Robust Face Mask Detection with Combined Frontal and Angled Viewed Faces

Ivan George L. Tarun[1], Vidal Wyatt M. Lopez[1], Patricia Angela R. Abu[1]
and Ma. Regina Justina E. Estuar[2]

[1]*Ateneo Laboratory for Intelligent Visual Environment, Dept. of Information Systems and Computer Science, Ateneo De Manila University, Katipunan Avenue, Quezon City, Philippines*
[2]*Ateneo Center for Computing Competency and Research,Ateneo De Manila University, Katipunan Avenue, Quezon City, Philippines*

Keywords: Face Mask Detection, Object Detection, Deep Learning, Computer Vision.

Abstract: One such protocol currently enforced by the Philippine government to combat COVID-19 is the mandatory use of face masks in public places. The problem however is that ensuring people follow this protocol is difficult to monitor during a pandemic due to other conflicting health protocols like social distancing and workforce reduction. This study therefore explores on the creation of deep learning models that consider both frontal and side view images of the face for face mask detection. In doing so, improvements to robustness were found when compared to using models that were previously trained on purely frontal images. This was accomplished by first relabeling a subset of images from the FMLD dataset. These images were then split into train, validation, and test sets. Four deep learning models (YOLOv5 Small, YOLOv5 Medium, CenterNet Resnet50 V1 FPN 512x512, CenterNet HourGlass104 512x512) were later trained on the training set of images. These four models were compared with three models (MobileNetV1, ResNet50, VGG16) that were trained previously on purely frontal images. Results show that the four models trained on the relabeled FMLD dataset offer an approximately 20% increase in classification accuracy over the three models that were previously trained on purely frontal images.

1 INTRODUCTION

New infectious diseases continue to emerge to this day with them contributing to the already long list of discovered illnesses. With that said, emerging infections (EIs) are infectious diseases that have just recently appeared, or they have already existed but only now are they quickly expanding in geographic range or occurrence (Petersen et al., 2018). There has been a constant fight to contain EIs throughout history due to their global burden of being among the leading causes of death and disability.

Moreover, infectious diseases have continued to evolve throughout history with different diseases having various potential to spread globally (Fauci, 2001). COVID-19 is one recent disease which is caused by the zoonotic virus known as the Severe Acute Respiratory Syndrome Coronavirus 2 or SARS-CoV-2 (Çelik et al., 2020). It is classified as a highly transmissible and pathogenic viral infection. COVID-19 first emerged in Wuhan, China at the end of 2019 and subsequently escalated into a global pandemic. Upon its first emergence, the novel coronavirus managed to kill more than 1,800 people and infect over 70,000 individuals in the first 50 days of the epidemic in China (Shereen et al., 2020). With this as context, the Philippines as of the beginning of February 2022, has surpassed 3.6 million total infections along with 3.5 million recoveries while 54,000 have died (Department of Health (Philippines), 2022).

One protocol currently enforced by the Philippine government to combat COVID-19 is the mandatory use of face masks in public places (Lazaro et al., 2020). The problem however is that ensuring people follow this protocol during a pandemic is difficult to monitor due to conflicting health protocols (social distancing & workforce reduction). Compliance also tends to lessen over time due to complacency (Choudhary et al., 2021). To help solve this, the Ateneo Laboratory for Intelligent Visual Environment (ALIVE) created a real-time face mask detection model for video feeds (Lopez et al., 2021). It checks if a person is wearing a mask or not, and if so, checks if the mask is medically approved or not. Still, the model is

Tarun, I., Lopez, V., Abu, P. and Estuar, M.
Robust Face Mask Detection with Combined Frontal and Angled Viewed Faces.
DOI: 10.5220/0010986000003179
In *Proceedings of the 24th International Conference on Enterprise Information Systems (ICEIS 2022) - Volume 1*, pages 462-470
ISBN: 978-989-758-569-2; ISSN: 2184-4992

limited by only being trained to detect face masks of those who look directly into the camera.

Given the limitations of the front-facing face mask detection model, this study aims to improve on the robustness of the computer vision models in (Lopez et al., 2021) by training new models that also consider angled or side view images of the face.

2 REVIEW OF RELATED LITERATURE

2.1 COVID-19 Health Protocols

From an international perspective, the World Health Organization (WHO) published a document titled "Mask use in the context of COVID-19: interim guidance" (World Health Organization, 2020). To summarize, the WHO recommends the general population to wear at least fabric masks in public settings where physical distancing of at least one meter cannot be maintained and ventilation is known to be poor. Meanwhile, only people with an increased risk of severe complications from contracting COVID-19 are recommended to wear medical masks. These people include those who are aged 60 and above along with those who have underlying comorbidities. As for those who are suspected or confirmed to have COVID-19, they should always wear a medical mask no matter the community setting.

Transitioning to a more localized perspective, the "Omnibus Guidelines on the Implementation of Community Quarantine in the Philippines Updated as of September 23, 2021" dictates comprehensive protocols on how the COVID-19 pandemic should be managed within the Philippines as prepared by the Inter-Agency Task Force for the Management of Emerging Infectious Diseases (Inter-Agency Task Force for the Management of Emerging Infectious Diseases, 2021). Mask use is also recommended by these guidelines, medical or non-medical, similar to that of the WHO. However, there are some key differences particularly that wearing of masks are mandatory in any setting (indoor or outdoor) outside one's own residence.

Overall, these protocols provide deeper context regarding the problem of this study which is the monitoring of the compliance of the public to COVID-19 health protocols. Also note that the technical details of these protocols vary per organization and that there is no universal answer regarding which is the correct one to follow. Multiple organizations nonetheless agree that the wearing of face masks is recommended in combating COVID-19. This therefore solidifies the reasoning behind this study which is to create a computer vision model that monitors the wearing of face masks.

2.2 Previous Attempts in Creating a Computer Vision Model for Detecting Face Masks

A study created a novel deep learning model for face mask detection (Loey et al., 2021). ResNet-50 (Residual Neural Network) with transfer learning and YOLO v2 (You Only Look Once) was used for feature extraction and detection of face masks respectively in the training, validation, and testing phases. The YOLO family of detectors were used for reasons of speed and performance. Mean Intersection over Union (IoU) was also used in the study to estimate the best number of anchor boxes and increase model performance. Data augmentation was also performed to increase the diversity of their dataset by flipping images horizontally. The study then compared two optimizer techniques used in improving detector performance, Stochastic Gradient Descent with Momentum (SGDM) and the Adam optimizer. Their results show that SGDM was better than Adam in training time, validation Root Mean Square Error (RMSE), and validation Loss. However, Adam was better in Mini-batch RMSE and Loss. Evaluation of the detector performance resulted in Adam being better than SGDM with an average precision (AP) of 0.81 while SGDM had 0.61 in all recall levels. Miss rates were also lower and better with Adam with it having a log-average miss rate of 0.4 while SGDM had 0.6. The proposed detector model was then compared with other related works wherein it was found that it performed the best even if AP only reached 81%.

Another study implemented a real-time face mask detection system for embedded systems (Lopez et al., 2021). This study comes from the aforementioned ALIVE laboratory which the current paper seeks to improve upon. The face mask detection system aims to do three class classification namely for wearing medically approved masks, non-medically approved masks, and wearing no mask. Medically approved masks include surgical masks and N95 respirators while non-medically approved masks consist of body parts covering the face (e.g., hands), scarfs, and cloth face masks. A comparative analysis was subsequently conducted between MobileNetV1, VGG16, and ResNet50 to determine which deep learning model to use for the face mask detection system. In doing so, the MAsked FAces (MAFA) dataset (Ge et al., 2017) was manually reclassified to fit the three aforementioned mask wearing types. The three mod-

els were then trained on this subdataset, also incorporating data augmentation, for 20 epochs each with a batch size of 32 and a learning rate of 1e-4. Results show that the MobileNetV1 achieved the highest validation accuracy of 79% followed by VGG16 (76%) and then ResNet50 (37%). After this, the models were ran on a Raspberry Pi 4 Model B (4GB RAM) embedded system and classified video captured from a webcam to see their real-time performance in terms of frames per second (fps). MobileNetV1 had the highest fps with 9.6, then ResNet50 (5.13 fps), and finally VGG16 (4.62 fps).

2.3 Mask Detection Datasets

The aforementioned MAFA dataset is one potential dataset that can be used (Ge et al., 2017). It is meant for masked face detection and consists of 30,811 images containing 35,806 masked human faces. Each masked face has six main attributes which are the locations of the face, eyes, masks, face orientation, occlusion degree, and mask type. Of interest to the current study is the face orientation attribute wherein five orientations were defined specifically left, front, right, left-front, and right-front. Next, the mask type attribute consists of four categories including Simple Mask (pure color), Complex Mask (complex textures or logos), Human Body (face covered by hand, hair, etc.) and Hybrid Mask (combinations of at least two mask types, or one mask type with eyes occluded by glasses).

One criticism of the MAFA dataset is that it is more suited for occlusion detection rather than mask detection (Nowrin et al., 2021). To solve this issue amongst others, a study created the Face Mask Label Dataset (FMLD) (Batagelj et al., 2021). It is a combination of the MAFA and WIDER FACE datasets (Yang et al., 2016). FMLD overall has 41,934 images containing 63,072 faces labeled as either correctly masked, incorrectly masked, or unmasked. Additional annotations are also available like gender, pose, and ethnicity.

2.4 Deep Learning Models

In order to determine the best deep learning models to use for the face mask detection, multiple architectures were benchmarked in this study which is further discussed in Section 3. Some of the benchmarked models include the CenterNet architecture (Zhou et al., 2019) which involves anchorless object detection. It replaces the traditional Non-Maximum Suppression (NMS) with a simpler, faster, and more accurate algorithm. This different approach models objects as a single point which is the center point of the bounding box. Keypoint estimation is then used to find the center points while regression is used for finding other object properties such as size, location, and pose. CenterNet operates on the insight that relevant box predictions can be determined depending on the location of their centers instead of their overlap with the object as in NMS. Less garbage predictions are therefore made compared to anchor-based detection along with removing the need for the computationally heavy NMS.

The rest of the benchmarked models incorporate the YOLOv5 architecture (Jocher et al., 2021). In general, YOLO works by dividing an image into a system of grids where each grid detects objects within itself. This system results in consuming less computational resources during operation. The main differences of YOLOv5 to its predecessors start with its backbone. YOLOv5 combines Cross Stage Partial Networks or CSPNet (Wang et al., 2020) with Darknet to form CSPDarknet as its backbone to extract features from an input image. CSPNet solves the duplicate gradient problems usually found in large-scale backbones and therefore leads to fewer parameters and floating-point operations per second. These further yield faster inference speed and reduced model size. As for the model neck, YOLOv5 utilizes the Path Aggregation Network or PANet (Liu et al., 2018) to generate feature pyramids that aggregate and pass features to the model head. PANet introduces a new feature pyramid structure which improves the propagation of low-level features by having an enhanced bottom-up path. All-in-all, PANet contributes better location accuracy for objects. The model head remains the same between YOLOv5 and YOLOv4 which is responsible for producing class predictions and bounding boxes. It is capable of multi-scale detection or handling small, medium, and large objects (Xu et al., 2021). Other differences present in YOLOv5 include the use of the Leaky Rectified Linear Unit and Sigmoid activation functions to overcome the vanishing gradient problem. The loss function has also switched to the Binary Cross-Entropy with Logits Loss function. These last two differences help YOLOv5 to learn faster and perform better.

3 METHODOLOGY

This section contains the necessary steps that were performed for this study namely Dataset Relabeling, Splitting the Dataset, Model Training, and Model Comparison. These are further elaborated in their respective subsections.

3.1 Dataset Relabeling

With the goal of this study to improve the models in (Lopez et al., 2021), the dataset required needs to be similar wherein there are classes for Medically Approved Masks, Non-Medically Approved Masks, and No Masks. These would be for the frontal view of the face. Another set of the three classes would be needed for the side or angled views of the face. Taking this into consideration, either the MAFA or FMLD datasets need to be relabeled in order to have the same set of classes in (Lopez et al., 2021).

Therefore, Dataset Relabeling was done based on the FMLD dataset. The makesense.ai tool (Skalski, 2019) was used to relabel a subset of images into six classes namely Front - Medical Mask, Front - Non Medical Mask, Front - No Mask, Side - Medical Mask, Side - Non Medical Mask, and Side - No Mask. Frontal images are defined as faces where all its parts are in view (both eyes, whole nose, and whole mouth). Side view images then are defined as either purely left or purely right images of the face with only some parts of the face in view (one eye, half nose, half mouth, one ear). Medical Masks include surgical masks and N95 respirators while Non Medical Masks consist of scarfs, cloth face masks, and construction masks. Sample images can be seen in Figure 1.

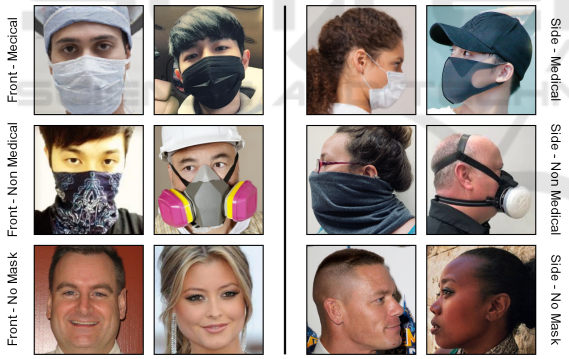

Figure 1: Sample images from the relabeled FMLD dataset.

3.2 Splitting the Dataset

Moving forward, the relabeled FMLD dataset was split into train, validation, and test sets. Eighty percent (80%) of the images were used as the train set, ten percent (10%) as the validation set, and ten percent (10%) as the test set. Care was taken to ensure that there was an equal number of images per class in each respective set. This was done through the train_test_split function with the stratify parameter of the scikit-learn package in Python (Pedregosa et al., 2011).

3.3 Model Training

After dataset splitting, four deep learning models were trained on the training set. Two of them are the YOLOv5 Small and YOLOv5 Medium models (Jocher et al., 2021). The other two are the CenterNet Resnet50 V1 FPN 512x512 and CenterNet HourGlass104 512x512 models from the Tensorflow 2 Object Detection API Model Zoo (Huang et al., 2017). Transfer learning was done with the four models pretrained on the COCO 2017 dataset. Each model also comes with their own standard set of data augmentations which were left as is. Next, the four models were each trained for 300 epochs. The YOLOv5 Small and Medium models had a batch size of 32 and 16 respectively with both having an input image size of 640 pixels. The CenterNet Resnet50 V1 FPN 512x512 and CenterNet HourGlass104 512x512 models then had a batch size of 16 and 4 respectively with both having an input image size of 512x512 pixels.

YOLOv5 models were chosen for their focus on speed while the two CenterNet models were chosen for their focus on performance. Mean Average Precision (mAP) at an Intersection over Union (IoU) of 0.5 to 0.95, or simply mAP@IoU=0.5:0.95 was then recorded during training using the validation set. The same goes for mAP@IoU=0.5. Total loss was also recorded for the training and validation sets. Training was done on a local desktop computer with an Nvidia GTX 1080 Ti GPU. The best and last weights of each model were saved afterwards. The best weights were chosen based on the highest value of mAP@IoU=0.5:0.95 across the 300 epochs.

3.4 Model Comparison

Following Model Training, the four models were compared with the three deep learning models (MobileNetV1, ResNet50, VGG16) from (Lopez et al., 2021) to see if training with side view images of the face would improve face mask detection from the sides. This was done by comparing the accuracy of the models in classifying the six classes of the relabeled FMLD dataset. Classification accuracy was chosen as the comparison metric because the three models from (Lopez et al., 2021) are classification models while the four models of this study are object detection models. Classification accuracy would be a common metric that can be measured for both types of models. One issue in measuring the classification accuracy for the three models from (Lopez et al., 2021) was that they were only trained to classify three classes (Medical Mask, Non Medical Mask, No Mask) while the relabeled FMLD dataset has six

classes as seen in Section 3.1. The eventual solution was to consider the predictions of the three models as correct or wrong depending on if they were able to predict the mask type of a face correctly no matter if the face had a frontal or side view. So for example, if the VGG16 model gave a prediction of Medical Mask and the ground truth label was Side - Medical Mask, then this was considered as correct.

The validation and test sets were then combined next in preparation for measuring the classification accuracy of all the seven models. This was done to increase the number of images per class so that classifying a single image correctly or wrongly would not affect the overall classification accuracy too much. Inference was then done with each of the seven models on the combined validation and test set wherein the class with the highest confidence score from the prediction was chosen as the predicted class for the particular image. Accuracy per class was calculated afterwards by dividing the number of correct predictions over the total number of images per class and then multiplying by 100 to get a percentage value. Calculating the classification accuracies along with Inferencing was done with Python. The best weights of the four models of this study were also used for Inferencing. The PyTorch package (Paszke et al., 2019) was used for Inferencing the YOLOv5 models while the Tensorflow package (TensorFlow Developers, 2021) was used for Inferencing the two CenterNet models.

4 RESULTS AND DISCUSSIONS

4.1 Relabeled FMLD Dataset

The completion of the dataset relabeling step produced a subdataset of images from the FMLD dataset that was subsequently used for this study. The relabeled FMLD dataset therefore consists of 300 images in total with 50 images per class. Each image only contains one face so that the dataset can be easily balanced during splitting. A breakdown of the relabeled FMLD dataset can be seen in Table 1.

Table 1: Breakdown of the relabeled FMLD dataset.

Class	Number of Samples
Front - Medical Mask	50
Front - Non Medical Mask	50
Front - No Mask	50
Side - Medical Mask	50
Side - Non Medical Mask	50
Side - No Mask	50
Total	300

4.2 Dataset Splits

Splitting the relabeled FMLD dataset resulted in three image sets. The training set thus contains 240 images or 80% of the total number of images. Equal distribution of the 240 images leads to 40 images per class. The validation set then consists of 30 images or 10% of the total images. This resolves to 5 images per class. Lastly, the test set also includes 30 images and also represents 10% of the total images. This too then gives 5 images per class.

4.3 Model Training

Figure 2 depicts the graphs of the mAP@IoU=0.5:0.95 and mAP@IoU=0.5 values for the four models of this study during training. It can be seen that the two CenterNet models have higher mAP values than the YOLOv5 models and they plateau earlier as well while the YOLOv5 models have a more gradual increase. Table 2 provides a summary of the mAP values of the four models wherein the best and last values are displayed along with the epoch where they occur. The highest value of mAP@IoU=0.5:0.95 (0.761451) was achieved by the CenterNet Resnet50 V1 FPN 512x512 model at Epoch 184. The YOLOv5 models have a lower value ranging from 0.05 to 0.08 to the CenterNet models. As for mAP@IoU=0.5, the highest value (0.994499) was achieved by the CenterNet HourGlass104 512x512 model at Epoch 33. The YOLOv5 models also have a lower value with a range of 0.01 to 0.02 to the CenterNet models.

Table 2: Summary of Mean Average Precision values.

	mAP @ IoU = 0.5:0.95		mAP @ IoU = 0.5	
Model	Best	Last	Best	Last
YOLOv5 Small	0.701899 Epoch 258	0.641859 Epoch 300	0.980495 Epoch 257	0.921729 Epoch 300
YOLOv5 Medium	0.681231 Epoch 300	0.681231 Epoch 300	0.974667 Epoch 284	0.896467 Epoch 300
CenterNet Resnet50 V1 FPN 512x512	**0.761451** **Epoch 184**	0.725012 Epoch 300	0.99057 Epoch 124	0.97454 Epoch 300
CenterNet HourGlass104 512x512	0.758944 Epoch 112	0.707889 Epoch 300	**0.994499** **Epoch 33**	0.912096 Epoch 300

Bold = Highest Value

The total loss curves of the YOLOv5 and CenterNet models are shown in Figures 3 and 4 respectively. It is evident that the CenterNet models were overfitting during training as the validation total losses started to fluctuate at around Epoch 30. These may also explain why the best values for mAP of the CenterNet models occur at the earlier epochs and not near the end. The YOLOv5 models on the other hand had

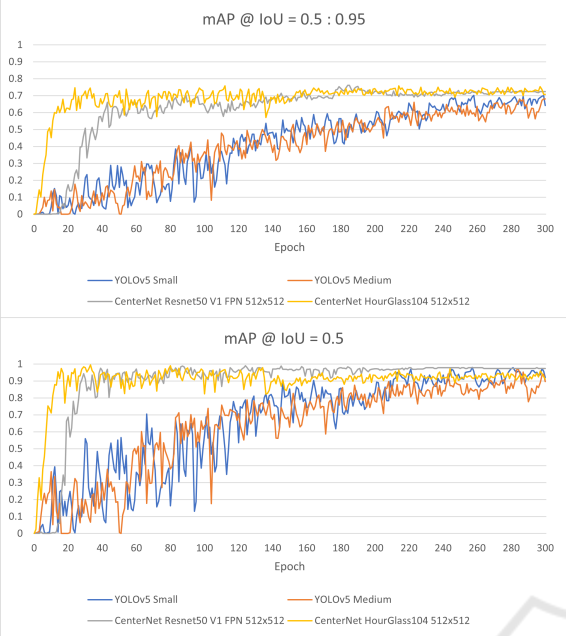

Figure 2: Mean Average Precision graphs of the four models of this study while training.

their validation total losses decreasing up to the very end of the 300 epochs and were not overfitting. Thus, their best mAP values also occur near the tail end of training.

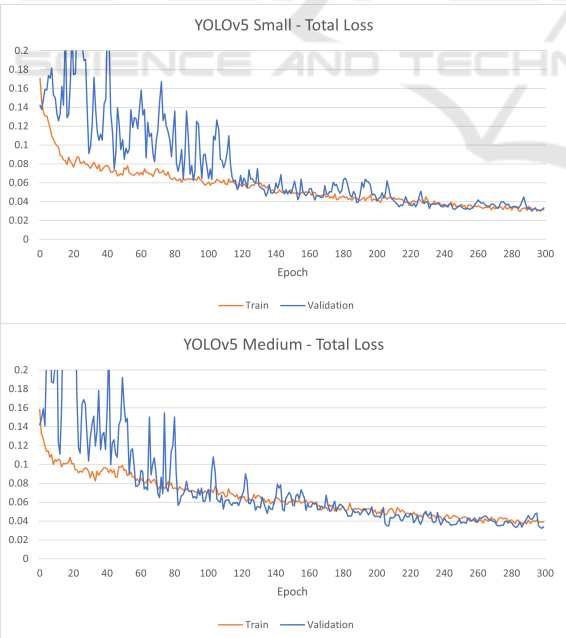

Figure 3: Total Loss curves of the YOLOv5 Models.

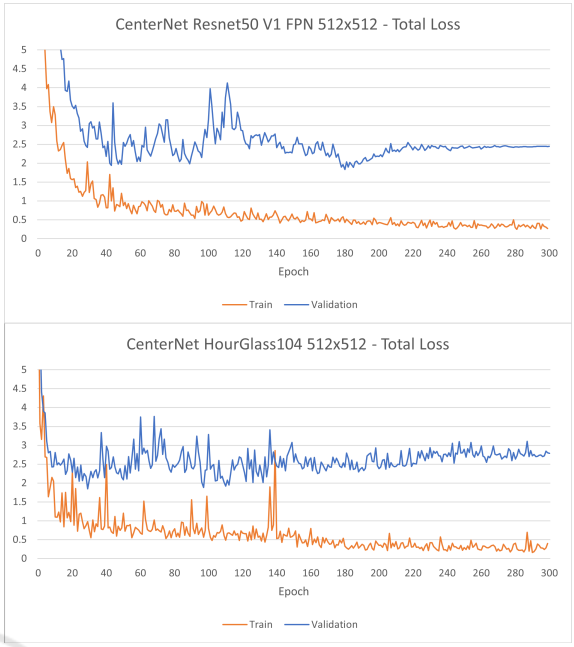

Figure 4: Total Loss curves of the CenterNet Models.

4.4 Model Comparison

The combination of the validation and test sets resulted into a set of 60 images total with 10 images per class. These 60 images were used for the model comparison of the seven models. Subsequently, Table 3 depicts a summary of the resulting classification accuracies per model and per class that were measured. Each cell contains the number of correctly predicted images, with respect to the model and class, out of the total number of images present in the class along with the corresponding accuracy.

For ease of readability, the six classes of the relabeled FMLD dataset will be abbreviated in the subsequent sections. The abbreviations are namely Front - Medical Mask (F-MM), Front - Non Medical Mask (F-NMM), Front - No Mask (F-NoM), Side - Medical Mask (S-MM), Side - Non Medical Mask (S-NMM), and Side - No Mask (S-NoM).

Bearing these technicalities in mind, the MobileNetV1 model seems to be an anomaly as it performed poorly. It managed to predict nine of the ten images correctly for the F-NMM and S-NMM classes. However, it failed to predict any of the ten images for the F-MM and S-MM classes. Further observation revealed that the reason behind this behavior was that the MobileNetV1 model was classifying majority of the images as Non Medical Mask hence the high accuracy in these classes. In the end, it managed to predict 21 of the 60 images in the combined validation and test set correctly which gives it an overall

Table 3: Summary of Classification Accuracies.

Model	Front - Medical Mask	Front - Non Medical Mask	Front - No Mask	Side - Medical Mask	Side - Non Medical Mask	Side - No Mask	Overall
MobileNetV1 (Lopez et al., 2021)	0/10 Accuracy = 0%	9/10 Accuracy = 90%	1/10 Accuracy = 10%	0/10 Accuracy = 0%	9/10 Accuracy = 90%	2/10 Accuracy = 20%	21/60 Accuracy = 35%
ResNet50 (Lopez et al., 2021)	9/10 Accuracy = 90%	**10/10 Accuracy = 100%**	6/10 Accuracy = 60%	7/10 Accuracy = 70%	**10/10 Accuracy = 100%**	4/10 Accuracy = 40%	46/60 Accuracy = 76.67%
VGG16 (Lopez et al., 2021)	**10/10 Accuracy = 100%**	**10/10 Accuracy = 100%**	6/10 Accuracy = 60%	4/10 Accuracy = 40%	**10/10 Accuracy = 100%**	3/10 Accuracy = 30%	43/60 Accuracy = 71.67%
YOLOv5 Small (This Study)	9/10 Accuracy = 90%	9/10 Accuracy = 90%	**10/10 Accuracy = 100%**	9/10 Accuracy = 90%	**10/10 Accuracy = 100%**	8/10 Accuracy = 80%	55/60 Accuracy = 91.67%
YOLOv5 Medium (This Study)	**10/10 Accuracy = 100%**	8/10 Accuracy = 80%	**10/10 Accuracy = 100%**	8/10 Accuracy = 80%	**10/10 Accuracy = 100%**	**10/10 Accuracy = 100%**	56/60 Accuracy = 93.33%
CenterNet Resnet50 V1 FPN 512x512 (This Study)	**10/10 Accuracy = 100%**	**10/10 Accuracy = 100%**	9/10 Accuracy = 90%	9/10 Accuracy = 90%	**10/10 Accuracy = 100%**	9/10 Accuracy = 90%	**57/60 Accuracy = 95%**
CenterNet HourGlass104 512x512 (This Study)	**10/10 Accuracy = 100%**	**10/10 Accuracy = 100%**	**10/10 Accuracy = 100%**	**10/10 Accuracy = 100%**	9/10 Accuracy = 90%	8/10 Accuracy = 80%	**57/60 Accuracy = 95%**

Bold = Highest Value in Column

accuracy of 35%.

Moving to the ResNet50 model, it had a generally acceptable performance. The model managed to predict all of the images correctly in the F-NMM and S-NMM classes. However, accuracy decreased with the F-NoM, S-MM, and S-NoM classes. The ResNet50 model had an accuracy of 76.67% all-in-all by predicting 46 of the 60 images correctly. It therefore has the highest accuracy out of the three models from (Lopez et al., 2021).

The VGG16 model also had an acceptable overall performance. It correctly predicted all of the images in the F-MM, F-NMM, and S-NMM classes. This has one more perfectly predicted class than that of the ResNet50 model but the VGG16 model had lower accuracy in the rest of the classes particularly in the S-MM and S-NoM classes. For this reason, it only has an overall accuracy of 71.67% which puts it second to the ResNet50 model. This therefore makes the order of the most accurate models coming from (Lopez et al., 2021) (from most to least) be the ResNet50 model in first followed by VGG16 and the MobileNetV1 in last. This is to be expected as (Agarwal and Mittal, 2019) and (Bianco et al., 2018) describe that the MobileNetV1 model intends to be lightweight and have lower accuracy while the ResNet50 and VGG16 models are deeper or have more layers and focus more on having high accuracy. Furthermore, the ResNet50 model tends to have higher accuracy than the VGG16 model because the ResNet50 model has an architecture that is designed to solve the vanishing or exploding gradient problem.

Proceeding to the YOLOv5 Small model, it can be said to have higher performance when compared to the other models. It predicted all of the images correctly in the F-NoM and S-NMM classes while its lowest scoring class was eight which is the S-NoM class. The rest of the classes had nine out of the ten images predicted correctly. It therefore has an overall accuracy of 91.67% by predicting 55 of the 60 images correctly.

As for the YOLOv5 Medium model, it too had significantly high performance compared to the other models. All of the images in the F-MM, F-NoM, S-NMM, and S-NoM classes were predicted correctly. The rest of the classes had eight of the ten images predicted as correct. Thus, it managed to predict 56 of the 60 images correctly, one more image than the YOLOv5 Small model, giving it an overall accuracy of 93.33%.

This leads to the CenterNet Resnet50 V1 FPN 512x512 model which had the highest overall performance. It only managed to perfectly predict three classes (F-MM, F-NMM, S-NMM) but the rest of the classes had nine out of the ten images correctly predicted. Fifty-seven (57) out of the 60 images were then predicted as correct resulting in an overall accuracy of 95%.

Lastly, the CenterNet HourGlass104 512x512 model also ties for the highest performance. The model predicted four classes perfectly namely the F-MM, F-NMM, F-NoM, and S-MM classes. Nine out of the ten images for the S-NMM class were correctly predicted while the S-NoM class had eight correct predictions. It too managed to correctly predict 57 of the 60 images equalling the CenterNet Resnet50 V1 FPN 512x512 model and therefore also giving it an overall accuracy of 95%. The order now of the most accurate models (from most to least of all seven models) comes to the CenterNet Resnet50 V1 FPN 512x512 and CenterNet HourGlass104 512x512 models tied for first, followed by the YOLOv5 Medium model, YOLOv5 Small, ResNet50, VGG16, and finally MobileNetV1. Again, these are expected because the CenterNet models are focused on performance while the YOLOv5

models focus more on speed. The CenterNet and YOLOv5 models should also have higher accuracies than the ResNet50, VGG16, and MobileNetV1 models from (Lopez et al., 2021) because the CenterNet and YOLOv5 models are trained on the relabeled FMLD dataset while the ResNet50, VGG16, and MobileNetV1 models are trained on a different dataset which is the reclassified MAFA dataset also found in (Lopez et al., 2021).

In summary, these results support the main rationale behind this study which was to find out if training deep learning models to also consider side view images of the face for face mask detection can improve the robustness of the front-facing face mask detection models that were already trained in (Lopez et al., 2021). It can be observed in Table 3 that the three models from (Lopez et al., 2021) (MobileNetV1, ResNet50, VGG16) had lower accuracy when it comes to detecting face masks from the side of a face. An exception would be the Side - Non Medical Mask class where high accuracy was surprisingly observed. Nevertheless, the four models (YOLOv5 Small, YOLOv5 Medium, CenterNet Resnet50 V1 FPN 512x512, CenterNet HourGlass104 512x512) of this study that were trained to also consider side view images of the face had indeed gained an improvement in accuracy when detecting face masks from the side of a face. Comparing the best overall accuracy from the models in (Lopez et al., 2021) (Resnet50 with 76.67%) with the models of this study (CenterNet Resnet50 V1 FPN 512x512 and CenterNet HourGlass104 512x512 with 95%), an overall improvement of $\approx$ 20% in accuracy was noticed.

5 CONCLUSION AND RECOMMENDATIONS

To conclude, the goal of this study was to improve upon the robustness of the deep learning models that were already created in (Lopez et al., 2021) for face mask detection. These models were intended to aid in monitoring compliance by checking if a person is wearing a mask or not, and if so, check if the mask is medically approved or not. They were limited however to only being trained in detecting face masks of those who look directly into the camera. Thus improvements were sought after by finding out if training new deep learning models to also consider side view images of the face can indeed improve upon the robustness of the front-facing face mask detection of the models from (Lopez et al., 2021). In doing so, dataset relabeling was performed resulting in the relabeled FMLD dataset with 300 images in total which are distributed equally across six classes namely Front - Medical Mask, Front - Non Medical Mask, Front - No Mask, Side - Medical Mask, Side - Non Medical Mask, and Side - No Mask. The relabeled FMLD dataset was then split into train, validation, and test image sets. Four deep learning models were then trained on the training set of images namely the YOLOv5 Small, YOLOv5 Medium, CenterNet Resnet50 V1 FPN 512x512, and CenterNet HourGlass104 512x512 models. Afterwards, these four models were compared with the three models (MobileNetV1, ResNet50, VGG16) from (Lopez et al., 2021) by measuring their classification accuracies on the combined validation and test set of the relabeled FMLD dataset. Results show that the four models of this study that were trained to also consider side view images of the face exhibited improvements in accuracy when detecting face masks from the side of a face. An overall increase of $\approx$ 20% was observed with the best models of this study against the best models from (Lopez et al., 2021) in classifying frontal and side view images. To this end, the goal of this study was achieved and it can therefore be said that the four models of this study are more robust than the models from (Lopez et al., 2021) when it comes to face mask detection.

As for future improvements to this study, possible recommendations can include the creation of a larger dataset containing more images since the relabeled FMLD dataset of this study is relatively small with only 300 images in total for six classes leading to 50 images per class. Incorporating ethnicity and/or time of day can also be added to this larger dataset. A more exhaustive list of deep learning models can be further benchmarked as well to find the best performing one. Another possible route would be to examine the effects of the different data augmentations that are available for use with the four models of this study as they were mainly left at the default settings. Lastly, attempting to produce a real-world application or setup by extending the methodology of this study serves as an additional possible avenue for continuation.

REFERENCES

Agarwal, T. and Mittal, H. (2019). Performance comparison of deep neural networks on image datasets. In *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pages 1–6.

Batagelj, B., Peer, P., Štruc, V., and Dobrišek, S. (2021). How to correctly detect face-masks for covid-19 from visual information? *Applied Sciences*, 11(5).

Bianco, S., Cadene, R., Celona, L., and Napoletano, P. (2018). Benchmark analysis of representative deep

neural network architectures. *IEEE Access*, 6:64270–64277.

Choudhary, O. P., Priyanka, Singh, I., and Rodriguez-Morales, A. J. (2021). Second wave of COVID-19 in India: Dissection of the causes and lessons learnt. *Travel Medicine and Infectious Disease*, 43:102126.

Department of Health (Philippines) (2022). COVID-19 Tracker.

Çelik, I., Saatçi, E., and Eyüboğlu, A. F. (2020). Emerging and reemerging respiratory viral infections up to Covid-19. *Turkish Journal of Medical Sciences*, 50(SI-1):557–562.

Fauci, A. S. (2001). Infectious Diseases: Considerations for the 21st Century. *Clinical Infectious Diseases*, 32(5):675–685.

Ge, S., Li, J., Ye, Q., and Luo, Z. (2017). Detecting masked faces in the wild with lle-cnns. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 426–434.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., and Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3296–3297.

Inter-Agency Task Force for the Management of Emerging Infectious Diseases (2021). Omnibus Guidelines on the Implementation of Community Quarantine in the Philippines as of September 23, 2021.

Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., NanoCode012, TaoXie, Kwon, Y., Michael, K., Changyu, L., Fang, J., V, A., Laughing, tkianai, yxNONG, Skalski, P., Hogan, A., Nadar, J., imyhxy, Mammana, L., AlexWang1900, Fati, C., Montes, D., Hajek, J., Diaconu, L., Minh, M. T., Marc, albinxavi, fatih, oleg, and wanghaoyang0106 (2021). ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support.

Lazaro, R. E., Tupas, E., Cabrera, R., and Villanueva, R. E. (2020). Wearing masks now mandatory. *Philstar.com*.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). Path aggregation network for instance segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8768.

Loey, M., Manogaran, G., Taha, M. H. N., and Khalifa, N. E. M. (2021). 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*, 65:102600.

Lopez, V. W. M., Abu, P. A. R., and Estuar, M. R. J. E. (2021). Real-time face mask detection using deep learning on embedded systems. In *2021 3rd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, pages 1–7.

Nowrin, A., Afroz, S., Rahman, M. S., Mahmud, I., and Cho, Y.-Z. (2021). Comprehensive review on facemask detection techniques in the context of covid-19. *IEEE Access*, 9:106839–106864.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Petersen, E., Petrosillo, N., Koopmans, M., Beeching, N., Di Caro, A., Gkrania-Klotsas, E., Kantele, A., Kohlmann, R., Koopmans, M., Lim, P.-L., Markotic, A., López-Vélez, R., Poirel, L., Rossen, J., Stienstra, Y., and Storgaard, M. (2018). Emerging infections—an increasingly important topic: review by the emerging infections task force. *Clinical Microbiology and Infection*, 24(4):369–375.

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., and Siddique, R. (2020). Covid-19 infection: Emergence, transmission, and characteristics of human coronaviruses. *Journal of Advanced Research*, 24:91–98.

Skalski, P. (2019). Make Sense. https://github.com/SkalskiP/make-sense/.

TensorFlow Developers (2021). Tensorflow. Specific TensorFlow versions can be found in the "Versions" list on the right side of this page. See the full list of authors "https://github.com/tensorflow/tensorflow/graphs/contributors" on GitHub.

Wang, C.-Y., Mark Liao, H.-Y., Wu, Y.-H., Chen, P.-Y., Hsieh, J.-W., and Yeh, I.-H. (2020). Cspnet: A new backbone that can enhance learning capability of cnn. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1571–1580.

World Health Organization (2020). Mask use in the context of COVID-19: interim guidance, 1 December 2020. Technical documents.

Xu, R., Lin, H., Lu, K., Cao, L., and Liu, Y. (2021). A forest fire detection system based on ensemble learning. *Forests*, 12:217.

Yang, S., Luo, P., Loy, C. C., and Tang, X. (2016). Wider face: A face detection benchmark. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5525–5533.

Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. In *arXiv preprint arXiv:1904.07850*.