Diverse Deep Neural Networks All Predict Human IT Well, After Training and Fitting

Katherine R. Storrs[1], Tim C. Kietzmann[2,3], Alexander Walther[3],
Johannes Mehrer[3], and Nikolaus Kriegeskorte[4]

Abstract

■ Deep neural networks (DNNs) trained on object recognition provide the best current models of high-level visual cortex. What remains unclear is how strongly experimental choices, such as network architecture, training, and fitting to brain data, contribute to the observed similarities. Here, we compare a diverse set of nine DNN architectures on their ability to explain the representational geometry of 62 object images in human inferior temporal (hIT) cortex, as measured with fMRI. We compare untrained networks to their task-trained counterparts and assess the effect of cross-validated fitting to hIT, by taking a weighted combination of the principal components of features within each layer and, subsequently, a weighted combination of layers. For each combination of training and fitting, we test all models for their correlation with the hIT representational dissimilarity matrix, using independent images and subjects. Trained models outperform untrained models (accounting for 57% more of the explainable variance), suggesting that structured visual features are important for explaining hIT. Model fitting further improves the alignment of DNN and hIT representations (by 124%), suggesting that the relative prevalence of different features in hIT does not readily emerge from the Imagenet object-recognition task used to train the networks. The same models can also explain the disparate representations in primary visual cortex (V1), where stronger weights are given to earlier layers. In each region, all architectures achieved equivalently high performance once trained and fitted. The models' shared properties—deep feedforward hierarchies of spatially restricted nonlinear filters—seem more important than their differences, when modeling human visual representations. ■

INTRODUCTION

One of the most striking achievements of the human visual system is our ability to recognize complex objects with extremely high accuracy. Recently, deep neural networks (DNNs) using feedforward hierarchies of convolutional features to process images have reached and even surpassed human category-level recognition performance (Lindsay, 2020; Kietzmann, McClure, & Kriegeskorte, 2018; He, Zhang, Ren, & Sun, 2016; Yamins & DiCarlo, 2016; Russakovsky et al., 2015). Despite being developed as computer vision tools, DNNs trained to recognize objects in images are also unsurpassed at predicting how natural images are represented in high-level ventral visual areas of the human and nonhuman primate brain (Lindsay, 2020; Xu & Vaziri-Pashkam, 2020; Bashivan, Kar, & DiCarlo, 2019; Ponce et al., 2019; Devereux, Clarke, & Tyler, 2018; Kubilius et al., 2018; Schrimpf et al., 2018; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Horikawa & Kamitani, 2017; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Yamins & DiCarlo, 2016; Güçlü & van Gerven, 2015; Agrawal, Stansbury, Malik, & Gallant, 2014; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014). There is some variability in the accuracy with which different recent DNNs can predict high-level visual representations (Xu & Vaziri-Pashkam, 2020; Zeman, Ritchie, Bracci, & de Beeck, 2020; Schrimpf et al., 2018), despite broadly high performance. It remains unclear how strongly network design choices, such as depth, architecture, task training, and subsequent model fitting to neural data, may contribute to the observed variations. There are several possible sources that can affect a DNN's high (or low) correlation with brain representations, and it is important to be able to tease these apart.

First, the architecture of a particular DNN model may cause its representations to be similar to those in the brain. For example, the architecture determines the spatial scale(s) of the image properties able to be represented within each layer. We can gain insight into the importance of such "baked in" knowledge by comparing the abilities of different architectures in their random, untrained state (Yamins et al., 2014). In the computer vision literature, deeper architectures have pushed the field toward higher object recognition accuracies (He et al., 2016; Szegedy et al., 2015; Simonyan & Zisserman, 2014), although more recently architectures have been devised that display equal or higher performance with orders of magnitude fewer parameters (Sandler, Howard,

[1]Justus Liebig University Giessen, Germany, [2]Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands, [3]MRC Cognition and Brain Sciences Unit, Cambridge, United Kingdom, [4]Columbia University

https://doi.org/10.1162/jocn_a_01755

Zhu, Zhmoginov, & Chen, 2018; Iandola et al., 2016). Does depth, across layers or across networks, help predict a model's correspondence with the brain?

Second, the task training received by a model may have led it to develop computational features that better match those in the visual cortex. It seems intuitive that the success of DNN models at predicting brain data is because in large part of the training the models receive on large data sets of natural images to do behaviorally relevant tasks such as object recognition (Yamins & DiCarlo, 2016). However, randomly weighted untrained DNN models are known to explain some variance in visual representations (Cichy et al., 2016; Güçlü & van Gerven, 2015), and at least one study reports higher performance for untrained than trained networks (Truzzi & Cusack, 2020). We can evaluate the contribution of training by comparing the ability of trained and untrained instances of the same architecture to predict brain data.

Finally, two trained models that have learned an identical set of features may nevertheless differ in their apparent similarity to brain representations if they contain different proportions of those features. For example, consider two hypothetical neural network models of low-level visual representations: Both networks have learned gabor-like oriented features, but one model contains an approximately equal number of features sensitive to each orientation, whereas the other has, through a quirk of its training data or task, dedicated most of its units to a single orientation. In this idealized example, the representations within both models span the same feature space, but the models will make very different predictions about, for example, how similar the evoked activity in cortical area V1 will be for probe stimuli containing different orientation distributions. We can evaluate to what extent the model has "the right features in the wrong proportions" by measuring how much its predictive power changes after allowing a linear reweighting of its features (Khaligh-Razavi, Henriksson, Kay, & Kriegeskorte, 2017). Many studies reporting high performance of DNNs as models of visual cortex allow linear reweighting of individual features (e.g., Bashivan et al., 2019; Ponce et al., 2019; Horikawa & Kamitani, 2017; Güçlü & van Gerven, 2015; Agrawal et al., 2014; Cadieu et al., 2014; Yamins et al., 2014), whereas others treat the representations within a layer of a network as fixed (e.g., Truzzi & Cusack, 2020; Zeman et al., 2020; Khaligh-Razavi & Kriegeskorte, 2014).

Depending on the particular research question, one may be more interested in model performance with or without fitting to brain data. For example, if we are interested in whether the distribution of visual features in the brain can be explained by its visual experience during development, we may prefer to compare models trained on different image diets in their unfitted states. On the other hand, if we are interested in the level of complexity encoded by neurons in a certain brain region, we may prefer to compare different, progressively complex, layers of a model after allowing the weighting of features in each layer to best fit the brain data.

Here, we systematically evaluate the contributions of task training and feature reweighting to the ability of models to predict representations of objects in the ventral stream, across nine state-of-the-art computer vision DNNs. We use representational similarity analysis (RSA) to evaluate the correspondence between fMRI brain activity patterns elicited by viewing object images and representations of those images in networks. We compare a diverse set of DNN models, varying widely in depth (8–201 layers; 25–825 processing steps), in terms of their ability to explain the representational geometry in human inferior temporal (hIT) cortex. Each model is tested both in an untrained randomly initialized state and after object-categorization training. We use principal component reweighting of features within each layer, and reweighting of layers, to best predict the hIT representation. After principal component analysis (PCA) and hIT fitting, each model is then tested on its ability to predict the hIT representational dissimilarity matrix (RDM) for an independent set of images in an independent set of subjects. This analysis ensures that the evaluation is not biased by overfitting to either images or subjects.

METHODS

Stimuli

Both human participants and neural networks were shown the same set of 62 colored images depicting faces, objects, and places, segmented and presented on a gray background of 427 × 427 pixels (see Figure 1A). The image set was constructed to include a balance of animate (faces and bodies) and inanimate (objects and scenes) stimuli, with animate stimuli further divided into human and animal faces/bodies, and inanimate stimuli divided into man-made and natural objects/scenes. Of the 20 human face images, 14 (seven male, seven female) were closely matched for low-level image statistics, depicting faces in a 30° semiprofile view with matched lighting and matched color-histogram profiles. The remaining face and nonface images contained greater pose and image variation and were a subset of those previously used in Kriegeskorte, Mur, Ruff, et al. (2008) and Kiani, Esteky, Mirpour, and Tanaka (2007).

Human fMRI Procedures

DNNs were tested against a preexisting human fMRI data set (Walther, Diedrichsen, Mur, Khaligh-Razavi, & Kriegeskorte, 2016; Walther, 2015), described below.

Participants

Participants were 24 healthy adult volunteers (15 female) naive to the goals of the experiment, with normal or corrected-

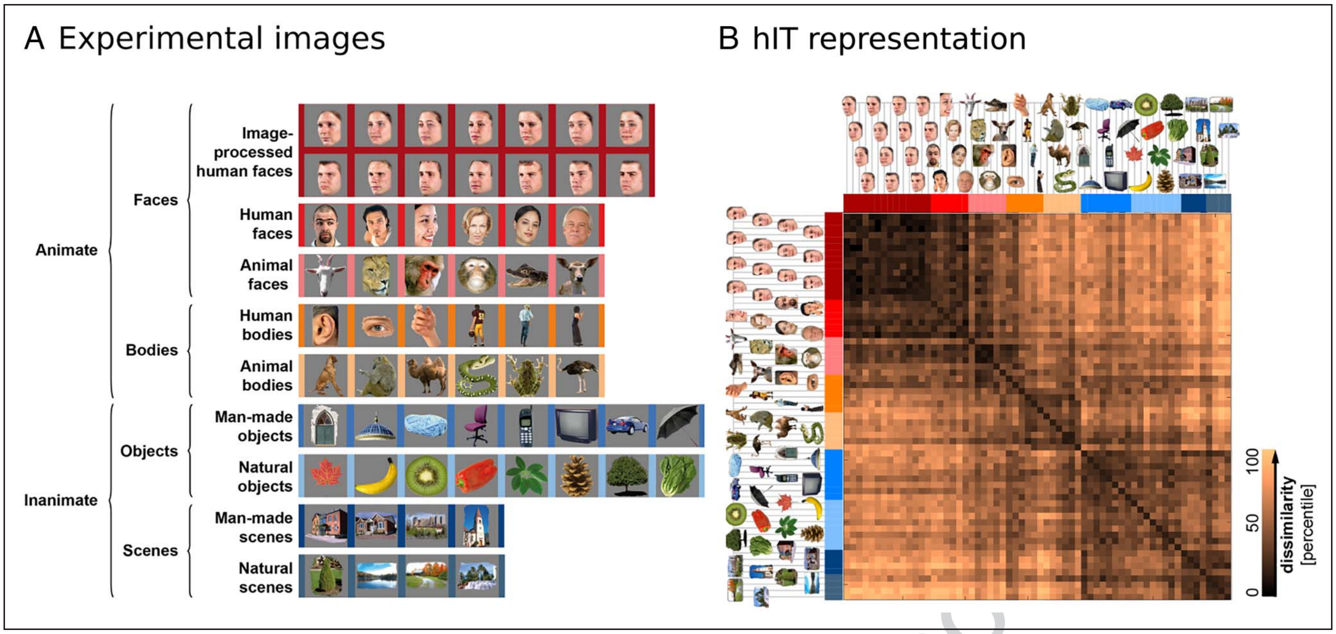

Figure 1. Human fMRI data set. Sixty-two stimulus images (A) were shown to 24 human participants in a rapid event-related fMRI experiment. An RDM was constructed for each participant from the cross-validated Mahalanobis distances between the multivoxel activation patterns elicited by each image in inferior temporal cortex (hIT). The average of all individual RDMs is shown in B, with dissimilarity between the hIT representation of each pair of images expressed as a percentile for visualization purposes.

to-normal vision. Participants gave informed consent, and the study was approved by the MRC Cognition and Brain Sciences Ethical Review Panel and conducted in accordance with the Declaration of Helsinki.

Image Familiarization

Before the main fMRI experiment, participants were familiarized with the stimuli and task outside the scanner by completing five runs of the 1-back task (i.e., five complete cycles through the stimulus set). Participants were also instructed to pay attention to the 62 images shown and try to commit them to memory. After completing the second of two fMRI sessions, their recall of the images was tested. One hundred twenty-eight isolated objects on gray backgrounds were shown to participants, of which 62 were the experimental stimuli. During this recall block, images were shown in a random order, with each image repeated twice, and participants were asked to identify which they had seen during the fMRI sessions. On average, recognition performance was excellent (mean accuracy = 92%; $d' = 3.58$, ±0.26 *SEM*), indicating that participants had attended to the experimental stimuli.

fMRI Task

During the fMRI scans, participants engaged in a 1-back task in which they were instructed to press a button if the present image was a repeat of that shown on the immediately previous trial. While performing the task, participants were asked to keep fixation on a central fixation cross. On each trial, one image was presented in the center of a gray screen, subtending 7° of visual angle. Images were shown for 500 msec with a trial-onset asynchrony of 3 sec. During each run, 56 of the stimuli were each presented once. The other six were presented twice, appearing as repeats in the 1-back task. Stimulus order was randomized within each run for each participant. The stimulus sequence also included 30 baseline trials with no image stimulus, comprising 5 blank trials at the beginning of each run, 5 at the end, and 20 randomly interspersed within stimulus trials.

MRI Measurements

Each participant undertook two scanning sessions on separate days, each consisting of 12 functional runs lasting approximately 5 min. Functional images were acquired on a Siemens Trim-Trio 3-T MRI scanner with a 32-channel head coil. For each functional run, we recorded 135 volumes containing 35 slices, each using a 2-D EPI sequence (repetition time = 2.18 sec, echo time = 30 msec, flip angle = 78°, voxel size: 2 mm *isotropic*, interleaved slice acquisition, GRAPPA acceleration factor: 2). We also acquired a high-resolution (1-mm *isotropic*) T1-weighted anatomical image in each session, using a Siemens magnetization prepared rapid gradient echo sequence.

Data Preprocessing

Image preprocessing was performed using SPM8 (www.fil.ion.ucl.ac.uk/spm/). For each participant, images from both sessions were jointly processed, after discarding the first two volumes of each run to prevent T1 saturation effects in the baseline of the regression coefficients. Preprocessing consisted of the following steps, in order: slice-scan-time correction, 3-D head motion correction by aligning to the first EPI of the first run of the first session, reslicing, and coregistration of the high-resolution anatomical images to the session mean EPI. No smoothing was performed.

ROI Definition

For each participant and hemisphere, early visual regions V1, V2, and V3 were defined using a cortical surface template projected into each subject's native volume (Benson, Butt, Brainard, & Aguirre, 2014). To define hIT, visually responsive voxels were first identified based on an independent functional localizer scan in which responses to 432 images of faces, places, objects, and scrambled objects were contrasted against baseline. An anatomical mask was then used to select the subset of those voxels that lay within the hIT region, using the FreeSurfer package (surfer.nmr.mgh.harvard.edu/), and any voxels belonging to V1, V2, or V3 were excluded.

Estimating Stimulus Response Patterns

Response patterns were calculated as in Walther, Nili, et al. (2016), using multivariate noise normalization to improve the reliability of dissimilarities subsequently measured between response patterns. Beta response weights were estimated using general linear model (GLM) with ordinary least squares. Time course data of each run were modeled using 62 stimulus predictors, separately for each subject and session. Six additional 1-back predictors were included to model repeated image stimuli in the 1-back task. For each run, we included six head motion predictors (3-D translation and rotation coordinates) and one intercept.

Before fitting, the time course data and the design matrix were filtered to remove low-frequency trends. Because cross-validated dissimilarity estimates, as used here to derive RDMs (details below), require two independent estimates of stimulus response vectors, two sets of GLMs were fitted for each session and subject (Walther, Nili, et al., 2016). For each of the 12 imaging runs, one GLM was fit on the data of the individual run, whereas another GLM was jointly fit on the data of the remaining runs, thereby keeping the GLM estimates independent. For the latter GLM, data from the included runs were concatenated for each stimulus (i.e., 11 entries per predictor in the design matrix), which stabilizes the regression weights. Nuisance regressors were modeled separately for each concatenated run. Finally, the $62 \times P$ stimulus response beta estimates from each GLM were normalized for multivariate spatial noise by the $P \times P$ variance–covariance matrix estimated from the time-course residuals, where P is the number of voxels.

Estimating Representational Geometry

To investigate visual representations in the brain, we used RSA (Nili et al., 2014; Kriegeskorte, Mur, & Bandettini, 2008). RSA characterizes the underlying representations of a given system via RDMs, consisting of dissimilarity estimates for all pairs of stimuli. The set of all pairwise distances describes the geometry of response patterns in high-dimensional activation space (Kriegeskorte & Diedrichsen, 2016) and can be used to compare representations across different systems (here, hIT and DNNs).

For each subject, an hIT RDM was computed by taking the cross-validated Mahalanobis distance (also known as the "crossnobis distance") between the patterns elicited in hIT by each pair of images (Walther, Nili, et al., 2016). We calculated leave-one-run-out cross-validated distances using the two sets of response pattern estimates derived from GLMs fitted to nonoverlapping imaging runs. Separate RDMs were derived from the left and right hemispheres and then averaged to create a single hIT RDM for each participant, of size 62 × 62 (1891 unique pairwise image dissimilarities). We repeated the same procedure using the response patterns elicited in primary visual cortex (V1) to derive V1 RDMs, which are the subject of later analyses.

DNN Models

Network Architectures and Training

We investigated nine deep convolutional neural network architectures representing various states of the art from the computer vision literature over the past 8 years (see Table 1). The architectures varied widely in the number of unique processing steps they involved (e.g., convolution, nonlinearity, pooling, batch normalization, concatenation), from 25 sequential processing steps (Alexnet; Krizhevsky, Sutskever, & Hinton, 2012) to 825 steps with branching nodes and skipping connections (Inception-Resnet-v2; Szegedy, Ioffe, Vanhoucke, & Alemi, 2017). Their sizes varied from 1.24 million parameters (Squeezenet; Iandola et al., 2016) to 138 million parameters (VGG-16; Simonyan & Zisserman, 2014).

All models were implemented in the Deep Learning Toolbox for MATLAB 2019b and were pretrained by their original developers on the Imagenet Large-Scale Visual Recognition Competition (ILSVRC; Russakovsky et al., 2015) data set. The ILSVRC training set consists of 1.2 million labeled images, and the networks' task is to categorize images as belonging to one of 1000 possible object and animal categories. All networks had near-identical training data and training tasks—slight differences were because of updates to the ILSVRC training set and image categories over the years and to differences in training strategies

Table 1. Details of the Nine Computer Vision DNNs Evaluated

Name	*Reference*	*Imagenet Top-5 Error*	*Depth (Layers)*	*Number of Parameters (Millions)*	*Layers Selected for Evaluation*
Alexnet	Krizhevsky et al. (2012)	20.9%	8	61.0	8 key layers: outputs of each convolutional or fully connected layer (after ReLU nonlinearity)
VGG-16	Simonyan and Zisserman (2014)	9.6%	16	138.0	16 key layers: outputs of each convolutional or fully connected layer (after ReLU nonlinearity)
Googlenet	Szegedy et al. (2015)	10.5%	22	7.0	13 key layers: outputs of first three convolutional layers, outputs of each branching "inception" module (after concatenation), and output of final fully connected layer
Resnet-18	He et al. (2016)	10.9%	18	11.7	10 key layers: output of first convolutional layer, outputs of each "residual block" (after addition), and output of final fully connected layer
Resnet-50	He et al. (2016)	7.1%	50	25.6	20 key layers: output of first convolutional layer, outputs of each "residual block" (after addition), and output of final fully connected layer
Squeezenet	Iandola et al. (2016)	19.4%	18	1.24	11 key layers: output of first convolutional layer, outputs of each "fire" module (after depth concatenation), and outputs of final convolutional and pooling layers
Densenet-201	Huang et al. (2017)	6.4%	201	20.0	103 key layers: output of first convolutional layer, outputs of each densely connected block (after depth concatenation), and output of final fully connected layer
Inception-Resnet-v2	Szegedy et al. (2017)	4.9%	164	55.9	50 key layers: outputs of first five convolutional layers, outputs of each "Inception-ResNet" module (after addition and ReLU), and output of final fully connected layer
Mobilenet-v2	Sandler et al. (2018)	9.7%	53	3.5	20 key layers: output of first convolutional layer, output of each residual or downsizing block (after batch normalization), and output of final fully connected layer

"Imagenet Top-5 Error" records the percentage of times the correct object label was not in the network's top five guesses for the public test set of the Imagenet 1000-way object classification database, used in the ILSVRC. Error rates are as reported for a single model and single crop of each test image for a PyTorch implementation of the models (from pytorch.org/docs/stable/torchvision/models.html), with the exception of InceptionResnet-v2, which is the single-model error rate reported in the original publication. Note that these error rates may differ from the model's ILSVRC result, because competition results are calculated on a confidential test image set and because competition submissions often use ensembles of networks and/or data augmentation at test time. The column "Layers Selected for Evaluation" briefly describes the criteria we used to select key processing steps within each network, to evaluate their match to neural representations in hIT cortex.

adopted by different research groups, such as which methods of data augmentation were used. Object classification error of the networks, as quantified by Top-5 error rate on the ILSVRC validation set, ranged from 20.9% error rate (Alexnet; Krizhevsky et al., 2012) to 4.9% error rate (Inception-Resnet-v2; Szegedy et al., 2017). This measure captures the proportion of test images for which the correct object category was not one of the network's top five guesses. Human Top-5 classification error rate for this data set is thought to be around 5.1% (Russakovsky et al., 2015).

In addition to analyzing the trained networks, we also created untrained randomly weighted versions of the same architectures, by replacing all weights and biases in each network layer by random numbers drawn from a Gaussian distribution with the same mean and standard deviation as the weights or biases in the trained network at the same layer. Analysis procedures were identical for trained and untrained networks.

Image Preprocessing

Before being input to neural networks, the 62 stimuli used in the fMRI experiment were resampled to the native input size of each network architecture (either 224 × 224, 227 × 227, or 299 × 299 pixels), and the mean red, green, blue pixel value of each network's training images was subtracted.

Layer Activations

For each architecture, we selected a subset of key layers to analyze, generally consisting of the outputs of convolutional or fully connected processing steps, after applying a nonlinearity. For architectures made up of "modules" or "blocks" of processing steps within which parallel branches of processing occurred, or over which skipping connections passed, we took as key layers the outputs of each submodule after all its inputs had been concatenated or added. Table 1 provides further detail on the criteria for key layer selection within each architecture. Network activation patterns in response to the 62 stimulus images were recorded for each of these key layers.

Comparing Human and DNN Image Representations

Ecologically Driven Principal Component Selection

The number of features in a layer varied by three orders of magnitude across layers and networks, from over 1 million unit activations in the early layers of some architectures, to 1000 in the output layers. Before reweighting features, we therefore used PCA to match the dimensionality of the representations across all layers and to bring the number of parameters to be fitted into a feasible range. For each layer of each network, a PCA was performed to extract the first 100 orthogonal dimensions explaining the most variance in the responses to a set of independent ecologically representative images. For this, a set of 3020 images was drawn from "Ecoset" (Mehrer, Spoerer, Jones, Kriegeskorte, & Kietzmann, 2021), a large-scale vision data set that mirrors the most common concrete nouns in the English language that describe basic level categories (such as dog, cat, table, etc.). Ecoset thereby represents categories that describe physical things in the world (rather than abstract concepts), which are of significance to humans. The image set used to calculate the PCA had no overlap with the experimental test set and were natural photographs with backgrounds (whereas test stimuli were isolated object images on gray backgrounds).

For each of the first 100 principal components, a principal component RDM (PC-RDM) was computed by taking the Euclidean distance between unit activation patterns elicited by each image pair, after projecting those activation patterns onto each principal component in turn (see Figure 2A). These 100 PC-RDMs, extracted for each layer of each model, were then fitted to human IT RDMs by cross-validation over both participants and images (see Figure 2B). Taking a weighted combination of RDMs derived from model features captures the representational geometry of a model version in which the strength or prevalence of those different features has been adjusted (Khaligh-Razavi & Kriegeskorte, 2014).

Cross-Validated Reweighting

We performed a two-stage reweighting procedure to fit model representations to human IT representations both within and across layers of a network and evaluated these fitted representations on held-out subjects and stimuli (Figure 2B). The first-level (within-layer) fitting linearly combined the 100 PC-RDMs within each layer of a model to create a single hIT-fitted RDM for that layer. The second-level fitting then linearly combined these single-layer RDMs across all layers of a network, to create a whole-network hIT-fitted RDM for that model. The weights for both levels of fitting were estimated within the same cross-validation procedure, ensuring that all fitting was performed on an independent set of subjects and stimuli from that used in model evaluation. As a result, the reported performance of the DNN models is based on their predictions of data for previously unseen subjects and images. In addition to model fitting, we calculated the performance of full-dimensionality unfitted versions of each layer, as well as the lower and upper bounds of the noise ceiling, all within the same cross-validation folds to ensure that all estimates were directly comparable. The noise ceiling is a measure of how well data from individual participants can predict the data from other participants. It provides upper and lower bounds on the expected performance of the true data-generating model, given the interparticipant variability in the data (Nili et al., 2014). Its upper bound indicates

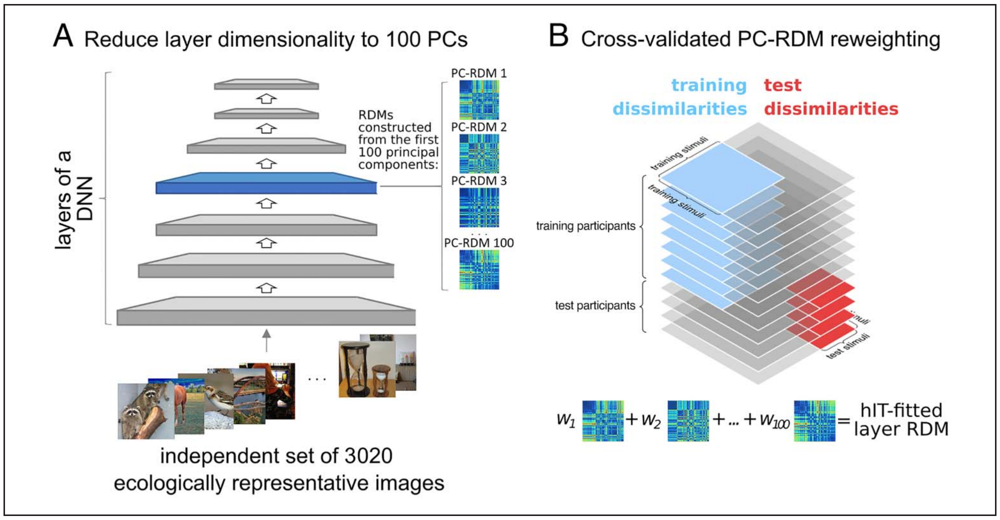
A Reduce layer dimensionality to 100 PCs
layers of a DNN
RDMs constructed from the first 100 principal components:
PC-RDM 1
PC-RDM 2
PC-RDM 3
. . .
PC-RDM 100
. . .
independent set of 3020 ecologically representative images
B Cross-validated PC-RDM reweighting
training dissimilarities
test dissimilarities
training stimuli
training stimuli
training participants
test participants
test stimuli
test stimuli
w_1 + w_2 + ... + w_{100} = hIT-fitted layer RDM

Figure 2. Ecologically driven dimensionality reduction and component reweighting. An independent data set (A) of 3020 images derived from object categories that are important to humans was constructed by sampling uniformly from the ecoset data set (Mehrer et al., 2021). We recorded the activations elicited by these images in each unit of each layer of each network. On the basis of the data obtained from all units of a given layer, we then ran a PCA to identify the first 100 orthogonal components that explained the most variance in the layer's response to these ecologically representative images. By projecting activation vectors onto each of the 100 PCs in turn, we constructed 100 PC-RDMs for the 62 experimental stimulus images, within each network layer, using Euclidean distance. (B) The 100 PC-RDMs in each layer were linearly combined using a cross-validated reweighting procedure to best predict the hIT fMRI representation ("first-level fitting"). All weights were fitted on data from both separate subjects and separate image stimuli from those on which they were tested. On each cross-validation fold, one nonnegative weight was assigned to each PC-RDM via least-squares fitting. Within the same cross-validation folds, an aggregate prediction for the whole network was then calculated by linearly weighting the per-layer fitted RDMs ("second-level fitting"; not shown).

the maximum possible performance for any model, within the given data set and analysis.

The full sequence of steps, run for all models within a single bootstrap resampling procedure, is as follows:

1. For each of 1000 bootstrap samples, resample both stimuli and subjects with replacement:

   a. For 200 cross-validation folds:

      i. Randomly assign 12 unique stimuli present in this bootstrap sample to be test stimuli. The test set always consists of data from exactly 12 unique images and typically contains repetitions of some. The training set contains the remaining images in this bootstrap sample and may include repetitions. Data from the same image never appear in both training and test sets.

      ii. Randomly assign five participants present in this bootstrap sample to be test participants. As with stimuli, the test set consists of data from exactly five individuals and may contain repetitions. The training set consists of data from the remaining individuals present in this bootstrap and may contain repetitions. The same subject never appears in both training and test sets.

      iii. Once data are separated into training and test sets:

         1. Create human target RDM: Average the data RDMs across training participants for the training stimuli to create a target RDM to which models will be fit.

         2. First-level (within-layer) hIT fitting: For each layer of each model, divide each of that layer's 100 PC-RDMs into separate training-stimuli and test-stimuli RDMs. Use nonnegative least squares regression to find the 100 weights that linearly combine the training-stimuli PC-RDMs to best predict the human target RDM. This fitted training-stimuli RDM will be used in the next step, for across-layer fitting. In addition, create a reweighted RDM capturing this layer's prediction for the test images, by combining the test-stimuli PC-RDMs using the weights obtained via the training-stimuli fit. This predicted test-stimuli RDM will be used to evaluate the performance of this layer. Here, we also compute an unfitted RDM for each layer based on the distances between test images in the full original feature space of each layer, with no dimensionality reduction or reweighting applied.

         3. Second-level (across-layer) hIT fitting: For each model, take the set of single-layer fitted training-stimuli RDMs calculated at the previous step. Use nonnegative least squares regression to linearly weight these first-level fitted RDMs to best predict the human target RDM. Construct a predicted RDM by using the resulting layer weights to combine the first-level-fitted test-stimuli RDMs from all layers. This whole-model hIT-fitted predicted RDM aggregates representations across all layers of a network, while allowing the influence of features to be linearly scaled both within and across layers to better match human representations. We also computed a whole-network predicted RDM using only second-level fitting, by applying the same across-layer reweighting to the unfitted per-layer RDMs described in the previous step.

This “unfitted” whole-model predicted RDM treats each layer as a fixed representation but allows the influence of each layer on the whole-network predicted RDM to be linearly scaled to better match human data.

4. Model evaluation: Evaluate the performance of each of the model predictions as the average Spearman correlation between the predicted RDM and each individual test participant’s RDM for test stimuli. Within a cross-validation fold, estimate the performance of the first-level fitted RDMs for each layer of each model, the unfitted per-layer RDMs for each layer of each model, and the two second-level fitted RDMs for each model, one of which combines the first-level fitted per-layer RDMs and the other of which combines the unfitted per-layer RDMs.
5. Noise ceiling calculation: Calculate the upper and lower bounds of the noise ceiling (Nili et al., 2014) by taking the correlation between each test participant’s test-stimuli RDM and the mean test-stimuli RDM averaged over either all participants (upper bound) or the training participants (lower bound). By calculating the noise ceiling within the model-reweighting cross-validation folds, we ensure that the lower bound estimates are correct for both fitted and unfitted models (Storrs, Khaligh-Razavi, & Kriegeskorte, 2020).

b. At the end of the 200 cross-validation folds, average the model performances (first and second levels) as well as noise ceiling estimates, to create a single estimate of each, for a given bootstrap sample.

RESULTS

We evaluated how well the representations of object images in each of nine diverse DNN architectures could predict those in hIT cortex. We analyzed performance both for each layer individually and when aggregating across layers in a network. We tested trained and untrained versions of each architecture as well as the effects of allowing linear reweighting of the principal feature components within each layer.

Object Recognition Training Modestly Improves Representational Correspondence with Human IT

First, we compared the hIT correlation of every layer in trained and untrained versions of each model (Figure 3). We found that training improved representational similarity to hIT, but by a perhaps surprisingly small degree. For each layer and model, we tested whether the distribution of differences in bootstrapped performance between the trained and untrained models contained zero, using a one-tailed test, with an alpha level of .05, uncorrected for multiple comparisons. Even using this liberal criterion, only five of the nine models contained layers in which trained performance was better than untrained, and Mobilenet was the only model to show significantly higher performance in most layers after training (see blue asterisks in Figure 3).

Notably, however, whereas untrained models showed similar hIT correlation across all their layers, the performance of trained models peaked for processing steps about one half to three fourths of the way from network input to output. This echoes previous findings of graded similarity between DNN representations and the human ventral stream (Xu & Vaziri-Pashkam, 2020; Zeman et al., 2020; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014) and suggests that the learning of natural image features of the correct complexity is responsible for the superior performance of later layers, rather than inherent architectural properties such as the spatial scale of the representations. Most models show a sharp decline in hIT match toward the final output layers, likely because their training target, a sparse vector indicating which of 1000 objects possible within the Imagenet ILSVRC challenge data set is present in an image, forces late layers not to represent degrees of similarity between images. There was substantial variation among models in how well the best layer correlated with the representation in hIT, but no model explained all of the explainable variance in the human data. A model could be considered to explain all explainable variance in a data set if it predicts human data as well as individual human participants can predict the data of other participants, as quantified by the lower bound of the noise ceiling (Storrs et al., 2020; Nili et al., 2014). For each layer of each model, we tested whether the bootstrap distribution of differences between the lower bound of the noise ceiling and the layer performance contained zero (one-tailed, $\alpha =$ .05, uncorrected for multiple comparisons). For all layers of all models, the lower bound of the noise ceiling was significantly higher than model performance. Although object-recognition training improves performance, the distribution of visual features learned by task-trained DNNs does not fully mirror those in human IT.

Reweighting Features within Trained Layers Dramatically Improves Correspondence with hIT

Next, we compared the hIT correlation of each layer of the trained networks, both in its full unfitted state and after model fitting, that is, reducing dimensionality via PCA on natural images and reweighting the principal components to better predict human representations using held-out images and participants (see Methods). Reducing and reweighting the feature space improved the correlation of a layer’s representations with that in hIT for virtually all layers of all networks (Figure 4). Five of the nine models contained at least one layer in which hIT correlation, after fitting, was not significantly lower than the lower bound of

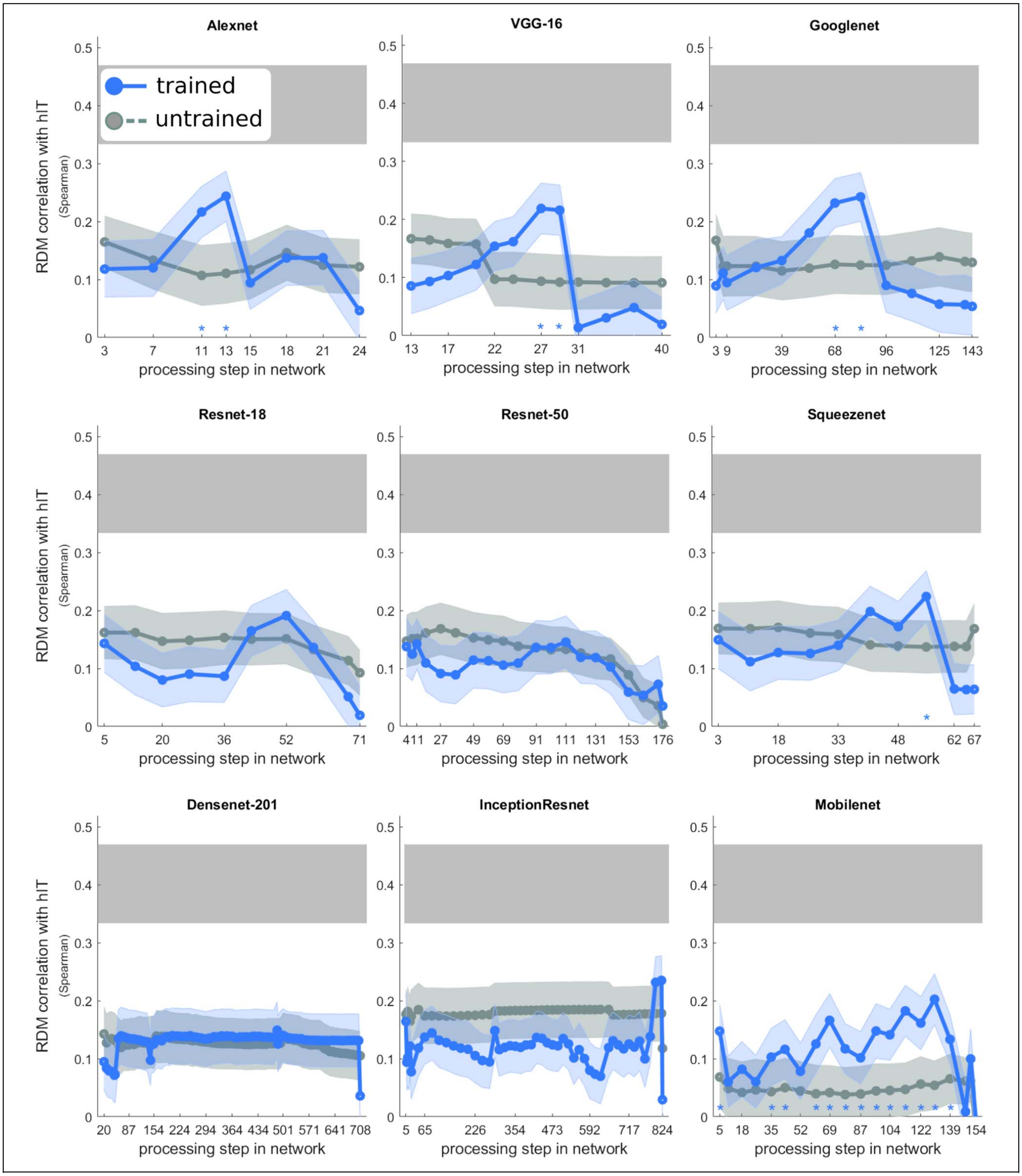

Figure 3. Object recognition training improves hIT correspondence in some layers of some models. Panels show the hIT correlation of the full representation in each layer of each architecture, with no dimensionality reduction or feature reweighting, for an object-recognition trained instance of each network (blue) and for a corresponding untrained instance with randomly initialized weights (gray). Each dot corresponds to one of the key layers selected for analysis (see Table 1) and indicates the mean of a distribution of 1000 bootstrap estimates of cross-validated layer performance, bootstrapped over both subjects and stimuli. For comparability across the diverse architectures, the "depth" of each layer is indicated in terms of the number of unique processing steps up to this point, such as convolution, batch normalization, pooling, or nonlinearity. Shaded regions indicate the standard deviation of the bootstrap distribution. The horizontal gray bar shows the lower and upper bounds of the noise ceiling, indicating the expected performance of the true model, given interparticipant variability in the data set. Blue asterisks above the x axis indicate that the representation in the trained network performs significantly better than that in the untrained network in this layer (α = .05, uncorrected). All layers perform significantly below the lower bound of the noise ceiling (α = .05, uncorrected).

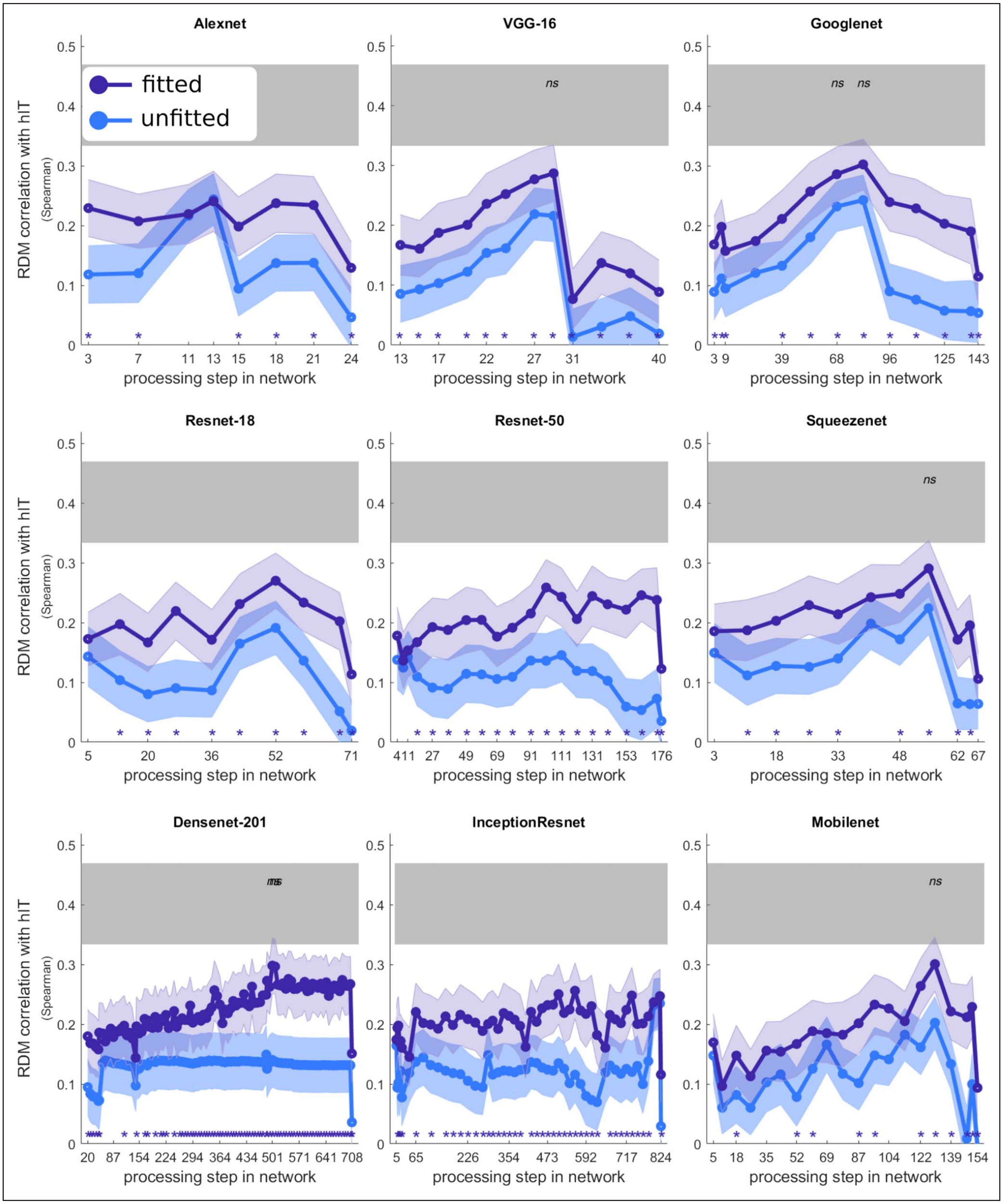

Figure 4. Reducing and reweighting the feature space dramatically improves hIT correlation across most layers in all network architectures. Each panel shows, for one model, the hIT correlation of the unfitted representation in the full original feature space (pale blue lines, same data as shown in Figure 3) and of the same feature space after reducing to 100 dimensions via ecologically driven dimensionality reduction (see Figure 2) and linearly reweighting those dimensions to fit hIT representations (cross-validated over both subjects and images). Blue asterisks indicate that the fitted layer performs significantly better than the original unfitted feature space. Layers in which the fitted representation is not significantly different from the lower bound of the noise ceiling are indicated by "*ns*"; all others perform significantly below the noise ceiling (α = .05, uncorrected).

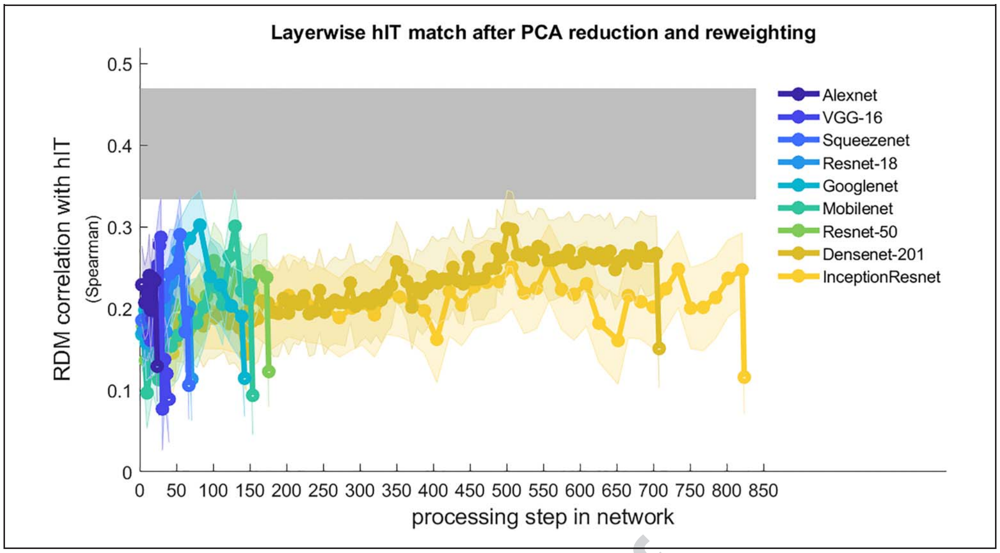

Figure 5. Depth is not the answer. Correlation with hIT representation for all layers of all networks, after "ecologically driven" PCA reduction and reweighting. Although many models reach representations in their late intermediate layers that well match hIT, increasing depth does not equate to increased hIT match, either within or between architectures, within this range of highly successful object-recognition computer vision models.

the noise ceiling (based on the bootstrapped distribution of differences, one-tailed, $\alpha = .05$, uncorrected; note that correcting for multiple comparisons would lower our threshold for considering a layer statistically indistinguishable from the noise ceiling and so would constitute a more liberal criterion than the test performed here).

Deeper Architectures Are Not Better Models of hIT

Despite the wide range of network depths across the nine architectures, spanning 25–825 unique processing steps, we found no compelling evidence that deeper networks were better models of hIT than shallower ones. Figure 5 shows the hIT correlation of each of the fitted layers (i.e., dark blue lines from Figure 4) for all models on the same axes for comparability. Although there was a clear improvement in hIT correspondence across early and intermediate layers within each model, the peak layer performance was similar across models of different depths. There was no correlation between depth of architecture and whole-network hIT match after combining representations across all layers via second-level fitting (see Figure 6), for either trained and within-layer fitted models ($r = .20$, $p = .61$, *ns*) or trained and within-layer unfitted models ($r = -.07$, $p = .86$, *ns*). Within the range of deep convolutional neural networks capable of high object-recognition accuracy, it does not appear that

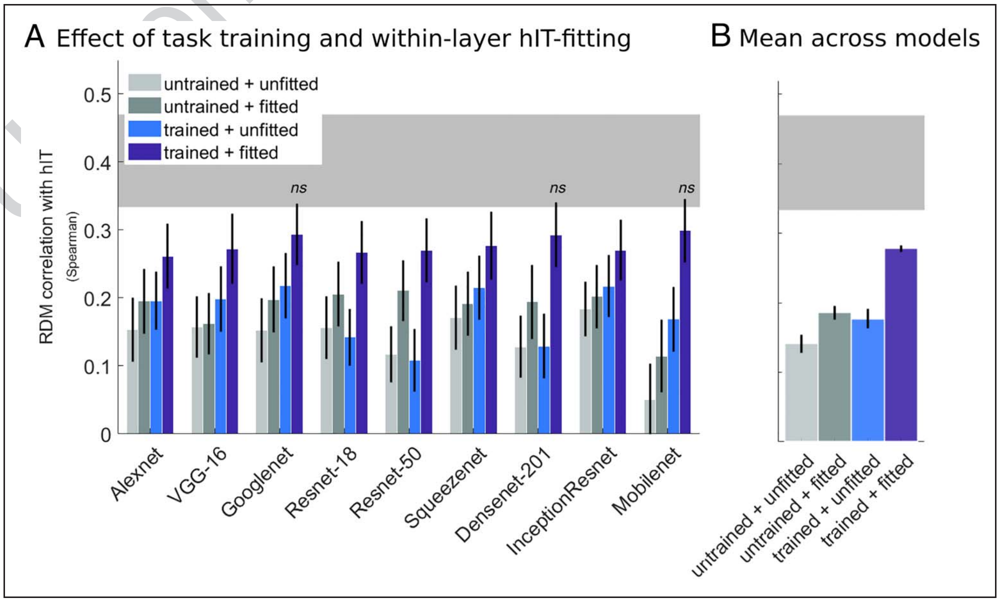

Figure 6. Training and within-layer fitting both improve correlation with hIT. (A) Bars show an estimate of the combined performance of all layers within each of the networks, obtained by second-level (across-layer) fitting in all cases. The fitting procedure is identical to that used to reweight principal components within each layer, except that it takes as input per-layer RDMs rather than PC-RDMs. Pale gray bars show the hIT match for the raw (unfitted) feature space of a randomly weighted instance of the network, dark gray bars show hIT match for the same random feature space after PCA reduction and within-layer reweighting, pale blue bars show hIT match for the unfitted feature space of the object-recognition-trained network, and dark blue bars show hIT match for the trained network after PCA reduction and within-layer reweighting. Error bars indicate the standard deviation of the bootstrap distribution. The horizontal gray bar indicates the lower and upper bounds of the noise ceiling. Models that do not perform significantly below the lower bound of the noise ceiling are indicated by "*ns*" within the noise ceiling; for all others, this comparison is significant, $\alpha = .05$, uncorrected. (B) Data from A averaged across all models. Error bars indicate the standard error of the mean across models.

greater depth leads to more brain-like representations (cf. Kubilius et al., 2018; Schrimpf et al., 2018).

A Combination of Training and Fitting Achieves a Good Match to hIT for Diverse DNNs

So far, we have considered each layer of each network as a separate candidate for predicting hIT. However, it is unlikely that the computational features across large parts of inferior temporal cortex correspond neatly to those in any single processing step of an artificial neural network. We therefore linearly combined the per-layer RDMs of each network to estimate the hIT correlation of the model considered as a whole, via a second-level fitting procedure (see Methods). The inputs to the second-level hIT fitting were either (i) the unfitted per-layer RDMs calculated from the full original feature space in each layer, that is, without dimensionality reduction or first-level (within-layer) fitting (light gray and light blue bars in Figure 6), or (ii) the hIT-fitted per-layer RDMs estimated via first-level fitting (involving both dimensionality reduction and linear reweighting, dark gray and dark blue bars). Both levels of fitting were performed within the same cross-validation loops, so that both within-layer and across-layer weights were always fitted based on the same split of training participant and stimulus data and tested on unseen subjects and stimuli.

This whole network analysis revealed that both object recognition training and subsequent hIT fitting improved the correspondence between model representations and those in hIT (Figure 6B). A 2 × 2 (training × fitting) ANOVA treating each of the nine DNN architectures as an independent observation revealed significant main effects of both training, $F(1, 35) = 33.82$, $p < .0001$, and fitting, $F(1, 35) = 43.69$, $p < .0001$. There was a significant interaction between training and fitting, such that within-layer fitting yields a larger benefit when applied to the features found in layers of trained than untrained models, $F(1, 35) = 6.51$, $p = .016$. This interaction between training and fitting suggests that training on the Imagenet object-recognition task causes models to develop features that capture aspects of real-world images that are important to their representation in hIT but does not cause models to learn the relative prevalence of these features seen in brain data.

To better understand model improvement effect sizes relative to the total explainable (nonnoise) variance in this data set, we calculated descriptive statistics by normalizing squared model correlations by the average squared correlations of the upper and lower noise ceilings (squared Spearman correlation = .16). On average, trained and within-layer fitted models explained 48.0% of the explainable rank variance in hIT representations (normalized squared Spearman correlation = .08). Starting from an untrained, unfitted model as a baseline, on average across models, task training produced a 57.5% increase in the proportion of explainable rank variance explained (taking the noise-ceiling normalized r^2 from .12 to .19). Similarly, hIT fitting of the untrained model produced a 73.4% increase (normalized r^2 from .12 to .21). Although these are substantial gains, the combination of training and hIT fitting achieved a superadditive boost. Compared to trained but unfitted networks, reweighting the features within each layer of trained networks led to a further 124% increase in the proportion of explainable rank variance explained (normalized r^2 from .21 to .48).

After training networks to classify objects, and linearly reweighting their learned features within and across layers, all nine DNN architectures yielded good models of hIT, explaining most of the explainable variance in the data (dark blue bars in Figure 6A). For three architectures (Googlenet, Densenet, and Mobilenet), the trained and hIT-fitted model was not statistically distinguishable from the lower bound of the noise ceiling, indicating performance on par with the ability of individual human brains to predict representational dissimilarities in other human brains (one-tailed test of whether the bootstrap distribution of differences contained zero, $\alpha = .05$, uncorrected). This does not mean that no superior model of the ventral stream can be found but only that no clear discrepancies can be found between representational geometries in brains and models, using the present set of image stimuli and brain data. To evaluate the relative contributions of architectural differences to model performances before and after reweighting, we ran a permutation test comparing all pairwise differences among trained but only second-level fitted (i.e., within-layer-unfitted) models to those among fully fitted models. Differences among the hIT correspondence of unfitted models were larger than those among the same models after they had undergone within-layer principal component reweighting (Cohen's $d = 1.49$, $p = .001$).

To gauge whether the remaining differences in performance among trained and fitted models were likely to be of theoretical interest, we performed an equivalence test (Lakens, 2017) in which we defined the variance in estimates of the lower bound of the noise ceiling as a naturally occurring variation in predicting individual human data. If a difference between two models falls outside the 95% confidence interval of this distribution, the models are more different from one another than human participants are from one another, which could be considered a threshold for the minimal potentially interesting model difference. For each pair of models, we tested whether the observed differences between the hIT correlation of the two models, across bootstrap samples, were significantly larger than the lower bound of the 95% confidence interval on noise ceiling variation and, at the same time, significantly lower than the upper bound of the confidence interval (Lakens, 2017). After training and hIT fitting, the differences between models proved statistically equivalent to the differences between human participants for all models ($\alpha = .05$, Bonferroni corrected for 36 pairwise comparisons among models). However, there were larger differences among pairs of trained but unfitted models, with differences in 8 of the 36 pairwise comparisons falling significantly outside the variability in the lower noise ceiling.

Did these differences point to the superiority of certain DNN architectures as models of the brain over others? If we think of each of these models as embodying different theoretical ideas about the essential properties of visual processing, some of those ideas are baked into the architecture of networks (e.g., the required depth, or spatial scales, of visual processing required), whereas others are implemented in the training task (e.g., that brain-like features emerge from exposure to natural image statistics or from the need to perform ecologically important tasks like object recognition). If the architectural components are the important differentiators of how well this set of models performs, then we might expect that models with good architectural properties might better predict brain data in both their trained and untrained states, which does seem indeed to be the case when looking at the unfitted model data. For within-layer-unfitted models, there was a positive correlation between hIT match before and after task training (Pearson skipped $r = .49$, 95% CI [0.16, 0.96], $\alpha = .05$, using the robust correlation toolbox for MATLAB [Pernet, Wilcox, & Rousselet, 2013]). However, this association evaporates after adjusting the preponderance of different features via fitting. There was no relationship between the performance of trained and untrained models after hIT fitting (Pearson skipped $r = .07$, *ns*). Thus, we see little evidence that some models are able to learn qualitatively better features thanks to their architectures (e.g., being uniquely able to learn features of the right spatial scales or complexities).

Our networks differed substantially in their ability to solve the challenges of object recognition, ranging from 20.9% to 4.9% error rate in object classification on the ILSVRC benchmark task (see Table 1). Among relatively shallow neural networks, models with higher object classification accuracy tend to provide feature spaces that can better predict the firing rates of neurons in macaque IT to object images (Yamins et al., 2014). Among deeper, higher-performing networks, this effect appears to saturate, and further improvements to classification accuracy no longer translate into higher performance as brain models (Schrimpf et al., 2018). We found no significant association, among either unfitted or fitted models, between accuracy on the ILSVRC object classification task and correlation with human IT (Pearson skipped $r = -.38$ (unfitted) and .32 (fitted), both *ns*). The commonalities among these diverse DNNs appear to matter more than their differences, when considered as models of human vision.

Model Dimensions Explaining Most Natural-Image Variation Better Explain Human IT and a Relatively Small Number of Dimensions Suffice

Our first-level (within-layer) fitting procedure consists of two steps, both of which potentially change the representational geometry: First, the full feature space of a layer is reduced to the 100 principal components accounting for the most variation in that layer's responses to ecologically representative images (Mehrer et al., 2021), and second, those principal components are linearly reweighted to best predict dissimilarities between images in human IT. To assess the effect of dimensionality reduction alone, we calculated the performance of a version of each model that had undergone the first (dimensionality reduction) step, but not the second (hIT-fitting) step. Within the main cross-validation procedure (see Methods), we created an RDM for each layer of each network by uniformly combining the layer's 100 PC-RDMs with equal weights and used these RDMs as a basis for the second-level (across-layer) fitting procedure. The resulting whole-network hIT correlations are shown by the central data points in Figure 7A. For comparison, to the left are shown performances using the "raw" layer representation, with no dimensionality reduction or reweighting, and to the right are performances using the reduced and hIT-fitted representation, both previously shown in Figure 6.

Despite reducing the dimensionality of the feature space by up to four orders of magnitude for some layers of some networks (e.g., in the earliest layers of VGG-16, from over 1 million units to only 100 principal dimensions), PCA improved the hIT correspondence of models (Figure 7A). A 2 × 2 (Dimensionality Reduction × Training) ANOVA revealed a main effect of both dimensionality reduction, $F(1, 35) = 6.39, p = .166$, and of training, $F(1, 35) = 15.14$, $p = .0005$. Post hoc tests on the effect of dimensionality reduction show that uniformly weighting the first 100 principal components within each layer led to higher hIT match than taking the full feature space, for both trained networks (paired-samples t test), $t(8) = 15.33, p < .0001$, and untrained networks, $t(8) = 3.23, p = .0121$. There was no interaction between training and dimensionality reduction, $F(1, 35) = 1.09$, *ns*. On average, for the trained networks, the Spearman correlation with hIT could be improved by 25.5% simply by taking dimensions within each layer that account for the most variation in the network's responses to natural images sampled from a set of object categories that appear frequently in human (visual) experience (Mehrer et al., 2021).

We also tested how robust model performance estimates were to changing the number of principal components used for the analysis. Although we have seen that dimensionality reduction helped models predict hIT representations (Figure 7A), networks with different intrinsic dimensionalities may nevertheless be differentially impacted when reduced to the same number of principal components. For example, the effective dimensionality of trained networks could be systematically lower than that of their untrained counterparts, which create random (and therefore approximately orthogonal) high-dimensional projections of their inputs. We therefore explored, within one example network, the effect of taking a wide range of different numbers of principal components for the within-layer reweighting procedure, ranging from 1 to 3000 (the

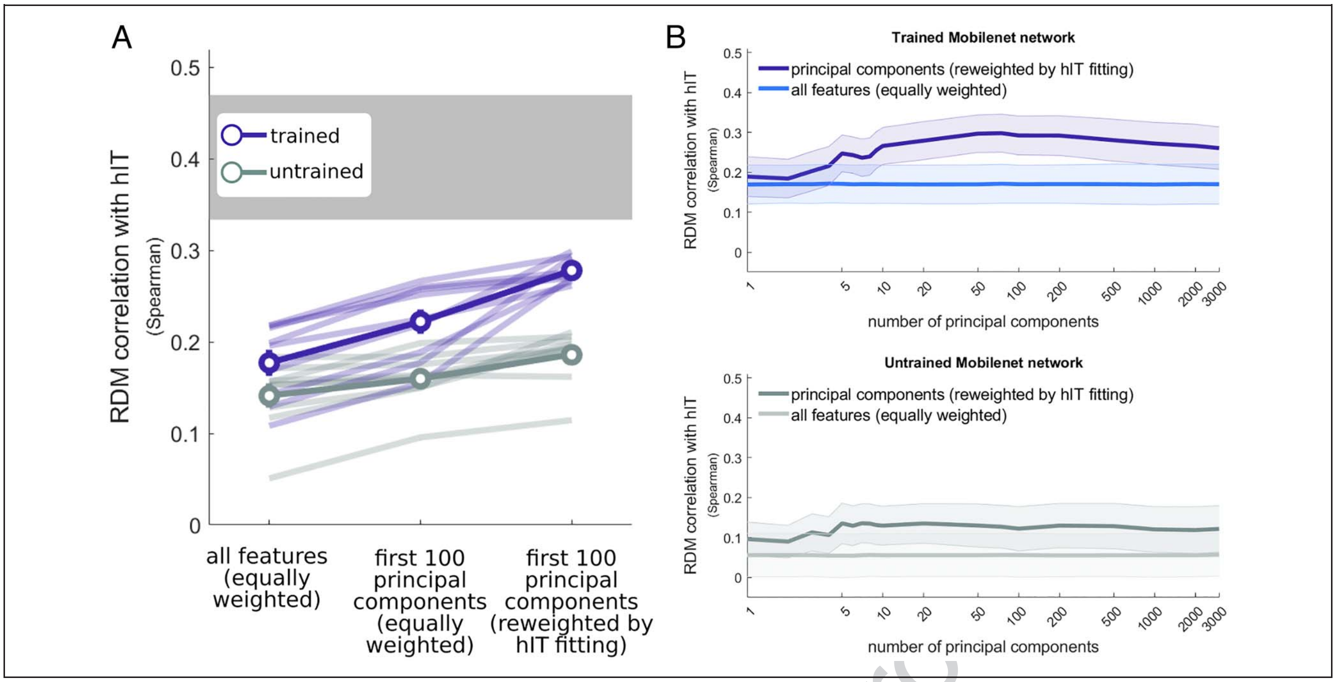

Figure 7. Finding dimensions of natural-image variation within each layer's feature space improve hIT match, and a small number of dimensions are sufficient. (A) Estimates of the whole-network hIT correlation for trained (blue) and random (gray) networks, derived by reweighting layer RDMs obtained from either (left) the full original feature space; (center) the first 100 components of that feature space, derived via PCA on an independent natural image set; and (right) after reweighting the principal components within each layer to predict hIT representations. (B) Estimates of the whole-network hIT correlation for one example network, Mobilenet, in its trained (top) and untrained (bottom) states, after reducing the feature space within each layer to various numbers of principal components (*x* axis). Whole-network performance is estimated by reweighting layer RDMs obtained from either (pale lines) the full original feature space or (dark lines) within-layer hIT-fitted representations using the specified number of principal components. Note that the *x* axis is logarithmic.

maximum possible was 3020, as determined by the size of the independent image set used to calculate principal components). We chose Mobilenet as the example network, because it achieved not only the highest hIT correspondence of all networks in its trained and fitted state but also the lowest hIT correspondence of all networks in its untrained and unfitted state. If untrained networks are disproportionately hampered by the limitations of taking 100 principal components, we expected to see improvements in the performance of untrained Mobilenet when a larger number was allowed. Figure 7B shows the hIT RDM correlation for trained and untrained Mobilenet, after both first-level (within-layer) and second-level (across-layer) fitting, using different numbers of PC-RDMs. Large improvements are evident when increasing from 1 to 10 dimensions, but plateau by 50 dimensions for both trained and untrained networks. One hundred principal components therefore appear more than sufficient to capture the portion of representational variance relevant for explaining hIT responses within this experimental image and fMRI data set.

The Same DNN Models Can Also Predict Representations in V1 Well, After Training and Fitting, by More Strongly Weighting Earlier Layers

Our primary interest is in finding a good model of the complex object representations in late ventral stream. However, DNNs contain a hierarchy of features, from simple to complex, and so also provide candidate models for earlier visual regions. Some previous research (Cadena et al., 2019) has reported that randomly weighted, untrained DNNs perform as well as trained ones in predicting representations in mouse visual cortex. Might object-recognition training therefore also provide less benefit to models when evaluated on the "simpler" representations in human primary visual cortex (V1), compared to hIT? V1 representational geometry measured during the same fMRI scanning sessions reveals a substantially different structure from that in hIT (Figure 8A), although intersubject consistency is similarly high in both regions. The average Spearman correlation between each participant's full V1 RDM and the average V1 RDM of all participants was $r = .381$, and in hIT, $r = .383$, but only $r = .163$ when trying to predict average hIT data from individual V1 data, or vice versa. Yet despite the distinct representational geometries, we find a strikingly similar pattern of model performances in both regions (Figure 8B).

Both object-recognition training and V1 fitting (cross-validated over both images and subjects) improved the ability of diverse DNNs to predict V1 representations. A 2 × 2 (Training × Fitting) ANOVA treating each of the nine DNN architectures as an independent observation revealed main effects of both training, $F(1, 35) = 15.40$, $p = .0004$, and of fitting, $F(1, 35) = 20.46$, $p = .0001$.

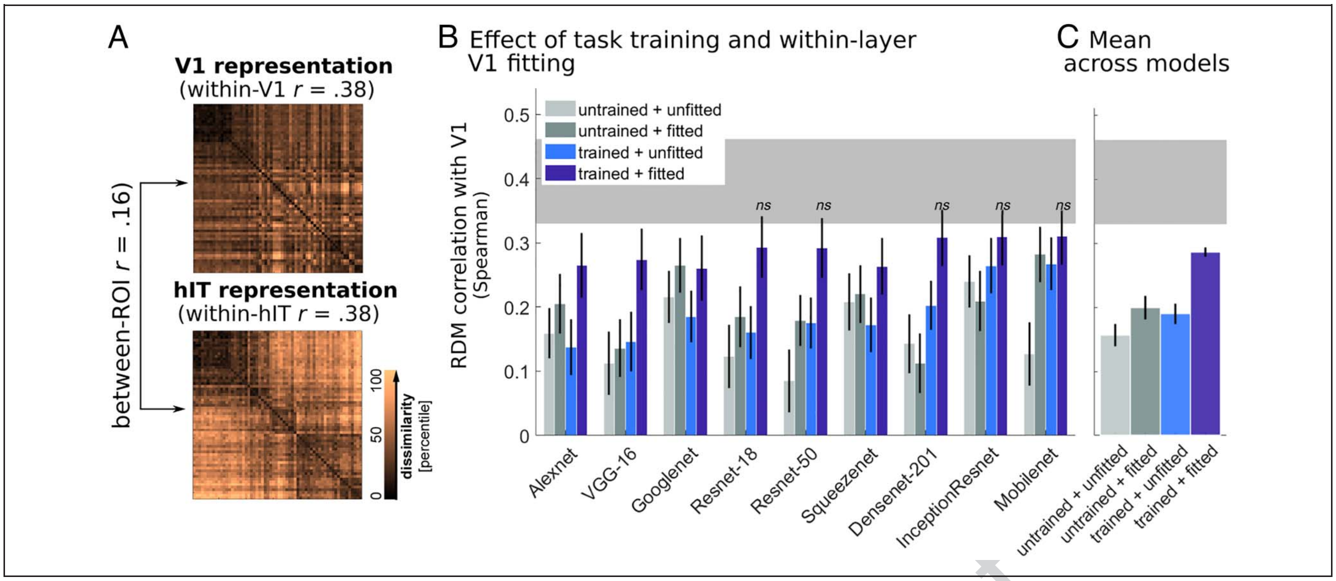

Figure 8. The same models are able to well explain the very different representations in V1. (A) Representational dissimilarity of the same 62 experimental images in V1 (top) and hIT (bottom; as in Figure 1B) averaged over participants. There was good interindividual consistency within each region, but there were substantial differences in representational geometry between regions. (B) Bars show an estimate of the combined performance of all layers within each of the networks, obtained by second-level (across-layer) fitting in all cases. Pale gray bars show the V1 match for the raw (unfitted) feature space of a randomly weighted instance of the network, dark gray bars show V1 match for the same random feature space after PCA reduction and within-layer reweighting on V1 data, pale blue bars show V1 match for the unfitted feature space of the object-recognition-trained network, and dark blue bars show V1 match for the trained network after PCA reduction and within-layer reweighting on V1 data. (C) Data from C averaged across all models. Conventions are as in Figure 6.

Unlike in hIT, however, the interaction between training and fitting was not significant, $F(1, 35) = 3.08$, $p = .09$. As in hIT, after training networks to classify objects and linearly reweighting their learned features within and across layers, all nine DNN architectures yielded good models of brain representations (dark blue bars in Figure 8B). For five architectures (Resnet-18, Resnet-50, Densenet, InceptionResnet, and Mobilenet), the trained and V1-fitted model was not statistically distinguishable from the lower bound of the noise ceiling, indicating similar performance to that of using individual human brain representations to predict others (one-tailed bootstrap percentile test of the null hypothesis that the difference is zero, $\alpha = .05$, uncorrected). A permutation test showed that differences in V1 correspondence were higher among unfitted models than their V1-fitted counterparts (Cohen's $d = 1.05$, $p = .002$).

To evaluate whether differences among models were larger than the differences among individual participants, we performed an equivalence test (Lakens, 2017). After training and V1 fitting, the differences between models proved statistically equivalent to the differences between human participants for all models ($\alpha = .05$, Bonferroni corrected for 36 pairwise comparisons among models). This result suggests that the variation among human RDMs is similar to that among model RDMs, although these should be interpreted with caution because the human data are affected by substantial measurement noise. However, there were larger differences among pairs of trained but unfitted models, with differences in 10 of the 36 pairwise comparisons falling significantly outside the variability in the lower noise ceiling. This representational variance across networks is likely caused both by architectural differences and by "individual differences" among networks trained from different random initializations of the same architecture (Mehrer, Spoerer, Kriegeskorte, & Kietzmann, 2020). With only a single trained instance of each architecture, we are not able to assess the relative contributions of each.

As in hIT, networks differed in their ability to predict brain representations in their raw states but performed similarly well after training and fitting. There was no evidence of a relationship between the ability of a DNN to predict V1 and its ability to predict hIT representations (Pearson skipped rs = .25 [for trained, fitted models], .06 [trained but unfitted], −.36 [untrained fitted], and .55 [untrained unfitted]; all values = *ns*, using the robust correlation toolbox for MATLAB [Pernet et al., 2013]).

Although all networks were able to predict both V1 and hIT representations well after combining representations across their layers (Figure 8B), we expect substantial differences in the specific layers that best explain each region (Zeman et al., 2020; Güçlü & van Gerven, 2015). To quantify the network depth within which performance peaked and compare across models with different numbers of layers, we first divided layer indices within each network by the maximum depth of the network, so that all networks had a nominal depth of 1. We then fitted piecewise polynomial splines to each network's layerwise hIT or V1 correlation, using MATLAB's smoothing splines function (smoothing factor = 0.999). Figure 9 (leftmost panels) shows the average of the resulting

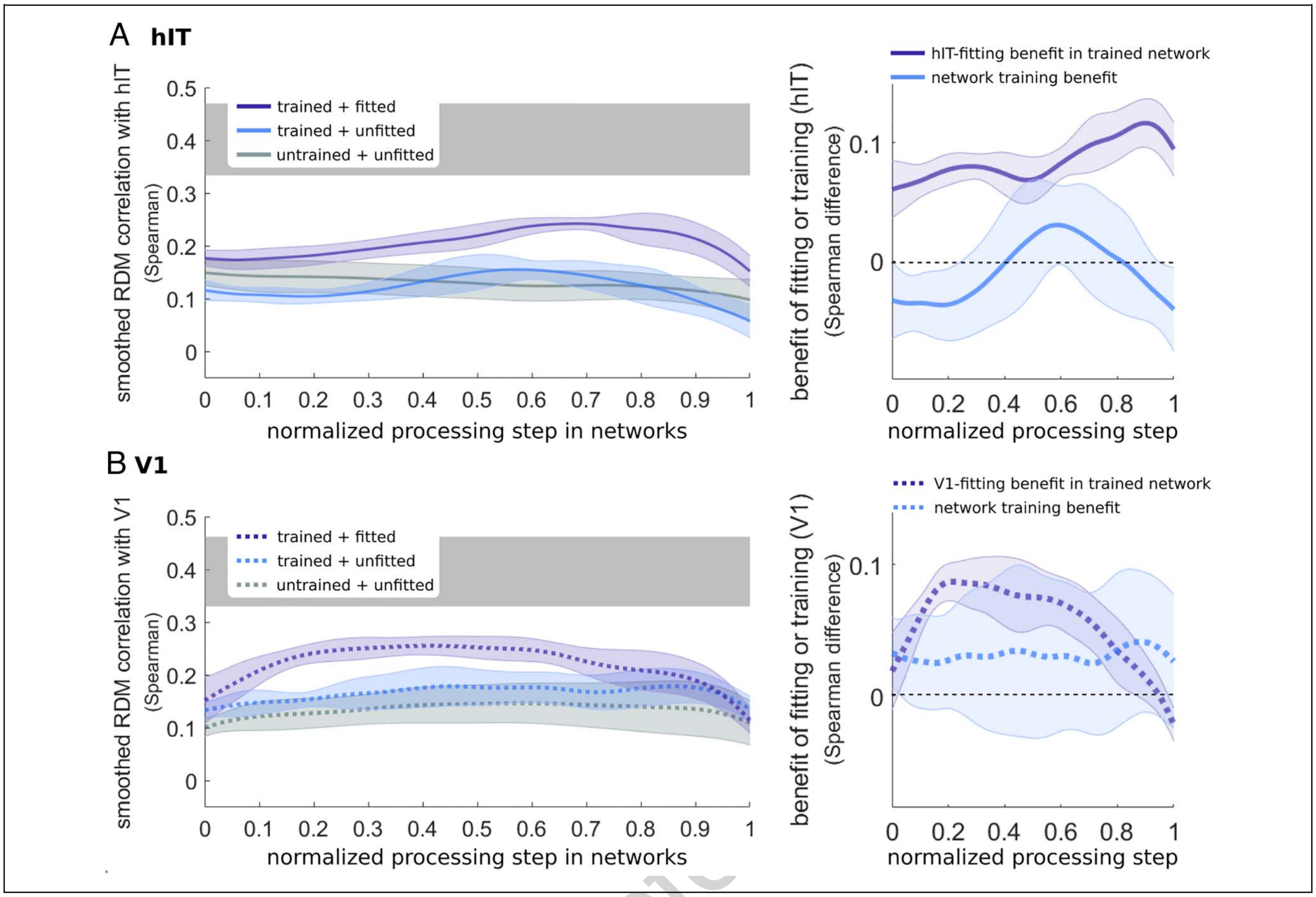

Figure 9. Fitting most improves the performance of deeper layers when predicting hIT and earlier layers when predicting V1. (A, Left) Mean and 95% confidence interval across nine DNNs of smooth splines fit to layerwise hIT correlation, normalized by network depth (raw values are shown in Figures 3 and 4), for (gray) untrained and unfitted models, (pale blue) trained but unfitted models, and (dark blue) trained and within-layer hIT-fitted models. (A, Right) Mean and 95% confidence interval of smooth splines fit to the difference in hIT correlation between either (pale blue) trained unfitted minus untrained unfitted networks or (dark blue) trained and fitted minus trained but unfitted networks. The horizontal gray bar indicates the lower and upper bounds of the noise ceiling. (B) The same models, evaluated on and fitted to primary visual cortex (V1) representations.

smoothed layerwise hIT and V1 correlations, before and after training and fitting. We compared the peaks of the smoothed layerwise brain correlations between regions. For trained but unfitted models, there was, surprisingly, no indication that later layers better predicted hIT data (mean location of peak layerwise performance, as a proportion of network depth = 0.56) than V1 (mean = 0.59; two-tailed paired-samples $t(8) = -0.10$, *ns*). However, after fitting each layer of the trained models to the respective brain region's representation, peak performance was reached robustly later in hIT ($M = 0.70$) than V1 ($M = 0.48$), $t(8) = 3.20$, 95% CI [0.06, 0.38], $p = .013$. To directly compare the effect of depth on fitting benefit in each region, we also calculated for each model and region the layerwise difference between brain-data correlation for a trained and fitted network, minus its trained but unfitted counterpart, and fitted a smoothing spline (Figure 9, rightmost panels). Fitting to brain data yielded larger benefits for deeper network layers when evaluated on hIT data (mean peak of smoothed fitting benefit = 0.76), than V1 ($M = 0.35$), $t(8) = 3.73$, 95% CI [0.16, 0.67], $p = .006$.

Together, these results support the intuition that earlier layers develop features more suited to predicting early visual cortical representations, whereas later layers contain features better able to explain representations in late object-selective inferior temporal cortex, likely because of differences in complexity, structure, and/or spatial scale across DNN layers. When we allow the relative influence of different features within a layer to be reweighted to better match brain data, earlier layers best predict V1, and later layers best predict hIT. The correspondence between network depth and cortical region is perhaps surprisingly subtle in this data set, however, because it is not evident in trained but unfitted models.

DISCUSSION

In this work, we investigated a diverse set of DNNs for their ability to predict representations estimated from fMRI data of hIT cortex. Comparing the predictive performance of untrained and object-recognition-trained network variants, we show that task training moderately improves correspondence with representations in hIT.

At a minimum, this suggests that structured visual features (e.g., containing spatial correlations) form a better basis for predicting those in the human brain and may point to the importance of ecologically relevant tasks in developing brain-like features. The effect of training is substantially amplified by model fitting, indicating that the relative prevalence of different features in hIT does not automatically emerge from the particular object-recognition task (ImageNet) used to train the networks. After task training and two-stage model fitting, the predictive performance of all networks, irrespective of depth, on data from unseen stimuli and participants, was similarly high, explaining 48% of the rank variance in hIT representations. This is similar to the proportion of variance DNNs have been found to explain in macaque electrophysiological data (Bashivan et al., 2019; Cadieu et al., 2014; Yamins et al., 2014).

Diverse architectures performed indistinguishably well as models of hIT, after training and fitting. The differences among trained (but unfitted) architectures did not correlate with other model metrics, such as network depth, object-recognition performance, hIT match before training, or ability to predict V1 representations. That is, we did not find any factors that caused certain architectures to be inherently better models of the ventral stream than others. This is perhaps unexpected, given that the set of DNNs could be thought of as implementing different computational theories about vision. For example, Liao and Poggio (2016) demonstrated that the Resnet architecture is equivalent to a recurrent network with certain constraints and therefore may be more biologically plausible than the others we consider. However, the lack of difference still informs theory in suggesting that the models' architectural specifics—depth, numbers of feature maps, sizes of filters, presence of skipping, or branching connections—matter less than their shared attributes. All nine networks are deep feedforward hierarchies of nonlinear features, with spatially restricted receptive fields whose size grows across layers. Although the present results cannot tease apart the contributions of each of these factors, previous work suggests that all are likely required in any successful model of the visual system (e.g., Khaligh-Razavi & Kriegeskorte, 2014; Riesenhuber & Poggio, 1999; Fukushima & Miyake, 1982; Hubel & Wiesel, 1962).

The evaluation of theories, and their implementations via computational models, against both brain and behavioral data is a very large endeavor. Our specific goal within this project was to systematically explore the effects of training and fitting for a diverse set of computational models, evaluated against neural representations, measured via fMRI, for one set of image stimuli. A recent study evaluating an overlapping set of DNNs against behavioral, rather than neural, data (Geirhos, Meding, & Wichmann, 2020) found similarly small differences between models. Convergently, both lines of work suggest that the architectural differences within the diverse family of modern supervised DNNs do not strongly impact their performance as models of human visual processing. This is a theoretically important finding that helps direct future work toward less explored factors, such as the computational role of recurrence (e.g., Kietzmann et al., 2019) and the learning objectives of biological brains (e.g., Storrs, Anderson, & Fleming, 2021; Marblestone, Wayne, & Kording, 2016).

Injecting Domain Knowledge through Training Helps Explain Brain Representations

A number of studies have shown that performance-optimized DNNs can explain representations in high-level regions of human and nonhuman primate ventral visual cortex (Xu & Vaziri-Pashkam, 2020; Kubilius et al., 2018; Schrimpf et al., 2018; Cichy et al., 2016; Yamins & DiCarlo, 2016; Güçlü & van Gerven, 2015; Kriegeskorte, 2015; Agrawal et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Our results here replicate and extend upon this widely appreciated finding by testing a large variety of models on the same hIT data set, while at the same time providing estimates for the relative effects of task training and model fitting.

Knowing the size of the effect of task training on a given network's predictive power serves an important role in our goal of understanding ventral stream computations. Following the normative approach, training models on an external task can help answer the question of "why" the ventral stream computes what it computes (Kietzmann et al., 2018). The present results indicate that encoding statistical structure or redundancy in images is important for recreating hIT-like representations, as randomly initialized untrained deep architectures performed significantly worse. Although a larger set of supervised and unsupervised training objectives must be investigated, exposure to natural images and the need to derive ecologically relevant information from those images (e.g., object identity) may provide important constraints on the kinds of visual features developed by brains and DNNs. At the same time, the dramatic benefits of model fitting suggest that training on the ImageNet ILSVRC challenge does not lead to the correct relative feature distribution.

In addition to helping answer "why" brain representations take the form they do, task training allows us to harness large training data sets to instill domain knowledge into models. Vision, like other feats of intelligence, requires knowledge about the world. In particular, recognition requires knowledge of what things look like. To explain task performance and high-level responses, therefore, a model needs the parametric capacity to store the requisite knowledge. One benefit of task training is that it allows experimenters to inject domain knowledge (e.g., about the structure of visual images) into models in a way that is entirely unfeasible in more hand-crafted or analytical models. The features learned through automated training on big data are "potential candidates" for the sorts of features likely represented in biological visual systems. Yet, the differences in predictive performance between random and trained versions of the same architecture are small, suggesting we should be cautious about interpreting the features learned by trained networks as being informative about the particular features represented in the ventral stream.

Reweighting Features Improves Correspondence to Brain Representations and Reveals Common Performance across Diverse Models

After finding dimensions of important variance, and reweighting to adjust the relative strength of those feature dimensions, good prediction of representations in either early (V1) or late (hIT) ventral cortex could be achieved by all models, with negligible differences in performance among very diverse architectures ranging from 8 to 201 layers in depth. The fitted and unfitted performance estimates provide us with different information, and experimenters may wish to use or not use fitting depending on their modeling objective.

The substantial benefit of fitting to brain data (124% improvement in hIT prediction, for trained models) should not be surprising, for at least two reasons. First, the object-recognition-trained networks had as their only requirement the classification of 1000 nameable objects, with a distribution highly unlike that found in the human visual diet—for example, the ILSVRC categories do not contain "person" or "face" (Mehrer et al., 2021; Russakovsky et al., 2015). The human ventral stream, in contrast, must subserve a wide range of behaviors beyond recognition such as navigation, interaction, and memory. For some research questions, we may be most interested in the performance of models without allowing feature reweighting. Model fitting (including encoding models and single- or two-stage RDM reweighting) always deviates models away from their "native" feature coding, by allowing the prevalence of different features to be adjusted (Khaligh-Razavi et al., 2017). If we are interested in which architectures, training objectives, and visual diets give rise to distributions of features similar to those found in the human visual system, the unfitted performance of models will be most informative.

A second consideration is that, even if an ideal DNN model of human IT were to exist, containing exactly the features and distribution of those features found in the brain, the measurement processes giving rise to data would bias the prevalence of measured features (Kriegeskorte & Diedrichsen, 2016). This provides one motivation for reweighting features—because we know that our measurement processes can introduce bias in feature sampling, requiring a model to match the measured prevalence of features might be too strict a criterion. Instead of reweighting existing features that emerge via task training, researchers have recently started using data from the human ventral stream to directly learn the network features themselves in end-to-end training on natural stimuli (Kietzmann et al., 2019; Seeliger, Ambrogioni, Güçlütürk, Güçlü, & van Gerven, 2019). Such procedures serve the important function of verifying that a given network architecture chosen is in principle capable of mirroring the right representational transitions observed in the brain.

The Future of DNNs as Models in Visual Neuroscience

The advent of high-performing object-recognition DNNs in computer vision has provided visual neuroscience with unprecedentedly good models for predicting visual responses in the human and nonhuman primate brain (Lindsay, 2020; Kietzmann et al., 2018; Schrimpf et al., 2018; Kriegeskorte, 2015; Yamins et al., 2014). Yet despite the achievements of such models, they are far from perfect models of biological vision, exhibiting fragility in the face of noise and other perturbations (Geirhos, Jacobsen, et al., 2020; Geirhos et al., 2017), an overreliance on textural information (Geirhos et al., 2018), and limited ability to predict brain responses to artificial stimuli (Xu & Vaziri-Pashkam, 2020). As a field, we are only scratching the surface by evaluating off-the-shelf feedforward DNNs trained on tasks devised by software engineers. Deep learning offers a powerful and flexible modeling framework based on biologically motivated elements.

Going forward, deep learning models in visual neuroscience will more broadly explore the space of objective functions, learning rules, architectures (Richards et al., 2019), and training diets (Mehrer et al., 2021). Recurrent networks can recycle neural resources to flexibly trade speed for accuracy in visual recognition and show great promise as models of temporal dynamics in visual cortex (van Bergen & Kriegeskorte, 2020; Kietzmann et al., 2019; Nayebi et al., 2018; Güçlü & van Gerven, 2017; Spoerer, McClure, & Kriegeskorte, 2017). Unsupervised learning objectives provide rich and ecologically feasible ways of getting complex knowledge about the visual world into the brain (Storrs et al., 2021; Storrs & Fleming, 2020). One of the central goals of computational visual neuroscience is a model that can predict neural representations in visual cortex at multiple levels of granularity, from single neuron responses to the aggregated population signals measured via fMRI, and can also predict the perceptual properties of our visual systems, as measured in behavioral experiments (Funke et al., 2020; Hebart, Zheng, Pereira, & Baker, 2020; Storrs & Kriegeskorte, 2020; Rajalingham et al., 2018; Schrimpf et al., 2018; Jozwik, Kriegeskorte, Storrs, & Mur, 2017). Models will be tested using larger data sets, with higher noise ceilings, and with stimuli designed to tease apart the differences between model predictions (Golan, Raju, & Kriegeskorte, 2019) and to minimize confounding low-level visual properties (Bracci, Ritchie, Kalfas, & Op de Beeck, 2019; Bracci & Op de Beeck, 2016). We have come a long way but are only just beginning to explore the full potential of deep learning in visual neuroscience.

Reprint requests should be sent to Katherine R. Storrs, Justus-Liebig-Universität Gießen, Fachbereich 06, Psychologie und Sportwissenschaft, Abteilung Allgemeine Psychologie, Otto-Behaghel-Strasse 10F, 35394 Giessen, Germany, or via e-mail: katherine.storrs@gmail.com.

Author Contributions

Katherine R. Storrs: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Visualization; Writing—Original draft; Writing—Review & editing. Tim C. Kietzmann: Conceptualization; Methodology; Writing—Review & editing. Alexander Walther: Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Writing—Original draft. Johannes Mehrer: Conceptualization; Formal analysis; Investigation; Methodology; Visualization; Writing—Review & editing. Nikolaus Kriegeskorte: Conceptualization; Funding acquisition; Investigation; Methodology; Resources; Supervision; Writing—Review & editing.

Funding Information

This work was supported by the UK Medical Research Council, grant number: MC-A060-5PR20. Nikolaus Kriegeskorte: H2020 European Research Council (https://dx.doi.org/10.13039/100010663), grant number: 261352. Johannes Mehrer: Vice Chancellor's Award by the Cambridge Trust. Katherine R. Storrs: Alexander von Humboldt Foundation fellowship (https://dx.doi.org/10.13039/100005156).

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. L. (2014). Pixels to voxels: Modeling visual representation in the human brain. *ArXiv Preprint ArXiv:1407.5104*.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*, eaav9436. https://doi.org/10.1126/science.aav9436, PubMed: 31048462

Benson, N. C., Butt, O. H., Brainard, D. H., & Aguirre, G. K. (2014). Correction of distortion in flattened representations of the cortical surface allows prediction of V1–V3 functional organization from anatomy. *PLoS Computational Biology*, *10*, e1003538. https://doi.org/10.1371/journal.pcbi.1003538, PubMed: 24676149

Bracci, S., & Op de Beeck, H. O. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *Journal of Neuroscience*, *36*, 432–444. https://doi.org/10.1523/JNEUROSCI.2314-15.2016, PubMed: 26758835

Bracci, S., Ritchie, J. B., Kalfas, I., & Op de Beeck, H. P. O. (2019). The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience*, *39*, 6513–6525. https://doi.org/10.1523/JNEUROSCI.1714-18.2019, PubMed: 31196934

Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., et al. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? Paper presented at *Advances in Neural Information Processing, Neuro AI Workshop*.

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*, e1003963. https://doi.org/10.1371/journal.pcbi.1003963, PubMed: 25521294

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755. https://doi.org/10.1038/srep27755, PubMed: 27282108

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, *8*, 1–12. https://doi.org/10.1038/s41598-018-28865-1, PubMed: 30006530

Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, *152*, 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001, PubMed: 27777172

Fukushima, K., & Miyake, S. (1982). Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets* (pp. 267–285). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-46466-9_18

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S., & Bethge, M. (2020). The notorious difficulty of comparing human and machine perception. *ArXiv Preprint ArXiv:2004.09406*. https://doi.org/10.32470/CCN.2019.1295-0

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., et al. (2020). Shortcut learning in deep neural networks. *ArXiv Preprint ArXiv:2004.07780*. https://doi.org/10.1038/s42256-020-00257-z

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: Object recognition when the signal gets weaker. *ArXiv Preprint ArXiv:1706.06969*.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *arXiv preprint arXiv:2006.16736*.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *ArXiv Preprint ArXiv:1811.12231*.

Golan, T., Raju, P. C., & Kriegeskorte, N. (2019). Controversial stimuli: Pitting neural networks against each other as models of human recognition. *ArXiv Preprint ArXiv:1911.09288*.

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*, 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015, PubMed: 26157000

Güçlü, U., & van Gerven, M. A. (2017). Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in Computational Neuroscience*, *11*, 7. https://doi.org/10.3389/fncom.2017.00007, PubMed: 28232797

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*, 1173–1185. https://doi.org/10.1038/s41562-020-00951-3, PubMed: 33046861

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*, 1–15. https://doi.org/10.1038/ncomms15037, PubMed: 28530228

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837, PubMed: 14449617

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50× fewer parameters and <0.5 MB model size. *ArXiv Preprint ArXiv:1602.07360*.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, 1726. https://doi.org/10.3389/fpsyg.2017.01726, PubMed: 29062291

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., & Kriegeskorte, N. (2017). Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, *76*, 184–197. https://doi.org/10.1016/j.jmp.2016.10.007, PubMed: 28298702

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*, e1003915. https://doi.org/10.1371/journal.pcbi.1003915, PubMed: 25375136

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, *97*, 4296–4309. https://doi.org/10.1152/jn.00024.2007, PubMed: 17428910

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2018). Deep neural networks in computational neuroscience. *BioRxiv*, 133504. https://doi.org/10.1101/133504

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences, U.S.A.*, *116*, 21854–21863. https://doi.org/10.1073/pnas.1905544116, PubMed: 31591217

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447, PubMed: 28532370

Kriegeskorte, N., & Diedrichsen, J. (2016). Inferring brain-computational mechanisms with models of activity measurements. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *371*, 20160278. https://doi.org/10.1098/rstb.2016.0278, PubMed: 27574316

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008, PubMed: 19104670

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*, 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043, PubMed: 19109916

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing Systems*, 1097–1105.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385. https://doi.org/10.1101/408385

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355–362. https://doi.org/10.1177/1948550617697177, PubMed: 28736600

Liao, Q., & Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv preprint arXiv:1604.03640*.

Lindsay, G. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 1–15. https://doi.org/10.1162/jocn_a_01544, PubMed: 32027584

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 94. https://doi.org/10.3389/fncom.2016.00094, PubMed: 27683554

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*, 1–12. https://doi.org/10.1038/s41467-020-19632-w, PubMed: 33184286

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences, U.S.A.*, *118*, e2011417118. https://doi.org/10.1073/pnas.2011417118, PubMed: 33593900

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., et al. (2018). Task-driven convolutional recurrent models of the visual system. In *Advances in neural information processing systems* (pp. 5290–5301).

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, *10*, e1003553. https://doi.org/10.1371/journal.pcbi.1003553, PubMed: 24743308

Pernet, C. R., Wilcox, R. R., & Rousselet, G. A. (2013). Robust correlation analyses: False positive and power validation using a new open source MATLAB toolbox. *Frontiers in Psychology*, *3*, 606. https://doi.org/10.3389/fpsyg.2012.00606, PubMed: 23335907

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell*, *177*, 999–1009. https://doi.org/10.1016/j.cell.2019.04.005, PubMed: 31051108

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks.

Journal of Neuroscience, *38*, 7255–7269. https://doi.org/10.1523/JNEUROSCI.0388-18.2018, PubMed: 30006365

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025. https://doi.org/10.1038/14819, PubMed: 10526343

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., et al. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*, 1761–1770. https://doi.org/10.1038/s41593-019-0520-2, PubMed: 31659335

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobilenetV2: Inverted residuals and linear bottlenecks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., et al. (2018). *Brain-Score: Which artificial neural network for object recognition is most brain-like? BioRxiv Preprint*. https://doi.org/10.1101/407007

Seeliger, K., Ambrogioni, L., Güçlütürk, Y., Güçlü, U., & van Gerven, M. A. (2019). End-to-end neural system identification with neural information flow. *BioRxiv*, 553255. https://doi.org/10.1101/553255, PubMed: 31659710

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551. https://doi.org/10.3389/fpsyg.2017.01551

Storrs, K. R., Anderson, B. L., & Fleming, R. W. (2021). Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 1–16. https://doi.org/10.1038/s41562-021-01097-6, PubMed: 33958744

Storrs, K. R., Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2020). Noise ceiling on the crossvalidated performance of reweighted models of representational dissimilarity: Addendum to Khaligh-Razavi & Kriegeskorte (2014). *BioRxiv*. https://doi.org/10.1101/2020.03.23.003046

Storrs, K. R., & Kriegeskorte, N. (2020). Deep learning for cognitive neuroscience. In *The cognitive neurosciences* (6th ed.). Cambridge, MA: MIT Press.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. Paper presented at *Thirty-First AAAI Conference on Artificial Intelligence*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Truzzi, A., & Cusack, R. (2020). Convolutional neural networks as a model of visual activity in the brain: Greater contribution of architecture than learned weights. In *Bridging AI and cognitive science*. ICLR.

van Bergen, R. S., & Kriegeskorte, N. (2020). Going in circles is the way forward: The role of recurrence in visual inference. *ArXiv Preprint ArXiv:2003.12128*. https://doi.org/10.1016/j.conb.2020.11.009, PubMed: 33279795

Walther, A. (2015). *Beyond brain decoding: Representational distances and geometries* (PhD Thesis). University of Cambridge.

Walther, A., Diedrichsen, J., Mur, M., Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2016). Sudden emergence of categoricality at the lateral–occipital stage of ventral visual processing. *Journal of Vision*, *16*, 407–407. https://doi.org/10.1167/16.12.407

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, *137*, 188–200. https://doi.org/10.1016/j.neuroimage.2015.12.012, PubMed: 26707889

Xu, Y., & Vaziri-Pashkam, M. (2020). *Limited correspondence in visual representation between the human brain and convolutional neural networks*. *BioRxiv*. https://doi.org/10.1101/2020.03.12.989376

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365. https://doi.org/10.1038/nn.4244, PubMed: 26906502

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111, PubMed: 24812127

Zeman, A. A., Ritchie, J. B., Bracci, S., & de Beeck, H. O. (2020). Orthogonal representations of object shape and category in deep convolutional neural networks and human visual cortex. *Scientific Reports*, *10*, 1–12. https://doi.org/10.1038/s41598-020-59175-0, PubMed: 32051467