Emergence of Emotion Selectivity in Deep Neural Networks Trained to Recognize Visual Objects

Authors: Peng Liu[1,2], Ke Bo[2], Mingzhou Ding[1]*, Ruogu Fang[1,3]*

[1]J. Crayton Pruitt Family Department of Biomedical Engineering, Herbert Wertheim College of Engineering, University of Florida, Gainesville, Florida, USA

[2]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, New Hampshire, USA

[3]Center for Cognitive Aging and Memory, McKnight Brain Institute, University of Florida, Gainesville, Florida, USA

*Co-corresponding authors: mding@bme.ufl.edu (MD) ruogu.fang@bme.ufl.edu (RF)

Abstract:

Recent neuroimaging studies have shown that the visual cortex plays an important role in representing the affective significance of visual input. The origin of these affect-specific visual representations is debated: they are intrinsic to the visual system versus they arise through reentry from frontal emotion processing structures such as the amygdala. We examined this problem by combining convolutional neural network (CNN) models of the human ventral visual cortex pre-trained on ImageNet with two datasets of affective images. Our results show that (1) in all layers of the CNN models, there were artificial neurons that responded consistently and selectively to neutral, pleasant, or unpleasant images and (2) lesioning these neurons by setting their output to 0 or enhancing these neurons by increasing their gain led to decreased or increased emotion recognition performance respectively. These results support the idea that the visual system may have the intrinsic ability to represent the affective significance of visual input and suggest that CNNs offer a fruitful platform for testing neuroscientific theories.

Author Summary:

What is the role played by sensory cortices in assessing the emotional significance of sensory input? This question is attracting increasing research interest. Recent work has found affect-specific neural representations in visual cortex. The origins of these representations are debated. According to the reentry hypothesis, these representations result from reentrant feedback arising from anterior emotion processing structures such as the amygdala. An alternative hypothesis holds that sensory cortex may have the intrinsic capacity to represent the emotional qualities of sensory input. We examined this problem by utilizing the convolutional neural networks (CNNs) trained to recognized visual objects as computational models of the primate ventral visual system. Emotionally charged images were divided into three broad categories (pleasant, neutral and unpleasant) and presented to the CNNs. Responses of artificial neurons to these images were found to exhibit robust emotion selectivity. Importantly, enhancing the neurons that were selective for a given emotion led to the increased ability in recognizing that emotion, whereas lesioning these neurons led to the decrease in that ability. This research lends support to the notion that emotional perception might be an intrinsic property of the visual cortex. It also underscores the CNNs' value in examining neuroscientific theories.

Introduction

Human emotions are complex and multifaceted and under the influence of many factors, including individual differences, cultural backgrounds, and the context in which the emotion is experienced (*1–5*). Still, a large number of people, across different cultures, different levels of education, and different socioeconomic backgrounds, experience similar feelings when viewing images of varying affective content (*6–9*). What fundamental principles in the functions of the human visual system underlie such universality requires elucidation.

Previous studies of emotion perception have primarily relied on empirical cognitive experiments (*10–12*). Some of them have focused on capturing human behavioral valence or arousal judgment on affective images (*13–16*), while others have recorded brain activities to look for neural correlates of affective stimuli processing (*17–21*). Despite decades of effort, how the brain transforms visual stimuli into subjective emotion judgments (e.g., happy, neutral, or unhappy) remains not well understood. The advent of machine learning especially artificial neural networks (ANNs) opens the possibility of addressing this problem using a modeling approach.

Artificial neural networks can project visual images to a feature space in which the activation patterns of hidden layers are the features used for object classification and recognition. One type of artificial neural network, convolutional neural networks (CNNs), owing to their hierarchical organization resembling that of the visual system, are increasingly used as models of visual processing in the primate brain (*22–26*). CNNs trained to recognize visual objects can achieve performance levels rivaling or even exceeding that of humans. Interestingly, CNNs trained on images from such databases as ImageNet (*27*) are found to demonstrate neural selectivity for a variety of stimuli that are not included in the training data. For instance, (*28*) showed that neurons in a CNN trained on ImageNet became selective for numbers without having been trained on any "number" datasets. Similarly, (*29*) demonstrated that a CNN, when trained on non-face objects, can develop a recognition performance for faces that significantly exceeds chance levels. These instances demonstrate that CNNs may possess recognition capabilities beyond the primary task they are trained on.

The role of the visual cortex in visual emotion processing is debated (*30*, *31*). (*32*) argued that emotion representation is an intrinsic property of the visual cortex. They used a CNN pre-trained on ImageNet to show that the model can accurately predict the emotion categories of affective images. (*20*), on the other hand, showed that the affective representations found in the visual cortex during affective scene processing might arise as the result of reentry from anterior emotion-modulating structures such as the amygdala. The goal of this study is to further examine this question using CNN models.

CNN models are well suited for addressing questions related to the human visual system. Among the many well-established CNN models, VGG-16 (*33*) has an intermediate level of complexity and is shown to have superior object recognition performance (*34*). Using VGG-16, recent cognitive neuroscience studies have explored how encoding and decoding of sensory information are hierarchically processed in the brain (*23*, *35*, *36*). (*23*) used VGG-16 to quantitatively demonstrate an explicit gradient of feature complexity encoded in the ventral visual pathway. (*35*) used VGG-16 to model the visual cortical activity of human participants viewing images of objects and demonstrated that activities in different layers of the model highly correlate with brain activities in different visual areas. (*36*) investigated qualitative similarities and differences between VGG-16 and other feed-forward CNNs in the representation of the visual object and showed these CNNs

exhibit multiple perceptual and neural phenomena such as the Thatcher effect (*37*) and Weber's law (*38*).

In this study, we mainly focused on VGG-16 pre-trained on ImageNet as the model of the human visual system and used AlexNet (*39*), which is another well-established CNN model of visual processing, to test whether the results can be replicated. Using two well-established affective image datasets: International Affective Picture System (IAPS) (*15*) and Nencki Affective Picture System (NAPS) (*16*), we examined whether emotion selectivity can spontaneously emerge in such systems and whether such emotion selectivity has functional significance. For each filter within a layer of the model, the emotional selectivity for the resulting feature map was established by first computing neural responses to three broad classes of images: pleasant, neutral, and unpleasant (tuning curves) at the level of each unit and then averaging these responses across all the units within the feature map. A feature map, also referred to as a neuron in what follows, is considered selective for a particular emotion if its tuning responses are robust and exhibit the strongest responses to images of that category from both datasets. To test whether these emotion-selective neurons have a functional role, we replaced the last 1000-unit object-recognition layer of the VGG-16 with a two-unit emotion-recognition layer and trained the connections to this layer to decode pleasant versus non-pleasant, neutral vs. non-neutral, and unpleasant vs. non-unpleasant images. Two neural manipulations were carried out: lesion and feature attention enhancements. Lesioning the neurons selective for a specific emotion is expected to degrade the network's performance in recognizing that emotion, whereas applying attention enhancement to the neurons selective for the emotion is expected to increase the network's performance in recognizing that emotion.

Results

We tested whether emotion selectivity can naturally arise in a CNN model trained to recognize visual objects. VGG-16 pre-trained on ImageNet data (*27*) was used for this purpose (see Fig 1). Filters/channels within a layer were referred to as neurons and responses from the units within the feature maps were averaged and treated as neuronal responses. Selectivity for pleasant, neutral, and unpleasant emotions was defined for each neuron based on its response profiles to images from two affective picture sets (IAPS and NAPS). The functional significance of these neurons was then assessed using lesion and attention enhancement methods.

Neuronal responses to emotional images in different convolutional layers

The tuning curve for a neuron is defined as the normalized mean response (tuning value) to pleasant, neutral, and unpleasant images in a given dataset plotted as a function of the emotion category. The maximum of the tuning curve indicates the neuron's preferred emotion category for that picture set. Fig 2A (top) shows the tuning curves of three neurons from the Convolutional Layer 3 (an early layer) for both IAPS and NAPS datasets. According to the definition above, these neurons are selective for pleasant, neutral, and unpleasant categories, respectively. For the top 100 images from IAPS and NAPS that elicited the strongest response in these neurons, Fig 2A (bottom) shows the valence distribution of these images. As can be seen, for these early layer neurons, while the pleasant neuron is more activated by images with high valence ratings (pleasant), for the neutral and unpleasant neurons, the patterns are less clear. For the neurons in Convolutional Layer 6 (a middle layer), however, as shown in Fig 2B, their emotion selectivity and the category of images they prefer show greater agreement. Namely, the pleasant neuron prefers predominately images with high valence (pleasant), the neural neuron prefers predominately images with intermediate valence (neutral), and the unpleasant neuron prefers predominately images with low valence

(unpleasant). The results for the three neurons from Convolutional Layer 13 (a deep layer) are similar to those from Layer 6; see Fig 2C.

Emotion selectivity in different convolutional layers

Whereas tuning value and tuning curve characterize a neuron's response to images from different emotion categories, the selectivity index (SI), which highlights the difference between responses to different emotion categories of images, is a better index for defining emotion selectivity. As shown in Fig 3A, emotion selectivity became stronger as one ascended the layers from early to deep, an effect that is especially noticeable for the IAPS datasets, supporting the notion that emotion differentiability increases as we go from earlier to deeper layers. In light of the computational principle that earlier layer neurons encode lower-level stimulus properties (e.g., shapes and edges) and deeper layer neurons encode higher-level properties such as semantic meaning (e.g., object identities) (*40–42*), the results in Fig 3A as well as Fig 2 suggest that from earlier to deeper layers, emotion as a higher level cognitive construct becomes progressively better defined and better differentiated.

To examine the role of the training to recognize objects in the foregoing observations, we performed the same analysis in a VGG-16 with randomly initialized weights (i.e., not trained to recognize objects). As seen in Fig 3A, emotion selectivity is generally low as evaluated by both datasets, and there is no clear layer-dependence in emotion selectivity, suggesting that the increased ability to represent and differentiate emotion in deeper network layers of the pre-trained VGG-16 is an ability acquired through the training for object recognition.

Generalizability of emotion-selective neurons

Fig 2 shows that a neuron can be tuned for the same emotion for both IAPS and NAPS datasets. A natural question is whether such neurons arise as the result of random chance or as an emergent property of the trained network. Further, based on the value of SI, all neurons are selectivity for one emotion or the other. Small SIs are likely subject to the influence of chance, and as such, neurons with small SIs should be removed from further consideration. How to determine the threshold for removal?

We performed two analyses to address the two questions. First, we rank-ordered neurons according to their SI values, removed certain percentages of neurons with small SI values, and attention-enhanced the remaining neurons (see next subsection) and observed the resulting performance improvement. The results in Fig 3C suggest that removing neurons whose SIs fell in the lower 20% (keeping 80%) is a reasonable threshold. Second, neurons determined to be emotion selective according to IAPS and that according to NAPS were subjected to an overlap analysis. Fig 3B (top) compares the number of neurons selective for the same emotion for both IAPS and NAPS datasets against the number of neurons to be expected from the overlap of two random sets of neurons. The former is consistently higher than the latter across all layers, with the effect becoming more prominent in deeper layers, suggesting that emotion selectivity generalizes across the two datasets and the generalizability is not due to chance.

What is the role of training to recognize visual objects in the generalizable emotion selectivity? To answer this question, we compared the number of emotion-selective neurons from the overlap analysis derived from pre-trained VGG-16 on ImageNet against that derived from randomly initialized VGG-16. Fig 3B (bottom) shows that for all emotion categories—pleasant, neutral, and unpleasant—the pre-trained network consistently demonstrated a higher number of emotion-

selective neurons in the later layers, especially from Layer 5 onwards. These findings suggest that emotion selectivity is an emergent property as the result of a neural network undergoing training for object recognition.

The functionality of emotion-selective neurons

To test whether emotion-selective neurons have a functional role, we followed (*43*) and replaced the last layer of the VGG-16, which originally contained 1,000 units for recognizing 1000 different types of objects, with a fully connected layer containing two units for recognizing two types of emotions. Three models were trained and tested for each of the two affective picture datasets: Model 1: pleasant versus non-pleasant, Model 2: neutral versus non-neutral, and Model 3: unpleasant versus non-unpleasant. Once these models were shown to have adequate emotion recognition performance (see Table 1), two neural manipulations were considered: feature attention enhancement and lesion. For feature attention enhancement (*44–46*), the gain of the neurons selective for a given emotion for both datasets is increased by increasing the slope of the ReLU activation function (see Methods) (*47–50*), whereas for lesion, the output of the neurons selective for a given emotion for both datasets was set to 0, which effectively removes the contribution of these neurons, i.e., they are lesioned. We hypothesized that (1) with attention enhancement, the network's ability to recognize emotion is increased (2) with lesioning, the network's ability to recognize emotion is decreased, and (3) such effects are not observed for modulating randomly selected neurons.

Feature attention enhancement: For IAPS images, Fig 4 compares performance changes after enhancing the emotion-selective neurons as well as enhancing the same number of randomly sampled neurons; see also Table 1. The optimal tuning strength for which we achieved the best performance enhancement was chosen for each layer in the plot. As one can see, for pleasant versus non-pleasant, neutral versus non-neutral, and unpleasant versus non-unpleasant emotions, enhancing the gain of the neurons selective for a specific emotion can significantly improve the emotion recognition performance of the CNN model for that emotion. Moreover, deeper layer attention enhancement tends to yield greater performance improvements than earlier layer attention enhancement. Increasing the gain in randomly selected neurons, however, shows either a marginal performance improvement or a significant performance decline. The feature-attention performance of emotion-selective neurons over random neurons is highly statistically significant in the middle and deeper layers ($p < 1.2e-02$). Fig 4 (right) shows the performance changes across layers as the tuning strength varied from 0 to 5.

We carried out the same analysis for the NAPS dataset in Fig 5. The results largely replicated that in Fig 4 for the IAPS dataset.

Lesion analysis: The functional importance of the emotion-selective neurons can be further assessed through lesion analysis (*51–54*). As shown in Fig 6 (see also Table 1), we compared the emotion recognition performance changes by setting the output from emotion-selective neurons to 0 as well as by setting the output of an equal number of randomly chosen neurons to 0. As can be seen, lesioning the emotion-selective neurons led to significant performance declines, especially for the deeper layers; the performance decline can be as high as 80%. In contrast, lesioning randomly selected neurons produces almost no performance changes. These results, replicated across both datasets, further support the hypothesis that emotion-selective neurons are important for emotion recognition, and the importance is higher in deeper layers than in earlier layers.

Discussion

It has been argued that the human visual system has the intrinsic ability to recognize the motivational significance of environmental inputs (*55*). We examined this problem using convolutional neural networks (CNNs) as models of the human visual system (*56–61*). Selecting the VGG16 pre-trained on images from the ImageNet as our model (*62–64*) and using two sets of affective images (IAPS and NAPS) as test stimuli, we found the existence of emotion-selective neurons in all layers of the model even though the model has never been explicitly trained to recognize emotion. Additionally, emotion selectivity becomes stronger and more consistent in the deeper layers, in agreement with prior literature suggesting that the deeper layers of CNNs encode higher-level semantic information. For VGG-16 with randomly initialized weights (i.e., not trained to recognize objects), however, no such effects were observed, suggesting that emotion selectivity may be an emergent property through network training. Applying two manipulations: feature attention enhancement and lesion, we can show further that the emotion-selective neurons are functionally significant, specifically: (1) after increasing the gain of emotion-selective neurons (e.g., feature attention enhancement), the network's performance in emotion recognition is enhanced relative to increasing the gain of randomly selected neurons and (2) in contrast, after lesioning the emotion-selective neurons, the network's performance in emotion recognition is degraded relative to lesioning randomly selected neurons. These performance differences are stronger and more noticeable in deeper layers than in earlier layers. In Fig F, Fig G, Fig H, and Fig I in S1 Text, we reported similar findings on the AlexNet, which is a simpler CNN that has also been used in numerous studies as a model of the ventral visual system (*65–68*). Together, our findings indicate that emotion selectivity can spontaneously emerge in CNN models trained to recognize visual objects, and these emotion-selective neurons play a significant role in recognizing emotion in natural images, lending credence to the notion that the visual system's ability to represent affective information may be intrinsic.

Affective processing in the visual cortex

The perception of opportunities and threats in complex visual scenes represents one of the main functions of the human visual system. The underlying neurophysiology is often studied by having observers view pictures varying in affective content. (*69*) reported greater functional activity in the visual cortex when subjects viewed pleasant and unpleasant pictures than neutral images. (*70*) showed the visual cortex has differential sensitivities in response to emotional stimuli compared to the amygdala. (*71*) demonstrated that emotional significance (e.g., valence or arousal) could modulate the perceptual encoding in the visual cortex. Two competing but not mutually exclusive groups of hypotheses have been advanced to account for emotion-specific modulations of activity in the visual cortex. The so-called reentry hypothesis states that the increased visual activation evoked by affective pictures results from reentrant feedback, meaning that signals arising in subcortical emotion processing structures such as the amygdala propagate to the visual cortex to facilitate the processing of motivationally salient stimuli (*72–74*). Recent work (*20*) provides support for this view. Using multivariate pattern analysis and functional connectivity, these authors showed that (1) different emotion categories (e.g., pleasant versus neutral and unpleasant versus neutral) are decodable based on the multivoxel patterns in the visual cortex and (2) the decoding accuracy is positively associated with the strength of connectivity from anterior emotion-modulating regions to ventral visual cortex. A second group of hypotheses states that the visual cortex may itself have the ability to code for the emotional qualities of a stimulus, without the necessity for recurrent processing (see (*75*) for a review). Evidence supporting this hypothesis comes from empirical studies in experimental animals (*76*, *77*) as well as in human observers (*78*), in which the extensive pairing of simple sensory cues such as tilted lines or sinusoidal gratings with emotionally relevant outcomes shapes early sensory responses (*79*). Beyond simple visual cues, recent computational work using deep neural networks has also suggested that the visual cortex

may intrinsically represent emotional value as contained in complex visual media such as video clips of varying affective content (32). Our findings reveal that emotion-selective neurons are present in all layers of two CNN models, which are computational representations of the visual cortex. These neurons play a crucial role in emotion recognition. This contributes to the growing computational evidence suggesting that the visual cortex may inherently possess the capability to evaluate the emotional significance of visual stimuli.

Neural selectivity in ANNs and the brain

That CNNs, or more generally ANNs, can be trained to recognize a large variety of visual objects has long been recognized. Remarkably, recent studies note that ANNs trained on recognizing visual objects can spontaneously develop selectivity for other types of input, including visual numbers and faces (*80*). The number sense is considered an inherent ability of the brain to estimate the quantity of certain items in a visual set (*81*, *82*). There is significant evidence demonstrating that the number sense exists in both humans (e.g., adults and infants) (*83–85*) and non-human primates (e.g., numerically naïve monkeys) (*86–88*). (*89*) found that number-selective units spontaneously emerged in a deep artificial neural network trained on ImageNet for object recognition. (*90*) demonstrated that number selectivity can even arise spontaneously in randomly initialized deep neural networks without any training. Both studies focused on the last convolutional layers, in which the number-selective units were found, and they also demonstrated that the emergence of number-selective units could result from the weighted sum of both decreasing and increasing the activity of some units. In addition, it is well known that face-selective neurons exist in humans (*91*) and non-human primates. (*80*) showed that neurons in a randomly initialized deep neural network without training could selectively respond to faces, and the neurons in the deeper layers are more selective. (*92*) demonstrated that brain-like functional segregation can emerge spontaneously in deep neural networks trained on object recognition and face perception and proposed that the development of functional segregation of face recognition in the brain is a result of computational optimization in the cortex. Augmenting this rapidly growing literature, our study demonstrates that emotion selectivity can emerge in deep artificial neural network models of the human visual system trained to recognize objects.

Layer dependence

Like the biological brain, the CNN model has a layered structure which allows the processing of information in a hierarchical fashion. Our layer-wise analysis showed that the extent and strength of emotion selectivity are a function of the model layers. Compared to the early layers, the deeper layers have larger portions of neurons that show emotion selectivity, and the selectivity is stronger, consistent with the previous observations that deeper layers of CNN models encode more abstract concepts. For example, (*40*, *93*) examined the internal representations of different layers in a CNN and found that deeper layers of the network tend to encode more abstract concepts, such as object parts and textures. The layered processing of emotional information may have several functional benefits. First, by processing visual information in hierarchical stages, the brain can quickly and efficiently respond to stimuli without the need for a complete and detailed analysis of the entire stimulus at once (*94–96*). This is especially important for the processing of emotionally salient stimuli, as quick and accurate emotional responses can be crucial for survival. Second, it would offer more flexibility for the processing of emotion at different levels of detail, which may depend on the perception task and the environmental context. For example, if the stimulus is perceived as significant or crucial for survival, it elicits a stronger and more widespread neural response, engaging multiple regions and processing stages. On the other hand, if the stimulus is not significant, it elicits a weaker and more limited neural response involving fewer regions or layers

and processing stages (*97–99*). Third, the integration of information from different levels allows for a more complete and nuanced representation of the visual stimulus and emotional response. This allows for the creation of a final representation that takes into account not just the visual properties of the stimulus but also its emotional significance and its impact on the individual (*100–102*). Lastly, by processing information in a layer-dependent manner, the brain can adapt and change the processing of information based on experience and learning (*103*). This allows the brain to refine its processing strategies and improve its performance over time (*104*).

Relation to prior literature

(*32*), to the best of our knowledge, is the first to examine emotion processing in deep neural networks. Their model, which is a modified AlexNet called the EmoNet, was shown to have the ability to classify affective images into 20 different emotion categories. Importantly, using a 20-way linear decoder, they further showed that neural activities in different layers of the network especially the deeper layers can differentiate different emotions in the input images, suggesting the existence of emotion selectivity neurons in CNNs. Building on this work, our main contributions are threefold: (1) confirming and characterizing emotion selectivity at the single filter (neuron) level, (2) demonstrating the functional significance of emotion-selective neurons through the application of lesion and attention enhancement methods, and (3) replicating the findings across two CNN models (VGG-16 and AlexNet) and two affective image sets (IAPS and NAPS).

Limitations and other considerations

Several limitations of our study should be noted. Firstly, emotion was divided into three broad categories: pleasant, unpleasant, and neutral. While this is in line with many neurophysiological studies in humans, future work should examine finer differentiations of emotion, e.g., joy, sadness, horror, disgust, and so on, and their neural representations in the brain. Secondly, there might be other factors (e.g., low-level features) that drive the emotion selectivity of neurons. Since we used grayscale images in this study, we can rule out color as a possible confounding low-level feature. Applying the GIST algorithm (*105*) to extract low-level features from images and the support vector machine (SVM) algorithm (*106*), we found that images from different emotion categories cannot be decoded from the low-level features; see Fig J in S1 Text. The impact of an image's object category and its emotion category on neural activation was examined by placing images in the IAPS and NAPS datasets into object categories based on the descriptions of the images (Fig LA and Fig MA in S1 Text) and applying Two-Way ANOVA tests to filter activations in the VGG-16 model. We found that the neurons responded more strongly to emotion categories than object categories and there were significant interactions between the two categories in deeper layers (Fig LB and Fig MB in S1 Text). We do note that, as the number of images in different object categories are relatively small in both affective datasets, this analysis should be viewed as preliminary. The influence of other factors such as the presence of faces and image animacy is more difficult to ascertain. Thirdly, although the present study is motivated by neuroscience questions, to what extent our results have a direct bearing on understanding brain function is unclear. Whereas previous work did compare activities in VGG-16 and other deep neural networks with neural recordings during object recognition (*67*, *107–109*), there is no study to date comparing activities in deep neural networks and neural recordings during emotion recognition. In this sense, this work's neural relevance should be considered speculative.

Materials and Methods

Affective picture sets

Two sets of widely used affective images were used in this study. The IAPS library includes 1,182 images covering approximately 20 subclasses of emotions such as joy, surprise, entrancement, sadness, romance, disgust, and fear. The NAPS library has 1,356 images that can be divided into similar subclasses. For both libraries, each image has a normative valence rating, ranging from 1 to 9, indicating whether the image expresses unpleasant, neutral, or pleasant emotions; the distributions of the valence rating from the two datasets were given in Fig AC(right) in S1 Text. In this study, for simplicity and following the common practice in human imaging studies of emotion (*20*, *110–112*), we classified images into three main categories based on their valence scores: "pleasant," "neutral," and "unpleasant." For images that fell near the boundary between categories, we used soft thresholds of 4.3±0.5 and 6.0±0.5 to determine their classification as either "unpleasant" or "neutral," or "neutral" or "pleasant." We also visually examined each image to confirm its category. Finally, any images that we could not confidently classify were marked as "unknown" and removed from the analysis. This process resulted in some differences in the number of images in each category from the original datasets. After this categorization, IAPS images were divided into 296 pleasant, 390 neutral, and 341 unpleasant images, and NAPS images into 352 pleasant, 477 neutral, and 281 unpleasant images (see Fig AB in S1 Text). These images were transformed from the original color images to grayscale images prior to the commencement of the study reported here. The goal was to remove color as a possible low-level visual feature confounding the emotion selectivity analysis.

The convolutional neural network model

VGG-16, a well-tested deep convolutional neural network for natural image recognition, was used in this study to evaluate emotion selectivity. It has 13 convolutional layers followed by three fully connected layers, with the last fully connected layer containing 1000 units for recognizing 1000 different types of visual objects. Each layer of VGG-16 contains a large number of filters/channels, the application of each of which results in a feature map consisting of a large number of units. For convenience, and to stress neurobiological relevance, these filters/channels were often referred to as artificial neurons or simply neurons in this paper. Each neuron is characterized by a ReLU activation function (see Fig A in S1 Text). Through this function, neurons within a given layer, upon receiving and processing the input from the previous layer, yield activation maps (i.e., feature maps) which become the input for the next layer. Previous studies have compared the activation patterns of the VGG-16 model with experimental recordings from both humans and non-human primates and found that early layers of the model behave similarly to early visual areas such as V1, whereas deeper layers of the model are more analogous to higher-order visual areas such as the object-selective lateral occipital areas (*22*, *113–115*).

In this study, VGG-16 was used in two ways. First, to examine whether emotional selectivity emerges in neurons trained to recognize objects, we took the VGG-16 model pre-trained on 1.2 million natural images from the ImageNet, presented affective pictures from the two aforementioned affective picture datasets to the model, and analyzed the activation profiles of neurons from each layer. The emotional selectivity of each neuron was determined from these activation profiles (see below). Second, to test the functionality of the emotion-selective neurons, we replaced the last layer of the VGG-16 with a two-unit fully connected layer and trained the connections to this two-unit layer to recognize two categories of emotion: pleasant versus non-pleasant, neutral versus non-neutral, or unpleasant versus non-unpleasant. The training of the last two-unit emotion recognition layer used cross-entropy as the objective function. It is worth noting that, aside from the last emotion-recognition layer, the other layers' weights in the VGG-16 network remained the same as that trained on the ImageNet data; in other words, they were frozen.

The training data and the testing data for the final 2-unit emotion recognition layer of our model were separate for IAPS and NAPS to avoid overfitting. Specifically, for each emotion category, we partitioned the images from both datasets into training, validation, and testing subsets at a ratio of 50%:25%:25%. We used a learning rate of $1e-3$, trained for 10 epochs, and set the batch size to 128. Finally, we employed the F1-score to assess the performance of our model in emotion recognition.

Emotion selectivity definition

We used two methods to evaluate the differential responses of a neuron to images from different emotion categories (pleasant, neutral, or unpleasant). Tuning value emphasizes the normalized response to images from the same category. It is used in Fig 2 to illustrate possible response profiles or tuning curves of different neurons. The selective index (SI), in contrast, emphasizes the difference between responses to images from one emotion category and those from other emotion categories. It is thus more suitable for quantifying the emotion selectivity of a neuron. Results reported in Figs 3 and 4 as well as in Fig F, Fig G, Fig H, and Fig I in S1 Text were done with the SI.

Tuning value calculation: We followed the method in (*43*) for calculating the tuning value in Fig 2. The tuning v focuses on the strength or magnitude of a neuron's response to a particular emotion, relative to its average response. The details can be found below.

The output from each filter also referred to as a neuron in this study, see Fig A in S1 Text, can be written as:

$$x^{lk} = (1+\alpha)\max[0, w^{lk} \times x^{l-1}] \quad (1)$$

where w^{lk} indicates the weights of the k^{th} filter in the l^{th} convolutional layer, and $*$ indicates mathematical convolution which applies matrix multiplication between w and the outputs X from the $(l-1)^{th}$ layer. Of note in Eq. (1) is that the ReLU activation function typically has a slope of 1 ($\alpha = 0$). Here in this work, the slope is a tunable parameter. By tuning the slope of the ReLU function, we change the gain of the neuron, simulating the effect of feature-based attention control (*43*, *53*).

Let $X_{i,j}^{lk}(n)$ represents the response of the unit located at coordinates (i,j) in the k^{th} filter in layer l to image n. Then

$$\bar{p}^{lk}(n) = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H} X_{i,j}^{lk}(n) \quad (2)$$

is the response to the image averaged across the entire filter. Here W and H represent the width and height of the feature map. Thus, the mean activity of the filter k in layer l in response to all images in a dataset can be formulated as:

$$\hat{p}^{lk} = \frac{1}{N}\sum_{n=1}^{N} \bar{p}^{lk}(n) \quad (3)$$

where N represents the total number of images in a given set. The tuning value of the filter is calculated according to

$$S_e^{lk} = \frac{\frac{1}{N_e}\sum_{n=1}^{N_e} \bar{p}^{lk}(n) - \hat{p}^{lk}}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\bar{p}^{lk}(n) - \hat{p}^{lk}\right)^2}} \qquad (4)$$

where S_e^{lk} represents the normalized activation of filter k in layer l in response to all images of emotion category e, where $e \in \{pleasant, neutral, unpleasant\}$. A neuron is considered selective for a specific emotion if the normalized activation for the images within that emotion category is highest among the three possible values. For example, if $S_{e=pleasant}^{lk}$=-0.1, $S_{e=neutral}^{lk} =$ 0.2, and $S_{e=unpleasant}^{lk}$=0.3, the artificial neuron k is considered selective for "unpleasant images".

Selectivity index calculation: Selectivity Index (SI) (*116*) is defined as follows. First, consider

$$d'\ (pleasant) = \frac{\bar{X}_{pleasant} - \frac{\bar{X}_{neutral} + \bar{X}_{unpleasant}}{2}}{\sqrt{\frac{\sigma^2_{pleasant} + \sigma^2_{neutral} + \sigma^2_{unpleasant}}{2}}}$$

$$d'\ (neutral) = \frac{\bar{X}_{neutral} - \frac{\bar{X}_{pleasant} + \bar{X}_{unpleasant}}{2}}{\sqrt{\frac{\sigma^2_{neutral} + \sigma^2_{pleasant} + \sigma^2_{unpleasant}}{2}}} \qquad \textbf{(5)}$$

$$\boldsymbol{d'\ (unpleasant) = \frac{\bar{X}_{unpleasant} - \frac{\bar{X}_{pleasant} + \bar{X}_{neutral}}{2}}{\sqrt{\frac{\sigma^2_{unpleasant} + \sigma^2_{pleasant} + \sigma^2_{neutral}}{2}}}}$$

where $X_{pleasant}$, $X_{neutral}$, and $X_{unpleasant}$ represents the mean response to the pleasant, neutral, and unpleasant categories, respectively; $\sigma^2_{pleasant}$, $\sigma^2_{neutral}$, and $\sigma^2_{unpleasant}$ represents the variance of the response to the pleasant, neutral, and unpleasant category, respectively. SI is the largest d' and the emotion that gives rise to the largest d' defines the emotion for which the neuron is selective.

Identification of emotion-selective neurons: To guard against spurious identification of emotion selectivity and ensure that neurons designated to be selective for an emotion do so for both datasets, we applied two analyses. First, we rank-ordered neurons according to their SI values, eliminated neurons with small SI values, and tested the emotion recognition performance under attention enhancement of the remaining neurons (see below). Increasing the percentage of neurons eliminated until we saw a significant change in performance. That percentage was then defined as the threshold for defining emotion selectivity within a dataset (see Fig 3C for an example of finding the threshold for the pleasant category on the IAPS dataset). Second, for neurons identified as

selective for certain emotions based on IAPS and that based on NAPS, we overlapped the two sets of neurons and considered the overlapped neurons to be the genuine emotion-selective neurons.

Testing the functionality of the emotion-selective neurons

Do the emotion-selective neurons defined above have a functional role? We applied two different approaches to examine this question: lesion and attention enhancement.

Lesion. If the emotion-selective neurons are functionally important, then lesioning these neurons should lead to degraded performance in recognizing the emotion of a given image. Here the lesion of a specific neuron is achieved by setting its output to 0 (namely, setting $\alpha = -1$ in Eq. (1)). In our experiments, we lesioned the neurons selective for a given emotion as well as randomly selected neurons in a particular layer and observed the changes in the emotion recognition performance of the model.

Attention enhancement. We further tested whether enhancing the activity of an emotion-selective neuron can lead to performance improvement in emotion recognition. Following (*43*), the strength of α was increased from 0 to 5 with interval step size 0.1, where $\alpha = 0$ is the conventional choice and $\alpha > 0$ represents increased neuronal gain (i.e., enhanced feature attention). According to the feature similarity gain theory, increasing the gain of a neuron leads to enhanced performance of the neuron in perceiving stimuli with the relevant features. In our experiments, we enhanced the neurons selective for a given emotion as well as randomly selected neurons in a particular layer and observed the changes in the emotion recognition performance of the model (*43*) (see Fig BA and Fig BB in S1 Text).

References

1. S. Kitayama, *Emotion and Culture: Empirical Studies of Mutual Influence* (American Psychological Association, Washington, DC, US, 1994).
2. E. D. McCarthy, The Social Construction of Emotions: New Directions from Culture Theory. *Sociology Faculty Publications* (1994).
3. S. J. Banks, K. T. Eddy, M. Angstadt, P. J. Nathan, K. L. Phan, Amygdala–frontal connectivity during emotion regulation. *Social Cognitive and Affective Neuroscience* 2, 303–312 (2007).
4. J. J. Gross, "Emotion regulation: Conceptual and empirical foundations" in *Handbook of Emotion Regulation, 2nd Ed* (The Guilford Press, New York, NY, US, 2014), pp. 3–20.
5. L. F. Barrett, M. Lewis, J. M. Haviland-Jones, *Handbook of Emotions* (Guilford Publications, 2016; https://books.google.com/books?id=cbKhDAAAQBAJ).
6. H. A. Elfenbein, N. Ambady, On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychol Bull* 128, 203–235 (2002).
7. S. Hareli, K. Kafetsios, U. Hess, A cross-cultural study on emotion expression and the learning of social norms. *Frontiers in Psychology* 6 (2015).
8. B. Q. Ford, I. B. Mauss, Culture and emotion regulation. *Curr Opin Psychol* 3, 1–5 (2015).
9. S. Olderbak, O. Wilhelm, Emotion perception and empathy: An individual differences test of relations. *Emotion* 17, 1092–1106 (2017).
10. R. S. Lazarus, *Emotion and Adaptation* (Oxford University Press, 1991).

11. J. A. Coan, *Handbook of Emotion Elicitation and Assessment* (Oxford University Press, New York, NY, US, 2007)*Handbook of emotion elicitation and assessment*.

12. V. LoBue, Behavioral evidence for a continuous approach to the perception of emotionally valenced stimuli. *Behavioral and Brain Sciences* 38, e79 (2015).

13. M. K. Greenwald, E. W. Cook, P. J. Lang, Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli. *Journal of Psychophysiology* 3, 51–64 (1989).

14. M. M. Bradley, P. J. Lang, Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 49–59 (1994).

15. P. Lang, International affective picture system (IAPS) : affective ratings of pictures and instruction manual. *undefined* (2005).

16. A. Marchewka, Ł. Żurawski, K. Jednoróg, A. Grabowska, The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database. *Behav Res* 46, 596–610 (2014).

17. T. Canli, Z. Zhao, J. E. Desmond, E. Kang, J. Gross, J. D. E. Gabrieli, An fMRI study of personality influences on brain reactivity to emotional stimuli. *Behavioral Neuroscience* 115, 33–42 (2001).

18. P. Vrticka, S. Simioni, E. Fornari, M. Schluep, P. Vuilleumier, D. Sander, Neural Substrates of Social Emotion Regulation: A fMRI Study on Imitation and Expressive Suppression to Dynamic Facial Signals. *Frontiers in Psychology* 4 (2013).

19. M. Résibois, P. Verduyn, P. Delaveau, J.-Y. Rotgé, P. Kuppens, I. Van Mechelen, P. Fossati, The neural basis of emotions varies over time: different regions go with onset- and offset-bound processes underlying emotion intensity. *Social Cognitive and Affective Neuroscience* 12, 1261–1271 (2017).

20. K. Bo, S. Yin, Y. Liu, Z. Hu, S. Meyyappan, S. Kim, A. Keil, M. Ding, Decoding Neural Representations of Affective Scenes in Retinotopic Visual Cortex. *Cerebral Cortex* 31, 3047–3063 (2021).

21. H. Saarimäki, Naturalistic Stimuli in Affective Neuroimaging: A Review. *Frontiers in Human Neuroscience* 15 (2021).

22. D. L. 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. *Proceedings of the National Academy of Sciences* 111, 8619–8624 (2014).

23. U. Güçlü, M. A. J. van Gerven, Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J. Neurosci.* 35, 10005–10014 (2015).

24. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat Neurosci* 19, 356–365 (2016).

25. A. H. Marblestone, G. Wayne, K. P. Kording, Toward an Integration of Deep Learning and Neuroscience. *Front. Comput. Neurosci.* 10 (2016).

26. B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, K. P. Kording, A deep learning framework for neuroscience. *Nat Neurosci* 22, 1761–1770 (2019).

27. J. Deng, W. Dong, R. Socher, L. Li, Kai Li, Li Fei-Fei, "ImageNet: A large-scale hierarchical image database" in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255.

28. K. Nasr, P. Viswanathan, A. Nieder, Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances* 5, eaav7903 (2019).

29. K. Dobs, A. Kell, J. Martinez, M. Cohen, N. Kanwisher, N. Kanwisher, Why Are Face and Object Processing Segregated in the Human Brain? Testing Computational Hypotheses with Deep Convolutional Neural Networks (2020).

30. P. Vuilleumier, M. P. Richardson, J. L. Armony, J. Driver, R. J. Dolan, Distant influences of amygdala lesion on visual cortical activation during emotional face processing. *Nat Neurosci* 7, 1271–1278 (2004).

31. M. G. Shuler, M. F. Bear, Reward Timing in the Primary Visual Cortex. *Science* 311, 1606–1609 (2006).

32. P. A. Kragel, M. C. Reddan, K. S. LaBar, T. D. Wager, Emotion schemas are embedded in the human visual system. *Science Advances* 5, eaaw4358 (2019).

33. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (2015).

34. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* 115, 211–252 (2015).

35. K. Seeliger, M. Fritsche, U. Güçlü, S. Schoenmakers, J.-M. Schoffelen, S. E. Bosch, M. A. J. van Gerven, Convolutional neural network-based encoding and decoding of visual object recognition in space and time. *NeuroImage* 180, 253–266 (2018).

36. G. Jacob, R. T. Pramod, H. Katti, S. P. Arun, Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat Commun* 12, 1872 (2021).

37. P. Thompson, Margaret Thatcher: A New Illusion. *Perception* 9, 483–484 (1980).

38. P. T. Sowden, "Psychophysics" in *APA Handbook of Research Methods in Psychology, Vol 1: Foundations, Planning, Measures, and Psychometrics* (American Psychological Association, Washington, DC, US, 2012)*APA handbooks in psychology*®, pp. 445–458.

39. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90 (2017).

40. M. D. Zeiler, R. Fergus, "Visualizing and Understanding Convolutional Networks" in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars, Eds. (Springer International Publishing, Cham, 2014), pp. 818–833.

41. G. Lee, Y.-W. Tai, J. Kim, Deep Saliency with Encoded Low level Distance Map and High Level Features. *arXiv:1604.05495 [cs]* (2016).

42. G. W. Lindsay, Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 1–15 (2020).

43. G. W. Lindsay, K. D. Miller, How biological attention mechanisms improve task performance in a large-scale visual system model. *eLife* 7, e38105 (2018).

44. J. H. R. Maunsell, S. Treue, Feature-based attention in visual cortex. *Trends in Neurosciences* 29, 317–322 (2006).

45. G. W. Lindsay, Feature-based Attention in Convolutional Neural Networks. *arXiv:1511.06408 [cs]* (2015).

46. C.-H. Yeh, M.-H. Lin, P.-C. Chang, L.-W. Kang, Enhanced Visual Attention-Guided Deep Neural Networks for Image Classification. *IEEE Access* 8, 163447–163457 (2020).

47. J. A. Cardin, L. A. Palmer, D. Contreras, Cellular mechanisms underlying stimulus-dependent gain modulation in primary visual cortex neurons in vivo. *Neuron* 59, 150–160 (2008).

48. E. Eldar, J. D. Cohen, Y. Niv, The effects of neural gain on attention and learning. *Nat Neurosci* 16, 1146–1153 (2013).

49. S. Jarvis, K. Nikolic, S. R. Schultz, Neuronal gain modulability is determined by dendritic morphology: A computational optogenetic study. *PLOS Computational Biology* 14, e1006027 (2018).

50. H. Bos, A.-M. Oswald, B. Doiron, Untangling stability and gain modulation in cortical circuits with multiple interneuron classes. bioRxiv [Preprint] (2020). https://doi.org/10.1101/2020.06.15.148114.

51. R. Aharonov, L. Segev, I. Meilijson, E. Ruppin, Localization of Function via Lesion Analysis. *Neural Computation* 15, 885–913 (2003).

52. L. J. Chareyron, D. G. Amaral, P. Lavenex, Selective lesion of the hippocampus increases the differentiation of immature neurons in the monkey amygdala. *Proceedings of the National Academy of Sciences* 113, 14420–14425 (2016).

53. G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, X.-J. Wang, Task representations in neural networks trained to perform many cognitive tasks. *Nat Neurosci* 22, 297–306 (2019).

54. S. Cohen-Zimerman, H. Khilwani, G. N. L. Smith, F. Krueger, B. Gordon, J. Grafman, The neural basis for mental state attribution: A voxel-based lesion mapping study. *Human Brain Mapping* 42, 65–79 (2021).

55. P. J. Lang, M. M. Bradley, B. N. Cuthbert, "Motivated attention: Affect, activation, and action" in *Attention and Orienting: Sensory and Motivational Processes* (Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 1997), pp. 97–135.

56. N. Kriegeskorte, Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science* 1, 417–446 (2015).

57. A. Brachmann, E. Barth, C. Redies, Using CNN Features to Better Understand What Makes Visual Artworks Special. *Frontiers in Psychology* 8 (2017).

58. K. Iigaya, S. Yi, I. A. Wahle, K. Tanwisuth, J. P. O'Doherty, Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features. *Nat Hum Behav* 5, 743–755 (2021).

59. L. E. van Dyck, R. Kwitt, S. J. Denzler, W. R. Gruber, Comparing Object Recognition in Humans and Deep Convolutional Neural Networks—An Eye Tracking Study. *Frontiers in Neuroscience* 15 (2021).

60. J. J. D. Singer, K. Seeliger, T. C. Kietzmann, M. N. Hebart, From photos to sketches - how humans and deep neural networks process objects across different levels of visual abstraction. *Journal of Vision* 22, 4 (2022).

61. J. Lee, M. Jung, N. Lustig, J.-H. Lee, Neural representations of the perception of handwritten digits and visual objects from a convolutional neural network compared to humans. *Human Brain Mapping* n/a (2023).

62. J. Kauramäki, I. P. Jääskeläinen, M. Sams, Selective Attention Increases Both Gain and Feature Selectivity of the Human Auditory Cortex. *PLOS ONE* 2, e909 (2007).

63. S. Moldakarimov, M. Bazhenov, T. J. Sejnowski, Top-Down Inputs Enhance Orientation Selectivity in Neurons of the Primary Visual Cortex during Perceptual Learning. *PLOS Computational Biology* 10, e1003770 (2014).

64. T. Pasternak, D. Tadin, Linking Neuronal Direction Selectivity to Perceptual Decisions About Visual Motion. *Annu Rev Vis Sci* 6, 335–362 (2020).

65. J. Kubilius, M. Schrimpf, K. Kar, R. Rajalingham, H. Hong, N. J. Majaj, E. B. Issa, P. Bashivan, J. Prescott-Roy, K. Schmidt, A. Nayebi, D. Bear, D. L. K. Yamins, J. J. DiCarlo, "Brain-like object recognition with high-performing shallow recurrent ANNs" in *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2019), pp. 12805–12816.

66. O. Rose, J. Johnson, B. Wang, C. R. Ponce, Visual prototypes in the ventral stream are attuned to complexity and gaze behavior. *Nat Commun* 12, 6723 (2021).

67. C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, D. L. K. Yamins, Unsupervised neural network models of the ventral visual stream. *Proc Natl Acad Sci USA* 118, e2014196118 (2021).

68. T. Bonnen, D. L. K. Yamins, A. D. Wagner, When the ventral visual stream is not enough: A deep learning account of medial temporal lobe involvement in perception. *Neuron* 109, 2755-2766.e6 (2021).

69. P. J. Lang, M. M. Bradley, J. R. Fitzsimmons, B. N. Cuthbert, J. D. Scott, B. Moulder, V. Nangia, Emotional arousal and activation of the visual cortex: An fMRI analysis. *Psychophysiology* 35, 199–210 (1998).

70. P. Rotshtein, R. Malach, U. Hadar, M. Graif, T. Hendler, Feeling or Features: Different Sensitivity to Emotion in High-Order Visual Cortex and Amygdala. *Neuron* 32, 747–757 (2001).

71. H. T. Schupp, J. Markus, A. I. Weike, A. O. Hamm, Emotional Facilitation of Sensory Processing in the Visual Cortex. *Psychol Sci* 14, 7–13 (2003).

72. D. Sabatinelli, M. M. Bradley, J. R. Fitzsimmons, P. J. Lang, Parallel amygdala and inferotemporal activation reflect emotional intensity and fear relevance. *Neuroimage* 24, 1265–1270 (2005).

73. P. J. Lang, M. M. Bradley, Emotion and the motivational brain. *Biol Psychol* 84, 437–450 (2010).

74. L. Pessoa, Emotion and Cognition and the Amygdala: From "what is it?" to "what's to be done?" *Neuropsychologia* 48, 3416–3429 (2010).

75. V. Miskovic, A. K. Anderson, Modality general and modality specific coding of hedonic valence. *Curr Opin Behav Sci* 19, 91–97 (2018).

76. N. M. Weinberger, Specific long-term memory traces in primary auditory cortex. *Nat Rev Neurosci* 5, 279–290 (2004).

77. Z. Li, A. Yan, K. Guo, W. Li, Fear-Related Signals in the Primary Visual Cortex. *Curr Biol* 29, 4078-4083.e2 (2019).

78. N. N. Thigpen, F. Bartsch, A. Keil, The malleability of emotional perception: Short-term plasticity in retinotopic neurons accompanies the formation of perceptual biases to threat. *Journal of Experimental Psychology: General* 146, 464–471 (2017).

79. V. Miskovic, A. Keil, Acquired fears reflected in cortical sensory processing: A review of electrophysiological studies of human classical conditioning. *Psychophysiology* 49, 1230–1241 (2012).

80. S. Baek, M. Song, J. Jang, G. Kim, S.-B. Paik, Face detection in untrained deep neural networks. *Nat Commun* 12, 7328 (2021).

81. D. Burr, J. Ross, A Visual Sense of Number. *Current Biology* 18, 425–428 (2008).

82. A. Nieder, The neuronal code for number. *Nat Rev Neurosci* 17, 366–382 (2016).

83. F. Xu, E. S. Spelke, Large number discrimination in 6-month-old infants. *Cognition* 74, B1–B11 (2000).

84. F. Xu, E. S. Spelke, S. Goddard, Number sense in human infants. *Dev Sci* 8, 88–101 (2005).

85. S. Santens, C. Roggeman, W. Fias, T. Verguts, Number Processing Pathways in Human Parietal Cortex. *Cerebral Cortex* 20, 77–88 (2010).

86. M. D. Hauser, S. Carey, L. B. Hauser, Spontaneous number representation in semi-free-ranging rhesus monkeys. *Proc Biol Sci* 267, 829–833 (2000).

87. H. Sawamura, K. Shima, J. Tanji, Numerical representation for action in the parietal cortex of the monkey. *Nature* 415, 918–922 (2002).

88. M. D. Hauser, F. Tsao, P. Garcia, E. S. Spelke, Evolutionary foundations of number: spontaneous representation of numerical magnitudes by cotton–top tamarins. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 270, 1441–1446 (2003).

89. K. Nasr, P. Viswanathan, A. Nieder, Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science Advances* 5, eaav7903.

90. G. Kim, J. Jang, S. Baek, M. Song, S.-B. Paik, Visual number sense in untrained deep neural networks. *Science Advances* 7, eabd6127 (2021).

91. N. Kanwisher, J. McDermott, M. M. Chun, The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *J. Neurosci.* 17, 4302–4311 (1997).

92. K. Dobs, J. Martinez, A. J. E. Kell, N. Kanwisher, Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advances* 8, eabl8913 (2022).

93. C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization. *arXiv:1611.03530 [cs]* (2017).

94. R. VanRullen, S. J. Thorpe, The time course of visual processing: from early perception to decision-making. *J Cogn Neurosci* 13, 454–461 (2001).

95. N. Srinivasan, R. Gupta, Rapid communication: Global-local processing affects recognition of distractor emotional faces. *Q J Exp Psychol (Hove)* 64, 425–433 (2011).

96. L. Cabral, B. Stojanoski, R. Cusack, Rapid and coarse face detection: With a lack of evidence for a nasal-temporal asymmetry. *Atten Percept Psychophys* 82, 1883–1895 (2020).

97. K. Zipser, V. A. F. Lamme, P. H. Schiller, Contextual Modulation in Primary Visual Cortex. *J. Neurosci.* 16, 7376–7389 (1996).

98. S. Tschechne, H. Neumann, Hierarchical representation of shapes in visual cortex—from localized features to figural shape segregation. *Front Comput Neurosci* 8, 93 (2014).

99. R. M. Willems, M. V. Peelen, How context changes the neural basis of perception and language. *iScience* 24, 102392 (2021).

100. M. M. Bradley, P. J. Lang, Affective reactions to acoustic stimuli. *Psychophysiology* 37, 204–215 (2000).

101. E. Harmon-Jones, P. A. Gable, C. K. Peterson, The role of asymmetric frontal cortical activity in emotion-related phenomena: A review and update. *Biological Psychology* 84, 451–462 (2010).

102. P. M. Niedenthal, A. Wood, Does emotion influence visual perception? Depends on how you look at it. *Cognition and Emotion* 33, 77–84 (2019).

103. G. Li, M. G. Forero, J. S. Wentzell, I. Durmus, R. Wolf, N. C. Anthoney, M. Parker, R. Jiang, J. Hasenauer, N. J. Strausfeld, M. Heisenberg, A. Hidalgo, A Toll-receptor map underlies structural brain plasticity. *eLife* 9, e52743 (2020).

104. A. L. Tierney, C. A. Nelson, Brain Development and the Role of Experience in the Early Years. *Zero Three* 30, 9–13 (2009).

105. A. Oliva, A. Torralba, Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision* 42, 145–175 (2001).

106. C. Cortes, V. Vapnik, Support-vector networks. *Mach Learn* 20, 273–297 (1995).

107. Tiago Marques, Martin Schrimpf, James J. DiCarlo, Multi-scale hierarchical neural network models that bridge from single neurons in the primate primary visual cortex to object recognition behavior. *bioRxiv*, 2021.03.01.433495 (2021).

108. N. A. Ratan Murty, P. Bashivan, A. Abate, J. J. DiCarlo, N. Kanwisher, Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat Commun* 12, 5540 (2021).

109. C. Uran, A. Peter, A. Lazar, W. Barnes, J. Klon-Lipok, K. A. Shapcott, R. Roese, P. Fries, W. Singer, M. Vinck, Predictive coding of natural images by V1 firing rates and rhythmic synchronization. *Neuron* 110, 1240-1257.e8 (2022).

110. W. Sato, T. Kochiyama, S. Yoshikawa, E. Naito, M. Matsumura, Enhanced neural activity in response to dynamic facial expressions of emotion: an fMRI study. *Brain Res Cogn Brain Res* 20, 81–91 (2004).

111. R. M. Cichy, D. Pantazis, A. Oliva, Resolving human object recognition in space and time. *Nat Neurosci* 17, 455–462 (2014).

112. P. T. Putnam, K. M. Gothard, Multidimensional Neural Selectivity in the Primate Amygdala. *eNeuro* 6 (2019).

113. C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, J. J. DiCarlo, Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology* 10, e1003963 (2014).

114. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci Rep* 6, 27755 (2016).

115. M. Eickenberg, A. Gramfort, G. Varoquaux, B. Thirion, Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage* 152, 184–194 (2017).

116. H. Lee, E. Margalit, K. M. Jozwik, M. A. Cohen, N. Kanwisher, D. L. K. Yamins, J. J. DiCarlo, Topographic deep artificial neural networks reproduce the hallmarks of the primate inferior temporal cortex face processing network. *bioRxiv*, 2020.07.09.185116 (2020).

Supporting information

S1 Text. Supplementary information file, including supplementary figures A-N and supplementary tables A-C.

Acknowledgments

Funding: This work was supported in part by the National Institutes of the National Institutes of Health/National Institute of Mental Health grants MH112558 (MD) and MH125615 (MD, RF), the National Science Foundation grant 1908299 (RF, MD) and 2318984 (RF, MD), the University of Florida Artificial Intelligence Research Catalyst Fund (RF, MD), the University of Florida Informatics Institute Graduate Student Fellowship (PL). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. None of the authors received a salary from the funders.

Author contributions:

Conceptualization: Peng Liu, Mingzhou Ding, Ruogu Fang

Data Curation: Peng Liu, Ke Bo

Formal Analysis: Peng Liu

Funding Acquisition: Peng Liu, Mingzhou Ding, Ruogu Fang

Investigation: Peng Liu, Ke Bo, Mingzhou Ding, Ruogu Fang

Methodology: Peng Liu, Ke Bo, Mingzhou Ding, Ruogu Fang

Project Administration: Mingzhou Ding, Ruogu Fang

Resources: Mingzhou Ding, Ruogu Fang

Software: Peng Liu

Supervision: Mingzhou Ding, Ruogu Fang

Validation: Peng Liu, Ke Bo, Mingzhou Ding, Ruogu Fang

Visualization: Peng Liu, Ke Bo

Writing – Original Draft Preparation: Peng Liu, Mingzhou Ding, Ruogu Fang

Writing – Review & Editing: Peng Liu, Ke Bo, Mingzhou Ding, Ruogu Fang

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: The data and code to replicate the key results are available at https://zenodo.org/records/10720551.

Figures and Tables

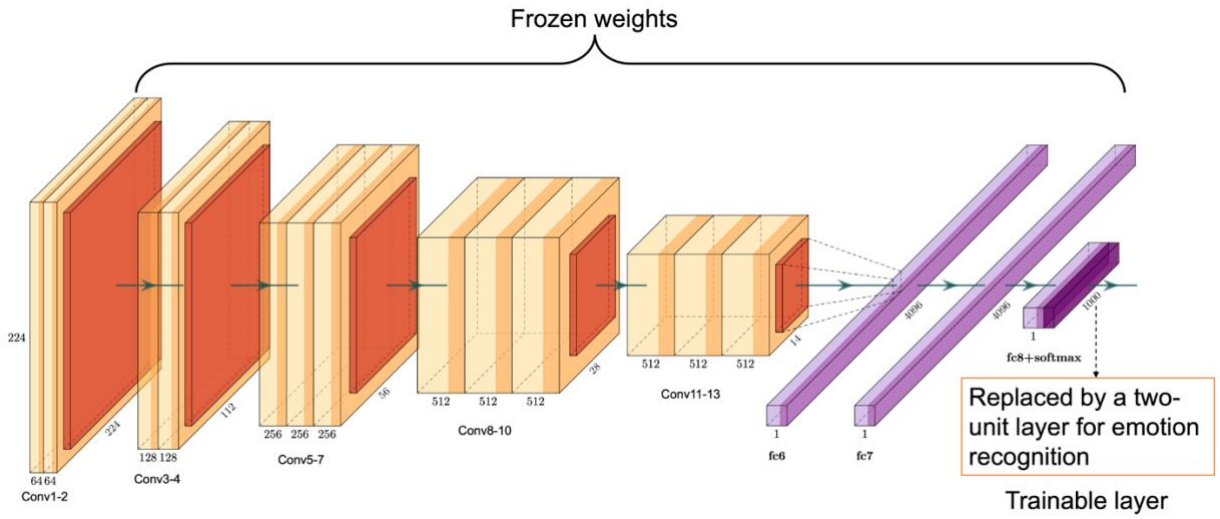


Fig 1. The architecture of the VGG-16 model. We used the VGG-16 pre-trained on ImageNet to model the visual system. VGG-16 has 13 convolutional layers and three fully connected (FC) layers. Each convolutional layer (light yellow color) is followed by a ReLU activation layer (yellow color) and a max-pooling layer (red color). Each FC layer (light purple color) is followed by a ReLU layer (purple color). The last FC layer is followed by a ReLU and a SoftMax layer (dark purple color). In the original VGG-16, the last layer was used to recognize 1000 different objects. In our model it was replaced by a two-unit layer whose connections to the preceding layer were trained to recognize different emotions: (1) pleasant vs. non-pleasant; (2) neutral vs. non-neutral; (3) unpleasant vs. non-unpleasant. Affective images in grayscale from two datasets (IAPS and NAPS) were presented to the model to define the emotion-selectivity of neurons in the convolutional layers. Lesion and attention enhancement were applied to assess these neurons' functional significance.

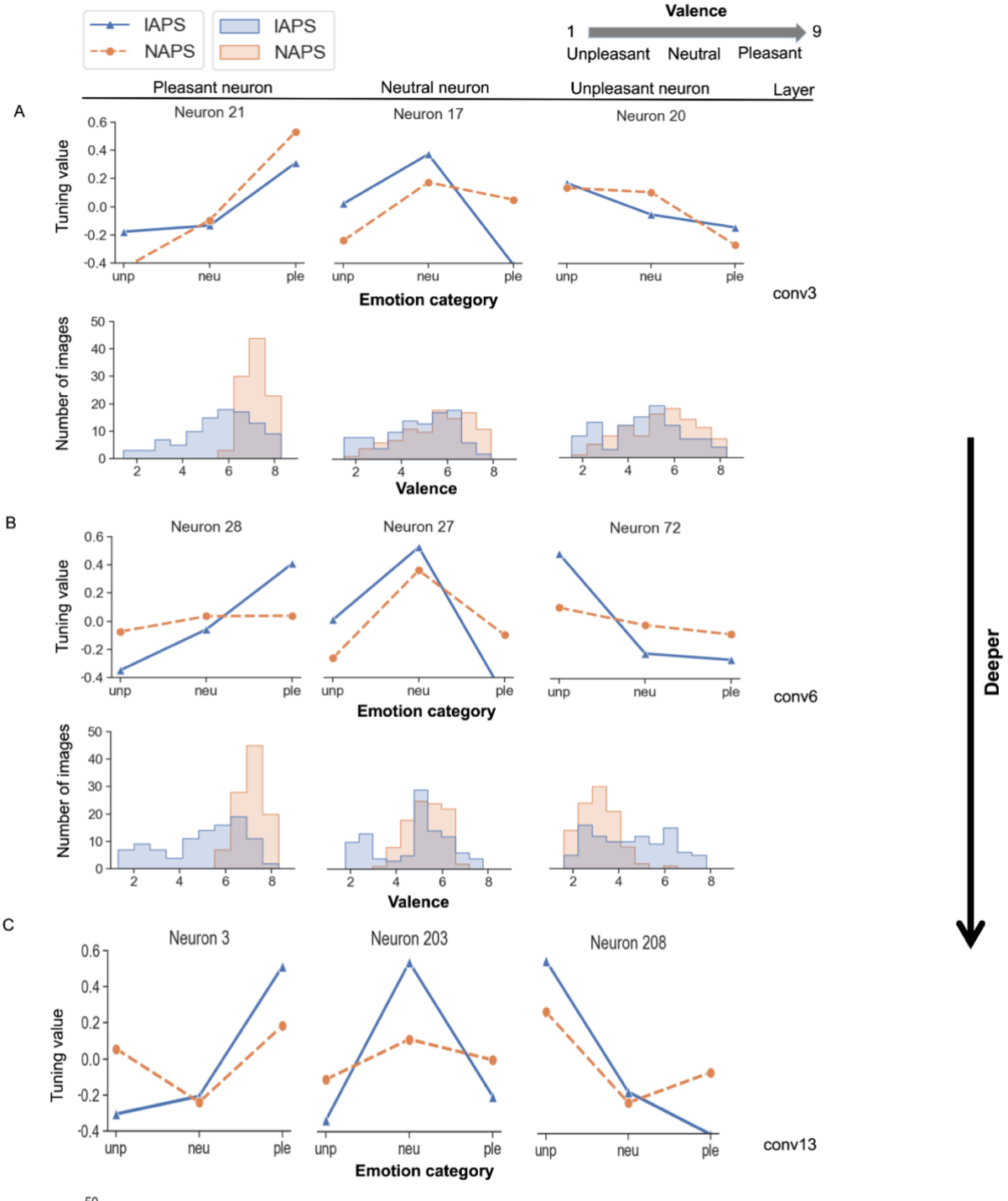


Fig 2. Tuning curves and emotion selectivity. (**A-C**) Tuning curves of example neurons from different convolutional layers (top panel) along with the valence distribution of the top 100 images that elicited the strongest responses for a given neuron.

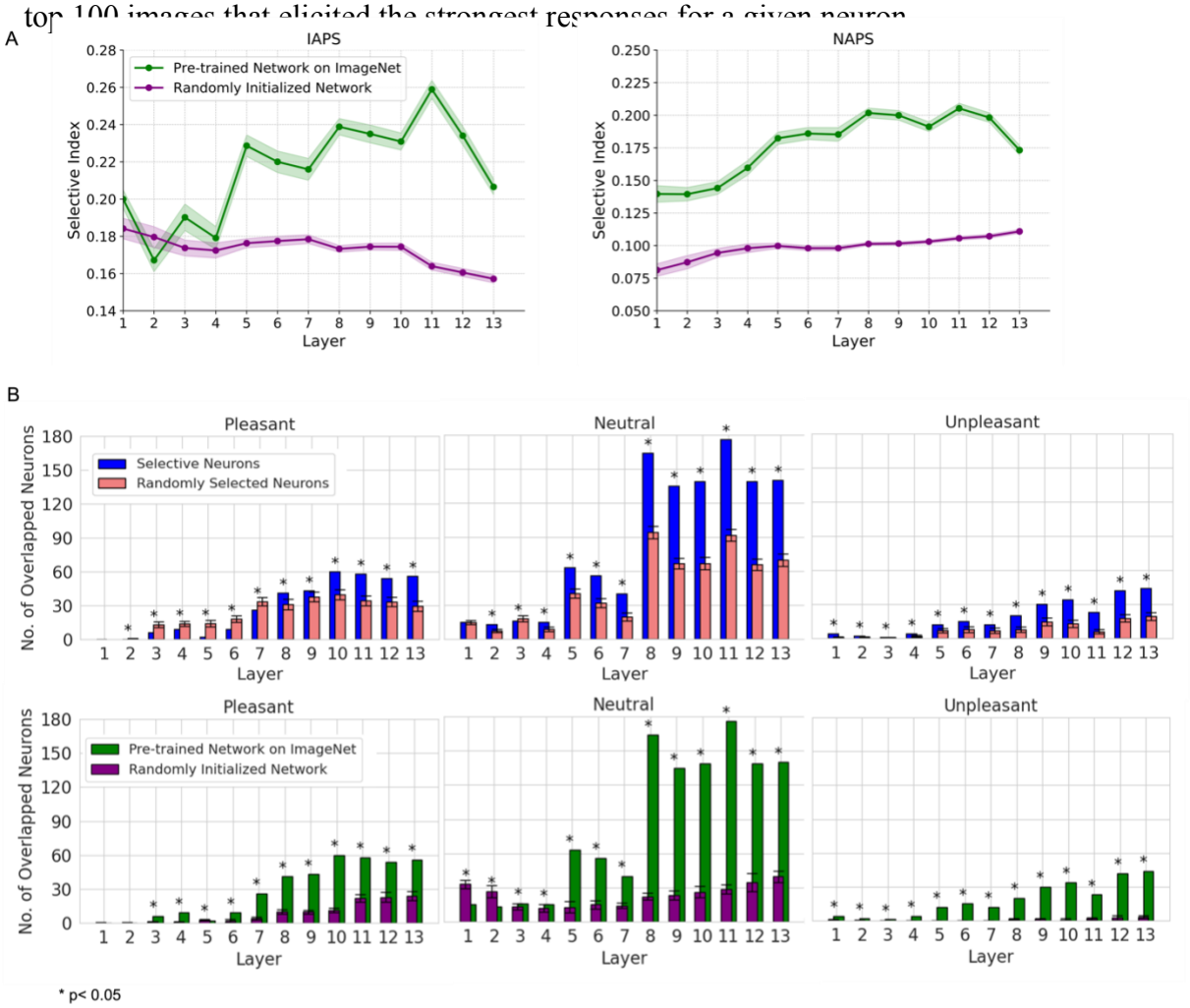


Fig 3. Emotion selectivity and its generalizability. (A) Emotion selectivity as a function of layer for IAPS and NAPS. (B-top) Number of neurons determined to be selective for a given emotion for both IAPS and NAPS datasets compared with the number of neurons in the overlap of two random sets of neurons. (B-bottom) The number of neurons determined to be selective for a given emotion for both IAPS and NAPS datasets in VGG-16 pretrained on ImageNet and with randomly initialized weights. (C) Removing successively larger percentages of neurons with small SI values and comparing the performance of attention-enhancing the remaining neurons yielded a threshold of 80% for determining emotion selectivity.

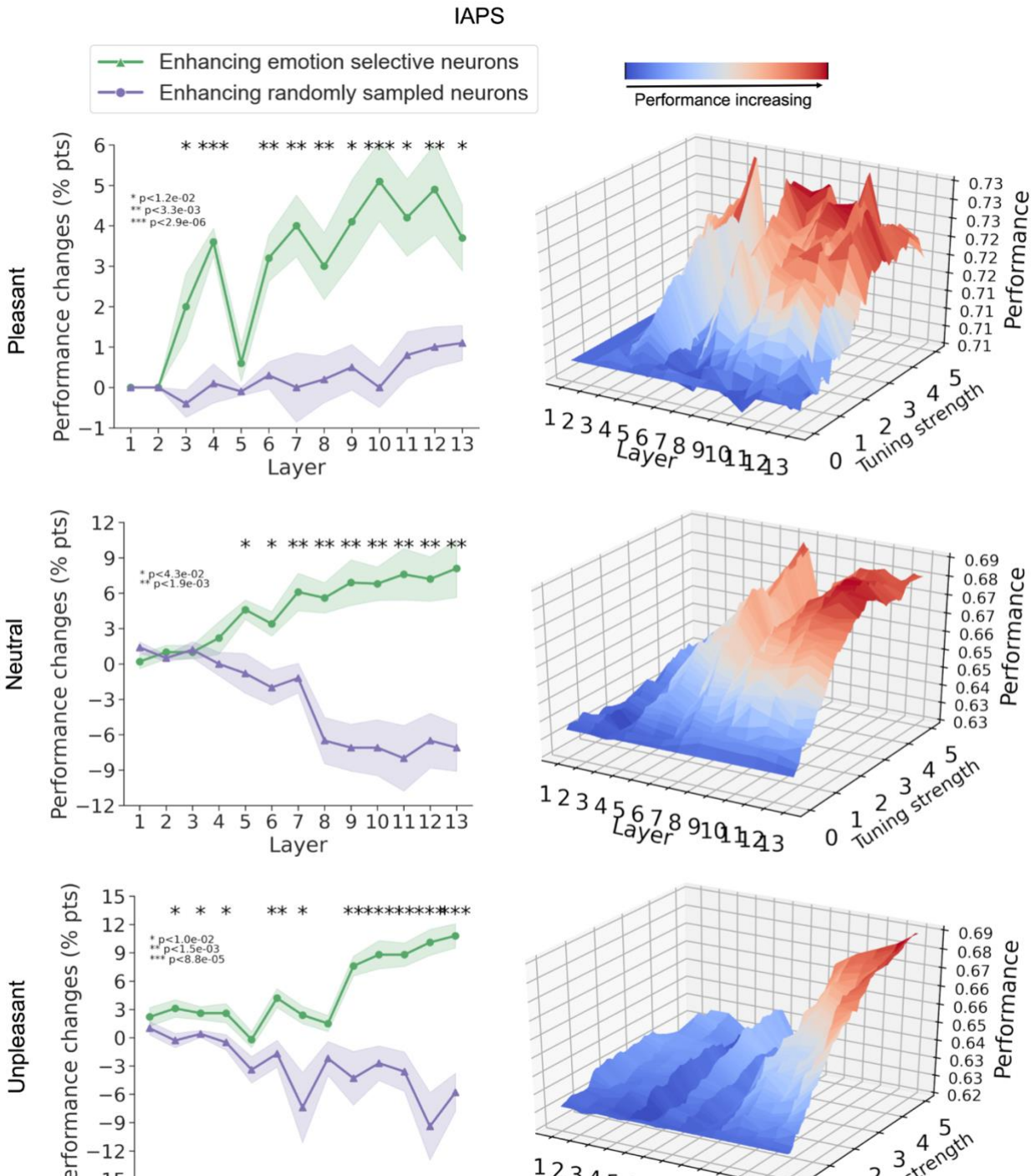


Fig 4. Effects of enhancing emotion-selective neurons and randomly selected neurons on *IAPS* dataset.

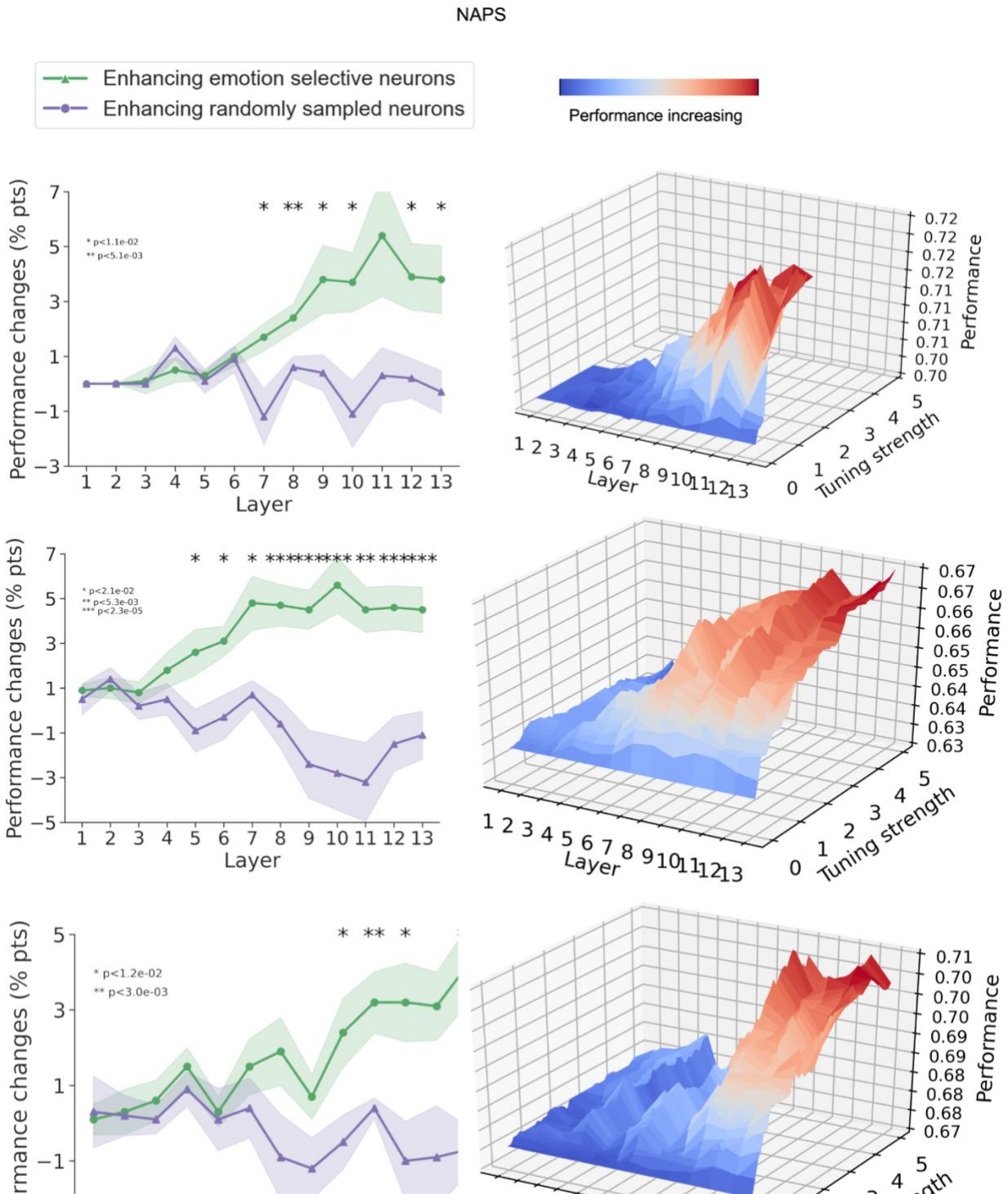


Fig 5. Effects of enhancing emotion-selective neurons and randomly selected neurons on *NAPS* dataset.

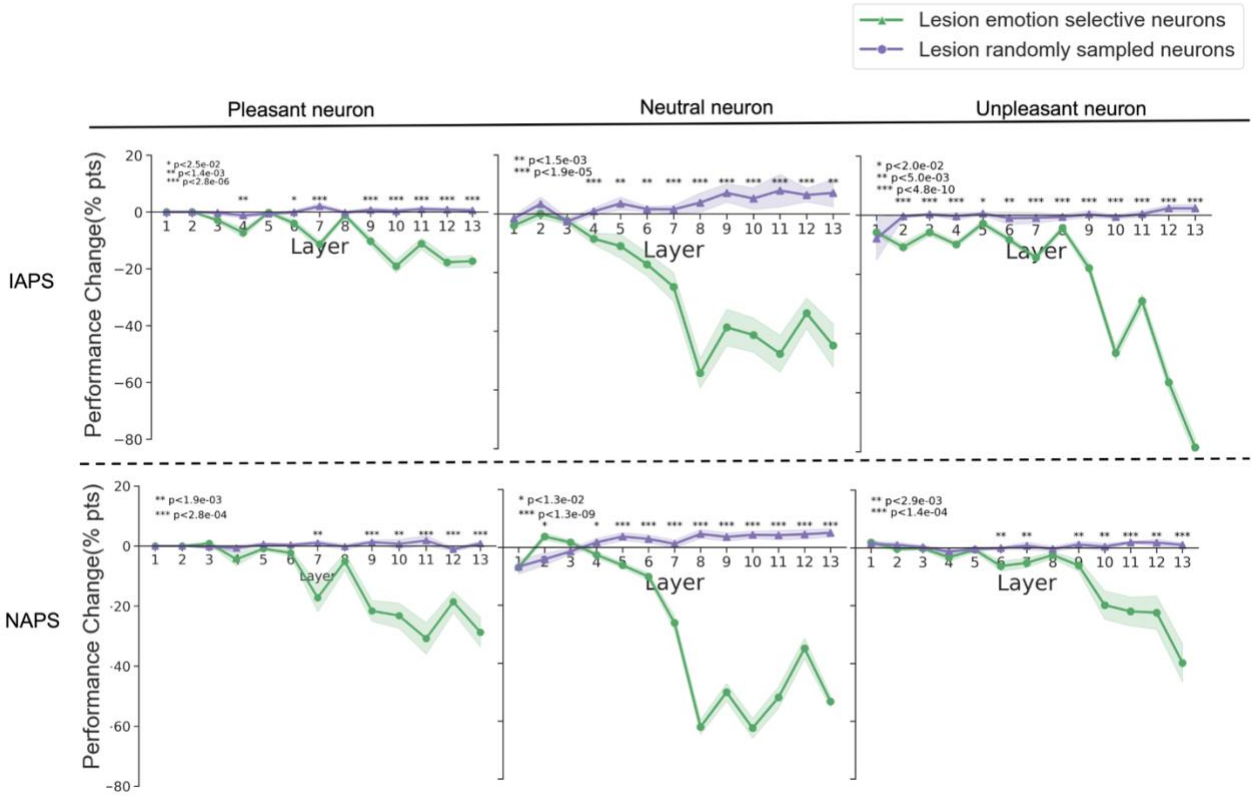


Fig 6. Lesion Analysis. Performance changes were compared between lesioning emotion-selective neurons and randomly selected neurons.

Table 1 Original and Enhanced and Lesioned Performance (F1-score) in VGG-16. The maximum performance changes for both enhancing and lesioning selective neurons across different layers are shown below.

Dataset	Emotion to Recognize	Original Performance	Enhanced Performance	Enh. Increased (%)	Lesioned Performance	Les. Decreased (%)
IAPS	Pleasant	0.70	0.73	4.29%	0.56	20%
	Neutral	0.63	0.69	9.52%	0.26	58%

	Unpleasant	0.62	0.69	11.29%	0.13	80%
NAPS	Pleasant	0.70	0.72	2.86%	0.49	31%
	Neutral	0.63	0.67	6.35%	0.25	61%
	Unpleasant	0.67	0.71	5.97%	0.41	39%

1111

Supplementary Materials

Emergence of Emotion Selectivity in Deep Neural Networks Trained to Recognize Visual Objects

Peng Liu[1,2], Ke Bo[2], Mingzhou Ding[1]*, Ruogu Fang[1,3]*

[1]J. Crayton Pruitt Family Department of Biomedical Engineering, Herbert Wertheim College of Engineering, University of Florida, Gainesville, FL, USA

[2]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA
[3]Center for Cognitive Aging and Memory, McKnight Brain Institute, University of Florida, Gainesville, FL, USA

***Corresponding author**: mding@bme.ufl.edu; ruogu.fang@bme.ufl.edu

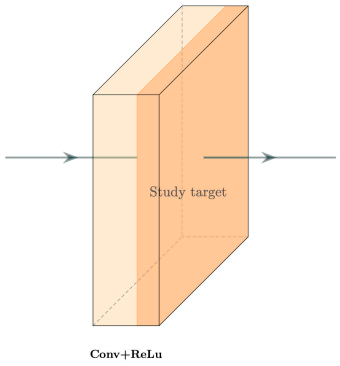
Supporting information Text

Eight topics related to the study reported in the main manuscript are addressed in this Supplementary Materials.

Topic 1. Additional details of model developments, image datasets and methods

The structure of artificial neurons, the number of images in each dataset, and valence distribution in each dataset are shown in Fig A. The information flow of applying convolution and ReLU function on an input, the details of how to enhance and lesion artificial neurons, and the datasets and networks used in the study are illustrated in Fig B.

A



B

Dataset	Pleasant	Neutral	Unpleasant	Total
IAPS	296	390	341	1,027
NAPS	352	477	281	1,110

C

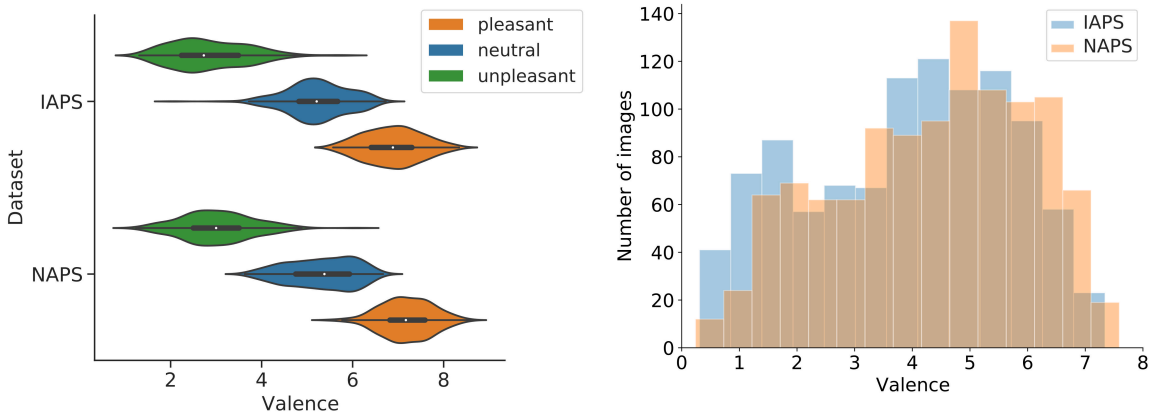
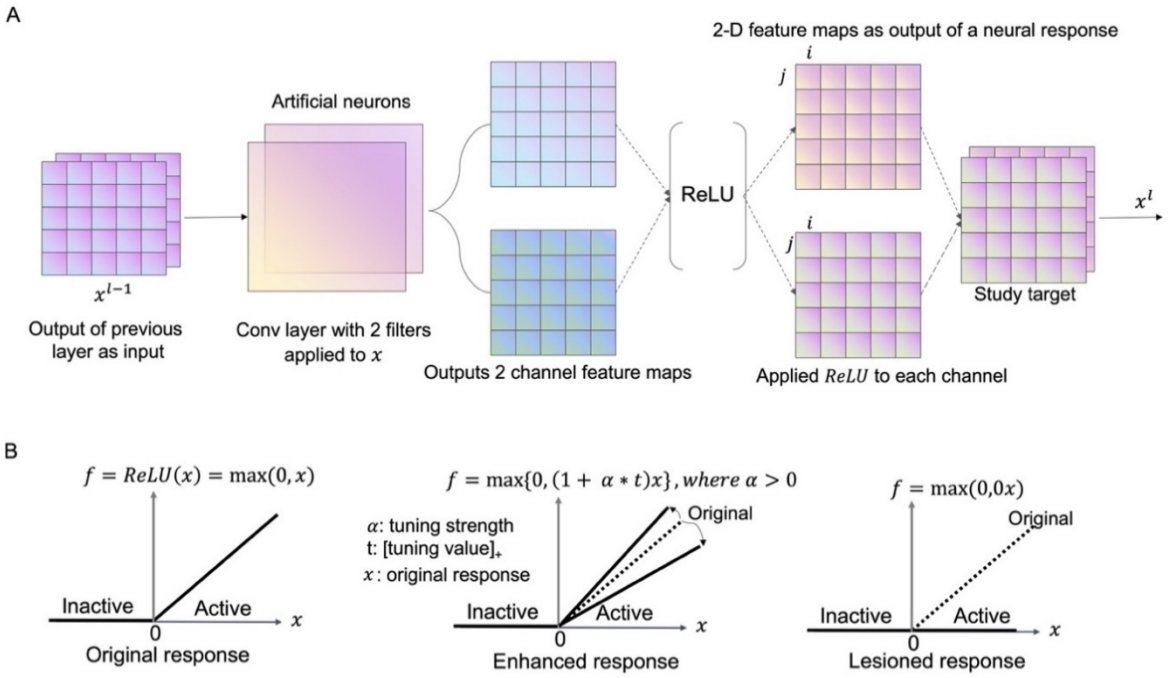


Fig A: Model development details and image datasets. (**A**) Each convolutional layer is followed by one *ReLU* layer, the output of which reflects the responses of the artificial neurons the convolutional layer. Thus, in this study, the output of the *ReLU* layer is our study target for understanding the activity of the artificial neurons. (**B**) It shows the number of images of each emotion category in the two datasets used in this study. Two datasets were treated equally for defining emotion-selective neurons and related lesion and attention manipulations. (**C**) It shows how the divided categorial images match the valence score originally rated by human subjects in the two datasets. The C (left) shows the valence score distribution and the boundary score between the pleasant and neutral category: 4.3 ± 0.5 and between the neutral and unpleasant category: 6.0 ± 0.5. The C (right) shows the number of images per valence score across two datasets. Basically, this figure illustrates the details of the model development and the affective image datasets.



C

The datasets and models used in study

#emotion categories	Dataset	The usage	Balanced #image per category(e.g., pleasant versus non-pleasant)	Associated model
Binary (2 categories) in each dataset	IAPS-pleasant	Tuning study, random sampling study, and lesion study	296	IAPS-pl
	IAPS-neutral		390	IAPS-ne
	IAPS-unpleasant		341	IAPS-un
	NAPS-pleasant		352	NAPS-pl
	NAPS-neutral		477	NAPS-ne
	NAPS-unpleasant		281	NAPS-un
-	IAPS	defining selectivity and overlapped neurons for each emotion category	See S. Fig.1B	No need model training
	NAPS			

Fig B: Further methodological details. (**A**) It shows that the information flow of applying convolution and $ReLU$ operation to the input x in layer l, which represents all the feature maps from the previous layer $l-1$. In this illustration, x is composed of two channels of feature maps. Two filters (referred to as artificial neurons) in the following convolutional layer are applied to x separately. Each filter convolution results in a channel of a new feature map, which then is passed through an activation function, $ReLU = \max(0, x)$. $ReLU$ function indicates which features after the convolutional operation is activated or inactivated. The active units are valued with positive numbers, and the inactive units are valued with zeros by the $ReLU$. Fundamentally, this figure indicates how an artificial neuron responds to a stimulus and how the response activations are calculated in a CNN. (**B**) The bold black line represents the $ReLU$ response (activation) values, and dash lines in the middle and right sub-figures represent the normal activation value of the $ReLU$. Three cases of activation behavior were investigated empirically. The normal excitation (left) is applying the original $ReLU$; the attention enhanced excitation (middle) is applying a positive weight α to the activation value x; the inhabited activation through lesion is setting the activation values to be zeros instead, which performs like a lesion study. (**C**) It summarizes the datasets and the models used in the study. The number of images in the binary model is balanced. The non-* category images were randomly selected from another two categories.

Topic 2. Additional analysis of the selective index

We examined the distribution and correlation of the selective index (SI) of selective neurons defined on dataset IAPS and NAPS, separately, in each layer and the number of selective neurons across layers by emotion category in IAPS and NAPS. The purpose is to answer the following questions: how many selective neurons there are in the network, how they depend on layers, and how the strength of selectivity depends on layers. The overall selective index is around 0.2 for IAPS and 0.15 for NAPS shown in Fig C. The correlation between IAPS-defined emotion SI and NAPS-defined SI was computed and the result was shown in Fig D. The left scatter plot, where neurons from all layers are combined, indicates a positive correlation (Pearson coefficient of 0.30) between IAPS- and NAPS-defined SI. The right plot shows that the correlation between the two SI indices increases as we move deeper into the network. This analysis further supports our claim that emotion selectivity is generalizable across the two datasets.

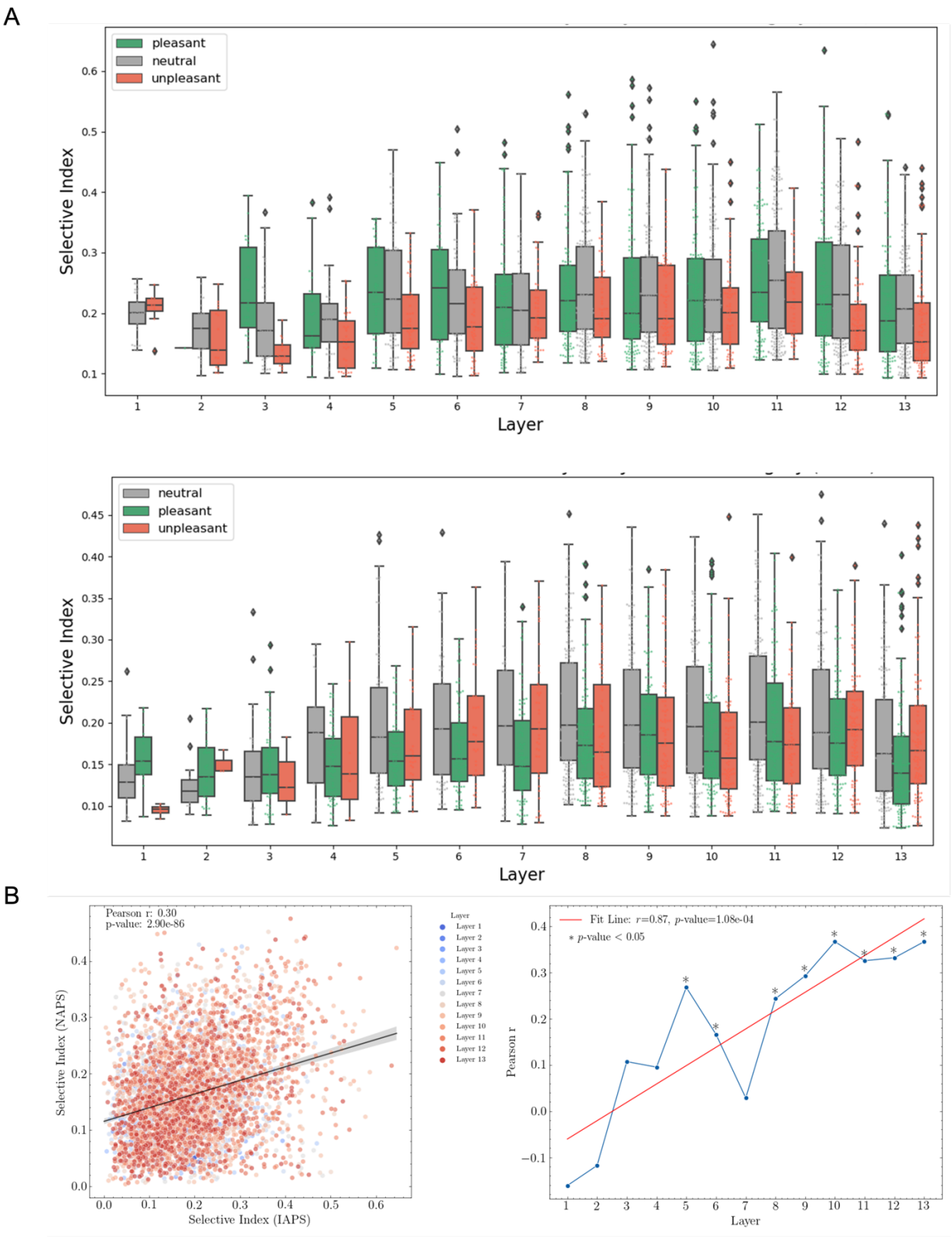


Fig C: Additional analysis of selective Index. (A) Distribution of selective indices across layers by emotion category in dataset IAPS (Top) and NAPS (Bottom). **(B)** The correlation between IAPS-defined SI and NAPS-defined SI. (Left) Neurons from all layers are combined. (Right) The layer-wise correlation was plotted.

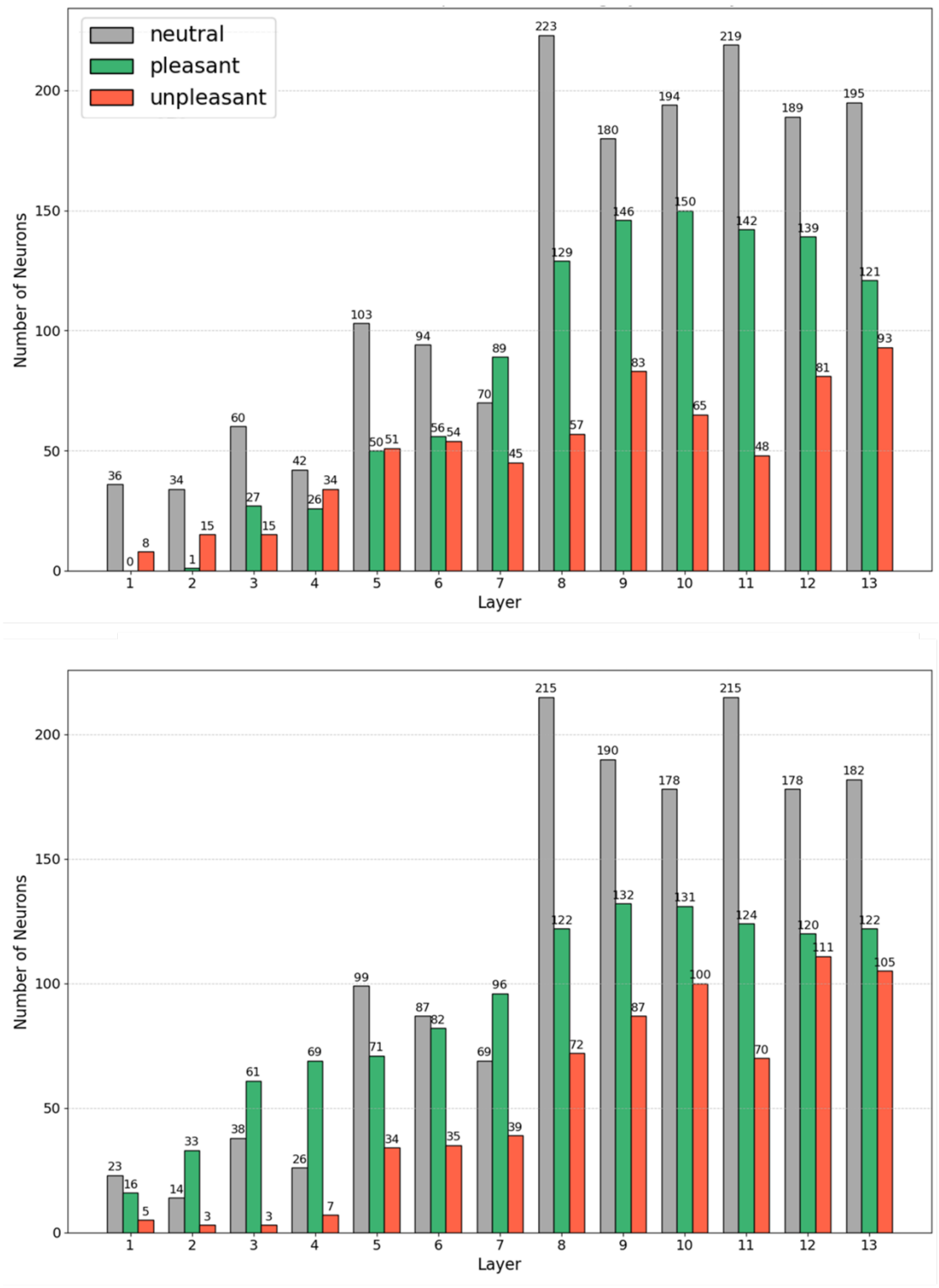


Fig D: Number of selective neurons across layers by emotion category in dataset IAPS (Top) and NAPS (Bottom).

Topic 3. Generalizability of emotion selectivity

We examined the functional generalization of selective neurons defined on IAPS and NAPS, separately, in Fig E. The purpose is to further verify whether the emotion selectivity defined on one dataset can be functionally generalized to another dataset. The result is consistent with other results obtained by enchaining on selective neurons with their selective index defined on the same dataset (either IAPS or NPAS). It further supports our claim that emotion selectivity shares a functional property between the two datasets.

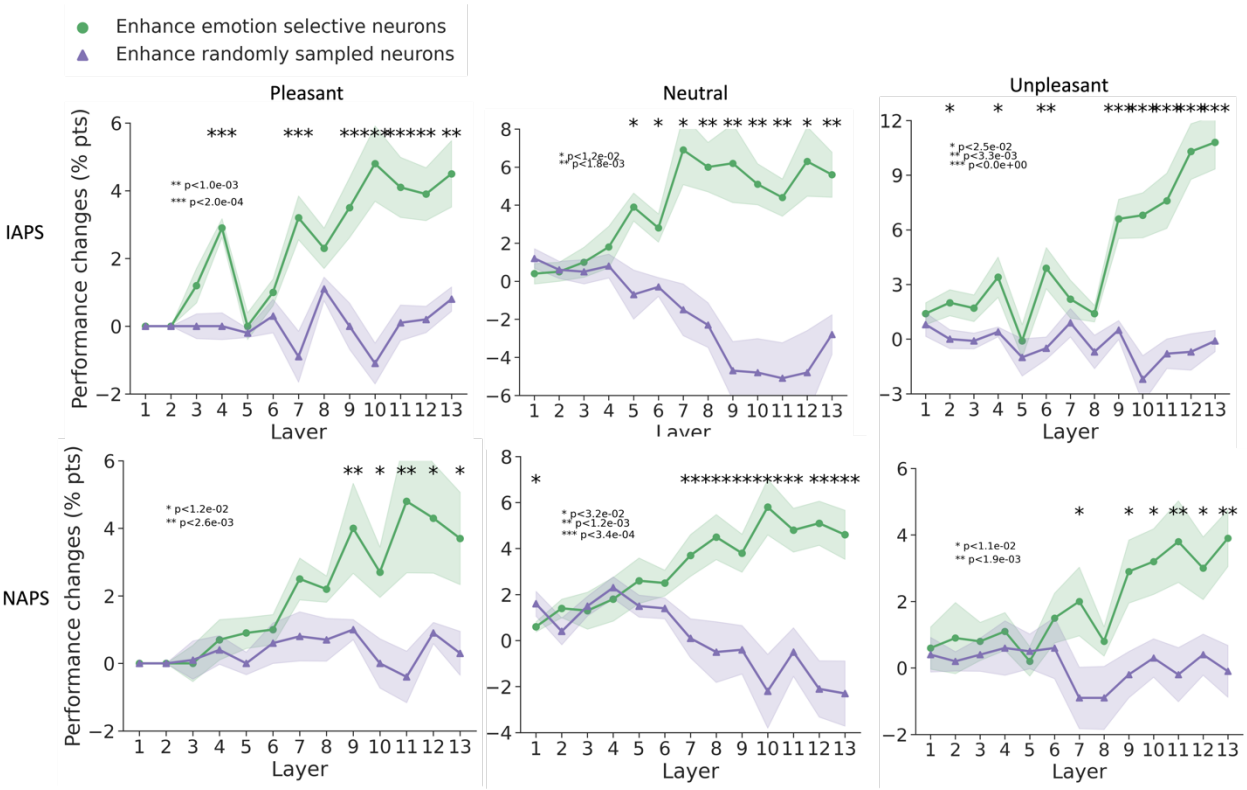


Fig E: Functional generalization analysis. (**Top**) We analyzed the enhancement of emotion-selective neurons (defined post-threshold) versus random neurons in each VGG-16 layer trained on IAPS. The selective index was derived from NAPS and tested on IAPS. (**Bottom**) We analyzed the enhancement of emotion-selective neurons (defined post-threshold) versus random neurons in each VGG-16 layer trained on NAPS. The selective index was derived from IAPS and tested on NAPS.

Topic 4. Result replication in AlexNet

We replicated the results in another network, AlexNet, shown in Fig F, Fig G, and Fig H, and Fig I. The purpose is to demonstrate that the emergence of emotion selectivity is not an idiosyncratic property of a specific deep neural network. We summarized the parallel results produced with VGG-16 and AlexNet in Table A and Table B.

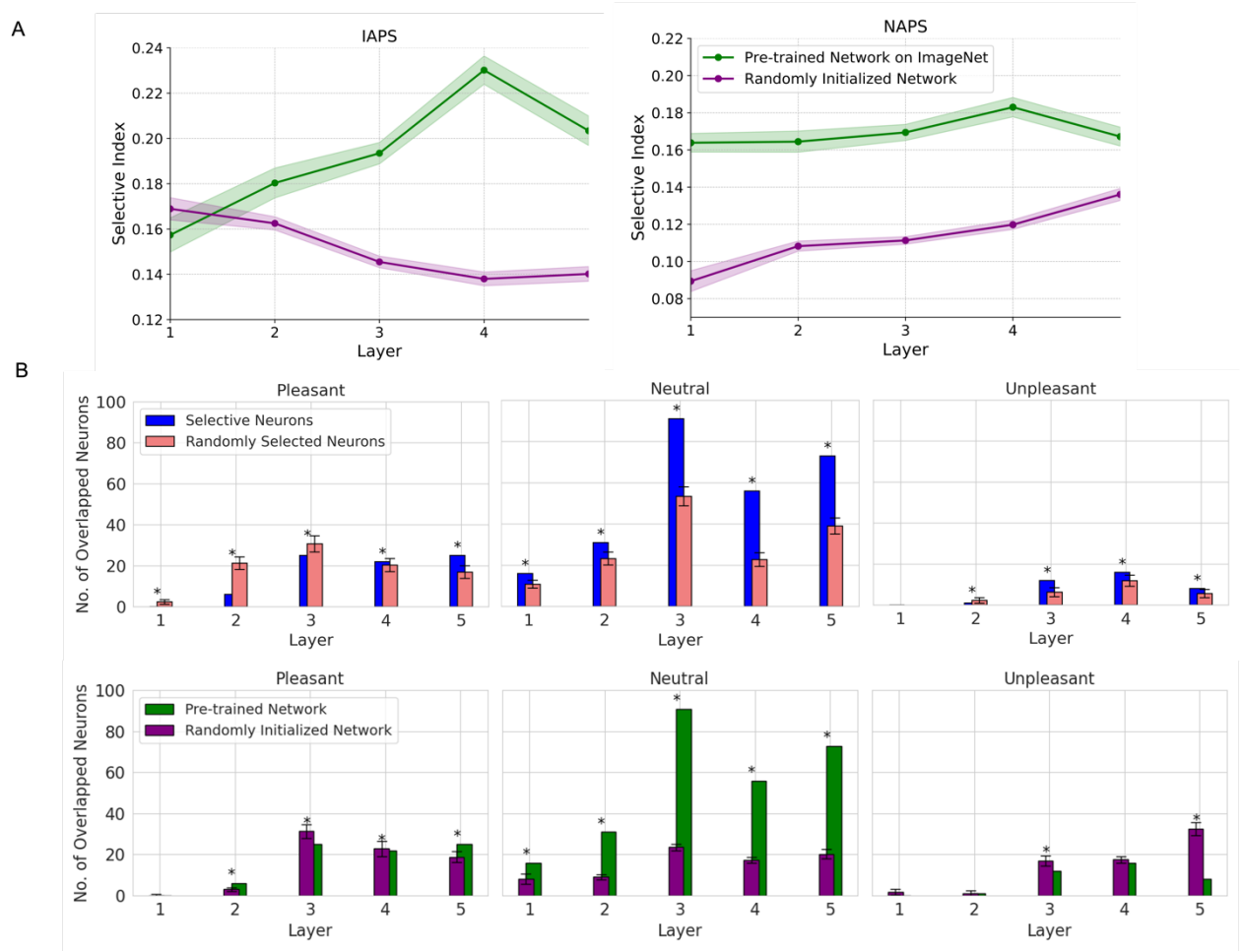


Fig F: Selective Index Quality (A) and Generalizability in AlexNet (B) of emotion selectivity across two datasets**.** The comparison of number of overlapped neurons derived from selective neurons and randomly selected neurons is plotted (B-top). The one of number of overlapped neurons derived from pre-trained AlexNet on ImageNet and initialized AlexNet network with random weights (B-bottom). The goal of this comparison is to demonstrate the significance of learned features from ImageNet in developing neuron selectivity. However, merely counting the overlapping neurons might not be adequate; we should also take into account the selectivity index quality. This is particularly important when the total number of neurons in a layer is small.

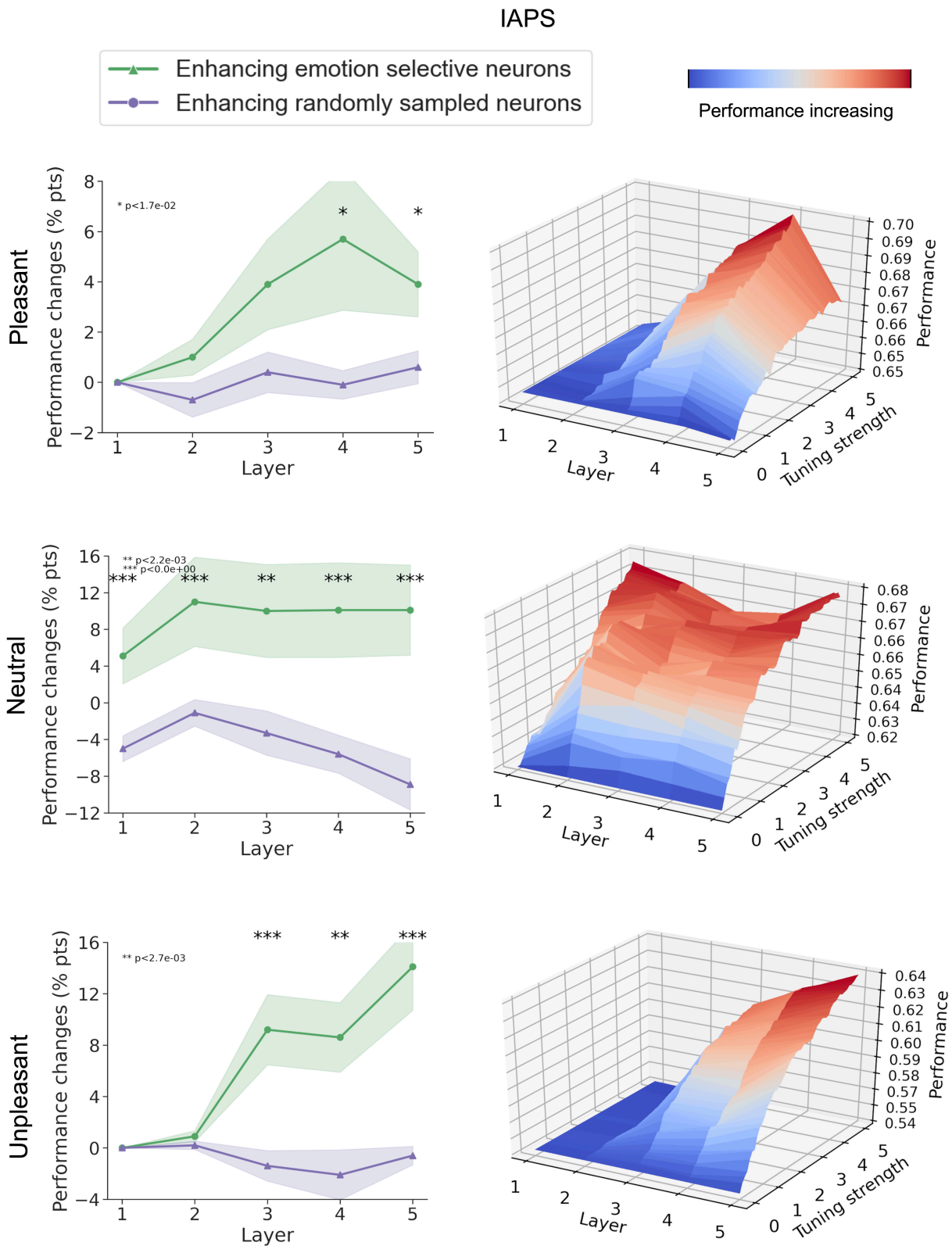


Fig G: Effects of attention-enhancing emotion-selective neurons and randomly selected neurons in AlexNet on IAPS dataset.

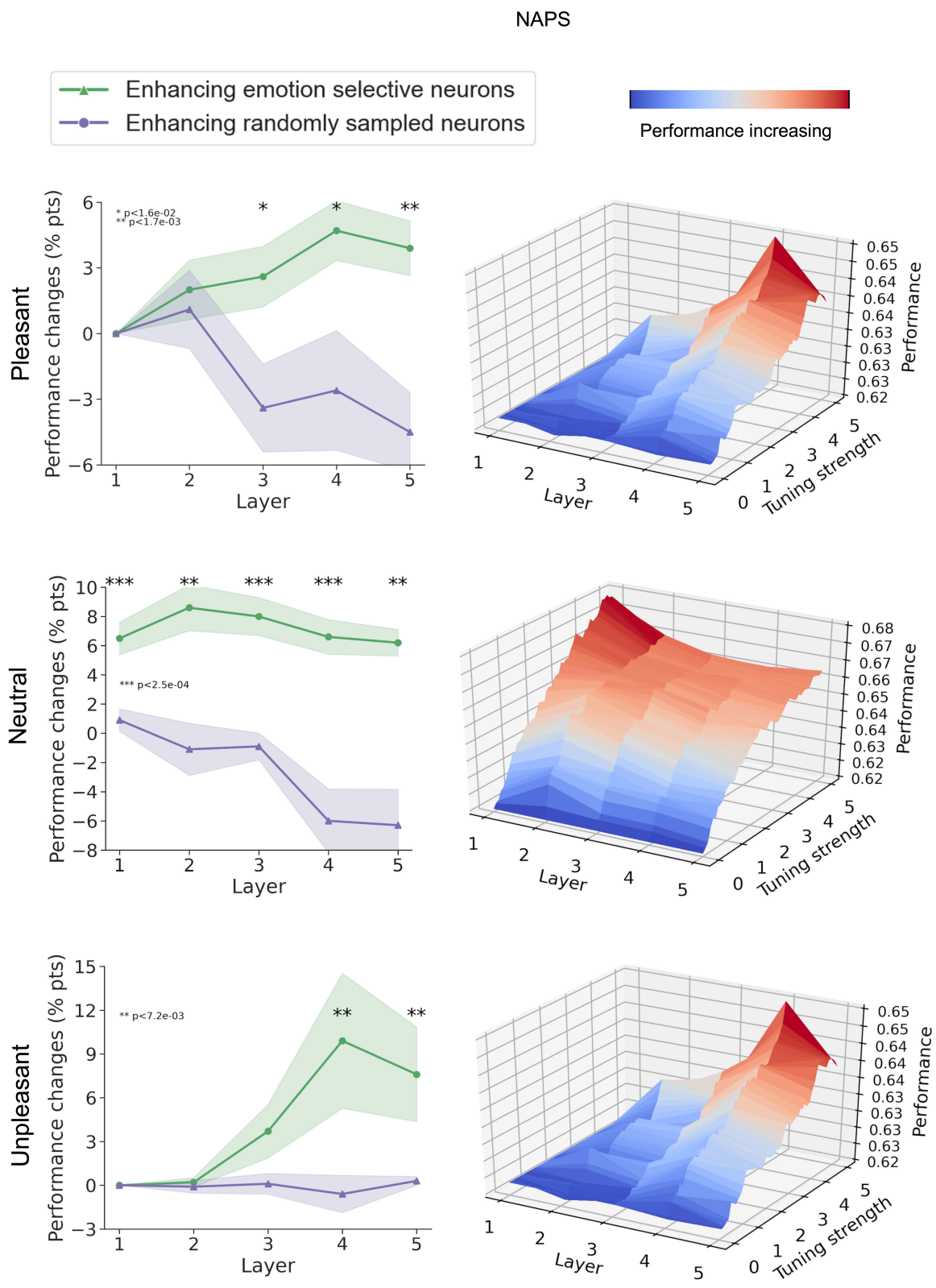


Fig H: Effects of attention-enhancing emotion-selective neurons and randomly selected neurons in AlexNet on NAPS dataset.

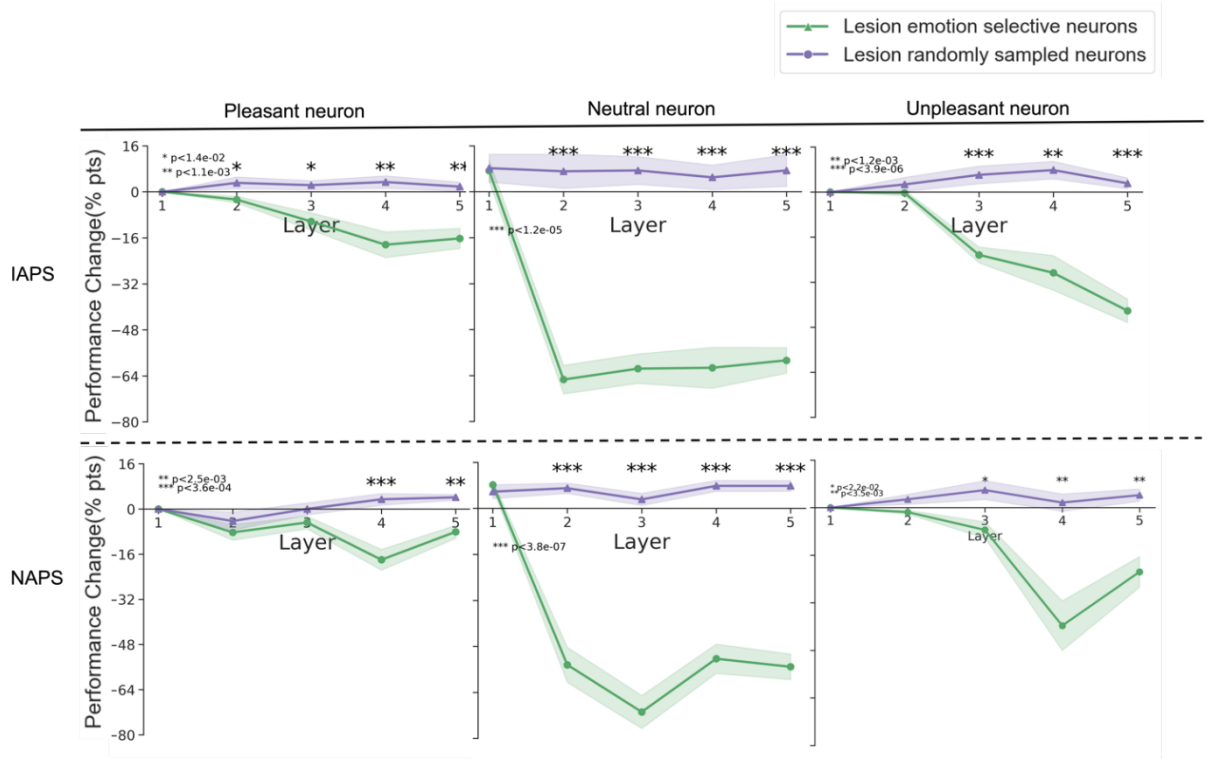


Figure I. Effects of lesioning emotion-selective neurons and randomly selected neurons in AlexNet. (**A**) IAPS dataset. (**B**) NAPS dataset**.**

Table A. Results comparison between VGG-16 and AlexNet.

Description	**VGG-16**	**AlexNet**
Selective Index Quality	Fig 3A	Fig C
Number of overlapped neurons	Fig 3B	Fig D
Enhance emotion- selective neurons	Fig 4	Fig G and Fig H
Lesion emotion-selective neurons	Fig 5	Fig I

Table B Original and Enhanced Performance (F1-score) in VGG-16 and AlexNet.

Network	Dataset	Emotion to Recognize	Original Performance	Enhanced Performance	Enh. Improvement (%)	Lesioned Performance	Les. Decreased (%)
VGG-16	IAPS	Pleasant	0.70	0.73	4.29%	0.56	20%
		Neutral	0.63	0.69	9.52%	0.26	58%
		Unpleasant	0.62	0.69	11.29%	0.13	80%
	NAPS	Pleasant	0.70	0.72	2.86%	0.49	31%
		Neutral	0.63	0.67	6.35%	0.25	61%
		Unpleasant	0.67	0.71	5.97%	0.41	39%
AlexNet	IAPS	Pleasant	0.65	0.70	7.69%	0.55	16%
		Neutral	0.62	0.68	9.68%	0.22	64%
		Unpleasant	0.54	0.64	18.52%	0.37	32%
	NAPS	Pleasant	0.62	0.65	4.84%	0.52	16%
		Neutral	0.62	0.68	9.68%	0.22	65%
		Unpleasant	0.62	0.65	4.84%	0.40	35%

Topic 5. Low-level features as possible confounding factors

low-level features were extracted from the images by using GIST algorithm. Pairwise emotion decoding was performed using (see Fig J) using SVM. The mean accuracy for both IAPS and NAPS datasets approximates the chance level, suggesting that low-level GIST features are insufficient for decoding emotion categories from images.

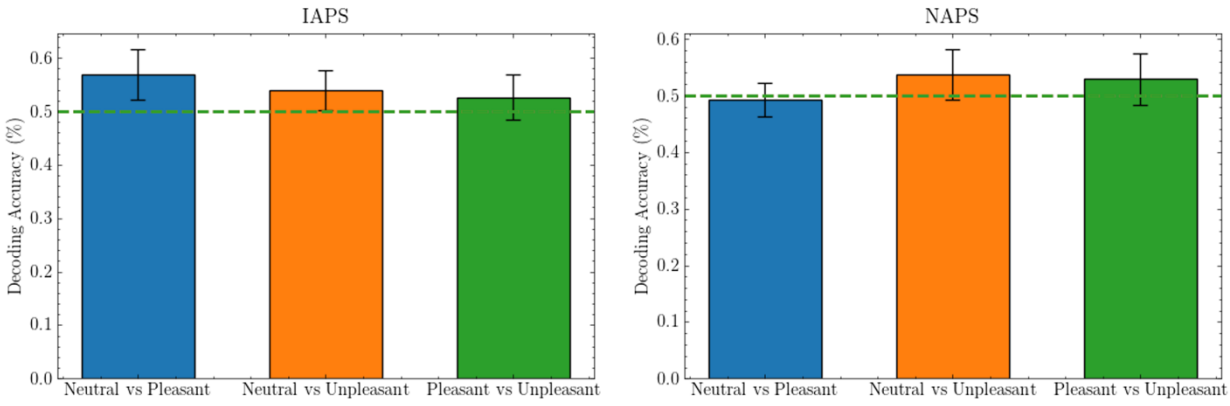


Fig J: Pairwise decoding results using low-level features (GIST). Dash line indicates the chance level performance (50%). The dashed lines indicate chance-level performance and bars representing the average accuracy across 10 iterations of 5-fold cross-validation.

Topic 6. Faces as possible confounding factors

The percentages of images involving faces in top 100 images (ranking based on neurons' activation to each image) that evoked the strongest response of selective neurons for each emotion category

(see Fig K) are: 16%, 62%, and 30% for pleasant, neutral, and unpleasant, separately. The analysis demonstrates that development of emotion selectivity in these neurons is unlikely affected by potential facial encoding that might arise during the training of the network on ImageNet.

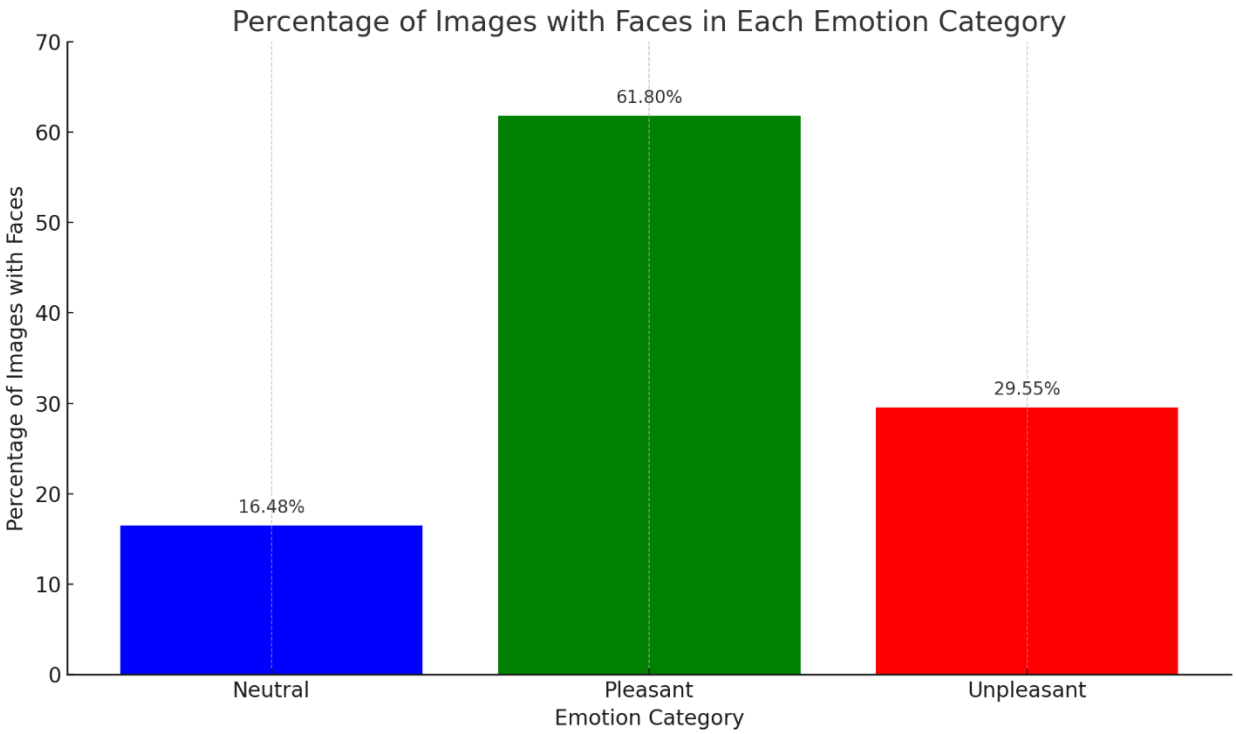


Fig K: Number of images involving faces in top 100 images that evoked the strongest response of emotion-selective neurons.

Topic 7. Animacy as possible confounding factors

Table C shows the mean valence and arousal of the top 100 images across selective neuron categories. The purpose of this analysis is to estimate how much valence and arousal relevant to the images evoked by the selective neurons are captured. The result shows the mean valence: 6.770, 5.180, and 2.898 and mean arousal: 5.055, 3.970, and 5.816 for top images that evoked the strongest responses in neurons selective for pleasant, neutral, and unpleasant emotion. More importantly, these images appear to contain both animate and inanimate content, suggesting that animacy might not be a confounding factor.

Table C Valence and Arousal of Top 100 Images Across Selective Neuron Categories. Note: Image ranking is based on neurons' activation to each image.

Neuron Selectivity	Mean Valence	Interpretation	Mean Arousal	Interpretation
Pleasant	6.770	Images evoke relatively positive or pleasant emotions.	5.055	Emotions of moderate intensity.
Neutral	5.180	Emotions neither particularly positive nor negative.	3.970	More subdued or calm emotions.
Unpleasant	2.898	Images evoke negative or unpleasant emotions.	5.816	Intense negative emotions (e.g., fear, distress).

Topic 8. Effects of emotion, object category, and their interaction on neuronal responses

As illustrated in Fig L-M, we examined, using a Two-Way ANOVA analyses, the impact of image emotion, image category, and their interaction on neuronal response. Object categories were identified based on the descriptions in the original datasets (refer to Fig LA and MA). Our findings reveal that the emotion category markedly affects neuronal activity in layers subsequent to the fifth (refer to Fig LB-top and MB-top), and the influence of the object category is increasing with layer depth but not significant. The interaction is significant in some deeper layers (refer to Fig LB-bottom and MB-bottom). It should be noted that this analysis should be viewed as preliminary, because the number of images in each object category is rather small, which may impact the analysis adversely. In addition, the selection of the images is also dataset-specific. For example, in the IAPS dataset, 15 images of dogs predominantly express negative emotions, whereas in the NAPS dataset, 35 images of dogs represent a mix of negative and positive emotions (see Fig N). This variance indicates the necessity for additional studies to comprehensively understand the interaction between image emotion and category.

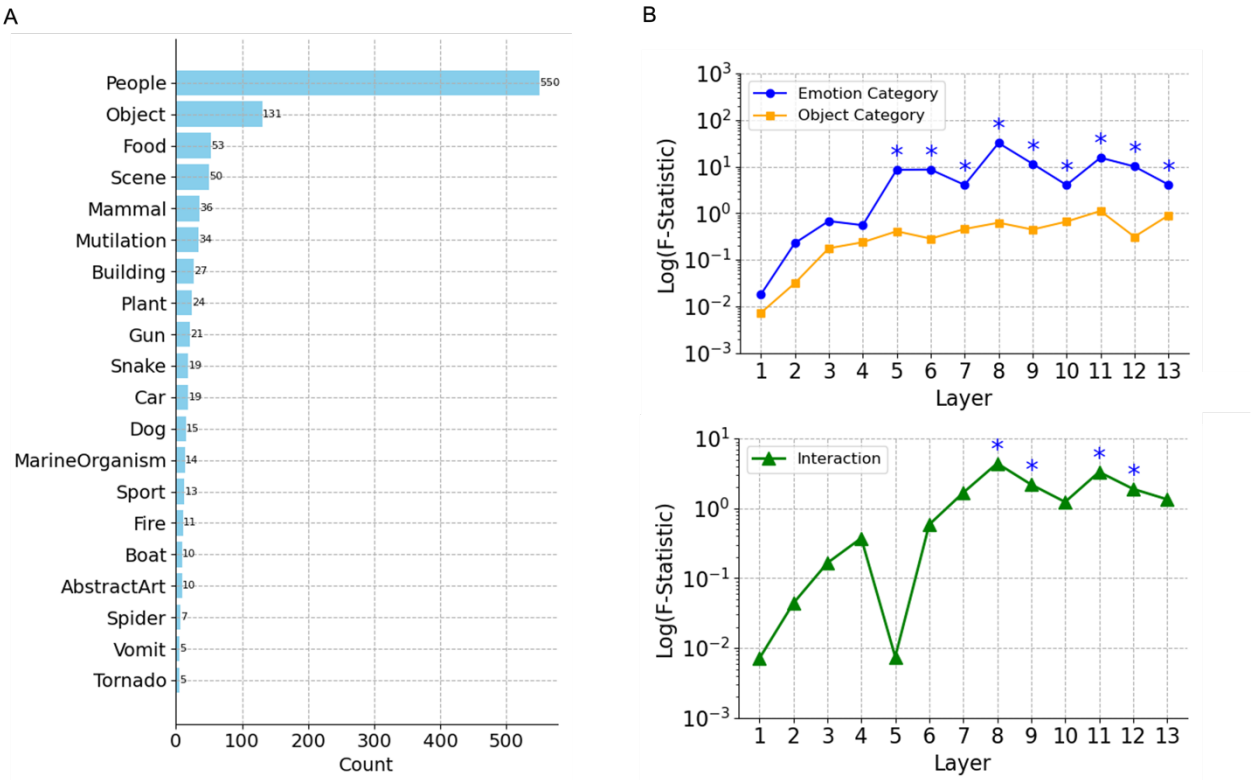


Fig L: Effects of emotion and object category on filter activity using *IAPS* images. **A.** The number of images in each of the top 20 object categories. **B. (top)** The F-statistic (log scale) of the effect of emotion and object category on filter activations across layers of the VGG-16 neural network. The statistics are obtained from a Two-Way ANOVA test, where the dependent variable is the filter activity in response to images. The plot reveals how each factor impacts the filter responses and how this influence changes from the input to deeper layers of the network; (**bottom**) The F-statistic (log scale) of the interaction between emotion dependent filter activation and object category dependent filter activation. The statistics are obtained from a Two-Way ANOVA test. * *indicates the influence is statistically significant.*

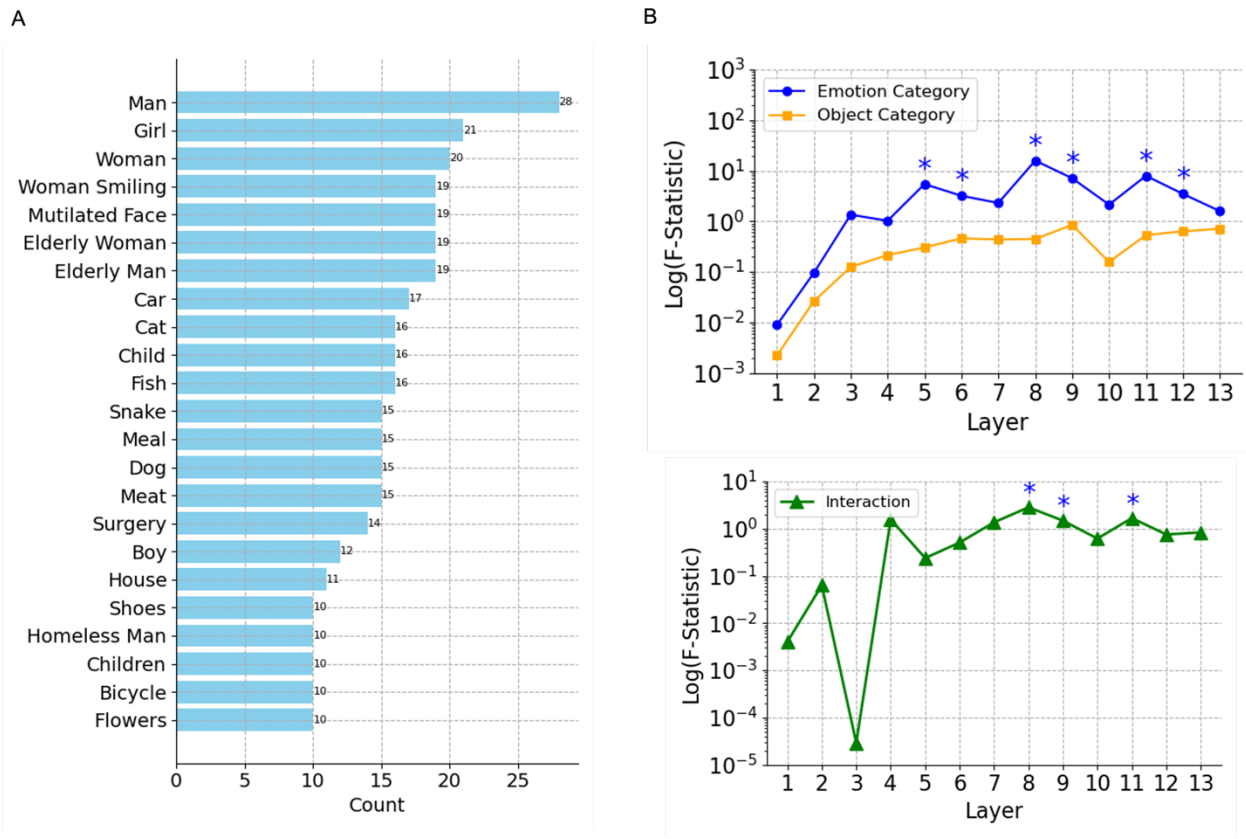


Fig M: **Effects of emotion and object category on filter activity using *NAPS* images**. **A.** The number of images in each of the top 23 object categories. **B. (top)** The F-statistic (log scale) of the effect of emotion and object category in filter activations across layers of the VGG-16 neural network. The statistics are obtained from a Two-Way ANOVA test, where the dependent variable is the filter activity in response to images. The plot reveals how each factor impacts the filter responses and how this influence changes from the input to deeper layers of the network; (**bottom**) The F-statistic (log scale) of interaction between emotion and object category. The statistics are obtained from a Two-Way ANOVA test. * *indicates the influence is statistically significant.*

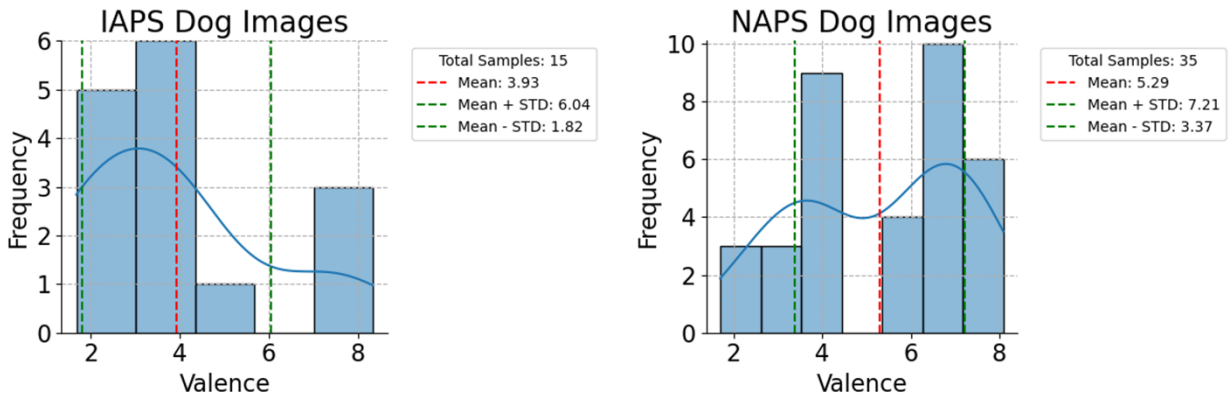


Fig N: Valence distributions of dog images in dataset IAPS (left) and NAPS (right).