Modeling Biological Face Recognition with Deep Convolutional Neural Networks

Leonard Elia van Dyck and Walter Roland Gruber

Abstract

■ Deep convolutional neural networks (DCNNs) have become the state-of-the-art computational models of biological object recognition. Their remarkable success has helped vision science break new ground, and recent efforts have started to transfer this achievement to research on biological face recognition. In this regard, face detection can be investigated by comparing face-selective biological neurons and brain areas to artificial neurons and model layers. Similarly, face identification can be examined by comparing in vivo and in silico multidimensional "face spaces." In this review, we summarize the first studies that use DCNNs to model biological face recognition. On the basis of a broad spectrum of behavioral and computational evidence, we conclude that DCNNs are useful models that closely resemble the general hierarchical organization of face recognition in the ventral visual pathway and the core face network. In two exemplary spotlights, we emphasize the unique scientific contributions of these models. First, studies on face detection in DCNNs indicate that elementary face selectivity emerges automatically through feedforward processing even in the absence of visual experience. Second, studies on face identification in DCNNs suggest that identity-specific experience and generative mechanisms facilitate this particular challenge. Taken together, as this novel modeling approach enables close control of predisposition (i.e., architecture) and experience (i.e., training data), it may be suited to inform long-standing debates on the substrates of biological face recognition. ■

INTRODUCTION

Face Recognition in the Brain

For decades, neuroscientists and psychologists have been intrigued by the behavioral and computational processes underlying biological face recognition. As especially faces of humans and nonhuman primates carry important social information (Freiwald, Duchaine, & Yovel, 2016), their perception could have developed as a highly specialized and effective form of object recognition (Kanwisher, 2000). Faces are processed in terms of physical and functional properties (Bruce & Young, 1986). On one side, face detection enables spotting interclass differences between faces and other objects and is often studied in the context of neuronal face selectivity (Tsao & Livingstone, 2008), that is, the functional specialization or "preference" of neurons for faces compared with a wide spectrum of other objects. On the other side, face identification relates to intraclass differences between individual face examples and involves numerous dimensional features that span a representational "face space" encoding the similarity of different identities, that is, their position in this multidimensional space in relation to examples of other identities (Valentine, Lewis, & Hills, 2016). In this review, we illustrate how the latest studies foreshadow that state-of-the-art computational encoding models, namely, deep convolutional neural networks (hereafter DCNNs), could be used to decipher biological face recognition.

A remarkable collection of neuroscientific studies has long identified face-selective neurons (Tsao, Freiwald, Tootell, & Livingstone, 2006; Perrett, Hietanen, Oram, & Benson, 1992; Desimone, 1991) and brain areas (Kanwisher & Yovel, 2006; Kanwisher, McDermott, & Chun, 1997; Perrett, Mistlin, & Chitty, 1987) as the neural basis for an assumed specialized treatment of face configurations (Freiwald, 2020; Hesse & Tsao, 2020; Freiwald & Tsao, 2010). Studies using fMRI have discovered that biological face recognition involves distributed networks of face-selective areas. This core face network comprises the fusiform face area, occipital face area, and posterior superior temporal sulcus (Haxby, Hoffman, & Gobbini, 2000; Kanwisher et al., 1997), and is supported by additional regions. The question of whether face selectivity arises through predisposition from innate mechanisms in the form of an automatic face bias (Johnson & Mareschal, 2001; Johnson, Dziurawiec, Ellis, & Morton, 1991), is acquired through experience leading to face expertise (Young & Burton, 2018; Gauthier & Bukach, 2007; Gauthier & Nelson, 2001), or an interaction of both (Srihasam, Vincent, & Livingstone, 2014; Srihasam, Mandeville, Morocz, Sullivan, & Livingstone, 2012), has sparked heated debates. Recent findings have steered this discussion in a completely new direction by indicating that faces may not be "special" after all, as their selectivity could arise as a byproduct of domain-general features (Vinken,

University of Salzburg, Austria

© 2023 Massachusetts Institute of Technology. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

Journal of Cognitive Neuroscience 35:10, pp. 1521–1537
https://doi.org/10.1162/jocn_a_02040

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

Konkle, & Livingstone, 2022; Bao, She, McGill, & Tsao, 2020) or optimization for a domain-specific task (Kanwisher, Gupta, & Dobs, 2023; Dobs, Yuan, Martinez, & Kanwisher, 2022; Yovel, Grosbard, & Abudarham, 2022b).

From a theoretical standpoint, the endeavor to disentangle the influences of nature and nurture on face recognition ultimately calls for more impactful and controlled experimental manipulations such as lesions, deprivations, or extensive exposure. However, from a practical standpoint, these designs are challenging and unethical to study in living beings and would not allow the examination of the brain's entire face processing system at once. Consequently, recent models from the field of deep learning could provide a suitable alternative by simulating such experiments in silico.

DCNNs as Models of Biological Object Recognition

DCNNs (e.g., LeCun, Bengio, & Hinton, 2015; Krizhevsky, Sutskever, & Hinton, 2012) are a subset of artificial neural networks and commonly used in the field of computer vision to solve a wide range of image classification tasks. The seminal discovery of simple and complex cells in the visual cortex (Hubel & Wiesel, 1959, 1962) once inspired these architectures, which stack convolutional and pooling operations to perform a series of linear and nonlinear transformations of input data through multiple layers of learnable filters. Since their invention, DCNNs have been rapidly adopted in the field of visual neuroscience. Today, because of their strong architectural and functional similarities to the ventral visual pathway, particularly in humans and nonhuman primates, DCNNs are established as state-of-the-art computational models of biological object recognition (Greene & Hansen, 2018; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Kheradpisheh, Ghodrati, Ganjtabesh, & Masquelier, 2016; Güçlü & van Gerven, 2015; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Yamins, Hong, Cadieu, & DiCarlo, 2013).

However, the general enthusiasm for these promising computational models has also raised critical voices from opponents of this research agenda. A part of this criticism relates directly to DCNNs and primarily addresses the shortcomings of the models, such as their need for large amounts of labeled data, their lack of generalization, their assumption of a relatively stable world, their lack of transparency as a "black box," and their reliance on fundamentally different processing mechanisms (e.g., Leek, Leonardis, & Heinke, 2022; Marcus, 2018). Another part of this criticism relates to the current research program that praises DCNNs as the best models of human object recognition. In a recent critical article, Bowers and colleagues (2022) point out various principled and practical problems of this research program and argue that it relies heavily on "prediction-based experiments," which focus on blindly explaining the highest possible variance rather than examining the actual underlying effects. They reason that, in contrast to "controlled experiments," this approach introduces various questionable practices in the field, such as biases in the selection of architectures, data sets, and parameters, as well as neglect of alternative explanations and hypotheses. In essence, they contend that DCNNs "share little overlap with biological vision" and "account for almost no results from psychological research" (Bowers et al., 2022).

Conversely, proponents of this research agenda argue that DCNNs are sophisticated models of biological vision that can be used to reveal novel and otherwise unobtainable insights. In a recent definition of this endeavor, Doerig and colleagues (2022) advocate that studies using artificial neural networks are often much like "controlled experiments," because they allow for close control over both the effects of architecture and experience, while even avoiding some of the limitations and biases inherent in classical paradigms. Following this chain of reasoning, DCNNs can be used to effectively investigate multiple levels, including behavior, computation, single units to complex dynamics, natural sensory information, naturalistic environments, and developmental processes. According to this view, dissimilarities between brains and artificial neural networks should not immediately refute the use of the latter, but rather point to intriguing mechanisms and causalities that require further exploration. Therefore, artificial neural networks provide a balance between the computational abstraction required and the biological detail needed to test complex hypotheses about the brain. Importantly, DCNNs can also be used to evaluate causal hypotheses that go far beyond "prediction-based experiments" by controlling factors that influence optimization for a particular task, thus isolating the minimum sufficient requirements for a specific phenomenon to occur (Kanwisher, Khosla, & Dobs, 2023). As their architecture, initial training, subsequent fine-tuning, and artificial neurons can be closely controlled and directly manipulated, DCNNs may provide a useful addition to the field of face perception research as encoding models of face recognition (reviewed here; see Figure 1) and decoding models for face reconstruction from neural data (VanRullen & Reddy, 2019). On the basis of this account, it seems highly reasonable to replace the brain as a targeted "black box" with a more accessible and controllable DCNN as a modeled "transparent box" that is still sufficiently complex but certainly easier to study.

This perspective provides a suitable starting point to explore the literature on modeling biological face recognition with DCNNs. With this in mind, our review is organized as follows: To begin with, we will summarize the behavioral and computational evidence supporting the idea that DCNNs are also useful models for biological face recognition, which involves both detection and identification. Then, we will outline two topical examples of how this modeling approach is being used to gain new insights. As we will see in a first spotlight, recent studies indicate that elementary face selectivity may arise automatically

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

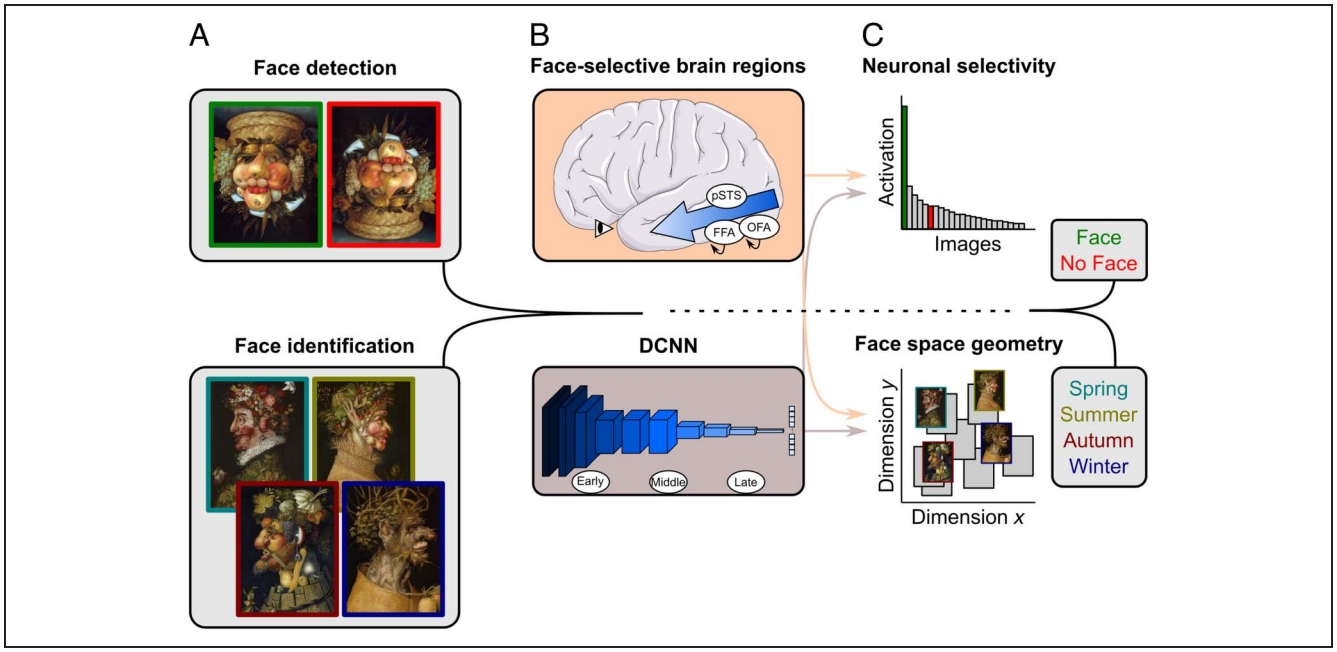

Figure 1. Face recognition in the brain and DCNNs. (A) Face detection and identification require recognition of interclass differences (i.e., faces and other objects) and intraclass differences (i.e., faces of various identities), respectively. (B) The brain contains many face-selective biological neurons and regions such as fusiform face area, occipital face area, and posterior superior temporal sulcus. DCNNs are able to develop similar face-selective artificial neurons throughout their early, middle, and late layers. (C) Face detection is often studied in the context of neuronal selectivity, which quantifies how much a given neuron is activated by faces compared with other objects. Face identification is usually investigated in the context of a multidimensional "face space," which encodes the similarity of identities along various dimensional features. Both investigations can be performed for biological and artificial neurons alike. Source: Paintings by Giuseppe Arcimboldo are for demonstration purposes only and publicly available at https://commons.wikimedia.org. Parts of the figure were adapted from SMART Servier Medical Art, licensed under a Creative Commons Attribution 3.0 unported license.

through feedforward hierarchies and even without visual experience. As we will see in a second spotlight, recent evidence also suggests that face identification is enhanced by identity-specific experience and additional generative mechanisms. Finally, we will revisit the results of the initial studies and discuss open questions and future directions of this rapidly evolving field.

DCNNs as Models of Biological Face Recognition

Are DCNNs Actually Useful Models to Study Biological Face Recognition?

To answer this question, we will evaluate the current evidence to determine whether DCNNs provide an adequate explanation for the different stages of this ability and thus constitute a suitable neuroscientific model (Doerig et al., 2022). We will focus primarily on the behavioral and computational levels (including findings on single units and complex dynamics), as there exists a considerable amount of literature in these areas. In addition, it should be noted that most studies aim to examine natural sensory information (e.g., neural data) recorded in naturalistic environments (e.g., naturalistic face images). For consistency in terminology, we will refer to DCNNs trained only on face images for face identification as "face-id models," trained on face and object images for face detection as "face-de models," trained only on nonface object images for object categorization as "object-cat models," and untrained randomly initialized as "untrained models."

Behavioral Level

At the behavioral level, comparisons between humans and DCNNs usually involve the evaluation of categorization accuracies, error patterns, similarity ratings, and their underlying representations in response to face stimuli under naturalistic and/or manipulated conditions.

Do Humans and DCNNs Perform Similarly in Face Recognition?

DCNNs have long matched human performance in face detection and identification (e.g., Farfade, Saberian, & Li, 2015; Sun, Wang, & Tang, 2014; Taigman, Yang, Ranzato, & Wolf, 2014). This success was facilitated by advances in deep learning and the compilation of large data sets of naturalistic face images (e.g., Cao, Li, Brandmeir, & Wang, 2021; Kemelmacher-Shlizerman, Seitz, Miller, & Brossard, 2016; Huang, Mattar, Berg, & Learned-Miller, 2008). For the particularly challenging task of face identification, it has been shown that face-id models perform equally well as forensic facial examiners (Phillips et al., 2018). A recent study by Dobs, Yuan, and colleagues (2022) indicated that face-id models, rather than face-de or object-cat models,

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

achieve human-level performance in an identity matching task. Notably, face-id models exhibited the most comparable representation of face stimuli to that of human face identification, as measured in behavioral multiarrangement and similarity-matching tasks. To investigate whether humans and DCNNs achieve this similar performance by similar means, Abudarham, Grosbard, and Yovel (2021) examined the role of critical features (e.g., eye shape, eyebrow thickness, and lip thickness) that have previously been isolated as critical features for humans as they can alter the perceived facial identity when manipulated (Abudarham & Yovel, 2016). Their results showed that face-id, but not object-cat models, use the same critical features as humans for face identification. In face-id models, the sensitivity to these features, as determined by the similarity between an original face example and versions with manipulated critical or noncritical features, correlated positively with behavioral face identification performance and computational identity selectivity toward late layers.

Therefore, it can be noted that DCNNs display humanlike performance in face recognition tasks, and there are reasons to believe that they process similar facial features to achieve this performance.

Do Humans and DCNNs Display Similar "Psychological Phenomena" of Face Recognition?

Humans exhibit several well-documented "psychological phenomena" of face recognition, such as the "face inversion effect" (e.g., Yin, 1969), the "Thatcher effect" (e.g., Thompson, 1980), face familiarity effects (e.g., Johnston & Edmonds, 2009), and the "other-race effect" (e.g., Bothwell, Brigham, & Malpass, 1989), which can be also tested in DCNNs. In this context, biological face recognition is often conceptualized as holistic template matching instead of purely atomistic feature processing (Freiwald et al., 2016; Tanaka & Farah, 1993). The "face inversion effect" leads to disproportionally reduced recognition performance and neural activity for inverted faces compared with inverted nonface objects (Yovel & Kanwisher, 2005; Rossion & Gauthier, 2002; Kanwisher, Tong, & Nakayama, 1998). This effect was already reported in simpler computational models (Hosoya & Hyvärinen, 2017; Farzmahdi, Rajaei, Ghodrati, Ebrahimpour, & Khaligh-Razavi, 2016) and more recently in various DCNNs (Dobs, Yuan, et al., 2022; Tian, Xie, Song, Hu, & Liu, 2022; Vinken et al., 2022; Yovel, Grosbard, & Abudarham, 2022a; Zeman, Leers, & de Beeck, 2022; Xu, Zhang, Zhen, & Liu, 2021). While recent evidence suggests that the "face inversion effect" manifests only in face-id models (Dobs, Yuan, et al., 2022), it has also been observed to a limited extent in face-de models (Tian et al., 2022; Xu et al., 2021), and in object-cat models that contained occasional faces (Vinken et al., 2022). The finding that DCNNs trained to simultaneously identify faces (i.e., different people) and nonface objects (i.e., different cups) misclassify and represent inverted faces as objects (Tian et al., 2022) could be also consistent with the hypothesis that upright and inverted faces are primarily processed by specialized face and object recognition systems in the brain (Yovel & Kanwisher, 2005; Haxby et al., 1999). Interestingly, inversion effects of nonface objects were reported to reach the level of faces, when the models are optimized for the corresponding identification task (Dobs, Yuan, et al., 2022; Yovel et al., 2022a). These findings fit well with previously reported nonface inversion effects in human experts for categories such as dogs and birds (Campbell & Tanaka, 2018; Diamond & Carey, 1986). This implies that such inversion effects may be primarily driven by domain-specific mechanisms originating from the optimization for fine-grained expert tasks, which challenges the notion of the unique nature of facial stimuli. Strikingly, in DCNNs, these perceptual characteristics were even observed at the level of local facial features with the "Thatcher effect," in which inverted eyes and mouth are perceived as obvious distortions in upright but blend in well in inverted presentations. In this context, face-id but not object-cat models showed an increased representational dissimilarity between normal and Thatcherized versions, which was higher for upright than for inverted presentations and generally increased toward late layers (Jacob, Pramod, Katti, & Arun, 2021).

Thus, it can be argued that DCNNs, and face-id models in particular, exhibit indicators of established "psychological phenomena" of human face recognition.

Does Familiarity Affect Face Identification in Humans and DCNN Similarly?

Familiarity effects demonstrate that face identification is generally easier for humans in the context of familiar compared with unfamiliar identities (Young & Burton, 2018). As humans were reported to use the same critical features mentioned earlier for the identification of familiar and unfamiliar faces (Abudarham, Shkiller, & Yovel, 2019), this prompts merely conceptual rather than perceptual differences. In an approach to simulate familiarity in DCNNs, Blauch, Behrmann, and Plaut (2021) tested face-id models before and after fine-tuning on new face identities. In face identification tasks on natural images, it was observed that familiarized face-id models surpass a simpler model, which combines principal component analysis (PCA) and linear discriminant analysis to capture bottom–up and top–down information (Kramer, Young, & Burton, 2018). Young and Burton (2021) argued that the relatively simple PCA + linear discriminant analysis model already provided valuable insights and was only exceeded by the superscale performance of DCNNs. However, returning to the desired balance between computational abstraction and biological detail in neuroscientific models (Doerig et al., 2022), it is crucial to gather evidence for these effects also in more complex models, as they allow for a wider range of phenomena, questions, and explanations that cannot be

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

addressed by simpler models (see Spotlights 1 and 2). The results by Blauch and colleagues (2021) highlighted that face experience in DCNNs increases their performance on unfamiliar faces but also their ability to learn novel identities. This supports the idea that face examples possess identity-general variability, which can be learned from other face configurations, but also identity-specific variability, which can only be learned through familiarization with the respective identity (see Spotlight 2). Consequently, face experience may enable rapid learning of identity-specific information and thereby enhance face identification for familiar faces. Another generalized form of familiarity effects in human face identification is the "other-race effect," which is characterized by a higher face identification performance for faces of familiar compared with unfamiliar ethnic groups. This effect was found in face-id models, based on the ethnicity of faces in the training sample (Tian, Xie, Hu, & Liu, 2021), but not in face-de or object-cat models (Dobs, Yuan, et al., 2022).

Therefore, it seems plausible that familiarity influences the ability of DCNNs to identify faces and may do so in a similar way as it does in humans. This broad spectrum of evidence suggests that DCNNs, and face-id models in particular, exhibit behavior similar to human face identification and, contrary to previous criticism (Bowers et al., 2022), may even account for various "psychological phenomena," such as the "face inversion effect," the "Thatcher effect," face familiarity effects, and the "other-race effect," under appropriate conditions.

Computational Level

At the computational level, comparisons between humans and DCNNs usually entail the investigation of encoded representations at the level of biological and artificial neurons, brain areas and model layers, and their processing mechanisms along the visual hierarchy.

Do Brains and DCNNs Encode and Process Information for Face Recognition Similarly?

For the task of face detection, Ratan Murty, Bashivan, Abate, DiCarlo, and Kanwisher (2021) demonstrated that face-de models outperform descriptive models and human experts in predicting fMRI activity of face-, body-, and place-selective visual brain areas to novel images. Moreover, meaningfully trained models, deeper architectures with more layers or recurrent connections, and broadly sampled training data with different object categories all increased the fit to neural data, contrary to untrained models, shallower architectures with fewer layers, and narrowly sampled training data with different face identities only. Adequate DCNN predictions were observed for single voxels, generalized well across participants, and were confirmed in an additional high-throughput screening approach. The finding that face-selective artificial neurons are able to predict their biological counterparts supports the hypothesis that they may also use similar underlying features for face recognition. Nonetheless, the increased predictability of object-cat models trained on more diverse data sets reveals that, in the context of face detection, these mechanisms appear to depend to some extent on domain-general visual properties, as observed in related studies (Vinken et al., 2022; Bao et al., 2020).

For the task of face identification, deeper insights into the functional similarity of biological and artificial computations were given by Wang, Cao, Brandmeir, Li, and Wang (2022), who revealed that just like the brain (Cao et al., 2021; De Falco, Ison, Fried, & Quian Quiroga, 2016), face-id models develop single- and multiple-identity-selective neurons that generalize well to highly abstract examples. Similar to the findings on face selectivity (Ratan Murty et al., 2021), identity-selective artificial and biological neurons were found to encode information in similar ways through region-based feature coding (Chang & Tsao, 2017), that is, earlier regions/layers encode the axes of a face space and later regions/layers are specialized for specific regions of this face space and thus identity selective. Subsequent analyses with selective lesioning of identity-selective artificial neurons revealed that artificial neurons tuned to single identities are more important for face identification than those tuned to multiple identities, which could also apply to their biological counterparts. Of course, unlike face-id models, the brain cannot be optimized for face identification alone, but must naturally solve other fine-grained expert tasks as well. This way, it could use the same domain-general or different domain-specific mechanisms for different expert tasks such as face and car identification. To test this with DCNNs, Kanwisher, Gupta, and colleagues (2023) trained a fully shared dual-task DCNN optimized for simultaneous face identification and object categorization, which has been shown to spontaneously develop functionally specialized face- and object-selective filters (Dobs, Martinez, Kell, & Kanwisher, 2022). Lesioning the top 20% of object-selective filters led to a more substantial deterioration in car identification performance than lesioning the same percentage of face-selective filters. Moreover, only a small portion of the most important car-selective filters overlapped with face-selective filters, whereas a more substantial portion of them overlapped with object-selective filters. The majority of car-selective artificial neurons were recycled from other previously nonselective artificial neurons. This provides computational evidence that fine-grained expertise is predominantly based on domain-specific mechanisms or task-specific optimization, rather than domain-general mechanisms. In addition, this implies that largely independently operating specialized processing systems could arise automatically in visual hierarchies that are optimized for different expert tasks (see Spotlight 1).

Accordingly, it can be argued that brains and DCNNs share similarities in the way they encode and process information for face recognition. Furthermore, these results provide preliminary evidence that face recognition may

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

require predominantly domain-general mechanisms, whereas face identification may require predominantly domain- or task-specific mechanisms.

Do Brains and DCNNs Encode Features in a Similar Hierarchical Fashion?

The previously reported critical features (Abudarham et al., 2019, 2021) may incorporate several dimensions of face stimuli. In accordance with this idea, face-id models were reported to generalize identity representations well to novel encounters of altered intrinsic, face-related features (e.g., emotional expression) but not so much for those of changed extrinsic, face-unrelated features (e.g., head position and illumination; Xu, Garrod, Scholte, Ince, & Schyns, 2018). In addition, other studies indicated that face-id models predict human face identification best if they use similar 3-D shape features of face scans (Daube et al., 2021). A considerable number of studies underline the following hierarchical coding of facial features in DCNNs, which roughly corresponds to that observed in human and nonhuman primate brains (e.g., Freiwald, 2020; Freiwald & Tsao, 2010): (1) Similar to object recognition, early layers encode low-level features such as edges, blobs, orientation, and color (LeCun et al., 2015) that may be closely aligned with domain-general features. (2) Middle layers encode view-specific features such as head orientation and illumination (Raman & Hosoya, 2020; Grossman et al., 2019). (3) Late layers encode patterns of view-invariant features that are related to gender and identity (Jozwik et al., 2022; Wang et al., 2022; Abudarham et al., 2021; Parde et al., 2021; Raman & Hosoya, 2020). (4) The final layer encodes a nonlinear identity benefit for familiar faces (Blauch et al., 2021), similar to that observed in nonhuman primates in later processing areas outside the core face network (Landi, Viswanathan, Serene, & Freiwald, 2021; Landi & Freiwald, 2017).

However, there are also partly contradictory results. Raman and Hosoya (2020) compared face-selective areas in nonhuman primates and face-id model layers regarding various tuning properties (i.e., view-invariance, shape-appearance, facial geometry, and contrast polarity). In middle areas, no single layer was able to predict all combined tuning properties and only view-specific tuning was similar. Surprisingly, face-id, object-cat, and untrained models all indicated increasing view-invariant identity selectivity toward late layers. In contrast, Grossman and colleagues (2019) reported no significant correlations for late but only for middle layers. Here, artificial neurons encoded especially view-specific features. View-invariance and identity selectivity emerged toward late layers in both face-id and object-cat but not untrained models and were not organized in a brain-like face space. However, given a careful interpretation, these two specific studies could also represent two sides of the same coin by highlighting the view-specific, pictorial tuning in middle (Grossman et al., 2019) and view-invariant identity-tuning in late regions and layers (Raman & Hosoya, 2020), although their disagreement could be explained by the comparison of different species (i.e., humans and macaques) and/or models (i.e., VGGFace and AlexNet). Other studies also reported significant but very small correlations between face-selective brain areas and face-id models based on their representational similarity (Tsantani et al., 2021).

On the basis of the current literature, it can be stated that DCNNs should be recognized as sophisticated computational models of biological face recognition that, depending on the task, encode and process information in a hierarchical manner that largely resembles the neural face recognition system of humans and nonhuman primates. However, as biological face recognition is far more complex than simple object labeling and involves objectives beyond physical properties, such as those on social content (Sutherland et al., 2013; Oosterhof & Todorov, 2008), it is likely that image-computable models still fail to capture several functional properties (e.g., Jiahui et al., 2022; Tsantani et al., 2021) and yet represent the best current approximation of this ability.

SPOTLIGHT 1: FACE SELECTIVITY EMERGES THROUGH FEEDFORWARD CASCADES

Emerging Selectivity

Debates on the origin of neural face selectivity hold that specialized face recognition may arise from nature (i.e., innate mechanisms), nurture (i.e., acquired experience), or the interaction of both. Studies with congenitally blind individuals indicated that face selectivity in occipitotemporal cortex may arise even in complete absence of visual experience (Ratan Murty et al., 2020; van den Hurk, Van Baelen, & Op de Beeck, 2017). Congenitally blind individuals were found to exhibit robust, functionally specialized neural responses to sounds associated with faces, bodies, places, and objects, which could be used to discriminate neural responses to the corresponding visual stimuli in sighted individuals (van den Hurk et al., 2017). Moreover, in this group of fully visually deprived individuals, such face-selective neural responses were observed even during haptic exploration of 3-D face stimuli (Ratan Murty et al., 2020). The striking similarities of functional specialization in congenitally blind and sighted individuals suggest that a category-selective map, including face selectivity and its topographic separation, may not require any visual inputs after all.

In this context, Xu and colleagues (2021) took advantage of the fact that architecture and training data can be separately controlled in DCNNs and tested whether domain-specific face experience is necessary for face selectivity to develop. Therefore, they created a fully face-deprived object-cat model by carefully removing all images that included occasional face stimuli (e.g., in the image background) from the model's training data set.

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

Interestingly, the face-deprived model performed only slightly worse in face detection and identification compared with standard face-de and face-id models. The results indicated that face-selective artificial neurons automatically emerged but were reduced in overall selectivity and sparseness. This automatic emergence of face selectivity in face-deprived models seems to rule out the need for acquired domain-specific face experience and supports previous evidence in nonhuman primates showing that face selectivity can develop automatically without domain-specific face experience (Sugita, 2008). Consistent with these findings, Vinken and colleagues (2022) showed that, in face patches of nonhuman primates, neural tuning for nonface objects is more predictive of face presence than neural tuning for faces itself. This reinforces the aforementioned assumption that face selectivity is predominantly based on domain-general features that are highly correlated with the category of faces. While the results made it clear that an object-cat model is also able to predict neuronal face selectivity in face patches, the highest similarity was surprisingly found in middle layers. This observation indicates that the reliable predictability of nonface object tuning for face presence could depend predominantly on local part-based features encoded in face patches. This aligns with the concept that face selectivity underlying face detection may arise either from the statistical properties of feedforward hierarchies, as in the case of purely feedforward DCNNs, or from acquired domain-general object experience.

To further disentangle these two possibilities, Baek, Song, Jang, Kim, and Paik (2021) investigated untrained models and reported similar emergence of face- and object-selective artificial neurons that shared many characteristics with biological neurons of nonhuman primates. In line with the aforementioned findings by Xu and colleagues (2021), the automatically emerging face selectivity increased across layers and was higher for global compared with local face configurations, whereas face experience sharpened face-selective artificial neurons and thereby increased their sparseness. Similar to previous work indicating that untrained models are able to perform object recognition (Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009; Pinto, Doukhan, DiCarlo, & Cox, 2009), this automatically emerging face selectivity could serve as the basis for successful face detection. More precisely, this implies that elementary face selectivity could potentially manifest even without visual experience through the brain's feedforward circuitry composed of simple and complex cells. Although the exact computational mechanisms for this phenomenon are not yet well understood and require further exploration, initial studies already point to several promising but ambiguous explanations. In any case, this automatically arising face selectivity could serve as a catalyst for later fine-tuning through domain-general and domain-specific experience to enable more demanding tasks such as face identification.

In accordance with this reasoning, object-cat models were reported to provide a better foundation for subsequent learning of a fine-grained identification of cars (Kanwisher, Gupta, et al., 2023) and birds (Yovel et al., 2022b) than face-id models. The inability of face-id models to adapt well to new expert tasks of different domains was found for both superordinate (i.e., related to bird species) and individual (i.e., related to bird identity) levels (Yovel et al., 2022b). This corroborates the idea that fine-grained expertise, as in the case of face identification, may rely primarily on domain-specific mechanisms or task-specific optimization but not just domain-general mechanisms or the "specialness" of face stimuli.

Disentangling Selectivity

To investigate how functional specialization with multiple selective systems could arise, Dobs, Martinez, and colleagues (2022) trained a face-id and object-cat model each in a single-task condition and a fully shared dual-task model with interleaved batches of faces and objects. As expected, single-task models performed worse on the respective other task, whereas the dual-task version succeeded in both tasks. Interestingly, however, a functional specialization with face- and object-selective artificial neurons arose automatically in dual-task models after middle layers and increased steadily toward late layers. Through lesioning these selective artificial neurons, this unsupervised task segregation was then confirmed a double dissociation that emerged only for meaningfully separated tasks and to a smaller degree for other nonface categories. This further suggests that some tasks (e.g., faces vs. objects) segregate better than others (e.g., food vs. objects), and hence, the former may rely on more distinct feature sets compared with the latter.

Although feedforward hierarchies in DCNNs enable the emergence of category-selective artificial neurons, this selectivity is not organized in spatial clusters as in functionally specialized brain areas. This is because invariant object recognition in these models is tied to the selectivity of entire convolutional feature maps rather than specific retinotopic receptive fields. However, techniques such as self-organizing maps (SOMs; Kohonen, 1990), that is, unsupervised artificial neural networks that learn to represent high-dimensional data in a low-dimensional space, or additional loss functions, for example, simulating the wiring costs between the artificial neurons of DCNNs, can be used to visualize and investigate the resulting topographies. In such an approach, Cowell and Cottrell (2013) used SOMs to replicate object-selective topographies reported in occipitotemporal cortex using fMRI (Spiridon & Kanwisher, 2002; Haxby et al., 2001). Similar to recent evidence highlighting the importance of domain-general mechanisms in the development of face selectivity (Vinken et al., 2022; Bao et al., 2020), their results suggested that a functionally specialized topography may also arise automatically through such mechanisms. Along

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

these lines, the particularly high selectivity for faces could be because of their high within-category and low between-categories similarity. These findings are further supported by recent unsupervised approaches, such as DCNN representations fed into SOMs (Doshi & Konkle, 2023) and topographic variational autoencoders (Keller, Gao, & Welling, 2021), which produce similar topographic maps. Although domain-general mechanisms may be sufficient to generate this object-selective topography in SOMs, this observation appears to be limited to face detection only, and arguably oversimplifies this phenomenon because of the models' assumption of a limited set of visual parameters and lack of consideration of hierarchical interactions (Blauch, Behrmann, & Plaut, 2022).

To expand on this idea, Blauch and colleagues (2022) examined interactive topographic networks that incorporate biologically plausible constraints. These networks were composed of object-cat models with additional recurrent layers, optimized to balance task performance and wiring costs to account for dependencies within the topographic map. As expected, these models displayed face-, object-, and scene-selective patches with columnar responses and efficient connectivity. Intriguingly, among various tested mechanisms, recurrence, separated excitatory and inhibitory artificial neurons, as well as purely excitatory feedforward connections, produced the most pronounced selective clusters. In another study, Lee and colleagues (2020) implemented a similar topographic DCNN that was optimized for object categorization while simultaneously minimizing the wiring cost of its artificial neurons. The resulting object-selective topography strongly resembled the functionally specialized organization of primate inferior temporal cortex, as especially face- and body-selective artificial neurons were organized in adjacent or overlapping clusters, and especially face-selective clusters were strongly interconnected to form a network across model layers.

To summarize, recent studies investigating face selectivity in DCNNs strengthen a computational account for the emergence of elementary face selectivity through neither purely innate mechanisms (i.e., hardwired from the beginning) nor specific experience (i.e., acquired later on) but rather the statistical properties of feedforward hierarchies. The development of face-selective neurons could be guided by additional selection pressure to categorize objects and minimize wiring effort, leading to their characteristic topographic organization.

SPOTLIGHT 2: FACE IDENTIFICATION IS ENHANCED BY IDENTITY-SPECIFIC EXPERIENCE AND GENERATIVE MECHANISMS

Identity-specific Experience

An important property of biological face recognition is its great robustness in challenging conditions. Especially familiarity is thought to enable humans to identify faces even under the most difficult conditions. In this regard, Noyes and colleagues (2021) demonstrated that the performance of face-id models decreases similarly to that of humans in the presence of deliberate disguise (Noyes & Jenkins, 2019). In face-id models, "face averaging" (i.e., averaging the representations of multiple examples of a given identity), which was proposed as a computational mechanism for biological face identification (Kramer, Ritchie, & Burton, 2015; Burton, Jenkins, Hancock, & White, 2005), led to enhanced within-identity matching in evasion disguise (i.e., correctly identifying the same identity even when the person is disguised to look nothing like themselves), but reduced between-identities matching in impersonation disguise (i.e., correctly distinguishing different identities, although one is disguised to look like the other). Curiously, performance on both within-identity and between-identities matching was enhanced by a contrast learning approach that emphasized differences in appearance and identity between individuals. This suggests that averaging multiple perceived face examples of a given identity may lead to a more robust representation for subsequent identification under changing or more difficult conditions.

DCNNs and Generative Models

As DCNNs are gradually entering face recognition research, the question arises whether they are indeed more suitable than preceding statistical or generative models (hereafter GMs; Tolba, El-Baz, & El-Harby, 2006). As an example, Jozwik and colleagues (2022) reported that face-id models do not predict human identity judgments significantly better than the Basel Face Model (BFM; Gerig et al., 2018; Paysan, Knothe, Amberg, Romdhani, & Vetter, 2009), a well-known 3-D morphable model based on principal components of 3-D face scans. Interestingly, both DCNNs trained on either natural images or generative examples predicted human judgments equally well. Although this may not surprise as both DCNNs and the BFM were tested on face stimuli that were generated by the BFM, this could also emphasize the ecological validity of this statistical tuning in both in vivo and in silico face spaces. These findings are also encouraged by comparisons with a similar 2-D morphable model, the widely used Active Appearance Model (AAM; Cootes, Edwards, & Taylor, 2001; Edwards, Taylor, & Cootes, 1998). The active appearance model was found to explain face-selective biological neurons equally well and even better in terms of identity-unrelated, extrinsic features (e.g., illumination) compared with face-id models (Chang, Egger, Vetter, & Tsao, 2021).

It is worth noting that the previously mentioned studies also used artificially generated face stimuli for testing, eliminating a key advantage of DCNNs. As reported by O'Toole, Castillo, Parde, Hill, and Chellappa (2018), DCNNs and GMs display differently constructed face spaces. Similar to how biological face recognition can

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

result in ambiguous face representations because of factors such as familiarity or viewing conditions, DCNNs encode not only the identity but also the quality of images in their face space. Thereby, low-quality and hard-to-recognize examples are positioned in the center of their face space, whereas higher-quality and easier-to-recognize examples are positioned at the outer edges. In contrast, GMs represent the prototypical face in the center of their face space because image quality is usually controlled here. This ability of DCNNs to learn and represent a variety of complex features directly from the data is invaluable in studies aimed primarily at examining face recognition processes in naturalistic settings. Because GMs can learn the underlying data distribution and generate new samples, they are particularly appealing in scenarios where there are only a few training examples. Accordingly, the combination of DCNNs and GMs could be particularly promising.

Analysis-by-Synthesis

Novel insights regarding the underlying mechanisms of face identification were recently obtained through a combination of DCNNs and GMs. Yildirim, Belledonne, Freiwald, and Tenenbaum (2020) implemented an efficient inverse graphics model, which combines a face-id model and the BFM. Instead of mapping a face image directly onto an identity label, this efficient inverse graphics model uses DCNN and GM components to go the other way around. First, a DCNN component is used to perform preliminary segmentation and normalization of faces, as well as to extract latent 3-D variables (e.g., shape, texture, and viewing parameters), and to perform standard face identification. Then, a GM component generates a 3-D representation based on the extracted latent variables. Finally, this representation is used to create a 2.5-D surface representation, followed by a 2-D segmented face image. This way, in contrast to common face-id models, this analysis-by-synthesis network not only analyzes a 2-D image and outputs the corresponding identity label (e.g., "Albert Einstein"), but instead feeds a generative model with facial parameters to infer a 3-D identity representation (e.g., a 3-D model of Albert Einstein's face), which is then used for identification. The model thereby infers parts of the face from the image, which cannot be directly extracted. The efficient inverse graphics model demonstrated remarkable performance by achieving an exceptional level of similarity to neural data of nonhuman primates ($r = .96$) at corresponding processing stages of the model and face patches with various hallmark features of face coding (such as view specificity, mirror symmetry, and identity specificity coding), when compared with standard DCNNs ($r = .36$), and thereby seemed to close some of the previously reported gaps in capturing the full hierarchy of biological face recognition (Grossman et al., 2019).

To summarize, recent studies investigating face identification in DCNNs indicate that this ability is enabled through domain-specific face experience for unfamiliar identities, that is, clustering inter-identity examples in a face space. Face identification is further enhanced through identity-specific experience for familiar faces, that is, clustering intra-identity examples closer together. Moreover, additional generative mechanisms may enable inferences that boost this capacity even more. In biological face recognition, these generative mechanisms could be found in recurrent and feedback connections.

CONCLUSION

In this review, we have summarized recent studies on modeling biological face recognition with DCNNs. On the basis of these findings, we conclude that they are useful models for studying biological face recognition under the right circumstances. To begin with, DCNNs are largely able to exhibit face recognition and identification behaviors similar to humans, and even account for several "psychological phenomena" from face recognition research such as the "face inversion effect," the "Thatcher effect," face familiarity effects, and the "other-race effect." In addition, DCNNs seem to encode and process face information in a highly similar manner compared with the ventral visual pathway and core face network of humans and nonhuman primates. This synthesis is consistent with another recent review on this topic by O'Toole and Castillo (2021), which describes four technical stages of face recognition learning in DCNNs: The first stage involves between-categories learning for object categorization, which defines domain-general object experience. The second and third stages involve within-category learning for face identification and adaptation to the specific characteristics of individuals and environments, which describe domain-specific face experience. The fourth stage of learning individual people corresponds to identity-specific experience. Crucially, the two spotlights presented add an additional stage before the first and after the final stage of this theoretical model (see Figure 2). This way, the foundation of face recognition learning in DCNNs may lie in the ability of feedforward hierarchies to enable the development of face selectivity even in the absence of visual experience, whereas the peak may consist of additional generative mechanisms that improve the effectiveness and robustness of face identification. Modeling biological face recognition with DCNNs has so far not only provided exciting novel insights, but its opportunity to control nature and nurture has opened up a new perspective on many levels for further investigation.

OPEN QUESTIONS AND FUTURE DIRECTIONS

The previous successes from modeling biological object recognition with DCNNs should be transferred and extended to the field of biological face recognition. In particular, more causal approaches, such as the recent experiments regarding in silico lesioning analyses (e.g., Kanwisher, Gupta, et al., 2023; Blauch et al., 2022; Dobs,

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

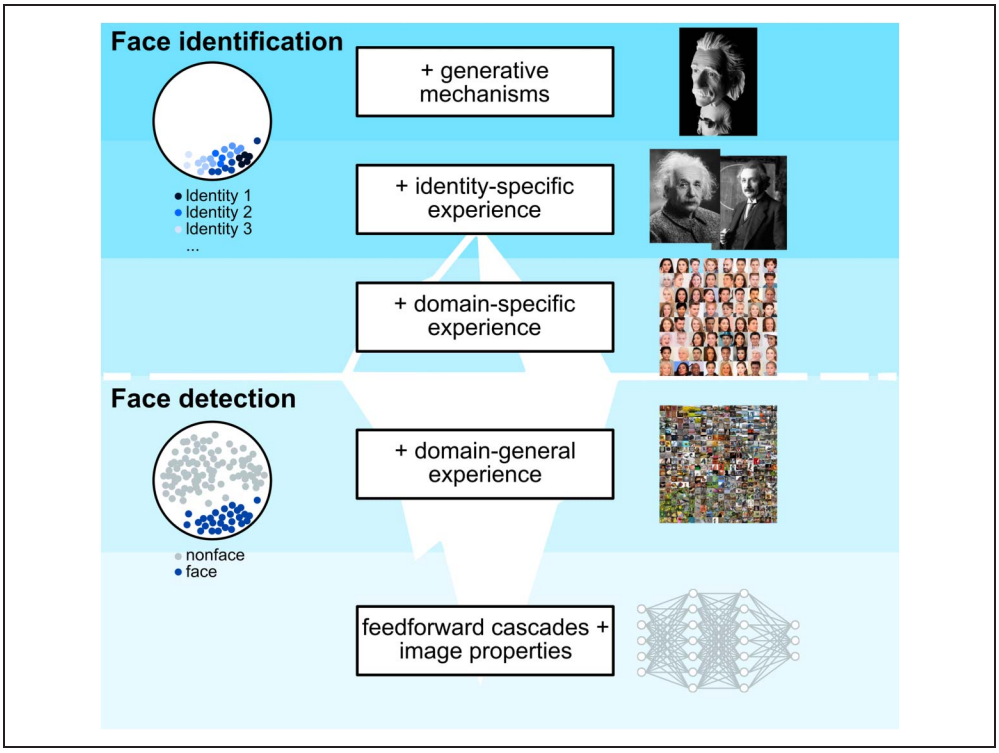

Figure 2. Summary model of face recognition learning in DCNNs. Elementary face selectivity, and hence face detection, arises automatically through feedforward processes without visual experience. Domain-general visual experience (i.e., various object categories) increases the ability to detect faces. Domain-specific visual experience (i.e., various face identities) further increases the ability to detect faces and serves as a basis for the ability to identify faces. Identity-specific experience (i.e., multiple examples of a given identity) enables robust face identification. Finally, this ability seems to peak with additional generative mechanisms, as the model is capable of inferring a 3-D representation from a 2-D image. Source: Images of objects obtained from the ImageNet database (Deng et al., 2009). To protect identities, the displayed non-famous faces were generated by an artificial neural network (https://thispersondoesnotexist.com/).

Martinez, et al., 2022; Wang et al., 2022), deprivation studies (e.g., Baek et al., 2021; Xu et al., 2021), and the synthesis of maximal activations (e.g., Vinken et al., 2022; Ratan Murty et al., 2021), should be further expanded to refute the criticisms of "prediction-based experiments" and to design primarily "controlled experiments" with adequate alternative explanations (Bowers et al., 2022). To shape this development in a targeted manner, it seems particularly important that future studies report aspects such as applied architectures, data sets, tasks, and hyperparameters with greater detail and transparency (see Table 1). On the one side, the advantage that architecture can be fully controlled in DCNNs should be further exploited. Instead of performing the same analyses with a similar model to control for differences in architecture, it might be helpful to also include particularly dissimilar models that should not demonstrate a given phenomenon. The addition of, for example, generative mechanisms, recurrent connections, biologically plausible connectivity, attentional mechanisms, and unsupervised learning could yield new insights about the influences of architecture. On the other side, control over the data set should also be further leveraged. Although more comprehensive and naturalistic data sets have the potential for better results (e.g., Allen et al., 2022; Hebart et al., 2019), the addition of more creative data sets, such as those used in recent work on fine-grained identification of nonface categories (Dobs, Yuan, et al., 2022; Yovel et al., 2022b), and customized data sets, such as artificial "controversial stimuli" created to reveal model preferences in synthesized images of mixed categories (Golan, Raju, & Kriegeskorte, 2020), could further strengthen the results and should be included.

Another promising line of future research involves comparisons with other measurements and populations. Although to date most findings are based on comparisons with fMRI data, other measurement techniques should be incorporated. As an example, comparisons with electrophysiological data could help to better understand how the representations encoded in the brain at different time points map to different model layers (e.g., Cichy et al., 2016) in the case of face recognition. As another example, comparisons of eye tracking data and model saliency maps could also shed more light on the attentional mechanisms (e.g., van Dyck, Denzler, & Gruber, 2022) underlying face detection and identification. Furthermore, as developmental studies showed that the gaze behavior of 4-month-old infants is more consistent with earlier model layers, whereas that of 12-month-old infants is more consistent with later layers (Kiat et al., 2022), this suggested a gradual shift toward greater abstraction may be especially interesting in the case of face recognition. Similarly, recent studies in young infants have shown that face selectivity develops early on (Kosakowski et al., 2022). Therefore, longitudinal studies could be used to link the stages of face recognition learning in the infant brain with those observed in DCNNs. In addition, comparisons involving patients with impaired face recognition or other species

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

Table 1. List of Reported Studies that Compared Biological Face Recognition and DCNNs

Study	*DCNN Model Family*	*Training Data Set*	*Training Task*	*Testing Data Set*	*Testing Task*
Abudarham et al. (2021)	GoogLeNet, VGG	ImageNet, VGGFace2	object-cat, face-id	custom	face-id
Abudarham et al. (2019)	OpenFace	CASIA-WebFace, FaceScrub	face-id	custom	face-id
Baek et al. (2021)	AlexNet	none	untrained	ImageNet, VGGFace2	encoding
Blauch et al. (2022)	interactive topographic network (based on ResNet)	ImageNet, VGGFace2, Places365	object-cat + face-id + scene-cat	–	face-id
Blauch et al. (2021)	VGG	ImageNet, LFIW, VGGFace2	untrained, object-cat, face-id	LFIW, VGGFace2	face-id
Chang et al. (2021)	AlexNet, CORnet, VGG	ImageNet, VGGFace2	object-cat, face-id	various	face-id
Colón et al. (2021)	ResNet	Universe	face-id	Karolinska Directed Emotional Faces	encoding
Daube et al. (2021)	ResNet (various objective functions)	artificially generated face stimuli	various	artificially generated face stimuli	face-id
Dobs, Martinez, et al. (2022)	VGG	ImageNet, VGGFace2	object-cat, face-id, dual-task	various	object-cat, face-id
Dobs, Yuan, et al. (2022)[a]	AlexNet, VGG	ImageNet, VGGFace2, others	untrained, object-cat, face-de, face-id	LFIW, VGGFace2	encoding
Doshi and Konkle (2023)[a]	AlexNet, SOMs	ImageNet	object-cat	various	encoding
Grossman et al. (2019)	VGG	ImageNet, VGGFace2	object-cat, face-id	various	encoding
Jacob et al. (2021)	VGG	ImageNet, VGGFace2	object-cat, face-id	IISc Indian Face Dataset	encoding, "Thatcher effect"
Jiahui et al. (2022)[a]	AlexNet, ResNet, VGG	ImageNet, MS-Celeb-1 M	object-cat, face-id	LFIW, custom	face-id, encoding
Jozwik et al. (2022)	AlexNet, VGG	artificially generated faces	various tasks	artificially generated faces	encoding, similarity judgments
Kanwisher, Gupta, et al. (2023)	VGG	ImageNet, VGGFace2	object-cat, face-id, car identification[b]	CompCars	car identification
Keles et al. (2021)	ResNet, VGG	Chicago Face Database, ImageNet, others	face-id	various	encoding, social judgments
Lee et al. (2020)[a]	topographic DCNNs (based on AlexNet)	ImageNet, LFIW, Places365, Open Images	object-cat	various	encoding
Noyes et al. (2021)	Sankaranarayanan et al. (2016)	CASIA-WebFace	face-id	FAÇADE	face-id

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

Table 1. (*continued*)

Study	*DCNN Model Family*	*Training Data Set*	*Training Task*	*Testing Data Set*	*Testing Task*
Parde et al. (2021)	ResNet	Universe	face-id	IARPA Janus Benchmark-C	face-id
Phillips et al. (2018)	various	various	face-id	custom	face-id
Raman and Hosoya (2020)	AlexNet, VGG	VGGFace2, (ImageNet, Oxford-102)	face-id	FEI face database, custom	encoding
Ratan Murty et al. (2021)	various	ImageNet, Places2, VGGFace2	object-cat, scene-cat, face-id	THINGS	face-de
Tian et al. (2022)	AlexNet, VGG	ImageNet, VGG-Face	object-cat, face-id	CASIA-WebFace, custom	encoding, "face inversion effect"
Tian et al. (2021)	VGG	VGG-Face	face-id	VGGFace2	face-id, "other-race effect"
Tsantani et al. (2021)	OpenFace	CASIA-WebFace, FaceScrub	face-id	custom	face-id
Vinken et al. (2022)[a]	Alexnet, GoogLeNet, ResNet, VGG	ImageNet	object-cat	custom	encoding
Wang et al. (2022)	VGG	CelebA, VGGFace2, (ImageNet)	face-id	custom	face-id
Xu et al. (2018)	ResNet	artificially generated face stimuli	face-id	artificially generated face stimuli	face-id
Xu et al. (2021)	AlexNet, ResNet	ImageNet, (Faces in the Wild, CASIA-WebFace)	object-cat	custom	face-de, face-id
Yildirim et al. (2020)	efficient inverse graphics model (based on AlexNet), VGG	artificially generated faces	face-id	face-identities-view	face-id
Yovel et al. (2022a)[a]	VGG	ImageNet, VGGFace2, others	object-cat, face-id, bird identification[b]	custom	encoding, "face inversion effect"
Yovel et al. (2022b)[a]	VGG	ImageNet, VGGFace2, others	object-cat, face-id, bird identification[b]	custom	bird identification
Zeman et al. (2022)[a]	AlexNet, DeepFace	ImageNet, CASIA-WebFace	object-cat, face-id	Mooney face online data set	face-de, "face inversion effect"

[a] Preprint at the time of writing.

[b] Different fine-tuning.

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

(e.g., monkeys or rodents) might be useful to extend the robustness of the current findings.

As previous studies have focused primarily on the physical properties of biological face recognition, future studies may also attempt to illuminate the functional properties. Initial studies with this aim have already demonstrated that social aspects of face recognition such as facial expression (Colón, Castillo, & O'Toole, 2021) and social judgments (Keles, Lin, & Adolphs, 2021) may already be partially encoded by associative learning in current face-id models.

Finally, it is worth noting that face recognition in real-world applications, such as those for security, biometric authentication, marketing, and healthcare, is much more challenging than in closely controlled experiments. Although the studies summarized in this review offer exciting insights for basic research at the intersection of neuroscience and artificial intelligence, the difficulty often lies in their practical implementation (for more technical perspectives, see Guo, Wang, Yan, Zheng, & Li, 2020; Li, Lin, Shen, Brandt, & Hua, 2015).

Reprint requests should be sent to Leonard Elia van Dyck, Department of Psychology, Centre for Cognitive Neuroscience, University of Salzburg, Hellbrunnerstr. 34, Salzburg, Austria, or via e-mail: lenny.vandyck@gmail.com.

Author Contributions

Leonard Elia van Dyck: Conceptualization; Funding acquisition; Visualization; Writing—Original draft; Writing—Review & editing. Walter Roland Gruber: Conceptualization; Funding acquisition; Supervision; Writing—Review & editing.

Funding Information

Universität Salzburg (https://dx.doi.org/10.13039/501100005644), grant number: Open Access Publication Fee.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

REFERENCES

Abudarham, N., Grosbard, I., & Yovel, G. (2021). Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition. *Cognitive Science*, *45*, e13031. https://doi.org/10.1111/cogs.13031, PubMed: 34490907

Abudarham, N., Shkiller, L., & Yovel, G. (2019). Critical features for face recognition. *Cognition*, *182*, 73–83. https://doi.org/10.1016/j.cognition.2018.09.002, PubMed: 30218914

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, *16*, 40. https://doi.org/10.1167/16.3.40, PubMed: 26928056

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*, 116–126. https://doi.org/10.1038/s41593-021-00962-x, PubMed: 34916659

Baek, S., Song, M., Jang, J., Kim, G., & Paik, S.-B. (2021). Face detection in untrained deep neural networks. *Nature Communications*, *12*, 7328. https://doi.org/10.1038/s41467-021-27606-9, PubMed: 34916514

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*, 103–108. https://doi.org/10.1038/s41586-020-2350-5, PubMed: 32494012

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2021). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition*, *208*, 104341. https://doi.org/10.1016/j.cognition.2020.104341, PubMed: 32586632

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2022). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *119*, e2112566119. https://doi.org/10.1073/pnas.2112566119, PubMed: 35027449

Bothwell, R. K., Brigham, J. C., & Malpass, R. S. (1989). Cross-racial identification. *Personality and Social Psychology Bulletin*, *15*, 19–25. https://doi.org/10.1177/0146167289151002

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., et al. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 1–74. https://doi.org/10.1017/S0140525X22002813, PubMed: 36453586

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327. https://doi.org/10.1111/j.2044-8295.1986.tb02199.x, PubMed: 3756376

Burton, A. M., Jenkins, R., Hancock, P. J. B., & White, D. (2005). Robust representations for face recognition: The power of averages. *Cognitive Psychology*, *51*, 256–284. https://doi.org/10.1016/j.cogpsych.2005.06.003, PubMed: 16198327

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*, e1003963. https://doi.org/10.1371/journal.pcbi.1003963, PubMed: 25521294

Campbell, A., & Tanaka, J. W. (2018). Inversion impairs expert budgerigar identity recognition: A face-like effect for a nonface object of expertise. *Perception*, *47*, 647–659. https://doi.org/10.1177/0301006618771806, PubMed: 29690836

Cao, R., Li, X., Brandmeir, N. J., & Wang, S. (2021). Encoding of facial features by single neurons in the human amygdala and hippocampus. *Communications Biology*, *4*, 1394. https://doi.org/10.1038/s42003-021-02917-1, PubMed: 34907323

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

Chang, L., Egger, B., Vetter, T., & Tsao, D. Y. (2021). Explaining face representation in the primate brain using different computational models. *Current Biology*, *31*, 2785–2795. https://doi.org/10.1016/j.cub.2021.04.014, PubMed: 33951457

Chang, L., & Tsao, D. Y. (2017). The code for facial identity in the primate brain. *Cell*, *169*, 1013–1028. https://doi.org/10.1016/j.cell.2017.05.011, PubMed: 28575666

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755. https://doi.org/10.1038/srep27755, PubMed: 27282108

Colón, 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. https://doi.org/10.1167/jov.21.4.4, PubMed: 33821927

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*, 681–685. https://doi.org/10.1109/34.927467

Cowell, R. A., & Cottrell, G. W. (2013). What evidence supports special processing for faces? A cautionary tale for fMRI interpretation. *Journal of Cognitive Neuroscience*, *25*, 1777–1793. https://doi.org/10.1162/jocn_a_00448, PubMed: 23859648

Daube, C., Xu, T., Zhan, J., Webb, A., Ince, R. A. A., Garrod, O. G. B., et al. (2021). Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns*, *2*, 100348. https://doi.org/10.1016/j.patter.2021.100348, PubMed: 34693374

De Falco, E., Ison, M. J., Fried, I., & Quian Quiroga, R. (2016). Long-term coding of personal and universal associations underlying the memory web in the human brain. *Nature Communications*, *7*, 13408. https://doi.org/10.1038/ncomms13408, PubMed: 27845773

Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, *2009*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Desimone, R. (1991). Face-selective cells in the temporal cortex of monkeys. *Journal of Cognitive Neuroscience*, *3*, 1–8. https://doi.org/10.1162/jocn.1991.3.1.1, PubMed: 23964801

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General*, *115*, 107–117. https://doi.org/10.1037/0096-3445.115.2.107, PubMed: 2940312

Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances*, *8*, eabl8913. https://doi.org/10.1126/sciadv.abl8913, PubMed: 35294241

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2022). Using deep convolutional neural networks to test why human face recognition works the way it does. *bioRxiv*. https://doi.org/10.1101/2022.11.23.517478

Doerig, A., Sommers, R., Seeliger, K., Richards, B., Ismael, J., Lindsay, G., et al. (2022). The neuroconnectionist research programme. *arXiv:2209.03718*. https://doi.org/10.48550/arXiv.2209.03718

Doshi, F. R., & Konkle, T. (2023). Visual object topographic motifs emerge from self-organization of a unified representational space. *bioRxiv*. https://doi.org/10.1101/2022.09.06.506403

Edwards, G. J., Taylor, C. J., & Cootes, T. F. (1998). Interpreting face images using active appearance models. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition* (pp. 300–305). https://doi.org/10.1109/AFGR.1998.670965

Farfade, S. S., Saberian, M. J., & Li, L.-J. (2015). Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval* (pp. 643–650). https://doi.org/10.1145/2671188.2749408

Farzmahdi, A., Rajaei, K., Ghodrati, M., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2016). A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans. *Scientific Reports*, *6*, 25025. https://doi.org/10.1038/srep25025, PubMed: 27113635

Freiwald, W. A. (2020). The neural mechanisms of face processing: Cells, areas, networks, and models. *Current Opinion in Neurobiology*, *60*, 184–191. https://doi.org/10.1016/j.conb.2019.12.007, PubMed: 31958622

Freiwald, W. A., Duchaine, B., & Yovel, G. (2016). Face processing systems: From neurons to real-world social perception. *Annual Review of Neuroscience*, *39*, 325–346. https://doi.org/10.1146/annurev-neuro-070815-013934, PubMed: 27442071

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*, 845–851. https://doi.org/10.1126/science.1194908, PubMed: 21051642

Gauthier, I., & Bukach, C. (2007). Should we reject the expertise hypothesis? *Cognition*, *103*, 322–330. https://doi.org/10.1016/j.cognition.2006.05.003, PubMed: 16780825

Gauthier, I., & Nelson, C. A. (2001). The development of face expertise. *Current Opinion in Neurobiology*, *11*, 219–224. https://doi.org/10.1016/S0959-4388(00)00200-2, PubMed: 11301243

Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schoenborn, S., et al. (2018). Morphable face models—An open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (pp. 75–82). https://doi.org/10.1109/FG.2018.00021

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences, U.S.A.*, *117*, 29330–29337. https://doi.org/10.1073/pnas.1912334117, PubMed: 33229549

Greene, M. R., & Hansen, B. C. (2018). Shared spatiotemporal category representations in biological and artificial deep neural networks. *PLoS Computational Biology*, *14*, e1006327. https://doi.org/10.1371/journal.pcbi.1006327, PubMed: 30040821

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., et al. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications*, *10*, 4934. https://doi.org/10.1038/s41467-019-12623-6, PubMed: 31666525

Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, *35*, 10005. https://doi.org/10.1523/JNEUROSCI.5023-14.2015, PubMed: 26157000

Guo, G., Wang, H., Yan, Y., Zheng, J., & Li, B. (2020). A fast face detection method via convolutional neural network. *Neurocomputing*, *395*, 128–137. https://doi.org/10.1016/j.neucom.2018.02.110

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2430. https://doi.org/10.1126/science.1063736, PubMed: 11577229

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*, 223–233. https://doi.org/10.1016/s1364-6613(00)01482-0, PubMed: 10827445

Haxby, J. V., Ungerleider, L. G., Clark, V. P., Schouten, J. L., Hoffman, E. A., & Martin, A. (1999). The effect of face inversion on activity in human neural systems for face and object perception. *Neuron*, *22*, 189–199. https://doi.org/10.1016/S0896-6273(00)80690-X, PubMed: 10027301

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., et al. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS One*, *14*, e0223792. https://doi.org/10.1371/journal.pone.0223792, PubMed: 31613926

Hesse, J. K., & Tsao, D. Y. (2020). The macaque face patch system: A turtle's underbelly for the brain. *Nature Reviews Neuroscience*, *21*, 695–716. https://doi.org/10.1038/s41583-020-00393-w, PubMed: 33144718

Hosoya, H., & Hyvärinen, A. (2017). A mixture of sparse coding models explaining properties of face neurons related to holistic and parts-based processing. *PLoS Computational Biology*, *13*, e1005667. https://doi.org/10.1371/journal.pcbi.1005667, PubMed: 28742816

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*. https://hal.inria.fr/inria-00321923

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, *148*, 574–591. https://doi.org/10.1113/jphysiol.1959.sp006308, PubMed: 14403679

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*, 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837, PubMed: 14449617

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications*, *12*, 1872. https://doi.org/10.1038/s41467-021-22078-3, PubMed: 33767141

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th International Conference on Computer Vision* (pp. 2146–2153). https://doi.org/10.1109/ICCV.2009.5459469

Jiahui, G., Feilong, M., di Oleggio Castello, M. V., Nastase, S. A., Haxby, J. V., & Gobbini, M. I. (2022). Modeling naturalistic face processing in humans with deep convolutional neural networks. *bioRxiv*. https://doi.org/10.1101/2021.11.17.469009

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, *40*, 1–19. https://doi.org/10.1016/0010-0277(91)90045-6, PubMed: 1786670

Johnson, M. H., & Mareschal, D. (2001). Cognitive and perceptual development during infancy. *Current Opinion in Neurobiology*, *11*, 213–218. https://doi.org/10.1016/S0959-4388(00)00199-9, PubMed: 11301242

Johnston, R. A., & Edmonds, A. J. (2009). Familiar and unfamiliar face recognition: A review. *Memory*, *17*, 577–596. https://doi.org/10.1080/09658210902976969, PubMed: 19548173

Jozwik, K. M., O'Keeffe, J., Storrs, K. R., Guo, W., Golan, T., & Kriegeskorte, N. (2022). Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proceedings of the National Academy of Sciences, U.S.A.*, *119*, e2115047119. https://doi.org/10.1073/pnas.2115047119, PubMed: 35767642

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, *3*, 759–763. https://doi.org/10.1038/77664, PubMed: 10903567

Kanwisher, N., Gupta, P., & Dobs, K. (2023). CNNs reveal the computational implausibility of the expertise hypothesis. *iScience*, *26*, 105976. https://doi.org/10.1016/j.isci.2023.105976, PubMed: 36794151

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*, *46*, 240–254. https://doi.org/10.1016/j.tins.2022.12.008, PubMed: 36658072

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*, 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997, PubMed: 9151747

Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition*, *68*, B1–B11. https://doi.org/10.1016/S0010-0277(98)00035-3, PubMed: 9775518

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *361*, 2109–2128. https://doi.org/10.1098/rstb.2006.1934, PubMed: 17118927

Keles, U., Lin, C., & Adolphs, R. (2021). A cautionary note on predicting social judgments from faces with deep neural networks. *Affective Science*, *2*, 438–454. https://doi.org/10.1007/s42761-021-00075-5, PubMed: 34966898

Keller, T. A., Gao, Q., & Welling, M. (2021). Modeling category-selective cortical regions with topographic variational autoencoders. *arXiv:2110.13911*. https://doi.org/10.48550/arXiv.2110.13911

Kemelmacher-Shlizerman, I., Seitz, S. M., Miller, D., & Brossard, E. (2016). The MegaFace benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4873–4882). https://doi.org/10.1109/CVPR.2016.527

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*, e1003915. https://doi.org/10.1371/journal.pcbi.1003915, PubMed: 25375136

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific Reports*, *6*, 32672. https://doi.org/10.1038/srep32672, PubMed: 27601096

Kiat, J. E., Luck, S. J., Beckner, A. G., Hayes, T. R., Pomaranski, K. I., Henderson, J. M., et al. (2022). Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis. *Developmental Science*, *25*, e13155. https://doi.org/10.1111/desc.13155, PubMed: 34240787

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*, 1464–1480. https://doi.org/10.1109/5.58325

Kosakowski, H. L., Cohen, M. A., Takahashi, A., Keil, B., Kanwisher, N., & Saxe, R. (2022). Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. *Current Biology*, *32*, 265–274. https://doi.org/10.1016/j.cub.2021.10.064, PubMed: 34784506

Kramer, R. S. S., Ritchie, K. L., & Burton, A. M. (2015). Viewers extract the mean from images of the same person: A route to face learning. *Journal of Vision*, *15*, 1. https://doi.org/10.1167/15.4.1, PubMed: 26067175

Kramer, R. S. S., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition*, *172*, 46–58. https://doi.org/10.1016/j.cognition.2017.12.005, PubMed: 29232594

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

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).

Landi, S. M., & Freiwald, W. A. (2017). Two areas for familiar face recognition in the primate brain. *Science*, *357*, 591–595. https://doi.org/10.1126/science.aan1139, PubMed: 28798130

Landi, S. M., Viswanathan, P., Serene, S., & Freiwald, W. A. (2021). A fast link between face perception and memory in the temporal pole. *Science*, *373*, 581–585. https://doi.org/10.1126/science.abi6671, PubMed: 34210891

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539, PubMed: 26017442

Lee, H., Margalit, E., Jozwik, K. M., Cohen, M. A., Kanwisher, N., Yamins, D. L. K., et al. (2020). Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*. https://doi.org/10.1101/2020.07.09.185116

Leek, C. E., Leonardis, A., & Heinke, D. (2022). Deep neural networks and image classification in biological vision. *Vision Research*, *197*, 108058. https://doi.org/10.1016/j.visres.2022.108058, PubMed: 35487146

Li, H., Lin, Z., Shen, X., Brandt, J., & Hua, G. (2015). A convolutional neural network cascade for face detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5325–5334). https://doi.org/10.1109/CVPR.2015.7299170

Marcus, G. F. (2018). Deep learning: A critical appraisal. *arXiv:1801.00631*. https://doi.org/10.48550/arXiv.1801.00631

Noyes, E., & Jenkins, R. (2019). Deliberate disguise in face identification. *Journal of Experimental Psychology: Applied*, *25*, 280–290. https://doi.org/10.1037/xap0000213, PubMed: 30730157

Noyes, E., Parde, C. J., Colón, Y. I., Hill, M. Q., Castillo, C. D., Jenkins, R., et al. (2021). Seeing through disguise: Getting to know you with a deep convolutional neural network. *Cognition*, *211*, 104611. https://doi.org/10.1016/j.cognition.2021.104611, PubMed: 33592392

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, U.S.A.*, *105*, 11087–11092. https://doi.org/10.1073/pnas.0805664105, PubMed: 18685089

O'Toole, A. J., & Castillo, C. D. (2021). Face recognition by humans and machines: Three fundamental advances from deep learning. *Annual Review of Vision Science*, *7*, 543–570. https://doi.org/10.1146/annurev-vision-093019-111701, PubMed: 34348035

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*, 794–809. https://doi.org/10.1016/j.tics.2018.06.006, PubMed: 30097304

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*, 15. https://doi.org/10.1167/jov.21.8.15, PubMed: 34379084

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., & Vetter, T. (2009). A 3D face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance* (pp. 296–301). https://doi.org/10.1109/AVSS.2009.58

Perrett, D. I., Hietanen, J. K., Oram, M. W., & Benson, P. J. (1992). Organization and functions of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, *335*, 23–30. https://doi.org/10.1098/rstb.1992.0003, PubMed: 1348133

Perrett, D. I., Mistlin, A. J., & Chitty, A. J. (1987). Visual neurones responsive to faces. *Trends in Neurosciences*, *10*, 358–364. https://doi.org/10.1016/0166-2236(87)90071-3

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences, U.S.A.*, *115*, 6171–6176. https://doi.org/10.1073/pnas.1721355115, PubMed: 29844174

Pinto, N., Doukhan, D., DiCarlo, J. J., & Cox, D. D. (2009). A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Computational Biology*, *5*, e1000579. https://doi.org/10.1371/journal.pcbi.1000579, PubMed: 19956750

Raman, R., & Hosoya, H. (2020). Convolutional neural networks explain tuning properties of anterior, but not middle, face-processing areas in macaque inferotemporal cortex. *Communications Biology*, *3*, 221. https://doi.org/10.1038/s42003-020-0945-x, PubMed: 32385392

Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J., & Kanwisher, N. (2021). Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nature Communications*, *12*, 5540. https://doi.org/10.1038/s41467-021-25409-6, PubMed: 34545079

Ratan Murty, N. A., Teng, S., Beeler, D., Mynick, A., Oliva, A., & Kanwisher, N. (2020). Visual experience is not necessary for the development of face-selectivity in the lateral fusiform gyrus. *Proceedings of the National Academy of Sciences, U.S.A.*, *117*, 23011–23020. https://doi.org/10.1073/pnas.2004607117, PubMed: 32839334

Rossion, B., & Gauthier, I. (2002). How does the brain process upright and inverted faces? *Behavioral and Cognitive Neuroscience Reviews*, *1*, 63–75. https://doi.org/10.1177/1534582302001001004, PubMed: 17715586

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

Spiridon, M., & Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron*, *35*, 1157–1165. https://doi.org/10.1016/s0896-6273(02)00877-2, PubMed: 12354404

Srihasam, K., Mandeville, J. B., Morocz, I. A., Sullivan, K. J., & Livingstone, M. S. (2012). Behavioral and anatomical consequences of early versus late symbol training in macaques. *Neuron*, *73*, 608–619. https://doi.org/10.1016/j.neuron.2011.12.022, PubMed: 22325210

Srihasam, K., Vincent, J. L., & Livingstone, M. S. (2014). Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nature Neuroscience*, *17*, 1776–1783. https://doi.org/10.1038/nn.3855, PubMed: 25362472

Sugita, Y. (2008). Face perception in monkeys reared with no exposure to faces. *Proceedings of the National Academy of Sciences, U.S.A.*, *105*, 394–398. https://doi.org/10.1073/pnas.0706079105, PubMed: 18172214

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). https://doi.org/10.1109/CVPR.2014.244

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*, 105–118. https://doi.org/10.1016/j.cognition.2012.12.001, PubMed: 23376296

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025

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

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, A: Human Experimental Psychology*, *46*, 225–245. https://doi.org/10.1080/14640749308401045, PubMed: 8316637

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception*, *9*, 483–484. https://doi.org/10.1068/p090483, PubMed: 6999452

Tian, J., Xie, H., Hu, S., & Liu, J. (2021). Multidimensional face representation in a deep convolutional neural network reveals the mechanism underlying AI racism. *Frontiers in Computational Neuroscience*, *15*, 620281. https://doi.org/10.3389/fncom.2021.620281, PubMed: 33776675

Tian, F., Xie, H., Song, Y., Hu, S., & Liu, J. (2022). The face inversion effect in deep convolutional neural networks. *Frontiers in Computational Neuroscience*, *16*, 854218. https://doi.org/10.3389/fncom.2022.854218, PubMed: 35615057

Tolba, A., El-Baz, A., & El-Harby, A. (2006). Face recognition: A literature review. *International Journal of Signal Processing*, *2*, 88–103.

Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021). FFA and OFA encode distinct types of face identity information. *Journal of Neuroscience*, *41*, 1952–1969. https://doi.org/10.1523/JNEUROSCI.1449-20.2020, PubMed: 33452225

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*, 670–674. https://doi.org/10.1126/science.1119983, PubMed: 16456083

Tsao, D. Y., & Livingstone, M. S. (2008). Mechanisms of face perception. *Annual Review of Neuroscience*, *31*, 411–437. https://doi.org/10.1146/annurev.neuro.30.051606.094238, PubMed: 18558862

Valentine, T., Lewis, M. B., & Hills, P. J. (2016). Face-space: A unifying concept in face recognition research. *Quarterly Journal of Experimental Psychology*, *69*, 1996–2019. https://doi.org/10.1080/17470218.2014.990392, PubMed: 25427883

van den Hurk, J., Van Baelen, M., & Op de Beeck, H. P. (2017). Development of visual category selectivity in ventral visual cortex does not require visual experience. *Proceedings of the National Academy of Sciences, U.S.A.*, *114*, E4501–E4510. https://doi.org/10.1073/pnas.1612862114, PubMed: 28507127

van Dyck, L. E., Denzler, S. J., & Gruber, W. R. (2022). Guiding visual attention in deep convolutional neural networks based on human eye movements. *Frontiers in Neuroscience*, *16*, 975639. https://doi.org/10.3389/fnins.2022.975639, PubMed: 36177359

VanRullen, R., & Reddy, L. (2019). Reconstructing faces from fMRI patterns using deep generative neural networks. *Communications Biology*, *2*, 193. https://doi.org/10.1038/s42003-019-0438-y, PubMed: 31123717

Vinken, K., Konkle, T., & Livingstone, M. (2022). The neural code for 'face cells' is not face specific. *bioRxiv*. https://doi.org/10.1101/2022.03.06.483186

Wang, J., Cao, R., Brandmeir, N. J., Li, X., & Wang, S. (2022). Face identity coding in the deep neural network and primate brain. *Communications Biology*, *5*, 611. https://doi.org/10.1038/s42003-022-03557-9, PubMed: 35725902

Xu, S., Zhang, Y., Zhen, Z., & Liu, J. (2021). The face module emerged in a deep convolutional neural network selectively deprived of face experience. *Frontiers in Computational Neuroscience*, *15*, 626259. https://doi.org/10.3389/fncom.2021.626259, PubMed: 34093154

Xu, T., Garrod, O., Scholte, S. H., Ince, R., & Schyns, P. G. (2018). Using psychophysical methods to understand mechanisms of face identification in a deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1976–1984). https://doi.org/10.1109/CVPRW.2018.00266

Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NeurIPS)* (pp. 3093–3101).

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, U.S.A.*, *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111, PubMed: 24812127

Yildirim, I., Belledonne, M., Freiwald, W., & Tenenbaum, J. (2020). Efficient inverse graphics in biological face processing. *Science Advances*, *6*, eaax5979. https://doi.org/10.1126/sciadv.aax5979, PubMed: 32181338

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*, 141–145. https://doi.org/10.1037/h0027474

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences*, *22*, 100–110. https://doi.org/10.1016/j.tics.2017.11.007, PubMed: 29254899

Young, A. W., & Burton, A. M. (2021). Insights from computational models of face recognition: A reply to Blauch, Behrmann and Plaut. *Cognition*, *208*, 104422. https://doi.org/10.1016/j.cognition.2020.104422, PubMed: 32800311

Yovel, G., Grosbard, I., & Abudarham, N. (2022a). Computational models of perceptual expertise reveal a domain-specific inversion effect for objects of expertise. *PsyArXiv*. https://doi.org/10.31234/osf.io/yv574

Yovel, G., Grosbard, I., & Abudarham, N. (2022b). Deep learning models of perceptual expertise support a domain-specific account. *bioRxiv*. https://doi.org/10.1101/2022.12.01.518342

Yovel, G., & Kanwisher, N. (2005). The neural basis of the behavioral face-inversion effect. *Current Biology*, *15*, 2256–2262. https://doi.org/10.1016/j.cub.2005.10.072, PubMed: 16360687

Zeman, A., Leers, T., & de Beeck, H. O. (2022). Mooney face image processing in deep convolutional neural networks compared to humans. *bioRxiv*. https://doi.org/10.1101/2022.03.21.485240

Downloaded from http://direct.mit.edu/jocn/article-pdf/35/10/1521/2157635/jocn_a_02040.pdf by guest on 26 April 2025