A contrastive coding account of category selectivity in the ventral visual stream

Jacob S. Prince [1*], George A. Alvarez [1], & Talia Konkle [1,2,3]

[1] *Department of Psychology, Harvard University, USA*
[2] *Center for Brain Science, Harvard University, USA*
[3] *Kempner Institute for Biological and Artificial Intelligence, Harvard University, USA*

* Corresponding author (jacob.samuel.prince@gmail.com)

ABSTRACT

Modular and distributed theories of category selectivity along the ventral visual stream have long existed in tension. Here, we present a reconciling framework, based on a series of analyses relating category-selective tuning within biological and artificial neural networks. We discover that, in models trained with contrastive self-supervised objectives over a rich natural image diet, visual category-selective tuning naturally emerges for classic categories of faces, bodies, scenes, and words. Further, lesions of these model units lead to selective, dissociable recognition deficits. Finally, these pre-identified units from a single model can predict neural responses in all corresponding face-, scene-, body-, and word-selective regions of the human visual system, even under a constrained sparse-positive encoding procedure. The success of this model indicates that the nature of category-selective tuning in the human brain (e.g. for faces or scenes) is dependent on the learning diet, and best understood in the context of the full scope of experienced visual input. Broadly, we offer a unifying theoretical account where category-selective tuning naturally emerges as a consequence of positive information routing through hierarchical population codes, in order to disentangle the statistics of visual experience.

Keywords: category selectivity, contrastive learning, feature tuning, functional specialization

INTRODUCTION

Our ability to see and recognize objects, people, and the broader environment around us is supported by representations along the ventral visual stream (1–6). Foundational discoveries have charted regions with category-selective tuning, evident for a few categories: faces (7–10), bodies (11, 12), scenes (13–15), and visually presented words (16–18). That is, recorded responses in single units and voxels respond systematically more on average to the neurons' preferred category, (e.g. images of faces), with only weak responses to other categories. The properties of these regions and the pressures guiding their emergence has been the subject of intense study for decades (19–31). What is the nature of the tuning in these regions that supports the selective responses across images? And, are these category-selective regions better understood distinctly from each other and the representation of other categories, or as interrelated regions that are part of a broader visual representational scheme?

A prominent view of these brain regions is as distinctive, independent, functional modules (3, 19, 32–35). Category-selective regions only exist for a few domains and are thus hypothesized to be different from other visual categories in some important way. For example, for faces, bodies, and scenes:

these categories may have particular ecological relevance and may emerge through built-in specialized architectures, supporting domain-specialized tasks such as face individuation (20, 36–38). For the case of word-selectivity, perhaps extensive visual training and a need to interface with long-range connectivity to language regions leads to the emergence of regions such as the visual word-form area (39–41). Brain regions with selectivities for different categories are typically studied in depth by different communities, and are assumed to have very different kinds of visual feature tuning from one another, related purely to their distinct domains (e.g. face-specialized, body-specialized, scene-specialized, letter-specialized). Indeed, some of the strongest evidence for the modularity of these regions comes from selective deficits following brain damage (42–49) or other forms of causal perturbation such as transcranial magnetic stimulation (50–52), electrical microstimulation (53), and pharmacological inactivation (54); see 55 for review). In this way, modular frameworks specify a 'labeled-line' type of visual representation (56), where, for example, a neuron's highly selective firing rates to face images is a direct indicator of its functional role in face processing.

However, modular accounts are restricted to explaining the representation of a few specific visual categories. Empirical evidence shows that not all visual categories have clustered areas of highly selective tuning along the ventral stream (57–59), leaving unspecified how all other categories are represented. Distributed coding accounts are the prominent alternative framework, where the relevant category information is assumed to be encoded in the activation pattern of the large-scale neural population (2, 60–62). Under this view, across the population of neurons, both high and low responses are critical to signal category information (much like how both 1's and 0's encode information in a binary code). These accounts sometimes assume the underlying visual feature tuning of units in object-responsive cortex is either fairly generic (e.g. geons or other building-block-like contours that apply to many categories to some degree; 63), or, that the tuning of any one unit may not be particularly interpretable without the context of the full population code (5). This latter perspective is a key element in untangling theories, which tend to de-emphasize the interpretation of the response properties of local units or features, in favor of studying the broader geometry of the population (64). The strongest empirical evidence for a distributed representational account comes from the pervasive category information present across regions of high-level visual cortex (2, 59, 61, 65–67). For example, both face and non-face categories can be decoded from both face-selective regions and non-face selective regions (2, 68). However, it is debated whether this extensive decodable information is all functionally relevant for recognition behavior (69–72).

These modular and distributed representational accounts have continued to develop in parallel, each contributing to a set of empirical puzzle pieces that constrain our understanding of category selectivity along ventral visual cortex. To date, it remains unclear the extent to which these theories are compatible or opposed. For example, one possibility is that there may simply be two modes of representation in the visual system, where a few special categories have more domain-specialized tuning and localist readout schemes, while all others have more generic tuning with accompanying distributed readout schemes. However, further pieces of empirical and computational evidence hint at a deeper unifying relationship between categories with selective regions and the representation of other objects – for example, based on their systematic topographic relationships on the cortical surface (23, 25, 28, 73–75). Along these lines, here we offer an updated account of category-selective regions, which provides specific insight into both the nature of their feature tuning and their function in the context of information readout.

To do so, we leverage a particular kind of deep convolutional neural network model (DNN) to operationalize this unifying perspective. Specifically, we used self-supervised instance-level contrastive learning objectives (76, 77), to train a standard DNN architecture over a large set of natural images (78, 79). The contrastive learning objective does not prioritize any special categories over other ob-

ject categories, nor does it even presuppose object categories at all. Instead, these self-supervised contrastive models learn to represent every experienced image distinctly from every other image in an embedding space (while being similar to itself under some transformations; see also 80–83). As a consequence, these models develop emergent category-level representations: images with similar visual characteristics tend to come from similar categories, and thus are naturally routed through the hierarchy to nearby locations in the representational space.

A key property of these contrastive models, relevant for the work here, is that the nature of the learned features is intrinsically *diet-dependent*. For example, training a contrastive model that experiences only images of faces will learn feature tuning that aims to discriminate among face image content. Training the model over a richer visual diet, like the ImageNet dataset used here, will provide learned features that aim to discriminate among all 1.2M images (e.g. 80). In other words, the set of images is critical for determining the nature of the feature tuning. Another important property of these models is that the set of units within a layer, as a whole, must jointly represent the entire input space. Thus, the nature of the feature tuning of any one unit is not only diet-dependent, but also related to the feature tuning of the other units in the layer. In these ways, the feature tuning of each unit in these contrastive networks is meaningfully related to both the scope of the visual input and the tuning of other units within and across layers.

The aim of the present work is to leverage these contrastive models to provide a possible computational explanation of the emergence and function of category-selective tuning. We show that a contrastive DNN model has emergent category-selective units, which lead to selective and dissociable recognition deficits when lesioned, and which can predict the high-dimensional response structure of diverse category-selective regions in the human brain. We further introduce a novel, sparse positive weighted voxel-encoding scheme, which reflects a more constrained linking procedure between biological and artificial neural network responses, under the hypothesis that the tuning direction of single model units (and neurons) is key for signaling image content. Broadly, we argue that category-selective regions are facets of a rich, diet-dependent, contrastive feature space. To propose a mechanistic account for these signatures, we introduce the concept of positive routing through hierarchical population codes, where units with different tuning serve to channel different content through successive stages into an increasingly untangled latent space.

RESULTS

Category-selective tuning emerges in models without category-specialized mechanisms

We first examined whether a contrastive DNN trained without any category-specific architectural motifs or task objectives would show emergent category-selective signatures that mirror those observed in the human ventral visual stream. Note that our use of the term 'category selectivity' throughout is specifically referring to the categories (domains) of faces, bodies, scenes, and visual word forms, following the well-characterized category-selective regions of the ventral stream.

We used a popular self-supervised learning objective (Barlow Twins; 76), to train a standard deep convolutional neural network model architecture (AlexNet; 78) using the ImageNet dataset (79). Barlow

Twins attempts to minimize the difference in latent space between different augmentations of an image, outputting a high-dimensional embedding of maximally independent dimensions[1] (see *Methods*).

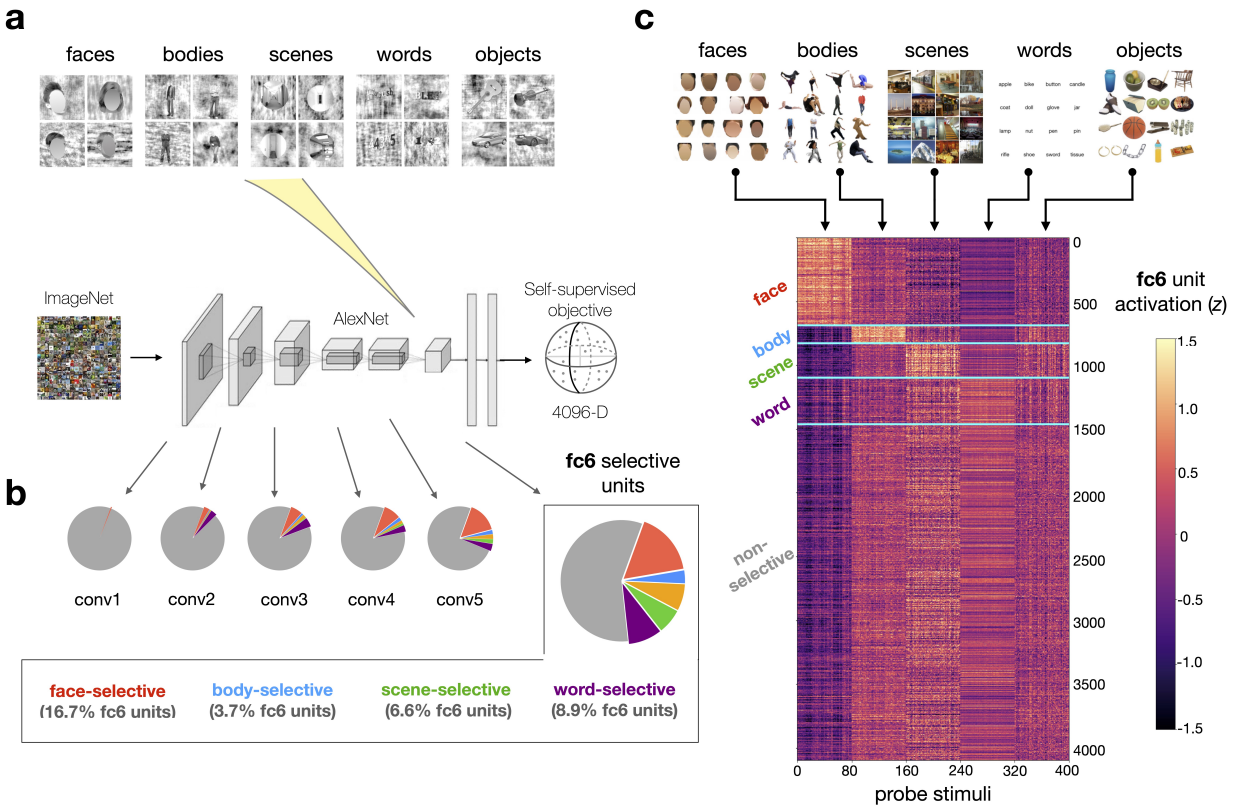


Figure 1: Emergent category selectivity in a self-supervised DNN. *(A) Localizer procedure for identifying category-selective units in the self-supervised model. (B) Pie charts summarize proportions of model units selective for faces, bodies, objects, scenes, and words in model layers conv1-5 and fc6 (enlarged). (C) Responses in model layer fc6 to the RGB probe set of face, body, scene, word, and object stimuli. N = 80 images per domain. Layer fc6 units (y axis) are sorted according to their selectivities as defined using the independent localizer stimuli. The 400-dimensional activation profile of each unit is z-scored for purposes of visualization. Images containing identifiable human faces are masked.*

To test for emergent category-selective tuning in the self-supervised model, we designed procedures to mirror the localization of category-selective regions in the human ventral visual stream in typical fMRI experiments. Specifically, we recorded the activations of every model unit in response to the same localizer image set (87) used to identify category-selective regions in the Natural Scenes Dataset (88), and then applied a statistical contrast to identify non-overlapping sets of model units that were selective for faces, bodies, scenes, and words (**Figure 1A**; see *Methods*).

With this localizer procedure, we were able to identify small groups of category-selective units in each layer with either robust face-selectivity, place-selectivity, body-selectivity, or word-selectivity (FDR-corrected $p < 0.05$; mean unit t-value for faces $= 9.29 \pm 4.47$ SD; scenes = 8.86 ± 3.34; bodies = 7.38 ± 2.17; words = 8.12 ± 2.74). The overall proportions of selective units tended to increase at

[1] Although the Barlow Twins objective has been described as energy- or covariance-based rather than contrastive (84), the objective function emphasizes distinctions across dimensions within a batch, and is thus contrastive with respect to dimensions of encoding. Indeed, the equivalences between these classes of self-supervised algorithms have been extensively validated by studying the properties of their gradients (85) and their generalization behavior (86). For our purposes, the important point is that the Barlow Twins algorithm yields representations that distinguish between instances.

progressively deeper layers (**Supplementary Figure 1**). Further, note that even in a single layer (e.g. fc6) it was possible to find different units with selectivity for each domain (**Figure 1B**). Thus, distinct category-selective units emerge in the context of a single population code, as operationalized by a hierarchical layer.

We next tested whether the units' selectivities generalized to an alternate standard localizer image set. This independent probe set consisted of 400 total images of faces, bodies, scenes, words, and objects, and differed in both low-level image statistics (e.g. color instead of grayscale) and high-level image content (e.g. containing diverse object types instead of only cars and guitars in the 'object' category). We observed that the selective units maintained their high responses to preferred stimuli (**Figure 1C**), suggesting that their category-selective tuning is robust and not dependent on a specific probe image set, as is also the case in the human brain. Thus, emergent category-selective tuning is found in a self-supervised DNN model.

To test for the generality of this effect, we repeated these analyses on an AlexNet model trained with a different self-supervised learning objective (Instance-Prototype Contrastive learning – IPCL; 77; see *Methods*). We again observed similar units with emergent category-selective tuning (**Supplementary Figure 1**).[2]

Finally, prior reports have indicated that units with face-selective tuning can emerge even in untrained networks (89). We also identified category-selective units in an untrained AlexNet architecture. However, we found that there were substantially fewer than in trained networks (e.g. 14.6% vs. 35.7% of layer fc6 units selective for faces, bodies, scenes, or words in untrained vs. trained models), the strength of these units' selectivities was far weaker (e.g. fc6 mean t-value across domains = 6.6 vs. 13.4, **Supplementary Figure 2A**), and their response properties did not generalize to the independent probe set (**Supplementary Figure 2B**). These results jointly suggest that training (i.e. visual experience) is necessary for DNN features to develop reliable selectivities for these high-level domains of interest.

In summary, the fMRI-inspired localizer procedure was able to robustly identify category-selective DNN units in these self-supervised contrastive models. These results demonstrate that units with category-selective tuning to faces, bodies, scenes, and words can emerge as a general consequence of contrastive learning over a rich natural image diet (as reflected in the ImageNet stimulus set). We emphasize that no categorical, semantic or other domain-specific learning pressures were involved in training these models.

Lesioning category-selective units yields dissociable, predictable deficits

Next we examined the functional role these category-selective DNN units have for image recognition. Neurophysiological evidence has shown dissociable recognition deficits arising from perturbation of the ventral visual stream (e.g. 44, 52, 53). If these units are acting as a functional module in the DNN, lesioning the units with face-selectivity (i.e. setting their activations to 0) should yield a selective deficit for recognition of face-like categories. Similarly, lesioning the scene-selective units should yield selective deficits for more scene-like categories, and so forth. And critically, these deficits should be dissociable. Alternatively, these category-selective units may be tuned arbitrarily in the layer's feature space, and bear no consistent functional role for the object categories that happen to drive high, selective responses.

[2]Note that 1000-way object categorization tasks are also richly discriminative, or contrastive, in nature. That is, they require learning features to support many-way fine-grained object discrimination. We observed that ImageNet-trained category-supervised models also showed emergent category-selective tuning; **Supplementary Figure 1**.

To explore the impact of lesions of category-selective model units, we first needed to instantiate a readout mechanism. Note that the self-supervised model is trained only to learn features which discriminate any image view from any other view. How is category information read out from this rich visual feature space? Traditional practices and theoretical frameworks focus on linear separability in the full population code, instantiated through fully-connected readout weights learned for each category (a distributed readout scheme). However, here, we constrained this procedure by adding sparsity regularization (L1 penalty) to the readout weights. This approach operationalizes a view where category information can be accessed without requiring connections to all other neurons. We offer that this readout method provides a more theoretically constrained and biologically plausible connectivity motif (see *General Discussion*). Note that all results below hold with the standard, more flexible fully-connected readout method.

A regularized linear readout function was trained for 10 epochs on top of the penultimate layer embedding (relu7; frozen backbone), and we measured the baseline recognition accuracy for each of the 1000 ImageNet categories (see **Figure 2A**). Averaging across categories, the mean top-5 classification accuracy was 61.5% $\pm$ 20.3% SD. The learned readout weights were extremely sparse (85.4% of weights with absolute magnitude < 0.001, versus 5.4% from unregularized readout with the same hyperparameters). Next, we lesioned all of the face-selective units across the model's relu layers (or body-selective units, or scene-selective units, etc.), and measured the change in recognition accuracy for each ImageNet category after each type of lesion.

Example ImageNet categories with the greatest recognition deficits from each type of lesion are shown in **Figure 2B**. Lesions to face-selective units led to the strongest deficits for several dog breed categories (Δ=64-80%; mean 4.3% units/layer lesioned); body-selective units lesions led to the strongest deficits for animals and articles of clothing (Δ=28-36%; mean 0.8% units/layer lesioned); scene-selective unit lesions led to deficits in recognizing scene categories and big object categories (Δ=64-80%; mean 1.3% units/layer lesioned); and word-selective unit lesions led to deficits in recognizing insect categories (Δ=40-48%; mean 1.7% units/layer lesioned). Note that there are no explicit 'face', 'body', 'word', or 'scene' categories in the 1000 categories of ImageNet, but these impaired categories are broadly consistent with expectations from the literature (e.g. animal images elicit a moderate to high response in face-selective regions, and big object categories elicit moderate to high responses in scene-selective regions; 3, 27, 90).

Are these recognition deficits selective and dissociable? In our next analysis, we identified the top 10 categories most impacted by each type of lesion using one half of the ImageNet validation set images, and then re-computed the recognition deficits using the independent half of the validation set. **Figure 2C** shows that categories impaired by one type of lesion were largely spared with other types of lesions. **Supplementary Figure 3A** shows this analysis for increasing numbers of categories, where the same trends hold. To quantify the dissociability of lesioning deficits across all ImageNet categories, we directly correlated the 1000-category cost profiles (i.e. the differences between unlesioned and lesioned top-5 accuracies) between all pairs of lesions. The cost profiles were broadly unrelated, and in some cases negatively related: as examples, the face vs. scene cost profile similarity was $r = -0.21$, face vs. word $r = -0.08$, and the average r between domain pairs $= 0.02 \pm 0.13$ SD (see **Supplementary Figure 3B-C**). Thus, these category-selective units across the model layers have different causal involvement in the recognition of different categories.

Are these recognition deficits predictable? Drawing on modular frameworks, we next tested the idea that the information-carrying signal in the network is related to high activation levels. For example, are the ImageNet categories that most activate the face-selective units also the hardest to recognize after lesioning? We tested for this relationship across all types of lesions, plotted in **Figure 2D**. Indeed,

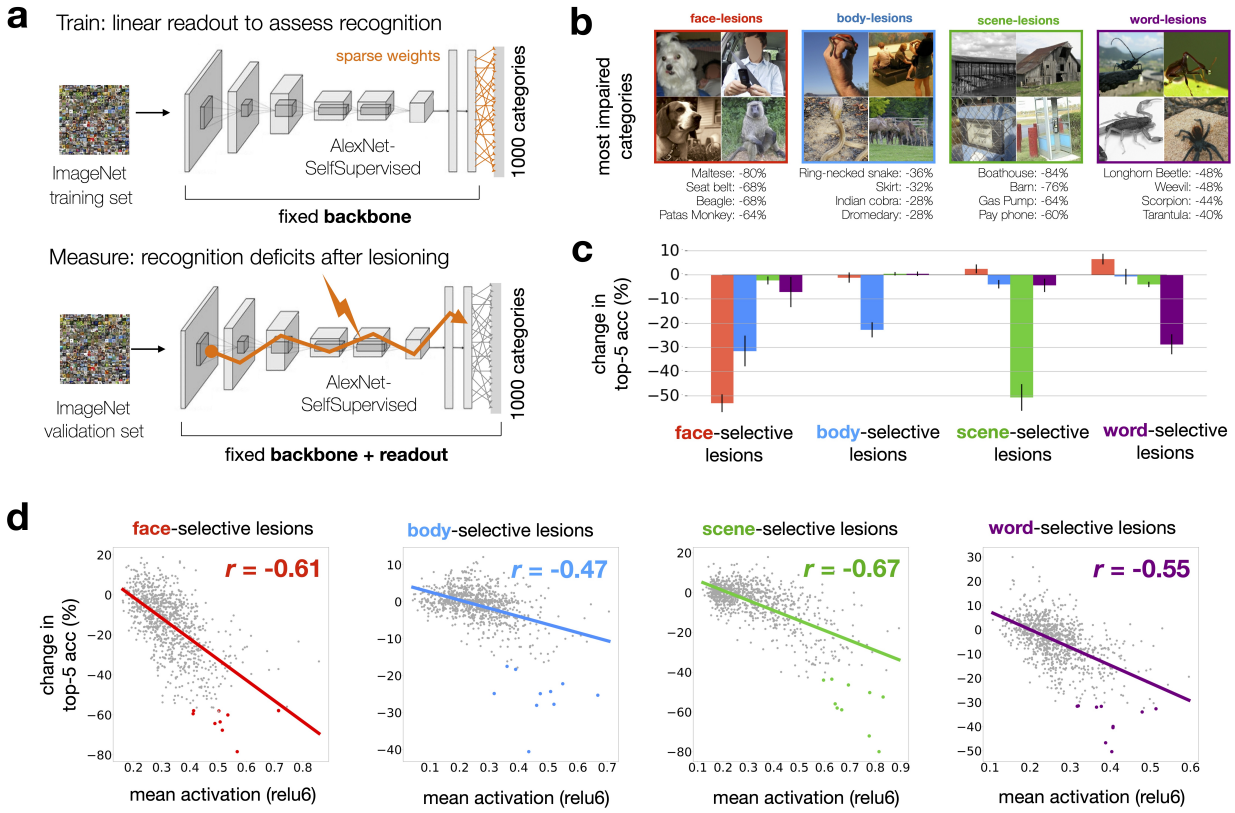


Figure 2: Impact of lesioning category-selective units on object recognition behavior. *(A) Procedure for measuring object recognition deficits from lesions to category-selective units in the self-supervised model. Baseline top-5 recognition accuracy is assessed via sparse readout from relu7 features. Recognition deficits reflect the drop in category accuracy following lesions to category-selective units from each domain. (B) Top 4 categories with greatest recognition deficits are shown for lesions to face-, scene-, body-, and word-selective units. Plotted exemplars are chosen from the ImageNet validation set and labeled with drop in top-5 accuracy between baseline and lesioned validation passes. (C) Bar graphs show the mean (± SEM) change in top-5 accuracy over the top 10 categories most impacted by lesions to the selective units of each domain. (D) Relationship between category activation and lesioning cost, for face-, body-, scene-, and word-selective units. Dots reflect the 1000 ImageNet categories. The x axis shows mean activation levels for each category (layer relu6 plotted) from the un-lesioned model. The y axis shows change in top-5 accuracy after lesioning. Colored points reflect the top 10 categories most impacted by lesions to each domain. Images containing identifiable human faces are masked.*

we observed consistent negative correlations (mean $r = -0.57 \pm 0.08$ SD) – the more an ImageNet category activates a group of category-selective units, the more difficult it is to recognize that category after lesioning. The strength and consistency of this relationship implies that positive activation magnitude is a reasonably reliable indicator of functional relevance.

As a control analysis, we tested the impact of random lesions. The entire procedure was repeated, this time targeting randomized indices of units that were numerically matched to the size of category-selective subsets in each relu layer. In this regime, we no longer observed meaningful variation in the degree of recognition deficits across categories. As a result, the correlations between activation magnitude and subsequent deficits after lesioning dropped to near zero (**Supplementary Figure 4B**). Thus, systematic deficits for some categories over others only occur when lesioning cohesive sets of category-tuned units across the layers.

These lesion experiments have two key implications. First, they reveal a link between a unit's activation magnitude and its functional relevance in recognition behavior. We hypothesize that this is partially a consequence of the relu nonlinearity, where information propagates from layer to layer through positive activation only. Second, these results demonstrate that functionally dissociable deficits can occur even with a domain-general contrastive objective, and within a general architecture (with no pre-specified modular pathways or branches). The results instead imply that during training, the model weights effectively form separable routes through the layer hierarchy, directing different kinds of image content to separable locations in the final embedding space.

Linking model and brain tuning with sparse positive encoding constraints

All experiments conducted so far have focused on a DNN model, identifying units with category-selective tuning, and charting their dissociable functional roles. How similar are these category-selective model units to category-selective regions of the brain? We next test the hypothesis that these pre-identified sets of model units from a single contrastive model can capture the response profiles and representational geometries of the classic category selective regions. This departs from many prior approaches, which typically focus on only one category (e.g. only faces, only scenes), often using category-specialized feature models (e.g. models that apply only to faces; 91–94), and, from approaches that train separate end-to-end DNN models to predict each brain ROI independently (30). Instead, here we test the hypothesis that the nature of the feature tuning across all category-selective regions can be understood jointly, within a single model, as a consequence of contrastive feature learning over a rich image diet, in order to distinguish all kinds of visual content.

While most standard approaches model each brain voxel as a linear combination of *all* features from a DNN layer, we instead fit encoding models for each category-selective region using *only* the subsets of DNN units with corresponding selectivity (e.g. only face-selective units are used for modeling FFA and OFA, **Figure 3A**). Further, we impose a second constraint on the encoding procedure, requiring that all learned encoding weights be sparse and positive (**Figure 3B**). Thus, for example, face-selective voxels can only be fit using a minimal positively-weighted combination of face-selective model units. These added forms of regularization strongly enhance the interpretability of the linking function.

We introduce the sparse positivity constraint on the encoding weights to help ensure that the learned tuning directions between model units and voxels are relatively aligned and in correspondence with each other. *Positivity* is important, as for example, scene-selective and face-selective voxels often have negatively correlated response profiles; standard linking procedures (e.g. linear regression) can apply negative weights to scene-selective DNN units to adequately model face-selective voxels (see **Supplementary Figures 5-6**). *Sparsity* is important to pressure towards as close to a one-to-one alignment between the learned tuning in models and voxels/neurons as possible, and to reduce the amount of feature remixing that happens through the linking procedure. In a way, here we have already implemented a hard sparsity constraint by restricting our voxel-wise encoding models to rely only a small prespecified set of units from the layer. But there are many face-selective voxels and face-selective units, so the additional L1 regularization pushes for an even tighter alignment between voxel tuning and DNN units. Together, these constraints help operationalize the underlying theory of the contrastive account, where the tuning directions of a DNN are meaningfully oriented in the population space, and correspond to the tuning evident across high-level visual cortex.

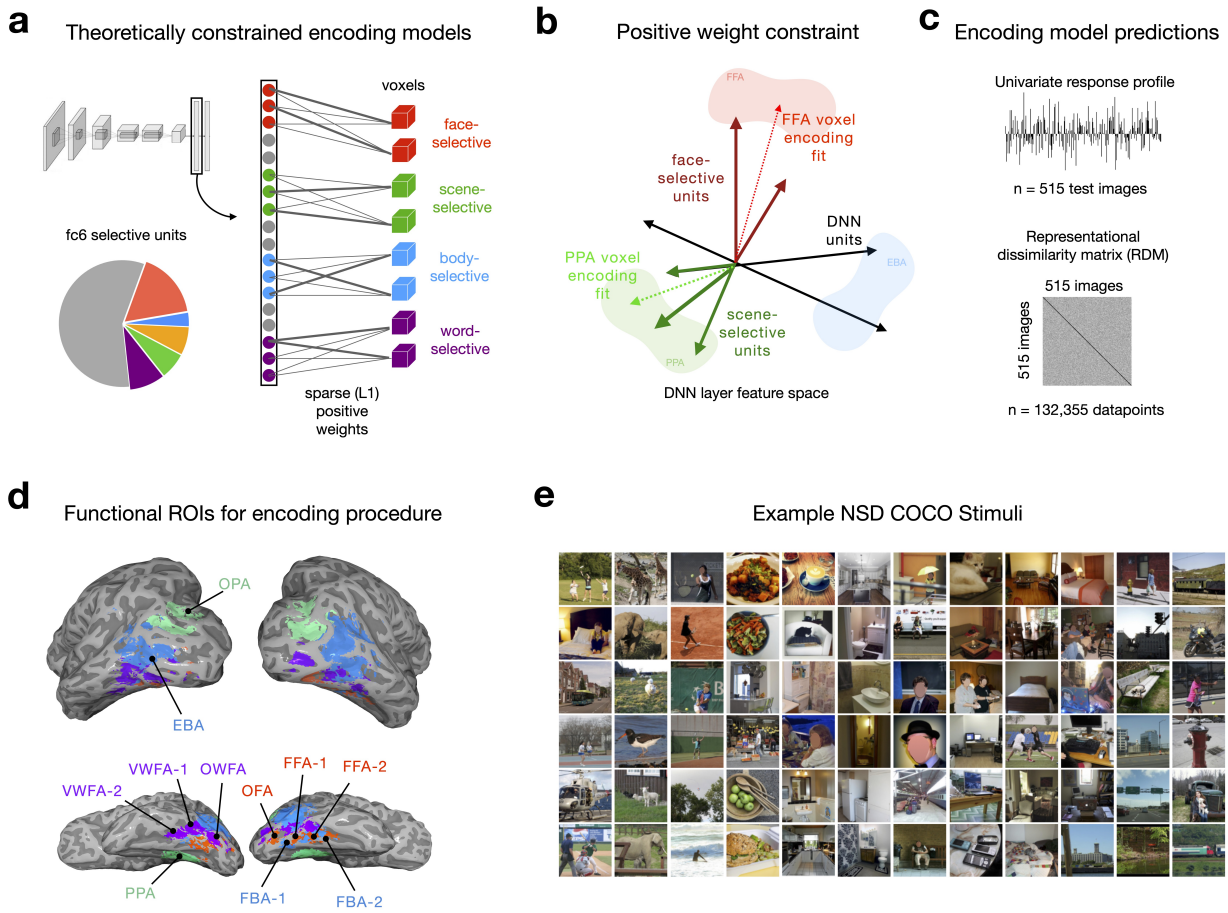


Figure 3: Sparse positive-weighted encoding models of category-selective ROIs*. (A) Schematic of the brain-model encoding procedure. DNN units and ROI surface vertices with corresponding selectivity are mapped to one-another via positive-weighted linear regression, and weights are regularized to be sparse (L1 penalty). (B) Impact of the positive weight constraint. In predicting a given brain vertex (e.g. from FFA, shaded red), DNN units with anticorrelated tuning directions to the target (e.g. from scene-selective units, dark green) cannot contribute to encoding fits. (C) Outputs of the encoding procedure, computed from model-predicted responses over the 515 test stimuli. (D) Data from the Natural Scenes Dataset (N = 8; representative subject plotted) used for the encoding procedure. 11 category selective ROIs are analyzed per subject, with preferences for faces (red), scenes (green), bodies (blue), and words (purple). (B) Examples from the test set of 515 Microsoft COCO stimuli from the NSD experiment. Images containing identifiable human faces are masked.*

Category-selective units account for representational signatures of category-selective ROIs

For our encoding analyses, we used the open-source Natural Scenes Dataset (88), targeting a set of 11 category-selective regions (**Figure 3D**). These include 3 face-selective regions (FFA-1, FFA-2, OFA), 3 body-selective regions (EBA, FBA-1, FBA-2), 2 scene-selective regions (PPA, OPA), and 3 visual word-form-selective regions (henceforth, "word-selective", VWFA-1, VWFA-2, OWFA). All analyses were done on each individual subject. For each voxel in each region, we fit a regularized (sparse positive-weighted) encoding model from the pre-identified model units in each layer with corresponding selectivity, using data from 1000 training images. Using an independent set of 1000 validation images, we then ran a subject-specific cross-validation procedure to identify the most predictive layer (see *Methods*, see **Supplementary Figures 5-6** for the complete layer-wise results).

Finally, we used these voxel-wise encoding models to generate predicted activations to a common test set of 515 images.

Our two key outcome measures were the correlation between the predicted and measured univariate response profiles (515 images) and multivariate representational dissimilarity matrices (RDMs, 132,355 pairwise comparisons). Thus, the NSD dataset offers a substantially richer characterization of item-level representational structure in each region than has previously been possible (though there are also clear limitations of this test set, see *General Discussion*).

Figure 4 shows results for all ROIs, for both univariate response profiles and multivariate RDMs. We found that the face, scene, body, and word-selective model units each accounted for substantial structure in their corresponding brain regions, for both univariate responses (e.g., 8-subject mean best-layer $r = 0.58$ for FFA-1, $r = 0.69$ for PPA, $r = 0.48$ for EBA, $r = 0.28$ for VWFA-1) as well as multivariate responses (e.g., 8-subject mean best-layer $r = 0.47$ for FFA-1, $r = 0.46$ for PPA, $r = 0.45$ for EBA, $r = 0.41$ for VWFA-1). Here we plot the results from the best-fitting layer, however we note that units from several mid-to-high layers achieved comparable levels of prediction; these results are not highly dependent on the choice of one model layer (see **Supplementary Figures 5-6** for layer-wise results; **Supplementary Figure 7** for summary of top-predicting layers). We also observe that the most predictive groups of DNN units are nearly always those with matched selectivity to the neural ROI (e.g. face-selective units predicting FFA, scene-selective units predicting PPA; **Supplementary Figures 5A, 6A**). Critically, this alignment arises only as a consequence of the sparse-positive encoding procedure – when repeating our analyses using unregularized linear regression, all domain-selective subsets are able to achieve reasonably high prediction levels (**Supplementary Figures 5B, 6B**). Thus, sparse-positive regularization is necessary to preserve a meaningful analogy between model and brain tuning directions during the encoding procedure.

Will any feature model be able to predict these brain data? We considered the category-selective units in untrained models, as well as two models of lower-level image statistics (Gist PC and Gabor features, see *Methods*). Across the 11 ROIs, these 3 feature models achieved substantially lower prediction levels than the contrastive DNN unit subsets, for both univariate (paired t-test of within-subject differences, 32 of 33 comparisons significant, Bonferroni corrected $p < 0.0014$) and multivariate comparisons (33 of 33 tests significant). Indeed, these simpler models accounted for virtually no structure of the brain region RDMs.

Finally, we tested whether a feature space designed specifically for face recognition might also show similar (or even better) emergent fits to face-selective brain regions. We trained an AlexNet model to perform supervised 3,372-way face recognition using the VGGFace2 dataset (95, see *Methods*). The later stages of this model can be thought of as operationalizing a face-specialized feature space, with feature tuning learned solely in the service of recognizing individuals, in a way that is performant over different viewpoints, lighting, expressions, etc. The trained VGGFace model achieved high accuracy at multi-way face individuation ($\sim$84% top-1 accuracy; linear SVM trained on penultimate layer activations to novel identities from the test set). We used the sparse positive mapping procedure to fit encoding models from the full activation matrix of each layer to each subject's face-selective voxels (FFA-1, FFA-2, and OFA).

The face-recognition model features were less predictive than the face-selective subsets of the contrastive model (FFA-1: mean r = 0.32 vs 0.58 for univariate prediction, r = 0.16 vs 0.47 for RDM prediction; same trend in FFA-2 and OFA; paired t-test statistically significant for all comparisons; Bonferroni corrected $p < 0.0014$; **Figure 4A-B**, red). This empirical result is consistent with other work comparing face-trained vs ImageNet-trained models (29, 77, 96, 97). Thus, not just any rich

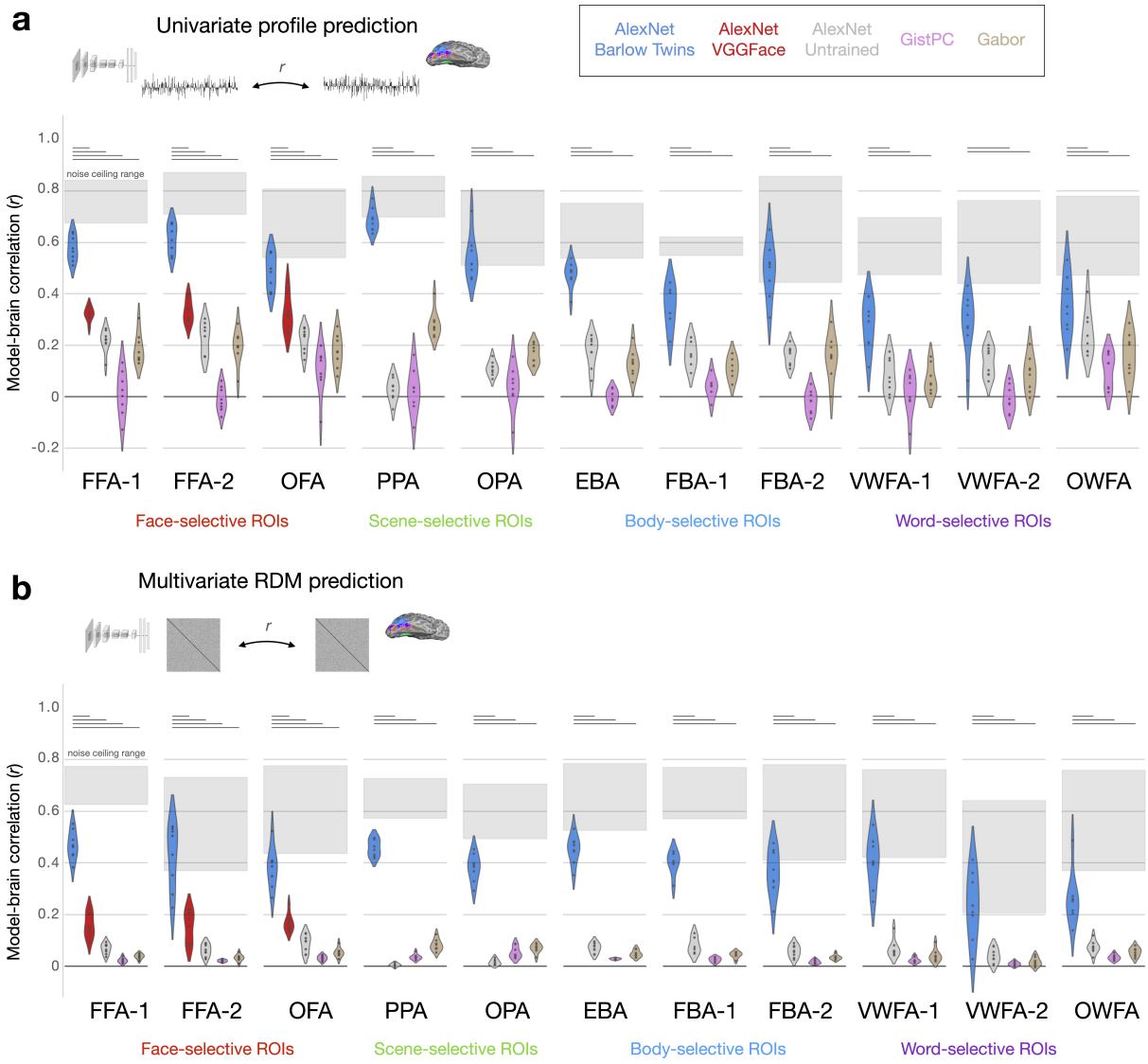


Figure 4: Representational correspondence between model and brain. *Encoding results for predicting univariate response profiles (A) and multivariate RDMs (B) are summarized for the 8 NSD subjects, for each DNN and low-level image statistic model. Plotted values reflect best-layer correlations, as defined using cross-validation (see Methods). Shaded regions show the range of subject-specific noise ceilings computed over the data from each ROI. Significance is assessed using paired t-tests over subject-wise prediction levels between the AlexNet Barlow Twins model and each other candidate model. Horizontal bars reflect significant effects favoring the Barlow Twins model; Bonferroni-corrected* $p < 0.0014$.

feature space can capture the representational structure of these regions—the face-selective tuning emerging in the contrastive feature space is both distinctive from other feature models and better for matching these brain data.

Overall, these results demonstrate that pre-identified unit subsets from a single DNN model can predict the response structure of 11 diverse category-selective regions, capturing both their univariate tuning over hundreds of natural images, and the multivariate similarity structure over hundreds of thousands of pairwise similarity comparisons.

A contrastive account of category-selective regions

This set of computational and empirical results offers a unifying account of category-selective tuning. Self-supervised contrastive learning over a rich image diet naturally yields rich discriminative, diet-dependent features. These features are not 'about' any one category – the learning is entirely self-supervised and does not presuppose categories. However, the set of features as a whole serve to differentiate all kinds of content. By implication, that model units learn robust face, body, scene, and word selectivity implies that these categories have image statistics that are particularly distinctive in the space of natural image views. Further, this tuning arises gradually across the hierarchical stages (**Supplementary Figure 1**), effectively routing images from their (tangled) pixel-based representation to increasingly distinctive and separable parts of the population code.

To provide a graphical intuition for this progressive untangling of implicit category information, **Figure 5** traces the representational similarity of a small probe set of images from four categories, across several hierarchical processing stages of the contrastive model (conv1, conv3, conv5, fc7; **Figure 5A**; see *Methods*). Each dot reflects an image, where images that are far apart in each multidimensional scaling plot reflects the degree that they evoke more distinct activation patterns across the units in a layer. Early stages show a more 'tangled' population code where images from different categories are intermixed, while later stages show clearer emergent separability of these categories.

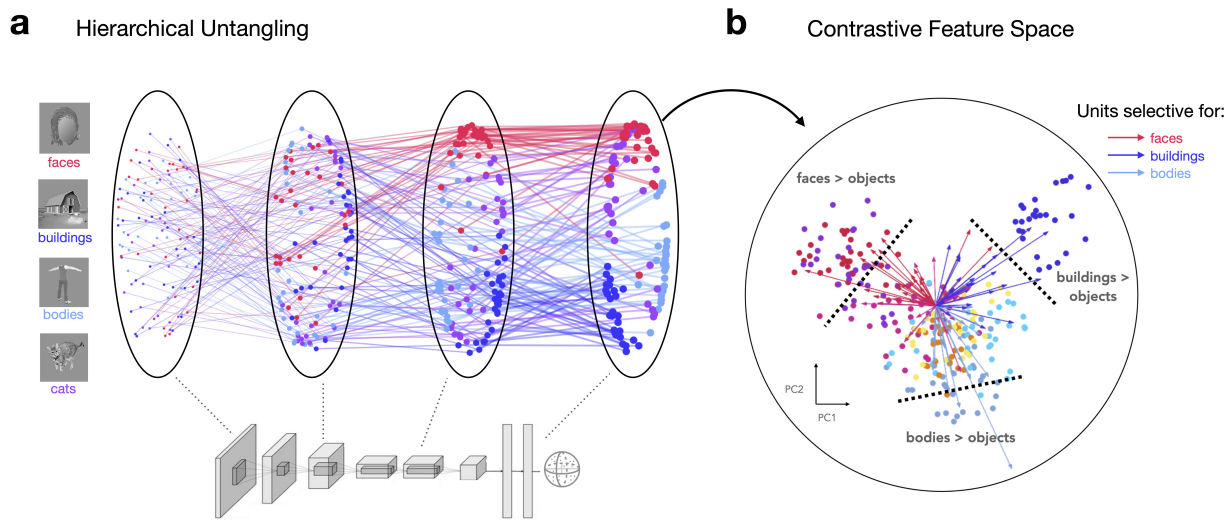


Figure 5: Low-dimensional projection of category-selective axes in a contrastive feature space. *(A) Image trajectories through the Barlow Twins hierarchical population code are assessed using a 4-category image set with 30 instances per category. Layers conv1, conv3, conv5, and fc7 are plotted. Each connected sequence of dots reflects the trajectory of an image's feature representations in MDS space (see Methods). (B) Delineation of category-selective regions in the PC space of layer fc6. Feature representations from the same set of probe stimuli are plotted in PC2 space; each image is a dot. Tuning orientations of the top 25 most selective units for faces, scenes, and bodies are projected into the 2-dimensional PC space of the ImageNet validation set. Images containing identifiable human faces are masked.*

Figure 5B aims to help convey how different category-selective tuning emerges within the same layer. The population code from the fc6 layer is plotted for a probe set of 8 categories (98), along the first two principal components derived from model responses to the ImageNet validation set. In this low-dimensional projection, images of faces, buildings, and bodies naturally have distinctive unit activations from each other as evidenced by the separable clustering. We took the pre-identified face, building, and body-selective model units (from the initial localizer procedure) and projected them into

the PC space, indicated by arrows. This visualization helps clarify that units with particular tuning are key for enabling images of faces, bodies, or scenes to separate from all other images in the model's representational space. Localizing these units with a statistical contrast (e.g. faces > objects) is akin to selecting units through a decision boundary in this PC space. These are illustrated with schematic dashed lines in the visualization, highlighting how category-selective units are different facets of this single, unified population code.

Note that this model was trained without presupposing the notion of face, body, and scene categories at all – thus, the progressive untangling and selective tuning of these categories is an emergent consequence of their particular image statistics, in the context of learning discriminative features over a rich natural image diet. In other words, the tuning of face-selective units, for instance, does not reflect tuning to faces independent of other categories (which presupposes a face category), but instead reflects tuning to faces in contrast to other categories. Consequently, the nature of category-selective tuning for each category must be understood in the context of the statistics of the visual input diet.

DISCUSSION

Here we present an updated computational-empirical account of category-selective tuning along the human ventral visual stream. We leverage self-supervised, contrastive, deep neural networks, trained over a rich image diet. We show that these models naturally learn features with emergent category-selective tuning, which yield functionally dissociable deficits when lesioned, and that these features can predict the response structure of diverse category-selective regions to hundreds of natural images. As such, we propose that category-selective tuning in the human brain is intimately tied to visual diet, and best understood in the context of the full scope of experienced input. This framework supports a unified theory of category selectivity, suggesting that regions like FFA and PPA arise as a consequence of positive information routing along the ventral hierarchy, which serves to untangle the statistics of the visual world in support of downstream readout. Next, we clarify the implications of this "contrastive coding" framework, explore the consequences of diet-dependent feature tuning, and discuss our novel approach to model-brain linking procedures.

A contrastive coding alternative to modular and distributed coding frameworks

The category-selective regions of visual cortex and their relationship to each other and the surrounding occipitotemporal cortex have been the subject of intense debate (e.g. 2, 3). The polar extremes of fully modular versus fully distributed coding each account for some but not all empirical patterns of data, leading to an unresolved tension. Here we offer a "contrastive coding" account, which we suggest can reconcile many points of conflict between modular and distributed viewpoints, drawing on insights from the current work into *tuning*, *readout*, and *routing*, as detailed below.

The first insight we offer regards the possible nature of the *tuning* in high-level visual cortex. Our account of feature tuning in these regions is neither domain-specialized (as is often assumed in more modular frameworks) nor generic or uninterpretable (as if often assumed in more distributed frameworks), but instead is diet-dependent, with learned selectivities that reflect the scope of visual input. We refer to this as "contrastive feature tuning", drawing on the principles of self-supervised contrastive learning objectives, where the learned feature tuning is directly determined by the set of visual inputs that need to be distinguished. The units in each layer work together as a population to represent and untangle the natural image diet (distributed-like), but individual unit tuning can be causally linked to recognition of specific content (modular-like). This is particularly true in later layers,

where category-selective units 'point to' a particular part of the natural image manifold. For example, we show that face-selective units encode a particular part of high-level, contrastive feature space where face images happen to cluster together, which is distinct from scene-selective, body-selective, and word-selective sub-spaces.

A second insight of the current work relates to how information is *read out* of a contrastive feature space. Within distributed coding frameworks, the information is assumed to be spread across the population, and is thus extracted through fully-connected linear readout. In a contrastive space, however, even though the features are part of a shared population code, we show that recognizing individual categories does not require fully-connected readout. That is, our implementation of a sparse readout module for category recognition demonstrates that connections to only a small subset of contrastive features can support recognition of any particular category.

The success of sparse readout from a contrastive feature space has important theoretical implications. First, it clarifies a middle position between modular and distributed frameworks, challenging the extreme inference that because there cannot be a neuron for every category in the limit, the code must be distributed and read out in a fully distributed manner. Second, from a biological plausibility perspective, sparse readout requires dramatically fewer connections at the interface between a high-level visual feature bank and a cognitive categorical representation. Third, theoretically, sparse-regularized readout offers an alternative, more constrained, measure of information evident in a population code than the current gold-standard of linear readout (5, 99; see 100, 101 for review).

The final insight of this work relevant to our contrastive coding framework is related to the concept of *positive routing* through hierarchical population codes. Functionally specialized deficits are a cornerstone of modular theories, with less clear connection to distributed theories. Our lesioning experiments relieve this tension, demonstrating that functionally specialized deficits can occur even in general models, trained with general architectures, and general objective functions. Our data indirectly show that units with similar category-selective tuning form a connected route through the hierarchy, serving to help direct different kinds of image content to separable locations in the final embedding space. This hypothesis is consistent with emerging mechanistic interpretability work showing that there are separable, sparse routes of computation through visual DNNs, and that learning leads to implicit modularization of the networks (102, 103). Indeed, Dobs et al. (104) have shown that joint training on face and object recognition naturally produces implicit modular routes through the network, without requiring pre-specified architectural branches. We also find evidence for implicitly learned routes in the network, but here, critically, this learned modularity is not induced by domain-specialized task objectives, but by the structure of the visual diet itself.

Implications of diet-depending feature tuning

Here we argue that the tuning of category-selective neurons in the human brain is fundamentally tied to the statistics of visual experience. Our analyses leverage ImageNet, whose scale and diversity are sufficient to learn brain-predictive features via contrastive objectives that support object recognition through sparse readout. However, our modeling framework is limited in several ways. First, ImageNet may not be a faithful reflection of the visual diet that humans experience. Recent efforts have introduced alternative large-scale datasets which may improve the ecological validity of models of the ventral stream (e.g. EcoSet, 105). Second, human OTC contains large swaths of cortex devoted to body representation (EBA) – but our model learns very few units with emergent body selectivity. Within our

modeling framework, we hypothesize that the visual diet may lack sufficient variation in body pose and social interactions to capture the full diversity of body-relevant features in human OTC (e.g. 106–111).

Another intriguing observation is that units with word-selective tuning emerge in models that do not have reading as a functional goal (see also 112), while in the human brain, the visual word-form area emerges during acquisition of reading skills (113). Our lesion analyses suggest that words presented on noisy backgrounds follow similar 'routes' as insect images. One account of this link between text and insects is simply that the ImageNet categories may lack sufficient close-up views of letters and alphabets, and that insects are simply the next nearest category in terms of image statistics. However, an alternative interpretation is that visual experience with thin-lined objects, such as insects, could act as a scaffold for the later formation of letter and word representations. Indeed, a literature search reveals that insects tend to evoke greater responses in the occipital word-form area than other object classes (e.g. chairs, cars; 114), hinting that the link between these kinds of content is not solely an artifact of the ImageNet diet.

Can this model help explain why only specific categories (such as face, body, scenes, and visual word forms), rather than others, show units with strong selectivity? According to the contrastive coding account, feature tuning hinges on the relative frequency and variation of experiences with these categories, and their intrinsic visual image statistics. Under this view, it's possible that certain artifact categories are more functionally defined (such as 'coffee maker'), with more heterogeneous feature representations, leading to a lesser degree of selective feature tuning than other categories. Such categories may even require a more distributed readout from the visual population code than more natural categories like 'granny smith apple.' There is surely even more complexity to acknowledge here, as empirical data show that *early* visual experience with categories is particularly critical for selective tuning to develop (e.g. 115–118). Similarly, deep neural networks demonstrate critical periods for categories experienced early on (119). A productive direction for future work is to more deeply link these models' input diets and training parameters with the rich literature on variation in visual experiences throughout development (120–126).

A key related question is whether the shared 515 test images in NSD are sufficiently representative of natural images to support the claim that category-selective regions have contrastive feature tuning (rather than domain-specialized features). On one hand, the stimuli provide a diverse sweep of natural image content, which has the advantage of not favoring a specific representational hypothesis. On the other hand, to validate the contrastive feature space, it would also be useful to examine complementary datasets containing many individual face images, body images, or visually presented word images. There is some existing evidence from such studies that supports the contrastive coding hypothesis: Murty et al. collected responses to 185 images of faces (n = 25) and non-faces (n = 160), showing that deep neural networks trained on the ImageNet dataset could explain substantial variance in FFA voxel tuning and geometry (better than many other models, including VGGFace-trained models; 29). Similarly, Vinken et al. studied responses of face-selective neurons in non-human primates (n = 447 face stimuli and 932 non-face object stimuli), finding that response profiles were better fit by ImageNet-trained models than face-trained models (97). Chang et al. made the same observation studying primate face patch responses to 2100 face images (127). And, when analyzing face-selective neuronal groups recorded intracranially in 33 human patients, Grossman et al. observed that ImageNet features provided a better match to brain RDMs than VGGFace features, for stimulus sets containing faces only (128).

Model-to-brain linking with sparse positive constraints

Encoding models are the current standard approach for modeling voxel responses, typically as a positively and negatively weighted sum of features across a DNN layer (96, 129–132), or in some cases, across multiple layers (133). This approach may be justified, given that voxel activity reflects the metabolic needs of tens of thousands of neurons (though the same degree of flexible reweighting is typically allowed for single neuron encoding models as well; 134, 135).

However, in line with our theory of DNNs as learning implicit hierarchical routes from pixels to untangled embedding spaces, we argue that a proper linking assumption should emphasize the degree of *emergent* alignment, with minimal feature reweighting and remixing allowed. Additionally, because only positive activations progress through the layer hierarchy in relu-networks, there is clear motivation for the positive-weighting constraint. A voxel-to-unit linking procedure that can flip the positive or negative tuning direction of a unit might add information to the linking process that will not be accessible to the DNN itself in subsequent layers. Indeed, the critical role of tuning direction in coding content distinctions aligns with neurophysiological research. Jaegle et al. have demonstrated that the orientation of tuning in the population code of IT cortex is a key signal of object identity, and that variations in magnitude along the same orientation are linked to contrast and attention, while holding identity constant (136).

Relatedly, our constrained sparse-positive mapping scheme breaks from common assumptions that the geometry of the space is the key determinant of downstream information readout. In such a scheme, uniformly rotating all the units' tuning directions would not affect the information that can be linearly decoded from the population. However, such rotations would likely alter the nature of category-selective tuning in the space, impacting which information gets routed from one layer to the next (since only positive activations travel forward). Thus, the sparse-positive linking procedure operationalizes the assumption that tuning directions are not arbitrary – rather, that they are critical to the function of the neural system.

To address a potential confusion, our experiments involve two different uses of sparsity: sparse readout (for ImageNet categorization), and sparse-positive linking constraints. These have different motivations. Sparse readout is focused on the internal structure of the model, dictating how one layer interacts with another. This is guided by biological considerations, such as the need for constraints on the number of 'wires' or connections required for successful readout. On the other hand, sparse-positive encoding constraints deal with the relationship between two systems – one artificial and one biological. Here, the objective is to naturally align tuning curves with minimal remixing and reweighting of the input data.

Finally, our work parallels ongoing efforts to model the topographic layout of the ventral visual stream using object recognition DNNs. Here we have focused on category-selective tuning in deep neural networks and their capacity to predict individual voxels in category-selective regions. Even though our model has no explicit 'regions', it nevertheless possesses a topological structure, with units showing similar tuning based on their connectivity to prior layers. Different approaches have sought to spatialize this network structure, by smoothly mapping the learned data manifold (74, 137) or by incorporating a topographic constraint into the model's loss function and architecture during the feature learning process itself (75, 138). These efforts complement our theoretical framework, as a smooth mapping of the contrastive features onto a cortical sheet with similar tuning nearby would directly facilitate sparse readout of category information via relatively localized connections.

In sum, the contrastive coding framework provides a new way to reconcile the theoretical debate regarding the nature of category-selective regions. Our work re-situates hallmark empirical studies of

regions like FFA and PPA in the context of an updated neuroconnectionist paradigm that focuses on how visual diet, feature tuning, and information readout interact along the visual hierarchy (139). Moving forward, this framework may support the discovery of new routes, selectivities, and content distinctions within occipitotemporal cortex, in a data-informed manner. Indeed, data-driven modeling of fMRI data has recently led to the discovery of food-selective responses within the ventral stream, hinting at the potential of similar approaches to reveal finer-grained aspects of visual representation and topography (140–142). Existing or hypothesized 'content routes' could be further probed using tailored stimulus sets (e.g. controversial stimuli, 143), and subject to new circuit decomposition techniques (e.g. 102), bringing advances in mechanistic interpretability research to bear on visual cognitive neuroscience. Broadly, we offer that a contrastive coding framework provides new traction for understanding how we recognize exemplars both from the well-studied visual domains, as well as thousands of other visual entities, in everyday life.

METHODS

Models

Our main analyses involved a standard AlexNet architecture, modified to have group-norm rather than batch-norm operations, which we have found helpful to stabilize self-supervised model training (following 77). The model was trained on the ImageNet dataset using the self-supervised Barlow Twins objective, with the same training procedures and hyperparameters employed in 76. The goal of the Barlow Twins objective is to reduce redundancy in the neural network's learned representations while preserving the informativeness of its features. In brief, the learning algorithm involves measuring the cross-correlation matrix between the embeddings of two identical networks fed with distorted versions of a batch of image samples, with the objective to make this matrix close to the identity.

For supplementary analyses with a similar contrastive objective function, the Instance Prototype Contrastive Learning (IPCL) model was applied (77). This self-supervised model also uses an AlexNet architecture but integrates group normalization layers. A forward pass involves creating five augmentations per image, passing them through the DNN backbone, and projecting them into a 128-dimensional L2-normed embedding space. The learning objective includes two components: driving the embeddings of augmented views towards their average representation, and simultaneously encouraging these representations to differ from those of recently encountered items.

For comparison between self-supervised and category-supervised feature spaces, an AlexNet model trained on the VGGFace2 dataset was employed (95). This large-scale dataset contains over 3.3 million images of more than 9,000 individuals, capturing diverse identity variations. The VGGFace trained model was trained using cross-entropy loss and SGD for 100 epochs, adopting the default image transformation scheme.

DNN localizer procedure

The DNN localizer procedure involved defining category-selective units within each layer of the network by analyzing responses to different categories of images in a localizer image set. Primary analyses relied on the vpnl-fLoc localizer image set (87), which contains 288 stimuli for each of the face, scene, body, object, and character (word) categories, plus 144 additional scrambled stimuli. Different spatial locations in convolutional layers were treated as different units; there was no activation pooling involved in identifying selective units. Face-selective units were identified by comparing

activations in response to face images against activations in response to images from each non-face category (bodies, scenes, objects, word, scrambled). Each comparison involved a two-sample t-test for each neuron, yielding a t-map and corresponding p-map reflecting relative preference for faces versus each non-face category. A False Discovery Rate (FDR) threshold was calculated and applied to define 'selective' face-preferring neurons for each domain comparison. The intersection of these neuron groups was then identified to define the final set of face-selective units within a layer. This process was repeated for scene-, body-, and word-selective units, ensuring no overlap among groups due to preference requirements.

The robustness of the localizer procedure was evaluated qualitatively by examining responses in layer fc6 to an independent RGB probe set that maintained the same categorical domains (n = 80 stimuli per category), to test for high activations to the preferred stimuli in the corresponding subsets of selective units. To quantify these effects, the generalizability of selectivity estimates was assessed by comparing the t-values from the initial localizer procedure with those derived from repeating the same procedure using the RGB probe images. The degree of generalization was reflected in the extent to which the selective units identified using the initial procedure also showed large-magnitude positive t-values in response to the RGB probe set. The same generalization analysis was also conducted on an untrained AlexNet Barlow Twins model.

Lesioning DNN selective units

This analysis aimed to understand the effect of perturbing groups of DNN units that are selective for each domain (e.g. faces or bodies) on recognition performance. Before introducing any lesions, the model's baseline recognition accuracy was measured by appending a linear readout layer to the relu7 model layer and training it for 10 epochs on ImageNet categorization. The readout weights were initialized following a normal distribution, with all other trainable parameters frozen. Training was performed using the entire ImageNet training set and employed a batch size of 512. To promote sparsity, an L1 penalty was applied (lambda = 10e-6). A OneCycle learning rate schedule was used, with the following hyperparameters: max_lr = 0.05, initial_lr = 0.001, pct_start = 0.3. After the training phase, recognition accuracy was assessed using the ImageNet validation set, and top-5 accuracy was used for all subsequent analyses.

To quantify the effect of lesions targeting a specific domain, a masking procedure was employed to 'lesion' the target selective units in relu layers 1-7, setting their output activations to zero. The indices of the lesioned units for each domain were those identified as category-selective in the DNN localizer procedure. Post-lesioning, the model's recognition performance was again evaluated, without any further training to assess recovery of category performance. After lesioning, the drop in recognition accuracy levels was evaluated through a second ImageNet validation pass.

To identify the k categories most affected by lesions to face, body, scene, and word-selective units, the accuracy drop was first calculated for each category using half of each ImageNet validation set. The top categories experiencing the most significant drops in accuracy were identified, and then the remaining portion of the validation set was used to compute final changes in accuracy due to lesioning. We also compared the full 1000-dimensional cost profiles between each pair of domain-targeted lesions, using Pearson similarity.

Lastly, we tested if activation levels within each group of selective units at baseline could predict subsequent recognition deficits following perturbation. To do so, we computed the correlation between each layer's 1000-dimensional category activation profile and the lesioning cost profile for each of the

face-, scene-, body-, and word-selective lesions.

Relating category selective regions in models and brains

We investigated if category selective units in the DNN could explain the high-dimensional response structure in diverse category selective brain regions. We used the Natural Scenes Dataset (NSD; N = 8; 88), containing measurements of over 70,000 unique stimuli from the Microsoft Common Objects in Context (COCO) dataset (144). We implemented a sparse positive-weighted encoding procedure to map between selective units and brain regions of interest (ROIs) with corresponding selectivities, taking into account both univariate and multivariate response signatures.

For our analysis, we focused on a subset of the dataset (16,515 stimuli; 8 subjects x 2,000 total subject-specific stimuli split equally between the training and validation sets, and a shared 515-stimulus test set). These stimuli, presented during a long-term recognition task, were viewed three times by each subject. We analyzed data from 11 category selective ROIs from occipitotemporal cortex (OTC), defined using an independent localizer procedure. These included face-selective regions (OFA, FFA-1, FFA-2), scene-selective regions (OPA, PPA), body-selective regions (EBA, FBA-1, FBA-2), and word-selective regions (OWFA, VWFA-1, VWFA-2). The same localizer stimuli were used to identify category-selective units in our DNN models.

BOLD responses in NSD were estimated using the GLMsingle toolbox (145). We accounted for potential session-to-session response inconsistencies by z-scoring the within-session response profiles in each surface vertex, before extracting responses to the training, validation, and testing stimuli. The data from repeated instances of each stimulus were averaged, and we implemented a reliability-based voxel selection procedure (146) to select vertices with stable response structure. We used noise ceiling signal-to-noise ratio (NCSNR, 88) as our reliability metric and included only voxels with NCSNR $>$ 0.3. Surface vertices from the same ROI across the two hemispheres (e.g. lh-PPA and rh-PPA) were concatenated prior to analysis.

Our constrained encoding procedure involved modeling each brain vertex as a linear combination of features from a DNN layer, using only DNN units with corresponding selectivity (e.g. only face-selective units are used for modeling FFA and OFA). All encoding weights were required to be sparse and positive. For the VGGFace model, these constraints only differed in that the entire layer's features were used for encoding. We conducted an additional set of analyses in the trained Barlow Twins model, comparing matched-selectivity encoding models to those with mismatched selectivities (e.g. mapping scene-, body-, and word-selective units to FFA).

The COCO stimuli were divided into training, validation, and test sets. Features were extracted from each DNN model for each COCO stimulus at each distinct layer of the network from conv3 onward. We used sklearn's Lasso function to fit each L1-regularized encoding model, enforcing non-negativity of the encoding weights using the 'positive' input argument. The default alpha value of 0.1 was used for all encoding fits in the ImageNet and VGGFace-trained models, and a value of 0.001 was used for encoding models from the untrained model.

Once an encoding model was fit, we then predicted the response of every ROI vertex to each image for the validation and test images, and computed two model-predicted outputs for comparison to the true brain responses: the ROI mean univariate response profile and the multivariate similarity structure ("veRSA", 77, 96). In determining maximal model-brain correspondence, our key metrics were derived from the best-fitting layer to a given brain region. We calculated the univariate and veRSA correlations

(Pearson r) between model predictions and actual brain data for each layer using the validation set. The layer with the highest correlation was separately selected for univariate and multivariate metrics. Using the independent test set of 515 stimuli, final univariate and veRSA correlations were then computed for the selected layers, providing an independent measure of the maximum correspondence between the model and brain region.

To test if basic image statistic models were able to capture representational structure in the category selective ROIs, we additionally computed encoding fits for Gabor and GistPC feature spaces. Gabor features were extracted in a 16 x 16 grid over the original, recognizable images (425 x 425 pixels) at 4 different scales, with 12, 8, 6, and 4 oriented Gist per scale, respectively (147). This procedure yielded a flattened feature dimension of 7680. GistPC model features were extracted by taking the first 50 principal components of this Gabor feature matrix. The encoding procedures were identical to those described above for the DNN models.

We estimated noise ceilings for each target brain ROI to provide a context for model performance results. We used a novel method based on generative modeling of signal and noise, termed GSN (148), which estimates multivariate Gaussian distributions characterizing the signal and noise structure of each ROI from each subject. Distinct noise ceilings for the univariate and RSA analyses were calculated through Monte Carlo simulations. This process involved first generating a noiseless 515-dimensional univariate response profile (or a 515 x 515 RDM) from the inferred signal distribution. These profiles (or RDMs) were then correlated with counterparts constructed from noisy measurements. The latter were generated by adding samples from the estimated signal and noise distributions, thereby effectively simulating the realistic observational conditions. The range of estimated noise ceilings for the 8 subjects are plotted separately for each outcome metric and each brain region.

Low-dimensional visualization of hierarchical contrastive feature spaces

To map the hierarchical evolution of category-selective tuning, we visualized the gradual emergence of categorical structure in the Barlow Twins model using a compact image set of four object categories: faces, buildings, bodies, and cats (98). Each category contained 30 examples, with equalized overall luminance and contrast levels across the 120 images, achieved via the SHINE toolbox.

For our hierarchical visualization, we focused on model layers conv1, conv3, conv5, and fc7. We initially condensed the activation data for each layer (120 images x n_features) via Principal Component Analysis (PCA) into a 10-component matrix (120 images x 10 PCs). These were subsequently merged into one matrix (480 images x 10 PCs), enabling the analysis of each image's 'representational trajectory' through the model (149). Using Pearson dissimilarity, we then constructed a 'meta-RDM' to encapsulate the pairwise similarity of all image features across all layers (480 images x 480 images). The meta-RDM served as input to multidimensional scaling (MDS), yielding a 2D projection of the entire representational geometry (480 images x 2 MDS dimensions).

This 2D representation was then lifted into a 3D visualization. Here, embeddings from disparate layers were grouped into different x-axis positions, symbolizing the progression from conv1 to fc7. The result was a 3D scatter plot where each dot signifies an input image, connected by lines to indicate its trajectory through the model layers. The proximity between dots within each layer's MDS space reflects the similarity of their corresponding embeddings in DNN feature space.

To demonstrate the demarcation of category-selective regions within this feature space, we created another low-dimensional visualization using the 120-image probe set. We first performed 2D PCA on

4096-dimensional layer fc6 activations computed for a subset of the ImageNet validation set (2000 total images; 2 per category), and then projected the fc6 response vectors for each of the 120 probe stimuli into this 2D space, with dots color-coded by category.

Lastly, we aimed to visualize the tuning directions of category-selective units within this low-dimensional projection. Using the output of the DNN localizer procedure (vpnl-fLoc image set, 87), we identified the indices of face-, body-, and scene-selective units in layer fc6 of the Barlow Twins model. We then depicted the oriented tuning vector of each of the 25 most selective units for these three domains. This was accomplished by creating a one-hot 4096-dimensional vector denoting the index of the unit, and computing its dot product with the (4096, 2)-dimensional PCA component matrix from the ImageNet validation stimuli. A constant scaling factor of 2.5e4 was applied to each one-hot vector before multiplication for visualization purposes. This ensured the resultant tuning vector's magnitude laid meaningfully onto the image data points in PC space, without altering the relative angles of the different tuning vectors.

Data Availability

Public image datasets used to train the models include: ImageNet (https://image-net.org/) and VGGFace2 (https://github.com/ox-vgg/vgg_face2). The Natural Scenes Dataset is available here: https://naturalscenesdataset.org/.

Code Availability

Code for all analysis and plotting routines is available on GitHub: https://github.com/jacob-prince/PROJECT_DNFFA and https://github.com/jacob-prince/jsputils.

Acknowledgments

This project was supported by NSF Career Grant #1942438 (TK) and an NDSEG fellowship (JSP). We thank all members of Harvard Vision Lab, K. Kay, K. Vinken, G. Tuckute, and N. Blauch for helpful conversations throughout the project and for feedback on the manuscript.

Author Contributions

Conceptualization: JSP and TK. Methodology: JSP, TK, and GAA. Investigation: JSP, TK, and GAA. Visualization: JSP and TK. Supervision: TK and GAA. Writing—original draft: JSP and TK. Writing—review and editing: JSP, TK, and GAA.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Mortimer Mishkin, Leslie G Ungerleider, and Kathleen A Macko. Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, 6:414–417, 1983.

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

[3] Nancy Kanwisher. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, 107(25):11163–11170, 2010.

[4] Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.

[5] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[6] Travis Meyer and Nicole C Rust. Single-exposure visual memory judgments are reflected in inferotemporal cortex. *elife*, 7:e32259, 2018.

[7] Nancy Kanwisher, Josh McDermott, and Marvin M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, 17(11): 4302–4311, 1997.

[8] Gregory McCarthy, Aina Puce, John C Gore, and Truett Allison. Face-specific processing in the human fusiform gyrus. *Journal of cognitive neuroscience*, 9(5):605–610, 1997.

[9] Aina Puce, Truett Allison, Maryam Asgari, John C Gore, and Gregory McCarthy. Differential sensitivity of human visual cortex to faces, letterstrings, and textures: a functional magnetic resonance imaging study. *Journal of neuroscience*, 16(16):5205–5215, 1996.

[10] Doris Y Tsao, Winrich A Freiwald, Roger BH Tootell, and Margaret S Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.

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

[12] Marius V Peelen and Paul E Downing. Selectivity for the human body in the fusiform gyrus. *Journal of neurophysiology*, 93(1):603–608, 2005.

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

[14] Geoffrey K Aguirre, Eric Zarahn, and Mark D'Esposito. An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron*, 21(2):373–383, 1998.

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

[16] Steven E Petersen, Peter T Fox, Abraham Z Snyder, and Marcus E Raichle. Activation of extrastriate and frontal cortical areas by visual words and word-like stimuli. *Science*, 249(4972): 1041–1044, 1990.

[17] Laurent Cohen, Stanislas Dehaene, Lionel Naccache, Stéphane Lehéricy, Ghislaine Dehaene-Lambertz, Marie-Anne Hénaff, and François Michel. The visual word form area: spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, 123(2):291–307, 2000.

[18] Thad A Polk, Matthew Stallcup, Geoffrey K Aguirre, David C Alsop, Mark D'esposito, John A Detre, and Martha J Farah. Neural specialization for letter recognition. *Journal of Cognitive Neuroscience*, 14(2):145–159, 2002.

[19] Hans P Op de Beeck, Ineke Pillet, and J Brendan Ritchie. Factors determining where category-selective areas emerge in visual cortex. *Trends in cognitive sciences*, 23(9):784–797, 2019.

[20] Lindsey J Powell, Heather L Kosakowski, and Rebecca Saxe. Social origins of cortical face areas. *Trends in cognitive sciences*, 22(9):752–763, 2018.

[21] Margaret S Livingstone, Michael J Arcaro, and Peter F Schade. Cortex is cortex: Ubiquitous principles drive face-domain development. *Trends in cognitive sciences*, 23(1):3, 2019.

[22] Michael J Arcaro and Margaret S Livingstone. On the relationship between maps and domains in inferotemporal cortex. *Nature Reviews Neuroscience*, 22(9):573–583, 2021.

[23] Rosemary A Cowell and Garrison W Cottrell. What evidence supports special processing for faces? a cautionary tale for fmri interpretation. *Journal of Cognitive Neuroscience*, 25(11): 1777–1793, 2013.

[24] Michael J Tarr and Isabel Gauthier. Ffa: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature neuroscience*, 3(8):764–769, 2000.

[25] David C Plaut and Marlene Behrmann. Complementary neural representations for faces and words: A computational exploration. *Cognitive neuropsychology*, 28(3-4):251–275, 2011.

[26] Sheng He, Haicheng Liu, Yi Jiang, Changming Chen, Qiyong Gong, and Xuchu Weng. Transforming a left lateral fusiform region into vwfa through training in illiterate adults. *Journal of Vision*, 9(8):853–853, 2009.

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

[28] Talia Konkle and Alfonso Caramazza. The large-scale organization of object-responsive cortex is reflected in resting-state network architecture. *Cerebral cortex*, 27(10):4933–4945, 2017.

[29] N Apurva Ratan Murty, Pouya Bashivan, Alex Abate, James J DiCarlo, and Nancy Kanwisher. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature communications*, 12(1):5540, 2021.

[30] Meenakshi Khosla and Leila Wehbe. High-level visual areas act like domain-general filters with strong selectivity and functional specialization. *bioRxiv*, pages 2022–03, 2022.

[31] Heather L Kosakowski, Michael A Cohen, Atsushi Takahashi, Boris Keil, Nancy Kanwisher, and Rebecca Saxe. Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. *Current Biology*, 32(2):265–274, 2022.

[32] Bradford Z Mahon and Alfonso Caramazza. What drives the organization of object knowledge in the brain? *Trends in cognitive sciences*, 15(3):97–103, 2011.

[33] Marius V Peelen and Paul E Downing. Category selectivity in human visual cortex: Beyond visual object recognition. *Neuropsychologia*, 105:177–183, 2017.

[34] Stefania Bracci, J Brendan Ritchie, and Hans Op de Beeck. On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, 105:153–164, 2017.

[35] Sebastian Moeller, Winrich A Freiwald, and Doris Y Tsao. Patches with links: a unified system for processing faces in the macaque temporal lobe. *Science*, 320(5881):1355–1359, 2008.

[36] Olivier Pascalis and David J Kelly. The origins of face processing in humans: Phylogeny and ontogeny. *Perspectives on psychological science*, 4(2):200–209, 2009.

[37] Frederik S Kamps, Cassandra L Hendrix, Patricia A Brennan, and Daniel D Dilks. Connectivity at the origins of domain specificity in the cortical face and place networks. *Proceedings of the National Academy of Sciences*, 117(11):6163–6169, 2020.

[38] Bradford Z Mahon. Domain-specific connectivity drives the organization of object knowledge in the brain. In *Handbook of Clinical Neurology*, volume 187, pages 221–244. Elsevier, 2022.

[39] Thomas Hannagan, Amir Amedi, Laurent Cohen, Ghislaine Dehaene-Lambertz, and Stanislas Dehaene. Origins of the specialization for letters and numbers in ventral occipitotemporal cortex. *Trends in cognitive sciences*, 19(7):374–382, 2015.

[40] Stanislas Dehaene and Ghislaine Dehaene-Lambertz. Is the brain prewired for letters? *Nature neuroscience*, 19(9):1192–1193, 2016.

[41] Jin Li, David E Osher, Heather A Hansen, and Zeynep M Saygin. Innate connectivity patterns drive the development of the visual word form area. *Scientific reports*, 10(1):18039, 2020.

[42] Michel Habib and Angela Sirigu. Pure topographical disorientation: a definition and anatomical basis. *Cortex*, 23(1):73–85, 1987.

[43] Russell Epstein, Edgar A DeYoe, Daniel Z Press, Allyson C Rosen, and Nancy Kanwisher. Neuropsychological evidence for a topographical learning mechanism in parahippocampal cortex. *Cognitive neuropsychology*, 18(6):481–508, 2001.

[44] Jason JS Barton, Daniel Z Press, Julian P Keenan, and Margaret O'Connor. Lesions of the fusiform face area impair perception of facial configuration in prosopagnosia. *Neurology*, 58(1): 71–78, 2002.

[45] Christine Schiltz, Bettina Sorger, Roberto Caldara, Fatima Ahmed, Eugene Mayer, Rainer Goebel, and Bruno Rossion. Impaired face discrimination in acquired prosopagnosia is associated with abnormal response to individual faces in the right middle fusiform gyrus. *Cerebral Cortex*, 16(4):574–586, 2006.

[46] N Lang, Juergen Baudewig, Kai Kallenberg, Andrea Antal, Svenja Happe, Peter Dechent, and W Paulus. Transient prosopagnosia after ischemic stroke. *Neurology*, 66(6):916–916, 2006.

[47] Jason JS Barton. Structure and function in acquired prosopagnosia: lessons from a series of 10 patients with brain damage. *Journal of neuropsychology*, 2(1):197–225, 2008.

[48] Laurence Dricot, Bettina Sorger, Christine Schiltz, Rainer Goebel, and Bruno Rossion. The roles of "face" and "non-face" areas during individual face perception: evidence by fmri adaptation in a brain-damaged prosopagnosic patient. *Neuroimage*, 40(1):318–332, 2008.

[49] Valentina Moro, Cosimo Urgesi, Simone Pernigo, Paola Lanteri, Mariella Pazzaglia, and Salvatore Maria Aglioti. The neural basis of body form and body action agnosia. *Neuron*, 60(2): 235–246, 2008.

[50] Cosimo Urgesi, Giovanni Berlucchi, and Salvatore M Aglioti. Magnetic stimulation of extrastriate body area impairs visual processing of nonfacial body parts. *Current Biology*, 14(23): 2130–2134, 2004.

[51] David Pitcher, Vincent Walsh, Galit Yovel, and Bradley Duchaine. Tms evidence for the involvement of the right occipital face area in early face processing. *Current Biology*, 17(18): 1568–1573, 2007.

[52] David Pitcher, Lucie Charles, Joseph T Devlin, Vincent Walsh, and Bradley Duchaine. Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Current Biology*, 19(4):319–324, 2009.

[53] Sebastian Moeller, Trinity Crapse, Le Chang, and Doris Y Tsao. The effect of face patch microstimulation on perception of faces and objects. *Nature neuroscience*, 20(5):743–752, 2017.

[54] Srivatsun Sadagopan, Wilbert Zarco, and Winrich A Freiwald. A causal relationship between face-patch activity and face-detection behavior. *Elife*, 6:e18558, 2017.

[55] Nancy Kanwisher and Jason JS Barton. The functional architecture of the face system: Integrating evidence from fmri and patient studies. *The Oxford handbook of face perception*, pages 111–129, 2011.

[56] Charles G Gross. Genealogy of the "grandmother cell". *The Neuroscientist*, 8(5):512–518, 2002.

[57] Alumit Ishai, Leslie G Ungerleider, Alex Martin, Jennifer L Schouten, and James V Haxby. Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 96(16):9379–9384, 1999.

[58] Paul E Downing, AW-Y Chan, Marius Vincent Peelen, CM Dodds, and N Kanwisher. Domain specificity in visual cortex. *Cerebral cortex*, 16(10):1453–1461, 2006.

[59] Roozbeh Kiani, Hossein Esteky, Koorosh Mirpour, and Keiji Tanaka. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of neurophysiology*, 97(6):4296–4309, 2007.

[60] Alumit Ishai, Leslie G Ungerleider, Alex Martin, and James V Haxby. The representation of objects in the human occipital and temporal cortex. *Journal of Cognitive Neuroscience*, 12 (Supplement 2):35–51, 2000.

[61] Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.

[62] Marieke Mur, Douglas A Ruff, Jerzy Bodurka, Peter De Weerd, Peter A Bandettini, and Nikolaus Kriegeskorte. Categorical, yet graded–single-image activation profiles of human category-selective cortical regions. *Journal of Neuroscience*, 32(25):8649–8662, 2012.

[63] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[64] SueYeon Chung and LF Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current opinion in neurobiology*, 70:137–144, 2021.

[65] David D Cox and Robert L Savoy. Functional magnetic resonance imaging (fmri)"brain reading": detecting and classifying distributed patterns of fmri activity in human visual cortex. *Neuroimage*, 19(2):261–270, 2003.

[66] Evelyn Eger, John Ashburner, John-Dylan Haynes, Raymond J Dolan, and Geraint Rees. fmri activity patterns in human loc carry information about object exemplars within category. *Journal of cognitive neuroscience*, 20(2):356–370, 2008.

[67] Andrew C Connolly, J Swaroop Guntupalli, Jason Gors, Michael Hanke, Yaroslav O Halchenko, Yu-Chien Wu, Hervé Abdi, and James V Haxby. The representation of biological classes in the human brain. *Journal of Neuroscience*, 32(8):2608–2618, 2012.

[68] Rebecca F Schwarzlose, Jascha D Swisher, Sabin Dang, and Nancy Kanwisher. The distribution of category and location information across object-selective regions in human visual cortex. *Proceedings of the National Academy of Sciences*, 105(11):4447–4452, 2008.

[69] Mona Spiridon and Nancy Kanwisher. How distributed is visual category information in human occipito-temporal cortex? an fmri study. *Neuron*, 35(6):1157–1165, 2002.

[70] Mark A Williams, Sabin Dang, and Nancy G Kanwisher. Only some spatial patterns of fmri response are read out in task performance. *Nature neuroscience*, 10(6):685–686, 2007.

[71] Leila Reddy and Nancy Kanwisher. Category selectivity in the ventral visual pathway confers robustness to clutter and diverted attention. *Current Biology*, 17(23):2067–2072, 2007.

[72] Gerwin Schalk, Christoph Kapeller, Christoph Guger, Hiroshi Ogawa, Satoru Hiroshima, Rosa Lafer-Sousa, Zeynep M Saygin, Kyousuke Kamada, and Nancy Kanwisher. Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain. *Proceedings of the National Academy of Sciences*, 114(46):12285–12290, 2017.

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

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

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

[76] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

[77] Talia Konkle and George A Alvarez. A self-supervised domain-general learning framework for human ventral stream representation. *Nature communications*, 13(1):491, 2022.

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

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

[80] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[81] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[82] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020. https://arxiv.org/abs/2006.09882.

[83] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[84] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann Lecun. On the duality between contrastive and non-contrastive self-supervised learning. *arXiv preprint arXiv:2206.02574*, 2022.

[85] Chenxin Tao, Honghui Wang, Xizhou Zhu, Jiahua Dong, Shiji Song, Gao Huang, and Jifeng Dai. Exploring the equivalence of siamese self-supervised learning via a unified gradient framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14431–14440, 2022.

[86] Weiran Huang, Mingyang Yi, and Xuyang Zhao. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021.

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

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

[89] Seungdae Baek, Min Song, Jaeson Jang, Gwangsu Kim, and Se-Bum Paik. Face detection in untrained deep neural networks. *Nature communications*, 12(1):7328, 2021.

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

[91] Le Chang and Doris Y Tsao. The code for facial identity in the primate brain. *Cell*, 169(6): 1013–1028, 2017.

[92] Ilker Yildirim, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. Efficient inverse graphics in biological face processing. *Science advances*, 6(10):eaax5979, 2020.

[93] Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):6456, 2021.

[94] Gareth J Edwards, Timothy F Cootes, and Christopher J Taylor. Face recognition using active appearance models. In *Computer Vision—ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II 5*, pages 581–595. Springer, 1998.

[95] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

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

[97] Kasper Vinken, Jacob S Prince, Talia Konkle, and Margaret Livingstone. The neural code for 'face cells' is not face specific. *bioRxiv*, pages 2022–03, 2022.

[98] Michael A Cohen, George A Alvarez, Ken Nakayama, and Talia Konkle. Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of neurophysiology*, 117(1):388–402, 2017.

[99] Chou P Hung, Gabriel Kreiman, Tomaso Poggio, and James J DiCarlo. Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310(5749):863–866, 2005.

[100] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fmri. *Neuroimage*, 56(2):400–410, 2011.

[101] Abdelhak Mahmoudi, Sylvain Takerkart, Fakhita Regragui, Driss Boussaoud, Andrea Brovelli, et al. Multivoxel pattern analysis for fmri data: a review. *Computational and mathematical methods in medicine*, 2012, 2012.

[102] Chris Hamblin, Talia Konkle, and George Alvarez. Pruning for interpretable, feature-preserving circuits in cnns. *arXiv preprint arXiv:2206.01627*, 2022.

[103] Michael A Lepori, Thomas Serre, and Ellie Pavlick. Break it down: evidence for structural compositionality in neural networks. *arXiv preprint arXiv:2301.10884*, 2023.

[104] Katharina Dobs, Julio Martinez, Alexander JE Kell, and Nancy Kanwisher. Brain-like functional specialization emerges spontaneously in deep neural networks. *Science advances*, 8(11): eabl8913, 2022.

[105] Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8):e2011417118, 2021.

[106] Stefania Bracci and Marius V Peelen. Body and object effectors: the organization of object representations in high-level visual cortex reflects body–object interactions. *Journal of Neuroscience*, 33(46):18247–18258, 2013.

[107] Stefania Bracci, Alfonso Caramazza, and Marius V Peelen. Representational similarity of body parts in human occipitotemporal cortex. *Journal of Neuroscience*, 35(38):12977–12985, 2015.

[108] Moritz F Wurm, Alfonso Caramazza, and Angelika Lingnau. Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. *Journal of Neuroscience*, 37(3):562–575, 2017.

[109] Leyla Tarhan and Talia Konkle. Sociality and interaction envelope organize visual action representations. *Nature Communications*, 11(1):3002, 2020.

[110] Leyla Tarhan, Julian De Freitas, and Talia Konkle. Behavioral and neural representations en route to intuitive action understanding. *Neuropsychologia*, 163:108048, 2021.

[111] Etienne Abassi and Liuba Papeo. The representation of two-body shapes in the human visual cortex. *Journal of Neuroscience*, 40(4):852–863, 2020.

[112] Daniel Janini, Chris Hamblin, Arturo Deza, and Talia Konkle. General object-based features account for letter perception. *PLOS Computational Biology*, 18(9):e1010522, 2022.

[113] Ghislaine Dehaene-Lambertz, Karla Monzalvo, and Stanislas Dehaene. The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLoS biology*, 16(3):e2004103, 2018.

[114] Bo Zhang, Sheng He, and Xuchu Weng. Localization and functional characterization of an occipital visual word form sensitive area. *Scientific reports*, 8(1):6723, 2018.

[115] Krishna Srihasam, Joseph B Mandeville, Istvan A Morocz, Kevin J Sullivan, and Margaret S Livingstone. Behavioral and anatomical consequences of early versus late symbol training in macaques. *Neuron*, 73(3):608–619, 2012.

[116] Margaret S Livingstone, Justin L Vincent, Michael J Arcaro, Krishna Srihasam, Peter F Schade, and Tristram Savage. Development of the macaque face-patch system. *Nature communications*, 8(1):14897, 2017.

[117] Michael J Arcaro, Peter F Schade, Justin L Vincent, Carlos R Ponce, and Margaret S Livingstone. Seeing faces is necessary for face-domain formation. *Nature neuroscience*, 20(10):1404–1412, 2017.

[118] Jesse Gomez, Michael Barnett, and Kalanit Grill-Spector. Extensive childhood experience with pokémon suggests eccentricity drives organization of visual cortex. *Nature human behaviour*, 3 (6):611–624, 2019.

[119] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2018.

[120] Richard Le Grand, Catherine J Mondloch, Daphne Maurer, and Henry P Brent. Early visual experience and face processing. *Nature*, 410(6831):890–890, 2001.

[121] Linda B Smith. Learning to recognize objects. *Psychological Science*, 14(3):244–250, 2003.

[122] Gal Raz and Rebecca Saxe. Learning in infancy is active, endogenously motivated, and depends on the prefrontal cortices. *Annual Review of Developmental Psychology*, 2:247–268, 2020.

[123] Alfredo F Pereira and Linda B Smith. Developmental changes in visual object recognition between 18 and 24 months of age. *Developmental science*, 12(1):67–80, 2009.

[124] Yuri Ostrovsky, Ethan Meyers, Suma Ganesh, Umang Mathur, and Pawan Sinha. Visual parsing after recovery from blindness. *Psychological Science*, 20(12):1484–1491, 2009.

[125] Jessica Sullivan, Michelle Mei, Andrew Perfors, Erica Wojcik, and Michael C Frank. Saycam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open mind*, 5: 20–29, 2021.

[126] Elizabeth S Spelke. *What Babies Know: Core Knowledge and Composition Volume 1*, volume 1. Oxford University Press, 2022.

[127] Le Chang, Bernhard Egger, Thomas Vetter, and Doris Y Tsao. Explaining face representation in the primate brain using different computational models. *Current Biology*, 31(13):2785–2795, 2021.

[128] Shany Grossman, Guy Gaziv, Erin M Yeagle, Michal Harel, Pierre Mégevand, David M Groppe, Simon Khuvis, Jose L Herrero, Michal Irani, Ashesh D Mehta, et al. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature communications*, 10(1):4934, 2019.

[129] Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, 33(10):2044–2064, 2021.

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

[131] Pulkit Agrawal, Dustin Stansbury, Jitendra Malik, and Jack L. Gallant. Pixels to voxels: Modeling visual representation in the human brain, 2014.

[132] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152: 184–194, 2017.

[133] Tom Dupre la Tour, Michael Lu, Michael Eickenberg, and Jack L Gallant. A finer mapping of convolutional neural network layers to the visual cortex. In *SVRHM 2021 Workshop@ NeurIPS*, 2021.

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

[135] Martin Schrimpf, 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. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv preprint*, 2018. doi: 10.1101/407007.

[136] Andrew Jaegle, Vahid Mehrpour, Yalda Mohsenzadeh, Travis Meyer, Aude Oliva, and Nicole Rust. Population response magnitude variation in inferotemporal cortex predicts image memorability. *Elife*, 8:e47596, 2019.

[137] T Anderson Keller, Qinghe Gao, and Max Welling. Modeling category-selective cortical regions with topographic variational autoencoders. *arXiv preprint arXiv:2110.13911*, 2021.

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

[139] Adrien Doerig, Rowan Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace Lindsay, Konrad Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. The neuroconnectionist research programme. *arXiv preprint arXiv:2209.03718*, 2022.

[140] Ian ML Pennock, Chris Racey, Emily J Allen, Yihan Wu, Thomas Naselaris, Kendrick N Kay, Anna Franklin, and Jenny M Bosten. Color-biased regions in the ventral visual pathway are food selective. *Current Biology*, 33(1):134–146, 2023.

[141] Nidhi Jain, Aria Wang, Margaret M Henderson, Ruogu Lin, Jacob S Prince, Michael J Tarr, and Leila Wehbe. Selectivity for food in human ventral visual cortex. *Communications Biology*, 6 (1):175, 2023.

[142] Meenakshi Khosla, N Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32(19):4159–4171, 2022.

[143] Tal Golan, Wenxuan Guo, Heiko H Schütt, and Nikolaus Kriegeskorte. Distinguishing representational geometries with controversial stimuli: Bayesian experimental design and its application to face dissimilarity judgments. *arXiv preprint arXiv:2211.15053*, 2022.

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

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

[146] Leyla Tarhan and Talia Konkle. Reliability-based voxel selection. *Neuroimage*, 207:116350, 2020.

[147] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[148] cvnlab. GSN: Generative modeling of signal and noise. https://github.com/cvnlab/GSN, 2022.

[149] Aylin Kallmayer, Jacob Prince, and Talia Konkle. Comparing representations that support object, scene, and face recognition using representational trajectory analysis. *Journal of Vision*, 20(11): 861–861, 2020.

A contrastive coding framework of category selectivity in the ventral visual stream

Jacob S. Prince, George A. Alvarez & Talia Konkle

Supplementary Information

Supplementary Figures

Supplementary Figure 1: Layer summary of localizer outcomes across DNNs

Supplementary Figure 2: Selectivity profiles do not generalize in an untrained DNN

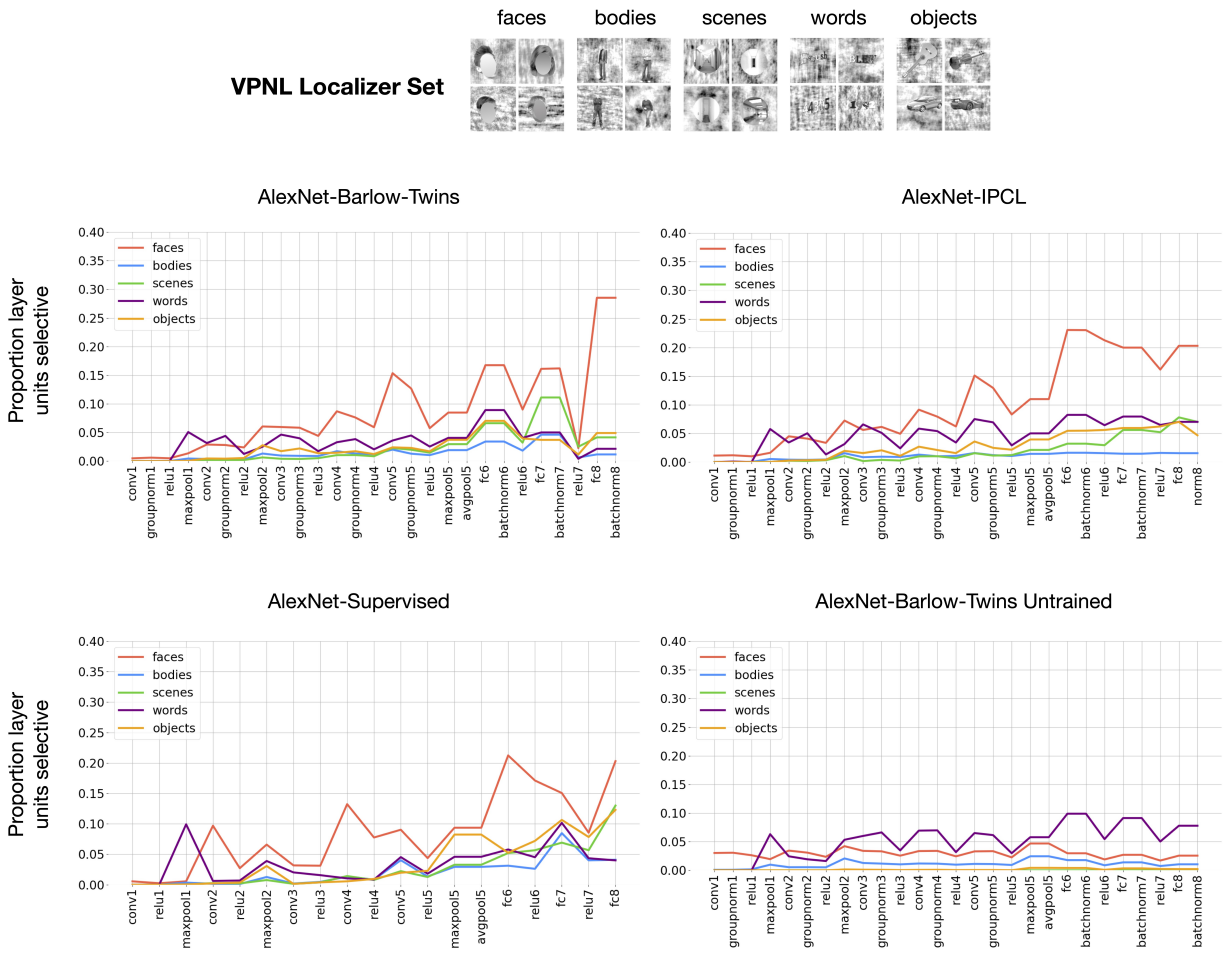
Supplementary Figure 3: Summary of selective, dissociable lesioning deficits

Supplementary Figure 4: Relationship between activation and recognition deficit for selective-unit and randomized lesions

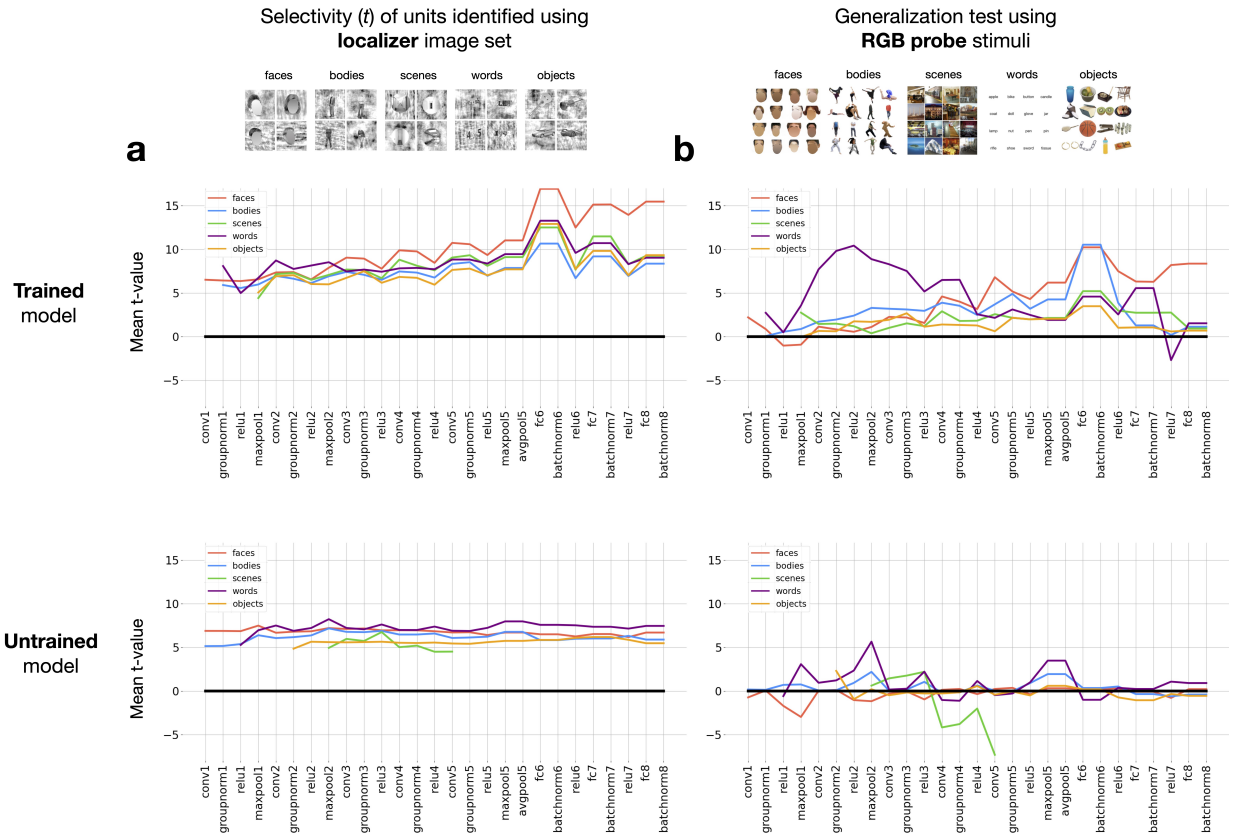
Supplementary Figure 5: Impact of sparse-positive readout (univariate profile prediction)

Supplementary Figure 6: Impact of sparse-positive readout (multivariate RDM prediction)

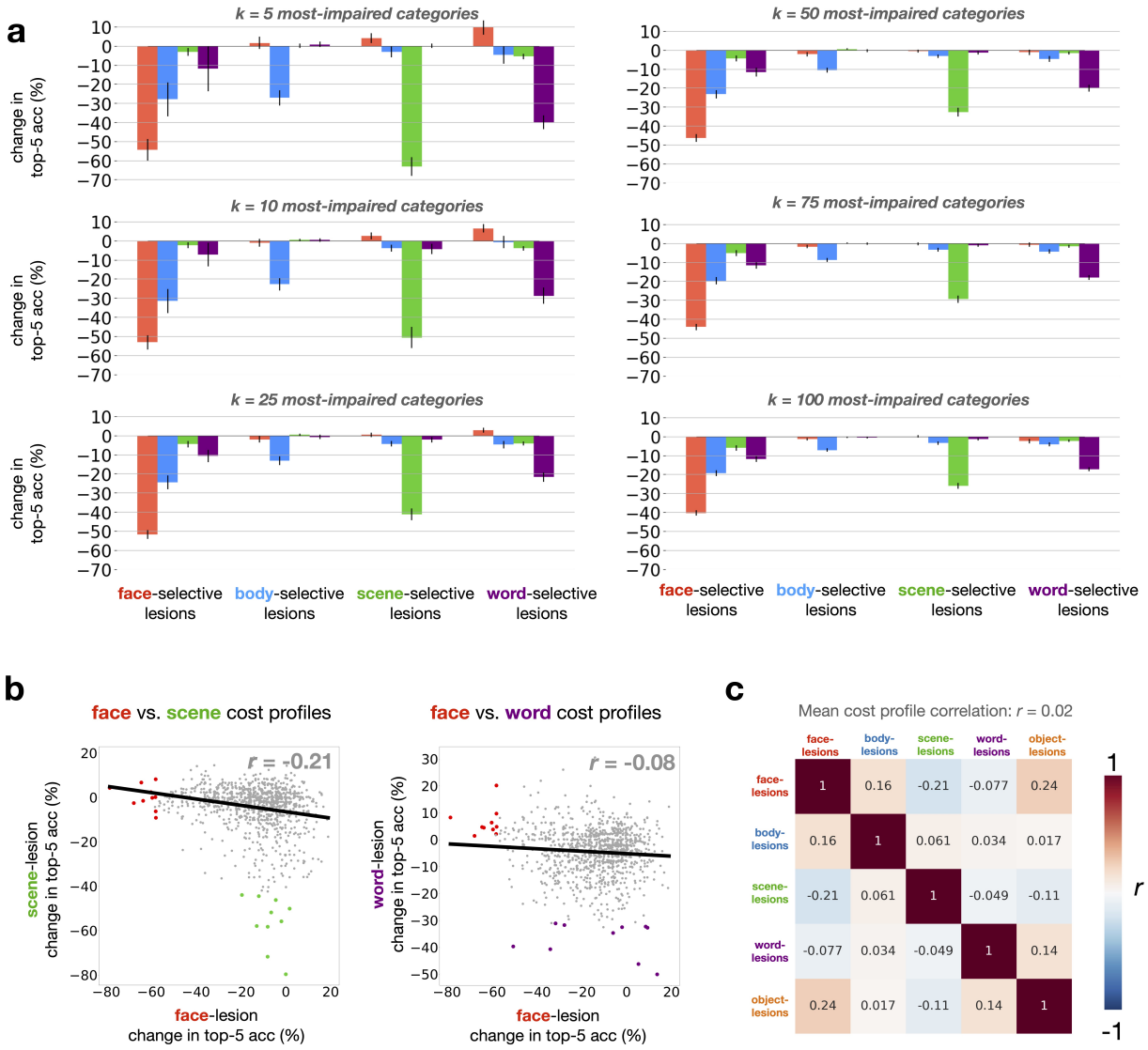
Supplementary Figure 7: Indices of the most brain-predictive DNN layers



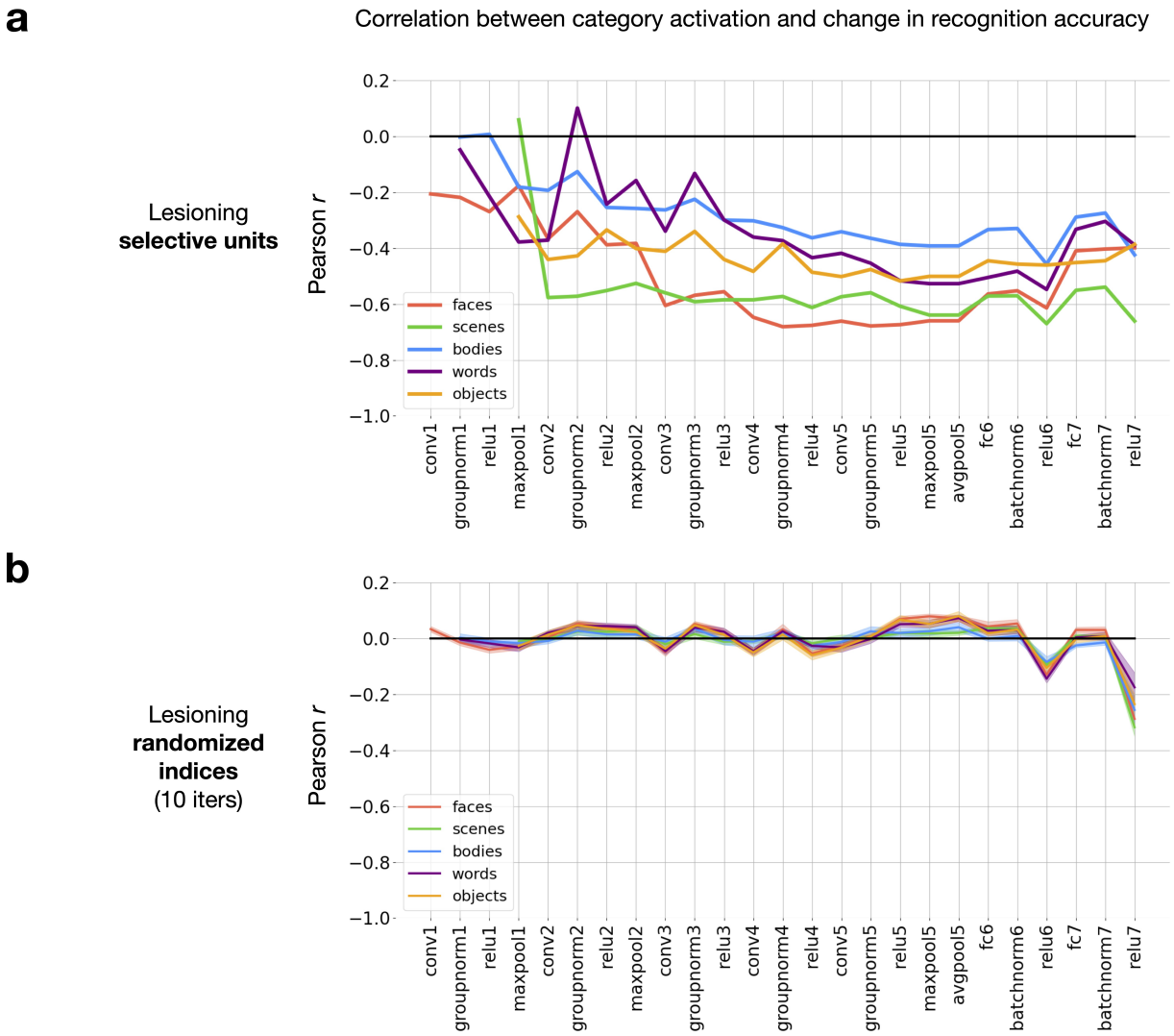
Supplementary Figure 1: Layer summary of localizer outcomes across DNNs. *Layerwise proportions of units selective for the 5 localizer domains are plotted for the trained AlexNet Barlow Twins model (top left), the AlexNet IPCL model (top right), a category-supervised AlexNet ImageNet model (TorchVision pretrained implementation; bottom left), and an untrained initialization of the Barlow Twins model (bottom right). Images containing identifiable human faces are masked.*



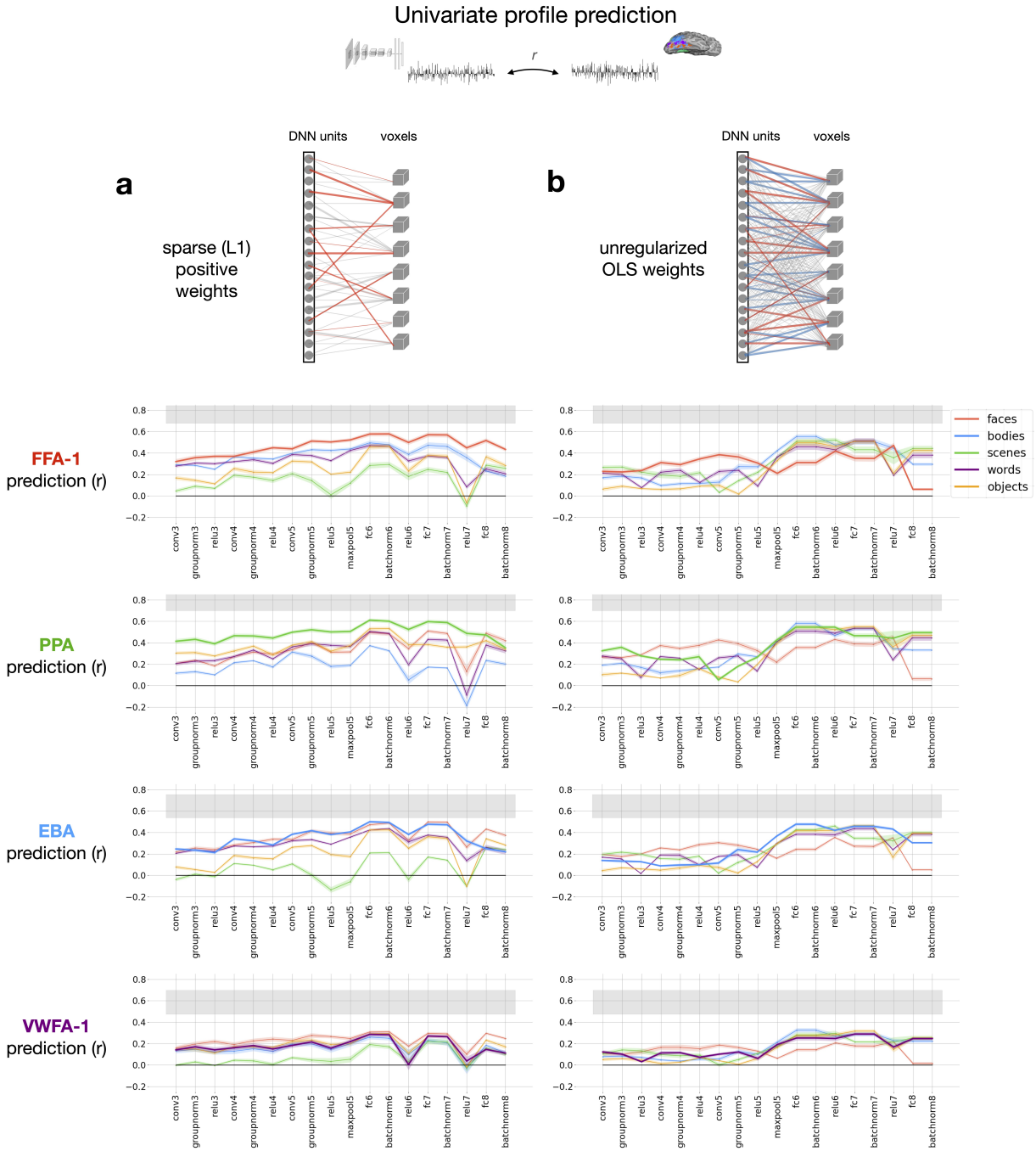
Supplementary Figure 2: Selectivity profiles do not generalize in an untrained DNN. *(A) Strength of units' selectivities for their preferred domains from the initial localizer run (vpnl-fLoc stimuli). Selectivities are summarized across model layers for the trained AlexNet Barlow Twins model (top) and the untrained model (bottom). The y-axis shows the mean t-statistic within each group of selective units (averaged both over units and over all statistical contrasts involving the target domain, e.g. faces vs. scenes, faces vs. bodies, etc.) (B) Generalization of selectivity to the independent RGB probe set. The t-values are reported from the same groups of units (defined using the vpnl-floc set), with the contrast t-values now computed over activations from the RGB image set. Images containing identifiable human faces are masked.*



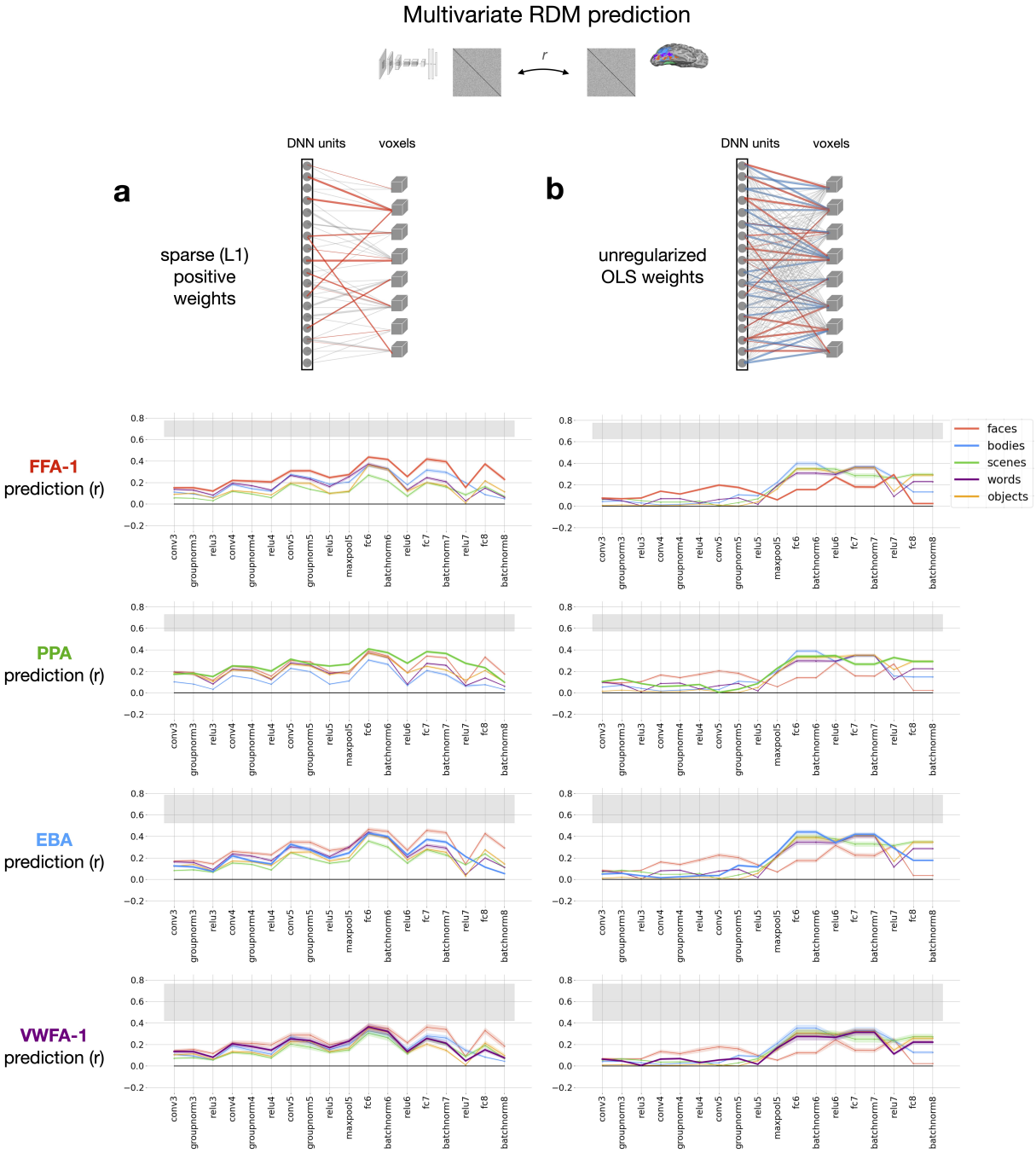
Supplementary Figure 3: Summary of selective, dissociable lesioning deficits. *(A) Bar graphs show the mean (± SEM) change in % top-5 recognition accuracy over the top k categories most impacted by lesions to the selective units of each domain. Indices of the k categories are identified using half the ImageNet validation set, and accuracies for plotting are computed using the other half. (B) Relationship between 1000-dimensional profile of category-level lesioning deficits for face vs. scene unit lesions (left) and face vs. word unit lesions (right). Colored dots reflect the top 10 categories impacted by each domain's lesion. (C) Full pairwise similarity matrix (r) comparing the 1000-dim profiles of lesioning deficits that arise from lesions to each domain of selective units.*



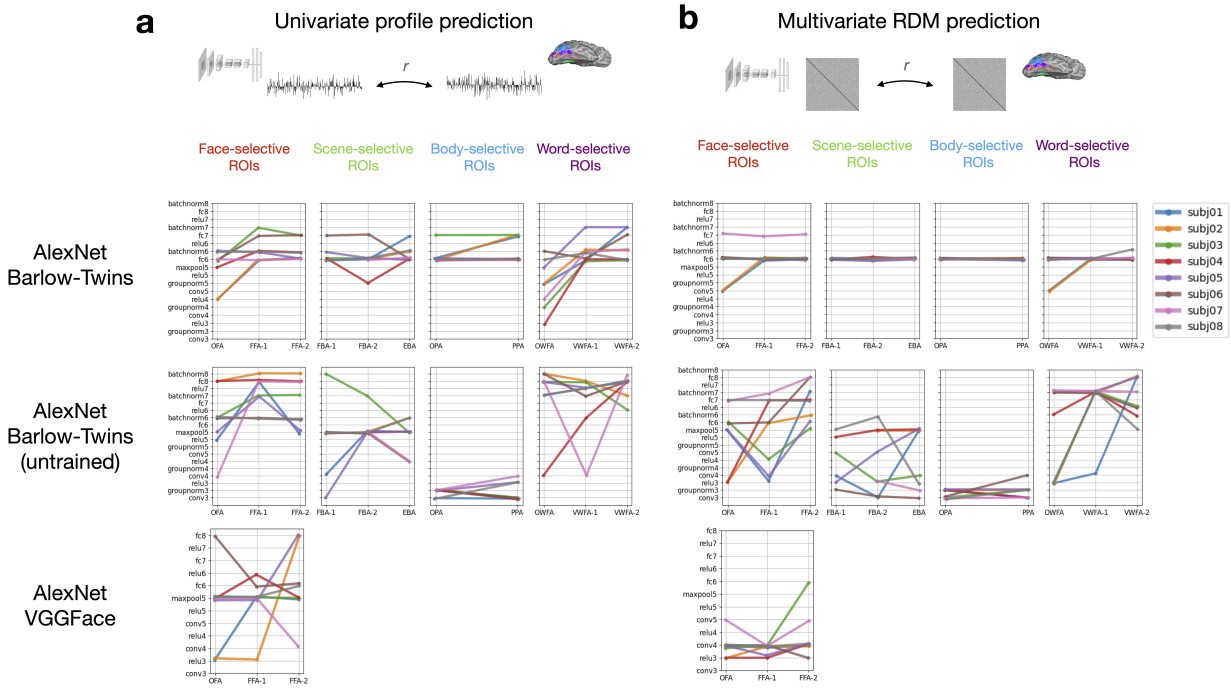
Supplementary Figure 4: Relationship between activation and recognition deficit for selective-unit and randomized lesions. *(A) Layer summary of the relationship between 1000-category profiles of mean activation magnitude (in the unlesioned model) and the changes in recognition accuracy observed after lesioning each domain's selective units. Data plotted are from the AlexNet Barlow Twins model, and lesions are applied to layers relu1-7. (B) The same analysis as (A), except that the indices of target units are randomized (within each layer) prior to lesioning. The randomized lesion experiment is repeated 10 times, with the mean ± SEM of resultant correlation values plotted.*



Supplementary Figure 5: Impact of sparse-positive readout (univariate profile prediction). *Layer summaries of encoding results are plotted for prediction of univariate response profiles in FFA-1, PPA, EBA, and VWFA-1 (N = 8 subjects, n = 515 test images). All data are from the AlexNet Barlow Twins model, with line colors representing the domain of selective units that are mapped onto the target ROI. Prediction results are compared between (A) the constrained encoding procedure involving sparse-positive weights only, and (B) an unregularized encoding procedure that uses ordinary least-squares (OLS) fitting, with no additional constraints. Shaded regions reflect the range of subject-specific noise ceilings computed over the univariate response profiles from each ROI.*



Supplementary Figure 6: Impact of sparse-positive readout (multivariate RDM prediction). *Layer summaries of encoding results are plotted for prediction of multivariate RDMs in FFA-1, PPA, EBA, and VWFA-1 (N = 8 subjects, n = 132,355 unique pairwise comparisons per RDM). All data are from the AlexNet Barlow Twins model, with line colors representing the domain of selective units that are mapped onto the target ROI. Prediction results are compared between (A) the constrained encoding procedure involving sparse-positive weights only, and (B) an unregularized encoding procedure that uses ordinary least-squares (OLS) fitting, with no additional constraints. Shaded regions reflect the range of subject-specific noise ceilings computed over the RDMs from each ROI.*



Supplementary Figure 7: Indices of the most brain-predictive DNN layers. *Indices of the most predictive DNN layers for univariate (A) and multivariate (B) encoding are identified using a validation set of 1000 subject-specific NSD stimuli. Lines reflect the outcomes for the 8 NSD subjects, across all 11 functional ROIs, for the three DNN models included in the encoding procedure.*