Aesthetic Evaluation of Ambiguous Imagery

Xi Wang
ETH Zürich
xi.wang@inf.ehz.ch

Zoya Bylinskii, Aaron Hertzmann
Adobe Research
bylinski@adobe.com,hertzman@dgp.toronto.edu

Robert Pepperell
Fovolab/Cardiff Met. Univ., UK
rpepperell@cardiffmet.ac.uk

In this work, we evaluated aesthetic preferences for ambiguous artwork, using GAN images. Do people prefer more-ambiguous imagery? We find that the answers depended heavily on the way that we performed the experiments. Our findings offer important lessons for anyone conducting aesthetic image evaluations, as well as questions for future research into how to perform such evaluations.

Ambiguity is an important feature of creative practice and its effects have been exploited frequently by Modernist and Contemporary writers and artists [6, 17]. Yet, in the history of art, highly ambiguous works have often proven controversial, being both the subject of public ridicule, while also being highly valued by art specialists and collectors. For example, the works of J. M. W. Turner, now revered as one of Britain's greatest and most popular painters, were regarded by many of his contemporaries as indecipherable noise.

In previous work, we developed an ambiguity measure for images, defined as the entropy of the noun histogram from crowdworkers' text descriptions of an image [21]. Measuring entropy for different viewing durations gives multiple entropies for an image. We demonstrated qualitative examples on GAN-generated images from among the most-popular ArtBreeder [19] images (Figure 1), showing ways that this measure could organize images by different kinds of ambiguity and indeterminacy. The next step was to see whether our ambiguity measure was somehow related to image aesthetics.

We hypothesized that our entropy measure would predict aesthetic preference. We expected that ambiguous GAN images—the types of images typically used in GAN artwork [8]—would be preferred over realistic images of mountains and cats.

1 Data collection and initial results

We first set up an Amazon Mechanical Turk (MTurk) Rating task, in the same manner as many previous aesthetic evaluation experiments, e.g., [9, 14, 18]. Using the same GAN image dataset as before [21], we asked MTurk raters to evaluate each image on a Likert 1-5 scale. We averaged these ratings to obtain per-image scores. To capture different kinds of image appreciation, we phrased the question in several different ways in different versions of the experiment, e.g., asking for which images are most "interesting", "powerful", or "engaging". To our surprise, the recognizable images were, on average, rated higher than the abstract and indeterminate images, across all three wordings.

We imagined several possible explanations for these results. We hypothesized that the crowdworkers were speeding through the task, and that forcing them to spend more time engaging with the images would improve appreciation, given that one's appreciation for ambiguous imagery changes over viewing duration [13]. We also imagined that framing the task in an art appreciation context might change their responses, e.g., the responses to "What image do you like" might be different from the responses to "What image should be in an art gallery?" Some previous studies found gallery tasks to be effective [7, 20], and also found value in asking participants to give titles to the works.

Hence, we formulated a Gallery Curation task in which MTurk crowdworkers were asked to select 5 images out of 50 for a hypothetical gallery show; we also required them to view each image for at least two seconds before providing each rating. Again, the results showed a clear preference for unambiguous artworks.

Submitted to 2021 NeurIPS Workshop on Creativity for Learning and Design

We then formulated a Ranking task, in which raters were asked to rank groups of 5 images from most- to least-favorite. Aggregating these results, the most unambiguous images were again preferred.

2 Clustering analysis

We hypothesized that, by averaging across all raters, we were ignoring variations in individual preferences. As an initial test, we divided up the raters in each test according to their average scores for a portion of our dataset that we had previously labeled as "AbstractFlat": images that are highly abstract patterns of color. For the first two tasks, no statistically-significant clusters emerged, whereas two clusters do appear in the Ranking task: on a 1-5 scale, abstract images were ranked on average 1.89 by participants in one cluster, and 3.57 by the other.

Encouraged by this outcome, we sought to build a more precise clustering model. We formulated a clustering version of the algorithm in [16], that jointly assigns raters to clusters while also estimating a per-cluster average ranking for each image. In this model, the dominant cluster of raters prefers recognizable images, but a significant minority prefers ambiguous imagery.

To check for overfitting, we tested this model against a single-cluster model on predicting new scores on hold-out data, and indeed found that the two-cluster model is more predictive for the ranking task, thereby validating the clustering. We also found that adding clusters did not improve predictions.

3 Discussion

In our experiments, we find that our entropy measure can predict aesthetic rating, once we cluster individual preferences. Note that we do not expect these features to outperform state-of-the-art predictors (e.g., [18]). Rather, our goal has been to study the role of ambiguity in aesthetics preference.

These results also show that **typical aesthetic evaluation methodologies used in graphics and vision can "fail" (or mislead) when there are significant variations in preferences, such as for ambiguous artworks.**

We find that many raters preferred ambiguous imagery, but they were in the minority. Previous research has found significant individual preferences in aesthetics [1, 2, 3, 4, 5, 11, 12, 22], and also used preferences in aesthetic prediction modeling [10, 15, 18]. And, of course, aggregate ratings from MTurk may be sufficient for many downstream tasks, but our results suggest that they are a poor measure of artistic qualities. That is, the kinds of images often shown as artwork, including GAN artworks, would generally perform poorly in these studies.

Moreover, **the formulation of the MTurk task changed the outcome of the experiment.** There are several differences between the tasks that could have caused this; we believe that the Ranking task yielded different results from the others because it forced the raters to spend more time in contemplation of each image. It may be in particular that **ambiguous images require longer viewing time for appreciation** [13].

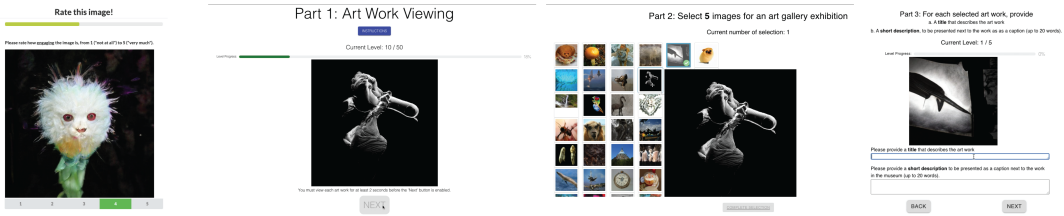
There are many possible causes and factors in these outcomes, including the task formulations, the backgrounds and demographics of the raters, the MTurk incentives, and the raters' individual preferences.

This points to the need for further research on how to perform and interpret aesthetic evaluation studies. What does it mean to evaluate aesthetic preferences, when our preferences can depend on so many factors and which can change as a function of time, experience, and context? If different experiments give different results, how do we know which results are more meaningful? When are the outcomes reliable and replicable, rather than being the artifact of a study design, and when do study results translate to real-world effects?

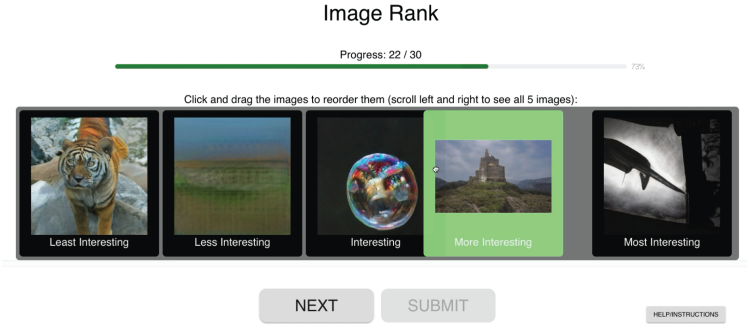
Until we better understand these complex issues, we recommend that research works based solely on MTurk evaluation of image outputs should take caution when interpreting data about aesthetic preferences.

Recognizable Dichotomous Indeterminate Abstract AbstractFlat

Figure 1: **Examples of the five categories from the Artbreeder stimuli set, from [21].** The set of 150 images used in our previous study were loosely categorised in five categories, which are “Recognizable”, “Dichotomous”, “Indeterminate”, "Abstract” and “AbstractFlat”.



(a) Rating task (b) Gallery Curation task



(c) Ranking task

Figure 2: **Aesthetic evaluations of the ArtBreeder stimuli set. a.** The simple Rating task. Participants were asked how interesting/powerful/engaging a stimuli is on a five-point Likert rating scale from 1 to 5. **b.** The Gallery Curation task where participants first view a set of images (part 1) and select five of them for a gallery show (part 2). At the end, they are asked to provide a title and description for the selected candidates (part 3). **c.** In the Ranking experiment, five images are presented in one trial, and participants are asked to drag them such that they are rearranged to a ranking based on their interestingness.

References

[1] Y. Bao, T. Yang, X. Lin, Y. Fang, Y. Wang, E. Pöppel, and Q. Lei. Aesthetic preferences for eastern and western traditional visual art: Identity matters. *Frontiers in Psychology*, 7:1596, 2016.

[2] N. J. Bullot and R. Reber. The artful mind meets art history: Toward a psycho-historical framework for the science of art appreciation. *Behavioral and brain sciences*, 36(2):123–137, 2013.

[3] I. L. Child, M. Cooperman, and H. M. Wolowitz. Esthetic preference and other correlates of active versus passive food preference. *Journal of Personality and Social Psychology*, 11(1):75, 1969.

[4] E. B. Coleman. Aesthetics as a cross-cultural concept. *Literature & Aesthetics*, 15(1), 2011.

[5] H. Eysenck. Personality factors and preference judgments. *Nature*, 148(3751):346–346, 1941.

Figure 3: **The most selected images in the Gallery Cration task in two groups.** Participants of the Gallery Curation task were divided in two groups based on their preference towards “AbstractFlat" images. The top row shows the most-selected images among those that do not prefer abstract flat images, and the bottom row are the images from those that do. The corresponding selection rates in percentage are shown below the images.

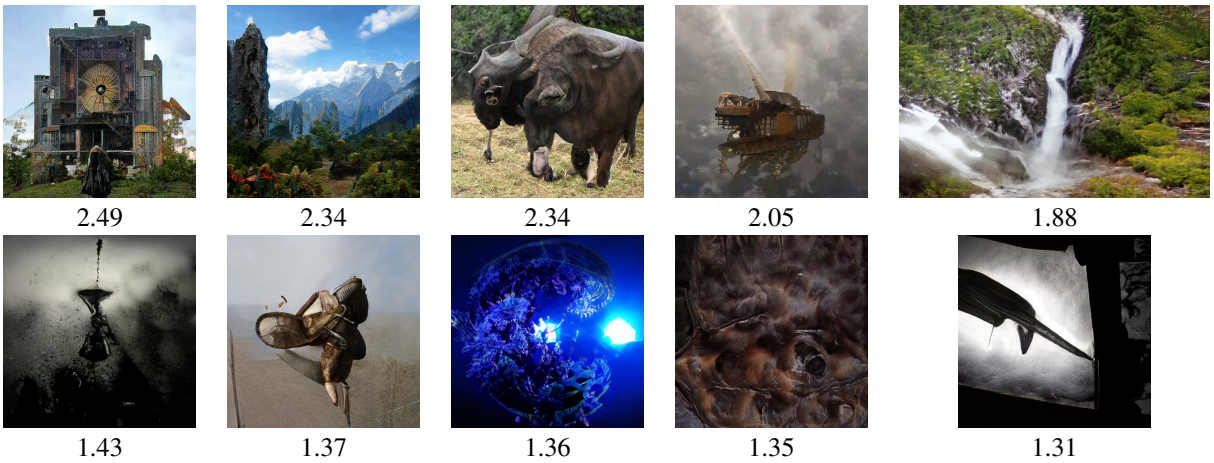


Figure 4: **The top-5 highest rank images of the two groups of the Ranking task.** Participants of the ranking task were divided in two groups based on their preference towards “AbstractFlat" images, and the top-5 highest ranked images in each group are shown here. The corresponding ranks measured by binary loss are shown below the images.

[6] D. Gamboni. *Potential images: Ambiguity and indeterminacy in modern art*. Reaktion Books, 2002.

[7] L. K. Graf and J. R. Landwehr. Aesthetic pleasure versus aesthetic interest: The two routes to aesthetic liking. *Frontiers in Psychology*, 8:15, 2017.

[8] A. Hertzmann. Visual indeterminacy in gan art. *Leonardo*, 53(4), 2020.

[9] K. Iigaya, S. Yi, I. A. Wahle, K. Tanwisuth, and J. P. O’Doherty. Aesthetic preference for art can be predicted from a mixture of low-and high-level visual features. *Nature Human Behaviour*, pages 1–13, 2021.

[10] A. Kapoor, J. C. Caicedo, D. Lischinski, and S. B. Kang. Collaborative personalization of image enhancement. *International journal of computer vision*, 108(1-2):148–164, 2014.

[11] P. Machotka. Esthetic judgment warm and cool: Cognitive and affective determinants. *Journal of Personality and Social Psychology*, 42(1):100, 1982.

[12] H. J. McWhinnie. A review of research on aesthetic measure. *Acta Psychologica*, 28:363–375, 1968.

[13] C. Muth, M. H. Raab, and C.-C. Carbon. Semantic stability is more pleasurable in unstable episodic contexts. on the relevance of perceptual challenge in art appreciation. *Frontiers in Human Neuroscience*, 10:43, 2016.

[14] P. O'Donovan, A. Agarwala, and A. Hertzmann. Color compatibility from large datasets. *ACM Transactions on Graphics*, 30(43), 2011.

[15] P. O'Donovan, A. Agarwala, and A. Hertzmann. Collaborative Filtering of Color Aesthetics. In *Proc. Computational Aesthetics (CAe)*, 2014.

[16] P. O'Donovan, J. Lībeks, A. Agarwala, and A. Hertzmann. Exploratory font selection using crowdsourced attributes. *ACM Trans. Graph.*, 33(4), July 2014.

[17] R. Pepperell. Seeing without objects: Visual indeterminacy and art. *Leonardo*, 39(5):394–400, 2006.

[18] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran. Personalized image aesthetics. In *Proc. ICCV*, 2017.

[19] J. Simon. Artbreeder (website), 2019. `artbreeder.com`.

[20] C. C. Steciuch, R. D. Kopatich, D. P. Feller, A. M. Durik, and K. Millis. Don't go with your gut: Exploring the role of motivation in aesthetic experiences. *Psychology of Aesthetics, Creativity, and the Arts*, 15(1):28, 2021.

[21] X. Wang, Z. Bylinskii, A. Hertzmann, and R. Pepperell. Toward quantifying ambiguities in artistic images. *ACM Trans. Appl. Percept.*, 17(4), Nov. 2020.

[22] T. Yang, S. Silveira, A. Formuli, M. Paolini, E. Pöppel, T. Sander, and Y. Bao. Aesthetic experiences across cultures: Neural correlates when viewing traditional eastern or western landscape paintings. *Frontiers in Psychology*, 10:798, 2019.