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, NH, USA 03755.

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

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

[4] Cognitive Science, Dartmouth College, NH, USA 03755.

[5] Dipartimento di Medicina Specialistica, Diagnostica e Sperimentale, Università di Bologna, Bologna, Italy 40138.

[6] *Lead Contact*

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

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

Abstract

Deep convolutional neural networks (DCNNs) trained for face identification can rival and even exceed human-level performance. The relationships between internal representations learned by DCNNs and those of the primate face processing system are not well understood, especially in naturalistic settings. We developed the largest naturalistic dynamic face stimulus set in human neuroimaging research (700+ naturalistic video clips of unfamiliar faces) to investigate this problem. DCNN representational geometries were weakly but significantly correlated with neural response geometries across the human face processing system. Intermediate layers better matched visual, face-selective cortices, and behavioral similarity judgments than the final fully-connected layers. Our results showed DCNNs captured only a small amount of the rich information in the neural representations during naturalistic face viewing. Future artificial neural networks trained with more ecological objective functions may help advance artificial intelligence toward the ultimate goal of mimicking human intelligence in naturalistic, real-world scenarios.

Introduction

Deep convolutional neural networks (DCNNs) that are trained for face identification can match or even exceed human-level performance (Parkhi et al., 2015; Phillips et al., 2018; Taigman et al., 2014). Do these models learn internal representations similar to those used by the human brain? Attempts to directly interpret the embedding spaces learned by DCNNs suggest the models may implicitly represent a variety of face features (O'Toole et al., 2018). Previous studies reported that deep layers of DCNNs show substantial similarity to neural responses in the ventral temporal cortex of nonhuman primates (Raman and Hosoya, 2020; Schrimpf et al., 2020; Yamins et al., 2014; Yildirim et al., 2020). Studies investigating human face processing reported similarity to DCNNs, and, recently, a growing literature advocates the use of DCNNs as models to understand the neural basis of face processing in humans (Dobs et al., 2022; Grossman et al., 2019; Kuzovkin et al., 2018; Murty et al., 2021; Tsantani et al., 2021). However, correlations reported between DCNNs and neural representations of faces are often weak in humans (Kuzovkin et al., 2018; Tsantani et al., 2021), and a recent study has challenged how well DCNNs capture face representation in the macaque brain (Chang et al., 2021).

Face processing relies on a network of spatially segregated but functionally interconnected brain regions. In the human brain, face-selective regions include areas in the occipital and ventral temporal cortex, the superior temporal sulcus, and prefrontal cortex (Fairhall and Ishai, 2007; Guntupalli et al., 2017; Haxby and Gobbini, 2011; Haxby et al., 2000; Jiahui et al., 2020a; Natu and O'Toole, 2011; Ubaldi and Fairhall, 2021; Visconti di Oleggio Castello et al., 2017, 2021). To localize the full network of face-selective regions, dynamic videos of faces are widely used because, in comparison to still images, they evoke more reliable brain responses, better capture the cortical extent of the face-selective regions, and better engage anterior temporal and frontal face-selective regions than still images (Fox et al., 2009; Pitcher et al., 2011). So far, the few studies that examined similarities between DCNNs and biological face networks (Grossman et al., 2019; Tsantani et al., 2021) concentrated their investigation on only a few face-selective regions (mainly the occipital and posterior temporal cortices) and have not examined the remaining regions or the overall face network organization. In addition, nearly all previous studies used static images of a limited variety of identities, reducing the number, distribution, and naturalness of the stimuli and, hence, the ecological validity of their findings.

In this study, we developed a stimulus set comprising 707 naturalistic video clips of unfamiliar faces (Visconti di Oleggio Castello, 2018) to investigate the similarity between DCNNs and the distributed neural system for face processing in humans. Our stimulus set of face videos is the largest used in neuroimaging studies to investigate face processing. Faces in these video clips vary across a broad spectrum of perceived gender, age, ethnicity, head orientations, and expressions, providing a rich

sampling of high-dimensional face space. Instead of limiting our analysis to a few face-selective regions, we compared the representations of these faces in DCNNs and cortical responses across the entire face network, including face regions in the ventral, dorsal, and anterior core system (Visconti di Oleggio Castello et al., 2017, 2021). In addition, we used behavioral tasks to measure the perceived similarity of the faces stimuli as well as labeling face features such as perceived gender, age, expression, and ethnicity. We exhaustively quantified to what extent the perceived similarities and face features can explain the representations of both DCNNs and neural responses across the face processing system.

Correlations between DCNNs and neural representation were low but had a meaningful cortical distribution with the highest values in visual and face-selective brain regions. DCNN-neural correlations were much higher for intermediate layers than for the final fully-connected layers, suggesting that the distillation of intermediate layer features into the fully-connected layer features, which is guided by the narrow objective function of maximizing identification of individuals, diverges from the human face system, which serves a far richer array of objectives. Representational structure in both the DCNNs and the human brain data were highly consistent, but the best-performing layers of DCNNs accounted for less than 3% of the meaningful variance in the human brain. Moreover, there was no evident correspondence between the DCNN layer structure and the hierarchical structure of face-selective areas along processing pathways, suggesting further that DCNNs do not model the succession of processing stages in higher-order visual pathways. In particular, DCNNs do not represent dynamic features and cognitive information that are an integral part of human face processing. Correlations between behavioral ratings of face similarity and DCNN representations in intermediate layers, but not the fully-connected layer, were high, indicating that DCNN intermediate layers do in fact model some important aspects of human face processing. Correlations between behavioral ratings of face similarity and neural representation, however, were also very low, albeit with a meaningful distribution in the face system, suggesting that neural representation reflects face information beyond what guides behavioral judgments. Our findings suggest that human neural representations of dynamic, naturalistic faces contain rich and high-dimensional information, and that current state-of-the-art DCNNs capture only a small portion of this richness. Future research is needed to fully decompose the components of information in the neural representations of naturalistic face viewing and advance the artificial intelligence networks with more ecological objective functions.

Results

Quantifying the similarity between face representation in DCNNs and human face areas

To investigate whether a DCNN with human-level face identification performance is a good model of the human face processing system, we examined the representational similarity between patterns of face-DCNN features and patterns of brain responses. Human subjects underwent fMRI scanning while viewing a sequence of 4 s dynamic, naturalistic video stimuli depicting 707 faces across a variety of perceived genders, ages, ethnicities, head orientations, and expressions. Current state-of-the-art fMRI functional localizers for defining functional face category selectivity use similar dynamic videos of faces in naturalistic settings (Fox et al., 2009; Pitcher et al., 2011). Brain data from all participants were functionally aligned using hyperalignment based on participants' brain activity measured while watching a commercial movie, the Grand Budapest Hotel (Visconti di Oleggio Castello et al., 2020). Hyperalignment aligns brain response patterns in a common high-dimensional information space to capture shared information encoded in idiosyncratic topographies and greatly increases intersubject correlation of local representational geometry (Feilong et al., 2018; Guntupalli et al., 2016, 2018; Haxby et al., 2011, 2020a; Nastase et al., 2017). fMRI responses were averaged over 4 s for each face video in each cortical vertex, and a representational dissimilarity matrix (RDM) was constructed capturing the pairwise relationships between response patterns for each face within 10-mm-radius searchlights (Figure 1). For each layer of the DCNN, we similarly constructed an RDM capturing the pairwise relationships between activation patterns for each of the 707 faces in our stimuli (Figure 1). For analysis of correlations between DCNN and neural RDMs, we calculated the average RDM across participants for each searchlight.

To characterize DCNN face representations, we first used InsightFace, a state-of-the-art deep face recognition package (https://github.com/deepinsight/insightface) and later compared these representations to those generated by two other DCNNs – AlexNet and VGG16. The InsightFace package includes face detection (RetinaFace), face alignment, and face recognition (ArcFace) steps. Using exactly the same video stimuli as input to the human neural system, this deep face recognition system generates a 512-dimensional feature embedding for each frame of each face video clip as the final output (Figure 1). To construct the DCNN RDM, we averaged the embeddings across frames for each face video clip and computed the pairwise similarities between the averaged embeddings for all 707 video clips. Averaging the DCNN feature embeddings across frames is analogous to averaging fMRI patterns of brain responses across time-points. We correlated the ArcFace RDM in each layer and the neural RDM in each cortical searchlight and found that the representational geometries were more similar in regions extending from the early visual cortex to other regions in the occipital lobe, in the ventral temporal cortex, along the superior temporal sulcus, and in higher-level regions in the frontal lobe for all ArcFace layers (see example maps of the late intermediate layer _plus45 and the

fully-connected layer fc1 in Figure 2). These regions overlap largely with the previously reported human face processing system consisting of multiple face-selective regions (Haxby and Gobbini, 2011; Haxby et al., 2000; Jiahui et al., 2020a; Visconti di Oleggio Castello et al., 2017, 2021). We used a dynamic face localizer to independently define face-selective regions in our group of participants (face vs. objects; Figure S3A & B) (Jiahui et al., 2020a; Pitcher et al., 2011) and calculated the mean correlations for face-selective and non-face-selective regions in each layer. We found that neural RDMs in face-selective regions were best modeled by the late-intermediate ArcFace layers, and correlations drastically dropped after layer _plus45 and reached their lowest values in the final fully-connected layers (Figure 2C).

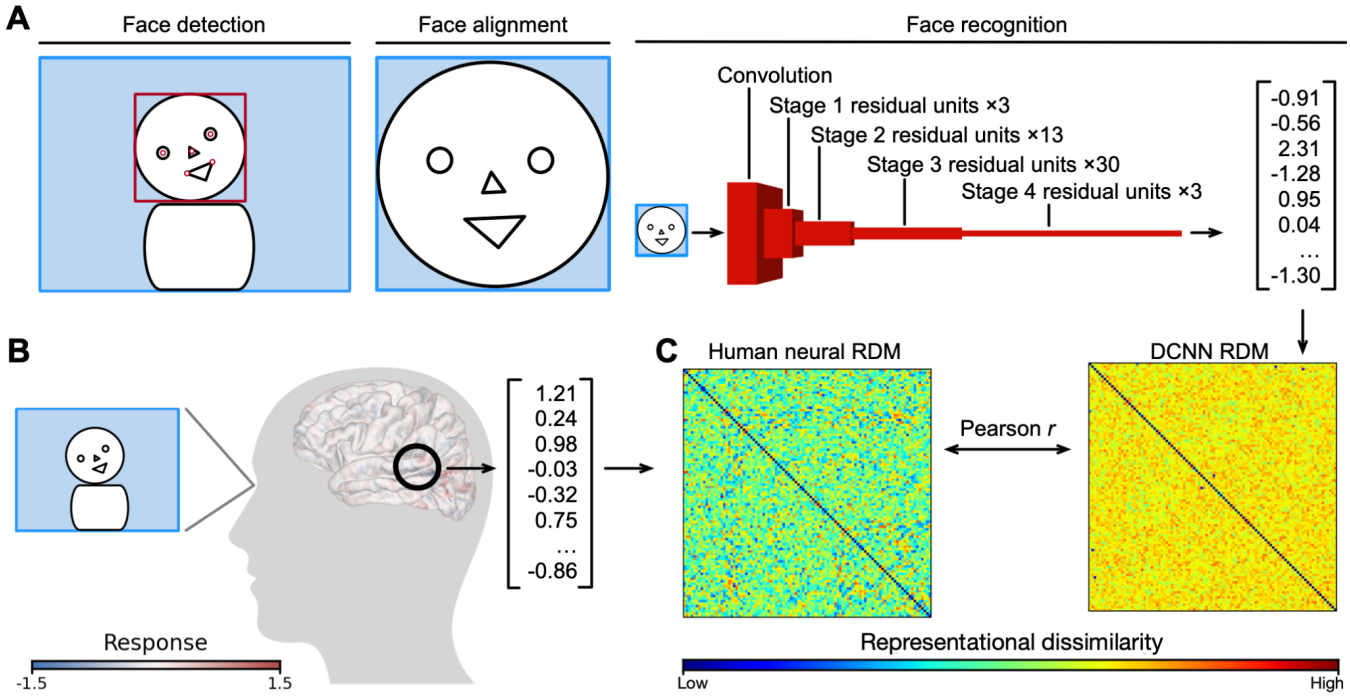

Figure 1. Schematic illustration of deep convolutional neural network (DCNN) and human neural representations of faces. Both the DCNNs and human subjects were presented with the same naturalistic face videos. **A.** The DCNN face recognition process comprised three steps. First, the face and its five key landmarks were automatically detected in each frame, and these landmarks were used to create the image of the aligned and cropped face. The cropped face image was then fed into the DCNN as input, and passed through convolutional layers, residual units, and fully-connected layers. The final output was a 512-dimensional embedding vector. Each video clip comprised 120 frames, and the corresponding 120 vectors were averaged to obtain an average embedding vector for each clip. Note that this illustrative example was based on the fully-connected layer of ArcFace. **B.** Human participants watched the face video clips in the fMRI scanner, and their brain responses were recorded. For each brain region (searchlight), the responses of multiple vertices in the region formed a spatial pattern, and the resulting pattern vector was considered the neural representation of the face clip for that brain region. **C.** For each brain region, we computed the dissimilarities between the pattern vectors of the 707 face clips, which formed a 707 × 707 representational dissimilarity matrix (RDM). Similarly, we created a DCNN RDM based on dissimilarities between the 707 embedding vectors. We used the Pearson correlation between the two RDMs to assess the similarity between DCNN and neural representations. Note that this figure is illustrative and not based on real data nor actual face clips used in the experiment.

Although correlations in the face-selective regions were significantly higher than the non-face-selective regions in both the peak intermediate layer and the final fully-connected layer, correlations with the peak intermediate layer were more than five times stronger than with the final fully-connected layer across face-selective regions. An additional analysis excluded the possibility that the low correlation was due to RSA's inherent assumption of equal weights or scales for all features comprising the two RDMs (Conwell et al., 2021; Kaniuth and Hebart, 2021; Khaligh-Razavi et al., 2017; Konkle and Alvarez, 2022) (see Material and Methods for details and Figure S4 for more results).

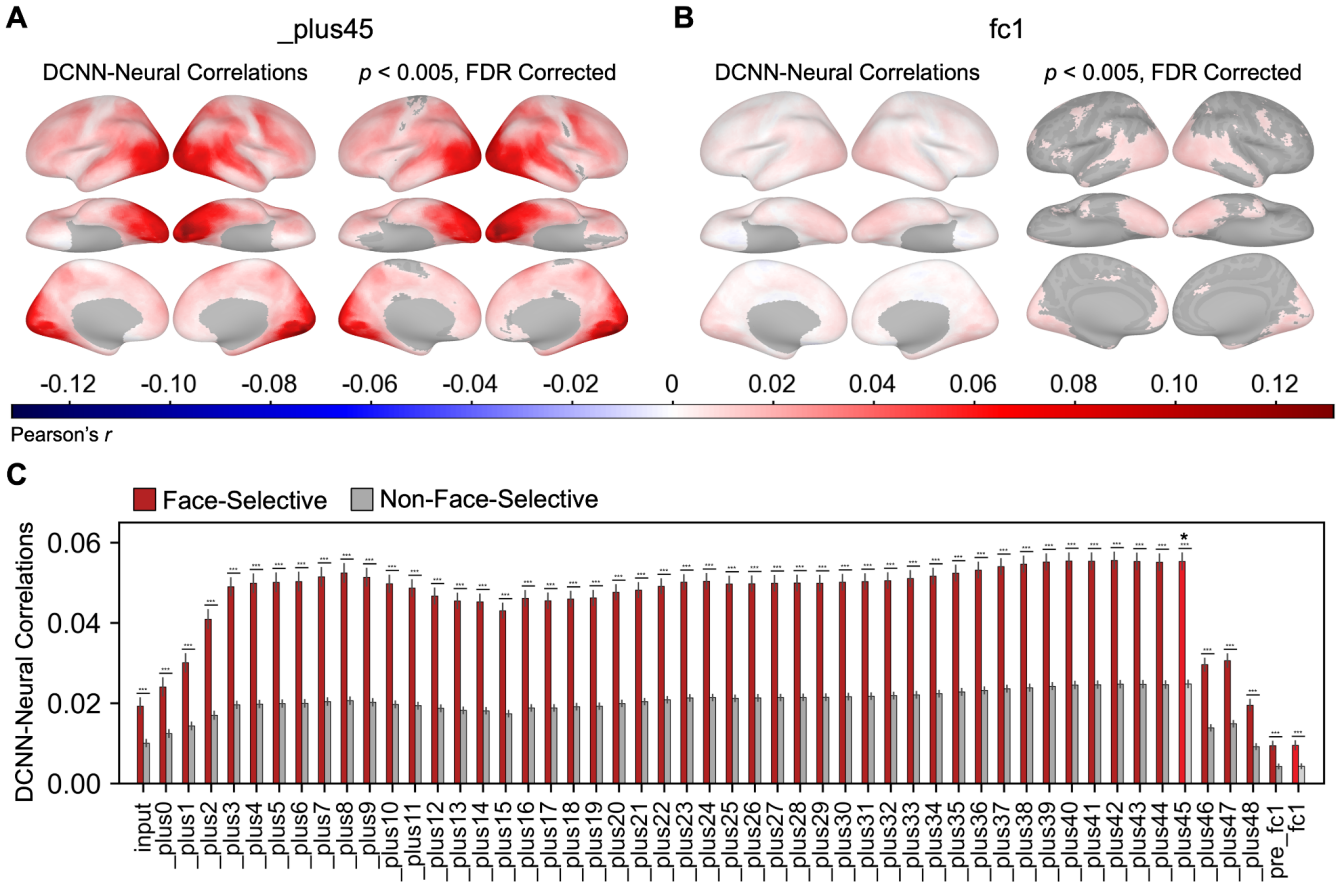

Figure 2. Correlations between ArcFace and neural RDMs. **A & B.** The DCNN-neural correlations across all cortical searchlights using RDMs in layer _plus45 (output of the last stage 3 residual unit) and fc1 (the last fully-connected layer). Both layers are highlighted in panel C. Correlations in the visual cortex, ventral temporal cortex, STS, and frontal regions were statistically significant for both layers (controlling FDR at $p < .005$, permutation test). **C.** Average correlations for face-selective regions (defined by a dynamic localizer, faces vs. objects, $t > 5$) and non-face-selective regions ($t <= 5$) plotted as red and gray bars respectively for each layer. The error bar length stands for one standard error of the mean estimated by bootstrap resampling of stimuli. Significance of the difference between the two bars was assessed via a permutation test randomizing stimulus labels. Layer _plus45 had the largest correlation with neural RDMs among all layers. *** $p < 0.001$.

Stimulus-related information content in face-DCNN and neural representational geometries

Correlations between ArcFace and neural representations were surprisingly low across all layers (Pearson's $r < 0.06$). The low correlations could be due to the representations being inherently dissimilar

between ArcFace and human participants (i.e., low “true” correlations), or due to excessive noise in data, which constrains the ceilings of possible correlations (i.e., low correlations due to low explainable variance). To examine to what extent our results are affected by noise, we measured the reliability of neural RDMs across subjects using Cronbach's alpha. Cronbach's alpha measures the amount of variance in neural RDMs that could be accounted for (i.e., variance not caused by noise), and it can be used as a noise ceiling for RDM correlations (Jiahui et al., 2020a). A high Cronbach's alpha means that RDMs from different participants are similar to each other, and the average RDM has high signal-to-noise ratio. Noise ceilings were high with maximum values exceeding 0.8 in early visual and 0.7 in posterior face-selective regions (Figure 3E & S9G). Noise ceilings in the anterior face regions were lower (0.1–0.4) but still several times higher than ArcFace-neural RSA correlation values in those same regions. Besides Cronbach’s alpha, additional decoding analyses revealed high cross-subject face identity decoding accuracies (over 80% accuracy in posterior face areas, Figure S6). Furthermore, the similarities of neural RDMs between areas of the face processing system replicated previous findings describing how face representations change from region to region (Guntupalli et al., 2017; Visconti di Oleggio Castello et al., 2017) (Figure S5). These analyses demonstrate that meaningful information about faces was encoded in brain responses, and further excluded the possibility that the weak correlations between the face-DCNN and neural representations were due to unreliable neural RDMs.

To examine whether more distributed brain activity patterns might better match the DCNN representations, we repeated this analysis with larger searchlight sizes (15 mm and 20 mm radius). Larger searchlight sizes only slightly improved the correlations, and the overall results remained weak (less than 2% variance, Figure S7 & S8).

To test whether a specific DCNN architecture had a significant effect on the representational similarities between the DCNN and the biological neural network, we performed a similar analysis using two other face-DCNNs (AlexNet and VGG16). Unlike ArcFace (Deng et al., 2019), which is based on a residual neural network (ResNet), AlexNet and VGG16 primarily comprise "classic" convolutional layers (see Material and Methods for details). We trained AlexNet and VGG16 with millions of face images and verified that they reached a satisfying performance level (see Material and Methods for details). We calculated DCNN-neural correlations for all layers in the other two face-DCNNs (AlexNet, and VGG16). Similar to ArcFace, intermediate layers of AlexNet and VGG16 showed substantially higher correlation values with neural RDMs than did the fully-connected layers, but those values were still very small (mean Pearson’s $r < 0.08$) compared to the estimated noise ceilings in the whole-brain, face-selective, and non-face-selective regions (Figure 3A, B, C, & D).

The weak correlations between face-DCNN and neural representations cannot be attributed to unreliable DCNN embeddings either. We evaluated the information content in the feature spaces for layers in the DCNNs by calculating correlations between RDMs for the layers of all three face-DCNNs (Figure 3F). Although the three face-DCNNs had different architectures, they shared highly similar representational geometries for the faces in our stimulus set, especially in the middle layers including ArcFace layer _plus45 (Pearson's $r > 0.7$), and these cross-DCNN similarities were layer specific. Correlations between ArcFace and the other two DCNNs in the last few layers and fully-connected layers were not as high as in the middle layers, but were still markedly higher (Pearson's $r > 0.3$) than the DCNN-neural correlations. Similar results for layer-specific DCNN representational geometries for objects have been shown previously (Kornblith et al., 2019; Mehrer et al., 2020).

Although the DCNN-neural correlations showed a meaningful cortical distribution, no meaningful mapping was evident between the layer structure of face-DCNNs and the sequence of face-selective areas in the human neural system for face representations (Figure S9B-D), suggesting that the sequence of representational geometries in the face-DCNN layers differs from the progression of representational geometries along the neural face pathways (Chang and Tsao, 2017; Freiwald and Tsao, 2010; Guntupalli et al., 2017; Visconti di Oleggio Castello et al., 2017).

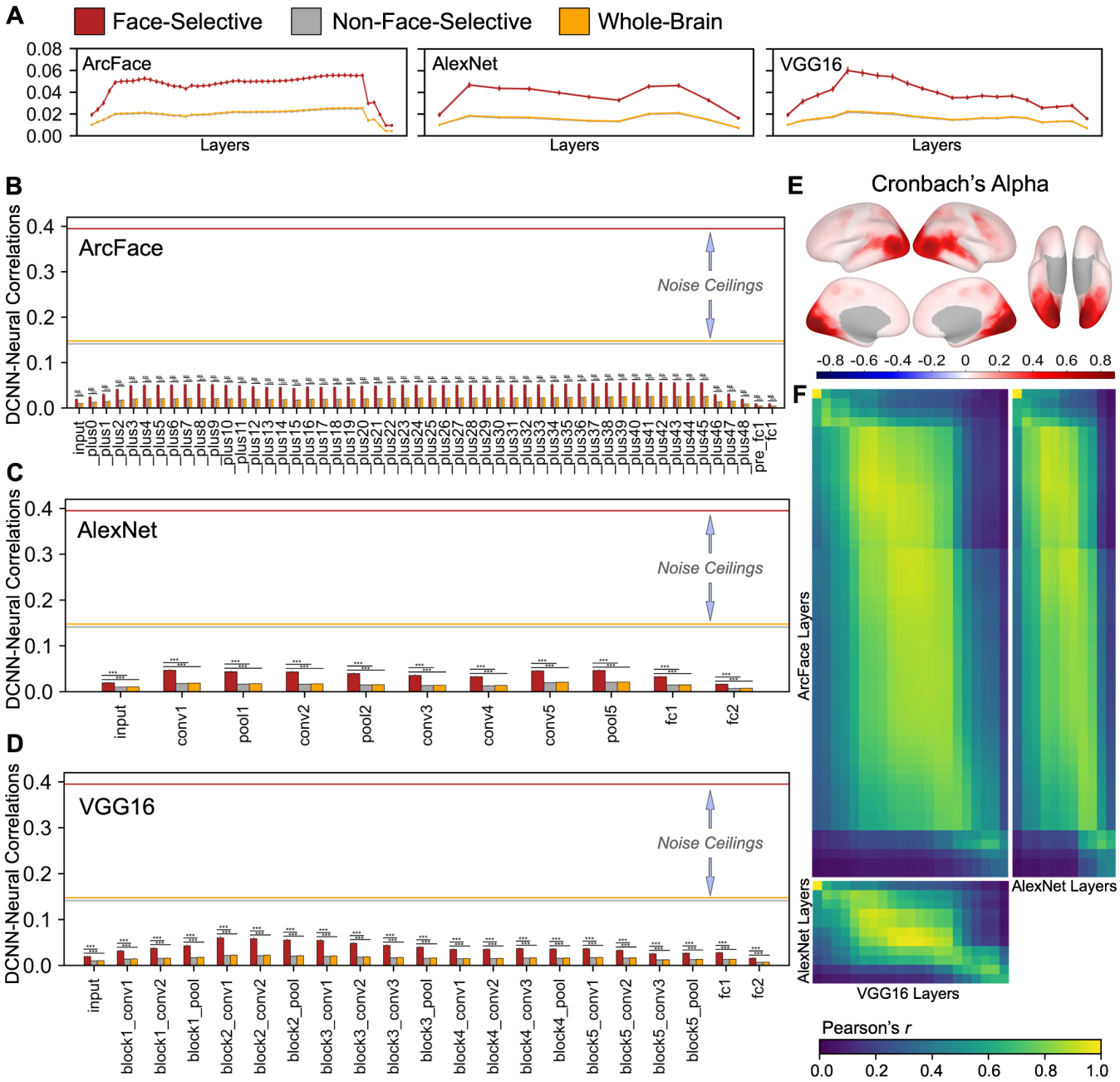

Figure 3. RSA results in layers of three face-DCNNs and cross correlations across the three DCNNs. **A.** Line plots of the mean DCNN–neural correlations in face-selective regions (red), non-face-selective regions (gray), and across the whole brain (orange) for each layer of the three face-trained DCNNs (ArcFace, AlexNet, and VGG16). (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 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 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. *** $p < 0.001$. **E.** Cronbach's alphas (noise ceilings) across all cortical searchlights. **F.** Correlations in each pair of layers in the three DCNN pairs. Layer orders are the same as those in panel A, B, C, & D.

Face representations in object-DCNNs

We next investigated whether DCNNs trained for general object category classification could explain more variance in the neural representations. Faces comprise a category in a broader object space (Bao et al., 2020; Long et al., 2018), and face-selective areas also subserve general object recognition (Haxby et al., 2001). Thus, it is possible that a DCNN trained on general object classification may model extra variance in the representational geometry of face areas. We used two DCNNs (AlexNet and VGG16) that had similar structures to face-DCNNs but were pre-trained with ImageNet images (Deng et al., 2009). Following the same analysis we used for face-DCNNs, we calculated correlations between RDMs built with feature embeddings for our face stimuli from each layer of the two object-DCNNs and neural representations in each searchlight across the cortical sheet. Mean correlation coefficients were calculated for the face-selective areas, non-face-selective areas, and the entire cortex. Surprisingly, correlations of object-DCNN and neural RDMs were similar to correlations of face-DCNN and neural RDMs (mean Pearson's $r < 0.08$) and similarly accounted for only a small fraction of the noise ceilings (Figure 4A, B & C). The layer-specific representational geometries of feature embeddings for the face stimuli were reliable over structures of the object-DCNNs as in the face-DCNNs. The two object-DCNNs shared highly similar, layer-specific representational geometries for the face videos, especially in the middle layers (Figure 4D, Pearson's $r > 0.8$).

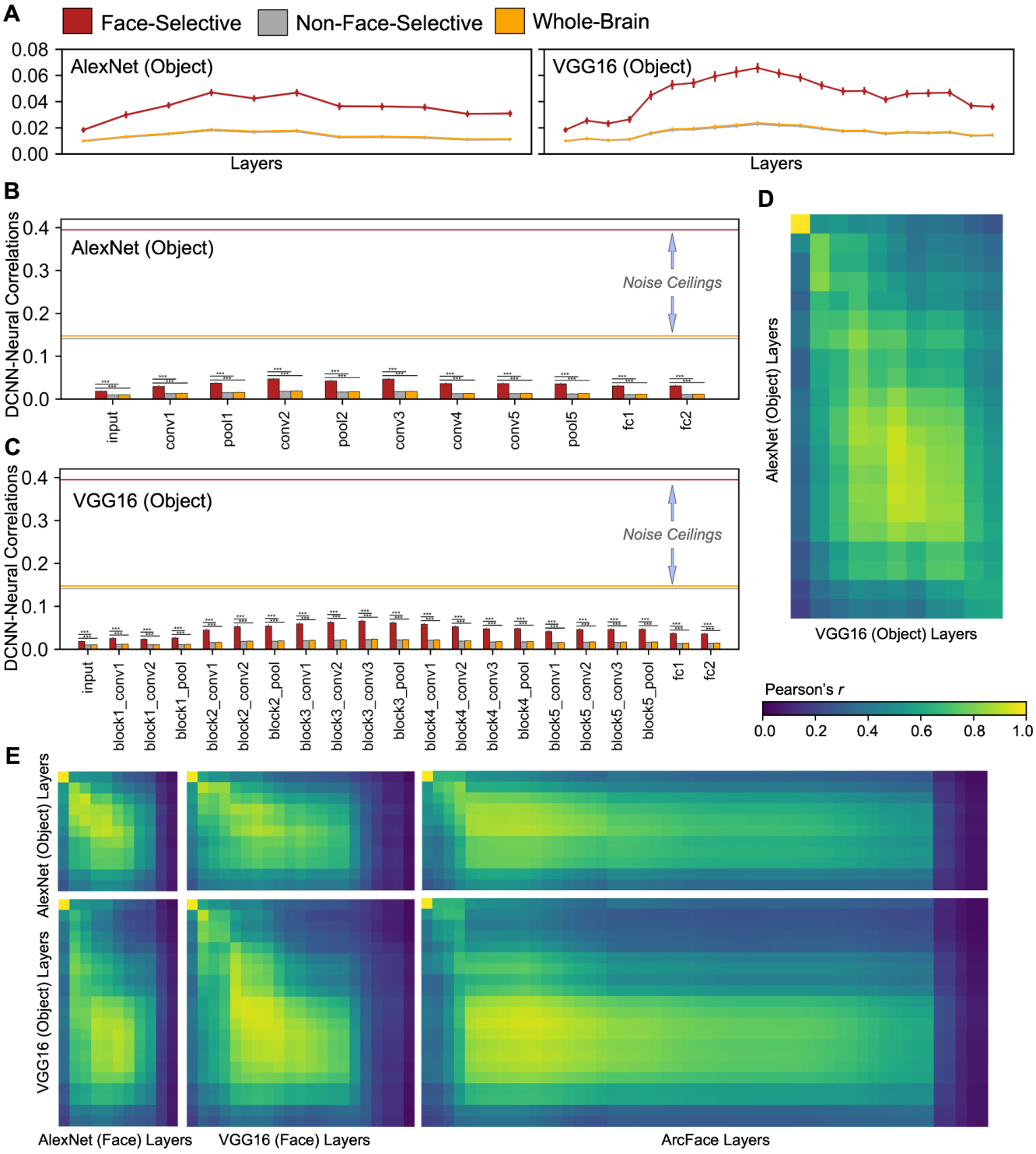

Figure 4. Layerwise RSA results for two object-DCNNs and cross correlations between face- and object-DCNNs. **A.** Line plots of mean DCNN-neural correlations in face-selective, non-face-selective, and whole-brain regions in layers of the two object-DCNNs (AlexNet, and VGG16). **B & C.** Bar plots of the same values as in panel A (mean DCNN-neural correlations in face-selective, non-face-selective, and whole-brain regions in layers of 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. *** $p < 0.001$. **D.** Correlations in each pair of layers in object-trained AlexNet and VGG16. Layer

orders are the same as those in panel A, B, & C. **E.** Correlations in each pair of layers in the six face- and object-DCNN pairs. Layers are displayed in the order as in panels A, B, and C in Figure 4 for the two object-DCNNs and in panel A, B, C and D in Figure 3 for the three face-DCNNs.

We next asked whether object-DCNNs and face-DCNNs share similar feature representations. To answer this question, we calculated correlations between RDMs calculated using feature embeddings in the layers of the three face-DCNNS with RDMs calculated using embeddings in the layers of the two object-DCNNs (Figure 4E). Unlike the analyses comparing representations within face- and object-DCNNs, representational geometries across differently trained networks were highly similar only in the early and middle layers, but quite dissimilar in the final few layers, especially in the final fully-connected layer (Pearson's $r < 0.08$).

We also examined DCNN-neural correlations in each individual face-selective ROI for each layer in both face- and object-DCNNs (Figure S9). Posterior regions had stronger correlations than anterior regions, and in anterior regions the right hemisphere tended to have relatively stronger correlations than the left hemisphere. None of the ROIs at any layer had correlations that approached the noise ceiling (Pearson's $r < 0.1$ in all cases).

Comparing behavioral similarity ratings to DCNNs and the human brain

Neither face-DCNNs nor object-DCNNs were good models for the biological neural representations of dynamic faces in naturalistic settings. We next investigated whether representational geometry based on subjective behavioral ratings of similarity could better capture the neural representations. Subjective similarities were measured using an arrangement task (Goldstone, 1994; Kriegeskorte and Mur, 2012; Nastase) with workers on Mechanical Turk (MTurk, https://www.mturk.com/). The stimuli used in single scanning runs (58 or 59 faces) were positioned outside of a circle at the beginning of the task, and MTurk workers were asked to arrange the stimuli inside the circle based on similarity of facial appearance. Thus, MTurk workers placed faces that they judged to be more similar closer to each other (Figure 5A). To retain the dynamic aspect of the stimuli, each image would expand when the cursor hovered over its thumbnail and play the 4 s video. The video automatically played once when the cursor hovered the first time, and participants could re-watch the video at any time if they right clicked the thumbnail.

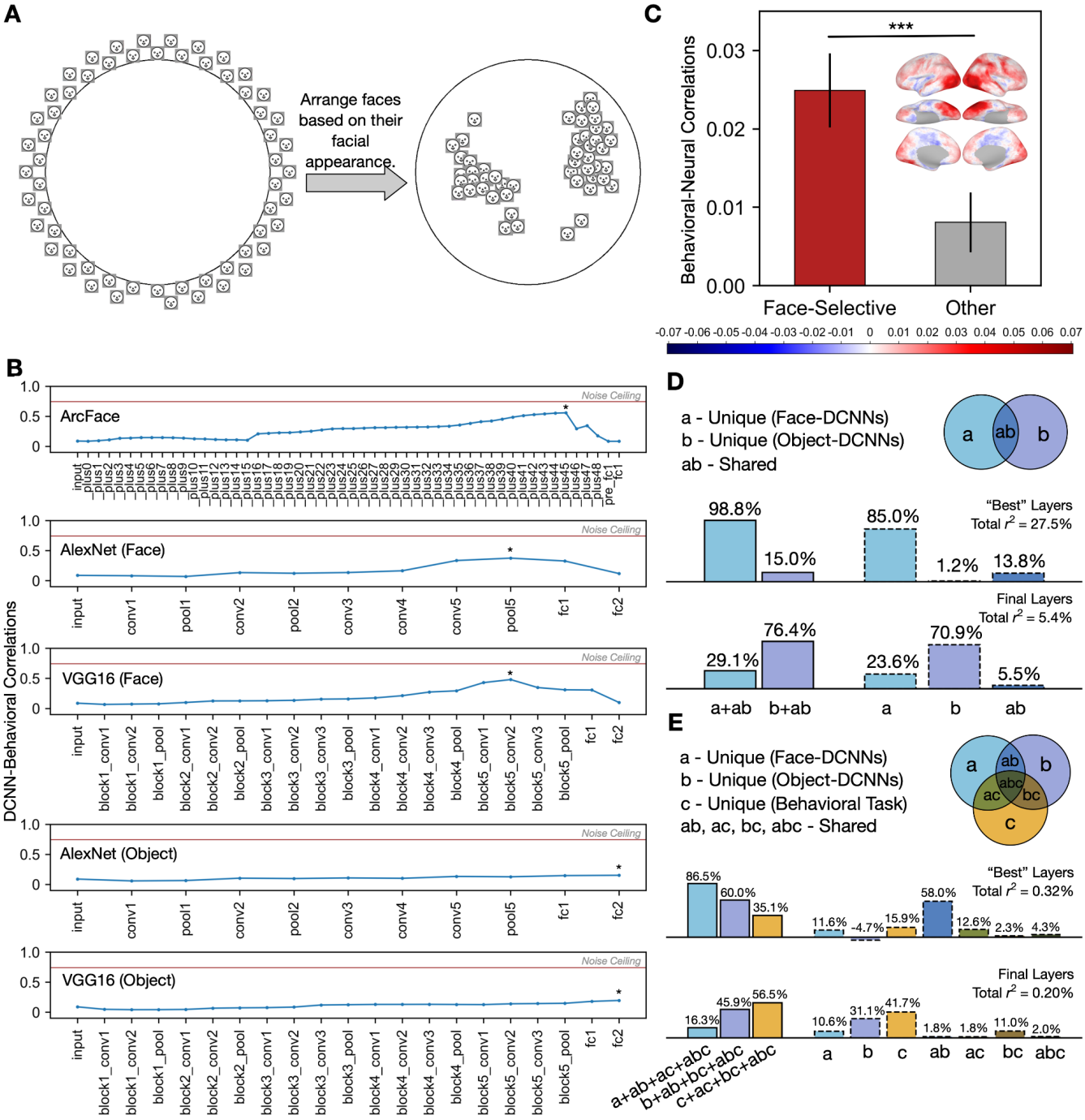

Figure 5. Behavioral arrangement task paradigm and results. **A.** Participants organized face stimuli based on facial appearance. **B.** Mean correlations across participants and runs between the behavioral RDM and DCNN RDM in each layer in all five DCNNs. The star marks the layer that has the highest 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.** Mean behavioral-neural correlations in face-selective and non-face-selective areas. The surface plot depicts the correlation values in the cortex. **D.** 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 all two 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 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. **E.** 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 two 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. For panel B, the error bars indicate standard error of the mean estimated by bootstrap resampling the stimuli for each layer (the error bars are too small to be visible in some cases). For panel C, the error bars indicate standard error of the mean estimated by bootstrap resampling the stimuli. Significance of the difference between the two bars was estimated using a permutation test randomizing the stimulus labels. *** $p < 0.001$.

Behavioral RDMs were constructed based on the distances between stimulus pairs for each MTurk worker, and individual RDMs were averaged across MTurk workers for each scanning run (see Material and Methods for details). Because these behavioral measures of perceptual similarity capture pairwise distances only for stimuli within runs, all comparisons to DCNN and neural representational geometries were made within runs. We calculated correlations between behavioral RDMs and RDMs for every layer in each DCNN within each run. Then, the within-run correlations were averaged across runs. Noise ceilings for the behavioral RDMs were also calculated within each run using Cronbach's alpha and averaged across runs. Correlations between behavioral and face-DCNN RDMs (ArcFace, face-AlexNet, face-VGG16) peaked in late intermediate layers, and the highest correlations were close to the noise ceiling (Figure 5B). By contrast, correlations between behavioral ratings and object-DCNNs were low in all layers. To test whether face- or object-DCNNs were good models of the behavioral representations, RDMs of the layers that had the strongest correlations ("best layers") in the three face-DCNNs (ArcFace, face-AlexNet, face-VGG16) and the two object-DCNNs (object-AlexNet, object-VGG16) were averaged within each category. RDMs of the final layers were similarly averaged. Variance partitioning analysis showed that the "best layers" of the face- and object-DCNNs explained 27.5% of the total variance of the behavioral RDMs (Figure 5D), due primarily to shared variance with face-DCNNs. By contrast, the final, fully-connected layers accounted for only 5.4% of the total variance.

We then analyzed the behavioral representational geometries as a model of the neural representations. Similarly to the analysis steps using DCNNs, RSA was performed for each searchlight across the entire cortex. Face-selective areas had significantly higher correlations than other areas (Figure 5C, permutation test, $p < 0.001$). To compare the amount of variance of the neural representations that was explained by the three types of models (behavioral task, face-DCNNs,

object-DCNNs), we performed a variance partitioning analysis in face-selective regions (see Material and Methods for details). Overall, the three models combined only explained 0.32% of the total neural variance even when using the layers from the face- and object-DCNN models that had the strongest correlations with neural representations ("best layers", see Material & Methods for details) (Figure 5E). The best layers increased the total explained variance from 0.20% to 0.32%, a ~60% increase but still only a small fraction (2.0%) of the noise ceiling. When the best layers were used, shared variance between face-DCNNs and object-DCNNs accounted for the major part of the total variance (58.0%, 0.19% of the neural RDM variance). When the final fully-connected layers were used, performance of the behavioral task and object-DCNNs explained most of the variance (41.7% and 31.1% , 0.08% and 0.06% of neural RDM variance, respectively), and face-DCNNs explained only 10.6%. Face-DCNNs had very little overlap with the other two models (1.8% with behavioral RDM and 1.8% with object-DCNNs), and the three models shared a very small portion of variance together (2.0%).

In summary, face-DCNNs captured variance of the behavioral face arrangement task in layers before the final fully-connected layers, but neither face-DCNNs, nor object-DCNNs, nor the behavioral arrangement task, nor any combination provided a strong model that captured the meaningful variance in the neural representations of dynamic faces.

Categorical features in the face representation of DCNNs and the human brain

To better understand the information represented in the representational geometries and investigate reasons for the divergence between the three types of representations (behavioral, DCNN, and neural), we collected ratings of face features for the stimuli from an independent group of MTurk workers. These face features included perceived gender, age, ethnicity, expression, and head orientation. The contribution of each face feature to the representational geometries of DCNNs, brain responses, and behavior was quantified by computing z-scored spatial distances within and between feature groups. Figure 6D shows an example that highlights the rationale of this analysis using multidimensional scaling (MDS) plots. A larger difference of between-group vs. within-group distances corresponds to a clearer division between the feature clusters (e.g., female/male).

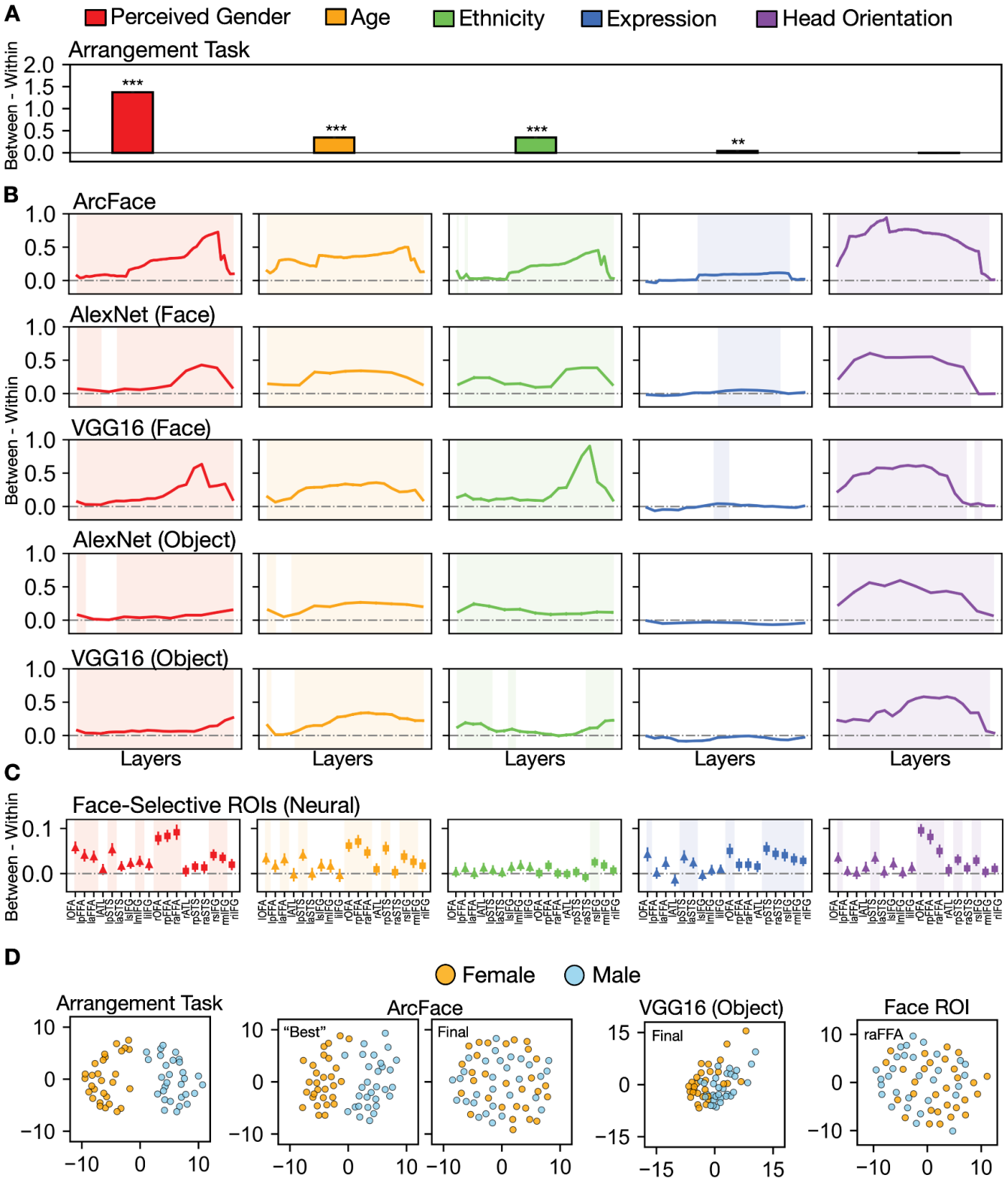

Figure 6. Five face features in behavioral, DCNN, and neural representations. **A, B, C.** Difference in the between- and within-group distance of perceived gender (red), age (orange), ethnicity (green), expression (blue), and head orientation (purple) in representational geometries of the behavioral arrangement task, each layer of the three face-trained DCNNs (ArcFace, AlexNet, VGG16), the two object-trained DCNNs (AlexNet, VGG16), and each face-selective ROI (bilateral OFA, pFFA, aFFA, ATL, pSTS, aSTS, sIFG, mIFG, iIFG). 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. The y-axis in panel C was scaled down to fit difference values in neural ROIs. Left triangles are nine face-selective ROIs in the left hemisphere, and right squares are face-selective ROIs in the right hemisphere. **D.** Example MDS plots using RDMs of the same run in the behavioral arrangement task, the "best" layer that showed highest correlation with the behavioral RDM (_plus45) in ArcFace, the final layer in ArcFace, the final layer in object-trained VGG16, and the right aFFA. Each dot is a stimulus. Orange and blue dots indicate perceived females and males, respectively. Behavioral and neural RDMs in this analysis were mean RDMs across participants. ** $p < 0.01$, *** $p < 0.001$.

Behaviorally, participants relied heavily on the perceived gender to cluster stimulus faces in the arrangement task, followed by age and ethnicity. Expression and head orientation were largely ignored (Figure 6A). Perceived gender, age, and ethnicity also contributed to the face-DCNN representations in late intermediate layers, but most of these contributions were reduced to low levels in the final few layers. A small but significant contribution of expression information was present in the middle layers and disappeared in the final few layers. Head orientation was more strongly represented in earlier intermediate layers, especially in ArcFace, and was not represented in the final few layers (Figure 6B, top three rows). Object-DCNNs did not reflect these face features except for head orientation in intermediate layers (Figure 6B, bottom two rows).

Compared to behavioral RDMs and DCNN RDMs in intermediate layers, neural RDMs contained less information for all face features. Analysis of face-selective ROIs revealed that face ROIs in the right hemisphere represented these features more prominently. Identity-related invariant categorical features such as perceived gender and age were significantly represented in bilateral face areas in the ventral temporal cortex (OFA, pFFA, aFFA), as well as in the pSTS and right IFG (Figure 6C). Expression was significantly represented in face areas in the OFA, pSTS, aSTS, and right IFG, but not the FFA. Head orientation was significantly represented in the OFA, right pFFA and mFFA, and the pSTS. The division of information represented in the ventral, dorsal, and frontal face-selective ROIs was consistent with previous results describing the distributed neural face processing system.

Discussion

State-of-the-art DCNNs trained to perform face recognition tasks have drawn considerable attention from researchers investigating face processing in humans and nonhuman primates. These artificial networks can identify faces at levels of accuracy that match or exceed human performance. Here, we investigated the extent to which DCNNs can model real-world face processing by measuring brain and behavioral responses to a large, varied set of naturalistic face videos. We found significant correlations between the DCNN and neural representational geometries with a cortical distribution that mirrored that of the human visual and face perception systems. Correlations were highest in visual cortex and

throughout the distributed face processing system, including all areas in the ventral, dorsal, and anterior core systems (Haxby et al., 2000; Jiahui et al., 2020a; Visconti di Oleggio Castello et al., 2017, 2021). Surprisingly, we found that, across three different face-DCNNs, the final fully-connected layers did not provide a good match to the neural representational geometries; rather, the intermediate layers provided a better model across the distributed face system in humans. This suggests that the transformations from intermediate DCNN layers to fully-connected layers, which distill visual information to optimize the network's objective function—identity recognition—diverge substantially from neural processes that subserve naturalistic face processing in humans. This finding resonates with work in natural language processing suggesting that the final layers of deep language models are overspecialized for their particular objective (e.g. word prediction): intermediate layers provide the best transferability to other linguistic tasks (Liu et al., 2019) and the best model for brain activity observed during natural language comprehension (Caucheteux et al., 2021; Toneva and Wehbe, 2019). Furthermore, different layers of the model did not systematically fit different regions of the distributed face processing system, suggesting the model layers do not recapitulate the succession of representational geometries, which reflect stages of processing in the human brain (Chang and Tsao, 2017; Freiwald and Tsao, 2010; Guntupalli et al., 2017; Visconti di Oleggio Castello et al., 2017).

The same stimuli were used as inputs to the DCNNs, for the fMRI experiment, and for the behavioral similarity task. Differences in representational geometry reflect differences in how the same input is processed by different systems. DCNNs embed each frame of the video stimuli in different feature spaces at each layer. We averaged the feature embeddings in each layer across all 120 frames in the 4 s videos. Similarly, we averaged the neural response patterns, measured every second, across all time-points in each video. State-of-the-art DCNNs do not model how face embeddings change over time, whereas the human face system does encode observed actions. In fact, agentic action is a major factor in the representational geometry of visual representation in the human brain (Haxby et al., 2020b; Nastase et al., 2017). A fruitful area for further development of future machine vision models may be incorporating features that capture how faces change over time. When such models are developed, we can reassess whether they produce representational geometries that more closely mirror the representational geometry of neural responses to naturalistic, dynamic faces.

Behavioral ratings of face similarity were dominated by categorical attributes of perceived gender, age, and ethnicity (Figure 6). These categorical variables play little role in individuation of face identity. In cognitive models of face perception, such categorical judgments precede processing for individuation (Bruce and Young, 1986). Though it remains controversial, many patients with prosopagnosia, who have impaired recognition of face identity, can still judge categorical attributes such

as perceived gender, age, expression, and gaze direction (Kress and Daum, 2003; Little et al., 2021; Richoz et al., 2015). Head orientation was not reflected in these ratings, and expression was only minimally reflected, suggesting that the raters did not prioritize these attributes when judging face similarity.

Representational geometry in the intermediate but not the fully-connected layer of face-DCNNs reflected these categorical attributes, suggesting that the transformations from intermediate to fully-connected layers effectively suppress information about categorical attributes to optimize individuation (see Figure S10 for a schematic illustration of how different information components are weighted in neural, DCNN, and behavioral RDMs). Contributions of categorical attributes to neural representations were significant but very weak. Perceived gender and age were reflected in the responses in ventral temporal, pSTS, and inferior frontal face areas. Expression was reflected in the representational geometries in the OFA, STS, and frontal areas. These results indicate that population responses in temporal and frontal face areas that afford individuation of faces also represent these categorical attributes of faces. It is worth noting that the individuation information used by face-DCNNs and neural responses are likely to be different. Low correlations between the representational geometries in face areas and fully-connected layers of face-DCNNs may be due, in part, to the multiplexing of categorical information with individuating information in the neural representations of faces. Similarly, the relatively stronger, albeit low, correlations with representational geometries in intermediate layers may be due, in part, to the finding that categorical and individuating information also are multiplexed in these intermediate face-DCNN layers. Due to a higher proportion of shared categorical information with neural representations in the intermediate layers, correlations were much stronger in these layers.

Our results show that both the fMRI-based representational geometries and the DCNN-based representational geometries are stable and information-rich, but less than 3% of that meaningful variance is shared. The challenge is to identify the meaningful information that is contained in the neural and face-DCNN representations but not shared. We focus here on two domains of information in dynamic videos that may play a large role in the variance that remains to be explained: information in facial movement and information derived from other cognitive processes that enrich face representations, such as social inferences, memory, and attention.

In comparison to static stimuli, dynamic stimuli dramatically alter the neural response to faces both in terms of tuning profiles and representational geometry. Response tuning to static, well-controlled stimuli in face patches are dominated by the presence or absence of faces or their static structural features (e.g. Chang & Tsao, 2017), but tuning for dynamic face stimuli is dominated by biological motion (Leopold and Park, 2020; McMahon et al., 2015; Park et al., 2017; Russ and Leopold, 2015). In

addition, dynamic faces are superior to static faces for the localization of face-selective areas (Fox et al., 2009; Hasson et al., 2004; Jiahui et al., 2020a; Pitcher et al., 2011). In a similar vein, representational geometry in ventral temporal cortex responses to static images of animals is dominated by animal category, but representational geometry for responses to videos of naturally behaving animals is dominated by animal behavior. Although animal category plays a significant role, it is dwarfed by the representation of behavior, which accounts for 2.5 times more variance (Haxby et al., 2020a; Nastase et al., 2017). Based on results cited here, the uniquely dynamic information in face videos may account for over two-thirds of the variance in neural representational geometries. Further research is needed to characterize precisely how these dynamics change the geometry of face representation. While static face stimuli may appear to be well-modeled by current DCNNs, naturalistic stimuli are required to reveal the deficiency of such models, and to focus our attention on how best to improve these models.

The neural representations of human faces reflect a variety of cognitive features. People automatically make inferences about novel faces — trustworthiness, competence, attractiveness — that can distort representational geometry (Oosterhof and Todorov, 2008; Todorov et al., 2008). Person knowledge plays a large role in the representation of familiar faces (Gobbini and Haxby, 2007; Gobbini et al., 2004; Leibenluft et al., 2004; Ramon and Gobbini, 2018; Visconti di Oleggio Castello et al., 2017, 2021). Familiarity is also known to distort face representations (Chauhan et al., 2020), and similarity of novel faces to familiar faces may influence perception and attribution. Faces also play a role in directing attention (Carlin et al., 2011; Friesen and Kingstone, 1998; Haxby et al., 2000; Hoffman and Haxby, 2000), and attention has a large effect on neural responses to faces (Furey et al., 2006; Jiahui et al., 2020b; Kanwisher and Wojciulik, 2000) that can be influenced by factors such as trait inferences, familiarity, and memory. Teasing apart the roles played by these different social and cognitive factors on human face representational geometry requires further research. Similarly, developing machine vision systems that incorporate dynamic and social features (expression, eye gaze, mouth movements, etc.) may enhance their power and utility for human-machine interactions.

Our behavioral measures were dominated by categorical attributes rather than features that underlie individuation. Better behavioral measures that reflect features for individuation may improve the correlation of behavior both with the fully connected layers in face-DCNNs and with neural representational geometries. Similarly, better modeling of neural representational geometry may help to tease out the part that is related to individuation separate from representations of categorical features, dynamic factors, social inferences, etc. Models that partition variance from these different sources could provide a better analysis of how well face-DCNNs map onto the human system. We think it is important to investigate this information in naturalistic dynamic stimuli that broadly sample face space rather than

to denature and impoverish the stimulus space, so that the role of the aspects of human face perception that are modeled by DCNNs can be placed properly in the full context of the much richer, more diverse human face perception system.

To summarize the results, Figure S10 illustrates our interpretation of the different components that played a role in the correlations of the behavioral, DCNNs, and human neural RDMs. Our results show that the behavioral RDM was mainly defined by information about categorical features, such as perceived gender and age. Similarly, RDMs from intermediate layers of face DCNNs mainly contained categorical feature information, while RDMs from the deep layers mainly contained information relevant for face individuation (but see (O'Toole and Castillo, 2021)). Thus, the high correlations between behavioral RDMs and the RDMs of the intermediate layers of face-DCNNs were mainly driven by the shared categorical information in both types of RDMs. The low correlations with deep layers were due to little face individuation information in the behavioral RDMs as well as little categorical feature information in the RDMs of deep layers.

On the other hand, the neural RDMs in the face-processing system contained all four kinds of information – categorical information (Figure 6), face individuation (e.g., Figure S6), dynamic information (dynamic faces are superior to static faces for the localization of face-selective areas), and information from other cognitive processes (e.g., social inference, memory, attention). However, because categorical information contributed the least to the neural RDMs (Figure 6), the shared information between behavioral and neural RDMs was limited. This low contribution of categorical information in neural RDMs can also explain the low correlations between neural RDMs and face-DCNNs in intermediate layers. Finally, the type of information used for face identification by DCNNs and the human face processing system may be different. For example, dynamic information or information from other cognitive processes is essential for the human face processing system (Fox et al., 2009; Oosterhof and Todorov, 2008; Pitcher et al., 2011; Todorov et al., 2008), while this type of information was not encoded by the face-DCNNs. These differences likely contributed to the low correlations between DCNN RDMs and neural RDMs.

Figure S10 suggests a framework for explaining the difference between our results and the results from previous studies that compared DCNNs to brain responses. Because dynamic information plays a major role in the geometry of brain representations (Haxby et al., 2020b; Nastase et al., 2017; Russ and Leopold, 2015), static images could generate higher correlation values between brain responses and DCNNs that do not use motion information (Daube et al., 2021; Grossman et al., 2019; Tsantani et al., 2021). Similarly, studies that used stimuli spanning superordinate categories (e.g., with multiple visual categories (Konkle and Alvarez, 2022; Murty et al., 2021)) would bias representations

towards categorical information, reducing the contribution of information that is needed for within-class individuation such as face identification.

Although face-DCNNs are trained on an exceptionally rich diversity of face images, face-DCNNs are optimized to encode these faces according to a very specific objective function: face identification. Face identification, however, is only one aspect of face processing in humans, which is flexible, highly contextualized, and ultimately supports social interaction. Building a representation of the uniqueness of the identity of a face takes a few hundred milliseconds (Castello and Gobbini, 2015), but is followed by sustained processing of a dynamic face in naturalistic viewing for gleaning other information for social cognition — changes of expression gaze, and head-orientation; speech-related mouth movements; inferences of intentions, social rank, social affiliation, reliability, and more. The human system for face perception is serving all of these goals during naturalistic viewing, and processes for face identification, besides playing only a small part that is finished quickly at the onset, may also be integrated with other functions in such a way that identification cannot be simply dissociated as a modular process. Perhaps in the future, artificial neural networks trained with more ecological objective functions (Daube et al., 2021; Hasson et al., 2020; Ranjan et al., 2017; Zhuang et al., 2021), requiring not just face recognition, but extending to facial dynamics, attention, memory, social context, and social judgments, will create face representations that afford a more ecologically-valid model that better captures the face processing system in humans. Our findings show that current state-of-the-art DCNNs are early-stage models for the human face processing system, and gaps exist between the current face DCNNs and the goal of developing *in silico* artificial intelligence models that mimic human intelligence in real-world, naturalistic scenarios.

Material and Methods

Participants

Twenty-one participants (mean age 27.3 years, range 22–31, 11 reported female) participated in the fMRI study. All participants had normal hearing and normal or corrected-to-normal vision, and no known history of neurological illness. The study was approved by the Dartmouth Committee for the Protection of Human Subjects.

Experimental Design

The Grand Budapest Hotel and localizer data were also used in prior work by Jiahui and colleagues (2020a). The MRI data acquisition parameters, preprocessing, and data analysis methods involving these two data sets are the same as in the previous publication.

The Grand Budapest Hotel. The full-length Grand Budapest Hotel movie was divided into six parts. Parts were divided at scene changes to keep the narrative of the movie intact. Participants watched the first part of the movie (~45 min) outside the scanner. Immediately thereafter, participants watched the remaining five parts of the movie in the scanner (~50 min, each part lasting 9–13 min) with audio. These data were curated and made publicly available for research use (Visconti di Oleggio Castello et al., 2020).

Hyperface. Video clips (707 clips, 4 s each) of individuals behaving naturally were created. The video clips were downloaded from YouTube and mostly comprised different people talking in interviews. Individuals in the clips varied widely in their identity, age, ethnicity, perceived gender, and head orientation. Audio channels were removed from the clips and the clips were cropped to remove unrelated text. The video clips were divided into 12 blocks (~59 clips per block) to match the 12 scanning runs and block order was counterbalanced across participants. In each run, participants were asked to watch the video clips (without fixation), shown continuously. After all clips in a run were shown, participants were tested with a brief four-trial memory check where they were asked to report whether a test clip was novel or was presented in the current run. Feedback was provided at the end of each run. Data from the memory test was removed from all analyses.

Dynamic localizer. Participants watched 3 s dynamic clips of faces, bodies, scenes, objects, and scrambled objects (Pitcher et al., 2011). The clips were presented continuously in 18 s blocks of each category, without blank periods between blocks. The blocks followed this order: an 18 s fixation period, five blocks of different categories (each lasting 18 s) in random order, an 18 s fixation period, five blocks of the categories in reversed order, and a final 18 s fixation period. Participants were required to press a button whenever they saw a repetition of a clip (five total in each run, one for each category). Four 234 s runs were collected for a total duration of 15:44.

Behavioral Arrangement Task. An independent group of 39 Amazon Mechanical Turk (MTurk) workers performed this task. Stimuli in a scanning run (59 stimuli for run 1-11 and 58 stimuli for run 12) were displayed as thumbnails outside a white circle on a gray background. When a trial began, the stimuli were arranged in randomized equidistant positions around the circumference of the circle. The first mouse hover triggered a larger and dynamic display of the video clip of that stimulus, and MTurk workers were able to rewatch the video by right clicking the mouse button. MTurk workers were instructed to arrange the thumbnails within the circle based on the similarity of the face appearance. To ensure a reasonable time for each participant to complete the experiment, we asked each of them to

perform three trials randomly selected from the total 12 trials. At least 10 different individuals completed each trial.

Behavioral Rating Task. Another independent group of 121 Amazon Mechanical Turk workers participated in the behavioral rating task. In each trial of the task, participants watched the video clip of a stimulus and rated the stimulus on five features: perceived gender (M/F), age (0-10, 11-20, 21-30, 31-40, 41-50, 51-60, 61-70, 70+), ethnicity (White, Black or African American, Asian, Indian, Hispanic or Latino, Other), expression (Neutral, Happiness, Surprise, Anger, Disgust, Sadness, Fear), and overall head orientation (Mostly Left, Mostly Center, Mostly Right). All 707 stimuli clips were divided into 15 independent experiment sessions (about 47 clips in each session), and each participant was assigned to one session to ensure the experiment could be completed in a reasonable amount of time. At least eight different individuals performed each session, and the final rating of each clip was the one that the most workers agreed on.

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 (Krizhevsky, 2014), VGG16 (Simonyan and Zisserman, 2015), and ResNet100 (He et al., 2016).

DCNN Models Trained for Face Recognition. All 3 Face DCNNs were trained using the MS-Celeb-1M dataset (Guo et al., 2016), 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 (Deng et al., 2019). 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 (Deng et al., 2019) to normalize the embedding vectors and weights.

We used the Adam optimizer for training (Kingma and Ba, 2017), 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 (Huang et al., 2007). 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, 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 (Esteban et al., 2019). T1-weighted images were corrected for intensity non-uniformity (Tustison et al., 2010) and skullstripped using antsBrainExtraction.sh. High resolution cortical surfaces were reconstructed with FreeSurfer (Fischl, 2012) using both T1-weighted and T2-weighted images, and then normalized to the fsaverage template based on sulcal curvature (Fischl et al., 1999). Functional data were slice-time corrected using 3dTshift (Cox, 1996), motion corrected using MCFLIRT (Jenkinson et al., 2002), distortion corrected using fieldmap estimate scans (one for each session), and then resampled to the fsaverage template based on boundary-based registration (Greve and Fischl, 2009). 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 (Power et al., 2014), 6 principal components from a combined cerebrospinal fluid and white matter mask (aCompCor) (Behzadi et al., 2007), and up to second-order polynomial trends.

Searchlight Hyperalignment. All three imaging datasets were hyperaligned (Feilong et al., 2018; Guntupalli et al., 2016, 2018; Haxby et al., 2011) 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.

Searchlight Representational Similarity Analysis. We performed a searchlight representational similarity analysis to quantify the similarity between DCNN and neural representational geometries. Embeddings derived from the final fully-connected layer and the intermediate layers were used to build the representational dissimilarity matrix (RDM) of the DCNN networks. In detail, the stimulus face and its five key landmarks were automatically detected in each frame to create the aligned and cropped face image. The cropped face image was then fed into the DCNN as input, and passed through the layers. Each video clip comprised 120 frames, and the corresponding 120 vectors were averaged to obtain an average embedding vector for each clip. Neural responses to each stimulus of the video clip were averaged over the duration of 4 s in each cortical vertex after adjusting for a 5 s hemodynamic delay, and the RDM was built using pattern similarity across clips for each 10 mm searchlight in each participant. The Hyperface stimulus set included 707 stimuli. This resulted in 707 × 707 RDMs for the DCNN layers and for each searchlight per participant, with each element of the RDM reflecting the correlation distance (1 – Pearson's r) between the response patterns elicited by the two stimuli in a pair (Figure 1). The neural RDMs were first averaged across participants in each searchlight, and Pearson's r values were calculated to measure the similarity between the model and neural RDMs across all surface searchlights to generate the whole-brain correlation map. To assess the statistical significance of whole-brain correlation maps, we performed a permutation test by shuffling the labels of the 707 stimuli prior to recomputing the RDMs 1000 times for each intermediate layer and 5000 times for the final fully-connected layer. The false discovery rate (FDR) was controlled at $p < 0.005$ to obtain whole-brain FDR corrected maps. For run-by-run analysis, RSA was done for each individual scanning run, and the correlation maps were averaged across runs.

Correlations in the Face-Selective ROIs and the Noise Ceiling. The face-selectivity map was estimated using hyperaligned localizer data. We calculated the univariate contrast map of faces vs. objects for each participant using the hyperaligned localizer data in the common model information space, averaged these to get the group face-selective map (Figure S3), and applied a conservative threshold of $t > 5$ to obtain the face-selective regions. Individual ROIs including the occipital face area (OFA), the posterior and anterior fusiform face areas (aFFA and pFFA), the anterior temporal face area (ATL), the posterior and anterior superior temporal sulcus (pSTS andaSTS), and three inferior frontal face areas (superior, middle, and inferior: sIFG, mIFG, and iIFG) bilaterally were localized by drawing a

disc of radius = 10 mm centered on the peak face-selective response (see also analysis with radius = 15 mm and 20 mm in Figure S7 & S8). Mean correlation coefficients were calculated for searchlights with centers within face-selective areas and non-face-selective areas for each layer of DCNNs. The correlation coefficient of each face-selective ROI was the value for the searchlight centered on the peak of face-selective response. Standard errors of the mean were calculated by bootstrapping the stimuli 1000 times for each intermediate layer and 5000 times for the final fully-connected layer. Statistical significance was assessed by permutation tests randomizing the stimulus labels 1000 times for each intermediate layer and 5000 times for the fully-connected layer.

The noise ceiling provides an estimate of the maximum possible correlation with the neural RDM predicted by the unknown true model (Nili et al., 2014). Because we averaged individuals' RDMs before RSA analysis, the noise ceiling was estimated by calculating Cronbach's alpha using neural RDMs across participants (Cronbach, 1951). Cronbach's alpha was used to describe the reliability of the neural RDMs across participants in each searchlight. To obtain noise ceilings for the face-selective areas, non-face-selective areas, and across the whole brain, mean alphas were calculated by averaging across vertices (corresponding to searchlight centers) in these regions. For the run-by-run analysis, the noise ceiling was estimated for each individual scanning run first and was averaged across runs to get the estimation of the overall noise ceiling.

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) (Kaniuth and Hebart, 2021; Khaligh-Razavi et al., 2017). 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.

Behavioral Arrangement Task RDMs and Noise Ceilings. Coordinates at the center point of the thumbnails were used to build behavioral RDMs for stimuli in each scan run for each participant. Each element of the behavioral RDMs reflected the Euclidean distance between the placements for a given pair of stimuli. Individual behavioral arrangement task RDMs were averaged across participants before further analysis. Because we averaged individual behavioral RDMs in each run before further analysis, the noise ceiling for each run was estimated using Cronbach's alpha across participants and averaged across runs.

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 (Groen et al., 2018; Hebart et al., 2018). 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.

References

Bao, P., She, L., McGill, M., and Tsao, D.Y. (2020). A map of object space in primate inferotemporal cortex. Nature *583*, 103–108. https://doi.org/10.1038/s41586-020-2350-5.

Behzadi, Y., Restom, K., Liau, J., and Liu, T.T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. NeuroImage *37*, 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042.

Bruce, V., and Young, A. (1986). Understanding face recognition. Br. J. Psychol. *77*, 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x.

Carlin, J.D., Calder, A.J., Kriegeskorte, N., Nili, H., and Rowe, J.B. (2011). A head view-invariant representation of gaze direction in anterior superior temporal sulcus. Curr. Biol. CB *21*, 1817–1821. https://doi.org/10.1016/j.cub.2011.09.025.

Castello, M.V. di O., and Gobbini, M.I. (2015). Familiar Face Detection in 180ms. PLOS ONE *10*, e0136548. https://doi.org/10.1371/journal.pone.0136548.

Caucheteux, C., Gramfort, A., and King, J.-R. (2021). GPT-2's activations predict the degree of semantic comprehension in the human brain. 2021.04.20.440622. https://doi.org/10.1101/2021.04.20.440622.

Chang, L., and Tsao, D.Y. (2017). The Code for Facial Identity in the Primate Brain. Cell *169*, 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011.

Chang, L., Egger, B., Vetter, T., and Tsao, D.Y. (2021). Explaining face representation in the primate brain using different computational models. Curr. Biol. *31*, 2785-2795.e4. https://doi.org/10.1016/j.cub.2021.04.014.

Chauhan, V., Kotlewska, I., Tang, S., and Gobbini, M.I. (2020). How familiarity warps representation in the face space. J. Vis. *20*, 18. https://doi.org/10.1167/jov.20.7.18.

Conwell, C., Prince, J.S., Alvarez, G.A., and Konkle, T. (2021). What can 5.17 billion regression fits tell us about artificial models of the human visual system? In SVRHM 2021 Workshop @ NeurIPS, p.

Cox, R.W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. Int. J. *29*, 162–173. .

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika *16*, 297–334. https://doi.org/10.1007/BF02310555.

Daube, C., Xu, T., Zhan, J., Webb, A., Ince, R.A.A., Garrod, O.G.B., and Schyns, P.G. (2021). Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. Patterns *2*, 100348. https://doi.org/10.1016/j.patter.2021.100348.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. ArXiv180107698 Cs.

Dobs, K., Martinez, J., Kell, A.J.E., and Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. Sci. Adv. *8*, eabl8913. https://doi.org/10.1126/sciadv.abl8913.

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods *16*, 111. https://doi.org/10.1038/s41592-018-0235-4.

Fairhall, S.L., and Ishai, A. (2007). Effective Connectivity within the Distributed Cortical Network for Face Perception. Cereb. Cortex *17*, 2400–2406. https://doi.org/10.1093/cercor/bhl148.

Feilong, M., Nastase, S.A., Guntupalli, J.S., and Haxby, J.V. (2018). Reliable individual differences in fine-grained cortical functional architecture. NeuroImage *183*, 375–386. https://doi.org/10.1016/j.neuroimage.2018.08.029.

Fischl, B. (2012). FreeSurfer. NeuroImage *62*, 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021.

Fischl, B., Sereno, M.I., and Dale, A.M. (1999). Cortical Surface-Based Analysis: II: Inflation, Flattening, and a Surface-Based Coordinate System. NeuroImage *9*, 195–207. https://doi.org/10.1006/nimg.1998.0396.

Fox, C.J., Iaria, G., and Barton, J.J.S. (2009). Defining the face processing network: Optimization of the functional localizer in fMRI. Hum. Brain Mapp. *30*, 1637–1651. https://doi.org/10.1002/hbm.20630.

Freiwald, W.A., and Tsao, D.Y. (2010). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. Science *330*, 845–851. https://doi.org/10.1126/science.1194908.

Friesen, C.K., and Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychon. Bull. Rev. *5*, 490–495. https://doi.org/10.3758/BF03208827.

Furey, M.L., Tanskanen, T., Beauchamp, M.S., Avikainen, S., Uutela, K., Hari, R., and Haxby, J.V. (2006). Dissociation of face-selective cortical responses by attention. Proc. Natl. Acad. Sci. U. S. A. *103*, 1065–1070. https://doi.org/10.1073/pnas.0510124103.

Gobbini, M.I., and Haxby, J.V. (2007). Neural systems for recognition of familiar faces. Neuropsychologia *45*, 32–41. https://doi.org/10.1016/j.neuropsychologia.2006.04.015.

Gobbini, M.I., Leibenluft, E., Santiago, N., and Haxby, J.V. (2004). Social and emotional attachment in the neural representation of faces. NeuroImage *22*, 1628–1635. https://doi.org/10.1016/j.neuroimage.2004.03.049.

Goldstone, R. (1994). An efficient method for obtaining similarity data. Behav. Res. Methods Instrum. Comput. *26*, 381–386. https://doi.org/10.3758/BF03204653.

Greve, D.N., and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. NeuroImage *48*, 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.

Groen, I.I., Greene, M.R., Baldassano, C., Fei-Fei, L., Beck, D.M., and Baker, C.I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. ELife *7*, e32962. https://doi.org/10.7554/eLife.32962.

Grossman, S., Gaziv, G., Yeagle, E.M., Harel, M., Mégevand, P., Groppe, D.M., Khuvis, S., Herrero, J.L., Irani, M., Mehta, A.D., et al. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. Nat. Commun. *10*, 4934. https://doi.org/10.1038/s41467-019-12623-6.

Guntupalli, J.S., Hanke, M., Halchenko, Y.O., Connolly, A.C., Ramadge, P.J., and Haxby, J.V. (2016). A Model of Representational Spaces in Human Cortex. Cereb. Cortex *26*, 2919–2934. https://doi.org/10.1093/cercor/bhw068.

Guntupalli, J.S., Wheeler, K.G., and Gobbini, M.I. (2017). Disentangling the Representation of Identity from Head View Along the Human Face Processing Pathway. Cereb. Cortex *27*, 46–53. https://doi.org/10.1093/cercor/bhw344.

Guntupalli, J.S., Feilong, M., and Haxby, J.V. (2018). A computational model of shared fine-scale structure in the human connectome. PLOS Comput. Biol. *14*, e1006120. https://doi.org/10.1371/journal.pcbi.1006120.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. ArXiv160708221 Cs.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Intersubject Synchronization of Cortical Activity During Natural Vision. Science *303*, 1634–1640. https://doi.org/10.1126/science.1089506.

Hasson, U., Nastase, S.A., and Goldstein, A. (2020). Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks. Neuron *105*, 416–434. https://doi.org/10.1016/j.neuron.2019.12.002.

Haxby, J.V., and Gobbini, M.I. (2011). Distributed Neural Systems for Face Perception. In Oxford Handbook of Face Perception, (Oxford, New York: Oxford University Press), pp. 93–110.

Haxby, J.V., Hoffman, E.A., and Gobbini, M.I. (2000). The distributed human neural system for face perception. Trends Cogn. Sci. *4*, 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. Science *293*, 2425–2430. https://doi.org/10.1126/science.1063736.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., and Ramadge, P.J. (2011). A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. Neuron *72*, 404–416. https://doi.org/10.1016/j.neuron.2011.08.026.

Haxby, J.V., Guntupalli, J.S., Nastase, S.A., and Feilong, M. (2020a). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. ELife *9*, e56601. https://doi.org/10.7554/eLife.56601.

Haxby, J.V., Gobbini, M.I., and Nastase, S.A. (2020b). Naturalistic stimuli reveal a dominant role for agentic action in visual representation. NeuroImage *216*, 116561. https://doi.org/10.1016/j.neuroimage.2020.116561.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity Mappings in Deep Residual Networks. ArXiv160305027 Cs.

Hebart, M.N., Bankson, B.B., Harel, A., Baker, C.I., and Cichy, R.M. (2018). The representational dynamics of task and object processing in humans. ELife *7*, e32816. https://doi.org/10.7554/eLife.32816.

Hoffman, E.A., and Haxby, J.V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. Nat. Neurosci. *3*, 80–84. https://doi.org/10.1038/71152.

Huang, G.B., Mattar, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage *17*, 825–841. https://doi.org/10.1006/nimg.2002.1132.

Jiahui, G., Feilong, M., Visconti di Oleggio Castello, M., Guntupalli, J.S., Chauhan, V., Haxby, J.V., and Gobbini, M.I. (2020a). Predicting individual face-selective topography using naturalistic stimuli. NeuroImage *216*, 116458. https://doi.org/10.1016/j.neuroimage.2019.116458.

Jiahui, G., Yang, H., and Duchaine, B. (2020b). Attentional modulation differentially affects ventral and dorsal face areas in both normal participants and developmental prosopagnosics. Cogn. Neuropsychol. *0*, 1–12. https://doi.org/10.1080/02643294.2020.1765753.

Kaniuth, P., and Hebart, M.N. (2021). Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior.

Kanwisher, N., and Wojciulik, E. (2000). Visual attention: insights from brain imaging. Nat. Rev. Neurosci. *1*, 91–100. https://doi.org/10.1038/35039043.

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., and Kriegeskorte, N. (2017). 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. https://doi.org/10.1016/j.jmp.2016.10.007.

Kingma, D.P., and Ba, J. (2017). Adam: A Method for Stochastic Optimization. ArXiv14126980 Cs.

Konkle, T., and Alvarez, G.A. (2022). A self-supervised domain-general learning framework for human ventral stream representation. Nat. Commun. *13*, 491. https://doi.org/10.1038/s41467-022-28091-4.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of Neural Network Representations Revisited. ArXiv190500414 Cs Q-Bio Stat.

Kress, T., and Daum, I. (2003). Developmental Prosopagnosia: A Review. Behav. Neurol. *14*, 109–121. https://doi.org/10.1155/2003/520476.

Kriegeskorte, N., and Mur, M. (2012). Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements. Front. Psychol. *3*. .

Krizhevsky, A. (2014). One weird trick for parallelizing convolutional neural networks. ArXiv14045997 Cs.

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J.-P., Baciu, M., Kahane, P., Rheims, S., Vidal, J.R., and Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. Commun. Biol. *1*, 1–12.

https://doi.org/10.1038/s42003-018-0110-y.

Leibenluft, E., Gobbini, M.I., Harrison, T., and Haxby, J.V. (2004). Mothers' neural activation in response to pictures of their children and other children. Biol. Psychiatry *56*, 225–232. https://doi.org/10.1016/j.biopsych.2004.05.017.

Leopold, D.A., and Park, S.H. (2020). Studying the visual brain in its natural rhythm. NeuroImage *216*, 116790. https://doi.org/10.1016/j.neuroimage.2020.116790.

Little, Z., Palmer, C.J., and Susilo, T. (2021). Intact gaze processing in developmental prosopagnosia. J. Vis. *21*, 2267. https://doi.org/10.1167/jov.21.9.2267.

Liu, N.F., Gardner, M., Belinkov, Y., Peters, M.E., and Smith, N.A. (2019). Linguistic Knowledge and Transferability of Contextual Representations. ArXiv190308855 Cs.

Long, B., Yu, C.-P., and Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. Proc. Natl. Acad. Sci. *115*, E9015–E9024. https://doi.org/10.1073/pnas.1719616115.

McMahon, D.B.T., Russ, B.E., Elnaiem, H.D., Kurnikova, A.I., and Leopold, D.A. (2015). Single-Unit Activity during Natural Vision: Diversity, Consistency, and Spatial Sensitivity among AF Face Patch Neurons. J. Neurosci. *35*, 5537–5548. https://doi.org/10.1523/JNEUROSCI.3825-14.2015.

Mehrer, J., Spoerer, C.J., Kriegeskorte, N., and Kietzmann, T.C. (2020). Individual differences among deep neural network models. Nat. Commun. *11*, 5725. https://doi.org/10.1038/s41467-020-19632-w.

Murty, N.Apurva.R., Bashivan, P., Abate, A., DiCarlo, J.J., and Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. Nat. Commun. *12*, 5540. https://doi.org/10.1038/s41467-021-25409-6.

Nastase, S.A. The Geometry of Observed Action Representation During Natural Vision. Ph.D. Dartmouth College.

Nastase, S.A., Connolly, A.C., Oosterhof, N.N., Halchenko, Y.O., Guntupalli, J.S., Visconti di Oleggio Castello, M., Gors, J., Gobbini, M.I., and Haxby, J.V. (2017). Attention Selectively Reshapes the Geometry of Distributed Semantic Representation. Cereb. Cortex *27*, 4277–4291. https://doi.org/10.1093/cercor/bhx138.

Natu, V., and O'Toole, A.J. (2011). The neural processing of familiar and unfamiliar faces: A review and synopsis. Br. J. Psychol. *102*, 726–747. https://doi.org/10.1111/j.2044-8295.2011.02053.x.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. PLOS Comput. Biol. *10*, e1003553. https://doi.org/10.1371/journal.pcbi.1003553.

Oosterhof, N.N., and Todorov, A. (2008). The functional basis of face evaluation. Proc. Natl. Acad. Sci. *105*, 11087–11092. https://doi.org/10.1073/pnas.0805664105.

O'Toole, A.J., and Castillo, C.D. (2021). Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning. Annu. Rev. Vis. Sci. *7*, null. https://doi.org/10.1146/annurev-vision-093019-111701.

O'Toole, A.J., Castillo, C.D., Parde, C.J., Hill, M.Q., and Chellappa, R. (2018). Face Space

Representations in Deep Convolutional Neural Networks. Trends Cogn. Sci. *22*, 794–809. https://doi.org/10.1016/j.tics.2018.06.006.

Park, S.H., Russ, B.E., McMahon, D.B.T., Koyano, K.W., Berman, R.A., and Leopold, D.A. (2017). Functional Subpopulations of Neurons in a Macaque Face Patch Revealed by Single-Unit fMRI Mapping. Neuron *95*, 971-981.e5. https://doi.org/10.1016/j.neuron.2017.07.014.

Parkhi, O.M., Vedaldi, A., and Zisserman, A. (2015). Deep Face Recognition. In Procedings of the British Machine Vision Conference 2015, (Swansea: British Machine Vision Association), p. 41.1-41.12.

Phillips, P.J., Yates, A.N., Hu, Y., Hahn, C.A., Noyes, E., Jackson, K., Cavazos, J.G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proc. Natl. Acad. Sci. *115*, 6171–6176. https://doi.org/10.1073/pnas.1721355115.

Pitcher, D., Dilks, D.D., Saxe, R.R., Triantafyllou, C., and Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. NeuroImage *56*, 2356–2363. https://doi.org/10.1016/j.neuroimage.2011.03.067.

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048.

Raman, R., and Hosoya, H. (2020). Convolutional neural networks explain tuning properties of anterior, but not middle, face-processing areas in macaque inferotemporal cortex. Commun. Biol. *3*, 1–14. https://doi.org/10.1038/s42003-020-0945-x.

Ramon, M., and Gobbini, M.I. (2018). Familiarity matters: A review on prioritized processing of personally familiar faces. Vis. Cogn. *26*, 179–195. https://doi.org/10.1080/13506285.2017.1405134.

Ranjan, R., Patel, V.M., and Chellappa, R. (2017). HyperFace: A Deep Multi-task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. ArXiv160301249 Cs.

Richoz, A.-R., Jack, R.E., Garrod, O.G.B., Schyns, P.G., and Caldara, R. (2015). Reconstructing dynamic mental models of facial expressions in prosopagnosia reveals distinct representations for identity and expression. Cortex J. Devoted Study Nerv. Syst. Behav. *65*, 50–64. https://doi.org/10.1016/j.cortex.2014.11.015.

Russ, B.E., and Leopold, D.A. (2015). Functional MRI mapping of dynamic visual features during natural viewing in the macaque. NeuroImage *109*, 84–94. https://doi.org/10.1016/j.neuroimage.2015.01.012.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., et al. (2020). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? BioRxiv 407007. https://doi.org/10.1101/407007.

Simonyan, K., and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. ArXiv14091556 Cs.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In 2014 IEEE Conference on Computer Vision and Pattern

Recognition, pp. 1701–1708.

Todorov, A., Said, C.P., Engell, A.D., and Oosterhof, N.N. (2008). Understanding evaluation of faces on social dimensions. Trends Cogn. Sci. *12*, 455–460. https://doi.org/10.1016/j.tics.2008.10.001.

Toneva, M., and Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). ArXiv190511833 Cs Q-Bio.

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A.L., McGettigan, C., and Garrido, L. (2021). FFA and OFA encode distinct types of face identity information. J. Neurosci. https://doi.org/10.1523/JNEUROSCI.1449-20.2020.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., and Gee, J.C. (2010). N4ITK: Improved N3 Bias Correction. IEEE Trans. Med. Imaging *29*, 1310–1320. https://doi.org/10.1109/TMI.2010.2046908.

Ubaldi, S., and Fairhall, S.L. (2021). fMRI response to automatic and purposeful familiar-face processing in perceptual and nonperceptual cortical regions. J. Neurophysiol. *125*, 1058–1067. https://doi.org/10.1152/jn.00481.2020.

Visconti di Oleggio Castello, M. (2018). Characterizing Feature Representations in the Human Face-Processing Network with Multivariate Analyses and Encoding Models ([Unpublished doctoral dissertation]. Dartmouth College).

Visconti di Oleggio Castello, M., Halchenko, Y.O., Guntupalli, J.S., Gors, J.D., and Gobbini, M.I. (2017). The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. Sci. Rep. *7*, 12237. https://doi.org/10.1038/s41598-017-12559-1.

Visconti di Oleggio Castello, M., Chauhan, V., Jiahui, G., and Gobbini, M.I. (2020). An fMRI dataset in response to “The Grand Budapest Hotel”, a socially-rich, naturalistic movie. Sci. Data *7*, 383. https://doi.org/10.1038/s41597-020-00735-4.

Visconti di Oleggio Castello, M., Haxby, J.V., and Gobbini, M.I. (2021). Shared neural codes for visual and semantic information about familiar faces in a common representational space. Proc. Natl. Acad. Sci. *118*. https://doi.org/10.1073/pnas.2110474118.

Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111.

Yildirim, I., Belledonne, M., Freiwald, W., and Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. Sci. Adv. *6*, eaax5979. https://doi.org/10.1126/sciadv.aax5979.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M.C., DiCarlo, J.J., and Yamins, D.L.K. (2021). Unsupervised neural network models of the ventral visual stream. Proc. Natl. Acad. Sci. *118*. https://doi.org/10.1073/pnas.2014196118.

Acknowledgements

We thank the authors of the InsightFace package for making their models and training data freely available for non-commercial research use. This work was supported by NSF grants 1607845 (J.V.H) and 1835200 (M.I.G).

Author contributions

Conceptualization, G.J., M.F., J.V.H., and M.I.G.; Methodology, G.J., M.F., J.V.H., and M.I.G.; Software, G.J., M.F., and S.A.N.; Formal Analysis, G.J. and M.F.; Investigation, G.J., M.F., and M.V.d.O.C.; Resources, J.V.H. and M.I.G.; Data Curation, G.J., M.F., and M.V.d.O.C.; Writing – Original Draft, G.J., M.F., J.V.H., and M.I.G.; Writing – Review & Editing, G.J., M.F., M.V.d.O.C., S.A.N., J.V.H., and M.I.G.; Visualization, G.J. and M.F.; Supervision, J.V.H. and M.I.G.; Funding Acquisition, J.V.H. and M.I.G.

Competing interests

The authors declare no competing interests.

Data and materials availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Further data and codes can be downloaded from https://github.com/GUO-Jiahui/face_DCNN. Additional data and materials that support the findings of this study are available on request from the corresponding authors G.J., M.F., and M.I.G.