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

Angelina Wang[0000−0001−9140−3523], Arvind Narayanan[0000−0001−7176−4479], and Olga Russakovsky[0000−0001−5272−3241]

Princeton University

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. To tackle this issue and to enable the preemptive analysis of large-scale dataset, we present our tool. REVISE (REvealing VIsual biaSEs) is a tool that assists in the investigation of a visual dataset, surfacing potential biases currently along three dimensions: (1) object-based, (2) gender-based, and (3) geography-based. Object-based biases relate to size, context, or diversity of object representation. Gender-based metrics aim to reveal the stereotypical portrayal of people of different genders. Geography-based analyses consider the representation of different geographic locations. REVISE sheds light on the dataset along these dimensions; 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 then 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: dataset bias, dataset analysis, computer vision fairness

1 Introduction

Computer vision dataset bias is a well-known and much-studied problem. In 2011, Torralba and Efros [60] 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 [9, 44, 47, 49, 54, 65], object contexts [13, 16, 52], object sub-categories [71], scenes [70], gender [10, 35], gender contexts [11, 69], skin tones [10, 63], geographic locations [16, 55] and cultures [16]. The downstream effects of these under-representations range from the more innocuous, like limited generalization of car classifiers [60], to the much more serious, like having deep societal implications in automated facial analysis [10, 26]. Efforts such as Datasheets for Datasets [23] 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.

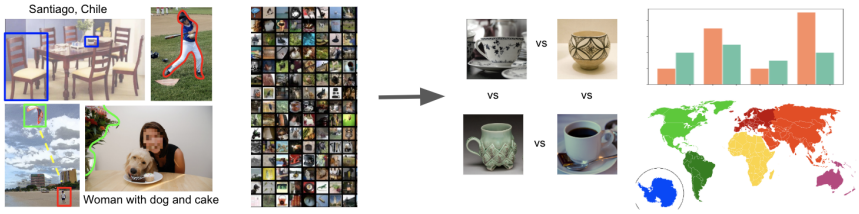

Fig. 1: Our tool takes in as input a visual dataset and its annotations, and outputs metrics, seeking to produce insights and possible actions.

However, this is only the beginning. It is often impossible to foresee the biases hiding in the data, and manual review is certainly 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. REVISE is a broad-purpose tool 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) gender-based and (3) geography-based. Object-based analysis is most familiar to the computer vision community [60], considering statistics about object frequency, scale, context, or diversity of representation. Gender-based analysis considers the representation of people of different genders within the dataset [23, 69]; such issues are gaining attention within the computer vision community. Future iterations of REVISE will include analysis of additional axes of identity. Finally, geography-based analysis considers the portrayal of different geographic regions within the dataset; this is a new but very important conversation within the community [55].

We imagine two primary use cases: (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.

Example Findings: 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 [43], OpenImages [39], YFCC100m [58]. Some examples of the kinds of automatic insights our tool has found include:

- In the object detection dataset COCO [43], 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). One way for the dataset creator to mitigate the problem is to query for images of `airplane` appearing in scenes of `mountains, desert, sky`.
- The OpenImages dataset [39] depicts 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, and

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) [58] 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. It is important to 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.

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 [53]) or pairwise (e.g., COCO [43]), or by scraping websites like Flickr (e.g., YFCC100m [58], OpenImages [39]). There is extensive preprocessing and cleaning done on the datasets. Human annotators, sometimes in conjunction with automated tools [70], 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 [60] and having the machine learning community learn from data collection approaches of other disciplines [8, 32]. Recent work [51] has looked at dataset issues related to consent and justice, and motivate enforcing Institutional Review Boards (IRB) approval for large scale datasets. Constructive solutions will need to combine automated analysis with human judgement as automated methods cannot yet understand things like the historical context of a 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 [32].

Computer vision tools: Hoiem et al. [27] built a tool to diagnose the weaknesses of object detector models in order to help improve them. More recently,

tools in the video domain [4] are looking into forms of dataset bias in activity recognition [56]. 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 [59, 60], many more forms of dataset bias are being increasingly noticed in the fairness domain [12, 45, 66]. There has been significant work looking at how to deal with this from the algorithm side [17, 18, 36, 62] with varying definitions of fairness [21, 24, 37, 50, 68] that are often deemed to be mathematically incompatible with each other [14, 38], but in this work, we look at the problem earlier in the pipeline from the dataset side.

Automated bias detectors: IBM's AI Fairness 360 [6] is an open-source toolkit that discovers biases in datasets and machine learning models. 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. On the other hand, REVISE looks directly at the dataset and its annotations to discover model-agnostic patterns. The Dataset Nutrition Label [28] 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. Swinger et al. [57] looks at automatic detection of biases in word embeddings, but we look at patterns in visual images and their annotations.

3 Tool Overview

REVISE is intended to be general enough to yield insights at varying levels of granularity, depending on the annotations available. We do use external tools and pretrained models [1, 30, 33, 34, 70] to derive some of our metrics, and acknowledge these models themselves may contain biases.

REVISE takes the form of a Jupyter notebook interface that allows exploration and customization of metrics. The analyses that can be performed depend on the annotations available:

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 are also beginning to use automated computer vision techniques to infer some semantic labels, like scenes.

Gender-based insights require gender labels of the people in the images. The tool is general enough that given labels of any groupings of people, such as racial groups, the corresponding analyses can be performed. In this paper we've limited our analyses to a grouping based on perceived binary gender because these labels already exist in the datasets we look at, even though it is not at all inclusive of all gender categories. We 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.

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

Metric	Example insight	Example action
Object counts	Within the supercategory `appliance`, `oven` and `refrigerator` are overrepresented and `toaster` is underrepresented	Query for more `toaster` images
Duplicate annotations	The same object is frequently labeled as both `doughnut` and `bagel`	Manually reconcile the duplicate annotations
Object scale	`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	`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	`Baseball glove` doesn't occur much outside of sports fields	Query images of `baseball glove` in different scenes like a sidewalk
Appearance diversity	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

Geography-based insights require country- or subregion-level annotations on where each image is taken. Information about the user who took each image would also be helpful — for example to determine if they were a local or a tourist. We do not have this user information in the datasets we analyze and instead infer it from the language and content of the tag captions.

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.

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 [19], and a number of object recognition datasets have been collected e.g., Caltech101 [20], PASCAL VOC [19], ImageNet [15, 53]. In Section 4.1 we introduce 6 such metrics reported by REVISE; in Section 4.2 we dive into the actionable insights we surface as a result, all summarized in Table 1.

4.1 Object-based Metrics

Object counts: Object counts in the real world tend to naturally follow a long-tail distribution [49, 54, 65]. But for datasets, there are two main views: reflecting the natural long-tail distribution (e.g., in SUN [64]) or approximately equal balancing (e.g., in ImageNet [53]). 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. Objects can also be grouped into hierarchical *supercategories*: e.g., an `appliance` supercategory encompasses the more granular instances of `oven`, `refrigerator`, and `microwave` in COCO [43]. By

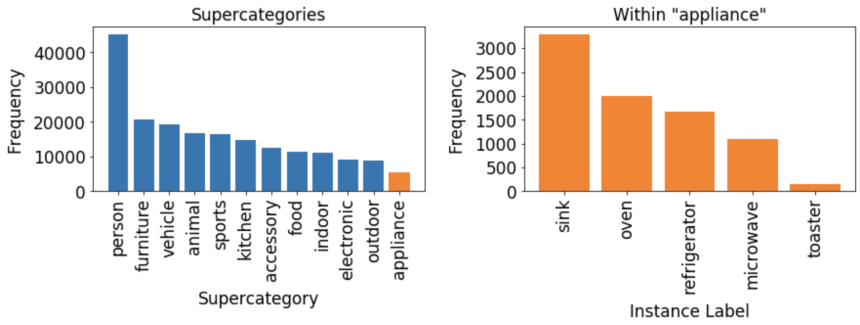

Fig. 2: `Oven` and `refrigerator` counts fall below the median of object classes in COCO; however, they are actually over-represented within the appliance category.

computing how frequently an object is represented both within its supercategory, as well as all objects, this allows for a fined-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 double as represented as the average object from the `appliance` class.

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 [40]. 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` [53]. 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 over-counting). REVISE automatically identifies such object instances, and in the OpenImages dataset [39] some examples of automatically detected pairs include `bagel` and `doughnut`, `jaguar` and `leopard`, and `orange` and `grapefruit`.

Object scale: It is well-known that object size plays a key role in object recognition accuracy [27, 53], as well as semantic importance in an image [7]. While many quantizations of object scale have been proposed [27, 43], 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).

Object co-occurrence: Object co-occurrence is a known contextual visual cue exploited by object detection models [22, 48], 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 Section 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%).

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 [25] trained on Places [70] to classify every image into one of 16 scene groups,[1] 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.

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 [67]. We use a ResNet-110 network [30] trained on CIFAR-10 [41] to extract a 64-dimensional feature representation of every instance bounding box, resized to 32x32 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) and further smaller than the average distance between two unrelated objects (6.48 ± 1.44). 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 Section 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 [53]. For the others the path

[1] 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`.

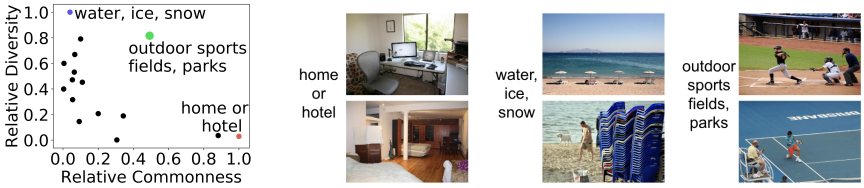

Fig. 3: The left shows the tradeoff for `furniture` in COCO between how much scenes increase appearance diversity (our goal) and how common they are (ease of collecting this data). To maximize both, `outdoor sports fields, parks` would be the most efficient way of augmenting this category. `Water, ice, 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 right are sample images of `furniture` from these scenes.

is less straight-forward. For datasets that come from web queries, following the literature [19, 43, 53] 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, 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. 3), allowing the user to make the final decision.

5 Gender-Based Analysis

We next look into potential discrepancies in various aspects of how each gender is represented, summarized in Table 2. The two datasets we have gender labels for are COCO and OpenImages. The gender labels in COCO are from [69], 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. We use the same methodology along with other gendered labels like "boy" and "girl" on OpenImages using

Table 2: Gender-based summary: investigating representation of different genders

Metric	Example insight	Example action
Contextual representation	Males occur in more outdoors scenes and with `sports` objects. Females occur in more indoors scenes and with `kitchen` objects.	Collect more images of females in outdoors scenes with `sports` objects, and vice versa for males.
Interactions	In images with musical instrument `organ`, males are more likely to be actually playing the `organ`.	Collect more images of females playing `organs`.
Appearance differences	Males in `sports uniforms` tend to be playing outdoor sports, while females in `sports uniforms` are often indoors or in swimsuits.	Collect more images of each gender with `sports uniform` in their underrepresented scenes.
Gender label inference	When gender is unlikely to be identifiable, people in images are by default labeled as male.	Prune these gender labels from the dataset so as not to reinforce societal stereotypes.

pre-existing annotations of individuals. In Section 5.1 we explain some of the metrics that we collect, and in Section 5.2 we discuss possible actions.

5.1 Gender-based Metrics

Contextual representation: We look into the contexts different genders tend to be featured in through object groups and scenes, with results in Appendix A.

Instance Counts and Distances: Analyzing the object instances themselves allows a more granualar understanding of gender biases in the dataset. In OpenImages we find that objects like `cosmetics`, `doll`, and `washing machine` are overrepresented with females, and objects like `rugby ball`, `beer`, `bicycle` are overrepresented with males. 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 $dist = \frac{\text{distance between p and o centers}}{\sqrt{\text{area}_p * \text{area}_o}}$ to estimate distance in the 3D world. 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

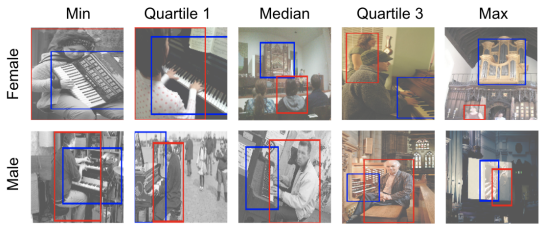

Fig. 4: 5 images from OpenImages for a person (red bounding box) of each gender pictured with an organ (blue bounding box) along the gradient of inferred 3D distances. Males tend to be featured as actually playing the instrument, whereas females are oftentimes merely in the same space as the instrument.

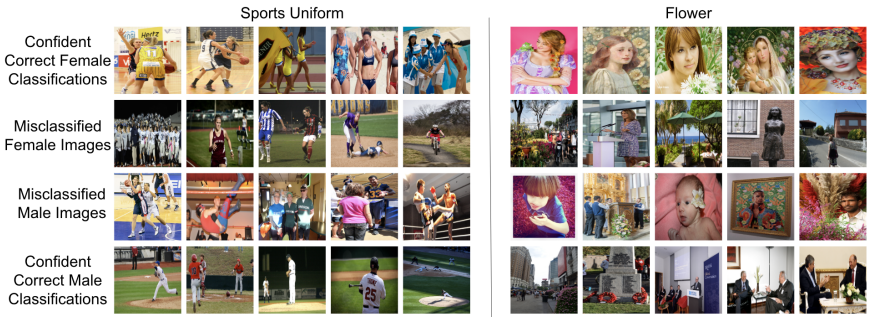

Fig. 5: Qualitative interpretation of what the visual model has learned for the `sports uniform` and `flower` objects between the two genders in OpenImages.

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. 4 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.

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 numbers, or even distances, aren't telling the whole story. This analysis is done by (1) extracting FC7 features from AlexNet [42] pretrained on Places [70] 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 [29, 31]) to prevent over-fitting, and then (3) fitting a Linear Support Vector Machine to see if it is able to learn a difference between images of the same object with different genders. To make sure the female and male images are actually linearly separable and the classifier isn't over-fitting, we look at both the accuracy as well as the ratio in accuracy between the SVM trained on the correctly labeled data and randomly shuffled data. In Fig. 5 we can see what the Linear SVM has learned on OpenImages for the `sports uniform` and `flower` categories. For `sports uniform`, males tend to be represented as playing outdoor sports like baseball, while females tend to be portrayed as playing an indoor sport like basketball or in a swimsuit. For `flower`, we see another drastic difference in how males and females are portrayed, where males pictured with a `flower` are in formal, official settings, whereas females are in staged settings or paintings.

Gender label inference: Finally, we examine the concerning practice of assigning gender to a person in the case where the person is far too small to be identifiable, or no face is even detected in the image. This is not to say that if these cases are not met it is acceptable to assign gender, 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 [11] 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 certain recognition tasks in color images [61]) or if automated face detection (using Amazon Rekognition [1]) fails. For COCO, we find that among images with a human whose gender is unlikely to be identifiable, 77% are labeled male. In OpenImages,[2] 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 [46]. 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.

5.2 Gender-based Actionable Insights

Compared to object-based metrics, the actionable insights for gender-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. 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. Gender biases that are systemic and historical are more problematic than others [5], 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 stereotypes. Each gender 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 look into to what extent their models have learned the dataset's biases before they are deployed.

For the metric of gender label inference, this brings up a larger question of which situations, if any, gender labels should ever be assigned. 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.

[2] Random subset of size 100,000

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

Metric	Example insight	Example action
Country distribution	Most images are from the USA, with very few from the countries of Africa	Collect more images from the countries of Africa
Local language analysis	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
Tag counts, appearances	`Wildlife` is overrepresented in Kiribati, and `mosque` in Iran	Collect other kinds of images representing these countries

6 Geography-Based Analysis

Finally, we look into the geography of the images, and the cultural biases that arise. We use the YFCC100m dataset [58] because of the geo-location data it contains. However, we use a different subset of the dataset for metrics that require more annotations, and explain each below.

6.1 Geography-based Metrics

Country distribution[3]**:** The first thing we look at is the geographical distribution of where images come from. Researchers have looked at OpenImages and ImageNet and found these datasets to be amerocentric and eurocentric [55], with models dropping in performance when being run on images from underrepresented locales. In the left side of Fig. 6 it immediately stands out that the USA is drastically overrepresented compared to other countries, with the continent of Africa being very sparsely represented.

Local language analysis[4]**:** 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. The right side of Fig. 6 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 [33, 34]. 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. We then investigate if there are appearance differences in how locals and tourists portray a country by

[3]Data subset: images with geo-location metadata

[4]Data subset: images with geo-location metadata and Flickr tags

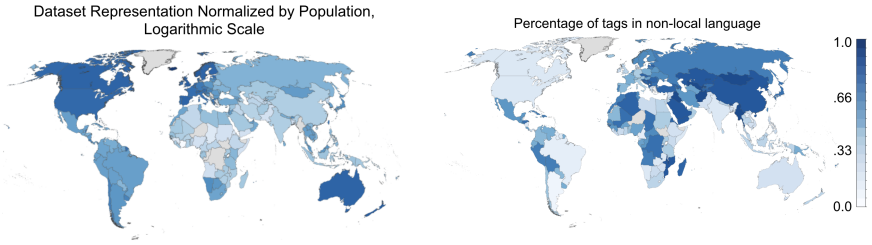

Fig. 6: Geographic distribution normalized by population (left) and percentage of tags in a non-local language (right) in YFCC100m. Even when underrepresented countries are imaged, it is not necessarily by someone local to that area.

automatically running visual models. Although our tool does not find any such notable difference, this kind of analysis can be useful on other datasets where a local's perspective is dramatically different than that of a tourist's.

Tag counts, appearances[5]**:** To gain insight into the actual content of what is being portrayed in images from country to country, we look at the tags assigned to each image. This allows us to discern if certain tags are over/under-represented between countries. We consider the frequency with which each 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, North Korea with `men` at 76x, Iran with `mosque` at 30x, Egypt with `politics` at 20x, and United States with `safari` at .92x. But because, as we've seen in previous sections, numbers don't always tell the full story, we also look into the appearances of how different subregions, as defined by the United Nations geoscheme [3], represent certain tags. DeVries et al. [16] showed that object-recognition systems perform worse on 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 Sec. 5.1 where we run a Linear SVM on the featurized images, this time performing 17-way classification between the different subregions. In Fig. 7 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. While this is a more innocent discrepancy, one could imagine how it may be important to know if other tags are represented differently across subregions so that models do not overfit to one particular subregion's representation of an object.

[5]Data subset: images with geo-location metadata and cleaned English tags from a list of 1540 from the Tag Prediction competition [2]. Because we are using this prexisting dataset of tags in order to meaningfully relate different images, we are excluding a large variety of images that have captions in a non-English language.

Fig. 7: 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.

6.2 Geography-based Actionable Insights

Much like the gender-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. 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 country. Dataset users should validate that their models are not overfitting to a particular country's representation by testing on more geographically diverse data. 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 et al. [32] 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 Conclusions

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, gender-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.

Acknowledgments

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, Zhiwei Deng, and Sadhika Malladi for their helpful comments, and Zeyu Wang, Deniz Oktay, and Nobline Yoo for testing out the tool and providing feedback.

References

1. Amazon rekognition https://aws.amazon.com/rekognition/
2. A Yahoo Flickr grand challenge on tag and caption prediction (2016), https://multimediacommons.wordpress.com/tag-caption-prediction-challenge/
3. United Nations statistics division - methodology (2019), https://unstats.un.org/unsd/methodology/m49/
4. Alwassel, H., Heilbron, F.C., Escorcia, V., Ghanem, B.: Diagnosing error in temporal action detectors. European Conference on Computer Vision (ECCV) (2018)
5. Bearman, S., Korobov, N., Thorne, A.: The fabric of internalized sexism. Journal of Integrated Social Sciences 1(1): 10-47 (2009)
6. Bellamy, R.K.E., Dey, K., Hend, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K.R., Zhang, Y.: AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943 (2018)
7. Berg, A.C., Berg, T.L., III, H.D., Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Sood, A., Stratos, K., Yamaguchi, K.: Understanding and predicting importance in images. Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
8. Brown, C.: Archives and recordkeeping: Theory into practices. Facet Publishing (2014)
9. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. arXiv:1710.05381 (2017)
10. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. ACM Conference on Fairness, Accountability, Transparency (FAccT) (2018)
11. Burns, K., Hendricks, L.A., Saenko, K., Darrell, T., Rohrbach, A.: Women also snowboard: Overcoming bias in captioning models. European Conference on Computer Vision (ECCV) (2018)
12. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain humanlike biases. Science **356**(6334), 183–186 (2017)
13. Choi, M.J., Torralba, A., Willsky, A.S.: Context models and out-of-context objects. Pattern Recognition Letters p. 853862 (2012)
14. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data (2017)
15. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
16. DeVries, T., Misra, I., Wang, C., van der Maaten, L.: Does object recognition work for everyone? Conference on Computer Vision and Pattern Recognition workshops (CVPRW) (2019)
17. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (2012)
18. Dwork, C., Immorlica, N., Kalai, A.T., Leiserson, M.: Decoupled classifiers for fair and efficient machine learning. arXiv:1707.06613 (2017)
19. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision (IJCV) (2010)

20. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. IEEE CVPR Workshop of Generative Model Based Vision (2004)
21. Gajane, P., Pechenizkiy, M.: On formalizing fairness in prediction with machine learning. arXiv:1710.03184 (2017)
22. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
23. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., III, H.D., Crawford, K.: Datasheets for datasets. ACM Conference on Fairness, Accountability, Transparency (FAccT) (2018)
24. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems (NeurIPS) (2016)
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. European Conference on Computer Vision (ECCV) (2016)
26. Hill, K.: Wrongfully accused by an algorithm. The New York Times (2020), `https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html`
27. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. European Conference on Computer Vision (ECCV) (2012)
28. Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K.: The dataset nutrition label: A framework to drive higher data quality standards. arXiv:1805.03677 (2018)
29. Hua, J., Xiong, Z., Lowey, J., Suh, E., Dougherty, E.R.: Optimal number of features as a function of sample size for various classification rules. Bioinformatics **21**, 15091515 (2005)
30. Idelbayev, Y.: (2019), `https://github.com/akamaster/pytorch_resnet_cifar10`
31. Jain, A.K., Waller, W.: On the optimal number of features in the classification of multivariate gaussian data. Pattern Recognition **10**, 365–374 (1978)
32. Jo, E.S., Gebru, T.: Lessons from archives: Strategies for collecting sociocultural data in machine learning. ACM Conference on Fairness, Accountability, Transparency (FAccT) (2020)
33. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016)
34. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
35. Kay, M., Matuszek, C., Munson, S.A.: Unequal representation and gender stereotypes in image search results for occupations. Human Factors in Computing Systems pp. 3819–3828 (2015)
36. Khosla, A., Zhou, T., Malisiewicz, T., Efros, A.A., Torralba, A.: Undoing the damage of dataset bias. European Conference on Computer Vision (ECCV) (2012)
37. Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., Schlkopf, B.: Avoiding discrimination through causal reasoning. Advances in Neural Information Processing Systems (NeurIPS) (2017)
38. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. Proceedings of Innovations in Theoretical Computer Science (ITCS) (2017)
39. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages (2017)

40. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalanditis, Y., Li, L.J., Shamma, D.A., Bernstein, M., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations (2016), https://arxiv.org/abs/1602.07332
41. Krizhevsky, A.: Learning multiple layers of features from tiny images. Technical Report (2009)
42. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems (NeurIPS) pp. 1097–1105 (2012)
43. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollar, P.: Microsoft COCO: Common objects in context. European Conference on Computer Vision (ECCV) (2014)
44. Liu, X.Y., Wu, J., Zhou, Z.H.: Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics (2009)
45. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. arXiv:1908.09635 (2019)
46. Moulton, J.: The myth of the neutral 'man'. Sexist Language: A Modern Philosophical Analysis pp. 100–116 (1981)
47. Oksuz, K., Cam, B.C., Kalkan, S., Akbas, E.: Imbalance Problems in Object Detection: A Review. arXiv e-prints p. arXiv:1909.00169 (Aug 2019)
48. Oliva, A., Torralba, A.: The role of context in object recognition. Trends in Cognitive Sciences (2007)
49. Ouyang, W., Wang, X., Zhang, C., Yang, X.: Factors in finetuning deep model for object detection with long-tail distribution. Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
50. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. Advances in Neural Information Processing Systems (NeurIPS) (2017)
51. Prabhu, V.U., Birhane, A.: Large image datasets: A pyrrhic win for computer vision? arXiv:2006.16923 (2020)
52. Rosenfeld, A., Zemel, R., Tsotsos, J.K.: The elephant in the room. arXiv:1808.03305 (2018)
53. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
54. Salakhutdinov, R., Torralba, A., Tenenbaum, J.: Learning to share visual appearance for multiclass object detection. Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
55. Shankar, S., Halpern, Y., Breck, E., Atwood, J., Wilson, J., Sculley, D.: No classification without representation: Assessing geodiversity issues in open datasets for the developing world. NeurIPS workshop: Machine Learning for the Developing World (2017)
56. Sigurdsson, G.A., Russakovsky, O., Gupta, A.: What actions are needed for understanding human actions in videos? International Conference on Computer Vision (ICCV) (2017)
57. Swinger, N., De-Arteaga, M., IV, N.H., Leiserson, M., Kalai, A.: What are the biases in my word embedding? Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES) (2019)
58. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM (2016)

59. Tommasi, T., Patricia, N., Caputo, B., Tuytelaars, T.: A deeper look at dataset bias. German Conference on Pattern Recognition (2015)
60. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
61. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: a large dataset for non-parametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(11), 1958–1970 (2008)
62. Wang, Z., Qinami, K., Karakozis, Y., Genova, K., Nair, P., Hata, K., Russakovsky, O.: Towards fairness in visual recognition: Effective strategies for bias mitigation. Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
63. Wilson, B., Hoffman, J., Morgenstern, J.: Predictive inequity in object detection. arXiv:1902.11097 (2019)
64. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
65. Yang, J., Price, B., Cohen, S., Yang, M.H.: Context driven scene parsing with attention to rare classes. Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
66. Yang, K., Qinami, K., Fei-Fei, L., Deng, J., Russakovsky, O.: Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. ACM Conference on Fairness, Accountability, Transparency (FAccT) (2020)
67. Yao, Y., Zhang, J., Shen, F., Hua, X., Xu, J., Tang, Z.: Exploiting web images for dataset construction: A domain robust approach. IEEE Transactions on Multimedia pp. 1771–1784 (2017)
68. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018)
69. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (2017)
70. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
71. Zhu, X., Anguelov, D., Ramanan, D.: Capturing long-tail distributions of object subcategories. Conference on Computer Vision and Pattern Recognition (CVPR) (2014)