COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019) e12729

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12729

Social Trait Information in Deep Convolutional Neural Networks Trained for Face Identification

Connor J. Parde,[a] Ying Hu,[a] Carlos Castillo,[b] Swami Sankaranarayanan,[b] Alice J. O'Toole[a]

[a]*School of Behavioral and Brain Sciences, The University of Texas at Dallas*
[b]*University of Maryland Institute for Advanced Computer Studies, University of Maryland*

Received 4 June 2018; received in revised form 8 January 2019; accepted 19 March 2019

Abstract

Faces provide information about a person's identity, as well as their sex, age, and ethnicity. People also infer social and personality traits from the face — judgments that can have important societal and personal consequences. In recent years, deep convolutional neural networks (DCNNs) have proven adept at representing the identity of a face from images that vary widely in viewpoint, illumination, expression, and appearance. These algorithms are modeled on the primate visual cortex and consist of multiple processing layers of simulated neurons. Here, we examined whether a DCNN trained for face identification also retains a representation of the information in faces that supports social-trait inferences. Participants rated male and female faces on a diverse set of 18 personality traits. Linear classifiers were trained with cross validation to predict human-assigned trait ratings from the 512 dimensional representations of faces that emerged at the top-layer of a DCNN trained for face identification. The network was trained with 494,414 images of 10,575 identities and consisted of seven layers and 19.8 million parameters. The top-level DCNN features produced by the network predicted the human-assigned social trait profiles with good accuracy. Human-assigned ratings for the individual traits were also predicted accurately. We conclude that the face representations that emerge from DCNNs retain facial information that goes beyond the strict limits of their training.

Keywords: Social trait perception; Face identification; Deep learning; Convolutional neural networks

Correspondence should be sent to Connor J. Parde, School of Behavioral and Brain Sciences, The University of Texas at Dallas, 800 West Campbell Rd., Richardson, TX 75080. E-mail: Connor.Parde@utdallas.edu

1. Introduction

1.1. Background

The human face is the basis for a wide variety of judgments. From just a single glance, we can recognize friends, family, and acquaintances. Faces also help us retrieve semantic or emotional information about others (e.g., name, career, relation). In addition to being the primary cue for identity, faces carry information useful for making visually derived semantic categorizations (Bruce & Young, 1986). For example, we spontaneously perceive the age, ethnicity, and gender of others (Cloutier, Mason, & Macrae, 2005; Macrae, Quinn, Mason, & Quadflieg, 2005). Among these categorizations are the personality inferences we make from faces (Bar, Neta, & Linz, 2006; Todorov, Said, Engell, & Oosterhof, 2008). Although the human tendency to infer personality traits from faces is well studied, much less is understood about the extent to which this information supports or interacts with the information used for face identification.

From a computational perspective, Oosterhof and Todorov (2008) and Walker and Vetter (2009) showed that human trait judgments can be modeled and manipulated reliably by computer graphics programs. These studies also demonstrated that faces contain quantifiable physical correlates for human social-trait judgments. In Oosterhof and Todorov (2008), participants labeled neutral-expression face images with words that related to social traits (e.g., sociable, mean, weird, confident). These responses were reduced to 13-dimensional social-trait vectors that described each face, which were then submitted to principal component analysis (PCA). This produced a *trait space* with two main axes. The first axis was interpreted as "trustworthiness/valence" and the second axis was interpreted as face "dominance." Next, Oosterhof and Todorov (2008) generated a face-shape space using three-dimensional laser scans of faces. This space was used to produce three-dimensional models of faces that participants rated for trustworthiness and dominance. Using the best linear fit, Oosterhof and Todorov (2008) found that the fiducial points in the face-shape model could be manipulated to alter trait judgments. The degree of these manipulations predicted whether human participants rated faces as more or less trustworthy or dominant. Social information can also be inferred from highly variable images (e.g., large changes in viewpoint, expression, illumination, and age), albeit with a third important dimension in the trait space which is related to youthful/attractiveness (Sutherland et al., 2013; Vernon, Sutherland, Young, & Hartley, 2014).

One limitation of Oosterhof and Todorov's (2008) model is that it considers only face shape and does not consider face reflectance or pigmentation. Walker and Vetter (2009) addressed this limitation with a model that included both the shape and reflectance of faces. They used 200 three-dimensional laser scans of faces to generate a three-dimensional morphable model (Blanz & Vetter, 1999) and collected trait ratings on a sample of faces. Using these ratings, they found the locations of faces in the space that had high values on specific traits and calculated the average position of these faces, as well as the direction of the vector toward this average from the global average in the space. To

generate faces with variable amounts of these traits, individual faces in the space were moved along this trajectory to alter (decrease or increase) the presence of the trait. Walker and Vetter (2009) showed that these manipulations could be controlled to affect social trait inferences made by humans.

Combined, the work of Oosterhof and Todorov (2008) and Walker and Vetter (2009) demonstrates that it is possible to use computer graphics models, driven by human ratings, to "generate" faces that elicit specific social trait judgments. Consistent with these *generative* findings, other studies have shown that for a limited number of traits, it is possible to learn a direct mapping from face *images* to trait judgments, without explicit control or knowledge of the underlying two- or three-dimensional structure of the face. For example, using deep convolutional neural networks (DCNNs), Lewenberg, Bachrach, Shankar, and Criminisi (2016) predicted social-trait judgments from face images. Lewenberg et al. (2016) crowd-sourced large numbers of "attribute" ratings from faces, on both objective (e.g., hair color, gender) and subjective (e.g., attractiveness, humor) aspects of facial appearance. Next, DCNNs were trained to classify images in a binary manner, according to the presence or absence of each attribute. Lewenberg et al. (2016) predicted objective attributes with very high accuracy (*gender* = 98.33%, *ethnicity* = 83.35%, *hair color* = 91.69%, *makeup* = 92.87%, *age* = 88.83%). Three subjective traits also were well-predicted, but to a lesser degree (*attractiveness* = 78.85%, *humorous* = 69.06%, *chubby* = 61.38%). In converging work, McCurrie et al. (2017) trained DCNNs to predict three subjective traits (trustworthiness, dominance, and IQ), as well as age. McCurrie et al. (2017) measured the proportion of shared variance (R^2) between the trait predictions computed from their model and the trait ratings assigned by human participants. McCurrie et al. (2017) found significant agreement in all cases (trustworthiness $R^2 = 0.43$, dominance $R^2 = 0.46$, age $R^2 = 0.74$, IQ $R^2 = 0.27$).

It has also been shown that the social-trait inferences made by humans can be predicted using the output from DCNNs trained for different tasks (Song, Linjie, Atalla, & Cottrell, 2017). Song et al. (2017) predicted human-assigned social traits from the principal components (PCs) of the feature responses produced by networks trained for object recognition, face identification, and facial landmark localization. They found that the highest correlation between human-assigned traits and predicted traits could be obtained using the features from a network trained for object recognition. This underscores the widespread availability of trait information in the visual features used to represent diverse images. However, the image set used by Song et al. (2017) did not control for emotional expressions or image characteristics (pose, illumination, etc.). These variables are known to influence the way humans perceive social traits (Said, Sebe, & Todorov, 2009). Thus, the social-trait prediction performance reported by Song et al. (2017) was based on a combination of features from face identity, expression, and image parameters. A more fine-tuned analysis is necessary to determine the extent to which each of these factors can contribute to the accuracy of the predictions. In this study, we focused primarily on properties of the face itself as a cue to social trait inferences.

Overall, the above studies indicate that DCNNs are capable of learning human-assigned social traits from face images, as well as attributes such as gender and age. The

features that support these trait inferences were made concrete by the work of Oosterhof and Todorov (2008) and Walker and Vetter (2009). For example, wide eyes are seen as trustworthy, whereas thin eyes and lips are seen as untrustworthy. These features are an integral part of a person's appearance and thereby a cue to their unique identity. An empirical indication that identity and trait information may be related comes from a study by Hassin and Trope (2000), who found that faces were judged to be more similar to one another when they were rated as having comparable social traits.

This returns us to the question we address here. How does the information in faces that gives rise to social-trait inferences relate to the information that specifies the identity of a face? To that end, we made use of a computational model trained exclusively to identify faces. We asked whether the face representations produced by this network retain the information needed to make inferences about a face's social-trait appearance. In other words, to what extent is it possible to use face representations optimized in an identity-trained neural network as a representation that supports human-like trait inferences?

Before proceeding, we digress briefly to define and discuss DCNNs, a class of hierarchical neural networks introduced by Krizhevsky, Sutskever, and Hinton (2012). DCNNs have changed the state-of-the-art in machine learning and have proven especially powerful for visual tasks such as face and object recognition (Krizhevsky et al., 2012; Taigman, Yang, Ranzato, & Wolf, 2014). These networks were designed originally to model the response properties of the primate ventral visual stream (Fukushima, 1988). DCNNs consist of layers of simulated neurons that alternately convolve and pool input, while expanding the representation in intermediary layers of the deep network (Hinton, Srivastava, Krizhevsky, Sutskever, & Salakhutdinov, 2012). The final layer(s) of DCNNs are typically fully interconnected and serve to compress the representation at the top level of the network into an abstract set of emergent features. These features form a highly compact representational code—usually on the order of a few hundred elements. It is important to note that the success of deep learning is in part due to the availability of extremely large, open-source, identity-labeled training datasets.

For face identification, a major accomplishment of DCNNs is their visual robustness to changes in viewpoint, illumination, expression, and appearance (AbdAlmageed et al., 2016; Chen, Patel, & Chellappa, 2016; Ranjan, Sankaranarayanan, Castillo, & Chellappa, 2017; N. Zhang et al., 2015). This is a consequence of the fact that DCNN architectures for face identification are trained on large numbers of identities and use multiple, variable images of each identity. DCNNs trained in this manner have consistently yielded state-of-the-art results on the well-tested Labeled Faces in the Wild and YouTube Faces datasets (Huang, Ramesh, Berg, & Learned-Miller, 2007; Parkhi, Vedaldi, & Zisserman, 2015; Schroff, Kalenichenko, & Philbin, 2015; Sun, Wang, & Tang, 2014; Wolf, Hassner, & Maoz, 2011; Zhou, Cao, & Yin, 2015).

For present purposes, the compact nature of the face representations produced by DCNNs, as well as the impressive ability of these representations to support identification, makes them an ideal tool for probing the co-dependence of trait and identity information in faces.

1.2. Overview

We tested the interdependence of identity and social-trait information in the top-level features of a DCNN trained for face identification. To study this, we collected social-trait ratings from participants for a large number of face images. Next, we obtained identity descriptors for each of the face images using a state-of-the-art face-identification DCNN (Sankaranarayanan, Castillo, Alavi, & Chellappa, 2016). We used the DCNN representations and the human-assigned trait ratings to address four questions. First, we asked whether the social-trait ratings we collected modeled a *trait space* similar to that described in previous research (Fiske, Cuddy, & Glick, 2006; Oosterhof & Todorov, 2008; Walker & Vetter, 2009; Wiggins, Phillips, & Trapnell, 1989). This was tested by interpreting the first two axes produced by a multivariate analysis of the human trait ratings.

Second, we verified that the network used in this study produced an identity code powerful enough to support recognition. Specifically, we computed Receiver Operating Characteristic (ROC) curves to measure identification performance on the faces in our dataset, using the top-level features produced by the network as the representation of each face.

Third, we asked whether the top-level descriptors produced for each image by the face-identification DCNN could be used to predict the human-assigned social trait ratings. We trained and tested a linear regression model with cross-validation and compared the similarity between the predicted trait profile and the human-generated trait profile for each face.

Fourth, we asked whether social traits could be predicted *individually*. To answer this, we measured the prediction error for each social trait.

2. Methods — human ratings

2.1. Participants

A total of 85 participants completed the data-collection portion of this experiment (20 males, 65 females, aged 18–30, $M = 21$). Participants consisted mostly of undergraduate students from The University of Texas at Dallas. Participants enrolled in a psychology course were compensated for their participation with one credit toward that course. Participants not enrolled in a psychology course received no compensation. Of the 85 participants, one male and one female were excluded from analysis, because their arrival time overlapped with additional participants and they completed the experiment in an alternate setting.

2.2. Face stimuli

The face stimuli presented in this study were selected from the Human ID Database (O'Toole et al., 2005). To control for effects attributable to the race of the stimulus, only images of Caucasian identities were used in this experiment. All images showed individuals with a neutral expression. A total of 280 images were selected, depicting 192 distinct identities. Some identities ($n = 83$) appeared in the dataset multiple times to allow for verification

of the network's face–identification accuracy (see Section 4.2). In total, 109 identities appeared once, 78 identities appeared twice, and five identities appeared three times. For the data collection task, to make the workload for participants manageable, the 280 images were divided into four sets containing 70 images each (19 male, 51 female). All participants were assigned randomly to rate one of the sets. There were no duplicated identities within each set. Participants were unfamiliar with the identities in the images they were shown.

2.3. Social traits

The list of social traits used in this study was derived from the Big Five Factors of Personality (Gosling, Rentfrow, & Swann, 2003). The social traits listed by Gosling et al. (2003) constitute a generalizable and representative selection of traits pertaining to individual personalities. Each trait in the Big Five is classified into one of five personality domains — openness, conscientiousness, extraversion, agreeableness, or neuroticism. For the data collection task, we selected 18 traits that were either from this list or were related to items from this list (see Fig. 1). We selected multiple traits from each of the five domains listed by Gosling et al. (2003). It should be noted that the traits we chose for this study are different than those used in previous work (e.g., Oosterhof & Todorov, 2008). The traits here were chosen to reflect a wide range of personality traits that fit theoretically into the Big Five axes, as well as face perception research that examines trait inferences. We consider the relationship of these terms to those used in previous work by constructing the trait space described in Section 4.1.

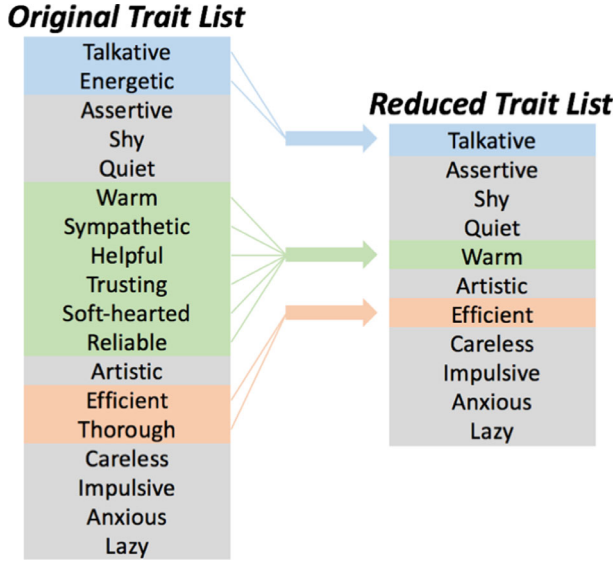

Fig. 1. List of social traits considered for this study. The original trait list shows the traits rated by participants. The reduced trait list shows the 11 unique ways participants used traits from this original list. Traits from the original list whose value correlated strongly with one another (>0.8 Cohen's *d*) were averaged together.

2.4. Procedure for human trait ratings

Participants in this study were asked to decide whether each item in a list of traits applied to a given face. Every participant viewed 70 face images (19 male, 51 female). Each face was presented on a computer screen next to a list of 18 social traits. For each trait, the participant was asked to select whether the trait: (a) applied perfectly, (b) applied somewhat, or (c) did not apply, to the face being shown. Previous studies examining social-trait inferences have used scales ranging from two points (Lewenberg et al., 2016) to nine points (Song et al., 2017), indicating that researchers' range of the response options has varied substantially. We used a 3-point scale to limit variance that might occur based on individual differences in the way participants used the scale. There was no default selection. Participants were able to advance to the next face only after making a selection for each trait. Participants were allowed as much time as they needed to rate each face. Fig. 2 provides an example page from this data collection task.

Trait ratings were collected from 19 and 23 participants per face image. The data comprised an $n \times m$ trait matrix, X, where n represents the number of faces and m represents the number of traits. Any given cell of the matrix $X_{i,j}$ contained the mean of participants' ratings for the jth trait on the ith face. The columns of this trait matrix were then converted into z-scores so that each face was represented by its deviation from the average ratings across the faces. The social-trait inferences assigned to different images of the

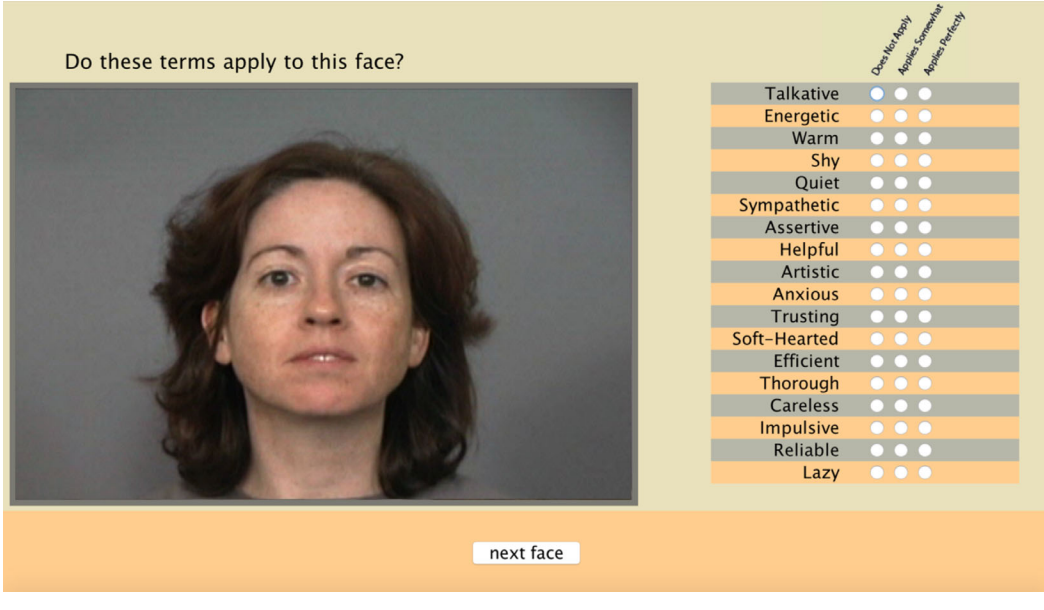

Fig. 2. A screenshot taken from the data collection task completed by participants. Participants viewed 70 (19 males, 51 females) faces in total, organized into sets according to gender. The list of traits was displayed next to each face. For each trait, participants were asked to select whether the trait "applies perfectly," "applies somewhat," or "does not apply" to the face being viewed.

same individual were more correlated than the trait inferences assigned to images showing different identities (same identity, $r = .56$; different identity, $r = .13$).

Several trait labels were applied very similarly by participants, indicating that at a higher level of abstraction, some subsets of traits measure the same underlying variable. The dimensionality of the trait space was reduced to minimize the effect of these duplicated traits within the dataset. To do this, traits that correlated highly with one another (Pearson correlation coefficient, $r > .80$) were averaged together and treated as a single trait. This reduced the final number of traits from 18 to 11 by combining (a) *talkative* and *energetic*; (b) *warm*, *sympathetic*, *softhearted*, *trusting*, *helpful*, and *reliable*; and (c) *efficient* and *thorough*. The traits *talkative* and *energetic* both imply levels of activity and were combined and relabeled as *talkative*. The traits *warm, sympathetic, soft-hearted, trusting, helpful,* and *reliable* all imply positive valence and were combined and relabeled as *warm*. The traits *efficient* and *thorough* both imply capability and were combined and relabeled as *efficient*. The social trait space considered in this study was comprised of the 11 traits that remained.

3. Methods for DCNN trait analysis

3.1. DCNN network specification

The neural network features for each face were generated from the DCNN described in Table 1. This network was trained for face identification and has achieved state-of-the-art performance on the IJB-A dataset (Klare et al., 2015). Parametric rectified linear units (PReLU) are used to compute the activation function. The final output layer consists of 512 units. The network was trained on the CASIA Webface database, which consists of 494,414 images depicting 10,575 identities (Yi, Lei, Liao, & Li, 2014).

Table 1
The deep network architecture of the face recognition network described by Sankaranarayanan et al. (2016). The analyses presented in this study examine the features extracted from the Fc7 layer of this network, for each of the images in our dataset

Deep Network Architecture		
Layer	Kernel Size/Stride	No. of Parameters
Conv1	11 × 11/4	35 k
Conv2	5 × 5/2	614 k
Conv3	3 × 3/2	885 k
Conv4	3 × 3/2	1.3 M
Conv5	3 × 3/1	885 k
Conv6	3 × 3/1	590 k
Fc6	1,024	9.4 M
Fc7	512	524 k
Fc8	10,575	5.5 M
Softmax loss		19.8 M

3.2. DCNN face representation

To carry out the simulation, the DCNN processed the 280 face images rated by participants to produce a unique 512-dimensional feature descriptor vector for each image. These feature descriptors were considered as the face identification network's representation of each face.

4. Analysis and results

4.1. Consistency of social-trait space with previous findings

First, we examined the underlying social-trait space from our human-assigned trait ratings to compare with previous findings (Fiske et al., 2006; Oosterhof & Todorov, 2008; Wiggins et al., 1989). This was important given that the social traits used in this study differed from those used in previous work. To this end, we submitted our participants' responses to a principal component analysis (PCA). PCA is a multivariate analysis technique that uses singular value decomposition to rotate the data and create a new space defined by orthogonal components that are ordered according to the percentage of variance they explain. Here, PCA was applied to the human trait inferences. The coordinates of each trait were plotted along the first two principal components and visualized to ensure parity with the existing literature (Fig. 3).

The first component produced by a PCA of the human-assigned traits was interpreted as *approachability* and separates traits such as *talkative*, *efficient*, and *warm,* from traits such as *anxious*, and *quiet*. The second component was interpreted as *dominance* and separates traits such as *assertive* and *impulsive* from traits such as *shy*, *quiet*, and *anxious*. These results are generally consistent with previous studies, which identified two primary social-trait components corresponding to slight variations on the axes we found: affiliation and dominance (Wiggins et al., 1989), warmth and competence (Fiske et al., 2006), and trustworthiness and dominance (Oosterhof & Todorov, 2008).

4.2. Face-identification accuracy

To test the DCNN for identification accuracy, we computed the cosine similarity between the 512-dimensional feature descriptors for all possible pairs of the 280 face images in our stimulus set. This resulted in an $n \times n$ similarity matrix, where n represents the number of face images. Identification accuracy was analyzed using an ROC curve to measure the separability between same-identity (match) and different-identity (non-match) image comparisons. The same-identity distribution contained the similarity scores for all image pairs in which the same individual was shown in both images. In all cases, these were two different images of the same person, taken over a span of weeks or months. The different-identity distribution consisted of all pairs of images depicting two

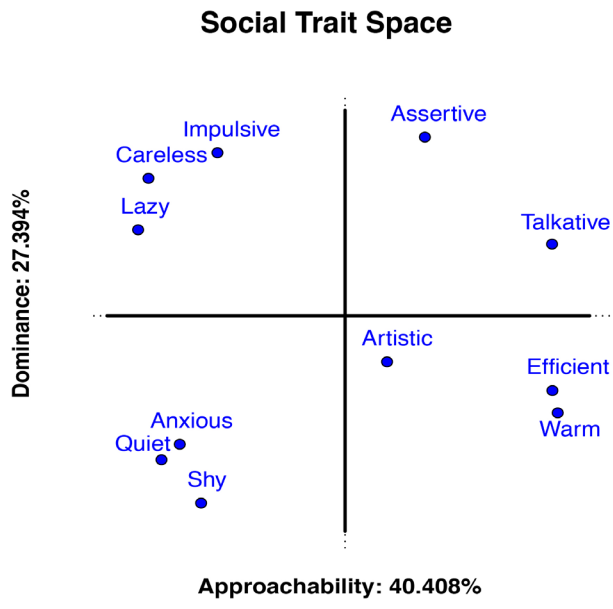

Fig. 3. Principal component analysis of the trait ratings collected from human participants. The 11 traits used to create our trait space are positioned according to their factor scores on the first (x-axis) and second (y-axis) components. The correlation between traits can be interpreted according to the angle between them, where a 0° angle implies perfect positive correlation, a 90° angle implies no correlation, and a 180° angle implies perfect negative correlation.

different individuals. We considered only scores from the upper triangle of the symmetric similarity matrix.

The ROC showed near-perfect performance on our stimulus set, as measured using the area under the curve (AUC = 0.995). This illustrates the high quality of the identity information in the top-level DCNN features.

4.3. Trait-profile predictions from DCNN face representations

A regression model was trained to predict the human-generated social-trait responses from the face representation generated by the face-identification DCNN. The trait ratings were represented in an $n \times m$ matrix, Y, where each of the n rows represented one of the 280 stimuli and each of the m columns represented one of the 11 traits. The face recognition features were represented in an $n \times k$ matrix, X, where each of the n rows again represented one of the 280 stimuli and each of the k columns represented one of the 512 top-level features. The regression model used to train the mapping from DCNN features to traits was:

$$Y = XB + E$$

where B represents the learned $k \times m$ weight matrix and E represents the error matrix. This model was trained using cross validation, wherein the regression model was learned using all but one of the stimulus images and was tested on the left-out image. In cases where the identity in the test image also appeared in the training images, all images of that identity were held out from the training set. This produced an $n \times m$ output matrix, $\overline{Y}$, similar to Y, but where all trait values were model predictions rather than human responses.

Human-generated trait vectors and computer-predicted trait vectors were compared using their cosine similarity. To test for statistical significance, we compared the average cosine similarity between the human- and computer-generated social-trait vectors to a null hypothesis distribution. This null distribution was created by a permutation test where the values within each column of the $n \times m$ trait matrix, Y, were reordered randomly. This effectively destroyed the relationship between social traits and stimuli, but preserved the underlying distribution of scores within each trait. The regression model was then re-computed. This was repeated for 1,000 iterations. After each iteration, the average cosine similarity between the 280 human-generated trait vectors and the 280 computer-predicted trait vectors was calculated.

The average cosine similarity between the human trait ratings and the computer-predictions (i.e., *true cosines)* was compared to the cosine similarities from the null distribution. The results show that the vectors of social-trait inferences predicted using the actual model were significantly more similar ($p < .001$) to the human-generated traits than the social-trait vectors predicted from the null distribution. There was no overlap between the mean of the true cosines ($M = 0.353$) and any of the cosines computed in the null distribution ($M = 0.078$). Thus, we conclude that the top-level features produced by the face-identification network contain information that supports human-generated social-trait ratings.

Next, we explored the degree to which social-trait information is distributed throughout the top-level features from the face-identification DCNN. This was tested by eliminating varying amounts of top-level DCNN features from the space and then re-computing the regression model. Deleting features containing information critical to social-trait prediction should reduce the cosine similarity between the human-generated trait vectors and the computer predictions. Alternatively, deleting "noisy" features (i.e., features consisting of only information extraneous to the social-trait prediction task) should *increase* the accuracy of the trait predictions.

The top-level features from the DCNN were ordered according to their contribution to the regression model. The contribution of each feature was assessed using its regression weights and was measured to be the sum of absolute values across the rows in the k (number of features) by m (number of traits) weight matrix. Rows with the lowest sums contributed the least to the regression model and rows with the highest sums contributed the most. Beginning with the lowest 5%, an increasing number of features were removed and the regression models were recomputed. This was repeated until the prediction accuracy, indicated by cosine similarity, began to decrease. A decrease in prediction accuracy occurred after removing 27.5% of the features. The remaining 72.5% of features represent the optimal set of predictors. The average cosine similarity between human-generated trait

vectors from this optimized regression model was significantly higher ($M = 0.533$, $p < .001$) than the average cosine similarities computed using the null distribution ($M = 0.078$). There was no overlap between the average of the true cosines and the cosines computed in the null distribution. These results show that social trait information is not evenly distributed throughout the top-level feature space learned by the face-identification DCNN. The importance of individual top-level DCNN features for either identification or for predicting social-trait inferences should be explored in-depth in future work. Here, we report the finding that accurate trait-inference prediction requires just a subset of the features from the overall top-level DCNN feature space.

4.4. Predicting individual traits from face representations

The previous regression experiment established that a profile of social-trait information is represented in the top-level features of the face-identification DCNN. We now address the extent to which individual social traits can be predicted in isolation. Regression-model predictions were computed for each of the 11 individual traits. The resulting $n \times 1$ dimensional trait vectors were compared to the corresponding columns from the original $n \times m$ trait matrix using the coefficient of determination (R^2). This provides a direct measure of the similarity between each predicted trait and its human-generated counterpart.

The coefficients of determination for predicting each trait individually are presented in Fig. 4. In Fig. 5, we plot the prediction accuracy for each trait contrasted against a null distribution. This figure shows that all of the social traits considered in this study were predicted at levels significantly above chance. The best-predicted traits here were

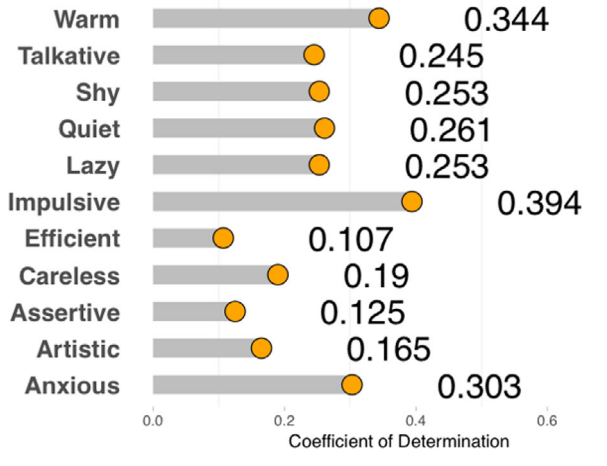

Fig. 4. Coefficients of determination (R^2) for each of the traits predicted by the top-level DCNN features. Traits are listed along the *y*-axis, and the coefficient of determination between predicted trait ratings and human-generated trait ratings are listed along the *x*-axis. The similarity between human-assigned traits and computer predictions was assessed both in terms of R^2 as well as prediction error (Fig. 5).

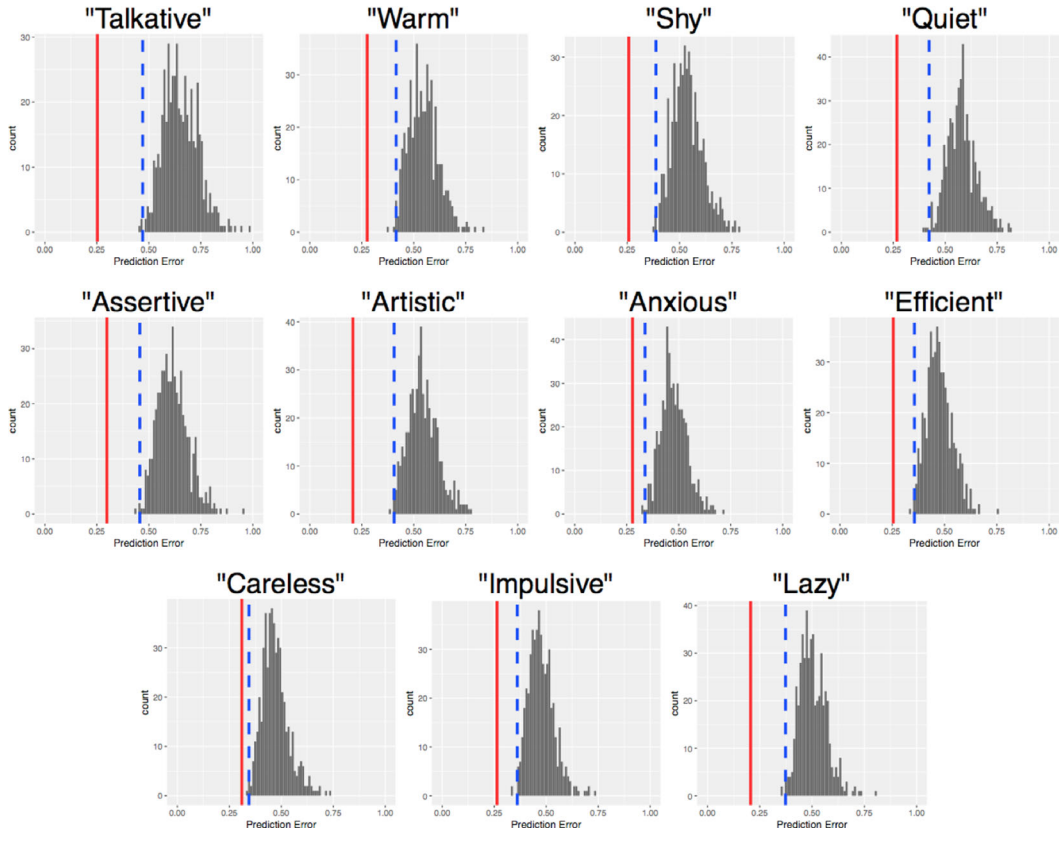

Fig. 5. Trait prediction error (difference from human ratings) when using the actual model (red line) versus a null distribution (gray). The dashed blue line represents a Bonferroni-corrected alpha level of $\alpha = 0.00225$ (0.025/11) on the null distribution. All traits were predicted with significantly less error when using the actual model than when using the null distribution.

impulsive, *anxious*, and *warm*. These social traits were predicted at levels comparable to the subjective trait predictions reported for the three traits examined in McCurrie et al. (2017). In the present case, however, predictions were made from a network trained for face identification, whereas in McCurrie et al. (2017), the network was trained explicitly for trait prediction.

5. Discussion

Human faces provide a rich source of information that can be used to identify individuals, as well as to form first impressions of strangers. Both topics have received considerable attention in the literature, but they have never been investigated simultaneously. In this study, we examined the co-existence of identity and trait information in faces. The

main findings are as follows. First, we demonstrate that information related to social and personality traits is retained in the top-level features of a DCNN trained for face identification. This finding complements previous work illustrating that DCNNs retain specific information about images that is not directly relevant for object/face recognition (Hong, Yamins, Majaj, & Dicarlo, 2016; Parde et al., 2017). Second, we show that DCNN face "identity" representations predict human-generated social-trait inferences, both at the individual-trait level and at the level of full trait profiles. Third, we show that the DCNN used in this study identifies faces with high accuracy, allowing us to consider the retention of social trait information in the context of a high-performing identification system. Fourth, we verify that the particular social trait ratings we collected in this study, which differ somewhat from previous studies, produce a social-trait space similar to those reported in previous work (Oosterhof & Todorov, 2008; Walker & Vetter, 2009).

To understand the implications of these findings, it is useful to return to what we know about the nature of the physical information that supports trait inferences and to expand our discussion beyond simple face structure. The literature offers evidence that three sources of information contribute to social-trait inferences (Sutherland et al., 2013): face structure (Oosterhof & Todorov, 2008; Walker & Vetter, 2009), emotional facial expressions (Said et al., 2009; Sutherland et al., 2013), and image characteristics (Todorov & Porter, 2014). As noted previously, evidence for the role of face structure in trait inferences comes from findings indicating that the manipulation of face structure, using computer graphics models, affects human trait judgments in predictable ways (Oosterhof & Todorov, 2008; Walker & Vetter, 2009). The fact that face structure contributes substantially to both face identity and trait inferences is perhaps not surprising, but makes it clear that information relevant for one of these tasks may also be relevant for the other.

In addition to face structure, both posed facial expressions and expressions that are derived from the structure of neutral faces have been shown to influence trait ratings (Oosterhof & Todorov, 2008). For example, Montepare and Dobish (2003) showed that deliberate changes in the emotional expression of a face influence social trait ratings. These differences in the perception of social traits reflect the structural resemblance of a face to a specific emotion (Said et al., 2009). Further, a recent study by Todorov and Porter (2014) showed that image characteristics, which include photometric variables such as illumination and image quality, are linked to social-trait perception. It has also been shown that DCNNs trained for face identification retain information pertaining to these image characteristics (Parde et al., 2017). Together, these studies provide further evidence that identity and social traits can coexist in a unitary representation. In this study, face emotion and image characteristics were controlled to reduce the noise associated with social-trait judgments. Because of this, the present results may underestimate the extent to which it is possible to predict social-trait inferences from more naturalistic stimuli (see Song et al., 2017 for an approach that allows face expression to support trait classification).

Returning to the question of how trait and identity information are related, it is important to remember that face structure is relevant for face identification, but that facial

expression and image characteristics are not. The finding that facial expression and image characteristics contribute to trait inferences makes it clear that first impressions are based not only on the face itself, but also on the specific circumstances in which a face is encountered. Although the behavioral and neural literatures reflect a conceptual division between social-trait and identity processing (Haxby, Hoffman, & Gobbini, 2000), advances in modeling face identification with DCNNs demonstrate that this division is not computationally necessary to account for the creation of a unitary face representation that can support both identity and trait perception.

DCNNs are designed to model the computational processes seen in the primate ventral visual stream (Fukushima, 1988; Krizhevsky et al., 2012). There are some questions concerning the validity of the claim that these networks are good models of ventral stream visual processing, given that their architectures are not well-suited for complex, sequential behavior (Edelman, 2016). However, research has shown that the response properties of units in the intermediate layers of a DCNN predict the response properties of neurons in primate visual area V4. Moreover, units in the top level of the DCNN predict the responses of neurons in the inferotemporal (IT) cortex (Yamins & DiCarlo, 2016; Yamins et al., 2014). Although interpreting the similarities between the ventral visual stream and DCNNs requires caution, the finding that these networks can accommodate both identity and social-trait information is consistent with recent evidence suggesting specific modifications to the distributed model of face processing (Haxby et al., 2000).

The distributed model posits a functional division of identity and social information from faces, whereby invariant identity information from faces (e.g., face structure) is processed in ventral stream face areas, and social (e.g., changeable aspects, including emotion, gaze, and head rotation) information is processed in the dorsal stream face areas (posterior Superior Temporal Sulcus). Although this theory is consistent with much existing data on the neural processing of faces (Allison, Puce, & McCarthy, 2000; Gobbini & Haxby, 2007; Puce, Allison, Bentin, Gore, & McCarthy, 1998), recent work suggests that the ventral pathway is also sensitive to some changeable information from faces. For example, Duchaine and Yovel (2015) suggest that the ventral temporal cortex processes emotional expression. More concretely, it has been shown that the fusiform face area (FFA) responds to faces regardless of whether people attend to the expression of the faces they view (Ganel, Valyear, Goshen-Gottstein, & Goodale, 2005). In addition, the intensity of observed facial expressions modulates neural responses in the FFA (Surguladze et al., 2003).

Further work is required to determine whether these responses arise purely from faces, or from faces and bodies in motion. However, combined with the present results, these neural findings advance support of the distributed model modification suggested by Duchaine and Yovel (2015). Specifically, this modification posits that ventral temporal representations support both identity and social judgments about faces. This modification does not alter the distributed model's original hypothesis that social information from facial motion is processed in the dorsal stream (pSTS). Rather, it suggests that both ventral and dorsal stream face areas can play a role in the social processing of faces. In considering DCNN properties in this context, it is worth noting that there is now strong evidence that

DCNNs retain specific information about face images, in addition to face categories (e.g., identity). That DCNNs are able to generalize well is paradoxical, given their enormous degrees of freedom. The topic of why DCNNs generalize is discussed in depth in O'Toole, Castillo, Parde, Hill, and Chellappa (2018) and C. Zhang, Bengio, Hardt, Recht, and Vinyals (2017). Presently, the surprising generalizability of DCNNs implies that these networks have the potential to model trait inferences based on expression and image-characteristics in addition to facial structure (cf., O'Toole et al., 2018; Parde et al., 2017). Future work should consider ways to dissect the sources of face-identity and face-image information that allow DCNNs to predict social trait inferences (see methods proposed by Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015).

In conclusion, we present the novel finding that social-trait information and identity information are not independent from one another. Understanding how social traits and identity are linked can provide clues to the structure of high-level visual information processing in general face perception. The simple and direct linear classification methods used in this study underscore the fact that trait information is readily accessible within the representations produced by DCNNs trained for face identification. The presence of this trait information may constrain theories of how neural codes for faces can serve multiple face processing tasks.

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

AbdAlmageed, W., Wu, Y., Rawls, S., Harel, S., Hassner, T., Masi, I., Choi, J., Lekust, J., Kim, J., Natarajan, P., & Nevatia, R. (2016). Face recognition using deep multi-pose representations. *2016 IEEE Conference on Applications of Computer Vision (WACV)*, (1), 1–9. https://doi.org/10.1109/WACV.2016.7477555.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: Role of the STS region. *Trends in Cognitive Sciences*, *4*(7), 267–278. https://doi.org/10.1016/S1364-6613(00)01501-1.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278. https://doi.org/10.1037/1528-3542.6.2.269.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques — SIGGRAPH '99* (pp. 187–194). https://doi.org/10.1145/311535.311556.

Bruce, V., & Young, A. (1986). Understanding face recognition Copyright. *British Journal of Psychology*, *77* (77), 305–327.

Chen, J. C., Patel, V. M., & Chellappa, R. (2016). Unconstrained face verification using deep CNN features. *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016*. https://doi.org/10.1109/wacv.2016.7477557

Cloutier, J., Mason, M. F., & Macrae, C. N. (2005). The perceptual determinants of person construal: Reopening the social-cognitive toolbox. *Journal of Personality and Social Psychology*, *88*(6), 885–894. https://doi.org/10.1037/0022-3514.88.6.885.

Duchaine, B., & Yovel, G. (2015). A revised neural framework for face processing. *Annual Review of Vision Science*, *1*(1), 393–416. https://doi.org/10.1146/annurev-vision-082114-035518.

Edelman, S. (2016). The minority report: Some common assumptions to reconsider in the modelling of the brain and behaviour. *Journal of Experimental & Theoretical Artificial Intelligence*, *28*(4), 751–776. https://doi.org/10.1080/0952813X.2015.1042534.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2006). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual. *Pattern Recognition*, *1*, 119–130.

Ganel, T., Valyear, K. F., Goshen-Gottstein, Y., & Goodale, M. A. (2005). The involvement of the "fusiform face area" in processing facial expression. *Neuropsychologia*, *43*(11), 1645–1654. https://doi.org/10.1016/j.neuropsychologia.2005.01.012.

Gobbini, M. I., & Haxby, J. V. (2007). Neural systems for recognition of familiar faces. *Neuropsychologia*, *45*(1), 32–41. https://doi.org/10.1016/j.neuropsychologia.2006.04.015.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, *37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1.

Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, *78*(5), 837–852. https://doi.org/10.1037/0022-3514.78.5.837.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223–233. https://doi.org/10.1016/S1364-6613(00)01482-0.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv Preprint arXiv:1207.0580*. 1–18. https://doi.org/arxiv:1207.0580

Hong, H., Yamins, D. L. K., Majaj, N. J., & Dicarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622. https://doi.org/10.1038/nn.4247.

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. *University of Massachusetts Amherst Technical Report*, *1*, 7–49. https://doi.org/10.1.1.122.8268

Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., & Jain, A. K. (2015). Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07–12–June*, (pp. 1931–1939). https://doi.org/10.1109/cvpr.2015.7298803

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1–9. https://doi.org/10.1016/j.protcy.2014.09.007

Lewenberg, Y., Bachrach, Y., Shankar, S., & Criminisi, A. (2016). Predicting personal traits from facial images using convolutional neural networks augmented with facial landmark information. In *IJCAI International Joint Conference on Artificial Intelligence, January* (pp. 1676–1682).

Macrae, C. N., Quinn, K. A., Mason, M. F., & Quadflieg, S. (2005). Understanding others: The face and person construal. *Journal of Personality and Social Psychology*, *89*(5), 686–695. https://doi.org/10.1037/0022-3514.89.5.686.

McCurrie, M., Beletti, F., Parzianello, L., Westendorp, A., Anthony, S., & Scheirer, W. J. (2017). Predicting first impressions with deep learning. In *Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017* (pp. 518–525). https://doi.org/10.1109/fg.2017.147

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior*, *27*(4), 237–254. https://doi.org/10.1023/A:1027332800296.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105.

O'Toole, A. J., Castillo, C. D., Parde, C., Hill, M. Q., & Chellappa, R. (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Sciences*, *22*(9), 794–809.

O'Toole, A. J., Harms, J., Snow, S. L., Hurst, D. R., Pappas, M. R., Ayyad, J. H., & Abdi, H. (2005). A video database of moving faces and people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 812–816. https://doi.org/10.1109/TPAMI.2005.90.

Parde, C. J., Castillo, C., Hill, M. Q., Colon, Y. I., Sankaranarayanan, S., Chen, J. C., & Otoole, A. J. (2017). Face and image representation in deep CNN features. In *Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017* (pp. 673–680). https://doi.org/10.1109/fg.2017.85

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British Machine Vision Conference 2015 (Section 3)* (pp. 41.1–41.12). https://doi.org/10.5244/c.29.41

Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *The Journal of Neuroscience*, *18*(6), 2188–2199. https://doi.org/10.1523/jneurosci.2161-10.2010.

Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. (2017). An all-in-one convolutional neural network for face analysis. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (pp. 17–24). https://doi.org/10.1109/fg.2017.137

Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*(2), 260–264. https://doi.org/10.1037/a0014681.

Sankaranarayanan, S., Castillo, C., Alavi, A., & Chellappa, R. (2016). Triplet similarity embedding for face verification. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (pp. 1–8). https://doi.org/10.1109/btas.2016.7791205

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07–12–June* (pp. 815–823). https://doi.org/10.1109/cvpr.2015.7298682

Song, A., Linjie, L., Atalla, C., & Cottrell, G. W. (2017). Learning to see people like people: Predicting social impressions of faces. arXiv:1705.04282v1.

Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation by joint identification-verification. *Advances in Neural Information Processing Systems*, 1988–1996, https://doi.org/10.1109/CVPR.2014.244.

Surguladze, S. A., Brammer, M. J., Young, A. W., Andrew, C., Travis, M. J., Williams, S. C. R., & Phillips, M. L. (2003). A preferential increase in the extrastriate response to signals of danger. *NeuroImage*, *19*(4), 1317–1328. https://doi.org/10.1016/S1053-8119(03)00085-5.

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127* (1), 105–118. https://doi.org/10.1016/j.cognition.2012.12.001.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1701–1708). https://doi.org/10.1109/cvpr.2014.220

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, *25*(7), 1404–1417. https://doi.org/10.1177/0956797614532474.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460. https://doi.org/10.1016/j.tics.2008.10.001.

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353–E3361. https://doi.org/10.1073/pnas.1409860111.

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, *9*(11), 1–13. https://doi.org/10.1167/9.11.12.

Wiggins, J. S., Phillips, N., & Trapnell, P. (1989). Circular reasoning about interpersonal behavior: Evidence concerning some untested assumptions underlying diagnostic classification. *Journal of Personality and Social Psychology*, *56*(2), 296–305.

Wolf, L., Hassner, T., & Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 529–534.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244.

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*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111.

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Learning face representation from scratch. *arXiv Preprint arXiv:1411.7923.* Available at: http://arxiv.org/abs/1411.7923

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *Deep Learning Workshop, 31st International Conference on Machine Learning*, 1–12.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). Understanding deep learning requires re-thinking generalization. arXiv Preprint arXiv:1611.03530.

Zhang, N., Paluri, M., Taigman, Y., Fergus, R., Bourdev, L., & Berkeley, U. C. (2015). Beyond frontal faces: Improving person recognition using multiple cues. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4804–4813.

Zhou, E., Cao, Z., & Yin, Q. (2015). Naive-deep face recognition: Touching the limit of LFW benchmark or not? *arXiv Preprint arXiv:1501.04690.* https://doi.org/10.1103/physrevd.91.045023