Closing the gap between single-unit and neural population codes: Insights from deep learning in face recognition

Connor J. Parde School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Y. Ivette Colón School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Matthew Q. Hill School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Carlos D. Castillo University of Maryland Institute of Advanced Computer Studies, University of Maryland, College Park, MD, USA

Prithviraj Dhar University of Maryland Institute of Advanced Computer Studies, University of Maryland, College Park, MD, USA

Alice J. O'Toole School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, TX, USA

Single-unit responses and population codes differ in the "read-out" information they provide about high-level visual representations. Diverging local and global read-outs can be difficult to reconcile with in vivo methods. To bridge this gap, we studied the relationship between single-unit and ensemble codes for identity, gender, and viewpoint, using a deep convolutional neural network (DCNN) trained for face recognition. Analogous to the primate visual system, DCNNs develop representations that generalize over image variation, while retaining subject (e.g., gender) and image (e.g., viewpoint) information. At the unit level, we measured the number of single units needed to predict attributes (identity, gender, viewpoint) and the predictive value of individual units for each attribute. Identification was remarkably accurate using random samples of only 3% of the network's output units, and all units had substantial identity-predicting power. Cross-unit responses were minimally correlated, indicating that single units code non-redundant identity cues. Gender and viewpoint classification required large-scale pooling of units—individual units had weak predictive power. At the ensemble level, principal component analysis of face representations showed that identity, gender, and viewpoint separated into high-dimensional subspaces, ordered by explained variance. Unit-based directions in the representational space were compared with the directions associated with the attributes. Identity, gender, and viewpoint contributed to all individual unit responses, undercutting a neural tuning analogy. Instead, single-unit responses carry superimposed, distributed codes for face identity, gender, and viewpoint. This undermines confidence in the interpretation of neural representations from unit response profiles for both DCNNs and, by analogy, high-level vision.

Introduction

The concept of a *feature* is at the core of psychological and neural theories of visual perception. The link between perceptual features and neurons has been an axiom of visual neuroscience since Lettvin et al. (1959) first described the receptive fields of ganglion cells as "bug perceivers." At low levels of visual processing, features are well defined and interpretable (e.g., vertical line, retinal location x). These feature codes are sparse, because they rely on a small number of specific neurons (Olshausen & Field, 1997). At higher levels of visual processing, where retinotopy gives way to categorical codes, the connection between receptive fields and features is unclear. "Face-selective" may describe a neuron's receptive field, but it provides

Citation: Parde, C. J., Colón, Y. I., Hill, M. Q., Castillo, C. D., Dhar, P., & O'Toole, A. J. (2021). Closing the gap between single-unit and neural population codes: Insights from deep learning in face recognition. *Journal of Vision*, *21*(8):15, 1–14, https://doi.org/10.1167/jov.21.8.15.

https://doi.org/10.1167/jov.21.8.15 Received December 18, 2020; published August 11, 2021 ISSN 1534-7362 Copyright 2021, The Authors

no information about the features used to encode a face.

A fundamental difference between retinotopic representations in early visual areas and the categorical representations that emerge in inferotemporal (IT) cortex is that the latter generalize across image variation (e.g., viewpoint). Image-based strategies analogous to those used in low-level vision were applied in early face-recognition algorithms (Turk & Pentland, 1991), which operated well on face images with limited variation in viewing conditions. Since 2014, deep convolutional neural networks (DCNNs) have overcome the limit of recognizing faces using image-based similarity (Sun et al., 2014; Schroff et al., 2015; Taigman et al., 2014). Similar to face codes in high-level visual cortex, DCNNs generalize over substantial image variation. Notably, neural codes for objects in IT cortex can be simulated using weighted combinations of DCNN output units, consistent with a "population doctrine" of neural coding (Eichenbaum, 2018; Saxena & Cunningham, 2019; Yamins et al., 2014; Yuste, 2015).

The parallels between primate vision and deep learning networks are by design (Fukushima, 1988; Krizhevsky et al., 2012). Here, we use a DCNN model to gain insight into how single-unit and population codes interface. DCNNs employ computational strategies similar to those used in the primate visual system and are trained extensively with real-world images. Although these networks have made significant progress on the problem of generalized face recognition, the face representation they create is poorly understood (Poggio et al., 2020). Approaches to dissecting this representation have been aimed at (a) uncovering the information retained in the descriptor (Hong et al., 2016; Parde et al., 2017), (b) probing the robustness of individual unit responses to image variation (Parde et al., 2017), (c) visualizing the receptive fields of *individual units* in the network code (Qin et al., 2018), and (d) visualizing the similarity structure of a population of *ensemble face representations* for images and identities (Hill et al., 2019). We consider each in turn.

First, it is now clear that face descriptors from DCNNs trained to identify faces retain a surprising amount of information about the original input image (Parde et al., 2017). Specifically, the output representation from DCNNs trained for face identification can be used to predict the viewpoint (yaw and pitch) and media type (still or video image) of the input image with high accuracy (O'Toole et al., 2018; Parde et al., 2017), as well as the illumination conditions (ambient or spotlight; Hill et al., 2019). Analogously, for object recognition, Hong et al. (2016) found that inferotemporal cortex in macaque ventral visual stream retains information about "categorically orthogonal" properties of the visual stimulus (e.g., pose and size), in addition to information about object category. Hong et al. (2016) also found that the top layer of a DCNN trained for object recognition retained this category-orthogonal information also. Therefore, deep networks achieve robust face/object identification, not by filtering out image-based information across layers of the network but by effectively managing it (DiCarlo & Cox, 2007; Hong et al., 2016). These findings are consistent with neuro-computational theory positing that ventral visual stream processing "untangles" face identity information from image parameters over successive layers of neural processing (DiCarlo & Cox, 2007). This perspective applies also to face identity processing. Consistent with the disentangling theory, DCNNs trained for identification produce a multipurpose representation of a face that can encode a wide range of functionally useful information (O'Toole & Castillo, 2021).

Second, given that DCNN descriptors contain image information, it is possible that the top-layer units separate identity and image information across different units of the face descriptor. Parde et al. (2017) tested this by probing the response properties of the top-layer units in a DCNN trained for face identification to either front-facing or three-quarter-view images of faces. Individual units did not respond consistently in either a view-specific or view-independent manner.

The third approach is to visualize the response preferences of units in the network (Qin et al., 2018) with the goal of translating them into perceptible images (Erhan et al., 2009; Ponce et al., 2019; Zeiler et al., 2011). This approach is useful for interpreting hidden units at lower layers of DCNNs, where unit activations can be linked to spatial locations within an input image. At higher levels of the network, however, unit responses are not bound to image locations and so image-based visualization may offer limited, and possibly misleading, insight into the underlying nature of the code (O'Toole & Castillo, 2021).

The fourth approach is to visualize the similarity structure of *ensembles* of DCNN unit activations, which correspond to the face descriptor codes for individual images. This reveals a highly organized *face space* (O'Toole et al., 2018; Valentine, 1991). Visualization was applied to face images of multiple identities that varied systematically in viewpoint and illumination (Hill et al., 2019). The resulting face space showed that images clustered by identity, identities separated into regions of male and female faces, illumination conditions (ambient vs. spotlight) nested within identity clusters, and viewpoint (frontal to profile) nested within illumination conditions. Deep networks trained with in-the-wild images, therefore, generate a highly structured representation of identity and image variation.

These four approaches focus either on single units or ensembles. Neither provides a complete account of how unit responses interact in a representational space to code information about faces. Here we used a deep learning network as a model of neural units in high-level visual cortex to examine the juxtaposition of single-unit and ensemble codes for face identity, gender, and viewpoint. At the unit level, we probed the distribution of face and image information in the network's face descriptors. We tested identification, gender classification, and viewpoint estimation in variably sized, randomly sampled subspaces of top-layer units from a DCNN trained for face identification. We then examined the minimum number of units needed for each task and the predictive power of individual units. At the ensemble level, we examined identity, gender, and viewpoint codes as directions in the representational space. We performed principal component analysis (PCA) on the face-descriptor vectors and analyzed the identity, gender, and viewpoint information coded by each principal component (PC). Combining neural units and ensembles, we examined the relationship between directions in the representational space defined by units and those defined by identity, gender, and viewpoint. We show that identity and image information commingle in individual unit responses but separate in the ensemble space. This challenges classical tuning analogies for neural units at high levels of visual processing (Hasson et al., 2020).

Method

Network

All data reported in the main text of the article are from a 101-layered face-identification DCNN based on a ResNet-101 architecture (He et al., 2016; Ranjan et al., 2017). This network performs with high accuracy across changes in viewpoint, illumination, and expression (cf. performance on IARPA Janus Benchmark-C [IJB-C]; Maze et al., 2018). Specifically, the network is based on the ResNet-101 (Wen et al., 2016) architecture. It was trained with the Universe dataset (Bansal et al., 2017a; Ranjan et al., 2018), which comprises three smaller datasets (UMDFaces: Bansal et al., 2017b; UMDVideos: Bansal et al., 2017a; MS-Celeb-1M: Guo et al., 2016) and includes 5,714,444 images of 58,020 identities. The network employs Crystal Loss (L2 Softmax) for training (Ranjan et al., 2018). Crystal Loss scale factor α was set to 50. ResNet-101 employs skip connections to retain the strength of the error signal over its 101-layer architecture (He et al., 2016). Once the training is complete, the final layer of the network is removed and the penultimate layer (512 units) is used as the identity descriptor. To test whether the results generalize across variation in network architecture, we performed a full replication of the reported results using an alternative face-identification network with a different architecture (Ranjan et al., 2018; Xie et al., 2017) (see Supplementary Section S1).

Face images

The test set consisted of images from the IJB-C dataset, which contains 3,531 subjects portrayed in 31,334 still images (10,040 non face images) and frames from 11,779 videos. For the present experiments, we used all still images in which our network could detect at least one face and for which viewpoint information was available. In total, we selected 22,248 (9,592 female; 12,656 male) face images of 3,531 (1,503 female; 2,028 male) identities. In images that contained multiple detectable faces, the image was segmented, and each face was considered independently.

Procedure

DCNN-generated representations of face images were obtained by processing the images through the DCNN to produce a 512-dimensional vector of unit activations at the penultimate layer of the network. This was considered the full-dimensional DCNN representation. The distribution of identity, gender, and viewpoint was examined across units in randomly sampled subspaces of varying dimensionalities (512, 256, 128, 64, 32, 16, 8, 4, and 2 units). For each dimensionality, 50 random samples were selected. In the following, we describe how identification and classification were tested for each face-image attribute in each of the sampled subspaces.

Identification

The DCNN was tested on its ability to determine if a pair of images showed the same identity or different identities. Same- and different-identity image pairs were created as follows. We worked with the available face images in two sets. This was done to avoid an exponential number of simulations as we randomly sampled varying numbers of units and computed identification accuracy for 50 random samples of each. Therefore, images were sampled from the test dataset and assigned randomly to Set A or Set B. Each set (A and B) contained 5,562 images of over 3,000 identities. No images were repeated, and there was substantial overlap of the identities across each set (2,729 identities appeared in both sets). Image pairs were then formed by comparing every image in Set A to every image in Set B, for a total of 30,935,844 image comparisons. Next, face-image descriptors were generated for all images in Sets A and B by processing the images through the DCNN. Similarity scores were computed for each image pair by measuring the cosine similarity of the DCNN-generated face-image descriptors. The similarity scores for same- and different-identity image pairs were used to compute a receiver operating characteristic (ROC) curve, and the area under the ROC (AUC) was used as a measure of accuracy.

Classification

Gender

Linear discriminant analysis (LDA) was used to classify face gender for each image in the dataset from the DCNN-generated representation. For each subspace simulation, an LDA was trained using the DCNN-generated representations of all images of 3,231 identities and was tested on a set of 300 held-out identities. This process was repeated, holding out a different set of 300 identities until all images were classified. Gender labels for each identity were verified by human raters. The final output values from the classifier were categorical gender predictions that were compared directly to the human-verified gender labels.

Viewpoint

Viewpoint was predicted using linear regression. Regression models were computed using the Moore–Penrose pseudo-inverse. We chose a linear model, rather than fitting a U-shaped function, because the DCNN we used (Ranjan et al., 2018) processed each face image twice, once in the original form and once after performing a symmetrical left–right flip (i.e., over the y-axis). The two face descriptors were then averaged together to form the face representation we analyze.[1] Therefore, viewpoint is functionally defined in terms of absolute value, from 0 to 113 degrees.

For each subspace simulation, a regression model was trained using the DCNN-generated representations of all images of 3,231 identities and tested on the remaining 300 identities. This process was repeated, each time reserving a different set of 300 identities, until viewpoint predictions had been assigned to each image representation. Ground truth for viewpoint was produced by the Hyperface system (Chen et al., 2016) and was defined as the absolute value of the deviation from a frontal pose, measured in degrees yaw (i.e., 0 = frontal, 90 = profile). Output predictions were continuous values corresponding to the predicted viewpoint in degrees yaw.

Permutations

Permutation tests were used to evaluate the statistical significance of the viewpoint and gender predictions. A null distribution was generated from the original data by randomly permuting values within each unit. Predictions made from the resulting permutations ($n = 1{,}000$) were compared to the true values from each classification test. All permutation tests were significant at $p < 0.001$, with no overlap between test value and null distribution.

Results

Unit-level analysis

Face identity

Identification accuracy proved robust in low-dimensional subspaces. Figure 1A shows face-identification accuracy as a function of the number of randomly selected units sampled from the face representation. Accuracy is near perfect in the full-dimensional space. A substantial number of units can be deleted with almost no effect on the identification

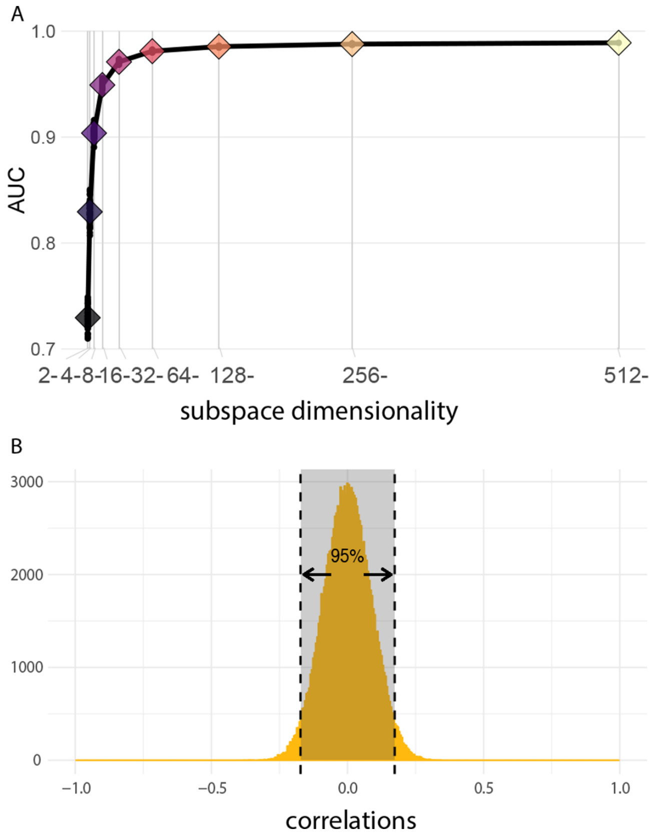

Figure 1. (A) Identification accuracy is plotted as a function of subspace dimensionality, measured as area under the ROC curve (AUC). Performance is nearly perfect (AUC ≈1.0) with the full 512-dimensional descriptor and shows negligible declines until subspace dimensionality reaches 16 units. Performance with as few as two units remains above chance. (B) Correlation histogram for unit responses across images indicates that units capture non redundant information for identification.

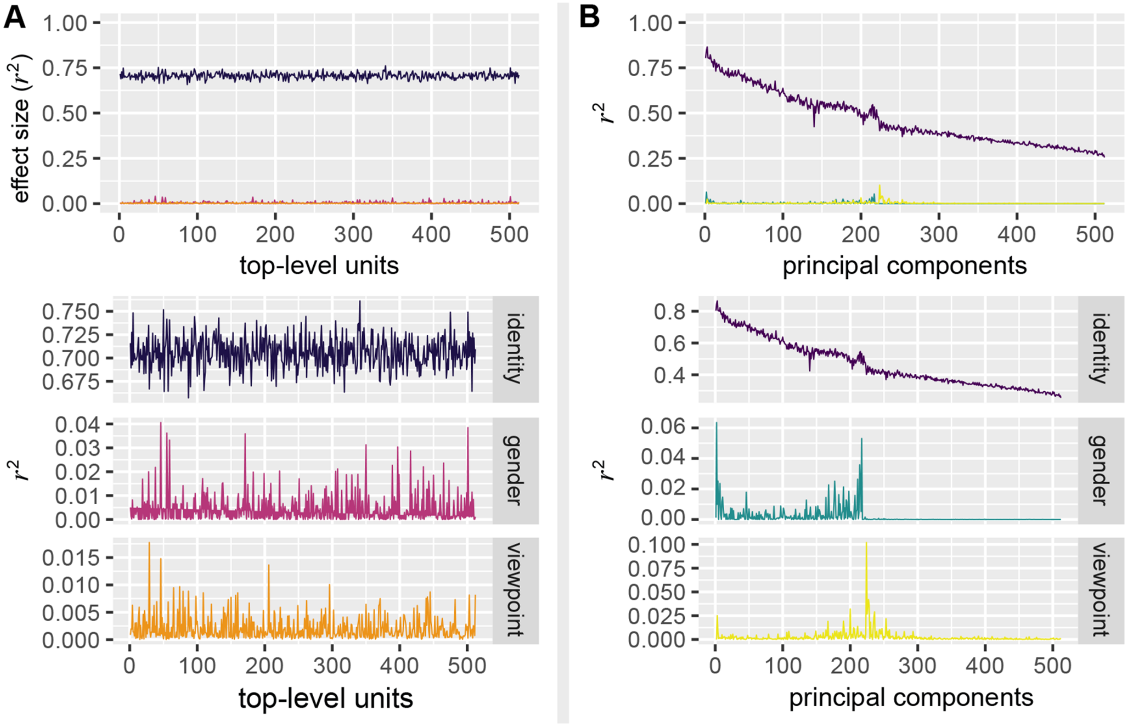

Figure 2. Effect sizes for units (A) and principal components (B) for identity, gender, and viewpoint. For both units and principal components, top panels illustrate the dominance of identity over gender and viewpoint. Lower panels show an approximately uniform distribution of effect sizes for units (A) and differentiated effect sizes for principal components (B) in all three attributes.

accuracy of the 3,000+ individuals in the test set. The first substantial drop in performance is seen at 16 units (≈3% of the full dimensionality). Accuracy remains high with as few as four units (AUC = 0.80) and is well above chance with two units (AUC = 0.72). This demonstrates that DCNN performance is robust with very small numbers of units and does not depend on the particular units sampled.

The remarkable stability of identification accuracy, even when randomly selecting very few top-layer units, is consistent with two types of codes. First, individual units might provide diverse identity cues. Combined, these cues could accumulate into a powerful code for identification. In this case, many individual units would show a measurable capacity for separating identities. Moreover, the identity information in individual units would be minimally correlated. Alternatively, the units might capture redundant, but effective, identity information. By this account, the response patterns of units would be highly correlated.

We found that individual units yield diverse, non redundant, solutions for identity. Figure 1B shows the distribution of response correlations for all possible pairs of top-level units across all images in the test set. The distribution is centered at zero, with 95% of correlations falling below an absolute value of 0.17. Therefore, units in the DCNN are minimally correlated and capture non redundant identity information.

Next, we quantified the identification capacity of individual units in the DCNN. Units with high identification capacity support maximal identity separation while minimizing the distance between same-identity images. Specifically, a unit has identity-separation power when its responses vary more between different-identity images than within sets of same-identity images. We applied the *F* statistic from analysis of variance (ANOVA) to each unit's responses to all images in the test set. The resulting *F* ratios provide an index of between- to within-identity variance. For each ANOVA, identity was the independent variable (with 3,531 levels, equal to one level for each identity in the test set), image was the random (observation) variable, and unit response was the dependent variable. All units separated identities, with *p* values less than 0.000098, which is $\alpha = 0.05$, Bonferroni corrected for 512 comparisons. This corrected alpha level is applied also in the gender and viewpoint simulations.

Next, we calculated the proportion of variance in a unit's response explained by identity variation (r^2 effect size). Figure 2 A (purple) shows the distribution of effect sizes across units (mean $r^2 = 0.691$, minimum = 0.6573, maximum = 0.7611). On average, 69.1% of the variance in individual unit responses is due to identity variation. This indicates that all units have a substantial capacity to separate identities.

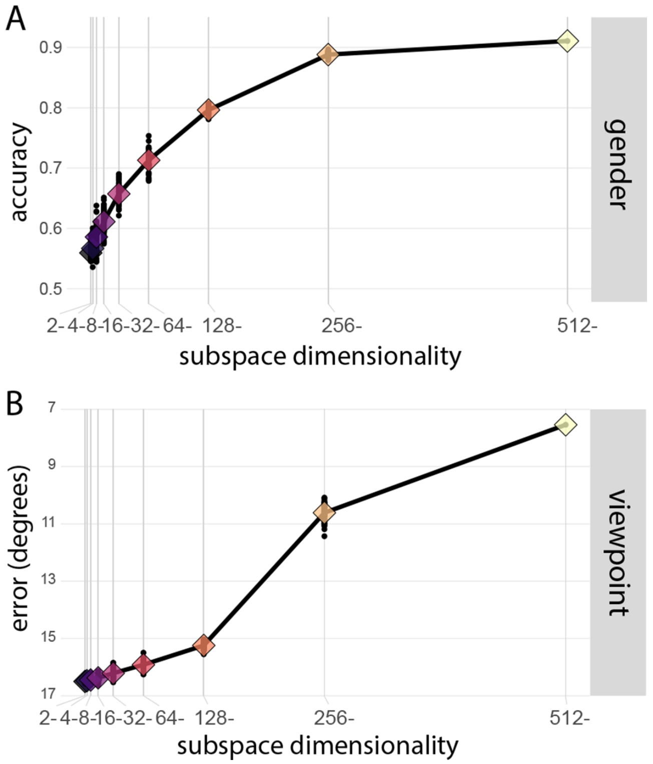

Figure 3. Gender and viewpoint prediction with variable numbers of randomly sampled units. Gender classification declines gradually (A) and viewpoint prediction declines rapidly (B) as sample size decreases. Mean performance across samples ($n = 50$) is shown with a diamond, colored by sample size. Because these performance measures are qualitatively different, they should not be compared in absolute terms (for comparison between gender, viewpoint, and identity, see effect sizes; Figure 2).

Gender

Gender-classification accuracy was measured in the same abbreviated subspaces sampled for the face-identification experiments. For each sample, LDA was applied to predict the gender (male or female) of each image in the test set from the unit responses. Using all units, gender classification was 91.1% correct. Classification accuracy declined steadily as the number of units sampled decreased (Figure 3A).

Next, the gender-separating capacity of each unit was measured. An ANOVA was performed for each unit, using gender as the independent variable. Overall, 71.5% of the units could separate images according to gender ($p < 0.000098$). However, gender accounted for only a small amount of the variance in unit responses. Figure 2A (pink) shows effect size across units (mean $r^2 = .0045$, minimum ≈ 0, maximum $= 0.041$). Notwithstanding the small effect sizes, the proportion of units with gender-separation power (71.5%) is meaningful. Only 5% (α level) of units would be significant for gender separation by chance. Overall, fewer units have predictive power for gender than for identity, and the predictive value of these units for gender is weaker.

Viewpoint

Linear regression models were trained to predict viewpoint using the abbreviated subspaces used for identification and gender classification. Error was measured as the difference between predicted and true yaw (in degrees). Figure 3B shows prediction error as a function of the number of randomly sampled units. Using all units, viewpoint was predicted within 7.35 degrees. Accuracy was at chance when subspace dimensionality fell to 32 units.

The viewpoint-separation capacity of each unit was assessed with ANOVA, using viewpoint as the independent variable. The absolute values of the true yaw measurements were binned into the following five categories: frontal (0°, 18°), near frontal (18°, 36°), half-profile (36°, 54°), near profile (54°, 72°), and profile (72°, 150°). To account for unequal group sizes, pooled sums-of-squares were used as the error term. Figure 2A (orange) shows the ANOVA effect size for each unit (mean $r^2 = 0.0020$, minimum ≈ 0, maximum $= 0.018$). In total, 54.7% of units separated images according to viewpoint ($p < 0.000098$), though effect sizes were small.

Single-unit summary

Multiple neural-like codes coexist within the same set of DCNN top-layer units. These are differentiated by the number of units needed to perform a task and by the predictive power of individual units for the task. First, all units provide strong cues to identity that are largely uncorrelated. Therefore, small numbers of randomly chosen units can achieve robust face identification. Second, gender is coded weakly in approximately 72% of the units. Accurate gender classification requires a larger number of units, because the set must include gender-predictive units, and these units must be combined. Third, even fewer units (about 50%) code viewpoint—each very weakly. Therefore, a large number of units are needed for accurate viewpoint estimation. All single-unit results were replicated using the output units from an alternative face-identification DCNN (see Supplemental Figures S1–S3).

Ensemble analysis: Identity, gender, and viewpoint

How do ensemble face representations encode identity, gender, and viewpoint in the high-dimensional

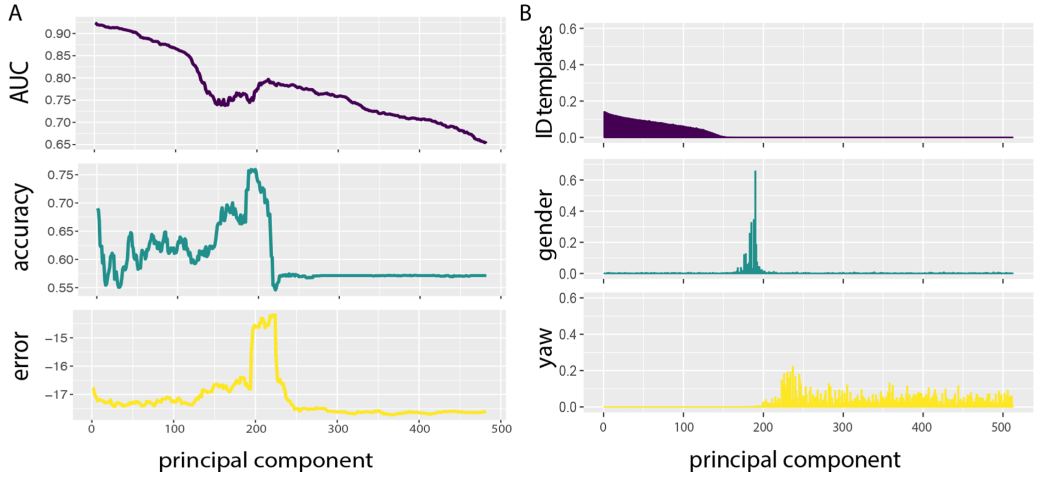

Figure 4. (A) Sliding windows of PCs used to predict identity (purple), gender (teal), and yaw (yellow) across the PC subspaces. Identification accuracy is highest when using early PCs. Gender and viewpoint classification are best when using subspaces with the highest effect sizes for gender and viewpoint separation, respectively. (B) Similarity between PCs and directions diagnostic for identity (purple), gender (teal), and yaw (yellow). Identity direction is the average similarity between identity templates and PCs. Gender direction is the linear discriminant line from the LDA for gender classification. Viewpoint direction is the weight vector from the linear regression for viewpoint prediction.

space created by the DCNN? Interpreting unit-based face-image codes as directions in this space requires a change in vantage point. A face-space representation Valentine (1991; O'Toole et al., 2018) was generated by applying PCA to the full ensemble of unit responses. The axes of the space (PCs) are ordered according to the proportion of variance explained by the ensemble face-image descriptors. We re expressed each face-image descriptor as a vector of PC coordinates. This captures a face-image representation in terms of its relationship to principal directions in the ensemble space.

For each PC, we measured identity, gender, and viewpoint separation using the ensemble-based image code. Effect sizes were computed for the PC-based codes using ANOVA, as was done for the unit-based codes. These appear in Figure 2B. Identity (purple) dominates gender (teal) and viewpoint (yellow) information, consistent with the unit code (Figure 2A). In contrast to the unit-based codes, effect sizes for individual PCs differentiate strongly by attribute. Effect sizes for identity are highest in PCs that explain the most variance in the space. Gender information peaks in two ranges of PCs (≈2–10 and ≈164–202). For viewpoint, effect sizes peak approximately between PCs 220 and 230. Therefore, face-image attributes separate into subspaces ordered roughly according to explained variance. Next, we show that these subspaces differ in their functional capacity to classify identity, gender, and viewpoint and that subspaces are organized to align with directions in the representational space diagnostic of face attributes.

Separation of attributes in the ensemble space could be organizational and/or functional, and measuring the classification capacity of attribute-related subspaces would be further evidence that ensemble codes functionally separate attributes. To test the functional separability of face-image attributes in different subspaces, we used sliding windows of 30 PCs at a time (1–30, 2–31, 3–32, etc.) to predict each attribute. Figure 4A shows that the accuracy of predictions for the three attributes differs with the PC range. As expected from an identity-trained system, identification accuracy is best in the subspaces that explain the most variance. Gender-classification accuracy is highest when using ranges of PCs that encompass the highest effect sizes for gender separation. Similarly, viewpoint prediction is most accurate with ranges of PCs that encompass the highest effect sizes for viewpoint separation.

To examine the organization of attributes in the ensemble code, we measured the alignment of face attributes with directions in the space. Specifically, we compared PC directions to directions diagnostic of identity, gender, and viewpoint. Identity direction was calculated by averaging the face descriptors for all images of an identity. Gender direction was the linear discriminant line from the LDA used for gender classification. Viewpoint direction was the vector of regression coefficients for viewpoint prediction.

Figure 4B (purple) shows the average of the absolute value of cosine similarities between each PC and all identity codes. Figure 4B (teal) shows the similarity between each PC and the gender direction, and Figure 4B (yellow) shows the similarity between each PC and the viewpoint direction. These plots reveal that identity information is distributed primarily across the first ≈150 PCs, gender information is distributed primarily across PCs ranked between 150 and 200, and viewpoint information is distributed primarily across PCs ranked greater than 200.

Consistent with the effect sizes computed for each PC, as well as the attribute predictions, this result shows that identity, gender, and viewpoint separate roughly into subspaces ordered according to explained variance in the DCNN-generated ensemble space. These subspaces separate attributes both organizationally and functionally and reflect a prioritization of identity over gender and of gender over viewpoint. This prioritization likely results from the optimization goals of the network and the combination of the statistical structure of the face image and full data set processed by the network. As the optimized variable, identity dominates the organization of the ensemble space. For viewpoint and gender, parceling out the contribution of the face image and the population of face images processed by the network would be challenging and would require additional data. We return to the question of optimization in the Discussion.

As with the single-unit results, all ensemble code results were replicated using an alternative network architecture (see Supplemental Figures S2 & S4). Additional details regarding the PCA of the full face space (including a scree plot) are included in the supplemental materials (see Supplemental Information Section S2).

Juxtaposed unit and ensemble codes

PCs capture directions that can be interpreted in terms of identity, gender, and viewpoint. How do these directions relate to the basis vectors that define the DCNN units? The answer to this question will tell us whether individual units "respond preferentially" to specific face-image attributes.

We calculated the cosine similarity between the PC directions and the unit directions (Unit 1: $[1, 0, 0 \ldots 0_{512}]$, Unit 2: $[0, 1, 0 \ldots 0_{512}]$, etc.). If a unit responds preferentially to viewpoint, gender, or identity, it will align closely with PCs related to a specific attribute. Alternatively, if semantically interpretable information (identity, gender, viewpoint) is confounded in a unit's response, it will yield a uniform distribution of similarities across the PCs related to each attribute.

The results indicate that unit responses confound identity, gender, and viewpoint. Figure 5 (top) shows a uniform distribution of similarities across PCs for a single unit. We found this for all of the 512 units (see Supplemental Information Section S3, or Extended Data on the Open Science Framework: https://osf.io/xewuk/). Figure 5 (bottom) shows a density plot of these similarities, separated by attribute. Identity, gender, and viewpoint information, which are separated in the high-dimensional space, are confounded in single-unit responses. Consistent with previous results, this finding replicated with an alternative face-identification DCNN (see Supplemental Figure S5). This undermines a classic tuning analogy for units. In isolation, individual units cannot be interpreted in terms of a specific attribute.

Discussion

Historically, neural codes have been characterized as either sparse or distributed. Sparse codes signal a stimulus via the activity of a small number of highly predictive units. Distributed codes signal a stimulus via the combined activity of many weakly predictive units. The DCNN's identity code encompasses fundamental mechanisms of both sparse (highly predictive single units) and distributed (powerful combinations of units) codes. This unusual combination of characteristics accounts for the DCNN's remarkable resilience to deleting units in the face representation (for the resilience of other parts of a DCNN, see Casper et al., 2019). Superimposed on the identity representation are standard distributed codes for gender and viewpoint and likely other subject and image variables. For these codes, ensembles, not individual units, make accurate attribute predictions.

The results reveal three distinct attribute codes (identity, gender, viewpoint) in one set of units. These codes vary in the extent to which they distribute information across units. Because multiple attribute codes share the same units, the label "sparse" or "distributed" must specify a particular attribute. In deep (higher) layers of DCNNs, where units respond to abstract combinations of low-level visual features, these shared codes may be common. If these codes exist in the primate visual system, they would likely be at higher levels of the visual processing hierarchy. In low-level visual areas (e.g., V1), neural receptive fields refer to locations in the retinotopic image and are more likely to act as single-attribute "feature detectors." We will return to this point shortly in the context of what a neurophysiologist would find in probing units in the DCNN.

Much of what appears complex in individual units is clear in the ensemble space. PCs separate attributes in the DCNN representation according to explained variance. This reflects network prioritization (identity > gender > viewpoint). PCs comprise a "special," interpretable rotation of the unit axes, because the face attributes are not represented equally in the face descriptors. The juxtaposition of unit and ensemble characteristics indicates that information coded by a deep network is in the representational space, not in any given projection of the space onto a particular set of axes (cf. Hasson et al., 2020); Szegedy et al. 2013.

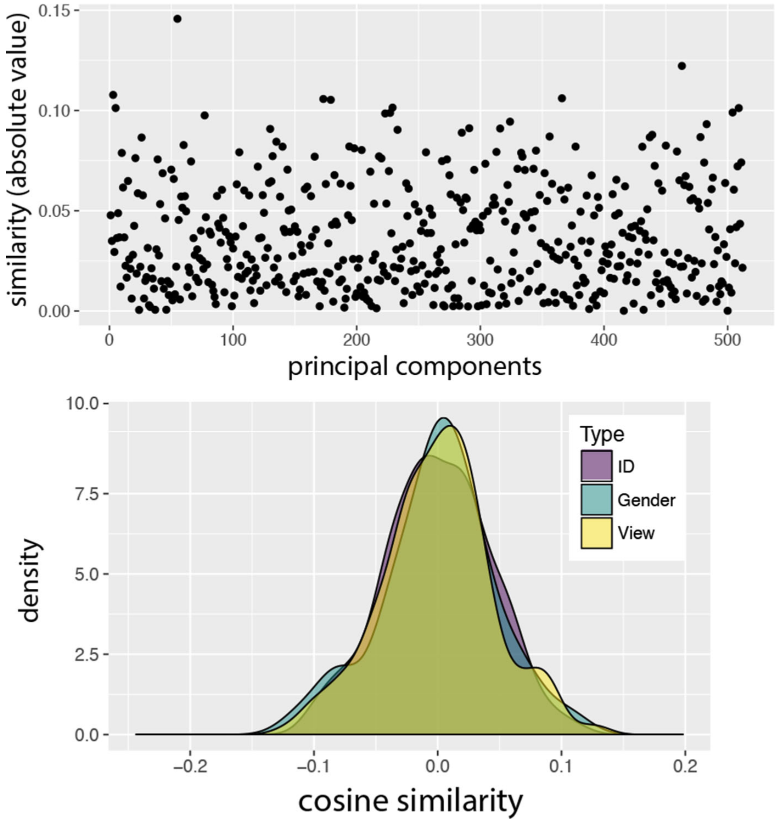

Figure 5. (Top) For a single example unit, absolute value of similarities between unit direction and each PC shows confounding of unit response with identity, gender, and viewpoint. (Bottom) Density plot of similarities between the example unit and PCs associated with identity (purple), gender (blue), and viewpoint (yellow). The distributions overlap almost completely, indicating that each type of information contributes to the unit's activation. This finding was consistent across all unit basis vectors.

How, then, are we to understand the units? The DCNN is optimized to separate identity, not to maximize the interpretability of the information that achieves identity separation. From a computational perspective, any orthogonal basis is as good as any other. Given the high dimensionality of the system and the stochastic nature of training, the likelihood of a DCNN converging on a semantically interpretable basis set is exceedingly low. Units serve the sole purpose of providing a set of basis axes that support maximal separation of identities in the space. There should be no expectation that the response of individual units be "tuned" to semantic features (Hasson et al., 2020). In isolation, units provide little or no information about the visual code that operates in the high-dimensional space. Instead, units must be interpreted in an appropriate population-based computational framework (Eichenbaum, 2018; Hasson et al., 2020; O'Toole et al., 2018; O'Toole & Castillo, 2021; ; Saxena & Cunningham, 2019; Yamins et al., 2014; Yuste, 2015).

How does this affect the way we interpret neural data? The literature is replete with reports of preferentially tuned neurons in face-selective cortex. Electrophysiological recordings differentiate face patches based on the tuning of neurons (e.g., PL: eyes, eye region, face outlines [Issa & DiCarlo, 2012]; ML: iris size, inter eye distance, face shape, and face views [Freiwald & Tsao, 2010]; AM: view-invariant identity [Freiwald & Tsao, 2010]; ML-MF: face shape parameters extracted from a computationally based active appearance model [Chang & Tsao, 2017]; and AM: appearance [reflectance/albedo] parameters from the same active appearance model [Chang & Tsao, 2017]). The problem with interpreting single-neuron responses is evident when we consider what a neurophysiologist would conclude by recording from top-layer units in the network we analyzed.

First, most of these units would appear to be "identity-tuned," preferring some identities (high activation) over others (low activation). However, our data show that each unit exhibits substantial identity-separation capacity (cf. effect sizes). Effect sizes take into account the full range of responses, instead of making only a "high" versus "low" response comparison. The neural-tuning analogy obscures the possibility that individual units can contribute to identity coding with a relatively low-magnitude response. This response, in the context of the responses of other units, *is* information in a distributed code. A neurophysiologist would find "identity-tuned units" here (in what is, essentially, a distributed code), only because identity modulates the individual unit responses so saliently. *These are not identity-tuned units; they are identity-separator units*. Moreover, what separates identity in these units is not likely to be semantically interpretable. This is due to the uncertain relationship between meaningful directions in the representational space and the arbitrary directions of the unit axes.

Second, no units would appear to be tuned to gender or viewpoint, because these attributes modulate the response of a unit only weakly in comparison to identity. Thus, it might appear to a neurophysiologist that viewpoint has been filtered out of the code. Instead, both the gender and viewpoint codes would be hidden from the neurophysiologist, despite the fact that the ensemble of units contains enough information for accurate classification of these attributes. From a neural tuning perspective, the undetected modulation of unit responses by viewpoint would imply that the units signal identity in a viewpoint-invariant way. This is correct, but it is a misleading characterization of the capacity of the units to code viewpoint.

Neurophysiological investigations of visual codes rely on neural-tuning data from single units in conjunction with population decoding methods. However, if multiple, commingled, distributed codes, such as those generated by the DCNN, exist in primate visual cortex, over emphasis on neural-tuning functions in high-level areas may be counter productive (Hasson et al., 2020; O'Toole & Castillo, 2021). Rather than characterizing neural units by the features or stimuli to which they respond (Bashivan et al., 2019), we should instead consider units as organizational axes in a representational space (Szegedy et al., 2013). The importance of a unit lies in its utility for separating items within a class, not in the interpretability of attributes that drive the unit to a high level of activation.

Incorporating these ideas into theories of neural encoding requires a shift in perspective from principles that have guided the analysis of neural data for decades. These principles have perhaps biased the interpretation of findings to preclude a number of plausible code types. We suggest the reconsideration of several assumptions that underlie these interpretations. First, neurons firing at any rate (not just at a high rate) can be fundamental to a representation—consistent with basic information theory. Second, any given neuron can code multiple stimulus attributes, via the potential for overlay of multiple variably distributed codes. Third, individual units in a distributed code can potentially *all* have high predictive power from quasi-independent (different) sources of information, as was the case for identity here. Good predictability of a stimulus attribute from small numbers of units does not preclude the existence of many more units with similar predictive capacity that may draw on diverse, non redundant sources of information. By analogy, it is possible to categorize an animal as a dog from seeing either its tail or its snout. Predictive power of a given unit, which is sometimes equated with selectivity, gets us only partway to understanding how a neural code works.

More generally, it is important to ask whether DCNN models operate in ways analogous to the primate visual system and human behavior. We consider this question from several perspectives. First, as noted, image-based information is retained in macaque IT cortex (Hong et al., 2016) and in top-layer DCNN representations (cf. present work; Hill et al., 2019; Hong et al., 2016; Parde et al., 2017). This unexpected property of the DCNN code—found also in high-level neural representations of objects—counters the assumption that the primate visual system must eliminate image-based information to recognize faces/objects across image variation. Instead, deep networks can accommodate diverse information about faces (identity, gender, viewpoint) in a unified code (DiCarlo & Cox 2007; O'Toole et al., 2018; O'Toole & Castillo, 2021).

A second point of comparison is provided by "brain-score" metrics that evaluate brain-machine similarity using a composite of neural and behavioral benchmarks from primate/human data (Schrimpf et al., 2018). The network architecture tested here is based on the ResNet-101 architecture, which scored in the top 3 of the 25 networks tested for brain similarity (Schrimpf et al., 2018). Notably, however, our results also replicated in a completely different DCNN architecture (see Supplemental Section S1). Other studies of face representations in DCNNs have likewise replicated across quite different architectures (e.g., Hill et al., 2019; Parde et al., 2017). Neurocomputational theory based on *direct fit models* posits that "computational capacity," rather than "network architecture," is the primary mechanism underlying the ability of DCNNs to generalize across viewpoint (Hasson et al., 2020). The replicability of face representation results across high-capacity DCNNs is predicted by direct fit theory.

Third, the present results are largely consistent with neurophysiological findings in face patches but point to a different interpretation of these data. Single-unit recordings in primate cortex show that invariance to viewpoint increases from face patches earlier in visual processing to those later in the processing stream. Freiwald & Tsao (2010) posited a progressive computation of an identity code via neural pooling, whereby the output of neurons in early face patches, which respond in a view-dependent way, is pooled in later face patches to create a fully view-invariant representation of identity. Consistent with Freiwald and Tsao (2010), DCNN units at the top of the processing hierarchy are far more strongly modulated by identity than by viewpoint. However, the DCNN units also code viewpoint but in a distributed way. In ascribing coding capacity only to high neural firing rates, a distributed coding of viewpoint across units would not be detected. Consequently, these neurophysiological results are consistent with DCNN representations.

Fourth, neurophysiological evidence supporting the claim of a progressive computation of view-invariant identity across processing layers is also consistent with data from deep networks. Abudarham and Yovel (2020), for example, traced the similarity of representations for images of faces across DCNN layers. Earlier layers in the network showed view specificity, whereas higher layers showed view invariance. However, consistent with the presence of image-based information in the face representation at the top of the network, similarity scores across head view decreased monotonically with increasing view disparity, indicating that view information was retained in the ensemble code. Concomitantly, Dhar et al. (2020) measured the *expressivity* of attributes (identity, age, sex, and yaw) across layers of a DCNN trained for face identification. Expressivity was defined as the degree to which a feature vector, in any network layer, specified an attribute. They found an increase in the expressivity of identity from the final pooling layer to the last fully connected layer. Thus, the finding that invariance to viewpoint increases from face patches earlier in visual processing to those later in the processing stream (Freiwald & Tsao, 2010) is consistent with the emergence of invariance across the layers of the DCNN—though again the interpretation of the underlying code differs.

Fifth, reconciling deep network codes with neurophysiological work showing that face images viewed by a macaque can be reconstructed from a linear combination of the activity from 205 face cells in face patches ML/MF and AM (Chang & Tsao 2017) is less certain. Comparing the present results to those of Chang and Tsao (2017) is difficult, because their goal was to reconstruct the image, rather than to identify the face from variable images. These are different tasks, and so additional data on both the computational and neural side may be needed to relate these findings.

Finally, from a human psychology perspective, a common criticism of the ecological validity of deep networks used for face recognition is that they require large numbers of identities (thousands) and images (millions) to converge. A closer look at the layered training used in deep networks, however, puts this large-scale training in perspective. Deep networks are trained with the goal of producing a system that can map any arbitrary face image onto a face-identity representation that can be used for identification. Once this large-scale training is complete, the identity nodes used in training are simply removed from the network. An additional training step is needed to learn a specific set of special (i.e., familiar) faces. One common method is to implement a simple (one-layer) linear network to map DCNN-generated face-image representations onto a new set of identity nodes (for a review of these methods, see O'Toole & Castillo, 2021). Thus, large-scale training is best considered general perceptual learning of faces, whereas this second step is best considered *familiar face learning* (O'Toole et al., 2018; O'Toole & Castillo, 2021). This latter maps well with psychological data showing that the average person is familiar with approximately 5,000 faces (Jenkins et al., 2018).

Open questions remain about how to optimize models of the visual processing of faces. In the present study, and in others (Colón et al., 2021; Dhar et al., 2020; Hill et al., 2019; Parde et al., 2017, 2019), face representations were built from deep learning algorithms trained explicitly to separate identities. In other work (Yildirim et al., 2020), networks have been trained to simultaneously decode an array of face parameters from images (i.e., shape and texture coefficients from a three-dimensional computer graphics model of faces, along with illumination direction and pose angle). In the identity-optimized DCNNs we tested, identity was strongly expressed in all units, whereas gender and viewpoint were distributed across multiple units. By optimizing for identity separation—the most finely resolved information in a face—the deep network's face representation naturally encompassed subject information (gender) and image parameters (pose and illumination) with no explicit need to optimize for these other attributes. A network trained to distinguish another type of image/subject attribute (e.g., gender) would likely encode that attribute in much the same way our network encodes identity. However, that network would be unlikely to capture the subtle visual cues that can differentiate individual people.

The application of high-performing deep networks to complex problems of visual recognition now provides a test bed for exploring types of codes that are not easily explored with single-unit or even multi-unit neural recording methods. These networks allow us simultaneous access to the units and representational spaces that support human levels of face recognition performance (Phillips et al., 2018). There may be much to learn from these networks about the possibilities of effective, neural-like computation.

Keywords: neural encoding, machine learning, perception

Acknowledgments

Funding provided by National Eye Institute Grant R01EY029692-01 to AOT and by the Intelligence Advanced Research Projects Activity (IARPA).

This research is based in part on 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 and 2019-022600002. 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.

Commercial relationships: none.

Corresponding author. Connor John Parde.
Email: connor.parde@utdallas.edu.
Address: 800 W. Campbell Rd, GR 41 Richardson, TX 75080, USA.

Footnote

[1] This procedure is common practice in the computational literature.

References

Abudarham, N., & Yovel, G. (2020). Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition. *bioRxiv*.

Bansal, A., Castillo, C., Ranjan, R., & Chellappa, R. (2017). The do's and don'ts for cnn-based face verification. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 2545–2554).

Bansal, A., Nanduri, A., Castillo, C. D., Ranjan, R., & Chellappa, R. (2017). Umdfaces: An annotated face dataset for training deep networks. In *IEEE International Joint Conference on Biometrics (IJCB)* (pp. 464–473). IEEE.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science, 364*(6439), 1–13.

Casper, S., Boix, X., D'Amario, V., Guo, L., Schrimpf, M., Vinken, K., & Krieman, G. (2019). Frivolous units: Wider networks are not really that wide. *arXiv preprint arXiv:1912.04783*.

Chang, L., & Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell, 169*(6), 1013–1028.e14.

Chen, J. C., Patel, V. M., & Chellappa, R. (2016). Unconstrained face verification using deep cnn features. In *IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1–9). IEEE.

Coln, Y. I., Castillo, C. D., & O'Toole, A. J. (2021). Facial expression is retained in deep networks trained for face identification. *Journal of Vision, 21*(4), 4, https://doi.org/10.1167/jov.21.4.4.

Dhar, P., Bansal, A., Castillo, C. D., Gleason, J., Phillips, P. J., & Chellappa, R. (2020). How are attributes expressed in face DCNNs? In *15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)* (pp. 85–92). IEEE.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333–341.

Eichenbaum, H. (2018). Barlow versus Hebb: When is it time to abandon the notion of feature detectors and adopt the cell assembly as the unit of cognition? *Neuroscience Letters, 680*, 88–93.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal, 1341*(3), 1.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science, 330*(6005), 845–851.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks, 1*(2), 119–130.

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision* (pp. 87–102). Springer, Cham.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron, 105*(3), 416–434.

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

Hill, M. Q., Parde, C. J., Castillo, C. D., Colon, Y. I., Ranjan, R., Chen, J.-C., . . . O'Toole, A. J. (2019). Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence, 1*(11), 522–529.

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience, 19*(4), 613.

Issa, E. B., & DiCarlo, J. J. (2012). Precedence of the eye region in neural processing of faces. *Journal of Neuroscience, 32*(47), 16666–16682.

Jenkins, R., Dowsett, A., & Burton, A. (2018). How many faces do people know? *Proceedings of the Royal Society B, 285*(1888), 20181319.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105). Lake Tahoe, CA: NIPS.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE, 47*(11), 1940–1951.

Maze, B., Adams, J. C., Duncan, J. A., Kalka, N. D., Miller, T., Otto, C., . . . Grother, P. (2018). IARPA Janus benchmark - c: Face dataset and protocol. *2018 International Conference on Biometrics (ICB)*, 158–165.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*(23), 3311–3325.

O'Toole, A. J., & Castillo, C. D. (2021). Face recognition by humans and machines: Three fundamental advances from deep networks. *Annual Reviews of Vision Science*.

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

Parde, C. J., Castillo, C., Hill, M. Q., Colon, Y. I., Sankaranarayanan, S., Chen, J. C., & O'Toole, A. J. (2017). Face and image representation in deep cnn features. In *12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (pp. 673–680). IEEE.

Parde, C. J., Hu, Y., Castillo, C., Sankaranarayanan, S., & O'Toole, A. J. (2019). Social trait information in deep convolutional neural networks trained for face identification. *Cognitive Science, 43*(6), e12729.

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

Poggio, T., Banburski, A., & Liao, Q. (2020). Theoretical issues in deep networks. *Proceedings of the National Academy of Sciences, 117*(48), 30039–30045.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell, 177*(4), 999–1009.

Qin, Z., Yu, F., Liu, C., & Chen, X. (2018). How convolutional neural network see the world: A survey of convolutional neural network visualization methods. *arXiv preprint arXiv:1804.11191*.

Ranjan, R., Bansal, A., Xu, H., Sankaranarayanan, S., Chen, J. C., Castillo, C. D., & Chellappa, R. (2018). Crystal loss and quality pooling for unconstrained face verification and recognition. *arXiv preprint arXiv:1804.01159*.

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

Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology, 55*, 103–111.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., . . . DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1891–1898).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Taigman, Y., Yang, M., Ranzato, M., & 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).

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience, 3*(1), 71–86.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology Section A, 43*(2), 161–204.

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision* (pp. 499–515).

Xie, S., Girshick, R., Doll´ar, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).

Yamins, D. L., 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.

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances, 6*(10), eaax5979.

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience, 16*(8), 487–497.

Zeiler, M. D., Taylor, G. W., & Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In *2011 international conference on computer vision* (pp. 2018–2025).