FACET: Fairness in Computer Vision Evaluation Benchmark

Laura Gustafson Chloe Rolland Nikhila Ravi Quentin Duval Aaron Adcock
Cheng-Yang Fu Melissa Hall Candace Ross
Meta AI Research, FAIR
facet@meta.com

Abstract

*Computer vision models have known performance disparities across attributes such as gender and skin tone. This means during tasks such as classification and detection, model performance differs for certain classes based on the demographics of the people in the image. These disparities have been shown to exist, but until now there has not been a unified approach to measure these differences for common use-cases of computer vision models. We present a new benchmark named FACET (**FA**irness in **C**omputer Vision **E**valua**T**ion), a large, publicly available evaluation set of 32k images for some of the most common vision tasks - image classification, object detection and segmentation. For every image in FACET, we hired expert reviewers to manually annotate person-related attributes such as perceived skin tone and hair type, manually draw bounding boxes and label fine-grained person-related classes such as disk jockey or guitarist. In addition, we use FACET to benchmark state-of-the-art vision models and present a deeper understanding of potential performance disparities and challenges across sensitive demographic attributes. With the exhaustive annotations collected, we probe models using single demographics attributes as well as multiple attributes using an intersectional approach (e.g. hair color and perceived skin tone). Our results show that classification, detection, segmentation, and visual grounding models exhibit performance disparities across demographic attributes and intersections of attributes. These harms suggest that not all people represented in datasets receive fair and equitable treatment in these vision tasks. We hope current and future results using our benchmark will contribute to fairer, more robust vision models. FACET is available publicly at* `https://facet.metademolab.com`.

1. Introduction

The ability of computer vision models to perform a wide range of tasks is due in no small part to large, widely used datasets. These large-scale datasets containing millions of images often have image-level labels such as ImageNet [16] or object-level annotations found in datasets such as MS-COCO [63] or Open Images [62]. Annotations are also used at the person-level, in datasets such as CelebA [64], UTK-Faces [98] and More Inclusive People Annotations (MIAP) [84]. These person-level annotations in particular enable a more fine-grained analysis and evaluation of model performance accross groups. Prior work using these person-level annotations to evaluate model fairness has shown that vision models learn societal biases and stereotypes, which negatively impact performance and cause downstream harms [87, 101, 74, 90, 88]. This makes fairness datasets particularly important as vision models continue to grow.

One weakness of existing fairness datasets is that they lack exhaustive and diverse demographic annotations that

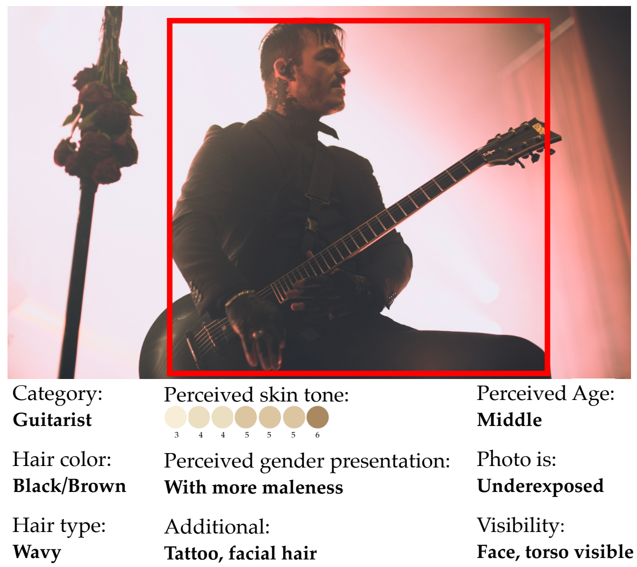

Figure 1: An example image and annotations from our dataset FACET . Every image in FACET contains annotations from expert reviewers on the primary class, sensitive attributes include *perceived gender presentation, perceived skin tone*, and *perceived age group* and additional visual attributes like hair color and type, visible tattoos, etc.

Size	– 32k images, 50k people
Evaluation Annotations	– 52-person related classes – bounding boxes around each person – person/hair/clothing labels for 69k masks
Protected Groups	– perceived skin tone – perceived age group – perceived gender presentation
Additional Person Attributes	– hair: color, hair type, facial hair – accessories: headscarf, facemask, hat – other: tattoo
Miscellaneous Attributes	lighting condition, level of occlusion

Table 1: Statistics on size of FACET and person annotations including labels for classification (e.g. soldier, teacher) and attributes such as hair color and perceived skin tone.

can support multiple vision tasks. For instance, while Open Images More Inclusive People Annotations (MIAP) [84] can be used for classification and detection, the labels are not particularly diverse as only perceived gender presentation and perceived age group are labeled. Image-level class labels are also sparse, with an incomplete set of true positives and true negatives per image. Another dataset, CelebA, contains many more person-level attributes but is primarily for face localization. In addition, CelebA contains many subjective and potentially harmful attributes e.g. attractive, big lips, chubby [25]. These weakenesses can greatly impact our ability to perform more fine-grained fairness analyses.

In this paper, we present FACET (Fairness in Computer Vision Evaluation Benchmark), a large-scale evaluation benchmark with exhaustive annotations for 32k images from Segment Anything 1 Billion (SA-1B) [59]labeled across 13 person attributes and 52 person classes. The 13 attributes include examples such as perceived skin tone, hair type, perceived age group; the 52 person classes include categories such as *hairdresser* and *reporter*. To ensure the annotations are both high quality and labeled by a diverse group of people, we used trained, expert annotators sourced from several geographic regions (North and Latin America, Middle East, Africa, East and Southeast Asia).

FACET enables a deeper analysis of potential fairness concerns and model biases for specific demographic axes. We can explore questions such as: 1) *Are models better at classifying people as skateboarder when their perceived gender presentation has more stereotypically male attributes?* 2) *Are open-vocabulary detection models better at detecting backpackers who are perceived to be younger?* 3) *Do standard detection models struggle to detect people whose skin appears darker?* 4) *Are these problems magnified when the person has coily hair compared to straight hair?* 5) *Do performance discrepancies differ across the detection and segmentation tasks?* These questions illustrate a few examples of how model biases can be explored at a deep, intersectional level using the exhaustive annotations in FACET. We use FACET to evaluate multiple state-of-the-art vision models to understand their fairness on demographic attributes (perceived gender presentation, perceived skin tone, perceived age group) as well as their existing demographic biases. FACET is publicly available at `https://facet.metademolab.com`.

Our contributions include:

– our new publicly available fairness benchmark FACET, containing 32k images from Segment Anything 1 Billion (SA-1B) [59], manually annotated with demographic and additional visual attributes labels by expert annotators
– 52 person-related class labels and manually drawn bounding boxes for every annotated person in every image (50k total people)
– person, clothing, or hair labels for 69k masks
– a benchmark for using FACET to compare different models, showing quantitative results and qualitative analyses on existing vision models using FACET

FACET is an evaluation-only benchmark. Using any of the annotations for training is strictly prohibited.

2. Related Work

Vision datasets that are annotated with apparent or self-reported demographic attributes are frequently used for studying model fairness. Table 2 compares FACET to other annotated datasets.

Classification Datasets such as [64, 98, 8, 57] are used to evaluate the gender and skin tone disparities in face recognition[1]. Gender Shades [8], for instance, showed that gender classification systems perform significantly worse on females compared to males and on darker skin compared to lighter skin using labels from annotators. These datasets cannot be used for tasks outside of facial recognition, e.g. object detection or image classification. Casual Conversations [43] is a dataset used for videos; this dataset was used to highlight disparities across gender and skin tone for the Deep Fake Detection challenge [23]. Geographic and income-based disparities have been evaluated as well [15, 89, 37], most commonly with the DollarStreet dataset [1, 80].

Detection/Segmentation [101] generated gender annotations via captions for MS-COCO. [94] annotated a subsection of pedestrians in BDD100k [97] for the task of pedestrian detection and found higher performance for lighter skin tones. However, these demographic annotations are often noisy and either lack annotator training or lack annotators altogether and are instead approximated from captions.

[1]We retain the same language used in the original papers, which is based on gender labels of the datasets that were audited.

Dataset *Task*	*Dataset Size*					*Apparent or Self-Reported Attributes*						
	#/people	#/images	#/videos	#/boxes	#/masks	gender	age	skin tone	race	lighting	additional	Tasks
UTK-Face[98]	20k	20k	–	–	–	Yes	Yes	No	Yes	No	No	–
FairFace[57]	108k	108k	–	–	–	Yes	Yes	No	Yes	No	No	–
Gender Shades[8]	1.2k	1.2k	–	–	–	Yes	Yes	Yes	No	No	No	–
OpenImages MIAP[84]	454k	100k	–	454k	*	Yes	Yes	No	No	No	No	C*DS*
[94] annotations for BDDK 100k [97]	16k	2.2k	–	16k	*	No	No	Yes	No	Yes	No	DS*
[100] annotations for COCO [63]	28k	16k	–	28k	28k	Yes	No	Yes	No	No	No	C*DS
Casual Conversations v1[43]	3k	N/A	45k	–	–	Yes	Yes	Yes	No	Yes	Yes	–
Casual Conversations v2 [42]	5.6k	N/A	26k	–	–	Yes	Yes	Yes	No	Yes	Yes	–
Ours – FACET	50k	32k	–	50k	69k	Yes	Yes	Yes	No	Yes	Yes	CDS

* represents tasks/annotations that are not included in the fairness portion of the dataset, but are included in the overall dataset.
e.g COCO has been used for multi-class classification [101, 92]

Table 2: Tasks and attribute annotations comparing existing datasets to FACET. These existing datasets were designed for fairness evaluations for other use cases, which is not to suggest that they are limited in use. The *tasks* (CDS) considered are **C**lassification of an image, **D**etection of a person or person-related objects, **S**egmentation of a person or person-related objects. For classification, we do not include the classification task of classifying protected or non-protected attributes of a person. For attributes, FACET does not include race as it is not a visually salient category, exacerbates bias [58] and misclassification has been shown to cause emotional distress [10]. Bounding boxes are denoted as bboxes.

More Inclusive Annotations for People (MIAP) [84], which is a subset of Open Images [62], is dataset that does focus on high quality, more complete person-level demographic annotations for bounding boxes. While MIAP is similar to FACET, it only has annotations for perceived gender presentation and perceived age group. FACET has far more exhaustive annotations spanning far more attributes. We have 13 attributes, including demographic attributes (perceived gender presentation, perceived age group, and perceived skin tone) as well as additional attributes such image quality (lighting and occlusion) and physical presentation (e.g. hair type, accessories, tattoos, etc.).

Best Practices Audits of popular computer vision datasets have found gender artifacts [66], a lack of geographic diversity [85], malignant stereotypes and NSFW content [12, 74, 6]. To combat these issues, there has been significant research about dataset development including tools [91], best practices for creating datasets [61, 82, 3, 52, 73, 42, 54, 81, 19, 83] and designing annotation tasks with crowdworkers [17, 14]. A large body of work also explored how researchers should document the intended use and considerations made when developing models [69], datasets [32, 76, 47, 67] and crowdsourced annotation tasks [21].

3. Benchmark Method

The goal of our benchmark is to evaluate and analyze how vision models perform across different demographic and additional attributes for different categories of people. This analysis requires (1) images that contain people with a diverse set of attributes and (2) images that contain people matching a variety of person-related categories. We focus on person-related categories such as occupations of people or person-related pasttimes, e.g. *doctor, basketball player, student, backpacker*, etc.). We prioritized a diverse set of categories for a more thorough analysis.

To generate the list of person-related categories, we use WordNet [68], which is a hierarchical database of language concepts. Each language concept is a single node in the hierarchy. For instance, the concept *apple* is a node with parent *edible fruit*. We take the language concept `person` and treat all of its children as potential categories, following [96]. We filter out offensive synsets noted in [96] and [12], generating 1,239 candidate synsets. We trim this list to 52 categories using the considerations below. Figure 2 shows a sample WordNet tree structure for two classes in FACET, and Figure 8 in Appendix A.2 shows the full hierarchy.

Connection to ImageNet-21k (IN21k) To ensure consistency with existing concepts used for computer vision model evaluation, we require our categories to overlap with the taxonomy of the widely used ImageNet-21k (IN21k) dataset [79]. This approach has been used to select object classes by other datasets [5] and follows previous work [34, 86]. This means models trained with IN21k can be evaluated out-of-the-box on FACET. IN21k is a long-tailed dataset, meaning many classes have very few images. We exclude categories with < 500 examples similar to [79] to ensure that models pre-trained on IN21k will transfer seamlessly to our evaluation set.

Concept Selection IN21k has overlapping classes with varying levels of specificity (i.e. `surgeon` is a subcategory of `doctor`). Following [49], we include classes with roughly the same "basic level." Using their findings of relative feature importance for classifying "basic level", we limit the depth in the WordNet hierarchy from the `person` node to 5, as a proxy for level of specificity. To alleviate ambiguity, we focus primarily on occupation/trade, sports, art, and leisure related categories of people. This leaves us with 52 categories. Our final list of concepts is shown in Table 17 in the Appendix.

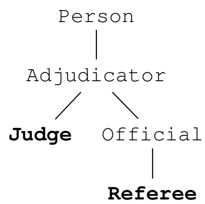

Figure 2: Example of WordNet tree structure relating the FACET classes `referee` and `judge` to the Person node.

4. Attribute Selection: Demographic and Additional Visual Attributes

FACET includes both demographic attributes and additional visual attributes. These exhaustively labeled, manually annotated attributes for all images in the dataset allows for evaluation of model performance and robustness at a fine-grained level. For example, we can investigate potential biases associated with a single attribute as well as at the intersection of multiple attributes. *Intersectionality* is a framework exploring how multiple attributes can actually magnify societal biases [13]; the exhaustive attributes in FACET means we can explore intersectional harms with respect to model fairness as well. Examples questions we can explore include "Do models struggle to classify people with tattoos?" (single attribute) and "Do models perform better for people with curly hair who appear to have perceived lighter skin tones than to those with perceived darker skin tones?" (intersection of attributes). See Appendix A.4 for the full list of attributes and their distributions.

4.1. Demographic Attributes

Perceived Skin Tone The Monk Skin Tone Scale [70], shown in Figure 3, was developed specifically for the computer vision use case. We intentionally use the Monk Skin Tone scale over the Fitzpatrick skin type [29], which was developed as means for determining one's likelihood of getting sunburn and lacks variance in darker skin tones [51, 72]. Fitzpatrick skin type has been shown to be unreliable for image annotation [36].

Skin tone as a spectrum. Skin tone is difficult to annotate[2] and can vary based on the lighting of the photo [55]. For this reason, we annotate skin tone as a spectrum. We gather annotations from three different annotators, allowing annotators to select as many skin tone values they feel best represent the person. This gives us a distribution over various skin tones. *We note that perceived skin tone is not a proxy for race or ethnicity, and should not be used as such.*

[2]Studies show even annotating one's own skin tone is difficult [27].

Figure 3: Monk Skin Tone Scale[70], an inclusive scale that includes 10 different skin tones.

Perceived Gender Presentation Annotators select whether they perceive a person as having more stereotypically female attributes, having more stereotypically male attributes, or having attributes outside of the gender binary. We annotate perceived gender presentation instead of gender, as gender cannot be determined purely from an image; attempting to do so can be harmful to groups who are misgendered [40]. A more thorough discussion is in Section 6.4.

Perceived Age Group We have three distinct perceived age group groups – *younger* are people perceived to be under 25 years old; *middle* are people perceived to be between 25-65 years old and *older* are people perceived to be over 65 years old. This follows the recommendations of [42] which matches the United Nation's breakdown of age [2], but we collapse `Adults 25-40` and `Middle-age Adults 41-65` into one category *middle*.

While it is impossible to tell a person's true age from an image, these numerical ranges are a rough guideline to delineate each perceived age group .

4.2. Additional Attributes

Hair color and hair type Because conv-nets are shown to recognize patterns and textures [30, 31, 33, 7] and hair types represent a range of different textures, we annotate the hair color and hair type.

Perceived lighting. Annotators labeled the lighting condition *on the person*. This annotation is important in part because it heavily impacts perceived skin tone [55]. These annotations can also guide the difficulty of the classification/detection problem, as models have been shown to have robustness vulnerabilities with respect to brightness [45, 53].

Additional attributes. We also annotate additional items relating to a person's appearance, using the recommendations of [42]. We condense the recommendations to the following list. These are **facial hair, head scarf** [3]**, hat/cap, eyewear (eyeglasses/sunglasses), face masks** [4]**, tattoos and a person's visibility** .

[3]The motivation for this annotation is from a finding of [86] that the concept *hijab* is predicted far more frequently for images with perceived lighter skin tones in UTK-Faces [98] than for those with perceived darker skin tones. It is unknown if this is a source of bias, as it is unknown whether or not there was a *hijab* in the photo.

[4]Many images in FACET include more face masks than prior works, such as ImageNet, due to the COVID-19 pandemic.

5. Annotation Method

5.1. Data Preprocessing

FACET is composed of images from Segment Anything 1 Billion (SA-1B). We preprocessed the dataset to create a rough pool of relevant examples (with a focus on high recall) before beginning the annotation process. We use caption and tags for each image to create a candidate pool of images to annotate. First, for each of the 52 selected categories, we created a list of related terms.

As each category corresponds to a WordNet synset, we use the lemmas for each synset to generate the related terms per concept. For categories with low frequency in the dataset, we supplement the list with related nouns (ie *flute* when looking for examples of *flutist*).

Separately, for each example (image, caption, tag), we tokenize and lemmatize the caption using the Natural Language Toolkit (NLTK) [65]. For instance, the caption "The person is playing the flute" gets lemmatized to {person, play, flute} (without stop words). We compute the overlap between the caption's lemmas + tags with the relevant term lists for each of the 52 categories to approximate which categories likely occur in each image. We select images with the most overlap for annotation.

5.2. Annotation Pipeline

Given the sensitivity of these labels, we took multiple steps to ensure high-quality annotations. Annotators completed stage-specific training before beginning labeling and perceived skin tone annotations underwent multi-review. Annotators could also mark *cannot be determined* for any image where they could not perceive the attribute. See Figure 6 in Appendix A.2.

Stage 1: *Filtering Images by Target Categories* First, annotators are tasked with quickly filtering images based on whether they contain people who belong to a subset of categories. Following the process described earlier in this section, we use the metadata for each image to create a shorter list of likely categories per photo. We do not use any classification or detection models to filter images to avoid potential model biases that can skew the data distribution. Annotators note the number of people in each image who match the specified categories. We exclude images marked with more than 5 people matching the target categories, given the time-intensive nature of annotating attributes for each person. This stage eliminates roughly 80% of the candidate images. For the remaining stages, we move from quick filtering to a focus on precision.

Stage 2: *Annotating Bounding Boxes* Annotators are tasked with drawing bounding boxes around each person in the image that matches **any** of the target 52 categories. For each bounding box, annotators mark a primary class, as well as a secondary class if necessary. The primary and secondary class structure alleviates potential overlap between categories. For example, a person playing the guitar and singing can match the category labels `guitarist` and `singer`. Furthermore, allowing two classes permits for representation of visually ambiguous classes, *e.g. a person in scrubs who could be a doctor or nurse*.

Stage 3: *Annotating perceived skin tone* We assign perceived skin tone annotations to its own step separate from other attributes to allow us to aggregate annotations from multiple raters. We choose to aggregate as one's own skin tone can affect the perceived skin tone of others [46, 28]. In this stage, annotators label the perceived skin tone of a specified person using the Monk Skin Tone Scale [70] (see Figure 3). We ask annotators to select at least 2 adjacent values and aggregate the results across three annotators. We report the number of times each skin tone was chosen.

Stage 4: *Annotating Remaining Attributes* In the final stage, annotators label the remaining attributes (see Section 4.2) for each person in the bounding boxes from Stage 2.

Stage 5: *Annotating SA-1BMasks* As FACET images come from SA-1B, which has images and masks, we label a subset of masks as `person, clothing, hair`. We do not collect exhaustive annotations for person-related masks in FACET ; we focus on annotating masks for people who are fully visible, with an attempt to balance demographic attributes. More details are given in Appendix A.2.1.

5.3. Annotator Diversity

We prioritized having a geographically diverse set of annotators following [56, 17] and sourced raters from varying regions to increase the diversity of annotations. Our annotators come from six different geographic regions to increase the diversity of the annotations, with one country per region. These regions (with country in parenthesis) include North America (United States), Latin American (Colombia), Middle East (Egypt), Africa (Kenya), Southeast Asia (Philippines) and East Asia (Taiwan). We show a more fine-grained breakdowns of annotators per region in Figure 4. We aimed for a roughly balanced number of raters per region but had disproportionate pass-rates of training across the various regions. We further describe our annotation process and annotators sourced in Appendix A.3, and answer the questions posed by CrowdWorkSheets [21] in Appendix C.

5.4. FACET Statistics

In this section we summarize the attribute and image breakdown of FACET. Table 3 shows the three demographic

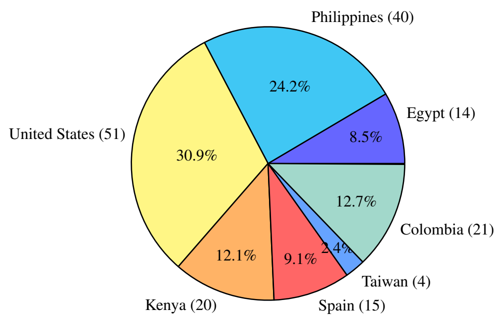

Figure 4: Breakdown of raters who passing training by percentage and by raw number, shown in parentheses.

Perceived or Apparent Attributes	*#/people*	%	*#/images*	%
gender presentation				
– more stereotypically F	10k	21%	8k	26%
– more stereotypically M	33k	67%	23k	72%
– non-binary	95	<1%	95	<1%
– unknown	6k	11%	5k	5%
Monk Skin Tone				
– 1	5k	10%	4k	13%
– 2	20k	41%	15k	48%
– 3	26k	53%	19k	61%
– 4	27k	54%	20k	63%
– 5	22k	44%	17k	54%
– 6	16k	33%	13k	40%
– 7	9k	18%	7k	23%
– 8	5k	10%	4k	13%
– 9	3k	6%	2k	7%
– 10	1k	3%	1k	3%
– unknown	18k	37%	13k	42%
age				
– younger	9k	18%	7k	23%
– middle	27k	55%	20k	64%
– older	3k	5%	2k	8%
– unknown	10k	21%	9k	27%

*Images can have multiple labels for each attribute, which is why numbers may not sum to 100%. F=femaleness; M=maleness

Table 3: Breakdown of representation of the demographic groups in the evaluation set.

groups and their corresponding number of occurrences in the evaluation set. The majority of perceived gender presentation annotations are people perceived to have more stereotypically male attributes, followed by people perceived as having more stereotypically female attributes. Perceived skin tone annotations essentially follow a normal distribution; the majority of annotations are in the range of skin tones 3-6. Appendix A.4 details more statistics about the FACET benchmark including the number of people per class and demographic attribute along with the frequency of additional attributes. Figure 5 shows an approximate

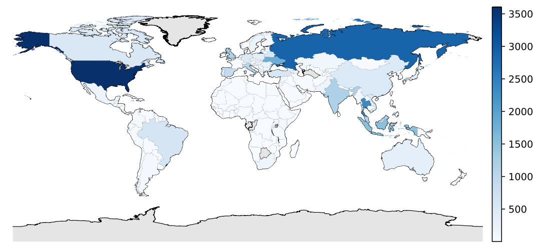

Figure 5: Approximate geographic distribution of the images in FACET.

geographic breakdown of the images in FACET. The geographic information was inferred from locations mentioned in the captions, so the distribution is approximate.

6. Fairness Evaluations using FACET

We use FACET to evaluate fairness by measuring performance disparities across different attributes for a selection of state-of-the-art vision models. Given a model f, a performance metric *recall*, a set of concepts $\mathcal{C}$, an attribute label l and a set of images $\mathcal{I}_l^C$, we compute:

$$\begin{aligned} disparity = recall(f(l_1, \mathcal{I}_{l_1}^C, \mathcal{C})) \\ - recall(f(l_2, \mathcal{I}_{l_2}^C, \mathcal{C})) \end{aligned} \quad (1)$$

As a concrete example, we can compute the disparity between people perceived as younger (l_1) versus people perceived as older (l_2) for the concept teacher ($\mathcal{C} = \{\text{teacher}\}$). Images $\mathcal{I}_{l_1}^C$ and $\mathcal{I}_{l_2}^C$ are images of teacher who are perceived to be younger and images of teachers who are perceived to be older, respectively. Disparity > 1 indicates the model performs better for images with label l_1 and disparity < 1 indicates the model performs better for label l_2. A perfectly fair model evaluated with FACET would have a disparity of 0, meaning it has the same performance across all images regardless of the associated attributes.

FACET is unique for two key reasons:

1. *Exhaustive attribute and class level evaluation:* FACET's annotations are exhaustive, meaning every person who matches a class in every image is annotated across all attributes through a rigorous annotation process. Datasets that include only the `person` class label and/or very sparse attribute labels risk bias leakage from the unlabeled people or difficulty performing a deep analysis due to the lack of exhaustive labels. These are not concerns with FACET.
2. *Annotations for multiple vision tasks:* Because every image is annotated with bounding boxes and person-related classes, multiple vision tasks can be evaluated and analyzed alongside the exhaustively annotated person attributes.

	Difference in Average Recalls		
Person Class	$M - F$	Person Class	$F - M$
gardener	16.4	dancer	21.7
craftsman	13.6	retailer	17.0
laborer	10.3	reporter	16.0
skateboarder	8.8	nurse	12.9
prayer	8.8	student	12.8
waiter	8.3	gymnast	8.5
speaker	5.4	painter	6.1
guitarist	4.0	hairdresser	5.2
singer	1.6	climber	5.1
lawman	1.4	horseman	4.5

Table 4: CLIP's performance disparity for the classes with the largest disparity across perceived gender presentation. The classes on the left indicate better performance for images with people who are perceived as having **more stereotypically male attributes**; results on the right indicate better performance for those perceived as having **more stereotypically female attributes**.

6.1. Classification

Are models better at classifying people as skateboarder when their perceived gender presentation has more stereotypically male attributes? To help answer this question and others like it, we evaluate standard image classification models with the FACET class labels. For classification evaluation, we only evaluate images with a single person. For images where a single person is labeled with multiple person classes, we treat both classes as valid labels. We evaluate classification using CLIP ViT-B/32 [77] in a zero-shot setting. The largest discrepancies for CLIP on the perceived gender presentation axis are shown in Table 4. Some of these classes parallel societal, gender-related biases (e.g. careers like nurses and hairdresser for those who are perceived with more femaleness [35, 71]). We show further analysis of CLIP across other demographic groups in Appendix A.5. We also show how we can use the FACET IN21k class overlap to evaluate an ImageNet21k pre-trained ViT [26].

6.2. Person Detection & Segmentation

6.2.1 Person Detection

We evaluate a Faster R-CNN model with a ResNet-50-FPN backbone [78] pretrained on COCO. During evaluation, we only keep the predicted boxes corresponding to the COCO `person` class. We treat the remaining boxes as class-agnostic, and we compute the average recall (AR) and mAR (mean average recall) metrics proposed in [50] with all predicted boxes and measure performance across the demographic attributes. For person detection, we focus on evaluating perceived skin tone and how model performance parallel societal biases [22, 60].

Monk Skin Tone (MST)	mAR	$AR_{0.5}$	$AR_{0.75}$
– 1	75.5	98.4	85.0
– 2	75.0	98.3	84.0
– 3	74.7	98.3	83.5
– 4	74.4	98.1	83.0
– 5	74.1	98.2	82.6
– 6	73.9	98.3	82.5
– 7	73.7	98.2	82.2
– 8	73.7	98.0	82.5
– 9	73.3	97.3	81.1
– 10	72.6	96.5	80.4

Average Recall (AR), with IoU values as subscripts. mAR is averaged across IoUs from 0.5 to 0.95, in increments of 0.05.

Table 5: Average recall (AR) on FACET for a ResNet-50 Faster R-CNN. The model has the best performance for MST=1, which is the lightest skin tone, and the lowest performance for MST=9 and 10, are the darkest skin tones. The largest disparity is for $AR_{0.75}$.

Do standard detection models struggle to detect people whose skin appears darker? We compute the AR across the predicted bounding boxes for perceived skin tone , as shown in Table 5. At *every* IoU for the ARs, the darkest perceived skin tone has the lowest performance. The gap between the highest and lowest performance is over 4 points at $AR_{0.75}$, which suggests Faster R-CNN does struggle more on precisely detecting those perceived with darker skin tones.

Does this problem magnify when, for instance, the person has coily compared to straight hair? We dig deeper into the previous results to investigate intersections of attributes. In Table 6 we measure mAR per hair type for people with the three lightest perceived skin tone versus the three darkest perceived skin tone. This unearths several concerning findings: for 456 hair types performance is higher for the lighter skin tones than the darker skin tones. These are fairly significant gaps; we over a 10 point different for hair type *dreads* at $AR_{0.75}$, for instance. This is a particularly interesting finding for two reasons. First, we see a 50x increase in the disparity for *dreads* across perceived skin tone from $AR_{0.5}$ to $AR_{0.75}$. This suggests Faster R-CNN can detect people with dreadlocks, but struggles to perform accurate localization as shown by the larger gap as the IoU threshold for AR increases. Second, dreadlocks are often associated with darker skin and a plethora of associated stereotypes [24, 9, 75, 4]). This means the likely association because *dreads* and darker skin tones in the training data interestingly combat this performance disparity.

lighter = {1, 2, 3} *darker* = {8, 9, 10}

Hair ↓ / Skin →	mAR lighter	mAR darker	$AR_{0.5}$ lighter	$AR_{0.5}$ darker	$AR_{0.75}$ lighter	$AR_{0.75}$ darker
coily	**76.7**	73.4	98.2	**98.5**	**87.3**	80.9
dreads	**77.1**	74.7	97.9	**98.1**	**94.8**	85.7
bald	**78.1**	71.5	**99.0**	96.7	**87.8**	77.5
straight	75.6	**76.1**	98.4	**99.1**	84.8	**85.6**
curly	**75.0**	74.8	98.5	**99.2**	**84.7**	83.7
wavy	**76.1**	75.8	98.6	**99.1**	**85.5**	84.8

Average Recall (AR), with IoU values as subscripts. mAR is averaged across IoUs from 0.5 to 0.95, in increments of 0.05.

Table 6: Average recall (AR) on FACET for a ResNet-50 Faster R-CNN. We show performance for the intersection of hair type and perceived skin tone. Performance is higher for lighter perceived skin tone for every hair type except straight and wavy.

6.2.2 Person Instance Segmentation

We evaluate a MaskR-CNN model with a ResNet-50-FPN backbone [44] pretrained on COCO. In the same pattern as for person detection, we only keep the predicted masks corresponding to the COCO `person` class, and compute AR in a class agnostic way. We use the IoU between predicted and ground truth mask for instance segmentation, opposed to between boxes used for detection.

Do performance discrepancies differ across the detection and segmentation task? We compare potential discrepancies across segmentation and detection of people. We evaluate MaskR-CNN for person detection and person segmentation separately. For consistency, we limit the evaluation for both detection and segmentation to the set of people who have a mask annotation. We compare the patterns of discrepancies in AR across perceived gender presentation for person detection and segmentation, as shown in Table 7. We notice that for both detection and segmentation, the performance disparities are largest at $AR_{0.75}$. We also observe slightly larger gaps in performance for detection compared to segmentation. In line with prior work [44], we find higher AR for person detection than instance segmentation. We describe the experimental setup in more detail in Appendix A.5.

perceived gender presentation	mAR *box*	mAR *mask*	$AR_{0.5}$ *box*	$AR_{0.5}$ *mask*	$AR_{0.75}$ *box*	$AR_{0.75}$ *mask*
– more stereotypically male attributes	78.3	72.2	99.3	98.1	88.0	84.6
– more stereotypically female attributes	75.6	70.8	99.0	97.5	84.7	82.9
– outside of gender binary	77.0	63.0	98.0	92.0	88.0	74.0

Average Recall (AR), with IoU values as subscripts. mAR is averaged across IoUs from 0.5 to 0.95, in increments of 0.05.

Table 7: We compare the AR on FACET for a ResNet-50 MaskR-CNN across the person detection and person instance segmentation tasks. The candidates box dictates the AR for person detection, box proposals, and $mask$ for segmentation, mask proposals.

6.3. Open World Detection & Visual Grounding

6.3.1 Open Vocabulary Detection

Next we evaluate open vocabulary detection using Detic [102]. We describe the experimental setup in detail in the Appendix A.5. For Detic, we focus on perceived age group.

Are open-vocabulary detection models better at detecting backpackers who are perceived to be younger? To be illustrative of disparities observed with FACET, we selected the three person related classes with the biggest disparity between groups: *trumpeter, backpacker and painter*. Detic exhibits perceived age group-based performance disparities for all 3 categories. The disparities are large, with an 15 point gap in mAR for the *backpacker* class. The disparities are also consistent across AR measurements for a specific class. Unlike for Faster R-CNN with perceived skin tone, we typically observe larger gaps for $AR_{0.5}$ than $AR_{0.75}$, suggesting that in this case there is perhaps more of a discrepancy in the classification of an object as in the scene as there is with the precision of the bounding box. We show all per-class disparities as well as the mean disparities across all 52 classes in Table 15 in Appendix A.5.

6.3.2 Visual Grounding

Lastly, we evaluate the visual grounding using OFA [93], a sequence-to-sequence vision-language model. We evaluate perceived age group disparities using the three person classes with the largest disparities which are nurse, gardener and guitarist. Results are in Table 15. OFA's largest disparity is nearly 27 points, observed in the nurse class. The disparities are also consistent across all IoU values; for every class, the best performing perceived age group label is consistent across all IoU values. We show the full table of per-class disparities as well as the disparities averaged across all classes in Table 15 in Appendix A.5.

6.4. Limitations

As the development of datasets for fairness analysis becomes more common, approaches and recommendations for how to do so in ethical and safe ways are also being

Person Class	Detic (detection) mAR	Detic $AR_{0.5}$	Detic $AR_{0.75}$	OFA (visual grounding) mAR	OFA $AR_{0.5}$	OFA $AR_{0.75}$
	backpacker			*gardener*		
– young	**45.4**	**55.3**	**47.4**	11.9	32.6	7.0
– middle	42.1	51.7	44.6	18.3	40.1	14.6
– older	29.8	35.3	33.3	**27.9**	**58.1**	**24.2**
	trumpeter			*solider*		
– young	22.8	26.7	25.6	**16.3**	9.2	**40.0**
– middle	29.5	34.8	31.4	**16.3**	**14.5**	33.8
– older	**38.4**	**45.5**	**38.2**	1.3	6.3	0.0
	drummer			*guitarist*		
– young	19.9	24.9	19.7	19.8	38.7	18.5
– middle	26.3	34.7	27.6	19.6	35.5	20.6
– older	**34.2**	**41.8**	**35.8**	**32.0**	**56.7**	**36.7**

Average Recall (AR), with IoU values as subscripts. mAR is averaged across IoUs from 0.5 to 0.95, in increments of 0.05.

Table 8: Per-class performance for Detic and OFA on a subset of FACET classes. The perceived age group with the highest performance per class is bolded.

increasingly explored [42, 32, 3, 83]. While we strongly believe FACET will help practitioners better understand sources of bias in their model, we note that translating real world concepts and demographic groups to dataset annotations is inherently imperfect.

First, while self-identification of concept classes and person-related attributes is preferred [83, 99], our adaptation of an existing dataset requires external annotations. To reduce these potential biases, we use highly trained annotators and avoid automated labeling methods like adapting existing captions, alt-text or model classifications [101, 92, 99, 39, 38, 84, 8]. Second, while generating sets of labels for each attribute, there is a trade-off between having more labels (wider representation) and opting for fewer, higher frequency labels (more statistical significance) [38]. This was extensively considered, and we acknowledge that, as with any paper using discrete labels, our labels for perceived gender presentation and perceived age group risk erasure of genders and ages that are not identifiable in our categorization [20]. For skin tone, we follow the Monk Scale [70], which shows better inclusivity of darker skin tones. For concept classes, we map to ImageNet classes to encourage easy adoption and to ensure mutually exclusive classes. Lastly, FACET and other fairness datasets are representative of the current time period and organizational infrastructure within which it was created [18, 32]. To address how this affects annotations and insights when performing evaluations, we include in Appendix C our responses to the CrowdWorkSheets [21] for FACET.

7. Discussion

We have seen rapid growth and impressive performance gains in computer vision across a number of tasks such as classification, detection, segmentation and visual grounding. Simultaneously, these models have learned societal biases and can perpetuate these harmful stereotypes in downstream tasks. We present FACET, a vision fairness benchmark that contains 32k annotated images of 50k people. People in the images are exhaustively labeled with demographic attributes, including perceived gender presentation, perceived skin tone and perceived age group, and additional attributes such as hair type and light exposure. Labeling demographic attributes requires thoughtful design, so we hired expert annotators and prioritized clean annotations. In addition to these attributes, FACET also has manual annotations for bounding boxes and person-related classes. These person-related classes, such as hairdresser and farmer, overlap with the ImageNet-21K (IN21k) vocabulary, meaning vision models that can be evaluated on IN21k can also seamlessly use FACET. We aimed to be extremely conscious and respectful with our annotations, while also acknowledging that there are limitations with this and similar fairness datasets. We are publicly releasing FACET to encourage and lower the barrier to entry to evaluating vision models for potential biases. We propose several ways that researchers can use FACET to evaluate their models for potential fairness concerns across a variety of common vision tasks.

Acknowledgments We would like to thank Adina Williams, Priya Goyal, Zachary Winstrom, Austin Miller, Nicolas Carion, Nisha Deo, Emily Dinan, Ahuva Goldstand, Somya Jain, Alexander Kirillov, Mallika Malhotra, Hanzi Mao, Ishan Misra, Laurens van der Maaten, and Mary Williamson.

References

[1] Dollar Street. `https://www.gapminder.org/dollar-street`. Accessed: 2023-03-08.

[2] Provisional Guidelines on Standard International Age Classifications. `https://unstats.un.org/unsd/publication/SeriesM/SeriesM_74e.pdf`. Accessed: 2023-03-08.

[3] Jerone TA Andrews, Dora Zhao, William Thong, Apostolos Modas, Orestis Papakyriakopoulos, Shruti Nagpal, and Alice Xiang. Ethical considerations for collecting human-centric image datasets. arXiv preprint arXiv:2302.03629, 2023.

[4] Patricia A Banks. No dreadlocks allowed: Race, hairstyles, and cultural exclusion in schools. Hairstyles, and Cultural Exclusion in Schools (February 12, 2021), 2021.

[5] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems, 32, 2019.

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963, 2021.

[7] Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. arXiv preprint arXiv:1904.00760, 2019.

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

[9] Ayana Byrd and Lori L Tharps. When black hair is against the rules. The New York Times, 1, 2014.

[10] Mary E. Campbell and Lisa Troyer. The implications of racial misclassification by observers. American Sociological Review, 72(5):750–765, 2007.

[11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020.

[12] Kate Crawford and Trevor Paglen. Excavating ai: The politics of training sets for machine learning. https://excavating.ai/, September 2019.

[13] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. u. Chi. Legal f., page 139, 1989.

[14] Florian Daniel, Pavel Kucherbaev, Cinzia Cappiello, Boualem Benatallah, and Mohammad Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. ACM Computing Surveys (CSUR), 51(1):1–40, 2018.

[15] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 52–59, 2019.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.

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

[18] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. On the genealogy of machine learning datasets: A critical history of imagenet. Big Data & Society, 8(2):20539517211035955, 2021.

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

[20] Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "gender" in nlp bias research, 2022.

[21] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. CrowdWorkSheets: Accounting for individual and collective identities underlying crowdsourced dataset annotation. In 2022 ACM Conference on Fairness, Accountability, and Transparency. ACM, jun 2022.

[22] Travis L Dixon and Keith B Maddox. Skin tone, crime news, and social reality judgments: Priming the stereotype of the dark and dangerous black criminal 1. Journal of Applied Social Psychology, 35(8):1555–1570, 2005.

[23] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. The deepfake detection challenge dataset. CoRR, abs/2006.07397, 2020.

[24] Saran Donahoo and Asia D Smith. Controlling the crown: Legal efforts to professionalize black hair. Race and Justice, 12(1):182–203, 2022.

[25] Samuel Dooley, Ryan Downing, George Wei, Nathan Shankar, Bradon Thymes, Gudrun Thorkelsdottir, Tiye Kurtz-Miott, Rachel Mattson, Olufemi Obiwumi, Valeriia Cherepanova, et al. Comparing human and machine bias in face recognition. arXiv preprint arXiv:2110.08396, 2021.

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[27] Steven Eilers, Daniel Bach, Rikki Gaber, Hanz Blatt, Yanina Guevara, Katie Nitsche, Roopal Kundu, and June Robinson. Accuracy of self-report in assessing fitzpatrick skin phototypes i through vi. JAMA dermatology (Chicago, Ill.), 149, 09 2013.

[28] Cynthia Feliciano. Shades of race: How phenotype and observer characteristics shape racial classification. American Behavioral Scientist, 60(4):390–419, 2016.

[29] Thomas B Fitzpatrick. Soleil et peau. J. Med. Esthet., 2:33–34, 1975.

[30] Leon Gatys, Alexander S Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. Advances in neural information processing systems, 28, 2015.

[31] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Texture and art with deep neural networks. Current opinion in neurobiology, 46:178–186, 2017.

[32] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. Communications of the ACM, 64(12):86–92, 2021.

[33] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In International Conference on Learning Representations, 2019.

[34] Priya Goyal, Adriana Romero Soriano, Caner Hazirbas, Levent Sagun, and Nicolas Usunier. Fairness indicators for

systematic assessments of visual feature extractors. arXiv preprint arXiv:2202.07603, 2022.

[35] Anthony G Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology, 74 6:1464–80, 1998.

[36] Matthew Groh, Caleb Harris, Roxana Daneshjou, Omar Badri, and Arash Koochek. Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm. Proc. ACM Hum.-Comput. Interact., 6(CSCW2), nov 2022.

[37] Laura Gustafson, Megan Richards, Melissa Hall, Caner Hazirbas, Diane Bouchacourt, and Mark Ibrahim. Pinpointing why object recognition performance degrades across income levels and geographies. arXiv preprint arXiv:2304.05391, 2023.

[38] Melissa Hall, Bobbie Chern, Laura Gustafson, Denisse Ventura, Harshad Kulkarni, Candace Ross, and Nicolas Usunier. Towards reliable assessments of demographic disparities in multi-label image classifiers. arXiv preprint arXiv:2302.08572, 2023.

[39] Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-language models performing zero-shot tasks exhibit gender-based disparities. arXiv preprint arXiv:2301.11100, 2023.

[40] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[41] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. Advances in neural information processing systems, 29, 2016.

[42] Caner Hazirbas, Yejin Bang, Tiezheng Yu, Parisa Assar, Bilal Porgali, Vítor Albiero, Stefan Hermanek, Jacqueline Pan, Emily McReynolds, Miranda Bogen, Pascale Fung, and Cristian Canton Ferrer. Casual conversations v2: Designing a large consent-driven dataset to measure algorithmic bias and robustness, 2022.

[43] Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton-Ferrer. Towards measuring fairness in AI: the casual conversations dataset. CoRR, abs/2104.02821, 2021.

[44] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017.

[45] Dan Hendrycks and Thomas G Dietterich. Benchmarking neural network robustness to common corruptions and surface variations. arXiv preprint arXiv:1807.01697, 2018.

[46] Mark E. Hill. Race of the interviewer and perception of skin color: Evidence from the multi-city study of urban inequality. American Sociological Review, 67(1):99–108, 2002.

[47] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. arXiv preprint arXiv:1805.03677, 2018.

[48] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label. Data Protection and Privacy, 12(12):1, 2020.

[49] Laura Hollink, Aysenur Bilgin, and Jacco van Ossenbruggen. Is it a fruit, an apple or a granny smith? predicting the basic level in a concept hierarchy, 2019.

[50] Jan Hendrik Hosang, Rodrigo Benenson, Piotr Dollár, and Bernt Schiele. What makes for effective detection proposals? CoRR, abs/1502.05082, 2015.

[51] John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. Reliability and validity of image-based and self-reported skin phenotype metrics. CoRR, abs/2106.11240, 2021.

[52] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 560–575, 2021.

[53] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. arXiv preprint arXiv:2211.01866, 2022.

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

[55] Praveen Kakumanu, Sokratis Makrogiannis, and Nikolaos Bourbakis. A survey of skin-color modeling and detection methods. Pattern recognition, 40(3):1106–1122, 2007.

[56] Shivani Kapania, Ding Wang, and Alex Taylor. A hunt for the snark: Annotator diversity in data practices. 2023.

[57] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1548–1558, 2021.

[58] Zaid Khan and Yun Fu. 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, pages 587–597, 2021.

[59] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023.

[60] Heather M Kleider-Offutt, Alesha D Bond, and Shanna EA Hegerty. Black stereotypical features: When a face type can get you in trouble. Current Directions in Psychological Science, 26(1):28–33, 2017.

[61] Bernard Koch, Emily Denton, Alex Hanna, and Jacob Gates Foster. Reduced, reused and recycled: The life of a dataset in machine learning research. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

[62] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. International Journal of Computer Vision, 128(7):1956–1981, 2020.

[63] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014.

[64] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision, pages 3730–3738, 2015.

[65] Edward Loper and Steven Bird. Nltk: The natural language toolkit. arXiv preprint cs/0205028, 2002.

[66] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. arXiv preprint arXiv:2206.09191, 2022.

[67] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. Documenting computer vision datasets: an invitation to reflexive data practices. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 161–172, 2021.

[68] George A Miller. Wordnet: a lexical database for english. Communications of the ACM, 38(11):39–41, 1995.

[69] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency, pages 220–229, 2019.

[70] Ellis Monk. Monk skin tone scale, 2019.

[71] Brian A Nosek, Mahzarin R Banaji, and Anthony G Greenwald. Harvesting implicit group attitudes and beliefs from a demonstration web site. Group Dynamics: Theory, Research, and Practice, 6(1):101, 2002.

[72] U.K. Okoji, S.C. Taylor, and J.B. Lipoff. Equity in skin typing: why it is time to replace the Fitzpatrick scale. British Journal of Dermatology, 185(1):198–199, 07 2021.

[73] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns, 2(11):100336, 2021.

[74] VU Prabhu and A Birhane. Large datasets: A pyrrhic win for computer vision. arXiv preprint arXiv:2006.16923, 3, 2020.

[75] Heather M Prendergast. Case 10: West indian/caribbean. Diversity and Inclusion in Quality Patient Care, pages 253–258, 2016.

[76] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In 2022 ACM Conference on Fairness, Accountability, and Transparency, pages 1776–1826, 2022.

[77] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[78] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015.

[79] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.

[80] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022.

[81] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. 2021.

[82] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. Do datasets have politics? disciplinary values in computer vision dataset development. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2):1–37, 2021.

[83] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker. How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. Proc. ACM Hum.-Comput. Interact., 4(CSCW1), may 2020.

[84] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. A step toward more inclusive people annotations for fairness. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. ACM, jul 2021.

[85] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. arXiv preprint arXiv:1711.08536, 2017.

[86] Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 804–814, 2022.

[87] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pages 701–713, 2021.

[88] Pierre Stock and Moustapha Cissé. Convnets and imagenet beyond accuracy: Explanations, bias detection, adversar-

ial examples and model criticism. CoRR, abs/1711.11443, 2017.

[89] Sruthi Sudhakar, Viraj Prabhu, Olga Russakovsky, and Judy Hoffman. Icon ˆ2: Reliably benchmarking predictive inequity in object detection. arXiv preprint arXiv:2306.04482, 2023.

[90] Schrasing Tong and Lalana Kagal. Investigating bias in image classification using model explanations. arXiv preprint arXiv:2012.05463, 2020.

[91] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. International Journal of Computer Vision, 130(7):1790–1810, 2022.

[92] Angelina Wang and Olga Russakovsky. Directional bias amplification. CoRR, abs/2102.12594, 2021.

[93] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. arXiv preprint arXiv:2202.03052, 2022.

[94] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection, 2019.

[95] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pages 38–45, 2020.

[96] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. ACM, jan 2020.

[97] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[98] Song Yang Zhang, Zhifei and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.

[99] Dora Zhao, Jerone T. A. Andrews, and Alice Xiang. Men also do laundry: Multi-attribute bias amplification, 2022.

[100] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 14830–14840, 2021.

[101] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. arXiv preprint arXiv:1707.09457, 2017.

[102] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX, pages 350–368. Springer, 2022.