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 and has been used as an optimization target for deep neural network models (DNNs) of the visual system. However, visual brain areas represent many different types of information, and optimizing for classification of object identity alone does not constrain how other information may be encoded in visual representations. Information about different scene parameters may be discarded altogether ("invariance"), represented in non-interfering subspaces of population activity ("factorization") or encoded in an entangled fashion. In this work, we provide evidence that factorization is a normative principle of biological visual representations. In the monkey ventral visual hierarchy, we found that factorization of object pose and background information from object identity increased in higher-level regions and strongly contributed to improving object identity decoding performance. We then conducted a 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. Notably, invariance to these parameters was not consistently associated with matches to neural and behavioral data, suggesting that maintaining non-class information in factorized activity subspaces is preferred to dropping it altogether. Thus, we propose that factorization of visual scene information is a widely used strategy in brains and DNN models thereof.

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, it remains opaque what aspects of the representations of these more performant models drive them to better match neural data. Moreover, beyond a certain threshold level of object classification performance, further improvement fails to produce a concomitant improvement in predicting primate neural responses[2,4,5]. This weakening trend motivates finding new

normative principles, besides object classification ability, that push models to better match primate visual representations.

One strategy for achieving high object classification performance is to form representations that discard (are invariant to) all information besides object class. However, high-level cortical neurons in the primate ventral visual stream are known to encode other variables in visual input besides object identity, such as object pose[6–9]. How do visual representations encode these multiple forms of information simultaneously? Here, we introduce methods to quantify the relationships between different types of visual information in a population code (e.g., object pose vs. camera viewpoint), and in particular the degree to which different forms of information are "factorized." Intuitively, if the variance driven by one parameter is encoded independently from, or uncorrelated with, the variance driven by other scene parameters, we say this code is factorized. Factorization is notably distinct from building invariance to scene parameters, which renders neural activity insensitive to variation in those parameters. Invariance is a zero-sum strategy: building invariance to some parameters improves the ability to decode others. Factorization, by contrast, can enable simultaneous decoding of many parameters at once, supporting diverse visually guided behaviors (e.g., spatial navigation, object manipulation or object classification)[10]. We note that our definition of factorization is closely related to the existing concept of manifold disentanglement[11,12] and can be seen as a generalization of disentanglement to high-dimensional visual scene parameters like object pose.

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 positively correlated with models' matches to neural and behavioral data. While 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 that capture scene understanding as performed by biological vision.

RESULTS

Disentangling object identity manifolds in population responses can be achieved by qualitatively different strategies, including: building invariance of responses to non-identity scene parameters and/or factorizing non-identity-driven response variance into isolated (factorized) subspaces (**Figure 1A**, left vs. center panels, cylindrical/spherical shaded regions represent object manifolds). Both strategies maintain an "identity subspace" in which object manifolds are linearly separable. In non-invariant and non-factorized representations, other variables like camera viewpoint also drive variance within the identity subspace, "entangling" the representations of the two variables (**Figure 1A**, right; viewpoint driven variance is mainly in identity subspace, orange flat shaded region). To formalize these different representational strategies, we introduced measures of factorization and invariance to scene parameters in neural population responses (**Figure 1B;** see Equations 1 and 2 in **Methods**). Concretely, invariance to a scene variable (e.g. object motion) is computed by measuring the degree to which varying that parameter alone changes neural responses, relative to the changes induced by varying other parameters (lower relative influence on neural activity corresponds to higher invariance). Factorization is computed by identifying the axes in neural population activity space that are influenced by varying the parameter of interest and assessing how much it overlaps the axes influenced by other parameters ("*a*" in **Figure 1B,C**)(lower overlap corresponds to higher factorization). For instance, a neural population in which one neural subpopulation encodes object

identity and another separate subpopulation encodes object position exhibits a high degree of factorization of those two parameters (however, note that factorization may also be achieved by neural populations with mixed selectivity in which the "subpopulations" correspond to subspaces, or independent orthogonal linear projections, of neural activity space rather than physical subpopulations). Factorization, unlike invariance, has the potential to enable the simultaneous representation of multiple scene parameters in a decodable fashion. Intuitively, factorization increases with higher dimensionality as this decreases overlap, all other things being equal (in the limit, the angle between points will approach 90° or a fully orthogonal code in high dimensions), and for a given finite, fixed dimension, factorization is mainly driven by the angle between this dimension and other variable subspaces which measures the degree of contamination (**Figure 1C**; square vs. parallelogram). In a simulation, the extent to which the variables of interest were represented in a factorized way (i.e. along orthogonal axes, rather than correlated axes) influenced the ability of a linear discriminator to successfully decode both variables in a generalizable fashion from a few training samples (**Figure 1C**).

Given the theoretically desirable properties of factorized representations, we next asked whether such representations are observed in neural data, and how much factorization contributes empirically to downstream decoding performance in real data. Specifically, we took advantage of an existing dataset in which the tested images independently varied object identity versus object pose plus background context[13]. We found that both V4 and IT responses exhibited more significant factorization of object identity information from non-identity information than a shuffle control (which accounts for effects on factorization due to changes in dimensionality between regions) (**Figure S1**; see **Methods**). Furthermore, the degree of factorization increased from V4 to IT (**Figure 2A**). Consistent with prior studies, we also found that invariance to non-identity information increased from V4 to IT in our analysis (**Figure 2A**, right, solid lines)[14]. Invariance to non-identity information was even more pronounced when measured in the subspace of population activity capturing the bulk (90%) of identity-driven variance, as a consequence of increased factorization of identity from non-identity information (**Figure 2A**, right, dashed lines). To illustrate the beneficial effect of factorization on decoding performance, we performed a statistical lesion experiment that precisely targeted this aspect of representational geometry. Specifically, 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**). Note that this transformation, designed to decrease factorization, acts on the angle between latent variable subspaces. The applied linear basis rotation leaves all other activity statistics completely intact (such as mean neural firing rates, covariance structure of the population, and its invariance to non-class variables) yet has the effect of strongly reducing object identity decoding performance in both V4 and IT (**Figure 2B)**. Our analysis shows that maintaining invariance alone in the neural population code was insufficient to account for a large fraction of decoding performance in the cortex; factorization of non-identity variables is key to the decoding performance achieved by V4 and IT representations.

We next asked whether factorization is found in deep neural network (DNN) model representations and whether this novel, heretofore unconsidered metric is a strong indicator of more brainlike models. When working with computational models, we have the liberty to test an arbitrary number of stimuli; therefore, we could independently vary multiple scene parameters at sufficient scale to enable computing factorization and invariance for each, and we explored factorization in DNN model representations in more depth than previously measured in existing neural experiments. To gain insight back into neural representations, we also assessed the ability of each model to predict separately collected neural and behavioral data. In this fashion, we may indirectly assess the relative significance of geometric properties like factorization and invariance to biological visual representations – if, for instance, models with more factorized representations consistently match

neural data more closely, we may infer that those neural representations likely exhibit factorization themselves (**Figure 3A**). To measure factorization, invariance, and decoding properties of DNN models, we generated an augmented image set, based on the images used in the previous dataset (**Figure 2A**), 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, allowing us to measure the variance induced by each parameter relative to the variance across all scene parameters (**Figure 3B**; 100 base scenes and 10 transformed images for each source of variation). We presented this large image dataset to models (4000 images total) to assess the relative degree of representational factorization of and invariance to each scene parameter. We conducted this analysis across a broad range of DNNs varying in architecture and objective as well as other implementational choices to obtain the widest possible range of DNN representations for testing our hypothesis. These included models using supervised training for object classification[15,16], contrastive self-supervised training[17,18], and self-supervised models trained using auxiliary objective functions[19–22] (see ***Methods*** and **Table S2**).

First, we asked whether, in the course of training, DNN models develop factorized representations at all. We found that the final 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 3C**, top row, rightward shift relative to black cross, a randomly initialized ResNet-50). By contrast, training DNNs produced mixed effects on invariance, typically increasing it for background and lighting but reducing it for object pose and camera viewpoint (**Figure 3C**, bottom row, leftward shift relative to black cross for left two panels). Moreover, we found that the degree of factorization in models correlated with the degree to which they predicted neural activity for single-unit IT data (**Figure 3C**, top row), which can be seen as correlative evidence that neural representations in IT exhibit factorization of all scene variables tested. Interestingly, we saw a different pattern for representational invariance to a scene parameter. Invariance showed mixed correlations with neural predictivity (**Figure 3C**, bottom row), suggesting that IT neural representations build invariance to some scene information (background and lighting) but not to others (object pose and observer viewpoint). Similar effects were observed when we assessed correlations between these metrics and fits to human behavioral data (rather than macaque neural data) (**Figure 3D**).

To assess the robustness of these findings to choice of images and brain regions used in an experiment, we conducted the same analyses across a large and diverse set of previously collected neural and behavioral datasets, from different primate species and visual regions (6 macaque datasets[13,23,24]: two V4, two IT, and two behavior; 6 human datasets[24–26]: two V4, two HVC, and two behavior; **Table S1**). Consistently, increased factorization of scene parameters in model representations correlated with models being more predictive of neural spiking responses, voxel BOLD signal, and behavioral responses to images (**Figure 4A**, 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 4C**, red and green circles). Our results were preserved when we re-ran the analyses using only the subset of models with the identical ResNet-50 architecture (**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 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**). Finally, we tested whether our results generalized across the particular image set used for computing the model factorization scores in the first place. Here, instead of relying on our synthetically generated images, where each scene parameter was directly

controlled, we re-computed factorization from two types of relatively unconstrained natural movies, one where the observer moves in an urban environment (approximates camera viewpoint changes)[27] and another where objects move in front of a fairly stationary observer (approximates object pose changes)[28]. Similar to the result found for factorization measured using augmentations of synthetic images, factorization of frame-by-frame variance (local in time, presumably dominated by either observer or camera motion; see **Methods**) from other sources of variance across natural movies (non-local in time) was correlated with improved neural predictivity in both macaque and human data while invariance to local frame-by-frame differences was not (**Figure 4B**; black versus gray bars). Thus, we have shown that a main finding – the importance of object pose and camera viewpoint factorization for achieving brainlike representations – holds across types of brain signal (spiking vs. BOLD), species (monkey vs. human), cortical brain areas (V4 vs. IT), images used in experiments (synthetic, grayscale vs. natural, color), and image sets for computing the metric (synthetic images vs. natural movies).

Our analysis of DNN models provides strong evidence that greater factorization of a variety of scene variables is consistently associated with a stronger match to neural and behavioral data. Prior work has identified similar correlation between object classification performance (measured fitting a decoder for object class using model representations) and fidelity to neural data[3]. A priori, it is possible that the correlations we have demonstrated between scene parameter factorization and neural fit can be entirely captured by the known correlation between classification performance and neural fits[2,3], as factorization and classification may themselves be correlated. However, we found that factorization scores significantly boosted cross-validated predictive power of neural/behavioral fit performance compared to simply using object classification alone (**Figure 5A**). Thus, considering factorization in addition to object classification performance improves upon our prior understanding of the properties of more brainlike models (**Figure 5B**).

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[3,29]. These include representations that are completely invariant to non-class information[11,12] and representations that retain a high-dimensional but factorized encoding of non-class information, which disentangles the representation of multiple variables (**Figure 1A**, left). Here, we presented evidence that factorization of non-class information is an important strategy used, alongside invariance, by the high-level visual cortex (**Figure 2**) and by DNNs that are predictive of primate neural and behavioral data (**Figures 3,4**). Prior work has indicated that building representations that support object classification performance, and representations that preserve high-dimensional information about natural images, are both important principles of the primate visual system[1,30] (though see Conwell et al.[31] who find that effective dimensionality alone may not be a robust predictor in high-level visual cortex). Our work unifies these two perspectives, as factorization of multiple scene variables enables high object identity decoding performance while preserving high-dimensional information for various other properties of the visual scene. Disentangled, factorized variable manifolds can enable task performance in the low-training sample regime (**Figure 1C**), which was the focus of recent work on few-shot object classification in the visual system[32]. Our work is also complementary to work on representational straightening of natural movie trajectories in the population space[33]. This work suggests that visual representations maintain a locally linear code of latent variables like viewpoint, while our work focuses on the global arrangement of the linear subspaces affected by different variables. The idea of local straightening of natural movies was found to be predictive for early visual

cortex neural responses but not necessarily for high-level visual cortex[34], where the present work suggests factorization may play a role.

Going forward, we expect factorization could prove to be a useful objective function for optimizing neural network models that better resemble primate visual systems. Complementing prior theoretical work on representational geometry for manifolds of a single variable, namely object class[12], we show that it is the orthogonal relationships between population subspaces encoding different variables that is a consistent principle of brain-like visual representations[35]. 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[36,37], (2) Separate brain subregions encode qualitatively different latent variables but using distributed representations within each region[38–40], (3) Each neuron encodes multiple variables in a distributed population code, such that the factorization of different variables is only apparent as independent directions when assessed in high-dimensional population activity space[36,41]. Future work can disambiguate among these possibilities by systematically examining ventral visual stream subregions[8,40,42] and single neuron tuning curves within them[43,44].

METHODS

Monkey Datasets. Macaque monkey datasets were of single-unit neural recordings[23], multi-unit neural recordings[13], and object recognition behavior[24]. Single-unit spiking responses to natural images were measured in V4 and anterior ventral IT[23]. The advantages of this dataset is that it contains well-isolated single neurons, the gold standard for electrophysiology. Furthermore, the IT recordings were obtained from penetrating electrodes targeting the anterior ventral portion of IT near the base of 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 are more easily placed such as CIT and PIT[13], 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 Supplementary Table S1). Finally, the monkey behavioral dataset came from a third study examining the image-by-image object classification performance of macaques and humans[24].

Human Datasets. Three datasets from humans were used, two fMRI datasets and one object recognition behavior dataset[4,24,25]. The fMRI datasets used different images (color versus grayscale) but otherwise used fairly similar number of images and voxel resolution in imaging. Human fMRI studies have found that different DNN layers tend to map to V4 and HVC human fMRI voxels[4]. The human behavioral dataset measured image-by-image classification performance and was collected in the same study as the monkey behavioral signatures[24].

Computational Models. In recent years, 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. 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, though some in our library utilized different architectures. The randomly initialized network control utilized ResNet-50 (see **Figure 3C, D**). The set of models we used is listed in **Table S2**.

Simulation of Factorized versus Non-factorized Representational Geometries. For the simulation in **Figure 1C**, 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 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 used in **Figure 2B**, 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 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 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 (considering the simulation in **Figure 1C** of two binary variables, this basis change of the neural data in **Figure 2B** is equivalent to turning a square into the maximally flat parallelogram, the degenerate one where all the points are collinear) .

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[17,18].

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 a 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 images 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)[27,28]. Here, the *all parameters* condition is simply the variance across all movie frames which in the case of MIT Moments in Time dataset includes variance across thousands of video clips taken in many different settings and in the case of the UT-Austin Egocentric movie dataset includes variance across only 4 movies but over long durations of time during which an observer translates extensively in an 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 and thus effectively captured variance induced by these non-spatial parameters[28]. In the UT Austin Egocentric movie set, new objects are encountered as the subject navigates around the urban landscape[27].

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)[45], computing dot product similarities of the representations of all pairs of images and measuring the Spearman correlation coefficient between these pairwise similarity matrices obtained from a given model and neural dataset, respectively.

Model Behavioral Signatures. We followed the approach of Rajalingham, Issa et al.[24] 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 used in the experiments of Rajalingham, Issa et al.[24], 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 and I2 vectors and matrices, respectively (both I1 and I2 results are reported).

Model Layer Choices. The scatter plots in **Figure 3C,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 even may 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 linear 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 estimate 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. 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

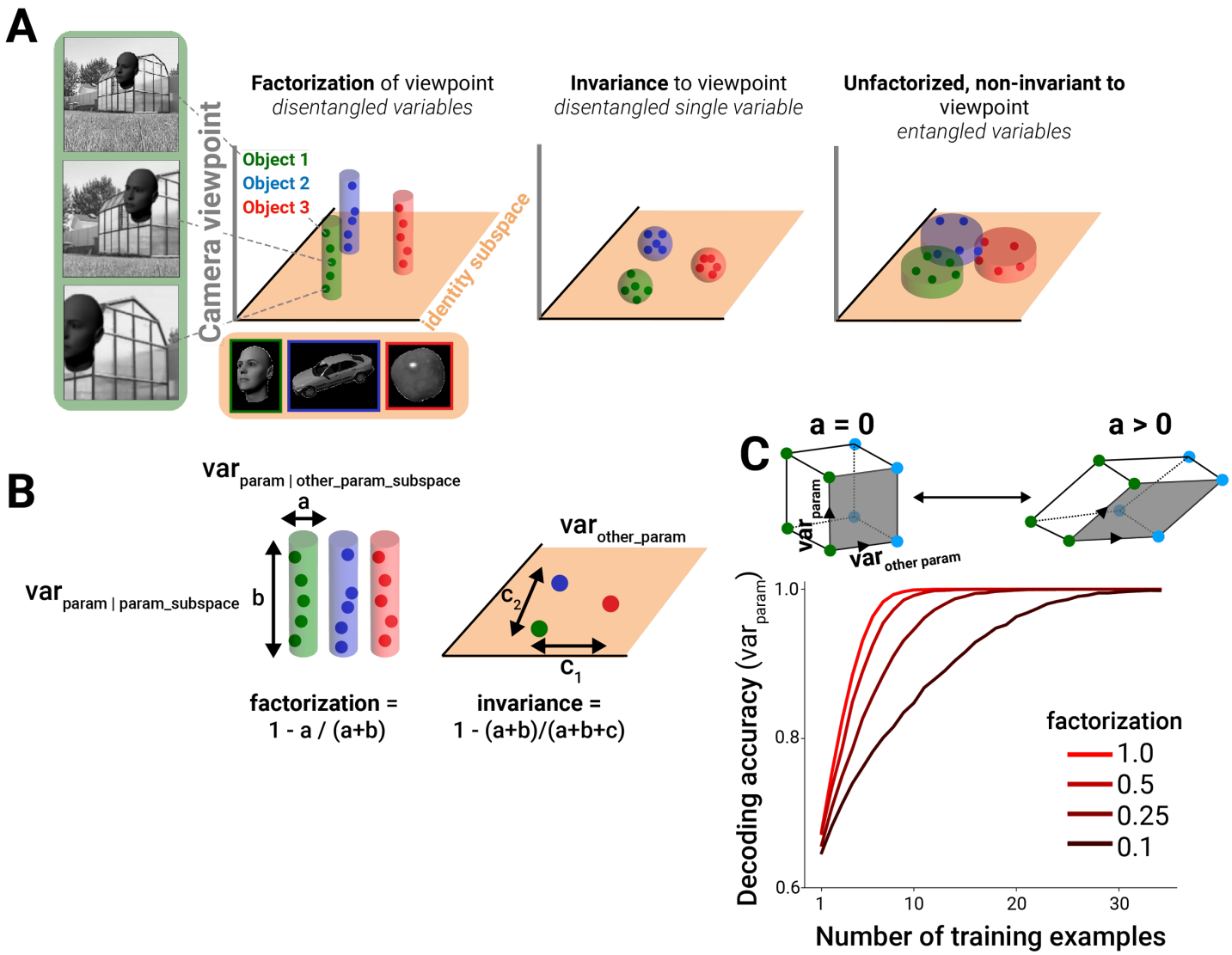


Figure 1. Framework for quantifying factorization in neural and model representations. (A) A subspace for encoding object identity in a linearly separable manner can be achieved by becoming invariant to non-class variables (compact spheres, middle column; colored dots represent example images within each class) or by encoding variance induced by non-identity variables in orthogonal neural axes to the identity subspace (extended cylinders, left column), but only the factorization strategy simultaneously represents multiple variables in a disentangled fashion. A code that is sensitive to non-identity parameters within the identity subspace corrupts the ability to decode identity (right column) (identity subspace in orange). (B) Variance across images within a class can be measured in two different linear subspaces: that containing the majority of variance for all other parameters *(a, "other_param_subspace")* and that containing the majority of the variance for that parameter *(b, "param_subspace")*. Factorization is defined as the fraction of parameter-induced variance that avoids the other-parameter subspace (left). By contrast, invariance to the parameter of interest is computed by comparing the overall parameter-induced variance to the variance in response to other parameters *(c, "var_other_param")* (right). (C) In a simulation of coding strategies for two binary variables out of 10 total dimensions that are varying (see **Methods**), a decrease in orthogonality of the relationship between the encoding of the two variables, or *a*>0 (going from a square to a parallelogram geometry; only 3 of 10 dimensions used in simulation are shown), despite maintaining linear separability of variables, results in poor classifier performance in the few training-samples regime.

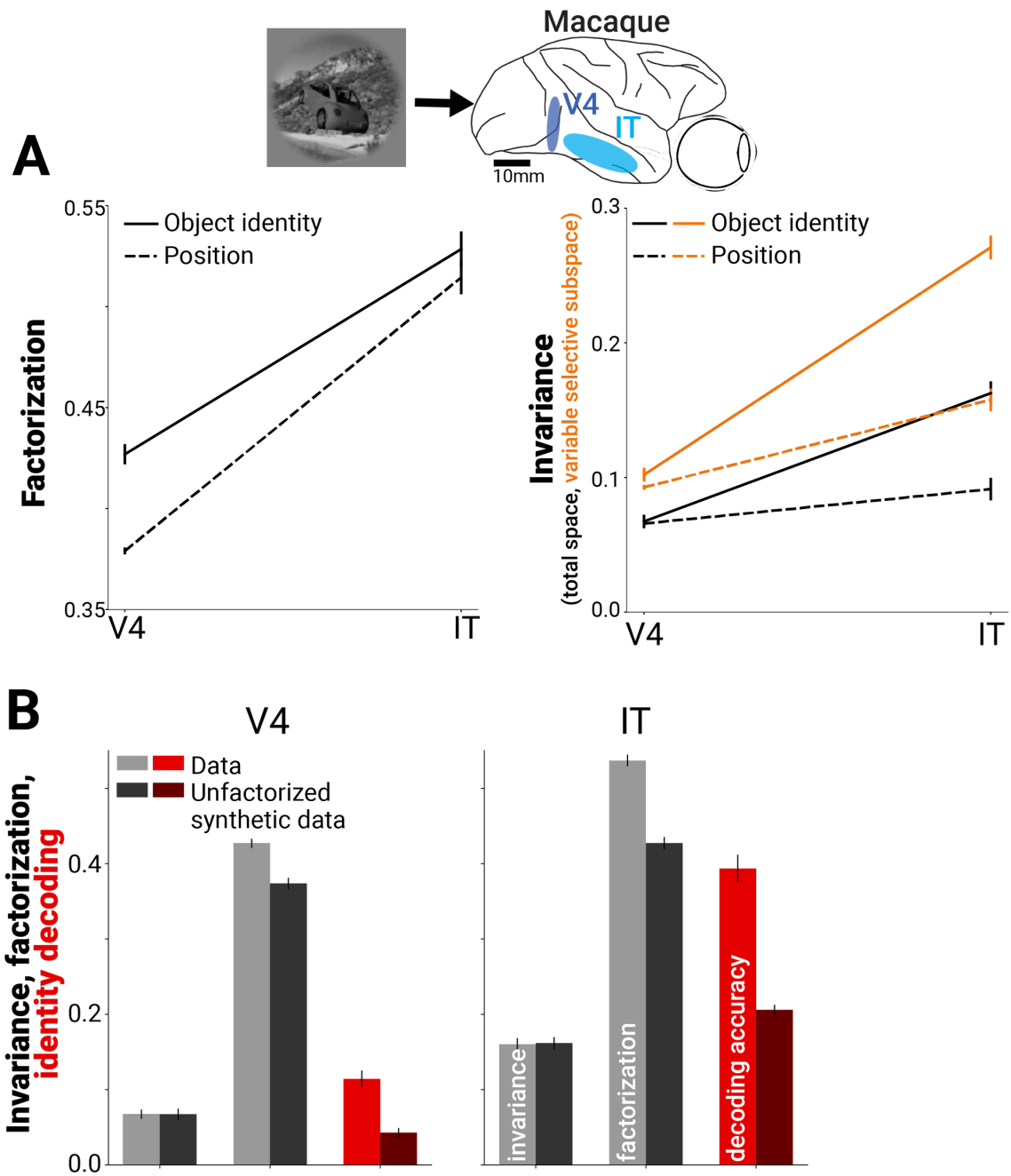


Figure 2. Benefit of factorization to neural decoding in macaque V4 and IT. (A) Factorization of object identity and position increased from macaque V4 to IT (dataset E1 – multiunit activity in macaque visual cortex)(left). Like factorization, invariance also increased from V4 to IT (note, "identity" refers to invariance to all non-identity position factors, solid black line)(right). Combined with increased factorization of the remaining variance, this led to higher invariance within the variable's subspace (orange lines), representing a neural subspace for identity information with invariance to nuisance parameters which decoders can target for read-out. (B) Applying a transformation to the data that rotated the relative positions of mean responses to object classes (see **Methods**), designed to preserve relevant activity statistics (including invariance to non-class factors) 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 = 1/64; n=128 multi-unit sites in V4 and 128 in IT) .

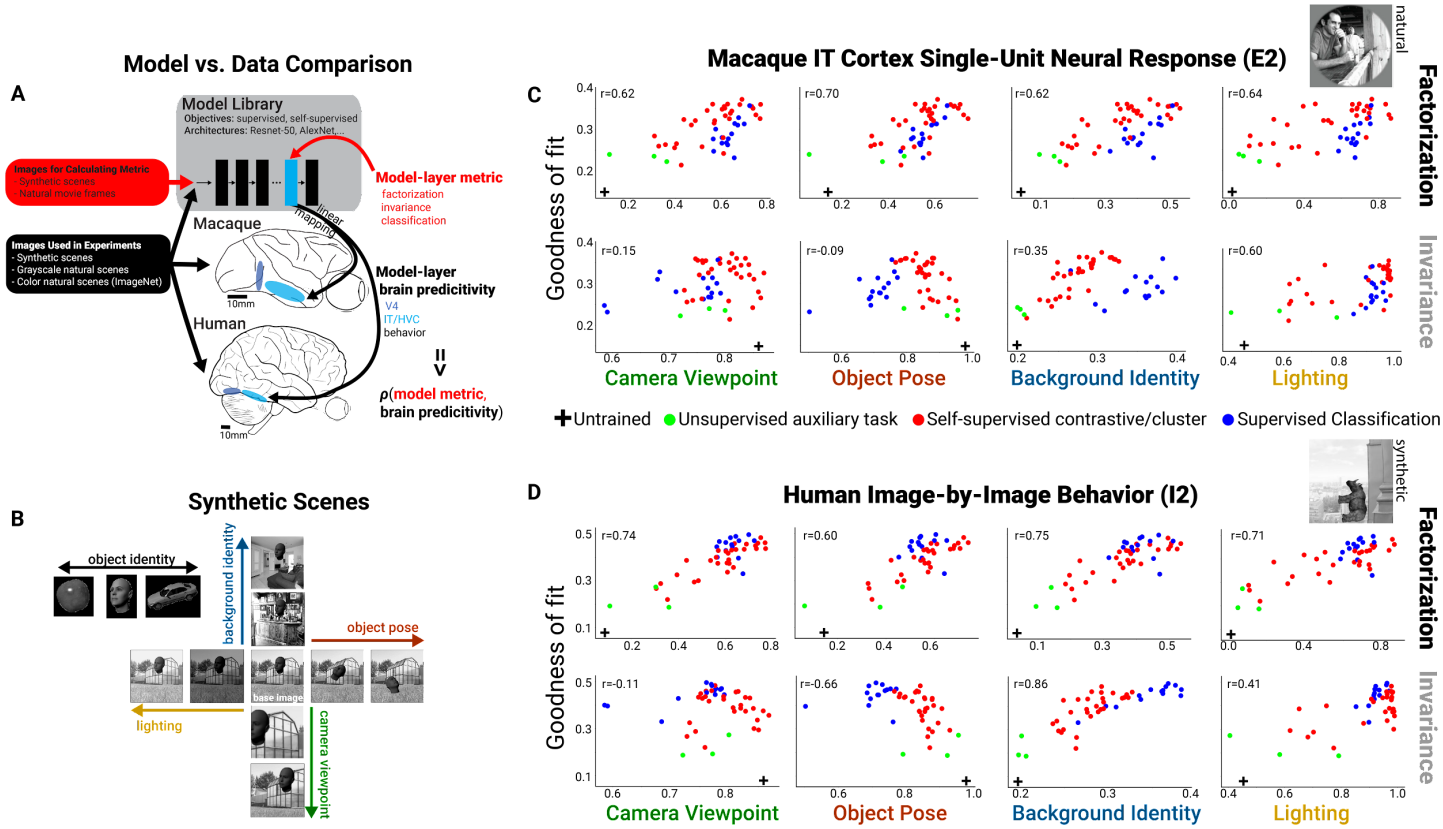


Figure 3. Measurement of factorization in DNN models and relationship to neural predictivity. (A) Schematic showing how metaanalysis on models and brain data was conducted by first computing various representational metrics on models and then measuring a model's predictive power across a variety of datasets. 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), and the across model correlation coefficient between a particular representational metric and neural predictivity for a dataset summarizes the potential importance of the metric in producing more brainlike models (see **Figure 4**). (B) For computing the representational metrics of factorization of and invariance to a scene parameter, variance in model responses was induced by individually varying each of four scene parameters (n=10 parameter levels) for each base scene (n=100 base scenes). (C) Scatter plots for example neural dataset (IT single-units, macaque E2 dataset) showing the correlation between a model's predictive power as an encoding model for IT neural data versus a model's ability to factorize or become invariant to 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 (rightward shift relative to yellow vertical dashed line). Invariance to background and lighting but not to object pose and viewpoint increased in trained models relative to the untrained control (rightward versus leftward shift relative to yellow vertical dashed line). (D) Same as (C) except for human behavior performance patterns across images (human I2 dataset).

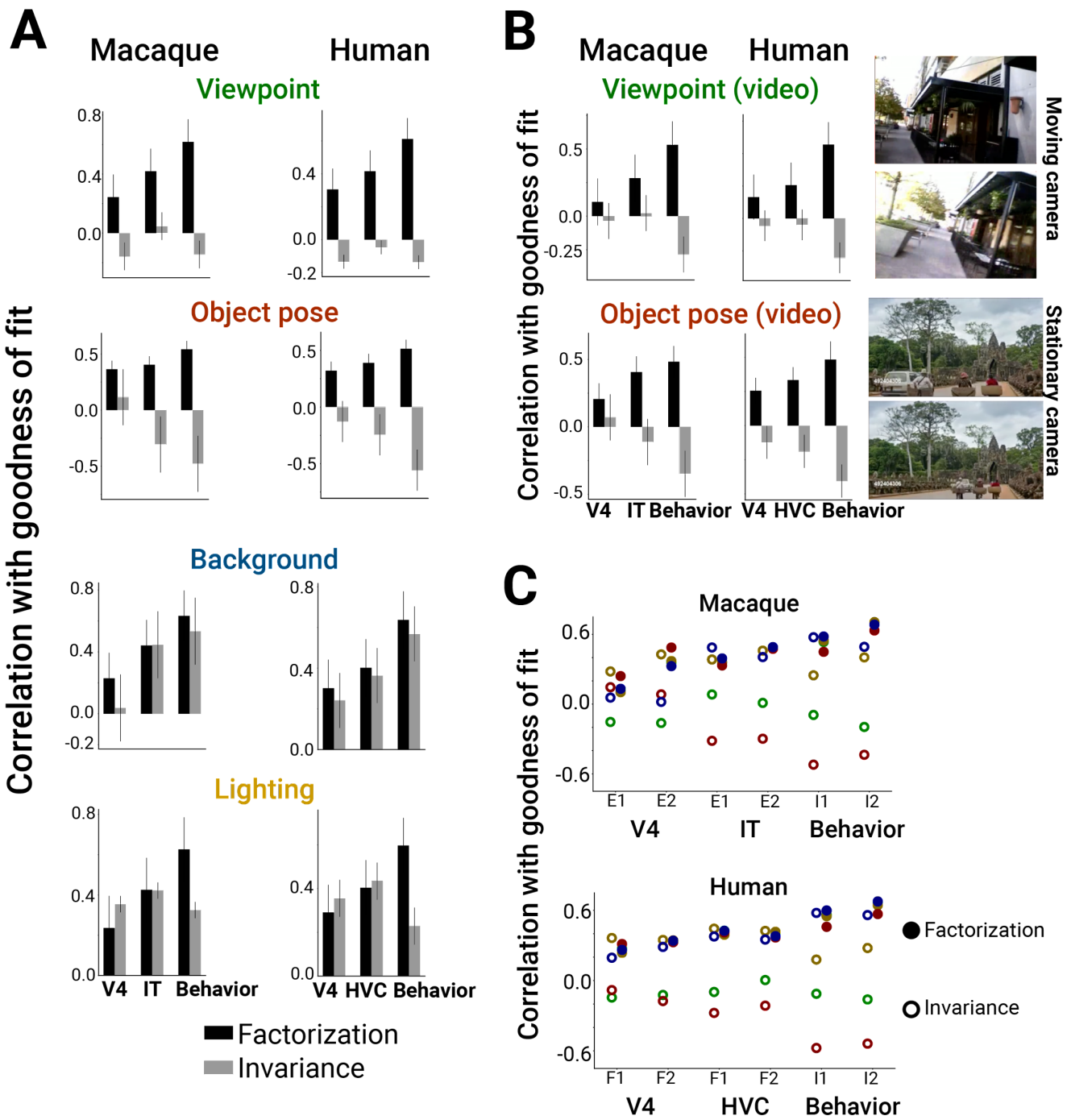


Figure 4. Scene parameter factorization correlates with more brainlike DNN models. (A) Factorization of scene parameters in model representations consistently correlated with a model being more brainlike across multiple independent datasets measuring monkey neurons, human fMRI voxels, or behavioral performance in both macaques and humans (left vs. right column)(gray bars). By contrast, increased invariance to camera viewpoint or object pose was not indicative of brainlike models (black bars). In all cases, model representational metric and neural predictivity score were computed by averaging scores across the last 5 model layers. (B) Recomputing camera viewpoint or object pose factorization from natural movie datasets that primarily contained camera or object motion, respectively (right: example movie frames; also see **Methods**), gave similar results for predicting which model representations would be more brainlike as computing these factorization scores using our synthetic images. (C) Summary of the results from (A) across datasets (x-axis) for invariance (open symbols) versus factorization (closed symbols).

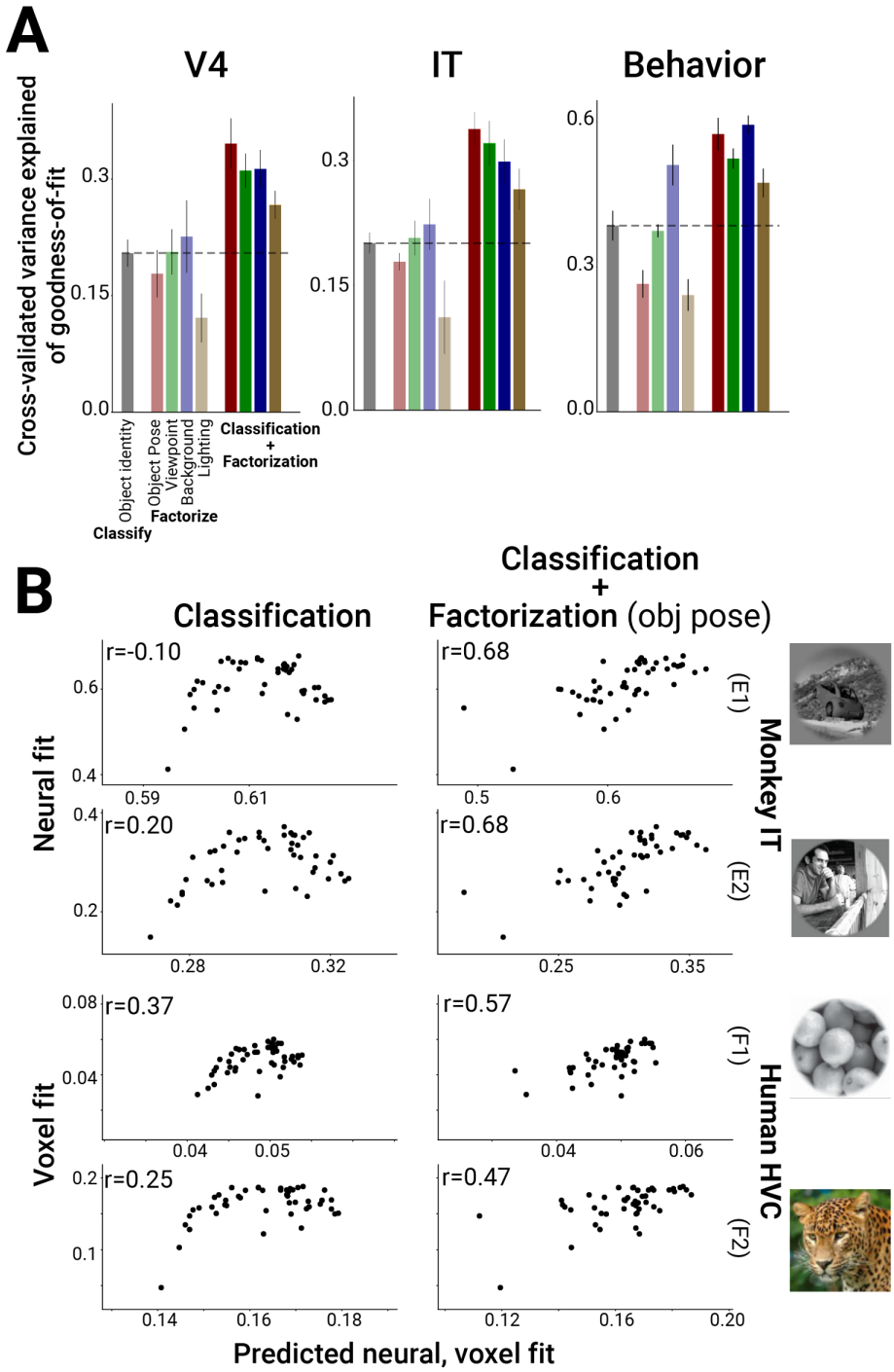


Figure 5. Scene parameter factorization correlates with more brainlike DNN models. (A) Average brain predictivity across datasets of classification (black solid bar) and factorization (colored solid bars) in a model representation. Adding factorization to classification in a regression model produced significant improvements in predicting the most brainlike models, cross-validated (across models) performance averaged across datasets (unfaded bars exceed dashed line for classification alone as a metric). (B) Example scatter plots for neural and fMRI datasets (IT multi-units & single-units, macaque E1 & E2; fMRI with grayscale & color images, human F1 & F2) 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 as a combined, predictive metric for brainlikeness of a model (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. Linsley, D. *et al.* Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. Preprint at https://doi.org/10.48550/arXiv.2306.03779 (2023).
6. Freiwald, W. A. & Tsao, D. Y. Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science* **330**, 845–851 (2010).
7. 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).
8. 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).
9. Peters, B. & Kriegeskorte, N. Capturing the objects of vision with neural networks. *Nat. Hum. Behav.* **5**, 1127–1144 (2021).
10. Johnston, W. J. & Fusi, S. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *Nat. Commun.* **14**, 1040 (2023).
11. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
12. Chung, S., Lee, D. D. & Sompolinsky, H. Classification and Geometry of General Perceptual Manifolds. *Phys. Rev. X* **8**, 031003 (2018).
13. 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).
14. 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).
15. 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).
16. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *ArXiv151203385 Cs* (2015).
17. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum Contrast for Unsupervised Visual Representation Learning. *ArXiv191105722 Cs* (2020).
18. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A Simple Framework for Contrastive Learning of Visual Representations. *ArXiv200205709 Cs Stat* (2020).
19. Tian, Y., Krishnan, D. & Isola, P. Contrastive Multiview Coding. *ArXiv190605849 Cs* (2019).
20. Doersch, C., Gupta, A. & Efros, A. A. Unsupervised Visual Representation Learning by Context

Prediction. *ArXiv150505192 Cs* (2016).
21. He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. in 2961–2969 (2017).
22. Donahue, J. & Simonyan, K. Large Scale Adversarial Representation Learning. in *Advances in Neural Information Processing Systems* vol. 32 (Curran Associates, Inc., 2019).
23. 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).
24. 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).
25. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
26. Shen, G., Horikawa, T., Majima, K. & Kamitani, Y. Deep image reconstruction from human brain activity. *PLOS Comput. Biol.* **15**, e1006633 (2019).
27. 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.
28. 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).
29. 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.
30. 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).
31. Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A. & Konkle, T. What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? 2022.03.28.485868 Preprint at https://doi.org/10.1101/2022.03.28.485868 (2023).
32. Sorscher, B., Ganguli, S. & Sompolinsky, H. Neural representational geometry underlies few-shot concept learning. *Proc. Natl. Acad. Sci.* **119**, e2200800119 (2022).
33. Hénaff, O. J. *et al.* Primary visual cortex straightens natural video trajectories. *Nat. Commun.* **12**, 5982 (2021).
34. Toosi, T. & Issa, E. Brain-like representational straightening of natural movies in robust feedforward neural networks. in *The Eleventh International Conference on Learning Representations* vol. 11 (2022).
35. Bernardi, S. *et al.* The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954-967.e21 (2020).
36. Field, D. J. What Is the Goal of Sensory Coding? *Neural Comput.* **6**, 559–601 (1994).
37. Chang, L. & Tsao, D. Y. The Code for Facial Identity in the Primate Brain. *Cell* **169**, 1013-1028.e14 (2017).
38. 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).
39. 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).
40. 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).
41. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
42. Kravitz, D. J., Saleem, K. S., Baker, C. I. & Mishkin, M. A new neural framework for visuospatial processing. *Nat. Rev. Neurosci.* **12**, 217–230 (2011).
43. 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).
44. Freiwald, W. A., Tsao, D. Y. & Livingstone, M. S. A face feature space in the macaque temporal lobe. *Nat. Neurosci.* **12**, 1187–1196 (2009).
45. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* **17**, 401–412 (2013).

SUPPLEMENT

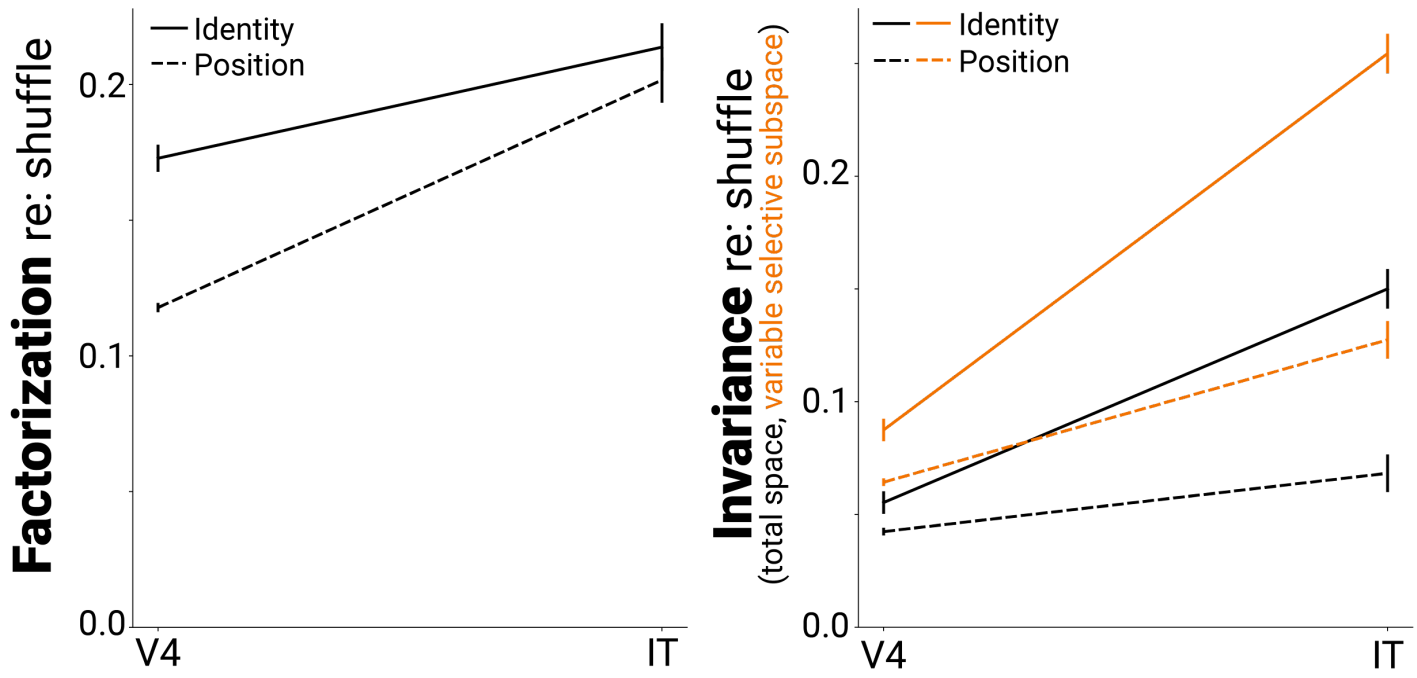
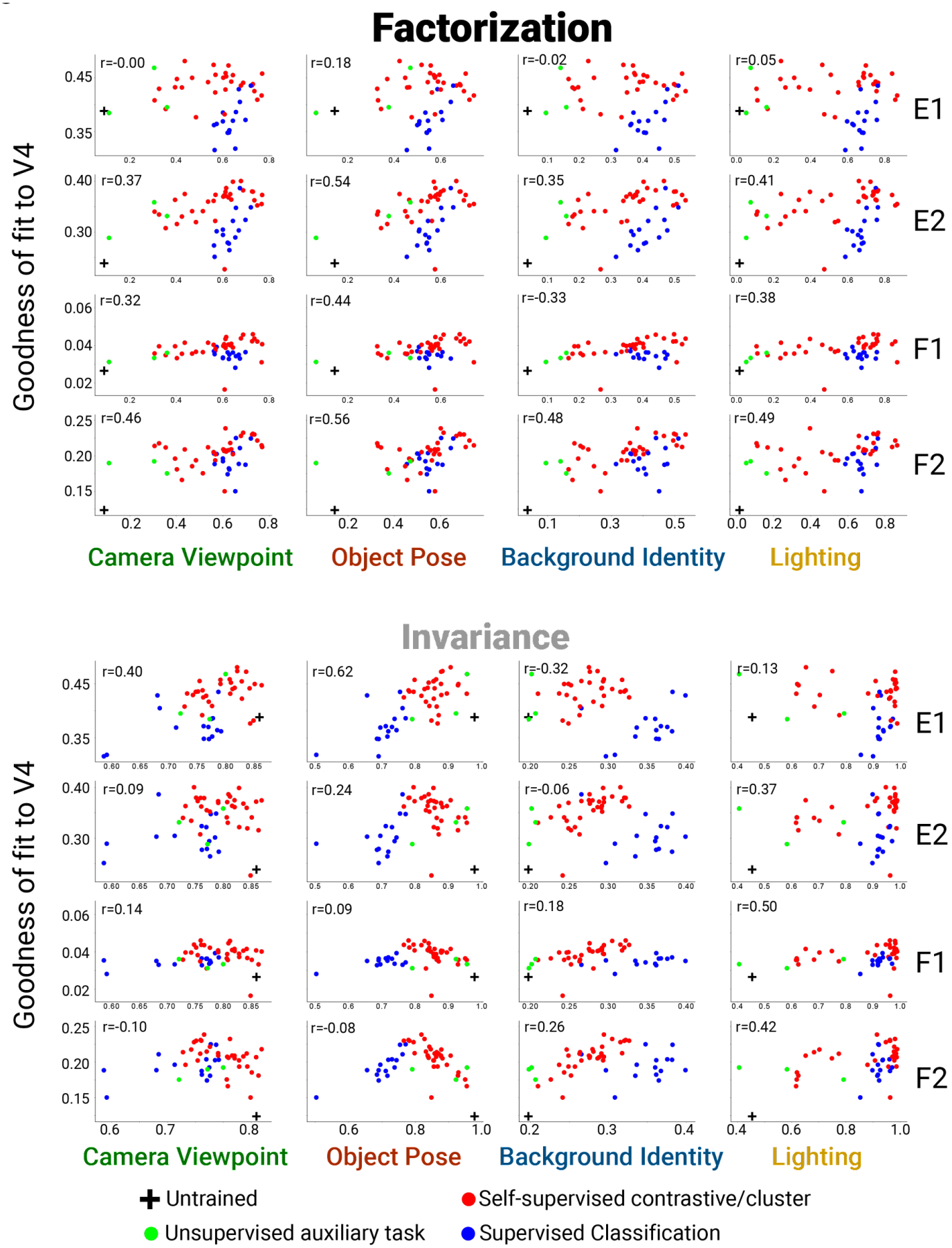
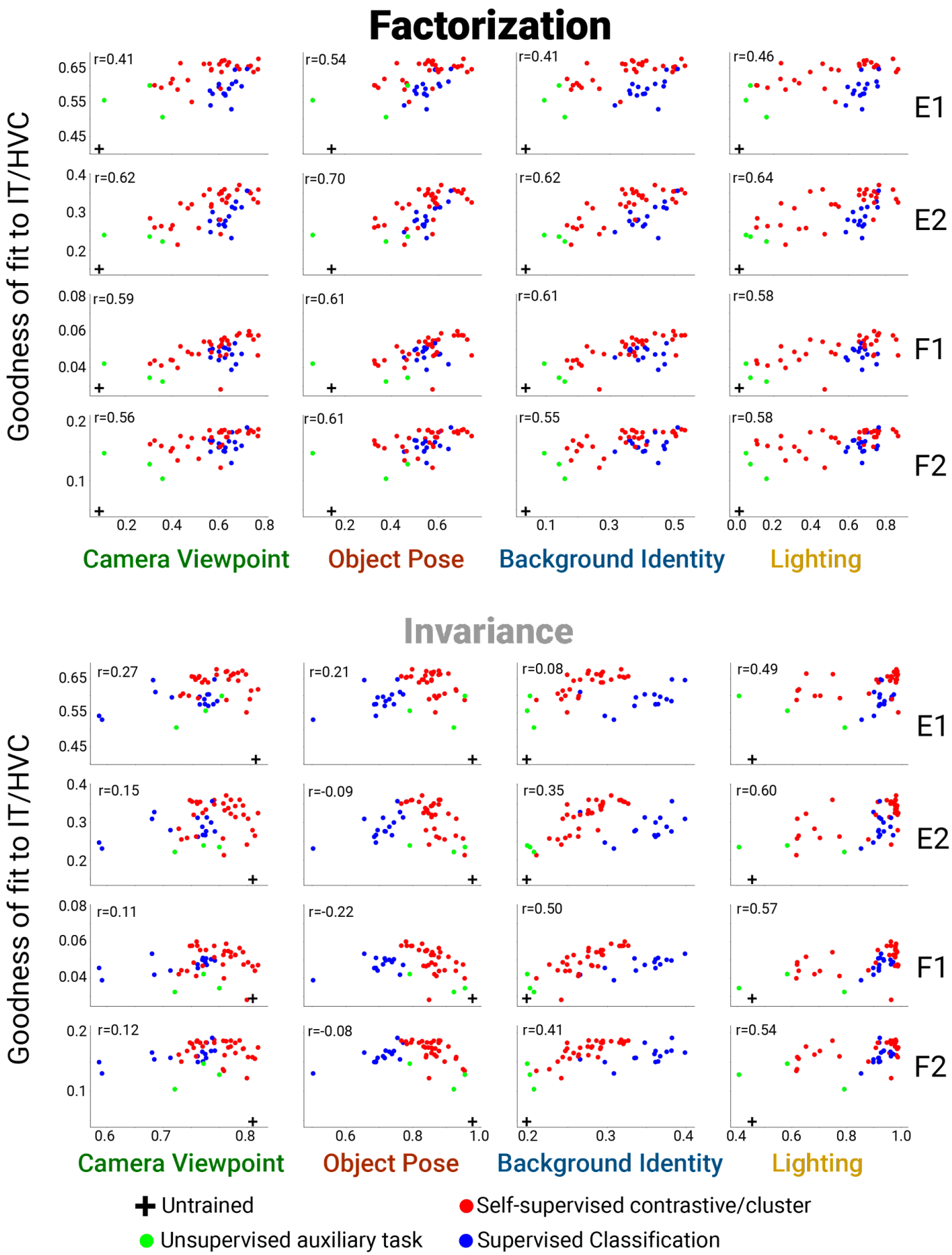


Figure S1. Factorization and invariance in V4 & IT neural data. Normalized factorization and invariance as in **Figure 2A** but after subtracting shuffle control for V4 and IT neural dataset. Shuffling the image identities of each population vector accounts for increases in factorization driven purely 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.


Factorization
Goodness of fit to V4
r=-0.00
r=0.18
r=-0.02
r=0.05
E1
r=0.37
r=0.54
r=0.35
r=0.41
E2
r=0.32
r=0.44
r=-0.33
r=0.38
F1
r=0.46
r=0.56
r=0.48
r=0.49
F2
Camera Viewpoint
Object Pose
Background Identity
Lighting
Invariance
Goodness of fit to V4
r=0.40
r=0.62
r=-0.32
r=0.13
E1
r=0.09
r=0.24
r=-0.06
r=0.37
E2
r=0.14
r=0.09
r=0.18
r=0.50
F1
r=-0.10
r=-0.08
r=0.26
r=0.42
F2
Camera Viewpoint
Object Pose
Background Identity
Lighting
Untrained
Self-supervised contrastive/cluster
Unsupervised auxiliary task
Supervised Classification


Factorization
Goodness of fit to IT/HVC
r=0.41
r=0.54
r=0.41
r=0.46
E1
r=0.62
r=0.70
r=0.62
r=0.64
E2
r=0.59
r=0.61
r=0.61
r=0.58
F1
r=0.56
r=0.61
r=0.55
r=0.58
F2
Camera Viewpoint
Object Pose
Background Identity
Lighting
Invariance
Goodness of fit to IT/HVC
r=0.27
r=0.21
r=0.08
r=0.49
E1
r=0.15
r=-0.09
r=0.35
r=0.60
E2
r=0.11
r=-0.22
r=0.50
r=0.57
F1
r=0.12
r=-0.08
r=0.41
r=0.54
F2
Camera Viewpoint
Object Pose
Background Identity
Lighting
Untrained
Self-supervised contrastive/cluster
Unsupervised auxiliary task
Supervised Classification

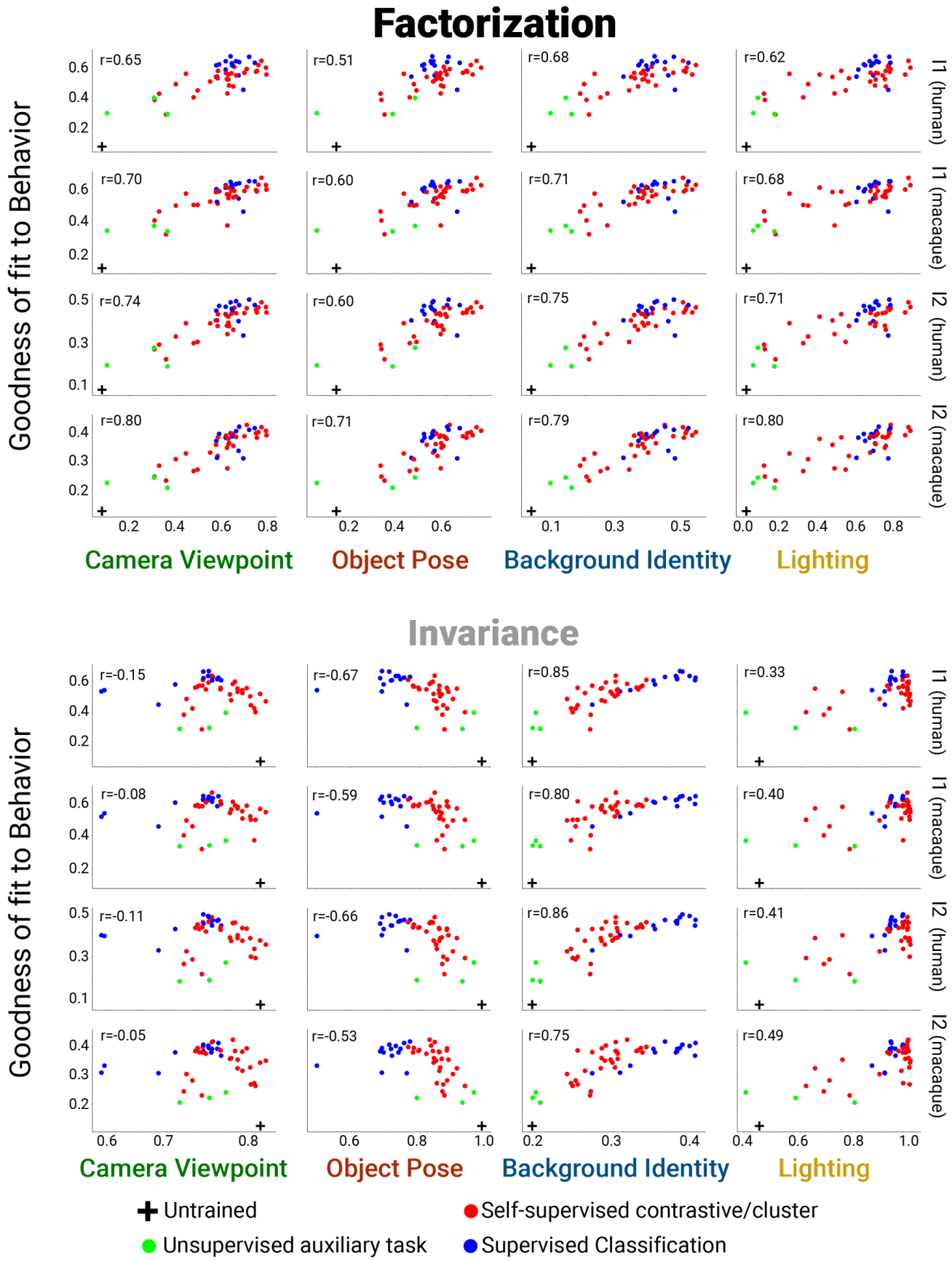

Factorization
Goodness of fit to Behavior
r=0.65
r=0.51
r=0.68
r=0.62
r=0.70
r=0.60
r=0.71
r=0.68
r=0.74
r=0.60
r=0.75
r=0.71
r=0.80
r=0.71
r=0.79
r=0.80
I1 (human)
I1 (macaque)
I2 (human)
I2 (macaque)
Camera Viewpoint
Object Pose
Background Identity
Lighting
Invariance
Goodness of fit to Behavior
r=-0.15
r=-0.67
r=0.85
r=0.33
r=-0.08
r=-0.59
r=0.80
r=0.40
r=-0.11
r=-0.66
r=0.86
r=0.41
r=-0.05
r=-0.53
r=0.75
r=0.49
I1 (human)
I1 (macaque)
I2 (human)
I2 (macaque)
Camera Viewpoint
Object Pose
Background Identity
Lighting
Untrained
Self-supervised contrastive/cluster
Unsupervised auxiliary task
Supervised Classification

Figure S2. Scatter plots for all datasets. Scatter plots as in **Figure 3C,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.

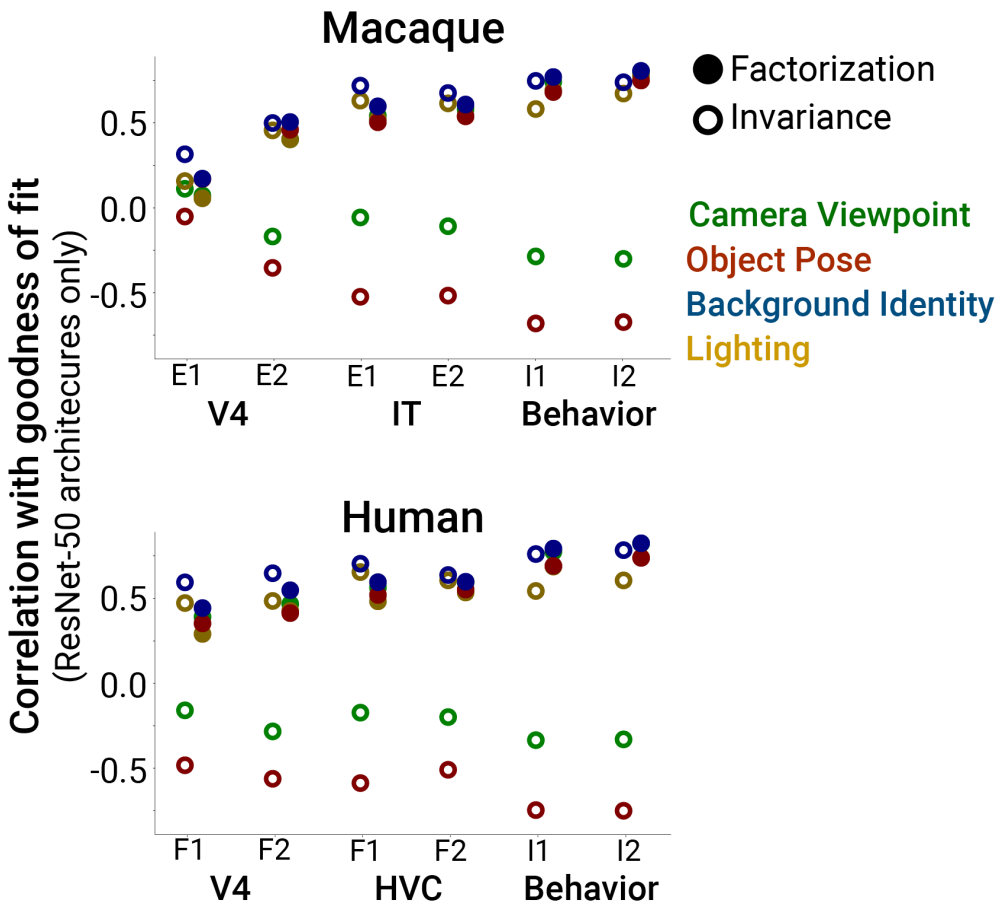


Figure S3. Predictivity of factorization and invariance restricting to ResNet-50 model architectures. Same format as **Figure 4C** except with the analyses restricted to using only models with the Resnet-50 architecture. The main finding of factorization of scene parameters in DNNs being generally positively correlated with better predictions of brain data is replicated using this architecture-matched subset of models, controlling for potential confounds from model architecture.

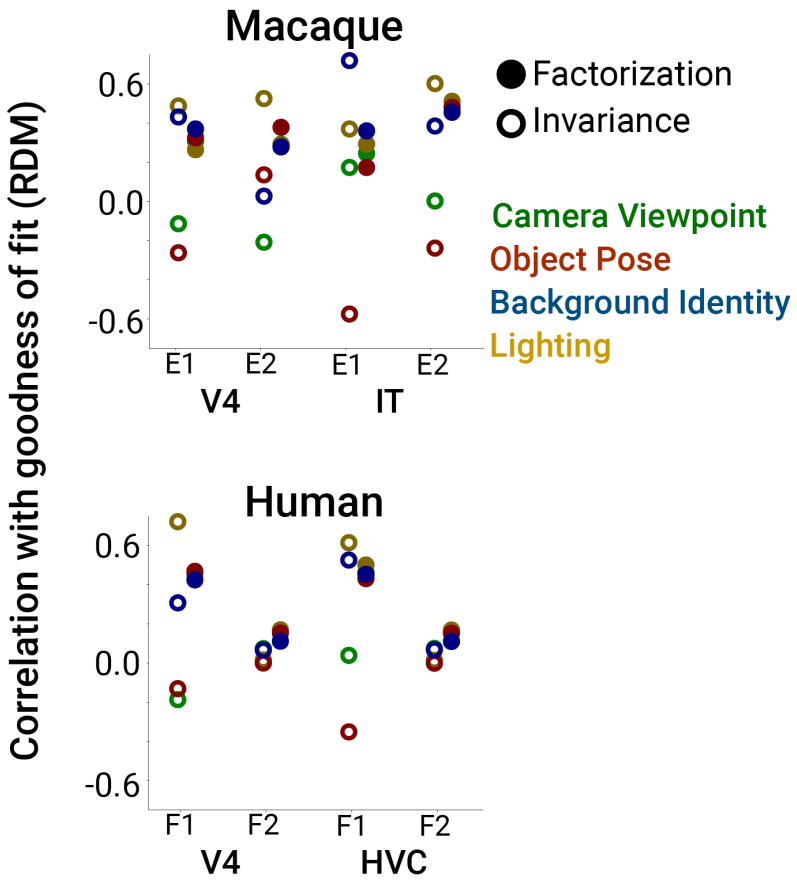


Figure S4. Predictivity of factorization and invariance for RDMs. Same format as **Figure 4C** except for predicting population representational dissimilarity matrices (RDMs) of macaque neurophysiological and human fMRI data (in the main analyses linear encoding fits of each single neuron/voxel were 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.

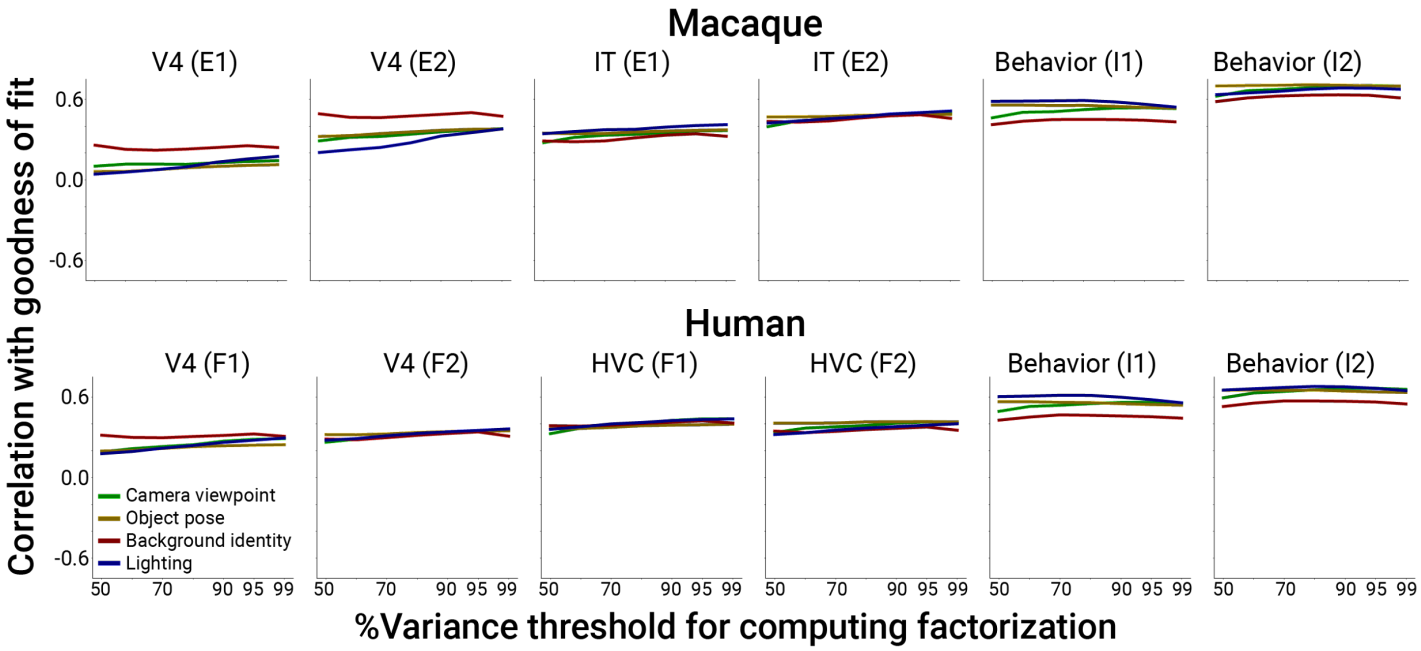


Figure S5. Effect on neural and behavioral predictivity of PCA threshold for computing factorization, Related to Figure 4. The % variance threshold used in the main text for estimating a PCA linear subspace capturing the bulk of the variance induced by all other parameters besides the parameter of interest 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).