UNIVERSAL DIMENSIONS OF VISUAL REPRESENTATION

Zirui Chen & Michael F. Bonner
Department of Cognitive Science
Johns Hopkins University, Baltimore, 21218
zchen160@jhu.edu, mfbonner@jhu.edu

ABSTRACT

Do neural network models of vision learn brain-aligned representations because they share architectural constraints and task objectives with biological vision or because they learn universal features of natural image processing? We characterized the universality of hundreds of thousands of representational dimensions from visual neural networks with varied construction. We found that networks with varied architectures and task objectives learn to represent natural images using a shared set of latent dimensions, despite appearing highly distinct at a surface level. Next, by comparing these networks with human brain representations measured with fMRI, we found that the most brain-aligned representations in neural networks are those that are universal and independent of a network's specific characteristics. Remarkably, each network can be reduced to fewer than ten of its most universal dimensions with little impact on its representational similarity to the human brain. These results suggest that the underlying similarities between artificial and biological vision are primarily governed by a core set of universal image representations that are convergently learned by diverse systems.

1 Introduction

Deep neural networks have a remarkable ability to simulate the representations of biological vision [15, 30, 5, 8]. However, due to their immense complexity, the principles that govern the brain-aligned representations of deep networks remain poorly understood.

A leading approach interprets neural network representations in terms of their architectures and task objectives, which are thought to function as key constraints on a network's learned representations [29, 24, 3, 7, 14]. However, an alternative possibility is that the brain-aligned representations of neural networks are not contingent on specific optimization constraints but instead reflect universal aspects of natural image representation that emerge in diverse systems [10, 13, 8].

Here we sought to determine if the representations that neural networks share with human vision are universal across networks. We examined over 200,000 dimensions of natural image representation in deep neural networks with varied designs. Our analyses revealed the existence of universal dimensions that are shared across networks and emerge under highly varied optimization conditions. Universal dimensions were observed across the full depth of network layers and across a variety of architectures and task objectives. Visualizations of these dimensions show that they do not simply encode low-level image statistics but also higher-level semantic properties. We next compared these dimensions to the representations of the human brain measured with fMRI, and we found that universal dimensions are highly brain-aligned and underlie conventional measures of representational similarity between neural networks and visual cortex. Together, these findings demonstrate the striking degree to which the shared properties of artificial and biological vision correspond to general-purpose representations that have little do to with the details of a network's architecture or task objective.

2 Results

2.1 Assessing universality and brain similarity

We sought to compare two fundamental quantities of representational dimensions in neural networks: 1) their universality across varied networks and 2) their similarity to human brain representations. Here we briefly describe how we computed these two quantities. A more detailed description is provided in the Methods.

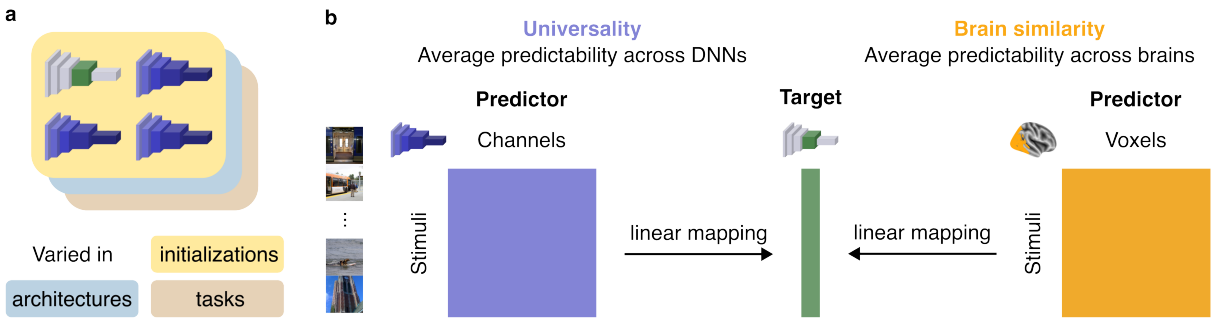

Figure 1: **Overview of method for computing universality and brain similarity of network dimensions. (a)** Four sets of deep neural networks were analyzed, including three sets of trained models that varied in either their random initializations, architectures, or task objectives and one set of untrained models with different initializations. **(b)** Universality and brain similarity were defined as the average prediction accuracy of a latent dimension from a target network when using the activations of other networks or the fMRI activations of the human brain as predictors. Dimensions that can be consistently predicted from the representations of other networks have high universality. Dimensions that can be consistently predicted from the representations of the human brain have high brain similarity.

As illustrated in Figure 1, we characterized universality and brain similarity by examining the activations of networks and the human brain to a large and diverse set of natural images from the Microsoft Common Objects in Context (COCO) image database [20]. For each latent dimension d of a network's activations (i.e., each principal component), we computed a universality metric by obtaining its average predictability from the activations of m other networks:

$$\text{Universality}_d = \text{avg}(\text{r}_{d,1},\ \text{r}_{d,2},\ \ldots,\ \text{r}_{d,m})\,, \tag{1}$$

where $\text{avg}(\cdot)$ is the average and $r_{d,m}$ is the prediction accuracy of a cross-validated linear regression with dimension d as the predictand and network m as the predictor. We performed these analyses on the principal components (PCs) of a network's activations so that each dimension d is sampled from an orthogonal basis rather than from a set of potentially redundant neurons. The motivation for this metric is to quantify the average degree to which a dimension in one network is shared with the representations of other networks. Dimensions that are shared across all networks will have universality scores close to 1, while dimensions that only emerge under specific network configurations will have universality scores close to 0.

For each dimension d of a network's activations, we also computed a brain-similarity metric by obtaining its average predictability from the fMRI activations of n human brains:

$$\text{Brain similarity}_d = \text{avg}(\text{r}_{d,1},\ \text{r}_{d,2},\ \ldots,\ \text{r}_{d,n})\,, \tag{2}$$

where $\text{avg}(\cdot)$ is the average and $r_{d,n}$ is the prediction accuracy of a cross-validated linear regression with dimension d as the predictand and subject n as the predictor. Here the motivation is to quantify the average degree to which a network dimension is shared with the representations of the human visual cortex. Dimensions that are shared between networks and humans will have brain similarity scores close to 1, while dimensions that are not shared with humans will have brain similarity scores close to 0.

We examined the universality of representational dimensions among several sets of vision networks that varied in their random initializations, architectures, or task objectives. The specific networks used for these analyses are described in 4.2.1 and listed in Tables S1 and S2. To assess brain similarity, we compared network representations with image-evoked fMRI responses from the Natural Scenes Dataset (NSD) [1], which is the largest existing fMRI dataset of natural scene perception in the human brain. We focused on a portion of this dataset that contains fMRI responses to 872 images shown to each of eight participants. This dataset is ideally suited for assessing whether the dimensions of natural image representations in neural networks can also be found in the representations of the human brain. For our main analyses, we focused on a general region of interest that included all voxels in visual cortex whose activity was modulated by the presentation of visual stimuli (Fig. 1).

2.2 Universality across initialization weights

We used universality and brain similarity to address a central question: Are there universal dimensions of natural image representation that are shared by neural networks and humans? We first performed these analyses for a setting in

which we naturally expected to find shared network dimensions. Specifically, we examined universality among a set of networks that were initialized with different random weights but were otherwise identical (i.e., same architecture, task, and training data). These networks were 20 ResNet-18 architectures trained on image classification using the Tiny ImageNet dataset [11, 26, 19]. For each dimension in each network layer, we computed its universality using the other networks as predictors. We examined network layers spanning the full range of model depth, except for the final classification layer. We iterated this analysis over a total of 36,596 dimensions across all networks.

As shown in the left panel of Figure 2, we observed universality scores spanning the full range from 0 to 1, with a high density of points around 0, indicating that most dimensions are idiosyncratic. The high density of idiosyncratic dimensions could reflect representations present at initialization that remain largely unchanged during training, or they could reflect unique representational strategies learned by specific network instances. In contrast, the universal dimensions at the other end of the scale reflect convergent representations that reliably emerge in all networks despite differences in their starting points. Notably, these universal dimensions account for a relatively small subset of the total number of network dimensions, as illustrated by the lower density of points at the high end of the universality axis.

We next compared these network dimensions to human brain representations, and we found that the universal dimensions exhibit exceptionally strong brain similarity scores (Fig. 2). This demonstrates that among the many network dimensions examined here, it is only those that are invariably learned by networks with different initial conditions that are also strongly shared with the representations of the human visual system.

In follow-up analyses, we found that the trend in Figure 2 was consistently observed in each network layer (Fig. 3), each individual network (Fig. S1), each individual fMRI subject (Fig. S2), and in multiple regions of interest in visual cortex (Fig. S3). The observation of this effect in all network layers shows that universality is not restricted to low-level features in early layers but instead extends across the full depth of the network. We emphasize that the universality and brain similarity metrics in our analysis pipeline are not guaranteed to be related to one another. Using simulated data, we can trivially obtain a range of universality and brain similarity scores while observing no positive relationship between these two metrics (Fig. S4). Together, these findings show that among a preponderance of idiosyncratic network dimensions, there exists a smaller subset of highly convergent dimensions that are learned by different network instances and are shared with human vision.

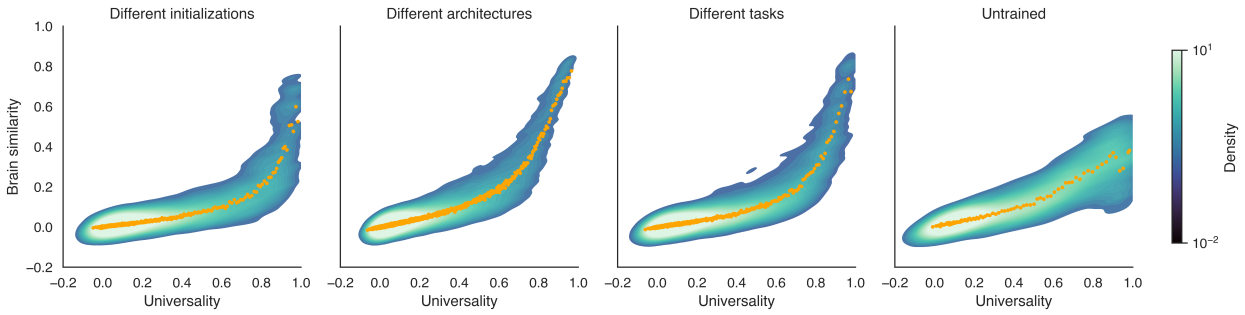

Figure 2: **Universality and brain similarity of network dimensions.** Universality and brain similarity were computed for representational dimensions in four sets of deep neural networks. These included three sets of trained networks with varied initializations, architectures, and task objectives and one set of untrained networks. These metrics were computed for the principal components of network activations extracted from the sampled layers of each network. Universality scores reflect the degree to which a representational dimension is shared across all networks in a set, and brain similarity scores reflect the degree to which a representational dimension is shared with the human visual system. Measurements of human visual cortex activity were obtained from the Natural Scenes fMRI Dataset using a general region of interest that included all visually responsive voxels [1]. Universality and brain similarity scores are plotted for all analyzed network dimensions. These plots show the density of dimensions on a logarithmic scale, with densities computed using kernel density estimation. The orange dots show the mean universality and brain similarity scores for equally sized quantiles of 100 dimensions along the x-axis. These plots show similar trends for all three sets of trained models (the first three plots on the left). Specifically, they exhibit a high density of points near the origin, showing that most dimensions are idiosyncratic to each network and are not shared with the human brain. However, they also contain a subset of dimensions with exceptionally high universality and brain similarity scores. These latter dimensions correspond to representations that are consistently learned by all networks within a set and are also strongly shared with the visual representations of the human brain. In contrast, untrained networks (right panel) can also have shared dimensions, but these shared untrained dimensions have relatively weak brain similarity scores.

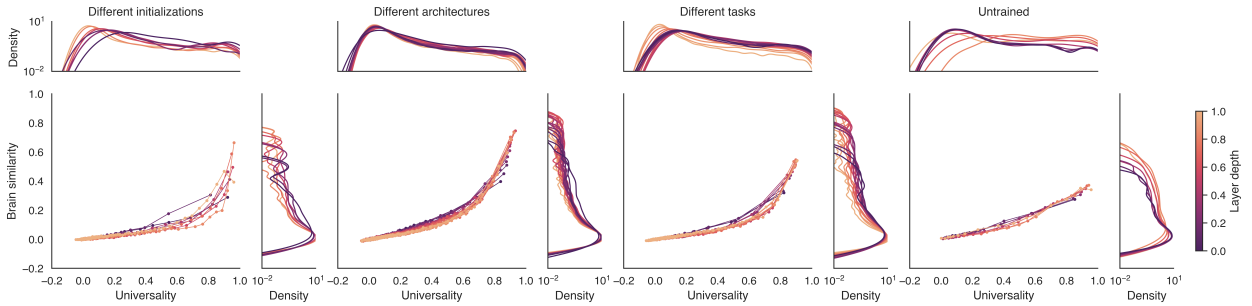

Figure 3: **Universality and brain similarity across network layers.** These plots show the universality and brain similarity scores for individual network layers spanning the full depth of each network. Four sets of deep neural networks were examined, including three sets of trained networks with varied initializations, architectures, and tasks and one set of untrained networks. The analyses are the same as in Figure 2, but here the results are plotted as the average values for individual layers, which are labeled according to their relative depth. Further details of these network layers are included in the supplement file model_layer.csv. Average universality and brain similarity scores were computed for equally sized quantiles of 100 dimensions along the x-axis for each layer. Panels on the sides of each plot show the density of dimensions on a logarithmic scale computed using kernel density estimation. As in Figure 2, these plots exhibit a high density of points near the origin, which means that across all sampled layers, most dimensions are idiosyncratic and are not shared with the human brain. However, the three sets of trained networks (first three plots on the left) also contain a subset of dimensions at the right end of each plot that have exceptionally high universality and brain similarity scores. Importantly, these layer-wise plots show that a consistent trend is observed across all sampled layers and that universal dimensions are not restricted to early network layers.

2.3 Universality across architectures and tasks

We next sought to determine if universal dimensions can be detected among networks with varied architectures and tasks. To this end, we quantified universality and brain similarity for two sets of models. The first was a set of models with different architectures but trained on the same task. Specifically, we examined 19 networks trained to perform ImageNet classification using varied architectures, including convolutional models, vision transformers, and MLP-Mixers. The second was a set of 9 models with the same architecture (ResNet-50) but trained on different tasks. The tasks included object classification and a variety of self-supervised tasks, such as contrastive learning, identifying image rotations, and solving jigsaw puzzles. For all models, we examined layers spanning the full network depth, except for the final layer. Further details about the models can be found in Tables S1 and S2. In total, we examined 149,743 dimensions in the set of varied architectures and 43,132 dimensions in the set of varied tasks.

The findings for these two sets of models were surprisingly consistent with those observed for models with varied initializations (Fig. 2). We again found that most dimensions are idiosyncratic (i.e., specific to a model) and not shared with the brain, as shown by the high density of points near the origin. Again, we also found that a subset of dimensions exhibit exceptionally high scores on both the universality and brain similarity metrics. These latter dimensions correspond to representations that reliably emerge across many models despite variations in their architectures and the tasks that they were trained to perform. Furthermore, the generality of these representations extends beyond artificial vision, as they are also strongly shared with the representations of the human visual system. Remarkably, these findings are highly similar when considering networks that vary in either architectures or task objectives. This suggests that the underlying similarities among the representations of these networks—as well as their similarities to human vision—are only weakly influenced by architecture and task but instead reflect highly general properties of image representations in deep networks.

We also note that there is a striking paucity of points in the upper left quadrant of the plots in Figure 2. Points in this quadrant would correspond to representations shared with the brain but learned only by networks with specific optimization constraints—namely, specific architectures or tasks. The lack of points in this quadrant suggests that the details of architectures and tasks have a relatively minor role in shaping the brain-aligned representations of neural networks.

In follow-up analyses, we again found that these results are highly robust—they were observed in each network layer (Fig. 3), each individual network (Fig. S1), each individual fMRI subject (Fig. S2), and in multiple regions of interest in visual cortex (Fig. S3). Together, these findings reveal the remarkable degree to which vision systems with varied architectures and task objectives can nonetheless converge on a set of general-purpose representations that are shared not only across models but also between artificial and biological vision.

2.4 Universality across untrained networks

Our analyses thus far have focused on sets of trained neural networks, and we have interpreted the dimensions in the upper right quadrant of the plots in Figure 2 as *learned* representations. However, we expect that shared dimensions can also be found among sets of untrained models due to statistical regularities in the activations that natural images elicit in networks with random filters. We thus wondered whether our findings for the trained networks could be explained by the statistics of image activations alone—without any need for learning—or whether they diverge from the trends observed in randomly initialized networks. To address this question, we examined 20 ResNet-18 architectures that were randomly initialized with the same seeds as the trained models presented in the left panel of Figure 2. We followed the same procedures as in the preceding analyses. For each dimension in each network layer, we computed its universality using the other networks as predictors, and we iterated this analysis over a total of 9,413 dimensions from all networks.

These analyses showed that, as expected, there is a wide range of universality scores for the untrained network dimensions, with some dimensions that are found in all networks (Fig. 2). These universal dimensions of untrained networks correspond to representations that consistently emerge when propagating natural images through a hierarchy of random convolutional filters. They are thus due to image statistics alone and not learned representational properties. However, importantly, the relationship between universality and brain similarity diverges from the relationship that was observed for trained models. Specifically, for untrained networks, we observe a shallow and approximately linear relationship between universality and brain similarity, whereas for trained networks, brain similarity exhibits a sharp nonlinear increase at the high end of the universality axis. As a result, the shared dimensions of untrained networks have substantially lower brain similarity scores than the shared dimensions of trained networks. As in the previous sets of analyses, we again found that these results were consistent in each network layer (Fig. 3), each individual network (Fig. S1), each individual fMRI subject (Fig. S2), and in multiple regions of interest in visual cortex (Fig. S3). In sum, when comparing the trends for trained and untrained networks, the findings demonstrate that the universal dimensions of trained networks reflect *learned* representational properties that cannot be explained by image statistics and random features alone.

2.5 Universality and the visual hierarchy

Previous work has shown that in many neural networks trained on natural images, the first layer contains general-purpose V1-like filters tuned to orientation, frequency, and color, whereas subsequent layers contain filters that appear to be increasingly specialized [18, 31]. This suggests the possibility that universality may only be prominent in early network layers and then rapidly diminish across the network hierarchy. To address this possibility, we examined the universality and brain similarity of network representations in individual layers along the full depth of each network. Figure 3 shows the results of these analyses for all sets of models. Across all sets of trained models, we found relatively similar distributions of universality scores at all sampled layers, with highly universal dimensions detected even in the deepest layers that we examined. Furthermore, these analyses show that the relationship between universality and brain similarity is consistent across layers. Thus, these findings suggest that at all levels of network depth, we can find general-purpose representations that are reliably learned by diverse networks and are strongly shared with the human brain.

2.6 Universal dimensions and high-level image properties

The findings from the previous section show that universal dimensions are not restricted to early network layers. However, these findings do not directly address the question of whether the shared dimensions of later layers represent high-level semantic properties or low-level image statistics, such as luminance gradients and spatial frequency distributions. To address this question, we performed exploratory visualization analyses of the universal dimensions in later network layers. Specifically, we examined the penultimate layer of the models from the varied-task set, as shown in the middle right panel of Figure 2. We focused on the varied-task set because the models in this set have the same ResNet-50 architecture, which allowed us to examine representations from the same targeted layer in all networks. We concatenated the dimensions from the penultimate layer across all networks and ranked their universality scores. We then selected the top 100 dimensions with the highest universality scores and visualized the image representations of these 100 dimensions in a 2D space using uniform manifold approximation and projection (UMAP) [21].

As shown in Figure 4, these representations exhibit rich high-level organization, with images grouped into clusters of semantically related items, such as people, sports, animals, and food. For comparison, we also generated UMAP embeddings of the 100 dimensions with the lowest universality scores from this layer. We found that these idiosyncratic dimensions exhibit no clear semantic organization (Fig. S5). We also generated UMAP embeddings of the top 100 universal dimensions in the penultimate layer of the untrained-network set, which is shown in the rightmost panel of Figure 2. In contrast to the trained networks, the universal dimensions of untrained networks emphasize prominent low-level features, such as coarse luminance gradients (Fig. S6). Together, these visualization analyses show that the universal representations of high-level layers in trained networks encode high-level image properties that group images

into semantically meaningful clusters. This suggests that there are common organizing principles of high-level image semantics that are universally learned.

Figure 4: **Two-dimensional visualization of high-level universal representations.** Image activations for the 100 most universal dimensions from a high-level network layer were embedded in two dimensions using uniform manifold approximation and projection. Specifically, image activations were obtained for the top 100 dimensions with the highest universality scores in the penultimate layer from the set of models trained on different tasks. This plot shows that universal dimensions do not simply reflect low-level image features but instead capture high-level properties that group images into semantically related clusters, some of which are highlighted here, including animals, food, sports, people, and vehicles. In contrast, Figure S5 shows a visualization of the 100 *least* universal dimensions from the same network layer, and it shows no clear semantic organization.

2.7 Universal dimensions and representational similarity analysis

Our findings thus far show that the universal dimensions of networks can be strongly predicted from human brain representations. We next sought to evaluate the effect of universal dimensions on a conventional representational similarity analysis (RSA). Specifically, we performed a targeted analysis to determine if universal dimensions drive the representational similarity scores obtained from comparisons of neural networks with visual cortex. To do so, we conducted a standard RSA on networks with representations reduced to low-dimensional subspaces of their most

universal dimensions. We performed this analysis for all sets of networks examined in Figure 2. Following the RSA procedures in [5], we split the stimuli into training and test sets. We selected the best-performing layer from each network on the training set and computed the final RSA score for the selected layers on the test set. We then reduced each network to a subset of its most universal dimensions and computed RSA scores for these reduced model representations. We analyzed the same general region of interest in visual cortex as in the preceding analyses. We found that even when the networks are reduced to just ten or five universal dimensions, their RSA scores exhibit little or no decrease—in fact, for all three sets of trained networks, they slightly improve (Fig. 5). Similar results were observed when performing these analyses in individual subjects (Fig. S7) and in other regions of interest (Fig. S8). These findings suggest that the representational similarities between neural networks and visual cortex are largely driven by subspaces of network dimensions that are universal.

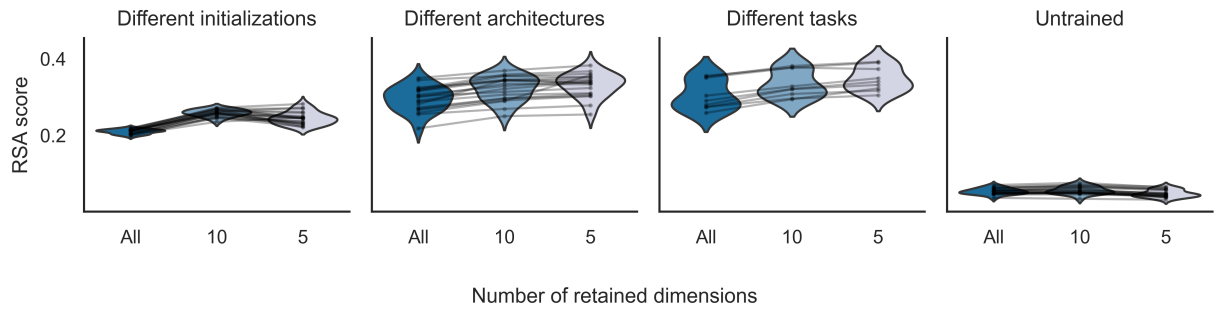

Figure 5: **Universal dimensions underlie the results of conventional representational similarity analyses** Representational similarity analysis (RSA) was used to compare the representations of neural networks and visual cortex. These analyses were performed using the same general region of interest in visual cortex and the same sets of neural networks as in Figures 2 and 3. Representational dissimilarity matrices (RDMs) were created by calculating Pearson correlation distances for pairwise comparisons of image representations within each network and each fMRI subject. RSA scores were obtained by calculating the Spearman correlation between the RDMs for a network and an fMRI subject. These RSA scores were averaged across subjects. For each network, the best-performing layer was selected using a set of training data, and a final RSA score was computed on held-out test data. These analyses show the results of RSA for networks whose representations were either intact or reduced to subspaces of their top ten or five universal dimensions. In these plots, each dot is a network, and lines connect different versions of a network containing either all, ten, or five dimensions. The violin plots show distributions of RSA scores across networks. Even after drastically reducing the networks to just ten or five universal dimensions, the RSA scores exhibit little or no decrease—in fact, for all three sets of trained networks, the RSA scores slightly improve. These results demonstrate that conventional measures of representational similarity between neural networks and visual cortex are largely driven by the subspaces of universal dimensions contained within each network. Similar trends were observed within each individual subject and in other regions of interest, as shown in Figures S7 and S8.

3 Discussion

Our work reveals universal dimensions of natural image representation that are learned by artificial vision systems and are shared with the human brain. These dimensions emerge in diverse neural networks despite variation in their architectures and task objectives. The role of these dimensions in vision appears to be general-purpose—they are not specialized for any single task but instead support many downstream objectives. Universal representations are found at all levels of visual processing in deep networks, from low-level image properties in early layers to high-level semantics in late layers. Together, these findings suggest that machines and humans share a set of general-purpose visual representations that can be learned under highly varied conditions.

Deep learning is now the standard framework for computational modeling in neuroscience, and many previous efforts have sought to understand these deep learning models in terms of their specialization: that is, what objectives they are specialized for, and what specific network characteristics underlie their similarities to the brain [29, 24, 3, 7, 14]. Our work views the representations of deep networks from a different perspective. Rather than searching for specific model characteristics that might be associated with stronger alignment with the brain, we sought to discover the elements of network representations that are instead *invariant* across models. Using this approach, we found that crucial aspects of deep network representations—those that are most strongly shared with the human brain—are, to a remarkable degree, independent of the network characteristics that many previous studies have emphasized. The invariance of these

representations implies that they are not primarily governed by the details of a network's architecture or task objective but instead by more general principles of natural image representation in deep vision systems [10, 13].

Our findings suggest several exciting directions for future work. First, our approach could be extended beyond vision models to examine the representational dimensions that are shared across vision and language. Previous work has shown that language-model embeddings of object names and scene captions are predictive of image representations in high-level visual cortex [4, 2, 6]. An open question is whether networks trained on language data alone learn the same universal dimensions of natural scene representation as image-trained networks. Second, our findings show that universal dimensions emerge in networks despite differences in the tasks that the networks are optimized to perform. This raises the intriguing possibility that universal dimensions could be hard-coded into networks at initialization, potentially making the learning process faster and more data efficient. Third, while previous work has revealed similarities between the visual cortex representations of humans and monkeys [17], we still know little about the degree to which representational dimensions may be universal or species-specific across mammalian vision. This question could be addressed by applying our approach to recordings of cortical responses to the same stimuli in different species.

In sum, our results show that the most brain-aligned representations of visual neural networks are universal and independent of a network's specific characteristics. What fundamental principles might explain the convergence of networks to universal dimensions? Theories of efficient coding suggest that frequency- and orientation-tuned filters are consistently observed in the first layer of vision systems because they constitute efficient bases that are adapted to the statistics of natural images [22, 27]. It remains an open question whether this efficient-coding hypothesis can be extended to a deep hierarchy, which could potentially explain universal dimensions as a consequence of optimal image encoding. An alternative possibility is that deep networks learn shared representations of the true generative factors in the visual world—e.g., the objects, materials, contexts, and optical phenomena that make a scene. This could be the case if the optimal strategy for solving challenging tasks on natural stimuli is to learn the invariant properties of reality [13], and it would suggest that the universal dimensions detected here reflect a shared internal model of the visual environment in machines and humans.

4 Methods

4.1 Natural Scenes Dataset

4.1.1 Stimuli and experimental design

The Natural Scenes Dataset (NSD) is a large-scale publicly available fMRI dataset on human vision that is described in detail in a previous report [1]. Here we briefly review the key attributes of this dataset. The NSD study sourced color natural scene stimuli from the Microsoft Common Objects in Context (COCO) database [20] and collected 7T fMRI responses (1.8mm voxels, 1.6s TR) from eight adult subjects who viewed these stimuli while performing a continuous recognition memory task, namely to respond if the presented image was seen before in the experiment. Each subject viewed approximately 10,000 stimuli with three repetitions, though some subjects saw fewer stimuli because they did not complete all scanning sessions. Among the stimuli, 872 "shared" images were viewed by all subjects at least once. We used these shared images and the corresponding fMRI data for our main analyses.

4.1.2 Data preprocessing

We used the NSD single-trial betas, preprocessed in 1.8-mm volume space and denoised with the GLMdenoise technique (`version 3`; `betas_fithrf_GLMdenoiseRR`). Subsequently, the betas were transformed to z-scores within each individual scanning session, as recommended by the authors [1]. For all analyses, we used the averaged betas across repetitions.

4.1.3 Regions of interest

Our main analyses focused on the `nsdgeneral` region of interest (ROI), which includes a large swath of visual cortex. This ROI, as defined in the NSD study [1], contains all voxels that were reliably modulated by the presentation of visual stimuli, comprising approximately 15,000 voxels in each subject. We also conducted follow-up analyses in smaller ROIs, including the `ventral`, `parietal`, `lateral` streams, using the "streams" ROIs provided in the NSD dataset.

4.2 Deep neural networks

4.2.1 Model sets

We examined four sets of DNN models to probe the universality of representational features across different factors of variation:

- Random seeds in trained models
- Architectures
- Task objectives
- Random seeds in untrained models.

The first set included 20 pretrained ResNet-18 models examined in a previous study of model hyperparameters [26]. These models were initialized with unique random seeds and trained on Tiny ImageNet [19]. We extracted features from nine rectified linear unit (ReLU) layers that span the depth of the ResNet-18 architecture, except for the final output layer. In total, we extracted 36,596 dimensions from this set.

The second set included 19 pretrained models with varied architectures, which included various convolutional networks, transformers, and MLP-Mixers (Table S1). All were trained on ImageNet for object classification [25] and obtained from the torchvision library [23, 28]. Given the variety of architectures in this set, the sampled layers included ReLU, normalization, attention, multi-layer perceptron, and other model-specific operations. The sampled layers span the full range of layer depth in each model (additional details are provided in Table S1). The complete list of these layers are included in the supplement file model_layer.csv. A total of 149,743 dimensions were extracted from this set.

The third set included 9 pretrained ResNet-50 models from the torchvision library [23] and the VISSL model zoo [9]. These models were trained to perform a variety of tasks on ImageNet images [25] (Table S2). A total of 43,132 dimensions were extracted from all ReLU layers, except for the final output layer.

The fourth set included 20 untrained ResNet-18 models [11] with different random weights, which were created using Kaiming normal initialization [12]. Each model had a unique random seed. A total of 9,413 dimensions were extracted from all ReLU layers. Note that the lower number of dimensions here relative to the set of trained models with different random seeds is due to the low-rank activation matrices of untrained networks.

4.2.2 Feature extraction

Before computing our universality and brain-similarity metrics, we first needed to extract a set of feature activations from each model layer. We sought to quantify these metrics for the *features* representations in each network rather than for the spatial representations. We thus applied global max-pooling to remove spatial information from the activations

of each model layer. For convolutional networks, pooling was applied across the height and width dimensions, and for networks with patch embeddings, pooling was applied across patch dimensions. To extract all orthogonal dimensions from each model layer, we first performed PCA on the activations to the 72,128 "unshared" images from NSD, retaining all PCs up to the matrix rank, computed with the default procedure in `torch.linalg.matrix_rank` in PyTorch [23]. We then transformed the 872 "shared" images from NSD to the PC basis and computed universality and brain similarity metrics for each PC.

4.3 Metrics

4.3.1 Universality

Our universality metric estimates the degree to which a representational dimension is shared across multiple DNNs. For a given dimension in a *target* network, we used cross-validated ridge regression to predict its activations as a linear combination of the activations from another *predictor* network. We performed this analysis using the same "shared" NSD images that were used to compute brain similarity (described in 4.3.2). The regressors consisted of activations concatenated across all sampled layers of the predictor network. The procedure for cross-validated ridge regression is described in 4.3.3. We computed the mean Pearson correlation between the predicted and actual responses of the target dimension across all cross-validation folds, and we repeated this process, using every network other than the target as the predictor. We then obtained the universality score by taking the median correlation across all predictor networks. We used median instead of mean to ensure that the final summary statistic was not driven by a small subset of predictor networks with exceptionally high or low scores. This entire procedure was repeated to obtain universality scores for all dimension in all sampled layers of the target network, and it was then repeated with each network as the target.

4.3.2 Brain similarity

Our brain similarity metric estimates the degree to which a dimension in a DNN can be predicted from fMRI responses measured in human visual cortex. Given a target dimension from a network, we used cross-validated ridge regression to predict its activations as a linear combination of the trial-averaged fMRI responses to the "shared" images in a single subject. We computed the mean Pearson correlation between the predicted and actual responses of the target dimension across all cross-validation folds, and we repeated this procedure for all fMRI subjects. Brain similarity is defined as the mean score across all subjects.

4.3.3 Cross-validated ridge regression

We computed universality and brain similarity scores using ridge regression with a nested cross-validation design. The outer loop of this cross-validation design had five folds. We fit the parameters of the ridge regression on four folds of training data. We first selected the optimal ridge penalty for each target dimension from values with equal logarithmic spacing between 10^{-3} to 10^{4}. The optimal ridge penalty was the one that yielded that best performance when applying leave-one-out cross-validation to the training data. We then fit the regression weights using the full set of training data and the optimal ridge penalty, and we applied these regression weights to generate predicted responses in the held-out fold of test data. Performance was evaluated as the correlation between the predicted and actual responses on the held-out test data. This procedure was repeated using all five folds as held-out test data, and the performance scores were average across folds.

4.3.4 Representational similarity analysis

We computed conventional representational similarity analysis (RSA) scores for comparisons of networks and visual cortex [16], adapting the procedure described in [5]. We split the 872 "shared" images into a training set and a test set of 436 images each. In each set, representational dissimilarity matrices (RDMs) were created by calculating Pearson correlation distances for pairwise comparisons of image representations within each network layer and each fMRI subject. For each network layer, we computed RDMs using the same globally pooled channel activations that were used to compute universality and brain similarity scores. RSA scores were obtained by calculating the Spearman correlation between the RDMs for a network layer and an fMRI subject, and these scores were averaged across subjects. For each network, the best-performing layer was selected based on the RSA scores in the training set, and a final RSA score was computed for each network using the selected layer in the held-out test set. We next examined the contribution of universal dimensions to the RSA scores by reducing each network to the subspace spanned by its top ten or five universal dimensions. Specifically, we reconstructed the test-set activations of each network using only the top ten or five most universal dimensions, and we re-computed the final RSA score on these reconstructed test data.

5 Data availability

The Natural Scenes Dataset is available at https://naturalscenesdataset.org/ [1].

6 Code availability

Code for all analyses in this study is available at https://github.com/zche377/universal_dimensions.

7 Acknowledgements

This research was supported in part by a JHU Catalyst Award to MFB and grant NSF PHY-2309135 to the Kavli Institute for Theoretical Physics (KITP).

References

[1] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. "A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence." In: *Nature neuroscience* 25.1 (2022), pp. 116–126.

[2] Michael F Bonner and Russell A Epstein. "Object representations in the human brain reflect the co-occurrence statistics of vision and language." In: *Nature communications* 12.1 (2021), p. 4081.

[3] Rosa Cao and Daniel Yamins. "Explanatory models in neuroscience, Part 2: Functional intelligibility and the contravariance principle." In: *Cognitive Systems Research* 85 (2024), p. 101200.

[4] Thomas A Carlson, Ryan A Simmons, Nikolaus Kriegeskorte, and L Robert Slevc. "The emergence of semantic meaning in the ventral temporal pathway." In: *Journal of cognitive neuroscience* 26.1 (2014), pp. 120–131.

[5] 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?" In: *bioRxiv* (2023). eprint: https://www.biorxiv.org/content/early/2023/07/01/2022.03.28.485868.full.pdf. URL: https://www.biorxiv.org/content/early/2023/07/01/2022.03.28.485868.

[6] Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. *Visual representations in the human brain are aligned with large language models*. 2024. arXiv: 2209.11737 [cs.CV]. URL: https://arxiv.org/abs/2209.11737.

[7] Adrien Doerig, Rowan P Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W Lindsay, Konrad P Kording, Talia Konkle, Marcel AJ Van Gerven, Nikolaus Kriegeskorte, et al. "The neuroconnectionist research programme." In: *Nature Reviews Neuroscience* 24.7 (2023), pp. 431–450.

[8] Eric Elmoznino and Michael F Bonner. "High-performing neural network models of visual cortex benefit from high latent dimensionality." In: *PLOS Computational Biology* 20.1 (2024), e1011792.

[9] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. *VISSL*. https://github.com/facebookresearch/vissl. 2021.

[10] Florentin Guth and Brice Ménard. "On the universality of neural encodings in CNNs." In: *ICLR 2024 Workshop on Representational Alignment*. 2024. URL: https://openreview.net/forum?id=ofEBFOrITI.

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

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

[13] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. "The platonic representation hypothesis." In: *arXiv preprint arXiv:2405.07987* (2024).

[14] Nancy Kanwisher, Meenakshi Khosla, and Katharina Dobs. "Using artificial neural networks to ask 'why'questions of minds and brains." In: *Trends in Neurosciences* 46.3 (2023), pp. 240–254.

[15] Nikolaus Kriegeskorte. "Deep neural networks: a new framework for modeling biological vision and brain information processing." In: *Annual review of vision science* 1 (2015), pp. 417–446.

[16] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. "Representational similarity analysis-connecting the branches of systems neuroscience." In: *Frontiers in systems neuroscience* (2008), p. 4.

[17] 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." In: *Neuron* 60.6 (2008), pp. 1126–1141.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[19] Ya Le and Xuan Yang. "Tiny imagenet visual recognition challenge." In: *CS 231N* 7.7 (2015), p. 3.

[20] 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: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755.

[21] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. "UMAP: Uniform Manifold Approximation and Projection." In: *The Journal of Open Source Software* 3.29 (2018), p. 861.

[22] Bruno A Olshausen and David J Field. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." In: *Nature* 381.6583 (1996), pp. 607–609.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. "Pytorch: An imperative style, high-performance deep learning library." In: *Advances in neural information processing systems* 32 (2019).

[24] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. "A deep learning framework for neuroscience." In: *Nature neuroscience* 22.11 (2019), pp. 1761–1770.

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge." In: *International journal of computer vision* 115 (2015), pp. 211–252.

[26] Konstantin Schürholt, Diyar Taskiran, Boris Knyazev, Xavier Giró-i-Nieto, and Damian Borth. "Model zoos: A dataset of diverse populations of neural network models." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 38134–38148.

[27] Eero P Simoncelli and Bruno A Olshausen. "Natural image statistics and neural representation." In: *Annual review of neuroscience* 24.1 (2001), pp. 1193–1216.

[28] Ross Wightman. *PyTorch Image Models*. https://github.com/rwightman/pytorch-image-models. 2019. DOI: 10.5281/zenodo.4414861.

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

[30] 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." In: *Proceedings of the national academy of sciences* 111.23 (2014), pp. 8619–8624.

[31] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. "Understanding neural networks through deep visualization." In: *arXiv preprint arXiv:1506.06579* (2015).

8 Supplementary material

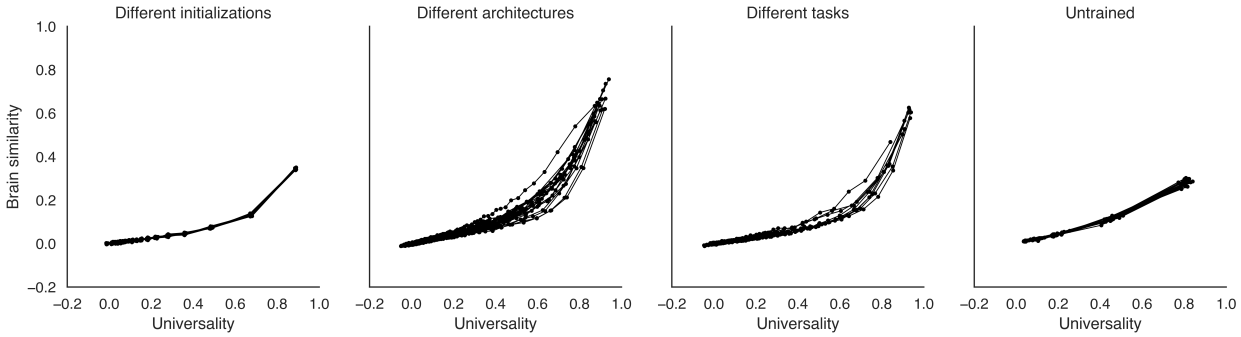

Figure S1: **Universality and brain similarity for individual networks.** These plots show the universality and brain similarity scores for individual networks. Four sets of deep neural networks were examined, including three sets of trained networks with varied initializations, architectures, and tasks and one set of untrained networks. The analyses are the same as in Figure 3, but here the results are plotted as the average values for individual networks. Average universality and brain similarity scores were computed for equally sized quantiles of 100 dimensions along the x-axis for each network. As in Figure 3, these plots show that a highly consistent trend is observed across all networks and that universal dimensions are not restricted to a subset of networks.

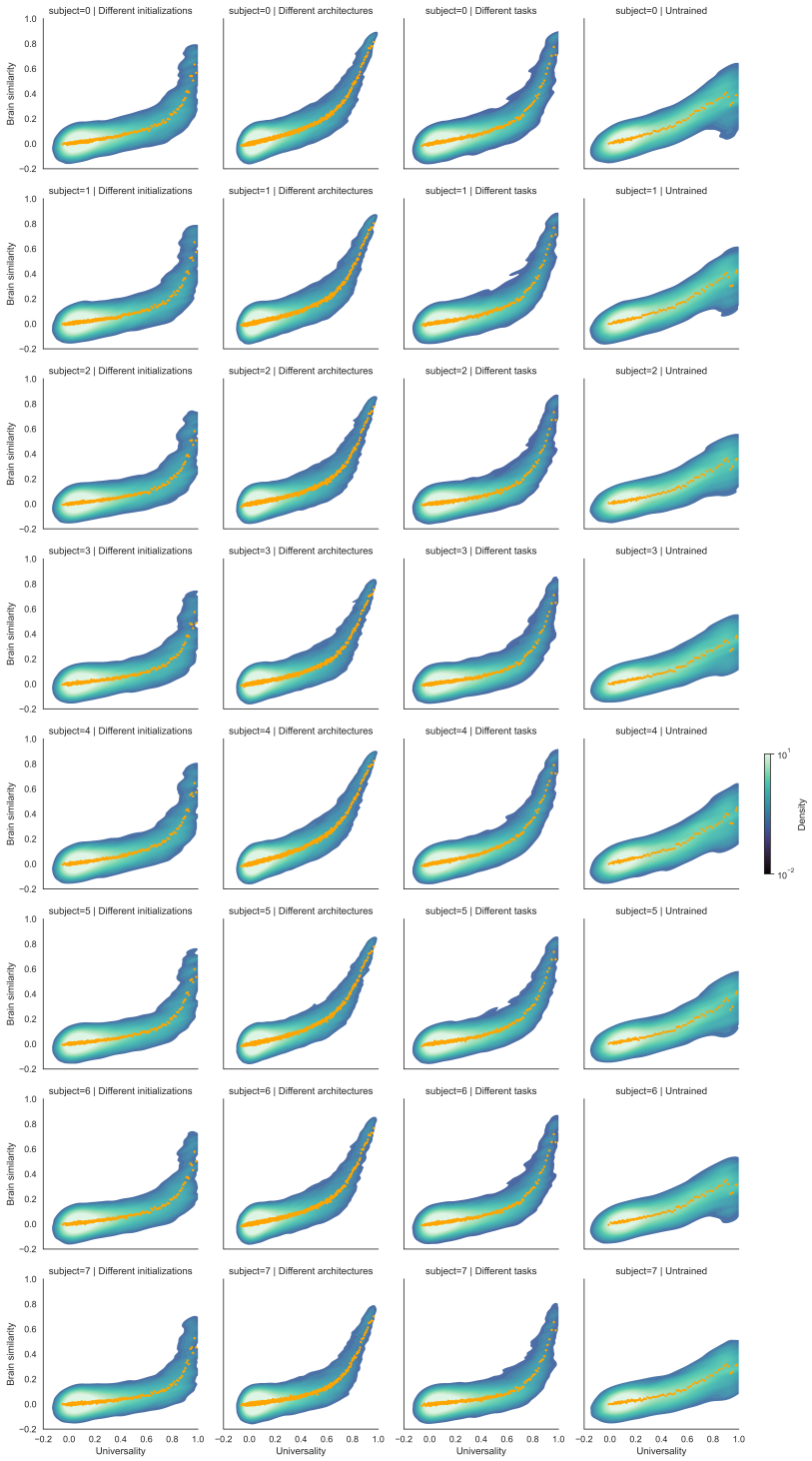

Figure S2: **Universality and brain similarity of network dimensions in individual subjects.** These plots show the relationship between universality and brain similarity in each subject from the fMRI dataset. The analyses are the same in Figure 2 but without averaging the brain similarity scores across subjects. As in Figure 2, these plots show the density of dimensions on a logarithmic scale computed using kernel density estimation. The orange dots show the mean universality and brain similarity scores for equally sized quantiles of 100 dimensions along the x-axis. These results demonstrate that the relationship between universality and brain similarity is highly consistent and robustly detected in all individual subjects.

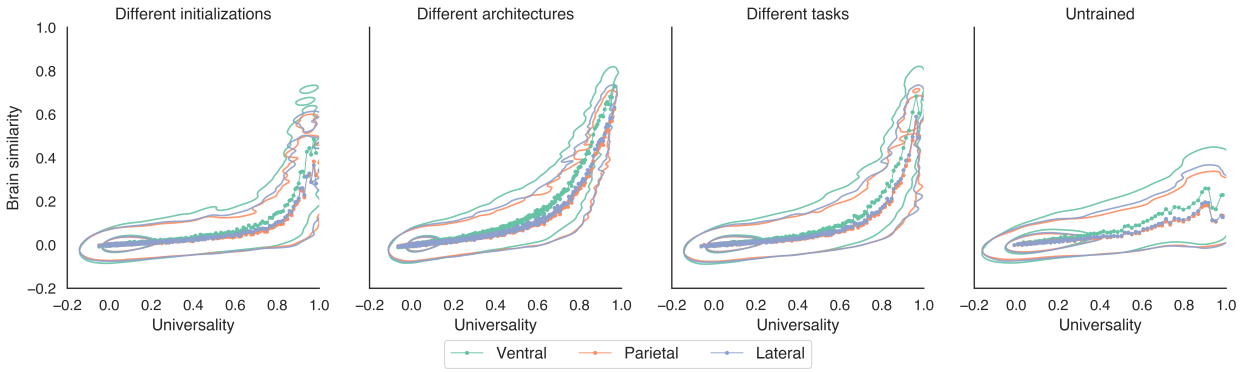

Figure S3: **Universality and brain similarity for multiple regions of interest in visual cortex.** These plots show the universality and brain similarity scores for three regions of interest: the ventral, parietal, and lateral streams. These regions are based on the "streams" masks as defined by the authors of the Natural Scenes Dataset study [1]. The analysis is the same as in Figure 2, but here brain similarity scores are computed using the three stream regions rather than the large `nsdgeneral` region from our main analyses. To visualize the results for all three regions in a single plot, the contours of the kernel density estimate plots are displayed here, rather than their density values. Average universality and brain similarity scores were computed for equally sized quantiles of 100 dimensions along the x-axis for each region. These plots show that a highly consistent trend is observed across all regions.

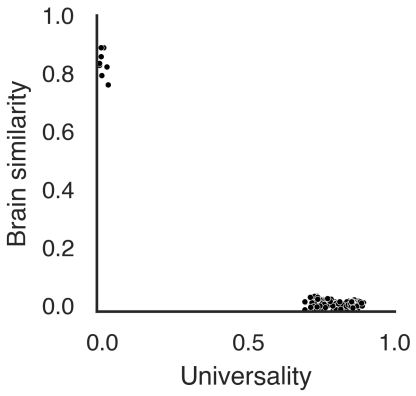

Figure S4: **Universality and brain similarity are not intrinsically correlated.** Universality and brain similarity scores were computed on simulated data to demonstrate that these metrics are not intrinsically correlated and can be trivially dissociated from one another. Data were generated for 20 simulated "subjects" and "networks." All subjects and a single network were created from a common matrix of orthonormal random variables with added Gaussian noise. The remaining networks were created from another matrix of orthonormal random variables with added Gaussian noise. These simulated data yield a subset of dimensions with low universality and high brain similarity and another set of dimensions with high universality and low brain similarity. The plot limits for both axes are set to [-0.02, 1] to make the data points near 0 visible.

Figure S5: **Two-dimensional visualization of model-specific representations.** Image activations for the 100 *least* universal dimensions from a high-level network layer were embedded in two dimensions using uniform manifold approximation and projection. Specifically, image activations were obtained for the 100 dimensions with the lowest universality scores in the penultimate layer from the set of ResNet-50 models trained on different tasks. In contrast to the universal dimensions visualized in Figure 4, this plot shows no clear semantic organization.

Figure S6: **Two-dimensional visualization of untrained models.** Image activations for the 100 most universal dimensions of untrained networks were embedded in two dimensions using uniform manifold approximation and projection. Specifically, image activations were obtained for the top 100 dimensions with the highest universality scores in the penultimate layer from the set of untrained ResNet-18 models with different random weights. In contrast to the universal dimensions visualized in Figure 4, this plot shows no clear semantic organization. Instead, the universal dimensions of untrained networks appear to emphasize low-level image properties, as demonstrated by the strong luminance gradient from left to right in this plot.

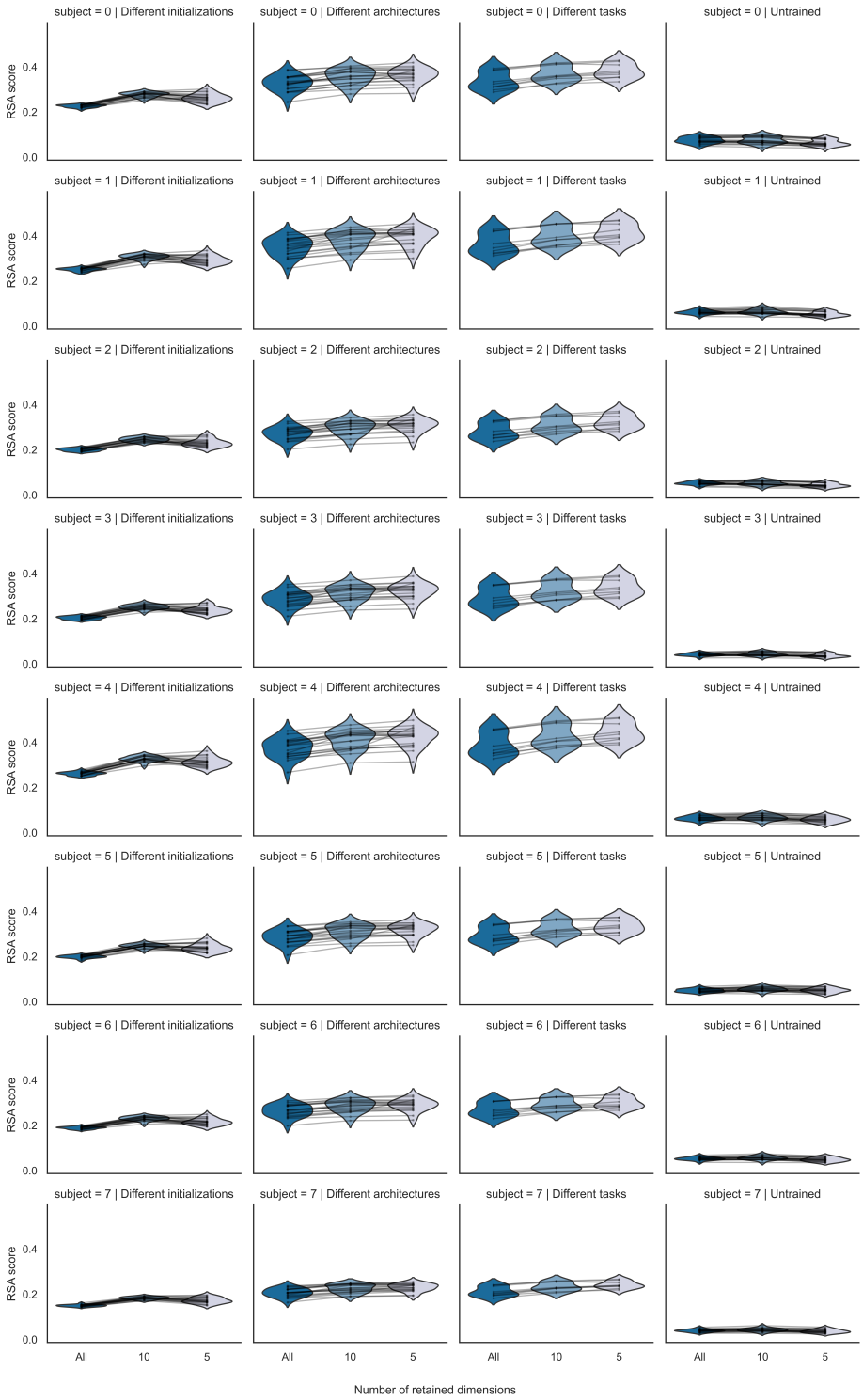

Figure S7: **Universal dimensions underlie the results of representational similarity analyses for individual subjects.** These plots show the results of representational similarity analyses (RSA) comparing networks with each subject from the fMRI dataset. The analyses are the same as in Figure 5 but without averaging the RSA scores across subjects. As in Figure 5, each dot is a network, whose representations were either intact or reduced to subspaces of their top ten or five universal dimensions, and the violin plots show distributions of RSA scores across networks. These results demonstrate that the subspaces of universal dimensions within each network consistently drive the representational similarity between neural networks and visual cortex across all individual subjects.

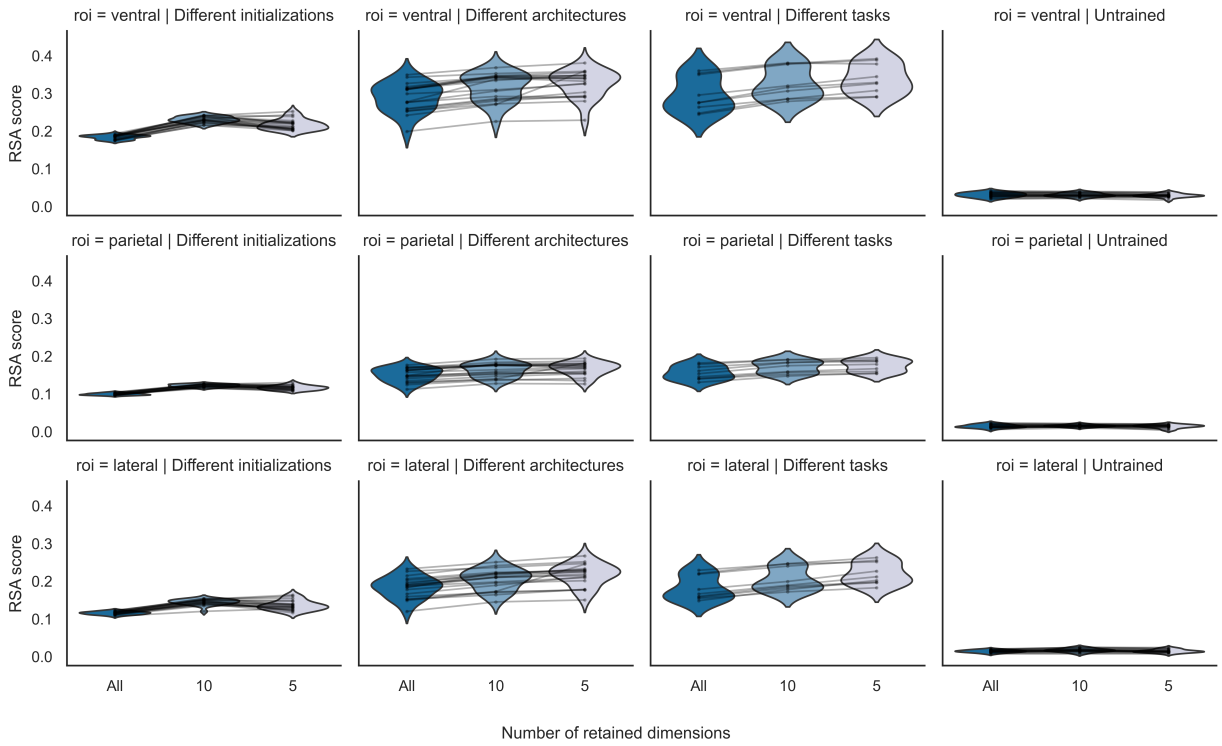

Figure S8: **Universal dimensions underlie the results of representational similarity analyses in multiple regions of interest.** These plots show the results of representational similarity analyses (RSA) comparing networks with the fMRI responses from three regions of interest: the ventral, parietal , and lateral streams. These regions are based on the "streams" masks as defined by the authors of the Natural Scenes Dataset study [1]. The analyses are the same as in Figure 5, but here the representational dissimilarity matrices (RDM) for the fMRI data are computed using the three stream regions rather than the large nsdgeneral region from our main analyses. As in Figure 5, each dot is a network, whose representations were either intact or reduced to subspaces of their top ten or five universal dimensions, and the violin plots show distributions of RSA scores across networks. These results demonstrate that the subspaces of universal dimensions within each network consistently drive the representational similarity between neural networks and visual cortex across all regions.

Architecture	Learning objective	Architecture type	Training data	Source
ResNet18	Object classification	Convolutional	ImageNet	PyTorch
ResNet50		Convolutional		
ResNeXT50_32x4d		Convolutional		
Wide_ResNet50_2		Convolutional		
AlexNet		Convolutional		
VGG16		Convolutional		
DenseNet121		Convolutional		
SqueezeNet1_1		Convolutional		
ShuffleNet_v2_x1_0		Convolutional		
ConveNeXt_tiny		Convolutional		
Swin_t		Transformer		
MaxVit_t		Transformer		
Cait_xxs24_224		Transformer		Timm
Coat_lite_tiny		Transformer		
Deit_tiny_patch16_224		Transformer		
Levit_128		Transformer		
Mixer_b16_224		MLP-Mixer		
ResMLP_12_224		MLP-Mixer		
Dla34		Convolutional		

Table S1: Networks from the set of trained models with varied architectures.

Architecture	Learning objective	Training setting	Training data	Source
ResNet50	Object classification	Supervised	ImageNet	PyTorch
	Jigsaw	Self-supervised		VISSL
	RotNet			
	ClusterFit			
	NPID++			
	PIRL			
	SimCLR			
	SwAV			
	DeepClusterV2			

Table S2: Networks from the set of trained models with varied training objectives.