*Computers, Materials & Continua*
DOI:10.32604/cmc.2022.017865
***Article***

Tech Science Press

# An Automated Real-Time Face Mask Detection System Using Transfer Learning with Faster-RCNN in the Era of the COVID-19 Pandemic

**Maha Farouk S. Sabir[1], Irfan Mehmood[2,*], Wafaa Adnan Alsaggaf[3], Enas Fawai Khairullah[3], Samar Alhuraiji[4], Ahmed S. Alghamdi[5] and Ahmed A. Abd El-Latif[6]**

[1]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Saudi Arabia
[2]Centre for Visual Computing, Faculty of Engineering and Informatics, University of Bradford, Bradford, U.K
[3]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, P.O. Box 23713, Saudi Arabia
[4]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[5]Department of Cybersecurity, College of Computer Science and Engineering, University of Jeddah, Saudi Arabia
[6]Department of Mathematics and Computer Science, Faculty of Science, Menoufia University, 32511, Egypt
*Corresponding Author: Irfan Mehmood. Email: irfanmehmood@ieee.org

Received: 15 February 2021; Accepted: 07 September 2021

**Abstract:** Today, due to the pandemic of COVID-19 the entire world is facing a serious health crisis. According to the World Health Organization (WHO), people in public places should wear a face mask to control the rapid transmission of COVID-19. The governmental bodies of different countries imposed that wearing a face mask is compulsory in public places. Therefore, it is very difficult to manually monitor people in overcrowded areas. This research focuses on providing a solution to enforce one of the important preventative measures of COVID-19 in public places, by presenting an automated system that automatically localizes masked and unmasked human faces within an image or video of an area which assist in this outbreak of COVID-19. This paper demonstrates a transfer learning approach with the Faster-RCNN model to detect faces that are masked or unmasked. The proposed framework is built by fine-tuning the state-of-the-art deep learning model, Faster-RCNN, and has been validated on a publicly available dataset named Face Mask Dataset (FMD) and achieving the highest average precision (AP) of 81% and highest average Recall (AR) of 84%. This shows the strong robustness and capabilities of the Faster-RCNN model to detect individuals with masked and un-masked faces. Moreover, this work applies to real-time and can be implemented in any public service area.

**Keywords:** COIVD-19; deep learning; faster-RCNN; object detection; transfer learning; face mask

## 1 Introduction

The governmental response of differing nations to control the rapid global spread of COVID-19 was to take necessary preventative measures [1] to avoid a majorly disruptive impact on economic and normal day-to-day activities. In various countries where an increased curve of COVID-19 cases are recorded, a lockdown for several months is implemented as a direct response [2]. To minimize people's exposure to the novel virus, many authorities like the World Health Organization (WHO) have laid down several preventative measures and guidelines, one such being that all citizens in public places should wear a face mask [3,4].

Before the pandemic of COVID-19, only a minority of people used to wear face masks mainly in an attempt to protect themselves from air pollution. Many other health professionals including doctors and nurses also wore face masks during their operational practices. In addition to wearing face masks, social distancing i.e., maintaining a distance of 3 ft from any other individual was suggested [4]. According to WHO, COVID-19 is a global pandemic and throughout the world, there are up to 22 million infected cases. Many positive cases are usually found in crowded places [4]. Due to the pernicious effect COVID-19 has on people [5], it has become a serious health and economic problem worldwide [6]. According to [7], it is observed that in more than 180 countries there are six million infected cases with a death rate of 3%. The reason behind this rapid spread of the disease is due to a lack of rule adherence regarding the preventative measures suggested, especially, in overcrowded, high populace areas. The usage of Personal Protective Equipment (PPE) has also been recommended by WHO. The production of PPE however is very limited in many countries [8]. In addition to COVID-19, another disease which includes Severe Acute Respiratory Syndrome (SARS) and the Middle East Respiratory Syndrome (MERS) are also large-scale respiratory diseases that occurred in recent years [9,10]. It is reported by Liu et al. [11] that exponential growth in COVD-19 cases is more than SARS. Therefore, the top priority of government is public health [12]. So, in order to help the global effort, the detection of face masks is a very crucial task.

Many scientists prescribed that these respiratory diseases can be prevented by wearing face masks [13]. Previous studies also show that the spread of all respiratory diseases can easily be prevented by wearing face masks [14–16]. Fortunately, Leung et al. [17] observed that the use of surgical face masks also prevents the spread of coronavirus. Using N95 and surgical masks in blocking the spread of SARS have an effective rate of about 91% and 68% respectively [18]. So, throughout the world, there are many countries where wearing masks is mandatory by governmental law. Many private organizations also follow the guidelines of wearing masks [19]. Furthermore, many public service providers only provide services to customers if they adhere to the face mask-wearing policy [12]. These rules and laws are imposed by the government in response to the exponential growth and spread of the virus and it is difficult to ensure that people are following rules. There are a lot of challenges and risks faced by different policymakers in controlling the transmission of COVID-19 [20]. To track people who violate the rules, there is a need for the implementation of a robust automated system. In France, the surveillance cameras of the Paris Metro Systems are integrated with new AI software to track the face masks of passengers [21]. Similarly, New software developed by French startup DatakaLab [22] produces statistical data by recognizing the people who are not wearing face masks which helps different authorities in predicting COVID-19's potential outbreaks. This need is also recognized in our research and we have developed an automated system that is well suited to detecting real-time violations of individuals not adhering to mask-wearing policies, in turn, assisting supervisory bodies. As in the era of Artificial Intelligence (AI) [23], there are various Deep learning (DL) [24–28] and Machine learning (ML) techniques [29] that are available to design such systems that prevent the transmission of this novel global pandemic [30]. Many techniques are used to design early prediction systems that can help to forecast the spread of disease. This will allow for the implementation of various controlling and monitoring strategies that are adopted to prevent the further spread of this disease. Many emerging technologies which include the Internet of Things (IoT), AI, DL, and ML are used to diagnose complex diseases and forecasting their early prediction like COVID-19 [31–35]. Many researchers have exploited AI's power to quickly detect the infections of COVID-19 [36] which includes the diagnosis of the virus through chest X-rays. Furthermore, face mask detection refers to the task of finding the location of the face and then detecting masked or unmasked faces [37]. Currently, there are many applications of face recognition and object detection in domains of education [38], autonomous driving [39], surveillance and so on [40].

In this research work, we have mainly considered one of the important preventative measures which are face masks to control the rapid transmission of COVID-19. Our proposed model is based on the Faster-RCNN object detection model, and it is suitable to detect violations and detect persons who are not wearing a face mask. The main contributions of this research work are given below:

- A novel deep learning model based on transfer learning with Faster-RCNN to automatically detect and localize masked and un-masked faces in images and videos
- A detailed analysis of the proposed model on primary challenging MS COCO evaluation metricsis also performed to measure the performance of the model
- This technique is not previously used and experimental analysis shows the capability of Faster-RCNN in localizing masked and un-masked faces
- Detailed analysis on real-time video of different frame rates is also performed which shows the capability of our proposed system in real-time videos.
- This system can be integrated with several surveillance cameras and assist different countries and establishments to monitor people in crowded areas and prevent the spread of disease.

The rest of the paper is categorized as follows; Section 2 explains related work, Section 3 presents methodology, Section 4 explains Results and Comparative Analysis followed by a Conclusion. Moreover, some sample images of the FMA dataset are given in Fig. 1.

**Figure 1:** Sample images of face mask dataset (FMA)

## 2 Related Work

Towards object detection and image recognition [41], many applicative advancements have taken place [40,42]. In various works, the main focus is on image reconstruction tasks and face recognition for identity verification, however, the main objective of this work is to detect the faces of individuals who are wearing or refraining from wearing masks. It should also be mentioned that this is within a real-time capacity, at various locations, with a focus on ensuring public safety from viruses and secondly, detecting individuals who violate the rules imposed by establishments or supervisory bodies.

Qin et al. [43] proposed the SRCNet classification network for the identification of face masks with an accuracy of 98.7%. In their work, they used three categories "correct wearing facemask", "incorrect wearing facemask", and faces with "no mask'. Ejaz et al. [44] used Principal Component Analysis (PCA) to recognize persons with masked and un-masked faces. It was observed in this research that PCA is not capable of identifying faces with a mask as its accuracy is decreased to 68.75%. Similarly, Park et al. [45] proposed the method to remove sunglasses from a human face, and then by using the recursive error compensation method the removed region was reconstructed.

Li et al. [46] used the object detection algorithm yolov3 to detect faces, based on darknet-19 which is a deep network architecture. For training purposes, they used WIDER FACE and Celebi databases, and later on, they validated their model on the FDDB database achieving accuracy results of 93.9%. In the same way, Din et al. [47] used Generative Adversarial Networks (GAN) based model to remove face masks from facial images, and then the region covered by the face mask is reconstructed using GAN. Nieto-Rodríguez et al. [48] proposed an automated system to detect the presence of surgical masks in operating rooms. The main objective of this system is to generate an alarm when a face mask is not worn by medical staff. This work has achieved 95% accuracy results. Khan et al. [49] proposed and designed an interactive model that can remove multiple objects from the given facial image such as a microphone and later on, by using GAN the removed region is reconstructed. Hussain et al. [50] use the architecture of VGG16 for recognizing and classifying the emotions from the face. In this work, KDEF database is used for the training of VGG16 and hence achieved an accuracy of 88%. Loey et al. [51] proposed a hybrid method for face mask classification which includes transfer learning models and traditional machine learning models. The proposed method is divided into two phases in which the first phase deals with feature extraction using ResNet 50, while the classification is performed in the second stage with Support Vector Machines (SVM), decision tree, and ensemble methods. They used three benchmark datasets for the validation of their proposed model and achieved the highest accuracy of 99.64% with SVM. Ge et al. [52] proposed a model along with a dataset to recognize masked and un-masked faces in the wild. MAFA a large face mask dataset that includes 35, 806 faces with a mask is introduced in this work. Moreover, they also proposed a model named LLE-CNN which is based on convolutional neural networks with three major modules. These modules include a proposal, embedding, and verification. They achieved an average precision of 76.1% with LLE-CNNs.

Furthermore, some researchers proposed different methods for detecting different accessories on faces by employing the use of image features and deep learning methods. These accessories commonly include glasses and hat detection. Jing et al. [53] proposed a method of glasses detection in a small area between eyes by using the information of edges in the image. Some traditional machine learning algorithms which include SVM and K-Nearest-Neighbor (KNN) are also used in the detection of different accessories from facial images [54–56]. Recently, deep learning methods are widely used in the detection of face accessories that are capable of extracting abstract and high-level information [57,58]. However, the different kinds of face masks on the face are also considered as facial accessories. Moreover, the conversion of low-quality images to high-quality images is necessary to increase the performance of classification and object detection methods [59–63]. For surveillance monitoring, Uiboupin et al. [64] adopted the approach of Super-Resolution (SR) based networks that utilize sparse representation for improving the performance of face recognition. Zou et al. [65] also adopted SR on low-resolution images to improve the performance of facial recognition and proved that there is a significant improvement in recognition performance by combining a face recognition model with Na et al. [62] improve object detection and classification performance by introducing the method of SR networks on cropped regions of candidates. However, these SR methods are either based on high-level representations or features of the face for improving accuracy of face recognition. In the case of facial image classification, especially regarding the automated detection of conditions with face masks, there have not been any report's published related to improvements in classification of facial images by employing deep-learning-based SR networks combined with networks for classification.

It is evident and observed from the above context that there is a very limited number of articles and research on face mask detection and there is also a need to improve existing methods. Additional experimentation and research on currently unused algorithms is also required. So, in the battle against COVID-19, we contribute to the body of mask recognition techniques utilizing the approach of transfer learning with the Faster-RCNN model.

## 3 Proposed Methodology

The proposed methodology is described below, Fig. 2 shows the architecture diagram of our proposed methodology:

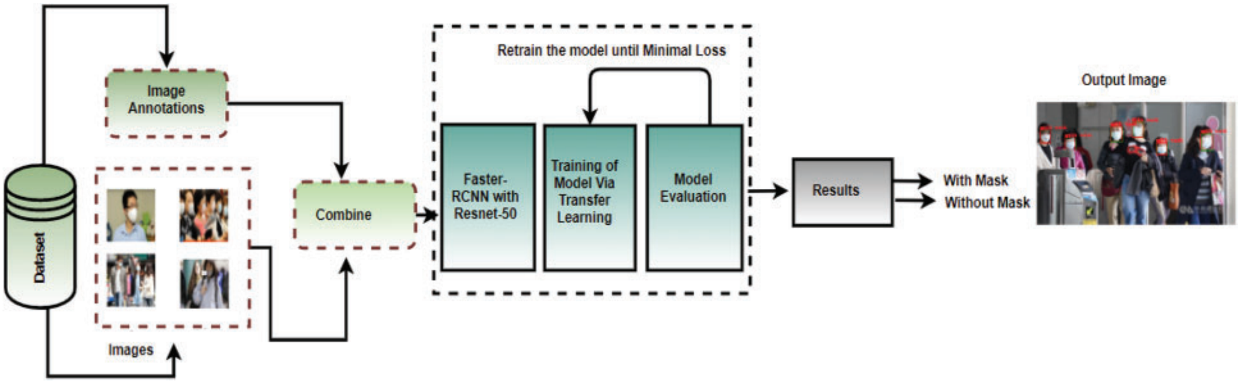
Image Annotations
Dataset
Images
Combine
Retrain the model until Minimal Loss
Faster-RCNN with Resnet-50
Training of Model Via Transfer Learning
Model Evaluation
Results
With Mask
Without Mask
Output Image

**Figure 2:** Schematic representation of the transfer learning with Faster-RCNN

### *3.1 Faster-RCNN Architecture*

Faster-RCNN is an extension of the Fast-RCNN model and it consists of two modules. The first module is based on the Region Proposal Network (RPN) which is simply a convolutional neural network that proposes different regions from an image. The second module is the detector which detects different objects based on the region proposals extracted by the first module. For object detection, it is a single-stage network. The attention mechanisms [66] helps the RPN network to tell the network of Faster-RCNN that where to look in the image.

### *3.2 Region Proposal Network (RPN)*

The input of the region proposal network can be an image of any size, and its output is different region proposals of a rectangular size that each has their own objectness score. It is a score generated for each region that shows whether the region contains an object or not. To generate region proposals, a small network slides over the output which is a convolutional feature map. A $n \times n$ spatial window is also taken as input by a small network. Every sliding window is mapped to a lower-dimensional feature (256-d for ZF or ZF-net [67] and 512-d for VGG, with ReLu [68] following). These features are passed to two fully connected layers named a box-regression layer and a box-classification layer.

### *3.3 Anchors*

At the location of each sliding window, multiple region proposals are predicted simultaneously, and for each location, the maximum possible proposal is denoted by $k$. All of these $k$ proposals are relative to $k$ reference boxes which are called anchors. Each anchor is associated with an aspect ratio and scale and it is centered at the sliding window.

### *3.4 Loss Functions*

To train the RPN network, a binary class label is used to determine whether it is an object or not to each anchor. The objective function which is to be minimized for an image is defined as:

$$L(\{pi\},\{ti\}) = \frac{1}{N_{cls}} \sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda \frac{1}{N_{reg}} \sum i\, p_i^*\, L_{reg}(t_i, t_i^*) \tag{1}$$

In the above equation, a mini-batch $i$ represents the index of an anchor and for each anchor, $p_i$ represents the predicted probability or output score of anchors being an object or not. If a positive anchor comes then $p_i^*$ which represents the ground truth is also one and it is zero if negative comes. In simple words, the first term in Eq. (1) is classification loss over two classes to determine whether it is an object or not an object. Similarly, the regression loss of bounding boxes is represented by the second term in Eq. (1). The four bounding box coordinates which are predicted by the model is represented by $t_i$ while the ground truth coordinates associated with a positive anchor is represented by $t_i^*$. The classification loss over two classes is represented by $L_{cls}$ and $L_{reg}\ (t_i,\ t_i^*) = R(t_i - t_i^*)$ is used for regression loss where $R$ represents the robust loss function (smooth L1) as defined in [69]. The regression loss is activated and represented by the term $p_i^*\ L_{reg}$ for positive anchors ($p_i^* = 1$) only and it is disabled if ($p_i^* = 0$). The outputs of the fully connected layers namely *cls* and *reg* comprised of $\{pi\}$ and $\{ti\}$. These terms are weighted by λ which is a balancing parameter and normalized by $N_{cls}$ and $N_{reg}$ respectively. $N_{cls}$ is the normalization parameter of mini-batch and $N_{reg}$ is the normalization parameter of regression loss which is equal to the number of locations of anchors. Moreover, for bounding box regression, the four coordinate's parameterizations are adopted [24]:

$$t_x = (x - x_a)/w_a,\ \ t_y = (y - y)/h_a \tag{2}$$

$$t_w = log(w/w_a),\ \ t_h = log(h - /h_a) \tag{3}$$

$$t^*x = (x^* - x_a)/w_a,\ \ t^*y = (y^* - y)/h_a \tag{4}$$

$$t^*w = log(w^*/w_a),\ \ t^*h = log(h^*/h_a) \tag{5}$$

where the box center coordinates are denoted by $x, w, y, h$ and also its width and height. The predicted bounding box, anchor bounding box, and the ground truth bounding box is denoted by $x,\ x_{a,},\ x^*$. The same is the case with $y,\ w$, and $h$. From an anchor box, this can be the same as a bounding regression box to the nearby ground truth. More specifically the width, height, and coordinates of the prediction box is represented by, $w,\ y,\ h$, for anchor box it is represented by $x_a, w_a, y_a, h_a$, and $x^*, w^*, y^*, h^*$ denotes the ground truth box coordinates.

### 3.5 Sharing of Features

Several ways are used to train the Faster-RCNN, such as, sharing features which include alternating training whereby the RPN network is trained first and then the proposal of regions generated by the RPN is used to train the Fast-RCNN. The alternative is to approximate joint learning via an ROI pooling layer used to differentiate w.r.t coordinates of boxes.

### 3.6 Transfer Learning Using Faster-RCNN

In this work, we utilize a transfer learning approach with Faster-RCNN. We start with a pre-trained Faster-RCNN model trained on the COCO-2017 dataset and then fine-tune the last layer to train a model on our custom dataset and the required number of classes [70]. The classifier is replaced with our classes which are "mask", "un-masked" faces, and background class. The backbone network here used is Resnet-50. The layers of Resnet-50 are not further train and will be kept frozen. Usually, in the concept of transfer learning by fine-tuning, the layers of the pre-trained network are kept frozen to prevent weight modification and avoid loss of information contained in pre-trained layers during future training. The layers of feature generation are fixed and there is the change in the only *cls* and *reg* layers. The total number of input channels specified is 256.

Moreover, there are many different anchor boxes, say *n* is given for each pixel with certain aspect ratios. The specification of anchor sizes and scale are 32, 24, 24, 16, and 8 respectively. The optimizer is set to Stochastic Gradient Descent (SGD). The learning rate is 0.005 with momentum and weight decay values are 0.9 and 0.0005 and the number of epochs is set to 20.

## 4 Results and Experiment Analysis

### 4.1 Dataset

In this research, the Face Mask Dataset (FMD) is used which is comprised of 853 images and their corresponding XML annotation files. Some image samples of the FMD dataset are shown in Fig. 1. The augmentation which includes Random horizontal flip is also applied to the images of the training set. Furthermore, the experimentation is performed on Google Colab with GPU in Python.

### 4.2 Results and Discussions

To evaluate the Faster-RCNN model, COCO-2017 evaluation metrics are used which include an Intersection of Union (IOU) score at different thresholds and the computing average precision and recall. The IOU score represents the overlapping and intersection area between actual and predicted bounding boxes divided by taking a union of both. To determine the value of IOU at which an object is inside the predicted bounding box we consider different thresholds. The challenge datasets which include PascalVoc and MS COCO show that the 0.5 IOU threshold is good enough. IOU is defined by Eq. (6):

$$IoU = \frac{area(Bp \cap Bgt)}{area(Bp \cup Bgt)} \tag{6}$$

A *Bp* and *Bgt* are the bounding boxes of predicted and actual. A detection is considered to be True Positive (TP) if a detection has an IOU greater than the threshold. A False Positive (FP) is considered to be the wrong detection because the IOU score is less than the threshold in this case. A case in which ground truth is not detected is False Negative (FN) while the corrected misdetection result is represented by True Negative (TN). Tab. 1 shows the Average Precision starting from IOU threshold 0.50 to 0.95 with a step size of 0.05 and with areas considered as "small" "medium"," large" and "all".

This is a primary and very challenging metric. The Average Precision (AP) in our experiment for small objects in the image in which an area is less than $32^2$ (on a scale of the pixel) is covered is 0.37. In this scenario the people standing very far away from the camera. The AP for the area of objects greater than $32^2$ and less than $96^2$ is 0.52 and are the medium objects in the image, while an area greater than $96^2$ is for large objects in the image and the AP for large objects is 0.81. Similarly, for areas equal to “all” it is 0.42 respectively. The maximum detections per image considered in our experiment is 100. Moreover, with this primary challenging metric, our Faster-RCNN model has achieved the highest AP of 0.81 respectively.

**Table 1:** Average precision at different scales and thresholds

| IOU threshold | Area | Maximum detection | Average precision |
|---|---|---|---|
| IoU = 0.50:0.95 | Small ($<32^2$) | 100 | 0.37 |
| IoU = 0.50:0.95 | Medium ($>32^2$ and $<96^2$) | 100 | 0.52 |
| IoU = 0.50:0.95 | Large ($>96^2$) | 100 | 0.81 |
| IoU = 0.50:0.95 | All | 100 | 0.42 |

Similarly, the Average Recall (AR) values are also considered with this primary challenging metric. Tab. 2 shows the AR values. If we consider maximum detections per image is 100, with an area greater than $96^2$, then AR is 0.84. Furthermore, AR values of IOU thresholds starting from 0.50 to 0.95, with a step size of 0.05 are also given in Tab. 2. For the medium objects having an area greater than $32^2$ and less than $96^2$, the AR values are 0.60. Similarly, for small objects, it is 0.44 respectively.

**Table 2:** Average recall at different scales and thresholds

| IOU threshold | Area | Maximum detection | Average recall |
|---|---|---|---|
| IoU = 0.50:0.95 | Small ($<32^2$) | 100 | 0.44 |
| IoU = 0.50:0.95 | Medium ($>32^2$ and $<96^2$) | 100 | 0.60 |
| IoU = 0.50:0.95 | Large ($>96^2$) | 100 | 0.84 |

Furthermore, if the maximum detections 1, 10, and 100 are considered for “all” areas then, AR achieved is 0.20, 0.46, and 0.50 respectively which is shown in Tab. 3.

**Table 3:** Average recall at a rate of different maximum detections

| IOU threshold | Area | Maximum detection | Average recall |
|---|---|---|---|
| IoU = 0.50:0.95 | All | 1 | 0.20 |
| IoU = 0.50:0.95 | All | 10 | 0.46 |
| IoU = 0.50:0.95 | All | 100 | 0.50 |

The other evaluation metric of MS COCO which is identical to PascalVoc is the AP at IOU threshold 0.5. So, in this case, the AP is 0.71 by considering the 100 detections per image with an area equal to "all". Another strict metric of MS-COCO is AP at the IOU threshold of 0.75. The AP achieved, in this case, is 0.47 with maximum detections per image of 100 as shown in Tab. 4.

**Table 4:** Average precision at PascalVoc metric and strict metric of MS COCO

| IOU threshold | Area | Maximum detection | Average precision |
|---|---|---|---|
| IoU = 0.50 | all | 100 | 0.71 |
| IoU = 0.75 | all | 100 | 0.47 |

Moreover, during the training of Faster-RCNN the different loss functions *vs.* the number of epochs and no of steps are plotted and In Fig. 3 the loss of classifier $L_{cls}$ and loss of bounding box regression $L_{box}$ is shown. The loss of objectness for region proposal by the RPN network and loss of regression box in the RPN network is also shown in Fig. 4. Figs. 3 and 4 graphs show the values of loss over many epochs. Similarly, the values of the same losses over many steps per epoch are also plotted which are shown in Figs. 5 and 6.

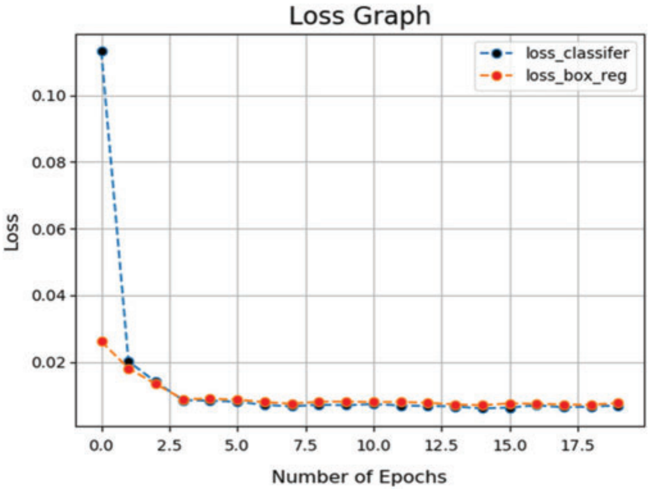

**Figure 3:** Graph of $L_{cls}$ and $L_{box}$ losses *vs.* Epochs

### *4.3 Analysis of Real-Time Video*

In the proposed work, we have also considered the detection of face masks in real-time videos. Videos consist of a stack of frames passing per second usually referred to as fps. If the real-time camera captures 30 frames per second, then it means there are 30 images in which the model needs to detect persons with masked and un-masked faces. The total time to detect each frame of video by our model is 0.17 s. Time analysis on videos of different fps is shown in Tab. 5.

It is observed from the above table that time decreases with the increase number of frames. Generally, the frame rate started with 30 fps is used in most of the videos.

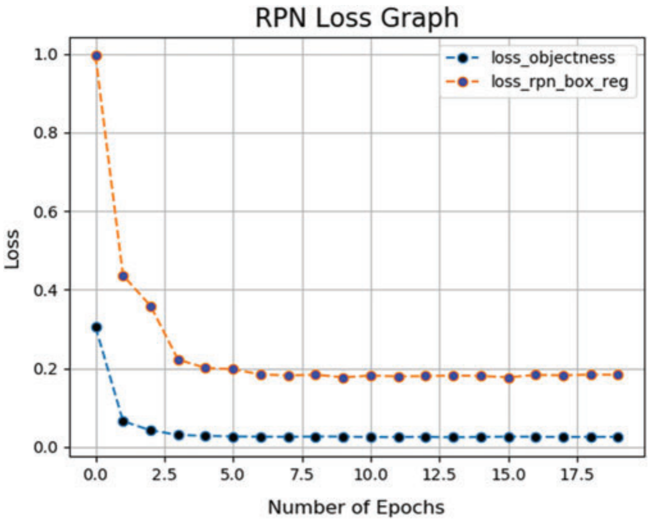

**Figure 4:** Graph of RPN loses *vs.* no of epochs

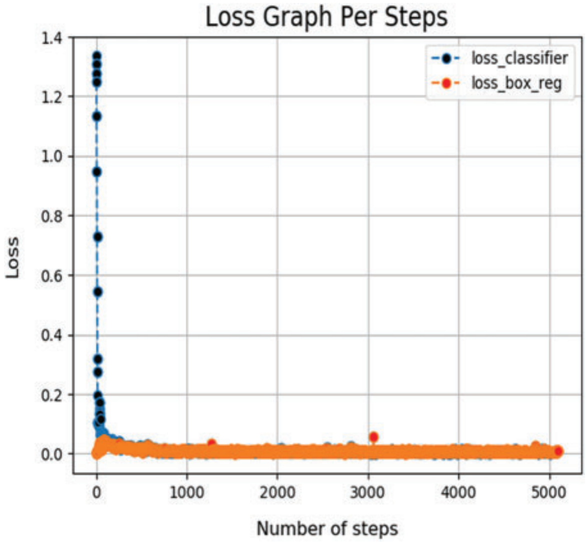

**Figure 5:** Graph of $L_{cls}$ and $L_{box}$ losses *vs.* no of steps per epoch

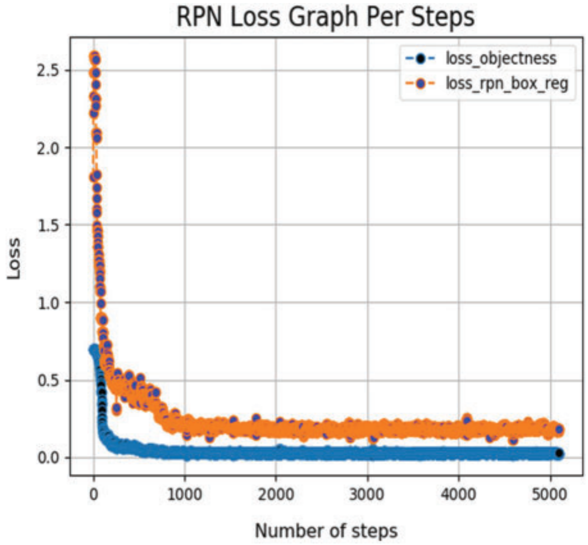

**Figure 6:** Loss of RPN network *vs.* no of steps per epoch

**Table 5:** Time Analysis for real-time videos

| Frame rate | Video duration | Time |
|---|---|---|
| 30 fps | 1 sec | 5.1 sec |
| 25 fps | 1 sec | 4.25 sec |
| 20 fps | 1 sec | 3.4 sec |
| 15 fps | 1 sec | 2.55 sec |

### *4.4 Comparison with Related Works*

The outcomes of our proposed Faster-RCNN with transfer learning is elaborated in previous sections. Our approach of using a transfer learning-based Faster-RCNN with Resnet-50 performs better in real-time face mask detection than previous models. Previous research articles mostly focus on the classification of masked and unmasked faces. The comparison of our approach to other models in this area is shown in Tab. 6. Altmann et al. [20] proposed LLE-CNN which is a CNN with three modules and achieved an average precision of 76.1% on the MAFA dataset which is a dataset of real face masks. The first module of LLE-CNN is the proposal module which is responsible for extracting candidate facial regions by combining two pre-trained CNN. All extracted regions are represented with high dimensional descriptors. After that, the second module named the Embedding module which uses a Linear Embedding (LLE) algorithm is used to convert these descriptors to the similarity-based descriptor. Lastly, the Verification module is employed for the identification of candidate faces followed by the utilization of unified CNN to jointly perform classification and regression tasks. Moreover, Alghamdi et al. [19] uses the hybrid approach to perform classification on masked and un-masked faces and achieves a classification accuracy of 99.64%. In their work, they use a deep learning approach by utilizing the architecture of ResNet50 for feature extraction followed by traditional ML algorithms to perform classification. The algorithms include decision tree, SVM, and ensemble learning. Similarly, Feng et al. [13] also perform a classification task on a face mask classification problem and achieves an accuracy of 70% by using the PCA algorithm. In the presented research, the object detection techniques are utilized and the highest AP and AR achieved is 81% and 84% respectively. We have analyzed the performance of our approach under the strict primary challenging metrics of MS COCO with different scales, IOU thresholds, and several detections per image. Moreover, some examples of detection results are also shown in Fig. 7.

**Table 6:** A comparative analysis of the proposed framework with existing work

| Authors | Method | Classification | Detection | Result |
|---|---|---|---|---|
| Loey et al. [51] | Hybrid | Yes | No | Accuracy = 99.64% |
| Ejaz et al. [44] | PCA | Yes | No | Accuracy = 70% |
| Din et al. [47] | GAN | Yes | Yes | - |
| Ge et al. [52] | LLE-CNNs | Yes | Yes | Average precision = 76.1% |
| **Proposed method** | **Transfer learning with Faster-RCNN** | **Yes** | **Yes** | **Average precision = 81% and Average recall = 84%** |

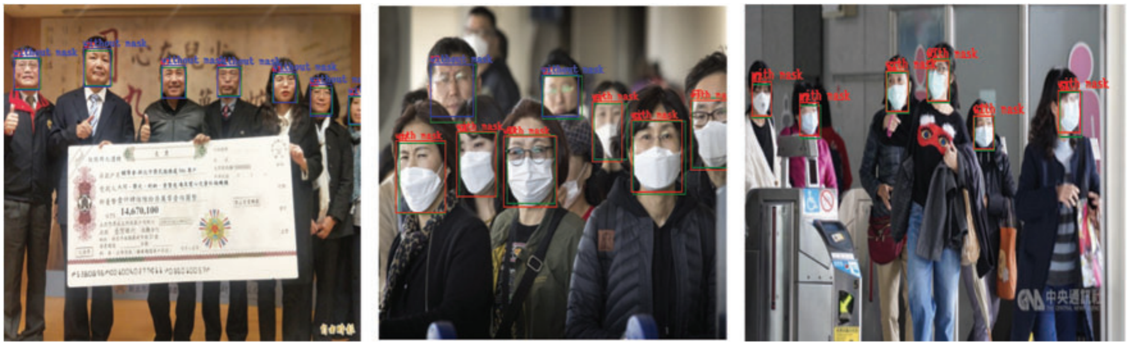

**Figure 7:** Detection results on images

## 5 Conclusion

In this paper, we proposed an automated system for the real-time detection of face masks to act as a preventative measure in controlling the rapid spread of COVID-19. This system helps the policymakers of different governmental authorities to track and monitor people who are not wearing face masks at public places in a bid to prevent the spread of the virus. Many countries have published statistics of COVID-19 cases that demonstrate the spread of COVID-19 is more than the observed value in crowded areas. The proposed model is based on a transfer learning approach with Faster-RCNN and achieved the highest AP and AR of 81% and 84% respectively. We analyze the performance of the proposed work with twelve primary challenging metrics of MS COCO. Furthermore, a detailed analysis of real-time videos of different frame rates is also presented. This work can be improved and extended by adding more diversity to a dataset and by applying other object detection algorithms which include a Single Shot Detector (SSD) in comparison with Faster-RCNN. Moreover, a generalized face recognition system while wearing a face mask can also be implemented.

**Funding Statement:** This work was supported King Abdulaziz University under grant number IFPHI-033-611-2020.

**Conflicts of Interest:** The authors declare that they have no conflicts of interest to report regarding the present study.

## References

[1] H. T. Rauf, M. I. U. Lali, M. A. Khan, S. Kadry, H. Alolaiyan *et al.,* "Time series forecasting of COVID-19 transmission in Asia pacific countries using deep neural networks," *Personal and Ubiquitous Computing*, vol. 6, no. 1, pp. 1–18, 2021.

[2] A. Sedik, A. M. Iliyasu, A. El-Rahiem, M. E. Abdel Samea, A. Abdel-Raheem *et al.,* "Deploying machine and deep learning models for efficient data-augmented detection of COVID-19 infections," *Viruses*, vol. 12, pp. 769, 2020.

[3] T. Akram, M. Attique, S. Gul, A. Shahzad, M. Altaf *et al.,* "A novel framework for rapid diagnosis of COVID-19 on computed tomography scans," *Pattern Analysis and Applications*, vol. 24, no. 11, pp. 1–14, 2021.

[4] S. P. Stawicki and S. C. Galwankar, "Winning together: Novel coronavirus (COVID-19) infographic," *Journal of Emergencies, Trauma, and Shock*, vol. 13, no. 2, pp. 103, 2020.

[5] A. Sedik, M. Hammad, F. E. Abd El-Samie, B. B. Gupta and A. A. Abd El-Latif, "Efficient deep learning approach for augmented detection of coronavirus disease," *Neural Computing and Applications*, vol. 21, pp. 1–18, 2021.

[6] A. M. Rahmani and S. Y. H. Mirmahaleh, "Coronavirus disease (COVID-19) prevention and treatment methods and effective parameters: A systematic literature review," *Sustainable Cities and Society*, vol. 64, pp. 102568, 2021.

[7] M. Sajjad, S. Khan, K. Muhammad, W. Wu, A. Ullah *et al.,* "Multi-grade brain tumor classification using deep CNN with extensive data augmentation," *Journal of Computational Science*, vol. 30, pp. 174–182, 2019.

[8] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Nevada, USA, pp. 770–778, 2016.

[9] P. A. Rota, M. S. Oberste, S. S. Monroe, W. A. Nix, R. Campagnoli *et al.,* "Characterization of a novel coronavirus associated with severe acute respiratory syndrome," *Science*, vol. 300, pp. 1394–9, 2003.

[10] Z. A. Memish, A. I. Zumla, R. F. Al-Hakeem, A. A. Al-Rabeeah and G. M. Stephens, "Family cluster of Middle East respiratory syndrome coronavirus infections," *The New England Journal of Medicine*, vol. 368, pp. 2487–94, 2013.

[11] Y. Liu, A. A. Gayle, A. Wilder-Smith and J. Rocklov, "The reproductive number of COVID-19 is higher compared to SARS coronavirus," *Journal of Travel Medicine*, vol. 27, pp. 1–4, 2020.

[12] Y. Fang, Y. Nie and M. Penny, "Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis," *Journal of Medicl Virology*, vol. 92, pp. 645–659, 2020.

[13] S. Feng, C. Shen, N. Xia, W. Song, M. Fan *et al.,* "Rational use of face masks in the COVID-19 pandemic," *The Lancet Respiratory Medicine*, vol. 8, pp. 434–436, 2020.

[14] B. J. Cowling, K. H. Chan, V. J. Fang, C. K. Cheng, R. O. Fung *et al.,* "Facemasks and hand hygiene to prevent influenza transmission in households: A cluster randomized trial," *Annals of Internal Medicine*, vol. 151, pp. 437–46, 2009.

[15] S. M. Tracht, S. Y. Del Valle and J. M. Hyman, "Mathematical modeling of the effectiveness of facemasks in reducing the spread of novel influenza a (h1n1)," *PLoS One*, vol. 5, pp. e9018, 2010.

[16] T. Jefferson, C. B. Del Mar, L. Dooley, E. Ferroni, L. A. Al-Ansary *et al.,* "Physical interventions to interrupt or reduce the spread of respiratory viruses," *Cochrane Database of Systematic Reviews*, vol. 2, pp. CD006207, 2011.

[17] N. H. L. Leung, D. K. W. Chu, E. Y. C. Shiu, K. H. Chan, J. J. McDevitt *et al.,* "Respiratory virus shedding in exhaled breath and efficacy of face masks," *Nature Medicine*, vol. 26, pp. 676–680, 2020.

[18] S. W. Sim, K. S. Moey and N. C. Tan, "The use of facemasks to prevent respiratory infection: A literature review in the context of the health belief model," *Singapore Medical Journal*, vol. 55, no. 3, pp. 160–7, 2014.

[19] A. Alghamdi, M. Hammad, H. Ugail, A. Abdel-Raheem, K. Muhammad *et al.,* "Detection of myocardial infarction based on novel deep transfer learning methods for urban healthcare in smart cities," *Multimedia Tools and Applications*, vol. 4, pp. 1–22, 2020.

[20] D. M. Altmann, D. C. Douek and R. J. Boyton. "What policy makers need to know about COVID-19 protective immunity," *Image and Vision Computing*, vol. 11, no. 2, pp. 1–25, 2021.

[21] K. He, X. Zhang, S. Ren and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Boston, MA, USA, pp. 1026–1034, 2015.

[22] H. Cao, H. Liu, E. Song, C. -C. Hung, G. Ma *et al.,* "Dual-branch residual network for lung nodule segmentation," *Applied Soft Computing*, vol. 86, pp. 105934, 2020.

[23] F. Afza, M. A. Khan, M. Sharif, S. Kadry, G. Manogaran *et al.,* "A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection," *Image and Vision Computing*, vol. 106, pp. 104090, 2021.

[24] M. Khan, S. Kadry, P. Parwekar, R. Damaševičius, A. Mehmood *et al.,* "Human gait analysis for osteoarthritis prediction: A framework of deep learning and kernel extreme learning machine," *Complex Intelligent Systems*, vol. 11, no. 3, pp. 1–27, 2021.

[25] M. Maqsood, S. Yasmin, I. Mehmood, M. Bukhari and M. Kim, "An efficient DA-net architecture for lung nodule segmentation," *Mathematics*, vol. 9, no. 13, pp. 1457, 2021.

[26] M. Bukhari, K. B. Bajwa, S. Gillani, M. Maqsood, M. Y. Durrani *et al.,* "An efficient gait recognition method for known and unknown covariate conditions," *IEEE Access*, vol. 9, pp. 6465–6477, 2020.

[27] M. Maqsood, M. Bukhari, Z. Ali, S. Gillani, I. Mehmood *et al.,* "A residual-learning-based multi-scale parallel-convolutions-assisted efficient CAD system for liver tumor detection," *Mathematics*, vol. 9, no. 10, pp. 1133, 2021.

[28] Z. Ali, A. Irtaza and M. Maqsood, "An efficient U-net framework for lung nodule detection using densely connected dilated convolutions," *The Journal of Supercomputing*, vol. 21, pp. 1–22, 2021.

[29] M. A. Khan, S. Kadry, Y. D. Zhang, T. Akram, M. Sharif *et al.,* "Prediction of COVID-19-pneumonia based on selected deep features and one class kernel extreme learning machine," *Computers & Electrical Engineering*, vol. 90, pp. 106960, 2021.

[30] S. Agarwal, N. S. Punn, S. K. Sonbhadra, P. Nagabhushan, K. Pandian *et al.,* "Unleashing the power of disruptive and emerging technologies amid COVID 2019: A detailed review," *Artificial Intelligence Review*, vol. 4, pp. 1–31, 2020.

[31] D. S. W. Ting, L. Carin, V. Dzau and T. Y. Wong, "Digital technology and COVID-19," *Nature Medicine*, vol. 26, pp. 459–461, 2020.

[32] S. K. Sonbhadra, S. Agarwal and P. Nagabhushan, "Target specific mining of COVID-19 scholarly articles using one-class approach," *Chaos*, *Solitons & Fractals*, vol. 140, pp. 110155, 2020.

[33] N. S. Punn, S. K. Sonbhadra and S. Agarwal, "Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and deepsort techniques," *Computers & Electrical Engineering*, vol. 90, pp. 126970, 2020.

[34] A. Sedik, A. M. Iliyasu, A. El-Rahiem, M. E. A. Samea, A. Abdel-Raheem *et al.,* "Deploying machine and deep learning models for efficient data-augmented detection of COVID-19 infections," *Viruses*, vol. 12, no. 7, pp. 769, 2020.

[35] N. S. Punn and S. Agarwal, "Crowd analysis for congestion control early warning system on foot over bridge," in *Twelfth Int. Conf. on Contemporary Computing (IC3)*, Noida, India, pp. 1–6, 2019.

[36] D. S. W. Ting, L. Carin, V. Dzau and T. Y. Wong, "Digital technology and COVID-19," *Nature Medicine*, vol. 26, pp. 459–461, 2020.

[37] A. M. Ismael and A. Sengur, "Deep learning approaches for COVID-19 detection based on chest X-ray images," *Expert Systems with Applications*, vol. 164, pp. 114054, 2021.

[38] K. Savita, N. A. Hasbullah, S. M. Taib, A. I. Z. Abidin and M. Muniandy, "How's the turnout to the class? A face detection system for universities," in *IEEE Conf. on e-Learning, e-Management and e-Services (IC3e)*, Melaka, Malaysia, pp. 179–18, 2018.

[39] D. -H. Lee, K. -L. Chen, K. -H. Liou, C. -L. Liu and J. -L. Liu, "Deep learning and control algorithms of direct perception for autonomous driving," *Applied Intelligence*, vol. 51, pp. 1–11, 2020.

[40] Z. -Q. Zhao, P. Zheng, S. -t. Xu and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[41] M. Rashid, M. A. Khan, M. Alhaisoni, S. -H. Wang, S. R. Naqvi *et al.,* "A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection," *Sustainability*, vol. 12, no. 12, pp. 5037, 2020.

[42] N. Punn and S. Agarwal, "Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks," *Multimedia Tools and Applications*, vol. 3, no. 4, pp. 1–19, 2020.

[43] B. Qin and D. Li, "Identifying facemask-wearing condition using image super-resolution with classification network to prevent COVID-19," *Sensors*, vol. 20, no. 18, pp. 5236, 2020.

[44] M. S. Ejaz, M. R. Islam, M. Sifatullah and A. Sarker, "Implementation of principal component analysis on masked and non-masked face recognition," in *1st Int. Conf. on Advances in Science, Engineering and Robotics Technology (ICASERT)*, Dhaka, Bangladesh, pp. 1–5, 2019.

[45] J. S. Park, Y. H. Oh, S. C. Ahn and S. W. Lee, "Glasses removal from facial image using recursive error compensation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 805–811, 2005.

[46] N. Wang, Q. Li, A. A. A. El-Latif, J. Peng and X. Niu, "An enhanced thermal face recognition method based on multiscale complex fusion for Gabor coefficients," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 2339–2358, 2014.

[47] N. U. Din, K. Javed, S. Bae and J. Yi, "A novel GAN-based network for unmasking of masked face," *IEEE Access*, vol. 8, pp. 44276–44287, 2020.

[48] A. Nieto-Rodríguez, M. Mucientes and V. M. Brea, "System for medical mask detection in the operating room through facial attributes," in *Iberian Conf. on Pattern Recognition and Image Analysis*, Santiago de Compostela, Spain, pp. 138–145, 2015.

[49] M. K. J. Khan, N. Ud Din, S. Bae and J. Yi, "Interactive removal of microphone object in facial images," *Electronics*, vol. 8, pp. 1115, 2019.

[50] S. A. Hussain and A. S. A. Al Balushi, "A real time face emotion classification and recognition using deep learning model," in *Journal of Physics: Conference Series*, vol. 3, no. 11, pp. 012087, 2020.

[51] M. Loey, G. Manogaran, M. H. N. Taha and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic," *Measurement*, vol. 167, pp. 108288, 2021.

[52] S. Ge, J. Li, Q. Ye and Z. Luo, "Detecting masked faces in the wild with lle-cnns," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Venice, Italy, pp. 2682–2690, 2017.

[53] Z. Jing and R. Mariani, "Glasses detection and extraction by deformable contour," in *Proc. 15th Int. Conf. on Pattern Recognition. ICPR-2000*, Barcelona, Spain, pp. 933–936, 2000.

[54] A. Fernandez, R. Casado and R. Usamentiaga, "A real-time big data architecture for glasses detection using computer vision techniques," in *3rd Int. Conf. on Future Internet of Things and Cloud (FiCloud)*, Rome, Italy, pp. 591–596, 2015.

[55] A. Fernández, R. García, R. Usamentiaga and R. Casado, "Glasses detection on real images based on robust alignment," *Machine Vision and Applications*, vol. 26, pp. 519–531, 2015.

[56] S. Du, J. Liu, Y. Liu, X. Zhang and J. Xue, "Precise glasses detection algorithm for face with in-plane rotation," *Multimedia Systems*, vol. 23, pp. 293–302, 2017.

[57] L. Shao, R. Zhu and Q. Zhao, "Glasses detection using convolutional neural networks," in *Chinese Conf. on Biometric Recognition*, Chengdu, China, pp. 711–719, 2016.

[58] Z. Xie, H. Liu, Z. Li and Y. He, "A convolutional neural network based approach towards real-time hard hat detection," in *IEEE Int. Conf. on Progress in Informatics and Computing (PIC)*, Suzhou, China, pp. 430–434, 2018.

[59] F. Zhang, F. Yang, C. Li and G. Yuan, "CMNet: A connect-and-merge convolutional neural network for fast vehicle detection in urban traffic surveillance," *IEEE Access*, vol. 7, pp. 72660–72671, 2019.

[60] S. Hao, W. Wang, Y. Ye, E. Li and L. Bruzzone, "A deep network architecture for super-resolution-aided hyperspectral image classification with classwise loss," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 4650–4663, 2018.

[61] P. Lu, L. Barazzetti, V. Chandran, K. Gavaghan, S. Weber *et al.,* "Highly accurate facial nerve segmentation refinement from CBCT/CT imaging using a super-resolution classification approach," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 1, pp. 178–188, 2017.

[62] B. Na and G. C. Fox, "Object detection by a super-resolution method and a convolutional neural networks," in *2018 IEEE Int. Conf. on Big Data (Big Data)*, Seattle, WA, USA, pp. 2263–2269, 2018.

[63] D. Cai, K. Chen, Y. Qian and J. -K. Kämäräinen, "Convolutional low-resolution fine-grained classification," *Pattern Recognition Letters*, vol. 119, pp. 166–171, 2019.

[64] T. Uiboupin, P. Rasti, G. Anbarjafari and H. Demirel, "Facial image super resolution using sparse representation for improving face recognition in surveillance monitoring," in *24th Signal Processing and Communication Application Conf. (SIU)*, Zonguldak, Turkey, pp. 437–440, 2016.

[65] W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Transactions on Image Processing*, vol. 21, pp. 327–340, 2011.

[66] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho and Y. Bengio, "Attention-based models for speech recognition," *Multimedia Tools and Applications*, vol. 6, no. 2, pp. 1–21, 2015.

[67] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conf. on Computer Vision*, London, UK, pp. 818–833, 2014.

[68] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Int. Conf. of Machine Learning*, NY, USA, pp. 1–6, 2010.

[69] R. Girshick, "Fast r-cnn," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1440–1448, 2015.

[70] A. Mehmood, M. A. Khan, M. Sharif, S. A. Khan, M. Shaheen *et al.,* "Prosperous human gait recognition: An end-to-end system based on pre-trained CNN features selection," *Multimedia Tools and Applications*, vol. 11, no. 7, pp. 1–21, 2020.