Concurrent emergence of view invariance, sensitivity to critical features, and identity face classification through visual experience: Insights from deep learning algorithms

Mandy Rosemblaum[*]
School of Neuroscience,
Tel Aviv University, Tel Aviv, Israel

Nitzan Guy[*]
School of Psychological Sciences,
Tel Aviv University, Tel Aviv, Israel

Idan Grosbard
School of Neuroscience,
Tel Aviv University, Tel Aviv, Israel

Libi Kliger
School of Psychological Sciences,
Tel Aviv University, Tel Aviv, Israel

Naphtali Abudarham
School of Psychological Sciences,
Tel Aviv University, Tel Aviv, Israel

Galit Yovel
School of Neuroscience,
Tel Aviv University, Tel Aviv, Israel
School of Psychological Sciences,
Tel Aviv University, Tel Aviv, Israel

Visual experience is known to play a critical role in face recognition. This experience is thought to enable the formation of a view-invariant representation by learning which features are critical to identify faces across views. Discovering these critical features and the type of experience that is needed to uncover them is challenging. A recent study revealed a subset of facial features that are critical for human face recognition. Furthermore, face-trained deep convolutional neural networks (DCNNs) were sensitive to these facial features. These findings enable us now to ask what type of face experience is required for the network to become sensitive to these human-like critical features, and whether it is associated with the formation of a view-invariant representation and face classification performance. To that end, we systematically manipulated the number of within-identity and between-identity face images during training and examined its effect on the network performance on face classification, view-invariant representation, and sensitivity to human-like critical facial features. Results show that increasing the number of images per identity, as well as the number of identities were both required for the simultaneous development of a view-invariant representation, sensitivity to human-like critical features, and successful identity classification. The concurrent emergence of sensitivity to critical features, view invariance and classification performance through experience implies that they depend on similar features. Overall, we show how systematic manipulation of the training diet of DCNNs can shed light on the role of experience in the generation of human-like representations.

Introduction

Object recognition is a computationally challenging task that humans resolve effortlessly. To successfully classify objects into different categories, the brain must create an identity-preserved representation that is tolerant to within-class changes, such as viewpoint, lighting, size, occlusion and so forth (DiCarlo & Cox, 2007). This is achieved by emphasizing features that

Citation: Rosemblaum, M., Guy, N., Grosbard, I., Kliger, L., Abudarham, N., & Yovel, G. (2025). Concurrent emergence of view invariance, sensitivity to critical features, and identity face classification through visual experience: Insights from deep learning algorithms. *Journal of Vision*, *25*(8):2, 1–12, https://doi.org/10.1167/jov.25.8.2.

https://doi.org/10.1167/jov.25.8.2 Received July 27, 2024; published July 1, 2025 ISSN 1534-7362 Copyright 2025 The Authors

remain unchanged across different variations, while disregarding features that vary across these variations (Raviv, Lupyan, & Green, 2022). The nature of the experience that is required for the visual system to learn which features are critical and generate a view-invariant representation has so far remained unknown.

Recent advancements in machine vision have successfully resolved the task of object and face recognition with deep convolutional neural networks. These algorithms, trained on thousands of images in a supervised or self-supervised manner, now perform on par with humans in face and object classification (Simonyan & Zisserman, 2014). Whereas the exact computations used by these algorithms and their similarity to the computations used by humans to resolve this task are unknown, recent studies have uncovered notable similarities between the representations generated by deep convolutional neural networks (DCNNs) and the human brain and mind (Abudarham, Grosbard, & Yovel, 2021; Cichy & Kaiser, 2019; Groen et al., 2018; O'Toole & Castillo, 2021). Thus, by studying the type of experience that is required to generate *human-like* representations in DCNNs, we can gain insights on the ingredients that are needed for these representations to emerge in humans.

In the current study, we adopted this approach to shed light on the visual experience necessary for creating a human-like, view-invariant representation of faces with DCNNs. The role of experience in human face recognition is well-established. Studies have shown that face recognition is better for familiar than unfamiliar faces (Young & Burton, 2018) and for faces from own race than other race faces for which we have greater visual experience (Laurence, Zhou, & Mondloch, 2016; McKone et al., 2019; Tanaka, Heptonstall, & Hagen, 2017). Moreover, developmental studies indicate that face recognition gradually improves with development, including the ability to generalize across different images of the same individual (Baker & Mondloch, 2019; Matthews & Mondloch, 2022). Many studies have emphasized the importance of experience with variable face images for successful face recognition (Baker & Mondloch, 2019; Honig, Shoham, & Yovel, 2022; Kramer, Jenkins, Young, & Burton, 2017; Ritchie & Burton, 2017). However, systematic manipulation of human real-life experience with faces is not possible and it is therefore hard to determine a direct link between the visual diet that humans are exposed to and its contribution to the generation of a face representation that enables their face recognition abilities.

In a recent set of studies, Abudarham and colleagues (Abudarham, Shkiller, & Yovel, 2019; Abudarham & Yovel, 2016; Abudarham & Yovel, 2020) discovered that humans are sensitive to a subset of facial features that are critical for face identification. Replacing these features changed the identity of a face (see Abudarham et al., 2019, figure 1). Moreover, Abudarham and colleagues (2016) found that human sensitivity to these critical features remained invariant across variations in head pose, which makes them potentially useful for view-invariant identity classification. They further revealed that face-trained, but not object-trained DCNNs, showed similar sensitivity to this subset of facial features. This indicates that experience with faces is necessary to learn to use these features for identity classification. These findings are also consistent with recent studies showing human-like face effects such as the face inversion effect and other-race effect in face-trained but not object-trained DCNNs (Dobs, Yuan, Martinez, & Kanwisher, 2023; Yovel, Grosbard, & Abudarham, 2023). Furthermore, sensitivity to these critical features and the generation of a view-invariant representation were found in higher layers of the face-trained network, whereas earlier layers showed no preference to this subset of face features and evidence for a view-specific face representation (Abudarham et al., 2021). This human-like representation enables us to link between humans and DCNNs view-invariant representations and examine the type of visual diet that is required for the development of successful identity classification.

To this end, in the current study we systematically manipulated the amount and type of visual-diet and examined its effect on the generation of a view-invariant representation, sensitivity to human-like critical features and performance on face identity in DCNNs. We manipulated experience by systematically increasing the number of within-identity images or between-identity images. This enabled us to assess the relative importance of between-identity and within-identity image variability. We then examined these models on the following measures: First, we examined performance of each of the trained DCNNs on a standard face verification benchmark, the Labeled Faces in the Wild, a face dataset that is commonly used to assess performance of face-trained DCNNs (Liao, Zhen, Dong, & Li, 2014) (Figure 1). Second, we measured the distance between faces that differ in head views to measure view-specific versus view invariance representations (Figure 2). We also measured the distance between faces that differ in noncritical features or critical features, to measure sensitivity to human-like critical features (Figure 3). These distances were measured based on the pixel-based representation or the identity-based representation of the images, which is generated in the last hidden layer of the fully trained face model. We then measured the similarity of the representation that was generated for faces in the last hidden layer of each trained network to the pixel-based representation and the identity-based representation of a fully trained face model. This enables us to determine whether increasing the amount of training makes the face representation more similar to the identity-based representation and less similar to the pixel-based representation. Moreover, concurrent emergence of view invariance and sensitivity to critical features along with improved performance on face identity verification as a function of the visual diet would suggest that they depend on similar features (Figure 4). It will further indicate the importance of experience to the generation of these features.

General methods

Model

We used VGG-16 (Simonyan & Zisserman, 2014) as a the base model, which we trained on different numbers of face images. We selected this model because it has been often used in previous studies (Abudarham et al., 2021; Blauch, Behrmann, & Plaut, 2021; Dobs et al., 2023). The representations used in the study are extracted from the penultimate layer (FC7).

Train dataset

We used the VGGFace2 dataset (Cao, Shen, Xie, Parkhi, & Zisserman, 2018) to train our networks. VGGFace2 is a large-scale face recognition dataset developed by the Visual Geometry Group at the University of Oxford. It contains over 3 million images of more than 8000 individuals, with each individual represented by several hundred images. The images were collected from a variety of sources and were annotated with bounding boxes and labels indicating the identity of the individuals.

Training protocol

We created 64 subsets of face images, which included 2, 5, 10, 50, 100, 200, 500, 1000 identities. For each of these number of identities we selected one, five, 10, 20, 50, 100, 200, or 300 images per identity. Because our test images were White, we also trained the DCNNs only on white faces. For the small training sets (1–100 identities, with all possible images per identity), we trained each DCNN on 30 different data sets to obtain robust performance measure of their representations/performance. The results were then averaged across the 30 networks. Representations were extracted also from the fully-trained model that was trained on the whole VGGface2 data set.

Stimuli

View-specific and view-invariant representations: To examine whether the network generates a view-specific or a view-invariant representation, we used images of 15 identities from the color FERET face-image dataset. For each identity, we selected four images: a “reference” frontal image, a second “frontal” image that is different from the reference image, a quarter-left image, and a half-left image. All images were of adult White males, well-lit, with no glasses, hats or facial hair. The images were cropped just below the chin to include only the face, including the hair and ears. This resulted in four types of face pairs: “Same-Frontal,” “Same-quarter view,” “Same- half view,” and “Different- Frontal” (See Figure 2A for examples of the four types of face pairs, the original face images used in the study were real faces not shown in this figure).

Critical features for face recognition: We used 25 face identities to generate image pairs. For each of the 25 identities, we used an original image, an image with modified critical features, and an image with modified noncritical features (for more information about how the face images were created see (Abudarham & Yovel, 2016). We also used a different unmodified image of the same person, which we used as a reference image. This allowed us to create four image pairs: the “Same” pair, which compares the reference image to the original image, the “Different” pair, which compares the reference image to a reference image of a different identity, the “Critical features” pair, which compares the reference image to the original image with different critical features, and the “Non-critical features” pair, which compares the reference image to the original image with different noncritical features. (See Figure 3A for examples of the four types of face pairs, the original face images used in the study were real faces not shown in the figure).

Performance measures

We measured the performance of the trained DCNNs on a face verification task using the standard Labeled Faces in the Wild (LFW) benchmark (Liao et al., 2014). LFW is a database of unconstrained face images used for testing performance level of face recognition algorithms. The data set contains more than 13,000 images of faces collected from the web, 1680 of the people pictured have two or more distinct photos in the data set. More information about this data base can be found here https://vis-www.cs.umass.edu/lfw/. We used 6000 pairs of face images. These pairs consist of positive pairs, where both images show the same person, and negative pairs, where the two images show different people. The goal of the face verification task is to determine if the two images in each pair belong to the same person or not. We assessed the models’ performance by measuring the cosine distance between the embeddings of pairs of faces. If the distance was smaller than a predetermined optimal threshold, the images were classified as the same person, otherwise they were classified as different. The accuracy values reported here reflect performance achieved using the optimal threshold for each model.

Extracting representations from DCNNs: To extract the representations that were generated by the DCNNs, we ran the trained models in evaluation mode on a predefined set of image stimuli. The face images were first aligned using the MTCNN face alignment algorithm (Xiang & Zhu, 2017). After alignment, the images were normalized with the standard ImageNet normalization (M = [0.5, 0.5, 0.5], SD = [0.5, 0.5, 0.5]). We first measured the pixel-based representations of all face images. We then examined the representations at the penultimate, fully connected (fc7) layer. This is the final hidden layer that generates the final representation that is transformed to the output probability layer.

Quantifying view-invariance of face-representations in DCNNs: We calculated the Euclidean distances between the penultimate layer (fc7) embeddings of the following pairs of faces: same identity faces—same view, same identity faces—quarter view, same identity faces—half view, and different identity faces—same view (as shown in Figure 2A) for 15 different identities. The face alignment procedure failed to detect four of the half-view faces, so we only had 11 face pairs in the frontal-half-view condition. The distance scores were normalized by dividing the measured distances by the maximal distance value in each run across all stimuli and conditions. This resulted in a normalized score that ranged from 0–1. These distance scores were calculated for each of the 64 DCNNs, as well as for a *pixel-based representation* based on the pixel values of the images and for the *identity-based representation* based on the values of the penultimate (last hidden – fc7) layer of the fully face-trained DCNN (Abudarham et al., 2021). Finally, to measure the similarity between each of the 64 trained DCNNs and the baseline models (*pixel-based* and *identity-based*), we calculated the Euclidian distance between the normalized mean distances of the four types of face pairs (dividing the distance of each pair by the sum of the distances of the four pairs) of each trained DCNN with each baseline models. Smaller distances indicate that the DCNN is more similar to the pixel model (Figure 2D) or the identity model (Figure 2E).

Measurement of sensitivity to critical features: We calculated the Euclidean distances between the representations of the following four conditions: Same, Non-Critical, Critical, and Different (See Figure 3A). Each condition includes 25 image pairs. Distances were calculated for a *pixel-based representation* based on the pixel values of the images and for the other representations based on the penultimate layer. We then performed the same analysis that is described in the previous section to measure the similarity of each of the 64 models with a pixel-based or an identity-based representations (see Supplementary Figure S2).

Results

Data reported in the results section can be found in this OSF link: https://osf.io/huzkp/?view_only=dfc6c75cc12b424d851794be43ca3f44. To assess the effect of the visual diet on the generation of a view-invariant representation of face identity, we trained a DCNN (VGG-16) with the following training diets: We created 64 subsets of face images, which included all possible combinations of two, five, 10, 50, 100, 200, 500, and 1000 identities and one, five, 10, 20, 50, 100, 200, and 300 images per identity. For models that are trained with a relatively smaller number of faces (1–100 identities models), we trained the model with thirty different sets of faces to avoid stimulus specific effects. The results were then averaged across the thirty models for each condition.

Effects of visual diet on face identity classification

We measured the performance of the DCNNs on a standard face verification task, Labeled Faces in the Wild (LFW) benchmark (Liao et al., 2014). For each of the networks we extracted the representation in the last hidden (penultimate) layer (fc7) and assessed performance on a same-different identity task (see methods). Figure 1 shows that accuracy improves for DCNNs trained on larger number of images. Both the number of different identities as well as the number of images per identity were needed to improve performance. Accuracy did not exceed 75% if the number of identities was below 10 for any number of images per identity (up to 300 images per identity) or if the number of images per identity was below five for any number of identities (up to 1000 identities) (See Supplementary Table S1 for report of performance level values that are shown in Figure 1). This suggests that identity face classification requires experience with images of different identities but also with different images of the same identity. This is further corroborated in an examination of a subset of the DNNs that are trained on the same total number of images ($n = 10{,}000$ images) but differ in the number of identities/number of images per identity, showing an overall similar level of performance (see Supplementary Figure S4A). We next assessed how this experience changes the representation from a view-specific to a view-invariant representation.

The emergence of a view-invariant representation

To evaluate whether a representation that is generated by a DCNN is view-specific or view-invariant, we

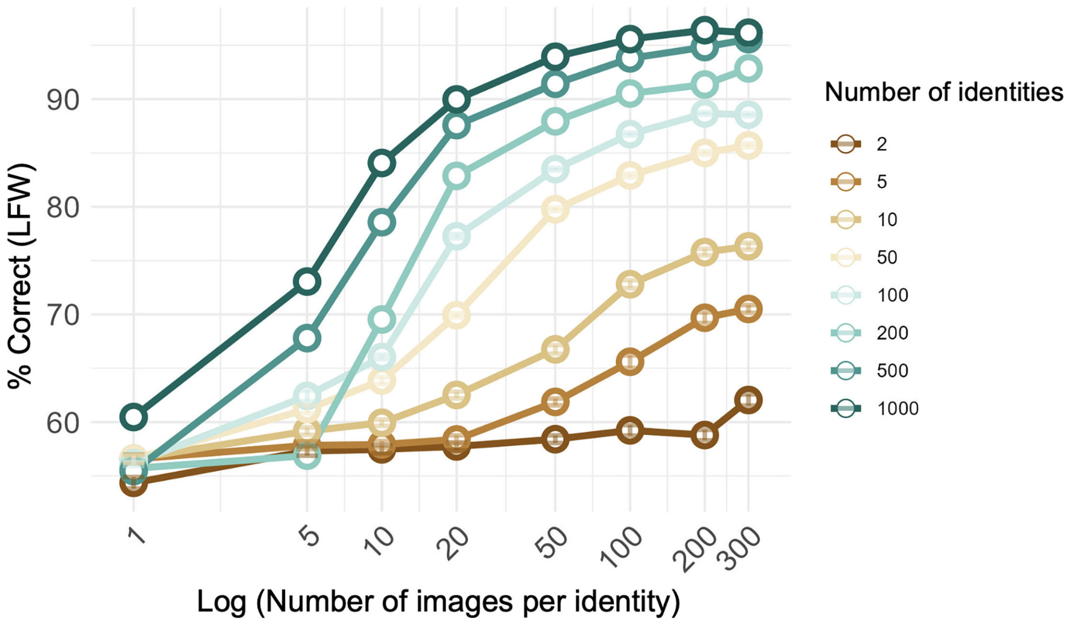

Figure 1. Accuracy on a face verification task with LFW benchmark. Performance gradually improves as the number of identities (color), and the number of images per identity (x-axis) increase. Chance level is 50%. Error bars indicate the standard error of the 30 samples that were tested for models that were trained on 100 identities or less.

used face images of 15 identities from which we generated the following four types of pairs: same identity: same view (frontal), same-identity: different view (frontal vs. quarter view), same-identity: different view (frontal versus half view), different identity-same view (frontal) (Figure 2A). We measured the Euclidean distance between the feature vectors of the four face pairs for two base-line representations: *Pixel-based representation*, which was the raw pixel values of the test face images. *Identity-based representation,* which was the representation in the penultimate layer of a fully face-trained DCNN (>8000 identities with approximately 300 images per identity, see methods). The distance between the pixel-based representations of each pair of faces showed a view-specific representation as indicated by a larger distance between same identity-different view face pairs (light blue bars) than between different identity-same view face pairs (red bar) (Figure 2B). As mentioned above, the face alignment procedure (see methods) failed to detect 4 of the half-view faces, for this reason we only included 11 face pairs in the half-view condition. The statistical analysis was therefore performed on these 11 identities across all conditions. A repeated measure analysis of variance (ANOVA) across the four face types reveal a significant effect of face type ($F(3,30) = 76.19$, $p < 0.001$). Post hoc comparisons reveal that all conditions were statistically different from one another ($p < 0.02$) (see Supplementary Table S2 for all statistical tests).

The distance between the representations of each pair of faces based on the penultimate layer of the fully-trained network revealed a view-invariant representation as manifested by a larger distance between different identity-same view face pairs than between same identity-different view face pairs (Figure 2C). A repeated measure ANOVA across the four face types reveal a significant effect of face type ($F(3,30) = 131.74$, $p < 0.001$). Post hoc comparisons reveal that all conditions were statistically different from one another ($p < 0.004$) (see Supplementary Table S2 for all statistical tests). On top of the larger distance between different identity same-view faces and same identity different view faces, we also see that the representation preserves the view-specific information, as evident by the larger distance between same identity faces that differ in head views in larger viewing than smaller viewing angles (the three blue bars). These findings are consistent with Hill and colleagues who proposed that the high-level identity representation of DCNNs preserves both identity and view information. (Hill et al., 2019).

Next, we measured the similarity between the representations generated for the same four face pairs in each of the trained models with the pixel-based and the identity-based representations. We did that by calculating the Euclidean distance between the distribution of the four types of face pairs of the *pixel-based* representation (Figure 2B) and *identity-based* representation (Figure 2C) with the distribution of the four types of face pairs in each of the 64 DCNN trained models (see Supplementary Figure S1). The similarity to the *pixel-based* distribution is presented in Figure 2D

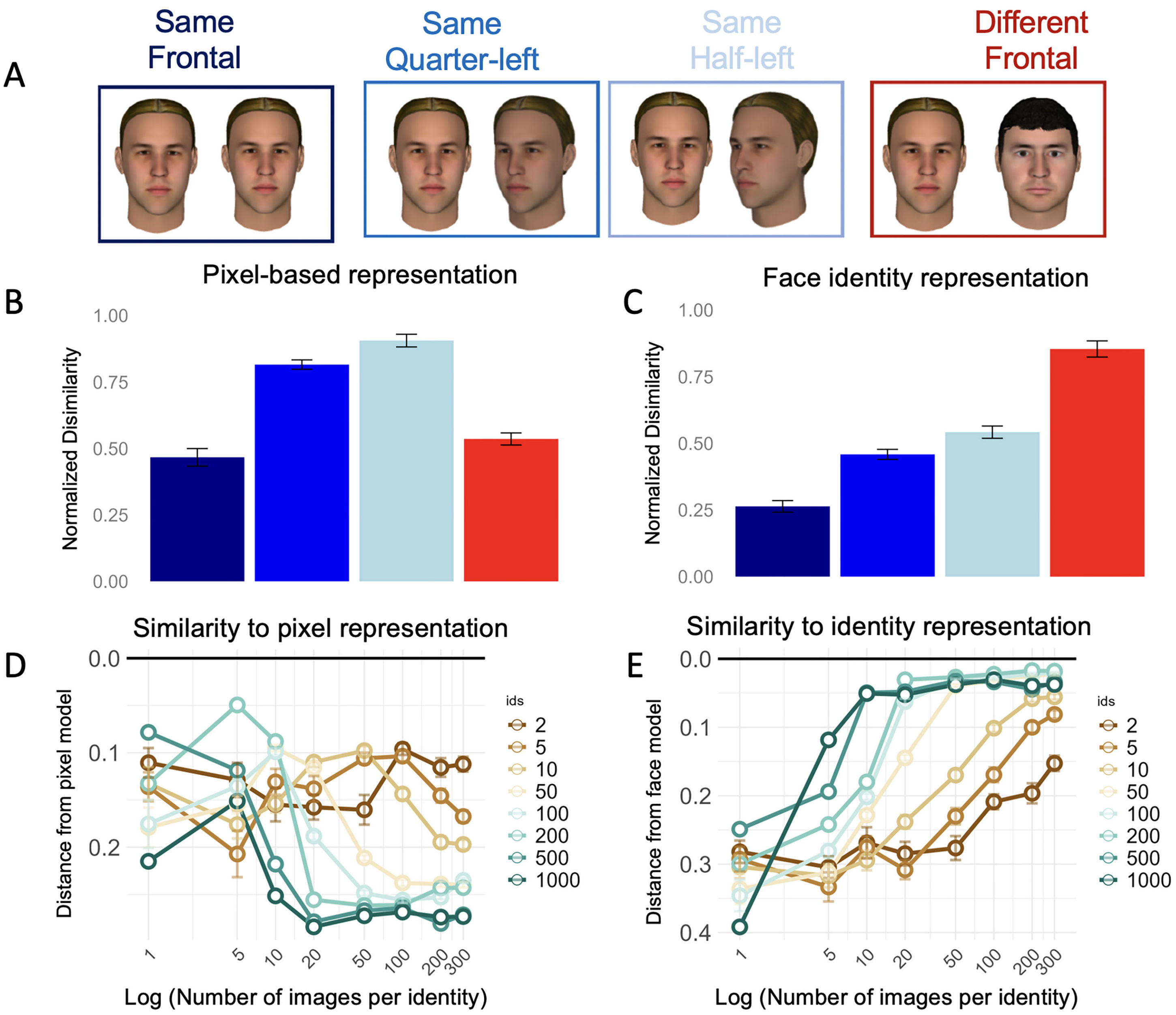

Figure 2. The emergence of a view-invariant face representation with visual experience. (**A**) The four types of face pairs used to test a view-invariant representation. Same identity same view (dark blue), same identity quarter-left (light blue), same identity half-left (lightest blue) and different identity same view (red). The DCNN was tested on real faces. The faces presented in the Figure are generated by a graphic software (FaceGen), to depict the type of headview changes that were used in the experiment. (**B**) Disimilarity scores (normalized distance) between the four types of face pairs based on the pixel layer reveals a view-specific representation. Error bars indicate the standard errors across images of different identities. (**C**) Disimilarity scores (normalized distance) between the four types of face pairs based on the penultimate layer of a fully face-trained DCNN reveals a view invariant representation. Error bars indicate the standard errors across images of different identities. (**D**) The similarity (measured by Euclidean distance where smaller values indicate higher similarity) of each DCNN with the pixel-based representations (panel B) is higher for DCNNs trained on smaller number of images. (**E**) The similarity (measured by Euclidean distance) of each DCNN with the identity-based representation (panel C) is higher for DCNNs trained on larger number of images. To see the representations of each DCNN see Supplementary Figure S1.

and to the *identity-based* distribution in Figure 2E. We found that DCNNs that are trained on smaller number of identities and images per identity generate representations that are more similar to a view-specific, image-based representation, and DCNNs that are trained on larger number of identities and images per identity generate a representation that is more similar to a view-invariant, identity-based representation. In particular, we see that a DCNN that is trained on large number of identities (500 or 1000) generate a view-invariant representation with only 10 images per identity. Examination of a subset of the DNNs that were trained on the same total number of images (10,000) but differed in the number of identities or number of images per identity show that both between identity and within identity variances contribute to the emergence of this view invariant representation (see Supplementary Figures S4B, S4C).

Sensitivity for human-like critical features

To evaluate whether the representations of DCNNs are sensitive to human-like view invariant critical features, we measured the distance between representations of four types of face pairs of 25 different identities (not included in the train set). Figure 3A shows an example of each type of face pairs: “Same identity” are different images of the same identity, “Non-critical features” are same identity face pairs in which noncritical features were replaced; “Critical features” are same identity face pairs in which critical features were replaced; and “Different identity” face pairs. Figure 3B shows the Euclidean distances between these face pairs based on their pixel-based representations. A repeated measure ANOVA reveal a significant effect of face type $F(3, 72) = 7.25$, $p = 0.001$. Post hoc comparisons revealed the pixel-based distances of same identity pairs were smaller than the distance between noncritical features pairs ($t(24) = 3.75$, $p = 0.002$), critical feature pairs ($t(24) = 4.18$, $p < 0.001$) and different identity pairs ($t(24) = 3.28$, $p = 0.006$). Importantly, there was no difference between the pixel-based distances of face pairs that differ in critical and non critical features (t(24) = 0.42, p = 0.68). These findings indicate that pixel information is not sensitive to human-like critical features more than noncritical features. Figure 3C shows the Euclidean distances between representations of the same face pairs, based on the penultimate layer of a fully face-trained DCNN (Abudarham et al., 2019; Abudarham et al., 2021). Here we see a much larger distance between faces that differ in critical features than faces that differ in noncritical features, indicating that the identity-based representation is sensitive to human-like critical features. We also show that faces that differ in critical features are as different as different identity faces, indicating that changing them is similar to changing the identity of a face. A repeated measure ANOVA across the four face types reveal a significant effect of face type $F(3, 75) = 175.51$, $p < 0.001$. Post hoc comparisons reveal that all conditions were statistically different from one another ($p < 0.001$) (see Supplementary Table S3 for all statistical tests), except face pairs that differ in critical features and different face pairs, which did not differ statistically ($t(24) = 0.684$, $p = 0.5$). These findings are consistent with our definition of critical features, which are features that changing them change the identity of the face.

Next, we measured the similarity of the representations to the four face pairs that were generated by each of the face-trained models with the pixel-based and identity-based representations. We calculated the Euclidean distance between the distribution of the four types of face pairs of the *pixel-based* representation (Figure 3B) and *identity-based* representation (Figure 3C) with the distribution of the four types of face pairs in each of the 64 DCNN trained models (see Supplementary Figure S2). The distances from the *pixel-based* distribution are presented in Figure 3D and from the *identity-based* distribution in Figure 3E. DCNNs that were trained on smaller number of images were more similar to the image-based representation showing no sensitivity to critical features over noncritical features, whereas DCNNs that were trained on a larger number of images were more similar to models that are sensitive to human-like critical features. Examination of a subset of the DNNs that were trained on the same total number of images (10,000) but differed in the number of identities or number of images per identity show that both between identity and within identity variances contribute to the sensitivity to critical features (see Supplementary Figures S4D, S4E).

Abudarham and Yovel (2016) suggested that humans are sensitive to critical features because they enable a view-invariant representation of face identity, which is needed for successful face recognition across different appearances of the same identity. To examine these correspondences, we computed sensitivity to noncritical features by subtraction of face pairs that differ in noncritical features from same identity faces (Figures 4A, 4B) and sensitivity to critical features by subtraction of face pairs that differ in critical features from same identity faces (Figures 4C, 4D). We then examined the strength of a linear relationship across all 64 DCNNs with accuracy on face verification based on the LFW benchmark (Figures 4A, 4C, y-axis) and with a measure that indicates a view invariant representation (Figures 4B, 4D, y-axis). The view invariant measure was calculated by subtracting the distance between different face pairs (red bar in Figure 2) from the distance between same identity different head view face pair (light blue in Figure 2). Thus a negative score indicates a view-specific representation and a positive score indicate a view-invariant representation.

Figure 4 shows a strong linear relationship between sensitivity to critical features and accuracy on the LFW benchmark (Figure 4C). It also shows a strong linear relationship between sensitivity to critical features and the emergence of a view-invariant representation (Figure 4D). There was no such linear relationship

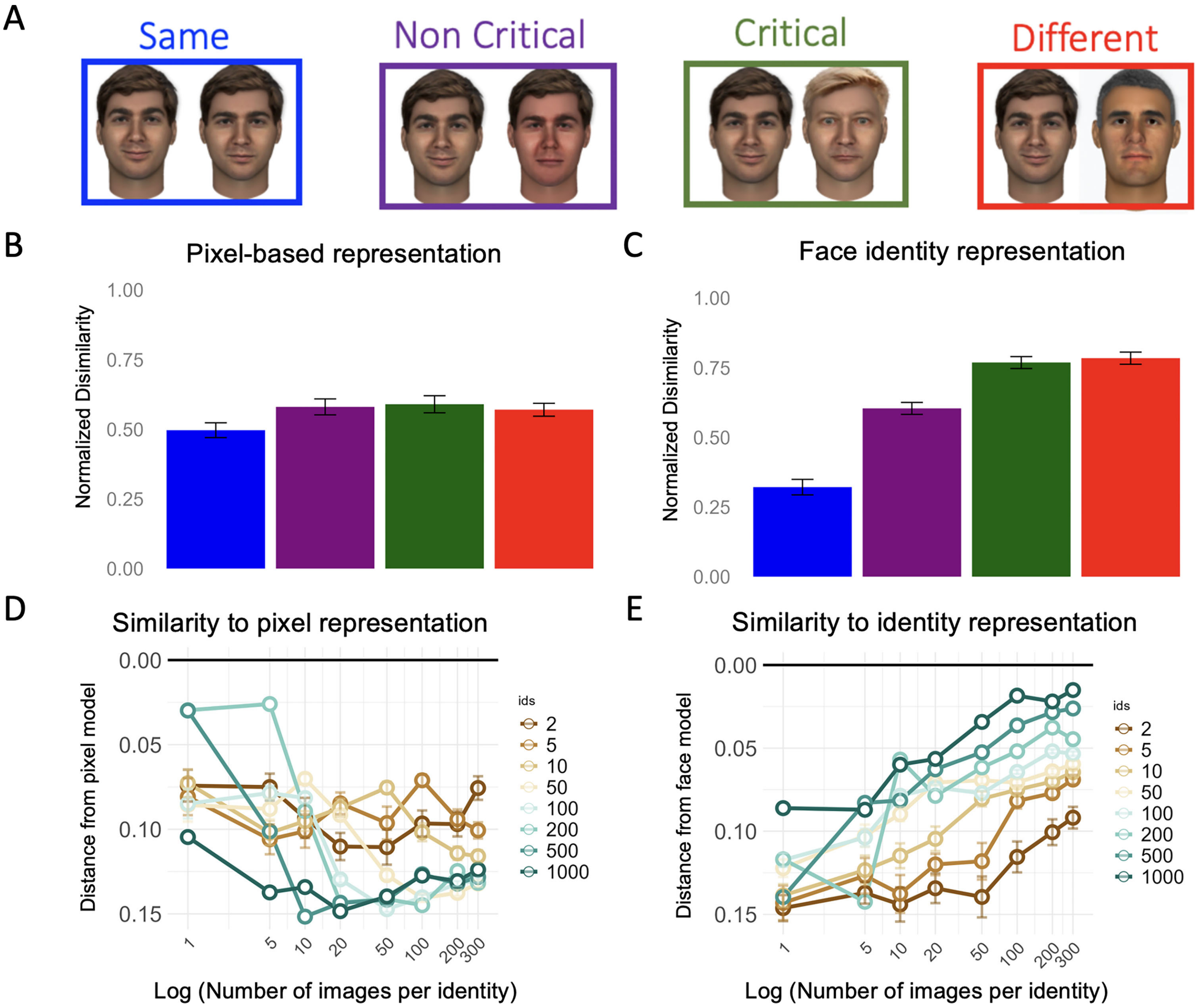

Figure 3. The emergence of sensitivity to critical features with visual experience: (**A**) The four types of face pairs that are used to test sensitivity for critical features. Same identity (blue), noncritical features changed (purple), critical features changed (green) and different identity (red). The DCNN was tested on real faces. The faces presented in the Figure were generated by a graphic software (FaceGen) to depict the feature manipulations that were performed on the real face images. (**B**) Disimilarity scores (normalized distance) between the four types of face pairs based on the pixel layer reveals no sensitivity to human-like critical features. Error bars indicate the standard errors across images of different identities. (**C**) Disimilarity scores (normalized distance) between the four types of face pairs based on the penultimate layer of a fully face-trained DCNN reveales high sensitivity to human-like critical features. Error bars indicate the standard errors across images of different identities. (**D**) The similarity (measured by Euclidean distance where lower values are higher similarity) of each DCNN to the pixel-based representations (panel B) is higher for DCNNs trained on smaller number of images. (**E**) The similarity (measured by Euclidean distance) of each DCNN to the identity-based representation (panel C) is higher for DCNNs trained on larger number of images. To see the representations of each DCNN see Supplementary Figure S2.

between sensitivity to noncritical features (distance between pairs of faces that differ in noncritical features) and performance on face verification task (Figure 4A) or the emergence of a view invariant representation (Figure 4B).

We calculated the linear relationship between sensitivity to noncritical features and sensitivity to critical features, with performance on LFW (Figures 4A, 4C) and the view invariant representation (Figures 4B, 4D). To compare the fit of the models, we

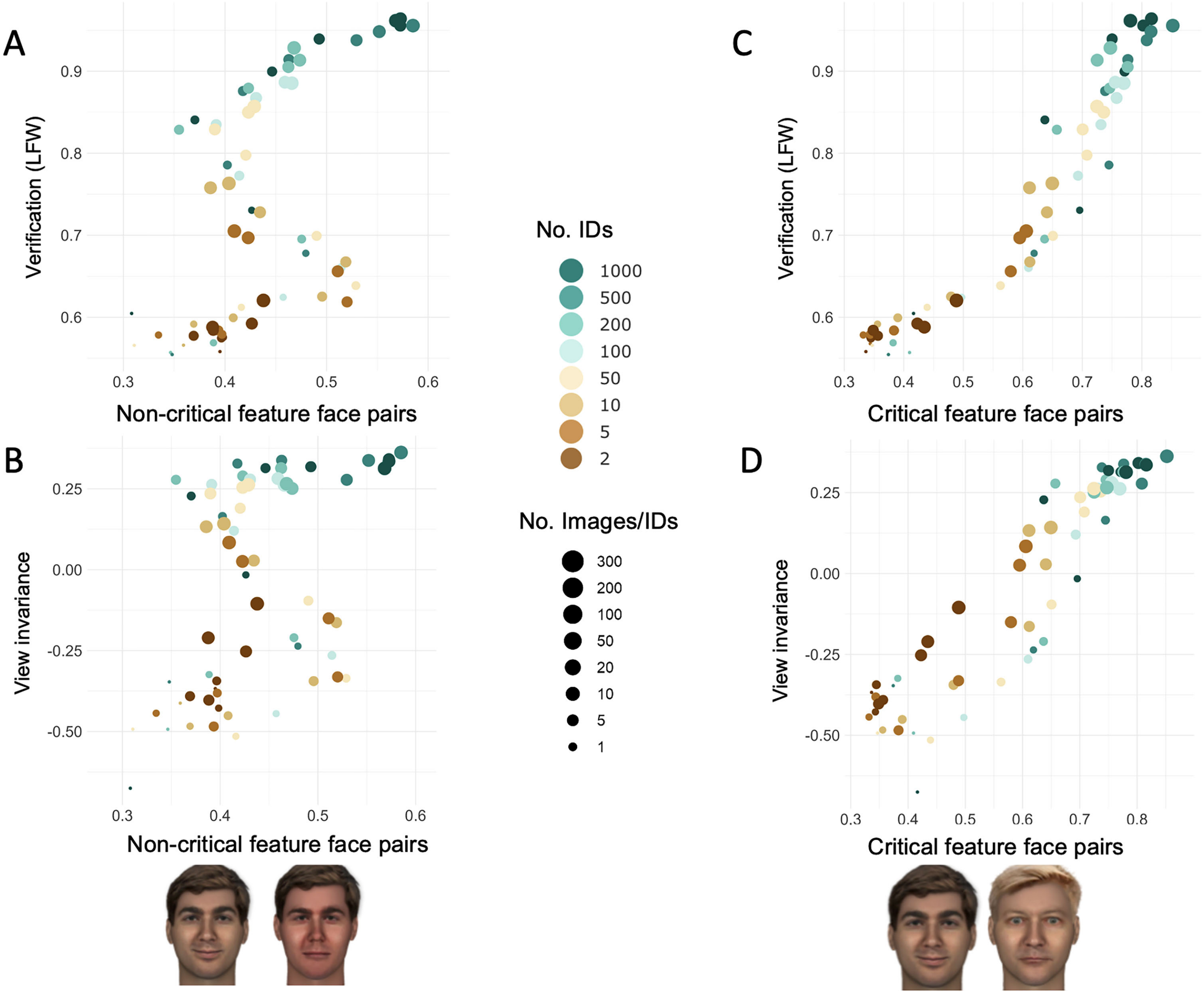

Figure 4. Sensitivity to human-like critical features is correlated with performance on face verification and the emergence of a view invariant representation: (**A**) Sensitivity to noncritical features, measured by the distance between same identity faces that differ in noncritical features, and performance on face verification task do not emerge concurrently as a function of experience. (**B**) Sensitivity to noncritical features does not emerge concurrently with a view-invariant representation as a function of experience. (**C**) Sensitivity to critical features, measured by the distance between same identity faces that differ in critical features, and performance on face verification task emerges concurrently as a function of experience. (**D**) Sensitivity to critical features emerges concurrently with a view-invariant representation as a function of experience.

computed the R^2 of each model and the variance of the residuals. As seen in our results, the linear relationship between the sensitivity to critical features (distance between pairs of faces the differ in critical features) with LFW and view-invariant is much stronger (larger variance and lower R^2) than the sensitivity to noncritical features (distance between pairs of faces that differ in noncritical features). Overall, these findings further support the relationship between the emergence of view invariance and performance on face identity task with the emergence of sensitivity to critical features but not with sensitivity to noncritical features. DCNNs that are sensitive to human-like critical features (larger distance between faces that differ in critical features) show better performance on a face verification task and a view invariant representation (bottom, right) (Table 1).

Predicted ~ predictor	R^2 adjusted	Variance (residuals)
LFW ~ Non Critical (Figure 4A)	0.3	0.69
View Invariant ~ Non Critical (Figure 4B)	0.21	0.77
LFW ~ Critical (Figure 4C)	0.89	0.1
View Invariant ~ Critical (Figure 4D)	0.86	0.14

Table 1. Results of a linear regression of the distance between face pairs that differ in critical and noncritical features and performance on LFW and the view-invariant representation.

Discussion

Successful face recognition depends on the ability to generalize across different images of the same identity and discriminate between images of different identities. The goal of the current study was to leverage the success of DCNNs in face recognition and their similarity to human-like representations (Abudarham et al., 2021; Blauch et al., 2021; Dobs et al., 2023), to examine whether success on a face verification task, the emergence of a view-invariant representation and sensitivity to human-like critical facial features emerge concurrently as a function of the amount and type of experience with faces. Our findings show that increasing both the number of images per identity and number of identities, concurrently improved verification accuracy, the emergence of a view-invariant representation and sensitivity to human-like critical facial features. These findings suggest a critical role for experience with faces in the generation of these representations.

For many years cognitive scientists and computer scientists have attempted to reveal the critical features that enable human-level face recognition performance. Despite the success of current machine learning algorithms to recognize faces at, or even above, human-level performance, it is still unknown which features are used by these algorithms to perform this task. Studies in humans revealed a subset of facial features for which humans showed high perceptual sensitivity. Furthermore, changing these features changed the identity of the face, indicating their importance for human face recognition (Abudarham et al., 2019). Abudarham and colleagues (2016) further suggested that these features enable a view-invariant representation, as they remain invariant across different head-views (Abudarham & Yovel, 2016). In the current study, we were able to link these two phenomena and their relationship with verification accuracy by showing that they emerge concurrently as a function of the amount of experience with faces during training (Figure 4, Supplementary Figure S4). Particularly, we found that increased sensitivity to changes in critical features was strongly associated with the emergence of view invariance (Figure 4C) and improved performance on a face verification task (Figure 4D) as a function of the amount of faces that DNNs were trained with. These findings suggest that these identity-based representations rely on similar features. Such a linear relationship was not found for sensitivity to changes in noncritical facial features (Figures 4A, 4B). These findings are consistent with the finding changes in noncritical features had a similar effect on the pixel-based and the identity-based representation, in contrast to the much larger distance that was found for changes in critical features in the identity relative to the pixel-based representation (Figures 3B, 3C). Taken together, sensitivity to critical features, a view-invariant representation and performance on face recognition emerge concurrently along the hierarchy of visual processing as well as with increased experience.

The relevance of these findings to human face recognition should be evaluated considering the nature of human experience with faces during development. Recent studies that have used head-mounted cameras on infants' foreheads during the first year of their life show that during this period, they were primarily exposed to three main identities from myriad of different appearances and head-view (Fausey, Jayaraman, & Smith, 2016). It is only later during development that the number of identities start increasing reaching a few thousands of familiar identities in adults (Jenkins, Dowsett, & Burton, 2018). Indeed, performance in face recognition improves slowly and requires several years to reach adult level performance (Matthews & Mondloch, 2022). To better learn about effects of human-like experience from face recognition algorithms, it is necessary to train the algorithms on a more human-like type of experience with faces, which is different from the training set and training protocols of current face recognition algorithms (Vong, Wang, Orhan, & Lake, 2024; Yoshida & Smith, 2008).

Another important difference between human and face recognition algorithms is that human face recognition primarily concerns the recognition of familiar faces (Burton, Bruce, & Hancock, 1999; Young & Burton, 2018), whereas face recognition algorithms are trained to classify untrained (unfamiliar) identities. In the current study, we used unfamiliar faces to test the representation of face recognition algorithms and learn about their ability to generalize to unlearned examples. However, if the goal of the human face recognition system is to only classify socially relevant familiar identities, computer algorithms that aim to model human face recognition should take this consideration into account. Given that an important aspect of familiar face recognition is their semantic representations, the development of familiar face recognition may be better modeled with multi-modal visual-semantic algorithms (Shoham, Grosbard, Patashnik, Cohen-Or, & Yovel, 2024; Vong et al., 2024).

Recent studies that examined human-like representations in DCNNs revealed other similarities between humans and DCNNs. This includes a much larger drop in performance for inverted than upright faces than the drop that is found for objects (Dobs et al., 2023; Yovel et al., 2023). The Thatcher illusion in which distorted faces look more similar to normal faces when they are inverted than upright was also found in face-trained but not object-trained DCNNs (Jacob, Pramod, Katti, & Arun, 2021). A drop in performance for the race of faces that the algorithms was not trained on (i.e. lower performance for Asian faces in a DCNN trained on White faces) is also typically found in DCNNs similar to the human other race effect (Dobs et al., 2023; O'Toole & Castillo, 2021). The approach that we used in the current study enables us now to ask what kind of experience is required for these human-like representations to emerge.

In summary, recent advances in machine learning that enable face recognition algorithms to reach human-level performance, and the similarity between the representations generated by humans and machines (Abudarham et al., 2021; Dobs et al., 2023; Jacob et al., 2021), offer us new computational tools to explore the factors that mediate human face recognition. Future studies will further investigate the contribution of more specific characteristics of face images, such as their pose, expression, and lighting, to the generation of a view-invariant representation and sensitivity to view-invariant human-like critical features.

Keywords: face recognition, deep learning, experience, view invariance, critical features

Acknowledgments

Supported by an ISF grant 917/21 to GY.

Commercial relationships: none.
Correspondence author: Galit Yovel.
Email: gality@tauex.tau.ac.il.
Address: School of Psychological Sciences, Sagol School of Neuroscience, Tel Aviv University, Tel Aviv 6139001, Israel.

*MR and NG contributed equally to this work.

References

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.

Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision, 16*(3), 40, https://doi.org/10.1167/16.3.40.

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*(9), e13031, https://doi.org/10.1101/2020.01.01.890277.

Baker, K. A., & Mondloch, C. J. (2019). Two Sides of Face Learning: Improving Between-Identity Discrimination While Tolerating More Within-Person Variability in Appearance. *Perception, 48*(11), 1124–1145, https://doi.org/10.1177/0301006619867862.

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.

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999). From Pixels to People: A Model of Familiar Face Recognition. *Cognitive Science, 23*(1), 1–31, https://doi.org/10.1207/s15516709cog2301_1.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 67–74, https://doi.org/10.1109/FG.2018.00020.

Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences, 23*(4), 305–317, https://doi.org/10.1016/j.tics.2019.01.009.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333–341, https://doi.org/10.1016/j.tics.2007.06.010.

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2023). Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences, 120*(32), e2220642120, https://doi.org/10.1073/pnas.2220642120.

Fausey, C. M., Jayaraman, S., & Smith, L. B. (2016). From faces to hands: Changing visual input in the first two years. *Cognition, 152*, 101–107, https://doi.org/10.1016/j.cognition.2016.03.005.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife, 7*, e32962, https://doi.org/10.7554/eLife.32962.

Hill, M. Q., Parde, C. J., Castillo, C. D., Colón, Y. I., Ranjan, R., Chen, J.-C., . . . O'Toole, A. J. (2019).

Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence, 1*(11), 522–529, https://doi.org/10.1038/s42256-019-0111-7.

Honig, T., Shoham, A., & Yovel, G. (2022). Perceptual similarity modulates effects of learning from variability on face recognition. *Vision Research, 201*, 108128, https://doi.org/10.1016/j.visres.2022.108128.

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*(1), 1872, https://doi.org/10.1038/s41467-021-22078-3.

Jenkins, R., Dowsett, A. J., & Burton, A. M. (2018). How many faces do people know? *Proceedings of the Royal Society B: Biological Sciences, 285*(1888), 20181319, https://doi.org/10.1098/rspb.2018.1319.

Kramer, R. S., Jenkins, R., Young, A. W., & Burton, A. M. (2017). Natural variability is essential to learning new faces. *Visual Cognition, 25*(4–6), 470–476.

Laurence, S., Zhou, X., & Mondloch, C. J. (2016). The flip side of the other-race coin: They all look *different* to me. *British Journal of Psychology, 107*(2), 374–388, https://doi.org/10.1111/bjop.12147.

Liao, S., Zhen, L., Dong, Y., & Li, S. Z. (2014). A benchmark study of large-scale unconstrained face recognition. *IEEE International Joint Conference on Biometrics,* 1–8, https://doi.org/10.1109/BTAS.2014.6996301.

Matthews, C. M., & Mondloch, C. J. (2022). Learning faces from variability: Four- and five-year-olds differ from older children and adults. *Journal of Experimental Child Psychology, 213*, 105259, https://doi.org/10.1016/j.jecp.2021.105259.

McKone, E., Wan, L., Pidcock, M., Crookes, K., Reynolds, K., Dawel, A., . . . Fiorentini, C. (2019). A critical period for faces: Other-race face recognition is improved by childhood but not adult social contact. *Scientific Reports, 9*(1), 12820, https://doi.org/10.1038/s41598-019-49202-0.

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.

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences, 26*(6), 462–483, https://doi.org/10.1016/j.tics.2022.03.007.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology, 70*(5), 897–905.

Shoham, A., Grosbard, I., Patashnik, O., Cohen-Or, D., & Yovel, G. (2024). Using Deep learning algorithms to disentangle visual and semantic information in human perception and memory. *Nature Human Behaviour, 8*(4).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint arXiv:1409.1556*.

Tanaka, J. W., Heptonstall, B., & Hagen, S. (2017). Perceptual expertise and the plasticity of other-race face recognition. In *Face Recognition* (pp. 121–139). Oxfordshire: Routledge.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science, 383*(6682), 504–511, https://doi.org/10.1126/science.adi1374.

Yoshida, H., & Smith, L. B. (2008). What's in view for toddlers? Using a head camera to study visual experience. *Infancy, 13*(3), 229–248, https://doi.org/10.1080/15250000802004437.

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences, 22*(2), 100–110, https://doi.org/10.1016/j.tics.2017.11.007.

Yovel, G., Grosbard, I., & Abudarham, N. (2023). Deep learning models challenge the prevailing assumption that face-like effects for objects of expertise support domain-general mechanisms. *Proceedings of the Royal Society B: Biological Sciences, 290*(1998), 20230093, https://doi.org/10.1098/rspb.2023.0093.

Xiang, J., & Zhu, G. (2017). Joint face detection and facial expression recognition with MTCNN. In *2017 4th International Conference on Information Science and Control Engineering (ICISCE)* (pp. 424–427). IEEE.