International Journal of Computer Vision (2022) 130:1790–1810
https://doi.org/10.1007/s11263-022-01625-5

REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets

Angelina Wang[1] · Alexander Liu[1] · Ryan Zhang[1] · Anat Kleiman[1] · Leslie Kim[1] · Dora Zhao[1] · Iroha Shirai[1] · Arvind Narayanan[1] · Olga Russakovsky[1]

Received: 17 July 2021 / Accepted: 27 April 2022 / Published online: 23 May 2022

Abstract
Machine learning models are known to perpetuate and even amplify the biases present in the data. However, these data biases frequently do not become apparent until after the models are deployed. Our work tackles this issue and enables the preemptive analysis of large-scale datasets. REvealing VIsual biaSEs (REVISE) is a tool that assists in the investigation of a visual dataset, surfacing potential biases along three dimensions: (1) object-based, (2) person-based, and (3) geography-based. Object-based biases relate to the size, context, or diversity of the depicted objects. Person-based metrics focus on analyzing the portrayal of people within the dataset. Geography-based analyses consider the representation of different geographic locations. These three dimensions are deeply intertwined in how they interact to bias a dataset, and REVISE sheds light on this; the responsibility then lies with the user to consider the cultural and historical context, and to determine which of the revealed biases may be problematic. The tool further assists the user by suggesting actionable steps that may be taken to mitigate the revealed biases. Overall, the key aim of our work is to tackle the machine learning bias problem early in the pipeline. REVISE is available at https://github.com/princetonvisualai/revise-tool.

Keywords Computer vision datasets · Bias mitigation · Tool

1 Introduction

Computer vision dataset bias is a well-known and much-studied problem. Torralba and Efros (2011) highlighted the fact that every dataset is a unique slice through the visual world, representing a particular distribution of visual data. Since then, researchers have noted the under-representation of object classes (Oksuz et al., 2019; Ouyang et al., 2016; Yang et al., 2014; Salakhutdinov et al., 2011; Buda et al., 2017; Liu et al., 2009), object contexts (Choi et al., 2012; Rosenfeld et al., 2018), object sub-categories (Zhu et al., 2014), scenes (Zhou et al., 2017), gender (Kay et al., 2015), gender contexts (Zhao et al., 2017; Burns et al., 2018), skin tones (Buolamwini & Gebru, 2018; Wilson et al., 2019), geographic locations (Shankar et al., 2017) and cultures (DeVries et al., 2019). The downstream effects of these under-representations range from the more innocuous, like limited generalization of car classifiers (Torralba & Efros, 2011), to the much more serious, like having deep societal implications in automated facial analysis (Buolamwini & Gebru, 2018; Hill, 2020). Efforts such as Datasheets for Datasets (Gebru et al., 2018) have played an important role in encouraging dataset transparency through articulating the intent of the dataset creators, summarizing the data collection processes, and warning downstream dataset users of potential biases in the data. However, this alone is not sufficient, as there is no algorithm to identify all biases hiding in the data, and manual review is not a feasible strategy given the scale of modern datasets.

Bias Detection Tool To mitigate this issue, we provide an automated tool for REvealing VIsual biaSEs (REVISE) in datasets (Fig. 1). REVISE is a broad-purpose tool that uses a combination of existing annotations and automated tools for surfacing the under- and different- representations hiding within visual datasets. For the current exploration we limit ourselves to three sets of metrics: (1) object-based, (2) person-based and (3) geography-based.

Object-based analysis is most familiar to the computer vision community (Torralba & Efros, 2011), as many of the popular visual recognition datasets are object-centric (Everingham et al., 2010; Russakovsky et al., 2015). Thus, these

Communicated by Diane Larlus.

✉ Angelina Wang
angelina.wang@princeton.edu

[1] Princeton University, Princeton, USA

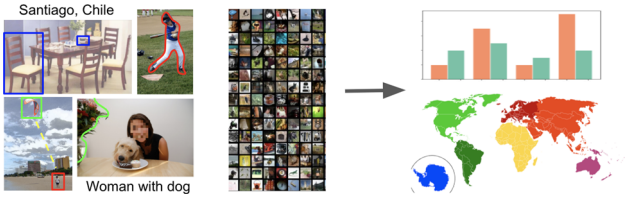

Fig. 1 Our tool takes in as input a visual dataset and its annotations, and outputs metrics, seeking to produce insights and possible actions

analyses focus on considering statistics about object frequency, scale, context, or diversity of representation.

Person-based analyses began to gain attention with research showing unequal performance for people of different genders and skin tones (Gebru et al., 2018; Zhao et al., 2017). This line of analysis considers the representation of people of different demographics within the dataset, and allows the user to assess what potential downstream consequences this may have in order to consider how best to intervene. It also builds on the object-based analysis by considering how the representation of objects with people of different demographic groups differs.

Finally, geography-based analysis considers the portrayal of different geographic regions within the dataset; this is a relatively new but very important conversation within the community (Shankar et al., 2017; DeVries et al., 2019). This axis of analysis is deeply intertwined with the previous two, as geography influences both the types of objects that are represented, as well as the different people that are pictured.

We imagine two primary use cases for our tool: (1) dataset builders can use the actionable insights produced by our tool during the process of dataset compilation to guide the direction of further data collection, and (2) dataset users who train models can use the tool to understand what kinds of biases their models may inherit as a result of training on a particular dataset.

Tool Inputs Our tool takes in as input a collection of images and their corresponding annotations. Which metrics can be computed depend on the annotations available, e.g., geolocation annotations are required to compute statistics about different geographical representation. To perform analyses beyond just the annotations provided, we also use external tools and pretrained models, such as Fasttext language detection (Joulin et al., 2016a, b), Places scene detection (Zhou et al., 2017), and automatic feature extraction (Idelbayev, 2019) to derive some of our metrics. Of course, both human (van Miltenburg et al., 2018) and model (Buolamwini & Gebru, 2018) annotations will contain their own sets of biases, such as being systematically more accurate for one group over another. By incorporating tools which are external to the input dataset, we are potentially introducing additional confounding biases into the analyses. However, this is still worth doing in order to generate more insightful analyses beyond the set of data and annotations immediately available to us. It is important to acknowledge and understand the biases that annotations, whether collected or automatically detected, will surely bring with them. For example, if the language classifier used on image labels to infer the locality of the photographer is over-predicting English, then one should understand these limitations and interpret their results accordingly, perhaps by performing additional manual verification. Thus, the presence of model cards to accompany any such models would help users understand the limitations and assumptions that come with their use (Mitchell et al., 2019). As with all findings surfaced by REVISE, the user will need to integrate their own understanding of annotation biases, and apply a critical lens in interpreting and acting on any results.

Tool Outputs REVISE automatically surfaces a variety of metrics that highlight unrepresentative or anomalous patterns in the dataset. To validate the usefulness of the tool, we have used it to analyze several datasets commonly used in computer vision: COCO (Lin et al., 2014), OpenImages (Krasin et al., 2017), YFCC100m (Thomee et al., 2016), and BDD100K (Yu et al., 2020). Some examples of the kinds of automatic insights our tool has found include:

- In the object detection dataset COCO (Lin et al., 2014), we use object bounding boxes to find that some objects, e.g., `airplane`, `bed` and `pizza`, are frequently large in the image. This is because fewer images of `airplanes` appear in the sky (far away; small) than on the ground (close-up; large). This may be a problem since object size plays a key role in recognition accuracy. One mitigation is to query for images of `airplane` appearing in scenes of `mountains, desert, sky`.
- In the OpenImages dataset (Krasin et al., 2017) we leverage gender annotations and people bounding boxes to find that labels exist for a large number of people who are too small in the image for human annotators to determine their gender; nevertheless, we found that annotators infer that they are `male` 69% of the time, especially in scenes of `outdoor sports fields, parks`. Computer vision researchers might want to exercise caution with these gender annotations so they don't propagate into the model.
- In the computer vision and multimedia dataset YFCC100m (Yahoo Flickr Creative Commons 100 million) (Thomee et al., 2016) we use geolocation annotations to find that images come from 196 different countries. However, we estimate that for around 47% of those countries—especially in developing regions of the world—the images are predominantly photos taken by

visitors to the country rather than by locals, potentially resulting in a stereotypical portrayal.

A benefit of our tool is that a user doesn't need to have specific biases in mind, as these can be hard to enumerate. Rather, the tool automatically surfaces unusual patterns. REVISE cannot automatically say which of these patterns, or lack of patterns, are problematic, and leaves that analysis up to the user's judgment and expertise. We note that "bias" is a contested term, and while our tool seeks to surface a variety of findings that are interesting to dataset creators and users, not all may be considered forms of bias by everyone.

Extension from Conference Version (Wang et al., 2020a) This work is an extension of prior work published at ECCV 2020 (Wang et al., 2020a). Our original work has raised awareness on the biases encoded in visual datasets (Steed & Caliskan, 2021; Fabbrizzi et al., 2021), as well as been used on new datasets (Zhao et al., 2021) that consider skin tone as the protected attribute. At the time of Zhao et al. (2021)'s analysis, our person-based measurements only worked for binary attributes, and thus their findings on skin tone are restricted to considering individuals as belonging to the categories of having either darker or lighter skin tone. However, with our new feature update that allows for ordinal attributes, we are able to measure more nuanced insights, as seen in Figs. 6 and 8. The major changes we have made in this extended work are as follows:

- Geography-based analyses are extended to incorporate external data sources like income and weather, with results on new BDD100k dataset in Sects. 6.1.5 and 6.1.6. They are also combined with the people-based analyses, with results in Sect. 6.1.3.
- Geography-based analyses now use customizable GeoJSON files to create a user-friendly interactive geographic interface, as shown in Fig. 3. Other user experience improvements include a video demonstration of select features on the Github repository, and the automatic generation of a static PDF to summarize the key findings on a dataset to accompany the dynamic tool.
- Person-based analyses are generalized to work beyond binary attributes. In addition, there is added functionality to perform regression analyses for attributes that are also ordinal, such as quantized age or skin tone, with results in Sects. 5.1.1, 5.1.2, and 5.1.3.
- Previous analyses have been strengthened: added automatic categorization of object labels to a corresponding supercategory with word embedding distances in order to make analyses more tractable to interpret (beginning of Sect. 4.1); incorporated permutation tests for statistical significance in person-based and geography-based analyses (Sect. 5.1.4).

2 Related Work

Data Collection Visual datasets are constructed in various ways, with the most common being through keyword queries to search engines, whether singular (e.g., ImageNet Russakovsky et al., 2015) or pairwise (e.g., COCO Lin et al., 2014), or by scraping websites like Flickr (e.g., YFCC100m Thomee et al., 2016, OpenImages Krasin et al., 2017). There is extensive preprocessing and cleaning done on the datasets. Human annotators, sometimes in conjunction with automated tools (Zhou et al., 2017), then assign various labels and annotations. Dataset collectors put in significant effort to deal with problems like long-tails to ensure a more balanced distribution, and intra-class diversity by doing things like explicitly seeking out non-iconic images beyond just the object itself in focus.

Dataset Bias Rather than pick a single definition, we adopt an inclusive notion of bias and seek to highlight ways in which the dataset builder can monitor and control the distribution of their data. Proposed ways to deal with dataset bias include cross-dataset analysis (Torralba & Efros, 2011) and having the machine learning community learn from data collection approaches of other disciplines (Jo & Gebru, 2020; Brown, 2014). Recent work Prabhu and Birhane (2020) has looked at dataset issues related to consent and justice; the authors advocate for enforcing Institutional Review Board (IRB) approval for large scale datasets. Beyond vision datasets, biases in datasets from other domains have also been interrogated, such as those from criminal justice (Bao et al. 2021), the U.S. Census (Ding et al., 2021), and vision language (Birhane et al., 2021). Although we have limited the scope of our work to the contents of visual datasets itself, there are much broader questions of fairness and justice to be considered regarding the role that datasets play (Denton et al., 2020; Paullada et al., 2020; Scheuerman et al., 2021; Peng et al., 2021). Constructive solutions will need to combine automated analysis with human judgement as automated methods cannot yet understand things like the historical context that led to an observed statistical imbalance in the dataset. Our work takes this approach by automatically supplying a host of new metrics for analyzing a dataset along with actions that can be taken to mitigate these findings. However, the responsibility lies with the user to select next steps. The tool is open-source, lowering the resource and effort barrier to creating ethical datasets (Jo & Gebru, 2020).

Computer Vision Tools Hoiem et al. (2012) built a tool to diagnose the weaknesses of object detector models in order to help improve them. More recently, there have been tools like TIDE (Bolya et al., 2020) which also surfaces object detection errors, as well as tools in the video domain (Alwassel et al.., 2018) looking into forms of dataset bias in activity recog-

nition (Sigurdsson et al., 2017). We similarly in spirit hope to build a tool that will, as one goal, help dataset curators be aware of the patterns and biases present in their datasets so they can iteratively make adjustments.

Algorithmic Fairness In addition to looking at how models trained on one dataset generalize poorly to others (Tommasi et al., 2015; Torralba & Efros, 2011), many more forms of dataset bias are being increasingly noticed in the fairness domain (Caliskan et al., 2017; Mehrabi et al., 2019; Yang et al., 2020). There has been significant work looking at how to deal with this from the algorithm side (Dwork et al., 2012; Khosla et al., 2012; Dwork et al., 2017; Wang et al., 2020b) with varying definitions of fairness (Kilbertus et al., 2017; Zhang et al., 2018; Pleiss et al., 2017; Gajane & Pechenizkiy, 2017; Hardt et al., 2016) that are often deemed to be mathematically incompatible with each other (Chouldechova, 2017; Kleinberg et al., 2017), but in this work, we look at the problem earlier in the pipeline from the dataset side.

Automated Bias Detectors To make the measurement of bias tractable, there have been many proposals of automated bias detectors. Many focus on measuring the outputs of models, for example, Facebook's Fairness Flow (Facebook, 2021) focuses on assessing model predictions for their contextual fairness, though also looks at the labels themselves. Similarly, IBM's AI Fairness 360 (Bellamy et al., 2018) discovers biases in machine learning models, and also looks into the datasets. However, its look into dataset biases is limited in that it first trains a model on that dataset, then interrogates this trained model with specific questions. Steed & Caliskan, (2021) is another work that quantifies bias in a model, in the form of biased associations present in unsupervised model representations. An increasingly popular line of work additionally uses Generative Adversarial Networks to create counterfactuals for the sake of quantifying bias in models (Denton et al., 2019; Sattigeri et al., 2019; Sharmanska et al., 2020; Balakrishnan et al.., 2020). Swinger et al. (2019) look at automatic detection of biases in word embeddings, but we look at patterns in visual images and their annotations. RUBi is an automated method for detecting unimodal model biases, but specifically for models trained in the Visual Question Answering (VQA) domain (Cadene et al., 2019). Different from much of this work, REVISE looks directly at the dataset and its annotations to discover more generalizable, model-agnostic patterns.

In terms of work that measures the bias of datasets, the Dataset Nutrition Label (Holland et al., 2018) is a recent project that assesses machine learning datasets. Differently, our approach works on visual rather than tabular data which allows us to use additional computer vision techniques, and goes deeper in terms of presenting a variety of graphs and statistical results. Amazon SageMaker Clarify (Amazon, 2021) also works to detect bias in training data, but only along the person-based axis, and not object nor geography. Similarly, Google's Know Your Data (Google People + AI Research, 2021) also aims to help mitigate bias issues in image datasets. However, their tool currently only works on TensorFlow image datasets, whereas REVISE will work for any local image dataset. This has the benefit of allowing dataset creators to iteratively query our tool during the development process of their dataset, as well as dataset users to apply it to a private or proprietary dataset.

3 Tool Overview

REVISE is a general tool intended to yield insights at varying levels of granularity. As input, it requires an image dataset and its associated annotations. Depending on the types of annotations available, the tool automatically computes a host of metrics, to be described in Sects. 4.1, 5.1, and 6.1, broken down by the axes of object, person, and geography.

The metrics are often situated to provide a user with anomalous patterns, such as when the size distribution of an object class is highly non-uniform, and correspondingly provide automatic data-driven insights on how one might correct for this distribution. However, a metric itself has no normative claim on its own; ultimately it is up to the user to determine whether the automatically-surfaced patterns deviate enough from an intended distribution that this would be a problem for the downstream application of models trained on the dataset.

Tool Practically, REVISE takes the form of a Jupyter notebook interface that allows interactive exploration, as shown in Figs. 2 and 3. For privacy reasons, all analyses are run on a user's local machine. By default, the code to compute the metrics are largely abstracted away. However, all code is open-sourced such that a user can perform any personal customization to the metrics to fit their intended use-case. After multiple rounds of iterations with potential users, we have added a number of features to increase adoptability and usability of the tool. For one, we have created a video that will demonstrate to potential users what the interface of the tool looks like, and allow them to get a sense of whether this would be useful for their purposes. We have also included a feature that automatically generates a summary PDF as a result of running the tool and exploring the notebook. This supplements the dynamic nature of the tool with a static component that summarizes the key findings. Additionally, the biggest hurdle for a user to get started with their own dataset is the process of setting their dataset up to be in the standardized format that our tool requires. To this end, we have created a comprehensive testing script that provides informative feed-

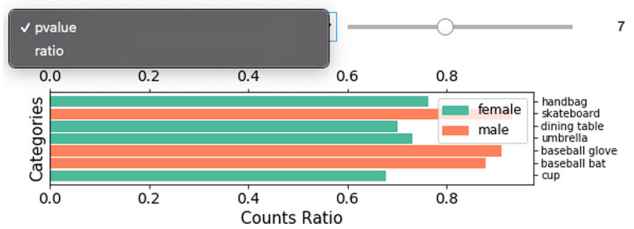

Fig. 2 Example interface of a metric in our notebook. A dropdown menus allow for sorting by p value or ratio, and a sliding bar allows adjusting the number of examples shown in the graph. Best viewed in color (Color figure online)

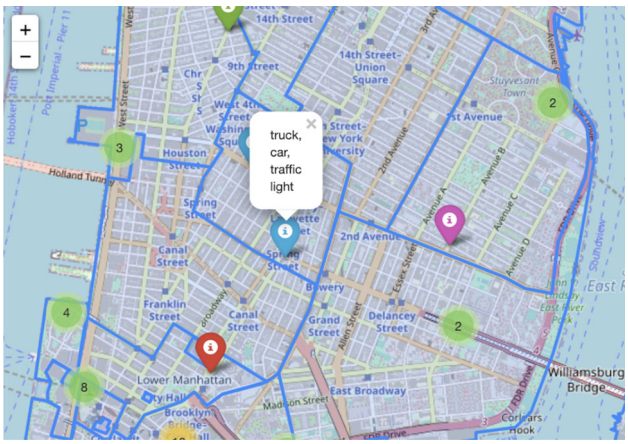

Fig. 3 Interface for exploring datasets with geography annotations. Interactive features allow viewing both image distribution by geography, as well as a bubble showing the labels of a specific image

back to ensure a user has inputted their dataset in the proper format before it is run through the tool.

Axes of Analyses The analyses that can ultimately be performed depend on the annotations available:

1. **Object-based insights** require instance labels and, if available, their corresponding bounding boxes and object category. Datasets are frequently collected together with manual annotations, but we also use automated computer vision techniques to infer some semantic labels, like scenes.
2. **Person-based insights** require sensitive attribute labels of the people in the images. The tool is general enough that given labels of any grouping of people, such as racial groups, the corresponding analyses can be performed. If the attribute labels are ordinal, such as quantized age or skin tone, additional regression analyses are available.
3. **Geography-based insights** flexibly allow for labels in two possible formats: (1) region labels as strings, e.g., "Portugal", "Nigeria", or (2) GPS latitude and longitude coordinates. By default the tool will use a global map, but users can override this with their own GeoJSON file.[1] The geography labels are analyzed in conjunction with the previously mentioned object and demographic labels, as well as external data source annotations. These can be at either the image-level, e.g., language of an image caption or region level, e.g., population size.

In the rest of the paper we will describe some insights automatically generated by our tool on various datasets, and potential actions that can be taken. Because of the interacting nature of biases along different axes, taking action through mitigation or rebalancing along one axis, e.g., gender, is likely to change the makeup of the dataset along another axis, e.g., geography. The user will need to exercise caution and determine which are the relevant axes to investigate, potentially applying the tool iteratively to observe for unintended side effects of interventions on the dataset. The metrics are all run fully automatically, but based on the statistically significant results that are surfaced by the tool, we pick out the interesting findings to present in this paper that demonstrate the flavor of insight each metric will provide.

4 Object-Based Analysis

We begin with an object-based approach to gain a basic understanding of a dataset. Much visual recognition research has centered on recognizing objects as the core building block (Everingham et al., 2010), and a number of object recognition datasets have been collected e.g., Caltech101 (Fei-Fei et al., 2004), PASCAL VOC (Everingham et al., 2010), ImageNet (Russakovsky et al., 2015; Deng et al., 2009). In Sect. 4.1 we introduce the metrics reported by REVISE; in Sect. 4.2 we dive into the actionable insights we surface as a result, all summarized in Table 1.

4.1 Object-Based Metrics

Of the metrics we will introduce, several (e.g., object counts, duplicate annotations, object scale) are commonly used by dataset collectors; others (e.g., scene or appearance diversity) are sometimes used during ad-hoc dataset examination but rarely quantified.

When the number of labels is very large (e.g., OpenImages contains 19,995) dataset analysis at the object level can be intractable to interpret. This motivates the need for higher-level *supercategories*: e.g., an `appliance` supercategory encompasses the more granular instances of `oven`,

[1] GeoJSON is a JSON-based standard for encoding boundary and region information through GPS data. GeoJSON files for many geographic regions are easily downloadable online, and can be readily converted from shapefiles, another type of geographic boundary file.

Table 1 Object-based summary: for image content and object annotations of COCO

Metric	Example insight	Example action
Object counts (Sect. 4.1.1)	Within the supercategory `appliance`, `oven` and `refrigerator` are overrepresented and `toaster` is underrepresented	Query for more `toaster` images
Duplicate annotations (Sect. 4.1.2)	The same object is frequently labeled as both `doughnut` and `bagel`	Manually reconcile the duplicate annotations
Object scale (Sect. 4.1.3)	`Airplane` is overrepresented as very large in images, as there are few images of airplanes smaller and flying in the sky	Query more images of `airplane` with `kite`, since they're more likely to have a small `airplane`
Object co-occurrences (Sect. 4.1.4)	`Person` appears more with unhealthy `food` like `cake` or `hot dog` than `broccoli` or `orange`	Query for more images of people with a healthier `food`
Scene diversity (Sect. 4.1.5)	`Baseball glove` doesn't occur much outside of sports fields	Query images of `baseball glove` in different scenes like a sidewalk
Appearance diversity (Sect. 4.1.6)	The appearance of `furniture` objects become more varied when they come from scenes like `water, ice, snow` and `outdoor sports fields, parks` rather than predominantly from `home or hotel`.	Query more images of `furniture` in `outdoor sports fields, parks`, since this scene is more common than `water, ice, snow`, and still contributes appearance diversity

`refrigerator`, and `microwave` in COCO (Lin et al., 2014). Most datasets, however, do not contain explicit mappings from labels to supercategories like COCO does. REVISE automatically bins labels to a set of predefined supercategories using the cosine similarity of word embeddings (Honnibal et al., 2020). Results of auto-generated mappings are returned to the user, sorted by confidence, and the user is free to override any of the mappings. In a random sample of labels from the OpenImages dataset mapped to the COCO supercategories, human validation finds this automatic binning strategy to be appropriate on 44 of 50 labels.

4.1.1 Object Counts

Object counts in the real world tend to naturally follow a long-tail distribution (Ouyang et al., 2016; Yang et al., 2014; Salakhutdinov et al., 2011). As for object counts in datasets, there are two main views: reflecting the natural long-tail distribution (e.g., in SUN Xiao et al., 2010) or approximately equal balancing (e.g., in ImageNet Russakovsky et al., 2015). Either way, the first-order statistic when analyzing a dataset is to compute the per-category counts and verify that they are consistent with the target distribution. By computing how frequently an object is represented both within its supercategory, as well as among all objects, this allows for a fine-grained look at frequency statistics: for example, while the `oven` and `refrigerator` objects fall below the median number of instances for an object class in COCO, it is nevertheless notable that both of these objects are around twice as frequent as the average object from the `appliance` class (Fig. 4).

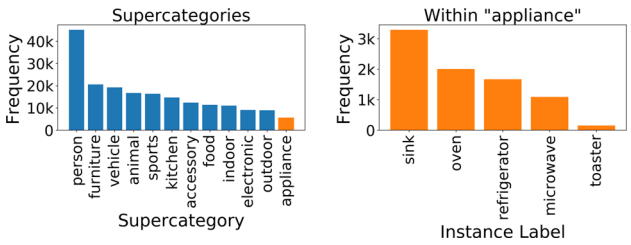

Fig. 4 `Oven` and `refrigerator` counts fall below the median of object classes in COCO; however, they are actually over-represented within the appliance category

4.1.2 Duplicate Annotations

A common issue with object dataset annotation is the labeling of the same object instance with two names (e.g., `cup` and `mug`), which is especially problematic in free-form annotation datasets such as Visual Genome (Krishna et al., 2016). In datasets with closed-world vocabulary, image annotation is commonly done for a single object class at a time causing confusion when the same object is labeled as both `trumpet` and `trombone` (Russakovsky et al., 2015). While these occurrences are manually filtered in some datasets, automatic identification of such pairs is useful for both dataset curators (to remove errors) and to dataset users (to avoid overcounting of either object). REVISE automatically identifies such object instances. In the OpenImages dataset (Krasin

et al., 2017) some examples of automatically detected pairs include bagel and doughnut, jaguar and leopard, and orange and grapefruit. In each case, the two labels are distinct (although visually similar) concepts, suggesting annotation errors.

4.1.3 Object Scale

It is well-known that object size plays a key role in object recognition accuracy (Russakovsky et al., 2015; Hoiem et al., 2012), as well as semantic importance in an image (Berg et al., 2012). While many quantizations of object scale have been proposed (Lin et al., 2014; Hoiem et al., 2012), we aim for a metric that is both comparable across object classes and invariant to image resolution to be suitable for different datasets. Thus, for every object instance we compute the fraction of image area occupied by this instance, and quantize into 5 equal-sized bins across the entire dataset. This binning reveals, for example, that rather than an equal 20% for each size, 77% of airplanes and 73% of pizzas in COCO are extra large ($> 9.3\%$ of the image area).

4.1.4 Object Co-occurrence

Object co-occurrence is a known contextual visual cue exploited by object detection models (Galleguillos et al., 2008; Oliva & Torralba, 2007), and thus can serve as an important measure of the diversity of object context. We compute all pairwise object class co-occurrence statistics within the dataset, and use them both to identify surprising co-occurrences as well as to generate potential search queries to diversify the dataset, as described in Sect. 4.2. For example, we find that in COCO, person appears in 43% of images containing the food category; however, person appears in a smaller percentage of images containing broccoli (15%), carrot (21%), and orange (29%), and conversely a greater percentage of images containing cake (55%), donut (55%), and hot dog (56%).

4.1.5 Scene Diversity

Building on quantifying the common context of an object, we additionally strive to measure the scene diversity directly. To do so, for every object class we consider the entropy of scene categories in which the object appears. We use a ResNet-18 (He et al., 2016) trained on Places (Zhou et al., 2017) to classify every image into one of 16 scene groups,[2] and identify objects like person that appear in a higher diversity of scenes versus objects like baseball glove that appear in fewer kinds of scenes (almost all baseball fields). This insight may guide dataset creators to further augment the dataset, as well as guide dataset users to want to test if their models can support out-of-context recognition on the objects that appear in fewer kinds of scenes, for example baseball gloves in a street.

4.1.6 Appearance Diversity

Finally, we consider the appearance diversity (i.e., intra-class variation) of each object class, which is a primary challenge in object detection (Yao et al., 2017). We use a ResNet-110 network (Idelbayev, 2019) trained on CIFAR-10 (Krizhevsky, 2009) to extract a 64-dimensional feature representation of every instance bounding box, resized to 32×32 pixels. We first validate that distances in this feature space correspond to semantically meaningful measures of diversity. To do so, on the COCO dataset we compute the average distance with $n = 500{,}000$ between two object instances of the same class (5.91 ± 1.44), and verify that it is smaller than the average distance between two object instances belonging to different classes but the same supercategory (6.24 ± 1.42), with a Cohen's D effect size of .23 and further smaller than the average distance between two unrelated objects (6.48 ± 1.44), with a Cohen's D effect size of .17. This metric allows us to identify individual object instances that contribute the most to the diversity of an object class, and informs our interventions in the next section.

4.2 Object-Based Actionable Insights

The metrics of Sect. 4.1 help surface biases or other issues, but it may not always be clear how to address them. We strive to mitigate this concern by providing examples of meaningful, actionable, and useful steps to guide the user.

For duplicate annotations, the remedy is straight-forward: perform manual cleanup of the data, e.g., as in "Appendix E" of Russakovsky et al. (2015). For the others the path is less straight-forward. For datasets that come from web queries, following the literature (Everingham et al., 2010; Russakovsky et al., 2015; Lin et al., 2014) REVISE defines search queries of the form "XX and YY," where XX corresponds to the target object class, and YY corresponds to a contextual term (another object class, scene category, etc.). REVISE ranks all possible queries to identify the ones that are most likely to lead to the target outcome, and we investigate this approach more thoroughly in "Appendix C".

For example, within COCO, airplanes have low diversity of scale and are predominantly large in the images. Our tool identifies that smaller airplanes co-occurred with objects like surfboard and scenes like mountains,

[2] Because top-1 accuracy for even the best model on all 365 scenes is 55.19%, but top-5 accuracy is 85.07%, we use the less granular scene categorization at the second tier of the defined scene hierarchy here. For example, aquarium, church indoor, and music studio fall into the scene group of indoor cultural.

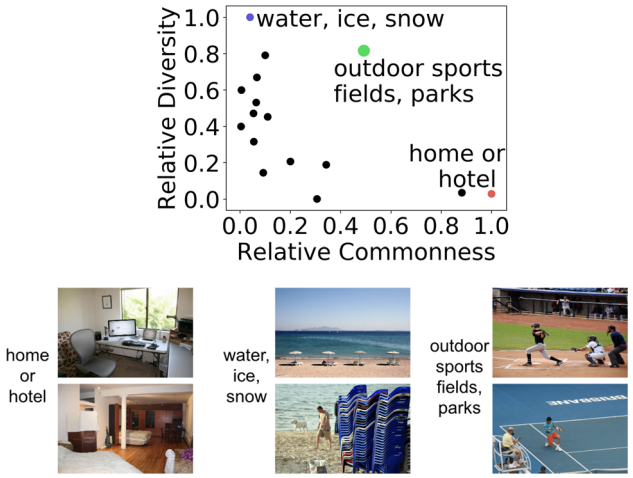

Fig. 5 The top shows the tradeoff for `furniture` in COCO between how much scenes increase appearance diversity (our goal) and how common they are (ease of collecting this data). To maximize both, `outdoor sports fields, parks` would be the most efficient way of augmenting this category. `Water, ice, snow` provides the most diversity but is hard to find, and `home or hotel` is the easiest to find but provides little diversity. On the bottom are sample images of `furniture` from these scenes

`desert, sky` (which are more likely to be photographed from afar). In other words, size matters by itself, but a skewed size distribution could also be a proxy for other types of biases. Dataset creators aiming to diversify their dataset towards a more uniform distribution of object scale can use these queries as a guide. These pairwise queries can similarly be used to diversify appearance diversity. `Furniture` objects appear predominantly in indoor scenes like `home or hotel`, so querying for furniture in scenes like `water, ice, snow` would diversify the dataset. However, this combination is quite rare, so we want to navigate the tradeoff between a pair's commonness and its contribution to diversity. Thus, we are more likely to be successful if we query for images in the more common `outdoor sports fields, parks` scenes, which also brings a significant amount of appearance diversity. The tool provides a visualization of this tradeoff (Fig. 5), allowing the user to make the final decision.

5 Person-Based Analysis

We next look into discrepancies in various aspects of how people of differing demographic attributes are represented, summarized in Table 2. The datasets we consider here are COCO (Lin et al., 2014), for which we have gender and skin tone annotations, and OpenImages (Krasin et al., 2017), for which we have gender annotations. In Sect. 5.1 we explain some of the metrics that we collect, and in Sect. 5.2 we discuss possible actions.

Gender Labels The gender labels in COCO are from Zhao et al. (2017), and their methodology in determining the gender for an image is that if at least one caption contains the word "man" and there is no mention of "woman", then it is a male image, and vice versa for female images. Zhao et al. (2021) compares the results of this schema to labels derived from external annotators. We use the same methodology along with other gendered labels like "boy" and "girl" on OpenImages' pre-existing annotations of individuals. It is important to acknowledge that the labels we are using are those of perceived binary gender, which is not inclusive of all gender categories. We will use the terms male and female to refer to binarized socially-perceived gender expression, and not gender identity nor sex assigned at birth, neither of which can be inferred from an image. In "Appendix A.1" we consider some of the problems that arise from using gender labels that have been inferred in this way.

Skin Tone Labels Our skin tone annotations for COCO come from Zhao et al. (2021), and are numbered 1-6 according to the Fitzpatrick scale (Fitzpatrick, 1988), where 1 is the lightest and 6 is the darkest. We use perceived skin tone as a poor proxy for race, and acknowledge that this risks reifying a particular inaccurate conception of race (Hanna et al., 2020). We consider skin tone as an ordinal variable, and analyze trendlines that result as we increase or decrease along this axis.

5.1 Person-Based Metrics

In this section, we will give representative findings for each metric that demonstrate the kind of insight our tool can provide. We start out by considering both gender and skin tone for COCO in Sects. 5.1.1 and 5.1.2, before transitioning to gender in OpenImages in Sects. 5.1.3 and 5.1.4.

5.1.1 Person Prominence

As our first line of analysis regarding how people of different demographic attributes are represented, we consider the proportion of an image a person takes up, as well as their distance from the center. We treat these two measures as a proxy for importance (Berg et al., 2012), where people who are larger and more to the center of an image are the focal point. We run the analysis for COCO on people differentiated both by gender and by skin tone. For gender, people who are male tend to take up more of the image ($.268 \pm .213$ for male vs $.138 \pm .148$ for female, with a Cohen's D effect size of .709) and be closer to the center ($.363 \pm .218$ for male vs $.510 \pm .250$ for female, with a Cohen's D effect size of .627). For people of different skin tones, in Fig. 6 we see that as

Table 2 Person-based summary: investigating representation of people with different demographic attributes

Metric	Example insight	Example action
Person Prominence (Sect. 5.1.1)	As the skin tone of the person in an image increases in darkness, the person is more likely to be smaller and further from the center.	Collect more images of people of different skintones as the subject of an image rather than in the background
Contextual representation (Sect. 5.1.2)	Male people occur in more outdoors scenes and with `sports` objects. Female people occur in more indoors scenes and with `kitchen` objects	Collect more images of female people in outdoors scenes with `sports` objects, and vice versa for male people
Instance counts and distances (Sect. 5.1.3)	In images with musical instrument `organ`, male people are more likely to be actually playing the `organ`	Collect more images of female people playing `organs`
Appearance differences (Sect. 5.1.4)	Male people in `sports uniforms` tend to be playing outdoor sports, while female people in `sports uniforms` are often indoors or in swimsuits.	Collect more images of each gender with `sports uniform` in their underrepresented scenes

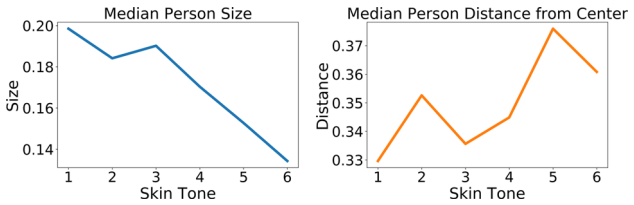

Fig. 6 In the COCO dataset, as a person's skin tone increases in darkness, that person is more likely to be smaller and further from the center. This indicates that people of darker skin tones are more likely to be in the background of an image rather than featured prominently. We used Jonckheere's trend test (Jonckheere, 1954) to show there is an a priori ordering to size and distance values by skin tone with p values of 2.11e−7 and .014, respectively

skin tone increases in darkness, the person is more likely to take up less of the image, as well as be further from the center. This indicates a bias against female people and people of darker skin tones as being less likely to be the focal point of an image.

5.1.2 Contextual Representation

Looking beyond just the person themselves, we consider the contexts that people with different demographic attributes tend to be featured in through the object groups they cooccur with, and the scenes they appear in. We first consider people of two different genders in COCO, and in Fig. 7 see that images with female people tend to be more indoors in scenes like `shopping and dining` and with object groups like `furniture`, `accessory`, and `appliance`. On the other hand, male people tend to be in more outdoors scenes like `sports fields` and `water, ice, snow`, and with object groups like `sports` and `vehicle`. These trends reflect gender stereotypes in many societies and can propagate into the models. While there is work on algorithmically intervening to break these associations, there are often too many proxy features to robustly do so. Thus it can be useful to intervene at the dataset creation stage.

Then, we consider these analyses in COCO along the ordinal variable of skin tone. In Fig. 8 we see statistically significant trends according to the Wald test on a non-zero slope of regression lines where people with lighter skin tones are more likely to be in `home or hotel` scenes and with object groups like `furniture`, and people with darker skin tones are more likely to be in `outdoor transportation` scenes and with object groups like `vehicle`. In the next metric we dig deeper into these object categories by considering the particular objects themselves.

5.1.3 Instance Counts and Distances

Analyzing object instances allows a more granular understanding of biases in the dataset. For example, in the previous metric on COCO we found `vehicle` objects to occur more with people of darker skin tones, and `furniture` with people of lighter skin tones. The specific `vehicle` objects that fit this trend are `motorcycle` and `bus`, while the specific `furniture` objects are `bed` and `couch`.

In OpenImages we find that objects like `cosmetics`, `doll`, and `washing machine` are overrepresented with female people, and objects like `rugby ball`, `beer`, `bicycle` are overrepresented with male people. However, beyond just looking at the number of times objects appear, we also look at the distance an object is from a person. We use a scaled distance measure as a proxy for understanding if a particular person, p, and object, o, are actually interacting with each other in order to derive more meaningful insight than just quantifying a mutual appearance in the same image. The distance measure we define is

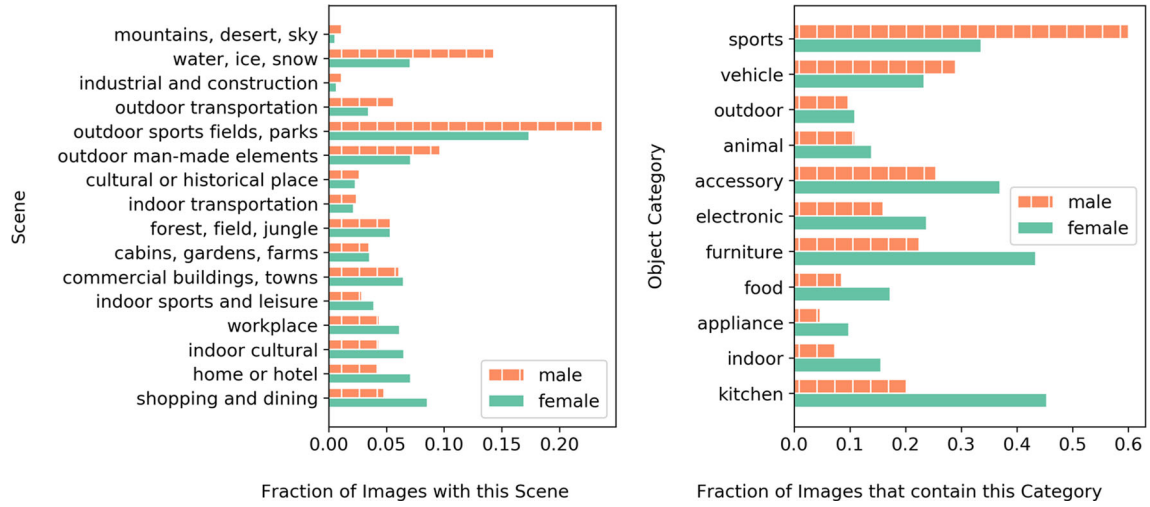

Fig. 7 Contextual information of images in COCO by gender, represented by fraction that are in a scene (left) and have an object from the category (right)

$$dist = \frac{\text{distance between p and o centers}}{\sqrt{\text{area}_\text{p} * \text{area}_\text{o}}} \quad (1)$$

to estimate distance in the 3D world, where area_p is measured on a normalized image of total area 1. In "Appendix B" we validate this notion that our distance measure can be used as a proxy interaction. We consider these distances in order to disambiguate between situations where a person is merely in an image with an object in the background, rather than directly interacting with the object, revealing biases that were not clear from just looking at the frequency differences. For example, `organ` (the musical instrument) did not have a statistically significant difference in frequency between the genders, but does in distance, or under our interpretation, relation. In Fig. 9 we investigate what accounts for this difference and see that when a male person is pictured with an organ, he is likely to be playing it, whereas a female person may just be near it but not necessarily directly interacting with it. Through this analysis we discover something more subtle about how an object is represented.

For these kinds of qualitative analyses, our tool necessarily can only serve as a focusing heuristic, as quantitative results such as the number of objects that have a statistically significant difference would be largely meaningless without human interpretation. In other words, we cannot automate the process of saying a statistical significant difference in distance is the result of a meaningful bias. However, by focusing attention on these statistically significant cases, our tool makes it actually tractable to measure for these biases.

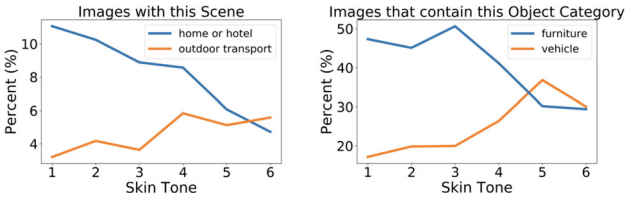

Fig. 8 We fit regression lines between co-occurrences of people with particular skin tones, and scenes and object categories. We show in the figure example categories where the Wald test has $p < .05$ that the slopes are non-zero, revealing trends that appear in image context as skin tone changes. On the left, we see that as an individual's skin tone increases in darkness, they are less likely to be pictured in `home or hotel` scenes, and more likely to be pictured in `outdoor transportation` scenes. On the right, we see that for object categories, people with darker skin tones are less likely to be pictured with `furniture` objects, and more likely to be pictured with `vehicle` objects

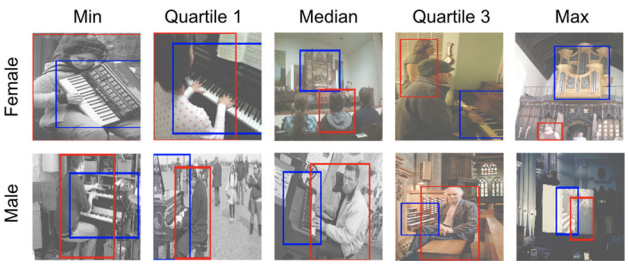

Fig. 9 5 images from OpenImages for a person (red bounding box) of each gender pictured with an organ (blue bounding box) along the gradient of inferred 3D distances. Male people tend to be featured as actually playing the instrument, whereas female people are oftentimes merely in the same space as the instrument (Color figure online)

5.1.4 Appearance Differences

We also look into the appearance differences in images of each gender with a particular object. This is to further disambiguate situations where occurrence counts, or even distances, aren't telling the whole story. This analysis is done by (1) extracting FC7 features from AlexNet (Krizhevsky

Fig. 10 Qualitative interpretation of what the visual model has learned for the `sports uniform` and `flower` objects between the two genders in OpenImages. "Confident Correct" are the images with the highest confidence scores

et al., 2012) pretrained on Places (Zhou et al., 2017) on a randomly sampled subset of the images to get scene-level features, (2) projecting them into $\sqrt{\text{number of samples}}$ dimensions (as is recommended in Hua et al., (2005), Jain & Waller, (1978)) to prevent over-fitting, and then (3) fitting a Linear Support Vector Machine with Sklearn's Pedregosa et al., (2011) default L2 regularization to see if it is able to learn a difference, as defined by classification accuracy of gender, between images of the same object with different genders. To make sure the female and male images are actually linearly separable and the classifier is not over-fitting, we randomly shuffle the data labels and use a permutation test to get statistical significance on our results (Ojala & Garriga, 2010). Our tool allows customizable sorting by both strength of the difference as well as statistical significance. In Fig. 10 we can see what the Linear SVM has learned on OpenImages for the `sports uniform` and `flower` categories. For `sports uniform`, male people tend to be represented as playing outdoor sports like baseball, while female people tend to be portrayed as playing an indoor sport like basketball or in a swimsuit. For `flower`, we see another drastic difference in how male and female people are portrayed, where the former are pictured with a `flower` in formal, official settings, and the latter are in staged settings or paintings.

5.2 Person-Based Actionable Insights

Compared to object-based metrics, the actionable insights for person-based metrics are less concrete and more nuanced. There is a tradeoff between attempting to represent the visual world as it is versus as we think it should be. For example, in contemporary societies, gender representation in various occupations, activities, etc. is unequal, so it is not obvious that aiming for gender parity across all object categories is the right approach. Biases that are systemic and historical are more problematic than others (Bearman et al., 2009), and this analysis cannot be automated. Further, the downstream impact of unequal representation depends on the specific models and tasks. Nevertheless, we provide some recommendations.

A trend that appeared in the metrics is that images frequently fell in line with common gender and racial stereotypes. Each group of people was under- or over-represented in a particular way, and dataset collectors may want to adjust their datasets to account for these by augmenting in the direction of the underrepresentations. Dataset users may want to audit their models, and investigate to what extent their models have learned the dataset's biases before they are deployed, as stereotypical correlations in the training data are likely to be amplified in model outputs (Zhao et al., 2017; Wang & Russakovsky, 2021).

6 Geography-Based Analysis

Finally, we turn to the geography of the images. We consider geography in the context of the object-based and person-based analyses from before, as well as additional axes. Geography uniquely interacts with both the types of objects that appear in images, as well as the demographics of the people. Because of these interactions, biases and problems around generalization have been shown to appear (Shankar et al., 2017; Gebru et al., 2017; DeVries et al., 2019).

In addition to COCO, for which we can derive geography labels on a subset of the images by querying the source of the images, i.e., Flickr, we also consider the global YFCC100m dataset[3] (Thomee et al., 2016), and the New York-centric BDD100K (Yu et al., 2020) self-driving car dataset.[4]

In Sect. 6.1 we present findings from our metrics, and in Sect. 6.2 we discuss what can be done about them (Table 3).

6.1 Geography-Based Metrics

In this section we analyze geography in the context of objects and people appearances, but also language, income, and weather. For distribution (Sect. 6.1.1), objects (Sect. 6.1.2), and language (Sect. 6.1.4) we look at the YFCC100m dataset, for people (Sect. 6.1.3) we look at COCO, and then for income (Sect. 6.1.5) and weather (Sect. 6.1.6) we look at BDD100K. Additionally, our analysis on geography by income is a case study into what our automated analyses in conjunction with an external data source of region-level labels may look like. One could also imagine plugging in a different external data source, e.g., region-level population size, and the tool would automatically run the same metrics along this axis instead.

[3] We use different subsets of the YFCC100m dataset depending on the particular annotations required by each metric.

[4] We consider the subset of the BDD100K dataset with images in New York City, which is a majority of the dataset.

Table 3 Geography-based summary: looking into the geo-representation of a dataset, and how that differs between different regions

Metric	Example insight	Example action
Geography distribution (Sect. 6.1.1)	Most images are from the USA, with very few from the countries of Africa	Collect more images from the countries of Africa
Geography by object (Sect. 6.1.2)	`Wildlife` is overrepresented in Kiribati, and `mosque` in Iran	Compare dataset frequencies to real-world frequencies; consider collecting other kinds of images representing these countries
Geography by people (Sect. 6.1.3)	Underrepresented regions like Africa and South Asia contain many of the images of people with darker skin tones	Collect more images from underrepresented regions to also diversify the people of different skin tones being represented
Geography by language (Sect. 6.1.4)	Countries in Africa and Asia that are already underrepresented are frequently represented by non-locals rather than locals	Collect more images taken by locals rather than visitors in underrepresented countries
Geography by income (Sect. 6.1.5)	Normalized by square mile, wealthier zip codes have more images, which also contain a different distribution of labels	Collect more images from zip codes with lower incomes
Geography by weather (Sect. 6.1.6)	Northern California has significantly less `snowy` images than New York City	Finetune a model on a weather distribution most similar to that in which it will be deployed

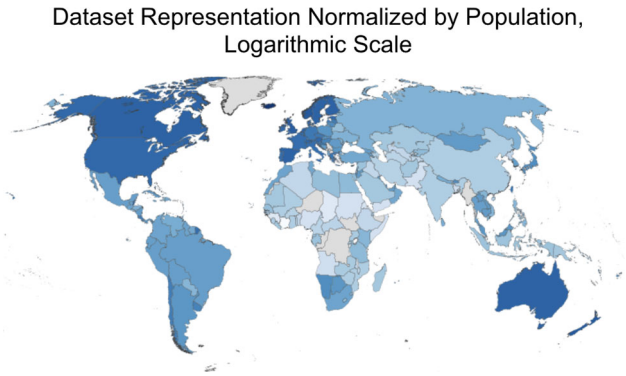

Fig. 11 Geographic distribution normalized by population in YFCC100m

6.1.1 Geographic Distribution

The first line of analysis is to look at the overall geographic distribution of a dataset. Researchers have looked at OpenImages and ImageNet and found these datasets to be Amerocentric and Eurocentric (Shankar et al., 2017), with models dropping in performance when being run on images from underrepresented locales. In Fig. 11 it immediately stands out that in the global YFCC100m dataset, the USA is drastically overrepresented compared to most other countries, with the continent of Africa being very sparsely represented. This can lead to generalization problems where a model may perform worse on image from a region it has not seen as much of Devries et al. (2019).

6.1.2 Geography by Object

In the YFCC100m dataset, we have access to image tags, which we treat as object labels. We combine our object-based analysis techniques with this geography data, allowing us to discern if certain labels are over- or under-represented between different areas. We then begin by considering the frequency with which each image tag appears in the set of a country's tags, compared to the frequency that same tag makes up in the rest of the countries. Some examples of over- and under-representations include Kiribati with `wildlife` at 86x, Iran with `mosque` at 30x, Egypt with `politics` at 20x, and United States with `safari` at .92x. We note that, as seen in the previous metric, this dataset is so skewed in terms of representation that most statistically significant underrepresentations are in the United States, as no other country has a high enough sample size. Additionally, whether these over- or under-representations are problematic enough to warrant intervention is entirely up to the user and their downstream task. We have normalized these tags by number of tag occurrences, and *not* by real-world distributions of the objects they mention, e.g., perhaps there are simply more `mosques` in Iran than other countries and this overrepresentation is innocuous and in fact a representative depiction of the country—it is up to the user to verify this.

We also look beyond the numbers themselves into the appearances of how different subregions, as defined by the United Nations geoscheme (United Nations Statistics Division, 2019), represent certain tags. Devries et al. (2019) showed that object-recognition systems perform worse on

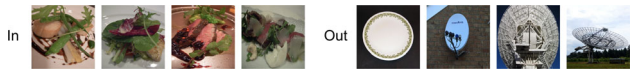

Fig. 12 A qualitative look at YFCC100m for what the visual model confidently and correctly classifies for images with the `dish` tag as in Eastern Asia, and out

images from countries that are not as well-represented in the dataset due to appearance differences within an object class, so we look into such appearance differences within a Flickr tag. We perform the same analysis as in Sect. 5.1 where we run a Linear SVM on the featurized images, this time performing 17-way classification between the different subregions. In Fig. 12 we show an example of the `dish` tag, and what images from the most accurately classified subregion, Eastern Asia, look like compared to images from the other subregions. Images with the `dish` tag tend to refer to food items in Eastern Asia, rather than a satellite dish or plate, which is a more common practice in other regions. This example is telling of a more pernicious problem than mis-identifying dishes, which is that of dialect differences between regions, and how that might affect the semantic meaning of a label. Disentangling homonyms will require computer vision systems to pay attention to the more subtle nuances of linguistics (Roll et al., 2018). It may be important to know if tags are represented differently across subregions so that models do not overfit to one particular subregion's representation of an object.

6.1.3 Geography by People

Next, we combine our COCO demographic skin tone annotations with geography labels. In Fig. 13 we see that images of people with darker skin tones tend to come from South Asia and Africa, but neither of these regions are very well-represented compared to images from the United States and Europe. In fact, while 85.5% of images of people with lighter skin tones (values 1–3) come from North America and Europe, this number is 58.2% for people with darker skin tones (values 4–6). Models that use this dataset may develop an understanding of people with darker skin tones that will be primarily informed by people from North America and Europe, which is a very small sample of people with darker skin tones in the world. Cultural practices differ among people across regions, and depending on the downstream application, it could be important that an understanding of a group not be informed only by the people in one geographic region.

For this particular dataset, we can additionally customize our tool to incorporate an external data source. Looking at the images only within the United States and binning by urban centers, as defined by the U.S. Census, we find that while 84.4% of images of people with lighter skin tones 1–3 are

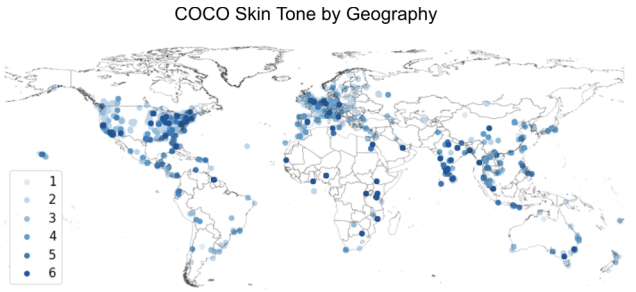

Fig. 13 Geographical distribution of COCO images based on skin tone annotations. Images from South Asia and Africa tend to contain people of darker skin tones, although the majority of images are coming from the United States and Europe

located in an urban area, 92.7% of images of people with darker skin tones 4–6 are located in an urban area.

6.1.4 Geography by Language

When we looked at the global distribution of the YFCC100m dataset, we saw an uneven distribution, with few images coming from countries in Africa and Asia. However, the locale of an image can be misleading, since if all the images taken in a particular country are only by tourists, this would not necessarily encompass the geo-representation one would hope for. Thus, here we combine our geography labels with language annotations. Fig. 14 shows the percentage of images taken in a country and captioned in something other than the national language(s), as detected by the fastText library (Joulin et al., 2016a, b). We use the lower bound of the binomial proportion confidence interval in the figure so that countries with only a few images total which happen to be mostly taken by tourists are not shown to be disproportionately imaged as so. Even with this lower bound, we see that many countries that are represented poorly in number are also under-represented by locals. To determine the implications in representation based on who is portraying a country, we categorize an image as taken by a local, tourist, or unknown, using a combination of language detected and tag content as an imperfect proxy. Harmful downstream effects of a lack of geographically diverse representation presented by locals has been shown in prior work (DeVries et al., 2019), which demonstrates how object recognition systems have trouble identifying common household items like `soap` and `spices` when they are not pictured in the traditionally Western way. We thus investigate if there are appearance differences in how locals and tourists portray a country by automatically running visual models. Although our tool does not find any such notable difference in the YFCC100m data, this kind of analysis can be useful on other datasets where a local's perspective is dramatically different than that of a tourist's.

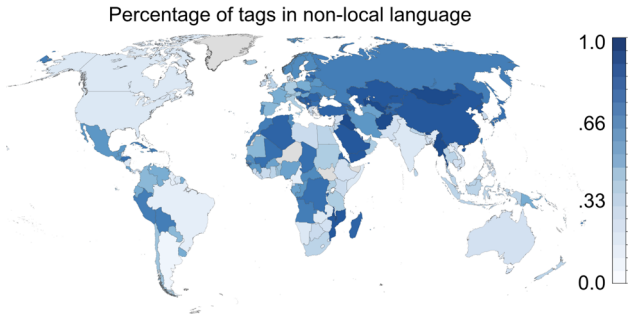

Fig. 14 Percentage of tags in a non-local language in YFCC100m. Even when underrepresented countries are imaged, it is not necessarily by someone local to that area

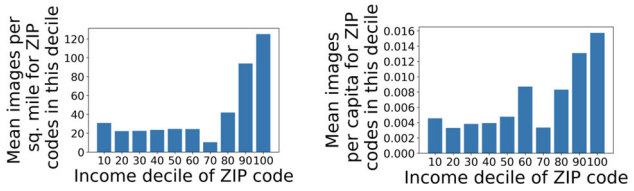

Fig. 15 ZIP codes with higher income are more represented in the BDD100K New York data

6.1.5 Geography by Income

Next, we consider how geography interacts with income. For this analysis, we focus on the portion of the BDD100K dataset in New York, and use income statistics by ZIP code (The United States Census Bureau, 2019; Keeping Track Online, 2019). This dataset was collected by crowdsourcing videos uploaded by drivers, a collection process that has the potential to introduce geographic or socioeconomic biases due to the self-selection of drivers.

To test whether this is the case, we divide the ZIP codes into deciles based on average income, and visualize how representation varies by income decline (Fig. 15). We see that there is a large difference in the number of images per square mile between the two wealthiest deciles and the rest. It is possible that some of this may be explained by the wealthier ZIP codes being in boroughs with a greater density of roads. Accordingly, we also visualize the mean images per capita rather than per square mile, and find that a large difference persists.

Such differences in representation can introduce biases or performance disparities in models trained on the data, because areas with different socioeconomic attributes are known to have systematic appearance differences (Gebru et al., 2017). As evidence of such appearance differences in the BDD100K data, we highlight in Fig. 16 that income correlates with the presence of both the `bicycle` and `pedestrian` label.

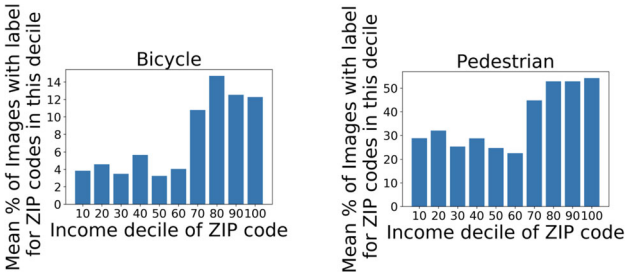

Fig. 16 ZIP codes with higher income are more likely to contain bicycles and pedestrians in the BDD100K New York data

6.1.6 Geography by Weather

In the BDD100K self-driving car dataset, we have access to weather tags on each image. Weather is a very relevant factor for this context of automated driving, as oftentimes datasets only contain weather in clear conditions (Sheeny et al., 2021), and thus have trouble generalizing to other weather conditions. Unsurprisingly, there are discrepancies between the weather distributions of images in the Northern California and New York City portions of this dataset, especially when looking at the `snowy` label, which is present at 0.3% for the former and 10% for the latter. It is important to be aware of these differences when deploying models in a setting different from the one they were trained in. We note that while we have distinguished between geography analyses by object and by weather, both are automatically run through the same technical functionality of the tool, as they are considering how the variation of per-image tags, i.e., object and weather labels, vary by region.

6.2 Geography-Based Actionable Insights

Much like the demographic-based actionable insights, those for geography-based are also more general and dependent on what the model trained on the data will be used for. Contextual knowledge about additional axes that may need to be considered, e.g., weather and language, can be easily incorporated into the tool as long as the appropriate annotations exist. This same contextual knowledge can help to navigate tensions that may occur if mitigation techniques for two different axes conflict with each other. Additionally, as we demonstrated with income, geography may often serve as a proxy for other characteristics, e.g., religion. In the absence of these kinds of annotations, geolocation can serve as a useful surrogate that may alert us to a socially significant discrepancy. Under- and over- representations can be approached in ways similar to before by augmenting the dataset, an important step in making sure we do not have a one-sided perspective of a region. Dataset users should validate that their models are not overfitting to a particular region's representation and image distribution by testing on

more geographically diverse data, especially on that which is representative of where a model will be deployed. The geographic distribution of a dataset is intricately linked to the representations of objects and the people in them. Because of this, we note that not all instances of distribution differences are problematic and certain findings of the tool, such as an underrepresentation of `safari` in the United States, may be entirely expected and not warrant any action to be taken. This will all depend on the use-case of the tested dataset.

It is clear that as we deploy more and more models into the world, there should be some form of either equal or equitable geo-representation. This emphasizes the need for data collection to explicitly seek out more diversity in locale, and specifically from the people that live there. Technology has been known to leave groups behind as it makes rapid advancements, and it is crucial that dataset representation does not follow this trend and base representation on digital availability. It requires more effort to seek out images from underrepresented areas, but as Jo & Gebru, (2020) discuss, there are actions that can and should be taken, such as explicitly collecting data from underrepresented geographic regions, to ensure a more diverse representation.

7 Discussion

REVISE is effective at surfacing and helping mitigate many kinds of biases in visual datasets. But we make no claim that REVISE will identify *all* visual biases. Creating an "unbiased" dataset may not be a realistic goal. The challenges are both practical (the sheer number of categories in modern datasets; the difficulty of gathering images from parts of the world where few people are online) and conceptual (how should we balance the goals of representing the world as it is and the world as we want it to be)?

The kind of interventions that can and should be performed in response to discovered biases will vary greatly depending on the dataset and applications. For example, for an object recognition benchmark, one may lean toward removing or obfuscating people that occur in images since the occurrence of people is largely incidental to the scientific goals of the dataset (Prabhu & Birhane, 2020; Yang et al., 2021). But such an intervention wouldn't make sense for a dataset used as part of a self-driving vehicle application. Rather, when a dataset is used in a production setting, interventions should be guided by an understanding of the downstream harms that may occur in that specific application (Barocas et al., 2019), such as poor performance in some neighborhoods. Making sense of which representations are more harmful for downstream applications may require additional data sources to help understand whether an underrepresentation is, for example, a result of a problem in the data collection effort, or simply representative of the world being imaged. Further, dataset bias mitigation is only one step, albeit an important one, in the much broader process of addressing fairness in the deployment of a machine learning system (Green & Hu, 2018; Birhane, 2021).

We also note that much of our analyses necessarily involves subdividing people along various socially-constructed dimensions. By operationalizing dynamic and non-discrete concepts such as gender and using skin tone as a proxy for race, we reify certain conceptions of these concepts (Hanna et al., 2020; Jacobs & Wallach, 2021) that harm certain groups, e.g., non-binary individuals (Scheuerman et al., 2020; Hamidi et al., 2018).

8 Conclusion

In conclusion, we present the REVISE tool, which automates the discovery of potential biases in visual datasets and their annotations. We perform this investigation along three axes: object-based, person-based, and
geography-based, and note that there are many more axes along which biases live. What cannot be automated is determining which of these biases are problematic and which are not, so we hope that by surfacing anomalous patterns as well as actionable next steps to the user, we can at least bring these biases to light.

Acknowledgements This work is partially supported by the National Science Foundation under Grant No. 1763642 and No. 1704444. We would also like to thank Felix Yu, Vikram Ramaswamy, and Zhiwei Deng for their helpful comments, and Zeyu Wang, Deniz Oktay, and Nobline Yoo for testing out the tool and providing feedback.

A Appendices

A.1 Gender Label Inference

An additional person-based metric we consider is gender label inference. Specifically, we note two especially concerning practices of assigning gender to a person who is too small to be identifiable, or no face is detected in the image. This is not to say that if these cases are not met it is acceptable to assign gender, as gender cannot be visually perceived by an annotator, but merely that assigning gender when one of these two cases is applicable is a particularly egregious practice. For example, it's been shown that in images where a person is fully clad with snowboarding equipment and a helmet, they are still labeled as male (Burns et al., 2018) due to preconceived stereotypes. We investigate the contextual cues annotators rely on to assign gender, and consider the gender of a person unlikely to be identifiable if the person is too small (below 1000 pixels, which is the number of dimensions that humans require to perform cer-

Fig. 17 Examples from OpenImages where annotators assigned gender to the person, but they should not have. The criteria used are that the person is either too small or has no face detected

tain recognition tasks in color images Torralba et al., 2008) or if automated face detection (we used Amazon Rekognition ("Amazon Rekognition" , n.d.), but note that any other face detection tool can be used) fails. For COCO, we find that among images with a human whose gender is unlikely to be identifiable, 77% are labeled male. In OpenImages,[5] this fraction is 69%. Thus, annotators seem to default to labeling a person as male when they cannot identify the gender; the use of male-as-norm is a problematic practice (Moulton, 1981). Further, we find that annotators are most likely to default to male as a gender label in `outdoor sports fields, parks` scenes, which is 2.9x the rate of female. Similarly, the rate for `indoor transportation` scenes is 4.2x and `outdoor transportation` is 4.5x, with the closest ratio being in `shopping and dining`, where male is 1.2x as likely as female. This suggests that in the absence of gender cues from the person themselves, annotators make inferences based on image context. In Fig. 17 we show examples from OpenImages where our tool determined that gender definitely should not be inferred, but was. Because attributes like skin tone can be inferred from parts of the image, such as a person's arm, we do not consider that attribute in this analysis.

This metric of gender label inference also brings up a larger question of which situations, if any, gender labels should ever be assigned (Scheuerman et al., 2020; Hamidi et al., 2018). However, that is outside the scope of this work, where we simply recommend that dataset creators should give clearer guidance to annotators, and remove the gender labels on images where gender can definitely not be determined. We note that while we picked out two criteria of when a person is too small and when there is no face detected to be instances in which gender inference is particularly egregious, there are many other situations that users may wish to delineate for their own purposes.

A.2 Validating Distance as a Proxy for Interaction

In Sect. 5.1, Instance Counts and Distances, we make the claim that we can use distance between a person and an object as a proxy for if the person, p, is actually interacting with the object, o, as opposed to just appearing in the same image with it. This allows us to get more meaningful insight as to how genders may be interacting with objects differently. The distance measure we define is $dist = \frac{\text{distance between p and o centers}}{\sqrt{\text{area}_p * \text{area}_o}}$, which is a relative measure within each object class because it makes the assumption that all people are the same size, and all instances of an object are the same size. To validate the claim we are making, we look at the SpatialSense dataset (Yang et al., 2019); specifically, at 6 objects that we hope to be somewhat representative of the different ways people interact with objects: `ball`, `book`, `car`, `dog`, `guitar`, and `table`. These objects were picked over ones such as `wall` or `floor`, in which it is more ambiguous what counts as an interaction. We then hand-labeled the images where this object cooccurs with a human as "yes" or "no" based on whether the person of interest is interacting with the object or not. We pick the threshold by optimizing for mean per class accuracy, where every distance below it as classified as a "yes" interaction and every distance above it as a "no" interaction. The threshold is picked based on the same data that the accuracy is reported for.

As can be seen in Table 4, for all 6 categories the mean of the distances when someone is interacting with an object is lower than that of when someone is not. This matches our claim that distance, although imperfect, can serve as a proxy for interaction. From looking at the visualization of the distribution of the distances in Fig. 18, we can see that for certain objects like `ball` and `table`, which also have the lowest mean per class accuracy, there is more overlap between the distances for "yes" interactions and "no" interactions. Intuitively, this makes some sense, because a `ball` is an object that can be interacted with both from a distance and from direct contact, and for `table` in the labeled examples, people were often seated at a table but not directly interacting with it.

A.3 Pairwise Queries

In Sect. 4.2, another claim we make is that pairwise queries of the form "`[Desired Object]` and `[Suggested Query Term]`" could allow dataset collectors to augment their dataset with the types of images they want. One of the examples we gave is that if one notices the images of `airplane` in their dataset are overrepresented in the larger sizes, our tool would recommend they make the query "`airplane` and `surfboard`" to augment their dataset, because based on the distribution of training samples, this combination is more likely than other kinds of queries to lead to images of smaller airplanes.

However, there are a few concerns with this approach. For one, certain queries might not return any search results.

[5] Random subset of size 100,000.

Table 4 Distances are classified as "yes" or "no" interaction based on a threshold optimized for mean per class accuracy

Object	# Labeled Ex.'s	Mean Per Class Acc (%)	"Yes" Distance mean ± std	"No" Distance mean ± std	Threshold
Ball	107	67	6.16 ± 2.64	8.54 ± 4.15	7.63
Book	27	78	2.45 ± 1.99	4.84 ± 2.24	3.88
Car	135	71	2.94 ± 3.20	4.59 ± 2.97	2.74
Dog	58	71	1.08 ± 1.12	2.07 ± 1.79	0.60
Guitar	40	88	0.90 ± 1.77	2.13 ± 1.21	1.61
Table	76	67	1.88 ± 1.19	3.28 ± 2.45	2.47

Visualization of the classification in Fig. 18. Distances for "yes" interactions are lower than "no" interactions in all cases, in line with our claim that smaller distances are more likely to signify an interaction

This is especially the case when the suggested query term is a scene category, such as `indoor cultural`, in which the query "`pizza` and `indoor cultural`" might not be very fruitful. To deal with this, we can substitute the scene category, `indoor cultural`, for more specific scenes in that category, like `classroom` and `conference`, so that the query becomes something like "`pizza` and `classroom`". When the suggested query term involves an object, there is another approach we can take. In datasets like PASCAL VOC (Everingham et al., 2010), the set of queries used to collect the dataset is given. For example, to get pictures of `boat`, they also queried for `barge`, `ferry`, and `canoe`. Thus, in addition to querying, for example, "`airplane` and `boat`", one could also query for "`airplane` and `ferry`", "`airplane` and `barge`", etc.

Another concern is there might be a distribution difference between the correlation observed in the data and the correlation in images returned for queries. For example, just because `cat` and `dog` cooccur at a certain rate in the dataset, does not necessarily mean they cooccur at this same rate in search engine images. However, our query recommendation rests on the assumptions that datasets are constructed by querying a search engine, and that objects cooccur at roughly the same relative rates in the dataset as they do in query returns; for example, that because `train` cooccurring with `boat` in our dataset tends to be more likely to be small, in images returned from queries, `train` is also likely to be smaller if `boat` is in the image. We make an assumption that for an image that contains a `train` and `boat`, the query "`train` and `boat`" would recover these kinds of images back, but it could be the case that the actual query used to find this image was "coastal transit." If we had access to the actual query used to find each image, the conditional probability could then be calculated over the queries themselves rather than the object or scene cooccurrences. It is because we don't have these original queries that we use cooccurrences to serve as a proxy for recovering them.

To gain some confidence in our use of these pairwise queries in place of the original queries, we show qualitative examples of the results when searching on Flickr for images

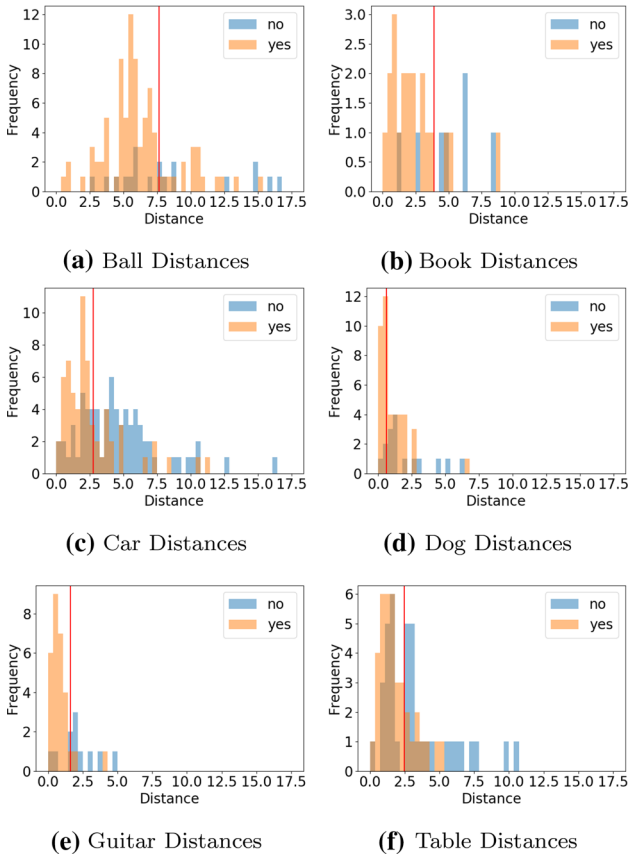

Fig. 18 Distances for the objects that were hand-labeled, orange if there is an interaction, and blue if there is not. The red vertical line is the threshold along which everything below is classified as "yes", and everything above is classified as "no"

that contain the tags of the object(s) searched. We show the results of querying for (1) just the object (2) the object and query term that we would hope leads to more of the object in a smaller size, and (3) the object and query term that we would hope leads to more of the object in a bigger size. In Figs. 19 and 20 we show the results of images sorted by relevance under the Creative Commons license. We can see that when we perform these pairwise queries, we do indeed have some level of control over the size of the object in the

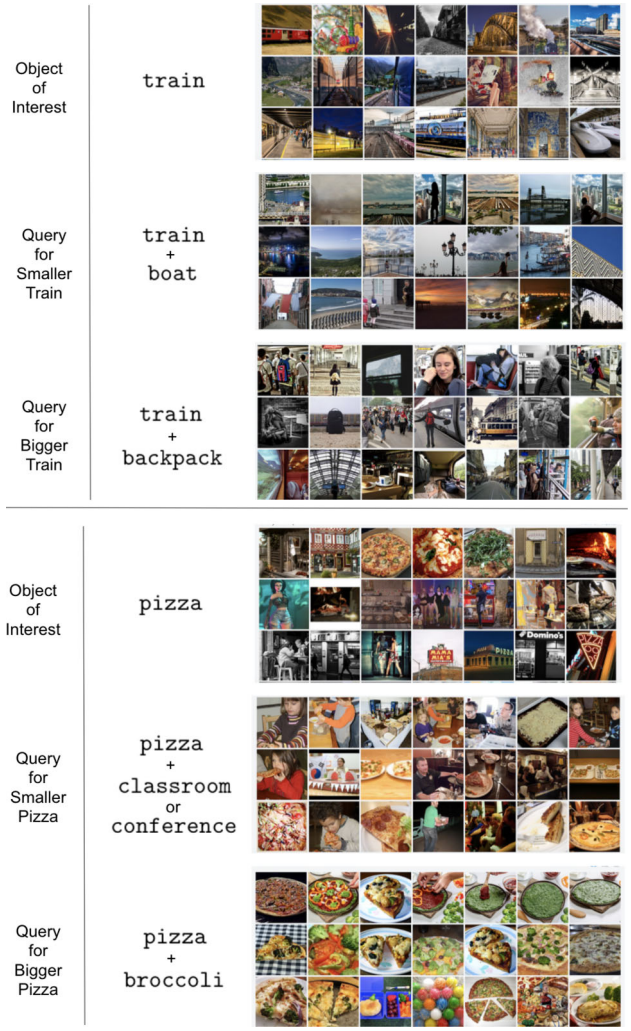

Fig. 19 Screenshots of top results from performing queries on Flickr that satisfy the tags mentioned. For `train`, when it is queried with `boat`, the `train` itself is more likely to be farther away, and thus smaller. When queried with `backpack`, the image is more likely to show travelers right next to, or even inside of, a `train`, and thus show it more in the foreground. The same idea applies for `pizza` where it's imaged from further in the background when paired with an `indoor cultural` scene, and up close with `broccoli`

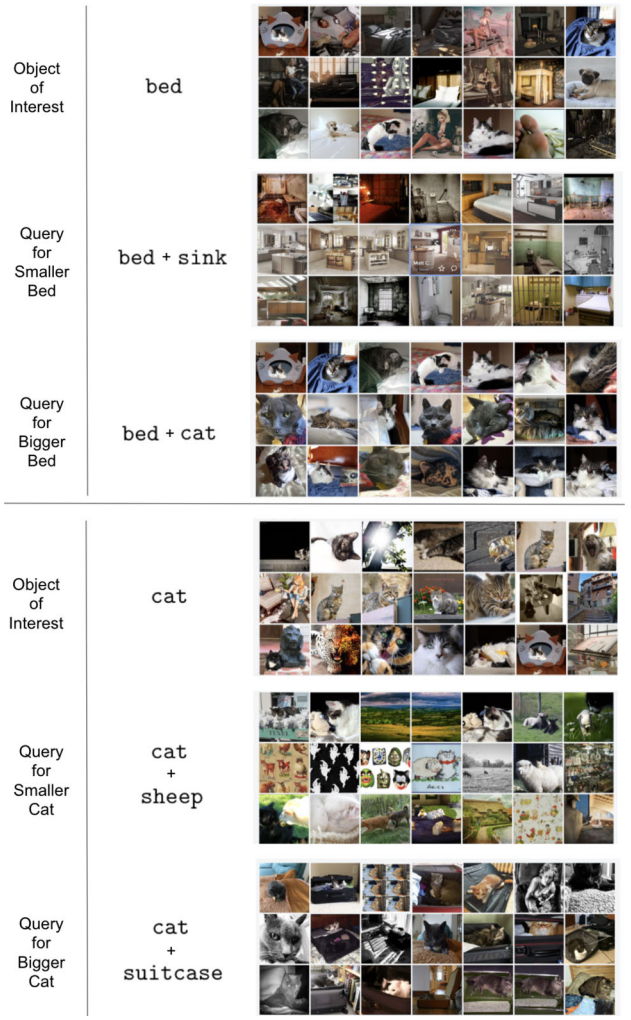

Fig. 20 Screenshots of top results from performing queries on Flickr that satisfy the tags mentioned. For `bed`, `sink` provides a context that makes it more likely to be imaged further away, whereas `cat` brings `bed` to the forefront. The same is the case when the object of interest is now `cat`, where a pairwise query with `sheep` makes it more likely to be further, and `suitcase` to be closer

resulting images. For example, "`pizza` and `classroom`" and "`pizza` and `conference`" queries (scenes swapped in for `indoor cultural`) return smaller pizzas than the "`pizza` and `broccoli`" query, which tends to feature bigger pizzas that take up the whole image. This could of course create other representation issues such as a surplus of `pizza` and `broccoli` images, so it could be important to use more than one of the recommended queries our tool surfaces. Although this is an imperfect method, it is still a useful tactic we can use without having access to the actual queries used to create the dataset.[6]

[6] We also looked into using reverse image searches to recover the query, but the "best guess labels" returned from these searches were not particularly useful, erring on both the side of being much too vague, such as returning "sea" for any scene with water, or too specific, with the exact name and brand of one of the objects.

References

Alwassel, H., Heilbron, F. C., Escorcia, V., & Ghanem, B. (2018). Diagnosing error in temporal action detectors. In *European conference on computer vision (ECCV)*.

Amazon. (2021). Amazon sagemaker clarify. Retrieved December 2, 2019, from https://aws.amazon.com/sagemaker/clarify/

Amazon rekognition. (n.d.). Retrieved December 2, 2019, from https://aws.amazon.com/rekognition/

Balakrishnan, G., Xiong, Y., Xia, W., & Perona, P. (2020). Towards causal benchmarking of bias in face analysis algorithms. In *European conference on computer vision (ECCV)*.

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., ... Venkatasubramanian, S. (2021). It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. arXiv:2106.05498.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. Retrieved December 2, 2019, from http://www.fairmlbook.org.fairmlbook.org.

Bearman, S., Korobov, N., & Thorne, A. (2009). The fabric of internalized sexism. *Journal of Integrated Social Sciences, 1*(1), 10–47.

Bellamy, R. K. E., Dey, K., Hend, M., Hoffman, S. C., Houde, S., Kannan, K., ... Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943.

Berg, A. C., Berg, T. L., III, H. D., Dodge, J., Goyal, A., Han, X., ... Yamaguchi, K. (2012). Understanding and predicting importance in images. In *Conference on computer vision and pattern recognition (CVPR)*.

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns, 2*, 100205.

Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv:2110.01963.

Bolya, D., Foley, S., Hays, J., & Hoffman, J. (2020). TIDE: A general toolbox for identifying object detection errors. In *European conference on computer vision (ECCV)*.

Brown, C. (2014). *Archives and recordkeeping: Theory into practices*. Facet Publishing.

Buda, M., Maki, A., & Mazurowski, M. A. (2017). A systematic study of the class imbalance problem in convolutional neural networks. arXiv:1710.05381.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM conference on fairness, accountability, transparency (FAccT)*.

Burns, K., Hendricks, L. A., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. In *European conference on computer vision (ECCV)*.

Cadene, R., Dancette, C., Ben-younes, H., Cord, M., & Parikh, D. (2019). RUBi: Reducing unimodal biases in visual question answering. In *Advances in neural information processing systems (NeurIPS)*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain humanlike biases. *Science, 356*(6334), 183–186.

Choi, M. J., Torralba, A., & Willsky, A. S. (2012). Context models and out-of-context objects. *Pattern Recognition Letters, 33*, 853–862.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 52*, 153–163.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR*.

Denton, E., Hanna, A., Amironesei, R., Smart, A., Nicole, H., & Scheuerman, M. K. (2020). Bringing the people back in: Contesting benchmark machine learning datasets. arXiv:2007.07399.

Denton, E., Hutchinson, B., Mitchell, M., Gebru, T., & Zaldivar, A. (2019). Image counterfactual sensitivity analysis for detecting unintended bias. In *CVPR workshop on fairness accountability transparency and ethics in computer vision*.

DeVries, T., Misra, I., Wang, C., & van der Maaten, L. (2019). Does object recognition work for everyone? In *Conference on computer vision and pattern recognition workshops (CVPRW)*.

Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. arXiv:2108.04884.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*.

Dwork, C., Immorlica, N., Kalai, A. T., & Leiserson, M. (2017). Decoupled classifiers for fair and efficient machine learning. arXiv:1707.06613.

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision (IJCV), 88*, 303–338.

Fabbrizzi, S., Papadopoulos, S., & Eirini Ntoutsi, I. K. (2021). A survey on bias in visual datasets. arXiv:2107.07919.

Facebook AI. (2021). Fairness flow. Retrieved from https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone/

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. In *IEEE CVPR workshop of generative model based vision*.

Fitzpatrick, T. B. (1988). The validity and practicality of sun-reactive skin types I through VI. *Archives of Dermatology, 6*, 869–871.

Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. arXiv:1710.03184.

Galleguillos, C., Rabinovich, A., & Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. In *Conference on computer vision and pattern recognition (CVPR)*.

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences of the United States of America (PNAS), 114*(50), 13108–13113. https://doi.org/10.1073/pnas.1700035114

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2018). Datasheets for datasets. In *ACM conference on fairness, accountability, transparency (FAccT)*.

Google People + AI Research. (2021). Know your data. Retrieved from https://knowyourdata.withgoogle.com/.

Green, B., & Hu, L. (2018). The myth in the methodology: Towards a recontextualization of fairness in machine learning. In *Machine learning: The debates workshop at the 35th international conference on machine learning*.

Hamidi, F., Scheuerman, M. K., & Branham, S. (2018). Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Conference on human factors in computing systems (CHI)*.

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. In *ACM conference on fairness, accountability, transparency (FAccT)*.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems (NeurIPS)*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *European conference on computer vision (ECCV)*.

Hill, K. (2020). Wrongfully accused by an algorithm. The New York Times. Retrieved from https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html.

Hoiem, D., Chodpathumwan, Y., & Dai, Q. (2012). Diagnosing error in object detectors. In *European conference on computer vision (ECCV)*.

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The dataset nutrition label: A framework to drive higher data quality standards. arXiv:1805.03677.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in python. https://doi.org/10.5281/zenodo.1212303

Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. R. (2005). Optimal number of features as a function of sample size for various classification rules. *Bioinformatics, 21*, 1509–1515.

Idelbayev, Y. (2019). Retrieved from https://github.com/akamaster/pytorchresnetcifar10

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. In *ACM conference on fairness, accountability, transparency (FAccT)*.

Jain, A. K., & Waller, W. (1978). On the optimal number of features in the classification of multivariate gaussian data. *Pattern Recognition, 10*, 365–374.

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *ACM conference on fairness, accountability, transparency (FAccT)*.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika, 41*, 133–145.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. *Human Factors in Computing Systems, 33*, 3819–3828.

Keeping Track Online. (2019). Median incomes. Retrieved from https://data.cccnewyork.org/data/table/66/median-incomes#66/107/62/a/a.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., & Torralba, A. (2012). Undoing the damage of dataset bias. In *European conference on computer vision (ECCV)*.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in neural information processing systems (NeurIPS)*.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In Proceedings of innovations in theoretical computer science (ITCS).

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-ElHaija, S., Kuznetsova, A., ... Murphy, K. (2017). Openimages: A public dataset for large-scale multilabel and multi-class image classification. Dataset available from https://github.com/openimages.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... Fei-Fei, L. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations. Retrieved from https://arxiv.org/abs/1602.07332

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical Report.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems (NeurIPS)* (pp. 1097–1105).

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., ... Dollar, P. (2014). Microsoft COCO: Common objects in context. In *European conference on computer vision (ECCV)*.

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, 39*, 539–550.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv:1908.09635.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2019). Model cards for model reporting. In *ACM conference on fairness, accountability, transparency (FAccT)*.

Moulton, J. (1981). The myth of the neutral 'man'. In *Sexist language: A modern philosophical analysis* (pp. 100–116).

Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research, 11*, 1833–1863.

Oksuz, K., Cam, B. C., Kalkan, S., & Akbas, E. (2019). Imbalance Problems in Object Detection: A Review. arXiv e-prints, arXiv:1909.00169. eprint: 1909. 00169

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*, 520–527.

Ouyang, W., Wang, X., Zhang, C., & Yang, X. (2016). Factors in finetuning deep model for object detection with long-tail distribution. In *Conference on computer vision and pattern recognition (CVPR)*.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2020). *Data and its (dis)contents: A survey of dataset development and use in machine learning research*. In *NeurIPS workshop: ML retrospectives, surveys, and meta-analyses*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Peng, K., Mathur, A., & Narayanan, A. (2021). Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in neural information processing systems (NeurIPS)*.

Prabhu, V. U., & Birhane, A. (2020). Large image datasets: A pyrrhic win for computer vision? arXiv:2006.16923.

Roll, U., Correia, R. A., & Berger-Tal, O. (2018). Using machine learning to disentangle homonyms in large text corpora. *Conservation Biology, 32*, 716–724.

Rosenfeld, A., Zemel, R., & Tsotsos, J. K. (2018). The elephant in the room. arXiv:1808.03305.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV), 115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Salakhutdinov, R., Torralba, A., & Tenenbaum, J. (2011). Learning to share visual appearance for multiclass object detection. In *Conference on computer vision and pattern recognition (CVPR)*.

Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., & Varshney, K. R. (2019). Fairness GAN. *IBM Journal of Research and Development, 63*, 3-1–3-9.

Scheuerman, M. K., Hanna, A., & Denton, E. (2021). Do datasets have politics? disciplinary values in computer vision dataset development. In ACM conference on computer-supported cooperative work and social computing (CSCW).

Scheuerman, M. K., Wade, K., Lustig, C., & Brubaker, J. R. (2020). How we've taught algorithms to see identity: Constructing race and gender in image databases for facial analysis. In *Proceedings of the ACM on human–computer interaction*.

Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., & Sculley, D. (2017). *No classification without representation: Assessing geodiversity issues in open datasets for the developing world*. In *NeurIPS workshop: Machine learning for the developing world*.

Sharmanska, V., Hendricks, L. A., Darrell, T., & Quadrianto, N. (2020). Contrastive examples for addressing the tyranny of the majority. arXiv:2004.06524.

Sheeny, M., Pellegrin, E. D., Mukherjee, S., Ahrabian, A., Wang, S., & Wallace, A. (2021). RADIATE: A radar dataset for automotive perception in bad weather. In *IEEE international conference on robotics and automation (ICRA)*.

Sigurdsson, G. A., Russakovsky, O., & Gupta, A. (2017). What actions are needed for understanding human actions in videos? In *International conference on computer vision (ICCV)*.

Steed, R., & Caliskan, A. (2021). Image representations learned with unsupervised pre-training contain human-like biases. In *Conference on fairness, accountability, and transparency (FAccT)*.

Swinger, N., De-Arteaga, M., IV, N. H., Leiserson, M., & Kalai, A. (2019). What are the biases in my word embedding? In *Proceedings of the AAAI/ACM conference on artificial intelligence, ethics, and society (AIES)*.

The United States Census Bureau. (2019). American community survey 1-year estimates, table s1903 (2005–2019). Retrieved from https://data.census.gov/.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Li, & L.-J. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM, 59*, 64–73.

Tommasi, T., Patricia, N., Caputo, B., & Tuytelaars, T. (2015). A deeper look at dataset bias. In *German conference on pattern recognition*.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Conference on computer vision and pattern recognition (CVPR)*.

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: A large dataset for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*(11), 1958–1970.

United Nations Statistics Division. (2019). United Nations statistics division - methodology. Retrieved from https://unstats.un.org/unsd/methodology/m49/.

van Miltenburg, E., Elliott, D., & Vossen, P. (2018). Talking about other people: An endless range of possibilities. In *International natural language generation conference*.

Wang, A., Narayanan, A., & Russakovsky, O. (2020). REVISE: A tool for measuring and mitigating bias in visual datasets. In *European conference on computer vision (ECCV)*.

Wang, A., & Russakovsky, O. (2021). Directional bias. In *International conference on machine learning (ICML)*.

Wang, Z., Qinami, K., Karakozis, Y., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Conference on computer vision and pattern recognition (CVPR)*.

Wilson, B., Hoffman, J., & Morgenstern, J. (2019). Predictive inequity in object detection. arXiv:1902.11097

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *Conference on computer vision and pattern recognition (CVPR)*.

Yang, J., Price, B., Cohen, S., & Yang, M.-H. (2014). Context driven scene parsing with attention to rare classes. In *Conference on computer vision and pattern recognition (CVPR)*.

Yang, K., Qinami, K., Fei-Fei, L., Deng, J., & Russakovsky, O. (2020). Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *ACM conference on fairness, accountability, transparency (FAccT)*.

Yang, K., Russakovsky, O., & Deng, J. (2019). Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *International conference on computer vision (ICCV)*.

Yang, K., Yau, J., Fei-Fei, L., Deng, J., & Russakovsky, O. (2021). A study of face obfuscation in imagenet. arXiv:2103.06191.

Yao, Y., Zhang, J., Shen, F., Hua, X., Xu, J., & Tang, Z. (2017). Exploiting web images for dataset construction: A domain robust approach. *IEEE Transactions on Multimedia, 19*, 1771–1784.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., ... Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM conference on AI, ethics, and society*.

Zhao, D., Wang, A., & Russakovsky, O. (2021). Understanding and evaluating racial biases in image captioning. In *CoRR*, arXiv:2106.08503.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*, 1452–1464.

Zhu, X., Anguelov, D., & Ramanan, D. (2014). Capturing long-tail distributions of object subcategories. In *Conference on computer vision and pattern recognition (CVPR)*.