bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

What computational model provides the best explanation of face representations in the primate brain?

Le Chang[1,2]*, Bernhard Egger[4,5], Thomas Vetter[4], Doris Y. Tsao[1,3,6]*

[1]Division of Biology and Biological Engineering, Computation and Neural Systems, Caltech, Pasadena, CA, 91125, USA.

[2]Institute of Neuroscience, Key Laboratory of Primate Neurobiology, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, 200031, China.

[3]Howard Hughes Medical Institute, Pasadena, CA, 91125, USA.

[4]Department of Mathematics and Computer Science, University of Basel, Basel, Switzerland

[5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA

[6]Lead Contact

*Correspondence: lechang@ion.ac.cn (L.C.), dortsao@caltech.edu (D.Y.T).

Summary

Understanding how the brain represents the identity of complex objects is a central challenge of visual neuroscience. The principles governing object processing have been extensively studied in the macaque face patch system, a sub-network of inferotemporal (IT) cortex specialized for face processing (Tsao et al., 2006). A previous study reported that single face patch neurons encode axes of a generative model called the “active appearance” model (Chang and Tsao, 2017), which transforms 50-d feature vectors separately representing facial shape and facial texture into facial images (Cootes et al., 2001; Edwards et al., 1998). However, it remains unclear whether this model constitutes the best model for explaining face cell responses. Here, we recorded responses of cells in the most anterior face patch AM to a large set of real face images, and compared a large number of models for explaining neural responses. We found that the active appearance model better explained responses than any other model except CORnet-Z, a feedforward deep neural network trained on general object classification to classify non-face images, whose performance it tied on some face image sets and exceeded on others. Surprisingly, deep neural networks trained specifically on facial identification did not explain neural responses well. A major reason is that units in the network, unlike neurons, are less modulated by face-related factors unrelated to facial identification such as illumination.

Keywords:
inferior temporal cortex; primate vision; face processing; neural coding; electrophysiology

Introduction

Primates are able to recognize objects invariant to changes in orientation and

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

position. Neurons in macaque face patch AM represent facial identity independent of head orientation (Freiwald and Tsao, 2010), therefore providing a unique opportunity to study how invariant object identity is represented in the brain. One intuitive computational strategy for invariant face recognition is to separate information about facial shape from that about facial texture. Changes in head orientation or expression can produce changes in facial shape but leave unaltered the underlying texture map of the face (arising largely from physical features such as skin pigmentation, the shape and thickness of eyebrows, eyes, lips, and so on). An effective computational approach to decouple shape and texture information contained in a face is the "active appearance model," a scheme for representing faces by projecting them onto two sets of axes, one describing the shape and one describing the shape-free appearance of a face (Cootes et al., 2001; Edwards et al., 1998). While some shape-related features can vary depending on facial identity (e.g., inter-eye distance), and some appearance-related features can vary for the same facial identity (e.g., illumination), the decoupling between shape and texture parameters accomplished by the active appearance model nevertheless approximately aligns with the needs of invariant face identification.

A recent study used facial images synthesized by an active appearance model to explore the coding scheme of AM face cells (Chang and Tsao, 2017). The study found that the active appearance model provides a remarkably simple account of AM activity: AM cells are approximately encoding linear combinations of axes of this model. However, this study left several issues unaddressed. First, the study used synthetic faces generated by the active appearance model rather than real faces. The code for real faces in the macaque brain may be different from that for synthetic faces. Furthermore, since the faces tested were directly controlled by parameters of the active appearance model, this may have given an unfair advantage to this model over other models for explaining face cell responses. Second, while the study compared the active appearance model to a few other models, it notably did not evaluate state-of-art deep networks trained on face recognition. Convolutional neural networks (CNNs) trained to perform face recognition now achieve close-to-human or even better performance (Parkhi et al., 2015; Taigman et al., 2014), naturally raising the question, how similar are the representations used by these artificial networks compared to those used by the primate face patch system?

Here, we set out to compare a large set of different models for face representation in terms of their power to explain neural responses from macaque face patch AM to pictures of real faces. The models tested include the original active appearance model used by Chang and Tsao (Chang and Tsao, 2017), referred to below as the "2D Morphable Model", an Eigenface Model (Sirovich and Kirby, 1987; Turk and Pentland, 1991), a 3D Morphable Model (Blanz and Vetter, 1999; Paysan et al., 2009), several CNN models (Krizhevsky et al., 2012; Parkhi et al., 2015; Simonyan and Zisserman, 2015; Kubilius et al., 2018), a β variational autoencoder (Higgins et al., 2017), and a model implementing Hebbian learning on V1-like representations (Leibo et al., 2017).

Results

We collected 2100 real faces from multiple online face databases, including the FERET face database (Phillips et al., 2000; Phillips et al., 1998b), CVL face database (Solina et al., 2003), MR2 face database (Strohminger et al., 2016), PEAL face database (Gao et al., 2008), AR face database (Martinez and Benavente, 1998),

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Chicago face database (Ma et al., 2015), and CelebA database (Yang et al., 2015) (Figure 1A). Responses of 159 face-selective cells in macaque face patch AM were recorded from two monkeys while presenting the facial images (Figure 1B). To find the optimal set of axes explaining neuronal responses, we extracted feature vectors from several different models, including 2D Morphable Model (Cootes et al., 2001), 3D Morphable Model (Blanz and Vetter, 1999), Eigenface Model (Sirovich and Kirby, 1987; Turk and Pentland, 1991), AlexNet (Krizhevsky et al., 2012), VGG-face (Parkhi et al., 2015), VGG-19 (Simonyan and Zisserman, 2015), CORnet (Kubilius et al., 2018), β-VAE (Higgins et al., 2017), and a model implementing Hebbian learning on V1-like representations (Leibo et al., 2017; Figure 1C). These models each parameterize faces using very different principles. The Eigenface model has the simplest form, consisting of principal components of pixel-level representations of facial images. The 2D and 3D Morphable Models are generative face models that convert a set of parameters into a facial image. AlexNet, VGGs, and CORnet are neural network models each trained on a different task: AlexNet, VGG-19, and the CORnets are all trained to classify images into 1000 non-face object categories. VGG-face is trained to identify 2,622 celebrities. The CORnet family includes three networks: CORnet-Z, CORnet-R, and CORnet-S. All three models have four areas that are identified with cortical areas V1, V2, V4, and IT. CORnet-Z has a purely feedforward structure, while CORnet-R and S contain recurrent connections within areas. β-VAE is a deep generative model that learns to faithfully reconstruct the input images, while being additionally regularized in a way that encourages individual network units to code for semantically meaningful variables. The Hebbian learning model is a biologically plausible model accounting for mirror-symmetric view tuning in face patch AL and view invariance in face patch AM. We chose these models because they are well known CNN models trained on object categorization (AlexNet, VGGs and CORnets), important computational models for face recognition (Eigenface, 2D Morphable Model, 3D Morphable Model, and Hebbian learning model), or a state-of-art neural network model for unsupervised disentangled representation learning (β-VAE).

To quantify how well each model can explain AM neuronal responses, for each model, we learned a linear mapping between features of that model and the neural population response vector. To avoid overfitting, we first reduced the dimensionality of each model by performing principal components analysis (PCA) on model responses to the 2100 faces, yielding N features for each face and each model. Then a 50-fold cross-validation paradigm was performed: responses of each neuron to 42*49=2058 faces were fit by linear regression using the N features, and then the responses of the neuron to the remaining 42 faces were predicted using the same linear transform. To quantify prediction accuracy, we compared the predicted population response vector to each face with the actual population response vector to the face as well as the population response vector to a random distractor face (Figure 1D). If the angle between prediction and target was smaller than that between prediction and distractor, the prediction was considered correct.

To compare different models, we used the top 50 PCs of features from each model, as in a previous study (Chang and Tsao, 2017). We found the best model was one of the CORnet models, CORnet-Z, followed by the 2D Morphable Model (Figure 2A). Interestingly, we found that VGG-face, a deep network trained to identify individual faces, performed worse than the other models, while CORnet-Z, which was trained to classify 1000 classes of non-face objects, performed better than any other model. A

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

confounding factor is that the images we used came from multiple face databases which have variable backgrounds. Some of the models may use background information more than other models for prediction. Hence performance differences between models could have been driven by representation of the background. In theory, the background should not be relevant to a model of face representation.

Thus we next extracted features from facial images without background (see Methods), and repeated the comparison of different models. The ordering of model performance was largely preserved after background removal (Figure 2B). However, after background removal, the performance of the 2D Morphable Model was not significantly different from CORnet-Z (p=0.76); this is consistent with the fact that the 2D Morphable Model only accounts for intensity variations of faces, but not background.

In the above two cases, we found the performance of the 3D Morphable Model was much lower than the 2D Morphable Model. However, there is an important difference between the 2D and 3D Morphable Models: the latter does not fit hair-related features. To compensate for this difference, we further tested the models on hairless facial images derived from fits using the 3D Morphable Model (Figure 2C, left). We performed the analysis using either 50 PCs or 110 PCs (the dimension of the 3D Morphable Model). In both cases, the 3D Morphable Model outperformed VGG-face, VGG-19, CORnet-R, CORnet-S, Eigenface (Figure 2C). In the case with 110 PCs, it even outperformed AlexNet, CORnet-Z and the 2D Morphable Model (Figure 2D). For faces without hair, the 2D Morphable Model also performed significantly better than all of the neural network models.

Furthermore, the use of the facial images generated by the 3D Morphable Model allowed us to test a Hebbian learning model recently proposed to account for face patch responses (Leibo et al., 2017). This model posits that the weights of face cells, learned through Hebbian learning, converge to the top PCs of the neuron's past inputs, and these inputs should generically constitute short movies of faces rotating in depth. The 3D Morphable Model allowed us to readily synthesize a set of facial images at multiple views and thus test the Hebbian model (see Methods). The Hebbian model performed better than the VGGs, CORnet-R and Eigenface models, comparably to CORnet-S (p=0.09 for 50 PCs and p=0.17 for 110 PCs), and worse than the 2D Morphable Model, 3D Morphable Model, AlexNet, and CORnet-Z (Figure 2C,D; note that the difference between the Hebbian model and AlexNet was not significant for 50 PCs, with p=0.06).

Finally, we performed a more detailed comparison between the 2D morphable model and another generative model, β-VAE, whose latent units are encouraged to encode semantically meaningful variables, otherwise known as disentangled variables. β-VAE contains an encoder that transforms the image input into a vector of disentangled latent variables and a decoder that transforms the vector back into an image. The encoder is implemented with a convolutional neural network, similar to other network models. A recent analysis of the same data set as in the present paper found that a subset of single AM cells have selectivity remarkably matched to that of single β-VAE latents, suggesting that AM and β-VAE have converged, at least partially, upon the same set of parameters for describing faces (Irina Higgins, personal communication). In particular, the single-neuron alignment between AM and β-VAE was better than that for any other model including the 2D morphable model. Thus we wanted to address in detail how β-VAE compares to the 2D morphable

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

model by the metric of neural population encoding. Due to variations in training parameters, 400 different β-VAE models were trained, each with 50 latent units. For the comparison in Figure 2A, we chose the β-VAE with the least encoding error. We further compared encoding performance of all 400 β-VAE models to that of the 2D morphable model. Close inspection revealed that some of the latent units have much smaller variance than other units in response to 2100 faces, thus we removed those units with variance <0.01. We found for both β-VAE and 2D Morphable model encoding errors decreased when the dimensionality increased, as one might expect, since more dimensions are available for capturing the information present in the neural responses, and the 2D Morphable model performed better than β-VAE models at matched feature dimensions (Figure 2E, $p<0.05$ in all cases, $p<0.01$ for dimension=10 or ≥18). However, among the 400 β-VAEs reported in Figure 2E, many models did not learn a well disentangled representation, hence failing at one of the optimisation objectives. Therefore, we also compared the 2D Morphable model to a subset of β-VAEs where the units were well disentangled (based on the unsupervised disentangled ranking (UDR) score for each β-VAE, see Methods), and found that the encoding performance difference was not significant at low dimensions, but the 2D Morphable model performed better for feature dimensions=6, 10 and 12 (Figure 2E, *inset*). Thus we conclude that by the metric of neural encoding performance, the 2D morphable model outperforms other generative models, including both β-VAE and Eigenface models, and performs similarly to disentangled β-VAEs at lower dimensions.

Next, we asked the complementary question: how well could we predict the model features by linear combinations of neural responses of face cells? The same procedure as the encoding analysis was followed, except that the respective roles played by model features and neural responses were reversed. The results for this decoding analysis were largely consistent with encoding analysis (Figure 3, Figure S1A). For the original images and images without the background, the 2D Morphable model performed better than all other models except CORnet-Z ($p=0.06$ for original images, $p=0.07$ after background removal). For images generated by the 3D Morphable model, the 2D Morphable model outperformed all other models except the 3D Morphable model ($p=0.68$ for 50-d model, $p<0.001$ for 110-d model). Comparing the 2D Morphable model with β-VAEs at matched dimensions, we found that the two models were comparable at dimensions≤18 (Figure 3E), but the 2D Morphable model performed better at higher dimensions. The fact that the performance of the two models were more comparable in both decoding and encoding at low dimensions (Figures 2E, 3E) suggests that with certain training parameters, β-VAE was efficient at extracting a small number of meaningful features, but the number of disentangled features discovered in this way may not be sufficient to achieve a good performance in decoding/encoding. To better illustrate the relationship between feature dimensionality, encoding/decoding errors and disentanglement, we quantified the quality of disentanglement achieved by β-VAE models by the UDR score as before. We found positive correlations between the encoding/decoding errors and UDR (Figure 2F and 3F), and a negative correlation between the number of informative features and UDR (Figure S1B). These results support the idea that the objective of disentanglement encourages the model to converge to a small number of informative features, at the expense of the overall explanatory power of the face code in the brain.

Overall, our results suggest that linear combinations of features of the 2D Morphable Model were closely related to the responses of face cells, achieving encoding and

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

decoding performance comparable to even the best neural network models developed recently (Figure 2 and 3), while at the same time using a simple and transparent representation that does not involve hundreds of thousands of "black box" parameters. This result also extends our previous finding from synthetic faces to real faces (Chang and Tsao, 2017). The similar performance of several of the models to the 2D Morphable Model (e.g., AlexNet and CORnet-Z in Figure 2A and 3A) raises the question whether these other models are simply linear transformations of the 2D Morphable Model, or whether they provide "additional" features that could help explain neural responses.

To address this question, we plotted both neural responses and model features in the space of the 2100 face stimuli, with each neuron represented by a 2100-d vector corresponding to the response of the neuron to the 2100 faces, and each feature represented by a 2100-d vector corresponding to the value of that feature across the 2100 faces. If the response of a neuron can be expressed as a linear combination of the top 50 features of a specific model, then the 2100-d response vector of the neuron should lie in a 50-d linear subspace spanned by the feature vectors for that model (since $\vec{R} = a_1\vec{f}_1 + \ldots a_{50}\vec{f}_{50}$, where $\vec{R}$ is the response vector of the neuron, and $\vec{f}_i$ is the ith feature vector). To test whether this is the case, we projected the response of each neuron onto the subspace spanned by each model (Figure 4A). The vector length of the projection quantifies how well linear combination of model features explains the responses. We performed PCA on the set of projection vectors to quantify how well model features explain neuronal responses at the population level. The top PCs identify directions in the 50-d subspace explaining the largest variance of neural data. Cumulative explained variance was then computed (Figure 4B-E). This analysis was performed using either facial images without background or hair-free reconstructions by the 3D Morphable Model as model inputs. For facial images without background, the 2D Morphable Model and CORnet-Z accounted for the most variance across all models, followed by AlexNet, CORnet-R, CORnet-S, Eigenface, 3D Morphable Model, VGG-19 and VGG-face (Figure 4B, solid lines). For hair-free reconstructions, the 2D Morphable Model accounted for the most variance, followed by CORnetZ, the 3D Morphable Model, AlexNet, CORnet-S, VGG-19, CORnet-R, Eigenface and VGG-face (Figure 4D, solid lines). The ordering was quite consistent with the previous quantification using encoding errors (cf. Figure 2).

Next, we wanted to know how different models deviate from the 2D Morphable Model and how much variance could be accounted by such deviations; in other words, *to what extent do the different face space models encode distinct face subspaces*? Towards this aim, we first orthogonalized each feature of one model to 50 features of the 2D Morphable Model, with the linear combinations of the orthogonalized features forming the orthogonalized subspace. Then we quantified variances of neural responses accounted by each orthogonal model, as we did previously for the full model (dashed lines in Figure 4B and D). Overall, we found variances explained by orthogonal models to be much smaller than that by full models (the gray trace indicates chance level computed by randomly shuffling neural responses to different faces), suggesting other models provide limited "additional" features to explain neural responses. For facial images without background, orthogonalized CORnet-Z accounted for the most variance across all models, followed by AlexNet, CORnet-R, CORnet-S, VGG-19, VGG-face, 3D Morphable Model and Eigenface (Figure 4B, dashed lines). For hair-free reconstructions, orthogonalized CORnet-Z accounted for the most variance across all models, followed by CORnet-S, AlexNet, CORnet-R,

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

VGG-19, 3D Morphable Model, VGG-face and Eigenface (Figure 4D, dashed lines). Furthermore, we asked the opposite question: what happens when the 2D Morphable Model is orthogonalized to other models? Again, we found variances explained by the orthogonalized 2D Morphable Model to be much smaller than the full 2D Morphable Model, indicating that the 2D Morphable Model features significantly overlap those of other models. However, a significant extent of non-overlap was also found (dashed lines in Figure 4C, E). For facial images without background, the most explained variance was achieved by the 2D Morphable Model orthogonalized to VGG-face, followed by VGG-19, 3D Morphable Model, Eigenface, CORnet-S, CORnet-R, AlexNet and CORnet-Z (Figure 4C, dashed lines). For hair-free reconstructions, the most explained variance was achieved by the 2D Morphable Model orthogonalized to VGG-face, followed by Eigenface, CORnet-R, VGG-19, CORnet-S, AlexNet, CORnet-Z and 3D Morphable Model (Figure 4E, dashed lines).

In the above analysis, models were compared with respect to the amount of variance they could explain in the neural responses. We also asked how well each model could explain features of other models through linear regression, independent of neural responses. To address this, we repeated the analysis of Figure 4, but instead of using real neurons, we constructed 159 simulated neurons by linear combinations of features of a particular model. These simulated responses were then projected into the subspace spanned by features of the same model (e.g., simulated neurons of AlexNet were projected into the subspace of AlexNet features) (Figure S2A, solid lines), as well as the subspace spanned by the same model features orthogonalized to the 2D Morphable Model (Figure S2A, dashed lines). As expected, the solid lines all approach 1, since the same models were used to both simulate and predict responses. Substantial variance was explained by features orthogonal to the 2D Morphable Model (Figure S2A, dashed lines), suggesting the models do provide additional features beyond those of the 2D Morphable Model. However, not all of these features help to explain neural responses. For example, the Eigenface model contains a sizable component orthogonal to the 2D Morphable Model (Figure S2A), but the amount of variance explained by the component of the Eigenface model orthogonal to the 2D Morphable Model was only slightly higher than chance level (Figure 4B). We also used the 2D Morphable Model to simulate neuronal responses, and asked how well 2D Morphable features orthogonal to each of the other models explain 2D Morphable features (dotted traces in Figure S2B, analogous to Figure 4C). The most variance was explained by the component of the 2D Morphable Model orthogonal to Vgg-face, while the least variance was explained by the component orthogonal to Eigenface. Finally, to compare all model pairs on equal footing, Figure S2C plots the amount of variance in each 50-feature face model explained by each of the other models. Interestingly, the 2D Morphable Model explained as much variance in CORnet-Z features as AlexNet. This shows that for the subspace of faces, an explicit generative model of face representation, the 2D Morphable Model, can do as good a job at explaining CORnet-Z features as a deep network explicitly built with similar architectural principles and training procedure (AlexNet).

Finally, we wanted to gain some insight into why AlexNet outperforms VGG-face in explaining neural responses, demonstrated by both the encoding/decoding analysis (Figure 2A-D, Figure 3A-D) and the explained variance analysis (Figure 4B, D, solid traces). This is surprising since AlexNet is not trained to classify any face images (albeit some images within the different training classes do contain faces), while Vgg-face is trained exclusively to identify face images. We started by computing similarity

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

matrices (Kriegeskorte et al., 2008) for neural population responses from face patch AM (Figure 5A1). Here each entry of the matrix represents the similarity between a pair of faces, quantified as correlation between population responses to the face pair. The same analysis was repeated with the top 50 PCs of deep features from AlexNet or VGG-face (Figure 5A2, A3). There is a clear difference between the similarity matrix for VGG face compared to those for AM and Alexnet features. Similarity matrices for both AM and AlexNet features show a dark cross with a bright center, but this is not the case for VGG face. We computed the difference between the Vgg-face and AlexNet matrices (Figure 5A4), and then shuffled the rows and columns according to the first principle component of the difference matrix (Figure 5A5). After sorting, positive entries tended to be located at the bottom-left and the upper-right corner (square outlines in Figure 5A5): here a positive difference indicates the two faces are more similar under features of VGG-face than AlexNet; therefore the faces at the opposite ends of PC1 are more likely to be confused by VGG-face. What do the faces at the two extremities look like? To examine the difference, we picked the first 100 faces and last 100 faces along the direction of PC1, and divided them into 20 groups of 10 faces. An average face after shape normalization was generated for each group (Figure 5B). We see an interesting difference: The first 10 groups of faces show inhomogeneous illumination--some parts of faces, such as cheeks and hair, are brighter than other parts of the face, such as the mouth, while the last 10 groups of faces appear more homogeneously illuminated.

In the analyses of Figure 5, we used a database, CAS-PEAL, which contains only Chinese faces. Is this observation unique to Chinese faces? We repeated the same analysis for 748 Caucasian faces. Similar to CAS-PEAL faces, we found that the face groups eliciting a much more similar representation by Vgg-face compared to AlexNet consisted of faces with unbalanced versus homogeneous illumination (Figure S3). In sum, we found that VGG-face is much less sensitive to illumination differences than both AM cells as well as AlexNet, and this likely contributes to the inferior ability of Vgg-face to predict AM responses compared to AlexNet.

Discussion

Face processing has been a subject of intense research effort in both visual neuroscience and computer vision, naturally raising the question, what, if any, computer vision model of face representation best matches that used by the primate brain. A recent paper found the 2D Morphable Model, a classic model of face representation from computer vision, could explain neural activity in face patches remarkably well (Chang and Tsao, 2017). At the same time, a number of groups have found that activity in deep layers of convolutional neural networks can explain significant variance of neural responses in ventral temporal cortex (Yamins et al., 2014; Kalfas et al., 2017; Yildirim et al., 2020; Schrimpf et al., 2018). Here, we extend those results by comparing the efficacy of a large number of different computational models of face representation to account for neural activity in face patch AM. We were especially interested in how the 2D morphable model, a simple and explicit graphical model, would compare to Vgg-face, a black box deep neural network dedicated to face recognition containing hundreds of thousands of parameters and trained on nearly a million (982,803) facial images. Our findings suggest that the 2D Morphable Model is better than most other models in explaining the neuronal representation of real faces including Vgg-face. For faces without background, the 2D Morphable Model allowed better linear coding of neural responses by model features than every model except CORnet-Z, whose performance it matched;

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

specifically, the 2D Morphable Model performed worse than CORnet-Z for faces with both hair and backgrounds, comparably to CORnet-Z for faces with no backgrounds, and better than CORnet-Z for faces without hair reconstructed by 3D Morphable Model. This is surprising, since the 2D Morphable Model is one of the oldest models (next to the Eigenface model). Comparison of the subspaces spanned by the different models revealed that of the different neural network models, CORnet-Z spanned the largest feature space in common with the 2D Morphable Model, both in absolute terms (Figure S2C, row 1) as well as for the purpose of explaining neural responses (Figure 4C, E). Furthermore, the 2D Morphable Model could explain CORnet-Z features as well as AlexNet, a deep network whose architectural principles and training procedure CORnet-Z explicitly emulates (Figure S2C, row 7). These results suggest that the face subspace portion of the representation learned by CORnet-Z may be interpreted in much simpler terms, as a shape appearance model. The results provide an important counter-example to the increasingly popular view that only distributed representations learned by multi-layer networks can well explain IT activity (Kietzmann et al., 2018; Lillicrap et al., 2020). Why a network trained on object classification should learn an approximation to a generative model of faces is an interesting question for future research.

Overall, the 2D Morphable Model is composed of two components: a shape component defined by positions of facial landmarks, and a shape-free appearance (or texture) component defined by the intensity distribution of the facial image after shape normalization. Neither component is exclusively bound to facial identity: a slight rotation or change in facial expression will alter the shape coordinates, while changes in lighting conditions will alter the appearance coordinates. Previously, we found that AM cells deal with the former source of variation by largely ignoring changes in shape dimensions (Chang and Tsao, 2017). Here, we find that for the latter source of variation, AM cells do represent lighting in population responses, consistent with the observation that our recognition of unfamiliar faces is susceptible to changes in lighting conditions (Young and Burton, 2018). We believe this finding partially explains the deviation of neural representation from models trained to largely ignore lighting, e.g., Vgg-face (Figure 5).

A recent theoretical study found that a deep neural network, when trained on certain tasks, gradually abandons information about the input unrelated to the task in its deep layers (Tishby and Zaslavsky, 2017). Thus artificial neural networks are unlikely to be fully identical to the brain, since the tasks both systems are trained on are unlikely to be identical. It makes sense that VGG-face is not able to distinguish illumination, since the identity of an individual doesn't depend on illumination. Why does AlexNet still contain information about illumination? It is possible this occurs because AlexNet has not been trained specifically on face identification, and illumination-related features are useful for more general object classification tasks (e.g., distinguishing a concave hole from a convex bump (Ramachandran, 1988)). In contrast, it appears that VGG-face is so specialized that any information unrelated to identity is filtered out in the end.

The fact that AM neurons are more consistent with the 2D morphable model than with VGG-face suggests macaque face cells are not over-specialized for facial identification, but rather provide an array of high-level information related to different aspects of faces in the visual field. Because the 2D Morphable Model retains all information needed to reconstruct a face, it should be able to perform well on any face-related task. Of course, preservation of information is not the only goal for visual

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

representations in the brain, as the Eigenface model should be the model most linearly related to the image input, but it does not perform as well as the 2D Morphable Model or AlexNet. The format of information representation is also important. Another example is β-VAE, which does not appear to preserve enough information to perform as well as the 2D Morphable Model for encoding/decoding neural responses at higher dimensions. At lower dimensions, however, β-VAE models have similar encoding/decoding performance as the 2D Morphable Model and furthermore show better single-neuron alignment with AM units (Irina Higgins, personal communication).

The three CORnet models were recently developed and compared in terms of their ability to explain general IT responses (not restricted to face patches) in another study (Kubilius et al., 2018). CORnet-Z is the simplest model, which has a purely feed-forward structure, while CORnet-R and CORnet-S are recurrent models. When compared for ability to explain a dataset containing IT responses to a set of general object stimuli with a large range of variations in orientation, size, and locations, it was found that CORnet-S performed best and CORnet-Z worst among the three CORnets (cf. Figure 3 in Kubilius et al., 2018). This is quite different from our results, where CORnet-Z was the best model for encoding face cell responses. This could be a result of stimulus selection: the recurrent layers could help establish invariance in complex/difficult situations, but may not be a big advantage in our case, since our stimuli are well aligned. This suggests the results of model comparison depend on the stimuli being used in the study.

Overall, our analyses comparing a large number of models in terms of their ability to explain responses of cells in face patch AM show that a simple and explicit generative face model, the 2D Morphable Model, performs surprisingly well—rivaling or surpassing the deep network classifiers considered. This result supports the hypothesis that deep networks may be generally understood as inverting generative models (Lin and Tegmark, 2016; Ho et al., 2018), and raises the possibility that mechanisms underlying general object recognition may be understood in similar explicit terms without relying on “black box” neural networks. Furthermore, the extremely poor performance of a deep network trained for face recognition in explaining face cell responses may give insight into face patch development, raising the possibility that it may be optimized not for face recognition per se but instead for face reconstruction supporting arbitrary face-related behaviors.

STAR Methods

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
Rhesus macaques (Macaca mulatta)	UC Davis primate research center	N/A
Software and Algorithms		
MATLAB	MathWorks	http://mathworks.com/
MatConvNet	VLFeat	http://www.vlfeat.org/matconvnet

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Basel face model	Gravis Research Group, University of Basel	https://github.com/unibas-gravis/basel-face-pipeline
Menpo	Intelligent Behaviour Understanding Group, Imperial College London	http://www.menpo.org
Other		
Tungsten Microelectrode	FHC	Lot #:221355

Experimental Model Details

Two male rhesus macaques (Macaca mulatta) of 7-10 years old were used in this study. Both animals were pair-housed and kept on a 14 hr/10hr light/dark cycle. All procedures conformed to local and US National Institutes of Health guidelines, including the US National Institutes of Health Guide for Care and Use of Laboratory Animals. All experiments were performed with the approval of the Caltech Institutional Animal Care and Use Committee (IACUC).

Face Patch Localization

Two male rhesus macaques were trained to maintain fixation on a small spot for juice reward. Monkeys were scanned in a 3T TIM (Siemens, Munich, Germany) magnet while passively viewing images on a screen. Feraheme contrast agent was injected to improve signal/noise ratio. Face patches were determined by identifying regions responding significantly more to faces than to bodies, fruits, gadgets, hands, and scrambled patterns, and were confirmed across multiple independent scan sessions. Additional details are available in previous publications (Freiwald and Tsao, 2010; Ohayon et al., 2012; Tsao et al., 2006).

Single-unit Recording

Tungsten electrodes (18–20 Mohm at 1 kHz, FHC) were back loaded into plastic guide tubes. Guide tubes length was set to reach approximately 3–5 mm below the dura surface. The electrode was advanced slowly with a manual advancer (Narishige Scientific Instrument, Tokyo, Japan). Neural signals were amplified and extracellular action potentials were isolated using the box method in an on-line spike sorting system (Plexon, Dallas, TX, USA). Spikes were sampled at 40 kHz. All spike data were re-sorted with offline spike sorting clustering algorithms (Plexon). Only well-isolated units were considered for further analysis.

Behavioral Task and Visual Stimuli

Monkeys were head fixed and passively viewed the screen in a dark room. Stimuli were presented on a CRT monitor (DELL P1130). The intensity of the screen was measured using a colorimeter (PR650, Photo Research) and linearized for visual stimulation. Screen size covered 27.7*36.9 visual degrees and stimulus size spanned 5.7 degrees. The fixation spot size was 0.2 degrees in diameter and the fixation window was a square with the diameter of 2.5 degrees. Images were presented in

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

random order using custom software. Eye position was monitored using an infrared eye tracking system (ISCAN). Juice reward was delivered every 2–4 s if fixation was properly maintained. For visual stimulation, all images were presented for 150 ms interleaved by 180 ms of a gray screen. Each image was presented 3–5 times to obtain reliable firing rate statistics. In this study, two different stimulus sets were used:

a) A set of 16 real face images, and 80 images of objects from nonface categories (fruits, bodies, gadgets, hands, and scrambled images) (Freiwald and Tsao, 2010; Ohayon et al., 2012; Tsao et al., 2006).

b) A set of 2100 images of real faces from multiple face databases, FERET face database(Phillips et al., 2000; Phillips et al., 1998b), CVL face database(Solina et al., 2003), MR2 face database(Strohminger et al., 2016), PEAL face database(Gao et al., 2008), AR face database(Martinez and Benavente, 1998), Chicago face database(Ma et al., 2015) and CelebA database(Yang et al., 2015). 17 online photos of celebrities were also included.

Quantification and Statistical Analysis

Selection of face selective cells

To quantify the face selectivity of individual cells, we defined a face-selectivity index as:

$$FSI = \frac{mean\ response_{face} - mean\ response_{nonface\ objects}}{mean\ response_{face} + mean\ response_{nonface\ objects}} \quad (1)$$

The number of spikes in a time window of 50-350 ms after stimulus onset was counted for each stimulus. Units with high face selectivity (FSI > 0.33) were selected for further recordings.

Extraction of facial feature from images

Each facial image was fed into the following models to extract corresponding features:

1) 2D Morphable Model

This is the same model as used in our previous paper (Chang and Tsao, 2017) and feature extraction followed the procedure of previous papers on active appearance modeling (Cootes et al., 2001; Edwards et al., 1998). First, a set of 80 landmarks were labeled on each of the 2100 facial images. Out of the 80 landmarks, 68 were automatically labeled using an online package ("menpo", http://www.menpo.org) and the remaining 12 were manually labeled. The positions of landmarks were normalized for mean and variance for each of the 2100 faces, and an average shape template was calculated. Then each face was smoothly warped so that the landmarks matched this shape template, using a technique based on spline interpolation (Bookstein, 1989). This warped image was then normalized for mean and variance and reshaped to a 1-d vector. Principal component analysis was carried out on positions of landmarks and vectors of shape-free intensity independently. Equal numbers of shape PCs and shape-free appearance PCs were extracted to compare with other models (25 shape/25 appearance PCs vs. 50 features of other

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

models; 55 shape/55 appearance PCs vs. 110 features of other models). This model was also used to generate images without background used in Figure 2B. In this case, we first morphed all 2100 faces to the shape template, defined a mask for the standard shape to remove the background, and then morphed the masked facial image back to the original shape.

2) 3D Morphable Model

We built a grayscale variant of the Basel Face Model (Paysan et al., 2009) from the original 200 face scans. The ill-posed 3D reconstruction from a 2D image was solved using (Schonborn et al., 2017) and the publicly available code from (Gerig et al., 2018). The first 50 principal components for the shape and color model respectively were adapted during the model adaptation process. The sampling-based method was initialized with the same landmarks as provided to the 2D Morphable Model. The pose was fixed to a frontal pose and the spherical harmonics illumination parameters were estimated robustly using (Egger et al., 2018) and the average illumination condition was fixed for the whole dataset. Note that the full complexity and flexibility of the 3DMM is not explored when analyzing frontal images only. Besides the model adaptation novel views where generated using the standard 3DMM pipeline by changing the head orientation and camera parameters. The images with varying head orientations were used to construct the Hebbian learning model (see below).

3) Eigenface model

PCA was performed on the original image intensities of 2100 faces and top 50/110PCs were extracted to compare with other models.

4) Pre-trained neural network models

We loaded 2100 facial images into the following pre-trained neural networks: (1) a MATLAB implementation of AlexNet: This network contains 8 layers: 5 convolutional layers and 3 fully connected layers, and has been pre-trained to identify a thousand classes of non-face objects. (2) a MATLAB implementation of Vgg-face neural network (Parkhi et al., 2015). This network contains 16 layers: 13 convolutional layers+3 fully connected layers, and has been pre-trained to recognize faces of 2622 identities. (3) a MATLAB implementation of Vgg-19 neural network (Simonyan and Zisserman, 2015). This network contains 19 layers: 16 convolutional layers and 3 fully connected layers, and has been pre-trained to identify a thousand objects. (4) a PyTorch implementation of CORnet (Kubilius et al., 2018). The CORnet family includes three networks: CORnet-Z, CORnet-R, and CORnet-S. All three models have four areas that are identified with cortical areas V1, V2, V4, and IT. CORnet-Z is the simplest model of the three, involving only feedforward connections, CORnet-R introduces recurrent dynamics within each area into the otherwise purely feed-forward network, and CORnet-S is the most complicated (containing the most convolutional layers and including skip connections), aiming to match neural and behavioral data. The three CORnet models have been pre-trained to identify a thousand classes of objects. Parameters of the first three pretrained networks were downloaded from: http://www.vlfeat.org/matconvnet/pretrained/. CORnets were downloaded from https://github.com/dicarlolab/CORnet. PCA was performed on activation of units in the penultimate layers (IT area in the case of CORnet), and top 50/110PCs were extracted to compare with other models.

5) β-VAE model

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We used the standard architecture and optimisation parameters introduced in (Higgins et al., 2017) for training the β-VAE. The encoder consisted of four convolutional layers (32x4x4 stride 2, 32x4x4 stride 2, 64x4x4 stride 2), followed by a 256-d fully connected layer and a 50-d latent representation. The decoder architecture was the reverse of the encoder. We used ReLU activations throughout. The decoder parametrised a Bernoulli distribution. We used Adam optimiser with 1e-4 learning rate and trained the models for 1 mln iterations using batch size of 16, which was enough to achieve convergence. The models were trained to optimise the following disentangling objective:

$$\mathcal{L}_{\beta-VAE} = E_{p(\mathbf{x})}[\, E_{q_\varphi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta KL(q_\varphi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z}))\,] \quad (2)$$

where p(x) is the probability of the image data, q(z|x) is the learnt posterior over the latent units given the data, and p(z) is the unit Gaussian prior with a diagonal covariance matrix.

For the β-VAE model the main hyperparameter of interest that affects the quality of the learnt latent units is the value of β. The β hyperparameter controls the degree of disentangling achieved during training, as well as the intrinsic dimensionality of the learnt latent representation (Higgins et al., 2017). Typically a β>1 is necessary to achieve good disentangling, however the exact value differs for different datasets. Hence, we trained 400 models with different values of β by uniformly sampling 40 values of β in the [0.5, 20] range. Another factor that affects the quality of disentangled representation is the random initialisation seed for training the models. Hence, for each β value, we trained 10 models from different random initialisation seeds, resulting in the total pool of 400 trained β-VAE.

The recently proposed Unsupervised Disentanglement Ranking (UDR) score (Duan et al., 2020) was used to select 51 model instances with the most disentangled representations (within the top 15% of UDR scores). The UDR score measures the quality of disentanglement achieved by trained β-VAE models by performing pairwise comparisons between the representations learnt by models trained using the same hyperparameter setting but with different seeds (Duan et al., 2020).

6) Hebbian learning model

This is a biologically plausible model recently proposed to explain view invariance in face cells (Leibo et al., 2017). V1-like features (C1-layer of HMAX model) were extracted from the facial images. PCA was performed on V1-like encodings of a single identity at different head orientations: the ith PC of the kth identity is denoted as w_i^k. The activation of the kth unit to a given face is $\mu^k(x) = \sum_{i=1}^{r} \langle x, w_i^k \rangle^2$, where x is the V1-like encoding of that face, and r is the number of PCs being used. In our experiment, we used rotated versions of the fitted 3D Morphable Models (from -90ºto 90º in 5ºincrements) as inputs to this model (Figure 2C, 3C), resulting in 2100 such units. PCA was performed on activation of the 2100 units, and the top 50/110PCs were extracted. Since the 3D Morphable Model only fits part of the face and may not provide a satisfactory explanation of neural responses to full faces, we only implemented the Hebbian model on 3D-fits of the original images and compared it to other models under the same condition (Figure 2C, 2D, 3C, 3D).

Quantification of encoding and decoding errors

For encoding analysis, responses of each neuron were first normalized to zero mean

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and unit variance. A 50-fold cross-validation paradigm was performed: 2100 faces were split into 10 groups of 42 faces. Responses of each neuron to 49 groups of faces were fit by a linear regression model using PCs of a set of features, and the responses of this neuron to the remaining group of 42 faces were predicted using the same linear transform. This process was repeated for all ten groups, so every single face had a predicted response. To quantify prediction accuracy, we examined the predicted responses to individual faces in the space of population responses, and compared this to either the actual response to the face (target) or that to a distractor face. If the angle between the predicted response and distractor response is smaller than that between the predicted response and target response, this was considered as a mistake. Encoding error was quantified as the frequency of mistakes across all pairs of target and distractor faces. Wilcoxon signed rank test was used to determine statistical significance of difference between two models.

For decoding analysis, features of each model dimension were first normalized to zero mean and unit variance. The same procedure used in the encoding analysis was employed, except that the respective roles played by neural responses and model features were reversed.

Similarity matrix

Based on the normalized population response, a similarity matrix of correlation coefficients was computed between the population response vectors to each of the n faces. For neural network models, top 50 PCs of activation of units in the penultimate layers of the networks were used to represent faces.

Figure Legends:

Figure 1. Stimulus and analysis paradigm.

A, 2100 facial photos from multiple face databases were used in this experiment. Three examples are shown. [Note that the facial images shown here and in Figures 2-4 are synthetically generated faces that serve as stand-ins for the actual example faces in order to satisfy bioRxiv's policy on the use of images of human faces; they are not from the database and were not shown to the monkeys.] B, Images were presented to the animal while recording from the most anterior face patch AM (anterior medial face patch). C, Each facial image was analyzed using 10 different models. The same number of features were extracted from units of different models using principal component analysis (PCA) for comparison. D, Different models were compared with respect to how well they could predict neuronal responses to faces. A 10-fold cross-validation paradigm was employed for quantification: 2100 faces were evenly distributed into 10 groups. Responses of each neuron to 9 groups were fit by linear regression using features of a particular face model, and the responses of this neuron to the remaining 210 faces were predicted using the same linear transform. To quantify prediction accuracy, we compared predicted responses to individual faces in the space of population responses to either the actual response to that face or that to a distractor face. If the angle between predicted response and target

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

response is smaller than that between predicted response and distractor response, the prediction is considered correct. All pairs of faces were used as both target and distractor and the proportion of correct predictions was computed.

Figure 2. Comparing how well different models of face coding can explain AM neuronal responses to facial images.

A, For each model, 50 features were extracted using PCA and used to predict responses of AM neurons. Encoding errors are plotted for each model. Error-bars represent s.e. for 2100 target faces (i.e., error was computed for each target face when comparing to 2099 distractors, and s.e. was computed for the 2100 errors). CORnet-Z performed significantly better than the other models ($p<0.001$ in all cases except from the 2D Morphable Model, $p<0.01$ between CORnet-Z and the 2D Morphable Model, Wilcoxon signed-rank test), and the 2D Morphable Model performed significantly better than the remaining models ($p<0.001$). B, To remove differences between models arising from differential encoding of image background, face images with uniform background were presented to different models (see Methods). CORnet-Z and 2D Morphable Model performed significantly better than the other models ($p<0.001$), with no significant difference between the two models ($p=0.76$). C, To create facial images without hair, each facial image in the database was fit using a 3D Morphable Model (left). [Note that the 3D morph fit shown here and in subsequent figures is a synthetically generated face that serves as a stand-in for the actual example 3D morph fit in order to satisfy bioRxiv's policy on the use of images of human faces.] The fits were used as inputs to each model. For example, a new 2-D Morphable Model was constructed by morphing the fitted images to an average shape. 50 features were extracted from each of the models using PCA for comparison. D, Same as C, but for 110 features. For 50 features, the 2D Morphable Model performed significantly better than the other models ($p<0.001$), while there was no significant difference between 3D Morphable Model and CORnetZ ($p=0.13$). For 110 features, the 3D Morphable Model outperformed all other models ($p<0.001$). E. Encoding errors for 400 β-VAEs after removing dimensions with variance<0.01 were compared with the 2D Morphable model at equivalent dimensions (equal number of shape and appearance dimensions were chosen for the 2D Morphable model). Wilcoxon signed-rank test was employed to compare the two models after performing 50-fold cross validation (*=$p<0.05$; **=$p<0.01$; n.s.=not significant). Inset, for the 51 most disentangled VAEs, subsets of features explaining the most variance of each model were compared to the 2D Morphable model at equivalent dimensions (since equal number of shape and appearance dimensions were selected for the 2D Morphable model, only even number of total dimensions were shown here). F. Encoding errors for all 400 β-VAEs were plotted against UDR score (left: full model; right: partial model after removing dimensions with variance<0.01).

Figure 3. Comparing how well AM neuronal responses to facial images can explain different models of face coding.

A, For each model, 50 features were extracted using PCA and responses of AM neurons were used to predict the model features. Decoding errors are plotted for

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

each model. Error-bars represent s.e. for 2100 target faces (i.e., error was computed for each target face when comparing to 2099 distractors, and s.e. was computed for the 2100 errors). CORnet-Z performed significantly better than the other models ($p<0.001$) except from the 2D Morphable Model ($p=0.06$, Wilcoxon signed-rank test), and the 2D Morphable Model performed significantly better than the remaining models ($p<0.001$). B, To remove differences between models arising from differential encoding of image background, face images with uniform background were presented to different models (see Methods). CORnet-Z and 2D Morphable Model performed significantly better than the other models ($p<0.001$), with no significant difference between the two models ($p=0.07$). C, To create facial images without hair, each facial image in the database was fit using a 3D Morphable Model (left). The fits were used as inputs to each model. For example, a new 2-D Morphable Model was constructed by morphing the fitted images to an average shape. 50 features were extracted from each of the models using PCA for comparison. D, Same as C, but for 110 features. For 50 features, the 2D Morphable Model and 3D Morphable model performed significantly better than the other models ($p<0.001$), with no significant difference between the two models ($p=0.68$). For 110 features, the 3D Morphable Model outperformed all other models ($p<0.001$). E. Decoding errors for 400 β-VAEs after removing dimensions with variance<0.01 were compared with the 2D Morphable model at equivalent dimensions (equal number of shape and appearance dimensions were chosen for 2D Morphable model). Wilcoxon signed-rank test was employed to compare the two models after performing 50-fold cross validation (*=$p<0.05$; **=$p<0.01$; n.s.=not significant). Inset, for the 51 most disentangled VAEs, subsets of features explaining the most variance of each model were compared to the 2D Morphable model at equivalent dimensions (since equal number of shape and appearance dimensions were selected for the 2D Morphable model, only even number of total dimensions were shown here). F. Decoding errors for all 400 β-VAEs were plotted against UDR score (left: full model; right: partial model after removing dimensions with variance<0.01).

Figure 4. Measuring explanatory subspace overlap between the 2D Morphable Model and other models

A, Analysis paradigm. Each neuron/model feature is represented by a single 2100-d vector, with each dimension representing response/feature value for one face. 50-d features of one model span a subspace of the 2100-d space. If responses are perfectly predicted by 50-d model features, i.e., $\vec{r} = M\vec{f}$, where $\vec{r}$ is an n-d neural response vector, M is an n x 50 matrix determined through linear regression, and $\vec{f}$ is the 50-d feature vector of the face, then neuron responses should span a subspace within that spanned by model features. Response vectors were projected onto each model subspace, and PCA was performed on the projected features (left). To compare two different face spaces, one of the model's features were first orthogonalized to the other model, and the orthogonalized features span another subspace (right). The same analysis as the original model can be performed for the orthogonalized space. B, Cumulative eigenvalues after PCA are plotted for 9 models (solid lines). Dashed lines indicate the results after othorgonalization to the 2D Morphable Model. Gray lines are results after randomly shuffling neuronal responses. C, The solid lines are the same as B, and dashed lines represent results of orthogonalizing the 2D Morphable Model to the other models. Image background was

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

removed before presenting the image to the network models (cf. Figure 2B). D and E, Same as B and C, but using reconstructions by the 3D Morphable Model as inputs.

Figure 5. Vgg-face features and AlexNet features show marked difference in coding illumination levels.

A, Similarity matrices were computed for 913 faces from CAS-PEAL database using AM population responses (A1) and features of two network models, AlexNet (A2) and Vgg-face (A3). Each entry indicates the correlation between representations of two faces. The difference between the two matrices derived from the network models was computed (A4). Rows and columns of the differential matrix were shuffled according to the first principal component of the difference matrix (A5). The red squares outline face pairs taken from the first and last 100 faces: these face pairs showed a significantly representational similarity by Vgg-face compared to AlexNet. B, First 100 faces and last 100 faces along the direction of PC1 were divided into 20 groups of 10 faces. An average face after shape normalization was generated for each group. [Note that images shown here are not actual faces of any individuals, but the average images of 10 faces after being morphed to the average shape, using the same algorithm as the 2D Morphable Model.]

Supplementary Information

Figure S1. Further analysis on encoding and decoding. Related to Figures 2 and 3.

A, Comparison between encoding and decoding errors for all models. 50 features were included in all cases. B, Relationship between the number of informative features (variance≥0.01) and UDR score for all 400 β-VAEs.

Figure S2. Direct comparison between feature spaces spanned by different models. Related to Figure 4.

A, Similar to Figure 4B, but instead of using responses of real neurons, 159 simulated neurons were constructed by linear combinations of features of a particular model, i.e., $\vec{r} = M\vec{f}$, where $\vec{r}$ is a response vector of 159 simulated neurons, M is a 159 x 50 matrix containing independent random variables following normal distribution N(0,1), and $\vec{f}$ is the 50-d feature vector of the face. These simulated responses were then projected into the subspace spanned by features of the same model (solid lines), as well as the subspace spanned by the same model features orthogonalized to the 2D Morphable Model (dashed lines). B, Simulated responses using 2D Morphable Model projected into the subspace by features of other models (solid lines), as well as the subspace spanned by 2D Morphable Model orthogonalized to those models (dashed lines). C, For each pair of two different models (X,Y), features of model X were fitted by features of model Y, both using 50 feature dimensions (PCs). Explained variances were then averaged across the 50 PCs of model X, weighted by the variance of the original features explained by each PC. The averaged explained variances of all model pairs were then color-coded and

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

plotted as a matrix, with its rows representing model Xs, and columns representing model Ys

Figure S3. Comparison between VGG-face and AlexNet for Caucasian faces. Related to Figure 5.

Same as Figure 5B, but for 748 Caucasian faces we presented. To attenuate the influence of diverse image backgrounds in multiple databases, we removed the background before presenting images to the networks (cf. Figure 2B). [Note that images shown here are not actual faces of any individuals, but the average images of 10 faces after being morphed to the average shape, using the same algorithm as the 2D Morphable Model.]

References:

Blanz, V., and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. Comp Graph, 187-194.
Bookstein, F.L. (1989). Principal Warps - Thin-Plate Splines and the Decomposition of Deformations. Ieee T Pattern Anal *11*, 567-585.
Chang, L., and Tsao, D.Y. (2017). The Code for Facial Identity in the Primate Brain. Cell *169*, 1013-1028 e1014.
Cootes, T.F., Edwards, G.J., and Taylor, C.J. (2001). Active appearance models. Ieee T Pattern Anal *23*, 681-685.
Duan, S., Matthey, L., Saraiva, A., Watters, N., Burgess, C., Lerchner, A., and Higgins, I. (2020). Unsupervised model selection for variational disentangled representation learning. ICLR *2020*.
Edwards, G.J., Taylor, C.J., and Cootes, T.F. (1998). Interpreting face images using Active Appearance Models. Automatic Face and Gesture Recognition - Third Ieee International Conference Proceedings, 300-305.
Egger, B., Schonborn, S., Schneider, A., Kortylewski, A., Morel-Forster, A., Blumer, C., and Vetter, T. (2018). Occlusion-Aware 3D Morphable Models and an Illumination Prior for Face Image Analysis. Int J Comput Vision *126*, 1269-1287.
Freiwald, W.A., and Tsao, D.Y. (2010). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. Science *330*, 845-851.
Gao, W., Cao, B., Shan, S.G., Chen, X.L., Zhou, D.L., Zhang, X.H., and Zhao, D.B. (2008). The CAS-PEAL large-scale Chinese face database and baseline evaluations. Ieee T Syst Man Cy A *38*, 149-161.
Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schonborn, S., and Vetter, T. (2018). Morphable Face Models - An Open Framework. Ieee Int Conf Automa, 75-82.
Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. ICLR *2017*.
Ho, N., Nguyen, T., Patel, A., Anandkumar, A., Jordan, M.I., and Baraniuk, R.G. (2018). Neural Rendering Model: Joint Generation and Prediction for Semi-Supervised Learning. arXiv *1811.02657*.
Kalfas, I., Kumar, S., and Vogels, R. (2017). Shape Selectivity of Middle Superior Temporal Sulcus Body Patch Neurons. eNeuro *4*.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Kietzmann, T., McClure, P., and Kriegeskorte, N. (2018). Deep Neural Networks in Computational Neuroscience. bioRxiv *doi: https://doi.org/10.1101/133504*.
Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci *2*, 4.
Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. NIPS *2012*.
Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D.L.K., and Dicarlo, J.J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. bioRxiv *doi: https://doi.org/10.1101/408385*
Leibo, J.Z., Liao, Q., Anselmi, F., Freiwald, W.A., and Poggio, T. (2017). View-Tolerant Face Recognition and Hebbian Learning Imply Mirror-Symmetric Neural Tuning to Head Orientation. Curr Biol *27*, 62-67.
Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., and Hinton, G. (2020). Backpropagation and the brain. Nat Rev Neurosci. Epub.
Lin, H., and Tegmark, M. (2016). Why does deep and cheap learning work so well? arXiv.
Ma, D.S., Correll, J., and Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. Behav Res Methods *47*, 1122-1135.
Martinez, A.M., and Benavente, R. (1998). The AR Face Database. CVC Technical Report *24*.
Ohayon, S., Freiwald, W.A., and Tsao, D.Y. (2012). What makes a cell face selective? The importance of contrast. Neuron *74*, 567-581.
Parkhi, O.M., Vedaldi, A., and Zisserman, A. (2015). Deep Face Recognition. British Machine Vision Conference *2015*.
Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3D Face Model for Pose and Illumination Invariant Face Recognition. Avss: 2009 6th Ieee International Conference on Advanced Video and Signal Based Surveillance, 296-301.
Phillips, P.J., Moon, H., Rizvi, S.A., and Rauss, P.J. (2000). The FERET evaluation methodology for face-recognition algorithms. Ieee T Pattern Anal *22*, 1090-1104.
Phillips, P.J., Wechsler, H., Huang, J., and Rauss, P. (1998a). The FERET database and evaluation procedure for face recognition algorithms. Image Vision Comput *16*, 12.
Phillips, P.J., Wechsler, H., Huang, J., and Rauss, P.J. (1998b). The FERET database and evaluation procedure for face-recognition algorithms. Image Vision Comput *16*, 295-306.
Ramachandran, V.S. (1988). Perception of shape from shading. Nature *331*, 163-166.
Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K.*, et al.* (2018). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? bioRxiv *doi: http://dx.doi.org/10.1101/407007.*
Schonborn, S., Egger, B., Morel-Forster, A., and Vetter, T. (2017). Markov Chain Monte Carlo for Automated Face Image Analysis. Int J Comput Vision *123*, 160-183.
Simonyan, K., and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv *1409.1556*.
Sirovich, L., and Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. J Opt Soc Am A *4*, 519-524.
Solina, F., Peer, P., Batagelj, B., Juvan, S., and Kovac, J. (2003). Color-based face detection in the "15 seconds of fame" art installation". Conference on Computer Vision / Computer Graphics Collaboration for Model-based Imaging, Rendering, image Analysis and Graphical special Effects, 10.
Strohminger, N., Gray, K., Chituc, V., Heffner, J., Schein, C., and Heagins, T.B.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(2016). The MR2: A multi-racial, mega-resolution database of facial stimuli. Behav Res Methods *48*, 1197-1204.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. Proc Cvpr Ieee, 1701-1708.

Tishby, N., and Zaslavsky, N. (2017). Deep Learning and the Information Bottleneck Principle. arXiv *1503.02406v1*.

Tsao, D.Y., Freiwald, W.A., Tootell, R.B., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670-674.

Turk, M.A., and Pentland, A.P. (1991). Face Recognition Using Eigenfaces. 1991 Ieee Computer Society Conference on Computer Vision and Pattern Recognition, 586-591.

Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A *111*, 8619-8624.

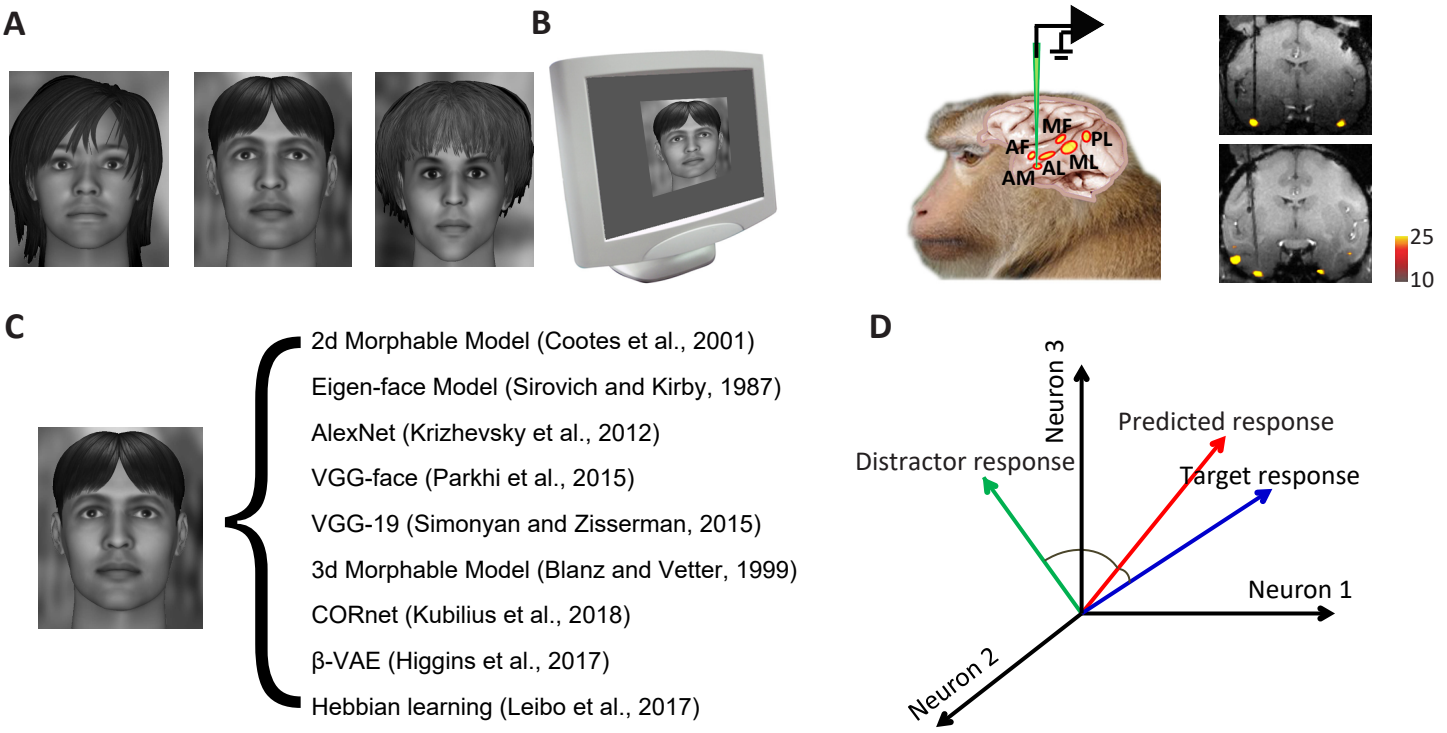
Yang, S., Luo, P., Loy, C.C., and Tang, X. (2015). From Facial Parts Responses to Face Detection: A Deep Learning Approach. IEEE International Conference on Computer Vision, 9.

Yildirim, I., Belledonne, M., Freiwald, W., and Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. Sci Adv *6*, eaax5979.

Young, A.W., and Burton, A.M. (2018). Are We Face Experts? Trends Cogn Sci *22*, 100-110.

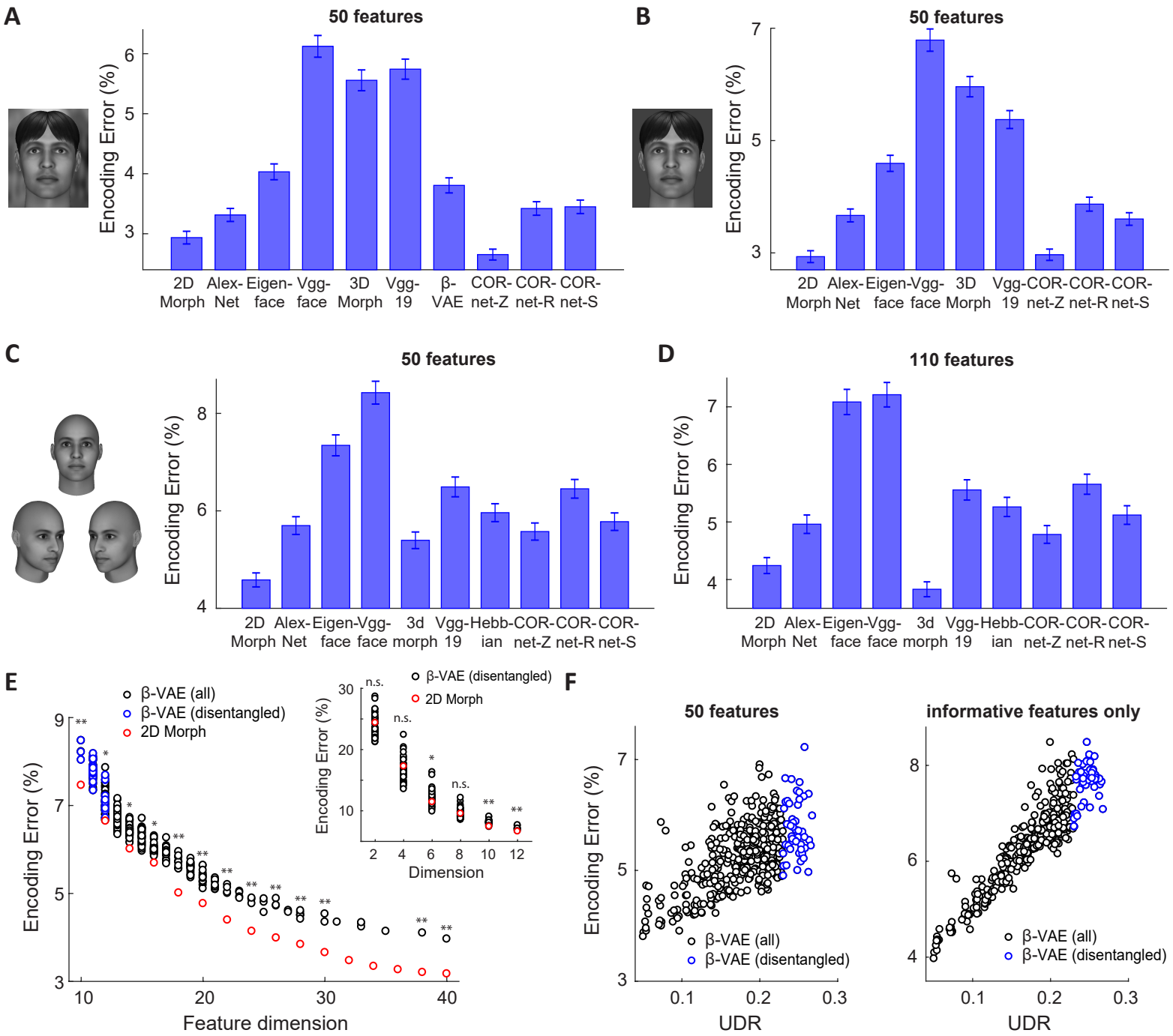
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1



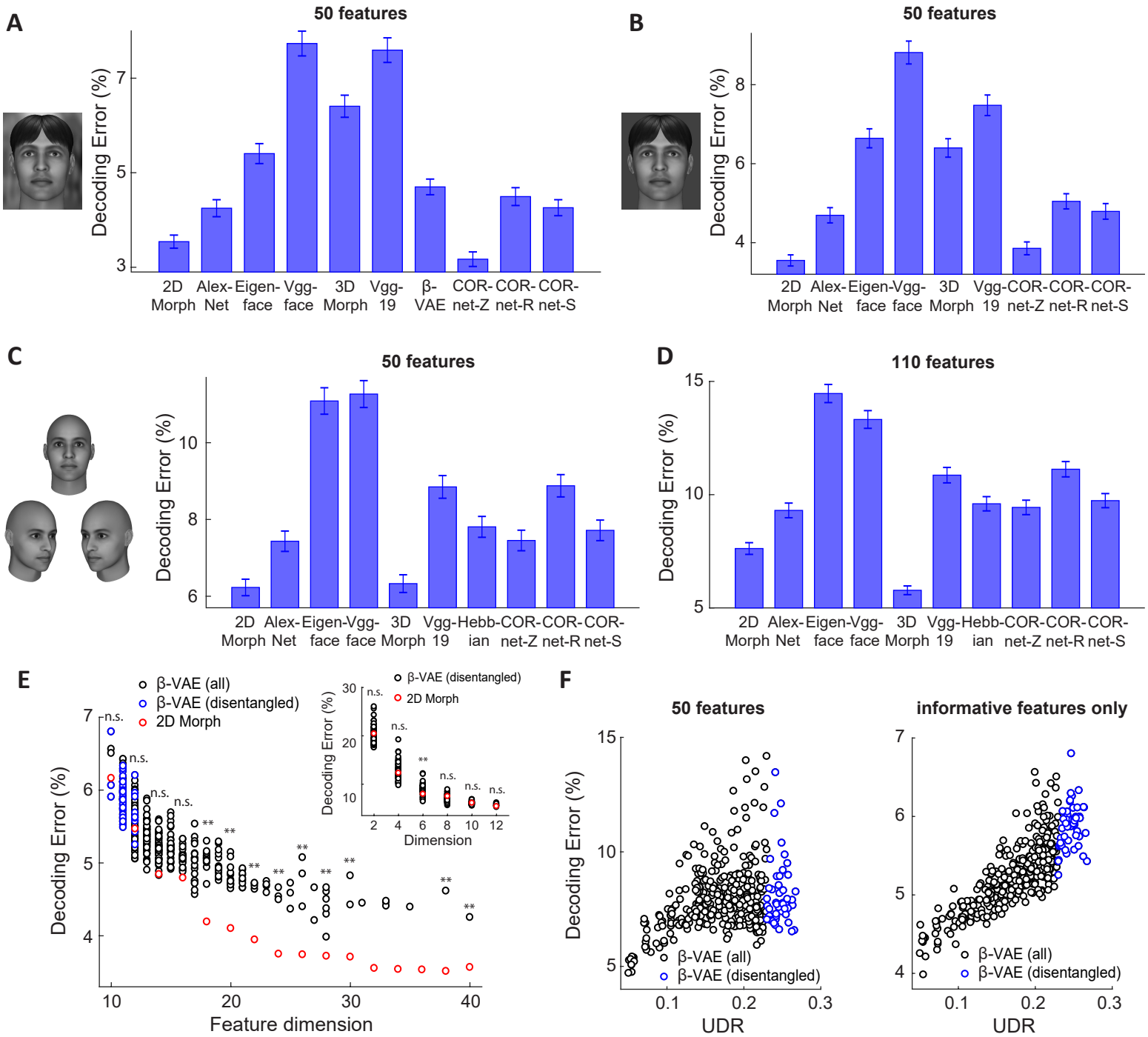
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2



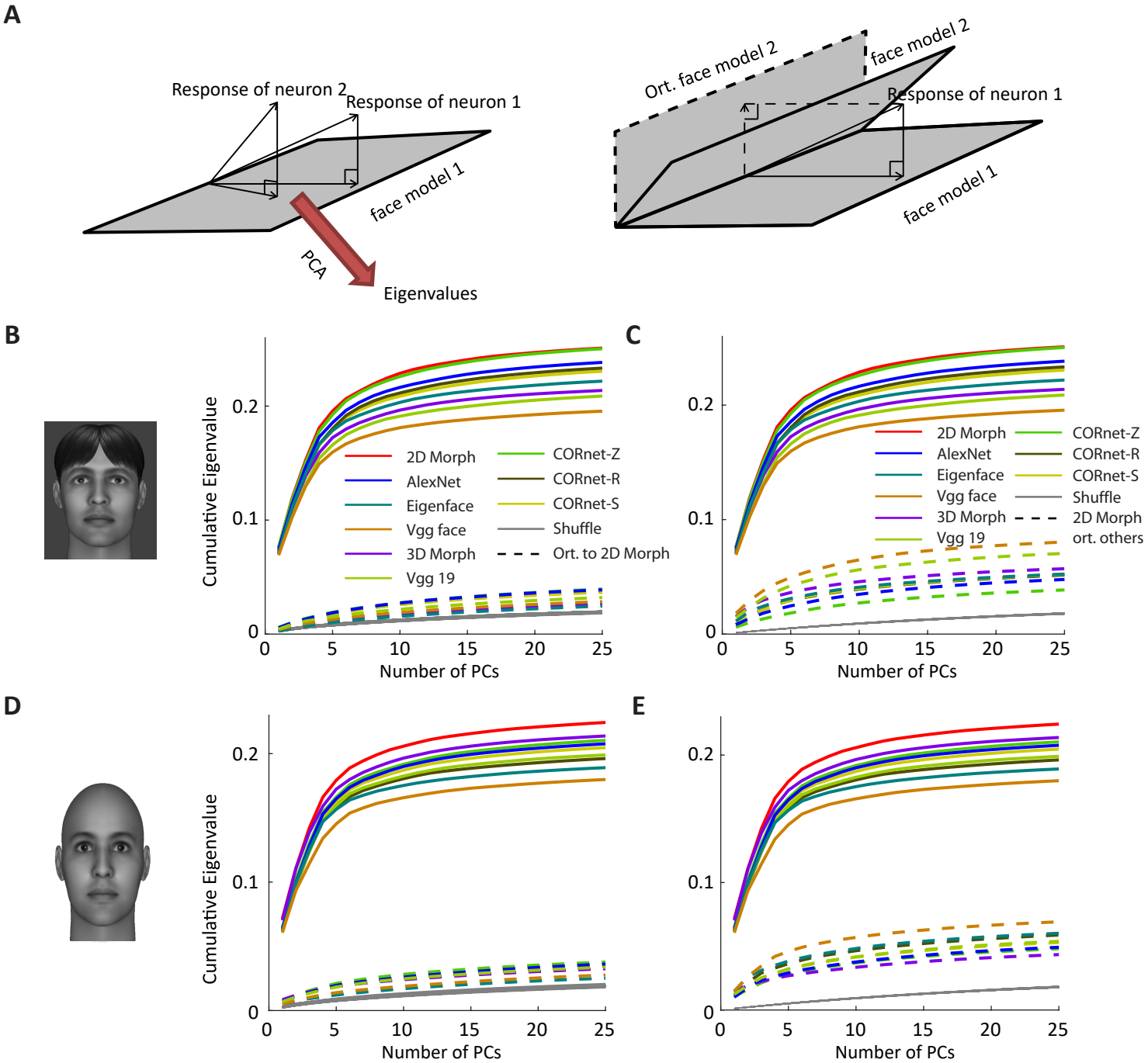
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3



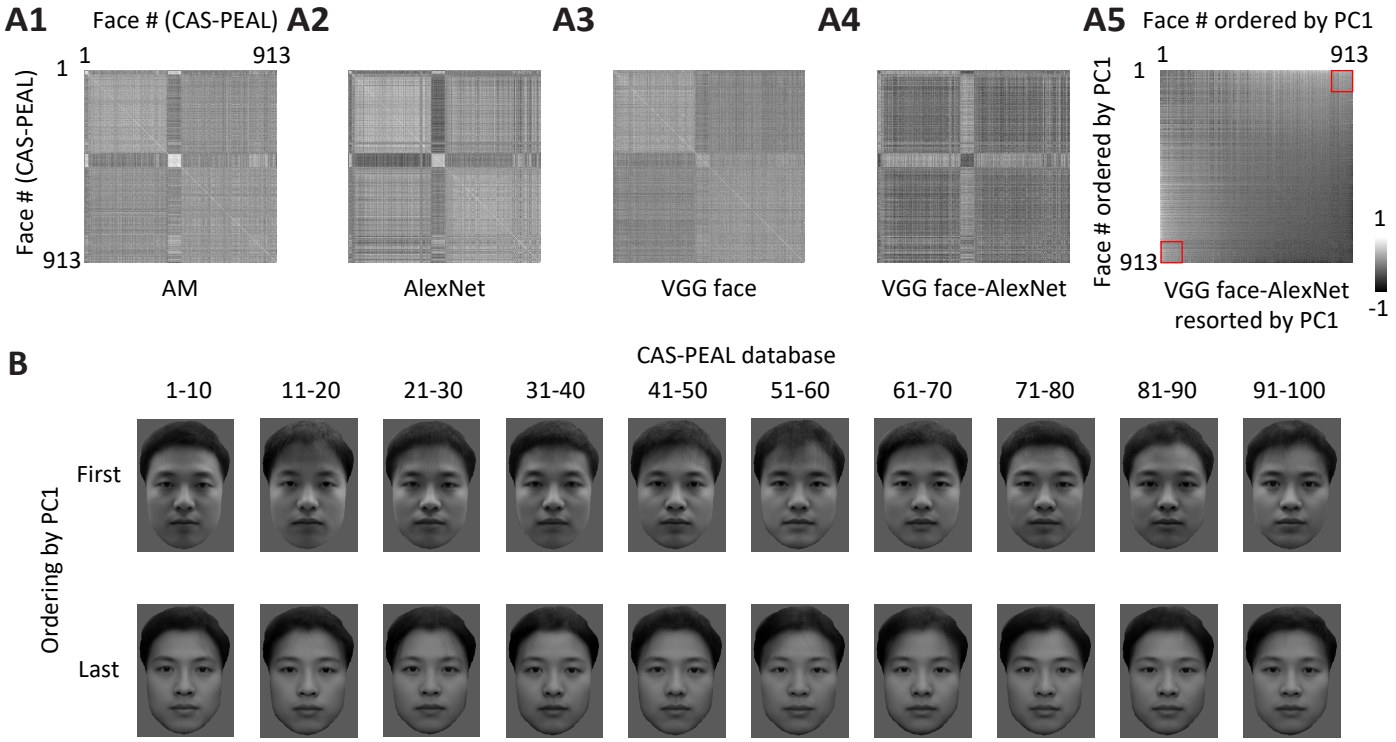
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4



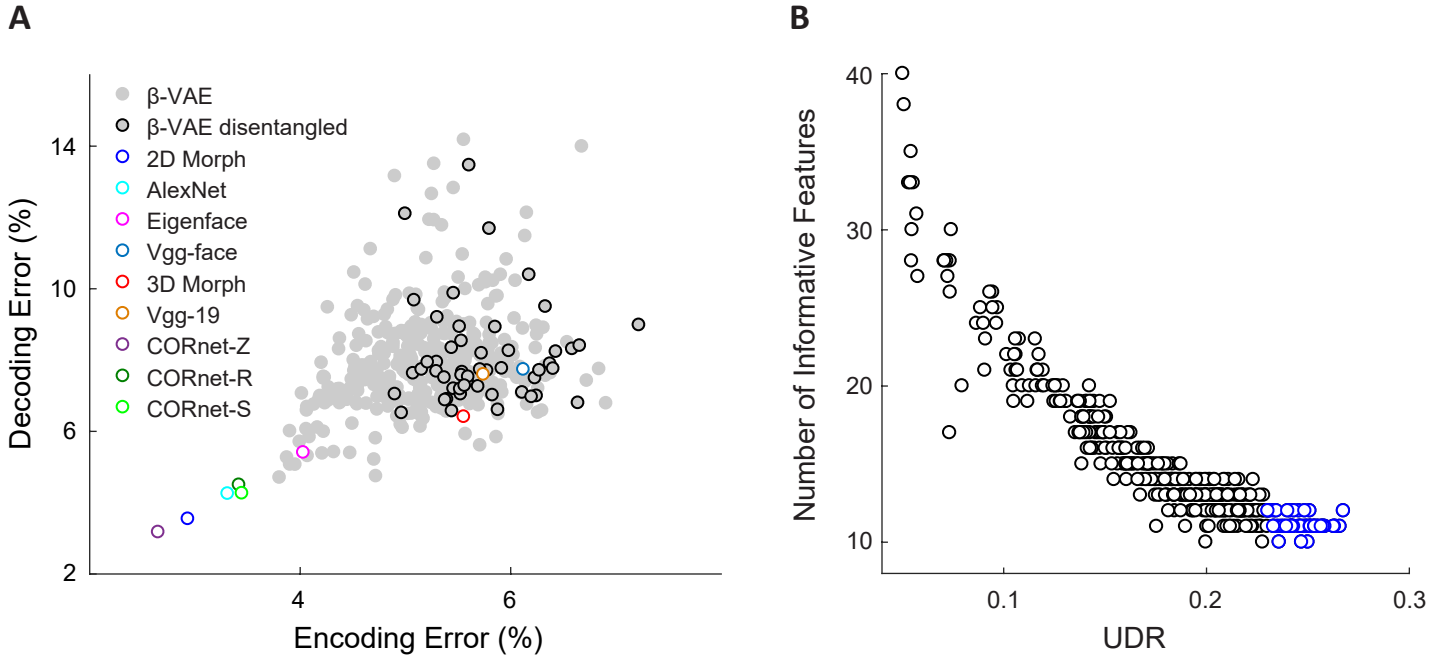
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 5



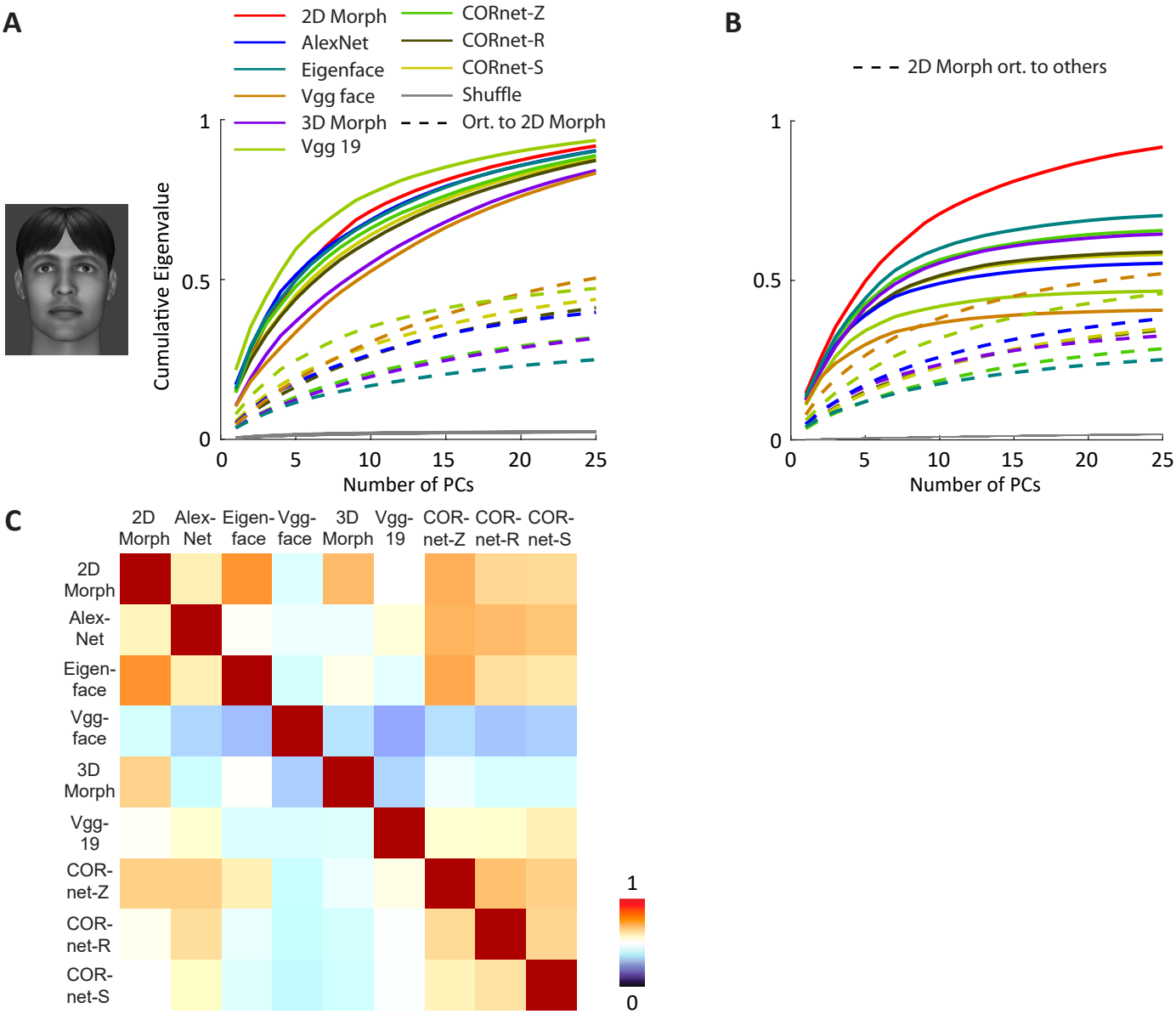
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S1



bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S2



bioRxiv preprint doi: https://doi.org/10.1101/2020.06.07.111930; this version posted June 8, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

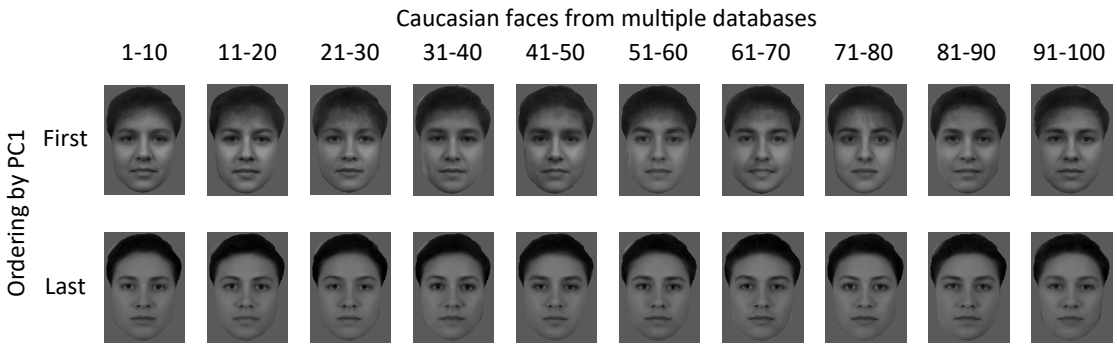



Figure S3