TOWARDS A THEORY OF EVALUATION FOR AESTHETIC PHENOMENON PROBLEMS IN COMPUTER VISION

Samuel Philip Goree

Submitted to the faculty of the Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics, Computing, and Engineering,
Indiana University

July 2023

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

David J. Crandall, Ph.D., Chair

Norman M. Su, Ph.D.

Christopher Raphael, Ph.D.

Kalani Craig, Ph.D.

Justin N. Wood, Ph.D.

Date of Defense:

July 5th 2023

ACKNOWLEDGMENTS

First, I thank my parents, Pam and Kimo, my brother Kai, my grandparents Arlene and Marvin and all the Chasek, Horowitz, Schulze and Simcox family members who have been cheering me on through this process.

Thank you so much Emma, for moving to Indiana, for every time you've called out my nonsense and every time you've told me to stop working during every unnecessary late night. I couldn't have done any of this without you.

Similarly, none of this would have been possible without the trust and support of my advisor, David. From our first meeting about a GRFP proposal to my defense, you have given me the benefit of the doubt and let me pursue every strange research direction I thought of, no matter how unrelated, and advocated for me and my work at every turn.

Additionally, I thank my other research and teaching mentors, Norman and J, and my committee members Chris, Kalani and Justin, for their guidance over the past five years, my coauthors, Bardia, Weslie and Gabe, for their contributions to this thesis, and all the other friends who have commiserated with me on Discord (or even in real life!) over the past years.

Finally, this thesis is based on work that was funded in part by the National Science Foundation through Graduate Research Fellowship, a CAREER award (IIS-1253549) and grant 2112635, 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.

Samuel Philip Goree

TOWARDS A THEORY OF EVALUATION FOR AESTHETIC PHENOMENON PROBLEMS IN COMPUTER VISION

When working with large cultural image datasets, many research topics require computational operationalization of subjective aesthetic phenomena, such as beauty, colorfulness or similarity. In this thesis, we argue that algorithms for these problems cannot be objectively evaluated. Instead, these problems demand more contextually situated and positional forms of evaluation. To that end, we present motivating studies in three different domains where aesthetic phenomenon problems occur: web design history, empirical color harmony and art historical periodization. In each of these areas, we demonstrate strategies for using statistical uncertainty and qualitative information to inform our computational approaches and validate their findings. Then, drawing on ideas from feminist theory, we engage in an extended critical exploration of algorithms for image aesthetic quality assessment. We show how quantitative performance metrics are insufficient for this problem, especially in personalized contexts. As an alternative, we prototype and pilot a method for qualitative evaluation in the context of smartphone photography.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivating Example: The Colorful Paintings of Mark Rothko

The American painter Mark Rothko is known for, among other things, a dramatic change in the style of his paintings mid-way through his career. Initially grounded in a surrealist style, he started to paint large color fields, which grew darker over time as he slid into depression [49]. Rothko's work is a popular test area for computational analysis of images specifically because this shift is so visual [326]. So how might we identify this shift automatically from images of his paintings?

Figure 1.1 shows several Mark Rothko paintings, plotted by year on the X axis and a measure of colorfulness on the Y axis. The five-year running average colorfulness is shown in black, while the output of an automatic periodization method, which we will describe and improve on in Chapter 5, is shown in blue. Immediately, we see the same story about Rothko's art presented visually: his early surrealist works change to color field paintings, which increasingly have low colorfulness values over time.

By this point, any art historian is likely (and rightly) recoiling. Treating images of paintings as data and visualizing them, even plotting trends over time, is disrespectful and commodifying — evoking the visual language of stock prices or productivity measures. Colorfulness is a complex, subjective human experience. Specific colors have different vi-

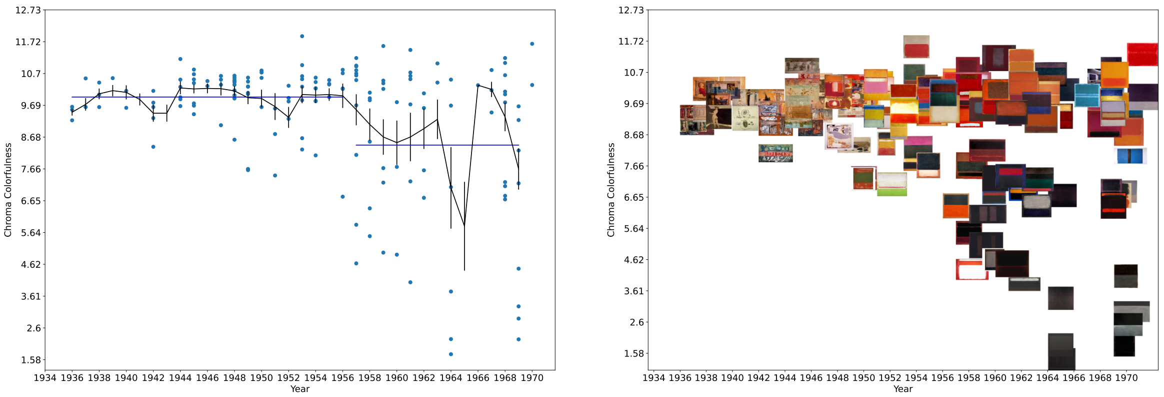


Figure 1.1: Mark Rothko paintings on WikiArt, plotted according to a measure of colorfulness ("chroma colorfulness") over time. *Left:* A scatterplot showing colorfulness values over time, with running mean and standard error bars in black and automatic periodization in blue. *Right:* The images that the scattered points represent, arranged similarly.

sual effects, emotional associations and cultural meanings that do not easily reduce to a scalar value, and arguably cannot be quantified. Today, however, the boundary between what can and cannot be quantified is more permeable than ever: human visual culture is increasingly digitized and decentralized [275], deep neural network generative models can produce seemingly expressive artistic images [266, 221] and computer vision researchers have developed methods for automatically interpreting the aesthetic properties of images, like the emotional impact of photographs [76], the color semantics of magazine covers [163, Ch. 3] or the complexity of a painting [181]. We claim that topics at this boundary are worthwhile to study, both through algorithmic modeling and critical engagement, not because measuring art is a good thing to do on its own, but because these methods can help us see works of art, and aesthetic phenomena like colorfulness, differently.

Any computer scientist likely had the opposite reaction: why bother spending time and compute resources applying computational methods to mature existing disciplines, like art and aesthetics, which do not have urgent problems in need of computational solutions? While this argument is also reasonable, we would argue in favor of the study of topics without urgent need for practical solutions because they give us room to consider our methods in full depth without the pressure to take shortcuts. Even if the methods we develop are ultimately not very useful for art historians (or another application area), studying impractical problems around phenomena like colorfulness creates a site for critical engagement which can teach us both about visual experience and about our research methods for less subjective problems. Given current crises regarding bias and injustice in computer vision [38, 317, 116, 54, 273, 85, 255], we believe engagement with these problems in more abstract domains like these is needed to explore longer-term solutions.

The critical engagement we are interested in occurs not in the modeling itself but in the problem formulation and evaluation: the way we describe the problem mathematically and determine how well a computational solution actually solves it. These components, especially evaluation mechanisms, are more infrastructural [44], making them more relevant to humanistic inquiry than the specific technical methods. To illustrate the connection, let us return to our Rothko example. The specific measure of colorfulness used for the Y axis values in this plot is a very simple metric: the average chroma of each image.

Formally, for an image in CIELa*b* color space (which we discuss further in Chapter 3), $I(x, y) = (L_{x,y}, a_{x,y}, b_{x,y})$ for $0 < x < W, 0 < y < H$, the chroma colorfulness is defined as follows:

$$C_{chroma}(I) = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \sqrt{a_{x,y}^2 + b_{x,y}^2} \tag{1.1}$$

While that is an easily measurable quantity, **does it measure colorfulness?** Imagine an observer, Alice, looks at this chart and notices that many of the high-chroma pixels from the early images actually correspond to the color of the canvas, rather than to the paint. She argues that this is an error: the early color field paintings, with large regions of vibrant color, should be much more colorful than the realist paintings that preceded them.

If our objective is to choose an elegant measure for arranging the paintings and call it colorfulness only as a shorthand, an inconsistency with human perception is not an issue — we only have to demonstrate the computational elegance of this metric. However, if we want a measure which correctly aligns with some notion of human-perceived "colorfulness," these sorts of disagreements become problematic.

Now, imagine we consult the research literature and find Hasler and Suesstrunk's "Measuring Colourfulness in Natural Images" [144] and take their recommended colorfulness measure, "Hasler-Suesstrunk colorfulness," defined as follows:

$$\begin{aligned} C_{Hasler}(I) &= \sigma_{ab} + 0.38\mu_{ab} \qquad (1.2) \\ \sigma_{ab} &= \sqrt{\sigma_a^2 + \sigma_b^2} \\ \mu_{ab} &= \sqrt{\mu_a^2 + \mu_b^2} \end{aligned}$$

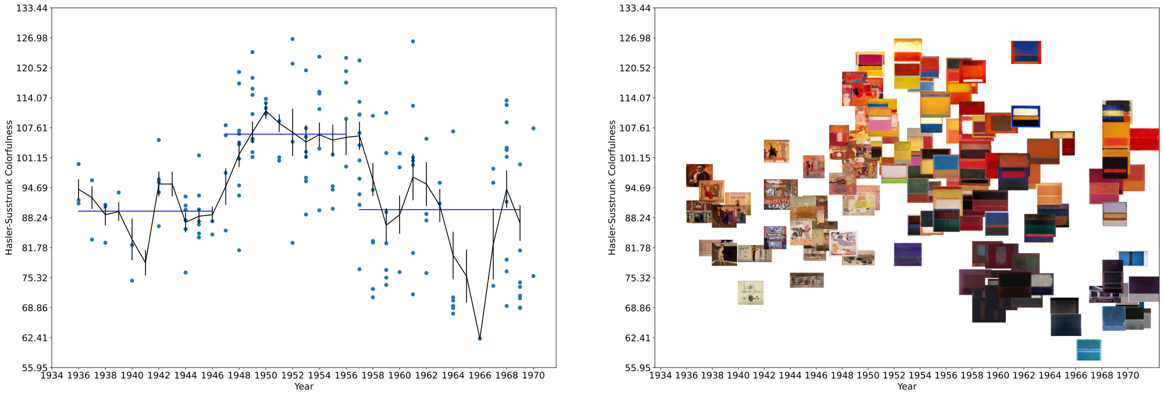


Figure 1.2: The same Mark Rothko paintings on WikiArt from Figure 1.1, plotted according to a different measure of colorfulness ("Hasler-Suesstrunk colorfulness") over time.

Where μ_a, μ_b are the channel-wise means along the a and b channels, respectively, and σ_a, σ_b are the channel-wise standard deviations along the a and b channels, respectively. They find that this metric is computationally simple, but achieves a high correlation ($r = 0.937$) with empirical measurements of human perceived colorfulness. Figure 1.2 shows the same plot as Figure 1.1, except using Hasler-Suesstrunk colorfulness.

This visualization tells a completely different story than the one in Figure 1.1. Instead of finding a two-segment periodization, it finds a three-segment one which separates Rothko's early surrealist works from his color field paintings. Interestingly, it also reveals nuances which aren't easily expressed in our short narrative: Rothko's work actually became much more colorful and abstract a few years before his first color field painting, and his final years included both some of his darkest paintings as well as some of his most colorful.

However, that does not necessarily make this function a correct measure of colorfulness. Imagine another observer, Bob, looks at this chart and points out *Red* (1968), which has a

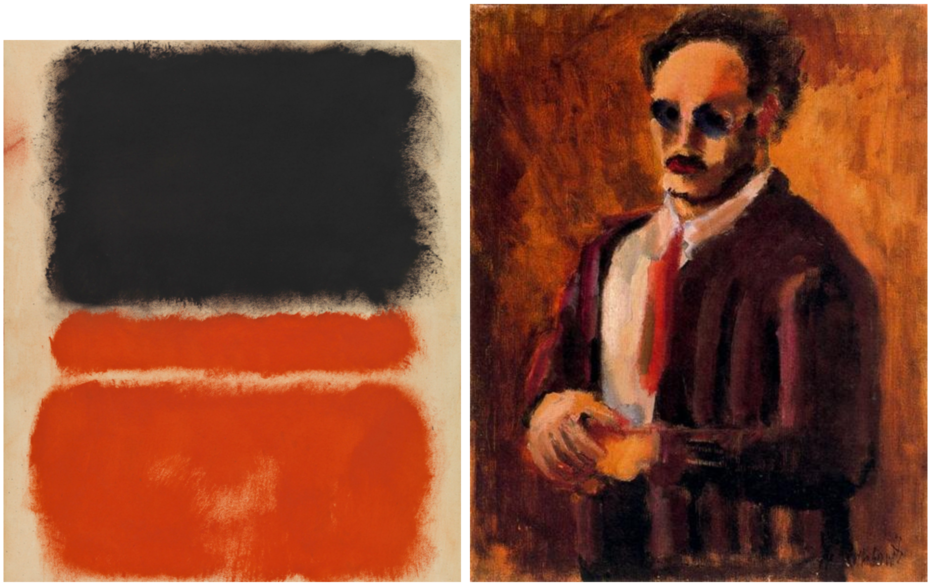


Figure 1.3: *Left:* Mark Rothko, *Untitled (Red)*, 1968. *Right:* Mark Rothko, *Self-Portrait*, 1936. Both images from WikiArt. Which of these paintings is more colorful?

colorfulness value over 110, is not nearly as colorful as Rothko's *Self-Portrait* (1936), but has a colorfulness score less than 100 (shown in Figure 1.3). He argues that colorfulness is really a contextual property, and depends on things like style and medium. The self-portrait, which mixes its pigments to form many different colors, is more colorful than the close-to-monochromatic hues of the color field painting.

Hasler and Suesstrunk arrived at their metric by surveying twenty human participants, likely American college students, asking them to rate the colorfulness of photographs on a 7-point scale. They then computed several combinations of the mean and standard deviation of the color channels, and found linear combinations with maximum correlation with image colorfulness scores derived from the human ratings. Color perception, and particularly hue sensitivity, varies from person to person [99]; if the authors had instead started from a different set of image features, used a collection of paintings instead of photographs or surveyed a different population, the best measure might have differed considerably. With that in mind, it seems hasty to say that this function measures "colorfulness" in the abstract.

What if we had a sophisticated computer program which could learn, using human-provided labels, to recognize specific genre and medium information and give nuanced, contextualized colorfulness evaluations? Imagine this program could even learn to predict the perception of each new human user. We could then use it to generate a personalized version of our plot above, based on predictions for each person's perception of colorfulness in each painting. Would this personalized metric constitute a more correct or true measure of colorfulness? And if such a program actually worked, would the analysis it yields, steeped in relativism, tell us anything useful or interesting about collections of paintings?

The key factor here is that by proposing and using a computational measure of colorfulness, we are implicitly theorizing, specifying what we think colorfulness is and implying an analysis of Rothko's paintings. While such a measure can be based on empirical psychology research, its implementation and usage for a given collection of images is ultimately an authored series of analysis choices, realized in code, that can be analyzed, criticized and revised by others. The process of developing such an analysis and arguing for its utility can be a worthwhile digital humanistic research process, even if it is never correct in a scientific sense.

In this Rothko example, we have seen an important pattern:

- An aesthetic phenomenon, which we experience in images, is difficult to measure.
- Computationally simple implementations do not align well with human perception.
- Using a more complex implementation improves performance with respect to perceptual data, but human perception still varies due to contextual and subjective factors; there is no one way that all humans experience the aesthetic phenomenon in question.
- These measures, rather than failed approaches to universal measures of aesthetics, offer different perspectives about the phenomenon under study; their performance is a matter of interpretation.

Completely missing from this discussion has been the limitations of our dataset and lack of contextual knowledge about art history in general and Rothko's life in particular. In Chapter 5, we will revisit this problem — splitting time into periods based on works of art — and put forward a method for uncertainty-aware periodization. Using Bayesian statistics, we can take into account the incompleteness of the data, and quantify our prior historical understanding, to see how certain we can be about the location of each period boundary. But some limitations of quantitative perspectives cannot be overcome. For example, a central aspect of Rothko's works is their enormous scale. While we can easily measure the dimensions of paintings, no measure of a disembodied digital image can convey the experience of standing under a looming canvas. Similarly, even the most useful measure of an aesthetic quality cannot substitute for actual human perception.

These sorts of aesthetic phenomenon problems do not only occur in cultural analytics. They are common in a variety of disciplines including information retrieval, recommender systems, artificial intelligence and computational design. We can imagine scenarios where humans might want to find or create things with particular visual qualities which they readily perceive, but cannot describe precisely. Solutions to these problems, even if they are not substitutes for human perception, can be useful in the creation of tools for artists and designers, as well as research tools for librarians, scholars and the public. Further, the boundary between these problems and the rest of computer vision is fuzzy. Many central computer vision problems, like object recognition or image captioning, have subjective, aesthetic aspects which are treated as noise.

Regardless of the specific application area, aesthetic phenomenon problems are serious challenges for the research methods of computer vision — how we go about formulating vision problems, developing algorithms, collecting data and evaluating performance in order to develop effective and reliable computational vision systems. Particularly, when we acknowledge that the underlying problem is subjective, increased performance on a quantitative metric no longer necessarily corresponds to a more useful or informative algorithmic system in general. These issues cannot be fixed with a better performance metric, because we can just ask the same kinds of questions about that metric. Eventually we have to confront the subjective factors and impose a perspective. Rather than impose our perspectives quietly, and ignore subjective differences until they become problematic, this thesis makes the case for an alternative theory of performance which explicitly handles subjectivity, grounded in Donna Haraway's feminist and posthumanist theory of situated knowledges [138].

By drawing on feminist theory, we do not seek to equate aesthetics or subjectivity with femininity, call computer vision sexist or discuss gender biases which exist in computer vision models in practice (though such biases do often exist, see [38, 317, 116]). Instead, we are drawing on a more theoretical side of feminism which challenges boundaries where one side is implicitly elevated above the other, like male/female, objective/subjective, scientific/humanistic or quantitative/qualitative [139, 138]. Challenging these hierarchies leads towards alternative epistemologies which ask us to reconsider fundamental assumptions of the scientific process. This analytical lens is particularly useful for aesthetic phenomenon problems because they already blur the boundary between subjective and objective. Feminist theory provides a robust explanation for why we cannot blur that boundary in one direction, quantifying the subjective, without also blurring it in the other direction and questioning the objectivity of the quantitative.

In this context, Haraway's epistemology theorizes that quantitative evaluations can at best provide *partial* evaluation of computer vision algorithms, both in the sense of incomplete as well as favoring some kinds of knowledge over others. We can access alternative perspectives on the effectiveness of systems by allowing users to experience them directly, as things in the world rather than abstract models, and interpreting their feedback. Though this approach requires more work than simply measuring accuracy or correlation on one or more test sets, we find that it offers several benefits over traditional quantitative evaluation. For example, it allows us to evaluate not only the performance of the resulting algorithm, but also the quality of the problem statement and data model. Typically, these elements are only ever evaluated by authors and peer reviewers. Paradoxically, by approaching algorithms for subjective problems qualitatively, we can come to evaluations which are *less* constrained by the subjectivity of the researchers, and less vulnerable to forms of technological determinism common in machine learning [129, 30].

Many of these ideas are not new to computing — subjective [200, Ch. 8], plural [25], situated [50] and feminist [21, 22, 94, 87] approaches to knowledge and evaluation have gained recent interest in computing. However, in this thesis we are not examining user-facing interfaces; rather we are using human subject evaluations for more basic research which might underlie a variety of applications, including computational science or user-facing products, in the future. As we will discuss in Chapter 9, this is highly unusual in computer vision research currently.

1.2 Outline

After discussing related work and defining what we mean by aesthetic phenomenon problems in Chapter 2, the remainder of the thesis forms two parts:

- In the first part (Chapters 3, 4 and 5), we motivate these theoretical issues through three research studies which involve subjective, aesthetic aspects of images:
    - In Chapter 3, we explore the history of web design, and present evidence of its homogenization over time. Leveraging both ethnographic interviews and large-scale image analysis, we show the value of integrating qualitative and computational approaches.
    - In Chapter 4, we present a probabilistic model of template-based color schemes, which we use to reverse-engineer online color scheme generators and test a hy-

pothesis regarding the hue-invariance of human color scheme preferences.

  - In Chapter 5, we describe a Bayesian model of artistic periods based on image features. We show how such a model handles subjective prior knowledge and affords an uncertain perspective on period boundaries in art history.

- In the second part (Chapters 6, 7 and 8), we engage in an extended discussion of a single test problem: image aesthetic quality assessment (IAQA), or the task of determining automatically if a photograph is a "good photo." Rather than present a new algorithm to solve this problem, we study the evaluation of such algorithms.

  - In Chapter 6, we situate IAQA within the history of quantified aesthetics, and using the concept of the aesthetic gap [76], characterize recent work in this space.
  - In Chapter 7, we interrogate personalization as an alternative approach to subjectivity in IAQA. Inspired by an argument from feminist aesthetics, we conduct a paired preference study and investigate when personal differences from average preferences are actually predictable based on particular demographic or image attributes.
  - In Chapter 8, we explore an alternative epistemological approach to subjectivity based on feminist human-computer interaction (HCI). We present a prototype smartphone camera app interface for evaluating the differences between IAQA models, and report preliminary user studies of its effectiveness.

- Finally, in Chapter 9, we discuss these findings, and connect our exploration to larger issues in computer vision and machine learning.

The purpose of this inquiry is not to criticize evaluation metrics or image measures in cultural analytics. Instead, it is a good-faith attempt to reconcile computer science, which relies on mathematical abstraction, with qualitative and interpretive inquiry. Computational and qualitative methods have a great deal to offer one another, but successful interdisciplinary exchange requires first engaging in basic algorithmic research built on human-centric methods. Echoing Haraway [139, 138], we find that the presumed dichotomies present around this topic, particularly those between quantitative/qualitative, objective/subjective and human/machine, are less rigid than we might assume.

By investigating these issues, this thesis offers the following theoretical contributions:

1. We define the space of aesthetic phenomenon problems and illustrate the properties of these problems, particularly in relation to subjectivity.

2. We explore ways to integrate qualitative information into computational approaches to these problems in three contexts: web design history, empirical color harmony and art historical periodization.

3. Drawing on concepts from feminist philosophy, we question the use of benchmark-based evaluation for aesthetic quality assessment and propose a human-centric, qualitative evaluation method as an alternative.

We also provide several technical contributions:

1. Chapter 3 introduces a novel method for measuring the difference between website layouts using distance metrics defined on X-Y tree decompositions.

2. Chapter 4 includes a method for using Gaussian mixture models to model the use of template-based color schemes in a dataset.

3. Chapter 5 demonstrates how a common dynamic programming-based algorithm can be used for segmenting image collections into historical periods.

4. Chapter 7 introduces an open-source web-based data annotation tool for pairwise image annotation.

5. Chapter 8 prototypes a method for embedding an image metric within a digital camera smartphone application.

Specific research questions will be articulated in each chapter, when appropriate.

1.3 Publications

This thesis is based in part on preliminary work from the following publications:

P1 "Studying Empirical Color Harmony in Design" Goree, S., Crandall, D. *Third Workshop on Computer Vision for Fashion, Art and Design at CVPR 2020.*

P2 "What Does it Take to Cross the Aesthetic Gap? The Development of Image Aesthetic Quality Assessment in Computer Vision." Goree, S. *International Conference on Computational Creativity (ICCC) 2021.*

P3 "Investigating the Homogenization of Web Design: A Mixed-Methods Approach." Goree, S., Doosti, B., Crandall, D., Su, N. *ACM CHI Conference on Human Factors in Computing Systems (CHI) 2021.*

P4 “Correct for Whom? Subjectivity and the Evaluation of Personalized Image Aesthetics Assessment Models” Goree, S., Khoo, W., Crandall, D. *AAAI Conference on Artificial Intelligence (AAAI) 2023.*

P5 “Situated Cameras, Situated Knowledges: Towards an Egocentric Epistemology for Computer Vision” Goree, S., Crandall, D. *Eleventh International Workshop on Egocentric Perception, Interaction and Computing at CVPR 2023.*

Additionally, two other works were written as part of my PhD studies:

P6 “‘It Was Really All About Books’ Speech-like Techno-Masculinity in the Rhetoric of Dot-Com Era Web Design Books” Goree, S., Crandall, D., Su, N. *ACM Transactions on Computer-Human Interaction (TOCHI).* Vol. 30, no. 2, 2023.

P7 “Attention is All They Need: Exploring the Media Archaeology of the Computer Vision Research Paper” Goree, S., Appleby, G., Crandall, D., Su, N. ArXiv preprint, under submission at ACM CSCW.

Chapter 2

Definitions and Related Work

Since the materials in this thesis cover a wide range of topics and touch several disciplines, we will introduce definitions and discuss recent work which is related to the whole thesis in this chapter, and introduce additional related work, as appropriate, in each subsequent chapter. First, we will explore the use of computer vision for cultural data, and develop a working definition for our central topic, aesthetic phenomenon problems, then connect this topic to discussions in computer science, philosophy, psychology, computational creativity and cultural analytics. Second, we introduce Haraway's theory of situated knowledges as a way of approaching subjectivity for these problems, and discuss recent work engaging with feminist epistemology in computing. Finally, we use three examples to explore how subjectivity plays a role in aesthetic phenomenon problems: the 18th century quantitative art criticism of Roger de Piles, the 20th century origins of colorimetry, and the 2010 computer vision paper "Predicting Facial Beauty Without Landmarks."

2.1 Aesthetic Phenomenon Problems

2.1.1 Aesthetic Phenomenon Problems in Computer Vision

Recent advances in computer vision offer new opportunities for studying visual culture. This topic has prehistory in highly geometric approaches to fine art images, such as the work of Stork and Johnson studying position illuminants in Baroque paintings [293], Irfan and Stork's authentication of Jackson Pollock paintings [159], Bernadini et al.'s 3D reconstruction of Michaelangelo's Florentine Pietà from image data [32] or Crandall and Snavely's analysis of networks of people and the cultural landmarks they photograph from Flickr data [69]. While geometric and geospatial approaches to cultural images are not beyond criticism, these sorts of studies are not the topic of our inquiry here.

Lev Manovich's concept of "Cultural Analytics" puts forward a more radical claim: that data analysis and visualization can be used to carry out analysis and interpretation of cultural images at scale [326, 223, 224]. There are now a variety of projects applying computational image analysis, often based on machine learning, to cultural images. For example, Elgammal and Saleh quantify the historical creativity of a work of art from visual features alone [96]. Thomas and Kovashka automatically infer the political alignments of web advertisements [303]. Karjus et al. measure the visual complexity of paintings over time [181]. Doosti et al. propose a method for studying the visual similarity of webpages [90]. These sorts of projects involve reasoning about visual qualities which lack an objective basis. We call these kinds of visual computing tasks "aesthetic phenomenon problems."

Aesthetic phenomenon problems are computational problems [66], which involve formalizing and computing measures of experiential concepts defined in terms of one or more human subjects. While this is not a category with firm boundaries, we find it is a useful framework for thinking about the commonalities between problems in different domains. Some examples include seemingly-impossible problems, like automatically deciding whether an object in a digital image is art, whether a sound captured in digital audio is sad or whether a computer game is fun. More mundane examples include metrics for color consistency [18], visible compression artifacts [209] or text legibility [292]. These problems, in general, cannot be formalized mathematically because people disagree as to whether a given formalization matches their perception. We can define algorithms to solve these problems, often fit to perceptual data using machine learning, but because of the lack of objective ground truth, evaluating their performance in general is very difficult.

The difficulty of evaluating algorithmic solutions to aesthetic phenomenon problems becomes even more important when those algorithmic solutions are then used to evaluate other algorithms. For example, consider the Frechet Inception Distance (FID), which measures the perceptual difference between two sets of images [153]. FID is now ubiquitous in the evaluation of image generation algorithms, even though it does not align well with the human judgments it was designed to mimic [210] and the typical algorithm used to compute it uses a statistically biased estimator, so FID scores evaluated on differently sized samples are not comparable [63]. These sorts of issues only emerge because there is no way to formally evaluate the FID — the authors who proposed it, Heusel et al. [153], only justify their approach by showing that the FID correlates with distortion under six specific image transformations.

There are a variety of statistical methods for working with inter-subject disagreement in perceptual data. When working with categorical labels, variation is typically measured using inter-rater reliability measures, such as Cohen's Kappa and its various generalizations in Fleiss' Kappa and Krippendorff's Alpha [331]. These metrics are designed for categorical annotation, but can be extended to ordinal or numerical annotations [192]. When faced with bounding box or segmentation data, as is common in computer vision, we typically use the Jaccard coefficient to measure annotator agreement [299].

Typically, after assessing the degree of inter-subjective variation, we perform aggregation to eliminate the subjective differences between annotator responses, usually by taking the mean, median or mode label. Recently, however, some researchers within natural language processing (NLP), driven by concerns about subjectivity, are turning away from aggregation. Their approach, called perspectivism, approaches annotation without assuming the existence of ground truth labels [25]. Basile et al. argue that machine learning problems exist on a spectrum from fully objective to fully subjective, and a problem's position on this spectrum can be quantified using dispersion and reliability measures [24].

In the perspectivist framework, datasets can be described as weak perspectivist approaches, which store distributions of labels instead of averages, and strong perspectivist approaches which save disaggregated data — where every annotation is treated as a separate data instance [24]. These methods are a direct response to concerns around algorithm bias arising from data annotation, and an extension to Bender and Friedman's case for data statements which make the annotator demographics and data provenance for NLP datasets explicit [29]. Disaggregated, perspectivist approaches show promise for tasks like emotion recognition [245] and hate speech detection [175] which, while not in the visual realm, are just as subjective as the problems discussed here. Interesting to our inquiry, approaches which fit the definition of weak perspectivism have been used in IAQA, the problem we discuss in Chapters 6, 7 and 8, for over a decade, and our approach in Chapter 7 qualifies as a "strong perspectivist" approach, though perspectivism was developed independently, in parallel to our work in that chapter.

Given a restricted space of documents to annotate, disaggregated annotation data begins to resemble the data for a recommender system. Within the space of recommender systems research, algorithms based on collaborative filtering yield elegant solutions to problems involving individual user preferences without engaging underlying aesthetic phenomena at all. These algorithms usually leverage a user-content score matrix (equivalent to the annotator-data matrix in perspectivism) to measure similarity between users and to predict scores on content seen by some users but not others [12]. Collaborative filtering algorithms have been extremely effective in systems where the space of data examples is constrained and densely annotated. But the "cold start" problem, handling unannotated examples, poses challenges for these methods [285], and many generalizations of recommender systems problems can be understood as aesthetic phenomenon problems. Additionally, when collaborative filtering-based recommender systems collect annotations for data examples which were shown to users based on the collaborative filtering algorithm itself, the system dynamics can influence annotations [68], which disconnects those annotations from the underlying aesthetic phenomenon they are supposed to reflect. For example, feedback loops unpredictably affect the ratings for songs mostly independently from user preferences [271]. For a less-scholarly examination of the relationship between collaborative filtering predictions and aesthetics, please see Tom Vanderbilt's *You May Also Like* [311].

2.1.2 Aesthetic Phenomenon Problems Beyond Computer Vision

By calling these phenomena “aesthetic” we appeal to the formalized study of aesthetics, as a “science of perception” first named in the 18th century work of Alexander Gottlieb Baumgarten [133]. To some extent, the issues we encounter surrounding this topic are similar to the problems discussed in 18th century aesthetics. For example, in his 1757 essay “Of the Standard of Taste,” Hume argues, against Baumgarten and his contemporaries, that a science of perception cannot really exist because aesthetic judgment is entirely subjective and not based on reason. Beauty, rather than a quality of objects, “exists merely in the mind which contemplates them; and each mind perceives a different beauty.” He also takes issue with reducing judgment to “geometrical truth and exactness,” since it would lead us to find a work which could be objectively called most beautiful or most ugly, which defeats the purpose of artistic criticism [156]. But in 1790, Kant’s *Critique of Judgment* responds, arguing that aesthetic judgment only differs from person to person because it is “bound up with interest,” meaning that we make judgments based other factors. But if we work backwards, rationally accounting for our external interests and biases, we can arrive at judgments which are completely *disinterested*, based only on the pure aesthetic emotion which is the same in all rational observers [176]. Many of the issues surrounding aesthetic phenomenon problems are connected to Kant’s claim of disinterested judgment. We will return to disinterestedness and its critics in Chapter 7.

There is also a rich history of research in psychology which attempts to measure and model aesthetic phenomena, going back to the psychophysics of Ernst Weber and Gustav

Fechner. Fechner famously formalized Weber's law, which relates the intensity of stimuli and sensations on a logarithmic scale [43]. Later, Fechner showed 347 subjects a series of rectangles and ellipses and asked them to choose the most appealing, and the rectangle with proportions drawn from the golden ratio was chosen the most frequently [128]. Fechner's psychological approach was groundbreaking, compared to earlier philosophical study of aesthetics, because it treats aesthetic judgment as a psychophysical process which can be observed indirectly in the laboratory, even if the judgment itself is subjective.

Other authors have approached aesthetic phenomena from a mathematical perspective, focused on elegance over consistency with human perception. The origins of this approach are in the work of George Birkhoff, who proposed that aesthetics in any discipline or medium could be measured through specific implementations of a single universal aesthetic measure based on order and complexity [39]. His approach was highly influential; for example, Moon and Spencer developed the first quantitative model of color harmony based on Birkhoff's theory [236], which went on to influence Matsuda [230] and the template-based color schemes we discuss in Chapter 4.

Within this tradition, Philip Galanter explores the larger space of possible aesthetic measures, defined as functions which perform aesthetic evaluation on digital objects (usually images). This space includes Birkhoff's aesthetic measure [39], all the models of aesthetic quality assessment we discuss in Chapter 6, and even aesthetic evaluation functions discovered via a genetic algorithm which do not resemble human concepts of aesthetics [113]. This unifying category is interesting because the only factor that differentiates aesthetic measures from other scalar functions over a space of digital objects is the way that we use them. Any such function can serve as an aesthetic measure if we use it on artistic images and interpret its scalar output as an aesthetic evaluation. This approach is surprisingly poststructuralist in its view — metrics are not inherently aesthetic measures, we have to interpret them as such — though Galanter does not identify his thought in this way. We take inspiration from this reading in Chapter 8. For more details on theories of computational aesthetics, particularly those based on information theory, please see Greenfield, who provides a history of the term "computational aesthetics" [130], or Jahanian, who gives a taxonomy of computational approaches to aesthetics [163, Ch. 2].

Designing algorithms to solve aesthetic phenomenon problems can be a fascinating intellectual exercise, in dialogue with the philosophy and psychology of perception. However, when our interest in these problems is one of necessity, approaches based on statistical modeling or machine learning become appealing. Such models can be fit to data and capture nuances in perception which are difficult to describe and model explicitly. The recent success of machine learning approaches, and particularly deep neural networks, as proxies for human perceptual measures in computer vision cannot be overstated [334].

In cultural analytics, many qualities we wish to study result in aesthetic phenomenon problems, as we attempt to operationalize those concepts algorithmically. For example, as mentioned earlier, Elgammal and Saleh study the creativity of historical art images [96]. Their approach starts with vectors based on Torresani's classeme features [305], trained on photographs, to quantify what the art images look like. Next, they define a measure of visual similarity between pairs of images and construct a weighted directed graph with edges from each image to similar images later in time. Finally, they quantify creativity using a weighted eigenvector centrality measure on that graph [96]. Each of these steps — defining image similarity, the influence graph and creativity on that graph — constitutes an aesthetic phenomenon problem which is largely justified through an appeal to mathematical elegance, qualitative examples of how paintings assigned high creativity scores agree with the artistic canon and a study investigating the effect of counterfactual dates. However, judging a measure of creativity based on its ability to detect works typically viewed as creative begs the question of whether canonically innovative works of art are actually as creative as the canon would suppose. For a more thorough statement of this argument in a different context, see Christine Battersby's discussion of Francis Galton's study of genius [26, Ch. 13].

2.2 Situated Knowledges and the Posthuman Subject

Aesthetic phenomenon problems are difficult to evaluate because they are subjective. But what is subjectivity and how do subjects gain knowledge of the world which differs from one another? In this section we will briefly introduce the work of the feminist theorist Donna Haraway, particularly her essays "Situated Knowledges: the Science Question in Feminism and the Privilege of Partial Perspective" [138] and "A Cyborg Manifesto" [139]. In these texts, Haraway puts forward a theory of knowledge and the subject which blurs the lines between human and machine, between subjective and objective knowledge and between qualitative and quantitative inquiry, making it an appropriate theoretical foundation for our topic.

In "Situated Knowledges" [138], Haraway discusses scientific objectivity and its fraught relationship with feminism. Many of the central questions of feminist theory involve questioning scientific facts — particularly those about women, their biological differences from men and how those differences should relate to social issues. Some feminists cite these conflicts as justification to throw out scientific inquiry itself as biased, but Haraway, who comes from a biology background [140], disagrees. She does not want a feminist critique of science to serve as "one more excuse for [feminists] not learning any post-Newtonian physics," or a justification for pseudoscience. Instead, she seeks to find a way of thinking about knowledge which admits both real scientific knowledge about the world as well as arguments against sexist findings in science [138].

To reconcile these perspectives, Haraway employs vision as a metaphor. She observes that science, when it separates a "view" of the world from the way that it was captured, performs a "god trick" — pretending that an observer's limited view can actually see everything from an omniscient god's-eye view. But all vision: human, animal or machine, is actually situated, limited and partial. We cannot see distant stars, bacteria or atoms as they truly are; we can only see them as they are captured by cameras, telescopes or other sensors, often put through data analysis systems to produce images designed specifically for our eyes. Haraway's position is not an attack on scientists, who typically acknowledge the limitations of their instruments and methods. Rather it is a critique of industry, government and the public, who perform god tricks when they treat the findings of scientists as completely true, detached from the limitations of their research methods. In Haraway's words:

> Infinite vision is an illusion, a god trick...We need to learn in our bodies, endowed with primate color and stereoscopic vision, how to attach the objective to our theoretical and political scanners in order to name what we are and

> what we are not...Objectivity turns out to be about particular and specific embodiment and definitely not about the false vision promising transcendence of all limits and responsibility. The moral is simple: only partial perspective promises objective vision. All Western cultural narratives about objectivity are allegories of the ideologies governing the relations of what we call mind and body, distance and responsibility. Feminist objectivity is about limited location and situated knowledge, not about transcendence and splitting of subject and object. It allows us to become answerable for what we learn how to see. [138]

In other words, humans and our hybrid technological-biological vision systems, are always part of the universe observing itself. A firm boundary between subjective and objective knowledge is part of a splitting of subject and object, which is tied up in older Western ideas about difference between humans and nature. But, as Haraway claims, "coming to terms with the agency of the 'objects' studied is the only way to avoid gross error and false knowledge of many kinds," [138] meaning that the human and nonhumans that we research can have their own ways of understanding the world, and a boundary between subject and object (with researchers as subjects and participants as objects) denies them that agency. Rejecting these connected boundaries does not make scientific truth subjective or relative, it actually makes findings more objective because we acknowledge the embodied, situated and plural reality of our knowledge. While computer scientists are very familiar with situated approaches to knowledge in the context of robotics where our agents do not get to have full vision of their environments [50], we have yet to understand that similar limitations also affect our research methods.

In "A Cyborg Manifesto," Haraway extends this epistemological perspective into a theory of subjectivity, and critique of several branches of feminism. For Haraway, human minds do not exist separately from bodies or the world; instead they are connected and conditioned by the technologies that shape our social selves and our worldviews. In Haraway's words, "we are all chimeras, theorized and fabricated hybrids of machine and organism...cyborgs." [139]. Haraway identifies three key boundaries which have broken down in the twentieth century to create cyborg subjectivity:

- *Human and animal:* understanding of evolutionary biology teaches us that humans are animals, not fundamentally different from other great apes.
- *Organism and machine:* Many of the behaviors of organisms, including thought, are actually quite mechanical and can be replicated by machines.
- *Physical and nonphysical:* Brains and computers are both physical devices, making ideas and virtual objects ultimately physical as well.

Donald Hall identifies Haraway's approach as a stark departure from humanism. The human subject, rather than a singular disembodied Cartesian mind, is conceptualized as a hybrid biological-technological system, embedded in larger ecologies. Science fiction stories about cyborgs — future people who are part human and part machine — allow us to grapple with this reality and understand ourselves as simultaneously people and things. [137, Ch. 4]. For an example of this shift, consider eyeglasses. We think of our vision systems as biological, but for many of us with visual conditions like myopia, we actually have hybrid biological-technological vision. If our glasses break, we have lost a fundamental part of ourselves and our ability to engage with the outside world, and need to rely on the optometry system and its global supply chains to return to everyday life. Other technologies, from televisions to personality tests, similarly interface with and extend the biological self.

While Haraway does not discuss aesthetics herself, her concept of the subject can be read as an approach to aesthetics. Our experience of a work of art is mediated through several lenses, including the literal biology of our eyes and brains (and possibly glasses), the technologies of image production and reproduction and the socially and historically situated structures of meaning which shape our interpretations. This view generally aligns with other contemporaneous feminist art and media criticism, such as Larua Mulvey's concept of the male gaze [240], Christine Battersby's critique of genius [26] or Carolyn Korsmeyer's critique of Kantian disinterestedness [190]. These works encourage us to not just think critically about the way that gender, race and other factors are depicted within works, but the way that they structure our view of those works, including our assumptions about the audience and artist and the cultural-industrial processes which brought those images to our attention in the first place.

The most important takeaway for our inquiry from this simple reading of Haraway is that under this concept of subjectivity, the solutions to aesthetic phenomenon problems are computational ways for us to see images. They are not more objective formulations of the phenomena under study, and they are not artificial subjects experiencing aesthetic phenomena themselves (though a theoretical future artificial subject could see works of art through them). Additionally, Haraway's approach explains how all of these shifts in our understanding of knowledge and the subject are linked. If we want blur the boundary between subjective and objective knowledge to treat aesthetic phenomena as quantifiable and approach culture through data analysis, we also have to blur it in the other direction and see computation and data analysis as interpretive processes carried out by human researchers, embodied and embedded in historical and material relations. Though it underlies much of our inquiry, we will return to situated knowledge and its implications for our evaluation of algorithms for aesthetic phenomenon problems, in Chapter 8.

Two key communities within computing have been exploring Haraway's theory of situated knowledges: feminist HCI and feminist digital humanities (DH). Our work here builds on both of these areas.

Situated approaches to knowledge were introduced to HCI by Bardzell and Bardzell [22], and have led to a wide variety of qualitative design studies which center the experiences of women and prioritize situated and embodied approaches to knowledge. For example, Sultana et al. study the lives of rural women in Bangladesh and make recommendations for how to design within, rather than against, patriarchal society and how to avoid treating users as passive victims in the design process [297]. Fiesler et al. examine the design of an online fanfiction archive developed primarily by and for women, and show how design decisions reflect and negotiate within the values of that subculture [106]. Su et al. analyze forum posts written by owners of sex dolls, and explore the way that these dolls, despite not being animate, function as embodied technologies of the self, allowing users to develop complex fictions exploring norms surrounding intimacy and "care of the self" [295].

Feminist theory entered the digital humanities at a similar time, though it has prehistory in the critical cartography of authors like Brian Harley [67, Ch. 14] and Wood and Fels [322]. These approaches question the disembodied objectivity of maps, and the ways that they express geopolitical power. Later critical approaches extend these arguments to data visualization. For example, Johanna Drucker locates data visualization in rhetoric, and describes the historical development of Western visual language for information, and its relationship to power [91]. d'Ignazio and Klein argue for explicitly feminist approaches to data visualization, which they operationalize through six principles questioning objectivity and binary thinking, centering context, embodiment and affect and explicitly discussing power and labor [94]. In *Data Feminism*, the same authors build on this framework with their concept of data visceralization, which elevates the emotional over the rational in our presentation of data. They advocate for physical and embodied representations of data, and explicit representations of uncertainty [87].

As discussed in Chapter 1, some digital humanities scholars are quite critical of cultural analytics methods. As articulated by Claire Bishop, digital methods reduce the complex judgments of art historians to simple statistical measures, ignores the difficult questions regarding construction of a canon in favor of easily accessible datasets and represents an incursion of neoliberalism into the humanities [40]. In a dialogue with Bishop, Drucker responds, pushing back on these points and arguing that the worthwhile humanistic inquiry which occurs in digital humanities work exists at the level of information infrastructure, and in constructing the systems of classification and data modeling which underlie the work [92]. This response invokes Bowker and Star's theory of infrastructure — the data schemas and category systems which structure both our understanding of the world and our data collection, but often fade into the background [44]. We engage with a similar concept, "subjectivity in the model" in the subsequent section.

Our approach builds on these uses of feminist methods, and runs parallel to other applications of feminist methods in AI, like recent work in human-robot interaction [320].

It is difficult to leverage computer vision in feminist HCI and DH in part because of the uncritical approach computer vision takes to knowledge and vision. Rather than a fundamental disciplinary boundary, we claim that computer vision methods for aesthetic phenomenon problems can be useful for applications in feminist HCI and DH contexts. But for such applications, our methods should be conceptualized and developed in a human-centric way with emphasis on the situated and uncertain nature of both our modeling and our evaluation. We explore several approaches: in Chapter 3, we treat the output of vision systems as only one perspective on the history of web design, in combination with the experiences of our veteran web designer participants. In Chapter 5, we explore ways of making the uncertainty inherent in analysis of cultural data explicit. In Chapter 8, we turn away from supposedly objective evaluation metrics and towards qualitative evaluations which occur in the real world.

2.3 Three Examples of Subjectivity in Aesthetic Phenomenon Problems

Aesthetic phenomenon problems resist objective mathematical formalization because the ways that we understand the underlying phenomena vary from person to person. One aspect of subjectivity in these problems comes from variation in individual perception, which can be based on a wide range of biological and cultural factors. We will refer to that aspect as subjectivity in the data. However, there is another aspect, which is more infrastructural [44], related to the authors' understanding of the problem. We will refer to that aspect as subjectivity in the model. In this section, we explore the interplay between these two aspects using three examples: early 18th century quantitative art criticism, early 20th century colorimetry and early 21st century computer vision study of women's faces. While these examples are not essential to our broader argument, we include them to illustrate our concept of aesthetic phenomenon problems, and show how it provides an interesting frame of analysis, even in these historical cases.

2.3.1 Example 1: Roger de Piles' *The Balance of Painters*

Arguably the earliest example of a quantitative approach to an aesthetic phenomenon problem can be found in the 18th century work of Roger de Piles, the art buyer for Louis XIV, in his books *The Balance of Painters* [80] and *The Art of Painting* [81]. In the former, de Piles arrives at quantitative ratings for 56 well known artists based on four characteristics: composition, design, color and expression. The tasks of quantifying these characteristics, including the final ranking, can be seen as aesthetic phenomenon problems, as they involve algorithmically formalizing aesthetic qualities. In the latter, he starts from these characteristics and develops one of the first theories of painting. Contrary to appearance, de Piles did not develop this approach as an objective rubric for evaluating artists, but as an explanation for his aesthetic opinions.

The Art of Painting provides a systematic analysis of different qualities of painting, with a focus on composition, design and coloring. While many sections deal with specifics, such as how to correctly depict facial expressions or draperies, other sections focus on more broad, philosophical question, such as whether art can be more beautiful than nature or whether painters are justified in representing the divine using human figures. In Chapter 28, de Piles discusses three sorts of knowledge which can be had about individual paintings:

> There are three several sorts of knowledge relating to pictures. The first consists in discovering what is good, and what is bad in the same picture: the second has respect to the name of the author: and the third is to know whether it is an original or a copy. [81]

By grouping these three kinds of knowledge together, de Piles implies that aesthetic judgements relate to empirical properties that can be discovered, much like authorship or authenticity. He goes on to describe why he believes such knowledge is important.

> The knowledge of principles helps one to find...the cause of the effects that we admire...Those that have not cultivated their minds by the knowledge of principles, or at least have some speculation of them, may however be sensible of the effects of a fine picture, but can never give a reason for the judgment they make. I have endeavoured, by my idea of a perfect Painter, to assist the natural light of the lovers of Painting, however, I do not pretend to make them penetrate into the detail of the parts of the art; that is rather the business of the Painters than of the curious: I would only put their minds in a good way towards knowledge, that they may, in general, be able to know what is good, and what is bad in a picture. [81]

This passage makes the important distinction between making aesthetic judgements and giving reasons for those judgements. He is not advocating a system where aesthetic evaluations are made mechanically based on objective characteristics, but only for finding evidence and reasons to explain such evaluations. Judgements of authorship and authenticity can by supported by the same evidence, by describing the manner of a painter in terms of objective features. Later analysis provides further evidence that the evaluations precede the evidence in de Piles' table: the amount de Piles wrote about each of the artists (which might serve as a proxy measure for how much he enjoys their work) does not correlate with any of the specific characteristics, but does correlate with the overall scores, implying that the final ranking of painters precedes the specific scores [118].

Through this system, de Piles appeals to the quantitative as a way to make his aesthetic evaluations seem more reasoned and objective. Such an urge towards rationality makes sense given his early Enlightenment context: the scientific revolution was quickly transforming the study of the natural world from an imprecise and noisy discipline of natural philosophy to the more rigorous modern scientific method. However, it is still unclear whether these scores reflect his personal taste, or the taste of his employer, Louis XIV.

In either case, we can see the distinction between subjectivity at the level of the model (which characteristics measure the quality of a painter? How should they be weighted?) and at the level of the data (how well-designed are the paintings of Caravaggio?). Allowing subjective judgments at the level of the data does not make the model objective, but it does make the model explicit in a way which gives us insight into the way that the author thought about the problem under study. By trying to be objective and formalizing his intuitions about what is valuable in a work of art, de Piles actually gives us a valuable insight into his own perspective, which has been analyzed by later scholars [118].

Interestingly, this set of ratings has also had a surprising amount of influence on economists studying the art market. Several groups of economists and economic historians have used de Piles' ratings as evidence for historical attitudes towards specific artists and used that quantification to measure the relationship between perceived quality, taste and price [125, 294, 288]. In other words, by quantifying his judgments, de Piles had a disproportionate amount of impact on later scholarship among economists interested in art for its commodity value, compared to his contemporaries. This legacy demonstrates one of the key dangers inherent in quantifying subjective qualities: associations between quantification and objectivity can lead people to treat the quantitative measure as a more objective approach to the concept under study, which at least for de Piles, is not the case.

2.3.2 Example 2: Subjectivity in Colorimetry

Our second example comes from the early 20th century history of colorimetry, the science of measuring subjective human color perception, and specifically the development of the 1931 Standard Observer model and CIE XYZ color space. Again, we will see a distinction between subjectivity in a the model and subjectivity in the data. Much like de Piles' model of quality in painting, the resulting model of colorimetry is not a scientific theory, but a designed model architecture, authored by a small number of Anglo-American scientists and fit to data.

Colorimetry as a discipline developed during the 19th century, but found maturity with the first meetings of the International Commission on Illumination (CIE) starting in 1913, culminating in the 1931 adoption of the CIE standard observer model for color perception [46]. The standard observer model relates tristimulus values, combinations of red, green and blue light, to specific pure wavelengths which humans see as identical [325]. It serves as the foundation for most subsequent digital image technologies, including computer vision. The development of the standard observer can be seen as an aesthetic phenomenon problem: "perceptually identical color" is an experiential judgment, which measurably varies between individuals. Driven by their goal of standardization, the CIE decided to ignore these variations, and construct a single definition of "normal" color vision.

Color perception has been the topic of a wide range of scientific research over the past century [325]. One key finding is that there is a great deal of individual variation in color perception. Some of that variation is due to a variety of biological factors like color deficiencies, natural variation in L:M cone ratios and changing cone sensitivity over individual lifespan [99]. Other aspects of variation are due to cultural factors, particularly native language, which effects human sensitivity on color recognition, categorization and detection tasks [302]. Given all of the sources of variation in color sensitivity, color perception is actually remarkably stable in people who do not have abnormal color perception due to a disorder like dichromacy [99]. One possible explanation for this stability is that color words form labels to ground different individuals' development in the same linguistically situated color categories, so humans learn to interpret roughly the same color information, despite all of the low-level differences [217].

Despite individual variation, we only use one colorimetric model for digital images: the CIE standard observer model. This model was initially developed based on the simultaneous research of William Wright and John Guild in the late 1920s. Guild studied the perception of seven individuals who worked in the National Physics Laboratory (six men and one woman) [132], while Wright studied ten anonymous "trichromats" [323]. Despite different apparatuses and experimental designs, they found roughly the same color matching curves — combinations of red, green and blue light which humans perceive as identical to each wavelength of light. The 1931 CIE standard observer model and XYZ color space are based on these two studies [323, 132, 325], their reported tristimulus values are shown in Figure 2.1.

Wright and Guild arrived at similar results, in part, because of historical factors: they

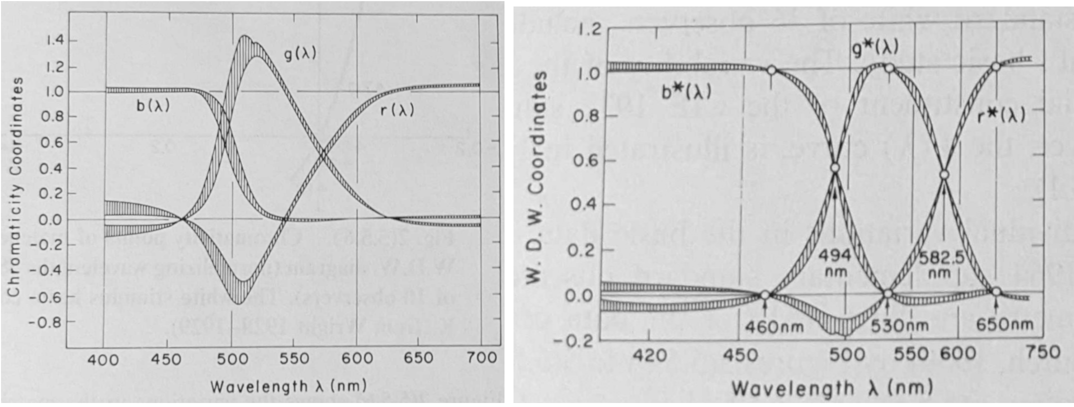


Figure 2.1: Tristimulus value ranges measured for seven participants by Guild (left) and for ten participants by Wright (right). Figures from [325].

were both determined to arrive at a single definition of "normal" color perception despite individual variation based on Herman von Helmholtz and Thomas Young's three dimensional color theory [169]. This approach responds to Munsell's groundbreaking color system a decade earlier [246], and exists in service of both emerging industrialized electric lighting and corporate advertising and branding, both of which required standardized color [41].

These authors are upfront about their motivations. Guild argues,

> The international photometric scale which governs the output of large industries concerned with the production of illuminants, necessitates the elevation of some particular set of visibility data to the dignity of a standard, to be used universally in all computations carried out for technical purposes. Hopeless confusion would arise if every lamp manufacturer, or every photometric standardising laboratory, employed units based on individual judgment as to the most accurate visibility data available at any particular time. [132]

Similarly, Wright speculates,

> It has become abundantly clear that, until data for a 'normal' eye have become standardised, the scope of colorimetric science will be strictly limited...I should not, therefore, be surprised if subsequent research shows some modifying influences to exist; but if such is the case it is safe to predict that any eccen-

> tricities will be of a small order and unlikely to affect the value of practical colorimetry. [323].

In other words, they do not arrive at a standard observer model as a scientific theory of human color vision, they develop it as an accurate-enough model to solve an important problem for scientific and industrial applications. In fact, this perspective was only possible because the German color scientists, who participated in the CIE and were interested in more psychological approaches to color, had been excluded due to official ostracization after World War I [169]. Later on, Wright described the way the CIE accepted their standard observer:

> Priest was the official American delegate and I think it was quite clear that he had come briefed to delay the adoption of any standard observer, since he thought we were rushing things. He in fact raised a succession of objections...Then overnight, T. Smith...and Guild would recalculate a lot of data to meet Priest's criticisms, and Priest would turn up next morning with something else to object to. In the end they wore Priest down and he accepted most of the proposals that Guild was going to put forward at the C.I.E. meeting [46]

Guild and Wright approached colorimetry from a particular perspective driven by a need for standardization, shaped by the sociopolitical process of a committee discussion. Despite its origins, the model they developed works, and has been vindicated by the countless successful applications of digital imaging and computer graphics in the decades since. We can imagine a more accurate model, which accounted for color effects like shine and glitter, or percecptual phenomena like afterimage or chromatic abberation in the human eye, as the German color scientists proposed [169], or personalized to account for different color sensitivity. But such a model may have had more barriers to adoption and created countless implementation headaches for later engineers.

2.3.3 Example 3: The 2010 ECCV Hot or Not Paper

Finally, we turn to an example closer to the present state of computer vision, the well-cited "Predicting Facial Beauty Without Landmarks" [126] of Gray et al. published at the 2010 European Conference on Computer Vision (ECCV). This paper (which we refer to as PFBWL) "aims to investigate and develop intelligent systems for learning the concept of female facial beauty and producing human-like predictors" [126], and introduced a dataset of relabeled images from `hotornot.com`, a 2000s-era website which allowed internet users to upload photos of themselves and rate others' face photos.

While we could criticize `hotornot.com` for its blatantly sexist premise, or admonish PFBWL for using images of peoples' faces without their consent, our interest is more subtle: how the authors approach the subjectivity of facial beauty. They treat both their intuition about beauty and `hotornot.com`'s particularly turn-of-the-millennium way of seeing faces as representative of a kind of objective "common sense" truth about how humans see each other, and make that intuition real through data. Our intent is not to criticize PFBWL or its authors, but instead to use the paper as an illustrative example of how subjectivity is often treated in computer vision research.

The `hotornot.com` website allowed users to rate others' face photos on a ten point attractiveness scale. According to *Time* magazine, the site's cofounders, James Hong and Jim Young, "thought of the idea in 2000 as they debated the attractiveness of a passing woman on the street. They decided to let the masses vote. Within a week of launching, the site has two million page views per day" [89]. The site pioneered the interaction pattern underlying social media platforms like Facebook and Instagram where users upload photos and then see and express their opinion of others' photos [89, 170].

Regardless of the gender of the users involved, the `hotornot.com` interaction pattern embodies a particularly masculine way of looking, one that resembles Laura Mulvey's concept of the male gaze [240]. The male gaze is a theoretical tool from feminist media studies which can be understood as the factor which differentiates nakedness from nudity. In the words of art critic John Berger, "a naked body has to be seen as an object in order to become a nude" [31], meaning that the drama within a nude portrait can only exist through the presence of an unpictured masculine observer immediately behind the image plane. For example, in Figure 2.2 left, we reproduce the Ingres painting *La Grande Odalisque*. The odalisque (a fictional concubine of an Ottoman sultan) looks directly towards the image plane, reacting to the viewer as if they are a man. This concept proves particularly useful for analyzing representations of women in visual culture, particularly cinematography, which is often warped around a hypothetical (presumably heterosexual) man watching from behind the image plane [240]. When our analysis of films or paintings ignores that hypothetical man, we implicitly assert the masculine way of looking as neutral or objective, echoing Haraway's concept of the god-trick.

PFBWL effectively automates this masculine way of looking at photos of women. The paper starts by sketching a history of facial beauty that begins with "artists and philosophers," flows through "social scientists," and ends with "computer scientists," because of "the need for more complex feature representations." This historical progression from qualitative to quantitative to computational is framed as moving from ill-defined to well-defined,

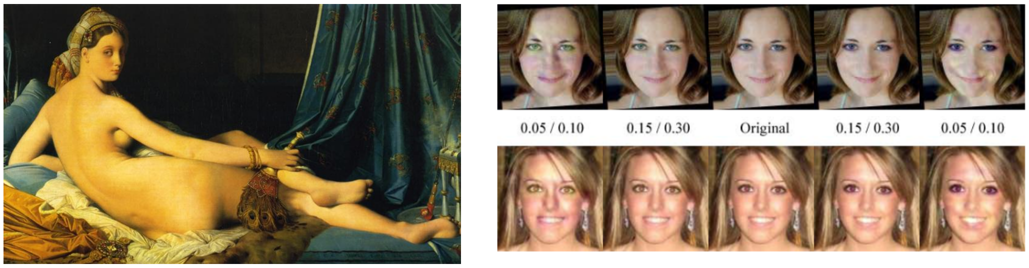


Figure 2.2: *Left:* Jean Auguste Dominique Ingres, *La Grande Odalisque* (1814) from Wikiart. The depicted woman stares towards the image plane as if there were a man immediately behind it. *Right:* Figure from PFBWL showing their method for optimizing the beauty of images, using their model. Notice that the main difference from left to right is the quantity of eyeliner.

and the paper's approach, which avoids manually-selected facial landmark points, advances this progression from many subjective definitions based on specific research findings to an objective definition of beauty learned from data. To conduct experiments, PFBWL asked 30 labelers about the relative beauty of `hotornot.com` faces.

The paper does not indicate who these labelers were or how they were instructed to label, but it does say that the labeling process was framed as a forced-choice task between pairs of face images. This rating system is a departure from the one on `hotornot.com`, which the authors explain by referencing the difficulty of assigning scores to images. "Each user will have a different system of rating images, and a user's rating of one image may be affected by the rating given to the previous image" [126]. Once many such pairwise labels are collected, the authors use gradient descent to find a vector $\mathbf{s}$ where each index is an absolute score s_i corresponding to each image i to minimize a cost function,

$$J(\mathbf{s}) = \sum_{i=1}^{M} \phi(s_i^+ - s_i^-) + \lambda \mathbf{s}^T \mathbf{s}.$$

In other words, for each comparison, scores incur a negative cost $\phi(d) = e^{-d}$ on the difference between the current scores for the two images, in addition to a quadratic cost based on the magnitude of the scores. This implies an underlying model where each image has a hidden, constant, true beauty score which obeys the transitive property (if a is more beautiful than b and b is more beautiful than c, then a is more beautiful than c), and the difference between the scores for two images ($s_i^+ - s_i^-$) determines the chance that a user will choose one over the other. Interestingly, this model resembles the Elo system in competitive chess, which also relates the probability of a pairwise comparison (match win) to a score difference [121]. The authors acknowledge that this is a simplification, as "each individual's opinion can be varied due to factors like culture, race and education" so they "focus on learning the common sense and leave further investigation on personal effects to future work."

Finally, the paper uses machine learning models, which are pre-AlexNet one and two-layer convolutional neural networks, to estimate scalar beauty from image regions. The models achieve test set correlation scores ranging from 0.134 to 0.458, which the authors assure us is not a problem because their goal is not to produce a highly accurate model, but to more objectively study beauty. So they use the dual form of their optimization problem to compute a "derivative of beauty" for any input image, which can then be used to optimize a photo's beauty; we reproduce their example in Figure 2.2 (right). Interestingly, their model focuses on the quantity of eyeliner, picking up on the makeup trend around 2010.

While the authors are able to find a meaningful trend in face image ratings data, their modeling assumptions about subjective difference deserves unpacking. In their attempt to find a solvable machine learning problem, the authors make the modeling assumption that "common sense" facial beauty is an objective property of images, which individuals observe unreliably. Subjective difference, rather than disagreement about what beauty is or how to describe it, is just a factor which causes humans to err and choose the objectively less beautiful face some of the time. The authors ignore subjectivity at the level of the model, and take whatever steps they can to minimize for subjectivity at the level of the data, assuming that individual ratings are noisy estimates of an underlying common sense.

While this discussion has been somewhat critical, our goal is not to disparage the authors' work in this paper. They made defensible modeling decisions and arrived at an interesting result. Instead, we seek to use this example to discuss the issues which can arise from lack of self-reflection regarding subjectivity. Their approach to `hotornot.com` leads them to reify its inherent male gaze and elevate a trend in their participants' rankings to an objective finding around how "people" rate facial beauty. A larger, more representative set of labelers might have improve their claims. But no increase in dataset size allows them to escape their culturally and historically situated concept of what beauty is and how it should be understood quantitatively.

In these three examples, we have illustrated two things. First we have shown our concept of aesthetic phenomenon problems and how it applies not only to current work in computer vision, but historical attempts to quantify the aesthetic qualities of images as well. Second, we have demonstrated the way that these problems involve subjectivity on two levels: the data, which encodes subjective differences between research participants, and the model, which is shaped by the subjectivity of the researchers. While we can use large sample sizes and measures of dispersion to assess subjectivity in the data, we can never really assess or account for our own subjectivity as researchers when approaching these problems.

Part I: Three Motivating Studies

Chapter 3

Measuring the Homogenization of Web Design

3.1 Introduction

Web design (i.e. the design of pages for the world wide web) is a discipline which merges technical, artistic, interactive and textual elements. Though it is much newer than other design disciplines, its thirty years of history can serve as a snapshot to help us understand the mechanisms underlying the development of design standards and practices [100, 51]. The web also has the benefit of the Internet Archive (`https://archive.org`), a repository of 780 billion web pages which can serve as a historical archive for study. However, many of the questions that we'd like to answer regarding web design rely on organizing pages within that archive based on their subjective qualities, like visual design similarity. In this chapter, we will explore ways for measuring the *visual homogeneity* of the web, and how computational measures and ethnographic inquiry can inform one another and deepen our understanding of web design practices. This hybrid approach to web design will serve as our first motivating example to illustrate the potential benefits of hybrid qualitative-computational methods.

We are interested in homogeneity because as the technical constraints on the web (e.g., bandwidth) have loosened and CSS and JavaScript have matured into expressive design tools, we would expect the variety of visual designs on the web to explode. Certainly the accessibility and power of such tools have led to an influx of creative remixes in at least the video medium [203]. Instead, at least anecdotally, the opposite has happened: a number of design blogs have posed questions like, "Why do all websites look the same now?" [239, 65, 233, 97, 11, 263, 213, 276, 5, 4, 208]. These posts typically point out common web design patterns and surmise that templates, common web frameworks, and the complexity required of today's sites(e.g., ensuring that they are accessible and responsive to multiple screen sizes) have led to the current state of designs. For some, homogenization is positive because it makes the Internet more usable and inclusive [276], while others argue that homogenization stifles creativity and negatively affects user experience by divorcing form from content [239].

For similar reasons, HCI research has begun to examine trends in website design. Chen et al. [61] speculate that web design has evolved in concert with technological advancements such as changing screen resolutions and sizes, new media formats, and the popularity of templates. These are described as distinct web "design periods." Moran [237] and Brage [48] address the rise of Web Brutalism, a related counter-trend against homogeneous designs. Additionally, there is a growing area of interest in the study of design history [102] and history is a growing area of interest in the study of web design [164]. Purely aesthetic design qualities have been shown to influence perceived usability [47, 232], making design history important for studying the past, present, and future of interaction. Some of this work is motivated by a belief that understanding the history of web design can expand designers' repertoire, leading to innovative designs drawing from the past, for example.

Our central research question in this chapter is:

RQ1 — How has the visual homogeneity of the Internet changed between 2003 and 2019?

We investigate these dates in particular because of the design periods observed by Chen et al. [61]: we believe that the homogenization trends that designers and scholars have observed may have occurred within and between the “Chaos,” “Formative,” and “Condensation” periods that they identify.

We use a mixed-methods approach to identify and explain homogenization in website design. We introduce and apply computational approaches to characterize visual properties of website designs, first focusing on a small dataset of 100 representative websites, and then scaling up to three larger datasets which contain over 200,000 snapshots of over 10,000 websites. To contextualize our computational findings, we conducted semi-structured interviews with 11 expert website design professionals (each having at least about 15 years of experience) involving their own historical portfolios. Our design history approach [88] is a first step to investigating, on a wide scale, the design evolution of the web. (However, we acknowledge that due to challenges of data availability and biases in our own experiences of the web (cf. Section 3.3.1), our sampling strategies tend towards sites from the American English-speaking web.) Additionally, our work contributes a modeling exercise to an existing conversation in the HCI community on how to quantify websites’ aesthetic qualities [162, 47, 197, 260, 264], and offers a novel method for measuring layout similarity using tree edit distances.

We find that average distance between website layouts in our datasets *declined signifi-*

cantly —43%— between 2010 and 2019, providing the first—to our knowledge—quantitative evidence that website designs are becoming more similar to each other. However, we do not believe homogenization is necessarily a cause for alarm. Consistent use of familiar design patterns can improve usability, since common patterns may have had rigorous usability testing and may already be familiar to new users [279, 310]. Websites which conform to style trends may also be more likely to satisfy accessibility standards [55] (although the converse is not necessarily true: accessible sites do not have to look visually homogeneous).

Yet we believe the design of technical systems supporting the web should be scrutinized for their potential to subtly change the space of legitimate designs. For example, the Mozilla Foundation argues that decentralization is essential to the long-term health of the Internet [110] to prevent companies from undermining privacy, openness, and competition. Arguably, the homogenization of design, may signal that a few corporations have gained influence over what constitutes proper design on web.

A preliminary version was published at ACM CHI in 2021 (Publication P3).

3.2 Related Work

Our work in this chapter lies at the intersection of two disciplines that crisscross HCI: design history and cultural analytics.

3.2.1 Design History

Design history argues that the seemingly mundane and anonymous work of countless engineers and designers is worthy of serious academic study [103]. While much work has focused on either design of objects and material culture or graphic design in printed material, the World Wide Web has also been analyzed by design historians. Engholm [101] argues that web design should be part of the corpus considered in design history and describes how concepts of genre and style from design history can be applied to the web. Chen et al.'s interaction design criticism sessions identified a series of design periods in the history of the web [61]. The latter argues that larger-scale work is needed to more holistically examine the design patterns in the history of the web. Our work takes a first step towards this vision, using quantitative methods to analyze large datasets in archives of the web to more broadly and systematically measure trends in web design.

Our research also aligns with efforts by design historians in constructing online galleries of web design, such as the Web Design Museum [191] and webmuseum.dk [3]. The process of acquiring, registering, and preserving websites as cultural artifacts is challenging to do at scale [102]. Research on information retrieval and extraction techniques such as ours may help future web design curators search the vast unorganized material available in collections like the Internet Archive [8].

3.2.2 Cultural Analytics

Online trends have been the subject of much study in cultural analytics research. Consumption patterns of social media videos [254] and photos [329] are significantly related to our cultural preferences. Various social media have been shown to predict trends in music as well [286]. Specific to web design, some work has analyzed raw HTML [260, 197, 162], but Javascript, CSS, and other code makes it difficult to study the visual design of a page without actually fully rendering it.

Visual analysis methods have been used in cultural analytics as well. Ushizima et al. [308] use visual features related to color, spatial organization, edges, and texture to quantify the visual differences between groups on Flickr. Metrics from computer vision have been used to identify stylistic and cultural trends in art history data. Saleh and Elgammal use metric learning on low-level image features to find metrics useful for classifying style, genre, and artist [270]. Influenced by their metric learning approach, we focus on finding useful interpretable representations and metrics specific to color and layout.

Cultural analytics researchers have also studied web design. Brady and Phillips investigate the relationships between color, balance, and usability [47], finding that websites modified to have spatially balanced pages and triadic color schemes are rated more usable than unmodified sites. Ben-David et al. [28] use a "visual distant reading" of color to study the websites of Yugoslavia between 1997 and 2000 and observe both a decrease in color diversity and a shift away from the colors of the national flag at the start of the Kosovo war. Cocciolo [64] automatically measures the quantity of text on web pages and finds that text density has been declining since a peak around 2005. Wobbrock et al. [321] find that the perceived credibility of news websites is heavily related to the visual design, and that the "overall gestalt of a page is responsible for participants' credibility judgments more than any single factor." Overall this work lends further credence to how web design is intimately entangled with its users and their societal context [61].

Recent work has used Convolutional Neural Networks (CNNs) to analyze visual design. CNNs have the advantage of being able to learn "holistic" and domain-specific feature representatives instead of relying on hand-engineered features. Jahanian et al. [164] use both color and CNN features to study web history, and Doosti et al. [90] use CNN features to classify sites based on genre and year. Our work builds directly on these last two papers—which address methods for quantitatively characterizing designs—by focusing on a specific research question about the homogeneity of the web and directly comparing learned and interpretable features. In addition, we conducted semi-structured interviews with web design professionals to better understand, contextualize, and validate the results suggested by our quantitative analysis.

3.3 Methods

We use a mixed-methods approach in our research, developing and employing computational methods to uncover large scale patterns of website design—namely, its increasing homogenization—and conducting semi-structured interviews with experienced web designers to identify sources of these patterns.

3.3.1 Computational Methods: Uncovering Large-scale Patterns in Website Design

One of our contributions is to try to quantitatively measure whether website design is becoming more homogeneous. We do this by collecting and automatically analyzing large-scale historical corpora of rendered images of web pages. Though computational analysis of the semantic structure and textual content of the web has yielded important insights (e.g., ties between organizations) [52], recent work suggests that the visual design of websites encodes information about changes in design standards, technological innovations, and aesthetics [90, 61]. We thus use rendered website screenshots as our main source of data (see Figure 3.1 for an example).

To try to measure visual design differences at scale, we developed computer vision-based distance metrics that try to evaluate perceptual similarity of website images, and then applied these metrics to our corpora. Due to the subjectivity of the problem, there is no general computational metric that can accurately predict human perception of visual similarity in all contexts. Given this lack of a gold standard metric, we applied two very different approaches, with orthogonal strengths and weaknesses, to help avoid basing any findings on biases or artifacts of the metrics themselves.

First, we measure distances in a mathematically parsimonious way using hand-engineered representations of color and layout based on edit distances. Unlike the deep features, these hand-designed metrics are constrained by our prior assumptions about the design features that may be relevant, but (also unlike deep features) they allow us to better interpret the reasons for change over time.

Then, using deep convolutional neural networks, we take a data-driven approach to learn classifiers that estimate the visual difference between web page images [90]. These deep learning models have become ubiquitous across nearly all problems in computer vision because they can learn complex features that may be difficult to describe algorithmically, including ones that predict human perceptual similarity [334]. However, this advantage is also their weakness: while these models typically perform very well, they are "black-boxes" [262] for which it is notoriously difficult to interpret the features they learn or how they make their decisions.

Data Collection

We collected a large-scale dataset of website images over time by using the historical crawls of the Internet Archive [8]. Just as there is no gold standard metric for comparing the visual similarity of websites, there is no obvious choice of URL corpus to use for our analysis. It may seem that the entire web (or a randomly-chosen subset) would be the ideal dataset, but such a corpus would not actually reflect the average web users' perception of the web: it would over-represent sites having many distinct pages, pages of the dark and deep web (which vastly outnumber those of the surface web [146]), and spam sites that arguably do not reflect the mainstream web. Moreover, random sampling of the web is difficult in practice because of its decentralized nature [136, 51], and the Internet Archive's collection, though impressively expansive, does not include a dense history of the entire web.

As a baseline dataset, we collect the homepages of large public corporations of the Russell 1000 stock index, as it existed in 2018 when we began the study. Indices such as Russell are commonly used to gauge the health of the overall economy, and we reasoned that this sample might represent overall web trends, albeit biased towards corporate designs.

As a check against that bias, we verify our findings with two additional datasets based on different selection criteria: Alexa rankings [7] and Webby Award nominated URLs [251]. Alexa's rankings of websites have been used extensively, but its methodology is proprietary and has changed over time: it originally counted page loads through a custom user-installed browser toolbar, but now uses a variety of data. Historical Alexa rankings are also incomplete and the Internet Archive does not have regular snapshots of the full rankings. Webby Award-nominated URLs belong to websites nominated for an award in any “Websites” category. Since the structure of the Webby awards has changed over time, this is also an inconsistent sample, and only includes websites owned by individuals or organizations who pay an entry fee. In total, our three datasets constitute 227,802 images of 10,482 websites.

Verifying our findings on separate corpora collected with different selection criteria helps reduce biases related to any one dataset, and allows us to compare design trajectories of these different sets as well. We use Russell as our baseline dataset, since its selection criteria are publicly available and consistent (unlike Alexa and the Webby awards described above, which use proprietary or subjective criteria), but we replicate our main results on all three. We believe that variety in collection methodology and the datasets themselves lends robustness to our findings.

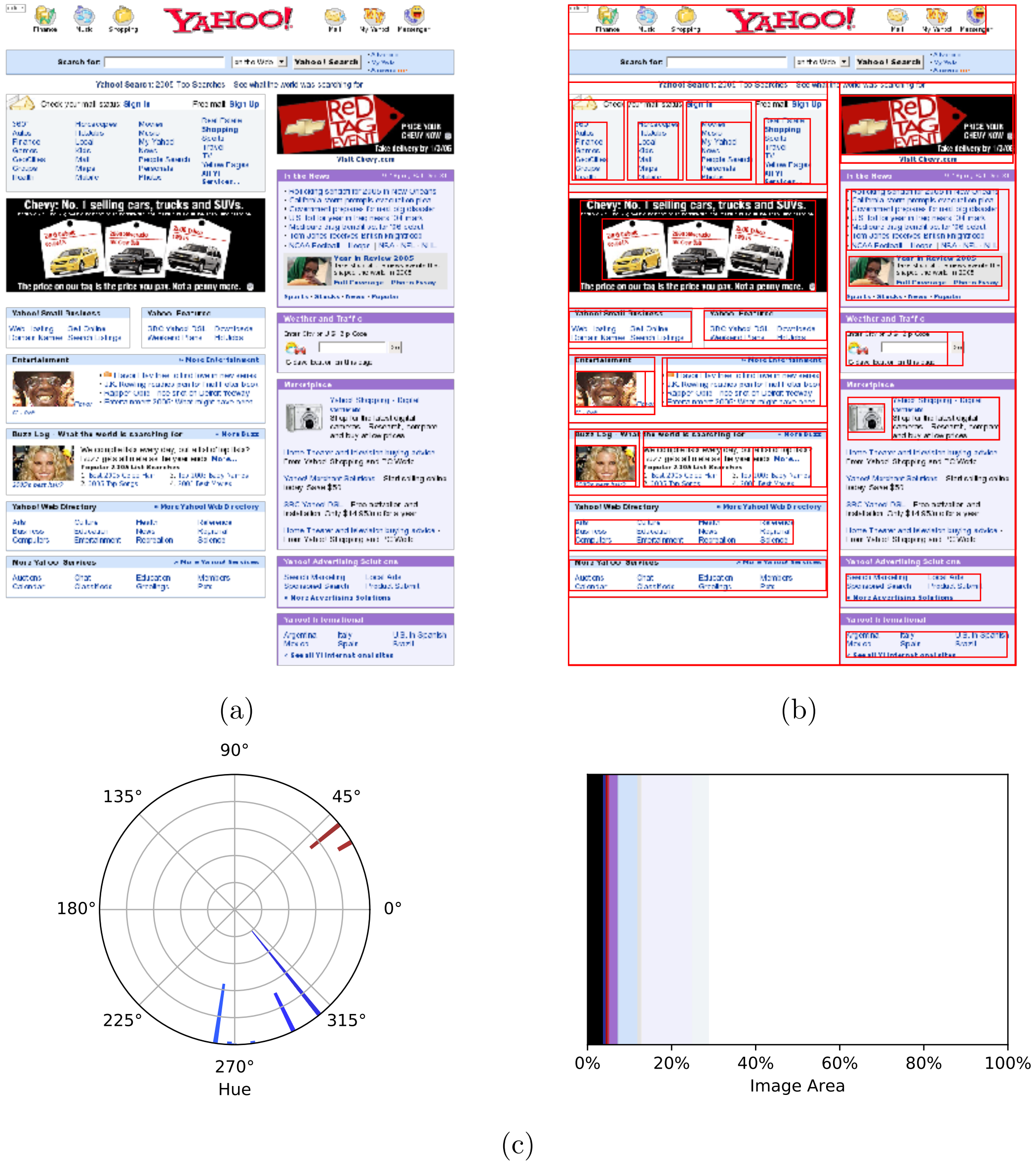
For each of the above three corpora of URLs, we use the Internet Archive [8] to fetch historical snapshots of the source code every 15 days, as available, from 2003–2017. For the Alexa and Webby sets, we do not look at the entire history of each site, but rather use snapshots within one year of each year that it appeared on the URL list. This makes this dataset reflect the changing population of popular and award-winning sites. Once we have downloaded the front-end source code for each site at each point in time, we render it as a 1200x1200 pixel image using PhantomJS [9].

Unfortunately, the Internet Archive does not have a dense historical collection for each website, and some sites have been indexed much less frequently than others. When working with data which is noisy and irregular, it is possible to find “trends” that are actually artifacts of methodological changes in the ranking and archiving processes. We thus conduct some experiments on what we call the *Dense Russell* dataset, which is a more complete dataset consisting of 100 websites of the Russell-1000 dataset which are available at least once every 15 days over the entire period 2003-2017. (This is not the same as the Russell 100 stock market index). Most of the 100 companies are *Consumer Services* (N=31) or *Technology* (N=29) according to the Russell 1000 categorization, while 15 are *Finance*, 13 are *Capital Goods*, 5 are *Energy*, 3 are *Health Care*, 2 are *Transportation*, and 2 are *Basic Industries.*

Color and Layout Metrics

Edit distance metrics are mathematically simple approaches to measuring similarity and difference. These metrics rely on the intuition that difference within a set of documents can be measured by defining a set of edits with associated costs and finding the minimum cost edit sequence required to transform one document into another. Originally proposed by Levenshtein [204], the edit distance between two binary strings is defined as the minimum number of reversals, insertions and deletions required to transform one string into another. The optimization problem inherent in this definition lends itself to a dynamic programming solution, and thus can be computed efficiently and exactly. These properties have led edit distances to be generalized to natural language strings [330] as well as other data structures.

For our hand-engineered distance metrics, we make the modeling assumption that the visual design of a site is captured by two major characteristics: color scheme and spatial layout. While other research [264] has focused on collecting specific numerical features



(a)

(b)

(c)

Figure 3.1: An example of our hand-engineered visual representations: (a) a sample website from the Alexa dataset, (b) nodes of the XY-tree decomposition, visualized as red boxes overlaid on the raw image, and (c) circular hue histogram and area plot (inspired by visualizations [28] showing the website's color distribution as fractions of the X-axis).

which capture these characteristics, we define distance metrics on our data representations directly. By using ideas from early query-by-image systems that search for images with similar color [267] and layout [227], we maintain the interpretability of our metrics without the added complexity of enumerating and weighting all the ways that two website images could look similar to one another.

For color, we work in the CIEL*a*b* color space, rather than RGB. Instead of representing colors using red, green and blue values, L*a*b* uses lightness, green-red and blue-yellow axes. It also has a guarantee of perceptual uniformity, meaning that pairs of colors which are equally distant, according to the CIE 1976 delta E metric [79], will look similarly different to most human observers [1].

We represent color schemes using color histograms (Figure 3.1), and use the Earth Mover's Distance (EMD) for measuring distances between them. Intuitively, the EMD captures the minimum amount of "work" it would take to transform one website image so that its color histogram matches that of another website image, where one unit of work increments or decrements the value of one pixel color channel, much like an edit distance. Unlike simpler metrics like cosine or Euclidean distance which simply compare histograms on a bin-by-bin basis, the EMD incorporates the fact that nearby bins—e.g., similar colors—should be considered more similar to each other than distant bins. Specifically, we use the CIEL*a*b* color space and measure the distance between two *individual* colors using the CIE 1976 delta E metric. Our CIEL*a*b* histograms have $100 \times 256 \times 256$ bins. Figure 3.2

[1] the delta E problem, of measuring perceptual color similarity, is also an aesthetic phenomenon problem and falls victim to the same issues we are discussing. In fact, the CIE has released two updated metrics since 1976 which incrementally improve the fit to human perceptual data. We use the CIE's metric somewhat uncritically here, as its mathematical simplicitly matches with our goal of a mathematically elegant metric.

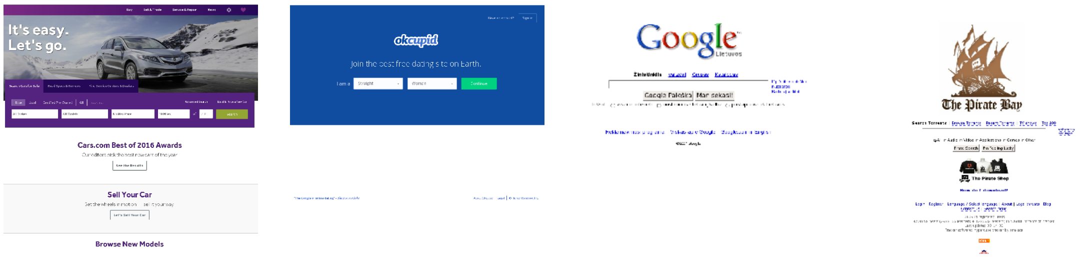
(right) shows an example of two non-identical sites with low color distance. The problem of finding the minimum cost transport between two histograms, which measures the value of the EMD, has a well-studied linear programming solution. We use the variant of EMD described in Rubner et al. [267] implemented by Pele and Werman [256].

We represent layout using XY-trees, which are created with a structure-based tree decomposition algorithm similar to that proposed in Ha et al. [135], although our implementation uses the algorithm from Reinecke et al. [264] (see Figure 3.1). The basic idea is to break up the page into a hierarchical structure of page elements by decomposing along gutters and solid colors of the page, creating a tree node for each nested image region. Trees also lend themselves to a type of edit distance, which in this case measures the amount of work needed to transform the layout tree of one image into another by inserting, deleting, or relabeling nodes. The cost of insertion or deletion is equal to the size of the region being inserted or deleted in pixels, and the cost of relabeling is the symmetric difference between the old and new regions (i.e. the number of pixels contained in either the old region or the new region but not both). We compute edit distances using an open-source implementation [154] of the Zhang-Shasha algorithm [332]. Measuring website layout difference using XY-tree edit distance is a novel contribution, to the best of our knowledge, though edit distances on XY-trees have been used before for document image retrieval [227]. Figure 3.2 (left) shows an example of two non-identical sites with low layout distance.

CNN Metric

We use deep learning with Convolutional Neural Networks (CNNs) to create a representation which automatically tries to quantify visual differences between websites. This representation is instead learned automatically from training data, and is not based on ideals of mathematical elegance like the color and layout metrics. To do this, we train a CNN model on a classification task: given a website image (taken at some point in the period 2003–2019), identify which of 100 companies each page belongs to. Rather than train on the full Russell dataset, we train on a 100-company subset because CNN models are known to be sensitive to class imbalance [53], so it is important that each of the classes has a roughly equal number of snapshots. We train such a classifier not to accurately solve the prediciton problem, but instead to learn a set of features which are helpful for reasoning about the branding of web pages [283]. We use a classification task, rather than something like an autoencoding [313] or contrastive learning [165] task, in order to maintain consistency with the earlier work of Doosti et al. [90].

In more detail, we train a CNN model with 273 randomly-sampled images from each of the 100 company websites, each resized to 227 × 227 pixels, which is standard for CNN approaches. We use a canonical model architecture, AlexNet [193], the same model used in [90] except that the final fully-connected layer has 100 outputs, corresponding to the 100 classes of our classification problem; please see [193] for network details. Each output is a number between 0 and 1 that indicates the estimated similarity to that class (website). We use these outputs as a feature vector, and quantify the difference between two



Similar sites by layout distance

Similar sites by color distance

Figure 3.2: Sample pairs of websites from the Alexa dataset, having *left:* low layout distance, because both have a single large element across the top half and minimal content below, and *right:* low color distance, since they both are mostly white with blue and black text.

website images as the Euclidean distance between their vectors. Since we are using our deep learning model not as an accurate classifier but as a "data mining model" that lets us characterize any website using a feature vector, we use an established and well-studied model architecture rather than newer, more complex techniques.

3.3.2 Qualitative Methods: Identifying Changes in Web Design Practices

Additionally, since images only hold a piece of the story of a website's design, and automated computer vision on large-scale image collections is a relatively blunt instrument, we also conduct a series of semi-structured interviews with experienced web designers. These interviews provide important historical context and link the computational trends to, for examples, specific changes in tools, techniques, and practices used in web design, as well as societal trends.

We sought out professionals and semi-professionals with at least 15 years of web design experience via snowball sampling over email and social media. We recruited 11 participants,

Participant ID	Web Design Experience	Job Title
P1	25 years	Coordinator of Instructional Design
P2	20 years	Retired Freelance Web Designer
P3	20 years	Systems Engineer
P4	14 years	Digital Editor & Web Designer
P5	20 years	Lecturer & Freelance Web Designer
P6	25 years	Vice President of Digital Marketing
P7	27 years	Graduate Student
P8	25 years	UX Strategist, Designer & Trainer
P9	22 years	Java & Web Development Trainer
P10	27 years	Web Accessibility Officer
P11	26 years	Web Designer & Joomla Certified Admin.

Table 3.1: Interview participant information.

listed in Table 3.1, of which 7 (64%) were men and 4 (36%) were women. Six (55%) were in North America, while 2 (18%) were in Europe, and 1 (9%) was in each of South America, Asia, and Oceania, respectively. Semi-structured interviews were conducted remotely over the Zoom videoconferencing platform between May and July 2020, and each interview lasted between one and two hours. Before each interview, the participant was asked to prepare a portfolio of 4–7 representative websites in which they had been involved in the design process. To avoid confirmation bias, we did not pose direct questions regarding any hypothesized homogenization of web design; rather, the semi-structured interview protocol focused on identifying factors which shaped the participants' design decisions regarding the visual appearance of their chosen websites. The portfolios proved useful in grounding participant stories in concrete details of their web design practices [280, p.88]. All interviews were transcribed and anonymized.

Analytic Approach

We analyze our interview transcripts using a constructivist grounded theory approach [58]. Transcripts were coded by the authors to identify emergent themes with a focus on tools, processes, technologies, and historical events which shaped the designs of specific sites. Constructivist grounded theory was particularly apt because its inductive approach draws from both a literature review (e.g., work by [61]) and multiple focal points in data, such as our computational analyses. We initially coded interviews with open codes, then with axial codes focused on concepts which may relate to broad trends in visual design practices. Some codes described the tools that played key roles in the design process (e.g., "JQuery," "Adobe Photoshop," "Wordpress") while other codes described professional, cultural, and technological shifts in the design process (e.g., "negotiation," "designing for web vs. print," and "mobile-first design"). Throughout the interviewing and analysis, an iterative process of memoing led to a final set of themes, which was the genesis of our reported qualitative findings.

3.3.3 Limitations

Our study has several limitations. Though we use two different modeling approaches for measuring the visual design similarity of pairs of websites, and compare results across three datasets, we cannot claim our measures of similarity are definitive. The CNN distance is taken from a model used to classify website identities and is not guaranteed to measure any particular visual qualities, although we find that it tends to correlate with our interpretable color and layout distances (see Section 3.4.1). All of these metrics constitute approaches to the aesthetic phenomenon problem of measuring design similarity between webpages and are necessarily imperfect and uncertain, but their agreement with one provides evidence towards their reliability.

Also, as we discussed above, it is difficult to collect a representative sample of the web—or even to define what a representative sample should be. We use three different datasets with different selection criteria to try to avoid artifacts of any single one, but our selections are nevertheless biased towards the English-language, and primarily American, web. Future work is needed to study how homogenization trends spread internationally and across language barriers.

Though our interview participants were involved with the designs of several of the specific websites in our datasets, we cannot say for certain that their experiences explain the trends we observed through our computational analyses. Our selection criteria, which prioritized designers who had experienced the whole temporal range of our sample, does not capture the changing nature of web design as a profession.

3.4 Results

We now turn to presenting the results of our mixed-methods analysis. We start by introducing our central finding: according to our measures, visual similarity on the web has been increasing, particularly since 2007. We then describe our investigations of potential underlying causes for this homogenization in layout and color as well as the significance of the time period from 2007. We organize these findings into three sections which investigate the relationships, respectively, between layouts and software libraries, color and image

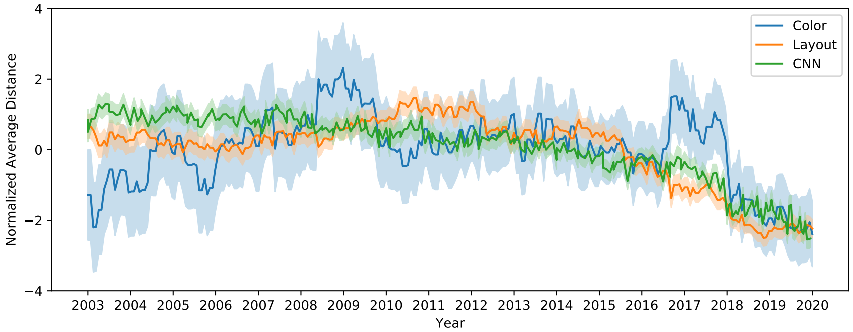


Figure 3.3: Average pairwise distance between the Dense Russell websites, plotted as a function of time, for color, layout, and CNN features. Shaded regions show 95% confidence intervals. To allow comparison across the three metrics (which have different dimensions), each plot was normalized to 0 mean and unit standard deviation. Lower values indicate more homogeneous.

content, and the 2007 catalyst for mobile support and responsive design.

3.4.1 Homogenization in the Dense Russell Subset Over Time

We begin our analysis by examining the Dense Russell subset (cf. Section 3.3.1), which includes 100 websites for which snapshots are consistently available every 15 days in the Internet Archive. For each 15-day period, we computed the color feature for each of the 100 websites, and computed the average Earth Mover's Distance (EMD) between all pairs of these sites. We also computed the average Euclidean distance between CNN features as well as the average tree distance between layout features at each time period. Figure 3.3 presents the results:

- Sites become less homogeneous (higher values) between 2003 and 2007 and more homogeneous (lower values) afterwards.

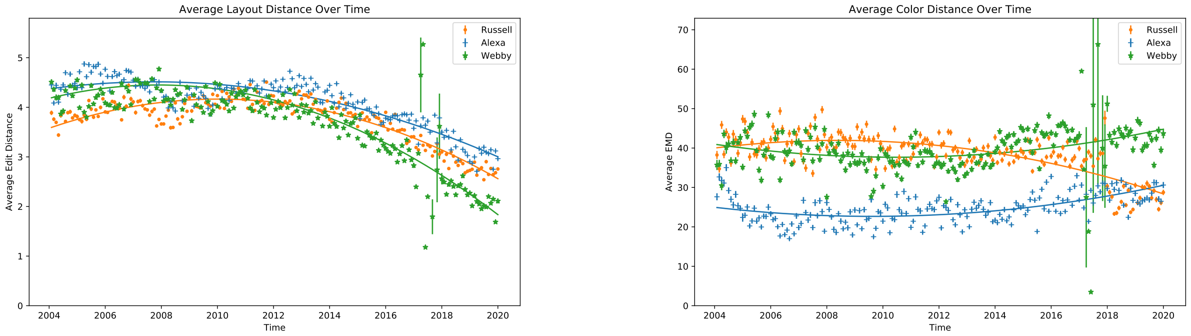


Figure 3.4: Average layout (left) and color (right) distances in all three datasets, with 95% confidence intervals and quadratic regression lines. All layout distance trends and the Russell color distance trend are negative with $p < 0.001$. The Alexa and Webby color distance trends are positive with $p < 0.001$. Large confidence intervals in the Webby data in 2017 are due to inconsistent data availability.

- Color distance declines 32% between 2008 and 2019.
- Layout distance declines 44% from 2010 to 2019.
- The CNN distance roughly follows the color distance metric with a decline of 30% between 2007 and 2019.

Overall, this suggests that *websites have homogenized since 2007, and that layout in particular has seen a significant decrease in diversity.*

We emphasize the large scale of this analysis: though it only contains 100 pages, each page has a snapshot every 15 days for 17 years (408 snapshots per site), for a total of about 40,800 snapshots. At each temporal snapshot, each of the pages is compared to every other ($\frac{100\times99}{2} = 4950$), for a total of about 2 million comparisons (4950×408).

As mentioned in Section 3.3.1, we curated additional datasets, collecting according to different criteria, in order to verify our computational results. For both the Alexa rankings and Webby Awards dataset, we randomly chose a subset of 100 sites that had data available for each month, computed the pairwise distances between them for each month, and plot the results over time with confidence intervals and trend lines in Figure 3.4. Notably, all of the layout trend lines have negative slope, indicating homogenization over time, and the layout trend for the Webby data is particularly steep. On the other hand, the Alexa and Webby datasets show different color trends, with Alexa decreasing in homogeneity and the Webby dataset remaining relatively consistent across time.

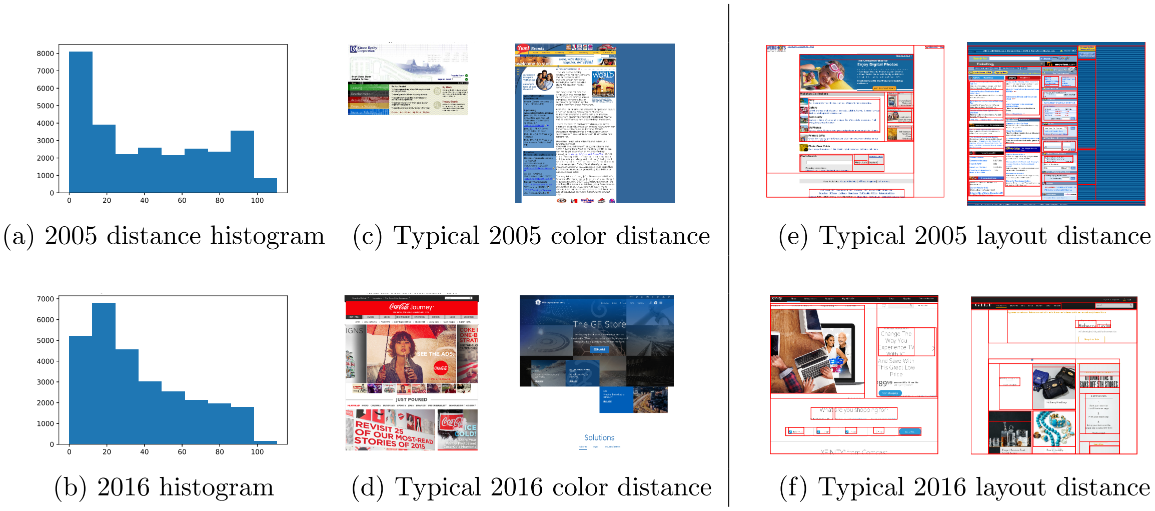
These results confirm the downward layout distance trend we observed with the Dense Russell subset, though each dataset has slightly different micro trends. The color distance plot highlights the noise in the Alexa and Webby data: since the set of highly-ranked Alexa websites changes over time, the mean similarity jumps up and down month to month. Meanwhile, the small number of Webby categories causes large confidence intervals in early years. The significantly lower color diversity of the Alexa data makes sense given the average color distributions in Figure 3.8: the Alexa websites have much more whitespace and fewer colored backgrounds than websites from the other two datasets.

One might argue that our trends align because they are cueing on the same image characteristics. To address this concern, we explored the relationship between our three distance metrics using multiple regression. The color and layout distances have no correlation with each other, as expected. We found that though the CNN distance is correlated with both color and layout distance, the coefficients are relatively small (0.2 for color and 0.18 for layout, respectively (both $p < 0.001$)), indicating that our CNN distance is incorporating additional information beyond color and layout. The CNN has access to far more information about the image, e.g., it can use texture and shape features which do not appear in either the color or layout representations.

3.4.2 Layout Similarity and Libraries

Our analysis of our large-scale website corpora in the last section showed evidence of homogenization, especially in the website layouts. But what is driving this effect? In this section, we find evidence that increasing dependence on libraries, frameworks, and content management systems (CMS) in web design spurred layout homogenization. Moreover, over different periods of time, particular libraries have exerted an oversized influence on layout across the web. We also find that increasing expectations for responsive, accessible, and usable sites has made it increasingly difficult to create unique, complex layouts that web designers of the past once could.

To give some intuition for our measure of layout similarity and how it has changed over time, Figure 3.5(e) shows a sample pair of websites from 2005 whose layout distance is the average distance in 2005, and (f) shows a sample pair in 2016 whose layout distance is the average of 2016. The red rectangles show the decomposition found by our layout analysis. As seen in these "prototypical" examples, early sites often use box-based layouts with fixed size and shape, while later layouts often have flat designs with fewer boxes and more images, presumably allowing them to scale to different sizes more easily.



(a) 2005 distance histogram (c) Typical 2005 color distance (e) Typical 2005 layout distance

(b) 2016 histogram (d) Typical 2016 color distance (f) Typical 2016 layout distance

Figure 3.5: Visualization of average color and layout distances in 2005 versus 2016. (a) and (b): Histograms of color distances in 2005 and 2016, respectively. Note that this is a histogram of distances between color distributions, with units of image area × CIE color distance. The observed change between 2005 and 2016 is due to a shift from colored backgrounds to off-white and image backgrounds. (c) A sample pair of images with the average color distance between pairs of sites in 2005. (d) Same, for average color distance in 2016. (e) and (f) Same, with average layout distance in 2005 and 2016, respectively.

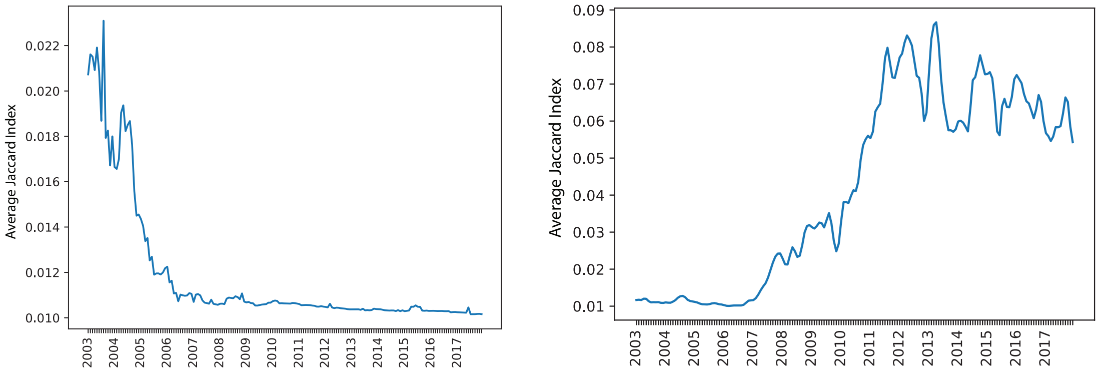
Dependence on Packaged Code

One hypothesis for the increasing similarity of website layouts is that the rise of libraries, frameworks, and content management system (CMS) templates makes it easy to create new sites with a given predefined "look." Indeed, many of the participants in our interviews mentioned the impact that these have had on the design of the web. For example, participants P6, P8, and P11 pointed to CMS templates allowing users without web design expertise to produce high quality sites. P1, P4, and P6 described instances where they made use of code snippets copied from other sites on the Internet, because they were not experienced enough with Javascript or it was simply easier to appropriate existing snippets.

Participant P5 identified an influx of software engineers into web design around 2012-2014 who over-use framework defaults due to a lack of design experience.

To test the hypothesis that reuse of website source code and/or the use of libraries, frameworks, and CMS templates may be driving the increasing website layout similarity, we performed analysis directly on the source code files of our Dense Russell dataset. We first compared the similarity of source code files themselves. For every pair of pages at a given time point, we used a string matching algorithm to compute the length of the substrings in common between the two source code files over the sum of their lengths. Figure 3.6(a) presents the results. Surprisingly, we found that the content of websites (including their native CSS and JavaScript code) have become *less* similar over time, despite that visual appearance has become more similar. One possible explanation is that as the demands for online information content increase, the diversity and complexity of that content increases as well. Another explanation is that over time the amount of native code has decreased, and instead web designers have increasingly used libraries instead of writing (or copying) native code.

To investigate library usage, we used Wappalyzer [10] to detect libraries and back-end platforms from front-end web page source code for each page at each time point of our Dense Russell dataset. The tool detects 101 JavaScript and 47 CSS libraries. For every possible pair of websites at each time point, we computed the Jaccard similarity of the sets of libraries they use (i.e., the ratio of the size of intersection of the two sets over the size of the union). This metric ranges from 0 if two pages do not share any libraries in common or do not use any libraries at all, to 1.0 if they use exactly the same libraries. As



(a) Similarity of code

(b) Similarity of library usage

Figure 3.6: Average pairwise similarity between the Dense Russell websites, plotted as a function of time, according to (a) similarity of front-end source code, and (b) similarity of library use. We see that the source code itself has become less similar over time, while websites have used more and more libraries in common.

Figure 3.6(b) shows, average similarity according to library usage has increased over time, especially after 2007, mirroring the results we found for visual similarity; the correlation between our CNN's measure of visual similarity and our measure of library similarity is 0.77 ($p < 0.001$).

Taken together, these results suggest that the increasing homogeneity of the visual web is not caused by increasing similarity of website content, but by increasing homogeneity in the choice of libraries that web designers use.

Library Monopolies

If increasing use of common libraries is causing websites to look more similar, then specific software library overlaps should predict for a decrease in visual layout distance. We perform multiple regression to predict layout similarity using binary features indicating whether two websites share each of the 16 most common libraries. The regression coefficients in Table 3.2 suggest that, yes, particular libraries are associated with increased similarity of web page layouts. Bootstrap correlates strongly with decreased layout distance, relative to other libraries. Other libraries, like SWFObject, which is used to embed Adobe Flash content, and jQuery tools, a Javascript user interface library, correlate with more varied layouts.

Several of our interview participants commented on the relationship between software tools and source code similarity. P5 and P10 commented on the low quality and complexity of code that has been auto-generated by a development environment (like Dreamweaver), CMS, or framework, indicating that the rise of libraries and frameworks may be contributing to the drop in code overlap. P4, P6, and P10 emphasize that these tools are essential time-savers to handle the complexity of the web, and P5 identifies a split between highly technical and non-technical tooling for the web: *"We have tools where we're splitting the middle. We have tools that are getting easier and easier... to use... where people can build websites with no knowledge of code whatsoever and communicate their ideas and live with whatever limitations those technologies put in place. And on the other end, if you want to get into the professional side of this, it is unbelievable the barriers to entry that have been put up."*

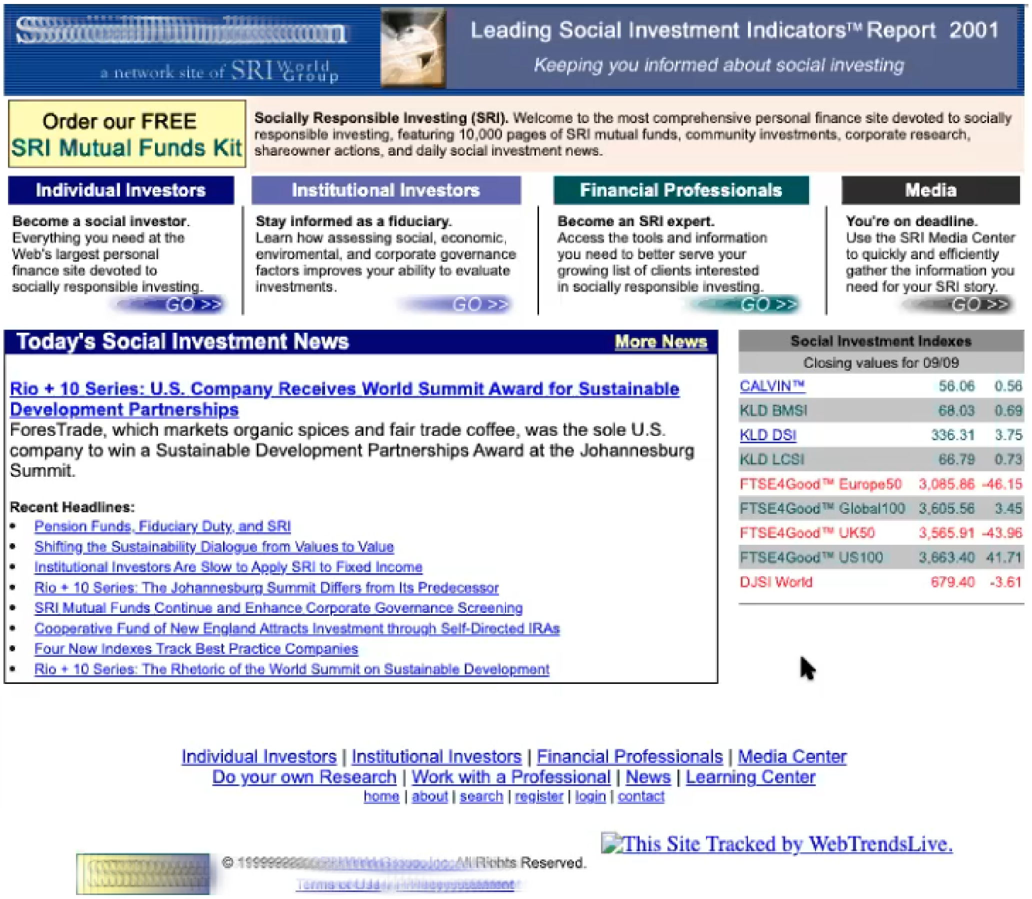
Shifting Layout Practices

Participants also discussed changing best practices in web design related to layout. For example, they recalled some of their considerations during the earlier days of the web: P1, P8, and P11 discussed using fixed-width wallpapers and background images for layout, P3 and P9 referenced a desire to keep content "above the fold"—within the first screenful of content (Figure 3.7)—so that users did not have to scroll, P5 described using search engines like Yahoo and AltaVista, which have highly complex layouts, as design examples, and P4, P5 and P9 described a process of designing a fixed layout in Photoshop and then "slicing" it into small images to put into an HTML table. These practices all but disappeared with the advent of CSS, which allowed sophisticated and reusable layout and styling of text content, and which was better for load times and search engine optimization.

In addition to saving time, libraries and frameworks often promote better usability and accessibility — a common topic in six of our interviews (P1, P2, P6, P7, P8, P10). P10 emphasized the deep relationship between accessibility, usability, and resilience: *"Make your designs as resilient as possible... They should work if somebody wants to blow the screen up to 300%, the interface should still work as intended."* Unlike the earlier mindset inherited from print design, web designers now see themselves designing flexible interfaces, not documents. P6 discussed that a web framework *"comes with a lot of accessibility code built in. So it's got ARIA [an accessibility standard for web applications] and stuff like that."*

3.4.3 Color Similarity and Technological/Cultural Constraints

While it stands to reason that web frameworks would influence the layout of websites, it is less clear if this would explain shifts in color, given that frameworks allow color to be customized. So how can we explain the homogenization of color schemes observed



a network site of SRI World Group

Leading Social Investment Indicators™ Report 2001

Keeping you informed about social investing

Order our FREE SRI Mutual Funds Kit

Socially Responsible Investing (SRI). Welcome to the most comprehensive personal finance site devoted to socially responsible investing, featuring 10,000 pages of SRI mutual funds, community investments, corporate research, shareowner actions, and daily social investment news.

Individual Investors

Become a social investor. Everything you need at the Web's largest personal finance site devoted to socially responsible investing.

GO >>

Institutional Investors

Stay informed as a fiduciary. Learn how assessing social, economic, enviromental, and corporate governance factors improves your ability to evaluate investments.

GO >>

Financial Professionals

Become an SRI expert. Access the tools and information you need to better serve your growing list of clients interested in socially responsible investing.

GO >>

Media

You're on deadline. Use the SRI Media Center to quickly and efficiently gather the information you need for your SRI story.

GO >>

Today's Social Investment News More News

Rio + 10 Series: U.S. Company Receives World Summit Award for Sustainable Development Partnerships

ForesTrade, which markets organic spices and fair trade coffee, was the sole U.S. company to win a Sustainable Development Partnerships Award at the Johannesburg Summit.

Recent Headlines:

- Pension Funds, Fiduciary Duty, and SRI
- Shifting the Sustainability Dialogue from Values to Value
- Institutional Investors Are Slow to Apply SRI to Fixed Income
- Rio + 10 Series: The Johannesburg Summit Differs from Its Predecessor
- SRI Mutual Funds Continue and Enhance Corporate Governance Screening
- Cooperative Fund of New England Attracts Investment through Self-Directed IRAs
- Four New Indexes Track Best Practice Companies
- Rio + 10 Series: The Rhetoric of the World Summit on Sustainable Development

Social Investment Indexes

Closing values for 09/09

CALVIN™	56.06	0.56
KLD BMSI	68.03	0.69
KLD DSI	336.31	3.75
KLD LCSI	66.79	0.73
FTSE4Good™ Europe50	3,085.86	-46.15
FTSE4Good™ Global100	3,605.56	3.45
FTSE4Good™ UK50	3,565.91	-43.96
FTSE4Good™ US100	3,663.40	41.71
DJSI World	679.40	-3.61

Individual Investors | Institutional Investors | Financial Professionals | Media Center
Do your own Research | Work with a Professional | News | Learning Center
home | about | search | register | login | contact

This Site Tracked by WebTrendsLive.

© [illegible] All Rights Reserved.
[illegible]

Figure 3.7: A screenshot from one of our interviews regarding layout practices. P5 explains, *"The boss was very, very concerned with above the fold... And so not a lot of white space and a lot of words and a lot of text that's a lot the same size."*

Library	Normalized Coefficient	N pairs
Bootstrap	-0.410	287
Font Awesome	-0.190	340
jQuery UI	-0.187	598
Underscore	-0.184	54
Moment.js	-0.173	25
Modernizr	-0.129	1489
ZURB Foundation	-0.007	52
Scriptaculous	0.055	31
yepnope	0.067	470
Prototype	0.089	62
jQuery	0.118	12523
React	0.199	36
Lightbox	0.204	23
SWFObject	0.373	1719
jQuery Tools	0.562	76

Table 3.2: Normalized regression coefficients for library overlap and layout distance. Negative numbers indicate that the presence of this library makes it more likely that two websites look similar, positive numbers indicate the opposite. N Pairs indicates the number of site pairs that had that library in common out of 65703 total pairs.

in Figure 3.3? In this section, we observe that the homogenization in color schemes is related to the shift from monochromatic, usually white backgrounds to more off-white backgrounds and image backgrounds, and we find that this shift comes from a combination of technological and cultural constraints.

By plotting the full histogram of color distances, rather than just the average, we observe that, year by year, the color distances start as a bimodal distribution with modes around 0 and 90, then converge to a single mode around 10 (see Figure 3.5). As backgrounds have a large number of pixels, they cause the distance to swing heavily towards low values when both sites have similar backgrounds. Photos, on the other hand, have many different

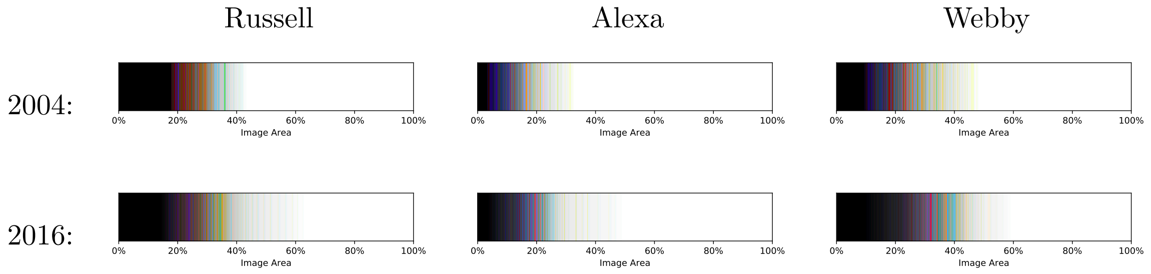


Figure 3.8: Average color distributions for three datasets in 2004 vs. 2016, in which bar width indicates the average fraction of each image made up of pixels of that color. In each dataset, the amount of dark and off-white increases and the amount of white and black decreases during this period.

colors of pixels, so their distances are less polarized. We can also look at the average color distributions across time (Figure 3.8) and observe that off-white and dark colors have increased while black and white have decreased, which supports the notion that off-white and photo backgrounds are more common.

Interview data connects color distribution changes to changing technological and cultural constraints. In terms of technological constraints, Participants P5 and P9 identify the "web-safe palette," an early design constraint to ensure that web colors would appear correctly on different monitors. P9 mentions, "*I tend to compose colors in multiples of 51. So I would say, oh, that's a 51, 51, 102. Yeah, I still tend to do that despite the fact that we don't need to use web safe colors anymore.*" P10 points to early differences between monitor gamuts (the range of colors that can be displayed) as a reason to avoid color on the early web. P4 and P5 mentioned that bandwidth limitations prevented them from using much image content in the early days of the web. Since images were large files, they would load slowly over poor connections and negatively impact user experience. P7 identifies CSS and

SVG as prerequisites to expressive use of color in web design, since they allowed a large degree of control over visual presentation without significantly increasing page load times.

Cultural constraints come from the changing nature of website design. P1 and P7 describe how, in the early days of the web, their stakeholders like clients or organizational leadership generally weren't interested in the details of the website. As the web has grown, these stakeholders have become increasingly invested in the look and feel of their organization's web presence, and website design and redesign has become a long, negotiated processes. Five participants (P1, P2, P3, P4, and P10) referenced negotiations spread over weeks or months focused on visual details like typeface or accent color during website redesigns. P1 describes the process for a website redesign in 2007: "*They had focus groups. They come in and they meet with large numbers of people and talk about their needs. And then they do a sitemap which kinda lays out the content, how it's going to be organized. And then... they gave us like five options of color schemes... And then our administration and faculty would basically vote on it.*" Because redesigns are more expensive and attract attention from stakeholders, designers are under more pressure to comply with branding guidelines and avoid unusual color combinations. P10 explains how a background color choice was made: "*This sky background kind of came out of the branding efforts to do with wings... And there was, this kind of illustrated the transition from IT running the show to marketing and communications running the show.*"

Wireframes — i.e., visual prototypes — are essential for facilitating these long design cycles. P2 describes wireframing with pencil and paper in 2004 while P8 references wireframing with software tools like InVision and Axure regularly from 2011 to the present.

These wireframes function as drafts: they serve as a fixed point and ground discussions through the design process for specialists who may not have the time or expertise to edit the source code directly.

3.4.4 Catalyst for Mobile Support and Responsive Design

We observed that the abrupt change in web design visual similarity after 2007 (Figure 3.3) coincides with the release of Apple's first-generation iPhone. We hypothesized that this is not coincidental: the rapid need for mobile support may have caused the web to grow more homogeneous as companies began using similar front-end libraries to support mobile platforms. In this section, we show that our findings do suggest that visual similarity of website design has moved in tandem with increasing mobile support. Mobile design and therefore responsive design became de rigueur in the industry, pushed in part by search engine optimization (SEO) strategies.

To track mobile support in company websites over time, we analyze the CSS code in each web page's source code. If any of the CSS code has the "`@mobile`" keyword, we flag that website as supporting mobile screens in their design. The correlation between the average website CNN similarity in each month and the corresponding rate of mobile support is 0.84 ($p < 0.001$), which suggests that websites which support mobile tend to look more similar than websites which do not.

Our interview data strongly supports this relationship between mobile support and visual homogenization. Participants P5, P6, and P10 pointed to the rise of mobile web browsers as a major factor in their decision to adopt the philosophy of responsive design (i.e., where content “responds” to the size and shape of the browser, rearranging to best use the screen space, usually with a CSS framework). While any web page design can be made responsive by hand, doing so requires tedious work specifying content rearrangement and resizing rules using CSS media queries, and it is difficult from a design perspective to create layouts that are visually appealing in several arrangements. P5 and P9 point towards Apple’s design decisions not to support Flash and emphasize scrolling over other forms of navigation on the iPhone. An increasing share of web traffic coming from mobile devices has led to a rise in “mobile-first” design practice that several participants (P2, P6, P8, P9, P10) would recommend to new designers. Interestingly, all five of these participants were quick to clarify that they do not personally use mobile-first practices in their design.

Four participants (P2, P4, P6, and P9) reference search engine optimization (SEO), and Google’s policies in particular, as factors in their work. While not claiming to be “SEO expert[s],” they identified strategies for improving SEO. P4 said that he stopped using Flash for interaction because Google would not index text inside Flash components. P4 and P6 point to Google search’s 2015 “Mobilegeddon” update, which suddenly prioritized search results which would display well on mobile, as the reason the web adopted responsive design. This shift corresponds with the second wave of library adoption in Figure 3.6. P2 and P9 say they use alt tags on images to improve SEO and P6 referenced using Google’s Accelerated Mobile Pages framework as a new SEO strategy, describing Google as “*the 800-pound gorilla at this point. They’re driving a lot of what happens on the web in terms of design and stuff.*”

3.5 Discussion

We set out to investigate a very straightforward question — has visual design on the web become more homogeneous? Answering this question turned out to be a much larger-scale endeavor than we had imagined. We first collected a large-scale set of over 200,000 historical snapshots (over 15 years) of over 10,000 websites. To avoid bias associated with any single selection criteria, the dataset consisted of websites selected from three different sources: corporate sites of the Russell 1000 stock index, winners and nominees for the Webby awards, and top-visited sites according to Alexa. We then developed novel computational methods for measuring and characterizing the similarity of website images. To avoid bias associated with any single measure, we developed three, one based on deep learning and two hand-engineered features that characterize color and spatial layout features, respectively.

Across datasets and metrics, this large-scale, quantitative analysis showed strong evidence that website designs—especially with respect to page layouts—have become more similar, starting between 2007 and 2010 and continuing until at least 2016. To understand why, we analyzed, again at large scale, the similarity of source code, the use of libraries, and support for mobile devices. We found that the use of a relatively small number of frameworks and libraries has expanded significantly, and that the use of similar libraries strongly correlates with visual similarity—suggesting that the rise of these tools may be contributing significantly to the homogenization that we observed.

But this large-scale quantitative analysis could not reveal what was causing the uptake of libraries, or other more subtle factors that might be driving homogenization. We thus recruited and interviewed 11 web design professionals, each having at least 15 years of experience, to reflect on the changes in their design process over time. These qualitative data contextualized our quantitative results and confirmed or introduced several explanations, including the rise of software libraries and frameworks, increased use of large images, and support for mobile devices.

Layout homogenization appears to have begun after the release of the iPhone. Our interviews suggest that as mobile browsers grew in market share, responsive design and the libraries and frameworks that support them became an essential part of professional web design. This transformation was driven by libraries promoted by major tech companies, such as Twitter's Bootstrap and Font Awesome which is incorporated into the BootstrapCDN (content delivery network), and jQuery and Modernizr which are both included in templates like Microsoft ASP.Net MVC. Responsive Web Design has changed how we experience the web [157] and largely replaced older best practices like placing content "above the fold." As developing for the responsive web requires more technical expertise, many designs shifted closer to library and framework defaults to avoid time-consuming, difficult development work.

Color homogenization offers a less clear story. Histogram data (Figures 3.1 and 3.8) suggests a shift from colored backgrounds to off-white backgrounds featuring images. Rather than being driven by the display limitations of new mobile devices, these changes respond to lifting bandwidth restrictions and increased support for CSS and SVG. While homogenization was the trend in the Russell 1000 sites, the opposite trend was true for the Alexa and Webby datasets, indicating that the shift to off-white backgrounds and large images does not consistently lead to color homogenization on less corporate websites.

We choose to reserve judgment on whether the trend of visual homogenization is good or bad. We suggest, however, that if the diversity of visual designs is indeed shrinking, this may limit the perceived repertoire of possible and legitimate designs that future website designers draw from, constraining the creativity and innovation of future websites.

Two decades ago, studies by Newman & Landay [244] noted that web designers used many informal, flexible representations of websites (e.g., sketches) for wireframing; these unconstrained, sometimes low-fidelity representations afford usual and unusual designs equally, limited only by the designer's imagination. Designers turned to media like paper for practical reasons: creating polished, high fidelity prototypes at the time was laborious.

Now, designers have a myriad of tools, including wireframing tools like InVision, Figma and Axure, CMS like Wordpress and Joomla, and frontend libraries and frameworks to rapidly prototype high fidelity websites. These tools allow designers to search for or choose from exemplar widgets and styles that follow the current landscape of *legitimate* designs and studied interaction patterns [310, 201], at the cost of making unusual designs more difficult to create [150]. The bias towards legitimate designs is helpful for both amateur and professional designers: using such tools and strategies leads to faster development cycles, reduced complexity, better accessibility, and use of learned affordances to achieve greater usability.

However, as HCI takes up other concerns such as value-sensitive, authentic, and reflective design [111, 296, 282], we may rethink how visual design (and closely related ideas about interaction) can go beyond concerns of usability, efficiency, or even marketing. We do not suggest turning back progress on, for example, creating inclusive websites or open-source software, but we should investigate how to enable enjoyable, diverse, and/or provocative forms within the space of inclusive design. For example, though animated GIFs may violate accessibility guidelines, the aesthetic qualities that make them popular indicate that we should not prematurely limit the creation of novel designs for accessibility reasons [120].

Turning to our broader discussion of subjective measures and cultural analytics, we see examples of several broad trends in these sorts of projects:

- A seemingly-monolithic entity, in this case the web, turns out to be difficult to define. Different perspectives on how to study the web "in general" lead to different conclusions (for example, 3.4 right).
- A simple perceptual property, visual design similarity, is quite difficult to measure. Different criteria for metrics (mathematically simple vs. data driven) lead to different metrics, with subtly different results.
- Qualitative approaches yield important context which help us unify the results of contrasting computational methodologies. Some trends, like layout homogenization, occur across these differences in perspective.

Ultimately, when computational methods operate within a qualitative epistemology, the inherent subjectivity of the underlying problem, and the issues with any one implementation, cease to present a serious problem. Different data collection strategies and different metrics offer contrasting partial perspectives, much like the contrasting perspectives of different research participants. By acknowledging the inherent limitations of each metric and interview, we can interpret and synthesize them into a narrative about the development of web design practices, which is more strongly grounded and supported by evidence than one based on qualitative or computational methods alone.

Chapter 4

A Probabilistic Model of Template-based Color Harmony

4.1 Introduction

In some circumstances, when subjectivity plays a factor in computer vision problems, a reasonable solution is to express subjective factors as a source of random variation, and take a probabilistic modeling approach instead. When we do so, probabilistic models, which can account for many types of noise due to unobservable factors, become very appealing. In this chapter, we examine a probabilistic model of color harmony which uses real datasets of art and design images as a proxy for the distribution of popular preference for color schemes in context.

Color harmony is a fundamental topic in color theory which asks the question: why are some combinations of colors more appealing to humans than others? A color harmony model predicts whether a given combination of colors will look appealing or not. Color harmony is largely inspired by the success of formal models of musical harmony [13], where a small-number ratio like $\frac{1}{2}$ or $\frac{2}{3}$ between the frequencies is a strong predictor of consonance. Since there is a long-standing association between musical tone and color in the western scientific tradition (Isaac Newton, for example, famously believed that sound and light were the same phenomenon operating at different timescales [253]), scholars across the arts and sciences have sought to discover models which predict pleasing color schemes. These

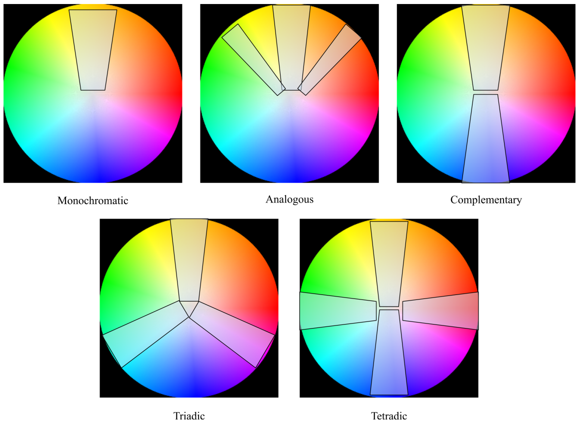


Figure 4.1: Five common color templates used by online color scheme generators. A color scheme from this template chooses colors within the shaded area, rotated to any center hue.

models range from highly qualitative theories such as Itten's system of seven contrasts [161] to more recent quantitative models like Matsuda's hue templates [230].

In recent years, color harmony models have inspired a number of online color scheme generators, which automatically choose colors for designers. Some generate harmonious color schemes stochastically [291, 36, 261] while others ask the user to choose a single color, then derive other harmonious colors (e.g., Adobe Color [6]). These generators choose colors using a set of color templates, shapes on the color wheel like the ones shown in Figure 4.1. These models rely on what we will refer to as the *hue-invariance hypothesis*, that certain angles on the color wheel produce harmonious color schemes regardless of hue. This claim is analogous to the way harmony works in music, as chords are consonant or dissonant regardless of the root pitch. Not all online color scheme generators subscribe to the hue-invariance hypothesis: some, like the ColourLovers platform, allow users to create, share, and search for color schemes without any underlying model [235].

Despite the elegance of these models and the convenience of online generators, studies of human color harmony preferences [298, 250] have shown that template-based models are not very accurate descriptions of human aesthetic preferences for color combinations: preferences vary between individuals and depend on the hue values (not just their relative positions in color-space) [250].

For this reason, we are interested in modeling template-based color scheme generators, both to reverse-engineer existing generators and extract the templates they use, as well as to examine image datasets and determine the extent to which they are template-based. Rather than explore these issues using data collected about abstract color schemes, in this chapter, we ask:

RQ2 — To what extent are color schemes extracted from real website, fashion and art images based on the same templates as online color scheme generators?

Based on the work of O'Donovan et al. [250], we hypothesize that color scheme models fit to template-based online color scheme generators will not assign high likelihoods to the color combinations present in real art and design images, but color scheme models based on human-designed color schemes will.

To investigate this topic, we fit Gaussian Mixture Models (GMMs) to hue-normalized representations of color schemes, and then using the resulting model to compute the likelihood of color schemes contained in images. While conceptually simple, we show that this approach is nevertheless powerful enough to uncover trends in real-world forms of visual design. In particular, we apply our model on large-scale, time-referenced datasets of three very different types of visual artifacts — websites, fashion and artworks — and show how it distinguishes between highly template-based and non-template-based color schemes, and uncovers clear temporal trends in all three domains. To our knowledge, we are the first to present a practical technique for quantifying color combinations independent of dominant hue. Please note that while these models seem similar to the IAQA models discussed in Chapters 6, 7 and 8, they are measuring likelihood of occurring in a training set, rather than likelihood of being classified as high quality.

A preliminary version of this chapter was presented
at CVPR CVFAD Workshop (Publication P1)

4.2 Methods

Our model assumes that an image is generated by sampling from a hierarchical mixture model. Intuitively, we are seeking to expose the underlying logic for which color combinations are present in a dataset of images, where each image is a collection of color-coordinated objects (like clothes, webpage layout elements or painted shapes) depicted by pixels with similar color values. We can model the low-level aspects as a mixture model over the pixels of the image, and the high-level aspects, which are our ultimate interest, as a mixture model over the color schemes in the dataset.

We approach this modeling process in a step-by-step maximum likelihood manner. Specifically, we need to extract the principal colors from the image, order those colors, convert them to a representation which is invariant to lightness, chroma, and hue translations, as well as lightness and chroma dilations, and apply a probabilistic model to those features via maximum likelihood estimation. If only a subset of the original image is of interest,

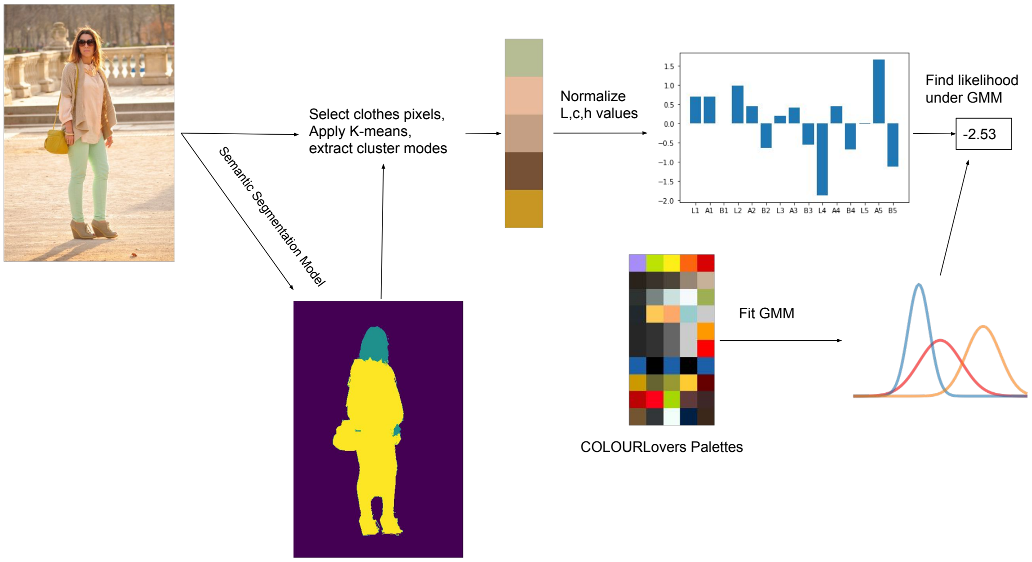


Figure 4.2: Our approach to compute the likelihood that a fashion color scheme came from a given color scheme dataset. We use the same process, without semantic segmentation, for website images.

as with fashion images, we apply semantic segmentation first to isolate the pixels which correspond to clothing. We now describe each of these steps.

Extracting Color Schemes: To extract color schemes from images, we use weighted K-means clustering on the image pixels, where the weight for each pixel is equal to its chroma value and $k = 5$ for consistency with O'Donovan et al. [250] and the Paletton generator [291]. Despite the existence of more complex methods [207, 83], we found that simply taking the (chroma-weighted) mode of each cluster served as an effective representation of the color scheme of an image.

Stabilization: In order to cluster together color schemes which were generated by the same color templates, we need a stable, hue-invariant representation which captures the relative positions of each color, while being invariant to translation or dilation of lightness, chroma, or hue. To this end, we apply the following stabilization procedure. Consider a color scheme $X = x_1, ..., x_k$ where each $x_i = l_i, a_i, b_i$ is a point in CIEL*a*b* colorspace (i.e. the cluster modes from the K-means clustering), and colors are ordered from most to least frequent in the original image. We express the chroma $c_i = \sqrt{a_i^2 + b_i^2}$ and hue $h_i = \text{atan2}(a_i, b_i)$ (two argument inverse tangent). We compute the hue-invariant lightness, hue, and chroma,

$$l'_i = \frac{l_i - \bar{l}}{\sqrt{\frac{1}{k}\Sigma_j^k (l_j - \bar{l})^2}}, \quad c'_i = \frac{c_i}{\bar{c}}, \quad h'_i = (h_i - h_1) \mod 2\pi,$$

where $\bar{l}$ and $\bar{c}$ indicate the mean lightness and chroma, respectively, guaranteeing a chroma mean of 1, principal hue of 0, lightness mean of 0, and standard deviation of 1. Finally, we convert back to rectangular coordinates, to avoid the hue discontinuity around 0.

$$a'_i = c'_i \cos(h'_i), \quad b'_i = c'_i \sin(h'_i)$$

This approach has two important consequences. First, it ensures that the patterns we observe show whether the images adhere to the template model and that they are unaffected by general trends in hue, chroma, or lightness, such as a trend towards lighter colors or more reds. Second, it makes color templates (e.g., Figure 4.1) linearly separable. We find that when tested on synthetic data generated from five templates, a softmax regression model trained to distinguish between color templates improves from 39% to 69% accuracy.

Modeling: To capture the different kinds of color templates used in a dataset, we use a Gaussian Mixture Model (GMM) over the $3k$ dimensional normalized features. Note that the modal regions of the distribution correspond to common color patterns on the color wheel, not colors themselves, and thus high likelihood examples exhibit common color patterns and low likelihood examples do not. We use Expectation Maximization to fit GMMs with 10 components.

Semantic Segmentation: For fashion images, in order to reason about the colors of clothes pixels, rather than background or skin pixels, we employed a semantic segmentation model [212] trained on the CFPD dataset [211].[1] We reduced the labels in the dataset from 22 to 3 classes: (1) skin, face, hair, and sunglasses, (2) background, and (3) all remaining labels. Our analysis is only conducted on pixels from the third class. Our model achieved 93% accuracy on the CFPD test set for this greatly simplified clothes parsing problem.

Datasets: We fit our model to three color schemes datasets, and then observe color scheme trends in three image datasets: website images, fashion images and art images. Our color scheme datasets include our own synthetic template-based color schemes which use the five templates in 4.1 ($n = 70,000$), the ColourLovers dataset ($n = 383,938$, sampled down to $n = 70,000$) of human-uploaded color schemes from [250] and schemes we scraped from colormind.io ($n = 70,000$), which generates schemes using a deep learning approach, inspired by the pix2pix architecture of Isola et al. [160, 261]. Our dataset of website images ($n = 50,232$) consists of screenshots from the Alexa top 500 US websites from 2004 to 2016 which we collected using the Internet Archive.[2] Our dataset of fashion images is the Street Fashion Styles dataset collected from posts to the website Chictopia between 2009 and 2016, collected by Gu et al. [131] ($n = 27,087$). Our dataset of art images is a random 25% subsample of the Wikiart dataset ($n = 39,393$). We filter the WikiArt dataset to works

[1] We adapted code from `https://github.com/minar09/Fashion-Clothing-Parsing`

[2] `https://archive.org`

dated between 1750 and 2020, and randomly subsample to bring its size in line with the other two test sets.

4.3 Results

After fitting our model to the three color scheme datasets, we can visualize the kinds of schemes each model prefers by visualizing the values of the means, and examining the most likely and unlikely test examples. The means are shown in Figure 4.3. Since they exist in a hue-invariant representation, we visualize their values for four hue rotations at a neutral lightness and chroma. While the Template and Colormind models learn a variety of multi-hue color schemes, the ColourLovers model only fits to monochromatic schemes with different combinations of lightness and chroma. As the first two datasets are based on color models with random elements, and the ColourLovers dataset is based on user submitted color schemes, we have effectively reproduced the result regarding the poor alignment between human preferences and template-based color schemes from [250].

We visualize several of these examples from the Wikiart and SFS datasets in Figure 4.4. The three models, despite being trained on different sources, assign high and low probability to the same sorts of color schemes: simple monochromatic and analogous color schemes have high probability and schemes with an out-of-place color are assigned low probability.

Our analysis of website, art and fashion images is shown in Figure 4.5. Comparing the models to one another, we see that the trends are somewhat similar from model to model, providing further evidence that they have learned similar concepts. The ColourLovers and

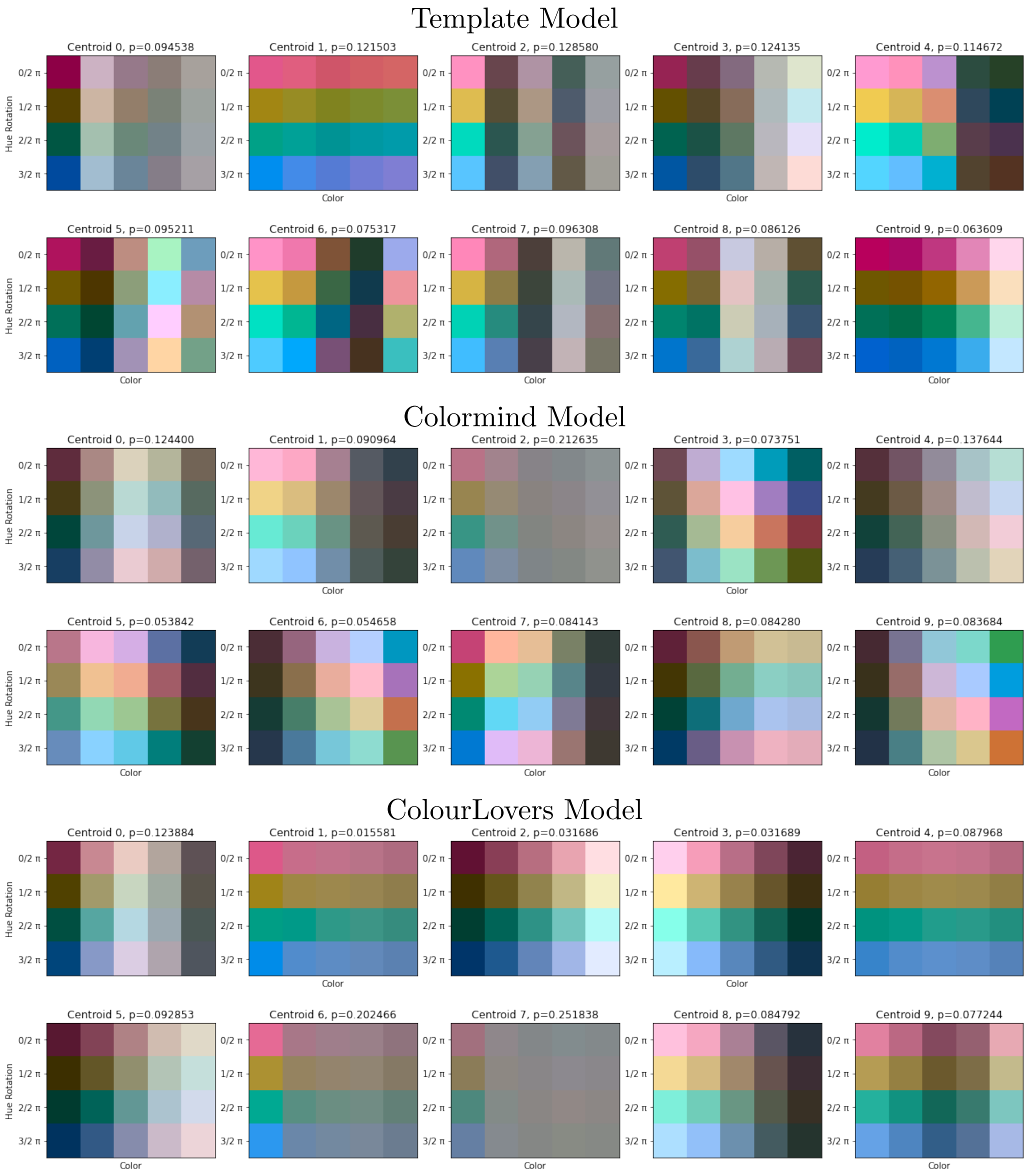


Figure 4.3: Learned means of our three GMMs. Since models are fit in a hue-invariant (and lightness/chroma normalized) representation, we visualize the mean at four hue rotation values.

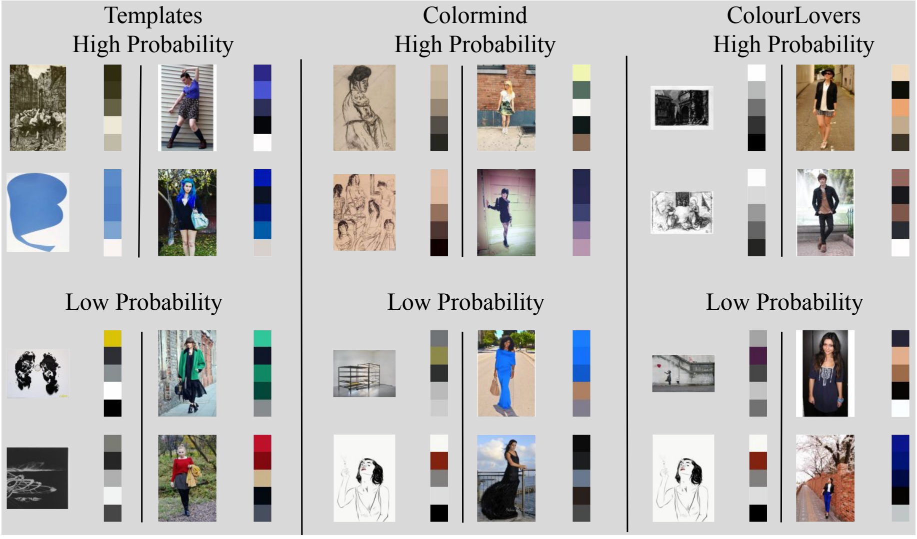


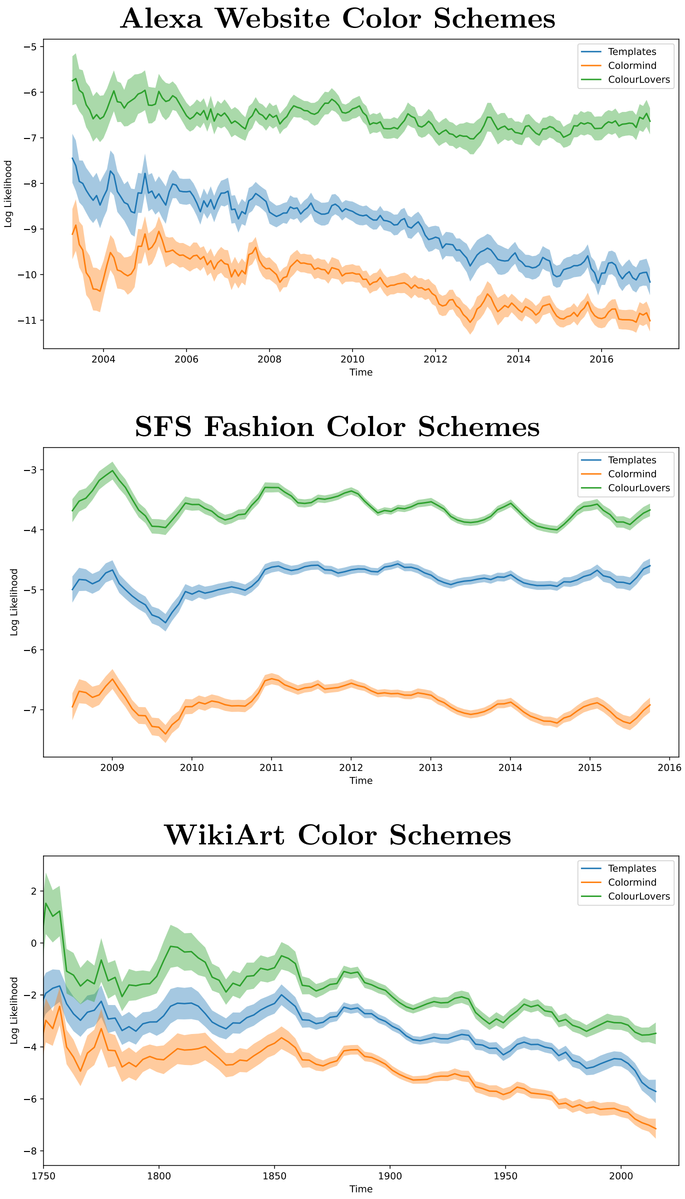
Figure 4.4: Examples of high and low probability color schemes under each model, drawn from WikiArt (left) and SFS Fashion (right). Each model assigns high probabilities to simple monochromatic and analogous color schemes, and low probabilities to color schemes which contain an unusual color element.

ColorMind models assign consistently higher and lower likelihood to color schemes, respectively, indicating that they have learned more or less diffuse definitions, respectively.

Additionally, we observe several trends in the model likelihoods over time:

- The Alexa data remains relatively consistent under the ColourLovers model, but decreases in likelihood over time under the other models. This trend corresponds to the shift described in Chapter 3 away from color backgrounds towards off-white backgrounds and more image content in websites starting in 2008.

- The WikiArt data generally decreases in likelihood over time as well, though from the examples we can see some of that decrease in likelihood is due to an increase in both photographic images and vibrant, synthetic colors. As WikiArt is a highly



Alexa Website Color Schemes

SFS Fashion Color Schemes

WikiArt Color Schemes

Figure 4.5: Average log likelihood for each three month sliding window of images in each of the Alexa and SFS datasets, and for each three year sliding window in the WikiArt dataset, according to each of our Gaussian mixture models. Shaded area shows standard error for each mean.

biased collection, we do not recommend reading closely into the micro-level changes.

- The SFS data remains relatively constant through the data period, and displays a degree of seasonality, with likelihoods increasing in winter months when color choices are more muted.

4.4 Discussion

In this chapter, we have presented results from applying probabilistic models of template-based color schemes, learned from abstract color scheme generators, to three art and design image datasets. We found, consistent with the results of O'Donovan et al. [250], that templates are not a good model of the color schemes which occur in real art and design images — models learned from human-provided color schemes assign higher likelihoods.

Interestingly, our models assign higher likelihoods to art images than fashion or website images. This trend makes intuitive sense: historically, the theory of template-based color schemes is inspired by studies of color theoretic practices used by fine artists [230, 161]. While fashion and website color schemes may use colors for a variety of design reasons, works of art are deploying color for largely aesthetic reasons.

Our work here comes with a variety of limitations and caveats. Particularly, there is potential for racial bias arising from skin/clothing parsing algorithms like the one we use, especially when color is the subject of research. While we did not notice significantly different accuracy based on race, we note that the CFPD dataset appears to over-represent American, European, and Asian women, and does not contain demographic labels to measure performance bias precisely. Similarly, the WikiArt dataset we study only includes art images which were preserved, photographed and deemed significant enough to upload to WikiArt, which is a highly nonrepresentative sample, primarily containing European art, across a variety of genres and media.

Within our larger investigation, this study demonstrates the potential for using color scheme data collected at scale from images as a proxy for popular color scheme preferences. Even though the aesthetic quality of color schemes is highly subjective and contextual, making it difficult to evaluate via perceptual studies, large image datasets can show us which color schemes individuals find good enough to use, and allow us, under a fair degree of uncertainty, to model the underlying aesthetic phenomenon as it manifests in context.

Chapter 5

Finding Historical Periods in Collections of Paintings: A Bayesian Approach

5.1 Introduction

A central topic in art history is the division of artists, paintings and styles into historical periods. While once the organizing principle of the discipline, periodiziation has fallen out of favor in recent decades [184]. One argument against artistic periods comes from Ernst Gombrich, who connects periodization to a Renaissance notion of progress and almost-scientific improvement in style over time, which invites ahistorical politicization, including narratives of cultural supremacy [122, Ch. 1]. Additional criticisms highlight the simplistic, essentializing nature of historical periods — artists do not all produce one style at a time and do not suddenly shift from style to style, except in the rare case that the artists understood their own work in terms of such broad historical structures, like some Florentine artists of the Renaissance [272]. Such criticisms have led contemporary art historians to specify and study individual works, artists and movements in historical context, without using a general system of broad historical periods [184].

Responding to these criticisms, some cultural analytics scholars have advocated a return to thinking about art broadly, viewing stylistic change over time in terms of its underlying network dynamics [274]. Particularly, Manovich argues that digital methods, used to study culture at scale, can avoid the earlier problems with classification and periodization [223].

Computational approaches afford a full view of art history, using continuous representations to embrace complexity and avoid essentializing narratives, categories or exemplars.

Between the two extremes of highly specific art history and broadly general cultural analytics, we are interested in finding a middle road which captures the benefits of both. In this chapter, we explore one possible approach by returning to the notion of periodization: looking at specific collections, and finding the measurable visual boundaries which exist within them. However, learning from critiques, we do not seek to split works of art into discrete categories. Instead, we propose a probabilistic approach, conceptualizing periods as fundamentally uncertain and fluid, using tools from Bayesian statistics.

While uncertainty is a major component of feminist approaches to data analysis [91, 94, 87], to the best of our knowledge our work is the first which considers a feminist motivation for the Bayesian perspective on probability and subjective belief. Bayesians view the probability of a statement as a measure of our degree of subjective certainty, not a measure of a random event. Subjective certainty, in this sense, is defined as the betting odds that we would accept as fair if we were to bet on whether a statement is really true [143].

Bayesian analysis follows the logic of Bayes rule, a probabilistic law proposed by the 18th century cleric Thomas Bayes [143]. Bayesian modeling holds that all of our beliefs about the world are fundamentally uncertain and subjective, but provides an objective process for updating them based on new data. In this chapter, we apply this view of modeling to artistic periods. We investigate the research question:

RQ3 — How can we use Bayesian probability to quantify uncertainty in automatic periodization of artwork images?

From this question, we emphasize that we are not answering art historical questions. We are interested in developing a modeling approach. We test our model using art data from an incomplete and biased data source, WikiArt, and do not seek to make art historical claims based on the results. We hope both our general approach and specific model can enable future art historical inquiry using more reliable data sources.

We propose segmenting a collection of images into subsets with distinct means in a visual feature space over time. Depending on the specific features, these models can find many different distributions over period boundaries, which can be combined with prior estimates based on historical characteristics. To enable this kind of art image analysis, we demonstrate an efficient algorithm based on dynamic programming to compute periodizations and their conditional posterior distributions.

5.2 Related Work

A variety of recent scholars have used scalar or vector measures of art images from the WikiArt dataset. Elgammal and Saleh quantify the historical creativity of a work of art by constructing a temporal graph of works and measure the likelihood of later works under the feature distribution of earlier ones [96]. Saleh et al. apply a similar framework to identify work-to-work and artist-to-artist influences through history [269]. De La Rosa and Suárez measure facial attractiveness in art images over time [78]. Several recent papers have applied information-theoretic measures: Sigaki et al. quantify the entropy and complexity of the art images of each period [287], Desikan et al. use information theory to devise measures of style and color [289] and Karjus et al. apply ensembles of compression complexity measures to study the complexity of paintings over time [181]. A variety of recent works have investigated different features and feature learning methods for art images; see Castellano and Vessio for a recent literature review [56].

Approaches to offline changepoint detection and estimation have been well-studied in statistics. The problem formulation, estimating a signal which changes discretely through time using only noisy estimates, was proposed at least as early as 1964 by Chernoff and Zacks [62], and the method we use is very similar to the one proposed in [327]. As this kind of problem occurs in many kinds of signals, similar algorithms have been recently applied to data as disparate as oil prices [57] and intramuscular electromyography [301]. This technique is less frequently used in computer vision; however, it has been recently applied to detect changes in traffic flow patterns using image data [166].

5.3 Methodology

5.3.1 Data Model

We assume a set of painting images $I_1, I_2, \ldots, I_N$ sorted in chronological order. We assume that the artist or collection of artists went through a progression of Π periods with distinct visual styles throughout their career, where Π is a known constant. Then each painting I_i comes from period $p_i \in \mathbb{Z}^+$, where $p_1 = 1$, $p_N = \Pi$ and $p_i \leq p_{i+1}$ for all $1 <= i <= N$, and the probability of switching periods, $Pr(p_i = p_{i-1} + 1)$, is constant.

We also assume that we have a feature extraction function ϕ with which we can compute visual features $y_i = \phi(I_i)$ for each i. The feature extraction process could generate a single highly interpretable measure, like the colorfulness measure used in Chapter 1, a more complex feature representation like the color schemes used in Chapter 4 or a feature space learned from data, like a pre-trained deep convolutional neural network [194]. We assume that within each period, the D dimensional visual features of each image y_i is drawn from a multivariate Gaussian with known diagonal covariance S and unknown mean,

$$y_i \sim \mathcal{N}(m_{p_i}, S)$$

where m_{p_i} is the unknown mean within the period p_i of image I_i. We assume that the period mean m_{p_i} is also drawn from a Gaussian,

$$m_{p_i} \sim \mathcal{N}(\mu, \Sigma)$$

Where μ and Σ are hyperparameters that depend on the choice of ϕ. Our task is to find the most likely $p_1, \ldots, p_N$ given $y_1 \ldots, y_N$ and Π, but unknown $m_1, \ldots, m_\Pi$.

5.3.2 Maximum Likelihood Solution

Using the above modeling assumptions, we can write the likelihood function as

$$\begin{aligned} & Pr(y_1 \ldots y_N | m_1, \ldots, m_\Pi, p_1, \ldots, p_N) \\ = & \prod_{i=1}^{N} Pr(y_i | m_{p_i}, p_i) \end{aligned}$$

We can estimate p_i by finding the values of $p_1, \ldots, p_N$ to maximize this function. Maximizing this likelihood is equivalent to minimizing a negative log likelihood,

$$\begin{aligned} & -\log Pr(y_1, \ldots, y_N | m_1, ..., m_\Pi, p_1, ..., p_N) \\ = & \sum_{i=1}^{N} -\log \mathcal{N}(y_i | m_{p_i}, S) \\ = & Z \sum_{i=1}^{N} (y_i - m_{p_i})^T S^{-1} (y_i - m_{p_i}) \end{aligned}$$

where S^{-1} is the inverse of the diagonal covariance S and Z is a constant that does not affect the maximization. Thus maximizing the likelihood is equivalent to choosing all of the p_i and m_{pi} to minimize the sum of square errors in each feature dimension, weighted by the corresponding value of S^{-1}, $\sum\limits_{i=1}^{N} \sum\limits_{d=1}^{D} \frac{1}{S_d}(y_{i,d} - m_{p_i,d})^2$. Since each m_{p_i} is estimated from y, we can rewrite this expression as a sum over periods:

$$\sum_{\pi=1}^{\Pi} \sum_{i \text{ s.t. } p_i = \pi} \sum_{d=1}^{D} \frac{1}{S_d}(y_{i,d} - \bar{y}_{\pi,d})^2$$

where $\bar{y}_{\pi,d}$ is the mean of each $y_{i,d}$ such that $p_i = \pi$. Since each $p_i \leq p_{i+1}$, we can restate that inner sum using a function:

$$SSE(i, j) = \sum_{k=i}^{j} \sum_{d=1}^{D} \frac{1}{S_d}(y_{k,d} - \bar{y}_{i,j,d})^2$$

where $\bar{y}_{i,j,d}$ is the mean of $y_{i,d}, \ldots, y_{j,d}$.

The optimal p can be computed using a dynamic programming algorithm, similar to an algorithm proposed by Richard Bellman for approximating curves using line segments [27]. Call our recursive subproblem $ML(i, j, n)$, which stores the minimum error of a segmentation between year i and year j with exactly n segment boundaries (i.e. $p_j - p_i = n$).

$$ML(i, j, 0) = SSE(i, j) \qquad (i \neq j)$$

$$ML(i, i, n) = 0 \qquad (n \geq 0)$$

$$ML(i, j, n) = \min_{i<k<j} ML(i, k, n-1) + ML(k, j, 0) \qquad (n > 0, i \neq j)$$

To compute the ML solution, fill in a table of values for ML, starting at the base cases for $i = j$ and $n = 0$ and compute the remaining values in order of ascending n from 1 to $\Pi - 1$. The ML solution for the full dataset is $ML(1, N, \Pi - 1)$.

5.3.3 Posterior and MAP Solution

A similar approach works for computing the maximum a posteriori (MAP) estimate, except instead of maximizing the likelihood, we maximize the posterior:

$$
\begin{aligned}
&Pr(p_1,\ldots,p_N|y_1,\ldots,y_N) \\
=&\frac{Pr(p_1,\ldots,p_N)}{Pr(y_1,\ldots,y_N)}Pr(y_1,\ldots,y_N|p_1,\ldots,p_N) && \text{By Bayes rule} \\
=&\frac{Pr(p_1,\ldots,p_N)}{Pr(y_1,\ldots,y_N)}\prod_{i=1}^{N}Pr(y_i|p_i)
\end{aligned}
$$

Similarly to the previous derivation, we group the product over N by period:

$$\frac{Pr(p_1,\ldots,p_N)}{Pr(y_1,\ldots,y_N)}\prod_{\pi=1}^{\Pi}\prod_{i\text{ s.t. }p_i=\pi}Pr(y_i|\pi)$$

and since the denominator is constant with respect to $p_1,\ldots,p_N$, its probability will be constant, so we can instead maximize a proportional expression:

$$\begin{aligned}
&\log Pr(p_1, \ldots, p_N) \prod_{\pi=1}^{\Pi} \prod_{i \text{ s.t. } p_i=\pi} Pr(y_i|\pi) \\
&= \log Pr(p_1, \ldots, p_N) \prod_{\pi=1}^{\Pi} \prod_{i \text{ s.t. } p_i=\pi} \prod_{d=1}^{D} Pr(y_{i,d}|\pi)
\end{aligned}$$

To compute $Pr(y_{i,d}|\pi)$, we must first marginalize over each $m_{\pi,d}$:

$$\begin{aligned}
&\log Pr(p_1, \ldots, p_N) \prod_{\pi=1}^{\Pi} \prod_{d=1}^{D} \int \prod_{i \text{ s.t. } p_i=\pi} Pr(y_{i,d}|m_{\pi,d}, S_d) Pr(m_{\pi,d}|\mu_d, \Sigma_d) dm_{\pi,d} \\
&= \log Pr(p_1, \ldots, p_N) + \sum_{\pi=1}^{\Pi} \sum_{d=1}^{D} \log \int \prod_{i \text{ s.t. } p_i=\pi} Pr(y_{i,d}|m_{\pi,d}, S_d) Pr(m_{\pi,d}|\mu_d, \Sigma_d) dm_{\pi,d} \\
&= \log Pr(p_1, \ldots, p_N) + \sum_{\pi=1}^{\Pi} \sum_{d=1}^{D} \log \int \mathcal{N}(y_{i,d}, \ldots, y_{j,d}|m_{\pi,d}, S_d) \mathcal{N}(m_{\pi,d}|\mu_d, \Sigma_d) dm_{\pi,d}
\end{aligned}$$

where $\{y_i, \ldots, y_j\} = \{y_i \text{ s.t. } p_i = \pi\}$, i.e. the features for images within period π.

The integral corresponds to the model evidence of the Gaussian distribution, which for a Gaussian prior has a closed-form solution [155]. As a subroutine to compute the evidence, we define a function Ev similar to SSE:

$$\begin{aligned}
Ev(i,j) &= \sum_{d=1}^{D} \log \int \mathcal{N}(y_{i,j,d}|m_d, S_d) \mathcal{N}(m_d|\mu_d, \Sigma_d) dm_d \\
&= \sum_{d=1}^{D} \log(\frac{1}{2\pi\Sigma_d})^{\frac{j-i}{2}} \sqrt{\frac{\lambda_d}{\lambda_{i,j,d}}} \exp(-\frac{1}{2}(\frac{y_{i,j}^T y_{i,j}}{S_d} + \lambda\mu_d^2 - \lambda_{i,j,d}\mu_{i,j,d}^2))
\end{aligned}$$

Where $\lambda_d = \frac{1}{\Sigma_d}$, $\mu_{i,j,d}$ is the mean of the posterior and $\lambda_{i,j,d}$ is the inverse variance of the posterior:

$$\lambda_{i,j,d} = \lambda_d + \frac{j-i}{S_d}$$
$$\mu_{i,j,d} = \frac{\lambda_d \mu + (j-i)\bar{y}_{i,j,d}/S_d}{\lambda_{i,j,d}}$$

similar to above, $\bar{y}_{i,j,d}$ is the mean of $y_{i,d}, \ldots, y_{j,d}$.

The optimal period assignment can be found via a similar dynamic programming algorithm, except instead of minimizing an expression in terms of SSE, we maximize an expression in terms of Ev:

$$MAP(i,j,0) = \log Pr(p_1, \ldots, p_N) + Ev(i,j) \qquad (i \neq j)$$
$$MAP(i,i,n) = 0 \qquad (n \geq 0)$$
$$MAP(i,j,n) = \max_{i<k<j} MAP(i,k,n-1) + MAP(k,j,0) \qquad (n > 0, i \neq j)$$

This table is computed in the same manner as the ML solution.

There are several ways we might extend this framework. For example, we may define more complex data models for each period which use other distributions or linear models of the year or other historical context. We also may adopt a non-uniform distribution for $Pr(p_i = p_{i-1} + 1)$ based on our prior knowledge about the likelihood of a period boundary in each year of the data. These extensions allow us to encode our subjective prior knowledge based on interpretation of historical context to be taken into account by the model.

5.3.4 Model Selection

Through the previous sections, we have not specified how we arrived at the number of period boundaries, Π. Call the model with Π period boundaries $\mathcal{M}_{\Pi}$. A full Bayesian treatment would compute the Bayes factor, $Pr(y|\mathcal{M}_{\Pi})$ for each value of Π [183]. However, that requires computing the likelihood of the data under each of $\binom{N-1}{\Pi-1}$ possible periodizations. We instead use the Bayesian information criterion (BIC), a value which captures the tradeoff between model complexity and error, for model selection [278]:

$$BIC(N) = \log ML(0, x_N, \Pi) - \frac{2\Pi - 1}{2} \log x_N$$

This equation assumes that the number of parameters in the model with $\Pi - 1$ period boundaries is $2\Pi - 1$, Π for the means and $\Pi - 1$ for the boundary locations. For higher dimensional data, that numerator becomes $\Pi d + \Pi - 1$ where d is the dimensionality of the data. We choose Π based on BIC values for the maximum likelihood case and use the same Π for the maximum a posterior case.

5.3.5 Results

The ML periodization is used to split the works of Mark Rothko in Chapter 1. Results for that data under two colorfulness metrics can be seen in Figures 1.1 and 1.2. While the ML approach produces a range of interesting results, it falls prey to the same kinds of criticisms as traditional art periods, as well as the criticisms of reasoning about art from necessarily incomplete data using uncertain metrics. The advantage of the Bayesian approach is its ability to express that uncertainty. Unfortunately, computing the full joint posterior over p is intractable, but we can compute the conditional posterior $Pr(p_i = p_{i-1} + 1)$ using the values of $MAP(\cdot, \cdot, \Pi - 2)$. In the figures below, we visualize discrete distributions over period boundaries, $Pr(\pi_1), \ldots, Pr(\pi_{\Pi-1}|\pi_{\Pi-2})$, where each π_a is the lowest index i such that $p_i = a$.

We evaluate this periodization method qualitatively by applying it to the work of single well known artists using a single scalar image feature, and visualizing the MAP periodization and conditional posterior distribution for each period boundary. When visualized, the effectiveness of this periodization algorithm is apparent. In this section, we show results for two artists: Mark Rothko and Pablo Picasso. Then, we show a multivariate approach which is able to identify periods in the Renaissance.

The Bayesian periodization of the Rothko paintings is shown in Figure 5.1. The conditional log-posterior is shown in the top plot (categorical at year resolution). The MAP periodizations are also shown as dotted vertical lines in the lower plot. From the conditional posterior, we can see that the period boundary in 1947 has a rounded peak, indicating that we are uncertain if the boundary should be there or in 1946 or 1948 instead, but we are relatively certain a boundary should fall in those years. The boundary in 1958 is more certain, and it is more likely to come earlier than later. The blue error bars show the uncertainty in our mean colorfulness model for each period, showing that the third period is much less certain than the previous two.

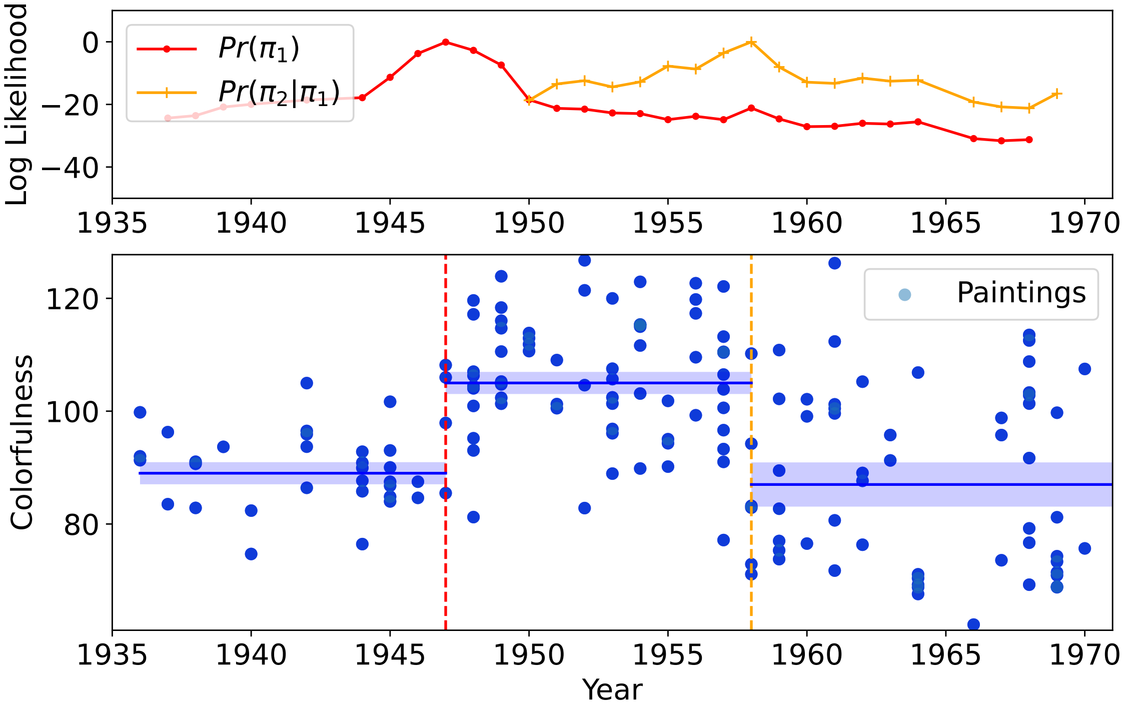


Figure 5.1: Periodization of the works of Mark Rothko, according to Hasler-Suesstrunk colorfulness. The lower plot shows the colorfulness and year of each painting, with dotted lines for the period boundaries, horizontal lines for the colorfulness model, with 95% confidence intervals. The upper plot shows the conditional posterior for each period boundary.

Looking at the work of another well-known painter, Figure 5.2 shows the works of Picasso on WikiArt, visualized using the mean color hue of the image:

$$H(I) = \begin{cases} 60 \cdot \frac{\bar{g}-\bar{b}}{d} \quad \mathrm{mod}\ 360 & \bar{r} > \max(\bar{g}, \bar{b}) \\ 60 \cdot \frac{\bar{b}-\bar{r}}{d} + 120 \quad \mathrm{mod}\ 360 & \bar{g} > \max(\bar{b}, \bar{r}) \\ 60 \cdot \frac{\bar{r}-\bar{g}}{d} + 240 \quad \mathrm{mod}\ 360 & \bar{b} > \max(\bar{r}, \bar{g}) \\ 0 & \text{o.w.} \end{cases}$$

Where $\bar{r}, \bar{g}, \bar{b}$ are the mean values of each RGB image channel scaled to unit range, respectively, and $d = \max(\bar{r}, \bar{g}, \bar{b}) - \min(\bar{r}, \bar{g}, \bar{b})$. Mean hue is an interesting metric to measure for the works of Picasso specifically because Picasso had a famous blue period in the years 1902–1904 and a rose period in the years 1905–1906, at the start of a well-documented, prolific and varied eight-decade artistic career [259].

The results of this analysis are visualized in Figure 5.2. The top left shows the works in question, the vertical spike towards the left corresponds to Picasso's well-known blue period. Much like our investigation of Rothko, the blue-dominant paintings do not strictly occur during the blue period. Some occur as early as 1899 (e.g. *La Chata*), or as late as 1907 (e.g. *Dance of the Veils*). The Bayesian information criterion yields similar values for periodizations with 3, 4, 5 and 6 periods, so we visualize all four. All of them find a boundary in 1906, and three of the four find boundaries in 1918 and 1930, indicating natural hue distribution changes in those years. The 6-period model interestingly identifies a short period in 1902–1906 with a hue mean further towards the blue and violet hues, matching the common narrative about Picasso in those years.

However, the boundaries between the latter periodizations are uncertain. The log likelihoods for neighboring years are relatively similar, especially in the years 1910–1930, when most of the mean hues are in the orange range, showing the uncertainty around period boundaries according to this metric. Similarly, confidence intervals around the means in each period show how our approach affords random omissions and errors in measurement.

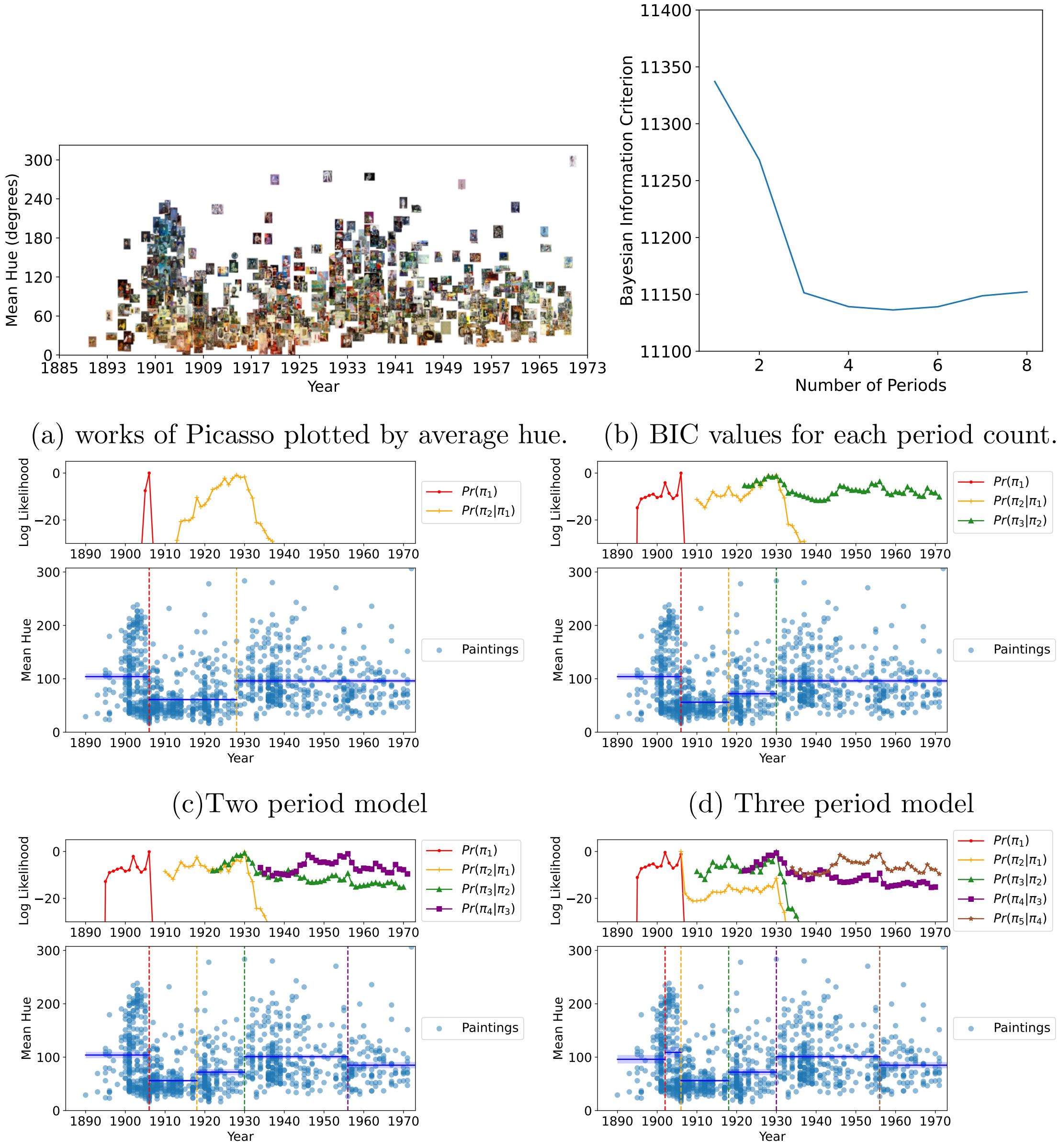
Finally, we turn to a multivarate example over a larger set of images: all the works on WikiArt dated between 1400 and 1700. This period highlights the traditional art historical period boundaries around the Renaissance, whose neighbors are often defined by terms of exclusion, like Gothic or Baroque [122, Ch. 8]. To keep the results visualizable, we use two scalar features: the Hasler-Suesstrunk colorfulness [144] and a measure of visual complexity proposed by Machado et al. the average value of the Canny edge detector on the grayscale image. Despite its simplicity, this metric achieves a Pearson $r = 0.76$ with human perception of visual complexity [220].

Using these 2-dimensional representations, we find the MAP solution with minimum BIC has five periods:

- (Start) 1400–1467: a period of high complexity and colorfulness, corresponding to the proto-Renaissance, including many manuscript illuminations.

- 1467–1492: a drastic drop in complexity and colorfulness corresponding to the early Renaissance, including some of the earliest works on paper on WikiArt.

- 1492–1527: a complexity minimum, exemplified by highly ordered works like the



(a) works of Picasso plotted by average hue.

(b) BIC values for each period count.

(c)Two period model

(d) Three period model

(e) Four period model

(f) Five period model

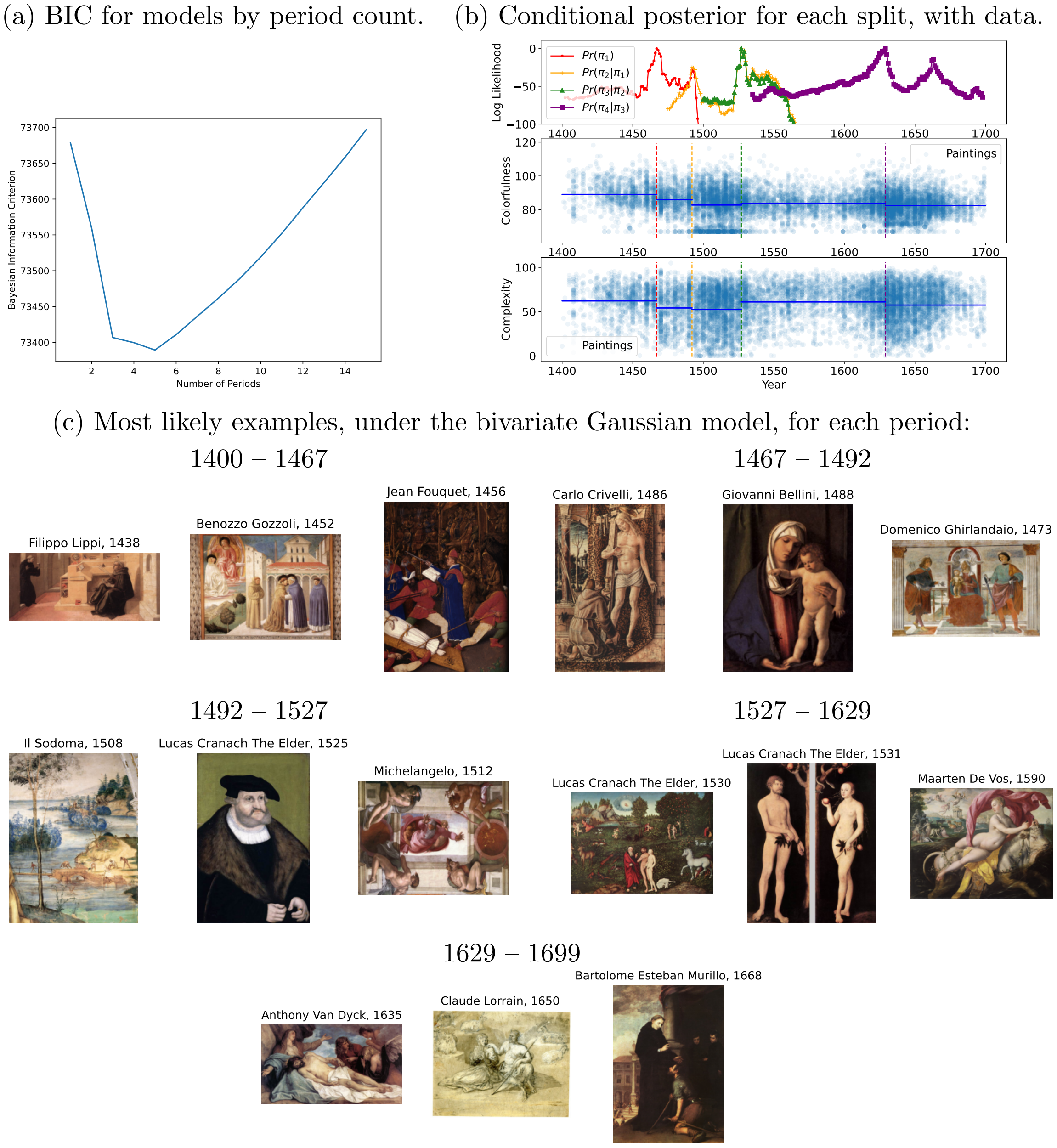
Figure 5.2: Four possible periodizations of the works of Picasso, according to mean hue.

Michaelangelo example. This period includes a large number of surviving works on paper.

- 1527–1629: a return to higher complexity without a corresponding return to colorfulness, including Mannerist paintings like the example by Maarten de Vos. This period contains significantly fewer paintings on WikiArt.
- 1629–1700 (End): a drop in both complexity and colorfulness, the Baroque. This trend is partially explained by the increased use of chiaroscuro, and the increasing number of East Asian paintings on WikiArt.

These periods show both the flexibility of our periodization method, as well as some of the limitations of working with specific features and data in this manner. Rather than simply splitting time into a single set of categories, we find that any number of periods between 3 and 7 are similarly informative for the data. Examining the four period model, our approach is able to capture specific distributional shifts which match traditional art historical narratives. The period boundaries are all relatively sharp unimodal distributions, indicating that the data actually shifts around that boundary, not gradually over time.

However, these sorts of analyses reveal that the distribution of artworks on WikiArt are driven primarily by the productivity and preferred medium of well-known artists — in years when artists whose work is well-documented in museum collections are productive, more works ends up digitized and uploaded. In this period, that means the work of Italian and Dutch artists is highly over-represented, and the visual culture of the rest of the world is drowned out.



(a) BIC for models by period count.

(b) Conditional posterior for each split, with data.

(c) Most likely examples, under the bivariate Gaussian model, for each period:

Figure 5.3: Periodization of all WikiArt paintings from 1400–1700. (a) shows the value of the BIC for each period count. (b) shows the minimum BIC periodization, with conditional posterior values for each boundary. (c) shows three examples closest to the mean colorfulness and complexity for each period.

5.4 Discussion

In this chapter, we have taken a Bayesian approach to artistic periods and presented a flexible probabilistic model which can be used to find either maximum likelihood periodization, or distributions over period boundaries. We applied this model to the work of two prolific and well-documented modern artists and the works available on WikiArt during the early Renaissance, and found that it produces periodizations which align with conventional art historical narratives, while also highlighting the relative uncertainty of period boundaries. This probabilistic approach is a less reductive way to arrive at artistic periods based on data without resorting to essentializing narratives of progress.

However, this method does not provide a path towards an objective or unbiased digital art history. No collection of images provides an objective or unbiased view of the past: myriad intentional and unintentional curatorial decisions made over time have influenced data availability. To some extent, any analysis of WikiArt a self-fulfilling prophecy because the collection of digitized works has been curated over time according to notions of art historical significance, and the trace of those notions is visible in the collection. For a thorough statement of this argument, see Amanda Wasielewski's book, *Computational Formalism* [319]. Beyond these arguments, no feature representation applies equally well to all images. Even the raw pixels of the photograph itself are an incomplete and ahistorical view of some works of art, as they only capture a view from one angle in a limited light spectrum, which (as we discussed in Chapter 2, has been engineered according to principles of colorimetry based on the perception of 20th century scientists [14]).

Instead, invoking Haraway [138], we believe this method provides a path towards a more explicitly situated form of digital art history. Given our positions as twenty-first century scholars, with uncertain prior knowledge about the past and incomplete, non-representative data, the best we can do is explicate our perspective. Taking a Bayesian approach directly supports that goal: priors allow us to state our subjective beliefs about the past, quantify their strength, then update those beliefs based on what data is available, without assuming the completeness of such data. A more historiographical approach based on formal Bayesian epistemology, such as that of Stephen Hartmann [143], would be an interesting direction for future work.

There are numerous opportunities to extend and expand this model based on other kinds of knowledge external to the images. For example, we can estimate how over- or under-sampled each artist and geographic area is over time and weight our examples based on those characteristics. We can also treat specific measures of subjective qualities like colorfulness as distributions, capturing the inherent subjective uncertainty around those qualities. We can also extend the model to avoid assuming a unimodal constant model with equal variance for each period, either by accounting for covariance between features or using linear or mixture models.

Discussion of Part I

Over the past three chapters, we have examined three cultural analytics studies which involve aesthetic phenomenon problems. Chapter 3 involves the visual similarity between webpages, Chapter 4 involves human preferences for color schemes and Chapter 5 involves the notion of an artistic period. All three of these qualities are subjective, and the latter two are the topic of debate around the hue-invariance hypothesis in color theory [161, 230, 250] and division of art history into stylistic periods [184, 122, 272].

In each of these chapters, we have taken different approaches to reconciling computational solutions with the fundamental qualitative nature of the underlying problems. In Chapter 3, we integrate ethnographic interviews into our data analysis and triangulate our findings. In Chapter 4, we use real art and design images as a proxy for popularity in context. In Chapter 5, we take a Bayesian approach, encode our qualitative knowledge into priors and predict distributions over answers.

However, in each of these studies, we neglected to explicitly evaluate our quantitative measures, as is typical in computer vision. We do not evaluate these measures because such evaluation is difficult, for a variety of reasons. In the following part, we will turn to a specific aesthetic phenomenon problem, image aesthetic quality assessment, and explore those reasons, and the broader issues surrounding evaluation, in detail.

Part II: Aesthetic Quality Assessment

Chapter 6

Aesthetic Quality Assessment and the Aesthetic Gap

6.1 Introduction

In this chapter, we introduce the test problem which we will discuss in the next three chapters: aesthetic quality assessment, or the task of determining whether a digital image is high or low aesthetic quality. Aesthetic quality assessment has a variety of useful applications — for example in computational design and automatic photo editing and curation. Our interest in this task, however, is more as a test problem, a prototypical aesthetic phenomenon problem. In this chapter, we introduce aesthetic quality assessment and trace its development.

Image aesthetic quality assessment (IAQA) seeks to apply machine learning to measure the aesthetic quality of images, usually by classifying them as “high” or “low” quality, based on the opinions of human raters, originally collected in 2006 from a photography challenge website [75, 185, 241], and more recently from a crowd worker platform [188, 265]. While aesthetics may seem solidly beyond the range of computation, there are a number of reasons computer vision researchers would like to have an image aesthetic quality measure, both for direct application in automatic photo curation and editing [205], as well as indirect use for the evaluation of image generative models [324] and image processing algorithms like computational bokeh effects [145].

From a computer vision perspective, this problem is interesting specifically because of how different it is from typical image classification tasks. Usually, computer vision seeks features and models which ignore the style of an image and only reason about its representational content. But in IAQA, we actually care more about the style. For example, in Chapter 8 we use several models which rely on the image Laplacian, the sum of the second partial derivatives of the image at each point, which detects areas of rapid change. While that property makes the Laplacian a good candidate feature for edge detection, it is noisy because it is too sensitive to sudden shifts in image intensity. In practice, the Laplacian is used for edge detection after applying a Gaussian blur filter, to remove false positives due to noise. But in IAQA, we are actually more interested in measuring the existing blur, so we work with the raw Laplacian. Similarly, Wang et al. explore which data augmentation strategies are "aesthetics-preserving," as typical image transformations like random cropping or color jittering might change the aesthetic quality of the image [318]. These are examples of the way that IAQA questions the boundary between signal and noise in computer vision, which leads towards different modeling decisions.

One concept which arises from this literature is the notion of the *aesthetic gap*. Roughly analogous to the semantic gap in information retrieval, which separates the low-level features of images like pixels and lines from the high-level features humans observe in images like objects and symbols [141], Datta et al. [76] define the aesthetic gap as separating "the information that one can extract from low-level visual data" and "the interpretation of emotions that the visual data may arouse in a particular user." This concept is particularly interesting because it places a boundary between problems which have solutions contained within the image itself, and those which require the involvement of human users, which resembles our concept of aesthetic phenomenon problems; the difference between the dimensions of a Rothko painting versus its feeling of scale, to reuse the example from Chapter 1. In this chapter, however, we turn the concept of the aesthetic gap back on IAQA and investigate the research question:

RQ4 — Does recent progress in aesthetic quality assessment actually constitute a cross over Datta's aesthetic gap?

To explore this topic, we present a historical narrative tracing the development of computer vision methods to measure aesthetic quality.

A preliminary version of this chapter was published at ICCC 2021 (P2)

6.2 Quantifying Aesthetics Before Computing

Taste varies from person to person, across time and place and is highly subject to influence, even in a laboratory setting [37]. Despite these challenges, aesthetics is one of the oldest topics of study in psychology, dating back to the 19th century work of the experimental psychologist Gustav Fechner. Fechner showed 347 subjects a series of rectangles and ellipses and asked them to choose the most appealing, and the rectangle with proportions drawn from the golden ratio was chosen the most frequently [128].

Fechner's work on aesthetics has been criticized by later psychologists and philosophers. For example, the 20th century Gestalt psychologist Rudolf Arnheim identifies a connection between Fechner's interest in measuring perception of beauty with his larger spiritual, cosmological and philosophical beliefs, and argues that Fechner's view of beauty as something which can be distilled down to one variable makes his findings related to art scientifically unreliable. “Just as Fechner's study does not tell us why people prefer the ratio of the golden section to others, so most of the innumerable preference studies carried out since his time tell us deplorably little about what people see when they look at an aesthetic object, what they mean by saying that they like or dislike it, and why they prefer the objects they prefer” [17].

Inquiry specifically into aesthetic measures, like the ones put forward by contemporary computer vision researchers, starts with the work of the 20th century American mathematician George Birkhoff. Birkhoff's 1933 book *Aesthetic Measure* puts forward a theory of aesthetic experience which divides it into three phases: first we recognize the complexity of a work, next we feel the sense that it is valuable, then finally we recognize the underlying order to which it adheres. Birkhoff claims these three properties: order (O), complexity (C) and value (M), can be related via an equation:

$$M = \frac{O}{C}$$

Where C is a measure of complexity, how difficult the work is to describe, and O is a measure of order, the degree to which the complexity is organized. For example, in the case of polygons, Birkhoff defines,

$$M = \frac{V + E + R + HV - F}{C}$$

Where V, E and R are binary features indicating vertical symmetry, equilibrium (visual balance) and rotational symmetry, respectively, HV is a feature with value 2, 1 or 0 based on whether the lines of the polygon lie on a simple network of parallel or perpendicular lines, F is the number of unsatisfactory visual characteristics, out of a list of 7, and C is the number of distinct lines containing at least one side of the polygon.

A surprising number of Birkhoff's shapes are cultural and political symbols, including the Christian and Celtic crosses, Star of David, triskelion and swastika. While these symbols are simple shapes and may only appear as coincidence, recent study of Birkhoff's antisemitism and involvement with Nazi Germany might give us reason to reconsider why Birkhoff, whose prior work was mostly in dynamical systems, would want to develop a quantitative system of objective beauty [242]. Regardless of his motivations, Birkhoff's concepts of order and complexity showcase a mathematician's view of beauty, where the most beautiful things are large, with many symmetries, but emerge out of elegant descriptions. Similar concepts of beauty, which exist in tension between simplicity and epistemic satisfaction, have been found in social studies of mathematicians [158].

Birkhoff's approach, like Fechner's, has been extremely influential, inspiring a century of computational approaches to aesthetics (e.g. Moon and Spencer's model of color harmony [236]), but it is poorly regarded by many philosophers. For example, Susanne Langer claims that the easily described nature of musical harmony has led to a great deal of hope that other aspects of art might be quantified and understood mathematically as well. However, “there is no use discussing the sheer nonsense or the academic oddities to which this hope has given rise, such as...the serious and elaborate effort of G.D. Birkhoff to compute the exact degree of beauty in any art work (plastic, poetic and musical) by taking the ‘aesthetic measure’ of its components and integrating these to obtain a quantitative value judgment” [199]. Langer goes on to argue that while musical sound is easy to describe, such description does not access the artistic qualities of music like motion, which exist in virtual space and time rather than in the physical sound.

Langer’s criticism of Birkhoff invokes a similar criterion to Datta et al.: the difference between the explicitly measurable qualities of an object and the virtual and experiential qualities which inform its aesthetics are quite similar to the idea of a semantic or aesthetic gap. While rather simplistic mathematical models like those of Birkhoff likely lack the capacity to model something comparable to a human’s aesthetic response, it is unclear whether more sophisticated computer vision models learned from data share that limitation.

6.3 Early Machine Learning Approaches

Contemporary study of aesthetics in computer vision begins with the simultaneous work of Datta et al. and Ke et al. in 2006. Despite both working at the same time, and in the same US state (Pennsylvania), these two groups of authors arrived the problem area from different conceptual directions and take different approaches within the context of image classification.

Datta et al. are determined to automatically learn from data which factors influence aesthetic value. They claim that, "in spite of the ambiguous definition of aesthetics...there exist certain visual properties which make photographs, *in general* more aesthetically beautiful" [75]. Their concept of aesthetic value originates from their data: over 3,000 images collected from the website `photo.net`, which allows users to upload their photos, and allows other users to rate them on "aesthetics" and "originality."[1] They cite two other sources on their understanding of aesthetics: the Oxford Advanced Learner's Dictionary and a book, Rudolf Arnheim's 1965 *Art and Visual Perception: A Psychology of the Creative Eye* [16]. Aesthetic quality assessment is framed in terms of image classification: they train decision trees and support vector machines to classify images into high and low aesthetics categories based on a variety of features extracted from images (e.g. measures of colorfulness, the photographic rule-of-thirds, image dimensions).

The decision to cite Arnheim pulls this approach towards psychological aesthetics, a field which exists in dialogue with both the work of earlier psychologists like Fechner, as well as the history of aesthetic philosophy. In a later survey paper [171], the same authors cement that link. They discuss the approaches of analytic philosophers like Nelson Goodman and Richard Wollheim, as well as recent work in neuroaesthetics by Semir Zeki, who claims that aesthetic experience can be identified and explained by activity in specific brain regions.

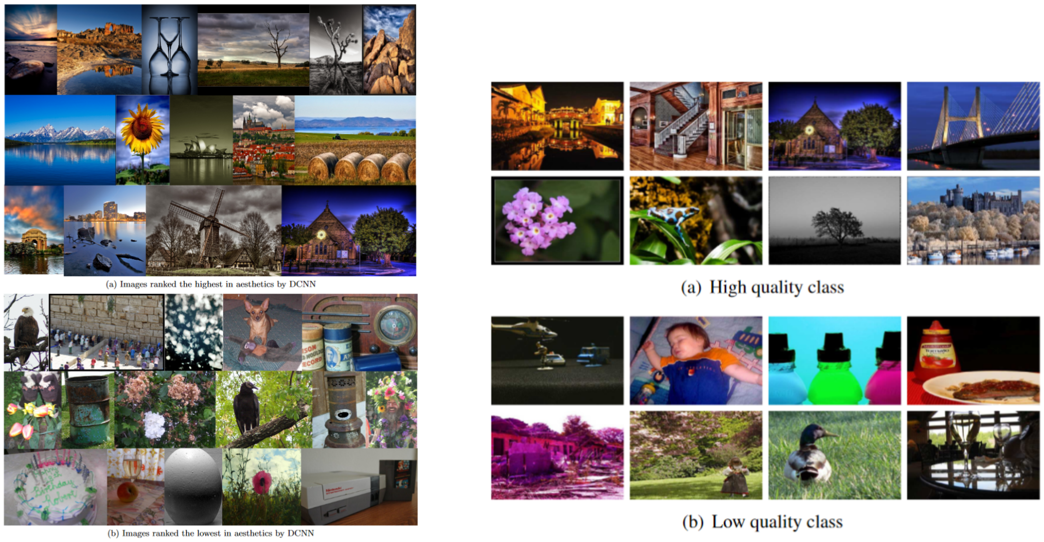
To contrast, Ke et al. [185] approach IAQA from the perspective of photo curation.

[1] `photo.net`, surprisingly, was not created by professional photographers, but by Philip Greenspun, a computer scientist at MIT interested in online communities.

Rather than psychological aesthetics, they ground their work in image quality assessment, an area of computer vision research concerned with measuring image noise and degradation [174]. Rather than making claims about philosophy, Ke et al. argue that a well-designed set of features may be used to reason about the subjective aspects of image quality, like the difference between professional and amateur photos. Their method makes use of images and ratings from the photo challenge website `DPChallenge.com`, which they divide into "professional" and "amateur" categories based on ratings. They cite two popular photography books to justify their choices of features, which include edge and color histograms, as well as Fourier transform-based blur metrics, which they use to train a Naive Bayes classifier.

Over the next six years, a variety of other publications emerged proposing different combinations of image features for solving the aesthetic quality assessment problem. While other scholars used similar approaches at first [76, 167], later authors shifted towards low-level features like GIST or SIFT descriptors due to an influential paper by Marchesotti et al. which made the case that hand-crafted features are ineffective because they are non-exhaustive, computationally expensive and rely on heuristic assumptions which may not generalize well [226].

The relationship between Datta, Ke and both earlier and later aesthetic thought is at the heart of our claims about the aesthetic gap. The work of Datta et al. is framed as an approach to computational aesthetics, but like Ke et al., they only measure how consistent a photograph is with common photography rules of thumb. Later work further conflates these two concepts of "aesthetic quality" by shifting to lower-level image features to better fit the



(a) Images ranked the highest in aesthetics by DCNN

(b) Images ranked the lowest in aesthetics by DCNN

(a) High quality class

(b) Low quality class

Lu et al. [214] Lee et al. [202]

Figure 6.1: Two figures from IAQA papers comparing high and low aesthetic quality images in their dataset.

dataset labels. However, inspection of the "high quality" and "low quality" images in these datasets makes it clear that the distinction between them is more of a stylistic difference than anything else. Figure 6.1 shows comparisons between high and low quality photos from two IAQA papers. The qualities shared by all of the photos labeled as "high quality" is evident: these are overwhelmingly photos of landscapes and flowers which prioritize color and explicit emotionality, the style that wins contests on `DPChallenge.com`. But adhering to this style is not synonymous with having a high quality photo. Photography can be aesthetically pleasing in as many ways as other art forms, and many genres of art photography like candid photography or photojournalism do not prioritize the use of such dramatic visual effects. In other words, these papers and datasets seem to conflate explicit emotionality with the potential to arouse emotion.

6.4 The AVA Dataset and Deep Learning

In 2012, two major events shifted the conversation around IAQA. First, in June, Murray et al. [241] released the Analysis of Visual Aesthetics (AVA) dataset, containing over 250,000 photos from `DPChallenge.com`, an order of magnitude larger than any existing dataset. They also released metadata, including rating distributions and category labels, where possible. Second, in October, Krizhevsky et al. [194] dramatically beat the benchmark on the ImageNet visual recognition challenge using a deep convolutional neural network (CNN). While deep learning had profound effects on computer vision as a whole, these two contemporaneous changes produced a paradigm shift in the study of IAQA.

Lu et al. [214] published the first paper applying deep learning to aesthetic image classification in 2014. They reiterate the argument from Marchesotti in favor of generic image features, and claim that deep features are even more generic, since they work with pixels directly. Lu et al. identify that the fixed input size of AlexNet makes it difficult to apply to images of many different dimensions in AVA, since cropping or warping might disrupt aesthetic quality, so they use two-column models. These models contain two "columns," each following the same network architecture, to learn from warped and cropped versions of the image simultaneously, then integrate the learned representations before the final fully-connected classifier layer. Neither this work, nor the generation of papers which followed their lead in applying CNNs to the AVA dataset [179, 335, 218], make much reference to the problem statement and its context at all, aside from acknowledging its highly subjective nature.

While CNNs do not carry all of the assumptions of things like measures of colorfulness or edge histograms, they are not blank slates either. The connectivity structure of convolutional and max-pooling layers within these networks encode several assumptions. For example, the assumption the salient features of an image occur in relatively small patches within the image, or that the presence of an activation is more significant than the absence. These are good assumptions for classifying between different types of objects or handwritten digits [194], but are not necessarily good for aesthetic judgment, which at least in the eyes of psychologists like Arnheim [16], relies on holistic, Gestalt phenomena.

In the past five years, several trends have emerged in IAQA. First, Kong et al. [188] suggest including user data to personalize image assessments [202, 71, 336, 172, 173], which Ren et al. [265] formalize into an active learning task. Second, different objectives beyond classification have emerged, including pairwise comparison [218, 202] and distribution learning [70, 104]. Finally, the binary classification accuracy benchmark on the AVA dataset has steadily increased, reaching over 91% (see Table 6.1).

Additionally, a new argument for this research area, related to curation and editing of photographs for social media, has emerged. Several recent authors make reference to the widespread popularity of social networking services [316], the exponential growth of online visual data [284, 202] and the growing need for automatic photo editing tools [316]. This claim for significance brings IAQA into the realm of AI-based creativity support tools, further increasing its relevance to the computational creativity community.

Our narrative in this section emphasizes the continuity between the current state-of-the-art in IAQA and the long history of aesthetics in other disciplines. There is a direct

Paper	Year	Metrics	Score
Murray et al. [241]	2012	Accuracy	67%
Tang et al. [300]	2013	Accuracy	92%
Lu et al. [214]	2014	Accuracy	71%
Lu et al. [215]	2015	Accuracy	75.4%
Kao et al. [179]	2015	MSE	0.45
Mai et al. [222]	2016	Accuracy	77.1%
Kong et al. [188]	2016	ρ, Accuracy	0.56, 77.3%
Zhou et al. [335]	2016	Accuracy	78.1%
Lv et al. [218]	2016	Mean AP	0.611
Wang et al. [318]	2016	Accuracy	76%
Kao et al. [178]	2017	Accuracy	78%
Ma et al. [219]	2017	Accuracy, F1	82.5%, 0.92
Ko et al. [187]	2018	ρ, accuracy	0.87, 82.2%
Fang et al. [104]	2018	Distribution metrics	0.12 (KL)
Jin et al. [168]	2018	Distribution metrics	0.381 (KL)
Sheng et al. [284]	2018	Accuracy	83.3%
Lee et al. [202]	2019	ρ, MASD, accuracy	0.92, 0.02, 91.5%
Wang et al. [316]	2019	ρ, MSE, accuracy	0.28, 0.69, 81.5%

Table 6.1: Benchmark score results on the AVA dataset.

continuity from classical to deep methods: Marchesotti et al. made their argument in favor of low-level image features before the advent of deep learning, and the first deep learning-based method of Lu et al. is framed as the natural extension of that argument. Even highly technical recent papers, which are quite distant from the philosophical motivations of authors like Datta et al., are implicitly weighing into a long conversation on the nature of art and beauty, which may have wide reaching implications. But do any of them really cross the aesthetic gap and reason about "the interpretation of emotions that the visual data may arouse in a particular user?"

6.5 Discussion

So far, we have traced the evolution of IAQA in computer vision from its prehistory in the work of Fechner and Birkhoff, its origins in psychological aesthetics and photographic rules of thumb and its shift from hand-engineered features to deep learning. We saw how its two goals rooted in computational aesthetics and image quality assessment merged over time, and how performance on the AVA dataset, which arguably only captures a specific, popular photographic style, has been treated as a stand-in for an algorithm's ability to measure aesthetic quality more generally. With that continuity in mind, we find it difficult to point to a specific paper or accuracy level where these approaches cross the aesthetic gap introduced by Datta et al. However, such a claim raises other questions about the nature of this gap.

For example, it is possible that the success of recent deep learning models on the AVA dataset demonstrates that there is no such gap: the neuroscientific arguments indicate that our aesthetic responses exist in a lower level of the visual system than we might believe [59], and it is possible that we actually make judgments based on simple visual statistics and only use higher cognitive processes to explain those judgments, much like de Piles' ranking of the artists discussed in Chapter 2. Such a finding would vindicate scholars like Birkhoff, who believed that a measure of aesthetics could be computed from measures of order and complexity, without regard for the emotions of the observer [39]. On the other hand, if we assume that an aesthetic gap does exist, and making aesthetic judgments requires algorithms which understand meaning and emotional attachment, that would cast further doubt on whether IAQA models are actually measuring aesthetics and whether accuracy on the AVA is a suitable measure of performance.

If deep learning models cannot overcome the aesthetic gap, how should we, as artificial intelligence researchers, proceed? It's not unreasonable to imagine a computationally creative agent which both interprets symbols and models emotional attachments enough to have something resembling an understanding of taste. But since taste is subjective, it is still unclear how to measure performance. Can a model have its own preferences, or should it merely predict the preferences of humans?

This last point reaches towards an important question regarding artificial intelligence and subjectivity. When the authors in this space describe IAQA as subjective, who do they imagine to be the subject? IAQA chooses to derive ground-truth labels from an average of many humans' aesthetic quality ratings, but such data risks conflating aesthetic quality with photo contest popularity. But, as we explore in the next chapter, attempting to predict the aesthetic preferences of each individual user comes with its own host of caveats and difficulties.

Chapter 7

Correct for Whom? Problems with Personalization

Over the past fifteen years, computer vision researchers have investigated techniques for image aesthetic quality assessment (IAQA). As discussed in the previous chapter, this research area emerged from image quality assessment [185] and computational aesthetics [75, 76]. 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 [241].

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 [70, 104] or few-shot personalization [265, 202, 71, 336, 172, 173]. To account for variation between users, some methods use auxiliary information from social media data [70], demographics [172, 173], or psychometrics [336] to better capture a given user's perspective. In parallel, other researchers [218, 202] 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 [258]. 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 [176]. 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. [75] 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 [171]. Similarly, when proposing their well-known AVA dataset for IAQA, Murray et al. [241] 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 [177], and Edmund Burke argued that light skin was naturally aligned with the beautiful while dark skin was closer to the sublime [15]. 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 [190], which has contributed to various forms of discrimination, including gender and racial bias in the artistic canon [26, 82]. As elaborated by Carolyn Korsmeyer [190],

> 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. Though Kant is not developing machine learning model architectures, this argument echoes both Haraway’s critique of god-tricks in science, as well as our concept of “subjectivity in the model” from Chapter 2.

Returning to machine learning, we can take inspiration from this philosophical debate and generate empirical research questions about personalized IAQA:

RQ5 — How well do the average aesthetic scores from an existing dataset actually predict new individuals’ judgments?

RQ6 — 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.

These issues are important to consider 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. [85] 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 Luc Ferry [105] 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. [188]. 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 intersubject 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

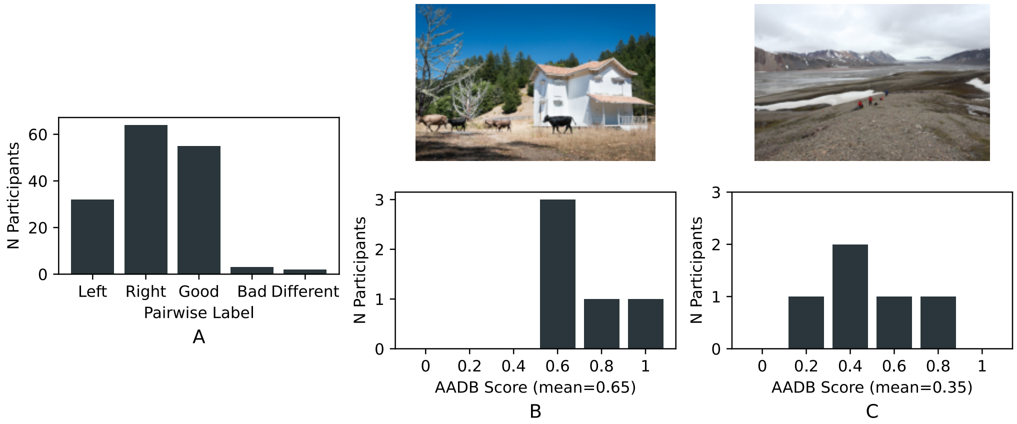


Figure 7.1: **Sample ratings from our dataset.** A pair of images from the AADB dataset, one of 20 "common" pairs shown to all participants in our study. *A:* the ratings from our participants using a labeling scheme with options for "both images are good," "both images are bad" and "these images are 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.

not simply a matter of gender or education level, for example.

A preliminary version of this chapter was published at AAAI 2023 (P4)

7.1 Related Work

As discussed in the previous chapter, many authors have approached IAQA in an objective, "average" framing over the past two decades. Recently, others have framed IAQA as a more subjective problem. Ren et al. [265] 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 [241] for pretraining, Lee and Kim [202] 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 [123].

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 [252, 216] for discussion. Böckenholt [42] 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.

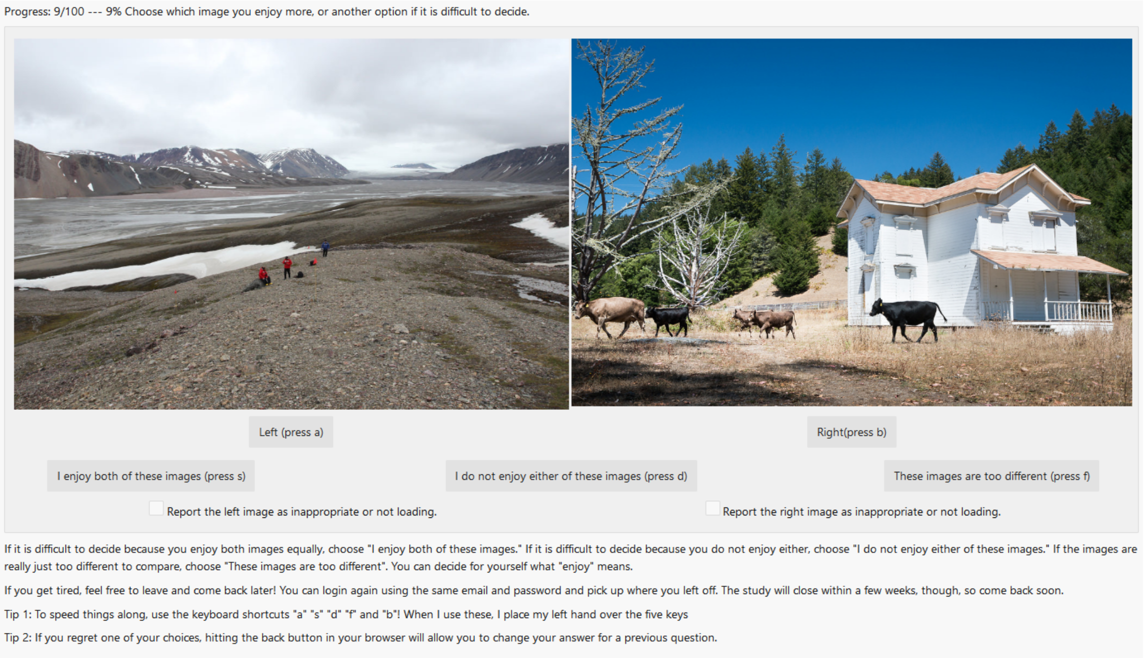
While research through data relabeling is a relatively unusual approach, it has begun to gain traction in machine learning. Beyer et al. [34] conduct a relabeling of the ImageNet validation set [84] to assess whether improved accuracy on ImageNet actually reflects progress on image classification. Kong et al. [189] develop a more general framework for studying relabeling and its effects on model performance.

7.2 Methods

7.2.1 Study and Collection Interface Design

To permit comparison with existing work, we collected new aesthetic labels for the existing AADB images [188] (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 [188]) use Amazon Mechanical



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 7.2: Screenshot of our labeling interface.

Turk (AMT), and collect aesthetic labels for individual images on a two- [300], five- [265], or ten-point [188] 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) [225]; 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 [72], psychology of art [234], and aesthetics of the everyday [33], and evokes the language of philosopher John Dewey's concept of the aesthetic: "experience as appreciative, perceiving and enjoying" [86]. Others have used prompts such as interestingness [134, 115], pleasing, harmonious [117] 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.

7.2.2 Recruitment and Data Collection

After receiving approval from our university's IRB, 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.[1] 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 partici-

[1] Frequency values for demographic characteristics can be found in our supplementary materials, available at `http://vision.soic.indiana.edu/wp/wp-content/uploads/suppelementary_materials_for_IAQA_and_feminist_aesthetics.pdf`.

pants' 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[2], 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.

7.2.3 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 [202], 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 [312] 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 a n-dimensional vector of real-valued scores $S = S_1, ..., S_n$ and a matrix of pairwise labels, $L = L_{i,j} \in \{-1, 0, 1\}$ where L is only defined

[2] These responses were excluded because they contained lengthy, nonsensical free text responses and chose each of the five options equally often; human participants typically chose "left" or "right" significantly more often than "these options are too different to compare." Excluded participants were still compensated.

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,

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

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.

7.3 Results

7.3.1 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

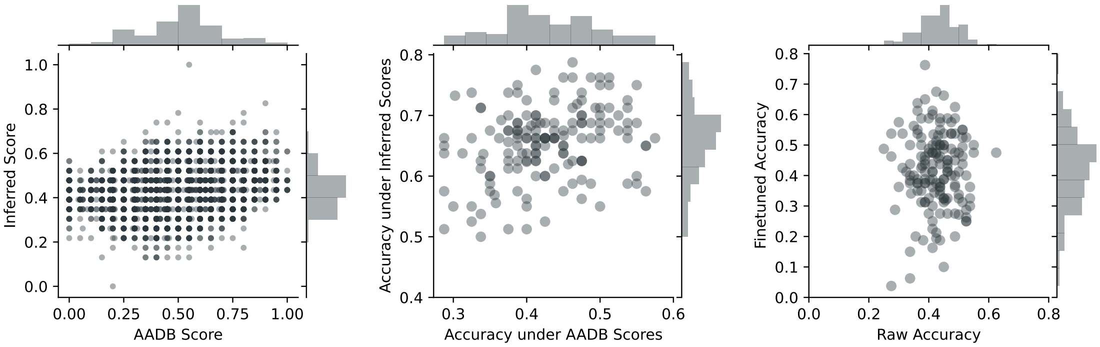


Figure 7.3: **Measuring agreement and variance in aesthetics scores.** *Left:* After converting our pairwise labels to inferred image scores, we plot them and measure their correlation ($r = 0.27$) vs. the original AADB labels. Data points indicate images. *Center:* Next, we convert both the AADB scores 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 [265], 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.

also test the performance of a personalized model inspired by that of Ren et al. [265] on our labels.

Comparing Single-Image Scores

Figure 7.3 (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 [202] report $\rho = 0.879$).

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 7.3 (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 by Ren et al. [265], 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 7.3 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.

7.3.2 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. [265], 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 [304] with regularization tradeoff parameter $\alpha = 0.0005$ to select relevant content variables [107]. 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.

The estimated coefficients are shown in Figure 7.4. 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.

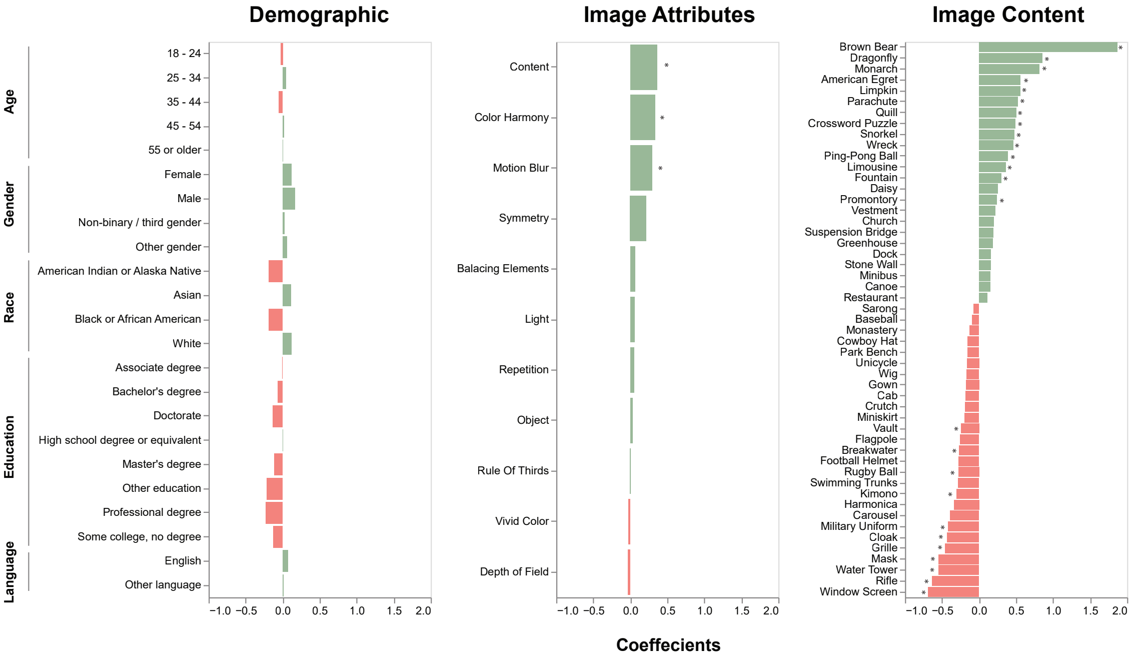


Figure 7.4: **Which properties of participants/images predict (dis)agreement between the AADB ground truth and our participants' ratings?** Regression coefficients for (left) demographic, (center) aesthetic, and (right) content indicate how much of the variance in consistency is explained by each attribute 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.

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. [75], 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.

7.3.3 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)

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 7.1, 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.

7.4 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. [188] and our relabeling. We find that this critique largely holds for IAQA data: while the original labels are usually better than random guessing, their predictive quality varies greatly from person to person, and modeling, especially with 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 [188] or the personalization model we used [265]. 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 [241], 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 [188] or FlickrAES [265]. 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.

Chapter 8

Design of an Interface for Participant Evaluation of Aesthetic Quality Assessment Models

8.1 Introduction

Over the previous two chapters, we have assembled an argument that computer vision evaluation methods are insufficient for handling the subjectivity inherent in aesthetic quality assessment. In Chapter 6, we raised the issue of choosing the subject whose taste we are emulating in aesthetic quality assessment. In Chapter 7, we explored the idea that the user could be the subject. However, attempts to model the differences in human preferences are very difficult to evaluate as our results start to depend on the sample of individual labelers and their respective concepts of aesthetics. A model which scores better for one group of people may not score better for a different group of people. We can avoid this problem by specifying. Instead of trying to model aesthetic preference in general, we model a specific person or group's taste. However, collecting large quantities of evaluation data for a specific user population is undesirable for many computer vision applications, as data collection is expensive and techniques are not often developed with specific users in mind.

In this chapter, we explore an alternative evaluation paradigm: treating IAQA models in the non-personalized formulation as a tool, and approaching its evaluation through a more contextualized and situated approach. Following Galanter [113], each aesthetic quality assessment model is an image measure, one element of the space of all functions which map images or collections of images to scalars. While some elements of this space are measures which arguably correspond to specific visual qualities (like Hasler-Suesstrunk colorfulness from Chapter 1, Birkhoff's aesthetic measure or the layout similarity metric based on tree edit distance from Chapter 3), most of them are unrecognizable to humans as measures of taste. From this perspective, determining whether any of these functions actually measure aesthetic quality is entirely external to the mathematics; a matter of interpretation and judgment which must be carried out by a human in context. So instead of measuring whether these functions correctly assign aesthetic quality scores to images, we determine whether they measure a quality which human subjects can recognize and interpret as aesthetic quality. We investigate the research question:

RQ7 — How do computing graduate students evaluate aesthetic quality assessment models when seen through a smartphone camera interface?

This question and line of thinking are directly rooted in Haraway's concept of situated knowledge. As discussed in Chapter 2, Haraway argues that when science separates a "view" of the world from the way that it was captured, it performs a "god trick," pretending to see everything from nowhere. Computer vision as a discipline often performs two such god tricks. First, literally, it treats sets of images as objective recordings of reality, detached from the cameras and photographers who take them. Second, more metaphorically, it treats its knowledge about the performance of models and algorithms as objective truth, separate from the data and methods which allow us to evaluate them. Many computer vision systems have substantial limitations: they are only ever approximately correct, only have limited knowledge of the world and faithfully reproduce the biases, both good and bad, of their training data [54]. The problem is not the existence of these limitations, which researchers often acknowledge, but the way the research system performs a god trick and transforms algorithmic tools which provide situated, uncertain knowledge about the world into arbiters of objective truth, justified by quantitative evaluations.

Haraway argues that knowledge being situated does not make truth relative. Instead, it requires a shift to feminist objectivity where real, objective knowledge about the world comes from specific physical, social and conceptual perspectives. We seek to carry out IAQA model evaluations from specific perspectives. To envision what this approach would look like, we turn to secondary literature which operationalizes Haraway's theory. Bhavnani, writing in the context of social science research [35], proposes three criteria for feminist objectivity:

1. Reinscription: Does the research method portray the participants as passive and powerless, or does it recast them as active agents?

2. Micropolitics: Does the research engage with the political relationships between researcher and participant?

3. Difference: Does the research engage with differences in perspective between participants?

Towards these goals, we introduce a qualitative method for participant evaluation of image measures and describe a pilot study exploring how participants interpret and evaluate four IAQA models. Our method relies on making the rather abstract aesthetic quality assessment models embodied and tangible via a smartphone camera app interface. Unlike ordinary camera apps, our interface lacks a shutter button and instead takes photos when the value of an aesthetic quality assessment algorithm exceeds a threshold. By placing the point of interaction within the familiar context of a camera interface, we are able to significantly broaden the kinds of perspectives which can be collected regarding model performance. Unlike quantitative evaluation methods which are typically carried out by computer scientists and machine learning engineers based on data from anonymous human participants, any sighted person with smartphone literacy skills can evaluate IAQA algorithms themselves using our interface. As a result, our work is also a contribution towards conversations around explainable AI.

Our approach to evaluation takes place in the world at the time the photograph is taken. There is no hidden photographer responsible for the images. Evaluating at the time of photography avoids the first god trick because we do not disconnect the very literal view of the world from its source. On the more metaphorical level, we also meet Bhavnani's criteria:

1. We recast the human subjects, who in other IAQA research are anonymous crowd workers, into active participants in the research process who are given space to express their nuanced views about aesthetics and cameras.

2. By giving participants evaluative agency, we reverse the typical power dynamic in machine learning where photographers and labelers are disconnected from the models derived from their data.

3. Qualitative analysis gives us the flexibility to handle subjective difference with nuance, rather than reducing it to a measure of dispersion or personalization technique.

Regardless of whether one accepts this feminist framing, we find that these evaluations form a helpful complement to benchmark-based evaluation particularly because individuals' judgments about the various IAQA algorithms are contextualized within their existing relationships to images, cameras and aesthetics. Engaging with participants in a qualitative, open-ended setting allows them to offer feedback on the problem formulation and potential applications, elements which are typically only ever evaluated by authors and conference reviewers, or users of applied products based on computer vision methods. This approach effectively allows us to evaluate subjective elements at both the level of the data as well as the level of the model (as discussed in Chapter 2), allowing researchers to reflect on the assumptions and potential biases present in these models.

Placing these algorithms in the context of a camera is particularly pertinent today, as the boundary between human and computational elements in smartphone photography becomes increasingly blurred. Both iPhones and many models of Android now support "live photo" or "best shot" modes where cameras actually take a short video when the shutter button is pressed, then automatically choose the "best" frame according to an aesthetic measure. The specific aesthetic measure used by Google Pixel phones (as of 2018) is a linear model based on features related to face quality, object optical flow, global motion blur and camera status, relying on a MobileNet-based object detector [333]. Despite the ubiquity of this feature, most smartphone users are not aware that their photos are produced somewhat collaboratively with a machine learning model. By using an application without a shutter button, we hope to make this functionality more conspicuous.

8.2 Related Work

In addition to the related work discussed in previous chapters, our work exists in dialogue with work in model interpretation and visualization, particularly user mental models of AI systems. Our approach is additionally informed by the study of egocentric computer vision and critical design approaches from HCI.

Approaches to visualization for computer vision models typically focus on offline approaches. These include visualizations of feature maps for classical methods [315], class activation maps derived from deep neural networks [281], nonlinear dimensionality reduction techniques like T-SNE [309] and UMAP [231] for visualizing feature spaces and more sophisticated agent-interaction explanations like the visual explanations of Hendricks et al. [149]. Outside of computer vision, there is a robust literature in visual analytics for interactive visualization for machine learning model development; see [268] for a literature review. The goals of visualization depend on user psychology; model explanations should cohere well with how users already mentally model the problem under study [206], and there is often a tradeoff between soundness and completeness in explanations which shapes user understanding [196]. We approach this problem very differently: instead of producing offline visual explanations, we allow the user to build their own explanations based on free interaction with the model.

Our usage of a handheld camera as a site for exploration of computer vision is inspired by egocentric (first person) computer vision, the study of imagery taken from a camera worn on a human [74], robot [229], animal [198] or car [328]. This research area situates the camera within the scene it depicts, which is atypical for computer vision and creates a variety of technical challenges regarding camera shakes and occlusion from hands [19]. This research area echoes earlier ideas from authors in robotics like Rodney Brooks regarding situated and embodied intelligence [50].

Finally, we build on the concepts of research through design [108], speculative design [93] and critical design [23, 20]. These closely-related methods use design practice as a form of research, exploring concepts from design theory, speculating about alternative futures and explicating the implicit assumptions behind technology. For example, Odom et al. apply these methods in their design of Photobox, a speculative slow technology which occasionally prints photos from a Flickr library, slowly amassing a collection over years, which challenges the notion that technology should be fast, easy and disposable [249]. Similarly, Pierce and Paulos's Inaccessible digital camera, a camera made of concrete which must be destroyed to extract the digital storage medium within, similarly questioning notions of functionality and disposability [257]. Recently, Karmann's Paragraphica continues in the legacy of speculative camera designs. This camera-like device lacks a lens or photo sensor, instead it generates photographs using an image generation model conditioned on location, time, temperature and maps data, questioning the mapping between photos and reality [182]. The important characteristic is that these designs are not potential future products. Instead, they are used to explore alternative design spaces and make abstract critiques of technology tangible, facilitating future designs in these spaces.

8.3 Interface Design and Development

Our design goal is to make aesthetic quality assessment algorithms tangible, so that participants can judge the algorithms' learned aesthetic preferences in a real-world setting. An initial sketch for the design is shown in Figure 8.1 (a). While we initially discussed designs with micro-controllers and LED feedback, we decided on an Android smartphone-based design, since smartphones offer both high-quality cameras and hardware acceleration for on-device deep neural network inference at relatively low cost. A low-fidelity prototype, an image we showed on a smartphone to test the concept, is shown in Figure 8.1 (b).

The application was developed on top of the existing open source project Open Camera, licensed via the GNU General Public License [142]. Starting from an existing camera application was helpful for learning the Android API and gave the project a functional camera user interface. We added three settings to the application:

1. "Aesthetics Capture Mode" removes the shutter button and starts a background process that takes and evaluates a photo once per second. If the estimated IAQA score is greater than a threshold, the photo is saved and a visual feedback animation for photo capture plays.

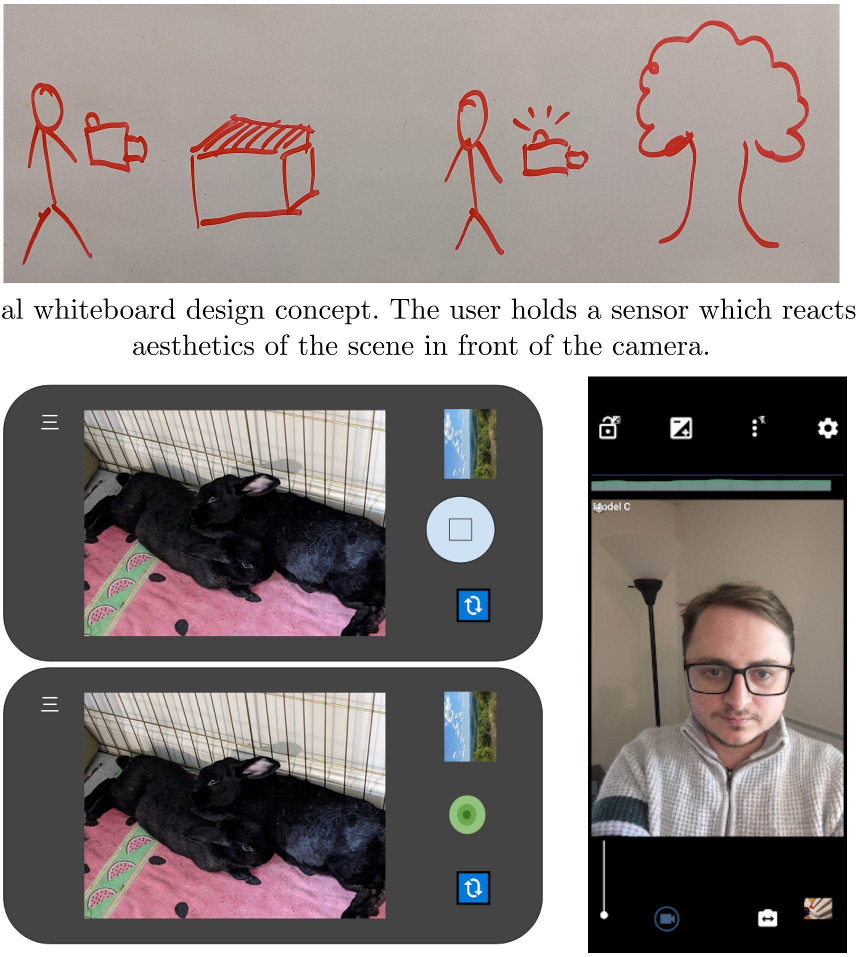
2. "Aesthetics Indicator Mode" starts the same background process, but adds a line plot to the top of the preview showing the IAQA model output. If aesthetics capture mode is also on, the graph shows a horizontal threshold line as well.

3. "Aesthetics AI Sensor" allows the user to choose one of four models, described below.

Removing the shutter button was a key decision rooted in critical design methods. Users assume they have full control of a camera, but many of the choices involved in photography are made automatically in cameras already [151], simulating different film ISO values [127] and creating computational Bokeh effects without adjusting focal length [145]. These auto settings are often adjusted based on computational image quality measures [145]. Users, however, expect to choose when photos are taken, and removing that aspect of user control could lead them to reconsider the way that computational measures of taste are already influencing photography, and to speculate about human-AI co-creativity in photography [77] and possible future AI art forms.

We experimented early in the development process with removing the camera preview and using different kinds of visual or haptic feedback, but found in an informal pre-pilot session with one user that taking photos without a preview was difficult. We also moved away from the visual indicator present in the early design (Figure 8.1 (b)), which looked too much like a shutter button, leading our pre-pilot user to tap it and expect to take a photo. Haptic feedback proved especially unusable because it was difficult to calibrate: either the phone would vibrate constantly, annoying the user and draining the battery, or vibrate too little to be helpful for interpreting the model output. As a result, we designed the minimal line plot shown in Figure 8.1 (c).

Our prototype has four model architectures which are characteristic examples of four different approaches to aesthetic quality assessment, inspired by different eras of research on this problem.



(a) Initial whiteboard design concept. The user holds a sensor which reacts to the aesthetics of the scene in front of the camera.

(b) Low fidelity prototype: a smartphone camera app without a shutter button

(c) Screenshot of the final prototype.

Figure 8.1: Design Iterations

A: A baseline model using the mean of the approximate image Laplacian as an aesthetic measure, i.e. for grayscale image $I[x, y]$ with width W and height H,

$$M_{\mathcal{A}} = \frac{1}{WH}\sum_{i=1}^{W}\sum_{j=1}^{H}(I * L)[i, j], L = \left[\begin{smallmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{smallmatrix}\right]$$

Where $*$ is the discrete two-dimensional convolution operator. The image Laplacian is highest in areas where there are sharp edges in pixel value, and goes down when an image is blurry, giving it a slight positive correlation with aesthetic quality.[1]

B: A linear model using hand-crafted features based on the early IAQA work of Ke et al. [185]. This model uses four sets of image transformations: the image Laplacian, the 4096-bin color histogram, the image Fourier transform and the lightness distribution. The first two transformations are further distilled by taking the mean feature map of the positive and negative classes and measuring the $\mathcal{L}_1$ distance from the test image to the mean for each class. In the case of the Fourier transform, we follow Ke et al. and measure the highest frequency bin with value greater than 5. For the lightness, we measure the width of the 98% mass distribution. While Ke et al. use a Naive Bayes classifier on these features, we use logistic regression for ease of deployment alongside the other three models.

C: A 2014-era deep neural network, based on the 8-layer AlexNet architecture [194], with a two-column approach similar to Lu et al. [214]. To avoid warping images to

[1] We also experimented with the variance of the Laplacian, but found that the mean led to a more interesting aesthetic measure, as it increased in the presence of visually interesting content, and decreased for blurry photos.

the 224 by 224 resolution required by the AlexNet architecture, this model has one AlexNet-style network for a center cropped local view of the full resolution image and another center cropped global view of the image downsized to 256 by 256. Following Lu et al., we concatenate the models' hidden representations before applying the final fully connected layer, allowing us to train the two columns jointly using stochastic gradient descent.

D: A more contemporary deep neural network approach, based on the 18-layer Resnet architecture [148] and trained using the Adam optimizer [186], without any other IAQA-specific modifications.

Both deep neural networks are randomly initialized using He initialization [147] and trained on the AVA dataset using a cross-entropy loss function and learning rate starting from 0.001 and slightly decaying multiplicatively each epoch by $1 - 10^{-7}$. We emphasize that our goal here is not to qualitatively test specific modeling decisions. Instead, we use these four models as examples which characterize four different approaches to the problem: a mathematically elegant metric, a highly engineered feature-based approach based on researcher intuition, a more data-driven model designed for the task, and a fully data-driven approach using a problem-agnostic model architecture.

Quantitative evaluation results for the models using the AVA dataset benchmark are shown in Table 8.2, and a visualization showing differences in model output distributions is shown in Figure 8.2. Following the IAQA literature, we report accuracy and ranking correlation (Spearman's ρ). Accuracy is somewhat misleading for the AVA dataset, which has almost a 70%–30% class imbalance towards positive examples, so we also include the AUC-ROC, a measure of the probability that a high quality photo will score higher than a low quality photo. All metrics are computed on the AVA test set. Despite the fact that the model architecture was not designed for the task, Model D performs the best across the board. While our first three approaches roughly mirror the reported accuracy in their respective papers, none of these models are as performant as more recent state-of-the-art methods: we encourage future work evaluating specific contemporary approaches.

We initially listed models using the last names of the first authors of the papers which proposed the architectures, but based on pre-pilot study feedback, we changed them to "Model A," "Model B," "Model C" and "Model D" to avoid giving the models human-like names. Additionally based on pre-pilot feedback, we made the threshold adaptive. Specifically, given a sequence of aesthetics ratings $y_1, ..., y_t$, an image is saved at time t if

$$y_t > 0.11 \sum_{i=1}^{10} y_{t-i}$$

In other words, a photo is taken every second, but it is only saved if the current photo is rated at least 10% higher than the average of the prior 10 photos.

8.4 User Study Methodology

After developing our camera interface, and receiving IRB approval, we conducted a pilot study of its capabilities. This study served two purposes: it allowed us to demonstrate that the core functionality of the interface worked and was usable by participants with

Participant	Age Range	Gender	Academic Background
P1	31-40	Woman	HCI, health informatics, computer science
P2	31-40	Man	security informatics, HCI
P3	31-40	Woman	HCI, data science
P4	21-30	Man	HCI, computer science
P5	31-40	Man	Electrical and computer engineering

Table 8.1: Study Participant Demographics. We specifically recruited graduate students in computing outside of computer vision, with a focus on students with HCI experience.

limited computer vision or machine learning experience. We conducted one 40-70 minute session with each of five participants between February and March 2023. Participants were recruited using paper fliers and university mailing lists, and paid US$15 for participation. All participants were graduate students in computing-related disciplines outside of computer vision, AI and machine learning; see Table 8.1 for details. We focused our recruitment on students with some graduate training in HCI, as designers and researchers outside of computer vision are the target user population for our application. We imagine a future use case where designers could try out different aesthetic measures for a potential software application using our interface as a visualization tool.

Sessions were conducted in public spaces on two college campuses, Luddy Hall at Indiana University in Bloomington and Dimond Library at the University of New Hampshire in Durham. Sessions were conducted according to a semi-structured protocol in three stages:

1. The facilitator briefly describes the premise for the study and guides the user through the interface. Then, the participant is asked to practice using it in the three different photo modes (as a standard smartphone camera, with aesthetics indicator visible and

with aesthetics capture activated).

2. Once the participant is comfortable using the interface, the participant is asked to wander around the space, take photos using each of the four models loaded onto the smartphone and think aloud to the facilitator about how the models take photos.

3. Once the participant affirms they have a good understanding of how the models are similar or different from one another, the participant and facilitator sit down and review all of the photos taken during the session and identify which models were responsible for the best and worst photos. Finally, the facilitator asks brief closing interview questions regarding similarities and differences between the models, specific peculiarities of each model, usability of the app and usefulness of qualitative evaluation for design.

Sessions were audio recorded, recordings were transcribed and all images taken by the research phone were saved. Our analysis followed a constructivist grounded theory approach [119, 58] through inductive content analysis [200, 98]. As this is a preliminary study, our goal is to develop a theoretical understanding of how participants interact with IAQA models, and establish methodological recommendations for future, similar studies. Grounded theory is an appropriate methodology for that task because it prioritizes theory generation over theory confirmation. Additionally, we collect multimodal data, including images and transcripts, which content analysis is well suited to approach. Practically, we took notes on our transcripts using a word processor comment feature about specific sections, cross-referencing saved images based on timestamps. After identifying several cross-session themes in open coding, we engaged in a second analysis of the interview transcripts and images to locate characteristic examples of each theme, with an emphasis on adjectives that participants used to describe models.

We see this study as a pilot, which serves as a proof of concept for our research method, and generates methodological insights for future similar studies. While five participants is not enough to make strong claims about the differences between the models or how users approach them, it is enough to confirm that the interface is functional, demonstrate that our qualitative approach can generate worthwhile insights and identify most of the interface's usability issues [247, 248].

8.5 Results

In this section, we describe the results of our user study sessions. We found, in line with other human-AI interaction research, that our participants tended to personify the algorithms they are asked to evaluate. However, they also treated them as puzzles, trying to guess how the different models were implemented. Participants additionally had a mixture of criticism and constructive feedback for both aesthetic quality assessment and our interface design, and their approaches reflected their differing personal relationships to smartphone cameras. We will discuss each of these themes in detail.

8.5.1 Personification and Reverse-Engineering

We found that three of our five participants (P1,P2,P4) had a tendency to personify both the models as well as the interface itself. For example, P4 observes that *"[model] B doesn't* *seem to be very excited, it's almost like, very stoic."* This observation reflects the fact that model B's output has lower variance than the others (see Figure 8.2; most output values fall around 50%, and P4 interpreted that quality as stoicism. P1 constantly turned to language around the models' likes and dislikes: *"C did not like this but A liked it...It definitely loves patterns. Like uniform patterns."* and interests: *"Now I cannot be too sure if it is [taking photos] because of the floor or if it is because of the chair because it was finding the floor very interesting."* When the interface crashed in the middle of the session, P2 responds, speaking to the app, *"You are unhappy."* While the way that computers function as social actors is well-known in HCI [243, 114, 95] and AI [95, 307], we emphasize that this personification is happening without the use of natural language or a human-like artificial persona. Just a letter name and a scalar measure of "preference" was enough to lead these participants towards these patterns.

To illustrate the differences our participants observed, we collected all of the words used to describe each model. We specifically find instances where participants describe the model's character or emotions, not its behavior, preferences or performance. Descriptive words are shown in the right column of Table 8.2. We can see that participants use terms like *"picky"* or *"stoic"* to describe model B's low variability, and terms like *"difficult to predict,"* *"random"* or *"like a cop camera"* to describe the high variability, and preference for red cars, of model D.

At the same time, all of our participants had a tendency to conceptually reverse-engineer the models, treating our task as a puzzle to be solved. For example, P5 offers a number of ideas: *"is that it? How close it is to the people?"* *"it's taking pictures of trees and...still* *objects,"* and *"does it also...try to understand the color?"* P3 speculates on the content of the training data: *"is it just trained on like landscapes and not people?"* P2 identifies that model B seems to be obeying photography rules: *"[model B] tries to actually pick larger objects that are sort of center focus, which would be like focal points...I have no idea if this uses like the rule of threes for how you frame up stuff or not."* Interestingly, Model B, based on Ke et al. [185], is explicitly designed based on these photographic rules of thumb. P2 also notes that model D prefers specific colors of cars: *"apparently it really likes blue or red vehicles. Which makes me think that's probably because, it's probably trained on those the most."*

Along these lines, P1 and P4 decided to engage in experiments, systematically varying specific factors to figure out what causes the model to take a photo. P1 wondered if model D knew the difference between a tree with leaves and a tree without: *"So it definitely detects plants...Let's try model D on a leafless plant. And see what it does...I don't want any building in the background, just the plant....It's because the background was like a little part of the building was coming in the picture. That's why it was clicking. It's not liking the tree at all."* Similarly, to investigate model B, P4 systematically rotated the camera while a regular brick pattern was in view: *"I have a hypothesis that maybe when it was like like right along the ground, like maybe when the lines were, sort of, in line. That's when it took it."*

This reverse-engineering approach led to feelings of confusion and disappointment when models C and D were revealed to be deep neural networks, not interpretable visual measures. While all our participants were aware of machine learning, it is unclear how familiar each one was with the limitations of different modeling approaches. P3 in particular dislikes that the models are not clearly nameable: *"Instead of model ABCD, how about you write something like model nature, model human, model bird?...Ok you give me three words, you know, what is model D's feature?"* P3 wants to evaluate the models in terms of what they actually measure, since she does not think there a single, general kind of aesthetic quality exists, and is frustrated that such descriptions are not possible for the deep neural network models. Interestingly, this lack of interpretability led P4 to express a desire to trust the model's taste over his own: *"assuming that machine learning models know much more than us, even though they might not think like us, but they have a larger set of data that's trying to feed them...I would want it to be more selective."* This comment echoes recent findings regarding a novel trend towards implicit trust of algorithmic systems as ultimate authorities [180].

When asked to choose their favorite model, three out of five participants chose model C, despite it not having the highest benchmark score. Their justifications centered around its "picky" behavior, taking fewer photos than other models. See Table 8.3 for details. To some extent, these differences in behavior which were important to our participants are visible in the output score distributions for the models shown in Figure 8.2.

8.5.2 Perspectives on Photography

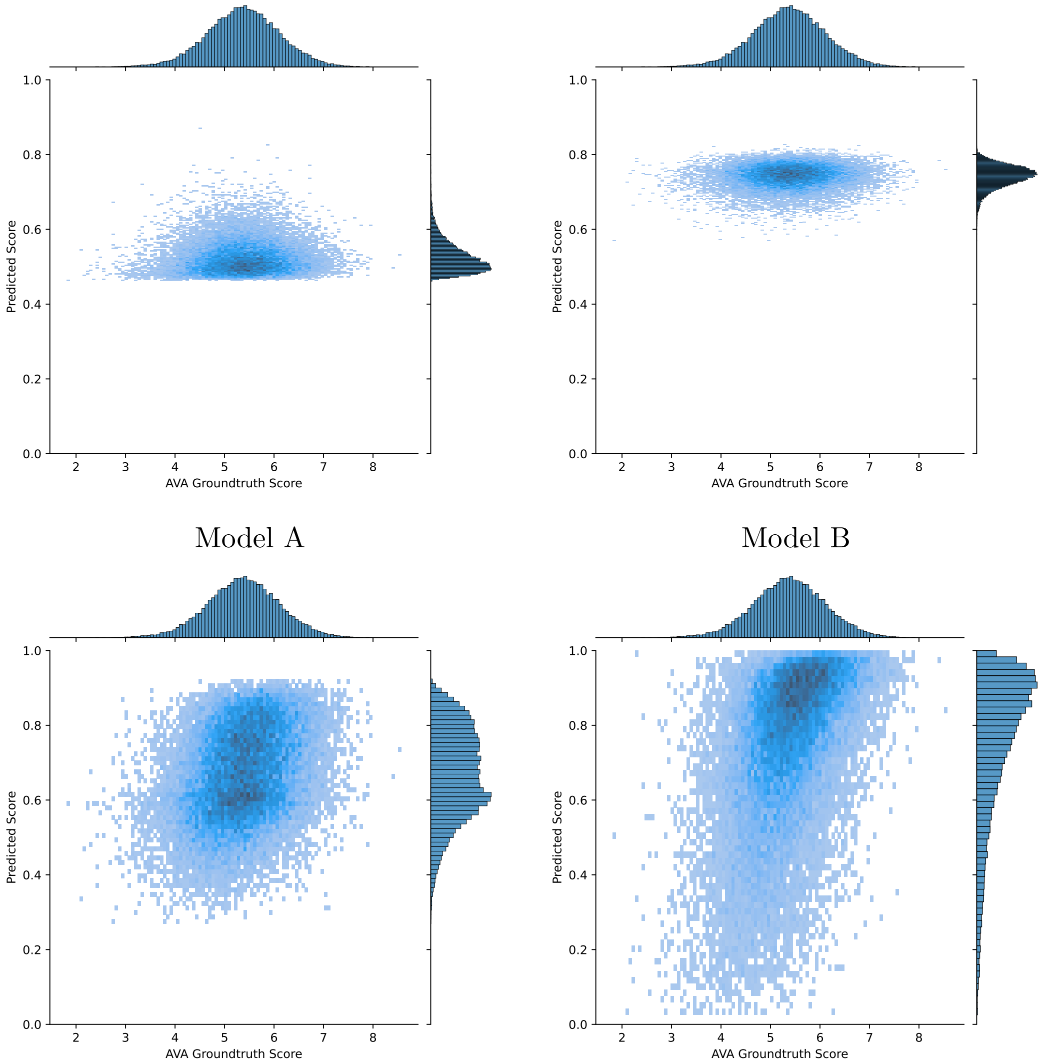
A recurring theme in our studies was each participant's personal relationship with smartphone photography and goals in taking pictures. These relationships vary considerably from person to person. One major difference in preferences surrounded the quantity of photos

Model	Acc.	ρ	AUC	Descriptions
A	0.600	0.047	0.530	*very selective* [P1], *greedy* [P2], *strange* [P4], *unreliable* [P4]
B	0.703	0.059	0.500	*low threshold* [P1], *picky* [P2,P4], *stoic* [P4], *unbothered* [P4]
C	0.708	0.296	0.546	*loves patterns* [P1], *picky* [P4], *understandable* [P5]
D	**0.740**	**0.473**	**0.605**	*unpredictable* [P1], *random* [P1], *like a cop camera* [P2], *likes most things* [P4], *object oriented* [P5]

Table 8.2: Accuracy, Ranking Correlation and AUC metrics for each model on the AVA test set, juxtaposed with the adjective descriptions used by our participants for each model.

Participant	Preference	Reasons
P1	B or D	*"I don't want a model that is so selective...I want to have pictures for me to sort and delete...D is at least choosing human faces, B is not even doing that."*
P2	C	*"you're not picking up everything, but you're also not having such a low reaction rate that you don't pick up anything."*
P3	None	*"I don't mind to click the shutter button because that is the certain moment and the angle I want to take it! I definitely need that moment! I don't want the camera to take it for me."*
P4	C	*"what I would want from that model is to take an unexpectedly nice picture, which means I would want it to be more selective if it's going in conjunction with the manual button."*
P5	C	*"model C is taking picture when there's some sort of like, nature...it's like a landscape photography...feature."*

Table 8.3: Responses when asked to choose a best model.



Model A

Model B

Model C

Model D

Figure 8.2: Output value vs. groundtruth joint and marginal distributions for each of our four aesthetic quality assessment models.

an automatic camera should take. P1 prefers taking more photos than necessary: *"I would rather want to have pictures there in my hand for me to sort and delete all the ones that are not good."* P3, on the other hand, dislikes taking too many photos and will even send friends blurry photos: *"Every photo is a good photo for me...if it is a little blurry, I still send because I don't want to take it the second time."* P2 observed that the photo quantity wasn't the only important factor: *"it's taking fewer...ok, it's taking photos of more things than I would, but it's like, if you're picking an object to take photos of, it's taking less photos than I would take of that object."* While he found some of those unusual photos worthwhile, *"I could use this in graphic design,"* most of them were unwanted.

Two participants referenced photo editing applications during the study. P3 references Meitu, an app described on the Google Play store as "Make your photos stunning and sensational! Whatever your beauty preference, do it all with Meitu!" [1]. P3 elaborates, *"You don't need to do any makeup, it's makeup for you! So a lot of girls including me like this because sometimes we don't need to make up, but we can make up here, you know. It makes me white, and it makes me clear, and it removed the dark part of my face or the environment or this is like, makes me younger."* The main appeal of this tool, for P3, is that it has a huge variety of filters and editor features so that each user can find the combination which looks best to them. She would not use any kind of photo tool unless it gave her that kind of aesthetic control.

P4 references several other applications, including the social media platform Instagram and VSCO, an app described on Google Play as "a leading photo and video editor that nurtures the creative journey with our library of 200+ premium quality presets and tools," [2].

P4 describes how he would edit one of the photos taken by our interface, *"this also could be considered aesthetic, like if I was trying to post this to Instagram, I'd like blow out the highlights and make it seem a little more make the background look a bit more even."* Editing is core to his photographic practice: *"Maybe for me, like I think of these as, like even the pictures that I click myself, I look at all of them as starting points and what can I make that goes beyond what I took."* This approach shapes the way he looks at the photos taken in this study: *"I don't understand some of the reasons for these shots. Like it could be made aesthetic...by aesthetic I mean things that could like possibly go on to Instagram."*

8.5.3 Perspectives on Aesthetic Quality Assessment

While it was not the primary topic of our study, several participants had positive feedback on our application design. P1 suggests that an automatic camera might be useful for people with disabilities, or to help take photos while in the car: *"if the model [is] distinguishing between the blurry and the non blurry and keeping only the one, that's good."* She remarked that the interface would be helpful for getting a *"human perspective"* on the functionality of each model. P2 similarly found it *"very helpful...you can actually get a feel for what it's doing."* P5 echoed that a buttonless camera would be helpful while driving, *"So you're driving in one hand, and if you're trying to take a picture of the scenery and other stuff, this feature will be really helpful."*

Participants also had a variety of criticism for our design and aesthetic quality assessment in general. P2 disliked the free-form nature of our study and recommended that we give specific tasks like in a photography class. He also was confused by the lack of effect exposure seemed to have on the different models. P5 found it difficult to tell the difference between models, and would prefer working with quantitative performance measures. P4 believes that the way computers and humans judge photos should remain complementary:

> *"I find it harder to figure out if a machine can think of aesthetics in the same way that humans do because, for a machine, [the photo] is the final picture, but for human that is not the final picture and we can always like step it up and make it look more interesting. So if it's a sort of discerning person, like probably a designer or a photographer, and they might just like be inspired... When working with AI...you work in conjunction, one doesn't replace the other and basically things that might take up a lot of time or like, instead of grunt work that you can leave to the AI and then you can use it as a sounding board or like an inspiration to get to something that's more refined and polished."*

In other words, even if an IAQA algorithm is used to evaluate photos, the final say regarding aesthetic quality should remain in the hands of the user. But other more tedious work like differentiating between blurry and non-blurry photos can be safely automated. Similarly, P3 was extremely critical of aesthetic quality assessment, as she believes that there is no need to waste research time on measures of aesthetics. Later, she explains her perspective:

> *So that shows the model's emotion? Then I need to satisfy the model not satisfy me. Yeah, this model used me, not I used the model...I need to understand the algorithm behind more like what makes this number peak? What brings the value down? If I don't know the calculation, I just don't understand what...[if] the algorithm of the model preference is not that good?...If the model itself is not good, I don't need to satisfy that model. Or maybe when, the moment I don't satisfy the model is the correct thing or is a good thing to do! You know, if the model is not the best one, there is no need to make it high."*

In other words, P3 gets to the heart of the concern shared by P4 regarding user control: if the camera has the final say on whether a photo is taken, it shifts the balance of power, placing the user's behavior under the algorithm's judgmental gaze.

8.6 Discussion

To summarize, we designed an interface for evaluating IAQA models in the real world, and conducted a pilot study of its effectiveness. We designed our interface as a camera application with no shutter button and settings for four models. In our pilot study, participants come to understand the different models by simultaneously personifying them and attempting to reverse-engineer their behavior. The resulting user epxeriences are shaped both by the implementation details of the models, as well as people's prior experience with smartphone cameras and relationships to images and aesthetics.

Personification is a well-known aspect of human-AI interaction. HCI research has shown that computers function as social actors [243], and this pattern extends to intelligent systems which function as media agents [114]. A similar phenomenon is well known in AI, called the ELIZA effect, after the 1964 text-based AI therapist developed by Joseph Wisenbaum, which human participants believed understood and had empathy for their problems even though it was only procedurally generating responses. Hamid Ekbia claims the Eliza effect is an example of a broader "attribution fallacy" where humans believe that a computer system has mental faculties and emotional states much like their own, even when the system demonstrably does not [95, p. 8]. While the ELIZA effect has been observed in chat programs, case-based reasoning systems [95, Ch.5] and social robots like Kismet [307], it is unusual in the scalar output of a binary classifier.

Participants' evaluations seem to have been shaped by the implementation differences between the models, the context and protocol of the study, as well as their differing backgrounds and prior relationships with smartphone cameras. Models A and B, due to their implementations, tended to produce values with lower variance, which participants read as selective, picky or stoic. Model D, which produced a wider range of values, was correspondingly seen as unpredictable or random. P1 and P4, who had computer science backgrounds, approached the evaluation process through experimentation, while P2, P3 and P5 took more observational approaches. The objects and backgrounds users had to photograph were implicitly determined by the environment: Luddy Hall in Bloomington was designed according to specific aesthetic principles, which came through in participant photographs. Finally, the design of the interface brought our work into comparison with other camera applications. P3 and P4 were frequent users of other smartphone photography tools, and compared our interface to those tools.

These differing interpretive factors are the heart of our concept of situated evaluation: participant evaluations are not fully determined by the objective characteristics of the models, but they are not fully subjective either. Instead, they are a product of the model performance, interface design, the research environment and our participants' backgrounds and approaches. In continuity with Haraway and her critique of god-tricks, we do not recommend attempting to eliminate these confounding factors. Instead, we recommend considering evaluations in context qualitatively. These contextual evaluations yield insights into our participants' evaluations of both the differing models, as well as the assumptions of IAQA.

These findings have a variety of limitations. Crucially, we cannot come to strong conclusions about the modeling work of other authors, as our models were not implemented exactly like the papers which inspired them. We also cannot make claims about which of these models is best aligned with computing graduate students' taste based on five participants. Given a larger sample, we might find that most participants agree in most contexts with the ordering created by accuracy on the AVA benchmark. But this qualitative approach still has the potential to offer contextual nuances which a scalar measure cannot provide.

We have several recommendations for future qualitative evaluations of IAQA models. First, we found that a semi-structured approach allowed the experimenter's off-hand remarks and follow-up questions to influence participant behaviors. The free-form nature of the sessions made participants a little uncomfortable, and they responded by trying to figure out what the facilitator wanted to hear. We recommend using a scripted and highly structured format, with specific tasks for the photography part of the session. We also recommend using as simple an interface as possible, rather than one which resembles a camera, to reduce confusion and potential crashes. To eliminate the confounding factor of different model output distributions, we recommend ensuring that all model outputs are near-identically distributed between 0 and 1, possibly by approximating the inverse cumulative distribution function for each model output distribution on a test set. Making multiple model outputs visible at the same time would also help participants to compare models with similar distributions: a slightly more complex design with multiple overlayed graphs could make it more obvious when a shift in the scene causes a higher result from one classifier and a lower result from another.

Finally, while we can center evaluation in the perspectives of people outside computer vision research, we cannot escape the implicit way that computer science approaches shape the development of models [85]. In other words, no matter how much feedback participants can offer, they will always be discussing measures developed by computer scientists. A more thoroughly feminist approach to this problem would take further steps to center views from the margins of computer science education, investigate participatory [277] and value sensitive [112] design approaches for IAQA. To take this decentering further, we also encourage the development of tools and platforms to allow individuals with minimal programming experience to develop measures of specific aesthetic qualities in images. This view of customized, rather than personalized, measures offers a more respectful, human-centered approach to subjectivity in these problems.

Chapter 9

Discussion and Future Work

In the previous chapters, we have discussed aesthetic phenomenon problems in computer vision and the difficulties which arise when evaluating them. In Chapters 3, 4 and 5 we gave examples of research studies which involve operationalizing the experiential qualities of image data, resulting in aesthetics phenomenon problems. In each, we showed ways to utilize additional qualitative and quantitative data sources to avoid the difficult problem of evaluating these algorithms directly. In Chapters 6, 7 and 8, we focused in on IAQA and interrogated ways of grappling with subjectivity in our evaluation. Our claim is that quantitative metrics computed on benchmarks only give us partial knowledge of the quality of our solutions to these problems: they can measure how well our metric aligns with broad trends in the decontextualized preferences of a population, but adequately considering subjective and contextual differences is much more challenging. To better account for these factors, we advocate qualitative evaluations, using methods from HCI, to consider IAQA models in context. In this chapter, we recapitulate several key themes, and for each one, offer recommendations for computational study of aesthetics phenomenon problems and possibilities for future work.

9.1 New Approaches to Cultural Images

In Chapter 3, we study the homogenization of web design using both measures of visual similarity alongside interviews with veteran web designers. While confirming an overall homogenization trend was our main result, we also arrived at a variety of more specific, nuanced findings — specific software library usage, increased use of images for color instead of backgrounds, the rise of high-fidelity prototyping tools and Google's shift to prioritize mobile — by combining several computational analyses on different definitions of the problem under study in the context provided by ethnographic interviews.

Similarly, in Chapter 4, we were able to approach a centuries-old question about color harmony from a new perspective: instead of only philosophizing or looking at human preferences for color schemes in a lab setting, we ask whether people actually use color schemes with these templates in images of web design, fashion and painting. While such inquiry does not lead to timeless scientific or philosophical truths, and there is no way to establish general claims about the way people use color based on photos from the Internet, it allows us to center our analysis in the context of real images.

Third, in Chapter 5 we approach the difficult issue of periods in art history from a new perspective, utilizing Bayesian methods to show the inherent uncertainty of such periods, and encode subjective degrees of belief, both in the lives of a specific artist and in a historical period. While these approaches are not a path to more objective category systems for art, they allow us to approach these systems as models, suggestions and ways of seeing which we can integrate into our qualitative models.

These three studies show examples of how computational methods can allow us to approach old problems in new ways. Instead of trampling over existing ways of seeing and knowing about culture, these methods allow us see differently, though modeling and visualization, and update our mental models to inform more specific inquiry. Aesthetic measures, rather than a way of automating away critical humanistic inquiry, serve as texts, grounding that inquiry. In my opinion, the process of debating concepts like layout similarity or visual complexity and how measures of them should actually work is a refreshing, contemporary way to think about visual experience.

There are also a wide variety of less-subjective ways that we can apply computation to cultural data. More geometric vision approaches, like those discussed at the start of Chapter 2 [293, 159, 32, 69], still have unstudied potential for analysis of art images. Visual intertextuality [109] and network analysis [238] are other area where computers can help us to understand old topics differently. While we have not considered these areas in our inquiry here, we would encourage taking a similarly skeptical approach towards objective truth about art in these contexts as well.

Unfortunately, participation in debates around computational approaches are only really open to those who are fluent in the language of computational modeling, and the culture of computing. While requiring scholars to learn an unfamiliar language and culture to participate in scholarly discourse is nothing new in the academic humanities, there is a need for future research on accessibility in this domain. Computational image analysis is significantly more difficult for non-experts to learn than similar fields like computational text analysis, in part because the methods are younger and less work has been done on making them available to a wide audience. We strongly encourage future work on tools for computational image analysis which are accessible to non-experts, and especially non-programmers.

9.2 Who is the Subject?

In Chapter 6, we raised a key question: when we say IAQA is a subjective problem, who is the subject? In the personalized statement of the problem discussed in Chapter 7, the subject is the user, and the modeling process attempts to predict specifically how they will feel about a given image. But for applications of IAQA where there is no single human user (for example, when evaluating image processing algorithms), this philosophical solution no longer works. In line with Haraway's critique of "god-tricks" in science and the feminist critique of Kantian disinterestedness, we believe assertions that a computational aesthetic measure is objective or universal amounts to computational construction of taste, where we elevate an uncertain and limited way of understanding images to the status of objective truth. Instead, models based on these assumptions predict a kind of popularity, which ignores individuals whose tastes deviate from the norm.

One could also argue that the algorithm itself is the subject. As discussed in Chapter 6, this requires making an artificial intelligence claim, that the computation happening inside a particular aesthetic measure is similar to the computation happening in a human brain. Our claims regarding personalization make this perspective more appealing — there are many algorithms which may resemble human emotional responses to images because there are many ways that humans emotionally respond to images, and even more ways that an algorithm could respond which are similar. However, since an aesthetic measure is not a person in any other sense, it is unclear why we would give weight to the output of an aesthetic measure which is only justified through its own taste. As P3 from Chapter 8 so succinctly put it, *"If the model itself is not good, I don't need to satisfy that model."*

Instead, we would venture that regardless of whose preference data is collected, the subject in IAQA is always at least partially the human researcher who decides how to frame the problem. The researcher has full design agency to shape the metric they develop according to their personal photographic preferences. This agency can be expressed in explicit ways, like how Ke et al. reference specific photographic rules of thumb, or more implicitly [185], like how Lu et al. discard the image categories from the AVA dataset [214].

In many cases, it is acceptable or even advantageous for the researcher to express their subjectivity through the design of an IAQA model. If the target use of the model is in an automatic editing tool, for example, having one specific style that a tool targets is an important design choice which can help establish an identity for the product. But in other contexts, we may want to decenter the researcher, avoid relying too heavily on their concept of aesthetic quality and model the taste of a particular human population instead. In that context, we should rely on qualitative research techniques rooted in ethnography and participatory design [200, 277] to incorporate the mental models of individuals in the target population into our computational modeling process. As discussed in Chapter 8, a gold standard for decentering the researcher would be creating easy-to-use tools for designing aesthetic measures which members of the target population could themselves use. We advocate for more research into customizable approaches for these problems, which give non-expert users the ability to design algorithms for these problems based on their perspectives, as opposed to personalized models which attempt to account for user subjectivity automatically. Design of these tools, as well as investigation of how non-experts choose to (or not to) measure perceptual qualities, would be an excellent direction for future research into aesthetic phenomenon problems.

9.3 Uncertainty and the Culture of Computer Vision

Another central theme has been the epistemic uncertainty inherent in aesthetic phenomenon problems. Uncertainty is not a component of Haraway's approach to knowledge, but it is a central addition of Drucker [91] and D'Ignazio and Klein [94], who use uncertainty to avoid false claims of objectivity in visualization and mapping. In Chapter 5, we connect their use of uncertainty to Bayesian statistics, which has its own tradition of quantifying subjective degrees of belief. Bayesian methods are typically employed in machine learning research to quantify the "beliefs" of artificial agents as they interact with a difficult-to-sense world, but they can also be useful for quantifying our beliefs as researchers about the problems that we are studying.

While computer vision is well-acquainted with uncertainty in data, uncertainty in our modeling and experimentation is a much newer concept. Just a few decades ago most computer vision research studied methods with mathematical performance guarantees, which are largely not subject to the feminist critiques of knowledge we discuss here. Recently, however, evaluation has shifted towards metrics computed on empirical benchmarks; but the assumption of exact, objective knowledge regarding algorithm performance has remained in place. This assumption is reinforced in computer vision papers using the visual rhetoric of the "results table," with a firm boundary between the "state-of-the-art" result and the rest.

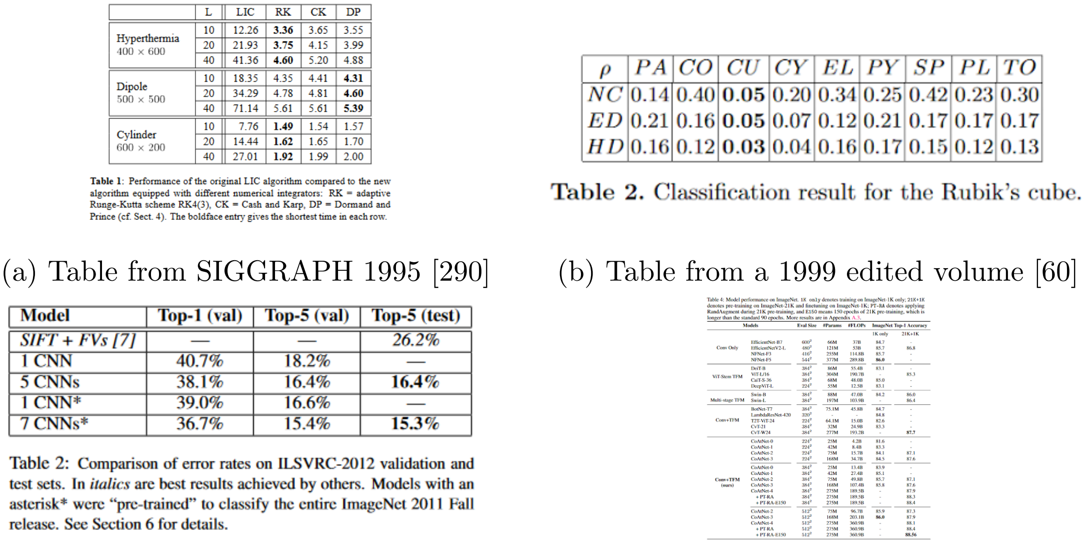
In a recent study of computer vision researchers[1] [124], we traced this shift. One participant, a senior computer vision researcher, had a combination of quantitative and qualitative evaluation in her paper from 2003: *"quantitative evaluation, you know, back in 2003 was still kind of in its infancy...I'm not sure that this [2003] paper has basically any comparison to competing methods which probably would be required today."* A second explains that in 1999, showing example output of his system was sufficient: *"instead of [Amazon] Mechanical Turk you just have the reviewers just eyeball the images."* In a recent blog post, Aaron Hertzmann echoes this sentiment: papers which present methods with clearly visible results are now asked by reviewers to include user studies to show that their work outperforms prior work [152]. In the satirical 2010 "Paper Gestalt" [314] paper, which attempts to use computer vision methods to distinguish between good and bad papers, large confusing tables were identified as a key feature of *bad* papers, not an essential feature of good ones.

So how did computer vision transform from a mathematical discipline based on geometry to an empirical discipline based on benchmarks? We can see the seeds of this transition as early as a debate at ICCV 1999 between Jitendra Malik and Olivier Faugeras [306]. In that debate, Malik argued that geometric methods had reached their limits, and computer vision should focus more on probabilistic modeling of perceptual factors, while Faugeras

[1] A paper about this study, P7, is currently under submission and available as a preprint at `https://arxiv.org/abs/2209.11200`. This study is not included as a chapter in this thesis due to its limited relevance to aesthetic phenomenon problems.

responded that empirical computer vision was unscientific, since it is unfalsifiable, and geometric methods based on rigorous mathematics were a better foundation for the discipline. Malik rebuts, arguing that computer vision is not a science, but a hybrid of mathematics, science and engineering. Regardless of who was right at the time, Malik's position was repeatedly justified by successes in probabilistic and learning-based computer vision shortly after.

The publication of Krizhevsky, Sutskever and Hinton's "ImageNet Classification with Deep Convolutional Neural Networks" in 2012 [194] marks a turning point for empirical evaluation. This paper is historically significant for setting off the deep learning revolution, and its design and writing served as a foundation for the thousands of deep learning-based computer vision papers that followed. The paper's central argument is that several "new and unusual features" lead deep convolutional neural networks to significantly outperform other methods. These features include rectified linear units (ReLU), GPU-based training, and regularization techniques like data augmentation and dropout. Neural network papers were obligated to use empirical evaluation, as there are insufficient theoretical guarantees for these models and they are difficult to compare otherwise. Over the following years, many papers followed, showing that deep convolutional neural networks outperform existing methods on other central problems like object detection and semantic segmentation. Because they follow Krizhevsky's argumentative form, these papers use comparisons on benchmarks to show their effectiveness. This style of table, with one number in bold, actually arises in computer graphics before entering computer vision; see Figure 9.1 (a) for an early example. Early graphics results tables primarily showed runtime comparisons, rather



	L	LIC	RK	CK	DP
Hyperthermia 400 × 600	10	12.26	**3.36**	3.65	3.55
	20	21.93	**3.75**	4.15	3.99
	40	41.36	**4.60**	5.20	4.88
Dipole 500 × 500	10	18.35	4.35	4.41	**4.31**
	20	34.29	4.78	4.81	**4.60**
	40	71.14	5.61	5.61	**5.39**
Cylinder 600 × 200	10	7.76	**1.49**	1.54	1.57
	20	14.44	**1.62**	1.65	1.70
	40	27.01	**1.92**	1.99	2.00

Table 1: Performance of the original LIC algorithm compared to the new algorithm equipped with different numerical integrators: RK = adaptive Runge-Kutta scheme RK4(3), CK = Cash and Karp, DP = Dormand and Prince (cf. Sect. 4). The boldface entry gives the shortest time in each row.

(a) Table from SIGGRAPH 1995 [290]

ρ	PA	CO	CU	CY	EL	PY	SP	PL	TO
NC	0.14	0.40	**0.05**	0.20	0.34	0.25	0.42	0.23	0.30
ED	0.21	0.16	**0.05**	0.07	0.12	0.21	0.17	0.17	0.17
HD	0.16	0.12	**0.03**	0.04	0.16	0.17	0.15	0.12	0.13

Table 2. Classification result for the Rubik's cube.

(b) Table from a 1999 edited volume [60]

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

(c) Table from NeurIPS 2012 [194]

Table 4: Model performance on ImageNet. 1K only denotes training on ImageNet-1K only; 21K+1K denotes pre-training on ImageNet-21K and finetuning on ImageNet-1K; PT-RA denotes applying RandAugment during 21K pre-training, and E150 means 150 epochs of 21K pre-training, which is longer than the standard 90 epochs. More results are in Appendix A.3.

Models		**Eval Size**	**#Params**	**#FLOPs**	**ImageNet Top-1 Accuracy**	
					1K only	21K+1K
Conv Only	EfficientNet-B7	600^2	66M	37B	84.7	-
	EfficientNetV2-L	480^2	121M	53B	85.7	86.8
	NFNet-F3	416^2	255M	114.8B	85.7	-
	NFNet-F5	544^2	377M	289.8B	**86.0**	-
ViT-Stem TFM	DeiT-B	384^2	86M	55.4B	83.1	-
	ViT-L/16	384^2	304M	190.7B	-	85.3
	CaiT-S-36	384^2	68M	48.0B	85.0	-
	DeepViT-L	224^2	55M	12.5B	83.1	-
Multi-stage TFM	Swin-B	384^2	88M	47.0B	84.2	86.0
	Swin-L	384^2	197M	103.9B	-	86.4
Conv+TFM	BotNet-T7	384^2	75.1M	45.8B	84.7	-
	LambdaResNet-420	320^2	-	-	84.8	-
	T2T-ViT-24	224^2	64.1M	15.0B	82.6	-
	CvT-21	384^2	32M	24.9B	83.3	-
	CvT-W24	384^2	277M	193.2B	-	**87.7**
Conv+TFM (ours)	CoAtNet-0	224^2	25M	4.2B	81.6	-
	CoAtNet-1	224^2	42M	8.4B	83.3	-
	CoAtNet-2	224^2	75M	15.7B	84.1	87.1
	CoAtNet-3	224^2	168M	34.7B	84.5	87.6
	CoAtNet-0	384^2	25M	13.4B	83.9	-
	CoAtNet-1	384^2	42M	27.4B	85.1	-
	CoAtNet-2	384^2	75M	49.8B	85.7	87.1
	CoAtNet-3	384^2	168M	107.4B	85.8	87.6
	CoAtNet-4	384^2	275M	189.5B	-	87.9
	+ PT-RA	384^2	275M	189.5B	-	88.3
	+ PT-RA-E150	384^2	275M	189.5B	-	88.4
	CoAtNet-2	512^2	75M	96.7B	85.9	87.3
	CoAtNet-3	512^2	168M	203.1B	**86.0**	87.9
	CoAtNet-4	512^2	275M	360.9B	-	88.1
	+ PT-RA	512^2	275M	360.9B	-	88.4
	+ PT-RA-E150	512^2	275M	360.9B	-	**88.56**

(d) Table from NeurIPS 2021 [73]

Figure 9.1: Visual development of the results table in computer vision. (a) is an early example from computer graphics. (b) is an early example from computer vision. (c) is from the highly influential 2012 AlexNet ImageNet classification paper [194], (d) is a 2021 state-of-the-art result on ImageNet, requiring a much larger number of comparisons [73].

than accuracy or quality evaluations. These tables start to appear in the computer vision literature for showing machine learning performance at least as early as 1999 (Figure 9.1 (b)).

Today, competition on major vision problems is fierce. For example, compare the table in Figure 9.1 (c), from a 2012 paper, to the table in Figure 9.1 (d), from a 2021 paper. The benchmark remains ImageNet, though performance has surged from 40% top-1 accuracy to over 85%, but the competing state-of-the-art includes dozens of models and differ by only fractions of a percent. While it is indisputable that the surge in accuracy values corresponds to an increase in performance, given the uncertain relationship between the ImageNet benchmark and real-world image classification problems, especially the subjective nature of our categorization systems themselves [44, 85, 34], simply measuring the highest scalar performance value should not guarantee that we have found a state-of-the-art approach. But each successive paper asserts that their approach is the state-of-the-art, putting their accuracy value in bold.

We encourage future work throughout machine learning exploring Bayesian approaches to subjectivity and epistemic uncertainty. These statistical methods allow us to relate uncertainty to subjective degrees of belief. We also encourage work connecting these ideas to perspectivist approaches to data annotation [24, 25]. These methods seem to be natural allies for foregrounding the evidence we have for our modeling decisions, as well as our degree of uncertainty (or certainty) around model performance across computer vision problems. While it may seem that benchmark-based evaluation is simply the way that research is done in computer vision, these developments are relatively recent. It is possible for researchers to adopt alternative evaluation methods and argumentative forms when appropriate.

9.4 Aesthetic Measures and Optimization Beyond Objective Evaluation

A topic which has come up several times is the relationship between aesthetic measures and the images which optimally activate them. Hume took issue with the idea of reducing judgment to geometric principles precisely because they imply an optimal, ideal beauty [156]. Galanter develops his theory of computational aesthetics with the expressed purpose of evolving works of art using genetic algorithms to optimize for them [113]. Some of our participants in Chapter 8 tried to methodically optimize for each measure in order to determine its taste. And in Chapter 6, we observe a characteristic style of dramatic nature photo, which achieves the highest ratings in the AVA dataset, which aligns well with the explanatory regression coefficients from Chapter 7. While the idea of optimizing for a particular measure resembles economic or managerial decision-making which is unacceptable to many artists and humanists, there may be interesting avenues for research into tools based on these ideas, even if the functions involved are not objective measures of any specific aesthetic phenomenon.

For example, these ideas lead towards a hypothesis: that we can operationalize difficult-to-describe visual styles using a mapping between computational measures and the images which optimally activate them. For example, given a large basis set of images, we could visualize a candidate measure by retrieving the images which optimally activate it, or select a measure by first selecting several images from this basis set and searching for a measure which rates those images highest. In fact, the $N \times M$ matrix of scores assigned to N images by M aesthetic measures resembles the user-content matrix studied in collaborative filtering [12, 285], as well as recent analysis of complexity measures as image features by Karjus et al. [181]. Future study of aesthetic measures as a space of style features which could be used to develop measures for specific desired, but difficult to describe, design aesthetics is a worthwhile area of inquiry.

An additional possible application area is in the understanding and explanation of image processing algorithms. Metrics for more objective concepts of image quality and degradation are already used in this space for evaluation [174, 145], but we advocate for something less normative. Aesthetic measures could be used to describe and explain how new image transformations, especially those which are parameterized by deep neural networks, actually affect the perceptual qualities of images, either individually or as an interpretable vector showing differences on several metrics, much like classeme features [305] are assembled out of weak image classifiers. These metrics should not be treated as an objective replacement for human evaluation, but can be used descriptively to quantify image transformations. Conversely, interpretable image transformations could be used to describe difficult-to-interpret aesthetic measures by applying transformations to optimize images for that measure.

9.5 Aesthetic Phenomenon Problems and Ethics

Our final key theme is one which is under-discussed in the prior chapters: the relationship between aesthetics and ethics, and the relationship between aesthetic phenomenon problems and larger ethical issues in computer vision. While discussing the long relationship between ethics and aesthetics is well beyond the scope of this thesis, it is worth noting that ethics and aesthetics both require us to make judgments, and it is easy for matters of taste to shape our ethical stances or for our moral judgments to affect our interpretations of works of art.

Centrally, aesthetic quality assessment algorithms differ from other social scientific or neuroscientific study of perception and taste because they do not simply describe how people have made judgments regarding aesthetic phenomena, they ask how computers *should* make those judgments in the future. Implicit in our notion of the aesthetic phenomenon problem is the assumption that computers should make new judgments similarly to the ways that humans have made judgments in the past. This assumption can lead, as it has in topics like facial recognition [54], to reproduction of historical biases if we do not actively intervene. As discussed in Chapters 5 and 7, the artistic canon shapes datasets like WikiArt and reflects historical concepts of beauty and genius intimately tied to race, nationality and gender [26, 190, 82, 319]. As developers of these models and algorithms, we have a degree of creative freedom to choose how the aesthetic qualities we seek to measure should be defined, and a responsibility to be upfront with our degree of authorship over our approaches. Attempting to model the contents of our data as neutrally as possible is not a neutral option in this context, as every aesthetic measure makes implicit normative judgments.

There is a serious danger, if we overstate the neutrality or certainty of our approaches to aesthetic phenomenon problems, that they will be misinterpreted as objective approaches to taste and beauty. In Chapter 8, P4 falls victim to this misinterpretation: *"assuming that machine learning models know much more than us, even though they might not think like us, but they have a larger set of data that's trying to feed them...I would want it to be more selective."* It is very tempting to trust artificial agents as omniscient aesthetic oracles, who have some inscrutable quantitative knowledge about what we *should* consider high quality, instead of interpolated knowledge about what we *have* considered high quality. Such approaches play into pseudo-scientific trends, such as the mythology surrounding the golden ratio as a numerical secret to beauty in art and architecture [228].[2]

[2] George Markowsky debunks this mythology. He points out that the golden ratio is an irrational number,

We strongly oppose the use of aesthetic measures to evaluate the work of human artists for economic or governance purposes. As we discussed in Chapter 1, such measures are dehumanizing and commodifying. Supposedly objective definitions of beauty can be highly politically charged, especially when concepts of ideal beauty intersect with femininity and race [240] or artistic value [26] to create narratives of cultural supremacy or decline. This thesis does not delve deeply into these issues, but we encourage future study from more explicitly socio-political perspectives on aesthetic phenomenon problems, especially related to the computational construction of taste discussed in Chapter 7.

Finally, we have only briefly touched issues regarding the politics of data collection for aesthetic phenomenon problems. In Chapter 8, we sketched an argument for how qualitative evaluation can challenge the imbalanced power relationship between researchers and participants, especially on crowd work platforms, we encourage further study and pursuit of that line of inquiry. We have not discussed the ethics of collecting large cultural image datasets. While increasing dataset size may help to remedy the biases present in smaller datasets and produce more reliable quantitative evaluations for these problems, there are difficult questions regarding both privacy and surveillance which arise in this area. These topics have been the subject of active debate for the past decade [45], and are not yet resolved [195]. Further inquiry into the privacy status of cultural images is needed in order to better govern work in this area.

and any attempt to measure it in art or architecture is at most wishful thinking [228].

9.6 Conclusion

As artificial intelligence technologies improve, the boundaries between subjective and objective qualities will continue to blur. This thesis takes a first step towards naming a collection of problems at that boundary and approaching them from a feminist, human-centered perspective. Despite the difficulties of working with these problems, and the unusual combinations of research methods they afford, we emphasize that computational methods offer fascinating, radically new ways of seeing culture. Pushing at the boundaries of evaluation in computer vision creates new opportunities for us to think more deeply about how we personally experience the visual world and how we can express different ways of seeing computationally.

Bibliography

[1] "Meitu photo editor and ai art on google play," https://play.google.com/store/apps/details?id=com.mt.mtxx.mtx retrieved May 15th 2023.

[2] "Vsco: Photo & video editor on google play," https://play.google.com/store/apps/details?id=com.vsco.cam retrieved May 15th 2023.

[3] "webmuseum.dk," 2009. [Online]. Available: http://webmuseum.dk/

[4] "Longing for innovation: Why do all websites look the same?" Article on webydo.com, Mar 2016, retrieved from https://cmd-t.webydo.com/longing-for-innovation-why do-all-websites-look-the-same-9c80f5c41c61?gi=25360bd9ad1b.

[5] "Why all new websites look the same," Blog post on fontbundles.net, May 2016, retrieved from https://fontbundles.net/blog/why-all-new-websites-look-the-same.

[6] "Adobe color," 2019, https://color.adobe.com/create.

[7] "Alexa: About us," https://www.alexa.com/about, 2019.

[8] "The internet archive," http://www.archive.org, 2019.

[9] "Phantomjs," http://www.phantomjs.org, 2019.

[10] "Wappalyzer," https://www.wappalyzer.com/, 2019.

[11] "Why do all websites look the same?" Blog post on bigtuna.com, 2019, retrieved from https://bigtuna.com/why-do-all-websites-look-the-same/.

[12] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE transactions on knowledge and data engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[13] J. Albers, *Interaction of color.* Yale Press, 1975.

[14] M. Anderson, R. Motta, S. Chandrasekar, and M. Stokes, "Proposal for a standard default color space for the internet-srgb." in *Color Imaging Conference*, vol. 6, 1996.

[15] M. Armstrong, ""the effects of blackness": Gender, race, and the sublime in aesthetic theories of burke and kant," *The Journal of Aesthetics and Art Criticism*, vol. 54, no. 3, pp. 213–236, 1996.

[16] R. Arnheim, *Art and visual perception: A psychology of the creative eye.* Univ of California Press, 1965.

[17] ——, "The other Gustav Theodor Fechner." in *A century of psychology as science*, S. Koch and D. E. Leary, Eds. American Psychological Association, 1985.

[18] A. Badano, C. Revie, A. Casertano, W.-C. Cheng, P. Green, T. Kimpe, E. Krupinski, C. Sisson, S. Skrøvseth, D. Treanor *et al.*, "Consistency and standardization of color in medical imaging: a consensus report," *Journal of digital imaging*, vol. 28, pp. 41–52, 2015.

[19] S. Bambach, S. Lee, D. J. Crandall, and C. Yu, “Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions,” in *ICCV*, 2015, pp. 1949–1957.

[20] J. Bardzell and S. Bardzell, “What is” critical” about critical design?” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2013, pp. 3297–3306.

[21] S. Bardzell, “Feminist hci: taking stock and outlining an agenda for design,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2010, pp. 1301–1310.

[22] S. Bardzell and J. Bardzell, “Towards a feminist hci methodology: social science, feminism, and hci,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 675–684.

[23] S. Bardzell, J. Bardzell, J. Forlizzi, J. Zimmerman, and J. Antanitis, “Critical design and critical theory: the challenge of designing for provocation,” in *Proceedings of the designing interactive systems conference*, 2012, pp. 288–297.

[24] V. Basile, F. Cabitza, A. Campagner, and M. Fell, “Toward a perspectivist turn in ground truthing for predictive computing,” *arXiv preprint arXiv:2109.04270*, 2021.

[25] V. Basile, T. Caselli, A. Balahur, and L.-W. Ku, “Bias, subjectivity and perspectives in natural language processing,” *Frontiers in Artificial Intelligence*, p. 116, 2022.

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

[27] R. Bellman, "On the approximation of curves by line segments using dynamic programming," *Communications of the ACM*, vol. 4, no. 6, p. 284, 1961.

[28] A. Ben-David, A. Amram, and R. Bekkerman, "The colors of the national Web: Visual data analysis of the historical Yugoslav Web domain," vol. 19, no. 1, pp. 95–106, 2016.

[29] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, 2018.

[30] J. J. Benjamin, A. Berger, N. Merrill, and J. Pierce, "Machine learning uncertainty as a design material: a post-phenomenological inquiry," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, 2021, pp. 1–14.

[31] J. Berger, *Ways of Seeing.* Penguin, 1972.

[32] F. Bernardini, H. Rushmeier, I. M. Martin, J. Mittleman, and G. Taubin, "Building a digital model of michelangelo's florentine piet?" *IEEE Computer Graphics and Applications*, vol. 22, no. 01, pp. 59–67, 2002.

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

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

[35] K.-K. Bhavnani, "Tracing the contours: Feminist research and feminist objectivity," in *Women's Studies International Forum*, vol. 16, no. 2. Elsevier, 1993, pp. 95–104.

[36] F. Bianchi, "Coolors," 2014, https://coolors.co/.

[37] G. Bignardi, T. Ishizu, and S. Zeki, "The differential power of extraneous influences to modify aesthetic judgments of biological and artifactual stimuli," *PsyCh Journal*, 2020.

[38] A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?" in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV).* IEEE, 2021, pp. 1536–1546.

[39] G. D. Birkhoff, *Aesthetic measure.* Harvard University Press, 1933.

[40] C. Bishop, "Against digital art history," *International Journal for Digital Art History*, no. 3, 2018.

[41] R. L. Blaszczyk, *The color revolution.* MIT Press, 2012.

[42] U. Böckenholt, "Thresholds and intransitivities in pairwise judgments: A multilevel analysis," *Journal of Educational and Behavioral Statistics*, vol. 26, no. 3, pp. 269–282, 2001.

[43] E. G. Boring, "The beginning and growth of measurement in psychology," *Isis*, vol. 52, no. 2, pp. 238–257, 1961.

[44] G. C. Bowker and S. L. Star, *Sorting things out: Classification and its consequences.* MIT press, 2000.

[45] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, communication & society*, vol. 15, no. 5, pp. 662–679, 2012.

[46] R. M. Boynton, "History and current status of a physiologically based system of photometry and colorimetry," *JOSA A*, vol. 13, no. 8, pp. 1609–1621, 1996.

[47] L. Brady and C. Phillips, "Aesthetics and usability: A look at color and balance," *Usability News*, vol. 5, no. 1, February 2003.

[48] E. Brage, "The rise of brutalism and antidesign," 2019. [Online]. Available: http://www.diva-portal.org/smash/get/diva2:1304924/FULLTEXT01.pdf

[49] J. E. Breslin, *Mark Rothko: a biography.* University of Chicago Press, 1993.

[50] R. A. Brooks, "New approaches to robotics," *Science*, vol. 253, no. 5025, pp. 1227–1232, 1991.

[51] N. Brügger, *The Archived Web: Doing History in the Digital Age.* The MIT Press, 2018.

[52] N. Brügger and R. Schroeder, *The Web as History: Using Web Archives to Understand the Past and the Present.* London: UCL Press, 2017.

[53] M. Buda, A. Maki, and M. A. Mazurowski, “A systematic study of the class imbalance problem in convolutional neural networks,” *Neural Networks*, vol. 106, pp. 249–259, 2018.

[54] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *FAccT.* PMLR, 2018, pp. 77–91.

[55] B. Caldwell, M. Cooper, L. G. Reid, and G. Vanderheiden, “Web content accessibility guidelines (wcag) 2.0,” *WWW Consortium (W3C)*, 2008.

[56] G. Castellano and G. Vessio, “Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview,” *Neural Computing and Applications*, vol. 33, no. 19, pp. 12 263–12 282, 2021.

[57] J. Chai, Q. Lu, Y. Hu, S. Wang, K. K. Lai, and H. Liu, “Analysis and bayes statistical probability inference of crude oil price change point,” *Technological Forecasting and Social Change*, vol. 126, pp. 271–283, 2018.

[58] K. Charmaz, *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis.* Pine Forge Press, 2006.

[59] A. Chatterjee and O. Vartanian, “Neuroscience of aesthetics,” *Annals of the New York Academy of Sciences*, vol. 1369, no. 1, pp. 172–194, 2016.

[60] A. Chella, V. D. Gesù, I. Infantino, D. Intravaia, and C. Valenti, "A cooperating strategy for objects recognition," in *Shape, Contour and Grouping in Computer Vision.* Springer, 1999, pp. 264–274.

[61] W. Chen, D. J. Crandall, and N. M. Su, "Understanding the aesthetic evolution of websites: Towards a notion of design periods," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 5976–5987. [Online]. Available: http://doi.acm.org/10.1145/3025453.3025607

[62] H. Chernoff and S. Zacks, "Estimating the current mean of a normal distribution which is subjected to changes in time," *The Annals of Mathematical Statistics*, vol. 35, no. 3, pp. 999–1018, 1964.

[63] M. J. Chong and D. Forsyth, "Effectively unbiased fid and inception score and where to find them," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6070–6079.

[64] A. Cocciolo, "The Rise and Fall of Text on the Web: A Quantitative Study of Web Archives," vol. 20, no. 3, 2015. [Online]. Available: https://eric.ed.gov/?id=EJ1077827

[65] M. Collins, "Web design trends: Why do all websites look the same?" Blog post on friday.ie, April 2016, retrieved from https://www.friday.ie/blog/why-do-all-websites-look-the-same/.

[66] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2022.

[67] D. Cosgrove and S. Daniels, *The iconography of landscape: essays on the symbolic representation, design and use of past environments*. Cambridge University Press, 1988, vol. 9.

[68] D. Cosley, S. K. Lam, I. Albert, J. A. Konstan, and J. Riedl, “Is seeing believing? how recommender system interfaces affect users’ opinions,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003, pp. 585–592.

[69] D. Crandall and N. Snavely, “Networks of photos, landmarks, and people,” *Leonardo*, vol. 44, no. 3, pp. 240–243, 2011.

[70] C. Cui, H. Fang, X. Deng, X. Nie, H. Dai, and Y. Yin, “Distribution-oriented aesthetics assessment for image search,” in *ACM SIGIR RDIR*, 2017, pp. 1013–1016.

[71] C. Cui, W. Yang, C. Shi, M. Wang, X. Nie, and Y. Yin, “Personalized image quality assessment with social-sensed aesthetic preference,” *Information Sciences*, vol. 512, pp. 780–794, 2020.

[72] D. Cyr, M. Head, and A. Ivanov, “Design aesthetics leading to m-loyalty in mobile commerce,” *Information & management*, vol. 43, no. 8, pp. 950–963, 2006.

[73] Z. Dai, H. Liu, Q. V. Le, and M. Tan, “Coatnet: Marrying convolution and attention for all data sizes,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 3965–3977, 2021.

[74] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *ECCV*, 2018, pp. 720–736.

[75] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*. Springer, 2006, pp. 288–301.

[76] R. Datta, J. Li, and J. Z. Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: An exposition," in *ICIP*. IEEE, 2008, pp. 105–108.

[77] N. Davis, "Human-computer co-creativity: Blending human and computational creativity," in *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 9, no. 6, 2013, pp. 9–12.

[78] J. De La Rosa and J.-L. Suárez, "A quantitative approach to beauty. perceived attractiveness of human faces in world painting," *International Journal for Digital Art History*, no. 1, 2015.

[79] C. I. de l'Eclairage, "Recommendations on uniform color spaces, color-difference equations, psychometric color terms," *Paris: CIE*, 1978.

[80] R. de Piles, *The Principles of Painting...: To which is Added, The Balance of Painters...* J. Osborn, 1743.

[81] ——, *The art of painting, with the lives and characters of above 300 of the most eminent painters...*, 3rd ed. London: Thomas Payne, 1750.

[82] K. Deepwell, "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)*, 2019.

[83] J. Delon, A. Desolneux, J. L. Lisani, and A. B. Petro, "Automatic color palette," in *ICIP*, 2005.

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

[85] E. Denton, A. Hanna, R. Amironesei, A. Smart, and H. Nicole, "On the genealogy of machine learning datasets: A critical history of imagenet," *Big Data & Society*, vol. 8, no. 2, p. 20539517211035955, 2021.

[86] J. Dewey, *Art as experience.* Penguin, 2005.

[87] C. D'ignazio and L. F. Klein, *Data feminism.* MIT press, 2020.

[88] C. DiSalvo, "The need for design history in HCI," *Interactions*, vol. 21, no. 6, pp. 20–21, Oct. 2014.

[89] E. Dockterman, "How 'hot or not' created the internet we know today," *Time Magazine*, 2014.

[90] B. Doosti, D. J. Crandall, and N. M. Su, "A deep study into the history of web design," in *Proceedings of the 2017 ACM on Web Science Conference*, ser.

WebSci '17. New York, NY, USA: ACM, 2017, pp. 329–338. [Online]. Available: http://doi.acm.org/10.1145/3091478.3091503

[91] J. Drucker, *Graphesis: Visual forms of knowledge production.* Harvard University Press Cambridge, MA, 2014, vol. 6.

[92] J. Drucker and C. Bishop, "A conversation on digital art history," *Debates in the digital humanities 2019*, pp. 321–334, 2019.

[93] A. Dunne and F. Raby, *Speculative everything: design, fiction, and social dreaming*, 2013.

[94] C. d'Ignazio and L. F. Klein, "Feminist data visualization." Workshop on Visualization for the Digital Humanities (VIS4DH), Baltimore. IEEE., 2016.

[95] H. R. Ekbia, *Artificial dreams: the quest for non-biological intelligence.* Cambridge University Press Cambridge, 2008, vol. 200, no. 8.

[96] A. Elgammal and B. Saleh, "Quantifying creativity in art networks," in *Proceedings of the Sixth International Conference on Computational Creativity June*, 2015, p. 39.

[97] D. Ellis, "All websites look the same," Blog post on novolume.co.uk, February 2015, retrieved from http://www.novolume.co.uk/blog/all-websites-look-the-same/.

[98] S. Elo and H. Kyngäs, "The qualitative content analysis process," *Journal of advanced nursing*, vol. 62, no. 1, pp. 107–115, 2008.

[99] K. J. Emery and M. A. Webster, "Individual differences and their implications for color perception," *Current Opinion in Behavioral Sciences*, vol. 30, pp. 28–33, 2019.

[100] I. Engholm, "Digital style history: The development of graphic design on the Internet," vol. 13, no. 4, pp. 193–211, 2002.

[101] ——, "Design history of the www: Website development from the perspective of genre and style theory," *Artifact*, vol. 1, no. 4, pp. 217–231, Jan. 2007. [Online]. Available: http://dx.doi.org/10.1080/17493460802127757

[102] ——, "Research-based online presentation of web design history: The case of webmuseum.dk," in *Web History*, N. Brügger, Ed. Peter Lang Inc., 2010.

[103] K. Fallan, *Design History: Understanding Theory and Method.* Berg, 2010.

[104] H. Fang, C. Cui, X. Deng, X. Nie, M. Jian, and Y. Yin, "Image aesthetic distribution prediction with fully convolutional network," in *ICMM.* Springer, 2018, pp. 267–278.

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

[106] C. Fiesler, S. Morrison, and A. S. Bruckman, "An archive of their own: A case study of feminist hci and values in design," in *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2016, pp. 2574–2585.

[107] V. Fonti and E. Belitser, "Feature selection using lasso," *VU Amsterdam research paper in business analytics*, vol. 30, pp. 1–25, 2017.

[108] J. Forlizzi, J. Zimmerman, and E. Stolterman, "From design research to theory: Evidence of a maturing field," *Proceedings of IASDR*, vol. 9, pp. 2889–2898, 2009.

[109] C. W. Forstall and W. J. Scheirer, "Statistical learning as a model for intertextuality," *Quantitative Intertextuality: Analyzing the Markers of Information Reuse*, pp. 23–52, 2019.

[110] M. Foundation, "Internet health report v. 1.0 2018," 2019. [Online]. Available: https://internethealthreport.org/2019

[111] B. Friedman, P. H. Kahn, and A. Borning, "Value Sensitive Design and Information Systems," in *Human-Computer Interaction and Management Information Systems: Foundations Advances in Management Information Systems*, ser. Advances in Management Information Systems. M.E. Sharpe, 2006, vol. 5, pp. 348–372. [Online]. Available: https://link.springer.com/chapter/10.1007/978-94-007-7844-3_4

[112] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren, "Value sensitive design and information systems," *Early engagement and new technologies: Opening up the laboratory*, pp. 55–95, 2013.

[113] P. Galanter, "Computational aesthetic evaluation: past and future," *Computers and creativity*, pp. 255–293, 2012.

[114] A. Gambino, J. Fox, and R. A. Ratan, "Building a stronger casa: Extending the computers are social actors paradigm," *Human-Machine Communication*, vol. 1, pp. 71–85, 2020.

[115] M. Gardezi, K. H. Fung, U. M. Baig, M. Ismail, O. Kadosh, Y. S. Bonneh, and B. R. Sheth, "What makes an image interesting and how can we explain it," *Frontiers in Psychology*, vol. 12, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.668651

[116] R. W. Gehl, L. Moyer-Horner, and S. K. Yeo, "Training computers to see internet pornography: Gender and sexual discrimination in computer vision science," *Television & New Media*, vol. 18, no. 6, pp. 529–547, 2017.

[117] H. A. Geller, R. Bartho, K. Thömmes, and C. Redies, "Statistical image properties predict aesthetic ratings in abstract paintings created by neural style transfer," *Frontiers in Neuroscience*, vol. 16, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnins.2022.999720

[118] V. Ginsburgh and S. Weyers, "De piles, drawing and color. an essay in quantitative art history," in *Artibus et Historiae*, vol. 23, no. 45, 2002, pp. 191–203.

[119] B. G. Glaser and A. L. Strauss, *Discovery of grounded theory: Strategies for qualitative research.* Routledge, 2017.

[120] C. Gleason, A. Pavel, H. Gururaj, K. M. Kitani, and J. P. Bigham, "Making GIFs Accessible," in *The SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS '20. New York, NY, USA: ACM, 2020.

[121] M. E. Glickman and A. C. Jones, "Rating the chess rating system," *Chance*, vol. 12, pp. 21–28, 1999.

[122] E. H. Gombrich, *Norm and form.* Phaidon, 1985.

[123] S. Goree, “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*, 2021.

[124] S. Goree, G. Appleby, D. Crandall, and N. Su, “Attention is all they need: Exploring the media archaeology of the computer vision research paper,” *arXiv preprint arXiv:2209.11200*, 2022.

[125] K. Graddy, “Taste endures! the rankings of roger de piles († 1709) and three centuries of art prices,” *The Journal of Economic History*, pp. 766–791, 2013.

[126] D. Gray, K. Yu, W. Xu, and Y. Gong, “Predicting facial beauty without landmarks,” in *European Conference on Computer Vision*, pp. 434–447.

[127] E. Gray, “Understanding auto iso in photography,” https://photographylife.com/what-is-auto-iso, retrieved May 18th 2023, 2018.

[128] C. D. Green, “All that glitters: A review of psychological research on the aesthetics of the golden section,” *Perception*, vol. 24, no. 8, pp. 937–968, 1995.

[129] D. Greene, A. L. Hoffmann, and L. Stark, “Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning,” 2019.

[130] G. Greenfield, “On the origins of the term ”computational aesthetics”,” in *Proceedings of the First Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, ser. Computational Aesthetics'05. Goslar, DEU: Eurographics Association, 2005, p. 9–12.

[131] X. Gu, Y. Wong, P. Peng, L. Shou, G. Chen, and M. S. Kankanhalli, “Understanding fashion trends from street photos via neighbor-constrained embedding learning,” in *ACM Multimedia*, 2017, pp. 190–198.

[132] J. Guild, “The colorimetric properties of the spectrum,” *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 230, no. 681-693, pp. 149–187, 1931.

[133] P. Guyer, “18th century german aesthetics,” *Stanford Encyclopedia of Philosophy*, 2007.

[134] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. V. Gool, “The interestingness of images,” in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 1633–1640.

[135] J. Ha, R. M. Haralick, and I. T. Phillips, “Recursive x-y cut using bounding boxes of connected components,” *Proceedings of 3rd International Conference on Document Analysis and Recognition*, vol. 2, pp. 952–955 vol.2, 1995.

[136] S. A. Hale and V. D. Alexander, "Live versus archive: Comparing a web archive and to a population of webpages," in *The Web as History: Using Web Archives to Understand the Past and the Present.*, 2017, pp. 45–61, oCLC: 1057656067.

[137] D. Hall, *Subjectivity.* Routledge, 2004.

[138] D. Haraway, "Situated knowledges: The science question in feminism and the privilege of partial perspective," in *Feminist theory reader.* Routledge, 1988, pp. 303–310.

[139] ——, *Simians, cyborgs, and women: The reinvention of nature.* Routledge, 2013.

[140] D. J. Haraway, "Crystals, fabrics, and fields: Metaphors that shape embryos. berkeley," 1976.

[141] J. S. Hare, P. H. Lewis, P. G. Enser, and C. J. Sandom, "Mind the gap: Another look at the problem of the semantic gap in image retrieval," in *Multimedia Content Analysis, Management, and Retrieval 2006*, vol. 6073. International Society for Optics and Photonics, 2006, p. 607309.

[142] M. Harman, "Opencamera," https://opencamera.org.uk/.

[143] S. Hartmann, J. Sprenger *et al.*, "Bayesian epistemology," in *Routledge companion to epistemology.* Routledge, 2010, pp. 609–620.

[144] D. Hasler and S. E. Suesstrunk, "Measuring colorfulness in natural images," in *Human vision and electronic imaging VIII*, vol. 5007. SPIE, 2003, pp. 87–95.

[145] W. Hauser, B. Neveu, J.-B. Jourdain, C. Viard, and F. Guichard, "Image quality benchmark of computational bokeh," *Electronic Imaging*, vol. 2018, no. 12, pp. 340–1, 2018.

[146] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, "Accessing the deep web: A survey," *Communications of the ACM*, vol. 50, no. 5, pp. 94–101, 2007.

[147] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[148] ——, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[149] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata, "Grounding visual explanations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 264–279.

[150] S. R. Herring, C.-C. Chang, J. Krantzler, and B. P. Bailey, "Getting inspired!: Understanding how and why examples are used in creative design practice," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 87–96.

[151] A. Hertzmann, "The choices hidden in photography," *Journal of Vision*, vol. 22, no. 11, pp. 10–10, 2022.

[152] ——, "The curse of performative user studies," Retrieved from https://aaronhertzmann.com/2023/04/27/user-studies.html on 5-24-23, 2023.

[153] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[154] T. Hinderson, "Zhang-shasha: Tree edit distance in python," 2014. [Online]. Available: https://github.com/timtadh/zhang-shasha

[155] P. D. Hoff, *A first course in Bayesian statistical methods.* Springer, 2009, vol. 580.

[156] D. Hume, "Of the standard of taste," in *Art and its Significance: An Anthology of Aesthetic Theory*, S. D. Ross, Ed. SUNY Press, 1994.

[157] A. Hussain and E. O. Mkpojiogu, "The effect of responsive web design on the user experience with laptop and smartphone devices," *Jurnal Teknologi*, vol. 77, no. 4, 2015.

[158] M. Inglis and A. Aberdein, "Beauty is not simplicity: An analysis of mathematicians' proof appraisals," *Philosophia Mathematica*, vol. 23, no. 1, pp. 87–109, 2015.

[159] M. Irfan and D. G. Stork, "Multiple visual features for the computer authentication of jackson pollock's drip paintings: beyond box counting and fractals," in *Image Processing: Machine Vision Applications II*, vol. 7251. SPIE, 2009, pp. 236–246.

[160] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[161] J. Itten, *The elements of color.* John Wiley & Sons, 1970.

[162] M. Y. Ivory, “An empirical foundation for automated web interface evaluation,” Ph.D. dissertation, UC Berkeley, 2001.

[163] A. Jahanian, *Quantifying aesthetics of visual design applied to automatic design.* Springer Theses, 2016.

[164] A. Jahanian, P. Isola, and D. Wei, “Mining visual evolution in 21 years of web design,” in *CHI EA*, 2017.

[165] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A survey on contrastive self-supervised learning,” *Technologies*, vol. 9, no. 1, p. 2, 2020.

[166] S. Jeong, S. Roh, and K. Sohn, “Multi-regime analysis for computer vision-based traffic surveillance using a change-point detection algorithm,” *IEEE Access*, vol. 9, pp. 40 980–40 995, 2021.

[167] W. Jiang, A. C. Loui, and C. D. Cerosaletti, “Automatic aesthetic value assessment in photographic images,” in *2010 IEEE International Conference on Multimedia and Expo.* IEEE, 2010, pp. 920–925.

[168] X. Jin, L. Wu, X. Li, S. Chen, S. Peng, J. Chi, S. Ge, C. Song, and G. Zhao, "Predicting aesthetic score distribution through cumulative jensen-shannon divergence," in *AAAI*, vol. 32, no. 1, 2018.

[169] S. F. Johnston, "The construction of colorimetry by committee," *Science in context*, vol. 9, no. 4, pp. 387–420, 1996.

[170] J. Joho, "Hotornot shaped the social web as we know it," *Mashable*, 2020.

[171] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.

[172] M. Kairanbay, J. See, and L.-K. Wong, "Towards demographic-based photographic aesthetics prediction for portraitures," in *International Conference on Multimedia Modeling*. Springer, 2018, pp. 531–543.

[173] ——, "Beauty is in the eye of the beholder: Demographically oriented analysis of aesthetics in photographs," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 2s, pp. 1–21, 2019.

[174] V. Kamble and K. Bhurchandi, "No-reference image quality assessment algorithms: A survey," *Optik*, vol. 126, no. 11-12, pp. 1090–1097, 2015.

[175] K. Kanclerz, M. Gruza, K. Karanowski, J. Bielaniewicz, P. Miłkowski, J. Kocoń, and P. Kazienko, "What if ground truth is subjective? personalized deep neural hate

speech detection," in *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 2022, pp. 37–45.

[176] I. Kant, *Critique of Judgment.* Hackett Press, 1987.

[177] ——, *Observations on the Feeling of the Beautiful and Sublime.* Univ of California Press, 2003.

[178] Y. Kao, R. He, and K. Huang, "Deep aesthetic quality assessment with semantic information," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1482–1495, 2017.

[179] Y. Kao, C. Wang, and K. Huang, "Visual aesthetic quality assessment with a regression model," in *ICIP.* IEEE, 2015, pp. 1583–1587.

[180] S. Kapania, O. Siy, G. Clapper, A. M. SP, and N. Sambasivan, "" because ai is 100% right and safe": User attitudes and sources of ai authority in india," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–18.

[181] A. Karjus, M. C. Solà, T. Ohm, S. E. Ahnert, and M. Schich, "Compression ensembles quantify aesthetic complexity and the evolution of visual art," *arXiv preprint arXiv:2205.10271*, 2022.

[182] B. Karmann, "Paragraphica – context to image (ai) camera," Retrieved from https://www.creativeapplications.net/objects/paragraphica-context-to-image-ai-camera/ on 6-8-23, 2023.

[183] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the american statistical association*, vol. 90, no. 430, pp. 773–795, 1995.

[184] T. D. Kaufmann, "Periodization and its discontents," *Journal of Art Historiography*, vol. 2, no. 2, 2010.

[185] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *CVPR*, vol. 1. IEEE, 2006, pp. 419–426.

[186] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[187] K. Ko, J.-T. Lee, and C.-S. Kim, "Pac-net: pairwise aesthetic comparison network for image aesthetic assessment," in *ICIP*. IEEE, 2018, pp. 2491–2495.

[188] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes, "Photo aesthetics ranking network with attributes and content adaptation," in *ECCV*. Springer, 2016, pp. 662–679.

[189] S. Kong, Y. Shen, and L. Huang, "Resolving training biases via influence-based data relabeling," in *International Conference on Learning Representations*, 2021.

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

[191] P. Kovář and O. Letocha, "Web design museum," 2019. [Online]. Available: https://www.webdesignmuseum.org/

[192] K. Krippendorff, "Computing krippendorff's alpha-reliability," 2011.

[193] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012.

[194] ——, “Imagenet classification with deep convolutional neural networks,” *NIPS*, vol. 25, pp. 1097–1105, 2012.

[195] V. Krotov and L. Silva, “Legality and ethics of web scraping,” 2018.

[196] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, “Too much, too little, or just right? ways explanations impact end users’ mental models,” in *2013 IEEE Symposium on visual languages and human centric computing.* IEEE, 2013, pp. 3–10.

[197] R. Kumar, A. Satyanarayan, C. Torres, M. Lim, S. Ahmad, S. R. Klemmer, and J. O. Talton, “Webzeitgeist: Design mining the web,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013.

[198] C. Ladha, N. Hammerla, E. Hughes, P. Olivier, and T. Ploetz, “Dog’s life: wearable activity recognition for dogs,” in *UBICOMP*, 2013, pp. 415–418.

[199] S. Langer, *Feeling and Form.* Scribner, 1953.

[200] J. Lazar, J. H. Feng, and H. Hochheiser, *Research methods in human-computer interaction.* Morgan Kaufmann, 2017.

[201] B. Lee, S. Srivastava, R. Kumar, R. Brafman, and S. R. Klemmer, “Designing with interactive example galleries,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2010, pp. 2257–2266.

[202] J.-T. Lee and C.-S. Kim, “Image aesthetic assessment based on pairwise comparison a unified approach to score regression, binary classification, and personalization,” in *ICCV*, 2019, pp. 1191–1200.

[203] L. Lessig, *Remix: Making Art and Commerce Thrive in the Hybrid Economy.* Penguin, 2008.

[204] V. I. Levenshtein *et al.*, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[205] C. Li, A. C. Loui, and T. Chen, “Towards aesthetics: A photo quality assessment and photo selection system,” in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 827–830.

[206] C. C. Liem, M. Langer, A. Demetriou, A. M. Hiemstra, A. Sukma Wicaksana, M. P. Born, and C. J. König, “Psychology meets machine learning: Interdisciplinary perspectives on algorithmic job candidate screening,” *Explainable and interpretable models in computer vision and machine learning*, pp. 197–253, 2018.

[207] S. Lin and P. Hanrahan, “Modeling how people extract color themes from images,” in *CHI*, 2013, pp. 3101–3110.

[208] B. Lisefski, “Data-driven design is killing our instincts,” Article on medium.com, August 2019, retrieved from https://modus.medium.com/data-driven-design-is-killing-our-instincts-d448d141653d.

[209] J. Liu, D. Liu, W. Yang, S. Xia, X. Zhang, and Y. Dai, “A comprehensive benchmark for single image compression artifact reduction,” *IEEE Transactions on image processing*, vol. 29, pp. 7845–7860, 2020.

[210] S. Liu, Y. Wei, J. Lu, and J. Zhou, “An improved evaluation framework for generative adversarial networks,” *arXiv preprint arXiv:1803.07474*, 2018.

[211] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, “Fashion parsing with weak color-category labels,” *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 253–265, 2013.

[212] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *CVPR*, 2015, pp. 3431–3440.

[213] L. Lu, “Why do all modern websites look the same?” Blog post on designroast.org, November 2015, retrieved from https://designroast.org/why-do-all-modern-websites-look-the-same/.

[214] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, “Rapid: Rating pictorial aesthetics using deep learning,” in *ACM Multimedia*, 2014, pp. 457–466.

[215] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, “Deep multi-patch aggregation network for image style, aesthetics, and quality estimation,” in *ICCV*, 2015, pp. 990–998.

[216] C. R. Luckett, S. L. Burns, and L. Jenkinson, “Estimates of relative acceptability from paired preference tests,” *Journal of Sensory Studies*, vol. 35, no. 5, p. e12593, 2020.

[217] G. Lupyan, R. A. Rahman, L. Boroditsky, and A. Clark, “Effects of language on visual perception,” *Trends in cognitive sciences*, vol. 24, no. 11, pp. 930–944, 2020.

[218] H. Lv and X. Tian, “Learning relative aesthetic quality with a pairwise approach,” in *ICMM*. Springer, 2016, pp. 493–504.

[219] S. Ma, J. Liu, and C. Wen Chen, “A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment,” in *CVPR*, 2017, pp. 4535–4544.

[220] P. Machado, J. Romero, M. Nadal, A. Santos, J. Correia, and A. Carballal, “Computerized measures of visual complexity,” *Acta psychologica*, vol. 160, pp. 43–57, 2015.

[221] A.-S. Maerten and D. Soydaner, “From paintbrush to pixel: A review of deep neural networks in ai-generated art,” *arXiv preprint arXiv:2302.10913*, 2023.

[222] L. Mai, H. Jin, and F. Liu, “Composition-preserving deep photo aesthetics assessment,” in *CVPR*, 2016, pp. 497–506.

[223] L. Manovich, "The science of culture? social computing, digital humanities and cultural analytics," *Journal of Cultural Analytics*, vol. 1, no. 1, 2016.

[224] ——, *Cultural analytics.* Mit Press, 2020.

[225] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," in *Computer graphics forum*, vol. 31, no. 8. Wiley Online Library, 2012, pp. 2478–2491.

[226] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka, "Assessing the aesthetic quality of photographs using generic image descriptors," in *ICCV.* IEEE, 2011, pp. 1784–1791.

[227] S. Marinai, E. Marino, and G. Soda, "Layout based document image retrieval by means of xy tree reduction," in *Eighth International Conference on Document Analysis and Recognition (ICDAR'05)*, Aug 2005, pp. 432–436 Vol. 1.

[228] G. Markowsky, "Misconceptions about the golden ratio," *The College Mathematics Journal*, vol. 23, no. 1, pp. 2–19, 1992.

[229] R. Martin-Martin, M. Patel, H. Rezatofighi, A. Shenoi, J. Gwak, E. Frankel, A. Sadeghian, and S. Savarese, "Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments," *PAMI*, 2021.

[230] Y. Matsuda, "Color design," *Asakura Shoten*, vol. 2, no. 4, p. 10, 1995.

[231] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.

[232] E. Michailidou, S. Harper, and S. Bechhofer, “Visual complexity and aesthetic perception of web pages,” 09 2008, pp. 215–224.

[233] D. Miller, “Why do all websites look the same?” Blog post on awareweb.com, April 2019, retrieved from https://www.awareweb.com/blog/why-do-all-websites-look-the-same.

[234] K. Millis, “Making meaning brings pleasure: the influence of titles on aesthetic experiences.” *Emotion*, vol. 1, no. 3, p. 320, 2001.

[235] D. Monsef, “Colourlovers,” 2005, https://www.colourlovers.com/.

[236] P. Moon and D. Spencer, “Geometric formulation of classical color harmony,” *JOSA*, vol. 34, no. 1, pp. 46–59, 1944.

[237] K. Moran, “Brutalism and antidesign,” Article on nngroup.com, November 2017, retrieved from https://www.nngroup.com/articles/brutalism-antidesign/.

[238] M. Moravec, “Network analysis and feminist artists,” *Artl@ s Bulletin*, vol. 6, no. 3, p. 5, 2017.

[239] B. Müller, “Why do all websites look the same,” Article on Medium.com, September 2018, retrieved from https://medium.com/s/story/on-the-visual-weariness-of-the-web-8af1c969ce73.

[240] L. Mulvey, “Visual pleasure and narrative cinema,” in *Visual and other pleasures.* Springer, 1989, pp. 14–26.

[241] N. Murray, L. Marchesotti, and F. Perronnin, “Ava: A large-scale database for aesthetic visual analysis,” in *CVPR.* IEEE, 2012, pp. 2408–2415.

[242] S. Nadis and S.-T. Yau, *A history in sum: 150 years of mathematics at Harvard (1825-1975).* Harvard University Press, 2013.

[243] C. Nass, J. Steuer, and E. R. Tauber, “Computers are social actors,” in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1994, pp. 72–78.

[244] M. W. Newman and J. A. Landay, “Sitemaps, Storyboards, and Specifications: A Sketch of Web Site Design Practice,” in *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*, ser. DIS ’00. ACM, 2000, pp. 263–274.

[245] A. Ngo, A. Candri, T. Ferdinan, J. Kocoń, and W. Korczynski, “Studemo: A non-aggregated review dataset for personalized emotion recognition,” in *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 2022, pp. 46–55.

[246] D. Nickerson, “History of the munsell color system and its scientific application,” *JOSA*, vol. 30, no. 12, pp. 575–586, 1940.

[247] J. Nielsen, “Estimating the number of subjects needed for a thinking aloud test,” *International journal of human-computer studies*, vol. 41, no. 3, pp. 385–397, 1994.

[248] ——, “Why you only need to test with 5 users,” 2000, https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/ retrieved May 15th 2023.

[249] W. Odom, M. Selby, A. Sellen, D. Kirk, R. Banks, and T. Regan, “Photobox: on the design of a slow technology,” in *Proceedings of the designing interactive systems conference*, 2012, pp. 665–668.

[250] P. O’Donovan, A. Agarwala, and A. Hertzmann, “Color compatibility from large datasets,” in *ACM SIGGRAPH*, 2011, pp. 1–12.

[251] I. A. of Digital Arts and Sciences, “Webby awards,” 2019. [Online]. Available: http://www.webbyawards.com

[252] M. O’Mahony and S. Wichchukit, “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*, vol. 66, pp. 146–152, 2017.

[253] G. O’Malley, “Literary synesthesia,” *The journal of aesthetics and art criticism*, vol. 15, no. 4, pp. 391–411, 1957.

[254] M. Park, J. Park, Y. M. Baek, and M. Macy, “Cultural values and cross-cultural video consumption on youtube,” *PloS one*, vol. 12, no. 5, 2017.

[255] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, “Data and its (dis) contents: A survey of dataset development and use in machine learning research,” *Patterns*, vol. 2, no. 11, p. 100336, 2021.

[256] O. Pele and M. Werman, “Fast and robust earth mover’s distances,” in *2009 IEEE 12th International Conference on Computer Vision*, Sep. 2009, pp. 460–467.

[257] J. Pierce and E. Paulos, “Making multiple uses of the obscura 1c digital camera: reflecting on the design, production, packaging and distribution of a counterfunctional device,” in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 2103–2112.

[258] W. S. Pluhar, *Translator’s Introduction to Kant’s Critique of Judgment.* Indianap. Hackett, 1987.

[259] A. Podoksik, *Pablo picasso.* Parkstone International, 2019.

[260] X. Qi and B. Davison, “Web page classification: Features and algorithms,” *ACM Comp. Surv.*, vol. 41, no. 2, pp. 12:1–12:31, Feb. 2009. [Online]. Available: http://doi.acm.org/10.1145/1459352.1459357

[261] J. Qiao, “Colormind,” 2017, http://colormind.io/.

[262] Z. Qin, F. Yu, C. Liu, and X. Chen, “How convolutional neural networks see the world — a survey of convolutional neural network visualization methods,” *Mathematical Foundations of Computing*, vol. 1, p. 149, 2018. [Online]. Available: http://aimsciences.org//article/id/324b6e02-74f8-4511-ae70-636b3cc0362f

[263] P. Regensburg, “5 reasons why most websites look the same today,” Blog post on blog.raincastle.com, March 2014, retrieved from https://blog.raincastle.com/bid/105839/5-Reasons-Why-Most-Websites-Look-the-Same-Today.

[264] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, and K. Z. Gajos, "Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2013, pp. 2049–2058.

[265] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *ICCV*, 2017, pp. 638–647.

[266] K. Roose, "An ai-generated picture won an art prize. artists aren't happy," *The New York Times*, vol. 2, p. 2022, 2022.

[267] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *IJCV*, vol. 40, no. 2, pp. 99–121, 2000.

[268] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim, "What you see is what you can change: Human-centered machine learning by interactive visualization," *Neurocomputing*, vol. 268, pp. 164–175, 2017.

[269] B. Saleh, K. Abe, R. S. Arora, and A. Elgammal, "Toward automated discovery of artistic influence," *Multimedia Tools and Applications*, vol. 75, pp. 3565–3591, 2016.

[270] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *International Journal for Digital Art History*, vol. 2, October 2016. [Online]. Available: https://journals.ub.uni-heidelberg.de/index.php/dah/article/view/23376

[271] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," *science*, vol. 311, no. 5762, pp. 854–856, 2006.

[272] M. Schapiro, H. W. Janson, and E. H. Gombrich, "Criteria of periodization in the history of european art," *New Literary History*, vol. 1, no. 2, pp. 113–125, 1970.

[273] M. K. Scheuerman, A. Hanna, and E. Denton, "Do datasets have politics? disciplinary values in computer vision dataset development," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–37, 2021.

[274] M. Schich, C. Song, Y.-Y. Ahn, A. Mirsky, M. Martino, A.-L. Barabási, and D. Helbing, "A network framework of cultural history," *science*, vol. 345, no. 6196, pp. 558–562, 2014.

[275] B. Schneider, "From clubs to affinity: The decentralization of art on the internet." [Online]. Available: https://web.archive.org/web/20120707101824/http://fourninetyone.com/2011/01/06/fromclubstoaffinity/

[276] Y. Schoen, "In defense of homogeneous design," Article on medium.com, March 2016, retrieved from https://medium.com/@yarcom/in-defense-of-homogeneous-design-b27f79f4bb87.

[277] D. Schuler and A. Namioka, *Participatory design: Principles and practices.* CRC Press, 1993.

[278] G. Schwarz, "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[279] A. Seffah, "The evolution of design patterns in hci: from pattern languages to pattern-oriented design," in *Proceedings of the 1st International Workshop on Pattern-Driven Engineering of Interactive Computing Systems.* ACM, 2010, pp. 4–9.

[280] I. Seidman, *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*, 3rd ed. Teachers College Press, 2006.

[281] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[282] P. Sengers, K. Boehner, S. David, and J. J. Kaye, "Reflective Design," in *Proceedings of the 4th Decennial Conference on Critical Computing: Between Sense and Sensibility*, ser. CC '05. ACM, 2005, pp. 49–58.

[283] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[284] K. Sheng, W. Dong, C. Ma, X. Mei, F. Huang, and B.-G. Hu, "Attention-based multi-patch aggregation for image aesthetic assessment," in *ACM Multimedia*, 2018, pp. 879–886.

[285] Y. Shi, M. Larson, and A. Hanjalic, "Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, pp. 1–45, 2014.

[286] B. Shulman, A. Sharma, and D. Cosley, "Predictability of popularity: Gaps between prediction and understanding," in *ICWSM*, 2016. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13129/12754

[287] H. Y. Sigaki, M. Perc, and H. V. Ribeiro, "History of art paintings through the lens of entropy and complexity," *Proceedings of the National Academy of Sciences*, vol. 115, no. 37, pp. E8585–E8594, 2018.

[288] C. Spaenjers, W. N. Goetzmann, and E. Mamonova, "The economics of aesthetics and record prices for art since 1701," *Explorations in Economic History*, vol. 57, pp. 79–94, 2015.

[289] B. Srinivasa Desikan, H. Shimao, and H. Miton, "Wikiartvectors: Style and color representations of artworks for cultural analysis via information theoretic measures," *Entropy*, vol. 24, no. 9, p. 1175, 2022.

[290] D. Stalling and H.-C. Hege, "Fast and resolution independent line integral convolution," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995, pp. 249–256.

[291] P. Stanicek, "Paletton," 2002, https://paletton.com/.

[292] H. Stigmar and L. Harrie, "Evaluation of analytical measures of map legibility," *The Cartographic Journal*, vol. 48, no. 1, pp. 41–53, 2011.

[293] D. Stork and M. Johnson, "Estimating the location of illuminants in realist master paintings computer image analysis addresses a debate in art history of the baroque," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 255–258.

[294] W. G. Studdert-Kennedy and M. Davenport, "The balance of roger de piles: a statistical analysis," *The Journal of Aesthetics and Art Criticism*, vol. 32, no. 4, pp. 493–502, 1974.

[295] N. M. Su, A. Lazar, J. Bardzell, and S. Bardzell, "Of dolls and men: Anticipating sexual intimacy with robots," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 26, no. 3, pp. 1–35, 2019.

[296] N. M. Su and E. Stolterman, "A Design Approach for Authenticity and Technology," in *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, ser. DIS '16. ACM, 2016, pp. 643–655.

[297] S. Sultana, F. Guimbretière, P. Sengers, and N. Dell, "Design within a patriarchal society: Opportunities and challenges in designing for rural women in bangladesh," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[298] F. Szabo, P. Bodrogi, and J. Schanda, “Experimental modeling of colour harmony,” *Color Research & Application*, vol. 35, no. 1, pp. 34–49, 2010.

[299] R. Szeliski, *Computer vision: algorithms and applications.* Springer Science & Business Media, 2010.

[300] X. Tang, W. Luo, and X. Wang, “Content-based photo quality assessment,” *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[301] M. S. Tenan, A. J. Tweedell, and C. A. Haynes, “Iterative assessment of statistically-oriented and standard algorithms for determining muscle onset with intramuscular electromyography,” *Journal of Applied Biomechanics*, vol. 33, no. 6, pp. 464–468, 2017.

[302] G. Thierry, P. Athanasopoulos, A. Wiggett, B. Dering, and J.-R. Kuipers, “Unconscious effects of language-specific terminology on preattentive color perception,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 11, pp. 4567–4570, 2009.

[303] C. Thomas and A. Kovashka, “Predicting the politics of an image using webly supervised data,” *Advances in neural information processing systems*, vol. 32, 2019.

[304] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[305] L. Torresani, M. Szummer, and A. Fitzgibbon, “Efficient object category recognition using classemes,” in *Computer Vision–ECCV 2010: 11th European Conference on*

Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11. Springer, 2010, pp. 776–789.

[306] B. Triggs, A. Zisserman, and R. Szeliski, *Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings.* Springer, 2000.

[307] S. Turkle, C. Breazeal, O. Dasté, and B. Scassellati, "Encounters with kismet and cog: Children respond to relational artifacts," *Digital media: Transformations in human communication*, vol. 120, 2006.

[308] D. Ushizima, L. Manovich, T. Margolis, and J. Douglas, "Cultural analytics of large datasets from flickr," in *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[309] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[310] D. K. Van Duyne, J. A. Landay, and J. I. Hong, *The Design of Sites: Patterns for Creating Winning Web Sites*, 2nd ed. Prentice Hall, 2007, oCLC: ocm70911156.

[311] T. Vanderbilt, *You May Also Like: Taste in an Age of Endless Choice.* Simon and Schuster, 2016.

[312] S. Vigna, "Spectral ranking," *Network Science*, vol. 4, no. 4, pp. 433–445, 2016.

[313] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[314] C. Von Bearnensquash, "Paper gestalt," *Secret Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2010.

[315] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "Hoggles: Visualizing object detection features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1–8.

[316] L. Wang, X. Wang, T. Yamasaki, and K. Aizawa, "Aspect-ratio-preserving multi-patch image aesthetics score prediction," in *CVPR Workshops*, 2019, pp. 0–0.

[317] T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez, "Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5310–5319.

[318] Z. Wang, S. Chang, F. Dolcos, D. Beck, D. Liu, and T. S. Huang, "Brain-inspired deep networks for image aesthetics assessment," *arXiv preprint arXiv:1601.04155*, 2016.

[319] A. Wasielewski, *Computational Formalism.* MIT Press, 2023.

[320] K. Winkle, D. McMillan, M. Arnelid, K. Harrison, M. Balaam, E. Johnson, and I. Leite, "Feminist human-robot interaction: Disentangling power, principles and

practice for better, more ethical hri," in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 2023, pp. 72–82.

[321] J. O. Wobbrock, A. K. Hsu, M. A. Burger, and M. J. Magee, "Isolating the effects of web page visual appearance on the perceived credibility of online news among college students," in *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, ser. HT '19. New York, NY, USA: ACM, 2019, pp. 191–200. [Online]. Available: http://doi.acm.org/10.1145/3342220.3343663

[322] D. Wood and J. Fels, *The power of maps.* Guilford Press, 1992.

[323] W. D. Wright, "A re-determination of the trichromatic coefficients of the spectral colours," *Transactions of the Optical Society*, vol. 30, no. 4, p. 141, 1929.

[324] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li, "Better aligning text-to-image models with human preference," *arXiv preprint arXiv:2303.14420*, 2023.

[325] G. Wyszecki and W. Stiles, *Color science: concepts and methods, quantitative data and formulae.* John Wiley & Sons, 1982.

[326] S. Yamaoka, L. Manovich, J. Douglass, and F. Kuester, "Cultural analytics in large-scale visualization environments," *Computer*, vol. 44, no. 12, pp. 39–48, 2011.

[327] Y.-C. Yao, "Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches," *The Annals of Statistics*, pp. 1434–1447, 1984.

[328] Y. Yao, M. Xu, C. Choi, D. J. Crandall, E. M. Atkins, and B. Dariush, "Egocentric vision-based future vehicle localization for intelligent driving assistance systems," in *ICRA*. IEEE, 2019, pp. 9711–9717.

[329] Q. You, D. Garcia-Garcia, M. Paluri, J. Luo, and J. Joo, "Cultural diffusion and trends in facebook photographs," in *ICWSM*, 2017. [Online]. Available: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15704/14800

[330] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[331] A. Zapf, S. Castell, L. Morawietz, and A. Karch, "Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate?" *BMC medical research methodology*, vol. 16, pp. 1–10, 2016.

[332] K. Zhang and D. Shasha, "Simple fast algorithms for the editing distance between trees and related problems," *SIAM J. Comput.*, vol. 18, pp. 1245–1262, 12 1989.

[333] L. Zhang and W. Hong, "Top shot on pixel 3," Retrieved from https://ai.googleblog.com/2018/12/top-shot-on-pixel-3.html on 6-8-23, 2018.

[334] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[335] Y. Zhou, X. Lu, J. Zhang, and J. Z. Wang, "Joint image and text representation for aesthetics analysis," in *ACM Multimedia*, 2016, pp. 262–266.

[336] H. Zhu, Y. Zhou, L. Li, Y. Li, and Y. Guo, “Learning personalized image aesthetics from subjective and objective attributes,” *IEEE Transactions on Multimedia*, 2021.

CURRICULUM VITA

Samuel Philip Goree

sgoree@iu.edu · samgoree.github.io

Education

2018 – **Indiana University** – Blomington, IN

2023 PhD in Informatics, Intelligent and Interactive Systems track.

Advisor: David Crandall

PhD Minor in Digital Arts and Humanities.

2013 – **Oberlin College** – Oberlin, OH

2017 BA in Computer Science and Musical Studies

Graduated with High Honors in Computer Science *GPA: 3.8*.

Honors and Scholarships

2020 – **NSF Graduate Research Fellowship**

2023 The NSF GRFP recognizes and supports outstanding graduate students in NSF-supported STEM disciplines who are pursuing research-based degrees at accredited US institutions.

Journal Articles

2022 **“It Was Really All About Books:” Speech-like Techno-Masculinity in the Rhetoric of Dot-Com Era Web Design Books**

Goree, S., Crandall, D., Su, N.

ACM Transactions on Computer-Human Interaction (TOCHI), Vol. 30, no. 2.

Peer-Revewed Conference Publications

2023 **Correct for Whom? Subjectivity and the Evaluation of Personalized Image Aesthetics Assessment Models**

Goree, S., Khoo, W., Crandall, D.

AAAI Conference on Artificial Intelligence 2023.

2022 **HyperNP: Interactive Visual Exploration of Multidimensional Projection Hyperparameters**

Appleby, G., Espadoto, M., Chen, R.,**Goree, S.**, Telea, A., Anderson, E., Chang, R.

Eurographics Conference on Visualization (EuroViz) 2022.

2021 **What Does it Take to Cross the Aesthetic Gap? The Development of Image Aesthetic Quality Assessment in Computer Vision.**

Goree, S.

International Conference on Computational Creativity (ICCC) 2021.

2021 **Investigating the Homogenization of Web Design: A Mixed-Methods Approach.**

Goree, S., Doosti, B., Crandall, D., Su, N.

ACM CHI Conference on Human Factors in Computing Systems (CHI) 2021.

2020 **Studying Empirical Color Harmony in Design**

Goree, S., Crandall, D.

Third Workshop on Computer Vision for Fashion, Art and Design at CVPR 2020.

2018 **Pain Town, an Agent-Based Model of Opioid Use Trajectories in a Small Community.**

Bobashev, G., **Goree, S.**, Frank, J., Zule, W.

Social, Cultural, and Behavioral Modeling (eds Thomson, R. et al.) 2018

Other Publications

2022 **Attention is All They Need: Exploring the Media Archaeology of the Computer Vision Research Paper**

Goree, S., Appleby, G., Crandall, D., Su, N. *arXiv Preprint*

2021 **The Limits of Colorization of Historical Images by AI**

Goree, S. *Hyperallergic.com*

2020 **Yes, Websites Really are Starting to Look More Similar**

Goree, S., Doosti, B., Crandall, D., Su, N. *The Conversation*

Work Experience

2021 **Summer Research Assistant**

NYU Abu Dhabi. Abu Dhabi, UAE.

Worked with sociologist Dr. Kangsan Lee, to study applications of computer vision in the analysis of contemporary art markets.

2017 – **Data Scientist**

2018 RTI International. Durham, NC.

Applied expertise in machine learning, advanced analytics, statistical modeling and web development to a variety of social science-motivated projects at a large research nonprofit.

2016 **Undergraduate Researcher**

Mentor: Robert Keller (Harvey Mudd College Summer REU).

Developed deep learning backend for jazz improvisation software ImproVisor with recurrent generative adversarial networks.

2015 **Undergraduate Researcher**

Mentor: Larry Medsker (Siena College Summer REU).

Implemented information extraction and text clustering techniques to analyze a corpus of letters to the New York State EPA.

Teaching Experience

Spring 2022 **Instructor of Record, INFO I-399: Python for Data Analysis**

Designed and taught a pre-professional class for Informatics undergraduate students on the Python data analysis and data science ecosystem, including an introduction to visualization and machine learning.

2018 – 2020 **Associate Instructor, INFO I-210: Information Infrastructure 1**

Teaching assistant, lab instructor and grader for a first course in programming for Informatics undergraduate students at Indiana University.

2015 – 2017 **TA/Grader**

Served as a TA and Grader for a variety of undergraduate courses at Oberlin College including CSCI 275: Programming Abstractions, CSCI 280: Algorithms, CSCI 383: Theory of Computer Science

Presentations

Jun. 2022 Buying a Work of Art or an Artist? Impossibility and Possibility of Predicting Price of Artwork.

NYU Art + Data Conference 2022

Sep. 2021 Confronting Subjectivity in Computer Vision for Art and Design History.

ICCC Doctoral Consortium

May 2021 Investigating the Homogenization of Web Design: A Mixed-Methods Approach.

Visualization Lab, Tufts University

Jun. 2020 Why Do All Websites Look the Same Now?
Emperor Design All Hands Meeting

Jun. 2019 Musical Interfaces, Metaphors and Online MIDI Keyboards.
Midwest Music and Audio Day

Theses

2017 Towards a Relative-Pitch Neural Network System for Chorale Composition and Harmonization.
Computer Science Honors Thesis, Oberlin College.

2017 Structure and Randomness in Iannis Xenakis' *Analogique A*.
Musical Studies Capstone Thesis, Oberlin College

Service

2018-2023 **Member** Graduate Informatics Student Association, Indiana Graduate Workers' Coalition

Peer Review ISMIR, ICLR, CHI, DIS, Leonardo.