PNAS

Supporting Information for

Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition

Katharina Dobs*, Joanne Yuan, Julio Martinez, Nancy Kanwisher*

*Corresponding authors: Katharina Dobs, Nancy Kanwisher
Email: katharina.dobs@psychol.uni-giessen.de, ngk@mit.edu

This PDF file includes:

Supplementary Note 1: Layer-wise analysis of face recognition performance, other-race and face inversion effect in VGG16

Methods. To compare the face recognition performance between humans and CNNs for each layer, we extracted activation from each layer of three VGG16 networks (Face-ID CNN; Obj-Cat CNN and Untrained CNN) and performed the same analysis. Specifically, we extracted the activation from each convolutional and fully-connected layer after the relu operation. When the layer was followed by a pooling layer, we extracted the activation from the pooling layer. For each model and layer, we computed the correlation distance between the activation patterns of each pair of images (Fig. 1B, right panel). The network's choice was determined by which of the two matching images had an activation pattern that was closest to the target image. We performed this analysis for the white female datasets upright and inverted, and for the unfamiliar white and Asian female datasets.

Face recognition performance on upright faces (Experiment 1). For the white female dataset, we found that the face-trained CNN began to outperform the object-trained CNN from the last convolutional layer onwards (Fig. S1; p=0, bootstrap test). The performance of the untrained and the object-trained CNNs did not vary much across layers. These results show that the late stages of the face-trained network outperform the object-trained network and approach human-level face recognition performance.

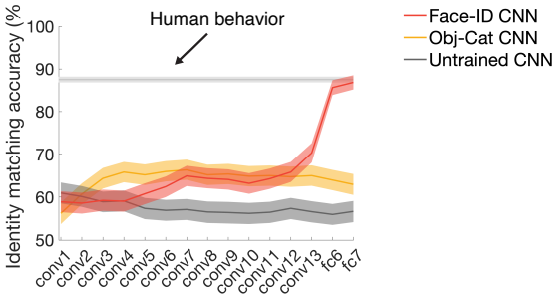

Fig. S1. Late layers of face-trained CNNs outperform object-trained and untrained CNNs and achieve humanlike performance. (a) Human performance (n=1,532) was 87.53% (gray horizontal line; chancel level was 50%). The face-trained CNN (red) outperformed the object-trained CNN (yellow) from the last convolutional layer (conv13) onwards (p=0, bootstrap test). The face-trained CNN further reached human performance in the penultimate fully-connected layer (fc7; p>0.4, bootstrap test). Networks trained on object categorization (yellow) performed better than untrained CNNs (gray) across all layers, but did not reach human-level recognition performance. Error bars denote bootstrapped 95% CI.

Other-race effect (Experiment 4). To measure the other-race effect in each layer, we compared the layer-wise accuracy on the target-matching task on the non-famous white female dataset to the accuracy on the Asian female dataset (Fig. S2). The performance of the white face-trained network on the white female dataset was significantly higher than for the Asian female dataset from the last convolutional layer onwards (all p=0, bootstrap test) and vice versa for the Asian face-trained CNN (all p<0.01, bootstrap test). Neither the object-trained nor the untrained CNN showed a significant difference between the two datasets (all p>0.2, bootstrap test).

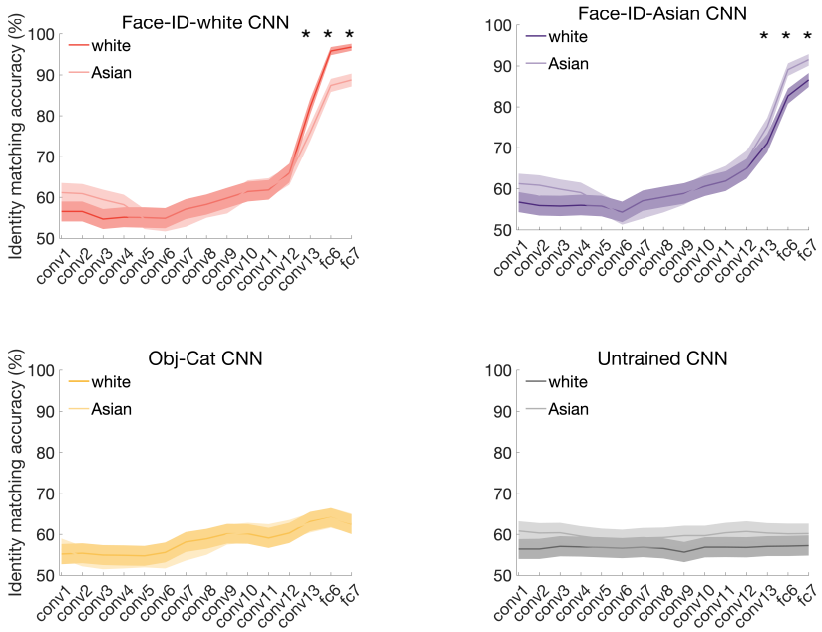

Fig. S2. Late layers of the face-trained CNNs show an other-race effect. For the face-trained CNN (Face-ID-white CNN in red; all Asian faces were removed from the training), the performance on the white female dataset was significantly higher than on the Asian female dataset from the last convolutional layer onwards. The opposite was true for the CNN trained on Asian identities (Face-ID-Asian CNN in purple). Neither the object-trained (Obj-Face-Cat CNN in yellow) nor the untrained (Untrained CNN in gray) CNN showed a significant difference between both datasets at any layer. Shaded areas denote bootstrapped 95% CI. Asterisks indicate significant differences (p<0.01, bootstrap test).

Face inversion effect (Experiment 5). We found similar results when comparing the performance on the upright and inverted white female dataset (Fig. S3). Only the face-trained, but not the object-trained or untrained, CNN showed a significant face inversion effect, i.e., improved performance for upright compared to inverted images, from the last convolutional layer onwards (all $p<0.01$, bootstrap test).

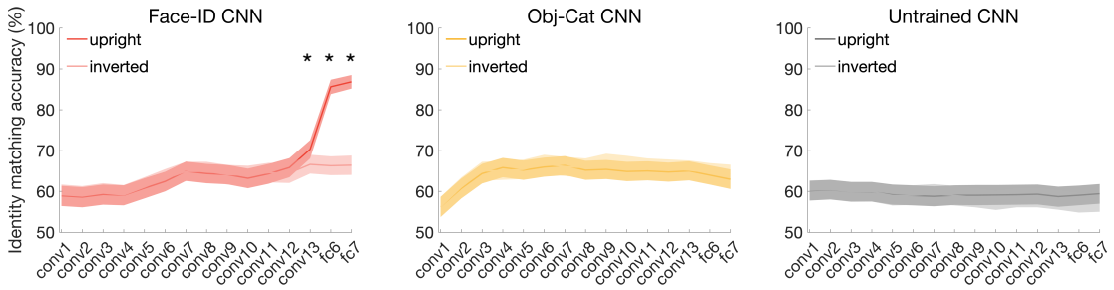

Fig. S3. Late layers of the face-trained CNN show a face inversion effect. The pattern of differences was very similar to the other-race effect. Only the face-trained CNN (Face-ID CNN in red) showed a significant difference in performance between upright and inverted face stimuli starting at the last convolutional layer. Shaded areas denote bootstrapped 95% CI. Asterisks indicate significant differences (bootstrap test, $p<0.01$).

Supplementary Note 2: Replication of results for face recognition performance, other-race and face inversion effect in Alexnet and ResNet

Methods. To test whether the main results in the paper generalize to other architectures, we compared face recognition performance of humans to Alexnet and ResNet-50. For each of these architectures, we trained three different networks: i) one trained on face discrimination, ii) one trained on object categorization, and iii) one untrained. Note that for computational efficiency, we restricted this analysis to the face-trained, object-trained and untrained CNNs (referred to as Face-ID CNN, Obj-Cat CNN, Untrained CNN, respectively). We used the same training stimuli, parameters and procedure to train the CNNs on face or object categorization. For both architectures, we extracted the activation from the last relu (Alexnet) or the last pooling (ResNet) layer preceding the classification layer. We performed this analysis for the white female identities upright and inverted, and for the non-famous white and Asian female identities, respectively.

Face recognition performance on upright faces (Experiment 1). The pattern of results on the white female dataset obtained for VGG16 were generally replicated with Alexnet and ResNet (Fig. S4). We found that the face-trained CNN (red) outperformed the object-trained CNN for both architectures (p=0, bootstrap test), while the object-trained CNN outperformed the untrained CNN (p=0, bootstrap test). However, neither Alexnet nor ResNet reached human performance (p=0, bootstrap test).

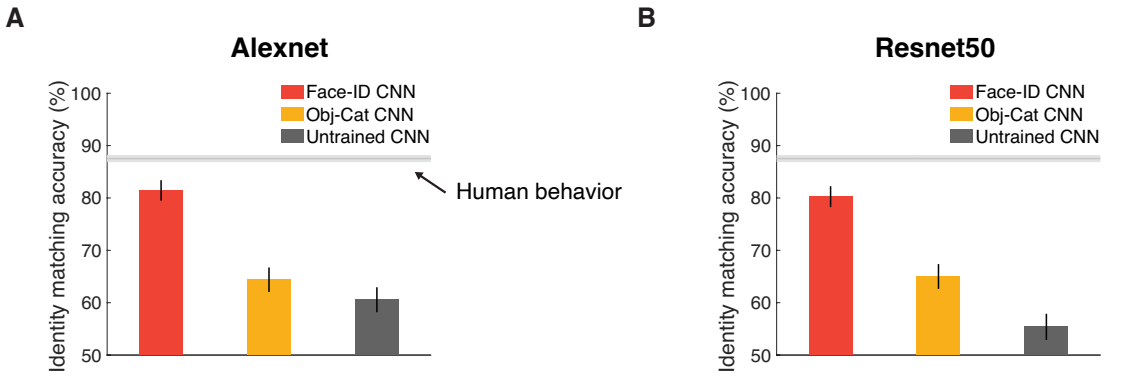

Fig. S4. Pattern of face recognition results replicated in Alexnet (A) and ResNet (B). Error bars denote bootstrapped 95% CI.

Face inversion effect (Experiment 5). To investigate the face inversion effect in Alexnet and ResNet, we ran the same analysis on the inverted face dataset (Fig. S5). As we found for VGG16, the performance of the face-trained Alexnet and ResNet on upright stimuli was significantly higher than for inverted stimuli (p=0, bootstrap test). In both architectures, neither the object-trained nor the untrained CNN showed a significant difference between the two datasets (all p>0.2, bootstrap test). These results show that both architectures, if trained on (upright) faces, also show a face inversion effect.

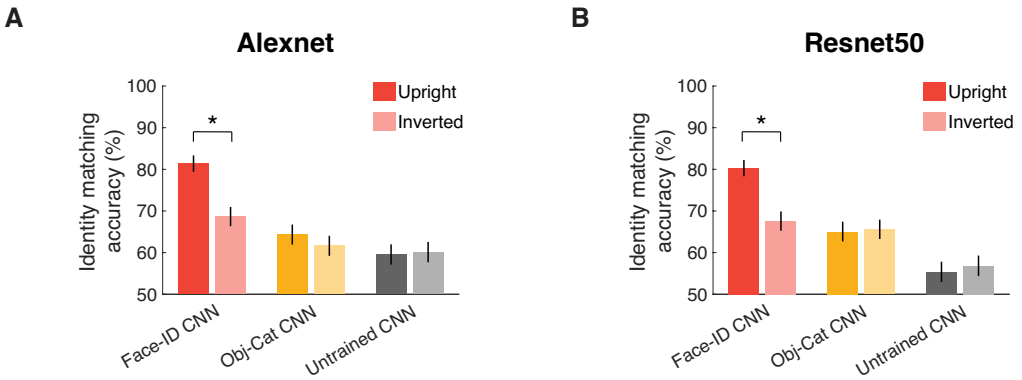

Fig. S5 | Face inversion effect in Alexnet and ResNet (Experiment 5). (A) For Alexnet trained on (upright) faces (red), the performance on the upright female dataset was significantly higher than on the inverted female dataset in the penultimate fully-connected layer. Neither the object-trained (yellow) nor the untrained (gray) Alexnet showed a significant difference between the two datasets. **(B)** The pattern of differences was very similar for ResNet and replicated the results for Alexnet and VGG16: Only the face-trained CNN showed a significant difference in performance between upright and inverted face stimuli in the penultimate fully-connected layer. Error bars denote bootstrapped 95% CI. Asterisks indicate significant differences (bootstrap test, p=0).

Other-race effect (Experiment 4). We found very similar results when comparing the performance on the non-famous white female and Asian female dataset (Fig. S6). Only the face-trained, but not the object-trained or untrained CNNs showed a significant other-race effect, that is improved performance for white compared to Asian face images in the penultimate fully-connected layer (p=0, bootstrap test).

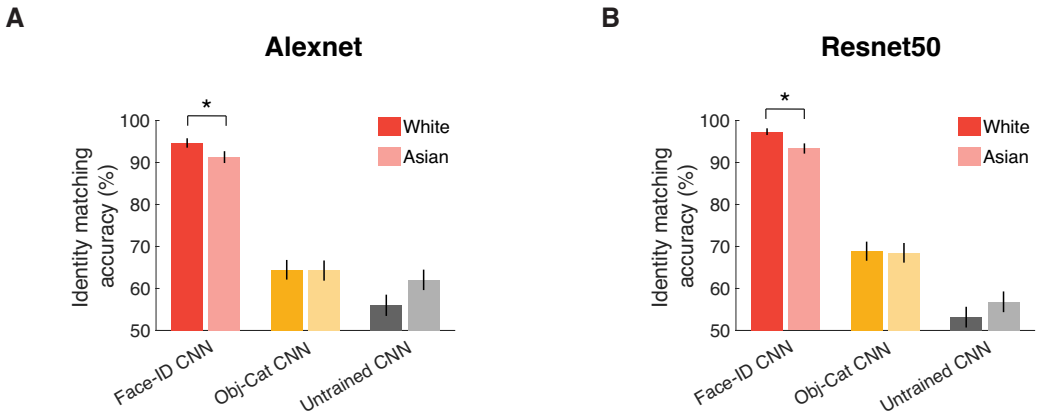

Fig. S6. Other-race effect in Alexnet and ResNet (Experiment 4). (A) For Alexnet trained on (predominantly white) faces (red), the performance on the white female dataset was significantly higher than on the Asian female dataset in the penultimate fully-connected layer. Neither the object-trained (yellow) nor the untrained (gray) Alexnet showed a significant difference between the two datasets. **(B)** The pattern of differences for the ResNet architecture was very similar to Alexnet and replicated the results for VGG16: Only the face-trained CNN showed a significant difference in performance between white and Asian face stimuli in the penultimate fully-connected layer. Error bars denote bootstrapped 95% CI. Asterisks indicate significant differences (bootstrap test, p=0).

Supplementary Note 3: Analysis of errors and trial-by-trial predictivity between humans and CNNs (Experiment 1)

Methods. The analyses so far show that CNNs trained on face recognition achieve accuracy levels similar to humans when tested on the same task. But do they achieve this in the same way? To address this question, we first asked whether humans and CNNs find the same triplets (i.e., the same specific combination of target and two matching images) difficult. To test whether triplets that the CNNs did not perform correctly were also harder for human participants, we separately analyzed the triplets for which the CNNs were correct versus incorrect.

Next, we asked how well each CNN predicts the behavioral choices on a trial-by-trial level (independent of the accuracy of those choices). To compare trial-wise performance between CNNs and human behavior, we first analyzed the behavioral choices. For each triplet, we computed the proportion of trials on which people chose match A (irrespective of whether match A was correct or incorrect). To get an estimate of the noise ceiling, we used split-half reliability, as follows. We divided the participants into 50 random splits, computed the triplet-wise choice proportion for each half, and correlated the two halves. The mean of these split-half correlations (across the 50 splits) plus or minus twice the standard deviation of the distribution served as upper or lower behavioral noise ceilings, respectively. For the CNNs, we transformed the representational distances between the target T and match A and the target T and match B obtained for each triplet into choice probabilities. For each triplet, we used Luce's choice axiom to calculate the probability of choosing match A over match B given target T. According to this axiom, the probability of choosing match A can be expressed as the conditional probability of selecting A given the target T:

$$P(A|T) = 1 - \frac{d(A,T)}{d(A,T) + d(B,T)}$$

where d(A,T) is the representational distance between the match A and the target image T, and d(B,T) is the corresponding representational distance between match B and target T. If the two matches in a triplet are equally similar to the target T (i.e., d(A,T) = d(B, T)), the probability of choosing A is 0.5. If the match A is very similar to the target T (e.g., the distance d(A,T) = 0.1) and the match B is very dissimilar (e.g., the distance d(B,T) = 0.9), the probability of choosing A is very high. To test how well the CNN's choice probabilities predict the behavioral proportional choices, we correlated the behavioral proportional choices of match A with the CNN's probabilities of choosing match A for all triplets. We further computed bootstrapped 95% CIs by bootstrapping the triplets and computing the correlations 10,000 times. We used bootstrap tests to compare the predictivity between CNNs.

Results. For the face-identity trained CNN, we indeed found that human performance was significantly better on triplets in which the CNN was correct (human performance 88.2%; 1355 triplets) than on triplets for which it was incorrect (human performance: 83.3%; 205 triplets; p=0, bootstrap test). Moreover, this difference in performance (4.9%) was significantly smaller for the CNN trained on object categorization (difference: 1.5%; p=0.02, bootstrap test) and the untrained CNN (difference: 1.6%; p=0.02, bootstrap test). Here, humans performed significantly but only slightly worse on triplets the CNNs performed incorrectly (human performance: 86.6% for both Obj-Cat and untrained CNN) than on triplets the CNNs performed correctly (human performance: 88.1% for Obj-Cat CNN; p=0.03, bootstrap test; 88.2% for Untrained CNN; p=0.02, bootstrap test). This finding suggests that the face-identity trained CNN not only achieves a similar recognition accuracy to humans, but also shows similar errors to humans.

Would a network trained on face detection better match human face behavior? We indeed found a significant difference of 3.2% in human performance on triplets in which the CNN trained on object and face categorization (Obj-Face-Cat CNN) was correct vs. incorrect (human performance: 88.6% (1047 triplets) vs. 85.4% (513 triplets); p=0, bootstrap test). The size of this difference was not significantly different from those found for any of the other CNNs (Face-ID CNN: p=0.2; Obj-Cat and Untrained CNN: p=0.1; bootstrap tests).

How well would each network predict human behavior on a trial-by-trial level independent of overall accuracy (Fig. S7)? Prior to correlating the CNN's choice probabilities with human choice

proportions, we measured how reliable human choices were using split-half reliability. Human choices across triplets were highly reliable (mean split-half correlation across 50 random splits: r=0.94). We next asked how well each of the CNNs would correlate with the behavioral choice proportions. We found that the face-identity trained CNN (Face-ID CNN: r=0.74) explained around 62% of the explainable variance (normalized by the split-half reliability) in the human behavioral choices, thereby outperforming the CNN trained on object and face categorization (Obj-Face-Cat CNN: r=0.37; p=0, bootstrap test), the CNN trained on object categorization (Obj-Cat CNN: r=0.29; p=0, bootstrap test) and an untrained CNN (Untrained CNN: r =0.18; p=0, bootstrap test). These results suggest that the face-identity trained CNN not only achieves human-level accuracy in face recognition, but also predicts human behavioral choices on a trial-by-trial level well.

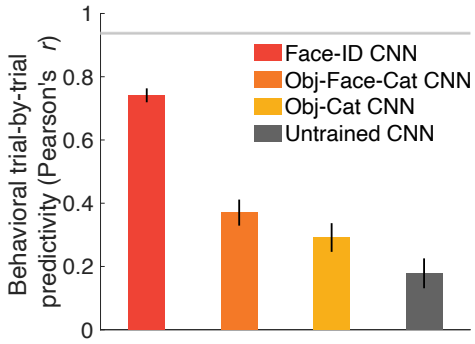

Fig. S7. Behavioral trial-by-trial predictivity by CNNs (Experiment 1). The face-trained CNN (red) best predicted human behavioral choices on a trial-by-trial level. Networks trained on object categorization and face detection (orange), or object categorization only (yellow) performed better than the untrained CNN (gray), but were not highly predictive of human behavioral choices. Error bars denote bootstrapped 95% CIs. The gray line represents the split-half reliability of the human behavioral choices (mean ±2*SD across 50 random splits).

Supplementary Note 4: Comparing human perceptual similarity in a similarity matching task to task-optimized CNNs (Experiment 3).

Participants. A set of 697 individual workers participated in the similarity-matching task (Fig. 2C) on Amazon Mechanical Turk. A total of 29 workers were excluded from the analysis due to overly fast responses (response time in more than five trials < 500ms or more than 10 trials < 800 ms). All workers were located in the United States. The average workers' age was between 25 and 34 years, 56% of workers were female, 42% were male and 2% reported 'other' or did not report their sex. The majority of the workers were white (71%), 19% were Black, 8% were Asian and 2% reported 'other' or did not report their race. For this task, workers were not restricted in the number of trials they could perform. All workers provided informed consent and were compensated financially for their time. The experimental protocol was approved by the Massachusetts Institute of Technology (MIT) Committee on the Use of Humans as Experimental Subjects (COUHES No 1806424985) and conducted following all ethical regulations for conducting behavioral experiments.

Stimuli and behavioral representational dissimilarities. To test whether the results from the multi-arrangement task from Experiment 2 would generalize to a different dataset and task, we conducted a similarity-matching task on Amazon Mechanical Turk. To construct this task, we chose one image of each of 60 unfamiliar male identities from the Flickr-Faces-HQ database (1). The stimuli included 60 young male identities of similar age (approximately between 20 and 30 years old) with a neutral facial expression. Participants were asked to choose which of two images was more similar to a third target image (i.e., triplet). Each of the possible triplets (60x59x58/2 for a total of 102,660 triplets) was sampled once (although some triplets were excluded, see Participants). The choice for a specific triplet provided two pairwise similarities: between the target and each of the matching images (e.g., when the choice for the triplet with target A and matches B and C was C, this would result in "1" for the pair A-C, and "0" for the pair A-B). Thus, the perceived similarity of each pair of face images was on average sampled ~100 times (i.e., each of the 102,660 triplets produced two of the 1,770 (60x59/2) pairwise similarities resulting in 116 (102,660/1,770) samples without exclusions). The proportional number of times that each target-match pair was chosen as the more similar pair was used as the similarity value for the pair and converted into a dissimilarity value by subtracting it from 1. We extracted the lower triangle excluding the diagonal from the resulting dissimilarity matrix (see Supplementary Note 5 for visualization of the behavioral RDM) to obtain a vector of pairwise dissimilarities. We then used the vector of dissimilarity values to compute the similarity with the CNNs.

To compute the noise ceiling in this task, we used split-half correlation. Since participants were not limited in the number of trials they could perform, participants contributed with varying degrees to the final dissimilarity matrix. To not bias the split-half correlation by participants who contributed more trials, we only used the first set of trials collected by each individual participant in the noise-ceiling calculation. We randomly split the participants into two halves 50 times and computed the correlation between the dissimilarity vectors based on the two halves for each split. We then used the mean correlation added and subtracted by twice the standard deviation of this set of correlations as noise ceiling.

Representational similarity analysis between humans and CNNs. As in Experiment 2, we obtained representational dissimilarities in CNNs, by presenting the same stimuli used for the human participants to the four CNNs. For each CNN, we extracted the activation patterns to each image separately from the penultimate layer (see Supplementary Note 7 for other layers) and computed the correlation distance (1 – Pearson's *r*) between each pair of activation patterns. This resulted in one RDM for each of the four CNNs (see Supplementary Note 6 for visualization of the RDMs).

To compute the similarity between the human RDMs and the RDMs obtained for the CNNs, we rank-correlated the human behavioral dissimilarity vector obtained from the pairwise

proportional choices across all triplets with the corresponding CNN dissimilarity vector constructed from the same stimuli.

Statistical inference. To measure statistical significance, we used the same bootstrap tests as in Experiment 2, but bootstrapped the dissimilarity vectors instead of participants. Specifically, we bootstrapped the dissimilarity values of the behavioral and CNN dissimilarity vectors 10,000 times and computed the rank correlation to obtain 95% CIs, and to compute a distribution of correlation differences.

Supplementary Note 5: RDMs of representational similarity analyses in VGG16

Methods. We used representational similarity analysis to compare the behavioral representational dissimilarity matrices (RDMs) obtained from the multi-arrangement task (Experiment 2; Fig. 2B) and the similarity-matching task (Experiment 3) to the RDMs of all four VGG16 models. Here, we plot the behavioral RDMs along with the RDMs obtained from the four VGG16 models for visual comparison.

Multi-arrangement task (Experiment 2). For the multi-arrangement task (Experiment 2; Fig. S8), the face-trained CNN (Face-ID CNN) best mirrors the structure of the behavioral RDMs. This similarity goes beyond the coarse distinctions between male and female and old and young faces. For example, within the old female faces, the Face-ID CNN also shows a high similarity between the third and the fourth identity. Note that while the object-trained and object-and-face-categorization trained CNNs also show some of the coarse categories (e.g., older male faces in the bottom right are highly similar to each other but distinct from young female faces), they do not capture all aspects of the fine-grained structure. For example, older female faces are more similar to young male faces (blue colors in the bottom left quadrant) in these CNNs than in human behavior and the Face-ID CNN (yellow colors in the bottom left quadrant). Interestingly, even the untrained CNN shows a trend for some of the aspects in the behavioral RDM, such as a high similarity between older female identities.

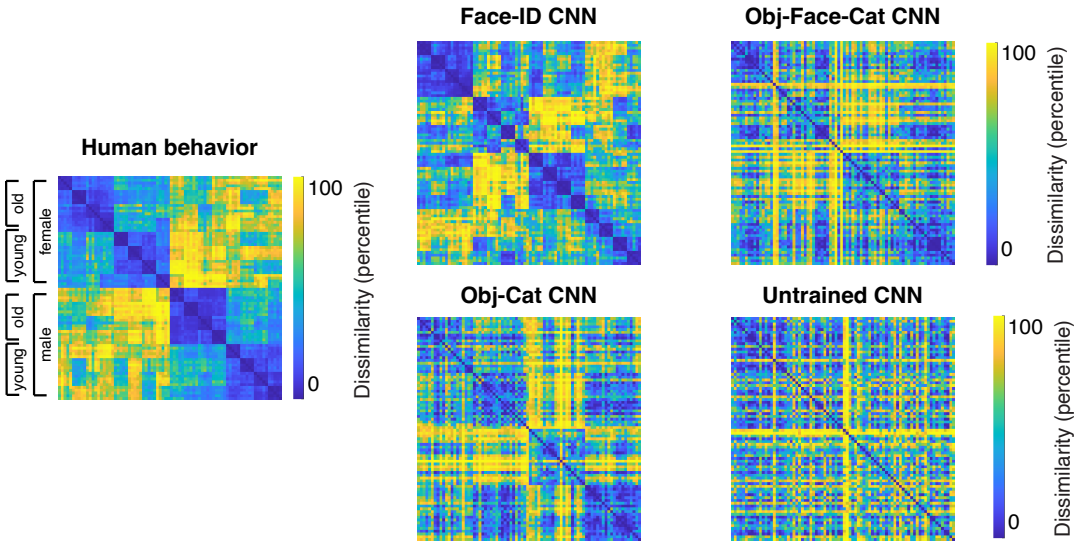

Fig. S8. Face-trained CNN best matches human perceptual similarity in a multi-arrangement task (Experiment 2). The stimuli used in the multi-arrangement task consisted of 5 stimuli for each of 16 identities for which half of them were old versus young and female versus male, respectively. In the RDM, the stimuli are arranged by identity (5 images per identity), age (young versus old) and gender (female versus male). The Face-ID CNN best matches the behavioral RDM. The Obj-Face-Cat and Obj-Cat CNN show some coarse aspects but do not match the fine-grained details of the behavioral RDM.

Similarity-matching task (Experiment 3). For the similarity-matching task (Experiment 3; Fig. S9), all 60 identities were young and male and we used one image per identity. As can be seen from the behavioral RDM, there is no clear structure in the RDM. As for the multi-arrangement ask, the RDM of the face-trained CNN appears most similar to the behavioral RDM. In particular, the other three CNNs show strong similarities for certain images with all other images (as can evident by blue lines in the matrices), which are not visible in the behavioral or the RDM of the Face-ID CNN.

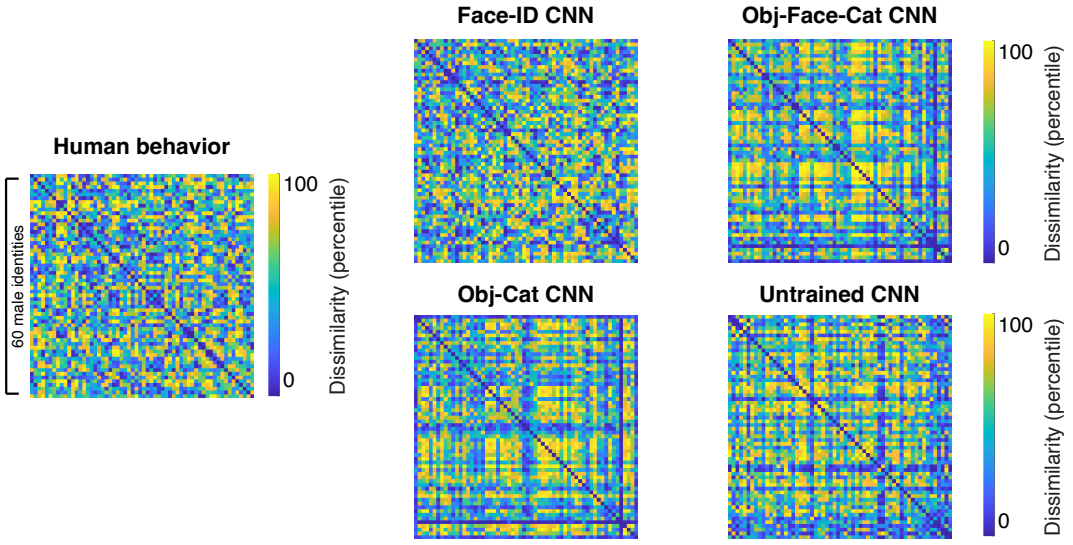

Fig. S9. Face-trained CNN best matches human perceptual similarity in a similarity-matching task (Experiment 3). The stimuli in the similarity-matching task consisted of 60 young, male identities (one image per identity). The Face-ID CNN best matches the behavioral RDM. The Obj-Face-Cat and Obj-Cat CNN show specific structures (e.g., strong similarity of a specific image with all other images) that are neither visible in the behavioral RDM nor the face-trained CNN.

Supplementary Note 6: Layer-wise representational similarity analysis in VGG16

Methods. We used representational similarity analysis to compare the behavioral representational dissimilarity matrices (RDMs) obtained from the multi-arrangement task (Experiment 2; Fig. 2B) and the similarity-matching task (Experiment 3; Fig. 2C) to the layer-wise RDMs of the VGG16 models trained on face identification and object categorization and the untrained VGG16 model. Specifically, to obtain layer-wise RDMs for each model, we computed the distance (i.e., 1 – Pearson's *r*) between the activation patterns extracted from each layer for the same stimuli.

Results. In the multi-arrangement task (Experiment 2; Fig. S10A), we find that correlations between the face-trained CNN (red) and human behavior increased with progressive layers in the network from the first convolutional layer (Spearman's *r*: 0.05), to mid-level convolutional layers (e.g., Conv8: Spearman's *r*: 0.16) to the last convolutional layer (Spearman's *r*: 0.34), the latter even reaching noise ceiling (i.e., the maximum correlation possible given the consistency across subjects; light-gray vertical bar). In contrast, the object-trained CNN (yellow) represented faces less similarly to humans (max. Spearman's *r*: 0.19), with correlations increasing slightly after the first 4-5 layers, reaching its maximum in the penultimate fully-connected layer. The representational dissimilarities of the untrained CNN (dark gray) showed a low correlation with human behavior across all layers (max. Spearman's *r*: 0.04). Thus, the later stages of processing in face-trained, but not object-trained or untrained, CNNs match human behavior well, suggesting that faces are similarly represented in human behavior and late stages of face-trained CNNs.

We find a very similar pattern for the similarity-matching task (Experiment 3; Fig. S10B).

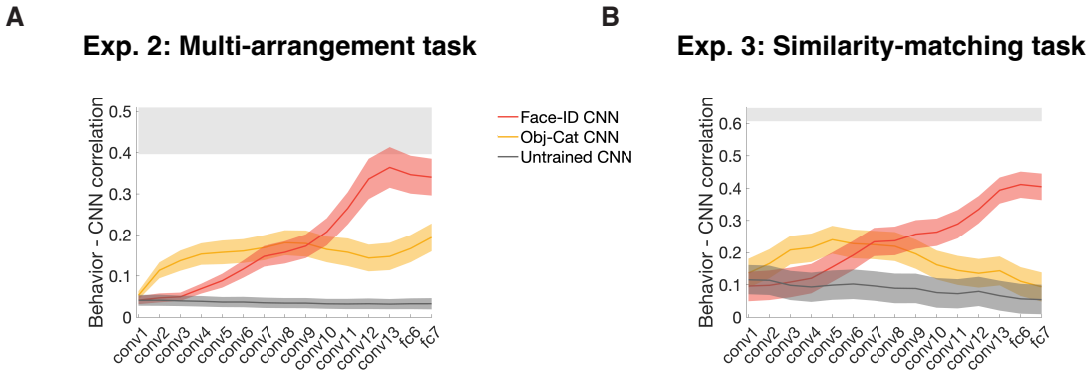

Fig. S10. Late layers of face-trained but not object-trained or untrained CNNs match human face behavior. (A) We performed RSA on all layers of the three VGG16 models and human behavioral similarities from the multi-arrangement task (Experiment 2). Late layers of the Face-ID CNN (red) matched human behavioral representational similarity best and reached the noise ceiling (light gray bar). Neither the untrained CNN (Untrained CNN, dark gray) nor the object-trained CNN (Obj-Cat CNN, yellow) matched human representational similarities. Shaded areas represent bootstrapped SEMs across subjects. Gray vertical bar represents noise ceiling. **(B)** The results in (A) were replicated on the similarity-matching task (Experiment 3) using on a distinct dataset of 60 unfamiliar male identities (one image each). Late layers of the Face-ID CNN (red) matched human behavioral representational similarity best, far outperforming the untrained CNN (gray) and the object-trained CNN (yellow). Shaded areas represent bootstrapped 95% CIs across dissimilarity values. Gray vertical bar represents noise ceiling.

Supplementary Note 7: Representational similarity analysis in Alexnet and ResNet

Methods. To test whether these results would generalize to other architectures, we compared the behavioral RDMs obtained from the multi-arrangement task (Experiment 2; Fig. 2A) and the similarity-matching task (Experiment 3; Fig. 2B) to the CNN RDMs of all three Alexnet and ResNet-50 models. Specifically, to obtain RDMs for each model, we computed the distance (i.e., 1 – Pearson's *r*) between the activation patterns extracted from the penultimate layer for the same stimuli.

Results. For the multi-arrangement task (Exp. 2; Fig. S11A), the correlations between the face-trained Alexnet (red) and human behavior were close to the noise ceiling (Spearman's *r*: 0.32, close to noise ceiling). In contrast, the object-trained CNN (yellow) represented faces less similarly to humans (Spearman's *r*: 0.18). The representational dissimilarities of the untrained CNN (dark gray) showed a low correlation with human behavior (Spearman's *r*: 0.04). We replicated this pattern of results for Alexnet in the similarity-matching task (Exp. 3; Fig. S11B). Thus, processing in face-trained, but not object-trained or untrained, Alexnet models match human behavior well.

ResNet trained on faces, objects and untrained showed a very similar pattern. In the multi-arrangement task (Exp. 2; Fig. S11A) the face-trained CNN (red) even reached the noise ceiling (Spearman's *r*: .36), while the object-trained (Spearman's *r*: .19) and the untrained (Spearman's *r*: .03) CNN achieved much lower correlations with human similarity representations. We again replicated this pattern of results for ResNet in the similarity-matching task (Exp. 3; Fig S11B).

Taken together, these findings suggest that faces are similarly represented in human behavior and face-trained feed-forward CNNs, irrespective of architecture.

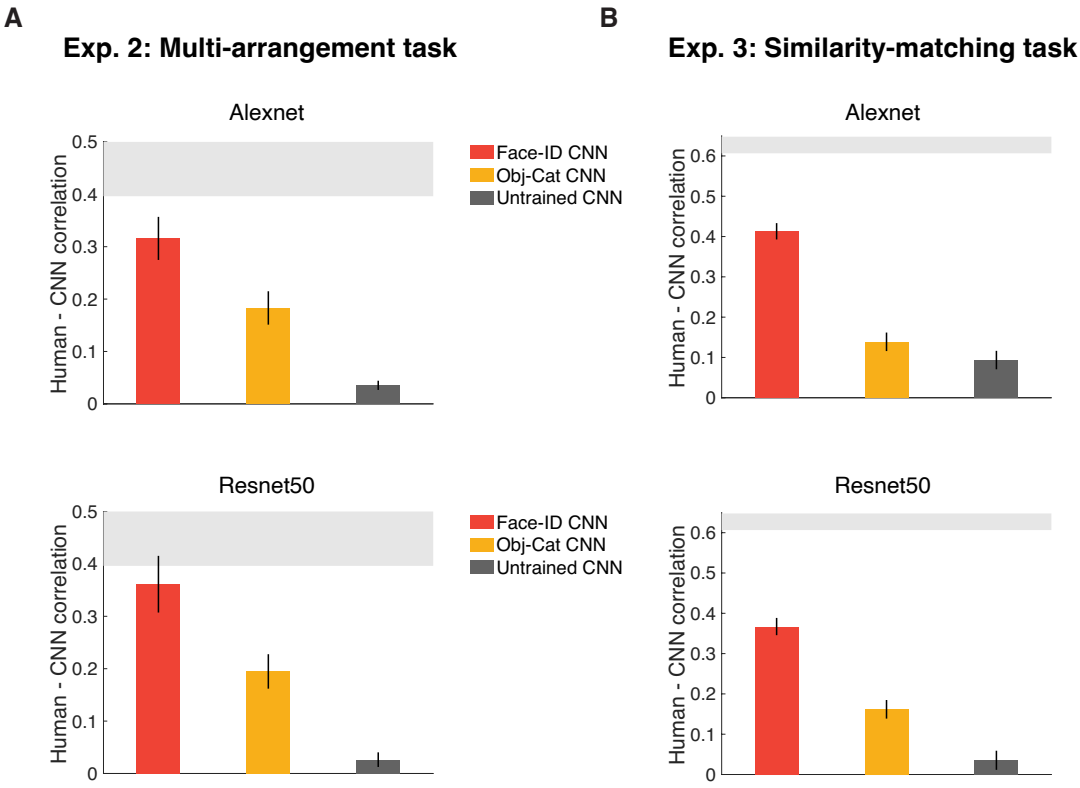

Fig. S11. Face-trained but not object-trained or untrained Alexnet and ResNet-50 architectures approach human face behavior. **(A)** We performed RSA by measuring the similarity between human behavioral similarities from a multi-arrangement task (Experiment 2) and layer-wise RDMs obtained from the three Alexnet and three ResNet-50 models. For both architectures, the face-trained models (red) matched human behavioral representational similarity best and approached noise ceiling (light gray bar). Neither the untrained CNN (dark gray) nor the object-trained CNN (yellow) matched human representational similarities. Error bars represent bootstrapped SEMs across subjects. **(B)** The results of in (A) were replicated in a similarity-matching task on Amazon Mechanical Turk (Experiment 3) based on a distinct dataset of 60 unfamiliar male identities (one image each). For both architectures, the Face-ID CNN (red) matched human behavioral representational similarity best, far outperforming the untrained CNN (gray) and the object-trained CNN (yellow). Error bars represent bootstrapped 95% CIs across dissimilarity values.

Table S1. Overview of experiments, participants and datasets

Human Experiment	**Human participants (n)**	**Testing platform**	**Task**	**Stimuli**	**Figure**
Exp. 1: Face recognition - upright	1,532	Amazon Mechanical Turk	Target-matching task	Set A	Fig. 1C
Exp. 2: Perceptual similarity	14	Meadows	Multi-arrangement task	Set B	Fig. 2B
Exp. 3: Perceptual similarity	668	Amazon Mechanical Turk	Similarity-matching task	Set C	Fig. 2C
Exp. 4A: Other-race effect - white participants	269	Amazon Mechanical Turk	Target-matching task	Set D	Fig. 3A
Exp. 4B: Other-race effect - Asian participants	102	Clickworker + Meadows	Target-matching task	Set D	Fig. 3A
Exp. 5: Face recognition - inverted	1,219	Amazon Mechanical Turk	Target-matching task	Set A inverted	Fig. 3B

Table S2. Overview of experimental datasets

Test Stimulus Set Name	Description	Link
Set A	200 face images (5 images of each of 40 celebrities)	https://osf.io/dbks3/
Set B	80 face images (5 images of each of 16 identities)	https://osf.io/gk6f5/
Set C	60 face images (1 image of each of 60 identities)	https://osf.io/dbks3/
Set D	400 face images (5 images of each of 40 white and 40 Asian identities)	https://osf.io/dbks3/
Set E	1000 face images (10 images of each 100 identities)	https://osf.io/dbks3/
Set F	1000 car images (10 images of each 100 car model/makes)	https://osf.io/dbks3/

Table S3. Overview of CNN experiments

CNN Experiment	CNNs	Analysis Method	Stimuli	Figure
Face recognition – upright (Exp. 1)	CNN 1-4	Target-matching task	Set A	Fig. 1C
Perceptual similarity (Exp. 2)	CNN 1-4	RSA	Set B	Fig. 2B
Perceptual similarity (Exp. 3)	CNN 1-4	RSA	Set C	Fig. 2C
Other-race effect (Exp. 4)	CNN 3-8	Target-matching task	Set D	Fig. 3A
Face recognition – inverted (Exp. 5)	CNN 1-4	Target-matching task	Set A inverted	Fig. 3B
Inverted face inversion effect	CNN 1, 9	SVM decoding	Set E	Fig. 4A
Car inversion effect	CNN 1, 3, 4, 10	SVM decoding	Set F	Fig. 4B

Table S4. Overview of trained and untrained CNNs

CNN #	CNN Name	Training Set	Link
CNN 1	Face-ID CNN	1,714 VGGFace2 classes	https://github.com/ox-vgg/vgg_face2
CNN 2	Obj-Face-Cat CNN	423 ImageNet classes + 1,714 VGGFace2 classes (assigned to one output class)	
CNN 3	Obj-Cat CNN	423 ImageNet classes	https://www.image-net.org/challenges/LSVRC/2012/index.php
CNN 4	Untrained CNN	None	
CNN 5	Face-ID-white CNN	1,654 VGGFace2 classes (white only)	
CNN 6	Face-ID-Asian CNN	1,654 Asian Face Dataset classes	https://github.com/X-zhangyang/Asian-Face-Image-Dataset-AFD-dataset
CNN 7	Obj-Face-Cat-white CNN	423 ImageNet classes + 1,654 VGGFace2 classes (white only; assigned to one output class)	
CNN 8	Obj-Face-Cat-Asian CNN	423 ImageNet classes + 1,654 Asian Face Dataset classes (assigned to one output class)	
CNN 9	Face-ID-inv CNN	1,714 VGGFace2 classes (inverted)	
CNN 10	Car CNN	1,109 combined CompCars dataset classes	http://mmlab.ie.cuhk.edu.hk/datasets/comp_cars/

Supplementary References

1. T. Karras, S. Laine, T. Aila, A Style-Based Generator Architecture for Generative Adversarial Networks in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), pp. 4401–4410.