Emerged human-like facial expression representation in a deep convolutional neural network

Liqin Zhou^a, Ming Meng^{a,b,*}, Ke Zhou^{a,*}

^a Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing 100875, China

^b Key Laboratory of Brain, Cognition and Education Sciences, Ministry of Education, China; School of Psychology, Center for Studies of Psychological Application, and Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou 510631, China

* Corresponding authors:

Ming Meng

School of Psychology, South China Normal University

E-mail: mingmeng@m.scnu.edu.cn

&

Ke Zhou

School of Psychology, Beijing Normal University

E-mail: kzhou@bnu.edu.cn

Abstract

Face identity and expression play critical roles in social communication. Recent research found that the deep convolutional neural networks (DCNNs) trained to recognize facial identities spontaneously learn features that support facial expression recognition, and vice versa, suggesting an integrated representation of facial identity and expression. In the present study, we found that the expression-selective units spontaneously emerged in a VGG-Face trained for facial identity recognition and tuned to distinct basic expressions. Importantly, they exhibited typical hallmarks of human expression perception, i.e., the facial expression confusion effect and categorical perception effect. We then investigated whether the emergence of expression-selective units is attributed to either face-specific experience or domain-general processing, by carrying out the same analysis on a VGG-16 trained for object classification and an untrained VGG-Face without any visual experience, both of them having the identical architecture with the pretrained VGG-Face. Although Similar expression-selective units were found in both DCNNs, they did not exhibit reliable human-like characteristics of facial expression perception. Taken together, our computational findings revealed the necessity of domain-specific visual experience of face identity for the development of facial expression perception, highlighting the contribution of nurture to form human-like facial expression perception. Beyond the weak equivalence between human and DCNNs at the input-output behavior, emerging simulated algorithms between models and humans could be established through domain-specific experience.

Keywords

Facial expression, Face identity, Domain-specific experience, Expression confusion, Categorical perception

Introduction

Facial identity and expression play important roles in daily life and social communication. When interacting with others, we can not only easily recognize who they are through their facial identity information, but also access their emotions from their facial expressions. An influential early model proposed that face identity and expression were processed separately via parallel pathways [1,2]. Configural information for encoding face identity and expression was different from each other [3]. Findings from several neuropsychological studies supported this view. Patients with impaired facial expression recognition still retained the ability to recognize famous faces [4,5], whereas patients with prosopagnosia (an inability to recognize the identity of others from their faces) could still recognize facial expressions [4–6]. Haxby et al. (2000) further proposed a distributed neural system for face perception, which emphasized a distinction between the representation of invariant aspects (e.g., identity) and changeable aspects (e.g., expression) of faces. According to this model, in the core system, lateral inferior occipitotemporal cortex (i.e., fusiform face area, FFA; occipital face area, OFA) and superior temporal sulcus (STS) may contribute to the recognition of facial identity and expression, respectively [8,9]. Patients with OFA/FFA damage have deficits for face identity recognition, and those with damage to the posterior STS (pSTS) suffer impairments in discriminating expression [10].

On the other hand, processing mechanisms of the human visual system for facial identity and expression recognition normally share face stimuli as inputs. That is, naturally, a face may contain both identity and expression information. Early visual processing of the same face stimuli would be the same for both identity and expression recognition, but it is unclear at what stage they may start to split. Amid increasing evidence to suggest an interdependence or interaction between face identity and expression processing [11–15], we hypothesize that any computational model that simulates human performance for facial identity and expression recognition must share common inputs for training. Moreover, if domain-specific face input is necessary to train a computational model that simulates human performance for facial identity and expression recognition, it would suggest that the split of identity and expression processing might occur at or after the domain-general visual processing stages. However, if no training or no domain-specific training of face inputs were needed for a computational model that simulates human performance, it would

suggest a dissociation between identity and expression processing at domain-general stages of visual processing.

Specifically, DCNNs has achieved human-level performance in object recognition of natural images. Investigations combining DCNNs with cognitive neuroscience further discovered similar functional properties between artificial and biological systems. For instance, there is a trend of high similarity between the hierarchy of DCNNs and primate ventral visual pathways [16,17]. Research relevant to this study revealed a similarity of activation patterns between face identity-pretrained DCNNs and human FFA/OFA [18]. Thus, DCNNs could be a useful model simulating the processes of biological neural systems. More recently, several seminal studies have found that the DCNNs trained for the recognition of facial expression spontaneously developed facial identity recognition ability, and vice versa, suggesting that integrated representations of identity and expression may arise naturally within neural networks, like humans do [19,20]. However, a recent study found that face identity-selective units could spontaneously emerge in a untrained DCNN [21], which seemed to cast a strong doubt on the role of nurture in developing face perception, as well as the abovementioned speculation. Indeed, when adopting a computational approach to examine the human cognitive function, a success in classifying different expressions only suggests the weak equivalence between DCNNs and humans at the input-output behavior in Marr's three-level framework, which does not necessarily mean that DCNNs and humans adopt the similar representational mechanisms (i.e., algorithms) to achieve a same computational goal [22]. Therefore, to explore whether a common mechanism may be shared by both artificial and biological intelligent systems, a much stronger equivalence should be tested by establishing additional relationships between models and humans, i.e., identities in algorithms between them [23].

Therefore, in the present study, we borrowed the cognitive approaches developed in human studies to explore whether the human-like facial expression recognition relied on face identity recognition, by using the VGG-Face, a typical DCNN pretrained for the face identity recognition task (hereafter referred to as pretrained VGG-Face). The pretrained VGG-Face was chosen because of its relatively simple architecture and evidence supporting its similar representations of face identity to those in the human ventral pathway [18]. The training process of VGG-Face has already determined units' selectivity for various features to optimize the network's

face identity recognition performance. If the pretrained VGG-Face could simulate the interdependence between facial identity and expression in the human brain, then it should spontaneously generate expression-selective units. The selective units should also be able to predict the expressions of new face images. However, as mentioned above, having an ability to correctly classify different expressions does not necessarily mean human-like perception of expressions. Here, we introduced morphed expression continua to test whether these units perceived morphed expression categorically in a human-like way.

Then, in order to answer what human-like expression perception depends on, we introduced two additional DCNNs. The first one is the VGG-16, a DCNN that has an almost identical architecture with the pretrained VGG-Face but was trained only for natural object classification. The other one is an untrained VGG-Face, which has identical architecture to the pretrained VGG-Face, but its weights are randomly assigned with no training (hereafter referred to as untrained VGG-Face). Comparisons among the three DCNNs would clarify whether the human-like expression perception relies on face (identity) recognition specific experience, or general object recognition experience, or merely the architecture of the network.

Materials and methods

Neural network models

The VGG-Face (https://www.robots.ox.ac.uk/~vgg/software/vgg_face/)[24] pretrained for recognizing 2622 face identities on a database with 2.6 million face images was used. It achieved state-of-the-art performance whilst requiring less data than other state-of-the-art models (DeepFace and FaceNet). The network consists of thirteen convolutional (conv) layers and three fully connected (FC) layers. All these sixteen layers are followed by a rectification linear unit (ReLU). The thirteen convolutional layers are distributed to five blocks. Each of the first two blocks consists of two consecutive convolutional layers followed by max pooling. Each of the latter three blocks consists of three consecutive layers followed by max pooling.

Besides, we used another two DCNNs (VGG-16 and untrained VGG-Face) for comparisons. The VGG-16 was trained for classifying 1000 object categories using the ILSVRC-2014 ImageNet database, which contains over 14 million natural visual images (<https://arxiv.org/abs/1409.1556>)[25]. It achieves a 92.7% top-5 test accuracy

in ImageNet and thus is one of the best models submitted to the ILSVRC-2014. The architecture of the VGG-16 is identical to the pretrained VGG-Face except that the last FC layer includes 1000 units for 1000 object classes rather than 2622 units for facial identities in the pretrained VGG-Face. The untrained VGG-Face preserved the fully identical architecture of the pretrained VGG-Face while randomly assigning the connective weights (Xavier normal initialization)[18,26] without any training experience.

Stimuli

Three stimulus sets were used in the study. Stimulus set 1 was used to detect expression-selective units in the DCNNs. It contained 624 facial expression images: 104 identities, each with 6 basic expressions (anger, disgust, fear, happiness, sadness, and surprise) [27,28]. In the stimulus set 1, seventy identities were from Karolinska Directed Emotional Faces (KDEF) database [29], and the other 34 identities from the NimStim Database [30]. Then, to validate the reliability of expression recognition capability of the expression-selective units, a second stimulus set — the Radboud Faces Database (RaFD) [31] was applied. The front view images of all the 67 identities in the RaFD were used, each identity contained the aforementioned six expressions. Furthermore, we applied AffectNet Database as the third stimulus set to test if the units could recognize facial expressions in real life stimuli. The AffectNet Database is by far the largest database of facial expressions in the wild, and its images were collected from the Internet [32]. Four thousand eight hundred manually annotated images from the AffectNet Database were used, and each expression included 800 images. In this dataset, the face identities across expressions are different.

For each dataset, the luminance and contrast of the images were matched by using the SHINE toolbox [33], and the face part were reserved with the background removed by using the facemorpher package (https://alyssaq.github.io/face_morpher/facemorpher.html). Then, the images were resized to 224×224 pixels.

The prototypic expressions used in the morphed expression discrimination task were from the stimulus set 1. All identities in the stimulus set 1 were used. Seven morph continua were tested in the present study, including happiness-anger, happiness-fear, anger-disgust, happiness-sadness, anger-fear, disgust-fear, and

disgust-sad. The number of morphed levels in each morph continuum was 201. The morphing process was conducted by using the facemorpher package.

Analysis of network units

Each DCNN was presented with the stimulus set 1 and the responses of the units in the final layer of the feature extraction network (conv5-3) were extracted to be analyzed. Like Nasr et al. (2019), a two-way non-repeated ANOVA with expression (6 facial expressions) and identity (104 identities) as factors was conducted to identify expression-selective units. The “expression-selective units” were referred to as those that exhibited a significant main effect of expression ($p \leq 0.01$) but no significant effect of identity ($p > 0.01$). For each expression-selective unit, the responses were normalized across all images in the stimulus set 1. After that, its tuning value for each expression was calculated by taking the difference between the average response to all images within the same expression and the average response to all images in the stimulus set, and then dividing the difference by the standard deviation of the responses across all images in the stimulus set [35]:

$$TV_i^k = \frac{\frac{1}{P_k} \sum_{p \in k} A_i^p - \frac{1}{P} \sum_{p=1}^P A_i^p}{\sqrt{\frac{1}{P} \sum_{p=1}^P (A_i^p - \frac{1}{P} \sum_{p=1}^P A_i^p)^2}}$$

Where TV_i^k is the tuning value of unit i to expression k . A_i^p is the normalized response of unit i to image p . P_k is the number of images that are labeled as expression k . P is the number of all images in the database. The tuning value reflects the extent to which a unit activates preferentially to images of a specific expression. For each unit, the expression with the highest tuning value is defined as its preferred expression.

In each DCNN, to test the reliability and generality of the expression recognition ability of expression-selective units, the SVC model trained on the stimulus set 1 was used to predict the expressions of images from the RaFD and AffectNet Database, separately. In the SVC model, the first six hundred principal components of the responses of expression-selective units were used. The prediction accuracy indicated the extent to which expression-selective units could classify facial expressions. Further, predicted expressions and true expressions of the images were used to construct the confusion matrix. And this confusion matrix was correlated with that of humans to measure how similar the expression recognition performance of

expression-selective units and humans were.

Morphed expression discrimination task

To test whether expression-selective units exhibit a human-like categorical perception of morphed facial expressions, we designed a morphed expression discrimination task that was comparable to the ABX discrimination task designed for human beings [36–38]. Taking happiness-anger continuum for example, the expression-selective units whose preferred expression was happiness or anger were selected to perform the task. At first, a binary SVC model was trained on the prototypic expressions (happy and angry expressions of all 104 identities in the stimulus set 1), then the trained SVC model was used to predict the expressions of morphed expression images (the middle 199 morph levels besides the two prototypic expressions). For each morph level, the identification frequency of anger was defined to be the network’s identification rate at the current expression morph level.

To quantitatively characterize the shape of the identification curve, we fitted Linear function, Quadratic function (poly2), and Logistic function to the curve, respectively. If the network perceived the morphed expressions like a human, the identification curve should be nonlinear and should show an abrupt category boundary. Thus, the goodness of fit (R^2) of the Logistic function (S-shaped) to identification curves should be the best.

Comparisons between different DCNNs

To test the dependence of human-like expression perception of expression-selective units on face identity recognition experience, we also introduced VGG-16 and untrained VGG-Face as controls. The VGG-16 is trained for natural object classification, and the untrained VGG-Face has no training experience. First, the expression classification performances of expression-selective units in different DCNNs were compared to explore whether these units in the pretrained VGG-Face recognized expressions better than the VGG-16 and untrained VGG-Face. Then, we assessed the differences of categorical perception of morphed expressions among the DCNNs by respectively comparing their goodness of fit (R^2) of the Logistic function and goodness of fit (R^2) of the Linear function. Finally, the Mann–Whitney U test was used for statistically evaluating the differences among the DCNNs.

256

257 **Results**

258 **1. Expression-selective units spontaneously emerge in the pretrained** 259 **VGG-Face.**

260 We first explored whether expression-selective units could spontaneously emerge
261 in the pretrained VGG-Face. Note that, the pretrained VGG-Face was trained for
262 recognizing 2622 identities with over two million face images [24]. It consists of
263 thirteen convolutional (conv) layers and three fully connected (FC) layers (Fig. 1A;
264 details in Methods). The first thirteen convolutional layers form a feature extraction
265 network which transforms images to a high-level representation, and the following
266 three fully-connected layers make up a classification network that classifies images
267 using the high-level representation [34].

268 Here, we tested the expression selectivity of each unit in the final layer of the
269 feature extraction network (conv5-3) with stimulus set 1 (details in Methods). The
270 stimulus set 1 is composed of 104 different facial identities selected from the
271 Karolinska Directed Emotional Faces (KDEF)[29] and NimStim [30] databases, and
272 each identity includes 6 kinds of basic expressions (i.e., anger, disgust, fear, happiness,
273 sadness, and surprise)[27,28]. Specifically, all 624 images in the stimulus set 1 were
274 presented to the pretrained VGG-Face and their activations in the conv5-3 layer were
275 extracted. First, we conducted a two-way non-repeated analysis of variance (ANOVA)
276 with identity and expression as factors to detect units selective to facial expression (p
277 ≤ 0.01) but not to face identity ($p > 0.01$). 1259 units out of the total 100,352 units
278 (1.25%) in the conv5-3 layer were founded to have expression selectivity. Then, for
279 each expression-selective unit, its tuning value [35] for each expression category was
280 calculated to measure whether and to what extent it preferred specific expression. As
281 shown in Fig. 1B, almost all units responded selectively to only one specific
282 expression and exhibited a tuning effect. Finally, to test whether the responses of
283 these expression-selective units provide sufficient information for successful
284 expression recognition, we performed principal components analysis (PCA) on these
285 unit activations to all images and selected the first six hundred principal components
286 to perform an expression classification task, using a support vector classification
287 (SVC) analysis with 104-fold cross-validation. The classification accuracy (mean $\pm$
288 standard error, 76.76% $\pm$ 1.59%) was much higher than the chance level (16.67%) and

much higher than the accuracy with randomly shuffled expression labels ($p = 1.8 \times 10^{-35}$, Mann–Whitney U test)(Fig. 1C). The results indicated that the expression-selective units spontaneously emerged in the pretrained VGG-Face that was pretrained for face identity recognition, which echoed previous findings [19,20].

2. Human-like expression confusion effect of the expression-selective units in the pretrained VGG-Face.

To examine the reliability of the expression-selective units, we used the classification model trained by stimulus set 1 to predict the expressions of images selected from the Radboud Faces Database (RaFD)[31]. The RaFD is an independent facial expression database including 67 face identities with different head and gaze directions, and only the front view images of each identity and each expression were used in the present study. The classification accuracy of the expressions from the RaFD was significantly higher than the chance level (accuracy = 67.91%, 95% confidential interval (CI) = [63.18%, 72.39%], image bootstrap with 10,000 replications)(Fig. 2A). It thus indicated that the expression-selective units in the pretrained VGG-Face had a reliable and general expression discriminability.

Subsequently, to test whether the response profiles of these units were like the human behavior, we calculated a confusion matrix of the response patterns of these units to each expression of the RaFD (Fig.2B) and correlated it with the behavioral confusion matrix of facial expression recognition from one previous human study [31], which also used expression stimuli from the RaFD (Fig. 2C). We found that the two confusion matrices were highly correlated ($r = 0.91$, $p = 1.6 \times 10^{-14}$). The similarity in the confusing effects of facial expressions between VGG-Face and humans was further confirmed by using the behavioral confusion matrix from another human study [30], whose expression stimuli were selected from the NimStim database ($r = 0.92$, $p = 4.1 \times 10^{-15}$)(Fig. 2D). In all three confusion matrices, the true label of each expression category was the most frequently predicted label and consistent across matrices, and the confusion pattern in each matrix was almost coincided with each other. For example, like humans, fear and surprise may be confused with each other, and disgust was mistaken for anger more often than for other expressions. However, we found that these units also confused sadness with anger and fear, which was different from those in the human studies. Overall, the results suggested a similar

expression confusion effect between the expression-selective units in the pretrained VGG-Face and humans.

3. Ecological validity of expression selectivity emerged in the pretrained VGG-Face.

The facial expressions in the stimulus set 1 and RaFD were collected from the same actors and/or with detailed instructions to express different emotions, which thus had limited ecological validity. If the expression-selective units can recognize expressions, they should also be able to recognize the real-life facial expressions with ecological validity. To verify this, we performed the same analysis on expression images selected from the AffectNet database, which is a large real-world facial expression database [32]. We selected 4800 images with manually annotated expressions, 800 for each basic expression. Note that, in the AffectNet database, the face identities across expressions are distinct. Again, we found that these expression-selective units can classify facial expressions much better than chance level accuracy = 29.52%, 95% CI = [28.27%, 30.79%], image bootstrap with 10,000 replications)(Fig. 4A). Besides, their confusion matrix for facial expressions selected from the AffectNet database also showed significant similarities with those from humans (for RaFD, $r = 0.62$, $p = 5.9 \times 10^{-5}$; for NimStim, $r = 0.67$, $p = 7.2 \times 10^{-6}$)(Fig. 4B). The reliable human-like confusion effect of facial expression recognition suggested that the expression-selective units in the pretrained VGG-Face can recognize facial expressions in a way as humans do.

4. The expression-selective units in the pretrained VGG-Face show human-like categorical perception for morphed facial expressions.

One may argue that the similarity in the expression confusion effect does not necessarily mean expression-selective units perceive expressions in a human-like way. It might result from the similarities in physical properties of the expression images, since the image-based PCA analysis (i.e., principal components based on pixel intensities and shapes) could also yield a confusion matrix similar to that of humans [39]. Therefore, to further confirm whether these units could behave a human-like psychophysical response function to facial expressions, we tested whether their responses exhibited categorical effect of facial expression perception by using

morphed expression continuum. Considering the generality of the categorical emotion perception in humans (Fig. 3A), we systematically tested the categorical effect in seven expression continua including happiness-anger, happiness-fear, anger-disgust, happiness-sadness, anger-fear, disgust-fear, and disgust-sadness. All of them have been tested in humans [36–38,40,41]. In detail, we designed a morphed expression discrimination task (Fig. 3B) that was comparable to the ABX discrimination task designed for humans [36–38]. The prototypic expressions were selected from the stimulus set 1. For each expression continuum, images of the two prototypic expressions were used to train an SVC model, and then the trained SVC model was applied to identify expressions of the morphed images. At each morph level of the continuum, the identification frequency of one out of the two expressions was defined as the units' identification rate at the current morph level. We hypothesized that if the selective units perceive expressions like humans, i.e., showing categorical effect, then the identification curves should be S-shaped.

As predicted, for all continua, the identification curves of the selective units were S-shaped (Fig. 3C). To quantify this effect, we fitted Linear, Quadratic (Poly2) and Logistic functions to each identification curve, respectively. If the units exhibited a human-like categorical effect, the goodness of fit (R^2) of the Logistic function to the curves should be the best. Otherwise, the goodness of fit of the Linear function to the curves should be the best if the units' response followed the physical changes in images. As illustrated in Fig. 3D and 3E, we found that all seven identification curves showed typical S-like patterns (Logistic vs. Linear: $p = 0.002$; Logistic vs. Poly2: $p = 0.002$, Mann–Whitney U test).

5. The human-like expression perception only spontaneously emerges in the DCNN with domain-specific experience (pretrained VGG-Face), but not in those with domain-general visual experience (VGG-16) or without any visual experience (untrained VGG-Face).

So far, we had demonstrated that the human-like perception of expression could spontaneously emerge in the DCNN pretrained for face identity recognition. However, how did these express-selective units achieve human-like expression perception? Specifically, it was still unknown whether the spontaneous emergence of the human-like perception of expression depended on the domain-specific experience

(e.g., face-related visual experience), a general natural object recognition experience, or even only the architecture of the DCNN.

To address this question, we introduced two additional DCNNs: VGG-16 and untrained VGG-Face. The architecture of VGG-16 is almost identical to the pretrained VGG-Face except that the last FC layer includes 1000 units rather than 2622 units. VGG-16 was trained for classifying 1000 object categories using natural object images from ImageNet [25], thus it only had object recognition experience. The untrained VGG-Face preserved the identical architecture of VGG-Face while randomly assigning the connective weights (Xavier normal initialization) [18,26], and had no training experience.

Images from the stimulus set 1 were also presented to VGG-16 and untrained VGG-Face respectively, and the responses of the units in the conv5-3 layer were extracted. Then, as that used in the pretrained VGG-Face, a two-way non-repeated ANOVA was performed to detect expression-selective units: 835 (0.83%) and 644 (0.64%) of the 100,352 total units were found to be expression-selective in the VGG-16 and in the untrained VGG-Face, respectively. It seemed that expression-selective units also spontaneously emerged in the VGG-16 with the experience of the natural visual objects, and even in the untrained VGG-Face without any visual experience.

Then, for each DCNN, images from the AffectNet Database were applied to test the reliability and generality of the units' expression recognition ability. As in the pretrained VGG-Face, the classification accuracies in these two DCNNs were higher than the chance level (VGG-16: accuracy = 23.29%, 95% CI = [22.10%, 24.48%]; untrained VGG-Face: accuracy = 21.60%, 95% CI = [20.46%, 22.79%], bootstrap with 10,000 replications)(Fig. 4A). Crucially, by further comparing the classification accuracies of these two DCNNs with the pretrained VGG-Face, we found that the performance of the expression-selective units in the pretrained VGG-Face was significantly much higher than VGG-16 ($p < 0.001$, Mann-Whitney U test) and untrained VGG-Face ($p < 0.001$). Besides, the accuracy of expression-selective units in the VGG-16 was better than the untrained VGG-Face ($p < 0.001$). The results revealed expression-selective units in the DCNNs, whether with face identity recognition experience or not, could classify facial expressions. More importantly, the face identity recognition experience was more beneficial than general object classification experience for the enhancement of the units' expression recognition

ability.

Further and importantly, for both DCNNs, the similarities of expression confusion effect between the expression-selective units and humans were tested by correlating their confusion matrices with those of humans (Fig. 2C and 2D, respectively). The confusion matrix of expression-selective units in neither of the two DCNNs resembled those of humans (Fig. 4C, VGG-16: $r_{\text{RaFD}} = 0.20$, $p_{\text{RaFD}} = 0.24$; $r_{\text{NimStim}} = 0.24$, $p_{\text{NimStim}} = 0.15$)(Fig. 4D, untrained VGG-Face: $r_{\text{RaFD}} = 0.18$, $p_{\text{RaFD}} = 0.29$; $r_{\text{NimStim}} = 0.20$, $p_{\text{NimStim}} = 0.23$). Collectively, only the expression-selective units in the pretrained VGG-Face presented a human-like expression confusion effect. These results implied that, at least for the recognition facial expression, the domain-specific training experience was necessary for a DCNN to behave human-like perception.

As the expression-selective units in the VGG-16 and untrained VGG-Face showed no similarity to human expression recognition, we hypothesized that they may not perceive expressions like humans. To verify this hypothesis, we further investigated whether the expression-selective units in these two DCNNs would show categorical perception effect by used the same analysis mentioned above. As shown in Fig. 4E, the expression-selective units from both the VGG-16 and untrained VGG-Face only presented a weak S-shaped trend in very few continua. By comparing their goodness of fitting with that of the pretrained VGG-Face, the identification curves of the expression-selective units in the VGG-16 and untrained VGG-Face showed a more obvious linear trend than that in the pretrained VGG-Face (Linear: pretrained VGG-Face vs. VGG-16: $p = 0.003$; pretrained VGG-Face vs. untrained VGG-Face: $p = 0.005$; VGG-16 vs. untrained VGG-Face: $p = 0.609$, Mann–Whitney U test)(Fig. 4F). Correspondingly, they presented a worse logistic trend than the pretrained VGG-Face (Logistic: pretrained VGG-Face vs. VGG-16: $p = 0.002$; pretrained VGG-Face vs. untrained VGG-Face: $p = 0.002$; VGG-16 vs. untrained VGG-Face: $p = 0.307$)(Fig. 4F). Taken together, the face identity recognition experience, which was domain-specific, helped expression-selective units in the DCNN to achieve human-like categorical perception of facial expressions, whereas the general object classification experience and the architecture itself may only help capture physical features of facial expressions.

Discussion and conclusion

The purpose of the current study was to evaluate whether the spontaneous emergence of human-like expression-selective units in DCNNs would depend on domain-specific visual experience. We found that the pretrained VGG-Face, a DCNN with irrelevant face identity recognition experience, could spontaneously generate expression-selective units. And these units allowed reliable human-like expression perception, including expression confusion effect and categorical perception effect. By further comparing the pretrained VGG-Face with VGG-16 and untrained VGG-Face, we found that although all the three DCNNs could generate expression-selective units, their performance of expression classification differed. The classification accuracy of the expression-selective units in the pretrained VGG-Face was the highest whereas that in the untrained VGG-Face was the lowest. More critically, only the expression-selective units in the pretrained VGG-Face showed apparent human-like expression confusion effect and categorical perception effect. Expression-selective units in both the VGG-16 and untrained VGG-Face did not perform similar to human perception, that is, they showed no human-like confusion effect and exhibited a continuous linear perception of morphed facial expressions instead of categorical perception. These results indicated that the human-like expression perception could only spontaneously emerge in the pretrained VGG-Face with domain-specific experience (i.e., visual experience of face identities), but not in the VGG-16 with task-irrelevant visual experience or the untrained VGG-Face without any visual experience. This finding supports the idea that human-like facial expression perception relied on face identity recognition experience.

Previous research has demonstrated that DCNNs could attain sensitivity to abstract natural features, such as number and face identity, by exposures to irrelevant natural visual stimuli [18,34,42], or even by randomly distributed weights without any training experience [21,43]. The DCNNs' innate sense of number was consistent with the spontaneous representation of visual numerosity in various species, including nonhuman primates [44] and birds [45]. Similarly, the emergence of face identity selective units in untrained DCNNs was in line with the face selectivity founded in 1-month monkey infants [46]. The spontaneous emergence of the selectivity of number and face identity in nonhuman and infant biological systems and untrained *in silico* DCNNs suggested that hard-wired connections of neural circuit were sufficient

to perceive numerosity and face identity. Recent studies further explained the innate ability to recognize face identity might result from that face identity information could be represented by generic object features [18,47]. Therefore, in both biological systems and DCNNs, the extraction of high-order information of natural features such as number and face configuration depended much more on the physical architecture of the networks rather than to the training experience. Consistently, in the present study we found that besides the pretrained VGG-Face, both the VGG-16 and untrained VGG-Face could generate expression-selective units owing to the network architecture.

However, we argue that the pretrained VGG-Face is fundamentally different from VGG-16 and untrained VGG-Face. Unlike number sense and face identity recognition, only the information conveyed by the expression-selective units in the pretrained VGG-Face was sufficient to explain expression confusion effect and categorical expression perception observed in humans. Since the architectures of all three DCNNs (pretrained VGG-Face, VGG-16, and untrained VGG-Face) were identical, their divergence of expression perception originated from distinct training experiences. Namely, the human-like expression perception in the pretrained VGG-Face depended on domain-specific visual experience. The necessity of domain-specific training experience was consistent with the discoveries in biological systems. Specifically, Infants were not born with categorical perception of facial expressions. For instance, they began to show true categorical perception of happy faces and fearful faces only when they were at least 7-month old [48–50], and their discriminability of some other expression continua might develop even later [51]. In addition, a study examining international adopted children revealed that early postnatal deprivation to other-race faces disrupted expression recognition and heightened amygdala response to out-group emotional faces relative to in-group faces [52], revealing the importance of early domain-specific experience for the development of racial-specific facial expression processing. There is also evidence directly support that familiarity and perceptual learning can improve categorical perception [53,54]. Therefore, the architectures of both biological neural systems and artificial neural systems are insufficient to approach adult-level facial expression perception. Concurrent face identity development [55] or domain-specific training experience is needed.

Why would the human-like expression perception in DCNN distinctly rely on

domain-specific experience compared to e.g., number sense and face identity recognition? We think that considering their distinct development or evolution in biological systems, the uniqueness of expression processing may originate from the difficulty of extracting abstract social information using generic natural features. The present findings revealed that while the expression-selective units in the pretrained VGG-Face extracted categorical/discontinuous expression information from morph continua, the expression-selective units in the VGG-16 and untrained VGG-Face merely extracted continuous linear information from visual features. The results indicated that the continuous representation of facial expression was architecture-dependent, but the categorical representation of facial expression was domain-specific experience dependent. Therefore, the face identity and facial expression were recognized by at least partly interconnected systems instead of separate parallel systems.

Indeed, the categorical perception of morphed expressions in the pretrained VGG-Face was in line with the categorical representation of expression in the amygdala, while the VGG-16 and untrained VGG-Face resembled pSTS which exhibited a continuous linear representation of morphed expressions [41,56]. The continuous representation of expression in the VGG-16 and untrained VGG-Face endowed the expression-selective units with a weak ability to recognize expressions, coinciding with the previous finding showing recognition of expressions at a certain level due to image-based principal components [39]. Thus, it indicated that expression representation in the ventral visual pathway, VGG-16 and untrained VGG-Face relied mainly on the similarities of physical properties of images with facial expressions, whereas expression representation in the amygdala and pretrained VGG-Face depended on the categorical information in faces (namely, social meaning) that was critical for making rapid and correct physiological responses to threat and danger. Based on these findings, we would suspect that the function of the core face network may be inborn, but the normal function of the extended face network would rely on postnatal domain-specific experience. However, future developmental study combined with advanced neuroimaging techniques for infants may be needed to confirm this hunch.

Together, theoretical contributions of the present study are two-folds. Firstly, our findings added strong evidence supporting DCNNs' potential to perform human-like representation. The spontaneous generation of human-like facial expression confusion

effect and categorical perception in the pretrained VGG-Face were in line with other similarities between DCNNs and humans, such as similar coding hierarchy as the feedforward visual-cortical network [16,17,57,58], number sense [34,43], object shape perception [59], face identity recognition [18,21], and perceptual learning [60]. Secondly and perhaps more importantly, our computational findings revealed the necessity of domain-specific visual experience of face identity for the development of facial expression perception, and suggested a biologically plausible model for internal brain processing of social information in addition to generic natural features after pretrained on domain-specific tasks, highlighting the contribution of nurture to form human-like facial expression perception. As there exists challenges to conduct human developmental research, such as ethical concerns, recruitments difficulties, participant attrition, and time-consuming (particularly in long-term longitudinal studies), the advantage of systematic comparisons among DCNNs at different levels showcases how DCNN could be appropriately utilized as a powerful tool for the study of human cognitive development. Beyond the weak equivalence between human and DCNNs at the input-output behavior, emerging simulated algorithms between models and humans could be established through domain-specific experience.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgements

This work was supported by the National Key R&D Program of China (2019YFA0709503), National Nature Science Foundation of China grant (31671133), National Basic Research Program.

Author contributions

L. Zhou and K. Zhou contributed to the conceptualization and methodology. L. Zhou contributed to the data curation and writing-original draft. K. Zhou and M. Meng contributed to the writing-review & editing and supervision of the project.

References

- [1] Bruce V, Young A. Understanding face recognition. *Br J Psychol* 1986;77:305–27.
<https://doi.org/10.1111/j.2044-8295.1986.tb02199.x>.
- [2] Bruce V. Influences of familiarity on the processing of faces. *Perception* 1986;15:387–97. <https://doi.org/10.1068/p150387>.
- [3] Calder AJ, Keane J, Young AW, Dean M. Configural information in facial expression perception. *J Exp Psychol Hum Percept Perform* 2000;26:527–51.
<https://doi.org/10.1037/0096-1523.26.2.527>.
- [4] Young AW, Newcombe F, Haan EHFD, Small M, Hay DC. Face perception after brain injury: Selective impairments affecting identity and expression. *Brain* 1993;116:941–59.
<https://doi.org/10.1093/brain/116.4.941>.
- [5] Humphreys GW, Donnelly N, Riddoch MJ. Expression is computed separately from facial identity, and it is computed separately for moving and static faces: Neuropsychological evidence. *Neuropsychologia* 1993;31:173–81.
[https://doi.org/10.1016/0028-3932\(93\)90045-2](https://doi.org/10.1016/0028-3932(93)90045-2).
- [6] Duchaine BC, Parker H, Nakayama K. Normal recognition of emotion in a prosopagnosic. *Perception* 2003;32:827–38. <https://doi.org/10.1068/p5067>.
- [7] Haxby J V., Hoffman EA, Gobbini MII. The distributed human neural system for face perception. *Trends Cogn Sci* 2000;4:223–33.
[https://doi.org/10.1016/S1364-6613\(00\)01482-0](https://doi.org/10.1016/S1364-6613(00)01482-0).
- [8] Winston JS, Henson RNA, Fine-Goulden MR, Dolan RJ. fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *J Neurophysiol* 2004;92:1830–9. <https://doi.org/10.1152/jn.00155.2004>.
- [9] Andrews TJ, Ewbank MP. Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *Neuroimage* 2004;23:905–13.
<https://doi.org/10.1016/j.neuroimage.2004.07.060>.
- [10] Fox CJ, Hanif HM, Iaria G, Duchaine BC, Barton JJSS. Perceptual and anatomic patterns of selective deficits in facial identity and expression processing. *Neuropsychologia* 2011;49:3188–200. <https://doi.org/10.1016/j.neuropsychologia.2011.07.018>.

-
- 619 [11] Redfern AS, Benton CP. Expression dependence in the perception of facial identity.
620 Iperception 2017;8:1–15. <https://doi.org/10.1177/2041669517710663>.
- 621 [12] Ganel T, Goshen-Gottstein Y. Effects of familiarity on the perceptual integrality of
622 the identity and expression of faces: The parallel-route hypothesis revisited. J Exp Psychol Hum
623 Percept Perform 2004;30:583–97. <https://doi.org/10.1037/0096-1523.30.3.583>.
- 624 [13] Yankouskaya A, Rotshtein P, Humphreys GW. Interactions between identity and
625 emotional expression in face processing across the lifespan: Evidence from redundancy gains. J
626 Aging Res 2014;2014:1–12. <https://doi.org/10.1155/2014/136073>.
- 627 [14] Gerlicher AMV, Van Loon AM, Scholte HS, Lamme VAF, Van der Leij AR.
628 Emotional facial expressions reduce neural adaptation to face identity. Soc Cogn Affect
629 Neurosci 2014;9:610–4. <https://doi.org/10.1093/scan/nst022>.
- 630 [15] Baseler HA, Harris RJ, Young AW, Andrews TJ. Neural responses to expression
631 and gaze in the posterior superior temporal sulcus interact with facial identity. Cereb Cortex
632 2014;24:737–44. <https://doi.org/10.1093/cercor/bhs360>.
- 633 [16] Yamins DLKK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ.
634 Performance-optimized hierarchical models predict neural responses in higher visual cortex.
635 Proc Natl Acad Sci U S A 2014;111:8619–24. <https://doi.org/10.1073/pnas.1403112111>.
- 636 [17] Bashivan P, Kar K, Dicarlo JJ, Figs S. Neural population control via deep image
637 synthesis. Science (80-) 2019;364. <https://doi.org/10.1126/science.aav9436>.
- 638 [18] Grossman S, Gaziv G, Yeagle EM, Harel M, Mégevand P, Groppe DM, et al.
639 Convergent evolution of face spaces across human face-selective neuronal groups and deep
640 convolutional networks. Nat Commun 2019;10:4934.
641 <https://doi.org/10.1038/s41467-019-12623-6>.
- 642 [19] O’Neill KC, Saxe R, Anzellotti S, Nell KCO, Saxe R, Anzellotti S. Recognition of
643 identity and expressions as integrated processes. PsyArXiv 2019:1–18.
644 <https://doi.org/10.31234/osf.io/9c2e5>.
- 645 [20] Colón YI, Castillo CD, O’Toole AJ. Facial expression is retained in deep networks
646 trained for face identification. J Vis 2021;21:4. <https://doi.org/10.1167/jov.21.4.4>.
- 647 [21] Baek S, Song M, Jang J, Kim G, Paik S. Spontaneous generation of face recognition
648 in untrained deep neural networks. BioRxiv Prepr 101101/857482 2019.

649 <https://doi.org/http://dx.doi.org/10.1101/857466>.

650 [22] Marr D. Vision: a computational investigation into the human representation and
651 processing of visual information. San Francisco: W.H. Freeman; 1982.

652 [23] Dawson MRW. Mind, body, world: Foundations of cognitive science. vol. 1.
653 Athabasca University Press; 2013.

654 [24] Parkhi OM, Vedaldi A, Zisserman A. Deep Face Recognition 2015:41.1-41.12.
655 <https://doi.org/10.5244/c.29.41>.

656 [25] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale
657 Image Recognition. 3rd Int Conf Learn Represent ICLR 2015 - Conf Track Proc 2014:1-14.

658 [26] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward
659 neural networks. J Mach Learn Res 2010;9:249-56.

660 [27] Ekman P, Cordaro D. What is meant by calling emotions basic. Emot Rev
661 2011;3:364-70. <https://doi.org/10.1177/1754073911410740>.

662 [28] Ekman P. An argument for basic emotions. Cogn Emot 1992;6:169-200.
663 <https://doi.org/10.1080/02699939208411068>.

664 [29] Lundqvist D, Flykt A, Öhman A. The Karolinska directed emotional faces (KDEF),
665 CD ROM from Department of Clinical Neuroscience, Psychology section. Karolinska Institutet
666 1998.

667 [30] Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare TA, et al. The
668 NimStim set of facial expressions: Judgments from untrained research participants. Psychiatry
669 Res 2009;168:242-9. <https://doi.org/10.1016/j.psychres.2008.05.006>.

670 [31] Langner O, Dotsch R, Bijlstra G, Wigboldus DHJJ, Hawk ST, van Knippenberg A.
671 Presentation and validation of the Radboud Faces Database. Cogn Emot 2010;24:1377-88.
672 <https://doi.org/10.1080/02699930903485076>.

673 [32] Mollahosseini A, Hasani B, Mahoor MH. AffectNet: A Database for Facial
674 Expression, Valence, and Arousal Computing in the Wild. IEEE Trans Affect Comput
675 2019;10:18-31. <https://doi.org/10.1109/TAFFC.2017.2740923>.

676 [33] Willenbockel V, Sadr J, Fiset D, Horne GO, Gosselin F, Tanaka JW. Controlling
677 low-level image properties: The SHINE toolbox. Behav Res Methods 2010;42:671-84.
678 <https://doi.org/10.3758/BRM.42.3.671>.

679 [34] Nasr K, Viswanathan P, Nieder A. Number detectors spontaneously emerge in a
680 deep neural network designed for visual object recognition. *Sci Adv* 2019;5:1–11.
681 <https://doi.org/10.1126/sciadv.aav7903>.

682 [35] Lindsay GW, Miller KD. How biological attention mechanisms improve task
683 performance in a large-scale visual system model. *Elife* 2018;7:1–29.
684 <https://doi.org/10.7554/eLife.38105>.

685 [36] Etcoff NL, Magee JJ. Categorical perception of facial expressions. *Cognition*
686 1992;44:227–40. [https://doi.org/10.1016/0010-0277\(92\)90002-Y](https://doi.org/10.1016/0010-0277(92)90002-Y).

687 [37] Calder AJ, Young AW, Perrett DI, Etcoff NL, Rowland D. Categorical perception
688 of morphed facial expressions. *Vis Cogn* 1996;3:81–118. <https://doi.org/10.1080/713756735>.

689 [38] Fujimura T, Matsuda YT, Katahira K, Okada M, Okanoya K. Categorical and
690 dimensional perceptions in decoding emotional facial expressions. *Cogn Emot* 2012;26:587–601.
691 <https://doi.org/10.1080/02699931.2011.595391>.

692 [39] Calder AJ, Burton AM, Miller P, Young AW, Akamatsu S. A principal component
693 analysis of facial expressions. *Vision Res* 2001;41:1179–208.
694 [https://doi.org/10.1016/S0042-6989\(01\)00002-5](https://doi.org/10.1016/S0042-6989(01)00002-5).

695 [40] Fox CJ, Moon SY, Iaria G, Barton JJSS, Young S, Iaria G, et al. The correlates of
696 subjective perception of identity and expression in the face network: An fMRI adaptation study.
697 *Neuroimage* 2009;44:569–80. <https://doi.org/10.1016/j.neuroimage.2008.09.011>.

698 [41] Harris RJ, Young AW, Andrews TJ. Morphing between expressions dissociates
699 continuous from categorical representations of facial expression in the human brain. *Proc Natl*
700 *Acad Sci U S A* 2012;109:21164–9. <https://doi.org/10.1073/pnas.1212207110>.

701 [42] Zhou C, Xu W, Liu Y, Xue Z, Chen R, Zhou K, et al. Numerosity representation in
702 a deep convolutional neural network. *J Pacific Rim Psychol* 2021;15:183449092110126.
703 <https://doi.org/10.1177/18344909211012613>.

704 [43] Kim G, Jang J, Baek S, Song M, Paik S-BB. Spontaneous generation of innate
705 number sense in untrained deep neural networks. *BioRxiv* 2019:857482.
706 <https://doi.org/10.1101/857482>.

707 [44] Viswanathan P, Nieder A. Neuronal correlates of a visual “sense of number” in
708 primate parietal and prefrontal cortices. *Proc Natl Acad Sci U S A* 2013;110:11187–92.

709 <https://doi.org/10.1073/pnas.1308141110>.

710 [45] Wagener L, Loconsole M, Ditz HM, Nieder A. Neurons in the Endbrain of

711 Numerically Naive Crows Spontaneously Encode Visual Numerosity. *Curr Biol*

712 2018;28:1090-1094.e4. <https://doi.org/10.1016/j.cub.2018.02.023>.

713 [46] Livingstone MS, Vincent JL, Arcaro MJ, Srihasam K, Schade PF, Savage T.

714 Development of the macaque face-patch system. *Nat Commun* 2017;8.

715 <https://doi.org/10.1038/ncomms14897>.

716 [47] Bao P, She L, McGill M, Tsao DY. A map of object space in primate inferotemporal

717 cortex. *Nature* 2020. <https://doi.org/10.1038/s41586-020-2350-5>.

718 [48] Kotsoni E, De Haan M, Johnson MH, Haanô M De, Johnson MH, De Haan M, et al.

719 Categorical perception of facial expressions by 7-month-old infants. *Perception*

720 2001;30:1115–25. <https://doi.org/10.1068/p3155>.

721 [49] Leppanen JM, Richmond J, Vogel-farley VK, Moulson MC, Nelson CA, Leppänen

722 JM, et al. Categorical representation of facial expressions in the infant brain. *Infancy*

723 2009;14:346–62. <https://doi.org/10.1080/15250000902839393>.

724 [50] Hoemann K, Wu R, LoBue V, Oakes LM, Xu F, Barrett LF. Developing an

725 Understanding of Emotion Categories: Lessons from Objects. *Trends Cogn Sci* 2020;24:39–51.

726 <https://doi.org/10.1016/j.tics.2019.10.010>.

727 [51] Lee V, Cheal JL, Rutherford MD. Categorical perception along the happy-angry

728 and happy-sad continua in the first year of life. *Infant Behav Dev* 2015;40:95–102.

729 <https://doi.org/10.1016/j.infbeh.2015.04.006>.

730 [52] Telzer EH, Flannery J, Shapiro M, Humphreys KL, Goff B, Gabard-Durman L, et al.

731 Early experience shapes amygdala sensitivity to race: An international adoption design. *J*

732 *Neurosci* 2013;33:13484–8. <https://doi.org/10.1523/JNEUROSCI.1272-13.2013>.

733 [53] Du Y, Zhang F, Wang Y, Bi T, Qiu J. Perceptual Learning of Facial Expressions.

734 *Vision Res* 2016;128:19–29. <https://doi.org/10.1016/j.visres.2016.08.005>.

735 [54] Beale JM, Keil FC, Elsevier C, Keil C. Categorical effects in the perception of faces.

736 *Cognition* 1995;57:217–39. [https://doi.org/10.1016/0010-0277\(95\)00669-X](https://doi.org/10.1016/0010-0277(95)00669-X).

737 [55] Dalrymple KA, Visconti Di Oleggio Castello M, Elison JT, Gobbini MI, Visconti

738 M, Elison JT, et al. Concurrent development of facial identity and expression discrimination.

739 PLoS One 2017;12:1–16. <https://doi.org/10.1371/journal.pone.0179458>.

740 [56] Harris RJ, Young AW, Andrews TJ. Dynamic stimuli demonstrate a categorical
741 representation of facial expression in the amygdala. *Neuropsychologia* 2014;56:47–52.
742 <https://doi.org/10.1016/j.neuropsychologia.2014.01.005>.

743 [57] Tripp BP. Similarities and differences between stimulus tuning in the
744 inferotemporal visual cortex and convolutional networks. *Proc Int Jt Conf Neural Networks*
745 2017;2017-May:3551–60. <https://doi.org/10.1109/IJCNN.2017.7966303>.

746 [58] Wen H, Shi J, Zhang Y, Lu KH, Cao J, Liu Z. Neural encoding and decoding with
747 deep learning for dynamic natural vision. *Cereb Cortex* 2018;28:4136–60.
748 <https://doi.org/10.1093/cercor/bhx268>.

749 [59] Kubilius J, Bracci S, Op de Beeck HP. Deep Neural Networks as a Computational
750 Model for Human Shape Sensitivity. *PLoS Comput Biol* 2016;12:1–26.
751 <https://doi.org/10.1371/journal.pcbi.1004896>.

752 [60] Wenliang LK, Seitz AR. Deep Neural Networks for Modeling Visual Perceptual
753 Learning. *J Neurosci* 2018;38:6028–44. <https://doi.org/10.1523/jneurosci.1620-17.2018>.
754
755

Figure captions

Fig. 1 Expression-selective units emerged in the pretrained VGG-Face. (A). The architecture of the VGG-Face. (B) The tuning value map of the expression-selective units in the pretrained VGG-Face. (C) The expression classification performance of the expression-selective units. The black dashed line represents the chance level. Error bars indicate SE. ***: $p \leq 0.001$.

Fig. 2 Human-like expression confusion effect of the expression-selective units in the pretrained VGG-Face. (A) The expression discriminability of the expression-selective units emerged in the pretrained VGG-Face for the RaFD. The black dashed line represents the chance level. (B) The confusion matrix of the expression-selective units for the RaFD. (C) Confusion matrix for the RaFD in human study (Adapted from Fig. 4 in Langner et al. 2010). (D) Confusion matrix for the NimStim set in human study (Adapted from Table 2 in Tottenham et al. 2009).

Fig. 3 Categorical perception of facial expressions of the expression-selective units in the pretrained VGG-Face. (A) Example facial stimuli used in morph continua (Happy-Anger). (B) The identification rates for the seven continua. The identification rates refer to the identification frequency of one out of the two expressions. Labels along the x-axis indicate the percentage of this expression in facial stimuli. Black dots represent true identification rates. Blue solid lines indicate fitting for the logistic function. (C) Goodness-of-fit (R^2) of each regression type for each expression continuum. The black dashed lines represent R^2 at 0.95 and 1.00, separately. (D) Mean goodness-of-fit (R^2) among expression continua. The R^2 in the Logistic regression was much higher than the other two regressions. Error bars indicate SE. *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$.

Fig. 4 Expression recognition of the expression-selective units for the expressions from the AffectNet in the pretrained VGG-Face, VGG-16, and untrained VGG-Face. (A) The expression discriminability of the expression-selective units for each DCNN. Expression classification of the expression-selective units in the pretrained VGG-Face is much better than in the VGG-16 and untrained VGG-Face. The black dashed line represents the chance level. (B-D) The confusion matrix of the expression-selective units in the pretrained VGG-Face (B), VGG-16 (C), and untrained VGG-Face (D). (E) The identification rates for the seven continua in the VGG-16 and the untrained VGG-Face, respectively. Black dots represent true identification rates. Blue solid lines indicate fitting for the logistic function. (F) The Goodness-of-fit (R^2) of each fit type for each DCNN. Logistic regression fits best for the pretrained VGG-Face than for the other two DCNNs, whereas linear regression fits worst for the pretrained VGG-Face. Error bars indicate SE. *: $p \leq 0.05$; **: $p \leq 0.01$; ***: $p \leq 0.001$.

A

B

C

A

B

C

D