Scientia Iranica D (2023) 30(3), 1058–1067

SCIENTIA IRANICA

Sharif University of Technology
Scientia Iranica
Transactions D: Computer Science & Engineering and Electrical Engineering
http://scientiairanica.sharif.edu

Wearing face mask detection using deep learning during COVID-19 pandemic

J. Khoramdel, S. Hatami, and M. Sadedel*

Faculty of Mechanical Engineering, Tarbiat Modares University, Tehran, Iran.

Received 19 September 2021; received in revised form 6 October 2022; accepted 13 February 2023

KEYWORDS
COVID-19;
Deep learning;
Object detection;
Face mask;
Convolutional neural networks.

Abstract. During the COVID-19 pandemic, wearing a face mask has been an effective way to prevent the spread of COVID-19. In a number of monitoring jobs, human workforce has been replaced with computers thanks to the outstanding performance of deep learning models. Monitoring the wearing of a face mask is another task that can be done by deep learning models with acceptable accuracy. The main challenge of this task, however, is the limited amount of data because of the quarantine constraints. This study investigated the capability of three state-of-the-art object detection neural networks to detect face mask for real-world applications. To this end, three models were employed: Single Shot Detector (SSD) and two versions of You Only Look Once (YOLO) including YOLOv4-tiny and YOLOv4-tiny-3l, among which the best was selected. In the proposed method, according to the performance of these models, the most viable model for real-world and mobile device applications, compared to other recent studies, was the YOLOv4-tiny model with the mean Average Precision (mAP) and Frames Per Second (FPS) of 85.31% and 50.66, respectively. These acceptable values were obtained using two datasets with only 1531 images in three separate classes.

1. Introduction

More than two years have passed since the emergence and prevalence of COVID-19 pandemic (known as the coronavirus) and it is still at its peak in some regions of the world. The pandemic rage continues to critically affects the situations in many countries. According to the instructions given by World Health Organization (WHO) [1], people are required to wear masks in all crowded and even secluded environments when in contact with others. While a majority of people follow this law satisfactorily, some others do not abide and put the lives of their own and other at risk. For this reason, many governments all around the world seriously pursue these violators after identifying them through face mask detection in smart cities [2]. One way to identify these people is to employ manual methods to monitor and control them; however, the application of this surveillance method is quite impossible and if possible, it is time-consuming, especially in crowded places. For this reason, the current research employed an automated method, i.e., use of artificial intelligence and deep learning. In addition, the use of object-based models in deep learning allows distinguishing among people wearing a mask correctly, those who do not wear a mask, and those who wear a mask incorrectly, each of whom can use different types of masks either.

*. *Corresponding author. Tel./Fax: +98 21 82884987*
E-mail addresses: j.khoramdel@modares.ac.ir (J. Khoramdel); soheila.hatami@modares.ac.ir (S. Hatami); majid.sadedel@modares.ac.ir (M. Sadedel)

doi: 10.24200/sci.2023.59093.6057

This method appears to be in compliance with health protocols for any place like hospitals.

The structure of this research is organized as follows. The next section briefly reviews the related studies on face mask detection. Section 3 presents more details of the used dataset and the proposed algorithm, gives further descriptions of the two datasets merged together, and presents more data on the models and networks used this research. Section 4 presents the final results and evaluations. Finally, Section 5 summarizes and concludes the study. The main contributions of the proposed method are listed in the following:

- Given the lack of dataset about the individuals adhering to hijab and other similar veils in the case of Iranian people, the proposed method helps achieve a suitable mean Average Precision (mAP) and Frames Per Second (FPS) using two prepared datasets with less similarity and even with a small number of images;
- In addition, this method does not require preprocessing during training and testing phases and instead, it uses three classes rather than two dissimilar other previous works to better simulate real-word situations;
- Finally, three different models are compared based on Convolutional Neural Network (CNN), the best of which enjoys low computational costs and volume space for storage to obtain a suitable model that is appropriate for mobile device applications.

2. Related works

Given that most parts of the faces are covered, it is quite difficult to distinguish people with mask. In other words, it is not possible to easily recognize the face and then find out how people wear mask (with mask, without mask, and incorrect mask). In fact, in [3], faces were identified using Multi-task Cascaded Convolutional Networks and in the output, five landmarks were marked on the face. Identification of these five landmarks means full recognition of people's faces. The method employed in this study to recognize the faces was practical in some ways, but it unfortunately cannot be easily used to identify faces with masks because many of these landmarks are covered. For example, in [4], the required dataset that included people with correct masks and incorrect masks were prepared using landmarks. Synthetically created, this dataset contained almost the same distribution of both classes. In this regard, the researchers used the method described in [5] as the basis of their work and put a mask on people's faces. They took into account 68 landmarks on a face without a mask and 12 landmarks on the mask alone. In case these 12 landmarks are placed on the landmarks that fit the face, it means that the mask has been worn correctly; otherwise, the mask should cover some specific parts of the face, meaning that the mask has not been worn correctly.

Nieto- Rodriguez et al. [6] proposed a system that triggered an alarm when identifying the personnel in the operating room who do not wear mandatory masks. The system consists of two face and mask detectors that use the tone in the Hue, Saturation, Value (HSV) color space. The proposed system reached a recall above 95% with a false positive below 5% for the detection of faces and surgical masks.

In [7], the authors used a model that consists of two parts. For feature extraction, they used ResNet-50 and for classification, they used three different classifiers including decision trees, Support Vector Machine (SVM), and ensemble algorithm. Also, they worked on three different publicly available datasets and obtained good results over these datasets described in detail, but failed to report any value for FPS to know about its real-time application.

In [8], four popular object detection algorithms, two of which belonging to You Only Look Once (YOLO) family namely YOLOv3 and YOLOv3-tiny, Single Shot Detector (SSD), and Faster Region-Based Convolutional Neural Network (Faster R-CNN) were used to recognize people with and without mask and for this purpose, a specific dataset of their own, i.e., Moxa3k [9] that contains 3000 images from Kaggle dataset of medical mask dataset [10] and others, was collected from relevant websites. It was concluded that YOLOv3-tiny was more suitable to make a good balance between the accuracy and real-world applications, hence a need for obtaining reasonable accuracy for real-time work with CCTV cameras, especially when detecting people in crowded places. Jignesh Chowdary et al. [11] utilized the InceptionV3 pre-trained model by removing its last layer and adding five layers to the fine-tuning the model. They also used Simulated Masked Face Dataset (SMFD) [12] that exhibited a good balance between two classes (simulated masked facial and unmasked facial images) for their work and achieved reasonable results for both testing and training phases.

In some other research works like in [13], some prepared libraries such as TensorFlow, Keras, and OpenCV were used that are simple yet not deep enough to extract features for complicated conditions and make them unsuitable for some real-world situations.

In [14], different models including Faster R-CNN, YOLOv3, YOLOv4, YOLOv5, and YOLOR were used for face mask detection and social distance determination as well. In addition, a dataset consisting of several different video frames with two classes was collected. The ViDMASK included 20,000 instances of people with mask and 2,500 instances without mask, clearly indicating the non-uniform distribution of the samples across different classes. It should be noted that although the number of the images in the ViDMASK is larger than that of Moxa3k, they were captured from videos. Hence, the variation between different frames is not significant.

3. The proposed method

In this section, the used datasets as well as the proposed method are described in detail.

3.1. Dataset

Here, for our purpose, we used two different prepared datasets. Of note, we made some changes to these two datasets to prepare them for our final results. The first dataset [15] contains 853 images in PNG, jpg, and jpeg formats. This dataset has three classes of people with correct mask, without mask, and incorrect mask. It contains people in different situations including a crowded environment where the number of people in the image is high and where there is only one person in the image. The annotation of these images by default are given in an XLM file. The file contains the name of the image folder, name of the image to which the class belongs, width, height, depth (RGB images), name of the image class (with mask, without mask, and incorrect mask), bounding box coordinate that includes $x_{\min}$, $y_{\min}$, $x_{\max}$, and $y_{\max}$ (the first two are the top-right coordinates of the bounding box and the next two are the bottom left coordinates of the bounding box). Therefore, in order to work with YOLO and SSD models, these XML files were converted into the txt files. The next dataset [10] contains 678 images in jpg format. The images in this dataset include three classes of people with correct mask, without mask, and incorrect mask. In addition, the image annotations in this dataset include annotations in the XLM and txt files. The XLM files of these images contain the same data as the previous dataset.

Finally, these selected datasets were merged. Our final dataset contains images in both jpg and PNG formats. Of all the dataset images, 80% are for the train data, and 20% for the test data. Figure 1 contains some samples of the dataset images.

Figure 2 shows the frequency of instances in each class according to which the frequency of the "incorrect mask" class instances is quite low compared to the two other classes. In the presented dataset, the numbers of "with mask", "without mask", and "incorrect mask" classes are 6322, 1377, and 247, respectively.

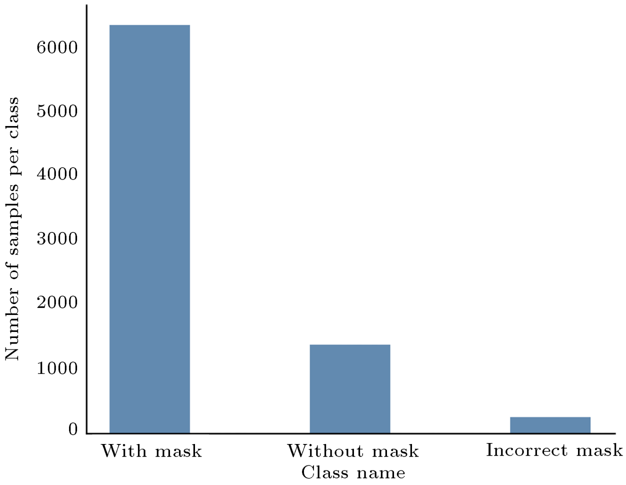


Figure 2. Frequency of each class in the face mask detection dataset.

3.2. Object detection

The primary objective of this paper is to develop a measure to monitor the necessary protocols for COVID-19 or any place that needs to obtain compliance with health protocols. To this end, in the first stage, an object detection algorithm was employed to detect and classify people who wore mask, who did not wear mask, and who wore the mask incorrectly. A number of algorithms have been introduced in recent years for object detection, and considerable progress has been made in this field. Since AlexNet won the ImageNet challenge in 2012 [16], CNNs have gained momentum in computer vision tasks. CNN-based approaches have gained significant superiority to the non-CNN-based algorithms [17] like HoG [18], SIFT [19], Haar feature-based object detection [20], etc. It should be noted that like in many other tasks, a tradeoff between speed and accuracy can also be found here. Some CNNs like Faster R-CNN have achieved high accuracy; however, their low speed at the run time [21] has remained a major drawback which makes it unreliable for real-time applications.

To speed up the object detection task, Redmon et

Figure 1. Some samples of the dataset images.

al. [22] developed YOLO in 2016. Inspired by YOLO, SSD was later developed in 2016 [23]. However, SSD provided higher accuracy and speed than YOLO. Four different versions of YOLO have been released to date, and each version was modified to obtain higher accuracy and run time speed. Each version of YOLO has a lighter version called YOLO-tiny. The YOLO-tiny network uses fewer convolutional layers than YOLO, leading to increase in speed at the expense of accuracy. However, when the dataset does not include enough samples for training, YOLO-tiny may exhibit a better performance than YOLO on that dataset. After SSD succeeded in outperforming YOLO by doing object detection on the multiple scales of the feature map to handle the problem with small size objects, YOLOv3 [24] was presented with this idea to perform object detection on multiple scales to obtain higher accuracy. YOLOv3 used three scales and nine anchor boxes per grid cell for this purpose, while YOLOv3-tiny uses only two scales and six anchor boxes per each grid cell. Adarsh et al. [25] remarked that while the mAP of YOLOv3-tiny was obtained as 33.2% on COCO dataset [26], that of YOLOv3 was obtained as 57.8% on the same dataset. All in all, YOLOv3-tiny can perform object detection 11 times faster than YOLOv3.

When YOLOv4 was proposed by Bochkovskiy et al. [27], two structures pertaining to YOLOv4-tiny were introduced. Same as YOLOv3-tiny, one of these structures uses two scales for object detection and six anchor boxes per grid cell (this version is called YOLOv4-tiny), while the other one uses three scales for object detection and nine bounding boxes per grid cell (this version is called YOLOv4-tiny-3l since it uses three scales for detection). The YOLOv4-tiny-3l is expected to be slower than YOLOv4-tiny and faster than YOLOv4. Since no comparisons can be made between the accuracy and speed of SSD, YOLOv4-tiny, and YOLOv4-tiny-3l, these networks will be trained on the above-mentioned dataset to choose the efficient object detector for this dataset in terms of speed and accuracy. After object detection on a single frame, instances of each class are counted and the distance between the detected people in the image space is calculated.

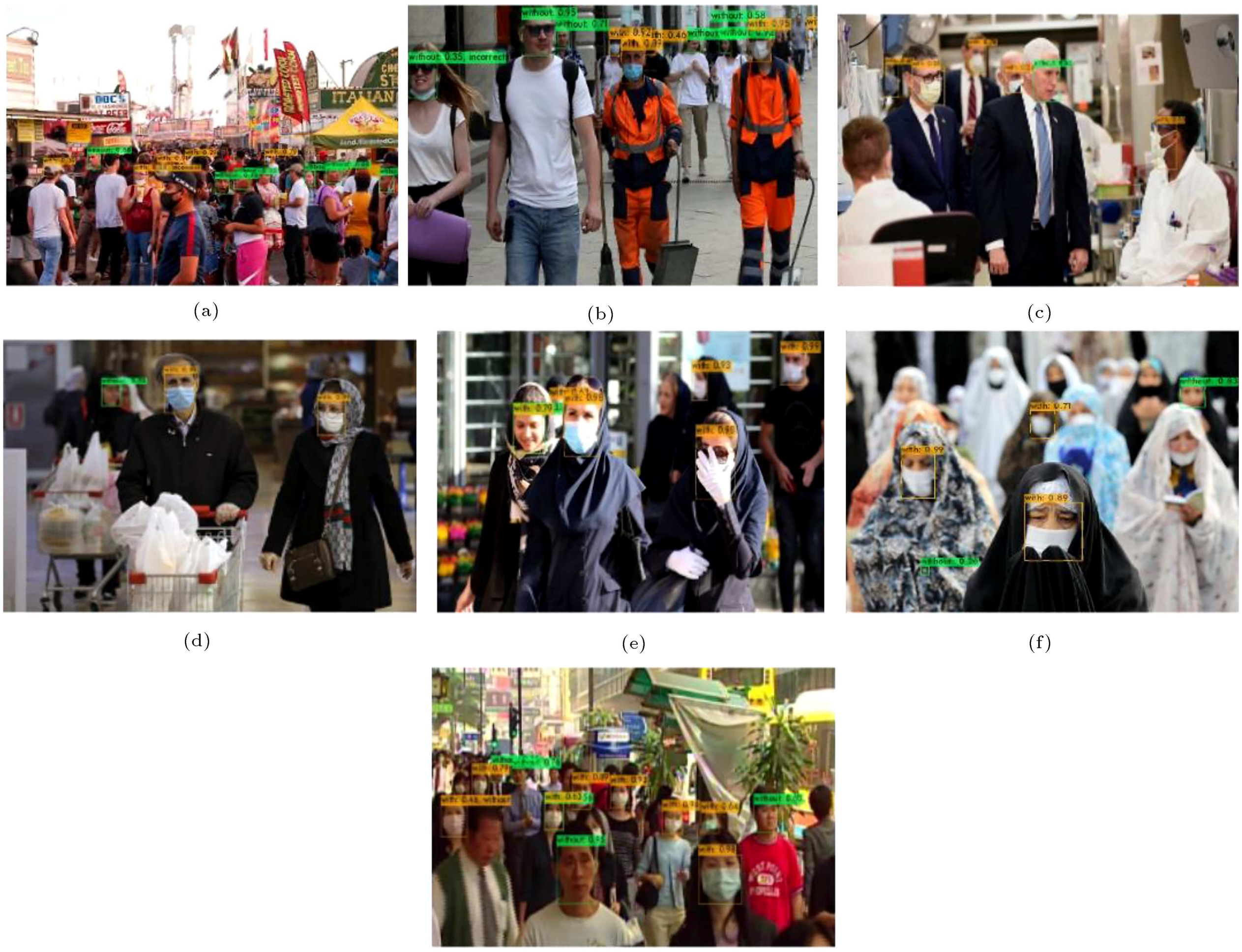
4. Experimental results

The training process was conducted on a device with one Tesla K80 GPU and a single-core hyper-threaded Xeon Processor 2.3 GHz. The resolution of the input image was 416 × 416 for both YOLOv4-tiny networks and 300×300 for the SSD. For the networks to converge faster and gain more generalization power, the pretrained weights on the COCO dataset were used as the initial weights for the training. The dataset for mask detection contained 1540 images, 1232 of which were used for training and 308 for validation. After training, seven images were collected from the internet as the test images. The prediction results for the YOLOv4-tiny-3l, YOLOv4-tiny, and SSD on these images are shown in Figure 3, Figure 4, and Figure 5, respectively. These test images contained both crowded and uncrowded cases. They were taken outdoors (a), (b), (e), and (g) and indoors (c), (d), and (f) under different lighting conditions. As observed in Figure 3(c), there was a person with incorrect mask in the image, but the YOLOv4-tiny-3l failed to identify that person and classified him as a person "without mask". However, Figure 4(c) shows the better performance of YOLOv4-tiny in this case in classifying that person correctly. In Figure 5, the bounding boxes for the classes "with mask", "without mask", and "incorrect mask" are shown with green, red, and yellow colors, respectively. As observed, the SSD model failed to detect any people belonging to the "incorrect mask" class, the same class with the least number of images in the dataset. While the dataset mainly includes the images of people from east Asia (Figure 1) and does not contain images of Iranian people, YOLOv4-tiny and YOLOv4-tiny-3l succeeded in face mask detection among the Iranian people (Figures 3(c), (d), and (e); Figure 4(c), (d), and (e)). The trained models are consequently applicable to surveillance in Iran. Regardless of whether the images are crowded or not, YOLOv4-tiny and YOLOv4-tiny-3l detected the desired classes correctly unless the faces were blurred or very far from the camera.

Table 1 shows Average Precision (AP) of the networks on the validation data. These results were obtained and calculated with an Intersection Over Union

Table 1. Accuracy of the models in the validation data.

	AP		
Class	**YOLOv4-tiny-3l**	**YOLOv4-tiny**	**SSD**
Incorrect mask	77.53%	90.48%	54.72%
With mask	94.87%	90.37%	79.47%
Without mask	78.62%	75.08%	47.20%
mAP	83.67%	85.31%	60.46%



(a) (b) (c)

(d) (e) (f)

(g)

Figure 3. The performance of YOLOv4-tiny-3l in test images.

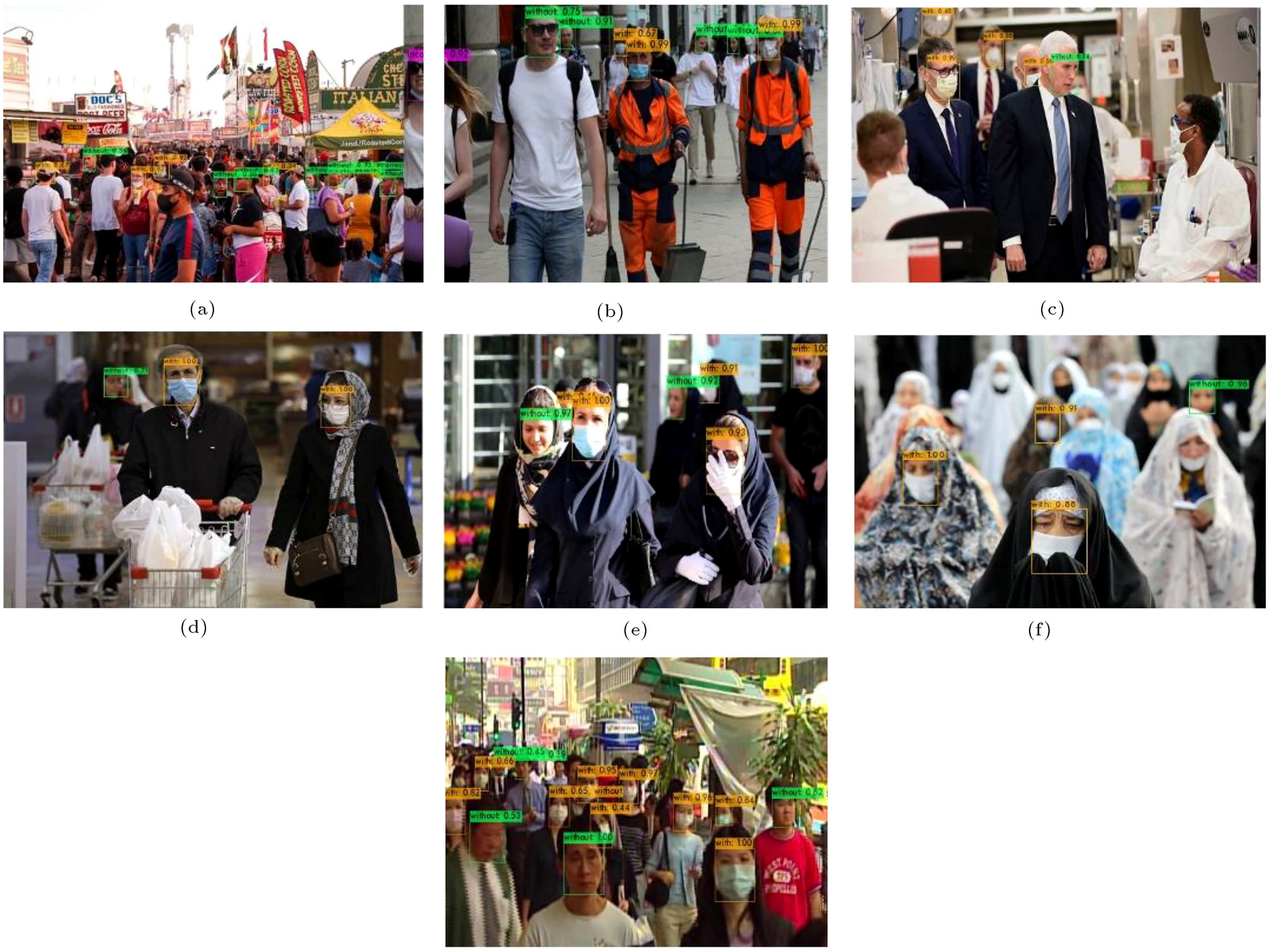
(IOU) threshold of 50%. Although YOLOv4-tiny-3l slightly outperformed YOLOv4-tiny in the two "with mask" and "without mask" classes, YOLOv4-tiny has a significantly better performance on the "incorrect mask" class which led to a better mAP. The YOLOv4-tiny achieved a desirable mAP, indicating that the object detection task on this dataset is applicable. On the contrary, the lower mAP of the two other models (YOLOv4-tiny-3l and SSD), which is heavier than YOLOv4-tiny and is expected to have higher accuracy, showing that the size of the dataset is not suitable for training these heavier models.

Table 2 lists the speed values of these three models at the test phase. As expected, the YOLOv4-tiny model, with the least number of parameters compared to the other two models, achieved the highest FPS. Of note, the FPS of all these three models is good enough for real-time applications. Since many surveillance cameras have a frame rate of 30, the FPS of all three models is good enough for real-time applications. Given that the average human walking speed is in the range of 1.2–1.4 m/s, it is unnecessary to process all the 30 frames taken by the camera and only consider one-third of the frames (10 frames).

Table 2. Accuracy of the models in the validation data.

Model	YOLOv4-tiny-3l	YOLOv4-tiny	SSD
Speed (FPS)	38.3	50.66	27.14

As observed in Figure 2, the "incorrect mask" class instances are much smaller in number than the instances of the "with mask" and "without mask". Hence, there may be no difference between "incorrect mask" and "without mask" classes in terms of the health and safety protocols. However, the visual characteristics of a person with an incorrect mask are more similar to a person with mask rather than a person without mask. To validate this assumption, we merged the "incorrect mask" and "without mask" classes into one class and then, trained the YOLOv4-tiny on "without mask" and "with mask" classes. The results of evaluating the model based on the validation data are given Table 3 according to which the mAP considerably decreased compared to the previous part.



(a) (b) (c)

(d) (e) (f)

(g)

Figure 4. The performance of YOLOv4-tiny in test images.

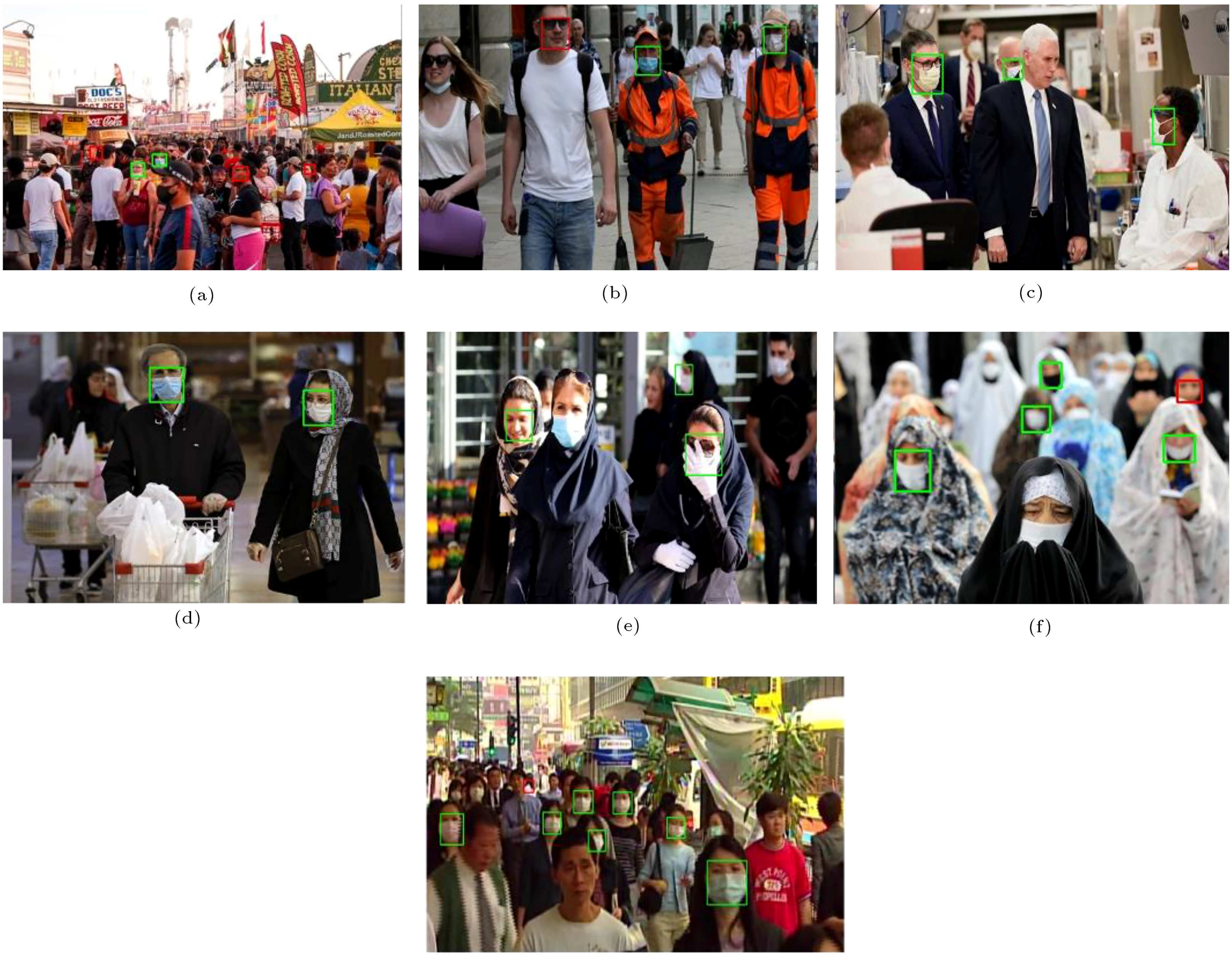
Table 3. Accuracy of the models in the validation data.

	AP
Class	**YOLOv4-tiny-3l**
With mask	89.82%
Without mask	61.33%
mAP	75.57%

Therefore, merging the "incorrect mask" and "without mask" classes would not be a good idea. This finding also implies another benefit of the trained models in this paper in comparison to the works that had considered two classes. Not only was the "incorrect mask" class detected but also the accuracy of the model in two other categories was enhanced. In other words, the trained models with three classes are more robust than their counterparts.

On the same dataset, Roy et al. [8] trained YOLOv3, YOLOv3-tiny, SSD, and Faster R-CNN [8]. The highest reported mAP value for the IOU threshold of 50% is 66.84% for YOLOv3 with the input image resolution of 608 × 608. Although the YOLOv3 model was a more complex model than the YOLOv4-tiny and YOLOv4-tiny-3l models, our trained models exhibited better performance. As mentioned earlier, the resolution of the input images was 416 × 416 that helped reduce the computational cost even more than the resolution of 608 × 608 and improved the speed at the testing phase. Hence, our model proved to be more accurate and faster than the trained model in [8].

Despite the claim made by Jignesh Chowdary et al. [11], regarding face mask detection, their methodology did not involve object detection. In fact, they did face detection at the first phase and then, after identifying the faces in the image, they were cropped and fed into an InceptionV3 model for the image classification task. They, however, did not mention which object detection they used for the first phase. If they were using CNNs for face detection in the first stage, running two heavy models (one for face detection and the other for face mask classification) in the cascade mode would not be efficient at all since it could slow down the prediction, especially in the



(a) (b) (c) (d) (e) (f) (g)

Figure 5. The performance of SSD in test images.

crowded images. Our results show that it is possible for an object detection network to accurately detect faces with mask and without mask. If they used the classic algorithms for face detection like Haar cascade, not only the face detection accuracy would decrease, especially in the crowded areas, but also they could not benefit GPU for parallel computations due to the nature of these algorithms. Consequently, the overall speed and accuracy would be highly affected. Finally, our proposed algorithm is superior to that proposed by Loey et al. [7] in both cases.

It should be noted that the proposed method in [3] is applicable only on small images (they used the resolution of 12×12 for the input images), meaning that their approach does not have the required scalability to detect faces that are far from the camera and it can be only used on the close-up faces in the image. This scalability issue was also raised in the study of Nieto-Rodriguez et al. [6].

As shown in Table 4, the proposed method was compared with a number of others proposed in recent studies. The method used in each of these recent studies is presented with certain advantages and disadvantages and finally their mAP and FPS. In the methodology section, each of the studies used different networks and architectures, and the results showed that all the work that has been done in this field has not changed the architectures or even the networks, except for some parts where only the parameters were changed. In fact, it is clear that in any recent work, as well as in the proposed method, a comparison has been made between the best networks and architectures.

One of the fundamental issues in each of these methods is the use of datasets with a large number and variety of images. Except one case, other cases have used prepared datasets and this is a great advantage for them. For example, there is no data on people wearing hijab and other different head and face coverings that exist in Iran. Thus, for this reason, we were forced to use two publicly available datasets, even though they did not resemble the data belonging to the case of people with hijab. Finally, some datasets used in

Table 4. Comparison of some recent studies with the proposed method.

Study	Method	Advantages	Disadvantages	mAP, FPS
Loey et al. [7]	ResNet50 (feature extractor) + SVM, ensemble, decision tree (classifier, using them as separate classifiers)	Using three prepared different datasets with 10000, 1570, and 13000 images, each	Using preprocessing, high computational cost, using a dataset with only one labeled face per image, not suitable for real-time applications, using a variety of synthetic data for testing, and containing only two classes (mask and no mask),	For SVM (the best) 99.64% (real-world dataset), 99.49% (synthetic dataset), 100% (synthetic dataset), FPS not reported
Roy et al. [8]	SSD, F-RCNN, YOLOv3, YOLOv3-tiny	Preparing and using a dataset with 3000 images	Achieving a lower mAP, using networks with high computational cost and utilizing some models not applicable in real time, and containing only two classes (mask and no mask)	56.27% (for YOLOv3-tiny, real-world dataset), 138
Asghar et al. [28]	(Preprocessing) DS-CNN (Depthwise separable convolutional neural network)	Using three prepared different datasets with 8000 combinations of the first two datasets and 3000 images for the third one	Not suitable for real-time applications (the FPS is not reported), using a variety of synthetic data for testing (8000 images), and containing only two classes (mask and no mask)	92% (synthetic dataset), FPS not reported
Ottakath et al. [14]	Faster R-CNN, YOLOv4, YOLOv5, YOLOR	ViDMASK dataset was collected and different models were trained and evaluated on ViDMASK and Moxa3k	Only two classes were considered; low accuracy was reported on Moxa3k	92.4% (YOLOR) on ViDMASK, 68.2% (YOLOv4-tiny) on Moxa3k, 139 (YOLOv4-tiny), 16.9 (YOLOR)
The proposed method	SSD, YOLOv4-tiny-3l, YOLOv4-tiny	No need for preprocessing, containing three classes (mask, no mask and incorrect mask), low computational cost, containing low space for storage, and suitable for mobile device applications	Dataset with a limited number of images	85.31% (real-world dataset), 50.66

these works contained images of real people's faces and artificially applied masks on their faces so that the synthetic data were prepared or used even for the testing phase.

In some cases, although the method used in [7,28] was discussed as real time, no value was reported for FPS. For this reason, this claim is almost unacceptable. Although high accuracy of 92.4% was reported on ViDMASK with YOLOR model in [14], the highest accuracy on Moxa3k was obtained with Faster R-CNN model with the accuracy rate of 74.7%. After Faster R-CNN which is one of the slowest models (16 FPS), the most accurate model was YOLOv4-tiny (68.22%), which was considerably faster than Faster R-CNN (139 FPS). It should be noted that these accuracy rates were reported in detecting two classes, while the more accurate results were obtained in our study for detecting three and two classes.

Other issues including heavy data preprocessing required by some methods constitute other unresolved drawbacks of previous studies. Data preprocessing is done before the training phase or even in some cases for the test phase. In some methods, networks and architectures with a high computational cost have been used, a flaw from which some of previous methods suffer. Given that the issue of face mask detection is a topic that requires an algorithm with light computational cost (for example, for use in mobile device applications), our proposed method presented an acceptable model that could give the best results with fewer parameters. Finally, the number of classes with real-world varieties has not been addressed sufficiently in the previous methods. The use of three classes is another advantage of the proposed method, which can be applied not only to the case of COVID-19 pandemic, but also to places such as hospitals, personnel who wear the mask incorrectly. Therefore, this method helps health protocols considerably.

5. Conclusions

The main objective of this paper was to evaluate the performance of three recent state-of-the-art object detection neural networks in terms of face mask detection task. According to the observation, despite the limited amount of data for face mask detection, it is still possible to detect and classify people in the images in three classes of "with mask", "without mask", and "incorrect mask". These networks exhibited acceptable performance on the test images even in the crowded areas. In addition, it is possible to use these networks for object detection in real-time scenarios.

Finally, it was concluded that for mobile device applications (with low space for storage) and, of course, for real-world images or video frames, the proposed method even with a small number and variety of data could strike a balance between the mean Average Precision (mAP) and Frames Per Second (FPS).

One of the options that helps improve this method in the future is to use datasets with a wide variety of images and their affinity to the Iranian people (if this method is to be defined, for example, for people with a special type of face). This method can also be used for places where they can be identified even with a mask on their face, for example, identification of people with mask or another cover on their face when entering the workplace. In the end, it is possible to benefit from datasets such as ViDMASK [14] and achieve better results for other environments in this article by balancing its data and using the optimized models.

References

1. "Coronavirus disease (COVID-19): masks", https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-masks (2020).

2. Rahman, M.M., Manik, M.M.H., Islam, M.M., et al. "An automated system to limit COVID-19 using facial mask detection in smart city network", *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp. 1–5 (2020). DOI: 10.1109/IEMTRONICS51293.2020.9216386

3. Zhang, K., Zhang, Z., Li, Z., et al. "Joint face detection and alignment using multitask cascaded convolutional networks", *IEEE Signal Process. Lett.*, **23**, pp. 1499–503 (2016).

4. Cabani, A., Hammoudi, K., Benhabiles, H., et al. "MaskedFace-Net - a dataset of correctly/incorrectly masked face images in the context of COVID-19", *Smart Heal.*, **19**, 100144 (2020).

5. "Dataset of face images flickr-faces-HQ (FFHQ)" (2022), https://github.com/NVlabs/ffhq-dataset.

6. Nieto-Rodriguez, A., Mucientes, M., and Brea, V. "System for medical mask detection in the operating room through facial attributes", *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 138—145 (2015).
DOI: 10.1007/978-3-319-19390-8-16.

7. Loey, M., Manogaran, G., Taha, M., et al. "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic", *Measurement*, **167**, 108288 (2021).

8. Roy, B., Nandy, S., Ghosh, D., et al. "MOXA: A deep learning based unmanned approach for real-time monitoring of people wearing medical masks", *Trans. Indian Natl. Acad. Eng.*, **5**, pp. 509–518 (2020).

9. "Moxa3k dataset", https://shitty-bots-inc.github.io/MOXA/index.html (2020).

10. Waghe, S. "Medical masks dataset", https://www.kaggle.com/shreyashwaghe/medical-mask-dataset (2020).

11. Jignesh Chowdary, G., Punn, N.S., Sonbhadra, S.K., et al. "Face mask detection using transfer learning of Inceptionv3", *Big Data Analytics*, pp. 81–90 (Springer International Publishing, 2020).

12. Prajnasb, "Observations", https://github.com/prajnasb/observations (2020).

13. Das, A., Ansari, M.W., and Basak, R. "COVID-19 face mask detection using tensorflow, keras and opencv", *2020 IEEE 17th India Council International Conference (INDICON)*, pp. 1–5 (2020).
DOI: 10.1109/INDICON49873.2020.9342585.

14. Ottakath, N., Elharrouss, O., Almaadeed, N., et al. "ViDMASK dataset for face mask detection with social distance measurement", *Displays*, **73**, 102235 (2022).

15. Larxel, "Face mask detection", https://www.kaggle.com/andrewmvd/face-mask-detection (2020).

16. Krizhevsky, A., Sutskever, I., and Hinton, G. "ImageNet classification with deep convolutional neural networks", *Neural Inf. Process. Syst.*, **25**, pp. 84–90 (2012).

17. Girshick, R., Donahue, J., Darrell, T., et al. "Region-based convolutional networks for accurate object detection and segmentation", *IEEE Trans. Pattern Anal. Mach. Intell.*, **38**, pp. 142–158 (2016).

18. Dalal, N. and Triggs, B. "Histograms of oriented gradients for human detection", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, **2** (2005).

19. Lowe, D.G. "Object recognition from local scale-invariant features", *Proceedings of the Seventh IEEE International Conference on Computer Vision*, **2**, pp. 1150–1157 (1999).

20. Viola, P. and Jones, M. "Rapid object detection using a boosted cascade of simple features", *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001*, pp. 511–518 (2001).

21. Ren, S., He, K., Girshick, R., et al. "Faster R-CNN: Towards real-time object detection with region proposal networks", *IEEE Trans. Pattern Anal. Mach. Intell.*, **39**, pp. 1137–1149 (2015).

22. Redmon, J., Divvala, S., Girshick, R., et al. "You only look once: unified, real-time object detection", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788 (2016).
DOI: 10.1109/CVPR.2016.91

23. Liu, W., Anguelov, D., Erhan, D., et al. "SSD: single shot multibox detector", *European Conference on Computer Vision*, pp. 21–37 (2016).
DOI: 10.1007/978-3-319-46448-0-2

24. Redmon, J. and Farhadi, A. "YOLOv3: an incremental improvement", *arXiv Prepr. arXiv1804.02767* (2018).

25. Adarsh, P., Rathi, P., and Kumar, M. "YOLOv3-tiny: Object detection and recognition using one stage improved model", *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 687–694 (2020).

26. Lin, T.-Y., Maire, M., Belongie, S., et al. "Microsoft coco: common objects in context", *European Conference on Computer Vision* (eds. Fleet, D., Pajdla, T., Schiele, B. and Tuytelaars, T.), pp. 740–755 (Springer International Publishing (2014).

27. Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y.M. "YOLOv4: Optimal speed and accuracy of object detection", *arXiv Prepr. arXiv2004.10934* (2020).

28. Asghar, M.Z., Albogamy, F.R., Al-Rakhami, M.S., et al. "Facial mask detection using depthwise separable convolutional neural network model during COVID-19 pandemic", *Front. Public Heal.*, **10** (2022).

Biographies

Javad Khoramdel is currently pursuing his MSc degree in Mechatronics at Tarbiat Modares University, Tehran, Iran in 2023. He received his BSc degree in Mechanical Engineering from K. N. Toosi University of Technology, Tehran in 2019. His research interests include computer vision, robotics, and deep learning.

Soheila Hatami received her MSc degree in Mechatronics from Tarbiat Modares University, Tehran, Iran in 2022. She obtained her BSc degree from University of Mohaghegh Ardabili, Ardabil, and Imam Khomeini International University (as a guest student), Qazvin in 2017, all in Electrical Engineering (power branch). Her research interests include deep learning, image processing, industrial automation, and autonomous vehicles.

Majid Sadedel is currently an Assistant Professor at Tarbiat Modares University, Tehran, Iran. He received his PhD degree from Tehran University, Tehran in 2016, MSc degree from Sharif University of Technology, Tehran in 2011, and BSc degree from Amirkabir University of Technology, Tehran in 2009, all in Mechanical Engineering. His research interests include robotics, artificial intelligence, mechatronics, and industrial automation.