Trends in Cognitive Sciences

CellPress
REVIEWS

Review

Face Space Representations in Deep Convolutional Neural Networks

Alice J. O'Toole,[1,*] Carlos D. Castillo,[2] Connor J. Parde,[1] Matthew Q. Hill,[1] and Rama Chellappa[2]

Inspired by the primate visual system, deep convolutional neural networks (DCNNs) have made impressive progress on the complex problem of recognizing faces across variations of viewpoint, illumination, expression, and appearance. This generalized face recognition is a hallmark of human recognition for familiar faces. Despite the computational advances, the visual nature of the face code that emerges in DCNNs is poorly understood. We review what is known about these codes, using the long-standing metaphor of a 'face space' to ground them in the broader context of previous-generation face recognition algorithms. We show that DCNN face representations are a fundamentally new class of visual representation that allows for, but does not assure, generalized face recognition.

Face Recognition in Deep Convolutional Neural Networks

Generalized face recognition is accomplished in the human brain by large-scale networks of neurons. Deep convolutional neural networks (DCNNs) accomplish this task in an analogous, albeit artificial, way [1–9]. These **feed-forward neural networks** (see Glossary) are constructed of multiple layers of simulated neurons. As with the human visual system, DCNNs begin with raw images, which they recode across multiple 'neural layers' that alternately convolve and pool input. Although the calculations executed by the simulated neurons in the network are simple (**convolutions and pooling**), enormous numbers of computations are used to convert an image of a face into a representation that supports identification. Across early layers of these networks, the number of coding parameters expands exponentially reaching into 10s of millions of parameters and beyond. The highly compact face representation that emerges at the top level of the DCNN consists of just a few hundred features and is robust across a wide range of image variation (e.g., pose, illumination, and expression [10,11]).

To be successful as a face recognition system, a DCNN must be trained with multiple variable images of a very large number of identities. This is now feasible due to the large data sets of labeled training data (faces and identities) on the Web [1,6,12,13] (Box 1). Despite the rapid and sustained progress of DCNNs, the 'visual nature' of the face representation that emerges at the top level of a DCNN is not understood, nor is it known how this representation operates effectively across changes in image parameters and facial appearance [14,15]. We use the term 'visual nature' to refer to the specific visual information that is encoded neurally (e.g., structure, features, color) and the form the information takes in the representation (e.g., viewer centered, object centered, hybrid code).

In this review, we begin with the framework of a **face space**, which has grounded computational and psychological approaches to face recognition for decades. Next, we trace the evolution of face representations in these models and discuss the strengths and limitations of each. We continue with a look at human versus machine comparisons over the past decade.

Highlights

Deep convolutional neural networks (DCNNs) are the first class of algorithm to achieve generalized face recognition across viewpoint, illumination, expression, and appearance.

Face space illustrates the progress of automated face recognition. Image-based models do not generalize across images or identity. Active appearance models represent identity, but do not model image generalization. DCNNs create a unitary space that houses both facial identity and face images.

Face representations in DCNNs are compact, with feature units that are not tuned to face or image properties (e.g., viewpoint) in any commonly understood way.

DCNN face spaces retain highly detailed information about face images, in addition to face identities.

Semantic interpretation of face representations in DCNN follows sparse trajectories in the space, rather than being interpretable by feature unit activation.

[1]School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA
[2]University of Maryland Institute for Advanced Computer Studies, College Park, MD, USA

*Correspondence:
otoole@utdallas.edu (A.J. O'Toole).

https://doi.org/10.1016/j.tics.2018.06.006

These provide insight into the face space representations created by computational algorithms, and consequently help us to understand their advantages and limitations as models of human face processing. Finally, we review what is known about the nature of the visual representation that emerges at the top level of DCNNs (see Figure 1, Key Figure, for an overview).

It becomes clear that DCNNs address critical shortcomings of previous models to produce a representation that is powerful and fundamentally different from existing hypotheses about face representations. The computational advances made by DCNNs offer insight into long-standing questions about how the visual system can turn retinotopic codes into representations that accomplish generalized face recognition. DCNNs also offer a novel model for understanding how the quality of a face representation evolves through personal experience.

Face Space Model of Human Face Recognition

Psychologists have relied on the idea of a face space as a theoretical metaphor for human face representation since it was introduced in the early nineties [16]. In its first abstract form, face representations were considered as points in a multidimensional face space, with the axes of the space representing facial features, and the average or prototype face at the center of the space. The distance between two faces in the space indicated the perceived similarity between the faces. This simple model accounted for interesting phenomena in human face processing, including face typicality [17] and other-race effects [18]. In the first computational instantiation of a face space, principal component analysis (PCA) was applied to scaled and aligned face images to produce an **image-based face space**. These models appeared simultaneously in psychology [19,20] and computer vision [21,22]. In psychology, image-based face spaces provided a testable instantiation of the effects predicted by the abstract face space model [16]. In computer vision, despite the strong limitation of image-based PCA to operate only within a

Box 1. Deep Convolutional Neural Networks for Face Recognition

Convolutional neural networks have been around for decades [59]. The neocognitron model from the 1980s [60,61] was inspired by neurons in cat visual cortex [62] and worked with cascaded layers of simulated neurons with diverse receptive fields. Advances in training methods such as backpropagation in the mid-80s enabled error correction to propagate across multiple layers of neurons [63]. Using architectures that resemble today's DCNNs, backpropagation was applied to handwritten digit recognition not long after backpropagation was introduced [63]. The main advance of DCNNs for faces is their ability to operate effectively on uncontrolled images captured 'in the wild' (see Figure I).

Why then, are DCNNs from the past few years so much more powerful than earlier DCNNs? It is now clear that speed and scale matter. High-end computational graphical processing units, the preferred computational engine of choice for DCNNs, are 30–50 thousand times faster than computers in the 80s. Thus, it is now feasible to execute large numbers of convolutions and to train networks with data sets that are orders of magnitude larger than those used previously. For deep networks, which have very large numbers of parameters, large data sets are not a luxury, but a necessity for the DCNN to converge and generalize well. Data sets on this scale were unimaginable even a decade ago, but are now freely available from the Web. Table I shows an overview of the data set sizes and parameter numbers for four important DCNN-based face recognition systems.

In addition to scale and speed, there have also been theoretical breakthroughs that alter the flow of information in DCNNs and change the way the internal representations are tuned to learning sets. New nonlinear activation functions (e.g., ReLU [64]) skip connections and allow for deeper networks. '**Loss functions**' have also improved with the introduction of softmax [65] and triplet loss [66]. These functions map the weights of the network and an input into a real value that represents the loss incurred when comparing the prediction to the ground truth.

Based on their performance, DCNNs are the top choice for face recognition applications. The fast pace of progress, however, has outpaced researchers' understanding of how the representations in DCNNs relate to those created by the visual system. When tens of millions of computations stand between a single input face image and the representation produced by these networks, this is perhaps not surprising.

Glossary

Active appearance model (AAM): a computational instantiation of a face space that uses PCA to analyze faces coded in terms of their deformation in shape and reflectance from the average face.

Convolution and pooling: a convolution takes a linear combination of neighboring pixels according to the weights expressed in a filter. Pooling combines the output of multiple convolutions through a simple operation, usually a maximum or an average. Neurons in the primate visual system implement a computation analogous to convolution. They do this by combining dendritic inputs and passing the sum of these inputs through a threshold function. If the sum of the weighted inputs exceeds the threshold, the neuron produces an action potential.

DCNN's face representation: defined as the pattern of activation produced by processing a face image through a trained DCNN at the top layer, just prior to the final layer that reads out the identification decision. It is a highly compact representation of the image in the form of a vector of a few hundred or a few thousand real numbers.

Face knowledge history: defined as the state of the DCNN parameters following the first stage of training. In this training, the DCNN parameters are set using a large number of images from a large number of faces. The face images in the training set and their corresponding identity labels are used to teach DCNNs to map multiple variable images of faces onto single identities.

Face learning history: defined by the state of the DCNN parameters following the second stage of training. In this training, the network parameters are tuned with the goal of optimizing the performance of the network to operate well on the data set(s) that will be used to test the system for face identification. The tuning mostly alters parameters in the top layers (high-level visual processing), leaving the lower layer (early visual processing) parameters largely stable.

Face space: refers to a metaphorical representation of how people perceive faces. The axes of the space represent

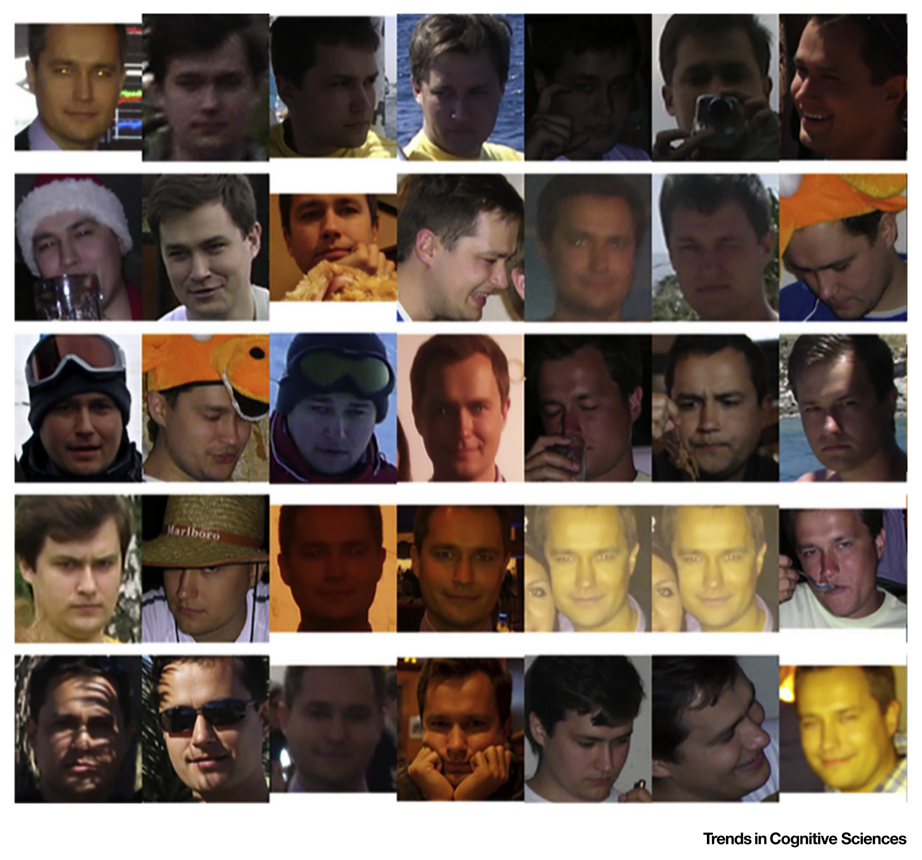


Figure I. Highly Variable Images Used to Test DCNNs. Example of the kinds of images that DCNNs are able to recognize from labeled faces in the wild [80]. This data set is a common benchmark for DCNN face recognition systems.

Table I. Four Important DCNN Face Recognition Systems

System	Architecture	Loss function	Training set size
Deep Face [5] (Facebook, 2013)	Five convolutional layers, two fully connected layers	softmax	4 million faces 4000 identities
VGG face [13] (2015)	Thirteen convolutional layers, two fully connected layers	softmax	2.6 million faces 2600 identities
FaceNet [6] (Google, 2015)	One convolution and 11 inceptions (modules associating convolutional layers and max pooling layers). Sixty-seven convolution layers and two fully connected layers	softmax and triplet loss	200 million faces 8 million identities
L2 softmax [9] (2017)	Uses ResNet101 has 100 convolutional layers that skip connections and two fully connected layers at the top	softmax and triplet loss	3.7 million faces 58 207 identities

single (frontal) viewpoint, this model provided the computational engine for the first generation of commercially viable face recognition systems [22].

By the mid-nineties, the shortcomings of image-based analysis of faces were apparent. Robust face recognition is impossible with image-based PCA. The development of 2D and 3D

the features with which faces are encoded (e.g., nose size). Each face has a value on each feature axis. The set of values specify a face's coordinates (position) in the space. The similarity between faces is modeled as the distance between them.

Feed-forward neural network: neural network that moves input in only one direction, that is, forward through one or more layers of processing nodes without looping or cycling backward through previous layers.

Image-based face space: a computational instantiation of the face space metaphor that is created by submitting images of a set of faces to a PCA. The axes of the space are the extracted principal components or eigenvectors. This model is sometimes referred to as eigenfaces [22].

Loss function (cost function): using the current weights of the network and a given input, the network will make an output prediction. A loss function generates a real number that represents the loss (or cost) incurred by making this prediction. Training a network is the process of adjusting the weights to minimize the loss over the entire training set.

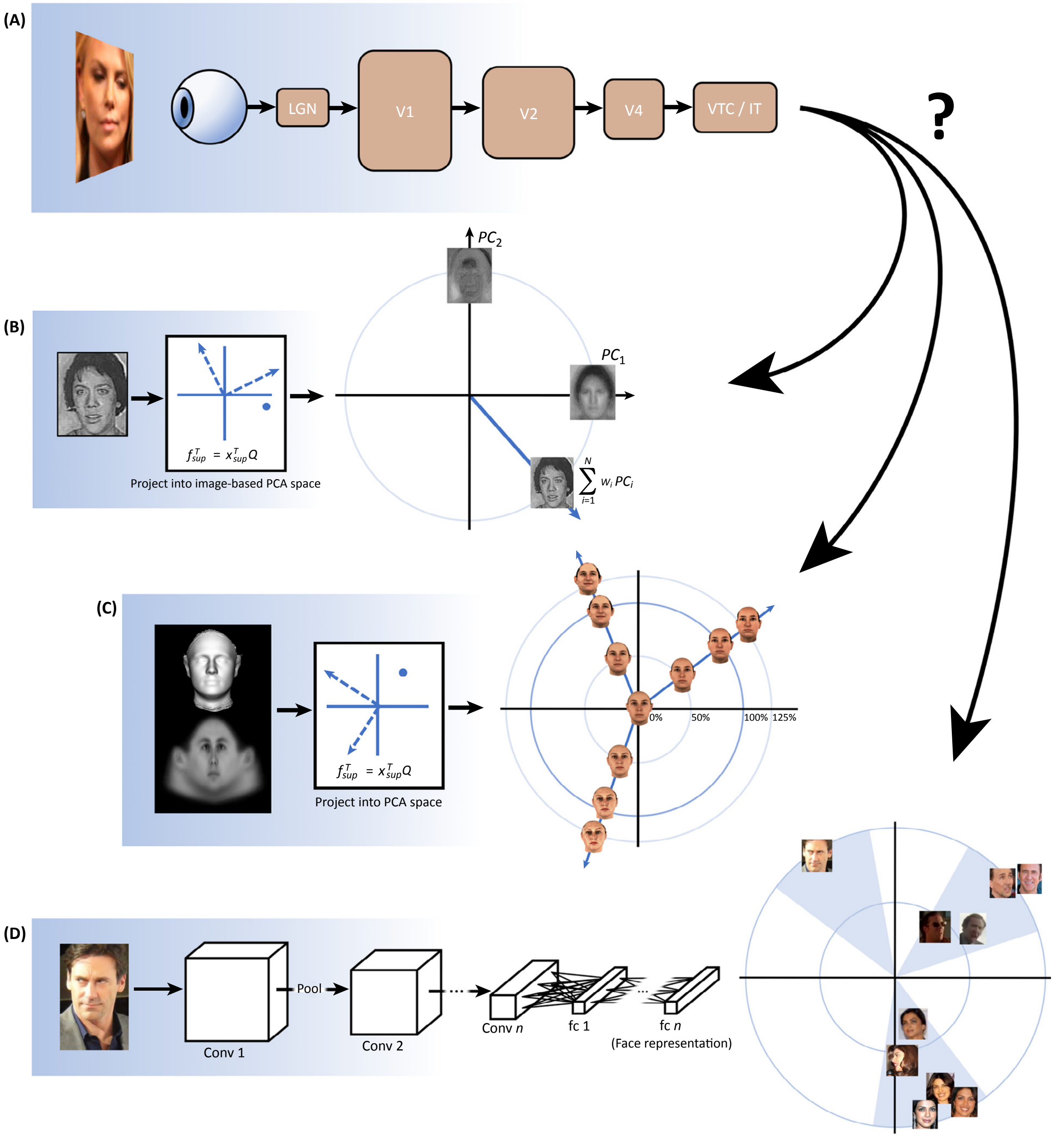
Opponent code: refers to a representation whereby features indicate an opposition along a continuum that is centered at an average or stable point.

Personal face learning history: defined as the knowledge the face recognition system has of the identities to be tested. The implementation of this training varies across DCNNs, but is commonly applied to a network that takes the final layer of the DCNN as input to a simple network that learns to map a new set of face images onto their respective face labels (e.g., unique identity codes). These are the identities that the network learns as individuals.

***t*-SNE Visualization:** *t*-SNE, (abbreviation for *t*-distributed stochastic neighbor embedding), is a nonlinear method of dimensionality reduction that uses gradient descent to preserve the distance between each point in a high-dimensional space while reducing the number of dimensions.

Key Figure

Evolution of Face Space in Computational/Neural Models



Trends in Cognitive Sciences

(See figure legend on the bottom of the next page.)

morphable models that operate independently on face shape (configuration) and reflectance (albedo) information in a face was an important step forward [10,23–25]. Morphable models, referred to in the computer science literature as **active appearance models** (AAMs) of facial synthesis [23], had a strong impact on computer graphics. The 3D morphable model [25] contains complete 3D surface and 2D reflectance information about faces and supports reconstruction of a 3D head from a single, high-quality frontal image. This enables synthesis of heads viewed from arbitrary viewpoints and under arbitrary illumination conditions. AAMs, however, model the appearance of a face and were never intended as face recognition systems. In addition, the impracticality of 3D data capture limited the use of these models in face recognition systems.

Despite limited value as a face recognition system, morphable models had a profound impact on psychological theory. A primary accomplishment of these models was that they solved the correspondence problem between any arbitrary face/head and the average face/head, with high-density point-by-point correspondence [25]. This yielded a novel face space that was completely 'morphable'. Any face could be morphed continuously into any other face across a trajectory in high-dimensional space. Using the average face as a reference, these models represent shape and reflectance variability across a population of faces, with each face coded as the deformation of its shape (δx, δy, δz) and reflectance (δr, δg, δb) from the average.

Applied to this deformable model of shape and reflectance, PCA produces a face space with properties that are quite interesting to psychologists. First, individual identities can be isolated in the face space as trajectories that begin at the origin of the space, pass through anticaricatures (less distinctive versions of a face), and terminate at the veridical face. Continuing this identity trajectory [11] beyond the original face produces an exaggerated (caricatured) version of a face. Second, the model accounted for high-level face adaptation effects, whereby adaptation to distorted faces (excessively wide or thin) produced an after-effect when viewing a subsequently presented normal face [26]. After-effects were extended to face and antiface pairs [27], as well as to race, gender, and facial expression [28]. Adaptation has proven relevant for understanding the neural processing of faces, both in fMRI [29] and in single-unit neurophysiology in primates [30].

Next, we explain why these previous models now fail to make progress on understanding human face processing.

Models Fail to Account for Recent Psychological Findings

There are strong differences in the robustness of human face recognition for familiar versus unfamiliar faces (for a review, see [31]). These differences make it clear that there are two fatal problems in using the morphable model as a psychological model of face recognition. First, the

Figure 1. (A) Schematic shows the progression of visual processing in the ventral stream from an image-based representation in early visual areas [lateral geniculate nucleus (LGN), V1, V2, V4] toward a categorical representation in the ventral temporal cortex (VTC)/inferior temporal (IT) cortex. Arrows point to the evolution of computational hypotheses about the nature of neural/psychological representations of faces. (B) Early image-based principal component analysis (PCA) models analyze pixel-based images. They are a good fit for early visual processing, but fail to generalize across variable images to categorical representations of identity. (C) Morphable models represent the deformation of face shape and pigmentation with respect to an average or prototype face, but do not model the variability of real-world images. They are powerful face synthesis tools, but are not useful for recognition of faces in natural viewing conditions. (D) Deep convolutional neural network (DCNN) models emulate the neural processing of the primate visual system and produce a face space that retains information about the categorical identity of faces and image properties with other categorical relevance. DCNN receptive fields across these layers for object recognition have been visualized in previous work [78]. PCA, principal component analysis.

model is based completely on information about the face itself (shape, reflectance). Modeled in this space, face recognition success/failure in any given case must be accounted for only in terms of the properties of a face in relation to other faces (similarity/confusability). There is no provision in this model for understanding the role of imaging conditions in recognition, because there is no way to model photometric effects (illumination, pose) in the context of the similarity space of faces.

This brings us to the second problem. There is no provision in these models for allowing some faces (familiar faces) to be better-represented than other faces (unfamiliar faces). Recent landmark papers in the psychology literature [31,32] make it clear that progress in understanding human face recognition depends on an ability to understand why and how familiar faces are recognized more robustly than unfamiliar faces. The severity of the challenge this problem poses to understanding human face recognition is illustrated in Figure 2A from [32]. Participants were asked to indicate the number of identities pictured among 40 images. On average, people responded that there were 7.5 identities. In fact, there are only two identities! With similarly challenging images of two Swedish celebrities, Swedish participants made no errors. On these same images, British participants again responded that there were more than seven identities present.

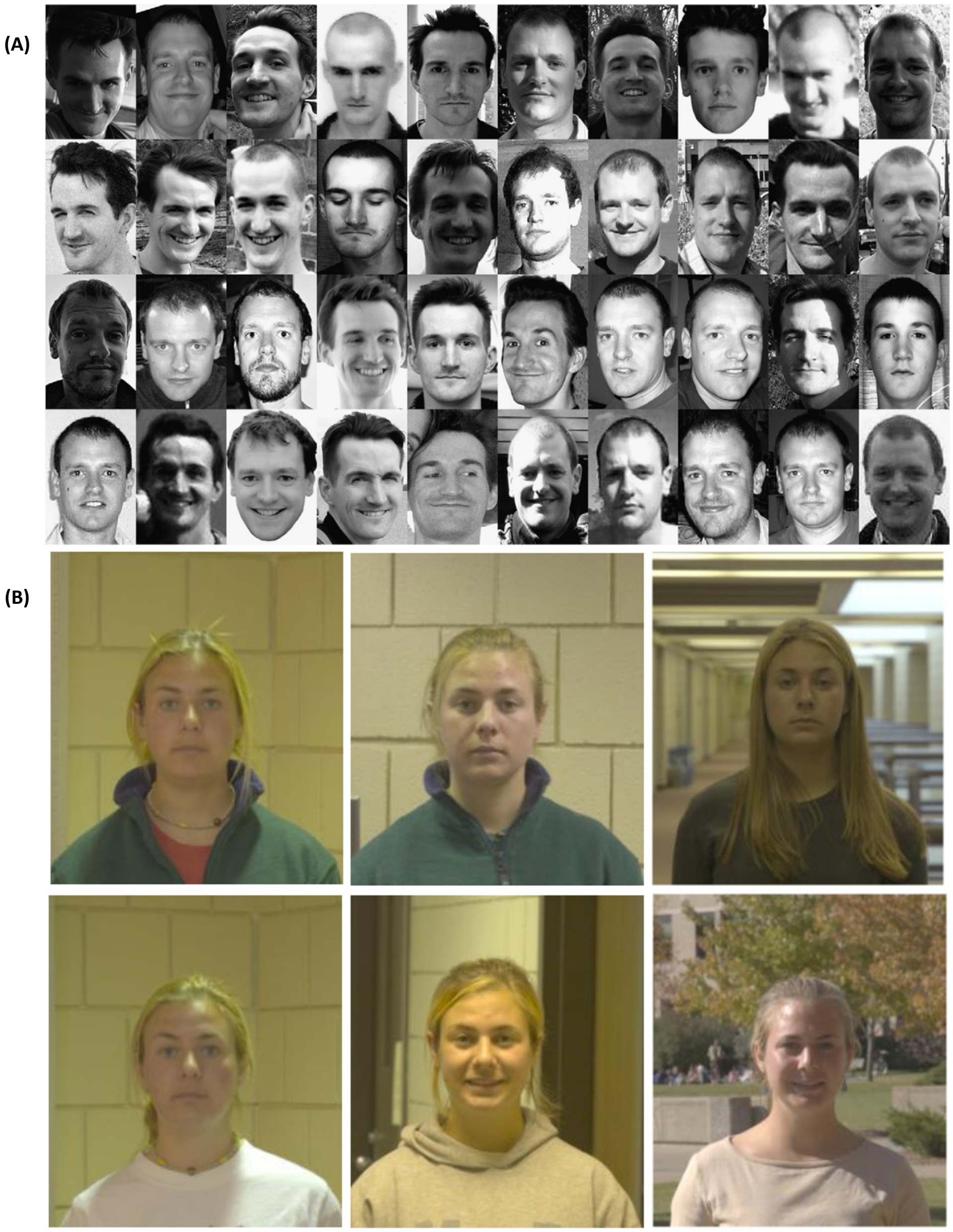
This experiment illustrates the difference between the human ability to ‘tell so many similar people apart’ and the human ability to ‘tell (a smaller number of) people together’. The former is long cited by psychologists to underlie human expertise for face recognition. The latter is a more recently appreciated human expertise that is limited to faces of the people we know well. In Figure 2A, we fail to see the multiple pictures of the two people together, because they are unfamiliar to us. If we knew them, it would be a simple task. This ability represents human face recognition at its best and is not a part of current models of human face recognition.

Next, we trace the progress of machine-based face recognition algorithms as models of human face processing. We approach this question by using human–machine performance comparisons as a guide.

Machine Face Recognition Approaches Human Performance: (2005–2013)

Beginning in 2005, computer-based face recognition systems began to close the gap between human and machine performance. This progress can be tracked by examining the difficulty of the problems on which machine performance was tested. Over the past two decades, large-scale evaluations of state-of-the-art face recognition algorithms, open to international competitors from academics and industry, have provided a look at the progression of task difficulty expected at any given time by a state-of-the-art face recognition algorithm [33–36]. These tests were augmented with systematic comparisons between human and algorithm performance beginning in 2005 (cf., for a review [37]).

In 2007, the state of the art for machines was to determine whether pairs of frontal images showed the same identity or different identities, when one image was taken under controlled illumination (passport style) and the second image taken under uncontrolled indoor illumination. At that time, the best three algorithms surpassed humans on face pairs prescreened to be challenging for the machines [38]. By 2012, the state of the art for machine recognition progressed to matching pairs of images with unconstrained variation in illumination (indoor and outdoor) and expression. Figure 2B shows six images of the same person, arranged (by column) into three pairs, from an international test of algorithms in [39]. Image pairs were divided into easy, more challenging, and extremely challenging categories, based on the performance



Trends in Cognitive Sciences

Figure 2. Perceiving Identity from Highly Variable Images. (A) How many different identities are pictured among these 40 images? People unfamiliar with the identities pictured guess that there are between seven and eight identities, when in fact there are only two. This figure, from [32], provides compelling evidence that familiarity with an individual is an important prerequisite for generalized face recognition. (B) There are six images of the same person here, grouped into three pairs (by column). For computational models of face recognition prior to 2013, the leftmost pair was considered 'easy', the middle pair was considered 'challenging', and the rightmost pair was considered 'extremely challenging' [39]. In 2012, machines were more accurate than humans at matching the identity of the easy and challenging pairs, but humans and machines were equally accurate on the extremely challenging pairs.

of a baseline algorithm. Human–machine comparisons for matching the identity of faces across pairs of images showed that machines were far better than humans on the easy and moderately challenging pairs [40]. On the extremely challenging pairs, machines and humans were equally matched. In no case were humans superior to the best algorithms.

In summary, as of 2012, the best face recognition algorithms performed as well as, or better than, humans – with two caveats. First, the human perceivers tested were not familiar with the people in the images. Second, state-of-the-art algorithms, up to 2012, worked only on frontal images with no ability to recognize people over changes in viewpoint. Thus, even as recently as 2012–2013, a viewpoint change of more than a few degrees of rotation was beyond the capability of the best algorithms.

Deep Networks and Generalized Face Recognition (2014–Present)

DCNNs trained for face recognition first appeared in 2014 and have moved the state of the art to recognition of uncontrolled images captured 'in the wild' (e.g., [3,5–8,13,41]). See Box 1 for an example of the kinds of images that DCNNs are able to recognize. Performance on data sets such as Labeled Faces in the Wild [42,43], IJB-A [44], and Mega-Face [45] offers evidence that face recognition by machines is beginning to attack the problem of generalized face recognition, including the ability to 'tell faces together' across widely variable changes in image and appearance.

One reason for the progress of DCNNs is that they are trained with millions of 'diverse' images of thousands of individuals 'in the wild'. In other words, the algorithm learns an identity from exposure to many images. Ideally, the diversity of images for each identity should span multiple pose, illumination, and expression conditions. It should also span appearance variables (e.g., age, make-up, hair styles, facial hair, glasses). To produce a face recognition system that generalizes well, the availability of this kind of 'high-quality' training data is necessary. It is worth noting that data quality, for DCNNs, is usually thought to reference the number and diversity of training exemplars and the accuracy of the training labels. It is not generally considered in terms of traditional image quality metrics (high resolution, optimal illumination, taken from a frontal viewpoint). Accurate labels are also critical and are provided by online crowd-sourcing, or in the case of commercial systems (e.g., Facebook), by labeled face-identity pairs available in social media.

In the next section, focusing on how DCNNs are trained, we show how these networks have the potential to model the evolution of an individual face's representation as it becomes familiar to the network.

From Unfamiliar to Familiar in a DCNN

Each of the steps needed to implement a DCNN has an analog in human face recognition. The first step is to train the DCNN with millions of images of thousands of people. This results in a network that converts a face image into a compact feature code, with the DCNN pushing the representation through a bottleneck of units in the penultimate layer of the network. In other words, the compact feature code emerges one layer down from the top layer of the network. The top layer contains the 'training identity' units of the network during its initial training. We can think of this penultimate layer, therefore, as the neural activation profile for a 'generic' incoming face image that was not part of the training set. Although the system is trained initially to identify the people in the training data set from a wide variety of images, the final identification layer (i.e., the 'training identity' units one layer beyond the compact feature layer) is simply removed at the end of the training. This initial training step gives the system a general knowledge of how facial appearance changes with image variation and can be thought of as the system's **face knowledge history**. Evidence for the 'general' nature of this initial training is that some DCNNs for face recognition are actually built on top of networks that have been trained for object recognition (see, e.g., [46]; for object networks, see, e.g., [47]). This stage of training may function primarily to learn general visual transformations across viewpoint and illumination. These transformations are applicable to both objects and faces.

Second, DCNNs are commonly fine-tuned with data (faces and labels) that are statistically similar to the images that one expects to encounter in the intended application. Often, these images are from the database for which the application is intended, but are of different people than those to be recognized in the application. In this second training, the learning rate is set to have low impact on the weights in lower layers of the network and higher impact on the weights in layers closer to the top of the network. The rationale is that processing in lower layers (closer to the 'retinotopic' image) should remain relatively stable across most/all sets of faces. Processing closer to the top of the network, however, should be more tuned to subsets of image types that the system expects to encounter in the application. This produces a **face learning history** with emergent higher-level features fine-tuned to the statistical structure of the training examples. This training might be used, for example, to model the quality of representations produced for different races of faces as a function of diversity in the population.

Once this second training is complete, the top layer of the network is again removed. In a final training, the faces that must be recognized as individuals are learned, but in a way that does not alter the weights within the DCNN itself. Instead, this last stage of training usually consists of training a simple one-layer network that maps the top-layer DCNN representations onto an identity label. This training can rely on all available 'training images' of each person to be learned and can be thought of as a **personal face learning history**.

How well a person is learned depends on the quality of the representation possible, given the face knowledge, the face history, and the quality of training data for the individual person at the final training phase. Some faces (familiar) will be better represented than others. Here, we refer to the code produced at the top layer of the network as the **DCNN's face representation**.

Next, we review what is known about the face representations that emerge from DCNN processing.

Probing Face Representations in DCNNs

What is in a DCNN feature? If the goal is to create a generalized representation of facial identity from image-based input, one computational strategy is to eliminate or filter out the nuisance information from an image to obtain a purely categorical code (identity). However, top-layer DCNN representations, which support face identification across image variation, retain surprisingly accurate information about the original input image [14] (see also [15] for a similar result in object recognition DCNNs). Using top-layer features from two recent face recognition networks [8,48], it is possible to predict the view (yaw, ±90°; pitch, on-center, or off-center), and media type (video frame or still image) of the actual image that was input to the network [14]. Specifically, a classifier network trained to predict face yaw from the DCNN face representation was accurate to within 8.6°. Pitch was predicted with an accuracy of 71.5% and media type was predicted with 93.3% accuracy. Electrophysiological recordings in the inferior temporal (IT) cortex of macaque monkeys [15] likewise show explicit coding of image properties (position, size, pose) for objects (Box 2). This is a surprising result in neurophysiology that is entirely consistent with the finding that yaw, pitch, and media type can be predicted from the

Box 2. Neural Perspectives: Going up!

There is broad consensus among visual neuroscientists that the computational end goal of early visual processing (V1, V2, V3, V4) is to create object representations that generalize across exemplars of a category and variable image parameters [15,67–69]. Representations must also retain information to distinguish among exemplars [70,71]. There is overwhelming evidence from functional neuroimaging and primate electrophysiology that face/object representations meet these criteria in the ventral temporal cortex (VTC), but fail to meet them at earlier stages of processing [71]. Thus, it may seem puzzling that category-based representations in the VTC retain low-level visual features with high fidelity. Category-orthogonal object properties (e.g., viewpoint), however, are represented more explicitly in the IT cortex of macaques than in earlier ventral visual stream areas [15]. If the goal of neural computation is the elimination of image-based information to arrive at a category-based representation, these findings force us to consider representations in the VTC in a different light.

Putting together the perspective of neural computation with the strategies used by deep learning, there is a way to resolve this paradox by redefining the computational goals of the visual system. In a paper that preceded modern DCNNs [70], the authors note that each layer along the progression of ventral system areas produces a high-dimensional representation space. In image-based spaces in early visual areas, the categorical structure of face/object identity is hopelessly confounded with image parameters [70]. Restricting our attention to faces, in simple terms, the organization (similarity structure) of faces in the face space is altered from one layer to the next. The goal of the progressive neural recoding, according to [70], is not to eliminate image parameters along the upward stream, but to refine the representational space to 'untangle' categories inherent in the input (e.g., identities, viewpoints, image conditions). A successful high-dimensional representation supports multiple subspace projections that allow for easy (linear) readout of face identities and other categorical data in the input.

The implications of this change in perspective are profound. Probing neural receptive fields or even population receptive fields in the VTC/IT may offer little or no insight into the nature of the visual code at work in the high-dimensional space defined by neurons. Instead, we need to understand how combinations of neuronal responses support projections that untangle all/much of the categorical information in the incoming image.

high-level face representations generated by DCNNs [14]. The result is also consistent with functional neuroimaging studies that find representations in human face-selective regions tuned to viewpoint (e.g., [49–51]) and other low-level image properties such as illumination [52], size [53], and position [54].

Focusing on single feature units at the top layer of a DCNN, we see clear differences in the extent to which DCNN codes for the individuals learned by the network are robust against viewpoint change [14]. Receptive field analysis of these units showed that view invariance was a property, not of the individual feature units in the representation, but of individual identities. Specifically, in a recent paper [14], an index of viewpoint invariance for feature units in the top level of a face recognition DCNN was developed. This index measured whether individual feature values for each identity changed systematically with viewpoint (frontal vs. profile). For some identities, most of the feature values differed systematically between frontal and profile views. However, for other identities, most feature values did not diverge for frontal and profile views.

Why would some faces be coded robustly and others in a view-dependent way? One possibility is that some/all physical characteristics of individual faces are more/less diagnostic across viewpoint (e.g., baldness, facial hair). Supporting this idea, among 25 000 test faces, Bono, with his ubiquitous blue-tinted spectacles, was one of the most robustly coded faces [14]. Another possibility is that the diversity of the available training images for a person modulates the stability of individual feature values. These two hypotheses are not mutually exclusive. DCNN face recognition algorithms, however, are usually trained and tested with unconstrained data sets. Consequently, there is rarely enough control on the number/diversity of images used to represent individual identities to evaluate the relative importance of these two factors in making a robust code for an individual face.

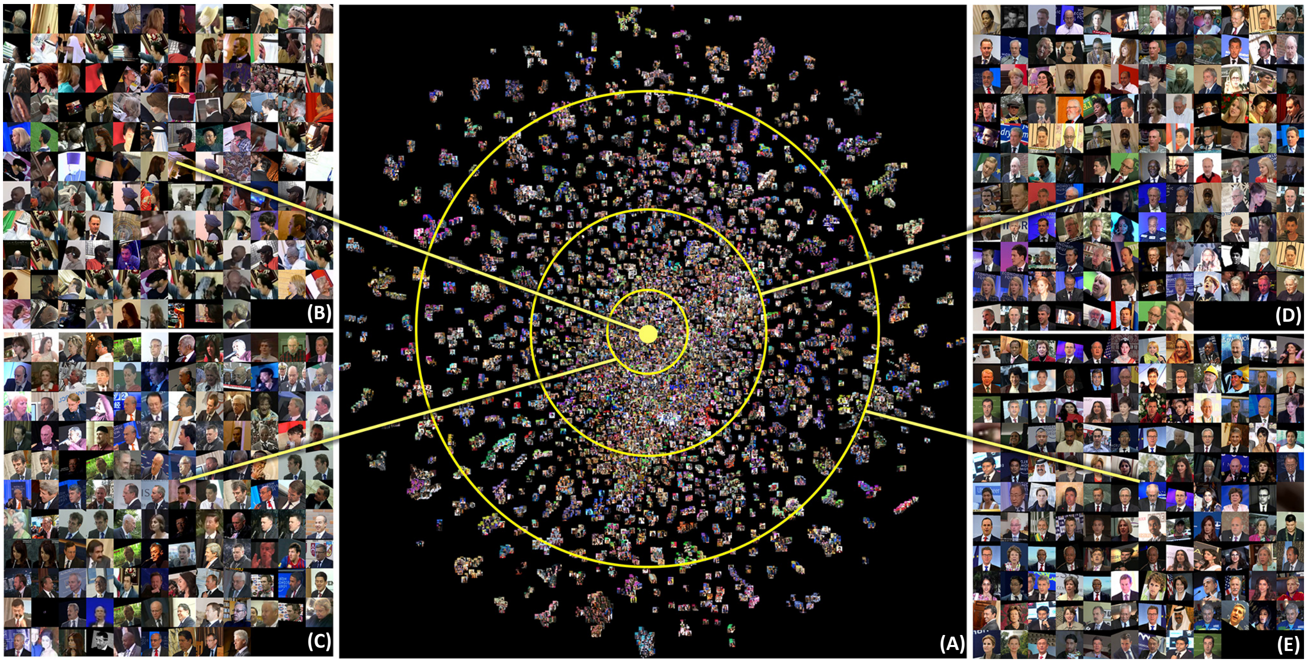
Notwithstanding, the finding that individual features in the top-level representation can operate both with and without tolerance for view change, depending on the data quality and personal experience, is striking. This new finding suggests a visual code wherein the individual features are not tuned to viewpoint in any commonly understood way (Box 2).

Next, we consider the global organization of the DCNN face space. This gives insight into the interpretation of the neural response properties of single feature units.

DCNN Face Space Makes Image Quality Explicit

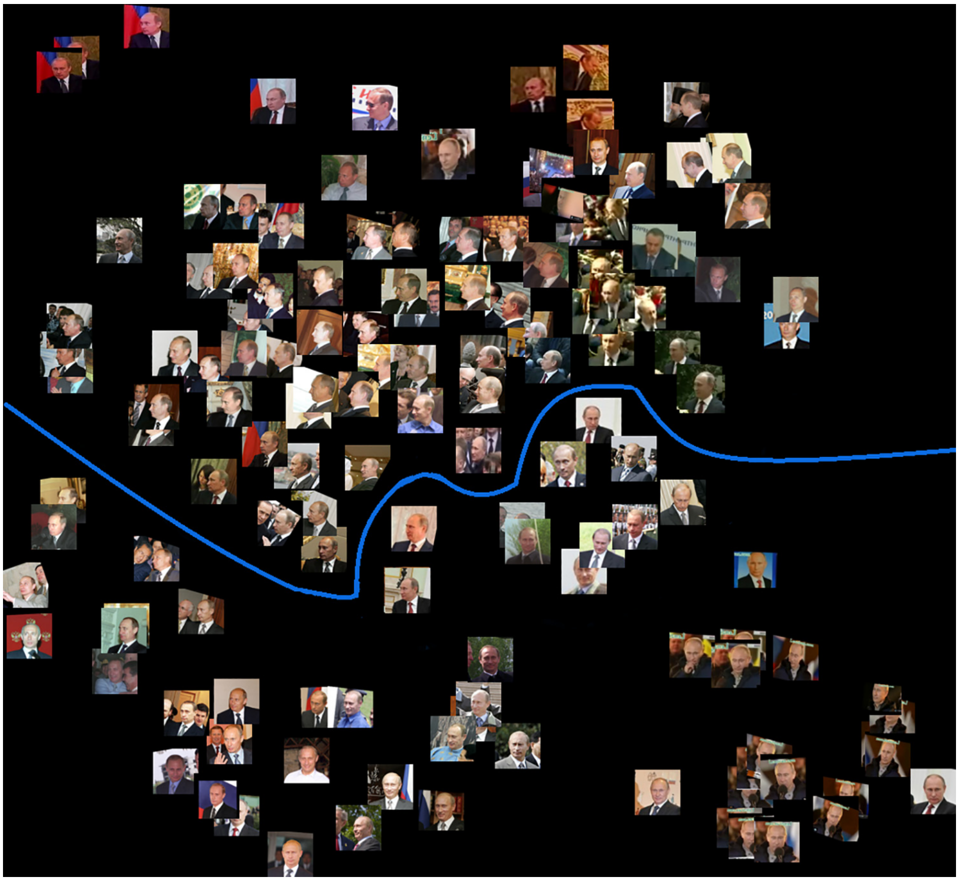
The face space that emerges from the top-level face representations in DCNNs shows a remarkable structure, with low-quality images (blurred, obscured, etc.) near the center of the space and progressively higher-quality images located more peripherally (see Figure 3, [14]). The figure also shows that face images located more peripherally tend to be frontal views, whereas off-center views are closer to the center of the space.

Figure 3 shows that the structure of DCNN face space differs from the prototype-centered spaces seen in image-based and morphable models. Instead, the center of the DCNN space acts as a kind of 'garbage dump' for poor-quality and unusable images. From a coding perspective, this suggests an **opponent code**, whereby face images close to the center of the space are there because the DCNN fails to code them across multiple contrastive dimensions. This unique DCNN structure may be due to the need to accommodate, at high



Trends in Cognitive Sciences

Figure 3. Structure of Deep Convolutional Neural Network (DCNN) Face Space for 25K images. (A) A face space visualization of 25 000 face images from [14]. Images close to the center of the space are of poor quality (B). Better-quality images radiate out from the center, with off-center views more prevalent in inner rings around the center of the space (B, C, D), and frontal views located more peripherally (E). Visualization is created with *t*-SNE [79] – a dimensionality reduction technique that uses stochastic gradient descent methods to preserve high-dimensional Euclidean distances between data points, while embedding them in low-dimensional space. (Note: the *t*-SNE is used for illustration purposes only: faces in panels B–E were selected using their vector norms in the high-dimensional deep feature space.) *t*-SNE, *t*-distributed stochastic neighbor embedding.



Trends in Cognitive Sciences

Figure 4. A Single Identity in the Deep Convolutional Neural Network (DCNN) Face Space. Example of the spatial layout of 140 images of a single identity in the DCNN space, visualized with *t*-SNE. The blue curve, which is hand-drawn, roughly separates the non-frontal and frontal images in the space. *t*-SNE, *t*-distributed stochastic neighbor embedding.

levels of abstraction, both the face identity and the individual images that comprise the model's experience with the face.

How do face identities and image data coexist in the unified space produced by DCNNs? The ***t*-SNE visualization** allows us to expand any section of the space to examine the local organization of image representations within an identity. Figure 4 shows face space locations of 140 in-the-wild images of a single identity [14]. Frontal and off-frontal views of the person separate roughly in the space (as indicated by the hand-drawn blue curve). Because the *t*-SNE analysis shows a low-dimensional representation of a high-dimensional space, we can assume that this structure explains substantial variation in the representation of this identity. This points to an image-driven structure hierarchically nested within identity, melding identity and images in a unitary space.

DCNN Representations Are Overkill

An intriguing feature of DCNNs is the highly compact nature of the face representation that emerges at the top layer. How could a few hundred or even a few thousand features represent such a large number of faces over such a wide variety of image and appearance variables? One possibility is that each top-layer feature plays a critical role in representing the uniqueness of a facial identity. This speculation appears incorrect. For object recognition networks, the dimensionality of the DCNN output layer can be reduced substantially with little or no effect on performance [55]. Beginning with multiple baseline networks, each with a 4096-dimensional output layer, a reduction down to 2048 features actually improved object recognition performance marginally [55]. A reduction of the output layer to 1024, and even to 128, features resulted in an average performance drop of only 2%!

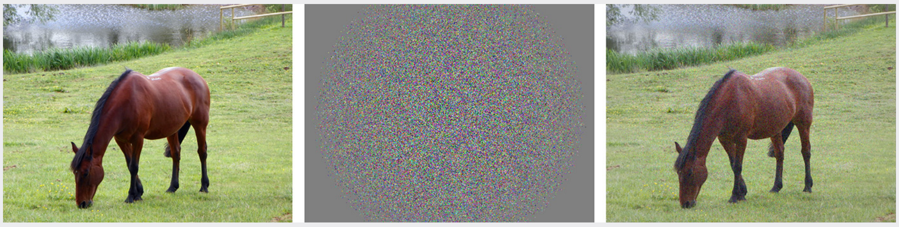
Interpreting this finding is challenging, but there are clues in a recent study [56] that examined the semantic meaning of individual top-level units in DCNNs. That study begins with the premise that traditional computer vision systems rely on feature extraction, whereby single features are interpretable as meaningful variations in the image input domain (e.g., 'x detectors'). The coordinates in this space provide a semantic 'read-out' of the stimulus (e.g., face) from the individual feature units (e.g., big nose, round face). For DCNNs, however, there is no distinction in semantic meaning between individual high-level feature values and random linear

Box 3. Little Networks, Little Problems; Big Networks, Big Problems

A caveat to the success of DCNNs is that these networks can be attacked by 'adversarial images' [72]. Specifically, small perturbations, added to an image that a DCNN would classify correctly, can cause the network to misclassify the image – often with high confidence. Remarkably, for human perceivers, the adversarial image can appear visually identical to the original (see Figure I). Adversarial attacks pose an enormous challenge for DCNN applications that operate on in-the-wild images. Imagine generating a perturbation to an input image in a self-driving car that causes a stop sign to be classified as a yield sign.

Why do adversarial attacks succeed? Simply put, CNNs learn a mapping from input image space to feature space. In this nonlinear transformation, there will be points that are close in image space, but distant in DCNN feature space. Adversarial attacks find these points, sometimes with the aid of the network itself (white box attack) and sometimes with the aid of other networks trained on similar data (black box attack). The deeper the network, the more nonlinear the mapping. Consequently, breakthroughs that have allowed for deeper networks that increase performance markedly have opened the networks up even more to adversarial vulnerabilities. Methods for combating these attacks are being developed with multiple strategies (e.g., [73–75]) that show some success.

From a visual system perspective, the finding that two perceptually identical images can be physically quite distinct is not new. For example, metamers in color vision, discovered in the 1800s, showed that the perceived color of a monochromatic light could be matched exactly by a linear combination of three other wavelengths [76]. Color metamers provided the first scientific evidence for the trichromatic nature of human color vision [77]. It remains to be seen whether an understanding of adversarial metamers for DCNNs will likewise result in a better understanding of the fundamental nature of the high-level visual representations in the primate visual system. However, it is clear that the existence of DCNN metamers, is not, in and of itself, a finding that disqualifies DCNNs as a potential model of the ventral temporal cortex.



Trends in Cognitive Sciences

Figure I. Example of an Adversarial Image. Image that a DCNN identifies as a horse with 88% confidence (left). Adversarial perturbation (center). Perturbed image (i.e., original image plus adversarial perturbation) (right). The same DCNN identifies this as a bicycle with 92% confidence!

combinations of features (i.e., random directions) in the space [56]. The authors conclude, therefore, that the space, rather than the individual units, contains the semantic stimulus information.

A fundamental implication of this change of perspective is that it throws doubt on the possibility of understanding the meaning of the high-level feature units (neurons) in terms of their response properties. Semantic characterization of units as ‘x detectors’ may give us little insight into the nature of the visual code (Box 2). Instead, the DCNN space presents a needle-in-the-hay-stack ‘search problem’ for meaning. The number of random directions in the space is so overwhelming that a search for semantic meaning must transcend the limits of probing the response properties of the feature units themselves. If meaning resides in directions through the space, it is not surprising that features can operate both in a view-invariant and in a view-dependent way [14].

If all random directions in the space are potentially meaningful semantically, even a low-dimensional DCNN face space (a few hundred dimensions) has more capacity than we need for the number of human faces we learn in a lifetime. The vastness, and perhaps ‘emptiness’, of the DCNN space, even when it is storing enormous numbers of images, may help to explain why, on occasion, face/object identification by DCNNs can go surprisingly astray (see Box 3).

From a neural perspective, in recent work [57], a measure of generalization accuracy as a function of representational complexity was developed. For object recognition, DCNNs scored better on this measure than the IT cortex, suggesting that DCNNs rival representations in primate IT cortex [57] (see Box 2).

Concluding Remarks

DCNNs have made progress on long-standing problems in computer vision that seemed intractable only 5 years ago. For their performance alone, DCNN representations are worthy of serious study. It has been shown that these models now perform at a level equal to the best humans (professional forensic face examiners and super-recognizers) [58]. Beyond that, however, DCNNs were designed, decades ago (see Box 1), to exploit the computational strategy used by the primate visual system. Only recently has it been possible to train these networks to find a complex mapping between images and categories, using enormous data sets of face images – sampled as they might be in a natural visual world. As a proof of principle, DCNNs qualify as a promising model of neurally-inspired face representations in high-level visual cortex. At present, our knowledge of how DCNNs work is limited, but the research paths ahead are open to exploration (see Outstanding Questions).

Acknowledgments

The authors were supported in part by the Intelligence Advanced Research Projects Activity (IARPA). This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

1. Simonyan, K. *et al*. (2013) Fisher vector faces in the wild. In *Procedings of the British Machine Vision Conference 2013*, BMVA. Vol. 2, pp. 8.1–8.12, https://doi.org/10.5244/C.27.8
2. Huang, G.B. *et al*. (2012) Learning hierarchical representations for face verification with convolutional deep belief networks. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2518–2525, IEEE
3. Sun, Y. *et al*. (2014) Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1891–1898, IEEE
4. Haoxiang, L. *et al*. (2015) A convolutional neural network cascade for face detection. In *IEEE Conference on Computer Vision and*

Outstanding Questions

Should DCNN models of faces be trained with objects and scenes to leverage global distinctions in the structure of the ventral temporal cortex for broad categorical distinctions (e.g., animate/inanimate; faces, objects, and place; retinal eccentricities) and local within-category specializations? Or, is a modular approach to learning faces more valid in neural terms?

Is there a fundamental difference between the lower or upper layers of DCNNs trained for object categorization and face identification? Or, do they generate similar feature spaces? If there is a difference, is it due to the different training tasks or to different sets of training images?

How can DCNN research impact and relate to feature spaces in the human brain? Can DCNNs model multiple face-selective regions?

In addition to the finding that DCNN-generated face representations contain information about image parameters, do they also contain information about superordinate face categories such as gender and age?

How should data quality be defined? What role does data quality play in the formation of face representations that generalize well across image variation and at what stage of training is data quality most important?

How do DCNN architectures, as well as loss functions and other processing decisions, affect the representation of faces in DCNNs?

Are all DCNN architectures created equal? Is there a critical mass of connections beyond which the nature of the emergent representation stays more or less stable? Or, do particular characteristics of network architecture matter?

If semantic information in faces resides in random directions through a neural space, in what way do we need to alter techniques in primate neurophysiology to begin to capture the neural coding of image semantics?

Pattern Recognition, pp. 5325–5334, IEEE. https://doi.org/10.1109/CVPR.2015.7299170

5. Taigman, Y. *et al*. (2014) DeepFace: closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708, IEEE. https://doi.org/10.1109/CVPR.2014.220
6. Schroff, F. *et al*. (2015) FaceNet: a unified embedding for face recognition and clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 815–823, IEEE
7. Chen, J.-C. *et al*. (2015) An end-to-end system for unconstrained face verification with deep convolutional neural networks. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 360–368, IEEE
8. Sankaranarayanan, S. *et al*. (2016) Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS 2016),* pp. 1–8, IEEE
9. Ranjan, R. *et al.* (2007) An all-in-one convolutional neural network for face analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1–6
10. Hancock, P.J.B. *et al.* (1996) Face processing: human perception and principal component analysis. *Mem. Cognit.* 24, 26–40
11. O'Toole, A.J. *et al.* (1999) Three-dimensional shape and two-dimensional surface reflectance contributions to face recognition: an application of three-dimensional morphing. *Vis. Res.* 39, 3145–3155
12. Hu, G. *et al*. (2015) When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pp. 384–392, IEEE
13. Parkhi, O.M. *et al*. (2015) Deep face recognition. In *Proceedings of the British Machine Vision Conference 2015*, BMVA. Vol. 1, pp. 41.1–41.12, https://doi.org/10.5244/C.29.41
14. Parde, C.J. *et al*. (2017) Face and image representation in deep CNN features. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 673–680, IEEE
15. Hong, H. *et al.* (2016) Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19, 613–622
16. Valentine, T. (1991) A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q. J. Exp. Psychol. A* 43, 161–204
17. Light, L.L. *et al.* (1979) Recognition memory for typical and unusual faces. *J. Exp. Psychol. Hum. Learn.* 5, 212–228
18. Malpass, R.S. and Kravitz, J. (1969) Recognition for faces of own and other race faces. *J. Pers. Soc. Psychol.* 13, 330–334
19. O'Toole, A.J. *et al.* (1988) A physical system approach to recognition memory for spatially transformed faces. *Neural Netw.* 1, 179–199
20. O'Toole, A.J. *et al.* (1993) Low-dimensional representation of faces in higher dimensions of the face space. *J. Opt. Soc. Am. A* 10, 405–411
21. Sirovich, L. and Kirby, M. (1987) Low-dimensional procedure for the characterization of human faces. *J. Opt. Soc. Am. A* 4, 519–524
22. Turk, M. and Pentland, A. (1991) Eigenfaces for recognition. *J. Cogn. Neurosci.* 3, 71–86
23. Cootes, T. *et al.* (2001) Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 681–685
24. Cootes, T.F. *et al.* (1995) Active shape models – their training and application. *Comput. Vis. Image Underst.* 61, 38–59
25. Blanz, V. and Vetter, T. (1999) A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH'99),* pp. 187–194, ACM Press
26. Webster, M.A. and Maclin, O.H. (1999) Figural aftereffects in the perception of faces. *Psychon. Bull. Rev.* 6, 647–653
27. Leopold, D.A. *et al.* (2001) Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat. Neurosci.* 4, 89–94
28. Webster, M.A. *et al.* (2004) Adaptation to natural facial categories. *Nature* 428, 557–561
29. Loffler, G. *et al.* (2005) fMRI evidence for the neural representation of faces. *Nat. Neurosci.* 8, 1386–1391
30. Leopold, D.A. *et al.* (2006) Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* 442, 572–575
31. Young, A.W. and Burton, A.M. (2018) Are we face experts? *Trends Cogn. Sci.* 22, 100–110
32. Jenkins, R. *et al.* (2011) Variability in photos of the same face. *Cognition* 121, 313–323
33. Phillips, P.J. *et al.* (2000) The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 1090–1104
34. Phillips, P.J. *et al*. (2005) Overview of the face recognition grand challenge. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 947–954, IEEE
35. Phillips, P.J. *et al.* (2010) FRVT 2006 and ICE 2006 large-scale experimental results. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 831–846
36. Phillips, P.J. *et al.* (2012) The good, the bad, and the ugly face challenge problem. *Image Vis. Comput.* 30, 177–185
37. Phillips, P.J. and O'Toole, A.J. (2014) Comparison of human and computer performance across face recognition experiments. *Image Vis. Comput.* 32, 74–85
38. O'Toole, A.J. *et al.* (2007) Face recognition algorithms surpass humans matching faces over changes in illumination. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 1642–1646
39. Phillips, P.J. *et al*. (2011) An introduction to the good, the bad, and the ugly face recognition challenge problem. In *2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*, pp. 346–353, IEEE
40. O'Toole, A.J. *et al.* (2012) Comparing face recognition algorithms to humans on challenging tasks. *ACM Trans. Appl. Percept.* 9, 1–13
41. Sun, Y. *et al.* (2016) Hybrid deep learning for face verification. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 1997–2009
42. Kumar, N. *et al*. (2009) Attribute and simile classifiers for face verification. In *Proceedings of the 12th IEEE International Conference on Computer Vision (ICCV)*, pp. 365–372, IEEE
43. Learned-Miller, E. *et al.* (2016) Labeled faces in the wild: a survey. In *Advances in Face Detection and Facial Image Analysis* (Kawulok, M., ed.), pp. 189–248, Springer
44. Whitelam, C. *et al*. (2017) IARPA Janus Benchmark-B Face Dataset. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 592–600, IEEE
45. Kemelmacher-Shlizerman, I. *et al*. (2016) The MegaFace Benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–10, IEEE
46. Ranjan, R. *et al.* (2017) HyperFace: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* Published online December 8, 2017. http://dx.doi.org/10.1109/TPAMI.2017.2781233
47. Krizhevsky, A. *et al*. (2012) ImageNet classification with deep convolutional neural networks. In *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems (Vol. 1)*, pp. 1097–1105, Curran Associates
48. Chen, J.C. *et al*. (2016) Unconstrained face verification using deep CNN features. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV 2016)*, pp. 1–9, IEEE
49. Kietzmann, T.C. *et al.* (2015) The occipital face area is causally involved in facial viewpoint perception. *J. Neurosci.* 35, 16398–16403
50. Kietzmann, T.C. *et al.* (2012) Prevalence of selectivity for mirror-symmetric views of faces in the ventral and dorsal visual pathways. *J. Neurosci.* 32, 11763–11772

51. Natu, V. *et al.* (2010) Dissociable neural patterns of facial identity across changes in viewpoint. *J. Cogn. Neurosci.* 22, 1570–1582

52. Grill-Spector, K. *et al.* (1999) Differential processing of objects under various viewing conditions in the human lateral occipital complex. *Neuron* 24, 187–203

53. Yue, X. *et al.* (2011) Lower-level stimulus features strongly influence responses in the fusiform face area. *Cereb. Cortex* 21, 35–47

54. Kay, K.N. *et al.* (2015) Attention reduces spatial uncertainty in human ventral temporal cortex. *Curr. Biol.* 25, 595–600

55. Chatfield, K. *et al*. (2014) Return of the devil in the details: delving deep into convolutional nets. *arXiv:1405.3531*

56. Szegedy, C. *et al*. (2014) Intriguing properties of neural networks. *arXiv:1312.6199*

57. Cadieu, C.F. *et al.* (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* 10, 1–18

58. Phillips, P.J. *et al.* (2018) Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proc. Natl. Acad. Sci. U. S. A* 115, 6171–6176

59. Lecun, Y. *et al.* (2015) Deep learning. *Nature* 521, 436–444

60. Fukushima, K. (1980) Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* 36, 193–202

61. Fukushima, K. *et al.* (1983) Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Trans. Syst. Man Cybern.* SMC-13, 826–834

62. Hubel, D.H. and Wiesel, T.N. (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* 160, 106–154

63. LeCun, Y. *et al.* (1989) Backpropagation applied to handwritten zip code recognition. *Neural Comput.* 1, 541–551

64. Hahnioser, R.H.R. *et al.* (2000) Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* 405, 947–951

65. Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, Springer

66. Weinberger, K.Q. and Saul, L.K. (2009) Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.* 10, 207–244

67. Guclu, U. and van Gerven, M.A.J. (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014

68. Eickenberg, M. *et al.* (2017) Seeing it all: convolutional network layers map the function of the human visual system. *Neuroimage* 152, 184–194

69. Yamins, D.L.K. *et al.* (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U. S. A.* 111, 8619–8624

70. DiCarlo, J.J. and Cox, D.D. (2007) Untangling invariant object recognition. *Trends Cogn. Sci.* 11, 333–341

71. Grill-Spector, K. and Weiner, K.S. (2014) The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* 15, 536–548

72. Goodfellow, I.J. *et al.* (2009) Measuring invariances in deep networks. *Adv. Neural Inf. Process. Syst. 22* 22, 646–654

73. Goodfellow, I. *et al.* (2014) Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pp. 1–9, Neural Information Processing Systems Foundation

74. Rawat, A. *et al.* (2017) Harnessing model uncertainty for detecting adversarial examples. In *Second Workshop on Bayesian Deep Learning (NIPS 2017)*, pp. 1–13, Neural Information Processing Systems Foundation

75. Li, X. and Li, F. (2017) Adversarial examples detection in deep networks with convolutional filter statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5775–5783, IEEE

76. Maxwell, J.C. (1857) XVIII. Experiments on colour, as perceived by the eye, with remarks on colour-blindness. *Trans. R. Soc. Edinb.* 21, 275–298

77. Young, T. (1802) The Bakerian lecture: on the theory of light and colours. *Philos. Trans. R. Soc. London* 92, 12–48

78. Zeiler, M.D. and Fergus, R. (2014) Visualizing and understanding convolutional networks. *arXiv:1311.2901*

79. van der Maaten, L. and Hinton, G. (2008) Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* 9, 2579–2605

80. Huang, G.B. *et al.* (2007) *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* In: http://w.tamaraberg.com/papers/Huang_eccv2008-lfw.pdf