Crowdsourcing Detection of Sampling Biases in Image Datasets

Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K. Thiruvathukal, Ming Yin*
Purdue University

ABSTRACT

Despite many exciting innovations in computer vision, recent studies reveal a number of risks in existing computer vision systems, suggesting results of such systems may be unfair and untrustworthy. Many of these risks can be partly attributed to the use of a training image dataset that exhibits sampling biases and thus does not accurately reflect the real visual world. Being able to detect potential sampling biases in the visual dataset prior to model development is thus essential for mitigating the fairness and trustworthy concerns in computer vision. In this paper, we propose a three-step crowdsourcing workflow to get humans into the loop for facilitating bias discovery in image datasets. Through two sets of evaluation studies, we find that the proposed workflow can effectively organize the crowd to detect sampling biases in both datasets that are artificially created with designed biases and real-world image datasets that are widely used in computer vision research and system development.

CCS CONCEPTS

• **Human-centered computing → Collaborative and social computing**.

KEYWORDS

sampling bias, crowdsourcing, image dataset, workflow design

ACM Reference Format:
Xiao Hu, Haobo Wang, Anirudh Vegesana, Somesh Dube, Kaiwen Yu, Gore Kao, Shuo-Han Chen, Yung-Hsiang Lu, George K. Thiruvathukal, Ming Yin. 2020. Crowdsourcing Detection of Sampling Biases in Image Datasets. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3366423.3380063

1 INTRODUCTION

Computer vision technologies have been applied to an increasingly wide range of applications from autonomous navigation, to medical image analysis, to precision agriculture [7, 9]. Despite many exciting innovations, recent studies reveal a number of risks in using existing computer vision systems, suggesting results of such systems may be unfair or untrustworthy. For example, major commercial facial analysis tools were shown to have substantial accuracy disparities for people of different genders or with different skin colors [3]. Visual semantic role labeling models were found to exhibit societal biases and stereotypes [23], such as frequently associating certain activity labels with specific gender (e.g., associate "cooking" with woman). Even worse, seemly accurate image classifiers may in fact have picked up spurious correlations between objects and irrelevant background information, rather than recognizing meaningful features of the objects [14].

Many of the risks embedded in modern computer vision systems can be partly attributed to the use of a training dataset that is *biased*. Indeed, the computer vision community has long recognized that many visual datasets present varying degrees of build-in bias due to factors such as photographic style of photographers and selection from dataset curators [18]. Using these biased datasets to train machine learning models for addressing different computer vision tasks naturally leads to the phenomenon of "bias in, bias out" and results in undesirable performance. Thus, to mitigate the fairness and trustworthy concerns in computer vision, it is critical to start the entire pipeline with high-quality visual datasets that, at least, are authentic representations of the visual world. In other words, being able to detect *sampling biases* of a dataset prior to developing models using the dataset is a key step in guarding against unfair or untrustworthy outcomes in computer vision.

While a few techniques have been developed to automatically detect dataset biases [19], the non-structured nature of visual data makes bias detection in image datasets particularly challenging. This is because no human-comprehensive attributes can be directly leveraged from the dataset to reason about the statistical associations between different features of the data. On the other hand, humans have the innate capability to understand images and identify patterns in images. This naturally leads us to ask, *can we leverage the wisdom of the crowd to detect sampling biases of image dataset?*

In this paper, we present such a human-in-the-loop approach to facilitate the bias detection in image datasets. Specifically, we present a crowdsourcing workflow which uses an image dataset provided by its curator as the input and outputs a list of statements by the crowd to represent sampling biases of the input image dataset. Our workflow will guide crowd workers to subsequently work on a series of three interconnected tasks: (1) inspect random samples of images from the input dataset and describe their similarity using a question-answer pair, (2) review separate random samples of images from the input dataset and provide answers to questions solicited from the previous step, and (3) judge whether statements of the image dataset that are automatically generated using the questions and answers collected accurately reflect the real world. This workflow is further augmented by back-end text processing techniques to deal with the noisy inputs from the crowd.

We conduct two studies to evaluate how the workflow enables the crowd to uncover biases, in both artificially-created image datasets with designed biases and real-world image datasets that have been frequently used in computer vision research and system

*S.-H. Chen is affiliated with Academia Sinica, and G. K. Thiruvathukal is affiliated with Loyola University. Correspondence should be directed to: hu440@purdue.edu, yunglu@purdue.edu, mingyin@purdue.edu.

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380063

development. Our results show that following our workflow, crowd workers can successfully discover all intentionally injected biases in the artificial datasets. More importantly, the crowd also effectively identifies a large number of meaningful sampling biases for real-world datasets, and their precision and recall in bias detection are 0.546 and 0.786, respectively, suggesting they are less precise but more complete than an individual expert in detecting biases.

2 RELATED WORK

The computer vision community has long recognized the influence of dataset biases on the performance of object detection and classification [11, 18]. Recent discussions on whether results produced by computer vision systems are appropriate and fair further draw attention to this issue and raise the awareness of the potential negative impact of dataset biases [3, 10, 23]. As a result, multiple efforts are undertaken to identify limitations of existing computer vision datasets and remedy issues that may result in problematic usage, including the biased representations of visual world in image datasets (e.g., [22]). So far, these efforts to improve the quality of image datasets are led by researchers, with relative limited participation from the general public, despite that the wisdom of the crowd has previously been utilized in many different tasks, including label annotation [16, 20], image segmentation [4], and semantic attribute generation [13, 17], to enhance computer vision systems.

In this paper, we explore a human-in-the-loop approach to actively engage the crowd to help detect biases in image dataset. The task of bias detection is inherently complex, and previous research in crowdsourcing has shown the success of decomposing complex tasks into small "micro-tasks" and engaging different crowds in working on different subtasks to collectively solve the grand problem [2, 5, 12]. Following this spirit, we decompose the bias detection task into a workflow of three interconnected steps: question generation, answer collection, and bias judgment. The design of the first step of our workflow—having the crowd to inspect image samples from a dataset and then generate question-answer pairs to describe similarity between these images—is inspired by both the recent research on visual question collection from the crowd to improve image understanding [1], and previous research on soliciting semantic attributes and clusters of images from the crowd [13, 17, 21].

3 CROWDSOURCING WORKFLOW

To leverage the wisdom of the crowd to detect potential biases hidden in image datasets, we propose a three-step crowdsourcing workflow. This workflow takes an image dataset that is provided by its curator as an input, and outputs a list of potential biases of this dataset. In this paper, we focus on detecting biases for image datasets that are constructed for facilitating the recognition of a particular type of object X (e.g., a dataset of car which enables a computer vision system to classify whether a car exists in an image). We further restrict on discovering *sampling biases* of image datasets (i.e., biases that are manifested as the image dataset fails to closely represent the real visual world). Figure 1 depicts our workflow, described in detail in the following subsections.

3.1 Step 1: Question Generation

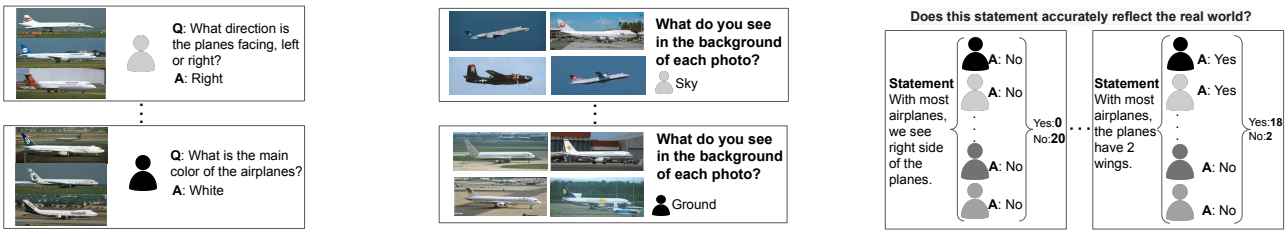
In essence, the problem of detecting sampling biases of an image dataset requires the identification of human-comprehensive attributes of images, and on these identified attributes, the distributions of attribute values observed within the given image dataset are *different* from those in the real visual world. For example, in a dataset of car images, "car type" can be considered as such a biased attribute if for the majority of images in this dataset, the type of car is sedan. A straightforward way to obtain these biased attributes is to have people (e.g., crowd workers) inspect the image dataset and find out attributes that contain biases. However, had the crowd workers been asked to inspect an entire set of image data (which often contains at least a few hundreds of images), they can be easily overloaded with the large number of images and may hardly find any meaningful biased attribute. Thus, to mitigate the information overload, we instead ask crowd workers to search for biased attributes by inspecting a *small portion* of images of the dataset, with the assumption that sampling biases for the entire dataset likely also exist in subsets.

Specifically, in each task of Step 1, we present a crowd worker with a set of n images that are randomly sampled from the input image dataset[1]. Workers are told that these images are collected to enable the automatic detection of the target object X (e.g., car), and they are asked to carefully inspect these images and find *similarities* between them. We intentionally ask workers to search for similarity between images rather than identifying biases, as similarity is an easier concept for laypeople to understand. In an early design of the workflow, we ask workers to provide names for those attributes on which they find similarity across the n images. We find, however, the quality of crowd-generated attributes following this design is not very high—crowd workers often input the names of common objects in the images or input attributes without sufficient explanations, which makes it difficult to interpret what exactly the attribute refers to (e.g., suggest "color" as an attribute without specifying whether it means the color of an object or the color of the background).

To solve this problem, inspired by recent efforts in collecting visual questions from the crowd [1], we redesign the first step and ask workers to describe the similarity they find across the n images using a *question-answer pair*, and the question in the pair is then used to characterize the attribute on which workers find similarity. This design allows us to obtain more contexts for the attribute, and thus confusion is decreased. More specifically, we ask workers to start their questions with "What," "Where," "When," or "How," as it has been showed that most questions generated by the crowd when describing images start with these interrogative words [1]. Workers are free to find similarities on any part of the images, including the objects and the background. We further instruct workers not to ask questions regarding the name or common characteristics of the target object X (e.g., "How many wheels does a car have?"). In other words, we nudge workers into identifying those "unusual" similarities across images which possibly imply biases.

Each worker is encouraged to generate as many unique questions as they can to describe the similarities among images shown to them. As different workers get different samples of the image dataset, the task of inspecting the entire image dataset is accomplished jointly by a group of workers. The output of Step 1, then, is a list of candidate biased attributes produced by the group of workers, in which each attribute is described through a question.

[1] n is a parameter of the workflow that can be tuned.



(a) Step 1: Question Generation **(b) Step 2: Answer Collection** **(c) Step 3: Bias Judgment**

Figure 1: A three-step crowdsourcing workflow to detect sampling biases presented in an image dataset.

Post-processing. Crowd workers may describe the same kind of similarity using different questions. To reduce the redundancy among the crowd-generated questions, we utilize spaCy, an open source natural language processing tool which is shown to have superior performance in dependency parsing [6], to conduct real-time text comparison and merge questions generated by workers. Specifically, given two questions, we remove all stop words in both questions and then compute a similarity score of the remaining sentences based on word embeddings using spaCy. These two questions will be merged if their similarity score is above a threshold, and the question with "higher quality"—quantified by having more noun phrases and dependent clauses—will be used to represent this group of questions. Through a pilot study, we find that the highest accuracy in question merging is achieved when merging two questions only if their similarity score is above 0.76. Thus, 0.76 is used as our similarity score threshold in determining whether to merge two questions.

3.2 Step 2: Answer Collection

Step 1 produces a list of *candidate* of biased attributes. However, similarities identified among n randomly sampled images may only capture "biases" within that particular sample, and further validation is needed to verify whether such a bias exists outside the specific sample (e.g., a worker may ask "What is the color of the car?" suggesting the color of car as a potential biased attribute, but in fact the worker may happen to have inspected a sample of mostly white cars). Thus, in Step 2, we use questions generated in Step 1 as inputs and collect answers to each of them based on different image data sampled from the given image dataset. Doing so, we can gauge on the value distributions for each of the attributes identified in Step 1 within the given image dataset, and thus filter those attributes that do not contain biases.

In particular, in each task, a worker will be presented with m images that are, again, randomly sampled from the input dataset[2], along with one question that is previously generated in Step 1. Workers are asked to carefully review the images and then answer the question using a simple word or phrase. If at least half of the m images share the same answer to the question, the worker is asked to enter that answer; otherwise, the worker can click a button to skip the question. By design, each question will be answered multiple times, each time with respect to a different sample of m images, to cover the entire dataset. Thus, together, Step 2 allows us to obtain a rough estimate of answer distribution for each of the questions generated in Step 1 within the given image dataset.

[2]Similar as before, m is another parameter of the workflow that can be tuned.

Post-processing. Similar as that in Step 1, for each question, workers may generate answers of similar meanings using different words. To reduce redundancy, we first enable auto-complete as a worker types the answer to a question, such that all existing answers to the *same* question will be shown as suggestions for the worker to consider as long as they contain the substring currently entered by the worker. Furthermore, after Step 2 is finished, we again use spaCy to identify similar answers to a question and merge them, and a list of final answers is produced for each question. The weight of each final answer to a question is then computed as the fraction of workers who provide that answer. Given a question, if the majority of workers choose to skip it, that means answers to this question are actually very diverse and therefore we consider it as not characterizing actual sampling bias of the dataset. On the other hand, if the highest weight is above a threshold τ for final answers to a question, the highest-weight answer will be selected and together with the question, they will be rephrased into a declarative statement s through a customized algorithm (e.g., "With most cars, they are family size."); the weight of the answer for this statement is denoted as w_s. Note that regardless of what w_s is, the rephrased declarative statement s always suggests that the selected answer is the *majority* answer by adding the part "With most X" where X is the name of the target object in the dataset (we will explain the rationale of this below in Section 3.3). The threshold τ can be set by the curators of the dataset to reflect the degree of biases that they are targeted at—the more they are interested in identifying attributes of images on which values are unbalanced to a smaller degree, the lower they should set the threshold τ.

3.3 Step 3: Bias Judgment

Finally, Step 3 takes the set of statements produced in Step 2 as inputs. Crowd workers are told that these statements describe a dataset of images of the target object X and are generated by previous workers inspecting the dataset. They are asked to judge, based on their common sense knowledge and subjective belief, whether each of the statements is true in the real visual world for images containing the target object X. To avoid biasing worker's mental model of the real visual world, we do not provide workers with any samples of the image dataset in this step. As mentioned earlier, each of the statements claims the "majority" value of an attribute for images in the dataset (e.g., "With most cars, they are family size." suggests the majority value of attribute "car size" is family size). Thus, given a statement s on a particular attribute, the *higher* fraction of workers indicating the statement as not accurately reflecting the real world (denote the fraction as f_s), the *more balanced* the real-world value distribution on this attribute should be. That

Figure 2: Inputs and outputs of the workflow in our two evaluation studies. Top panel: sample images of the image datasets used in Study 1 (the airplane dataset) and Study 2 (the car dataset); bottom panel: Top 10 "biases" with distinct meanings that are detected by the crowd using our workflow for each dataset. Each bias is coded into one of the 4 categories: Known bias (KB), additional bias (AB), unbiased similarity (US) or unrelated (U). KB and AB are considered correct detection of sampling biases (highlight in green), while US and U are considered incorrect detection (highlight in red).

is, sorting the statements in decreasing order of f_s, statements on attributes whose real-world value distributions are more balanced should rank higher in the list. This helps us to differentiate attributes that potentially reflect actual sampling biases of the dataset from attributes that describe common characteristics of the target object X and thus do not reflect biases (e.g., the attribute "number of wheels" for a car; the value distribution on this attribute should be very unbalanced and thus statements on this attribute should rank low on the list). Given that in a statement, the selected answer for the attribute is not always the majority answer with respect to the input dataset, the final output of Step 3 is a list of the statements sorted by the decreasing order of $f_s \cdot w_s$. In this way, statements that concern attributes whose real-world value distributions should be balanced (i.e., high f_s) *yet* whose value distributions within the given dataset are highly unbalanced (i.e., high w_s) will rank high on the list, hence the ranking of the statements roughly reflects the degree of biases. This list will be returned to dataset curators after removing the "With most X" part in each statement, so that further investigation can be conducted on the dataset with respect to those biases that are detected by the crowd.

3.4 Additional Workflow Control

As our workflow places significant requirement on the capability of reading and writing in English, we use a short English language test to filter workers who have limited English proficiency. For tasks in Step 1 and 2, workers are required to watch a short video which explains the interface and presents instructions on how to complete the task before they work on the task. On the other hand, task interface for Step 3 is straightforward so only textual instructions are provided. In addition, while our current design assumes that *all* images in the input dataset will be inspected collectively by a group of workers in Step 1, and each question produced in Step 1 will be answered with respect to samples of images that cover the *entire* dataset by another group of workers in Step 2, this needs not to be the case, especially for datasets of substaintially large scale. In fact, to increase the capability of the workflow to scale to large datasets, assuming sampling bias of a dataset also exists in its randomly-sampled subsets, we can restrict Step 1 and Step 2 on any manageable portion of the input dataset.

4 EVALUATION

To evaluate the effectiveness of the proposed workflow, we conduct two studies in which crowd workers are recruited from Amazon's Mechancial Turk (MTurk) to discover sampling biases in image datasets following our workflow.

4.1 Study 1: Uncover Injected Biases

First, as a proof of concept, we manually create a small image dataset in a way such that it exhibits sampling biases on a few attributes. Then, using this dataset as the input to our workflow, we explore whether these "injected" biases can be detected by the crowd, and how well the workflow works in general.

4.1.1 Dataset. We select 120 images from the class of *airplane* images of Caltech 101 [8] such that within this set of images, biases exhibit on 4 attributes:

- *Direction of airplane*: 100% of images in this dataset have airplanes pointing to the right.
- *Status of airplane*: 80% of images in this dataset have airplanes parking on the ground.
- *Size/Type of airplane*: 80% of images have airplanes as medium- or large-size commercial airplanes.
- *Color of airplane*: 70% of images have airplanes that are mostly white.

Figure 2 (upper left) shows some sample images in this dataset.

4.1.2 Experiment Procedure and Statistics. We post tasks for each step of our workflow as Human Intelligence Tasks (HITs) on MTurk. In particular, in Step 1, we set $n = 3$ and randomly divide the entire set of 120 images into 40 samples of 3 images. In total, 14 workers inspect these samples and produce 116 questions, which are eventually merged into a list of 42 questions by the NLP tool spaCy. In Step 2, we set $m = 4$ and then randomly divide the entire set of 120 images into 30 samples of 4 images. For each sample of 4 images, each of the 42 questions has been answered once by some

worker, and a total of 229 final answers are generated for the 42 questions by 44 workers. We set $\tau = 0$ to look for all potential biased attributes of the image dataset, regardless of how unbalanced the value distribution on the attribute is within the given dataset. Then, based on the answers collected in Step 2, 32 questions are finally rephrased into bias statements, each of which are reviewed by 20 workers in Step 3. Therefore, the output of Study 1 is a ranked list of 32 statements of the dataset which are believed by the crowd as representing sampling biases of the airplane dataset.

4.1.3 Results. We evaluate the performance of our workflow on whether it enables the crowd to detect reasonable biases of the input dataset, and how each step of the workflow facilitates the detection of these biases. In particular:

- Can the crowd identify meaningful biased attributes in Step 1?
- Through collecting answers to each question in Step 2, can we remove questions that do not actually reflect biases of the dataset?
- Does the ranking of statements produced in Step 3 reflect how likely each statement represents real sampling bias of the dataset?

First, we look into whether the crowd can uncover potential biases in image dataset through our workflow. Figure 2 (lower left) presents the top 10 "biases" with distinct meanings[3] that are detected by the crowd for the airplane dataset following our workflow. It is observed that the final list of 32 statements of dataset biases not only includes *all* 4 biases of the dataset that are intentionally designed by us, but also some additional biases of this dataset that are *not* due to our intentional design. For example, the crowd find that most airplane images are taken during the daytime, which indeed represent a bias of the dataset (118 out of 120 images are taken during the day). In other words, we confirm that using the proposed workflow to organize the crowd, the crowd *can* effectively detect biases in an image dataset.

We next aim to understand how each step of the workflow has facilitated the detection of biases. To see how Step 1 has allowed crowd workers to generate questions that describe potential biases and how Step 2 has enabled the filtering of questions that do not reflect actual biases, two co-authors of the paper code each of the 116 crowd-generated questions produced in Step 1, and each of the 32 statements produced in Step 2, into one of the 4 categories:

- *Known bias*: Biases that are intentionally injected by us to the dataset.
- *Additional bias*: Biases that actually exist in the dataset but are not intentionally injected by us.
- *Unbiased Similarity*: Common characteristics that are shared by typical airplanes which do not represent biases (e.g., "how many wings do planes have?").
- *Unrelated*: Questions or statements that describe neither actual sampling biases of the dataset nor similarity.

Intuitively, questions and statements that are categorized as "known bias" or "additional bias" are considered as *correct* detection of biases, while questions and statements that are coded as the other two categories are considered as incorrect detections.

More specifically, each annotator first codes the questions and statements independently, and the agreement between annotators is found to be high (i.e., Cohen's kappa values are 0.668 and 0.737, for the coding of 116 questions and 32 statements, respectively). Then, to aggregate the coding results, the two annotators discuss with each other to address categorizations for those questions and statements that they code differently, and produce a final category for each of them that both annotators agree. Figure 3a compares the distributions of categories for the 116 questions produced by crowd workers before merging, the 42 questions obtained after merging, and the 32 statements that are generated at the end of Step 2. It is found that in Step 1, before merging questions with similar meanings, 70.7% of the questions generated by the crowd (82 out of 116 questions) capture known biases or additional biases, and thus represent correct detection of dataset biases. Not surprisingly, the fraction of questions that reflect correct detection of biases decreases to 52.4% after question merging, as the merge mostly occurs when multiple questions are generated to describe the same type of biases. However, after Step 2, the percentage of statements that represent actual biases increases back to 65.6%—this is mainly because when collecting answers to each question, the crowd tends to skip answering those unrelated questions, and therefore a significant number of unrelated questions have been filtered and are not rephrased into a statement about the dataset.

Finally, to see whether the ranking produced in Step 3 provides reliable information on how likely each statement reflects actual sampling biases of the dataset, we compute the precision and recall for the top k ($1 \leq k \leq 32$) bias statements[4] and plot the precision-recall curve in Figure 3b. The area under this prevision-recall curve (AUPRC) is high (i.e., AUPRC=0.893), suggesting that Step 3 ranks correct bias statements higher on the output list—in fact, on the ranked list, 80% of the top 10 statements characterize correct biases while 60% of the bottom 10 statements capture incorrect biases.

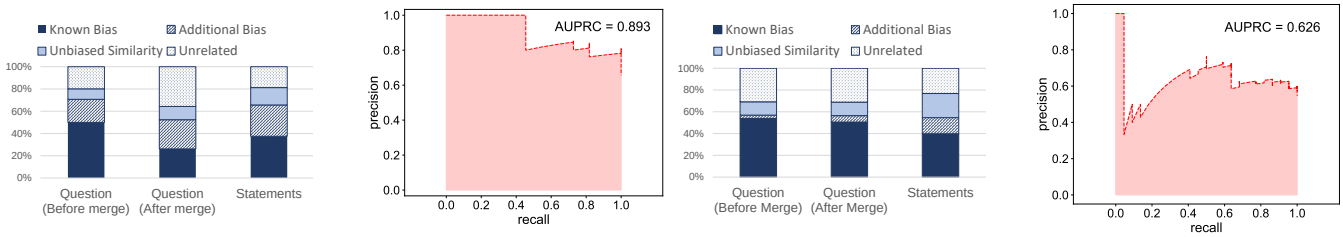
In sum, our evaluation in Study 1 suggests that with respect to the artificially created small image dataset of airplanes, our workflow can effectively guide the crowd to discover the biases of this dataset, and each step of the workflow serves as an essential component in facilitating the detection of these biases.

4.2 Study 2: Detect Bias in Real-World Datasets

We now move on to examine how our workflow performs when it is used to guide the crowd to detect biases in real-world image datasets. Compared to the artificial image dataset that we use in Study 1, detecting biases in real-world image datasets presents at least two new challenges. First, the size of real-world image datasets is often large, and detecting biases for these large image datasets places high requirement on the scalability of the workflow. Second, the sampling biases presented in real-world image datasets may be less extreme compared to the biases that we injected into our artificial dataset, so it is interesting to see whether the crowd can still detect these subtle biases using our workflow.

[3]Despite our best effort to remove redundancy, the list of 42 questions produced in Step 1 after merging still contains similar questions, which further results in bias statements with similar meanings in Step 3.

[4]When computing the precision and recall, we increase the recall at the k-th statement only if the k-th statement reflects a correct bias *and* this bias has not been described in any of the top $k-1$ statements.



(a) Study 1: categories **(b) Study 1: PR curve** **(c) Study 2: Categories** **(d) Study 2: PR curve**

Figure 3: Evaluation of workflow performance. 3a and 3c: distributions of correct and incorrect detection of biases for questions generated in Step 1 and statements produced in Step 2; 3b and 3d: precision-recall curves for the ranked list output in Step 3.

4.2.1 Dataset. For Study 2, we use the class of car images from ImageNet [15]—an image database that has been extensively used in computer vision research and product development—as our input dataset. This dataset contains a total of 1,300 car images, Figure 2 (upper right) shows samples of images in it.

To obtain the ground-truth biases of this dataset, we recruit 6 students who have conducted research on biases in image datasets for at least 2 years to serve as "experts" to inspect this dataset. Specifically, each expert gets a random 10% sample of images in the car dataset, and the expert is asked to independently review these images and record as many biases that they have noticed in this sample as possible. On average, each expert records 10.3 biases. An aggregated list of 23 biases is then generated by merging similar biases identified by individual experts, and two experts further go through this list to categorize whether each bias actually represents meaningful sampling bias of the dataset. It is found that 21 out of the 23 biases can be considered as correct biases. We thus use these 21 biases as the final list of expert-identified biases of the dataset.

4.2.2 Experiment Procedure and Statistics. Different from that in Study 1, in this Study, we only ask crowd workers to jointly inspect a *portion* of the input dataset (rather than the entire dataset) to increase the scalability of our workflow. In particular, in Step 1, we take a random 50% sample of the input dataset and divide them into samples of 3 images (i.e., $n = 3$). 44 workers inspect these samples and produce 620 questions, which have been further merged into 154 questions by spaCy. In Step 2, we then take a random 60% of images from the rest of the dataset (hence it is 30% of the entire input dataset) and divide them into samples of 4 images (i.e., $m = 4$). 478 workers take Step 2 tasks to review these 4-image samples and together, they generate 787 final answers to all 154 questions, among which 108 are rephrased into bias statements. Finally, the list of 108 statements is reviewed by 27 workers in Step 3, and therefore the output of Study 2 is a ranked list of 108 statements of the dataset which are believed by the crowd as representing sampling biases of the car dataset.

4.2.3 Results. Figure 2 (lower right) presents the top 10 "biases" with distinct meanings that are detected in this car dataset by the crowd following our workflow. We find that among the 21 expert-identified biases, 15 of them (71.4%) are also detected by the crowd, while the crowd also find 7 additional biases that is not detected by the experts (e.g., "No car have license plate on front").

Then, similar as before, we ask two of our experts to independently classify each of the 620 crowd-generated questions in Step 1 and each of the 108 statements produced in Step 2 into one of the 4 categories, where "known bias" here is defined as the expert-identified biases, and "additional bias" refers to biases that are identified only by the crowd but not by the experts. The annotation agreement between the two experts is relatively high (i.e., Cohen's kappa values are 0.690 and 0.632, for the coding of 620 questions and 108 statements, respectively). Again, after the independent coding, the two experts discuss with each other to address disagreement. Figure 3c shows final results of category distributions. Considering the *union* of known biases and additional biases as the correct biases, we find that in Step 1, 56.9% (or 56.5%) of the questions generated by the crowd before merging (or after merging) represent correct sampling biases of the dataset. Unrelated questions are effectively filtered through Step 2 (though questions characterizing unbiased similarity do not), and the percentage of statements that represent correct biases is 54.6% at the end of Step 2. Figure 3d further shows the precision-recall curve for the ranked list produced as the output of Study 2, and the curve has a relatively high AUPRC of 0.626, indicating the ranking is reasonable.

Finally, to understand how well the crowd performs in detecting dataset biases as compared to the experts, we compute the precision and recall for the 108 biases that are generated by the crowd following our workflow, and for the biases identified by each individual expert, separately. It is found that the crowd's precision and recall are 0.546 and 0.786, respectively, while on average, the precision and recall for an expert are 0.883 and 0.333. In other words, compared to an expert, biases detected for a real-world image dataset by the crowd following our workflow have a lower but acceptable precision, and they are much more complete.

5 CONCLUSIONS

In this paper, we present a crowdsourcing workflow to organize laypeople to detect potential sampling biases in an image dataset. Through two sets of evaluation studies, we find that the crowd can effectively discover many reasonable biases of image datasets using our workflow. Our results highlight the promise of bringing humans into the loop to improve the quality of visual datasets and increase the fairness and trustworthy of computer vision results. In the future, we are interested in conducting comparative studies to rigorously examine how parameters of the workflow (e.g., n, m and τ) influence its performance (i.e., precision and recall of the detected biases). Evaluations with substantially larger datasets (e.g., at the level of hundreds of thousands of images) are also needed to provide guidance on further scaling the current workflow (e.g., what's the minimum portion of dataset that needs to be inspected by the crowd to allow reliable detection of biases?).

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[2] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 313–322.

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

[4] Ferran Cabezas, Axel Carlier, Vincent Charvillat, Amaia Salvador, and Xavier Giro-i Nieto. 2015. Quality control in crowdsourced object segmentation. In *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 4243–4247.

[5] Lydia B Chilton, Greg Little, Darren Edge, Daniel S Weld, and James A Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1999–2008.

[6] Jinho D Choi, Joel Tetreault, and Amanda Stent. 2015. It depends: Dependency parser comparison using a web-based evaluation tool. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 387–396.

[7] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115.

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2007. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding* 106, 1 (2007), 59–70.

[9] Jochen Hemming and Thomas Rath. 2001. PA—Precision agriculture: Computer-vision-based weed identification under field conditions using controlled lighting. *Journal of agricultural engineering research* 78, 3 (2001), 233–243.

[10] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3819–3828.

[11] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. 2012. Undoing the damage of dataset bias. In *European Conference on Computer Vision*. Springer, 158–171.

[12] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 4017–4026.

[13] Genevieve Patterson and James Hays. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2751–2758.

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[16] Hao Su, Jia Deng, and Li Fei-Fei. 2012. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[17] Tian Tian, Ning Chen, and Jun Zhu. 2017. Learning attributes from the crowdsourced relative labels. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[18] Antonio Torralba, Alexei A Efros, et al. 2011. Unbiased look at dataset bias.. In *CVPR*, Vol. 1. Citeseer, 7.

[19] Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2017. FairTest: Discovering unwarranted associations in data-driven applications. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 401–416.

[20] Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 319–326.

[21] Michael J Wilber, Iljung S Kwak, and Serge J Belongie. 2014. Cost-effective hits for relative similarity comparisons. In *Second AAAI conference on human computation and crowdsourcing*.

[22] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2019. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. *arXiv preprint arXiv:1912.07726* (2019).

[23] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.