PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES OPEN ACCESS

Modeling naturalistic face processing in humans with deep convolutional neural networks

Guo Jiahui[a,1,2], Ma Feilong[a,1,2], Matteo Visconti di Oleggio Castello[b], Samuel A. Nastase[c], James V. Haxby[a], and M. Ida Gobbini[d,e,2]

Edited by Marlene Behrmann, University of Pittsburgh, Pittsburgh, PA; received March 16, 2023; accepted September 11, 2023

Deep convolutional neural networks (DCNNs) trained for face identification can rival and even exceed human-level performance. The ways in which the internal face representations in DCNNs relate to human cognitive representations and brain activity are not well understood. Nearly all previous studies focused on static face image processing with rapid display times and ignored the processing of naturalistic, dynamic information. To address this gap, we developed the largest naturalistic dynamic face stimulus set in human neuroimaging research (700+ naturalistic video clips of unfamiliar faces). We used this naturalistic dataset to compare representational geometries estimated from DCNNs, behavioral responses, and brain responses. We found that DCNN representational geometries were consistent across architectures, cognitive representational geometries were consistent across raters in a behavioral arrangement task, and neural representational geometries in face areas were consistent across brains. Representational geometries in late, fully connected DCNN layers, which are optimized for individuation, were much more weakly correlated with cognitive and neural geometries than were geometries in late-intermediate layers. The late-intermediate face-DCNN layers successfully matched cognitive representational geometries, as measured with a behavioral arrangement task that primarily reflected categorical attributes, and correlated with neural representational geometries in known face-selective topographies. Our study suggests that current DCNNs successfully capture neural cognitive processes for categorical attributes of faces but less accurately capture individuation and dynamic features.

artificial neural network | deep neural network | face identification | naturalistic stimuli | hyperalignment

Significance

Faces in real life convey categorical attributes (e.g., age), unique identities, and dynamic information (e.g., expression, attention). Deep convolutional neural networks (DCNNs) can be trained to individuate faces, but individuation may be only a small part of naturalistic face perception. Our study compared representations of naturalistic, dynamic faces in DCNNs, cognitive tasks, and brain responses measured with functional magnetic resonance imaging (fMRI). Our results show that intermediate DCNN representations capture categorical attributes of faces that match cognitive and neural representations but later DCNN representations that extract view-invariant identity do not, suggesting that DCNNs provide a good model for early cognitive and neural face processing of categorical attributes but are a poor model for individuation and for extended processing of dynamic features.

Deep convolutional neural networks (DCNNs) that are trained for face identification can match or even exceed human-level performance (1–3). Do these models learn internal representations of faces similar to human cognitive and neural representations? Attempts to directly interpret the embedding spaces learned by DCNNs suggest that the models may implicitly represent a variety of face features (4). Previous studies reported that representations of objects and faces in deep layers of DCNNs show substantial similarity to neural responses in the ventral temporal cortex of nonhuman primates (5–8). Recent studies reported similar face representations in DCNNs and the human brain (9–13). Nearly all prior studies, however, used static face images with short display times (hundreds of milliseconds). One study so far (13) that used dynamic naturalistic video clips of faces with longer presentation times (3 s) reported weak correlations between face representations in DCNNs and the brain.

Although face perception processes operate on both still images and videos, the quick processing of static images with rapid display times and the more extended processing of longer dynamic videos may engage different cognitive processes and brain responses. Early processing of still images affords individuation of identity but is only a small part of more extended face processing in naturalistic settings. Recognition of identity appears to be achieved in under 400 ms, but people continue to watch faces intently long after identity is established. The extended processing of naturalistic, dynamic faces may elaborate information that relates inferences of state of mind to social cognitive and semantic context. In support of this view, neural responses to dynamic videos reveal a richer information space that is not evident in responses to static images (14–18). It is currently unclear whether DCNNs capture these additional levels of information about faces.

To test the utility of DCNNs as models of human cognitive and neural representations of dynamic, naturalistic faces, we developed a stimulus set comprising 707 naturalistic 4 s video clips of unfamiliar faces (19). This face stimulus set, alongside the accompanying fMRI data, is one of the largest currently available in the neuroimaging literature. Faces in these video clips vary across a broad spectrum of perceived gender, age, ethnicity, head orientations, and expressions, providing a rich sampling of the high-dimensional face space. We analyzed this dynamic face stimulus set in terms of representational geometries produced

Author contributions: G.J., M.F., J.V.H., and M.I.G. designed research; G.J., M.F., and M.V.d.O.C. performed research; G.J., M.F., and S.A.N. contributed new reagents/analytic tools; G.J. and M.F. analyzed data; J.V.H. and M.I.G. resources, supervision, funding acquisition; and G.J., M.F., M.V.d.O.C., S.A.N., J.V.H., and M.I.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]G.J. and M.F. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jiahui.guo@dartmouth.edu, feilong.ma@dartmouth.edu, or mariaida.gobbini@unibo.it.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2304085120/-/DCSupplemental.

Published October 17, 2023.

by DCNNs, by behavioral measures of perceived similarity and categorical attributes, and by fMRI measures of neural responses. To ensure that our results were not dependent on a specific DCNN architecture, we repeated all analyses using three separate face-DCNNs. In a behavioral arrangement task, raters placed thumbnails of face videos in a two-dimensional field according to perceived similarity. In a second behavioral task, raters judged categorical attributes of the face images (gender, ethnicity, age, expression, and head orientation). Instead of limiting our analysis to a few face-selective regions as in previous studies (mainly the occipital and posterior temporal cortices), we compared face representations between DCNNs and cortical responses across the entire face-processing network, including regions in the ventral, dorsal, and anterior core system (20, 21).

Representational geometries derived from DCNN, behavioral, and neural measures were all highly reliable, providing a strong foundation and high noise ceilings for investigating their interrelationships. Correlations between representations in DCNNs and the behavioral arrangement task were high, approaching the noise ceiling. Further analysis with feature ratings showed that representational geometries produced by both DCNNs and the behavioral arrangement task were dominated by categorical face attributes. Even though the final, fully connected layers of DCNNs are optimized for view-invariant recognition of identity, their correlations with behavioral and neural geometries were markedly weaker than were correlations with late-intermediate layers, suggesting that the human cognitive and neural processes for face individuation are poorly modeled by DCNN processes for face individuation. Correlations of neural representational geometries with DCNN and behavioral representational geometries were significant, albeit low, with a meaningful cortical distribution. The highly reliable but unexplained variance in neural representational geometries may reflect face information beyond categorical attributes, such as dynamic information that is not captured by the behavioral tasks or by DCNNs, or it may reflect face-identity information that is used by the human brain but not by DCNNs. Overall, our results show that current DCNNs successfully model representations of categorical face attributes but support our hypothesis that their utility for modeling human cognitive and neural representations of dynamic, naturalistic faces may be limited to this early stage of processing and not extend to information embedded in dynamic information and to human processes for face individuation.

Results

Reliable Face Representations in DCNNs, Human Behavior, and the Brain. To investigate shared information in DCNNs, human behavior, and the human brain, we characterized the representations of 707 naturalistic face video clips with multiple high-performing DCNNs, a behavioral arrangement task of perceived similarity, and fMRI data.

To derive DCNN face representations, we first used InsightFace, a state-of-the-art deep face recognition package (https://github.com/deepinsight/insightface). This package includes face detection (RetinaFace), face alignment, and face recognition (ArcFace) steps (Fig. 1*A*) and is currently the industry standard for face identification. We compared these representations to those in two other face-trained DCNNs (AlexNet and VGG16) and two object-trained DCNNs with the same architecture (AlexNet and VGG16) (22).

In the behavioral arrangement task, workers on Mechanical Turk (MTurk, https://www.mturk.com/) arranged videos according to perceived face similarities. 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 the similarity of facial appearance (Fig. 1*C*). To retain the dynamic aspect of the stimuli, each stimulus 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 rewatch the video at any time if they right-clicked the thumbnail.

In the fMRI experiment, human participants underwent scanning while viewing a sequence of 4 s dynamic, naturalistic video stimuli (Fig. 1*D*). Current state-of-the-art fMRI localizers for defining functional face category selectivity use similar dynamic videos of faces in naturalistic settings (23, 24). Brain data from all participants were functionally aligned using hyperalignment based on participants' brain activity (*SI Appendix*, Fig. S1) measured while watching a commercial movie, the Grand Budapest Hotel (25). 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 (16, 26–30).

Representational dissimilarity matrices (RDMs) were constructed for DCNNs, behavioral similarity arrangements, and neural responses using similar methodologies to characterize pairwise distances between face video clips (see *Materials and Methods* for details). We first assessed the reliability of the information content in the RDMs.

We compared RDMs of different DCNN architectures by calculating correlations between layers within each of the three face-DCNNs (Fig. 1*B*). Although the three face-DCNNs had different architectures, they shared highly similar representational geometries for faces in our stimulus set, especially in the middle layers (Pearson's $r > 0.7$). Similar correlations between face- and object-DCNNs were found for intermediate layers, but fully connected layers from face-DCNNs were mostly uncorrelated with object-DCNNs (*SI Appendix*, Fig. S3). These cross-similarities were layer specific, which extended previous results showing layer-specific DCNN representational geometries for objects using object-trained DCNNs (31, 32). Correlations between ArcFace and the other two face-DCNNs in the last few layers and fully connected layers also were significant (Pearson's $r > 0.3$. Correlations between the final layers of ArcFace and AlexNet, ArcFace and VGG16, and between AlexNet and VGG16 were 0.46, 0.35, and 0.66, respectively.) but lower than in the middle layers. We found similarly reliable RDMs for the two object-DCNNs (*SI Appendix*, Fig. S3*A*).

Next, we calculated the noise ceiling for the behavioral arrangement task. Since each behavioral trial showed only face videos from one scanning run (~60 faces), this task measured RDM similarity across workers for stimuli within scanning runs. The noise ceiling was calculated using Cronbach's alpha (33), which was computed first across participants within each run and then averaged across runs. A high Cronbach's alpha means that RDMs from different participants are similar to each other and that the average RDM has a high signal-to-noise ratio. The mean Cronbach's alpha value was 0.74, indicating highly similar behavioral arrangements across participants.

We then measured the reliability of neural RDMs across subjects. Noise ceilings were high with maximum values exceeding 0.8 in early visual and 0.7 in posterior face-selective regions (Fig. 1*D* and *SI Appendix*, Fig. S12*G*). Noise ceilings in the anterior face regions were around 0.1 to 0.4. To further demonstrate that identity information was reliably encoded in the neural representations, we conducted between-subject decoding analyses. For this analysis, we split participants into training and test groups with two different strategies: split-half (dividing participants into two equally sized training

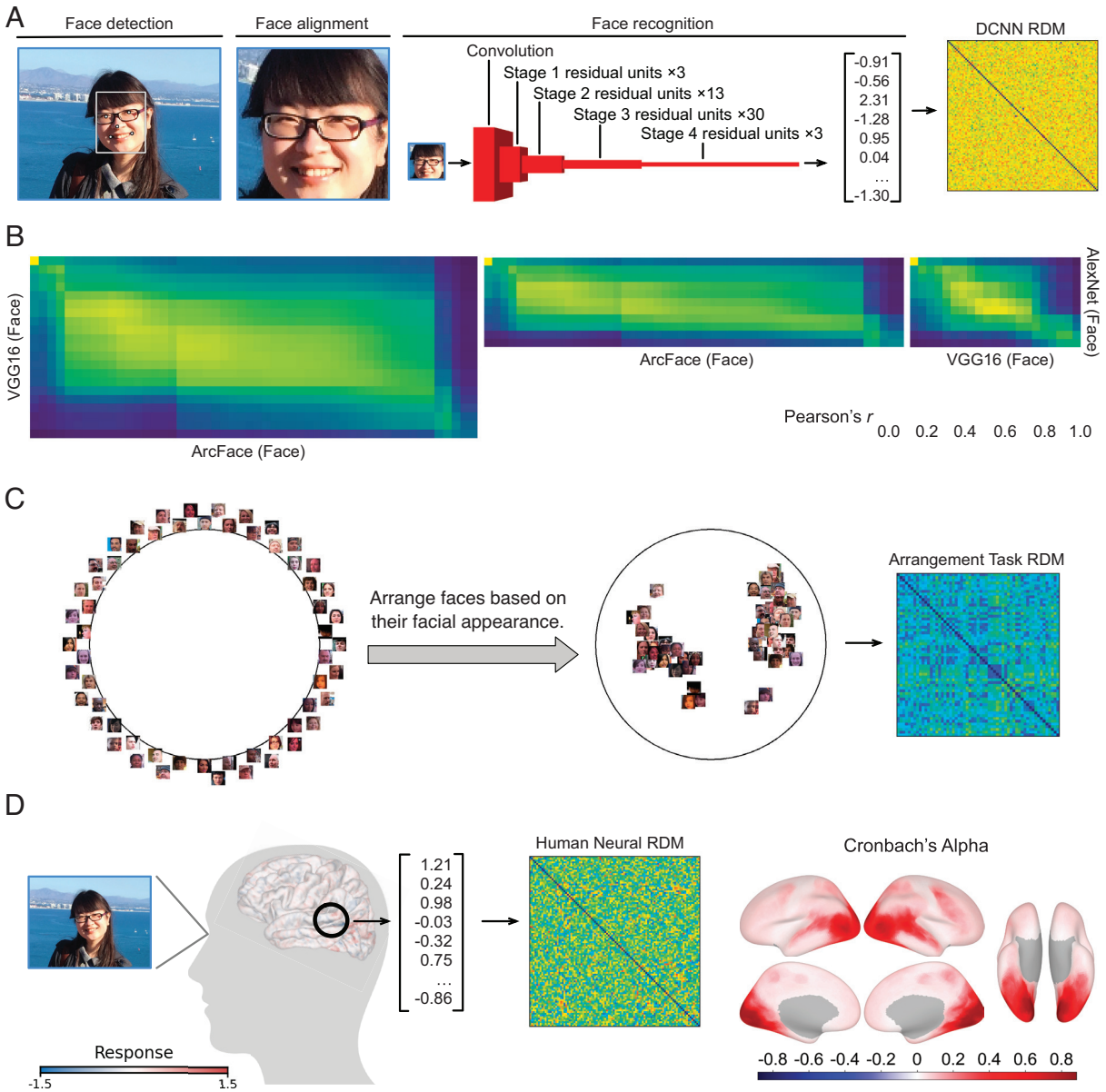

Fig. 1. Schematic illustration of representational dissimilarity matrices (RDMs) and reliabilities of DCNNs, behavioral performances, and human neural responses. (*A*) The DCNN face recognition process comprised three steps. First, the face and its five key landmarks were automatically detected in each 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. We computed the dissimilarities between the embedding vectors of the 707 face clips to build a 707 × 707 RDM. Note that DCNNs and human subjects were presented with the same naturalistic face videos, and this illustrative example was based on the fully connected layer of ArcFace. (*B*) Correlations in each pair of layers within each of the three face-DCNN pairs. (*C*) In the behavioral arrangement task, MTurk workers organized face stimuli based on facial appearance, and behavioral RDMs were constructed based on the distances between stimulus pairs. Note that this figure is illustrative and not based on real data in the experiment. Mean Cronbach's alpha across participants was high (0.74). (*D*) Human participants watched face video clips in the fMRI scanner, and their brain responses were recorded. For each brain region (searchlight), 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. For each brain region, we computed the dissimilarities between the pattern vectors of the 707 face clips, which formed a 707 × 707 RDM. The surface plot depicts Cronbach's alphas (noise ceilings) of brain RDMs across all cortical searchlights.

and test groups) and leave-one-out (iteratively holding out one participant at a time as the test participant). Specifically, we examined whether the activation pattern for each stimulus face was more similar to the pattern for that face in other participants' brains than to the patterns for other faces. Results with both methods generated high identity decoding accuracies, especially in face-selective areas (over 80% accuracy in posterior face areas, *SI Appendix*, Fig. S4), suggesting reliable encoding of identity information in the neural data across participants. Furthermore, similarities of neural RDMs between areas of the face processing system replicated previous findings describing how face representations change from region to region (20, 34) (*SI Appendix*, Fig. S5). Overall, these results showed that meaningful information for faces was reliably encoded in local patterns of fMRI responses in cortical face processing areas.

Strong Correlations between DCNN and Human Behavioral Representations. We applied representational similarity analysis (RSA) to investigate relationships between RDMs based on DCNN features in different layers and on human behavioral representations (see the section above for an overview of the behavioral experiment). Correlations between face-DCNN RDMs (ArcFace, face-AlexNet, face-VGG16) and behavioral

RDMs peaked in late-intermediate layers, and the highest correlations were close to the noise ceiling (Fig. 2*A*). These high correlation values demonstrate that face-DCNN feature spaces for our face video stimuli capture information in human cognitive representations. By contrast, correlations between object-DCNN RDMs and behavioral RDMs were low across all layers (*SI Appendix*, Fig. S3*B*). Taken together, these results show that the type of the DCNNs and the image statistics of training datasets (face-DCNNs vs. object-DCNNs) have a stronger effect than the specific DCNN architecture when modeling human behavioral representations (35).

To further compare face- and object-DCNNs, we conducted a variance partitioning analysis that quantified how much variance in behavioral representational geometries could be accounted for by face- and object-DCNNs. Results showed that the layers of the face- and object-DCNNs with the strongest correlations ("best layers") explained 27.5% of the total variance of the behavioral RDMs, due primarily to face-DCNNs (27.2% of the total explained variance). By contrast, the final, fully connected layers accounted for only 5.4% of the total variance of the behavioral RDMs, which was primarily due to object-DCNNs (4.1% of the total explained variance) (*SI Appendix*, Fig. S13).

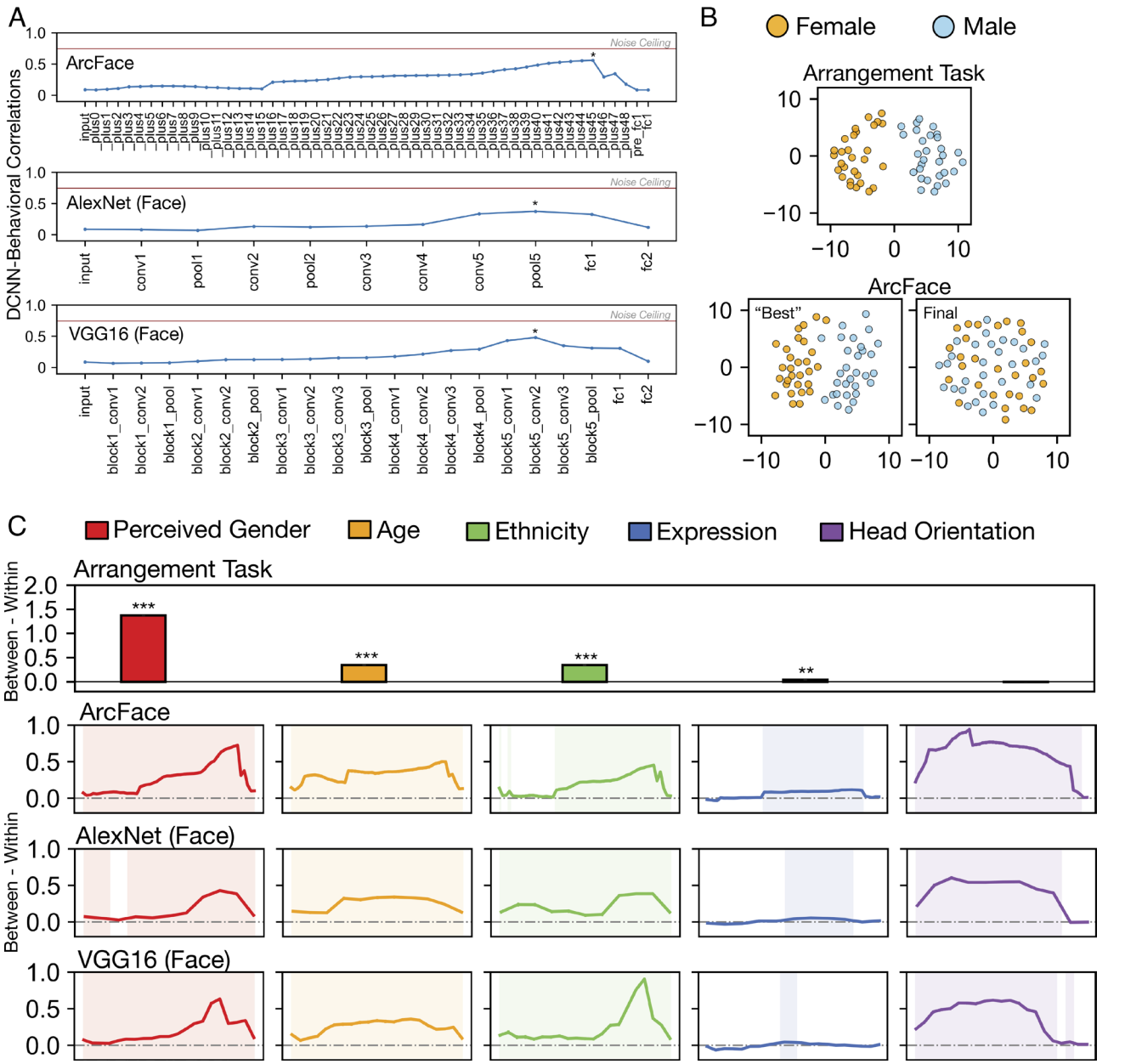

Fig. 2. Correlations between DCNN and behavioral RDMs and between neural and behavioral RDMs. (*A*) Mean correlations across participants and runs between the behavioral and DCNN RDMs in each layer in all three face-DCNNs. The star marks the layer that has the highest correlation with the behavioral task in each DCNN. The red horizontal line in each subplot represents the mean noise ceiling of the behavioral arrangement task across runs. (*B*) Example MDS plots using RDMs of the same run in the behavioral arrangement task, the "best" layer that showed the highest correlation with the behavioral RDM (_plus45) in ArcFace, and the final layer in ArcFace. 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. (*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, and each layer of the three face-trained DCNNs (ArcFace, AlexNet, VGG16). These differences were calculated within each run and then averaged across runs. Shaded layers show significant differences in the between- vs. within-group test ($P < 0.05$, permutation test, one-tail). Error bars indicate the SEM estimated by bootstrap resampling the stimuli (the error bars are too small to be visible in some cases). $^{**}P < 0.01$ and $^{***}P < 0.001$.

We investigated the nature of the face information that is shared across DCNN and behavioral arrangement RDMs by examining the role of categorical face attributes—perceived gender, age, ethnicity, expression, and head orientation—in face-DCNN and behavioral arrangement RDMs. The contribution of each face feature to the representational geometries was quantified by computing z-scored spatial distances within and between feature groups. Fig. 2*B* 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).

Behavioral arrangement RDMs were largely driven by perceived gender, followed by age and ethnicity of faces (Fig. 2*C*). Expression played only a minor role in similarity arrangements, and head orientation played no role. Gender, age, and ethnicity were most strongly reflected in the late-intermediate layers of face-DCNNs (Fig. 2*C*) but only weakly reflected in object-DCNNs (*SI Appendix*, Fig. S3*D*). Head orientation, by contrast, was more strongly reflected in early intermediate layers of face-DCNNs. Thus, the reliable representational geometries in late-intermediate layers of face-DCNNs carry information about cognitive representations that reflect major face categorical attributes. Importantly, representations of these categorical attributes strongly contribute to the similarities between face-DCNNs and the behavioral clustering of perceived similarity.

Correlations of Neural Representations with DCNN and Cognitive Representations. We analyzed relationships between neural representational and DCNN geometries, on the one hand, and between neural representational and cognitive geometries, on the other hand. We first correlated neural RDMs in all cortical searchlights with the ArcFace RDMs in each layer. Results showed that 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 Fig. 3 *A* and *B*). These regions largely correspond to the previously reported human face processing system consisting of multiple face-selective regions (20, 21, 33, 36, 37).

We independently defined face-selective regions using a dynamic face localizer (face vs. objects; *SI Appendix*, Fig. S6) (24, 33) 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. Correlations dropped drastically after layer _plus45 and reached their lowest values in the final fully connected layers (Fig. 3*C*). 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, and similarly for other category-selective visual areas (e.g., body-selective areas. *SI Appendix*, Fig. S16).

Next, we tested whether a specific DCNN architecture had a significant effect on the similarities between DCNN and human neural representations, we performed a similar analysis using two other face-DCNNs (AlexNet and VGG16) and found similar results across layers (Fig. 3*D*). Following the same analysis we used for face-DCNNs, we calculated correlations between RDMs in each layer of the two object-DCNNs and neural representations in each searchlight across the cortical sheet. Similarly, mean correlation coefficients were calculated for the face-selective and non-face-selective areas. Correlating object-DCNN and neural RDMs generated comparable correlations to those between face-DCNN and neural RDMs (*SI Appendix*, Fig. S3*C*), indicating that DCNN-neural RDM correlations were not influenced by the type of the DCNNs (face-DCNNs vs. object-DCNNs).

For all DCNNs, intermediate layers provided a markedly better model for the neural representations than final fully connected layers. This finding is consistent with previous work examining face representation in the brain and DCNNs (5, 10, 38, 39). Interestingly, however, for all DCNNs, correlations between any layer and neural representations accounted for only a fraction of the meaningful variance (all Pearson's $r < 0.1$). The correlation values were especially low compared to the reliable noise ceilings of neural representations (*SI Appendix*, Figs. S7 and S8, the noise ceiling was ~0.4 on average for face-selective areas, with some areas exceeding 0.7, and the DCNN-neural correlations were always <0.1). Additionally, 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 (*SI Appendix*, Fig. S12), suggesting that the sequence of representational geometries in the face-DCNN layers differs from the progression of representational geometries along the neural face pathways (20, 34, 40, 41). The DCNN-neural correlations in each individual face-selective ROI for each layer in both face- and object-DCNNs also showed that none of the ROIs had correlations that approached the noise ceiling (Pearson's $r < 0.1$ in all cases). An additional variance partitioning analysis revealed that the variance in neural RDMs is minimally explained by DCNNs, behavioral models, or by a combination of the two (*SI Appendix*, Fig. S13).

We conducted an additional analysis to investigate whether the low correlations were due to RSA's inherent assumption of equal weights or scales for all features comprising the two RDMs (42–45) (see *Materials and Methods* for details). This additional analysis generated similar results as using classic RSA, excluding the possibility that the low correlations are due to the particular assumption of RSA (*SI Appendix*, Fig. S9). To examine whether more distributed brain activity patterns might lead to a better match between the DCNN and neural 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 explained, *SI Appendix*, Figs. S10 and S11).

Correlations between the behavioral arrangement and neural representations in the searchlight analysis consistently showed face-selective areas had significantly higher correlations than other areas (Fig. 3*E*, permutation test, $P < 0.001$). However, the actual correlation values remained small ($r < 0.1$), suggesting that a major difference existed between clustering according to facial similarity and the extended processing of dynamic faces.

The strength of categorical face attributes in neural representational geometries showed a distribution across face-selective ROIs (Fig. 3*G*) that was consistent with known specialization. Face ROIs in the right hemisphere represented these features more prominently than did face ROIs in the left hemisphere. Identity-related invariant categorical face features, such as perceived gender and age, were significantly represented in bilateral face areas in the ventral temporal cortex [OFA (occipital face area), pFFA, aFFA (posterior and anterior fusiform face areas)], as well as in the pSTS (posterior superior temporal sulcus) and right IFG (inferior frontal face areas) (Fig. 3 *F* and *G*). Expression was significantly represented in face areas in the OFA, pSTS, aSTS (posterior and anterior superior temporal sulcus), and right IFG, but not the FFA. Head orientation was significantly represented in the OFA, right pFFA and mFFA, and the pSTS. Although neural representations

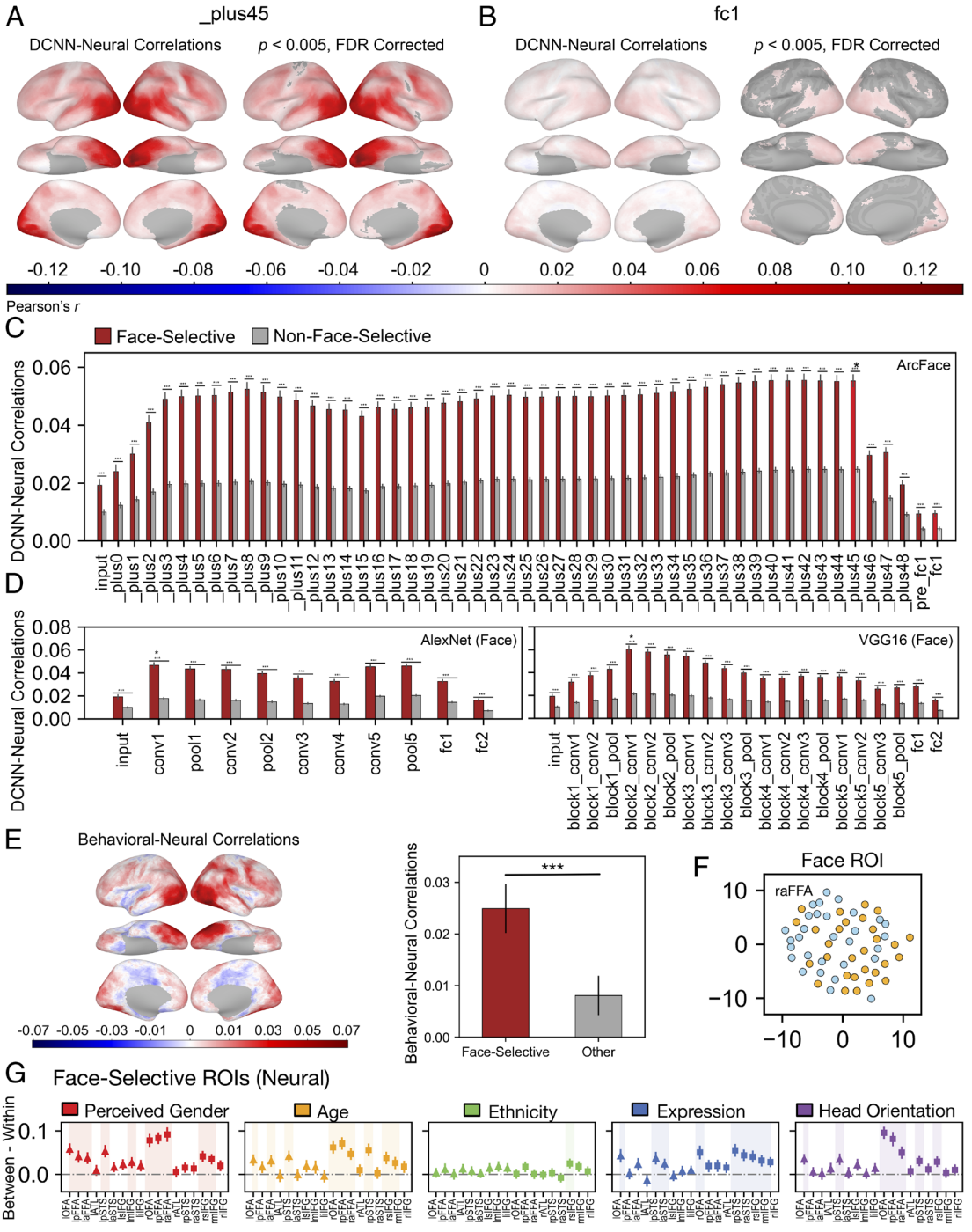

Fig. 3. Correlations between DCNN and neural RDMs. (*A* and *B*) The DCNN-neural correlations across all cortical searchlights using RDMs in layer _plus45 (output of the last stage 3 residual unit) and fc1 (the last fully connected layer). Both layers are highlighted in panel C. Correlations in the visual 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 ($t <= 5$) are plotted as red and gray bars respectively for each layer. The error bar length stands for one SEM estimated by bootstrap resampling of stimuli. The 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. (*D*) Average correlations for face-selective regions and non-face-selective regions for each layer in the two face-DCNNs. Regions, significance, and the color code were defined the same as in panel *C*. Stars indicate the layers that had the largest correlations. (*E*) The neural-behavioral correlation values in the cortex and the mean behavioral-neural correlations in face-selective (red) and non-face-selective (gray) areas. The error bars indicate the SEM estimated by bootstrap resampling the stimuli. The significance of the difference between the two bars was estimated using a permutation test randomizing the stimulus labels. (*F*) Example MDS plots using RDMs for the same run in the right aFFA. Each dot is a stimulus. Orange and blue dots indicate perceived females and males, respectively, the same as Fig. 2*B*. Neural RDMs in this analysis were mean RDMs across participants. (*G*) 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 each face-selective regions of interest (ROI, bilateral OFA, pFFA, aFFA, ATL, pSTS, aSTS, sIFG, mIFG, and iIFG). These differences were calculated within each run and then averaged across runs. Shaded ROIs show significant differences in the between- vs. within-group test ($P < 0.05$, permutation test, one-tail). The significance of the difference was estimated based on a random permutation test randomizing the stimulus labels. Error bars represent one SEM estimated by bootstrap resampling stimuli. Left triangles are nine face-selective ROIs in the left hemisphere, and right squares are face-selective ROIs in the right hemisphere. ***$P < 0.001$.

contained significant information for all categorical attributes, this categorical information was more weakly represented than in behavioral and DCNN representations in intermediate layers.

Discussion

State-of-the-art DCNNs trained to perform face identification 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. Previous neuroscientific studies mainly focused on face representations that are produced by rapid processing of still images. One previous study (13) investigated the representations produced by more extended processing of naturalistic, dynamic faces. Here, we investigated the extent to which DCNNs can model real-world face processing by comparing representational geometries produced by DCNNs to representational geometries produced by brain and behavioral responses to a large, varied set of naturalistic face videos.

Our results showed that DCNN, behavioral, and neural representational geometries were stable and information-rich. Face-DCNNs and behavioral representations of perceived similarities captured shared information about face categorical attributes. This face categorical knowledge was strongly represented in late-intermediate layers of face-DCNNs. By contrast, the final fully connected layers, which are optimized for face identification, did not carry much categorical information. Brain responses to the face videos had representational geometries that were highly reliable across participants, and reliabilities were highest in face-selective cortical areas. Information in the neural representational geometries was significantly correlated with the information in DCNN and behavioral geometries, albeit weakly. The DCNN-neural correlations had a meaningful cortical distribution, following the full distributed face system in occipital, temporal, and frontal cortices, and were much stronger for late-intermediate layers than for the final fully connected layers. Correlations between DCNN representational geometries and the other two (cognitive and neural) geometries were much stronger for late-intermediate layers too, suggesting that the optimization in the final fully connected DCNN layers for recognition of identity is a poor model for how human cognitive and neural systems individuate faces. Although the identification ability of face DCNNs successfully generalizes across stimulus sets (*SI Appendix*, Fig. S2), the representations of the final fully connected layers may nonetheless be biased toward specific training datasets. The idiosyncratic image statistics of these training datasets clearly diverge from the statistical structure learned (in evolution or development) by humans, contributing to the discrepancies observed between humans and DCNNs.

The maximal local correlations between face-DCNN and neural representations show that at most only 3% of the meaningful variance (as defined by the noise ceiling) is shared. It is unclear what information in the highly reliable neural representational geometries is unaccounted for by the face-DCNNs. 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 the rapid processing of briefly presented static stimuli, extended processing of dynamic stimuli dramatically alters the neural response to faces both in terms of tuning profiles and representational geometry. Response tuning to static, well-controlled stimuli in face patches is dominated by the presence or absence of faces or their static structural features (40), but tuning for dynamic face stimuli is dominated by biological motion (14, 18, 46–48). In addition, dynamic faces are superior to static faces for the localization of face-selective areas (23, 24, 33, 49), indicating that they better or more fully engage face-related neural processes. In a similar vein, representational geometry in the ventral temporal cortex for static images of animals is dominated by animal category, but representational geometry for 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 (16, 30). Further research is needed to precisely characterize how these dynamics change the geometry of face representation. While the processing of static face stimuli may appear to be well-modeled by current DCNNs, extended processing of naturalistic stimuli may reveal the deficiency of such models and help focus our attention on how best to improve them.

Temporally extended face processing with dynamic videos may recruit a variety of cognitive features. People automatically make inferences about novel faces—trustworthiness, competence, and attractiveness—that can distort representational geometry (50, 51). Person knowledge plays a large role in the representation of familiar faces (20, 21, 52–55). Familiarity is also known to distort face representations (56, 57), and similarity of novel faces to familiar faces may influence perception and attribution. Faces also play a role in directing attention (37, 58–60), and attention has a large effect on neural responses to faces (61–63) 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 interaction.

Behavioral performance in the arrangement task was dominated by major categorical face attributes of perceived gender, age, and ethnicity. These categorical variables play little role in individuation of face identity. In cognitive models of face perception, such categorical judgments precede processing for individuation (64). 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 (65–67). Thus, the categorical face information that is captured by late-intermediate face-DCNN layers and is correlated with cognitive task performance is weakly related to individuation. Processes for face individuation in late, fully connected layers are powerful but less related to human cognitive and neural processes. We performed an additional behavioral similarity rating experiment to measure similarities between faces based on individuation attributes rather than categorical attributes (see the *Materials and Methods* section for details). Results showed that dissimilarity matrices based on these ratings correlated very weakly with DCNN dissimilarity matrices (*SI Appendix*, Figs. S14 and S15), unlike behavioral dissimilarity matrices that included categorical differences (Fig. 2*A*). These results corroborate our finding that categorical information was the primary driver of the correlations between the behavioral arrangement task and face-DCNN representations.

To summarize the results, Fig. 4 illustrates our interpretation of the different components that played a role in the correlations of the DCNN, behavioral, and human neural RDMs. This figure proposes that 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

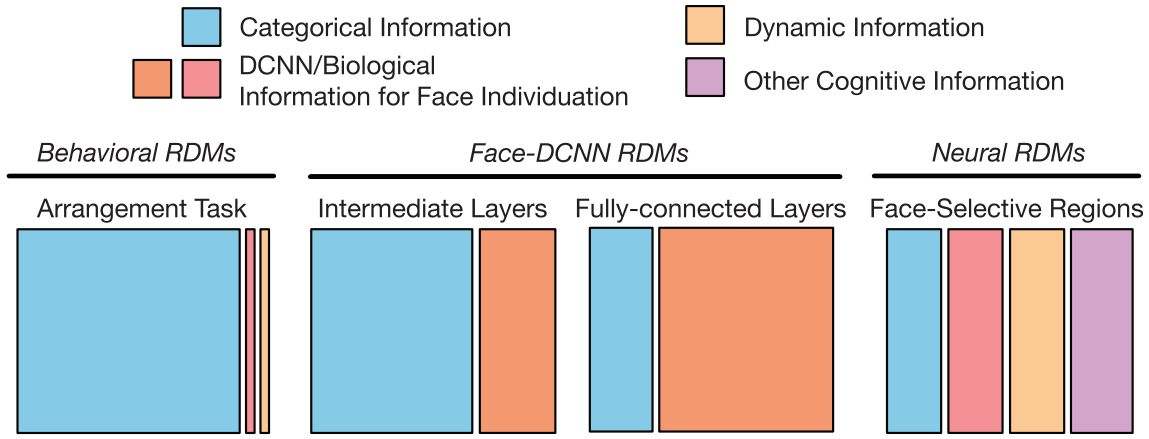

Fig. 4. Schematic illustration of hypothesized components of information in RDMs. In RDMs for the behavioral arrangement task and RDMs for the intermediate layers of face-DCNNs, categorical information (blue) played a major role. Most categorical information was factored out in the fully connected layers of face-DCNNs, and the information for face individuation (orange) became the most prominent part. Neural RDMs in face-selective regions contained similarly weak categorical information as in the final fully connected layers. Successful individuation of faces may only rely on part of facial features, and the information for face individuation in neural RDMs (red) may be different from that in face-DCNNs. Different from face-DCNNs, neural RDMs also contain dynamic (yellow) and cognitive information (purple) that plays an essential role in human face processing. Note that an extra component existed in all RDMs that stood for the unexplained variance and it was omitted for cleaner display.

other hand, the neural RDMs in the face-processing system contained all four kinds of information—categorical information, face individuation (e.g., *SI Appendix*, Fig. S4), 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). Because categorical information was only one component in the neural RDMs, 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. Similar magnitudes of correlation were found between the representations of the behavioral pairwise comparisons and the neural representations (*SI Appendix*, Figs. S14 and S15). Behavioral similarity ratings may reflect the influence of dynamic, individuation, and other attributes of faces on human behavior. These types of information all contribute to the complex neural representations of dynamic, naturalistic faces. Finally, there is shared face identification information between DCNN and the neural RDMs, but the type of information used for face identification in the late, fully connected DCNN layers and in the human face processing system could be quite different. When we narrowed our focus to a smaller stimulus group, matched based on their superordinate categorical information, we observed a decrease in correlation between DCNN and neural representations compared to using the full stimulus set, but the correlations were still substantial, particularly in the later intermediate layers (*SI Appendix*, Figs. S14 and S15). On the other hand, dynamic information or information from other cognitive processes is essential for the human face processing system (23, 24, 50, 51), while this type of information is largely ignored by the face-DCNNs. These differences likely contributed to the low correlations between DCNN RDMs and neural RDMs.

The Fig. 4 suggests a framework for explaining the difference between our results and the results from previous studies that compared DCNNs to brain responses in rapid static face processing tasks. Because dynamic information plays a role in the geometry of brain representations (16, 17, 48), static images could generate higher correlation values between brain responses and DCNNs that do not use motion information (10, 13, 68). Similarly, studies that used stimuli spanning superordinate categories [e.g., with multiple visual categories (12, 45)] would bias representations toward categorical information, reducing the contribution of information that is needed for within-class individuation such as face identification.

None of the existing behavioral tasks or computational models that we tested resulted in a strong alignment with neural representations. There are several possible ways forward. First, future work could examine whether subject-specific behavioral judgments better account for variance in neural RDMs in the same subjects (69). Here, for practical reasons, behavioral data were not collected in the same subjects that participated in the fMRI experiment. Second, arrangement tasks with different behavioral goals beyond simple similarity in facial appearance could provide access to features in the neural representational space that are currently inaccessible or underexplored. Any behavioral judgment task necessarily reduces complex, high-dimensional neural representations associated with face stimuli into a low-dimensional (often unidimensional) behavioral output. By definition, no single low-dimensional behavioral task can explain high-dimensional neural representations that support many different tasks. However, these options are limited due to practical considerations. The large number of stimuli in each trial imposes a high cognitive load, making it challenging for participants to use more nuanced information within a reasonable timeframe.

Although face-DCNNs are trained on an exceptionally large set 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 (70), 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 (68, 71–73), requiring not just face recognition, but extending to facial dynamics, attention, memory, social context, and social judgments, will learn face representations that afford a

more ecologically valid model that better captures the representations of the face processing system in humans.

Materials and Methods

Participants. Twenty-one participants (mean age 27.3 y, range 22 to 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. All participants provided written informed consent.

Experimental Design. The Grand Budapest Hotel and localizer data were also used in prior work by Jiahui and colleagues (33). The MRI data acquisition parameters, preprocessing, and data analysis methods involving these two datasets 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 to 13 min) with audio. These data were curated and made publicly available for research use (25).

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 were removed from all analyses.

Dynamic localizer. Participants watched 3 s dynamic clips of faces, bodies, scenes, objects, and scrambled objects (24). 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 MTurk workers performed this task. Stimuli in a scanning run (59 stimuli for run 1 to 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 MTurk 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 to 10, 11 to 20, 21 to 30, 31 to 40, 41 to 50, 51 to 60, 61 to 70, and 70+), ethnicity (White, Black or African American, Asian, Indian, Hispanic or Latino, and Other), expression (Neutral, Happiness, Surprise, Anger, Disgust, Sadness, and Fear), and overall head orientation (Mostly Left, Mostly Center, and 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.

Behavioral pairwise comparison task. We designed a behavioral pairwise comparison task to limit the use of categorical features and prioritize dynamic, individuation features. Four additional independent groups of online workers (25 workers in each group, total = 100) on Prolific (https://www.prolific.co/) completed this task. Stimuli in the four groups were selected to minimize within-group differences in categorical information (white females, age 21 to 30, happy, facing right: 18 clips; black males, age 21 to 30, neutral expression, facing right: 15 clips; black females, age 21 to 40, happy or neutral faces, facing right: 16 clips; white males, age 21 to 30, neutral faces, facing right: 20 clips). Dynamic stimuli were presented side-by-side on the screen. After the video clips played once, a slider bar (range 1 to 10) below the two clips could be dragged to indicate the similarity of the two faces. The videos were replayed on loop until the "Next" button was hit to move to the next trial. Each pair of faces was displayed twice with counterbalanced left and right positions on the screen.

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 prescan normalization, fat suppression, multiband (i.e., simultaneous multislice) acceleration factor of 4 (using blipped CAIPIRINHA), and no in-plane acceleration (i.e., GRAPPA acceleration factor of one): TR/TE = 1,000/33 ms, flip angle = 59°, resolution = 2.5 mm^3 isotropic voxels, matrix size = 96 × 96, FoV = 240 × 240 mm, 52 axial slices with full brain coverage and no gap, anterior–posterior phase encoding. See SI for details on the MRI data acquisition parameters.

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 (74), VGG16 (75), and ResNet100 (76). See *SI Appendix* for details on training of the DCNN.

Data Analysis.

Preprocessing. MRI data were preprocessed using fMRIPrep version 1.4.1 (77). The following confound variables were regressed out of the signal in each run: six motion parameters and their derivatives, global signal, framewise displacement (78), 6 principal components from a combined cerebrospinal fluid and white matter mask (aCompCor) (79), and up to second-order polynomial trends. See *SI Appendix* for details on the preprocessing steps.

Searchlight hyperalignment. All three imaging datasets were hyperaligned (26–29) based on responses to the Grand Budapest Hotel (*SI Appendix*, Fig. S1). See *SI Appendix* for details on the steps for hyperalignment.

Searchlight RSA. We performed a searchlight RSA 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 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 (Fig. 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 1,000 times for each intermediate layer and 5,000 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 performed for each individual scanning run, and the correlation maps were averaged across runs.

Correlations in the category-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 (*SI Appendix*, Fig. S6), and applied

a conservative threshold of $t > 5$ to obtain the face-selective regions. Other category-selective ROIs were localized following the same steps based on the contrasts of bodies, scenes, and objects vs. all the other categories, respectively. Individual face-selective ROIs including the OFA, the aFFA and pFFA, the anterior temporal face area (ATL), the posterior and anterior superior temporal sulcus (pSTS and aSTS), and three IFG (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 *SI Appendix*, Figs. S10 and S11). 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. SEM were calculated by bootstrapping the stimuli 1,000 times for each intermediate layer and 5,000 times for the final fully connected layer. Statistical significance was assessed by permutation tests randomizing the stimulus labels 1,000 times for each intermediate layer and 5,000 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 (80). Because we averaged individuals' RDMs before RSA analysis, the noise ceiling was estimated by calculating Cronbach's alpha using neural RDMs across participants (81). 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) (43, 44). We tested whether relaxing this assumption might yield larger DCNN-neural correlations. See SI for detailed reweighting steps.

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). See *SI Appendix* for details on the steps for the decoding 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.

Behavioral pairwise comparison task RDMs and noise ceilings. Behavioral pairwise RDMs were constructed based on the slider bar ratings for each participant in each stimulus group. Each element of the behavioral pairwise RDMs reflected the perceived similarity rating between a given pair of stimuli. Individual behavioral pairwise comparison task RDMs were averaged across participants before further analysis. RSA results were averaged across the four stimuli groups as well as the noise ceilings that were estimated based on Cronbach's alpha across participants in each group.

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. See *SI Appendix* for details on the steps for the variance partitioning analysis.

MDS. MDS was used to visualize the representational geometry of face stimuli for different DCNNs, behavioral arrangements, and neural ROIs. See *SI Appendix* for detailed explanations.

Data, Materials, and Software Availability. All data needed to evaluate the conclusions in the paper are present in the paper and/or *SI Appendix*. Further data and codes can be downloaded from https://github.com/GUO-Jiahui/face_DCNN. Previously published data were used for this work (25).

ACKNOWLEDGMENTS. We thank Manon de Villemejane and Carlo Cipolli for their invaluable contributions to this research. We thank the authors of the InsightFace package for making their models and training data freely available for noncommercial research use. This work was supported by NSF grants 1607845 (J.V.H) and 1835200 (M.I.G) and by NIMH grant 5R01MH127199 (J.V.H. and M.I.G.).

Author affiliations: [a]Center for Cognitive Neuroscience, Dartmouth College, Hanover, NH 03755; [b]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720; [c]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544; [d]Department of Medical and Surgical Sciences, University of Bologna, Bologna 40138, Italy; and [e]Istituti di Ricovero e Cura a Carattere Scientifico, Istituto delle Scienze Neurologiche di Bologna, Bologna 40139, Italia

1. O. M. Parkhi, A. Vedaldi, A. Zisserman, *Deep face recognition* (British Machine Vision Association, 2015), pp. 41.1–41.12.
2. P. J. Phillips *et al.*, Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6171–6176 (2018).
3. Y. Taigman, M. Yang, M. Ranzato, L. Wolf, "DeepFace: Closing the gap to human-level performance" in *IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH, USA, 2014), pp. 1701–1708.
4. A. J. O'Toole, C. D. Castillo, C. J. Parde, M. Q. Hill, R. Chellappa, Face space representations in deep convolutional neural networks. *Trends Cogn. Sci.* **22**, 794–809 (2018).
5. R. Raman, H. Hosoya, Convolutional neural networks explain tuning properties of anterior, but not middle, face-processing areas in macaque inferotemporal cortex. *Commun. Biol.* **3**, 1–14 (2020).
6. M. Schrimpf *et al.*, Brain-Score: Which artificial neural network for object recognition is most brain-like? bioRxiv [Preprint] (2020). https://doi.org/10.1101/407007 (Accessed 28 July 2021).
7. D. L. K. Yamins *et al.*, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
8. I. Yildirim, M. Belledonne, W. Freiwald, J. Tenenbaum, Efficient inverse graphics in biological face processing. *Sci. Adv.* **6**, eaax5979 (2020).
9. K. Dobs, J. Martinez, A. J. E. Kell, N. Kanwisher, Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. Adv.* **8**, eabl8913 (2022).
10. S. Grossman *et al.*, Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 4934 (2019).
11. I. Kuzovkin *et al.*, Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Commun. Biol.* **1**, 1–12 (2018).
12. N. Apurva *et al.*, Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat. Commun.* **12**, 5540 (2021).
13. M. Tsantani *et al.*, FFA and OFA encode distinct types of face identity information. *J. Neurosci.* **41**, 1952–1969 (2021) 10.1523/JNEUROSCI.1449-20.2020.
14. S. H. Park *et al.*, Functional subpopulations of neurons in a macaque face patch revealed by single-unit fMRI mapping. *Neuron* **95**, 971–981.e5 (2017).
15. S. H. Park *et al.*, Parallel functional subnetworks embedded in the macaque face patch system. *Sci. Adv.* **8**, eabm2054 (2022).
16. S. A. Nastase *et al.*, Attention selectively reshapes the geometry of distributed semantic representation. *Cereb. Cortex* **27**, 4277–4291 (2017).
17. J. V. Haxby, M. I. Gobbini, S. A. Nastase, Naturalistic stimuli reveal a dominant role for agentic action in visual representation. *NeuroImage* **216**, 116561 (2020).
18. B. E. Russ, K. W. Koyano, J. Day-Cooney, N. Perwez, D. A. Leopold, Temporal continuity shapes visual responses of macaque face patch neurons. *Neuron* **111**, 903–914.e3 (2023), 10.1016/j.neuron.2022.12.021.
19. M. Visconti di Oleggio Castello, "Characterizing feature representations in the human face-processing network with multivariate analyses and encoding models," Doctoral dissertation (Dartmouth College, 2018).
20. M. Visconti di Oleggio Castello, Y. O. Halchenko, J. S. Guntupalli, J. D. Gors, M. I. Gobbini, The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. *Sci. Rep.* **7**, 12237 (2017).
21. M. Visconti di Oleggio Castello, J. V. Haxby, M. I. Gobbini, Shared neural codes for visual and semantic information about familiar faces in a common representational space. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2110474118 (2021).
22. J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database" in *IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL, USA, 2009), pp. 248–255.
23. C. J. Fox, G. Iaria, J. J. S. Barton, Defining the face processing network: Optimization of the functional localizer in fMRI. *Hum. Brain Mapp.* **30**, 1637–1651 (2009).
24. D. Pitcher, D. D. Dilks, R. R. Saxe, C. Triantafyllou, N. Kanwisher, Differential selectivity for dynamic versus static information in face-selective cortical regions. *NeuroImage* **56**, 2356–2363 (2011).
25. M. Visconti di Oleggio Castello, V. Chauhan, G. Jiahui, M. I. Gobbini, An fMRI dataset in response to "The Grand Budapest Hotel", a socially-rich, naturalistic movie. *Sci. Data* **7**, 383 (2020).
26. M. Feilong, S. A. Nastase, J. S. Guntupalli, J. V. Haxby, Reliable individual differences in fine-grained cortical functional architecture. *NeuroImage* **183**, 375–386 (2018).
27. J. S. Guntupalli *et al.*, A model of representational spaces in human cortex. *Cereb. Cortex* **26**, 2919–2934 (2016).
28. J. S. Guntupalli, M. Feilong, J. V. Haxby, A computational model of shared fine-scale structure in the human connectome. *PLOS Comput. Biol.* **14**, e1006120 (2018).
29. J. V. Haxby *et al.*, A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).

30. J. V. Haxby, J. S. Guntupalli, S. A. Nastase, M. Feilong, Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife* **9**, e56601 (2020).
31. S. Kornblith, M. Norouzi, H. Lee, G. Hinton, Similarity of neural network representations revisited. arXiv [Preprint] (2019). https://doi.org/10.48550/arXiv.1905.00414. (Accessed 26 October 2020).
32. J. Mehrer, C. J. Spoerer, N. Kriegeskorte, T. C. Kietzmann, Individual differences among deep neural network models. *Nat. Commun.* **11**, 5725 (2020).
33. G. Jiahui *et al.*, Predicting individual face-selective topography using naturalistic stimuli. *NeuroImage* **216**, 116458 (2020).
34. J. S. Guntupalli, K. G. Wheeler, M. I. Gobbini, Disentangling the representation of identity from head view along the human face processing pathway. *Cereb. Cortex* **27**, 46–53 (2017).
35. C. Conwell, J. S. Prince, K. N. Kay, G. A. Alvarez, T. Konkle, What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? bioRxiv [Preprint] (2023). https://doi.org/10.1101/2022.03.28.485868 (Accessed 16 August 2023).
36. J. V. Haxby, M. I. Gobbini, "Distributed neural systems for face perception" in *Oxford Handbook of Face Perception, Oxford Library of Psychology*, (Oxford University Press, 2011), pp. 93–110.
37. J. V. Haxby, E. A. Hoffman, M. I. Gobbini, The distributed human neural system for face perception. *Trends Cogn. Sci.* **4**, 223–233 (2000).
38. M. R. Greene, B. C. Hansen, Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS Comput. Biol.* **14**, e1006327 (2018).
39. M. F. Bonner, R. A. Epstein, Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Comput. Biol.* **14**, e1006111 (2018).
40. L. Chang, D. Y. Tsao, The code for facial identity in the primate brain. *Cell* **169**, 1013–1028.e14 (2017).
41. W. A. Freiwald, D. Y. Tsao, Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* **330**, 845–851 (2010).
42. C. Conwell, J. S. Prince, G. A. Alvarez, T. Konkle, "What can 5.17 billion regression fits tell us about artificial models of the human visual system?" in *NeurIPS* (SVRHM, 2021). https://openreview.net/forum?id=i_xiyGq6FNT.
43. P. Kaniuth, M. N. Hebart, Feature-reweighted representational similarity analysis: A method for improving the fit between computational models, brains, and behavior. *Neuroimage* **257**, 119294 (2022).
44. S.-M. Khaligh-Razavi, L. Henriksson, K. Kay, N. Kriegeskorte, Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *J. Math. Psychol.* **76**, 184–197 (2017).
45. T. Konkle, G. A. Alvarez, A self-supervised domain-general learning framework for human ventral stream representation. *Nat. Commun.* **13**, 491 (2022).
46. D. A. Leopold, S. H. Park, Studying the visual brain in its natural rhythm. *NeuroImage* **216**, 116790 (2020).
47. D. B. T. McMahon, B. E. Russ, H. D. Elnaiem, A. I. Kurnikova, D. A. Leopold, Single-unit activity during natural vision: Diversity, consistency, and spatial sensitivity among AF face patch neurons. *J. Neurosci.* **35**, 5537–5548 (2015).
48. B. E. Russ, D. A. Leopold, Functional MRI mapping of dynamic visual features during natural viewing in the macaque. *NeuroImage* **109**, 84–94 (2015).
49. U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, R. Malach, Intersubject synchronization of cortical activity during natural vision. *Science* **303**, 1634–1640 (2004).
50. N. N. Oosterhof, A. Todorov, The functional basis of face evaluation. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 11087–11092 (2008).
51. A. Todorov, C. P. Said, A. D. Engell, N. N. Oosterhof, Understanding evaluation of faces on social dimensions. *Trends Cogn. Sci.* **12**, 455–460 (2008).
52. M. I. Gobbini, J. V. Haxby, Neural systems for recognition of familiar faces. *Neuropsychologia* **45**, 32–41 (2007).
53. M. I. Gobbini, E. Leibenluft, N. Santiago, J. V. Haxby, Social and emotional attachment in the neural representation of faces. *NeuroImage* **22**, 1628–1635 (2004).
54. E. Leibenluft, M. I. Gobbini, T. Harrison, J. V. Haxby, Mothers' neural activation in response to pictures of their children and other children. *Biol. Psychiatry* **56**, 225–232 (2004).
55. M. Ramon, M. I. Gobbini, Familiarity matters: A review on prioritized processing of personally familiar faces. *Vis. Cogn.* **26**, 179–195 (2018).
56. V. Chauhan, I. Kotlewska, S. Tang, M. I. Gobbini, How familiarity warps representation in the face space. *J. Vis.* **20**, 18 (2020).
57. M. Visconti di Oleggio Castello, M. Taylor, P. Cavanagh, M. I. Gobbini, Idiosyncratic, retinotopic bias in face identification modulated by familiarity. *eNeuro* **5**, ENEURO.0054-18.2018 (2018).
58. J. D. Carlin, A. J. Calder, N. Kriegeskorte, H. Nili, J. B. Rowe, A head view-invariant representation of gaze direction in anterior superior temporal sulcus. *Curr. Biol.* **21**, 1817–1821 (2011).
59. C. K. Friesen, A. Kingstone, The eyes have it! Reflexive orienting is triggered by nonpredictive gaze *Psychon. Bull. Rev.* **5**, 490–495 (1998).
60. E. A. Hoffman, J. V. Haxby, Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nat. Neurosci.* **3**, 80–84 (2000).
61. M. L. Furey *et al.*, Dissociation of face-selective cortical responses by attention. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 1065–1070 (2006).
62. G. Jiahui, H. Yang, B. Duchaine, Attentional modulation differentially affects ventral and dorsal face areas in both normal participants and developmental prosopagnosics. *Cogn. Neuropsychol.* **37**, 1–12 (2020).
63. N. Kanwisher, E. Wojciulik, Visual attention: Insights from brain imaging. *Nat. Rev. Neurosci.* **1**, 91–100 (2000).
64. V. Bruce, A. Young, Understanding face recognition. *Br. J. Psychol.* **77**, 305–327 (1986).
65. T. Kress, I. Daum, Developmental prosopagnosia: A review. *Behav. Neurol.* **14**, 109–121 (2003).
66. A.-R. Richoz, R. E. Jack, O. G. B. Garrod, P. G. Schyns, R. Caldara, 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 (2015).
67. Z. Little, C. J. Palmer, T. Susilo, Intact gaze processing in developmental prosopagnosia. *J. Vis.* **21**, 2267 (2021).
68. C. Daube *et al.*, Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns* **2**, 100348 (2021).
69. I. Charest, R. A. Kievit, T. W. Schmitz, D. Deca, N. Kriegeskorte, Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14565–14570 (2014).
70. M. Visconti di Oleggio Castello, M. I. Gobbini, Familiar face detection in 180ms. *PLOS One* **10**, e0136548 (2015).
71. U. Hasson, S. A. Nastase, A. Goldstein, Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron* **105**, 416–434 (2020).
72. R. Ranjan, V. M. Patel, R. Chellappa, HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. arXiv [Preprint] (2017). https://doi.org/10.48550/arXiv.1603.01249. (Accessed 13 November 2021).
73. C. Zhuang *et al.*, Unsupervised neural network models of the ventral visual stream. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2014196118 (2021).
74. A. Krizhevsky, One weird trick for parallelizing convolutional neural networks. arXiv [Preprint] (2014). https://doi.org/10.48550/arXiv.1404.5997. (Accessed 16 September 2021).
75. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv [Preprint] (2015). https://doi.org/10.48550/arXiv.1409.1556. (Accessed 31 January 2021).
76. K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks. arXiv [Preprint] (2016). https://doi.org/10.48550/arXiv.1603.05027. (Accessed 19 November 2020).
77. O. Esteban *et al.*, fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111 (2019).
78. J. D. Power *et al.*, Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage* **84**, 320–341 (2014).
79. Y. Behzadi, K. Restom, J. Liau, T. T. Liu, A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).
80. H. Nili *et al.*, A toolbox for representational similarity analysis. *PLOS Comput. Biol.* **10**, e1003553 (2014).
81. L. J. Cronbach, Coefficient alpha and the internal structure of tests. *Psychometrika* **16**, 297–334 (1951).