PNAS

Supporting Information for

Modeling naturalistic face processing in humans with deep convolutional neural networks

Authors

Guo Jiahui[1†*], Ma Feilong[1†*], Matteo Visconti di Oleggio Castello[2], Samuel A. Nastase[3], James V. Haxby[1], M. Ida Gobbini[4,5,6*]

Affiliations

[1] Center for Cognitive Neuroscience, Dartmouth College, Hanover, NH 03755, USA.

[2] Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA.

[3] Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA.

[4] Department of Medical and Surgical Sciences, University of Bologna, Bologna 40138, Italy.

[5] Istituti di Ricovero e Cura a Carattere Scientifico (IRCCS) Istituto delle Scienze Neurologiche di Bologna, Bologna 40139, Italia.

[6] *Lead Contact*

[†]*These authors contributed equally to this work.*

[*]*Correspondence: jiahui.guo@dartmouth.edu, feilong.ma@dartmouth.edu, mariaida.gobbini@unibo.it*

Material and Methods

MRI Data Acquisition

All data were acquired using a 3 T Siemens Magnetom Prisma MRI scanner with a 32-channel head coil at the Dartmouth Brain Imaging Center. CaseForge headcases were used to minimize head motion. BOLD images were acquired in an interleaved fashion using gradient-echo echo-planar imaging with pre-scan normalization, fat suppression, multiband (i.e., simultaneous multi-slice; SMS) acceleration factor of 4 (using blipped CAIPIRINHA), and no in-plane acceleration (i.e., GRAPPA acceleration factor of one): TR/TE = 1000/33 ms, flip angle = 59°, resolution = 2.5 mm^3 isotropic voxels, matrix size = 96 x 96, FoV = 240 x 240 mm, 52 axial slices with full brain coverage and no gap, anterior–posterior phase encoding. At the beginning of each run, three dummy scans were acquired to allow for signal stabilization. The T1-weighted structural scan was acquired using a high-resolution single-shot MPRAGE sequence with an in-plane acceleration factor of 2 using GRAPPA: TR/TE/TI = 2300/2.32/933 ms, flip angle = 8°, resolution = 0.9375 x 0.9375 x 0.9 mm voxels, matrix size = 256 x 256, FoV = 240 x 240 x 172.8 mm, 192 sagittal slices, ascending acquisition, anterior–posterior phase encoding, no fat suppression, and with 5 min 21 s total acquisition time. A T2-weighted structural scan was acquired with an in-plane acceleration factor of 2 using GRAPPA: TR/TE = 3200/563 ms, flip angle = 120°, resolution = 0.9375 x 0.9375 x 0.9 mm voxels, matrix size = 256 x 256, FoV = 240 x 240 x 172.8 mm, 192 sagittal slices, ascending acquisition, anterior–posterior phase encoding, no fat suppression, and lasted for 3 min 21 s. At the beginning of each session (The Grand Budapest Hotel, the Hyperface stimulus, and the localizer task), a fieldmap scan was collected for distortion correction.

DCNN Models

We used five DCNN models in our analysis: three DCNNs trained for face recognition and two DCNNs trained for object recognition. These DCNNs cover a wide range of commonly used “classic” and state-of-the-art DCNN architectures, including AlexNet (1), VGG16 (2), and ResNet100 (3).

DCNN Models Trained for Face Recognition. All 3 Face DCNNs were trained using the MS-Celeb-1M dataset (4), one of the largest publicly available datasets for face recognition. The training dataset contains approximately 10 million images sampled from 100,000 top celebrity identities from a knowledge base comprising 1 million celebrities. We used a curated version of the dataset as provided by the InsightFace package, which contains 85,742 identities and 5.8 million aligned face images with 112 × 112 resolution. These images were divided into 45,490 batches with 128 images each during training.

The main architecture of the three face-DCNNs are ResNet100, AlexNet, and VGG16, respectively. The ResNet100 face-DCNN was the pre-trained ArcFace model provided by the InsightFace package (labeled as "LResNet100E-IR,ArcFace@ms1m-refine-v2", https://github.com/deepinsight/insightface/wiki/Model-Zoo#3-face-recognition-models), commonly known as the pre-trained ArcFace DCNN.

We trained the other two DCNNs also with the ArcFace loss function, one of the most effective loss functions for face recognition (5). To facilitate convergence, for each of these two DCNNs, we first trained it for 4 epochs using the softmax loss function, and then fine-tuned it for another 16 epochs using the ArcFace loss function. For the softmax loss function, we followed the steps of (5) to normalize the embedding vectors and weights.

We used the Adam optimizer for training (6), with an initial learning rate of 2^{-10}. We used a procedure which we call "prestige" to choose the optimal learning rate. That is, for each epoch, we trained two replicas of the DCNN with different learning rates: one with the same learning rate as the previous epoch, and the other with half the previous learning rate. After both replicas had finished training, we only kept the one with a smaller loss. This procedure allows us to train our DCNNs with a satisfactory convergence speed. We also repeated the training of each network twice with different initializations, and only used the one with smaller loss.

We assessed the performance of the two DCNNs we trained using the Labeled Faces in the Wild (LFW) dataset (7). Specifically, we used the curated version of the dataset provided by the InsightFace package, which contained 6000 pairs of images.

We used a 10-fold cross-validation scheme to evaluate the prediction accuracy of our DCNNs. For each pair of images, we computed the similarity of their embedding vectors, and predicted whether they were the same identity or not based on a threshold. For each cross-validation fold, the threshold was chosen based only on training data.

The classification accuracy was 98.75% and 98.45% (chance accuracy: 50%) for the face AlexNet and face VGG16 (Figure S2), respectively, which were comparable to previous results based on DCNNs that had similar architectures (https://paperswithcode.com/sota/face-verification-on-labeled-faces-in-the).

DCNN Models Trained for Object Recognition. We used pretrained AlexNet and VGG16 models provided by the torchvision package (https://pytorch.org/vision/stable/models.html), which were optimized for object recognition based on the ImageNet dataset. Although those networks were not specifically trained for face recognition, they were able to classify face identity to some extent, with an

accuracy of 68.43% and 68.30% for the object AlexNet and object VGG 16, respectively (chance accuracy: 50%).

Data Analysis

Preprocessing. MRI data were preprocessed using fMRIPrep version 1.4.1 (8). T1-weighted images were corrected for intensity non-uniformity (9) and skullstripped using antsBrainExtraction.sh. High resolution cortical surfaces were reconstructed with FreeSurfer (10) using both T1-weighted and T2-weighted images, and then normalized to the fsaverage template based on sulcal curvature (11). Functional data were slice-time corrected using 3dTshift (12), motion corrected using MCFLIRT (13), distortion corrected using fieldmap estimate scans (one for each session), and then resampled to the fsaverage template based on boundary-based registration (14). After these steps, functional data were in alignment with the fsaverage template based on cortical folding patterns. The following confound variables were regressed out of the signal in each run: six motion parameters and their derivatives, global signal, framewise displacement (15), 6 principal components from a combined cerebrospinal fluid and white matter mask (aCompCor) (16), and up to second-order polynomial trends.

Searchlight Hyperalignment. All three imaging datasets were hyperaligned (17–20) based on responses to the Grand Budapest Hotel (Figure S1). We first built a common model information space where patterns of fMRI responses to the Grand Budapest Hotel movie were aligned across subjects. Whole-cortex transformation matrices for each individual were calculated using a searchlight-based algorithm to project each participant's cortical space into the common model information space. Transformation matrices were calculated for all 15 mm radius searchlights in each brain using an iterative procedure and Procrustes alignment, and then aggregated into a single matrix for each hemisphere. Transformation matrices for each participant were used to transform their Hyperface and dynamic localizer data into the common model space, so that all three imaging datasets were functionally aligned in the same common model information space.

Reweighting Features Prior to RSA. RSA has the strong assumption that all features contribute equally to generate an RDM (e.g. all cortical vertices in a searchlight are equally important when computing pattern similarity between two conditions) (21, 22). We tested whether relaxing this assumption might yield larger DCNN-neural correlations. We developed a novel approach that best matched the DCNN features to the brain responses before performing RSA. First, the DCNN features were matched to the brain responses by performing singular value decomposition (SVD) on the covariance matrix between the DCNN features and brain features. This step corresponds to bringing the DCNN features into a space with reduced dimensionality N that best matches brain responses. Second,

ridge regression was used to predict responses for each brain feature (vertex) from the reduced-dimension DCNN features. Third, the fitted ridge model was used to generate predictions of the brain responses from DCNN features for left-out data. This procedure ultimately yields a reweighted subspace of the original DCNN feature space that best predicts brain features. Finally, these predicted features were used to generate RDMs, which were then analyzed as reweighted DCNN RDMs.

This process was performed using nested cross-validation. The 12 scanning runs were separated into two sets (11 training runs and one test run) for both the brain and DCNN (ArcFace) responses. The training runs were used to estimate the transformation for the shared components and the ridge regression parameters. The hyperparameters (number of dimensions N and ridge regularization parameter α) were chosen based on a nested leave-one-run-out loop within the 11 training runs. We performed a grid search on N and α by testing 18 evenly distributed values from 5 to 90 for N, and 29 evenly distributed values from 10^{-7} to 10^{7} on a logarithmic scale for α. The regression model was trained to yield the best R^2. The best model was then used to generate RDMs for the left-out run. This analysis was performed within each searchlight.

Cross-subject Identity Decoding. The cross-subject identity decoding analysis was done as a binary classification task with a simple one-nearest-neighbor classifier across all searchlights (10 mm radius). We performed the analysis using a split-half cross-validation scheme. That is, we divided the subjects into two groups (training and test). For each face and each group, we computed an average response pattern across subjects in the group. We assessed whether the average pattern of a face for the training group is more similar to that of the same face for the test group compared to a different face (2-alternative forced choice; chance accuracy = 50%). We repeated this for all pairs of faces and averaged the accuracy. Because there are many different ways to split the subjects into two groups, we also repeated the split-half procedure 100 times and averaged the accuracy across repetitions.

We also performed a similar analysis using leave-one-subject-out cross-validation instead of split-half cross-validation. Each time, we computed the average response patterns of 20 subjects and compared with the left-out test subject.

Note that the RSA analysis is based on the average of all 21 subjects rather than half of the subjects or a single subject, and the data quality is superior to the average patterns used in this classification analysis.

Variance Partitioning Analysis. Variance partitioning analysis based on multiple linear regression was used to quantify the unique contributions of each model taking into consideration the contribution of other models. In detail, this analysis was done to tease apart the contribution of face- and object-trained DCNNs in explaining variance of the behavioral arrangement task RDM, and to separate the relative

contributions of face- and object-trained DCNNs, as well as the behavioral arrangement task performance in explaining variance of the neural RDM for a given face-selective region. In the first analysis (comparing behavioral RDMs and two types of DCNN RDMs), the off-diagonal elements of the behavioral RDM were assigned as the dependent variable, and the off-diagonal elements of the two DCNN models were assigned as independent variables (predictors). In the second analysis (comparing neural, behavioral, and two types of DCNN RDMs), the dependent variable was the off-diagonal elements of the mean RDM across vertices in face-selective cortex, and the independent variables were the off-diagonal elements in the two DCNN models and the behavioral RDM. Both analyses were performed within runs and the variance was averaged across runs as the final result. To obtain unique and shared variance for each model, in the former analysis, three multiple regression analyses were run in total. The three analyses included the full model that had both DCNNs as predictors and two reduced models that contained an individual DCNN as the predictor. For the latter analysis, seven multiple regression analyses were run including one full model that had behavioral and two DCNN models as predictors, as well as six reduced models that had either combinations of two models from the three (behavioral, two DCNNs) or one individual model alone as the predictor. By comparing the adjusted explained variance (adjusted R^2) of the full model and the reduced models, variance that was explained by each model independently could be inferred (23, 24). The variance partitioning analysis was conducted using the "vegan" package of R (https://cran.r-project.org/web/packages/vegan/vegan.pdf).

RSA between the behavioral arrangement task RDMs and RDMs of DCNNs were carried out for each run and correlations were averaged across runs. "Best layers" for the behavioral arrangement task were layers that had the strongest correlations with the behavioral RDMs, including _plus45 of ArcFace, pool5 of face AlexNet, block5_conv2 of face VGG16, fc2 of object AlexNet, and fc2 of object VGG16. In the run-by-run RSA analysis between neural RDMs and RDMs of DCNNs, "best layers" were _plus42 of ArcFace, conv1 of face AlexNet, block2_pool of face VGG16, conv2 of object AlexNet, and block3_pool of object VGG16. These layers were very similar with the "best layers" in the RSA analysis using all 707 stimuli (_plus42 of ArcFace, conv1 of face AlexNet, block2_conv1 of face VGG16, conv2 of object AlexNet, and block3_conv3 of object VGG16).

Multidimensional Scaling. Multidimensional scaling (MDS) was used to visualize the representational geometry of face stimuli for different DCNNs, behavioral arrangements, and neural ROIs. Pairwise correlation distance matrices were computed for stimuli pairs in each of the spaces in a run-by-run manner. Metric MDS was used to project the stimuli onto a 2-dimensional space and these locations were color-coded based on the behavioral ratings. MDS was also used to visualize representational geometry of the 18 face-selective ROIs (Figure S5). Pairwise correlation distance matrices were

computed for face ROI pairs based on the 707-by-707 RDM of each face ROI in a second-order RSA manner.

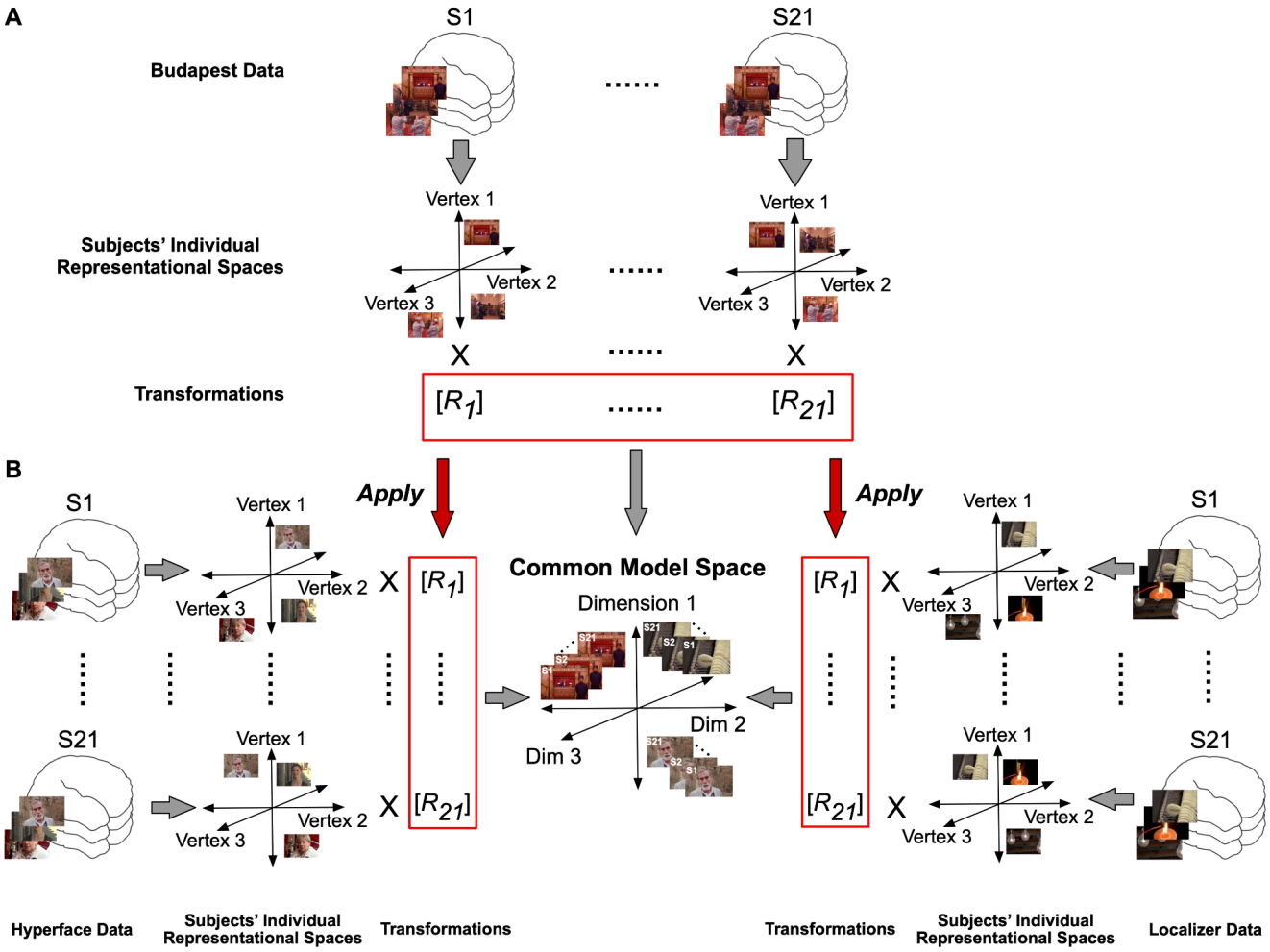

Figure S1. Schematic of the hyperalignment procedure. **A.** Transformation matrices were calculated by hyperaligning each participant's responses to the Grand Budapest Hotel movie to the common model space. **B.** The transformation matrix derived from the Grand Budapest Hotel movie for each participant was then applied to the Hyperface data. **C.** The transformation matrix derived from the Grand Budapest Hotel movie for each participant was also applied to the localizer runs. Thus, after hyperalignment, the Budapest data, the Hyperface data, and the localizer data were all functionally aligned to the common model space.

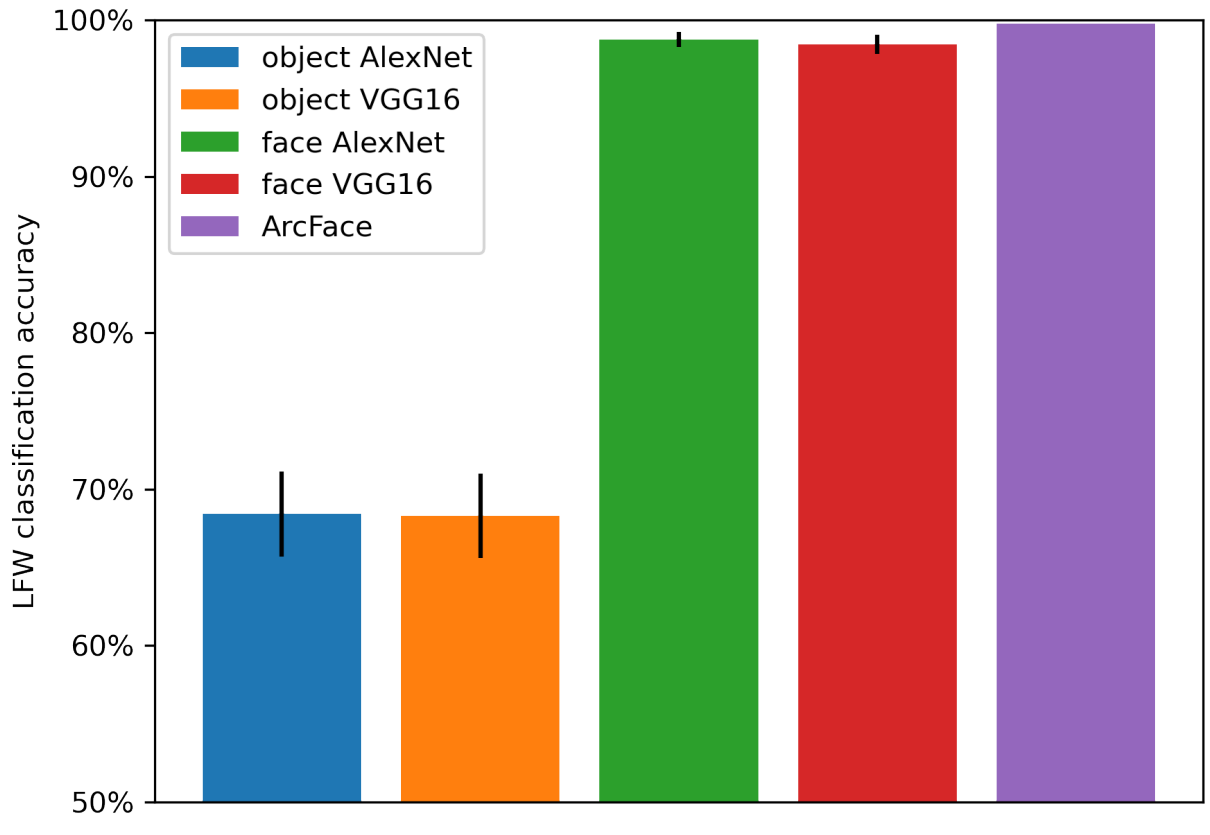

Figure S2. DCNN performance on face recognition. Using the Labeled Faces in the Wild (LFW) dataset, we assessed the face recognition performance of the DCNNs used in our analysis. The benchmarking task was to tell if two faces were the same person or not based on the similarity of their embeddings, and thus the chance accuracy was 50%. The dataset was divided into 10 folds and a leave-one-out cross-validation was used to determine the threshold (i.e. the similarity level to determine if two faces are the same or not). The accuracy was 68.43% and 68.30% for the object AlexNet and object VGG16, respectively, and 98.75% and 98.45% for the face AlexNet and face VGG16, respectively. The accuracy was 99.77% for the pre-trained ArcFace, which was provided by the InsightFace package. Error bars are the standard deviation of the accuracy across 10 cross-validation folds.

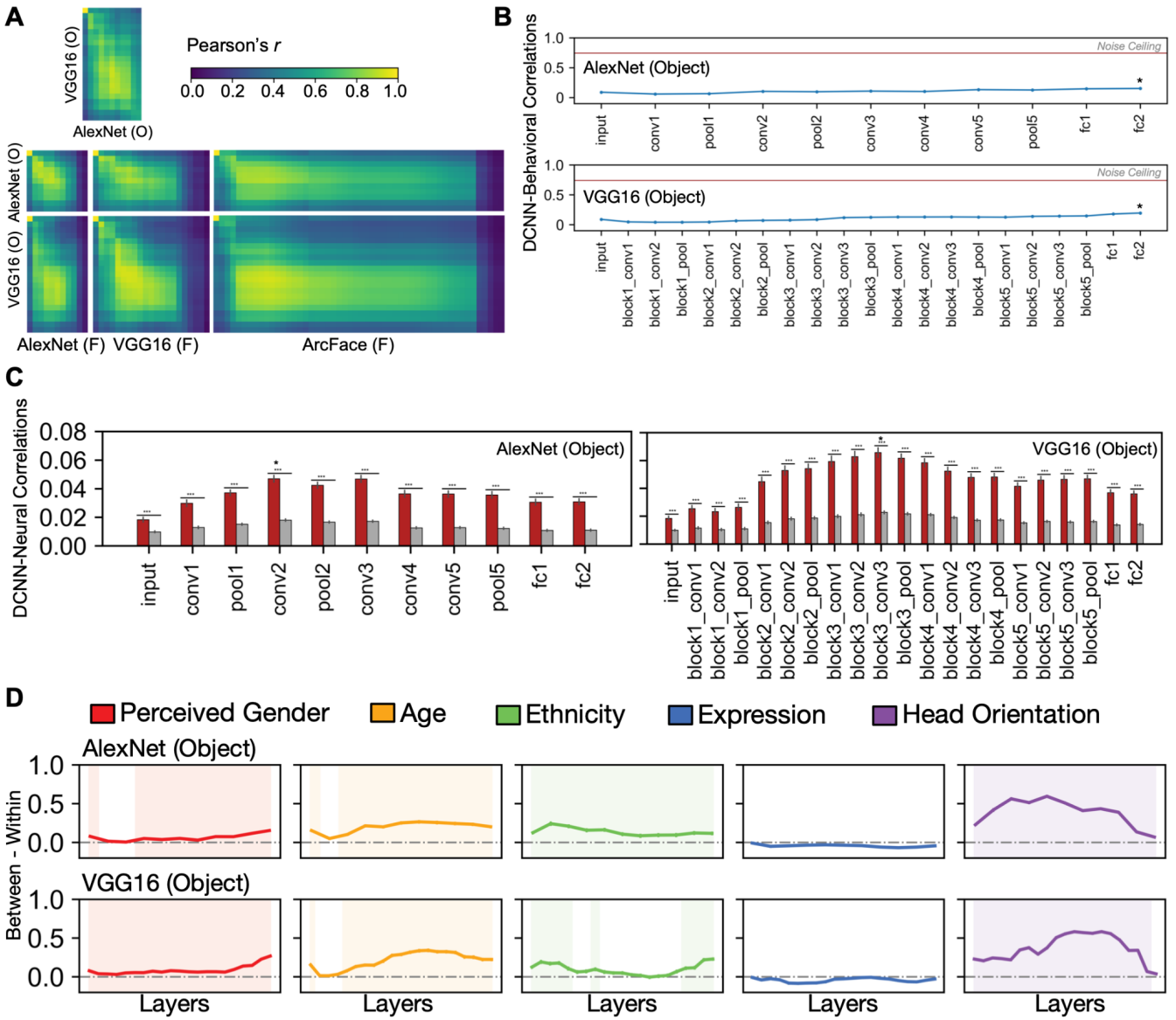

Figure S3. Object-DCNN representations. **A.** Correlations in each pair of layers in the two object-trained AlexNet and VGG16, and across the six face- (ArcFace, AlexNet, and VGG16) and object-DCNN pairs. **B.** Mean correlations across participants and runs between the behavioral RDM and DCNN RDM in each layer in the two object-DCNNs. The star marks the layer that has the highest correlation with the behavioral task in each DCNN. The red horizontal line in each subplot represents the mean noise ceiling of the behavioral arrangement task across runs. **C.** Average correlations for face-selective regions and non-face-selective regions for each layer in the two object-DCNNs. Regions, significance, and the color code were defined the same as in Figure 3C. Stars indicate the layers that had the largest correlations. **D.** Difference in the between- and within-group distance of perceived gender (red), age (orange), ethnicity (green), expression (blue), and head orientation (purple) in representational geometries of the behavioral arrangement task, each layer of the two object-trained DCNNs (AlexNet,

VGG16). These differences were calculated within each run and then averaged across runs. Shaded layers and ROIs show significant differences in the between- versus within-group test ($p < 0.05$, permutation test, one-tail). Significance of the difference was estimated based on a random permutation test randomizing the stimulus labels. Error bars represent one standard error of the mean estimated by bootstrap resampling stimuli. *** $p < 0.001$.

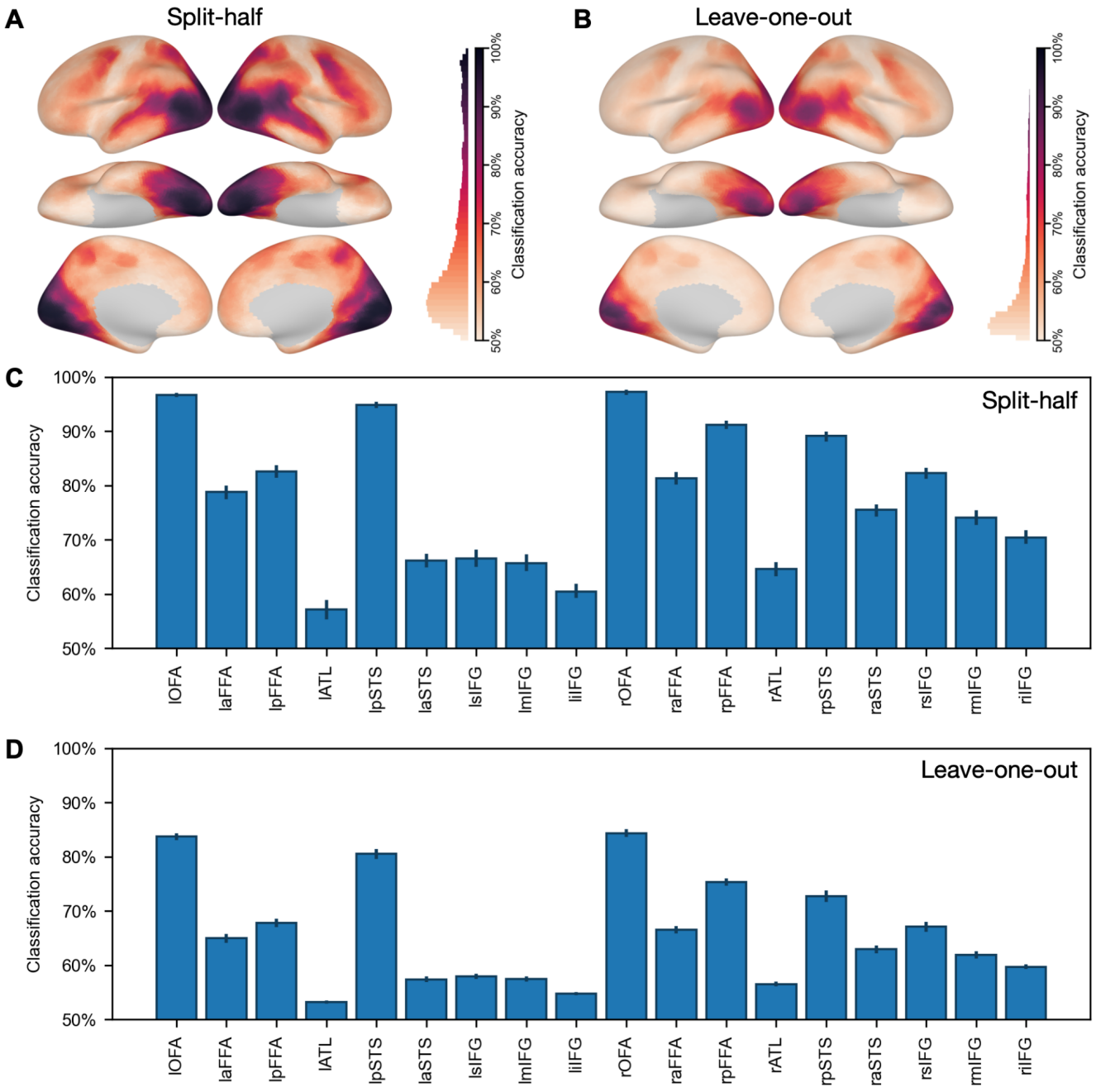

Figure S4. Between-subject identity decoding accuracies. A. Searchlight analysis of between-subject identity classification. This classification analysis was a binary classification task with a simple one-nearest-neighbor correlation-based classifier. We used a searchlight radius of 10 mm, similar to our other analyses. We used a split-half cross-validation scheme. That is, we divided the subjects into two groups (training and test). The classification task is binary classification (2-alternative forced choice). For each face and each group, we computed an average response pattern across subjects in the group. Each time, we assessed whether the test group's average pattern of a certain face was more similar to the training group's pattern to the same face than to a different face (chance accuracy = 50%). We repeated

this procedure for each pair of faces and averaged the accuracy across face pairs. Because there are multiple ways to split subjects into groups, we also repeated the split-half procedure 100 times and averaged the accuracy across repetitions. Classification accuracy was high for visual and face-selective regions. B. Similar analysis based on a leave-one-subject-out cross-validation scheme. Each time, we computed the average response patterns of 20 subjects and compared these patterns with those of the left-out test subject. The accuracies were lower compared with A, because the data from only a single subject was used as test data, but nonetheless clearly significant. Note that even in the split-half condition, there were only 10 or 11 subjects per group. Our RSA analysis was based on the average across all 21 subjects, and thus the average patterns have higher quality than those used in A. C. Split-half classification accuracy in face-selective regions. D. Leave-one-subject-out classification accuracy in face-selective regions.

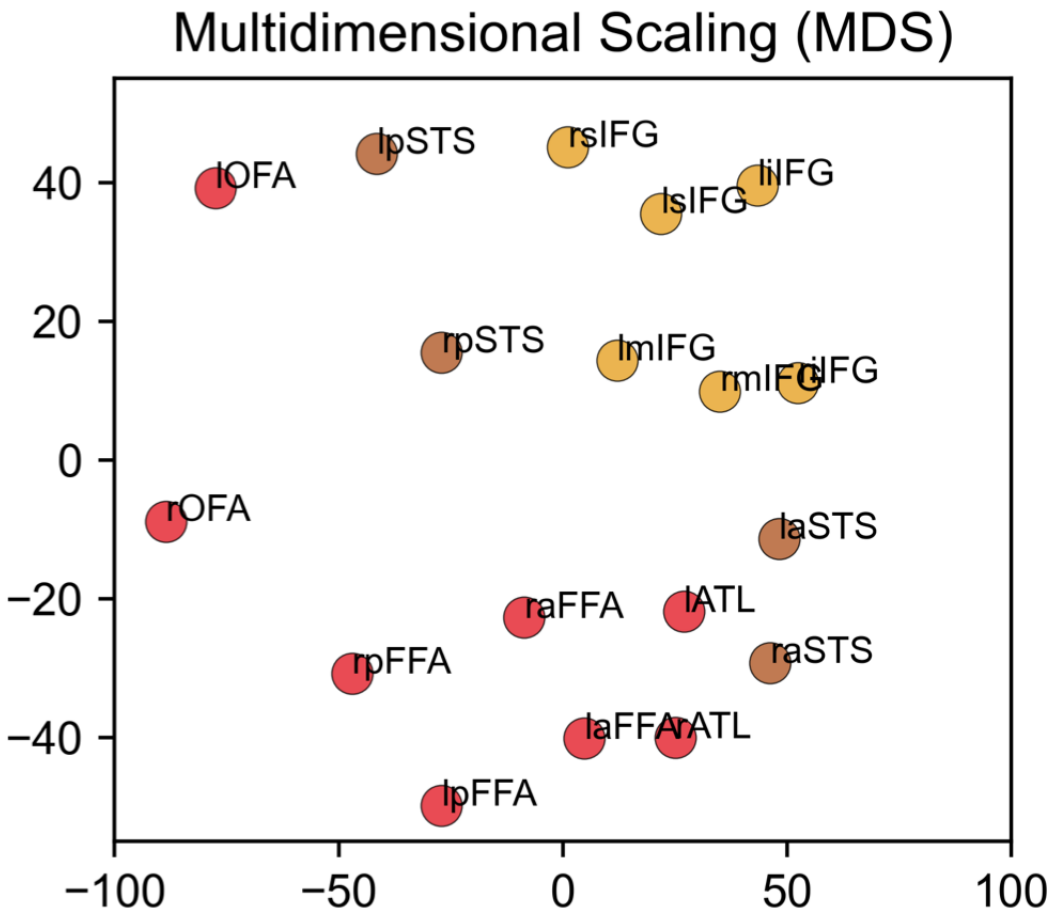

Figure S5. MDS plot of distances between RDMs in 18 bilateral face-selective ROIs. Each dot represents one face-selective ROI (10 mm radius with the peak at the center). These face ROIs were localized using the faces-vs-objects contrast with the dynamic localizer. Red markers denote ROIs in ventral temporal cortex, orange markers denote ROIs in lateral temporal cortex, and yellow markers denote ROIs in frontal cortex. The clustering of ROIs follows the anatomical and functional hierarchy of the face-processing network.

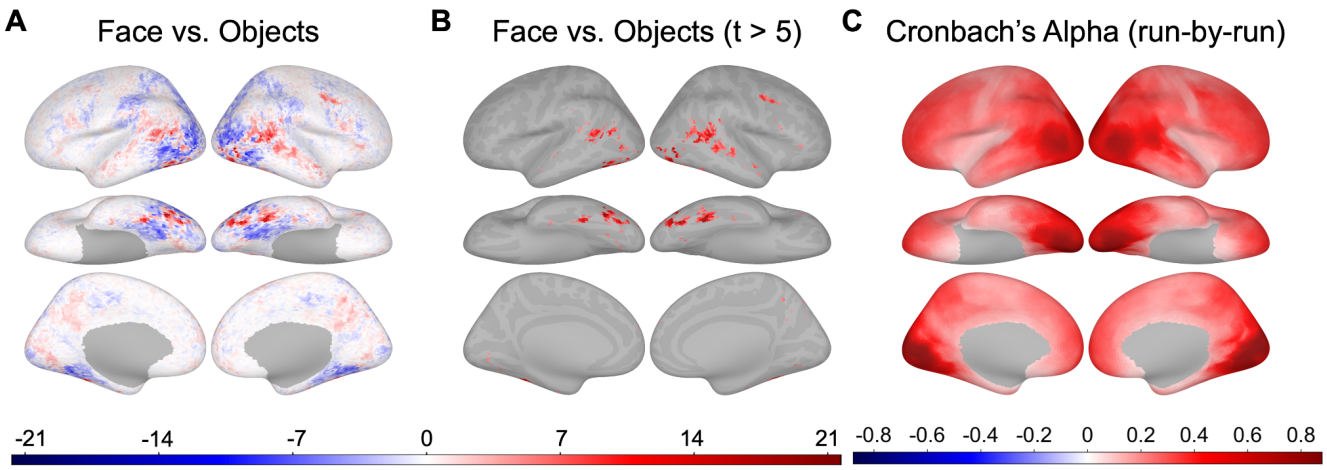

Figure S6. Contrast maps and noise ceilings. **A.** The whole-brain faces-vs-objects contrast *t*-map was calculated for each individual using hyperaligned dynamic localizer data. Then the *t*-maps were averaged across participants. The darker the red color was, the stronger the responses were to faces. **B.** The face-selective regions thresholded at $t > 5$. **C.** Whole-brain Conbach's alpha map (run-wise). Reliabilities of neural RDMs across participants were calculated for each searchlight for each run, and the mean Cronbach's alpha map was derived by averaging maps across runs.

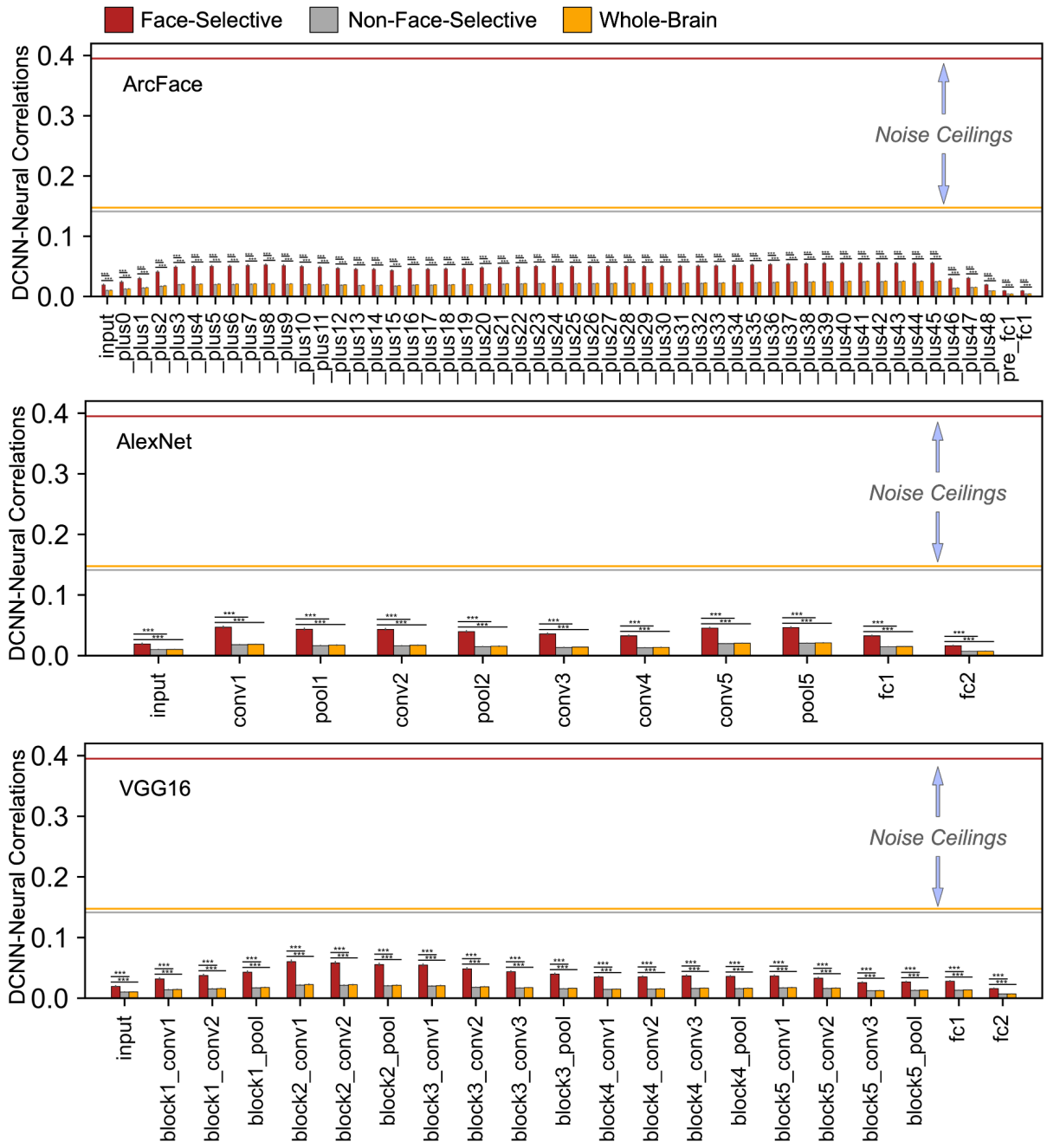

Figure S7. RSA results in layers of three face-DCNNs with noise ceilings. Bar plots of the mean DCNN-neural correlations in face-selective, non-face-selective, and whole-brain regions in layers of the three face-DCNNs with noise ceilings. Horizontal lines in each panel represent the mean noise ceiling in face-selective, non-face-selective, and whole-brain searchlights. These lines are coded with the same colors in the legends. In all three panels, error bars represent one standard error of the mean estimated by bootstrap resampling stimuli. Significance of the difference was estimated based on a permutation test randomizing the stimulus labels. *** $p < 0.001$.

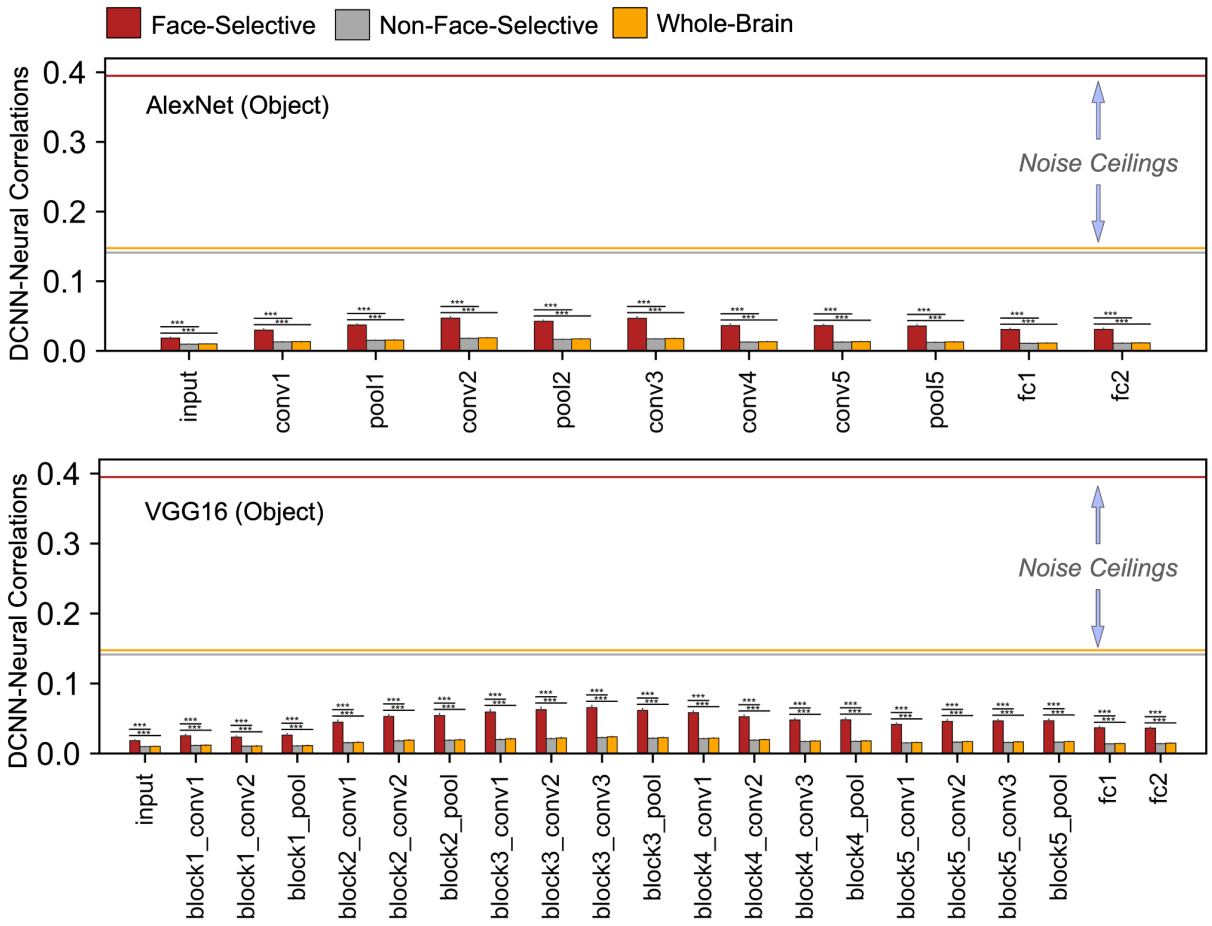

Figure S8. RSA results in layers of two object-DCNNs with noise ceilings. Bar plots of the mean DCNN-neural correlations in face-selective, non-face-selective, and whole-brain regions in layers of the two object-DCNNs with noise ceilings. Horizontal lines in each panel represent the mean noise ceiling in face-selective areas, non-face-selective areas, and across the whole brain. These lines are color coded according to the legend at top. In panels A, B, and C, the error bars indicate standard error of the mean estimated by bootstrap resampling stimuli. Significance of the difference was assessed based on a permutation test randomizing the stimulus labels. *** $p < 0.001$.

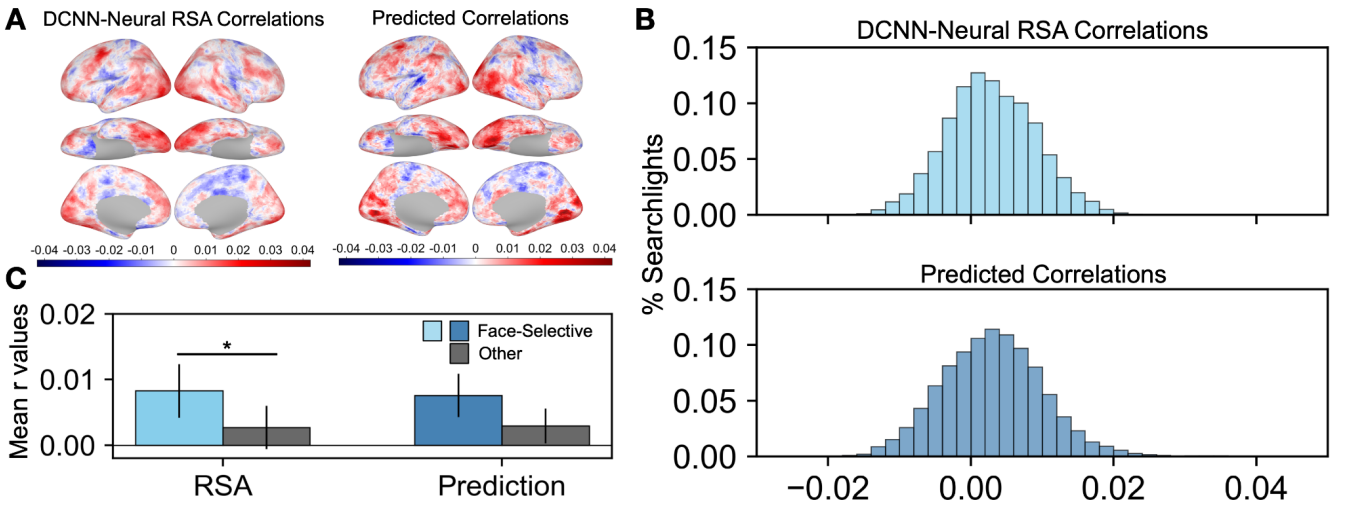

Figure S9. RSA and prediction analysis. **A.** Left panel shows the whole-brain ArcFace-neural RSA map (run-wise). Right panel shows the whole-brain prediction correlation map. Because prediction analysis was done in a leave-one-run-out cross-validation procedure (see Material & Methods for details), the RSA was recalculated using a run-wise procedure for comparison. In detail, RDMs were computed for each run in each searchlight, and the resulting correlations were averaged across runs to generate the map in the left panel. **B.** Histograms of ArcFace-neural RSA correlations calculated using data in each run and averaged across all twelve runs are plotted at bin size of 0.002 in the upper panel. Histograms of correlations between predicted and original RDMs are plotted at the same bin size in the lower panel. **C.** Average correlations of face-selective regions ($t > 5$) and non-face-selective regions ($t <= 5$) using RSA and prediction maps. The error bars indicate standard error of the mean estimated by bootstrap resampling of the stimuli (5000 bootstraps). Significance of the difference between the two bars was assessed using a permutation test randomizing the stimulus labels (5000 permutations). * $p <$ 0.05.

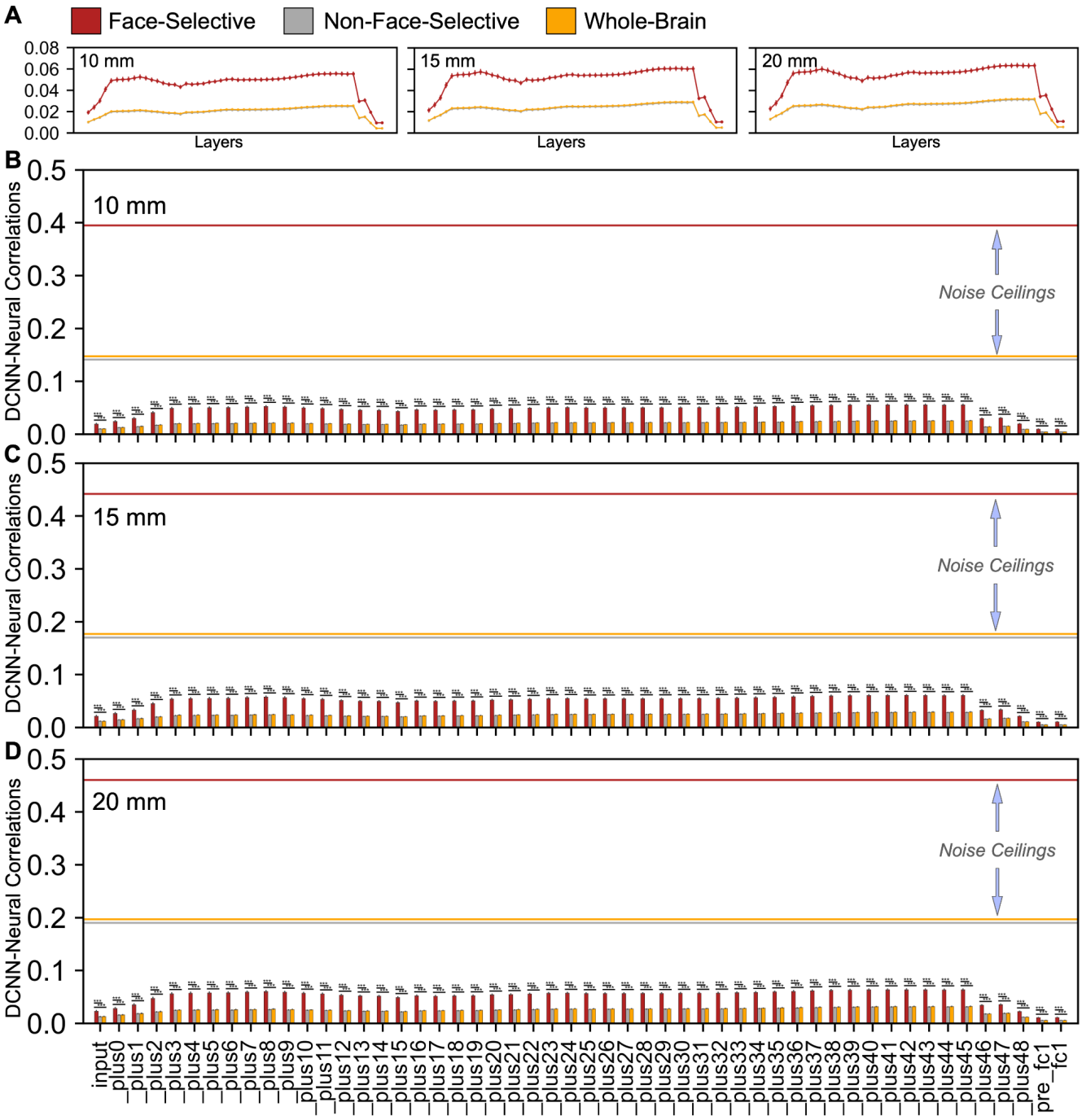

Figure S10. RSA results in layers of ArcFace with different searchlight sizes. **A.** Line plots of the mean DCNN–neural correlations in face-selective regions (red), non-face-selective regions (gray), and across the whole brain (orange) for each layer of ArcFace with a searchlight radius of 10 mm, 15 mm, and 20 mm (note that the gray and orange lines are largely overlapped). **B, C, & D.** Bar plots of the same values as in panel A (mean DCNN-neural correlations in face-selective, non-face-selective, and whole-brain regions in layers of ArcFace) with noise ceilings. Horizontal lines in each panel represent the mean noise ceiling in face-selective, non-face-selective, and whole-brain searchlights. These lines are coded

with the same colors in the legends. In all four panels, error bars represent one standard error of the mean estimated by bootstrap resampling stimuli. Significance of the difference was estimated based on a permutation test randomizing the stimulus labels. *** $p < 0.001$. Larger searchlight sizes slightly increased DCNN-neural correlations, but the overall correlations remained weak.

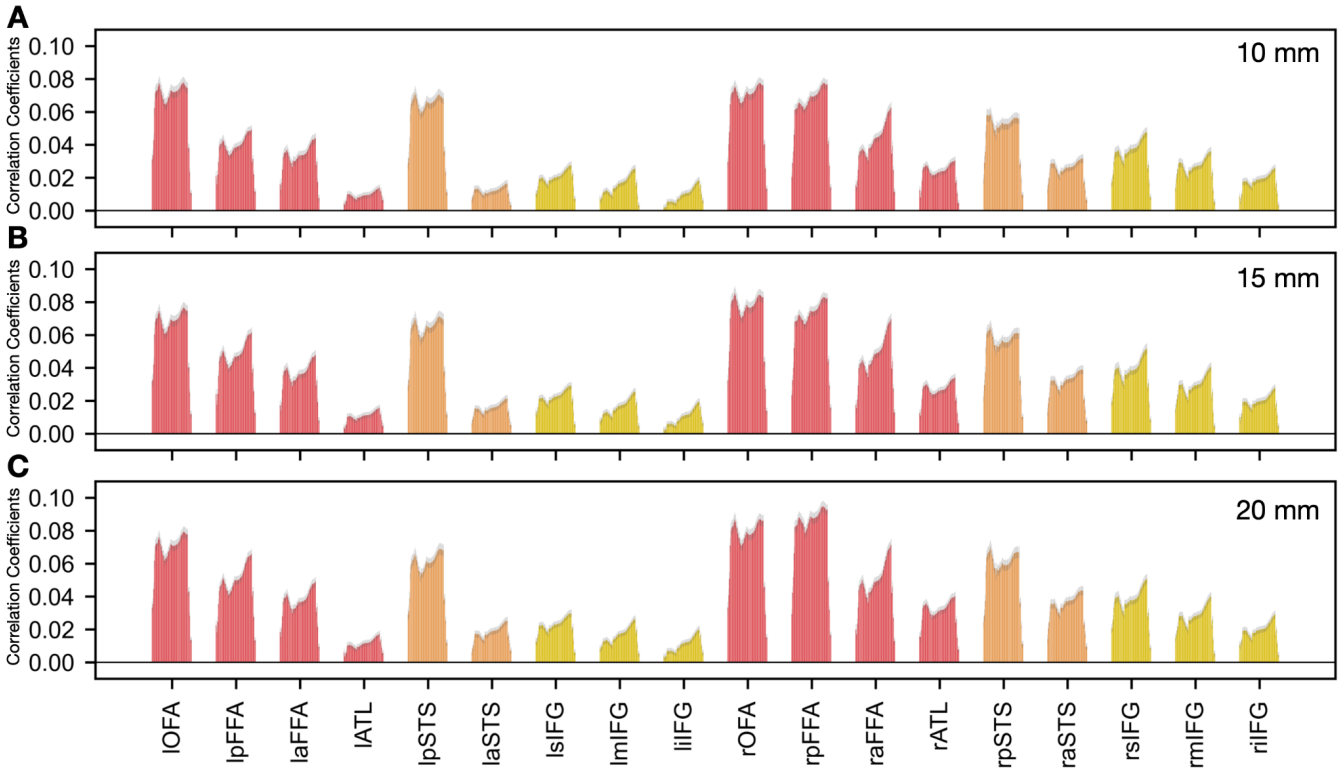

Figure S11. RSA results in individual face-selective ROIs with different ROI sizes. **A, B, & C.** RSA correlations for each layer of ArcFace in individual face-selective ROIs with ROI size of r = 10 mm, 15 mm, and 20 mm. Error bars indicate standard error of the mean estimated by bootstrap resampling the stimuli (1000 bootstraps).

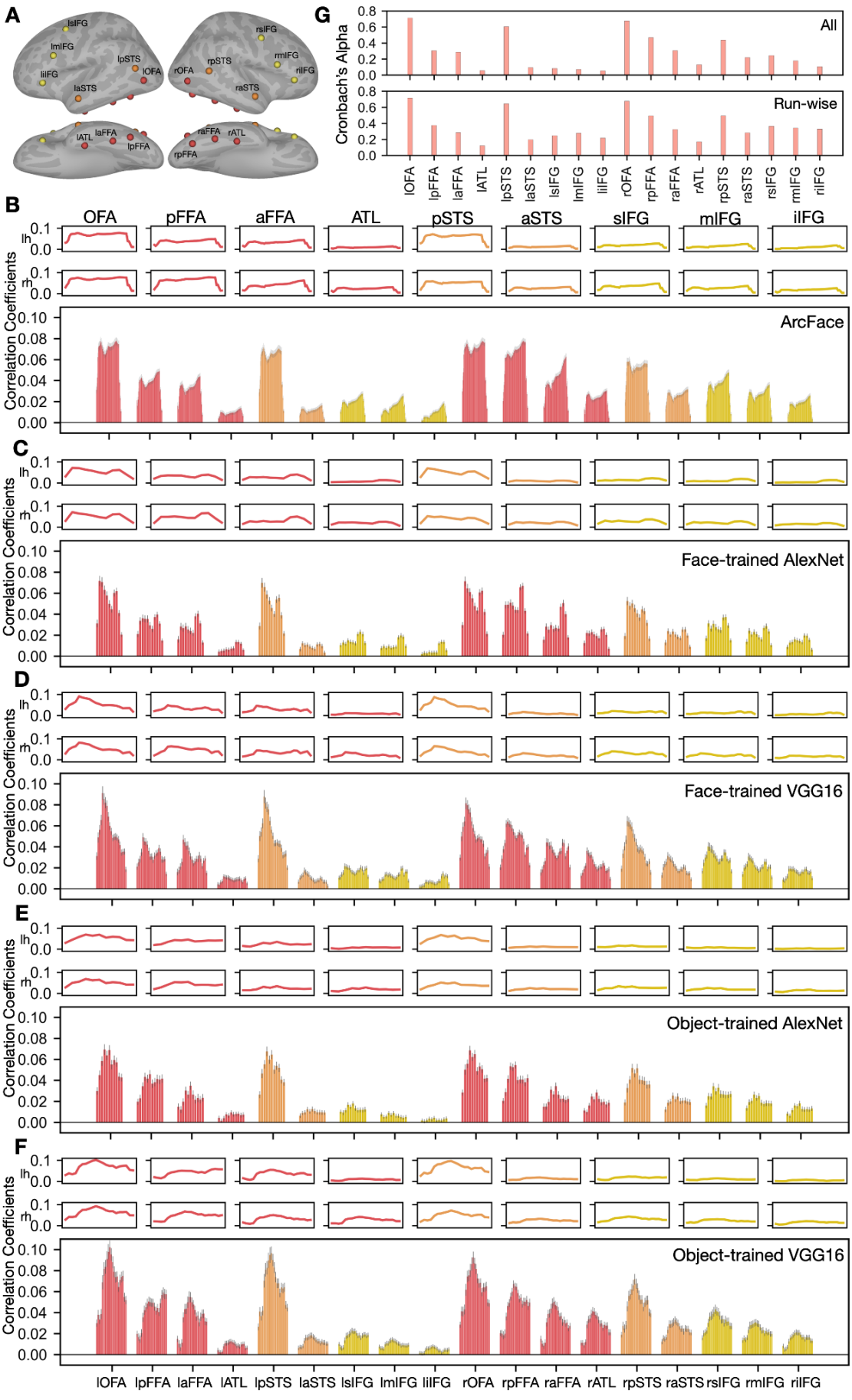
A
lsIFG
lmIFG
lpSTS
lOFA
liIFG
laSTS
lATL
laFFA
lpFFA
rpFFA
raFFA
rATL
rpSTS
rOFA
rsIFG
rmIFG
riIFG
raSTS
G
Cronbach's Alpha
All
Run-wise
lOFA lpFFA laFFA lATL lpSTS laSTS lsIFG lmIFG liIFG rOFA rpFFA raFFA rATL rpSTS raSTS rsIFG rmIFG riIFG
B
OFA pFFA aFFA ATL pSTS aSTS sIFG mIFG iIFG
lh
rh
Correlation Coefficients
ArcFace
C
Face-trained AlexNet
D
Face-trained VGG16
E
Object-trained AlexNet
F
Object-trained VGG16
lOFA lpFFA laFFA lATL lpSTS laSTS lsIFG lmIFG liIFG rOFA rpFFA raFFA rATL rpSTS raSTS rsIFG rmIFG riIFG

Figure S12. RSA results and noise ceilings in individual face-selective ROIs. **A.** Peak coordinates of 18 bilateral face-selective ROIs across both hemispheres. Red markers denote ROIs in ventral temporal cortex, orange markers denote ROIs in lateral temporal cortex, and yellow markers denote ROIs in frontal cortex. **B, C, D, E, & F.** RSA correlations for each layer of ArcFace, AlexNet (face-trained), VGG16 (face-trained), AlexNet (object-trained), VGG16 (object-trained) respectively in individual face-selective ROIs. Error bars indicate standard error of the mean estimated by bootstrap resampling the stimuli (1000 bootstraps). The line plots at the top of each panel match the values from the bar plot below. The line plots are included to more clearly display the profile of correlations across DCNN layers. **G.** Noise ceilings (Cronbach's alphas) in each face-selective ROI. Noise ceilings were calculated using all stimuli at once (All, upper panel), and within each run and averaged across runs (Run-wise, lower panel). The DCNN-neural correlations were much lower than the noise ceiling of that ROI across all layers.

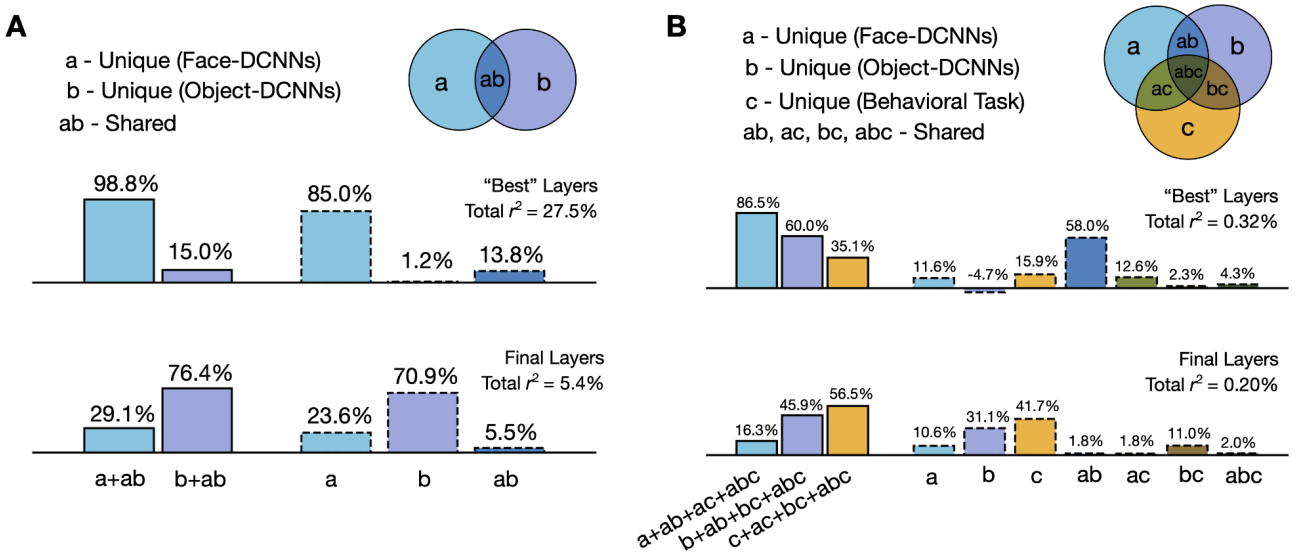

Figure S13. Variance partitioning analysis. **A.** Results of the variance partitioning analysis on the RDMs of face-DCNNs (light blue) and the object-DCNNs (purple) explaining variance of the behavioral task representations. Variance noted as unique or shared (numbers above the bars) are percentages of the total variance explained by both models combined (total r^2). The first two bars (solid edges) show the total variance that each DCNN category (face or object) explained, and the rest of the bars (dotted edges) show the unique and shared variance that each DCNN category explained. In the upper panel, face- and object-DCNN RDMs were the mean of the RDMs in layers with the highest correlations with the behavioral RDM ("best" layers, marked with stars in panel B. _plus45, pool5, block5_conv2, fc2, and fc2 in ArcFace, face AlexNet, face VGG16, object AlexNet, and object VGG16 accordingly). In the lower panel, face- and object-DCNN RDMs were the mean of the RDMs in the final layer. **B.** Results of the variance partitioning analysis on the RDMs of face-DCNNs (light blue), object-DCNNs (purple), and behavioral task (yellow) explaining variance of the neural representations in face-selective areas ($t >$ 5). Variance noted as unique or shared (numbers above the bars) are percentages of the total variance explained by all three models combined (total r^2). The first three bars (solid edges) stand for the total variance that DCNNs and the behavioral task explained, and the rest of bars (dotted edges) stand for the unique and shared variance that each DCNN category and the behavioral task explained. In the upper panel, face- and object-DCNN RDMs were the mean of the RDMs in layers that had the highest correlations with the neural RDM in the run-by-run analysis ("best" layers, _plus42, conv1, block2_pool, conv2, block3_pool in ArcFace, face-trained AlexNet, face-trained VGG16, object-trained AlexNet, and object-trained VGG16 accordingly). In the lower panel, face- and object-DCNN RDMs were the mean of the RDMs in the final layer.

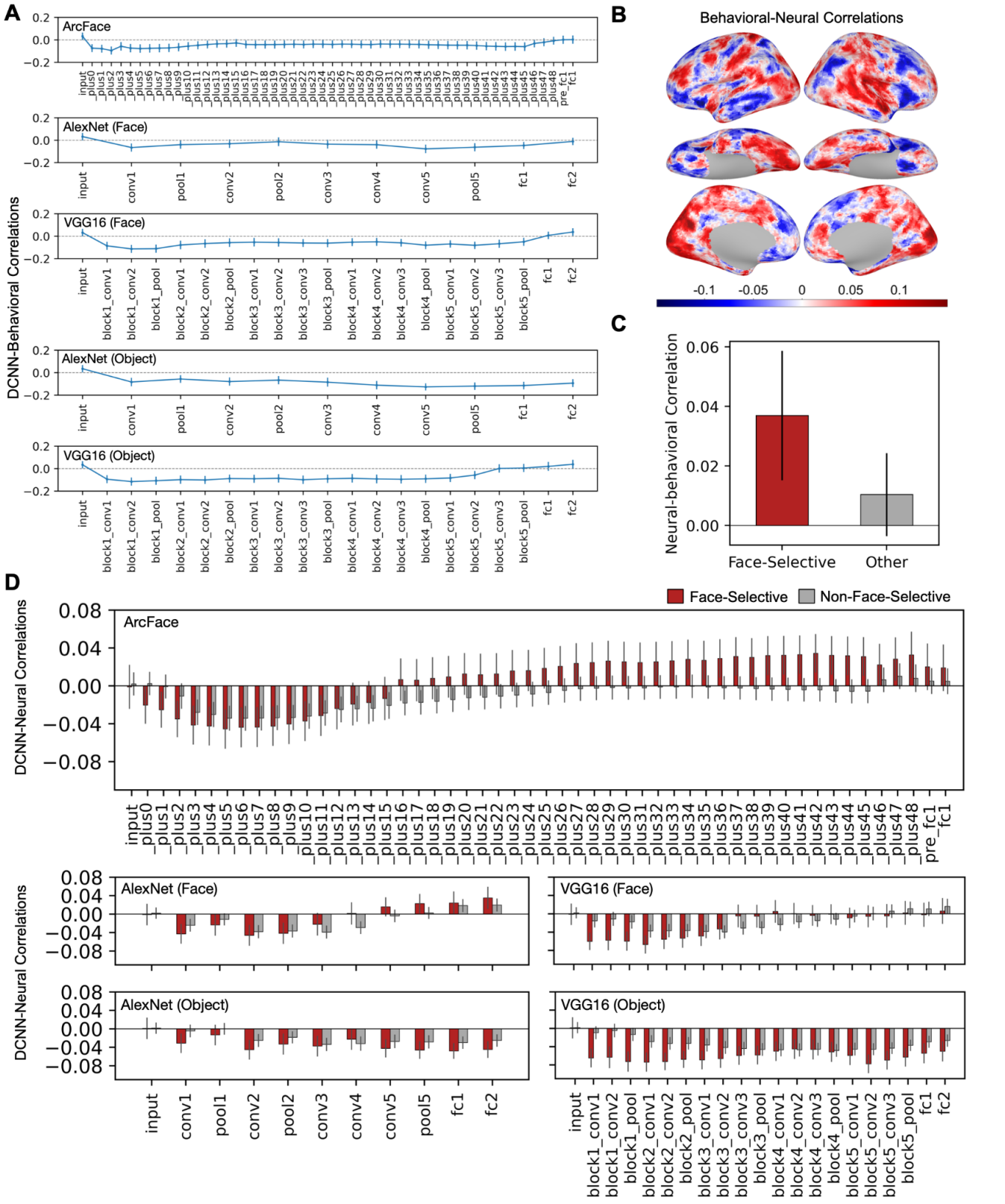

Figure S14. Correlations based on stimuli in behavioral pairwise comparison experiments. **A.** Mean correlations between RDMs based on behavioral similarity ratings and RDMs based on DCNN features in each layer, across participants and the four groups (white females ages 21-30, generally happy, facing

right, 18 clips; black males age 21-30, neural expression, facing right, 15 clips; black females age 21-40 with happy or neutral faces, facing right, 16 clips; white males age 21-30 with neutral faces, facing right, 20 clips). All correlations are below 0.039 and only a small fraction of the noise ceiling of 0.86. The gray line stands for zero. **B.** The neural-behavioral correlation values in the cortex. **C.** The mean behavioral-neural correlations in face-selective (red) and non-face-selective (gray) areas. Note that the behavioral-neural correlations in face-selective areas for individuation are comparable (non-significant difference) to the behavioral-neural correlations based on the arrangement task which reflected mostly categorical attributes. The error bars indicate standard error of the mean estimated by bootstrap resampling the stimuli. **D.** Bar plots of the mean DCNN-neural correlations, restricted to the within-group RDMs for the faces selected to minimize differences in categorical attributes in face-selective and non-face-selective regions in layers of the three face-DCNNs and two object-DCNNs. In all five panels, error bars represent one standard error of the mean estimated by bootstrap resampling stimuli.

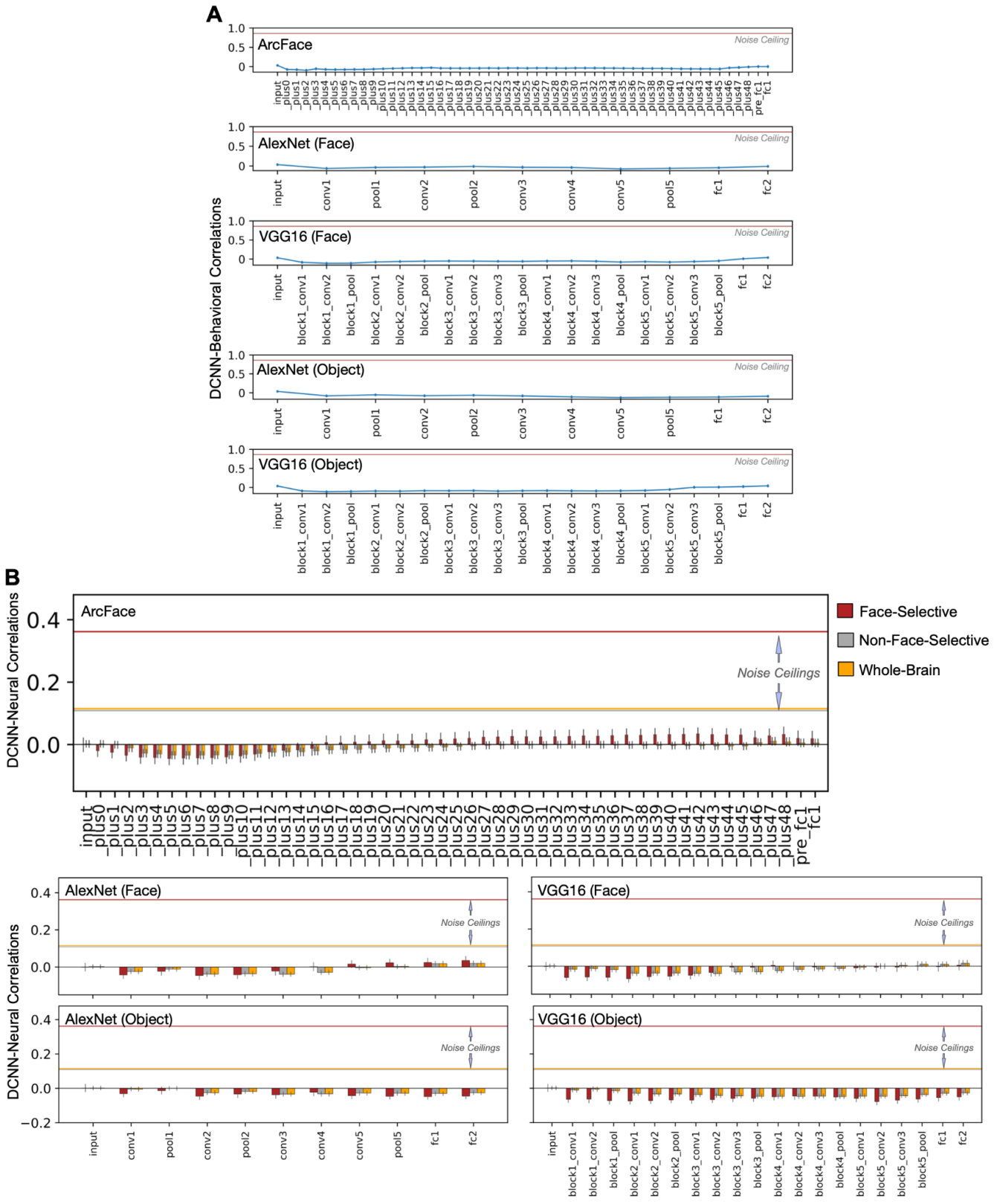

Figure S15. Correlations based on stimuli in behavioral pairwise comparison experiments with noise ceilings. **A.** Mean correlations between RDMs based on behavioral similarity ratings and RDMs based on DCNN features in each layer, across participants and the four groups. Noise ceilings based on the

cross-participant reliability of the similarity ratings (Cronbach's alpha) were over 0.8, showing that DCNNs captured only a very small portion of the meaningful variance in behavioral ratings reflecting individuation of faces. **B.** Bar plots of the mean DCNN-neural correlations, restricted to the within-group RDMs for the faces selected to minimize differences in categorical attributes, in face-selective, non-face-selective, and whole-brain regions in layers of the three face-DCNNs and two object-DCNNs with noise ceilings. Noise ceilings based on the cross-participant reliability of the neural RDMs (Cronbach's alpha) were over 0.35 for the face-selective areas and over 0.1 for the non-face-selective areas and the whole brain, showing that DCNNs captured only a very small portion of the meaningful variance in neural representational geometry even when that variance is restricted to information for individuation.

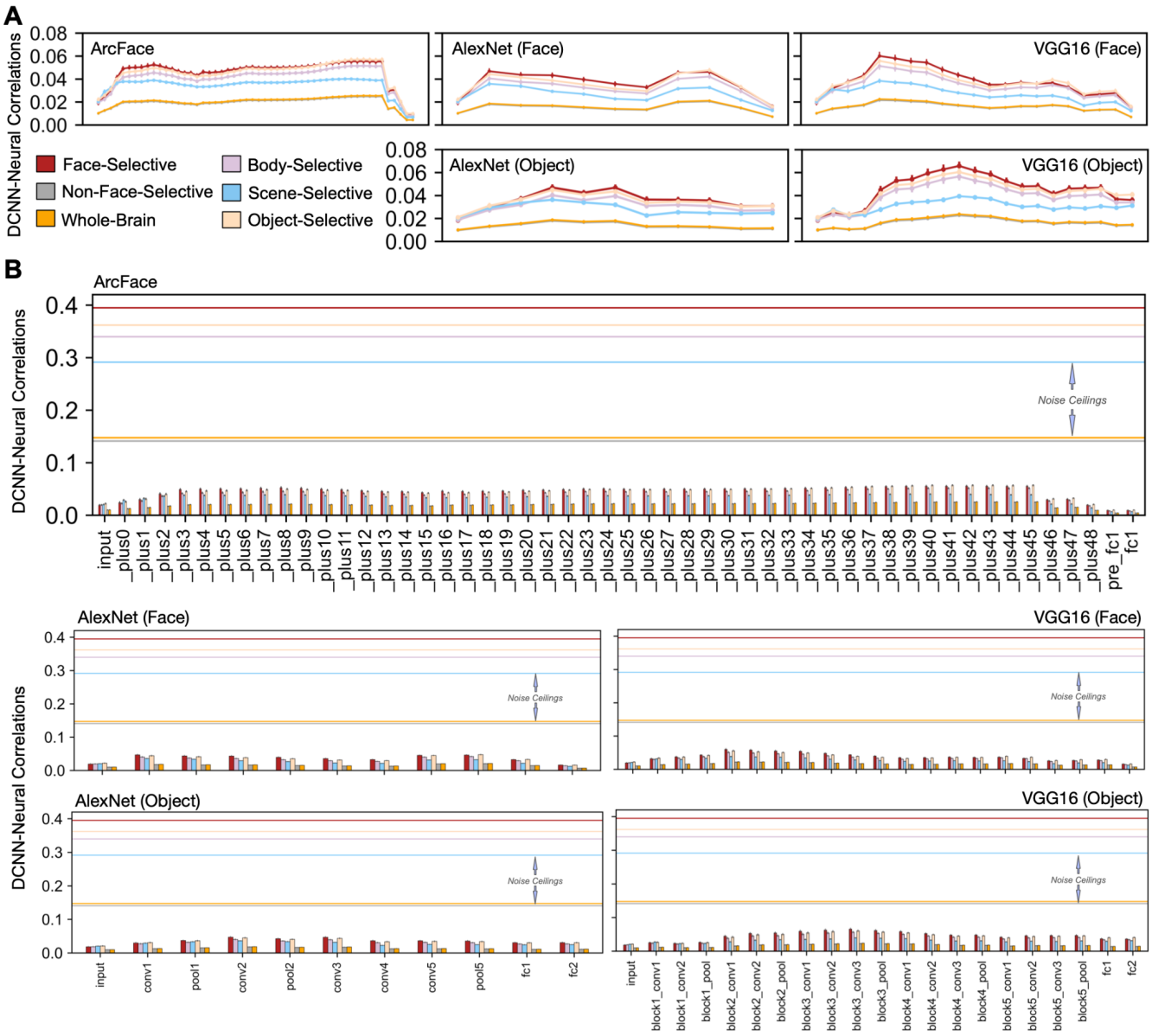

Figure S16. RSA results for high-level category-selective areas across layers of DCNNs with noise ceilings. **A.** Line plots of the mean DCNN–neural correlations in face-selective regions (red), body-selective regions (purple), scene-selective regions (blue), object-selective regions (yellow), non-face-selective regions (gray), and across the whole brain (orange) for each layer of the five DCNNs. **B.** Bar plots of the same values as in panel A (mean DCNN-neural correlations in face-selective, body-selective, scene-selective, object-selective, non-face-selective, and whole-brain regions in layers of the five DCNNs) with noise ceilings. Horizontal lines in each panel represent the mean noise ceilings of those regions. These lines are coded with the same colors in the legends. In both panels, error bars represent one standard error of the mean estimated by bootstrap resampling stimuli.

SI References

1. A. Krizhevsky, One weird trick for parallelizing convolutional neural networks. *ArXiv14045997 Cs* (2014) (September 16, 2021).
2. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv14091556 Cs* (2015) (January 31, 2021).
3. K. He, X. Zhang, S. Ren, J. Sun, Identity Mappings in Deep Residual Networks. *ArXiv160305027 Cs* (2016) (November 19, 2020).
4. Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. *ArXiv160708221 Cs* (2016) (September 16, 2021).
5. J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *ArXiv180107698 Cs* (2019) (March 4, 2020).
6. D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2017) (September 16, 2021).
7. G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: A database for studying face recognition in unconstrained environments (2007).
8. O. Esteban, *et al.*, fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111 (2019).
9. N. J. Tustison, *et al.*, N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med. Imaging* **29**, 1310–1320 (2010).
10. B. Fischl, FreeSurfer. *NeuroImage* **62**, 774–781 (2012).
11. B. Fischl, M. I. Sereno, A. M. Dale, Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System. *NeuroImage* **9**, 195–207 (1999).
12. R. W. Cox, AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res. Int. J.* **29**, 162–173 (1996).
13. M. Jenkinson, P. Bannister, M. Brady, S. Smith, Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. *NeuroImage* **17**, 825–841 (2002).
14. D. N. Greve, B. Fischl, Accurate and robust brain image alignment using boundary-based registration. *NeuroImage* **48**, 63–72 (2009).
15. J. D. Power, *et al.*, Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).
16. Y. Behzadi, K. Restom, J. Liau, T. T. Liu, A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).
17. M. Feilong, S. A. Nastase, J. S. Guntupalli, J. V. Haxby, Reliable individual differences in fine-grained cortical functional architecture. *NeuroImage* **183**, 375–386 (2018).
18. J. S. Guntupalli, *et al.*, A Model of Representational Spaces in Human Cortex. *Cereb. Cortex* **26**, 2919–2934 (2016).
19. J. S. Guntupalli, M. Feilong, J. V. Haxby, A computational model of shared fine-scale structure in the human connectome. *PLOS Comput. Biol.* **14**, e1006120 (2018).
20. J. V. Haxby, *et al.*, A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. *Neuron* **72**, 404–416 (2011).
21. P. Kaniuth, M. N. Hebart, "Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior" (2021).
22. S.-M. Khaligh-Razavi, L. Henriksson, K. Kay, N. Kriegeskorte, Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *J. Math. Psychol.* **76**, 184–197 (2017).
23. I. I. Groen, *et al.*, Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife* **7**, e32962 (2018).
24. M. N. Hebart, B. B. Bankson, A. Harel, C. I. Baker, R. M. Cichy, The representational dynamics of task and object processing in humans. *eLife* **7**, e32816 (2018).