Published in final edited form as:

Nat Hum Behav. 2021 June ; 5(6): 743–755. doi:10.1038/s41562-021-01124-6.

Aesthetic preference for art can be predicted from a mixture of low- and high-level visual features

Kiyohito Iigaya[*,1], **Sanghyun Yi**[1], **Iman A. Wahle**[1], **Koranis Tanwisuth**[†,1], **John P. O'Doherty**[1]

[1]Division of Humanities and Social Sciences, California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125

Abstract

It is an open question whether preferences for visual art can be lawfully predicted from the basic constituent elements of a visual image. Here, we developed and tested a computational framework to investigate how aesthetic values are formed. We show that it is possible to explain human preferences for a visual art piece based on a mixture of low- and high-level features of the image. Subjective value ratings could be predicted not only within but also across individuals, with a regression model with the same set of interpretable features. We also show that the features which predict aesthetic preference can emerge hierarchically within a deep convolutional neural network trained only for object recognition. Our findings suggest that human preferences for art can be explained at least in part as a systematic integration over the underlying visual features of an image.

Introduction

From ancient cave paintings to digital pictures posted on Instagram, the expression and appreciation of visual art is at the core of human experience. As Kant famously pointed out, art is both subjective and universal.[1] Each individual person may have his/her own taste, but a given piece of art can also appeal to a large number of people across cultures and history. This subjective universality raises a fundamental question: should artistic tastes be likened to the inscrutable, idiosyncratic, and irreducible, or is it possible to deduce lawful and generalizable principles by which humans form aesthetic opinions?

The nature of aesthetic judgment has long been subject to empirical investigation.[2–10] Some studies have focused on the visual and psychological aspects of art might influence aesthetics (e.g., see[3,5,9,11–13]), while other work has highlighted the brain regions whose activity level correlates with aesthetic values (e.g.,[14,15]). However, attaining a mechanistic understanding of how humans compute aesthetic judgments in the first place from the raw visual input has thus far proved elusive.

[*]corresponding author: kiigaya@caltech.edu.

[†]present address: Department of Psychology, University of California, Berkeley, 2121 Berkeley Way, Berkeley, CA 94720

Author Contributions

K.I. and J.P.O. conceived and designed the project. K.I., S.Y., I.A.W., K.T., performed experiments and K.I., S.Y., I.A.W., K.T., J.P.O. analyzed and discussed results. K.I., S.Y., I.A.W., J.P.O. wrote the manuscript.

Competing interests

The authors declare no competing interests.

A long-standing finding, which partly supports the idiosyncrasy of preference formation, is that prior experience with a specific stimulus can influence value judgment, such as the role of prior episodic memories involving the item, or prior associative history.[2–5,16–18] However, while the influence of past experience on current preference is undeniable, humans can express preferences for completely novel stimuli, suggesting that value judgments can be actively and dynamically computed.

It is an open question how the brain can transform realistically complex stimuli into a simple subjective value. The brain takes a massively high-dimensional input (e.g., a complex art image) and eventually reduces this input to a one-dimensional scalar output (e.g., how much do I like this?). Little is known about how dimensionality reduction can be performed at this scale, while generating reliable output (preference ratings) for all kinds of visual input.

In machine-learning, classification problems (e.g., dog vs. non-dog) are typically solved by projecting an input to a *feature space*.[19] Each feature is a useful attribute that guides the classification of the input. Features can be engineered by taking easily observable characteristics of an object (e.g., its height and weight), or in other cases can be implicitly generated in a more abstracted and less easily interpretable manner (e.g., the activation patterns of hidden layers in deep artificial neural networks).

While previous studies have hinted at the use of such a feature-based framework, in those prior studies the features involved were salient and obvious properties of a stimulus (e.g., multi-attribute artificial stimuli including the movement and the color of dots,[20–22] or items that are suited to a functional decomposition such as food odor[23] or nutritive components;[24] see also[25,26]). However, in the case of visual imagery, the sheer visual complexity of one art piece, as well as the enormous variation between pieces, renders the task of identifying the relevant features that underpin this process exceedingly challenging. In addition, it is not even clear if features are extracted and used for aesthetic judgment in the first place. Moreover, even if relevant features are identified, it is unknown to what extent people may idiosyncratically select the features they use to shape their preferences and how they weigh those features to generate value judgments. Finally, the manner by which a complex visual image gets transformed into relevant features and then into a subjective value, is unclear.

Here, we aimed to establish a general mechanism that could underpin the construction of aesthetic preference. We first extracted features of an art image that have been theorized to play a role in aesthetic valuation.[9,11,27–29] These features reflect subjective judgments about an image, and as such, we deemed them to be “high-level” features, as they required human judgment to determine their presence in an image. We augmented this with a bottom-up process that extracted visual features derived from each image’s statistics and visual properties, a feature set we labeled as “low-level”. We then used ratings from human participants’ across a large set of painting and photography images to ascertain the extent to which we could predict art preferences using our image feature set. Finally, we applied a deep convolutional neural network (DCNN) to establish the degree to which features for computing visual preference might emerge spontaneously while processing visual images in an (approximately) brain-like architecture.

Results

Linear feature summation (LFS) model predicts human valuation of visual art

Participants were asked to report how much they liked various pieces of art (images of paintings). The data were collected from both in-lab (N=7) and online participants using Amazon Mechanical-Turk (N=1359). In-lab participants were recruited from the local community. These participants visited our lab in person and performed the task in a standard laboratory setting, while online participants performed the task over the internet.

On each trial, participants were presented with an image of a painting on a computer screen and asked to report how much they liked it on a scale of 0 (not at all) to 3 (very much) (Figure 1A). Each of the in-lab participants rated all of the paintings without repetition (1001 different paintings), while online participants rated approximately 60 stimuli, each drawn randomly from the image set. The stimulus set consisted of paintings from a broad range of art genres (Figure 1B), and each online participant saw images that were taken with equal proportions from different genres to avoid systematic biases related to style and time-period.

Using this rating data, we tested our hypothesis that the subjective value of an individual painting can be constructed by integrating across features commonly shared across all paintings. For this, each image was decomposed into its fundamental visual and emotional features. These feature values are then integrated linearly, with each participant being assigned a unique set of features weights from which the model constructs a subjective preference (Figure 1C). This model embodies the notion that subjective values are computed in a feature space, whereby overall subjective value is computed as a weighted linear sum over feature content (Figure 1DE). We refer to this model as the Linear Feature Summation (LFS) model.

The LFS model extracts various low-level visual features from an input image using a combination of computer vision methods (e.g.,[27]). This approach computes numerical scores for different aspects of visual content in the image, such as the average hue and brightness of image segments, as well as the entirety of the image itself, as identified by machine learning techniques, e.g., Graph-Cuts[30] (Details of this approach are described in the Methods section). Thus, we note that the LFS model performs a (weighted) linear sum over features, where features can be constructed non-linearly.

The LFS model also includes more abstract or “high-level” attributes that are likely to contribute to valuation. For this, we introduced three features based on previous studies:[28,29] the image is ‘abstract or concrete’, ‘dynamic or still’, ‘hot or cold’, as well as a fourth high-level feature concerning whether the image had a positive or negative emotional valence. Note that “valence” is not necessarily synonymous with valuation: if a piece of art denotes content with a negative emotional tone (e.g., Edvard Munch’s “The Scream”), it can still be judged to have a highly positive subjective value by the art appreciator. We hypothesized that these high-level features are constructed in downstream units using low-level features as input (Figure 1C). However, because we do not know the value of these high-level features a priori, following previous studies[28,29] we invited participants with familiarity and experience

in art (n=13) to provide subjective judgments about the presence of each of these features in each of the images in our stimulus set (though we note a previous study found that artistic experience did not affect feature annotations[28]). We took the average score over these experts' ratings as the input into the model representing the content of each high-level attribute feature for each image.

The final output of the model is a linear combination of low- and high- level features. We assumed that weights over the features are fixed for each individual, which is a necessary requirement to derive generalizable conclusions about the features used to generate valuation across images. As our high-level features were annotated by humans, we treat low-level and high-level features equally, in a non-hierarchical manner, in order to determine the overall predictive power of our LFS model.

We first determined a minimal set of features that can reliably capture rating scores across participants in order to gain insights into aesthetic preference universality. For this, we performed a group-level lasso regression on the data we collected in our in-depth in-lab study (n=7; each rated all 1001 images) using all of the low-level and high-level features that we constructed. By doing so, we removed from consideration those features that do not provide useful predictive information, ultimately selecting the features most uniquely predictive of subjective value, leaving 9 low-level and 4 high-level attribute features. The features include some low-level features computed from the entire images such as the 'mean hue contrast' and the 'blurring effect' as well as some low-level features computed using segmentation methods such as the 'position and the size of the largest segment', in addition to high-level features (please see the Methods for more details). Note that while the integration weights can be tuned for individual participant(s), the feature values for each image remain consistent for all participants.

We then asked how a linear regression model with these features can predict an individual's liking for visual art. Remarkably, we found that we can predict subjective ratings in both a within-, and out-of-, participants manner; the model predicts subjective value not only when we trained the model's weights on the same participant (using a cross-validated procedure) but also when we trained the weights on other in-lab participants, and even when we trained the weights in an entirely independent sample of online participants (Figure 2A. Within-participant prediction, Participant 1: $r = 0.81$, $p < 0.001$ by permutation test, participant 2: $r = 0.42$, $p < 0.001$ by permutation test, participant 3: $r = 0.60$, $p < 0.001$ by permutation test, participant 4: $r = 0.34$, $p < 0.001$ by permutation test, participant 5: $r = 0.33$, $p < 0.001$ by permutation test, participant 6: $r = 0.69$, $p < 0.001$ by permutation test, participant 7: $r = 0.45$, $p < 0.001$ by permutation test. Across-participant prediction using in-lab data, Participant 1: $r = 0.67$, $p < 0.001$ by permutation test, participant 2: $r = 0.36$, $p < 0.001$ by permutation test, participant 3: $r = 0.56$, $p < 0.001$ by permutation test, participant 4: $r = 0.23$, $p < 0.001$ by permutation test, participant 5: $r = 0.27$, $p < 0.001$ by permutation test, participant 6: $r = 0.58$, $p < 0.001$ by permutation test, participant 7: $r = 0.37$, $p < 0.001$ by permutation test. Across-participant prediction using online data, Participant 1: $r = 0.74$, $p < 0.001$ by permutation test, participant 2: $r = 0.35$, $p < 0.001$ by permutation test, participant 3: $r = 0.58$, $p < 0.001$ by permutation test, participant 4: $r = 0.23$, $p < 0.001$ by permutation test, participant 5: $r = 0.27$, $p < 0.001$ by permutation test, participant 6: $r =$

0.67, $p < 0.001$ by permutation test, participant 7: $r = 0.36$, $p < 0.001$ by permutation test. Please also see Supplementary Figure 1).

Similarly, we found that we could reliably predict value ratings for online participants (Figure 2B), not only when training the model on the online participants' data (using leave-one-out cross-validation, mean $r = 0.46$, $p < 0.001$ by permutation test) but also when the model had been trained using in-lab participants' data (mean $r = 0.44$, $p < 0.001$ by permutation test). The difference between these two predictive accuracy was not significant ($p = 0.08$ by permutation test). We also tested the extent to which we can predict value from the low-level attributes alone. Removing the high-level features impaired predictive performance somewhat, but yielded highly significant prediction nonetheless (Figure 2B, mean $r = 0.39$, $p < 0.001$ by permutation test). The difference between the predictions using both low- and high-level features and the prediction using low-level features alone was significant ($p < 0.001$ by permutation test). These results suggest that a non-negligible proportion of the variance in participants' aesthetic ratings can be captured using simple visual features, and can be generalized across people and the art genres that we tested here.

Although we could predict each individual's ratings by training the model on the ratings of others, the degree to which each individual could be predicted from the pooled weights of other participants varied considerably. This suggests that while a common generic model of feature integration can predict individual liking ratings to a surprisingly high degree, there are also likely to be individual differences in how particular features are weighted, which reflects personal aesthetic tastes.

We therefore asked how the model's integration weights for each participant can be varied across participants if we fit the model to each participant separately. We found that the estimated model's feature weights varied across in-lab participants, though all seven of them assigned the largest positive value to the concreteness feature (Figure 2C and Supplementary Figure 2). This preference for concreteness generalized to many of the participants in our much larger scale dataset of online participants. However, we also found that a significant number of participants showed a small or even negative weight on the concreteness feature (Figure 3A).

To better understand the heterogeneity of aesthetic computation across our sample, we aimed to identify potential clusters of individuals that might use features similarly, using the large-scale online participants' data. We fit the LFS model weights to each individual participant, and then fit a Gaussian mixture model to the estimated weights over participants. By comparing the Bayes Information Criteria score between models with a different number of Gaussians, we identified three clusters in the data (Figure 3A,B). Clusters 1 and 2 show somewhat opposing preferences, while cluster 3 shows a distinct preference altogether (Figure 3C). Consistent with our in-lab dataset, the majority of individuals (78%) in our online dataset belonged to Cluster 1, showing a positive weight on the concreteness feature (Figure 3A,B), apparently preferring images of scenery and impressionism (Figure 3D). The remainder belong to one of two other groups: cluster 2 (7%) exhibited a strong preference for dynamic images (e.g., cubism), while cluster 3 (15%) had a large negative weight on concreteness and a positive weight on valence, exhibiting a preference for abstract art and

color fields (Figure 3A,B,D). We also noticed that the difference between clusters was not well described by art categories.

One potential concern might be that the model's performance relies on the preferences for a particular art genre over the other (e.g., people may like impressionism over cubism); however, the same model trained on all images captures significant variations in preference *within* each art genre after taking out the effect of genre preferences (Supplementary Figure 3), suggesting that the model captures variations in subjective preference both within and across genres. Indeed, we found in a representation dissimilarity analysis over visual stimuli, that low-level features seem to capture art genres, but high-level features go beyond the genres (Supplementary Figure 4).

Although our model can capture significant variance in aesthetic liking judgements, it by no means captures everything. We compared our model's out-of-participant performance with the average correlations in ratings between participants, showing that there is significant variance in ratings that the model fails to capture in all of the art genres (Supplementary Figure 5). The latter provides an estimate of a noise ceiling, that is, the variance in ratings that can in principle be predicted from the data. The difference between the model's predictions and the average ratings shows that there is still significant remaining variance that the model fails to capture in all of the art genres.

The above results are based on a linear regression of low- and high-level features, but we also considered the possibility that high-level features are comprised of low-level features (illustrated in Figure 1C). To assess this, we probed the degree to which a linear combination of low-level features could predict the annotated ratings of high-level features. For this, we trained a linear support vector machine using all low-level features as input, and indeed we found that variance ascribed to high-level features could be predicted by low-level features (Figure 2D, concreteness: accuracy = 0.80, $p < 0.001$ by permutation test, dynamics: accuracy = 0.64, $p < 0.001$ by permutation test, temperature: accuracy 0.79, $p < 0.001$ by permutation test, valence: accuracy = 0.75, $p < 0.001$ by permutation test). This suggests that high-level features can be constructed using objective elements of the images, rather than subjective sensations, although the construction may well depend on additional nonlinear operations.

Finally, to test for the effects of the salience of features within the image on the behavioral prediction, we calculated a saliency map for each stimulus using the standard saliency toolbox.[31] Then we re-calculated visual features (11 global features, 20 segmented features) with the saliency map, simply by filtering the features through the saliency map (please see Method for details). We added these saliency-weighted features to the original feature set, and performed linear regression analysis. We however found that the salience map filtered features did not improve the model's predictive accuracy (Supplementary Figure 6).

The LFS model also predicts human valuation of photographs

One potential concern we had was that our ability to predict artwork rating scores using this linear model might be somehow idiosyncratic due to specific properties of the stimuli used in our stimulus-set. To address this, we investigated the extent to which our findings

generalize to other kinds of visual images by using a new image database of 716 images;[32] Figure 4A), this time involving photographs (as opposed to paintings) of various objects and scenes, including landscapes, animals, flowers, and pictures of food. We obtained ratings for these 716 images in a new m-Turk sample of 382 participants. Using the low-level attributes alone (these images were not annotated with high-level features), the linear integration model could reliably predict photograph ratings (Figure 4B). The model performed well when trained and tested on the photograph database ($r = 0.24$, $p < 0.001$ by permutation test), but to our surprise, the same model (as trained on photographs) could also predict the ratings for paintings that we collected in our first experiment ($r = 0.23$, $p < 0.001$ by permutation test), and vice versa (a model trained on the painting ratings could predict photograph ratings. $r = 0.11$, $p < 0.001$ by permutation test). Of note, accuracy was reduced if trained on paintings and tested on photographs (though still highly above chance), suggesting that the photographs enabled improved generalization (possibly because the set of photographs were more diverse). We stress that here, in all cases the model was trained and tested on completely separate sets of participants.

We also tested whether the inclusion of high-level features can improve the model's predictive accuracy in the photograph dataset. Using the support vector machine that is trained on high-level features in the visual art dataset, we estimated binarized high level features in the photograph dataset. We then tested how the model with both low- and high-level features predict ratings in photograph dataset. We found that the full model with both high- and low-level features performs significantly better than the prediction from average ratings, though the direct comparison between the model with low-level alone and the full model did not yield statistical significance (Supplementary Figure 7). This indicates that abstract high-level features can contain information that is generalized across different images, which enables the model to go beyond the average ratings for each image.

A deep convolutional neural network (DCNN) model predicts human liking ratings for visual art

We now have shown that our LFS model can capture subjective preference for visual art; however, as we selected the model's features using a mixture of prior literature and bottom-up machine learning tools, we do not know if this strategy has any biological import. In particular, 1) because we handpicked the LFS model's features, it is not clear if and how a neural system learns to represent these features. It is unlikely that an actual neural system (e.g., human brain) is trained on the features explicitly. Rather, if the LFS model represents a biologically plausible computation, features should emerge out of training on value judgements without explicitly trained on features. Also, 2) it is not clear what kind of network architecture is sufficient to achieve the LFS model's computation. Specifically, it is unknown how a network architecture could end up representing low-level and high-level features hierarchically and integrating them to construct subjective value.

To address these issues, we utilized a standard deep convolutional neural network (DCNN; VGG 16[33]), that had been pre-trained for object recognition with ImageNet.[34] This allows us to test if the computation of the LFS model can be realized in a standard feed-forward network. We used this network with fixed pre-trained weights in convolutional layers, but

trained the weights for the last three fully-connected layers on averaged liking ratings. Mirroring the results of our LFS model, we found that our DCNN model can predict human participants' liking ratings across all participants (Figure 5A. Participant 1: $r = 0.68$, $p < 0.001$ by permutation test, participant 2: $r = 0.34$, $p < 0.001$ by permutation test, participant 3: $r = 0.42$, $p < 0.001$ by permutation test, participant 4: $r = 0.19$, $p < 0.001$ by permutation test, participant 5: $r = 0.18$, $p < 0.001$ by permutation test, participant 6: $r = 0.49$, $p < 0.001$ by permutation test, participant 7: $r = 0.29$, $p < 0.001$ by permutation test). This shows that it is indeed possible to predict preferences for visual art using a deep-learning approach without explicitly selecting stimulus features. In a supplementary analysis, we also opened the convolutional layers to training, but saw no improvement.

The LFS model's features emerge spontaneously in the DCNN model's hidden layers

We turn now to ask whether the features used in our LFS model are spontaneously encoded in the neural network. Mirroring our illustration of the LFS model in Figure 1C, we hypothesized that low-level visual features would be represented in early layers of the DCNN, while more abstract high-level attributes would be represented in later layers of the network. For our investigation, we performed decoding analyses to predict high- and low-level feature values using the activation patterns in each hidden layer. We first reduced the dimensions of each layer using principal component analysis (PCA), and using the top principal components (PCs) that capture 80% of the variance of each layer, we trained a regression model for a given variable that we aimed to predict (e.g., ratings or a feature) using the PCs of each layer.

We first tested to see if we can predict subjective liking ratings using the hidden layer activation patterns. We were able to decode subjective ratings across all layers, but noted that decodability gradually increased for layers deeper in the network (Figure 5B). This came as a surprise since all but the last three layers (layers 1 to 13, out of 16) were pre-trained not on the rating scales being decoded, but on image classifications alone using ImageNet, hinting at a tight relationship between value coding and visual recognition.

We then tested to see how the hidden layers related to the LFS model's features. This analysis showed that hidden layers could predict all 23 features included in the LFS model. Consistent with our hypothesis, six (of the 19) putative low-level features tested were represented more robustly in early layers, as shown by a significantly negative decoding slope across layers (Figure 5CD). We also found four (out of 19) low-level features had a decoding accuracy that increased as a function of the depth of the layer, suggesting those low-level features, in fact, may be better identified as high-level features. However, we note that the overall predictive accuracy of these features was low compared to those showing negative slopes. These positive slope features include: "the presence of a person", "the mass center for the largest segment", "mass variance of the largest segment", and "entropy in the 2nd largest segment", all of which require relatively complex computations (e.g., segmentation and the identification of the location of the segments) compared to the ones showing negative slopes (e.g., average saturation). We note that this result is consistent with a previous electrophysiological and computational modeling study in macaques,[35] which reported that the position of an object on the screen is more robustly represented in

higher visual areas and deeper layers, as position identifications of a segment and an object likely involve similar computations. These object-related features were also referred to as 'low-level' features,[35] in line with our original reference. The other 9 low-level features tested did not show either a strong positive or negative slope.

Similarly, for the putative high-level features, we found that two (of 4) features were more robustly represented in later layers (Figure 5EF). However, "temperature", which was labelled as a high-level feature, showed a significant negative slope. Given that this feature is based on color palettes in the image (i.e., whether the color palette is hot or cold), this feature's variance may already be well captured by low-level image statistics. The fourth putative high-level feature, valence, did not show either an increasing or decreasing trend in decoding across layers. Thus, the DCNN allows a more principled means to identify low and high-level features, enabling us to label 6 features as low-level (based on greater representation of those features in earlier layers of the network), and 6 as high-level features, indicated by representations that are present to a greater extent in later layers of the network. These LFS model-based analyses on the DCNN sheds light into what are often-considered-to-be "black box" computations in deep artificial neural networks, and may provide an empirical definition of computational complexity in feature extraction of visual as well as other sensory inputs.

Taken together, our DCNN analyses suggest that our conceptualized LFS model (Figure 1C) is, in fact, a natural consequence of training the neural network on object recognition and predicting subjective aesthetic value, without requiring any explicit feature engineering.

Discussion

Whether we can lawfully account for personal preferences in the aesthetic appreciation of art has long been an open question in the arts and sciences.[1,2,4,7] Here, we addressed this question by engineering a hierarchical linear feature summation (LFS) model that generates subjective preference according to a weighted mixture of explicitly designed stimulus features. This model was verified with both in-depth lab-based small scale behavioural experiments and large-scale on-line behavioral experiments, and contrasted to a deep convolutional neural network (DCNN) model. We found that it is indeed possible to predict subjective valuations for both paintings and photography using the same feature set, and we demonstrate hierarchical feature representations in a DCNN that predicts aesthetic valuations.

Our results indicate that linearly integrating a small set of visual features can explain human preferences for paintings and photography. Not only is it possible to predict an individual's ratings based on that particular individual's prior ratings for other images, but we also found that this strategy allowed us to predict one individual's preferences from the preferences of others, even for novel stimuli. This is achievable likely because the majority of participants shared substantial variance in their preferences, and the model efficiently extracted this, as shown by our clustering analysis whereby one dominant cluster was found to account for the majority of participants' liking ratings. Our results are consistent with a number of empirical

aesthetics studies proposing that statistical properties of images can account for aesthetic values (e.g.,[12,36,37]).

We also found that the LFS model with the same visual feature set can predict subjective values for both visual art and for diverse photographic stimuli. This suggests that the features used for visual aesthetic judgement may not be domain-specific but universal, relying on a small set of visual features shared across visual stimuli. Our findings also hint that the extraction of these features might be a natural consequence of developing a visual system. We found that a DCNN model trained on object recognition and valuation represents those features throughout the hidden layers. Further studies could investigate whether such feature-extractions and feature-based value judgement are universal computations not only in visual processing but also other sensory domains such as in audition and olfaction.

The cluster analysis we ran on the large-scale online study showed that there is variation in preferences across individuals. A substantial component of that variation is whether or not participants liked concrete art or abstract art: the majority assigned large positive weights to concreteness, while the others assigned large negative weights. This indicates that concreteness alone is explaining a substantial part of the variance, and accounting for variation in preferences across groups of individuals. However, it should be noted that while concreteness does account for a substantial portion of variance in people's preferences, other high-level features also play an important role, including dynamics and valence.

We also note that, though such high-level features, including concreteness, can be used to predict preference, much if not most of the significant variance explained by such high-level features can also be explained directly as a linear combination of a number of low-level visual features. This is consistent with the idea that many low-level sensory features (that are present in early layers in DCNN) are transformed into a smaller number of task-specific high-level features (that are present in deeper layers in DCNN), which are in turn used to predict subjective value.

It is also important to note there was in addition variance across individuals in preferences that the model did not capture well.[38,39] Thus, while art preferences share some commonalities across clusters of individuals, there is in addition some degree of individual variability. We found that the degree of commonality in subjective ratings in our study was in a similar range to previous studies (e.g., the average correlation between our M-turk samples was 0.45, which is similar to[38]).

We nonetheless should stress that in our study there are two limitations. One is that most of our in-lab and online participants are not art experts and it thus remains possible that artistically experienced people might judge artworks differently. The second is that we only covered a relatively narrow subset of art genres, leaving open the possibility that there may be some art genres for which the model may not perform well. That said, we also validated our models using a wide range of photographs, indicating the potential generalizability of our findings even beyond drawings. In fact, previous studies (e.g.,[37]) suggest that artworks and natural scenes share some statistical regularities, of which our model might be able to take advantage.

It should also be noted that the predictive power of our model varied across participants. One possibility is that some participants were more reliable/consistent in reporting their ratings. Unfortunately, we did not present the same stimuli multiple times to each participant, making it difficult to directly test the consistency of participants' choices. However we did present the same set of stimuli to each participant. We thus directly tested how ratings of each painting were similar across participants. To test this, we computed the average ratings over n-1 participants of each painting and computed correlation between the average ratings and the ratings of the remaining participant. We performed this for each participant, and found that the correlation systematically co-varied with the within participant predictions of the model. This suggests that variability in predictability is largely due to noise in participant's preferences.

Here, utilizing a set of interpretable visual and emotional features, we showed that these features are employed by individuals to make value judgments for art. We note that this is by no means a complete enumeration of the features used by humans. For instance, the semantic meaning of a painting, its historical importance, as well as memories of past experiences elicited by the painting, are also likely to play important roles (see, e.g.,[5,9]). Thus, rather than offering a feature catalogue, our findings shed light on the general principles by which feature integration yields to aesthetic valuation. However, the features that we identified are likely to be important, particularly as we utilized a reasonably large set of potential features in our initial feature set which was subsequently narrowed down to a set of only the most relevant features.

We found evidence using a DCNN model that the features engineered for use in the LFS model spontaneously emerge in a neurally plausible manner. Our deep network model was not explicitly trained on any of the LFS model's features, nor were the convolutional hidden layers of the network trained on liking ratings (only the later fully connected layers were trained using rating scores). Nevertheless, we were still able to identify LFS features from the hidden layers of the network, which suggests that the features used for aesthetic valuation likely emerge spontaneously through more basic and generalizable aspects of visual development. Further, those features may well be utilized for a wide range of visual tasks, including object classification, prediction, and identification. Thus, we speculate that these findings suggesting a common feature space shared across different tasks may provide insights into transfer learning[40] in machine learning.

One important consideration is whether linear feature operations are sufficient to describe the computations underlying aesthetic valuation. Notably, the highly non-linear deep network did not substantively outperform the simple linear model. However, in the LFS model, the feature extraction process itself is not necessarily linear (e.g., segmentation). As such, our results do not rule out the possibility of nonlinearity in feature extraction processes in the brain, but they do suggest that the final feature value integration for computing subjective art valuation can be approximated by a linear operation. This computational scheme resonates with a widely-used machine-learning technique referred to as the kernel method, whereby inputs are transformed into a high-dimensional feature space in which categories are linearly separable,[41,42] as well as with high-dimensional task-related variables represented in the brain.[43]

Although the deep neural network approach has been successfully applied to a wide range of machine learning problems (e.g.,[44–46]), the underlying computational mechanisms that deep neural networks leverage in order to attain high performance across domains are opaque and often poorly understood. Here, we provide evidence that the hidden layers in the deep network encode low-level and high-level features relevant for computing aesthetic visual preferences in a hierarchical manner, which are utilized to produce coherent behavioral outputs. Thus our study provides a clear link from distributed neuronal computations to interpretable, explicit, hierarchical feature representations. Our study thus highlights the merits of a model-based analysis of artificial neural networks in order to better understand the nature of the computations implemented therein. We however caution that we do not claim that the DCNN model necessarily provides a plausible account of actual neural computations going on in the brain. Unlike a DCNN which is exclusively feedforward in its connections between layers, the brain is heavily recurrent, and thus is likely to be better approximated by networks with recurrent architecture.

The present findings offer a mechanism through which artistic preferences can be predicted. It is of course important to note that aesthetic experience more broadly defined goes beyond the simple one-dimensional liking rating (a proxy of valuation) that we study here (e.g.,[4,7,9]), and that judgments can be context-dependent.[47] Art is likely to be perceived along many dimensions, of which valuation is but one, with some dimensions relying more on idiosyncratic experience than others. Nevertheless, we speculate that just as it is possible to explain aesthetic valuation in terms of underlying features, many other aspects of the experience of art can also likely be decomposed into more atomic feature-based computations, with different dimensions employing different weights over those features. Indeed, subjective value can itself be considered to be a “feature” in a feature space, albeit a high-level one, alongside other judgments that might be made about a piece of art. Further, although we did not find evidence for this in the present study, it is undoubtedly the case that various psychological processes such as attention are likely to dynamically modulate the relative weights over features that construct subjective value, as well as modulating underlying neural activity.[26]

To conclude, our findings demonstrate that once relevant features are extracted from a visual image, it is possible to dynamically construct values that can account for actual art preferences. Thus, far from being inscrutable and idiosyncratic, with modern computational and machine-learning tools, aesthetic valuation can be lawfully described and its mechanisms illuminated.

Methods

Participants

Participants provided informed consent for their participation in the study, which was approved by the Caltech IRB. Participants received monetary compensation according to their task lengths.

Online study: A total of 1936 volunteers (female: 883 (45.6%). age 18–24 yr: 285 (14.8%); 25–34 yr: 823 (42.8%); 35–44 yr: 435 (22.6%); 45 yr and above: 382 (19.8%))

participated in our on-line studies in the Amazon Mechanical Turk (M-turk). 1545 of them participated in the ART task, and 391 of them participated in the AVA photo task. Among these, participants who missed trials and failed to complete 50 trials were excluded from our analyses, leaving us with online 1359 participants in the ART task data and 382 participants in the AVA photo task data.

In-lab study: Seven volunteers (female: 3. age 18–24 yr: 1; 25–34 yr: 4; 35–44 yr: 2. 4 Asian, 3 Caucasian) were recruited to our in-lab study.The in-lab participants did not include lab members but were instead recruited from the local community in Pasadena. Seven participants completed master's degree or higher. None of the participants possessed an art degree. Six of the participants reported that they visit art museums less than once a month, while one participant reported visiting art museums at least once but less than four times a month.

Additionally, thirteen art-experienced participants (female: 6. age 18–24 yr: 3; 25–34 yr: 9; 35–44 yr: 1) were invited to evaluate the high-level feature values. These participants for annotation were primarily recruited from the ArtCenter College of Design community.

Stimuli

Painting stimuli were taken from the visual art encyclopedia www.wikiart.org. Using a script that randomly selects images in a given category of art, we downloaded 206 or 207 images from four categories of art (825 in total). The categories were 'Abstract Art', 'Impressionism', 'Color Fields', and 'Cubism'. We randomly downloaded images with each tag using our custom code in order to avoid subjective bias. For the in-lab studies, we supplemented this database with an additional 176 paintings that were used in a previous study.[29] As per the journal's request, in the main figures, we present images purchased from Alamy.com. The color field paintings are different from our original stimuli due to unavailability.

Picture images were taken from the Aesthetic Visual Analysis (AVA) dataset. This dataset consists of images from multiple online photo contests. We took images from the following categories (about 90 images from each): 'Animals', 'Floral', 'Nature', 'Sky', 'Still Life', 'Advertisement', 'Sky', and 'Abstract Pictures'. In a total of 716 images were used.

Tasks

Behavioural task

Liking rating task: On each trial, participants were presented with an image of the artwork (in the Art-liking Rating Task: ART) or a picture image (in the AVA photo task) on the computer screen. Participants reported within 6 seconds how much they like the artwork (or the picture image), by pressing buttons corresponding to a scale that ranged from 0, 1, 2, 3, where 0 = not like at all, 1= like a little, 2 = like, and 3 = strongly like, presented at the bottom of the image. Each of the on-line ART participants performed on average 57 trials of the rating task, followed by a familiarity task in which they reported if they could recognize the name of the artist who painted the artwork for the same images that they reported their liking ratings. The images for each online participant were drawn to balance different art

genres. On-site ART participants performed 1001 trials of rating tasks. On-site participants had a chance to take a short break approximately every 100 trials. Each of the on-line AVA photo task participants performed on average 115 trials of the rating task.

Feature annotation: The four high-level features were annotated in a manner following.[28,29] On each trial, participants were asked about the feature value of a given stimulus, ranged from −2, −1, 0, 1, 2. Following,[28] example figures showing extreme feature values are always shown on the screen as a reference (please see Supplementary Figure 8). Each participants completed four separate tasks (for four features) in a random order, where each task consists of 1001 trials (with 1001 images).

Linear feature summation model (LFS model)—We hypothesized that subjective preferences for visual stimuli are constructed by the influence of visual and emotional features of the stimuli. As its simplest, we assumed that the subjective value of the i-th stimulus v_i is computed by a weighted sum of feature values $f_{i,j}$:

$$v_i = \sum_{j=0}^{n_f} w_j f_{i,j} \quad (1)$$

where w_j is a weight of the j-th feature, $f_{i,j}$ is the value of the j-th feature for stimulus i, and n_f is the number of features. The 0-th feature is a constant $f_{i,0} = 1$ for all i's.

Importantly, w_j is not a function of a particular stimulus but shared across all visual stimuli, reflecting the *taste* of a participant. The same taste (w_j's) can also be shared across different participants, as we showed in our behavioral analysis. The features $f_{i,j}$ were computed using visual stimuli; we used the same feature values to predict liking ratings across participants. We used the simple linear model Eq.(1) to predict liking ratings in our behavioral analysis (please see below for how we determined features and weights).

As we schematically showed in Figure 1, we hypothesized that the input stimulus was first broke down into low-level features and then transformed into high-level features, and indeed we found that a significant variance of high-level features can be predicted by a set of low-level features.

Features—Because we did not know apriori what features would best describe human aesthetic values for visual art, we constructed a large feature set using previously published methods from computer vision augmented with additional features that we ourselves identified using additional existing machine learning methods.

Visual low-level features introduced in[27]: We employed 40 visual features introduced in.[27] We do not repeat descriptions of the features here; but briefly, the feature sets consist of 12 global features that are computed from the entire image that include color distributions, brightness effects, blurring effects, and edge detection, and 28 local features that are computed for separate segments of the image (the first, the second and the third largest segments). Most features are computed straightforwardly in either HSL (hue, saturation, lightness) or HSV (hue, saturation, value) space (e.g. average hue value).

One feature that deserves description is a blurring effect. Following,[27,48] we assumed that the image I was generated from a hypothetical sharp image with a Gaussian smoothing filter with an unknown variance σ. Assuming that the frequency distribution for the hypothetical image is approximately the same as the blurred, actual image, the parameter σ represents the degree to which the image was blurred. The σ was estimated by the Fourier transform of the original image by the highest frequency, whose power is greater than a certain threshold.

$$f_{blur} = max\left(k_x, k_y\right) \propto \frac{1}{\sigma}$$

where $k_x = 2(x - n_x/2)/n_x$ and $k_y = 2(y - n_y/2)/n_y$ with (x, y) and (n_x, n_y) are the coordinates of the pixel and the total number of pixel values, respectively. The above max was taken within the components whose power is larger than four.[27]

The segmentation for this feature set was computed by a technique called kernel Graph-Cut.[30,49] Following,[27] we generated a total of at least six segments for each image using a C^{++} and MATLAB package for kernel graph cut segmentation.[49] The regularization parameter that weighs the cost of cut against smoothness was adjusted for each image in order to obtain about six segments. Please see[27,49] for the full description of this method and examples.

Of these 40 features, we included all of them in our initial feature set except for local features for the third-largest segment, which were highly correlated with features for the first and second-largest segments and were thus deemed unlikely to add unique variance to the feature prediction stage.

Additional Low-Level Features: We assembled the following low-level features to supplement the set by Li & Chen.[27] These include both global features and local features. Local features were calculated on segments determined by two methods. The first method was statistical region merging (SRM) as implemented by,[50] where the segmentation parameter was incremented until at least three segments were calculated. The second method converted paintings into LAB color space and used k-means clustering of the A and B components. While the first method reliably identified distinct shapes in the paintings, the second method reliably identified distinct color motifs in the paintings.

The segmentation method for each feature is indicated in the following descriptions. Each local feature was calculated on the first and second-largest segments.

Local Features:

- Segment Size (SRM): Segment size for segment i was calculated as the area of segment i over the area of the entire image:

$$f_{\text{segment size}} = \frac{\text{area segment } i}{\text{total area}} \quad (2)$$

- HSV Mean (SRM): To calculate mean hue, saturation and color value for each segment, segments were converted from RGB to HSV color space.

$$f_{\text{mean hue}} = \text{mean(hue values in segment } i) \tag{3}$$

$$f_{\text{mean saturation}} = \text{mean(saturation values in segment } i) \tag{4}$$

$$f_{\text{mean color value}} = \text{mean(color values in segment } i) \tag{5}$$

- Segment Moments (SRM):

$$f_{\text{CoM X coordinate}} = \frac{\sum_{k \in \text{segment } i} x_k}{\text{area segment } i} \tag{6}$$

$$f_{\text{CoM Y coordinate}} = \frac{\sum_{k \in \text{ segment } i} y_k}{\text{area segment } i} \tag{7}$$

$$f_{\text{Variance}} = \frac{\sum_{k \in \text{segment } i} (x_k - \bar{x})^2 + (y_k - \bar{y})^2}{\text{area segment } i} \tag{8}$$

$$f_{\text{Skew}} = \frac{\sum_{k \in \text{segment } i} (x_k - \bar{x})^3 + (y_k - \bar{y})^3}{\text{area segment } i} \tag{9}$$

where $(\bar{x}, \bar{y})$ is the center of mass coordinates of the corresponding segment.

- Entropy (SRM):

$$f_{\text{entropy}} = -\sum_j \left(p_j * \log_2(p_j)\right) \tag{10}$$

where p equals the normalized intensity histogram counts of segment i.

- Symmetry (SRM): For each segment, the painting was cropped to maximum dimensions of the segment. The horizontal and vertical mirror images of the rectangle were taken, and the mean squared error of each was calculated from the original.

$$f_{\text{horizontal symmetry}} = \frac{\sum_{x,y \in \text{segment}} \left(\text{segment}_{x,y} - \text{horizontal\_flip(segment)}_{x,y}\right)^2}{\#\text{ pixels in segment}} \tag{11}$$

$$f_{\text{vertical symmetry}} = \frac{\sum_{x,y \in \text{segment}} \left(\text{segment}_{x,y} - \text{vertical\_flip(segment)}_{x,y}\right)^2}{\#\text{ pixels in segment}} \tag{12}$$

- R-Value Mean (K-Means): Originally, we took the mean of R, G, and B values for each segment, but found these values to be highly correlated, so we reduced these three features down to just one feature for mean R value.

$$f_{\text{R-value}} = \text{mean(R-values in segment)} \tag{13}$$

- HSV Mean (K-Means): As with SRM generated segments, we took the hue, saturation, and color value means of segments generated by K-means segmentation as described in equations 2–4.

Global Features:

- Image Intensity: Paintings were converted from RGB to grayscale from 0 to 255 to yield a measure of intensity. The 0–255 scale was divided into five equally-sized bins. Each bin count accounted for one feature.

$$f_{\text{intensity count bin } i \in \{1,4\}} = \frac{\#\ \text{pixels with intensity} \in \left\{\frac{255(i-1)}{5}, \frac{255i}{5}\right\}}{\text{total area}} \tag{14}$$

- HSV Modes: Paintings were converted to HSV space, and the modes of the hue, saturation, and color value across the entire painting were calculated. While we took mean HSV values over segments in an effort to calculate overall-segment statistics, we took the mode HSV values across the entire image in an effort to extract dominating trends across the painting as a whole.

$$f_{\text{mode hue}} = \text{mode(hue values in segment } i) \tag{15}$$

$$f_{\text{mode saturation}} = \text{mode(saturation values in segment } i) \tag{16}$$

$$f_{\text{mode color value}} = \text{mode(color values in segment } i) \tag{17}$$

- Aspect (width-height) Ratio:

$$f_{\text{aspect ratio}} = \frac{\text{image width}}{\text{image height}} \tag{18}$$

- Entropy: Entropy over the entire painting was calculated according to equation 9.

In addition, we also annotated our image set with whether or not each image included a person. This was done by manual annotation, but it can also be done with a human detection algorithm (e.g., see[51]). We included this presence-of-a-person feature in the low-level feature set originally, though we found in our DCNN analysis that the feature shows a signature of a high-level feature. As we showed in the main text, classifying this feature as a low-level feature or as a high-level feature does not change our results. We also note that the presence-of-a-person was not included when we analyzed photography ratings.

High-Level Feature Set[28,29]: We also introduced features that are more abstract and not easily computed by a simple algorithm. In,[28] Chatterjee et al. pioneered this by introducing 12 features (color temperature, depth, abstract, realism, balance, accuracy, stroke, animacy, emotion, color saturation, complexity) that were annotated by human participants for 24 paintings, in which the authors have found that annotations were consistent across participants, regardless of their artistic experience. Vaidya et al.[29] further collected annotations of these feature sets from artistically experienced participants for an additional 175 paintings and performed a principal component analysis, finding three major components that summarize the variance of the original 12 features. Inspired by the three principal components, we introduced three high-level features: concreteness, dynamics, and temperature. Also, we introduced valence as an additional high-level feature.[52] The four high-level features were annotated in a similar manner to the previous studies.[28,29] We took the mean annotations of all 13 participants for each image as feature values.

Salience-map filtered Feature Set: In order to test the influence of visual salience of each image on features, we constructed features that are filtered with salience maps. For this, we computed a salience map of each stimulus using the salience toolbox.[31] Then we recalculated feature values using pixel data multiplied with the salience maps. We included 11 global features (average hue, average saturation, average color value, image intensity (5 bins), R mean, G mean, B mean), as well as 20 local features (10 largest and 10 second largest segments of average hue, average saturation, average color value, horizontal center of mass, vertical center of mass, variance, skewness, entropy, left-right symmetry, and top-bottom symmetry). Segments were created by taking everywhere the salience map was not zero, and using the boundaries function to create polygons of these non-zero areas.

Identifying the shared feature set that predicts aesthetic preferences: The above method allowed us to have a set of 83 features in total that are possibly used to predict human aesthetic valuation. These features are likely redundant because some of them are highly correlated, and many may not contribute to decisions at all. We thus sought to identify a minimal subset of features that are commonly used by participants. We performed this analysis using MATLAB Sparse Gradient Descent Library (https://github.com/hiroyuki-kasai/SparseGDLibrary). For this, we first orthogonalized features by sparse PCA.[53] Then we performed a regression with a LASSO penalty at the group level using the seven in-lab participants' behavioral data with a function *group – lasso – problem*. We used Fast Iterative Soft Thresholding Algorithm (FISTA) with cross-validation. After eliminating PC's that were not shared by more than one participant, we transformed the PC's back to the original space. We then eliminated one of the two features that were most highly correlated ($r^2 > 0.5$) to obtain the final set of shared features.

The identified shared features from the in-lab behavioral participants are the following: the concreteness, the dynamics, the temperature, the valence, the global hue contrast from,[27] the global brightness contrast from,[27] the blurring effect from,[27] the vertical center of largest segment using Graph-cut from,[27] the average saturation of the second-largest segment using Graph-cut from,[27] the blurring contrast between the largest and the second largest segments using Graph-cut in,[27] the size of largest segment using SRM, width-height ratio, and the

presence of a person. Interestingly, most local features that were computed by Graph-Cut ended up being included as a result of the LASSO procedure, while other local features relying on other segmentation methods ended up being removed.

Representation similarity analyses: We computed a representation dissimilarity matrix over visual art stimuli using low-level or high-level features alone. The distance was measured by one minus correlation.

Model fitting: We tested how our shared-feature model can predict human liking ratings using out-of-sample tests. All models were cross-validated in twenty folds, and we used ridge regression unless otherwise stated. Hyperparameters were tuned by cross-validation. In one analysis, we trained our regression model with a shared feature set on the average ratings of six in-lab participants and tested on the remaining participant. We calculated the Pearson correlation between model predictions (pooled predictions from all cross-validation sets) and actual data, and defined it as the predictive accuracy. In another analysis, we trained our regression model on one participant and tested on the same participant (randomly partitioned to 20 folds). In a further analysis, we also trained our model on the average ratings of on-line participants and tested it on in-lab participants, thereby ensuring complete independent between our training data and testing data. We also tested our model on on-line participants. In one analysis, we trained our model on $n - 1$ participants and tested on the remaining participant (leave-one-out). In another, we trained our model on in-lab participants and tested on-line participants. We also tested our model with low-level features only. For this, we trained our model with all low-level features with a lasso penalty on $n - 1$ on-line participants and tested on the remaining online participant.

We also estimated individual participant's feature weights by fitting a linear regression model with the shared feature set to each participant. For illustrative purposes, the weights were normalized for each participant by the maximum feature value (concreteness) in Figures 1G.

We further tested whether the low-level feature set could predict high-level features. For this, we used a support vector machine with cross-validation (fitcsvm in MATLAB).

We performed the same analysis with our dataset for photo contest images (AVA). Because we do not have high-level feature ratings fro the photo contest database and we also did not obtain manual ratings of the presence of a person for that image database, we only used low-level features (minus the presence of a person feature) for analyzing this data. In one analysis, we trained our model on the AVA dataset and tested it on the AVA dataset. In another analysis We trained our model on the average ratings of online ART dataset and tested it on the AVA dataset. Also, we trained our model on the AVA dataset ratings and tested it on the online ART dataset ratings. We used all low-level features with a strong lasso penalty and cross-validation to avoid over-fitting. The hyperparameters were optimized by cross-validation.

The significance of the above analyses was measured by generating a null distribution constructed by the same analyses but with permuted image labels. The null distribution was

construed by 10000 permutations. The chance level was determined by the mean of the null distribution.

Cluster analysis: We explored individual differences in the feature space by applying a clustering analysis of on-line participants. For this, we first transformed our shared-feature set to principal component (PC) space and fit our model in the PC space. The obtained weights were normalized by using each participant's weight with the maximum magnitude. Then we fit a Gaussian mixture model with a different number of nodes to the resulting weights, assuming that the off-diagonal terms of the covariance are zero because the fit is in the PC space. We used an expectation-maximization method with 100 different random starting points, using MATLAB's function fitgmdist. We compared the Bayesian Information Criterion scores for each fit. We then compared the results of models with the number of clusters set to n=1,2,3,4,5,6 and took the model with the minimum BIC score. This turned out to be a model with n=3 clusters. Hard clustering was used, where each data point in the weight space was assigned to the component yielding the highest posterior probability.

Deep Convolutional Neural Network (DCNN) analysis

Network architecture: The deep convolutional neural network (DCNN) we used consists of two parts. An input image feeds into convolutional layers from the standard VGG-16 network that is pre-trained on ImageNet. The output of the convolutional layers then projects to fully connected layers. This architecture follows the current state-of-the-art model on aesthetic evaluation.[32,54]

The details of the convolutional layers from the VGG network can be found in;[33] but briefly, it consists of 13 convolutional layers and 5 intervening max pooling layers. Each convolutional layer is followed by a rectified linear unit (ReLU). The output of the final convolutional layer is flattened to a 25088-dimensional vector so that it can be fed into the fully connected layer.

The fully connected part has two hidden layers, where each layer has 4096 dimensions. The fully connected layers are also followed by a ReLU layer. During training, a dropout layer was added with a drop out probability 0.5 after every ReLU layer for regularization. Following the current state of the art model,[54] the output of the fully connected network is a 10-dimensional vector that is normalized by a softmax. The output vector was weighted averaged to produce a scalar value[54] that ranges from 0 to 3.

Network training: We trained our model on our in-lab behavioral data set by tuning weights in the fully connected layers. Training all layers on the ART and/or AVA data set are also possible, but we found that the model's predictive performance does not improve even if we trained all layers. We employed 10-fold cross-validation to benchmark the art rating prediction.

The model was optimized using a Huber loss metric, which is robust to outliers.[55] The average rating of each image among the in-lab participants was used as the ground truth.

We used stochastic gradient descent (SGD) with momentum to train the model. We used a batch size of 100, a learning rate of 10^{-4}, the momentum of 0.9, and weight decay of 5×10^{-4}. The learning rate decayed by a factor of 0.1 every 30 epochs.

To handle various sizes of images, we used the zero-padding method. Because our model could only have a 224 × 224 sized input, we first scaled the input images to have the longer edges be 224 pixels long. Then we filled the remaining space with 0 valued pixels (black).

We used Python 3.7, PyTorch 0.4.1.post2, and CUDA 9.0 throughout the analysis.

Retraining DCNN to extract hidden layer activations: We also trained our network on a single fold ART data in order to obtain a single set of hidden layer activations. We prevented over-fitting by stopping our training when the model performance (Pearson correlation between the model's prediction and data) reached the mean correlation from the 10-folds cross-validation.

Decoding features from the deep neural network: We decoded the LFS model's features from hidden layers by using linear (for continuous features) and logistic (for categorical features) regression models. We considered the activations of outputs of ReLU layers (total of 15 layers). First, we split the data into ten folds for the 10-fold cross-validation. In each iteration of the cross-validation, because dimensions of the hidden layers are much larger ($64 \times 224 \times 224 = 3211264$) than the actual data size, we first performed PCA on the activation of each hidden layer from the training set. The number of principal components was chosen to account for 80% of the total variance. By doing so, each layer's dimension was reduced to less than 536. Then the hidden layers' activations from the test set were projected onto the principal component space by using the fitted PCA transformation matrices. The regularization coefficient for the regression was tuned by doing a grid search, and the best performing coefficient for each layer and feature was chosen based on the scores from the 10-folds cross-validation. We tested for a total of 19 features, including all 18 features that we used for our fMRI analysis,[56] as well as the simplest feature that was not included into our fMRI analysis (as a result of our group-level feature selection) but that was also of interest here: the average hue value. In a supplementary analysis, we also explored whether adding 'style matrices' of hidden layers[57] to the PCA-transformed hidden layer's activations can improve the decoding accuracy; however, we found the style matrices do not improve the decoding accuracy. Sklearn 0.19.2 on Python 3.7 was used.

Reclassifying features according to the slopes of the doodling accuracy across hidden layers: In our LFS model, we classified putative low-level and high-level features simply by whether a feature is computed by a computer algorithm vs annotated by humans respectively. In reality, however, some putative low-level features are more complex in terms of how they could be constructed than other lower level features, while some putative high-level features could in fact be computed straightforwardly from raw pixel inputs. Using the decoding results of the features from hidden layers in the DCNN, we identified DCNN-defined low-level and high-level features. For this, we fit a linear slope to the estimated decoding accuracy vs hidden layers. We permuted layer labels 10,000 times and performed the same analysis to construct null distribution as described earlier. We classified a feature as

high-level if the slope was significantly positive at $p < 0.001$, and we classified a feature as a low-level feature if the slope was signifcantly negative at $p < 0.001$.

The features showing negative slopes were: the average hue, the average saturation, the average hue of the largest segment using GraphCut, the average color value of the largest segment using GraphCut, the image intensity in bin 1, the image intensity in bin 3, and the temperature.

The features showing positive slopes were: the concreteness, the dynamics, the presence of a person, the vertical coordinate of the mass center for the largest segment using the Graph Cut, the mass variance of the largest segment using the SRM, the entropy in the 2nd largest segment using SRM. All of these require relatively complex computations, such as localization of segments or image identification. This is consistent with a previous study showing that object-related local features showed a similar increased decodability at a deeper layer.[35]

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The code that support the findings of this study are available from the corresponding author upon reasonable request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Peter Dayan, Shin Shimojo, Pietro Perona, Lesley Fellows, Avinash Vaidya, Jeff Cockburn and Logan Cross for discussions and suggestions. We also thank Seiji Iigaya and Erica Iigaya for drawing color fields paintings presented in this manuscript. This work was supported by NIDA grant R01DA040011 and the Caltech Conte Center for Social Decision Making (P50MH094258) to JOD, the Japan Society for Promotion of Science the Swartz Foundation and the Suntory Foundation to KI, and the William H. and Helen Lang SURF Fellowship to IW. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

1. Kant I Critique of judgment (Hackett Publishing, 1987).
2. Fechner GT Vorschule der aesthetik, vol. 1 (Breitkopf & Härtel, 1876).
3. Ramachandran VS & Hirstein W The science of art: A neurological theory of aesthetic experience. Journal of consciousness Studies 6, 15–51 (1999).
4. Zeki S Inner vision: An exploration of art and the brain (2002).
5. Leder H, Belke B, Oeberst A & Augustin D A model of aesthetic appreciation and aesthetic judgments. British journal of psychology 95, 489–508 (2004). [PubMed: 15527534]
6. Biederman I & Vessel EA Perceptual pleasure and the brain: A novel theory explains why the brain craves information and seeks it through the senses. American scientist 94, 247–253 (2006).

7. Chatterjee A Neuroaesthetics: a coming of age story. Journal of cognitive neuroscience 23, 53–62 (2011). [PubMed: 20175677]
8. Shimamura AP & Palmer SE Aesthetic science: Connecting minds, brains, and experience (OUP USA, 2012).
9. Palmer SE, Schloss KB & Sammartino J Visual aesthetics and human preference. Annual review of psychology 64, 77–107 (2013).
10. Leder H & Nadal M Ten years of a model of aesthetic appreciation and aesthetic judgments: The aesthetic episode–developments and challenges in empirical aesthetics. British Journal of Psychology 105, 443–464 (2014). [PubMed: 25280118]
11. Chatterjee A Prospects for a cognitive neuroscience of visual aesthetics. Bull. Psychol. Art 4 (2003).
12. Bar M & Neta M Humans prefer curved visual objects. Psychological science 17, 645–648 (2006). [PubMed: 16913943]
13. Van Paasschen J, Zamboni E, Bacci F & Melcher D Consistent emotions elicited by low-level visual features in abstract art. Art & Perception 2, 99–118 (2014).
14. Cela-Conde CJ et al. Activation of the prefrontal cortex in the human visual aesthetic perception. Proceedings of the National Academy of Sciences 101, 6321–6325 (2004).
15. Kawabata H & Zeki S Neural correlates of beauty. Journal of neurophysiology 91, 1699–1705 (2004). [PubMed: 15010496]
16. Weber EU & Johnson EJ Constructing preferences from memory. The Construction of Preference, Lichtenstein S & Slovic P,(eds.) 397–410 (2006).
17. Wimmer GE & Shohamy D Preference by association: how memory mechanisms in the hippocampus bias decisions. Science 338, 270–273 (2012). [PubMed: 23066083]
18. Barron HC, Dolan RJ & Behrens TE Online evaluation of novel choices by simultaneous representation of multiple memories. Nature neuroscience 16, 1492 (2013). [PubMed: 24013592]
19. Bishop CM Pattern recognition and machine learning (springer, 2006).
20. Kahnt T, Heinzle J, Park SQ & Haynes J-D Decoding different roles for vmpfc and dlpfc in multi-attribute decision making. Neuroimage 56, 709–715 (2011). [PubMed: 20510371]
21. Mante V, Sussillo D, Shenoy KV & Newsome WT Context-dependent computation by recurrent dynamics in prefrontal cortex. nature 503, 78 (2013). [PubMed: 24201281]
22. Pelletier G & Fellows LK A critical role for human ventromedial frontal lobe in value comparison of complex objects based on attribute configuration. Journal of Neuroscience 39, 4124–4132 (2019). [PubMed: 30867258]
23. Howard JD & Gottfried JA Configural and elemental coding of natural odor mixture components in the human brain. Neuron 84, 857–869 (2014). [PubMed: 25453843]
24. Suzuki S, Cross L & O'Doherty JP Elucidating the underlying components of food valuation in the human orbitofrontal cortex. Nature neuroscience 20, 1780 (2017). [PubMed: 29184201]
25. Hare TA, Camerer CF & Rangel A Self-control in decision-making involves modulation of the vmpfc valuation system. Science 324, 646–648 (2009). [PubMed: 19407204]
26. Lim S-L, O'Doherty JP & Rangel A Stimulus value signals in ventromedial pfc reflect the integration of attribute value signals computed in fusiform gyrus and posterior superior temporal gyrus. Journal of Neuroscience 33, 8729–8741 (2013). [PubMed: 23678116]
27. Li C & Chen T Aesthetic visual quality assessment of paintings. IEEE Journal of selected topics in Signal Processing 3, 236–252 (2009).
28. Chatterjee A, Widick P, Sternschein R, Smith WB & Bromberger B The assessment of art attributes. Empirical Studies of the Arts 28, 207–222 (2010).
29. Vaidya AR, Sefranek M & Fellows LK Ventromedial frontal lobe damage alters how specific attributes are weighed in subjective valuation. Cerebral Cortex 1–11 (2017). [PubMed: 28365777]
30. Rother C, Kolmogorov V & Blake A Grabcut: Interactive foreground extraction using iterated graph cuts. In ACM transactions on graphics (TOG), vol. 23, 309–314 (ACM, 2004).
31. Walther D & Koch C Modeling attention to salient proto-objects. Neural networks 19, 1395–1407 (2006). [PubMed: 17098563]

32. Murray N, Marchesotti L & Perronnin F Ava: A large-scale database for aesthetic visual analysis. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, 2408–2415 (IEEE, 2012).
33. Simonyan K & Zisserman A Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
34. Deng J et al. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, 248–255 (Ieee, 2009).
35. Hong H, Yamins DL, Majaj NJ & DiCarlo JJ Explicit information for category-orthogonal object properties increases along the ventral stream. Nature neuroscience 19, 613 (2016). [PubMed: 26900926]
36. Mallon B, Redies C & Hayn-Leichsenring GU Beauty in abstract paintings: perceptual contrast and statistical properties. Frontiers in human neuroscience 8, 161 (2014). [PubMed: 24711791]
37. Graham D & Field D Statistical regularities of art images and natural scenes: Spectra, sparseness and nonlinearities. Spatial vision 21, 149–164 (2008).
38. Vessel EA, Maurer N, Denker AH & Starr GG Stronger shared taste for natural aesthetic domains than for artifacts of human culture. Cognition 179, 121–131 (2018). [PubMed: 29936343]
39. Vessel EA & Rubin N Beauty and the beholder: Highly individual taste for abstract, but not real-world images. Journal of vision 10, 18–18 (2010).
40. Bengio Y Deep learning of representations for unsupervised and transfer learning. In Proceedings of ICML workshop on unsupervised and transfer learning, 17–36 (2012).
41. Leshno M, Lin VY, Pinkus A & Schocken S Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. Neural networks 6, 861–867 (1993).
42. Hofmann T, Schölkopf B & Smola AJ Kernel methods in machine learning. The annals of statistics 1171–1220 (2008).
43. Rigotti M et al. The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585 (2013). [PubMed: 23685452]
44. Krizhevsky A, Sutskever I & Hinton GE Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, 1097–1105 (2012).
45. Esteva A et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542, 115 (2017). [PubMed: 28117445]
46. LeCun Y, Bengio Y & Hinton G Deep learning. nature 521, 436 (2015). [PubMed: 26017442]
47. Brieber D, Nadal M & Leder H In the white cube: Museum context enhances the valuation and memory of art. Acta psychologica 154, 36–42 (2015). [PubMed: 25481660]
48. Ke Y, Tang X & Jing F The design of high-level features for photo quality assessment. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, 419–426 (IEEE, 2006).
49. Salah MB, Mitiche A & Ayed IB Multiregion image segmentation by parametric kernel graph cuts. IEEE Transactions on Image Processing 20, 545–557 (2010). [PubMed: 20716502]
50. Nock R & Nielsen F Statistical region merging. IEEE Transactions on pattern analysis and machine intelligence 26, 1452–1458 (2004). [PubMed: 15521493]
51. Zhu Q, Yeh M-C, Cheng K-T & Avidan S Fast human detection using a cascade of histograms of oriented gradients. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, 1491–1498 (IEEE, 2006).
52. We thank Avi Vaidya and Lesley Fellows for this suggestion.
53. Hein M & Bühler T An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse pca. In Advances in Neural Information Processing Systems, 847–855 (2010).
54. Murray N & Gordo A A deep architecture for unified aesthetic prediction. arXiv preprint arXiv:1708.04890 (2017).
55. Huber PJ Robust estimation of a location parameter. Ann. Math. Statist. 35, 73–101 (1964).
56. Iigaya K, Yi S, Wahle IA, Tanwisuth K & O'Doherty JP Aesthetic preference for art emerges from a weighted integration over hierarchically structured visual features in the brain. bioRxiv (2020).

57. Gatys LA, Ecker AS & Bethge M Image style transfer using convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2414–2423 (2016).

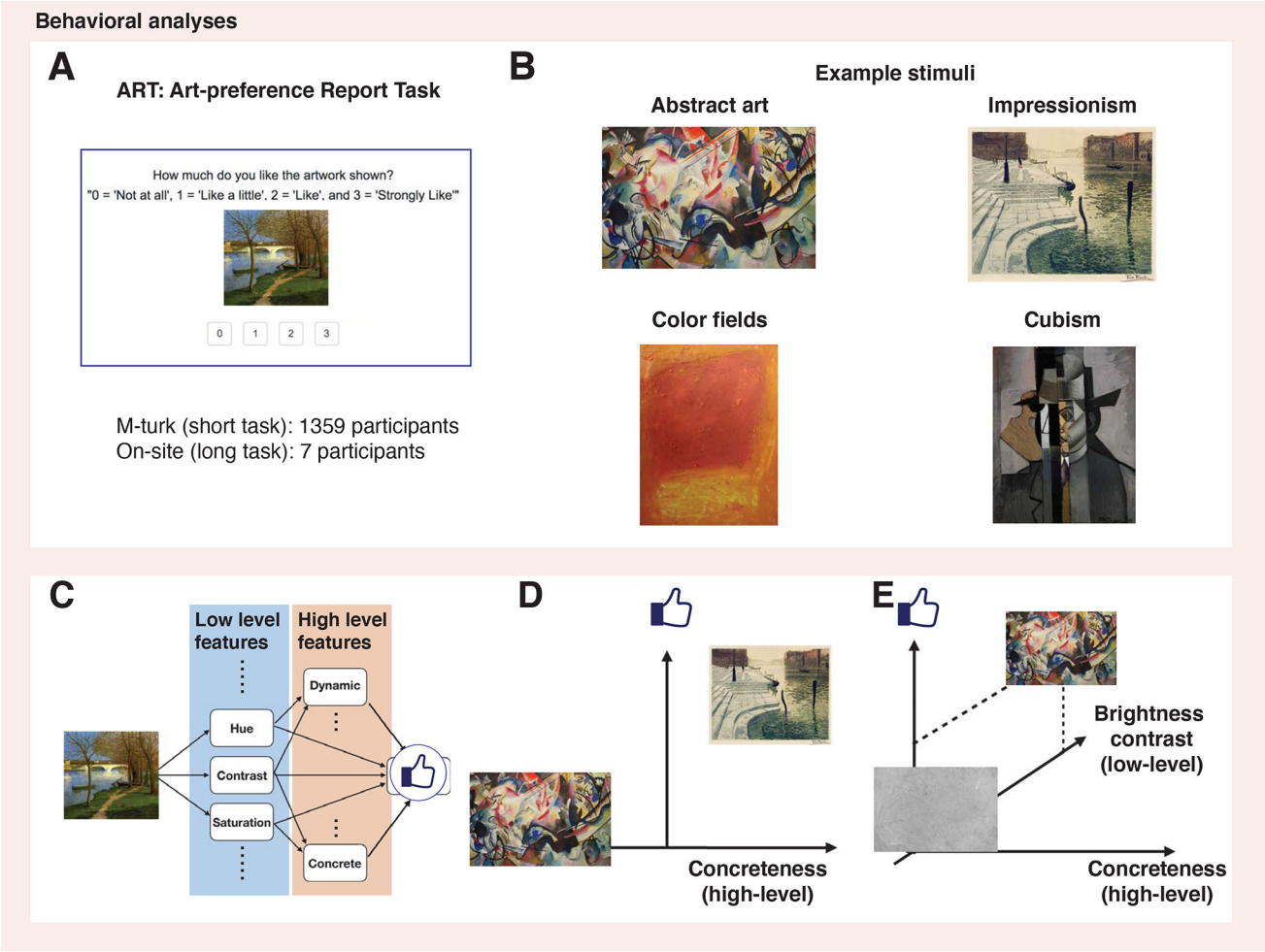

Figure 1:

Testing the linear feature summation (LFS) model that constructs aesthetic value of visual stimuli. (**A**). The task (ART: art-liking rating task). Participants were asked to report how much they like a stimulus (a piece of artwork) shown on the screen using a four-point Likert rating ranging from 0 to 3. (**B**). Example stimuli: Cubism, Impressionism, Abstract art and Color Fields, and supplemented with art stimuli previously used.[29] Each m-turk participant performed approximately 60 trials, while in-lab participants performed 1001 trials (one trial per image). (**C**). Schematic of the LFS model. A visual stimulus (e.g., artwork) is decomposed into various low-level visual features (e.g., mean hue, mean contrast), as well as high-level features (e.g., concreteness, dynamics). We hypothesized that high-level features are constructed from low-level features, and that subjective value is constructed from a linear combination of all low and high-level features. (**D**). How features can help construct subjective value. In this example, preference was separated by the concreteness feature. (**E**). In this example, the value over the concreteness axis was the same for four images; but another feature, in this case, the brightness contrast, could separate preferences over art. Due to copyrights, some paintings presented here are not identical to what we used in our studies. Credit: History and Art Collection, ART Collection, Aleksandra Konoplya, Alamy Stock Photo, RISD Museum.

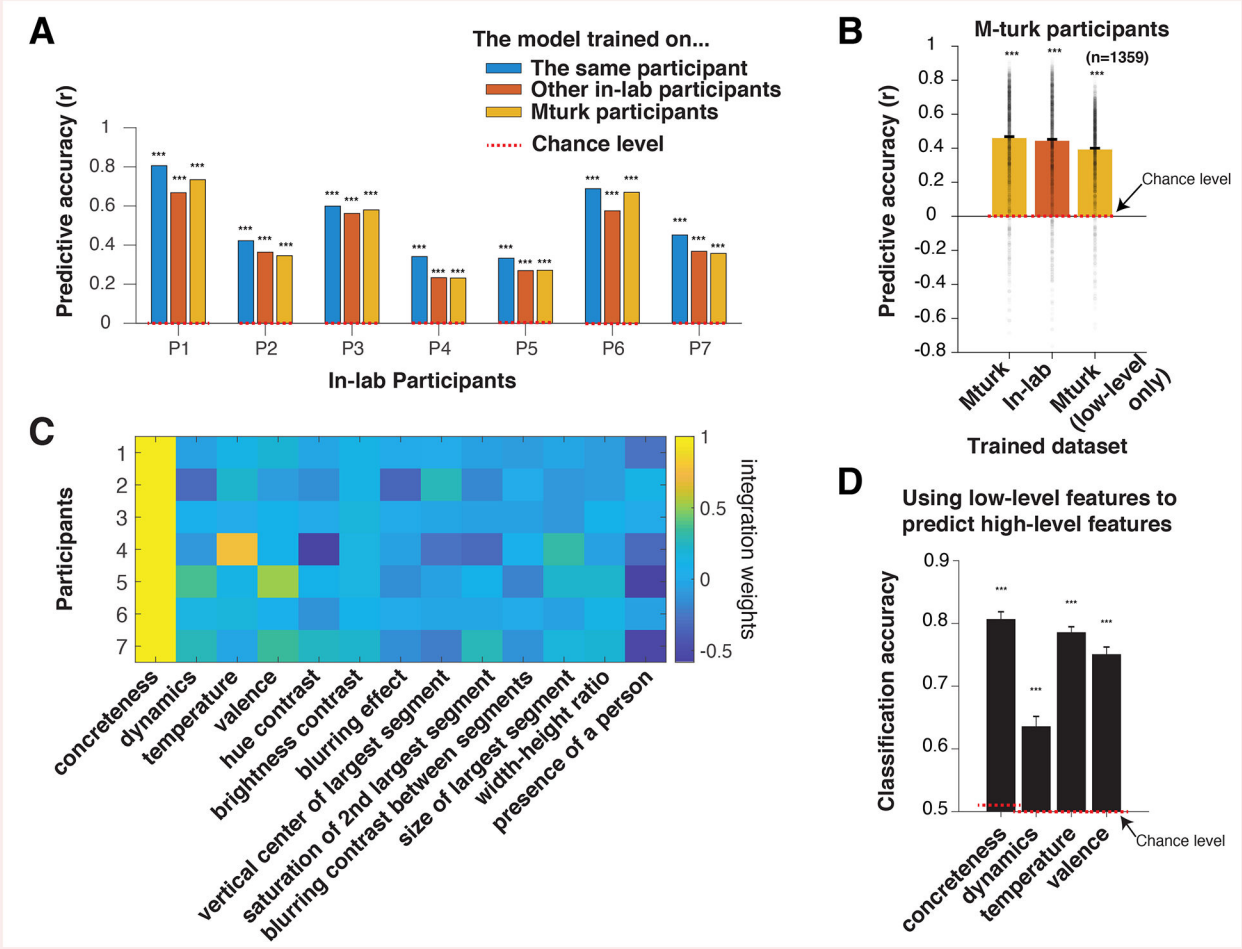

Figure 2:

The LFS model successfully predicts the subjective value of paintings. (**A**). The LFS model with shared features captured in-lab participants' art liking ratings. The predictive score, defined by the Pearson correlation coefficient between the model's out-of-sample prediction and actual ratings, was significantly greater than chance for all subjects who performed the task in the lab. The model was trained on six participants and tested on the remaining participant (blue), trained and tested on the same participant (red), and trained on on-line participants and tested on in-lab participants (yellow). In-lab subjects performed a long task with 1001 trials. Statistical significance was tested against a null distribution of correlation scores constructed by the same analyses with permuted image labels. The chance level (the mean of the null distribution) is indicated by the dotted lines (at 0). The same set of features (shown in **C**) was used throughout the analysis. (**B**). Our model also successfully accounted for the on-line participants' liking of the art stimuli. We trained the model on all-but-one participants and tested on the remaining participants (left). We also fit the model separately to in-lab participants and tested it independently on all on-line participants (middle). The model predicted liking ratings significantly in all cases, even when we used low-level attributes alone (right). Each on-line participant performed approximately 60 trials. The

error bars show the mean and the SE over participants, while the dots indicate individual participants. The chance level (the mean of the null distribution constructed in the same manner as F) is indicated by the dotted line. (**C**). Weights on shared features that were estimated for in-lab participants. We estimated weights by fitting individual participants separately. (**D**). The low-level features can predict the variance of high-level features. Classification accuracy (high or low values, split by medians) are shown. Note that though the prediction is highly significant, there is still a small amount of variance remaining that is unique to high-level features. The chance level (the mean of the null distribution) is indicated by the dotted line. The error bars indicate the standard errors over cross-validation partitions. In all panels, three stars indicate $p < 0.001$ against permutation tests.

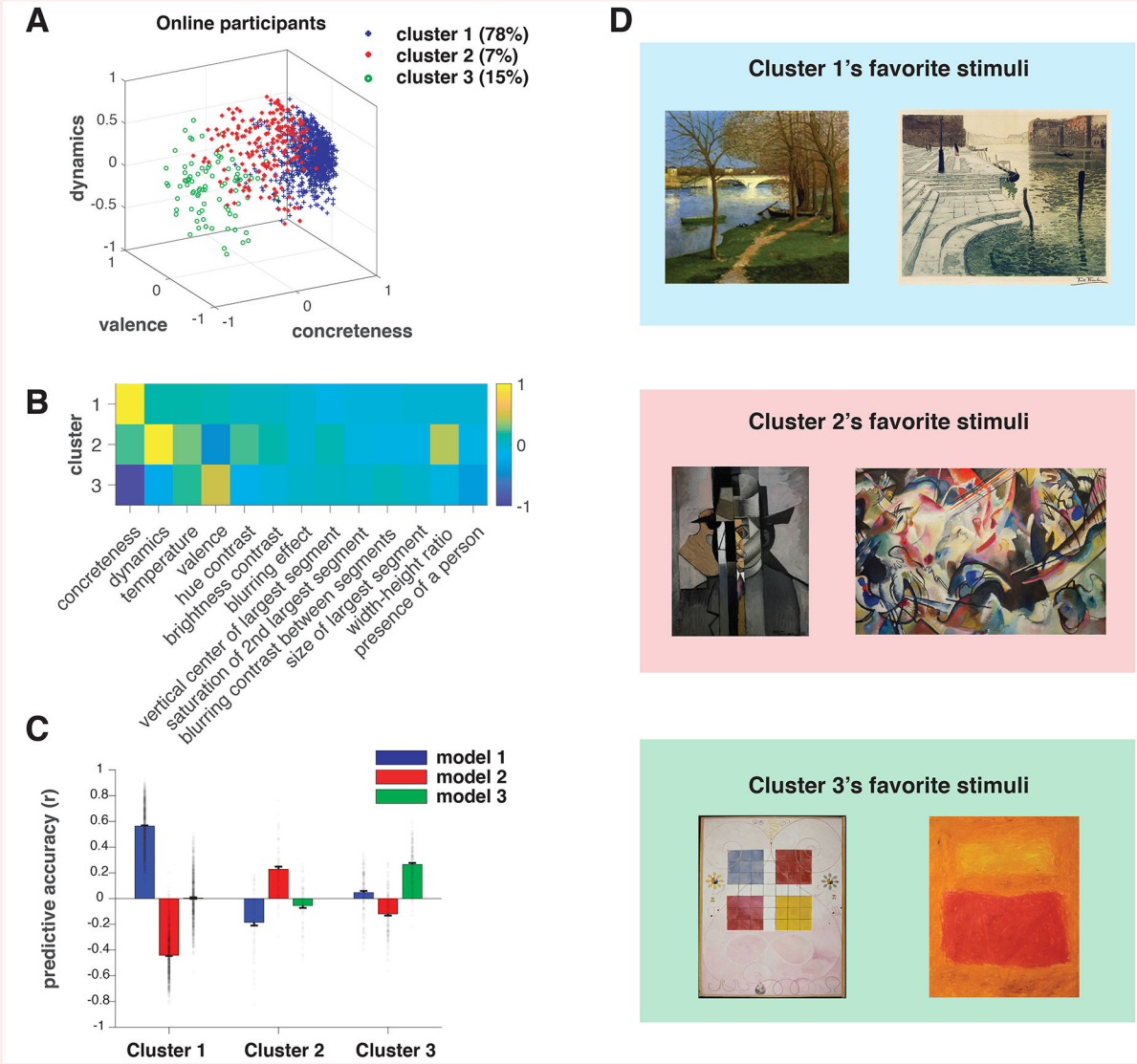

Figure 3:

Cluster analysis in the feature space suggests the existence of distinct groups of individuals who vary in their preference computations across our online sample. (**A**). The estimated feature weights of all participants, colored by cluster membership. We fit the LFS model to each individual online participant. We then performed a clustering analysis on the estimated weights using a Gaussian mixture model. The number of Gaussians was optimized by comparing Bayes Information Criterion (BIC) scores. Estimated weights from three features are shown for illustration. (**B**). The estimated feature weights at the center of each cluster. Cluster 1 assigns a large positive value to concreteness, while cluster 3 assigns a large negative value to concreteness. Cluster 2 has a distinctively large weight on the dynamic. (**C**). Predictive accuracy of participants in each cluster, using a model with the mean of each gaussian as its parameters. The result suggests that Cluster 1 and 2 have conflicting preferences, while cluster 3 is rather distinct. The error bars indicate the mean and SE

of participants, while the dots indicate individual participants. (**D**). Example stimuli that were preferred by each cluster of participants. The stimuli preferred by participants in cluster 1 include realistic landscape paintings, some of which are from impressionism. The stimuli preferred by cluster 2 include abstract, complex paintings e.g., in cubism. Cluster 3's favorite stimuli include simple paintings in color fields and abstract art. Art images are purchased from Alamy.com. Due to copyrights, color field paintings presented here are not identical to what we used in our studies. Credit: History and Art Collection, ART Collection, LatitudeStock, Volgi archive / Alamy Stock Photo, RISD Museum.

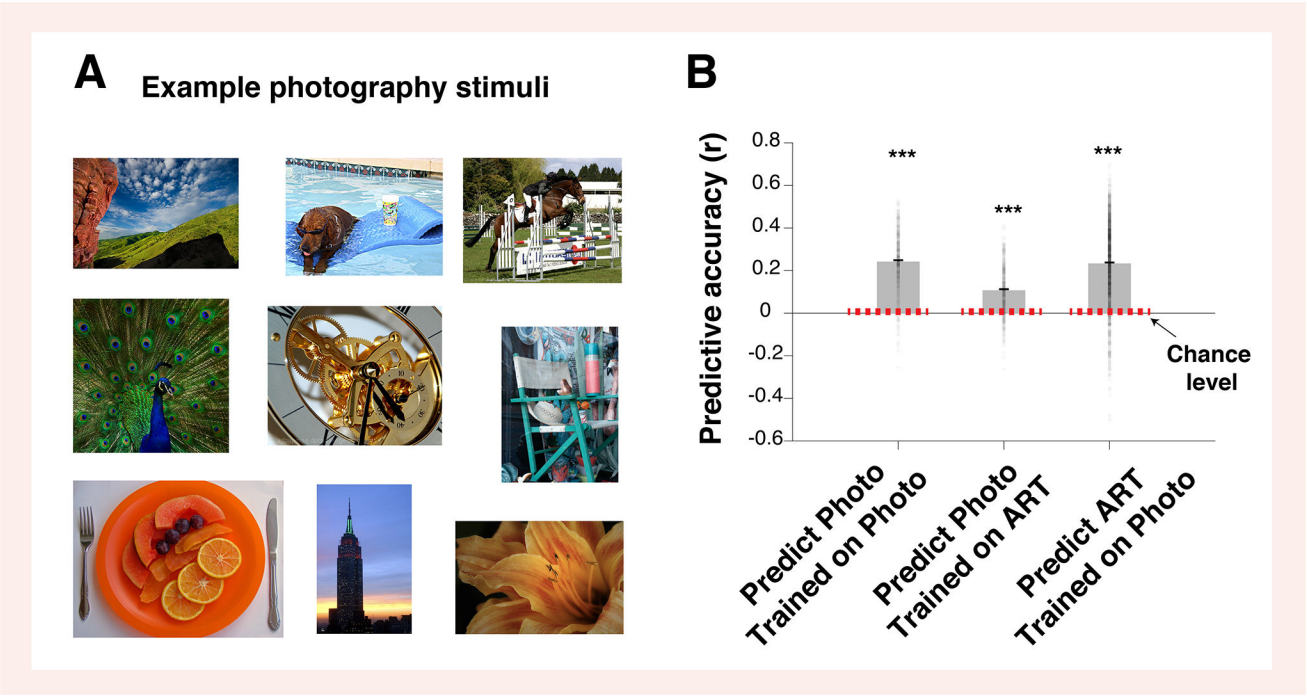

Figure 4:

The LFS model also predicts subjective liking ratings for various kinds of photographs. (**A**). Example stimuli from the photography dataset. We took a wide range of images from the online photography (AVA) dataset,[32] and ran a further on-line experiment in new M-Turk participants (n = 382) to obtain value ratings for these images. Images are from AVA dataset. (**B**). A linear model with low-level features alone captured liking ratings for photography. This model when trained on liking ratings for photography (data from the current experiment) also captured liking ratings for paintings (data from the previous experiment described in Figure 1), and the model trained on liking ratings for paintings could also liking ratings for photography. We note that in all cases the model was trained and tested on completely separate sets of participants. The significance was tested against the null distribution constructed from the analysis with permuted image labels. The error bars indicate the mean and the SE, while the dots indicate individual participants.

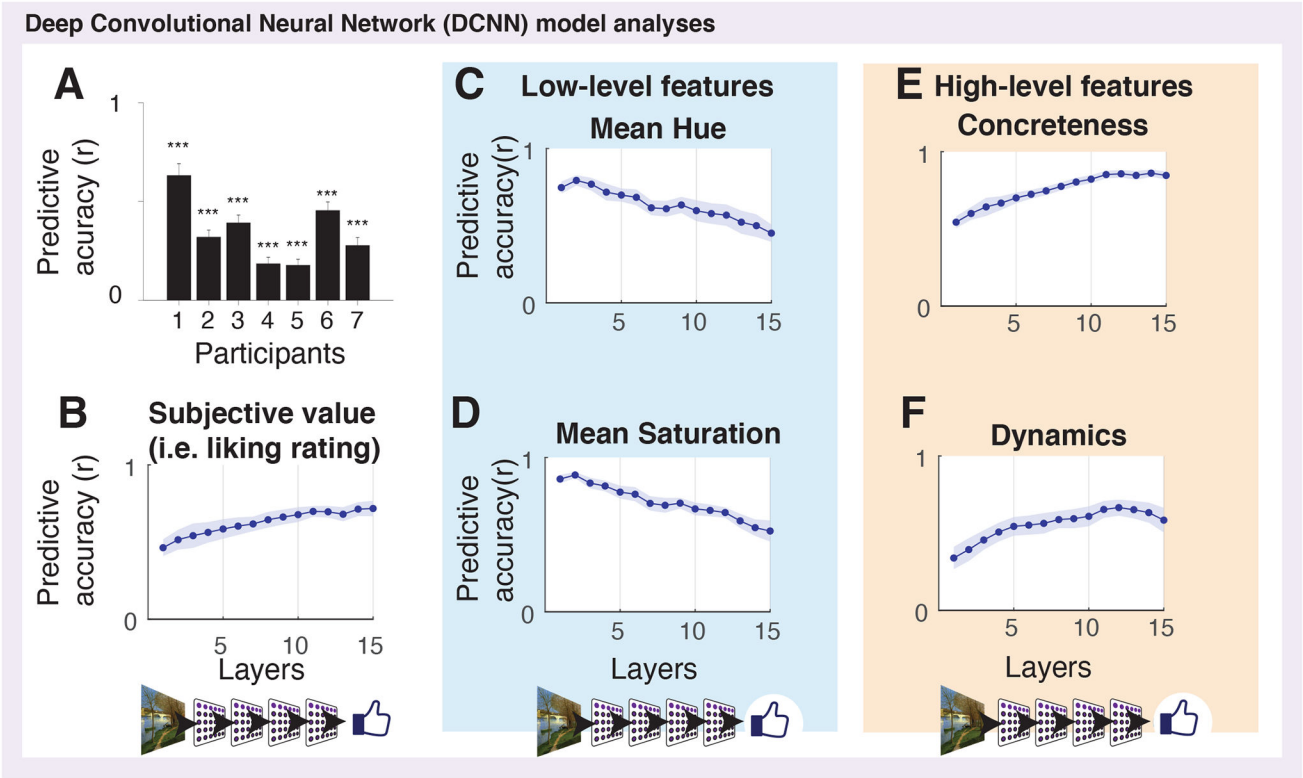

Figure 5:

A deep convolutional neural network (DCNN) can predict subjective values (i.e., liking ratings) of art stimuli, and the features that we introduced to our LFS model spontaneously emerge in the hidden layers of the network. We utilized a standard convolutional neural network (VGG 16[33]) that came pre-trained on object recognition with ImageNet,[34] consisting of 13 convolutional and three fully connected layers. We trained the last three fully connected layers of our network on average art liking scores without explicitly teaching the network about the LFS model's features. (**A**).The neural network could successfully predict human participants' liking ratings significantly greater than chance across all participants. The significance ($p < 0.001$, indicated by three stars) was tested by a permutation test. (**B**). We found that we can decode average liking ratings using activation patterns in each of the hidden layers. The predictive accuracy was defined by the Pearson correlation between (out-of-sample) model's predictions and the data. For this, we used a (ridge) linear regression to predict liking ratings from each hidden layer. We first reduced the dimensions of each layer with a PCA, taking top PCs that capture 80% of the variance in each layer. The accuracy gradually increases over layers despite the fact that most layers (layers 1–13) were not trained on liking ratings but on ImageNet classifications alone. (**C,D**). When performing the same analysis with the LFS model's features, we found some low-level visual features with significantly decreasing predictive accuracy over hidden layers (e.g., the mean hue and the mean saturation). We also found that a few computationally demanding low-level features showed the opposite trend (see the main text). (**E,F**). We found some high-level visual features with significantly increasing predictive accuracy over hidden layers (e.g., concreteness and dynamics). We also found that temperature, which we

introduced as a putative high-level feature, actually shows the opposite trend, likely because it is a color-based feature that can be straightforwardly computed from pixel data. Credit: History and Art Collection / Alamy Stock Photo.