

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** (**FAirness in Computer Vision EvaluATion**), 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

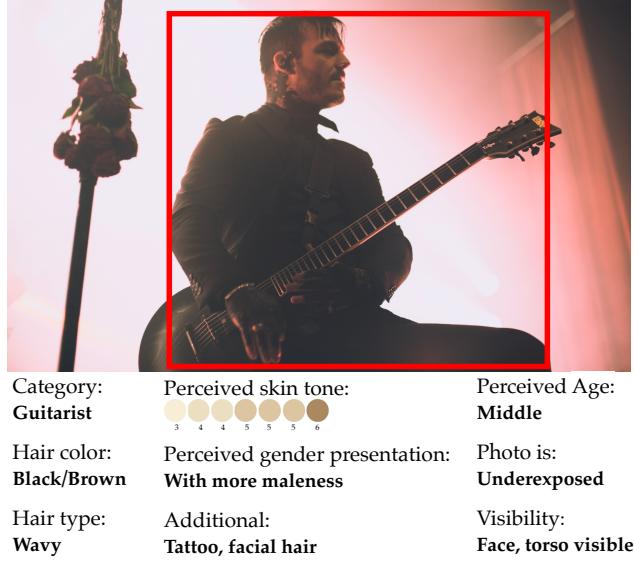


Figure 1: An example image and annotations from our dataset FACET. Every image in FACET contains annotations from expert reviewers on the primary class, sensitive attributes including *perceived gender presentation*, *perceived skin tone*, and *perceived age group*, alongside additional visual attributes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

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 across 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

<i>Size</i>	– 32k images, 50k people
<i>Evaluation Annotations</i>	<ul style="list-style-type: none"> – 52-person related classes – bounding boxes around each person – person/hair/clothing labels for 69k masks
<i>Protected Groups</i>	<ul style="list-style-type: none"> – perceived skin tone – perceived age group – perceived gender presentation
<i>Additional Person Attributes</i>	<ul style="list-style-type: none"> – hair: color, hair type, facial hair – accessories: headscarf, face mask, hat – other: tattoo
<i>Miscellaneous Attributes</i>	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 weaknesses 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 illus-

trate 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 recognition¹. 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.

¹We 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 boxes.

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 past-times, 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.

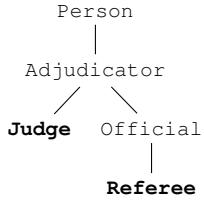


Figure 2: Example of WordNet tree structure relating the FACET classes `referee` and `judge` to the `Person` node.

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 allow 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 annotate² 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.*

²Studies 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**.

³The 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.

⁴Many 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-1B Masks* As FACET images come from Segment Anything 1 Billion (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 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

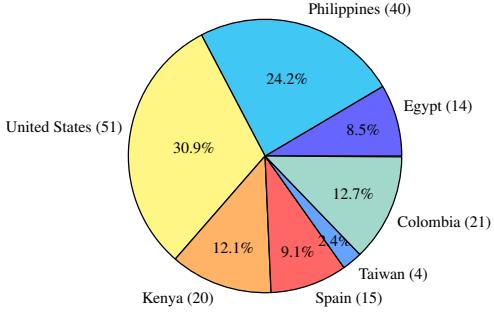


Figure 4: Breakdown of raters who passed training by percentage and by raw number, shown in parentheses.

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

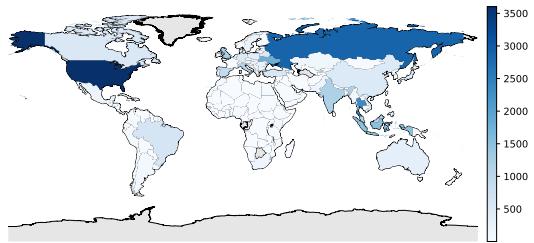


Figure 5: Approximate geographic distribution of the images in FACET.

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:

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

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 teachers 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. higher performance on careers like nurses and hairdressers for those who are perceived with more femininity [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 parallels 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, which are the darkest skin tones. The largest disparity between skin tones 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; for instance, we see a nearly 10 point difference for hair type *dreads* at AR_{0.75}. 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 between *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		AR _{0.5}		AR _{0.75}	
	lighter	darker	lighter	darker	lighter	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 Mask R-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 masks 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 Mask R-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.

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.

<i>perceived gender presentation</i>	mAR		AR _{0.5}		AR _{0.75}	
	<i>box</i>	<i>mask</i>	<i>box</i>	<i>mask</i>	<i>box</i>	<i>mask</i>
– 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 Mask R-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.

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 a 15 point gap in mAR for the *backpacker* class. The disparities are also consistent across AR measurements for a specific class. Unlike what we observed for Faster R-CNN with perceived skin tone, we typically observe larger gaps for AR_{0.5} than AR_{0.75}. This suggest that in open-world detection, there is perhaps more of a discrepancy in the classification/identifying a person for a given category than there is in 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 visual grounding using OFA [93], a sequence-to-sequence vision-language model. We evaluate perceived age group disparities using three person classes with large 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 and best performing perceived age group label for every class 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 being increas-

Person Class	Detic (detection)			OFA (visual grounding)		
	mAR	AR _{0.5}	AR _{0.75}	mAR	AR _{0.5}	AR _{0.75}
<i>backpacker</i>						
- 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
<i>trumpeter</i>						
- 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
<i>drummer</i>						
- 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.

ingly 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, Valeria 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 Balestrieri, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdza, 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. Okojo, 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 2: 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 Chau- mond, Clement Delangue, Anthony Moi, Pierrick 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, Jerome 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.

A. Appendix

A.1. Category Selection

Imageability and Salience of the Categories When choosing the categories for FACET, we considered the “imageability” of our concepts, from [96]. However, we found that this did not transfer well to our use case. First, we found that many of the ‘highly imageable’ concepts include classes directly related to a demographic attribute. (For instance, *black woman* (*n09637339*) has an imageability score of 5 out of 5.) Additionally, many highly imageable concepts are abstract, meaning they are easy to imagine but hard to classify. As an example, it is easy to imagine what the concept *mother* may look like, but it is hard to determine if someone is a “mother” from a photo. (*Is any person perceived with having stereotypically feminine attributes with a child in a photo presumed to be a mother?*)

Class Hierarchy and Representation We show the full connection of our chosen concepts in WordNet in their relation to the Person synset. Figure 8 shows the full connection of our chosen concepts in WordNet in their relation to the Person synset. All relevant sub-trees and intermediate synsets are shown. We can see that many of the classes in FACET share the same parent node. We also note that no class in FACET is a direct descendant of another class. This demonstrates that there is no overlap between classes. Table 17 shows the representation of each class in the evaluation set.

A.2. Annotation Pipeline

We describe in more detail the annotation pipeline we use for FACET.

A.2.1 Annotation Pipeline Design

Preprocessing Figure 6 shows the pre-processing steps of the captions to create the candidate set of images to annotate. We ‘score’ each caption for each category based on the overlap of relevant words for the category and caption. We sample captions with the highest ‘score’ per category. We choose the candidate images for FACET from a set of roughly 6 million images.

We select a starting set of images for annotation such that we expect the portion of images that pass stage 1 to be roughly class balanced. To approximate the probability that images with overlap per category are true positives, we sample 50 images per category and annotate the true positives. We use this frequency to determine how much to over sample a specific category. As we continue the annotation process, for additional rounds, we sample images with overlap based on the categories that are under-represented in the dataset thus far. We note that many categories did not have

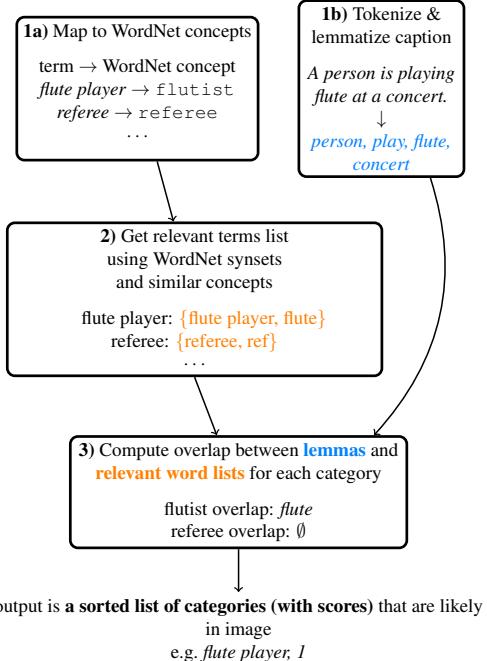


Figure 6: Label annotation pipeline: The preprocessing steps before beginning the annotation pipeline. **In 1a)** we map all of the person-related classes to concepts in WordNet. We denote WordNet concepts in a different font. (See Section 3 for a full description on WordNet concepts and synsets). **In 1b)** we tokenize and lemmatize the captions to produce a list of **lemmas**. **In 2)** for each of the 52 categories, we generate a **relevant words list** using WordNet synsets. Lastly **in 3)** we compute overlap between the **lemmas** and **relevant words list** and select images to annotate which have high overlap.

enough images with matching relevant words and as such we did not achieve equal representation of all categories.

Annotation Stages Figure 7 shows the four separate annotation tasks of the main annotation pipeline. Breaking the annotation process into multiple sub-tasks allows for more fine-grained control. For Stage 1, we focus on speed, and ask annotators to spend little time per task. To increase speed, we group multiple images with the same target categories into a single task with a default value of *0 people match the categories*, and ask the annotator to label each image. We separate Stages 3 and 4, so that we can gather multiple annotations for apparent skin tone only. We separate these stages from Stage 2 to simplify the task for annotators, such that they only need consider the perceived demographic attributes for one person at a time. Additionally, this allows the annotators in later stages to quality assure the annotations from earlier stages, as described in Section A.3.2.

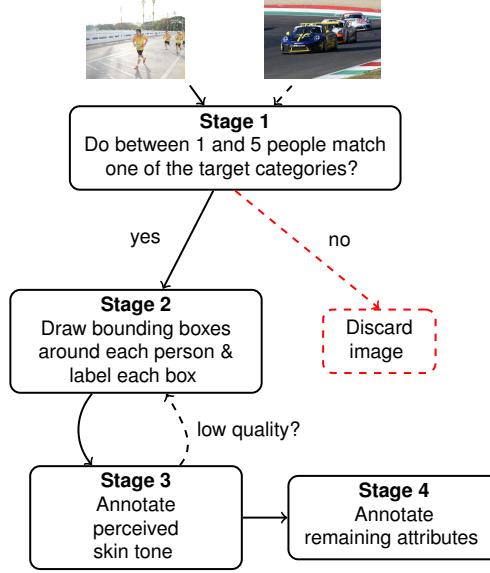


Figure 7: Image annotation pipeline: The four stages of the main annotation pipeline. The image on the left can be fully annotated; the image on the right does not contain the target categories and gets excluded after Stage 1. Dashed lines show paths that do not advance to the next stage.

Mask Annotations We collect labels for SA-1B mask annotations separately after completing the annotation pipeline. First, we select candidate people from FACET with attempts made to balance the number of people per demographic group. Next, we select a candidate set of masks to annotate by collecting the set of masks inside the bounding boxes for these people. For each mask and FACET bounding box in which it resides, we asked annotators if the mask corresponds to *the person’s body*, *the person’s hair* or *an item of clothing on the person* for a given person, as denoted by a bounding box. Annotators did not make any modifications to the masks, e.g. change the shape. Annotators were told to only select a class if the mask covers the entire item; masks for a portion of the person, or part of item of clothing were not labeled. Additionally, annotators were only told to select a class if it met the label for the person described by the bounding box; masks for people, hair, clothing inside of the bounding box but belonging to a different person were not labeled. Thus, each mask is attached to a specific person in FACET. The breakdown of the masks per image is given in A.4.2. The breakdown of masks per given demographic or additional attribute is given in Table 18.

A.3. Annotation Quality Assurance (QA)

A.3.1 Annotator Quality Assurance using Training

Before completing any annotations used in FACET, annotators were trained for each stage separately. We trained

annotators by giving them a sample set of tasks and comparing their annotations to a known golden set. For Stages 1, 3 and 4 (image filtering, perceived skin tone annotation, other perceived attributes annotation), annotators passed the training step if the recall of their annotations compared to a fixed golden set was above a quality threshold. This threshold was set for each stage depending on the difficulty of the task. For Stage 2 (drawing bounding boxes), QA was done per annotator to assess the quality of boxes. We provided feedback to annotators individually and only graduated the annotators once they addressed the feedback. A manual IoU threshold of 0.85 between an annotator and the golden set was used. Annotators under that threshold were not manually reviewed, as we found that this correlated with extremely poor box quality, and these annotators did not graduate training. Before feedback, we noticed that many annotators were drawing bounding boxes that included objects the person was holding (*e.g. guitar*) as opposed to tightly around the person. After manual review and feedback, the quality of the annotations was much higher and consistent.

A.3.2 Annotation Quality Assurance using Multi-Review and Quality Checks

In addition to implementing a multi-review process for the perceived skin tone annotations of each target person as discussed in Section 5.2, we used Stage 3 to QA the bounding boxes drawn by the annotators. The annotators in Stage 3 were asked whether the bounding box for the person in the task was drawn tightly around the person. If – for any bounding box in the image – any of the three annotators marked that the bounding box was not tight, the image was placed back in Stage 2 of the pipeline to be re-annotated.

A.4. Dataset Statistics and Breakdown

A.4.1 Attribute Representation

We detail the attribute breakdown for the remaining annotations in FACET. Table 9 details the statistics for the remaining person annotations. Table 10 shows the results of the robustness annotations with breakdowns on occlusion level and lighting condition.

A.4.2 Image statistics

We measure the statistics of images beyond specific attributes. Figure 9 shows the number of annotated people per image; less than one third of the images contain more than one person. Figure 10 shows the person box size as a fraction of total image size, broken down by the number of people in the image. All images in FACET are used for detection. Images with only one person are used for classification and visual grounding. **For masks, the 69k labeled masks span 18k people in 17k images of FACET.** Each

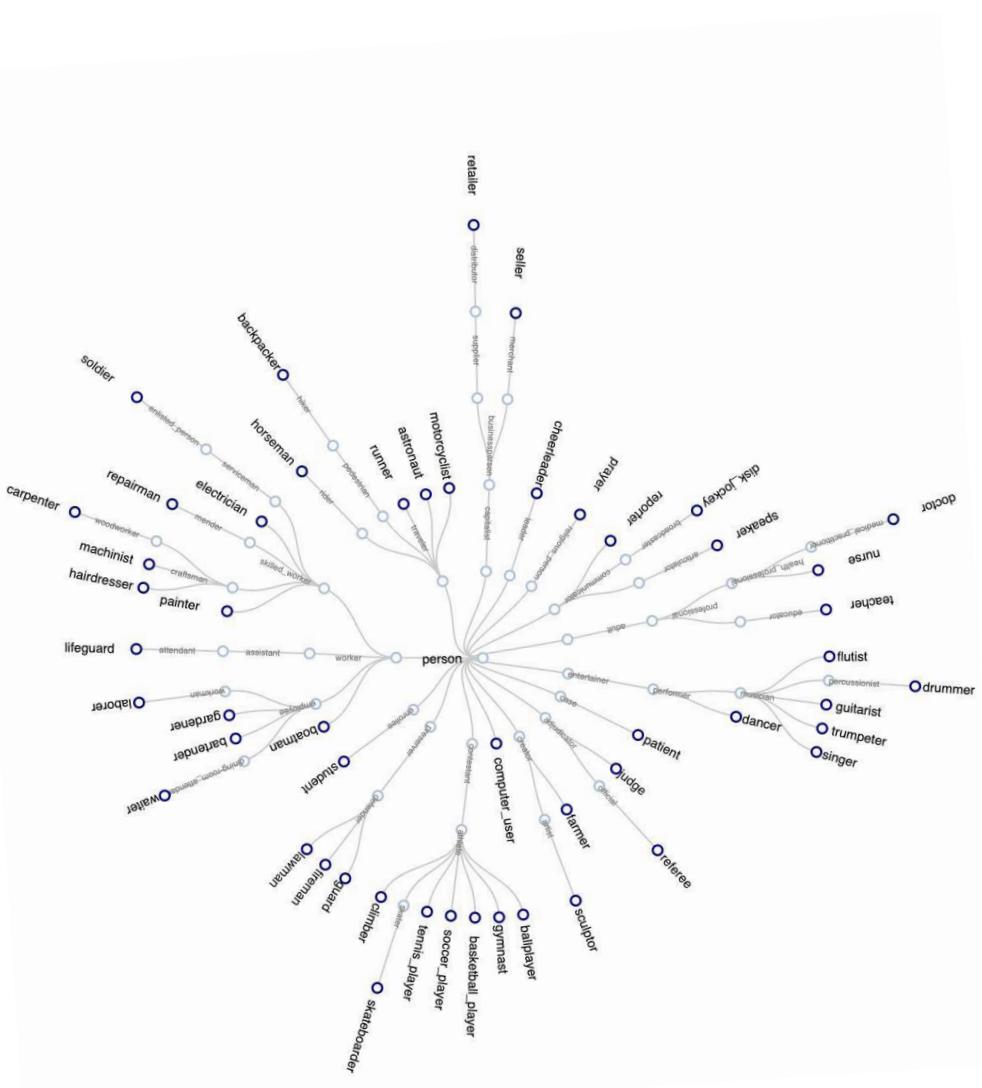


Figure 8: WordNet hierarchy of the FACET classes in relation to the `Person` synset. Classes are mapped to the `Person` synset (center) by their hypernyms (parents). Classes (leaves) are marked in blue. Grey nodes correspond to an intermediate hyponyms.

person with associated labeled masks has an average of 4 masks.

A.5. Evaluation

A.5.1 Dataset Setup

- For image classification, we limit the evaluation to examples in FACET that only contain one person. This helps alleviate ambiguities in performance. With this setup, we can consider the performance of the model on an image equivalent to performance of the model on the image for a specific set of attributes. There are 21k images in FACET that meet this criteria.

- For person and open world detection, we use all examples in FACET.
 - For person segmentation, and the corresponding person detection baseline, we only use images and people inside each image that had a person mask - 11k people.
 - For visual grounding, we only use examples in FACET with one person, as OFA predicts only one bounding box.

	<i>people</i>	<i>%</i>	<i>images</i>	<i>%</i>	
<i>Hair color</i>	black	17k	34%	13k	42%
	blonde	3k	6%	3k	8%
	brown	11k	22%	9k	29%
	red/orange	547	1%	518	2%
	colored	269	1%	265	1%
	grey	2k	4%	2k	6%
	unknown	20k	40%	15k	46%
<i>Hair type</i>	wavy	9k	19%	8k	26%
	curly	761	2%	735	2%
	straight	19k	37%	15k	47%
	coily	458	1%	435	1%
	dreadlocks	296	1%	282	1%
	bald	1k	2%	965	3%
	unknown	23k	45%	16k	52%
<i>Additional Annotations</i>	eyeware	5k	11%	5k	15%
	headscarf	2k	5%	2k	6%
	tattoo	705	1%	672	2%
	cap	14k	29%	10k	33%
	facial-hair	6k	12%	5k	17%
	mask	3k	6%	2k	7%

Table 9: Statistics on the remaining person attributes: *hair color*, *hair type*, *presence of additional features* in FACET. Annotators could mark multiple hair colors and types for a single person.

	<i>label</i>	<i>people</i>	<i>%</i>	<i>images</i>	<i>%</i>
<i>Lighting Condition</i>	overexposed	941	2%	890	3%
	well-lit	40k	80%	27k	85%
	dimly-lit	11k	22%	9k	28%
	underexposed	1k	3%	1k	4%
	unknown	878	2%	849	3%
<i>Visibility</i>	minimal	7k	15%	7k	21%
	face	15k	30%	12k	38%
	torso	36k	73%	25k	78%

Table 10: Robustness annotations.

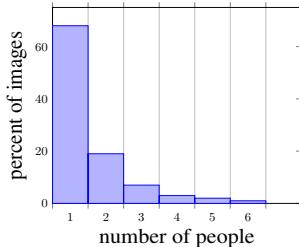
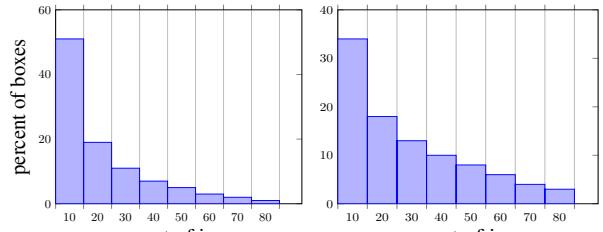


Figure 9: Histogram of number of people per image in FACET.



(a) all images (b) one person images

Figure 10: Histogram of person bounding box size as a percentage of total image size.

ID	Category	Attribute	Prediction
1	Dancer	+F	Dancer
2	Gardener	+M	Gardener
3	Dancer	U	Gardener
4	Dancer	+M	Dancer
5	Guitarist	+M	Gardener
6	Dancer	+M	Guitarist
7	Dancer	NB	Dancer
8	Dancer	+F	Dancer
10	Gardener	+F	Guitarist
11	Dancer	+F	Guitarist
12	Guitarist	+F	Guitarist
13	Dancer	+F	Dancer
14	Gardener	+F	Guitarist
15	Guitarist	U	Dancer

ID	Category	Attribute	Prediction
7	Dancer	NB	Dancer

recall for non binary presentation, dancer 1.0

ID	Category	Attribute	Prediction
4	Dancer	+M	Dancer
6	Dancer	+M	Guitarist

ID	Catagory	Attribute	Prediction
1	Male	Height	recall for more maleness, dancer is 0.5

ID	Category	Attribute	Prediction
1	Dancer	+F	Dancer
8	Dancer	+F	Dancer
11	Dancer	+F	Guitarist
13	Dancer	+F	Dancer

recall for *more femaleness*, dancer 0.75

Figure 11: Example of how we score classification models for FACET.

A.5.2 Choice of Metric

We choose to focus on recall as it allows us to only consider examples with a specific demographic attribute or set of attributes. We choose to avoid a metric that would take into account false positives, as for some evaluations it is not clear what a false positive would mean. For example, for person detection, it is not obvious which demographic attribute a false positive would correspond to. *What demographic attributes would we consider a predicted false positive person to have?* While it might make sense for images with only one person to assume they had the same demographic attributes as the ground truth person in the photo, it is even less clear what the correct assumption would be to make if there were multiple people in the photo. To avoid this ambiguity, we focus on recall.

Classification We compare performance on a per-class basis, as overall performance metrics can hide disparities – i.e. the model could have large biases but in opposite directions for two classes, which would yield an overall performance disparity of 0. We also don’t want the metric to be influenced by the prevalence of the class for the group. We focus on the recall ($R@1$) for the group and class for our evaluation. This is equivalent to the accuracy for the specific (class, attribute) pair. We do not account for true negatives or false positives. Figure 11 visualizes our metric. We note that there are multiple approaches to calculating a metric per class – e.g. we could also look at the accuracy for the class when looking at all examples of the protected group, which is why detail the specifics of our considered metric.

Alignment with traditional fairness metrics The difference in recall we measure is equivalent to *equality of opportunity* [41] – larger differences in recall are further from equality of opportunity. *Equalized odds* [41] is an extension of this, which analyzes the true negative rate (TNR). For example for CLIP, the largest difference in TNR between people perceived as having more stereotypically male and more stereotypically female attributes is for **retailers**, with 3.8 difference in TNR. This suggests that people perceived as having more stereotypically female attributes are over-predicted as **retailers**. The next largest difference between these two groups for CLIP is for **tennis player**, with a difference in TNR of -3.0, suggesting that people perceived as having more stereotypically male attributes are over-predicted as **tennis players**.

A.5.3 Classification

Experimental Setup In order to have maximum control over the experiment, we evaluate classification models on images in FACET that only contain one annotated person. By filtering out images with > 1 person, we are left with 21k images. We look at the per-class disparities between two groups only if both groups have at least 50 examples. We analyze CLIP based on recall.

ImageNet-21k Pretraining As FACET categories overlap with ImageNet classes, we can evaluate models trained on ImageNet-21k out of the box. We take the max score over the FACET classes from the ImageNet class predictions. Table 12 shows a comparison of performance discrepancies across perceived age group for CLIP ViT B/32 and a ViT B/16 pre-trained on IN21k from [79].

Architecture Choice

Person Class	#	CLIP ViT B/32			ViT B/16 IN21k				
		Y	M	O	#	Y	M	O	
Top for CLIP	seller	1	57.5	72.8	86.2	9	47.2	53.4	59.3
	ballplayer	2	60.6	75.5	-	2	57.6	77.4	-
	guitarist	3	70.3	80.2	65.5	10	45.5	47.9	36.4
	speaker	4	17.6	28.5	30.6	4	13.7	25.7	30.6
	laborer	5	49.0	52.7	61.7	3	48.1	52.9	66.0
Top for ViT	painter	21	56.5	51.0	53.9	1	37.0	43.1	57.8
	ballplayer	2	60.6	75.5	-	2	57.6	77.4	-
	laborer	5	49.0	52.7	61.7	3	48.1	52.9	66.0
	speaker	4	17.6	28.5	30.6	4	13.7	25.7	30.6
	guard	7	44.6	32.9	-	5	48.5	31.7	-

Table 11: Per-class performance for CLIP and a ViT pre-trained on ImageNet-21k. A subset of FACET classes are shown. The perceived age groups with the highest performance discrepancy per class are bolded. (Y is *young*, M is *middle*, O is *older*). The top five classes with the biggest discrepancies per model are shown. # corresponds to the rank for class in terms of magnitude of the discrepancy. Lower number indicates larger discrepancy. We note that most classes are in both of the models’ top 10 classes with the largest discrepancies.; 2 classes are in both models’ top 5. Recall for class and perceived age group pairings with less than 50 samples are not reported.

Person Class	#	ResNet IN21k			ViT IN21k				
		Y	M	O	#	Y	M	O	
Top for ResNet	laborer	1	35.6	38.1	55.3	3	48.1	52.9	66.0
	guard	2	49.5	30.5	-	6	48.5	31.7	-
	painter	3	38.9	35.9	53.9	1	37.0	43.1	57.8
	ballplayer	4	62.1	79.3	-	2	57.6	77.4	-
	craftsman	5	67.2	78.4	81.8	12	74.6	78.7	81.8
Top for ViT	painter	3	38.9	35.9	53.9	1	37.0	43.1	57.8
	ballplayer	4	62.1	79.3	-	2	57.6	77.4	-
	laborer	1	35.6	38.1	55.3	3	48.1	52.9	66.0
	speaker	15	20.6	25.9	24.6	4	13.7	25.7	30.6
	guard	2	49.5	30.5	-	5	48.5	31.7	-

Table 12: Per-class performance for a ViT and ResNet pre-trained on ImageNet-21k. A subset of FACET classes are shown. The perceived age groups with the highest performance discrepancy per class are bolded. (Y is *young*, M is *middle*, O is *older*). The top five classes with the biggest discrepancies per model are shown. # corresponds to the rank for class in terms of magnitude of the discrepancy. Lower number indicates larger discrepancy. Recall for class and perceived age group pairings with less than 50 samples are not reported.

A.5.4 Person Detection

We use a pre-trained Faster R-CNN with a ResNet50 FPN backbone pretrained on COCO for person detection.

Additional Results Table 13 shows person detection results across perceived gender presentation and perceived

<i>Demographic Group</i>	mAR	AR _{0.5}	AR _{0.75}
perceived gender presentation			
– more stereotypically maleness	74.4	97.8	83.1
– more stereotypically femaleness	72.2	97.9	80.7
– outside of gender binary	71.2	97.9	76.8
perceived age group			
– younger	73.9	98.3	82.6
– middle	74.3	98.0	83.1
– older	74.8	98.5	84.5

Table 13: Average recall (AR) on FACET for a ResNet50 Faster R-CNN. Mean AR (mAR) averages across IoUs from 0.5 to 0.95 in increments of 0.05; AR_{0.5} and AR_{0.75} refer to IoU at 0.5 and 0.75.

<i>Monk Skin Tone (MST)</i>	mAR	AR _{0.5}	AR _{0.75}
1	85.4	99.0	93.3
2	84.6	98.8	92.1
3	84.4	98.7	91.6
4	84.2	98.6	91.3
5	84.0	98.6	91.2
6	84.0	98.7	91.2
7	83.8	98.6	91.1
8	84.1	98.6	91.5
9	83.6	98.6	90.9
10	82.8	98.2	90.1

Table 14: Average recall (AR) on FACET for a ResNet50-backbone DETR model. Mean AR (mAR) averages across IoUs from 0.5 to 0.95 in increments of 0.05; AR_{0.5} and AR_{0.75} refer to IoU at 0.5 and 0.75.

age group. Table 14 shows person detection results for a DETR [11] model with a ResNet50 backbone for perceived skin tone.

A.5.5 Person Segmentation

We use a Mask R-CNN [44] with a ResNet50 FPN backbone pretrained on COCO for person detection and instance segmentation. For this experiment, we only evaluate people in images if they have a mask annotated as `person` as well. This leaves us with 11k examples (people). For boxes, we compute the IoU of the predicted box to the human-labeled bounding box in FACET. For masks, we compute the IoU of the predicted mask to the Segment Anything-generated, annotator verified, mask in Segment Anything 1 Billion (SA-1B) [59]. Annotators verified and labelled the mask as `person`, and were instructed only to do so if the mask was around the entire person (similar to bounding boxes in FACET). Annotators did not make any updates to the mask boundary.

A.5.6 Open World Detection

Experimental Setup We use Detic [102] trained on IN21-k with a SWIN-B backbone for open world detection. For the CLIP embeddings, we use the prompt ‘a person who is a {}’ opposed to the ‘a {}’ used in the original paper. As we focus on recall, we do not use a confidence threshold for DETIC’s predictions. Similarly we allow multiple class predictions per box. We take the top 100 predictions per image to compute AR.

Additional Results Table 15 shows the per class disparities for all classes for perceived age group.

A.5.7 Visual Grounding

We evaluated OFA [93]. For OFA, we used the pretrained version OFA_{large} in the HuggingFace Transformers library [95]; we did not perform any additional finetuning. We used beam-search with 5 beams, top-p=0.6 and limited the generation to a maximum of 100 new tokens. We prompted OFA with the input (e.g. “Which region does the text {person class} describe?”). Because OFA produces a single bounding box per prompt, we only evaluated images that contained no more than one person instance per person class. 7858 images were excluded because they contained multiple instances per class. We show the average recall across different IoUs and for different perceived age group labels in Table 15.

	mAR			AR _{0.5}			AR _{0.75}			mAR			AR _{0.5}			AR _{0.75}		
	young	middle	older	young	middle	older	young	middle	older	young	middle	older	young	middle	older	young	middle	older
astronaut	64.0	70.2	-	80.0	85.3	-	60.0	73.7	-	0.0	30.3	-	0.0	51.4	-	0.0	32.9	-
backpacker	45.4	42.1	29.8	55.3	51.7	35.3	47.4	44.6	33.3	7.7	5.9	11.5	15.5	11.2	22.0	6.8	5.8	9.8
ballplayer	43.8	45.4	61.8	46.7	48.3	63.6	45.8	46.9	63.6	43.9	38.0	58.0	72.7	67.2	100.0	50.6	41.5	80.0
bartender	81.7	75.4	-	83.3	85.4	-	83.3	80.5	-	0.0	12.5	-	0.0	25.0	-	0.0	8.3	-
basketball player	61.2	66.7	-	69.0	74.7	-	64.9	70.4	-	26.2	24.2	-	46.4	41.3	-	25.0	26.0	-
boatman	69.4	59.9	64.0	87.7	79.1	82.3	77.8	65.3	68.8	1.4	1.9	1.5	4.5	6.2	8.1	0.6	0.9	0.0
carpenter	67.9	64.8	81.7	71.4	73.5	91.7	71.4	68.9	87.5	0.0	2.4	0.0	0.0	5.6	0.0	0.0	2.2	0.0
cheerleader	13.3	12.7	-	15.6	13.5	-	14.8	12.6	-	20.0	12.0	-	41.9	20.0	-	16.1	15.0	-
climber	76.6	74.4	67.5	91.9	92.4	75.0	81.8	81.0	75.0	0.0	1.2	0.0	0.0	2.5	0.0	0.0	0.8	0.0
computer user	72.9	66.3	68.4	81.0	77.8	73.7	76.3	67.5	68.4	7.9	5.9	8.6	12.9	10.3	14.3	8.2	6.2	7.1
craftsman	44.5	47.0	56.9	48.6	52.1	61.9	44.8	48.3	58.6	33.8	39.1	40.7	55.2	62.6	66.9	37.9	45.7	47.6
dancer	77.2	71.1	75.6	91.4	85.4	87.5	83.7	77.5	78.1	37.6	32.0	24.3	68.8	55.7	57.1	39.0	37.1	28.6
disk jockey	77.2	68.4	-	82.1	78.8	-	79.1	72.5	-	3.5	3.8	-	6.8	6.5	-	4.1	3.0	-
doctor	74.6	77.4	75.7	86.2	88.7	81.0	76.6	79.8	78.6	33.6	30.9	38.0	55.2	52.1	60.8	40.2	33.6	45.1
drummer	19.9	26.3	34.2	24.9	34.7	41.8	19.7	27.6	35.8	5.0	3.8	0.7	9.2	8.1	1.7	4.6	3.3	0.0
electrician	56.3	51.4	48.6	62.8	62.5	57.1	62.8	54.4	57.1	0.0	1.2	0.0	0.0	1.6	0.0	0.0	1.6	0.0
farmer	81.5	81.1	85.4	95.9	96.6	99.1	86.9	88.4	93.0	6.2	5.0	6.6	12.8	9.9	13.7	5.1	4.4	5.5
fireman	86.3	76.4	76.4	96.2	90.1	85.7	90.4	82.6	85.7	14.0	14.7	22.0	26.7	32.9	60.0	13.3	12.5	20.0
flutist	32.1	40.5	51.0	35.4	47.5	54.8	35.4	43.7	54.8	15.0	10.5	11.7	31.8	19.9	20.8	9.1	9.9	12.5
gardener	82.3	78.6	86.8	98.3	94.7	100.0	90.0	84.4	97.3	11.9	18.3	27.9	32.6	40.1	58.1	7.0	14.6	24.2
guard	81.9	80.2	88.5	94.3	90.6	97.5	89.4	87.2	95.0	14.1	15.2	19.2	34.0	31.9	38.5	9.6	12.5	11.5
guitarist	75.9	79.3	79.5	90.5	93.7	95.1	80.0	84.4	86.3	19.8	19.6	32.0	38.7	35.5	56.7	18.5	20.6	36.7
gymnast	87.7	85.5	-	96.2	95.6	-	92.4	89.9	-	10.0	8.5	-	19.1	17.1	-	9.8	9.1	-
hairdresser	76.8	79.4	79.0	94.1	96.9	92.9	82.4	79.9	83.3	15.2	13.3	12.1	28.0	24.3	23.5	12.0	12.8	14.7
horseman	70.9	62.1	64.5	85.4	75.7	80.0	77.2	67.9	70.0	13.4	14.5	11.0	36.5	38.8	30.0	3.8	5.4	0.0
judge	25.7	31.3	28.3	28.6	35.3	33.3	28.6	33.8	33.3	-	10.4	0.0	-	25.0	0.0	-	3.6	0.0
laborer	75.3	73.1	74.4	88.4	85.8	86.1	79.9	78.9	79.9	23.1	21.9	28.9	44.0	46.2	58.6	22.0	17.4	24.3
lawman	71.5	70.1	67.1	79.0	77.7	74.3	75.5	74.6	70.6	20.2	21.1	22.6	42.0	43.1	46.2	18.3	18.8	21.5
lifeguard	41.8	46.1	52.5	51.7	54.9	62.5	47.5	49.8	62.5	7.5	7.0	0.0	19.7	17.9	0.0	2.8	5.2	0.0
machinist	60.0	49.9	41.1	63.9	56.5	44.4	63.9	52.2	44.4	21.7	21.3	23.3	34.8	35.5	41.7	26.1	25.0	25.0
motorcyclist	57.9	52.7	51.9	81.6	78.2	69.2	60.9	54.2	57.7	21.9	15.5	19.2	50.0	37.0	37.5	12.3	9.6	20.8
nurse	83.4	81.5	81.7	95.6	93.9	91.3	90.5	86.1	82.6	31.8	24.8	34.5	52.2	43.6	50.0	37.2	26.0	40.0
painter	54.0	58.9	68.6	60.8	66.3	73.8	58.2	62.3	73.8	18.0	15.6	17.7	30.1	29.3	27.6	23.3	16.1	20.4
patient	64.1	66.9	67.1	87.0	85.6	86.5	65.6	69.2	68.3	28.5	26.5	26.6	50.3	47.6	45.2	29.7	27.6	28.0
prayer	82.8	83.0	85.2	96.0	95.2	95.2	89.0	89.5	89.5	0.0	2.7	2.8	0.0	5.5	4.3	0.0	1.8	2.9
referee	70.2	77.5	84.9	75.5	85.3	91.4	73.6	80.9	88.6	19.6	20.4	21.4	40.8	40.1	45.7	16.3	19.8	22.9
repairman	71.2	61.7	65.2	77.6	69.7	71.0	75.0	65.5	69.6	20.1	17.9	17.0	39.5	32.8	30.4	18.4	19.1	17.4
reporter	21.7	22.9	25.0	23.7	25.7	29.2	22.4	23.7	25.0	9.2	5.2	4.5	19.7	13.0	6.9	7.0	3.9	3.4
retailer	33.3	35.0	52.2	40.9	43.2	59.5	33.6	38.6	54.1	1.0	2.5	3.1	2.8	6.5	6.9	0.0	1.5	3.4
runner	90.9	85.9	91.1	99.2	95.2	100.0	97.7	90.6	100.0	7.5	8.3	0.0	21.6	21.1	0.0	3.9	4.7	0.0
sculptor	74.5	73.0	85.0	81.8	82.4	95.8	77.3	77.6	83.3	2.4	2.3	0.0	5.9	5.1	0.0	0.0	3.1	0.0
seller	73.0	73.2	74.5	87.4	87.0	88.5	82.0	79.0	80.8	7.8	8.2	9.8	16.8	16.6	21.1	6.2	6.9	8.8
singer	80.6	80.9	85.0	88.8	88.2	96.1	85.0	85.5	88.2	5.1	3.9	1.7	10.3	7.3	5.8	4.5	3.6	0.0
skateboarder	40.7	43.1	-	45.1	46.5	-	43.4	46.1	-	21.5	23.9	-	46.9	49.7	-	18.8	22.1	-
soccer player	81.8	82.9	-	90.6	91.7	-	86.2	87.1	-	26.7	22.6	-	49.1	42.8	-	24.6	21.4	-
soldier	65.1	63.4	51.4	72.8	72.7	56.8	69.7	68.3	54.1	16.3	1.2	40.0	33.8	6.2	9.2	14.5	0.0	
speaker	83.0	80.8	85.1	89.3	88.5	93.1	87.7	85.6	89.1	2.0	1.7	2.1	4.6	3.2	3.7	1.9	1.7	1.6
student	60.6	71.1	-	69.8	80.9	-	64.6	74.4	-	29.0	25.3	0.0	51.9	44.7	0.0	33.8	25.0	0.0
teacher	83.4	81.0	80.0	96.6	90.3	87.5	93.1	85.8	87.5	28.1	22.2	15.0	51.6	39.8	50.0	29.0	24.8	0.0
tennis player	94.2	93.8	-	98.9	98.9	-	97.2	97.8	-	32.5	33.8	-	60.0	62.2	-	32.7	34.4	-
trumpeter	22.8	29.5	38.4	26.7	34.8	45.5	25.6	31.4	38.2	5.3	5.1	3.6	11.6	10.3	5.1	2.3	5.7	5.1
waiter	76.2	77.6	-	92.4	92.9	-	83.3	82.5	-	5.2	4.2	-	10.4	8.6	-	4.2	4.0	-
avg	64.6	64.0	68.2	74.1	74.4	76.4	68.6	67.9	72.4	14.7	14.5	14.0	28.5	27.8	26.2	13.8	14.1	15.0

(a) Results for Detic

(b) Results for OFA

Table 15: The average recall (AR) results for Detic (detection) and OFA (visual grounding) across the 52 person-related classes for each perceived age group label. The highest recall numbers are bolded.

B. Data Card

We provide a data card for FACET, following the guidance of [48].

FACET https://facet.metademolab.com	
FACET is a large, publicly available evaluation set of 31,702 images for the most common vision problems - image classification, object detection, segmentation and visual grounding . People in FACET are annotated with person-related attributes such as perceived skin tone and hair type , bounding boxes and labeled with fine-grained person-related classes such as <i>disk jockey</i> or <i>guitarist</i> .	
Overview	
Publisher	Meta AI Research, FAIR
Authors	Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, Candace Ross
Contact	facet@meta.com
Funding & Funding Type	Industry
License	Custom license, see dataset download agreement
Applications	
Dataset Purpose	Evaluate computer vision models to detect potential fairness concerns
Key Application	<i>Computer Vision, Fairness and Robustness</i>
Primary Motivations	Give researchers a tool to help understand model fairness. Allow researchers to investigate how the demographic attributes of a person in the image correlates with model performance. FACET supports common vision tasks, with annotations for classification, detection, and segmentation.
Intended Audience	Researchers aiming to detect potential fairness concerns and biases in their trained vision models.
Suitable Use Case	FACET is for evaluation only.
Data Type	
Primary Data Type	Images
Primary Annotation Types	Manually gathered annotations for: <ul style="list-style-type: none">• Bounding boxes• Category labels for the bounding boxes• A series of demographic, robustness, and additional attributes for the person in the bounding box• Manually annotated labels for masks from Segment Anything 1 Billion (SA-1B[59]). These masks were automatically generated by the Segment Anything Model (SAM).

FACET Data Card	
Data SnapShot	<ul style="list-style-type: none"> • 31,702 images • exhaustive annotations for 49,551 people • 52 categories for people that include occupations, athletes, artists, etc • 13 attributes annotated for person including demographic attributes such as <i>perceived gender presentation</i> robustness annotations such as <i>lighting condition</i> and additional attributes such as <i>hair color</i> • 3 mask labels <i>person</i>, <i>clothing</i>, <i>hair</i> for 69k masks. Masks and mask labels are not exhaustive. 17k people in 14k images have labelled masks. Additional unlabeled masks from SA-1B are compatible with FACET .
Data Sources	Images come from SA-1B.
Annotation format	<p>JSON files of COCO formatted⁵ annotations for the bounding boxes and masks are provided, and person attribute annotations are given in a CSV file.</p> <p><code>coco_boxes.json</code>: COCO-style JSON file containing bounding boxes for people in FACET. The <code>id</code> field for each annotation is the same as <code>person_id</code>. <code>Category_id</code> corresponds to the primary class for the person.</p> <p><code>coco_masks.json</code>: COCO-style JSON file containing Segment Anything Model (SAM) generated masks for people in FACET. Each annotation has an additional <code>person_id</code> field, corresponding to the <code>person_id</code> in <code>annotations.csv</code> and <code>coco_boxes.json</code>. Categories for masks are one of [person, clothing, hair]. Masks are non-exhaustive.</p> <p><code>annotations.csv</code>: A CSV containing the annotations per person is be provided. Each item in the annotation file contains:</p> <ol style="list-style-type: none"> 1. Reference information: <ul style="list-style-type: none"> • <code>filename</code> • <code>person_id</code>: unique integer representing the annotation 2. Task information: <ul style="list-style-type: none"> • <code>class1</code>: This is the primary category the person matches. <i>Cannot be None</i>. • <code>class2</code>: This is the secondary category the person matches. <i>Can be None</i>. • <code>bounding_box</code>: Person bounding box.

⁵<https://cocodataset.org/#format-data>

FACET Data Card

3. Demographic Attribute annotations.
 - *perceived gender presentation* : All of the following annotations will be given in a binary fashion: [gender_presentation_masc, gender_presentation_non_binary, gender_presentation_fem, gender_presentation_na]
 - *perceived skin tone* : Each annotator's annotations are considered per MST in a binary fashion. Annotations from all annotators are summed into a single value per MST, so the value at MST_i may be greater than 1. Values will be given for all of the following: [skin_tone_1, ..., skin_tone_10, skin_tone_na]
 - *perceived age group* : all of the following annotations are included in a binary fashion: [age_presentation_young, age_presentation_middle, age_presentation_older, age_presentation_na]
4. Additional Attribute information: All binary values.
 - *hair color*: [hair_color_black, hair_color_red, hair_color_blonde, hair_color_brown, hair_color_colored, hair_color_grey, hair_color_na]
 - *hair type*: [hair_type_wavy, hair_type.curly, hair_type_coily, hair_type_straight, hair_type_bald, hair_type_dreadlocks, hair_type_na]
 - *other items*: [has_eyewear, has_headscarf, has_tattoo, has_cap, has_facial_hair, has_mask]
5. Robustness Annotations: All binary values.
 - *lighting condition*: [lighting_overexposed, lighting_underexposed, lighting_well_lit, lighting_dimly_lit]
 - *visibility*: [visible_minimal, visible_torso, visible_face]

C. FACET CrowdWorkSheets

To further describe our annotation process, we answer the questions posed in CrowdWorkSheets[21].

C.1. Task Formulation

At a high level, what are the subjective aspects of your task? Annotating the *perceived* attributes of a person is by nature subjective. For perceived skin tone we expected the annotations would be subjective and have high variance. To account for this, we gather annotations from three annotators and release the cumulative results of all three. For subjectivity across the other attributes and labeling classes, we provided annotators with diverse representations of each attribute or class in the guidelines to try to minimize annotator bias.

What assumptions do you make about annotators? How did you choose the specific wording of your task instructions? What steps, if any, were taken to verify the clarity of task instructions and wording for annotators? To qualify for the annotation task, annotators had to pass a strong English requirement. For the annotation of perceived skin tone only, we had a more lenient English requirement to increase the diversity of the annotators, and additionally translated the annotation instructions into Spanish.

As we were annotating images, we provided visual examples for all of the annotations and classes. We sourced multiple examples per attribute (e.g. brown hair) and class (e.g. doctor), with at least one example for someone with more stereotypical maleness with the attribute and someone with more stereotypical femaleness with the attribute. For classes, we sourced multiple examples of someone who would qualify for a given class (*e.g. for dancer we sourced images of both a ballerina and a break-dancer*). For given examples for the Monk Skin Tone (MST) scale, we sourced four examples per MST value, and attempted to capture some of the diversity within a specified MST value.

What are the precise instructions that were provided to annotators? The goal of the project is to build a dataset that helps determine if Computer Vision models have biases based on the apparent attributes of the person in the photo. We are creating an image classification dataset that also contains labels of the apparent protected attributes of the people in the image. The dataset is for evaluation only, and is to help better analyze and detect potential biases. The protected attributes will not in any way be used for training a model. We are not collecting any biometric information about the people in the photos.

- 1. Target category classification:** Given an image, and a target category, we aim to determine if the image is

a good representation for the category. The annotators will mark whether or not there is a person in the photo matching the category, and if so if there are ≤ 5 people who match this category. The categories will be all people related - such as doctor, soccer player, etc. Multiple images will be shown per task to annotate. The default response will be ‘No person matches this category’.

- 2. Bounding boxes and classification labels for people:** Given an image, draw bounding boxes around all people who match any of the list of categories. For each bounding box around a person, mark which category they belong to. If they belong to multiple categories, you should mark the second category under ‘secondary category’.
- 3. Apparent skin tone annotations** Given an image, with a bounding box around a person, annotate the person’s apparent skin tone. You may select as many skin tones from the list as you feel appropriate. If it is not possible to tell the skin tone from the photo, please mark cannot be determined. Please select at least two values for the skin tone, and make sure that the values that you select are consecutive. If it is too hard to determine the annotation, mark the values it appears and cannot be determined. Zoom in (option + mouse scroll) as necessary in order to determine the skin tone.
- 4. Apparent attribute annotations** Given an image, with a bounding box around a person, annotate the given apparent attributes of the person. For each category, see the examples given. If it is not possible to determine the attribute from the photo, please mark cannot be determined. Apparent lighting condition is on the person: Please indicate how the lighting is with respect to the person in the bounding box. If the lighting is between two categories, mark both.

C.2. Selecting Annotations

Are there certain perspectives that should be privileged? If so, how did you seek these perspectives out? No. N/A

Are there certain perspectives that would be harmful to include? If so, how did you screen these perspectives out? Harmful perspectives would include annotators who had a clear bias in their annotations. We screened these perspectives out by using training, and only including production raters who had high accuracy on the training set. Annotators with consistent bias would likely not have been able to get a high enough accuracy on the training to graduate.

Were sociodemographic characteristics used to select annotators for your task? If so, please detail the process. If you have any aggregated sociodemographic statistics about your annotator pool, please describe. Do you have reason to believe that sociodemographic characteristics of annotators may have impacted how they annotated the data? Why or why not? We sourced geographically diverse annotators from the following 7 countries during our annotation process: United States, Philippines, Egypt, Colombia, Taiwan, Spain and Kenya. The breakdown of annotators per region is shown in Figure 4 in the main text.

If you have any aggregated sociodemographic statistics about your annotator pool, please describe. Do you have reason to believe that sociodemographic characteristics of annotators may have impacted how they annotated the data? Why or why not? Other sociodemographic statistics about our annotator pool were not available.

Consider the intended context of use of the dataset and the individuals and communities that may be impacted by a model trained on this dataset. Are these communities represented in your annotator pool? The FACET benchmark is to be used for evaluation purposes only. The underlying images in FACET are geographically diverse. To incorporate geographic diversity into our annotation process, we sourced annotators from 7 countries across regions.

C.3. Platform and Infrastructure Choices

What annotation platform did you utilize? At a high level, what considerations informed your decision to choose this platform? Did the chosen platform sufficiently meet the requirements you outlined for annotator pools? Are any aspects not covered? We used a proprietary annotation platform.

What, if any, communication channels did your chosen platform offer to facilitate communication with annotators? How did this channel of communication influence the annotation process and/or resulting annotations? For Stage 2 (drawing and labeling bounding boxes for person classes), annotators' annotations were compared to a golden set and were required to achieve IoU above 85% to pass. After these training stages, annotations were manually reviewed and the annotators were given feedback for improvement. Following this, if annotators had high quality labels when spot-checked, they graduated to annotating images for the final benchmark.

We provided annotators individualized feedback during their training for drawing bounding boxes on a daily basis. Our vendor communicated to annotators common types of mistakes that we witnessed during training, and the corresponding corrections.

How much were annotators compensated? Did you consider any particular pay standards, when determining their compensation? If so, please describe. Annotators were compensated with an hour wage set per country.

C.4. Dataset Analysis and Evaluation

How do you define the quality of annotations in your context, and how did you assess the quality in the dataset you constructed? For each task, annotators were first placed into training for the task. They were asked to annotate a large number of examples per task. We hand annotated the same examples and used our annotations as the ground truth measured the accuracy per annotator. Annotators were graduated from training when their accuracy reached above a given threshold. For the task requiring annotators to draw bounding boxes around people, annotators were only graduated after we manually spot checked the their bounding boxes to ensure quality. During the perceived skin tone annotation task, we asked annotators if they agreed with the class label, and grade the quality of the given bounding box. If one of the three annotators disagreed with the class label or bounding box, the annotation was removed, and the image added to the queue of images for task 2 (drawing bounding boxes).

Have you conducted any analysis on disagreement patterns? If so, what analyses did you use and what were the major findings? We pointed out common mistakes during weekly meetings with the vendor. While in training, we noticed consistent mistakes among annotators that we corrected before graduation. The most common mistake was around drawing the bounding boxes: many annotators during training would draw bounding boxes that included objects the person was holding *e.g guitar*. With the weekly meetings and individualized feedback, we were able to address this.

How do the individual annotator responses relate to the final labels released in the dataset? For perceived skin tone only, we sourced three annotations per person in the dataset. We release the annotations from all three annotators, giving a distribution over perceived skin tone per person in the dataset. We believe that a distribution more accurately describes a person's perceived skin tone than a single value.

C.5. Dataset Release and Maintenance

Do you have reason to believe the annotations in this dataset may change over time? Do you plan to update your dataset? At this time we do not plan to have updates for this dataset. We will allow users to flag any images that may be objectionable content, and remove objectionable content if found.

Are there any conditions or definitions that, if changed, could impact the utility of your dataset? The FACET benchmark contains examples for many different types of professions, athletes, artists, etc. If over time the way these occupations look shifts, this could impact the dataset. As a concrete example, there are a number of images in the dataset that were taken since the beginning of the COVID-19 pandemic. Many doctors and nurses in the dataset are wearing much more PPE than in images of doctors and nurses from before the COVID-19 pandemic.

Will you attempt to track, impose limitations on, or otherwise influence how your dataset is used? If so, how?
The FACET benchmark is for evaluation purposes ONLY. Using FACET annotations for training is strictly prohibited. Users must agree to the terms of use before downloading the dataset.

Were annotators informed about how the data is externalized? If changes to the dataset are made, will they be informed? No. No.

Is there a process by which annotators can later choose to withdraw their data from the dataset? If so, please detail. No.

D. Fine-grained dataset statistics

Person Class	Total	FACET Class Statistics															
		Perceived Gender Presentation			Perceived Skin Tone										Perceived Age Group		
		stereotypical maleness	stereotypical femaleness	non-binary presentation	1	2	3	4	5	6	7	8	9	10	younger	middle	older
lawman	4609	3768	403	3	560	2363	2881	2642	1825	1215	615	322	166	74	387	3151	144
laborer	3030	2208	378	1	112	577	887	1171	1171	1269	844	508	291	136	297	1643	193
boatman	2147	1074	742	5	137	741	991	995	794	573	296	134	86	31	482	906	147
guard	1851	1597	121	4	306	1045	1208	1047	714	470	246	119	62	27	317	1181	48
backpacker	1738	1006	458	4	167	771	1010	951	761	475	193	93	50	28	361	842	53
basketball player	1680	1479	134	2	307	869	991	845	632	461	413	359	332	165	492	1056	3
tennis player	1663	1058	488	0	147	805	1152	1262	1002	617	234	126	90	57	360	1081	6
farmer	1632	823	539	1	50	208	335	466	635	816	681	450	216	87	129	844	227
soldier	1561	1336	75	0	204	766	892	802	578	463	281	130	66	22	237	972	39
singer	1518	1013	428	14	240	824	1013	931	677	399	184	140	93	46	357	984	89
dancer	1475	510	812	10	207	644	863	798	716	431	214	113	67	26	567	644	32
speaker	1470	1119	282	1	152	789	1093	1050	755	392	182	92	49	25	134	987	207
motorcyclist	1468	822	302	4	55	384	518	583	539	478	249	104	59	22	213	577	56
repairman	1430	1187	54	1	65	480	681	815	753	676	373	150	60	18	126	836	78
seller	1342	699	533	6	74	359	555	705	757	642	379	192	74	28	205	758	184
ballplayer	1316	1145	62	1	104	584	743	781	635	473	253	168	112	51	214	834	12
guitarist	1279	1115	87	3	138	678	843	816	596	330	139	73	50	26	233	802	116
computer user	1267	597	322	2	176	641	818	785	608	358	152	71	34	12	258	449	24
soccer player	1233	1102	34	1	113	521	692	711	559	364	200	126	128	76	322	732	5
craftsman	1127	785	220	4	75	321	467	598	631	627	389	210	92	36	117	599	188
nurse	1124	322	535	3	115	368	505	536	529	399	188	86	27	6	169	547	24
drummer	1006	744	162	3	114	428	534	483	388	331	222	150	99	42	256	530	68
skateboarder	1000	818	88	1	82	468	635	650	463	281	136	62	37	14	360	465	1
painter	983	590	251	0	77	318	460	530	506	420	246	123	56	22	168	435	129
fireman	933	674	34	0	68	270	358	391	237	192	77	22	10	7	55	512	14
patient	896	408	275	0	75	280	389	472	486	444	242	102	41	16	131	368	127
horseman	884	491	290	1	152	538	592	484	287	127	54	26	12	5	181	512	22
doctor	861	361	313	1	86	343	450	462	410	284	145	69	21	6	107	441	43
prayer	810	444	265	3	58	223	307	355	394	357	195	99	51	18	104	358	124
referee	776	694	38	1	88	417	539	547	374	186	80	38	23	10	54	584	35
student	747	379	247	1	92	241	322	365	367	316	163	93	62	26	319	264	5
runner	654	469	117	3	88	320	415	379	291	126	45	23	27	16	134	403	19
gymnast	635	252	316	1	116	348	424	366	322	145	43	28	17	6	300	233	2
retailer	561	296	234	0	53	198	301	298	298	196	90	41	14	5	114	332	39
climber	551	355	92	2	59	231	306	301	251	155	73	26	13	8	107	261	4
trumpeter	530	451	36	3	63	308	336	304	212	145	74	41	37	18	89	316	56
lifeguard	529	398	62	0	20	160	232	286	229	186	103	52	28	8	118	273	8
electrician	505	415	7	0	9	100	140	188	182	175	101	52	35	9	47	270	9
gardener	499	266	173	1	45	187	257	265	235	197	108	56	33	16	66	245	79
reporter	473	302	145	1	75	281	324	269	204	116	50	18	13	4	77	302	24
hairdresser	461	342	85	3	32	143	209	257	242	237	145	75	35	17	69	294	43
machinist	413	329	30	0	33	173	223	252	191	168	89	34	20	7	42	241	20
cheerleader	410	78	314	0	77	191	292	268	205	88	38	18	12	3	246	117	5
waiter	350	204	109	1	34	184	245	220	177	120	51	24	18	7	68	224	7
disk jockey	318	228	27	1	43	162	200	194	127	77	37	27	20	10	67	167	2
flutist	312	247	41	0	38	152	192	184	154	118	77	43	16	4	50	189	32
astronaut	289	165	14	0	15	72	89	78	58	18	2	0	0	2	5	158	2
carpenter	268	230	7	0	11	82	124	147	129	131	87	52	25	9	20	160	27
sculptor	240	187	21	0	10	76	104	120	107	107	78	50	24	5	24	144	27
teacher	216	116	76	1	28	104	141	142	108	76	36	16	10	4	31	150	9
judge	101	67	28	0	11	50	76	71	44	21	6	3	1	0	8	71	12
bartender	57	37	14	0	5	27	42	36	29	19	7	3	1	7	41	1	

Table 17: Number of people for each person class and demographic group in FACET.

FACET Mask Statistics			
	person	clothing	hair
perceived gender presentation			
with stereotypical maleness	6608	32103	3788
with stereotypical femaleness	4127	18136	3346
non-binary presentation	50	223	36
cannot be determined	72	193	13
perceived skin tone			
MST 1	2198	10687	1389
MST 2	5154	24328	3496
MST 3	6121	28825	4263
MST 4	5651	26583	3889
MST 5	4849	22738	3349
MST 6	3816	17931	2452
MST 7	2542	11845	1544
MST 8	1619	7564	922
MST 9	1216	5727	666
MST 10	521	2481	293
cannot be determined	2839	11844	1611
perceived age group			
younger	4145	19440	3107
middle	5443	25458	3319
older	1134	5352	733
cannot be determined	135	405	24
Hair color			
black	4053	18137	3323
brown	2726	12205	2267
blonde	1024	4633	952
red/orange	148	674	136
colored	84	340	96
grey	747	3519	559
cannot be determined	2885	14863	485
Hair type			
wavy	2090	9526	1897
curly	241	1141	253
straight	5141	22109	4395
coily	178	750	158
dreadlocks	113	522	109
bald	265	1167	81
Unknown	3626	19129	905
Additional attribute			
eyeware	1509	6993	957
headscarf	665	3634	256
tattoo	184	926	143
cap	3305	18209	797
facial hair	1511	7382	963
mask	591	3271	377

Table 18: Number of masks per type for the attributes in FACET. For perceived skin tone, hair color, hair type, and additional attributes, a person in FACET can be marked with multiple values; therefore, the sum of the masks over the group of attributes may be greater than the total number of masks.