Anyone Can Skateboard: Tracing Sources of Gender Biases in Classifiers

Nicole Meister Dora Zhao

Abstract

There is a growing body of work across machine learning examining the propagation of social biases in both datasets and trained models. One commonly studied bias is gender bias. While gender bias is easier to quantify in tabular data where gender identity is well-defined, the task becomes more difficult with unstructured data, such as images and videos, as researchers often have to rely on proxies to determine gender. Furthermore, the labels taken as "ground-truth" in these experiments can be sources of biases themselves. To understand the role of contextual objects and label disagreement in gender bias, we conduct four key analyses. First, we identify the top objects contributing to gender prediction by training a logistic regression classifier on ground-truth one-hot encodings of objects in an image. Second, we visualize the model's attention in gender classification and quantitatively measure where the model focuses. Next, we quantify the objects' impact on model performance by training a model with objects obscured. Our goal is to understand how models perform when different gender cues are removed. Finally, we analyze whether methods for aggregating ground-truth labels are systematically erasing the perspectives of different annotators. Our code can be found at https://github.com/dorazhao99/cos-529/.

1. Introduction

Gender bias, stemming from both under-representation of certain genders and gender contexts (e.g., woman riding skateboard), exists in all types of data. This contributes to machine learning models that not only perpetuate but also amplify these biases [38, 39, 42]. Machine learning models often take advantage of correlations between co-occurring objects [13]; however, in doing so, they risk inadvertently encoding social biases found in datasets [28, 31, 36, 41]. Without properly quantifying and actively reducing reliance on such correlations, deploying models can magnify harmful stereotypes.

In fact, these gender biases have already led to downstream harms in deployed applications. For instance, Schwemmer *et al.* [33] empirically test the Google image tagging system and discover that images of female politicians were labeled differently than their male counterparts. Similarly, inquiries into image search engines found that these systems return stereotypical results that mirror harmful gender roles [19, 25, 27]. As machine learning models are becoming more readily deployed, it is of increasing importance that researchers identify where biases may arise and take steps to mitigate or eliminate these biases.

In our work, we focus on trying to quantify how different aspects of training datasets, ranging from the ground-truth labels to the image features, can contribute to gender biases. To ground our work, we define gender bias using the framework of representational harms [2, 3]. Specifically, we consider two aspects of representational harms: stereotyping, (e.g., different object distributions in the image content [37], use of spurious correlations during inference [17, 39, 42]) and disparate performance based on protected attributes [5, 41].

This line of work is particularly challenging for the following reasons.

Ill-defined tokens. In other domains that use tabular or textual data, researchers are able to not only easily identify but also quantify how much each token (e.g., a feature in tabular data, a word in a sentence) contributes to gender bias [4]. In computer vision, there are no well-defined tokens, making it harder to characterize bias and pinpoint its sources.

Annotator disagreement. In addition to poorly-defined tokens, fairness researchers fail to consider label uncertainty, which occurs when annotators disagree on the "ground-truth" label [14,29,30]. This is in part due to dataset creators only releasing aggregated labels and dataset users placing blind trust in ground-truth labels. Since gender annotations are inferred labels for social constructs and thus highly influenced by annotators' perceptions, label uncertainty in gender classification is a critical issue.

Our contributions. This work tackles these challenges in studying gender bias by exploring two main ideas. First, to address the challenge of ill-defined tokens, we study how contextual objects influence gender classification by quantifying how much these objects contribute to gender classification and gender bias. Second, we examine how common

data collection practices, such as aggregating annotator-level labels using majority voting [41], can introduce biases into the ground-truth annotations.

Our key contributions include:

- **Top objects contributing to gender prediction.** For increased understanding of an object's role in gender classification, we train a logistic regression classifier on an image's ground-truth one-hot encodings of objects. The classifier's weights correspond directly to the object's role in classification. From this interpretable model, we identify the top objects contributing to gender prediction sorted by importance.
- **Visualize model attention in gender classification.** To qualitatively understand what models focus on when predicting gender, we visualize saliency maps of the gender classification model. We conduct a quantitative analysis by calculating the average IOU of the person segmentation mask and activated region in the saliency map.
- **Mask out different objects to understand their role in gender classification.** As an extension of our analysis on saliency maps, we also train models using images with different objects obscured. This allows us to experimentally determine the object's role in gender classification.
- **Examine biases arising from ground-truth labels** We analyze the inter-rater agreement between the singular gender label and the multiple annotator-level labels for each instance. In particular, we examine the agreement both at the individual and demographic group level.

2. Related Work

Dataset biases. The presence of bias in datasets has been well-studied. These biases include imbalanced demographic representation [5, 40, 41], stereotypical portrayals [6, 33, 36], and even the presence of harmful or offensive content [30]. However, given the large-scale nature of current image datasets, it is often difficult to uncover these biases. Recent work [37] has proposed tools to make identifying visual biases in image datasets more tractable.

We aim to build upon these existing tools and offer more in-depth analyses as to where biases may arise. To do so, we incorporate both more interpretable results via saliency maps and analyses of annotator disagreement. While tools to better understand data distributions within computer vision datasets exist, there remains a large gap in understanding how objects in an image (e.g., flowers in female images) contribute to gender prediction. Further, to measure the success in identifying which image features contribute to gender bias, quantifying the change in gender accuracy if an identified object is masked out is needed.

Understanding model biases. There is also interest in understanding where biases are arising in models. One proposed method is to experimentally manipulate features in an image to isolate sources of bias. This has been done using simple image processing techniques [24], finding counterfactual examples manually [35], or using generative adversarial networks (GANs) to synthetically manipulate attributes [1, 11]. Another approach is to use interpretability methods, such as attention heatmaps, to understand where the model is focusing when making a prediction. In a similar vein to Hendricks *et al.*'s [17] work on bias mitigation in image captioning, we are interested in visualizing what objects models are focusing on when classifying gender.

Annotator biases. One underexplored source of bias is in the ground-truth labels themselves. Annotator disagreement tends to arise for two reasons: the first is due to natural, random variation and the second is systematic differences in perceptions, often influenced by the demographics of the annotators themselves [26]. At the moment, most datasets use the majority over annotator labels to be the ground-truth. However, recent work [9, 14, 29] suggests that collapsing the disagreement may be removing important signals. Given that our gender classifier relies on crowdsourced labels, we explore how annotator disagreement may also be a source of bias.

3. Method

Dataset. We focus our analysis on the Common Objects in Context (COCO) dataset [21]. COCO is widely used across different computer vision tasks [12, 18, 23]; this makes it particularly important to investigate dataset biases that may be present [8]. Further, COCO has been the testbed for previous work on bias mitigation [39, 42].

Specifically, we use the newly published [41] crowdsourced perceived gender expression annotations on COCO 2014's validation set. In total, the annotations encompass 15,762 images and 28,315 `person` instances. When necessary, we split the dataset into training, validation, and test sets consisting of 6,468, 2,156, and 2,156 images respectively. As identified by Zhao et al. [41], the distribution is skewed with males comprising 68.9% of our train set and 70.5% of the test set.

Ethical considerations. It should be noted that while training a gender classifier is critical for understanding gender bias and mitigating downstream effects, gender classifiers are fundamentally imperfect. Often times, gender is reduced to a simplistic binary that can be harmful to individuals from the trans and / or non-binary community who may not fit into these narrow categories [15, 20]. While it

is important to understand where gender biases may arise in automated systems, we do not condone the use of automated gender recognition in practice. The purpose of a gender classifier in this project is for the study of gender bias propagation and not to be used in practice.

Model. Our gender classification model uses a ResNet-50 [16] pre-trained on ImageNet [10] as its backbone. The model is optimized with stochastic gradient descent (SGD) and a batch size of 200. The ResNet-50 has a final fully connected layer mapping the 2048 size hidden layer to a single output value corresponding to the gender classification. We arbitrarily assign "male" as equal to 0 and "female" to 1. After optimizing the hyperparameters using grid search (learning rate: $0.1, 0.001, 0.0001, 0.00001$; weight decay: $0, 0.01, 0.001, 0.0001$), the optimal parameters for the model are a learning rate of 0.001. SGD is used with a momentum of 0.9 and zero weight decay. The input images are resized to 224 on the smaller dimension, randomly cropped to 224×224 and randomly flipped horizontally during training. The model achieves an AUC of 90.87.

4. Gender Biases from Contextual Objects

4.1. Investigating Spurious Correlations

Our first course of action is to examine the distribution of objects in images labelled as "female" versus "male." This allows us to see if there are any overt interpretable spurious correlations the model exploits. Then, we train a logistic regression classifier on ground-truth one-hot encodings of objects to identify the most influential (top 10) objects contributing to male or female gender prediction sorted by importance.

Understanding the distribution of contextual objects. We start by analyzing the difference in object occurrences with each gender. To do so, we count the number of object occurrences using the COCO object annotations over all images that have been classified as either "male" or "female." We then visualize the top 10 largest percent differences in object occurrences between "male" and "female" images.

We find that the top 10 percent differences range from 2 to 3% difference between the same object in female images versus the object in male images (Fig. 1). While these object distributions may not seem incredibly large, in future sections we explore how gender classifiers exploit these small differences and how these spurious correlations contribute to differences in gender classification.

Top objects contributing to gender prediction. For increased understanding of the role of objects in gender classification, we train a model without deep features (i.e., no CNN). More specifically, we train a logistic regression classifier on ground truth one-hot encodings of objects in an image. This classifier's weights will correspond directly to the object's role in classification. Thus, we can identify the most relevant objects that contribute to gender prediction sorted by importance. Since the model predicts "male" (0) or "female" (1), the *lowest* 10 classifier weights correspond to the top 10 objects most useful to classify "male" which include `baseball glove`, `skateboard`, and `tie`. The *highest* 10 classifier weights correspond to the top 10 objects most useful to classify "female" include `teddy bear`, `handbag`, and `vase` (see Fig. 2).

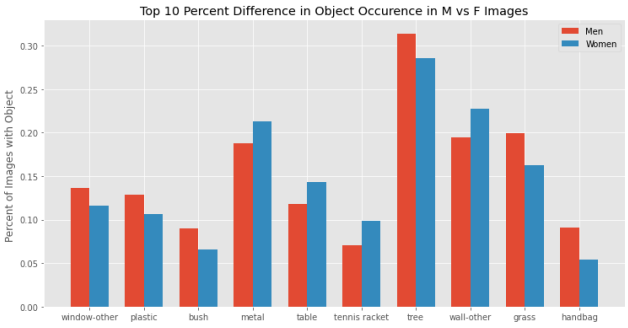

Figure 1. **Distribution of Objects in Male/Female Images.** This analysis aims to understand the distribution of object instances for female vs male images. This bar chart displays the top 10 largest percent differences in object occurrences between images classified as female vs male.

4.2. Visualizing Model Attention

We now visualize saliency maps of the gender classification model. Specifically, we use Class Activation Maps (CAMs) [34], which are saliency maps shown to expose the implicit attention of neural networks, highlighting "the most informative image regions relevant to the predicted class."

Using the CAMs, we conduct quantitative analysis to understand how often the model looks at the person segmentation. Then, to understand *why* a model's attention would not be directed towards the person when classifying their gender, we qualitatively examine the resulting saliency maps.

Evaluating Saliency Maps (CAMs) in Gender Classification. First, we quantify whether models are focusing on people when predicting gender. To do so, we calculate the overlap (IOU) between the activated regions of the CAMs (i.e., where the model is looking to identify gender) and the person segmentation mask. We define the activated region as the 'red' regions of the CAM and create a mask for that region (see Fig. 3 center). To mask out the activated regions of the CAM, red regions of the heatmaps are defined by the range of pixel values in the HSV colorspace with a lower bound of $lb = (85, 90, 100)$ and upper bound $ub = (130, 255, 255)$. From this analysis, we find that the average IOU between the activated region of the CAM and the person segmentation mask is 0.156 ± 0.1. We visualize

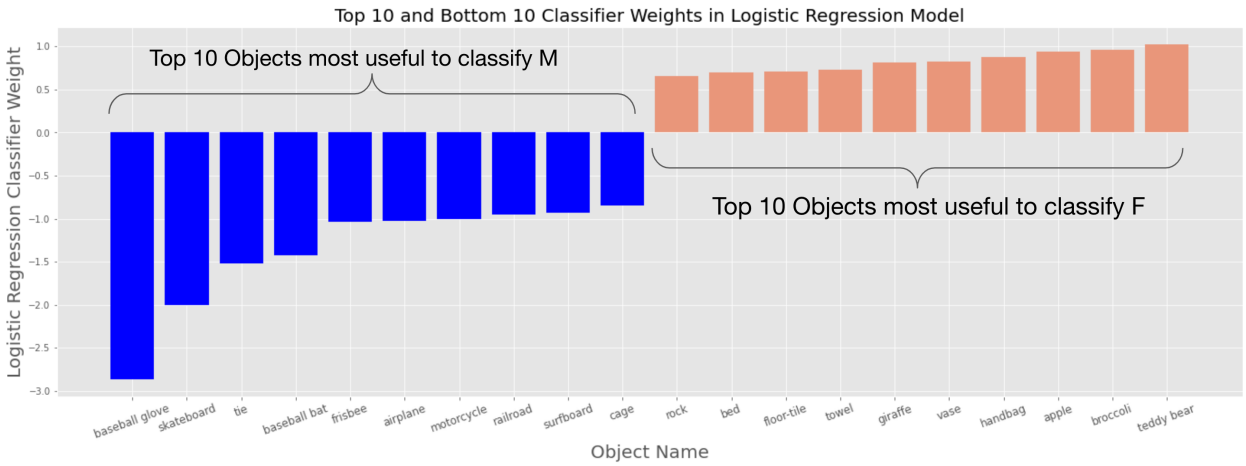

Figure 2. **Top 10 objects with strongest cuest for prediction male and female.** Top ten and bottom ten classifier weights in a logistics regression classifier trained on ground-truth one-hot encodings of objects in training images to predict "male" (M, 0) or "female" (F, 1). The objects with the strongest cues for "male" and "female" are `baseball glove` and `teddy bear`, respectively.

Figure 3. **IOU between Activated CAM and Person Segmentation Mask** To conduct a quantitative analysis of the model's attention (left), the IOU between the activated region of the CAM (center) and person segmentation mask (right) is calculated. This particular image has IOU of 0.156.

	Female	Male
All	0.208 ± 0.15	0.134 ± 0.13
No Top 10	0.208 ± 0.15	0.142 ± 0.14
Top 10	0.207 ± 0.16	0.123 ± 0.13

Table 1. **IOU CAM analysis.** We report the average and standard deviation of the overlap (IOU) between the activated regions of the CAMs and the person segmentation mask. "No Top 10" refers to images that *do not* contain the top 10 objects identified in Sec. 4.1 while "Top 10" refers to images that *do* contain the top 10 objects. The average IOU for all images is 0.156 ± 0.143.

an image with an IOU of 0.156 in Fig. 3 to illustrate that this average overlap is quite small.

We also conduct t-tests between the IOU for female and male images. We find there is a statistically significant difference ($p < 0.05$) between the average IOU for female images, 0.208, and average IOU for male images, 0.134 (Tab. 1). While this may be due to the fact that there are more male images (1620) than female images (536) in the annotated COCO validation set, it may also suggest that gender bias is more prevalent for images with males. Additionally, when comparing the IOU between male images containing the top 10 objects most useful to classify "male", as identified in Sec. 4.1, versus without the top 10 objects, there is a statistically significant difference in the IOU when the contextual objects are in the image (Fig. 5) compared to when the objects are not (Fig. 4). This suggests that when the top 10 objects appear in the image, the model's attention is less focused on the person. This indicates the model is exploiting spurious correlations to predict gender and thus is a source of gender bias. The difference is not statistically significant for female images.

Qualitative analysis of CAMs. Next, we seek to understand *why* models are not focusing on people when predicting gender. In Fig. 4 and 5, we visualize CAMs for predicting the class "female" and "male." More specifically, in Fig. 4 we visualize CAMs from images that *do not* contain the top 10 objects most useful for classifying gender that we determined in Sec. 4.1. For these images, we observe that the activated regions of CAMs focus on either the face, body, or body and face and we provide visual examples of this pattern. We also visualize CAMs from images that *do* contain the top 10 objects most useful for classifying gender (Fig. 5). From the activated regions in these CAMs, we display examples of when the model's attention is focused on other contextual objects to classify gender. For example, the model focus on objects, such as `horse`, `oven`, `teddy bear`, and `bed`, to classify "female" and objects such as `skateboard` and `motorcycle` to classify "male".

While CAMs provide valuable insight in identifying spe-

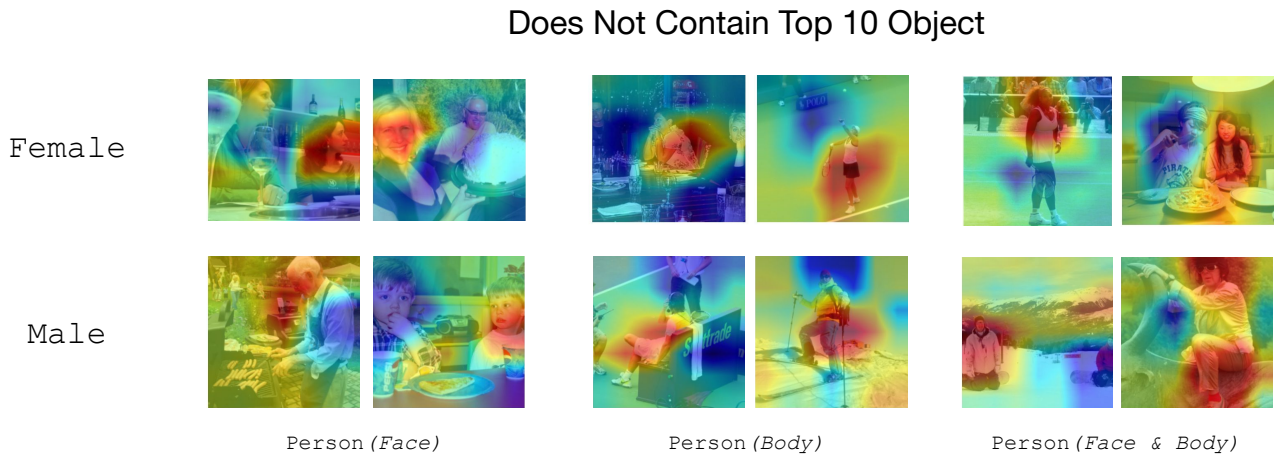

Figure 4. **CAMs of images without the Top 10 Objects useful to classify male or female.** We visualize images containing females and males and observe instances when the model's attention is focused on the face (left), body (center), and face & body (right).

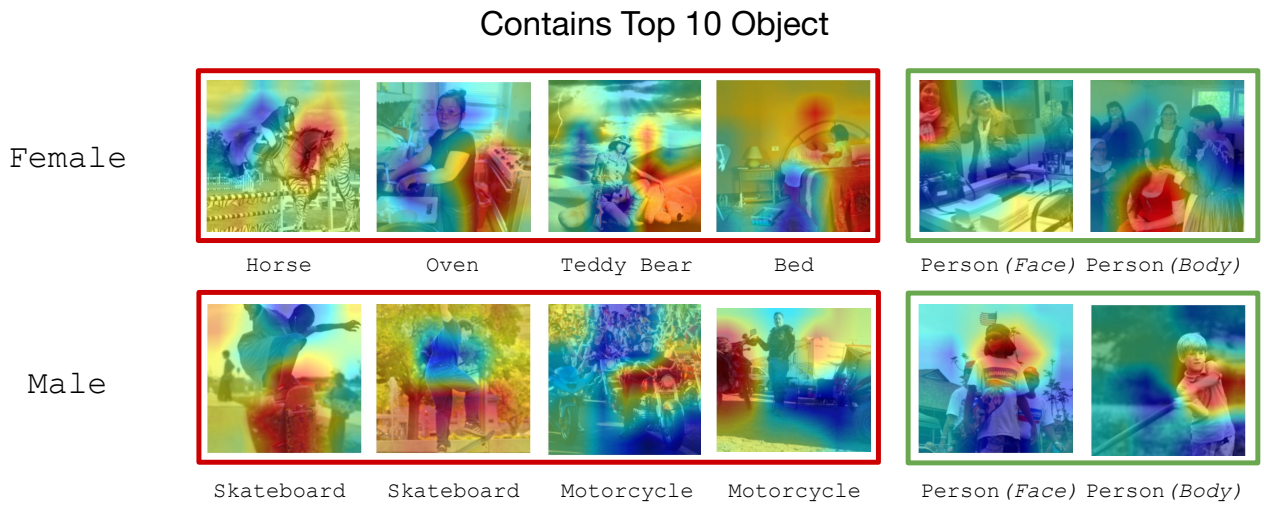

Figure 5. **CAMs of images with the Top 10 Objects useful to classify male or female.** We visualize images containing females and males. The red box contains images for which the model's attention is focused on contextual objects and not the person. The model's attention and reliance on spurious correlations to infer gender is a potential source for bias. In the green box, we highlight instances where the model's attention is focused on the person.

cific images where the model is not looking at the person to identify gender (Fig. 4), CAMs without overlapping activated regions and segmentation masks may be attributable to other variations in the image (i.e. scene, size of person, etc.), not just gender bias. This is evidenced by the large standard deviation in the IOU of CAMs. Therefore, we conclude that this analysis needs to be paired with the contextual masking experiments that we conduct and describe in the following section.

4.3. Masking Contextual Objects

To understand an object's role in gender prediction, we can remove (mask out) activated objects (i.e., objects identified by the classifier weights and objects activated by the CAMs) to observe the impact on gender accuracy. For masking out contextual objects, we use the object segmentations provided from the COCO dataset and set the pixels within the segmentation to the average pixel color in the image.

Experimental set-up. In addition to the gender classifier trained on unaltered images (BASELINE), we include four

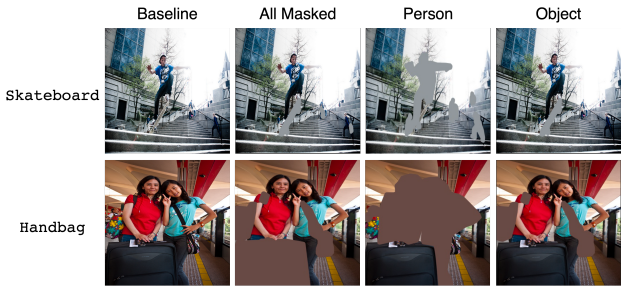

Figure 6. **Visualization of training data examples.** We provide examples of training data with no masked out contextual objects (Baseline), all objects except for person (All Masked), person objects masked, and gendered objects masked. The top row has the gendered object `skateboard` masked out, and the bottom row has `handbag`.

	Overall AUC	Female Acc.	Male Acc.
BASELINE	90.87	75.16	89.61
ALL MASKED	61.85	41.98	74.41
PERSON	65.70	44.50	75.53
SKATEBOARD	82.09	65.57	83.49
HANDBAG	82.06	64.31	83.29

Table 2. **Performance of models trained with varying number of contextual objects masked out.** The number in the first column is AUC $\times$ 100 for classification ability. The last two columns are accuracy scores (%) disaggregated by ground-truth gender labels. We calibrate our threshold with predictions on the validation set.

additional models. First, we mask out all people in the image to see how well a classifier can predict gender using *only* contextual objects. Next, for illustrative purposes, we choose two objects that previous analyses identified as being highly predictive of gender. Specifically, for SKATEBOARD, we mask out `skateboard`– one of the top objects that are useful in classifying "male". Similarly, we mask out `handbag`, which is a strong cue for "female" (HANDBAG). Finally, as a control, we study the impact of contextual objects as a whole on gender classification by masking out all objects, except person (ALL MASKED). We provide illustrative examples of the data used to train the respective models in Fig. 6.

Here, we use area under the ROC curve (AUC) to evaluate the overall performance of our gender classifier. We report AUC $\times$ 100. In addition, we conduct a disaggregated evaluation of the accuracy scores based on the ground-truth gender label. To calculate accuracy, we calibrate our classification threshold based on our validation set. Following prior work [38], we choose this threshold such that the proportion of predicted positive labels matches that of the ground-truth. The results are in Tab. 2.

Results. We start by making two general observation. First, we find that BASELINE performs the best across all of the models with an AUC of 90.87. Second, the disaggregated evaluation reveals that our classifier performs considerably better on images of males compared to females. For ALL MASKED, the difference in accuracy between the two groups is 32.43%.

We now consider the model performance when we mask out contextual objects. For SKATEBOARD and HANDBAG, we see similar drops in performance with AUCs of 82.09 and 82.06 respectively. The decrease is largely attributable to the drop in accuracy when classifying images of females. In fact, for both models, the disaggregated accuracy score for females dropped by 9.59% for SKATEBOARD and 10.85% for HANDBAG. Given that `handbag` objects are strong cues for "female," we expect to see a relatively large decrease in female accuracy scores for HANDBAG.

Finally, we look at the performance of ALL MASKED and PERSON. While both perform worse than BASELINE, we find that masking out all contextual objects (excluding people) has a greater negative impact on AUC (61.85) compared to masking out only people (65.70). Our findings indicate that the models are relying heavily on contextual objects to make gender predictions, more so than even relying on the person's appearance. This is in line with our findings from Sec. 4.2 that the average IOU between person segmentations and activated CAMs is quite small. Further, this suggests that models are likely to exploit spurious correlations when predicting gender and are thus also likely to reproduce existing social biases found in the data.

5. Gender Biases from Aggregated Labels

We now discuss the findings from our experiments on examining what kinds of biases are present in the crowdsourced demographic annotations for COCO. In part, these biases arise because dataset creators take the aggregate over the provided annotations, which may systematically erase the perspectives of certain individuals or particular groups. For example, in the COCO demographic annotations [41], the instance-level label is the majority over three or more annotations from different crowdsource workers.

To empirically evaluate what biases label aggregation introduce, we follow the methodology presented by Prabhakaran *et al* [29]. First, we consider how representational biases may be introduced on the individual annotator level. We then look at the same phenomenon but across different demographic groups.

5.1. Individual Level Biases

We start by examining whether the ground-truth COCO demographic labels exclude any individual annotator's perspective. For each worker, we calculate the Cohen's Kappa agreement score between their labels with the respective

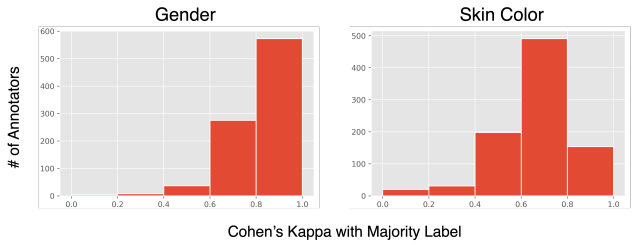

Figure 7. **Individual annotator agreement distribution.** Histograms representing the frequency distribution of individual annotator agreement with the majority label for the COCO demographic annotations dataset. Higher Cohen's Kappa scores indicate that the majority label is able to accurately capture the perspectives of individual annotators. The scores for perceived gender expression are on the left and skin color on the right.

majority labels for those instances. Typically used to measure inter-rater reliability, here Cohen's Kappa measures how well the majority labels align with the perspectives of individual annotators.

In Fig. 7, we plot the Cohen's Kappa scores for perceived gender expression annotations. On average, the agreement for perceived gender expression is 0.827 ± 0.14. This indicates that the majority labels are sufficiently representative of the perspectives for most individual annotators. We also consider the agreement for the observed skin color of the individual– another demographic attribute captured in the COCO annotations. Compared to perceived gender expression, there is more disagreement ($\kappa = 0.667 \pm 0.18$). The difference in agreement may be because perceived skin color is annotated on a more granular scale (i.e., rated 1 through 6 using the Fitzpatrick Skin Type scale) compared to perceived gender expression, which is a binary task. Further, this may indicate that there are strong societal stereotypes about what physical attributes indicate different genders.

5.2. Demographic Level Biases

Next, we turn to analyzing whether the majority labels equally capture the perspectives of different demographic groups. While it is problematic if majority labels do not capture individual annotators, it is particularly concerning if the majority labels are systematically excluding certain demographic groups [29].

The COCO demographic annotations also provide racial and gender identities for each of the annotators. Using these annotations, we calculate the Cohen's Kappa scores for each racial (White, Black, Asian, and Latine) and gender (Male, Female) group (Fig. 8). We also conduct an intersectional analysis, looking at each combined race-gender group (Fig. 7). In addition to calculating the Cohen's Kappa score with respect to the majority label, we conduct three one-way ANOVA tests to ascertain whether the differences are statistically significant.

From these analyses, we make three key observations. First, we find more disagreement with the majority level when disaggregating by racial groups compared to gender identity groups. In fact, male and female annotators have similar Cohen's Kappa scores of 0.815 ± 0.08 and 0.816 ± 0.08 respectively. For racial groups, Black annotators have lower Cohen's Kappa scores ($\kappa = 0.807 \pm 0.11$). In Fig. 8, we observe that the distribution of Kappa scores is skewed left for Black annotators, indicating there is a subset of annotators' perspectives that are not represented in the aggregated labels. Second, by looking at each race-gender group, we further disentangle which annotator perspectives are not represented. We find that Black males have the least agreement with the majority ($\kappa = 0.773 \pm 0.13$). In contrast, Black female annotator perspectives are most represented in the majority labels as they have the highest mean score of 0.833 ± 0.09. Finally, from the ANOVA tests, we find that these differences are not statistically significant at the $p = 0.05$ level. It is important to note that, even though the results are not statistically significant, it does not guarantee that the ground-truth labels are not a source of bias.

Overall, compared to Prabhakarn *et al.*'s [29] analysis of hate-speech, sentiment, and emotion annotations, we find there is more agreement between the aggregated majority and annotator-level for demographic labels. We suspect this discrepancy arises because the COCO demographic annotation task asked for judgement on physically observable traits (i.e., perceived gender expression and skin color). In contrast, the NLP tasks studied before were asking annotators for more subjective judgements, which are more likely to differ on the individual and the group level. Were we to analyze annotator-level labels for more subjective computer vision tasks (e.g., labeling the `Attractive` attribute in the CelebA dataset [22]), the ground-truth labels would likely be a larger source of bias.

6. Limitations

We report two limitation with our analysis. First, while human annotations for gender expression have been shown to be better than automatically derived annotations [41], we are still relying on proxies for the actual attribute of interest: gender identity. In addition, when using human annotations, we also inherit any of the biases that annotators may have [7, 26]. Even though our analysis indicates that there is considerable agreement amongst the annotators who provided the demographic labels, this does not guarantee that the sample is representative or that we are not excluding certain groups altogether. In particular, for the task of perceived gender, we note a lack of representation of transgender and / or non-binary individuals in the annotator population. This limitation highlights the need for collect-

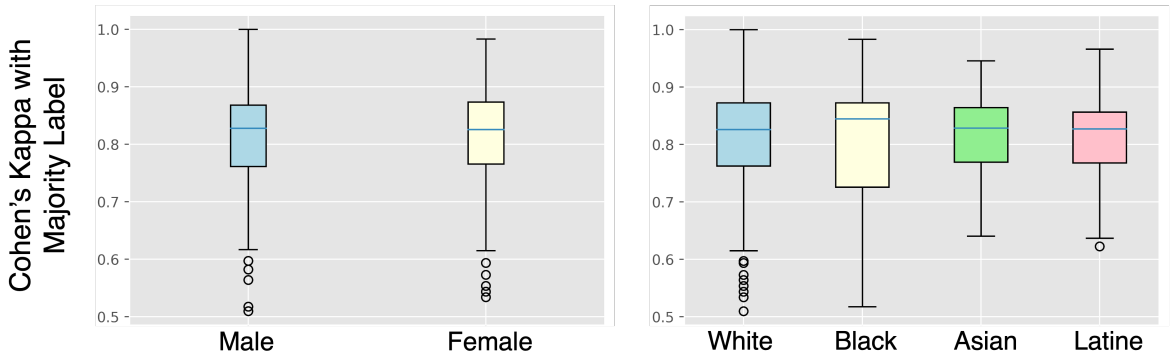

Figure 8. **Distribution of annotator agreement with the majority labels by annotator gender and race.** The distributions are disaggregated by the demographic groups– gender (left) and racial (right) – of the annotators.

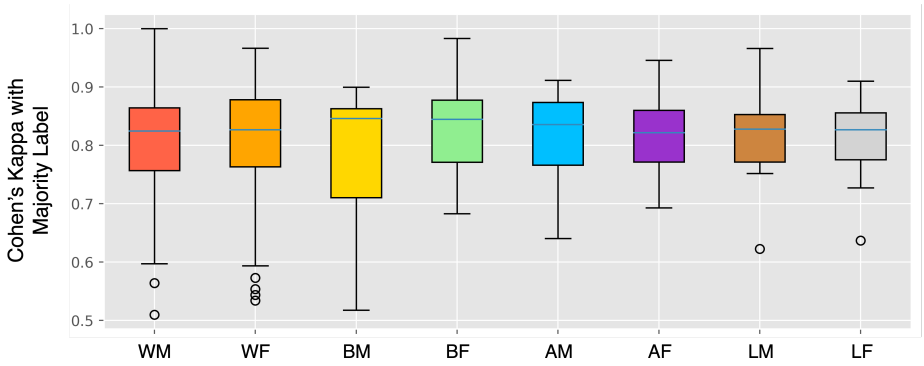

Figure 9. **Annotator agreement disaggregated by intersectional demographic groups.** Distribution of annotator agreement with the majority labels, disaggregated by the intersectional race (W, B, A, L) and gender (M, F) identities of the annotators.

ing self-reported demographics of the pictured individuals rather than relying on third-party judgements.

Second, our analysis focuses on individuals who appear strongly feminine or masculine. The COCO demographic annotations exclude `person` instances that were labeled as both "male" and "female." Also, annotators were also given the option to label gender expression as "unsure." As a result, we only have access to a filtered subset of all images. Nonetheless, we note that real-world gender classifiers are likely to be trained on binary attributes that have stronger physical cues for gender [32]. Thus, our conclusions are still relevant for real-world use cases.

7. Conclusion and Future Work

These results provide clarity on how contextual objects and label uncertainty contribute to gender bias. First, we identified the top 10 objects contributing to male and female gender prediction and analyzed their impact on gender classification which can be used to appropriately balanced and representative datasets. Subsequently, from our quantitative CAM analysis, we find that models look at contextual objects less for females, suggesting that gender bias may be more prevalent for images with males. These findings are experimentally verified through our analysis using masked objects. Finally, our analysis of the ground-truth labels suggest that the demographic labels we are using may be excluding some perspectives, namely Black male annotators.

Looking forward, while this project is scoped to gender bias, it can be applied to other forms of bias including ethnicity and race as Zhao *et al.* [41] also provide skin-color annotations on COCO. Furthermore, more extensive CAM analysis can be conducted to better understand where the model is looking when it looks at the person (i.e., identifying when the model look at facial features versus hair versus clothing to identify gender).

Studying sources of social biases in computer vision systems is critical. As these systems become increasingly ubiquitous, it is crucial to ensure that they do not not propagate and encode existing biases. Our work elucidates which contextual objects are utilized in gender classification by conducting quantitative and qualitative analysis, and provides clarity on the impact of annotator disagreement.

References

[1] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards causal benchmarking of bias in face analysis algorithms. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[2] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. In *Annual Conference of the Special Interest Group for Computing, Information and Society (SIGCIS)*, 2017. 1

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 1

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Neural Information Processing Systems (NeurIPS)*, 2016. 1

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency (FAccT)*, 2018. 1, 2

[6] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 2

[7] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021. 7

[8] Kathleen Creel and Deborah Hellman. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision making systems. In *Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021. 2

[9] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *arXiv preprint arXiv:2110.05719*, 2021. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 3

[11] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv e-prints*, pages arXiv–1906, 2019. 2

[12] Francis Ferraro, Nasrin Mostafazadeh, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell, et al. A survey of current datasets for vision and language research. *arXiv preprint arXiv:1506.06833*, 2015. 2

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[14] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Conference on Human Factors in Computing Systems (CHI)*, 2021. 1, 2

[15] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018. 2

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[17] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[18] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6), 2019. 2

[19] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *ACM Conference on Human Factors in Computing Systems (CHI)*, 2015. 1

[20] Os Keyes. The misgendering machines: Trans/HCI implications of automatic gender recognition. *ACM Conference on Computer Supported Cooperative Work (CSCW)*, 2018. 2

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014. 2

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015. 7

[23] Shervin Minaee, Yuri Y Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[24] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney. Understanding unequal gender classification accuracy from face images. *arXiv preprint arXiv:1812.00099*, 2018. 2

[25] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. 1

[26] Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. How do we talk about other people? group (un) fairness in natural language image descriptions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 106–114, 2019. 2, 7

[27] Jahna Otterbacher, Jo Bates, and Paul Clough. Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 chi conference on human factors in computing systems*, pages 6620–6631, 2017. 1

[28] Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis)

contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. 1

[29] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Díaz. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*, 2021. 1, 2, 6, 7

[30] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020. 1, 2

[31] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, 2017. 1

[32] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019. 8

[33] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6:2378023120967171, 2020. 1, 2

[34] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 3

[35] Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018. 2

[36] Emiel Van Miltenburg. Stereotyping and bias in the flickr30k dataset. *arXiv preprint arXiv:1605.06083*, 2016. 1, 2

[37] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. REVISE: A tool for measuring and mitigating bias in visual datasets. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[38] Angelina Wang and Olga Russakovsky. Directional bias amplification. In *International Conference on Machine Learning (ICML)*, 2021. 1, 6

[39] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[40] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In *ACM Conference on Fairness, Acountability and Transparency (FAT*)*, 2020. 2

[41] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6, 7, 8

[42] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 1, 2