COGNITIVE NEUROSCIENCE

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

Liqin Zhou[1], Anmin Yang[1], Ming Meng[2,3]*, Ke Zhou[1]*

Recent studies found that the deep convolutional neural networks (DCNNs) trained to recognize facial identities spontaneously learned features that support facial expression recognition, and vice versa. Here, we showed that the self-emerged expression-selective units in a VGG-Face trained for facial identification were tuned to distinct basic expressions and, importantly, exhibited hallmarks of human expression recognition (i.e., facial expression confusion and categorical perception). We then investigated whether the emergence of expression-selective units is attributed to either face-specific experience or domain-general processing by conducting the same analysis on a VGG-16 trained for object classification and an untrained VGG-Face without any visual experience, both 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. Together, these 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.

INTRODUCTION

Facial identity and expression play important roles in daily life and social communication. When interacting with others, we can easily recognize who they are through their facial identity information and 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 differed (*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.* (*7*) 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) and 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 in face identity recognition, and those with damage to the posterior STS (pSTS) suffer impairments in expression recognition (*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 contains 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 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, deep convolutional neural networks (DCNNs) have 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 to recognize 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 an untrained DCNN (*21*), which seemed to cast substantial doubt on the role of nurture in developing face perception and the above-mentioned speculation. 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

[1] Beijing Key Laboratory of Applied Experimental Psychology, Faculty of Psychology, Beijing Normal University, Beijing 100875, China. [2] Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents (South China Normal University), Ministry of Education, Guangzhou 510631, China. [3] Guangdong Key Laboratory of Mental Health and Cognitive Science, School of Psychology, South China Normal University, Guangzhou 510631, China.
*Corresponding author. Email: mingmeng@m.scnu.edu.cn (M.M.); kzhou@bnu.edu.cn (K.Z.)

humans adopt similar representational mechanisms (i.e., algorithms) to achieve the 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., similarity in algorithms between them (*23*).

Therefore, in the present study, we borrowed the cognitive approaches developed in human research 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 a 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, to answer the question of what the 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 an 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.

RESULTS

Expression-selective units spontaneously emerge in the pretrained VGG-Face

We first explored whether expression-selective units could spontaneously emerge in the pretrained VGG-Face. The pretrained VGG-Face was trained with more than 2 million face images to recognize 2622 identities (*24*). It consists of 13 convolutional (conv) layers and 3 fully connected (FC) layers (Fig. 1A). The first 13 convolutional layers form a feature extraction network that transforms images to a goal-directed high-level representation, and the following 3 FC layers form a classification network to classify images by converting the high-level representation into classification probabilities (*25*).

Since the final layer (conv5-3) of the feature extraction network represents the highest level representation (*26*, *27*) and has the largest receptive field among all convolutional layers, we tested the expression selectivity of each unit in this layer using stimulus set 1 to explore whether a DCNN could spontaneously generate facial expression–selective "neurons" (see Materials and Methods for details). Stimulus set 1 consisted of 104 different facial identities selected from the Karolinska Directed Emotional Faces (KDEF) (*28*) and NimStim (*29*) databases, and each identity has six basic expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise) (*30*, *31*). All 624 images in stimulus set 1 were presented to the pretrained VGG-Face, and their activations in the conv5-3 layer were extracted. First, we conducted a two-way nonrepeated analysis of variance (ANOVA) with identity and expression as factors to detect units selective to facial expression ($P \leq 0.01$) but not to face identity ($P > 0.01$). The units meeting the criteria were defined as the expression-selective units. Of the total 100,352 units, 1259 units (1.25%) in the conv5-3 layer were found to be expression selective. Then, for each expression-selective unit, its tuning value (*32*) for each expression category was calculated to measure whether and to what extent it preferred a specific expression. As shown in Fig. 1B, almost all units responded selectively to only one specific expression and exhibited a tuning effect. Last, to test whether the responses of these expression-selective units provide sufficient information for successful expression recognition, we performed principal components analysis (PCA) on the activations of these units to all images in stimulus set 1 and selected the first 600 principal components (PCs) to perform an expression classification task using a support vector classification (SVC) analysis with 104-fold cross-validation. The 600 PCs could explain nearly 100% variance of the expression-selective features (fig. S1). We found that the classification accuracy (mean ± SE, 76.76 ± 1.59%) of the expression-selective units was much higher than the chance level (16.67%) and much higher than the classification accuracy of images 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 VGG-Face pretrained for face identity recognition, which echoed previous findings (*19*, *20*).

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 using stimulus set 1 to predict the expressions of images selected from the Radboud Faces Database (RaFD) (*33*). The RaFD is an independent facial expression database including 67 face identities with different head and gaze directions. Only the front-view expressions of each identity were used in the present study (i.e., stimulus set 2). The prediction accuracy of the expressions from stimulus set 2 was significantly higher than the chance level [accuracy = 67.91%; 95% confidence interval (CI), 63.18 to 72.39%, bootstrapped with 10,000 iterations] (Fig. 2A). We also changed the number of PCs from 50 to 600 to explore whether the number of PCs influenced the prediction performance. As shown in fig. S2A, the prediction accuracy remained relatively stable as the number of PCs changed. It thus indicated that the expression-selective units in the pretrained VGG-Face had a reliable expression discriminability.

Subsequently, to test whether the expression representation of these units was similar to humans, we presented the same face images of stimulus set 2 to both human participants (experiment 1, see Materials and Methods for details) and the pretrained VGG-Face and calculated the confusion matrices of facial expression recognition, respectively (Fig. 2, B and C). Although the mean classification accuracy of the human participants (73.47%) was significantly higher than that of the pretrained VGG-Face, the error patterns of the two confusion matrices were highly correlated (Kendall's $\tau = 0.48$, $P = 5.4 \times 10^{-4}$). For instance, in both confusion matrices, fear and surprise might be confused with each other, disgust was frequently mistaken for anger, and anger was often mistaken for sadness. Overall,

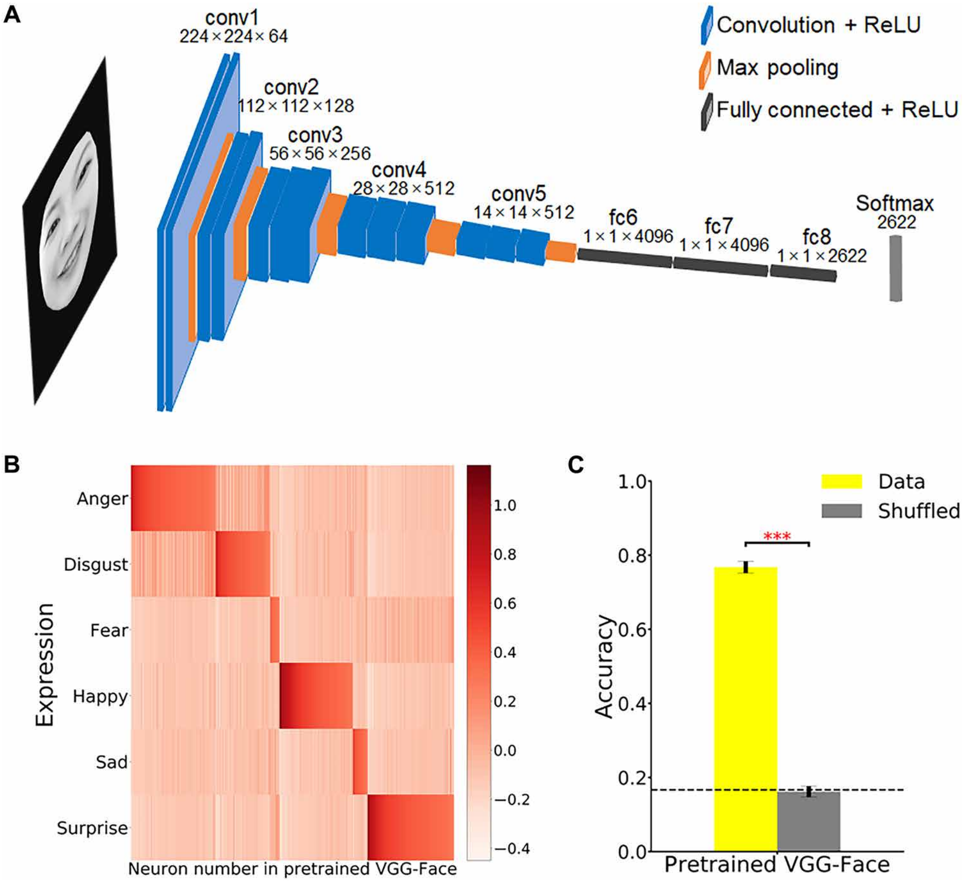

Fig. 1. Expression-selective units emerged in the pretrained VGG-Face. (**A**) The architecture of the VGG-Face. An example face image (for demonstration purposes only) is shown. Photo credit: Liqin Zhou, Beijing Normal University. ReLU, rectification linear unit. (**B**) The tuning value map of the 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. ***$P \leq 0.001$.

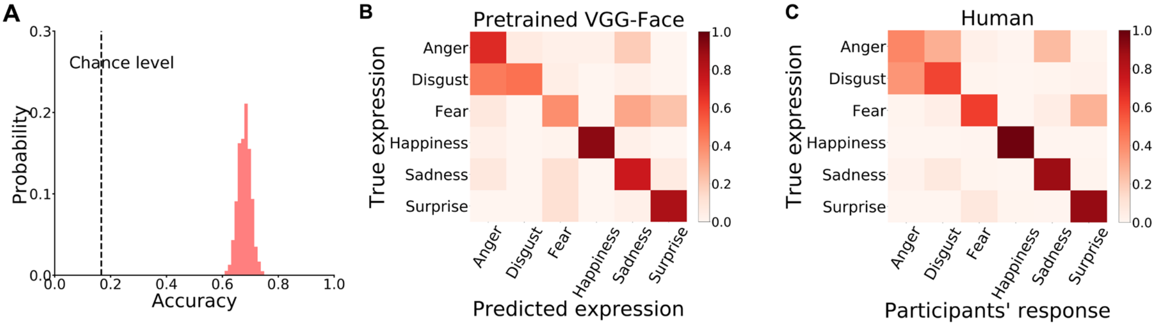

Fig. 2. Human-like expression confusion effect of the expression-selective units in the pretrained VGG-Face for stimulus set 2. (**A**) The expression discriminability of the expression-selective units emerged in the pretrained VGG-Face. The black dashed line represents the chance level. (**B**) The confusion matrix of the expression-selective units in the pretrained VGG-Face for stimulus set 2. (**C**) Human confusion matrix for stimulus set 2.

the results suggested a similar expression confusion effect between the expression-selective units in the pretrained VGG-Face and humans.

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

The facial expressions in stimulus set 1 and stimulus set 2 were collected from the same identities in the laboratory-controlled environment and 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 generated stimulus set 3 by selecting 4800 images with manually annotated expressions from the AffectNet database—a large real-world facial expression database (*34*). Each basic expression included 800 images. Note that, in stimulus set 3, the face identities across expressions are different. By using the same SVC model trained

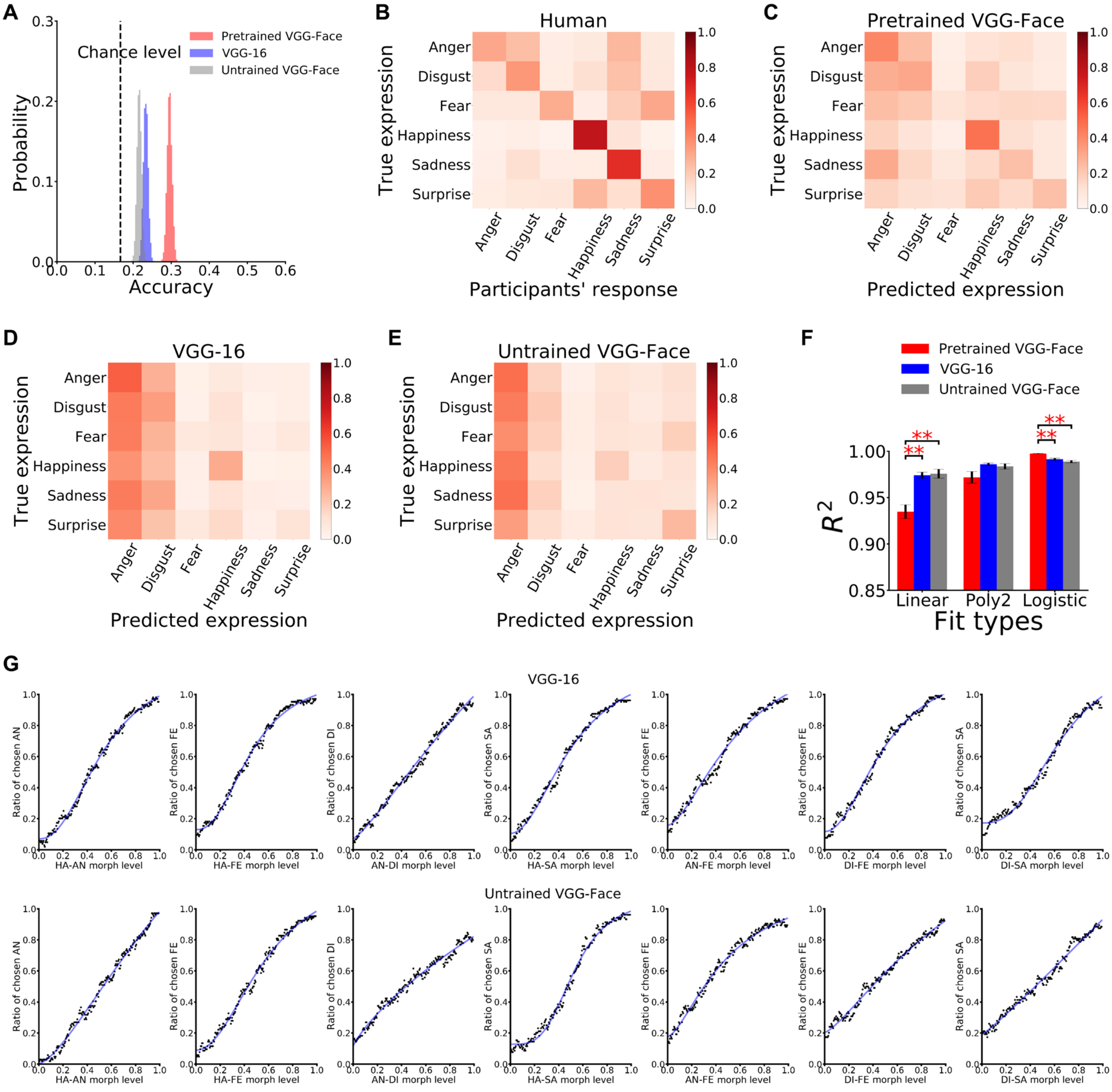

Fig. 3. Expression recognition of the expression-selective units in the pretrained VGG-Face, VGG-16, and untrained VGG-Face for stimulus set 3. (**A**) The expression discriminability of the expression-selective units in each DCNN. Expression classification of the expression-selective units in the pretrained VGG-Face is much better than in the VGG-16 and untrained VGG-Face. The black dashed line represents the chance level. (**B**) Human confusion matrix for stimulus set 3. (**C** to **E**) The confusion matrix of the expression-selective units in the pretrained VGG-Face (C), VGG-16 (D), and untrained VGG-Face (E) for stimulus set 3. (**F**) The goodness of fit (R^2) of each fit type for each DCNN. Logistic regression fits better for the pretrained VGG-Face than for the other two DCNNs, whereas linear regression fits the worst for the pretrained VGG-Face. Error bars indicate SE. $^{**}P \leq 0.01$. (**G**) 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. HA, happiness; AN, anger; FE, fear; DI, disgust and SA, sadness.

with stimulus set 1, we found that the prediction accuracy of the expressions from stimulus set 3 was also significantly higher than the chance level (accuracy = 29.56%; 95% CI, 28.31 to 30.85%, bootstrapped with 10,000 iterations) (Fig. 3A). Similarly, we also obtained the confusion matrices for both human participants (experiment 2, see Materials and Methods for details) and the pretrained VGG-Face (Fig. 3, B and C). Again, the error patterns of the two confusion matrices were highly correlated (Kendall's τ = 0.27, *P* = 0.037), although the mean classification accuracy of the human participants (46.76%) was higher than that of the pretrained VGG-Face. 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 humans do, even for real-life face images.

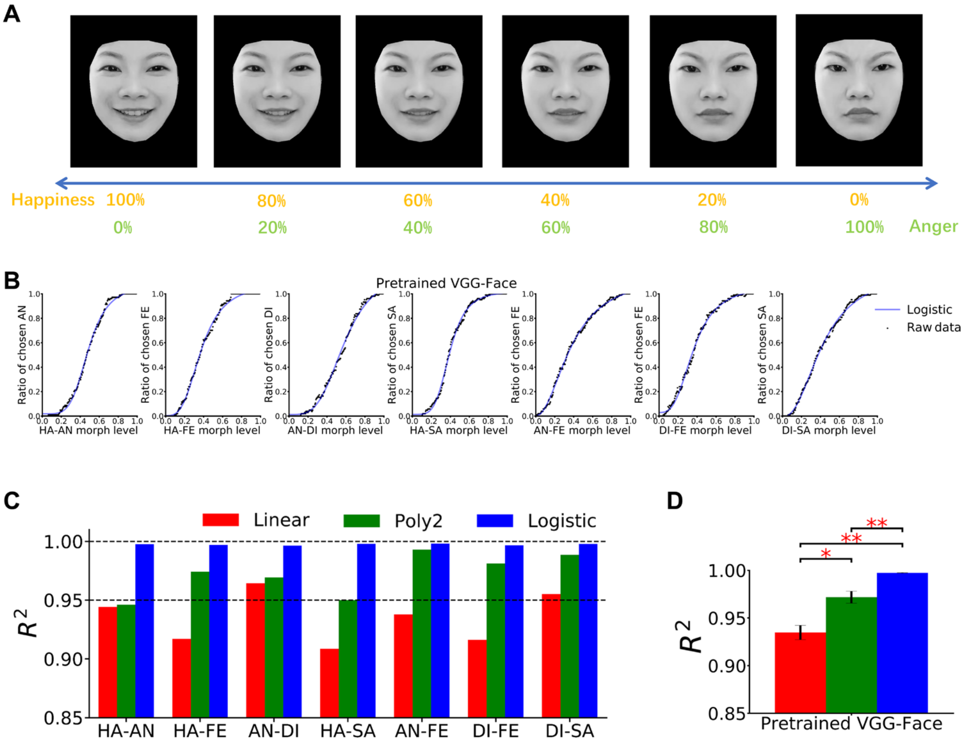

Fig. 4. Categorical perception of facial expressions of the expression-selective units in the pretrained VGG-Face. (**A**) Example facial stimuli used in a morph continuum (happiness-anger). An example face image (for demonstration purposes only) is shown. Photo credit: Liqin Zhou, Beijing Normal University. (**B**) The identification rates for the seven continua. The identification rates refer to the identification frequency of one 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. *$P \leq 0.05$ and **$P \leq 0.01$.

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

One may argue that the similarity in the expression confusion effect does not necessarily mean that 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 (i.e., PCs based on pixel intensities and shapes) could also yield a confusion matrix similar to that of humans (*35*). Therefore, to further confirm whether these units could exhibit a human-like psychophysical response to facial expressions, we tested whether their responses showed a categorical perception of facial expressions by using morphed expression continua. Considering the generality of the categorical emotion perception in humans, 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*–*40*). In detail, we designed a morphed expression discrimination task (Fig. 4A) that resembled the ABX discrimination task designed for humans (*36*, *39*, *40*). The prototypic expressions were selected from 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 of the two expressions was defined as the units' identification rate at the current morph level. We hypothesized that if the selective units perceived 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 expression-selective units in the pretrained VGG-Face were S-shaped (Fig. 4B). 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. 4 (C and D), we found that all seven identification curves showed typical S-like patterns (logistic versus linear: $P = 0.002$ and logistic versus Poly2: $P = 0.002$, Mann-Whitney *U* test).

The human-like expression perception only spontaneously emerged 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 expression-selective units achieve human-like expression perception? Specifically, it was still unknown whether the spontaneous emergence of the human-like expression perception 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 the VGG-16 is almost identical to the pretrained VGG-Face except that the last FC layer includes 1000 units rather than 2622 units. The VGG-16 was trained to classify 1000 object categories using natural object images from ImageNet (*41*); thus, it only had object-related visual experience. The untrained VGG-Face preserved the identical architecture of the VGG-Face while randomly assigning the connective weights (Xavier normal initialization) (*18*, *42*), and had no training experience.

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

Then, for each of the two DCNNs, images from stimulus set 3 were applied to test the reliability and generality of the expression recognition ability of expression-selective units. The classification accuracies in these two DCNNs were also higher than the chance level (pretrained VGG-16: accuracy = 23.33%; 95% CI, 22.13 to 24.54%; untrained VGG-Face: accuracy = 21.60%; 95% CI, 20.44 to 22.79%, bootstrap with 10,000 replications) (Fig. 3A). Crucially, we found that the classification accuracy of the expression-selective units in the pretrained VGG-Face was significantly much higher than those in the pretrained VGG-16 ($P < 0.001$, Mann-Whitney U test) and untrained VGG-Face ($P < 0.001$), and the classification accuracy of the expression-selective units in the pretrained VGG-16 was better than those in the untrained VGG-Face ($P < 0.001$). These results were relatively stable when changing the number of PCs in the SVC model (fig. S2B). The results revealed that expression-selective units in the DCNNs, whether with face identity recognition experience or not, could classify facial expressions. The face identity recognition experience was more beneficial than general object classification experience for the enhancement of the units' expression recognition ability.

Furthermore, for both DCNNs, the similarities of expression confusion effect between the expression-selective units and humans were tested by correlating their error patterns with that of human participants. The error patterns of expression-selective units in neither of the two DCNNs resembled that of human (Fig. 3D, VGG-16: Kendall's $\tau = -2.3 \times 10^{-3}$, $P = 0.986$; Fig. 3E, untrained VGG-Face: Kendall's $\tau = 0.02$, $P = 0.872$). 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 facial expression recognition, the domain-specific training experience was necessary for a DCNN to develop the human-like perception.

As the expression-selective units in the pretrained 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 a categorical perception effect by performing the same ABX discrimination task as that used in the pretrained VGG-Face. As shown in Fig. 3G, the expression-selective units from both the pretrained VGG-16 and untrained VGG-Face only presented a weak S-shaped trend in very few continua. By comparing their goodness of fit with that of the pretrained VGG-Face, the identification curves of the expression-selective units in the pretrained VGG-16 and untrained VGG-Face showed a more obvious linear trend than that in the pretrained VGG-Face (linear: pretrained VGG-Face versus pretrained VGG-16: $P = 0.003$; pretrained VGG-Face versus untrained VGG-Face: $P = 0.005$; pretrained VGG-16 versus untrained VGG-Face: $P = 0.609$; Mann-Whitney U test) (Fig. 3F). Correspondingly, they presented a significantly weaker logistic trend than the pretrained VGG-Face (logistic: pretrained VGG-Face versus pretrained VGG-16: $P = 0.002$; pretrained VGG-Face versus untrained VGG-Face: $P = 0.002$; and pretrained VGG-16 versus untrained VGG-Face: $P = 0.307$) (Fig. 3F). Together, the face identity recognition experience, which was domain specific, helped expression-selective units in the DCNN to achieve a 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.

In addition, we generated three new stimuli sets, including the scrambled (fig. S3B), contrast-negated (fig. S3C), and inverted (fig. S3D) versions of the face images in stimulus set 3 (fig. S3A) and conducted further control analyses to explore the possible contribution of low-level features (e.g., texture, brightness, edge, and gradient) to the expression recognition of the expression-selective units. For example, the inverted face retains all low-level features of the upright faces. We tested whether the expression-selective units of the three DCNNs could reliably classify the expressions of the three new stimulus sets. As shown in table S1 and fig. S3 (E to G), for all the three stimulus sets, the classification accuracies of the expression-selective units in all three DCNNs decreased significantly, near the chance level (all accuracies: <20.20%; chance level: 16.67%). Therefore, it is unlikely that the low-level features in the face images were simply the determining factors for the emergence of the expression-selective units.

DISCUSSION

The purpose of the current study was to evaluate whether the spontaneously emerged human-like expression-selective units in DCNNs would depend on domain-specific visual experience. We found that the pretrained VGG-Face, a DCNN with visual experience of face identity, could spontaneously generate expression-selective units. In addition, 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.

It should be noted that, in our study, the classification accuracies of the expression-selective units in the pretrained VGG-Face were worse than the performance of expression recognition in humans, which was consistent with a recent finding showing that the identity-trained DCNN retained expression information but with expression recognition accuracies far below human performance (*20*). The reason for the decreased expression recognition performance deserves future investigation, although it is beyond the scope of the present study. Critically, although the expression classification accuracy of the identity-trained DCNN was much lower than that of humans, the confusion effect and categorical perception revealed in our study mirrored the findings in humans, suggesting that the expression-selective units that emerged in the identity-trained DCNN represent facial expressions in a manner similar to humans.

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*, *27*, *43*) or even by randomly distributed weights without any training experience (*21*, *26*). 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 found in 1-month-old monkeys (*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 the neural circuit were sufficient to perceive numerosity and face identity. Recent studies further explained that the innate ability to recognize face identity might result from the idea 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 the 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 the 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 the categorical perception of facial expressions. For instance, they began to show the true categorical perception of happy faces and fearful faces only when they were at least 7 months 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 supporting the notion 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 while the continuous representation of facial expression might be architecture dependent, the categorical representation of facial expression might be domain-specific experience dependent.

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 (*38*, *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 PCs (*35*). 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. On the basis of 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, the theoretical contributions of the present study are twofold. First, 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 (*26*, *27*), object shape perception (*59*), face identity recognition (*18*, *21*), and perceptual learning (*60*). Second 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 being pretrained with domain-specific tasks, highlighting the contribution of nurture to form human-like facial expression perception. As there exist challenges in conducting human developmental research, such as ethical concerns, recruitment difficulties, participant attrition, and it being time-consuming (particularly in long-term longitudinal studies), the advantage of systematic comparisons among DCNNs at different levels showcases how DCNN could be appropriately used as a powerful tool for the study of human cognitive development. Beyond the weak equivalence between humans and DCNNs at the input-output behavior, emerging simulated algorithms between models and humans could be established through domain-specific experience.

MATERIALS AND METHODS

Neural network models

The VGG-Face (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 while requiring less data than other state-of-the-art models (DeepFace and FaceNet). The network consists of 13 convolutional layers and 3 FC layers. All these 16 layers are followed by a rectification linear unit. The 13 convolutional layers are distributed into 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 more than 14 million natural visual images (https://arxiv.org/abs/1409.1556) (*41*). 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*, *42*) 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 six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) (*30*, *31*). In stimulus set 1, 70 identities were from Karolinska Directed Emotional Faces (KDEF) database (*28*), and the other 34 identities were from the NimStim database (*29*). Then, to validate the reliability of the expression recognition capability of the expression-selective units, a second stimulus set—images from the Radboud Faces Database (RaFD) (*33*)—was applied. The front view images of all the 67 identities in the RaFD were used, and each identity contained the aforementioned six expressions. Furthermore, we applied images from the AffectNet database as stimulus set 3 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 collected in the real-world environment (*34*). A total of 4800 manually annotated images from the AffectNet database were used (800 images for each expression). In this database, the face identities across expressions are different.

For each stimulus set, the luminance and contrast of the images were matched by using the SHINE toolbox (*61*), and the face part was 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 stimulus set 1. All identities in 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.

In addition, we generated three new stimuli sets, including the scrambled (fig. S3B), contrast-negated (fig. S3C), and inverted (fig. S3D) versions of the face images in stimulus set 3 (fig. S3A) for the control analyses. The scrambled face image was generated by dividing the original face image into 12 × 10 blocks and shuffling the central 45 blocks of the image that covered the face area. The contrast-negated face image was generated by reversing the luminance values of the original face image. The inverted face image was generated by flipping the original face image upside down. Then, the images of all three new stimuli sets were resized to 244 × 224 pixels.

Human behavioral experiment

Experiment 1: Expression classification task on stimulus set 2

Participants. Twenty healthy college students (19 females: mean age = 20.35, SD = 1.68) participated in experiment 1. All had normal or corrected-to-normal vision. The study was approved by the Ethical Committee of the Beijing Normal University, and all participants provided written informed consent before the experiment.

Stimulus and procedure. The stimuli were 402 front view images of stimulus set 2 (67 identities from the RaFD database, each with six expressions). Participants were instructed to classify each image into one of the six facial expressions. Each image was tested twice for each participant.

Analysis. The confusion matrix was calculated for each participant, which was a 6-by-6 matrix with the rows representing the true expressions (ground truth) and the columns representing the expressions discriminated by the participant. The element (*i, j*) of the confusion matrix indicated the ratio of how many times the expression *i* was recognized as the expression *j*. The final confusion matrix of human participants was defined as the average of the confusion matrices of all participants.

Experiment 2: Expression classification task on stimulus set 3

Participants. Thirty-six healthy college students (29 females; mean age = 20.00, SD = 1.55) participated in experiment 2. All had a normal or corrected-to-normal vision. The study was approved by the Ethical Committee of the Beijing Normal University, and all participants provided written informed consent before the experiment.

Stimulus and procedure. The stimuli were 4800 images of stimulus set 3 from the AffectNet database (800 images for each of the six expressions). All images were randomly divided into eight groups of 600 images, of which each expression included 100 images. Each participant completed the expression discrimination task of one to eight groups of images. Participants were instructed to classify each image into one of the six facial expressions. Last, each image was classified by eight participants.

Analysis. There were 38,400 trials in total (4800 different images, each repeated eight times). As each participant only completed the expression classification of several groups of images, we pooled the data of all participants to calculate the confusion matrix. The confusion matrix was a 6-by-6 matrix with the rows representing the true expressions (ground truth) and the columns representing the expressions discriminated by participants. The element (i, j) of the confusion matrix indicated the ratio of how many times the expression i was recognized as the expression j across participants.

Analysis of network units

Each DCNN was presented with 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. Similar to Nasr *et al.* (*27*), a two-way nonrepeated ANOVA with expression (six facial expressions) and identity (104 identities) as factors was conducted to identify the 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 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 SD of the responses across all images in the stimulus set (*32*)

$$\mathrm{TV}_i^k = \frac{\frac{1}{P_k}\sum_{p\in k} A_i^p - \frac{1}{P}\sum_{p=1}^{P} A_i^p}{\sqrt{\frac{1}{P}\sum_{p=1}^{P}\left(A_i^p - \frac{1}{P}\sum_{p=1}^{P} A_i^p\right)^2}}$$

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, and 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.

To test the reliability of the expression recognition ability of expression-selective units in the pretrained VGG-Face, the SVC model trained on stimulus set 1 was used to predict the expressions of images from stimulus set 2. To further test the generality of the expression recognition ability of expression-selective units and the necessity of the domain-specific experience, for each DCNN, the SVC model trained on stimulus set 1 was used to predict the expressions of images from stimulus set 3. In the SVC model, the first 600 PCs of the responses of expression-selective units were used. The reasons for choosing the first 600 PCs were as follows: (i) the number of PCs should be less than the number of the images to avoid overfitting the training data, and (ii) to unify the DCNNs' component number, the number of PCs should also be no larger than the least number of the expression-selective units among the DCNNs (i.e., 644 units in the untrained VGG-Face). Meanwhile, the first 600 PCs could explain nearly 100% variance of the expression selective features (fig. S1). The prediction accuracy of the SVC model indicated the extent to which expression-selective units could correctly classify facial expressions. Further, predicted expressions and true expressions of the images were used to construct the confusion matrix. We then quantify the similarity of the error patterns between the expression-selective units and humans by calculating the pairwise Kendall rank correlation of the error rates (i.e., vectorized off-diagonal misclassification rates of the confusion matrices).

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*, *39*, *40*). Taking the 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 stimulus set 1) and 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 the 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 the goodness of fit (R^2) of the logistic function and the goodness of fit (R^2) of the linear function. Last, the Mann-Whitney U test was used to statistically evaluate the differences among the DCNNs.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at https://science.org/doi/10.1126/sciadv.abj4383

View/request a protocol for this paper from *Bio-protocol*.

REFERENCES AND NOTES

1. V. Bruce, A. Young, Understanding face recognition. *Br. J. Psychol.* **77**, 305–327 (1986).
2. V. Bruce, Influences of familiarity on the processing of faces. *Perception* **15**, 387–397 (1986).
3. A. J. Calder, J. Keane, A. W. Young, M. Dean, Configural information in facial expression perception. *J. Exp. Psychol. Hum. Percept. Perform.* **26**, 527–551 (2000).

4. A. W. Young, F. Newcombe, E. H. F. D. Haan, M. Small, D. C. Hay, Face perception after brain injury. Selective impairments affecting identity and expression. *Brain* **116**, 941–959 (1993).
5. G. W. Humphreys, N. Donnelly, M. J. Riddoch, Expression is computed separately from facial identity, and it is computed separately for moving and static faces: Neuropsychological evidence. *Neuropsychologia* **31**, 173–181 (1993).
6. B. C. Duchaine, H. Parker, K. Nakayama, Normal recognition of emotion in a prosopagnosic. *Perception* **32**, 827–838 (2003).
7. J. V. Haxby, E. A. Hoffman, M. I. I. Gobbini, The distributed human neural system for face perception. *Trends Cogn. Sci.* **4**, 223–233 (2000).
8. J. S. Winston, R. N. A. Henson, M. R. Fine-Goulden, R. J. Dolan, fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *J. Neurophysiol.* **92**, 1830–1839 (2004).
9. T. J. Andrews, M. P. Ewbank, Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *Neuroimage* **23**, 905–913 (2004).
10. C. J. Fox, H. M. Hanif, G. Iaria, B. C. Duchaine, J. J. S. S. Barton, Perceptual and anatomic patterns of selective deficits in facial identity and expression processing. *Neuropsychologia* **49**, 3188–3200 (2011).
11. A. S. Redfern, C. P. Benton, Expression dependence in the perception of facial identity. *Iperception* **8**, 2041669517710663 (2017).
12. T. Ganel, Y. Goshen-Gottstein, T. Ganel, Effects of familiarity on the perceptual integrality of the identity and expression of faces: The parallel-route hypothesis revisited. *J. Exp. Psychol. Hum. Percept. Perform.* **30**, 583–597 (2004).
13. A. Yankouskaya, P. Rotshtein, G. W. Humphreys, Interactions between identity and emotional expression in face processing across the lifespan: Evidence from redundancy gains. *J. Aging Res.* **2014**, 1–12 (2014).
14. A. M. V. Gerlicher, A. M. Van Loon, H. S. Scholte, V. A. F. Lamme, A. R. Van der Leij, Emotional facial expressions reduce neural adaptation to face identity. *Soc. Cogn. Affect. Neurosci.* **9**, 610–614 (2014).
15. H. A. Baseler, R. J. Harris, A. W. Young, T. J. Andrews, Neural responses to expression and gaze in the posterior superior temporal sulcus interact with facial identity. *Cereb. Cortex* **24**, 737–744 (2014).
16. D. L. K. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
17. P. Bashivan, K. Kar, J. J. DiCarlo, Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
18. S. Grossman, G. Gaziv, E. M. Yeagle, M. Harel, P. Mégevand, D. M. Groppe, S. Khuvis, J. L. Herrero, M. Irani, A. D. Mehta, R. Malach, Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 4934 (2019).
19. K. C. O'Nell, R. Saxe, S. Anzellotti, K. C. O. Nell, R. Saxe, S. Anzellotti, Recognition of identity and expressions as integrated processes. (PsyArXiv, 2019).
20. Y. I. Colón, C. D. Castillo, A. J. O'Toole, Facial expression is retained in deep networks trained for face identification. *J. Vis.* **21**, 4 (2021).
21. S. Baek, M. Song, J. Jang, G. Kim, S.-B. Paik, Spontaneous generation of face recognition in untrained deep neural networks. bioRxiv:857466 (2019).
22. D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (W.H. Freeman, 1982).
23. M. R. W. Dawson, *Mind, Body, World: Foundations of Cognitive Science* (Athabasca Univ. Press, 2013).
24. O. M. Parkhi, A. Vedaldi, A. Zisserman, Deep face recognition, in *Proceedings of the British Machine Vision Conference* (BMVA Press, 2015), pp. 41.1–41.12.
25. R. Yamashita, M. Nishio, R. K. G. Do, K. Togashi, Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **9**, 611–629 (2018).
26. G. Kim, J. Jang, S. Baek, M. Song, S.-B. Paik, Visual number sense in untrained deep neural networks. *Sci. Adv.* **7**, eabd6127 (2021).
27. K. Nasr, P. Viswanathan, A. Nieder, Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Sci. Adv.* **5**, eaav7903 (2019).
28. D. Lundqvist, A. Flykt, A. Öhman, The Karolinska directed emotional faces–KDEF, CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet (1998).
29. N. Tottenham, J. W. Tanaka, A. C. Leon, T. McCarry, M. Nurse, T. A. Hare, D. J. Marcus, A. Westerlund, B. J. Casey, C. Nelson, The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Res.* **168**, 242–249 (2009).
30. P. Ekman, An argument for basic emotions. *Cogn. Emot.* **6**, 169–200 (1992).
31. P. Ekman, D. Cordaro, What is meant by calling emotions basic. *Emot. Rev.* **3**, 364–370 (2011).
32. G. W. Lindsay, K. D. Miller, How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife* **7**, e38105 (2018).
33. O. Langner, R. Dotsch, G. Bijlstra, D. H. J. J. Wigboldus, S. T. Hawk, A. van Knippenberg, Presentation and validation of the Radboud Faces Database. *Cogn. Emot.* **24**, 1377–1388 (2010).
34. A. Mollahosseini, B. Hasani, M. H. Mahoor, AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.* **10**, 18–31 (2017).
35. A. J. Calder, A. M. Burton, P. Miller, A. W. Young, S. Akamatsu, A principal component analysis of facial expressions. *Vision Res.* **41**, 1179–1208 (2001).
36. N. L. Etcoff, J. J. Magee, Categorical perception of facial expressions. *Cognition* **44**, 227–240 (1992).
37. C. J. Fox, S. Y. Moon, G. Iaria, J. J. S. S. Barton, S. Young, G. Iaria, J. J. S. S. Barton, S. Y. Moon, G. Iaria, J. J. S. S. Barton, The correlates of subjective perception of identity and expression in the face network: An fMRI adaptation study. *Neuroimage* **44**, 569–580 (2009).
38. R. J. Harris, A. W. Young, T. J. Andrews, Morphing between expressions dissociates continuous from categorical representations of facial expression in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 21164–21169 (2012).
39. A. J. Calder, A. W. Young, D. I. Perrett, N. L. Etcoff, D. Rowland, Categorical perception of morphed facial expressions. *Vis. Cogn.* **3**, 81–118 (1996).
40. T. Fujimura, Y. T. Matsuda, K. Katahira, M. Okada, K. Okanoya, Categorical and dimensional perceptions in decoding emotional facial expressions. *Cogn. Emot.* **26**, 587–601 (2012).
41. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014).
42. X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks. *J. Mach. Learn. Res.* **9**, 249–256 (2010).
43. C. Zhou, W. Xu, Y. Liu, Z. Xue, R. Chen, K. Zhou, J. Liu, Numerosity representation in a deep convolutional neural network. *J. Pac. Rim Psychol.* **15**, 1–11 (2021).
44. P. Viswanathan, A. Nieder, Neuronal correlates of a visual "sense of number" in primate parietal and prefrontal cortices. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11187–11192 (2013).
45. L. Wagener, M. Loconsole, H. M. Ditz, A. Nieder, Neurons in the endbrain of numerically naive crows spontaneously encode visual numerosity. *Curr. Biol.* **28**, 1090–1094.e4 (2018).
46. M. S. Livingstone, J. L. Vincent, M. J. Arcaro, K. Srihasam, P. F. Schade, T. Savage, Development of the macaque face-patch system. *Nat. Commun.* **8**, 14897 (2017).
47. P. Bao, L. She, M. Mcgill, D. Y. Tsao, A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
48. E. Kotsoni, M. De Haan, M. H. Johnson, M. De Haanô, M. H. Johnson, M. De Haan, M. H. Johnson, M. De Haanô, M. H. Johnson, M. De Haan, M. H. Johnson, Categorical perception of facial expressions by 7-month-old infants. *Perception* **30**, 1115–1125 (2001).
49. J. M. Leppanen, J. Richmond, V. K. Vogel-farley, M. C. Moulson, C. A. Nelson, J. M. Leppänen, J. Richmond, V. K. Vogel-farley, M. C. Moulson, C. A. Nelson, Categorical representation of facial expressions in the infant brain. *Infancy* **14**, 346–362 (2009).
50. K. Hoemann, R. Wu, V. LoBue, L. M. Oakes, F. Xu, L. F. Barrett, Developing an understanding of emotion categories: Lessons from objects. *Trends Cogn. Sci.* **24**, 39–51 (2020).
51. V. Lee, J. L. Cheal, M. D. Rutherford, Categorical perception along the happy-angry and happy-sad continua in the first year of life. *Infant Behav. Dev.* **40**, 95–102 (2015).
52. E. H. Telzer, J. Flannery, M. Shapiro, K. L. Humphreys, B. Goff, L. Gabard-Durman, D. D. Gee, N. Tottenham, Early experience shapes amygdala sensitivity to race: An international adoption design. *J. Neurosci.* **33**, 13484–13488 (2013).
53. Y. Du, F. Zhang, Y. Wang, T. Bi, J. Qiu, Perceptual learning of facial expressions. *Vision Res.* **128**, 19–29 (2016).
54. J. M. Beale, F. C. Keil, Categorical effects in the perception of faces. *Cognition* **57**, 217–239 (1995).
55. K. A. Dalrymple, M. Visconti Di Oleggio Castello, J. T. Elison, M. I. Gobbini, Concurrent development of facial identity and expression discrimination. *PLOS ONE* **12**, e0179458 (2017).
56. R. J. Harris, A. W. Young, T. J. Andrews, Dynamic stimuli demonstrate a categorical representation of facial expression in the amygdala. *Neuropsychologia* **56**, 47–52 (2014).
57. B. P. Tripp, Similarities and differences between stimulus tuning in the inferotemporal visual cortex and convolutional networks, in *2017 International Joint Conference Neural Networks (IJCNN)* (IEEE, 2017), pp. 3551–3560.
58. H. Wen, J. Shi, Y. Zhang, K. H. Lu, J. Cao, Z. Liu, Neural encoding and decoding with deep learning for dynamic natural vision. *Cereb. Cortex* **28**, 4136–4160 (2018).
59. J. Kubilius, S. Bracci, H. P. Op de Beeck, Deep neural networks as a computational model for human shape sensitivity. *PLOS Comput. Biol.* **12**, 759 (2016).
60. L. K. Wenliang, A. R. Seitz, Deep neural networks for modeling visual perceptual learning. *J. Neurosci.* **38**, 6028–6044 (2018).
61. V. Willenbockel, J. Sadr, D. Fiset, G. O. Horne, F. Gosselin, J. W. Tanaka, Controlling low-level image properties: The SHINE toolbox. *Behav. Res. Methods* **42**, 671–684 (2010).

Acknowledgments: We thank J. Liu at the Tsinghua Laboratory of Brain & Intelligence for extremely helpful comments on our study. **Funding:** This work was supported by the National Science and Technology Innovation 2030 Major Program (2021ZD0203803), National Natural Science Foundation of China (31871136), and National Key R&D Program of China (2019YFA0709503). **Author contributions:** L.Z. and K.Z. contributed to the conceptualization and methodology. L.Z. designed and implemented the models. L.Z. and A.Y. contributed to the data collection and data analysis. L.Z., K.Z., and M.M. discussed all aspects of the analysis and figures. L.Z. wrote the original draft. K.Z. and M.M. contributed to the writing—review and editing and supervision of the project. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The data and code are archived in Zenodo at https://doi.org/10.5281/zenodo.5583295.

Submitted 12 May 2021
Accepted 2 February 2022
Published 23 March 2022
10.1126/sciadv.abj4383