Unmasking Inequality: Analyzing the Fairness of Face Mask Detection

Group KoMoFiHu: Olusanmi Hundogan, Christian Moll, Ivan Kondyurin, Giacomo Fiorentini
{i.kondyurin,c.e.moll,g.fiorentini,o.a.hundogan}@uu.nl
Utrecht University
Utrecht, the Netherlands

ABSTRACT

Many researchers have shown how datasets which are biased can propagate the bias to downstream tasks. In this report we investigate the effects of the well-known ImageNet dataset for image classification. We employ a YOLOv4 model which was pretrained on this biased dataset and examine the effects on object detection as downstream task. We show that an object detection model retains their bias even if finetuned on a balanced dataset. This effect is not necessarily visible with common Fairness metrics such as TPR or PPV, but on average precision (AP).

ACM Reference Format:
Group KoMoFiHu: Olusanmi Hundogan, Christian Moll, Ivan Kondyurin, Giacomo Fiorentini. 2021. Unmasking Inequality: Analyzing the Fairness of Face Mask Detection. In *Proceedings of Utrecht University (INFOMHCML'2021).* Utrecht University, 8 pages.

1 INTRODUCTION

Consider the following scenario. You and four other friends decide to visit a mall in broad daylight. As soon as you enter the alarm bell rings. The people around you stare at you with unease and security guards approach you. They decide to question two of you. All of you are let go as it becomes clear that this was a case of a false-positive detection.

This scenario raises multiple questions. How could a security system react immediately after your entrance? What makes two of your group stand out as more dangerous than the rest? And most importantly, is it likely to happen again? This scenario does not require any sensitive attributes nor details about the target object the model was scanning for to explain why automated object detection systems need to be reliable, fair and explainable. Yet, these aspects are often neglected in favor of efficiency. In this paper we try to address the latter two questions in detail. First, by answering if object detection models can maintain a bias which is ingrained in the dataset. Second, by explaining why some detections fail and other succeed.

We aim to determine whether the bias in original object detection model can penetrate into downstream tasks, such as finding a particular object, thus causing allocative harm in certain domains. For this research we have chosen the task of face mask detection, because fallacious object detection in this domain can lead to unfair treatment and violation of some rights due to erroneous fines or denied access to some facilities in case wearing a mask is mandatory. This problem has recently become increasingly important due to the COVID-19 pandemic and associated restrictions.

To prove that object detection model trained on ImageNet[1] dataset preserves its bias, we compare its performance after finetuning for the task on face mask detection on three types of datasets: an original, biased face mask dataset, a balanced version w.r.t race and gender, and augmented debiased dataset in which we performed image upscaling.

We show that certain unfairness persists even when the model is fine-tuned on the debiased data. We regard this as evidence for bias penetration. However, we also prove that debiasing procedure can improve the model's performance for some originally underrepresented groups. Also, the by-product of our experiment is a new version of face mask detection dataset which is labeled with sensitive attributes and equalized in terms of race and gender. It can be used for future experiments in debiasing object detection models as well as in implementing more fair practical solutions.

Section 1.1 provides a brief overview of literature regarding the observed bias in ImageNet, most common fairness criteria for Machine Learning models, and the explainability methods, in particular D-RISE, that we are using to facilitate the interpretation the model's performance. In section 2 we describe the available dataset, the preprocessing steps that we have taken, the debiasing procedure and the technique for visual explanations. In section 3 we provide the scores for biased, debiased, and augmented models. After that, potential limitations of our model are discussed.

1.1 Related Literature

1.1.1 Bias in ImageNet. ImageNet is one of the largest publicly available image datasets containing around 14 million manually annotated images which is the reason for its far-reaching popularity. Many state-of-the-art image-classifying and object recognition AI models are pre-trained on ImageNet due to its enormous size generally leading to good performance. However, a recent study conducted by Steed and Caliskan has shown that ImageNet contains various stereotypical biases such as race and gender, likely attributable to data imbalance [22]. More specifically, their findings call out for caution to computer vision practitioners when it comes to transfer learning since pre-trained models potentially propagate biases contained in training data into downstream tasks. Depending on the field of application this can cause various harms. An example are face recognition systems for which various studies already have shown that these are susceptible to bias [7]. Independent benchmarks by the Gender Shades project and the National Institute of

[1] ImageNet dataset: https://www.image-net.org/

INFOMHCML'2021, April 2021, Utrecht, the Netherlands
© 2021 Utrecht University
ACM ISBN xxxxxxx.
https://doi.org/xxxxxxx

Standards and Technology (NIST) have demonstrated that facial recognition technology exhibits racial and gender bias and have suggested that current facial recognition programs can be wildly inaccurate and still struggle to identify black faces [14].

A study by Facebook, Twitter, and LinkedIn has examined how training data bias propagation can affect object recognition systems in particular [13]. Even though people per se were not subject of the detection interest within the study, they have shown that object detection systems might not work well for everybody depending on demographics. Popular datasets such as ImageNet, COCO[2], and OpenImages[3] were found to have skewed geographic distribution since the majority of images come from Europe and North America, whereas relatively few images come from populous regions in Africa, South America, and Central and Southeast Asia [11].

Our study puts the human in a central role since the detection of interest is if a face mask is present in a persons face or not. To the best of our knowledge, no research so far examined the effect of training data bias propagation to the object recognition task with the objects of interest being worn by a person.

1.1.2 *Fairness in Machine Learning.*

Machine Learning/AI systems deployed in production can cause various harms when exhibiting biases. It is not difficult to see the unfairness when an AI system within the recruitment procedure favors males over females [10] or another AI system assisting judges rates black people being more likely to re-offend than white people [19]. An AI system that would fail more often to detect face masks in the face for black people than for white peoples is similarly unfair. Not only that, but also harmful when such a system is responsible e.g. to grant access to locations such as the subway. These kinds of harms are of allocative nature and are caused when a system withholds certain groups an opportunity or resource [2].

The topic of fairness is gaining popularity in machine learning in recent years. This seems quite reasonable when considering the particular fields in which machine learning is applied. Over the past years/decade, the deployment of data-driven AI systems in high-stakes fields is soaring. To name a few, such fields include healthcare, loan-granting, hiring and education [3] [16]. While research progress of fairness on an individual level is slowed by various difficulties of measuring individual fairness, some valuable metrics emerged to define and measuring fairness on a group level [2]. There are three major fairness metrics: **Independence**, **Separation**, **Sufficiency**. Independence, also known as demographic or statistical parity, measures the rate of favorable outcomes of individuals across groups. Separation, also known as equal opportunity or equalized odds and positive rate parity, is measured by comparing the true positive rates (TPR) and false positive rates (FPR) across groups. Equal opportunity is achieved when the TPR across groups is equal and equalized odds is achieved when both TPR and FPR are equal across groups. Sufficiency, also known as predictive parity, is measured by comparing the precision/positive predicted value (PPV) across groups.

Under mild assumptions any two of these three criteria are mutually exclusive [9]. Defining fairness with any of those criteria therefore depends on the application at hand. For our study we aim to measure fairness in terms of separation for the following reason: If a person wears a mask or not is a hard fact and no subject of debate since the images clearly display people wearing a mask or not. We therefore can consider the ground truth to be unbiased which leads to the selection of the separation criteria that is referred to as bias preserving. A fair model detecting face masks in peoples faces should perform equal in terms of TPR and FPR regardless of the sensitive attributes race and gender.

However, the object detection task differs in certain ways when looking at metric in comparison to usual binary classification. In object detection, a true positive is a correct detection, a false positive is a wrong detection and a false negative when a ground truth is not detected. True negatives generally not apply because there are in theory an infinite amount of correctly not detected bounding boxes within an image. Hence, it is not possible to compute FPR. We therefore focus on Equal Opportunity as main metric.

The average precision (AP) metric, which is the approximated area under the precision-recall curve will also be reported. This metric allows to quickly quantify precision and recall across various decision thresholds. As the metric combines precision and recall, it covers notions of separation and sufficiency.

1.1.3 *Interpretability and Explainability.*

Being able to interpret a machine learning model makes it possible to explain why and how it comes to its conclusions and decisions. Besides the prior mentioned fairness criteria to determine the overall fairness of a model, being able to explain the decisions of a model helps to get to the bottom of the models fair or unfair behavior [12].

Image classification and object detection models are typically Deep Neural Networks (DNN). DNNs are considered black boxes due to their multi-layer nonlinear structure [5]. They consist of many different units that all together perform many complex mathematical operations impossible to grasp for humans by simply looking at specific parts or results. Neural networks often contain millions of neurons. Therefore, post-hoc methods are required to interpret such systems and explain their results. Various interpretation methods were developed specifically for DNNs and two specific types are referred to as **feature visualization** and **pixel attribution** methods. Both types of methods generate explanations in a visual manner that provide an aid to understand which aspects of a certain image leads a model to its decisions.

For the creation of our visual explanations we decided to use a pixel attribution method named D-RISE [20]. D-RISE is a perturbation based method that generates saliency maps by repeatedly masking different areas of a given image to highlight the important pixels that led to the models detection. Since D-RISE does not need access to model internals it is more general and agnostic to the particular type of object detectors in contrast to other methods [20]. Furthermore, it is one of the only methods that is specifically made for object detection models. The majority of pixel attribution methods are rather made for the image classification task and can not be directly applied to the detection task. The advantage of D-RISE being made for object detectors is therefore that it enables to generate saliency maps for every detection made in a certain image. Section 2.3 goes into more detail about the functionality and specific configurations we have used for D-RISE to create our visual explanations.

[2] COCO dataset: https://cocodataset.org/home

[3] OpenImages dataset: https://opensource.google/projects/open-images-dataset

1.1.4 Object Detection Models. In simple terms, object detection is the task of locating the presence of an object of a certain class within an image. Object detection models are trained on datasets consisting of image and annotation files. Typically, every image is accompanied by a distinct annotation text file. The annotation files specify for each object that the model should be able to identify the respective class and bounding box coordinates of the object in the image. Plenty high performance object detection models exists with most of them having an architecture based on deep convolutional neural networks (CNN). The model chosen for this study is one of the most popular object detection algorithms called You Only Look Once (YOLO). The YOLO model, initially published by Redmon et al. in 2016 [21] underwent multiple evolutionary steps to this day. From the first version, also denoted as YOLOv1, four more major versions, YOLOv2 to YOLOv5 were published since then, with a couple of sub-versions. The fourth version of YOLO has been released in April 2020 and introduced in a paper by Bochkovskiy, Wang, and Liao[4] which continued Redmon's work.

We decided to use YOLOv4 since it is the last implementation that was pretrained on ImageNet. More specifically, the model uses the CSPDarkNet53 backbone, which is part of the open source framework Darknet. In contrast, YOLOv5[4] uses a different backbone and is not pre-trained on ImageNet and therefore not useful for our research question.

A notable improvement of YOLOv4 in comparison to its predecessors is the implementation of the *bag of freebies*. The term refers to a set of methods applied during the training procedure that aim to improve the models accuracy considerably [4]. Especially interesting are the subset of methods that perform data augmentation which aim to increase the variability of images in order to improve the generalization of the model training. Among others, this subset includes photometric and geometric distortion methods that creates new instance of an image by adjusting brightness, hue, contrast, saturation and by rotating, flipping and randomly scale or crop an image. These and the other bag of freebie methods can potentially assure a high accuracy of the model, even with an relatively small fine-tuning dataset.

Another important factor to note is that the publicly available pre-trained YOLOv4 model is not able to detect surgical face masks out of the box. This fact made it necessary to fine-tune the model on a face mask dataset. Fortunately, fine-tuning requires substantially less training data than training a model from scratch. Many face mask datasets are publicly available e.g. on Kaggle. However, this step potentially induces biases contained in the fine-tuning set, such as group imbalance, different image resolutions and bounding box sizes across groups. We see those aspects as potential confounding factors that complicate drawing conclusions about the effect of the actual ImageNet bias on the results. Subsequent sections will explain in detail our efforts to mitigate those confounding factors.

1.2 Research Question

As previously discussed, there is profound evidence that image based AI systems exhibit biases inherited in training data. This can cause allocative harms such as that a specific system perform significantly worse for minority groups. However, these harms may not apply if the fine-tuning dataset is fairly balanced. The question therefore becomes: Do models that are pre-trained on ImageNet propagate their bias through downstream tasks? With our study we aim to show that some deeply rooted ImageNet biases prevail despite efforts to post-hoc modify the model to be more fairly balanced. Ultimately, we want to answer this question by analyzing the results of our fine-tuned YOLOv4 model in accordance to the chosen fairness criteria.

2 METHODS

2.1 Data collection

2.1.1 Available resources overview. As mentioned above, various datasets for face mask detection on people are publicly available on Kaggle, with most of them being less than one year old. Among the most popular ones are *Face Mask Detection* by Larxel that contains 853 images annotated with boxes and three category labels and *Face Mask Detection Dataset* by the user Wobot Intelligence that contains 6024 images, also annotated. We have chosen the Wobot Intelligence dataset [23]. The decision was made for several reasons. First of all, it is the largest face mask dataset that contained images supplemented with bounding boxes for objects. Besides, the dataset contains rich annotations in JSON format that not only contain the labels and bounding boxes but also indicate how many people are visible on a certain image. On considerable amount of images there is only one person, which facilitates the task of labeling sensitive attributes. The dataset already shows some diversity in terms of race and gender. Additionally, it contains over 10 different sub-categories of masks, such as colored or surgical, which is beneficial for the robustness of the model, since the training data can include different possible types of face covers. Our research problem required the analysis of categorical bias with respect to several sensitive attributes, namely race, skin color, and gender, to draw conclusions about fairness. This information was not included in any of the available face mask datasets, including the Wobot Intelligence dataset, which necessitated additional labeling of our dataset of choice. Since the link between the annotated bounding box and corresponding person is often ambiguous in images with multiple people, we decided to confine out research to a subset of 2165 single-person images. They were automatically extracted by filtering those that only contain one of the following labels: "face no mask" and "face with mask". Our labeling approach is explained in detail in subsequent section 2.1.2.

2.1.2 Labeling. In total we labeled 2063 single-person images by their perceived race, skin and gender. It is important to point out that we only provide perception-based labels, since neither ground truth nor any factual information such as name or birthplace were available for this data. The difficulties and potential limitations for each attribute will be addressed below.

To label races, we combined the approach suggested by recent research in racial categories for the task of computer vision [17] with the traditional U.S Census Bureau race classification [8] which adheres to the 1997 Office of Management and Budget (OMB) standards on race and ethnicity. The latter points out the following races: White, Black or African American, Asian, American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander. However, in the available datasets there were only a few isolated examples

[4] YOLOv5 GitHub Repository: https://github.com/ultralytics/yolov5

of American Indian and Native Hawaiian people, and the search of several hundred images of masked people in these categories turned out to be unproductive. Therefore, during the labelling procedure we combined these two groups, as well as other groups with insufficient number of instances, into "Other". Additionally, the Asian category turned out to be significantly more diverse than any other category in terms of skin color, hair, eye shape, and other characteristics. Following the classifications suggested by Khan and Fu [17], we decided to split this group into two more integral subgroups, namely South Asian and East Asian (referred to as Asian in [17]). This subdivision was of particular importance for the goal of our current task, because in East Asia people tend to be more accustomed to wearing masks in public, and therefore more high-quality photos are available. The group of South Asian people, on the contrary, is more exposed to potential bias due to a smaller number of masked people photos publicly available. In the end, we used the following classification and encoding: Caucasian, Black, East Asian, South Asian, and Other. The skin color labels were provided based on Fitzpatrick scale, as suggested by Buolamwini and Gebru[6]. To simplify the classification and ensure that the resulting categories are large enough to compare the model performance on them, we decided to use a map th Fitzpatrick scale to a simplified scale with three categories: white (corresponding to Fitzpatrick skin types I and II), brown (skin types III and IV) and black (types V and VI). A similar technique was used by Buolamwini and Gebru with a binary scale for skin types I-III and IV-VI. It was noted during the labelling that despite a strong correlation between the race and the skin color, most racial categories included people from at least two skin color categories, thus showing that the multi-dimensional approach makes the classification more refined. For gender labelling we adapted a simplified binary scale used by Buolamwini and Gebru [6] despite the obvious limitations mentioned in that work. The gender was assigned based on perceived characteristics, such as facial and body features, hair type, facial hair, and sometimes clothes. The four human annotators relied on their subjective perception of these three categories. However, when a controversial instance was found, it was marked as "-1" in the respective category and discussed. If no agreement on most plausible label was achieved, this instance was marked as controversial and not included in the training data. Table 1 shows the count of images for each of the sensitive attributes.

Table 1: the counts of the full dataset pivoted by Race and Sex. "Others" are not included.

	Mask_on	With Mask	Without Mask
Sex	Race		
Female	Asian	258	81
	Black	51	45
	Caucasian	532	110
	South-Asian	82	60
Male	Asian	178	78
	Black	82	54
	Caucasian	381	123
	South-Asian	65	62

2.2 Data pre-processing

As established prior, the value of the YOLO-family lies in their use for transfer-learning purposes. The model can be easily modified for use-cases that it has not seen before. Furthermore, it does not require much data to learn reasonable detections. In this study we make use of these capabilities by fine-tuning the model on our tasks setting. As explained, the model will be trained on detecting either faces without mask or masks. This differs from YOLO's default configuration which detects 80 classes from the COCO dataset. The model will be trained on our face mask dataset. However, we will use three different fine-tuning configurations. The first configuration will be trained on a dataset which has varying distributions across the protected attributes. The distribution is a sample from the initial distribution and therefore reflects the true distribution of the dataset.

The second configuration will be trained on a subset of the first configuration's dataset. Here, we ensure that each protected attribute has the same amount of data points in the training data. With this configuration we can rule out any bias-effects that the fine-tuning process itself may introduce. For this purpose, we segment the data by Race, Gender and Label. We then take the size of each segment and select the minimal value as a reference to sample data points from each segment. We remove those data points from the initial training data[5]. We will refer to this dataset as balanced or debiased dataset. The distribution of each set is shown in Figure 1.

2.2.1 Confounding factors. As part of the data pre-processing step for this project, we examined various augmentations and transformations that could increase the quality of our training dataset. For this purpose we investigated the effects of pixelation, bounding-box to image ratios, upscaling and downscaling images and conducted a number of informal experiments. The experiments show that the pixelation factor reduces the model's ability to detect objects. However, the bounding box size in relation to the full image and the resolution also play a role. Generally, if the bounding box size falls below a certain threshold of roughly 1% of the entire image the model will potentially not be able to detect the face. However, this does not necessarily hold for high-resolution images as they typically remain more detectable after downscaling below the threshold. The reason for these effects is that downscaling removes valuable information from the image and as YOLO scales every image to a fixed size, this information is irreversibly lost for low-resolution images. These experiments reveal just some of the factors that may have an impact on face mask detection. This is important, because any of these characteristics may be concentrated in one of the majority or minority groups. In order to counter some of these effects we upscale images whose width and length were below 1000x1000 pixels prior to training. We use bicubic interpolation, which is more pronounced than YOLO's bilinear upscaling. With our tests, we were not able to prove that exceedingly large images impacted the model for the purpose of training and testing. Hence, we omit the downscaling procedure.

[5] The removal might lead to some races or classes being underrepresented in the fine-tuning dataset. However, this does not happen to the detriment of our research question but in fact supports it by making the unbalanced dataset *more unbalanced.*

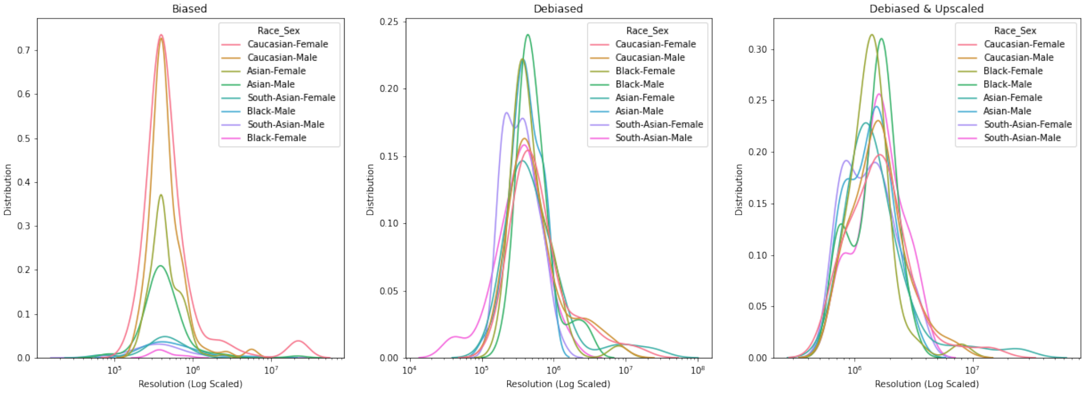

Figure 1: the distribution of the three fine-tuning sets on a logarithmic count scale. The legend shows the distribution of pixel counts per image for each race-gender segment of the data. Set 1 is unequally distributed in image counts, Set 2 is balanced, Set 3 is balanced and upscaled.

2.3 Visual Explanations

As discussed in section 1.1.3, we have decided to use D-RISE as an interpretability method for creating visual explanations to examine the model's detection behavior. The main idea behind D-RISE is to measure the effect of masking randomized regions of an image for every prediction made on the original image. By measuring the changes of the prediction behavior overall masked images for every initial prediction the pixel importances can be determined. The quality of a given visual explanation depends on three factors: the number of masks, the size and the amount of the masking areas within the masks. The amount of masks is defined by N, the size is determined by a scaling dividend s applied on the needed mask resolution and the number of areas by a probability p. Important to note is that there is no one-size-fits-all configuration that ensures precise visual explanations for every image. For example, larger masking areas usually result in quite good explanations for objects with a larger bounding box but fail to be precise for smaller bounding boxes. Generally, the more masks are used the better the visual explanations. We decided to adjust named parameters according to the images we aim to explain. Section 3.4 presents a selection of generated visual explanations for various detections made by our fine-tuned model.

3 RESULTS

3.1 Experimental Setup

For the experimental setup, we use Alexey's YOLOv4 implementation and follow the author's configuration instructions on how to train and detect custom objects. Hence, we deviate from the standard configuration by changing the maximum batches to 6000 iterations. We set the network size to 416 for both with and height. We also specify the number of objects to detect to 2 classes in each of the three YOLO-layers. All the last convolutional layers prior to each YOLO layer have a fixed number of 21 filters. As a pre-trained backbone, we download the weight-file for all convolutional layers of the original model by Bochkovskiy, Wang, and Liao[6]. We also set the batch size to 32 and subdivisions to 16.

The next hyperparameters are default settings of the darknet framework. However, we report them as is common practice in the literature. Noteworthy hyperparameters were a learning rate of 0.0013, a momentum of 0.949 and a decay of 0.0005 following a polynomial decay learning rate scheduling strategy. The steps to re-adjust the learning rate were set to 4800 and 5400 with a burn-in period of 1000.

We train a dedicated model for each fine-tuning set and evaluate their performance based on their mean average precision score on the part of the test data which was not in the test data after every thousand iteration step. We then compute all the established metrics on the balanced test dataset. As there's no real positive outcome in detecting faces or face masks we compute the weighted average versions of each metric. The weighting is based on the support (true labels) for each class and resembles the macro versions of TPR and PPV but removes the effect of class imbalances. Every model is trained for the full 6000 iterations although many peaked already after around 3000 at around 96% mean average precision.

We will also examine the area of the precision recall curve. Optimally, the areas should be the same for every group. This equality would align with the Equal-Odds fairness criterion. The amount of AP is of lesser importance due to the lack of a negative outcome. Lastly, we also report the averaged intersection over union (IoU) metric. This metric computes the overlap between predicted bounding box and ground truth.

3.2 Model fine-tuned on biased dataset

The results in Table 2 show that most models have a high average accuracy. The best accuracy was achieved for South Asian (0.99) as every single prediction was correct. This result is unexpected as South Asians are an underrepresented group in the training set. This bias may hint at biases that are rooted in the initial pre-training

[6] All weight files can be found on the Darknet GitHub page by AlexeyAB: https://github.com/AlexeyAB/darknet

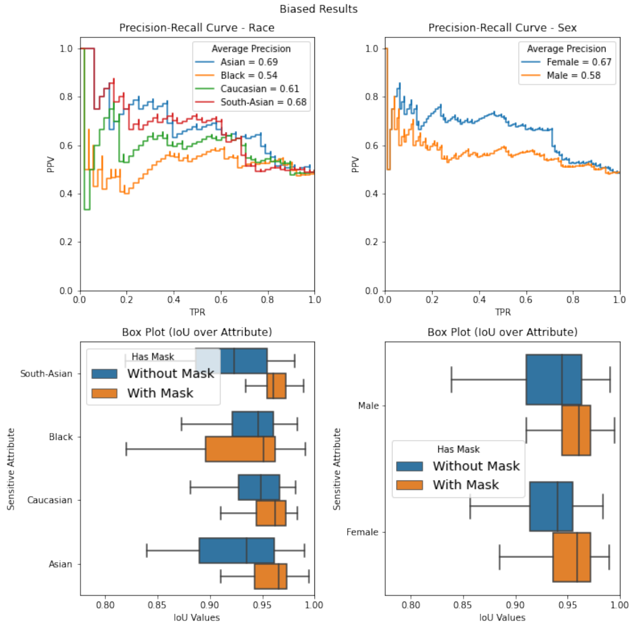

Figure 2: precision-recall curves and box plots of different groups.

dataset ImageNet. The least performing group is the Black subgroup when it comes to accuracy (0.935) and TPR (0.935). These results are expected as the Black Female group is indeed the least represented group in configuration 1. However, if we take a more pronounced look at the IoU values we see strong differences. The class-wise precision-recall curve in Figure 2 shows disparate confidences for TPR thresholds below 0.8. For Race, the S.Asian and Asian groups yield the highest AP while the Black subgroup the lowest. The Gender also reflects the disparity in accordance with the distribution. Namely, that males are less represented in the dataset than females. Figure 2 also displays model-uncertainty for South-Asian and Asian subgroups without masks, and the Black subgroup with masks. These uncertainties are either rooted in societal biases and/or in label imbalances. To answer this and further questions we turn towards configuration 2.

Table 2: the prediction results for every race and gender in case of fine-tuning on biased dataset.

Race	AP	IoU	Acc	TPR	PPV
Asian	0.690	0.929	0.942	0.940	0.947
Black	0.539	0.896	0.935	0.935	0.935
Cauc.	0.608	0.942	0.950	0.949	0.954
S.Asian	0.676	0.937	0.990	0.990	0.990

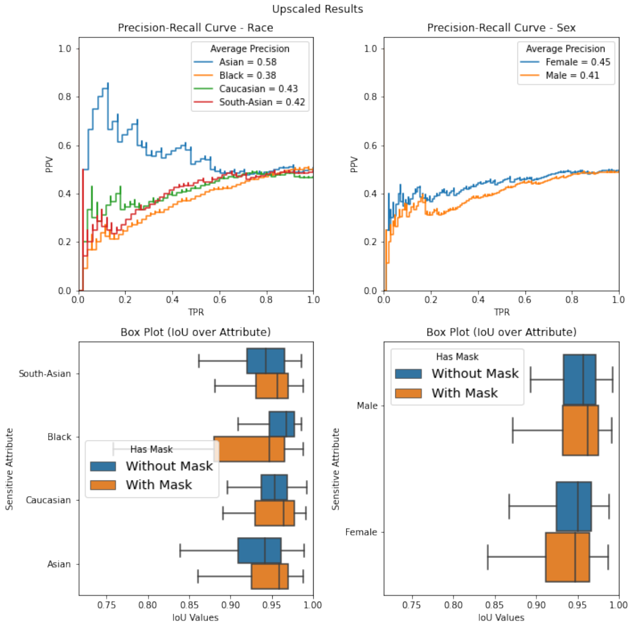

Figure 3: precision-recall curves and box plots of different groups.

3.3 Model fine-tuned on debiased dataset

In order to examine the effects of the pre-trained backbone regardless of the fine-tuning data, we show the results of configuration 3. In Figure 3 the disparities are less pronounced with the exception of the Asian subgroup. In terms of average precision, the Black subgroup retains its position at the lower end. The other subgroups have similar average precisions. However, the precision-recall curve reveals that the Asian subgroup remains highly precise even if the confidence threshold for detections is lowered. This shows that the model remains highly confident for Asian subgroups. The Asian subgroup is followed by the Caucasian subgroup. This hints at biases that may come from the initial ImageNet pretraining dataset. The dataset might have more depictions of face masks being worn by Asians or Caucasians. The IoU boxplot display more alignment in the respective distributions. However, Black Mask wearers still stand out.

For Gender, the precision-recall curves behave similarly to the prior configuration. However, the IoU distributions are more aligned. Hence, when it comes to Gender, the model performs the same regardless of the person wearing a mask or not and seemingly regardless of Gender. However, this aggregated view obscures intersectional effects as Table 3 shows. This table shows how TPR and PPV are still high for all classes. However, when examining the IoU there are differences for Black and White Males as their IoU values are generally lower than the average. With respect to AP, Caucasian females perform the strongest across all detection thresholds and Caucasian males the worst.

3.4 Explanations

In this section we present a set of saliency maps generated with D-RISE to explain a selection of detections made by our model. For

Table 3: the prediction results for every race and gender in case of fine-tuning on debiased dataset.

		AP	IoU	Acc	TPR	PPV
Race	Sex					
Asian	F	0.683	0.935	0.981	0.980	0.981
	M	0.534	0.939	0.980	0.980	0.980
Black	F	0.441	0.920	0.942	0.941	0.942
	M	0.352	0.890	0.944	0.942	0.948
Cauc.	F	0.428	0.920	0.944	0.941	0.948
	M	0.451	0.879	0.893	0.891	0.893
S.Asian	F	0.418	0.923	1.000	1.000	1.000
	M	0.463	0.928	0.981	0.980	0.981

the explanations we have considered the model fine-tuned on the upscaled and debiased dataset.

The first column in Figure 4 presents a set of images where the model correctly detects people wearing face masks. The images leads to the assumption that important parts for detecting face masks are the visible edges of the mask in combination with some skin and either parts of the ear or eyes.

The set of images in the second column show that correctly detecting no mask can be associated with the model looking at the lower part of the face. More specifically, the model seems to associate a visible mouth and nose with the person not wearing a mask. The third column shows instances where the model wrongly detected a mask. As previously stated, the visible edge of the mask in combination with some skin seem to be an aspect that leads to detect masks. Taking a closer look at those images it is visible that areas are highlighted containing edges that the model apparently confused with edges belonging to a mask.

In the first image of the fourth column no mask was detected. Still, the model correctly set the bounding box on the persons head. Unfortunately, The saliency map does not give a good explanation for this wrong detection. While examining the original image it gets apparent that a vintage like filter is applied. Also, the background of the image is quite blurred and a strong camera focus is put on the person. It is likely that such edited photographs are generally difficult for object detector models to handle. Nonetheless, this is the only image within the test set where detecting the mask was unsuccessful. The masks in the other two images of the fourth column were correctly detected by the model but with a relatively low confidence (around 0.4 and 0.2). The low confidence is presumably caused by the faces of the people within the images not being clearly visible. However, the saliency maps give quite good explanations for the correct detections.

4 DISCUSSION

In this paper we have shown that a Bias remains in a fine-tuned YOLO model even if the fine-tuning dataset was carefully balanced. Furthermore, we show that there are inter-sectional effects which further yielded differences in detection accuracies. Surprisingly, common fairness metrics such as PPV and TPR did not necessarily revealed these differences. This is most likely, because we chose the default detection threshold of the Darknet framework for the evaluation which lies at 25%. Apparently, even low confidence detection were often correct. However, because the chance level for two classes is not high we can't assume that these low confidence statements are legitimate guesses. However, the AP measure accounted for this insufficiency as it summarizes the precision-recall curve for varying thresholds. This indeed revealed some interesting results. Hence, for the object detection task it is important to look at more fairness criteria than typically discussed. This type of metric may help evaluating fairness for object detection situations in which training images are more realistic and less pristine such as surveillance camera footage. A fair object detection model should be fair regardless of the confidence threshold. The results show more uncertain decisions for the Black subgroup. This inability to detect to detect persons from the Black community has been known since Buolamwini and Gebru influential paper on this matter[6]. Recently, researchers such as Yee, Tantipongpipat, and Mishra and Hazirbas et al. dealt with some common downstream tasks such as facial cropping or gender classification. Our research results align with most results in this research field [24, 15]. However, the intersectional perspective also revealed unexpected results. Mainly, that White Males appear to perform second to last. It is not entirely clear why this is the case. This effect may be explained by any of the limitations of this paper.

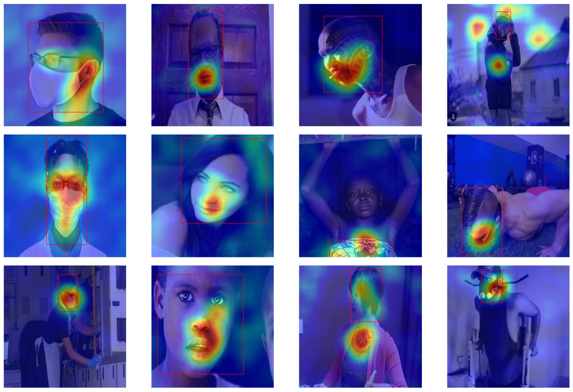

Figure 4: D-RISE saliency maps. The columns from left to right contain respectively: correctly detected faces with masks, correctly detected faces without masks, wrongly detected faces with masks. The final column contains wrongly detected face without mask and faces with masks detected correctly albeit with low confidence

4.1 Limitations

In this section, we will now address the limitations of our research. First, as previously mentioned, the labelling of the dataset was based on attributes perceived by the authors. For future research, a better result could be achieved through the help of an expert on the topic or by gathering data from the subjects of the pictures themselves. Second, due to the variety of sources and pictures in our dataset, the average difficulty of recognizing a particular subject might be very different among different groups, especially considering the size of the bounding box compared to the overall picture. This should therefore be considered a confounding factor that we were unable to address within our research. Third, while the resolution of the images themselves was partly addressed through our experiments, the impact of oversized images within our model's performance remains unknown. We were unable to address whether these outliers had a largely positive, negative, or average impact on the performance of our model and we, therefore, decided not to down-scale them to avoid insufficiently justified dataset pre-processing. Future work could involve further testing to compute the impact of these images, and if necessary address it by whatever means possible.

4.2 Ethical Considerations

The most important ethical consideration a researcher has to make concerning this work but also Fairness of computer vision models as a whole is the racialization of people. Race as a concept is not a biological term but rather a social construct and always a perceived attribute of ones character. Hence, labelling a dataset of individuals without awareness of this is deeply problematic. A researcher should always either prefer a self-statement of the individual or a more biologically grounded categorization like ethnicity if possible. Similar ideas apply to Gender. Since no reliable information about race and gender was provided, in the course of this project we had to base our assumptions on subtle, often culturally-motivated features of people. Despite the fact that such labelling is necessary for comparing the performance and making models more fair, its nature remain somewhat unethical, for this is the viewpoint we aim to avoid in the end. So, adding sensitive labels without reliable source is undesirable. However, when such features as gender, name, and age are available from the data, another serious ethical concern arises, namely the problem of anonymity. This is a trade-off that needs to be considered.

5 CONCLUSION

With this research we were able to show that biases can remain in data even if fine-tuned on data that was priorly debiased. This result is important, because it shows why it is important to take fairness early on into account. These fairness considerations are often not carefully thought out and may therefore lead to involuntary discriminatory systems or other falacies that plague the fairness research community. Object detection is often seen as a special version of a classification task. However, it has considerable differences such as the lack of true negatives or confidence thresholds. Hence, we showed that TPR and PPV alone cannot capture the idiosyncracies of object detection models, when dealing with fairness. The work also highlights the need for more fairness considerations within the object detection literature and more broadly transfer learning as a whole.

REFERENCES

[1] Alexey. *AlexeyAB/Darknet*. July 3, 2021. URL: https://github.com/AlexeyAB/darknet (visited on 07/03/2021).

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. "Fairness and Machine Learning". In: (), p. 253.

[3] Solon Barocas and Andrew D. Selbst. "Big Data's Disparate Impact Essay". In: *California Law Review* 104.3 (2016), pp. 671–732. URL: https://heinonline.org/HOL/P?h=hein.journals/calr104&i=695 (visited on 06/29/2021).

[4] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. *YOLOv4: Optimal Speed and Accuracy of Object Detection*. Version 1. Apr. 22, 2020. arXiv: 2004.10934 [cs, eess]. URL: http://arxiv.org/abs/2004.10934 (visited on 06/16/2021).

[5] Vanessa Buhrmester, David Münch, and Michael Arens. "Analysis of Explainers of Black Box Deep Neural Networks for Computer Vision: A Survey". In: *arXiv:1911.12116 [cs]* (Nov. 27, 2019). arXiv: 1911.12116. URL: http://arxiv.org/abs/1911.12116 (visited on 07/03/2021).

[6] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Ed. by Sorelle A. Friedler and Christo Wilson. Vol. 81. Proceedings of Machine Learning Research. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 77–91. URL: http://proceedings.mlr.press/v81/buolamwini18a.html.

[7] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: (), p. 15.

[8] US Census Bureau. *About Race*. The United States Census Bureau. Section: Government. URL: https://www.census.gov/topics/population/race/about.html (visited on 07/03/2021).

[9] Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *arXiv:1703.00056 [cs, stat]* (Feb. 28, 2017). arXiv: 1703.00056. URL: http://arxiv.org/abs/1703.00056 (visited on 07/05/2021).

[10] Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women". In: *Reuters* (Oct. 10, 2018). URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (visited on 07/05/2021).

[11] Terrance DeVries, Ishan Misra, and Changhan Wang. "Does Object Recognition Work for Everyone?" In: (), p. 8.

[12] Finale Doshi-Velez and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning". In: *arXiv:1702.08608 [cs, stat]* (Mar. 2, 2017). arXiv: 1702.08608. URL: http://arxiv.org/abs/1702.08608 (visited on 07/05/2021).

[13] Share on Facebook, Share on Twitter, and Share on LinkedIn. *Researchers Show That Computer Vision Algorithms Pretrained on ImageNet Exhibit Multiple, Distressing Biases*. VentureBeat. Nov. 3, 2020. URL: https://venturebeat.com/2020/11/03/researchers-show-that-computer-vision-algorithms-pretrained-on-imagenet-exhibit-multiple-distressing-biases/ (visited on 05/27/2021).

[14] Share on Facebook, Share on Twitter, and Share on LinkedIn. *NIST benchmarks show facial recognition technology still struggles to identify Black faces*. VentureBeat. Sept. 9, 2020. URL: https://venturebeat.com/2020/09/09/nist-benchmarks-show-facial-recognition-technology-still-struggles-to-identify-black-faces/ (visited on 07/01/2021).

[15] Caner Hazirbas et al. *Towards Measuring Fairness in AI: The Casual Conversations Dataset*. Apr. 6, 2021. arXiv: 2104.02821 [cs]. URL: http://arxiv.org/abs/2104.02821 (visited on 07/02/2021).

[16] Kenneth Holstein et al. "Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?" In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, May 2, 2019, pp. 1–16. ISBN: 978-1-4503-5970-2. URL: https://doi.org/10.1145/3290605.3300830 (visited on 06/29/2021).

[17] Zaid Khan and Yun Fu. "One Label, One Billion Faces: Usage and Consistency of Racial Categories in Computer Vision". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Mar. 3, 2021), pp. 587–597. DOI: 10.1145/3442188.3445920. arXiv: 2102.02320. URL: http://arxiv.org/abs/2102.02320 (visited on 07/03/2021).

[18] Larxel. *Face Mask Detection*. https://www.kaggle.com/. 2020. URL: https://kaggle.com/andrewmvd/face-mask-detection (visited on 07/03/2021).

[19] Jeff Larson Mattu Julia Angwin. *Machine Bias*. ProPublica. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (visited on 07/05/2021).

[20] Vitali Petsiuk et al. "Black-box Explanation of Object Detectors via Saliency Maps". In: *arXiv:2006.03204 [cs]* (June 10, 2021). arXiv: 2006.03204. URL: http://arxiv.org/abs/2006.03204 (visited on 07/04/2021).

[21] Joseph Redmon et al. "You Only Look Once: Unified, Real-Time Object Detection". In: *arXiv:1506.02640 [cs]* (May 9, 2016). arXiv: 1506.02640. URL: http://arxiv.org/abs/1506.02640 (visited on 07/02/2021).

[22] Ryan Steed and Aylin Caliskan. "Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 3, 2021, pp. 701–713. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445932. arXiv: 2010.15052. URL: https://doi.org/10.1145/3442188.3445932 (visited on 05/10/2021).

[23] Wobot Intelligence. *Face Mask Detection Dataset*. https://www.kaggle.com/. 2020. URL: https://kaggle.com/wobotintelligence/face-mask-detection-dataset (visited on 07/03/2021).

[24] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. *Image Cropping on Twitter: Fairness Metrics, Their Limitations, and the Importance of Representation, Design, and Agency*. May 18, 2021. arXiv: 2105.08667 [cs]. URL: http://arxiv.org/abs/2105.08667 (visited on 07/05/2021).