Bias behind the Wheel: Fairness Testing of Autonomous Driving Systems

XINYUE LI, Peking University, Beijing, China
ZHENPENG CHEN, Nanyang Technological University, Singapore, Singapore
JIE M. ZHANG, King's College London, London, United Kingdom
FEDERICA SARRO, University College London, London, United Kingdom
YING ZHANG and XUANZHE LIU, Peking University, Beijing, China

This article conducts fairness testing of automated pedestrian detection, a crucial but under-explored issue in autonomous driving systems. We evaluate eight state-of-the-art deep learning-based pedestrian detectors across demographic groups on large-scale real-world datasets. To enable thorough fairness testing, we provide extensive annotations for the datasets, resulting in 8,311 images with 16,070 gender labels, 20,115 age labels, and 3,513 skin tone labels. Our findings reveal significant fairness issues, particularly related to age. The proportion of undetected children is 20.14% higher compared to adults. Furthermore, we explore how various driving scenarios affect the fairness of pedestrian detectors. We find that pedestrian detectors demonstrate significant gender biases during night time, potentially exacerbating the prevalent societal issue of female safety concerns during nighttime out. Moreover, we observe that pedestrian detectors can demonstrate both enhanced fairness and superior performance under specific driving conditions, which challenges the fairness-performance tradeoff theory widely acknowledged in the fairness literature. We publicly release the code, data, and results to support future research on fairness in autonomous driving.

CCS Concepts: • **Software and its engineering → Software creation and management**; • **Computing methodologies → Machine learning**;

Additional Key Words and Phrases: Fairness testing, pedestrian detection, autonomous driving

ACM Reference format:
Xinyue Li, Zhenpeng Chen, Jie M. Zhang, Federica Sarro, Ying Zhang, and Xuanzhe Liu. 2025. Bias behind the Wheel: Fairness Testing of Autonomous Driving Systems. *ACM Trans. Softw. Eng. Methodol.* 34, 3, Article 82 (March 2025), 24 pages.
https://doi.org/10.1145/3702989

Xinyue Li, Ying Zhang, and Xuanzhe Liu are also affiliated with the Key Lab of High Confidence Software Technologies (Peking University), Ministry of Education, Beijing, China.
This work was supported by the National Natural Science Foundation of China under the grant number 62325201, as well as by the Center for Data Space Technology and Systems at Peking University.
Authors' Contact Information: Xinyue Li, Peking University, Beijing, China; e-mail: xinyueli@stu.pku.edu.cn; Zhenpeng Chen (corresponding author), Nanyang Technological University, Singapore, Singapore; e-mail: zhenpeng.chen@ntu.edu.sg; Jie M. Zhang, King's College London, London, United Kingdom; e-mail: jie.zhang@kcl.ac.uk; Federica Sarro, University College London, London, United Kingdom; e-mail: f.sarro@ucl.ac.uk; Ying Zhang, Peking University, Beijing, China; e-mail: ying.zhang@pku.edu.cn; Xuanzhe Liu, Peking University, Beijing, China; e-mail: liuxuanzhe@pku.edu.cn.

ACM 1557-7392/2025/3-ART82
https://doi.org/10.1145/3702989

1 Introduction

Autonomous driving systems are on track to become the predominant mode of transportation in the future [3]. However, these systems are susceptible to software bugs [37], which can potentially result in severe injuries or even fatalities for both pedestrians and passengers. The unfortunate incident in 2018 involving an autonomous vehicle from Uber serves as a stark reminder of these risks [2]. Given the safety-critical nature of autonomous driving systems, they have garnered substantial attention from the software testing community [74].

Extensive research efforts have been devoted to the testing of autonomous driving systems. For example, Tian et al. [63] introduced DeepTest, which applies image transformation to simulate potential camera noise in autonomous driving scenarios. Zhang et al. [77] developed DeepRoad, a **Generative Adversarial Network (GAN)**-based approach that generates test images from real-world driving scenes. Zhou et al. [81] proposed DeepBillboard, a system for generating adversarial billboards to induce potential steering errors in autonomous vehicles.

Although significant testing efforts have been made, to the best of our knowledge, the study of fairness testing for autonomous driving systems remains under investigation in the literature. From the **Software Engineering (SE)** perspective, fairness is considered a non-functional software property, making it an important subject for testing [62, 74]. Fairness testing, as an emerging domain within software testing, seeks to uncover fairness issues in software systems [26].

Fairness issues in autonomous driving systems, such as a higher accuracy in detecting male pedestrians compared to females, can perpetuate discriminatory outcomes and unequal treatment based on gender. This can result in harm to individuals belonging to marginalized groups, further exacerbating existing social inequalities. Therefore, it is crucial to prioritize fairness testing in autonomous driving systems.

To fill the gap, we conduct fairness testing of eight state-of-the-art **Deep Learning (DL)**-based pedestrian detectors that have been extensively studied in the research community. Our main focus is to quantitatively assess performance disparities in these detectors across diverse demographic groups, which are widely recognized as group fairness issues [27]. To enable fairness testing, we manually enrich four widely adopted real-world datasets with gender, age, and skin tone labels, resulting in a collection of 8,311 real-world images annotated with 16,070 gender labels, 20,115 age labels, and 3,513 skin tone labels. Using these labeled datasets, we assess the group fairness of existing pedestrian detectors and also explore how commonly studied driving scenarios (including various brightness, contrast, and weather conditions) impact the fairness of these detectors.

Our study reveals the following findings: *(1)* Overall, state-of-the-art pedestrian detectors exhibit significant bias regarding age. On the four datasets examined, the undetected proportions for children surpass those for adults by an average of 20.14%. However, the overall performance of these pedestrian detectors in detecting males and females and dark-skin and light-skin groups does not exhibit a large difference, with only a 1.19% and 0.44% gap in undetected proportions. *(2)* The studied pedestrian detectors reveal significant gender biases during night time, with a higher proportion of females going undetected compared to males. This situation may aggravate existing societal concerns about female safety during nighttime outings. *(3)* In contrast to the commonly accepted fairness-performance tradeoff, our findings suggest that pedestrian detectors can achieve enhanced fairness and detection performance under specific driving scenarios, such as those with higher brightness levels.

To summarize, we make the following contributions:

—We conduct the first comprehensive study on fairness testing of autonomous driving systems across various datasets and demographic attributes, evaluating eight widely studied DL-based pedestrian detectors and uncovering significant fairness issues.

—We augment four real-world datasets with manually labeled demographic information, resulting in 8,311 images with 16,070 gender labels, 20,115 age labels, and 3,513 skin tone labels.
—We publicly release the data, demographic labels, and code used in this study [9] to facilitate future research on fairness of autonomous driving systems.

2 Background and Related Work

This study resides at the intersection of two increasingly important SE topics: software fairness and autonomous driving testing. To provide the necessary context, we begin by reviewing the background knowledge and relevant prior research in these areas.

2.1 Software Fairness

Fairness has gained significant attention in the SE community since its initial exploration by SE researchers in 2008 [36]. There have been various definitions of fairness in the literature. In this article, we focus on group fairness, a concept extensively studied in software fairness research [15, 16, 22, 23, 29, 40, 44, 52, 73, 75, 76]. Notably, group fairness closely aligns with legal regulations on fairness [11], such as the adherence to the four-fifths rule, a cornerstone of US anti-discrimination law [13, 34, 35, 69, 72]. Consequently, testing and prioritizing group fairness when building software has emerged as an essential ethical duty and requirement for software engineers [22].

In the context of group fairness, certain personal characteristics that require protection against unfairness during decision-making are called sensitive attributes, also known as protected attributes [31,43, 74]. Well-recognized sensitive attributes include race, sex, age, pregnancy, familial status, disability status, and more [26, 65]. These sensitive attributes typically partition a population into distinct groups: a privileged group and an unprivileged group [26]. Group fairness entails the equal treatment of these groups by the same **Machine Learning (ML)** model. However, in practice, members of unprivileged groups often experience systematic disadvantages, resulting from unfair ML models. For instance, in the context of a pedestrian detection task, if age is deemed a sensitive attribute, the predictive model may exhibit a bias favoring the adult group over the child group. In this scenario, the adult group is considered the privileged group, while the child group becomes the unprivileged one.

Recently, Chen et al. [26] have presented a comprehensive survey of fairness testing research and analyzed its trend. This survey points out that the majority of existing work revolves around tabular data [15, 16, 28, 40, 73]. For example, Biswas and Rajan [15] evaluated fairness of ML models on a crowdsourced platform using tabular datasets covering tasks such as credit risk prediction, income prediction, marketing, and loan application. Similarly, Chen et al. [28] conducted an empirical study on the group fairness achieved by state-of-the-art bias mitigation methods across eight commonly used tabular data-driven decision tasks. In contrast, our article centers on fairness testing for pedestrian detection in autonomous driving systems. We specifically examine three sensitive attributes (i.e., gender, age, and skin tone) that are recognizable in autonomous driving datasets. These sensitive attributes have been demonstrated to be the most widely considered ones in the fairness testing literature [26].

2.2 Autonomous Driving Testing

Autonomous driving testing is a hot SE research topic, and researchers have proposed various testing techniques for autonomous driving systems [41, 55, 63, 77, 81, 83]. For instance, Tian et al. [63] proposed DeepTest, a novel technique using image transformations to emulate potential camera disturbances encountered in driving environments. Zhang et al. [77] presented DeepRoad, employing GANs to craft test images derived from actual driving scenarios. Zhou et al. [81]

introduced DeepBillboard, aiming to generate adversarial billboards that could lead to steering mistakes in autonomous vehicles. Guo et al. [41] developed LiRTest, marking the first automated testing technique for LiDAR-equipped autonomous driving software.

While a substantial body of knowledge focuses on assessing the robustness and correctness properties of autonomous driving systems [74], to the best of our knowledge, only a few studies have explored the fairness properties, particularly in the pedestrian detection domain within autonomous driving, indicating that this area remains under-explored.

Pedestrian detection is a crucial process that identifies pedestrians within street-level images by providing their predicted locations along with corresponding bounding boxes and confidence scores [33, 78]. Despite its significance, research on the fairness of pedestrian detection within autonomous driving is limited. Brandao [17] explored fairness in pedestrian detection, concentrating on classic ML-based methods. These classic techniques, reliant on manually defined features, have been eclipsed by DL-based pedestrian detection approaches, now prevalent in the autonomous driving domain. Wilson et al. [70] focused on skin tone bias, confined their analysis to a single dataset and two general object detectors. Similarly, Kogure et al. [46] explored age bias using a small-scale dataset and a detection method that is no longer state-of-the-art.

In summary, current fairness studies in pedestrian detection suffer from a lack of variety in the pedestrian detectors evaluated, the datasets used, and the range of sensitive attributes explored. Furthermore, no previous work has explored how different environmental characteristics (e.g., brightness, contrast, and weather conditions) affect fairness.

To address this knowledge gap, our article presents a comprehensive empirical study on revealing fairness issues in pedestrian detection. We conduct experiments using eight popular DL-based detection methods and four diverse testing datasets, encompassing different scenarios determined by a variety of factors such as brightness, contrast, and weather conditions. We focus on three widely considered sensitive attributes, i.e., gender, age, and skin tone. The scale and diversity of our experiments enable us to provide comprehensive insights into the fairness circumstances of the existing pedestrian detectors.

3 Experimental Design

This section introduces our **Research Questions (RQs)** and experimental design.

3.1 RQs

We aim to answer the following RQs.

RQ1 (Overall Fairness): To what extent do widely studied pedestrian detectors exhibit unfairness concerning common sensitive attributes? This RQ explores the performance difference of widely studied pedestrian detectors when they are applied to different demographic groups, characterized by sensitive attributes including gender, age, and skin tone.

RQ2 (Fairness in Different Scenarios): What fairness do pedestrian detectors achieve in different brightness, contrast, and weather conditions? We further investigate the fairness of commonly studied pedestrian detectors in different autonomous driving scenarios by considering a variety of brightness (RQ2.1), contrast (RQ2.2), and weather conditions (RQ2.3) [63, 77, 81].

Figure 1 illustrates our experimental settings to answer these RQs. In the following, we introduce the pedestrian detectors, benchmark datasets, evaluation metric, statistical analysis, and experimental details.

3.2 Pedestrian Detectors

In recent years, DL has revolutionized pedestrian detection approaches. We focus our analysis on eight DL-based pedestrian detectors that are widely studied in the autonomous driving community

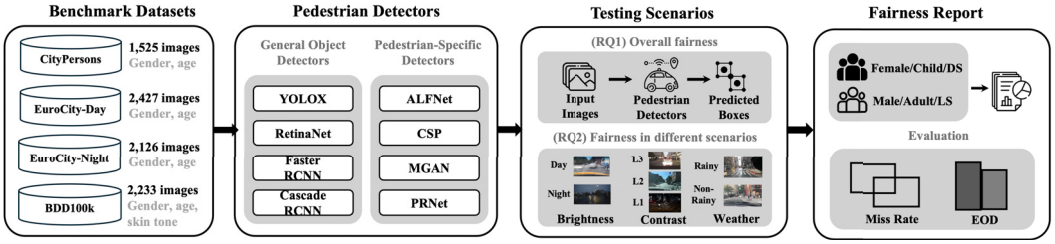

Fig. 1. Overview of our experimental settings.

Table 1. Pedestrian Detectors

Type	Detector	Backbone	Source
General object detectors	YOLOX	-	MMDetection [10]
	RetinaNet	X-101-64x4d-FPN	
	Faster RCNN	X-101-64x4d-FPN	
	Cascade RCNN	X-101-64x4d-FPN	
Pedestrian-specific detectors	ALFNet	ResNet50	ALFNet [1]
	CSP	ResNet50	Pedestron [4]
	MGAN	VGG16	
	PRNet	ResNet50	PRNet [3]

[20, 82]. These detectors are pre-trained DL models that researchers and practitioners can directly use for pedestrian detection tasks. They can be classified into two categories: general object detectors and pedestrian-specific detectors [42]. Next, we briefly introduce each category and the corresponding pedestrian detectors.

Table 1 provides an overview of these detectors. "Detector" shows the name of a pedestrian detector; "Backbone" represents the pre-trained deep neural network used for extracting features from input images; "Source" indicates the framework/toolkit name and its source for a given pedestrian detector.

General Object Detectors. General detectors can detect various objects such as cars, traffic lights, and pedestrians. They have great generalization ability but lack pedestrian-specific adaptation [42]. They can be categorized into two categories: two-stage and one-stage detectors [82]. Two-stage detectors propose regions before feature extraction and classification, achieving high accuracy but slower speed; one-stage detectors complete all operations in one step, providing faster speed but lower accuracy. Both types involve tradeoffs and are widely used for pedestrian detection. Hence, this article selects detectors from both categories. For one-stage detectors, we adopt the widely studied *YOLOX* [39] (a faster extension of the YOLO series [59] used in Apollo's autonomous driving systems [5]) and *RetinaNet* [48] (which addresses the class imbalance problem). For two-stage detectors, we employ the *Faster RCNN* [60] (one of the pioneering detectors in the RCNN family) and *Cascade RCNN* [19] (which achieves higher accuracy through a cascade of multiple CNNs to refine region proposals). These detectors have been extensively used in the autonomous driving literature [18, 20, 42, 71].

Pedestrian-Specific Detectors. Pedestrian-specific detectors use additional pedestrian-related information to improve detection performance [42]. In this study, we investigate state-of-the-art pedestrian-specific detectors, including ALFNet [50], CSP [1], MGAN [54], and PRNet [61]. *ALFNet*

Table 2. Benchmark Datasets

Name	Sensitive Attributes	#Images	Time
CityPersons	Gender, age	1,525	Day
EuroCity-Day	Gender, age	2,427	Day
EuroCity-Night	Gender, age	2,126	Night
BDD100k	Gender, age, skin tone	2,233	Day, night

uses progressive detection heads on SSD [49] to refine initial anchors for improved detection accuracy. *CSP* introduces an anchor-free approach by locating center points and scaling pedestrians. *MGAN* uses visible-area bounding-box information to guide attention mask generation for occluded pedestrian detection. *PRNet* presents a novel progressive refinement network for occluded pedestrian detection.

3.3 Benchmark Datasets

3.3.1 Dataset Selection. We perform our experiments on four real-world datasets that have been extensively studied by researchers to evaluate the performance of pedestrian detectors in autonomous driving [20, 42]. These datasets consist of street-level images captured by cameras mounted on autonomous vehicles, showcasing pedestrians in diverse poses, sizes, and occlusion scenarios. Table 2 presents details about these datasets, including the sensitive attributes, the number of images in each dataset, and the respective time of day when these images were captured. Next, we briefly introduce each dataset:

— *CityPersons* dataset [79] stands as the most widely studied benchmark for evaluating pedestrian detectors [1, 20, 42, 50, 54, 61]. Its test set includes 1,525 images captured across six cities, showcasing diverse weather conditions and street scenes.
— *EuroCityPersons* dataset [18] contains 4,553 images gathered from seven European cities, encompassing both day and night time captures. The dataset can be categorized into two sets: 2,427 images captured during the day, and 2,126 images captured at night, referred to as the *EuroCity-Day* dataset and the *EuroCity-Night* dataset, respectively.
— *Berkeley Driving* dataset (a.k.a., *BDD 100k* dataset) [71] is an extensive driving dataset, including 2,233 images from 40 classes that are typical of driving scenes. These images were captured from four different cities, providing various times of the day. Notably, this dataset showcases a greater diversity of pedestrians than the CityPersons and EuroCityPersons datasets, including individuals with varied skin tones.

3.3.2 Sensitive Attribute Labeling. We focus on three sensitive attributes: gender, age, and skin tone. These attributes are identifiable in autonomous driving images and are recognized as the three most extensively studied sensitive attributes in fairness testing literature [26]. To enable fairness analysis, we need datasets with labels that indicate these sensitive attributes of the humans depicted in the images. Among the datasets investigated herein, the only sensitive attribute already labeled is the skin tone (i.e., light-skin tone and dark-skin tone) for the BDD100k dataset.

We manually label gender and age for each of the datasets considered in this study. For skin tone, we only use the BDD100k dataset. As described in Section 3.3.1, this dataset shows a greater diversity in skin tones compared to other datasets. This diversity is linked to the geographic locations where the datasets were collected. Specifically, BDD100k, collected in the United States, exhibits more varied skin tones. In contrast, other datasets from European cities, such as CityPersons collected in Germany, have significantly fewer dark-skin individuals. As reported [6–8], there is only around 1%

Table 3. Cohen's κ Values for Labeling Gender and Age

Dataset	κ_{gender}	κ_{age}
CityPersons	0.814	0.847
EuroCity-Day	0.800	0.925
EuroCity-Night	0.870	0.847
BDD100k	0.854	0.828

Table 4. Cohen's κ Values and Corresponding Agreement Levels of Inter-Rater Agreement

κ Values	Agreement Level
<0	No agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1	Almost perfect agreement

representation of individuals with dark skin in Germany. During our labeling process, we indeed encountered scarcity of dark-skin individuals in the CityPersons, EuroCity-Day, and EuroCity-Night datasets, leading to our decision not to label these three datasets with skin tone attributes.

The labeling process involves two annotators to minimize the influence of labeling bias. We focus on images that align with the widely adopted "reasonable subset" principle [33], meaning that we label images containing labeled pedestrians with a height of at least 50 pixels and little to no occlusion. For such images, human annotators can label the sensitive attributes with high confidence [17]. Using the filtered datasets, the two annotators independently label the gender and age attributes for each image. For gender, we follow previous studies [17] and consider only two labels: male and female. As for age, in line with the literature [17, 46], we classify pedestrians into two labels: child and adult, based on their physical characteristics depicted in the images.

To ensure the reliability of the labeling procedure, both annotators independently label the gender and age attributes for each image. We use Cohen's kappa (κ) [30], a widely adopted metric for measuring inter-rater agreement, during the independent labeling process [10, 25, 51, 66, 68, 80]. The obtained κ values for gender and age attributes in each of the four datasets are summarized in Table 3. According to the literature [47], a κ value between 0.81 and 1 signifies almost perfect agreement, as shown in Table 4, while a value between 0.61 and 0.8 indicates substantial agreement. In our labeling process, we achieve substantial agreement in gender labeling for the EuroCity-Day dataset and almost perfect agreement for all other tasks. This high level of agreement underscores the reliability of our labeling procedure [10, 25, 51, 66, 68, 80]. In cases where the two annotators encounter conflicts, an arbitrator is involved in the discussion to reach a consensus. After the labeling process, the summary of the number of labeled pedestrian instances for each dataset is presented in Table 5.

3.3.3 Scenario Labeling. To deeply explore the fairness of pedestrian detectors across various driving scenarios (i.e., different brightness, contrast, and weather conditions), we also need images labeled with scenario information. We therefore classify the images containing labeled pedestrians

Table 5. Number of Labeled Pedestrian Instances per Dataset

	Gender		Age		Skin Tone	
Dataset	Male	Female	Adult	Child	Light-Skin	Dark-Skin
CityPersons	2,357	1,822	4,568	233	-	-
EuroCity-Day	1,726	1,646	4,498	100	-	-
EuroCity-Night	1,265	1,318	4,165	68	-	-
BDD100k	3,457	2,479	6,293	190	2,724	789
Overall	8,805	7,265	19,524	591	2,724	789

(5,933 out of the total 8,311 images in all datasets) into different scenarios, to enable our analysis in RQ2. An alternative approach to achieving this purpose is via generating images with different scenarios using existing test generation techniques from the autonomous driving literature. We do not choose this approach because generated images are not real images and can suffer from unnaturalness [74].

Labeling Brightness. Brightness represents the overall lightness or darkness of the image. To distinguish the brightness of the images, we use the time-of-day labels provided in the dataset annotations. Specifically, we categorize the images into "day time" and "night time." As shown in Table 2, the CityPersons and EuroCity-Day datasets consist entirely of day time images, while the EuroCity-Night dataset consists entirely of night time images. These datasets are straightforward to categorize based on time of day, though they do not provide more granular brightness labels. For the BDD100k dataset, we use the detailed "timeofday" labels provided by Wilson et al. [70], which include separate annotations for "day time" (covering dawn, dusk, and full daylight) and "night time." We apply these labels directly in our study to distinguish between brightness conditions.

Labeling Contrast. Contrast is the difference in brightness between objects in an image. To quantify the contrast of each image, we use the RMS contrast measurement [56], a standard measure in the computer vision literature [12, 45, 57]. To apply the RMS measurement, we first need to convert all images into the gray-scale mode [56]. Then, we calculate the RMS contrast value for each image based on the converted version. A higher RMS contrast value indicates a greater contrast. To classify the images into different contrast levels, we identify the maximum RMS contrast value (which is 89.45) and the minimum RMS contrast value (which is 11.42) among all images. Then, we evenly divide this RMS contrast values range into three classes (each level can have sufficient images for statistical analysis), each covering an interval of 26.01 units (calculated as (max-min)/3). Each class represents a contrast level, labeled from level 1 to level 3, with higher levels indicating images with higher contrast. Then, we categorize the images into their respective contrast levels based on their RMS contrast values.

Labeling Weather Conditions. Common weather conditions studied in the autonomous driving literature include rain, fog, and snow [55, 63, 77, 81]. However, our datasets rarely contain images depicting fog and snow. This is due to the fact that our datasets are collected from real-world scenarios where fog and snow are infrequently encountered. The limited samples of snowy or foggy weather pose challenges for statistical analysis. As a result, we focus on rain as the weather condition of interest. Two annotators independently classify images containing labeled pedestrians into two categories: rainy and non-rainy. During this process, 1,856 images are not annotated because neither the two annotators nor the arbitrator could accurately distinguish the weather conditions.

Table 6. Number of Images in Different Brightness Conditions, Contrast Levels, and Weather Conditions

Brightness Conditions		Contrast Levels			Weather Conditions	
Day Time	Night Time	Level 1	Level 2	Level 3	Rainy	Non-Rainy
4,409	1,524	1,163	3,933	837	277	3,800

To measure inter-rater agreement during manual labeling, we also use Cohen's kappa (κ). The κ value is 0.813, indicating almost perfect agreement [47]. This high level of agreement confirms the reliability of our labeling procedure. After scenario labeling, the summary of the number of images under different brightness, contrast, and weather conditions is presented in Table 6.

3.4 Evaluation Metric

There have been well-established quantitative measures for group fairness in the literature. The most widely adopted fairness measures include **Statistical Parity Difference (SPD)**, **Equal Opportunity Difference (EOD)**, and **Average Odds Difference (AOD)** [26, 27, 73]. Let a sensitive attribute be A, with 0 as the unprivileged group and 1 the privileged group; let the real classification label be Y and the predicted label $\hat{Y}$, with 0 as the unfavorable class and 1 as the favorable class. In addition, we use Pr to denote probability.

SPD quantifies the difference in the probabilities of favorable outcomes between unprivileged and privileged groups:

$$SPD = Pr[\hat{Y} = 1|A = 0] - Pr[\hat{Y} = 1|A = 1]. \tag{1}$$

EOD quantifies the difference in the **True-Positive (TP)** rates between unprivileged and privileged groups:

$$EOD = Pr[\hat{Y} = 1|A = 0, Y = 1] - Pr[\hat{Y} = 1|A = 1, Y = 1]. \tag{2}$$

AOD quantifies the average difference between the **False-Positive (FP)** rates and TP rates for unprivileged and privileged groups:

$$\begin{aligned} AOD = &\frac{1}{2}(|Pr[\hat{Y} = 1|A = 0, Y = 0] - Pr[\hat{Y} = 1|A = 1, Y = 0]| \\ &+ |Pr[\hat{Y} = 1|A = 0, Y = 1] - Pr[\hat{Y} = 1|A = 1, Y = 1]|). \end{aligned} \tag{3}$$

In the context of pedestrian detection, both SPD and EOD measure the disparity in proportions of successfully detected pedestrians between privileged and unprivileged groups. They also both express the difference in **Miss Rates (MRs)** between privileged and unprivileged groups. *MR* is the most commonly studied performance metric in pedestrian detection [17, 46], which quantifies the proportion of undetected pedestrians. Formally, it is calculated as follows:

$$MR = 1 - \frac{TP}{TP + FN}, \tag{4}$$

where *TP* refers to the number of successfully detected ground-truth bounding boxes, and *FN* denotes the number of undetected ground-truth bounding boxes. Pedestrian detectors generate bounding box locations and confidence scores for recognized "person" instances in images. To assess whether a given ground-truth bounding box is successfully detected, the standard method in the literature is to use the **Intersection over Union (IoU)** metric [33]. The IoU metric quantifies the degree of overlap between the ground-truth bounding box and the detected bounding box.

If the IoU value is greater than 50%, the ground-truth bounding box is considered successfully detected [33]. Otherwise, it is classified as undetected.

The calculation of AOD requires precise FP information, referring to instances where members of the negative class (non-pedestrians) are incorrectly classified as the positive class (pedestrians). As described in Section 3.3.2, we adhere to the standard practice in the literature, where we focus on a "reasonable subset" of pedestrians within the images. This approach presents challenges in calculating precise FPs for each group because the negative class (non-pedestrian) may also include instances that belong to the positive class (pedestrian). Therefore, we do not consider AOD in this study.

In summary, we use both SPD and EOD as fairness measures for evaluating pedestrian detectors. Since these two measures yield identical values for pedestrian detection, for the remainder of the article, we present only *EOD*.

3.5 Statistical Analysis

To assess the extent to which any observed unfairness is statistically significant (i.e., whether there is a significant difference in the MR between privileged and unprivileged groups), we use the two-proportion z-test [21]. This statistical test is widely used to analyze differences between proportions [53, 64]. A result is deemed significant only if the obtained p-value falls below a pre-determined threshold (in our case, 0.05, a widely accepted threshold in the fairness literature [27, 28]). For instance, in evaluating whether there exists a difference between the MRs for male (MR_{male}) and female (MR_{female}) individuals detected by a pedestrian detector, the null hypothesis assumes that MR_{male} is equal to MR_{female}. If the resulting p-value is lower than 0.05, we reject the null hypothesis, indicating a significant difference between MR_{male} and MR_{female}.

3.6 Experimental Details

The experiments are implemented based on open source frameworks of each pedestrian detector. For general object detectors, we select the pre-trained models with the highest accuracy from the MMDetection model zoo [10]. For pedestrian-specific detectors, we employ pre-trained models available from their respective public repositories [1, 3, 4].

To ensure the reliability of our results, all experiments are repeated 10 times. The final results are derived by calculating the average across these 10 iterations.

All experiments are performed on a platform equipped with 64 GB RAM, 2.5 GHz Intel Xeon (R) v3 Dual CPUs, and one NVIDIA GeForce RTX 2080 Ti GPU. YOLOX, RetinaNet, Faster RCNN, and Cascade RCNN are implemented using PyTorch 1.8.1 and Python 3.7 on Ubuntu 18.04 LTS, following the MMDetection configuration [24]. CSP and MGAN use PyTorch 1.10.0 and Python 3.8 on Ubuntu 18.04 LTS, adhering to the Pedestron configuration [4]. ALFNet [50] and PRNet [61] are implemented using Keras 2.0.6, Tensorflow 1.4.0, and Python 2.7 on Ubuntu 16.04 LTS.

4 Results

This section answers our RQs based on experimental results.

4.1 RQ1: Overall Fairness

RQ1 investigates the overall fairness of eight state-of-the-art pedestrian detectors regarding gender, age, and skin tone. First, for each detector, we compute the MR for different demographic groups and calculate EOD based on the MR results over all the datasets that we study. We also use the two-proportion z-test to determine the significance of any observed unfairness, as described in Section 3.5. Table 7 presents the results, with significant unfairness results highlighted in shading. In the following, we analyze the results for gender, age, and skin tone, respectively.

Table 7. (RQ1) Overall Fairness in Pedestrian Detection across Gender, Age, and Skin Tone

Detectors	MR Male	MR Female	EOD (Gender)	MR Adult	MR Child	EOD (Age)	MR Light-Skin	MR Dark-Skin	EOD (Skin)
YOLOX	9.78%	10.50%	−0.72%	12.64%	42.47%	−29.83%	5.21%	3.80%	1.41%
RetinaNet	10.72%	12.59%	−1.87%	14.36%	44.33%	−29.97%	8.33%	4.44%	3.90%
Faster RCNN	3.80%	4.13%	−0.32%	5.24%	26.06%	−20.82%	5.91%	3.30%	2.62%
Cascade RCNN	3.87%	4.16%	−0.28%	5.12%	26.57%	−21.44%	5.21%	3.04%	2.17%
ALFNet	30.86%	32.90%	−2.04%	36.62%	53.47%	−16.85%	42.55%	43.98%	−1.43%
CSP	33.49%	35.25%	−1.76%	37.74%	50.42%	−12.68%	61.38%	64.77%	−3.39%
MGAN	29.76%	30.97%	−1.21%	33.58%	46.53%	−12.95%	52.94%	54.88%	−1.94%
PRNet	40.28%	41.62%	−1.34%	44.69%	61.25%	−16.56%	59.65%	59.44%	0.21%
Average	20.32%	21.52%	−1.19%	23.75%	43.89%	−20.14%	30.15%	29.71%	0.44%

Statistically significant biases, indicated by EOD (i.e., MR difference), are shaded. On average, detectors display comparable MRs for female and male pedestrians, as well as for dark-skin and light-skin individuals. However, concerning age, significant bias is observed, as pedestrian detectors exhibit a 20.14% higher MR for children compared to adults.

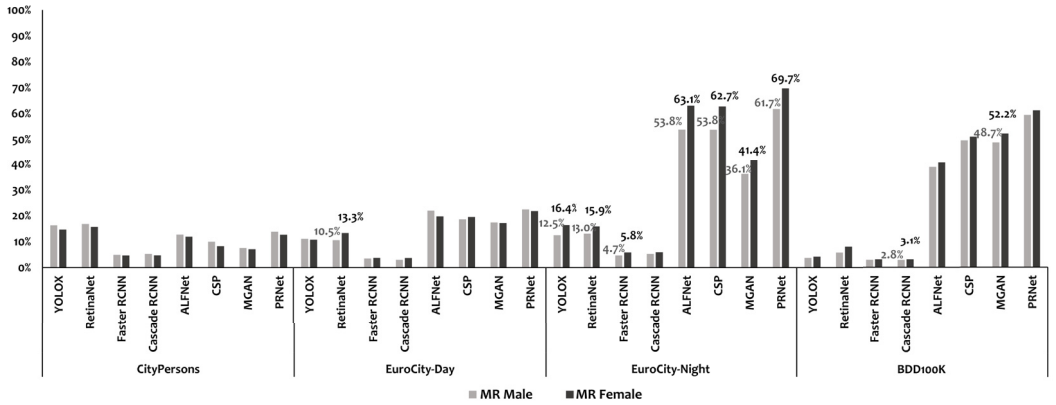

Fig. 2. (RQ1) MRs of pedestrian detectors for females and males across datasets. Statistically significant gender biases are indicated by labeled MR values. In CityPersons and EuroCity-Day datasets with only day time data, only one detector in the EuroCity-Day dataset exhibits significant gender bias. However, in the EuroCity-Night dataset, seven out of eight detectors show significantly higher MRs for females, revealing bias in female detection.

Gender. As shown in Table 7, on average, the MR difference between female and male pedestrians is merely 1.19% (p-value > 0.05), indicating that this difference is not statistically significant.

Furthermore, we analyze the MR difference achieved by each pedestrian detector across the four datasets used in our study. Figure 2 illustrates the results. Significant gender biases (i.e., significant MR differences) are indicated by labeled MR values. We observe that in the CityPersons and EuroCity-Day datasets, containing only day time data, only one detector in the EuroCity-Day dataset exhibits significant differences in MRs between females and males. For the remaining results, there are no notable differences in MRs between genders. However, results from the EuroCity-Night dataset present a contrasting observation, where seven of eight detectors exhibit a significantly higher MR for females, indicating bias in detecting females. In the BDD100k dataset that includes both day time and night time images, the MR difference is less pronounced. These observations motivate us to hypothesize that brightness conditions may influence the fairness of pedestrian detectors, which is further investigated in RQ2.

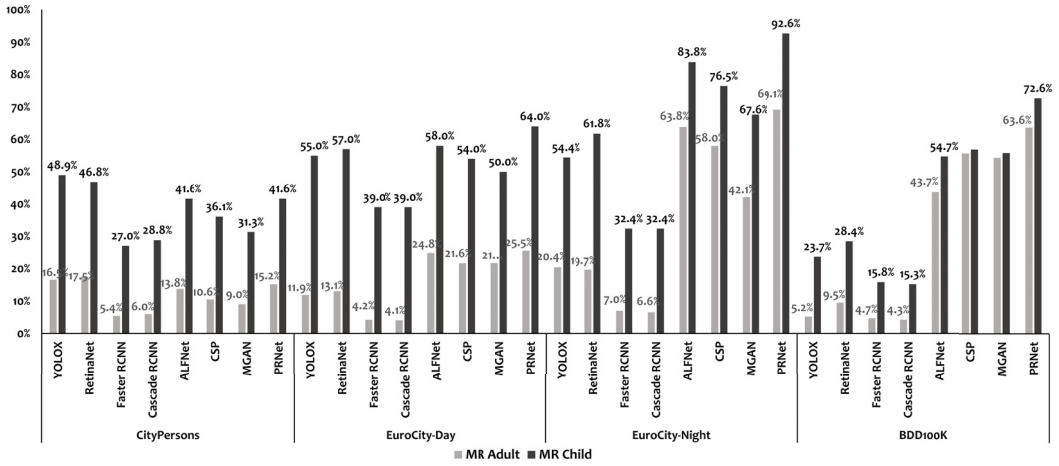

Fig. 3. (RQ1) MRs of pedestrian detectors for children and adults across datasets. Statistically significant age biases are labeled with MR values. In 30 out of 32 scenarios (comprising 4 datasets and 8 detectors), children have significantly higher MRs than adults.

Age. As observed in Table 7, pedestrian detection exhibits large age bias, with all studied detectors demonstrating significantly higher MRs for children compared to adults (p-value < 0.05). On average, the MR difference between children and adults is 20.14%.

Furthermore, we illustrate the MR difference across four datasets in Figure 3, with significant age biases indicated by labeled MR values. We observe that the MR of children is consistently higher than that of adults across all the datasets and all detectors. In particular, out of the total 32 results (combinations of 4 datasets and 8 pedestrian detection models), 30 exhibit a statistically significant MR difference (p-value < 0.05). This indicates that the bias favoring adults is not specific to a particular dataset or detector, highlighting a strong unfairness between adults and children.

The age bias may be attributed to the inherent challenge of detecting small objects, owing to the limited information provided by small bounding boxes [20, 32, 46, 78]. Given that children generally have smaller bodies compared to adults, their bounding boxes in the images also tend to be smaller. To demonstrate this, we analyze the distribution of bounding box sizes for pedestrians detected and undetected by all pedestrian detectors, as well as the distribution of ground-truth bounding box sizes for both adults and children. The results, presented in Figure 4, reveal a correspondence between these distributions. Specifically, the undetected bounding boxes and the ground-truth bounding boxes for children do not exceed 400 pixels in height and 200 pixels in width. This analysis shows that children, as well as undetected pedestrians, tend to have smaller bounding boxes.

Skin Tone. As displayed in Table 7, current pedestrian detectors exhibit minimal bias between light-skin and dark-skin individuals, with an average MR difference of just 0.44%. Specifically, among the four pedestrian-specific detectors, none show significant skin tone bias, whereas among the four general object detectors, three display significant skin tone bias, with MR differences ranging from 2.17% to 3.9%. These disparities are notably smaller than those observed in age bias.

> *Answer to RQ1*: On average, the eight state-of-the-art pedestrian detectors that we study exhibit no significant performance difference across gender and skin tone, but notable bias across age. In particular, the detectors show a 20.14% higher MR for children compared to adults.

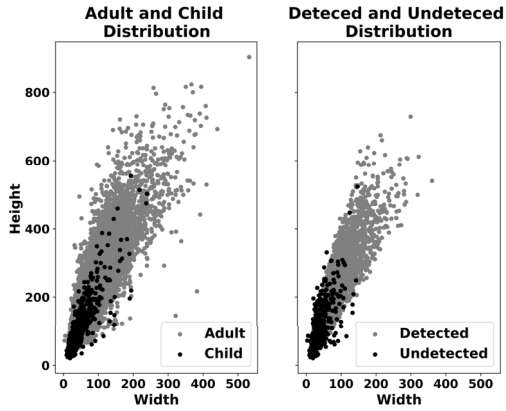

Fig. 4. (RQ1) Bounding box size distributions of adults and children (left) and bounding box size distributions of undetected and detected pedestrians (right). We observe that both children and undetected pedestrians tend to have smaller bounding boxes.

4.2 RQ2: Fairness in Different Scenarios

RQ2 evaluates the fairness of state-of-the-art pedestrian detectors under different real-world autonomous driving scenarios.

4.2.1 RQ2.1: Different Brightness Conditions. As introduced in Section 3.3.3, we consider two brightness conditions: day time and night time.

We first evaluate the overall MRs of the eight pedestrian detectors under day time and night time. The results, presented in Table 8, show a noticeable increase in average MRs during night time compared to day time for each demographic group. For example, the average MRs for males and females at night are 33.20% and 36.57%, respectively, compared to 17.24% and 17.32% during the day. A similar pattern is observed for age and skin tone. This indicates that the transition from day time to night time influences the performance of pedestrian detectors, with statistically significant higher MRs observed at night.

Then, we investigate whether the performance is equally decreased for different demographic groups. Specifically, we explore the fairness change from day time to night time. Table 8 shows the results, with statistically significant biases (i.e., MR differences) emphasized in shading.

Regarding gender, we observe a shift in the MR difference between males and females from day time to night time. During the day, there is only a slight −0.09% difference for males and females. However, during night time, the difference increases to −3.37% with statistically significant, indicating a notable change in fairness. In the night time condition, all pedestrian detectors exhibit higher MRs for females compared to males, with six of them showing statistically significant differences. This potentially worsens concerns regarding female safety during nighttime out, a prevalent societal issue [67].

For age, all detectors exhibit significant biases during both day time and night time. Moreover, the MR difference for children and adults increases from day time to night time, with the average difference increasing from −22.45% during the day to −25.48% at night, which is statistically significant. This suggests a higher probability of children being undetected during night time.

For skin tone, the MR difference between dark-skin and light-skin groups increases from day time to night time, with the average difference increasing from 0.15% at day time to 3.16% at night time. Nevertheless, the overall skin tone bias is not statistically significant during both day time and night time.

Table 8. (RQ2.1) MRs and EOD of Each Pedestrian Detector under Day Time and Night Time

Gender						
Detectors	Day Time			Night Time		
	MR Male	MR Female	EOD	MR Male	MR Female	EOD
YOLOX	9.73%	9.38%	0.35%	9.99%	14.53%	−4.54%
RetinaNet	10.42%	11.79%	−1.37%	11.99%	15.48%	−3.49%
Faster RCNN	3.66%	3.59%	0.07%	4.41%	6.06%	−1.66%
Cascade RCNN	3.58%	3.68%	−0.10%	5.11%	5.87%	−0.76%
ALFNet	24.52%	24.39%	0.13%	57.29%	63.42%	−6.14%
CSP	26.76%	26.38%	0.38%	61.57%	67.09%	−5.51%
MGAN	25.48%	26.29%	−0.81%	47.59%	47.76%	−0.17%
PRNet	33.73%	33.07%	0.66%	67.63%	72.33%	−4.70%
Average	17.24%	17.32%	−0.09%	33.20%	36.57%	−3.37%
Age						
Detectors	Day Time			Night Time		
	MR Adult	MR Child	EOD	MR Adult	MR Child	EOD
YOLOX	10.80%	40.77%	−29.97%	17.81%	54.93%	−37.12%
RetinaNet	12.73%	41.92%	−29.19%	18.97%	61.97%	−43.01%
Faster RCNN	4.50%	25.00%	−20.50%	7.34%	33.80%	−26.46%
Cascade RCNN	4.52%	25.58%	−21.05%	6.81%	33.80%	−26.99%
ALFNet	26.52%	49.42%	−22.90%	65.10%	83.10%	−18.00%
CSP	28.64%	46.92%	−18.28%	63.42%	76.06%	−12.64%
MGAN	27.87%	43.46%	−15.59%	49.72%	69.01%	−19.30%
PRNet	34.80%	56.92%	−22.13%	72.60%	92.96%	−20.36%
Average	18.80%	41.25%	−22.45%	37.72%	63.20%	−25.48%
Skin Tone (LS: Light Skin, DS: Dark Skin)						
Detectors	Day Time			Night Time		
	MR LS	MR DS	EOD	MR LS	MR DS	EOD
YOLOX	4.89%	4.03%	0.86%	7.06%	2.52%	4.53%
RetinaNet	7.44%	4.33%	3.11%	13.38%	5.04%	8.34%
Faster RCNN	5.36%	3.13%	2.23%	9.00%	4.20%	4.80%
Cascade RCNN	4.84%	2.99%	1.86%	7.30%	3.36%	3.94%
ALFNet	37.05%	38.66%	−1.61%	73.48%	73.95%	−0.47%
CSP	55.90%	59.85%	−3.95%	92.21%	92.44%	−0.22%
MGAN	47.12%	49.70%	−2.58%	85.64%	84.03%	1.61%
PRNet	54.39%	54.63%	−0.24%	89.29%	86.55%	2.74%
Average	27.31%	27.16%	0.15%	47.17%	44.01%	3.16%

Statistically significant unfairness results are shaded. We find that reduced brightness conditions not only decrease the performance of pedestrian detectors but also exacerbate their biases. Notably, while pedestrian detectors generally do not exhibit significant gender bias during day time, biases against females become pronounced with six out of eight detectors showing significant biases during night time.

> *Answer to RQ2.1*: Lower brightness conditions not only diminish the performance of pedestrian detectors but also exacerbate their bias. Particularly, during day time, pedestrian detectors generally do not exhibit significant gender bias, whereas six out of eight detectors demonstrate significant biases against females during night time.

Table 9. (RQ2.2) MRs and EOD of Each Pedestrian Detector under Different Contrast Levels

Gender									
Detectors	Level 3			Level 2			Level 1		
	Male MR	Female MR	EOD	Male MR	Female MR	EOD	Male MR	Female MR	EOD
YOLOX	6.30%	5.21%	1.09%	9.42%	10.42%	−1.00%	13.99%	14.71%	−0.72%
RetinaNet	7.89%	9.46%	−1.56%	10.47%	12.76%	−2.29%	13.99%	14.24%	−0.25%
Faster RCNN	3.83%	2.34%	1.49%	3.66%	4.16%	−0.50%	4.34%	5.32%	−0.98%
Cascade RCNN	3.43%	2.44%	0.98%	3.65%	4.26%	−0.62%	5.12%	5.01%	0.11%
ALFNet	33.33%	31.67%	1.66%	30.28%	32.66%	−2.38%	31.09%	34.74%	−3.65%
CSP	40.11%	37.41%	2.70%	32.61%	34.80%	−2.19%	31.54%	35.45%	−3.90%
MGAN	38.52%	38.89%	−0.38%	29.25%	31.23%	−1.98%	24.61%	24.10%	0.51%
PRNet	48.41%	46.97%	1.43%	39.45%	41.52%	−2.06%	36.92%	38.11%	−1.19%
Average	22.73%	21.80%	0.93%	19.85%	21.48%	−1.63%	20.20%	21.46%	−1.26%
Age									
Detectors	Level 3			Level 2			Level 1		
	Adult MR	Child MR	EOD	Adult MR	Child MR	EOD	Adult MR	Child MR	EOD
YOLOX	6.91%	32.89%	−25.98%	12.05%	40.92%	−28.87%	18.23%	53.23%	−35.00%
RetinaNet	10.28%	36.84%	−26.56%	13.85%	41.94%	−28.10%	18.70%	56.45%	−37.75%
Faster RCNN	4.00%	23.68%	−19.69%	5.08%	26.09%	−21.00%	6.56%	27.42%	−20.86%
Cascade RCNN	3.33%	19.74%	−16.41%	5.05%	25.58%	−20.52%	6.48%	33.87%	−27.39%
ALFNet	34.26%	52.63%	−18.37%	35.42%	54.73%	−19.32%	42.24%	50.00%	−7.76%
CSP	40.72%	53.95%	−13.23%	36.92%	52.17%	−15.25%	38.72%	42.74%	−4.02%
MGAN	40.38%	44.74%	−4.35%	33.20%	48.85%	−15.65%	30.68%	40.32%	−9.65%
PRNet	48.33%	68.42%	−20.09%	43.46%	63.17%	−19.71%	46.67%	50.81%	−4.14%
Average	23.53%	41.61%	−18.08%	23.13%	44.18%	−21.05%	26.04%	44.35%	−18.32%
Skin Tone (LS: Light Skin, DS: Dark Skin)									
Detectors	Level 3			Level 2			Level 1		
	LS MR	DS MR	EOD	LS MR	DS MR	EOD	LS MR	DS MR	EOD
YOLOX	4.15%	4.91%	−0.75%	5.75%	3.35%	2.40%	2.52%	2.17%	0.35%
RetinaNet	7.27%	5.28%	1.99%	8.60%	4.18%	4.41%	10.08%	2.17%	7.91%
Faster RCNN	5.04%	4.91%	0.14%	6.16%	2.51%	3.65%	6.72%	2.17%	4.55%
Cascade RCNN	4.01%	4.53%	−0.52%	5.64%	2.09%	3.55%	5.04%	4.35%	0.69%
ALFNet	37.09%	36.60%	0.49%	42.62%	45.40%	−2.78%	72.27%	71.74%	0.53%
CSP	53.26%	57.74%	−4.47%	62.40%	65.48%	−3.08%	90.76%	97.83%	−7.07%
MGAN	49.26%	50.19%	−0.93%	52.30%	54.81%	−2.51%	84.03%	82.61%	1.42%
PRNet	54.15%	54.72%	−0.56%	59.92%	59.21%	0.71%	86.55%	89.13%	−2.58%
Average	26.78%	27.36%	−0.58%	30.42%	29.63%	0.80%	44.75%	44.02%	0.73%

On average, detectors exhibit the most biased results (i.e., the largest absolute value of EOD) under level 2, while demonstrating the fairest outcomes under level 3.

4.2.2 RQ2.2: Different Contrast Levels. Following the roadmap outlined in RQ2.1, we begin by comparing the overall MRs of pedestrian detectors across different contrast levels, as presented in Table 9. As explained in Section 3.3.3, we categorize driving scenarios into three contrast levels, with a higher level indicating greater contrast. However, we do not observe a consistent pattern in the results. Specifically, concerning light-skin and dark-skin pedestrians, we note a decrease in MRs overall with increasing contrast. Nonetheless, this pattern is not observed across other demographic groups.

We then examine the shifts in fairness, as evidenced by the trends in EOD with increasing contrast, outlined in Table 9. Our analysis reveals a consistent pattern: On average, detectors exhibit the most biased results (i.e., the largest absolute value of EOD) under level 2, while demonstrating the fairest outcomes under level 3.

Specifically, for gender, detectors achieve −1.63% EOD under level 2 (most biased), −1.26% under level 1, and 0.93% under level 3 (fairest). For age, detectors achieve −21.05% EOD under level 2 (most biased), −18.32% under level 1, and 18.08% under level 3 (fairest). For skin tone, detectors achieve 0.80% EOD under level 2 (most biased), 0.73% under level 1, and −0.58% under level 3 (fairest).

This finding is further supported by the observation that under level 2, the highest number of detectors exhibit significant biases for gender, age, and skin tone, whereas under level 3, the fewest detectors show significant biases. Specifically, under level 2, five detectors demonstrate significant biases regarding gender, whereas under level 3, only one detector does so. Under level 2, eight detectors all demonstrate significant biases regarding age, whereas under level 3, seven detectors do so. Under level 2, four detectors show significant biases regarding skin tone, whereas under level 3, no detector exhibits significant biases.

> *Answer to RQ2.2*: We classify driving scenarios into three contrast levels and observe that while there is not a clear pattern in the overall detection performance change with contrast variation, the highest contrast level consistently produces the fairest detection results across gender, age, and skin tone.

4.2.3 RQ2.3: Different Weather Conditions. As described in Section 3.3.3, we consider two weather conditions: non-rainy and rainy.

Table 10 presents the MRs of eight pedestrian detectors under the two conditions, with statistically significant unfairness results shaded. Overall, the MR for each demographic group increases in rainy weather conditions. This escalation could be attributed to droplets covering the camera and disrupting the detectors. A follow-up question is whether rainy weather fairly increases the MR for different demographic groups.

From Table 10, we observe that rainy weather may potentially mitigate bias in pedestrian detectors. Specifically, under non-rainy conditions, all eight detectors exhibit significant bias against children, while under rainy conditions, three out of four pedestrian-specific detectors no longer show significant age bias any more. Similarly, under non-rainy conditions, two of eight detectors display significant bias against skin tone, whereas under rainy conditions, none do. However, the improvement in fairness due to rainy weather is marginal. The EOD difference between rainy and non-rainy conditions is 1.89% for gender, −0.29% for age, and 1.13% for skin tone.

> *Answer to RQ2.3*: Rainy weather conditions decrease overall detection performance to a large extent but have a subtle impact on the fairness of pedestrian detectors. Specifically, three out of four pedestrian-specific detectors no longer exhibit significant age bias under rainy weather. However, the improvement in fairness is marginal.

5 Discussion

5.1 Fairness-Performance Tradeoff

It is widely acknowledged that fairness improvement usually comes at the cost of ML performance (e.g., accuracy) [28, 38, 44, 69]. Therefore, developers need to grapple with the challenge of optimizing ML performance without compromising fairness, encapsulating this tension as the "fairness-performance tradeoff."

Nonetheless, we have observed that pedestrian detectors exhibiting greater fairness (i.e., lower absolute EOD values) can achieve superior performance (i.e., lower overall MRs) under certain

Table 10. (RQ2.3) MRs and EOD of Each Pedestrian Detector under Non-Rainy and Rainy Weather Conditions

Gender						
Detectors	Non-Rainy Weather			Rainy Weather		
	MR Male	MR Female	EOD	MR Male	MR Female	EOD
YOLOX	9.58%	8.95%	0.63%	8.33%	5.17%	3.16%
RetinaNet	10.62%	11.35%	−0.74%	10.09%	8.05%	2.04%
Faster RCNN	3.68%	3.29%	0.39%	4.39%	2.87%	1.51%
Cascade RCNN	3.68%	3.48%	0.20%	4.39%	2.87%	1.51%
ALFNet	23.52%	22.92%	0.60%	42.54%	40.80%	1.74%
CSP	25.29%	24.35%	0.94%	50.44%	44.25%	6.19%
MGAN	24.63%	25.26%	−0.62%	50.88%	50.57%	0.30%
PRNet	32.73%	31.64%	1.09%	51.75%	50.57%	1.18%
Average	16.72%	16.41%	0.31%	27.85%	25.65%	2.20%
Age						
Detectors	Non-Rainy Weather			Rainy Weather		
	MR Adult	MR Child	EOD	MR Adult	MR Child	EOD
YOLOX	10.45%	39.95%	−29.50%	9.40%	30.00%	−20.60%
RetinaNet	12.49%	40.65%	−28.16%	13.61%	40.00%	−26.39%
Faster RCNN	4.36%	23.09%	−18.73%	4.86%	30.00%	−25.14%
Cascade RCNN	4.53%	23.79%	−19.25%	4.38%	25.00%	−20.62%
ALFNet	25.00%	47.34%	−22.34%	45.06%	75.00%	−29.94%
CSP	26.86%	44.11%	−17.25%	48.78%	65.00%	−16.22%
MGAN	26.49%	41.11%	−14.62%	53.00%	65.00%	−12.00%
PRNet	33.19%	53.58%	−20.39%	53.32%	75.00%	−21.68%
Average	17.92%	39.20%	−21.28%	29.05%	50.63%	−21.57%
Skin Tone (LS: Light Skin, DS: Dark Skin)						
Detectors	Non-Rainy Weather			Rainy Weather		
	MR LS	MR DS	EOD	MR LS	MR DS	EOD
YOLOX	4.19%	3.47%	0.72%	16.94%	13.51%	3.42%
RetinaNet	6.56%	3.81%	2.74%	21.77%	13.51%	8.26%
Faster RCNN	4.59%	2.60%	1.99%	15.32%	10.81%	4.51%
Cascade RCNN	3.93%	2.60%	1.33%	16.13%	8.11%	8.02%
ALFNet	34.90%	36.92%	−2.02%	66.13%	64.86%	1.26%
CSP	53.05%	57.02%	−3.97%	79.03%	89.19%	−10.16%
MGAN	44.68%	47.83%	−3.15%	72.58%	78.38%	−5.80%
PRNet	51.99%	52.86%	−0.87%	77.42%	81.08%	−3.66%
Average	25.49%	25.89%	−0.40%	45.67%	44.93%	0.73%

Statistically significant unfairness results are shaded. We find that rainy weather does not largely impact fairness regarding gender, age, and skin tone.

environmental conditions. For instance, in the results of RQ2.1, we find that during day time compared to night time, the eight pedestrian detectors achieve better detection results while simultaneously displaying reduced absolute EOD values related to age, gender, and skin tone. Similarly, the results of RQ2.2 reveal that the highest level of contrast consistently leads to improved fairness regarding gender, age, and skin tone, while also demonstrating the best overall detection performance in specific cases. These findings offer a positive counterpoint to previous fairness-performance tradeoff theory, suggesting that we can enhance both detection performance and fairness by adjusting the brightness and contrast of captured images.

Table 11. Sensitivity Analysis of Age Bias under Varying Proportions of Adult Data

Detectors	Adult Data Proportion 80%			Adult Data Proportion 60%			Adult Data Proportion 40%			Adult Data Proportion 20%		
	MR Adult	MR Child	EOD	MR Adult	MR Child	EOD	MR Adult	MR Child	EOD	MR Adult	MR Child	EOD
YOLOX	12.68%	42.47%	−29.79%	12.92%	42.47%	−29.55%	12.73%	42.47%	−29.74%	14.06%	42.47%	−28.41%
RetinaNet	14.25%	44.33%	−30.08%	14.31%	44.33%	−30.02%	14.72%	44.33%	−29.61%	15.07%	44.33%	−29.26%
Faster RCNN	5.20 %	26.06%	−20.86%	5.38 %	26.06%	−20.68%	5.30 %	26.06%	−20.75%	5.29%	26.06%	−20.77%
Cascade RCNN	5.07 %	26.57%	−21.49%	5.24 %	26.57%	−21.32%	5.24 %	26.57%	−21.33%	5.02%	26.57%	−21.55%
ALFNet	36.35%	53.47%	−17.11%	35.88%	53.47%	−17.58%	36.27%	53.47%	−17.20%	34.12%	53.47%	−19.35%
CSP	37.51%	50.42%	−12.92%	37.03%	50.42%	−13.40%	37.35%	50.42%	−13.07%	34.95%	50.42%	−15.47%
MGAN	33.18%	46.53%	−13.35%	32.94%	46.53%	−13.60%	32.87%	46.53%	−13.66%	30.65%	46.53%	−15.89%
PRNet	44.40%	61.25%	−16.85%	44.18%	61.25%	−17.08%	44.05%	61.25%	−17.20%	42.09%	61.25%	−19.17

Statistically significant unfairness results are shaded. The analysis shows that the observed age bias remains significant across all distributions, indicating that our findings are robust even under substantial variations in the dataset's demographic composition.

5.2 Impact of Data Imbalance

Imbalanced data distribution, particularly the under-representation of certain demographic groups such as children, may pose a threat to the validity of our findings. To evaluate the robustness of our results under varying data distributions, we perform a sensitivity analysis on age, where the representation of children is most limited and exhibits notable bias across all pedestrian detectors.

In this analysis, we systematically vary the proportion of adult samples by randomly sampling them at 80%, 60%, 40%, and 20% of their original size, while maintaining the number of samples for children constant. This creates four new data distributions with varying adult-to-child ratios. We then re-evaluate the age bias for each distribution. The results, presented in Table 11, show that the observed age bias remains significant across all these distributions, indicating that our findings are robust even under substantial variations in the dataset's demographic composition.

5.3 Implications

5.3.1 Implication for Researchers. Our empirical results offer valuable research opportunities for researchers. *(1) Fairness improvement of autonomous driving systems via image editing.* Our findings reveal that brightness and contrast of images can impact the fairness of pedestrian detection. Thus, to improve fairness, a practical solution is to design specific post-processing image editing techniques to adjust contrast and brightness of captured input images, such as increasing brightness and contrast levels to dynamically counterbalance existing bias toward children and female pedestrians in low-brightness and low-contrast conditions. Moreover, previous discussions in Section 5.1 show that these solutions also have the potential to improve detection performance. *(2) Real-time adaptive fairness improvement of autonomous driving systems.* Adapting software systems to increasingly dynamic environments poses significant challenges in software development, an area that has been an important task in SE research [58]. Our empirical findings contribute to this ongoing task by introducing a new challenge: designing real-time adaptive fairness improvement methods for autonomous driving systems. Specifically, our research reveals that autonomous driving systems exhibit greater biases under specific driving scenarios influenced by changes in brightness and contrast. Given the common occurrence of such changes in real-world driving conditions, there is an urgent need to develop specific methods to ensure the consistent fairness of autonomous driving systems in response to these environmental dynamics. *(3) Fairness improvement of autonomous driving systems via multi-objective optimization.* Given our finding that the fairness-performance tradeoff may not hold in the context of pedestrian detection, researchers have the opportunity to develop multi-objective optimization strategies for model training, enabling the concurrent optimization of fairness and detection performance. At the same time, based on previous findings

regarding brightness and contrast, a possible approach could involve integrating fairness constraints during model training by incorporating contrast-aware or brightness-aware regularization techniques. The model could penalize overly confident predictions in low-contrast or low-brightness conditions, encouraging it to treat predictions more cautiously in these scenarios. Additionally, augmenting data by enhancing these attributes in images of under-represented groups (e.g., children in low-brightness conditions) could help the model generalize better. Synthetic image generation that varies brightness and contrast during training may further expose the model to diverse brightness and contrast conditions, potentially reducing bias. *(4) Addressing inherent limitations for detecting small objects of pedestrian detectors.* For age bias, it is important to address the inherent limitations of pedestrian detectors, particularly their difficulty in accurately detecting small pedestrians. These implications underline the importance of factoring these elements into effective fairness improvement strategies. *(5) Open datasets and annotations for fairness assessment.* There is a shortage of available datasets annotated with a wide range of demographic labels for pedestrian detection. To address this constraint, we have undertaken the task of manually annotating four datasets with information related to gender, age, and skin tone, and publicly released these annotated datasets. It is worth noting that there are additional demographic attributes that also need to be considered in the context of fairness, such as disability status and religion. We encourage researchers and practitioners to join us in the effort to promote fairness in pedestrian detection by creating and sharing datasets that encompass various demographic attributes.

5.3.2 Implication for Developers. Fairness is a critical non-functional requirement for software applications, but our study demonstrates the existence of significant bias in state-of-the-art pedestrian detectors. It is imperative for developers to prioritize their efforts in this domain. Addressing these concerns goes beyond enhancing the quality of autonomous driving systems; it also serves to safeguard autonomous driving companies from ethical, reputational, financial, and legal repercussions that may arise in the event of violations of anti-discrimination laws. Moreover, it is crucial for developers of autonomous driving systems to consider the influence of various environmental factors on fairness during the development process.

5.3.3 Implication for Policy Makers. Governments have a role in raising awareness about the potential biases in autonomous driving systems. *(1) Regulatory measures.* Autonomous driving systems play a crucial role in ensuring human safety. Our results reveal that fairness issues exist in modern pedestrian detection models. Therefore, it is essential for policy makers to enact regulations and standards that safeguard the rights of all individuals and address these concerns appropriately. *(2) Enhanced protection for vulnerable pedestrians.* Policies should also aim to address safety issues for vulnerable pedestrians such as children and females, who are, as revealed by this study, more likely to be overlooked by the current detection models.

5.4 Threats to Validity

Manual Labeling. The labeling process involves possible subjectivity, posing threats to the validity of the analysis. To mitigate this threat, two annotators and an arbitrator are involved in the labeling process: First each image is independently labeled by two annotators, and in cases where conflicts in labeling arise, we seek the expertise of an arbitrator to resolve such discrepancies and arrive at a consensus. The inter-rater agreement between the two annotators is high, which demonstrates the reliability of the labeling schema and procedure adopted herein.

Selection of Pedestrian Detectors. Our study is based on eight pedestrian detectors, which may lead to possible bias in this study. To mitigate this threat, we select representative pedestrian detectors based on two considerations. On one hand, the selected detectors are widely studied in pedestrian detection and autonomous driving literature [20, 71, 78, 79]. On the other hand, the selected models

cover the two typical types of pedestrian detection methods (i.e., one-stage detection and two-stage detection), ensuring a comprehensive representation of the techniques used in the field.

Selection of Datasets. The choice of datasets presents a potential threat to validity. Current literature on autonomous driving sometimes relies on automated techniques to simulate adverse weather conditions such as fog, rain, and snow. However, these synthetic images may not accurately reflect real-life conditions [74] (e.g., simply simulating fog by applying a set of filters over the original image, without considering the complex transformations that occur in actual road scenes under such conditions). Although real-world datasets for adverse weather like fog now exist [14], accurately annotating demographic attributes of pedestrians under foggy conditions remains challenging in the context of this study. These weather effects can obscure important features such as skin tone and blur physical outlines, making it difficult to discern attributes like gender and age. To address these concerns, we utilize four real-world datasets extensively explored in autonomous driving research, comprising a total of 8,311 images. Additionally, we have augmented these datasets with 16,070 gender labels, 20,115 age labels, and 3,513 skin tone labels. The extensive scale of data ensures the reliability of our results and mitigates the limitations associated with synthetic images.

Selection of Evaluation Measures. The fairness measures that we employ may introduce potential limitations. To mitigate this concern, we consider three commonly used fairness measures: SPD, EOD, and AOD. Upon careful discussion, we conclude that AOD may not be suitable for our study (Section 3.4), leading us to focus on SPD and EOD. We demonstrate that SPD and EOD lead to equivalent observations for pedestrian detection. Ultimately, we opt to use EOD (i.e., SPD) for our analysis, as it involves the comparison of the MRs of different demographic groups. This ensures a consistent evaluation with the current pedestrian detection research, as MR is the most widely adopted metric for measuring the performance of pedestrian detectors in the literature [17, 46].

Selection of Sensitive Attributes. The selection of sensitive attributes may introduce potential limitations to our study. To mitigate this concern, we focus on attributes that are most identifiable in autonomous driving images and are recognized as the three most extensively studied sensitive attributes in the fairness literature: gender, age, and skin tone [43]. While other sensitive attributes, such as disability status, precise age determinations for young and elderly individuals, or nationality, are relevant to fairness considerations, they present significant challenges for consistent identification from typical vehicle-mounted camera data due to ambiguous feature judgment. In contrast, gender, age (adult and child), and skin tone are more readily observable characteristics in such images, allowing for more reliable annotations.

6 Conclusion

This article presents the first comprehensive study on fairness testing of eight state-of-the-art pedestrian detectors, using four widely studied testing datasets. We investigate the fairness aspects of these detectors regarding gender, age, and skin tone. Furthermore, we conduct an in-depth analysis of fairness in various driving scenarios. Our findings reveal significant bias in the current pedestrian detectors, particularly toward children. Additionally, pedestrian detectors demonstrate significant gender biases during night time, potentially exacerbating the prevalent societal issue of female safety concerns during night-time out. Regarding skin tone, we observe balanced detection performance for light-skin and dark-skin individuals overall. As part of our contribution, we publicly release large-scale real-world pedestrian detection datasets with gender, age, and skin tone labels. These datasets aim to facilitate the future fairness research in autonomous driving. The insights gained in this study can pave the way for more fair and unbiased autonomous driving systems in the future.

Data Availability Statement

Our GitHub repository [9] contains datasets, sensitive attribute labels, scripts, and results of this work to facilitate replication and extension.

References

[1] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. 2019. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5187–5196.

[2] Daisuke Wakabayashi. 2018. Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. Retrieved from https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html

[3] Andrea Stocco, Michael Weiss, Marco Calzana, and Paolo Tonella. 2020. Misbehaviour prediction for autonomous driving systems. In *Proceedings of the 42nd International Conference on Software Engineering (ICSE '20)*, 359–371.

[4] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. 2022. Pedestrian detection: Domain generalization, CNNs, transformers and beyond. arXiv:2201.03176. Retrieved from https://arxiv.org/abs/2201.03176

[5] GitHub. 2023. Apollo. Retrieved from https://github.com/ApolloAuto/apollo-model-yolox

[6] Wikipedia. 2024. Afro-Germans. Retrieved from https://en.wikipedia.org/wiki/Afro-Germans

[7] Wikipedia. 2024. Demographics-of-Germany. Retrieved from https://en.wikipedia.org/wiki/Demographics_of_Germany

[8] Mohamad Moslimani, Christine Tamir, Abby Budiman, Luis Noe-Bustamante, and Lauren Mora. 2024. Facts about the U.S. Black Population. Retrieved from https://www.pewresearch.org/social-trends/fact-sheet/facts-about-the-us-black-population/

[9] GitHub. 2024. Replication Package. Retrieved from https://github.com/xinyuelxy/Bias-Behind-the-Wheel-Fairness-Testing-of-Autonomous-Driving-Systems

[10] Zhenpeng Chen, Yanbin Cao, Yuanqiang Liu, Haoyu Wang, Tao Xie, and Xuanzhe Liu. 2020. A comprehensive study on challenges in deploying deep learning based software. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '20)*, 750–762.

[11] Solon Barocas and Andrew D. Selbst. 2016. Big Data's disparate impact. *California Law Review* 104 (2016), 671.

[12] Md. Al-Amin Bhuiyan and Abdul Raouf Khan. 2018. Image quality assessment employing RMS contrast and histogram similarity. *The International Arab Journal of Information Technology* 15, 6 (2018), 983–989.

[13] D. Biddle. 2005. *Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing*. Gower.

[14] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. 2020. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '20)*, 11679–11689.

[15] Sumon Biswas and Hridesh Rajan. 2020. Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '20)*, 642–653.

[16] Sumon Biswas and Hridesh Rajan. 2021. Fair preprocessing: Towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21)*, 981–993.

[17] Martim Brandao. 2019. Age and gender bias in pedestrian detection algorithms. arXiv:1906.10490. Retrieved from https://arxiv.org/abs/1906.10490

[18] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M. Gavrila. 2019. EuroCity Persons: A novel benchmark for person detection in traffic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 8 (2019), 1844–1861.

[19] Zhaowei Cai and Nuno Vasconcelos. 2018. Cascade R-CNN: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6154–6162.

[20] Jiale Cao, Yanwei Pang, Jin Xie, Fahad Shahbaz Khan, and Ling Shao. 2020. From handcrafted to deep features for pedestrian detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2020), 4913–4934.

[21] G. Casella and R. L. Berger. 2007. *Statistical Inference* (2nd. ed.). Duxbury Press.

[22] Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in machine learning software: Why? How? What to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21)*, 429–440.

[23] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: A way to build fair ML software. In *Proceedings of the 28th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '20)*, 654–665.

[24] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. 2019. MMDetection: Open MMLab detection toolbox and benchmark. arXiv:1906.07155. Retrieved from https://arxiv.org/abs/1906.07155

[25] Zhenpeng Chen, Huihan Yao, Yiling Lou, Yanbin Cao, Yuanqiang Liu, Haoyu Wang, and Xuanzhe Liu. 2021. An empirical study on deployment faults of deep learning based mobile applications. In *Proceedings of the 43rd IEEE/ACM International Conference on Software Engineering (ICSE '21)*, 674–685.

[26] Zhenpeng Chen, Jie M. Zhang, Max Hort, Mark Harman, and Federica Sarro. 2024. Fairness testing: A comprehensive survey and analysis of trends. *ACM Transactions on Software Engineering and Methodology* 33, 5 (2024), 137:1–137:59.

[27] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2022. MAAT: A novel ensemble approach to addressing fairness and performance bugs for machine learning software. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '22)*, 1122–1134.

[28] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2023. A comprehensive empirical study of bias mitigation methods for machine learning classifiers. *ACM Transactions on Software Engineering and Methodology* 32, 4 (2023), 106:1–106:30.

[29] Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. 2024. Fairness improvement with multiple protected attributes: How far are we? In *Proceedings of the 46th IEEE/ACM International Conference on Software Engineering (ICSE '24)*, 160:1–160:13.

[30] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1960), 37–46.

[31] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv:1808.00023. Retrieved from https://arxiv.org/abs/1808.00023

[32] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. 2009. Pedestrian detection: A benchmark. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 304–311.

[33] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2012. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (2012), 743–761.

[34] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012*. Shafi Goldwasser (Ed.), ACM, 214–226. DOI: https://doi.org/10.1145/2090236.2090255

[35] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, 259–268.

[36] Anthony Finkelstein, Mark Harman, S. Afshin Mansouri, Jian Ren, and Yuanyuan Zhang. 2008. "Fairness analysis" in requirements assignments. In *Proceedings of the 16th IEEE International Requirements Engineering Conference (RE '08)*, 115–124.

[37] Joshua Garcia, Yang Feng, Junjie Shen, Sumaya Almanee, Yuan Xia, and Qi Alfred Chen. 2020. A comprehensive study of autonomous vehicle bugs. In *Proceedings of the 42nd International Conference on Software Engineering (ICSE '20)*, 385–396.

[38] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward Pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM '22)*, 316–324.

[39] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. Yolox: Exceeding yolo series in 2021. arXiv:2107.08430. Retrieved from https://arxiv.org/abs/2107.08430

[40] Usman Gohar, Sumon Biswas, and Hridesh Rajan. 2023. Towards understanding fairness and its composition in ensemble machine learning. In *Proceedings of the 45th IEEE/ACM International Conference on Software Engineering (ICSE '23)*, 1533–1545.

[41] An Guo, Yang Feng, and Zhenyu Chen. 2022. LiRTest: Augmenting LiDAR point clouds for automated testing of autonomous driving systems. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA '22)*, 480–492.

[42] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. 2020. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '20)*, 11323–11332.

[43] Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. 2024. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Journal on Responsible Computing* 1, 2 (2024), 1–52.

[44] Max Hort, Jie M. Zhang, Federica Sarro, and Mark Harman. 2021. Fairea: A model behaviour mutation approach to benchmarking bias mitigation methods. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21)*, 994–1006.

[45] Vinit Jakhetiya, Weisi Lin, Sunil Prasad Jaiswal, Ke Gu, and Sharath Chandra Guntuku. 2018. Just noticeable difference for natural images using RMS contrast and feed-back mechanism. *Neurocomputing* 275 (2018), 366–376.

[46] Shunsuke Kogure, Kai Watabe, Ryosuke Yamada, Yoshimitsu Aoki, Akio Nakamura, and Hirokatsu Kataoka. 2022. Age should not matter: Towards more accurate pedestrian detection via self-training. In *Proceedings of the AAAI Workshop on Artificial Intelligence with Biased or Scarce Data (AIBSD '22)*, 33, 1 (2022). Retrieved from https://www.mdpi.com/2813-0324/3/1/11

[47] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.

[48] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2980–2988.

[49] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single shot multibox detector. In *Proceedings of the 14th European Conference on Computer Vision (ECCV '16)*, 21–37.

[50] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. 2018. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV '18)*, 618–634.

[51] Xuanzhe Liu, Diandian Gu, Zhenpeng Chen, Jinfeng Wen, Zili Zhang, Yun Ma, Haoyu Wang, and Xin Jin. 2023. Rise of distributed deep learning training in the big model era: From a software engineering perspective. *ACM Transactions on Software Engineering and Methodology* 32, 6 (2023), 156:1–156:26.

[52] Giang Nguyen, Sumon Biswas, and Hridesh Rajan. 2023. Fix fairness, don't ruin accuracy: Performance aware fairness repair using AutoML. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '23)*, 502–514.

[53] Nan Niu, Wentao Wang, and Arushi Gupta. 2016. Gray links in the use of requirements traceability. In *Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE '16)*, 384–395.

[54] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. 2019. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4967–4975.

[55] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. 2019. DeepXplore: Automated whitebox testing of deep learning systems, Commun. ACM 62, 11 (2019), 137–145. Retrieved from https://doi.org/10.1145/3361566

[56] Eli Peli. 1990. Contrast in complex images. *Journal of the Optical Society of America A* 7, 10 (Oct. 1990), 2032–2040.

[57] Denis G. Pelli and Peter Bex. 2013. Measuring contrast sensitivity. *Vision Research* 90 (2013), 10–14.

[58] Nauman A. Qureshi and Anna Perini. 2009. Engineering adaptive requirements. In *Proceedings of the 2009 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems (SEAMS '09)*, 126–131.

[59] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788.

[60] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.

[61] Xiaolin Song, Kaili Zhao, Wen-Sheng Chu, Honggang Zhang, and Jun Guo. 2020. Progressive refinement network for occluded pedestrian detection. In *Proceedings of the 16th European Conference on Computer Vision (ECCV '20)*, 32–48.

[62] Zeyu Sun, Zhenpeng Chen, Jie Zhang, and Dan Hao. 2024. Fairness testing of machine translation systems. *ACM Transactions on Software Engineering and Methodology* 33, 6 (2024), 156.

[63] Yuchi Tian, Kexin Pei, Suman Jana, and Baishakhi Ray. 2018. DeepTest: Automated testing of deep-neural-network-driven autonomous cars. In *Proceedings of the 40th International Conference on Software Engineering (ICSE '18)*, 303–314.

[64] Alvaro Veizaga, Mauricio Alférez, Damiano Torre, Mehrdad Sabetzadeh, and Lionel C. Briand. 2021. On systematically building a controlled natural language for functional requirements. *Empirical Software Engineering* 26, 4 (2021), 79.

[65] Yuxuan Wan, Wenxuan Wang, Pinjia He, Jiazhen Gu, Haonan Bai, and Michael R. Lyu. 2023. BiasAsker: Measuring the bias in conversational AI system. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '23)*, 515–527.

[66] Chao Wang, Zhenpeng Chen, and Minghui Zhou. 2023. AutoML from software engineering perspective: Landscapes and challenges. In *Proceedings of the 20th IEEE/ACM International Conference on Mining Software Repositories (MSR '23)*, 39–51.

[67] Xin Wang, Ivan Ka Wai Lai, and Kun Wang. 2023. Do young women travellers really consider the risk of sexual harassment during night travel? Evening travel vs midnight travel. *Tourism Review* 78, 1 (2023), 58–71.

[68] Jinfeng Wen, Zhenpeng Chen, Yi Liu, Yiling Lou, Yun Ma, Gang Huang, Xin Jin, and Xuanzhe Liu. 2021. An empirical study on challenges of application development in serverless computing. In *Proceedings of the 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '21)*, 416–428.

[69] Michael L. Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking fairness: A trade-off revisited. In *Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS '19)*, 8780–8789.

[70] Benjamin Wilson, Judy Hoffman, and Jamie H. Morgenstern. 2019. Predictive inequity in object detection. arXiv:1902.11097. Retrieved from https://arxiv.org/abs/1902.11097

[71] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '20)*, 2633–2642.

[72] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS '17)*, 962–970.

[73] Jie M. Zhang and Mark Harman. 2021. "Ignorance and prejudice" in software fairness. In *Proceedings of 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE '21)*, 1436–1447.

[74] Jie M. Zhang, Mark Harman, Lei Ma, and Yang Liu. 2022. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering* 48, 2 (2022), 1–36.

[75] Mengdi Zhang and Jun Sun. 2022. Adaptive fairness improvement based on causality analysis. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE '22)*, 6–17.

[76] Mengdi Zhang, Jun Sun, Jingyi Wang, and Bing Sun. 2023. TestSGD: Interpretable testing of neural networks against subtle group discrimination. *ACM Transactions on Software Engineering and Methodology* 32, 6 (2023), 137:1–137:24.

[77] Mengshi Zhang, Yuqun Zhang, Lingming Zhang, Cong Liu, and Sarfraz Khurshid. 2018. DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems. In *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE '18)*, 132–142.

[78] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. 2016. How far are we from solving pedestrian detection? In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '16)*, 1259–1267.

[79] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. CityPersons: A diverse dataset for pedestrian detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '17)*, 4457–4465.

[80] Yixuan Zhang, Shangtong Cao, Haoyu Wang, Zhenpeng Chen, Xiapu Luo, Dongliang Mu, Yun Ma, Gang Huang, and Xuanzhe Liu. 2024. Characterizing and detecting WebAssembly runtime bugs. *ACM Transactions on Software Engineering and Methodology* 33, 2 (2024), 37:1–37:29.

[81] Husheng Zhou, Wei Li, Zelun Kong, Junfeng Guo, Yuqun Zhang, Bei Yu, Lingming Zhang, and Cong Liu. 2020. DeepBillboard: Systematic physical-world testing of autonomous driving systems. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering (ICSE '20)*, 347–358.

[82] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object detection in 20 years: A survey. *Proceedings of the IEEE* 111, 3 (2023), 257–276.

[83] Feilong Zuo, Zhengxiong Luo, Junze Yu, Zhe Liu, and Yu Jiang. 2021. PAVFuzz: State-sensitive fuzz testing of protocols in autonomous vehicles. In *Proceedings of the 58th ACM/IEEE Design Automation Conference (DAC '21)*, 823–828.

Received 4 April 2024; revised 5 October 2024; accepted 14 October 2024