Correct for Whom? Subjectivity and the Evaluation of Personalized Image Aesthetics Assessment Models

Samuel Goree[1], **Weslie Khoo**[2], **David J. Crandall**[2]

[1] Department of Informatics, Indiana University [2] Department of Computer Science, Indiana University
sgoree@iu.edu, weskhoo@iu.edu, djcran@indiana.edu

Abstract

The problem of image aesthetic quality assessment is surprisingly difficult to define precisely. Most early work attempted to estimate the average aesthetic rating of a group of observers, while some recent work has shifted to an approach based on few-shot personalization. In this paper, we connect few-shot personalization, via Immanuel Kant's concept of disinterested judgment, to an argument from feminist aesthetics about the biased tendencies of objective standards for subjective pleasures. To empirically investigate this philosophical debate, we introduce PR-AADB, a relabeling of the existing AADB dataset with labels for pairs of images, and measure how well the existing ground truth predicts our new pairwise labels. We find, consistent with the feminist critique, that both the existing ground truth and few-shot personalized predictions represent some users' preferences significantly better than others, but that it is difficult to predict when and for whom the existing ground truth will be correct. We thus advise against using benchmark datasets to evaluate models for personalized IAQA, and recommend caution when attempting to account for subjective difference using machine learning more generally.

Introduction

Over the past fifteen years, computer vision researchers have investigated techniques for image aesthetic quality assessment (IAQA). This research area emerged from image quality assessment (Ke, Tang, and Jing 2006) and computational aesthetics (Datta et al. 2006; Datta, Li, and Wang 2008). Originally, the goal was to classify a photograph as either "high quality" or "low quality," trying to predict an average of many labelers' judgments of the photo (Murray, Marchesotti, and Perronnin 2012).

Recently, however, some researchers have claimed that aesthetic quality is *fundamentally subjective* — not an attribute of the image itself but of a human user's perception of that image. These authors have begun to pose the problem in terms of distribution learning (Cui et al. 2017; Fang et al. 2018) or few-shot personalization (Ren et al. 2017; Lee and Kim 2019; Cui et al. 2020; Zhu et al. 2021; Kairanbay, See, and Wong 2018, 2019). To account for variation between users, some methods use auxiliary information from social media data (Cui et al. 2017), demographics (Kairanbay, See, and Wong 2018, 2019), or psychometrics (Zhu et al. 2021) to better capture a given user's perspective. In parallel, other researchers (Lv and Tian 2016; Lee and Kim 2019) have turned away from aesthetic quality as a real or boolean-valued score assigned to each image and towards a pairwise comparison between two images.

From a computer science perspective, these sorts of changes to a problem statement might seem minor, however philosophically attempting to account for the subjectivity of a user takes IAQA in a highly unusual direction for machine learning which we believe is worth examining closely.

This idea, that subjective difference exists but can be rationally explained, has roots in the work of the 18th century philosopher Immanuel Kant. Kant claims that when we call an object beautiful, we imply not just that we like it, but that all other rational people should feel the same way about that object. This position assumes that the subjective conditions for judgment are essentially the same among all rational people — a central assumption in Kant's philosophical system (Pluhar 1987). Disagreements over matters of taste only exist because they are "bound up with interest," meaning that they are made based on external factors like our desires, future gratification or pleasure in looking (Kant 1790). But, if we look beyond those personal interests, we find a universal *disinterested* judgment.

IAQA is steeped in Kantian ideas about interested and disinterested judgment. Early papers in this area attempt to access a universal kind of aesthetic quality in photographs, and ascribe individual variation to noise, similar to the way that Kant ascribes individual variation to personal interest. For example, Datta et al. (2006) claim that certain visual characteristics cause images to be, in general, more aesthetically appealing, and cite Kant and discuss his concept of taste in a later paper (Joshi et al. 2011). Similarly, when proposing their well-known AVA dataset for IAQA, Murray, Marchesotti, and Perronnin (2012) observe that the score distributions for images usually look fairly Gaussian, indicating that the mean score is a good estimate of the overall quality of the image. In this way, IAQA research treats the mean of several individuals' judgments as a universal disinterested judgment, abstracted from any one rater's particular perspective. Likewise, personalized models purport to add that perspective back in by accounting for deviation due to external factors like demographics, personality, preferences for aesthetic qualities or specific photo content. While relying on Kant's framework gives IAQA a strong philosophical basis, it also opens it up to critique.

One such critique comes from feminist philosophy. Kant very deliberately assumes that the subjective conditions for aesthetic judgment are common to all rational observers (i.e. we all have the same common sense ideas about beauty). However, feminist philosophers have observed that the supposedly universal, rational ideas advocated by Enlightenment thinkers like Kant included some ideas deeply rooted in those thinkers' worldviews, which are naturally limited by historical and cultural context. For example, Kant argued that women have a natural affinity for the beautiful and decorative while men have a natural affinity for the sublime and inspiring (Kant 1764), and Edmund Burke argued that light skin was naturally aligned with the beautiful while dark skin was closer to the sublime (Armstrong 1996). These claims are rooted in 18th century European views of race and gender and are clearly not true across space and time. To reconcile the supposed rationality and universality of their views with very real differences in perspectives held by those on the margins of society, these philosophers tended to dismiss alternative views, especially those of women and non-Europeans, as irrational or incomplete (Korsmeyer 2004), which has contributed to various forms of discrimination, including gender and racial bias in the artistic canon (Battersby 1989; Deepwell 2019). As elaborated by Korsmeyer (2004),

> Seeking to establish standards for artistic enjoyment can be seen as an attempt to regulate and homogenize pleasures according to a gauge that reflects distinct class bias, not to mention national and racial preferences. In promulgating the existence of standards for subjective pleasures, the preferences of people who were already culturally accredited, as it were, became the standards to be emulated. Ideas about taste and beauty, no matter how assiduous the attempt to universalize standards and to "purify" them of bias and prejudice, seem ineluctably to absorb reigning social values.

In other words, when people attempt to establish objective standards for subjective pleasures, no matter how objective or rational they attempt to be, those standards reflect the social values of the society that creates them.

Returning to machine learning, we can take inspiration from this philosophical debate and generate empirical research questions about personalized IAQA: how well do the average aesthetic scores from an existing dataset actually predict new individuals' judgments? And when and for whom can we accurately predict disagreement between the average scores and the individuals' judgments? The Kantian position would predict that the average scores perform similarly well for all users, and that features describing the image and labeler's interest could be used to predict disagreement, while the feminist position would predict that the average scores perform better for some users than others, but that those differences in taste are the result of differences in perspective, and cannot be inferred from specific features. Note that while this argument comes from feminist theory, we are drawing on its more theoretical side; our objective is not to investigate whether such models are sexist.

These issues are important because assumptions that we make while collecting data about image aesthetics might become self-fulfilling prophecies. In the context of image classification, Denton et al. (2021) argue that establishing benchmark datasets like ImageNet constitutes the "computational construction of meaning," where a somewhat arbitrary classification scheme ends up serving as an objective framework for interpreting the meaning of images. We worry that the data collection schemes used in IAQA may constitute the computational construction of taste. As Ferry (1993) argues, the concept of personal taste is itself an early modern invention, linked to humanism, rather than a fundamental fact of nature. We worry that subtle choices in data collection may inadvertently legitimize certain differences in aesthetic preference and delegitimize others.

To study these questions, we introduce PR-AADB, a new set of labels for a subset of the images from the AADB dataset of Kong et al. (2016). While our labels describe the same images, our dataset has several important differences: we collect pairwise labels instead of numerical scores, each user labels 20 "training" image pairs common to all users and 80 "testing" image pairs which are unique to that user, and we collect additional information about our participants including demographics and how they went about labeling. Since this is a relatively small dataset, containing labels for 16,548 image pairs drawn from 8,835 of the 9,958 images of the original AADB dataset, we see these modifications not as an improvement over the original labels, but as a means to critically evaluate the assumption of disinterestedness in IAQA and as additional testing for few-shot personalization.

We find, consistent with the feminist position, that average aesthetic quality labels are poor predictors of our participants' preferences. In addition, there is a high amount of inter-subject variance in the prediction quality, indicating that the ground truth represents some users' tastes significantly better than others. However, we do not find that demographic, style or content factors explain these disagreements. In other words, the ground truth inherently reflects some peoples' tastes better than others, but determining whose taste is not simply a matter of gender or education level, for example.

Related Work

Several authors have proposed datasets for image aesthetics analysis as an objective classification or regression problem. The early work in IAQA of Datta et al. (2006) and Ke, Tang, and Jing (2006) used relatively small datasets of images with binary quality labels. Murray, Marchesotti, and Perronnin (2012) released the Analysis of Visual Aesthetics (AVA) dataset, which contains over 250,000 photos from DPChallenge.com along with metadata such as rating distributions and category labels. This dataset continues the original framing of IAQA as a classification problem, where the ground truth label (high or low quality) is calculated from the average of many human ratings. In 2016, Kong et al.

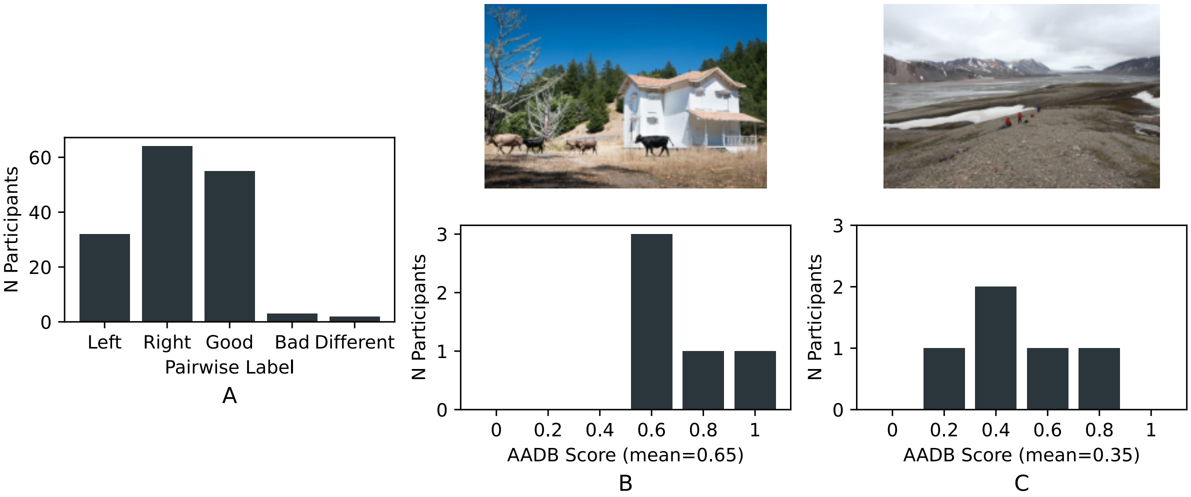


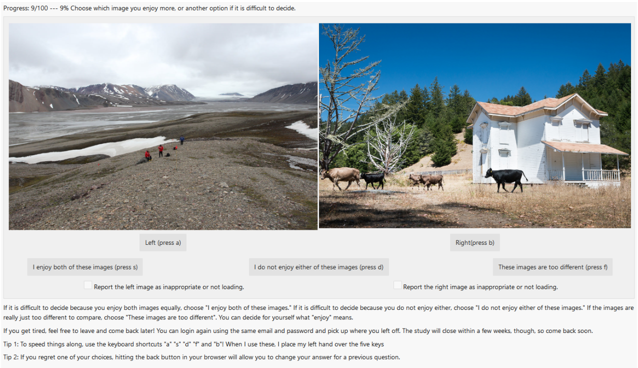
Figure 1: **Sample ratings from our dataset.** A pair of images from the AADB dataset, one of 20 “common” pairs shown to all participants in our study. *A:* the ratings from our participants using a labeling scheme with options for “both images are good,” “both images are bad” and “these images are too different to compare.” *B and C:* single-image ratings from AADB. Note that the single-image average for the left image is higher in AADB, but more of our participants preferred the right.

proposed the Aesthetic Attribute Database (AADB) which includes both overall aesthetics ratings and 11 aesthetic attributes (e.g. blur, depth of field) for each image.

Recently, others have framed IAQA as a more subjective problem. Ren et al. (2017) introduce the personalized image aesthetics task through the Flickr-AES dataset, which contains user-by-user ratings for each image. Using the larger AVA dataset (Murray, Marchesotti, and Perronnin 2012) for pretraining, Lee and Kim (2019) achieve better performance with a pairwise approach, using an eigenvector method to infer rankings from comparisons. However, prior work argues that labels from the AVA, AADB, and Flickr-AES datasets fail to capture the concept of aesthetic quality broadly, and instead capture a *specific* “aesthetic” photographic style common on photo-sharing websites (Goree 2021).

Outside of computer science, and particularly in food science, preference studies are common, and there is a rich history of debate on which sorts of preference study designs are most reliable; see (O’Mahony and Wichchukit 2017; Luckett, Burns, and Jenkinson 2020) for discussion. Böckenholt (2001) finds that when participants have difficulty appraising their own preferences, their ratings for single stimuli can be inconsistent, and advocates for designs involving pairwise preferences which allow participants to express their uncertainty.

While research through data relabeling is a relatively unusual approach, it has begun to gain traction in machine learning. Beyer et al. (2020) conduct a relabeling of the ImageNet validation set (Deng et al. 2009) to assess whether improved accuracy on ImageNet actually reflects progress on image classification. Kong, Shen, and Huang (2021) develop a more general framework for studying relabeling and its effects on model performance.


Progress: 9/100 --- 9% Choose which image you enjoy more, or another option if it is difficult to decide.

Left (press a)
Right(press b)
I enjoy both of these images (press s)
I do not enjoy either of these images (press d)
These images are too different (press f)

□ Report the left image as inappropriate or not loading.

□ Report the right image as inappropriate or not loading.

If it is difficult to decide because you enjoy both images equally, choose "I enjoy both of these images." If it is difficult to decide because you do not enjoy either, choose "I do not enjoy either of these images." If the images are really just too different to compare, choose "These images are too different". You can decide for yourself what "enjoy" means.

If you get tired, feel free to leave and come back later! You can login again using the same email and password and pick up where you left off. The study will close within a few weeks, though, so come back soon.

Tip 1: To speed things along, use the keyboard shortcuts "a" "s" "d" "f" and "b"! When I use these, I place my left hand over the five keys

Tip 2: If you regret one of your choices, hitting the back button in your browser will allow you to change your answer for a previous question.

Figure 2: Screenshot of our labeling interface.

Methods

Study and Collection Interface Design

To permit comparison with existing work, we collected new aesthetic labels for the existing AADB images (Kong et al. 2016) (instead of new images scraped from the web). We chose this dataset because of its relatively small size, thorough annotation, and prominence in the literature.

We began with a pilot study to tune our labeling protocol and interface. Most recent IAQA data collection studies (including the original AADB (Kong et al. 2016)) use Amazon Mechanical Turk (AMT), and collect aesthetic labels for individual images on a two- (Tang, Luo, and Wang 2013), five- (Ren et al. 2017), or ten-point (Kong et al. 2016) scale. However, we found that individual aesthetic quality opinions tend to lack precision: users do not have a universal point of reference for how appealing a 8/10 image would be versus a 6/10 image, for example. Instead of asking participants to label individual images, we found it better to present pairs of images and ask them to choose a preference between the two. Pairwise methods have long been used in image quality assessment (but not aesthetic quality assessment) (Mantiuk, Tomaszewska, and Mantiuk 2012); seeing images in pairs gives participants grounding because they are not evaluating an image's quality in the abstract, but instead relative to another image.

We also use a specific prompt: "Choose which image you enjoy more, or another option if it is difficult to decide," where the other options are "I enjoy both of these images," "I do not enjoy either of these images," and "These images are too different." The term "enjoy" grounds the label in the personal experience of the participant, rather than an abstract notion of aesthetic quality or beauty. This prompt contrasts with the one used for the original AADB labeling, "rate this photo w.r.t its aesthetic and select attributes to explain why this image is of high or low aesthetic." While one might argue that these prompts are measuring different qualities (i.e. there is more to aesthetic quality than just enjoyment), the term "aesthetic" is highly ambiguous. The term "enjoy" has been used to specify the sensory aspects of the aesthetic experience in several disciplines, including HCI (Cyr, Head, and Ivanov 2006), psychology of art (Millis 2001), and aesthetics of the everyday (Besson 2017), and evokes the language of philosopher John Dewey's concept of the aesthetic: "experience as appreciative, perceiving and enjoying" (Dewey 2005). Others have used prompts such as interestingness (Gygli et al. 2013; Gardezi et al. 2021), pleasing, harmonious (Geller et al. 2022) to define the aesthetic experience. Future studies could compare how different prompts could result in different responses from individuals.

We show each participant a small number of "common" image pairs, which are the same for everyone, and a larger number of "unique" image pairs, which are only shown to one participant. The common pairs provide a controlled training set for few-shot personalization. For example, future researchers could exclude specific pairs from the training set to measure their effect on the personalized model. The unique image pairs provide coverage of AADB, which allows us to both measure consistency between our participants' responses and the original labels, and to create a robust test set to evaluate few-shot personalization.

Recruitment and Data Collection

After receiving approval from our university's human subjects review board, we recruited participants through a combination of university mailing lists and social media with the following inclusion criteria: (1) At least 18 years old, (2) Located in the United States, (3) Not visually impaired. We split our data collection into two parts: a short screener survey with standard demographic questions, and a longer survey using the labeling process described in the previous section. Frequency values for demographic characteristics can be found in our supplementary materials. We provided compensation for each participant to label 20 "common" image pairs and 80 "unique" image pairs.

We took several measures to avoid unreliable participants: splitting our survey into two parts (screener and longer survey), CAPTCHA protection, free response questions and an analysis of label distributions. We filtered out hundreds of auto-generated responses to our screener survey and ended up discounting 11 responses on the longer survey which both submitted questionable free text responses and possessed unusual label distributions (e.g. a uniform distribution over the five responses). The high degree of agreement on some common image pairs (e.g. for pair 17 over 80% prefer image B while only 5% prefer image A) indicates that it is unlikely many participants are answering randomly.

Data was collected between November 10th, 2021 and January 5th, 2022. Out of 237 participants who responded to our call and were sent a survey link, 181 labeled at least one data pair and 176 completed the 100 labels required to receive payment. We included a set of three free-text questions in the middle of the survey, both to gauge our participants' reasoning and to evaluate whether each participant was answering questions in good faith. Upon manual examination of the free-text questions and response distributions, we excluded the data of 11 participants whose responses seemed to be generated by automated survey completion software, leaving 165 participants in the final released dataset. We collected labels for 16,548 pairs of images in total, sampled from the 9,958 images in AADB.

Comparing Across Label Structures

For each image pair (a, b) evaluated by a human subject we convert our five response categories into scalar pairwise labels $\{-1, 0, 1\}$, where -1 corresponds to a preference for a, 1 corresponds to b, and 0 corresponds to the three other options. Using a method similar to the one from (Lee and Kim 2019), we also find an estimated single-image labeling. This method relies on constructing a matrix L of comparisons where $L_{a,b} \in \{-1, 0, 1\}$ corresponds to the preference label, and then computing the first principal eigenvector of L. This eigenvector constitutes a spectral ranking (Vigna 2016) of the images, much like the Elo score or Pagerank. To make the scores more directly comparable to the AADB scores, we scale the resulting scores to fall between 0 and 1 by subtracting the minimum and dividing by the range.

Subsequently, we define accuracy between pairwise labels and real-valued image scores as follows. For a set of images $1, ..., n$ with real-valued scores $s_1, ..., s_n$ and pairwise labels from each participant, $L_{i,j} \in \{-1, 0, 1\}$ where L is only defined for pairs $(i, j) \in P, |P| = m < n^2$, we compute the accuracy of the scores to the labels using a thresholded indicator function,

$$\text{Acc}(S, L) = \frac{1}{m} \sum_{(i,j)\in P} \begin{cases} I(S_i - S_j < -t) & L_{i,j} = -1 \\ I(-t \leq S_i - S_j \leq t) & L_{i,j} = 0 \\ I(S_i - S_j > t) & L_{i,j} = 1, \end{cases}$$

where I has value 1 if the argument is true and 0 otherwise, and t is a threshold. In other words, if the score for image i is higher than the score for image j by at least t, we predict that the participant will choose image i, and if the difference is within the threshold, we predict that the participant will either like or dislike both images. We use a threshold $t = 0.075$, chosen *post hoc* to maximize the average accuracy of the AADB ground truth for our participant labels, the most generous possible value.

Results

Comparing PR-AADB and AADB Ground Truth

First, we evaluate the consistency between the aesthetics scores published with the original AADB dataset and the preference labels provided by our participants. Since these datasets cannot be compared directly, we first use our pairwise labels to infer image scores and evaluate their ranking correlation with the AADB labels, then we use both sets of image scores to infer "generic" pairwise labels. This kind of experiment is possible because our participants labeled the exact images from the AADB training and test sets. Finally, we also test the performance of the personalized model of Ren et al. (2017) on our labels.

Comparing Single-Image Scores: Figure 3 (left) presents the joint distribution of the single-image aesthetic scores from AADB, and the single-image scores inferred from our participants' pairwise labels using the eigenvector method. Even though both sets of scores are aggregate estimates of the aesthetic quality of the same images, their correlation is only 0.27. Importantly, their ranking correlation (Spearman's ρ) is also only 0.27, which is significantly lower than state-of-the-art model performance (Lee and Kim (2019) reports $\rho = 0.879$).

Comparing Pairwise Labels: Using the scheme described in the methods, we measure accuracy for the AADB scores as well as the scores we just inferred from our labels. Figure 3 (center) presents the joint distribution of accuracy scores on each participant. The AADB scores produce accuracy values which vary from 0.2875 to 0.575, a difference of almost 30%. For 12 of our participants, this is worse than random guessing. The scores inferred from our labels produce a similar amount of variance, ranging from 0.5 to 0.7875, and accuracy on the two is only somewhat correlated ($r = 0.30$). This suggests not that the original AADB labels are poor, but that there is no single set of real-valued aesthetic quality scores which would perform well for everyone.

Evaluating Model Performance: We also tested the deep learning-based personalized IAQA model introduced in Ren et al. (2017), which predicts a raw aesthetics score using a model trained on the Flickr-AES dataset, and then fine-tunes the prediction using a support vector machine (SVM) regressor to predict the residual between the raw and personalized aesthetic score. The SVM makes use of aesthetic attribute features (learned from the original AADB aesthetic attribute labels) and content features (from a clustering of ImageNet feature vectors) to inform its prediction. We adapt this model to predict pairwise labels, rather than aesthetic score residuals, by using an SVM classifier.

While we find that the raw predictions perform similarly to the AADB ground truth scores (42.6% accuracy vs. 42.7% accuracy), when we fit the personalized model to the 20 common image pairs for each user, we find that the average accuracy on the remaining 80 image pairs does not change significantly, but the variance greatly increases (Figure 3 right) from a standard deviation of 0.065 to 0.129. Further, the fine-tuned accuracy scores do not correlate with the original accuracy scores or the accuracy under the AADB ground truth. We speculate that the performance of a fine-tuned model depends both on whether the training images are similar to the testing images and whether the set of aesthetic and content attributes are good descriptors of an individual's taste.

These experiments indicate that while the AADB ground truth labels and our participants' judgments are somewhat correlated, there is a high degree of variance in both. If different users had been logged in to Mechanical Turk when the AADB was collected or their prompt had been phrased differently, the ground truth, and thus the algorithms which perform well, could have been radically different. By chance, we end up with a dataset that is more representative of some of our participants' preferences than others, and using few-shot learning to fine-tune a personalization model increases that variance, which might have positive or negative effects, depending on the user.

Explaining Label Disagreements

In this section, we turn to our second question: when and for whom can we accurately predict disagreement between these two sets of labels? We use logistic regression analysis to examine three possible explanatory factors: demographic differences, difference in preference for aesthetic attributes, and specific image content. Rather than use these variables as features to predict aesthetic quality directly, our target variable is whether the AADB ground truth and our participants' pairwise label will be consistent or inconsistent for each image pair (using the thresholding scheme described in the methods). As a result, we use logistic regression as a statistical analysis tool, not as a predictive machine learning model.

To describe demographic differences, we create dummy variables for demographic labels: age, gender, race, level of education, and first language (coded as either English or other). To describe formal aesthetic differences between the images in a pair, we use the absolute difference (i.e. $|r_1 - r_2|$ for ratings r_1, r_2) of the 11 AADB aesthetics ratings (e.g. color harmony rating or symmetry rating). For differences in the image content, inspired by Ren et al. (2017), we use an off-the-shelf classifier (ResNet18) to classify images using the 1000 ImageNet classes, then create 1000 binary variables where each feature is 1 if the corresponding class is within the top 3 predicted classes for either image, but not both, and 0 otherwise. While using ImageNet in this manner is potentially objectionable for treating image class predictions as a measure of image content, we use it to maintain consistency with the IAQA literature, rather than as an endorsement of the ImageNet categories. We select a subset of the 1000 content features by first removing 178 classes which are never predicted, then using LASSO (L_1 regularized) logistic regression (Tibshirani 1996) with regularization tradeoff parameter $\alpha = 0.0005$ to select relevant content variables (Fonti and Belitser 2017). With this alpha value, we select 120 of the 842 remaining classes as potentially relevant variables. Using the 24 demographic variables, 11 aesthetic attribute variables and 120 content variables, we fit an un-regularized regression model.

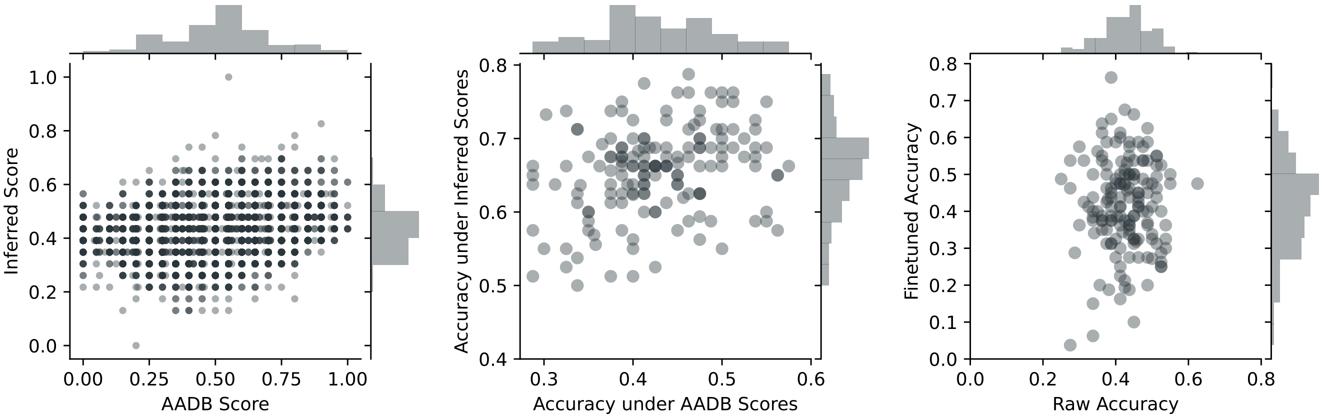


Figure 3: **Measuring agreement and variance in aesthetics scores.** *Left:* After converting our pairwise labels to inferred image scores, we plot them and measure their correlation ($r = 0.27$) vs. the original AADB labels. Data points indicate images. *Center:* Next, we convert both the AADB scores and the scores inferred from our labels to "generic" pairwise labels using the scheme described in the methods, we compare their accuracy for each participant ($r = 0.30$). Data points indicate participants. *Right:* Using the model from (Ren et al. 2017), we compute raw and personalized predictions for each image and compare their accuracy for each participant. Data points indicate participants. In each plot, points are rendered at low alpha and are represented by 'o'. Darker colors represents high density of data points in that area.

The estimated coefficients are shown in Figure 4. Our regression model is a poor predictive model with pseudo-R^2 of 0.023, i.e. our model only explains 2.3% of the inconsistency, though several of the coefficients are significant.

To our surprise, none of the demographic characteristics significantly predict consistency with the original AADB labels. While the coefficients for race show a noticeable difference between White and Asian participants and those from other racial groups, our sample, mostly drawn from a mailing list at a research university in the midwest United States, is not a representative sample of the greater population, and we hesitate to make strong claims based on a small sample.

Three of the aesthetic attributes — good content, color harmony, and good motion blur — have significant positive coefficients, which indicates that a high difference in those attributes between the two images increases the likelihood that our participants' judgments will be consistent.

Many of the content features have significant positive and negative coefficients, indicating that our participants were very sensitive to photo content. Some image classes, such as brown bear, dragonfly monarch, and limpkin, predict consistency, while window screen, rifle, mask, and military uniform predict less consistency. The number of positive coefficients associated with animals indicates that nature photographs produce consistent judgments while the negative coefficients indicate that photos containing military-related content are more controversial. We must note that these content labels were produced by an automatic classifier and not a human labeler, so they are noisy and may indicate the presence of visual patterns rather than exact objects.

Stepping back, regression analysis shows that consistency with the ground truth varies greatly from person to person, but that the differences are mostly not explained by demographics, aesthetic attributes, or visual features. Consistent with the claims of Datta et al. (2006), a few characteristics (e.g. natural subjects or color harmony) lead people to consistently find some images to be of higher aesthetic quality, however, there are other characteristics which are controversial and lead people to disagree on their quality.

Analysis of Free-Text Responses

We asked our participants two free-response questions during the survey: "How are you choosing between images?" and "Do you find yourself relying more on the content of the images (like the objects or people pictured) or the style (like whether the picture is blurry or if it is colorful)?" For the first question, we identified five categories of responses (we share representative quotes and participant ID numbers):

1. Personal preference, e.g. "Instinct" (P85), "My own personal preferences" (P105), "Would I consider them keepers" (P255)
2. Formal qualities, e.g. "content, composition, color" (P115), "Composition, centering, and lighting" (P52)
3. Content, e.g. "first impression; i think i prefer scenes over people" (P170), "My feeling. I like nature and greens. And I also like to see pictures of people having fun (not for work)" (P110)
4. Literal responses, e.g. "For some, I used a keyboard and for some, I used the mouse to select the correct options" (P7), "choosing either a or b." (P25)
5. A combination of these, e.g. "However I want, it's a study of aesthetics. I always pick the dogs, and I like colors, colors are fun" (P86), "Im gravitating towards elements i like in photography such as architecture, landscapes, and animals, if none of these are present I will tend to choose the picture that seems more visually interesting/intentionally composed." (P97)

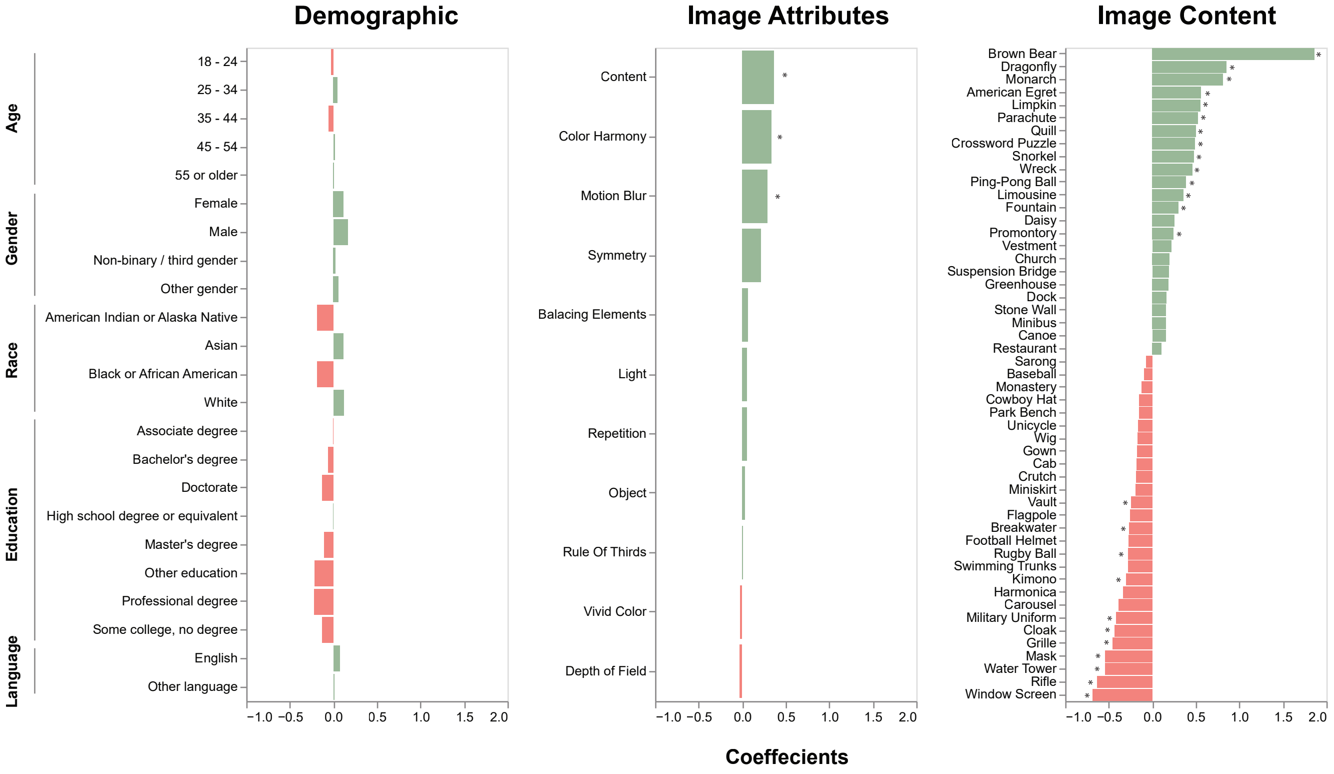


Figure 4: **Which properties of participants/images predict (dis)agreement between the AADB ground truth and our participants' ratings?** Regression coefficients for (left) demographic, (center) aesthetic, and (right) content indicate how much of the variance in consistency is explained by each attribute of the image or rater. For binary variables, the number of image pairs for which that variable is true are shown. Stars indicate coefficients for which we reject the null hypothesis at $p = 0.05$. A full regression table is included in our supplementary materials.

For the second question, we roughly grouped these responses into four categories, "content" (N=59), "style" (N=21), "both" (N=73), and "unclear" (N=3).

Taken together, these results indicate that our choice of prompt decoupled our participants' concept of aesthetic quality from a specific visual style. It also highlights the wide range of possible interpretations of words like "enjoy" and "aesthetics" which subtly change the concept under study (even though we did not use the term "aesthetics," it was often mentioned by participants). For example, given the pair of images in Figure 1, we can imagine one participant choosing the image on the left because they love cows while another participant chooses the image on the right because the stark landscape gave them with a feeling of awe, and both are valid responses, given their interpretations of the prompt and equally valid forms of aesthetic judgment.

Discussion

To summarize our results, inspired by an argument from feminist aesthetics, we collected new labels and conducted a statistical analysis of differences between the AADB labels of Kong et al. (2016) and our relabeling. We find that this critique largely holds: while the original labels are usually better than random guessing, their predictive quality varies greatly from person to person, and few-shot personalization only increases that variance.

Next, we asked if demographic, aesthetic, or content attributes could predict whether the AADB groundtruth will be consistent with a participant's preference for a given image pair. We find that these factors explain only a small amount of the variance in consistency, but there are specific aesthetic and content features, like a brown bear in one image or a difference in level of motion blur, which are informative. That means we do not find that the label disagreements are easily explained by demographic factors like gender or education level.

Our goal here is not to criticize the original AADB dataset (Kong et al. 2016) or the personalization model we used (Ren et al. 2017). We are using a different study design, with a different prompt, so we would not expect the original image scores to perfectly predict our results. Our data is also not a strict improvement on the original labels, which exist to show the relationship between ratings for overall aesthetics and aesthetic attributes, which we did not investigate.

Instead, we believe that our data and analysis show the profound difficulty of making personalized aesthetic quality predictions using machine learning. In the non-personalized formulation of the task, the prediction target is an objective kind of aesthetic quality based on popular consensus using a large sample size (Murray, Marchesotti, and Perronnin 2012), which smooths out the variance in individual interpretations to create a stable machine learning problem. However, by doing so, it ignores so many of the interesting and meaningful psycho-social phenomena which give aesthetics its depth. But accounting for subjectivity is not a matter of estimating a predictable deviation from an objective, average viewpoint.

In this way, IAQA mirrors other scientific problems. For example, we have simple physical laws which explain the behavior of a magnet, but if we try to infer the magnetic moments of its constituent atoms, the problem becomes significantly more complex, and there is no simple adjustment to the macro-level laws (i.e. average aesthetics assessments) which predicts the micro-level behavior (i.e. individual preferences for images). The analogy only goes so far, however, since we do not believe there are necessarily scientific laws which predict human aesthetic judgment.

Approaching personalization through few-shot learning results in a problem with almost-unmanageable amounts of variance, both between individual images and between users due to unobservable subjective factors. As the free-text responses show, knowing which factors are important to a participant requires knowing how they interpreted the prompt in addition to their specific aesthetic preferences, and it is unclear whether those degrees of freedom can be captured in a small number of ratings. Further, the high degree of inter-subject variance makes testing personalization algorithms difficult. Evaluating models by their accuracy (or ranking correlation, etc.) on a test set assumes that the test data is identically distributed to data from the real world, and if our labelers are not representative of some real world population (where representation is a matter of interpretive perspective and taste, not just demographics), we run the risk that our test accuracy ceases to be meaningful.

Thus, we recommend against evaluating personalized models based on their performance on a benchmark such as AADB (Kong et al. 2016) or FlickrAES (Ren et al. 2017). Such evaluations will vary tremendously based on the surveyed individuals and a model which is able to account for the differences in perspective present in such a dataset will not necessarily be able to account for the myriad of factors which affect preferences held by humans in general, and may be ill-suited to the kinds of subjective differences in another population. We encourage future work to investigate evaluating IAQA algorithms through user studies of specific populations, without the goal of producing general models of aesthetic preference. We also encourage future investigation into the potential downstream social consequences of predicting aesthetic preferences in, for example, social media contexts.

Conclusion

In summary, recent models for aesthetic quality assessment have investigated a problem formulation based on few-shot personalization, which has its roots in Kantian ideas about subjective difference due to interest. These ideas have been subject to criticism, as attempts to establish objective formulations for subjective taste tend to unintentionally reflect dominant cultural standards. Empirically, this issue manifests in IAQA when we attempt to adjust from the objective formulation, based on an average of several users' preferences, to the subjective formulation, based on individual preferences: the averages tend to predict the judgments of some individuals significantly better than others. We also have released our data and the source code for our labeling interface to assist other authors in performing studies of specific user populations, rather than comparing against less population-specific existing published benchmark datasets.

More generally, this finding indicates that the subjectivity of the problem creates a high degree of variance when sampling aesthetic preference data which makes benchmark-based evaluation unreliable. Accounting for subjective difference in machine learning problems appears to be a very difficult task, as many factors affect human perspectives, and deploying algorithmic methods based on seemingly-reasonable assumptions may constitute the computational construction of taste, enforcing reductive models of subjective difference and changing the underlying concepts under study. More study of subjectivity in machine learning and its evaluation is needed.

Acknowledgments

This work was supported in part by the National Science Foundation (through a Graduate Research Fellowship), Google, the U.S. Navy (N00174-19-1-0010) and the U.S. Department of Defense (W52P1J-22-9-3009). The views and conclusions contained in this document are those of the authors and should not be interpreted as those of any of the sponsors.

References

Armstrong, M. 1996. "The Effects of Blackness": Gender, Race, and the Sublime in Aesthetic Theories of Burke and Kant. *The Journal of Aesthetics and Art Criticism*, 54(3): 213–236.

Battersby, C. 1989. *Gender and genius: Towards a feminist aesthetics*. Indiana University Press.

Besson, A. 2017. Everyday aesthetics on staycation as a pathway to restoration. *International Journal of Humanities and Cultural Studies*, 4.

Beyer, L.; Hénaff, O. J.; Kolesnikov, A.; Zhai, X.; and Oord, A. v. d. 2020. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*.

Böckenholt, U. 2001. Thresholds and intransitivities in pairwise judgments: A multilevel analysis. *Journal of Educational and Behavioral Statistics*, 26(3): 269–282.

Cui, C.; Fang, H.; Deng, X.; Nie, X.; Dai, H.; and Yin, Y. 2017. Distribution-oriented aesthetics assessment for image search. In *ACM SIGIR RDIR*, 1013–1016.

Cui, C.; Yang, W.; Shi, C.; Wang, M.; Nie, X.; and Yin, Y. 2020. Personalized image quality assessment with Social-Sensed aesthetic preference. *Information Sciences*, 512: 780–794.

Cyr, D.; Head, M.; and Ivanov, A. 2006. Design aesthetics leading to m-loyalty in mobile commerce. *Information & management*, 43(8): 950–963.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2006. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 288–301. Springer.

Datta, R.; Li, J.; and Wang, J. Z. 2008. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*, 105–108. IEEE.

Deepwell, K. 2019. Beauty and its shadow: a feminist critique of disinterestedness. *Feminist Aesthetics and Philosophy of Art: Critical Visions, Creative Engagements. Nueva York: Springer Netherlands (Ed. original: 2011).*

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Denton, E.; Hanna, A.; Amironesei, R.; Smart, A.; and Nicole, H. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2): 20539517211035955.

Dewey, J. 2005. *Art as experience*. Penguin.

Fang, H.; Cui, C.; Deng, X.; Nie, X.; Jian, M.; and Yin, Y. 2018. Image aesthetic distribution prediction with fully convolutional network. In *ICMM*, 267–278. Springer.

Ferry, L. 1993. *Homo aestheticus: the invention of taste in the democratic age*. University of Chicago Press.

Fonti, V.; and Belitser, E. 2017. Feature selection using lasso. *VU Amsterdam research paper in business analytics*, 30: 1–25.

Gardezi, M.; Fung, K. H.; Baig, U. M.; Ismail, M.; Kadosh, O.; Bonneh, Y. S.; and Sheth, B. R. 2021. What Makes an Image Interesting and How Can We Explain It. *Frontiers in Psychology*, 12.

Geller, H. A.; Bartho, R.; Thömmes, K.; and Redies, C. 2022. Statistical image properties predict aesthetic ratings in abstract paintings created by neural style transfer. *Frontiers in Neuroscience*, 16.

Goree, S. 2021. What does it take to cross the aesthetic gap? the development of image aesthetic quality assessment in computer vision. In *Proceedings of the 12th International Conference on Computational Creativity*.

Gygli, M.; Grabner, H.; Riemenschneider, H.; Nater, F.; and Gool, L. V. 2013. The Interestingness of Images. In *2013 IEEE International Conference on Computer Vision*, 1633–1640.

Joshi, D.; Datta, R.; Fedorovskaya, E.; Luong, Q.-T.; Wang, J. Z.; Li, J.; and Luo, J. 2011. Aesthetics and emotions in images. *IEEE Signal Processing Magazine*, 28(5): 94–115.

Kairanbay, M.; See, J.; and Wong, L.-K. 2018. Towards Demographic-Based Photographic Aesthetics Prediction for Portraitures. In *International Conference on Multimedia Modeling*, 531–543. Springer.

Kairanbay, M.; See, J.; and Wong, L.-K. 2019. Beauty is in the eye of the beholder: Demographically oriented analysis of aesthetics in photographs. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2s): 1–21.

Kant, I. 1764. *Observations on the Feeling of the Beautiful and Sublime*. Univ of California Press.

Kant, I. 1790. Critique of Judgment. In Ross, S. D., ed., *Art and its Significance: An Anthology of Aesthetic Theory*. SUNY Press.

Ke, Y.; Tang, X.; and Jing, F. 2006. The design of high-level features for photo quality assessment. In *CVPR*, volume 1, 419–426. IEEE.

Kong, S.; Shen, X.; Lin, Z.; Mech, R.; and Fowlkes, C. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 662–679. Springer.

Kong, S.; Shen, Y.; and Huang, L. 2021. Resolving Training Biases via Influence-based Data Relabeling. In *International Conference on Learning Representations*.

Korsmeyer, C. 2004. *Gender and aesthetics: An introduction*. Routledge.

Lee, J.-T.; and Kim, C.-S. 2019. Image Aesthetic Assessment Based on Pairwise Comparison A Unified Approach to Score Regression, Binary Classification, and Personalization. In *ICCV*, 1191–1200.

Luckett, C. R.; Burns, S. L.; and Jenkinson, L. 2020. Estimates of relative acceptability from paired preference tests. *Journal of Sensory Studies*, 35(5): e12593.

Lv, H.; and Tian, X. 2016. Learning relative aesthetic quality with a pairwise approach. In *ICMM*, 493–504. Springer.

Mantiuk, R. K.; Tomaszewska, A.; and Mantiuk, R. 2012. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*, volume 31, 2478–2491. Wiley Online Library.

Millis, K. 2001. Making meaning brings pleasure: the influence of titles on aesthetic experiences. *Emotion*, 1(3): 320.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2408–2415. IEEE.

O'Mahony, M.; and Wichchukit, S. 2017. The evolution of paired preference tests from forced choice to the use of 'No Preference' options, from preference frequencies to d' values, from placebo pairs to signal detection. *Trends in Food Science & Technology*, 66: 146–152.

Pluhar, W. S. 1987. *Translator's Introduction to Kant's Critique of Judgment*. Indianap. Hackett.

Ren, J.; Shen, X.; Lin, Z.; Mech, R.; and Foran, D. J. 2017. Personalized image aesthetics. In *ICCV*, 638–647.

Tang, X.; Luo, W.; and Wang, X. 2013. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8): 1930–1943.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

Vigna, S. 2016. Spectral ranking. *Network Science*, 4(4): 433–445.

Zhu, H.; Zhou, Y.; Li, L.; Li, Y.; and Guo, Y. 2021. Learning Personalized Image Aesthetics from Subjective and Objective Attributes. *IEEE Transactions on Multimedia*.