Prototype principle explains aesthetics for general visual experiences beyond isolated objects

Yi-Chia Chen (0000-0002-8321-8595; yichiachen@ucla.edu)[1*],
Shuhao Fu (0000-0003-2672-5125; fushuhao@g.ucla.edu)[1],
Derek Feng (0000-0002-7860-0968; derek.feng@aya.yale.edu)[2],
Moriah Taylor (moriahtaylor095@gmail.com)[1],
Jeff Chang (jchang0330@gmail.com)[1],
Xiaoyang Chi (xiaoyangchi@ucla.edu)[1],
Hongjing Lu (0000-0003-0660-1176; hongjing@g.ucla.edu)[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,
1285 Franz Hall, Box 951563, Los Angeles, CA 90095-1563

Running Head : Prototype preferences beyond object perception
Funding Source : This project was funded by the National Science Foundation BSC-1655300 awarded to HL.
Acknowledgements : We thank Felix Chang and Keith J. Holyoak for helpful conversations.
Word Count : 3000 words (Introduction + Results + Discussion)

Abstract

A salient aspect of our daily visual experiences is the aesthetic impressions. Why do we enjoy some views but not others? Researchers have found visual attributes that we enjoy seeing for colors, shapes, and objects, and even for human faces and bodies; however, aesthetic preferences for holistic visual experiences—views that encompass many different objects and attributes—remain largely undescribed. What principles govern holistic visual aesthetics beyond preferences for certain objects or features? We found computational evidence in support of the prototype principle for visual aesthetics. We measure each view's prototypicality based on its average similarity to a large reference set of images in feature spaces from pretrained deep convolutional neural networks (AlexNet and VGG-16). For views consisting of inanimate content, the more prototypical the high-level visual representations are, the more aesthetically pleasing the views appear. This prototype effect suggests a functional answer to why we like what we like: We like typical visual experiences that are common, safe, or neutral. Thus, aesthetic preferences from holistic visual experiences are systematic, explainable, and reflect the underlying organization of visual representations of the world. Our method demonstrates a new way to explore both vision and aesthetics using computational models—besides investigating specific visual properties or isolated objects, we can also explore the organization of visual representations and aesthetic experiences holistically, getting us one step closer to understanding how we see, what we like, and why we like what we like in everyday life.

Keywords: aesthetics; prototype; deep learning; visual scene

1. Introduction

Whether we are walking in our own neighborhoods or flying to the opposite side of the world, people invest time, resources, and effort to see new sceneries. Besides expanding knowledge, a major purpose of such behaviors is to experience the beauty of what we see. This massive popularity of "sightseeing" comes with abundant mysteries, along with one central question: Why do some views look better than others?

1.1 Regularities in visual aesthetics

A common approach to understand "what looks good" is to ask how specific attributes influence aesthetic experiences, 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). These properties were commonly studied with simple stimuli, such as shapes (Gartus & Leder, 2013; Silvia & Barona, 2009), objects (Halberstadt & Rhodes, 2003), and abstract compositions (Locher et al., 2005). While simple and direct, this line of research demonstrated meaningful consensus and regularities in aesthetics across people. Nevertheless, it left an important question unanswered: How do we explain aesthetic impressions in *holistic visual experiences* from the real world, which encompass many objects and features?[1]

Some studies pioneered in exploring holistic visual aesthetics. One route taken was to circumvent the difficulties in characterizing holistic visual experiences as testable factors and explore neural correlates of aesthetics (Kaiser, 2022; Vessel et al., 2019; Yue et al., 2007). While this approach described what substrates realized the mental computations for aesthetics (at the implementation level; Marr, 1982), it provides limited insight for the computations themselves (at the computational and algorithmic levels). Other studies alternatively investigated influences from experimenter-selected factors, such as naturalness (Kaplan et al., 1972), symmetry (Damiano et al., 2023), and contour properties (Farzanfar & Walther, 2023). While this approach produces intriguing findings, it is constrained to specific dimensions according to researchers' intuitions.

1.2 The present study: The aesthetic prototype effect

The present study moved from exploring specific featural preferences to understanding general principles for holistic visual aesthetics that may provide not only answers to "what we like", but also "why we like what we like". We focused on a pervasive

[1] The caveats of reductionist approach in visual aesthetics in fact came up in researching scene perception as well: Just as it has not proved possible to combine featural aesthetic effects to predict holistic visual aesthetics, it has not been possible to explain holistic scene perception based solely on visual features selected or defined by researchers (Epstein & Baker, 2019).

phenomenon that connects multiple aspects of vision: the *prototype effect*, in which people show prioritizations for central representations of categories. Specifically, visual preferences for category prototypes have been observed for 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 (Posner & Keele, 1968; Solso & Raynis, 1979), and dynamic stimuli such as biological motion (Chen et al., 2023). This robust phenomenon not only reveals how feature processing led to category-based organizations of visual representations, but also provides natural explanations to aesthetic experiences, such as deviance detection (for alternatives, see Vogel et al., 2021). Thus, we asked: Can we explain aesthetics with a prototype principle for rich visual experiences?

To answer this question, we overcame two difficulties: First, prototypes are typically defined within an experimenter-selected category. However, how holistic visual experiences are categorized in visual systems is empirically unanswered. Thus, we did not limit the present study to certain categories of scenes, and instead, used prototype effects to reveal the categorical structures. That is, the discovery of a prototype effect would suggest the existence of a visual category (or a cluster of several categories) among the tested images. To start, we explored the prototype preferences in the broadest possible category—the full diversity of visual experiences—whether or not they belong to categories previously tested (i.e., specific scene types or objects). Second, it is not straightforward to empirically measure prototypicalities of holistic visual experiences. We decided against subjective ratings and estimated prototypicality using representations discovered in a data-driven fashion from deep neural networks (DNNs). While it remains debated whether DNNs accurately model human vision (Baker et al., 2018; Szegedy et al., 2013), they have undoubtedly uncovered a subset of useful feature representations for common tasks such as object recognition (Rajalingham et al., 2018; Yamins et al., 2014).

In six studies, we used diverse realistic photographs to capture rich visual experiences and tested the relationship between prototypicalities and people's aesthetic impressions from these views. To rule out specific alternative explanations, we examined three samples of observers and four diverse image sets, and used features discovered in two DNN models that were pretrained with different image sets.

2. Study 1: An Aesthetic Prototype Effect for Inanimate Visual Experiences

2.1 Method

2.1.1 Overview. The prototypicality of a set of images were estimated by comparing their DNN features to each image in a separate and larger image set. Human observers rated aesthetic impressions for images in the target set to assess the effect of prototypicality on aesthetic values (Figure 1). We tested images containing social

information or only inanimate information as two conditions since they tend to engage different visual processing (Caramazza & Shelton, 1998). 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 (Baek et al., 2021; Kim et al., 2021; Ramanujan et al., 2020).

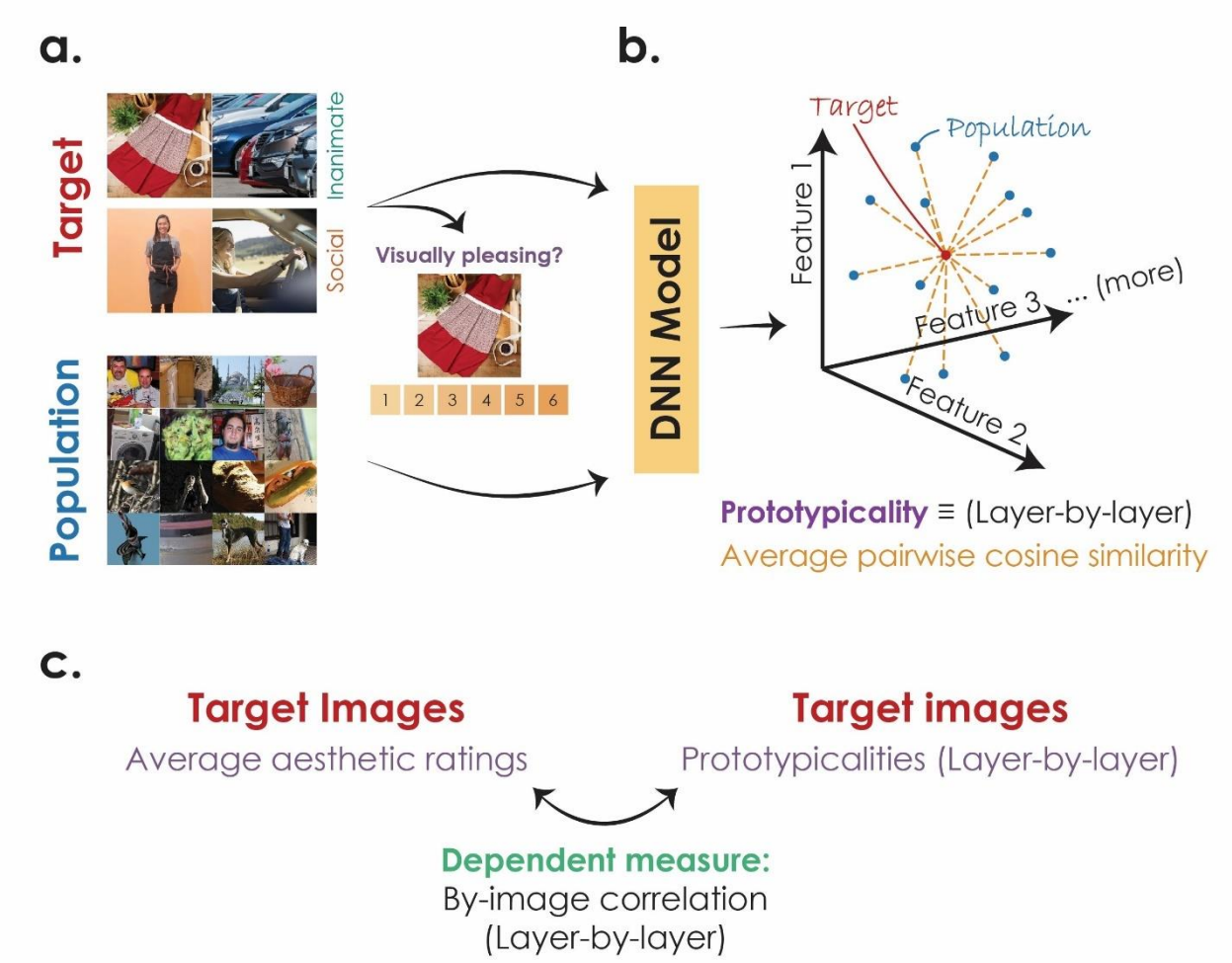

Figure 1 Procedure Overview. An overview of the procedure is depicted here: (a) Two sets of images—the target and population image set—were used to measure prototype preference. A hundred human observers rated the images in the target set on their aesthetic values. (b) DNN features for images in both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

2.1.2 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 .

2.1.3 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 with the next eligible images (sorted by file names). Thus, they included both inanimate and social images. These images were then resized to 256 x 256 px and cropped to retain a random 224 x 224 px region.

2.1.4 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). An intuitively 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.

2.1.5 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.

2.1.6 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

2.1.7 The pretrained visual feature model. We chose DNN models to extract visual features in images that has been proved useful for object recognition. AlexNet (Krizhevsky, 2014; Krizhevsky et al., 2017) was used due to its popularity and simplicity relative to other models. 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.

2.1.8 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.

2.1.9 Prototypicality estimation. To measure the prototypicality of each image (Figure 1b), we separately analyzed the features from the activities in each layer of the DNN models. For each image in the target set, we measured the pairwise cosine similarities between its feature vector and the feature vector for each image in the population set. Then, we averaged the target image's similarities to all population images. A higher similarity value indicates that the target image is more closely aligned with the average representation of the visual scenes depicted in the population images. 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.

2.1.10 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. For human responses, 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.

2.2 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. These results are publicly shared (https://osf.io/mqhxg/?view_only=e8e6d8435f2749558ea1fde16bd4c951). From inspecting the figures, three observations emerged: (a) The pretrained and control models led to different results. (b) For the pretrained model, the results differed between the inanimate and social images. (c) The inanimate images showed positive correlations only in layers above Conv-2 of pretrained AlexNet. These observations were confirmed by a repeated-measure ANOVA: The weights x layers x image types interaction was significant ($F(12, 1188)=22.36$, $p<.001$, $\eta_p^2=.184$), 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 prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (ts(99)>2.80, ps<.006, ds>0.28) except for Conv-1 ($t(99)=1.86$, $p=.066$, $d=0.19$). For the social images, the correlations were similar across the models, not significantly different in Conv-2 and Conv-4 (ts(99)<1.57, ps>.119, ds<0.16), and significantly more negative with the pretrained model in the rest of the layers (ts(99)>3.05, ps<.003, ds>0.30). Thus, the more prototypical an inanimate image is based on representations extracted later than Conv-2, the more aesthetically pleasing it appears. This aesthetic prototype effect was not observed in the earlies layer nor for images containing social content.

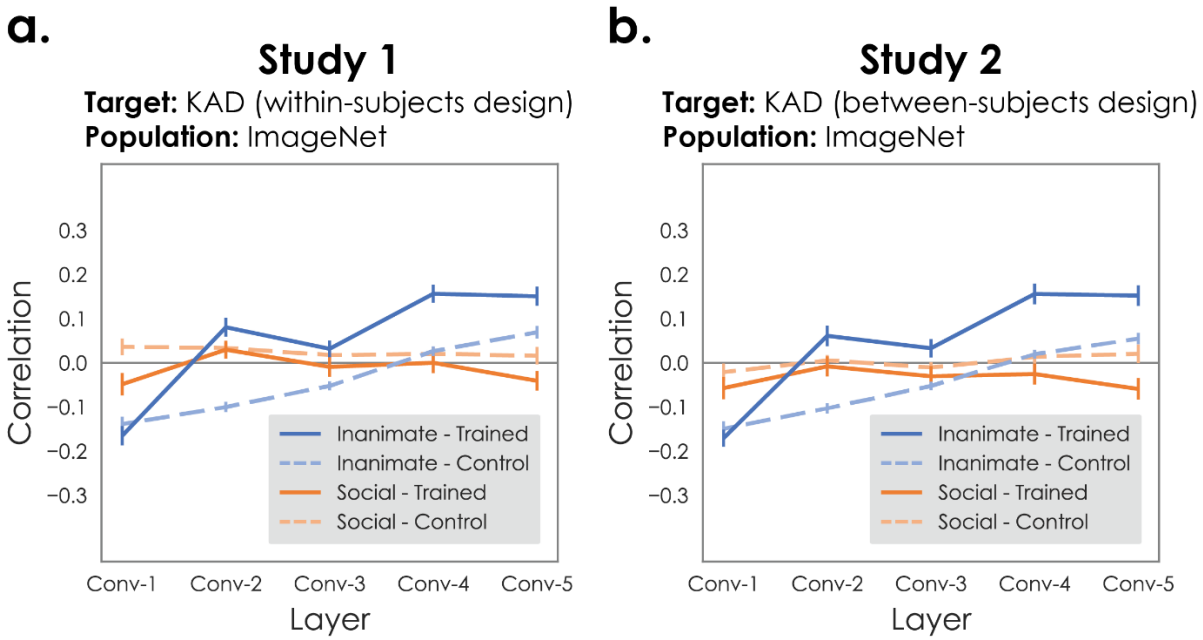

Figure 2 Results from Study 1 and 2. For (a) Study 1(within-subjects design) and (b) Study 2 (between-subjects design), correlations between human aesthetic ratings (z-scored) and image prototypicalities from each convolutional layer of AlexNet are plotted separately for four conditions. All error bars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

3. Study 2: Between-subjects Replication

Could the differences found between inanimate and social images arise from task demands, such as applying different strategies upon noticing the two categories? We ruled out this possibility with a between-subjects replication experiment.

3.1 Method

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

3.1.1 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.

3.1.2 Experiment procedure. All procedures were identical to Study 1 except that the observers were randomly and evenly assigned to the inanimate or social image

condition, and only rated the 78 images and 35 repeats in their respective conditions.

3.1.3 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).

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

3.2 Results and discussion

The results are plotted in Figure 2b. The same observations from Study 1 were apparent: Positive correlations were observed for layers above Conv-2 only for the pretrained AlexNet and the inanimate images. These observations were confirmed by the mixed-measure ANOVA. The weights x layers x image types interaction was significant ($F(12, 2376)=36.70$, $p<.001$, $\eta_p^2=.156$), along with all the two-way interactions (Fs>40.00, ps<.001, η_p^2s>.170). Critically, for the inanimate images, the correlations between the ratings and the prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (ts(99)>4.00, ps<.001, ds>0.40) except for Conv-1 ($t(99)=1.45$, $p=.150$, $d=0.15$). For the social images, the correlations were similar across the models, not significantly different in Conv-2 ($t(99)=1.20$, $p=.233$, $d=0.12$), and significantly more negative with the pretrained model in the rest of the layers (ts(99)>2.03, ps<.045, ds>0.20). Thus, after eliminating possible demand characteristics caused by rating both inanimate and social images, we replicated the findings from Study 1: The more prototypical an inanimate image is based on visual representations extracted in layers later than Conv-2, the more aesthetically pleasing it appears. This aesthetic prototype effect was again not observed for low-level visual presentations nor for images containing social information.

4. Study 3: Generalization to An Alternative Image Population

How general is the aesthetic prototype effect? We replicated the effect with a more

diverse population image set to assess whether the prototype preference operates over varying content.

4.1 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.

4.2 Results and discussion

The results are plotted in Figure 3a and 3b, where similar patterns emerged. Positive correlations were observed for layers above Conv-2 only with inanimate images using the pretrained model. This observation was confirmed by the ANOVAs: For the within-subjects experiment, the weights x layers x image types interaction 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). For the inanimate images, the correlations between the ratings and the prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (ts(99)>8.40, ps<.001, ds>0.84) except for Conv-1, where the correlation was more negative with the pretrained model ($t(99)=9.20$, $p<.001$, $d=0.92$). For the social images, the correlations were significantly more *negative* with the pretrained model than in the control model in all layers (ts(99)>2.31, ps<.023, ds>0.23).

For the between-subjects experiment, the weights x layers x image types interaction 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). For the inanimate images, the correlations between the ratings and the prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (ts(99)>8.65, ps<.001, ds>0.86) except for Conv-1, where the correlation was more negative with the pretrained model ($t(99)=8.39$, $p<.001$, $d=0.84$). For the social images, the correlations were similar across the models, not significantly different in Conv-3 ($t(99)=1.66$, $p=.100$, $d=0.17$), and significantly more negative with the pretrained model in the rest of the layers (ts(99)>2.16, ps<.033, ds>0.21. The results from both Study 1 and 2 were thus replicated in Study 3, and

generalized using AVA as the population image set.

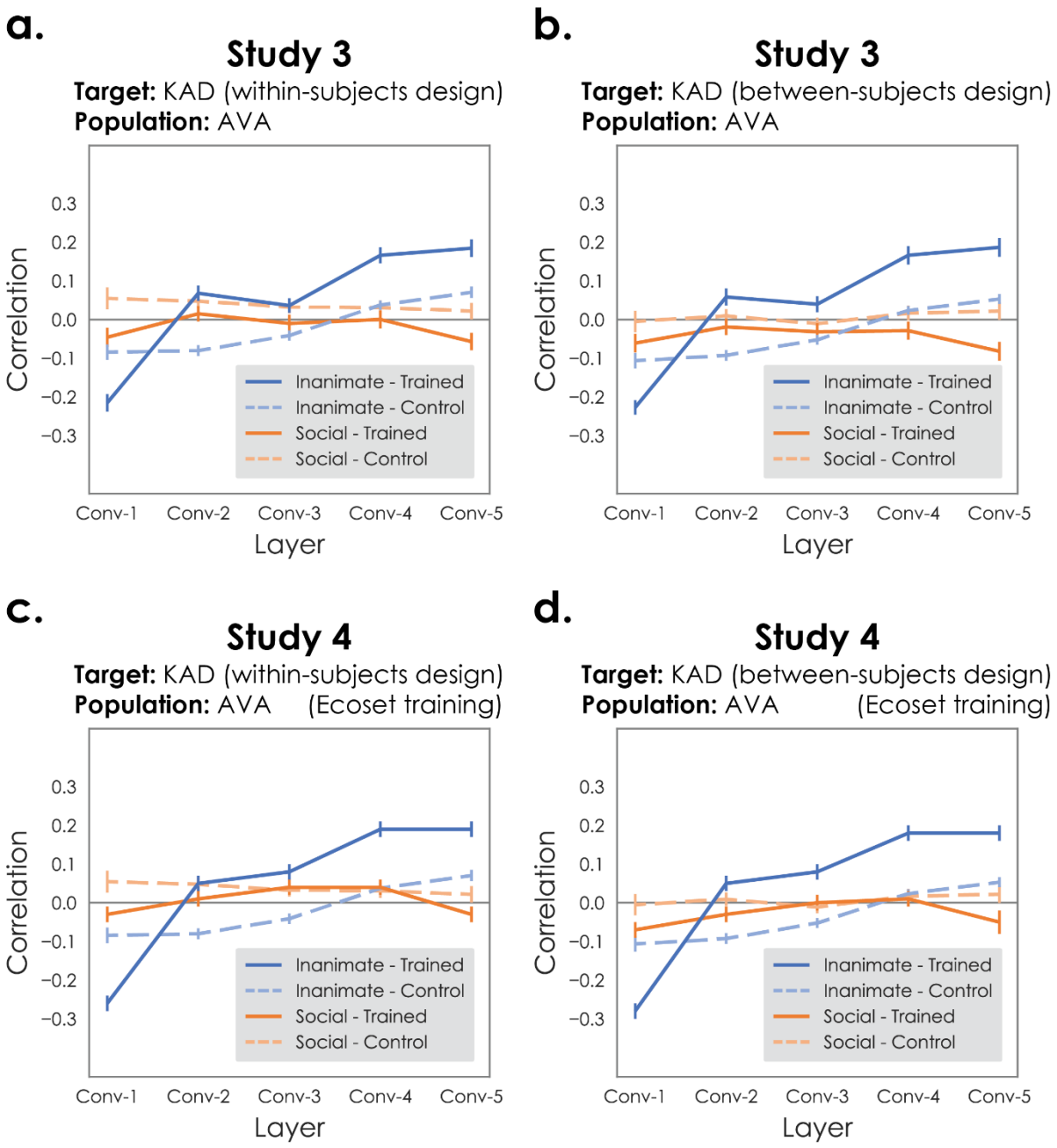

Figure 3 Results from Study 3 and 4. In the same manner as in Figure 2, correlations between aesthetics and prototypicalities are plotted for Study 3 and 4, where AVA served as the population image set. Aesthetic ratings from the within-subjects experiment were used in (a) and (c), and those from the between-subjects experiment were used in (b) and (d). AlexNet pretrained with ImageNet was used in (a) and (b), and AlexNet pretrained with the ecoset (including people categories) was used in (c) and (d).

5. Study 4: Generalization to a Model Trained with People Categories

Could the differences found between inanimate and social images arise from a lack of social features extracted by a model that was not trained to classify people? We ruled out this possibility using AlexNet pretrained with different people categories.

4.1 Method

The same analyses from Study 3 were replicated with an AlexNet that was pretrained to classify the categories in a new image dataset termed ecoset, which captures categories relevant to humans (Mehrer et al., 2021). The categories were selected based on frequent and concrete words and included inanimate objects, animals, and people (e.g., hairbrush, chipmunk, and child). Only the five convolutional layers were included in the analyses.

4.2 Results and discussion

The results are plotted in Figure 3c and 3d, where similar patterns emerged. Positive correlations were observed for layers above Conv-2 only with inanimate images using the pretrained model. This observation was confirmed by the ANOVAs: For the within-subjects experiment from Study 1, the weights x layers x image types interaction was significant ($F(4, 396)=199.59$, $p<.001$, $\eta_p^2=.668$), along with all the two-way interactions (Fs>50.00, ps<.001, η_p^2s>.340). For the inanimate images, the correlations between the ratings and the prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (ts(99)>9.50, ps<.001, ds>0.95) except for Conv-1, where the correlation was more negative with the pretrained model ($t(99)=14.98$, $p<.001$, $d=1.50$). For the social images, the correlations were similar across the models, not significantly different in Conv-3 and Conv-4 (ts(99)<0.95, ps>.350, ds<0.10), and significantly more *negative* with the pretrained model in the rest of the layers (ts(99)>2.50, ps<.015, ds>0.25).

For the between-subjects experiment from Study 2, the weights x layers x image types interaction was also significant ($F(4, 792)=212.65$, $p<.001$, $\eta_p^2=.518$), along with all the two-way interactions (Fs>45.00, ps<.001, η_p^2s>.190). For the inanimate images, the correlations between the ratings and the prototypicalities were significantly more positive with the pretrained model than with the control model in all layers (ts(99)>10.00, ps<.001, ds>1.00) except for Conv-1, where the correlation was more negative with the pretrained model ($t(99)=16.26$, $p<.001$, $d=1.63$). For the social images, the correlations were similar across the models, not significantly different in Conv-3 and Conv-4 (ts(99)<0.90, ps>.380, ds<0.09), and significantly more *negative* with the pretrained model in the rest of the layers (ts(99)>2.65, ps<.010, ds>0.25). The findings were thus replicated even with a model

pretrained with people categories.

6. Study 5: Generalization to a Diverse Target Image Set

In the previous four 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. We ruled out this possibility by generalizing the prototype effect to a new target set that contained dramatic emotional content.

6.1 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). The pretrained AlexNet model was the standard version which was trained on ImageNet.

6.1.1 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 Studies 1 thorough 3. 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.

6.1.2 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.

6.2 Results and discussion

The results are plotted in Figure 4, where a clear pattern emerged. The pretrained model showed more positive correlations compared to the control models for all conditions. This observation was 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, ps<.001, η_p^2s>.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 (ts(325)>16.10, ps<.001, ds>0.89). The same target images with AVA as the population set led to similar results, where the two-way interactions of weights x layers were significant (Fs>160.00, ps<.001, η_p^2s>.300), and the correlations were significantly more positive with the pretrained model than with the control model in all layers (ts(325)>9.95, ps<.001, ds>0.55). Thus, using a different set of images containing emotional content, the aesthetic prototype effect for inanimate images was again replicated, demonstrating its generality.

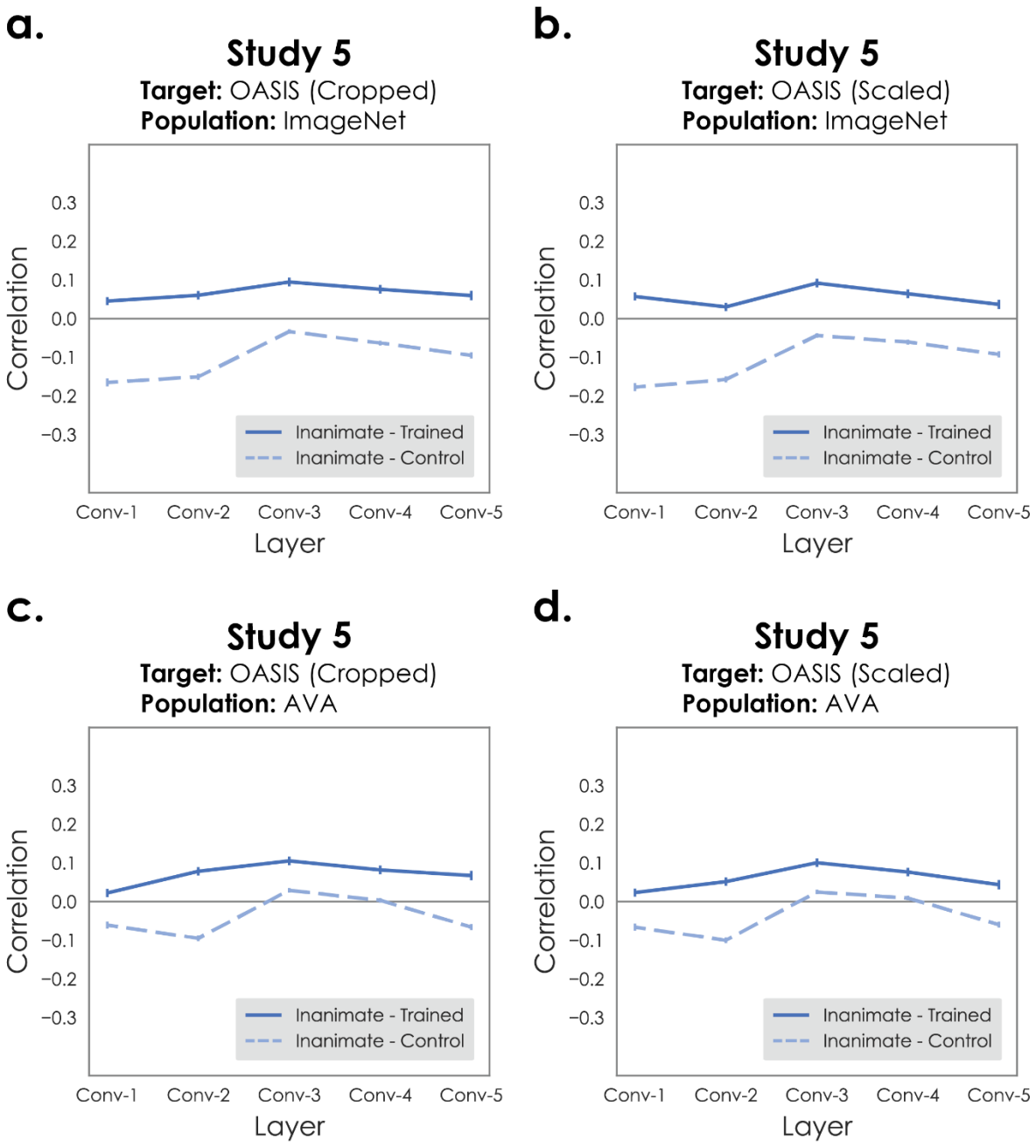

Figure 4 Results from Study 5. In the same manner as in Figure 2, correlations between aesthetics and prototypicalities are plotted for Study 5, where OASIS was used as the target image set. The population image set was ImageNet for (a) and (b), and was AVA for (c) and (d). Cropped OASIS images were used for (a) and (c), and scaled OASIS images were used for (b) and (d).

7. Study 6: Generalization to An Alternative DNN Model

We next generalized the findings using a different DNN model: 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 large number of features would have exhausted our hardware memory capacity at the time of this study.

7.1 Method

We conducted the same analyses using a larger DNN model, VGG-16. Given the large number of layers in VGG-16, we used only 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 (the inanimate KAD images and the cropped OASIS images) and one population set (the AVA).

7.2 Results and discussion

The results are plotted in Figure 5, where a clear pattern emerged. The pretrained model showed more positive correlations than the control model in later layers. This observation was confirmed by two separate repeated-measure ANOVAs. For the inanimate KAD images, the weights x layers interaction was significant ($F(4, 99)=172.12$, $ps<.001$, $\eta_p^2=.635$), and the correlations were significantly more positive with the pretrained model than with the control model in all tested layers ($ts(99)>2.80$, $ps<.001$, $ds>0.25$) 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 interaction was also significant ($F(4, 325)=564.84$, $ps<.001$, $\eta_p^2=.635$), and the correlations were significantly more positive with the pretrained model than with the control model in all tested layers ($ts(325)>6.90$, $ps<.001$, $ds>0.35$) except for Conv-3.3, where no difference was observed ($t(325)=0.10$, $p=.918$, $ds=0.01$). Thus, using a different DNN model, the aesthetic prototype effect was again replicated.

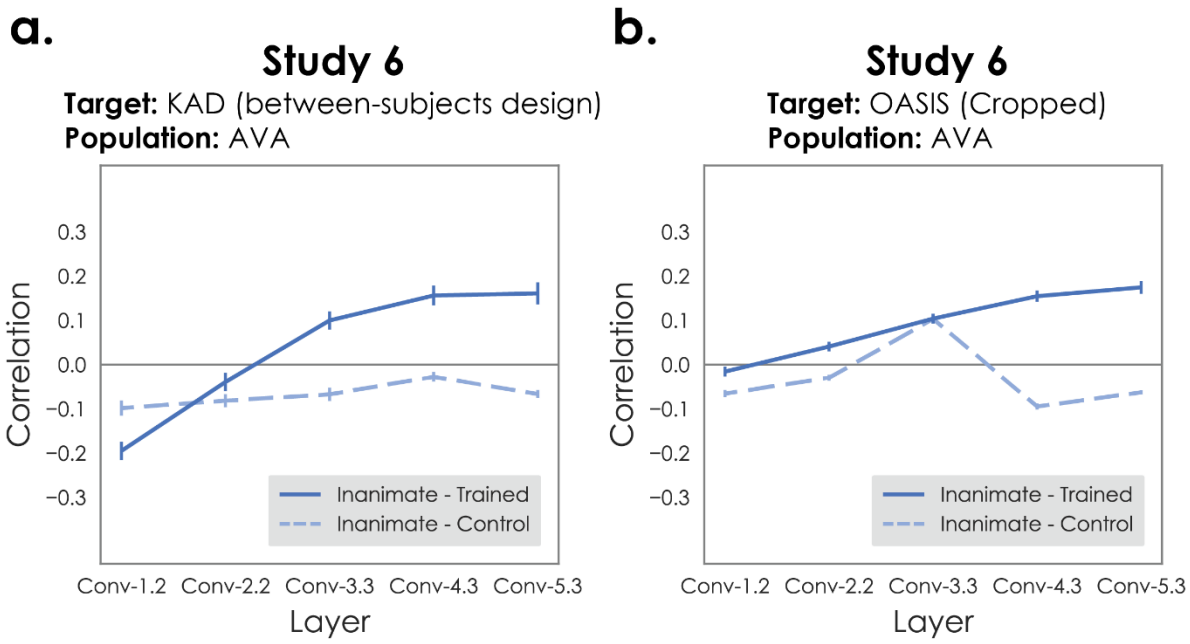

Figure 5 Results from Study 6. In the same manner in Figure 2, correlations between aesthetics and prototypicalities are plotted for Study 6, where VGG-16 was used to extract visual features and AVA was the population set. The target sets were (a) inanimate KAD images, and (b) cropped OASIS.

8. General Discussion

What principle governs aesthetic impressions arising from holistic visual experiences? We demonstrated a robust aesthetic prototype effect: The more prototypical an inanimate image is, the more aesthetically pleasing it appears (Studies 1 and 2). This preference can be observed with different population image sets (Study 3), for both emotional and non-emotional images (Study 5), and does not rely on features only discovered in specific DNN models (Study 4 and 6). Our findings also reveal possible stages of visual processing that gives rise to the prototype preference. Across six 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 (Khaligh-Razavi & Kriegeskorte, 2014). Also, the prototype effect was absent based on representations at Conv-1, suggesting the lack of participation from early vision.

The correlations found were robust yet small. However, we cannot conclude that the prototype principle matters only a little. While we asked the qualitative question of whether prototype effects exist, this study did not seek to perfectly model prototypicality and thus could underestimate effect sizes. Additional research is required to assess the strength of prototype effects in holistic visual experiences.

8.1 Why do we like prototypical inanimate visual experiences?

Why do we like inanimate visual experiences with prototypical high-level representations? This preference may arise simply as a byproduct in the fashion of a fluency effect: Visual processing of prototypical features are 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 positive experience (Reber et al., 2004; Winkielman et al., 2006), either because it serves as an internal reward for successful recognition, or because the positive experience from ease of processing is misinterpreted as positive evaluation for the visual experience.

An intriguing alternative explanation is that aesthetic preferences 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 visual experiences may thus aid people to seek beneficial environments.

Why then was the same aesthetic principle absent for social content? Perhaps the specialized visual processes for social content do not construct categories that lead to prototype prioritizations. Alternatively, other social factors (e.g., expressions) may overpower the prototype preferences. Future studies are required to explore these possibilities.

8.2 Relations to other aesthetic phenomena

How does the current findings 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? Specific features people like may correlate 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 prototype principle. If this is the case, one may be able to find DNN representations that predict visual prototypicality, aesthetic preferences, and the presence of these known features.

The prototype preference may also explain individual aesthetic tastes. Life-long personal visual diets may influence how stored visual experiences are distributed in a representational space, shifting the prototype in different directions. Thus, the prototypicality of an image varies based on past visual exposures, and leads to differences in aesthetic preferences. Since visual diets can vary considerably in each person's micro-environment, this explanation 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).

8.3 Conclusion

Taking advantage of the data-driven DNN model features, the present study was able to reveal robust aesthetic preferences for holistic visual experiences containing high-level prototypical features. This proves that aesthetic impressions triggered by holistic visual experiences are systematic, explainable, and reflect the underlying organization of visual representations.

Declarations

Funding
This project was funded by the National Science Foundation BSC-1655300 awarded to HL.

Conflicts of interest/Competing interests
The authors declare no conflicts of interest.

Ethics approval
The study was approved by the UCLA Institutional Review Board.

Consent to participate
Informed consent was obtained from all individual participants included in the study.

Consent for publication:
Not applicable.

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

Authors' 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.

Acknowledgments

We thank Felix Chang and Keith J. Holyoak for helpful conversations.

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.

Baek, S., Song, M., Jang, J., Kim, G., & Paik, S. B. (2021). Face detection in untrained deep neural networks. *Nature Communications, 12*:7328, 1–15.

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.

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. (2023). Aesthetic preferences for prototypical movements in human actions. *Cognitive Research: Principles and Implications, 8*:55, 1–13.

Damiano, C., Wilder, J., Zhou, E. Y., Walther, D. B., & Wagemans, J. (2023). The role of local and global symmetry in pleasure, interest, and complexity judgments of natural scenes. *Psychology of Aesthetics, Creativity, and the Arts, 17,* 322–337.

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

Farzanfar, D., & Walther, D. B. (2023). Changing what you like: Modifying contour properties shifts aesthetic valuations of scenes. *Psychological Science, 34,* 1101–1120.

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.

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.

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.

Kaiser, D. (2022). Characterizing dynamic neural representations of scene attractiveness. *Journal of Cognitive Neuroscience, 34,* 1988–1997.

Kaplan, S., Kaplan, R., & Wendt, J. S. (1972). Rated preference and complexity for natural and urban visual material. *Perception & Psychophysics, 12,* 354–356.

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.

Kim, G., Jang, J., Baek, S., Song, M., & Paik, S. B. (2021). Visual number sense in untrained deep neural networks. *Science Advances, 7*:eabd6127, 1–9.

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.

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.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* W.H. Freeman.

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.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences, 118*:e2011417118, 1–9.

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.

Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., & Rastegari, M. (2020). What's hidden in a randomly weighted neural network? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 11893–11902.

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.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? Personality and Social Psychology Review, 8, 364–382.

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.

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

Vessel, E. A., Isik, A. I., Belfi, A. M., Stahl, J. L., & Starr, G. G. (2019). The default-mode network represents aesthetic appeal that generalizes across visual domains. *Proceedings of the National Academy of Sciences, 116,* 19155-19164.

Vogel, T., Ingendahl, M., & Winkielman, P. (2021). The architecture of prototype preferences: Typicality, fluency, and valence. *Journal of Experimental Psychology: General, 150,* 187–194.

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.

Yue, X., Vessel, E. A., & Biederman, I. (2007). The neural basis of scene preferences. *Neuroreport, 18,* 525-529.