bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Face dissimilarity judgements are predicted by representational distance in deep neural networks and principal-component face space

Kamila M. Jozwik*[1,✉], Jonathan O'Keeffe*[2,✉], Katherine R. Storrs*[3,✉], and Nikolaus Kriegeskorte[4,✉]

[1]Department of Psychology, University of Cambridge, Cambridge, UK.
[2]Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK.
[3]Department of Experimental Psychology, Justus Liebig University, Giessen, Germany.
[4]Departments of Psychology and Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, USA.

Despite the importance of face perception in human and computer vision, no quantitative model of perceived face dissimilarity exists. We designed an efficient behavioural task to collect dissimilarity and same/different identity judgements for 232 pairs of realistic faces that densely sampled geometric relationships in a face space derived from principal components of 3D shape and texture (Basel Face Model, BFM). In a comparison of 15 models, we found that representational distances in deep neural networks (DNNs) and Euclidean distances within BFM space predicted human judgements best. A face-trained DNN explained unique variance over simpler models and was statistically indistinguishable from the noise ceiling. Sigmoidal transformation of distances improved performance for all models. Identity judgements were better predicted by Euclidean than angular or radial distances in BFM space. DNNs provide the best available image-computable models of perceived face dissimilarity. The success of BFM space suggests that human face perception is attuned to the natural distribution of faces.

face perception | face similarity | face identification | Basel Face Model | deep neural networks

Correspondence: ***jozwik.kamila@gmail.com***

Recognizing people by their faces is crucial to human social behaviour. Despite much work on the neural and behavioural signatures of face perception (see e.g. (1–4)), there is currently no quantitative model to predict how alike two faces will look to human observers. Advances in deep learning have yielded powerful artificial systems for face and object recognition (5–7), and 3D modelling and rendering techniques make it possible to systematically explore the space of possible faces (8–10). Here we investigate perceived dissimilarity among large sets of realistic faces and test a wide range of models.

Since faces of different people are structurally highly similar and vary along continuous dimensions (nose length, jaw width, etc.), it is helpful to think of faces as forming a continuous "face space" (11–13). A face space is an abstract space in which each face occupies a unique position. The dimensions of the space represent the physiognomic features that vary between faces. The origin of the multidimensional space is defined as the average face: the central tendency of the population of all faces. For an individual, this reference point is thought to reflect the sample of faces encountered in natural experience (9). Each face can then be thought of as a vector of features.

We used the Basel Face Model (BFM) (8), a widely used model in both computer graphics and face perception research (e.g. (14, 15)). The BFM is a 3D generative graphics model that produces nearly photorealistic face images from latent vectors describing shape and texture of the surfaces of natural faces. The model is based on principal components analysis (PCA) of 3D photo scans of 200 adult faces (8).

We asked to what extent the latent description of the BFM model and a range of image-computable models can predict the dissimilarity of two faces as perceived by humans. In addition, we asked to what extent the models can capture whether two images will be judged by people as depicting the same person. A model predicting identity judgements would have to capture the fact that people are able to discount various variations in appearance of the same person's face, such as the changes caused by ageing, weight fluctuations, and tanning.

Previous research has debated the relative contribution of shape vs textural information for face identification (16, 17). The BFM is defined in terms of two separate PCA spaces, one controlling the surface shape of the face via its 3D mesh, and the other controlling the texture and colouration of the face via its RGB texture map (Figure 1a). This enables us to ask whether human judgements are explained better by the coordinates within the shape or the texture subspaces in the BFM.

The distance in the BFM face space is not the only way to predict the perceived dissimilarity between two faces. We compare a range of representational models that enable us to measure the distance between the representations of any two particular face images. The models include raw pixel intensities, GIST features, 3D-mesh vertex coordinates used to render the faces, as well as computer-vision models such as HMAX and deep convolutional neural networks (DNNs). DNNs now rival human performance at visual face identification. Face-identification-trained DNNs have been shown to predict neural activity well in face-selective human brain regions recorded intracranially (18). Do these networks also capture subtle perceptual dissimilarity

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

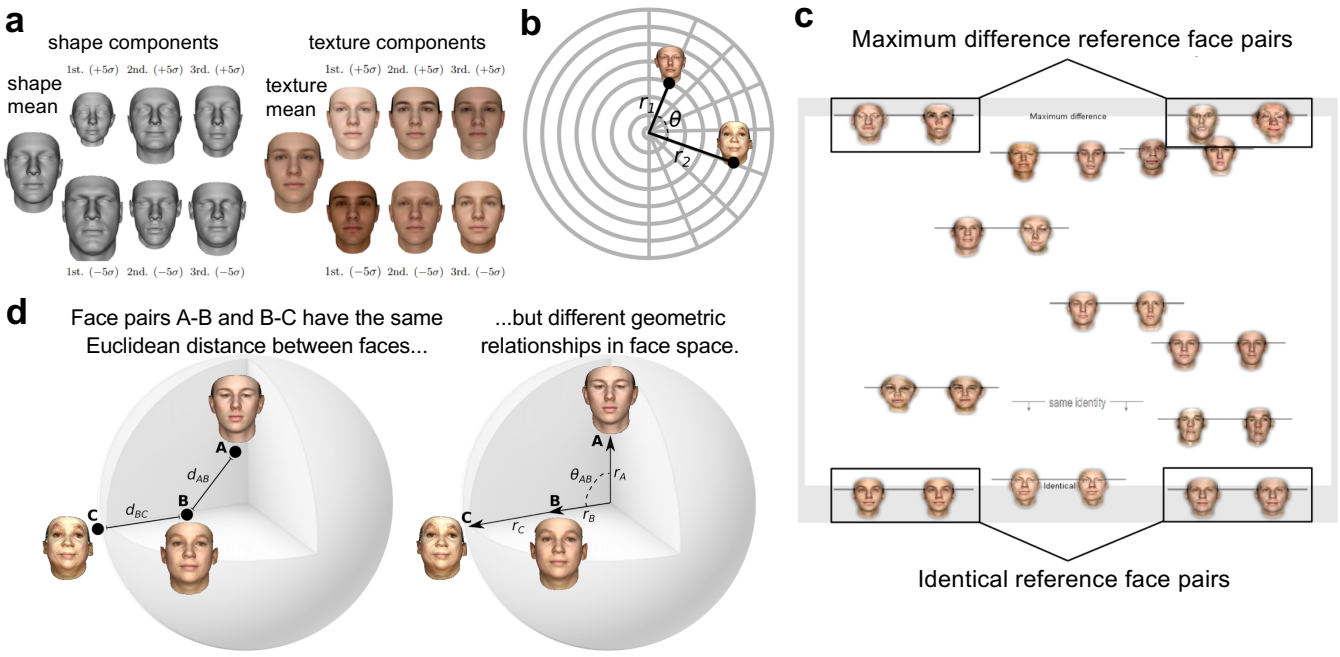

Fig. 1. Selecting pairs of faces from the BFM, and measuring perceived face dissimilarity and identity.
a) Illustration of the generative Basel Face Model (BFM), in which faces are described by separate components specifying their 3D shape (left) and texture (right). Both shape and texture components have a mean shape or texture (leftmost column), that can be changed by manipulating principal components. The first three PCs within each subspace are shown here (right three columns). Reproduced with permission from (8).
b) Stimulus selection. We defined stimuli as pairs of vectors in the BFM with radial lengths r_1 and r_2, and angle between them (θ). We sampled all unique combinations from 8 θ values and 8 radius values to obtain 232 face pairs. See Methods for details.
c) Behavioural experiment method. Participants positioned the face pairs along the vertical axis of the screen according to their relative dissimilarity. Faces were arranged in random subsets of eight pairs, as in the example shown. The vertical position of the line linking face pairs determined their precise location. As points of reference for participants, two example face pairs depicting "Maximum difference" (top) and two example face pairs depicting "Identical" (bottom) were shown. For each subset, participants also placed a "same identity" bar indicating the point below which pairs of images appear to depict the same identity.
d) Relationships between pairs of faces in a face space like the BFM can be thought of in terms of the Euclidean distance between them (left), or their geometric relationships relative to the centre of the face space (right). If perceived dissimilarity can be predicted from the Euclidean distance alone, then face pair AB should look exactly as similar to one another as face pair BC. However, if observers take the angular and radial geometry of face space into account, they may have substantially different similarities.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

relationships? Is the ranking of face pairs in these networks similar to that in human judgements? Does the visual diet of these networks affect their ability to capture human dissimilarity judgements?

To gain more insight into how humans judge face dissimilarity and identity, we designed a novel task for efficiently obtaining high-fidelity dissimilarity and identity judgements. Previous studies used multi-arrangement task for individual images of objects (19) but not pairs of images. We selected stimulus pairs to systematically sample geometric relations in statistical face space, exhaustively creating all combinations from a large set of facial vector lengths and angles (Figure 1b and 2a), see Methods). This experimental design allowed us to carefully test how well the BFM's statistical space captured human judgements, and what the geometric relationship was between distances in the BFM and distances in human perceptual space. During the task, participants arranged pairs of face images on a large touch-screen according to how similar they appeared, relative to anchoring face pairs at the top and bottom of the screen and relative to other adjusted pairs (Figure 1c). This task yielded a superior measure to standard dissimilarity ratings in three ways: 1) it produced a fine-grained continuous measure of face dissimilarity within each pair (the vertical position at which the pair was placed on the screen), 2) it was efficient (many pairs could be placed within a single trial) and 3) it was robust, since judgements are anchored relative to multiple visual references simultaneously (both the extreme anchor pairs provided above and below the sorting arena and the other adjustable pairs within each trial). Participants also placed a horizontal bar on each trial that indicated the separation between face pairs that appeared to depict different individuals, and pairs that appeared to depict different instances of "the same person". We sought to model both the continuous aspects of human face perception (graded dissimilarity) and its categorical aspects (same/different identity).

Results

Participants (N=26) were highly reliable in their dissimilarity judgements using the novel arrangement task (mean correlation between participants = 0.80, mean correlation for the same participant between sessions = 0.85, stimulus set A experiment), providing a high-quality dataset with which to adjudicate between candidate models. We repeated the same experiment with a subset of the same participants (N=15) six months later, with a new independently sampled face set fulfilling the same geometric relations as the original stimulus set (stimulus set B experiment, see Methods). Participants in the stimulus set B experiment were also highly reliable in their dissimilarity judgements (mean correlation between participants = 0.79). This level of replicability allowed us to evaluate to what extent dissimilarity judgements depend on idiosyncrasies of individual faces, and to what extent they can be predicted from geometric relations within a statistical face space.

Face dissimilarity judgements can be well predicted by distance in a statistical face space. We first asked how well human face dissimilarity judgements could be predicted by distances within the Basel Face Model (BFM), the principal-components face space from which our stimuli had been generated. Since we had selected face pairs to exhaustively sample different geometric relationships within the BFM, defined in terms of the angle between faces and the radial distance of each face from the origin, we were able to visualise human dissimilarity ratings in terms of these geometric features (Figure 2b). The human ratings bore a strong resemblance to the patterns of the Euclidean distances among our stimuli (Figure 2a). Given this, we plotted dissimilarity judgements for each face pair as a function of the Euclidean distance in the BFM. To test how well the BFM approximates face dissimilarity judgements we tested functions that best described the relationship between the behavioural dissimilarity judgements and the BFM. We plotted the predictions of each fitted model over the data and compared the models. If the BFM is a perfect approximator of face dissimilarity judgements a linear function would best describe the relationship between face dissimilarity judgements and the Euclidean distances in the BFM (Figure 2c). We do not find this assumption to be completely true as the sigmoidal function better describes the relationship between face dissimilarity judgements and the Euclidean distances in the BFM (Figure 2c). The sigmoidal relationship between the BFM and perceived distances suggests that observers have maximal sensitivity to differences between faces occupying moderately distant points in the statistical face space, at the expense of failing to differentiate between different levels of dissimilarity among very nearby or very far apart faces. This latter result may be related to the fact that faces with very large Euclidean distances in the BFM look slightly caricatured to humans. We observed similar results in the stimulus set B experiment (face dissimilarity judgements using different face pairs with the same geometrical properties as in the stimulus set A experiment, see Methods for details, Figure 2c). This result suggests that the sigmoidal relationship between the BFM and perceived distances is observed regardless of face pairs sampled. Overall, the BFM is a good, but not perfect, approximator of face dissimilarity judgements.

Face identity judgements can be well predicted from the Euclidean distance in BFM. We also asked humans to judge whether each pair of faces depicted the same or different identity, and examined human identity thresholds in relation to the Euclidean distance between faces in the BFM. We found that moderately dissimilar faces are often still perceived as having the same identity (Figure 3a, Figure 4). We observed similar results in the stimulus set B experiment (Figure 3c, Figure 4). The examples of face pairs judged as the same identity are shown in Figure 4. This result may be related to humans having a high tolerance to changes in a personal appearance due to age, weight fluctuations, or skin complexion depending on the season.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

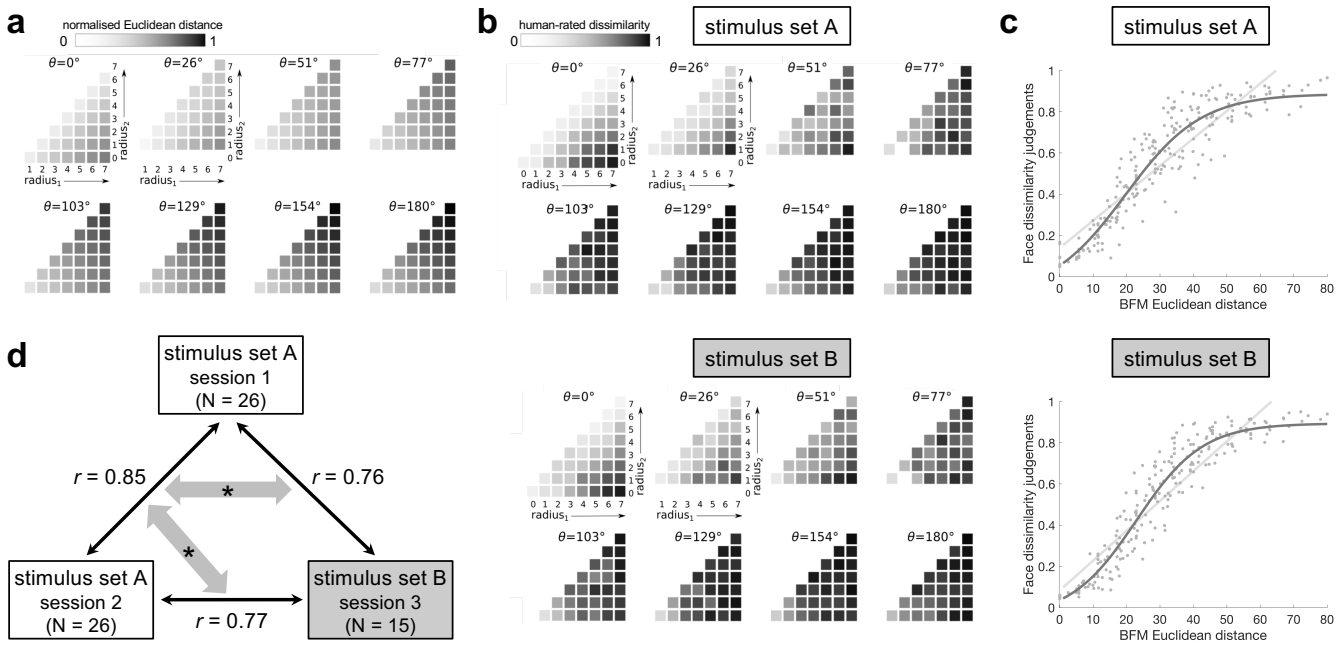

Fig. 2. Face dissimilarity judgements as a function of distances in the BFM.
a) Euclidean distances within the BFM for each face-pair in the stimulus sets. Sets A and B had identical BFM geometries, but used different specific face exemplars. Each grid shows pairs of face vectors separated by a specific angle (θ), and arranged by the lengths of each of the two vectors ($radius_1$ and $radius_2$). Radii were sampled in eight evenly-spaced steps, where step 0 = a length of zero, and step 7 = a length of 40 units in the BFM. Normalised Euclidean distance is indicated by darkness, from white (zero) to black (maximum distance within the stimulus set).
b) Human face dissimilarity judgements for each face pair, for stimulus set A (top) and B (bottom). Plotting conventions are as in 2a, except that darkness indicates human-rated dissimilarity, from white (identical faces) to black (maximally dissimilar faces), averaged over participants and trials.
c) Face dissimilarity judgements (y-axis) as a function of Euclidean distance in the BFM (x-axis) for stimulus set A (top) and B (bottom). Each dot represents the mean dissimilarity rating for one pair of faces, averaged across participants and trials. The pale grey line represents the fit of a linear function and the dark grey line represents the fit of a sigmoidal function to the data.
d) Replicability of face dissimilarity judgements between sessions 1 and 2 (using stimulus set A and the same participants), and session 3 (using stimulus set B and a subset of the participant group). Values are Pearson correlation coefficients, averaged across participants. Grey arrows with asterisks indicate significantly different correlations (two-sample t-test, p<0.05).

Face pairs in the BFM can be analyzed in terms of their geometric characteristics relative to the centre of the face space or as the Euclidean distance between them. Therefore, we tested alternative predictors of face identity judgements: geometry in the BFM (θ, absolute difference between r_1 and r_2) and the Euclidean distance. We could predict whether two faces will be classified as the same individual by each of the predictors, however, the Euclidean distance in the BFM predicted identity judgements better than the angular and radial geometry of face space.

Relative geometry within BFM is approximately but not exactly perceptually isotropic. As face pairs used in the stimulus set A and stimulus set B experiments had the same relative geometries but different face exemplars we could test whether the relative geometry within BFM is perceptually isotropic. To address that, we tested face dissimilarity judgements replicability by correlating average judgements across participants in stimulus sets A and B. Participants completed two sessions of the stimulus set A experiment and a subset of participants (15 out of 26) completed the third session of the stimulus set B experiment. If the relative geometry within BFM is isotropic then the correlation between stimulus set A and B experiments should be the same as the correlation between two sessions of the stimulus set A experiment. The correlation between two sessions of the stimulus set A experiment is 0.85, the correlation between stimulus set A session 1 and stimulus set B session 3 experiment is 0.76, and the correlation between stimulus set A session 2 and stimulus set B session 3 experiment is 0.77 (Figure 2d). These results suggest that the relative geometry within BFM is approximately, but not exactly, perceptually isotropic and we do not have strong evidence against isotropy. The stimulus set B experiment was performed 6 months after the stimulus set A experiment therefore the differences in correlations between session could be attributed to the longer time between sessions with different face exemplars.

Face dissimilarity judgements can be well predicted by a DNN trained on either faces or objects. We measured the dissimilarity of face representations within each pair in the activation space of 16-layer deep neural networks (DNNs) of the VGG-16 architecture (20), trained on either face identification (21) or on object categorisation (20). Both networks were implemented in the Matconvnet toolbox for Matlab and were pretrained on their respective tasks by the original authors.
To gain an intuitive understanding of the face pair arrangement performed by humans and DNNs, we visualised the mean ranking of face pairs from the face pairs perceived as maximally different to the ones that were perceived as the same (plotting

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

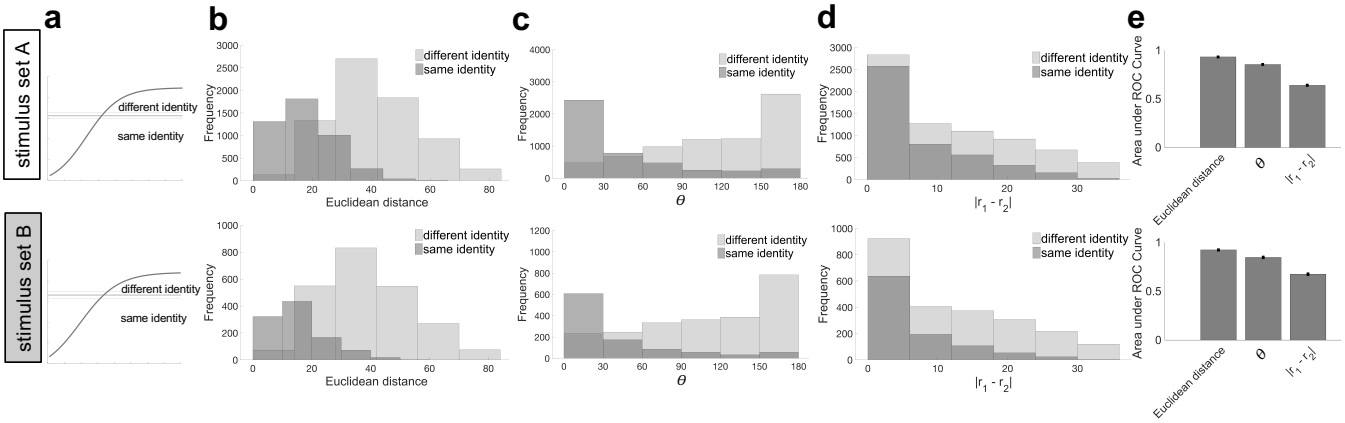

Fig. 3. Identity judgements as a function of geometry within the BFM.
a) Threshold for judging faces as belonging to the same/different identity, visualised relative to similarity judgements. Curved line shows the sigmoidal fit to face dissimilarity judgements (from Figure 2c) in stimulus set A (top) and B (bottom). Thick horizontal line shows the mean placement of the "different identity" threshold bar, across participants and trials; thinner lines above and below indicate the standard error of the mean over participants.
b) Histogram of how frequently face-pairs were judged as having the same identity (dark grey) or different identity (light grey), as a function of their Euclidean distance in the BFM.
c) Histogram of same and different identity judgements as a function of angle (θ) between faces in the BFM.
d) Histogram of same and different identity judgements as a function of the absolute difference between vector lengths in BFM (r_1 and r_2).
e) Summary of how well each of the three BFM metrics in b-d discriminates face pairs judged as having the same vs different identity. Bars show area under the receiver operating characteristic (ROC) curve calculated based on identity judgements using Euclidean distance, θ, and absolute difference between r_1 and r_2.

every 20 face pairs for clarity, see Figure 4 for the stimulus set A and B experiments).

To determine how face dissimilarity judgements performed by humans related to the dissimilarity of face image features from one of the current best models of face perception (VGG-Face), we displayed the dissimilarity ratings in VGG-Face alongside human face dissimilarity judgements. Visual inspection suggests that VGG-Face arranged face pairs somewhat differently than humans do. To quantify this, we computed the Pearson correlation between ranks of each subject dissimilarity judgements and VGG-Face. We also computed the Pearson correlation between ranks of each subject dissimilarity judgements and the mean dissimilarity judgements. The mean correlation between VGG-Face rankings and each human face dissimilarity rankings was 0.40 (Figure 4). Is the correlation between face pair arrangement across participants and VGG arrangement lower or higher than the correlation of each subject dissimilarity judgements and the mean dissimilarity judgements? We found that the former is the case as the correlation between each subject dissimilarity judgements rank and the mean dissimilarity judgements was 0.63. We observed similar results in the stimulus set B experiment (Figure 4) with the mean correlation between VGG-Face rankings and each subject face dissimilarity rankings being 0.35 and the correlation between each subject dissimilarity judgements rank and mean dissimilarity judgements rank being 0.60 (Figure 4). This result suggests that the VGG approximates the ranks of face pairs to a certain extent, but not fully, considering the between-subject variability.

Configural information and high-level person characteristics poorly predict perceived face dissimilarity. After establishing that face dissimilarity judgements can be predicted from the Euclidean distance relatively well and that VGG-Face can approximate face dissimilarity ranks in humans to a similar level as another human face dissimilarity ratings, we wanted to test a wide range of models to examine whether there is one model that best explains face dissimilarity judgements or there are multiple models that can explain the data equally well.

All models tested are schematically presented in Figure 5a. Several models were based on the BFM: BFM shape dimensions, BFM texture dimensions, full BFM (texture and shape dimensions together), one-dimensional projections in BFM onto which height, weight, age and gender were loaded most strongly, and angle in BFM between two face vectors. Alternative models consisted of a 3D mesh model, RGB pixels, GIST, and face configurations ("0th order" configuration (location of 30 key points such as eyes, nose, mouth), "1st order" configuration (distances between key points), and "2nd order" configuration (ratios of distances between key points)). Finally, the last class of models consisted of DNNs (VGG-16 architecture) trained on either object recognition or face identity recognition.

We inferentially compared each model's ability to predict face dissimilarity judgements, in both their raw state and after fitting a sigmoidal transform to model-predicted dissimilarities, using a procedure cross-validated over both participants and stimuli (see Methods). The highest-performing model was the VGG deep neural network trained on face identification, which was the only model to predict human responses as well as the responses from other participants (i.e. not significantly below the noise ceiling; Figure 5b). However, several other models had high performances that were not statistically different from that of VGG-Face: Euclidean distance in BFM shape subspace, GIST, an object-trained Alexnet DNN, full BFM space, and BFM texture subspace (Figure 5b, top). Therefore there is no one best model, but several different models are equally good at explaining face dissimilarity data. Performing the same analysis on the independent stimulus set B experiment revealed good

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

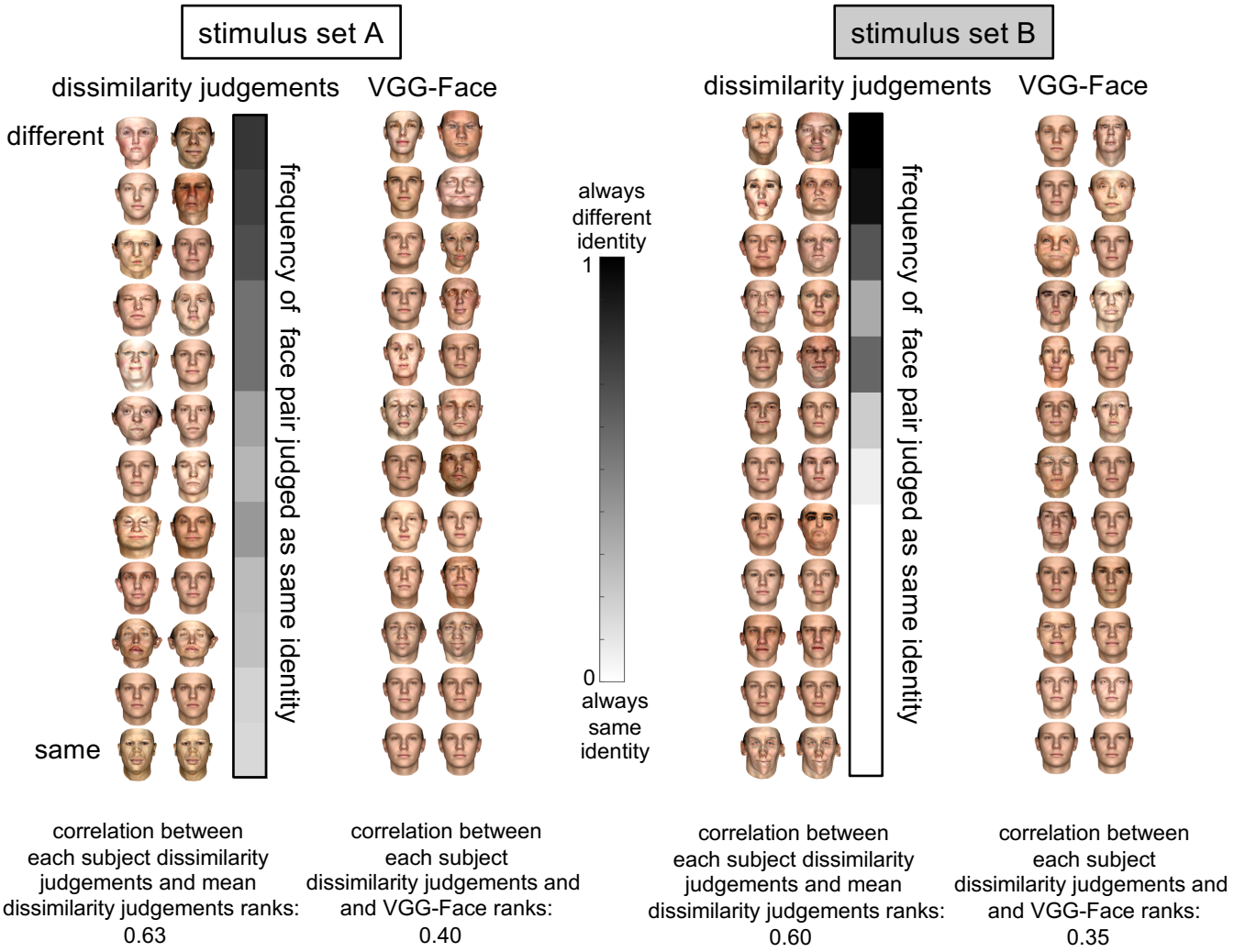

Fig. 4. Face dissimilarity rankings by humans, and by a deep neural network trained on face identity recognition (VGG-Face).
Within stimulus sets A (left) and B (right), leftmost column displays face pairs according to their average rated dissimilarity by humans from most dissimilar (top) to most similar (bottom), visualising every 20th face pair in each ranked set. Rightmost column shows the same visualisation of rankings, performed based on the Euclidean distance between representations in a DNN trained on face recognition (VGG-Face). Colorbars to the right of human dissimilarity rankings show the frequency with which participants classified the corresponding face pair as depicting the same identity, from white (0; faces were always judged as having the same identity) to black (1; faces were always judged as having different identities).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reproducibility of the model rankings, even though the individual faces are different (Figure 5b, bottom). VGG-Face again achieved the highest performance, but was not significantly superior to several other models: VGG-Object, Alexnet, GIST, full BFM space, or the shape or texture subspaces of BFM. Most models reached the noise ceiling in this second dataset, but this is likely because there was greater overall measurement noise, due to a smaller sample size and one rather than two experimental sessions.

There are substantial computational differences between the several models that all predict human perceived face dissimilarity well. Do they explain shared or unique variance in human judgements? To address this question we performed a unique variance analysis on all models. Several models explained a significant amount of unique variance, with VGG-Face explaining the most unique variance in both stimulus set A and stimulus set B experiments (Figure 5c). If some models explain unique variance, perhaps combining them would explain more overall variance in face dissimilarity judgements? To address this question, we combined all models into one model via linear weighting, and asked whether this combined model explains more variance than each of the models alone. Model weights were assigned within the same procedure individual models were evaluated, cross-validating over both participants and stimuli. We found that in both datasets, the combined weighted model reached high performance, but did not exceed the performance of the best individual model (Figure 5b).

Models based on BFM or DNN feature spaces outperformed most others, including models based on the face perception literature (angle in 'face space', and configurations of facial features) and two 'baseline' models (based on pixels or 3D face meshes). The success of the V1-like GIST model is surprising and may be due not to its unique explanatory power, but its high shared variance with more complex models for the image set used, although it is consistent with previous work finding that Gabor-based models explain variance in face matching experiments (22) and explain almost all variance in the face- and other complex shape-matching experiments when stimuli are tightly controlled (23). A person-attributes model, consisting only of the four dimensions which capture the highest variance (among the scanned individuals) in height, weight, age, and gender, did not perform well. This finding may seem surprising given that an earlier systematic attempt to predict face dissimilarity judgements from image-computable features found that dissimilarity was best predicted by weighted combinations of features that approximated natural high-level dimensions of personal characteristics such as age and weight (24). However, it seems that people use other or more than socially relevant dimensions when judging face dissimilarity in the experiment presented here.

Some may find it surprising that VGG trained on faces did not perform better than VGG trained on objects (as elaborated on in the discussion section). For both VGG trained on faces and VGG trained on objects late intermediate layers explained most variance in face dissimilarity ratings in the stimulus set A experiment and the stimulus set B experiment (Figure 6a). Our finding is consistent with a previous study that showed late and intermediate layers of object-trained VGG explaining more variance in object similarity judgements (25). Late intermediate layers were also the only layers that explained unique variance (up to 0.3% of total variance) in both the stimulus set A and stimulus set B experiments where we compared each model individually (Figure 6b). These results suggest that similar stages of processing are important for explaining both total and unique variance by VGG-Object and VGG-Face.

All models better predicted human responses after fitting a sigmoidal function to their raw predicted distances, and produced a greater relative improvement for more poorly-performing models, but did not substantially affect model rankings (Figure 5b).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

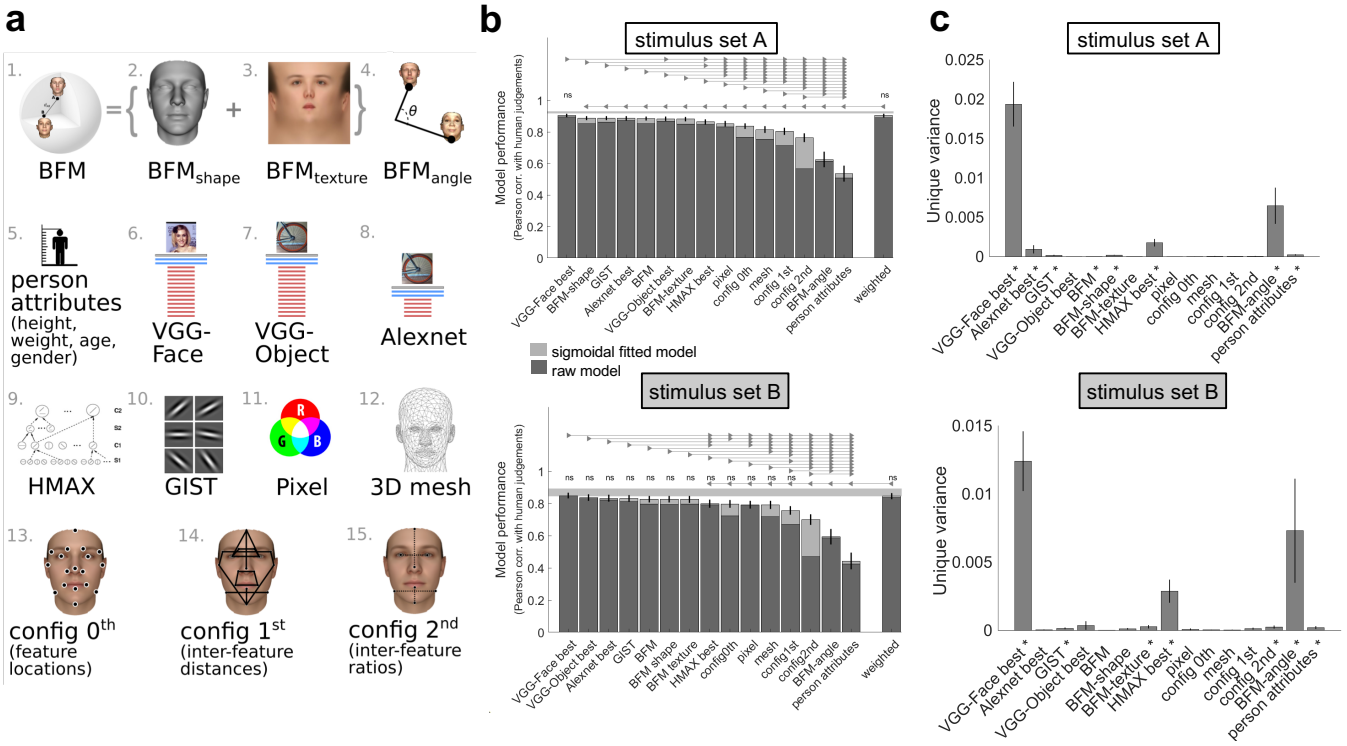

Fig. 5. Comparing diverse models in their ability to predict face dissimilarity judgements.
a) Schematic illustration of models compared (see Methods Table 1 for full details). Several models were based on the morphable model from which faces were generated: 1. Euclidean distance within the full BFM coordinate space; 2. distance in the BFM's shape subspace; 3. in the BFM's texture subspace; 4. angle between faces in the full BFM; and 5. distance within the height, weight, age and gender dimensions of the BFM. Deep neural network (DNN) models consisted of a 16 layer VGG architecture trained on either (6.) faces or (7.) objects, and (8.) an 8-layer Alexnet architecture trained on objects. Alternative models were: 9. a shallower HMAX neural network; 10. GIST image descriptors; 11. raw pixel values; 12. raw 3D face mesh; and configural models: 13. "0th order" configuration (location of 30 key points such as eyes, nose, mouth); 14. "1st order" configuration (distances between key points); and "2nd order" configuration (ratios of distances between key points).
b) Ability of each model to predict face dissimilarity judgements in stimulus set A (top) and B (bottom). Bars show Pearson correlation between human-judged face dissimilarity and face-pair distance within each model. The dark lower region of each bar shows performance for raw model distances, while the paler upper region shows additional performance gained if model distances are transformed by a compressive nonlinearity (a sigmoidal function fitted to data from training participants and face-pair stimuli). All models were significantly correlated with human data ($p < 0.05$ corrected). The grey bar represents the noise ceiling, which indicates the expected performance of the true model given the noise in the data. The final bar shows the performance of a linear weighted combinations of all models, fitted using non-negative least-squares. Fitting of sigmoidal transforms and linear reweighting was performed within the same cross-validation procedure, fitting and evaluating on separate pools of both participants and stimuli. Error bars show the standard error of the mean (95% confidence interval over 1,000 bootstrap samples). Horizontal lines show pairwise differences between model performance ($p < 0.05$, Bonferroni corrected across all comparisons). Models connected by triangular arrow markers indicate a significant difference, following the convention in (26), with the direction of the arrow marker showing which model is superior. All statistical tests shown were performed on the sigmoidally-transformed version of each model, since these had the highest performance in all cases.
c) Unique variance in face dissimilarity judgements computed using a hierarchical general linear model (GLM) for stimulus set A (top) and B (bottom). For each model, the unique variance is computed by subtracting the total variance explained by the reduced GLM (excluding the model of interest) from the total variance explained by the full GLM, using non-negative least squares to find optimal weights. Models that explain significant unique variance are indicated by an asterisk (one-sided Wilcoxon signed-rank test, $p < 0.05$ corrected). Error bars show the standard error of the mean based on single-subject unique variance.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

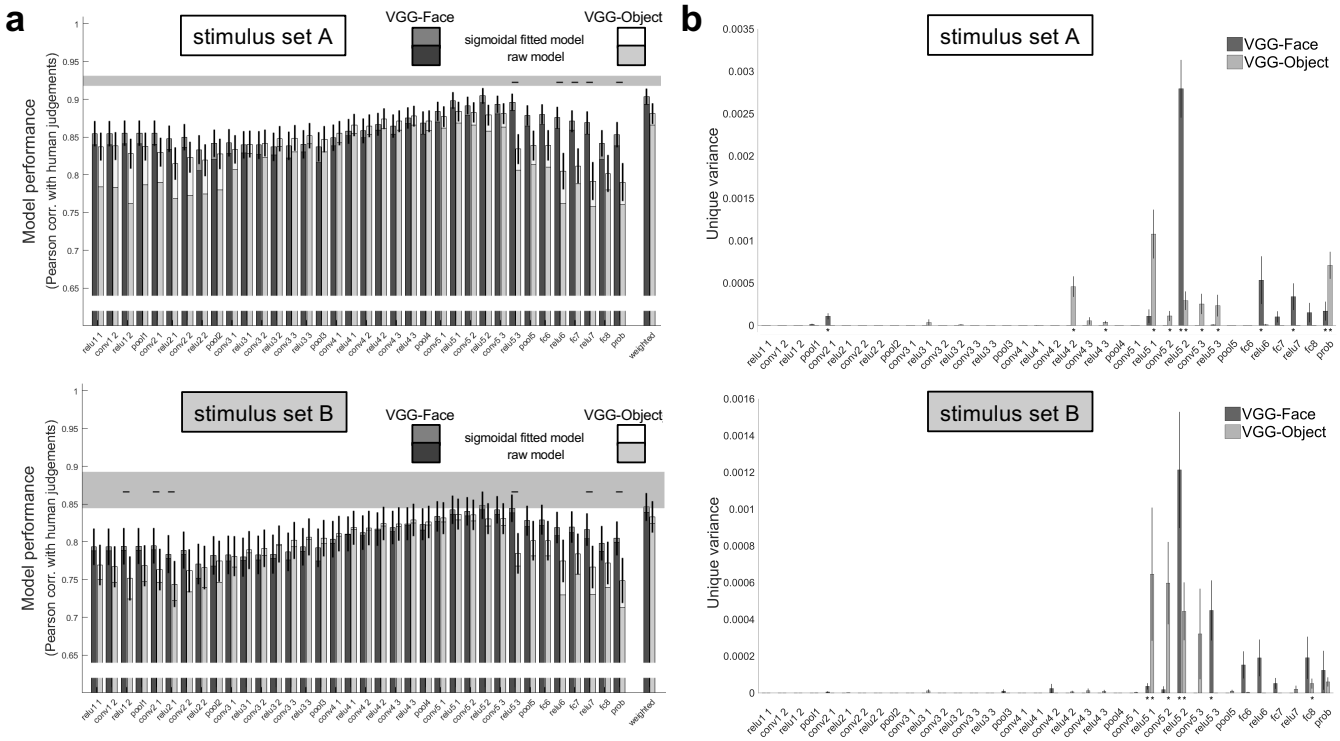

Fig. 6. Ability of each layer within VGG-Face and VGG-Object neural networks to predict human face dissimilarity judgements.
a) Correlations between human dissimilarity judgements in stimulus set A (top) and B (bottom) and Euclidean distance within each layer of the same deep neural network trained either to recognise faces (VGG-Face) or objects (VGG-Object). All key processing steps within each network are shown, including application of a non-linearity ('relu'), convolutional layers ('conv'), max-pooling ('pool'), and fully-connected layers ('fc'). The darker lower part of each bar shows performance of raw predicted distances, and paler upper parts show the same after fitting a sigmoidal transform, cross-validated over both participants and stimuli. The final bars show the performance of a linearly-weighted combination of all layers. Conventions are as in Figure 5b.
b) Unique variance in face dissimilarity judgements explained by each layer of each network, for stimulus set A (top) and in the stimulus set B experiment (bottom) using the same conventions as in Figure 5c).

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Discussion

We have shown that the Euclidean distance in the BFM is a good approximator of human dissimilarity judgements. To our knowledge, this is the first time the BFM has been validated as providing quantitative predictions of perceived face dissimilarity and identity. The BFM was previously shown to capture face impressions (27) and personality traits (28). The BFM's statistical face model is derived from separate principal components analyses of 3D face structure and of facial texture and colouration. It is, therefore, a more sophisticated statistical model than earlier PCA-based face space models derived from 2D images, which only moderately well predict face dissimilarity (29). In our study, BFM has a dual role of being a good model and stimulus generator.

The success of the Euclidean distance alone to predict both dissimilarity and identity is striking, given that psychological face space accounts have assigned particular importance to the geometric relationships of faces relative to a meaningful origin of the space (2, 11, 12, 30, 31). For example, it has been reported that there are larger perceptual differences between faces that span the average face than not (32). Extensive behavioural and neuropsychological work has sought to relate the computational mechanisms underlying face perception to geometric relationships in neural or psychological face space. It has been proposed that face-selective neurons explicitly encode faces in terms of vectors of deviation from an average face, based on evidence from monkey electrophysiology (33, 34) and human psychophysical adaptation (2) although alternative interpretations of the latter have been made (35, 36). Our comprehensive sampling of face pairs with the full range of possible geometric relationships was tailored to reveal the precise manner in which distances from the origin, and angular distances between faces, affect perceived dissimilarity. Yet both dissimilarity and identity data were explained best by simply the Euclidean distance, with geometric relationships in face space accounting for no additional variance. The BFM angle model did explain unique variance, however, the amount of variance explained was not enough to explain additional variance when combined with the BFM Euclidean distance model. Our results do not contradict previous studies, but suggest that effects of relative geometry may be more subtle than previously thought, when probed with large sets of faces that vary along diverse dimensions, rather than stimulus sets constructed to densely sample single or few dimensions (e.g. (31, 32)). Lastly, distances within BFM appear approximately but not exactly perceptually isotropic, as face dissimilarity judgements with different face exemplars but the same Euclidean distances and relative geometries were highly correlated, but less so than dissimilarity judgements with the same exemplars. One confounding factor here is however that the stimulus set B experiment was performed 6 months after the stimulus set A experiment, therefore, the differences in correlations between sessions could be attributed to the longer time between sessions with different face exemplars.

Distance within the BFM is not a perfect predictor of perceived dissimilarity. Firstly, like all morphable models, the BFM describes only the physical structure of faces, and so cannot account for effects of viewing factors such as pose and illumination, nor of familiarity, which we know to be substantial (cf. (30)). Relatedly, we did not explore people's ability to parse structural differences between faces from sources of accidental differences between face images that are important for face recognition (lighting (37, 38) and viewpoint(37, 39)). It is hard to predict how differences in lighting or viewpoint would affect the performance of the models but it may further differentiate image-based models (e.g. GIST) from BFM that is invariant and capture the distance between faces in principal-components space. DNNs lie on a continuum between image-dependent and invariant models as they learn partly invariant features during training. It is non-trivial that within the domain of highly controlled frontal well-lit viewing conditions tested here, the BFM better predicts perceived dissimilarity than other structure-only models, such as geometric mesh dissimilarity. Secondly, the BFM distance is imperfect as a dissimilarity predictor in that there remains unexplained variance that is reliable across individual observers but not captured by any BFM model. There are several possible reasons for this. The BFM has limitations as a morphable model, for example, it is based on the head scans of only 200 individuals, and this sample is biased in several ways, for example towards white, relatively young, faces. The sub-optimal performance could also be due to fundamental limitations shared by any physically-based model (30), such as its inability to capture perceptual inhomogeneities relating to psychologically relevant distinctions such as gender, ethnicity, or familiarity. It would be interesting to test in the future whether newer morphable face space models capture more of the remaining variance in human dissimilarity judgements (9). The task presented here provides an efficient way to test the perceptual validity of future face space models.

We found that humans often classify pairs of images as depicting the same identity even with relatively large distances in the BFM. Two faces may be perceptibly different from one another, while nevertheless appearing to be "the same person." The ability of the visual system to generalise identity across some degree of structural difference may be analogous to invariance to position, size and pose in object recognition (40). Face images generated from a single identity form a complex manifold as they may vary in age, weight, expression, makeup, facial hair, skin tone colour, and more. Given that we need to robustly recognise identity despite changes in these factors, it may not be surprising that there is a high tolerance for facial features when we judge one's identity. The stimulus set contained very dissimilar faces, which provided an anchor for people's definition of "different" and may influence moderately-dissimilar faces to look quite alike, in comparison. Participants seemed to interpret person identity quite generously, possibly imagining whether this face could be the same person if they aged, got

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tanned, or lost weight. "The same person" may be a not precisely defined concept, however people seem to agree what that concept means as they were consistent in judging the same/different identity boundary. Interestingly, the "different identity" boundary was close to the saturation point of face dissimilarity psychometric function. This result could be related to people dismissing all "different individuals" as completely-different and focusing their fine gradations of dissimilarity only within the range of faces that could depict the same identity. In our current experiment, the identity and face dissimilarity judgements are entangled and future experiments are needed to dissociate them.

Our data show clearly that some models of face dissimilarity are worse than others. Simply taking the angle between faces in the BFM is a poor predictor, as is a set of higher-order ratios between facial features. Perhaps surprisingly, the model consisting only of the four dimensions which capture the highest variance in height, weight, age, and gender performed poorly. Age and gender were shown to explain variance in face MEG representations (41) and we show that they do explain variance in face dissimilarity judgements task, however to a lesser extent than better-performing models. It seems that people use other or more than socially relevant dimensions when judging face dissimilarity.

Among highly performing models, we found that several explain face dissimilarity judgements similarly well. One of the models that explains a surprisingly large amount of variance is GIST. It has been previously shown, that Gabour-based models explain face representations well (14, 22). The models compared contain quite different feature spaces. For example, object-trained and face-trained VGG models learn distinctly different feature detectors (6), yet explain a similar amount of variance in human face dissimilarity judgements. Both object-trained and face-trained VGG models also explain a similar amount of variance in human inferior temporal cortex (42), and object-trained VGG explains variance in early MEG responses (43). The "face space" within a face-trained DNN organises faces differently than they are arranged in the BFM's principal components, for example, clustering low-quality images at the "origin" of the space, eliciting lower activity from all learned features (30).
It is perhaps remarkable that distances within the BFM are approximately as good at capturing perceived face dissimilarities as image-computable DNNs. Distances within the BFM contain no information about either the specific individuals concerned, or the image-level differences between the two rendered exemplars. DNNs, on the other hand, are image-computable and thus capture differences between the visible features in the specific rendered images seen by participants. The high success of the relatively impoverished BFM representation may highlight the importance of statistical face distributions to human face perception. After all, the BFM simply describes the statistical dissimilarity between two faces, expressed in units of standard deviations within the sample of 200 head-scanned individuals. The power of this statistical description is consistent with previous evidence for the adaptability of face representations, coming from face aftereffects (2, 34), the "own-race effect" (44, 45), and inversion and prototype effects (46)

We expected that a 16-layer VGG deep neural network trained to discriminate faces would likely be a better model than the same architecture trained on objects. However, this was not the case. There are a couple of possible explanations for this surprising result. The first is that although there is no "human" category in the ILSVRC dataset on which the VGG-object network was trained (7), there are images with faces in them (e.g. the classes "T-shirt" and "bowtie"), some classes have more than 90% images with faces (e.g. the classes "volleyball" and "military uniform"), and as many as 17% of all images in the dataset have at least one face (47). Therefore, the network may still have learned facial features, even without being explicitly trained on face discrimination. Indeed, DNNs trained with a spatial correlation loss in addition to a classification objective, developed "face patches" when trained on the same ILSVRC dataset (48). Even when all faces were completely removed from the training set, a face-deprived DNN was still able to categorize and discriminate faces (49). However, the face-deprived training affected the degree of DNN's face selectivity and DNN's ability to replicate the face-inversion effect(49). Another possibility is that human face dissimilarity judgements are based on general-purpose descriptions of high-level image structure, which are not specific to faces. Consistent with our behavioural results, VGG trained on faces was not beneficial in explaining neural recordings of faces (50). These results are consistent with our finding that a general object recognition model seems to be sufficient to develop features that are diagnostic of face dissimilarity. Emerging classes of models, such as inverse graphics model based on DNNs (10), could be tested as additional model candidates in the future.

One of the reasons for the equally high performance of disparate models is that, for our stimulus set, several models made highly correlated predictions, making it difficult to discriminate between them based on the current data. Model dissociation was also found to be difficult when studying the representations of face dissimilarity in human fusiform face area, where Gabor filters model performed similarly to face space sigmoidal ramp-tuning model (14). Stimulus optimisation methods could be used in the future to identify sets of stimuli for which current well-performing candidate models make maximally dissimilar predictions (26, 51, 52).

We conclude that deep neural networks provide the best available models of perceived facial similarity, and indeed a face-trained DNN fully explains human judgements within two independent datasets. Meanwhile, the more moderate success of a

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

principal-components face space emphasises the importance of the natural distribution of faces in human face perception.

Methods

Stimuli. Each face generated by the BFM corresponds to a unique point in the model's 398-dimensional space (199 shape dimensions, and 199 texture dimensions), centred on the average face. The relative locations of any pair of faces can therefore be summarised by three values: the length of the vector from the origin to the first face r_1, the length of the vector from the origin to the second face r_2, and the angle between the two face vectors θ (see Figure 1b). To create a set of face pairs spanning a wide range of relative geometries in face space, we systematically sampled all pairs of 8 possible vector length values (29 unique combinations) combined with 8 possible angular values. Possible angular values were eight uniform steps between 0 and 180 degrees, and possible vector lengths were eight uniform steps between 0 and 80 units in the BFM. This yielded 232 unique relative geometries. For each relative geometry, we then sampled two random points in the full 398-dimensional BFM space that satisfied the given geometric constraints. We generated two separate sets of face pairs with the same relative geometries but different face exemplars, by sampling two independent sets of points satisfying the same geometric constraints. The two sets (stimulus set A and stimulus set B) were used as stimuli in separate experimental sessions (see "Psychophysical face pair-arrangement task").

Participants. Human behavioural testing took place over three sessions. Twenty-six participants (13 female) took part in sessions 1 and 2, and a subset of 15 (6 female) took part in session 3. All testing was approved by the MRC Cognition and Brain Sciences Ethics Committee and was conducted in accordance with the Declaration of Helsinki. Volunteers gave informed consent and were reimbursed for their time. All participants had normal or corrected-to-normal vision.

Psychophysical face pair-arrangement task. The procedure in all sessions was identical, the only difference being that the same set of face pair stimuli was used in sessions 1 and 2, while session 3 used a second sampled set with identical geometric properties. Comparing the consistency between sessions 1, 2, and 3 allowed us to gauge how strongly human judgements were determined by geometric relationships in face space, irrespective of the individual face exemplars.

During an experimental session, participants were seated at a comfortable distance in front of a large touch-screen computer display (43" Panasonic TH-43LFE8-IR, resolution 1920x1080 pixels). On each trial, the participant saw a large white "arena", with a randomly arranged pile of eight face pairs in a grey region to the right-hand side (see Figure 1c). The two faces within each pair were joined together by a thin bar placed behind the faces, and each pair could be dragged around the touch-screen display by touching. Each face image was rendered in colour with a transparent background and a height of 144 pixels (approximately 7.1cm on screen).

The bottom edge of the white arena was labelled "Identical" and the top edge was labelled "Maximum difference". Two example face pairs were placed to the left and to the right of the "Identical" and "Maximum difference" labels to give participants reference points on what identical and maximally different faces look like. The maximally different example faces had the largest geometric distance possible within the experimentally sampled geometric relationships (i.e. the Euclidean distance in the BFM = 80) in contrast to identical faces (i.e. the Euclidean distance in the BFM = 0). The same example pairs were used for all trials and participants.

Participants were instructed to arrange the eight face pairs on each trial vertically, according to the dissimilarity of the two faces within the pair. For example, two identical faces should be placed at the very bottom of the screen. Two faces that look as different as faces can look from one another should be placed at the very top of the screen. Participants were instructed that only the vertical positioning of faces would be taken into account (horizontal space was provided so that face pairs could be more easily viewed, and so that face pairs perceived as being equally similar could be placed at the same vertical location). On each trial, once the participant had finished vertically arranging face pairs by dissimilarity, they were asked to drag an "identity line" (see Figure 1c) on the screen to indicate the point below which they considered image pairs to depict "the same person". Once eight face pairs and the identity line were placed, participants pressed the "Done" button to move to the next trial. Each session consisted of 29 trials.

Representational similarity analysis. We used representational similarity analysis (RSA) to evaluate how well each of a set of candidate models predicted human facial (dis)similarity judgements (53). For every model, a model-predicted dissimilarity was obtained by computing the distance between the two faces in each stimulus pair, within the model's feature space, using the model's distance metric (see "Candidate models of face dissimilarity"). Model performance was defined as the Pearson correlation between human dissimilarity judgements and the dissimilarities predicted by the model. We evaluated the ability to predict human data both for each individual model and for a linearly weighted combination of all models. To provide an estimate of the upper bound of explainable variance in the dataset, we calculated how well human data could be predicted by data from other participants, providing a "noise ceiling".

Noise ceilings, raw model performance, sigmoidally-transformed model performance, and reweighted combined model performance were all calculated within a single procedure, cross-validating over both participants and stimuli (54). On each of

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20 cross-validation folds, 5 participants and 46 face pairs were randomly assigned as test data, and the remaining stimuli and participants were used as training data. On each fold, a sigmoidally-transformed version of each model was created, by fitting a logistic function to best predict dissimilarities for training stimuli, averaged over training participants, from raw model distances. Also on each fold, a reweighted combined model was created using non-negative least-squares to assign one positive weight to each of the individual models, to best predict the dissimilarity ratings for training stimuli, averaged over training participants. We then calculated, for each raw model, each sigmoidally transformed model, and for the combined reweighted model, the Pearson correlation with the model's predictions for test stimuli for each individual test subject's ratings. The average correlation over test participants constituted that model's performance on this cross-validation fold. The upper bound of the noise ceiling was calculated within the same fold by correlating each test subject's test-stimulus data with the average test-stimulus data of all test participants (including their own). The lower bound was calculated by correlating each test subject's test-stimulus data with the average for all training subject's test-stimulus data (54, 55). Means and confidence intervals were obtained by bootstrapping the entire cross-validation procedure 1,000 times over both participants and stimuli.

We first determined whether each model was significantly different from the lower bound of the noise ceiling, by assessing whether the 95% confidence interval of the bootstrap distribution of differences between model and noise ceiling contained zero (54, 55), Bonferroni corrected for the number of models. Models that are not significantly different from the lower bound of the noise ceiling can be considered as explaining all explainable variance, given the noise and individual differences in the data. We subsequently tested for differences between the performance of different models. We defined a significant pairwise model comparison likewise as one in which the 95% confidence interval of the bootstrapped difference distribution did not contain zero, Bonferroni corrected for the number of pairwise comparisons.

Unique variance analysis. We used a hierarchical general linear model (GLM) to evaluate unique variance explained by best performing models (56). For each model, the unique variance was computed by subtracting the total variance explained by the reduced GLM (excluding the model of interest) from the total variance explained by the full GLM. We performed this procedure for each participant and used non-negative least squares to find optimal weights. A constant term was included in the GLM model. We performed a one-sided Wilcoxon signed-rank test to evaluate the significance of the unique variance contributed by each model across participants.

Candidate models of face dissimilarity. We considered a total of 15 models of face dissimilarity, each consisting of a set of features derived from the face image, the BFM coordinates, or 3D mesh, and a distance metric (see Table 1).

Basel Face Model. We considered four variant models based on the principal-component space provided by the BFM: (1) "BFM Euclidean" took the Euclidean distances between faces in the full 398-dimensional BFM space; (2) "BFM-shape" took the Euclidean distances only within the 199 components describing variations in the 3D shape of faces; (3) "BFM-texture" took the Euclidean distances only within the 199 separate components describing variations in the RGB texture maps that provided faces' pigmentation and features; and (4) "BFM angle" which took the cosine distance between face vectors in the full 398-dimensional space. For face pairs where cosine distance was undefined, because one face lay at the origin of BFM space, the angle between the two faces was defined as zero for the purposes of model evaluation.

To more fully explore the relationship between apparent dissimilarity and placements of faces in the full BFM space, we also considered linear and sigmoidal functions as candidates for predicting the relationship between the Euclidean distance in the BFM and face dissimilarity judgements. We estimated each model's predictive performance as the Pearson correlation between the fitted model's predicted dissimilarities and the dissimilarities recorded by the subject. We tested for significant differences between linear and sigmoidal function fits using a two-sided Wilcoxon signed-rank test. For each subject, we fitted the model to half of the data (session 1) and measured the predictive accuracy of the model in the second half of the data (session 2). The predictive accuracies were averaged across participants.

Person attributes. The BFM provides the axes onto which the height, weight, age, and gender of the 3D scanned participants most strongly loads. By projecting new face points onto these axes, we can approximately measure the height, weight, age and gender of each generated face. The "Person attributes" model took the Euclidean distance between faces, after projecting faces onto these four dimensions.

Models based on 3D face structure. Face perception is widely thought to depend on spatial relationships among facial features (4, 17, 60, 61). We calculated the Euclidean distance between the 3D meshes that were used to render each face ("Mesh" model). We also used the geometric information within each face's mesh description to calculate a first, second, and third-order configural model of facial feature arrangements, following suggestions by (60) and others (e.g. (17)) that face perception depends more strongly on distances or ratios of distances between facial features than raw feature locations. We selected 30 vertices on each face corresponding to key locations such as the centre and edges of each eye, the edges of the mouth, nose, jaw, chin, and hairline (see schematic in Figure 5a), using data provided in the BFM. The positions of these 30 vertices on each 3D face mesh formed the features for the "0th order" configural model. We then calculated 19 distances between horizontal

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Candidate models				
Model name	Description	Reference	Distance metric	Computed from
1. BFM	Morphable face space combining PCA subspaces of structural and textural components from 200 3D face scans.	(8)	Euclidean	BFM PCA coordinates
2. BFM shape	PCA subspace of only the structural components from 200 3D face scans.	(8)	Euclidean	BFM PCA coordinates
3. BFM texture	PCA subspace of only the textural components from 200 3D face scans.	(8)	Euclidean	BFM PCA coordinates
4. BFM angle	PCA subspace of only the structural components from 200 3D face scans.	(8)	Cosine	BFM PCA coordinates
5. Person attributes	Loading of PCA coordinates onto height, weight, age and gender vectors.	(8)	Euclidean	BFM PCA coordinates
6. VGG-Face best	Highest-performing layer of 16-layer deep neural network trained on face identification.	(20)	Euclidean	RGB image
7. VGG-Object best	Highest-performing layer of 16-layer deep neural network trained on object recognition.	(21)	Euclidean	RGB image
8. AlexNet best	Highest-performing layer of 8-layer deep neural network trained on object recognition.	(57)	Euclidean	RGB image
9. HMAX best	Highest-performing layer in a 4-layer cortically-inspired neural network.	(58)	Euclidean	RGB image
10. GIST	Gabor-based summary of contrast energy at different orientations and scales.	(59)	Euclidean	RGB image
11. Pixel	Raw image data.	n/a	Euclidean	RGB image
12. Mesh	Raw 3D mesh data.	n/a	Euclidean	3D mesh
13. Config 0th	Locations of key facial features (0th order configural information).	(60)	Euclidean	3D mesh
14. Config 1st	Distances between key facial features (1st order configural information).	(60)	Euclidean	3D mesh
15. Config 2nd	Ratios of distances between key facial features (2nd order configural information).	(60)	Euclidean	3D mesh

Table 1. Candidate models of face dissimilarity.

and vertically aligned features (e.g. width of nose, length of nose, separation of eyes), which formed the "1st order" configural model. Finally, we calculated 19 ratios among these distances (e.g. ratio of eye separation to eye height; ratio of nose width to nose length), which formed the "2nd order" configural model.

Deep neural networks. We used a state-of-the-art 16-layer convolutional neural network (VGG-16), trained on millions of images to recognize either object classes (20) or facial identities (21). The dissimilarity predicted by DNN models was defined as the Euclidean distance between activation patterns elicited by each image in a face pair in a single layer. To input to DNN models, faces were rendered at the VGG network input size of 124x124 pixels, on a white background, and preprocessed to subtract the average pixel value of the network's training image set.

Low-level image-computable models. As control models, we also considered the dissimilarity of two faces within in terms of several low-level image descriptors: (1) Euclidean distance in raw RGB pixel space; (2)

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Euclidean distance within a "GIST" descriptor, image structure at four spatial scales and eight orientations (https://people.csail.mit.edu/torralba/code/spatialenvelope/); and (3) HMAX, a simple four-layer neural network (http://cbcl.mit.edu/jmutch/hmin/). For comparability with the images seen by participants, all low-level image-computable models operated on faces rendered on a white background at 144x144 pixel resolution.

ACKNOWLEDGEMENTS

This research was supported by the Wellcome Trust [grant number 206521/Z/17/Z] awarded to KMJ; the Alexander von Humboldt Foundation postdoctoral fellowship awarded to KMJ; the Alexander von Humboldt Foundation postdoctoral fellowship awarded to KRS; the Wellcome Trust and the MRC Cognition and Brain Sciences Unit. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

COMPETING FINANCIAL INTERESTS

The authors declare that they have no competing interests.

AUTHOR CONTRIBUTIONS

JOK, KMJ and NK designed the experiments. KMJ collected the data. KMJ, JOK and KRS performed the analyses. KMJ and KRS wrote the paper. All authors edited the paper. NK supervised the work.

DATA AND CODE AVAILABILITY

The datasets and code generated during the current study are available from the corresponding author on reasonable request.

References

1. Nancy Kanwisher. Domain specificity in face perception. *Nature neuroscience*, 3(8):759–763, 2000.
2. Gillian Rhodes and David A Leopold. Adaptive norm-based coding of face identity. *The Oxford handbook of face perception*, pages 263–286, 2011.
3. Doris Y Tsao and Margaret S Livingstone. Mechanisms of face perception. *Annu. Rev. Neurosci.*, 31:411–437, 2008.
4. Martha J Farah, Kevin D Wilson, Maxwell Drain, and James N Tanaka. What is" special" about face perception? *Psychological review*, 105(3):482, 1998.
5. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.
6. Katherine R. Storrs and Nikolaus Kriegeskorte. Deep learning for cognitive neuroscience. In David Poeppel, George R. Mangun, and Michael S. Gazzaniga, editors, *The Cognitive Neurosciences*. MIT Press, 2020.
7. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
8. Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee, 2009.
9. Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models – past, present and future. *arXiv*, 2020.
10. Ilker Yildirim, Mario Belledonne, Winrich Freiwald, and Josh Tenenbaum. Efficient inverse graphics in biological face processing. *Science Advances*, 6(10), 2020.
11. Tim Valentine. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A*, 43(2):161–204, 1991.
12. Tim Valentine, Michael B Lewis, and Peter J Hills. Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology*, 69(10):1996–2019, 2016.
13. A Mike Burton and John R Vokey. The face-space typicality paradox: Understanding the face-space metaphor. *The Quarterly Journal of Experimental Psychology Section A*, 51(3):475–483, 1998.
14. Johan D. Carlin and Nikolaus Kriegeskorte. Adjudicating between face-coding models with individual-face fMRI responses. *PLOS Computational Biology*, 13(7):e1005604, 2017.
15. Neil MT Houlsby, Ferenc Huszár, Mohammad M Ghassemi, Gergő Orbán, Daniel M Wolpert, and Máté Lengyel. Cognitive tomography reveals complex, task-independent mental representations. *Current Biology*, 23(21):2169–2175, 2013.
16. Richard Russell, Irving Biederman, Marissa Nederhouser, and Pawan Sinha. The utility of surface reflectance for the recognition of upright and inverted faces. *Vision research*, 47(2):157–165, 2007.
17. Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94 (11):1948–1962, 2006.
18. Shany Grossman, Guy Gaziv, Erin M Yeagle, Michal Harel, Pierre Mégevand, David M Groppe, Simon Khuvis, Jose L Herrero, Michal Irani, Ashesh D Mehta, et al. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature communications*, 10(1):1–13, 2019.
19. Marieke Mur, Mirjam Meys, Jerzy Bodurka, Rainer Goebel, Peter a. Bandettini, and Nikolaus Kriegeskorte. Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, 4:1–22, 2013.
20. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, 2014.
21. Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. *arXiv*, 2015.
22. Irving Biederman and Peter Kalocsais. Neurocomputational bases of object and face recognition. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352 (1358):1203–1219.
23. Xiaomin Yue, Irving Biederman, Michael C Mangini, Christoph von der Malsburg, and Ori Amir. Predicting the psychophysical similarity of faces and non-face complex shapes by image-based measures. *Vision research*, 55:41–46, 2012.
24. Mark Steyvers and Tom Busey. Predicting similarity ratings to faces using physical descriptions. *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges*, pages 115–146, 2000.
25. Kamila M. Jozwik, Nikolaus Kriegeskorte, Katherine R. Storrs, and Marieke Mur. Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, 8:1726, 2017.
26. Tal Golan, Prashant C Raju, and Nikolaus Kriegeskorte. Controversial stimuli: pitting neural networks against each other as models of human recognition. *arXiv*, 2019.
27. Ryan M. Stolier, Eric Hehman, Matthias D. Keller, Mirella Walker, and Jonathan B. Freeman. The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences*, 115 (37):9210–9215, 2018.
28. Selma Carolin Rudert, Leonie Reutner, Rainer Greifeneder, and Mirella Walker. Faced with exclusion: Perceived facial warmth and competence influence moral judgments of social exclusion. *Journal of Experimental Social Psychology*, 68:101–112, 2017.
29. Rainer Scheuchenpflug. Predicting face similarity judgements with a computational model of face space. *Acta psychologica*, 100(3):229–242, 1999.
30. Alice J O'Toole, Carlos D Castillo, Connor J Parde, Matthew Q Hill, and Rama Chellappa. Face space representations in deep convolutional neural networks. *Trends in cognitive sciences*, 22(9): 794–809, 2018.
31. Kieran Lee, Graham Byatt, and Gillian Rhodes. Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological science*, 11(5):379–385, 2000.
32. Volker Blanz, Al ice J O'toole, Thomas Vetter, and Heather A Wild. On the other side of the mean: The perception of dissimilarity in human faces. *Perception*, 29(8):885–891, 2000.
33. Le Chang and Doris Y. Tsao. The Code for Facial Identity in the Primate Brain. *Cell*, 169(6):1013–1028.e14, 2017.
34. David A Leopold, Igor V Bondar, and Martin A Giese. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, 442(7102):572–575, 2006.
35. David A Ross, Mickael Deroche, and Thomas J Palmeri. Not just the norm: Exemplar-based models also predict face aftereffects. *Psychonomic bulletin & review*, 21(1):47–70, 2014.
36. Katherine R Storrs and Derek H Arnold. Not all face aftereffects are equal. *Vision research*, 64:7–16, 2012.
37. H Hill. Information and viewpoint dependence in face recognition. *Cognition*, 62(2):201–222, 1997.
38. Chang Hong Liu, Charles A Collin, A.Mike Burton, and Avi Chaudhuri. Lighting direction affects recognition of untextured faces in photographic positive and negative. *Vision Research*, 39(24): 4003–4009, 1999.
39. Vicki Bruce. Changing faces: Visual and non-visual coding processes in face recognition. *British Journal of Psychology*, 73(1):105–116, 1982.
40. James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.
41. Katharina Dobs, Leyla Isik, Dimitrios Pantazis, and Nancy Kanwisher. How face perception unfolds over time. *Nature Communications*, 10(1):1258, 2019.
42. Kamila Maria Jozwik, Martin Schrimpf, Nancy Kanwisher, and James J. DiCarlo. To find better neural network models of human vision, find better neural network models of primate vision. *bioRxiv*, 2019.
43. Kamila Maria Jozwik, Michael Lee, Tiago Marques, Martin Schrimpf, and Pouya Bashivan. Large-scale hyperparameter search for predicting human brain responses in the Algonauts challenge. *bioRxiv*, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2021.04.09.438859; this version posted April 10, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

44. John C Brigham and Roy S Malpass. The role of experience and contact in the recognition of faces of own-and other-race persons. *Journal of social issues*, 41(3):139–155, 1985.
45. Robert K Bothwell, John C Brigham, and Roy S Malpass. Cross-racial identification. *Personality and Social Psychology Bulletin*, 15(1):19–25, 1989.
46. Guy Wallis. Toward a unified model of face and object recognition in the human visual system. *Frontiers in psychology*, 4:497, 2013.
47. Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A Study of Face Obfuscation in ImageNet. *arXiv*, 2021.
48. Hyodong Lee, Eshed Margalit, Kamila M Jozwik, Michael A Cohen, Nancy Kanwisher, Daniel L K Yamins, and James J DiCarlo. Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 2020.
49. Shan Xu, Yiyuan Zhang, Zonglei Zhen, and Jia Liu. The face module emerges from domain-general visual experience: a deprivation study on deep convolution neural network. *bioRxiv*, 2020.
50. Le Chang, Bernhard Egger, Thomas Vetter, and Doris Y. Tsao. What computational model provides the best explanation of face representations in the primate brain? *bioRxiv*, 2020.
51. Zhou Wang and Eero P Simoncelli. Maximum differentiation (mad) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision*, 8(12):8–8, 2008.
52. Kamila Maria Jozwik, Ian Charest, Nikolaus Kriegeskorte, and Radoslaw Martin Cichy. Animacy dimensions ratings and approach for decorrelating stimuli dimensions. *Conference on Cognitive Computational Neuroscience*, 2017.
53. Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLoS Computational Biology*, 10(4), 2014. ISBN: 1553-7358 (Electronic)\n1553-734X (Linking).
54. Katherine R Storrs, Tim C Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. Diverse deep neural networks all predict human it well, after training and fitting. *bioRxiv*, 2020.
55. Katherine R Storrs, Seyed-Mahdi Khaligh-Razavi, and Nikolaus Kriegeskorte. Noise ceiling on the crossvalidated performance of reweighted models of representational dissimilarity: Addendum to khaligh-razavi & kriegeskorte (2014). *bioRxiv*, 2020.
56. Tim C Kietzmann, Courtney J Spoerer, Lynn Sörensen, Radoslaw M Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence required to capture the dynamic computations of the human ventral visual stream. *Proceedings of the National Academy of Sciences*, 2019.
57. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
58. Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11):1019–1025, 1999.
59. Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.
60. Gillian Rhodes. Looking at faces: First-order and second-order features as determinants of facial appearance. *Perception*, 17(1):43–63, 1988.
61. Jessica Taubert, Deborah Apthorp, David Aagten-Murphy, and David Alais. The role of holistic processing in face perception: Evidence from the face inversion effect. *Vision research*, 51(11): 1273–1278, 2011.