Masked Faces in Context (MASON) for Masked Face Detection and Classification

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Helena Shield

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Advisors: Dr Ming Jiang; Dr Catherine Qi Zhao

Jan, 2023

Acknowledgements

I would first like to thank my adviser Dr. Catherine Zhao and Dr. Ming Jiang for their incredible support throughout my time here. This project and thesis would not have been possible without your guidance and encouragement. I would also like to thank Dr. Ingrid Schneider and Dr. Ce Yang for serving on my committee. I am also grateful to the MNRI group that provided the funding and support for this project and for me during my time at UMN. Finally I'd like to thank my family, friends, and peers who have been a source of inspiration and comfort as I worked towards this accomplishment.

Abstract

As the SARS-CoV-2 virus mutated and spread around the world, scientists and public health officials were faced with the responsibility of making health recommendations as they studied the novel disease in real time. One such recommendation was the use of face masks of varying types as a method of reducing disease spread in public spaces. Evaluating the effectiveness of such measures requires accurate data collection of the proper facemask usage. The use of computer vision models to detect and classify face mask usage can aid in the collection process by monitoring usage in public spaces. However, training these models requires accurate and representative datasets. Pre-COVID-19 datasets and synthetic datasets have limitations that affect the accuracy of models in real world settings such as inaccurate representations of occlusion and limited variety of subjects, settings, and masks. In this work we present a new dataset Masked Faces in Context (MASON) of annotated real-world images focusing on the time period of 2020 to the present and baseline detection and classification models that outperforms the current state of the art. This dataset better snapshots mask wearing under covid with greater representation of different age groups, mask types, common occlusion items such as face shields, and face position. Our experiments demonstrate increased accuracy in face mask detection and classification.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In 2019, the novel virus SARS-CoV-2 spread around the world leaving scientists and public health officials scrambling. The resulting disease, known as COVID-19, posed a formidable adversary in its infectiousness and deadliness [3]. Developing disease control guidelines with a limited understanding of the precise virus mechanisms was necessary, and could be adjusted as the effect was observed. One such method was the implementation of masking mandates [4]. However, lacking proper tools to evaluate public mask-wearing itself makes it nearly impossible to analyze the effects of mask usage on disease spread.

One promising approach is to use computer vision detection and classification models to collect mask-wearing data in an accurate and timely manner. These types of models have been successfully applied in other areas such as autonomous vehicles, airport security, and visual health diagnoses [5, 6, 7]. An application in mask usage analyses would require training a model to detect faces and classify the masking, which in turn requires appropriate training data. Unfortunately, with the COVID-19 pandemic being the first widespread pandemic in the modern world, there are limited datasets suited to this problem. Many are either synthetic adaptations of face datasets or created before the pandemic, both of which come with concerns about bias and efficacy.

To aid in this challenge, we developed a new dataset Masked Faces in Context (MASON), and a baseline model for mask detection that can detect masked and unmasked faces in real-world images. In this new dataset, we focused on collecting real-world images off the web that reflect common masking practices under COVID-19. The images are presented in two forms: full images with bounding box labels for the visible faces in the image, and cropped images to only contain visible faces with labels for masking type. In our comparisons with existing datasets, we found models trained on MASON outperform that of existing datasets in both the face detection and masking classification tasks.

The following chapters are organized as follows:

- **Chapter 2** covers a literature review of previous work including an analysis of existing datasets consisting of masked faces and their limitations.
- **Chapter 3** covers our methods of creating MASON to address identified limitations.
- **Chapter 4** covers the results of our contributions and comparisons to previous work.
- **Chapter 5** covers the conclusion, limitations of our work, and future steps.

Chapter 2

Previous Work

While masking classification and masked face detection methods used the existing datasets such as MaskedFace-Net [8, 9], RMFD [10, 11], and MAFA [12, 13], a larger number chose to collect, annotate, and use smaller privately collected datasets [14, 15, 16, 17, 18, 19, 20, 21, 22]. These datasets often contained less than 2000 cropped images of masked and unmasked faces, and some contained less than 150 unique subjects. This pattern indicates that the drawbacks outlined below are significant to the face detection and mask classification problem and need to be addressed.

In this section, we review existing masked face datasets with their strengths and limitations. We then present an overview of the existing state-of-the-art face mask detection models associated with these datasets.

Existing Datasets

The three most commonly referenced masking datasets are the 2020 synthetic dataset MaskedFace-Net, the 2017 occlusion dataset Masked Faces (MAFA), and the 2020 masking dataset Real World Masked Face Dataset (RMFD). This section describes these datasets in detail and outlines their strengths and weaknesses regarding the masked face detection and classification problems.

MaskedFace-Net

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

Figure 2.1: Example images from the MaskedFace-Net dataset.

The MaskedFace-Net dataset created in 2020 consists of 133,782 face images taken from the face dataset Flickr-Faces-HQ (FFHQ) each with a blue surgical mask digitally superimposed over the face. About half of these masks are correctly placed, forming the Correctly Masked Face Dataset (CMFD). The other half, the Incorrectly Masked Face Dataset (IMFD), either shows an uncovered chin, uncovered nose, or uncovered nose and mouth [8].

Since MaskedFace-Net uses the well-established FFHQ as a base, many strengths from FFHQ apply to MaskedFace-Net as well. Namely, the large size and diversity of subjects [23]. However, there are some significant limitations to this dataset. Since the dataset was formed by digitally adding a mask to every face, each mask has the same appearance. The variety in pattern, shape, texture, lighting, and color that appear in real-world masks is not reflected in this dataset. The second major issue with synthetic datasets such as MaskedFace-Net is the handling of occlusion. The masks in this dataset appear over occluding objects such as hands, microphones, and glasses. Occlusion or non-forward-facing face pose in the original face image can also cover key facial landmarks used for the mask placement leading to unrealistic warping of the mask. Finally, as purely a masking classification dataset, this set only contains close-up headshots of people's faces, removing the potential for noisy backgrounds or other interacting objects and occlusions as well as lacking labels for masked face detection.

Figure 2.1 exemplifies some of the drawbacks listed above. Images (a) and (c) show misaligned masks on side-facing faces and image (b) has unnatural warping down the center of the mask as well as an extra sliver across the women's nose. The next three images show how the dataset does not account for occluding objects with the masks covering a microphone in image (d), a bus seat in image (e), and the man's hands in image (f). This collection of images also demonstrates the lack of variety in masks and the focus on clear headshots.

Overall these deficiencies make it more difficult to train effective models for masking classification because the data isn't representative of what real-world mask-wearing looks like.

Masked Face Dataset (MAFA)

Figure 2.2: Example images from the MAFA dataset.

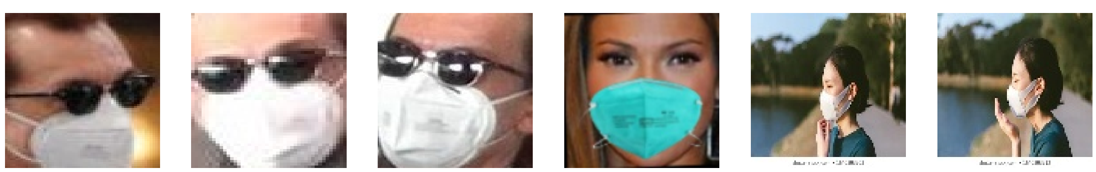
Masked Face Dataset (MAFA) was created in 2017 before the Covid-19 pandemic to aid in the detection of faces under occlusion. It consists of 30,811 images containing 35,806 masked faces [12]. The labels for this dataset consist of information regarding occlusion, gender, race, face orientation, mask location, and mask type. For this dataset, "mask" is more broadly defined as any face-occluding object.

The detailed labeling for this dataset makes it a good candidate for the masked face detection problem. However, because the dataset was created with a more broad purpose in mind, it does not correspond well to the masking classification problem. Some "masked" images do not include any masks but rather other sources of occlusion, making it impossible to differentiate between adequate and inadequate masking by Covid-19 standards with the currently available labels. The images with masked faces in 2017 tended towards occupational masking such as healthcare or construction as masking was not as widespread as it became post Covid-19. Focusing on these groups resulted in many of the photos being stock photo images which tend towards non-occluded forward-facing faces and less noisy backgrounds.

As seen in Figure 2.2 images (a), (b), and (c) feature healthcare and construction stock photos with unnaturally posed faces and the masks most commonly associated with each profession. Images (d), (e), and (f) are all "masked" images with a non-mask occluding object such as a fan, a hand, and a theater mask.

All in all, MAFA contains a limited sample of the variety of people, masks, and settings applicable to Covid-19.

Real-World-Masked-Face-Dataset (RMFD)



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

Figure 2.3: Example images from the Real World Masked Face Recognition Dataset (RWMFD) and Simulated Masked Face Recognition Dataset (SMFRD).

Real World Masked Face Dataset (RMFD) is broken into three sub-datasets: the Masked Face Detection Dataset (MFDD), the Real-world Masked Face Recognition Dataset (RMFRD), and the Simulated Masked Face Recognition Dataset (SMFRD) [10]. MFDD contains 24,771 real-world images collected from the internet of human faces along with annotations of the masking classification and face location. The second group, RMFRD, contains 5,000 real-world masked images and 90,000 real-world unmasked images of 525 public figures. The only annotations for this group are the masking classification for each image. The final group, SMFRD, is a synthetic mask dataset that consists of existing face datasets such as Labeled Faces in the Wild (LIW) and Webface with masks digitally added to each face. This dataset contains 500,000 images of 10,000 digitally masked faces.

Of the three datasets in this family, the MFDD appears to be the most appropriate for the masking classification and masked face detection problems as it has suitable annotations and is relatively large. However, this dataset was not made publicly available for use. Between RMFRD and SMFRD, RMFRD has the advantage of using real-world images while SMFRD struggles with many of the same limitations as MaskedFace-Net as it is digitally generated. Despite its large size, RMFD is limited by its unbalanced nature between masked and unmasked images and lacks diversity in its subjects since it only contains 525 unique faces. The images are also not all headshots and the labels do not reflect a difference between cropped images and non-cropped images.

Figure 2.3 exemplifies some of the drawbacks listed above. Images (a), (b), (c), (d), and (e) from RMFD are an example of a collection of images containing the same face. While the first three display some variation in the face pose of the subject, everything else appears about the same. The last two are also of the same subject and even though they are not as cropped as images (a), (b), and (c), the annotations are the same. Image (d) is an example image from SMFRD where the blue mask has been digitally added onto the woman's face.

In summary, MFDD is unavailable for our tasks, RMFD lacks diversity and balance, and the synthetic option SMFRD maintains the same design issues as MaskedFace-Net that make it unsuitable for these tasks.

Face Detection

The localization of faces in images is a well studied problem in object detection. Many of the commonly known state-of-the-art deep learning models such as the one stage detector YOLO [24] (and variants) and the two-stage detector RCNN [25] (and variants) are commonly used in general object detection applications thanks to their accuracy and speed [26, 27, 28, 29]. With the highly variant and often low resolution nature of faces in images, more face-specific detection methods exist as well. HyperFace [30] builds on these ideas but also integrates in face landmark localization, pose estimation, and gender recognition. RetinaFace [31] also applies face landmark localization in addition to 2D and 3D reconstruction. These methods were a step above earlier methods that used Haar-like features [32]. For problems specifically dealing with lower resolution faces, approaches that use generative adversarial networks (GANs) [33], training detectors are different scales [34], and custom loss functions with deep pyramid single shot face detectors [35].

The majority of the new approaches to masked face detection for classification use with an off-the-shelf face detector such as YOLO family [16, 18, 21, 36, 22] or RCNN family [13]. Other works that made further improvements to their detectors through training utilized backbones from the ResNet family such as RetinaFace and [1, 37, 38]. To compare the performance of our new dataset to the existing datasets, we opted to also use this approach and used ResNet50 as the backbone for our baseline.

Masking Classification

Image classification is another well studied problem. AlexNet, VGG16, and VGG19 are common CNN architectures trained on large datasets and are often used in classifications problems [26]. Another such network, MobileNetv2 [39] uses an inverted residual with a linear bottleneck making more efficient than its predecessor for easier real-life applications. One of the more popular models, ResNet50 [40] harnesses residual learning to improve on the existing deep neural network architectures. Combining these networks with transfer learning techniques is widely used in image classification problems [41, 42, 43].

For the masking classification task, many of the new approaches to masked face classification use transfer learning with a pretrained model such as the VGG family [14, 16, 9], MobileNet family [15, 18, 20, 9], or ResNet family [21, 44, 22, 45]. Some opted to train a custom CNN from scratch [11, 19], or use an existing pretrained model for feature extraction and trained custom neural net for classification [16]. A different approach used by the baseline model of the MaskedFace-Net dataset, among others, detects facial landmarks and determines mask label based on which landmarks are visible [46, 47]. For our comparisons, we opted to use the most popular approach among the masking classifiers which applies transfer learning to a pretrained backbone. With the even distribution of backbone choice among the existing methods and no clear frontrunner, we opted to use an Xception backbone for its ease of use and competitive performance [2].

Chapter 3

Method

As outlined in Chapter 2, the major challenges we identified in existing datasets are

- Synthetic datasets improperly handle occlusion, any side or down-turned face orientation, and lack the mask variety seen in the real world.
- Older datasets are heavily skewed to masks in an occupational setting and stock photos, lacking the mask and subject variety seen in the real world during the pandemic. The masks can also sometimes be inappropriate for disease prevention.
- Newer datasets are lacking in diversity in their subjects.

To address these concerns, we have compiled a new dataset Masked Faces in Context (MASON) which is more representative of everyday public mask usage worldwide.

This section will cover the methodology for the creation of the new dataset. We will also summarize the key characteristics of the dataset and compare it to the existing datasets outlined in the Previous Works chapter.

Data Collection

Figure 3.1: Examples of images collected for the new dataset. These show a variety of age, setting, face angle, occlusion, and mask appearance.

The images were manually collected from search engines and image websites such as Google Images, Bing Images, Duckduckgo Images, Flickr, Wikipedia commons, and Pixnio under the “Creative Commons licenses”. Keywords in the search filtered the images to recent (2020 - present) pictures of people wearing masks.

To capture a snapshot of mask-wearing in the wild requires diversity in areas such as people, location, orientation, mask appearance, and occlusions. This was done through the use of key search words in each category, looking for a minimum of 25 images collected per keyword. Though the keywords would target a specific category, they would often also have a secondary or tertiary effect in other categories that needed to be considered as well.

Location words such as countries and cities affect people’s appearance in both clothing and ethnicity as well as provide different backgrounds. More specific locations such as schools, parks, and malls can affect the age of the subjects as well as their physical pose orientations. For example, a “school” image is likely to show children while a “mall” image can have people of all ages. A picture from public transportation such as a train is likely to have people sitting or standing while looking down at their phones while a picture from a park can have someone biking or rollerblading. Spacious locations might see people standing further apart while images of protests tend to have many people standing close together. Densely populated images also saw high levels of occlusion.

Despite seeing a variety in masks naturally through search, having both identified diversity in masks as a limitation of existing datasets and given their importance in both the masked face detection and masking classification tasks, we wanted to ensure we targeted the many different types specifically with descriptor keywords such as cloth, colorful, KN95, and surgical. Similarly, occlusion handling from common objects was ensured with keywords such as sunglasses, beards, hats, head coverings, and face shields.

With this approach, the dataset features a wide variety of naturally occurring face orientations, mask types, occlusion, and people.

Data Annotation

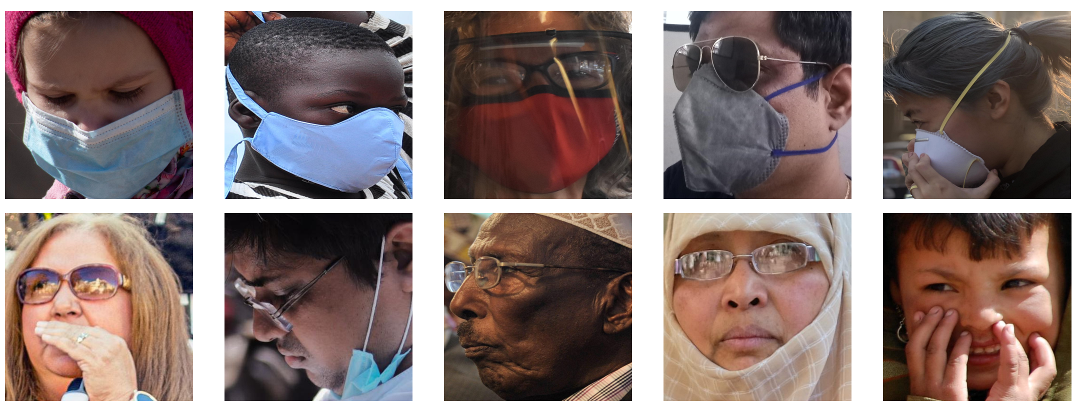


Figure 3.2: Examples of cropped images collected from the new dataset separated into masked and unmasked faces. These include a range of ages, face positions, and occlusion items such as glasses and face shields.

For the annotations we wanted address both the masked face detection and masking classification problems. Masked face detection requires specific face location in the overall image and masking classification requires categorical labels for masked or not masked.

The first step in annotating the new dataset was to locate all of the visible faces in the image, masked or unmasked. This was done manually with LabelImg by drawing bounding boxes on each face and labeling each box as “masked” or “not masked”. The labels were saved in a .txt file which included the masking label followed by the bounding box upper right corner pixel location followed by the width and height of the bounding box. The “masked” category only included correctly masked faces as described by the CDC [4], with all of the incorrectly or partially masked faces falling in the “not masked” category. Examples of the full images are shown in Figure 3.1.

The second step of annotation was to extract all of the faces and divide them by their masking label. To do this we used the bounding boxes and labels to crop the images around each face and sort them into folders corresponding to each label. Examples of the cropped images are shown in Figure 3.2.

The final result is a collection of 2058 original images with corresponding labels detailing the location of each face and the face’s masking classification. It is then also presented as 7716 individual face images sorted in the “Masked” and “Not Masked” categories.

Comparisons

Dataset	**# Images**	**# Face Images**	**Synthetic**	**Labels**
MaskedFace-Net	133,782	133,782	Yes	“Masked”/“Incorrectly Masked”
MAFA	30,811	35,806	Some	Face location, occlusion level, occlusion type
Real-World-Masked -Face-Dataset	92203	92203	Some	“masked”/“not masked”
MASON	1627	5903	No	face location, masked or not masked

Table 3.1: Logistics of previous and new datasets. While many of the datasets contain a larger number of images, they struggle to suit the Covid-19 mask detection problem due to their synthetic nature, inappropriate labels, and/or unrepresentative content.

Table 3.1 summarizes some of the logistical differences between the previous datasets and our new dataset MASON. Though the new dataset is the smallest in the total number of images, the other improvements it makes over existing datasets makes it best suited to the problems outlined in this paper. Compared to the synthetic datasets, it avoids the occlusion and warping issues that come with digitally placing masks over faces. It also more accurately captures masking under the Covid-19 pandemic in terms of mask appearance and people behavior. The labels for the new dataset are suited to both the face detection and masking classification problems while MaskedFace-Net and RWMFD only cater to the classification problem and MAFA only addresses the detection problem.

Dataset	**Mask Variety**	**Image Type**	**Year**
MaskedFace-Net	Only Surgical Mask	Headshots only	2020
MAFA	Variety	Headshots and Stock Photos	2017
Real-World-Masked -Face-Dataset	Variety	Headshots or closeups	2020
MASON	Variety	In-the-wild, and Headshots	2020-2022

Table 3.2: Summary of previous and new dataset features. The new dataset presents a more representative snapshot of mask wearing under Covid-19 than the previous 3 datasets.

Table 3.2 breaks down further logistical differences between MASON and the existing MaskedFace-Net, MAFA, and RMFD. MASON is the first to include real-world images from the Covid-19 pandemic featuring and labeling multiple people per image. Since the focus of the new dataset is on in-the-wild images taken during the Covid-19 pandemic, the dataset more accurately captures mask-wearing behavior compared to the datasets published at the beginning of the pandemic in 2020.

To further break down the dataset comparisons, each dataset was analyzed for subject comparisons, face pose comparisons, and mask appearance comparisons.

In the subject comparisons we looked primarily at the age of the subjects and the data type. The age groups were visually sorted into “child”, “adult”, and “elder” and counted in a sample of each of the datasets. In the context of Covid-19, it is even more important that every age group is represented as age is considered a major factor in Covid-19 risk [3]. Synthetic images were also counted along with age as different age groups will fit and wear masks differently.

For face pose comparisons we looked at whether each face was facing forwards, downwards, sideways, or a combination of the groups. The direction will significantly affect the appearance of the faces and masks.

In the mask appearance comparisons we looked at common occlusions and mask variety. We noted the occurrence of three common occlusion items that can interfere with mask appearance: face shields, eyewear, and head coverings. We also observed mask variety outside of the blue surgical mask commonly featured in synthetic datasets.

Subject Age and Type

Dataset	**Synthetic**	**Child**	**Adult**	**Elder**
MaskedFace-Net	100.0%	24.5%	70.0%	5.5%
MAFA	1.5%	8.6%	91.4%	0.5%
RMFD	9.5%	0.5%	98.5%	1.0%
MASON	0%	12.1%	75.0%	16.9%

Table 3.3: Subject comparison between MaskedFace-Net, MAFA, RMFD, and MASON looking at a breakdown of the images containing different age groups and whether an image was synthetically made or not.

As seen in Table 3.3, MASON has a greater age variety and is the only dataset to contain no synthetic data. This inclusion of different age groups is important in creating an accurate snapshot of mask-wearing in public. The new dataset especially improves on the proportion of images containing elderly people – a group heavily impacted by the Covid-19 pandemic. The second most diverse dataset is MaskedFace-Net which is adapted from the well established FFHQ. A comparisons of the performance of these two will be in Chapter 4.

Face Pose

Dataset	Forward	Side	Down
MaskedFace-Net	87.0%	11.0%	4.5%
MAFA	69.0%	35.0%	8.1%
RMFD	75.5%	19.5%	5%
MASON	38.3%	48.0%	25.0%

Table 3.4: Face pose comparison between each dataset looking at the breakdown of the different possible face poses (forwards, sideways, downwards) where a mask is still visible.

The second area of comparison was the direction person's head was facing (forwards, sideways, downwards) as seen in Table 3.4. Of the datasets, MASON is the most evenly distributed in face positions where a mask could be seen. A mask's appearance can vary with the angle at which it is seen so the inclusion of variety in the face pose is important to be able to detect masks in public settings where people's face direction cannot be controlled. Many of the existing datasets such as MaskedFace-Net and RMFD focused primarily on the forward-facing "headshot" style images which, even under common data augmentation techniques, do not cover the appearance of a sideways-facing mask.

Mask Appearance

Dataset	Non-Surgical	Face Shield	Eyewear	Head Covering
MaskedFace-Net	0.0%	0.0%	20.0%	10.5%
MAFA	50.3%	0.0%	12.7%	28.4%
RMFD	17.0%	0.0%	2.5%	3.5%
MASON	54.8%	5.4%	42.8%	42.2%

Table 3.5: Mask appearance comparison between datasets looking at the inclusion of common occluding objects and the variety of mask type.

The third major area of comparison is mask appearance as shown in Table 3.5. In addition to common occlusion item such as eyewear and head coverings, the Covid-19 pandemic also popularized face shields as a disease prevention tool which can also cover and distort the face in images. MASON is the only dataset to contain face shields in the images and has the greatest proportion of images that contain the two other common occlusion items. Though MaskedFace-Net includes many of the occlusion items, its synthetic nature means the items do not occlude the mask properly in the images. We also compared the proportion of images that used a non-surgical mask and found MASON had the greatest variety in that area as well.

These three comparisons demonstrate key improvements the MASON has over the existing datasets and the resulting performance gain can be seen in Chapter 4

Chapter 4

Results and Discussions

This section covers the experiments, results, and discussions comparing the newly created MASON dataset with the existing MaskedFace-Net, MAFA, and RMFD datasets. The comparisons will highlight overall performance gains in both the masked face detection problem and the masking classification problem as well as specific case analyses.

To evaluate the efficacy of the new dataset in masked face detection, we compared the performance of a model trained on MASON to that of one trained on the existing datasets. Of the existing datasets, only MAFA had appropriate labels for this task.

The baseline model consisted of FAN [1], the face-focused network built off of RetinaNet using ResNet50 as a backbone. It inputs an image and returns bounding box locations for detected faces. For these experiments, the models were trained on each of the datasets for 50 epochs.

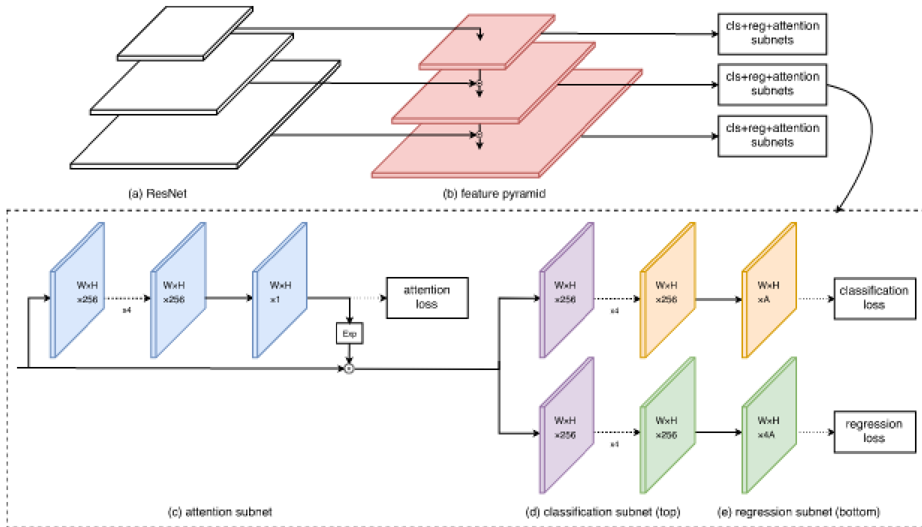


Figure 4.1: Overview of the FAN architecture [1]

The models were first evaluated for accuracy and mAP. They then were tested on a separately collected test set which was broken down into each of the categories outlined in Chapter 3 and compared on the percentage of faces accurately detected per category.

To evaluate the new MASON dataset for masking classification, we compared the performance of a model trained on MASON to those trained on existing datasets. For this task, MaskedFace-Net and RMFD were the two existing datasets with appropriate labeling.

The model used in this experiment consists of the Xception backbone commonly used in similar transfer learning classification applications. It inputs a cropped image of a face, determines the probability the image is of a masked face or an unmasked face, and outputs the binary classification label corresponding to the greater probability. For this experiment, a model was trained for each of the datasets for 5 epochs.

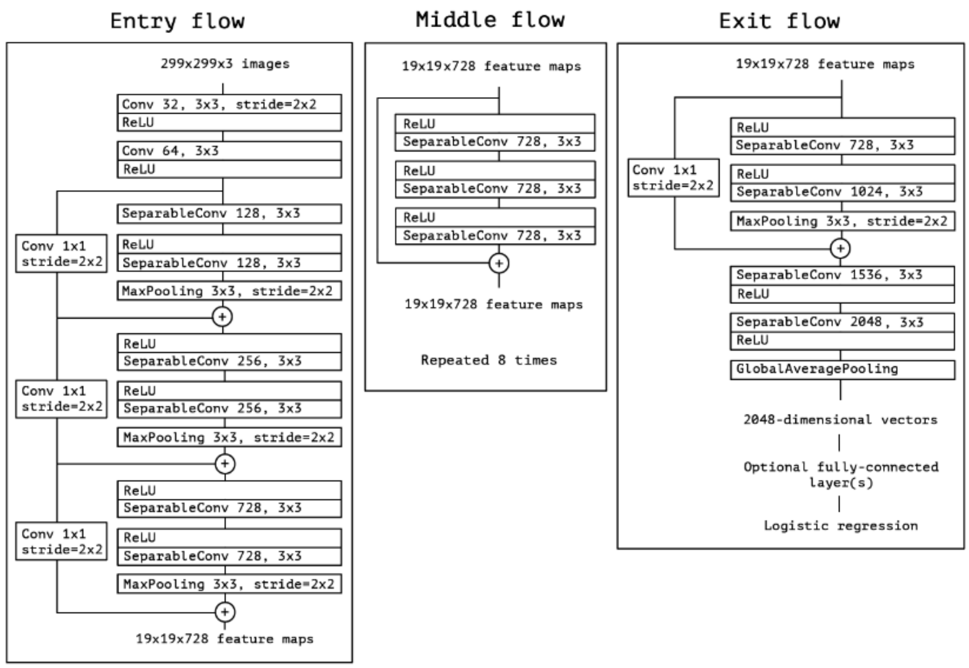


Figure 4.2: Overview of the Xception architecture [2]

To compare the strengths and weaknesses of each of the individually trained models, we tested the models on a separately gathered test set and looked at the accuracy, precision, recall, F1 score, ROC, and AUC performance for each model. The test set was then broken down into the categories outlined in Chapter 3 for more specific comparisons.

Detection Performance

Dataset	mAP	Accuracy
MAFA	0.1221	0.52
MASON	0.2558	0.86

Table 4.1: Accuracy and mAP of the MAFA-trained and MASON-trained models.

MASON MAFA Ground Truth

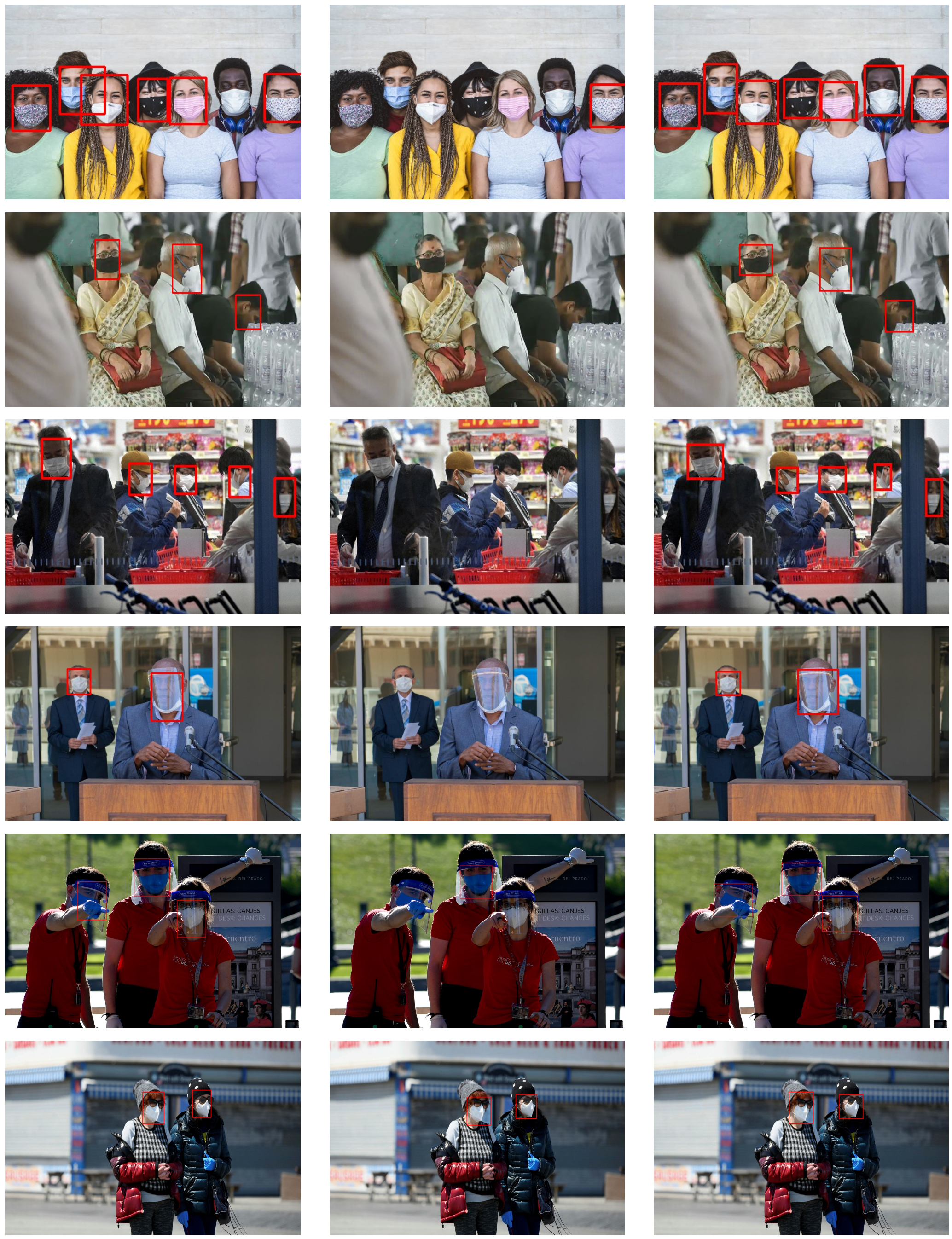


Table 4.2: Examples of test images run on the MASON-trained model and MAFA-trained model. MASON was able to detect more faces in challenging circumstances such as occlusion, face shields, and varied face poses.

As seen in Table 4.1, the model trained on the MASON dataset greatly outperforms the one trained on the MAFA dataset in both metrics. The key improvement areas can be further demonstrated in the images in Table 4.2. The model trained on MASON can handle a more unusual-looking mask, a side profile, and a face occluded by eye and head coverings while the MAFA-trained model fails in those cases. These correspond directly to the improvements outlined in the last chapter where MASON has proportionally more variety in mask appearance and a greater proportion of images with a sideways pose and common occluding objects.

We then qualitatively broke down the test set into the specific categories identified in Chapter 3 as areas of improvement: subject age, subject face pose, and subject occlusion.

Age

Dataset	**Child**	**Adult**	**Senior**
MAFA	0.416	0.301	0.358
MASON	0.714	0.681	0.791

Table 4.3: Percentage of each age category accurately detected by each model.

The first category is subject age. As outlined in Chapter 3, one of the focuses of the new dataset was to have a better distribution of subjects across the three age groups: child, adult, and senior. To evaluate efficacy in this area, we manually classified each of the faces in the testing set into the three categories, and looked at how well each of the models detects in each category. As seen in Table 4.3, the MASON trained model outperforms the MAFA trained model in every category, most significantly the senior category which, as seen in Table 3.3, was barely represented in the MAFA dataset.

Face Pose

Dataset	Forward	Sideways	Downward
MAFA	0.401	0.352	0.367
MASON	0.818	0.655	0.700

Table 4.4: Percentage of each face pose category accurately detected by each model.

The second category is the subject face pose. Using the same approach as for the "age" category, we first manually classified each of the faces in the test set as forwards facing, sideways facing, and/or downwards facing and then compared the detection accuracies between the two models for each. As seen in Table 4.4, MASON outperforms MAFA in every category, most significantly in the downward-facing face pose category which is the most underrepresented in the MAFA dataset in Table 3.4. As the lack of more naturally posed faces was a critique of the existing datasets, this improvement demonstrates how MASON is better representative of real life face poses.

Mask Appearance

Dataset	Face Shield	Glasses	Head Covering	Non Surgical Mask
MAFA	0.531	0.414	0.423	0.392
MASON	0.714	0.770	0.731	0.685

Table 4.5: Percentage of each mask appearance category accurately detected by each model.

The third category is mask appearance which covers both common occlusions and mask type. For each of the most common occlusion types we identified, which face shields, glasses, and head coverings, we manually classified each of the faces in the test set and compared the performance of the two models within each category. We then also looked at the performance of each of the two models on masks that did not appear to be the common blue surgical mask. As seen in Table 4.5 the MASON trained model outperformed the MAFA trained model in very category.

Overall the MASON-trained detection model performed better than the MAFA-trained detection model in every category. With the Covid-19 focused masks, subject age, face pose, and occlusion, the MASON dataset appears to be more suited to the face detection task under Covid-19 than the older MAFA dataset.

Classification Performance

Dataset	**Accuracy**	**Label**	**Precision**	**Recall**	**F1-Score**
MaskedFace-Net	0.69	Mask	**0.91**	0.62	0.74
		No Mask	0.50	**0.86**	0.63
RMFD	0.85	Mask	0.83	0.97	0.90
		No Mask	0.90	0.56	0.69
MASON	**0.90**	Mask	0.88	**0.98**	**0.93**
		No Mask	**0.94**	0.70	**0.80**

Table 4.6: Accuracy, precision, recall, and f1-score of baseline model trained on each of the datasets. In accuracy and f1-score the new MASON dataset outperforms the existing datasets.

As seen in Table 4.6, the baseline model trained on MASON showed a significant increase in overall accuracy compared to the existing datasets. It also performed the best in F1 scores for both the masked and not-masked classes, indicating a better balance of precision and recall than the other two models despite not outperforming them in both categories.

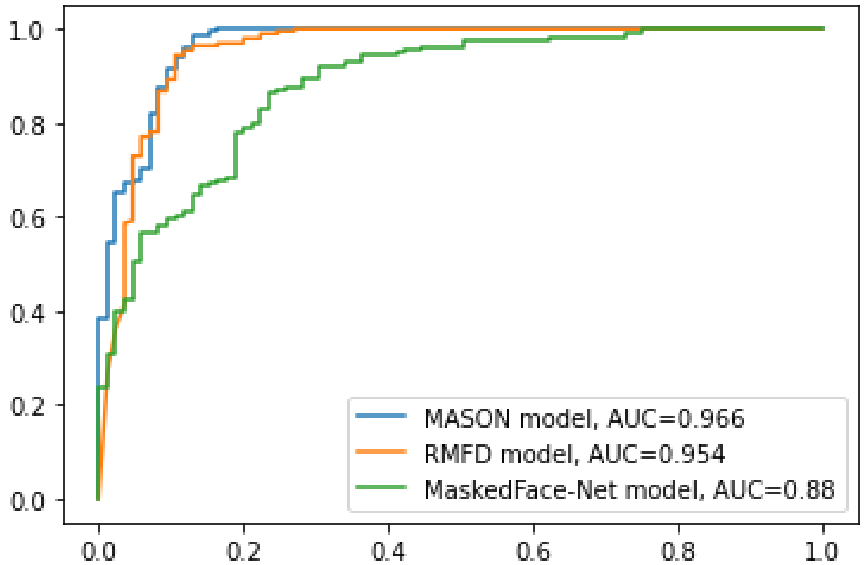


Figure 4.3: ROC and AUC score for each of the models.

Figure 4.3 visualizes a more comprehensive comparison between the models. The graph shows the True Positive by False Positive rate for each of the three classifiers (ROC) and calculates their respective areas under the curve (AUCs). For each classifier, AUC indicates the likelihood that a positive case will be more likely to be classified as positive than a negative case [48]. As shown in the figure, the AUC of the MASON-trained classifier is greater than that of the RMFD and MaskedFace-Net-trained classifier, indicating better performance.

Pixels / **Dataset**	$< 37^2$	$37^2 - 70^2$	$70^2 - 145^2$	$> 145^2$
MaskedFace-Net	0.714	0.737	0.604	0.716
RMFD	0.724	0.859	0.906	0.902
MASON	**0.819**	**0.907**	**0.911**	**0.944**

Table 4.7: Accuracy of each model broken down by image resolution. The model trained on the MASON outperforms the models trained on the other two in every category. The most significant improvement lies in the low resolution group.

We then looked at accuracy with respect to image resolution. In real-world images, people appear at different distances from the camera, resulting in low resolution for the further away faces. In Table 4.7 we compared the difference in accuracy between each of the datasets on four resolution groups. In all four groups, the MASON outperformed the existing datasets with the most significant improvement being in the smallest image group with a width and height of fewer than 37 pixels. This is a good indicator that for applications with low image resolution from varying depths, MASON will be more effective at classifying masked faces.

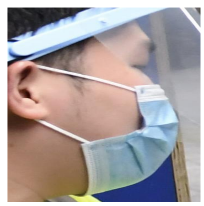
Dataset	 headscarf	 side orientation	 occlusion	 colorful mask
MaskedFace-Net	No Mask	No Mask	No Mask	No Mask
RMFD	Mask	No Mask	No Mask	Mask
MASON	No Mask	Mask	Mask	Mask
Ground Truth	No Mask	Mask	Mask	Mask

Figure 4.4: Examples of mask classification in cases of hair coverings, side face orientation, occlusion, and unusual mask coloring.

The key areas where MASON performed better than existing datasets were images with occlusion, images featuring the side profile of the face, and unusually colored or patterned masks. Figure 4.4 shows examples of these categories. The first image is an example of a face partially occluded by a head covering (in this case a head scarf) and glasses. The second shows a sideways-facing face with a face shield occlusion. The third is also sideways-facing and also has occlusion from the glasses as well as from another person's shoulder. The fourth has no occlusions but a colorful, patterned mask. Each of these examples were correctly classified by the MASON trained model and incorrectly classified by at least one of the other two models.

We then followed these qualitative observations with a break down of our quantitative results into the three categories outlined in Chapter 3: Age of subject, face pose, and mask appearance.

Age

Dataset	Child	Adult	Senior
MaskedFace-Net	0.628	0.644	0.821
RMFD	0.948	0.926	0.881
MASON	**0.961**	**0.948**	**0.925**

Table 4.8: Accuracy of each model broken down by subject age. The model trained on MASON outperforms that of the other two in every category and exhibits more balanced performance across each category.

The first category is subject age separated into child, adult, and senior. As outlined in Chapter 3 Table 3.3, MASON is more evenly distributed than the other datasets across the age groups, most notably in the senior group. For this test we manually sorted the subjects in the test images into child, adult, and senior groups and looked at the accuracy of each model for every group. In Table 4.8, the MASON-trained model outperforms the other two in every group, most significantly in the senior group, indicating that the MASON model is more representative of the general population than models trained on previous datasets.

Face Pose

Dataset	Forward	Sideways	Downward
MaskedFace-Net	0.751	0.613	0.517
RMFD	0.927	0.901	0.850
MASON	**0.949**	**0.944**	**0.933**

Table 4.9: Accuracy of each model broken down by face pose of the subject.

The next category is subject face pose which can be described as forwards facing, sideways facing, and/or downwards facing. The MASON-trained model again outperforms the other two in every category and more evenly across categories as seen in Table 4.9. While MaskedFace-Net and RMFD see a drop in performance of almost 10% in the underrepresented downward category compared to the other two categories, the MASON-trained model sees a smaller drop of about 1%.

Mask Appearance

Dataset	Face Shield	Glasses	Head Covering	Non Surgical Mask
MaskedFace-Net	0.571	0.655	0.808	0.515
RMFD	0.837	0.920	0.936	**0.977**
MASON	**0.898**	**0.931**	**0.974**	**0.977**

Table 4.10: Accuracy of each model broken down mask and occlusion type. The three occlusion categories were the most common sources of occlusion as described in Chapter 3. The mask type describes any mask that does not appear to be the most common blue surgical mask.

The final category is mask appearance which covers subject occlusion from common sources and mask variety. As shown in Table 4.10, the MaskedFace-Net dataset which only uses blue surgical masks performs significantly worse than RMFD and MASON which include a wide spread of masks. It also shows that the only model trained on face shields, the MASON-trained model, achieved almost 90% accuracy on faces covered by face shields compared to the RMFD-trained model at almost 84% and MaskedFace-Net-trained model at 57%. These improvements as well as the overall success of the MASON trained model in every category highlight its strengths in masking classification in the wild images.

With the new MASON dataset, we saw improvements in overall model performance and notably in the specific areas we identified as limitations in the existing datasets. Subject age, face orientation, occlusion, and mask variety were all areas where MASON contributed to a significant improvement in accuracy while maintaining performance in areas where the existing datasets were strong. We also saw a significant increase in performance for low-resolution images which further demonstrates the dataset's effectiveness in real-world mask classification applications.

Chapter 5

Conclusion & Future Steps

This chapter summarizes the newly created dataset MASON for face and mask detection during the Covid-19 pandemic and discusses its impact potential outside of its field. It then reviews the experimental results, limitations of the work, and potential future steps.

Conclusion

The newly created Masked Faces In Context (MASON) dataset provides a COVID-19 era snapshot of public masking with masked face detection and masking classification labels. Existing masking datasets were limited in their subject diversity and lacking in Covid-19 specific appearance and behavior such as the use of face shields. With the inclusion of more recent images and focusing on diversity, MASON overcomes these limitations and fills a gap in Covid-19 studies, and serves as an effective tool for model training and evaluation. Comparisons between models trained on the existing datasets and MASON show the new dataset to be an improvement in both the masked face detection and mask classification problems.

This work is currently being used in a proof of concept for mask-wearing evaluation studies in a public park. By tracking the movement of people along with their masked status, studies can more quickly understand mechanisms for Covid-19 spread and the effectiveness of masks as a prevention method. With the application of these models in public spaces such as public transportation and indoor businesses, policymakers can more easily and quickly see the effects of shutdown, masking, or social distancing policies and determine the necessary steps to control disease spread. These models can also provide people with a more specific analysis of risk in different public spaces and allow people to make more informed decisions about their movements. We believe our dataset will benefit epidemiologists in their studies of current and new diseases and public policymakers in their public health decisions as well as the general public.

Limitations and Future Steps

While we have baseline results demonstrating an improvement over the existing datasets, there is still room for optimization in both masked face detection and mask classification. Though they demonstrated overall improvement, both the classification and detection models struggled with lower-resolution images. Further training and model design tuning can improve the accuracy over the baseline presented in this paper.

The natural next steps for this work would be applications in public spaces. Running these models on real-time video will bring up further computational considerations and frame-to-frame continuity. These models will need to be able to run at an FPS fast enough to accurately capture every face. With people appearing in a series of frames, it will need strategies to not double-count people. Finally, the baseline output of bounding boxes with classification labels may require more post-processing for specific study analyses.

References

[1] Jianfeng Wang, Ye Yuan, and Gang Yu. Face attention network: An effective face detector for the occluded faces, 2017.

[2] François Chollet. Xception: Deep learning with depthwise separable convolutions, 2016.

[3] U.S. Department of Health and Human Services. Basics of covid-19, Mar 2022.

[4] U.S. Department of Health and Human Services. Mask guidance, Mar 2022.

[5] Jicheol Park, Jin-Hee Lee, and Sang Hyuk Son. A survey of obstacle detection using vision sensor for autonomous vehicles. In *2016 IEEE 22nd International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, pages 264–264, 2016.

[6] Andreas Opitz and Andreas Kriechbaum-Zabini. Evaluation of face recognition technologies for identity verification in an egate based on operational data of an airport. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–5, 2015.

[7] Rola EL SALEH, Sambit BAKHSHI, and Amine NAIT-ALI. Deep convolutional neural network for face skin diseases identification. In *2019 Fifth International Conference on Advances in Biomedical Engineering (ICABME)*, pages 1–4, 2019.

[8] 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*, 2020.

[9] Busra Kocacinar, Bilal Tas, Fatma Patlar Akbulut, Cagatay Catal, and Deepti Mishra. A real-time cnn-based lightweight mobile masked face recognition system. *IEEE Access*, 10:63496–63507, 2022.

[10] Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, Heling Chen, Yu Miao, Zhibing Huang, and Jinbi Liang. Masked face recognition dataset and application, 2020.

[11] Jumana Waleed, Thekra Abbas, and Taha Mohammed Hasan. Facemask wearing detection based on deep cnn to control covid-19 transmission. In *2022 Muthanna International Conference on Engineering Science and Technology (MICEST)*, pages 158–161, 2022.

[12] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. Detecting masked faces in the wild with lle-cnns. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 426–434, 2017.

[13] Indah Agustien Siradjuddin, Reynaldi, and Arif Muntasa. Faster region-based convolutional neural network for mask face detection. In *2021 5th International Conference on Informatics and Computational Sciences (ICICoS)*, pages 282–286, 2021.

[14] Sahana Rangasrinivasan, Sri Lohitha Bhagam, Nair K. Athira, Kondapi Niharika, Anjuna D. Raj, and T. Anjali. Covimask: A novel face mask type detector using convolutional neural networks. In *2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, pages 24–27, 2022.

[15] Nazmus Shakib Shadin, Silvia Sanjana, and Daiyan Ibrahim. Face mask detection using deep learning and transfer learning models. In *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pages 196–201, 2022.

[16] Fauzan Firdaus and Rinaldi Munir. Masked face recognition using deep learning based on unmasked area. In *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, pages 1–6, 2022.

[17] Ishwarya TM and K. Naveen Durai. Detection of face mask using convolutional neural network. In *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 2008–2012, 2022.

[18] Aneesh. R, Prasanth. K, and P. Madhavan. Social distancing and face mask detection for covid prevention. In *2022 8th International Conference on Smart Structures and Systems (ICSSS)*, pages 1–4, 2022.

[19] Dinesh Sharma and Geetam Singh Tomar. Face mask detection analysis for covid19 using cnn and deep learning. In *2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT)*, pages 239–252, 2022.

[20] Ruchi Sharma, Anmol Sharma, Raunak Jain, Soham Sharma, and Sreshtha Singh. Face mask detection using artificial intelligence for workplaces. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1003–1008, 2022.

[21] Rupali Kolpe, Shubham Ghogare, M.A. Jawale, P. William, and A.B. Pawar. Identification of face mask and social distancing using yolo algorithm based on machine learning approach. In *2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1399–1403, 2022.

[22] Vivek Aswal, Omkar Tupe, Shifa Shaikh, and Nadir N. Charniya. Single camera masked face identification. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 57–60, 2020.

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks.

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[25] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, 2016.

[26] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 502–511, 2019.

[27] Bilel Benjdira, Taha Khursheed, Anis Koubaa, Adel Ammar, and Kais Ouni. Car detection using unmanned aerial vehicles: Comparison between faster r-cnn and yolov3. In *2019 1st International Conference on Unmanned Vehicle Systems-Oman (UVS)*, pages 1–6, 2019.

[28] N. Murali Krishna, Ramidi Yashwanth Reddy, Mallu Sai Chandra Reddy, Kasibhatla Phani Madhav, and Gaikwad Sudham. Object detection and tracking using yolo. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1–7, 2021.

[29] Rakhsith L. A, Anusha K. S, Karthik B. E, Arun Nithish D, and Kishore Kumar V. A survey on object detection methods in deep learning. In *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1619–1626, 2021.

[30] Rajeev Ranjan, Vishal M. Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, 2019.

[31] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2020.

[32] Moh. Edi Wibowo, Ahmad Ashari, Ardacandra Subiantoro, and Wahyono Wahyono. Human face detection and tracking using retinaface network for surveillance systems. In *IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–5, 2021.

[33] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Finding tiny faces in the wild with generative adversarial network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21–30, 2018.

[34] Peiyun Hu and Deva Ramanan. Finding tiny faces. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1522–1530, 2017.

[35] Rajeev Ranjan, Ankan Bansal, Jingxiao Zheng, Hongyu Xu, Joshua Gleason, Boyu Lu, Anirudh Nanduri, Jun-Cheng Chen, Carlos D. Castillo, and Rama Chellappa. A fast and accurate system for face detection, identification, and verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):82–96, 2019.

[36] Benedetta Liberatori, Ciro Antonio Mami, Giovanni Santacatterina, Marco Zullich, and Felice Andrea Pellegrino. Yolo-based face mask detection on low-end devices using pruning and quantization. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 900–905, 2022.

[37] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2017.

[38] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild, 2019.

[39] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[41] Zhi Yi and Yuyang Wang. Transfer learning on interstitial lung disease classification. In *2021 International Conference on Signal Processing and Machine Learning (CONF-SPML)*, pages 199–205, 2021.

[42] Ling Shao, Fan Zhu, and Xuelong Li. Transfer learning for visual categorization: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 26(5):1019–1034, 2015.

[43] Ahmet Haydar Örnek, Mustafa Çelik, and Murat Ceylan. Mask detection from face images using deep learning and transfer learning. In *2021 15th Turkish National Software Engineering Symposium (UYMS)*, pages 1–4, 2021.

[44] Omar Adel Muhi, Mariem Farhat, and Mondher Frikha. Transfer learning for robust masked face recognition. In *2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–5, 2022.

[45] Devrim Kayali, Kamil Dimililer, and Boran Sekeroglu. Face mask detection and classification for covid-19 using deep learning. In *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6, 2021.

[46] B. Thaman, T. Cao, and N. Caporusso. Face mask detection using mediapipe facemesh. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 378–382, 2022.

[47] 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. *Computer Modeling in Engineering & Sciences*, 124(3):1049–1059, 2020.

[48] Jin Huang and C.X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3):299–310, 2005.