Behavioral/Cognitive

Intracranial Electroencephalography and Deep Neural Networks Reveal Shared Substrates for Representations of Face Identity and Expressions

Emily Schwartz,[1] Arish Alreja,[2,3,4,5] R. Mark Richardson,[6,7] Avniel Ghuman,[2,5,8] and Stefano Anzellotti[1]
[1]Department of Psychology and Neuroscience, Boston College, Chestnut Hill, Massachusetts 02467, [2]Center for the Neural Basis of Cognition, Carnegie Mellon University/University of Pittsburgh, Pittsburgh, Pennsylvania 15213, [3]Neuroscience Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, [4]Machine Learning Department, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, [5]Department of Neurological Surgery, University of Pittsburgh Medical Center Presbyterian, Pittsburgh, Pennsylvania 15213, [6]Department of Neurosurgery, Massachusetts General Hospital, Boston, Massachusetts 02114, [7]Harvard Medical School, Boston, Massachusetts 02115, and [8]Center for Neuroscience, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

According to a classical view of face perception (Bruce and Young, 1986; Haxby et al., 2000), face identity and facial expression recognition are performed by separate neural substrates (ventral and lateral temporal face-selective regions, respectively). However, recent studies challenge this view, showing that expression valence can also be decoded from ventral regions (Skerry and Saxe, 2014; Li et al., 2019), and identity from lateral regions (Anzellotti and Caramazza, 2017). These findings could be reconciled with the classical view if regions specialized for one task (either identity or expression) contain a small amount of information for the other task (that enables above-chance decoding). In this case, we would expect representations in lateral regions to be more similar to representations in deep convolutional neural networks (DCNNs) trained to recognize facial expression than to representations in DCNNs trained to recognize face identity (the converse should hold for ventral regions). We tested this hypothesis by analyzing neural responses to faces varying in identity and expression. Representational dissimilarity matrices (RDMs) computed from human intracranial recordings ($n = 11$ adults; 7 females) were compared with RDMs from DCNNs trained to label either identity or expression. We found that RDMs from DCNNs trained to recognize identity correlated with intracranial recordings more strongly in all regions tested—even in regions classically hypothesized to be specialized for expression. These results deviate from the classical view, suggesting that face-selective ventral and lateral regions contribute to the representation of both identity and expression.

Key words: deep neural networks; face identity recognition; face processing; facial expression recognition; intracranial electroencephalography

Significance Statement

Previous work proposed that separate brain regions are specialized for the recognition of face identity and facial expression. However, identity and expression recognition mechanisms might share common brain regions instead. We tested these alternatives using deep neural networks and intracranial recordings from face-selective brain regions. Deep neural networks trained to recognize identity and networks trained to recognize expression learned representations that correlate with neural recordings. Identity-trained representations correlated with intracranial recordings more strongly in all regions tested, including regions hypothesized to be expression specialized in the classical hypothesis. These findings support the view that identity and expression recognition rely on common brain regions. This discovery may require reevaluation of the roles that the ventral and lateral neural pathways play in processing socially relevant stimuli.

Received June 20, 2022; revised Mar. 25, 2023; accepted Apr. 17, 2023.

Author contributions: E.S., A.A., R.M.R., A.G., and S.A. designed research; A.A., R.M.R., and A.G. performed research; E.S. and S.A. analyzed data; E.S. and S.A. wrote the paper.

This work was supported by a CAREER Grant from National Science Foundation Career Grant 1943862 to S.A., National Institutes of Health Grants R01-MH-107797 and R21-EY-030297 to A.G., and National Science Foundation Grant 1734907 to A.G. We thank the patients for participating in the iEEG experiments and the University of Pittsburgh Medical Center Presbyterian epilepsy monitoring unit staff and administration for assistance and cooperation with our research. We also thank Michael Ward for assistance with the data collection.

The authors declare no competing financial interests.

Correspondence should be addressed to Stefano Anzellotti at stefano.anzellotti@bc.edu.

https://doi.org/10.1523/JNEUROSCI.1277-22.2023

Introduction

Humans are exposed to a multitude of faces every day; each face provides rich information about an individual's identity and emotion. The social importance of faces makes it critical that we understand how we recognize others and their facial expressions.

According to an established hypothesis (henceforth called the "classical" view), face identity and facial expression are processed by distinct, specialized pathways (Bruce and Young, 1986; Haxby et al., 2000). In this view, face-selective regions in ventral temporal cortex ("ventral stream") are specialized for identity recognition, while face-selective regions in lateral temporal cortex ("lateral stream") are specialized for expression recognition (Haxby et al., 2000). Indeed, previous studies indicate that the ventral stream plays a key role in face identity recognition. Response patterns in the ventral stream can be used to decode face identity (Nestor et al., 2011; Anzellotti et al., 2014; Ghuman et al., 2014; Axelrod and Yovel, 2015; Dobs et al., 2018; Li et al., 2019; Boring et al., 2021), and participants with face recognition deficits have reduced structural connectivity in ventral regions (Thomas et al., 2009). In parallel, other studies indicate that the lateral stream plays a role in expression recognition. Facial expression valence can be decoded from a region in lateral temporal cortex: the face-selective posterior superior temporal sulcus (pSTS; Peelen et al., 2010; Skerry and Saxe, 2014). Additionally, patients with pSTS damage experience expression recognition deficits (Fox et al., 2011), suggesting a causal role of the lateral stream in expression recognition.

While these findings support the involvement of the lateral stream in expression recognition, they do not rule out that the ventral stream might also play a role. Similarly, results suggesting ventral stream involvement in identity recognition do not rule out that the lateral stream might contribute to identity recognition. Considering this, an alternative hypothesis suggests that identity and expression are not necessarily independent neural mechanisms (Duchaine and Yovel, 2015). The ventral and lateral streams, instead, might differ in whether they represent form or motion (Duchaine and Yovel, 2015; Pitcher and Ungerleider, 2021). Consistent with this alternative, facial expression can be decoded in ventral face-selective regions (Skerry and Saxe, 2014; Li et al., 2019), and face identity can be decoded in lateral regions (face-selective pSTS; Hasan et al., 2016; Anzellotti and Caramazza, 2017; Dobs et al., 2018). Furthermore, behavioral studies find correlations between expression and identity recognition abilities (Connolly et al., 2019).

Even considering this evidence, it is still possible that ventral and lateral streams might be specialized for identity and expression recognition, respectively. Behavioral correlations between recognition abilities might result from differences in upstream regions before face processing separates into ventral and lateral streams. Furthermore, ventral representations specialized for identity might contain a small amount of expression information that would support fMRI decoding, and vice versa. Compatible with this possibility, computational studies using deep convolutional neural networks (DCNNs) found that identity-trained networks encode some expression information (Colón et al., 2021), and vice versa (Schwartz et al., 2023). In fact, one study found that, in contrast to untrained DCNNs and DCNNs trained to recognize nonface objects, DCNNs trained to recognize face identity have expression-selective units that share similarities with human expression recognition, making similar errors (Zhou et al., 2022). Together with our results, this suggests that identity and expression recognition might share common mechanisms both in the brain and in DCNNs. While DCNNs trained to recognize identity encode some expression information (and vice versa; Colón et al., 2021; Schwartz et al., 2023), DCNNs trained to recognize identity and DCNNs trained to recognize expression still have distinct representations (Fig. 1; see Materials and Methods). If the classical view is correct, representational dissimilarity matrices (RDMs) from identity-trained DCNNs should correlate with RDMs from ventral regions, and, symmetrically, RDMs from expression-trained DCNNs should correlate with RDMs from lateral regions. Critically, there would need to be an interaction between DCNN type (identity or expression trained) and brain region. By contrast, if ventral and lateral regions contribute to both identity and expression recognition, then one would anticipate that the DCNNs should correlate with both ventral and lateral regions, and that there would not necessarily be an interaction between DCNN type and brain region. Furthermore, these conclusions hold if the models either equally correlate with the regions or if one model outperforms the other for both sets of regions. We test this directly by analyzing neural responses measured with intracranial electroencephalography (iEEG) to faces varying in identity and expression. Comparing the representational geometry of neural responses in ventral and lateral regions to the representational geometry in DCNNs trained to recognize identity and expression, we examine whether RDMs extracted from these DCNNs correlate differentially with RDMs based on responses in face-selective electrodes in ventral and lateral regions.

Materials and Methods

Participants

The experimental protocols were approved by the Institutional Review Board of the University of Pittsburgh. Written informed consent was obtained from all participants. Participants were a subset of patients selected a priori from Li et al. (2019) and Boring et al. (2021), who performed two variations of the face individuation task. Eleven human patients (7 females; mean age, 31.8 years; SD, 9.89) underwent surgical placement of electrocorticographic (surface and depth) electrodes for seizure onset localization. One subject was initially excluded because of noisy data (as determined with a reliability analysis described in the Temporal localizer subsection). None of the subjects showed evidence of epileptic activity on electrodes located in the ventral and lateral temporal lobes.

Experimental design and statistical analysis

Stimuli. Subjects viewed face images from the Karolinska Directed Emotional Faces (KDEF) dataset (Lundqvist et al., 1998). The KDEF dataset consists of 4900 images depicting 70 individuals (50% female) showing seven different expressions from five different angles. The following expression categories were included in the experiment: happy, sad, afraid, angry, and neutral. Each combination of a face identity and a facial expression was shown in different viewpoints, including 0° (frontal view), 45° (left and right views), and 90° (profile; left and right views).

Experimental paradigm. Before completing the main task, participants completed a functional localizer task (Li et al., 2019; Boring et al., 2021). Subjects were shown images of faces, houses, bodies, words, hammers, and phase-scrambled faces. More details about the design of the functional localizer can be found in the studies by Li et al. (2019) and Boring et al. (2021). The data from the functional localizer was used to identify electrodes that respond selectively to faces. An electrode was deemed face selective using the criteria described in the Electrode localization section.

Two different sets of participants completed two different versions of the experiment (Li et al., 2019; Boring et al., 2021), which we will refer to as A and B. In both experiments, each trial began with a face image presented for 1000 ms. This was followed by a 500 ms intertrial interval, during which a fixation cross was presented at the center of the screen.

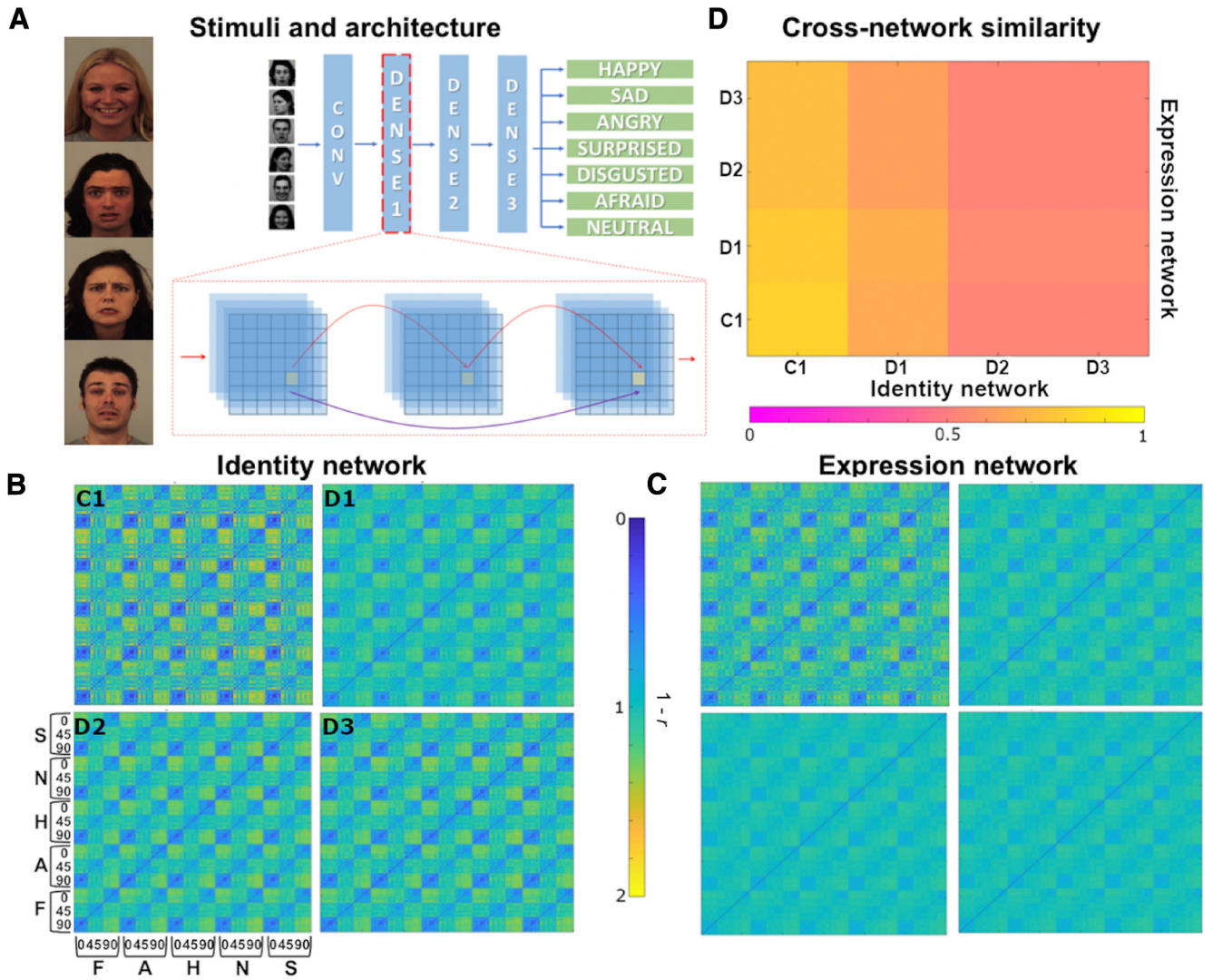

Figure 1. Face representations in a DenseNet trained to recognize identity or expression. ***A***, KDEF stimuli (AF27HAS, AM01AFS, AF06ANS, AM29AFS) and neural network architecture examples. ***B***, RDMs of the identity DenseNet features from KDEF images used in version A of the experiment. ***C***, RDMs of the expression DenseNet features from KDEF images used in version A of the experiment. ***D***, Kendall τ values between identity DenseNet RDMs and expression DenseNet RDMs. Each tick on the horizontal axis represents an identity DenseNet RDM, and each tick on the vertical axis represents an expression DenseNet RDM. C1, Conv 1; D1, dense block 1; D2, dense block 2; D3, dense block 3.

Subjects were instructed to press a button to identify whether the presented face was male or female. Subjects were asked to respond as quickly and as accurately as possible. A set of 10 practice trials was executed before the start of the experiment.

In experiment A, each subject performed one session containing 600 trials. Subjects viewed a set of stimuli that contained eight identities, five expressions, and five viewpoint angles (left/right profile, left/right 45°, and frontal). Each stimulus was presented three times within a session. In experiment B, subjects performed at least two sessions and viewed a different subset of KDEF stimuli. Subjects viewed a set of stimuli that contained 40 identities, five expressions, and three viewpoint angles (profile, 45°, and frontal). Each stimulus was shown only once per session.

Data preprocessing. Data were preprocessed at the University of Pittsburgh. Further details can be found in the studies by Li et al. (2019) and Boring et al. (2021). The data analyzed here contain 14 depth electrodes and 11 surface electrodes. Depth electrodes and surface electrodes were used to record local field potentials at 1000 Hz. Reference and ground electrodes were distantly placed from the recording electrodes subdurally and having contacts oriented toward the dura. Surface area of the recording site was similar across grid and strip electrode contacts. In this manuscript, "electrode contacts" will be referred to as "electrodes." There were no consistent differences in neural responses observed between the grid and depth electrodes. To extract single-trial potential signals, the raw data were bandpass filtered, preserving the frequencies from 0.2 to 115 Hz. This step was implemented using a fourth-order Butterworth filter. After removing slow and linear drift as well as high-frequency noise, a 60 Hz line noise was also removed with 55–65 Hz as the stop band. Single-trial potentials were time locked to the stimulus onset for the trial with the signal sampled at 1000 Hz.

Raw data were also inspected to identify and reduce artifacts. There were no ictal events detected. The mean maximum amplitude across all trials was computed, and any trials with a maximum amplitude 5 SDs above the mean were discarded. Trials that had a difference of $\geq$25 μV between back-to-back sampling instances were discarded as well. This resulted in $<$1% of trials being removed.

Electrode localization. The location of the electrodes (Fig. 2*A*) was determined using an automated method that was used to coregister grid electrodes and electrode strips (Hermes et al., 2010). Patient high-resolution postoperative CT scans were coregistered with anatomic MRI scans to section electrode contacts before patients underwent surgery and implantation of the electrodes. Preoperative and postoperative imaging scans were also used to localize stereo EEG electrodes. Face-selective electrodes were identified by analyzing data from a functional localizer, during which participants were shown images of faces, bodies, words, hammers, houses, and scrambled faces. An electrode was defined as face

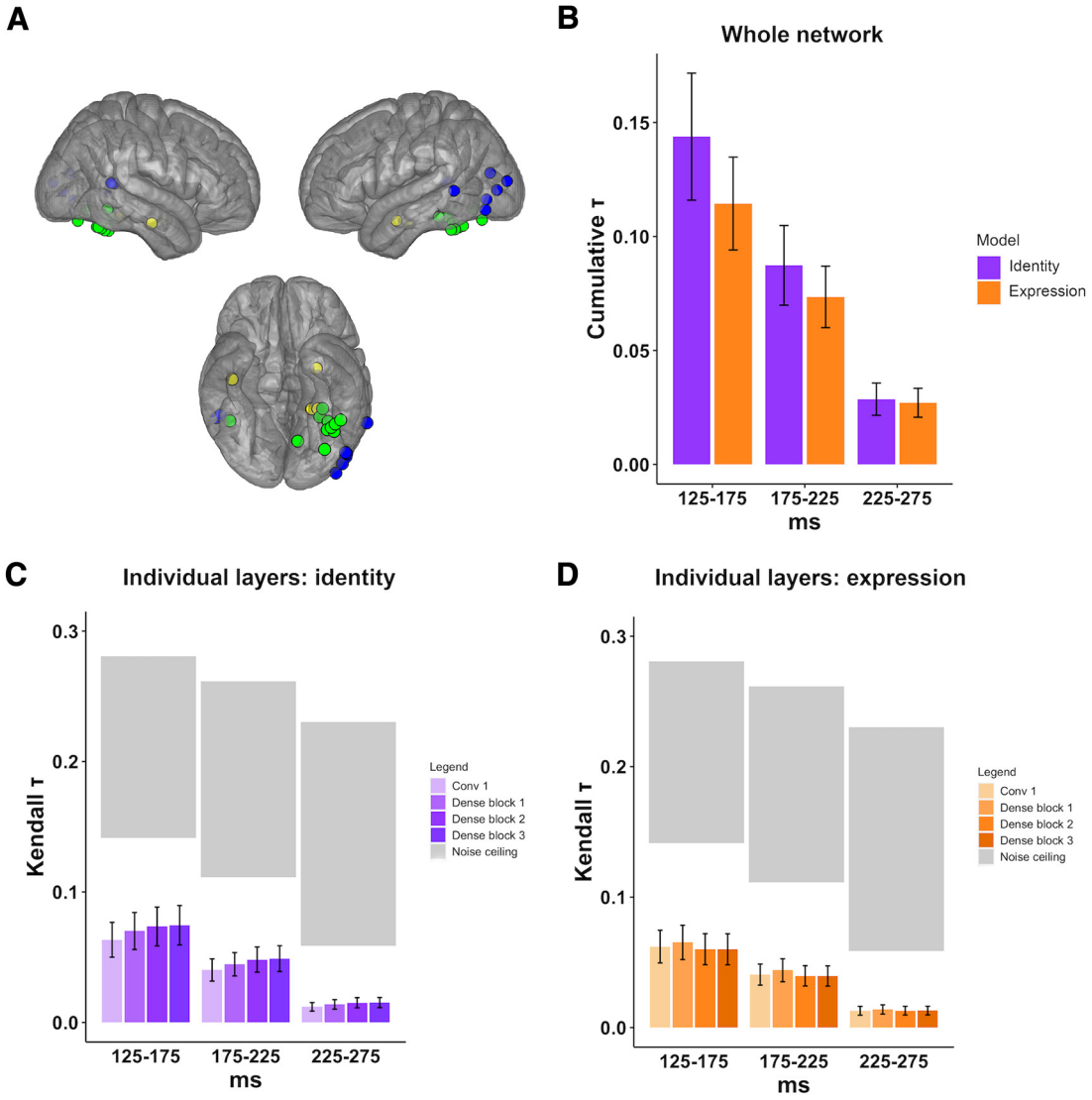

Figure 2. Face-selective electrodes and Kendall τ_B correlations between their representational similarity and the representational similarity in DenseNet layers. ***A***, Face-selective electrode locations ($n = 24$). ***B***, Semipartial τ_B values were computed to examine contribution across layers. This is plotted as a cumulative value obtained from each model and averaged over electrodes. SEM bars are depicted. ***C***, Kendall τ_B values between face-selective iEEG RDMs and layer feature RDMs from the identity DenseNet averaged over electrodes ($n = 24$). SEM bars are depicted. ***D***, Kendall τ_B values between face-selective iEEG RDMs and layer feature RDMs from the expression DenseNet averaged over electrodes ($n = 24$). SEM bars are depicted.

selective if its temporal response patterns could be used to decode faces from other object categories significantly above chance (Li et al., 2019; Boring et al., 2021).

Deep convolutional neural network models. DCNNs were implemented to model the neural data. Each network was trained to perform one task only, either identity recognition or expression recognition. Therefore, identity-trained models will be referred to as identity DCNNs, and the expression-trained models as expression DCNNs. For both the identity and expression DCNNs, we used a densely connected architecture (DenseNet; Huang et al., 2017; Fig. 1*A*) as well as a residual neural network (ResNet-18) architecture.

The identity DCNNs were trained to label identities using the CelebA dataset (http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html). CelebA consists of >300,000 images. To match the size of the dataset used for the two networks, a subset of CelebA was used. The subset contained 28,709 images for training and an additional 3589 images labeled for testing, containing a total of 1503 identities. These identities were randomly chosen, ensuring that at least 20 images were available for each identity. All images were sized 48 × 48 pixels and grayscale.

The expression DCNNs were trained to label expressions using face images from the Facial Expression Recognition 2013 (FER2013) dataset (Goodfellow et al., 2013). The dataset is split to contain 28,709 images specified for training and 3589 images labeled as "public test" for validation. All images were sized 48 × 48 pixels and grayscale.

Once trained, the DCNNs were tested on their ability to perform identity and expression recognition using the KDEF dataset (Lundqvist et al., 1998). This was done by freezing the weights of DCNNs and extracting the activations of units in the last convolutional layer of each network for each of the images. Activations for the different images were then used as the inputs to a simple readout layer. To test for identity labeling, the readout layer was trained on all KDEF images except from one expression category (85.7% train, 14.3% test). The left-out expression category was then used to test the ability of the network to label identity. Cross-validation was performed so that each expression category could be left out for training (seven testing sets) and performances were averaged. To test for expression labeling, images from seven identities were held out of the training set for the readout layer (90% train, 10% test). Cross-validation was performed so that each set of 7 identities could be left out for training (10 testing sets), and performances were averaged.

The DenseNet trained to recognize identity achieved an accuracy of 26.5% on a left-out subset of CelebA, and the DenseNet trained to recognize expression achieved an accuracy of 63.5% on a left-out subset of

FER2013 (Schwartz et al., 2023) Using the DenseNet, each network was able to transfer to KDEF for the task it was trained to perform (identity DenseNet on identity recognition: accuracy = 95.2%, chance level = 1.42%; expression DenseNet on expression recognition: accuracy = 81.9%, chance level = 14.2%). The identity DenseNet was able to label facial expression on the KDEF dataset with an accuracy of 77.7%. The expression DenseNet was able to label face identity on the KDEF dataset with an accuracy of 89.7%.

To facilitate comparison with previous studies, we additionally trained an identity and an expression DCNN based on the ResNet architecture (ResNet-18; He et al., 2016). The ResNet-18 networks were trained using the same datasets that were used for the DenseNets. The ResNet-18 trained to recognize identity achieved an accuracy of 28.0% on a left-out subset of CelebA, and 91.5% on KDEF (chance level = 1.42%). The ResNet-18 trained to recognize expressions achieved an accuracy of 61.3% on a left-out subset of FER2013, and 66.4% on KDEF (chance level = 14.2%). When transferring to the different tasks, the identity ResNet-18-labeled facial expression and the expression ResNet-18-labeled identity with accuracies of 55.7% and 80.1% on KDEF, respectively. Therefore, both DCNNs performed better than chance on left-out images from the datasets that they were trained on, as well as on images from the KDEF dataset. However, they did not transfer to KDEF as well as the DenseNets. A ResNet-18 pretrained on ImageNet to perform object recognition (henceforth referred to as the object ResNet-18) was implemented as an additional model comparison. Details on the training can be found in the study by He et al. (2016). The object ResNet-18 was trained using images in RGB mode. Since the identity and expression DCNNs were trained using grayscale images, we modified the weights of the conv1 layer here by summing over the dimension of the input channels. The object ResNet-18 was able to label identity and expression on KDEF images with accuracies of 96.2% (chance level = 1.42%) and 61.4% (chance level = 14.2%), respectively. A randomly initialized DenseNet and Resnet-18 (same architectures as trained DCNNs) were also used as additional control analyses.

Training and testing datasets comparisons. Since we could not access a sufficiently large dataset including both identity and expression labels, the identity DCNNs and expression DCNNs were trained using two different datasets. It is possible that the testing dataset (KDEF) might be more similar to one of the two training datasets (either CelebA or FER2013). If this is the case, the networks for which the training and testing datasets are more similar might perform better. To test this possibility, both training datasets were compared with the testing dataset by evaluating the similarity between image representations using features from the object ResNet-18 (see "Deep neural network models"). The object-trained ResNet-18 was used to extract feature representations from different layers for images in the identity and expression training datasets, and for images in the testing dataset. For each layer, Pearson r correlation coefficients were computed between the features of image pairs where one image is taken from the testing dataset (KDEF) and one from either the identity or expression training dataset (this analysis was performed separately for each training dataset). Correlations were computed for one channel at a time and averaged across channels. This was done for 100 different randomly chosen image pairs, and the correlations were averaged across the pairs. For each layer, this procedure yielded a measure of the similarity between the training and testing datasets based on the features in that particular layer. To estimate the robustness of the results, the analysis was conducted 10 times, each time selecting a different randomly chosen set of image pairs.

When comparing images from CelebA and KDEF, this analysis yielded mean values of 0.1254, 0.3606, 0.1894, 0.2430, and 0.1244 for conv1, and hidden layers 1−4 respectively. When comparing images from FER2013 and KDEF, this analysis yielded mean values of 0.1708, 0.4263, 0.2141, 0.2739, and 0.0848 for conv1, and hidden layers 1–4, respectively.

The similarity between the training datasets and the testing dataset is comparable. In addition, neither of the training datasets is more similar to the testing dataset for all layers of the object ResNet-18. If anything, the FER2013 dataset shows greater similarity to KDEF for most layers. Therefore, if the CelebA-trained networks were to better account for neural responses, it would be unlikely that this is because of CelebA being more similar to the testing dataset (KDEF).

Representational similarity analysis: comparison between DCNNs. Before comparing the representations in DCNNs to neural responses, we sought to quantify the differences in the representations learned by the identity DCNNs and by the expression DCNNs. Transfer-learning tests conducted in a previous study demonstrate that these DCNNs learn representations that can be used to perform the other task with above-chance accuracy (Schwartz et al., 2023). For example, representations in layers of the expression of DenseNet could be used to read out the identity of faces (Schwartz et al., 2023). However, this does not imply that the identity and expression DenseNets have the same representations.

To test the similarity of the representations in the two DCNNs, we used representational similarity analysis (RSA). We analyzed the representations in multiple hidden layers of the neural networks. Specifically, features from either four or five hidden layers were extracted: the first convolutional layer, and the last layer in each of the three dense blocks (after shrinkage) or the last layer in each of the four residual blocks. For each of these layers, we calculated RDMs using a three-step procedure. First, we extracted feature vectors for all KDEF images used in the experiment. Next, we mean centered the feature vectors by calculating and subtracting the mean feature vector across all KDEF images. Finally, for all pairs of images, we calculated the correlation distance between their mean-centered feature vectors (correlation distance is $1 - r$, where r is Pearson's correlation). In experiment B, information about viewpoint only included the viewpoint angle, without distinguishing between left and right viewpoints, therefore the feature vectors for the left and right viewpoints were averaged (i.e., left and right profile views averaged, left and right half views averaged).

This procedure produced RDMs of size 200×200 for experiment A, and RDMs of size 600×600 for experiment B (Fig. 1*B*,*C*). Note that, as described in the Experimental paradigm subsection, the sizes of the RDMs are different in the two experiments because different subsets of the KDEF images were used in experiment A and experiment B. In the end, the Kendall τ rank correlation coefficient (τ_B)was used to compute the similarity between the RDMs from different layers in the two different DCNNs. A 4×4 cross-network similarity matrix for the trained DenseNets is shown in Figure 1*D*.

Representational similarity analysis of neural data. To retain as much data as possible, we initially performed an analysis on all of the face-selective electrodes, including those from participants who were shown each stimulus once. In this analysis, we computed separate RDMs for each of three temporal windows (125–175, 175–225, and 225–275 ms). This specific temporal range was chosen based on previous studies on the temporal dynamics of visual face perception (Barbeau et al., 2008). As discussed in more detail later, it remains possible that some face information might be encoded in later time windows as well (Ghuman et al., 2014; Li et al., 2019; Boring et al., 2021; see also the Temporal localizer subsection). For each temporal window per each electrode, we extracted a 50-dimensional vector, such that the value for each dimension reflects the amount of measured response in the corresponding millisecond of the 50 ms window. RDMs were obtained by following the same procedure used for the DCNN RDMs, using correlation distance to determine the dissimilarity between the response patterns for each pair of stimuli. As in the RSA for the DCNNs, the average response over all stimuli was subtracted from each stimulus response to remove any baseline that is stimulus independent.

In addition to this, we performed an RSA restricted to highly reliable responses from electrodes located in the fusiform gyrus ($n = 7$). Highly reliable electrodes and time windows were identified following the procedure described below. We then extracted patterns of response from each reliable electrode and time window, and performed the RSA following the same approach described in the previous paragraph, comparing neural RDMs to RDMs extracted from the DenseNet models.

Temporal localizer. We sought to identify time windows during which face-selective electrodes show the most reliable responses. The data time series was segmented using disjoint, successive time windows of 50 ms. The first window was centered at 0 ms post-stimulus onset,

and the last at 500 ms. Therefore, the windows included time points starting from 25 ms before stimulus onset to 525 ms after onset. Disjoint windows were used to reduce the number of multiple comparisons. To identify which of the time windows contained relatively less noise compared with the amount of base signal, all presentations of the neural response of a stimulus were correlated within a specific time window. An average correlation over all time windows for that stimulus was obtained as well. The average correlations across all time windows were then subtracted from the time window-specific correlations. A paired t test was performed between the correlation of the stimulus responses for a given time window and the average correlation of the mean response averaged over all time windows for that stimulus ($p < 0.05$). This determined which of the time windows contained response patterns whose test–retest reliability was significantly higher than average. To correct for multiple comparisons, all t tests were Bonferroni corrected. One electrode was excluded from the RDM analysis because of not containing any time windows with reliable responses ($p > 0.05$).

Representational similarity analysis: comparison between neural activity and DCNNs. We next aimed to compare the neural representations in specific time windows to the representations in the DCNNs. In particular, we evaluated the extent to which RDMs computed using the identity DCNNs and using the expression DCNNs correlate with RDMs based on the iEEG measurements (Kriegeskorte and Kievit, 2013; Khaligh-Razavi and Kriegeskorte, 2014). To calculate the concordance between the RDMs of the DCNN and the neural RDMs, we performed two types of analysis. In the first type of analysis, we compared the RDMs extracted from neural data to RDMs extracted from individual layers of the identity DCNNs and of the expression DCNNs, calculating the Kendall τ_B. Since negative variance-explained values are uninterpretable, any negative τ_B correlations were set to 0 (Fang et al., 2022). The smallest of the negative values was −0.003. This affected a total of 37 of 360 τ values. This procedure was repeated using RDMs from the object ResNet-18 and untrained DCNNs as well. However, since this analysis compares neural representations to the representations in one DCNN layer at a time, one limitation of this analysis is that it does not capture the overall correspondence between neural data and the representations across all layers of a DCNN jointly.

Comparing neural representations to DCNN representations one layer at a time does not reveal to what extent different layers of the DCNN encode redundant information or unique information. To address this question, we introduced a new type of analysis using semipartial Kendall τ rank correlation (Kim, 2015) to evaluate the overall correspondence between the RDMs extracted from the neural data and each of the identity and expression DCNNs when considering jointly the representations in all layers of the DCNNs.

Semipartial correlations measure the strength of the relationship between two variables (i.e., between the neural RDM and the first hidden block RDM) while controlling for the effects of other variables (i.e., the initial convolutional RDM). Within each DCNN model, the semipartial τ_B was calculated for each layer, controlling for the effect of the previous layers. Then, the semipartial τ_B values were summed to obtain a cumulative τ_B value. This allows one to control for redundancy between the layers, evaluating the overall similarity between the models and the data without inflating the τ_B values.

After calculating the semipartial τ_B values between the face-selective electrodes and the identity and expression DCNNs, we performed model comparison using Bayes factor to potentially establish evidence for the absence of differences between the ability of DCNN to account for neural responses (Keysers et al., 2020). This was done to evaluate the statistical evidence for the possibility that there is no difference between the identity DCNN's representational similarity to the neural representations and the expression DCNN's representational similarity to the neural representations (and more precisely, that they come from a same distribution. The analysis with Bayes factor was performed using the set of all face-selective electrodes to maximize statistical power.

Relative contribution of identity and expression. Next, we set out to test whether different sets of electrodes were more strongly correlated with one DCNN over the other. The dataset included electrodes located in the ventral stream as well as electrodes located in lateral temporal regions. If ventral regions are specialized for identity recognition and lateral regions are specialized for expression recognition, ventral electrodes might have a greater cumulative τ_B with the identity DCNN, while lateral electrodes might have a greater cumulative τ_B with the expression DCNN. Alternatively, electrodes in ventral and lateral regions might be similar in terms of their relative correspondence to the identity DCNN and to the expression DCNN.

To compare the relative similarity of neural RDMs in individual electrodes to the RDMs of the identity and expression DCNNs, each electrode at each time window was plotted as a point in a 2D space, where the coordinate along the x-axis was determined by the cumulative Kendall τ_B between the electrode RDM and the identity DCNN RDM, and the coordinate along the y-axis was determined by the cumulative Kendall τ_B between the RDM of the electrode and the expression of DCNN RDM. If ventral electrodes have comparatively higher Kendall τ_B values with the identity DCNN, and lateral electrodes have comparatively higher Kendall τ_B values with the expression DCNN, the two sets of electrodes should fall on lines with different slopes, where the slopes correspond to the ratio between the cumulative τ_B for the identity model and the cumulative τ_B values for the expression model. In particular, electrodes in the ventral stream that are comparatively better explained by the identity DCNN should fall on a line that is closer to the identity axis, while electrodes in the lateral stream should fall on a line that is closer to the expression axis (despite electrodes varying in how well they are explained overall). This would demonstrate the presence of an interaction between DCNN model type and brain region (in line with the classical view). By contrast, if all the electrodes fall on the same line, it means that the relative performance of the identity and expression DCNN models at explaining neural responses is similar for the two streams (in contrast with the classical view), demonstrating the absence of an interaction between DCNN model type and brain region.

Frequentist tests are designed to test for the presence of significant interactions, but a lack of significant effects does not demonstrate no interaction. This makes it challenging to test for the absence of an interaction. However, Bayesian tests are built in such a way that they can evaluate the strength of evidence for the absence of an effect. Thus, a Bayesian approach is implemented to evaluate the relative support for a model in which all the electrodes fall on the same line compared with a model in which the electrodes can fall on two separate lines, one for each stream.

To statistically test whether ventral and lateral electrodes fall on lines with different slopes, we fit the data with two competing linear regression models: one model with two separate slopes for the ventral and lateral electrodes, and one model with a single slope. We then performed model selection with the Bayesian information criterion (BIC) to determine which linear regression model provides a better account for the data. A lower BIC score signifies the better model. The difference between BIC scores, $\delta = \mathrm{BIC}_{\mathrm{separate}} - \mathrm{BIC}_{\mathrm{combined}}$, determines the size of the effect: a difference >10 denotes strong evidence for the better model (Raftery, 1995).

To further examine the ratio between identity and expression model performance for the DenseNet models, we calculated an index ranging from $-\infty$ to ∞, where negative values indicate that the neural representations correlate more with representations in the expression DCNN, and positive values indicate that they correlate more with the identity DCNN. To accomplish this, for each electrode and time window, we calculated the index log ratio (LR) = $\log(\tau_{\mathrm{id}}/\tau_{\mathrm{exp}})$. This was then plotted as a histogram where LRs between $-\infty$ and 0 represent expression-preferring electrode/time window combinations and LRs between 0 and $+\infty$ represent identity-preferring electrode/time window combinations. When conducting comparisons with the DenseNets, three electrodes had cumulative τ_B values ≤ 0 in one time window; therefore, the log ratio could not be calculated, and they were not included in the log ratio histogram (Fig. 3*B*).

Data availability

The code for training and extracting neural network features for both the identity and expression models can be found here: https://github.com/els615/3DenseNets.

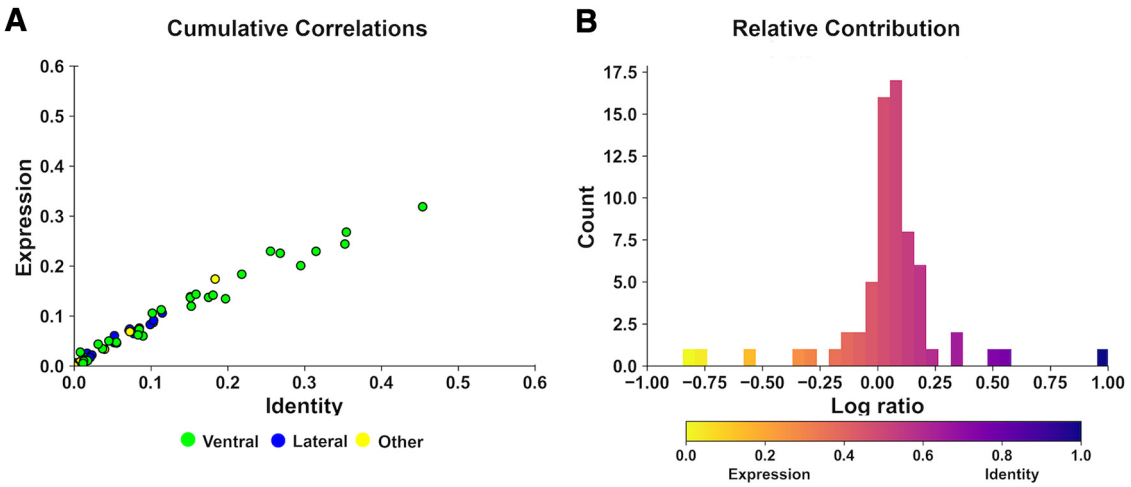

Figure 3. Variation across individual electrodes. ***A***, Scatter plot comparing τ_B values from identity and expression DenseNet models matched on electrode ($n = 24$) and time window. The neural response of each electrode was segmented into 3 time periods, generating 72 data points. ***B***, Histogram showing relative contribution of identity and expression DenseNet models (69 datapoints; 3 electrodes had one time window dropped). Expression-preferring electrodes have a log ratio from $-\infty$ to 0 while identity-preferring electrodes have a log ratio from 0 to $+\infty$.

Results

Representations in deep networks trained for identity and expression recognition

We compared the representations in two DCNNs with the same DenseNet architecture (Fig. 1*A*) where one network was trained to recognize identity and the other was trained to recognize expression. For each network, we calculated RDMs using activations in the first convolutional layer, and in the last layer of each dense block (Fig. 1*B*,*C*). To compare the feature representations across the two DCNNs, the similarity between the RDMs was computed using Kendall τ_B (Fig. 1*D*). Early layers were more similar to one another compared with late layers. The τ_B values between the DCNNs steadily decreased from layer to layer, indicating that the representations in the two DCNNs become increasingly different in later layers. A similar pattern was found when comparing the identity and expression ResNet-18 representations.

Localization of face-selective electrodes

After probing the representations of faces in the DCNNs, we localized face-selective electrodes to analyze the neural representations of the same set of face stimuli. Of the 1079 total electrodes across 11 participants, 25 were found to be face selective (2.3%). Of these, 25 electrodes, 12 were located in the ventral stream (defined as the ventral portion of the temporal cortex and of the occipital cortex anterior to area V2) with 10 of them being located in the fusiform [as determined with Neurosynth (Yarkoni et al., 2011); to confirm that these electrodes were located within Brodmann area 37, we additionally used MarsBar (Brett et al., 2002)]. However, one of these face-selective electrodes did not surpass our reliability analysis and was removed from further analyses. The remaining 24 electrodes are shown in Figure 2*A*. Eight of the face-selective electrodes were located in the lateral stream (defined as the lateral temporal cortex and lateral occipital cortex anterior to area V2, including V3d, V5, and the superior temporal sulcus), and four of the electrodes were located in regions outside the ventral and lateral streams, and thus were labeled as "other."

Comparison between face-selective neural responses and deep networks

Having identified the face-selective electrodes, we next sought to compare representations in these electrodes to representations in the trained DenseNet models. To this end, for each electrode and time window, we computed neural RDMs, and we compared them to the RDMs extracted from the DenseNets using Kendall τ_B. This analysis revealed that representational similarity between the model RDMs and the neural RDMs decreased from the 125–175 ms time window to the 225–275 ms time window (Fig. 2*C*,*D*) for both the identity and expression models (this might be because of a decline in the reliability of the signal; see Discussion). However, within each time window, Kendall τ_B values were comparable for both DenseNets (Fig. 2*C*,*D*).

To probe more rigorously the representational similarity between neural responses and the identity and expression DCNNs overall, we used a novel approach, which consists of calculating a cumulative Kendall τ_B value between neural responses and multiple layers of a DCNN combined (for details, see Materials and Methods). While the cumulative Kendall τ_B value between the identity DenseNet and neural responses was numerically higher than the expression DenseNet (Fig. 2*B*), the difference showed weak evidence for one model over the other (Bayes factor, 0.412–0.441).

To evaluate the robustness of our results, we then repeated our analysis using ResNet-18 for our model. Following the same approach as the DenseNet analysis, RDMs were extracted from the ResNet-18 and compared with each neural RDM. Similar to the DenseNet results, representational similarity between the ResNet-18 RDMs and the neural RDMs decreased from the 125–175 ms time window to the 225–275 ms time window (Fig. 4*A*, *B*), for both the identity and expression models. For almost all time windows, the identity ResNet-18 outperformed the expression ResNet-18 (Fig. 4*A*,*B*). The Bayes factor was performed on the cumulative Kendall τ_B values. This again found weak evidence for one model over the other (Bayes factor, 0.444–0.503).

Previous work (Storrs et al., 2021) found similar amounts of correspondence between trained and untrained neural network models and neural RDMs (unless tuning was used). Consistent with this, untrained DCNNs show similar correspondence with neural responses in this study. While identity and expression DCNNs yielded different representations (Fig. 1*B*–*D*), these differences did not capture corresponding differences between the neural responses in ventral and lateral regions. The untrained DenseNet layer correlations to the neural data had values ranging between 0.0567–0.0623, 0.0355–0.0416, and 0.0105–0.0124 for

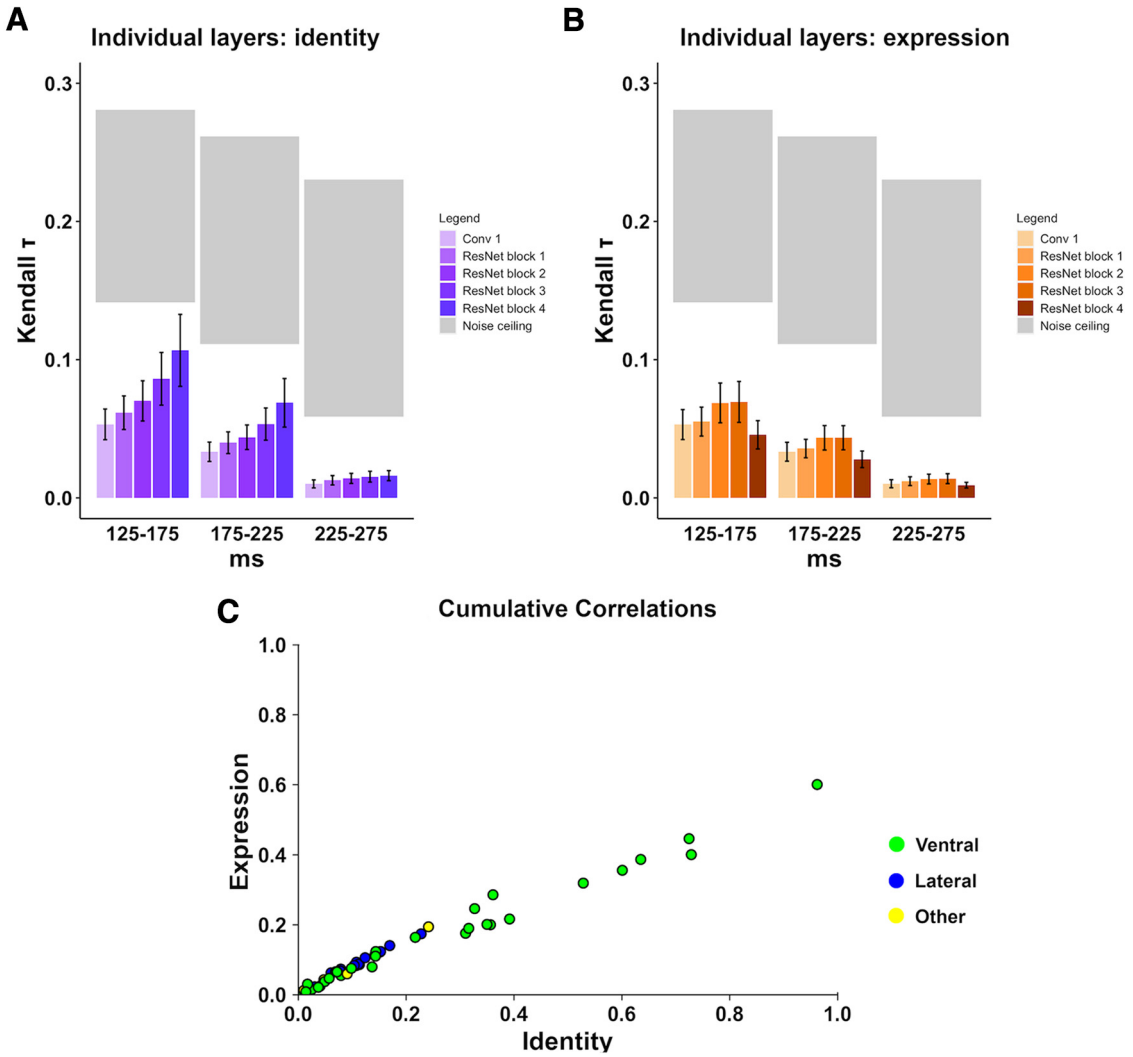

Figure 4. Face-selective electrodes and Kendall τ_B correlations between their representational similarity and the representational similarity in ResNet-18 layers. ***A***, Kendall τ_B values between face-selective iEEG RDMs and layer feature RDMs from the identity ResNet-18 averaged over electrodes ($n = 24$). SEM bars are depicted. ***B***, Kendall τ_B values between face-selective iEEG RDMs and layer feature RDMs from the expression ResNet-18 averaged over electrodes ($n = 24$). SEM bars are depicted. ***C***, Scatter plot comparing τ_B values from identity and expression ResNet-18 models matched on electrodes ($n = 24$) and time window. The neural response of each electrode was segmented into 3 time periods, generating 72 data points.

time windows 125–175, 175–225, and 225-275 ms, respectively. Neural responses showed a lower correspondence but similar pattern with the untrained ResNet-18 (Table 1). The ResNet-18 model that was pretrained on object recognition (Table 1) also performed comparably to the identity DCNNs. Overall, however, the identity ResNet-18 outperformed the pretrained object network.

Examining relative similarity in individual electrodes

The pattern of results observed across all face-selective electrodes might arise from averaging electrodes with distinct properties: ventral temporal electrodes in regions specialized for identity recognition, with greater representational similarity to the identity DCNN, and lateral temporal electrodes in regions specialized for expression recognition, with greater representational similarity to the expression DCNN. Alternatively, the representations measured by electrodes in both ventral and lateral temporal regions might be similar in terms of the extent to which they correlate with activations in the identity and expression DCNNs, respectively.

Table 1. ResNet-18 and Neural responses

	Conv 1	Hidden layer 1	Hidden layer 2	Hidden layer 3	Hidden layer 4
125–175 ms					
Identity ResNet-18	0.0532	0.0616	0.0702	0.0861	0.1067
Expression ResNet-18	0.0530	0.0552	0.0687	0.0694	0.0456
Object ResNet-18	0.0523	0.0695	0.0772	0.0728	0.1014
Untrained ResNet-18	0.0540	0.0525	0.0478	0.0422	0.0372
175–225 ms					
Identity ResNet-18	0.0333	0.0399	0.0438	0.0534	0.0688
Expression ResNet-18	0.0334	0.0357	0.0435	0.0435	0.0279
Object ResNet-18	0.0329	0.0433	0.0468	0.0449	0.0664
Untrained ResNet-18	0.0336	0.0331	0.0306	0.0272	0.0233
225–275 ms					
Identity ResNet-18	0.0101	0.0127	0.0141	0.0153	0.0161
Expression ResNet-18	0.0103	0.0121	0.0136	0.0139	0.0092
Object ResNet-18	0.0100	0.0134	0.0145	0.0140	0.0185
Untrained ResNet-18	0.0101	0.0099	0.0093	0.0083	0.0071

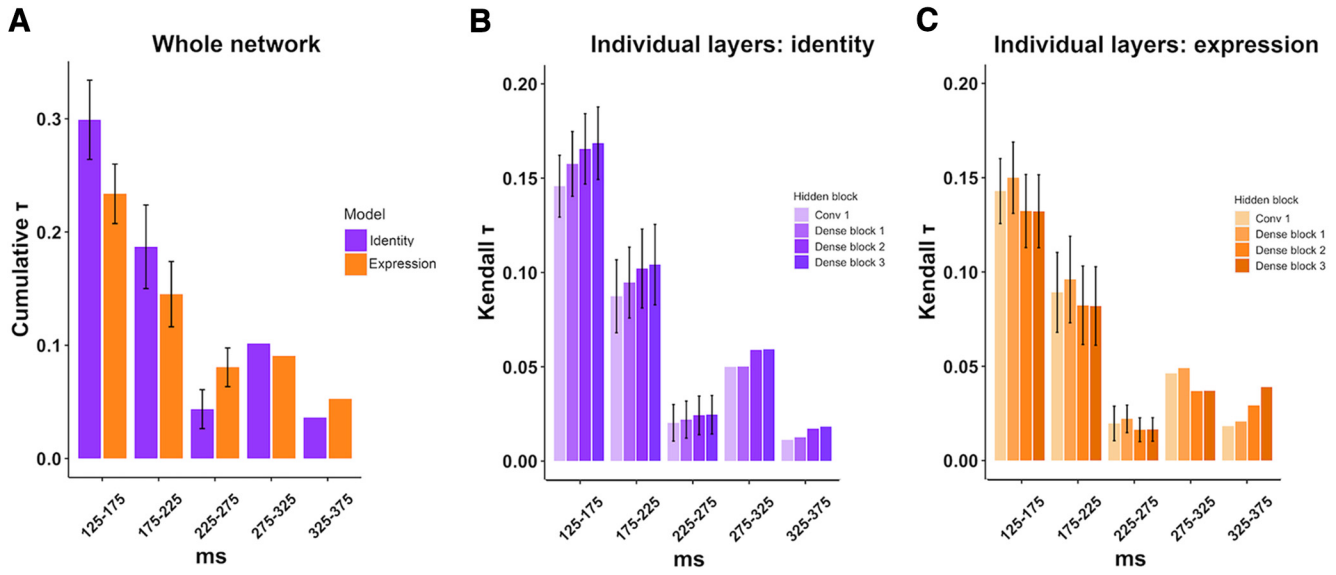

Figure 5. Representational similarity Kendall τ_B correlations between fusiform electrode responses and DenseNet deep network layers. ***A***, Semipartial τ_B values were computed to examine contributions across layers for fusiform electrodes ($n = 7$) in time windows showing high reliability (see Materials and Methods, Temporal localizer subsection). This is plotted as a cumulative value obtained from each model and averaged over electrodes. SEM bars are depicted for time windows with more than one electrode. ***B***, Kendall τ_B values between fusiform iEEG RDMs and layer feature RDMs from the identity DenseNet averaged over electrodes ($n = 7$). SEM bars are depicted for time windows with more than one electrode. ***C***, Kendall τ_B values between fusiform iEEG RDMs and layer feature RDMs from the expression DenseNet averaged over electrodes ($n = 7$). SEM bars are depicted for time windows with more than one electrode.

We investigated this question with two converging analyses. First, for each electrode we evaluated how similar the electrode is to the identity DenseNet RDM and then how similar it is to the expression DenseNet RDM. We used these two values as coordinates for a scatter plot (Fig. 3*A*). If ventral electrodes have comparatively higher cumulative Kendall τ_B values with the identity DenseNet, and lateral electrodes have comparatively higher cumulative Kendall τ_B values with the expression DenseNet, the two sets of electrodes should fall on two different lines with different slopes. Instead, all observed electrodes were located along one line, showing a similar ratio of expression τ_B to identity τ_B (Fig. 3*A*). To quantify this, we used the BIC (lower values indicate a better model) to compare a model with separate slopes for the ventral and lateral electrodes separately ($BIC_{separate} = -198.13$) to a model with a single slope for both the ventral and lateral electrodes ($BIC_{combined} = -514.41$, $BIC_{separate} - BIC_{combined} = 316.28$). Differences >10 in BIC values are interpreted as providing strong evidence in favor of the model with lower BIC (i.e., the combined model; Raftery, 1995). All electrodes (surface and depth) fall on the same line (Fig. 3*A*), suggesting that they have a similar ratio of match to the identity and expression DenseNet models.

The same procedure was repeated using the ResNet-18 model. In accordance with the DenseNet results, the face-selective electrodes were located along one line, showing a similar ratio of expression τ_B to identity τ_B (Fig. 4*C*).The Bayesian information criterion analysis confirmed the following: a model with separate slopes for the ventral and lateral electrodes separately had a smaller BIC ($BIC_{separate} = -191.91$) compared with a model with a single slope for both the ventral and lateral electrodes ($BIC_{combined} = -478.17$, $BIC_{separate} - BIC_{combined} = 286.26$). Again, this suggests that there is strong evidence in favor of a model where ventral and lateral face-selective electrodes are modeled with a single slope.

Next, we computed an index capturing the relative contribution of RDMs from the expression DenseNet and RDMs from the identity DenseNet to account for neural RDMs. The index ranges from $-\infty$ to ∞: negative values indicate a greater contribution of the expression DCNN, while positive values indicate a greater contribution of the identity DCNN (for details, see Materials and Methods). The distribution of index values is shown in Figure 3*B*.

Comparison between fusiform neural responses and deep networks

As a final step, we performed an additional analysis restricted to highly reliable responses in face-selective electrodes located in the fusiform, a region known to play a key role in face perception (Kanwisher et al., 1997). Several electrodes in this region ($n = 7$) had highly reliable responses across multiple time windows. The τ_B values for fusiform-located electrode comparisons were averaged across electrodes for each time window. Figure 5*A* shows results examining the contribution across layers of each DenseNet model. The τ_B values are higher compared with the average of all face-selective electrodes shown in Figure 2*B–D*, but follow a similar pattern. Within most time windows, the identity DenseNet model displayed a numerically larger cumulative semipartial τ_B compared with the expression DenseNet model when examining fusiform electrodes (as was the case in the analysis with all face-selective electrodes as well). This difference between the DenseNet models was greatest in the 125–175 ms range.

Figure 5*B* shows fusiform responses correlated with individual layers of the identity DenseNet, and Figure 5*C* shows fusiform responses correlated with individual layers of the expression DenseNet. Similar to the face-selective pattern mentioned above, both identity and expression DenseNet models were best able to explain neural responses in the 125–175 ms range, followed by 175–225 ms, and then 225–275 ms when averaging data over multiple electrodes. The 275–325 and 325–375 ms windows are included for single electrodes that showed reliable responses during one of the two periods. Similar to the face-selective electrodes again, within each time window, later layers of the identity DenseNet model outperformed earlier layers. This was not the case for the expression DenseNet model.

Discussion

According to a classical view in the field, face identity and facial expression are processed by separate mechanisms (Bruce and Young, 1986), as follows: identity is processed by regions in ventral temporal cortex, while expression is processed by regions in lateral temporal cortex (Haxby et al., 2000). If this is the case, features optimized to recognize facial expression should better capture the similarity between neural responses in lateral regions, and features optimized to recognize face identity should better capture responses in ventral regions. Thus, the classical view would predict that RDMs from identity-trained DCNNs should correlate more with RDMs from the ventral regions, and RDMs from expression-trained DCNNs should correlate more with RDMs from the lateral regions. However, this was not what we found: both identity and expression DCNNs were able to explain neural responses in ventral and lateral regions. The identity DCNNs outperformed the expression DCNNs in both sets of regions (although this difference was not found to be significant).

These results cannot be dismissed as being because of noise. First, if the data were too noisy, we would have encountered poor correlations between the DCNN models and neural responses. However, Kendall τ_B values in this study were comparable to those in other studies (Higgins et al., 2021). It should also be noted that the Kendall τ_B for both the identity and expression DCNN models were close to zero in later time windows, indicating that values found in earlier time windows were not just because of the method used. Second, statistical analysis using BIC revealed that the results provide strong support for the hypothesis that the relative contribution of identity and expression DCNN models is similar for ventral and lateral electrodes (Fig. 3). Finally, when restricting our analysis to electrodes and time windows with very reliable responses, the pattern of results was unchanged (Fig. 5).

Successful transfer learning can be difficult because of potential differences in the data distribution between source and target datasets (Madan et al., 2022). Thus, it is important to determine whether the neural networks trained for their respective tasks can successfully generalize to the KDEF dataset. Both the identity and the expression DCNNs yielded high accuracies on the KDEF dataset. Despite that the expression DCNN labeled expressions with a lower accuracy than the identity DCNN labeled identity, its accuracy was well within the human range [from 72% (Goeleven et al., 2008) to 89.2% (Calvo and Lundqvist, 2008)] for the DenseNet. For this reason, while it is difficult to rule out domain shift problems entirely, it is unlikely that the accuracy difference between the two DCNNs is because of a failure of transfer.

Instead, the difference might be driven by task difficulty. Some facial expressions can be ambiguous even for human observers (Aviezer et al., 2012; Guo, 2012; Tarnowski et al., 2017). While expression recognition performance on KDEF ranges from 72% (Goeleven et al., 2008) to 89.2% (Calvo and Lundqvist, 2008), human observers are very accurate (>90%) at recognizing identity (Bruce, 1982; Burton et al., 1999), even in the presence of changes in viewpoint.

The gender recognition task that the participants performed might have affected their neural responses, and in turn, their correspondence with the DCNN models. Previous work demonstrated that attention selectively enhances face representations (Dobs et al., 2018). It could be argued that the gender detection task is more similar to an identity task. However, gender provides only limited information about identity, and gender information can be decoded from neural responses earlier than identity information (Dobs et al., 2019). Despite this, we cannot entirely rule out that the gender recognition task might have differentially engaged identity recognition mechanisms, potentially enhancing the amount of identity information in face-selective regions.

Nonetheless, our findings are still difficult to reconcile with the classical view. If ventral regions are specialized for identity and lateral regions are specialized for expressions, we would expect that a gender task would enhance ventral responses and leave lateral responses unaffected (or suppressed). Instead, we find that lateral responses show robust correlations with the identity DCNN. A gender recognition task can only enhance identity representations in lateral regions if there can be identity representations in those regions to begin with. Therefore, the correspondence between the identity DCNN and lateral regions challenges the view that representations of identity and expressions are separate.

If ventral and lateral regions are not specialized respectively for the recognition of identity and expression, do they serve the same functional role? If not, what are their functional differences? Studies using combined transcranial magnetic stimulation (TMS) and fMRI (Pitcher et al., 2014) suggest that the pSTS might receive inputs from both regions responding to motion and regions encoding shape information. In addition, there is evidence for pSTS involvement in audiovisual integration (Nath and Beauchamp, 2012; Anzellotti and Caramazza, 2017; Rennig and Beauchamp, 2022). Considering this evidence, we speculate that lateral temporal regions along the superior temporal sulcus might host the convergence of static visual information, dynamic visual information, and auditory information.

In seeming contrast to the proposal that recognition of face identity and facial expression share common neural mechanisms, some previous studies reported patients with dissociations between these two abilities. For example, Hornak et al. (1996) reported a case of a patient with impaired recognition of expressions but spared recognition of identity. However, the patient had damage in ventral frontal cortex, not in lateral temporal cortex. As proposed in the study by Calder (2011), the processing of identity and expressions might diverge at later stages, but they might still rely on common regions in posterior temporal cortex. In a more recent study (Jansari et al., 2015), one patient (DY) with acquired prosopagnosia showed identity recognition deficits, but relatively intact expression recognition as tested with FEEST (Facial Expressions of Emotion: Stimuli and Tests; Young et al., 2002). However, DY did have difficulty recognizing anger (Jansari et al., 2015), indicating some impairment for expression recognition. In addition, while DY was impaired relative to controls at recognizing the identity of upright faces, his performance was similar to that of controls when distinguishing inverted faces and fractured faces, suggesting that he might rely on featural information (Jansari et al., 2015). Such featural information might have also been sufficient to distinguish between the different emotions in FEEST. This possibility is consistent with the previously reported finding that anger recognition is particularly affected by face inversion (Bombari et al., 2013; Fig. 2): in DY, an impairment for configural face processing might have led to the observed difficulties for recognizing the identity of upright faces and also to his disproportionate difficulty for recognizing anger. The present findings are part of broader research efforts indicating that information about object category and other object properties coexist in common regions within temporal cortex (Hong et al., 2016). A relevant study reported that speaker identity and speech content can be decoded in the superior temporal

cortex (Formisano et al., 2008; Bonte et al., 2014). Together, these studies reveal that some sets of tasks rely on shared brain regions, while others are implemented by distinct neural substrates. Recent work is beginning to investigate what are the optimal ways of structuring and sharing representations across multiple different tasks (Zamir et al., 2018; Schwartz et al., 2023).

It is important to note that this study is affected by some limitations. First, the DCNNs were trained using two different datasets. It would be preferable to use a training dataset that included both identity and expression labels, but we were unable to find one such dataset with a sufficient number of images. To mitigate this concern, the training datasets we used (FER2013 and CelebA) are similar in that they include images with a broad range of variation in viewpoint and pose. The DCNNs trained with the two datasets both achieved high performance on the KDEF dataset. It is worth mentioning that even if we had used a single dataset with labels for both identity and expression, the same dataset could include very different expressions but similar identities (or vice versa). Therefore, ensuring that the transfer accuracy of the DCNN is high is essential to determine whether the training procedure was successful for both identity and the expression tasks.

The Bayes factor analysis only showed weak evidence for the abilities of identity DCNN to explain the neural responses compared with the expression DCNNs when evaluating Kendall τ_B values for all of the face-selective electrodes together (Fig. 2*B*, DenseNet). It is possible that there would be stronger evidence if more data could have been included in the analysis, but, given the number of data points available, the evidence for this difference is only weak. However, even if the difference between the two models were strong, this would not alter the conclusion that the results challenge the classical view: Figure 2*B* includes electrodes from both the ventral and lateral streams, and the BIC scores strongly favored a single-line fit for both streams (Figs. 3*A*, 4*C*).

We found that neither the identity nor the expression DCNN models accounted for a large proportion of the variance in later time windows (Figs. 2, 4*A*,*B*), suggesting that the DCNN models we used do not fully capture the structure of face representations. This conclusion is in line with work showing that feedforward DCNNs do not offer a complete account of representational similarity between different images of objects (Xu and Vaziri-Pashkam, 2021). Models that incorporate recurrence are promising candidates to improve the concordance with neural representations (Kar et al., 2019; Kietzmann et al., 2019). Additional studies are needed to test whether they provide a better characterization of neural responses to faces in later time windows.

Recent findings have suggested that object-trained DCNN models can explain similar or greater variance in neural responses to face stimuli compared with face-trained DCNN models (Grossman et al., 2019; Chang et al., 2021; Ratan Murty et al., 2021). Some research groups have interpreted this to mean that face-selective cells are not entirely domain specific (the "domain-general view"; Vinken et al., 2022). Alternatively, it has also been proposed that face-selective cells may have a generalist-like function (the "generalist view"; Chang et al., 2021), in the sense that these cells might support multiple face perception tasks (e.g., recognition of expressions, age). If this is the case, DCNNs that encode features that can support several different face perception tasks would show more similarity to neural representations of face images. In turn, DCNNs trained to perform object recognition might encode such a variety of features because they are trained with many different object classes that vary widely in shape, color, and texture. This could explain why object recognition models show more similarity of neural responses to face images.

In our study, the ResNet-18 trained to perform face identity recognition and the ResNet-18 trained to perform object recognition performed similarly in terms of their correlation with the neural data. It is possible that this could be because of face-selective regions encoding domain-general features. Alternatively, if a ResNet-18 model was trained to perform multiple face tasks, rather than just a single task, it is possible that this face-specific model would significantly outperform the object-trained ResNet-18. This would suggest that face-selective regions do encode domain-specific features that support multiple different face tasks. Our study is not designed to discriminate between the domain-general view and the generalist view. However, our results are at least consistent with the generalist view, suggesting that face-selective regions contribute to both identity and expression recognition. Future studies will need to be implemented to distinguish between these two alternatives.

Although we did not observe differences between the ventral and lateral streams in terms of their correlations with identity and expression DCNNs, comparing the representations learned by these DCNNs in more detail remains an interesting question for future research. Methods that localize the regions of an image that are important for a given classification (Selvaraju et al., 2017) might offer cues about features that are key for both identity and expression recognition, and features that might be uniquely relevant for one of the two tasks.

Last, iEEG is a correlational method. Therefore, we are unable to demonstrate that representations recorded by lateral electrodes causally contribute to identity recognition, or that representations recorded by ventral electrodes causally contribute to expression recognition. Studies using causal methods (e.g., TMS; Pitcher et al., 2007) will be needed to establish the causal involvement of these representations for face perception. Even considering these limitations, the findings challenge the view for which lateral regions are specialized for expression recognition while ventral regions are specialized for identity and converge with recent evidence to suggest that face identity and facial expressions share common neural substrates.

References

Anzellotti S, Caramazza A (2017) Multimodal representations of person identity individuated with fmri. Cortex 89:85–97.

Anzellotti S, Fairhall SL, Caramazza A (2014) Decoding representations of face identity that are tolerant to rotation. Cereb Cortex 24:1988–1995.

Aviezer H, Trope Y, Todorov A (2012) Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science 338:1225–1229.

Axelrod V, Yovel G (2015) Successful decoding of famous faces in the fusiform face area. PloS one 10:e0117126.

Barbeau EJ, Taylor MJ, Regis J, Marquis P, Chauvel P, Liégeois-Chauvel C (2008) Spatiotemporal dynamics of face recognition. Cereb Cortex 18:997–1009.

Bombari D, Schmid PC, Schmid Mast M, Birri S, Mast FW, Lobmaier JS (2013) Emotion recognition: the role of featural and configural face information. Q J Exp Psychol (Hove) 66:2426–2442.

Bonte M, Hausfeld L, Scharke W, Valente G, Formisano E (2014) Task-dependent decoding of speaker and vowel identity from auditory cortical response patterns. J Neurosci 34:4548–4557.

Boring MJ, Silson EH, Ward MJ, Richardson RM, Fiez JA, Baker CI, Ghuman AS (2021) Multiple adjoining word-and face-selective regions in ventral temporal cortex exhibit distinct dynamics. J Neurosci 41:6314–6327.

Brett M, Anton J-L, Valabregue R, Poline J-B (2002) Region of interest analysis using an SPM toolbox. Neuroimage 16 [Suppl 1]:497 [abstract].

Bruce V (1982) Changing faces: visual and non-visual coding processes in face recognition. Br J Psychol 73:105–116.
Bruce V, Young A (1986) Understanding face recognition. Br J Psychol 77:305–327.
Burton AM, Wilson S, Cowan M, Bruce V (1999) Face recognition in poor-quality video: evidence from security surveillance. Psychol Sci 10:243–248.
Calder AJ (2011) Does facial identity and facial expression recognition involve separate visual routes. In: The Oxford handbook of face perception (Calder AJ, Rhodes G, Johnson MH, Haxby JV, eds), pp 427–448. Oxford, UK: Oxford UP.
Calvo MG, Lundqvist D (2008) Facial expressions of emotion (KDEF): identification under different display-duration conditions. Behav Res Methods 40:109–115.
Chang L, Egger B, Vetter T, Tsao DY (2021) Explaining face representation in the primate brain using different computational models. Curr Biol 31:2785–2795.e4.
Colón YI, Castillo CD, O'Toole AJ (2021) Facial expression is retained in deep networks trained for face identification. J Vis 21(4):4, 1–10.
Connolly HL, Young AW, Lewis GJ (2019) Recognition of facial expression and identity in part reflects a common ability, independent of general intelligence and visual short-term memory. Cogn Emot 33:1119–1128.
Dobs K, Schultz J, Bülthoff I, Gardner JL (2018) Task-dependent enhancement of facial expression and identity representations in human cortex. Neuroimage 172:689–702.
Dobs K, Isik L, Pantazis D, Kanwisher N (2019) How face perception unfolds over time. Nat Commun 10:1258.
Duchaine B, Yovel G (2015) A revised neural framework for face processing. Annu Rev Vis Sci 1:393–416.
Fang M, Poskanzer C, Anzellotti S (2022) Pymvpd: a toolbox for multivariate pattern dependence. Front Neuroinform 16:835772.
Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322:970–973.
Fox CJ, Hanif HM, Iaria G, Duchaine BC, Barton JJ (2011) Perceptual and anatomic patterns of selective deficits in facial identity and expression processing. Neuropsychologia 49:3188–3200.
Ghuman AS, Brunet NM, Li Y, Konecky RO, Pyles JA, Walls SA, Destefino V, Wang W, Richardson RM (2014) Dynamic encoding of face information in the human fusiform gyrus. Nat Commun 5:5672.
Goeleven E, De Raedt R, Leyman L, Verschuere B (2008) The Karolinska directed emotional faces: a validation study. Cogn Emot 22:1094–1118.
Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee D-H (2013) Challenges in representation learning: a report on three machine learning contests. In: Neural information processing: 20th international conference, ICONIP 2013 (Lee M, Hirose A, Hou Z-G, Kil RM, eds), pp 117–124. Berlin: Springer.
Grossman S, Gaziv G, Yeagle EM, Harel M, Mégevand P, Groppe DM, Khuvis S, Herrero JL, Irani M, Mehta AD, Malach R (2019) Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. Nat Commun 10:4934.
Guo K (2012) Holistic gaze strategy to categorize facial expression of varying intensities. PLoS One 7:e42585.
Hasan BAS, Valdes-Sosa M, Gross J, Belin P (2016) "Hearing faces and seeing voices": amodal coding of person identity in the human brain. Sci Rep 6:37494.
Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. Trends Cogn Sci 4:223–233.
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition CVPR 2016, pp 770–778. Piscataway, NJ: IEEE.
Hermes D, Miller KJ, Noordmans HJ, Vansteensel MJ, Ramsey NF (2010) Automated electrocorticographic electrode localization on individually rendered brain surfaces. J Neurosci Methods 185:293–298.
Higgins I, Chang L, Langston V, Hassabis D, Summerfield C, Tsao D, Botvinick M (2021) Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. Nat Commun 12:1–14.
Hong H, Yamins DL, Majaj NJ, DiCarlo JJ (2016) Explicit information for category-orthogonal object properties increases along the ventral stream. Nat Neurosci 19:613–622.
Hornak J, Rolls E, Wade D (1996) Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. Neuropsychologia 34:247–261.
Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition CVPR 2017, pp 4700–4708. Piscataway, NJ: IEEE.
Jansari A, Miller S, Pearce L, Cobb S, Sagiv N, Williams A, Tree J, Hanley R (2015) The man who mistook his neuropsychologist for a popstar: when configural processing fails in acquired prosopagnosia. Front Hum Neurosci 9:390.
Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.
Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat Neurosci 22:974–983.
Keysers C, Gazzola V, Wagenmakers E-J (2020) Using bayes factor hypothesis testing in neuroscience to establish evidence of absence. Nat Neurosci 23:788–799.
Khaligh-Razavi S-M, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain it cortical representation. PLoS Comput Biol 10:e1003915.
Kietzmann TC, Spoerer CJ, Sörensen LK, Cichy RM, Hauk O, Kriegeskorte N (2019) Recurrence is required to capture the representational dynamics of the human visual system. Proc Natl Acad Sci U S A 116:21854–21863.
Kim S (2015) ppcor: an r package for a fast calculation to semi-partial correlation coefficients. Commun Stat Appl Methods 22:665–674.
Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci 17:401–412.
Li Y, Richardson RM, Ghuman AS (2019) Posterior fusiform and midfusiform contribute to distinct stages of facial expression processing. Cereb Cortex 29:3209–3219.
Lundqvist D, Flykt A, Öhman A (1998) The Karolinska directed emotional faces (KDEF) [CD ROM], Vol 91, p 630. Stockholm, Sweden: Department of Clinical Neuroscience, Psychology Section, Karolinska Institutet.
Madan S, You L, Zhang M, Pfister H, Kreiman G (2022) What makes domain generalization hard? arXiv:2206.07802. https://doi.org/10.48550/arXiv.2206.07802.
Nath AR, Beauchamp MS (2012) A neural basis for interindividual differences in the mcgurk effect, a multisensory speech illusion. Neuroimage 59:781–787.
Nestor A, Plaut DC, Behrmann M (2011) Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. Proc Natl Acad Sci U S A 108:9998–10003.
Peelen MV, Atkinson AP, Vuilleumier P (2010) Supramodal representations of perceived emotions in the human brain. J Neurosci 30:10127–10134.
Pitcher D, Ungerleider LG (2021) Evidence for a third visual pathway specialized for social perception. Trends Cogn Sci 25:100–110.
Pitcher D, Walsh V, Yovel G, Duchaine B (2007) TMS evidence for the involvement of the right occipital face area in early face processing. Curr Biol 17:1568–1573.
Pitcher D, Duchaine B, Walsh V (2014) Combined tms and fmri reveal dissociable cortical pathways for dynamic and static face perception. Curr Biol 24:2066–2070.
Raftery AE (1995) Bayesian model selection in social research. Soc Methodol 25:111–163.
Ratan Murty NA, Bashivan P, Abate A, DiCarlo JJ, Kanwisher N (2021) Computational models of category-selective brain regions enable high-throughput tests of selectivity. Nat Commun 12:5540.
Rennig J, Beauchamp MS (2022) Intelligibility of audiovisual sentences drives multivoxel response patterns in human superior temporal cortex. Neuroimage 247:118796.
Schwartz E, O'Nell K, Saxe R, Anzellotti S (2023) Challenging the classical view: recognition of identity and expression as integrated processes. Brain Sci 13:296.
Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on

computer vision and pattern recognition CVRP 2017, pp 618–626. Piscataway, NJ: IEEE.

Skerry AE, Saxe R (2014) A common neural code for perceived and inferred emotion. J Neurosci 34:15997–16008.

Storrs KR, Kietzmann TC, Walther A, Mehrer J, Kriegeskorte N (2021) Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. J Cogn Neurosci 33:2044–2064

Tarnowski P, Kolodziej M, Majkowski A, Rak RJ (2017) Emotion recognition using facial expressions. Procedia Comput Sci 108:1175–1184.

Thomas C, Avidan G, Humphreys K, Jung K-j, Gao F, Behrmann M (2009) Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. Nat Neurosci 12:29–31.

Vinken K, Konkle T, Livingstone M (2022) The neural code for "face cells" is not face specific. bioRxiv 483186. https://doi.org/10.1101/2022.03.06.483186.

Xu Y, Vaziri-Pashkam M (2021) Limits to visual representational correspondence between convolutional neural networks and the human brain. Nat Commun 12:2065.

Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC, Wager TD (2011) Large-scale automated synthesis of human functional neuroimaging data. Nat Methods 8:665–670.

Young AW, Perrett D, Calder A, Sprengelmeyer R, Ekman P (2002) Facial expressions of emotion: stimuli and tests (FEEST). Bury St Edmunds, UK: Thames Valley Test Company.

Zamir AR, Sax A, Shen W, Guibas LJ, Malik J, Savarese S (2018) Taskonomy: disentangling task transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition CVPR 2018, pp 3712–3722. Piscataway, NJ: IEEE.

Zhou L, Yang A, Meng M, Zhou K (2022) Emerged human-like facial expression representation in a deep convolutional neural network. Sci Adv 8: eabj4383.