A Benchmark on Masked Face Recognition

Pedro Vidal*, Roger Leitzke Granada†,
Gustavo Führ†, Vanessa Testoni†, David Menotti*
*Department of Informatics, Federal University of Paraná, Curitiba, PR, Brazil {pbqv20, menotti}@inf.ufpr.br
†unico - idTech, Brazil {roger.granada, gustavo.fuhr, vanessa.testoni}@unico.io

***Abstract*—With the COVID-19 pandemic's emergency, using facial masks and contactless biometric systems became even more relevant to reduce the risk of contamination. Several direct and indirect problems gained relevance with the pandemic. Among them, masked face recognition (MFR) aims to recognize a person even when the person is wearing a face mask. Some state-of-the-art algorithms that work well for unmasked faces have suffered a severe performance drop when receiving masked faces as input. In this sense, the scientific community proposed approaches and competitions related to this topic. In this paper, we introduce a comparative study of four prominent solutions pipelines that use different techniques to tackle the masked face recognition problem, proposed by Huber et al. [1], Neto et al. [2], Boutros et al. [3], and Hsu et al. [4]. The performance evaluation was conducted on a real masked face database (MFR2 [5]), and using synthetic masks in three mainstream databases (LFW, AgeDB-30, and CFP-FP). We report results regarding unmasked-masked (U-M) and masked-masked (M-M) face verification performance. The unmasked-unmasked (U-U) scenario was also reported as a baseline to evaluate the drop of the selected models on non-occluded face verification. We further analyze the obtained results, generating a comprehensive comparative study of the selected approaches.**

I. Introduction

Face recognition is one of the core topics in the field of computer vision. Due to its non-intrusive nature and high discriminability for identity authentication, face recognition has been a prevalent biometric technique, and it has been used in a wide range of applications like video surveillance, security, and access control [6], [7].

The onset of the COVID-19 pandemic and the proliferated use of protective masks resulted in an indirect challenge to the facial recognition pipelines since they were designed to discriminate identities using full face information (i.e., non-occluded faces). When such systems receive faces occluded by masks as inputs, which hides valuable biometric cues, their overall performances drop significantly. Moreover, the use of facial recognition has become increasingly important because it fits in the scope of contactless biometric systems and can help avoid the virus spread (w.r.t fingerprint authentication, for instance).

This context motivates the research on masked face recognition, as the challenge of masked faces for automatic face recognition (FR) models was stated by studies conducted by the National Institute of Standards and Technology (NIST) [8], Department of Homeland Security [9], and the scientific community [10], [11]. These studies confirmed that wearing masks significantly negatively affects the accuracy of FR systems.

After a new volume of research [12], NIST promotes a second comparative study [13] involving post-pandemic algorithms, thus potentially designed with covered faces in mind, and concludes the performance of face recognition with face masks is comparable to the state of the art on unmasked images in mid-2017. In this sense, the community has developed new studies aiming to solve the masked face recognition (MFR) problem, resulting in different methods [1]–[4], [14], competitions, and case study papers [12], [15]–[17].

With the masked face recognition problem in mind, this work presents a comparative study of four prominent methods, namely by Huber et al. [1], Neto et al. [2], Boutros et al. [3] and Hsu et al. [4], that use different techniques to tackle this problem. We evaluate the models's performance on the face verification task using a real masked database, named Masked Faces in Real-World for Face Recognition (MFR2) [5]. and three mainstream databases (i.e., LFW [18], AgeDB-30 [19], and CFP-FP [20]) augmented with synthetic masks.

We report results regarding unmasked-masked (U-M) and masked-masked (M-M) face verification performance. In (U-M) scenery, a single face is covered by the mask, while both faces are covered by masks in the (M-M) scenery. We also report the unmasked-unmasked (U-U) scenery to evaluate the drop of the selected models on common face verification compared to a state-of-the-art non-masked face recognition pipeline. We aim with these experiments to present a vivid benchmark of the area, contrasting approaches that are not explicitly compared in the literature and evidencing the diverse techniques used to approach this problem. To the best of our knowledge, there is no such benchmark in the literature.

The remainder of this paper is structured as follows. Section II presents related works. Section III describes the methodology. Section IV shows results and a discussion about them. Section V concludes and points out some future work directions.

II. Related works

In recent years, research on face recognition has continued pushing the state of the art on non-occluded face recognition [21]–[24]. Motivated by the negative effect of the protective masks on facial recognition performance [8]–[10], [25], solutions to this problem were proposed recently by several works. Some works focused on presenting solutions to detect the presence of a mask on a face [26], [27] while not addressing the recognition of an occluded face. Nizam Ud Din et al. [28] presented a GAN-based solution to unmask a

face, which is achieved by using a pipeline with two stages, where initially, a model produces a binary segmentation for the masked region. Then the second stage removes the mask and synthesizes the affected region with fine details while retaining the global coherency of face structure. Anwar and Raychowdhury [5] finetuned a model on synthetically masked images using the Inception-ResNet v1 [29] with the triplet-loss FaceNet [30]. In that work, the authors proposed an open-source tool, MaskTheFace, to generate synthetic images, so that an effective masked face recognition system was trained on a large dataset of masked faces. A new, but small, collected from the web, ready-to-use real-world aligned masked face dataset of identities, proposed by the same authors, MFR2 was used to evaluate the model trained with synthetic images.

Studies of comparative nature were also published, highlighted by the work proposed by Jeevan et al. [31], which conducts a series of experiments by testing existing CNN architectures available in the literature and reports possible changing in loss functions, architectures, and training methods to enhance the masked face recognition performance. Also, two major competitions related to MFR were conducted: the IJCB-MFR-2021 [32] and the ICCV21-MFR [33]. The former evaluated all models using a private dataset containing masked faces, named MFR2. As described in the competition paper, the most competitive submissions are based on ResNet [34] architectures and the ArcFace [21] loss as a foundation, using synthetic data in the training process. The latter proposed two main tracks: a track for models trained on MS-Celeb-1M [35] and Glint360K [36] datasets and another track for models trained on WebFace260M [37]. The objective of having two tracks is to conduct a fair comparison. Models were evaluated on a private large-scale real masked faces dataset.

Following the competition's trends, recently published methods use ResNet as the feature extraction architecture and the MS1MV2 [38] dataset as the base training set. Huber et al. [1] propose a different approach, which employs Knowledge Distillation (KD) to produce similar embeddings for masked and unmasked faces. In their approach, the pre-trained teacher and student networks receive the same image, but the student network has a probability of 0.5 to add a synthetic mask on the face. The difference between the teacher and the student network embeddings is added to the total loss function inducing the embeddings to be similar even when a face is wearing a mask. To generate similar embeddings for masked and unmasked faces, Boutros et al. [3] propose an Embedding Unmasking Model (EUM) that operates on top of facial recognition systems, thus not requiring the existent facial recognition to be retrained with to deal with masked faces. They use a loss function called Self-restrained Triplet Loss (SRT) to minimize the distance between pairs of unmasked and masked face embeddings of the same person and to maximize the distance between pairs of masked face embeddings of two different people.

Neto et al. [2] propose a multi-task network that consists of one component for facial recognition, and another for mask face detection. They trained the model using a contrastive learning technique, with the MS1MV2 dataset augmented with synthetic masked faces. Both the aforementioned approaches achieve competitive results when compared to the solutions submitted for the IJCB-MFR-2021 competition. Hsu et al. [4] train a RestNet-100 with augmented masked synthetic images and conducts experiments evaluating different loss functions for tackling the MFR problem. The trained models were evaluated using different datasets that reproduce different face recognition challenges, e.g., changes in pose, illumination and expression, cross-age and low resolution images.

In this work, we selected the last four mentioned approaches to compose our comparative study. Hence, Section III describes in more detail each method that delineates our benchmark.

III. Methodology

In this section we describe in more detail the methods in the set of experiments carried out, present the datasets used and the method chosen to add synthetic masks to the images. We also describe the metrics and protocols used in the evaluation.

A. Selected approaches

In [1], the proposed approach employs the Knowledge Distillation (KD) technique using a training paradigm aiming to produce *embeddings* of masked faces that are similar to those of unmasked faces for a given subject. The proposed solution uses a pre-trained face recognition model as the teacher model combined with a state-of-the-art loss [22]. The proposed models are trained with images of the same subject wearing, and not wearing, a mask, such that the resulting model can handle with both situations, while the KD process ensures that the yielded embeddings of masked face images are similar to unmasked face ones of the same subject. In that work, two training strategies were investigated: Mask-invariant Low Guidance (MaskInv-LG) and Mask-invariant High Guidance (MaskInv-HG). In the former strategy, the weight that parameterize the importance between the face embedding of the teacher and the student network remains constant in the training, while in the latter, the weight is increased in the final stages of the training, aiming to emphasize the adaption of the network to the masked data. Moreover, that work also evaluated a model without the use of KD, named ElasticFace-Arc-Aug, where the student network is trained independently using the ElasticFaceLoss [22] on faces augmented with a synthetic mask using the probability 0.5. From the experimental results, the authors concluded the Mask-inv HG model yields the best results in the used benchmarks.

In [3], a new loss function and a new face recognition model are proposed, named Self-restrained Triplet (SRT) and Embedding Unmasking Model (EUM), respectively. The EUM architecture consists of a Fully Convolutional Neural Network (FCNN) operating on top of face recognition models, receiving as input embeddings of a masked face (generated by a ResNet-100 [39]). Its output is a new embedding, which is similar to the one generated by the Convolutional Neural

Network (CNN) if it received the same unmasked face as input. For such aim, the FCNN is trained with the SRT, which is similar to a triplet loss [40]. The main goal of SRT is to decrease the distance between genuine pairs (same person with and without mask) while keeping the distance between imposters (two different people). This goal is supported by results pointed by recent studies [8], [11] that show that embeddings extracted from genuine pairs without masks have their distance increased (at some extend) when one or both faces are masked, while the same result is not observed to imposter pairs, where their distance are kept.

In [2], a contrastive learning called FocusFace is employed in a multitask way to train a masked face recognition (MFR) model, which also detects a mask on the face. Generally speaking, the training process consists of feeding masked and unmasked face images to a CNN (ResNet-100 [39]), which computes the embeddings for each input image. The last layer of the original network was replaced by two parallel fully-connected layers with different lengths. The smaller one is trained with cross-entropy loss for mask detection, while the larger one is trained using the ArcFace loss [21] for face recognition. To complete the loss formulation, a third component is also included as the mean squared error between the two embeddings.

In [4], a Resnet-100 was used as the feature extractor and trained separately with five different loss functions, named Center Loss, the Marginal Loss, the Angular Softmax Loss, the Large Margin Cosine Loss and the Additive Angular Margin Loss (ArcFace). The base training set used the MS1MV2 dataset augmented with synthetic masked images, for all losses. They further benchmark the trained pipelines, using facial recognition datasets that address three different challenges: variations on pose, illumination and expression (PIE), cross-age and low resolution images, by using the IJB-C-IJB-B, FG-Net and SCface for each purpose respectively. The network trained with the ArcFace loss, hereafter called MaskInvArcface, achieved better results in all benchmarks. Hence, they further tested this network for recognizing faces on a real masked face dataset, demonstrating the effectiveness of the approach. Due to the highest results achieved by ArcFace, we select this model for our study.

B. Datasets for the benchmark & method to add synthetic masks to faces

To compare the approaches on masked data we used the LFW [18], AgeDB-30 [19], and CFP-FP [20] datasets, which are mainstream publicly available datasets that cover different challenges for face recognition. The LFW is an in-the-wild unconstrained face recognition dataset, that contains 13,233 images of 5749 different individuals. We follow the original protocol, which contains 3000 genuine pairs and 3000 impostor pairs. AgeDB is focused on comparison of images across age, and the most challenging scenario was used, named AgeDB-30, which has a gap of 30 years between the face images of the individuals, and contains 3000 genuine pairs and 3000 impostor pairs. On the other hand, the CFP-FP dataset contains images to evaluate differences in face pose (i.e., frontal and profile faces). The protocol contains 3500 genuine pairs and 3500 impostor pairs.

Fig. 1. Samples of unmasked reference and masked probe with synthetic masks added on (columns 1, 2, and 3) and a real mask (column 4). Columns 1 and 3 are impostor pairs from the AgeDB-30 [19] and LFW [18] respectively, column 2 and 4 are a genuine pairs from the CFP [20] MFR2 [5].

Given that these datasets do not have pairs of unmasked-masked individuals, we generate the masked versions of the images, using the publicly available method of the JDAI-CV toolkit [41], which was used with the same purpose on the ICCV21-MFR competition [33]. In order to generate the masked face, a 3D face is reconstructed on the input 2D face image. The UV texture map, the face geometry and the camera pose are obtained, and a facial mask is projected into the UV space. Due to the failure to detect landmark, the method was unable to generate masked images from a few images of the CFP dataset, thus the pairs related to those images were excluded, resulting in a reduced number of 3347 genuine pairs and 3353 impostors pairs to be compared. Meanwhile, for the AgeDB-30 and LFW the method generates well all the masked images, not affecting the original protocol.

Due to the lack of publicly available real large scale masked face datasets [31], we evaluate the models on a small scale real masked dataset, by using the MFR2, that consists of masked images collected from the web of 53 identities with a total of 269 images. We use the provided list of 848 pairs of images to be compared. Some of the pairs are shown in Fig. 1. With this dataset we can analyze the face verification performance of the models when exposed to real images, since all the evaluated models were originally trained with synthetic images.

C. Metrics and protocols used for evaluating the models

To conduct the evaluation we collect the publicly available codes from Huber et al. [1][1], Boutros et al. [3][2], Neto et al. [2][3], and Hsu et al. [4][4]. The images of all the used datasets were aligned and cropped using the Multi-task Cascaded Convolutional Networks (MTCNN) [29] detector. After this preprocessing step, all the dataset images were inputted into the four models resulting in four sets of 512-D feature vectors. In the verification phase, to decide if a pair of images is an impostor or a genuine, the cosine similarity scores between the pairs of images are taken.

[1]https://github.com/fdbtrs/Masked-Face-Recognition-KD
[2]https://github.com/fdbtrs/Self-restrained-Triplet-Loss
[3]https://github.com/NetoPedro/FocusFace
[4]https://github.com/AvLab-CV/Face_Mask_Generator

TABLE I

RESULTS OF THE UNIFIED BENCHMARK RELATED TO THE FOUR SELECTED MFR PIPELINES FOLLOWING THE AGEDB-30 DATASET PROTOCOL, IN TERMS OF UNMASKED-UNMASKED (U-U), UNMASKED-MASKED (U-M), MASKED-MASKED (M-M) FACE VERIFICATION PERFORMANCE.

No masks	EER	ZeroFMR	FRM1000	FMR100	FMR10
Arcface [21]	03.7	15.5	09.1	05.9	02.4
MaskInvArcface [4]	04.1	28.6	26.2	08.4	02.7
MaskInv-HG [1]	02.7	15.2	08.7	03.7	01.7
FocusFace [2]	05.1	32.1	25.0	12.9	03.3
EUM [3]	06.2	41.2	32.7	18.4	07.1
Mask vs. No-Mask					
Arcface [21]	13.6	86.8	67.8	41.8	16.5
MaskInvArcface [4]	10.7	62.6	58.8	30.1	10.9
MaskInv-HG [1]	04.7	26.1	23.4	09.6	03.1
FocusFace [2]	10.2	65.6	53.6	33.5	10.4
EUM [3]	18.0	92.8	77.8	50.7	32.7
Mask vs. Mask					
Arcface [21]	16.5	87.9	78.9	56.0	24.5
MaskInvArcface [4]	14.1	82.3	64.5	46.0	17.6
MaskInv-HG [1]	06.1	31.8	28.4	16.5	04.5
FocusFace [2]	11.9	69.0	58.3	41.6	13.4
EUM [3]	20.5	93.9	82.9	62.0	32.5

TABLE II

RESULTS OF THE UNIFIED BENCHMARK RELATED TO THE FOUR SELECTED MFR PIPELINES FOLLOWING THE CFP DATASET PROTOCOL, IN TERMS OF UNMASKED-UNMASKED (U-U), UNMASKED-MASKED (U-M), MASKED-MASKED (MM) FACE VERIFICATION PERFORMANCE.

No masks	EER	ZeroFMR	FRM1000	FMR100	FMR10
Arcface [21]	04.0	14.6	13.0	07.4	02.5
MaskInvArcface [4]	04.7	25.1	15.9	08.7	03.4
MaskInv-HG [1]	02.9	20.1	08.6	04.4	01.9
FocusFace [2]	13.7	66.5	46.8	30.3	15.4
EUM [3]	07.7	56.1	32.9	18.9	06.5
Mask vs. No-Mask					
Arcface [21]	18.7	77.0	64.0	50.6	26.9
MaskInvArcface [4]	14.9	66.2	50.1	37.1	17.7
MaskInv-HG [1]	10.3	48.4	39.3	20.1	10.4
FocusFace [2]	30.2	87.5	79.2	61.3	40.6
EUM [3]	26.8	96.6	91.0	74.0	45.2
Mask vs. Mask					
Arcface [21]	18.2	95.9	79.4	54.6	25.0
MaskInvArcface [4]	15.7	72.5	65.8	44.3	19.3
MaskInv-HG [1]	11.0	43.1	38.3	24.2	11.6
FocusFace [2]	26.7	91.9	88.6	60.9	37.3
EUM [3]	24.4	98.4	88.2	70.1	40.2

The results of the benchmark are reported in terms of the Equal Error Rate (EER), which determine the common value of the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) of a biometric system. We also report the value of False Non-Match Rate (FNMR) at three operation points: FMR1000, FMR100 and FMR10, which refer to the points which provide the lowest FNMR for a False Match Rate (FMR) $<0.1\%$, $<1.0\%$, $<10.0\%$ respectively.

IV. RESULTS

In this section, we present and discuss the results of the models considered in the experiments. Tables I, II, III, and IV summarize the face verification performance of the selected models in terms of the metrics described in Section III, in the scenarios where the reference image is unmasked and the probe image is masked (U-M) and where both images are masked (M-M), following the AgeDB-30, CFP, LFW protocol adopted by ArcFace [21] and the MFR2 original protocol [5].

We also report the case where the pair of images in the comparison are both unmasked (U-U), to evaluate how the training procedure with augmented masks affects the performance of common face verification. This evaluation is relevant since the system trained to recognize masked faces should be resilient to the absence of masks as well, being capable of using the additional information a maskless face presents [31].

On the AgeDB-30, as shown in Table I, MaskInv-HG performs better on all aspects, having an FMR1000 of 23.4 and 08.7 in the (U-M) and (U-U) scenarios, respectively, presenting even better results in comparison to the non-masked face recognition pipeline (Arcface) in the non-occluded scenario (U-U), which achieves a higher FMR1000 of 09.1. This performance gain can be correlated to the use of Elastic Face Loss, that advanced the SOTA (such as ArcFace [21] and MagFace [24]) on six challenging mainstream benchmarks in the training phase of the MaskInv-HG solution. In the (M-M) scenario, the MaskInv-HG solution achieves an FMR1000 of 28.4, resulting in the worst scenario for the model. As described by the authors of the model, "this is the case as the masked vs. masked setting benefits less from the main goal of the MaskInv solution, which is to create face representations similar between masked and unmasked faces, while the masked vs. masked setting require only the similarity between masked faces" [1]. However, the method significantly outperforms the ArcFace model by a large margin, which achieves a higher FMR1000 of 78.9.

The models FocusFace [2] and MaskInvArcFace [4] achieve better results when compared to the Arcface model in the (U-M) and (M-M) scenarios, corroborating that the training procedure boosts the performance in these scenarios. When compared to MaskInvArcface, in the (U-M) scenario, FocusFace presents better performance on the FMR1000 metric (53.6 vs. 58.8) but has a higher FMR100 (33.5 vs. 30.1). In the (M-M) scenario, FocusFace performs better on all the metrics. In contrast, in the (U-U) scenario, FocusFace and MaskInvArcface have a significantly higher FMR1000 of 25.0 and 26.2, respectively, when compared with ArcFace, which achieves 09.1, revealing the trade-off between higher accuracy in masked scenarios but less accuracy in the non-occluded one. In contrast, the MaskInv-HG performs better in the two masked scenarios than the FocusFace and MaskInvArcface and without sacrificing performance in the non-occluded scenario. The EUM models do not enhance the performance of masked faces on the evaluation set.

For the CFP dataset (Table II), the MaskInv-HG consistently achieved better results in all scenarios, including the (U-U) scenario, but having the gap related to the MaskInvArcface in the (U-M) and (M-M) scenarios reduced, when compared to the AgeDB-30 results. This is demonstrated by the FRM1000 of 58.8 vs. 23.4 achieved by the MaskInvArcface and MaskInv-HG, respectively, on the AgeDB-30 with un-

masked reference and masked probes, and the FMR100 of 50.1 vs. 39.3 achieved on the CFP with unmasked reference and masked probes by the same methods respectively.

When compared to the SOTA non occluded pipeline (Arcface), the models MaskInv-HG and MaskInvArcface perform better in the masked scenarios, presenting a better FMR1000 of 50.1 - 39.2 and 65.8 - 38.3 respectively, in the (U-M) and (M-M) scenarios, when compared to the FRM1000 of 64.0, 79.4 achieved by the Arcface model in the aforementioned scenarios.

As observed by Hsu et al. [4], the Arcface and MaskInvArcface performance on cross-age is the lowest among the other factors, such as pose, illumination and expression. This can be observed on the higher FMR presented in all masked scenarios of the AgeDB-30 dataset, in comparison to the CFP and LFW masked results. On the other hand, despite having a lower FMR1000 on CFP (U-U) scenario in relation to the AgeDB-30, the MaskInv-HG solution archive higher FMR1000 values on the two masked protocols (U-M) and (M-M), FRM100 of 39.3 and 38.0, when compared to the FMR1000 of 23.4, 28.4 on the AgeDB-30, showing that a cross-pose facial verification can be significantly degraded on masked faces, especially for this solution. For the FocusFace and EUM approaches, this degradation in performance is more pronounced by observing that the solution's FMR on all operation points is higher than the Arcface, showing that the training paradigms do not enhance the masked face verification performance of this dataset.

We can verify similar behavior to the other datasets' evaluation for the LFW dataset benchmark (Table III). The MaskInv-HG achieves the best performance on all protocols, followed by MaskInvArcface, which in this specific dataset, yields a lower FMR1000 of 07.7 and 12.8 in the (U-U) and (U-M) scenarios, respectively, when compared to the ArcFace, which achieves higher FMR100 of 11.8 and 17.0 in these same scenarios, respectively. Moreover, MaskInvArcFace does not sacrifice performance in the non-occluded one, presenting a similar FRM100 of 0.9 against the 1.0 of the Arcface model.

As already verified on the CFP benchmark, for the LFW benchmark, the FocusFace does not enhance the masked face recognition performance compared to Arcface. We can verify the same situation for the EUM model, which notably does not boost the masked face recognition on our evaluation set.

Finally, we evaluate the solutions in a real masked dataset named MFR2, presented in Table IV. As verified in all the other benchmarks, the MaskInv-HG solution followed by the MaskInvArcface presents better results, proving that the synthetic approach used to train the original models boosts the performance on real data. In third came the Arcface pipeline, which achieves slightly better results when compared to FocusFace, and EUM.

TABLE III
RESULTS OF THE UNIFIED BENCHMARK RELATED TO THE FOUR SELECTED MFR PIPELINES FOLLOWING THE LFW DATASET PROTOCOL, IN TERMS OF UNMASKED-UNMASKED (U-U), UNMASKED-MASKED (U-M), MASKED-MASKED (MM) FACE VERIFICATION PERFORMANCE.

No masks	EER	ZeroFMR	FRM1000	FMR100	FMR10
Arcface [21]	00.7	01.4	01.0	00.6	00.3
MaskInvArcface [4]	00.7	01.3	00.9	00.5	00.2
MaskInv-HG [1]	00.3	00.6	00.3	00.3	00.2
FocusFace [2]	01.3	06.4	02.3	01.4	00.5
EUM [3]	01.4	07.1	03.8	01.6	00.7
Mask vs. No-Mask					
Arcface [21]	03.9	14.3	11.8	07.0	03.0
MaskInvArcface [4]	02.8	08.4	07.7	03.7	02.2
MaskInv-HG [1]	01.9	03.6	02.9	02.2	01.6
FocusFace [2]	06.3	36.5	16.3	09.8	05.5
EUM [3]	08.9	48.5	41.9	23.4	08.1
Mask vs. Mask					
Arcface [21]	05.4	23.2	17.0	09.1	03.9
MaskInvArcface [4]	04.3	16.0	12.8	06.6	03.4
MaskInv-HG [1]	02.8	06.1	05.6	03.5	02.4
FocusFace [2]	07.6	34.8	21.4	12.9	07.3
EUM [3]	07.9	60.9	31.1	18.4	06.8

TABLE IV
RESULTS OF THE UNIFIED BENCHMARK RELATED TO THE FOUR SELECTED MFR PIPELINES FOLLOWING THE MFR2 DATASET PROTOCOL.

Unmasked-Masked	EER	ZeroFMR	FRM1000	FMR100	FMR10
Arcface [21]	07.8	26.2	26.2	13.7	06.6
MaskInvArcface [4]	06.8	14.9	14.9	11.6	06.6
MaskInv-HG [1]	05.4	08.3	08.3	07.5	04.5
FocusFace [2]	15.3	26.4	26.4	19.3	15.3
EUM [3]	12.0	40.1	40.1	29.0	12.7

V. CONCLUSIONS

In this study, we verify the performance drop when existing models trained on non-occluded faces are exposed to masked faces during evaluation by evaluating the Arcface model on masked data, corroborating the need for tailored solutions to masked face recognition compared in this document. With this benchmark, we contrast four diverse approaches not explicitly compared in the literature, evidencing the diverse techniques that can be used to deal with this problem. We especially highlight the MaskInv-HG solution, which archives consistently better results in the masked scenarios of all the datasets while retaining similar and even better performances concerning the SOTA Arcface non-occluded pipeline, in the unmasked-unmasked scenario.

As a natural extension of the comparative work done here, we identified two paths: 1) To evaluate computer vision models such as Vision Transformers [42], DeiT (Data-Efficient Image Transformers) [43] and the recent ConvNeXt [44] as backbones to the masked face recognition model; 2) To collect, using a script, images from the Internet of celebrities and politicians using and not using real masks, aiming for a fair and natural comparison of the studied models. Moreover, this dataset will be publicly released since it will be constructed using only public images.

ACKNOWLEDGMENTS

This work was supported by a tripartite-contract, i.e., unico - idTech, UFPR (Federal University of Paraná), and FUNPAR (*Fundação da Universidade Federal do Paraná*).

REFERENCES

[1] M. Huber, F. Boutros, F. Kirchbuchner, and N. Damer, "Mask-invariant face recognition through template-level knowledge distillation," in *FG 2021*, 2021, pp. 1–8.

[2] P. C. Neto, F. Boutros, J. R. Pinto, N. Darner, A. F. Sequeira, and J. S. Cardoso, "Focusface: Multi-task contrastive learning for masked face recognition," in *FG 2021*, 2021, pp. 01–08.

[3] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Self-restrained triplet loss for accurate masked face recognition," *Pattern Recognition*, vol. 124, p. 108473, 2022.

[4] G.-S. J. Hsu, H.-Y. Wu, C.-H. Tsai, S. Yanushkevich, and M. L. Gavrilova, "Masked face recognition from synthesis to reality," *IEEE Access*, vol. 10, 2022.

[5] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," *arXiv preprint arXiv:2008.11104*, 2020.

[6] D. Gorodnichy, S. Yanushkevich, and V. Shmerko, "Automated border control: Problem formalization," in *2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*. IEEE, 2014, pp. 118–125.

[7] G. Lovisotto, R. Malik, I. Sluganovic, M. Roeschlin, P. Trueman, and I. Martinovic, "Mobile biometrics in financial services: A five factor framework," *University of Oxford, Oxford, UK*, 2017.

[8] M. Ngan, P. Grother, and K. Hanaoka, "Ongoing face recognition vendor test (frvt) part 6a: Face recognition accuracy with masks using pre- covid-19 algorithms," NIST - National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., 2020.

[9] B. Batagelj, P. Peer, V. Štruc, and S. Dobrišek, "How to correctly detect face-masks for covid-19 from visual information?" *Applied Sciences*, vol. 11, no. 5, p. 2070, 2021.

[10] N. Damer, F. Boutros, M. Sußmilch, F. Kirchbuchner, and A. Kuijper., "Extended evaluation of the effect of real and simulated masks on face recognition performance," *IET Biometrics*, vol. 10, no. 5, 2021.

[11] N. Damer, J. H. Grebe, C. Chen, F. Boutros, F. Kirchbuchner, and A. Kuijper, "The effect of wearing a mask on face recognition performance: an exploratory study," in *BIOSIG*, 2020.

[12] B. Wang, J. Zheng, and C. L. P. Chen, "A survey on masked facial detection methods and datasets for fighting against covid-19," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 3, pp. 323–343, 2022.

[13] M. Ngan, P. Grother, and K. Hanaoka, "Draft supplement - ongoing face recognition vendor test (frvt): Part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms," NIST - National Institute of Standards and Technology - USA, Tech. Rep., 2022.

[14] P. C. Neto, F. Boutros, J. R. Pinto, M. Saffari, N. Damer, A. F. Sequeira, and J. S. Cardoso, "My eyes are up here: Promoting focus on uncovered regions in masked face recognition," *arXiv:2108.00996*, 2021.

[15] B. Fu, F. Kirchbuchner, and N. Damer, "The effect of wearing a face mask on face image quality," in *FG 2021*, 2021, pp. 1–8.

[16] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper, "Real masks and spoof faces: On the masked face presentation attack detection," *Pattern Recognition*, vol. 123, p. 108398, 2022.

[17] S. Seneviratne, N. Kasthuriarachchi, S. Rasnayaka, D. Hettiachchi, and R. Shariffdeen, "Does a face mask protect my privacy?: Deep learning to predict protected attributes from masked face images," in *AI 2021: Advances in Artificial Intelligence*, 2022, pp. 91–102.

[18] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. Technical Report 07-49, October 2007.

[19] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *CVPRW*, 2017.

[20] S. Sengupta, J. Cheng, C. Castillo, V. Patel, R. Chellappa, and D. Jacobs, "Frontal to profile face verification in the wild," in *WACV*, 2016.

[21] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019.

[22] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," *arXiv preprint arXiv:2109.09416*, 2021.

[23] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper, "Mixfacenets: Extremely efficient face recognition networks," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[24] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 225–14 234.

[25] M. Ngan, P. Grother, and K. Hanaoka, "Ongoing face recognition vendor test (frvt): Part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms," NIST - National Institute of Standards and Technology - USA, Tech. Rep., 2020.

[26] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic," *Measurement*, vol. 167, p. 108288, 2021.

[27] B. Qin and D. Li, "Identifying facemask-wearing condition using image super-resolution with classification network to prevent covid-19," *Sensors*, vol. 20, no. 18, p. 5236, 2020.

[28] N. U. Din, K. Javed, S. Bae, and J. Yi, "A novel gan-based network for unmasking of masked face," *IEEE Access*, vol. 8, 2020.

[29] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[30] D. Zeng, R. Veldhuis, and L. Spreeuwers, "A survey of face recognition techniques under occlusion," *IET Biometrics*, vol. 10, no. 6, 2021.

[31] G. Jeevan, G. C. Zacharias, M. S. Nair, and J. Rajan, "An empirical study of the impact of masks on face recognition," *Pattern Recognition*, vol. 122, p. 108308, 2022.

[32] F. Boutros *et al.*, "Mfr 2021: Masked face recognition competition," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2021, pp. 1–10.

[33] J. Deng, J. Guo, X. An, Z. Zhu, and S. Zafeiriou, "Masked face recognition challenge: The insightface track report," in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1437–1444.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[35] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision – ECCV 2016*, 2016, pp. 87–102.

[36] X. An, X. Zhu, Y. Xiao, L. Wu, M. Zhang, Y. Gao, B. Qin, D. Zhang, and Y. Fu, "Partial FC: training 10 million identities on a single machine," *CoRR/arXiv*, vol. abs/2010.05222, 2020.

[37] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, and J. Zhou, "Webface260m: A benchmark unveiling the power of million-scale deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[38] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 919–928.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[40] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[41] J. Wang, Y. Liu, Y. Hu, H. Shi, and T. Mei, "Facex-zoo: A pytorch toolbox for face recognition," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3779–3782.

[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[43] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *ICML*, vol. 139, July 2021.

[44] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," *arXiv preprint arXiv:2201.03545*, 2022.