2 and V4 layers of CORnet-Z, and in the V1 layer of CORnet-S. Only three cases showed a perfect overlap score of 1, while seven cases had a score of 0. Overall, the results suggest that the two methods generally select non-overlapping sets of units.

2.3.2 Retained Units after Pruning Fit the Behavior Data Better than Number-Detector Units

Figure 2.2 shows that the retained units after pruning provide a better fit for modeling the number RDM compared to the full set of units, highlighting the limitations of using the full set in classical RSA. Additionally, in most cases—except for V2 and V1 in CORnet-S DeWind, which is specifically trained for number discrimination—the number-detector units selected by ANOVA, as commonly studied in traditional literature, perform worse than the retained units after pruning. In 8/22 cases, they are even worse than the full set of units. This demonstrates that the number-detector units are not important for capturing the population-level representation of numerosity in human behavioral data.

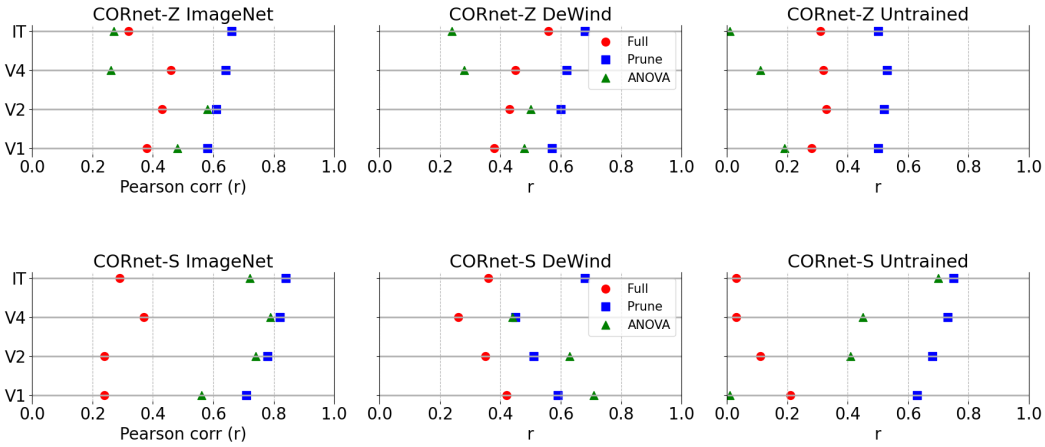


Figure 2.2: The Pearson correlations, quantified using RSA, from the full set of units, the retained units after pruning, and the number-detector units identified by ANOVA. The missing data points are due to an insufficient number of minimum units (2) required to compute the correlations.

2.4 Discussion

It has been emphasized that cognitive and behavioral functions emerge from the collective dynamics of neural populations, rather than isolated neuronal activity (Yuste 2015), underscoring the importance of population-level analysis for understanding complex behaviors like numerosity discrimination. In this context, RSA has become a widely used method for comparing representa-

tional spaces from human brain activity, behavioral data, and computational models. However, classical RSA assumes that each feature is equally important, making it difficult to interpret the contribution of individual features. In our case, classical RSA does not provide insights into the importance of number-detector units identified in CNNs. To address this limitation, we applied pruning, a feature selection technique used to identify and retain the most relevant components of a model—such as specific weights or activations—that best align with behavioral data. Pruning revealed that the number-detector units are not essential for contributing to the representation of numerosity at the population level, casting doubt on their significance for numerosity discrimination. These findings are also consistent with previous work by Mistry et al. (2023), which demonstrated that training a CNN for numerosity discrimination significantly reorganizes the number-detector units, and by Chapalain et al. (2024), which showed that dot-pattern-tuned units do not generalize to object-number information in photorealistic stimuli.

Future work could use explainable AI techniques to decode the semantic content within the subset of units selected by both methods. Extending the work on more ecological, natural image datasets (Upadhyay and Varma 2023; Hou et al. 2024) or exploring the language domain (Shah et al. 2023; Wennberg and Henter 2024; Varma et al. 2024) are also promising directions.

Chapter 3

Sparsity-guided Pruning to Preserve Representational Geometry and Model Human Similarity Judgments

Abstract While deep neural networks are increasingly adopted in cognitive sciences, they are often computationally expensive and contain irrelevant information for downstream tasks. In contrast to pruning approaches that aim to maintain classification accuracy, we present a pruning method to compress entire models while preserving their representation geometry. The target representational space can derived from a neural network or from human similarity space. Our method involves eliminating sparse, rarely activated components throughout the entire network architecture. We show that a deep model's representational space can be preserved or minimally altered when sparse features are removed, producing a compact model for network distillation and predicting human similarity judgments.

Publication status This chapter is extracted from a manuscript in preparation. The manuscript contains the extension of the works that were presented as two conference abstracts: 1) **Truong, N.**, & Hasson, U. (2024). Pruning sparse features for cognitive modeling. In *The 7th annual conference on Cognitive Computational Neuroscience*, and 2) **Truong, N.**, Bavaresco, A., & Hasson, U. (2023). The impact of rarely-firing nodes in neural networks on representational geometry and predictions of human similarity judgments. In *Conference on Cognitive Computational Neuroscience 2023* (pp. 1025-1028). The code will be published in the preprint version.

3.1 Introduction

3.1.1 Sparseness in the brain and in deep neural networks

Biological and artificial neural systems can exhibit sparse activations, where only a small subset of units responds to a given input, in contrast to dense representations. In biology, sparse coding has been observed across domains, including visual responses in visual cortex and odor responses in hippocampus and amygdala (Rolls and Tovee 1995; Kehl et al. 2024). In vision, sparsity is often attributed to metabolic efficiency and selective population coding. For example, population sparseness can be extremely high in monkey visual cortex, with ¡1% of neurons active for a stimulus (S. Tang et al. 2018).

Modern deep neural networks (DNNs) trained on image classification tasks also produce highly sparse representations. In VGG-16, units in the deepest fully connected layers respond, on average, to only 65% of the input images, with some units not activating for more than 90% of the images Hengyuan Hu et al. 2016. However, sparsity in DNNs is not produced by the same pressures shaping neurobiological systems. Instead, it is at least a byproduct of overparameterization: models are trained with far more parameters than are required to perform the task, as demonstrated by extensive research on pruning (see review H. Cheng et al. 2024; F. Chen et al. 2024; Menghani 2023; Y. He and Xiao 2023; Marinó et al. 2023; Lê et al. 2023). In particular, pruning motivated by the "lottery ticket hypothesis" (Frankle and Carbin 2018) has shown that small subnetworks within the original model, which consist of a subset of the original weights, can reach the same accuracy of the full network.

Overall, while neurobiological sparsity is functional, sparsity in DNNs is at least partially a side effect of overparameterization. Units that contribute to sparsity in DNNs do not necessarily contribute to population coding, suggesting that sparsity in DNNs should be treated differently than in biological systems. This is especially relevant when using pretrained models to study the geometry of specific semantic categories (e.g., the dataset of Peterson et al. 2018), because units inactive for a category cannot support within-category distinctions and should not involve in the analysis. Furthermore, if different categories activate different subsets of DNN units, this would suggest that the network organizes information in partially disjoint subspaces, corresponding to different semantic domains (Bavaresco and Fernández 2025). Thus, when studying the representational geometry of a DNN, across or within categories, it is useful to distinguish between active and inactive units.

3.1.2 Correlation Retaining Iterative Structural Pruning (CRISP)

There is currently no pruning method designed to approximate a target representational geometry (e.g., from human behavior or from the internal geometry of the model itself) by removing sparse, non-informative units. To address this gap, we develop CRISP (Correlation Retaining Iterative Structural Pruning), a structured, activation-based, global, iterative pruning method that compresses pretrained CNNs by removing units and filters without fine-tuning and without changing the weights of retained units. CRISP ranks candidates from the whole network by activation sparsity (e.g., high "percentage of zeros", following Hengyuan Hu et al. 2016) and iteratively prunes while monitoring preservation of representational similarity: after each step it recomputes activations and derives an RDM, and continues until the correlation with a reference RDM until the alignment drops below a threshold. Because it preserves similarity structure, CRISP isolates the subspaces that are most important for stimulus discrimination while facilitating explainability at the unit/filter level. Specifically, consistently inactive features can reveal what the model does not use to encode within a stimulus domain.

3.2 Methods

3.2.1 Model and Datasets

We used VGG-16 architecture Simonyan and Zisserman 2014 pretrained on ImageNet Deng, Dong, Socher, L.-J. Li, K. Li, et al. 2009. To evaluate the alignment between the model and human judgments, we used data from Peterson et al. 2018, which includes pairwise-similarity judgments for six datasets, each containing 120 images. These datasets were Animals, Fruits, Furniture, Vegetables, Transportation (referred to as Automobiles in the original paper) and Various. The latter combining images from the other categories. For each dataset, ten participants judged the similarity of all image pairs, rating them on a scale from 1 (completely dissimilar) to 10 (very similar). This process resulted in six matrices of average human ratings, each with dimensions of 120×120. The judgment scores and relevant images were kindly provided by the authors.

3.2.2 Pruning criterion

We used the Percentage of Zeros (PoZ) as the pruning signal. PoZ quantifies the proportion of zero activations after applying the ReLU function, reflecting

a unit's inactivity across a dataset Hengyuan Hu et al. 2016.

For a fully connected layer, PoZ is computed per unit as:

$$\mathrm{PoZ}_j = \frac{1}{B} \sum_{b=1}^{B} \mathbb{I}(a_b^{(j)} = 0), \tag{3.1}$$

where a_b^j is the post-ReLU activation of unit j for sample b, B is the total number of samples, and $\mathbb{I}(\cdot)$ is the indicator function.

For convolutional layers, PoZ is computed per feature map as:

$$\mathrm{PoZ}_k = \frac{1}{BHW} \sum_{b=1}^{B} \sum_{h=1}^{H} \sum_{w=1}^{W} \mathbb{I}(a_{bhw}^{(k)} = 0), \tag{3.2}$$

where $a_{bhw}^{(k)}$ is the activation at location (h, w) in feature map k for sample b, and $H \times W$ is the spatial size of the feature map.

Implementation of CRISP

We first consider the target representation geometry comes from the model itself, operating by a representational similarity/dissimilarity matrix computed from activations of certain layers in the model. Our pruning method, CRISP, is an iterative, structural pruning approach that eliminates units based on their PoZ, while preserving the representational geometry of the original network above a certain pre-defined threshold, while not altering the weights of all the retained units. At each iteration, the units or filters with the highest PoZ are structurally removed from the model. This involves permanently eliminating the unit from the architecture, not just zeroing the weights.

After each pruning step, we compute the representational similarity matrix (RSM) from penultimate-layer activations and evaluate its Spearman correlation ρ with the original RSM. Pruning continues as long as ρ remains above a user-defined threshold ρ_{target}. After each pruning step, PoZ is recomputed for all remaining units: even though the retained weights remain fixed in magnitude across iterations, removal of units changes the information flow in the network and therefore the PoZ of units in all layers deeper than those of the removed unit.

Input:
1. Network N_0
2. The layer that we want to approximate its geometry
3. Threshold Spearman ρ_{target} for stopping the pruning process
4. Granularity G indicating number of units to be removed in each iteration

`// e.g. 1 for small models; 50 for large models`

Output: Pruned sub-network N_p that preserves representational geometry at the chosen layer with $\rho \geq \rho_{\text{target}}$

Data: Dataset D

Compute the original representational similarity matrix RSM_0 using N_0 on D
Initialize $N \leftarrow N_0$
stop_flag $\leftarrow$ False

while *stop_flag = False* **do**
  Pass D through N and extract activations from all layers
  Compute PoZ (percentage of zeros) for all units in N
  Identify the G units with the highest PoZ

  Create a new network N' by structurally removing units G from N
  Compute the representational similarity matrix RSM' at the chosen layer of N'
  Compute $\rho = \text{Spearman}(RSM', RSM_0)$

  if $\rho \geq \rho_{target}$ **then**
    Update $N \leftarrow N'$ `// Accept the removal`
  else
    stop_flag $\leftarrow$ True `// Stop when removal reduces geometry below threshold`

Return final pruned network $N_p \leftarrow N$

Application of CRISP in VGG-16

We applied CRISP to a pretrained VGG-16 model in two ways, corresponding to two targets of representational geometry. First, we guided pruning by RSMs produced from each category's image-by-image similarity matrices as computed from the entire (full) embedding matrix. Second, we used human similarity judgments for the images within each category as guidance. Here, we focused only on embeddings in `fc1` and `fc2` layers and evaluated whether it is possible to derive pruned versions of the model that better align with

human similarity judgments. Due to the large architecture of VGG-16 (over 10,000 prunable units), we removed 50 units or feature maps per iteration to reduce computational cost.

3.3 Results

We tracked representational similarity across pruning iterations for fc1 and fc2, both relative to the original unpruned network (Figure 3.1) and relative to HSJs (Figure 3.2), separately for each semantic category in Peterson et al. (2018). For both layers, RSA relative to the original model remains high across a large range of pruning iterations, nearly at ceiling for fc1, indicating that substantial removal can be performed without strongly changing the representational geometry. Beyond approximately 100 iterations (depends on the datasets), RSA drops significantly, marking a transition where further pruning rapidly decrease the geometry. Alignment with HSJs shows a similar pattern: correlations remain largely stable during first around 100 iterations, and then collapse near the point where model geometry diverges from the original. Across panels, the 'various' category degrades earliest, which makes sense as a sanity check because it contains mix categories. Overall, the results suggest that the core, relevant geometry that represent a certain category is supported by a relatively small set of units and feature maps with low PoZ that can be retained under heavy pruning.

These preliminary results serve as a proof of concept for further analyses in the future. Pruning could be extended to earlier layers to test the utility of CRISP. Explainability analyses should also be included to decode the semantics of retained and removed units or feature maps, revealing what information is important for preserving the geometric structure within a semantic category in pretrained models.

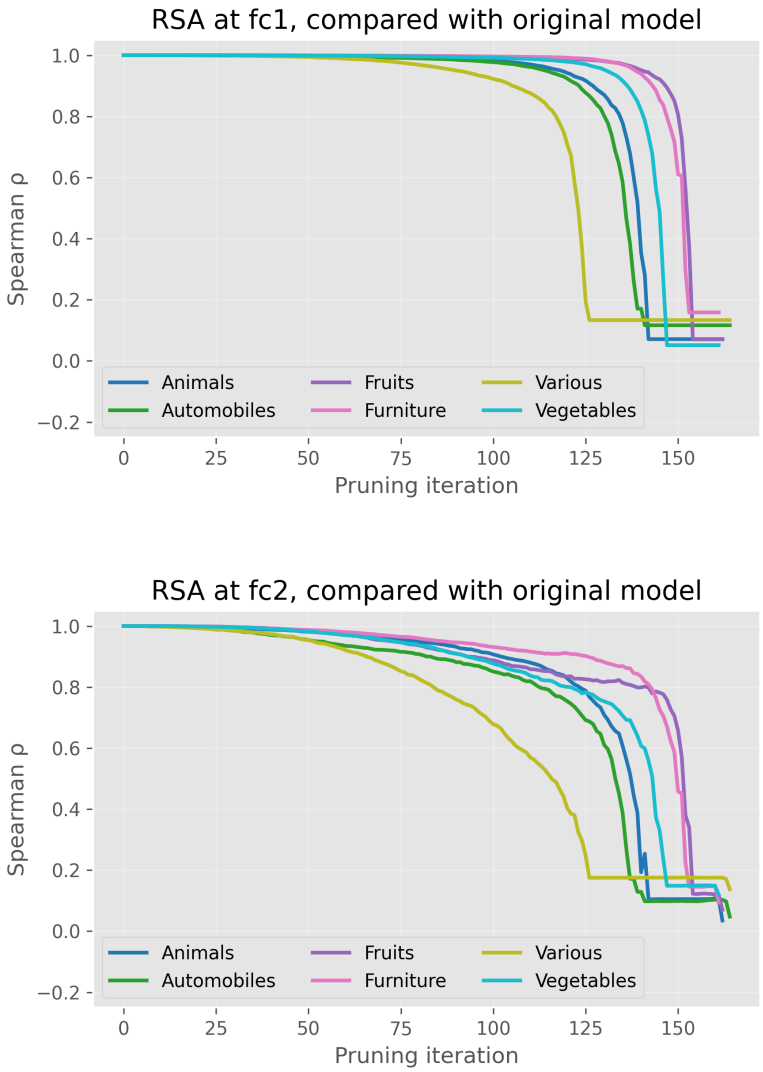


Figure 3.1: Tracking the representational geometry of the pruned models against that of the original model. The chosen layers are fc1 (top) and fc2 (bottom) from the fully connected layers. Each pruning iteration consists of 50 units or feature maps that have the highest PoZ at that iteration.

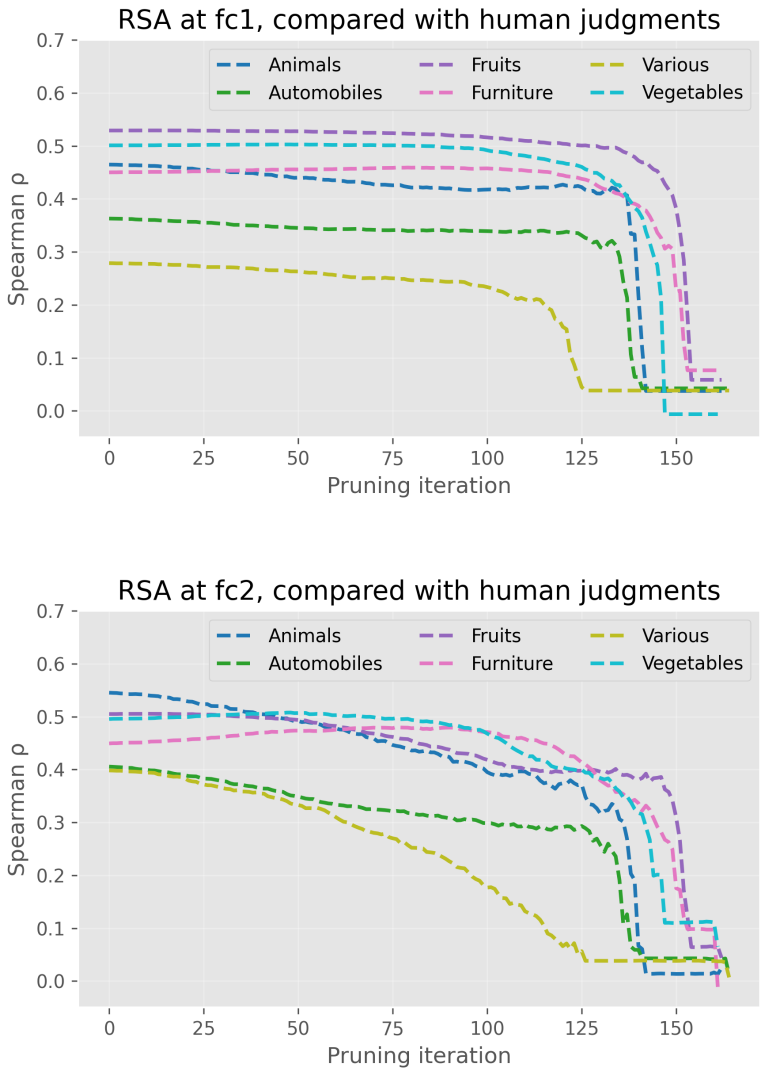


Figure 3.2: Tracking the representational geometry of the pruned models against human similarity judgments. The chosen layers are fc1 (top) and fc2 (bottom) from the fully connected layers. Each pruning iteration consists of 50 units or feature maps that have the highest PoZ at that iteration.

Chapter 4

Investigating Action Topography in Visual Cortex and Deep Artificial Neural Networks

Abstract High-level visual cortex contains category-selective areas embedded within larger-scale topographic maps like animacy and real-world size. Here, we propose action as a key organizing factor shaping visual cortex topography and assess the ability of topographic deep artificial neural networks (DANNs) in capturing this organization. Using fMRI, we examined responses to images of body-parts and objects with different degrees of action properties. In left lateral occipitotemporal cortex, we identified a topographically-organized action gradient, with overlapping activations for bodies, hands, tools, and manipulable objects along a dorsal-posterior to ventral-anterior axis, culminating at the intersection of body parts and objects exhibiting higher action properties. Multivariate analyses confirmed action as a crucial organizing principle, while shape and animacy dominated ventral occipitotemporal cortex and DANNs, which exhibited no action-based organization. Our proposed action dimension serves as a further organizing principle of object categories, advancing understanding of visual cortex organization and its divergence from DANN-based models.

Code `osf.io/ctmbx/`
`github.com/DavideCortinovis/Action-topography-in-visual-cortex`

Publication status : The author of the thesis contributed to the topographic modeling section in the paper: Cortinovis, D., **Truong, N.**, Op de Beeck, H., & Bracci, S. (2025). Investigating action topography in visual cortex and deep artificial neural networks. *Nature Communications*. This chapter contains the full abstract and introduction of the paper, and additionally the methods, results, and discussion involving the spatial organization of visual cortex and topographic models. Although the fMRI analyses are not the contributions of the author of this thesis, they are included in this chapter to support understanding of the topographic modeling section.

4.1 Introduction

Topography – the systematic, spatial organization in which neurons (or voxels) with similar functional properties are located near one another in the cortex – is ubiquitous throughout the cortex, from the retinotopy and pinwheels of primary visual cortex to the complex somatotopic organization of body parts in the so-called motor homunculus in M1 (Durbin and Mitchison 1990; Wandell et al. 2007; Penfield and Boldrey 1937). In occipitotemporal cortex (OTC), a topographic organization of functionally selective areas has been shown, with areas responding preferentially to ethologically-relevant categories such as faces, body parts, words, and scenes, mirrored along the ventral and lateral OTC, and forming a consistent spatial arrangement across participants (Kanwisher 2010; H. P. Op de Beeck et al. 2008; Taylor and Downing 2011; Grill-Spector and Weiner 2014).

Several accounts have tried to explain this organization by highlighting the role of different features that map object space onto the two-dimensional cortical sheet, leading to the emergence of functionally selective areas. These features span from low-level principles like eccentricity, to mid-level properties (e.g., curvature, aspect-ratio, texture), and to semantic principles like animacy and real-world size (Gomez et al. 2019; I. Levy et al. 2001; R. Malach et al. 2002; Yue et al. 2020; P. Bao et al. 2020; Coggan and Tong 2023; Jagadeesh and Gardner 2022; Kriegeskorte et al. 2008; Konkle and Oliva 2012). Some of these dimensions appear to be repeated across ventral and lateral OTC, explaining the mirrored organization of category-selective areas (Hasson et al. 2003; Konkle and Caramazza 2013; Silson et al. 2015). Remarkably, the representational space of higher-level layers in DANNs trained on object recognition captures the same object dimensions observed in the visual cortex (e.g., animacy, aspect-ratio – but see Yargholi and H. Op de Beeck 2023 – shape, real-world size; Khaligh-Razavi and Kriegeskorte 2014; P. Bao et al. 2020; H. O. d. Beeck et al. 2023; Zeman et al. 2020; T. Huang et al. 2022). Moreover, topo-

graphic DANNs – architectures that incorporate biologically inspired spatial constraints – develop category-selective responses (e.g., for faces, bodies, and scenes) that mirror the topographic organization found in the visual cortex (Blauch et al. 2022; Margalit et al. 2024; Z. Lu et al. 2025).

Notably, accumulating evidence suggests that despite the fact that lateral and ventral OTC show a similar mirrored object topography, their underlying representational space might be better explained by different object dimensions (Lingnau and Downing 2015; Wurm and Caramazza 2022). For instance, the left lateral OTC shows sensitivity to categories characterized by their action-related properties such as hands and tools, whose underlying selectivity is spatially adjacent to, and partially overlaps with, one another (Bracci and H. O. d. Beeck 2016; Mahon, Milleville, et al. 2007; Weiner and Grill-Spector 2010; Bracci, Cavina-Pratesi, et al. 2012). Hands and tools differ in many visual and semantic properties, such as their shape and animacy; eccentricity and real-world size accounts also cannot explain this pattern of results as this effect does not extend to other object categories sharing similar eccentricity or real-world size (Bracci and Peelen 2013; Striem-Amit et al. 2017). Instead, this evidence suggests that another dimension plays a role in shaping the topographic organization of visual cortex object space: action (Bracci and Peelen 2013).

The present study aims to investigate the principles underlying the organization of functionally selective areas, with a focus on how behaviorally relevant action properties of objects shape the spatial organization and content of representations in ventral and lateral OTC. We conducted an fMRI experiment where participants viewed images of body parts and objects varying in their degree of action properties (Matić et al. 2020).

Using univariate and multivariate analyses on fMRI data, along with representational predictions based on human similarity judgments, we tested how action dimensions interact with other proposed dimensions and compared results in human visual cortex with DANNs. Our results show a dissociation between ventral and lateral OTC in both topography and representational space. Action—alongside shape and animacy—emerged as a key principle explaining the arrangement of categories in lateral OTC, while animacy best explained topography and representational content in ventral OTC and in DANNs, which in turn did not show any action-related organization. These results demonstrate that action is a fundamental organizing dimension of OTC, and that further developments are necessary for current computational models to fully capture both topography and function of high-level visual cortex.

4.2 Methods

4.2.1 fMRI experiment and analyses

Participants

19 participants took part in the fMRI experiment (11 females, sex self-reported, mean age 25.6 years, standard deviation 6.06). One male participant was excluded due to head motion exceeding one voxel. All participants were right-handed except one, all had normal or corrected-to-normal vision, and no history of neurological disorder. All participants gave informed consent. The Ethics Committee of the University of Trento approved the procedure.

Stimuli

The stimulus set included 6 categories (Figure 4.1). Part of the images were used in Matić et al. (2020). The set comprised 3 body-parts (hands, headless bodies, and faces), 3 inanimate object categories (tools, manipulable objects, and non-manipulable objects), and chairs as a control category. Each object category was associated with a different degree of action-related properties. Tools were defined as hand-held objects that are typically used to physically and directly act on another object or surface (e.g., hammer); therefore, tools are not only graspable and manipulable, but also serve as action-effectors, akin to our hands (Bracci and Peelen 2013). Manipulable objects are objects that can be grasped, lifted, and manipulated but are not usually used as action-effectors (e.g., glass). Finally, non-manipulable objects were defined as large objects that cannot be grasped nor manipulated (e.g., bed). To control for low- and mid-level visual features, the object categories were matched for their perceived shape and orientation (Figure 4.1). In addition, tools and manipulable objects were matched for real-world size, ensuring that any difference between the two categories cannot be attributed to their actual size. Each category included 12 grey-scale images with a white background of 400x400 pixels.

Scanning procedure

In the fMRI experiment we collected 8 runs per participant. Each run lasted 400 sec (200 volumes). Each image was presented for 0.4 s, with an ISI of 0.266 s, in blocks of 8 s (i.e., 12 images per block). For each subject and for each run, a fully randomized sequence of all conditions was repeated 4 times, with a fixation block of 16 seconds at the beginning, in the middle (between sequences), and at the end of each run. Stimuli were presented with the Psychophysics Toolbox package (Brainard and Vision 1997) in MATLAB

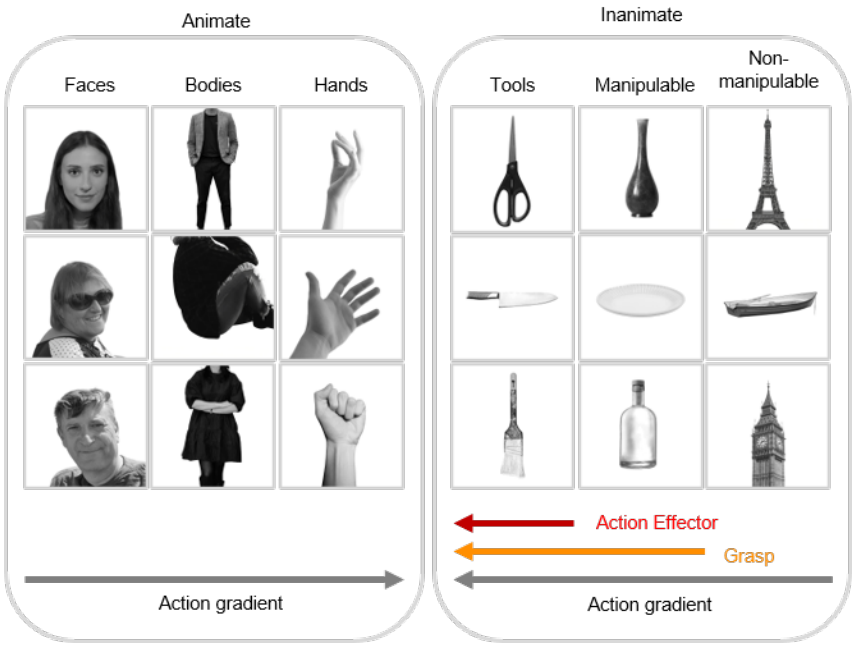


Figure 4.1: **Stimulus set.** Images were divided into 6 categories varying along two dimensions, animacy and action. For inanimate objects, action was characterized by two properties, action-effector (red) and graspability (orange). The three inanimate objects were matched for visual shape and orientation, to avoid confounds based on the overall shape (e.g., the elongation) of the stimuli. All images in this figure have been replaced with photographs obtained from Unsplash (`https://unsplash.com`), which provides images under a license allowing free commercial use without permission. Images were selected to be visually similar to the original stimuli. Face images were replaced with photographs of individuals who provided explicit consent for publication.

(2021b) (The MathWorks). Images were projected onto a screen (8 x 8 degrees of visual angle) and shown to the participants through a mirror mounted on the head coil. Participants were instructed to fixate their gaze on the fixation cross in the middle of the screen and press a button whenever the same image was repeated twice in a row within each block. The repeating image appeared once per block.

The imaging parameters and preprocessing process are described in the full paper.

Category overlap analysis

We measured the amount of voxel overlap between the activation clusters for each condition, separately for ventral and lateral OTC. To do that, we selected two masks using a combination of functional and anatomical criteria; specifically, we used the Neuromorphometrics atlas (Neuromorphometrics, Inc.) to define regions within ventral and lateral OTC; ventral OTC included the fusiform gyrus and the parahippocampal gyrus, whereas lateral OTC included the inferior and middle occipital gyri and the inferior and middle temporal gyri; within these anatomical regions we selected all the active voxels with a contrast of all conditions vs. baseline with a liberal threshold ($p < .05$ uncorrected); these masks, which contain only the voxels modulated by visual information, were used for the subsequent analysis. To compute the overlap analysis, we calculated the number of active voxels within each of the two masks for each condition vs. all remaining conditions (e.g., hands vs. all others) with a more conservative threshold ($p < .001$ uncorrected at the voxel level and $p < .05$ FDR-corrected at the cluster level). Applying a cluster correction ensures that only contiguous voxels with a meaningful minimum size are considered for the analysis. The resulting active voxels were employed to compute the overlap index which was calculated pairwise for all possible combination of categories by taking the number of voxels common to two clusters (for instance, the voxels that are active for both hands and tools) and dividing it by the number of voxels of the smaller of the two clusters. An index of 0 indicated no overlap between two categories, whereas an index of 1 indicates that the smaller cluster of a category falls completely within the bigger cluster of the other category. Following previously adopted approaches (e.g. X. Luo et al. 2024), we calculated the overlap at the group level. Overlap analysis at the group level may introduce smoothing that overestimate the amount of overlap between categories; however, previous comparisons of overlap analyses based on single subjects vs. group analyses revealed little differences in the results between the two (Cant and Xu 2012); moreover, the use of relatively conservative thresholds and the use of selective contrasts ensure the control of overestimation of overlap effects.

4.2.2 Topographic networks[1]

We tested a computational model of the spatial organization of ventral visual cortex to test its possible convergence or divergence with the spatial organization of visual cortex.

[1]The thesis's author developed these analyses.

As standard artificial neural networks do not have topographic constraints, we selected a further recently developed family of models that implement some constraints within their architecture to mimic the topographic organization of visual cortex (Margalit et al. 2024). These models – called Topographic Deep Artificial Neural Networks or TDANNs – were based on a ResNet-18 architecture and were trained with a self-supervised contrastive learning task on the ImageNet dataset (T. Chen et al. 2020). Prior to training, a mapping of units is implemented within each layer of the network, so that each unit has a corresponding 2D coordinate that maps them into a 2D grid that represents their physical distance. During training, a spatial loss function (together with the self-supervised task loss) is introduced: this function constraints nearby units to have correlated firing patterns to the same features within the dataset, so that the units that have similar functional properties will fall close in the simulated physical space. A parameter called α in the spatial loss function indicates how much the neighbouring units must be correlated with each other; following Margalit et al. (2024), we used a value of $\alpha = 0.25$, as it has been demonstrated to be the optimal value for the emergence of VTC-like topographic organization. These networks include 8 layers implementing topographic constraints, with different surface areas across layers to simulate the hierarchy of the ventral visual stream, from V1 to high-level VTC. We use five different random initializations of the network weights.

Data analyses

Univariate We performed simulated univariate analysis by testing the topographic organization and selectivity profile of the five different random initializations of the network in response to our six object categories; most analyses were conducted on the last layer that qualitatively showed the clearest clustering by categories, which we called VTC-like layer (as in Margalit et al. 2024). Specifically, we tested 1) the clustering of units selective for the different object categories within the simulated physical cortical space in the VTC-like layer and 2) the selectivity profile of the top-50 most selective neurons for each category in the VTC-like layer.

Overlap To examine whether object categories in the VTC-like layer of the TDANN exhibit a similar relationship to those found in the OTC, we measured the overlap in selectivity between units across different conditions. We followed the method introduced by Margalit et al. (2024). Specifically, the simulated cortical sheet was partitioned into 1 mm wide square sections. In each section, we assessed the proportion of units that were selective ($t > 3.5$) for two categories (e.g., hands and tools, hands and faces, etc.) in pairs. The

overlap between these categories was determined by analysing the frequency of selectivity co-occurrence of the two categories within each section. Essentially, if the selectivity frequency for one category can predict the selectivity for the other, the unit populations are considered to overlap. This overlap is measured using an index that ranges from 0 to 1: a score of 0 means the presence of units selective for one category (e.g., hands) always predicts the absence of units selective for the other (e.g., tools); a score of 0.5 indicates no predictability between the two categories; and a score of 1 signifies perfect overlap, where the presence of units selective for one category always coincides with the presence of the other category.

4.3 Results

To investigate how action-related properties influence object topography in visual cortex, we designed a stimulus set organized along two dimensions: animacy (body parts vs. inanimate objects) and action. Specifically, the three inanimate categories vary along two action-related properties: action effector and graspability (Figure 4.1). Tools are both action effectors and graspable; manipulable objects are graspable but not effectors; and non-manipulable objects are neither effectors nor graspable. The three body parts also differed in action relevance: low for faces, higher for bodies, and highest for hands. We first present results for the action-related organization of visual cortex, and then test if topographic models capture the same organization.

4.3.1 Action properties differentially shape object topography in ventral and lateral OTC

To investigate object space organization in ventral and lateral OTC (VOTC and LOTC, respectively), we first mapped the activation response for each category (versus all others, $t > 3.5$, $p < .05$ FDR corrected at the cluster level) onto the whole-brain surface (Figure 4.2a). Beyond replicating the known parallel organization of category selective responses in lateral and ventral OTC (Pillet et al. 2024), the whole-brain analysis confirmed a dissociation between the VOTC and LOTC in the left hemisphere (Figure 4.2a) based on the activation patterns for object classes with varying degrees of action-related information. Whereas in VOTC we found the typical medial-to-lateral animacy division with no overlap between animate and inanimate categories (Grill-Spector and Weiner 2014), in LOTC we observed overlapping responses between animate and inanimate conditions with a different degree of action properties. From dorsal-posterior to ventral-anterior, we observed selective

and partly overlapping activations for bodies, hands, tools, and manipulable objects, with a convergence and high degree of overlap for the animate and the inanimate categories characterized by the highest degree of action properties: hands and tools. The action-based organization was particularly evident when comparing activations of inanimate objects. Specifically, we found a consistent action-related gradient in LOTC, with a smooth transition across the cortical surface where the activation to object categories characterized by different action properties changes systematically according to the two action-related properties. This gradient was characterized by a large activation cluster for tools which are both action-effector and graspable, a smaller cluster for manipulable objects which are only graspable, and no significant activation for non-manipulable objects which are neither action effector nor graspable; the opposite pattern was observed in VOTC, with a larger cluster for non-manipulable relative to manipulable objects, which in turn revealed a larger activation relative to tools. Unlike the left hemisphere, the right hemisphere did not show any action-related organization, as neither tool nor object selectivity were observed. In the remainder of the paper, all analyses refer to the left hemisphere.

These results were further confirmed by the overlap analysis, which allowed us to further assess the spatial relationship between categories, with the underlying rationale that spatial proximity and overlap in the cortex suggest shared features (John Brendan Ritchie et al. 2024). We quantified the extent of activation overlap between each category by calculating an overlap index for each pairwise combination of regions, separately for the ventral and lateral OTC (see methods, Figure 4.2b). The index represents the number of voxels common between the areas, varying from 0 (no voxels in common) to 1 (the smaller area falls completely within the larger). In LOTC, from dorsal-posterior to ventral-anterior a large overlap could be observed between hands and bodies (0.68), between hands and tools (0.45), and between tools and manipulable objects (1.0, where manipulable objects fall completely within the larger tool cluster), but no overlap could be observed for the other combinations. On the contrary, in VOTC, no overlap could be observed between animate and inanimate categories, nor between faces and hands; inanimate objects, instead, presented a strong overlap with each other, with tools falling completely within the manipulable object cluster (1.0), and manipulable showing an extended overlap with non-manipulable objects (0.88), thus further confirming the opposite gradient in LOTC and VOTC for objects characterised by a different degree of action properties. A schematic visualization of category overlap is shown in Figure 4.2b.

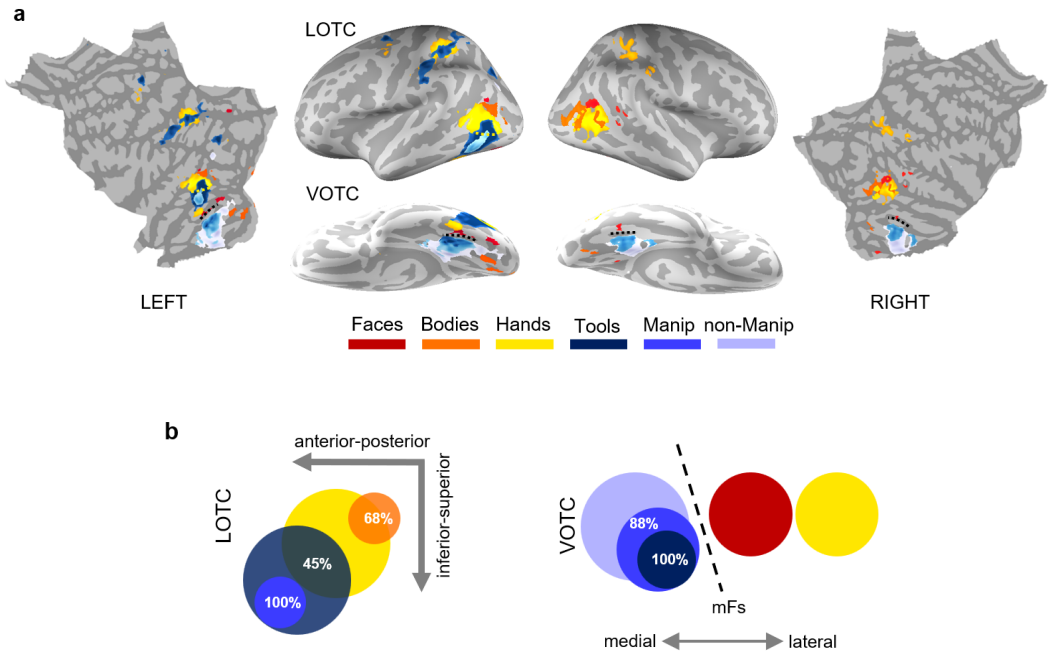


Figure 4.2: **Action-related topography of occipitotemporal cortex.** a) Whole-brain results. Response for each category (vs. all) was visualized on a freesurfer average brain surface using BrainSurfer (https://www.mathworks.com/matlabcentral/fileexchange/91485-brainsurfer), with a threshold of $t > 3.5$ ($p < 0.05$ FDR corrected at the cluster level), excluding activations within early visual cortex (approximately V1-V2-V3) to focus on the regions of interest in LOTC and VOTC. Color-coded dashed lines indicate overlap between activations. The black dashed line indicates the mid-fusiform sulcus. b) Category overlap visualization. The size of each circle represents the approximate size of the category-selective cluster in VOTC and LOTC in the left hemisphere.

4.3.2 Topographic DANNs successfully mimic animacy division in VOTC but fail to replicate action-based topography in LOTC[2]

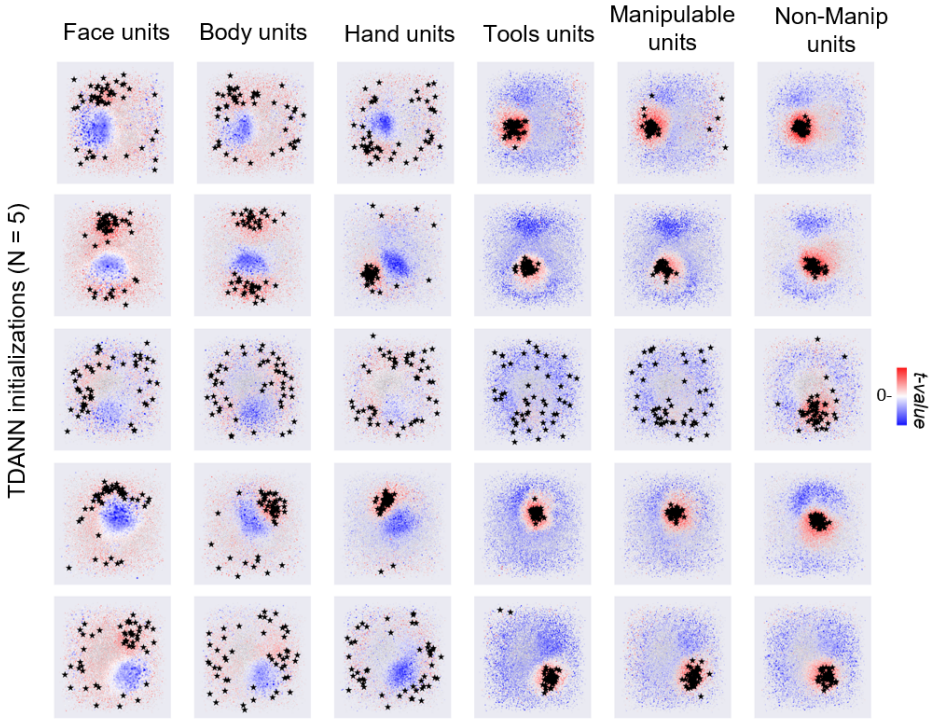


Figure 4.3: **Spatial distribution of each category** (as defined by t-values) on the simulated cortical space of the VTC-like layer of five random initializations of the TDANN. Rows correspond to each of the five initializations. Stars represent the location of the top-50 most selective units for that category. Category-selective units (positive t values) are shown in red while units not selective for that category (negative t values) are shown in blue.

The above results show that the lateral and ventral OTC are characterized by a different topographic organization: whereas in VOTC the animacy of objects strongly drives the organization of representations giving rise to the well-documented animacy division, in LOTC the topographic organization is driven by the degree of object action properties with a gradient from posterior-superior to anterior-inferior. Here, we test whether topographic deep artificial neural networks (TDANNs), a type of computational model developed

[2]The thesis's author developed these analyses.

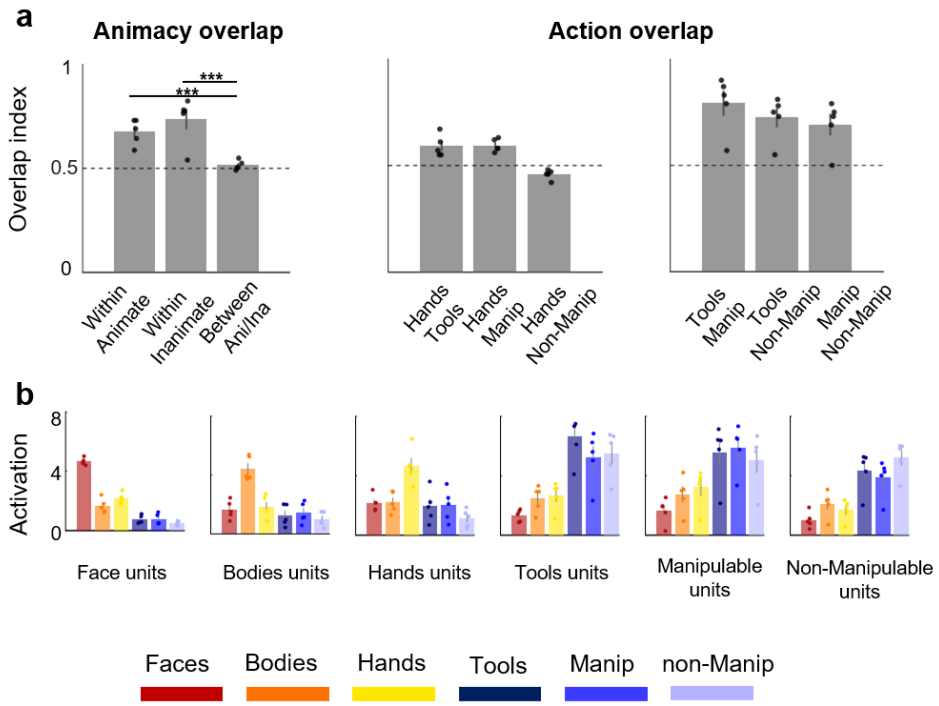


Figure 4.4: **TDANNs replicates animacy but not action-related organization of OTC.** a) Overlap analysis. Statistical significance was assessed using permutation tests (10,000 randomizations on the mean overlap score across initializations). Stars represent statistical significance at the minimum resolvable p-value (p = .0001), corresponding to the 10,000-permutation limit. Error bars correspond to ± 1 SEM across the random initializations. Black dashed line represents baseline (overlap of 0.5 means no correlation between the presence of two categories). Each data point represents the value from a single TDANN initialization (n = 5 model initializations). b) Selectivity profile of the top-50 most selective units for each category (red = faces; orange = bodies; yellow = hands; dark blue = tools; blue = manipulable objects; light blue = non-manipulable objects), based on the activation of the VTC-like layer (as in a). Each data point corresponds to one TDANN model initialization (n = 5 model initializations). Error bars indicate ± 1 SEM across model initializations. A baseline overlap of 0.5 denotes chance-level correspondence between category-selective units.

to capture the topographic organization of ventral visual cortex (Margalit et al. 2024), can mimic the action-related organization observed in lateral OTC. TDANNs allow testing whether a model designed to capture general topographic organization as a by-product of minimizing wiring length can account for object topography in visual cortex (Chklovskii, Schikorski, et al. 2002), thus suggesting that brain-like representations and their spatial organization can co-emerge with biologically inspired spatial constraints.

The network architecture was based on a ResNet-18 backbone, pre-trained with a self-supervised contrastive-learning object recognition task (T. Chen et al. 2020). We tested five different random initializations of the network's weights. We fed the networks with the images from our experiment and extracted the activation maps for each topographic layer, selecting the last VTC-like layer for further analyses (consistent with Margalit et al. 2024). A unit was defined as selective if its response for a specific category passed a set threshold (defined as $t > 3.5$, with a contrast of category > all). This uncorrected threshold was chosen for visualization purposes only (Figure 4.3). The subsequent functional selectivity analysis was performed on the first 50 most selective units. To investigate whether TDANNs replicate the topography and functional profile of category activations in visual cortex, we visualized their respective spatial distribution in the simulated cortical space and plotted the activation profiles for the 50 most selective units per category. Results are shown in Figure 4.3. Despite some variations between the five initializations – especially in the clustering's strength – two main findings could be observed (Figure 4.3): first, in all networks, units selective for animate and inanimate objects formed separate clusters, such that when a unit responded to a body-part it did not respond to an inanimate object and vice versa; second, no organization based on action properties was observed. Specifically, tools and hands did not activate the same units, and no smooth overlap based on action properties was found among the three object categories.

To quantify these observations and compare TDANNs with brain results, we performed an overlap analysis (as in Margalit et al. 2024). Specifically, we measured the co-occurrence of units selective for each category by using an overlap score ranging from 0 (the presence of one category always predicts the absence of the other) to 0.5 (no relationship) to 1 (perfect co-occurrence). Statistical significance was tested via 10,000 permutation tests. Results (Figure 4.4a) confirmed significant overlap within animate (score: 0.68, $p < .001$) and inanimate (score: 0.74, $p < .001$) categories relative to the between-category overlap (animate-inanimate, score: 0.51). In other words, units that responded to a body part or an inanimate object also responded significantly to other categories within the same superordinate class. Second, the overlap score between

action effector categories such as hands and tools (score: 0.59) was not significantly higher than the overlap between hands and other manipulable objects (score: 0.594, p = .37), as well as the overlap between tools and manipulable objects (score: 0.79) was not significantly larger relative to the overlap between tools and non-manipulable objects (score: 0.72, p = .24), nor relative to the overlap between manipulable and non-manipulable (score: 0.72, p = 0.33) thus, showing no action-related organization in TDANNs.

Visual exploration of Figure 4.3 suggests that, in addition to the separation between animate and inanimate categories, there seem to be additional differences in the organization of categories. Specifically, whereas the spatial distribution of units selective for the different body parts seem a bit scattered around, the inanimate objects mostly activated a similar portion of the cortical space. To investigate the functional profile of the TDANN units, we extracted the activation profiles for the 50 most selective units for each category and plotted the results (Figure 4.4b shows results averaged across the five initializations). Here, the focus was not on unit selectivity per se (e.g., do tool units respond to tools more than all other categories) but rather the degree to which a unit that responds to one category also responds to other categories (e.g., do tool units respond to other categories as well?). Overall, the results show that while a certain degree of category-selectivity could be found for the different body-parts, as different units selectively activated for each body part independently from the other body parts, the top-units for each inanimate object category responded to the other inanimate objects to a similar degree. Indeed, the selectivity of units chosen based on their response for faces, bodies, and hands was significantly higher for their preferred category compared to all other categories (for all contrasts, $p < .001$; permutation test n = 10,000). In contrast, units selected for their response to tools, manipulable and non-manipulable objects did not differ in selectivity across other inanimate object categories (for all contrasts, $p > .05$), while being more selective for their preferred category than for the animate categories (for all contrasts, $p < .001$; permutation test n = 10,000). Thus, similar to what we observed in visual cortex, TDANNs units that respond to one inanimate object category do also respond to the other inanimate object categories, but differently from human VTC, we did not observe any differential response gradient from high to low (tools > manipulable > non-manipulable) as observed in LOTC or from low to high (non-manipulable > manipulable > tools) as observed in VOTC. Finally, differently from visual cortex, units that respond to tools did not seem to activate hand units, thus confirming results from the TDANNs overlap analysis.

Overall, these results show that TDANNs primarily distinguish between

animate from inanimate objects, with additional functional selectivity for individual body-parts, and a weaker, or absent, distinction among inanimate object categories. These results mirror the pattern of overlap found in VOTC, which also showed a separation between animate and inanimate object categories, with further clustering for hands and faces. However, no action gradient organization, as found in LOTC, could be observed in TDANNs. Altogether, these analyses on networks implementing biologically-inspired topographic constraints reveal their ability to capture visual features important to distinguish animacy and to capture – to a certain extent – the selectivity for body-parts, but cannot replicate the action-related organization observed in visual cortex.

4.4 Discussion

Our study identifies action as a fundamental dimension shaping the topographic organization of the visual cortex. We demonstrate that the left lateral occipitotemporal cortex (LOTC) exhibits a dorsal-posterior to ventral-anterior gradient where body parts and inanimate objects are topographically organized based on their action-related properties. The combination of action effector and graspability contributes to explain the spatial organization of voxels that show a preferential response to bodies, hands, tools, and manipulable objects (Downing et al. 2001; Bracci, Ietswaart, et al. 2010; Pillet et al. 2024; Orlov et al. 2010; Chao et al. 1999; J. Almeida et al. 2023). While DANNs replicate aspects of ventral stream organization (e.g., animacy), they entirely lack the action-related topography observed in lateral OTC. Together, our results show that the action dimension is an important organizing principle of lateral OTC and highlight remaining gaps between biological and artificial systems.

Previous work emphasised how the combination of multiple object dimensions and principles may result in the topography-by-selectivity that is observed in high-level visual cortex (Grill-Spector and Weiner 2014; Arcaro and M. Livingstone 2024; Bracci and H. P. Op de Beeck 2023; Contier et al. 2024; Huth et al. 2012; Mahon and Caramazza 2011; H. P. O. d. Beeck, Pillet, et al. 2019; Peelen and Downing 2017; Prince et al. 2024; J Brendan Ritchie et al. 2025; Magri et al. 2020), with proposals stressing the role of shape, animacy, and real-world size1 (Konkle and Oliva 2012; Konkle and Caramazza 2013; H. P. O. d. Beeck, Torfs, et al. 2008), among others. Previous studies have already shown the relevance of action in explaining aspects of LOTC object space (Lingnau and Downing 2015; Kabulska et al. 2024; Tarhan and Konkle 2020; Tucciarelli et al. 2019). For example, overlapping responses in left LOTC between tools and hands, or tools and graspable food might reflect

shared end-effector properties and action-related affordances (Bracci, Cavina-Pratesi, et al. 2012; John Brendan Ritchie et al. 2024). Our results are in line with these previous findings and lift them up to a whole new level by revealing that a large-scale topographic organization is responsible for these earlier findings. More specifically, this approach enables us to move beyond post-hoc interpretations of visual cortex category organization (e.g., faces in lateral FG, tools in medial FG), allowing us to generate novel predictions about the spatial organization of new object categories – to be tested in future experiments – that share similar action-related features. Based on where these categories fall within a multidimensional feature space, we can predict their alignment within the topographic layout of OTC.

Furthermore, we demonstrate that lateral and ventral OTC represent different object features, with their topographic organization exhibiting opposing response patterns that depend on the degree of action properties associated with objects. In left LOTC, the action-based topography culminated at the intersection between animate (hands) and inanimate (tools) as both being end-effectors. Dorsally and posteriorly, hands overlap with bodies and inferiorly and anteriorly, tools overlap with manipulable objects which share with tools grasping properties but not end-effector properties. This organization is consistent across participants (even in unsmoothed, native surface) and cannot be explained by differences in object size or shape as tools and manipulable objects are matched for real-word size and all object categories are controlled for their overall shape. The opposite object pattern can be observed in VOTC, with higher and more extended activation for non-manipulable than manipulable objects, and tools being embedded within the manipulable object cluster in medial VOTC. These findings challenge views that tool representations in VOTC reflect action-related properties (Mahon, Milleville, et al. 2007), suggesting instead that they encode general object features – such as surface properties or weight (Cant and Goodale 2007; Gallivan et al. 2014) – shared across manipulable and non-manipulable objects to support recognition of inanimate objects in general rather than tools specifically (Cortinovis, Peelen, et al. 2025; Mahon and J. Almeida 2024).

DANNs results revealed both convergence and divergence with the functional and spatial organization of the visual cortex. Prior studies using topographic artificial neural networks or self-organizing maps have shown that principles like minimization of wiring length yield emergent macro- and mesoscale structures resembling those in visual cortex, including clusters for faces, bodies, scenes, and objects, and large-scale gradients of animacy and real-world size (Blauch et al. 2022; Z. Lu et al. 2025; Margalit et al. 2024; Cowell and Cottrell 2013; Doshi and Konkle 2023; Y. Zhang et al. 2024). Here, we confirm that

while these networks capture the large-scale clusters based on animacy, and to a certain extent the category clusters for faces, bodies and hands, they could not capture the action-based object topography and the category clusters for the three inanimate object categories.

This failure may stem from DANNs' reliance on mid-level visual features — such as shape and texture—that often correlate with object category in natural datasets. While this works well for animate categories (possibly because of curvature features; Long et al. 2018), it breaks down for inanimate categories when visual features are controlled, as in our study. In these cases, DANNs may default to encoding lower-level properties like orientation or aspect ratio, leading to weak category-specific clustering for inanimate objects. Thus, a tight control of visual features is especially important when comparing visual cortex and DANNs, as the two systems may represent objects in an apparent similar way but actually use different visual features that are confounded in the natural environment or uncontrolled stimulus sets (Bracci, Mraz, et al. 2023; Mahner et al. 2025).

The failure of these models in capturing action topography might stem from the fact that they are trained with static images that lack sensitivity to temporal dynamics, predictive processing, and temporal integration that humans naturally rely on Lake, Ullman, et al. (2017). Moreover, aside from motion, human action perception is shaped also by social context and affordances (Chartouny et al. 2024), factors that are entirely absent from current DANN models (Lake, Ullman, et al. 2017). For instance, the comparison between DANNs and visual cortex is especially revealing when considering the case of shape: while both systems are sensitive to aspects of shape, such as elongation and aspect-ratio, shape information might be used for different purposes: exclusively for categorization in DANNs, where shape is indicative of category membership, and for more varied behaviorally-relevant goals in the brain, such as grasping, manipulation, and functional use of objects. This divergence may arise because DANNs are trained on passive visual tasks (e.g., classification), whereas biological vision is inherently linked to action planning and sensorimotor experience. A promising direction may involve training models through reinforcement learning in embodied agents, where tasks are grounded in action. For example, agents could learn to evaluate an object's graspability or identify the specific parts relevant for grasping and functional use or learning actions in social contexts while interacting with humans (Chartouny et al. 2024; Yang et al. 2023). Overall, while TDANNs represent a step forward in modelling visual cortex organization, we point to the necessity of using more ecological, varied tasks – including beyond object or action classification – and the inclusion of biological constraints (**qian2024local**) to fully

model OTC spatial organization (but see Finzi et al. 2023).

In summary, this study demonstrates the critical role of the action dimension as an organizing principle of object representations in lateral occipitotemporal cortex. While artificial neural networks successfully replicated animacy-based organization, they failed to capture the action-based topography observed in the brain, despite their prominence in human functional organization. These findings underscore the importance of behaviorally relevant object properties in shaping the visual cortex's topography and advance our understanding of how multidimensional representations support object vision in the human brain.

Chapter 5

Beyond Topography: Topographic Regularization Improves Robustness and Reshapes Representations in Convolutional Neural Networks

Abstract Topographic convolutional neural networks (TCNNs) are computational models that can simulate aspects of the brain's spatial and functional organization. However, it is unclear whether and how different types of topographic regularization shape robustness, representational structure, and functional organization during end-to-end training. We address this question by comparing TCNNs trained with two local spatial losses applied to a penultimate-layer topographic grid: *i*) Weight Similarity (WS), whose objective penalizes differences between neighboring units' incoming weight vectors, and *ii*) Activation Similarity (AS), whose objective penalizes differences between neighboring units' activation patterns over stimuli. Compared to control non-topographic models, both regularizers tended to improve robustness against most types of input perturbations, and reduce activation-sparsity. WS, but not AS, consistently produced increased robustness to weight perturbations and signatures of functional localization. Together, these results show that local topographic regularization can improve robustness during end-to-end training while systematically reshaping representational structure.

Code `github.com/tlmnhut/weight_vs_act_toponet`

Publication status Under review (April 2026).

5.1 Introduction

Topographic models are computational models that provide an analogy to the cortical organization of the brain by instantiating a cortical-like map. In these models, each unit is assigned a position in a 2D grid (Poli et al. 2023), which defines a notion of distance between grid units. A topographic loss term can then be defined, which encourages spatially local similarity or smoothness in units' responses (or parameters) across the grid. It has been shown that when a task loss (e.g., cross-entropy) is optimized jointly with a topographic loss, both shallow (Jacobs and Jordan 1992) and deep (Blauch et al. 2022; Margalit et al. 2024; Z. Lu et al. 2025; Qian et al. 2026; Doshi and Konkle 2023; Deb et al. 2025; Y. Zhang et al. 2024; Rathi et al. 2024; Binhuraib et al. 2025; Al-Tahan et al. 2025) topographic networks can produce competitive task performance, while producing spatially organized responses that resemble cortical functional maps. For example, they produce angular and orientation preference such as those observed in early visual cortex, as well as category-selective clusters analogous to those reported in higher-level visual cortex (e.g., Blauch et al. 2022; Margalit et al. 2024; Z. Lu et al. 2025; Qian et al. 2026; Doshi and Konkle 2023; Deb et al. 2025; Y. Zhang et al. 2024). These models can also capture spatial biases in human behavior, recognizing objects more accurately when they appear in locations where they are most frequently experienced (Z. Lu et al. 2025). Outside the vision domain, topographic modeling has reproduced tonotopic organization in auditory cortex (Al-Tahan et al. 2025), and higher-level language organization e.g., Rathi et al. 2024; Binhuraib et al. 2025. Topographic networks can also be more strongly pruned, suggesting a sparser distribution of weight values (Poli et al. 2023; Deb et al. 2025).

The correlated activation patterns produced by topographic models carry parallels to brain activity, where nearby neurons often fire together, with correlations sometimes exceeding $r = 0.4$ in certain cortical circuits (e.g., Zohary et al. 1994; Hansen et al. 2012). Although such correlations reduce the degrees of freedom and representational capacity in biological neural populations (Zohary et al. 1994), they may also offer benefits such as robustness, as redundant neurons can compensate for one another (Harris and Mrsic-Flogel 2013). A similar trade-off has been discussed for computational models: on the one hand correlated units encode overlapping information and can impair decoding performance (Abbott and Dayan 1999), but on the other, moderate correlations can act as a form of regularization, producing more compact representations or giving priority to more informative features (Poli et al. 2023). In summary,

correlations constrain capacity but may provide computational advantages.

While topographic models are a topic of emerging interest in computational and systems neuroscience (Margalit et al. 2024; Rathi et al. 2024; Z. Lu et al. 2025; Lee et al. 2020; Blauch et al. 2022; Doshi and Konkle 2023; Y. Zhang et al. 2024; Krug et al. 2023; Hannagan 2021; Keller et al. 2021; Qian et al. 2026; D. Jiang et al. 2024; Deb et al. 2025; Dehghani et al. 2024; Finzi et al. 2023; Achterberg et al. 2023; Binhuraib et al. 2025; D. Zhou et al. 2025; Bashivan et al. 2025; Kamila Maria Jozwik et al. 2023; Mehrer, Spoerer, Kriegeskorte, et al. 2020; Al-Tahan et al. 2025; Cortinovis, Truong, et al. 2025), there is a substantial gap in our understanding of the computational consequences of imposing topographic regularization. This is an important open question not only for computational neuroscience, where topographic networks are mainly studied for their ability to reproduce brain-like spatial organization of responses, but also for machine learning, where the impacts of these locally induced correlations are poorly understood.

Going beyond a focus on spatial organization, we therefore ask whether these correlations provide any computational advantages, such as improved robustness to noise, and how they, in turn, shape the representations learned by topographic networks. To our knowledge, neither of these issues has been systematically examined in controlled comparisons to date. Specifically, we address these questions using two related criteria:

1. **Network robustness**: We introduce internal noise by perturbing the readout weight matrix connecting the penultimate (topographic) layer to the classification layer (Cheney et al. 2017; Arechiga and Michaels 2018; Savva et al. 2023), and external noise by corrupting input images at test. We test how resilient topographic models are compared to non-topographic models in terms of both representational stability and classification accuracy .

2. **Representational properties**: We compare topographic models to size-matched non-topographic control models; we quantify unit-level sparsity and entropy, similarity of weights and activations, the dimensionality of the latent space, and the tendency of models to produce category-selective "expert units". We further characterize the spatial organization of representations by measuring the smoothness of topographic maps and the extent to which activity is functionally localized (i.e., whether similarly responding units are spatially clustered). Finally, we test whether topographic regularization changes functional organization by quantifying orientation, eccentricity, and angular tuning in topographic and non-topographic networks.

To understand the computational effects of topographic regularization, we

use local spatial losses and end-to-end training. Prior work has often relied on global spatial losses that include a distance-dependent term over all unit pairs. These global losses directly encourage high activation similarity between nearby units as well as low activation similarity between distant ones (e.g., Poli et al. 2023; Margalit et al. 2024). They therefore effectively *impose* functional localization by suppressing long-range correlations. This makes localized clusters an outcome of the training objective rather than an emergent property. We therefore restrict the regularizer to local interactions only (e,g., Z. Lu et al. 2025). In addition, we train topographic networks end-to-end, because post hoc projection methods operating on frozen pretrained features (e.g., self-organized mappings, Doshi and Konkle 2023; D. Jiang et al. 2024; Krug et al. 2023; Y. Zhang et al. 2024) cannot address whether topographic regularization impacts feature learning itself.

Regarding the technical implementation, correlated activations in topographic models are often produced using an objective that encourages similarity between neighboring units' activation patterns (e.g., Lee et al. 2020; Margalit et al. 2024; Rathi et al. 2024; Poli et al. 2023). An alternative is to impose similarity constraints on incoming weights (Z. Lu et al. 2025), which may be more biologically plausible: correlated activity should emerge due to shared afferent connectivity instead of being directly enforced at the level of activations. For this reason we also avoid explicit lateral connections, as correlations in the brain often reflect shared afferent input (Shadlen and Newsome 1998). To directly contrast these two positions we compare two local topographic losses (Figure 5.1): **Activation Similarity (AS)**, whose objective is to produce correlated activations in adjacent units, and **Weight Similarity (WS)**, whose objective is to produce similar afferent weight vectors in adjacent units. In both cases, we consider only local neighbors.

We find that AS and WS training produced qualitatively different computational outcomes. Compared with AS and control (non-topographic) models, WS training produced representations that are more robust to perturbations of the readout weight matrix and stronger functional localization, with similarly responding units positioned at closer spatial distances. AS training produced representations that were more robust to degraded inputs, showed weaker functional localization than WS, and did not necessarily form smoother topographic maps as spatial smoothness strength increased.

As compared to control models, AS and WS produced different emphases on angular, orientation and symmetry tuning. Interestingly this was not necessarily accompanied by a reduced effective dimensionality of the network activation, as for the lowest level of the topographic regularizer, the overall effective dimensionality of topographic models could match or exceed those of the con-

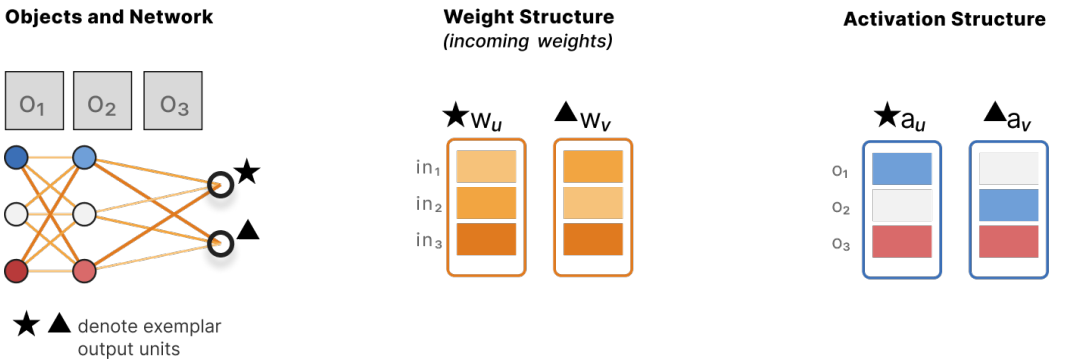


Figure 5.1: **Overview of main concepts**. Objects (O_1–O_3) are passed to a neural network. Two example output units ($\star = u$ and $\blacktriangle = v$) are indicated. For each example unit, an *activation vector* $(\mathbf{a}_u, \mathbf{a}_v)$ summarizes that unit's responses across objects, and a *weight vector* $(\mathbf{w}_u, \mathbf{w}_v)$ summarizes the incoming weights from the preceding layer. Activation similarity is computed by correlating activation vectors across objects. Weight similarity is computed as a distance between weight vectors.

trol models. Taken together, this suggests that topographic regularization does not simply reorganize the weight matrix to maintain classification, but produces a markedly different feature space.

5.2 Related work

5.2.1 Topographic regularization in artificial neural networks (ANNs)

Several studies have examined how spatially organized response patterns, often accompanied by increased local correlations, can be induced in ANNs by applying topographic objectives. One approach is end-to-end topographic training, which introduces a topographic loss term directly into the training objective. Typically, a joint loss combines a task loss (e.g., cross-entropy for classification) with an additional topographic spatial term. Early work of Jacobs and Jordan (1992) introduced a spatial loss in small-scale, fully connected networks, penalizing the weight magnitude proportionally to the physical distances between the pre- and post-synaptic units' assigned positions. As a result, adjacent units showed similar tuning. Others (e.g., Poli et al. 2023) introduced a loss that enforces an inverse relationship between the activation similarity of spatially arranged convolutional filters and their distances. This improved robustness

to pruning while maintaining task performance. Related approaches that produce cortical-like spatial organization include penalizing connections between spatially-remote units (Blauch et al. 2022), introducing lateral pooling between neighboring filters (Qian et al. 2026), and adding topographic structure to variational autoencoders (Keller et al. 2021). Topography can also emerge without an explicit spatial loss when kernels in deeper layers are defined as averages of partially overlapping kernels from a preceding layer, producing smooth transitions across kernel maps (Bashivan et al. 2025).

Other work used a hybrid approach Margalit et al. (2024), in which unit locations are optimized based on activation similarity computed from a pretrained model and then held these locations fixed during subsequent training. This model reproduced spatial properties of both early and high-level visual cortical features and outperformed standard networks on biological benchmarks. Similarly, Z. Lu et al. (2025) trained a fully topographic model by maximizing weight vector similarity between immediate neighboring units, and showed that the resulting networks capture spatial biases including center-periphery preferences.

Another class of topographic models is based on post hoc projection of pretrained representations, for example, via self-organizing maps (SOMs). In these approaches, a fixed, pretrained feature space is mapped onto a 2D grid with the objective of producing spatial clusters of similarly-responding units (Doshi and Konkle 2023; D. Jiang et al. 2024; Krug et al. 2023; Y. Zhang et al. 2024). Conceptually, these methods learn a spatial embedding of *existing* representations: they operate on frozen features and therefore cannot address how topographic regularization shapes feature learning during end-to-end task training (Tyson N. Aflalo and M. S. A. Graziano 2006).

5.2.2 Impact of topography on representational structure

Beyond producing correlated responses, topographic regularization also influences the representational structure of ANNs. Several studies have reported that such regularization reduces the effective dimensionality of latent representations (Deb et al. 2025; Margalit et al. 2024; Qian et al. 2026) and improves robustness to pruning (Poli et al. 2023; Deb et al. 2025).

At the same time, inter-unit correlations are sometimes considered detrimental in machine learning research. For example, the Barlow Twins architecture (Zbontar et al. 2021) maximizes agreement between paired views (original and distorted image) while penalizing correlated activity across units as computed from the off-diagonal terms of the batch-wise cross-correlation matrix.

The study demonstrates that reducing correlations improves classification accuracy. In addition similarity between incoming weight vectors is often considered a negative computational property, as minimizing such correlations improves classification accuracy (Cogswell et al. 2015; Rodríguez et al. 2016; G. Jin et al. 2020; Z. Wang et al. 2020). Reducing redundancy across convolutional filters has been reported to produce similar benefits (H. Zhang et al. 2025).

5.3 Methods

5.3.1 Models and datasets

MNIST. The model used for MNIST (LeCun 1998) training was a relatively shallow CNN. It consisted of two convolutional layers: the first with 32 output channels and the second with 64 channels, each using a 3×3 kernel. Both convolutional layers were followed by a ReLU activation function and 2×2 max-pooling. After the second convolutional layer (`conv2`), global average pooling was applied to each of the 64 feature maps, producing a 64-dimensional feature vector for each input image. This vector was then fed into a fully connected layer, `fc1`, which mapped the 64-dimensional vector to 121 units. The output of `fc1` was connected to a second fully connected layer, `fc2`, which produced the final 10 logits for classification, corresponding to the 10 MNIST classes. To reduce overfitting, dropout with a rate of 0.5 was applied after `fc1`.

CIFAR-10. We used the standard CIFAR-10, a 10-class dataset of 32×32 images (Krizhevsky, Hinton, et al. 2009). The CNN model used for image classification consisted of four convolutional layers with batch normalization applied after each. The number of channels in the first three layers were: 32, 64, and 128, each followed by max-pooling *stride* $= 2$. The fourth convolutional layer consisted of 256 channels and was followed by a global average pooling layer ($n = 256$ values). The last two layers were two fully connected layers. The first (`fc1`) consisted of 121 units, with dropout (0.3), and the second (`fc2`) mapped feature activations to the 10 output classes.

The exact same model definitions were used for the topographic models and the control models, for CIFAR-10 and MNIST. The only difference was that in the topographic models, the 121 `fc1` units were arranged on a 11×11 grid to which a spatial loss function could be applied as described below.

5.3.2 Spatial loss: weight-similarity and activation-similarity

Weight similarity. For training with a weight-similarity objectives, we used a joint loss function, combining the standard cross-entropy loss term, $\mathcal{L}_{\text{CE}} = \text{cross-entropy}(\text{output}, \text{target})$, and a spatial loss term $\mathcal{L}_{\text{spatial}}$. For weight-similarity, the spatial loss term was designed to increase similarity among immediately adjacent weight vectors in the 11×11 grid structure. To compute $\mathcal{L}_{\text{spatial}}$, we indexed the 121 incoming weight vectors of `fc1` on an 11×11 grid. Each grid position corresponds to one `fc1` unit with an incoming weight vector of dimension 64 (MNIST) or 256 (CIFAR-10). For each of the 121 grid cells, the immediate neighbors were identified. Then, for each cell, the L_2 norm (Euclidean distance) was computed between the weight vector of that cell and those of each neighboring cell. These distances were summed across all cells and divided by the number of neighboring cells, producing a single value indicating the average pairwise distance across the grid. The joint loss function was therefore $\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{spatial}}$, where λ is a weighting factor that sets the contribution of the spatial loss. We evaluated the impact of the spatial loss term under six weighting levels, with $\lambda \in \{0.1, 0.3, 0.5, 1, 2, 3\}$.

Activation similarity. The activation-similarity (AS) spatial loss term produced a single value indicating the similarity of each unit to its immediate neighbors. For AS, similarity was defined as $1 - r$, where r is the Pearson's correlation between the two units' activation vectors across minibatch samples. Here too `fc1` was the topographic layer, defined by reshaping 121 units to 11×11 grid.

Training parameters. MNIST models were trained using the Adam optimizer with a learning rate of $\eta = 0.001$ for 15 epochs. Initial evaluations showed that all models converged to a training-set accuracy of approximately 97% under moderate regularization strength. The CIFAR-10 model was trained using the same optimizer parameters for 30 epochs, reaching similar training accuracy of around 93% across the three model conditions.

We trained 10 independently initialized models for each λ level under both WS and AS regularization, resulting in 60 WS models and 60 AS models in total. Additionally, we trained 10 control models without topographic regularization, which are equivalent to $\lambda = 0$. In all analyses, mean statistics were computed from each set of 10 models from the same λ.

5.3.3 Robustness tests

Robustness of representational geometry. After training the control, AS and WS models on MNIST and CIFAR-10, we extracted the weight matrix connecting the topographic penultimate layer to the classification layer (a 10×121 matrix in both cases). We consider each of this matrix's 10 row vectors as a *class weight vector; (CWVs)*. From these CWVs, we computed a class-weight representational similarity matrix (RSM), which is a 10×10 matrix of pairwise similarity between CWVs (Lake, Zaremba, et al. 2015; Nayak et al. 2019; Filus and Domańska 2024; Filus and Domańska 2025). This was treated as the baseline representational geometry.

To evaluate the robustness of this representation, we conducted several perturbation tests in which Gaussian noise (four intensity levels) was added to the 10×121 class-weight matrix, and the RSM was recomputed. We refer to these as perturbed RSMs. Each analysis was repeated 100 times, and we report the mean results.

We determined the impact of noise using two metrics. First, we computed the second-order similarity as the cosine similarity between the upper triangles of the unperturbed and perturbed RSMs. This indicates how much the representational geometry was impacted by noise. This was computed separately for the WS, AS, and control models. Second, we evaluated the drop in classification accuracy caused by the addition of noise, relative to the unperturbed models.

Robustness to input corruption. For both WS and AS models, we evaluated their performance under various corruptions, with the noise applied to test-set images (training images were not corrupted). In all cases, corrupted images were normalized to the mean and standard deviation of MNIST and CIFAR-10 training sets. The noise interventions consisted of adding white noise, pink noise, and salt-and-pepper noise. White noise was introduced by adding to each pixel a random value from the standard normal distribution. Pink (1/f) noise was generated such that the power spectral density of the signal is inversely proportional to its frequency. Salt-and-pepper noise converted a proportion of randomly chosen image pixels to either black or white. Examples of each type are given in Supplementary Figure 21. Each noise intervention was applied at five different intensities.

5.3.4 Orientation and eccentricity tuning

After training MNIST and CIFAR-10, we evaluated the responses of the trained models on a standard stimulus set typically used for retinotopic mapping of

human occipital cortex. To study angular and orientation tuning we presented the trained networks with a rotating wedge (36 positions, angle extent 10°, radius = 14 pixels). To study eccentricity tuning, we presented the network with ring images (13 different radius levels). For each unit in the grid this produced a 36-element series for the wedge angle and a 13-element series for ring eccentricity.

Orientation analysis. To describe angular tuning in topographic units, for each unit we measured responses to the 36 wedge stimuli. We applied a Fast Fourier Transform to each unit's 36-point response profile and extracted the spectral power for the first five harmonics (cycles 1–5). These components reflect different angular periodicities: cycle 1 indicates preference for a single direction, cycle 2 for 180° symmetry consistent with orientation tuning, and cycle 4 indicates fourfold (90°) angular periodicity. We defined the *dominant harmonic* for each unit as the harmonic (cycle 1-5) with the largest spectral power (excluding the DC component; i.e., the mean value of the signal). These dominant harmonic labels were assigned to the 11×11 grid.

Eccentricity analysis. To study eccentricity response profiles, we analyzed each unit's response to the 13 ring stimuli of increasing eccentricity by fitting a linear model to the 13 response values. Units with a Pearson correlation coefficient of $|r| > 0.8$, were labeled as `increasing` or `decreasing` depending on the sign of r. For units not showing a linear profile, we evaluated if the response was selective to a particular eccentricity, indicating a band-pass response. To test this, we fitted a four-parameter Gaussian function (baseline, amplitude, center, width) to the unit's 13-point response profile. The quality of fit was determined using the coefficient of determination (R^2), and a unit was categorized as showing a `bandpass` response when $R^2 > 0.5$. Profiles that did not meet any of the above criteria were labeled `flat`.

5.3.5 Functional localization

For each trained model, we measured localization to understand how closely located were correlated units. We first computed the correlation between every pair of units' activations across the test images. Given a chosen correlation threshold, we marked a pair of units as connected if their correlation exceeded that threshold, producing a binary connectivity matrix (connected vs. not connected). Next, for each unit, we identified all other connected units and computed the mean Euclidean distance to those. In a last step, we averaged these per-unit distance values across all units in the grid to obtain a single

localization score for the model at that threshold. Smaller values indicate that strongly correlated units are more spatially clustered.

5.3.6 Weight correlations and activation correlations

After WS and AS training, we computed, for each unit in the 11×11 grid, its average incoming-weight correlation with neighboring units. Incoming weights refer to the weight matrix from the preceding global-average pooling layer to the 121-unit grid layer. The correlation $r_{ij}^{(\mathrm{In})}$ between the incoming weight vectors of units i and j was the Pearson correlation of the two vectors:

$$r_{ij}^{(\mathrm{In})} = \frac{\sum_k (w_{ik} - \bar{w}_i)(w_{jk} - \bar{w}_j)}{\sqrt{\sum_k (w_{ik} - \bar{w}_i)^2 \ \sum_k (w_{jk} - \bar{w}_j)^2}}\,, \tag{5.1}$$

where w_{ik} and w_{jk} denote the k-th afferent weight into units i and j, respectively, and $\bar{w}_i$ and $\bar{w}_j$ are the means of their incoming weight vectors (dimension 64 for MNIST; 256 for CIFAR-10). For each unit i, we then computed the mean correlation over its immediate (Moore) neighborhood S_i. The neighborhood size $|S_i|$ was 3, 5, or 8 for corner, edge, and interior units, respectively:

$$R_i^{(\mathrm{In})} = \frac{1}{|S_i|} \sum_{j \in S_i} r_{ij}^{(\mathrm{In})} \tag{5.2}$$

We evaluated activation correlations using the same logic described above for computing of incoming weight correlations. The difference was that for each unit-pair, we computed the correlation of their activation profiles for a minibatch of images. This allowed computing, for each unit, its average neighbourhood correlation. It also allowed studying the entire distribution of pairwise correlations.

Both the AS and WS spatial loss terms operate locally, by encouraging each unit to be similar to its immediate neighbors in the grid. This differs from global objectives Poli et al. (2023), which use a spatial loss that encourages activation similarity to decrease with grid distance across *all* unit pairs. Global objectives penalize cases where highly correlated units are positioned far apart. They therefore discourage the formation of multiple, disconnected clusters of similarly-responding units, and instead produces spatially contiguous regions of response-similar units.

5.3.7 Expert unit analysis

Following prior work e.g., Fedzechkina et al. 2025, we define expert (class-selective) units whose activations reflect one-vs-rest discrimination of a target

class. For each unit in the 11×11 topographic grid layer and each class, we quantified discriminability using the area under the receiver operating characteristic curve (AUC), computed from the unit's post-ReLU activations over the test set. The AUC estimates the probability that the unit's activation for an input from the target class exceeds activations for inputs from any other class.

We use two AUC thresholds. Units with $AUC > 0.70$ were classified as moderately class-selective; they produce higher activation to the target class in more than 70% of random class–nonclass comparisons. Units with $AUC > 0.90$ were classified as strongly class-selective. For each condition, we counted the number of units meeting these criteria, and normalized by the total number of units ($n = 121$) to obtain a proportion. Condition-level values were obtained by averaging the corresponding measures across the ten independently trained models for each condition.

To quantify the distribution of expert units over classes, each expert unit was assigned to the class for which it achieved its maximal AUC. This produced a distribution of expert units for each model family. We then computed the entropy of this distribution and normalized it by the maximum entropy for ten classes ($log10$). We refer to this quantity as *expertise balance*; it equals 1.0 when expert units are uniformly distributed across classes and lower otherwise.

Finally, we quantified the selectivity of expert units by computing, for each unit, the difference between its highest AUC (corresponding to the most discriminative class) and its second-highest AUC across classes. This AUC gap shows how exclusively a unit discriminates its preferred class relative to alternative classes. We refer to this as *expertise selectivity*.

5.4 Results

5.4.1 Accuracy

Control models were generally associated with the highest accuracy, followed by AS. For MNIST, WS slightly outperformed the control models at the lowest lambda level, while AS outperformed at $\lambda = 0.1 - 0.5$ (Figure 5.2). For CIFAR-10, accuracy of the topographic models was often lower than that of controlled ones. Stronger topographic regularization produced a drop in accuracy up to 2%, and more strongly for WS. This drop in accuracy was also observed in previous end-to-end topographic models (Margalit et al. 2024; Z. Lu et al. 2025; Rathi et al. 2024).

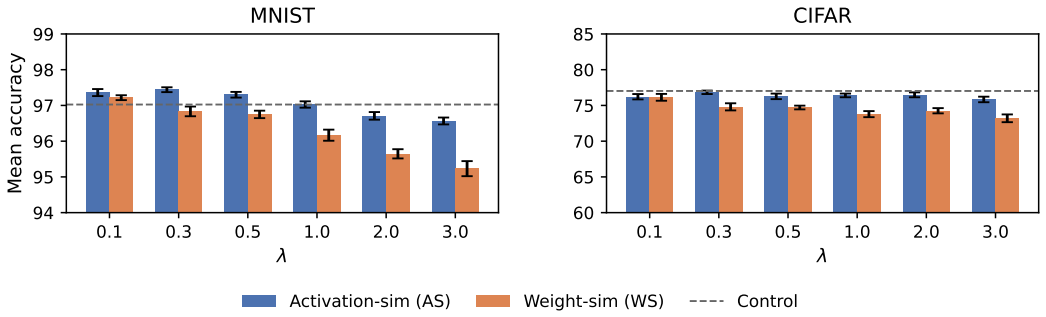


Figure 5.2: **Test accuracy as a function of topographic regularization strength** λ. Mean classification accuracy for control (dashed), AS (blue), and WS (orange) models is shown for the different values of λ. Error bars indicate $\pm$ s.e.m.

5.4.2 Robustness

Robustness of representational geometry. We added Gaussian noise of varying magnitudes to the weight matrix connecting the topographic and classification layers, then evaluated the resulting representation by constructing a representational similarity matrix (RSM) from the class-weight vectors. The intervention showed that WS training resulted in a more robust representational geometry than for AS and control models (see Figure 5.3). First, the second-order similarity between the baseline and perturbed RSMs was consistently higher for WS models, for both MNIST and CIFAR-10. WS models, particularly when trained with larger λ, were consistently top-ranked, showing less degradation in representational geometry. The control models showed the least robustness.

Robustness of representation was also evident in the accuracy drop statistics: for MNIST, WS models showed a relatively small drop of 0–5% point, but AS models showed larger drops of 5–20%. A similar pattern was observed for CIFAR-10, where the accuracy drop for AS was, on average, more than that of WS. In both datasets, WS outperformed the control models as well. These results suggest that WS training stabilizes representations under perturbation, which in turn also better maintain classification performance.

Robustness to image corruption Figure 5.4 shows, for each level and type of noise, which model family (AS, WS, control) was most robust to input corruption. For both datasets, AS tended to be the most robust. (Figure 22 presents the full breakdown, presenting accuracy changes compared to baselines by noise type, noise level and λ.)

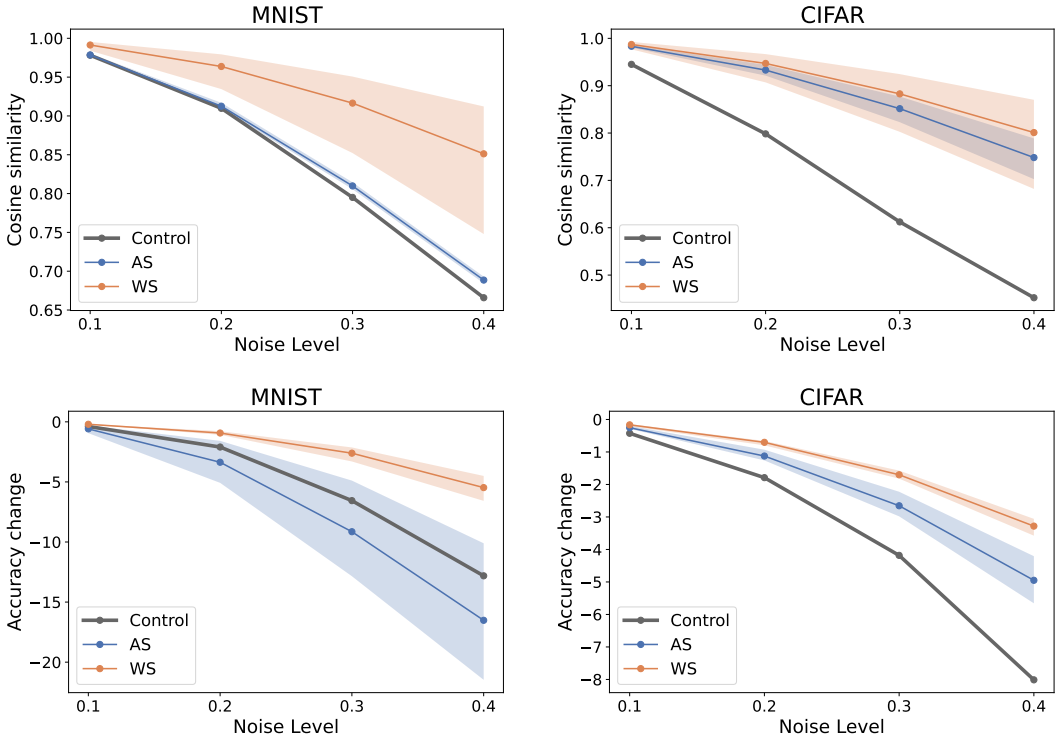


Figure 5.3: **Robustness to weight perturbations as a function of noise levels.** Robustness is evaluated by changes in representational geometry (top row) and test accuracy (bottom row) for increasing levels of additive weight noise. Representational geometry is defined as the representational similarity matrix (RSM) computed from class weight vectors (CWVs) in the readout layer. Cosine similarity between original and perturbed representations is shown in the top row; corresponding changes in test accuracy relative to baseline (non-perturbed) models are shown in the bottom row. Shaded regions indicate min–max range across λ levels. Values computed separately for each λ level and model are reported in Supplementary Figure 20.

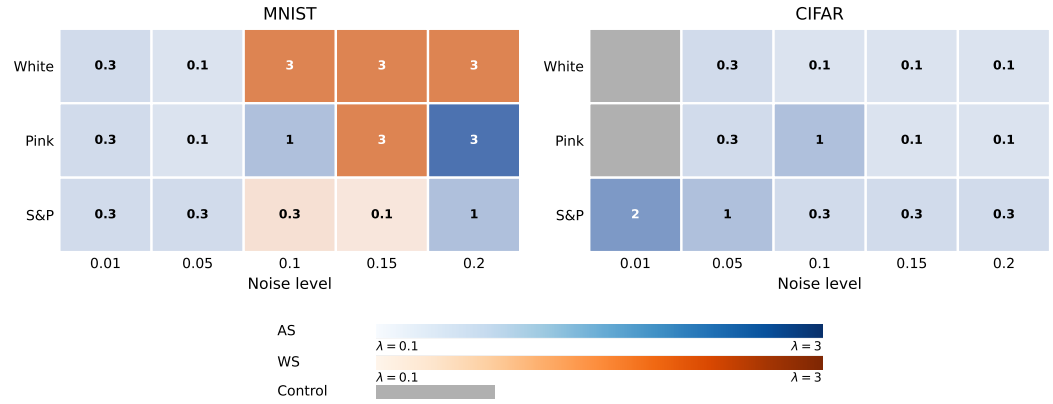


Figure 5.4: **Model-group robustness under input corruptions.** For each dataset and corruption type (white, pink, and salt-and-pepper), the most robust model group is defined as the one showing the smallest drop in test accuracy relative to the uncorrupted baseline. Each cell reports the λ value associated with the winning group. A full breakdown of accuracy drops across λ and corruption intensities is reported in Supplementary Figure 22.

It should be noted that at the lowest regularization strength ($\lambda = 0.1$) in MNIST, AS and WS showed stronger robustness to input corruption and stronger representational robustness than control models, but similar classification accuracy. Prior studies have tried training for robustness, finding this typically reduces accuracy (Hendrycks and Dietterich 2019; Lopes et al. 2019; Tsipras et al. 2018; Su et al. 2018). Our findings suggest that moderate topographic regularization can achieve a similar aim without a strong trade-off.

5.4.3 Activation entropy, sparsity and effective dimensionality

The variance or entropy of a unit's activations is often treated as an indicator of its functional importance within an ANN. Higher entropy reflects greater sensitivity to input variation, and lower-entropy elements are targets for pruning (Polyak and Wolf 2015; J. Wang et al. 2021; J.-H. Luo and J. Wu 2017; Liao et al. 2023; Z. Lu et al. 2025). Similarly, a unit's tendency to remain inactive across inputs, quantified as PoZ (Percentage of Zeros), is used as a pruning criterion, with higher PoZ indicating less functional importance (Huan Hu et al. 2016). We analyzed the entropy and PoZ of the grid units in AS, WS, and control models. Entropy was computed from each unit's pre-ReLU activations across the 10,000 images test images. PoZ was computed post-

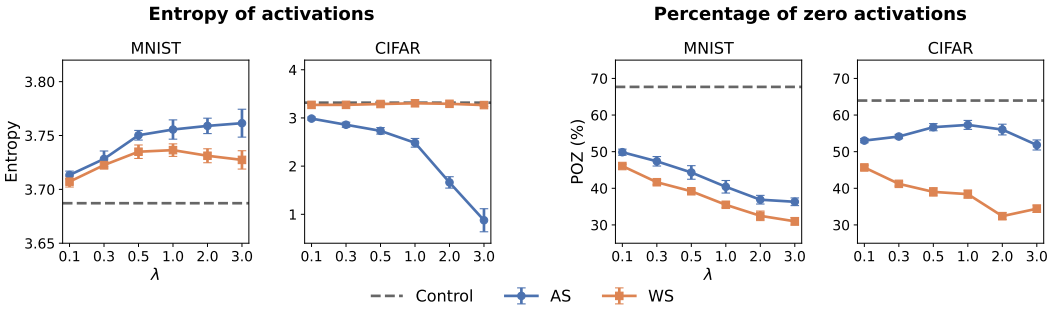


Figure 5.5: **Unit-level entropy and Percentage-of-Zero activations as a function of topographic regularization strength (λ).** The two left plots show average entropy of pre-ReLU unit activations for MNIST and CIFAR-10 across values of λ. The two right plots show the average Percentage-of-Zero (PoZ) of post-ReLU unit activations. Error bars indicate ± s.e.m.

ReLU, reflecting the fraction of inputs for which a unit's activation was zero. We computed the average entropy and PoZ across units within each grid. As shown in Figure 5.5, for MNIST, WS models present slightly lower entropy than AS models, but for CIFAR-10 they are significantly higher. Regarding PoZ, WS models show the lowest across all λ in both datasets, and both AS and WS showed their PoZ far below that of control. This suggests that incorporating similarity in either weights or activations during training leads to less sparse models.

To determine how these unit-level properties translate into the organization of information in the latent space, we examined the Effective Dimensionality (ED) of model activations (e.g, Margalit et al. 2024; Qian et al. 2026; Deb et al. 2025). Using the activations evoked by the 10,000 test images, we computed ED from the covariance matrix of all images to estimate the overall dimensionality of the representation. In addition, we computed within-class ED by isolating activations for each class and averaging ED across classes. We find that AS and WS models produced different effects on the dimensionality of latent representations (Figure 5.6). For both datasets, AS models were associated with higher overall effective dimensionality than WS and control models, meaning that information resided in a larger number of latent dimensions. These results suggest that WS training produces a more redundant, low-dimensional class representations. As expected, increasing λ led to a gradual reduction in both overall and within-class effective dimensionality, though the trend was stronger for WS. We also note that for CIFAR-10, WS produced higher overall effective dimensionality than control models at the lower λ levels. This shows that stronger local correlations do not necessarily

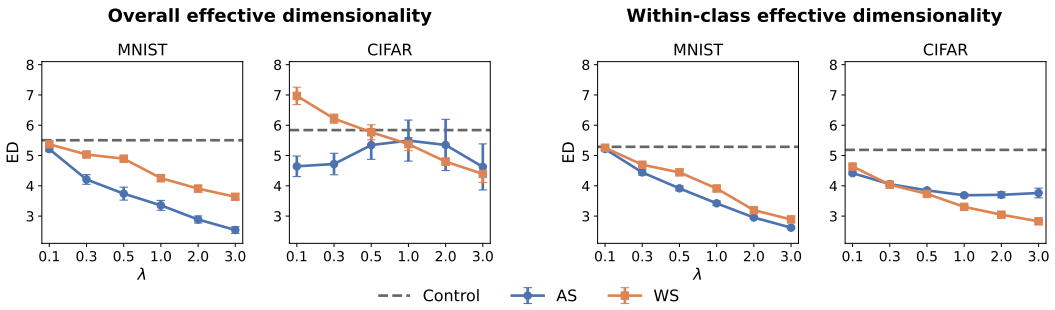


Figure 5.6: **Effective dimensionality of topographic representations**. Overall effective dimensionality (ED; left) and within-class effective dimensionality (right) of the fc1 layer as a function of the spatial regularization strength λ. Overall ED is computed from the activation covariance across all images; within-class ED is computed for each class separately and then averaged across classes. Error bars denote s.e.m, and dashed lines indicate control models.

result in reduced dimensionality.

5.4.4 Functional localization metrics

Functional co-localization

To evaluate to what extent units with similar firing patterns were positioned closely on the topographic grid, we defined two units as belonging to the same functional network if their activation patterns exceeded a correlation threshold α. We then computed the average Euclidean distance between connected units (see Methods). Figure 5.7 shows the results for WS and AS (for simplicity, the figure presents the mean and $\pm 1SD$ over responses of α; details are presented in Supplementary Figure 23). As the figure shows, for both MNIST and CIFAR-10, WS was associated with smaller distances between correlated units.

Spatial autocorrelation of unit activations

To understand the spatial organization produced by WS and AS training, we quantify activation smoothness, weight similarity, and activation correlations among neighboring and non-neighboring units on the topographic grid. We first quantified the spatial smoothness of activation maps using Moran's I, which is a spatial smoothness statistic Moran 1950, previously used in related work e.g., Rathi et al. 2024. Positive values indicate smoother transitions among neighboring units, negative values indicate spatial dispersion (e.g., high-

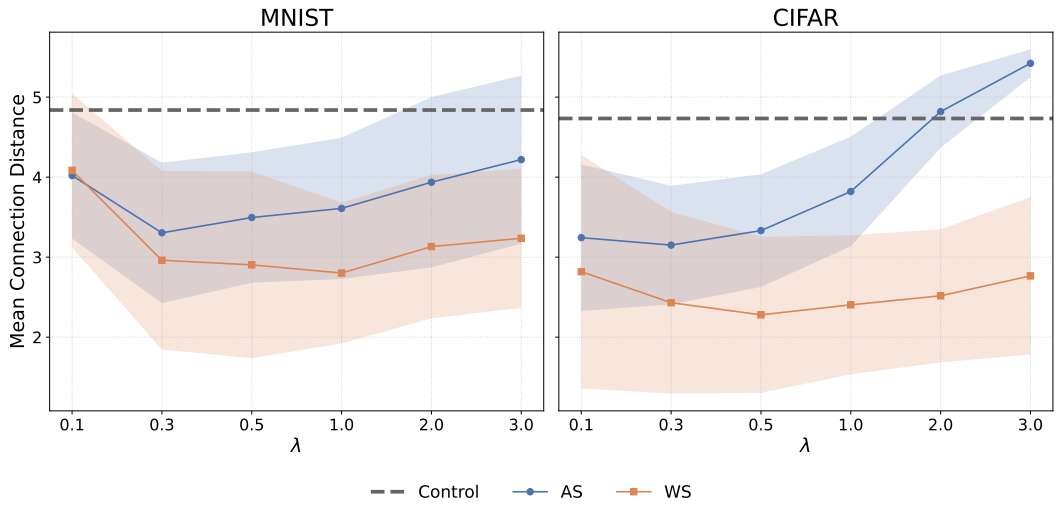


Figure 5.7: **Correlated-unit distance in the topographic grid as a function of topographic regularization strength (λ).** Mean spatial distance between pairs of units whose activity-correlation exceeds a threshold $\alpha \in 0.1, 0.3, 0.5, 0.6, 0.7, 0.8$. Solid lines indicate distances averaged across correlation thresholds α; shaded regions denote ± 1 SD across α thresholds. Distances computed separately for each correlation threshold and model are reported in Supplementary Figure 23.

low activation transitions between neighbors), and zero indicates randomness. We applied this metric to the pre-ReLU activation maps produced by each image in each model, and averaged within AS, WS, and control models. Figure 5.8 presents the results for MNIST, and similar findings were found for CIFAR-10 (see Supplementary Figure 24).

As Figure 5.8C shows, WS models produced consistently positive Moran's I values, which increased monotonically with λ. However, this monotonic trend does not hold for AS models, which shows that stronger spatial regularizations are not necessarily lead to smoother topographic maps. The control models approximated zero as expected. To visually appreciate the distribution of local activation similarity, Figure 5.8A shows activation maps from three randomly selected WS models trained under regularization strength of $\lambda = 0.1, 0.3, 0.5$. It is evident that both WS and AS training produces substantial spatial autocorrelation even at $\lambda = 0.1$.

To understand these patterns, we analyzed activation correlations at both local and global scales. At the neighborhood level, WS training produced slightly stronger correlations between adjacent units than AS across all values of λ, even though its objective did not encourage activation similarity (Figure

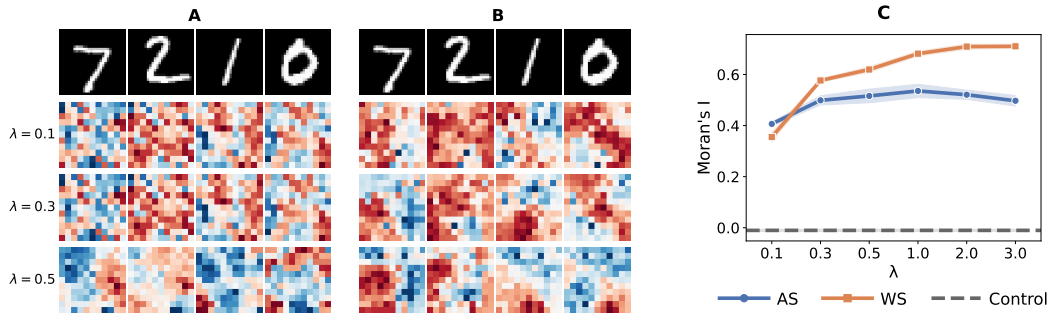


Figure 5.8: **Spatial smoothness of activation maps under activation- and weight-similarity training**. Sample activation maps from a topographic grid layer are shown for WS-trained models (Panel A) and AS-trained models (Panel B) at three values of λ. Panel C shows Moran's I, which quantifies spatial autocorrelation of grid activations. Similar patterns are observed for CIFAR-10 (Appendix Figure 24).

5.9, A–B). The control models show approximate zero correlations. At the level of all unit pairs (Figure 5.9, C–D), AS and WS training produced qualitatively different correlation structures: AS produced either a bimodal distribution or near-perfect correlation ($r \approx 1$) distribution. This means that the AS objective was achieved by strongly coupling a subset of units while leaving others weakly correlated. WS, in contrast, produced a flatter distribution with higher correlations across a broader range of unit pairs. Both AS and WS increased local similarity among incoming weight vectors relative to control (Supplementary Figure 25), though more strongly for WS as would be expected from its training objective.

5.4.5 Reorganization of angular and eccentricity tuning under topography

Relative to control models, topographic regularization resulted in changes to tuning profiles. The strongest effects were found when analyzing orientation-like angular responses and for the balance between central and peripheral eccentricity coding. For angular tuning, topographic regularization changed the balance between orientation tuning (`cycle=2`) and symmetry tuning (`cycle=4`) responses (Figure 5.10). For CIFAR-10, both AS and WS increased the prevalence of orientation tuning and reduced the number of units displaying symmetry tuning. For MNIST, the opposite patterns are observed: both AS and WS decrease the former and increase the latter. Thus, topographic training reweighted angular harmonics, but this depended on dataset properties

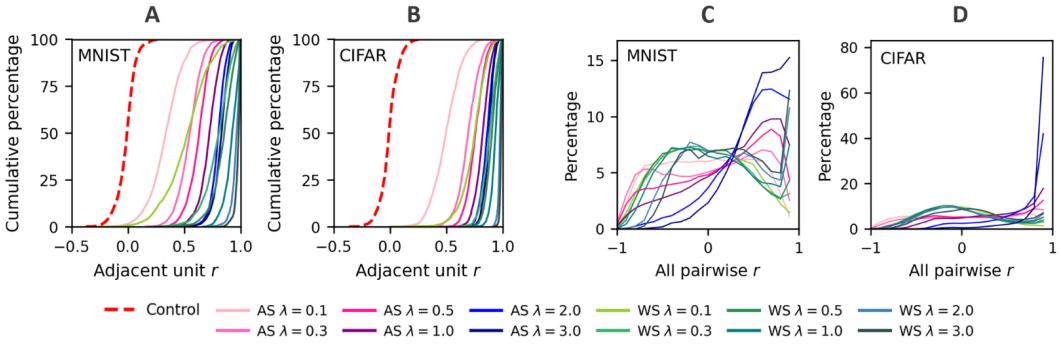


Figure 5.9: **Different patterns of correlated activations under activation- and weight-similarity training**. Cumulative distributions of pairwise activation correlations between adjacent units are shown for MNIST (Panel A) and CIFAR-10 (Panel B). Distributions of correlations computed across all unit pairs are shown for MNIST and CIFAR-10 (Panels C and D).

(MNIST vs. CIFAR-10) and the form of topographic regularization. Supplementary Figure 26 shows the spatial distribution of angular tuning profiles in the topographic grid for WS-trained models.

A reorganization was also observed for eccentricity tuning (Figure 5.11). Here, tuning profiles reflect radial gain functions: *decreasing* responses correspond to filters that weight central locations more strongly; *increasing* responses weight peripheral locations more strongly. The most salient effect was that, in both datasets, both AS and WS strongly reduced centrally preferring (decreasing) responses as compared to control. In MNIST, this reduction was accompanied by an increase in monotonically increasing responses, indicating greater sensitivity to peripheral locations.

5.4.6 Expert units

All models developed expert units that discriminated specific categories. We defined two levels of expertise: moderate-expertise units that systematically showed higher activity for one category in more than 70% of random category–noncategory comparisons, and high-expertise units, corresponding to a 90% threshold. The main finding, as shown in Figure 5.12, was that for both MNIST (Panel A) and CIFAR-10 (Panel D), AS models produced a larger proportion of high-expertise units as compared to WS and control. Regarding moderate-expertise, AS and WS are on par.

We next examined how expert units were distributed across categories by quantifying the balance of expertise (Figure 5.12, B, E). Overall, in both datasets the balance tends to decrease in stronger λ, showing the effect of

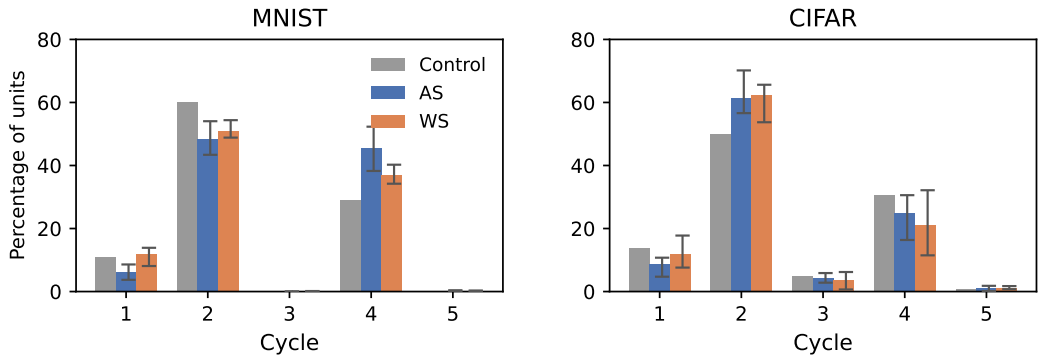


Figure 5.10: **Distribution of angular tuning profiles relative to control models under topographic training**. Percentage of units associated with each angular sensitivity profile (Cycles 1–5) is shown. Cycle 1: classic polar-angle tuning; Cycle 2: orientation-like tuning; Cycles 3–5: higher angular frequencies. Bars indicate mean percentages averaged across the six values of the regularization parameter λ; vertical whiskers indicate the full min–max range across λ. In both datasets, orientation-like tuning (Cycle 2) is the most prevalent profile, and AS and WS models differ from control models in the relative distribution of units across profiles. Full cycle-response data, shown separately for each value of λ, are reported in Figure 27.

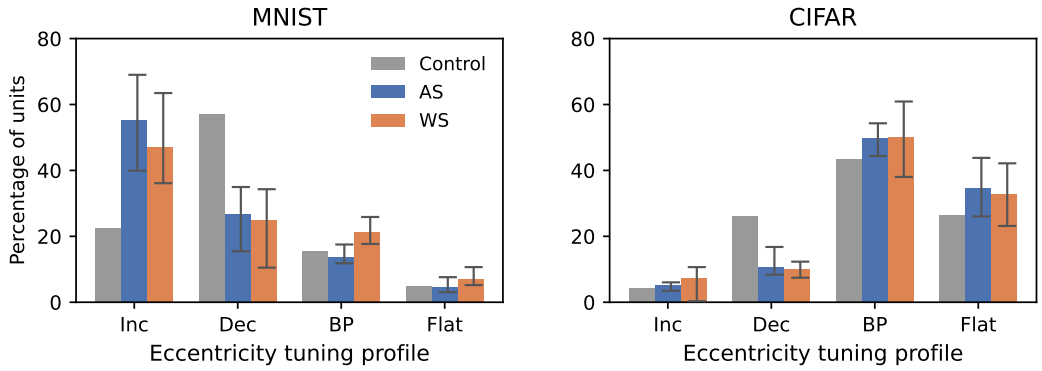


Figure 5.11: **Distribution of eccentricity tuning profiles relative to control models under topographic training**. Percentage of units showing increasing (Inc), decreasing (Dec), band-pass (BP), or flat eccentricity tuning profiles is shown. Bars indicate mean percentages averaged across values of the regularization parameter λ; vertical whiskers indicate the full min–max range across λ. In both datasets, AS and WS models differ from control in the relative distribution of units across eccentricity tuning profiles. Eccentricity data, shown separately for each value of λ, are reported in Figure 28.

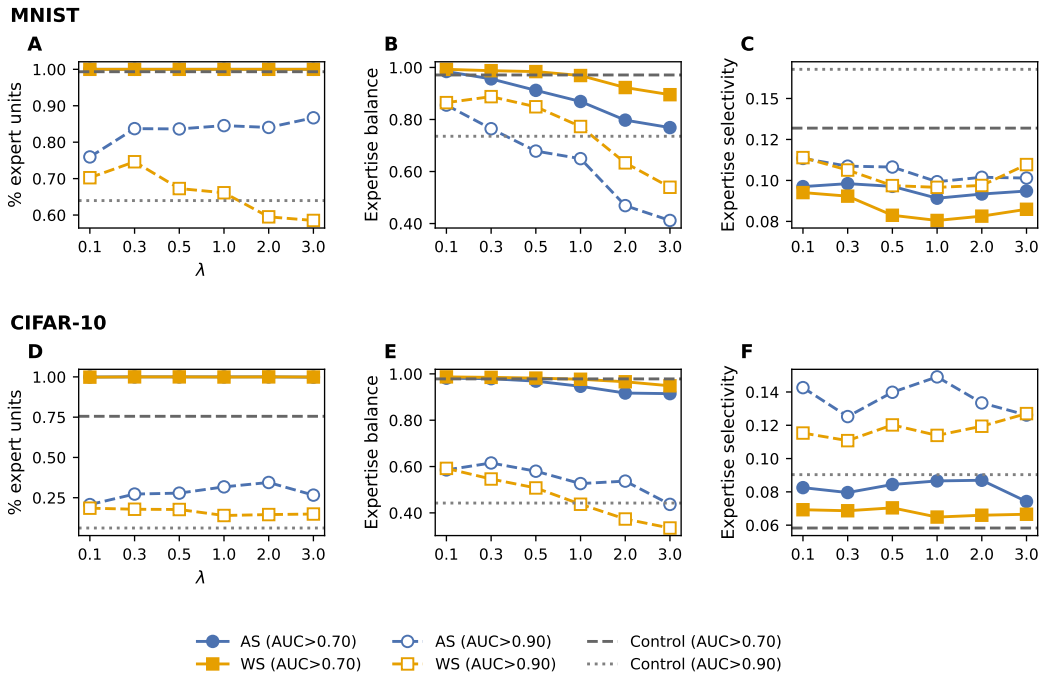


Figure 5.12: **Expert unit prevalence, balance, and selectivity**. Expert unit prevalence, defined as the proportion of units whose category discriminability exceeds a threshold (AUC $>$ 0.70 or AUC $>$ 0.90), is shown for MNIST (Panels A) and CIFAR-10 (Panels D). Expertise balance, quantified as the normalized entropy of expert units across categories, is shown in Panels B and E. Expertise selectivity, defined as the difference between the highest and second-highest AUC values for each unit, is shown in Panels C and F. See Methods for formal definitions of all measures.

skewing the distribution of expert units. For CIFAR-10, the balance was quite similar for moderate-expertise, while AS is slightly higher than WS and control for high-expertise. For MNIST, the patterns for high-expertise balance are opposite, where AS is lower than WS and control.

Finally, we evaluated the selectivity of expert units by isolating them and quantifying the difference between the response to the most preferred category and the second-most preferred category (Figure 5.12, panels C and F). We find that for both MNIST and CIFAR-10 the relative selectivity of expert units did not differ strongly among conditions. Taken together, the results show that topographic training strongly impacted the proportion and distribution of expert units. However, this depended on an interaction of dataset properties (MNIST vs. CIFAR-10) and the topographic loss itself.

5.5 Discussion

We studied how topographic regularization impacts robustness and representational characteristics of shallow CNNs when applied during end-to-end training. We focused on two local topographic losses: activation similarity (AS), which encourages Pearson correlation between neighboring units' activation vectors (computed across inputs), and weight similarity (WS), which encourages similarity among adjacent units' incoming (afferent) weight vectors. Our main objectives were to evaluate (1) robustness to readout-weight perturbations and input corruptions, and (2) representational properties including the entropy, sparsity, similarity profiles of the weights and activations, effective dimensionality, the emergence of class-selective (expert) units, and the spatial organization of responses on the 2D grid including smoothness and functional localization.

The following main findings emerge: (1) Overall, AS and WS models produced greater robustness to input corruption as compared to control models; (2) WS models were more robust to perturbations of the readout (classifier) weight matrix, accompanied by lower activation sparsity (lower PoZ); (3) Both AS and WS models could produce smooth topographic maps, however, WS achieved so with stronger functional localization; (4) Both AS and WS changed the distribution of pairwise correlations, consistent with the objectives of the two regularizer, but AS did so by aggressively driving very high pairwise correlations.

5.5.1 Topographic regularization improves robustness to noise

One key finding emerging from our study is that topographic regularization can produce models that are more robust to both weight perturbations and input corruptions. This is important because several approaches for achieving robustness against noisy inputs have used training on corrupted data (e.g., Sietsma and Dow 1991; Lopes et al. 2019). In prior work, robustness against parameter perturbations (Cheney et al. 2017; Arechiga and Michaels 2018; Savva et al. 2023) typically relies on specifically tailored loss terms (Tsai et al. 2021). In contrast, we find that robustness can be a byproduct of the topographic regularization itself. This suggests that a biologically inspired regularizer can be a viable strategy for improving robustness (see J. Liu and Y. Jin 2023 for review on robustness methods). Specifically, for both CIFAR-10 and MNIST, WS models showed advantages over AS and control models in terms of robustness to weight perturbation. In particular, representational similarity matrices of WS-trained models were more stable under weight perturbation, and clas-

sification accuracy degraded less. On the other hand, AS models were often more robust against image corruption for both MNIST and CIFAR-10. These results extend prior work showing that certain types of topographic architectures are more robust against adversarial attacks (Qian et al. 2026; Bashivan et al. 2025; D. Zhou et al. 2025).

Importantly, we observe a non-monotonic relationship between robustness and model performance. For MNIST, AS models trained with a weak regularization ($\lambda = 0.1$ and 0.3) produced higher accuracy than control models (Figure 5.2, left). Notably, the same setting also produced stronger robustness than controls. This pattern contrasts with prior work reporting that improved accuracy is associated with reduced robustness, particularly in adversarial-robustness settings (Su et al. 2018; Tsipras et al. 2018). Our findings point to a potentially important exception, where a moderate topographic regularization can, under some settings, improve both classification performance and robustness.

Our findings also show that the relative accuracy of topographic models and non-topographic controls depends on regularization strength (λ). In fact, previous studies report mixed findings, with some reporting reduced accuracy (Margalit et al. 2024; Z. Lu et al. 2025; Rathi et al. 2024) or slightly improved performance (Poli et al. 2023; X.-J. Zhang et al. 2025; Deb et al. 2025; Rathi et al. 2024) depending on setting. These spatial loss functions differed in their topographic objective, including local similarity between unit-weights or activations (Z. Lu et al. 2025; Deb et al. 2025), and global distance-weighted similarity penalties on activations (Margalit et al. 2024; Poli et al. 2023) or weight magnitudes (X.-J. Zhang et al. 2025; Jacobs and Jordan 1992). While the literature on this point is mixed, our findings suggest that the strength of topographic regularization is a key determinant.

5.5.2 Topographic regularization reshapes model representations

As noted in the introduction, in machine-learning practice, the presence of correlations among unit activations or among incoming weight vectors is often discouraged due to reduced feature diversity (Zbontar et al. 2021; Z. Wang et al. 2020) and because models often aim to de-correlate representations (Cogswell et al. 2015; Rodríguez et al. 2016; G. Jin et al. 2020). This emphasis differs from the role of correlations in biological systems, where correlated activity can serve to amplify specific signals (e.g., Shadlen and Newsome 1998). In artificial networks, however, similar effects can be achieved without correlated populations, for instance, by using high downstream readout weights. Still, some

machine-learning approaches intentionally produce structured redundancy, for example, by developing methods to cluster similar weights or activations into groups that enforce within-group similarity, with the intention of providing a basis for subsequent model compression (S. Han et al. 2015; D. Zhang et al. 2018; Neill et al. 2020; Wen et al. 2016). The results of the AS condition are most relevant to this literature as we find that it produces a substantial proportion of units with highly correlated activation profiles, suggesting that it might be a useful loss term for subsequent pruning or compression. While we have not specifically studied the pruning or compression potential of topographic models, several studies have already shown that topographic models are more amenable to pruning (Poli et al. 2023; Deb et al. 2025; Z. Lu et al. 2025).

The changes we find in representational organization were accompanied by changes to single-unit responses properties. Specifically, WS training was associated with significantly higher unit entropy than AS training in CIFAR-10, while in MNIST the entropy values are quite similar to those of AS and control. WS training also produced a lower PoZ than that of AS and control in both datasets. Higher entropy indicates greater response variability across inputs. The lower PoZ suggests reduced sparsity because units are active for a larger fraction of inputs. Both statistics depended on the strength of regularization, and differed for AS and WS models.

Consistent with prior reports (Margalit et al. 2024; Qian et al. 2026; Deb et al. 2025), we found that the topographic regularizer could reduce the effective dimensionality of the representations as λ increases (Figure 5.6). Both AS and WS training tend to produce lower effective dimensionality than control models. However, not all topographic conditions were associated with lower overall effective dimensionality; specifically, for CIFAR-10, both AS and WS showed higher dimensionality than control at certain levels of λ, and AS models do not follow a monotonic decreasing trend. In contrast, *within-class* effective dimensionality was consistently at or below control levels. This suggests that the topographic regularization mainly reduces within-class variation.

Lower-dimensional representations have been linked to increased robustness, as they are more resilient to various types of perturbations (Sanyal et al. 2018; Awasthi et al. 2020). Related findings in non-topographic models also suggest that encouraging similarity in learned features or activations improves robustness (Nassar et al. 2021; Gourtani and Meratnia 2024). Importantly, our results (Figure 5.6) suggest that robustness is not necessarily accompanied by lower dimensionality: in some cases, topographic regularization was associated with both higher overall effective dimensionality and greater robustness.

When analyzing the spatial organization of activity, we found that while

both AS and WS models produced high spatial smoothness (Moran's I), WS achieved so with shorter spatial distances between highly correlated units. AS models, interestingly, did not show a monotonic increase in spatial smoothness when the strength of spatial regularization (λ) increases, suggesting that over-smoothing can hurt the smoothness of activation maps in topographic models.

Regarding the emergence of expert units, we find that topographic regularization influences unit-level specialization by changing the number and distribution of class-selective expert units. The difference in the proportion of expert units between WS and AS, especially at high λ levels, suggests that both the type of loss function and the regularization strength control the production of category-selective elements. The AUC gap between the top and second-best class was similar across conditions, suggesting that topographic regularization mainly affects expert prevalence and class coverage rather than increasing unit selectivity.

5.5.3 Implications and future directions

An important implication for machine-learning applications is that moderate topographic regularization strengths can improve both task performance and robustness to noise. This suggests that modulating local correlations may be beneficial in some settings, even when brain-like spatial smoothness is not an objective in itself. This can be simply implemented by adding a topographic regularizer to the training objective.

Another direction for future research is to evaluate pruning in different types of topographic models (Poli et al. 2023; Z. Lu et al. 2025; X.-J. Zhang et al. 2025; Deb et al. 2025; Blauch et al. 2022). This can be done not only to achieve compression, but also to identify subnetworks that maintain high performance. Another possibility is to evaluate the existence of "winning tickets" as suggested by the lottery ticket hypothesis (Frankle and Carbin 2018), to test whether topographic networks are associated with sparser winning tickets. Extending topographic regularization to additional layers, including convolutional ones, could improve robustness to adversarial inputs, as suggested in Bashivan et al. (2025) and D. Zhou et al. (2025). In parallel, scaling topographic training to large-scale datasets could help facilitate comparisons of noise robustness between models and neural data (Jang, McCormack, et al. 2021; Jang and Tong 2024). Finally, we note that we studied topographic regularization in a supervised classification setting, where the spatial losses were jointly optimized with cross-entropy. More generally, topographic regularizers can be combined with many sorts of objectives, including unsupervised and self-supervised losses. An important direction for future research is to understand if the effects observed here generalize beyond supervised training.

Conclusions

5.6 Summary of contributions

This thesis studies cognitive-inspired approaches, implemented through structured pruning and topographic constraints, as methods for aligning representations between human and DNNs, and shaping the internal representations of DNNs beyond task accuracy. The objectives are to test whether models capture targeted aspects of human representational space, and to characterize how different constraints change what models represent and how robust or interpretable those representations become.

Part I develops pruning as both a predictive, explanatory and diagnostic tool for alignment to a target representational geometry, which could be behavioral or synthetic similarity judgments, or the geometry from the models themselves. Across the three studies in Part I, we show that pruning is an effective tool to select a subset of model activations that align with human similarity judgments, often better than using the full activations, as is commonly done in the literature. Moreover, the remaining/kept subset can be interpreted as relevant or irrelevant for modeling a certain semantic category such as animals, furnitures, etc. Study 1.1 demonstrates that the kept dimensions of convolutional layers can be projected back into image space to generate heatmaps that highlight the visual information most relevant for explaining human comparisons, offering an interpretable tool different from usual methods producing saliency heatmaps. In Study 1.2, by capturing similarity in the numerosity domain, we find that alignment is better explained by representational geometry than by isolated expert units. In Study 1.3, we contribute a geometry-guided structured pruning procedure that removes redundant channels or units while preserving a target representational geometry, extending earlier supervised pruning methods into a more general framework.

Part II evaluates what fine-grained cortical organization that topographic models can capture, and what computational advantages that topographic constraints offer. Specifically, in Study 2.1, we test whether a leading topographic model can capture the action-related dimension in occipitotemporal

cortex. While the model captures broad divisions such as animacy, it fails to produce an action-related gradient, highlighting a limitation of current generic spatial constraints and motivating additional inductive biases. In Study 2.2, going beyond the focus on topographic map visualization, we compare two common local constraints - Activation Similarity (AS) and Weight Similarity (WS) - and show that they yield different representational and computational consequences. Importantly, both can improve robustness to input perturbations and parameter noises, with AS generally producing clearer benefits on the former and WS on the latter. Moreover, AS and WS shape representations differently, affecting localization and feature tuning.

Overall, the thesis evaluates biologically motivated constraints not only by alignment performance or topographic map visualization, but by how they reshape the internal representational geometry, support interpretability, and improve robustness, helping clarify when and how such constraints can move models toward more human-like grounded representations.

5.7 Connecting to the broader landscape of literature

5.7.1 Low dimensional structure in human and model representations

A consistent finding across the pruning studies in Part I is that only a restricted subspace of a model's representation is relevant for a given semantic domain. This idea is aligned with previous evidence investigating the multidimensional representations in both behavioral and neural data. Analyses of large-scale behavioral datasets such as THINGS (Martin N Hebart, Zheng, et al. 2020) consistently show that the vast majority of variance in human similarity judgments is captured by a small number of interpretable axes. Specifically, a limited number of sparse, interpretable dimensions are sufficient to predict odd-one-out judgments across thousands of everyday objects. This low-dimensional structure is not confined to human similarity judgments: a series of studies using the fMRI and MEG data collected with the THINGS dataset (Martin N Hebart, Dickter, et al. 2019; Martin N Hebart, Contier, et al. 2023) have shown that these same dimensions are spatially mapped onto the visual cortex in topographic arrangements (Contier et al. 2024), and that they exhibit distinct, dissociable time-resolved profiles during visual processing (Teichmann et al. 2026), suggesting that dimensionality reduction is a fundamental organizing principle of cortical representation. Beyond object

perception, this principle extends to action recognition: a limited subset of action-related dimensions can account for the broad range of everyday actions humans perform (Bockes et al. 2025).

The pruning framework developed in Part I provides a computationally explicit method for identifying the analogous low-dimensional subspaces within artificial neural networks. Where approaches such as SPoSE (Martin N Hebart, Zheng, et al. 2020) and related embedding models recover interpretable dimensions by training directly on behavioral data, AIS-guided pruning and CRISP isolate these subspaces within pretrained networks without retraining, revealing which portions of an existing model's representational geometry carry human-relevant structure and which are redundant or misaligned. This connection also points to an important limitation. Recent work by Mahner et al. (2025) shows that the dimensions spontaneously recovered from deep network representations tend to reflect purely visual properties such as color, texture, or shape, rather than the semantic and functional properties that dominate human behavioral dimensions. Our results in Study 2.1 is consistent with these findings and reflect them at the neural level: current artificial neural networks (including topographic models or even networks trained in action recognition) fail to reproduce the action-related gradient in lateral OTC, despite capturing more visually-based divisions such as shape. This also suggests that action may potentially constitute a demanding future benchmark for closing the alignment gap between models and cortex.

Taken together, these converging findings - from behavioral embeddings, neural topography, time-resolved decoding, and structured pruning - suggest that dimensionality reduction is not just a computational convenience, but reflects a fundamental organization of representational geometry of the brain and, to an important but incomplete degree, of artificial neural networks.

5.7.2 Biological and artificial pruning

An interesting parallelism can be drawn between pruning in biological and artificial neural networks. While DNNs do not necessarily replicate the brain, and DNN pruning primarily serves engineering objectives, many works in machine learning related to pruning mention biological synaptic pruning as a source of inspiration (e.g. Bellec et al. 2017; Gerum et al. 2020; Zhao and Zeng 2021; B. Han et al. 2024). In our work, although we applied structured pruning to activations or feature maps, at the implementation level it involved the removal of weights, analogous to synapses in the brain.

In biological development, the brain initially overproduces synaptic connections and subsequently eliminates the weaker, less-used ones, a process that has been shown to refine neural circuits, improve signal-to-noise ratio, and

support the emergence of sparse, efficient coding (Neniskyte and Gross 2017; Riccomagno and Kolodkin 2015). This process is not random: competition for limited resources means that active, informative synapses are preserved while redundant ones are removed (Stephan et al. 2012). The pruning framework developed in Part I of this thesis operates under a similar logic. AIS-guided pruning and CRISP do not remove weights based on magnitude alone, but instead identify and retain the subspace of a model's representation that is most informative for a specific target geometry, which are human similarity judgments or the model's own representational structure. This mirrors the biological pruning more closely than standard magnitude-based approaches, which have been argued to lack biological plausibility because small weight magnitude does not imply low functional importance (Scholl et al. 2021).

A further point of convergence may be the functional consequences of pruning. In the brain, adolescent synaptic pruning in the prefrontal cortex helps improve high-level cognitive functions, such as working memory and goal-directed behavior, at the cost of some flexibility and generalization (Blakemore 2008; Averbeck 2022). Similarly, pruned recurrent networks show more stable dynamics and more accurate in reinforcement learning tasks, but learn new tasks more slowly than unpruned networks (Averbeck 2022). In modeling human similarity judgments, pruning a pretrained network for a given semantic category improves out-of-sample prediction within that domain, but the kept subspace is category-specific and does not necessarily generalize across domains (Bavaresco, Truong, et al. 2025). This suggests that the pruned representations are more specialized and less flexible, consistent with the functional consequences of pruning as a mechanism that finetunes representations to a specialized target structure at the expense of broader utility. In both the brain and models, pruning is not only a compression tool, but also a mechanism that shapes the representation toward sparse, task-relevant coding.

5.7.3 From spatial layout to computational advantages in topographic models

The topographic modeling work presented in this thesis is part of a broader and rapidly growing effort to understand what architectural constraints are necessary and sufficient to reproduce the spatial organization of ventral visual cortex, and potentially of the cortex more generally. A number of end-to-end topographic models based on deep artificial neural networks have been proposed in recent years, differing in the specific form of the spatial constraint they impose. One of the most prominent is the model of Margalit et al. (2024), which we adopt in Study 2.1 to test its ability to capture a novel organizational

dimension. While this model successfully reproduces broad divisions such as animacy and real-world size, we find that it fails to produce the action-related gradient we identify in lateral occipitotemporal cortex, suggesting that generic spatial smoothness constraints, however powerful, do not rule out the inductive biases needed to account for the fine-grained functional architecture of high-level visual cortex. Other recent models have explored alternative formulations of the spatial loss or relaxed standard architectural assumptions. Deb et al. (2025) demonstrated that topographic constraints can generalize beyond the visual domain, capturing spatial organization in auditory cortex as well, suggesting that the organizing principles may reflect domain-general properties of cortical computation rather than vision-specific solutions. Z. Lu et al. (2025) proposed a model that moves away from the weight-sharing structure of standard convolutional networks, a feature widely regarded as biologically implausible, by optimizing correlations among neighboring weight vectors, also reproducing retinotopic and category-selective organization. Despite their differences in implementation, all of these models converge on the same result: spatially constrained deep networks replicate (to an extent) the known organization of ventral visual cortex and beyond.

A potential limitation shared by most of these approaches, including the method in this thesis, is that the spatial organization they produce is imposed at some levels, rather than discovered: the spatial loss directly penalizes deviations from local similarity, and this must result in some forms of clustering. An old study (Sirosh and Miikkulainen 1994) and a recent study (Qian et al. 2026) challenges this assumption by showing that explicit spatial losses may not be necessary at all. Qian et al. (2026) demonstrated that incorporating local lateral connections into the architecture, without any direct spatial loss, is sufficient for the network to develop category-selective clusters organized along large-scale gradients resembling those observed in visual cortex. This result raises an interesting question about the origin of cortical topography: whether it is shaped by optimization pressure on spatial layout, or whether it is shaped as an emergent consequence of local connectivity.

Our work contributes to the topographic modeling literature in a distinct but parallel way. Rather than treating topographic modeling purely as a replication of the spatial layout of the cortex, we investigate topographic constraints as computational tools that shape the internal representational geometry of networks and confer functional advantages beyond map formation. Study 2.2 shows that inducing topographic organization improves robustness to both input perturbations and parameter noise. In this sense, the choice of spatial loss is not just a technical detail but provides a computational benefit to the model. This finding bring extra engineering benefits to the computational cog-

nitive neuroscience community, and may contribute as a bridge to connect to machine learning community.

Because the landscape of topographic models is diverse, with many different technical implementations, and because some existing works investigated adversarial noise with naturalistic data, whereas we used non-adversarial noise with small-scale, simple datasets, we cautiously extrapolate our findings to other models. We speculate that any method that produces high correlation in neighboring weights will benefit parameter-noise robustness. Methods that impose activation similarity at a global level (Margalit et al. 2024; Poli et al. 2023) may result in very high correlations between any pairs of units across the topographic grid, as suggested in our local AS cases. In these models, since they achieved slightly lower or on-par accuracy with non-topographic models, they may rely on different magnitudes of activation in the topographic layers to satisfy the classification or self-supervision task, or employ decorrelation at the readout layer. Moreover, because we found that stronger spatial regularization does not necessarily lead to smoother topographic maps, we speculate that, if one optimizes for smoothness, spatial regularization strength should be searched as a normal hyperparameter.

5.8 Relationship between pruning and topography

Since the thesis presents two parallel topics, here we investigate the relationship between the two by reviewing the use of pruning deep neural networks in topographic models. Topographic models, regardless of whether they are trained to increase correlation in activities or not, produce smoother activities than non-topographic baselines, which may increase redundancy and thus create a basis for pruning. Poli et al. (2023) applied weight-based magnitude pruning (L2) to convolutional filters and found that their topographic models can be pruned more than baseline models given the same accuracy drop. A similar finding is reported in Deb et al. (2025), where they applied L1 weight pruning. Z. Lu et al. (2025) lesioned the 25% lowest-entropy units in their topographic maps, resulting in only a small decrease in classification performance. We speculate that topographic models that have lower effective dimensionality than baseline models (e.g., our models, Margalit et al. 2024; Qian et al. 2026), encode more redundancy in their representations, and therefore could be pruned more than baseline models to achieve the same accuracy.

Another application of pruning in topographic models is to use pruning as a lesion tool on category-selective areas that specialize in certain functions,

to test whether the model can still function without the missing units, and to compare the findings to the neuroscience literature. Tyson N Aflalo and M. S. Graziano (2006) lesioned the hand-selective units in their model of motor cortex and found that the model reorganizes the functions in other units to reallocate hand-selective responses, consistent with observations in the brain. Cowell and Cottrell (2013) removed units that were maximally activated for face and house images, and found that the remaining units were still sufficient for classifying the two categories, replicating findings in humans (Haxby et al. 2001). Blauch et al. (2022) reports the opposite pattern: lesioning selective units for certain categories (face, object, scene) leads to a strong drop in accuracy for the lesioned category, and a slight drop in other categories, suggesting that functional organization is highly specialized but not strictly modular.

Since minimization of wiring length is one hypothesis explaining the formation of topography (Chklovskii and Koulakov 2004), pruning, which removes neuronal connections thus reduces wiring length, could be considered as a method to implement learning mechanisms in topographic models. Jacobs and Jordan (1992) showed that penalizing the weights between distant units drives the emergence of topography resembling organization in early visual cortex, but without explicitly pruning these weights from the model. Achterberg et al. (2023) explicitly pruned weak weights as a learning mechanism, showing that the model mirrors similar structural and functional organization in many cortical areas beyond vision. X.-J. Zhang et al. (2025) employed sparse evolutionary training, where weights between units are grown and pruned to satisfy a task loss and a topographic loss that penalizes long distance connections (similar to Jacobs and Jordan 1992). They find that their models can improve task performance, form task-specific modules, and simulate the distribution of neuronal connections in animals such as Ciona intestinalis and Caenorhabditis elegans (a type of sea vase and worm). Deb et al. (2025) mentioned synaptic pruning in the brain as inspiration to develop their methods (although they do not employ pruning as a mechanism to achieve topography), while Poli et al. (2023) cited synaptic pruning as motivation to perform their pruning experiments.

5.9 General limitations and future directions

The limitations and future directions of each study were discussed within each chapter, therefore, here we sketch the overall limitations and future directions across all presented studies. Regarding the theoretical limitations, we have not examined the nuance of human similarity judgments that can change depending on context and task (Murphy and Medin 1985; Roads and Love 2024). For

example, in an odd-one-out task with "banana", "blackberry", and "android", the odd one out could be "android" if a person considers the other two words as fruits, but it could be "banana" if a person considers the other two words as smartphones. Moreover, we use average population similarity judgments, so we do not capture individual differences that may result in selecting different subsets of the model via pruning (Carroll and J.-J. Chang 1970; Hönekopp 2006; Simmons and Estes 2008; Ichien et al. 2019). We note that our method can be applied to individual representational structure without adding any technical details.

Regarding the technical limitations, the pretrained model choices may need to be diversified, as the current models in the thesis are all CNNs. For example, vision transformers are also common models of vision (Oota et al. 2023; J. Tang et al. 2023; Adeli et al. 2025; Nguyen et al. 2025). This can lead to more comprehensive comparisons; however, vision transformers are not necessarily better aligned with human vision compared to traditional CNN models (Q. Zhou et al. 2022; Linsley, Rodriguez Rodriguez, et al. 2023; Linsley, Feng, et al. 2025). Moreover, the current models are trained with supervised learning, so adding self-supervised models could also help extend the work to investigate the effect of tasks on representational changes (Bakhtiari et al. 2021; Konkle and Alvarez 2022). Another limitation is in the metric measurements: we always employ RSA, which has certain drawbacks such as loss of important stimulus–response information, second-order confounds, and cases where high alignment scores still coexist with unhuman-like failures in DNNs (Dujmović et al. 2023; Viviani 2021; Gao et al. 2025). Methods such as linear regression between activations and behavior/brain data, or Centered Kernel Alignment, could complement RSA and compensate for some of these limitations (Kornblith et al. 2019).

Regarding future directions, in Study 1.1, we showed that pruning feature maps in convolutional networks against human similarity judgments can provide a mechanistic explanation of how people compare objects, therefore, future work may apply the same method to neural data, extending this framework beyond the behavioral level. This would make it possible to test whether neural representations in specific brain regions are sufficient to explain object comparisons and to identify which areas best capture the representational structure underlying human similarity judgments. In Study 1.2, we applied pruning in the numerosity domain and found that number-detector units are not critical for preserving population-level representations. Future work should better characterize the semantic content of the number-detector units, as well as the units retained or removed by pruning, using explainable AI approaches. In Study 1.3, we used pruning to isolate a subnetwork in later layers that pre-

served the representational geometry of either the original network or human similarity judgments. In future work, pruning should be extended to the whole network to test the broader utility of the method. Explainability analyses may be applied to decode the semantic content of retained and removed units or feature maps, thus revealing what information is important for preserving the geometric structure of a semantic category in pretrained models.

In Study 2.1, we found that a prominent topographic model could not capture the action dimension in the ventral visual stream. This suggests that future models of action-related organization may need to be trained on more ecological tasks, such as fine-grained action discrimination, or even on embodied data from simulations or robots interacting with objects in the real world. In Study 2.2, we showed that correlation topographic constraints can improve robustness to noise. A direction for future work is to scale these approaches to larger datasets and compare their noise robustness more directly with neural data. It will also be important to test whether these effects generalize across a wider range of architectures, including recurrent or transformer-based models, and across alternative training objectives such as self-supervised learning.

Another general future direction is extending to the language domain, as large language models now not only process language but can surprisingly model many vision tasks as well (Doerig et al. 2025; Conwell 2024; Bavaresco and Fernández 2025). Analyzing neural data is also necessary, so that we can move toward the underlying mechanisms that generate behavior (Charest et al. 2014).

Bibliography

Abbott, Larry F and Peter Dayan (1999). "The effect of correlated variability on the accuracy of a population code". In: *Neural computation* 11.1, pp. 91–101.

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. (2023). "Gpt-4 technical report". In: *arXiv preprint arXiv:2303.08774*.

Achterberg, Jascha, Danyal Akarca, DJ Strouse, John Duncan, and Duncan E Astle (2023). "Spatially embedded recurrent neural networks reveal widespread links between structural and functional neuroscience findings". In: *Nature Machine Intelligence* 5.12, pp. 1369–1381.

Adeli, Hossein, Sun Minni, and Nikolaus Kriegeskorte (2025). "Transformer brain encoders explain human high-level visual responses". In: *arXiv preprint arXiv:2505.17329*.

Aflalo, Tyson N and Michael SA Graziano (2006). "Possible origins of the complex topographic organization of motor cortex: reduction of a multidimensional space onto a two-dimensional array". In: *Journal of Neuroscience* 26.23, pp. 6288–6297.

— (2006). "Possible Origins of the Complex Topographic Organization of Motor Cortex: Reduction of a Multidimensional Space onto a Two-Dimensional Array". In: *Journal of Neuroscience* 26.23, pp. 6288–6297. DOI: 10.1523/JNEUROSCI.0768-06.2006.

Almeida, Jorge, Alessio Fracasso, Stephanie Kristensen, Daniela Valério, Fredrik Bergström, Ramakrishna Chakravarthi, Zohar Tal, and Jonathan Walbrin (2023). "Neural and behavioral signatures of the multidimensionality of manipulable object processing". In: *Communications Biology* 6.1, p. 940.

Arcaro, Michael and Margaret Livingstone (2024). "A whole-brain topographic ontology". In: *Annual Review of Neuroscience* 47.

Arechiga, Austin P and Alan J Michaels (2018). "The effect of weight errors on neural networks". In: *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, pp. 190–196.

Attarian, Maria, Brett D Roads, and Michael C Mozer (2020). "Transforming neural network visual representations to predict human judgments of similarity". In: *arXiv preprint arXiv:2010.06512*.

Averbeck, Bruno B (2022). "Pruning recurrent neural networks replicates adolescent changes in working memory and reinforcement learning". In: *Proceedings of the National Academy of Sciences* 119.22, e2121331119.

Awasthi, Pranjal, Himanshu Jain, Ankit Singh Rawat, and Aravindan Vijayaraghavan (2020). "Adversarial robustness via robust low rank representations". In: *Advances in Neural Information Processing Systems* 33, pp. 11391–11403.

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

Bao, Pinglei, Liang She, Mason McGill, and Doris Y Tsao (2020). "A map of object space in primate inferotemporal cortex". In: *Nature* 583.7814, pp. 103–108.

Bao, Wanqian and Uri Hasson (2024). "Identifying and interpreting non-aligned human conceptual representations using language modeling". In: *arXiv preprint arXiv:2403.06204*.

Barrow, Harry G, Alistair J Bray, and Julian ML Budd (1996). "A self-organizing model of "color blob" formation". In: *Neural Computation* 8.7, pp. 1427–1448.

Bashivan, Pouya, Reza Bayat, Adam Ibrahim, Amirozhan Dehghani, and Yifei Ren (2025). "Learning adversarially robust kernel ensembles with kernel average pooling". In: *Expert Systems with Applications* 266, p. 126017.

Battleday, Ruairidh M, Joshua C Peterson, and Thomas L Griffiths (2021). "From convolutional neural networks to models of higher-level cognition (and back again)". In: *Annals of the New York Academy of Sciences* 1505.1, pp. 55–78.

Bavaresco, Anna and Raquel Fernández (2025). "Experiential Semantic Information and Brain Alignment: Are Multimodal Models Better than Language Models?" In: *arXiv preprint arXiv:2504.00942*.

Bavaresco, Anna, Nhut Truong, and Uri Hasson (2025). "Modeling human concepts with subspaces in deep vision models". In: *ACM Transactions on Interactive Intelligent Systems* 15.4, pp. 1–25.

Beeck, Hans Op de et al. (2023). "Category trumps shape as an organizational principle of object space in the human occipitotemporal cortex". In: *Journal of Neuroscience* 43.16, pp. 2960–2972.

Beeck, Hans P Op de, Ineke Pillet, and J Brendan Ritchie (2019). "Factors determining where category-selective areas emerge in visual cortex". In: *Trends in cognitive sciences* 23.9, pp. 784–797.

Beeck, Hans P Op de, Katrien Torfs, and Johan Wagemans (2008). "Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway". In: *Journal of Neuroscience* 28.40, pp. 10111–10123.

Bellec, Guillaume, David Kappel, Wolfgang Maass, and Robert Legenstein (2017). "Deep rewiring: Training very sparse deep networks". In: *arXiv preprint arXiv:1711.05136*.

Binhuraib, Taha, Greta Tuckute, and Nicholas Blauch (2025). "Topoformer: brain-like topographic organization in Transformer language models through spatial querying and reweighting". In: *arXiv preprint arXiv:2510.18745*.

Blakemore, Sarah-Jayne (2008). "Development of the social brain during adolescence". In: *Quarterly journal of experimental psychology* 61.1, pp. 40–49.

Blauch, Nicholas M, Marlene Behrmann, and David C Plaut (2022). "A connectivity-constrained computational account of topographic organization in primate high-level visual cortex". In: *Proceedings of the National Academy of Sciences* 119.3, e2112566119.

Bockes, André, Martin N Hebart, and Angelika Lingnau (2025). "Revealing Key Dimensions Underlying the Recognition of Dynamic Human Actions". In: *Communications Psychology* 3.1, p. 149.

Bonhoeffer, Tobias and Amiram Grinvald (1991). "Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns". In: *Nature* 353.6343, pp. 429–431.

Bracci, Stefania and Hans Op de Beeck (2016). "Dissociations and associations between shape and category representations in the two visual pathways". In: *Journal of Neuroscience* 36.2, pp. 432–444.

Bracci, Stefania, Cristiana Cavina-Pratesi, Magdalena Ietswaart, Alfonso Caramazza, and Marius V Peelen (Mar. 2012). "Closely overlapping responses to tools and hands in left lateral occipitotemporal cortex". en. In: *J. Neurophysiol.* 107.5, pp. 1443–1456.

Bracci, Stefania, Magdalena Ietswaart, Marius V Peelen, and Cristiana Cavina-Pratesi (June 2010). "Dissociable neural responses to hands and non-hand body parts in human left extrastriate visual cortex". en. In: *J. Neurophysiol.* 103.6, pp. 3389–3397.

Bracci, Stefania, Jakob Mraz, Astrid Zeman, Gaëlle Leys, and Hans Op de Beeck (2023). "The representational hierarchy in human and artificial vi-

sual systems in the presence of object-scene regularities". In: *PLoS computational biology* 19.4, e1011086.

Bracci, Stefania and Hans P Op de Beeck (2023). "Understanding human object vision: a picture is worth a thousand representations". In: *Annual review of psychology* 74.1, pp. 113–135.

Bracci, Stefania and Marius V Peelen (Nov. 2013). "Body and object effectors: the organization of object representations in high-level visual cortex reflects body-object interactions". en. In: *J. Neurosci.* 33.46, pp. 18247–18258.

Brainard, David H and Spatial Vision (1997). "The psychophysics toolbox". In: *Spatial vision* 10.4, pp. 433–436.

Cant, Jonathan S and Melvyn A Goodale (2007). "Attention to form or surface properties modulates different regions of human occipitotemporal cortex". In: *Cerebral cortex* 17.3, pp. 713–731.

Cant, Jonathan S and Yaoda Xu (2012). "Object ensemble processing in human anterior-medial ventral visual cortex". In: *Journal of Neuroscience* 32.22, pp. 7685–7700.

Cao, Steven, Victor Sanh, and Alexander M Rush (2021). "Low-complexity probing via finding subnetworks". In: *arXiv preprint arXiv:2104.03514*.

Carroll, J Douglas (1976). "Spatial, non-spatial and hybrid models for scaling". In: *Psychometrika* 41.4, pp. 439–463.

Carroll, J Douglas and Jih-Jie Chang (1970). "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition". In: *Psychometrika* 35.3, pp. 283–319.

Castaldi, Elisa, Manuela Piazza, Stanislas Dehaene, Alexandre Vignaud, and Evelyn Eger (July 2019). "Attentional amplification of neural codes for number independent of other quantities along the dorsal visual stream". In: *eLife* 8. DOI: 10.7554/elife.45160. URL: https://doi.org/10.7554/elife.45160.

Chao, Linda L, James V Haxby, and Alex Martin (1999). "Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects". In: *Nature neuroscience* 2.10, pp. 913–919.

Chapalain, Thomas, Bertrand Thirion, and Evelyn Eger (Sept. 2024). "Trained deep neural network models of the ventral visual pathway encode numerosity with robustness to object and scene identity". In: *bioRxiv*. DOI: 10.1101/2024.09.05.611433. URL: https://doi.org/10.1101/2024.09.05.611433.

Charest, Ian, Rogier A Kievit, Taylor W Schmitz, Diana Deca, and Nikolaus Kriegeskorte (2014). "Unique semantic space in the brain of each beholder predicts perceived similarity". In: *Proceedings of the National Academy of Sciences* 111.40, pp. 14565–14570.

Chartouny, Augustin, Keivan Amini, Mehdi Khamassi, and Benoıt Girard (2024). "A new paradigm to study social and physical affordances as model-based reinforcement learning". In: *Cognitive Robotics* 4, pp. 142–155.

Chen, Fanghui, Shouliang Li, Jiale Han, Fengyuan Ren, and Zhen Yang (2024). "Review of Lightweight Deep Convolutional Neural Networks: F. Chen et al." In: *Archives of Computational Methods in Engineering* 31.4, pp. 1915–1937.

Chen, Ting, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PmLR, pp. 1597–1607.

Cheney, Nicholas, Martin Schrimpf, and Gabriel Kreiman (2017). "On the robustness of convolutional neural networks to internal architecture and weight perturbations". In: *arXiv preprint arXiv:1703.08245*.

Cheng, Hongrong, Miao Zhang, and Javen Qinfeng Shi (2024). "A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cheng, Yu, Felix X Yu, Rogerio S Feris, Sanjiv Kumar, Alok Choudhary, and Shi-Fu Chang (2015). "An exploration of parameter redundancy in deep networks with circulant projections". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2857–2865.

Chklovskii, Dmitri B and Alexei A Koulakov (2004). "Maps in the brain: what can we learn from them?" In: *Annu. Rev. Neurosci.* 27.1, pp. 369–392.

Chklovskii, Dmitri B, Thomas Schikorski, and Charles F Stevens (2002). "Wiring optimization in cortical circuits". In: *Neuron* 34.3, pp. 341–347.

Cichy, Radoslaw Martin, Aditya Khosla, Dimitrios Pantazis, Antonio Torralba, and Aude Oliva (June 2016). "Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence". In: *Scientific Reports* 6.1. DOI: 10.1038/srep27755. URL: https://doi.org/10.1038/srep27755.

Coggan, David D and Frank Tong (June 2023). "Spikiness and animacy as potential organizing principles of human ventral visual cortex". en. In: *Cereb. Cortex* 33.13, pp. 8194–8217.

Cogswell, Michael, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra (2015). "Reducing overfitting in deep networks by decorrelating representations". In: *arXiv preprint arXiv:1511.06068*.

Cohen, Yarden, Tatiana A Engel, Christopher Langdon, Grace W Lindsay, Torben Ott, Megan AK Peters, James M Shine, Vincent Breton-Provencher, and Srikanth Ramaswamy (2022). "Recent advances at the interface of neu-

roscience and artificial neural networks". In: *Journal of Neuroscience* 42.45, pp. 8514–8523.

Contier, Oliver, Chris I Baker, and Martin N Hebart (2024). "Distributed representations of behaviour-derived object dimensions in the human visual system". In: *Nature Human Behaviour* 8.11, pp. 2179–2193.

Conwell, Colin (2024). "Is visual cortex really "language-aligned"? Perspectives from Model-to-Brain Comparisons in Human and Monkeys on the Natural Scenes Dataset". PhD thesis. Harvard Medical School.

Cortinovis, Davide, Marius V Peelen, and Stefania Bracci (2025). "Tool representations in human visual cortex". In: *Journal of Cognitive Neuroscience* 37.3, pp. 515–531.

Cortinovis, Davide, Nhut Truong, Hans Op de Beeck, and Stefania Bracci (2025). "Investigating action topography in visual cortex and deep artificial neural networks". In: *Nature Communications*.

Cowell, Rosemary A. and Garrison W. Cottrell (Nov. 2013). "What Evidence Supports Special Processing for Faces? A Cautionary Tale for fMRI Interpretation". In: *Journal of Cognitive Neuroscience* 25.11, pp. 1777–1793. DOI: 10.1162/JOCN_A_00448.

Cunningham, James P (1978). "Free trees and bidirectional trees as representations of psychological distance". In: *Journal of mathematical psychology* 17.2, pp. 165–188.

Deb, Mayukh, Mainak Deb, and N Murty (2025). "TopoNets: High performing vision and language models with brain-like topography". In: *arXiv preprint arXiv:2501.16396*.

Dehaene, Stanislas (Apr. 2011). *The number sense*. OUP USA.

Dehghani, Amirozhan, Xinyu Qian, Asa Farahani, and Pouya Bashivan (2024). "Credit-based self organizing maps: training deep topographic networks with minimal performance degradation". In: *The Thirteenth International Conference on Learning Representations*.

Demircan, Can, Tankred Saanum, Leonardo Pettini, Marcel Binz, Blazej Baczkowski, Christian Doeller, Mona Garvert, and Eric Schulz (2024). "Evaluating alignment between humans and neural network representations in image-based learning tasks". In: *Advances in Neural Information Processing Systems* 37, pp. 122406–122433.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Cision and Pattern Recognition*. Ieee, pp. 248–255.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, None Kai Li, and None Li Fei-Fei (June 2009). "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. DOI:

10.1109/cvpr.2009.5206848. URL: https://doi.org/10.1109/cvpr.2009.5206848.

DeWind, Nicholas K., Geoffrey K. Adams, Michael L. Platt, and Elizabeth M. Brannon (Sept. 2015). "Modeling the approximate number system to quantify the contribution of visual stimulus features". In: *Cognition* 142, pp. 247–265. DOI: 10.1016/j.cognition.2015.05.016. URL: https://doi.org/10.1016/j.cognition.2015.05.016.

Dijk, Jelle A van, Maartje C de Jong, Gio Piantoni, Alessio Fracasso, Mariska J Vansteensel, Iris IA Groen, Natalia Petridou, and Serge O Dumoulin (2022). "Intracranial recordings show evidence of numerosity tuning in human parietal cortex". In: *Plos one* 17.8, e0272087.

Ding, Zhuokun, Paul G Fahey, Stelios Papadopoulos, Eric Y Wang, Brendan Celii, Christos Papadopoulos, Andersen Chang, Alexander B Kunin, Dat Tran, Jiakun Fu, et al. (2025). "Functional connectomics reveals general wiring rule in mouse visual cortex". In: *Nature* 640.8058, pp. 459–469.

Doerig, Adrien, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest (2025). "High-level visual representations in the human brain are aligned with large language models". In: *Nature Machine Intelligence* 7.8, pp. 1220–1234.

Doshi, Fenil R. and Talia Konkle (June 2023). "Cortical Topographic Motifs Emerge in a Self-Organized Map of Object Space". In: *Science Advances* 9.25, eade8187. DOI: 10.1126/sciadv.ade8187.

Downing, Paul E, Yuhong Jiang, Miles Shuman, and Nancy Kanwisher (2001). "A cortical area selective for visual processing of the human body". In: *Science* 293.5539, pp. 2470–2473.

Dujmović, Marin, Jeffrey S Bowers, Federico Adolfi, and Gaurav Malhotra (2023). *Obstacles to inferring mechanistic similarity using Representational Similarity Analysis*. Universitätsbibliothek Johann Christian Senckenberg.

Durbin, Richard and Graeme Mitchison (1990). "A dimension reduction framework for understanding cortical maps". In: *Nature* 343.6259, pp. 644–647.

Edwards, Laura A., Jennifer B. Wagner, Charline E. Simon, and Daniel C. Hyde (Sept. 2015). "Functional brain organization for number processing in pre-verbal infants". In: *Developmental Science* 19.5, pp. 757–769. DOI: 10.1111/desc.12333. URL: https://doi.org/10.1111/desc.12333.

Erwin, Ed, Klaus Obermayer, and Klaus Schulten (1995). "Models of orientation and ocular dominance columns in the visual cortex: A critical comparison". In: *Neural computation* 7.3, pp. 425–468.

Fedzechkina, Masha, Eleonora Gualdoni, Sinead Williamson, Katherine Metcalf, Skyler Seto, and Barry-John Theobald (2025). "ExpertLens: Activation steering features are highly interpretable". In: *arXiv preprint arXiv:2502.15090*.

Filus, Katarzyna and Joanna Domańska (2024). "Extracting coarse-grained classifiers from large Convolutional Neural Networks". In: *Engineering Applications of Artificial Intelligence* 138, p. 109377.

— (2025). "What is the doggest dog? Examination of typicality perception in ImageNet-trained networks". In: *Neural Networks* 188, p. 107425.

Finzi, Dawn, Eshed Margalit, Kendrick Kay, Daniel LK Yamins, and Kalanit Grill-Spector (2023). "A single computational objective drives specialization of streams in visual cortex". In: *bioRxiv*, pp. 2023–12. DOI: 10.1101/2023.12.19.572460. URL: https://www.biorxiv.org/content/10.1101/2023.12.19.572460v1.

Flechas Manrique, Natalia, Wanqian Bao, Aurelie Herbelot, and Uri Hasson (Dec. 2023a). "Enhancing Interpretability Using Human Similarity Judgements to Prune Word Embeddings". In: *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi. Singapore: Association for Computational Linguistics, pp. 169–179. DOI: 10.18653/v1/2023.blackboxnlp-1.13. URL: https://aclanthology.org/2023.blackboxnlp-1.13.

— (Oct. 2023b). "Enhancing Interpretability using Human Similarity Judgements to Prune Word Embeddings". In: *Proceedings of BlackboxNLP at EMNLP 2023*.

Frankle, Jonathan and Michael Carbin (2018). "The lottery ticket hypothesis: Finding sparse, trainable neural networks". In: *arXiv preprint arXiv:1803.03635*.

Gallivan, Jason P, Jonathan S Cant, Melvyn A Goodale, and J Randall Flanagan (2014). "Representation of object weight in human ventral visual cortex". In: *Current Biology* 24.16, pp. 1866–1873.

Gao, Chuanji, Gang Chen, Svetlana V Shinkareva, and Rutvik H Desai (2025). "Is Representational Similarity Analysis Reliable? A Comparison with Regression". In: *arXiv preprint arXiv:2511.00395*.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann (2020). "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11, pp. 665–673.

Geirhos, Robert, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel (2018). "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness". In: *International conference on learning representations*.

Gerum, Richard C, André Erpenbeck, Patrick Krauss, and Achim Schilling (2020). "Sparsity through evolutionary pruning prevents neuronal networks from overfitting". In: *Neural Networks* 128, pp. 305–312.

Gomez, Jesse, Michael Barnett, and Kalanit Grill-Spector (June 2019). "Extensive childhood experience with Pokémon suggests eccentricity drives organization of visual cortex". en. In: *Nat. Hum. Behav.* 3.6, pp. 611–624.

Goodhill, Geoffrey J (2007). "Contributions of theoretical modeling to the understanding of neural map development". In: *Neuron* 56.2, pp. 301–311.

Gou, Jianping, Baosheng Yu, Stephen J Maybank, and Dacheng Tao (2021). "Knowledge distillation: A survey". In: *International journal of computer vision* 129.6, pp. 1789–1819.

Gourtani, Saeed Khalilian and Nirvana Meratnia (2024). "Improving robustness of compressed models with weight sharing through knowledge distillation". In: *2024 IEEE 10th International Conference on Edge Computing and Scalable Cloud (EdgeCom)*. IEEE, pp. 13–21.

Goyal, Anirudh and Yoshua Bengio (2022). "Inductive biases for deep learning of higher-level cognition". In: *Proceedings of the Royal Society A* 478.2266, p. 20210068.

Grill-Spector, Kalanit and Kevin S Weiner (2014). "The functional architecture of the ventral temporal cortex and its role in categorization". In: *Nature Reviews Neuroscience* 15.8, pp. 536–548.

Han, Bing, Feifei Zhao, Yi Zeng, and Guobin Shen (2024). "Developmental plasticity-inspired adaptive pruning for deep spiking and artificial neural networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47.1, pp. 240–251.

Han, Song, Huizi Mao, and William J Dally (2015). "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding". In: *arXiv preprint arXiv:1510.00149*.

Hannagan, Thomas (Nov. 2021). "Reset Networks: Emergent Topography in Networks of Convolutional Neural Networks". In: *bioRxiv*. Preprint, version 2. DOI: 10.1101/2021.11.19.469308. URL: https://www.biorxiv.org/content/10.1101/2021.11.19.469308v2.

Hansen, Bryan J, Mircea I Chelaru, and Valentin Dragoi (2012). "Correlated variability in laminar cortical circuits". In: *Neuron* 76.3, pp. 590–602.

Harris, Kenneth D and Thomas D Mrsic-Flogel (2013). "Cortical connectivity and sensory coding". In: *Nature* 503.7474, pp. 51–58.

Harvey, Ben M, Barrie P Klein, Natalia Petridou, and Serge O Dumoulin (2013). "Topographic representation of numerosity in the human parietal cortex". In: *Science* 341.6150, pp. 1123–1126.

Hassabis, Demis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick (2017). "Neuroscience-inspired artificial intelligence". In: *Neuron* 95.2, pp. 245–258.

Hasson, Uri, Michal Harel, Ifat Levy, and Rafael Malach (Mar. 2003). "Large-scale mirror-symmetry organization of human occipito-temporal object areas". en. In: *Neuron* 37.6, pp. 1027–1041.

Haxby, James V, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini (2001). "Distributed and overlapping representations of faces and objects in ventral temporal cortex". In: *Science* 293.5539, pp. 2425–2430.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034.

— (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He, Yang and Lingao Xiao (2023). "Structured pruning for deep convolutional neural networks: A survey". In: *IEEE transactions on pattern analysis and machine intelligence* 46.5, pp. 2900–2919.

Hebart, Martin N, Oliver Contier, Lina Teichmann, Adam H Rockter, Charles Y Zheng, Alexis Kidder, Anna Corriveau, Maryam Vaziri-Pashkam, and Chris I Baker (2023). "THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior". In: *Elife* 12, e82580.

Hebart, Martin N, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker (2019). "THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images". In: *PloS one* 14.10, e0223792.

Hebart, Martin N, Charles Y Zheng, Francisco Pereira, and Chris I Baker (2020). "Revealing the multidimensional mental representations of natural objects underlying human similarity judgements". In: *Nature human behaviour* 4.11, pp. 1173–1185.

Hebb, Donald O (1949). "Organization of behavior: A neurophysiological theory". In: *(No Title)*.

Hendrycks, Dan and Thomas Dietterich (2019). "Benchmarking neural network robustness to common corruptions and perturbations". In: *arXiv preprint arXiv:1903.12261*.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean (2015). "Distilling the knowledge in a neural network". In: *arXiv preprint arXiv:1503.02531*.

Hodges Jr, JL (1958). "The significance probability of the Smirnov two-sample test". In: *Arkiv för matematik* 3.5, pp. 469–486.

Hönekopp, Johannes (2006). "Once more: is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness." In: *Journal of Experimental Psychology: Human Perception and Performance* 32.2, p. 199.

Hou, Kuinan, Marco Zorzi, and Alberto Testolin (2024). "Estimating the distribution of numerosity and non-numerical visual magnitudes in natural scenes using computer vision". In: *arXiv preprint arXiv:2409.11028*.

Hu, Hengyuan, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang (July 2016). "Network Trimming: A Data-Driven Neuron Pruning Approach towards Efficient Deep Architectures". en. In: *arXiv:1607.03250 [cs]*. arXiv: 1607.03250. URL: http://arxiv.org/abs/1607.03250.

Hu, Huan, Guillermo Penna, and Vishal Monga (2016). "Network trimming: A data-driven neuron pruning approach towards efficient deep architectures". In: *arXiv preprint arXiv:1607.03250*.

Huang, Gao, Zhuang Liu, and Kilian Q. Weinberger (2016). "Densely Connected Convolutional Networks". In: *CoRR* abs/1608.06993. arXiv: 1608.06993. URL: http://arxiv.org/abs/1608.06993.

Huang, Taicheng, Yiying Song, and Jia Liu (July 2022). "Real-world size of objects serves as an axis of object space". en. In: *Commun. Biol.* 5.1, p. 749.

Humphries, Colin, Einat Liebenthal, and Jeffrey R Binder (2010). "Tonotopic organization of human auditory cortex". In: *Neuroimage* 50.3, pp. 1202–1211.

Huth, Alexander G, Shinji Nishimoto, An T Vu, and Jack L Gallant (2012). "A continuous semantic space describes the representation of thousands of object and action categories across the human brain". In: *Neuron* 76.6, pp. 1210–1224.

Hyde, Daniel C., David A. Boas, Clancy Blair, and Susan Carey (Nov. 2010). "Near-infrared spectroscopy shows right parietal specialization for number in pre-verbal infants". In: *NeuroImage* 53.2, pp. 647–652. DOI: 10.1016/j.neuroimage.2010.06.030. URL: https://doi.org/10.1016/j.neuroimage.2010.06.030.

Ichien, Nicholas, Hongjing Lu, and Keith J Holyoak (2019). "Individual differences in judging similarity between semantic relations". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 41.

Izard, Véronique, Ghislaine Dehaene-Lambertz, and Stanislas Dehaene (Feb. 2008). "Distinct cerebral pathways for object identity and number in human infants". In: *PLoS Biology* 6.2, e11. DOI: 10.1371/journal.pbio.0060011. URL: https://doi.org/10.1371/journal.pbio.0060011.

Jacobs, Robert A and Michael I Jordan (1992). "Computational consequences of a bias toward short connections". In: *Journal of cognitive neuroscience* 4.4, pp. 323–336.

Jagadeesh, Akshay V and Justin L Gardner (2022). "Texture-like representation of objects in human visual cortex". In: *Proceedings of the National Academy of Sciences* 119.17, e2115302119.

Jang, Hojin, Devin McCormack, and Frank Tong (2021). "Noise-trained deep neural networks effectively predict human vision and its neural responses to challenging images". In: *PLoS biology* 19.12, e3001418.

Jang, Hojin and Frank Tong (2024). "Improved modeling of human vision by incorporating robustness to blur in convolutional neural networks". In: *Nature Communications* 15.1, p. 1989.

Jeon, Ikhwan and Taegon Kim (2023). "Distinctive properties of biological neural networks and recent advances in bottom-up approaches toward a better biologically plausible neural network". In: *Frontiers in Computational Neuroscience* 17, p. 1092185.

Jha, Aditi, Joshua C Peterson, and Thomas L Griffiths (2023). "Extracting low-dimensional psychological representations from convolutional neural networks". In: *Cognitive science* 47.1, e13226.

Ji, Jiaming, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, et al. (2025). "AI alignment: A contemporary survey". In: *ACM Computing Surveys* 58.5, pp. 1–38.

Jiang, Dunhan, Tianye Wang, Shiming Tang, and Tai-Sing Lee (Nov. 2024). "Computational Constraints Underlying the Emergence of Functional Domains in the Topological Map of Macaque V4". In: *bioRxiv*. Preprint, version 1. DOI: 10.1101/2024.11.30.626117. URL: https://www.biorxiv.org/content/10.1101/2024.11.30.626117v1.

Jiang, Xiaolong, Shan Shen, Cathryn R Cadwell, Philipp Berens, Fabian Sinz, Alexander S Ecker, Saumil Patel, and Andreas S Tolias (2015). "Principles of connectivity among morphologically defined cell types in adult neocortex". In: *Science* 350.6264, aac9462.

Jin, Gaojie, Xinping Yi, Liang Zhang, Lijun Zhang, Sven Schewe, and Xiaowei Huang (2020). "How does weight correlation affect generalisation ability of deep neural networks?" In: *Advances in Neural Information Processing Systems* 33, pp. 21346–21356.

Jozwik, Kamila M, Nikolaus Kriegeskorte, Katherine R Storrs, and Marieke Mur (2017). "Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments". In: *Frontiers in psychology* 8, p. 1726.

Jozwik, Kamila Maria, Hyodong Lee, Nancy G. Kanwisher, and James J. DiCarlo (2023). "First Steps in Using Topographic Deep Artificial Neural Network Models to Generate Hypotheses about Not-yet-detected Functional Neural Clusters in the Ventral Stream". In: *2023 Conference on Cognitive Computational Neuroscience*. URL: https://api.semanticscholar.org/CorpusID:261596768.

Kaas, Jon H (1997). "Topographic maps are fundamental to sensory processing". In: *Brain research bulletin* 44.2, pp. 107–112.

Kabulska, Zuzanna, Tonghe Zhuang, and Angelika Lingnau (2024). "Overlapping representations of observed actions and action-related features". In: *Human Brain Mapping* 45.3, e26605.

Kaniuth, Philipp and Martin N Hebart (2022a). "Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior". In: *NeuroImage* 257, p. 119294.

— (Aug. 2022b). "Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior". In: *NeuroImage* 257, p. 119294. DOI: 10.1016/j.neuroimage.2022.119294. URL: https://doi.org/10.1016/j.neuroimage.2022.119294.

Kanwisher, Nancy (2010). "Functional specificity in the human brain: a window into the functional architecture of the mind". In: *Proceedings of the national academy of sciences* 107.25, pp. 11163–11170.

Karami, Alireza (2024). "The representation of numerosity in the human brain and machines". PhD thesis. DOI: https://doi.org/10.15168/11572_402591.

Karami, Alireza, Elisa Castaldi, Evelyn Eger, Martin Hebart, and Manuela Piazza (Nov. 16, 2025). "Numerosity Is Directly Sensed and Dynamically Transformed in the Human Brain: Evidence from MEG-MRI Fusion". In: *bioRxiv*. DOI: 10.1101/2025.11.15.687894. URL: https://doi.org/10.1101/2025.11.15.687894.

Karami, Alireza, Elisa Castaldi, Evelyn Eger, and Manuela Piazza (July 9, 2025). "Distinct neural representational geometries of numerosity in early visual and association regions across visual streams". In: *Communications Biology* 8.1, p. 1029. DOI: 10.1038/s42003-025-08395-z. URL: https://doi.org/10.1038/s42003-025-08395-z.

Karami, Alireza, Nhut Truong, and Manuela Piazza (Jan. 1, 2025). "Investigation of Numerosity Representation in Convolution Neural Networks". In: *CCN 2025 Proceedings*. DOI: 10.32470/4j01408. URL: https://doi.org/10.32470/4j01408.

Kehl, Marcel S, Sina Mackay, Kathrin Ohla, Matthias Schneider, Valeri Borger, Rainer Surges, Marc Spehr, and Florian Mormann (2024). "Single-neuron representations of odours in the human brain". In: *Nature* 634.8034, pp. 626–634.

Keller, T. Anderson, Qinghe Gao, and Max Welling (Oct. 2021). "Modeling Category-Selective Cortical Regions with Topographic Variational Autoencoders". In: *arXiv*. Preprint, v2: Dec 19, 2021; SVRHM workshop @ NeurIPS 2021. DOI: 10.48550/arXiv.2110.13911. arXiv: 2110.13911 [q-bio.NC]. URL: https://arxiv.org/abs/2110.13911v2.

Khaligh-Razavi, Seyed-Mahdi and Nikolaus Kriegeskorte (Nov. 2014). "Deep supervised, but not unsupervised, models may explain IT cortical representation". In: *PLoS Computational Biology* 10.11, e1003915. DOI: 10.1371/journal.pcbi.1003915. URL: https://doi.org/10.1371/journal.pcbi.1003915.

Kim, Gwangsu, Jaeson Jang, Seungdae Baek, Min Song, and Se-Bum Paik (2021a). "Visual number sense in untrained deep neural networks". In: *Science advances* 7.1, eabd6127.

— (Jan. 2021b). "Visual number sense in untrained deep neural networks". In: *Science Advances* 7.1. DOI: 10.1126/sciadv.abd6127. URL: https://doi.org/10.1126/sciadv.abd6127.

King, Marcie L, Iris IA Groen, Adam Steel, Dwight J Kravitz, and Chris I Baker (2019). "Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images". In: *NeuroImage* 197, pp. 368–382.

Kobylkov, Dmitry, Uwe Mayer, Mirko Zanon, and Giorgio Vallortigara (2022a). "Number neurons in the nidopallium of young domestic chicks". In: *Proceedings of the National Academy of Sciences* 119.32, e2201039119.

— (Aug. 2022b). "Number neurons in the nidopallium of young domestic chicks". In: *Proceedings of the National Academy of Sciences* 119.32. DOI: 10.1073/pnas.2201039119. URL: https://doi.org/10.1073/pnas.2201039119.

Kohonen, Teuvo (1982). "Self-organized formation of topologically correct feature maps". In: *Biological cybernetics* 43.1, pp. 59–69.

Konkle, Talia and George A Alvarez (2022). "A self-supervised domain-general learning framework for human ventral stream representation". In: *Nature communications* 13.1, p. 491.

Konkle, Talia and Alfonso Caramazza (2013). "Tripartite organization of the ventral stream by animacy and object size". In: *Journal of Neuroscience* 33.25, pp. 10235–10242.

Konkle, Talia and Aude Oliva (2012). "A real-world size organization of object responses in occipitotemporal cortex". In: *Neuron* 74.6, pp. 1114–1124.

Kornblith, Simon, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton (2019). "Similarity of neural network representations revisited". In: *International conference on machine learning*. PMlR, pp. 3519–3529.

Kriegeskorte, Nikolaus (Jan. 2008). "Representational similarity analysis – connecting the branches of systems neuroscience". In: *Frontiers in Systems Neuroscience*. DOI: 10.3389/neuro.06.004.2008. URL: https://doi.org/10.3389/neuro.06.004.2008.

— (2015). "Deep neural networks: a new framework for modeling biological vision and brain information processing". In: *Annual review of vision science* 1.1, pp. 417–446.

Kriegeskorte, Nikolaus, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini (2008). "Matching categorical object representations in inferior temporal cortex of man and monkey". In: *Neuron* 60.6, pp. 1126–1141.

Krizhevsky, Alex, Geoffrey Hinton, et al. (2009). "Learning multiple layers of features from tiny images". In: URL: https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Krug, Valerie, Raihan Kabir Ratul, Christopher Olson, and Sebastian Stober (2023). "Visualizing deep neural networks with topographic activation maps". In: *HHAI 2023: Augmenting Human Intellect*. IOS Press, pp. 138–152.

Kruskal, J. B. (Mar. 1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1, pp. 1–27. DOI: 10.1007/bf02289565. URL: https://doi.org/10.1007/bf02289565.

Kruskal, Joseph B (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". In: *Psychometrika* 29.1, pp. 1–27.

Kubilius, Jonas, Stefania Bracci, and Hans P Op de Beeck (2016). "Deep neural networks as a computational model for human shape sensitivity". In: *PLoS computational biology* 12.4, e1004896.

Kubilius, Jonas, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib J. Majaj, Elias B. Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo (Oct. 2019). "Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs". In: arXiv.

Kubilius, Jonas, Martin Schrimpf, Aran Nayebi, Daniel Bear, Daniel L. K. Yamins, and James J. DiCarlo (Sept. 2018). "CORNET: Modeling the neural mechanisms of core object recognition". In: *bioRxiv (Cold Spring*

Harbor Laboratory). DOI: 10.1101/408385. URL: https://doi.org/10.1101/408385.

Kutter, Esther F, Jan Bostroem, Christian E Elger, Florian Mormann, and Andreas Nieder (2018). "Single neurons in the human brain encode numbers". In: *Neuron* 100.3, pp. 753–761.

Lake, Brenden M, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman (2017). "Building machines that learn and think like people". In: *Behavioral and brain sciences* 40, e253.

Lake, Brenden M, Wojciech Zaremba, Rob Fergus, and Todd M Gureckis (2015). "Deep neural networks predict category typicality ratings for images". In: *Proceedings of the annual meeting of the cognitive science society*. Vol. 37.

Lasne, Gabriel, Manuela Piazza, Stanislas Dehaene, Andreas Kleinschmidt, and Evelyn Eger (May 2019). "Discriminability of numerosity-evoked fMRI activity patterns in human intra-parietal cortex reflects behavioral numerical acuity". In: *Cortex* 114, pp. 90–101. DOI: 10.1016/j.cortex.2018.03.008. URL: https://doi.org/10.1016/j.cortex.2018.03.008.

Lê, Minh Tri, Pierre Wolinski, and Julyan Arbel (2023). "Efficient neural networks for tiny machine learning: A comprehensive review". In: *arXiv preprint arXiv:2311.11883*.

LeCun, Yann (1998). "The MNIST database of handwritten digits". In: *http://yann.lecun. com/exdb/mnist/*. Accessed 2026-01-10.

Lee, Hyodong, Eshed Margalit, Kamila M. Jozwik, Michael A. Cohen, Nancy Kanwisher, Daniel L. K. Yamins, and James J. DiCarlo (July 2020). "Topographic Deep Artificial Neural Networks Reproduce the Hallmarks of the Primate Inferior Temporal Cortex Face Processing Network". In: *bioRxiv*. Preprint. DOI: 10.1101/2020.07.09.185116. URL: https://www.biorxiv.org/content/10.1101/2020.07.09.185116v1.

Levy, I, U Hasson, G Avidan, T Hendler, and R Malach (May 2001). "Center-periphery organization of human object areas". en. In: *Nat. Neurosci.* 4.5, pp. 533–539.

Levy, Robert B and Alex D Reyes (2012). "Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex". In: *Journal of Neuroscience* 32.16, pp. 5609–5619.

Liao, Zhu, Victor Quétu, Van-Tam Nguyen, and Enzo Tartaglione (2023). "Can unstructured pruning reduce the depth in deep neural networks?" In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1402–1406.

Lindsay, Grace W. (Sept. 2021). "Convolutional neural networks as a model of the visual system: past, present, and future". In: *Journal of Cognitive*

Neuroscience 33.10, pp. 2017–2031. DOI: 10.1162/jocn\{_}a\{_}01544. URL: https://doi.org/10.1162/jocn_a_01544.

Lingnau, Angelika and Paul E Downing (May 2015). "The lateral occipitotemporal cortex in action". en. In: *Trends Cogn. Sci.* 19.5, pp. 268–277.

Linsley, Drew, Pinyuan Feng, and Thomas Serre (2025). "Better artificial intelligence does not mean better models of biology". In: *Trends in Cognitive Sciences*.

Linsley, Drew, Ivan F Rodriguez Rodriguez, Thomas Fel, Michael Arcaro, Saloni Sharma, Margaret Livingstone, and Thomas Serre (2023). "Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex". In: *Advances in Neural Information Processing Systems* 36, pp. 28873–28891.

Liu, Jia and Yaochu Jin (2023). "A comprehensive survey of robust deep learning in computer vision". In: *Journal of Automation and Intelligence* 2.4, pp. 175–195.

Livingstone, Margaret S and David H Hubel (1984). "Anatomy and physiology of a color system in the primate visual cortex". In: *Journal of Neuroscience* 4.1, pp. 309–356.

Loftus, Geoffrey R and Michael EJ Masson (1994). "Using confidence intervals in within-subject designs". In: *Psychonomic Bulletin & Review* 1.4, pp. 476–490.

Long, Bria, Chen-Ping Yu, and Talia Konkle (2018). "Mid-level visual features underlie the high-level categorical organization of the ventral stream". In: *Proceedings of the National Academy of Sciences* 115.38, E9015–E9024.

Lopes, Raphael Gontijo, Dong Yin, Ben Poole, Justin Gilmer, and Ekin D Cubuk (2019). "Improving robustness without sacrificing accuracy with patch gaussian augmentation". In: *arXiv preprint arXiv:1906.02611*.

López-Cardona, Ángela, Sebastián Idesis, Mireia Masias-Bruns, Sergi Abadal, and Ioannis Arapakis (2025). "Brain-Language Model Alignment: Insights into the Platonic Hypothesis and Intermediate-Layer Advantage". In: *arXiv preprint arXiv:2510.17833*.

Lou, Jianxun, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu (2022). "TranSalNet: Towards perceptually relevant visual saliency prediction". In: *Neurocomputing* 494, pp. 455–467.

Lu, Haidong D and Anna W Roe (2008). "Functional organization of color domains in V1 and V2 of macaque monkey revealed by optical imaging". In: *Cerebral Cortex* 18.3, pp. 516–533.

Lu, Zejin, Adrien Doerig, Victoria Bosch, Bas Krahmer, Daniel Kaiser, Radoslaw M Cichy, and Tim C Kietzmann (2025). "End-to-end topographic

networks as models of cortical map formation and human visual behaviour". In: *Nature Human Behaviour*, pp. 1–17.

Luo, Jian-Hao and Jianxin Wu (2017). "An entropy-based pruning method for cnn compression". In: *arXiv preprint arXiv:1706.05791*.

Luo, Xiangqi, Mingyang Li, Jiahong Zeng, Zhiyun Dai, Zhenjiang Cui, Minhong Zhu, Mengxin Tian, Jiahao Wu, and Zaizhu Han (2024). "Mechanisms underlying category learning in the human ventral occipito-temporal cortex". In: *NeuroImage* 287, p. 120520.

Magri, Caterina, Talia Konkle, and Alfonso Caramazza (2020). "The contribution of object size, manipulability, and stability on neural responses to inanimate objects". In: *bioRxiv*, pp. 2020–11.

Mahner, Florian P, Lukas Muttenthaler, Umut Güçlü, and Martin N Hebart (2025). "Dimensions underlying the representational alignment of deep neural networks with humans". In: *Nature Machine Intelligence* 7.6, pp. 848–859.

Mahon, Bradford Z and Jorge Almeida (2024). "Reciprocal interactions among parietal and occipito-temporal representations support everyday object-directed actions". In: *Neuropsychologia* 198, p. 108841.

Mahon, Bradford Z and Alfonso Caramazza (2011). "What drives the organization of object knowledge in the brain?" In: *Trends in cognitive sciences* 15.3, pp. 97–103.

Mahon, Bradford Z, Shawn C Milleville, Gioia A L Negri, Raffaella I Rumiati, Alfonso Caramazza, and Alex Martin (Aug. 2007). "Action-related properties shape object representations in the ventral stream". en. In: *Neuron* 55.3, pp. 507–520.

Malach, Eran, Gilad Yehudai, Shai Shalev-Schwartz, and Ohad Shamir (2020). "Proving the lottery ticket hypothesis: Pruning is all you need". In: *International Conference on Machine Learning*. PMLR, pp. 6682–6691.

Malach, Rafael, Ifat Levy, and Uri Hasson (Apr. 2002). "The topography of high-order human object areas". en. In: *Trends Cogn. Sci.* 6.4, pp. 176–184.

Margalit, Eshed, Hyodong Lee, Dawn Finzi, James J DiCarlo, Kalanit Grill-Spector, and Daniel LK Yamins (2024). "A unifying framework for functional organization in early and higher ventral visual cortex". In: *Neuron* 112.14, pp. 2435–2451.

Marinó, Giosué Cataldo, Alessandro Petrini, Dario Malchiodi, and Marco Frasca (2023). "Deep neural networks compression: A comparative survey and choice recommendations". In: *Neurocomputing* 520, pp. 152–170.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John

Etchemendy, et al. (2025). "Artificial intelligence index report 2025". In: *arXiv preprint arXiv:2504.07139*.

Mathis, Mackenzie Weygandt and Alexander Mathis (2025). "Joint modelling of brain and behaviour dynamics with artificial intelligence". In: *Nature Reviews Neuroscience*, pp. 1–14.

Matić, Karla, Hans Op de Beeck, and Stefania Bracci (Dec. 2020). "It's not all about looks: The role of object shape in parietal representations of manual tools". en. In: *Cortex* 133, pp. 358–370.

Mehrer, Johannes, Ben Lonnqvist, Anna Mitola, Abdulkadir Gokce, Paolo Papale, and Martin Schrimpf (2025). "Model-Guided Microstimulation Steers Primate Visual Behavior". In: *arXiv preprint arXiv:2510.03684*.

Mehrer, Johannes, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann (2021). "An ecologically motivated image dataset for deep learning yields better models of human vision". In: *Proceedings of the National Academy of Sciences* 118.8, e2011417118.

Mehrer, Johannes, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann (2020). "Individual differences among deep neural network models". In: *Nature communications* 11.1, p. 5725.

Menghani, Gaurav (2023). "Efficient deep learning: A survey on making deep learning models smaller, faster, and better". In: *ACM Computing Surveys* 55.12, pp. 1–37.

Mistry, Percy K., Anthony Strock, Ruizhe Liu, Griffin Young, and Vinod Menon (June 2023). "Learning-induced reorganization of number neurons and emergence of numerical representations in a biologically inspired neural network". In: *Nature Communications* 14.1. DOI: 10.1038/s41467-023-39548-5. URL: https://doi.org/10.1038/s41467-023-39548-5.

Momennejad, Ida (2023). "A rubric for human-like agents and NeuroAI". In: *Philosophical Transactions of the Royal Society B* 378.1869, p. 20210446.

Moran, Patrick AP (1950). "Notes on continuous stochastic phenomena". In: *Biometrika* 37.1/2, pp. 17–23.

Mur, Marieke, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter A Bandettini, and Nikolaus Kriegeskorte (2013). "Human object-similarity judgments reflect and transcend the primate-IT object representation". In: *Frontiers in psychology* 4, p. 128.

Murphy, Gregory L and Douglas L Medin (1985). "The role of theories in conceptual coherence." In: *Psychological review* 92.3, p. 289.

Muttenthaler, Lukas, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith (2022). "Human alignment of neural network representations". In: *arXiv preprint arXiv:2211.01201*.

Muttenthaler, Lukas, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine Hermann, Andrew Lampinen, and Simon Kornblith (2023). "Improving neural network representations using human similarity judgments". In: *Advances in neural information processing systems* 36, pp. 50978–51007.

Nasr, Khaled, Pooja Viswanathan, and Andreas Nieder (2019a). "Number detectors spontaneously emerge in a deep neural network designed for visual object recognition". In: *Science advances* 5.5, eaav7903.

— (May 2019b). "Number detectors spontaneously emerge in a deep neural network designed for visual object recognition". In: *Science Advances* 5.5. DOI: 10.1126/sciadv.aav7903. URL: https://doi.org/10.1126/sciadv.aav7903.

Nassar, Matthew R, Daniel Scott, and Apoorva Bhandari (2021). "Noise correlations for faster and more robust learning". In: *Journal of Neuroscience* 41.31, pp. 6740–6752.

Nauhaus, Ian, Kristina J Nielsen, Anita A Disney, and Edward M Callaway (2012). "Orthogonal micro-organization of orientation and spatial frequency in primate primary visual cortex". In: *Nature neuroscience* 15.12, pp. 1683–1690.

Nayak, Gaurav Kumar, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty (2019). "Zero-shot knowledge distillation in deep networks". In: *International Conference on Machine Learning*. PMLR, pp. 4743–4751.

Neill, James O', Greg Ver Steeg, and Aram Galstyan (2020). "Compressing deep neural networks via layer fusion". In: *arXiv preprint arXiv:2007.14917*.

Neniskyte, Urte and Cornelius T Gross (2017). "Errant gardeners: glial-cell-dependent synaptic pruning and neurodevelopmental disorders". In: *Nature Reviews Neuroscience* 18.11, pp. 658–670.

Nguyen, Xuan-Bac, Hojin Jang, Xin Li, Samee U Khan, Pawan Sinha, and Khoa Luu (2025). "Bractive: A brain activation approach to human visual brain learning". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

O'Reilly, Randall C (1998). "Six principles for biologically based computational models of cortical cognition". In: *Trends in cognitive sciences* 2.11, pp. 455–462.

Obermayer, Klaus, Helge Ritter, and Klaus Schulten (1990). "A principle for the formation of the spatial structure of cortical feature maps." In: *Proceedings of the National Academy of Sciences* 87.21, pp. 8345–8349.

Ohki, Kenichi, Sooyoung Chung, Prakash Kara, Mark Hübener, Tobias Bonhoeffer, and R Clay Reid (2006). "Highly ordered arrangement of single neurons in orientation pinwheels". In: *Nature* 442.7105, pp. 925–928.

Oota, Subba Reddy, Zijiao Chen, Manish Gupta, Raju S Bapi, Gaël Jobard, Frédéric Alexandre, and Xavier Hinaut (2023). "Deep neural networks and brain alignment: Brain encoding and decoding (survey)". In: *arXiv preprint arXiv:2307.10246*.

Op de Beeck, Hans P, Johannes Haushofer, and Nancy G Kanwisher (2008). "Interpreting fMRI data: maps, modules and dimensions". In: *Nature Reviews Neuroscience* 9.2, pp. 123–135.

Orlov, Tanya, Tamar R Makin, and Ehud Zohary (2010). "Topographic representation of the human body in the occipitotemporal cortex". In: *Neuron* 68.3, pp. 586–600.

Ororbia, Alexander, Ankur Mali, Adam Kohan, Beren Millidge, and Tommaso Salvatori (2024). "A review of neuroscience-inspired machine learning". In: *arXiv preprint arXiv:2403.18929*.

Palazzo, Simone, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah (2020). "Decoding brain representations by multimodal learning of neural activity and visual features". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11, pp. 3833–3849.

Park, Joonkoo (Dec. 2022). "Flawed stimulus design in additive-area heuristic studies". In: *Cognition* 229, p. 104919. DOI: 10.1016/j.cognition.2021.104919. URL: https://doi.org/10.1016/j.cognition.2021.104919.

Patel, Gaurav H, David M Kaplan, and Lawrence H Snyder (2014). "Topographic organization in the brain: searching for general principles". In: *Trends in cognitive sciences* 18.7, pp. 351–363.

Peelen, Marius V and Paul E Downing (2017). "Category selectivity in human visual cortex: Beyond visual object recognition". In: *Neuropsychologia* 105, pp. 177–183.

Penfield, Wilder and Edwin Boldrey (1937). "Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation." In: *Brain: A journal of neurology*.

Peterson, Joshua C, Joshua T Abbott, and Thomas L Griffiths (2018). "Evaluating (and improving) the correspondence between deep neural networks and human representations". In: *Cognitive science* 42.8, pp. 2648–2669.

Piazza, Manuela, Véronique Izard, Philippe Pinel, Denis Le Bihan, and Stanislas Dehaene (Oct. 2004). "Tuning curves for approximate numerosity in the human intraparietal sulcus". In: *Neuron* 44.3, pp. 547–555. DOI: 10.1016/j.neuron.2004.10.014. URL: https://doi.org/10.1016/j.neuron.2004.10.014.

Pillet, Ineke, Begüm Cerrahoğlu, Roxane Victoria Philips, Serge Dumoulin, and Hans Op de Beeck (2024). "The position of visual word forms in the

anatomical and representational space of visual categories in occipitotemporal cortex". In: *Imaging neuroscience* 2, pp. 1–28.

Poli, Maxime, Emmanuel Dupoux, and Rachid Riad (2023). "Introducing topography in convolutional neural networks". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 1–5.

Polyak, Adam and Lior Wolf (2015). "Channel-level acceleration of deep face representations". In: *IEEE Access* 3, pp. 2163–2175.

Prince, Jacob S, George A Alvarez, and Talia Konkle (2024). "Contrastive learning explains the emergence and function of visual category-selective regions". In: *Science Advances* 10.39, eadl1776.

Pulvermüller, Friedemann, Rosario Tomasello, Malte R Henningsen-Schomers, and Thomas Wennekers (2021). "Biological constraints on neural network models of cognitive function". In: *Nature Reviews Neuroscience* 22.8, pp. 488–502.

Qian, Xinyu, Amir Ozhan Dehghani, Asa Borzabadi Farahani, and Pouya Bashivan (2026). "Local lateral connectivity is sufficient for replicating cortex-like topographical organization in deep neural networks". In: *Nature Communications*.

Radosavovic, Ilija, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár (2020). "Designing network design spaces". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10428–10436.

Rathi, Neil, Johannes Mehrer, Badr AlKhamissi, Taha Binhuraib, Nicholas M Blauch, and Martin Schrimpf (2024). "TopoLM: brain-like spatio-functional organization in a topographic language model". In: *arXiv preprint arXiv:2410.11516*.

Riccomagno, Martin M and Alex L Kolodkin (2015). "Sculpting neural circuits by axon and dendrite pruning". In: *Annual review of cell and developmental biology* 31.1, pp. 779–805.

Richie, Russell and Sudeep Bhatia (Aug. 2021). "Similarity Judgment Within and Across Categories: A Comprehensive Model Comparison". In: *Cognitive Science* 45.8. ISSN: 1551-6709. DOI: 10.1111/cogs.13030. URL: http://dx.doi.org/10.1111/cogs.13030.

Ringach, Dario L, Patrick J Mineault, Elaine Tring, Nicholas D Olivas, Pablo Garcia-Junco-Clemente, and Joshua T Trachtenberg (2016). "Spatial clustering of tuning in mouse primary visual cortex". In: *Nature communications* 7.1, p. 12270.

Ritchie, J Brendan, Susan G Wardle, Maryam Vaziri-Pashkam, Dwight J Kravitz, and Chris I Baker (2025). "Rethinking category-selectivity in human visual cortex". In: *Cognitive neuroscience*, pp. 1–28.

Ritchie, John Brendan, Spencer T Andrews, Maryam Vaziri-Pashkam, and Chris I Baker (2024). "Graspable foods and tools elicit similar responses in visual cortex". In: *Cerebral Cortex* 34.9, bhae383.

Roads, Brett D and Bradley C Love (2021). "Enriching imagenet with human similarity judgments and psychological embeddings". In: *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 3547–3557.

— (2024). "Modeling similarity and psychological space". In: *Annual Review of Psychology* 75.1, pp. 215–240.

Rodríguez, Pau, Jordi Gonzalez, Guillem Cucurull, Josep M Gonfaus, and Xavier Roca (2016). "Regularizing cnns with locally constrained decorrelations". In: *arXiv preprint arXiv:1611.01967*.

Rolls, Edmund T and Martin J Tovee (1995). "Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex". In: *Journal of neurophysiology* 73.2, pp. 713–726.

Saenz, Melissa and Dave RM Langers (2014). "Tonotopic mapping of human auditory cortex". In: *Hearing research* 307, pp. 42–52.

Sanyal, Amartya, Varun Kanade, Philip HS Torr, and Puneet K Dokania (2018). "Robustness via deep low-rank representations". In: *arXiv preprint arXiv:1804.07090*.

Savva, Andreas G, Theocharis Theocharides, and Chrysostomos Nicopoulos (2023). "Robustness of Artificial Neural Networks Based on Weight Alterations Used for Prediction Purposes". In: *Algorithms* 16.7, p. 322.

Scholl, Carolin, Michael E Rule, and Matthias H Hennig (2021). "The information theory of developmental pruning: Optimizing global network architectures using local synaptic rules". In: *PLoS computational biology* 17.10, e1009458.

Schrimpf, Martin, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. (2018a). "Brain-score: Which artificial neural network for object recognition is most brain-like?" In: *BioRxiv*, p. 407007.

Schrimpf, Martin, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo (Sept. 2018b). "Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?" In: *bioRxiv (Cold Spring Harbor Laboratory)*. DOI: 10.1101/407007. URL: https://doi.org/10.1101/407007.

Shadlen, Michael N and William T Newsome (1998). "The variable discharge of cortical neurons: implications for connectivity, computation, and information coding". In: *Journal of neuroscience* 18.10, pp. 3870–3896.

Shah, Raj, Vijay Marupudi, Reba Koenen, Khushi Bhardwaj, and Sashank Varma (2023). "Numeric magnitude comparison effects in large language models". In: *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6147–6161.

Shen, Hua, Tiffany Knearem, Reshmi Ghosh, Kenan Alkiek, Kundan Krishna, Yachuan Liu, Ziqiao Ma, Savvas Petridis, Yi-Hao Peng, Li Qiwei, et al. (2024). "Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions". In: *arXiv preprint arXiv:2406.09264* 2406, pp. 1–56.

Shepard, Roger N (1962). "The analysis of proximities: multidimensional scaling with an unknown distance function. I." In: *Psychometrika* 27.2, pp. 125–140.

Shoham, Doron, Mark Hübener, Silke Schulze, Amiram Grinvald, and Tobias Bonhoeffer (1997). "Spatio–temporal frequency domains and their relation to cytochrome oxidase staining in cat visual cortex". In: *Nature* 385.6616, pp. 529–533.

Sietsma, Jocelyn and Robert JF Dow (1991). "Creating artificial neural networks that generalize". In: *Neural networks* 4.1, pp. 67–79.

Silson, Edward Harry, Annie Wai-Yiu Chan, Richard Craig Reynolds, Dwight Jacob Kravitz, and Chris Ian Baker (Aug. 2015). "A retinotopic basis for the division of High-Level scene processing between lateral and ventral human occipitotemporal cortex". en. In: *J. Neurosci.* 35.34, pp. 11921–11935.

Silver, Michael A and Sabine Kastner (2009). "Topographic maps in human frontal and parietal cortex". In: *Trends in cognitive sciences* 13.11, pp. 488–495.

Simmons, Sabrina and Zachary Estes (2008). "Individual differences in the perception of similarity and difference". In: *Cognition* 108.3, pp. 781–795.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Sirosh, Joseph and Risto Miikkulainen (1994). "Modeling cortical plasticity based on adapting lateral interaction". In: *The Neurobiology of Computation: Proceedings of the Third Annual Computation and Neural Systems Conference*. Springer, pp. 305–310.

Stephan, Alexander H, Ben A Barres, and Beth Stevens (2012). "The complement system: an unexpected role in synaptic pruning during development and disease". In: *Annual review of neuroscience* 35, pp. 369–389.

Striem-Amit, Ella, Gilles Vannuscorps, and Alfonso Caramazza (May 2017). "Sensorimotor-independent development of hands and tools selectivity in

the visual cortex". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114.18, pp. 4787–4792.

Su, Dong, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao (2018). "Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models". In: *Proceedings of the European conference on computer vision (ECCV)*, pp. 631–648.

Sucholutsky, Ilia, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Iris Groen, Jascha Achterberg, et al. (2023). "Getting aligned on representational alignment". In: *arXiv preprint arXiv:2310.13018*.

Swindale, NV (1996). "The development of topography in the visual cortex: a review of models". In: *Network: Computation in neural systems* 7.2, pp. 161–247.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna (2015). "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567. arXiv: 1512.00567. URL: http://arxiv.org/abs/1512.00567.

Al-Tahan, Haider, Mayukh Deb, Jenelle Feather, and N Murty (2025). "End-to-end Topographic Auditory Models Replicate Signatures of Human Auditory Cortex". In: *arXiv preprint arXiv:2509.24039*.

Tan, Mingxing and Quoc Le (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR, pp. 6105–6114.

Tang, Jerry, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth (2023). "Brain encoding models based on multimodal transformers can transfer across language and vision". In: *Advances in Neural Information Processing Systems* 36, pp. 29654–29666.

Tang, Shiming, Yimeng Zhang, Zhihao Li, Ming Li, Fang Liu, Hongfei Jiang, and Tai Sing Lee (2018). "Large-scale two-photon imaging revealed super-sparse population codes in the V1 superficial layer of awake monkeys". In: *Elife* 7, e33370.

Tarhan, Leyla and Talia Konkle (2020). "Sociality and interaction envelope organize visual action representations". In: *Nature Communications* 11.1, p. 3002.

Tarigopula, Priya, Scott Laurence Fairhall, Anna Bavaresco, Nhut Truong, and Uri Hasson (2023). "Improved prediction of behavioral and neural similarity spaces using pruned DNNs". In: *Neural Networks* 168, pp. 89–104.

Taylor, John C and Paul E Downing (2011). "Division of labor between lateral and ventral extrastriate representations of faces, bodies, and objects". In: *Journal of Cognitive Neuroscience* 23.12, pp. 4122–4137.

Teichmann, Lina, Martin N Hebart, and Chris I Baker (2026). "Dynamic representation of multidimensional object properties in the human brain". In: *Journal of Neuroscience*.

Truong, Nhut, Dario Pesenti, and Uri Hasson (2024). "Explaining Human Comparisons using Alignment-Importance Heatmaps". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 46.

Tsai, Yu-Lin, Chia-Yi Hsu, Chia-Mu Yu, and Pin-Yu Chen (2021). "Formalizing generalization and adversarial robustness of neural networks to weight perturbations". In: *Advances in Neural Information Processing Systems* 34, pp. 19692–19704.

Tsipras, Dimitris, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry (2018). "Robustness may be at odds with accuracy". In: *arXiv preprint arXiv:1805.12152*.

Tucciarelli, Raffaele, Moritz Wurm, Elisa Baccolo, and Angelika Lingnau (2019). "The representational space of observed actions". In: *elife* 8, e47686.

Tversky, Amos (1977). "Features of similarity." In: *Psychological review* 84.4, p. 327.

Upadhyay, Neha and Sashank Varma (2023). "CNN models' sensitivity to numerosity concepts". In: *The 3rd Workshop on Mathematical Reasoning and AI at NeurIPS'23*.

Varma, Sashank, Emily M Sanford, Vijay Marupudi, Olivia Shaffer, and R Brooke Lea (2024). "Recruitment of magnitude representations to understand graded words". In: *Cognitive Psychology* 153, p. 101673.

Viswanathan, Pooja and Andreas Nieder (2013a). "Neuronal correlates of a visual "sense of number" in primate parietal and prefrontal cortices". In: *Proceedings of the National Academy of Sciences* 110.27, pp. 11187–11192.

— (June 2013b). "Neuronal correlates of a visual "sense of number" in primate parietal and prefrontal cortices". In: *Proceedings of the National Academy of Sciences* 110.27, pp. 11187–11192. DOI: 10.1073/pnas.1308141110. URL: https://doi.org/10.1073/pnas.1308141110.

Viviani, Roberto (2021). "Overcoming bias in representational similarity analysis". In: *arXiv preprint arXiv:2102.08931*.

Von der Malsburg, Chr (1973). "Self-organization of orientation sensitive cells in the striate cortex". In: *Kybernetik* 14.2, pp. 85–100.

Wagener, Lysann, Maria Loconsole, Helen M Ditz, and Andreas Nieder (2018a). "Neurons in the endbrain of numerically naive crows spontaneously encode visual numerosity". In: *Current Biology* 28.7, pp. 1090–1094.

— (Apr. 2018b). "Neurons in the endbrain of numerically naive crows spontaneously encode visual numerosity". In: *Current Biology* 28.7, 1090–1094.e4.

DOI: 10.1016/j.cub.2018.02.023. URL: https://doi.org/10.1016/j.cub.2018.02.023.

Wandell, Brian A, Serge O Dumoulin, and Alyssa A Brewer (2007). "Visual field maps in human cortex". In: *Neuron* 56.2, pp. 366–383.

Wang, Haofan, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu (2020). "Score-CAM: Score-weighted visual explanations for convolutional neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 24–25.

Wang, Jielei, Ting Jiang, Zongyong Cui, and Zongjie Cao (2021). "Filter pruning with a feature map entropy importance criterion for convolution neural networks compressing". In: *Neurocomputing* 461, pp. 41–54.

Wang, Zhennan, Canqun Xiang, Wenbin Zou, and Chen Xu (2020). "Mma regularization: Decorrelating weights of neural networks by maximizing the minimal angles". In: *Advances in Neural Information Processing Systems* 33, pp. 19099–19110.

Weiner, Kevin S and Kalanit Grill-Spector (Oct. 2010). "Sparsely-distributed organization of face and limb activations in human ventral temporal cortex". en. In: *Neuroimage* 52.4, pp. 1559–1573.

Wen, Yandong, Kaipeng Zhang, Zhifeng Li, and Yu Qiao (2016). "A discriminative feature learning approach for deep face recognition". In: *European conference on computer vision*. Springer, pp. 499–515.

Wennberg, Ulme and Gustav Eje Henter (2024). "Exploring Internal Numeracy in Language Models: A Case Study on ALBERT". In: *arXiv preprint arXiv:2404.16574*.

Wichmann, Felix A and Robert Geirhos (2023). "Are deep neural networks adequate behavioral models of human visual perception?" In: *Annual review of vision science* 9.1, pp. 501–524.

Willshaw, David J and Christoph Von Der Malsburg (1976). "How patterned neural connections can be set up by self-organization". In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 194.1117, pp. 431–445.

Wong, YC, HC Kwan, WA MacKay, and JT Murphy (1978). "Spatial organization of precentral cortex in awake primates. I. Somatosensory inputs." In: *Journal of neurophysiology* 41.5, pp. 1107–1119.

Wurm, Moritz F and Alfonso Caramazza (Feb. 2022). "Two 'what' pathways for action and object recognition". en. In: *Trends Cogn. Sci.* 26.2, pp. 103–116.

Xie, Saining, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He (2017). "Aggregated residual transformations for deep neural networks". In: *Pro-*

ceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500.

Yamins, Daniel L K and James J DiCarlo (Feb. 2016). "Using goal-driven deep learning models to understand sensory cortex". In: *Nature Neuroscience* 19.3, pp. 356–365. DOI: 10.1038/nn.4244. URL: https://doi.org/10.1038/nn.4244.

Yamins, Daniel L. K., Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo (May 2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the National Academy of Sciences* 111.23, pp. 8619–8624. DOI: 10.1073/pnas.1403112111. URL: https://doi.org/10.1073/pnas.1403112111.

Yamins, Daniel LK and James J DiCarlo (2016). "Using goal-driven deep learning models to understand sensory cortex". In: *Nature neuroscience* 19.3, pp. 356–365.

Yang, Xintong, Ze Ji, Jing Wu, and Yu-Kun Lai (2023). "Recent advances of deep robotic affordance learning: a reinforcement learning perspective". In: *IEEE Transactions on Cognitive and Developmental Systems* 15.3, pp. 1139–1149.

Yargholi, Elahe' and Hans Op de Beeck (Apr. 2023). "Category trumps shape as an organizational principle of object space in the human occipitotemporal cortex". en. In: *J. Neurosci.* 43.16, pp. 2960–2972.

Yue, Xiaomin, Sophia Robert, and Leslie G Ungerleider (2020). "Curvature processing in human visual cortical areas". In: *NeuroImage* 222, p. 117295.

Yuste, Rafael (July 2015). "From the neuron doctrine to neural networks". In: *Nature reviews. Neuroscience* 16.8, pp. 487–497. DOI: 10.1038/nrn3962. URL: https://doi.org/10.1038/nrn3962.

Zbontar, Jure, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny (2021). "Barlow twins: Self-supervised learning via redundancy reduction". In: *International conference on machine learning*. PMLR, pp. 12310–12320.

Zeman, Astrid A, J Brendan Ritchie, Stefania Bracci, and Hans Op de Beeck (Feb. 2020). "Orthogonal representations of object shape and category in deep Convolutional Neural Networks and human visual cortex". en. In: *Sci. Rep.* 10.1, p. 2453.

Zhang, Dejiao, Haozhu Wang, Mario Figueiredo, and Laura Balzano (2018). "Learning to share: Simultaneous parameter tying and sparsification in deep learning". In: *International Conference on Learning Representations*.

Zhang, Huihuang, Haigen Hu, Deming Zhou, Xiaoqin Zhang, and Bin Cao (2025). "Compact CNN module balancing between feature diversity and redundancy". In: *Neural Networks* 188, p. 107456.

Zhang, Richard, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang (2018). "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595.

Zhang, Xi and Xiaolin Wu (Nov. 2020). *On numerosity of deep neural networks*. URL: https://arxiv.org/abs/2011.08674.

Zhang, Xin-Jie, Jack Murdoch Moore, Ting-Ting Gao, Xiaozhu Zhang, and Gang Yan (2025). "Brain-inspired wiring economics for artificial neural networks". In: *PNAS nexus* 4.1, pgae580.

Zhang, Yiyuan, Ke Zhou, Pinglei Bao, and Jia Liu (2024). "A biologically inspired computational model of human ventral temporal cortex". In: *Neural Networks* 178, p. 106437.

Zhao, Feifei and Yi Zeng (2021). "Dynamically optimizing network structure based on synaptic pruning in the brain". In: *Frontiers in Systems Neuroscience* 15, p. 620558.

Zhou, Deming, Yuetong Fang, Zhaorui Wang, and Renjing Xu (2025). "TD-SNNs: Competitive Topographic Deep Spiking Neural Networks for Visual Cortex Modeling". In: *arXiv preprint arXiv:2508.04270*.

Zhou, Qiongyi, Changde Du, and Huiguang He (2022). "Exploring the brain-like properties of deep neural networks: a neural encoding perspective". In: *Machine Intelligence Research* 19.5, pp. 439–455.

Zohary, Ehud, Michael N Shadlen, and William T Newsome (1994). "Correlated neuronal discharge rate and its implications for psychophysical performance". In: *Nature* 370.6485, pp. 140–143.

Appendix for Chapter 1

.0.1 Histogram of image-level correlations between Ecoset and ImageNet produced AIS maps

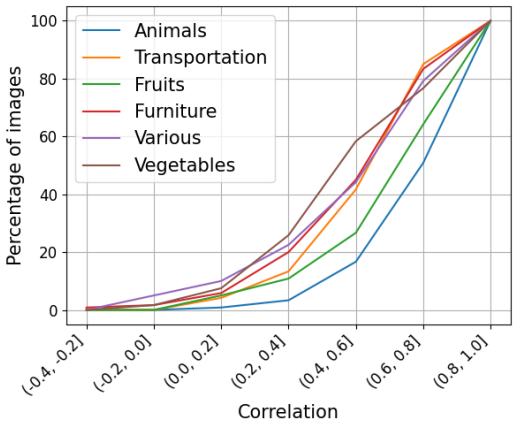


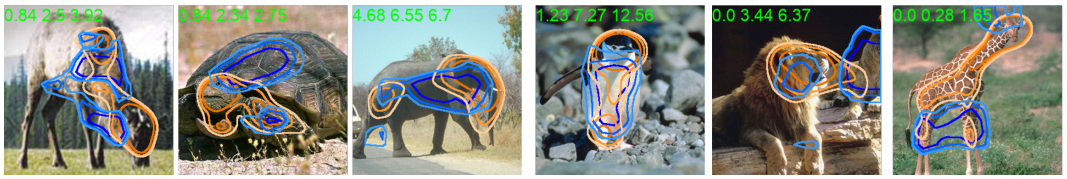
Figure 13: Cumulative histogram of correlations between heatmap' values created by ImageNet-trained and Ecoset-trained models.

.0.2 TranSalnet and AIS maps: Additional Images

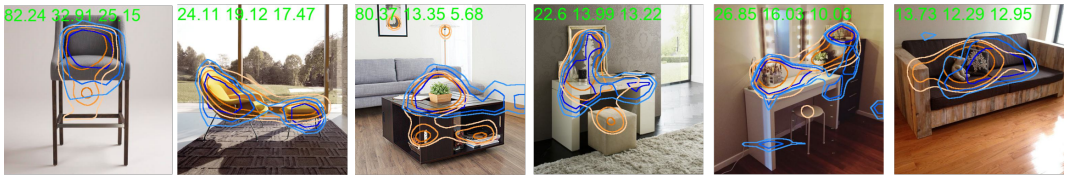
Additional images showing the overlap between the heatmaps created by Alignment Importance Scores (blue contours) and the saliency maps from TranSalNet (orange contours). Contours indicate the 5%, 10%, and 15% most important pixels, with increasing color intensity respectively. Relative Risk values computed from top 5%, 10% and 15% pixels in each map are printed on the top of each images.

Animals: High Relative Risk Ratio

Animals: Low Relative Risk Ratio

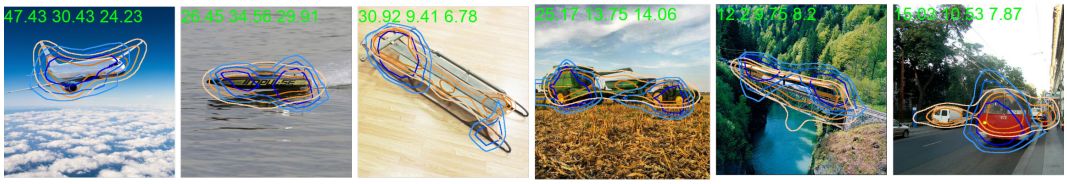


Furniture: High Relative Risk Ratio

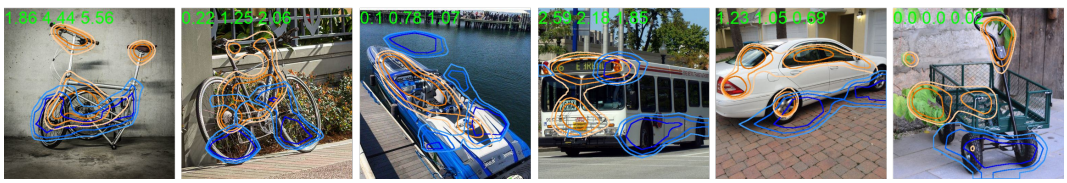


Furniture: Low Relative Risk Ratio

Transportation: High Relative Risk Ratio



Transportation: Low Relative Risk Ratio



.0.3 TranSalNet performance

The image, below, adapted from Lou et al. (2022) shows performance of Translanet in prediction of human gaze. The figure presents the original image, the human gaze location (Ground Truth), and the gaze predictions made by Translanet, when trained on two different vision models.

.0.4 Precision-Recall curves for Ecoset-produced images

See Figure 15.

.0.5 Production of second-order-isomorphism image-specific heatmaps in Tarigopula et al. (2023)

An image was masked using a sliding mask to evaluate how the masking of each image section impacted the 2OI between the DNN RDM and Brain RDM.

Figure 16 describes the main steps in the analysis. All 144 images in Set2 of King et al. (2019) were passed through a pruned DNN to extract embeddings. From these we constructed a baseline Representational Dissimilarity Matrix, (RDM_{DNN_base}). The correlation between RDM_{DNN_base} and RDM_{Brain} constituted ($2OI_{base}$). The masking procedure was applied to a target image and applied as follows. Masks were square 0-filters, and their sizes were set the range 24-56 pixels in intervals of 4 pixels (9 mask sizes in all). We used variable sizes to be sensitive to features of different granularity. The stride step size was set to 4 pixels for all filter sizes. We added zero padding to the edges of images as required depending on the size of each masking filter. As described in the main text, the perturbation to $2OI_{base}$ induced by each mask (computed as $2OI_{mask}$) was assigned to a 4×4 area at the center of the mask. This produced 8 perturbation values for the center of each set of 8 masks, of which we selected the value associated with the maximum absolute value (i.e., the negative or positive value that departed maximally from zero).

This entire procedure was applied to DNN embeddings extracted from VGG-19 DNNs whose embeddings were pruned as supervised by brain RDMs, or to embeddings derived from a non-pruned version of VGG-19 (internal control). In all cases we used a VGG-19 pretrained model as provided in Pytorch.

To visualize the perturbation scores we colored the 4×4 area in the center of each mask to avoid overlapping colors. Green colors denote a positive score, meaning the masking the given area produced a drop in 2OI, whereas red denotes the converse. Figure 17 presents more sample results.

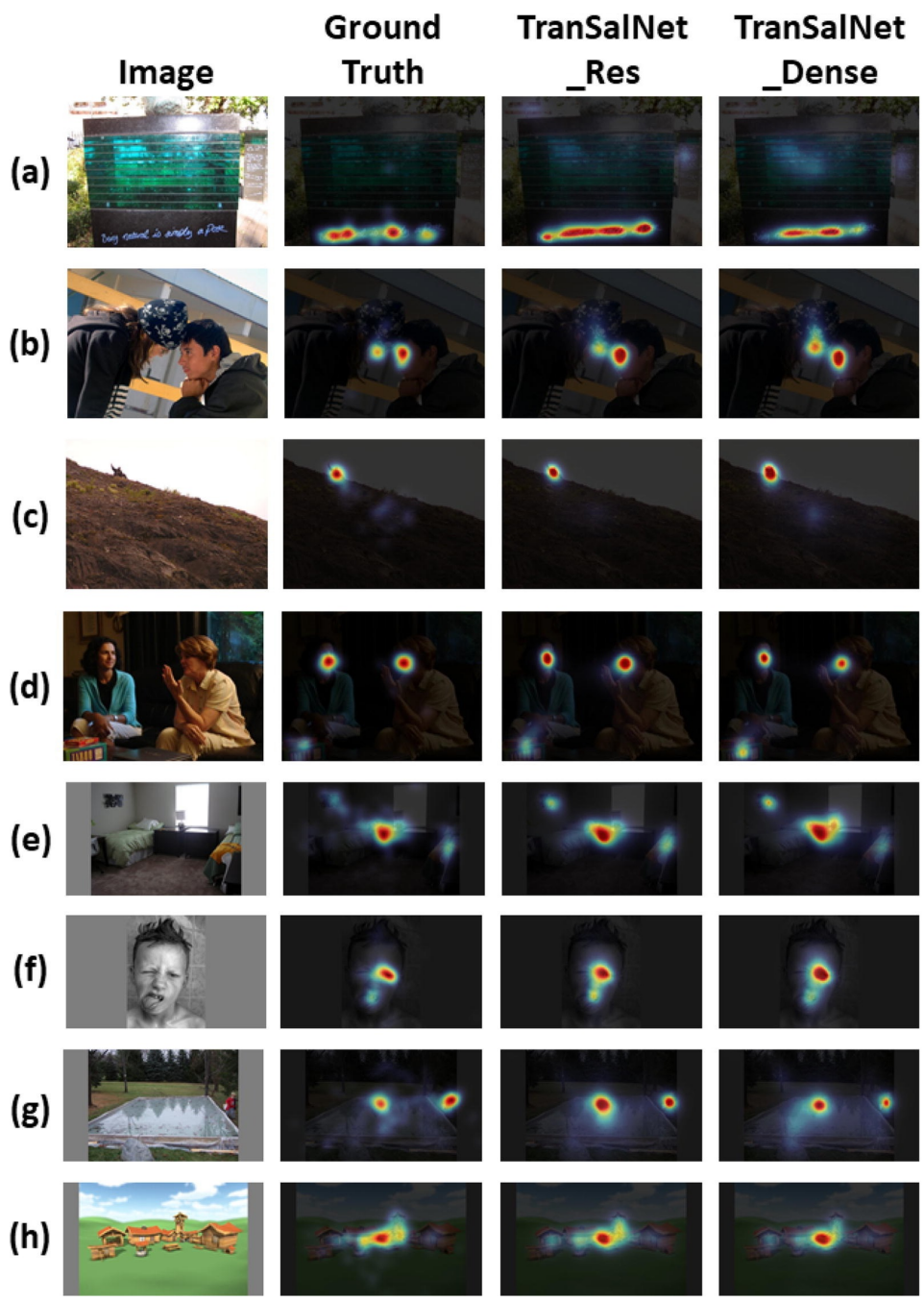
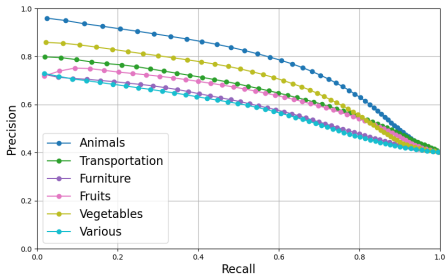
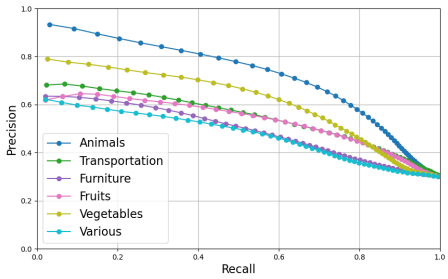


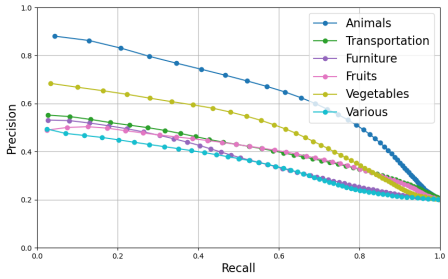
Figure 14: Figure adapted from Lou et al. (2022). https://doi.org/10.1016/j.neucom.2022.04.080. Original figure licensed CC-BY.



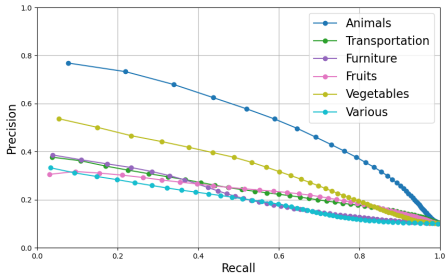
(a) Target thresholded at 60th percentile



(b) Target thresholded at 70th percentile



(c) Target thresholded at 80th percentile



(d) Target thresholded at 90th percentile

Figure 15: **Precision-Recall Curves for different thresholds of the target variable.** The target variable was heatmap values produced from AIS scores computed from Ecoset training. The predicting variable were saliency map values from obtained from TranSalNet.

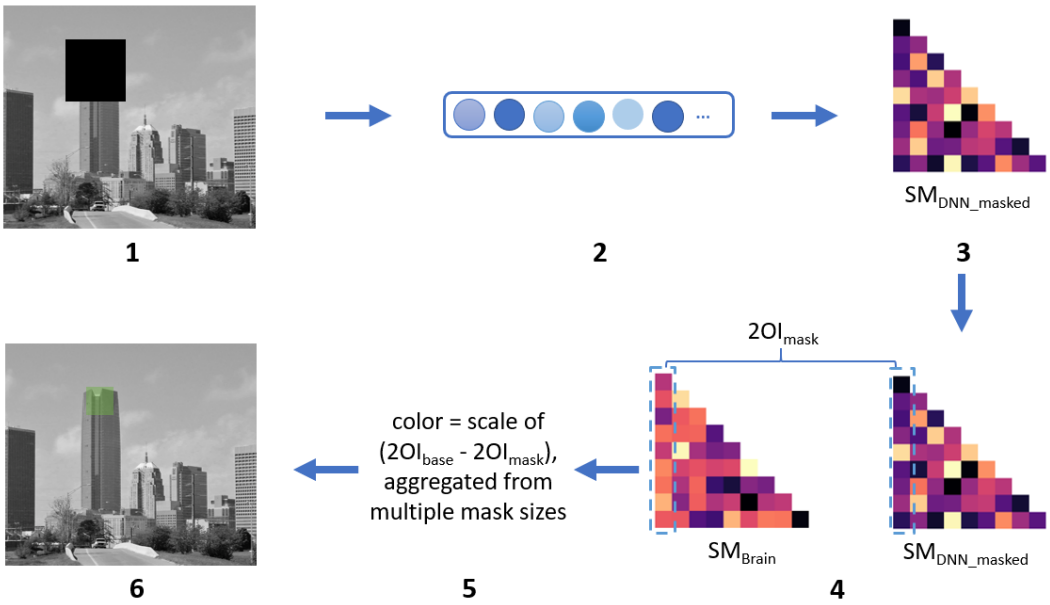


Figure 16: **Main steps in producing 2OI-perturbation heatmap in Tarigopula et al. (2023)**. 1. A section of the target image is masked. 2. The masked image is passed through the DNN and the image embeddings are extracted. 3. A Similarity Matrix (inverse version of RDM) is constructed to reflect the distance between the masked image and all other images (SM_{DNN_masked}); only correlations involving the target image are considered from this point on. 4. A 2OI value is computed by relating this set of correlation values to the set computed from brain data (SM_{Brain}). Those correlations involving the target image (here, e.g., Image 1) are delineated in the Figure by a dashed blue square. The two sets of correlation values are related via R^2 coefficient of determination. 5. The difference between $2OI_{mask}$ and $2OI_{base}$ is stored as the impact of the mask. 6. The magnitude of the difference is mapped onto a color scale.

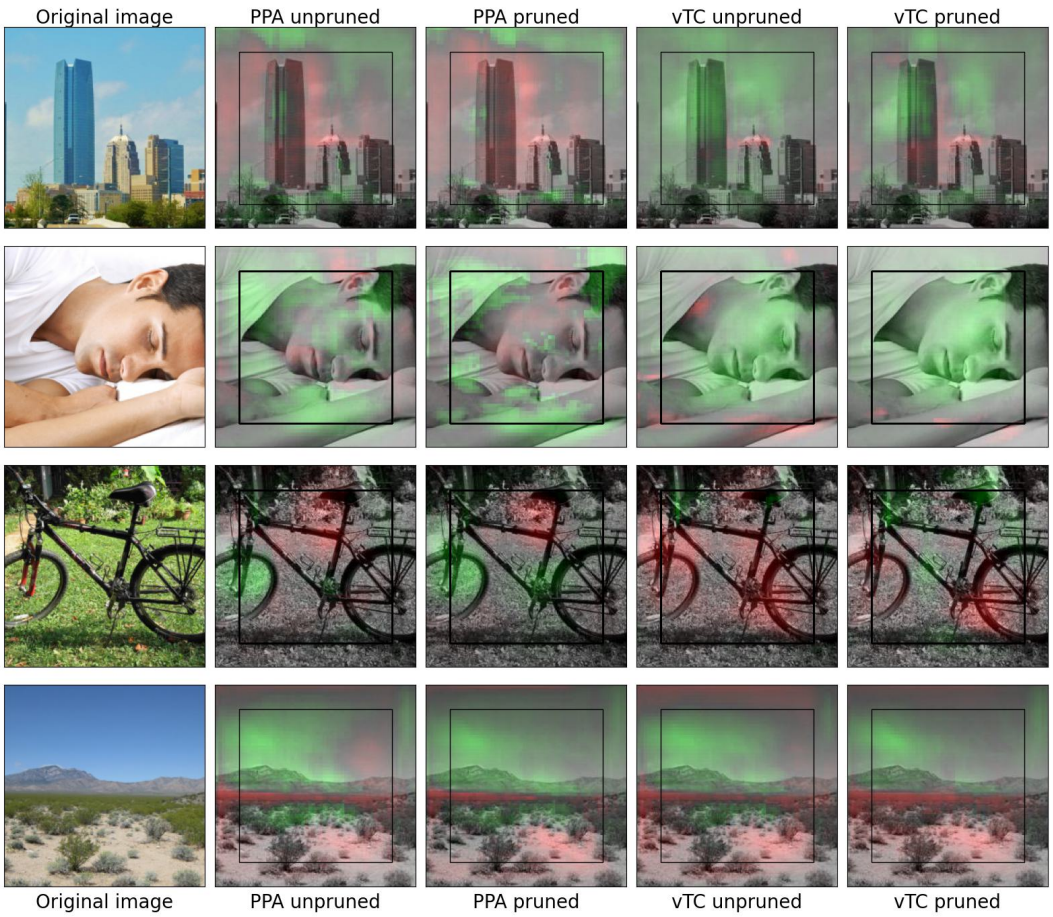


Figure 17: Additional sample heatmaps from Tarigopula et al. (2023) showing the contribution of each image section to second order isomorphism between a DNN RDM and a Brain RDM. More results can be downloaded from Github: *github.com/tlmnhut/Visualize_PrunedDNN_by_HumanSim/tree/main/results/grid*

Appendix for Chapter 5

.0.6 Training Dynamics

To evaluate how AS and WS regularization impacted training, we evaluated the cross-entropy loss and spatial-loss trajectory over the training epochs for $\lambda = 0.1$. Figure 18 shows the results for MNIST, and Figure 19 shows similar findings for CIFAR-10. For both MNIST and CIFAR-10, the trajectory of cross-entropy reduction (and accuracy) were highly similar, for all three types of models. This suggests that the different models learn the differentiation between classes at similar rates. With respect to the spatial loss terms, in MNIST, both AS and WS showed a strong increasing trend in a first few epochs, then the spatial losses dropped towards the end of training. In CIFAR-10, the AS-loss monotonically decreased, while WS-loss only decreased in the later half of the training process. For both datasets, accuracy remained at a relatively high level (50% of initial level) and began to slow down when training accuracy was already high. These data suggest that the inclusion of AS and WS regularization did not strongly impact the dynamics of classification accuracy or those of cross entropy loss during training. They also suggest that both AS and WS loss functions can produce an initial trade-off between the spatial and cross-entropy objectives, perhaps because the topographic regularization harms the feature learning required for classification. Once the features are learned, the spatial objective is more easily satisfied.

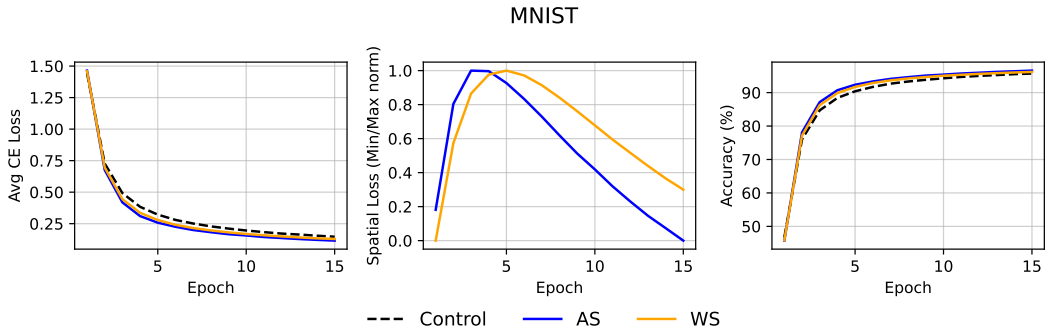


Figure 18: MNIST Train stats: Train-set accuracy and loss terms.

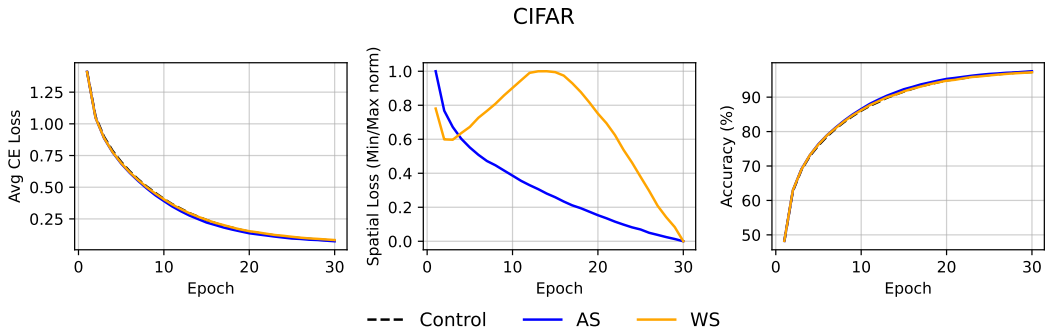


Figure 19: CIFAR-10 Train stats: Train-set accuracy and loss terms.

.0.7 Supplementary Figures

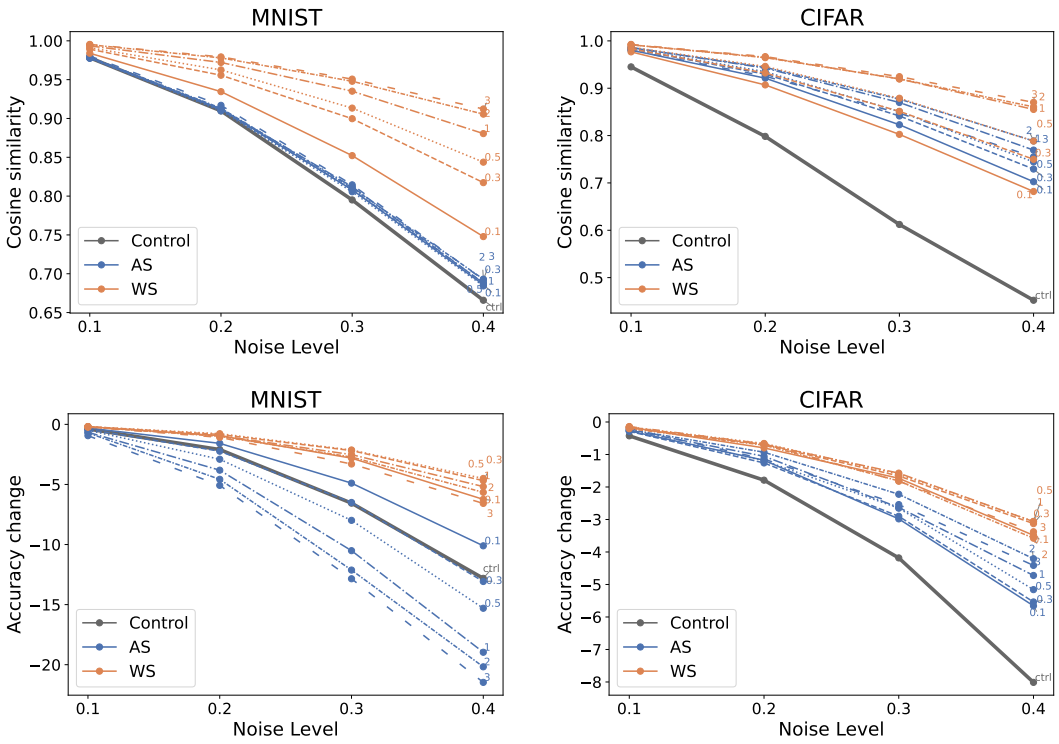
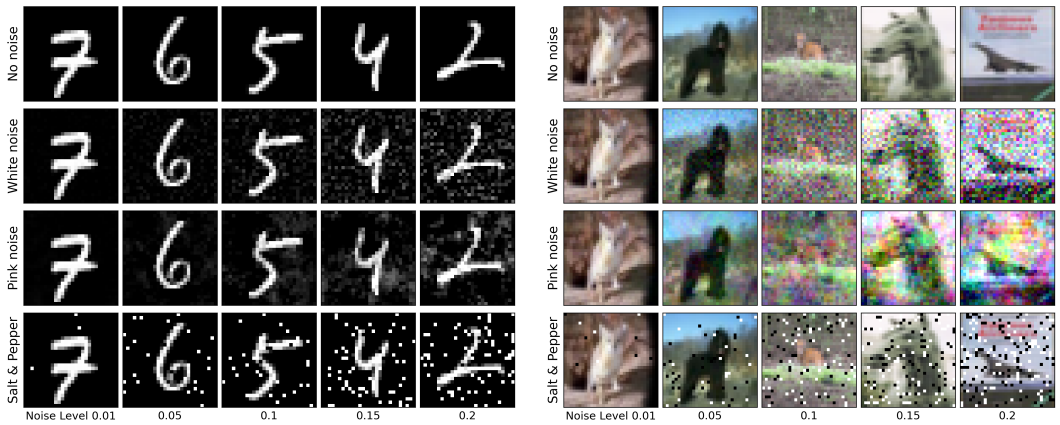


Figure 20: **Robustness to weight perturbations evaluated across individual values of λ.** Robustness is assessed by changes in representational geometry (top row) and test accuracy (bottom row) under increasing levels of additive noise applied to the final classification-layer weights. Representational geometry is defined as the 10×10 cosine-similarity matrix computed from category-level weight vectors, and robustness is quantified as the similarity between the original and perturbed matrices. Results are shown separately for MNIST (left) and CIFAR (right), and for control, AS, and WS models at individual values of the regularization parameter λ (indicated by line style).



(a) MNIST (b) CIFAR-10

Figure 21: **Examples of noise types and noise levels**. Representative MNIST (a) and CIFAR-10 (b) images are shown under different noise types (white noise, pink noise, and salt-and-pepper noise) and increasing noise levels. Rows: noise types; columns: noise level. For reference, noise-free inputs are presented in the top row.

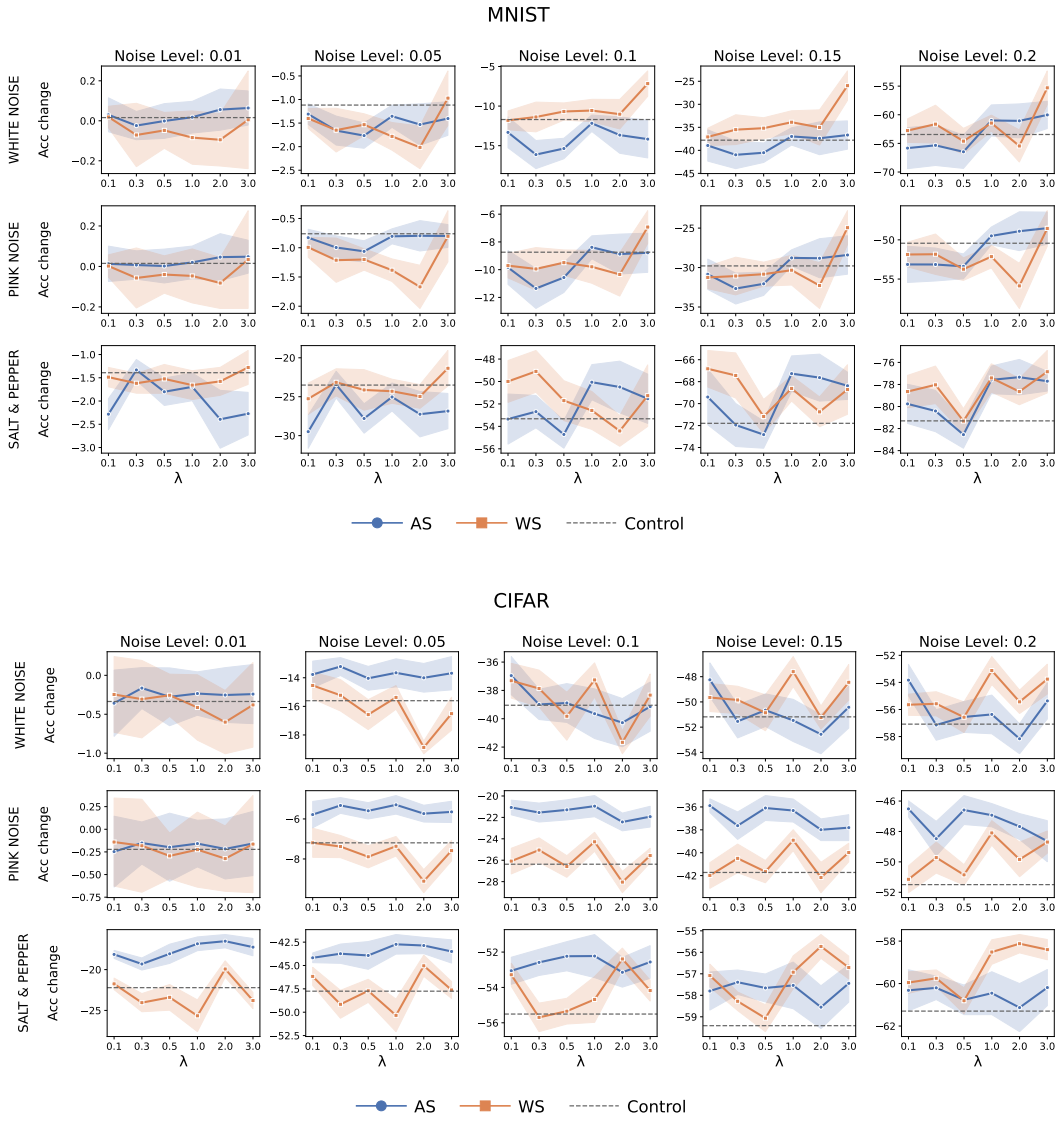


Figure 22: **Accuracy changes under input corruption across noise type, noise level, and topographic regularization**. Changes in classification accuracy compared to baseline (non-perturbed) models are shown as a function of λ for models trained with AS, WS, and control objectives. Results are shown for MNIST (top) and CIFAR (bottom), for three types of corruption (white noise, pink noise, and salt-and-pepper; rows), and for increasing noise levels (columns). Accuracy values are compared to the corresponding noise-free baseline (accuracy after adding noises minuses before adding). Shaded regions indicate $\pm s.e.m$ across runs.

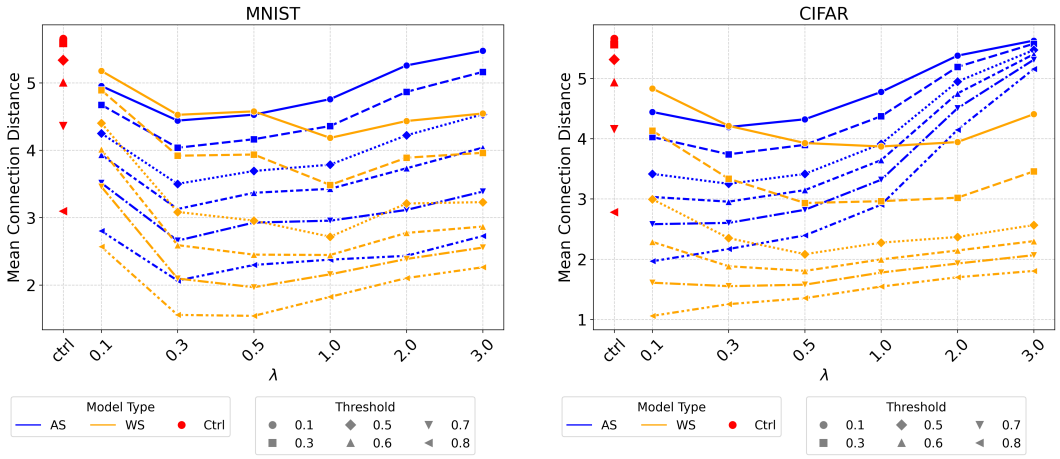


Figure 23: **Correlated-unit distances across individual correlation thresholds**. Mean spatial distance between pairs of units whose activity correlation exceeds a threshold $\alpha \in \{0.1, 0.3, 0.5, 0.6, 0.7, 0.8\}$ is shown for MNIST (left) and CIFAR-10 (right). Separate curves corresponding to individual correlation thresholds.

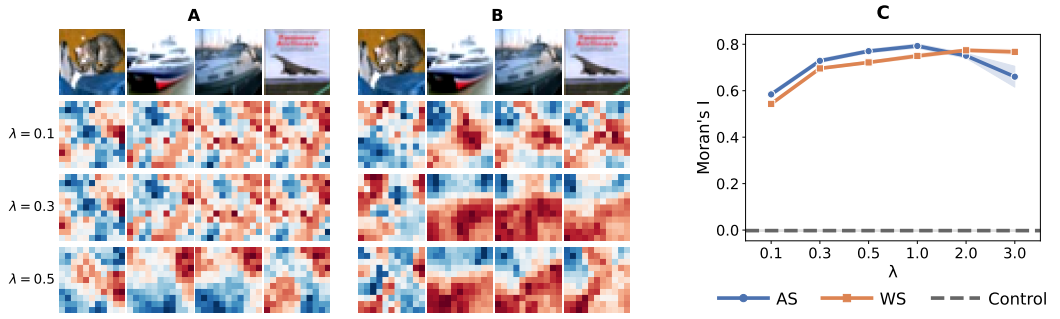


Figure 24: **Spatial smoothness of activation maps under activation- and weight-similarity training for CIFAR-10**.

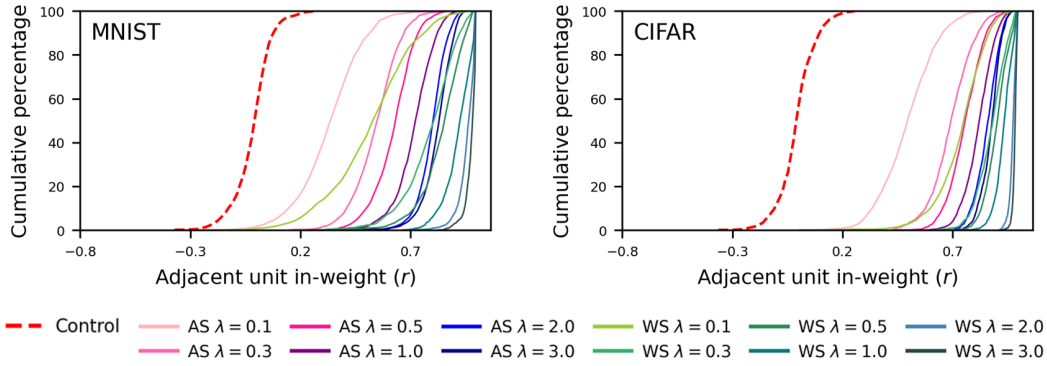


Figure 25: **Incoming weight correlations between adjacent units**. Cumulative distributions of average pairwise correlations between incoming weight vectors of adjacent units. For each unit, the mean correlation is computed form all adjacent neighbors. Distributions are shown for control, AS, and WS models across values of the regularization parameter λ.

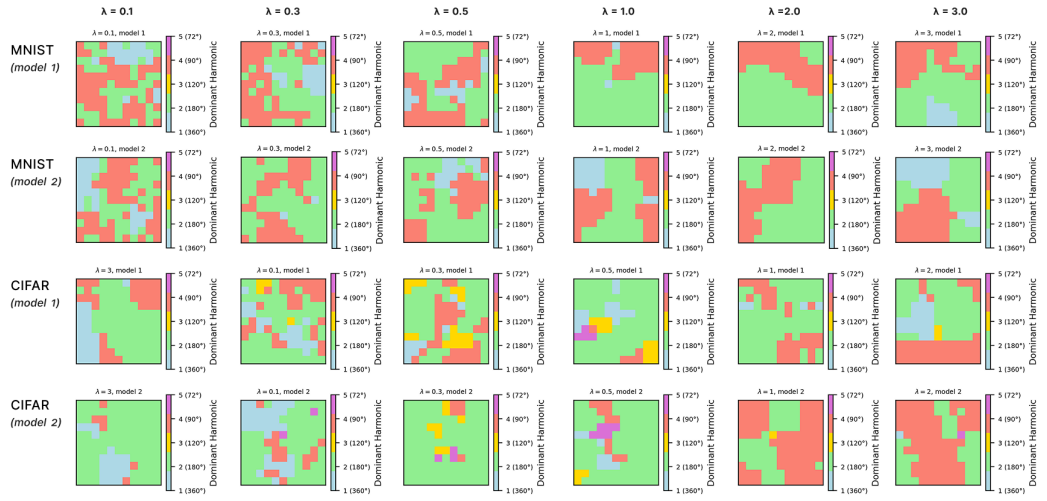


Figure 26: **Angular response properties under weight-similarity training**. Topographic maps show the dominant angular response type for units in the grid across values of the regularization parameter λ under WS training. Angular response types correspond to five harmonic components with periodicities of 360°, 180°, 120°, 90°, and 72° (Cycles 1–5), indicated by color. Results are shown for MNIST and CIFAR-10. Two randomly trained models shown per dataset. Columns correspond to values of λ.

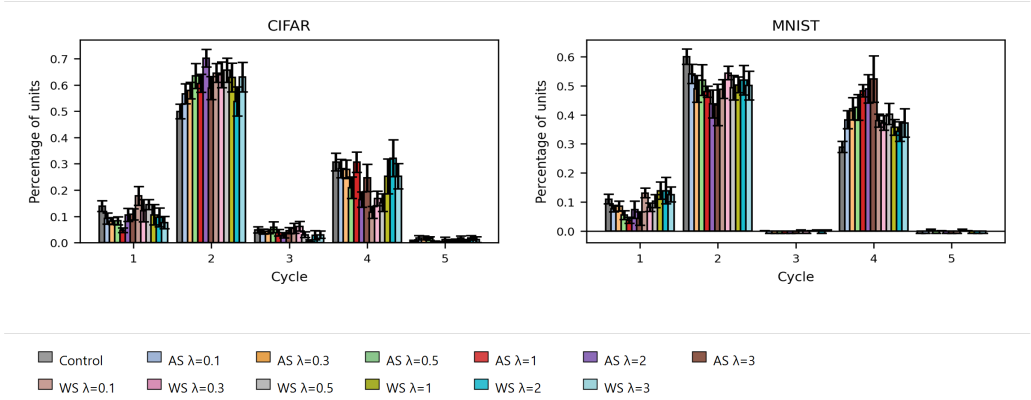


Figure 27: **Angular tuning profiles across training conditions and regularization strengths** (λ). Percentage of units assigned to each angular tracking profile (Cycles 1–5) is shown. Cycles correspond to harmonic response types with periodicities of 360°, 180°, 120°, 90°, and 72°, respectively. Within each cycle, bar order is control, followed AS and WS, at increasing values of λ. Bars indicate mean percentages across runs; error bars denote variability across runs.

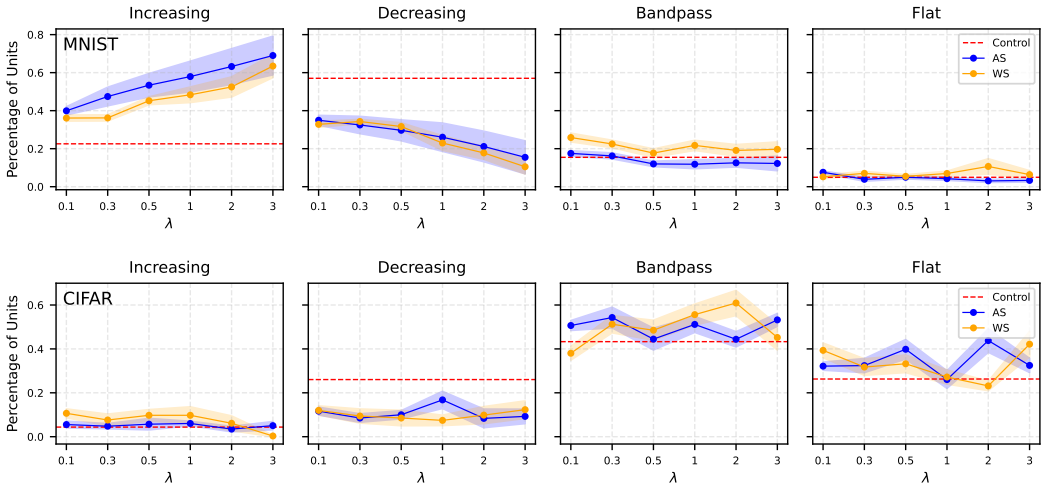


Figure 28: **Eccentricity tuning profiles across training conditions and regularization strengths**. Percentage of units showing increasing, decreasing, band-pass, or flat eccentricity tuning profiles. Each panel corresponds to one eccentricity tuning profile, as defined in the main text. *Increasing* and *decreasing* profiles correspond to monotonic changes in response magnitude with eccentricity and reflect units with positive or negative radial gain, respectively.