Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20226-7

Bias-aware face mask detection dataset

Alperen Kantarcı[2] · Ferda Ofli[1] · Muhammad Imran[1] · Hazım Kemal Ekenel[2]

Received: 11 January 2023 / Revised: 14 August 2024 / Accepted: 4 September 2024

Abstract
In December 2019, a novel coronavirus (COVID-19) spread so quickly around the world that many countries had to set mandatory face mask rules in public areas to reduce the transmission of the virus. To monitor public adherence, researchers aimed to rapidly develop efficient systems that can detect faces with masks automatically. However, the lack of representative and novel datasets posed challenges for training efficient models. Early attempts to collect face mask datasets did not account for potential race, gender, and age biases. Therefore, the resulting models show inherent biases toward specific race groups, such as Asian or Caucasian. In this work, we present a novel face mask detection dataset that contains images posted on Twitter during the pandemic from around the world. Unlike previous datasets, the proposed Bias-Aware Face Mask Detection (BAFMD) dataset contains more images from underrepresented races and age groups to mitigate the problem of the face mask detection task. We perform experiments to investigate potential biases in widely used face mask detection datasets and illustrate that the BAFMD dataset yields models with better performance and generalization ability. The dataset is publicly available at https://github.com/Alpkant/BAFMD.

Keywords Face mask detection · Bias · Social media · Dataset · Computer vision · Deep learning

✉ Ferda Ofli
fofli@hbku.edu.qa

Alperen Kantarcı
kantarcia@itu.edu.tr

Muhammad Imran
mimran@hbku.edu.qa

Hazım Kemal Ekenel
ekenel@itu.edu.tr

[1] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

[2] Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey

Published online: 02 October 2024

1 Introduction

The rapid worldwide spread of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2) or COVID-19 created a global pandemic. More than 127 million cases were confirmed within a year [1] because of the virus. Medical experts, public health agencies, and governments worldwide recommended a series of prevention measures, such as social distancing, travel bans, country-wide lockdowns, and wearing face masks in public spaces [2]. Practical measures, such as face masks, have been adopted for more extended periods. Computer vision researchers and practitioners rapidly started developing automatic detection methods due to this massive increase in face mask usage, as existing face detection methods struggled to detect faces with masks. Since monitoring and screening applications of face mask detection systems help society prevent virus transmission, it became essential to develop an accurate and fair face mask detection system.

Face mask detection has been an understudied sub-topic within face detection research until the COVID-19 pandemic. The earliest work on occluded face detection focused on occlusions such as glasses, hands covering the lower part of the face, and pollution-masks [3–6]. Moreover, these early works only focused on Asian countries where face mask usage was already common even before the COVID-19 pandemic because of excessive air pollution and SARS-associated coronavirus [7]. Therefore, when the pandemic started, researchers combined available datasets [8] that contained Asian people wearing face masks with other standard face detection datasets, such as WIDER [9], or they tried to produce datasets with artificial face masks. Although these approaches showed better performance for the masked face detection task, their application in the real-world setting remained limited mainly due to imbalanced race distribution in the datasets. Biased data leads to biased models that may not be applicable to certain population segments, e.g., people with dark skin color. Such issues potentially raise ethical concerns about the fairness of automated systems. Therefore, a system that will be used in daily life across the world should be trained with a more representative and demographically balanced dataset to mitigate biases [10, 11]. A study [12] shows that most existing large-scale face databases are biased towards "lighter skin" faces, e.g., Caucasian, compared to "darker" faces, e.g., Black. However, such a study has not been conducted on face mask detection datasets. In our observations, we noticed a clear selection bias toward Asian faces, as the most famous face mask datasets are collected in Asia.

In this paper, we address the critical need for a representative face mask detection dataset, with a particular focus on bias and fairness. To this end, we introduce a new dataset that offers a more balanced distribution across gender, race, and age, utilizing images sourced globally from Twitter. This dataset is made publicly available[1] to support and facilitate future research. We present demographic statistics of the dataset using advanced face attribute prediction methods. Additionally, we conduct experiments using existing face mask detectors alongside our newly proposed model on widely-used face mask detection datasets. Our model leverages the YOLO-v5 object detector in conjunction with our balanced dataset. Finally, we demonstrate that our bias-aware dataset enables models to generalize better and achieve superior performance in detection tasks compared to state-of-the-art face mask detection models.

[1] The dataset is available at https://github.com/Alpkant/BAFMD

2 Related work

Face occlusion [13–15], object detection [16, 17], and face detection [18, 19] are well-researched fields that can provide good baselines for developing face mask detection systems. However, face mask detection has received limited attention among the detection tasks and was studied within the broader occluded face detection problem. Therefore, only a few datasets were available when the COVID-19 pandemic started. During the pandemic, researchers published numerous studies in the face mask detection field. These studies mostly focus on either collecting new datasets or combining different datasets to obtain a representative face mask detection dataset as listed in Table 1. However, the high cost of annotating a new dataset prevented most of the researchers from collecting face mask datasets. Thus, researchers focused on either creating artificial face masks on face images [20] or refining the annotations of the publicly available face occlusion datasets [21].

Previous works that proposed a combination of datasets [21, 22] mostly use the MAFA dataset [3], which was collected from the Internet in 2017 as a face occlusion detection dataset. MAFA contains various face occlusions, including face masks. However, most of the images are collected from Asian countries, where face masks are widely used by the population. This is also the case for many different face mask detection datasets, such as MFDD [22]. Having a racial bias in the training dataset is a huge drawback for creating universal face mask detection models as they would be biased towards specific race groups. In contrast, we collected images from different ethnicities and age groups to create a more representative dataset. Furthermore, our dataset contains variety of face mask designs and textures, that increased during the COVID-19 pandemic.

Figure 1 visualizes the diverse nature of the face masks, while Fig. 2 shows some sample images available in the proposed dataset. This way, our dataset ensures that trained face mask detection models are capable of detecting faces from different ethnicities and age groups with face masks of not only white and blue, as typically used in previous years, but of different colors and shapes. The MAFA dataset also contains many incorrect annotations, as shown in [21]. Therefore, modifying the MAFA dataset to create a new face mask detection dataset requires fixing the incorrect annotations.

One of the initial works on face mask datasets was presented in [22] which proposed three different datasets for masked face recognition and face mask detection. The authors propose a Masked-Face Detection Dataset (MFDD), which is the extended version of the MAFA

Table 1 We compare different face mask detection datasets which contain bounding box annotations for the detection task

Dataset name	#Mask	#No Mask	#Images	Mask type	Image source	Ethnicity
MAFA [3]	35,806	911	30,811	Real	Google + Bing	Asian
FMD [23]	3,232	840	853	Real	Unknown	Asian
MFDD [22]	24,771	Unkown	4,343	Real	[24] + Internet	Unknown
FMLD* [21]	29,532	33,540	41,934	Real	MAFA + WIDER	Asian + Caucasian
MMD [25]	6,758	2,309	6,024	Real	Internet	Various
MaskedFace-Net [20]	67,049	66,734	133,783	Artificial	FFHQ [26]	Various
ISL-UFMD [27]	10,698	10,618	21,816	Real	Internet	Various
BAFMD (ours)	13,492	3,118	6,264	Real	Twitter	Various

(*) symbol indicates that the corresponding dataset only proposes annotations for existing datasets

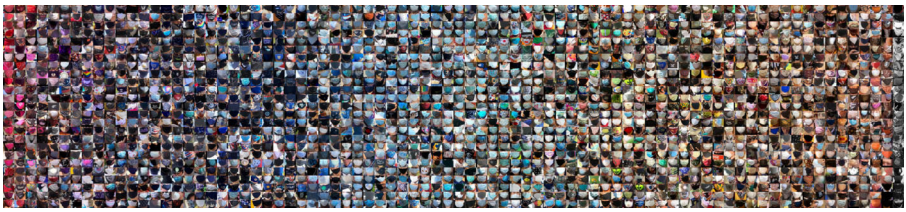


Fig. 1 [Best viewed in color]. Example face mask images available in the proposed dataset. Unlike simulated or pre-pandemic datasets, various colors and textures of the face masks are present

and WIDER datasets for face mask detection. They also propose the Real-world Masked-Face Recognition Dataset (RMFRD) and the Simulated Masked-Face Recognition Dataset (SMFRD) for masked-face recognition. Unfortunately, only a subset of these datasets are publicly available. Furthermore, training models with simulated images can be problematic due to the high domain difference between real and artificial masks.

Another dataset that filters previously proposed datasets to create a more refined one is proposed in [21]. Authors annotate the MAFA [3] and WIDER [9] datasets in the context of the COVID-19 pandemic and with respect to placement-correctness of face mask, gender, ethnicity, and pose. Their annotations show that the MAFA dataset contains mostly Asian and the WIDER dataset contains mostly Caucasian faces. This is problematic because the trained models might associate mask usage with races, as MAFA contains masked faces and WIDER mainly contains faces without masks.

Face Mask Detection (FMD) dataset [23] is proposed for a Kaggle competition during the pandemic. Images were collected from the Internet. They are annotated for three classes: with mask, without mask, and mask worn incorrectly. Medical Mask Detection (MMD) dataset [25] has been acquired from the Internet with attention to the diversity of ethnicities, ages, and regions. All images have been manually curated and annotated. It covers 20 classes of different accessories including faces with a mask, without a mask, or with an incorrectly worn mask.

MaskedFace-Net dataset [20] is an artificially created dataset using a deformable mask model and facial landmarks, similar to SMFRD [22]. Face images are collected from Flickr-Faces-HQ [26] (FFHQ) dataset. Then, digitally created mask models are placed on the mouth area of the given face images and annotated according to the correct mask usage.

Fig. 2 Example images from the Bias-Aware Face Mask Detection (BAFMD) dataset

Finally, more recently, researchers collected images from publicly available face datasets (i.e., FFHQ [26], CelebA [28], LFW [29]), YouTube, and web crawling from websites to create Interactive Systems Labs Unconstrained Face Mask Dataset (ISL-UFMD) [27]. Having diverse and multiple sources of images naturally increases the variability of ethnicity, age, and gender within the dataset. Unlike ISL-UFMD, we quantitatively measure specific attributes of the faces to increase the diversity and reduce possible biases in our dataset in a systematic manner.

3 Proposed dataset

Race and gender biases are well-known but an understudied topic for the face mask detection task. Our primary focus is to gather images that are as representative as possible to reduce dataset bias for a specific ethnicity, age, or gender. To this end, we first collected publicly posted images from Twitter by using keywords related to COVID-19 prevention measures and face masks during the pandemic. Tweet collection was initially restricted to Los Angeles County as it is the second most diverse place in the United States according to the Racial and Ethnic Diversity Index of Census Bureau [30]. Therefore, it is a suitable location to obtain a diverse collection. We ran a state-of-the-art (SOTA) face detector [19] to eliminate images without faces. Then, we manually labeled faces with and without masks by annotating facial bounding box locations and mask usage. We used LabelImg [31] labeling tool for annotations. By using this manually labeled data, we trained YOLO-v5 [32], which is a SOTA object detection model. The trained YOLO-v5 model is utilized to speed up our data annotation process by adopting a semi-automatic label annotation pipeline to estimate candidate bounding boxes and class labels.

For developing a representative and demographically balanced face mask dataset, having a balanced ratio of different faces is important. In our image collection, we created race, age group, and gender predictions of people. We employed FairFace [10], which is a SOTA face attribute classifier trained on a balanced race and gender face attribute dataset. FairFace requires MTCNN [33] face detector due to its training pipeline. Therefore, we only produced predictions for faces that could be detected with MTCNN [33]. FairFace defines seven race groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. It also has an option to define five race groups by combining Middle Eastern and White, as well as East Asian and Southeast Asian. Aside from race predictions, we also used gender and age group predictions. FairFace [10] uses the following age groups: 0-2, 3-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, and 70+. To balance the racial distribution and get more images from underrepresented ethnicities, we expanded our location filter to include images from 56 different countries, such as Kenya, Canada, Vietnam, and Turkey. The final dataset comprises 6,264 images, which contain 13,492 faces with masks and 3,118 faces without masks. Unlike most previous face mask detection datasets, which contain only one face per image, the high number of faces indicates that our dataset also contains crowded scenes. Moreover, our dataset captures high pose and illumination variations. Figure 2 shows images from our dataset, which we named as Bias-Aware Face Mask Detection (BAFMD) dataset. As our dataset is collected from public social media data, some of the data might not be usable depending on the user. The dataset is provided under the MIT License, and users must accept the terms of usage before acquisition[2]. Our dataset mainly contains images of people taking and sharing photographs on Twitter. It also contains images taken by the media

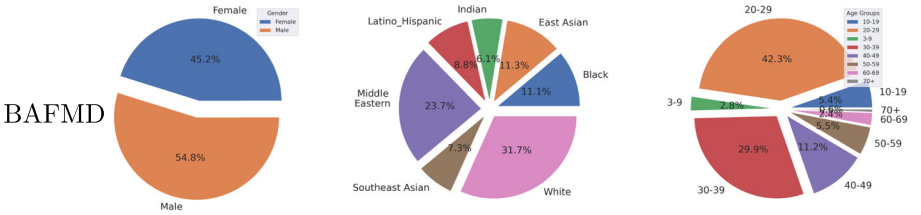
[2] BAFMD terms of usage https://bit.ly/BAFMD-terms-of-usage

with professional cameras. Therefore, it includes high-resolution and low-resolution images with challenging and unrestricted environments. Table 1 shows the difference between our proposed dataset and the other datasets. We compare our dataset with the MAFA [3] dataset, a well-known and widely used face mask detection dataset. Specifically, we compare the ratios of race, gender, and age groups. Having a more balanced dataset in terms of race, gender, and age groups creates less bias for the trained models [10]. As illustrated in Fig. 3, our dataset achieves more balanced ratios across race, gender, and age groups.

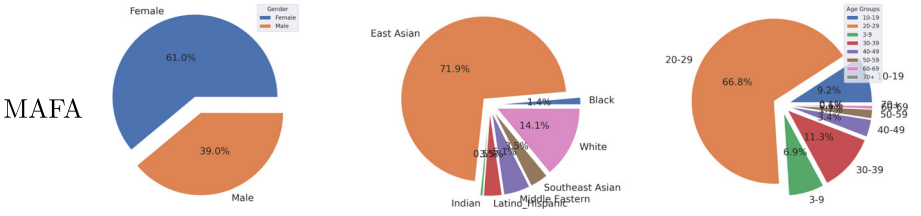
For reproducibility, we define training and testing sets of the dataset[1]. To create a test set, we used the statistics given race predictions of the FairFace model. We kept the test set proportional to the racial, gender, and age group ratios. We used 25% of the faces as the testing set. In the end, we got 5,466 training images and 798 testing images with a similar racial distribution. As stated above, FairFace [10] requires face images cropped by MTCNN [33]. Therefore, in order to produce reliable predictions from FairFace, we use MTCNN on our dataset to crop images. As MTCNN cannot detect all masked faces, only a subset of the dataset can be used for facial attribute prediction. We assume this subset would be sufficient to give information about the entire dataset.

4 Masked face detection

In this work, we use a state-of-the-art object detection architecture, YOLO-v5 [32], for training a facial mask detection model. Multiple research analyzed the performance of different



(a) FairFace Gender Predictions (b) FairFace Race Predictions (c) FairFace Age Predictions



(d) FairFace Gender Predictions (e) FairFace Race Predictions (f) FairFace Age Predictions

Fig. 3 FairFace analysis tool pipeline has been executed over all images of the BAFMD and MAFA datasets. FairFace gender, race and age group predictions for TFMD dataset are presented in a, b, c, respectively. Similarly, for MAFA dataset gender, race and age group predictions are presented in d, e, f, respectively

single- and two-stage object detection models on face mask detection datasets. Moreover, many studies in the field investigated face detection and classification using two separate networks. In this work, we compare six face and face mask detection models. We propose to use YOLO-v5 model as a face mask detector and compare the YOLO-v5 model with five different state-of-the-art face and face mask detectors. YOLO-v5 model is an extension to YOLO-v3 [16] model. YOLO object detectors divide images into a grid system. Each cell in the grid is responsible for detecting objects within itself. A single forward pass of the model yields multiple bounding boxes and their class prediction probabilities. Therefore, they provide faster and better object detection results compared to the other object detectors. YOLO-v5 contains multiple new features over YOLO-v3, such as Path aggregation network [34] and Cross Stage Partial Network [35].

In classical object detection, millions of images are annotated; therefore, bigger models like YOLO-v5 Extra Large can be trained. We train the YOLO-v5 Small model due to limited number of images in face mask detection datasets, and initialize training with the pretrained model weights. For each experiment, we start with a learning rate of 0.001 and use the learning rate scheduler of YOLO-v5. We train each model up to 450 epochs with an early stopping criterion to avoid overfitting.

For comparing YOLO-v5 model with the state-of-the-art face mask detectors, we use MTCNN [33], Baidu [36], AIZooTech [24], RetinaFace [19], YoloV4-P6-Facemask [37], ICE-YoloX [38], and AntiCov [39]. For all of the detectors, we use the default hyperparameters proposed in their paper or code. MTCNN [33] is one of the most popular and successful face detector which consists of three cascaded neural networks. Baidu [36] detector is based on PyramidBox [36] single-shot face detector. PyramidBox [36] implements several strategies to use context information to improve the face detection results. AIZooTech [24] is one of the first proposed face mask detection networks. It is a single-shot detector customized for the face mask detection problem. RetinaFace [19] is a single-shot multi-level face localization method that performs pixel-wise face localization. We use the RetinaFace model with ResNet-50 [40] backbone network. The AntiCov [39] is a customized one-stage face detector based on RetinaFace [19]. The AntiCov is much faster and lighter than RetinaFace in order to deploy the model on end devices with limited computation power. YoloV4-P6-Facemask model [37] is based on a Scaled-YOLOv4 model, which tries to optimize the scaling of backbone for the given task. ICE-YoloX [38] is a network that combines YoloX[41] with a proposed channel-enhanced feature pyramid network (CE-FPN) and specifically designed to address limitations in YoloX networks. The main aim of ICE-YoloX is to reduce the computational overhead while keeping the performance of the YoloX backbone the same.

5 Experiments

In this section, we first present the metrics used to assess the performance of the face mask detection methods. Then, we conduct experiments to evaluate the performance of different face mask detection methods on widely used face mask detection datasets and our BAFMD dataset. Additionally, we test different face mask detection methods on different datasets while changing the training dataset to assess the representativeness, i.e. generalization capability, of the training datasets. Finally, we consider the risks of using social media images where the contents can be removed in time. To observe the effect of this phenomenon, we test the performance of our model with respective to the changing number of training images. In all experiments, we use standard object detection performance metrics, mean average preci-

sion (mAP) and mean average recall (mAR), which has been proposed in [42] and adopted with different object detection benchmarks [43]. Calculation of the mAP requires the computation of the Intersection over Union (IoU) for each class.We calculate IoU by using area of our prediction (P) and area of ground truth (G) bounding box for an object. Following the many object detection research and competitions, we consider a prediction as *True Positive* (TP) if its IoU score is greater than 0.5. Moreover, in order to investigate the localization of the detectors in a more challenging situation, we report the average mAP result of detectors where the IoU threshold is between 0.5 and 0.95 with 0.05 increments. For mAR we average recall over a range of IoU thresholds from 0.5 to 1.0.

5.1 Same-dataset experiments

A considerable amount of face mask detection models are trained with a combination of MAFA and WIDER datasets because they contain a high number of images and were readily available at the start of the COVID-19 pandemic. However, as the MAFA dataset included some noisy annotations, combining the MAFA and WIDER dataset required more work. In FMLD dataset [21], the authors proposed a combination of MAFA and WIDER dataset by annotating both datasets manually. Therefore, they created a better dataset for training face mask detectors. In order to be comparable with previous work, we use MAFA, WIDER, FMLD and our Bias-Aware Face Mask Detection (BAFMD) dataset. In our experiments, we compare our model against MTCNN [33], Baidu [36], AIZooTech [24], RetinaFace-AntiCov [39], YOLOv4-P6-FaceMask [37], and ICE-YoloX [38]. As explained in Sections 2 and 4, these models and datasets are widely used for face mask detection.

We train all the face detectors using the available data and classes in the datasets, in order to test well-known face mask detectors and our proposal of using YOLO-v5 [32]. For all the training, default parameters given by the authors of the models are used and all of them has been trained until they converge. We trained and tested each method on official training and testing sets of the datasets.

The results in Table 2 show that in the MAFA dataset most of the proposed detectors achieve 80 to 85 $mAP_{0.5}\%$. However, in the WIDER dataset, the performances range from

Table 2 Mean average precision for IoU 0.5 ($mAP_{0.5}\%$), mean mAP from IoU 0.5 to 0.95 ($mAP_{0.5:.05:0.95}\%$) and mean average recall from IoU 0.5 to 1.0 results of different face detection models on MAFA [3], WIDER [9], FMLD [21] and BAFMD datasets

Method	Dataset			
	MAFA	WIDER	FMLD	BAFMD
MTCNN [33]	42.5 / 21.8 / 23.7	85.6 / 66.4 / 68.0	65.8 / 42.5 / 44.1	34.3 / 17.9 / 18.6
Baidu [36]	59.4 / 38.7 / 36.3	88.5 / 69.1 / 67.4	77.2 / 57.5 / 56.1	58.7 / 37.4 / 35.8
AIZooTech [24]	85.1 / 66.7 / 68.5	89.3 / 68.9 / 70.0	86.5 / 66.2 / 67.5	76.4 / 57.3 / 58.0
RetinaFace [19]	81.2 / 62.7 / 64.2	**99.4 / 78.8 / 79.0**	91.9 / 70.8 / 69.4	73.6 / 53.6 / 55.8
AntiCov [39]	84.9 / 65.0 / 66.9	93.7 / 72.9 / 73.5	87.8 / 68.1 / 69.0	78.1 / 66.7 / 67.1
YoloV4 [37]	83.4 / 65.4 / 63.0	99.3 / 78.0 / 76.5	92.0/ 70.9 / 72.0	79.3 / 65.4 / 66.0
ICE-YoloX [38]	85.9 / 66.3 / 66.4	91.8 / 74.2 / 75.0	90.6 / 69.2 / 70.3	82.4 / 63.8 / 64.4
Ours	**87.3 / 67.1 / 69.5**	92.0 / 74.2 / 75.3	**92.2 / 71.7 / 73.0**	**86.8 / 67.6 / 68.2**

WIDER face is a well known face detection dataset and does not include any mask annotation. Other datasets contain both mask and no mask classes. Please note that FMLD [21] dataset is combination of MAFA [3] and WIDER [9] datasets. Model that performed best on each dataset is highlighted in bold

85 to 99 $mAP_{0.5}\%$. This is an indication of the difficulty of the face mask detection problem. As the FMLD dataset is a combination of both MAFA and WIDER, the performances of models on this dataset are higher than MAFA but lower than WIDER. In our proposed dataset, the performances of different models are slightly worse than the MAFA dataset, which implies the difficulty of the dataset. Our proposed YOLO-v5 model outperforms other detectors on three out of four datasets. Moreover, the performance of the YOLO-v5 model is more stable across different datasets than other detectors.

5.2 Cross-dataset experiments

Many widely used, publicly available face mask detection datasets are racially imbalanced and contain images from specific regions of the world, such as Asia. In order to create a better and more representative dataset, we collected images from all around the world while keeping a balanced racial distribution. To test the representativeness of the datasets, we train RetinaFace [19] and our proposed method of using YOLO-v5 on FMLD and BAFMD datasets separately. We chose the FMLD dataset as it combines MAFA and WIDER face datasets which are among the popular datasets on face detection and face mask detection. For both datasets we use their official training and testing sets. We used the same hyperparameters as in the within-dataset experiments. Table 3 shows that the performance of both RetinaFace and our model decreases when trained on one dataset and tested on another. When models are trained on FMLD and tested on BAFMD, the drop in $mAP_{0.5}\%$ is nearly 15%. On the other hand, when models are trained on BAFMD and tested on FMLD, the drop in $mAP_{0.5}\%$ is nearly 7%. This trend is also the same in $mAP_{0.5:.05:0.95}\%$. In this challenging metric, the performance is nearly 20% lower, but the performance difference between each train-test pair is similar to the $mAP_{0.5}\%$. This experiment shows that a more representative and racially balanced dataset, such as BAFMD, can lead to better generalization. Therefore, using BAFMD may serve as a better training set for general face mask detectors. Apart from the better performance, training with a balanced dataset enables models to have less accuracy discrepancy among all race and gender groups, as shown in FairFace study [10].

Table 3 RetinaFace [19] and our method have been trained on both FMLD [21] and BAFMD datasets to assess their performance on a dataset that have not been trained

Method	Training set	Test set	$mAP_{0.5}\%$	$mAP_{0.5:.05:0.95}\%$	$mAR\%$
RetinaFace	BAFMD	BAFMD	73.6	52.7	55.8
RetinaFace	FMLD	FMLD	91.9	70.6	69.4
RetinaFace	BAFMD	FMLD	84.0	66.2	68.5
RetinaFace	FMLD	BAFMD	60.2	42.7	41.0
Ours	BAFMD	BAFMD	86.9	67.6	68.2
Ours	FMLD	FMLD	92.2	71.7	73.0
Ours	BAFMD	FMLD	84.5	63.8	65.6
Ours	FMLD	BAFMD	72.9	53.0	52.7

First four rows show the performance of RetinaFace [19] model when trained and tested on different sets. On the other hand last four rows show the performance of our model in the same settings. We also show the same-dataset test performances to highlight the performance drop on cross-dataset tests

5.3 Robustness to volatile social media data

Every day, social media users share thousands of photos to express their ideas or show what is happening around them. In many social media platforms users can control with whom to share their content. For example, a user can share their photo publicly and then make it private so that only the people that they allow can see. Moreover, users can delete or edit their shared content anytime. Therefore, social media content constantly changes and the acquisition and processing of this content should also adapt to this changing environment. In order to adapt these changes, methods similar to [44] can be utilized on top the face detection methods. As our proposed dataset contains images from Twitter, we can not expect to retrieve the entire dataset completely as time passes and the number of samples that can be accessed through the shared links is likely to decrease by time.

In order to assess the performance of our models against the removal of data in time, we trained different face mask detection models using fractions of the same training and validation sets of our BAFMD dataset. Six experiments were held by using 30%, 40%, 50%, 60%, 80%, and 100% of all training and validation samples, while the test set is kept fixed to be able to assess the performance fairly. The removed samples were chosen randomly in order to maintain a consistent distribution across different splits. This experimental setup indicates the potential performance drop for the researchers who would like to develop a face mask detection system using the BAFMD dataset.

For face mask detection, we used our proposed YOLO-v5 [32] model. In order to make the comparisons fair, we used the same hyperparameters for all the trainings. In Table 4, we show the performance of our models with respect to different amount of training data. When all of the available data is used for the training 86.9% $mAP_{0.5}$ is achieved. Removing 10% of the training images drops the performance by 2% to 3% in terms of $mAP_{0.5}$. Therefore, the results indicate that a small percentage of the dataset can still provide a sufficient amount of information to train a successful face mask detector. The results also show sensitivity to the number of training data for face mask detection models.

6 Conclusions

We studied the problem of face mask detection during the COVID-19 pandemic with a particular focus on dataset bias. The face mask detection problem has been an understudied

Table 4 For each training we keep randomly selected images of training and validation sets

Percentage of images	$mAP_{0.5}\%$	$mAP_{0.5:.05:0.95}\%$	$mAR\%$
100%	**86.88**	**67.62**	**68.20**
80%	84.12	60.55	61.25
60%	82.46	57.53	58.70
50%	81.52	55.89	57.05
40%	80.75	55.48	56.20
30%	79.20	53.56	54.10

First column shows percentage of images that has been kept for training to the original size of the dataset. *For each training we present* $mAP_{0.5}\%$, $mAP_{0.5:.05:0.95}\%$ *and* $mAR\%$ *results.* $mAP_{0.5}$ *shows the mAP when threshold is 0.5 for the IoU. For mAR the recall values are averaged for all recall values from 0.5 to 1.0 IoU*

sub-problem of face and object detection. In order to help society during the COVID-19 pandemic, many researchers across the world rapidly focused on the problem. However, majority of the earlier work has simply focused on training new architectures with the limited number of face occlusion datasets.

In this work, we introduced a novel face mask detection dataset named as Bias-Aware Face Mask Detection (BAFMD) dataset. To the best of our knowledge, it is the first face mask detection dataset that has been collected with a focus on mitigating demographic bias. Unlike most publicly available datasets, our dataset contains real-world face mask images with a more balanced distribution across different demographics, e.g., gender, race and age.

Moreover, our experimental results on multiple publicly available datasets show that the proposed YOLO-v5 model has comparable or superior performance to the proposed methods for face mask detection. On our proposed dataset, model performance slightly dips, suggesting its difficulty. When training and testing face detection models on different datasets, there's a significant performance drop. Which shows the difficulty of generalization in the face mask detection problem. A more balanced dataset, BAFMD, can lead to better model generalization and reduce racial and gender performance disparities. Additionally, our studies confirm that using a smaller portion (like 90%) of the training data only reduces performance slightly, revealing the sensitivity of face mask detection models to the volume of training data.

Funding Open Access funding provided by the Qatar National Library. The publication of this article was funded by Qatar National Library.

Availability of Data Collected data is available for non-commercial research upon request. Details of data request can be found in https://github.com/alpkant/bafmd

Declarations

Conflicts of Interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical Approval This paper does not contain any studies with human participants or animals performed by any of the authors.

References

1. Geneva: World Health Organization (2020) WHO COVID-19 Dashboard. (Online; Accessed on 30 Mar 2021) . https://covid19.who.int/
2. Güner HR, Hasanoğlu İ, Aktaş F (2020) COVID-19: Prevention and control measures in community. Turk J Med Sci 50(SI–1):571–577
3. Ge S, Li J, Ye Q, Luo Z (2017) Detecting masked faces in the wild with lle-cnns. In: Proc. of CVPR. pp 2682–2690
4. Zhang T, Li J, Jia W, Sun J, Yang H (2018) Fast and robust occluded face detection in ATM surveillance. Pattern Recogn Lett 107:33–40

5. Chen Y, Song L, Hu Y, He R (2018) Adversarial occlusion-aware face detection. In: 2018 IEEE 9th International conference on biometrics theory, applications and systems
6. Wang J, Yuan Y, Yu G (2017) Face attention network: an effective face detector for the occluded faces. arXiv:1711.07246
7. LeDuc JW, Barry MA (2004) SARS, the first pandemic of the 21st century. Emerg Infect Dis 10(11):26
8. Bhambani K, Jain T, Sultanpure KA (2020) Real-time face mask and social distancing violation detection system using YOLO. In: IEEE Bangalore humanitarian technology conference
9. Yang S, Luo P, Loy C-C, Tang X (2016) Wider face: a face detection benchmark. In: Proc. of CVPR
10. Karkkainen K, Joo J (2021) FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proc. of WACV
11. Terhörst P et al (2022) A comprehensive study on face recognition biases beyond demographics. IEEE Trans Technol Soc 3(1):16–30
12. Merler M, Ratha N, Feris RS, Smith JR (2019) Diversity in faces. arXiv:1901.10436
13. Burgos-Artizzu XP, Perona P, Dollár P (2013) Robust face landmark estimation under occlusion. In: Proc. of ICCV
14. Xia Y, Zhang B, Coenen F (2016) Face occlusion detection using deep convolutional neural networks. Int J Pattern Recognit Artif Intell 30(09):1660010
15. Kumar A, Kumar M, Kaur A (2021) Face detection in still images under occlusion and non-uniform illumination. Multim Tools Appl 80(10):14565–14590
16. Farhadi A, Redmon J (2018) Yolov3: an incremental improvement. In: Computer vision and pattern recognition
17. Tan M, Pang R, Le QV (2020) Efficientdet: scalable and efficient object detection. In: Proc. of CVPR
18. Jiang H, Learned-Miller E (2017) Face detection with the faster R-CNN. In: 2017 12th IEEE International conference on automatic face & gesture recognition (FG 2017). IEEE, pp 650–657
19. Deng J, Guo J, Ververas E, Kotsia I, Zafeiriou S (2020) Retinaface: Single-shot multi-level face localisation in the wild. In: Proc of CVPR pp 5203–5212
20. Cabani A, Hammoudi K, Benhabiles H, Melkemi M (2021) MaskedFace-Net-A dataset of correctly/incorrectly masked face images in the context of COVID-19. Smart health 19
21. Batagelj B, Peer P, Štruc V, Dobrišek S (2021) How to correctly detect face-masks for COVID-19 from visual information? Appl Sci 11(5):2070
22. Wang Z eta al (2020) Masked face recognition dataset and application
23. Larxel (2020) Face mask detection. Kaggle
24. Chiang D (2020) Face mask detection. GitHub
25. Humans in the Loop: Medical mask dataset. (2020). https://humansintheloop.org/resources/datasets/medical-mask-dataset/
26. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proc of CVPR pp 4401–4410
27. Eyiokur FI, Ekenel HK, Waibel A (2022) Unconstrained face mask and face-hand interaction datasets: building a computer vision system to help prevent the transmission of COVID-19. Signal, image and video processing. pp 1–8
28. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proc of ICCV
29. Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report 07-49, University of Massachusetts, Amherst
30. Census Bureau (2021) Racial and ethnic diversity in the United States: 2010 Census and 2020 Census. Accessed 29 Aug 2021. https://bit.ly/Census-Bureau-Racial
31. Labs H (2022) LabelImg. GitHub
32. Jocher G et al (2021) ultralytics/yolov5: v5.0 - YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations. Zenodo. https://doi.org/10.5281/zenodo.4679653
33. Xiang J, Zhu G (2017) Joint face detection and facial expression recognition with MTCNN. In: International conference on information science and control engineering
34. Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proc. of CVPR
35. Wang C-Y, et al (2020) CSPNet: a new backbone that can enhance learning capability of CNN. In: Proc of CVPRW
36. Tang X, Du DK, He Z, Liu J (2018) Pyramidbox: a context-assisted single shot face detector. In: Proc of ECCV
37. Mokeddem ML, Belahcene M, Bourennane S (2023) Covid-19 risk reduce based yolov4-p6-facemask detector and deepsort tracker. Multimed Tools Appl 82(15):23569–23593

38. Chen J, Zhang X, Tang Y, Yu H (2023) Ice-yolox: research on face mask detection algorithm based on improved yolox network. J Supercomput 1–22
39. Deepinsight (2020) RetinaFace anti cov face detector. GitHub
40. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proc. of CVPR
41. Ge Z, Liu S, Wang F, Li Z, Sun J (2021) Yolox: Exceeding yolo series in 2021. arXiv:2107.08430
42. Everingham M et al (2015) The pascal visual object classes challenge: a retrospective. Int J Comput Vision 111(1):98–136
43. Lin T-Y et al (2014) Microsoft coco: common objects in context. In: Proc of ECCV. Springer
44. Yan C, Zhang Y, Zhang Q, Yang Y, Jiang X, Yang Y, Wang B (2022) Privacy-preserving online automl for domain-specific face detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp 4134–4144