Prototype preference is a principle of aesthetics for real-world scene perception

Yi-Chia Chen (0000-0002-8321-8595)[1*],
Shuhao Fu (0000-0003-2672-5125)[1],
Derek Feng (0000-0002-7860-0968)[2],
Moriah Taylor[1],
Jeff Chang[1],
Xiaoyang Chi[1],
Hongjing Lu (0000-0003-0660-1176)[1,3]

[1]Department of Psychology, University of California, Los Angeles
[2]Department of Statistics & Data Science, Yale University
[3]Department of Statistics, University of California, Los Angeles

*Corresponding author, yichiachen@g.ucla.edu

Running Head : Aesthetic preference for scene prototype

Declarations of interest : This project was funded by National Science Foundation BSC-1655300 awarded to HL.

Abstract

A salient aspect of our daily visual experiences is aesthetic impressions triggered by the scenes that we view. Why do we enjoy viewing some scenes but not others? Researchers have found salient visual features that we enjoy seeing for colors, shapes, and objects, and even for human faces and bodies; however, aesthetic preferences among realistic scenes—views that encompass many different objects and features—remain poorly understood. What principles govern these aesthetic preferences for complex visual scenes? We tackled this question by taking advantage of recent advances in computational models of human vision, and demonstrated that prototypicality—how typical a scene is among others—predicts aesthetic impressions. We modeled scene prototypicality by computing the similarity between a specific scene and a large reference set of other scenes using embeddings extracted by pretrained deep neural networks, AlexNet and VGG-16. For scenes consisting of inanimate content, the more prototypical the high-level visual representations are, the more aesthetically pleasing the scene appears. This aesthetic prototype effect indicates that aesthetic preferences among visual scenes are systematic, explainable, and reflect the underlying organization of visual scene representations. Our method demonstrates a new way to explore aesthetics—besides investigating specific visual properties, we can also explore the holistic role of visual processing for realistic scenes in creating aesthetic experiences, and integrate it into our understanding of everyday vision.

Keywords: aesthetics; prototype; deep learning; visual scene

Whether we are walking in our own neighborhoods or flying to the opposite side of the world, people invest a huge amount of time and resources in efforts to see new scenes. Besides expanding knowledge, a major purpose of such behaviors is to experience the beauty of what we see. Despite the massive popularity of "sightseeing", visual aesthetics has remained a marginalized topic in psychology. In the few studies exploring this area, scientists have mostly focused on specific visual features such as color (Martindale & Moore, 1988; Palmer & Schloss, 2010), curvature (Bar & Neta, 2006), complexity (Berlyne, 1970), object size (Chen et al., 2022; Konkle & Oliva, 2011; Linsen et al., 2011), and various kinds of orientation (Avrahami et al., 2004; Chen, et al., 2018; Latto et al., 2000; Mather, 2012; Palmer et al., 2008). These properties have been studied in the context of simple stimuli, such as shapes (Gartus & Leder, 2013; Silvia & Barona, 2009) or objects (Bar & Neta, 2006; Halberstadt & Rhodes, 2003), as well as abstract compositions (Locher et al., 2005). A central mystery remains: How does visual processing give rise to the beauty of *realistic visual experiences*, which encompass many objects and features?

The lack of studies addressing the role of the visual system in scene aesthetics has a straightforward explanation[1]: Vision scientists are only starting to understand scene perception itself (e.g., Bonner & Epstein, 2018; Son, Walther, & Mack, 2021). Just as it has not proved possible to explain holistic scene perception based solely on visual features selected or defined by researchers (Epstein & Baker, 2019), it had not been possible to combine featural aesthetic effects to predict the aesthetic experiences for scenes. However, recent breakthroughs in deep neural networks (DNN) have inspired new ways to discover relevant representations, and have led to models of how these representations potentially generate perception of scenes (Donahue et al., 2014; Zhou et al., 2014). Here we ask: Can we use DNN-extracted visual features to understand how we see the beauty of scenes?

Deep neural networks as models of human visual systems?

A variety of studies have evaluated the extent to which DNN-extracted features reflect how humans see (for recent reviews, see Kanwisher, et al., 2023; Lindsay, 2021; c.f., Bowers, et al., in press). Although DNNs have been found to overly rely on textures (Baker et al., 2018) and to make mistakes unlike human errors (Szegedy et al., 2013), they nonetheless provide useful tool for feature representations, evaluated based on their predictions of both human behaviors (Rajalingham et al., 2018) and brain activity (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014).

Using DNN representations to approximate properties of human visual systems affords several advantages. First, neuroimaging methods cannot easily distinguish brain activities for feedforward visual processing from recurrent processing that projects higher-

[1] "Scenes" in vision science sometimes refers to a narrower category of wide scale visual experience—sights of places. Here, we aimed to include all scales of visual experiences from the real world that often involved multiple objects and features.

level cognitive representations. It is therefore difficult to isolate processes underlying seeing from the influence of thinking. This distinction is critical for the current study because we aimed to understand the role of the visual system in aesthetics, and thus needed to exclude effects from higher-level cognition (e.g., thinking). Second, DNNs provide real-time, spatially precise, and noiseless responses to input images, escaping the trade-off between spatial and temporal precision that arises with neuroscientific methods. This high precision allows item-by-item comparisons between stimuli with various similar properties. Third, it is more economical to use DNNs than to collect data from human brains, allowing the use of large stimulus sets, as in this study. Accordingly, using DNNs as models of the human visual system provides unique information that is unavailable with any other methods.

The present study: The aesthetic prototype effect

The present study aimed to fill the gap between studies of specific featural preferences and the investigation of general principles for scene aesthetics, with the use of the rich information that DNNs extract from scenes. We focused on a widely observed phenomenon that connects multiple aspects of visual processing: the *prototype effect*, in which people show prioritizations for the central representation of a category. In the context of visual aesthetics, preferences for category prototypes have been observed for human faces (Langlois & Roggman, 1990; Ryali et al., 2020), biological organisms (Halberstadt & Rhodes, 2003; Younger, 1990), man-made objects (Landwehr, Labroo, & Herrmann, 2011; Whitfield & Slatter, 1979), abstract shapes (Solso & Raynis, 1979), dot patterns (Posner & Keele, 1968), and even in dynamic stimuli such as biological motion (Chen, Pollick, & Lu, under review). Critically, all past studies that demonstrated prototype preferences have involved single objects that vary along limited feature dimensions. Here, in five studies, we asked whether the prototype effect is an aesthetic principle that operates in complex visual experiences, in which millions of feature dimensions can be extracted from DNNs.

Traditional prototype effects are typically measured within a specific category of objects (e.g., chairs), often predefined by the experimenters. Given the difficulties in determining how diverse visual experiences are categorized in human minds, in the present study we took a data-driven approach. Thus, we did not limit the present study to scenes from a specific set of basic-level categories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Instead, we used prototype effects as indications of how scene representations are organized into categories in the human visual system. In other words, the discovery of a prototype effect suggests the existence of a visual category among the tested scenes. This approach does not require making assumptions about scene categorization, and allowed us to explore the prototype preferences in the broadest possible category—the full diversity of visual experiences, whether or not they belong to predefined sets of categories.

Study 1: An Aesthetic Prototype Effect for Inanimate Scenes

We measured the prototype effect using two sets of scenes: The prototypicality of scenes in the target set were estimated by comparing their DNN features to each scene in a separate population set. A large sample of human observers rated aesthetic impressions for images in the target set to assess the effect of prototypicality on aesthetic values (see Figure 1 for an overview of our method). To confirm that our results were based on meaningful visual representations, we also employed a control model which maintained the same architecture as the pretrained DNN model, but the model parameters were randomly permuted.

Method

Target image set—Keyword Aesthetics dataset. An image set—named Keyword Aesthetic dataset (KAD)—consisting of 78 inanimate and 78 social images (Figure 1a) was constructed following these steps: (a) 700 words were randomly selected from the top 40,000 frequent words based on the British National Corpus word frequency database (Leech & Rayson, 2014). (b) These words were used as Google Image search keywords with the search tool option of size set to "large", and the top 100 results for each keyword were examined. (c) The first image passing these criteria (Chen et al., 2022) was selected: Related to the keyword, easily interpretable as a real photograph without visible alterations, larger than 550 x 550 px, has at least one distinct object (excluding uniform textures), with content that is not obviously emotional, and does not include any symbols (e.g., a brand mark or dollar sign) or text. (d) For the inanimate images, the image must not contain any part of realistic or cartoon depictions of people, body parts, or animals. For the social images, the image must contain at least one human eye visible enough to tell if it was opened or closed. (e) Images were retained only if the corresponding keywords resulted in the selection of both inanimate and social images. This procedure yielded 78 inanimate images and 78 social images from 78 keywords. The images were then resized to their respective smallest sizes that were still larger than 550 x 550 px and cropped to retain a random 500 x 500 px region. This random cropping was used to diversify the framing in the image set, since photographs online were often selectively framed by people in ways that may reflect certain biases. Only images that still met the selection criteria in (d) and (e) after cropping were included. The images were then scaled down to the size of 224 x 224 px, which was the input image size the models required. All images used in this study are publicly available in the OSF repository here: https://osf.io/mqhxg/?view_only=e8e6d8435f2749558ea1fde16bd4c951 .

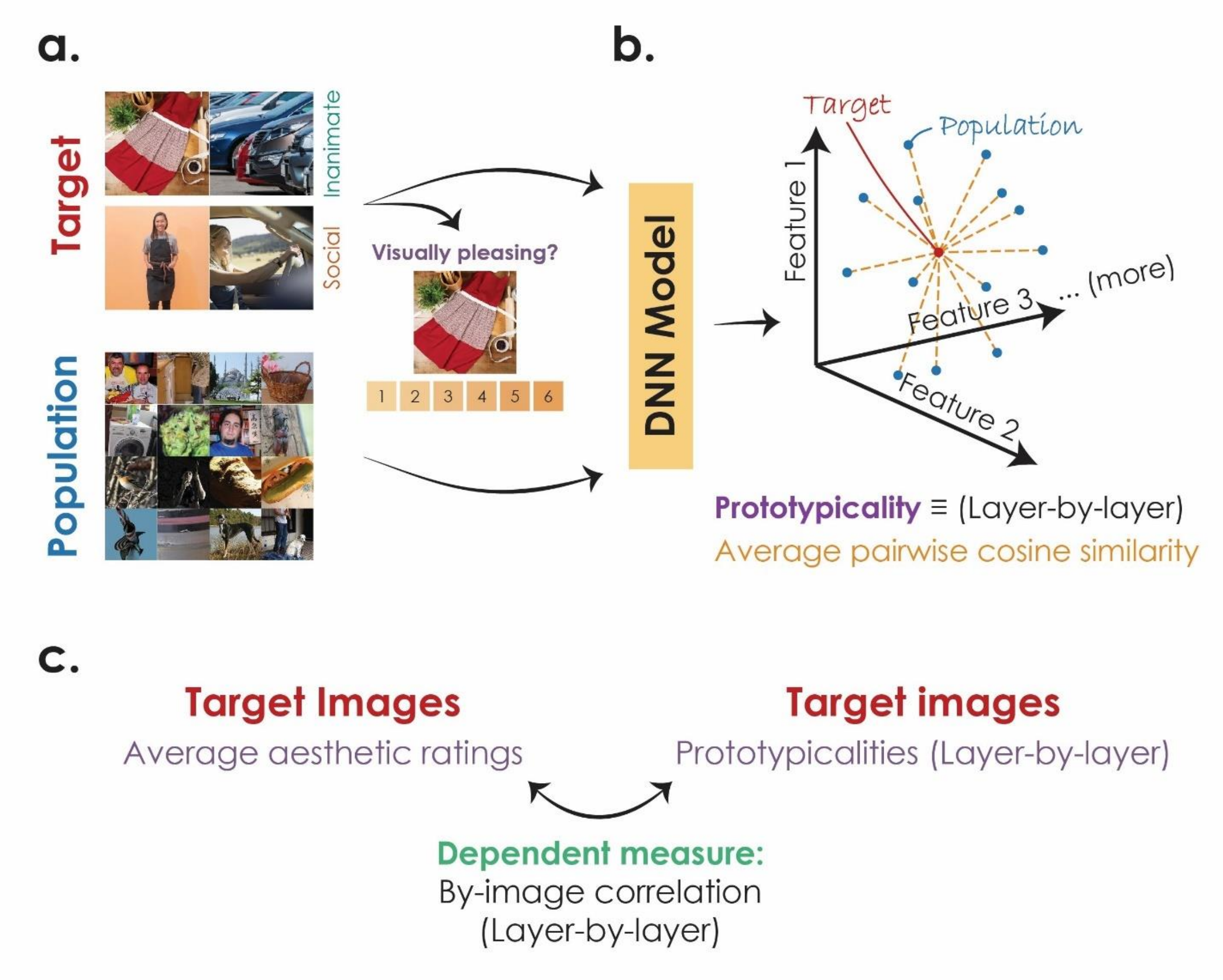

Figure 1. An overview of the procedure is depicted here: (a) Two sets of scenes—the target and population image set—were used to measure prototype preference. A hundred human observers rated the scenes in the target set on their aesthetic values. (b) DNN features for scenes in both image sets were extracted. We estimated the prototypicality of each scene in the target set by calculating its average cosine similarity with each of the scene in the population set. The prototypicalities were estimated separately for each layer of the DNN models. (c) We then calculated the by-image Spearman's rank correlation between the target scenes' average aesthetic ratings by humans and estimated prototypicalities by model simulations.

Population image set—ImageNet. A subset of 1,000 images from the ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015) was used as the population set (Figure 1a). This subset was formed by randomly selecting one image per class from the 1,000 classes in the validation set and replacing images that did not appear to be unaltered realistic photographs by the next eligible images (sorted by file names).

Thus, they included both inanimate and social scenes. These images were then resized to 256 x 256 px and cropped to retain a random 224 x 224 px region.

Observers. A convenience sample of 100 naïve undergraduate students from the University of California, Los Angeles (UCLA) community (81 females, 17 males, 1 other gender, 1 undisclosed, M_{age}=20.6, SD_{age}=2.5, $range_{age}$=[18, 31]; all with normal or corrected-to-normal vision) completed a 30-minute within-subjects online experiment in exchange for course credit. An additional 30 observers participated but were removed based on predetermined criteria (see details in the Observer exclusions section below). A large sample size was used, as effect size estimation was technically difficult due to the novelty and complexity of the analyses. The reliability of the effects was evaluated carefully by internal replications across different observer samples, image sets, and models. The study was approved by the UCLA Institutional Review Board.

Experiment procedure. Observers were directed to a website where stimulus presentation and data collection were controlled via custom software written in HTML, CSS, JavaScript, JQuery, and PHP. Observers were not allowed to participate using phones or tablets. After completing a CAPTCHA task (using the hCaptcha service: https://www.hcaptcha.com/), they were asked to maximize their window size, informed about their task, quizzed about their understanding of the instructions, and provided their consent.

During each trial, observers viewed each image from the preprocessed target set one-by-one and were asked to rate "how visually pleasing you find each image to be", and "In other words, how good/beautiful do you think the image looks". They rated each image on a 6-point Likert scale with labels (certainly pleasing, probably pleasing, guess pleasing, guess not pleasing, probably not pleasing, and certainly not pleasing). We chose to use multiple descriptions of aesthetic judgements that all pointed to the idea of beauty to prevent participants from overanalyzing a particular term. Research has also supported the use of "pleasingness" and "beautiful" scales to obtain judgments from the single dominant dimension of aesthetics (Augustin et al., 2012; Jacobsen et al., 2004; Russel & George, 1990).

Before the formal experiment began, the observers first practiced using the rating scale on one practice image. They then rated the 78 inanimate images and 78 social images (preprocessed with above procedures), mixed in a different random order for each observer. A random selection of 35 inanimate images and 35 social images (different for each observer) was then repeated a second time. The ratings from these repeats were only used to measure test-retest reliability for the purpose of observer exclusion, and were otherwise discarded in the main analyses. Observers were given a self-paced break halfway through the rating task. At the end of the experiment, observers answered a series of debriefing questions to ensure they had completed the experiment without any issues.

Additional questionnaires were administered for the purposes of other studies and were not analyzed for this study.

Observer exclusions. Thirty observers were excluded based on criteria decided before data collection began, with some observers triggering more than one criterion: six observers did not follow the instructions; three observers reported that they did not take the experiment seriously; one observer spent less than 0.5 second on at least one page of the instructions; three observers had a browser viewport smaller than 800px × 600px; four observers had at least one trial with the image not fully in view during the rating task; seven observers hid the experiment browser tab more than three times during the trials; five observers gave the same rating to more than 15 consecutive trials; one observer took longer than 120 seconds or shorter than 0.3 seconds to respond for more than four trials in either condition; ten observers had test-retest reliabilities lower than 0.5 in either condition; and four observers took too long to complete the experiment (two SDs longer from the mean duration of all observers before exclusions). Detailed documentation of all exclusions is publicly available in the OSF repository here: https://osf.io/mqhxg/?view_only=e8e6d8435f2749558ea1fde16bd4c951

The pretrained visual feature model. We chose deep neural network (DNN) models—one of the model classes that show strong recognition performance and resemblance to human visual system—to extract important visual features in scenes that people use to perform daily tasks. AlexNet (Krizhevsky, 2014 ; Krizhevsky et al., 2017) was used due to its popularity, simplicity relative to other models, robust performance in various tasks (e.g., Donahue et al., 2014), and multifaceted similarity to human visual systems (Schrimpf et al., 2020). The model was implemented in PyTorch 1.12.1 with pretrained weights fixed from training on the ImageNet object classification task. Visual features for each image in both target and population image sets were the embeddings in each layer, obtained by forward passes through the models. Each image resulted in 13 layers (including Convolutional, ReLU, and Max-pooling layers) of visual features represented in matrices of different sizes. (We did not include the classifier layers, i.e., the Dropout and Fully-connected layers, because the control model was not discriminative across different images at those layers.) The matrix from each layer for each image was then flattened to a vector for prototypicality estimation detailed in a later section. For example, on Layer Conv-5, the embeddings of an image were flattened from a 256 x 13 x 13 matrix to a vector of size 43264.

The control model. The control model was identical to the pretrained AlexNet, except that its parameters were permuted. The permutation was done separately for each layer, as well as for the weights and biases. Since the permutation process was random, we ran 100 different iterations of the control model and averaged the analysis results. The procedure for visual feature extractions was identical to that performed on the pretrained model reported above.

Prototypicality estimation. To measure the prototypicality of each image (Figure 1b), we separately analyzed the embedding features in each layer of the DNN models. For each image in the target set, we measured the pairwise cosine similarities between its embedding vector and the embedding vector for each image in the population set. Then, we averaged the target image's similarities to all population images. The higher the averaged value is, the more similar the target image is to other scenes in the population image set. This measure is thus an index of prototypicality. For each image in the target image set, we obtained a prototypicality score from each layer of the pretrained model, and 100 prototypicality scores from each layer of the 100 iterations of the control model.

Relation between prototypicality and aesthetic ratings. With the estimated prototypicality index, we tested its effect on aesthetic impressions by assessing Spearman's rank correlation (Figure 1c). Again, this analysis was done separately for each layer. Importantly, we also analyzed inanimate and social images separately, as specialized perceptual processes are likely involved with social images with animated agents (Caramazza & Shelton, 1998). The rating z-scores were calculated within each subject and each image type separately. We then correlated image prototypicalities with each observer's aesthetic rating z-scores. For the control model, we averaged the estimated prototypicality indices from 100 iterations before conducting the correlation. We then analyzed the correlation coefficients using a three-way repeated-measure ANOVA (including 2 weights (pretrained/control) x 13 layers x 2 image types (inanimate/social)). Following the results from the ANOVA, planned paired-sample t-tests comparing the pretrained and the control models indicated whether we found a reliable prototype effect in each condition.

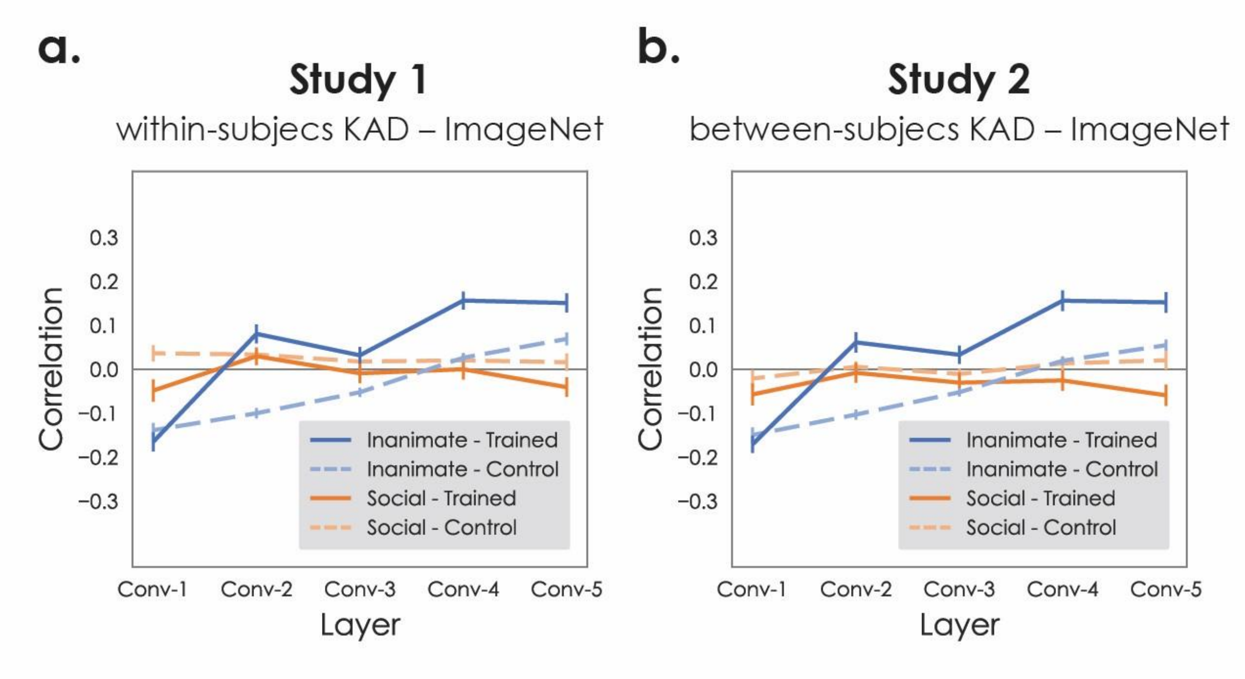

Figure 2. For (a) Study 1(within-subjects design) and (b) Study 2 (between-subjects design), correlations between aesthetic z-scores and image prototypicalities from each convolutional layer of AlexNet are plotted separately for four conditions. All error bars represent between-subjects 95% confidence intervals.

Results and discussion

The average correlation coefficients for the convolutional layers are plotted in Figure 2a. The results from the ReLU and Max-pooling layers were similar to the corresponding convolutional layers and thus were omitted from all figures and reporting for clarity and conciseness. Detailed results from ReLU and Max-pooling layers are publicly available in the OSF repository (https://osf.io/mqhxg/?view_only=e8e6d8435f2749558ea1fde16bd4c951). From inspecting the figures, three observations emerged: (a) There was a difference between the pretrained and the control models. (b) For the pretrained model, the results differed between the inanimate and social images. (c) The inanimate images showed positive correlations only in the layers above Conv-2 of pretrained AlexNet. These observations were confirmed by the repeated-measure ANOVA: The weights x layers x image types three-way interaction was significant (F(12, 1188)=22.36, p<.001, η_p^2=.184), along with all the two-way interactions (Fs>70.00, ps<.001, η_p^2s>.400).

Critically, for the inanimate images, the correlations between the ratings and the image prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (ts>2.80, ps<.006, ds>0.28) except for Conv-1 (t=1.86, p=.066, d=0.19). For the social images, the correlations were similar across the pretrained and the control models, not significantly different in Conv-2 and Conv-4 (ts<1.57, ps>.119, *d*s<0.16), and significantly more negative with the pretrained model in the rest of the layers (*t*s>3.05, *p*s<.003, *d*s>0.30). Thus, the more prototypical an inanimate scene is based on representations extracted in layers later than Conv-2, the more aesthetically pleasing it appears. This aesthetic prototype effect was not observed for low-level visual representations in early convolution layers nor for scenes containing social information.

Study 2: Between-subjects Replication

Although the differences found between inanimate and social scenes in study 1 are intriguing, it is nevertheless possible that they were caused by task demands from rating both inanimate and social scenes, such as applying different strategies upon noticing these two categories. In this study, we ruled out this possibility with a between-subjects replication experiment.

Method

The method of Study 2 was identical to that of Study 1 except as noted below.

Observers. Another sample of 200 naïve observers of the same nature as Study 1 (165 females, 33 males, 2 other genders, M_{age}=20.6, SD_{age}=2.7, $range_{age}$=[18, 48]) completed a 20-minute between-subjects experiment. An additional 47 observers participated but were removed based on predetermined criteria (see details in the Observer exclusions section below). The sample size was chosen so that each of the two conditions has a sample size that was identical to that used in Study 1.

Experiment procedure. All procedures were identical to Study 1 except that the observers were randomly assigned to the inanimate or social image condition, and only rated the 78 images and 35 repeats in their respective conditions.

Observer exclusions. Forty-seven observers were excluded based on criteria decided before data collection began, with some observers triggering more than one criterion: six observers did not follow the instructions; eighteen observers reported that they did not take the experiment seriously; seven observers had a browser viewport smaller than 800px × 600px; nine observers had at least one trial with the image not fully in view during the rating task; five observers hid the experiment browser tab more than three times during the trials; six observers gave the same rating to more than 15 consecutive trials; eight observers took longer than 120 seconds or shorter than 0.3 second to respond for more than four trials; nine observers had test-retest reliabilities lower than 0.5; and five observers took too long to complete the experiment (two SDs longer from the mean duration from all observers before exclusions).

Analysis. Instead of a repeated-measure ANOVA, a mixed-measure ANOVA was conducted with the image type as a between-subjects factor.

Results and discussion

The results are plotted in Figure 2b in the same format as for Study 1. From inspecting the figures, the same three observations from Study 1 were apparent. (a) There was again a difference between the pretrained and the control models. (b) For the pretrained model, the results differed between the inanimate and social images. (c) The inanimate images showed positive correlations only in the layers above Conv-2 of pretrained AlexNet. These observations were confirmed by the mixed-measure ANOVA. The weights x layers x image types three-way interaction was significant ($F(12, 2376)=36.70$, $p<.001$, $\eta_p^2=.156$), along with all the two-way interactions (*F*s>40.00, *p*s<.001, η_p^2s>.170). Critically, for the inanimate images, the correlations between the ratings and the image prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (*t*s>4.00, *p*s<.001, *d*s>0.40) except for Conv-1 ($t=1.45$, $p=.150$, $d=0.15$). For the social images, the correlations were similar across the models, not significantly different in Conv-2 ($t=1.20$, $p=.233$, $d=0.12$), and significantly more negative with the pretrained model in the rest of the layers (*t*s>2.03, *p*s<.045, *d*s>0.20). Thus, after eliminating any possible demand characteristics caused by rating both inanimate and social images, we replicated the findings from Study 1: The more prototypical an inanimate scene is based on visual representations extracted in layers later than Conv-2, the more aesthetically pleasing it appears. This aesthetic prototype effect was not observed for low-level visual presentations nor for scenes containing social information.

Study 3: Generalization to An Alternative Image Population

How general is the aesthetic prototype effect for inanimate scenes? In this study, we replicated the effect once again with a more diverse population image set to assess whether the prototype preference operates over diverse images.

Method

Using both the within- and between-subjects experiment data in Study 1 and 2, the same analyses were replicated with a new population image set— Aesthetic Visual Analysis (AVA; Murray et al., 2012). A subset of 995 images was used. We chose AVA because it is a large and diverse dataset that spans a wide range of aesthetic values from various thematic photography challenges. We selected a subset of images that maximized the range of aesthetic values following these steps: (a) All images from grayscale challenges or that were grayscale were excluded. (b) Remaining images were binned into 8 bins (1-2, 2-3, …, 8-9) based on average aesthetic votes (on a scale from 1-10). (c) The images in the two extreme bins (1-2, and 8-9) were all selected (rather than randomly sampled) due to the small size (3 images and 22 images respectively). (d) 162 images were randomly sampled from each of the remaining bins to achieve a subset size of around 1,000 images. In the end, 995 images were chosen, resized to 256 x 256 px, and cropped to retain a random 224 x 224 px region.

Results and discussion

The results are plotted in Figure 3a and 3b using the same human data collected in Study 1 and 2, but recomputed the prototypicality index using the population image set from AVA rather than ImageNet. Again, very similar patterns emerged in both within- and between-subjects experiments. For inanimate images, positive correlations between aesthetic ratings and prototypicalities were observed for the layers above Conv-2 of the pretrained AlexNet. This pattern did not emerge for social images or with the control model in which model parameters were randomly permuted. These observations were confirmed by the ANOVAs: For the within-subjects experiment, the three-way interaction of the weights x layers x image types was significant ($F(12, 1188)=85.16$, $p<.001$, $\eta_p^2=.462$), along with all the two-way interactions (Fs>60.00, ps<.001, η_p^2s>.400). Critically, for the inanimate images, the correlations between the ratings and the image prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (ts>8.40, ps<.001, ds>0.84) except for Conv-1, where the correlation was more negative with the pretrained model ($t=9.20$, $p<.001$, $d=0.92$). For the social images, the correlations were significantly more *negative* with the pretrained model than in the control model in all layers (ts>2.31, ps<.023, ds>0.23).

For the between-subjects experiment, the three-way interaction of weights x layers x image types was also significant ($F(12, 2376)=102.81$, $p<.001$, $\eta_p^2=.342$), along with all the two-way interactions (Fs>50.00, ps<.001, η_p^2s>.200). Critically, for the inanimate images, the correlations between the ratings and the image prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (ts>8.65, ps<.001, ds>0.86) except for Conv-1, where the correlation was more negative with the pretrained model ($t=8.39$, $p<.001$, $d=0.84$). For the social images, the correlations were similar across the models, not significantly different in Conv-3 ($t=1.66$, $p=.100$, $d=0.17$), and significantly more negative with the pretrained model in the rest of the layers (ts>2.16, ps<.033, ds>0.21.

The results from both Study 1 and Study 2 were thus replicated in Study 3, and generalized using AVA as the population image set. These findings demonstrate the robustness of the prototype preference for inanimate images.

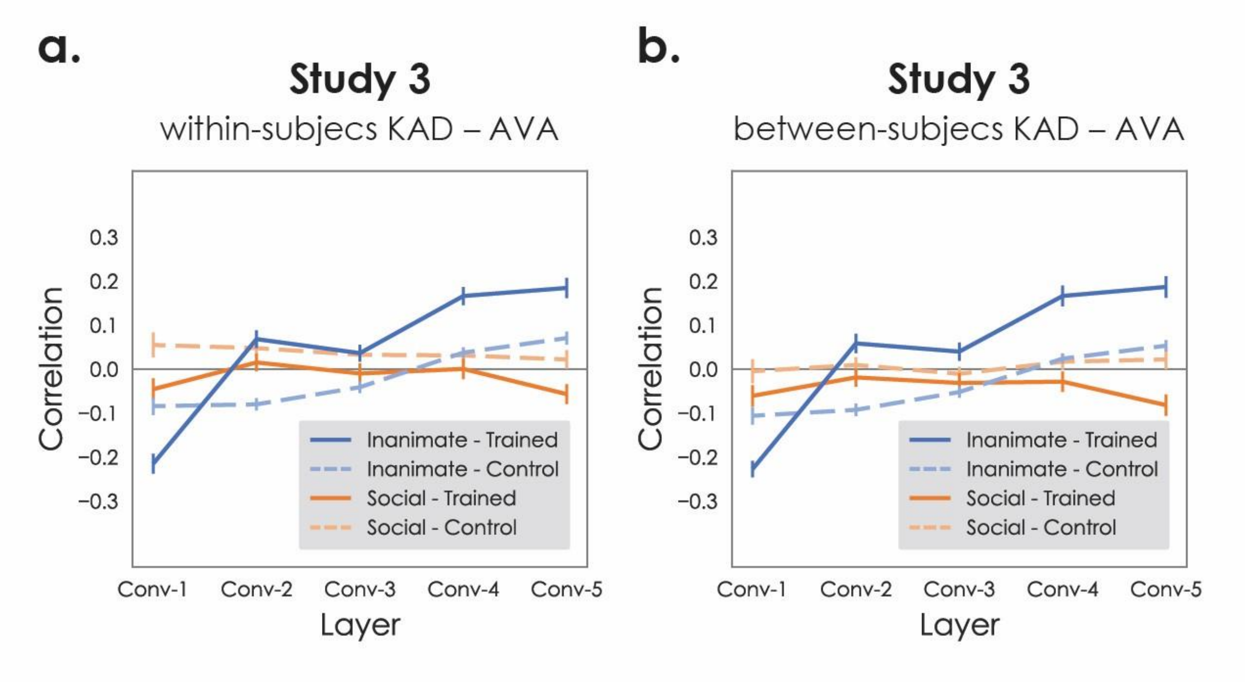

Figure 3. In the same manner as in Figure 2, correlations between aesthetics and prototypicalities are plotted for Study 3, where AVA served as the population image set. Aesthetic ratings from the within-subjects experiment are used in (a), and those from the between-subjects experiment are used in (b).

Study 4: Generalization to a Diverse Target Image Set

Can the prototype preference be observed in a large variety of inanimate scenes? In the previous three studies, we used a target image set that was specifically collected to contain no emotional images. Given the apparent link of emotions to aesthetics, it is possible that the prototype effect can only explain aesthetic experience when effects of emotions are removed. In this study, we sought to rule out this possibility by generalizing the prototype effect to a new target set that contained dramatic emotional content.

Method

All analyses conducted in Studies 1 through 3, involving two population image sets, were replicated with a new target image set—Open Affective Standardized Image Set (OASIS; Kurdi et al., 2017), with previously collected aesthetic ratings (Brielmann & Pelli, 2019).

Target image set—OASIS. A subset of 84 inanimate images from OASIS was used as the target image set. We chose OASIS for two reasons. First, aesthetic ratings were available (Brielmann & Pelli, 2019). Second, the images were specifically selected to be emotional, a criterion opposite to that of the KAD we tested in Study 1. We examined the subset of 225 images that were rated by the largest sample from Brielmann and Pelli (2019), and used the 84 images that contained only inanimate content, according to the same criteria used for the inanimate KAD. All images were resized from their original size of 500 x 400 pixel (px) to 280 x 224 px, and were either cropped from the center to 224 x 224 px, or scaled horizontally (with distortions) to 224 x 224 px, forming the cropped and scaled OASIS datasets respectively. Since we were using the aesthetic ratings collected by Brielmann & Pelli (2019) (see the next subsection for details), we chose to crop the images from the center regions (rather than random regions as in above experiments) based on the intuitive assumption that the aesthetic ratings were more likely based on the content in the center.

Aesthetic ratings. Aesthetic ratings of OASIS were collected in a past study (Brielmann & Pelli, 2019) from a group of diverse observers tested on the online crowdsourcing platform Amazon Mechanical-Turk. We used a subset of data by only including observers that were clinically normal and who rated the subset of images with the largest sample size. To maintain a consistent criteria of data quality, we excluded an additional four observers who gave the same rating to more than 15 consecutive trials (the same criterion adopted in Studies 1 and 2). This procedure led to the inclusion of ratings from a total of 326 observers. We then converted the ratings to z-scores within each subject, based on only the subset of images used in our analyses.

Results and discussion

The results are plotted in Figure 4a, 4b, 4c, and 4d for four conditions in which images in the OASIS target set were transformed via cropping or scaling, and the population images set were ImageNet or AVA dataset. From inspecting the figures, a clear pattern emerged. There was a difference between the pretrained and the control models for all conditions, with the pretrained model showing more positive correlations. These observations were confirmed by four separate repeated-measure ANOVAs. For both cropped and scaled OASIS target images with ImageNet as the population set, the two-way interactions of weights x layers were significant (*Fs*>140.00, *p*s<.001, $\eta_p{}^2$s>.300). The correlations between the ratings and the image prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (*t*s>16.10, *p*s<.001, *d*s>0.89). The same target images with AVA as the population set led to similar results, where the weights x layers two-way interactions were once again significant (*Fs*>160.00, *p*s<.001, $\eta_p{}^2$s>.300), and the correlations were significantly more positive with the pretrained model than with the control model in all layers (*t*s>9.95, *p*s<.001, *d*s>0.55). Thus, using a different set of images containing emotional content, the aesthetic prototype effect for inanimate scenes was once again replicated, demonstrating its generality.

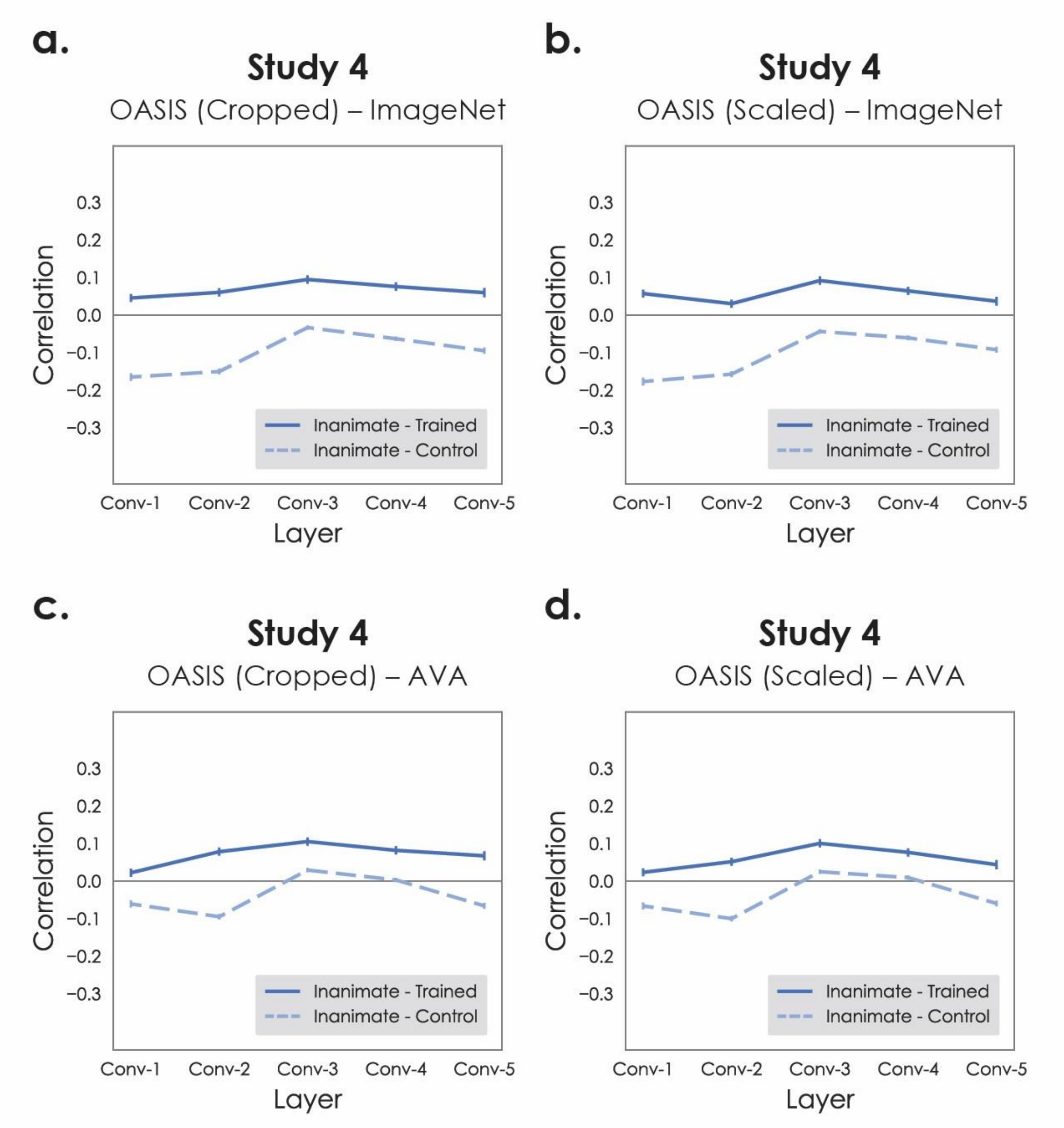

Figure 4. In the same manner as in Figure 2, correlations between aesthetics and prototypicalities are plotted for Study 4, where OASIS was used as the target image set. The population image set was ImageNet for (a) and (b), and was AVA for (c) and (d). Cropped OASIS images were used for (a) and (c), and scaled OASIS images were used for (b) and (d).

Study 5: Generalization to An Alternative DNN Model

AlexNet was used in the above four studies because of its popularity and robustness; however, an array of newer DNN models are available. We next generalized the findings using VGG-16 (Simonyan & Zisserman, 2014), chosen for its higher similarity to human behaviors and brain activities (Schrimpf et al., 2020) compared to AlexNet. We did not choose other DNN models, such as VGG-19 (Simonyan & Zisserman, 2014) or ResNet-50 (He et al., 2016), because their heavy computational loads would have been beyond our capacity at the time of this study.

Method

We conducted the same analyses using a larger DNN model, VGG-16. Given the large number of layers included in VGG-16, we used the latest sublayer for the five convolution layers (i.e., Conv-1.2, Conv-2.2, Conv-3.3, Conv-4.3, and Conv-5.3). We also focused on replication analyses on two target sets, using only AVA as the population set: The inanimate KAD images from Study 2 and the cropped OASIS images from Study 4.

Results and discussion

The results are plotted in Figure 5a and 5b. From inspecting the figures, a clear pattern emerged. There was a difference between the pretrained and the control models, with the pretrained model showing more positive correlations in later layers. These observations were confirmed by two separate repeated-measure ANOVAs. For the inanimate KAD images, the two-way interactions of weights x layers were significant (F(4, 99)=172.12, ps<.001, η_p^2=.635), and the correlations were significantly more positive with the pretrained model than with the control model in all tested layers (ts>2.80, ps<.001, ds>0.25) but Conv-1.2, where the correlation was more negative with the pretrained model (t(99)=6.12, p=.005, ds=0.61). For the cropped OASIS images, the weights x layers two-way interactions were also significant (F(4, 325)=564.84, ps<.001, η_p^2=.635), and the correlations were significantly more positive with the pretrained model than with the control model in all tested layers (ts>6.90, ps<.001, ds>0.35) except for Conv-3.3, where no difference was observed (t(99)=0.10, p=.918, ds=0.01). Thus, using a different DNN model, the aesthetic prototype effect was once again replicated: People prefer inanimate scenes with more prototypical high-level visual representations.

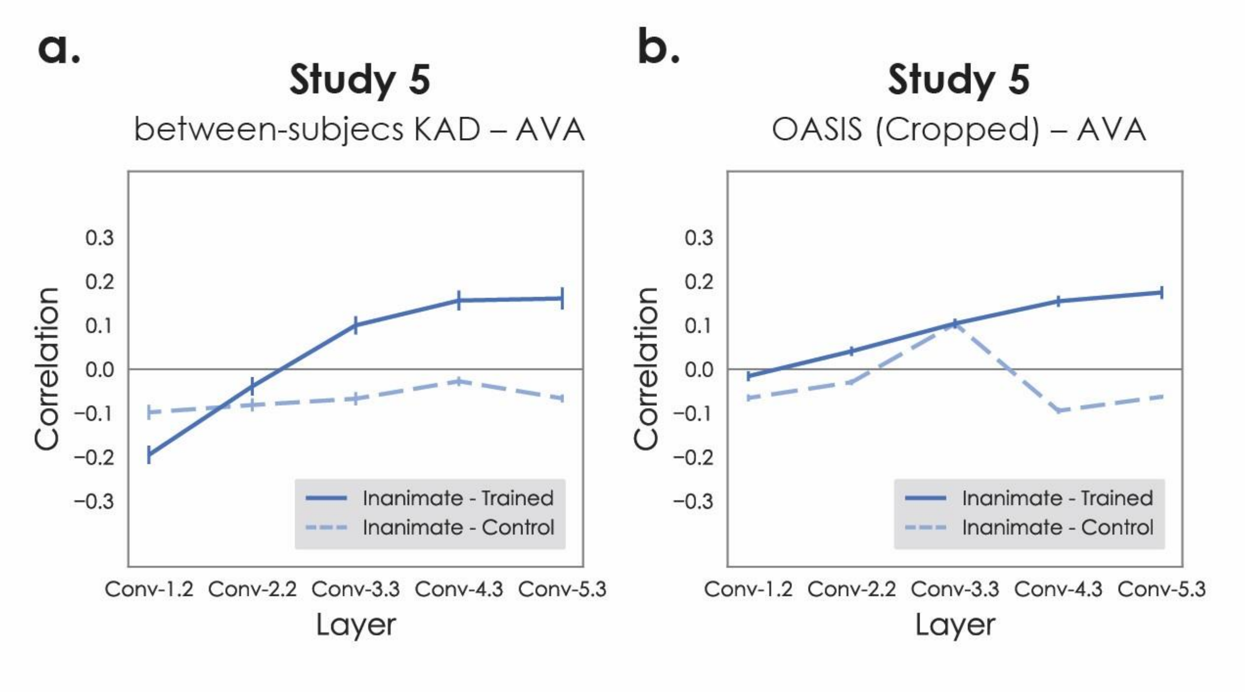

Figure 5. In the same manner in Figure 2, correlations between aesthetics and prototypicalities are plotted for Study 5, where VGG-16 was used to extract visual features and AVA was the population set. The target sets were (a) inanimate KAD images, and (b) cropped OASIS.

General Discussion

What principle governs aesthetic experiences triggered by realistic scenes? With multiple internal replications, we demonstrated a robust aesthetic prototype effect: The more prototypical an inanimate scene is, the more aesthetically pleasing it appears (Studies 1 and 2). This preference is not limited to specific stimuli, but can be observed with different population image sets (Study 3), and for both emotional and non-emotional images (Study 4). The prototype effect also does not depend on visual features extracted by a specific DNN model, as it was observed with features extracted from either AlexNet or VGG-16 (Study 5). Using permuted control models, we found that the prototype preference relies on visual representations optimized for performing object recognition tasks. Furthermore, this aesthetic principle appears to have an important boundary condition, as the same preference was not found for images containing social content.

Because layers of DNN models have been found to correspond to the hierarchy of visual areas in human brain, our findings suggest the possible stage of visual processing that gives rise to the prototype preference in aesthetic experience. Across five studies, the prototype effect was consistently observed with representations at later convolution layers around Conv-4 and Conv-5, which likely corresponds to high-level visual processing right before reaching the inferior temporal cortex (Khaligh-Razavi & Kriegeskorte, 2014). We also consistently observed a lack of prototype effect based on representations at Conv-1, suggesting the lack of participation from early visual processing.

Why do we like prototypical scenes?

Why do we like inanimate scenes with prototypical high-level visual representations? It is possible that this phenomenon arises simply as a byproduct of the non-aesthetic prototype effect and the aesthetic fluency effect: Visual processing of prototypical features tend to be prioritized and more fluent due to their centrality in the representational space (Posner & Keele, 1968; Reber et al., 1998; Winkielman et al., 2006). This perceptual fluency in turn leads to a more positive experience (Reber et al., 2004; Winkielman et al., 2006), either because it functionally serves as an internal reward for successful scene recognition, or because the positive experience from ease of processing is misinterpreted as positive scene evaluation.

An intriguing alternative explanation is that aesthetic preferences in fact served functions in our evolutionary past. Prototypical features suggest frequently encountered, known, and safe or neutral environments (e.g., everyday sightings of paths and homes), whereas atypical features may suggest unusual situations that require careful behavioral responses (e.g., a forest fire, a bloody room, a polluted lake). A relative aversion to atypical scenes may thus aid people in seeking environments that are more beneficial.

It is then quite curious that the same aesthetic principle was consistently absent for scenes containing social content. One might expect that prototypical social scenes would also lead to fluent processing or suggest safer encounters. It is possible that our analyses were insensitive to prototypicality for social scenes due to the way the DNN models used here were trained. Both AlexNet and VGG-16 were trained on ImageNet to perform an object classification task. This task involves few if any social categories. Although the images in ImageNet include many incidental human appearances, they may not exhibit the critical features required to distinguish different non-social categories. This supervised training with a limited task scope could bias the DNN models to focus on extracting informative features of inanimate objects, and thus result in representations that are less useful for predicting aesthetic experiences for scenes involving humans. Future studies should explore this issue by using models trained on stimuli and tasks with social components.

Relations to other aesthetic phenomena

How does the aesthetic prototype effect for realistic scenes relate to many past discoveries of featural preferences? We prefer stimuli that are blue (Palmer & Schloss, 2010), curvy (Bar & Neta, 2006), symmetrical (Jacobsen & Höfel, 2002), inward facing (Chen et al., 2018; Palmer et al., 2008), moderately complex (Martindale et al., 1988), and of canonical visual sizes (Chen et al., 2022; Konkle & Oliva, 2011; Linsen et al., 2011). Is the prototype preference an independent phenomenon, or does it connect to these various preferences? It is possible that specific features people like are correlated with prototypical representations. For example, prototypical features may be of moderate complexity, and thus a preference for moderate complexity is in fact a symptom of a more general aesthetic principle of prototypical preference. If this is the case, one may be able to find DNN representations that predict scene prototypicality, aesthetic preferences, and the presence of these known features.

The prototype preference may also explain mysterious individual differences in aesthetics—our individual aesthetic tastes. Life-long personal visual diets may influence how stored scenes are distributed in a representational space, shifting the prototype in different directions. Thus, the prototypicality of a scene will vary based on past visual exposures, and lead to differences in aesthetic preferences. Since visual diets can vary considerably in each person's micro-environment, this explanation of individual tastes is consistent with the findings that people's tastes do not form multiple clusters, but simply deviate in idiosyncratic ways from a common consensus (Chen et al., 2022).

Conclusion

Taking advantage of the unique information provided by DNN models, the present study was able to reveal a regularity in aesthetic experiences from diverse realistic scenes, without assuming any specific experimenter-defined features. A robust aesthetic preference for prototypes, arising from high-level visual processing, emerged as a general principle governing aesthetic experiences for inanimate scenes. Thus, aesthetic experiences triggered by complex realistic scenes are systematic, explainable, and reflect the underlying organization of visual scene representations.

Open Practices Statement

All materials, code, and data can be downloaded here: https://osf.io/mqhxg/?view_only=e8e6d8435f2749558ea1fde16bd4c951 .

Acknowledgments

We thank Felix Chang and Keith J. Holyoak for helpful conversations. This project was funded by the National Science Foundation BSC-1655300 awarded to HL.

Author Contributions

YCC, DF, and HL formulated the idea. YCC, SF, DF, JC, MT, and HL designed the research. YCC, JC, and MT prepared the materials. YCC and MT programmed and conducted the experiments. YCC, SF, DF, XC, and HL performed the analyses. YCC wrote the manuscript with all authors' input. HL acquired financial support.

References

Augustin, M. D., Wagemans, J., & Carbon, C. C. (2012). All is beautiful? Generality vs. specificity of word usage in visual aesthetics. *Acta Psychologica, 139,* 187–201.

Avrahami, J., Argaman, T., & Weiss-Chasum, D. (2004). The mysteries of the diagonal: Gender-related perceptual asymmetries. *Perception & Psychophysics, 66,* 1405–1417.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology, 14*:e1006613, 1–43.

Bar, M., & Neta, M. (2006). Humans prefer curved visual objects. *Psychological Science, 17,* 645–648.

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Perception & Psychophysics, 8,* 279–286.

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS computational biology, 14*:e1006111, 1–31.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (in press). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*.

Brielmann, A. A., & Pelli, D. G. (2019). Intense beauty requires intense pleasure. *Frontiers in Psychology, 10*:2420, 1–17.

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience, 10,* 1–34.

Chen, Y. -C., Chang, A., Rosenberg, M. D., Feng, D., Scholl, B. J., & Trainor, L. J. (2022). "Taste typicality" is a foundational and multi-modal dimension of ordinary aesthetic experience. *Current Biology, 32,* 1837–1842

Chen, Y. -C., Colombatto, C., & Scholl, B. J. (2018). Looking into the future: An inward bias in aesthetic experience driven only by gaze cues. *Cognition, 176,* 209–214.

Chen, Y. -C., Deza, A., & Konkle, T. (2022). How big should this object be? Perceptual influences on viewing-size preferences. *Cognition, 225*:105114, 1–11.

Chen, Y. -C., Pollick, F., & Lu, H. (under review). Aesthetic preferences for prototypical movements in human actions.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). DeCAF: A deep convolutional activation feature for generic visual recognition. *Proceedings of Machine Learning Research, 32,* 647–655.

Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science, 5,* 373–397.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 770–778.

Gartus, A., & Leder, H. (2013). The small step toward asymmetry: Aesthetic judgment of broken symmetries. *i-Perception, 4,* 361–364.

Halberstadt, J. B., & Rhodes, G. (2003). It's not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin & Review, 10,* 149–156.

Jacobsen, T., Buchta, K., Köhler, M., & Schröger, E. (2004). The primacy of beauty in judging the aesthetics of objects. *Psychological Reports, 94,* 1253–1260.

Jacobsen, T., & Höfel, L. E. A. (2002). Aesthetic judgments of novel graphic patterns: Analyses of individual judgments. *Perceptual and Motor Skills, 95,* 755–766.

Kanwisher, N., Khosla, M., & Dobs, K. (in press). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends in Neurosciences*.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*: e1003915, 1–29.

Konkle, T., & Oliva, A. (2011). Canonical visual size for real-world objects. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 23–37.

Krizhevsky, A. (2014). *One weird trick for parallelizing convolutional neural networks*. arXiv. https://doi.org/10.48550/arXiv.1404.5997

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

Kurdi, B., Lozano, S., & Banaji, M. R. (2017). Introducing the open affective standardized image set (OASIS). *Behavior Research Methods, 49,* 457–470.

Landwehr, J. R., Labroo, A. A., & Herrmann, A. (2011). Gut liking for the ordinary: Incorporating design fluency improves automobile sales forecasts. *Marketing Science, 30,* 416–429.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1,* 115–121.

Latto, R., Brian, D., & Kelly, B. (2000). An oblique effect in aesthetics: Homage to Mondrian (1872–1944). *Perception, 29,* 981–987.

Leech, G., & Rayson, P. (2014). *Word frequencies in written and spoken English*. Routledge.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience, 33,* 2017–2031.

Linsen, S., Leyssen, M. H., Sammartino, J., & Palmer, S. E. (2011). Aesthetic preferences in the size of images of real-world objects. *Perception, 40,* 291–298.

Locher, P., Overbeeke, K., & Stappers, P. J. (2005). Spatial balance of color triads in the abstract art of Piet Mondrian. *Perception, 34,* 169–189.

Martindale, C., & Moore, K. (1988). Priming, prototypicality, and preference. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 661–670.

Martindale, C., Moore, K., & West, A. (1988). Relationship of preference judgments to typicality, novelty, and mere exposure. *Empirical Studies of the Arts, 6,* 79–96.

Mather, G. (2012). Aesthetic judgement of orientation in modern art. *i-Perception, 3,* 18–24.

Murray, N., Marchesotti, L., & Perronnin, F. (2012). AVA: A large-scale database for aesthetic visual analysis. *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2408–2415.

Palmer, S. E., Gardner, J. S., & Wickens, T. D. (2008). Aesthetic issues in spatial composition: Effects of position and direction on framing single objects. *Spatial Vision, 21,* 421–449.

Palmer, S. E., & Schloss, K. B. (2010). An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences, 107,* 8877–8882.

Posner, M. l., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience, 38,* 7255–7269.

Reber, P. J., Stark, C. E. L., & Squire, L. R. (1998). Cortical areas supporting category learning identified using functional MRI. *Proceedings of the National Academy of Sciences, 95,* 747–750.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8,* 382–439.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision, 115,* 211–252.

Russell, P. A., & George, D. A. (1990). Relationships between aesthetic response scales applied to paintings. *Empirical Studies of the Arts, 8,* 15–30.

Ryali, C. K., Goffin, S., Winkielman, P., & Yu, A. J. (2020). From likely to likable: The role of statistical typicality in human social assessment of faces. *Proceedings of the National Academy of Sciences, 117,* 29371–29380.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K. & DiCarlo, J. J. (2020). *Brain-score: Which artificial neural network for object recognition is most brain-like?* bioRxiv. https://doi.org/10.1101/407007

Silvia, P. J., & Barona, C. M. (2009). Do people prefer curved objects? Angularity, expertise, and aesthetic preference. *Empirical Studies of the Arts, 27,* 25–42.

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv. https://doi.org/10.48550/arXiv.1409.1556

Solso, R. L., & Raynis, S. A. (1979). Prototype formation from imaged, kinesthetically, and visually presented geometric figures. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 701–712.

Son, G., Walther, D. B., & Mack, M. L. (2021). Scene wheels: Measuring perception and memory of real-world scenes with a continuous stimulus space. *Behavior Research Methods, 54,* 444–456.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). *Intriguing properties of neural networks*. arXiv. https://doi.org/10.48550/arXiv.1312.6199

Whitfield, T. A., & Slatter, P. E. (1979). The effects of categorization and prototypicality on aesthetic choice in a furniture selection task. *British Journal of Psychology, 70,* 65–75.

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because they are easy on the mind. *Psychological Science, 17,* 799–806.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111,* 8619–8624.

Younger, B. (1990). Infant categorization: Memory for category-level and specific item information. *Journal of Experimental Child Psychology, 50,* 131–155.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. *Advances in Neural Information Processing Systems, 27,* 1–9.