computers

MDPI

Article

Robust Face Mask Detection by a Socially Assistive Robot Using Deep Learning

Yuan Zhang [1], Meysam Effati [1,*], Aaron Hao Tan [1] and Goldie Nejat [1,2,*]

[1] Autonomous Systems and Biomechatronics Laboratory (ASBLab), Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON M5S 3G8, Canada
[2] KITE Research Institute, Toronto Rehabilitation Institute, University Health Network, Toronto, ON M5G 2A2, Canada
* Correspondence: meysam.effati@utoronto.ca (M.E.); nejat@mie.utoronto.ca (G.N.)

Abstract: Wearing masks in indoor and outdoor public places has been mandatory in a number of countries during the COVID-19 pandemic. Correctly wearing a face mask can reduce the transmission of the virus through respiratory droplets. In this paper, a novel two-step deep learning (DL) method based on our extended ResNet-50 is presented. It can detect and classify whether face masks are missing, are worn correctly or incorrectly, or the face is covered by other means (e.g., a hand or hair). Our DL method utilizes transfer learning with pretrained ResNet-50 weights to reduce training time and increase detection accuracy. Training and validation are achieved using the MaskedFace-Net, MAsked FAces (MAFA), and CelebA datasets. The trained model has been incorporated onto a socially assistive robot for robust and autonomous detection by a robot using lower-resolution images from the onboard camera. The results show a classification accuracy of 84.13% for the classification of no mask, correctly masked, and incorrectly masked faces in various real-world poses and occlusion scenarios using the robot.

Keywords: socially assistive robots; autonomous face mask detection; human–robot interactions; COVID-19 pandemic; extended ResNet-50

Citation: Zhang, Y.; Effati, M.; Tan, A.H.; Nejat, G. Robust Face Mask Detection by a Socially Assistive Robot Using Deep Learning. *Computers* **2024**, *13*, 7. https://doi.org/10.3390/computers13010007

Academic Editor: Paolo Bellavista

Received: 29 November 2023
Revised: 20 December 2023
Accepted: 21 December 2023
Published: 23 December 2023

1. Introduction

The coronavirus disease (COVID-19) has resulted in over 6.9 million deaths and 772 million infections worldwide as of November 2023 [1]. Recent studies during the pandemic have shown a positive trend towards the incorporation of embodied social robots for workplaces [2,3], public buildings [4,5], and healthcare settings [6,7]. Furthermore, a virtual robot has also been used to persuade people to wear masks, wash their hands, social distance, and promote COVID-19 vaccinations [8].

To prevent the spread of the virus, the World Health Organization (WHO) has suggested that people wear masks in public places [9]. Research has shown that wearing face masks can decrease the transmission and spread of infection [10,11]. Wearing masks in public indoor and outdoor spaces has been regulated as mandatory in most countries, and to date, with the easing of regulations, a number of healthcare, transit, education, and government facilities still require or strongly encourage masks. The commonly used masks people wear to reduce the spread of respiratory droplets from infected people while providing breathability include surgical masks, medical masks, and KN95 or N95 masks [12].

To enforce the mandatory wearing of face masks while reducing the time-consuming and costly task for humans to check people as they enter buildings, outdoor venues, etc., automated monitoring systems have been integrated into security cameras [13], tablets [14], and robots [15–20]. Recent detection systems have also used deep learning to classify whether people are wearing or not wearing masks (binary classification) in 2D images and videos [21–30]. Only a handful of face mask detection systems can detect not only if a

person is wearing a mask, but if they are wearing the mask properly [23,31]. Namely, these detection systems are able to detect that the mask fully covers the nose, mouth, and chin according to mask-use rules [32]. However, these systems have not yet been integrated for real-time mask detection on a robot to ensure that masks are properly worn. Therefore, a real-time autonomous deep learning-based mask detection system that can robustly detect and classify whether a mask is worn properly or if it only covers part of the face is needed in order to be implemented in robots.

In this paper, we present an autonomous, real-time, two-step deep learning (DL) method based on an extended ResNet-50 that can detect and classify face masks. Our method can identify the following classes: (1) if a mask is worn correctly and the nose, mouth, and chin are fully covered by the mask; (2) there is no mask on the face; and (3) a mask is worn incorrectly (partially worn) or the face is covered by other means (e.g., hand or hair). This novel mask detection method is incorporated and evaluated in real time on a physical socially assistive robot for a variety of different mask wearing and lighting conditions using lower-resolution 2D images from the robot's onboard camera.

2. Related Works

In this section, we discuss the mask detection methods using either machine learning or deep learning methods for the classification of faces with masks or without masks. Furthermore, we present a handful of robots that have incorporated mask detection systems for autonomous classification.

2.1. Machine Learning Methods for Mask Detection

Machine learning techniques have been used in mask detection to determine whether people are wearing masks in healthcare situations prior to COVID-19. For example, in [33], a system using 2D images that detected whether mandatory medical masks were worn in operating rooms was presented. This system used the LogitBoost method, developed using the AdaBoost learning algorithm and an additive logistic regression model [34]. It triggered an alarm when the system detected that a person was not wearing a surgical mask. The system consisted of a face detector and a mask detector. The color filter in each detector found the position, size, and threshold of the image features. The Labeled Faces in the Wild (LFW) dataset [35] was used in the detector's training phase, while the detector's test phase used the CMU Frontal Face Test Set [28,36] and the BAO dataset [37]. This face mask detection system was evaluated on the BAO dataset and a dataset the authors collected, and the system achieved a true positive rate (sensitivity) of approximately 95%.

In [38], the Viola–Jones detection method using the AdaBoost learning algorithm [34] was combined with the principal component analysis (PCA) algorithm to detect faces with masks or without masks. The PCA algorithm was used for feature extraction during the training process. This method used a combination of the ORL face database [39] and the authors' own captured images for training and testing. The average accuracies for masked face recognition and non-masked face recognition were determined to be 72% and 95%, respectively.

The machine leaning approaches discussed above only use binary classification (mask/no mask). Furthermore, they have only been tested on a priori collected 2D images datasets and have not yet been extended to real-time detection.

2.2. Deep Learning Methods for Mask Detection

In [40], a locally linear embedding convolutional neural network (LLE-CNN) was used to detect masked faces through the recovery of missing facial cues and the suppression of non-facial cues in the feature subspace. A new dataset named MAFA containing faces with various orientations and occlusions was created and used for both training and testing. The method combined two pre-trained CNNs for facial region extraction, an embedding module for facial cue recovery and a verification module to identify facial regions and refine their positions. Experiments were conducted on the proposed LLE-CNNs using the

MAFA dataset, and the results show an average precision of up to 76.4% in the detection of masked faces with facial regions occluded.

In [31], the RetinaFaceMask framework was proposed. The framework consisted of a feature pyramid network (FPN) which fused low- and high-level information through up-sampling and a context attention module (CAM) that combined high-level semantic information and detected the face mask state. Moreover, a cross-class object removal algorithm was developed to reject low-confidence and high-union-intersection predictions. The datasets used for both training and testing were the AIZOO Face Mask Dataset [34] and MaskedFAces for Face Mask Detection (MAFA-FMD) [22,33]. RetinaFaceMask had a mean average precision (mAP) of 68.3% for detecting faces with correctly worn masks, incorrectly worn masks, and no masks among the MAFA-FMD dataset.

In [21], a face mask detector that could detect multi-view faces was incorporated onto the Jetson Nano embedded system-on-module (SoM). The detector had two main modules: a fast face detector based on the Fast Face-CPU (FFCPU) deep learning architecture [41] for multi-view face detection and a slim CNN architecture with an attention module for mask/no-mask classification. The WIDER FACE dataset [42], the Simulated Masked Face Dataset (SMFD) [43], and the LFW Dataset [35] were used for training, while the testing was conducted on the Face Detection Data and Benchmark (FDDB) datasets [44]. The multi-view face mask detector achieved 96.6% accuracy for the FFCPU face detector and 99.72% for the face mask classification.

Off-the-shelf object detection methods using pre-trained networks have also been used for mask detection [22–29]. For example, in [22], the YOLOv3 object detection method with a backbone of Darknet-53 was used to classify people wearing masks or not wearing masks. The detection layers in YOLOv3 were adjusted for the detection of smaller faces, and the SoftMax loss function was chosen to maximize the inter-class feature differences and improve the computational speed by reducing the feature dimensions on detection layers. The training was conducted on two benchmark databases: the WIDER FACE [42] and the Celeb Faces Attributes (CelebA) [45] datasets. Tests were performed on the FDDB dataset [44,46]. The results showed that the method performed well on small faces and that the highest test accuracy achieved was 93.9%.

In [23], a face-mask-wearing identification method was developed that combined image super-resolution and classification networks (SRCNets). This method had three main steps: (1) facial detection and cropping, where all image facial areas were detected through a multitask cascaded CNN [47] and then cropped; (2) an image super-resolution (SR) network that outputted enhanced facial area images with pixel resolution of 224×224 pixels; and (3) a face-mask-wearing identification network which used MobileNet-v2 for classification. The dataset used during training and evaluation was the Medical Masks Dataset [48]. Three categories for face-mask-wearing status were identified: no face-mask-wearing, correct face-mask-wearing, and incorrect face-mask-wearing. An accuracy of 98.7% was achieved on the same dataset used during training.

In [24], a deep learning model based on YOLO-v2 with ResNet-50 was proposed to detect and annotate medical face masks in 2D images. The proposed model had two stages: feature extraction based on the ResNet-50 transfer learning model and medical face mask detection using YOLOv2. The datasets used to train the model were the Medical Masks Dataset [48] and the Face Mask Dataset (FMD) [49]. By introducing the mean Intersection over Union (IoU) for the estimation of the best anchor box number, the Adam optimizer achieved an average precision of 81% in finding and localizing medical-masked faces in 2D images.

In [25], a mask detection system was proposed that used (1) a color 2D principal component analysis–convolutional neural network (C2D CNN) for face detection; (2) a special convolutional architecture for mask recognition; and 3) the AlexNet architecture for classification. The datasets used for training and testing were the Real-World Masked Face Dataset (RMFD) [43] and Celeb Faces Attributes (CelebA) [45]. With the proposed mask

detection system, people with masks and without masks in the datasets were successfully classified (no accuracy results were reported).

In [26], the MobileNetV2 architecture was proposed for real-time face mask detection. The MobileNetV2 classified human faces with and without masks. It used four public datasets for both training and testing: Face Mask Lite Dataset [43], RMFD [43], MaskedFace-Net [50], and Face Mask Detection [51]. The experiments compared MobileNetV2 to other deep learning models such as ResNet-50, DenseNet, and VGG16, and the results showed that MobileNetV2 achieved the highest accuracy of 99% in both training and testing on the same datasets.

In [29], a deep learning model was presented that combined single-stage and two-stage classifiers to detect whether a person was wearing a face mask or not. The backbone of the network used a pretrained ResNet-50 classification model to extract information from 2D images and convert them into a feature map. A training pipeline was used for fine-tuning, and a deployment pipeline was used for real-time face detection and extraction. An identity predictor was used to retrieve the personal identification of people without masks. MAFA [40] and an unbiased dataset that the authors created were used for both training and testing. The experiments compared the detection performances of ResNet-50, AlexNet, and MobileNet. The results showed that the proposed ResNet-50 model achieved a 98.2% accuracy for binary mask detection on the datasets.

A Face Mask and Social Distancing Detection vision system was proposed in [27]. The system used pre-trained models of DenseNet, InceptionV3, MobileNet, MobileNetV2, ResNet-50, VGG-16, and VGG-19 for comparison purposes. The models with the highest accuracy rates (ResNet-50, VGG-16, and VGG-19) were then applied in an embedded vision system for testing. The dataset used for both training and validation was collected by the authors and included 2D images of faces with or without masks. Real-time testing results demonstrated that all the models could detect human faces and classify them successfully. It was reported that the system performed with an accuracy of 100% and a sensitivity of 99%.

An integrated method using stacked ResNet-50 and YOLOv5 for social distance monitoring and face mask detection was presented in [28]. A dual shot face detector (DSFD) was used to extract faces, and Stacked ResNet-50 was used to determine whether people were wearing masks or not. YOLOv5 was used to detect social distancing by detecting their positions and calculating distances using the density-based spatial clustering of applications with the noise (DBSCAN) clustering method. A dataset for both training and testing was collected from various sources, such as Google Images and RMFD [43]. Comparison experiments were performed with other DL methods such as Mobile Net V3, Inception V3, and ResNet-50. The proposed method outperformed these other methods with a testing accuracy of 84.13% using binary cross-entropy.

In [30], the automated detection of face masks and classification of mask types was presented using thermal images. Two deep learning models were adapted for the detection task: (1) the compact YOLOv5 "nano" model and (2) RetinaNet. A semi-supervised CNN with a Convolutional Autoencoder was used for mask classification. The models were trained on an extensive dataset comprising thermal images from different types of thermal cameras, showcasing people wearing three different kinds of masks. The YOLOv5 model had a mean average precision (mAP) higher than 97% and a precision rate of approximately 95% for the detection of face masks.

In [52], a two-stage mask detection method was proposed to monitor face mask rule violations in images captured by indoor web cameras. In the first stage, human tracking was conducted using DeepSORT [53] integrated with YOLOv3 [54] for real-time tracking. A multi-task cascaded convolutional neural network (MTCNN) was used to localize facial landmarks, focusing on the lower facial region and covering the noise, lips, and jawline in bounding boxes. The second stage involved k-means clustering of the nose area to classify faces as masked or unmasked. If the three largest clusters to be found fell within the "skin" RGB range, the face was classified as unmasked. The dataset included a combi-

nation of public face mask datasets and synthesized datasets. The method was tested on 5504 images, yielding an average accuracy of 97.13% in differentiating between masked and unmasked faces.

In [55], a two-stage face mask detection method based on the Faster R-CNN framework with MobileNet V2 was presented to classify face images into three categories: "with mask", "without mask", or "worn incorrectly". The method utilized ResNet-50 as a region proposal network (RPN) to generate anchor proposals. The dataset comprised 853 images of 3 classes (e.g., with mask, without mask, and mask worn incorrectly) from a Kaggle dataset [49]. These images underwent preprocessing, including resizing and conversion into tensors, followed by image augmentation to enhance training accuracy. The method achieved an overall mAP of 45% over 30 epochs in the testing phase; however, prediction results were not reported for the three different classes.

2.3. Autonomous Mask Detection by Robots

Only a handful of mask detection methods have been developed for use by robots and their onboard cameras [15–20].

The mobile rover Thor had a 2D camera that provided videos for mask detection [15]. It used an image generator (IG) that collected videos from various spaces, a human subject detector (HSD) that detected the presence of human subjects, a face detector and extractor (FD), and a mask detector (MD) that classified the detected face as masked or unmasked. A mask detection approach was used, consisting of (1) a deep learning model that integrated ResNet-50 with FPN for person detection, (2) multi-task convolutional neural networks (MT-CNN) for face detection and extraction, and (3) a CNN for masked and unmasked classification. The datasets used for training included the Microsoft Common Objects in Context (COCO) [56], CelebA [45], WIDER FACE [42], and Custom Mask Community Dataset [57] datasets. The mask detection system was implemented on the robot, and videos were collected by the robot from public spaces. Evaluation took place using these videos. The reported accuracy for classifying faces with/without masks was 81.31%.

In [16], a mask detection system that used a deep learning method with an improved YOLOv3 algorithm was developed for the small humanoid robot NAO. In the proposed method, an improved Darknet-53 with an EfficientNet attention mechanism (to increase image resolution) was used as the backbone. Distance-IoU (DioU) [58] was used to evaluate the normalized distance between the predicted bounding box and the true box. The training and validation dataset consisted of images from the internet and the authors' own dataset of masked and unmasked faces. The NAO robot captured images of people front-facing its 2D camera and then provided these images to the trained model to obtain the classification prediction (face without or with mask). The results showed that the mAP was 86.92% accurate on the validation set.

In [17], the NAO robot was used to promote face mask usage in public spaces during the pandemic. The robot used an external webcam and a single shot detector (SSD) with a ResNet-10 architecture as its backbone for mask detection. An experiment was conducted with the robot providing verbal and visual feedback based on mask detection results. Green LEDs and positive messages indicated correct usage, while red LEDs and warnings were used for incorrect or no mask usage. The results from the study showed that the detection had a 95% accuracy rate and that 87.5% of participants responded positively to the robot's interactions, highlighting the effectiveness of social robots in encouraging health and safety measures.

In [18], a mask detection feature was added to the Pepper robot. Pepper could simultaneously scan five faces as a group to determine whether people were wearing masks or not. The mask detection method was based on an SSD structure for face detection and the four deep learning frameworks of PyTorch, TensorFlow, Keras, and Caffe for mask and non-mask classification [59]. The datasets used for training were WIDER FACE and MAFA. No results were reported.

In [19], a deep learning mask detection method that identified people without masks and warned them of the probability of receiving a fine was incorporated into an outdoor security robot. The security robot had six cameras that provided a 360-degree view which could detect people without masks at distances of 4–24 feet. The robot took 2D images of people without masks and sent them to a cloud server for personal identification.

In [20], a custom autonomous medical assistant robot was developed to assist healthcare facilities in managing COVID-19 preventive measures. The robot used a 2D camera with a ConvNet model for face mask detection. The ConvNet model was trained on two image datasets containing people with and without masks. An experiment was conducted with the robot autonomously navigating healthcare facilities, monitoring mask usage, and promoting social distancing. While the robot's operational efficacy in detecting masks and encouraging social distancing was affirmed, the study did not provide quantitative results.

2.4. Summary of Limitations

From the aforementioned literature review, a number of existing mask detection methods have been developed, which have only provided binary classification (mask/no mask) using either 2D RGB images [12–16,18,19,52,55], thermal images [30], video streams [15], or real-time data from webcams [13,16,19,28]. However, there is limited research determining if people are properly covering their noses, mouths, and chins with well-fitted masks [23,31], 55]. Furthermore, [55] did not report separate results for such a class [31]. Moreover, there are only a few methods that deal with real-time autonomous image capturing, pre-processing, and detection with robots [15,16,18,19] and the use of low-resolution images [15,16,18,19]. Namely, the majority of DL methods have only been tested on datasets that include images that have already been cropped and are of front-facing people (i.e., directly facing the camera). These datasets do not reflect real-world scenes, which can have different/changing lighting conditions, varying face poses, and face occlusions during capturing.

Many mask detection methods use high-resolution images while training and testing their models; however, robots have cameras that provide lower-resolution images (e.g., Pepper's onboard camera has a 640 × 480 pixel image resolution with a maximum signal-to-noise ratio of 6 dB [60]), which may cause detection and classification errors. To date, to the authors' knowledge, there are only three mask detection systems integrated into real robots; however, they only provide binary mask classification in real-world settings [15,17,20].

In this paper, we propose a novel autonomous mask detection and classification system for a socially assistive robot that uses lower-resolution images to detect in real time whether people are wearing masks properly. The robot is able to detect and classify non-binary mask classification categories. Our system consists of two deep learning classifiers. The first is used to classify people with/without face masks, and the second is used to classify correctly/incorrectly worn face masks. Both classifiers use our new extended ResNet-50 structure that includes additional layers to improve its accuracy in real-world conditions (varying lighting, occlusion scenarios, and face poses) and to minimize overfitting to the training data. The extended ResNet-50 structure takes advantage of transfer learning by using pre-trained ResNet-50 weights for training and classification.

3. Autonomous Mask Detection and Classification Methodology

The overall structure of the proposed Robust Face Mask Detection and Classification system using our novel extended ResNet-50 structure is presented in Figure 1. It contains three main modules: (1) image pre-processing, (2) face detection, and (3) classification (consisting of Classifier 1 and Classifier 2). Raw images are pre-processed into square image inputs, where human faces are located at the center of the images and sent to the Face Detection module. Facial features are identified and provided to the classification module. The output of the classification module is the class of the image of interest, either: (1) face with no mask, (2) face with correctly worn mask, and (3) face with incorrectly worn

mask/face covered by other means. The main modules of our architecture are discussed in detail below.

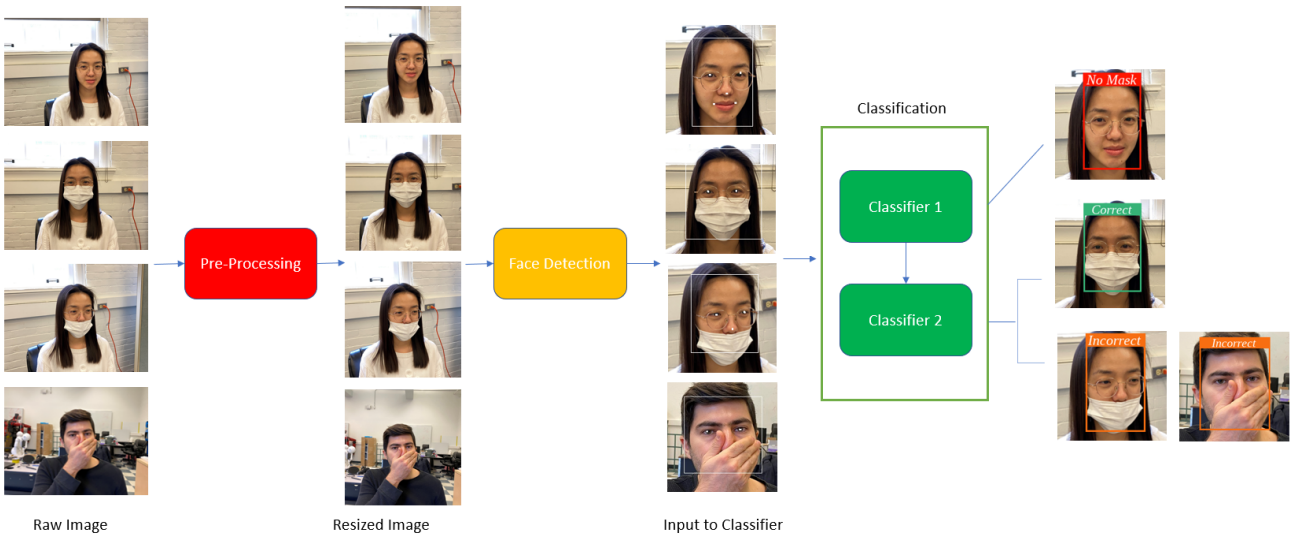

Figure 1. Proposed mask detection and classification system.

3.1. Pre-Processing Module

The purpose of the image pre-processing module is to resize the obtained $w \times h$ input rectangular images. When w is greater than h, the image is resized into a square $h \times h$ image, where w is the width and h is the height of the image. Raw images are segmented to ensure that the person is located in the center of an image using contour detection of the head and shoulders. Specifically, an OpenCV-based contour detection method utilizing the Haar Cascade classifier was used for head and shoulder detection [61]. Once the head and shoulder contours are detected, the image is cropped and adjusted to center these contours.

3.2. Face Detection Module

The multitask cascaded convolutional neural network for joint face detection and alignment is used to obtain facial areas under various conditions [47]. Several potential candidate windows of the positions of five facial landmarks, i.e., two eyes, nose, and the two corners of the mouth, are generated, and the highly overlapped candidates are then merged using non-maximum suppression (NMS). The output is a bounding box of the face and the positions of the five facial landmarks. The images are then cropped in size to 224×224 pixels.

3.3. Classification Module

We developed an extended ResNet-50 structure for both of our classification models (Classifier 1 and Classifier 2) in order to detect accurate facial features for extraction. As there are subtle differences between the facial features when people are correctly wearing masks versus incorrectly wearing masks, or when covering the mouth and nose by other means (e.g., a hand or hair), the feature extraction network's depth is important [62]. Each classifier consists of an extended ResNet-50 two-stage CNN. We use two-stage CNNs herein as they have been shown to achieve higher detection and classification accuracy compared to single-stage CNNs [63], as well as being able to incorporate robust feature extraction when subtle variations exist [64].

ResNet-50 is a 50-layer convolutional neural network with a 7×7 input layer followed by a 3×3 MaxPool layer, 4 stages of various convolutional layers, and a 1 Average Pool layer [65]. In general, ResNet-50 has been shown to have the highest accuracy among the deep CNN models, including InceptionV3, MobileNetV3, VGG-16, and VGG-19, which have all been trained on the ImageNet database [66]. The ResNet50 ImageNet pre-trained

weights are used for our deep CNN models. Our proposed extended ResNet-50 architecture is shown in Figure 2. The corresponding convolutional layers with their kernel sizes and output sizes are also presented. We added a flattening layer to transition from the ResNet model to a fully connected layer. This flattening layer converts all the multi-dimensional input tensors into a single, long, continuous, linear vector. Then, a fully connected dense layer with a rectified linear unit (ReLU) [67] activation function is used to classify the extracted features from the connected neurons. The dense layer performs matrix vector multiplication, while the ReLU activation, $f(x) = argmax(0, x)$, is a piecewise, non-linear activation function that outputs x if positive or zero otherwise [67]. ReLU is used since the neurons will only be deactivated if the output of the linear transformation is less than 0, which helps to prevent exponential growth in computation. ReLU also has fewer vanishing gradient problems in the hidden layers of the structure, which are between the input and output layer in the figure. A dropout layer is incorporated after the fully connected dense layer to reduce overfitting [68]. Another dense layer with SoftMax classifier [69] is used for classification. The softmax function takes the input vector and normalizes it into a probability distribution; thus, it outputs the probabilities of the mask classes.

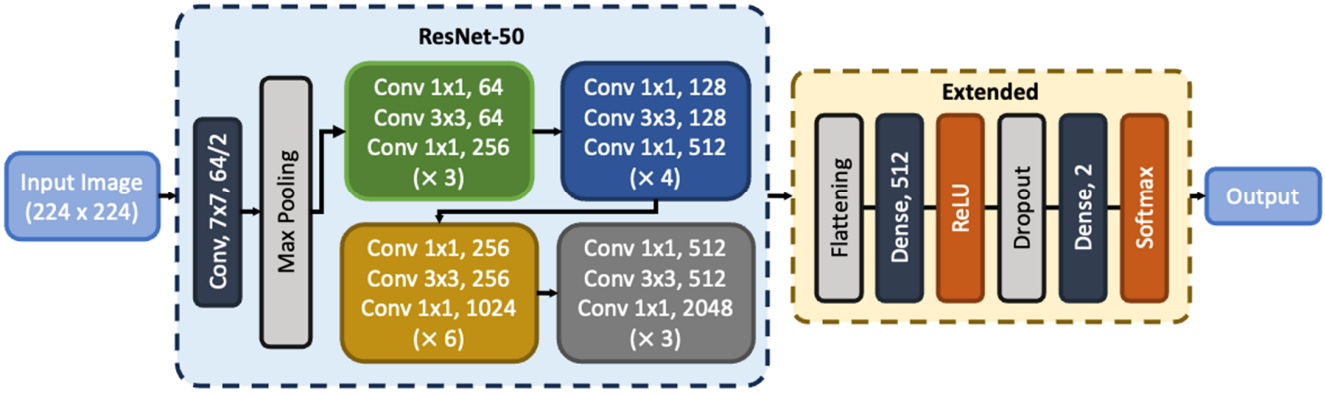

Figure 2. The proposed architecture of extended ResNet-50 Model.

We set all the layers in the ResNet-50 as non-trainable layers to take advantage of the pre-trained weights. Table 1 presents the number of parameters used for classification in our architecture. Namely, the total number of parameters is 24,637,826, while the number of trainable parameters in the extended layers is 1,050,114. The numbers of parameters for Classifier 1 and Classifier 2 are the same, as both classifiers use the same number of training images. If the output probability of the softmax function in Classifier 1 has classified an image as a face with no mask, the model will (1) output the classification result as "No Mask", and the classification process stops. If it has classified the image as "face with a mask," Classifier 2 is initiated to determine whether the mask is: (1) correctly worn (Correct) or (2) incorrectly worn/occluded by other means (Incorrect).

Table 1. Classification module parameter summary for both Classifier 1 and Classifier 2.

Layers	Output Shape	Parameter Count
ResNet-50	(None, 2048)	23,587,712
Flatten_1	(None, 2048)	0
Dense_2	(None, 512)	1,049,088
Dropout_1	(None, 512)	0
Dense_3	(None, 2)	1026
Total parameters: 24,637,826		
Trainable parameters:1,050,114		
Trainable parameters:1,050,114		
Non-trainable parameters: 23,587,712		

4. Training

The extended ResNET models for Classifier 1 and Classifier 2 were trained through Google Colab using GPUtil-1.4.0. Both Classifier 1 and Classifier 2 were developed in Python 3.9.

4.1. Training Dataset

Several public datasets were combined for the training of our proposed mask detection and classification system. These included: (1) MaskedFace-Net [50], (2) MAsked FAces (MAFA) dataset [40], and (3) Celeb Faces Attributes (CelebA) [45]. Example images are shown in Figure 3.

Figure 3. Example images of people with (**a**) correctly masked and incorrectly masked faces from MaskedFace-Net [50]; (**b**) faces covered by other means from MAFA [40]; and (**c**) unmasked faces from CelebA [45]. Images are from the public domain or permissions have been obtained.

MaskedFace-Net [50]: The MaskedFace-Net dataset [50] has 133,783 human face images, including 67,049 images of correctly masked faces (CMFD) and 66,734 images with incorrectly masked faces (IMFD). The images in this dataset have resolutions of 1024 × 1024 pixels and the majority of them are front-facing. We used 3000 images from CMFD and 3000 images from IMFD for the Classifier 1 training dataset. Another 6000 images from CMFD, as well as 4000 images of which 2000 were of masks covering only the mouth and chin and 2000 were of masks covering only the nose and chin, were selected from IMFD for the Classifier 2 training dataset.

MAsked FAces (MAFA) dataset [40]: The MAFA dataset [40] consists of 30,811 images of occluded faces, including occlusion by arms, hair, sunglasses, etc. This dataset has images with various resolutions, ranging from 188 × 188 pixels to 3500 × 2329 pixels. The majority of people are front-facing or slightly turned to the side. Two thousand images of people under various occlusions were selected for the incorrectly worn mask/face covered by other means class for the Classifier 2 training dataset.

CelebA dataset [45]: The CelebA dataset [45] has 202,599 facial images of celebrities under various situations, including unmasked faces that are mostly front-facing or slightly turned to the side. Six thousand images were selected as the no-mask class for the Classifier 1 training dataset from CelebA, with various resolutions from 178 × 218 pixels to 2560 × 1440 pixels.

We used a total of 24,000 images for training. There were 6000 images for faces with masks and 6000 images for faces with no masks in the training dataset for Classifier 1. For Classifier 2, there were 6000 correctly masked facial images, 2000 images occluded by other means, 2000 incorrectly masked faces with masks covering only the mouth and chin, and 2000 incorrectly masked faces with masks covering only the nose and chin to prevent dataset imbalance. A 70% and 30% training and validation split was used.

4.2. Training Procedure

Classifier 1 was trained to distinguish between images of faces with masks and faces without masks. For this, the binary cross-entropy loss function [70] was utilized to differentiate the predicted probabilities for these two categories against their true values. Training involved a batch size of 32 and a learning rate of 0.0001 over 250 epochs. An early stopping criterion was implemented to halt training if the validation loss remained constant for 25 epochs. Training for Classifier 1 took approximately 64.96 h. Classifier 2 was then trained to further categorize images identified as 'face with a mask' by Classifier 1. This classifier predicted whether the mask was worn correctly or incorrectly/obscured. The same training parameters and loss function used for Classifier 1 were applied to Classifier 2, ensuring consistency in the learning approach. Classifier 2 underwent training for approximately 16.24 h.

Both classifiers shared the same number of parameters, as they were trained on identical sets of images. The training process was conducted on a computer with an x64-based processor, 11th Gen Intel Core (TM) i7-1185G7, 16 GB RAM, and a certified OpenGL graphics card operating on Ubuntu 16.04 (Linux) and Python 2.7 with the Python SDK (Naoqi package 2.5.7.1) [71,72].

4.3. Classifier 1: Training and Validation Results

The loss and accuracy for Classifier 1 during both the training and validation phases are presented in Figure 4a,b. Classifier 1 showed convergence in its loss function, with the final training loss at 0.1 and the validation loss at 0.05 after 200 epochs. Classifier 1 also achieved a training accuracy of 97.80% and a validation accuracy of 97.3%.

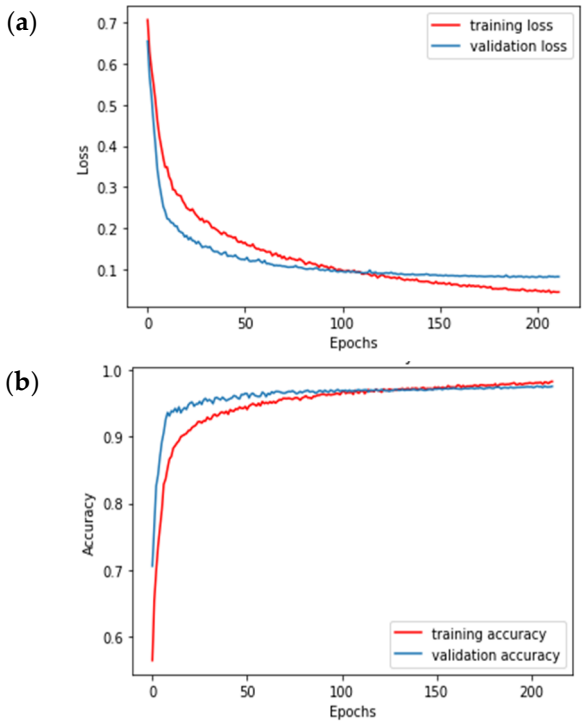

Figure 4. Classifier 1: (**a**) loss and (**b**) accuracy graphs.

4.4. Classifier 2: Training and Validation Results

Figure 5a,b present the loss and accuracy for Classifier 2. Classifier 2 achieved a final training loss of 0.01 and a validation loss of 0.15 after 50 epochs. Furthermore, Classifier 2 achieved a training accuracy of 99.62% and a validation accuracy of 95.93%.

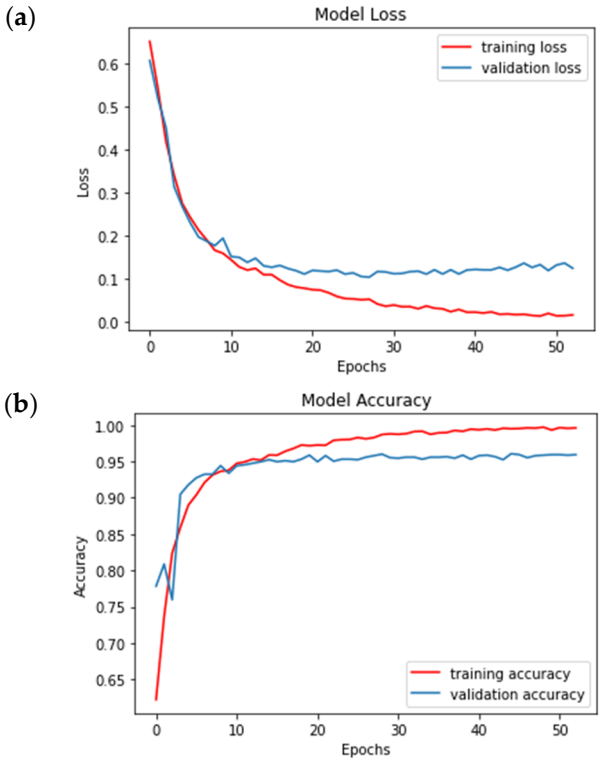

Figure 5. Classifier 2: (**a**) loss and (**b**) accuracy graphs.

5. Comparison Study

We conducted a feasibility comparison study between our extended ResNet-50 classifiers and the standard ResNet-50 classifier [66] for mask classification. Both methods were trained using the procedure presented in Section 4. Testing was conducted on a new unseen dataset consisting of a total of 128 images (4032 × 3024 pixel resolution) collected from members in our lab under different mask-wearing scenarios, which included faces with no masks, faces with correctly worn masks, and faces with incorrectly worn masks/faces covered by other means. Example images are shown in Figure 6.

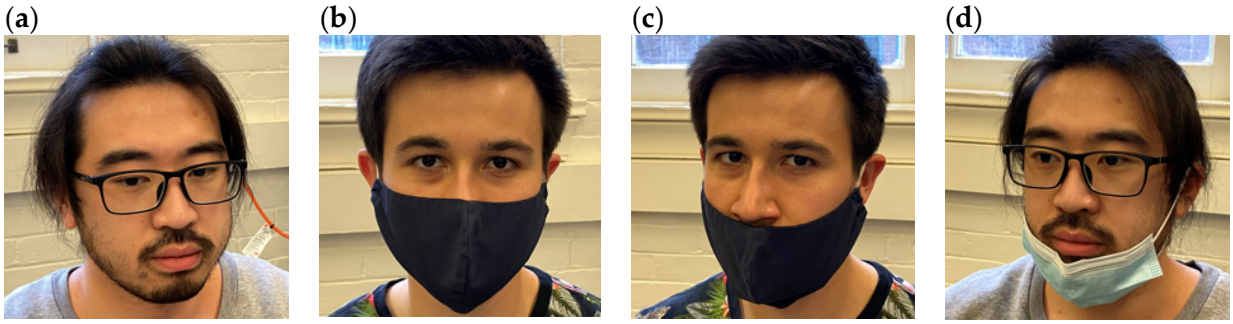

Figure 6. Example images from our comparison dataset with: (**a**) no mask, (**b**) correctly worn mask, (**c**) incorrectly worn mask (covering mouth and chin), and (**d**) incorrectly worn mask (covering chin only).

Table 2 presents the test accuracies of both classifiers for our extended ResNet-50 and the ResNet-50 methods. As can be seen in the table, the testing accuracies of 97.66% and

95.79% for Classifier 1 and Classifier 2 were higher for our extended ResNet-50 method than the standard ResNet-50 method. In particular, our Classifier 2 showed a classification improvement of 12.3% over ResNet-50 in classifying mask wearing.

Table 2. Testing accuracy comparison for ResNet-50 and our Extended ResNet-50.

Method	Classifier 1	Classifier 2
ResNet-50	95.31%	85.26%
Extended ResNet-50	97.66%	95.79%

6. Robot Experiments

We conducted mask detection experiments on the Pepper robot from Softbank Robotics to investigate the performance of our overall proposed autonomous detection and classification system under real-world conditions.

6.1. Robot Setup

Pepper has an onboard 2D camera located on its forehead with a resolution of 640 × 480 pixels. The camera was used to capture the images used by our mask detection and classification system via the getImageRemote function [73] and the Image module [74]. The images were sent (through a Wi-Fi channel used for the communication between the computer and the robot) to the same computer used for training to perform the entire classification process in real time.

6.2. Experiment Procedure

In order to investigate the robustness of our mask detection and classification system, the Pepper robot was situated for a day in front of our lab during the COVID pandemic. Twenty people (university students and staff) at various times approached and stood in front of Pepper, as shown in Figure 7. Various conditions were tested, including the following scenarios:

1. Mask covering parts of the nose, mouth, and chin;
2. Long hair partially covering the face or mask;
3. Hand covering the face or mask;
4. Different head rotations with respect to the robot (looking down, up, or to the sides);
5. Different mask colors and shapes;
6. Mask covering the forehead;
7. No mask worn;
8. Different lighting conditions (dim to bright lighting).

A finite-state machine was used to determine the behaviors of the robot. When the person stood in front of the robot, the robot would greet them, ask them to sanitize their hands, and then let them know it was checking for a face mask and prompt them to look at its camera. Then, it would state the result of its classification and either say good-bye (in the case of a mask worn correctly) or provide further instructions such as (1) to visit the front desk (in the case of no mask) or (2) to put on their mask properly (in the case of a mask worm incorrectly/face covered by other means). A video of our experimental procedure highlighting examples for the three mask-wearing classes is presented on our YouTube channel (https://www.youtube.com/watch?v=zcIb9sM_O10 (accessed on 29 September 2022)).

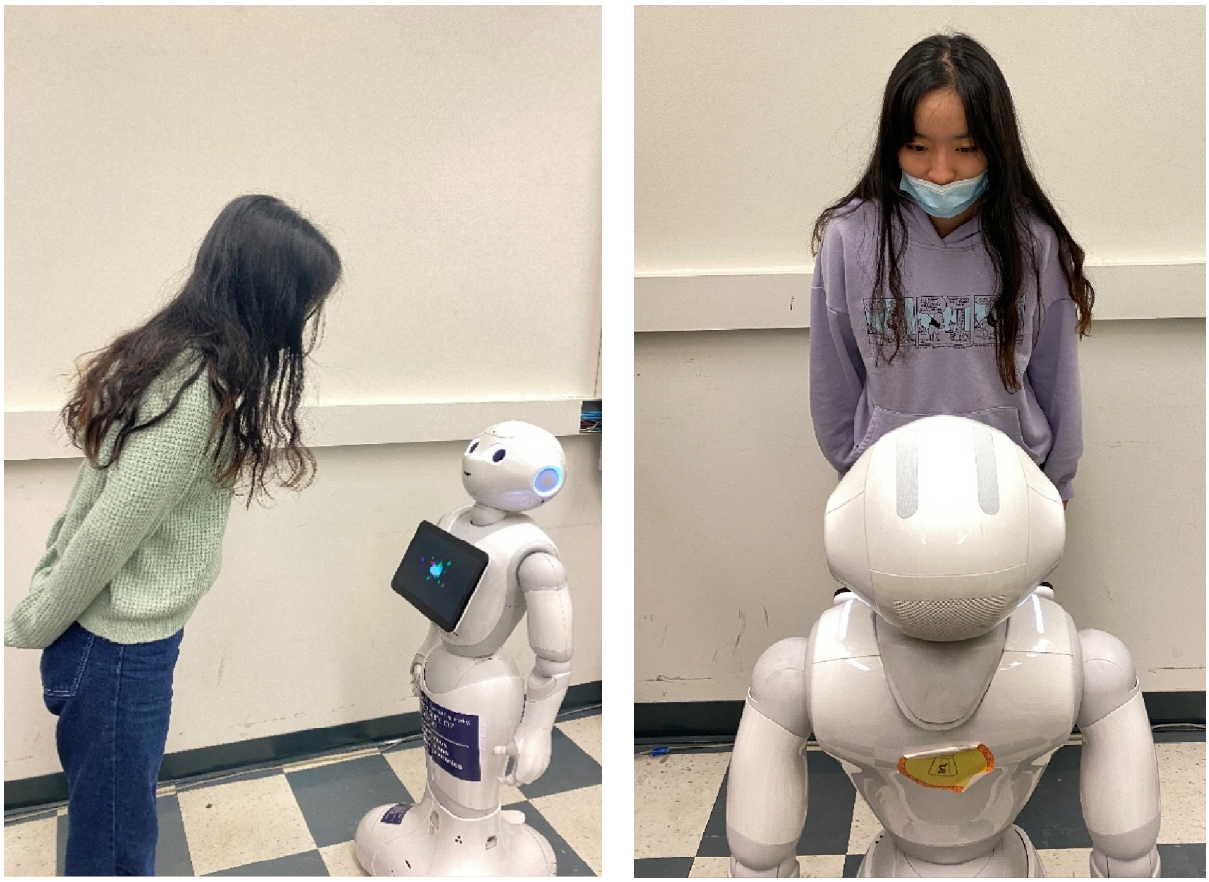

Figure 7. People standing in front of the Pepper robot for mask detection and classification.

6.3. Results

One hundred and twenty-six images were taken by Pepper and classified by our mask detection and classification system. The classification accuracy, precision, recall, and F1 scores for both classifiers are presented in Table 3. Example images and their classification labels are also shown in Figure 8. The average processing time for mask classification by the robot was 0.11 s.

Table 3. Mask classification results for robot experiment.

	Classifier 1	**Classifier 2**
Accuracy	91.27%	86.27%
Precision	95.05%	75.47%
Recall	94.12%	97.56%
F1 score	94.58%	85.11%

In general, the results show high accuracy, precision, recall, and F1 scores for both classifiers in terms of varying head poses and lighting conditions in low-resolution images taken by the robot. In Figure 8, it can be seen that our classifiers were able to identify the three mask classes for these conditions. For Classifier 1, when the mask covered only the chin, the classifier was able to identify the correctly worn mask class to be sent to Classifier 2, except when the lighting was too dim, at which point it was classified as no mask. This difference can be seen in the two images presented in Figure 8h, column 3, where for the second image, the lighting is visibly lower. In general, Classifier 2 was able to detect incorrectly worn masks as shown in Figure 8a,d–f. There were some conditions when the face was covered by other means and Classifier 2 was not able to detect the correct class. For example, in Figure 8b, column 2, where long hair was used to cover the nose, mouth,

and chin region of the face; the dark hair fully covered these features similarly to how a mask would. However, as can be seen in the rest of Figure 8b, the classifier was able to detect when there was a mask partially occluded by hair (column 1); when just hair was covering the mouth and chin, but not the nose (column 3); or when hair was covering the face at a descending angle, differently than a mask with symmetric ear loops would (column 4).

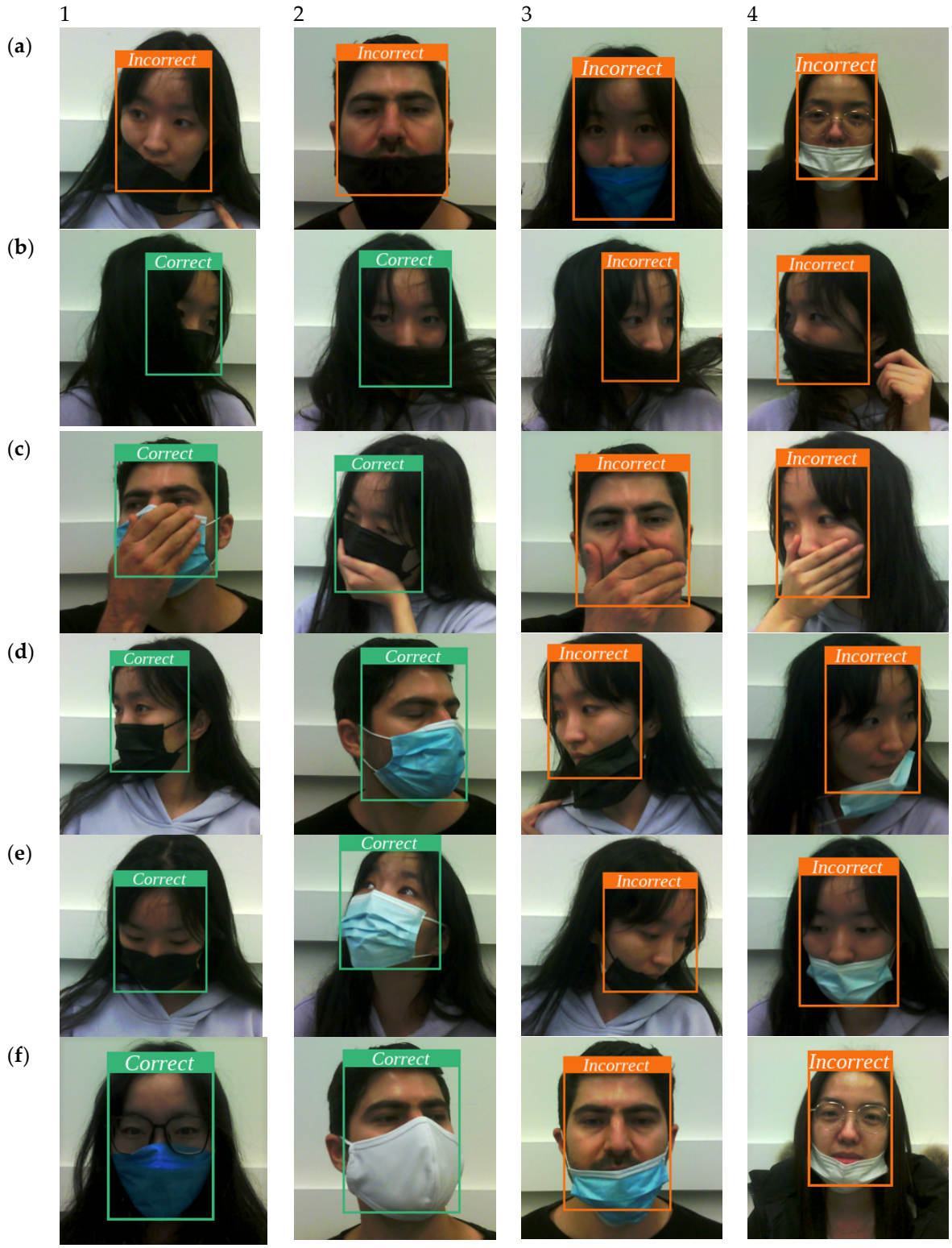

Figure 8. *Cont.*

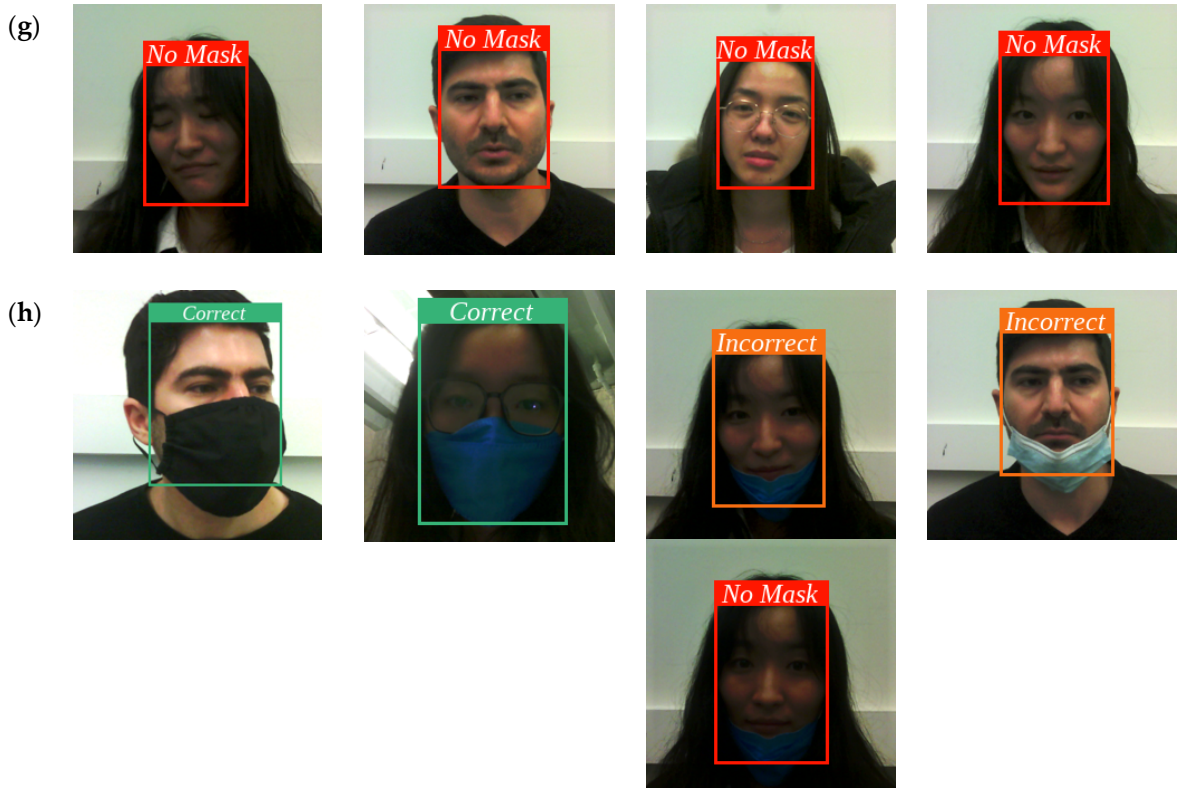

Figure 8. Mask classification results on images taken by Pepper under varying conditions. Incorrect = face with incorrectly worn mask/face covered by other means, Correct = face with correctly worn mask, and No Mask = face without mask. (**a**–**g**) presents example results, (**h**) presents different lighting conditions. Each column is denoted by 1–4.

We achieved a classification accuracy of 91.27% for Classifier 1 and 86.27% for Classifier 2, respectively. The overall accuracy for the three mask classes was 84.13%. Our experimental results are higher than the evaluation accuracy of 81.31% achieved for the only reported robot-based binary mask classification system [15].

6.4. Comparison of Single-Step and Two-Step Methods

We further compared our two-step, extended mask detection and classification ResNet-50 method (Classifiers 1 and 2) with other classifiers to highlight its performance during real-time robot deployment. Namely, we compared our method with (1) a method consisting of only a single extended ResNet-50 classifier and (2) the Super Resolution and Classification Network (SRCNet) method in [23], which is based on MobileNet-v2. For both these methods, a single classifier was used to determine the aforementioned three classes. These methods were trained using the datasets and training procedure in Section 4. Testing was performed using the same 126 images obtained by the Pepper robot. Accuracy, precision, recall, and F1 scores for all three methods are presented in Table 4. Figure 9 shows the confusion matrix for these methods.

Table 4. Accuracy, precision, recall, and F1 scores for: a single extended ResNet-50 classifier, our two-step extended ResNet-50 method, and SRCNet.

	Single Extended ResNet50 Classifier	Our Two-Step Extended ResNet50 Method (Classifier 1 and 2)	SRCNet
Accuracy	77.78%	84.13%	65.87%
Precision	76.32%	82.23%	65.95%
Recall	74.58%	83.96%	64.44%
F1 score	75.27%	82.75%	64.15%

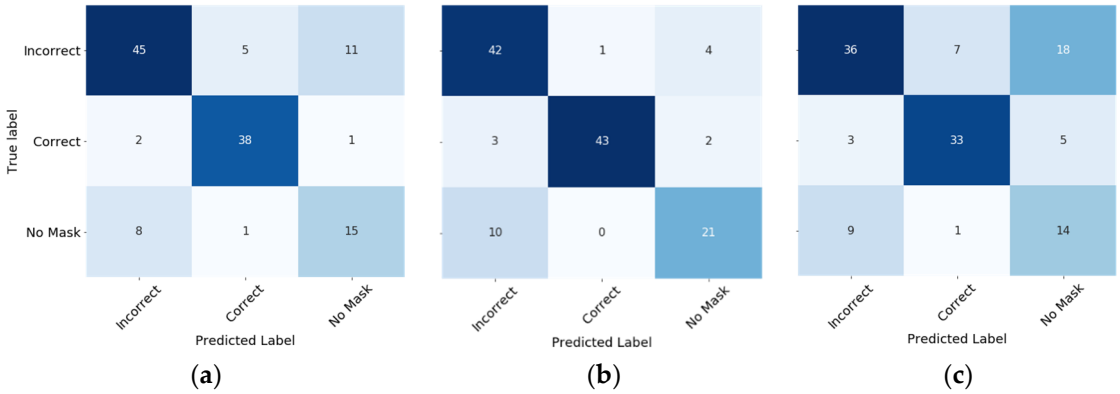

Figure 9. Confusion matrix comparison for (**a**) single extended ResNet-50 classifier, (**b**) our two-step extended ResNet-50 method, and (**c**) SRCNet.

Our two-step extended ResNet-50 method outperformed both methods on all four metrics for the classification of the three mask classes. The main difference in classification between all three methods was for the incorrectly-worn-mask class, where the single extended ResNet-50 classifier and SRCNet both had higher instances of wrongly predicting this class as the no-mask class or correctly-worn-mask class. Our method was more robust for this type of mask class.

7. Conclusions

In this paper, we present a novel, two-step, deep learning mask detection and classification system for socially assistive robots. The deep learning structure is based on our unique extended ResNet-50 structure for real-time classification of three classes: face with no mask, face with correctly worn mask, and face with incorrectly worn mask/face covered by other means. We obtained an overall training and validation accuracy of 97.8% and 97.3%, and a testing accuracy of 97.66%, on an unseen dataset under varying face and lighting conditions. The trained deep CNN models were integrated using the Pepper robot for real-world experiments with people under varying conditions, including different lighting situations, head poses, and face occlusions. Experimental results using the robot's onboard low-resolution camera showed a classification accuracy of 84.13%, outperforming single-classifier methods. The presented results are promising, as our system outperformed existing approaches. Our future work will consist of further testing and analysis, such as performing multiple training repetitions and optimization over a range of epochs. We will deploy the robot in larger public places for autonomous mask detection of crowds of people. This includes entrances at our university, local hospitals, and healthcare centers.

Author Contributions: Conceptualization, Y.Z., M.E. and G.N.; methodology, Y.Z., M.E. and G.N.; software, Y.Z.; validation, Y.Z., M.E. and G.N.; formal analysis, Y.Z. and M.E.; investigation, Y.Z. and M.E.; resources, G.N.; data curation, Y.Z. and M.E.; writing—original draft preparation, Y.Z., M.E., A.H.T. and G.N.; writing—review and editing, A.H.T. and G.N.; visualization, Y.Z., M.E., A.H.T. and

G.N.; supervision, G.N.; project administration, G.N.; funding acquisition, G.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by AGE-WELL Inc., the Canada Research Chairs (CRC) Program, the Natural Sciences and Engineering Research Council of Canada (NSERC) and NSERC CREATE HeRo fellowship.

Institutional Review Board Statement: This study is part of a research project that received approval from the University of Toronto Ethics Committee (ethics protocol code 41193 and approval date of 2022-11-21).

Informed Consent Statement: Informed consent was obtained from all individual participants included in the study and whose images were used in the manuscript.

Data Availability Statement: The public datasets used in this paper are available in the public domain. The MaskedFace dataset is available under license at https://creativecommons.org/licenses/by-nc-sa/4.0/, accessed on 1 September 2022, and the images used in Figure 3 are from the authors Dorian, rdoroshenko, Yada Liao, and Senterpartiet (Sp), and are available under license: https://creativecommons.org/licenses/by/2.0/, accessed on 1 September 2022. Written permission was obtained for the images in Figure 3 from the MAFA and CelebA datasets.

Acknowledgments: We would like to thank our study participants.

Conflicts of Interest: The authors declare that this study received funding from AGE-WELL Inc. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication. The authors declare no conflicts of interest.

References

1. World Health Organization. Coronavirus Disease (COVID-19). Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019?gclid=CjwKCAjw_L6LBhBb%2520EiwA4c46umlkirJ1KOaEq4v4SAUw8blELjV2hpge91FIa33ZFPae7WaxShqzBoCOmQQAvD_B%2520wE (accessed on 6 April 2022).
2. Savela, N.; Latikka, R.; Oksa, R.; Kortelainen, S.; Oksanen, A. Affective Attitudes Toward Robots at Work: A Population-Wide Four-Wave Survey Study. *Int. J. Soc. Robot.* **2022**, *14*, 1379–1395. [CrossRef] [PubMed]
3. (Sam) Kim, S.; Kim, J.; Badu-Baiden, F.; Giroux, M.; Choi, Y. Preference for robot service or human service in hotels? Impacts of the COVID-19 pandemic. *Int. J. Hosp. Manag.* **2021**, *93*, 102795.
4. Sathyamoorthy, A.J.; Patel, U.; Paul, M.; Savle, Y.; Manocha, D. COVID surveillance robot: Monitoring social distancing constraints in indoor scenarios. *PLoS ONE* **2021**, *16*, e0259713. [CrossRef] [PubMed]
5. Shah, S.H.H.; Steinnes, O.M.H.; Gustafsson, E.G.; Hameed, I.A. Multi-agent robot system to monitor and enforce physical distancing constraints in large areas to combat COVID-19 and future pandemics. *Appl. Sci.* **2021**, *11*, 7200. [CrossRef]
6. Turja, T.; Taipale, S.; Niemelä, M.; Oinas, T. Positive Turn in Elder-Care Workers' Views Toward Telecare Robots. *Int. J. Soc. Robot.* **2022**, *14*, 931–944. [CrossRef]
7. Getson, C.; Nejat, G. The adoption of socially assistive robots for long-term care: During COVID-19 and in a post-pandemic society. *Healthc. Manag. Forum* **2022**, *35*, 301–309. [CrossRef]
8. Augello, A.; Città, G.; Gentile, M.; Lieto, A. A Storytelling Robot Managing Persuasive and Ethical Stances via ACT-R: An Exploratory Study. *Int. J. Soc. Robot.* **2021**, *15*, 2115–2131. [CrossRef]
9. World Health Organization. Advice for the Public on COVID-19. Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public (accessed on 16 October 2021).
10. Ueki, H.; Furusawa, Y.; Iwatsuki-Horimoto, K.; Imai, M.; Kabata, H.; Nishimura, H.; Kawaoka, Y. Effectiveness of Face Masks in Preventing Airborne Transmission of SARS-CoV-2. *mSphere* **2020**, *5*, e00637-20. [CrossRef]
11. Brooks, J.T.; Butler, J.C. Effectiveness ofMaskWearing toControl Community Spread ofSARS-CoV-2. *Ann. Intern. Med.* **2021**, *174*, 335–343.
12. Centers for Disease Control and Prevention. Masks and Respirators. Available online: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/types-of-masks.html (accessed on 1 August 2022).
13. YouTube. Hikvision Mask Detection Solution. Available online: https://www.youtube.com/watch?v=FagQhPkrrws (accessed on 6 April 2022).
14. TRISTATE LOW VOLTAGE SUPPLY. Covid-19 Tablet Face, Mask, and Temperature Detection. Available online: https://tristatetelecom.com/productdetailI2.aspx?dataid=NEXUS (accessed on 6 April 2022).
15. Snyder, S.E.; Husari, G. Thor: A deep learning approach for face mask detection to prevent the COVID-19 pandemic. In Proceedings of the SoutheastCon 2021, Atlanta, GA, USA, 10–13 March 2021.
16. Li, Y.; Yan, J.; Hu, B. Mask Detection Based on Efficient-YOLO. In Proceedings of the 2021 40th Chinese Control Conference (CCC), Shanghai, China, 26–28 July 2021; pp. 4056–4061.

17. Sierra M, S.D.; Sergio, D.; Insuasty, M.; Daniel, D.E.; Munera, M.; Cifuentes, C.A. Improving Health and Safety Promotion with a Robotic Tool: A Case Study for Face Mask Detection. In Proceedings of the Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, Stockholm, Sweden, 13–16 March 2023; pp. 716–719.
18. SoftBank Robotics. New Feature: Pepper Mask Detection. Available online: https://www.softbankrobotics.com/emea/en/blog/news-trends/new-feature-pepper-mask-detection (accessed on 6 April 2022).
19. SMP Robotics-Autonomous Mobile Robot. Face Mask Detection Robot with a Voice Warning of a Fine for Not Wearing It in the Public Area. Available online: https://smprobotics.com/usa/face-mask-detection-robot/ (accessed on 6 April 2022).
20. Vibhandik, H.; Kale, S.; Shende, S.; Goudar, M. Medical Assistance Robot with capabilities of Mask Detection with Automatic Sanitization and Social Distancing Detection/Awareness. In Proceedings of the 6th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 1–3 December 2022; pp. 340–347.
21. Putro, M.D.; Nguyen, D.L.; Jo, K.H. *Real-Time Multi-View Face Mask Detector on Edge Device for Supporting Service Robots in the COVID-19 Pandemic*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; Volume 12672.
22. Li, C.; Wang, R.; Li, J.; Fei, L. *Face Detection Based on YOLOv3*; AISC; Springer: Singapore, 2020; Volume 1031.
23. Bosheng, Q.; Li, D. Identifying Facemask-Wearing Condition Using Image Super-Resolution with Classification Network. *Sensors* **2020**, *20*, 5236.
24. Loey, M.; Manogaran, G.; Taha, M.H.N.; Khalifa, N.E.M. Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. *Sustain. Cities Soc.* **2021**, *65*, 102600. [CrossRef] [PubMed]
25. Gupta, S.; Sreenivasu, S.V.N.; Chouhan, K.; Shrivastava, A.; Sahu, B.; Manohar Potdar, R. Novel Face Mask Detection Technique using Machine Learning to control COVID'19 pandemic. *Mater. Today Proc.* **2023**, *80*, 3714–3718. [CrossRef] [PubMed]
26. Boulila, W.; Alzahem, A.; Almoudi, A.; Afifi, M.; Alturki, I.; Driss, M. A Deep Learning-based Approach for Real-time Facemask Detection. In Proceedings of the 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Pasadena, CA, USA, 13–16 December 2021; pp. 1478–1481.
27. Teboulbi, S.; Messaoud, S.; Hajjaji, M.A.; Mtibaa, A. Real-Time Implementation of AI-Based Face Mask Detection and Social Distancing Measuring System for COVID-19 Prevention. *Sci. Program.* **2021**, *2021*, 8340779. [CrossRef]
28. Walia, I.S.; Kumar, D.; Sharma, K.; Hemanth, J.D.; Popescu, D.E. An integrated approach for monitoring social distancing and face mask detection using stacked Resnet-50 and YOLOv5. *Electronics* **2021**, *10*, 2996. [CrossRef]
29. Sethi, S.; Kathuria, M.; Kaushik, T. Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread. *J. Biomed. Inform.* **2021**, *120*, 103848. [CrossRef] [PubMed]
30. Kowalczyk, N.; Sobotka, M. Mask Detection and Classification in Thermal Face Images. *IEEE Access* **2023**, *11*, 43349–43359. [CrossRef]
31. Fan, X.; Jiang, M. RetinaFaceMask: A Single Stage Face Mask Detector for Assisting Control of the COVID-19 Pandemic. In Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Melbourne, Australia, 17–20 October 2021; pp. 832–837.
32. Canada. P. H. A. of Canada: Government of Canada. Available online: https://www.canada.ca/en/public-health/services/publications/diseases-conditions/covid-19-mask-fit-properly.html (accessed on 6 April 2022).
33. Nieto-Rodriguez, A.; Mucientes, M.; Brea, V.M. System for medical mask detection in the Operating Room Through Facial Attributes. In *Pattern Recognition and Image Analysis, Proceedings of the 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, 17–19 June 2015*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9117, pp. 138–145.
34. Viola, P.; Jones, M.J. Robust Real-Time Face Detection. *Int. J. Comput. Vis.* **2004**, *57*, 137–154. [CrossRef]
35. Huang, G.B.; Mattar, M.; Berg, T.; Learned-Miller, E.; Learned-Miller, E. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. *Eccv2008*. 2008. Available online: https://inria.hal.science/inria-00321923/document (accessed on 1 September 2022).
36. Rowley, H.A.; Member, S.; Baluja, S.; Kanade, T. Neural Network-Based Face Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 23–38. [CrossRef]
37. Frischholz, R. Bao Face Database at the Face Detection Homepage. 2012. Available online: https://facedetection.com/ (accessed on 1 August 2022).
38. Ejaz, M.S.; Islam, M.R.; Sifatullah, M.; Sarker, A. Implementation of Principal Component Analysis on Masked and Non-masked Face Recognition. In Proceedings of the 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 3–5 May 2019; Volume 2019, pp. 1–5.
39. ORL (Our Database of Faces). Available online: https://paperswithcode.com/dataset/orl (accessed on 2 January 2022).
40. Ge, S.; Li, J.; Ye, Q.; Luo, Z. Detecting masked faces in the wild with LLE-CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 426–434.
41. Putro, M.D.; Jo, K.H. Fast Face-CPU: A Real-time Fast Face Detector on CPU Using Deep Learning. In Proceedings of the 29th International Symposium on Industrial Electronics (ISIE), Delft, The Netherlands, 17–19 June 2020; pp. 55–60.
42. Yang, S.; Luo, P.; Loy, C.C.; Tang, X. WIDER FACE: A face detection benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5525–5533.
43. Wang, Z.; Huang, B.; Wang, G.; Yi, P.; Jiang, K. Masked Face Recognition Dataset and Application. *IEEE Trans. Biom. Behav. Identity Sci.* **2023**, *5*, 298–304. [CrossRef]

44. Jain, V.; Learned-Miller, E. *Fddb: A Benchmark for Face Detection in Unconstrained Settings*; UMass Amherst Technical Report; University of Massachusetts: Amherst, MA, USA, 2010.
45. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
46. Loey, M.; Manogaran, G.; Taha, M.H.N.; Khalifa, N.E.M. A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. *Meas. J. Int. Meas. Confed.* **2021**, *167*, 108288. [CrossRef]
47. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
48. Medical Mask Dataset: Humans in the Loop. Available online: https://humansintheloop.org/resources/datasets/medical-mask-dataset/ (accessed on 18 February 2022).
49. Larxel. Face Mask Detection Dataset. Available online: https://www.kaggle.com/datasets/andrewmvd/face-mask-detection (accessed on 2 January 2022).
50. Cabani, A.; Hammoudi, K.; Benhabiles, H.; Melkemi, M. MaskedFace-Net—A dataset of correctly/incorrectly masked face images in the context of COVID-19. *Smart Health* **2021**, *19*, 100144. [CrossRef] [PubMed]
51. Face Mask Detection. Available online: https://www.kaggle.com/andrewmvd/face-mask-detection (accessed on 2 January 2022).
52. Zereen, A.N.; Corraya, S.; Dailey, M.N.; Ekpanyapong, M. *Two-Stage Facial Mask Detection Model for Indoor Environments*; Springer: Singapore, 2021; Volume 1309.
53. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017; pp. 3645–3649.
54. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
55. Shrestha, H.; Megha, S.; Chakraborty, S.; Mazzara, M.; Kotorov, I. *Face Mask Recognition Based on Two-Stage Detector*; LNNS; Springer Nature: Cham, Switzerland, 2023; Volume 715.
56. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. LNCS 8693-Microsoft COCO: Common Objects in Context. In *Computer Vision–ECCV 2014, Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
57. Custom Mask Community Dataset. Available online: https://github.com/prajnasb/observations (accessed on 2 January 2022).
58. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
59. Chiang, D. Detect Faces and Determine Whether People Are Wearing Mask. 2020. Available online: https://github.com/AIZOOTech/FaceMaskDetection (accessed on 1 September 2022).
60. Aldebaran 2.0.6.8 Documentation. Pepper-2D Cameras-Aldebaran 2.0.6.8 Documentation. Available online: http://doc.aldebaran.com/2-0/family/juliette_technical/video_juliette.html#d-camera-juliette (accessed on 18 April 2022).
61. Lienhart, R.; Maydt, J. An extended set of Haar-like features for rapid object detection. *IEEE Int. Conf. Image Process.* **2002**, *1*, 900–903.
62. Srinivas, M.; Bharath, R.; Rajalakshmi, P.; Mohan, C.K. Multi-level classification: A generic classification method for medical datasets. In Proceedings of the 17th International Conference on E-health Networking, Application & Services (HealthCom), Boston, MA, USA, 14–17 October 2015; pp. 262–267.
63. Lohia, A.; Kadam, K.D.; Joshi, R.R.; Bongale, A.M. Bibliometric Analysis of One-stage and Two-stage Object Detection. *Libr. Philos. Pract.* **2021**, *2021*, 1–33.
64. Deng, Z.; Cao, M.; Rai, L.; Gao, W. A two-stage classification method for borehole-wall images with support vector machine. *PLoS ONE* **2018**, *13*, e0199749. [CrossRef]
65. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1717–1724.
66. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
67. Agarap, A.F. Deep Learning using Rectified Linear Units (ReLU). *arXiv* **2018**, arXiv:1803.08375.
68. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
69. Hinton, G.; Osindero, S.; Yee-Whye, T. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *1554*, 1527–1554. [CrossRef]
70. Zhang, Z.; Sabuncu, M.R. Generalized cross entropy loss for training deep neural networks with noisy labels. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 8778–8788.
71. SoftBank Robotics-Group. Downloads Softwares: Softbank Robotics. Available online: https://www.softbankrobotics.com/emea/en/support/pepper-naoqi-2-9/downloads-softwares (accessed on 13 March 2022).
72. NAOqi APIs-Aldebaran 2.4.3.28-r2 Documentation. Aldebaran Documentation What's New in Naoqi 2.4.3? Available online: http://doc.aldebaran.com/2-4/naoqi/index.html (accessed on 13 March 2022).

73. ALVideoDevice-Aldebaran 2.1.4.13 Documentation. Aldebaran Documentation. Available online: http://doc.aldebaran.com/2-1/naoqi/vision/alvideodevice.html (accessed on 13 March 2022).
74. Image Module-Pillow (PIL Fork) 9.0.1 Documentation. Image Module. Available online: https://pillow.readthedocs.io/en/stable/reference/Image.html (accessed on 13 March 2022).