Face Mask Detection Using CNN: Final Report

Team 12

Bao Tian Fu - 1004207119
Yu Chen (Jordan) Liu - 1004077973
Shu Cheng (Alex) Yeh - 1003959335
Shuocheng (Shirley) Zhang - 1003906213

Word count: 2452

1.0 Introduction

Since the start of the COVID-19 pandemic, the medical community has been recommending universal mask wearing as a measure to prevent the spread of the disease [1]. Over the months, a substantial increase in scientific evidence has led to more wide-spread recognition and adoption of mask wearing policies [2]. In addition, different mask types are often recommended under a variety of settings. For example, most supermarkets allow customers to wear homemade or surgical masks, while N95 respirators are usually reserved for healthcare providers [3].

Currently, the enforcement of most mask mandates is performed manually by front-desk staff or security. The motivation behind this project is to automate the screening process by developing a real-time mask detection model that not only identifies whether a mask is worn properly, but also the type of mask being worn (Figure 1). This model can also be used to support COVID-19 research by determining whether the degree of disease transmission in a public setting is influenced by the different types of masks present. Because machine learning can learn relevant features from examples and detect classes of objects in images, it is a suitable tool for this application.

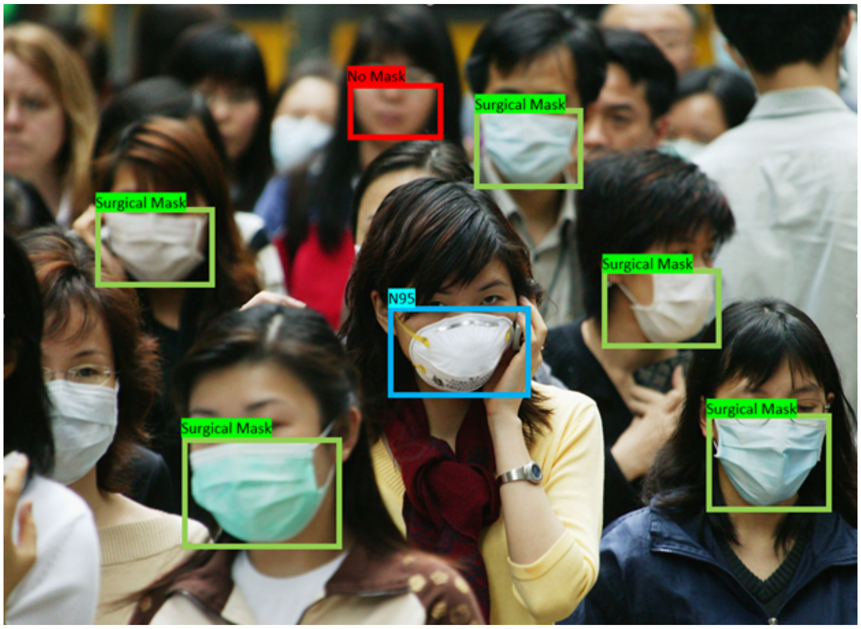


Fig. 1: Sample model output that detects and labels different mask-wearing status

2.0 Illustration

Figure 2 describes the pipeline constructed. A Mask Wearing Model is used to detect if the test subject's mouth, nose, and chin are covered, which suggests correctly worn masks [2]. If the mask is worn correctly, the image is passed onto a Mask Type Model, which detects the type of mask being worn. Any coverage of nose and lip from the subject besides N95 or Surgical will be classified as Other in the Mask Type Model.

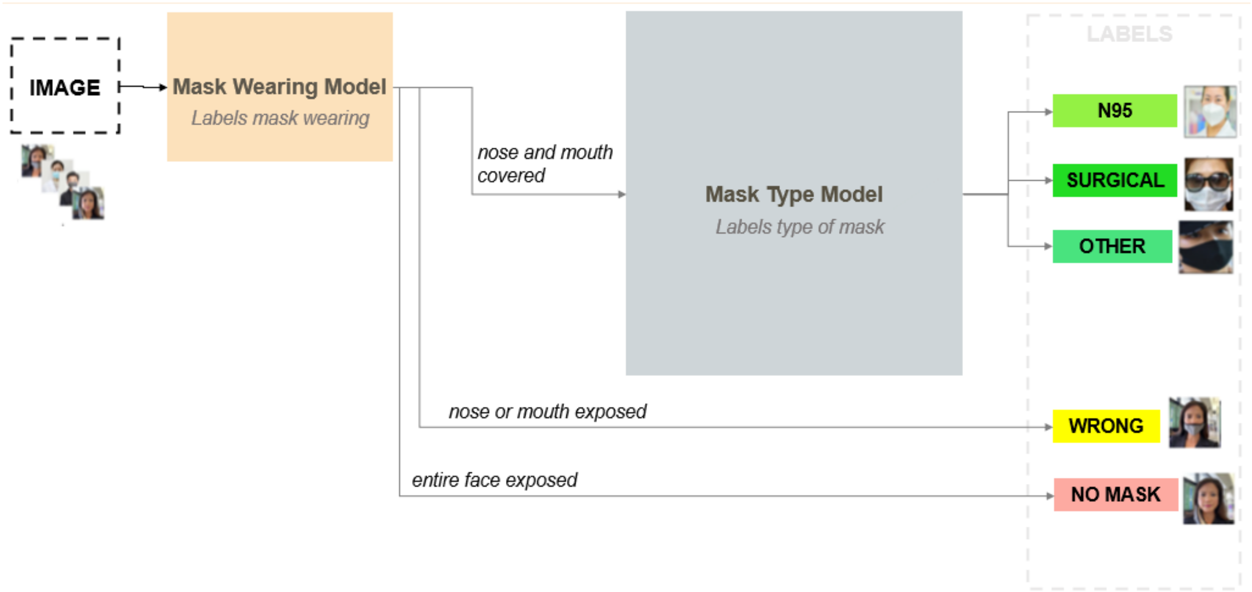


Fig. 2. An illustration of the mask detection workflow

3.0 Background & Related Work

Given the significance of mask-wearing in the past year, a considerable amount of work has been done to build and improve machine learning models for face mask detection. One 2020 study used a Deep Neural Network Single Shot Multibox Detector for feature detection and a pretrained Mobilenet V2 classifier for image classification. This SSDMNV2 model detects face masks on front-facing faces with an accuracy of 92.64%, and it can be used in embedded devices for real-time applications [3]. Another approach proposed in 2020 used the CNN ResNet-50 for feature extraction and YOLO v2 for face mask detection. With the help of anchor boxes that support the detection of multiple overlapping objects, This learning model can recognize medical mask-wearing faces with an average precision (AR) of 81% [4].

Although both models can effectively differentiate masked faces from unmasked faces, they are unable to classify the type of mask being worn. Therefore, this project's goal is to fill the gap by developing models that can accomplish both tasks.

4.0 Data Processing

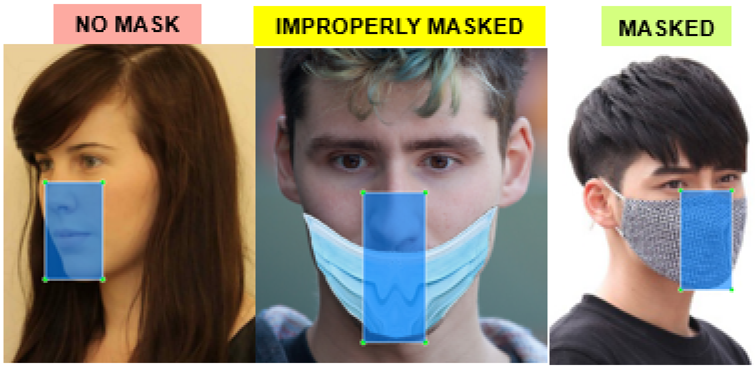
Raw images were collected from the 2020 MaskedFace-Net Dataset[5][6], the 2020 SSDMNV2 Model Dataset [7], the 2021 FMR-DB Dataset [8], and the ShutterStock repository. These datasets had thousands of images, and images were already classified into mask-wearing and non-mask-wearing categories. The drawback was that they were greatly unbalanced; there were very few “N95” mask images and “incorrectly worn” mask images. To increase the amount of training samples, artificial images: data with pictures of masks overlaid on raw exposed-face images were generated.

The raw data was then cleaned by discarding any unusable images that meets any one of the following criterias:

- Blurry image
- Over 80% of the face was cropped out
- Over 80% of the mask is not visible
- Type of mask cannot be identified by a human

For single-subject images, images containing more than one person were cropped to have only one person in the frame. To support detection of multiple masks in a frame, 100 multiple-subject images were selected and placed in a separate folder.

Next, ground truth was labeled manually using the labelImg [9] tool. The ground truth consisted of two parts: bounding boxes around the regions of interest and a text label matching either the type of mask worn or how the mask is being worn (correct, incorrect, or no mask). Note that to obtain the best performance for the Mask Wearing Model, the region of interest can only be limited to the subject’s nose to the mouth.



Single subject

Fig. 3. Sample of Mask Wearing Model data ground truth

For the Mask Type Model, the region of interest was defined as the visible area of the mask. The type of mask was also labeled manually using human judgement, due to the raw images not having these labels.

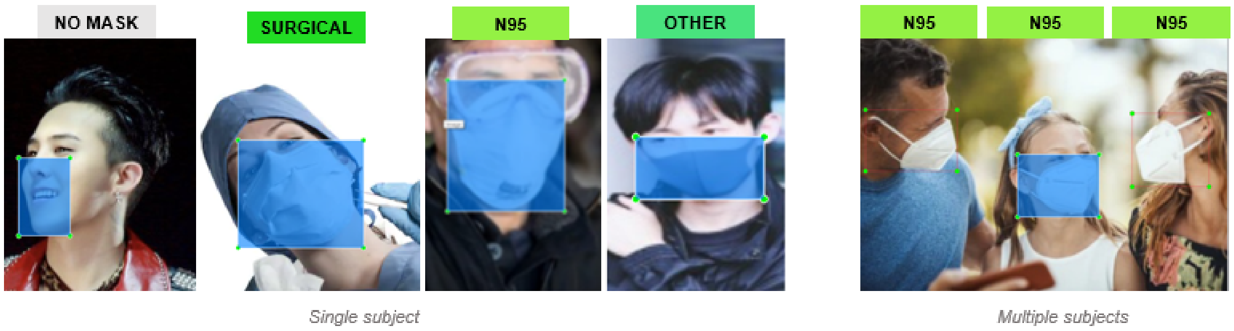


Fig. 4. Sample of Mask Type Model data ground truth

Due to the imbalance in image quantity, larger classes were undersampled to create a final balanced dataset. To mitigate bias from undersampling, a python script was written and used to shuffle each folder's data randomly prior to selection.

The Mask Wearing and Mask Type datasets were split into training, validation, and test sets at a 51-13-36, and 60-20-20 ratio respectively.

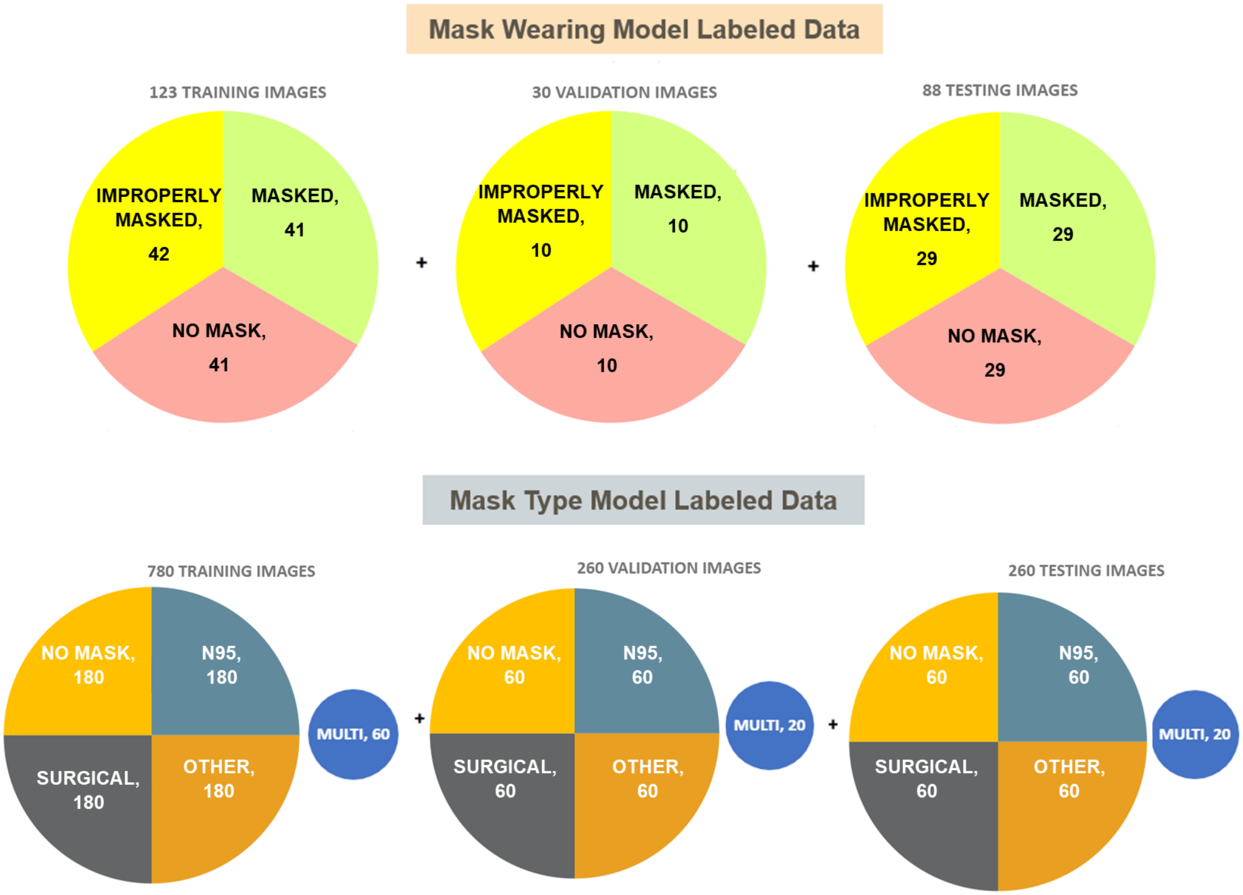


Fig. 5. Breakdown of the final dataset

The final dataset was balanced; containing a total of 241 ready-to-use images for training, validation, and testing purposes for the Mask Wearing Model, and 1300 images for the Mask Type Model.

5.0 Architecture

The team implemented a SSD-MobileNetV2 model for both Mask Wearing and Mask Type detection tasks (Figure 6). MobileNetV2 is a CNN classification model that replaces traditional convolutional layers with depth-wise separable convolutions which are much faster and more efficient [10]. It also uses bottleneck layers to uncompress, filter, and recompress the data to extract more information with fewer computations. In addition, it uses residual connections to help with the flow of gradients through the network. Each layer also utilizes batch normalization and an activation function of ReLU6. The full MobileNet V2 architecture consists of 17 building blocks followed by a regular 1×1 convolution, a global average pooling layer, and a classification layer.

To detect the location of masks in addition to classifying them, the team combined MobileNetV2 with a Single Shot MultiBox Detector (SSD) [11]. SSD builds upon features that are extracted by MobileNetV2 to further predict bounding box locations as well as class probabilities. Furthermore, SSD combines predictions from multiple feature maps to detect objects of various sizes. In addition, SSD-MobileNetV2 is a light-weight model that can run on mobile devices with real-time results, which will allow the team to implement the model as an iOS app.

The team downloaded a pretrained model with a total of 267 layers from the Tensorflow Object Detection API, and applied transfer learning techniques to tune the hyper-parameters as well as train the model with custom datasets. The final model used Relu_6 activation function with learning rate of 0.0008 and a batch size of 24. The team also used RandomCrop as an augmentation technique and dropout to counter overfitting.

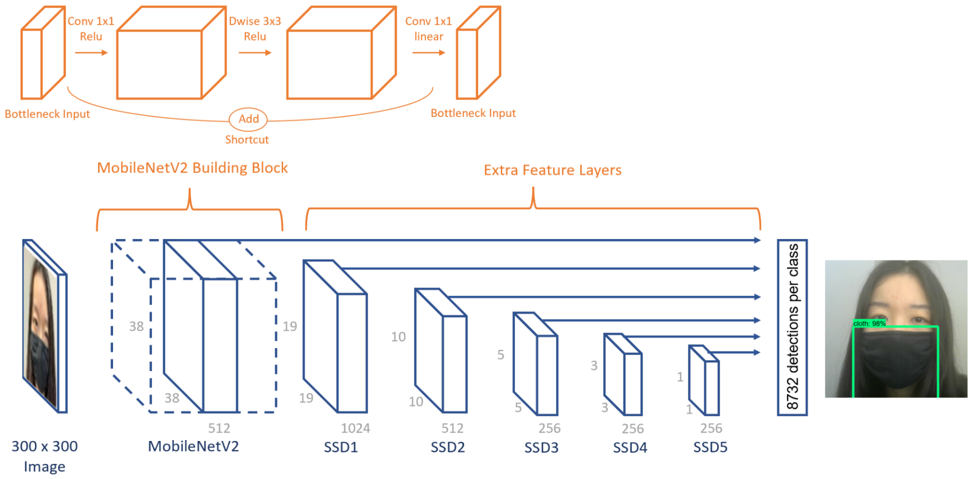


Fig. 6. An illustration of the SSD-MobileNetV2 architecture

6.0 Baseline Model

The baseline of the Mask Wearing Model is a three-layer CNN classification model. The model was able to achieve a 70% testing accuracy with a precision score of 0.73 and a recall score of 0.64. However, the team observed that the model could not efficiently classify whether the mask was worn correctly or not as shown in Figure 7, which was an essential requirement for the primary model.

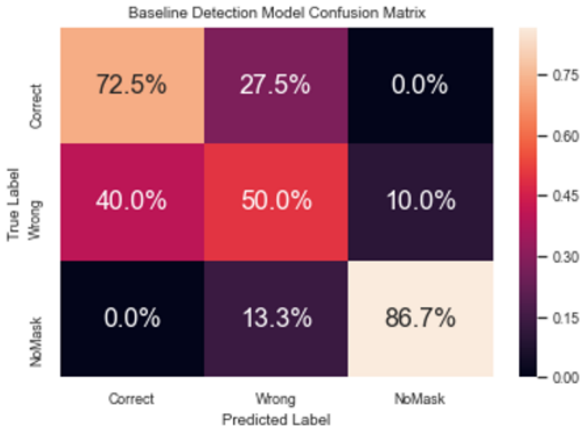


Fig. 7: Confusion matrix of the baseline for the Mask Wearing Model

The baseline for the Mask Type Model was first designed as a CNN multilayer classification model. Due to its simplicity, the resulting test accuracy was 55%, which was inadequate even for a baseline. Afterwards, the team tried transfer learning using AlexNet followed by a simple 3-layer CNN, and the model test accuracy increased to 70%. This baseline model was selected due to its simplicity in architecture compared to the primary model, while holding strong performance in classification implementations. This model has a precision score of 0.70 and a recall score of 0.70.

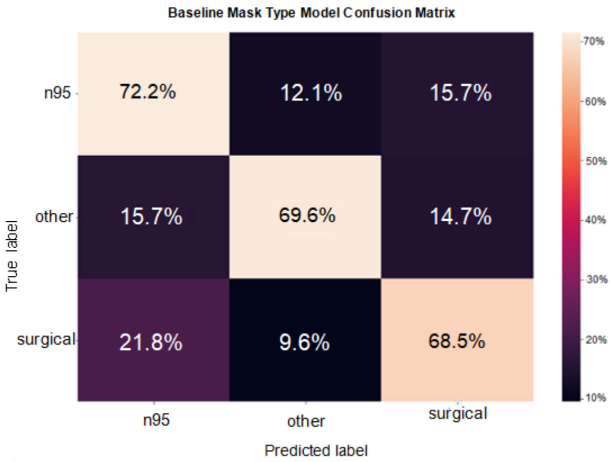


Fig 8. Confusion matrix of the baseline for the Mask Type Model

7.0 Quantitative Results

The performance of the primary Mask Wearing Model can be seen from the confusion matrix shown in Figure 9. In comparison to the baseline model confusion matrix shown in Figure 7, the primary model was able to differentiate correctly worn from incorrectly worn more efficiently and precisely. Additionally, the primary model achieved an average precision score of 0.9 and an average recall score of 0.83, which was a 17% and 19% improvement from the baseline.

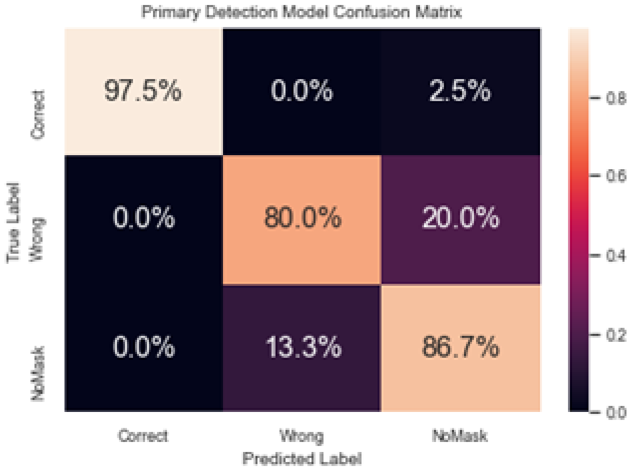


Fig. 9. Confusion Matrix of the Primary Mask Detection Model.

Simultaneously, the training of the Mask Type Model was terminated after 14,000 steps when the validation and training losses started to plateau (Figure 10). The final model achieved a Mean Average Precision (mAP) of 0.82 and an Average Recall (AR) score of 0.84 (Figure 11), a 0.12

to 0.14 improvement from the baseline 3-layer CNN with AlexNet model mentioned above. It should be noted that the team chose mAP and AR as the evaluation metrics because they represent an average over all classes and/or Intersection over Union (IoU) thresholds, which is especially suitable for quantifying object detection performance.

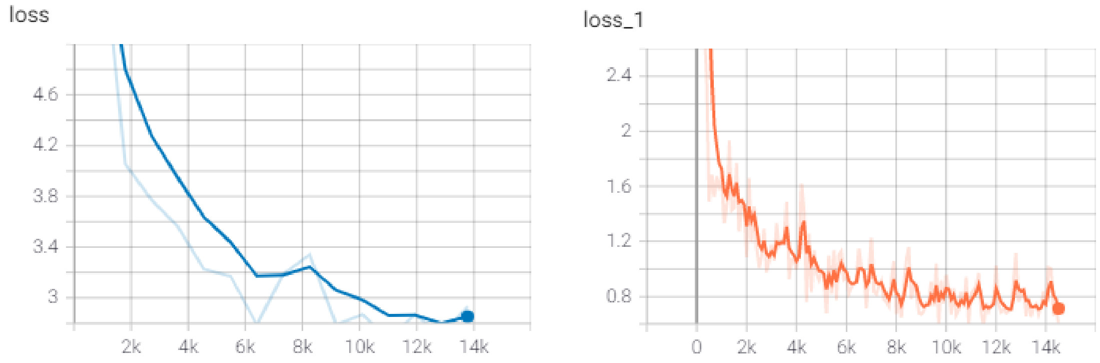


Fig. 10. Validation (left) and training (right) loss for the Mask Type Model

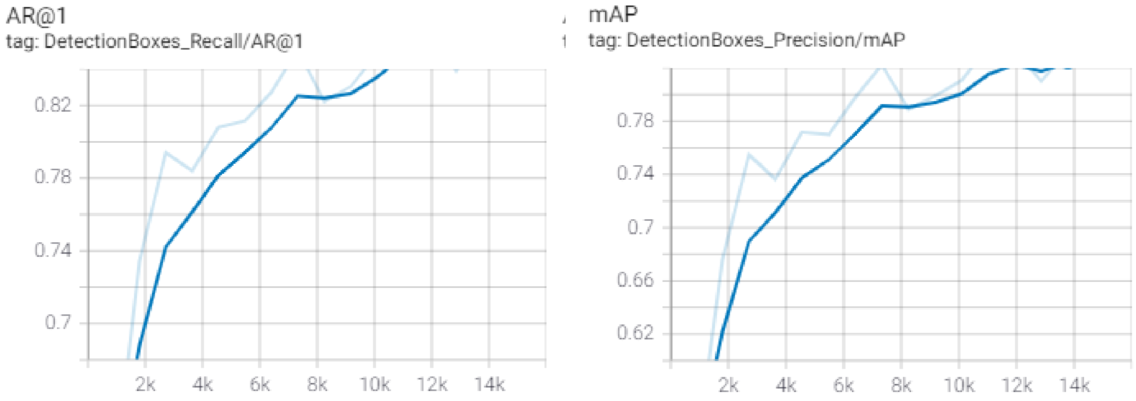


Fig. 11. Average recall (left) and precision (right) curve for the Mask Type Model

9.0 Qualitative Results

The Mask Wearing Model exhibits excellent performance when connected to a live webcam. As shown in Figure 12 and Figure 13, the model shows consistent accuracy for multiple faces, varying positions, and different mask-wearing styles.

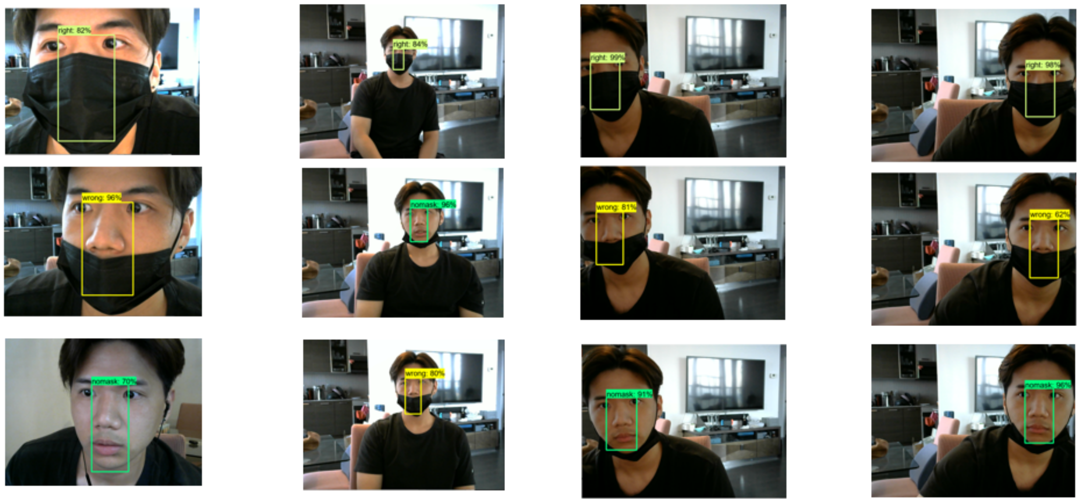


Fig. 12. Detection results of Mask Wearing Model with different positions of the subject.

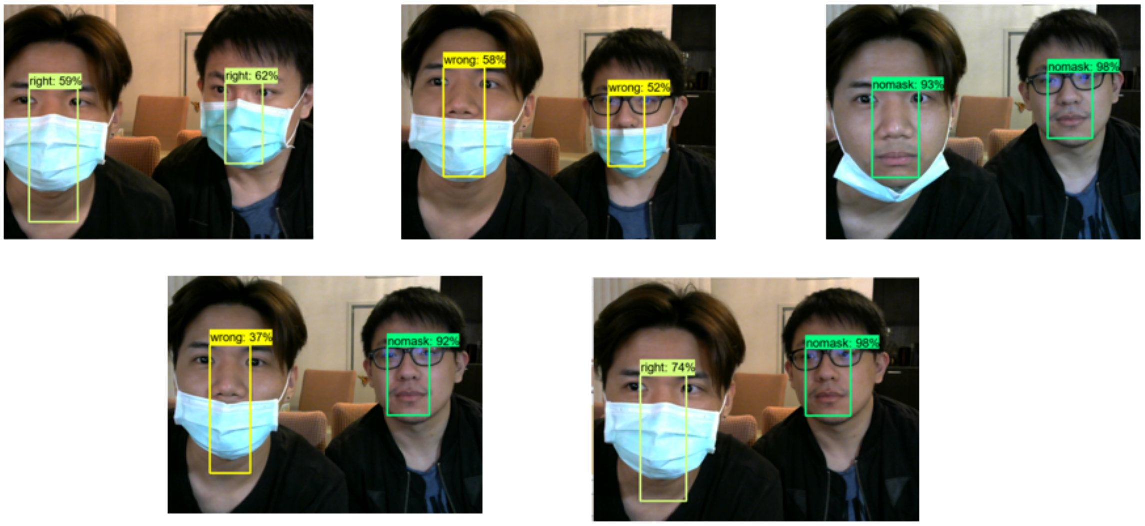


Fig. 13. Detection results of Mask Wearing Model with different number of subjects.

Similarly, the Mask Type Model performed very well on static images as illustrated in Figure 14. Bounding box locations and mask types are detected accurately and with high confidence on multiple faces at different angles and scaling. The team also tested the model on various edge cases and noticed two specific scenarios that are worth noting. In the top left image, the model performed extremely well by detecting overlapping N95 and surgical masks on the same face. In the bottom left image, the model did not perform as well and did not detect the mask wearing faces in the background, perhaps because those faces are too blurry and not front-facing enough.

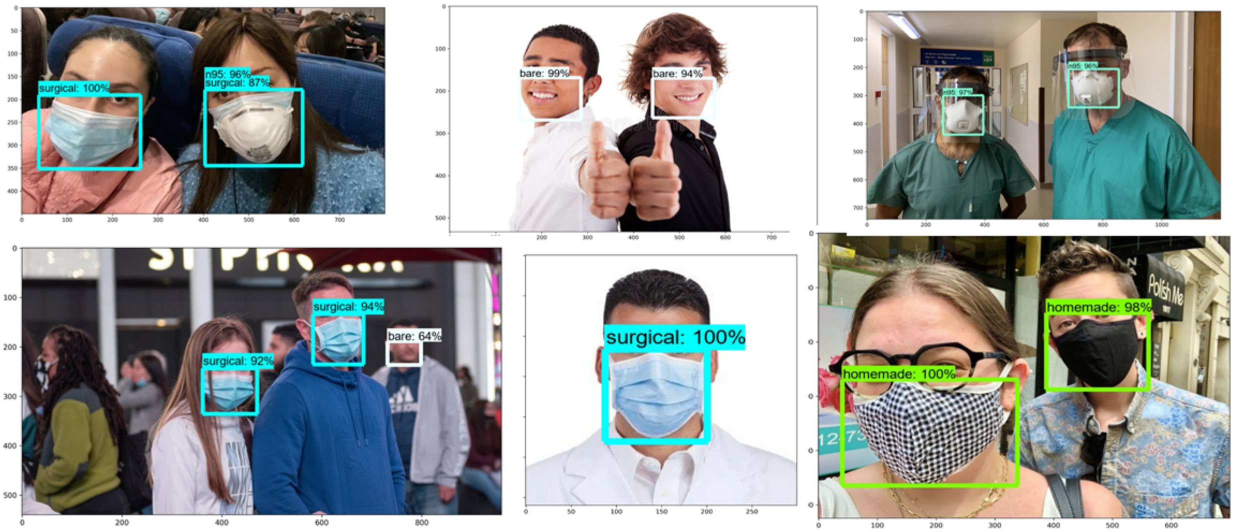


Fig. 14: Detection results of Mask Type Model on static images

Similar to the Mask Wearing model, the Mask Type model can accurately detect multiple types of masks in a live webcam as shown in Figure 15. A frame rate of over 30 FPS allows for real-time detection with minimum lagging.

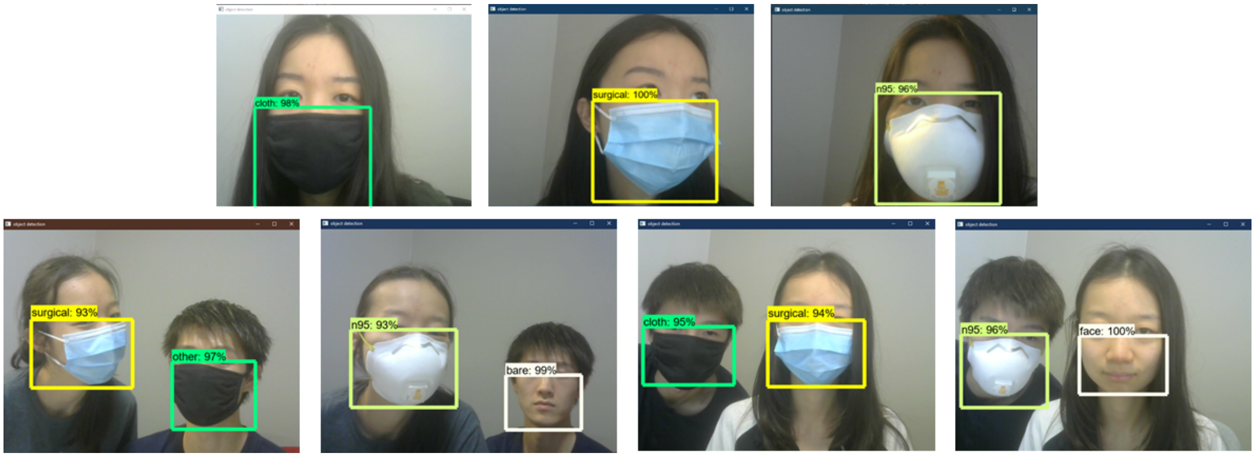


Fig. 15. Live detection results of Mask Type Model

10.0 Evaluation Results on New Data

The goal of this project was to develop a tool that replaces manual screening to enforce mask mandates. Therefore, the consistency in performance was critical especially on edge cases where masks could appear in different locations, sizes, and angles in the image. As discussed in the previous section, live video from a webcam was fed to the model frame by frame, and the team evaluated the model's performance on new data. The team paid extra attention to the model's performance on edge cases, and such cases with incorrect predictions were added to the training dataset. The model was then trained with the updated dataset. Executing this process iteratively, the team was able to eliminate most edge cases, which maintained the model's performance on never-seen data.

Furthermore, the team designed an iOS application to evaluate the model's performance in actual use cases of enforcing mask mandates(Figure 16, Figure 17). The application can also collect feedback from a broader demographic, which may include edge cases that the team failed to consider.

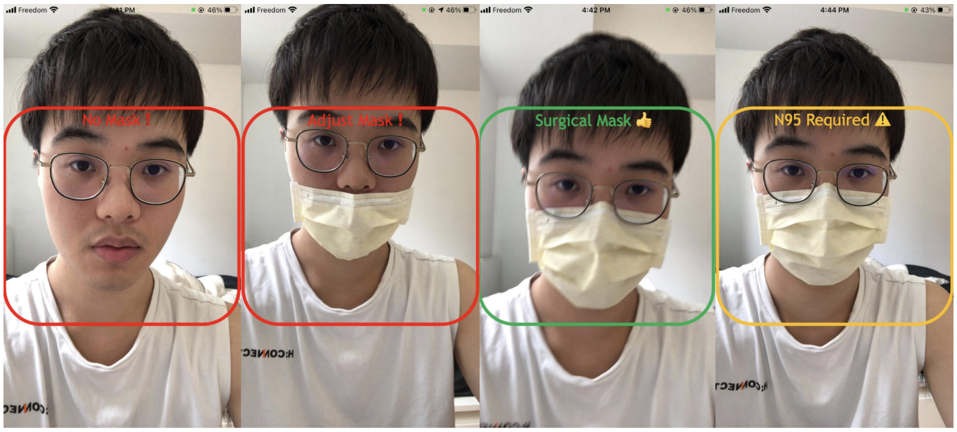
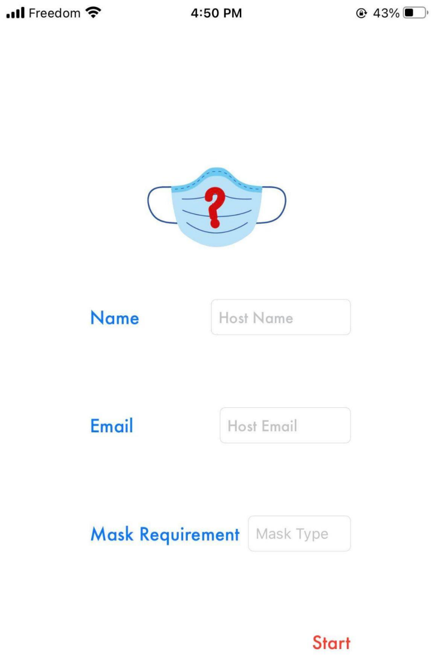


______Fig. 16 Screenshots taken from the iOS mobile application



Name Host Name

Email Host Email

Mask Requirement Mask Type

Fig. 17 User interface of the iOS mobile application

11.0 Discussion

Based on the results, the team believes the two models are suitable for solving the proposed problem. Their superior performances can be credited to the chosen SSD-MobilenetV2 object detection model.

Throughout the design process, the team noted some key advantages of the SSD-MobilenetV2 model. For its performances, the extremely lightweight model was capable of performing accurate real-time detection. The model also loaded smoothly on mobile devices without heating up the hardware significantly. After being trained on images with different subject scaling, rotated mask orientation and multiple masks, the model detected most edge cases successfully.

Despite the advantages of the chosen model, the project was much more difficult than what the team expected initially, and most obstacles were about generating enough quality data. Due to the limited quantity of the "incorrectly masked" class images for the Mask Wearing Model, the team initially included blurry images and images that had advertisement text or other source of biases, which drastically hindered the performance (Figure 18, Figure 19).

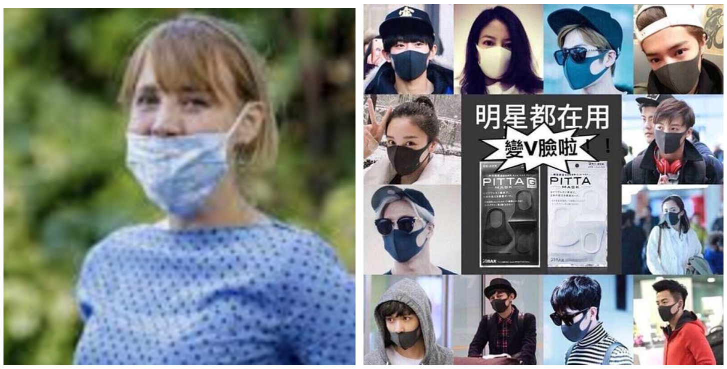


Fig. 18,19: Blurry image and image with text hinders performance

However, having a sufficient and balanced amount of images of each class was required for training the model as the team saw huge improvements in accuracy and precision when tripling the number of images per class for the Mask Type Model. Hence, the team put considerable effort into generating balanced datasets. When images of a specific class were in short supply, images of team members were inserted to the database.

The team also noted cases where the Mask Type Model mistakenly identified a surgical mask as a N95 or vice versa, as sometimes similarities between those two types of masks confuse human eyes. Unfortunately, the only solution for this problem is to keep expanding the dataset which is out of the scope of this project.

Overall, considering the difficulty of this project, the team has developed an highly accurate model that is deployable to different platforms.

12.0 Ethical Considerations

The municipal bylaw of mandatory mask wearing in public spaces has raised concerns towards its ethical standpoint in many countries [12]. Public resistance to face coverings can undermine the usage of a mask identification system both within and beyond pandemic control. Moreover, ethical consideration may arise in the process of collecting people's faces during the usage of this model. Therefore, extra caution must be taken to prevent unauthorized access. In addition, the team will do their best to ensure that the data collection process does not violate any ethical or portrait right issues by obtaining photographs from authorized public resources.

13.0 Code

Primary Mask Wearing Model:
https://drive.google.com/drive/folders/1MH1Gmyd_Lrpn6jD3ih_hXc-NYv_1dKTw?usp=sharing

Primary Mask Type Model:
https://drive.google.com/drive/folders/1CqRUFq5GW--6YsqpW3gWHFq3hnTXeLg9?usp=sharing

14.0 References

1. M. Rizvi, “Learn Image Classification on 3 Datasets using Convolutional Neural Networks (CNN),” Analytics Vidhya. 18-Feb-2020 [Online]. Available: https://www.analyticsvidhya.com/blog/2020/02/learn-image-classification-cnn-convolutional-neural-networks-3-datasets/. [Accessed: 31-May-2021]
2. “How to Safely Wear and Take Off a Cloth Face Covering,” Centers for Disease Control and Prevention, 07-May-2021. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-to-wear-cloth-face-coverings.html. [Accessed: 02-Jun-2021]
3. P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, and J. Hemanth, “SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2,” Sustainable Cities and Society, vol. 66, p. 102692, Mar-2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7775036/. [Accessed: 31-May-2021].
4. M. Loey, G. Manogaran, M. H. Taha, and N. E. Khalifa, “Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection,” Sustainable Cities and Society, vol. 65, p. 102600, Feb-2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7658565/. [Accessed: 31-May-2021]
5. Adnane Cabani, Karim Hammoudi, Halim Benhabiles, and Mahmoud Melkemi, "MaskedFace-Net - A dataset of correctly/incorrectly masked face images in the context of COVID-19", Smart Health, ISSN 2352-6483, Elsevier, 2020, DOI:10.1016/j.smhl.2020.100144. [Accessed: 25-June-2021]
6. Karim Hammoudi, Adnane Cabani, Halim Benhabiles, and Mahmoud Melkemi,"Validating the correct wearing of protection mask by taking a selfie: design of a mobile application "CheckYourMask" to limit the spread of COVID-19", CMES-Computer Modeling in Engineering & Sciences, Vol.124, No.3, pp. 1049-1059, 2020, DOI:10.32604/cmes.2020.011663. [Accessed: 25-June-2021]
7. Z. Wang, G. Wang, and B. Huang, “Masked Face Recognition Dataset and Application,” arxiv.org [Online]. Available: https://arxiv.org/abs/2003.09093. [Accessed: 01-Jun-2021]
8. Antonio Costantino Marceddu, Bartolomeo Montrucchio, September 10, 2020, "Facial Masks and Respirators Database (FMR-DB)", IEEE Dataport, doi: https://dx.doi.org/10.21227/wg71-v415. [Accessed 25-June-2021]
9. Tzutalin. LabelImg. Git code (2015). https://github.com/tzutalin/labelImg. [Accessed 25-June-2021]
10. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510-4520, doi: 10.1109/CVPR.2018.00474. [Accessed 23-June-2021]
11. Liu W. et al. (2016) SSD: Single Shot MultiBox Detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2. [Accessed 23-June-2021]
12. C. Obregon, “THE ETHICS OF MANDATORY MASKS,” Aug-2020 [Online]. Available: https://www.researchgate.net/publication/343523816_THE_ETHICS_OF_MANDATORY_MASKS. [Accessed: 01-Jun-2021]