SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Performance Evaluation of State-of-The-Art 2D Face Recognition Algorithms on Real and Synthetic Masked Face Datasets

Mohammad Amir Khan[1], Ahmed Rimaz Faizabadi[1], Muhammad Mahabubur Rashid[1,*], Hasan Firdous Zaki[1]

[1] Dept. of Mechatronics Engineering, International Islamic University Malaysia, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 27 January 2023
Received in revised form 27 February 2023
Accepted 29 March 2023
Available online 21 April 2023

ABSTRACT

Face recognition systems based on Convolutional neural networks have recorded unprecedented performance for multiple benchmark face datasets. Due to the Covid-19 outbreak, people are now compelled to wear face masks to reduce the virus's transmissibility. Recent research shows that when given the masked face recognition scenario, which imposes up to 70% occlusion of the face area, the performance of the FR algorithms degrades by a significant margin. This paper presents an experimental evaluation of a subset of the MFD-Kaggle and Masked-LFW (MLFW) datasets to explore the effects of face mask occlusion against implementing seven state-of-the-art FR models. Experiments on MFD-Kaggle show that the accuracy of the best-performing model, VGGFace degraded by almost 40%, from 82.1% (unmasked) to 40.4% (masked). On a larger-scale dataset MLFW, the impact of mask-wearing on FR models was also up to 50%. We trained and evaluated a proposed Mask Face Recognition (MFR) model whose performance is much better than the SOTA algorithms. The SOTA algorithms studied are unusable in the presence of face masks, and MFR performance is slightly degraded without face masks. This show that more robust FR models are required for real masked face applications while having a large-scale masked face dataset.

Keywords:
Facemasks; Data Sets; Face Recognition; Convolutional Neural Networks

1. Introduction

It is not uncommon for Face Recognition (FR) systems to be shown facial features such as eyes, noses, and mouths that are not impeded by anything else on the face. A wide variety of conditions, however, necessitate masks that wholly or partially obscure the faces of those who wear them. Pandemics, laboratories, medical procedures, and excessive pollution are just a few of these widespread occurrences. The best way to prevent COVID-19 is to wear masks and practice social distancing. Since every country in the world now mandates protective face masks in public places, researchers have had to dig deeper into how these facial recognition systems work while faces are covered up. Face recognition is an issue since the obstructed parts are essential for face detection and recognition [1]. However, secure authentication solutions that rely heavily on FR could be in

[*] *Corresponding author.*
E-mail address: mahbub@iium.edu.my

https://doi.org/10.37934/araset.30.2.225242

jeopardy from new regulations. Most of the recent algorithms are concerned with determining not whether a mask conceals a face and mask face detection. It is of the utmost importance to have a method for authenticating individuals who wear masks without revealing their faces. The impact of changing scenarios on existing FR-based authentication systems and the newly proposed Masked Face Recognition (MFR) system is not trivial and will be the subject of an investigation in this paper.

In recent years, deep learning technologies have made enormous strides in theoretical understanding and actual use. FR systems increasingly use deep learning models since it is a cutting-edge research area in computer vision. In light of the COVID-19 pandemic, a compelling situation has arisen in which the performance of masked facial recognition systems must be examined. The masks reduce face visibility by up to 70 percent [26] leaving just the eyes and forehead visible. The mask face datasets are scarce, and researchers are applying artificial masks on the unmasked face. Such a dataset is called a synthetic or simulated mask face dataset.

Ngan *et al.,* [2] used the original unmasked images to set a baseline for accuracy, digitally applied a mask to the face, and then evaluated the face recognition algorithms. Wang *et al.,* [3] used a feature pyramid network in their proposed face attention network (FAN). The distinctive layers of this neural network were utilized to resolve faces of varied sizes, providing distinct attention areas aimed at each feature layer. As a result, several more studies based on changes in attention strategies have been proposed in the literature, using supplementary network models to define a facial region of interest (ROI) on behalf of the feature extraction. However, when face masks are used, the attention maps fail to identify the ROI because they approximate facial appearances.

In addition, several approaches to recovering the clean faces hidden behind the obscured ones have been presented. Occlusion-resistant faces can be encoded, and the occluded part can be restored using the Robust LSTM-Autoencoders (RLA) model, which Zhao *et al*., [4] proposed. Occluded regions were removed using the Iterative Closest Point (ICP) approach in the work of King *et al.,* [5]. They restored the image using an arithmetical approximation of curves to deal with the obstructed areas. In contrast to partial face and occlusion, which only obscure some facial features used for face recognition, masking the significant facial cues makes masked face identification far more difficult. Face-occluded pictures can be effectively synthesized and recognized by GAN (Generative Adversarial Network) BoostGAN for large-pose variation and simultaneous corrupted regions [6]. However, the recovery methods for unknown identities under large occlusion are doubtful and poorly established for the FR system.

Ding *et al.,* [7] developed a new hidden part revealing (HPR) model to find the latent facial portion unaffected by face mask use. Geng *et al*., [8] presented a Domain-Constrained Ranking (DCR) loss based on a cross-domain center-based ranking algorithm. Two centers are created for each identity, one for the entire face picture and another for the masked face image. For masked facial features to be pushed closer to their full-face counterparts, the DCR is used to force them. First, face completeness is explicitly enforced, and then knowledge is transferred from an already pre-trained generic face recognition model using knowledge purification, as proposed by Li *et al*., [9]. using features learned only from the area around the eyes of the face pictures. Li *et al*., [9] presented a Convolutional Block Attention Module (CBAM) for masked face identification.

However, these works are not evaluated on real mask face datasets, and even if evaluated, the datasets are very small in close-set conditions. Despite the importance of masked face recognition in today's world, there appears to be a lack of relevant literature on the impact of real and synthetic face masks on MFR. Many attempts do not replicate the real-world scenario or open-set FR system for MFR. Further, to the best of our knowledge, there is no attempt to gauge the state-of-the-art (SOTA) FR algorithms in the presence of face masks.

Considering the gaps identified in the literature, a framework for evaluating SOTA FR algorithms are developed for faces covered with masks. A real-mask and synthetic mask face dataset are curated and human-verified for training and testing MFR systems. A complete experimental protocol for these datasets is developed to measure the MFR system. Apart from evaluating SOTA algorithms, a proposed MFR system is developed and benchmarked. Both simulated or synthetic masked face MLFW datasets and real masked face MFD-k datasets have been used to evaluate the effectiveness of these algorithms. To summarize, our major contributions are:

i. Detailed Investigation on the effect of large occlusion through face masks on SOTA face recognition algorithms with proposed evaluation framework on real and synthetic datasets.
ii. Curated a real masked face dataset for training and testing and developed a complete evaluation protocol.
iii. We have developed a synthetic masked face dataset from VGGFace2 with four different mask types and proposed a MaskVGGFace2- mini dataset for training and ablation study.
iv. Developed a masked face recognition algorithm and benchmarked MFR datasets for uncovered faces and faces with masks with further recommendations to improvise the MFR system.

The paper is organized into section 1, an Introduction covering the significance and background study of masked face recognition, highlighting existing gaps and contributions of this paper. Following section 2 is about the methodology of FR algorithm selection with evaluation framework, dataset creation, benchmarking protocols, and proposed MFR algorithm. Section 3 describes results and discussion for evaluation of the SOTA algorithm not trained with face covered with face masks on real and synthetic datasets along with proposed MFR results trained on masked faces. Finally, conclusions are made in section 4.

2. Methodology

Many CNN models have reached the pinnacle of face recognition performance. The aim is to see if the models can withstand the occurrence of masks on their faces and see their effect on the model. To correlate masked face recognition to unmasked face recognition using systems trained solely on standard images of faces without face masks. There are numerous face-recognition algorithms in the literature. SOTA FR algorithms are selected based on their performance on a benchmark dataset called Labelled Faces in the Wild (LFW) [10]. We only evaluate seven well-known CNN models showing impressive results on LFW. The method for the SOTA evaluation framework and each FR algorithm used in the evaluation is discussed in subsection 2.1.

2.1. SOTA Algorithms and Proposed Evaluation Framework

We evaluated seven state-of-the-art algorithms with significant benchmark performance on the LFW dataset for our research. However, these algorithms are not trained on masked faces, so their performance in changing scenarios must be investigated. For evaluation purposes, we use both real and synthetic datasets. The real masked face dataset, the MFD-Kaggle dataset, is curated and human-verified from internet sources [11]. An MLFW [12] is used to evaluate MFR on synthetic masked faces, as discussed in section 2.2. Along with the method for a benchmarking protocol for evaluating MFR.

The seven models used are, DeepFace by Taigman *et al.,* [13]. presented a deep CNN architecture called DeepFace in 2014. For the first time, architecture was able to attain on the LFW dataset an accuracy of 97.35 %, which was nearly as good as human performance in an unconstrained situation. As a result of substituting the traditional face recognition procedural pipelines with a few novel ones, they extracted face representations layer by layer. On a dataset of 4.4 million faces from 4000 people, they trained a 9-layer deep network architecture on this data. With the help of three neural networks, they could attain this level of accuracy.

Another model used was DeepID, a DCNN architecture described by Wang *et al.,* [14], which obtained 97.45% accuracy on the LFW dataset. They suggested that the CNN architecture could extract a feature vector known as DeepID by learning around 10,000 faces in the preliminary layers lowering the activation functions progressively. DeepID2 was subsequently introduced to minimize the inter-and intra-individual variability in the face in an enhanced model of DeepID. Later, they proposed other CNNs, based on GoogLeNet and VGGNet, termed DeepID3, inspired by these networks. These networks are deeper than GoogLeNet and VGGNet, but not the deepest. The accuracy of this architecture's implementation on the LFW dataset was 99.53%.

The FaceNet DCNN architecture produced by Schroff *et al.,* [15] in 2015 was built on the GoogLeNet architecture. It was trained on more than 200 million training images and eight million distinct face personalities, where it achieved an accuracy of almost 99.63% on LFW. To minimize variation between and among faces, they came up with the idea of creating a model that learns the feature maps directly. The triplet loss function derived from face blobs was used with a triplet mining approach to achieve the greatest possible L2 distance across examples from identical individuals. In 2015 Parkhi *et al.,* [16] proposed VGGFace and devised a method for assembling an extensive dataset with only a small amount of manual annotation. On this dataset, VGGNet was trained using the identical triplet loss strategy proposed by FaceNet and attained a 98.95% accuracy rate for verification on the LFW dataset.

The Dlib is based on the ResNet-34 model. However, *Wang et al*., [17] slightly modified the network to prune some layers for efficiency, where 29 convolutional layers were proposed instead of the original ResNet structure. Dlib is trained on FaceScrub [18], VGGFace, and web-scraped data. Face images are represented as 128- dimensional vectors in this algorithm, which expects inputs of 150x150x3. Dlib achieved an accuracy rate of 99.38% on the LFW dataset. OpenFace [19] employs a variant of the nn4 network used by FaceNet. Their modified nn4.small2 variation uses fewer parameters and is optimized for their smaller dataset, both of which are derived from the GoogLeNet architecture. Since OpenFace makes use of FaceNet's triplet loss, the network gives an embedding on the unit hypersphere, and similarity is represented by the Euclidean distance. For the neural network input, OpenFace does a straightforward 2D affine adjustment to make the eyes and nose seem roughly in the same positions. It uses the combined training data of CASIA-WebFace and FaceScrub, the two largest publicly available labeled face recognition datasets used in academic research. It has a 97.53% accuracy rate on the LFW dataset.

The ResNet34 model is also used to build ArcFace. It takes vector inputs that are 112x112x3 and gives back an embedding with 512 dimensions and the same ArcFace loss function [20]. On the LFW data set, the accuracy of the first study on ArcFace was 99.83%. Golwalkar and Mehendale *et al.,* [21] suggested a system that utilizes the deep metric learning technique and our own FaceMaskNet-21 deep learning network to produce 128-d encodings that assist in the face identification procedure from static images, live video streams, as well as static video files. The ArcFace model has an accuracy of 99.82% on the LFW dataset. The summary of all the seven selected algorithms with their performance on LFW is listed in Table 1. The algorithm selection criterion was based on DeepFace performance on LFW of 97.35%. All other popular selected algorithms have better performance than DeepFace.

Table 1
SOTA FR algorithms used for MFR evaluation

Sl. No.	Reference	Year	Network	LFW (Accuracy %)
1	Taigman *et al., [*13]	2014	DeepFace	97.35%
2	Wang *et al.,* [14]	2015	DeepID	99.52%
3	Schroff *et al.,* [15]	2015	FaceNet	99.63%
4	Parkhi *et al.,* [16]	2015	VGGFace	98.95%
5	Amos *et al.,* [19]	2016	OpenFace	97.53%
6	Wang *et al.,* [17]	2017	Dlib	99.38%
7	Deng *et al., [*20]	2017	ArcFace	99.82%

The proposed evaluation framework for masked faces on synthetic and real datasets is built on the Sefik *et al.,* [22] LightFace platform. The overall SOTA evaluation system pipeline and the evaluation methodology are depicted in Figure 1. After feeding each candidate algorithm every image from the evaluation dataset, embeddings are derived from those algorithms. The distance between any two entities in the list is calculated from respective embeddings using the list of a pair as specified by the evaluation protocol. Cosine distances are used as the metric, and the 2-sigma threshold setting technique typically employed in the FR application is employed here. Subsection 2.2 discusses how the evaluation datasets and protocols are obtained and their specifics.

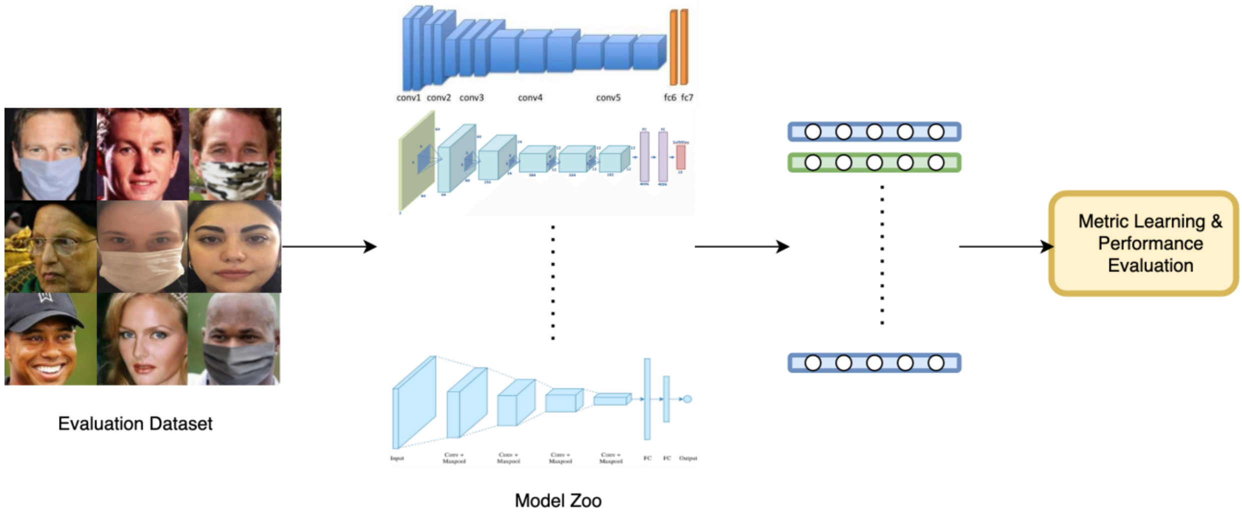


Fig. 1. Overview and methodology for the evaluation framework

2.2. Dataset Formulation

It is best to utilize a standard test dataset when benchmarking algorithms so that researchers may compare the outcomes directly. Research on Masked Face Recognition (MFR) necessitates lots of data from databases for training and testing. As MFR is relatively new, a common approach is to utilize existing face recognition datasets and simulate the faces with synthetically fitted masks. However, the performance of FR on synthetic masked faces cannot be trivially transferred to real-case scenarios. They contain many more variations in different sizes of facial masks, types, and styles (N95 vs. surgical masks). Different colors, textures, and the different ways of mask fitting can be caused by either the change of facial pose or individual preferences.

Therefore, we explore two datasets for evaluation: the curated real masked face dataset MFD-k with its evaluation protocol and the MLFW dataset, which is a synthetic dataset. The difference between the real mask face dataset and the synthetic face dataset is that the real mask face dataset contains facial images with actual masks. At the same time, the synthetic masked face dataset has software-generated masks applied to the face.

The real masked face dataset is derived from an eight-part Kaggle dump of masked face images [11]. It was about 500GB per part, and there was a lot of noise and inconsistency in the naming conventions. It also lacks proper documentation. We curated this dataset by organizing each class in a separate directory. The automated script using the help of file labels is developed to segregate identities. However, due to label inconsistencies, the mix-ups result in large noise. Hence every class was checked by a human, and a clean dataset was obtained. Any class with fewer than four images is discarded. Finally, a clean, curated, usable human-verified real masked face dataset was obtained, named MFD-k.

In the real masked face dataset, MFD-k, each identity is represented by at least four different types of photographs in each setup, as shown in Figure 2. The first image displays a face that has been appropriately masked, followed by two photos with faces with only the mouth and chin masked, respectively, and finally, an unmasked image. This dataset has over 251K images for more than 28,000 unique identities, including those with short hair, glasses, self-occlusion, and racial and gender variances. Most of the participants are between the ages of 25 and 40.

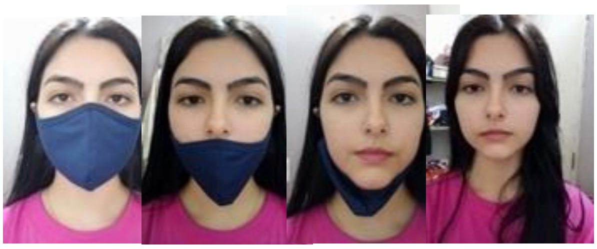


Fig. 2. MFD-k dataset of one identity with four mask settings

The inter and intra-class variation in datasets for light conditions, different types of masks, and background clutter makes this dataset more challenging. Figure 3 shows examples of images in the MFD-k dataset. The curated Kaggle masked face dataset MFD-k can be used for MFR training and testing. The small challenging dataset, MFD-k_test, with its evaluation protocols, is separated from the MFD-k and used to test the MFR algorithm performance on real mask faces. The methodology used for the MFD-k curation process is illustrated in Figure 4.

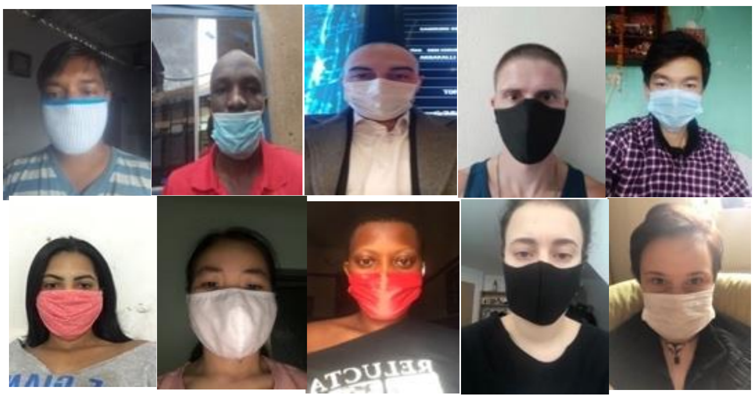


Fig. 3. Examples of images in the MFD-Kaggle dataset

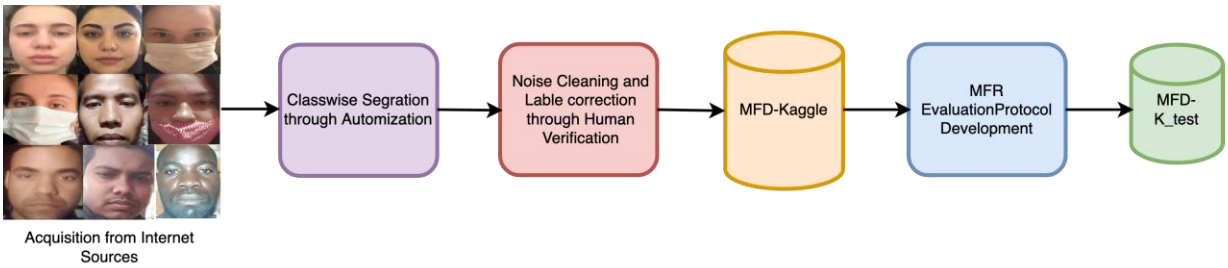

Acquisition from Internet Sources
Classwise Segration through Automization
Noise Cleaning and Lable correction through Human Verification
MFD-Kaggle
MFR EvaluationProtocol Development
MFD-K_test

Fig. 4. Methodology to curate real masked face train and test dataset from Kaggle

The second dataset utilized in this paper for testing the MFR algorithm was a synthetic version of a well-known benchmark dataset LFW called MLFW by Deng *et al.,* [12]. The MLFW dataset has 12000 images and 5749 identities. The masks are synthetically generated using the software on selected images with various types of masks and patterns. There are 6,000 verified protocol pairs, with an equal number of positive and negative examples. Since it is derived from non-masked faces from It is built from the Cross-Age LFW (CALFW) dataset, a comparison between MFR and non-MFR traditional algorithms may be more fruitful on this dataset. The MLFW dataset is not assessed on the state-of-the-art 2D FR models. The sample and different challenges from MLFW datasets are shown in Figure5 The first group of the MLFW dataset is used to determine if the face recognition model can correctly identify two faces, out of which one is wearing a mask. As a result, only one mask is given to each pair of faces. The second testing group is identifying the same person with different face masks. Moreover, the last subgroup comprises negative pairs with the same face masks. The total number of positive pairs equals the number of negative pairs.

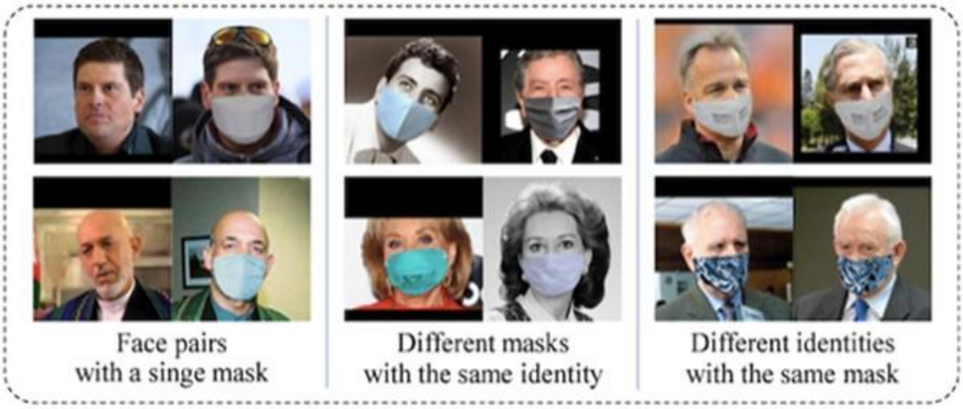

Face pairs with a singe mask
Different masks with the same identity
Different identities with the same mask

Fig. 5. Examples of Masked LFW Dataset [12]

Figure 6 shows the process by which the MaskVGGFace2 is created. We applied five different types of face masks to the VGGFace2 dataset. The type 1 mask is a cloth face mask, type 2 is a medical face mask with two different colors, blue and green, the type 3 mask is the KN-95 mask, and the type 4 mask is the N-95. Then simulated masked face dataset is merged with the original images without face masks. Further, cleaning and preprocessing resulted in a dataset with more than 5.5 million images and over 8000 classes called MaskVGGFace2. The MaskVGGFace2 dataset has a variety of images with various ethnicity, gender, pose, and illumination, illustrated in Figure 7 with different face masks used in this dataset.

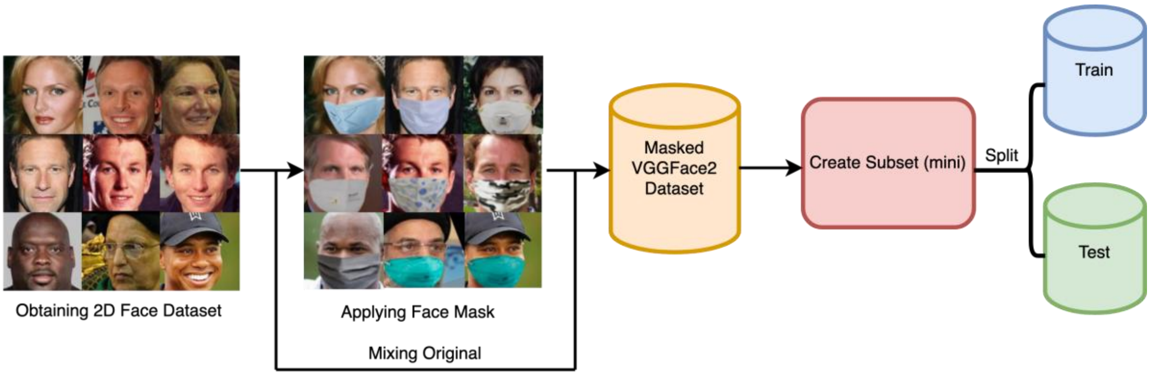


Fig. 6. Methodology to create a synthetic masked face dataset

Using real-world mask datasets continues to be a significant barrier to the MFR system's effectiveness. The availability of data augmentation and face masking technologies is necessary to test MFR algorithms on a range of real-world masks, including textured masks. For this reason, it is necessary to construct a sizable dataset of synthetic faces. It is made on VGGFace2 with a facial mask generator.

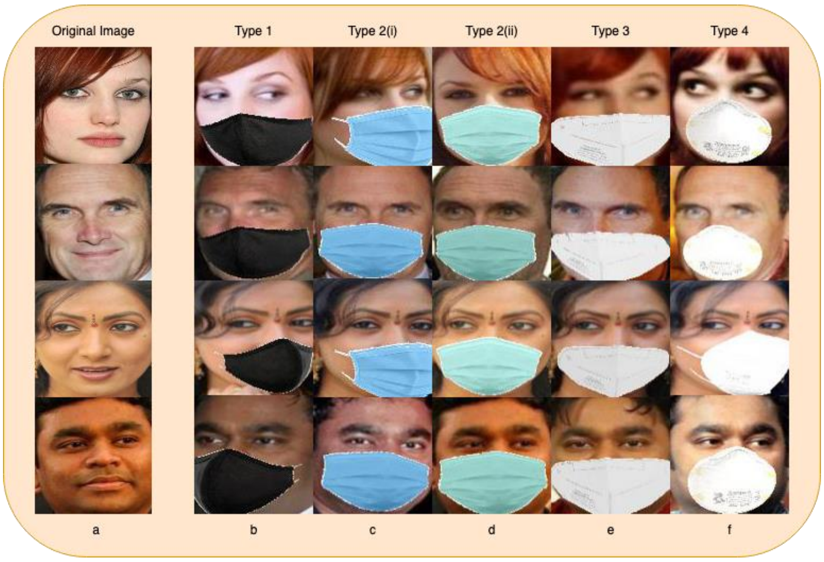


Fig. 7. Different types of face mask images in the MaskVGGFace2 dataset

The MaskVGGFace2-mini dataset is a subset created from the MaskVGGFace2. It has 2500 classes and more than 150K images. The purpose is to have a similar impact size as MFD-k for studying real synthetic mask faces comparable to different MFR algorithms. The MFD-k_test is a test dataset for real masked faces, and MLFW is used for synthetic masked face evaluation. The evaluation protocol for MLFW is described by Deng *et al.,* [13]. The summary of all the datasets formed for the study and their details are given in Table 2.

Table 2
Details of MFR datasets used in this paper

Dataset	Size (#Images)	Identities	Purpose
MFD-k	250,993	29,504	Train (MFR)
MaskVGGFace2	5,163,762	8,631	Train (MFR)
MaskVGGFace2-mini	153,886	2,500	Train- Ablation Study
MFD-k_test	240 (28,680 pairs)	30	Test (MFR)
MLFW	12000	5749	Test (MFR)

The MFD-k_test is a dataset with three protocols being developed to benchmark the robustness of MFR algorithms on real mask face datasets. The three protocols are unmasked-masked (UM) verification and unmasked-unmasked (UU pairs) non-occluded face verification. The third and final protocol is all pair verification with 28,680 total pairs, as in Table 3. The UM protocol tests the performance of faces with masks in a gallery of face images with no face mask. In this protocol, each pair has one image with a face mask and another without a face mask. The UU pair is to benchmark the performance of conventional FR operations with both faces without wearing a face mask. The third protocol of all pairs portrays the strength of MFR in changing scenarios. Table 3 displays the number of pairs employed by each protocol.

Table 3
Details of MFD-K_test evaluation protocol

Protocol	Positive Pairs	Negative Pairs	Total Pairs
Unmasked-Masked	14160	360	14400
Unmasked-Unmasked	7080	60	7140
All-Pair	28320	360	28680

MFR tests are used to evaluate the performance of the SOTA and trained FR models. The MFD-k_test and MLFW datasets are used for evaluation. We evaluate existing SOTA models that use a similar reporting procedure for their best results in the earlier works. That is an image pair can be given to the 1:1 verification task to see if the images show faces that belong to the same person. There are 3000 similar and 3000 non-similar image pairs in the MLFW standard protocol used for 1:1 verification for synthetic face masks. For real masked faces, three different protocols mentioned above are used. We evaluated the strength of SOTA algorithms and proposed MFR on both datasets for benchmarking and recommendations. The face datasets are preprocessed before training and evaluation, as described in section 2.3.

2.3 Data Preprocessing Pipeline

The pipeline for preprocessing the masked face dataset is depicted in Figure 8. It comprises four operational stages: face detection, face alignment, crop, and resize. Face detection in the presence of a face mask is challenging. The face detector receives a masked face image as an input and is responsible for detecting the masked face in the picture. The four different face detector algorithms are tested for the purpose.

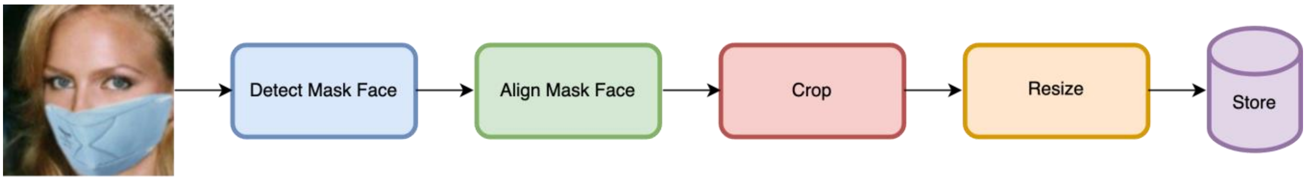


Fig. 8. Masked Face Dataset Pre-processing Pipeline

We have evaluated the MTCNN [23], Dlib [17], FaceBoxes [24], and RetinaFace [25] detectors on the MFD-Kaggle and MLFW datasets. The recognized face is aligned, and after aligning the photos, they are cropped to the desired input size. Accurate cropping is crucial. Otherwise, face recognition performance would deteriorate due to noisy background or partially cropped face region.

Although out of the scope of this paper, it is worth noting that our dataset preprocessing ablative study found that popular face detectors like MTCNN and Dlib did not perform well on the masked face datasets. The FaceBoxes could process all the images of MFD-k, and the RetinaFace could handle over 99% of the MLFW dataset, as mentioned in Table 4.

Table 4
Performance of various Face detectors on masked face datasets

Face Detector	Mean Error
MTCNN	23.333%
Dlib (DNN)	6.670%
RetinaFace	0.200%
FaceBoxes	0.0002%

2.4 Proposed MFR

We trained a ResNet model on the MaskVGGFace2-mini dataset. The model used is ResNet18. ResNet18 is an architecture that consists of 18 deep layers. This network's architecture was designed to make it possible for a significant number of convolutional layers to operate well. On the other hand, the output of a network almost always suffers when additional deep layers of complexity are added to it. Each module contains a total of four convolutional layers, except the one-to-one convolutional layer. There are a total of 18 layers, including the initial convolutional layer, the final fully connected layer, and all levels in between.

Because of this, the model is generally referred to as ResNet-18. Roughly 11 million trainable parameters are available in ResNet18. It is made up of CONV layers that have filters that are 3x3 in size (just like VGGNet). There are just two pooling layers utilized across the entirety of the network; one is located at the beginning of the network, and the other is located after the network. The accuracy performance of the ResNet-18 model on the LFW dataset is 96.4%.

The ResNet-18 is selected as the backbone CNN architecture for the proposed MFR since it is used in most SOTA FR algorithms as a preferred backbone like ArcFace and Dlib. The Hyperparameters for training used are detailed in Table 5:

Table 5
Hyperparameters for the proposed MFR model with ResNet-18 backbone

Hyperparameter	Value
Batch Size	32
Loss Function	Cross Entropy Loss
Epochs	50
Learning Rate	0.001
Momentum	0.5
Optimizer	SGD
Dropout	No

We adopted a transfer learning approach from ImageNet weights, as shown in Figure 9. The modified ResNet-18 backbone MFR model is developed using the Pytorch framework. The MaskVGGFace2-mini dataset with 2500 identities is split into 70% for training, 20% for validation, and 10% for testing. The complete setup of deep learning development and evaluation is carried out on Intel i7 10th Generation 2.9GHZ with 24 GB RAM using 3060Ti RTX 12 GB GPU.

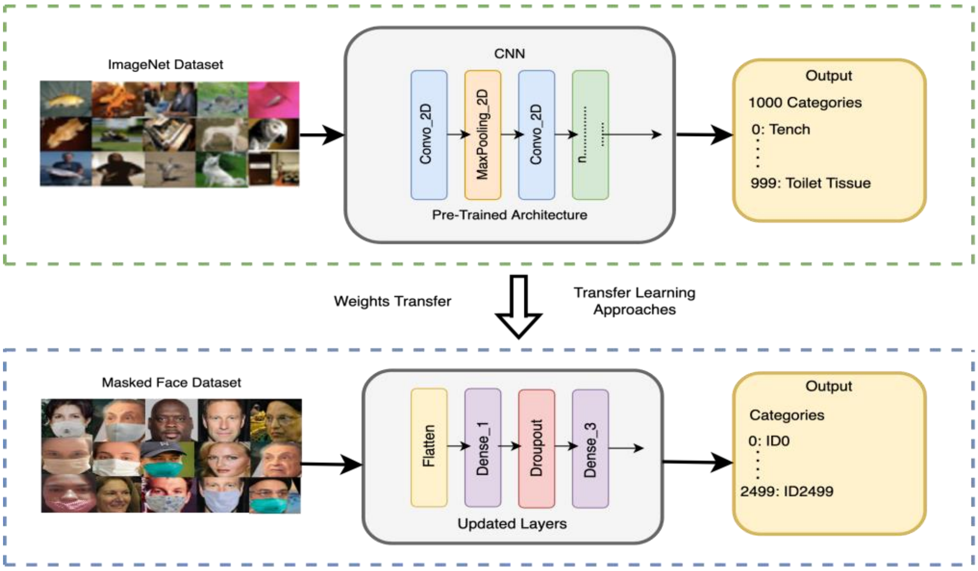


Fig. 9. Proposed MFR model development methodology

3. Results

This section discusses the results of the proposed MFR-trained model and seven SOTA algorithms according to the evaluation method described in section 2. The performance evaluation of SOTA FR with the proposed MFR algorithm on a real masked face dataset is done in section 3.1. and on the synthetic dataset in section 3.2. Section 3.3 details the performance of the proposed MFR training and testing on synthetic MFD. An elaborate discussion of eight FR systems on both datasets is carried out in section 3.4.

3.1 Performance Evaluation on Real Masked Face Dataset (MFD-k_test)

The performance of eight FR algorithms on MFD-k is shown in Figure 10. The plot bars illustrate how the algorithms fared on masked, no occlusion unmasked, and all pair protocols. A subset created from real masked faces from MFD-k_test is used for this examination. The dataset has 60 unique identities, which ensures that the dataset is evenly split between men and women. The unmasked-masked (UM), unmasked-unmasked pairs (UU), and all pair protocols are used as described in the methodology subsection 2.2.

From Figure 10, This first set of bar graphs depicts the proposed MFR with ResNet. A ResNet model was trained using a simulated masked face dataset, MaskVGGFace2-mini. The results show improved performance in masked UM protocol and All-pair protocol. This performance gain is because, unlike conventional face recognition models, which were trained using unmasked images, this one is fed with data containing partially visible masked faces and uncovered faces. However, comparing the UU protocols unmasked-unmasked combination to the best-performing VGGFace and ArcFace models, we also notice that the MFR model performance is degraded on faces not wearing masks. The unmasked-masked pair and the all-pair of the ResNet based proposed MFR model are the best compared to all traditional FR algorithms.

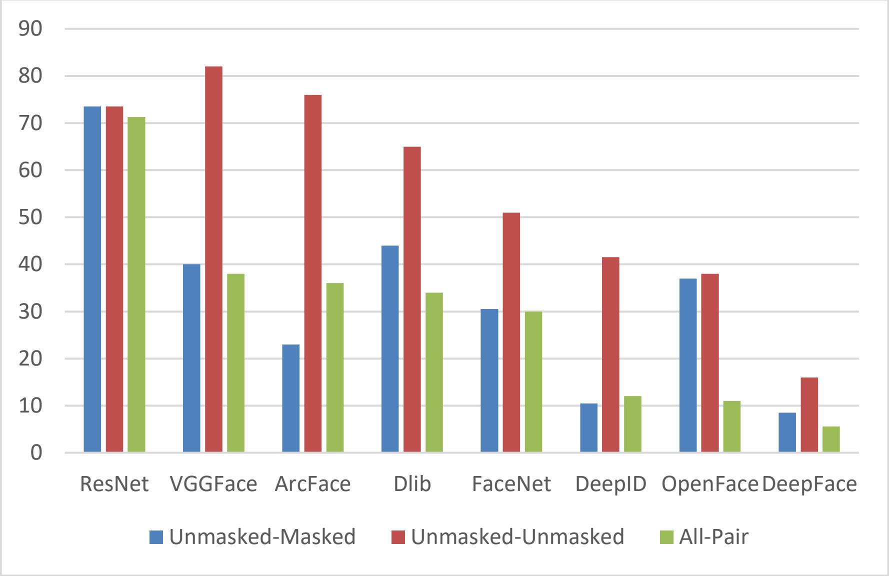


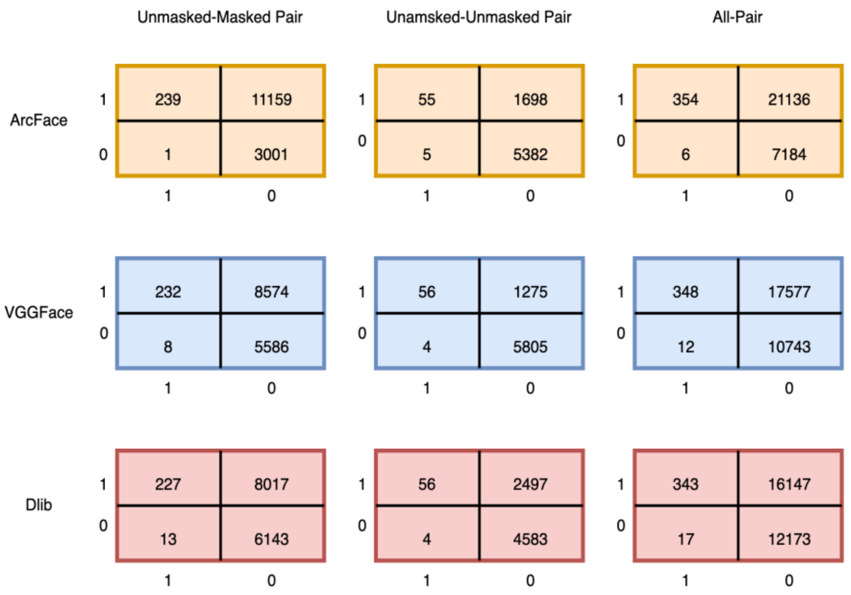
Fig. 10. Face Recognition Performance on MFD-k Dataset

The MFR task severely impacted three FR models, DeepID, OpenFace, and Deep face, where the recorded accuracies are lower than the random classifier performance. DeepFace recorded the most insufficient accuracy for the All-pair protocol. This degraded performance may be because of a failed 3D alignment procedure, which constitutes an important module in the algorithm that attempts to use Delauney triangulation to align the 2D-aligned cropped image, creating a 3-dimensional model from a generic 2D to a 3D model generator. This procedure plots the 67 fiducial points that have failed to owe to occlusion. We hypothesize that this is the primary cause of the failure. The failure can be attributed to large occlusion caused by facial masks. Compared to VGG-Face, FaceNet accuracy suffers due to its smaller output feature vector of 128. At the same time, VGG-Face has an output feature vector of 2622, which is much larger and more suited to distinguishing between faces.

The ArcFace performance is good as it uses a sophisticated deep metric learning method to learn the margin that separates two distinct faces effectively. Therefore, most face mask recognition algorithms use ArcFace due to its ability to generate discriminative embeddings. The UM protocol of unmasked-masked pair performance of Dlib is better than VGG-Face and ArcFace because its face recognition model is based on the highly accurate ResNet as a backbone. At the same time, it was trained on a large dataset with a high degree of variations, including the mixture of VGGFace2, Face Scrub dataset, and custom scrap dataset from the internet.

In addition, we observe that on real mask faces, the overall performance of the face recognition algorithm using UM unmasked-masked 1:1 protocol and all pair evaluation is worse than the random classifier, as shown in Figure 10. These results illustrate the significant ineffectiveness of the FR algorithms when dealing with the heavily occluded faces due to face mask-wearing. This can be compelling in real-world environments, especially for security-sensitive FR applications such as border control and private building access.

The confusion matrix for the three best SOTA algorithms on all three protocols is shown in Figure 11. It is clear that the precision is less than 5%, and the F1-score is within 10% of all algorithms across all protocols. The same metric of a 2-sigma fixed threshold is applied across all the protocols. The percentage of false positives across the algorithm is higher for UM protocol than for the UU protocol. It establishes that the SOTA FR performance degrades badly in the presence of face masks.



		Unmasked-Masked Pair		Unamsked-Unmasked Pair		All-Pair	
		1	0	1	0	1	0
ArcFace	1	239	11159	55	1698	354	21136
	0	1	3001	5	5382	6	7184
VGGFace	1	232	8574	56	1275	348	17577
	0	8	5586	4	5805	12	10743
Dlib	1	227	8017	56	2497	343	16147
	0	13	6143	4	4583	17	12173

Fig. 11. Confusion matrix for ArcFace, VGGFace, and Dlib across all three protocols on MFD-k_test

3.2 Performance Evaluation on Synthetic Masked Face Dataset (MLFW)

The performance of seven 2D FR algorithms on the MLFW dataset can be seen in Figure 12. Mask LFW performance is shown along with the unmasked face LFW dataset to examine the difference in the performance of the FR algorithms before and after applying a simple synthetic mask on faces.

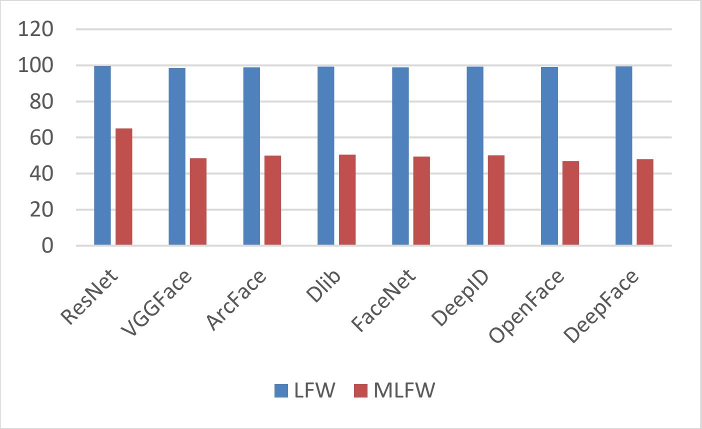


Fig. 12. 2DFace Recognition Performance on LFW and MLFW

As evident from Figure 12, the ResNet model has done better than the random classifier making its performance better than other than the traditional FR algorithms. The traditional 2D FR algorithm performance on the synthetic mask face dataset is less than the random classifier. At the same time, all the 2D face recognition algorithms have performed very well on the LFW dataset but could not handle the synthetic occlusion introduced in the dataset.

3.3 Performance of proposed MFR on MaskVGGface2-mini

Figure 10 shows the performance of the SOTA FR algorithms and proposed MFR with ResNet-18. As we can see, the performance of the SOTA FR algorithms is almost like a random classifier. The results of protocols UM unmasked-masked and the All-pair are very low, less than 50%. With the proposed model trained on ResNet-18, we saw that the results of the UM unmasked-masked and All-pair protocols have very high accuracy in comparison. For UM pair, the accuracy is 73.54%, and All-pair is 71.28%. For UU (unmasked-unmasked) pair, the performance is less than VGGFace, 71.58%, and VGGFace, more than 80%. This performance drop shows us that the model with a mask on the face considers the mask as the part of the face hence the embedding extracted is of poor quality.

The training and test accuracy and loss curves of the proposed MFR are plotted in Figures 13 and 14, respectively. We observed that the training accuracy of the ResNet-18 model is 99%, and the test accuracy is 77%. We can see that the test accuracy is low compared to the training accuracy due to overfitting and poor generalization of the data. The training dataset is sourced from VGGFace2, which has low-quality images and is complex. Due to the complex nature of the residual structure, the training is time-consuming and difficult. Also, the gradient of the residual data is not stable and sometimes leads to unreliable results. However, it shows significant improvement with just 50 epoch training, a small dataset, and a simple loss function compared to the SOTA FR algorithms training with millions of images. The confusion matrix of the proposed MFR is depicted in Figure 15 and can be compared with Figure 11. It clearly shows fewer errors on the ResNet-based proposed MFR.

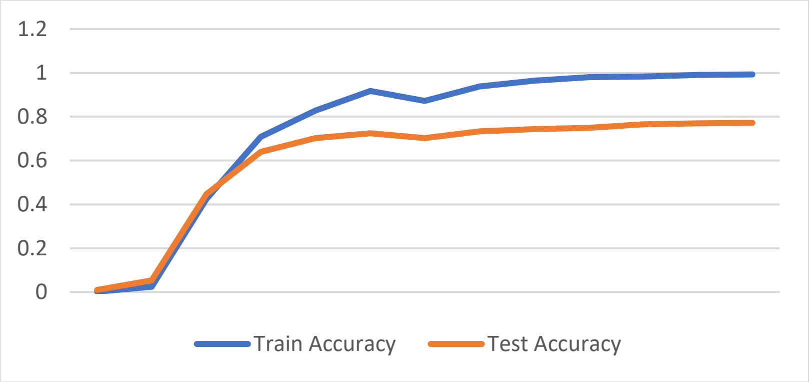


Fig. 13. The training and test accuracy of the proposed MFR model

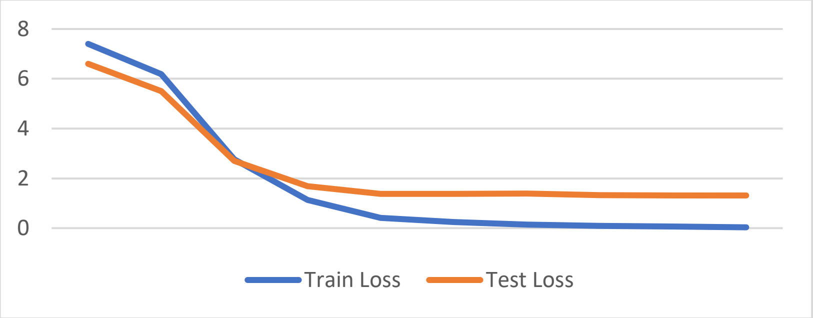
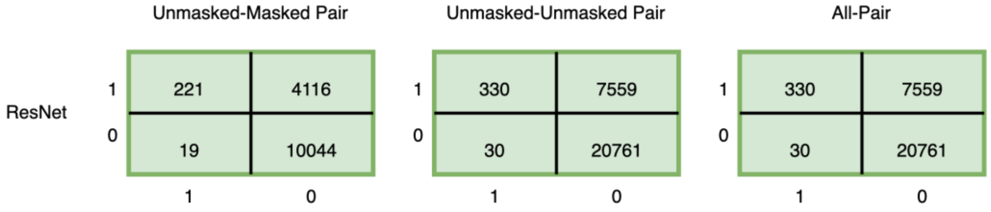


Fig. 14. The loss curves for training and testing on the proposed MFR model



		Unmasked-Masked Pair			Unmasked-Unmasked Pair			All-Pair	
ResNet	1	221	4116	1	330	7559	1	330	7559
	0	19	10044	0	30	20761	0	30	20761
		1	0		1	0		1	0

Fig. 15. The confusion matrix of ResNet-18 based proposed MFR across all protocols

3.4 Discussion

The algorithms not trained on masked images behave as a random classifier when exposed to mask images. When the seven face recognition algorithms are evaluated against synthetic and real datasets, their performance is almost like a random classifier. However, when the models are trained on masked datasets and again evaluated against synthetic and real datasets, the real dataset performance is better than the synthetic dataset. The reason for this is that in the case of a synthetic dataset, the model considers the mask to be a part of the face, whereas, in a real dataset, the model completely distinguishes between a face and a mask. As a result, systems that recognize disguised faces are in high demand.

On the other hand, when models are trained on a masked dataset and then re-evaluated on both a synthetic and a real mask face dataset, the real dataset comes out on top as the mask in the real mask face dataset is very prominent. Hence there is a need to efficiently tackle the problem of masks present on the face and improve their performance up to the performance of the existing face recognition systems. In short, the performance of the proposed MFR model is superior to SOTA algorithms on both real and synthetic masked face datasets. To determine the quality of the FR model, the separability of these algorithms for positive and negative pairs is analyzed, as shown in Figure 16. All distances of a given protocol positive pair fall to the left for low distances and negative pair curves on the right with high distances. Observing the peaks of these two curves for UM and All pair protocols results in poor separability resulting in a less discriminative FR algorithm. However, the separability for uncovered faces, as indicated in Figure 16 (b, e, and h), are more separable and discriminative. They resulted in better performance for non-occluded faces without the mask. However, it fails for masked faces. The proposed MFR is more discriminative and powerful across all protocols, as evident from Figure 17.

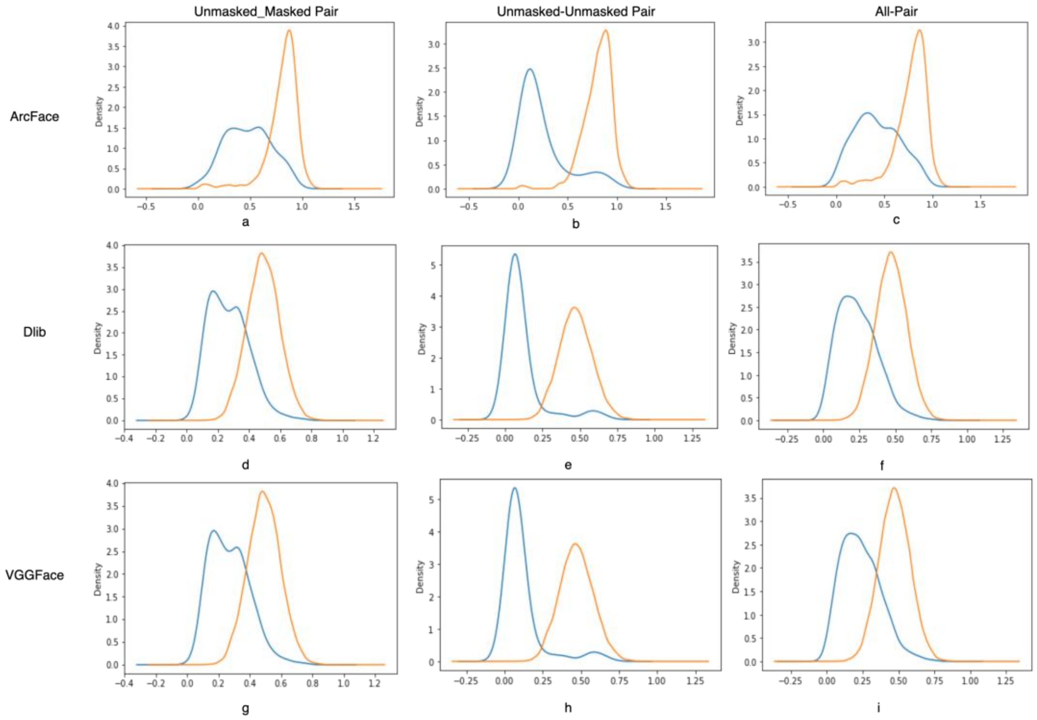


Fig. 16. The separability of positive(genuine) and negative(imposter) pairs for SOTA FR

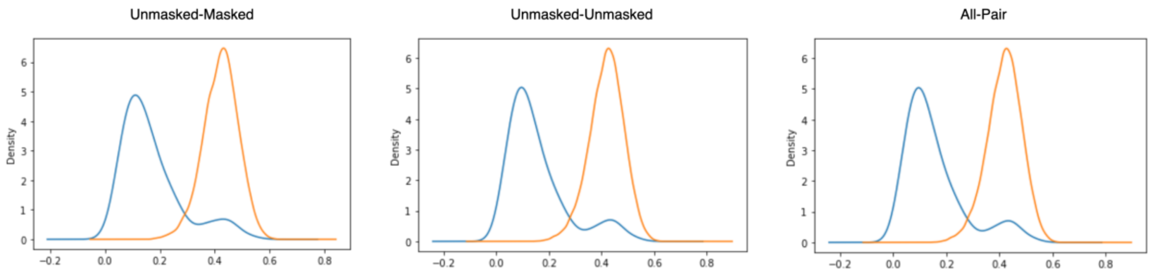


Fig. 17. The separability of positive(genuine) and negative(imposter) pairs for proposed MFR

The proposed MFR accuracy is significantly increased by over 25% for UM pair and All pair protocols in the presence of face masks. However, if keenly observed, the uncovered faces without masks are decreased by at least 4%. To overcome this, we recommended an ensemble approach of non-MFR and proposed an MFR model. Another approach is to introduce a sophisticated loss function with occlusion awareness. The performance of the proposed MFR on MLFW is not very promising since the type of face mask variation in MaskVGGFace2 lacks variations. Groups two and three of evaluation pairs used in MLFW protocols are very challenging for RGB-based 2D CNN models. However, we recommend using a mix of real and synthetic masked large dataset training can be vital to overcome this issue.

4. Conclusions

Considering the recent pandemic, people worldwide have begun donning face masks. Most face recognition systems are built to recognize the uncovered face and are trained on images where the subject is not obscured. Our research analyzed the effectiveness of previously published CNN architectures for generic face recognition in changing scenarios. Compared to the seven SOTA methods, the performance of the ResNet-based MFR model trained on the masked face dataset is noticeably better for both masked-umasked and all-pair. Analyzing the performance of various face recognition models on the MLFW dataset, we discovered that the masked face identification accuracy is as good as a random classifier and is approximately 50% of the standard accuracy of the 2D face recognition methods, with ResNet coming out on top.

The model's performance on the mask face dataset, where it was trained, is affected in different ways by synthetic and real masks. In the case of a synthetic dataset, the mask may be mistaken for real skin because the simulated masks are so close in appearance. In contrast to how well it performed during evaluation, the model on the real dataset can reliably tell the difference between a face and a mask, resulting in better overall performance. As a result of the performance degradation induced by both real and synthetic masks, we conclude that there is an opportunity for improvement in face recognition models.

Acknowledgment

The authors thank NVIDIA for donating the GPU used for this research.

References

[1] Jeevan, Govind, Geevar C. Zacharias, Madhu S. Nair, and Jeny Rajan. "An empirical study of the impact of masks on face recognition." *Pattern Recognition* 122 (2022): 108308. https://doi.org/10.1016/j.patcog.2021.108308

[2] Ngan, M. L., P. J. Grother, and K. K. Hanaoka. "Ongoing face recognition vendor test (FRVT) Part 6A: Face recognition accuracy with masks using pre-COVID-19 algorithms. 10.6028/NIST." (2020). https://doi.org/10.6028/NIST.IR.8331

[3] Wang, Qiangchang, and Guodong Guo. "DSA-Face: Diverse and sparse attentions for face recognition robust to pose variation and occlusion." *IEEE Transactions on Information Forensics and Security* 16 (2021): 4534-4543. https://doi.org/10.1109/TIFS.2021.3109463

[4] Zhao, Fang, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. "Robust LSTM-autoencoders for face de-occlusion in the wild." *IEEE Transactions on Image Processing* 27, no. 2 (2017): 778-790. https://doi.org/10.1109/TIP.2017.2771408

[5] King, Davis E. "Dlib-ml: A machine learning toolkit." *The Journal of Machine Learning Research* 10 (2009): 1755-1758.

[6] Duan, Qingyan, and Lei Zhang. "BoostGAN for occlusive profile face frontalization and recognition." *arXiv preprint arXiv:1902.09782* (2019).

[7] Ding, Feifei, Peixi Peng, Yangru Huang, Mengyue Geng, and Yonghong Tian. "Masked face recognition with latent part detection." In *Proceedings of the 28th ACM international Conference on multimedia*, pp. 2281-2289. 2020. https://doi.org/10.1145/3394171.3413731

[8] Geng, Mengyue, Peixi Peng, Yangru Huang, and Yonghong Tian. "Masked face recognition with generative data augmentation and domain constrained ranking." In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2246-2254. 2020. https://doi.org/10.1145/3394171.3413723

[9] Li, Chenyu, Shiming Ge, Daichi Zhang, and Jia Li. "Look through masks: Towards masked face recognition with de-occlusion distillation." In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 3016-3024. 2020. https://doi.org/10.1145/3394171.3413960

[10] Huang, Gary B., Marwan Mattar, Tamara Berg, and Eric Learned-Miller. "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments." In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. 2008.

[11] Roman, Kucev. “500 GB of Images for Face Mask Detection. Part 1.” 500 GB of images for Face Mask Detection. Part 1. Kaggle, June 14, 2021. https://www.kaggle.com/datasets/tapakah68/medical-masks-part1.

[12] Zhu, Zheng, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu et al. "Masked face recognition challenge: The webface260m track report." *arXiv preprint arXiv:2108.07189* (2021).

[13] Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. "Deepface: Closing the gap to human-level performance in face verification." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701-1708. 2014. https://doi.org/10.1109/CVPR.2014.220

[14] Ouyang, Wanli, Xiaogang Wang, Xingyu Zeng, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li et al. "Deepid-net: Deformable deep convolutional neural networks for object detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2403-2412. 2015. https://doi.org/10.1109/CVPR.2015.7298854

[15] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. "Facenet: A unified embedding for face recognition and clustering." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815-823. 2015. https://doi.org/10.1109/CVPR.2015.7298682

[16] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. "Deep face recognition." (2015). https://doi.org/10.5244/C.29.41

[17 King, Davis E. "Dlib-ml: A machine learning toolkit." *The Journal of Machine Learning Research* 10 (2009): 1755-1758.

[18] Ng, Hong-Wei, and Stefan Winkler. "A data-driven approach to cleaning large face datasets." In *2014 IEEE international conference on image processing (ICIP)*, pp. 343-347. IEEE, 2014. https://doi.org/10.1109/ICIP.2014.7025068

[19] Amos, Brandon, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. "Openface: A general-purpose face recognition library with mobile applications." *CMU School of Computer Science* 6, no. 2 (2016): 20.

[20] Wang, Kai, Xiaojiang Peng, Jianfei Yang, Debin Meng, and Yu Qiao. "Region attention networks for pose and occlusion robust facial expression recognition." *IEEE Transactions on Image Processing* 29 (2020): 4057-4069. https://doi.org/10.1109/TIP.2019.2956143

[21] Duan, Qingyan, and Lei Zhang. "Look more into occlusion: Realistic face frontalization and recognition with boostgan." *IEEE transactions on neural networks and learning systems* 32, no. 1 (2020): 214-228. https://doi.org/10.1109/TNNLS.2020.2978127

[22] Serengil, Sefik Ilkin, and Alper Ozpinar. "Lightface: A hybrid deep face recognition framework." In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1-5. IEEE, 2020. https://doi.org/10.1109/ASYU50717.2020.9259802

[23] . Zhang, Kaipeng, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. "Joint face detection and alignment using multitask cascaded convolutional networks." *IEEE signal processing letters* 23, no. 10 (2016): 1499-1503. https://doi.org/10.1109/LSP.2016.2603342

[24] Zhang, Shifeng, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z. Li. "Faceboxes: A CPU real-time face detector with high accuracy." In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1-9. IEEE, 2017. https://doi.org/10.1109/BTAS.2017.8272675

[25 Chavda, Amit, Jason Dsouza, Sumeet Badgujar, and Ankit Damani. "Multi-stage CNN architecture for face mask detection." In *2021 6th International Conference for Convergence in Technology (i2ct)*, pp. 1-8. IEEE, 2021. https://doi.org/10.1109/I2CT51068.2021.9418207

[26] B. Hayes. NIST Launches Studies into Masks' Effect on Face Recognition Software. The National Institute of Standards and Technology (NIST), August 4, 2020. https://www.nist.gov/news-events/news/2020/07/nist-launches-studies-masks-effect-face-recognition-software.