Cognition 208 (2021) 104341

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Computational insights into human perceptual expertise for familiar and unfamiliar face recognition

Nicholas M. Blauch[a,b,*], Marlene Behrmann[b,c], David C. Plaut[b,c]

[a] Program in Neural Computation, Carnegie Mellon University, United States of America
[b] Neuroscience Institute, Carnegie Mellon University, United States of America
[c] Department of Psychology, Carnegie Mellon University, United States of America

ARTICLE INFO

Keywords:
Familiarity
Face recognition
Expertise
Invariance
Deep convolutional neural network

ABSTRACT

Humans are generally thought to be experts at face recognition, and yet identity perception for unfamiliar faces is surprisingly poor compared to that for familiar faces. Prior theoretical work has argued that unfamiliar face identity perception suffers because the majority of identity-invariant visual variability is idiosyncratic to each identity, and thus, each face identity must be learned essentially from scratch. Using a high-performing deep convolutional neural network, we evaluate this claim by examining the effects of visual experience in untrained, object-expert and face-expert networks. We found that only face training led to substantial generalization in an identity verification task of novel unfamiliar identities. Moreover, generalization increased with the number of previously learned identities, highlighting the generality of identity-invariant information in face images. To better understand how familiarity builds upon generic face representations, we simulated familiarization with face identities by fine-tuning the network on images of the previously unfamiliar identities. Familiarization produced a sharp boost in verification, but only approached ceiling performance in the networks that were highly trained on faces. Moreover, in these face-expert networks, the sharp familiarity benefit was seen only at the identity-based output probability layer, and did not depend on changes to perceptual representations; rather, familiarity effects required learning only at the level of identity readout from a fixed expert representation. Our results thus reconcile the existence of a large familiar face advantage with claims that both familiar and unfamiliar face identity processing depend on shared expert perceptual representations.

1. Introduction

Faces are perhaps the most important class of visual stimuli for humans, and adult humans have developed substantial expertise for their perception (Diamond & Carey, 1986), performing effortless recognition and recall of associated identity-specific semantic information for a very large number of known individuals. However, the nature of this expertise has been the subject of multiple substantive debates. Researchers have long argued as to whether human expertise for faces is supported by a modular neural and cognitive mechanism dedicated to face recognition (Kanwisher, McDermott, & Chun, 1997; Kanwisher & Yovel, 2006; Tsao & Livingstone, 2008) or whether it arises through domain-general learning rules which could equally be applied to other categories such as artificial "Greeble" stimuli (Gauthier & Tarr, 1997; Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999b), or birds and cars (Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999a; Gauthier, Skudlarski, Gore, & Anderson, 2000) under appropriate task demands (Tarr & Gauthier, 2000). The disagreement is not whether humans are experts at face recognition, it is whether this expertise is domain-general or domain-specific.

In addition to this ongoing debate, an even more basic claim has recently been called into question, challenging the tenet that humans are experts at face recognition. Young and Burton (2018) argued that expertise for face recognition is restricted to familiar faces, and that perceptual performance with unfamiliar faces does not meet the qualifications for expertise. The evidence they offer for this proposition comes from a body of research showing that humans perform more poorly at processing the faces of unfamiliar versus familiar individuals. For example, across four experiments requiring participants to match unfamiliar faces, performance was highly error prone, especially when matches varied in viewpoint and expression (Bruce et al., 1999). As illustrated in Fig. 1, it can be quite difficult to determine whether two images of unfamiliar individuals are of the same identity (Fig. 1A), but if we are familiar with them, the task becomes substantially easier

* Corresponding author at: Baker Hall 342C, 4825 Frew St, Pittsburgh, PA 15213, United States of America.
E-mail address: blauch@cmu.edu (N.M. Blauch).

https://doi.org/10.1016/j.cognition.2020.104341
Received 9 November 2019; Received in revised form 22 May 2020; Accepted 26 May 2020
Available online 23 June 2020

A Australian celebrities **B** American celebrities

Fig. 1. Verifying the identity of images of unfamiliar faces can be much harder than doing so for familiar faces. Most American readers will be familiar with the American celebrities on the right, but not with the Australian celebrities on the left. The face verification task requires the participant to determine whether pairs of images are of the same or a different identity. The top row shows difficult identity matches, and the bottom row shows difficult identity non-matches.

(Fig. 1B). Beyond verification, when asked to sort photographs of two individuals into identity-specific piles, participants familiar with the identities correctly sorted the photos into two piles, whereas individuals unfamiliar with the identities used an average of seven piles; whereas images of different identities were rarely confused as the same identity, images of the same identity were frequently separated into multiple piles, reflecting the failure to group highly variable images of the same identity together (Jenkins, White, Van Montfort, & Mike Burton, 2011). Finally, in addition to demonstrating that unfamiliar face identity processing was less robust than that of familiarized faces, Megreya and Burton (2006) demonstrated that upright unfamiliar and familiar face matching accuracy correlated only weakly and non-significantly ($r = .277$, $p > .05$), whereas the correlation between performance on upright unfamiliar face matching and inverted familiar face matching was strong and highly significant ($r = .673$, $p < .01$), leading the authors to title their article *Unfamiliar faces are not faces*. Together, these results indicate that unfamiliar and familiar face perception may be quite different, and perhaps recruit qualitatively different perceptual mechanisms.

To account for these results, Kramer, Young, Day, and Burton (2017) and Kramer, Young, and Burton (2018) developed a computational model of face recognition of unfamiliar and familiar faces. This model falls within the class of Active Appearance Models (Cootes, Taylor, Cooper, & Graham, 1995; Cootes, Edwards, & Taylor, 1998) in which the goal is first to account for variations in face landmark position, and then to derive and analyze a shape-free appearance representation. Specifically, their model requires human input for a semi-automated assignment of landmarks to positions along key facial locations (i.e., locating the outline of the lips, nose, and eyes). Images are then linearly aligned to the average shape representation, and the resulting aligned images are analyzed for "shape-free" texture/appearance. To simulate unfamiliar recognition, Kramer and colleagues performed principal components analysis (PCA) on the texture representations of a set of familiar individuals, and projected images of unfamiliar individuals into this space. In order to simulate familiar recognition, they performed a linear discriminant analysis (LDA) on the PCA representation, yielding a PCA + LDA space. Intriguingly, whereas the PCA + LDA space separated familiar individuals well, the PCA space alone did a very poor job at separating unfamiliar individuals (Kramer et al., 2018). In contrast, the PCA space was shown to capture non-identity attributes such as race and gender (Kramer et al., 2017), which humans robustly perceive in unfamiliar individuals. The researchers argued that this model helps explain why human observers struggle at unfamiliar face recognition, but are robust at familiar face recognition: the majority of within-identity face variability is idiosyncratic and must be learned for each individual separately. Taking all of this into consideration, Young and Burton (2018) claimed that the recognition of unfamiliar faces does not meet the criteria for expertise, which is characterized by high accuracy and by relative automaticity of performance.

The claim that human face expertise is limited to familiar faces has been met with sharp disagreement from researchers who view face perception broadly as a specific instance of developed visual expertise. For example, Sunday and Gauthier (2018) argued that humans are experts at unfamiliar face recognition when compared to the appropriate baseline of general object recognition, and that expertise is not determined by performance level per se but by the extent of development of the perceptual skill. Such development can be induced for novel stimuli in the laboratory (e.g. Greebles; Gauthier et al., 1999a), and is associated with characteristic error patterns, such as the inversion effect assumed to indicate configural processing which is greater for learned than unlearned stimulus categories. Rossion (2018) further argued that humans are expert at all forms of *visual* face recognition, also pointing to key error patterns found for unfamiliar faces—including the inversion effect (Valentine, 1988), the other race effect (Bothwell, Brigham, & Malpass, 1989), and the composite face illusion (Young, Hellawell, & Hay, 2013)—as well as neuropsychological evidence, where the substantially better performance of normal adults compared to that of individuals with prosopagnosia (both congenital (Behrmann & Avidan, 2005) and acquired (Damasio, Damasio, & Van

Hoesen, 1982)) demonstrates that unfamiliar face recognition is a highly and specifically developed skill. To explain the relative improvement for familiar faces, Rossion (2018) suggested that the advantage "may be based on associated semantic, affective, and lexical (rather than visual) processes/representations" (p. 471). Finally, Abudarham, Shkiller, and Yovel (2019) demonstrated that the same critical features (Abudarham & Yovel, 2016) are used for both unfamiliar and familiar face recognition, suggesting that the difference between unfamiliar and familiar face recognition may be conceptual rather than perceptual.

In the current work, we attempt to combine many aspects of these various accounts in simulations that clarify the extent of human expertise in unfamiliar and familiar face recognition, and the reason why a strong advantage is seen for the latter over the former. Although Kramer et al. (2018) claimed that much of the variability of face images is idiosyncratic, we argue that they substantially underestimated human performance on unfamiliar face recognition, thereby overestimating the share of idiosyncratic variability in face recognition. Moreover, given that their model requires human input at the alignment stage—a process which obscures the representations required to perform such landmarking—the model does not provide a good evaluation of the expertise underlying human unfamiliar face recognition.

Given recent successes in deep learning for classification of object and face images (Cao, Shen, Xie, Parkhi, & Zisserman, 2018; Krizhevsky, Sutskever, & Hinton, 2012; Parkhi, Vedaldi, & Zisserman, 2015; Simonyan & Zisserman, 2015), we hypothesized that a deep convolutional neural network (DCNN) trained on faces would be capable of achieving human-level performance on both unfamiliar and familiar face recognition. Testing performance on ambient images from Labeled Faces in the Wild (Huang, Ramesh, Berg, & Learned-Miller, 2007), we find that a DCNN trained on thousands of face identities substantially outperforms a fully automated Active-Appearance Model conceptually similar to that used by Kramer et al. (2017) and Kramer et al. (2018). We go on to analyze the aspects of visual experience that are necessary to achieve high performance in the network. If the necessary visual experience is both extensive and face-specific, it would suggest that unfamiliar face recognition, like familiar face recognition, is a specific learned expertise (Sunday & Gauthier, 2018).

To determine the extent to which human-level face verification performance depends on extensive experience with faces per se, and not just with general object categories, we manipulated the pretraining conditions—both the extent and domain of visual experience—of the DCNN before testing it on a new set of unfamiliar faces. To understand the extent to which idiosyncratic visual experience with specific identities is critical, we fine-tuned the network on a training set of images of the previously unfamiliar face identities, and then tested verification on the same images for which we calculated unfamiliar verification performance. To examine the perceptual level at which successful performance emerges, we tested verification performance using representations at several layers throughout the network. We also examined the extent to which familiarity effects depend on perceptual comparisons, as opposed to learning to map fixed perceptual features to identity representations. Finally, we compared the performance of our network directly with data obtained from humans performing a difficult face verification task in order to confirm that our conclusions about the role of prior experience, the extent of expertise for unfamiliar recognition, and the computational role of familiarity apply to the typical human observer.

2. Methods

2.1. A fully-automated shape-free linear texture analysis model

We performed a conceptual replication of the model used by Kramer et al. (2017) and Kramer et al. (2018). Rather than determine landmarks for thousands of images by hand, we opted for a fully automated approach. We used a pretrained dense feature-based active appearance model to compute shape-free appearance representations. This model—implemented in the menpofit toolbox (Antonakos, Alabort-I-Medina, Tzimiropoulos, & Zafeiriou, 2015)—was fit on 3283 manually-landmarked images across multiple databases. In training, the model learns an alignment based on the Lucas-Kanade algorithm. The exact pretrained model can be found at https://menpofit.readthedocs.io/en/stable/api/menpofit/aam/load_balanced_frontal_face_fitter.html. Unfamiliar face representations were obtained directly from the appearance representation of the model, defined as the principal component scores of a PCA solution taken over post-aligned pixels in the training images. Familiar face representations were obtained as the linear projection along the normal vector of an optimally separating hyper-plane obtained through linear discriminant analysis (LDA) trained on a set of training images of the familiarized identities, as in Kramer et al. (2017) and Kramer et al. (2018).

2.2. A deep convolutional neural network model of visual recognition

Convolutional neural networks (CNNs) are a broad class of machine learning models producing state-of-the-art performance in both computer vision (Krizhevsky et al., 2012) and in predicting brain responses in macaque (Yamins et al., 2014) and human visual cortex (Khaligh-Razavi & Kriegeskorte, 2014), as well as in predicting human behavioral similarity ratings. Rather than being hand-coded, CNNs learn representations from data, and most commonly, from associating data with appropriate labels through supervised learning. The defining characteristic of a CNN is the convolutional layer, which contains a set of filters with a fixed, restricted spatial receptive field, which are applied to all locations in the input (i.e., convolved with the input) to produce a set of feature maps (one per filter). The restricted spatial receptive field was inspired by this well-known property of V1 neurons first discovered by Hubel and Wiesel (1959) (LeCun, Bottou, Bengio, & Haffner, 1998; Zeiler & Fergus, 2014). While there is no known mechanism by which the brain explicitly computes convolution, the convolution operator has proven to be useful compared with non-convolutional locally-connected layers, due to a massive reduction in model complexity through an inductive bias (i.e., prior) that image features found in one location may be found in other locations. Most CNNs contain a pooling operation following each convolution that induces some spatial invariance to local shifts of the input data. Similarly to convolution, the pooling operation was inspired the discovery of "simple" and "complex" cells in primary visual cortex, where complex cells respond to the preferred stimulus over a larger range than simple cells, appearing to implement an OR operator over simple cells in nearby regions (Hubel & Wiesel, 1962). However, pooling is not a defining characteristic of CNNs and many state-of-the-art models forego pooling (Springenberg, Dosovitskiy, Brox, & Riedmiller, 2015), instead using stride > 1 in convolutional layers to progressively downsample feature maps as network depth increases (He, Zhang, Ren, & Sun, 2016). Deep convolutional neural networks (DCNNs) are simply CNNs constrained to contain at least two "hidden" (learned) layers of features between input and output layers. The hierarchical organization of DCNNs containing multiple hidden layers is broadly inspired by the hierarchical organization of the visual cortex (see, e.g. Felleman & Van Essen, 1991; Yamins & DiCarlo, 2016).

2.2.1. Architecture

We used the VGG-16 DCNN architecture in all of our simulations (Simonyan & Zisserman, 2015), shown in Fig. 2. This architecture achieved state-of-the-art performance in ImageNet object recognition at the time of its publication, and has also been demonstrated to be a highly effective architecture for face recognition (Parkhi et al., 2015). As seen, the network contains 5 convolutional "blocks", containing 2, 2, 3, 3, and 3 convolutions per layer, respectively, following by max pooling and positive rectification. For simplicity, we refer to the output

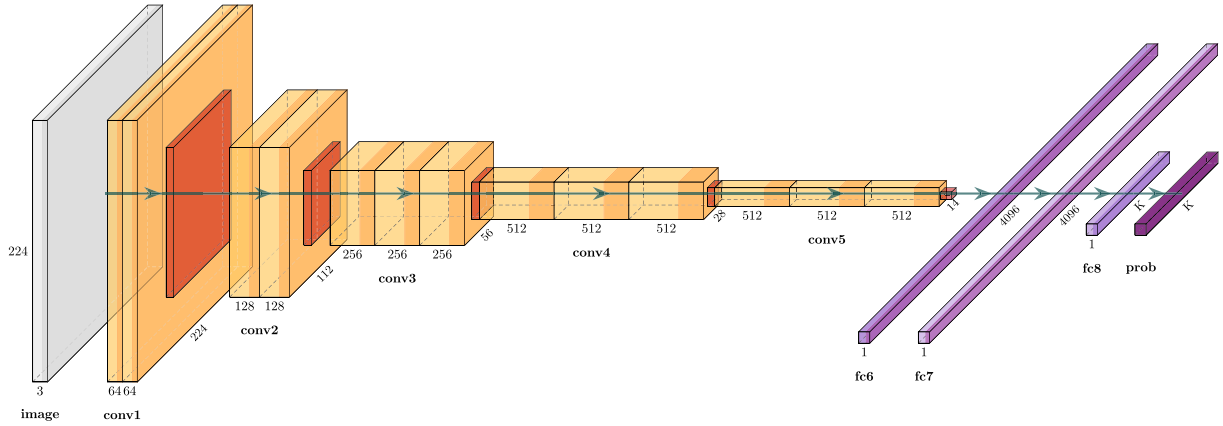

Fig. 2. Architecture of the VGG-16 deep convolutional neural network (DCNN) (Simonyan & Zisserman, 2015) (schematic produced using code at https://doi.org/10.5281/zenodo.2526396). The DCNN takes a 224 × 224 × 3 input image and transforms it in a hierarchical fashion to a set of output class probabilities. Convolutional blocks (conv1, conv2, …, conv5) contain 2 or 3 convolutional layers which do not downsample the spatial resolution of their input (i.e., stride of 1), followed by pooling. The convolutional blocks are following by three fully-connected layers, the last of which contains 1 unit per known identity. The activations in the last layer fc8 are transformed with the softmax function to a probability distribution, represented in layer prob. Operations are colored as following: convolution in light yellow, pooling in dark orange, linear transformation in light purple, rectification in dark yellow following convolution or purple following linear transformation, and finally softmax in dark purple. Arrows indicate the flow of information. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the *n*th convolutional block as conv*n*. The convolutional layers are followed by 3 fully-connected layers, where the first two (fc6 and fc7) are subject to rectification, and the last (fc8) is subject to a softmax operation, converting unit activations into an explicit probability distribution over known categories/identities. For simplicity, we refer to the rectified output of the first two fully-connected layers as fc6 and fc7, the pre-softmax output of the last fully-connected layer as fc8, and the post-softmax probability distribution as prob.

2.2.2. Pretraining

As a means of assaying the nature of pre-existing expertise needed for face recognition, we simulated three initial states of the network: 1) randomly initialized, 2) pretrained on objects, and 3) pretrained on faces. For pretraining on objects, we used a subset of 584 categories from the ImageNet large-scale image categorization challenge (Russakovsky et al., 2015) for which entry-level labels were available (although not used here) (Ordonez, Deng, Choi, Berg, & Berg, 2013). The images were divided into a training set (for adapting the network weights) and a validation set (for adapting the learning rate to avoid overfitting) as provided by the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC2012). For pretraining on faces, we used the VGGFace2 database (Cao et al., 2018), and selected a subset of identities that resulted in a close match in total number of images with our ImageNet database, and that did not overlap with the other databases we used for verification experiments. We manually created a validation set using 10% of the training data, such that the total images in the training and validation sets closely matched the numbers for the ImageNet set.

For each network, we used back-propagation to perform stochastic gradient descent in cross-entropy error, adapting the network weights to minimize the discrepancy between the identity activation generated by the network when presented with each image in the training set (using minibatches of 256 images sampled randomly without replacement each epoch) and the correct identity label for that image. An initial learning rate of 0.01 was allowed to decrease 4 times by a factor of 10 upon reaching a stable plateau in performance on the validation set, before performing early stopping at the 5th plateau, up to 50 epochs of training. All models converged within 50 epochs.

Pretraining was performed using the PyTorch neural network modeling package (Paszke et al., 2017). Code for setting up our image databases from the original ImageNet and VGGFace2 databases, for training the models, and for performing and visualizing the results of simulations will be made available on the GitHub page for this project (https://www.github.com/viscog-cmu/familiarity_sims).

2.3. Modeling familiarization through fine-tuning DCNNs

2.3.1. Experiments on Labeled Faces in the Wild

After pretraining, we performed fine-tuning of each network on a new set of face identities (familiarization), using the deep-funneled (aligned) images (Huang, Mattar, Lee, & Learned-Miller, 2012) of the Labeled Faces in the Wild (LFW) database (Huang et al., 2007). Identities with at least 18 images were selected and 10 images were held out for the test set. Verification was tested before, throughout, and after fine-tuning, where verification before fine-tuning corresponds to *unfamiliar* performance, and verification after fine-tuning corresponds to *familiar* performance. In the first epoch of fine-tuning (after testing unfamiliar verification), we appended new identity units to the existing ones, so that the network could learn to identify the new individuals. Here, fine-tuning refers to stochastic gradient descent back-propagated through the fully-connected layers only, with the weights of earlier convolutional layers held fixed. The network was not trained on verification explicitly, but rather only on identification of the new set of identities/categories. A fixed standard learning rate of 0.01, momentum of 0.9, and duration of 50 epochs were used, as there were too few images to permit the use of a validation set. As seen in Fig. 4A, the networks converged within this training period, and did not exhibit epoch-dependent over-fitting.

2.3.2. Experiments on VGGFace2 test set

In further experiments, we fine-tuned the face- or object-pretrained models on a new set of face identities in the test set of VGGFace2. We constructed several sub-databases from the set of 500 identities in order to simulate varying forms of experience with novel identities. We created datasets using 10, 50, or 100 identities, and set aside 100, 20, or 10 images per identity for verification testing, respectively, such that there were always 1000 verification images, and thus 499,500 verification pairs. We then set aside 10 images per identity for a validation set to control the learning rate during fine-tuning. Finally, from the remaining images available for each identity, we selected 1, 10, 50, 100, or 400 images to be trained on, such that the 100 image set contained all of the images in the 50 image set. In contrast to experiments using Labeled

Faces in the Wild, here we scheduled the learning rate exactly as we did in the pretraining phase, starting from a value of 0.01, and reducing it by a factor of 10 upon stable plateau of validation set accuracy, to a minimum of 10e−5 after which early stopping was performed. We set an extremely liberal maximum number of epochs of 1000 to ensure convergence of all models. Additionally, we varied the network layer at which fine-tuning started; whereas the LFW simulations started fine-tuning at layer fc6, here we examined fine-tuning beginning at the start of each convolutional block (conv1, conv2, …, conv5) and each fully connected layer (fc6, …,fc8).

2.4. Assessing identity perception through face identity verification

To perform face verification, we adopted a threshold-free similarity-based approach that can be applied to an arbitrary *m*-dimensional feature representation, including input images, shape-free texture components, and DCNN layer activations. First, given a set of feature responses $[x_1, \ldots, x_n]$ over images, the cosine distances between all test-set images were computed as $D_{i,j} = \cos(x_i, x_j)$ and then normalized to a range of [0,1]. A range of thresholds $\theta_k \in [0,1]$ was then used to compute a matrix of same/different judgments $Y_k = D > \theta_k$. The Y_k matrices were then compared to the true same/different matrix to compute true positive and false positive rates t_k and f_k. The vectors t and f constitute a Receiver-Operating-Characteristic (ROC) curve, and the area under the curve (AUC) was computed with numerical integration. Finally, we computed $d' = \sqrt{2}\,Z(AUC)$ where $Z(\cdot)$ is the inverse cumulative distribution function (CDF) of the standard normal distribution. This approach was applied to image pixels, the output of each layer of the VGG-16 network, and the appearance representations of the Active Appearance Model (AAM) before and after projection into an LDA space. Outputs were taken after pooling and rectification in each convolutional block (conv1, …,conv5), after rectification in the first-two fully-connected layers (fc6, fc7), and before and after the softmax operation in the final fully-connected layer (fc8, prob., respectively).

2.4.1. LFW

For LFW simulations, verification is reported across the entire set of 237,705 image pairs.

2.4.2. VGGFace2-test

For VGGFace2-test simulations, we divided the set of 499,500 image pairs into 20 non-overlapping segments. For each segment, we computed verification *d′* as described above, and report the mean calculated over segments. Error bars show 95% confidence intervals obtained from bootstrapping the sample of 20 values.

2.5. A human behavioral experiment on unfamiliar face verification

To provide a quantitative basis for evaluating the models, we carried out a behavioral experiment in which participants were presented with pairs of images of unfamiliar individuals and rated how likely it was that the images were of the same individual.

2.5.1. Participants

Twenty-three undergraduate students (15 female, mean age 20 yrs) from Carnegie Mellon University provided informed consent to participate in the experiment in exchange for course credit, in accordance with the protocol approved by the Institutional Review Board. Participants all reported normal or corrected-to-normal vision and were either Caucasian or raised in environments with many Caucasian individuals.

2.5.2. Stimuli

We used a version of VGG-16 pretrained on the original VGGFace dataset (Parkhi et al., 2015) to select difficult matching and non-matching image pairs in a dataset of unfamiliar Australian celebrities (Dunn, Ritchie, Kemp, & White, 2019). This dataset contains 40 identities with approximately 50 images per identity. Notably, there is no overlap in either VGGFace or the Australian Celebrities datasets with the VGGFace2 dataset that we used to train the DCNNs, as we explicitly removed all overlapping identities from the VGGFace2 dataset before any training. To select difficult image pairs, we selected the 1000 matching identity pairs with the largest cosine distance at the penultimate layer (fc7), and the 1000 non-matching identity pairs with the smallest cosine distance at the penultimate layer. For each participant, we randomly selected 200 of these pairs for each of the matching and non-matching conditions to yield 400 total trials per participant.

2.5.3. Procedure

Participants were seated approximately 60 cm from a computer screen and stimulus size was computed in terms of degrees of visual angle calibrated for either of 2 Dell LCD monitors. Participants completed a face verification experiment in which, on each trial, two face images were shown simultaneously to the left and right of the center of the screen for up to 10s. Participants were instructed to compare the similarity of the perceived identity by providing a 1–7 rating using the keyboard to indicate how similar the two faces were, where 7 indicated that the identities were definitely the same, and 1 indicated that the identities were definitely different. Following key press, a 500 ms interval ensued before the start of the next trial. Participants completed up to four sessions of the same 400-trial sequence, each taking approximately 15 min, performed back-to-back on the same day. No feedback was provided and thus, no information about the face identities was given to the participants. Prior to the start of each session, instructions were provided and the participant was given ten practice trials to acclimate to the experimental setup.

We then tested the verification ability of the face-trained and object-trained versions of VGG-16, as used in Methods 2.3.1 and Results 3.2–3.6, measuring unfamiliar verification performance of the pre-trained models, before any fine-tuning on the identities used in the behavioral experiment. Specifically, following earlier analyses, for each participant, we computed network performance by first computing cosine distances between *fc7* representations of each pair of trials for a given participant, extracting the area under the ROC curve, and converting this area to *d′* as the network performance for a given participant's set of trials.

2.6. An algorithm for combining perceptual and identity representations in face verification

To directly model the human task of performing face verification on images of unknown familiarity, we developed an algorithm that could use either perceptual or identity representations depending on their relative informativeness for a given face pair. The rationale for the rule is that face verification can be performed trivially if the two identities can be determined with confidence, and otherwise requires a more detailed perceptual comparison. We thus implemented a criterion *C* for determining whether to use identity or perceptual representations, based on the sum of output probability maximums over the two images. That is, identity representations are used if $\max(p_{ID}(x_1)) + \max(p_{ID}(x_2)) > C$, and otherwise, perceptual representations are used. The identity comparison can be made either by an explicit distance computation between probability distributions for the first and second image, or by verifying that the maximally active identity for each image is the same. We chose to implement the latter, as it makes a weaker commitment to the specific localist identity representation used in DCNNs, requiring only that whatever identity representation is used, it must be able to provide an index into the most probable identity; this seems to be a minimal requirement of any model of human identity representations. The criterion *C* was fit on a set of training images in order to maximize the area under the ROC curve for identity verification, for networks before and after familiarization. Fitting was

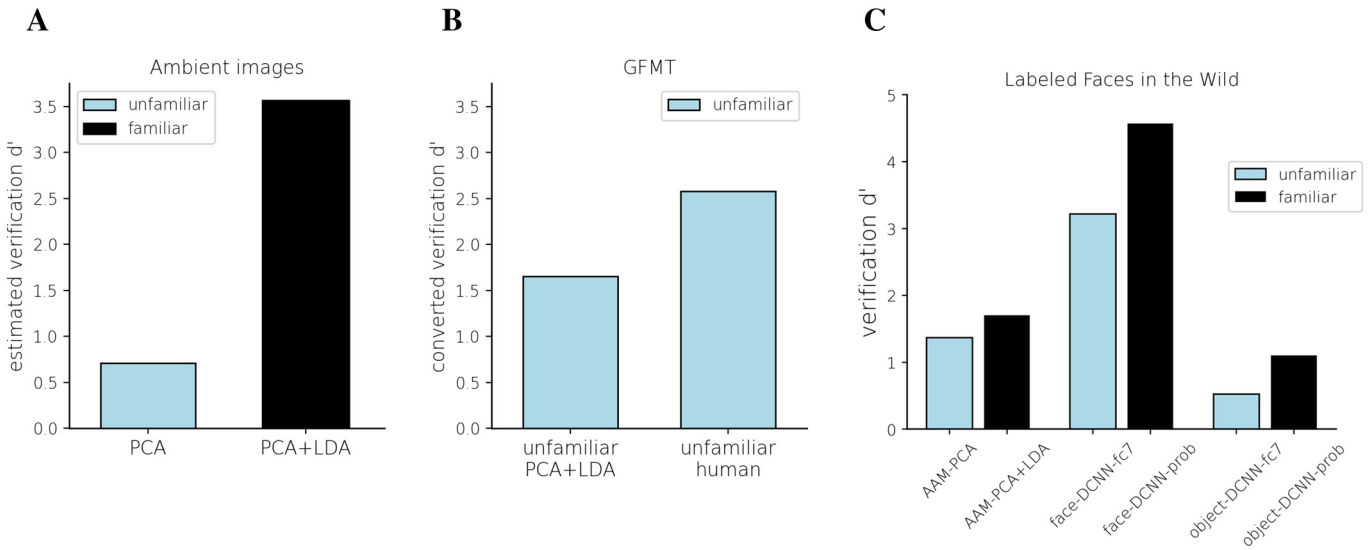

Fig. 3. The model of (Kramer et al., 2018) underestimates human-level unfamiliar face recognition and is outperformed by a face-trained, but not an object-trained DCNN. In A., we estimated *d′* from their distance measurements. In B., we converted their reported hit and false alarm rates to *d′*, which notably were not reported for the PCA model but only a PCA + LDA model fit on a separate set of identities from the ones tested. In C., we constructed an Active Appearance Model (AAM) similar to that used by (Kramer et al., 2018) but with fully-automated labeling of landmarks, and compared its performance on face verification of deep-funneled images of Labeled Faces in the Wild with a deep convolutional neural network model trained on faces (face-DCNN), or objects (object-DCNN), before and after familiarization.

performed with 5-fold cross-validation on the set of image pairs used in the behavioral experiment and associated simulation. Performance was computed before and after familiarization using the same face-trained network as in earlier experiments.

3. Results

3.1. Humans perform better than a shape-free texture model at unfamiliar face recognition

To get a better sense of the performance of the Kramer et al. (2018) model's ability to verify unfamiliar and familiar faces, we reanalyzed their main results. Given that the probabilities of hits and false alarms were not provided, we simulated them from a normal distribution estimate of the face distances provided in Fig. 11 of Kramer et al. (2018). We then performed an ROC analysis to compute *d′*. We estimated $d' = 0.707$ for the unfamiliar PCA space, and $d' = 3.56$ for the familiar PCA + LDA space, shown in Fig. 3A. These results demonstrate a severe deficit for unfamiliar face verification, in contrast to reasonably accurate verification of familiar face identities.

To understand whether their model accounted for human-level performance in unfamiliar face matching, we next reanalyzed their results for the Glasgow Face Matching Test. Notably, here the AAM PCA + LDA model was used to compute unfamiliar face recognition, using the LDA trained on the faces used in their previous experiment. From provided values of hit and false alarm rates p_{hit} and p_{fa}, we computed $d' = Z(p_{hit}) - Z(p_{fa})$, where $p_{hit} = p(\text{say same} \mid \text{same})$, $p_{fa} = p(\text{say same} \mid \text{different})$. The results of the model, $p_{hit} = .82$ and $p_{fa} = (1 - .78)$, yield $d' = 1.65$. For the human data on the same task, $p_{hit} = .92$, $p_{fa} = (1 - .88)$, we obtained $d' = 2.58$ (Fig. 3B). Kramer et al. (2018) did not provide analogous data for the PCA-only model; given the results of the earlier experiment, where *d′* was estimated as 0.707, we can only assume that this model performed more poorly than the unfamiliar PCA + LDA model. This suggests that some of the task-relevant (i.e., not purely statistical) variability is generic and not idiosyncratic, as it may be learned from an LDA solution on other faces.

3.2. A face-trained DCNN performs better than a shape-free texture model on both unfamiliar and familiar faces

We next tested an Active Appearance Model conceptually similar to the one used by Kramer et al. (2017) and Kramer et al. (2018), however with fully-automated (rather than semi-automated) landmark labeling; the model is otherwise identical, with a PCA being performed on the shape-free appearance, submitted to an LDA over training images for familiar faces only. We compared performance with that of a DCNN trained on either objects or faces. We assessed performance on a standard face verification baseline—Labeled Faces in the Wild (Huang et al., 2007)—and used the deep-funneled images, which have been computationally aligned using an approach based on deep learning (Huang et al., 2007), providing a helpful starting point for landmark labeling and alignment in the AAM.

As shown in Fig. 3C, both the AAM and DCNN models show a familiar face advantage. However, the face-trained DCNN performs substantially better with unfamiliar faces than the AAM model does even with familiar faces. When the task is unconstrained, the 2D automated AAM shows its weakness as a model of human face perception in comparison with the face-trained DCNN. In contrast, an object-trained DCNN performed worse at both unfamiliar and familiar face recognition, demonstrating that model complexity per se cannot account for the increase in performance of the DCNN relative to the AAM; rather, the specific experience of the face-trained network allows the network to achieve higher performance on both unfamiliar and familiar faces.

3.3. Face domain experience is necessary to learn to recognize new face identities robustly

Given that the face-trained, but not the object-trained DCNN performed well on verification of both unfamiliar and familiar individuals, we next sought to better understand how specific aspects of visual experience shaped DCNN identity perception. To do so, we pretrained two networks on roughly the same number of objects or faces (as in the last section), and randomly initialized a third network. We then fine-tuned each network to recognize new identities. The training and validation accuracy throughout learning for these three networks are shown in Fig. 4A.

Whereas the face-trained network quickly and robustly learned to

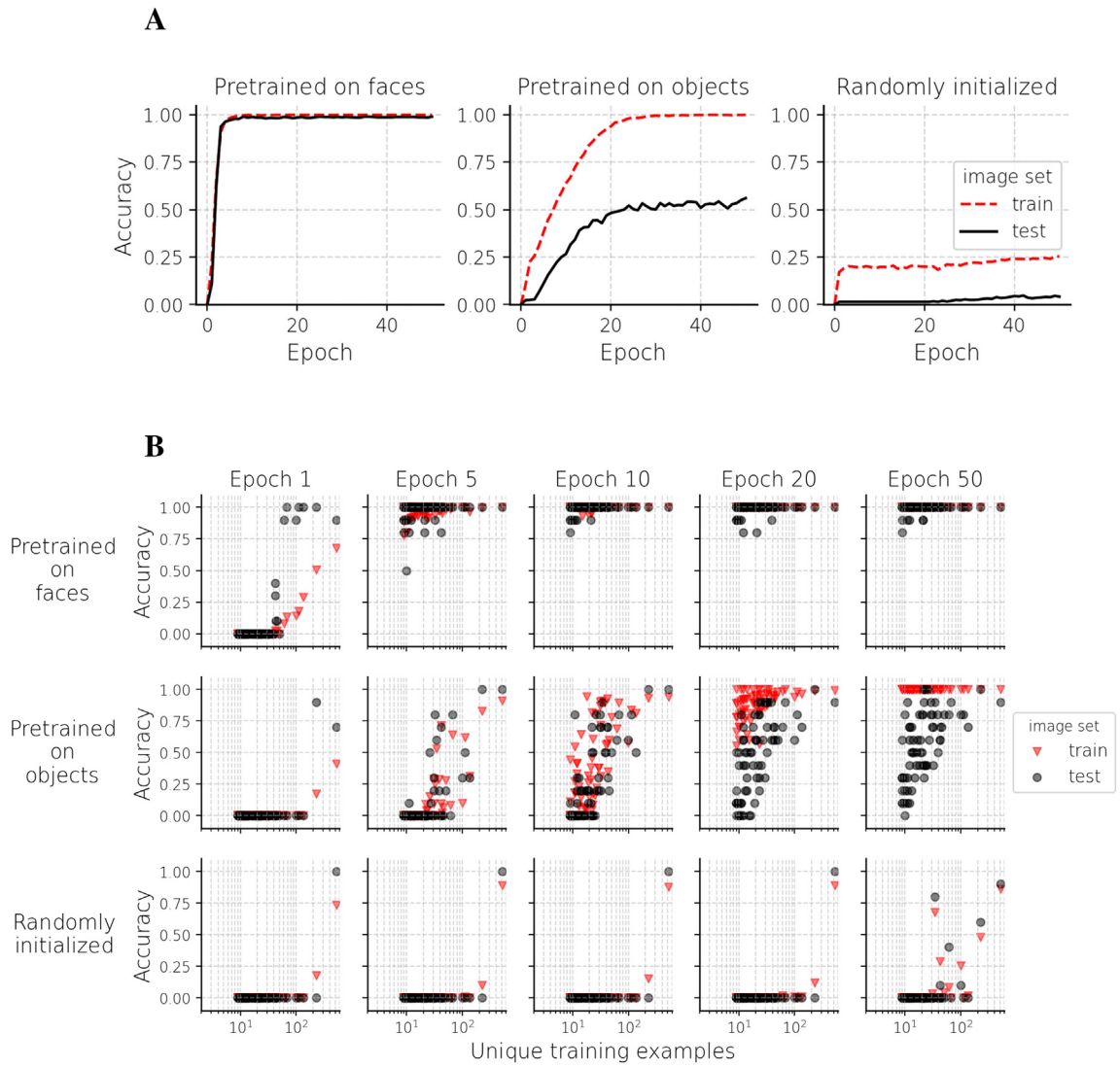

Fig. 4. Familiarizing three DCNNs with a novel set of identities. Networks pre-trained on faces, objects, or nothing (randomly initialized) were fine-tuned on novel identities in Labeled Faces in the Wild. In A., we plot performance of each network throughout training on training and held-out testing images collapsed across all new identities. In B., we plot accuracy for each new identity separately, vs. the number of unique training examples for each identity, shown for a representative sample of epochs throughout the course of familiarization.

categorize both seen and unseen images for the new identities, the object-trained network learned considerably more slowly, and generalization of its knowledge to held-out images was low. Despite the poorer performance of the object-trained network, it still performed better than the randomly initialized network. Performance plotted per identity across a range of per-identity experience is shown in Fig. 4B. These results demonstrate that the face-trained network learned to recognize novel faces robustly based on very little identity-specific data, achieving accuracy > =75% with as few as 10 training images per identity, and 100% accuracy with as few as 25 training images per identity. In contrast, the accuracy of the object-trained network depended much more strongly on the amount of experience, and remained below 90% even for the majority of individuals for which there were over 50 unique training images. These findings support our hypothesis that a high-fidelity featural description of faces—learned from a wealth of experience—is necessary to be able to group together highly variable ambient images of familiar faces for successful recognition. In other words, idiosyncratic experience on its own is not sufficient for good performance, especially when limited data is available for a given identity. The findings also support the notion that some of the features relevant to face identity recognition can be learned generically from pretraining on objects, but that this amount is relatively small, and object learning does not provide a sufficiently robust description for generalizing to unseen face images based on limited experience.

3.4. Face domain experience is necessary for robust unfamiliar verification performance

While recognition accuracy provides a good assessment of the ability of our networks to learn new faces, a different approach is needed to examine unfamiliar face processing abilities, and to compare them with familiar face processing abilities on the same images. We adopted the same approach as used in human studies: a face identity verification task in which the goal is to determine whether two faces are of the same identity. We computed a verification score based on the pairwise distances in a given representation, and applied this metric to image pixels and to each layer in the network, allowing us to determine which layer's representation most effectively discriminates between identities, and providing a measure of the extent to which performance is based on image or low-level statistics.

Verification performance for face- and object-trained networks, as well as for a randomly initialized network, are shown in Fig. 5, before

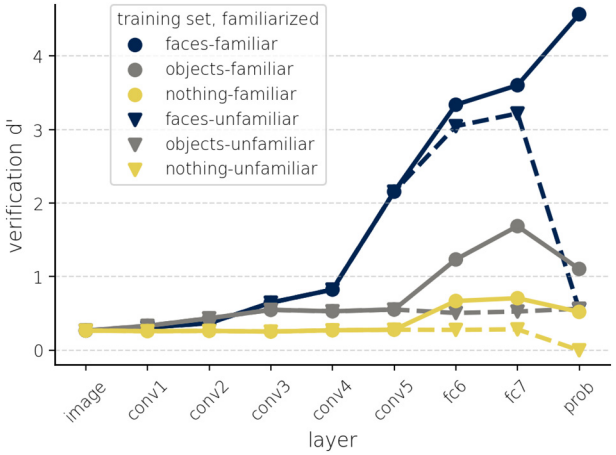

Fig. 5. Familiar and unfamiliar face verification by DCNNs with different training distributions matched in total number of images. Cosine distance matrices were computed over images for each layer separately, before and after familiarization. *Unfamiliar* representations were computed immediately following pretraining, and *familiar* representations were computed for the same images after 50 epochs of fine-tuning on a separate training set of images for the novel identities. d' was estimated with an ROC-based analysis (see Methods).

and after familiarization. First, examining performance of the face-trained network on a set of unfamiliar faces, shown in the dashed blue line, we can see that identity verification performance improves with increasing depth in the network ($d' > 3$ in fc7) up until the final, explicit probability layer, where performance sharply drops off ($d' < 1$ in prob). Notably, the network does not yet have an identity representation for the images it is verifying, thus, this result implies that the probability distribution over familiar faces is substantially less informative for unfamiliar face identity perception than the network's high-level perceptual representation in the penultimate layer. In contrast to the face-trained network, the untrained and object-trained networks show very little improvement with depth in the network, with all $d' < 0.7$. These results demonstrate that a similarity space developed through hierarchical computations in a network which has learned to recognize a large number of face identities naturally captures a substantial amount of identity-invariance for unfamiliar identities, placing same-identity images closer together than different-identity images to support good verification performance. These results challenge claims that within-identity facial image variability is entirely idiosyncratic; rather, the within-identity variability learned for familiar faces allows for large improvements in verifying unfamiliar face identities.

3.5. *Each fine-tuned layer shows a small familiarity advantage, but the final identity-based representation shows a qualitatively larger advantage*

One benefit of our method is that it allows us to compare unfamiliar and familiar performance on the same images of the same individuals. Fine-tuning was performed only on the weights into the fully-connected and output layers, and, as shown in Fig. 5, all of these layers show a small boost of familiarity. However, the large familiarity effect observed in many studies comparing human unfamiliar and familiar face processing is seen specifically at the output layer, which performed the best after familiarization but performed very poorly on unfamiliar faces. The randomly initialized and object-trained networks demonstrated some improvement with familiarity, but neither came close even to the *unfamiliar* performance of the face-trained network. These results suggest that familiarization of a sufficiently developed representation (i.e., one learned through prior experience with faces) allows the verification task to be performed on the basis of an identity representation, which, in the network, approaches orthogonality in the limit of perfect identification. In contrast, unfamiliar faces must be processed on the basis of more overlapping perceptual representations, which nonetheless untangle a substantial amount of invariance related to the perception of identity.

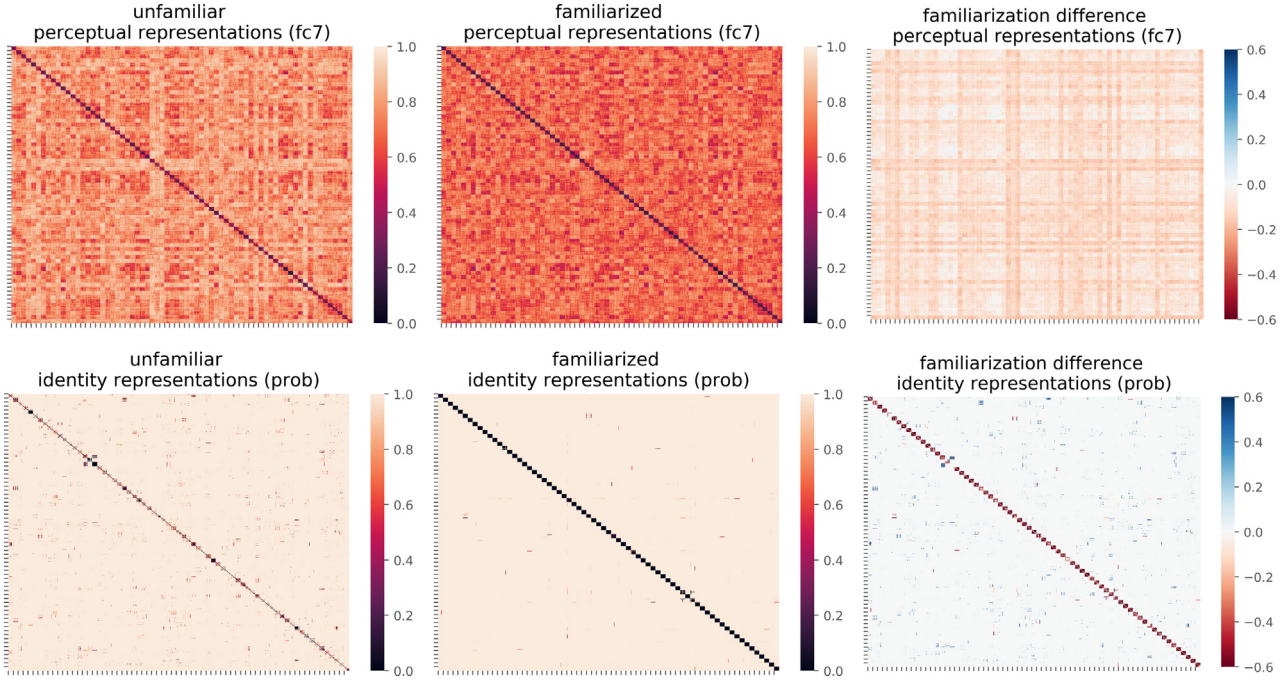

Fig. 6. Distance matrices of perceptual and identity representations in a face-trained DCNN before and after familiarization. Cosine distances were computed over images, with images sorted by identity (10 images per identity). The top row shows distances for the highest-level perceptual representations (fc7), and the bottom row shows distances for the softmax-probability identity representations (prob). The left-most column shows unfamiliar distance matrices, the middle column shows familiarized distance matrices, and the right column takes the difference (familiar – unfamiliar). (For best viewing of familiarization difference plots, the reader is referred to the online color version of the article.)

3.6. Familiarity makes images of the same individual look more similar, but minimally affects distances between different individuals

Another key aspect of familiarity effects in face perception is that, when asked to sort ambient images of multiple identities into separate groups, people are highly accurate for familiar faces, but for unfamiliar faces they make many misses (failures to group together same-identity images) yet few false alarms (failures to separate different-identity images) (Jenkins et al., 2011). This suggests that the human face-similarity space is largely sufficient to tell unfamiliar faces apart, but requires experience with individuals in order to group together highly variable images of the same person. To examine whether similar effects are seen in the network, we plotted the cosine similarity matrices for the penultimate fully-connected layer (fc7) and the output layer (prob) of the face-trained network which entered the ROC analyses used to generate *d'* values in Fig. 5, for familiar and unfamiliar faces, as well as the difference of these, shown in Fig. 6.

The unfamiliar distance matrix of fc7 shows a relatively accurate form, with low-distance clusters near the diagonal where the identities are the same, and mostly larger distances for different-identity pairs. Familiarization cleans up the non-matching areas of the distance matrix, resulting in greater overall uniformity across non-matching identities, despite greater overall similarity. However, familiarization does not affect the diagonal for the most part. In contrast, the output (prob) layer correctly places non-matching identities far apart before familiarization; however, it also fails to assimilate most pairs of matching identity images. After familiarization, this layer learns a near-perfect representation, with virtually all of the difference emerging as the assimilation of matching identity images. This result suggests that familiarity may provide a separate, identity-based similarity space which more readily groups together highly variable images of the same identity into a common representation than does the perceptual space (e.g., that of fc7).

3.7. Increasing face experience improves both perception of unfamiliar faces and learning of novel identities

An important aspect of the expertise account of unfamiliar (and familiar) face recognition is that performance is dependent not only on *some* experience with faces, but on *substantial* experience with faces. To test this, we trained the network on 1%, 10%, or 100% of the identities in the VGGFace2 database and tested unfamiliar and familiar verification, as well as accuracy throughout learning. Fig. 7A demonstrates increasing verification performance for both unfamiliar and familiar faces when training on a larger number of face identities. Fig. 7B shows that the increase in verification performance of unfamiliar and familiar faces in fc7, and familiar faces in the output identity layer, is roughly linear with the log of the fraction of identities that are pretrained. Further, the left panel demonstrates a qualitative performance difference on novel faces with an increasing number of pretrained faces. Specifically, whereas the network trained on all the face identities shows a sharp increase in familiarized performance at the output layer—the previously described recognition-based verification advantage—the networks trained on less data do not show this effect. The right panel demonstrates that this qualitative change is explained by a greater slope in the relationship between verification of familiar faces in the output layer compared to that of verification of unfamiliar and familiar faces in fc7. In sum, these results indicate that face training alone is not sufficient to achieve both high performance in unfamiliar face recognition, and a qualitatively large familiar face advantage at the output layer; rather, both effects are enhanced through *substantial* face experience, as present in the network trained on all the identities.

3.8. Perceptual learning is not necessary for the familiar face advantage

To determine whether the familiar face advantage depends on perceptual learning—that is, fine-tuning of perceptual features—rather than the mapping between perceptual features and output identity nodes, we fine-tuned either the full network, the fully-connected layers only (as in earlier simulations), or the output identity mapping only. For these simulations, we used the VGGFace2-test set in order to test a larger number of identities and images per identity. Here we used a validation set in order to train to convergence with a learning rate that decays upon plateau, as in pretraining, and then computed confidence intervals of verification performance by performing bootstrapping on a sample of non-overlapping segments of pairs of a different set of verification test images. Verification results for each network at the output and fc7 layers are shown in Fig. 8, when pretraining on objects versus faces, using either 10 or 100 newly familiar identities.

Examining the output layer verification performance of the face pretrained network (top left subplot in each figure panel), for each number of identities used in fine-tuning, the verification performance of the network when fine-tuned starting at conv1, fc6 or fc8 did not differ significantly (means within 95% confidence interval of fc8 performance). In contrast, irrespective of the number of identities used in fine-tuning, the output-layer performance of the object-pretrained network improved with fine-tuning beginning earlier in the network, especially for larger numbers of familiarization images per identity. These results demonstrate that perceptual learning is not necessary for a large advantage in verifying familiar faces. Rather, this advantage requires only learning to map highly variable perceptual representations of each identity to a common identity representation, making the

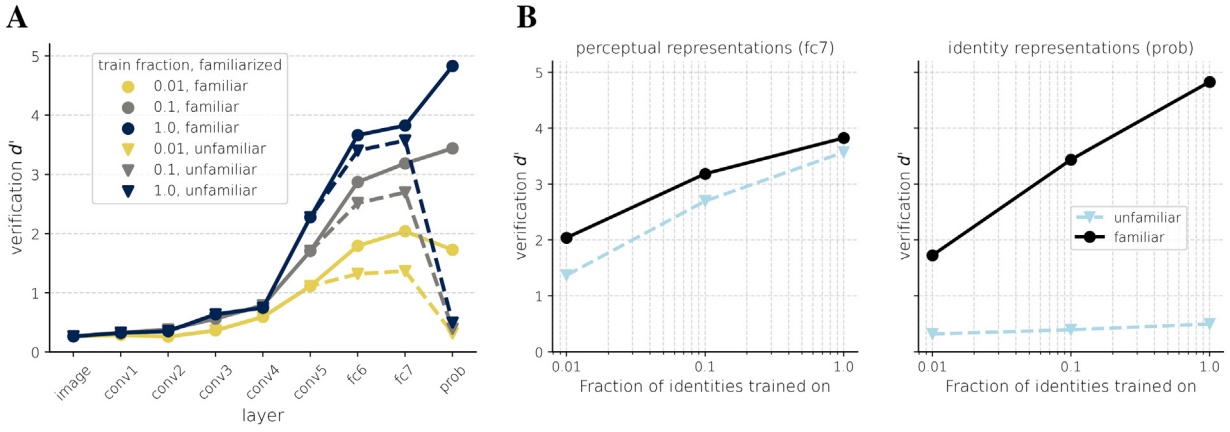

Fig. 7. Unfamiliar and familiar face verification measured in networks varying in the extent of face experience prior to familiarization. A fraction of 0.01, 0.1, or 1.0 of the total identities were used, and corresponding results for unfamiliar and familiarized face recognition are plotted as a function of layer (A) and fraction of identities (B) for high-level perceptual and identity representations. In B, a $\log_{10}$ X-scale is plotted against a linear Y-scale.

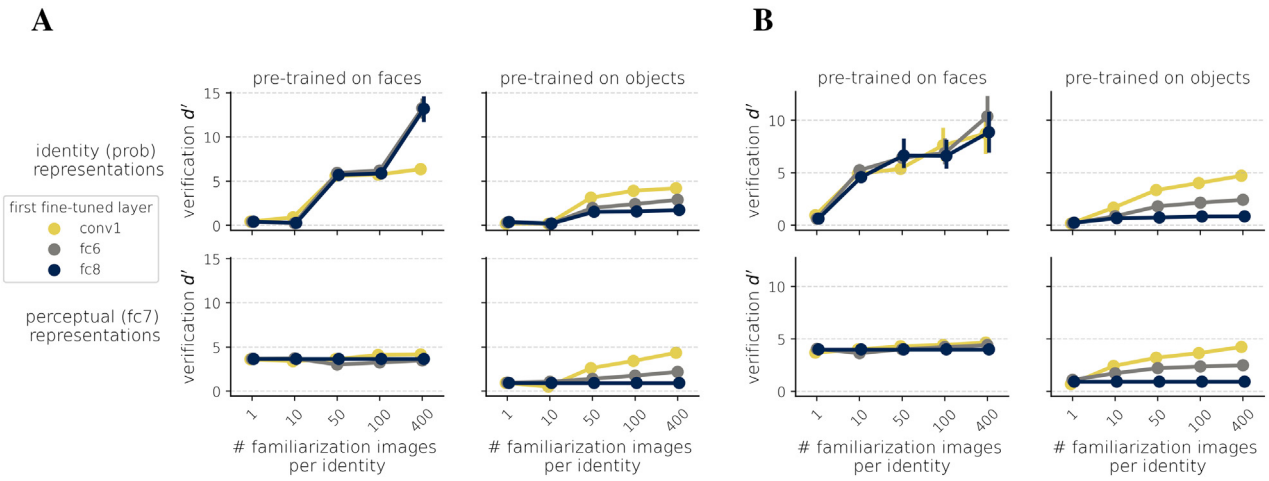

Fig. 8. The effect of experience with familiarized identites on familiar face verification, depending on the point in the network where fine-tuning begins: *conv1*, where the entire network is adapated; *fc6*, the type of fine-tuning used in the LFW experiment; and *fc8*, where only the final classifier layer is adapted. Fine-tuning on 10 identities is shown in A., and on 100 identities in B. Within both A. and B., columns vary the domain of pretraining (faces, objects), and rows vary the layer from which verification is computed (prob or fc7).

information relevant to the verification task explicit. Whether the perceptual representations have actually been adapted to the newly familiar faces is therefore irrelevant to this effect.

Further, across a broad range of training conditions, we replicated the earlier finding that prior experience with faces is necessary to achieve high verification performance, here using up to 400 familiarization images per 100 individuals. Thus, while such substantial familiarity permitted huge gains in face verification in the face expert network ($d' \approx 8$ compared to $d' < 5$ before familiarization; 100 identities, 400 images each, first fine-tuned layer fc8), performance of the most heavily familiarized object network ($d' \approx 5$; 100 identities, 400 images each, first fine-tuned layer conv1) only barely surpassed the performance of the unfamiliar face expert network, which achieved equivalent familiarized performance with only 10 images of the same individuals ($d' \approx 6$, identity representations, first fine-tuned layer fc8), and achieved greater performance with 50 images per identity, without modification to perceptual features ($d' > 6$, identity representations, first fine-tuned layer fc8). Thus, while familiarity may provide large gains in verification performance through the development of a confident identity representation, truly robust familiarized performance depends on having an expert perceptual face mechanism, which may be learned generically from other face identities.

3.9. Humans and a face-trained DCNN perform similarly on unfamiliar verification of challenging pairs

Finally, we sought to confirm that the DCNN simulation performs comparably to humans, to support our claim that its results are relevant to understanding the human perceptual system. We selected a set of challenging face pairs with an independent face-trained DCNN, and tested both humans and the face-trained and object-trained DCNNs used in the main LFW simulations. Human and DCNN performance is shown in Fig. 9. Human performance was evaluated for each of 4 sessions verifying the same pairs of unfamiliar face identities, whereas the networks were evaluated only once following pretraining. In the first session, humans showed a trend toward better performance than the DCNN ($t = 1.82$, $p = 0.0824$) at this unfamiliar verification task. Humans showed improvement over sessions even in the absence of feedback, and performance in each of the second through fourth sessions was significantly better than that of the DCNN (all $ps < 0.001$). Notably, the state of the DCNN that performed at this level of unfamiliar face recognition was sufficient to learn to perfectly verify the same images following familiarization on a separate set of images (see Fig. 10). In

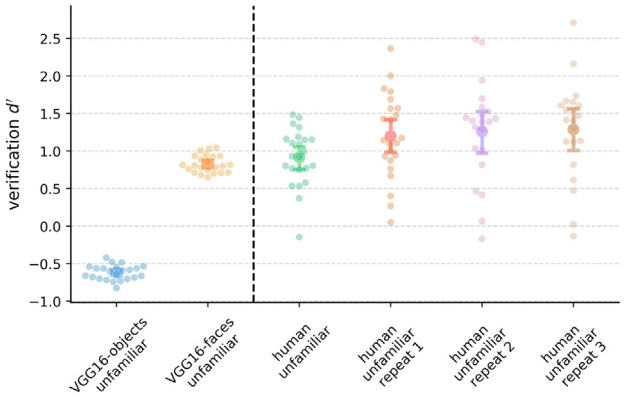

Fig. 9. Comparing human and DCNN unfamiliar verification performance on a challenging set of face image pairs from a dataset of Australian local celebrities. Unfamiliar verification performance of VGG-16 DCNN pretrained on objects or faces is shown on the left. Humans performed the same verification task 4 times, and performance is plotted for both the first totally unfamiliar session, and each of the three repeat sessions.

demonstrating similar performance of humans and the DCNN at unfamiliar verification—and if anything, slightly better performance of humans—these results validate our use of this model in making a claim for perceptual expertise underlying both unfamiliar and familiar faces.

3.10. A simple cognitive rule allows for optimal combination of perceptual and identity information in the service of identity verification of images of arbitrary familiarity

Two potential critiques of our work thus far are 1) no single representation in the model performs similarly to humans in both unfamiliar and familiar face recognition; rather, human unfamiliar face recognition is modeled well by the highest perceptual layer (fc7) whereas the benefits of familiarity in verification are seen when examining the explicit probability layer (prob), and 2) it assumes a localist identity representation which seems at odds with the distributed semantic, biographical, and episodic representation many believe constitutes a human identity representation. Regarding the first point, our claim is that humans have access to both forms of representations and can easily determine the more useful one for a pair of images, a process which is likely automated by the confidence of identity recognition of each individual. We developed a simple cognitive rule-

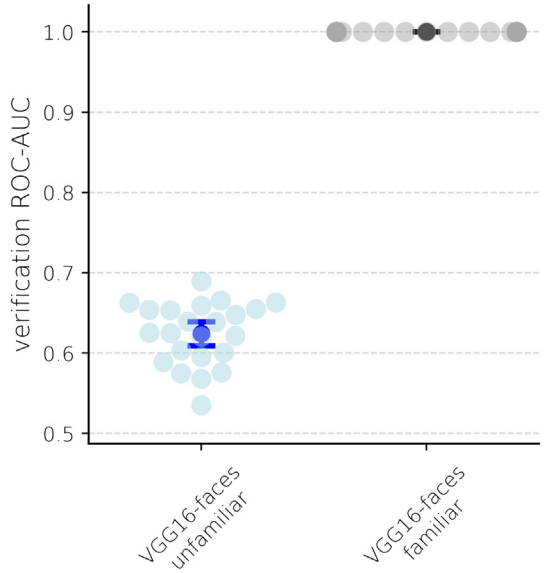

Fig. 10. Verification using a cognitive rule that flexibly determines whether to use perceptual or identity representations. Results are shown for the face-pretrained network before and after familiarization, evaluated on the same images as in the behavioral experiment. We plot area under the ROC curve (ROC-AUC) here instead of d', as $d' = \infty$ for the familiarized network on this small set of image pairs.

based model to demonstrate how this might be done. We performed this analysis as a proof of concept that perceptual and identity representations can be optimally combined on the fly for human-like unfamiliar and familiar face identity verification, and hold no strong commitment to the specifics of the rule. As described in Methods 2.6, the intuition is straightforward: given two images, if the system is highly confident about the identity of either image, or relatively confident about both, it should decide based on whether the identities match; otherwise, it should perform the task by comparing the perceptual representations elicited by each face. This rule also addresses the second point, as it softens the assumption of a localist identity representation, rather requiring only that whatever identity representation is used, it yields a confidence value for and index to the most probable known identity for a given face image.

Using the same images as in the human behavioral experiment, as shown in Fig. 9, we implemented the cognitive rule and evaluated it before and after familiarization. Results are shown in Fig. 10. As expected, this unitary mechanism recognizes difficult pairs of unfamiliar faces at above-chance levels and produces a sharp benefit for familiar faces, reaching perfect performance on the small set of images used.

4. Discussion and conclusions

The hypothesis that humans are not experts at unfamiliar face recognition has gained considerable attention recently, in part due to a host of studies demonstrating deficits on challenging face verification tasks for unfamiliar versus familiar faces. But the question remains: Are we as poor at unfamiliar face recognition as has been suggested by some (Young & Burton, 2018), or are we perceptual experts for faces regardless of familiarity (Rossion, 2018; Sunday & Gauthier, 2018)? Is the majority of face variability idiosyncratic to each identity (Kramer et al., 2018) or more generic across individual face identities? If the variability is more generic, do humans simply fail to learn this variability in the service of unfamiliar face perception?

Kramer et al. (2018) have recently put forth a model of face perception to explain both the robust human performance in recognizing familiar faces as well the poorer performance on unfamiliar faces. However, although this model does exhibit a familiar face advantage, we demonstrated that it substantially underestimates human-level performance on unfamiliar faces, and in so doing, may have led to a false rejection of human perceptual expertise in unfamiliar face recognition (Young & Burton, 2018). We showed that a conceptually similar active appearance model, using fully-automated computational face alignment, performed much more poorly than a deep convolutional neural network (DCNN) at both unfamiliar and familiar face recognition, but only if the DCNN was trained for face recognition (Fig. 3). Strikingly, the performance of the DCNN on *unfamiliar* faces was substantially better than the AAM model's performance on familiar faces. In contrast, the accuracy of a face-trained DCNN on unfamiliar and familiar face verification was comparable to that of humans (Fig. 9). The high performance of the DCNN justified an exploration of it in greater detail in order to better understand the experience necessary for human-level unfamiliar face identity verification, and why humans display a sharp improvement in the verification of familiar identities.

The high performance of the face-trained DCNN on unfamiliar face recognition suggests that a large share of the variability in face images that is relevant to recognition is generic across faces. The boost in performance on familiar faces confirmed the result of Kramer et al. (2018) that an additional share of variability is idiosyncratic and must be learned for each face. The next question is how much of the necessary generic variability is specific to faces versus being potentially learnable from other natural object categories? The answer to this question is important for our understanding of unfamiliar face recognition: if high performance is dependent on experience with faces and cannot be achieved through generic experience with natural object images, it reinforces the idea that unfamiliar face recognition performance is the product of a specific learned expertise with faces per se.

To answer this question, we compared the face-trained DCNN to one trained on a size-matched database of objects and to a randomly initialized one. The object-trained DCNN performed slightly better than the randomly initialized network at unfamiliar face recognition and much better at familiar face recognition. However, it performed strikingly worse than the face-trained network at both unfamiliar and familiar face recognition, with even familiarized recognition substantially worse than the unfamiliar recognition of the face trained network. This result suggests that the majority of generic identity-preserving face variability is not also generic across a broader class of natural objects. Further, these results demonstrate that a representation trained to capture generic face variability through experience with face images is important for human-level unfamiliar face verification, for learning to recognize new familiar individuals, and for robustly verifying those familiarized individuals.

Given strong evidence that the *domain* of experience is important for developing expertise in both unfamiliar and familiar face recognition, we next asked how performance depended on the *extent* of experience recognizing faces. To do so, we varied the number of identities seen in pretraining. We found that performance in verification of *both* unfamiliar and familiar faces consistently improved with experience over an increasing number of identities. This result strengthens the expertise account of general face recognition, and weakens the claim that faces must be learned one at a time (Young & Burton, 2018). Lastly, we extended our simulations to a broader range of familiarization conditions, varying both how many layers of the network were allowed to be fine-tuned, as well as the specific number of identities and images per identity shown to the network. In doing so, we found that the features learned from a large set of face identities were sufficient to learn an accurate mapping over multiple examples to arbitrary new identities, and this mapping (learned through familiarization) produced robust familiarity advantages regardless of whether perceptual learning took place. Thus, we argue that perceptual learning is not necessary for the familiar face advantage. Although, in humans, perceptual learning (Collins & Behrmann, 2020) and associated neural changes (Collins, Robinson, & Behrmann, 2018; Dobs, Isik, Pantazis, & Kanwisher, 2019)

may very well occur for familiar faces, we argue that these effects are not necessary to develop a familiar face advantage in identity perception. Rather, the ability to map an image of a familiar face to a pre-existing identity/biographical representation provides a more effective means of matching faces than a perceptual comparison, and this identity-based matching is possible only for familiar faces. However, we want to be clear that it is possible that familiarization could produce greater perceptual orthogonalization of different identities through recurrent dynamics or vector length normalization (Liu et al., 2017; Ranjan, Castillo, & Chellappa, 2017)[1]; however, our findings show that this is not necessary, likely because the same features are relevant across a wide range of faces. This idea has been supported by Abudarham et al. (2019) who show that the same critical perceptual features are used by humans for recognizing familiar and unfamiliar faces, as well as by a deep neural network. Thus, we reject the claim that each face identity must be learned from scratch (Young & Burton, 2018); rather, a lifetime of face learning allows for new faces to be rapidly familiarized based on little to no perceptual learning.

Finally, we devised a task to compare human performance with that of a DCNN on challenging image pairs chosen by an independently trained DCNN. Our results showed that the network was only slightly worse than humans at difficult unfamiliar face recognition. Given that DCNNs have recently been shown to perform on par with trained human facial examiners, which both performed better than untrained students (Phillips et al., 2018), the gap in performance seen here may be a result of using another DCNN—albeit one trained on entirely different face identities—to select hard images. Given the common architecture of the networks, it is possible that some image pairs are more difficult for these DCNNs than for humans, and vice-versa, and that some untested image pairs would be more difficult for humans than for the DCNN. Further, while our approach guaranteed that unfamiliar faces were totally unfamiliar to the network, some participants may have had some familiarity with some of the local Australian celebrities used as unfamiliar identities, which would give the human population a further advantage. Finally, the ability of humans to make multiple fixations and perform detailed featural comparisons between image pairs provides a further advantage not available to the network, which computed a single perceptual representation for each image, independent of the other image it was compared to. A more fine-grained analysis of the *differences* between humans and DCNNs in face recognition is an important open question for further research. Intriguingly, humans improved substantially over sessions of repeated trials even without feedback, suggesting that they were able to learn about the unfamiliar faces in the absence of any explicit cues to identity. This result suggests that, while pre-existing identity-specific representations are unavailable for unfamiliar individuals, they may be rapidly constructed without explicit cues to identity, perhaps by building episodic representations throughout the course of the behavioral experiment, which can be mentally clustered into a set of possible identities. In so doing, humans may integrate guesses about identity clusters with comparisons of perceptual representations to improve performance. An integration of identity information with existing perceptual representations may also explain how learning new faces creates apparent changes in an existing face space that are selective for the learned identities (Collins & Behrmann, 2020).

Given the results of our simulations and behavioral experiments, we return to the question of expertise in human unfamiliar face recognition. Let us consider an analogy of a professional golfer. As professional golfers routinely play the most challenging golf courses in the world, it is not hard to imagine that they get better at playing these courses with experience—whether via examination of a map, engagement with word-of-mouth advice from local experts, or experience through gameplay. In this sense, the professional golfer must learn the idiosyncrasies of each course—just as Young and Burton (2018) argue humans must do for novel familiar faces. Imagine now that a cognitive psychologist interested in golf expertise asks professional golfers to play a round of 18 holes on 6 courses that they have never played, and prevents them from seeing a map or gaining any other knowledge of the course besides what they experience as they play it. After this first round, they are allowed to study the course in depth, and play as many practice rounds as they can in a week before coming in for a final test of their performance. It would not be surprising if the pro golfers performed better with experience, as a result of learning the course's geography and other relevant details which were entirely unknowable in the first round. Imagine further that untrained golfers, amateur golfers, and professional tennis players (who are not also professional golfers) are brought in for the same experiment. It again would be unsurprising if the final, familiarized score of these less-skilled golfers improved from the earlier baseline, but yet, it *would* be surprising if their score reached even the baseline performance of the professional golfer. Given this, is the professional golfer a golf expert, or only an expert at familiar courses? In our view, the golfer's expertise is evident both in their high baseline performance level relative to other well-defined groups, as well as in their ability to rapidly learn the idiosyncrasies of the new course in the service of maximizing performance. We believe that the case of face expertise parallels the case of the golfer. Like golf courses, faces have idiosyncrasies which must be learned. But these idiosyncrasies do not represent the dominant variability of faces (which can be learned from other faces). By learning this generic variability, unfamiliar face perception is enhanced, and idiosyncratic variability can be learned rapidly in the service of familiar face perception.

[1] See Supplementary material for more detail.

CRediT authorship contribution statement

Nicholas M. Blauch: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, Writing - review & editing. **Marlene Behrmann:** Conceptualization, Methodology, Project administration, Supervision, Writing - review & editing. **David C. Plaut:** Conceptualization, Methodology, Project administration, Supervision, Writing - review & editing.

Acknowledgements

N.M.B. was supported by NIH training grant 5T90DA022762-13 to Robert E. Kass, and a Carnegie Mellon Neuroscience Institute Presidential Fellowship.

Appendix A. Supplementary data

Data from all simulations is made available at https://doi.org/10.1184/R1/12275381, and code relevant to reproducing the simulations and plotting the results is made available at https://github.com/viscog-cmu/familiarity_sims. Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104341.

References

Abudarham, N., Shkiller, L., & Yovel, G. (2019). Critical features for face recognition. *Cognition, 182*, 73–83. URL https://doi.org/10.1016/j.cognition.2018.09.002.

Abudarham, N., & Yovel, G. (Feb 2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision, 16*(3), 40. URL http://jov.arvojournals.org/article.aspx?doi=10.1167/16.3.40.

Antonakos, E., Alabort-I-Medina, J., Tzimiropoulos, G., & Zafeiriou, S. P. (2015). Feature-based lucas-kanade and active appearance models. *IEEE Transactions on Image Processing, 24*(9), 2617–2632. URL www.menpo.org.

Behrmann, M., & Avidan, G. (2005). Congenital prosopagnosia: Face-blind from birth. *Trends in Cognitive Sciences, 9*(4), 180–187.

Bothwell, R. K., Brigham, J. C., & Malpass, R. S. (Mar 1989). Cross-racial identification. *Personality and Social Psychology Bulletin, 15*(1), 19–25. URL http://journals.sagepub.com/doi/10.1177/0146167289151002.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P.

(1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*(4), 339–360. URL http://doi.apa.org/getdoi.cfm?doi=10.1037/1076-898X.5.4.339.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. *International conference on automatic face and gesture recognition.* URL http://www.robots.ox.ac.uk/.

Collins, E. C., & Behrmann, M. (2020). Exemplar learning reveals the representational origins of expert category perception. *Proceedings of the National Academy of Sciences, 20*(117), 11167–11177. https://www.pnas.org/content/117/20/11167.

Collins, E. C., Robinson, A. K., & Behrmann, M. (2018). Distinct neural processes for the perception of familiar versus unfamiliar faces along the visual hierarchy revealed by EEG. *NeuroImage, 181*(June), 120–131. URL https://doi.org/10.1016/j.neuroimage.2018.06.080.

Cootes, T., Edwards, G., & Taylor, C. (1998). Active appearance models. *European Conference on Computer Vision, 2*, 484–498. URL https://www.cs.cmu.edu/~efros/courses/AP06/Papers/cootes-eccv-98.pdf.

Cootes, T., Taylor, C., Cooper, D., & Graham, J. (1995). Active shape models-their training and application. *Tech. Rep* (pp. 1). . URL https://pdfs.semanticscholar.org/f731/b6745d829241941307c3ebf163e90e200318.pdf.

Damasio, A. R., Damasio, H., & Van Hoesen, G. W. (1982). Prosopagnosia: Anatomic basis and behavioral mechanisms. *Neurology, 32*(4), 331–441.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Tech. Rep* (pp. 2). . URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.555.3596&rep=rep1&type=pdf.

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications, 10*(1), https://doi.org/10.1101/442194 URL.

Dunn, J. D., Ritchie, K. L., Kemp, R. I., & White, D. (Jul 2019). Familiarity does not inhibit image-specific encoding of faces. *Journal of Experimental Psychology: Human Perception and Performance, 45*(7), 841–854.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*(1), 1–47.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience, 3*(2), 191–197. URL http://www.nature.com/doifinder/10.1038/72140.

Gauthier, I., & Tarr, M. J. (Jun 1997). Becoming a "Greeble" expert: Exploring mechanisms for face recognition. *Vision Research, 37*(12), 1673–1682. URL https://linkinghub.elsevier.com/retrieve/pii/S0042698996002866.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999a). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience, 2*(6), 568–573. URL https://doi.org/10.1038/9224.

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999b). Activation of the middle fusiform "face area" increases with expertise in recognizing novel objects. *Nature Neuroscience, 2*(6), 568–573. URL http://www.biac.duke.edu/education/courses/spring03/cogdev/readings/I.Gauthieretal(1999).pdf.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE computer society conference on computer vision and pattern recognition 2016-Decem* (pp. 770–778). .

Huang, G. B., Mattar, M. A., Lee, H., & Learned-Miller, E. (2012). *Learning to align from scratch.* NIPS. URL http://www.openu.ac.il/home/hassner/data/lfwa/.

Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep., University of Massachusetts, Amherst. URL. http://vis-www.cs.umass.edu/lfw/.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology, 148*, 574–591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology, 160*(1), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837.

Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (Dec 2011). Variability in photos of the same face. Cognition 121 (3), 313–323. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027711002022.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience 17 (11), 4302–11. URL. http://www.ncbi.nlm.nih.gov/pubmed/9151747.

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences, 361*(1476), 2109–2128.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*(11).

Kramer, R. S., Young, A. W., & Burton, A. M. (Mar 2018). Understanding face familiarity. Cognition 172, 46–58. URL https://linkinghub.elsevier.com/retrieve/pii/S0010027717303074.

Kramer, R. S., Young, A. W., Day, M. G., & Burton, A. M. (2017). Robust social categorization emerges from learning the identities of very few faces. *Psychological Review, 124*(2), 115–129.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems,* 1–9.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). SphereFace: Deep hypersphere embedding for face recognition. In: Proceedings - 30th IEEE conference on computer vision and pattern recognition, CVPR 2017. Vol. 2017-Janua. pp. 6738–6746. URL. https://github.com/wy1iu/sphereface.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition, 34*(4), 865–876.

Ordonez, V., Deng, J., Choi, Y., Berg, A. C., & Berg, T. L. (2013). From large scale image categorization to entry-level categories. In: Proceedings of the IEEE international conference on computer vision. pp. 2768–2775. URL. http://www.cs.unc.edu/~vicente/files/entrylevel.pdf.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. Proceedings of the British machine vision conference 2015 (section 3), 41.1–41.12. URL. http://www.bmva.org/bmvc/2015/papers/paper041/index.html.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Facebook, Z. D., ... Lerer, A. (2017). *Automatic differentiation in PyTorch.* NIPS1–4.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., ... O'Toole, A. J. (Jun 2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America, 115*(24), 6171–6176.

Ranjan, R., Castillo, C. D., & Chellappa, R. (2017). L2-constrained Softmax loss for discriminative face verification. ArXiv (1703.09507). URL. http://arxiv.org/abs/1703.09507.

Rossion, B. (2018). Humans are visual experts at unfamiliar face recognition. *Trends in Cognitive Sciences, 22*(6), 471–472.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. International Journal of Computer Vision 115 (3), 211–252. URL. https://www.cv-foundation.org/openaccess/content_cvpr_2015/app/1A_026_ext.pdf.

Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition.* International Conference on Learning Representations. URL http://www.robots.ox.ac.uk/.

Springenberg, J. T., Dosovitskiy, A., Brox, T., Riedmiller, M., Dec 2015. Striving for simplicity: The all convolutional net. In: 3rd international conference on learning representations, ICLR 2015 - Workshop track proceedings. URL http://arxiv.org/abs/1412.6806.

Sunday, M. A., & Gauthier, I. (2018). Face expertise for unfamiliar faces: A commentary on Young and Burton's "are we face experts?". *Journal of Expertise, x*(x), 1–7.

Tarr, M. J., & Gauthier, I. (2000). FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience, 3*(8), 764–769.

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience, 31*, 411–437.

Valentine, T. (Nov 1988). Upsidedown faces: A review of the effect of inversion upon face recognition. British Journal of Psychology 79 (4), 471–491. URL http://doi.wiley.com/10.1111/j.2044-8295.1988.tb02747.x.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature Neuroscience 19 (3), 356–365. URL. http://www.nature.com/doifinder/10.1038/nn.4244%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26906502.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences of the United States of America 111 (23), 8619–24. URL. http://www.ncbi.nlm.nih.gov/pubmed/24812127%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4060707.

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences, 22*(2), 100–110.

Young, A. W., Hellawell, D., & Hay, D. C. (Dec 2013). Configurational information in face perception. Perception 42 (11), 1166–1178. URL http://journals.sagepub.com/doi/10.1068/p160747.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *EECV. 1. EECV* (pp. 818–833).

Supplementary material:
Computational insights into human perceptual expertise for familiar and unfamiliar face recognition

Nicholas M. Blauch, Marlene Behrmann, David C. Plaut
Carnegie Mellon University

Data from all simulations is made available at https://doi.org/10.1184/R1/12275381, and code relevant to reproducing the simulations and plotting the results is made available at https://github.com/viscog-cmu/familiarity_sims.

1 Learning vector-length normalized face representations: effects on perceptual face verification for familiar and unfamiliar faces.

Normalizing the vector-length of face representations has proven to be an effective method for achieving representations which are more separable across identity using a cosine (i.e. angular) distance metric[3]. Very similar work has reformulated the softmax loss as an angular softmax loss with margin, to encourage angular separation of features lying on a hyper-sphere manifold [1]. We implemented the method of [3], where the vector length of `fc7` representations is normalized and scaled to a value of $\alpha = 40$ for each image. As shown in the paper, the specific value of α is not crucial so long as it is sufficiently large to allow enough surface area to ensure separability of the classes. This approach was used in the best performing model in [2].

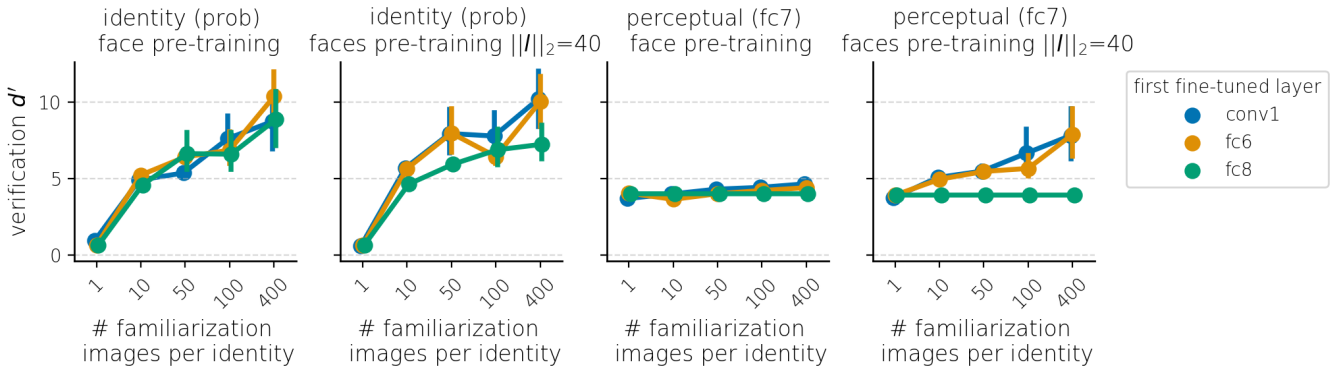

Figure S1: Comparing face verification from identity and perceptual representations in face-pretrained networks.$||l||_2 = 40$ networks were additionally fine-tuned (fc6-fc8) with the constraint that the vector length of fc7 representations was normalized to 40 for all images. Whereas strong improvements in verification with familiarity are seen only in the output layer for the standard face-pretrained network, in the l2-normalized network, familiarity gains are also seen in the perceptual representations. However, the best verification is still seen in the output identity representations, and this performance is equivalent to the standard network. Similarly, performance in the perceptual representation layer (given by the green line in two righmost plots) is similar for unfamiliar faces.

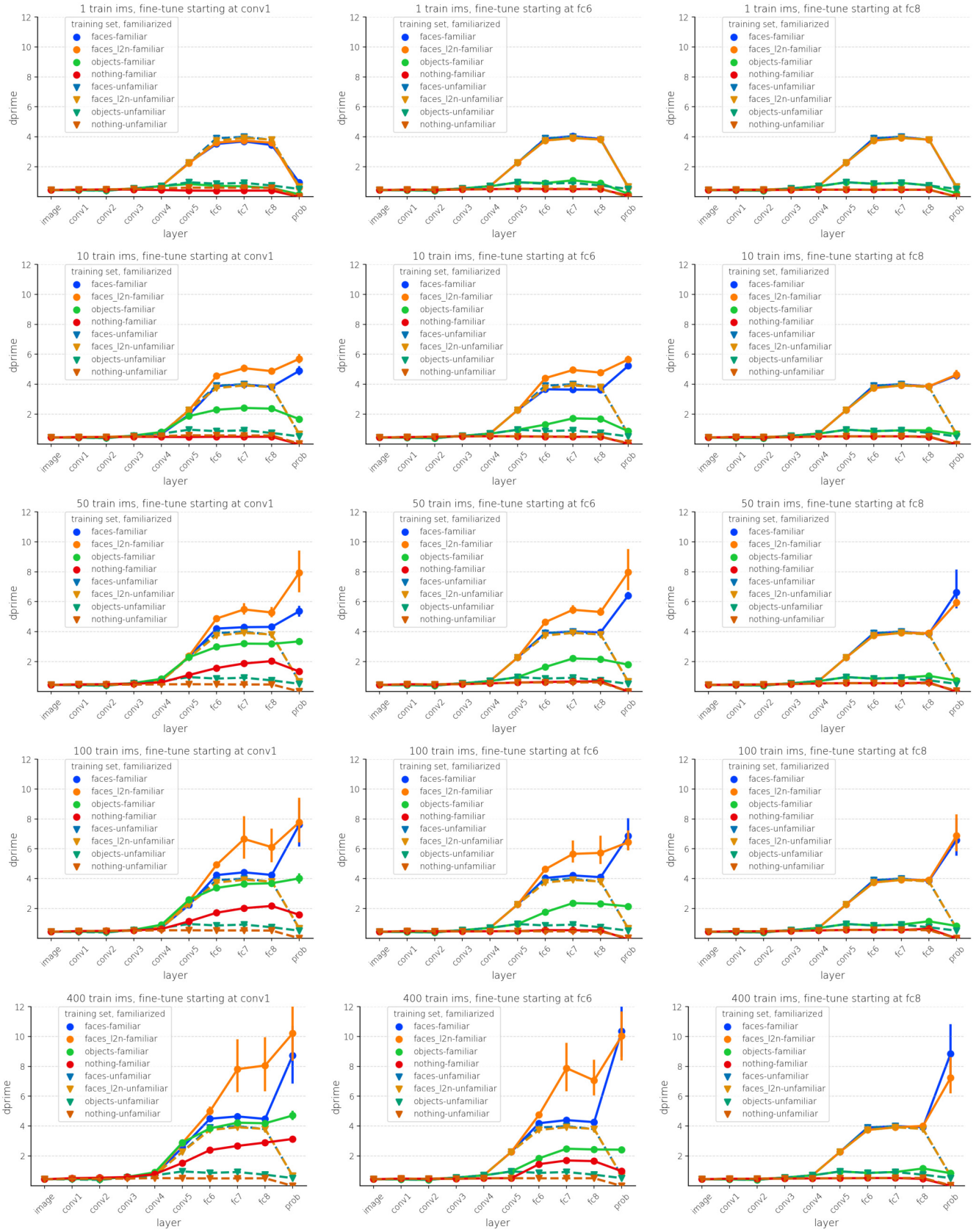

Figure S2: Sub-figures show an increasing number of training images per identity from top to bottom, with starting fine tuning later in the network from left to right. The l2-normalized network, shown in orange, show a large familiar face advantage that extends into perceptual representations. However, unfamiliar performance is equivalent across l2-normalized and standard softmax face-trained networks.

Normalizing the vector length of perceptual representations allows for a larger degree of orthogonalization of face identities in the perceptual layer, allowing for gains in verification performance in the fully-connected layers for familiar identities that are qualitatively similar to the gains seen only in the output probability layer in the standard softmax face-trained network. Unfamiliar performance is unaffected by this change. The results demonstrate that the standard face-trained DCNN with learned readout from fixed perceptual representations performs approximately as well as the end-to-end fine-tuned l2-normalized face-trained DCNN. But, the l2-normalized face-trained DCNN demonstrates that qualitatively large gains in perceptually-based verification are possible given learning from vector-length normalized face representations. Given the similar verification performance between perceptual representations in the l2-normalized network and identity representations in the standard network, it appears the l2-normalized is mainly serving to orthogonalize the representations of familiar identities (as motivated in the original paper [3]), which in the standard network is accomplished via the learned mapping and softmax over identity units.

References

[1] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. SphereFace: Deep hypersphere embedding for face recognition. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 6738–6746, 2017.

[2] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, J. C. Chen, C. D. Castillo, R. Chellappa, D. White, and A. J. O'Toole. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, 115(24):6171–6176, jun 2018.

[3] R. Ranjan, C. D. Castillo, and R. Chellappa. L2-constrained Softmax Loss for Discriminative Face Verification. *ArXiv*, (1703.09507), 2017.