Factorized visual representations in the primate visual system and deep neural networks

Jack W. Lindsey[1,2,] and Elias B. Issa[1,2,*]

1. Zuckerman Mind Brain Behavior Institute, Columbia University
2. Department of Neuroscience, Columbia University
*. Correspondence: elias.issa@columbia.edu

ABSTRACT

Object classification has been proposed as a principal objective of the primate ventral visual stream. However, optimizing for object classification alone does not constrain how other variables may be encoded in high-level visual representations. Here, we studied how the latent sources of variation in a visual scene are encoded within high-dimensional population codes in primate visual cortex and in deep neural networks (DNNs). In particular, we focused on the degree to which different sources of variation are represented in non-overlapping ("factorized") subspaces of population activity. In the monkey ventral visual hierarchy, we found that factorization of object pose and background information from object identity increased in higher-level regions. To test the importance of factorization in computational models of the brain, we then conducted a detailed large-scale analysis of factorization of individual scene parameters – lighting, background, camera viewpoint, and object pose – in a diverse library of DNN models of the visual system. Models which best matched neural, fMRI and behavioral data from both monkeys and humans across 12 datasets tended to be those which factorized scene parameters most strongly. In contrast, invariance to object pose and camera viewpoint in models was negatively associated with a match to neural and behavioral data. Intriguingly, we found that factorization was similar in magnitude and complementary to classification performance as an indicator of the most brainlike models suggesting a new principle. Thus, we propose that factorization of visual scene information is a widely used strategy in brains and DNN models.

INTRODUCTION

Artificial deep neural networks (DNNs) are the most predictive models of neural responses to images in the primate high-level visual cortex[1,2]. Many studies have reported that DNNs trained to perform image classification produce internal feature representations broadly similar to those in areas V4 and IT of the primate cortex, and that this similarity tends to be greater in models with better classification performance[3]. However, beyond a certain threshold level of object classification performance, further improvement fails to produce a concomitant improvement in predicting primate neural responses[2,4]. This weakening trend motivates finding new normative principles, besides object classification ability, that push models to better match primate visual representations. In particular, high-level cortical areas in the ventral visual cortex are known to encode other properties of visual input besides object identity[5–8]. Here, we introduce a framework for understanding the relationships between different types of visual information in a population code (e.g., object identity, pose, and camera viewpoint).

Intuitively, if the variance driven by one parameter is encoded independently from the variance driven by other scene parameters, we say this code is factorized. Factorization is a useful strategy for representing information about multiple scene parameters, to support different visually guided behaviors, in an easily decodable fashion[9]. Here, we found evidence for factorized representations in high-level primate visual areas. Testing across a broad library of DNN models that varied in their architecture and training objectives, we found that factorization of scene parameters in DNN feature representations was associated with models' matches to neural and behavioral data. Whereas simple invariance to some scene parameters (background scene and lighting conditions) predicted neural fits, invariance to others (object pose and camera viewpoint) did not. Our results generalized across both monkey and human datasets using different measures (neural spiking, fMRI, and behavior; 12 datasets total) and could not be accounted for by models' classification performance. Thus, we suggest that factorized encoding of multiple behaviorally-relevant scene variables is an important consideration in building more brainlike models of visual scene representations.

RESULTS

Decoding object identity from population responses can be enhanced by invariance of responses to non-identity scene parameters or by factorizing non-identity-driven response variance into isolated (factorized) subspaces (**Figure 1A**). To formalize these notions, we introduced measures of factorization and invariance to scene parameters in neural population responses (see Equations 1 and 2 in **Methods**). Factorization, unlike invariance, has the potential to enable the simultaneous representation of multiple scene parameters in a decodable fashion. To clarify the benefits of factorization for decoding, we simulated a simple task requiring simultaneous decoding of multiple variables. The extent to which the variables of interest were represented in a factorized way (i.e., along orthogonal axes, rather than correlated axes, in population activity space) influenced the ability of a linear discriminator to successfully decode both variables in a generalizable fashion from a few training samples (**Figure 1B**).

We next asked whether previously collected neural data from the macaque high-level visual cortex (**Table S1**) exhibited factorized structure[10]. Specifically, we took advantage of an existing dataset in which the tested images independently varied object identity versus object pose plus background context. We found that both V4 and IT responses exhibited more significant factorization of object identity information from non-identity information than a shuffle control (**Figure S1)**, and that the degree of factorization increased from V4 to IT (**Figure 1C**). Consistent with prior studies, we also found that invariance to non-identity information was increased from V4 to IT in our analysis (**Figure 1D**, black lines)[11]. Invariance to non-identity information was even more pronounced when measured in the subspace of population activity containing the bulk (90%) of identity-driven variance, as a consequence of increased factorization of identity from non-identity factors (**Figure 1D**, orange lines). To illustrate the particular benefit of factorization for decoding performance, we analyzed a transformed neural representation obtained by rotating the population data so that inter-class variance more strongly overlapped with the principal components of the within-class variance in the data (see **Methods**). This transformation, designed to decrease factorization while leaving invariance to non-class variables and other activity statistics intact (such as mean neural firing rates and covariance structure of the population) had the effect of significantly reducing object identity decoding performance in both V4 and IT (**Figure 1E)**.

We next sought to explore the high-level representation of other forms of visual information besides object identity. Existing experiments, however, have not recorded neural responses to image datasets that independently vary other scene parameters besides object identity at sufficient scale to enable an analysis like the above. Hence, we turned to a modeling-based approach and studied the degree of factorization of and invariance to specific scene parameters in representations learned by

DNNs (**Figure 2A**). We generated an augmented image set, based on the images used in the previous dataset (**Figure 1C**), in which we independently varied the foreground object identity, foreground object pose, background identity, scene lighting, and 2D scene viewpoint. Specifically for each base image from the original dataset, we generated sets of images that varied exactly one of the above scene parameters while keeping the others constant (**Figure 2B**; 100 base scenes and 10 transformed images for each source of variation). By presenting this image dataset to models (4000 images total), we could compute the relative degree of representational factorization and invariance for each scene parameter. We conducted this analysis across a broad range of DNNs varying in architecture and objective as well as other implementational choices. These included models using supervised training for object classification[12,13], contrastive self-supervised training[14,15], and self-supervised models trained using auxiliary objective functions[16–19] (see **Methods** and **Table S2**).

First, we observed that the final representational layers of trained networks exhibited consistent increases in factorization of all tested scene parameters relative to a randomly initialized (untrained) baseline with the same architecture (**Figure 2C**, top row, rightward shift relative to black cross, a randomly initialized ResNet-50). Moreover, we found that models' factorization scores correlated with the degree to which they predicted neural responses to natural images for single-unit IT data (**Figure 2C**, top row). Interestingly, we saw a different pattern for invariance to a scene parameter. First, training produced mixed effects on invariance, typically increasing it for background and lighting but reducing it for object pose and camera viewpoint (**Figure 2C**, bottom row, leftward shift relative to black cross for left two panels). Second, invariance across models showed mixed correlations with neural predictivity (**Figure 2C**, bottom row).

Similar patterns were observed across a large number of previously collected neural and behavioral datasets from different primate species and visual regions (6 macaque datasets[10,20,21]: two V4, two IT, and two behavior; 6 human datasets[21–23]: two V4, two HVC, and two behavior; **Table S1**). Consistently, increased factorization of scene parameters in late model layers correlated with models being more predictive of neural spiking, voxel BOLD signal, and behavioral responses to images (**Figure 3A**, black bars; see **Figure S2** for scatter plots across all datasets). Although invariance to appearance factors (background identity and scene lighting) correlated with more brainlike models, invariance for spatial transforms (object pose and camera viewpoint) consistently did not (zero or negative correlation values; **Figure 3A**, gray bars). **Figure 3C** summarizes these results across datasets. Our results were preserved when we re-ran the analyses using only the subset of models with the identical architecture (ResNet-50) (**Figure S3**) or when we evaluated model predictivity using representational dissimilarity matrices of the population (RDM) instead of linear regression fits of individual neurons or fMRI voxels (**Figure S4**). Furthermore, the main finding of a positive correlation between factorization and neural predictivity was robust to the particular choice of PCA threshold we used to quantify factorization (**Figure S5**).

Next, we tested whether our results generalized across image sets used for computing the model factorization scores. Here, instead of relying on our synthetically generated images, we re-computed factorization from two datasets of natural movies, one in which the observer moves in an urban environment (approximates camera viewpoint changes)[24] and another in which objects move in front of a fairly stationary observer (approximates object pose changes)[25]. Similar to the previous results using augmentations of naturalistic static images, factorization of frame-by-frame variance (local in time) from other sources of variance (non-local in time) across natural movies was correlated with improved neural predictivity in both macaque and human data, while invariance to these parameters was not (**Figure 3B**; black versus gray bars).

It is possible that the observed correlations between scene parameter factorization and neural fit could be entirely captured by the known correlation between classification performance and neural fits[2,3]. However, we found that factorization significantly boosted cross-validated predictive power

over simply using classification alone (**Figure 3D**), rectifying the saturating correlation between classification performance and neural fits (**Figure 3E**).

DISCUSSION

Object classification, which has been proposed as a normative principle for the function of the ventral visual stream, can be supported by qualitatively different representational geometries[3,26]. These include representations that are completely invariant to non-class information[27,28] and representations that retain a high-dimensional but factorized encoding of non-class information. Here, we presented evidence that factorization of non-class information is an important strategy used, along with invariance, by the high-level visual cortex and by DNNs that are predictive of primate neural and behavioral data. Concurrent work has shown that DNN models with high-dimensional embeddings of natural images yielded better fits to neural data[29]. Our work complements this finding and provides a potential interpretation, namely that high-dimensional representations are employed by visual areas in order to maintain orthogonal encodings for different sources of scene variation. We note that the degree of factorization measured in neural data is significantly greater than that of a shuffle control with the same dimensionality, indicating that the factorized encoding found in cortical responses goes beyond what would be inherited from a random high-dimensional representation (**Figure S1**).

Going forward, we expect factorization could prove to be a useful objective function for optimizing neural network models that better resemble primate visual systems. Our results complement prior theoretical studies that show benefits of orthogonal encoding of different sources of variance for generalization performance of trained decoders[9,30]. An important limitation of our work is that we do not specify the details of how a particular scene parameter is encoded within its factorized subspace. Neural codes could adopt different strategies resulting in similar factorization scores at the population level, each with some support in visual cortex literature: (1) Each neuron encodes a single latent variable[31,32], (2) Separate brain subregions encode qualitatively different latent variables but using distributed representations within a region[33–35], (3) Each neuron encodes multiple variables in a distributed population code, such that the factorization of different variables is only apparent when assessed in high-dimensional population activity space[31,36]. Future work can disambiguate among these possibilities by systematically examining subregions of the ventral visual stream and single-neuron tuning curves within them[37,38].

METHODS

Monkey datasets. We used three sources of data from macaque monkeys, corresponding to single-unit neural recordings[20], multi-unit neural recordings[10], and object recognition behavior[21]. Single-unit spiking responses to natural images were measured in V4 and anterior ventral IT[20]. These IT recordings were obtained from penetrating electrodes targeting the anterior ventral portion of IT near the base of the skull, reflecting the highest level of the IT hierarchy. On the other hand, the multi-unit dataset was obtained from across IT with a bias toward where multi-unit arrays were more easily placed such as CIT and PIT[10], complementing the recording locations of the single-unit dataset. An advantage of the multi-unit dataset using chronic recording arrays is that an order of magnitude more images were tested per recording site (see dataset comparisons in **Table S1**). Finally, the monkey behavioral dataset came from a third study examining the image-by-image object classification performance of macaques and humans[21].

Human datasets. Three datasets from humans were used: two fMRI datasets and one object recognition behavior dataset[4,21,22]. The fMRI datasets used different images (color versus grayscale) but otherwise used similar number of images and voxel resolution in imaging. The human behavioral

dataset measured image-by-image classification performance and was collected in the same study as the monkey behavioral signatures[21].

Computational models. A variety of approaches to training DNN vision models have been developed that learn representations that can be used for downstream classification (and other) tasks (see **Table S2** for a list of models used and corresponding references). Models differ in a variety of implementational choices including in their architecture, objective function, and training dataset. In the models we sampled, objectives included supervised learning of object classification (AlexNet, ResNet), self-supervised contrastive learning (MoCo, SimCLR), and other unsupervised learning algorithms based on auxiliary tasks (e.g., reconstruction, or colorization). A majority of the models that we considered relied on the widely used, performant ResNet-50 architecture trained on the ImageNet dataset, though some in our library utilized different architectures. The randomly initialized network control utilized ResNet-50 (see **Figure 2C,D**).

Simulation of factorized versus non-factorized representational geometries. For the simulation in **Figure 1B**, we generated data in the following way. First we randomly sampled the values of N=10 binary features. Feature values corresponded to positions in an N-dimensional vector space as follows: each feature was assigned an axis in N-dimensional space, and the value of each feature (+1 or -1) was treated as a coefficient indicating the position along that axis. All but two of the feature axes were orthogonal to the rest. The last two features, which served as targets for the trained linear decoders, were assigned axes whose alignment ranged from 0 (orthogonal) to 1 (identical). In the noiseless case, factorization (according to our definition) of these two variables with respect to one another is given by subtracting the square of the cosine of the angle between the axes from 1. We added Gaussian noise to the positions of each data point and randomly sampled K positive and negative examples for each variable of interest to use as training data for the linear classifier (a support vector machine).

Macaque neural data analyses. For the shuffle control used as a null model for factorization, we shuffled the object identity labels of the images (**Figure S1)**. For the transformation of the multi-unit neural dataset used in **Figure 1E**, we computed the principal components of the mean neural activity response to each object class ("class centers"), referred to as the inter-class PCs. We also computed the principal components of the data with corresponding class centers subtracted from each activity pattern, referred to as the intra-class PCs. We transformed the data by applying to the class centers a change of basis matrix that rotated each inter-class PC into the corresponding (according to the rank of the magnitude of its associated eigenvalue) intra-class PC. That is, the class centers were transformed by this matrix, but the relative positions of activity patterns for a given class were fixed. This transformation has the effect of preserving intra-class variance statistics exactly from the original data and of preserving everything about the statistics of inter-class variance except its orientation relative to intra-class variance. That is, the transformation is designed to affect (specifically decrease) factorization while controlling for all other statistics of the activity data that may be relevant to object classification performance.

Scene parameter variation. Our generated scenes consisted of foreground objects imposed upon natural backgrounds. To measure variance associated with a particular parameter like the background identity, we randomly sampled ten different backgrounds while holding the other variables (e.g., foreground object identity and pose constant). To measure variance associated with foreground object pose, we randomly varied object angle from [-90, 90] along all three axes

independently, object position on the two in-plane axes, horizontal [-30%, 30%] and vertical [-60%, 60%], and object size [×1/1.6, ×1.6]. To measure variance associated with camera position, we took crops of the image with scale uniformly varying from 20% to 100% of the image size, and position uniformly distributed across the image. To measure variance associated with lighting conditions we applied random jitters to the brightness, contrast, saturation, and hue of an image, with jitter value bounds of [-0.4, 0.4] for brightness, contrast, and saturation and [-0.1, 0.1] for hue. These parameter choices follow standard data augmentation practices for self-supervised neural network training, as used, for example, in the SimCLR and MoCo models tested here[14,15].

Factorization and invariance metrics. Factorization and invariance were measured according to the following equations:

$$factorization_{param} = 1 - var_{param|other\_param\_subspace} / var_{param} \quad (1)$$

$$invariance_{param} = 1 - var_{param} / var_{all\ param} \quad (2)$$

Variance induced by a parameter (var_{param}) is computed by measuring the variance (summed across all dimensions of neural activity space) of neural responses to the 10 augmented versions of the base images where the augmentations are those obtained by varying the parameter of interest. This quantity is then averaged across the 100 base images. The variance induced by all parameters is simply the sum of the variances across all images and augmentations. To define the "other-parameter subspace," we averaged neural responses for a given base image over all augmentations using the parameter of interest and ran PCA on the resulting set of averaged responses. The subspace was defined as the space spanned by top PCA components containing 90% of the variance of these responses. Intuitively, this space captures the bulk of the variance driven by all parameters other than the parameter of interest (due to the averaging step). The variance of the parameter of interest *within* this "other-parameter subspace," $var_{param|other\_param\_subspace}$, subspace, was computed the same way as var_{param}, but using the projections of neural activity responses onto the other-parameter subspace.

Natural movie factorization metrics

For natural movies, variance is not induced by explicit control of a parameter as in our synthetic scenes but implicitly, by considering contiguous frames (separated by 200ms in real time) as reflective of changes in one of two motion parameters (object versus observer motion) depending on how stationary the observer is (MIT Moments in Time movie set: stationary observer; UT-Austin Egocentric movie set: nonstationary observer)[24,25]. Here, the *all parameters* condition is simply the variance across all movie frames. In the case of MIT Moments in Time dataset, this includes variance across thousands of video clips taken in many different settings. In the case of the UT-Austin Egocentric movie dataset, this includes variance across only 4 movies but over long durations of time during which an observer translates extensively in an environment (3-5 hours). Thus, movie clips in the MIT Moments in Time movie set contained new scenes with different object identities, backgrounds, and lightings, capturing variance induced by these non-spatial parameters[25]. In the UT Austin Egocentric movie set, new objects are encountered as the subject navigates around the urban landscape[24].

Model neural encoding fits. Linear mappings between model features and neuron (or voxel) responses were computed using ridge regression (with regularization coefficient selected by cross validation) on a low-dimensional linear projection of model features (top 300 PCA components computed using images in each dataset). We also tested an alternative approach to measuring representational similarity between models and experimental data based on representational

similarity analysis (RSA)[39], computing dot product similarities of the representations of all pairs of images and measuring the Spearman's rank correlation coefficient between these pairwise similarity matrices obtained from a given model and neural dataset (**Figure S4**).

Model behavioral signatures. We followed the approach of Rajalingham, Issa et al.[21] We took human and macaque behavioral data from the object classification task and used it to create signatures of image-level difficulty (the "I1" vector) and image-by-distractor-object confusion rates (the "I2" matrix). We did the same for the DNN models, extracting model "behavior" by training logistic regression classifiers to classify object identity in the same image dataset from the experiments of Rajalingham, Issa et al.[21], using model layer activations as inputs. Model behavioral accuracy rates on image by distractor object pairs were assessed using the classification probabilities output by the logistic regression model, and these were used to compute I1 and I2 metrics as was done for the true behavioral data. Behavioral similarity between models and data was assessed by measuring the correlation between the entries of the I1 vector and I2 matrix (both I1 and I2 results are reported).

Model layer choices. The scatter plots in **Figure 2C,D** and **Figure S2** use metrics (factorization, invariance, and goodness of neural fit) taken from the final representational layer of the network (the layer prior to the logits layer used for classification in supervised network, prior to the embedding head in contrastive learning models, or prior to any auxiliary task-specific layers in unsupervised models trained using auxiliary tasks). However, representational geometries of model activations, and their match to neural activity and behavior, vary across layers. This variability arises because different model layers correspond to different stages of processing in the model (convolutional layers in some cases, and pooling operations in others), and may even have different dimensionalities. To ensure that our results do not depend on idiosyncrasies of representations in one particular model layer and the particular network operations that precede it, summary correlation statistics in all other figures (**Figure 3** and **Figures S3-S5**) show the results of the analysis in question averaged over the five final representational layers of the model. That is, the metrics of interest (factorization, invariance, neural fits, behavioral similarity scores) were computed independently for each of the five final representational layers of each model, and these five values were averaged prior to computing correlations between different metrics.

Correlation of model predictions and experimental data. A Spearman's rank correlation coefficient was calculated for each model layer x biological dataset combination (6 monkey datasets and 6 human datasets). Here, we do not correct for noise in the biological data when computing the correlation coefficient, as this would require trial repeats (for computing intertrial variability) that were limited or not available in the fMRI data used. In any event, normalizing by the data noise ceiling applies a uniform scaling to all model prediction scores and does not affect model comparison, which only depends on ranking models as being relatively better or worse in predicting brain data. Finally, we estimated the effectiveness of model factorization or invariance in combination with model object classification performance for predicting model neural and behavioral fit by performing a linear regression on the particular dual metric combination (e.g., classification plus object pose factorization) and reporting the Spearman correlation coefficient of the linearly weighted metric combination (**Figure 3D**). The correlation was assessed on held-out models (80% used for training, 20% for testing) and the results were averaged over 100 randomly sampled train/test splits.

ACKNOWLEDGEMENTS

This work was performed on the Columbia Zuckerman Institute Axon GPU cluster and via generous access to Cloud TPUs from Google's TPU Research Cloud (TRC). JWL was supported by the DOE CSGF (DE–SC0020347). EBI was supported by a Klingenstein-Simons fellowship, Sloan Foundation fellowship, and Grossman-Kavli Scholar Award. We thank Erica Shook for comments on a previous version of the manuscript. The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

JWL and EBI designed the research. JWL performed the computational modeling and data analysis. JWL and EBI wrote the manuscript. EBI supervised the research.

FIGURES

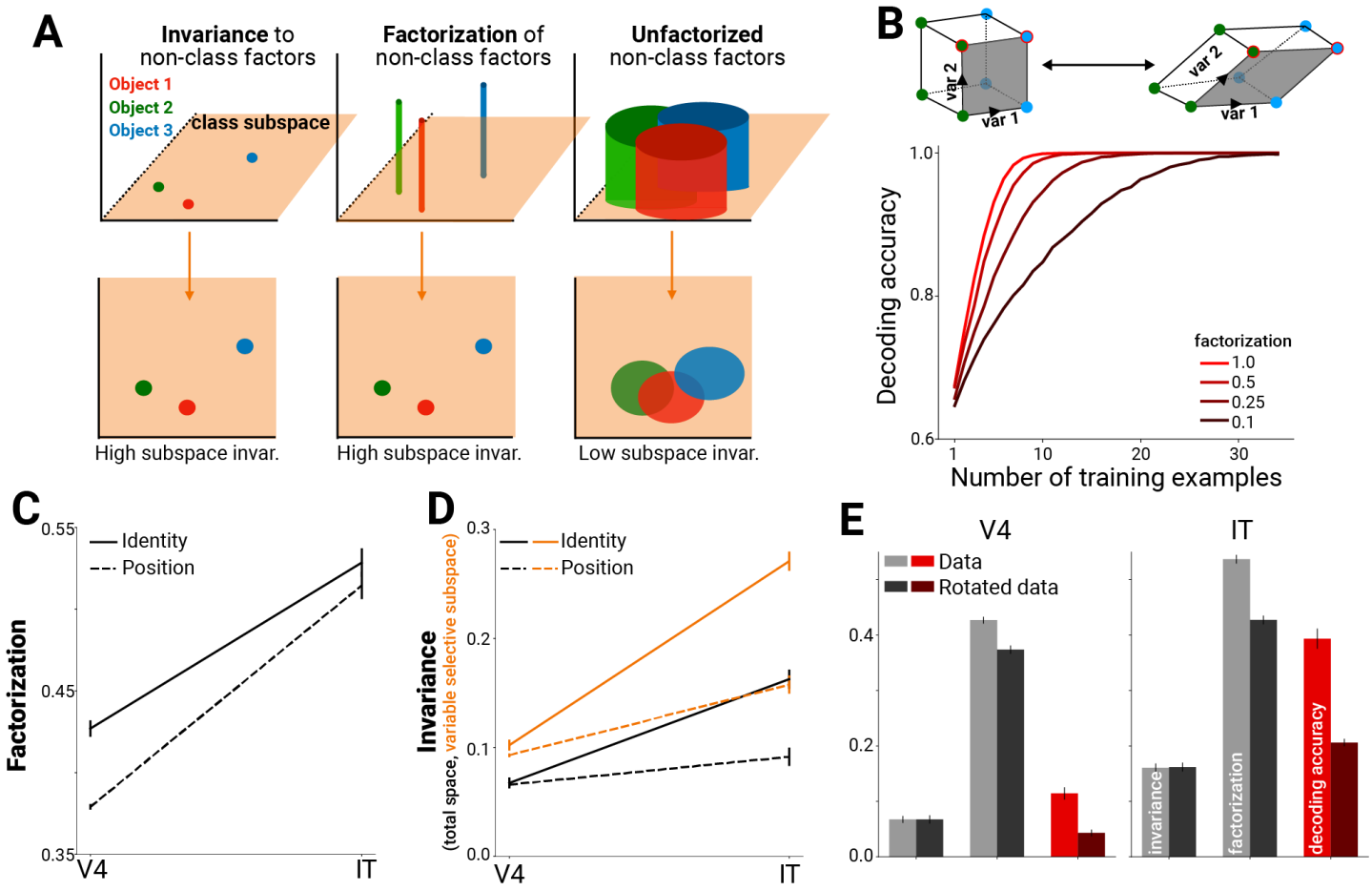


Figure 1. Framework and measurement of factorization in macaque V4 and IT. (A) Schematic illustrating three possible representations of multiple object classes in population activity space. Axes represent neural firing rates, or linear combinations thereof. Shapes (sphere or cylinder) depict the distribution of responses to images from each object class. Bottom row depicts the subspace of activity space that captures between-class variance. Representations which factorize within-class variance from between-class variance (middle column) are invariant to non-class factors when projected into this subspace. (B) In a simulated classification task requiring decoding of two binary variables (see **Methods**), a decrease in factorization – orthogonality of the relationship between the encoding of the two variables (square vs. parallelogram) – resulted in worse classifier performance, particularly in the low training sample regime (i.e., consider the case of training on the two data points encircled in red and attempting to generalize to the non-encircled dots in the bottom edge of the parallelogram). (C) Factorization of object identity from other sources of image variability and position from other sources of image variability increased from macaque V4 to IT (multiunit activity in macaque visual cortex from dataset E1). (D) Like factorization, invariance to non-identity and non-position factors also increased from V4 to IT (black lines). Within the subspace capturing the variance due to the variable of interest, effective invariance to other factors was even higher and exhibited a greater increase from V4 to IT than invariance over the whole population activity space (orange lines, higher slope than corresponding black lines). (E) Applying a transformation to the data that rotated the relative positions of mean responses to object classes, designed to preserve relevant activity statistics (including invariance to non-class factors; see **Methods**) while decreasing factorization of class information from non-class factors, has the effect of reducing object class decoding performance (light vs. dark red bars; chance performance = 1/64, using n=128 multi-unit sites in V4 and 128 in IT for decoding) .

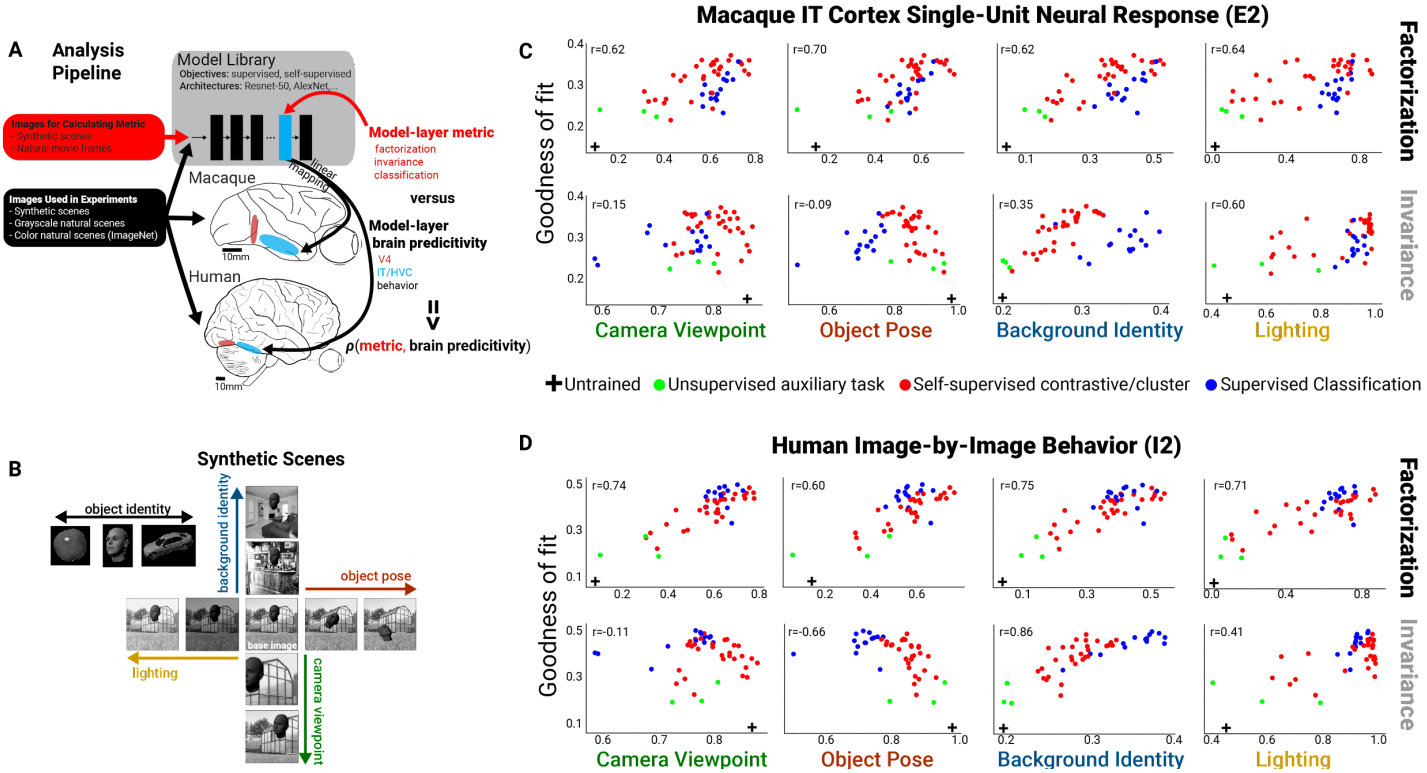


Figure 2. Measurement of factorization in DNN models and relationship to neural predictivity. (A) Schematic showing how our analysis on models and neural/behavioral data was conducted. First, we computed various representational metrics on model layers and measured a model layer's ability to predict neural and behavioral data across a variety of datasets. The combination of model-layer metric and model-layer dataset predictivity for a choice of model, layer, metric, and dataset specifies the coordinates of a single dot on the scatter plots in (C). (B) To compute factorization of and invariance to a scene parameter, we measured variance in model responses to sets of images obtained by individually varying each of four scene parameters (n=10 parameter levels) for each base image (n=100 base images, which contained varied objects and backgrounds). (C) Scatter plots for an example neural dataset (IT single-units, macaque E2 dataset) showing the correlation between a model's ability to predict IT single-unit neural data versus a model's ability to factorize (top row) or become invariant to (bottom row) different scene parameters (each dot is a different model, using each model's penultimate layer). Note that factorization in trained models is consistently higher than that for an untrained, randomly initialized Resnet-50 DNN architecture (top row, rightward shift relative to black cross). Invariance to background and lighting but not to object pose and viewpoint increased in trained models relative to the untrained control (bottom row, rightward versus leftward shift relative to black cross). Dot color indicates different classes of model training objective. (D) Same as (C) except with the y-axis replaced by the model layer's ability to predict human behavioral performance patterns on an image classification task (human I2 dataset).

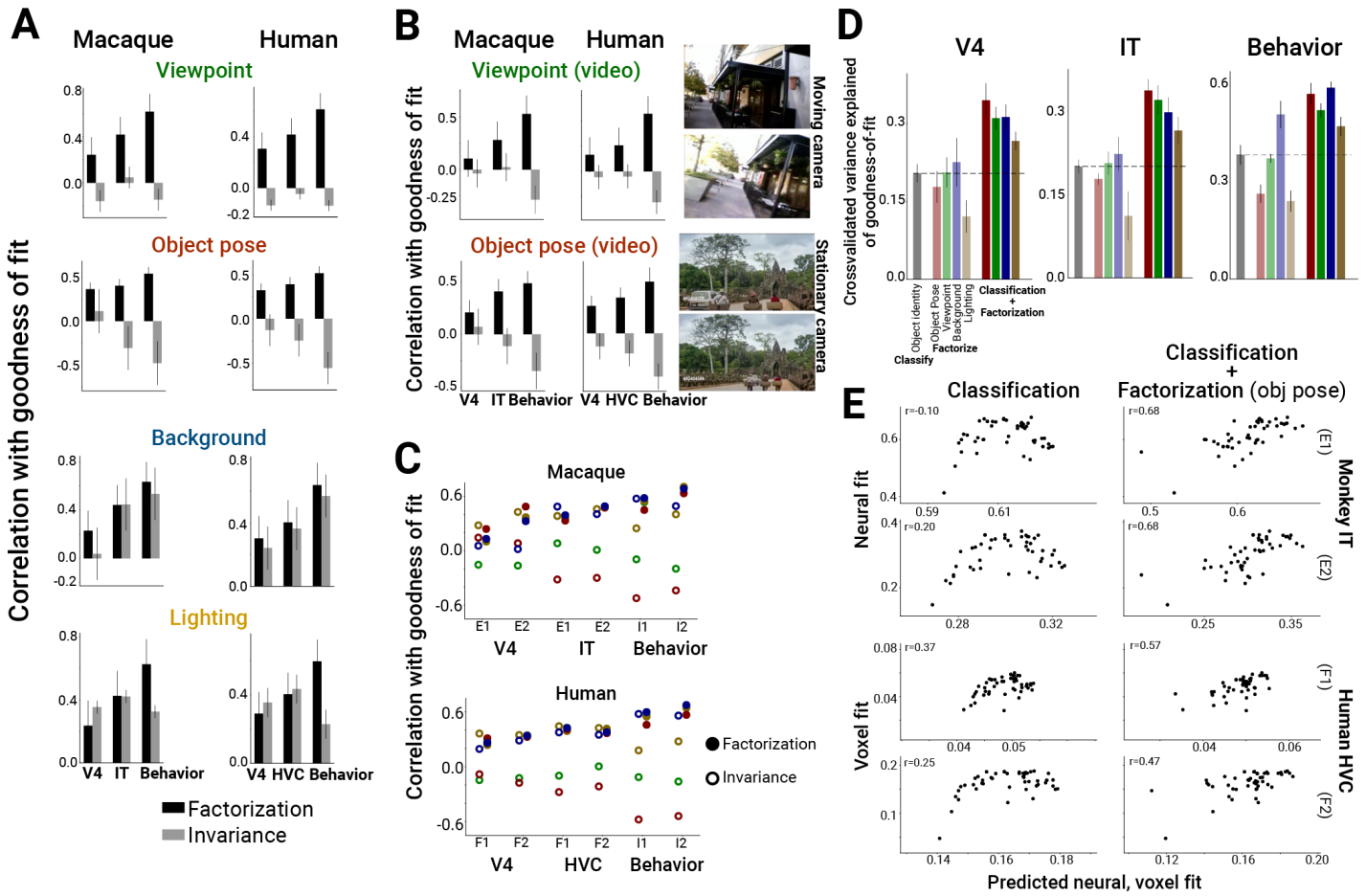


Figure 3. Scene parameter factorization correlates with more brainlike DNN models. (A) Factorization of scene parameters in model representations consistently correlated with a model better matching neural and behavioral data across multiple independent datasets (monkey neural data, human fMRI data, or behavioral performance in both macaques and humans) (black bars). Increased invariance to background and lighting (gray bars, bottom two rows), but not camera viewpoint or object pose (gray bars, top two rows), was also indicative of brainlike models. In all cases, model representational metric and neural predictivity score were computed by averaging scores across the last 5 model layers (see **Methods**). Error bars indicate confidence intervals (one standard deviation) obtained by bootstrapping over the choice of network models. (B) Recomputing camera viewpoint or object pose factorization from natural movie datasets that primarily contained camera or object motion, respectively (right: example movie frames; see **Methods**), gave similar results as in (A). (C) Summary of the results from (A) across primate datasets (x-axis) for invariance (open symbols) versus factorization (closed symbols) of different scene variables (colors, same convention as (A)). (D) Average (across datasets) degree to which classification (faded black bar) and factorization (faded colored bars) predicted neural and behavioral matches. Adding factorization to classification in a regression model produced significant improvements in predicting the most brainlike models (solid colored bars exceed dashed line for classification alone as a metric). All values indicate cross-validated variance explained on held-out models by a regression model trained (on a subset of models) to predict neural and behavioral matches based on the indicated quantity (or quantities). (E) Example scatter plots for neural and fMRI datasets (macaque IT multi-unit & single-unit responses corresponding to datasets E1 & E2, respectively; human fMRI responses to grayscale & color images corresponding to datasets F1 & F2, respectively) showing a reversing trend in neural (voxel) predictivity for models that are increasingly good at classification (left column). This saturating/reversing trend is no longer present when adding object pose factorization to classification of object identity as an additional regressor (right column).

REFERENCES

1. Cadieu, C. F. *et al.* Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Comput. Biol.* **10**, e1003963 (2014).
2. Schrimpf, M. *et al.* Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv* 407007 (2020) doi:10.1101/407007.
3. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* 201403112 (2014) doi:10.1073/pnas.1403112111.
4. Nonaka, S., Majima, K., Aoki, S. C. & Kamitani, Y. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *iScience* **24**, 103013 (2021).
5. Freiwald, W. A. & Tsao, D. Y. Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science* **330**, 845–851 (2010).
6. Hong, H., Yamins, D. L. K., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
7. Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G. & Mishkin, M. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* **17**, 26–49 (2013).
8. Peters, B. & Kriegeskorte, N. Capturing the objects of vision with neural networks. *Nat. Hum. Behav.* **5**, 1127–1144 (2021).
9. Bernardi, S. *et al.* The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954-967.e21 (2020).
10. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *J. Neurosci.* **35**, 13402–13418 (2015).
11. Rust, N. C. & DiCarlo, J. J. Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *J. Neurosci.* **30**, 12978–12995 (2010).
12. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).
13. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *ArXiv151203385 Cs* (2015).
14. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. *ArXiv191105722 Cs* (2020).
15. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *ArXiv200205709 Cs Stat* (2020).
16. Tian, Y., Krishnan, D. & Isola, P. Contrastive Multiview Coding. *ArXiv190605849 Cs* (2019).
17. Doersch, C., Gupta, A. & Efros, A. A. Unsupervised Visual Representation Learning by Context Prediction. *ArXiv150505192 Cs* (2016).
18. He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. in 2961–2969 (2017).
19. Donahue, J. & Simonyan, K. Large Scale Adversarial Representation Learning. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).
20. Rust, N. C. & DiCarlo, J. J. Balanced Increases in Selectivity and Tolerance Produce Constant Sparseness along the Ventral Visual Stream. *J. Neurosci.* **32**, 10170–10182 (2012).
21. Rajalingham, R. *et al.* Large-Scale, High-Resolution Comparison of the Core Visual Object

Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *J. Neurosci.* **38**, 7255–7269 (2018).

22. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
23. Shen, G., Horikawa, T., Majima, K. & Kamitani, Y. Deep image reconstruction from human brain activity. *PLOS Comput. Biol.* **15**, e1006633 (2019).
24. Lee, Y. J., Ghosh, J. & Grauman, K. Discovering important people and objects for egocentric video summarization. in *2012 IEEE Conference on Computer Vision and Pattern Recognition* 1346–1353 (2012). doi:10.1109/CVPR.2012.6247820.
25. Monfort, M. *et al.* Moments in Time Dataset: one million videos for event understanding. Preprint at https://doi.org/10.48550/arXiv.1801.03150 (2019).
26. Nayebi, A. *et al.* Goal-Driven Recurrent Neural Network Models of the Ventral Visual Stream. *bioRxiv* 2021.02.17.431717 (2021) doi:10.1101/2021.02.17.431717.
27. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
28. Chung, S., Lee, D. D. & Sompolinsky, H. Classification and Geometry of General Perceptual Manifolds. *Phys. Rev. X* **8**, 031003 (2018).
29. Elmoznino, E. & Bonner, M. F. High-performing neural network models of visual cortex benefit from high latent dimensionality. 2022.07.13.499969 Preprint at https://doi.org/10.1101/2022.07.13.499969 (2022).
30. Sorscher, B., Ganguli, S. & Sompolinsky, H. Neural representational geometry underlies few-shot concept learning. *Proc. Natl. Acad. Sci.* **119**, e2200800119 (2022).
31. Field, D. J. What Is the Goal of Sensory Coding? *Neural Comput.* **6**, 559–601 (1994).
32. Chang, L. & Tsao, D. Y. The Code for Facial Identity in the Primate Brain. *Cell* **169**, 1013-1028.e14 (2017).
33. Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H. & Livingstone, M. S. A Cortical Region Consisting Entirely of Face-Selective Cells. *Science* **311**, 670–674 (2006).
34. Lafer-Sousa, R. & Conway, B. R. Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nat. Neurosci.* **16**, 1870–1878 (2013).
35. Vaziri, S., Carlson, E. T., Wang, Z. & Connor, C. E. A Channel for 3D Environmental Shape in Anterior Inferotemporal Cortex. *Neuron* **84**, 55–62 (2014).
36. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
37. Leopold, D. A., Bondar, I. V. & Giese, M. A. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* **442**, 572–575 (2006).
38. Freiwald, W. A., Tsao, D. Y. & Livingstone, M. S. A face feature space in the macaque temporal lobe. *Nat. Neurosci.* **12**, 1187–1196 (2009).
39. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).

SUPPLEMENT

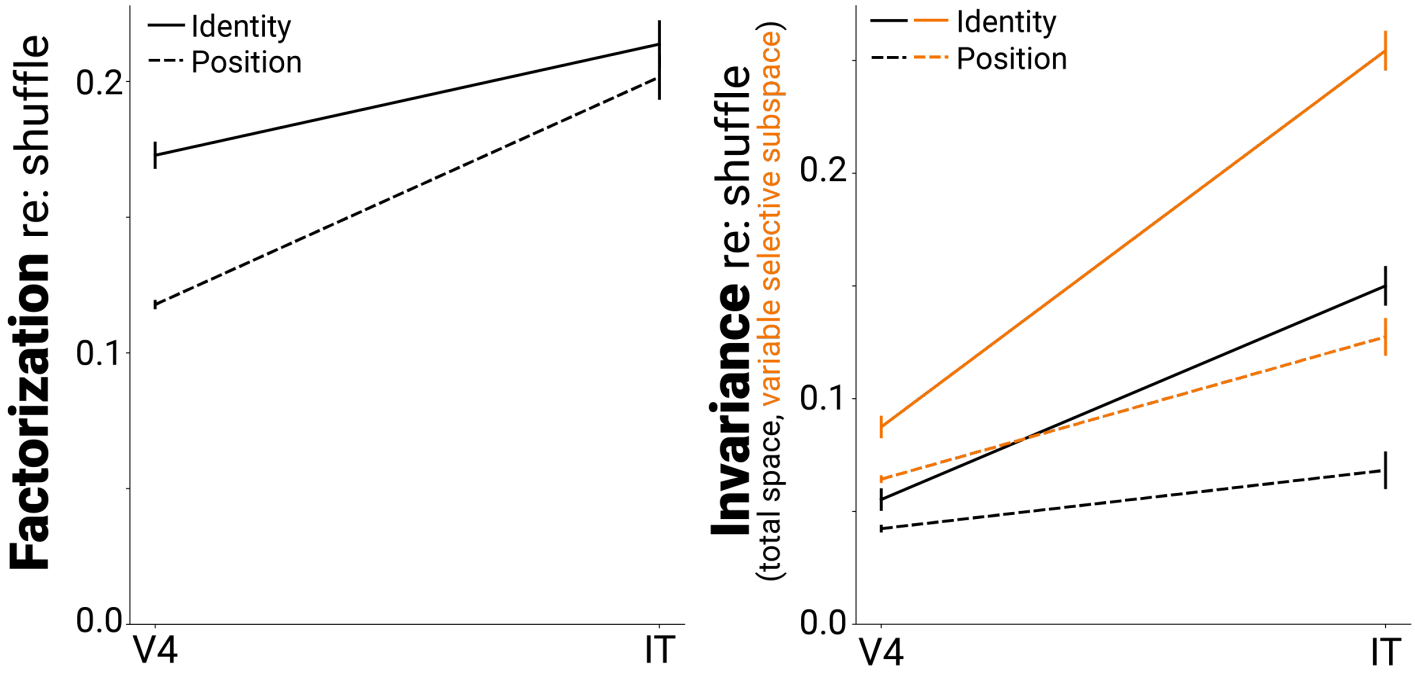
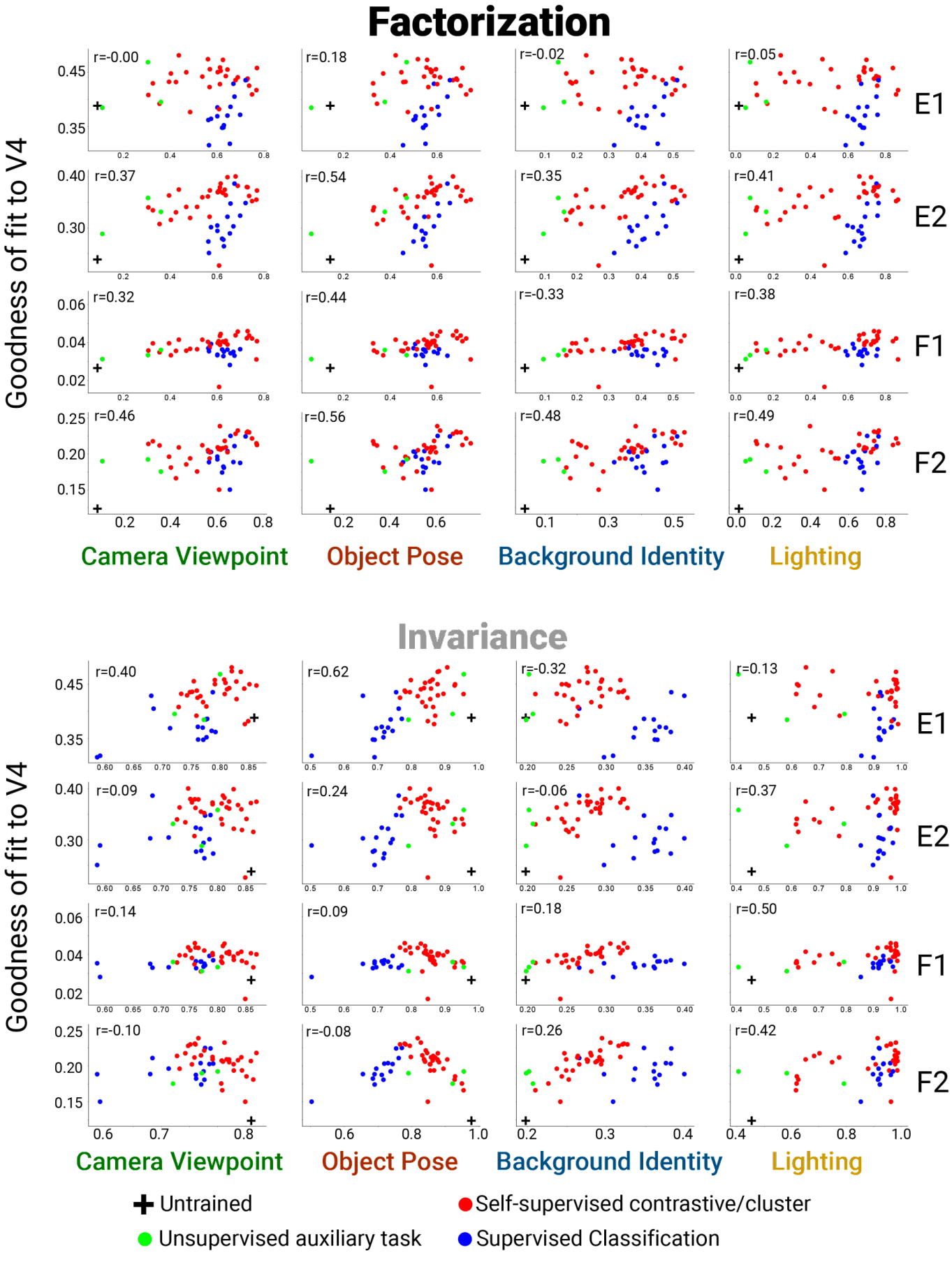
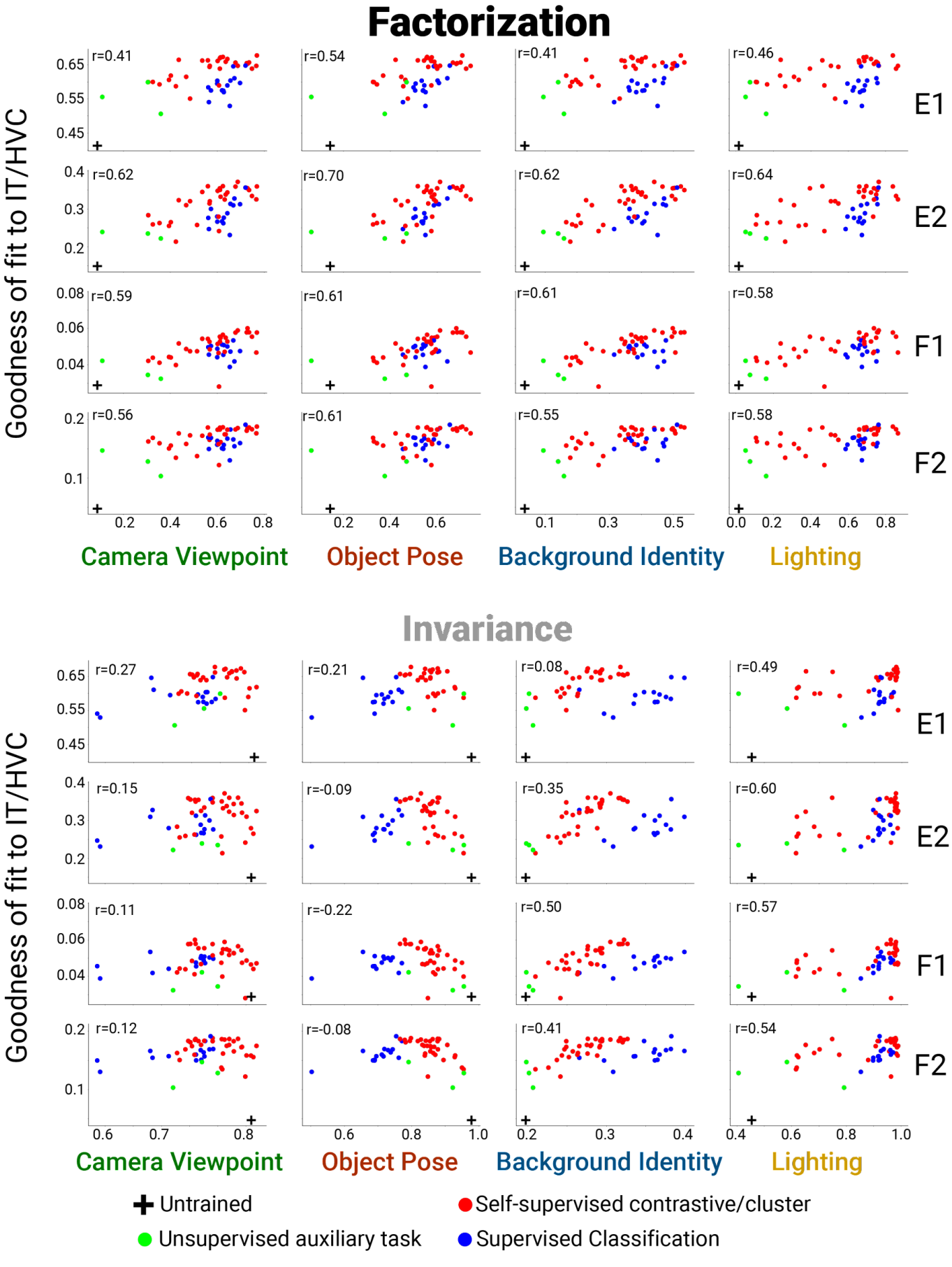


Figure S1. Scatter plots for all datasets. Normalized factorization and invariance as in **Figure 1C,D** but after subtracting shuffle control for V4 and IT neural dataset. Shuffling the image identities of each population vector accounts for increases in factorization driven purely by changes in the covariance statistics of population responses between V4 and IT. However, normalized factorization scores remain significantly above zero for both brain areas.





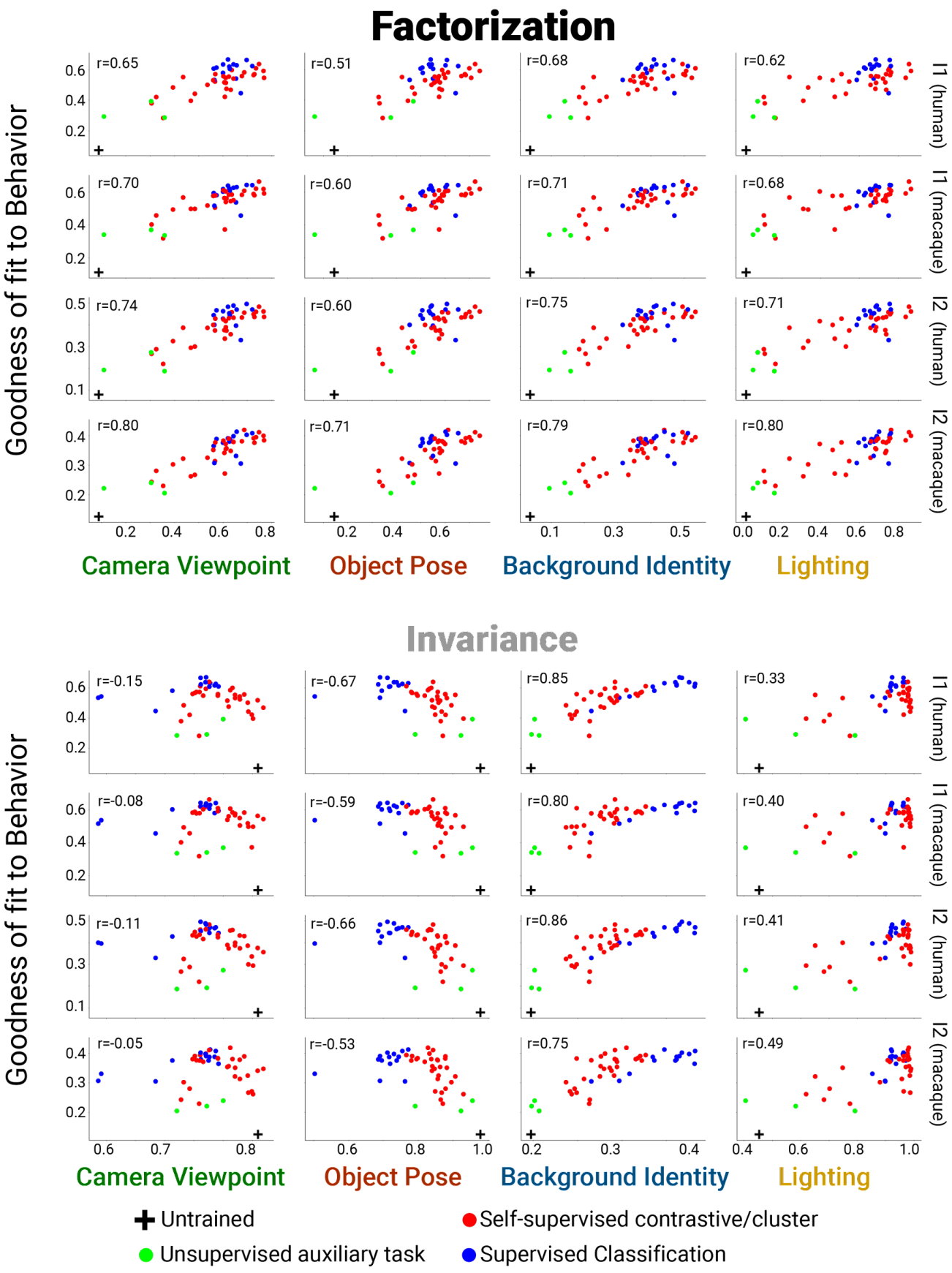


Figure S2. Scatter plots for all datasets. Scatter plots as in **Figure 2C,D** for all datasets. Brain metrics (y-axes) by panel are: (A) macaque neuron/human voxel fits in V4 cortex, (B) macaque neuron/human voxel fits in ITC/HVC, and (C) macaque/human per-image classification performance (I1) and image-by-distractor class performance (I2). In all panels, the plots in the top half use DNN factorization scores on the x-axis while the bottom half use DNN invariance scores.

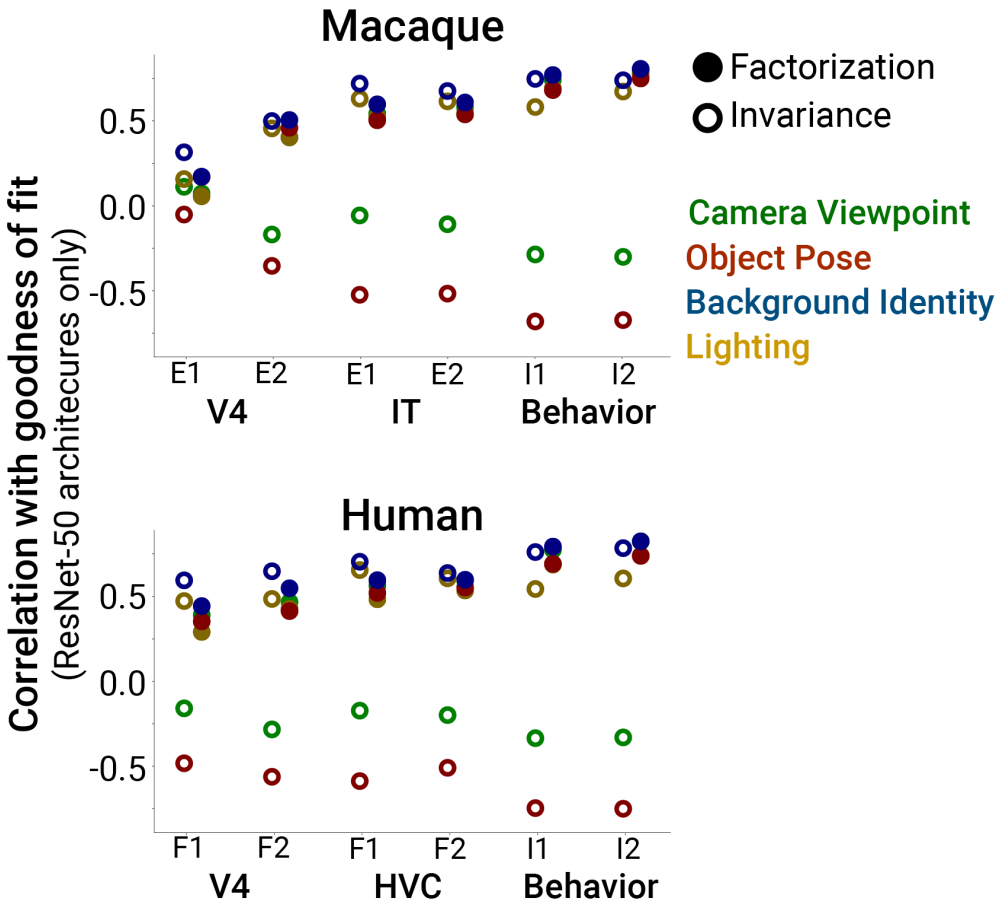


Figure S3. Predictivity of factorization and invariance restricting to ResNet-50 model architectures. Same format as **Figure 3C** except with the analyses restricted to using only models with the Resnet-50 architecture. The main finding of factorization of scene parameters in DNNs being generally positively correlated with better predictions of brain data is replicated using this architecture-matched subset of models, controlling for potential confounds from model architecture.

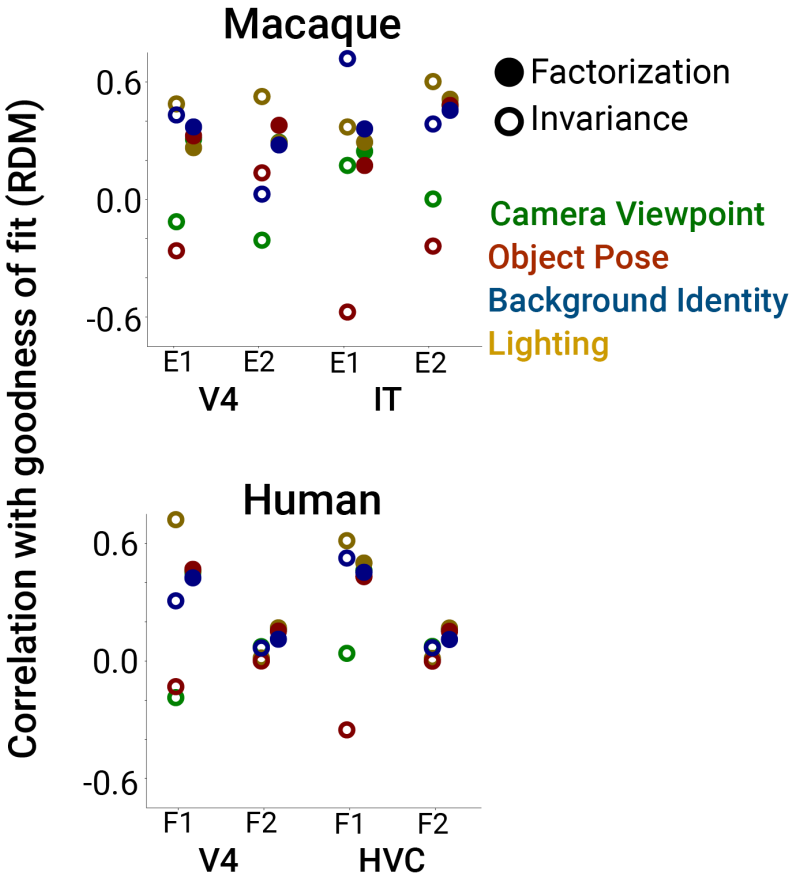


Figure S4. Predictivity of factorization and invariance for RDMs. Same format as **Figure 3C** except for predicting population representational dissimilarity matrices (RDMs) of macaque neurophysiological and human fMRI data (in the main analyses linear encoding fits of each single neuron/voxel were used to measure brain predictivity of a model). The main finding of factorization of scene parameters in DNNs being positively correlated with better predictions of brain data is replicated using RDMs instead of neural/voxel goodness of fit.

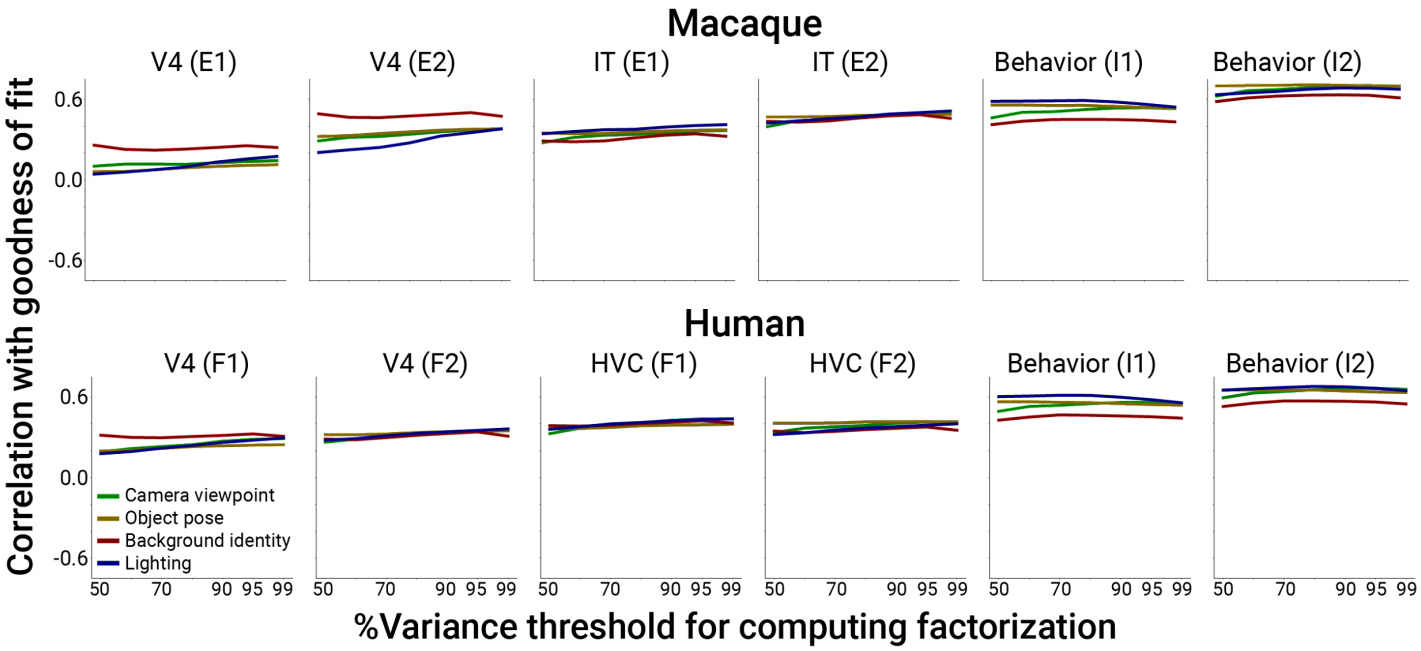


Figure S5. Effect on neural and behavioral predictivity of PCA threshold for computing factorization, Related to Figure 3. The % variance threshold used in the main text for estimating a PCA linear subspace capturing the bulk of the variance induced by all other parameters besides the parameter of interest is somewhat arbitrary. Here we show that results of our main analysis change little if we vary this parameter from 50-99%. In the main text, a PCA threshold of 90% was used for computing factorization scores.

Dataset	Key	#neurons, voxels	#subj	Image Stimuli	#images
DiCarlo-Majaj-Hong 2015[1] Macaque V4 multi-unit activity	E1	128	2	6°, grayscale, synthetic	5760
DiCarlo-Majaj-Hong 2015[1] Macaque IT multi-unit activity	E1	168	2	6°, grayscale, synthetic	5760
DiCarlo-Rust 2012[2] Macaque V4 single neuron	E2	143	2	5°, grayscale, natural	300
DiCarlo-Rust 2012[2] Macaque IT single neuron	E2	142	2	5°, grayscale, natural	300
DiCarlo-Rajalingham-Issa 2018[3] Macaque behavior Image-level classification	I1	N/A	5	6-8°, grayscale, synthetic	240
DiCarlo-Rajalingham-Issa 2018[3] Macaque behavior Image x class confusion matrix	I2	N/A	5	6-8°, grayscale, synthetic	240
Gallant-Kay 2008[4] Human V4 fMRI (dataset)	F1	2,557	2	20°, grayscale, natural	1870
Gallant-Kay 2008[4] Human HVC fMRI (dataset)	F1	1,286	2	20°, grayscale, natural	1870
Horikawa-Kamitani 2019[5] Human V4 fMRI (dataset)	F2	3,377	3	12°, color, natural	1250
Horikawa-Kamitani 2019[5] Human HVC fMRI (dataset)	F2	14,465	3	12°, color, natural	1250
DiCarlo-Rajalingham-Issa 2018[3] Human behavior Image-level classification	I1	N/A	1472	6-8°, grayscale, synthetic	240
DiCarlo-Rajalingham-Issa 2018[3] Human behavior Image x class confusion matrix	I2	N/A	1472	6-8°, grayscale, synthetic	240

Table S1. Datasets used for measuring similarity of models to the brain. Datasets from both macaque and human high-level visual cortex as well as high-level visual behavior were collated for testing the brainlikeness of computational models. For neural and fMRI datasets, the features in the model were used to predict the image-by-image response pattern of each neuron or voxel. For behavior datasets, the performance of linear decoders built atop model representations were compared to performance per image of macaques and humans.

Model	Architecture	Loss Function	Customization
SimCLR[6]	ResNet-50	Self-supervised (contrastive)	--------
SimCLR[6]	ResNet-50	Self-supervised (contrastive)	2x wide
SimCLR[6]	ResNet-152	Self-supervised (contrastive)	2x wide
SimCLR[6]	ResNet-50	Self-supervised (contrastive)	no projection head
SimCLR[6]	ResNet-50	Self-supervised (contrastive)	only crop augmentations
SimCLR[6]	ResNet-50	Self-supervised (contrastive)	only crop augmentations, temperature 0.2
SimCLR[6]	ResNet-50	Self-supervised (contrastive)	only crop augmentations, temperature 0.05
SimCLR[6]	ResNet-50	Self-supervised (contrastive)	only crop and blur augmentations
SimCLR[6]	ResNet-50	Self-supervised (contrastive)	only crop and non-hue color jitter augmentations
SimCLR[6]	ResNet-50	Self-supervised (contrastive)	only crop, blur, and non-hue color jitter augmentations
MOCO[7]	ResNet-50	Self-supervised (contrastive)	--------
MOCO v2[8]	ResNet-50	Self-supervised (contrastive)	--------
MOCO v2[8]	ResNet-50	Self-supervised (contrastive)	only crop augmentations
MOCO v2[8]	ResNet-50	Self-supervised (contrastive)	only crop, color jitter, and grayscale augmentations
MOCO v2[8]	ResNet-50	Self-supervised (contrastive)	only crop augmentations, all image inputs preprocessed to grayscale
MOCO v2[8]	ResNext-50	Self-supervised (contrastive)	--------
MOCO v2[8]	ResNet-18	Self-supervised (contrastive)	--------
Instance discrimination[9]	ResNet-50	Self-supervised (image discrimination)	--------
InfoMin[10]	ResNet-50	Self-supervised (contrastive)	--------
InfoMin[10]	ResNext-101	Self-supervised (contrastive)	--------

InfoMin[10]	ResNext-152	Self-supervised (contrastive)	--------
SwAV[11]	ResNet-50	Self-supervised (cluster)	--------
Deep clustering v2[12]	ResNet-50	Self-supervised (cluster)	--------
BYOL[13]	ResNet-50	Self-supervised (no negative examples)	--------
BYOL[13]	ResNet-50	Self-supervised (no negative examples)	only crop augmentations during training
BYOL[13]	ResNet-50	Self-supervised (no negative examples)	only crop and blur augmentations
BYOL[13]	ResNet-50	Self-supervised (no negative examples)	without color jitter augmentation
BYOL[13]	ResNet-50	Self-supervised (no negative examples)	without grayscale augmentation
BYOL[13]	ResNet-50	Self-supervised (no negative examples)	batch size 64
BYOL[13]	ResNet-50	Self-supervised (no negative examples)	batch size 512
Relative patch location[14]	ResNet-50	Auxiliary task (determine relative positions of image patches)	--------
Rotation prediction[14]	ResNet-50	Auxiliary task (infer rotations that were applied given a set of images)	--------
Colorization[15]	ResNet-50	Auxiliary task: (colorize grayscale images)	--------
Jigsaw puzzle[16]	ResNet-50	Auxiliary task: (determine relative positions of image patches)	--------
Big BiGAN[17]	ResNet-50	Auxiliary task (autoencoder objective with	--------

		reconstruction error parameterized using a neural network discriminator)	
ResNet[18]	ResNet-50	Supervised (classification)	--------
ResNet[18]	ResNet-50	Supervised (classification)	MOCO data augmentations used during training
ResNet[18]	ResNet-50	Supervised (classification)	no data augmentation used during training
ResNet[18]	ResNet-18	Supervised (classification)	--------
ResNet[18]	ResNet-101	Supervised (classification)	--------
Wide ResNet[19]	ResNet-50	Supervised (classification)	--------
AlexNet[20]	AlexNet	Supervised (classification)	--------
GoogLeNet[21]	GoogLeNet	Supervised (classification)	--------
Inception-v3[22]	Inception-v3	Supervised (classification)	--------
DenseNet[23]	DenseNet-169	Supervised (classification)	--------
DenseNet[23]	DenseNet-121	Supervised (classification)	--------
VGG[24]	VGG-11	Supervised (classification)	--------
VGG[24]	VGG-13	Supervised (classification)	--------
VGG[24]	VGG-16	Supervised (classification)	--------
VGG[24]	VGG-19	Supervised (classification)	--------
MobileNet[25]	MobileNet	Supervised (classification)	--------
SqueezeNet[26]	SqueezeNet-10	Supervised (classification)	--------
SqueezeNet[26]	SqueezeNet-11	Supervised (classification)	--------
ResNet[27]	ResNext-50	Supervised (classification)	--------

ResNet[27]	ResNet-101	Supervised (classification)	--------
MnasNet[28]	MnasNet_05	Supervised (classification)	--------
MnasNet[28]	MnasNet_10	Supervised (classification)	--------
ShuffleNet[29]	ShuffleNet_05	Supervised (classification)	--------
ShuffleNet[29]	ShuffleNet_10	Supervised (classification)	--------

Table S2. Models tested. For each model, we measured representational factorization and invariance in each of the final five representational layers of the model as well as evaluating their brainlikeness using the datasets in **Table S1.**

REFERENCES

1. Majaj, N. J., Hong, H., Solomon, E. A. & DiCarlo, J. J. Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *J. Neurosci.* **35**, 13402–13418 (2015).
2. Rust, N. C. & DiCarlo, J. J. Balanced Increases in Selectivity and Tolerance Produce Constant Sparseness along the Ventral Visual Stream. *J. Neurosci.* **32**, 10170–10182 (2012).
3. Rajalingham, R. *et al.* Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *J. Neurosci.* **38**, 7255–7269 (2018).
4. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
5. Shen, G., Horikawa, T., Majima, K. & Kamitani, Y. Deep image reconstruction from human brain activity. *PLOS Comput. Biol.* **15**, e1006633 (2019).
6. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *ArXiv200205709 Cs Stat* (2020).
7. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. *ArXiv191105722 Cs* (2020).
8. Chen, X., Fan, H., Girshick, R. & He, K. Improved Baselines with Momentum Contrastive Learning. *ArXiv200304297 Cs* (2020).
9. Wu, Z., Xiong, Y., Yu, S. & Lin, D. Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination. *ArXiv180501978 Cs* (2018).
10. Tian, Y. *et al.* What makes for good views for contrastive learning. *ArXiv200510243 Cs* (2020).
11. Caron, M. *et al.* Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. *ArXiv200609882 Cs* (2020).
12. Caron, M., Bojanowski, P., Joulin, A. & Douze, M. Deep Clustering for Unsupervised Learning of Visual Features. *ArXiv180705520 Cs* (2019).
13. Grill, J.-B. *et al.* Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. *ArXiv200607733 Cs Stat* (2020).
14. Gidaris, S., Bursuc, A., Komodakis, N., Pérez, P. & Cord, M. Boosting Few-Shot Visual Learning with Self-Supervision. Preprint at https://doi.org/10.48550/arXiv.1906.05186 (2019).
15. Zhang, R., Isola, P. & Efros, A. A. Colorful Image Colorization. Preprint at https://doi.org/10.48550/arXiv.1603.08511 (2016).
16. Noroozi, M. & Favaro, P. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. Preprint at https://doi.org/10.48550/arXiv.1603.09246 (2017).
17. Donahue, J. & Simonyan, K. Large Scale Adversarial Representation Learning. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).
18. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *ArXiv151203385 Cs* (2015).
19. Zagoruyko, S. & Komodakis, N. Wide Residual Networks. Preprint at https://doi.org/10.48550/arXiv.1605.07146 (2017).
20. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).

21. Szegedy, C. *et al.* Going Deeper with Convolutions. *ArXiv14094842 Cs* (2014).
22. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. Preprint at https://doi.org/10.48550/arXiv.1512.00567 (2015).
23. Huang, G., Liu, Z., van der Maaten, L. & Weinberger, K. Q. Densely Connected Convolutional Networks. Preprint at https://doi.org/10.48550/arXiv.1608.06993 (2018).
24. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Preprint at https://doi.org/10.48550/arXiv.1409.1556 (2015).
25. Howard, A. G. *et al.* MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Preprint at https://doi.org/10.48550/arXiv.1704.04861 (2017).
26. Iandola, F. N. *et al.* SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. Preprint at https://doi.org/10.48550/arXiv.1602.07360 (2016).
27. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated Residual Transformations for Deep Neural Networks. *ArXiv161105431 Cs* (2017).
28. Tan, M. *et al.* MnasNet: Platform-Aware Neural Architecture Search for Mobile. Preprint at https://doi.org/10.48550/arXiv.1807.11626 (2019).
29. Zhang, X., Zhou, X., Lin, M. & Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. Preprint at https://doi.org/10.48550/arXiv.1707.01083 (2017).