The neural code for ‘face cells’ is not face specific

Kasper Vinken[1]✉, Talia Konkle[2], and Margaret Livingstone[1]

✉ kasper_vinken@hms.harvard.edu

[1] Department of Neurobiology, Harvard Medical School, Boston, MA 02115
[2] Department of Psychology, Harvard University, Cambridge, MA 02478

1 Abstract

‘Face cells’ are visual neurons that selectively respond more to faces than other objects. Clustered together in inferotemporal cortex, they are thought to form a network of modules specialized in face processing by encoding face-specific features. Here we reveal that their category selectivity is instead captured by domain-general attributes. Analyzing neural responses in and around macaque face patches to hundreds of objects, we discovered graded tuning for non-face objects that was more predictive of face preference than was tuning for faces themselves. The relationship between category-level face selectivity and image-level non-face tuning was not predicted by color and simple shape properties, but by domain-general information encoded in deep neural networks trained on object classification. These findings contradict the long-standing assumption that face cells owe their category selectivity to face-specific features, challenging the prevailing idea that visual processing is carried out by discrete modules, each specialized in a semantically distinct domain.

2 Introduction

High-level visual areas in the ventral stream contain neurons that are considered category-selective because they respond more to images of one category than to images of others. The most compelling examples are so-called “face cells”, which selectively respond to faces and form a system of clusters throughout the inferotemporal cortex (IT) (Tsao *et al.*, 2006; Moeller, Freiwald and Tsao, 2008). These clusters are large enough to reliably manifest themselves as face-selective patches in functional imaging studies, where they are surrounded by non-face-selective regions (Kanwisher, Mcdermott and Chun, 1997; Tsao *et al.*, 2003). Whether category selectivity signifies a neural code based on domain-specific, semantically meaningful features, or whether the encoded information merely correlates with specific object categories, is subject to debate (Bracci and Beeck, 2016; Long, Yu and Konkle, 2018; Bao *et al.*, 2020; Murty *et al.*, 2021).

The tuning of face cells has been extensively investigated using face images (Taubert, Wardle and Ungerleider, 2020). These studies showed that face cells are driven by the presence of one or more face-specific features, or face parts, in their canonical configuration, such as the eyes, mouth, hair, and outline of a face (Bruce, Desimone and Gross, 1981; Perrett, Rolls and Caan, 1982; Desimone *et al.*, 1984). Individual face cells are sensitive to the position and arrangement of face parts (Leopold, Bondar and Giese, 2006; Freiwald, Tsao and Livingstone, 2009; Issa and DiCarlo, 2012; Chang and Tsao, 2017), bringing about a tuning for viewpoint and/or identity (Freiwald and Tsao, 2010), and selectivity for upright faces (Taubert, G. Van Belle, *et al.*, 2015; Taubert, Goedele Van Belle, *et al.*, 2015). However, previous reports also found reliable face-cell responses to some non-face objects (Tsao *et al.*, 2006) that did not contain face-specific parts, and even smaller, but still graded, tuning to other non-face images (Meyers *et al.*, 2015).

If face cells care only about the category of faces, then what explains the tuning for other objects in the low firing rate regime? Based on the literature we suggest two hypotheses. A first potential explanation (hypothesis H1; orthogonal multiplexing) is that face cells show a multiplexed tuning of face-specific information (i.e., with no reference to other objects), and independent category-orthogonal information that is uncorrelated with responses to faces. Indeed, non-categorical object properties, such as position, size, and viewpoint, are explicitly represented in IT population responses (Hong *et al.*, 2016). Likewise, shape information was partially dissociable from semantic category information in human imaging studies (Bracci and Beeck, 2016; Zeman *et al.*, 2020). A second hypothesis (H2; domain-general graded tuning) is that face selectivity is not categorical or semantic per se, but it depends on visual characteristics that are strongly represented in faces, yet still present to a graded extent in images of other objects. These attributes are domain-general, that is, not specifically about faces, but more about general image statistics which can be meaningfully applied to any category and may be harder to interpret intuitively than, say, actual face parts.

Here, we differentiate between these two hypotheses by testing whether the degree of category-level face selectivity (i.e., the defining property of face cells), is predictable from non-face responses. If non-face tuning and face selectivity are unrelated (H1; orthogonal multiplexing), we should not be able to predict face selectivity from object responses. However, if category-level selectivity and image-level selectivity are explained by the same stimulus attributes (H2; domain-general graded tuning), then a neuron's face selectivity should be predictable from its responses to non-face objects.

By analyzing neural responses in macaque IT to hundreds of face and non-face stimuli, in combination with computational models, we found that the responsiveness to non-face objects depends in a graded manner on domain-general visual characteristics that also explain the face selectivity of the neural site. Thus, face selectivity does not reflect a semantic or parts-based code, but a preference for visual image characteristics which may not be intuitively interpretable but are present to a larger extent in faces than in non-faces. This conclusion is consistent with the notion that face cells are not categorically different from other neurons, but that they together form a spectrum of tuning profiles in a shared space of image attributes, approximated by representations in later DNN layers.

3 Results

For most of the results, we analyzed recordings at 448 single and multi-unit sites, in central IT (in and around the middle lateral [ML] & middle fundus [MF] face regions) of 6 macaque monkeys, in response to 1407 images: 447 faces and 960 non-face objects with no clear semantic or perceptual association with faces (**Fig. 1a**). The majority of the central IT sites showed, on average, higher responses to faces than to non-face objects (332/448~74%). There was, however, overlap between responses to faces and non-faces: on average 783.2 (SD=227.8, min=55, max=960) non-face images elicited a stronger response than the minimum face response. We quantified the face selectivity of each neural site by calculating a face d' selectivity index, which expresses the expected response difference between a face and a non-face object in standard deviation units (see Methods; values > 0 indicate a higher average response to faces). The higher the d' magnitude, the more consistent the response difference between faces and non-face objects. The average face d' was 0.84 (SD=1.17) and ranged between -1.45 and 4.42. Note that the face d' distribution suggests a continuum rather than a discrete pattern (**Fig. 1a**, right inset). Even highly face-preferring sites still showed reliable graded responses to the images of non-face objects (**Fig. 1a**, lower rows in blue rectangle). We asked how well each neural site and its face selectivity could be characterized solely from the responses to these non-face images. A separate analysis (**Fig. S3**) was done on recordings from anterior IT (anterior lateral face region AL; N=57, 2 monkeys) using a smaller subset of 186 face and 412 non-face images. The majority of these anterior recording sites showed higher responses to faces (52/57~91%) and the average face d' was 1.53 (SD=0.94), ranging between -0.57 and 3.37.

3.1 Responses to non-face objects predict category-level face selectivity

We first addressed whether the tuning for non-face objects is directly related to the face selectivity of a neuron. If it is, the pattern of responses to non-face objects should predict the degree of face selectivity. To test this, we took for each neural site the vector of non-face responses and standardized (z-scored) the values to remove the effects of mean firing rate and scale (SD of firing rate). Next, we fit a linear regression model to predict the measured face d' values, using the standardized responses to non-face objects as predictor variables (see Methods). The results in **Fig. 1b** show that, using all 960 non-face object images, the model explained 75% of the out-of-fold variance in neural face d' (R^2=0.75, p<0.0001, 95% CI [0.71,0.78], Pearson's r=0.88). This means that the response patterns associated with exclusively inanimate, non-face objects can explain most of the variability in face selectivity between neural recording sites. The explained variance increased monotonically as a function of the number of non-face images used to predict face selectivity, starting from ~20% for a modest set of 25 images **Fig. 1c**. Thus, image-level selectivity for non-face objects is determined by characteristics related to the neural site's category-level face selectivity (H2), and not (just) by category-orthogonal information (H1).

Interestingly, the pattern of responses to individual face images was less predictive of face selectivity (R^2=0.49, p<0.0001, 95% CI [0.43,0.55], Pearson's r=0.71), even when the number of non-face images was subsampled to match the number of face images (**Fig. 1c**; face selectivity predicted from 1000 subsamples of 447 non-face images: mean R^2=0.69, min R^2=0.60, max R^2=0.76). A likely explanation is that, as a more homogeneous category, face images provide less information about the coarse tuning profile of a neuron.

The fact that the non-face response pattern predicts face selectivity, implies that some non-face objects elicit stronger responses in neural sites that prefer faces and/or that other images elicit stronger responses in non-face preferring sites. To quantify this, for each non-face image we took the vector of z-scored responses (columns from the blue rectangle in **Fig. 1a**) and correlated it with the vector of face d' values. A positive correlation means that the image tended to elicit a higher response in more face-selective neural sites, and vice versa for a negative correlation. For brevity we use the term 'faceness score' to refer to this correlation from here on. **Fig. 2a** shows all non-face object images sorted by their faceness score, and **Fig. 2b** confirms that the gradient of faceness scores indeed reflects face similarity in the neural representation. The smooth gradient of faceness scores indicates that the information about face d' is shared across many objects, and not limited to actual face features incidentally present in one or a few

objects. Visual inspection of **Fig. 2a** suggests that objects with a positive faceness score tended to be tannish/reddish and round, whereas negative faceness score objects tended to be elongated or spiky. We then calculated for each object the following six properties: elongation, spikiness, circularity, and Lu'v' color coordinates (see Methods). As predicted, the faceness scores correlated negatively with object elongation and spikiness, and positively with object circularity, redness (u'), and yellowness (v'; **Fig. 2c**, bars). These results suggest that, for face-selective sites, non-face object responses will also be negatively correlated with elongation and spikiness, and positively with circularity, u', and v'. To confirm this, we calculated for each neural site these six property-response correlations (using all 960 non-face object images), and correlated each of the resulting six variables with face d'. The results (**Fig. 2c**, red dots) matched the pattern of the correlations between object property and faceness score (**Fig. 2c**, green bars).

These analyses raise the question whether such simple object properties can explain the gradient of face selectivity found in the neural sites, as well as non-face tuning. Indeed, the majority of face cells are tuned to elongation/aspect ratio (Freiwald, Tsao and Livingstone, 2009), and the featural distinction between spiky versus stubby-shaped objects has recently been suggested as one of the two major axes in IT topography, including face patches (Bao *et al.*, 2020). Similarly, properties like roundness, elongation, and star-like shape were shown to account for object representations outside face selective regions in anterior IT (Baldassi *et al.*, 2013). We fit a model (same methods as for **Fig. 1b**) to predict face d' as a linear combination of the property-response correlations used for obtaining the red dots in **Fig. 2c**. **Fig. 2d** shows that this model explained ~5% of the out-of-fold variance in observed face d' (R^2=0.05, p<0.0001, 95% CI [-0.03,0.13], Pearson's r=0.33), falling short of the 75% explained by the non-face object responses themselves. Thus, while the data support a relation between image-level non-face tuning and category-level face selectivity, only a fraction of this link was explained by tuning to color and simple shape properties.

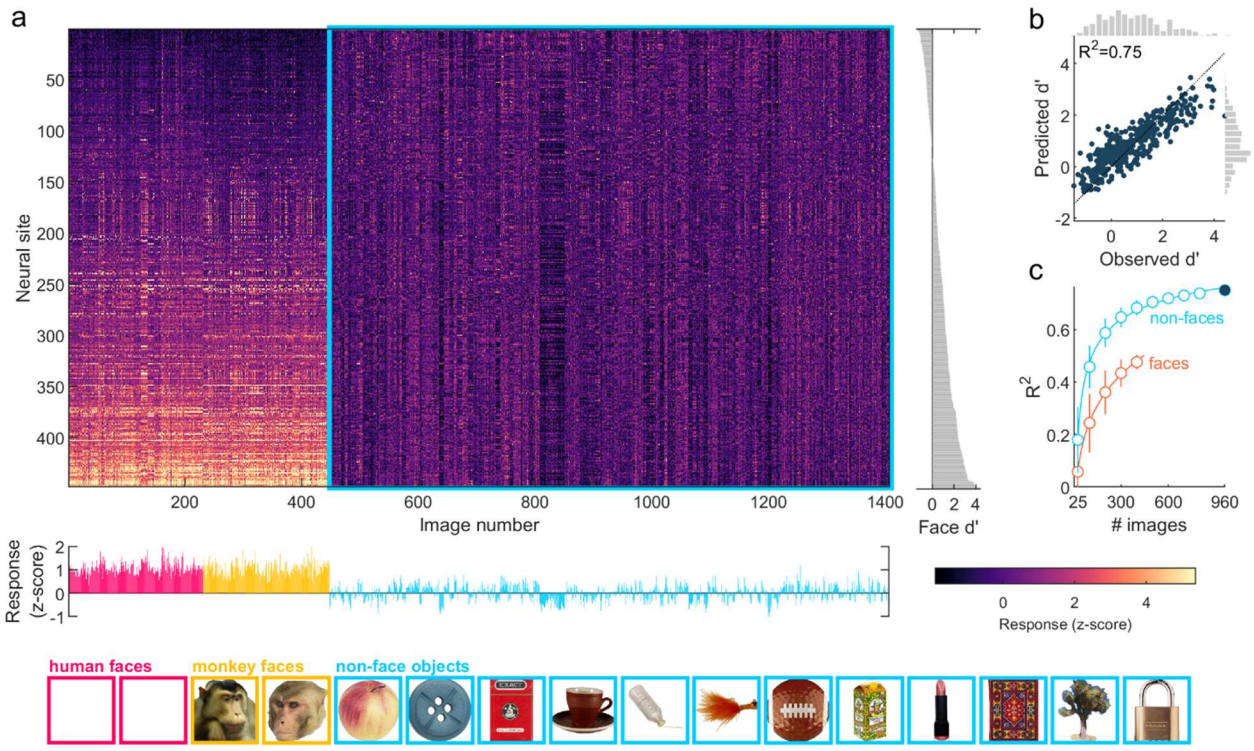

Fig. 1. The response pattern to non-face objects reveals the face selectivity of neurons. (a) Responses of 448 central IT sites (top) and population averages (middle) to 230 human faces (pink), 217 monkey faces (yellow), and 960 inanimate non-face objects (blue, see examples at the bottom). Responses were normalized (z-score) per site using the means and SDs calculated from non-face object images only (blue rectangle). Sites were sorted by face selectivity (face d'). **(b)** Scatter plot of face d' values predicted (using leave-one-session/array-out cross-validation) from the pattern of responses to all 960 non-face objects (blue rectangle in (a)), versus the observed face d'. Each marker depicts a single neural site. The dotted line indicates y=x. **(c)** Out-of-fold explained variance as a function of the number of non-face (blue) or face (orange) images used to predict face d' (means +- SD for randomly subsampling images 1000 times). The filled marker indicates the case using all 960 non-face object images, shown in (b).

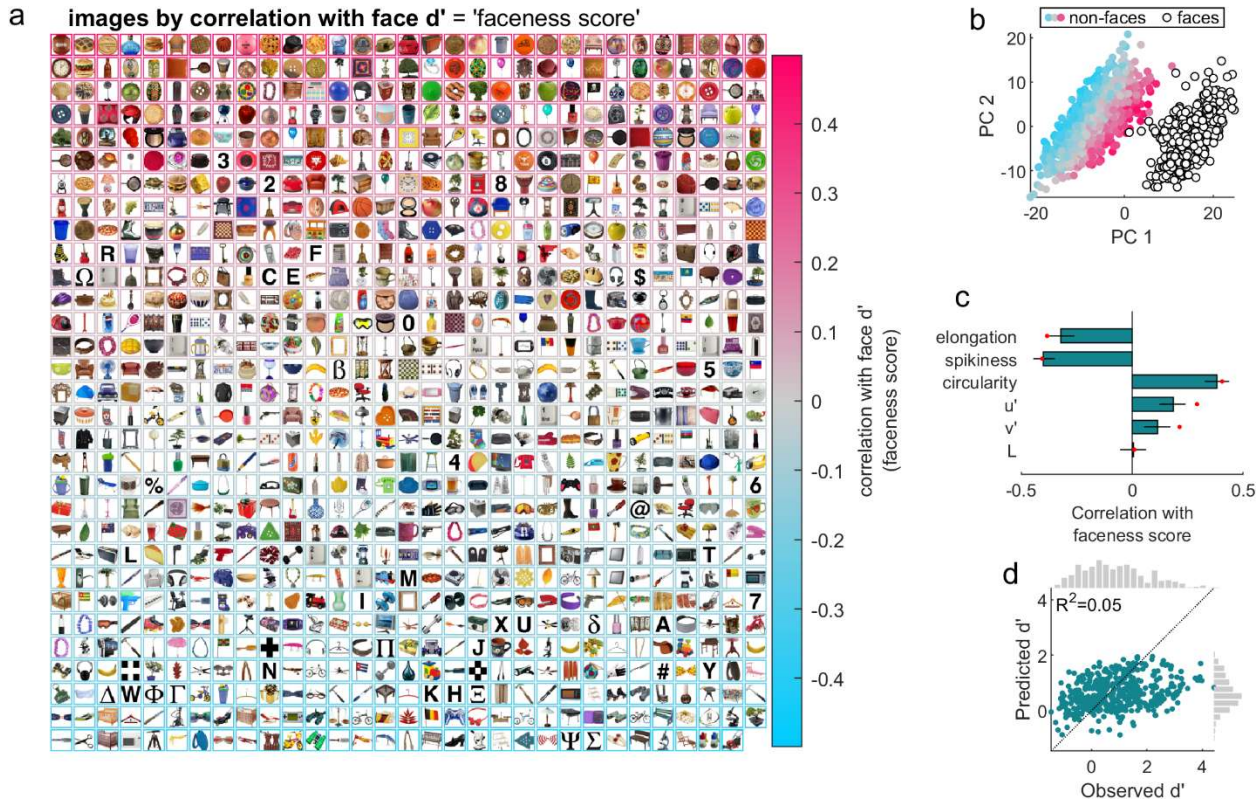

Fig. 2. Color and simple shape properties contribute little to predicting face selectivity. (a) All 960 non-face objects, sorted by the correlation of their response vector with face d'. We refer to this correlation as the faceness score of the object. **(b)** Depiction of all 1407 images in the first two principal components of the neural space (calculated using all 448 sites, z-scored using all images). The markers representing non-face objects are colored according to their faceness score (colorbar in (a)). **(c)** Correlations between faceness score and various object properties (error bars: 95% bootstrap confidence intervals, calculated by resampling images). Red markers: correlations between face d' and property-response correlations (i.e. the values calculated from correlating for each neural site the responses to the 960 non-face object images with the corresponding object properties). **(d)** Scatter plot of observed face d' values and the values predicted (using leave-one-session/array-out cross-validation) from the pattern of correlations between neural responses to non-face objects and each of the seven object properties of (c). Each dot depicts a single recording site. The dotted line indicates y=x.

3.2 Category-level face selectivity and image-level tuning for non-face objects share a common encoding axis

We next asked whether the link between face selectivity and non-face tuning could be better explained by statistical regularities encoded in convolutional deep neural networks (DNNs). We used a DNN trained on multi-way object categorization [ImageNet (Russakovsky *et al.*, 2015)] with the architecture "Inception" (Szegedy *et al.*, 2015) [comparable results were obtained with other architectures, including AlexNet (Krizhevsky, Sutskever and Hinton, 2012), VGG-16 (Simonyan and Zisserman, 2014), and ResNet-50 (He *et al.*, 2015)]. The statistical regularities, or DNN features, encoded by the network are not necessarily intuitively interpretable, like face parts or spikiness, but they have proven to explain a substantial amount of variance in IT responses (Cadieu *et al.*, 2014; Khaligh-Razavi and Kriegeskorte, 2014; Yamins *et al.*, 2014; Kalfas, Kumar and Vogels, 2017; Kalfas, Vinken and Vogels, 2018; Pospisil, Pasupathy and Bair, 2018). The linear mapping between DNN activations and neural responses is referred to as a DNN encoding model, which provides an estimate of the direction ("encoding axis") in the DNN representational space associated with the response gradient. If image-level non-face tuning and category-level face selectivity are explained by common image characteristics, then the encoding model fit only on responses to non-face images should also predict face selectivity, and possibly also image-level face tuning.

We calculated the explained variance in face d' for the non-face encoding models of each DNN layer (**Fig. 3a**) and found that it increased from 4% for the input pixel layer, up to a highest value of 58% (R^2=0.58, p<0.0001, 95% CI [0.52,0.63], Pearson's r=0.77; **Fig. 3a,b**) for *inception-4c* (yellow marker in **Fig. 3a**). Thus, the *inception-4c* encoding model, which we will continue using from here on, explained over ten times (.58/.05) more variance in face d' than the color and shape properties of **Fig. 2**. This implies that in the *inception-4c* representational space, a single non-face encoding axis largely captures both image-level tuning for non-face objects and category-level selectivity between faces and non-faces. The fact that the explained variance in face d' is low for early DNN layers, consistent with the limited predictive power of color and simple shape properties, and increases to its maximum in *inception-4c*, implies that higher-level image characteristics are required to explain the link between face selectivity and non-face tuning.

Unlike non-face encoding models, face encoding models were worse at predicting face selectivity (*inception-4c* face encoding model: R^2=-0.85, 95% CI [-1.03,-0.67], Pearson's r=0.57; non-face encoding model trained on a matched number of non-face stimuli: R^2=0.53, 95% CI

210 [0.48,0.58], Pearson's r=0.74). The negative R^2 value for the face encoding model was the result
211 of a global underestimation of face d' (M=-0.42, p<0.0001, 95% CI [-0.48,-0.36]), that is, the face
212 encoding model overall predicted lower responses in face-selective cells for faces than for non-
213 faces. This is consistent with the finding that the pattern of responses to face images was less
214 predictive of face selectivity than was the pattern of responses to non-faces (see previous
215 section).

216 The fact that the non-face encoding model could predict the degree of face selectivity
217 suggests that the model's encoding axis captures the close relationship between image-level non-
218 face tuning and category-level face selectivity. This leads to the interesting prediction that, for
219 highly face selective neural sites, the non-face encoding axis could be approximated by a face
220 versus non-face classification axis. We first reduced the *inception-4c* space to 200 principal
221 components, which was sufficient for almost maximally explaining face d' (**Fig. 3c**), and then
222 computed each axis (see Methods). On average, the cosine similarity between the non-face
223 encoding axis and classification axis was higher than chance (M=0.28, 95% CI [0.25,0.31], with
224 the 95th percentile chance interval=[-0.14,0.14]; **Fig. 3d**) and positively correlated with face d'
225 (Spearman's ρ=0.66, p<0.0001, 95% CI [0.60,0.71]). For highly face selective neural sites (top
226 10% of face d', N=45), the average cosine similarity was 0.66 (95% CI [0.63,0.69]). This further
227 strengthens the evidence that image-level non-face tuning and category-level face selectivity are
228 explained by common image characteristics encoded by later DNN layers.

229 Consistent with lower face d' prediction accuracy of the face encoding model, the cosine
230 similarity between the face encoding axes and the classification axis was on average within the
231 bounds of chance (M=-0.04, 95% CI [-0.05,-0.03], with chance=[-0.14,0.14]) and only slightly
232 correlated with face d' (Spearman's ρ=0.14, p<0.0001, 95% CI [0.11,0.17]).

233 Up until here we have only shown evidence that face selectivity is supported by a
234 domain-general code inferred from non-face responses. Beyond category selectivity, tuning
235 for individual faces could also be explained by the same attributes, in which case the non-
236 face encoding axis should predict image-level responses for faces. Indeed, both non-face
237 encoding models and face encoding models were able to predict response patterns of the
238 other category (Pearson's r=0.25 95% CI [0.24,0.27] and 0.21 95% CI [0.19,0.22],
239 respectively), but not as well as within-domain predictions (Pearson's r=0.55 95% CI
240 [0.53,0.56] and 0.55 95% CI [0.54,0.57]; **Fig. S1a-c**). Given the high dimensionality of the
241 encoding model, better within-domain predictions likely reflect overfitting on the training

domain. That is, in order to better fit the training data, an imperfect model may end up using non-robust attributes which don't extrapolate well outside the training domain. Indeed, using synthetic data with a single non-face encoding axis for both face and non-face images, we reproduced the observed discrepancy in within-domain versus out-of-domain predictions, in line with the overfitting account (**Fig. S1d-f**). We further reasoned that if overfitting leads to the inclusion of non-robust attributes, the effects of these attributes may balance each other out when all neural sites are considered together. Indeed, all fit non-face encoding models together captured the geometry of the neural population representation well, even for face images (**Fig. 4**). Therefore, the same attributes that explain responses for non-faces, also largely explain the neural representations of individual faces in addition to category-level face selectivity.

Finally, we examined how much of the variance in face d' was explained by all three previous models combined: object response patterns (**Fig. 1b**), color and shape properties (**Fig. 2d**), and the non-face DNN encoding model (**Fig. 3b**). With the predictions from each of these three combined in a metamodel (see Methods), the explained variance reached 80% (R^2=0.80, $p<0.0001$, 95% CI [0.76,0.83], Pearson's r=0.89), which is a substantial amount because for simulated data the expected value, given the number of stimuli and neural noise, was 85% +-1.5% (**Fig. S2**).

Up to this point, our analyses focused on neural sites located in central IT. It is possible that a true semantic-categorical representation only emerges in the more anterior parts of IT. To test this, we examined data from 57 neural sites recorded in and around the anterior face patch AL, using a smaller stimulus set. We found that, like neurons in central IT, face selectivity in anterior face patch AL was linked to tuning for non-face objects (**Fig. S3**).

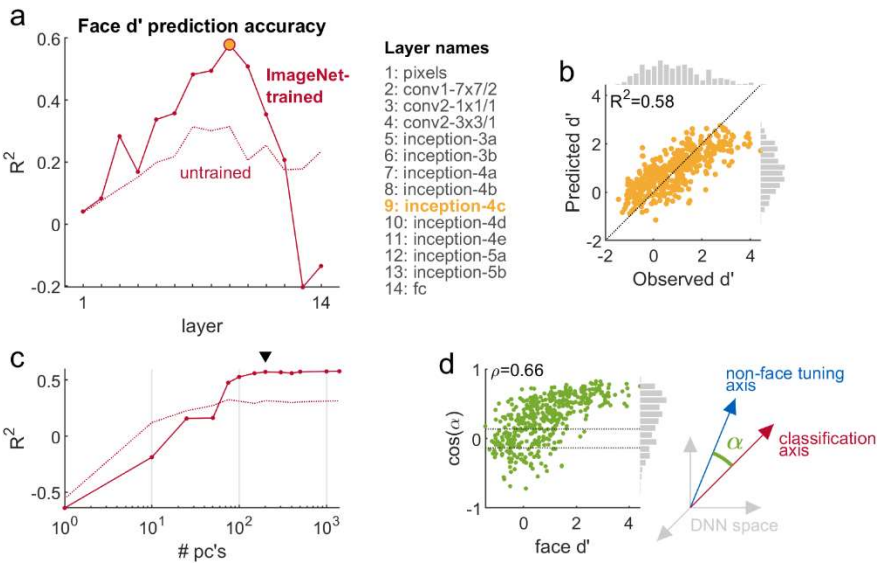

Fig. 3. DNN encoding models fit exclusively on non-face responses predict category-level face selectivity. (a) The amount of variance in neural face d', explained by non-face encoding models [successive layers of Inception (Szegedy *et al.*, 2015), see Methods]. For each neural site, we fit fourteen separate encoding models/axes on responses to non-face objects: one for each DNN layer, starting from the input space in pixels. Dotted line: models based on untrained DNN. **(b)** Scatter plot of the observed face d' values and the values predicted by the trained *inception-4c* layer model (yellow marker in (a)). Each dot depicts a single neural site. The dotted line indicates y=x. **(c)** Explained variance in face d' for *inception-4c* non-face encoding model predictions as a function of the number of *inception-4c* principal components. At 200 principal components, indicated by the black triangle, the encoding model reached >98% of its maximum explained variance. Same conventions as in (a). **(d)** Scatter plot of cosine similarity between each neural site's non-face encoding axis in *inception-4c* space, and a single face versus non-face classification axis (y-axis, see cartoon inset on the right), by face d' (x-axis). Horizontal lines indicate the 95th percentile interval of cosine similarities between random axes. The marginal histogram shows the distribution of cosine similarities.

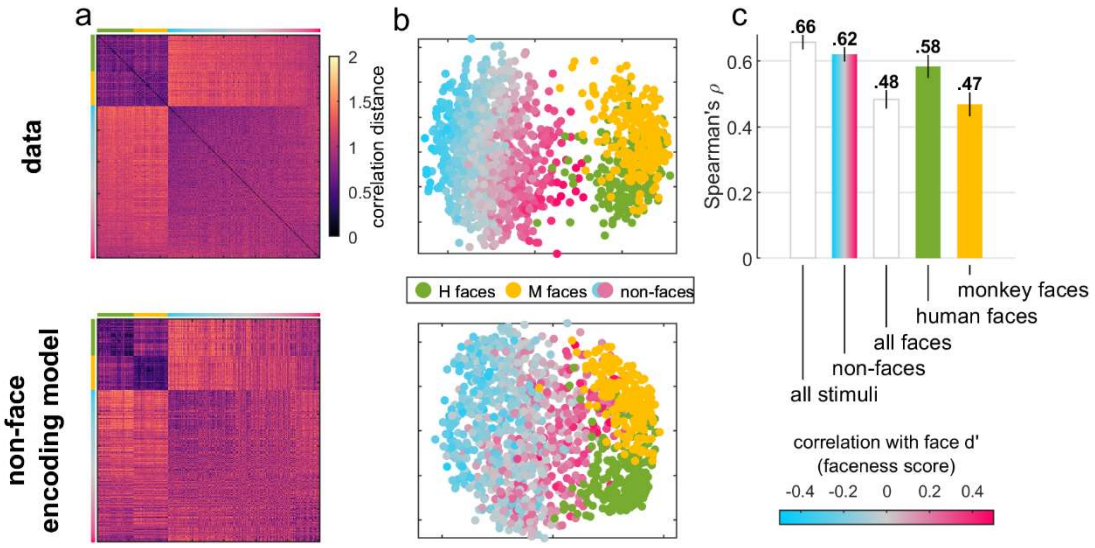

Fig. 4. Representational geometry reproduced by the *inception-4c* non-face encoding model. (a) Dissimilarity matrices based on the correlation distance between response vectors for each pair of the 1407 stimuli (non-faces are sorted by faceness score). Top: using neural responses (z-scored per site). Bottom: using non-face encoding model predictions (z-scored per site). **(b)** Visualization of dissimilarity matrices using classical multidimensional scaling. **(c)** Rank correlation between off-diagonal dissimilarities from the neural data and non-face encoding model, for different subsets of stimuli. Error bars: 95% confidence intervals from jackknife standard error estimates (resampling neural sites).

3.3 The DNN non-face encoding axis of face cells correlates with image characteristics of faces

Having found that the non-face encoding models explain face selectivity, we wondered how face-related the non-face encoding axes estimated by these models are. We coupled a generative neural network [BigGAN (Brock, Donahue and Simonyan, 2019)], to the neural encoding model and synthesize images for each neural site (Murty *et al.*, 2021). This procedure explores a space of naturalistic images for ones that optimally activate the modeled neuron. We hypothesized that if the image characteristics that best explain responses to non-face objects are face-related, the generator would likely synthesize a face. **Fig. 5a** shows one synthesized image for each of the 448 neural sites, generated using the *inception-4c* encoding model of **Fig. 3** (see Methods). Images synthesized for more face-selective neural sites tended to consist of a round region with small dark spots, often in a face-like configuration. More interestingly, these images were likely to be images of dog faces rather than of tan, round, non-face objects. Indeed, the

synthesized images were increasingly more likely to be classified as a dog for neural sites with higher face d' (**Fig. 5b**).

To visualize these trends, we performed two more synthesis experiments. First, we maximized the slope of predicted responses as a function of face d' rank, which produced a tan, round dog face, whereas minimizing the slope produced a spiky abstract object (**Fig. 5c**). Second, we divided all neural sites into 9 face d' bins, and synthesized images that maximized the average predicted response for each bin. Bins of neural sites with a face d' >0.75 consistently showed images of a tan dog face (**Fig. 5d**). Thus, we conclude that the non-face encoding axes revealed increasing preference for faces with increasing neural face selectivity. This finding is consistent with the idea that face selectivity can be explained by tuning to general statistical regularities that are most strongly represented in (animal/dog) face images but also in images of other objects.

We should note here that both the network of our encoding model (Inception) and the generator used for image synthesis (BigGAN) were originally trained on ImageNet, which is heavily biased towards dogs (120 categories), compared to primates (20 non-human primate categories; 0 human categories). Although there are no explicit "human" or "face" categories, 17% of the 1.4 million images do contain at least one human face, divided between other categories such as particular types of clothing or sports (Yang *et al.*, 2019). This representation of human faces in ImageNet is sufficient for BigGAN to be able to synthesize human faces, as evidenced by the images generated in previous work (Murty *et al.*, 2021). However, a critical difference with that study is that we did not include a single face image in the data set to fit the encoding model, so it may not be surprising that the networks fell back to their inherent biases and generated dog faces rather than monkey or human faces.

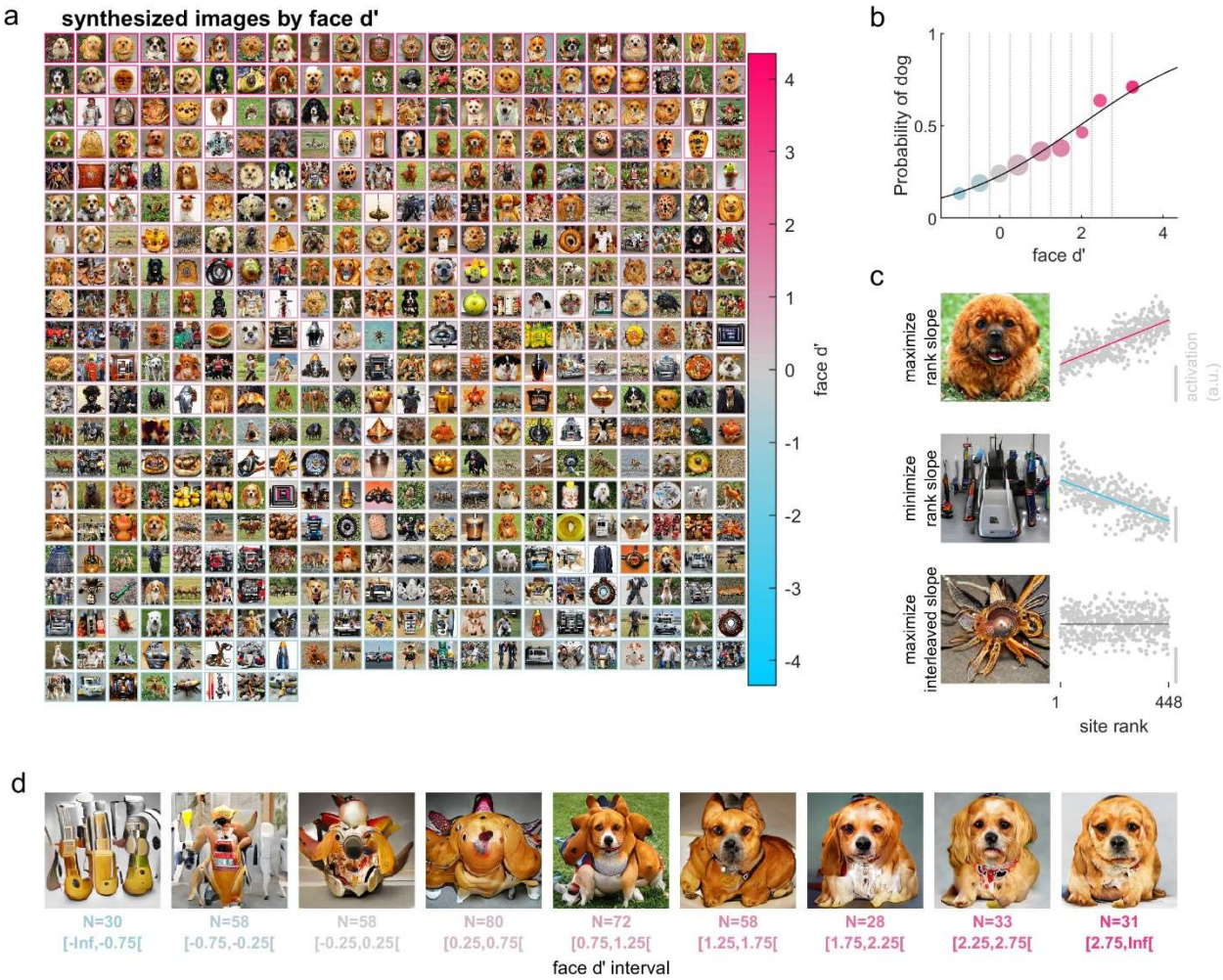

Fig. 5. Image synthesis reveals face-relatedness of the non-face encoding axis. (a) Synthesized images for each neural site using the *inception-4c* encoding model of **Fig. 3** coupled to BigGAN (Brock, Donahue and Simonyan, 2019). Images are sorted by the face d' of the corresponding neural site. **(b)** Proportion of synthesized images classified as a dog breed by resnet-101 (He *et al.*, 2015), as a function of neural face d'. Black line: logistic regression fit. Circles: averages per neural face d' bin [vertical lines indicate bin edges, same as (d)], colored according to average face d' value [see color bar in (a)], size is proportional to the number of sites. **(c)** Top/middle: image synthesized to maximize/minimize the slope of activations by neural face d' rank. Maximizing the slope synthesized images predicted to produce a higher response in face-selective sites, and a lower response in non-face-preferring sites (and vice versa for minimizing the slope). Bottom: to show that generating a dog face is not just a consequence of maximizing any slope, we maximized the slope after reversing the order of even ranks (the x-axis shows the unaltered rank order). **(d)** Images synthesized to maximize the average activation separately for each neural face d' interval [same bins as (b)].

3.4 The DNN non-face encoding axis captures the face inversion effect and selectivity for contextual face cues

Do the non-face encoding models also capture other properties associated with face cells? One such property is that face cells respond more vigorously to faces in their canonical, upright orientation (Tsao *et al.*, 2006; Taubert, Goedele Van Belle, *et al.*, 2015). To test whether our *inception-4c* non-face encoding model would replicate this, we presented 200 images of macaque (100) and human (100) faces to the model in upright and inverted orientations. These images were different from the faces used in the actual neural recordings, and thus not used in the calculation of neural face selectivity. The modeled neurons preferred upright faces (M=0.05, p<0.0001, 95% CI [0.04,0.06]), and this preference correlated positively with the face d' of the actual neural responses used to generate each model (Pearson's r=0.58, p<0.0001, 95% CI [0.52,0.63]; **Fig. 6a,b**).

Interestingly, when the non-face encoding model was based on an untrained DNN (i.e., with random weights), it still showed a preference for upright faces (M=0.026, p<0.0001, 95% CI [0.021,0.031]) and the correlation between the face-inversion effect and face d' was still significantly positive, albeit lower (Pearson's r=0.15, p=0.0011, 95% CI [0.06,0.24]; **Fig. 6b**, compare blue bars). Thus, the face-inversion effect can be inferred from response patterns to non-face images, even without the prior of a DNN trained with ImageNet, which contains faces.

More recently, our lab found that face selective neurons are also more responsive to parts of an image where contextual cues indicate a face ought to be, even if the face itself was occluded (Arcaro, Ponce and Livingstone, 2020). Using the same manipulated images as Arcaro and colleagues (2020), our non-face encoding model predicted a preference for non-face objects positioned where a face ought to be (M=0.13, p<0.0001, 95% CI [0.11,0.14]). This face-context preference was correlated with the face d' calculated from the actual neural responses (Pearson's r=0.48, p<0.0001, 95% CI [0.41,0.55]; **Fig. 6b,c**). **Fig. S4** shows the model-predicted response maps for all 15 manipulated images, averaged for the neural sites with a face d' >1.75.

Interestingly, with an untrained DNN, there no longer was a face-context preference (M=-0.04, p<0.0001, 95% CI [-0.06,-0.03]), and the effect of face context was no longer positively correlated with the neural face d' (Pearson's r=-0.17, p=0.0003, 95% CI [-0.25,-0.09]; **Fig. 5b**, compare red bars). This suggests that, unlike the face-inversion effect, the face-context preference cannot be directly inferred from response patterns to the non-face objects included in our stimulus set, because the association between faces and bodies requires prior experience

with faces in conjunction with bodies. Given that our encoding model was strictly feedforward, these results also confirm that such context effects can be implemented by the feedforward inputs to a neuron, shaped by experience, and that they do not require top-down processes as is often assumed.

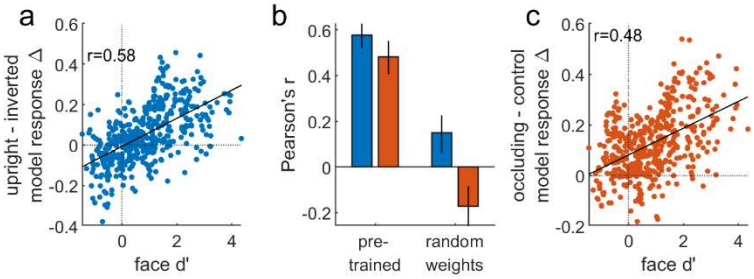

Fig. 6. An encoding model calculated from non-face responses prefers upright faces and objects in positions where a face ought to be. (a) Difference in response to upright versus inverted faces predicted by the *inception-4c* encoding model of **Fig. 3**, as a function of the face d' of the corresponding neural site. Black line shows least squares fit; Pearson's r is shown in the upper left corner. **(b)** Correlation between neural face d' and the model-predicted upright-inverted response difference (blue), or the model predicted occluding-control response difference (red, see (c)). Error bars: 95% bootstrap confidence intervals, calculated by resampling neural sites. Left: models based on an ImageNet-trained DNN (same as (a) and (c)). Right: model based on untrained DNN. **(c)** Same as (a), but for the difference in response to non-face objects occluding a face versus the same objects in control positions (see **Fig. S4** for corresponding heatmaps).

4 Discussion

In this study, we investigated the tuning for non-face objects in neural sites in and around face-selective regions ML/MF and AL of macaque IT. The neural sites revealed a graded spectrum of category-level face selectivity, ranging from not face selective to strongly face selective. We found that the tuning to non-face images was explained by information linearly related to face-selectivity, rather than by face-orthogonal information: the pattern of responses to non-faces could predict the degree of face selectivity across neural sites, while the pattern of responses to faces was significantly worse (**Fig. 1**). The non-face objects for which a higher response was predictive of a higher degree of face selectivity were significantly rounder, redder, more yellow, less spiky, and less elongated. However, these interpretable object properties only explained a fraction of the link between non-face tuning and face selectivity (**Fig. 2**). Instead, image attributes represented in higher layers of an image-classification-trained DNN could best

400 explain this link: the DNN encoding axis estimated from responses to non-face objects could
401 predict the degree of face selectivity (**Fig. 3**) and the representational geometry of all images (**Fig.**
402 **4**). Thus, the visual characteristics that best explained responses to non-face objects also
403 predicted the response difference between faces and non-faces. When coupled with image
404 synthesis using a generative adversarial network, the non-face encoding model showed that the
405 image characteristics that explained non-face tuning were more face-related for neural sites with
406 higher face selectivity (**Fig. 5**).

407 These results imply that non-face tuning of face cells is determined by domain-general
408 image characteristics that also explain category-level face selectivity (H2). Therefore, at its core,
409 face selectivity in the ventral stream should not be considered a semantic code dissociable from
410 visual attributes, as has been previously claimed for category representations (Bracci and Beeck,
411 2016; Zeman *et al.*, 2020). Nor is it a face-specific code that relies on the presence and
412 configuration of face parts (Bruce, Desimone and Gross, 1981; Perrett, Rolls and Caan, 1982;
413 Desimone *et al.*, 1984; Leopold, Bondar and Giese, 2006; Freiwald, Tsao and Livingstone, 2009;
414 Issa and DiCarlo, 2012; Chang and Tsao, 2017). Instead, face selectivity seems to emerge from
415 a combination of domain-general visual characteristics, that are also present in non-face objects,
416 resulting in responses that are graded, regardless of image category.

417 The fact that the face-selective encoding axis could be more accurately inferred from
418 response patterns to non-face objects than from response patterns to faces implies that to fully
419 understand the tuning of face cells, we need to take into consideration responses to non-face
420 objects. This idea is a substantial departure from traditional approaches, which first use non-face
421 images to identify face cells, but then use only faces to further characterize face-cell tuning
422 (Freiwald, Tsao and Livingstone, 2009; Freiwald and Tsao, 2010; Issa and DiCarlo, 2012), or use
423 models that represent only face-to-face variation that does not apply to non-face objects (Chang
424 and Tsao, 2017). Whether the face encoding axis does include attributes that vary only among
425 faces remains an open question. We did find that the non-face encoding axis explained responses
426 to faces less well than the face encoding axis (and vice versa), but this was an expected
427 consequence of model overfitting. Similarly, the semantic-categorical or parts-based views of face
428 selectivity may reflect an “overfitting” to categorical or parts-based stimulus sets. Thus, we believe
429 that an experimental bias towards the “preferred” category leads to a category-specific bias in
430 understanding. The key point is that features that apply to only faces are not a sufficient
431 explanation of face cells, and that understanding tuning in the context of all objects in a domain-
432 general way is required.

An interesting departure from the focus on faces are illusory face images, which share a perceptual experience with faces and engage face-selective regions, while clearly lacking face-specific properties such as skin color or the round shape of a face (Taubert *et al.*, 2017; Taubert, Wardle and Ungerleider, 2020; Wardle *et al.*, 2020). Another study, from our lab, focused on non-face images with contextual associations to faces, showing that face cells are sensitive to these cues (Arcaro, Ponce and Livingstone, 2020). Here, we went one step further by using non-face images, without any semantic or perceptual association with faces, to study the tuning of face cells to domain-general visual attributes.

What could these visual attributes be? In highly face-selective sites, the image characteristics that explained non-face tuning were face-related, as evidenced by our image synthesis analysis (despite not being face-specific, as discussed above). They were not entirely low-level, spatially localized features, because non-face encoding models based on pixels or earlier DNN layers did not explain the link between non-face tuning and face selectivity as well as later DNN layers did. These deeper layers encode image statistics which tend to correlate with the presence of high-level visual concepts, such as object parts, body parts, or animal faces (Zhou *et al.*, 2018), but they are not discrete or categorical in nature (Long, Yu and Konkle, 2018). Recently, Bao and colleagues (2020) dubbed the DNN principal component that best separated face versus non-face stimuli “stubby versus spiky”. While such labels may facilitate communication, they also may come with a false sense of understanding. In our data, spikiness and circularity correlated with the faceness score of an object, yet these properties explained only negligible variance in face selectivity (**Fig. 2**). Thus, unlike face parts or other interpretable object properties, the image attributes that explain the link between non-face tuning and face selectivity are difficult (if not impossible) to name or interpret intuitively.

The link between non-face tuning and face selectivity was captured well regardless of the level of face selectivity. This suggests that there is no categorical difference between face cells and non-face cells, but a graded difference in terms of how well their visual tuning aligns with the high-level concept of faces. This conclusion resonates with the previous finding that the middle face region shows a gradual, monotonic decrease in face selectivity from its center (Aparicio, Issa and DiCarlo, 2016). The alignment of the tuning axis of individual IT neurons with high-level concepts such as faces likely reflects a combination of the proto-organization of the visual system, and the statistical regularities of the experienced environment (Srihasam, Vincent and Livingstone, 2014; Arcaro and Livingstone, 2017; Long, Yu and Konkle, 2018; Arcaro, Ponce and Livingstone, 2020). Thus, face selectivity in IT neurons, and by extension category selectivity in

general, is best understood in terms of general principles and operations, rather than in terms of a discrete patchwork of semantic correlates.

Previous studies have successfully used DNN-based encoding models to characterize the tuning in category-selective visual areas and predict response patterns for a held-out test set. However, in these studies all image categories were represented in both the training and test set of the encoding model (Yamins *et al.*, 2014; Güçlü and van Gerven, 2015; Kalfas, Kumar and Vogels, 2017; Murty *et al.*, 2021), while some used only face images for both training and testing (Chang *et al.*, 2021). For this reason, these studies cannot distinguish whether category selectivity in brain responses was determined by domain-specific features (e.g., face parts) that correlate well with DNN activations, or by domain-general image attributes that generalize across category boundaries.

Finally, we want to emphasize that a domain-general neural code does not necessarily mean that face-selective regions cannot be considered domain-specific. Indeed, in this work we did not address the question of how the information encoded by these neurons, which was face selective, is used in downstream areas. However, face patch perturbation studies did in fact show evidence that non-face perception is affected, albeit less than face perception (Moeller *et al.*, 2017; Sadagopan, Zarco and Freiwald, 2017). From that point of view, it might be more accurate to think of category-selective regions as “domain-biased”, with a generalist function, rather than domain-specific.

5 Supplemental Information

5.1 Image-level generalization

Both non-face encoding models and face encoding models performed less well at predicting out-of-category image-level tuning (albeit still substantially higher than chance level of 0; **Fig. S1a-c**, compare green/blue lines). In principle this could imply that image-level face tuning and non-face tuning are partially explained by independent image characteristics. Note that this would not contradict our domain-general graded tuning hypothesis (H2), which pertains to category-level face selectivity. However, we suggest that a more parsimonious explanation for the reduced image-level generalization is encoding model overfitting (caused by neural noise, limitations of the encoding model, fitting procedure, and images themselves), rather than evidence for distinct kinds of tuning. To validate this reasoning empirically, we created synthetic data with

a known ground truth, by using encoding models fit on non-faces (i.e., a single encoding axis) to simulate responses to both face and non-face images. We used a different network architecture (i.e., AlexNet; Krizhevsky, Sutskever and Hinton, 2012) for generating the synthetic data, to simulate the fact that the network model used to infer an encoding axis is only an imperfect model of the brain. Face and non-face encoding models fit on these synthetic data produced the same pattern of results as the real data (**Fig. S1d-f**). Thus, even when the activation of a unit is driven by a single non-face encoding axis, it can appear that the responses for faces and non-faces are explained by independent image characteristics, based on the images used to estimate the encoding model. These simulations thus also serve to highlight that reduced out-of-category generalization of an encoding model is difficult to interpret, and does not directly imply distinct feature tuning, particularly when extrapolating to very different image sets.

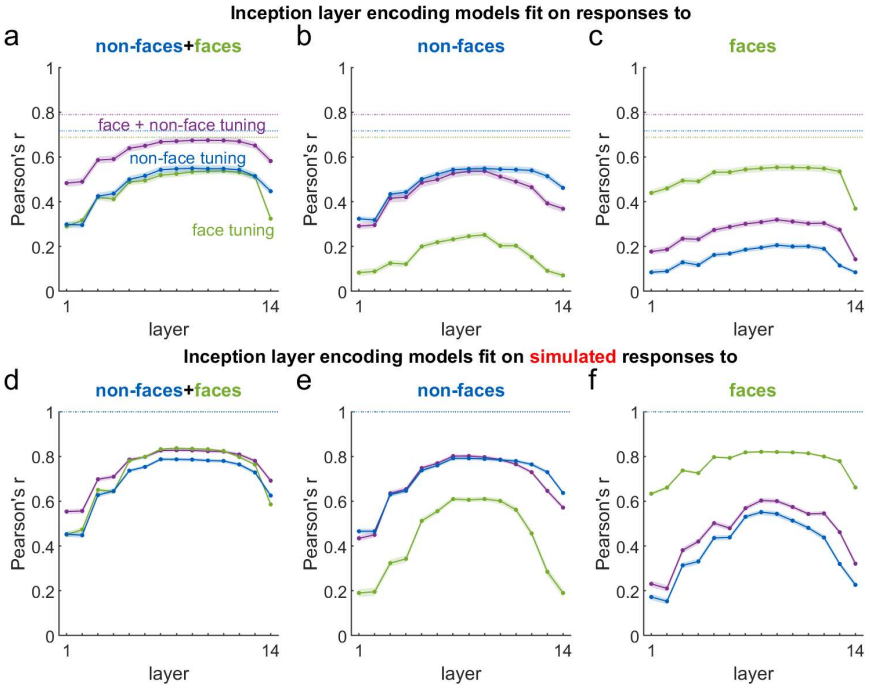

Fig. S1. Image-level generalization of encoding models fit on responses to faces, non-faces, or both. (a) Image-level accuracies of DNN encoding models [successive layers of Inception (Szegedy *et al.*, 2015), see Methods] fit on responses to both non-face and face images. Accuracy is quantified as the Pearson's r between observed and out-of-fold predicted responses, plotted separately for non-faces (blue), faces (green), and both (purple). Error bounds: 95% bootstrap confidence intervals, calculated by resampling neural sites. Horizontal lines: noise ceiling. **(b)** - **(c)** Same as (a), but for DNN encoding models fit to only non-face images (b), or only face images (c). **(d)** - **(f)** Same as in (a) - (c), but for synthetic data simulated from a different encoding model with a single, non-face encoding axis, based on the pool-5 layer of AlexNet

(Krizhevsky, Sutskever and Hinton, 2012). Using a different DNN/layer for simulating data was critical for replicating the drop in performance for out-of-category generalization observed in (a) - (c).

5.2 Metamodel

We compared whether the DNN encoding model (**Fig. 3b**) and tuning for color and shape properties (**Fig. 2d**) capture information on face selectivity above and beyond what can be predicted directly from non-face object responses (**Fig. 1b**). We built a metamodel combining the predictions of all three models (see Methods). If the explained variance of the metamodel exceeds the explained variance of the best individual model (i.e., R^2=0.75), then the individual models capture not entirely overlapping information. This analysis showed that sequentially adding the predictions from the non-face encoding model (**Fig. 3b**) and from the color and shape properties (**Fig. 2d**) to those of a direct fit on non-face object responses (**Fig. 1b**), increased the explained variance in face d' by another 3.5% and 1.1% respectively, leading to a total of 80% for the full metamodel (R^2=0.80, p<0.0001, 95% CI [0.76,0.83], Pearson's r=0.89, **Fig. S2**). Thus, the non-face encoding model and color and shape properties, capture information about face d' that was not captured by the direct fit on non-face object responses.

Overall, these models leave about 20% of the variance in face d' unexplained. However, there is a limit to what we can learn about a neuron's face selectivity even if the model were perfect, simply because neural responses are inherently noisy and because we used a limited number of non-face stimuli. In an extreme case, where none of the stimuli happen to excite the neuron reliably, nothing could possibly be learned. To estimate empirically how these factors limit the explainable variance in face selectivity, we simulated responses directly from the *inception-4c* non-face encoding model, and added Gaussian noise to match the reliability of the original neural responses. With these simulated data, the non-face encoding model used for inference is also the model that generated the responses (for both faces and non-faces). Thus, the explained variance in face d' estimated for these simulated data should give an estimate of what is theoretically inferable about face d', given neural noise and the stimulus set. The average explained variance in face d', calculated from 30 simulations of the metamodel fit on simulated data, was 85% +- 1.5% (**Fig. S2**, dashed line). This suggests that, for the real data, the metamodel in fact explains ~94% (.80/.85) of the explainable variance in face d', given the neural noise and the finite set of non-face stimuli.

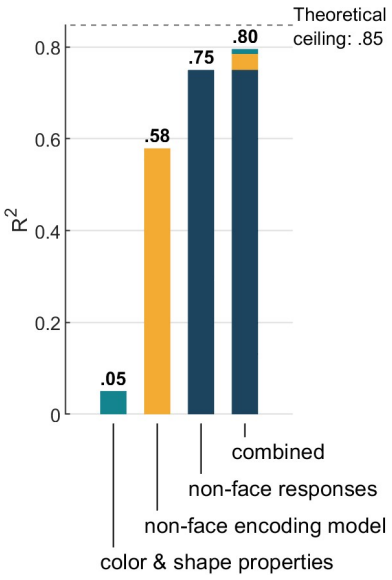

Fig. S2. Comparison of predictions from each individual model and a metamodel combining all three. Comparison of explained variance in face d' with previous models: color and shape properties (**Fig. 2d**), the *inception-4c* non-face encoding model, a direct fit on non-face object responses (**Fig. 1b**), and a metamodel combining predictions of all three. The horizontal line labeled "theoretical ceiling" indicates the expected value for the metamodel fit onto data simulated by the *inception-4c* non-face encoding model (see Methods).

5.3 Non-face tuning is also linked to category-level face selectivity in anterior IT

It is possible that a true semantic-categorical representation only emerges in the more anterior parts of IT. To test this, we examined data from 57 neural sites recorded in and around the anterior face patch AL for comparison with the sites recorded in and around ML from the results above. For the AL recordings, we used a smaller subset of 186 face and 412 non-face images, and so restricted the central IT data to the same stimulus subset for this comparison. With the smaller number of non-face images, the metamodel (see Methods, analogous to **Fig. S3**) explained 63% of the total variance in face d' (AL and central IT sites combined, R^2=0.63, p<0.0001, 95% CI [0.58,0.68], Pearson's r=0.80; **Fig. 6**). For the AL sites, the Pearson correlation between predicted and observed face d' values was 0.62 (p<0.0001, 95% CI [0.37,0.75]). To assess whether this correlation was different from central IT sites, we took a subsample of all central IT sites by matching the central IT site with the closest value in face d' for each AL site. For this d'-matched central IT sample, the correlation between predicted and observed face d'

values was 0.70 ($p<0.0001$, 95% CI [0.55,0.81]), which was comparable to the correlation obtained for AL sites (difference in Pearson's $r=0.08$, $p=0.48$, 95% CI [-0.10,0.34]). Thus, face selectivity in anterior face patch AL is linked to tuning for non-face objects, similar to neurons in central IT.

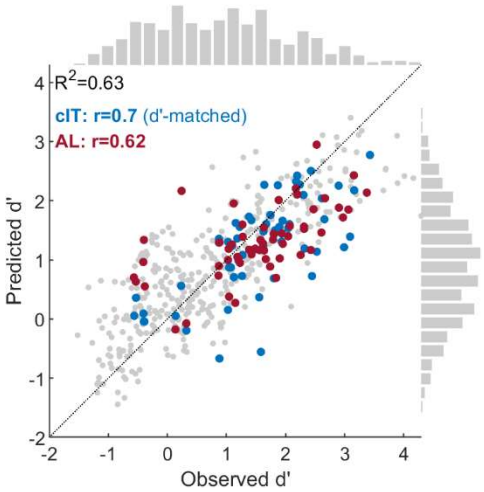

Fig. S3. The relation between face selectivity and tuning for non-face objects extends to the anterior lateral face patch. Scatter plot of face d' values predicted by a meta-model combining predictions from image properties, the *inception-4c* encoding model, and non-face object responses (see **Fig. 3c**), versus the observed face d'. Each marker depicts a single neural site. Neural sites from AL are indicated in red. A subset of central IT sites, each selected to best match an AL site on face d', is indicated in blue. The dotted line indicates y=x.

5.4 Response maps for **Fig. 6**

Fig. S4. Encoding model response maps for manipulated images with non-face objects in positions where a face ought and ought not to be (Arcaro, Ponce and Livingstone, 2020). Average response maps were calculated for neural sites with face d' > 1.75. Regions of pixels used to calculate the occluding-control response difference in **Fig. 6** are delineated with full white contours (object occluding a face) and dashed white contours (object in control position).

6 Methods

6.1 Animals

Eight adult male macaques were used in this experiment. Seven were implanted with chronic microelectrode arrays in the lower bank of the superior temporal sulcus: four monkeys at the location of the middle face region (ML & MF) and two monkeys at the location of the anterior face region (AL). One monkey had a recording cylinder for acute recordings implanted roughly over the middle face region. All procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee and conformed to NIH guidelines provided in the Guide for the Care and Use of Laboratory Animals.

6.2 Behavior

The monkeys were trained to perform a fixation task. They were rewarded with drops of juice to maintain fixation on a fixation spot in the middle of the screen (LCD monitor 53 cm in front of the monkey). Gaze position was monitored using an ISCAN system (Woburn, MA). MonkeyLogic (https://monkeylogic.nimh.nih.gov/) was used as the experimental control software. As long as the monkey maintained fixation, images were presented at a size of 4-6 visual degrees and at a rate of 100 ms on, 100-200 ms off. Images were presented foveally for acute recordings and at the center of the mapped receptive field for chronic recordings.

6.3 Recording arrays

Five monkeys were implanted with 32 channel floating microelectrode arrays (Microprobes for Life Sciences, Gaithersburg, MD) in the middle face region, identified by a functional magnetic resonance imaging (fMRI) localizer (see below). One monkey had an acute recording chamber positioned over the middle face region (identified by fMRI), and neuronal activity was recorded using a 32 channel NeuroNexus Vector array (Ann Arbor, MI) that was inserted each recording day. The two remaining monkeys were implanted with 64 channel NiCr microwire bundle arrays (McMahon *et al.*, 2014; Microprobes for Life Sciences, Gaithersburg, MD) in the anterior lateral face region, identified by fMRI localizer in one monkey and based on anatomical landmarks in the other (Arcaro *et al.*, 2020).

6.4 fMRI-guided array targeting

In all but one monkey, the target location of face patches was identified using fMRI. Monkeys were scanned in a 3-T Tim Trio scanner with an AC88 gradient insert using 4-channel surface coils (custom made by Azma Maryam at the Martinos Imaging Center), using a repetition time (TR) of 2 s, echo time (TE) of 13ms, flip angle (α) of 72°, iPAT=2, 1 mm isotropic voxels, matrix size 96 × 96 mm, 67 contiguous sagittal slices. Before each scanning session, monocrystalline iron oxide nanoparticles (MION; 12 mg/kg; Feraheme, AMAG Parmaceuticals, Cambridge, MA, USA) was injected in the saphenous vein to enhance contrast and measure blood volume directly (Vanduffel *et al.*, 2001; Leite *et al.*, 2002). To localize face-selective regions, 20s blocks of images of either faces or inanimate objects were presented in randomly shuffled order, separated by 20s of neutral gray screen. Additional details are described in Arcaro et al. (2017; 2020).

6.5 Stimuli

The stimuli used in this study are a subset of the images with objects on a white background that were also presented in (Ponce *et al.*, 2019). Most of those images were from (Konkle *et al.*, 2010), but the human and monkey face images were from our lab. The subset we used are 960 images of inanimate objects that do not represent a face (e.g. no jack-o'-lanterns, masks, toys with a head) and 447 close-up images of human and macaque faces, which varied in identity and viewpoint, with or without headgear or personal protective equipment worn by humans in the lab. For the model-simulated experiments we used separate images not used in the neural recordings. The face-inversion experiment was simulated with a set of 100 human and 100 monkey face images which included a background. The face-context experiment was simulated using the images of the original study, where objects in each scene were copied and pasted over the faces (Arcaro, Ponce and Livingstone, 2020).

6.6 Data analysis

6.6.1 Firing rates

We defined the neural response as the spike rate in the 100 ms time window starting at a latency of 50-100 ms after image onset. The exact latency of the response window was determined for each site individually, by calculating the image-level response reliability at each of the 51 latencies between 50 and 100 ms and picking the latency that maximized that reliability. Firing rates were trial averaged per image, resulting in one response vector per neural site. For the acute recordings the images were randomly divided in batches of 255 images, which were presented sequentially to the monkey in separate runs. For these sessions, run differences in median responses were equalized to remove slow trends in responsiveness that were unrelated to the stimuli. Only sites with a response reliability >0.4 were included in the analyses.

6.6.2 Response reliability

The firing-rate reliability was determined per neural site. First, for each image the number of repeated presentations (trials) were randomly split in half. Next, the responses were trial averaged to create two response vectors, one per half of the trials. These two split-half response vectors were then correlated, and the procedure was repeated for 100 random splits to compute an average correlation r. The reliability ρ was computed by applying the Spearman-Brown correction as follows:

$$\rho = \frac{2r}{(1+r)}$$

6.6.3 Face selectivity

Face selectivity was quantified by computing the d' sensitivity index comparing trial-averaged responses to faces and to non-faces:

$$d' = \frac{\mu_F - \mu_{NF}}{\sqrt{(\sigma_F^2 + \sigma_{NF}^2)\frac{1}{2}}}$$

where μ_F and μ_{NF} are the across-stimulus averages of the trial-averaged responses to faces and non-faces, and σ_F and σ_{NF} are the across-stimulus standard deviations. This face d' value quantifies how much higher (positive d') or lower (negative d') the response to a face is expected to be compared to an object, in standard deviation units.

6.6.4 Explained variance

To assess how accurately a model can predict face selectivity (face d'), we calculated the coefficient of determination R^2, which quantifies the proportion of the variation in the observed face d' values that is explained by the predicted face d' values:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where y_i are the observed values, $\bar{y}$ the mean of the observed values, and $\hat{y}_i$ the predicted values. Note that R^2 will be negative when the observed values y_i deviate more from the predicted values $\hat{y}_i$ than from their own mean $\bar{y}$.

6.6.5 Statistical inference

P-values were calculated using permutation tests, based on 10000 iterations. For R^2 and correlations, which calculate the correspondence between two variables, permutation testing was performed by randomly shuffling one of the two variables. For the paired difference between two correlations, the condition labels were randomly shuffled for each pair of observations. 95% Confidence intervals were calculated using the bias corrected accelerated bootstrap (DiCiccio and Efron, 1996), based on 10000 iterations.

6.7 Models

6.7.1 Predicting face selectivity from non-face response patterns

A linear support vector regression model was fit to predict face d' values from response patterns to non-face objects (using the MATLAB 2020a function *fitrlinear*, with the SpaRSA solver and default regularization). The responses of each neural site were first normalized (z-scored) using the mean and standard deviation of responses to non-face objects only. Prediction accuracy was evaluated on out-of-fold predictions using leave-one-session/array-out cross validation: the test partitions were defined as either all sites from the same array (chronic recordings), or all sites from the same session (acute recordings). This ensured that no simultaneously recorded data were ever split over the training and test partitions.

6.7.2 Color and shape properties

For each image, the following properties were computed from the non-background pixels: elongation, spikiness, circularity, and Lu'v' color coordinates. Object elongation was defined based on the minimum Feret diameter F_{min} and the maximum Feret diameter F_{max}, as follows: $1 - \frac{F_{min}}{F_{max}}$. Spikiness was defined based on the object area A_{obj} and the area of the convex hull of the object A_{hull}, as follows: $1 - \frac{A_{obj}}{A_{hull}}$. Circularity was defined using the object area and the object perimeter P_{obj}, as follows: $\frac{4A_{obj}\pi}{P_{obj}^2}$. Lu'v' color coordinates were computed assuming standard RGB.

6.7.3 DNN encoding model

The DNN encoding model was based on a convolutional neural network, used for extracting lower to higher-level image attributes, or DNN features, and a linear mapping between these DNN features and neural responses. The neural network had the architecture named "Inception" (Szegedy *et al.*, 2015), and was trained on the ImageNet dataset (Russakovsky *et al.*, 2015) to classify images into 1000 object categories. We used the pre-trained version of Inception that comes with the MATLAB 2020a Deep Learning Toolbox. Comparable results were obtained using other ImageNet-trained convolutional DNN architectures. Fourteen separate encoding models were created from the Inception network, each based on a subsequent processing step (layer) in the hierarchy: the input layer (pixels), the outputs of the first three convolutional layers, the outputs of each of the nine inception modules, and the output of the final fully connected layer. We refer to each of these encoding models by the name of the processing

step (layer) that they were based on. The outputs of each DNN layer were normalized per channel using the standard deviation and mean across all 1408 images (and across locations for pixels and convolutional layers). Next, the dimensionality of the outputs was reduced by applying principal component analysis using all 1407 images. Finally, a linear support vector regression model was fit to predict neural responses from the principal components of the normalized DNN activations (using the MATLAB 2020a function *fitrlinear*, with the SpaRSA solver and regularization parameter lambda set to 0.01; before fitting the predictors were centered on the mean of the training fold and the responses were centered and standardized using the mean and SD of the training fold). Performance was evaluated on out-of-fold predicted responses. For encoding models fit only on non-faces/faces, we used 10-fold cross validation for the non-face/face images. In this case, the predicted responses for images which were not included in any of the training folds, were computed as the average of the out-of-fold predictions. To compute predicted face d' values for the models, we calculated face d' using out-of-fold predicted responses.

To directly test the alignment between the non-face encoding axis and the face versus non-face classification axis, we first reduced the *inception-4c* representational space to 200 principal components, which was sufficient for almost maximally explaining face d' (**Fig. 3c**). The face versus non-face classification axis was computed by regressing a categorical dummy variable onto these first 200 principal components of the *inception-4c* layer activations. The non-face encoding axis for each neural site was computed by regressing the non-face responses onto the first 200 principal components (i.e., identical to the non-face encoding model but in this case restricting the DNN principal components).

6.7.4 Metamodel

The metamodel was a linear regression model fit to predict the observed face d' values from the predicted face d' values of three different models: (1) face d' predicted from non-face response patterns, (2) from the color and shape property-response correlations, and (3) from the *inception-4c* DNN encoding model. The model was fit using the same linear regression and the exact same cross validation (i.e., same folds) that we also used for predicting face selectivity from non-face response patterns.

6.7.5 Experiment simulations

The *inception-4c* encoding model was used to simulate neurons with a single image tuning axis, estimated from the actual neural responses to non-face images. For these simulations all 960 non-face images were used to fit the encoding model per neural site.

Face-inversion experiment

The outputs of the modeled neural sites were computed for 100 natural images with human faces and 100 with macaque faces, as well as the vertically flipped version of each image.

Occluded faces experiment

In the original study, the authors obtained spatial response maps by presenting large images (16x16 visual degrees) at different positions relative to the monkey's gaze, and recorded neural responses at each of these positions (Arcaro, Ponce and Livingstone, 2020). To simulate this experiment, we used the same images, and took crops of a size equivalent to 5x5 visual degrees using an 80x80 grid of crop-center positions spanning the entire image. The crops near the edge of the image were padded with white pixels, which was the background color used in the original experiment. Each of these crops was fed to the encoding model to compute the output of the modeled neural sites and obtain spatial response maps with a resolution of 80x80. These response maps were then resized to the resolution of the stimulus image using bicubic interpolation. For comparing simulated responses to face and control positions, we used the average values of the response maps within the face and control regions (white contour lines in **Fig. 5d**).

6.7.6 Image synthesis

The *inception-4c* encoding model was used to synthesize images that highly activated the non-face tuning axis of each neural site. All 960 non-face images were used to fit the encoding model, which was then coupled to a generative adversarial network, called BigGan (Brock, Donahue and Simonyan, 2019), following the same procedure as Murty and colleagues (Murty *et al.*, 2021). BigGan is a network trained to generate natural-looking images that each belong to one of the 1000 ImageNet object categories. We used a pre-trained version (available on Tensorflow hub, https://tfhub.dev/deepmind/biggan-256/2), which produces 256x256 pixel images in its top layer, based on an input class vector y and a latent noise vector z. The top layer of the generator was cropped to 224x224 pixels for coupling with the Inception-based encoding model. The class vector y was randomly initialized as 0.05*softmax(n), where n is a vector of

random samples from the truncated standard normal distribution between [0, 1]. The latent noise vector z was randomly sampled from a truncated standard normal distribution between [−2, 2]. To synthesize an image, both *y* and *z* were optimized [using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.001]. The objective functions were either to (a) maximize the activation of a single modeled neural site, or (b) maximize/minimize the slope of the activations of all modeled sites as a function of the face d' rank of the original neural responses, or (c) maximize the average activation of a group of modeled neural sites (binned by neural face d'). To minimize the effect of the random starting points *y* and *z*, the synthesis was first run in a warm-up procedure consisting of 10 runs of 100 steps, with each run using newly sampled values for *y* and *z*. The average of the 10 end-values of *y* and *z* were then used as the starting points for a run of 1000 steps to synthesize an image. This procedure was repeated 10 times, resulting in 10 synthesized images, from which we retained the single image that maximized the objective function.

7 References

Allen, E. J. *et al.* (2022) 'A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence', *Nature Neuroscience*, 25(1), pp. 116–126. doi: 10.1038/s41593-021-00962-x.

Aparicio, P. L., Issa, E. and DiCarlo, J. J. (2016) 'Neurophysiological organization of the middle face patch in macaque inferior temporal cortex.', *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 36(50), pp. 12729–12745. doi: 10.1523/JNEUROSCI.0237-16.2016.

Arcaro, M. J. *et al.* (2020) 'Anatomical correlates of face patches in macaque inferotemporal cortex', *Proceedings of the National Academy of Sciences*, 117(51), pp. 32667–32678. doi: 10.1073/pnas.2018780117.

Arcaro, M. J. and Livingstone, M. S. (2017) 'A hierarchical , retinotopic proto- organization of the primate visual system at birth', pp. 1–24. doi: 10.7554/eLife.26196.

Arcaro, M. J., Ponce, C. and Livingstone, M. (2020) 'The neurons that mistook a hat for a face', (MI), pp. 1–19.

Baldassi, C. *et al.* (2013) 'Shape Similarity, Better than Semantic Membership, Accounts for the Structure of Visual Object Representations in a Population of Monkey Inferotemporal Neurons', *PLoS Computational Biology*, 9(8). doi: 10.1371/journal.pcbi.1003167.

Bao, P. *et al.* (2020) 'A map of object space in primate inferotemporal cortex', *Nature*, 583(7814), pp. 103–108. doi: 10.1038/s41586-020-2350-5.

Bracci, X. S. and Beeck, H. Op De (2016) 'Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways', 36(2), pp. 432–444. doi: 10.1523/JNEUROSCI.2314-15.2016.

Brock, A., Donahue, J. and Simonyan, K. (2019) 'Large scale GaN training for high fidelity natural image synthesis', *7th International Conference on Learning Representations, ICLR 2019*, pp. 1–35.

Bruce, C., Desimone, R. and Gross, C. G. (1981) 'Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque', *Journal of Neurophysiology*, 46(2), pp. 369–384. doi: 10.1152/jn.1981.46.2.369.

Cadieu, C. F. *et al.* (2014) 'Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition', *PLoS Computational Biology*, 10(12). doi: 10.1371/journal.pcbi.1003963.

Chang, L. *et al.* (2021) 'Explaining face representation in the primate brain using different computational models', *Current Biology*, 31(13), pp. 2785-2795.e4. doi: 10.1016/j.cub.2021.04.014.

Chang, L. and Tsao, D. Y. (2017) 'The Code for Facial Identity in the Primate Brain', *Cell*, 169(6), pp. 1013-1028.e14. doi: 10.1016/j.cell.2017.05.011.

Desimone, R. *et al.* (1984) 'Stimulus-selective properties of inferior temporal neurons in the macaque', *Journal of Neuroscience*, 4(8), pp. 2051–2062. doi: 10.1523/jneurosci.04-08-02051.1984.

DiCiccio, T. J. and Efron, B. (1996) 'Bootstrap confidence intervals', *Statistical Science*, 11(3), pp. 151–160. doi: 10.1214/ss/1032280214.

Freiwald, W. A. and Tsao, D. Y. (2010) 'Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System', *Science*, 330(6005), pp. 845–851. doi: 10.1126/science.1194908.

Freiwald, W. A., Tsao, D. Y. and Livingstone, M. S. (2009) 'A face feature space in the macaque temporal lobe', *Nature Neuroscience*, 12(9), pp. 1187–1196. doi: 10.1038/nn.2363.

Güçlü, U. and van Gerven, M. A. J. (2015) 'Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream', *Journal of Neuroscience*, 35(27), pp. 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015.

He, K. *et al.* (2015) 'Deep Residual Learning for Image Recognition', *Multimedia Tools and Applications*, pp. 1–17. doi: 10.1007/s11042-017-4440-4.

Hong, H. *et al.* (2016) 'Explicit information for category-orthogonal object properties increases along the ventral stream', 19(4). doi: 10.1038/nn.4247.

Issa, E. B. and DiCarlo, J. J. (2012) 'Precedence of the Eye Region in Neural Processing of Faces', *Journal of Neuroscience*, 32(47), pp. 16666–16682. doi: 10.1523/JNEUROSCI.2391-12.2012.

Kalfas, I., Kumar, S. and Vogels, R. (2017) 'Shape Selectivity of Middle Superior Temporal Sulcus Body Patch Neurons', *Eneuro*, 4(June), p. ENEURO.0113-17.2017. doi: 10.1523/ENEURO.0113-17.2017.

Kalfas, I., Vinken, K. and Vogels, R. (2018) 'Representations of regular and irregular

shapes by deep Convolutional Neural Networks, monkey inferotemporal neurons and human judgments', *PLOS Computational Biology*, 14(10), p. e1006557. doi: 10.1371/journal.pcbi.1006557.

Kanwisher, N., Mcdermott, J. and Chun, M. M. (1997) 'The Fusiform Face Area : A Module in Human Extrastriate Cortex Specialized for Face Perception', 17(11), pp. 4302–4311.

Khaligh-Razavi, S. M. and Kriegeskorte, N. (2014) 'Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation', *PLoS Computational Biology*, 10(11). doi: 10.1371/journal.pcbi.1003915.

Konkle, T. *et al.* (2010) 'Conceptual distinctiveness supports detailed visual long-term memory for real-world objects.', *Journal of Experimental Psychology: General*, 139(3), pp. 558–578. doi: 10.1037/a0019165.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012) 'ImageNet Classification with Deep Convolutional Neural Networks', *Advances In Neural Information Processing Systems*, pp. 1–9. doi: http://dx.doi.org/10.1016/j.protcy.2014.09.007.

Leite, F. P. *et al.* (2002) 'Repeated fMRI using iron oxide contrast agent in awake, behaving macaques at 3 Tesla.', *NeuroImage*, 16(2), pp. 283–294. doi: 10.1006/nimg.2002.1110.

Leopold, D. A., Bondar, I. V. and Giese, M. A. (2006) 'Norm-based face encoding by single neurons in the monkey inferotemporal cortex', *Nature*, 442(7102), pp. 572–575. doi: 10.1038/nature04951.

Long, B., Yu, C. and Konkle, T. (2018) 'Mid-level visual features underlie the high-level categorical organization of the ventral stream', (7). doi: 10.1073/pnas.1719616115.

McMahon, D. B. T. *et al.* (2014) 'One month in the life of a neuron: longitudinal single-unit electrophysiology in the monkey visual system', *Journal of Neurophysiology*, 112(7), pp. 1748–1762. doi: 10.1152/jn.00052.2014.

Meyers, E. M. *et al.* (2015) 'Intelligent information loss: The coding of facial identity, head pose, and non-face information in the macaque face patch system', *Journal of Neuroscience*, 35(18), pp. 7069–7081. doi: 10.1523/JNEUROSCI.3086-14.2015.

Moeller, S. *et al.* (2017) 'The effect of face patch microstimulation on perception of faces and objects', *Nature Neuroscience*, 20(5), pp. 743–752. doi: 10.1038/nn.4527.

Moeller, S., Freiwald, W. A. and Tsao, D. Y. (2008) 'Patches with Links: A Unified System for Processing Faces in the Macaque Temporal Lobe', *Science*, 320(5881), pp. 1355–1359. doi: 10.1126/science.1157436.

Murty, N. A. R. *et al.* (2021) 'Computational models of category-selective brain regions enable high-throughput tests of selectivity', *Nature Communications*, 12(1). doi: 10.1038/s41467-021-25409-6.

Perrett, D. I., Rolls, E. T. and Caan, W. (1982) 'Visual neurones responsive to faces in the monkey temporal cortex', *Experimental Brain Research*, 47(3), pp. 329–342. doi: 10.1007/BF00239352.

Ponce, C. R. *et al.* (2019) 'Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences', *Cell*, 177(4), pp. 999-1009.e10. doi: 10.1016/j.cell.2019.04.005.

Pospisil, D. A., Pasupathy, A. and Bair, W. (2018) ''Artiphysiology' reveals V4-like shape tuning in a deep network trained for image classification', *eLife*, 7, pp. 1–31. doi: 10.7554/eLife.38242.

Russakovsky, O. *et al.* (2015) 'ImageNet Large Scale Visual Recognition Challenge', *International Journal of Computer Vision*, 115(3), pp. 211–252. doi: 10.1007/s11263-015-0816-y.

Sadagopan, S., Zarco, W. and Freiwald, W. A. (2017) 'A causal relationship between face-patch activity and face-detection behavior', *eLife*, 6, pp. 1–14. doi: 10.7554/eLife.18558.

Simonyan, K. and Zisserman, A. (2014) 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *Information and Software Technology*, 51(4), pp. 769–784. doi: 10.1016/j.infsof.2008.09.005.

Srihasam, K., Vincent, J. L. and Livingstone, M. S. (2014) 'Novel domain formation reveals proto-architecture in inferotemporal cortex.', *Nature neuroscience*, 17(12), pp. 1776–1783. doi: 10.1038/nn.3855.

Szegedy, C. *et al.* (2015) 'Going deeper with convolutions', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June, pp. 1–9. doi: 10.1109/CVPR.2015.7298594.

Taubert, J., Van Belle, G., *et al.* (2015) 'Neural Correlate of the Thatcher Face Illusion in a Monkey Face-Selective Patch', *Journal of Neuroscience*, 35(27), pp. 9872–9878. doi: 10.1523/JNEUROSCI.0446-15.2015.

Taubert, J., Van Belle, Goedele, *et al.* (2015) 'The effect of face inversion for neurons inside and outside fMRI-defined face-selective cortical regions.', *Journal of neurophysiology*, 113(5), pp. 1644–55. doi: 10.1152/jn.00700.2014.

Taubert, J. *et al.* (2017) 'Face Pareidolia in the Rhesus Monkey', *Current Biology*, 27(16), pp. 2505-2509.e2. doi: 10.1016/j.cub.2017.06.075.

Taubert, J., Wardle, S. G. and Ungerleider, L. G. (2020) 'What does a "face cell" want?'', *Progress in Neurobiology*, 195(May), p. 101880. doi: 10.1016/j.pneurobio.2020.101880.

Tsao, D. Y. *et al.* (2003) 'Faces and objects in macaque cerebral cortex.', *Nature neuroscience*, 6(9), pp. 989–995. doi: 10.1038/nn1111.

Tsao, D. Y. *et al.* (2006) 'A cortical region consisting entirely of face-selective cells.', *Science (New York, N.Y.)*, 311(5761), pp. 670–4. doi: 10.1126/science.1119983.

Vanduffel, W. *et al.* (2001) 'Visual motion processing investigated using contrast agent-enhanced fMRI in awake behaving monkeys', *Neuron*, 32(4), pp. 565–577. doi: 10.1016/S0896-6273(01)00502-5.

Wardle, S. G. *et al.* (2020) 'Rapid and dynamic processing of face pareidolia in the human brain', *Nature Communications*, 11(1), pp. 1–14. doi: 10.1038/s41467-020-18325-8.

Yamins, D. L. K. *et al.* (2014) ‘Performance-optimized hierarchical models predict neural responses in higher visual cortex’, *Proceedings of the National Academy of Sciences*, 111(23), pp. 8619–8624. doi: 10.1073/pnas.1403112111.

Yang, K. *et al.* (2019) ‘A Study of Face Obfuscation in ImageNet’, *arXiv*.

Zeman, A. A. *et al.* (2020) ‘Orthogonal Representations of Object Shape and Category in Deep Convolutional Neural Networks and Human Visual Cortex’, *Scientific Reports*, 10(1), pp. 1–12. doi: 10.1038/s41598-020-59175-0.

Zhou, B. *et al.* (2018) ‘Interpreting Deep Visual Representations via Network Dissection’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/TPAMI.2018.2858759.