ICON^2: Reliably Benchmarking Predictive Inequity in Object Detection by Identifying and Controlling for Confounders

Sruthi Sudhakar[*]
Columbia University
sruthi@cs.columbia.edu

Viraj Prabhu
Georgia Tech
virajp@gatech.edu

Olga Russakovsky
Princeton University
olgarus@cs.princeton.edu

Judy Hoffman
Georgia Tech
judy@gatech.edu

Abstract

As computer vision systems are being increasingly deployed at scale in high-stakes applications like autonomous driving, concerns about social bias in these systems are rising. Analysis of fairness in real-world vision systems, such as object detection in driving scenes, has been limited to observing predictive inequity across attributes such as pedestrian skin tone [41], and lacks a consistent methodology to disentangle the role of confounding variables e.g. *does my model perform worse for a certain skin tone, or are such scenes in my dataset more challenging due to occlusion and crowds? In this work, we introduce ICON^2, a framework for robustly answering this question. ICON^2 leverages prior knowledge on the deficiencies of object detection systems to* identify *performance discrepancies across sub-populations, compute* correlations *between these potential confounders and a given sensitive attribute, and* control *for the most likely confounders to obtain a more reliable estimate of model bias. Using our approach, we conduct an in-depth study on the performance of object detection with respect to income from the BDD100K driving dataset, revealing useful insights.*

1. Introduction

Computer vision models today are being deployed in high-stakes applications such as self-driving cars and job hiring. These vision models rely on datasets for learning which often contain undesirable biases that are perpetuated and in some cases amplified by these models, which can result in inequitable performance for certain sub-populations [4, 10, 49]. Within computer vision, fairness assessment has been limited to the relatively simple task of image classification [9, 34, 40], with limited work in evaluating fairness in more complex tasks such as object deteciton [41] In this work, we introduce a methodology for inspecting biases in *object detection models* trained on driving datasets.

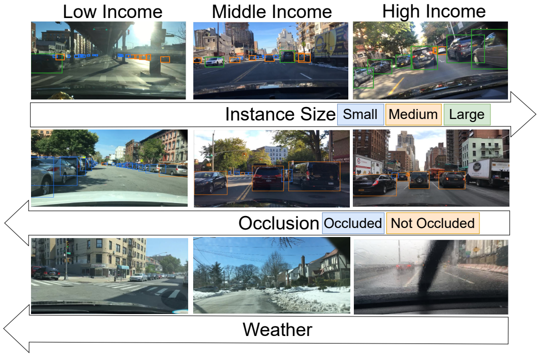

Figure 1. **Overview.** We seek to reliably benchmark performance discrepancies in object detection systems used for autonomous driving across driving scenes from varying income levels. We propose ICON^2, an automated framework that only uses test data and no retraining to Identify and Control for Confounders (such as instance size, occlusion, or weather) while evaluating predictive inequity, yielding more reliable insights.

Identifying biases in detection systems is challenging due to complex performance metrics in combination with contextual [32] and co-occurrence biases [49] across visually diverse driving scenarios. Prior work [41] has primarily relied on directly computing performance discrepancies across sub-populations to evaluate fairness across a sensitive attribute (in their case, skin tone). In practice however, this does not provide *reliable* insights to inform mitigation solutions. For example, simply knowing that average precision of a detector is 2 points higher in high-income neighborhoods than in low-income ones may not tell us the complete story, without controlling for possible confounders: *e.g.* were a higher percentage of images from low-income areas collected at nighttime, making detection harder? Similarly, were images form high-income areas less crowded or occluded, which made detection easier? Not controlling for these may lead to flawed conclusions that lead to misguided interventions (such as collecting more data from low-income regions, rather than addressing the underlying data collection protocol discrepancies). While other high-stakes applications such as medicine have long-emphasized the importance of considering confounders when assessing bias [33], autonomous driving research is yet to adopt this

[*] Corresponding author. Work done at Georgia Tech.

protocol. In this work, we provide a principled framework to identify and control for such confounding factors when measuring predictive inequity across a sensitive attribute.

Concretely, we leverage prior knowledge on known deficiencies of object detection systems, such as small objects or nighttime scenes [5, 13, 26, 47], and seek to *quantify* the extent to which such attributes impact the observed system biases. Our algorithm, that we call Identify and Control for Confounders (ICON^2), first *ranks* a set of such attributes according to their potential impact on performance, and then *validates* this ranking by measuring the performance disparity while *controlling* for the effects of top-ranked attributes.

Using our approach, we conduct an in-depth study on the detection performance disparity across income levels on the BDD100K [46] dataset for autonomous driving. We consider innate object detection deficiencies as potential confounders, including object size, time of day, occlusion, and weather. Our method first identifies the most likely confounders, *e.g.* finding that a larger proportion of small cars in images from lower-income regions likely contributes to worse performance. We then control for the likely confounders to see if it explains away the observed bias. Our analysis illuminates the complexity assessing fairness in object detection, and takes a step towards robust benchmarking that may guide more effective interventions.

2. Related Work

Fairness Toolkits. Several analysis tools [3,8,14,17,18,37] exist to discover biases in datasets and models. For example, Google's Know Your Data [8] focuses on discovering correlations between attribute labels and metadata labels in TensorFlow image datasets. However, this work only focuses on ground truth label correlations, limiting the discovery of bias. Extending this idea, REVISE [37] focuses on examining the dataset to discover biases along three dimensions: (1) object-based, (2) person-based, and (3) geography-based. However unlike us these works do not study complex tasks like object detection, and do not evaluate the impact of such correlations on performance.

Fairness in Object Detection. While much prior work in vision has investigated bias in image classification systems [24, 32, 48], object detection models differ in model architecture and performance metrics, necessitating their own study. Prior work has investigated object detection datasets more generally [29, 30], however to our knowledge only one prior work has touched on fairness in object detection. Wilson et al. [41] studied the predictive inequities between light-skinned and dark-skinned pedestrians using two SOTA two-stage detection systems. They aim to identify the role of two possible reasons for such inequity: time of day and occlusion, but propose a retraining-based strategy to control for these. We take inspiration from this work and develop a general methodology to reliably measure bias by controlling for various factors without retraining.

Common Deficiencies in Object Detection. There have been several works aimed at understanding factors that hurt object detection performance [6,19,27,45,47]. We describe these factors in Sec 3.3 and leverage these works to perform our study of bias in object detection.

3. Approach

We now introduce our approach to reliably benchmark predictive inequity in object detection for self-driving.

3.1. Preliminaries

A key aspect of our approach is to consider different underlying factors that can be used to describe aspects of the data. We refer to these quantities as ***attributes***, which can be any additional categorical variables with which the data is annotated. Example attributes include: time of day, object size, geolocation, camera resolution, *etc*. In addition, we consider an attribute as ***sensitive*** if it corresponds to quantities that have legal or ethical implications. Example sensitive attributes include: income level, gender, race, *etc*. For clarity, we refer to attributes across which we are *not* interested in inspecting bias, but which may still act as potential confounders, as ***explanatory*** attributes.

Fairness. We define a model to be *unbiased* or *fair with respect* to any attribute if the performance of the system (measured by average precision in our case) is independent of the attribute. For an attribute, A, with three values, a, b, c, the goal would be to have equal performance:

$$AP_a = AP_b = AP_c \tag{1}$$

This definition is a generalization of the equality of accuracies fairness metric [36].

Computing AP per Attribute Value. For simplicity of notation, we present our approach considering a single class and note that it can be applied independently per-class. For a given category and sensitive attribute A, we compute AP for each value taken by the sensitive attribute AP_{a_i} by only considering images that have positive ground truth annotations for the class of interest **and** for the sensitive attribute value, a_i. This strategy has a long history in object detection challenge datasets such as PASCAL [11] and MSCOCO [23] that use this approach to compute AP across object sizes (for values "small", "medium", and "large").

However, AP is known to be highly sensitive to the number of positive samples in the evaluation set. Let N_i denote the number of positive instances corresponding to attribute value a_i. Let $R_i(c)$ denote recall (fraction of instances detected with a confidence of at least c) of the object detection model over this subset and $F_i(c)$ the corresponding number of false positives. Standard precision for group $P_i(c)$ is:

$$P_i(c) = \frac{R_i(c) \times N_i}{R_i(c) \times N_i + F_i(c)} \tag{2}$$

Algorithm 1 Explanatory Attribute Ranking Algorithm

```
Require: Explanatory Attribute Set $\mathcal{E}$
Require: Sensitive Attribute $(A)$
  VAR $= [\ ]$   ▷ variance of explanatory attribute ProxyAPs
  for $e_j \in \mathcal{E}$ do   ▷ explanatory attribute: e.g., size
      ▷ AP per explanatory attribute value
      Compute $AP_{e_j} \quad \forall e_j \in E$
      ▷ Empirical Distribution
      Compute $P(e_j|a_i) \quad \forall e_j \in E, \forall a_i \in A$
      ▷ ProxyAP per sensitive value
      $\text{ProxyAP}^E_{a_i} = \sum_{e_j \in E} P(e_j|a_i) \cdot AP_{e_j}$
      $\mu^E_A = \frac{1}{|A|} \sum_{a_i \in A} \text{ProxyAP}^E_{a_i}$
      $\sigma^2(\text{ProxyAP}^E_A) = \frac{1}{|A|} \sum_{a_i \in A} (\text{ProxyAP}^E_{a_i} - \mu^E_A))^2$
      VAR $= [$VAR$,\ \sigma^2(\text{ProxyAP}^E_A)]$
  end for
  return sorted(VAR) ▷ Rank explanatory attributes
```

To *meaningfully* compare AP values across sets with potentially different numbers of positive instances, we leverage the normalization strategy proposed by Hoeim *et al.* [16], which computes *normalized precision* $P_{N,i}(c)$, by replacing N_i with a constant N, where N is the mean of N_i across all values of a_i:

$$P_{N,i}(c) = \frac{R_i(c) \times N}{R_i(c) \times N + F_i(c)} \tag{3}$$

These normalized precision values are then interpolated and averaged across recall values to produce AP_{a_i}.

Variance in Attribute Performance. To quantify the impact of an attribute, A on performance, we consider the *variance* of performance across all attribute values, $a_i \in A$

$$\mu_A = \frac{1}{|A|} \sum_{a_i \in A} AP_{a_i} \tag{4}$$

$$\sigma^2(AP_A) = \frac{1}{|A|} \sum_{a_i \in A} (AP_{a_i} - \mu_A)^2 \tag{5}$$

Attributes with high performance variance, $\sigma^2(AP_A)$, have large differences in performance across different attribute values. A sensitive attribute with high performance variance suggests a possible model bias across that attribute.

3.2. Identifying and Controlling for Confounders

We now introduce our ICON^2, our two-stage framework that first identifies and ranks potential confounders, and then controls for them to obtain more reliable estimates of predictive inequity.

3.2.1 Ranking Potential Confounders

Given a sensitive attribute, A with non-zero attribute variance $\sigma^2(AP_A)$ (see Eq. 5) we aim to provide a potential explanation. To do so, our approach first ranks a set of explanatory attributes by their potential impact on model performance across the sensitive attribute. To compute this ranking, we define a proxy performance metric. While the variance of an explanatory attribute $\sigma^2(AP_E)$, indicates the degree to which performance varies as a function of the explanatory attribute values, $e_i \in E$, this measure alone is insufficient to explain any observed variance for our sensitive attribute, A, as it doesn't account for the relationship between the two.

Instead, we consider the conditional distribution of an explanatory attribute given the sensitive attribute value, $P(E|a_i)$. Intuitively, if the explanatory attribute is independent of the sensitive attribute we would find that the conditional distribution does not differ based on the sensitive attribute value, so that $P(E|a_i) = P(E|a_j) \quad \forall a_i, a_j \in A$.

For an explanatory attribute to impact the performance variance of a sensitive attribute, we need both for the explanatory attribute to have high performance variance and for the sensitive attribute to be dependent on the explanatory attribute. To capture these two notions, we introduce a metric that we call the *Proxy AP*, which for a sensitive attribute value a_i and explanatory attribute E is given by:

$$\text{ProxyAP}^E_{a_i} = \sum_{e_j \in E} P(e_j|a_i) \cdot AP_{e_j} \tag{6}$$

ProxyAP captures performance for a sensitive group a_i as a function of the explanatory attribute performance weighted according to the distribution of the explanatory attribute under the sensitive group. Next, we compute the performance variance (Eq. 5) across proxyAP values, $\sigma^2(\text{ProxyAP}^E_A)$, in order to quantify the influence of an explanatory attribute E on model performance discrepancy across a sensitive attribute A.

Finally, we sort explanatory attributes in our set in decreasing order of $\sigma^2(\text{ProxyAP}^E_A)$. This provides us with a list of explanatory attributes to investigate ranked by importance, which we utilize in the next step of our method. Our ranking approach is summarized in Algorithm 1.

3.2.2 Controlling for Confounders

The explanatory attribute ranking algorithm helps to quickly arrive at directions of study for understanding and mitigating performance discrepancies. To obtain a reliable estimate of predictive inequity with respect to a sensitive attribute, we need to control for top-ranked explanatory attribute and see if they explanin away the high variance in sensitive attribute performance.

To control for the explanatory attribute, we consider instances that are just part of one explanatory attribute value, e_j, **and** part of a sensitive attribute value, a_i. We can compute performance, AP, on this set of instances (AP_{a_i,e_j}). Next, to determine whether controlling for this explanatory attribute reduces variance, we use Eq. 5 to compute variance across settings of the sensitive attribute, $\sigma^2(AP_{A,e_j})$, for a single explanatory attribute value. Finally, we take the mean of the variance across settings of the explanatory attribute, $\mu(\sigma^2(AP_{A,e_j})\ \forall e_j \in E)$, and compare this average to the original variance across the sensitive attribute before controlling, $\sigma^2(AP_A)$.

If the variance in performance after controlling for the explanatory attribute, $\mu(\sigma^2(AP_{A,e_j}) \forall e_j \in E)$, is less than the original variance in performance, $\sigma^2(AP_A)$, there is strong evidence that a possible reason for the measured performance discrepancy in the sensitive attribute is due to the known poor performance of this explanatory attribute in object detection. However, if the variance still persists, it shows that while this explanatory attribute may be correlated with the sensitive attribute, it is not substantially affecting the performance of the model.

3.3. Common Explanatory Attributes in Detection

Our approach relies on a set of explanatory attributes. We propose to leverage the vast literature on factors that are common failure modes of object detection systems, as a guide to further investigate and explain performance discrepancies across a sensitive attribute. We now detail a (non-exhaustive) list of factors which are known to negatively impact detection performance: Instance Size [1,5,21,47], Occlusion [12,27,31], Crowded scenes [7,19,44], Illumination [6,42], Weather [13,26], Contrast [20,45], Scene Layout [2,38], Aspect Ratio [25,35,39].

Many of these attribute values are implicit with object ground truth labels (*e.g.* instance size, aspect ratio, contrast, scene illumination, occlusion, and crowds can all be exactly or approximately measured). Other quantities like the weather or time of day can be usually extracted from the meta data associated with the captured images.

4. Experiments

4.1. Dataset & Implementation Details

We evaluate our framework on the BDD100K Driving Dataset [46] which contains bounding box annotations for 100k images and 13 classes (‘bicycle’, ‘bus’, ‘car’, ‘motorcycle’, ‘other person’, ‘other vehicle’, ‘pedestrian’, ‘rider’, ‘traffic light’, ‘traffic sign’, ‘trailer’, ‘train’, ‘truck’), from diverse scene types, weather conditions, and times of day.

We study how object detection performance varies across the sensitive attribute of income level (A). We limit our data to New York City to normalize for cost of living. To obtain income annotations, we follow REVISE [37], and correlate the GPS coordinates for each image (available for 68% of the dataset) with a median income ranging from $21.4k up to $250k. After associating these images with a median income value, we construct three equally-sized income bins: low-income ($21.4k-$62.1k), middle-income ($62.1k-$93.4k), and high-income ($93.4k-$250k).

		AP_a ($A =$ Income)			
Class	AP	low	middle	high	$\sigma(AP_A)$
All Classes	36.5	38.4 ±0.1	37.3 ±0.2	35.4 ±0.1	1.52
Car	49.8	48.7 ±0.1	49.6 ±0.1	53.2 ±0.1	2.38
Pedestrian	34.4	38.6 ±0.2	37.3 ±0.2	29.7 ±0.1	4.80
Truck	45.0	44.8 ±0.3	45.7 ±0.3	48.5 ±0.2	1.90
Traffic Sign	37.3	38.0 ±0.1	37.8 ±0.1	35.2 ±0.1	1.53
Traffic Light	25.1	25.6 ±0.1	25.8 ±0.1	23.5 ±0.1	1.27
Bus	48.2	51.3 ±0.4	49.3 ±0.5	46.3 ±0.4	2.49
Motorcycle	24.3	26.7 ±0.8	25.5 ±0.8	23.9 ±0.6	1.43
Bicycle	25.4	33.5 ±0.7	26.6 ±0.8	23.6 ±0.3	5.05

Table 1. Analysis of model performance on Income attribute across all 8 classes (row 1), and on each class independently in BDD100K. For each row we measure overall AP within each of the three sensitive attributes (‘low’, ‘middle’ and ‘high’ income) and standard deviation in performance across these values.

We use the Detectron2 library [43] and employ with a Faster-RCNN model [28] pretrained on ImageNet [30] with a ResNet [15] + FPN [22] backbone. We finetune this model on the 70k training set images in BDD100K with GPS annotations and evaluate on the validation set of 20k images. We limit our evaluation to 8 classes (‘bicycle’, ‘bus’, ‘car’, ‘motorcycle’, ‘pedestrian’, ‘traffic light’, ‘traffic sign’, ‘truck’), removing classes with very few examples or high label noise. To verify the effectiveness of the trained detector, we compute mean average precision (mAP) across all 8 classes in the BDD100K validation set, observing an mAP of 36.5, which is on-par with modern baselines.

4.2. Results

First, we compute both example performance per explanatory attribute (AP_{e_j}) as well as the distribution of each explanatory attribute by income level ($P(E|a_i)$) for all classes. A visualization of the results for the Car class (Fig. 2) and the Pedestrian class (Fig. 3) are provided as reference. We restrict our study in the main paper to the ‘car’, ‘pedestrian’, and ‘truck’ classes (rest in appendix) as they exhibit a potential bias (indicated by high variance) and have sufficient instances in each sensitive value even after controlling for the explanatory attribute, allowing for reliable AP estimates.

Next, we compute mean AP on the validation set over the 8 classes from BDD100K at a 95% confidence interval for each income value (‘low’, ‘middle’, and ‘high’), and the variance in performance of these values as described in Sec. 3. Results are presented in Table 1. As shown, we get mAP values of 38.4 ±0.1, 37.29 ±0.2, and 35.43 ±0.1 for the low, middle, and high income values respectively (Table 1, row 1). Notice that as income increases, performance *decreases*. This inverse relationship between performance and

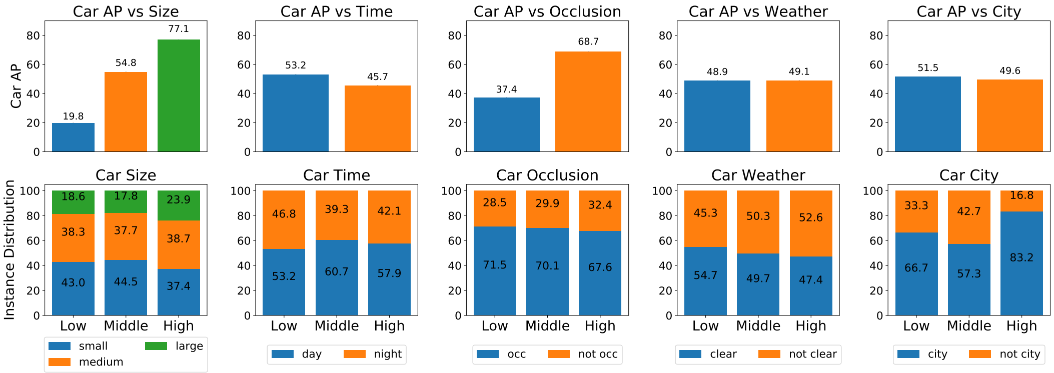

Figure 2. Analysis of the **Car** class in BDD100K. Performance vs explanatory attribute (AP_{e_i}, *top row*) and the distribution of the explanatory attribute values within each sensitive attribute value ($P(E|a)$, *bottom row*).

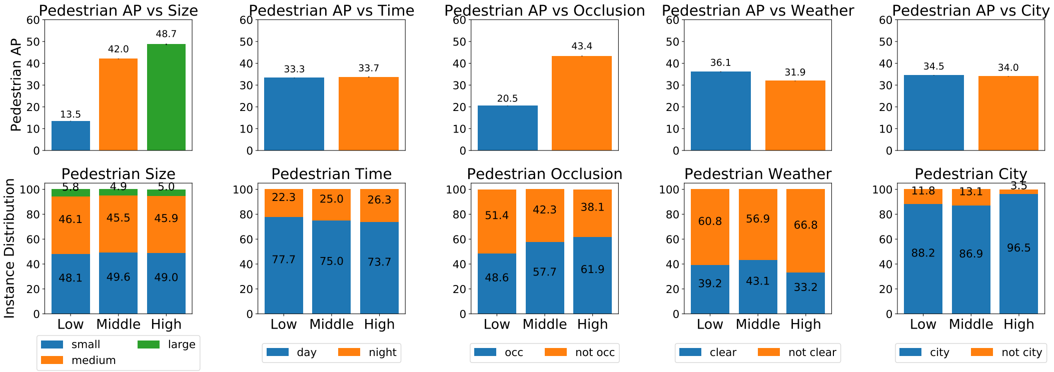

Figure 3. Analysis of **Pedestrian** class in BDD100K. Performance vs explanatory attribute (AP_{e_i}, *top row*) and the distribution of the explanatory attribute values within each sensitive attribute value ($P(E|a)$, *bottom row*).

income suggests that the model could be performing worse on higher income groups, but does not account for underlying confounders that may explain away this discrepancy. Moreover, each of the 8 classes do not follow the same relationship (*e.g.* income and performance are directly correlated for the car class). We now proceed to run our proposed method ICON^2 to discover and control for confounders in order to obtain reliable fairness estimates.

4.2.1 Ranking Confounders

To simplify analysis, we consider one class at a time. For each class, we identify and rank explanatory attributes that could explain away the observed performance gap across income levels. We consider five (non exhaustive) explanatory attributes (described in Sec 3.3) as possible confounders, to study further (instance 'size', 'time of day', 'occlusion', 'weather', and 'scene') for which we have dataset annotations. We follow the strategy described in Algorithm 1 to rank potential confounders in descebding order of performance variance in ProxyAP. We report these ordered values for the car, pedestrian, and truck, in the second column of Table 2. As seen, our preliminary ranking indicates that size is likely to be the largest confounder for the car and truck classes, whereas for pedestrian it is likely to be occlusion. Note that large $\sigma(\text{ProxyAP}^E)$ (greater than 1) for the top ranked attribute indicates that E is a more probable explanation for variance in performance across income. We hypothesize that attributes with lower rankings and a smaller $\sigma(\text{ProxyAP}^E)$ do not contribute substantially to the observed performance discrepancy.

4.2.2 Controlling for Confounders

Next, we follow the methodology proposed in Section 3.2.2 and compute $\mu(\sigma^2(AP_{A,e_j}))$ for the 'car', 'pedestrian', and 'truck' classes. To do so, we compute controlled AP (results for car and pedestrian classes reported in Fig. 4 and Fig. 5), and summarize the reduction in variance after controlling for each confounder in Table 2. We find:

Car: The 'size' explanatory attribute is ranked as the most probable (Table 2a) reason for the large (2.38, Table 1) variance in performance for the car class. Notice that after controlling for size, the variance in performance reduced by 1.18 (Table 2a), which is about a 50% reduction in variance! This validates that a significant contributing factor to

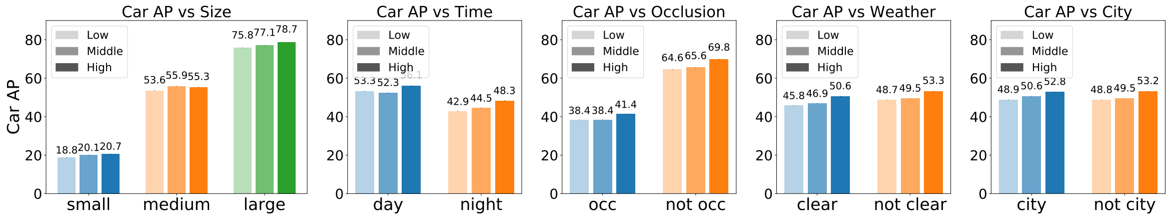

Figure 4. **Controlled AP** of the **Car** class across income levels while controlling for 5 explanatory attributes.

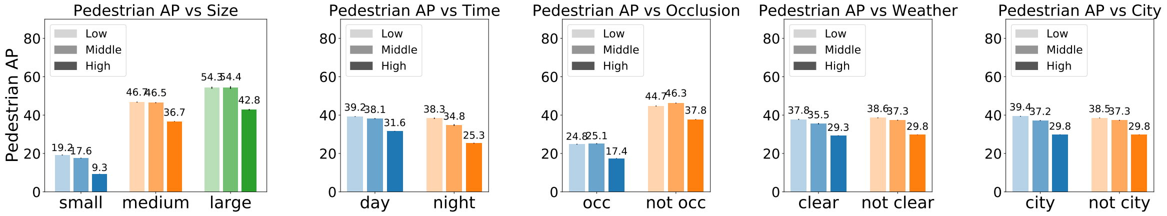

Figure 5. **Controlled AP** of the **Pedestrian** class across income levels while controlling for 5 explanatory attributes.

$\sigma(AP_A) = \mathbf{2.38}$

	E	σ(ProxyAP^E)	$\mu(\sigma(AP_{A,e_j}))$	Δ
1	size	**1.65**	**1.20**	**1.18**
2	occlusion	0.51	2.24	0.14
3	time	0.23	2.38	0.00
4	scene	0.20	2.17	0.22
5	weather	0.00	2.50	-0.11

(a) Car

$\sigma(AP_A) = \mathbf{4.80}$

	E	σ(ProxyAP^E)	$\mu(\sigma(AP_{A,e_j}))$	Δ
1	occlusion	**1.27**	**4.45**	**0.35**
2	size	0.21	5.89	-1.09
3	weather	0.17	4.59	0.21
4	scene	0.02	4.85	-0.05
5	time	0.01	5.42	-0.62

(b) Pedestrian

$\sigma(AP_A) = \mathbf{1.90}$

	E	σ(ProxyAP^E)	$\mu(\sigma(AP_{A,e_j}))$	Δ
1	size	**1.21**	**1.02**	**0.89**
2	scene	0.44	1.86	0.05
3	weather	0.37	1.82	0.08
4	occlusion	0.34	2.15	-0.25
5	time	0.09	2.24	-0.34

(c) Truck

Table 2. We show reduction in variance after controlling for the 5 chosen explanatory attributes on 3 classes (car, pedestrian, and truck). Notice that the top ranked attribute consistently leads to the largest reduction in variance in performance. For car and truck this corresponds to the size explanatory attribute while for pedestrian it corresponds to the occlusion explanatory attribute.

the previously measured predictive inequity across income levels is the fact that the object size distribution varies significantly across income regions.

Pedestrian: Unlike car and truck, the top ranked attribute for the pedestrian class is 'occlusion'. After controlling for occlusion, we notice that the variance reduces by 0.35, which is about ∼10% (Table 2b). This validates that there exists some spurious correlations between the occlusion and income attribute. However, it is likely that there are other attributes that are also causing performance gaps. For example, controlling for weather (the 3rd ranked attribute) reduces variance by 0.21 (full results in appendix).

Truck: Similarly, the 'size' explanatory attribute is ranked as the most probable reason for the 1.9 variance in performance for the truck class. We notice that variance in performance reduces by 0.89 (Table 2c), which is also about a 50% reduction in variance! Therefore, small objects are spuriously correlated with income and partially explain away the observed performance discrepancies across income levels for the truck class.

Overall, this study shows that it is important to consider common performance reducing attributes to help us understand how explanatory attributes that affect object detection systems spuriously correlate with a sensitive attribute due to dataset imbalances, and how these correlations reflect in performance discrepancies in the model. Furthermore, by controlling for certain explanatory attributes one may notice that performance discrepancies decrease, revealing possible factors that affect the sensitive attribute. This in-depth analysis can provide insight into sources of performance discrepancies and serve as a guide for understanding and leading future mitigation efforts. Note that it is possible to further the study by controlling for more than one explanatory attribute at a time, however, it is important to ensure that enough data remains in each subset to guarantee reliable AP evaluations before making comparisons.

5. Conclusion

We propose a framework to reliably benchmark bias in object detection systems used for autonomous driving. We highlight the necessity of investigating common failures in object detection systems as possible confounders, and further provide steps to understand how these common deficiencies affect performance with respect to the sensitive attribute of interest. We present an in-depth study on the performance of object detection with respect to income from the BDD100K dataset, and highlight our findings on possible reasons for measured biases.

References

[1] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[2] Sid Yingze Bao, Min Sun, and Silvio Savarese. Toward coherent object detection and scene layout understanding. *Image and Vision Computing*, 29(9):569–579, 2011.

[3] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, Oct. 2018.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.

[5] Chenyi Chen, Ming-Yu Liu, Oncel Tuzel, and Jianxiong Xiao. R-cnn for small object detection. In Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato, editors, *Computer Vision – ACCV 2016*, pages 214–230, Cham, 2017. Springer International Publishing.

[6] Winston Chen and Tejas Shah. Exploring low-light object detection techniques. *CoRR*, abs/2107.14382, 2021.

[7] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[8] Marie Pellat Daniel Smilkov, Nikhil Thorat and Ludovic Peran. Know your data. `https://knowyourdata.withgoogle.com/`, 2019.

[9] Abhijit Das, Antitza Dantcheva, and Francois Bremond. Mitigating bias in gender, age and ethnicity classification: a multi-task convolution neural network approach. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[10] Terrance Devries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? *ArXiv*, abs/1906.02659, 2019.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[12] Tianshi Gao, Benjamin Packer, and Daphne Koller. A segmentation-aware object detection model with occlusion handling. In *CVPR 2011*, pages 1361–1368, 2011.

[13] K. Garg and S.K. Nayar. Detection and removal of rain from videos. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I, 2004.

[14] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, et al. Amazon sagemaker clarify: Machine learning bias detection and explainability in the cloud. *arXiv preprint arXiv:2109.03285*, 2021.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[16] Derek Hoiem, Yodsawalai Chodpathumwan, and Qieyun Dai. Diagnosing error in object detectors. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *Computer Vision – ECCV 2012*, pages 340–353, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[17] Jonathan Tannen Isabel Kloumann. Facebook's fairness flow. `https://ai.facebook.com/blog/how-were-using-fairness-flow-to-help-build-ai-that-works-better-for-everyone`, 2021.

[18] Arvind Krishnakumar, Viraj Prabhu, Sruthi Sudhakar, and Judy Hoffman. Udis: Unsupervised discovery of bias in deep visual recognition models. In *British Machine Vision Conference (BMVC)*, November 2021.

[19] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 878–885 vol. 1, 2005.

[20] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[21] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[22] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.

[23] 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 *ECCV*, 2014.

[24] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3384–3393. PMLR, 10–15 Jul 2018.

[25] Subhransu Maji and Jitendra Malik. Object detection using a max-margin hough transform. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1038–1045, 2009.

[26] Srinivasa G. Narasimhan and Shree K. Nayar. Vision and the atmosphere. *International Journal of Computer Vision*, 48:233–254, 2004.

[27] Julian Pegoraro and Roman P. Pflugfelder. The problem of fragmented occlusion in object detection. *CoRR*, abs/2004.13076, 2020.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region

proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[29] Olga Russakovsky, Jia Deng, Zhiheng Huang, Alexander C. Berg, and Li Fei-Fei. Detecting avocados to zucchinis: What have we done, and where are we going? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[31] Kaziwa Saleh, Sándor Szénási, and Zoltán Vámossy. Occlusion handling in generic object detection: A review. *CoRR*, abs/2101.08845, 2021.

[32] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[33] Andrea C Skelly, Joseph R Dettori, and Erika D Brodt. Assessing bias: the importance of considering confounding. *Evidence-based spine-care journal*, 3(01):9–12, 2012.

[34] Sruthi Sudhakar, Viraj Prabhu, Arvind Krishnakumar, and Judy Hoffman. Mitigating bias in visual transformers via targeted alignment. In *British Machine Vision Conference (BMVC)*, November 2021.

[35] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *2009 IEEE 12th International Conference on Computer Vision*, pages 606–613, 2009.

[36] S. Verma and J. Rubin. Fairness definitions explained. In *IEEE/ACM International Workshop on Software Fairness (FairWare), IEEE. pp. 1–7.*, 2018.

[37] Angelina Wang, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. In *ECCV*, 2020.

[38] Tao Wang, Xuming He, Songzhi Su, and Yin Guan. Efficient scene layout aware object detection for traffic surveillance. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 926–933, 2017.

[39] Xiaoyu Wang, Ming Yang, Shenghuo Zhu, and Yuanqing Lin. Regionlets for generic object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[40] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[41] Benjamin Wilson, Judy Hoffman, and Jamie H. Morgenstern. Predictive inequity in object detection. *ArXiv*, abs/1902.11097, 2019.

[42] Cheng-En Wu, Yi-Ming Chan, Chien-Hung Chen, Wen-Cheng Chen, and Chu-Song Chen. Immvp: An efficient daytime and nighttime on-road object detector. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5, 2019.

[43] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.

[44] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] Xin Xu, Nan Mu, Hong Zhang, and Xiaowei Fu. Salient object detection from distinctive features in low contrast images. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3126–3130, 2015.

[46] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving video database with scalable annotation tooling. *CoRR*, abs/1805.04687, 2018.

[47] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. *CoRR*, abs/1912.10664, 2019.

[48] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018.

[49] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *CoRR*, abs/1707.09457, 2017.

Appendix

A. Limitations

We attempt to explain measured biases by considering common detection failures. This has two key limitations. First, we require a set of known biases with associated labels within our data to perform the specified computations. This is not overly restrictive as many of the explanatory attributes we consider are implicitly defined by standard ground truth labels for detection (*i.e.,* object size, occlusion) or are commonly collected as meta data (*i.e.,* time of day). However, undoubtedly defining more attributes and possible factors of variance will lead to more explanation of observed bias. Our framework can be used with any number of explanatory attributes and is not limited by the 5 we illustrate in this work. The second key limitation is that we can not guarantee that our framework reveals *all* possible sources of bias. This will be limited both by our explanatory attribute set as well as the fact that we currently only consider one explanatory attribute at a time. Future expansions of our approach may consider combinations of attributes. The dataset we study, BDD100K, did not have

	Traffic Sign		Traffic Light		Bus		Motorcycle		Bicycle	
Rank	E	σ(ProxyAP^E)	E	σ(ProxyAP^E)	E	σ(ProxyAP^E)	E	σ(ProxyAP^E)	E	σ(ProxyAP^E)
1	size	0.214	time	0.230	size	0.881	size	0.372	size	0.715
2	weather	0.159	size	0.205	occlusion	0.700	occlusion	0.293	weather	0.456
3	occlusion	0.119	scene	0.061	weather	0.242	weather	0.138	time	0.388
4	time	0.069	weather	0.060	scene	0.239	scene	0.032	occlusion	0.201
5	scene	0.021	occlusion	0.068	time	0.060	time	0.023	scene	0.041

Table 3. Explanatory attribute output ranking for the remaining 5 classes in the BDD-100K dataset. Note: all proxyAP variance values are smaller than the discovered explanatory attributes for car, pedestrian, truck (Table 2) and hence we did not find a compelling explanation for our sensitive attribute (income), however for the bus, motorcycle, bicycle classes the counts are too low to compute reliable AP values.

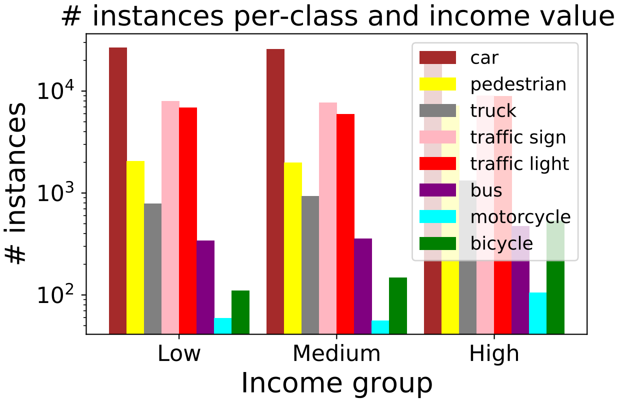

Figure 6. Instance counts of each class in each income value (*i.e.* low, middle, and high income). Note traffic sign, traffic light, bus, motorcycle, and bicycle have very few instances for low/middle income (there are even fewer instances when controlling for explanatory attributes).

sufficient samples across combinations of explanatory attributes to compute meaningful AP values. Finally, we note that we discover correlations and do not claim that there exists a definitive casual relationship between a chosen explanatory attribute and a sensitive attribute.

Bias Mitigation. The insights revealed by our framework may be used to guide subsequent mitigation strategies. Say our method discovers an explanatory variable E that explains a large proportion of performance variance across a sensitive attribute A. The discovered explanation may imply either: 1) model bias and/or 2) data bias. Consider, for example, that for the sensitive attribute, A, of income level, one discovers that explanatory attribute, E, of object size explains a significant amount of the performance discrepancy across income levels. This would imply an appropriate intervention would be to leverage a detection model with stronger performance on smaller objects. As a second example, consider an explanatory attribute of time of day. In our experiments we found that the dataset collection had disproportionately more night-time images in the low income regions, possibly related to how the dataset was collected. Rather than being an innate aspect of the scene types in the different geographic regions, this implies a bias in the data collection process. Hence, an effective mitigation strategy would be to collect a more equitably distribution of day and nighttime images across all socioeconomic regions.