Skin Deep: Investigating Subjectivity in Skin Tone Annotations for Computer Vision Benchmark Datasets

Teanna Barrett
Howard University
Washington, DC, USA
teanna.barrett@bison.howard.edu

Quan Ze Chen
University of Washington
Seattle, WA, USA
cqz@cs.washington.edu

Amy X. Zhang
University of Washington
Seattle, WA, USA
axz@cs.uw.edu

ABSTRACT

To investigate the well-observed racial disparities in computer vision systems that analyze images of humans, researchers have turned to skin tone as a more objective annotation than race metadata for fairness performance evaluations. However, the current state of skin tone annotation procedures is highly varied. For instance, researchers use a range of untested scales and skin tone categories, have unclear annotation procedures, and provide inadequate analyses of uncertainty. In addition, little attention is paid to the positionality of the humans involved in the annotation process—both designers and annotators alike—and the historical and sociological context of skin tone in the United States. Our work is the first to investigate the skin tone annotation process as a sociotechnical project. We surveyed recent skin tone annotation procedures and conducted annotation experiments to examine how subjective understandings of skin tone are embedded in skin tone annotation procedures. Our systematic literature review revealed the uninterrogated association between skin tone and race and the limited effort to analyze annotator uncertainty in current procedures for skin tone annotation in computer vision evaluation. Our experiments demonstrated that design decisions in the annotation procedure such as the order in which the skin tone scale is presented or additional context in the image (i.e., presence of a face) significantly affected the resulting inter-annotator agreement and individual uncertainty of skin tone annotations. We call for greater reflexivity in the design, analysis, and documentation of procedures for evaluation using skin tone.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies** → *Computer vision*;

KEYWORDS

skin tone annotation, model evaluation, fairness benchmark datasets, facial recognition, computer vision

ACM Reference Format:
Teanna Barrett, Quan Ze Chen, and Amy X. Zhang. 2023. Skin Deep: Investigating Subjectivity in Skin Tone Annotations for Computer Vision Benchmark Datasets. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, Article 111, 15 pages. https://doi.org/10.1145/3593013.3594114

FAccT '23, June 12–15, 2023, Chicago, IL, USA

ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594114

1 INTRODUCTION

Computer vision (CV) technologies for analyzing and processing visual data about humans, such as facial recognition, body detection [85], and skin condition classification [43], are being rapidly deployed in both everyday settings such as phone unlocking [35], as well as high stakes situations such as criminal identification [59] or skin disease detection [53]. Implementations of CV in the wild have exposed major discrepancies in performance, particularly the misclassification and misidentification of Black people in the United States [73]. While disparities in performance should already concern technologists, the material impacts of these inaccuracies along with the mistreatment of Black people by institutions such as police and healthcare systems compound the urgency of the problem.

One approach to address these issues is to test models on data disaggregated by race. The exemplary work in evaluating performance bias in face recognition [12] criticized this approach, pointing out that racial categories, which are highly contextualized by social attributes and vary widely in definition across societies around the world, are not a good fit for CV tasks. This is because these models only use general features of the face or perform skin detection to locate humans in an image. In other words, computer vision models only care about data that is skin deep. The social determinants that define a large part of race are not recorded in CV datasets.

To address the mismatch between deeply contextual race annotations and systems that only consider the visible aspects of race, some CV researchers have recently sought to use skin tone annotations instead. However, this newly adopted and relatively untested approach has also received criticism due to limitations of existing annotation scales [12] and measures [46]. In particular, the discrete categories of some of these scales do not represent the full range of human skin tones. In addition, skin tone annotation procedures vary widely from project to project. There has been little attention placed on annotator uncertainty and disagreement and how this may vary according to different annotation procedure designs [32]. Despite researcher consensus that skin tone is at least a more objective annotation than race, skin tone still carries subjective social meaning beyond color values, which has yet to be explored.

In this work, we investigate skin tone annotation processes for computer vision benchmark datasets as a sociotechnical project that embeds subjective understandings of skin tone. We first survey state-of-the-art usage of skin tone annotations in computer vision through a systematic literature review of CV papers utilizing skin tone annotations in the last five years (2017–2022). From a

qualitative analysis of 50 papers from notable venues including IEEE Computer Vision and Pattern Recognition (CVPR) and Neural Information Processing Systems (NeurIPS), we find a wide variety of procedures and use of skin tone categories and an overall lack of documentation around annotator characteristics and procedures. We also identify areas for improvement in transparency around skin tone distribution and analysis of annotator uncertainty. Finally, we interrogate the conflation between skin tone and race found in a significant number of papers.

Following our literature review, we conducted an experiment in which we collected skin tone annotations while varying certain configurations of the annotation process such as the type of scale, ordering of scale levels, and type of image annotated. We also collected the self-reported skin tone of each human annotator to investigate annotator positionality on this task and used a range-based annotation method to isolate different measures of uncertainty. After a deployment with 165 participants, we found that the order of the skin tone scale and the type of annotated image had a significant impact on the inter-rater and individual uncertainty of annotators, while we did not find a significant impact regarding scale type or annotators' self-reported skin tone. We document the results of our literature review, experiment data, and experiment code in an online repository.[1]

Our work demonstrates the importance of reflexivity when designing and implementing skin tone annotation procedures. If the humans involved in the skin tone annotation process do not acknowledge their subjective social biases, it is difficult to ensure the reliability of skin tone annotations across the field. In turn, unclear or limited reliability of skin tone annotations can affect the robustness of fairness performance evaluations [19] in CV and hinder the development of solutions for racialized disparities in CV models. We conclude with a discussion of design implications for CV researchers engaging in skin tone annotation.

2 BACKGROUND AND RELATED WORK

2.1 Impacts of Computer Vision for Analyzing Images of Humans

Computer vision (CV) is a sub-domain of machine learning (ML) in which images, videos, and other visual data are analyzed to derive specific information about the inputs. While a variety of data and insights have been the focus of CV, a large portion of the field is concerned with processing humans in visual data. Efforts in human-concerned CV tasks typically aim to surveil [4], identify (face recognition [86]), and assess physical attributes (skin condition classification [43]). As the accuracy of these systems improve, human-concerned CV technology has increasingly been deployed in public life, including in major institutions such as education, government, and healthcare.

While these technologies may perform well in experimental and testing environments, the implementation of CV in the wild has exposed major performance gaps. In particular, there is a clear pattern of face recognition misidentifications of Black users and subjects in the United States [44, 73]. Black people have been unjustly arrested as a result of being misidentified by police departments' facial recognition systems [73]. Another example is the low detection rates of skin diseases on darker skin tones [20, 27]. In addition, during the height of the pandemic, remote students took exams proctored by facial analysis technology. Students with darker complexions had to shine additional light on their faces because these proctoring platforms were not detecting them [78]. In all of these examples, the disparate low performance of the CV systems mirrors and contributes to the systemic racism in every arena of American society, including policing [81], medicine [37], and education [78]. As the harmful outcomes of CV technology towards Black people become more frequent and well-known, CV practitioners and researchers have a responsibility to investigate and address the racialized performance gaps in their systems [11].

2.2 Approaches to Investigate Disparities in Computer Vision Performance

The current approach in understanding and addressing performance gaps in machine learning is to test models on data disaggregated by race [41]. Although most technologies do not use the protected class of race as a feature [67], data points can have additional race labels to directly compare the performance of a model between different races. While this approach has successfully highlighted quantifiable disparities in tasks such as loan risk assessments [67], race annotations are not ideal for evaluating disparate performance in CV tasks. The hallmark work Gender Shades [12] was one of the first to evaluate performance bias in face recognition and clearly explain the unreliability of race annotations as a tool for evaluating disparity in CV. Race is more than phenotypical features and its categories are highly contextual [12]. Yet, human-concerned CV operates only on surface level information. For example, many face recognition models use general features of the face to execute their tasks. Furthermore, many models utilize skin detection to locate a human in images. In contrast, when it comes to race, every society defines the boundaries of the phenotypical attributes of race differently. If these annotations are not self-reported, annotators or researchers must use their personal view of racial categories to decide how image subjects are categorized [41], which threatens the reliability of the annotations.

This creates a tension in which race annotations which, as described, are a highly contextual record of perceptions and biases, are used to evaluate systems that only contend with the surface level aspects of race. This mismatch introduces cases in which a dataset with image subjects of all the same race may have widely different appearances, or a multi-racial image dataset of subjects may have similar appearances. Limited insights can be gained from a performance evaluation disaggregated by race in these two dataset instances. If they were instead annotated by skin tone, the performance of the CV systems can be evaluated based on how these models actually "see" race.

2.3 Origin, Limitations, and Sociotechnical Aspects of Skin Tone Annotations

The most popular measure for skin tone is the Fitzpatrick Skin Type (FST) which was created and proposed by Thomas Fitzpatrick [29] for analyzing sun burns. This set of six color swatches was meant to only be accompanied with written descriptions of the quality of the

[1] https://github.com/Social-Futures-Lab/skin-deep

skin and the corresponding reactions to sun exposure. Following the FST scale, an automatically calculated continuous measure called Individual Typology Angle (ITA) was developed to measure skin tone in relation to sun exposure sensitivity more precisely [21]. Even within the original scope of these skin tone measures, racial perspectives were embedded in the process. The first iteration of FST [29] and ITA [16] were explicitly meant for Caucasians. Even as these measures were updated, they still did not represent skin tones of Black or Asian populations as well [83]. The legacy of neglect for darker skin tones is evident in the continued lack of darker skin tones represented in dermatology image datasets [6, 49]. Gender Shades [12] was one of the first works to use skin tone annotations in CV to evaluate the performance of commercial facial recognition technologies. Since then, use of skin tone annotations in CV has grown, and new scales and measures exist such as the Monk Skin Tone (MST) scale, released by Google Research in 2022 [71]. MST has ten color swatches and was designed in collaboration with the sociologist Ellis Monk [63] to explicitly represent a broader range of people. However, to date, there has not been a public evaluation of the MST scale or study of its implementation.

In addition to improving the representation of skin tone scales, an understanding of the deeper sociological context of skin tone stratification is an important part of addressing the limitations of skin tone annotation. One of the primary contributors to the Monk Skin Tone scale has extensively studied how attitudes towards individuals with lighter and darker skin tones from the time of slavery in the United States persists in contemporary American society [63–65]. Racism is integral to skin tone stratification, but skin tone also has unique and pervasive social and political impacts of its own. There is a plethora of literature on the need to transparently articulate the assumptions, biases, and perspectives that are embedded in building datasets for machine learning as a whole [22, 23, 39]. Given the complex social reality of skin tone in the U.S., there is a specific need to investigate the implicit and explicit assumptions, biases, and perspectives in the skin tone annotation process as well.

2.4 Issues with Skin Tone Annotations for Computer Vision Evaluation

While skin tone is a viable annotation attribute for computer vision, its potential is limited by the current state of skin tone annotation processes. The primary problem identified by researchers is the imprecise nature of the annotation scales and measurements [12, 46]. The discrete categories in scales like FST do not sufficiently represent the possible skin tones of human subjects. Another aspect that has not been deeply explored is annotator and annotation uncertainty. There has been some documentation and analysis of the uncertainty of skin tone annotations [9, 14, 15, 32, 33, 80, 85], though our literature review finds this is rare. These works primarily provide analyses such as inter-rater reliability or some other measure of consensus or (dis)agreement. However, there is little investigation of the implicit biases or social environments that could also impact the certainty and consistency of the annotations. As many researchers have accepted skin tone as a more objective annotation than race, there has been little investigation into how skin tone annotations may also carry social meaning. Through our experiment, we provide empirical data on how different annotation processes such as differing scales and social aspects such as annotator positionality may impact annotation uncertainty.

Publication Venue	No. of Search Results	No. of Final Papers
IEEE Xplore (IWBF, BBIS, WACV, CVPR, ICB, FG, Access, +20 more)	99	37
ACM Digital Library (FAccT, CSCW)	38	4
NeurIPS	11	4
Patterns	4	1
IS&T EI	3	2
ECCV	1	1
Preprint (no search conducted)		1
Total	157	50

Table 1: Number of papers returned from a keyword search from each publication venue, along with the resulting papers in the final set of 50.

3 LITERATURE REVIEW OF SKIN TONE ANNOTATION IN COMPUTER VISION

We conducted a literature review to understand the current state of skin tone annotation procedures for computer vision datasets. As part of this literature review, we identified the common measurements and annotation procedures, annotator information, race metadata collected about the dataset subjects, and uncertainty analyses for skin tone annotations in the past five years (2017–2022). The complete annotated literature review is available in our project repository online.

3.1 Method

3.1.1 Researcher Positionality. In line with the feminist practice of reflexivity [7], we articulate how the identity, perspectives and social standing of the researcher informs our survey. In particular, the selection, pruning, and analysis of the skin tone annotation procedures are guided by the first author's understanding of skin tone and race. The literature review was conducted by the first author, who is Black, based in the United States and has a darker skin tone. As a member of the computer science academic community, the first author is also able to approach the survey from a technical understanding of dataset collection and annotation.

3.1.2 Identifying Skin Tone Annotation Procedures in Computer Vision Papers. To begin, we formulated a collection of keywords related to skin tone or skin color annotation. We collected 26 initial papers via manual searching on Google Scholar for papers that included the phrase "skin tone annotation." Papers that detailed a skin tone annotation procedure in their methodology were recorded, and additional papers that also detailed skin tone annotation procedures were found through citation trails. From perusing these papers, we chose the phrases: "skin tone", "skin type", "Fitzpatrick skin type", "darker skin", and "lighter skin". As many top computer vision publication venues are part of IEEE, we chose to conduct a search on all of IEEE's Xplore digital library. In addition to the

IEEE venues, we included the publication venues of the original 26 papers, excluding one which was a preprint. With the selection of keywords and conferences, the author completed an inclusive "OR" search of the keywords, title, and abstract in the IEEE Xplore library and the ACM Digital Library (ACM DL). We also conducted Google Scholar searches with the publication's name for venues not available through IEEE Xplore or ACM DL. Table 1 has the full breakdown of venues and number of search results from each.

3.1.3 Pruning. With a corpus of over 100 papers, the first author read over the abstract and dataset sections of each paper to remove works that did not conduct skin tone annotation procedures. Of the papers that were removed, many conducted imaging or sensing experiments rather than computer vision tasks. One paper written in Portuguese was omitted due to our own language limitations. Several makeup and fashion recommendation papers were removed because the skin tone annotations referred to the texture or condition of the skin (e.g., oily or containing acne) as opposed to complexion. Finally, we found many papers that did not annotate skin tone for their training or testing datasets. In these papers, CV tasks were paired with calls for skin tone diversity and the acknowledgement of disparate results between lighter-skin subjects and darker-skin subjects, yet ultimately the papers did not incorporate any skin tone annotations or evaluations. The prevalence of acknowledgement followed by inaction was noted as a concerning trend that disqualified a majority of the papers that were collected by the keyword and abstract search. After these stages of pruning, 50 papers remained that were analyzed for their skin tone annotation procedures and analysis.

3.1.4 Qualitative Analysis. In Appendix A, we list all the dimensions along which we annotated the papers, including their definition and an example of what was coded. Our dimensions were partly informed by a prior literature review done by Scheuerman et al. [74] on race and gender annotations in image datasets. We adapted their approach for the topic of skin tone annotations. First, we collected basic information about the skin tone annotation dataset, such as the CV task and the number of subjects in the dataset. Next, we collected information about the annotation procedure, including who or what annotated the dataset and what scale was used for the range of skin tones. We then marked whether the paper described any analysis of the skin tone annotations for uncertainty. This included any procedure or metric to measure or evaluate the (dis)agreement, consensus, or reliability of annotations. We finally marked whether there were any race annotations or metadata intended to represent the race, ethnicity, or nationality of an image subject.

3.2 Results

We broadly grouped the papers in our literature review into three categories for further comparison, as some papers reported on or made use of certain dimensions while others did not. The first category, 'Robust Skin Tone Annotation Process', refers to papers that provide a detailed account of their skin tone annotation process, annotation uncertainty analysis, and distribution of skin tones represented in their dataset. Our second category, 'Ambiguous Skin Tone Annotation Process', comprises processes that did not include one or more of the components of a robust skin tone annotation process. These two categories are mutually exclusive and together comprise the full set of 50 papers. Our third orthogonal category of 'Skin Tone and Race Annotation Process' selects papers with processes which collected or annotated both skin tone and race data. For each category, we report on the dimensions and highlight trends. A summary of findings is in Table 2.

3.2.1 Robust Skin Tone Annotation Process. There were only 9 papers out of the 50 total in our literature review with a detailed description of process, uncertainty analysis, and reporting of skin tone distribution.

Dataset: The papers in this category were mostly conducting skin tone classification or skin condition classification and created skin tone annotations to evaluate their performance ($n = 4$). Even for the tasks that were not explicitly about skin classification, the goal of creating a "diverse" training and testing dataset was a central motivation for collecting skin tone annotations. This prioritization of diversity in the dataset limited dataset size—for instance, Wilson et al. [85] noted their lack of image subjects with darker-skin tones limited the size of the dataset as they desired a balanced sample. Another important constraint was time and labor for the skin tone annotations. The largest datasets in this category of works were achieved partially due to automated annotations [9] or a large pool of annotators [14]. Excluding the two largest datasets, which have more than 200,000 images and were not collected by the authors themselves, the datasets in this category had an average size of 7,789 images.

Procedure: All of the works except one had human annotators involved in the process. Some combined approaches such as automated procedures to compare the work done by subject matter expert annotators and non-expert crowd work annotators [32]. The majority of the works also had multiple annotators involved in the annotation process. The authors recognized and reported how annotations from different annotators had differences that had to be reconciled [32]. Authors used many methods to coordinate multiple annotators. For procedures that included crowd workers, platforms such as Amazon Web Services Mechanical Turk (AWS MTurk), Scale AI, and Centaur Labs were used. Sometimes authors weighed the annotations based on annotators' experience and reliability. All of the manual procedures referenced the Fitzpatrick Skin Type (FST) scale. In two works, authors calculated Individual Topology Angle (ITA) values and then binned the skin tone values into six, FST-inspired categories from lighter skin tones to darker tones [15, 32]. In one work, even when human annotators provided FST annotations, the annotations were combined into two bins: "lighter" or "darker" [14]. Finally, three papers referenced FST but had annotators place images into lighter and darker categories.

Uncertainty Analysis: The uncertainty analysis in these works mostly operated as a tool for determining the final value, for instance, determining an annotation was reliable if it had majority consensus [85]. Another paper relied on an opaque "dynamic consensus" functionality of the chosen annotation platform to determine final ratings [33]. The newest iteration [32] included interrater reliability metrics such as Pearson correlation coefficients and qualitative confusion matrices between different annotators. Papers used statistical measures such as Krippendorff's alpha [80], Cohen's k [14], and normalized standard deviation [9] to evaluate

Category	Findings	Papers
Robust Skin Tone Annotation Process	• Smaller datasets, emphasis on skin tone diversity • Mostly manual, use of multiple human annotators per item • Use of FST scale albeit with simplifications to "lighter", "darker" categories • Reported uncertainty analyses	[9, 14, 15, 32, 33, 38, 46, 80, 85]
Ambiguous Skin Tone Annotation Process	• Face recognition or health classification tasks were common • 1/3 were automatically annotated using ITA; the rest used human annotators, oftentimes the authors themselves • Mentioned FST scale but often used custom categories instead • Typically lacked information about annotation process or distributions • Lacked uncertainty analyses	[1–3, 5, 8, 10, 12, 18, 25, 26, 30, 34, 36, 42, 43, 45, 47, 48, 51, 54–58, 60–62, 66, 68–70, 75–77, 79, 82, 84, 86–89]
Skin Tone and Race Annotation Process	• Race is used to contextualize skin tone or equate with skin tone • Lacked uncertainty analyses	[12, 18, 26, 38, 46, 56, 61, 62, 66, 76, 82, 84, 88, 89]

Table 2: Summary of findings from analyzing prior skin tone annotation procedures and datasets.

the overall (dis)agreement in the skin tone annotations. While the computed metrics varied, all serve as a step towards transparency and discussion of the limitations of skin tone annotations as an exact measurement. Evaluating (dis)agreement between annotators was the common approach for understanding annotation uncertainty.

Summary: The papers with more robust processes valued having multiple human annotators to provide more reliable annotations. They also explained their procedures for determining consensus and many went a step further to acknowledge and evaluate (dis)agreement across annotations. However, this annotation process was difficult to scale because of time, labor, and compensation constraints. Most critically, the use of FST as the annotation scale led to greater uncertainty and imbalance in distribution, leading authors to simplify the skin tone categories.

3.2.2 Ambiguous Skin Tone Annotation Process. The majority of the papers ($n = 41$) detailed their skin tone annotation process ambiguously. Many of these works did not **A**) conduct uncertainty analysis ($n = 34$), **B**) include a description of the skin tone annotation process ($n = 10$), **C**) mention the skin tone annotation scale that was used ($n = 8$), or **D**) clearly mention who were the skin tone annotators ($n = 7$).

Dataset: The most common tasks were face recognition ($n = 7$), health tasks such as heart rate tracking and skin condition classification ($n = 8$), synthetic image generation ($n = 4$), and fashion recommendations ($n = 3$). We note that face recognition researchers have taken interest in evaluation using skin tone, after the publication of major disparities in performance [12]. The ongoing interest in health ML paired with dermatological roots of skin tone annotation may also explain the prevalence of health-related tasks. Datasets varied in size from millions of images to 5 videos. The larger to mid-sized datasets utilized previously made datasets ($n = 22$), with CelebA [52] being the most frequently used dataset ($n = 5$). Newly made datasets were primarily composed of online images ($n = 4$) or collected in-person ($n = 6$).

Procedure: Some works automated their skin tone annotation processes ($n = 10$), with ITA being the most popular measure. All except one paper simplified the continuous values from ITA into ranges, with at most 8 bins. However, similar to the robust annotation papers, the majority of the works relied on human annotators. Many manual annotations were done by the creators of the dataset ($n = 7$). However, only two of these papers [36, 55] described how the authors annotated skin tone. Even then, there was little to no description of any training, the positionality of the annotators, or the annotation environment (e.g., the use of reference images or color swatches). The majority of these works referenced the FST scale. However, unlike the robust annotation papers, the author annotators developed custom skin tone categories based on common skin color descriptions (e.g., dark, brown, medium, fair, white). In a few cases, authors outsourced annotations to subject-matter-experts [5] or crowdworkers [58, 76, 84]. Finally, there were a few papers that did not record information about their skin tone annotators. In addition, the majority of the works did not provide any information about the distribution of the skin tone categories represented in their dataset. In one paper, the authors made use of FST, finding the third FST tone was the only skin tone in their dataset [77]; this bias was then left as is.

Uncertainty Analysis: Unfortunately, none of these works reported any metrics or commentary about the uncertainty, inter-rater reliability, or disagreement of the skin tone annotations.

Summary: It is a positive sign to see papers in recent years using skin tone annotations for evaluation and sometimes noting biased distributions in their datasets. However, the lack of information the about annotation processes in the majority of the papers in our literature review demonstrates that skin tone annotations are not yet seen as an integral part of dataset quality or an important attribute to evaluate performance of CV tasks. This affirms the need for developing standards for conducting and reporting on skin tone annotations.

3.2.3 Skin Tone and Race Annotation Process. We chose to more deeply examine a subset of the papers ($n = 13$) that we noticed blur the boundaries between more objective skin tone annotations and socially-specific race metadata. In some cases, the presence of

race metadata is not detrimental to the intended goal of evaluating skin tone representation in datasets. On the other hand, certain conflations of skin tone and race reinforce harmful practices of essentializing race.

Dataset: All of these works used relatively small datasets. There was a fairly even distribution of computer vision tasks such as face recognition [12, 54, 61], heart rate extraction [82, 88], and gender classification [62, 66].

Procedure and Race Annotations: We found three ways skin tone and race data were used in the annotation process. The race, ethnicity, or nationality of the image subject was used to: **A**) provide general demographic information and was not involved in the skin tone annotation process, **B**) inform the selection of data before the skin tone annotation process, or **C**) categorize skin tone. ***A)*** *Race Metadata for Context*: The majority of works provided aggregated racial data to provide more context about the individuals in their dataset ($n = 6$). ***B)*** *Race Metadata for Data Selection*: Three papers used race metadata to determine which images were selected for skin tone annotation [12, 82] or represented in the final dataset [38]. Every work provided explanation as to why the race of the image subjects were factored into the process. For example, Howard et al. [38] only used images of Black and White subjects because other races present in the subject pool were not well represented. In the other paper, the authors collected images from African and Nordic countries because "African countries typically have darker-skinned individuals whereas Nordic countries tend to have lighter-skinned citizens" [12]. But the authors also noted in their paper the limitation of associating skin tone with race or ethnicity.

C) *Race = Skin Tone*: Finally, four papers used race language for the skin tone annotation values [26, 56, 62, 88]. In one paper, one of the skin tone annotation categories was defined as "Asian-skin" [26]. In this case, the racialized framing of the skin tone annotations contradict the more objective purpose of skin tone annotations. Although the authors refer to skin color of the modeled subjects, additional elements such as hairstyle, eye color and facial features are important to the design of their study. By only recording skin tone, the other features encoded into the subjects are obfuscated, and the complex racial content of their subjects are uninterrogated. In a second paper, the skin tone annotation categories were "Caucasians", "Yellows", and "Blacks" [88], where one skin tone had a more overtly stereotypical description that can be assumed to be an association with individuals from Asian-descent. This assumption is confirmed by the fact that the authors referred to the image subjects as "White people, black people, and yellow people." Finally, another paper describes their skin tone annotation categories as "Black", "South Asian", "Northern Asian", and "White" [56]. While these categories offer more distinction than the previous works, the racializing of the annotation still contradicts the intended objective purpose of skin tone annotations. The authors rationalize their choice by asserting that the ethnic homogeneity of the different regions can extend to skin tone homogeneity; however, skin tone annotation is fundamentally visual. **Uncertainty Analysis**: Only two works reported any uncertainty evaluations [38, 46]. These works were classified as robust annotation procedures and also notably did not use racial language in the skin tone annotation labels. As for the other works, the lack of uncertainty analysis is consistent with the larger lack of engagement with ambiguity in the annotation process.

Summary: This category of papers is the clearest demonstration of why skin tone annotation processes are sociotechnical projects. The creators' perspectives on skin tone and race were embedded throughout the whole process, both explicitly and implicitly.

4 EXPERIMENTING WITH SKIN TONE ANNOTATION DESIGN

As can be seen, works varied widely in their skin tone annotation processes when reported, though there was also a lack of reporting about these processes in many papers. We note that many of the manually deployed practices have not been tested or compared against each other. In addition, it is unclear how different design decisions may impact annotator uncertainty or (dis)agreement. Thus, we performed an annotation experiment to investigate how different designs decisions in the annotation process may affect the resulting annotations.

4.1 Method

Annotation Tool: We based the design of our annotation tool on a range-based annotation system, Goldilocks [17], which utilizes a two-stage process to collect range ratings on a continuous scale (Figure 1). One goal in our experiment was to explore limitations of the level of granularity afforded by existing scales. Thus, we used continuous ratings to give annotators full freedom in assigning their ratings while referencing a scale rather than being tied to the established levels on a specific skin tone scale. Additionally, like in Goldilocks, each annotator provided both upper and lower bounds rather than a single rating value. This allowed for a per-annotator estimate of uncertainty around each annotation, improving our insights into uncertainty beyond disagreement metrics. To ground the pre-existing scales (either MST [71] or FST [28]), we displayed the color swatches that define the scale levels as anchors under our continuous scale. As MST and FST do not specify distances between levels, we placed anchors uniformly across our scale. We also provided an anchor in the form of an example image drawn from the dataset of the task images.

Experiment Setup: For an annotator on a rating task, the most important factors they will engage with are the scale and the items being rated. Thus, we set up the following conditions (denoted in SMALL-CAPS) to test how differences here might affect the skin tone annotations produced:

- Scale Type: We tested the 6-point FST [28] scale (FITZ) and the 10-point MST [63] scale (MONK).
- Scale Order: Since each scale is mapped to a [0, 1] range, we tested varying the scale order to be either *lighter to darker* (LD) where lower values represent lighter skin tones, or *darker to lighter* (DL) where lower values represent darker skin tones. Scale values are presented left-to-right for increasing values.
- Image Type: We tested 2 different types of images, based on whether a whole face was visible. Images from the SKIN condition contained images of skin conditions without faces or with partial facial features [33] while images from the FACE condition contained portrait photos with the subject's face [54].

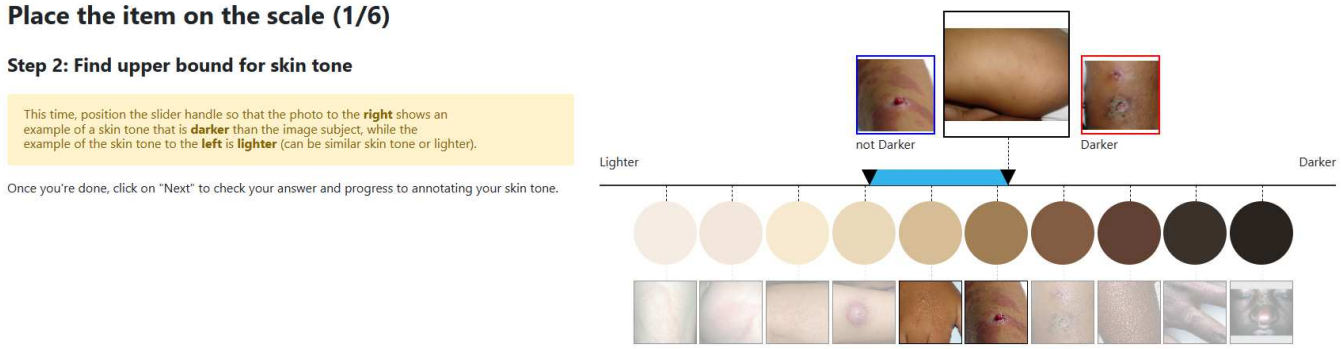


Figure 1: A screenshot of the annotation interface being used to annotate an example image in the SKIN dataset.

In our experiments, we tested all combinations above, resulting in 8 total combinations for how the annotation tool could be configured. Across all conditions, annotators were asked to first familiarize themselves with the tool by completing a tutorial using the tool under that configuration, with feedback mechanisms on their answer similar to gated instructions used in crowd task training [50], before proceeding to the annotation task. To investigate annotator positionality, we asked annotators to mark their own skin tone along the scale using the annotation tool after completing the tutorial. Based on prior recommendations [38], we provide tips for assessing their skin tone.

Image Dataset: We drew images from two datasets: the Fitzpatrick17k [32] (SKIN) and the IARPA Janus Benchmark-C (IJB-C) [54] (FACE). The Fitzpatrick17k dataset is a 17,000 image dataset of medical images of skin conditions from a variety of publicly available medical textbooks. The IJB-C dataset holds over 138,000 publicly available images from the Internet. Both datasets were manually annotated using FST and included annotations from crowdworkers annotators. We curated images from each dataset to achieve a spread across skin tones, resulting in 12 selected images for the FACE dataset and 11 images for the SKIN dataset.Annotators were then randomly assigned to annotation task sessions where they annotated one group of 6 images under one of our 8 conditions.

Annotator Recruitment: For the main component of our annotation study, we recruited 160 U.S.-based annotators from AWS MTurk with the criteria of having completed at least 1000 tasks with a 95% or higher approval rate. Crowd annotators were paid $8.50 per task ($2.50 base pay and $1.00 bonus per annotation) for a 30 minute task, and were not allowed to participate in more than one condition. We conducted manual quality control checks for spam behavior and redeployed instances where this was observed. After removing incomplete tasks, we had a final set from 153 crowd annotators. Annotators were asked to provide general demographic information to determine the representation of our participant sample along race and gender identities in the U.S. context. This enabled us to diagnose and correct for demographic skew in our sample. After the AWS MTurk deployment, we noticed that our recruited annotator population skewed heavily towards those who self-reported as White (84.2%). Prior work surveying MTurk workers has identified similar demographic imbalances [24]. To correct our demographic skew to be more in line with the U.S. adult population [13], we augmented the crowd annotators with an additional sample of annotators recruited through social channels and personal networks, focusing on increasing the representation of those who identified as non-White. These annotators were paid in the form of a $10 gift card and were assigned one of the conditions randomly. At the end we recruited 12 additional non-crowd annotators.

The final demographic distribution of our full annotator pool of 165 participants was: 78.2% White, 7.9% Asian, 6.7% Black, 4.8% Latino, 0.6% Native-American, and 1.8% multiple. The gender distribution of annotators was 40.6% female, 58.8% male, 0.6% non-binary. We also asked participants to self-report their own skin tone using the annotation tool; the distribution is shown in Appendix B. We discuss limitations of our sample in Section 6.

Designer and Annotator Positionality: A limitation of prior literature was the opaqueness of the designer positionality. When the authors did not describe the decision-making behind their process, we could only guess what social, political, and ethical perspectives informed the design. By stating our positionality as designers, we hope to not only reflect on how our identities impacted our design but directly engage with the limits of our perspectives. The study was scoped to the U.S. as all of the authors reside in the U.S. and are at U.S. institutions. This informed decisions such as the ethnicity and gender categories we used in our survey. We also required annotators to be U.S.-based, so that they are also situated in the racial and skin tone-stratification contexts of the U.S.

For Black populations in the U.S., skin tone has been used since the Atlantic Slave Trade to allocate degrees of social power to enslaved people [64]. Even today, the associations to skin tone are reflected in sentencing trends in the U.S. justice systems [63]. Given the context, there is great responsibility placed on an annotator when they select skin tone for themselves and image subjects. Thus, we surveyed annotators about their comfort with the task of annotating skin tone (Appendix C). Finally, as mentioned, we asked annotators to mark their own skin tone—this enabled us to investigate another aspect of annotator positionality.

4.2 Findings

4.2.1 RQ 1: Do the scales (FITZ, MONK) correlate with each other for measuring skin tones? For our first research question, we wanted

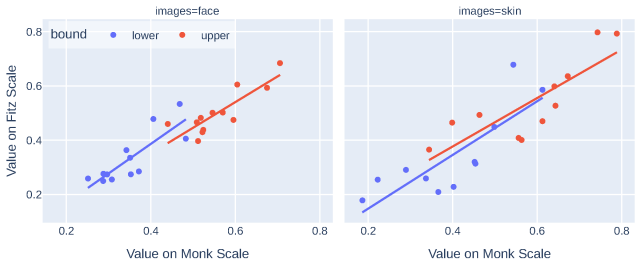


Figure 2: (RQ1) Correlation between the two scale types for both bounds. Each data point indicates the aggregated skin tone rating for the corresponding bound on an image. Higher values indicate darker skin tones. Values produced by DL conditions are remapped onto a light-to-dark scale before aggregating.

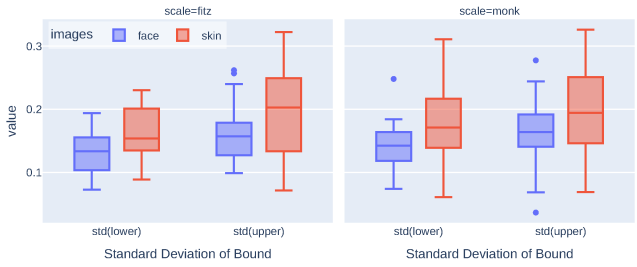


Figure 3: (RQ2) Box plot of agreement (measured as standard deviation of each bound, lower values indicate more agreement) for each scale type and image type. The only significant variable is the *image type* (shown here as red and blue colors).

to confirm whether there is general agreement in the values produced between the two skin tone annotation scale types we used. Across both image types (FACE and SKIN), we found high positive correlation for the upper ($R^2 = 0.721$ and 0.680 respectively) and lower bounds ($R^2 = 0.692$ and 0.662 respectively) produced by annotators using the two scales (Figure 2). This result largely serves as a check to validate that both skin tone scales were indeed able to capture differences across a range of skin tones and that annotators were generally able to utilize our annotation interface with existing scales for annotating skin tones.

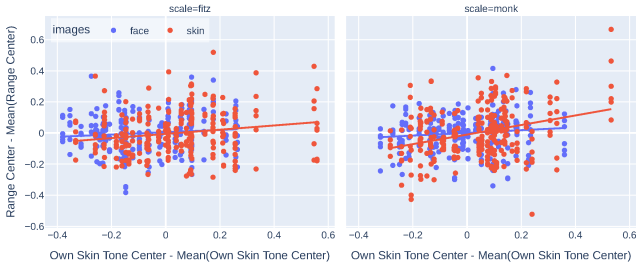
4.2.2 RQ 2: Does scale type, ordering of scale, or image type affect the agreement between annotators? To examine whether our control variables affected agreement between annotators, we used a linear model multi-way ANOVA test to compare the effect of the independent variables: scale type (FITZ, MONK), scale order (LD, DL), and image type (FACE, SKIN) variables as well as any pairwise interactions, on the dependent variable of standard deviation of the upper and lower bounds. The standard deviation of each bound is used as it is a common way to characterize inter-annotator agreement in a continuous rating scale setting (Figure 3).

We found that for both lower and upper bounds, the only significant variable (at $p = 1.5 \times 10^{-3} < 0.05$, and $p = 2.9 \times 10^{-3} < 0.05$ for each bound respectively) that affected annotator agreement was the **image type**, with images in the FACE condition showing more agreement than that of SKIN. As skin tone annotation is most commonly conducted on datasets involving portrait shots with visible faces, this finding may suggest that skin tone annotation on less common image types, like images of skin patches without faces, may result in lower agreement between annotators as they have less context to draw from. However, the inclusion of faces also potentially biases annotators towards using other contextual features such as race and ethnicity, which may have affected their consistency.

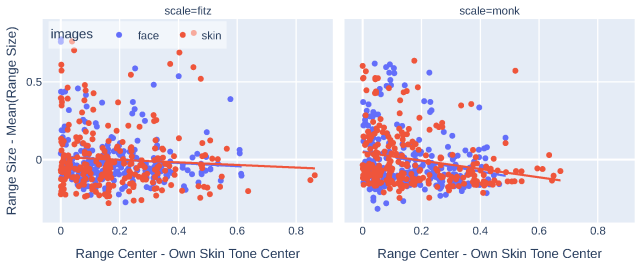
Figure 4: (RQ3) Box plot of uncertainty (measured as the size of each range, lower values indicate more certainty) for each scale type and image type. We see significant effects on uncertainty from both the *scale order* and the *scale type* × *image type* interaction.

4.2.3 RQ 3: Does scale type, ordering of scale, or image type affect the uncertainty of each annotator? While previously we explored the effects of our controlled variables on agreement *between* different annotators, here we examined whether the configuration of annotation system affects each annotator's own *individual* uncertainty during annotation. As we utilized a range-based annotation system, we were able to quantify this uncertainty by examining the *size* of the ranges produced by each annotator. Similar to the section before, we conducted an ANOVA test to compare our independent variables and pairwise interactions against the *size* of the range produced by each annotator (Figure 4).

We found two significant effects: the **scale order** ($p = 0.029 < 0.05$) and the interaction of **scale type** × **image type** ($p = 0.010 < 0.05$). Of the latter interaction effect, we found that there was a statistically significant difference between the pairings (FITZ x FACE) and (MONK x FACE) at $p = 0.024 < 0.05$ (identified via Tukey's HSD). Thus, we conclude that the order in which the skin tone is presented can affect individual annotators' own evaluation of their uncertainty. However, *how* it affected uncertainty seems to depend on the type of image and scale being used. Overall, we found that using a *dark-to-light* scale ordering tended to result in lower individual uncertainty. However, when annotating SKIN images, this trend was not observed for the Fitzpatrick scale. We hypothesize that this may have been the result of multiple factors at work: As a large majority of our annotators self-reported lighter skin tones, when utilizing our annotation process which establishes the lower bound first, a *dark-to-light* scale may result in better estimation of the lower bound from more easily contrasting skin



(a) Value bias: x-axis shows the annotator's self-reported skin tone relative to the mean across all annotators. y-axis shows the annotator's judgment of skin tone of each data point relative to the respective mean of that item in the associated condition. Higher = darker skin tone.



(b) Uncertainty bias: x-axis shows the distance between the annotated item's skin tone and the annotator's self-reported skin tone. y-axis shows the uncertainty of the item relative to the mean uncertainty of that item in the respective condition

Figure 5: (RQ4) Two scatter plots showcasing potential correlations between the annotators' own self-reported skin tone and the annotations produced by them.

tones, reducing the size of the final range. On the other hand, the lack of additional race and ethnicity context in the SKIN task may have worked in conjunction with the FITZ scale's specialization for directly annotating skin tone on skin patches, reducing this effect.

4.2.4 RQ 4: Does the annotator's own self-reported skin tone (positionality) bias their uncertainty or agreement? Finally, we examined whether the annotators' self-reported skin tone correlated with how they produced skin tone annotations. Specifically, we tested for two forms of potential biases that could occur. For **value bias**, we looked at whether the annotator's own skin tone (relative to the mean across all annotators) correlated with their annotations (relative to the mean across annotations produced under the same experiment conditions). A positive or negative correlation here would indicate that an annotator biases their annotations towards or away from their own skin tone. As shown in Figure 5a, we note a very weak (not significant) positive trend, with the most prominent $R^2 = 0.100$ observed for SKIN images using the MONK scale. This suggests that annotators may potentially bias their annotated skin tone towards their own. However, we also note the caveat that data points get much sparser the further we get from the mean skin tone, which we hypothesize is likely due to the demographic concentration of our annotators.

For **uncertainty bias**, we examined whether the difference between the annotated image's skin tone and the annotator's own skin tone correlated with their self-reported uncertainty in the form of the size of their range. A positive or negative correlation here would indicate that an annotator is more or less uncertain the further the annotated image's skin tone is from their own skin tone. As shown in Figure 5b, we note a very weak (not significant) negative trend, with the most prominent $R^2 = 0.056$ observed for SKIN images using the MONK scale. This may suggest that annotators are potentially more certain when annotating images that have a skin tone different from their own. However, as before, any potential trends may be a result of the lack of annotators on the darker end of the skin tone spectrum.

5 DISCUSSION AND RECOMMENDATIONS

Associations and Differences between Skin Tone and Race. The degree to which race data is embedded in the skin tone annotation process often reveals a Western hegemonic understanding of the association of skin tone to race. While we indeed observed correlations between self-reported skin tone and race in our annotation study, such correlations are far from a perfect association between the two. We argue that skin tone annotations are not just a proxy to associate certain ethnic groups to skin tone categories; instead they capture a spectrum of additional complexions not afforded by simple ethnic groups and race demographics. This can be especially salient for underrepresented populations within the Black and Asian diasporas. In papers that have noted the lack of representation of skin tones of Black or Asian populations [83], authors have suggested including skin tones to better represent the extent of possible skin tones among these populations. Throughout the literature review, we observed implicit and explicit associations between skin tone and race that manifested in the designers' decision-making. We encourage researchers seeking to engage with datasets involving skin tone and race to take care in capturing these associations. For instance, they could capture annotator positionality through annotator self-reported skin tone and demographic information in conjunction with the collection of annotations.

Uncertainty Present in Skin Tone Annotations. In both our literature review and our own experiments, we have observed that even the more objective measurement of skin tone is still associated with factors of uncertainty. Within our study, aspects like the ordering of the scale were observed to affect individual uncertainty, while the type of image affected collective agreement. In practice, the aspect of uncertainty in annotations can often be overlooked, leading to over-estimates of the capabilities of downstream models [31]. While design factors of our study limited our ability to further examine the mechanisms behind the observed uncertainty, our results demonstrate the need for recruiting a diverse pool of annotators when conducting skin tone annotations and collecting multiple annotations per item so that uncertainty can be identified. We also urge anyone collecting skin tone annotations for modeling to be cognizant about how uncertainty may limit the actual capabilities of models produced.

Transparency in Skin Tone Annotation Procedures. As noted in our literature review, more than 60% of the corpus had ambiguously documented procedures. The lack of transparency

and consistent reporting in skin tone annotation processes is a major issue for current work using skin tone annotation. In addition, our review highlighted components of the annotation process that could be standardized or at least explained in a more meaningful way. Coupled with our findings of factors that affected annotation results in the annotation study, we believe that providing transparency around the procedures and configuration decisions in skin tone annotation will be an important step if we are to use skin tone data as ground truth for ensuring the fairness of algorithms and datasets. Indeed, even the most robust skin tone annotation procedures in prior work reported some level of unreliability in their annotations, meaning that transparency will be an important way for downstream users to understand the limitations resulting from the subjectivity of skin tone annotation.

Consider Effects from Scale Order. In our annotation experiments, we found that arranging the scale from *darker to lighter* somewhat surprisingly correlated with a significant reduction in individual annotators' uncertainty. The hegemony of the lighter to darker hierarchy associated to skin tone stratification might have played a role in creating implicit biases leading to this trend. For example, FST is encoded with category values that increase in value from the lighter to darker skin tones. In the literature review, even when the scale was not FST, scale values were often ordered from lighter to darker categories. In a society in which lighter skin tones yield a social proximity to whiteness, ordering skin tone with lighter skin first enforces the hierarchy of social power. Since our tool establishes the lower boundary first, starting with a skin tone spectrum default of lighter to darker shades could mean that our set of majority-White U.S.-based annotators produce a less precise rating when exploring the spectrum of lighter skin tones. On the other hand, when the annotator is presented the less common darker-to-lighter scale, the novelty may have encouraged the annotators to consider their annotations with more intention. We pose that the interaction between how scales are explored and ordered should be an aspect to which researchers collecting skin tone annotations pay greater attention.

Consider Effects from Dataset Type. Another area of sensitivity we identified was the type of image data involved in the annotation. In our own experiments, the SKIN dataset involved the classification of skin tones based on images of skin disease rather than the more common classification of portrait photos. Some annotators expressed in their free-response feedback that they were caught off guard by the unsettling nature of such images even though their nature was indicated in our recruitment, task prompts, and consent forms. Given this, we recommend taking additional care around the task design and training phases when conducting annotations on potentially sensitive image types.

More generally though, our experiments exposed the sensitivity of skin tone annotation processes to the types of images being annotated. In our results in Section 4.2.2, we hypothesized that the additional context provided through the portraits relating to the subject's race and ethnicity may be the reason higher agreement is observed. Krishnapriya et al. [46] went so far as to suggest that human annotators were more consistent than the fully automated system because they factored in the race of the image subjects in their skin tone analysis.

6 LIMITATIONS AND FUTURE WORK

One limitation of our study arises from the scope of the annotation task and the recruitment of annotators. While our study explored whether potential differences exist between different image types, we only utilized a limited sub-sample of images of each type and measured along the two most commonly used skin tone scales. A larger scale study involving a larger set of images that control for additional factors would be helpful in shedding light on *how* different factors affect annotations rather than our exploration of *whether* they do.

Additionally, a majority of our annotators were recruited through online crowdsourcing platforms, leading to an observed imbalance in the annotator demographics (Section 4.1). This is not surprising, as online crowd work platforms have long been known to have a skewed demographic breakdown compared to the general population, presenting challenges for studies involving subjectivity [24, 40, 72]. Even though our additional recruitment yielded a more representative final demographic distribution, the majority of annotators overall still identified as White, contributing to an over-representation of lighter skin tones in self-reported ratings (Appendix B). In addition, we were not able to recruit subject-matter experts (i.e., dermatology) to provide an expert-informed source for us to evaluate how differences we observed impacted the final quality or correctness of annotations themselves in relation to scales like FST. We note that works in our literature review have used dermatologists to validate annotations [12] or provide a set of annotations as a contribution to inter-rater uncertainty analyses and consensus scores [32].

Finally, we did not compare our manual annotations against fully automated skin tone evaluation metrics and systems based on pixel data. While human-centric skin tone data and scales (like FST and MST) remain widely used, there is a body of work that proposes the use of automated metrics and systems to establish skin tone as a more objectively-defined concept. As the act of annotation itself is a human and socially-situated activity, exploring similarities and differences between automation-centered metrics and systems versus human-defined skin tone judgments is an interesting avenue for future work.

7 CONCLUSION

The central goal of this work was to investigate the social subjectivity inherent in skin tone annotations for computer vision evaluation. We achieved this through cataloguing how subjectivity was addressed in prior annotation processes, as well as through experimentation. Our literature review showed great variability in annotation procedures, an overall lack of documentation, and a lack of engagement with subjectivity. Our experiment findings indicated that many factors, like the type of data and the configuration of the annotation process, can affect aspects like agreement and uncertainty around the data produced by skin tone annotation. We contribute to the effort to address disparities in CV models by investigating how factors related to the data, annotators, and annotation process can all impact skin tone annotation outcomes. We call upon the broader community to commit to moving towards more robust, principled, and well-documented processes when working with CV tasks involving skin tones.

ACKNOWLEDGMENTS

This work was supported by an NSF REU Supplement under award #2120497. We thank all of our experiment study participants for taking part in our study. We also thank the members of the Social Futures Lab at UW CSE, the UW DUB Summer REU program, and the Affective Biometrics Lab at Howard University for feedback on the work and piloting of our experiment. We are also grateful for preliminary resources provided by the Google Monk Scale team.

REFERENCES

[1] Ali Al-Naji and Javaan Chahl. 2017. Simultaneous tracking of cardiorespiratory signals for multiple persons using a machine vision system with noise artifact removal. *IEEE journal of translational engineering in health and medicine* 5 (2017), 1–10.

[2] Ali Al-Naji and Javaan Chahl. 2018. Remote optical cardiopulmonary signal extraction with noise artifact removal, multiple subject detection & long-distance. *IEEE Access* 6 (2018), 11573–11595.

[3] Mohammed Aledhari, Rehma Razzak, Reza M Parizi, and Gautam Srivastava. 2021. Multimodal machine learning for pedestrian detection. In *2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring)*. IEEE, 1–7.

[4] Denise Almeida, Konstantin Shmarko, and Elizabeth Lomas. 2022. The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks. *AI and Ethics* 2, 3 (2022), 377–387.

[5] Janet Anderson, Charles Otto, Brianna Maze, Nathan Kalka, and James A Duncan. 2019. Understanding confounding factors in face detection and recognition. In *2019 International Conference on Biometrics (ICB)*. IEEE, 1–8.

[6] Oneida Arosarena. 2015. Options and challenges for facial rejuvenation in patients with higher fitzpatrick skin phototypes. *JAMA Facial Plastic Surgery* 17, 5 (2015), 358–359.

[7] Mariam Attia and Julian Edge. 2017. Be (com) ing a reflexive researcher: a developmental approach to research methodology. *Open Review of Educational Research* 4, 1 (2017), 33–45.

[8] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems* 32 (2019).

[9] Keivan Bahmani, Richard Plesh, Chinmay Sahu, Mahesh Banavar, and Stephanie Schuckers. 2021. SREDS: A dichromatic separation based measure of skin color. In *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 1–6.

[10] Josh Beal, Hao-Yu Wu, Dong Huk Park, Andrew Zhai, and Dmitry Kislyuk. 2022. Billion-scale pretraining with vision transformers for multi-task visual representations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 564–573.

[11] Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns* 2, 2 (2021), 100205.

[12] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[13] United States Census Bureau. 2020. 2020 Census Results. https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-results.html.

[14] L Elisa Celis and Vijay Keswani. 2020. Implicit diversity in image summarization. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–28.

[15] Cheng-Chun Chang, Shi-Tien Hsing, Yung-Chi Chuang, Chien-Ta Wu, Tung-Jing Fang, Kuan-Fu Chen, and Bill Choi. 2018. Robust skin type classification using convolutional neural networks. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2011–2014.

[16] Alain Chardon, Isabelle Cretois, and Colette Hourseau. 1991. Skin colour typology and suntanning pathways. *International journal of cosmetic science* 13, 4 (1991), 191–208.

[17] Quan Ze Chen, Daniel S Weld, and Amy X Zhang. 2021. Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[18] Cynthia M Cook, John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. 2019. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1, 1 (2019), 32–41.

[19] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.

[20] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. 2022. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Science advances* 8, 31 (2022), eabq6147.

[21] S Del Bino, J Sok, E Bessac, and F Bernerd. 2006. Relationship between skin response to ultraviolet exposure and skin color type. *Pigment cell research* 19, 6 (2006), 606–614.

[22] Emily Denton, Mark Díaz, Ian Kivlichan, Vinodkumar Prabhakaran, and Rachel Rosen. 2021. Whose ground truth? accounting for individual and collective identities underlying dataset annotation. *arXiv preprint arXiv:2112.04554* (2021).

[23] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the people back in: Contesting benchmark machine learning datasets. *arXiv preprint arXiv:2007.07399* (2020).

[24] Djellel Eddine Difallah, Elena Filatova, and Panagiotis G. Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (2018).

[25] Manan Doshi, Jimil Shah, Rahul Soni, and Soni Bhambar. 2022. FHP: Facial and Hair Feature Processor for Hairstyle Recommendation. In *2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC)*. IEEE, 1–4.

[26] Rafael Falcon, Mauro Patti, Stanislas Brochard-Garnier, G Pacianotto Gouveia, S Torres Acevedo, Thelma Bergot, Rick Alarcon, Corentin Bomstein, Hervé Macudzinski, P Maitre, et al. 2022. Image quality evaluation of video conferencing solutions with realistic laboratory scenes. *Electronic Imaging* 34, 9 (2022), 318–1.

[27] Todd Feathers. 2021. Google's new dermatology app wasn't designed for people with darker skin. *Retrieved August* 10 (2021), 2022.

[28] Thomas B. Fitzpatrick. 1988. The Validity and Practicality of Sun-Reactive Skin Types I Through VI. *Archives of Dermatology* 124, 6 (06 1988), 869–871. https://doi.org/10.1001/archderm.1988.01670060015008

[29] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.

[30] Manuel B Garcia, Teodoro F Revano, Beau Gray M Habal, Jennifer O Contreras, and John Benedic R Enriquez. 2018. A pornographic image and video filtering application using optimized nudity recognition and detection algorithm. In *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)*. IEEE, 1–5.

[31] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 388, 14 pages. https://doi.org/10.1145/3411764.3445423

[32] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. 2022. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. *arXiv preprint arXiv:2207.02942* (2022).

[33] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1820–1828.

[34] Yağmur Güçlütürk, Umut Güçlü, Katja Seeliger, Sander Bosch, Rob van Lier, and Marcel A van Gerven. 2017. Reconstructing perceived faces from brain activations with deep adversarial neural decoding. *Advances in neural information processing systems* 30 (2017).

[35] Ammar Haider and Nosheen Sabahat. 2022. A Usability and Accuracy Measurement of Smartphones Face Recognition. In *2022 2nd International Conference on Artificial Intelligence (ICAI)*. IEEE, 19–25.

[36] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton Ferrer. 2021. Towards measuring fairness in ai: the casual conversations dataset. *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2021).

[37] John Hoberman. 2012. *Black and blue: The origins and consequences of medical racism*. Univ of California Press.

[38] John J Howard, Yevgeniy B Sirotin, Jerry L Tipton, and Arun R Vemury. 2021. Reliability and validity of image-based and self-reported skin phenotype metrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 4 (2021), 550–560.

[39] Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 306–316.

[40] Gabriella Kazai, J. Kamps, and Natasa Milic-Frayling. 2012. The face of quality in crowdsourcing relevance labels: demographics, personality and labeling accuracy. *Proceedings of the 21st ACM international conference on Information and knowledge management* (2012).

[41] Zaid Khan and Yun Fu. 2021. One label, one billion faces: Usage and consistency of racial categories in computer vision. In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*. 587–597.

[42] Hannah Kim, Girmaw Abebe Tadesse, Celia Cintas, Skyler Speakman, and Kush Varshney. 2022. Out-of-distribution detection in dermatology using input perturbation and subset scanning. In *2022 IEEE 19th International Symposium on*

Biomedical Imaging (ISBI). IEEE, 1–4.
[43] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. 2019. Estimating skin tone and effects on classification performance in dermatology datasets. *arXiv preprint arXiv:1910.13268* (2019).
[44] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on information forensics and security* 7, 6 (2012), 1789–1801.
[45] Reeta Koshy, Anisha Gharat, Tejashri Wagh, and Siddesh Sonawane. 2021. A Complexion based Outfit color recommender using Neural Networks. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*. IEEE, 1–7.
[46] KS Krishnapriya, Gabriella Pangelinan, Michael C King, and Kevin W Bowyer. 2022. Analysis of Manual and Automated Skin Tone Assignments. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 429–438.
[47] Camila Laranjeira da Silva, João Macedo, Sandra Avila, and Jefersson dos Santos. 2022. Seeing without looking: Analysis pipeline for child sexual abuse datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2189–2205.
[48] Minh-Ha Le, Md Sakib Nizam Khan, Georgia Tsaloli, Niklas Carlsson, and Sonja Buchegger. 2020. Anonfaces: Anonymizing faces adjusted to constraints on efficacy and security. In *Proceedings of the 19th Workshop on Privacy in the Electronic Society*. 87–100.
[49] JC Lester, JL Jia, L Zhang, GA Okoye, and E Linos. 2020. Absence of images of skin of colour in publications of COVID-19 skin manifestations. *British Journal of Dermatology* 183, 3 (2020), 593–595.
[50] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H Lin, Xiao Ling, and Daniel S Weld. 2016. Effective crowd annotation for relation extraction. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 897–906.
[51] Xin Liu, Ziheng Jiang, Josh Fromm, Xuhai Xu, Shwetak Patel, and Daniel McDuff. 2021. MetaPhys: few-shot adaptation for non-contact physiological measurement. In *Proceedings of the conference on health, inference, and learning*. 154–163.
[52] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.
[53] MWP Maduranga and Dilshan Nandasena. 2022. Mobile-based skin disease diagnosis system using convolutional neural networks (CNN). *IJ Image Graphics Signal Process.* 3 (2022), 47–57.
[54] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. 2018. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*. IEEE, 158–165.
[55] Daniel McDuff, Xin Liu, Javier Hernandez, Erroll Wood, and Tadas Baltrusaitis. 2021. Synthetic Data for Multi-Parameter Camera-Based Physiological Sensing. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 3742–3748.
[56] Daniel McDuff, Shuang Ma, Yale Song, and Ashish Kapoor. 2019. Characterizing bias in classifiers using generative models. *Advances in neural information processing systems* 32 (2019).
[57] Anay Mehrotra and L Elisa Celis. 2021. Mitigating bias in set selection with noisy protected attributes. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 237–248.
[58] Hanna F Menezes, Arthur SC Ferreira, Eanes T Pereira, and Herman M Gomes. 2021. Bias and Fairness in Face Detection. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. IEEE, 247–254.
[59] Arjun Menon. 2023. Leveraging Facial Recognition Technology in Criminal Identification. *Interdisciplinary Innovations and Developments towards Smart and Sustainable Industries https://doi. org/10.13052/rp-978-87-7022-828-2* (2023).
[60] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. *arXiv preprint arXiv:1901.10436* (2019).
[61] Shiksha Mishra, Puspita Majumdar, Muskan Dosi, Mayank Vatsa, and Richa Singh. 2021. Dual sensor indian masked face dataset. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 1–8.
[62] David A Molina, Leonardo Causa, and Juan Tapia. 2020. Reduction of bias for gender and ethnicity from face images using automated skin tone classification. In *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 1–5.
[63] Ellis P Monk. 2019. The color of punishment: African Americans, skin tone, and the criminal justice system. *Ethnic and Racial Studies* 42, 10 (2019), 1593–1612.
[64] Ellis P Monk Jr. 2021. The unceasing significance of colorism: Skin tone stratification in the United States. *Daedalus* 150, 2 (2021), 76–90.
[65] Ellis P Monk Jr, Jerry Kaufman, and Yadira Montoya. 2021. Skin tone and perceived discrimination: Health and aging beyond the binary in NSHAP 2015. *The Journals of Gerontology: Series B* 76, Supplement_3 (2021), S313–S321.
[66] Vidya Muthukumar. 2019. Color-theoretic experiments to understand unequal gender classification accuracy from face images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.
[67] Cathy O'neil. 2017. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.
[68] Chunjong Park, Anas Awadalla, Tadayoshi Kohno, and Shwetak Patel. 2021. Reliable and trustworthy machine learning for health using dataset shift detection. *Advances in Neural Information Processing Systems* 34 (2021), 3043–3056.
[69] Jisoo Park, Hyungjoon Kim, Seonmi Ji, and Eenjun Hwang. 2018. An automatic virtual makeup scheme based on personal color analysis. In *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication*. 1–7.
[70] PRH Perera, ESS Soysa, HRS De Silva, ARP Tavarayan, MP Gamage, and KMLP Weerasinghe. 2021. Virtual Makeover and Makeup Recommendation Based on Personal Trait Analysis. In *2021 3rd International Conference on Advancements in Computing (ICAC)*. IEEE, 288–293.
[71] Google Research. 2022. Developing the Monk Skin Tone Scale. https://skintone.google/the-scale.
[72] Joel Ross, Lilly C. Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (2010).
[73] Tate Ryan-Mosley. 2021. The new lawsuit that shows facial recognition is officially a civil rights issue. https://www.technologyreview.com/2021/04/14/1022676/robert-williams-facial-recognition-lawsuit-aclu-detroit-police/.
[74] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW1 (2020), 1–35.
[75] Zaineb Shah, Syed Ayaz Ali Shah, Aamir Shahzad, Ahmad Fayyaz, Shoaib Khaliq, Ali Zahir, and Goh Chuan Meng. 2022. Deep Learning-Based Forearm Subcutaneous Veins Segmentation. *IEEE Access* 10 (2022), 42814–42820.
[76] Tomáš Sixta, Julio CS Jacques Junior, Pau Buch-Cardona, Eduard Vazquez, and Sergio Escalera. 2020. Fairface challenge at eccv 2020: Analyzing bias in face recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, 463–481.
[77] Radim Spetlík, Jan Cech, and Jiri Matas. 2018. Non-contact reflectance photoplethysmography: Progress, limitations, and myths. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 702–709.
[78] Shea Swauger. 2020. Our bodies encoded: Algorithmic test proctoring in higher education. *Critical digital pedagogy* (2020).
[79] Muhammad Uzair Tariq, Arup Kumar Ghosh, Karla Badillo-Urquiola, Abhiditya Jha, Sanjeev Koppal, and Pamela J Wisniewski. 2018. *Designing light filters to detect skin using a low-powered sensor*. IEEE.
[80] Yuushi Toyoda, Gale Lucas, and Jonathan Gratch. 2021. Predicting Worker Accuracy from Nonverbal Behaviour: Benefits and Potential for Algorithmic Bias. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*. 25–30.
[81] Alex S Vitale. 2021. *The end of policing*. Verso Books.
[82] Wenjin Wang, Albertus C Den Brinker, Sander Stuijk, and Gerard De Haan. 2016. Algorithmic principles of remote PPG. *IEEE Transactions on Biomedical Engineering* 64, 7 (2016), 1479–1491.
[83] Olivia R Ware, Jessica E Dawson, Michi M Shinohara, and Susan C Taylor. 2020. Racial limitations of Fitzpatrick skin type. *Cutis* 105, 2 (2020), 77–80.
[84] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. 2017. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 90–98.
[85] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. 2019. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097* (2019).
[86] Seyma Yucer, Furkan Tektas, Noura Al Moubayed, and Toby P Breckon. 2022. Measuring hidden bias within face recognition via racial phenotypes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 995–1004.
[87] Longhao Zhang, Xipeng Pan, Huihua Yang, and Lingqiao Li. 2020. On Open-Set, High-Fidelity and Identity-Specific Face Transformation. *IEEE Access* 8 (2020), 224643–224653.
[88] Xiaobiao Zhang, Xiaoyi Feng, and Zhaoqiang Xia. 2019. Analysis of Factors on BVP Signal Extraction Based on Imaging Principle. In *Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications*. 48–55.
[89] Qian Zheng, Ankur Purwar, Heng Zhao, Guang Liang Lim, Ling Li, Debasish Behera, Qian Wang, Min Tan, Rizhao Cai, Jennifer Werner, et al. 2022. Automatic facial skin feature detection for everyone. *arXiv preprint arXiv:2203.16056* (2022).

A LITERATURE REVIEW DIMENSIONS OF ANALYSIS

Below in Table 3, we list all dimensions the first author used to qualitatively analyze the papers in our literature review. We also provide a more detailed definition of that dimension along with an example of an annotation for a paper. The full dataset of papers along with the annotated values for each dimension is provided at https://github.com/Social-Futures-Lab/skin-deep/.

Dimension	Definition	Examples
	Dataset Information	
Task	The intended Computer Vision task that will be trained or tested by the annotated dataset.	Facial Recognition
Year	The year the paper was published.	2017
Dataset Source	The source of the images or videos in the annotated dataset.	Flickr, CelebA
Publicly Available	The resulting dataset (including the annotations) is available to the public	Yes/No
No. of Subjects	The number of individual human subjects represented in the dataset	15, 20
No. of Images/Videos	The number of images or videos are a part of the dataset	100, 40/300, 40 videos
	Annotation Procedure	
Procedure Description	An explanation of the skin tone scale definition, annotation platform and procedural restrictions	"The apparent skin tone and lighting attributes were labeled by a group of human evaluators [58]."
Annotator	The individual or individuals who conducted the skin tone annotations.	Author(s), Experts
Scale/Measurement	The numerical scale or categories used to describe the range of skin tones represented by the annotations.	Fitzpatrick scale, "lighter" and "darker"
Annotation Distribution	The proportion of each skin tone category present in the annotated dataset.	Lighter (50%), Darker (50%)
	Annotation Uncertainty	
Analysis Description	An explanation of the methods, calculations, or qualitative findings used to evaluate inter-rater agreement or uncertainty	"We compute the Cohen's k-coefficient for all pairs of participants with more than 5 common images in their surveys [14]."
Analysis Results	Any quantitative metric for the uncertainty of the skin tone annotations.	"k-coefficient: 0.58 (median: 0.62)"
	Race Annotations or Metadata	
Racial Categories	The categories used to describe the various races or ethnicities of the dataset subjects.	Black, South Asian, Western European
Category Distribution	The proportion of each racial category presented in the annotated dataset.	Black (100%)
	Additional Information	
Comments	Relevant excerpts, thoughts or external work that provide context to the annotation process in the work.	The work also used a dataset that was already annotated for skin tone.

Table 3: The definitions and examples of all the dimensions evaluated in the literature review.

B ANNOTATOR SELF-REPORTED SKIN TONE DISTRIBUTIONS

We see a diverse spread of skin tone values as self reported by annotators, with a noted bias towards the lighter side of each scale overall (Figure 6). Skin tone values can present a space that is much richer than simple racial demographics. We can also observe differences between the two scales, where, despite the entire continuous range [0, 1] being available in both cases, the less-granular FITZ scale results in less resolution compared to the MONK scale that contains more levels.

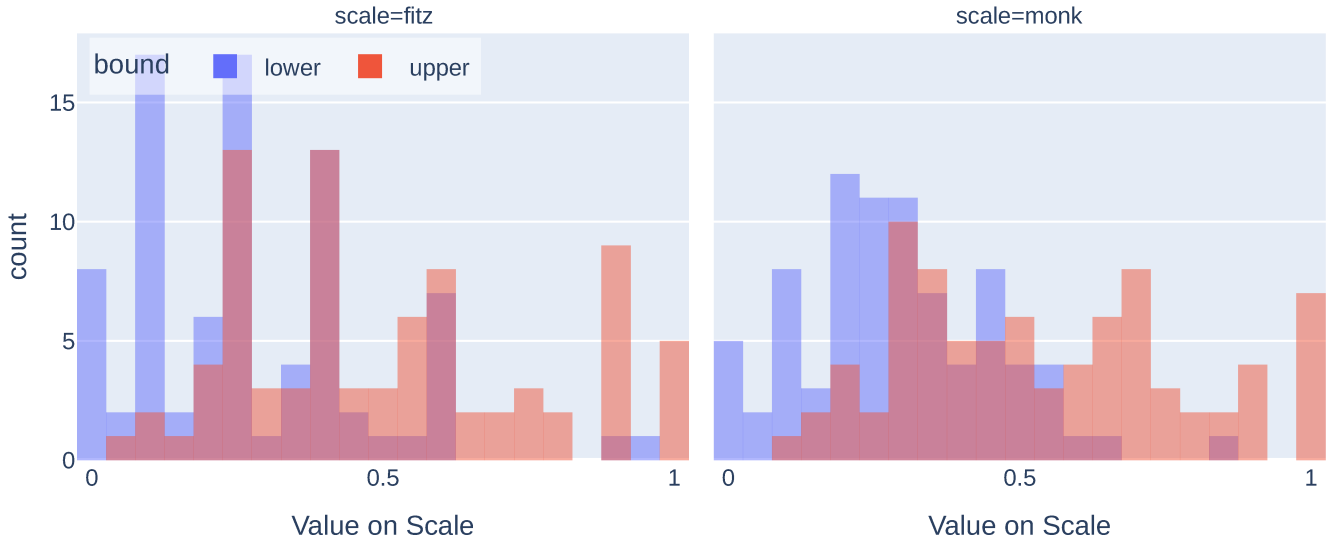


Figure 6: A histogram of the distribution of self-reported skin tone values (upper and lower bounds) shown for each scale. Lower values indicate lighter skin tone while higher values indicate darker.

C POST-ANNOTATION SURVEY

After the main annotation task the participants were prompted to answer the following statements with likert scale response of agreement (strongly disagree, disagree, neutral, agree, strongly agree):

- I am confident in the labeling of my skin tone.
- I am confident in my annotations of the images.
- Throughout history people have been categorized by the color of their skin by other individuals placed in positions of social power. These categorizations were used to enforce harmful social hierarchies based on racism and colorism (i.e. Jim Crow era in Southern US). Even today in the United States, people have different life experiences based on the long-term impact of skin tone categorizations (i.e. arrest rates in the United States). Given this particular context, indicate how much you agree with the following statements: I considered the race and/or ethnicity of the image subject while annotating.
- I felt comfortable annotating the skin tone of the given images in this task.
- I felt comfortable annotating my **own** skin tone.

To further clarify the statements, short definitions of key words were provided as well.

- Confident: certain to a significant degree
- Consider: take into account, especially before making a decision

C.1 Results from Annotator Self-Reported Experiences

We examined the survey results to explore how the participants perceived the task of annotating skin tone. As can be seen in the results from Figure 7, overall most annotators were comfortable with the idea of annotating skin tone in general (both for the images and themselves). We also observed that compared to the other questions, more annotators self reported disagreement with the statement that race and/or ethnicity were a factor in their annotation consideration.

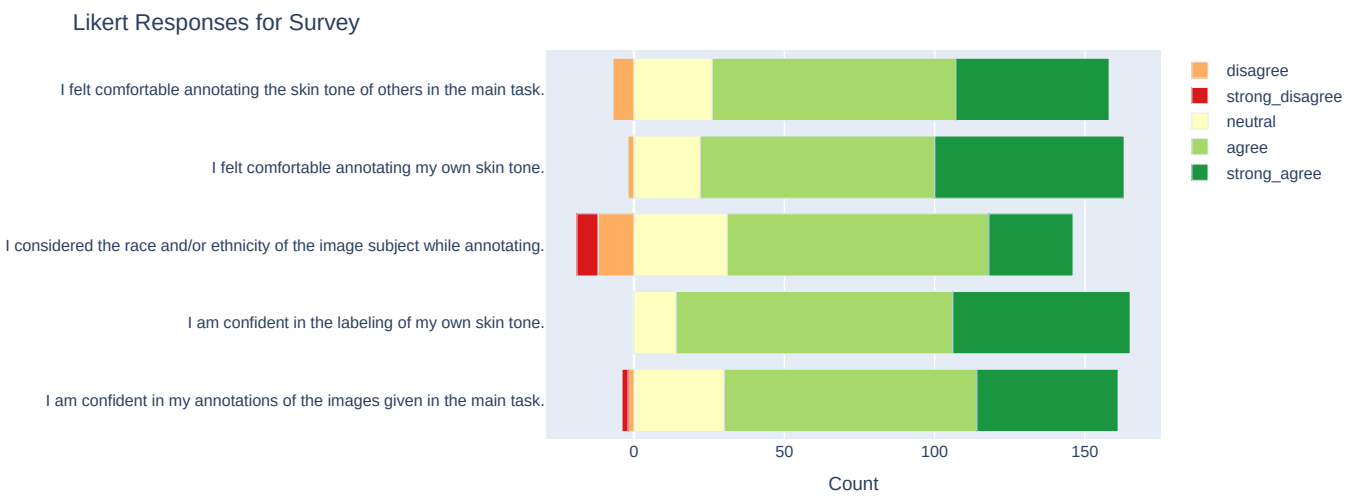


Figure 7: Diverging stacked bar chart of responses to the survey.