




Extended evaluation of the effect of real and simulated masks on face recognition performance

Naser Damer^{1,2}  | Fadi Boutros^{1,2}  | Marius Süßmilch¹ | Florian Kirchbuchner¹ | Arjan Kuijper^{1,2} 

¹Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany

²Mathematical and Applied Visual Computing, Darmstadt, Germany

Correspondence

Naser Damer, Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany.
Email: naser.damer@igd.fraunhofer.de

Funding information

Bundesministerium für Bildung und Forschung, Grant/Award Number: ATHENE; Hessisches Ministerium für Wissenschaft und Kunst, Grant/Award Number: ATHENE
Open Access funding enabled and organized by Projekt DEAL.

Abstract

Face recognition is an essential technology in our daily lives as a contactless and convenient method of accurate identity verification. Processes such as secure login to electronic devices or identity verification at automatic border control gates are increasingly dependent on such technologies. The recent COVID-19 pandemic has increased the focus on hygienic and contactless identity verification methods. The pandemic has led to the wide use of face masks, essential to keep the pandemic under control. The effect of mask-wearing on face recognition in a collaborative environment is currently a sensitive yet understudied issue. Recent reports have tackled this by using face images with synthetic mask-like face occlusions without exclusively assessing how representative they are of real face masks. These issues are addressed by presenting a specifically collected database containing three sessions, each with three different capture instructions, to simulate real use cases. The data are augmented to include previously used synthetic mask occlusions. Further studied is the effect of masked face probes on the behaviour of four face recognition systems—three academic and one commercial. This study evaluates both masked-to-non-masked and masked-to-masked face comparisons. In addition, real masks in the database are compared with simulated masks to determine their comparative effects on face recognition performance.

1 | INTRODUCTION

In hygiene-sensitive scenarios, such as the current COVID-19 pandemic, the importance of contactless and high-throughput operations is escalating, especially at crowded facilities such as airports. An existing accurate and contactless identity verification method is face recognition. However, covering the face with a facial mask has been forced in public places in many countries during the COVID-19 pandemic to fight the spread of the contagious disease. This fact can influence the performance, and thus trust, of the face recognition system and generate questions about its functionality under a situation where the individuals' faces are masked.

In the scope of face-detection algorithms, face occlusion has been addressed by many researchers [1]. Additionally, there has been a growing interest in developing solutions targeting occlusion-invariant face recognition [2]. However,

most such works have addressed general face occlusions typically appearing in the wild—for example, partial captures and sunglasses. Given the current situation of the COVID-19 pandemic, it is crucial to analyse the exact effect of wearing facial masks on the behaviour and performance of face recognition systems under a collaborative use-case verification. This work aims to evaluate and analyse this effect as a needed effort to enable the development of solutions addressing accurate face verification under these scenarios. To that end, this paper presents the following contributions:

- A database based on a realistically variant collaborative face capture scenario. This database contains three sessions per subject, and each includes three capture variations. The database includes face images with and without masks in addition to face images with simulated masks.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *IET Biometrics* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

- A study of the behaviour of three widely studied academic face recognition solutions and one commercial off-the-shelf face recognition solution (COTS) when encountering masked faces compared with the typical baseline of not wearing a face mask.
- An evaluation of verification performance when comparing each masked face with others, both real and simulated, to estimate the validity of having masked and not-masked reference images.
- An investigation of the validity of using face images with synthetically simulated masks in evaluating face recognition systems for deployment on real masked faces. This is performed by comparing the effects of simulated and real masked face images in our face recognition performance data.

Our conclusion includes pointing out the strong signs of the negative effect on face recognition performance, stressing the need for appropriate evaluation databases and recognition solutions. We also point out the difficulty in assuming that simulated masks represent real masks well in face recognition evaluation studies. Our evaluation also points out that comparing masked faces might be more accurate than comparing masked and non-masked faces for some face recognition systems.

This paper is an invited extension to the paper [3], which achieved the best paper award at BIOSIG 2020 [4]. This paper is extended on the invited conference paper as follows:

- The database presented here and used for the provided study is extended in size and variation. The presented database includes 48 subjects rather than 24 as in the invited work. The new database also includes a set of augmented samples that represent faces with simulated masks.
- The evaluation in this work expands on the invited paper by studying the effect of verifying masked-to-masked face pairs in comparisons with no-mask pairs and mask-to-no-mask pairs. Thus, an experimental probe is provided for masked or synthetically masked reference images.
- This study also extends the invited paper by including a study on the effect of simulated masks on face recognition performance in comparison with the effect of real masks in the provided database.
- As an addition to the common verification performance measures presented here and in the invited paper, we provide an indication of generalizability by reporting a measure of genuine and imposter score separability, the Fisher discriminant ratio (FDR), for all experimental setups.

2 | RELATED WORK

Many operational challenges are faced in the deployment of face recognition solutions. Issues related to attacks on face recognition systems are considered the most important of

these challenges and receive most of the research attention. Such attacks can be morphing attacks [5, 6], presentation attacks (spoofing) [7], or different unconventional attacks [8]. Face recognition deployability is also affected by the biometric sample capture [9] and presentation [10], including face occlusions. Occluded face detection is a widely studied challenge in the domain of computer vision. A study by Optiz et al. [1] is a clear example of occluded face detection—they targeted the accurate detection of occluded faces by presenting a solution based on a novel grid loss. Ge et al. [11] focussed on detecting faces (not face recognition) with mask occlusions in in-the-wild scenarios. Their research included face-covering objects in a wider perspective rather than facial masks worn for medical or hygiene reasons. Such studies are highly relevant to face recognition because the detection of faces (while wearing masks) is an essential preprocessing step where face recognition systems might fail, as shown later in our experiments.

As discussed, the detection of occluded faces is one of the challenges facing the deployment of biometric face systems. However, the biometric recognition of occluded faces is a more demanding challenge. One of the recent works to address this issue is Song et al.'s [2] research aiming to improve face recognition performance under general occlusions. The approach presented by Song et al. [2] targets localizing and abandoning corrupted feature elements, which might be associated with occlusions, from the recognition process. A very recent work presented by Wang et al. [12] focussed on masked faces. In a short and underdetailed presentation, their work introduced crawled databases for face recognition and detection as well as simulated masked faces. The authors claim to have increased verification accuracy from 50% to 95%; however, they did not provide any information on the baseline used, the proposed algorithmic details, or clarity on the evaluation database used. A recent preprint by Anwar and Raychowdhury [13] has presented a database that includes 296 face images, partially with real masks, of 53 identities. The images in the database can be considered captured under in-the-wild conditions, as they are crawled from the Internet and do not represent a collaborative face recognition scenario. They proposed fine-tuning an existing face recognition network to achieve better evaluation performance. On a larger scale, the National Institute of Standards and Technology (NIST), as a part of the ongoing Face Recognition Vendor Test (FRVT), has published a specific study (FRVT—Part 6A) on the effect of face masks on the performance on face recognition systems provided by vendors [14]. The NIST study concluded that the algorithm accuracy with masked faces declined substantially. One of the main study limitations is the use of simulated masked images under the questioned assumption that their effect represents real face masks, an issue that this article tackles. All the mentioned studies did not include an evaluation of masked-to-masked face verification, which might motivate requiring an additional masked reference image, an issue also tackled in this article.

Under the current COVID-19 pandemic, an explicitly gathered database and the evaluation of real face masks on collaborative face recognition is essential. This issue, as cleared

above, also includes the need to study the appropriateness of using simulated face masks for automatic face recognition performance evaluation and assess the performance drop when comparing masked face pairs. These issues are all addressed in this work.

3 | THE DATABASE

The collected database aims to enable the analyses of face recognition performance on masked faces and motivate future research in the domain of masked face recognition. The presented database is an ongoing effort aiming at enabling large-scale face recognition evaluation. Our presented data represents a collaborative, however varying, scenario. The targeted scenario is that of unlocking personal devices or identity verification at automatic border control gates. This motivated the variations in masks, illumination, and background in the presented database.

The participants were requested to capture the data on three different days, not necessarily consecutive. Each of these days is considered one session. In each of these sessions/days, three videos, with a minimum duration of 5 s, are collected by the subjects. The videos are collected from a static webcam (not handheld), while the users are requested to look at the camera, simulating a face recognition-based login scenario. The data is collected by participants at their residences during the pandemic-induced home-office period. The data is collected during the day (indoor, daylight), and the participants were asked to remove eyeglasses when they were considered to have very thick frames. To simulate a realistic scenario, no restrictions on mask type, background, or any other restrictions were imposed. The three captured videos each day were as follows: (1) not wearing a mask with no additional electric lighting (illumination); this scenario is referred to as baseline (BL); (2) wearing a mask with no additional electric lighting (illumination); this scenario is referred to as the first masked scenario (M1); (3) wearing a mask with the room's existing electric lighting (illumination) turned on; this scenario is referred to as the second masked scenario (M2). Given that wearing a mask might result in varying shadow and reflection patterns, the variation in the illumination in M1 and M2 is considered.

The reference (R) data is the data from the first session (day), consisting of the baseline reference (BLR), the mask reference of the first capture scenario (M1R), and the mask reference of the second capture scenario (M2R). The masked

references (M1R and M2R) are joined into one masked reference subset, M12R. The probe data (P) is the data of the second and third sessions (days), including the no-mask baseline probe (BLP), mask probe of the first capture scenario (M1P), mask probe of the second capture scenario (M2P), and combined probe data (M1P and M2P) noted as M12P. The first second of each video was neglected to avoid possible biases related to subject interaction with the capture device. After the one-second skip, three seconds were considered. From these three seconds, 10 frames were extracted with a gap of 9 frames between each consecutive frame, knowing that all videos are captured at a frame rate of 30 frames per second. To allow the synthetically added masks to be comparatively evaluated, the data is augmented with additional subsets that include the images with simulated masks. As synthetically masked references, simulated masks are added to the BLR images to create the simulated mask reference (SMR) subset. For probes, the BLP subset is augmented with simulated masks to create the simulated mask probe (SMP) subset. The details of adding the simulated masks are presented in Section 5.2.

The total number of participants in the presented database is 48, compared with 24 in the first version of the database [3]. All subjects participated in all three required sessions. Table 1 presents an overview of the database structure in terms of the number of participants, sessions, and extracted frames from each video. Figure 1 shows samples of the database; please note that only samples from the SMP are shown, as there is no consistent visual difference between SMR and SMP subsets.

4 | FACE RECOGNITION

We analyse the performance of three academic face recognition solutions and one commercial face recognition solution to present a wide view of the effect mask-wearing on face recognition performance. The three academic algorithms are ArcFace [15], VGGFace [16], and SphereFace [17]. The commercial solution is the MegaMatcher 12.1 SDK [18] from Neurotechnology. The following is detailed information about the selected face recognition solutions.

SphereFace is chosen because it achieved competitive verification accuracy on Labeled Faces in the Wild (LFW) [19] 99.42% and YouTube Faces (YTF) [20] 95.0%. SphereFace contains 64 convolutional neural network layers trained on the CASIA-WebFace data set [21]. SphereFace is trained with the angular softmax loss function (A-softmax). The key concept

Session Data Split	Session 1: References					Session 2 and 3: Probes				
	BLR	M1R	M2R	M12R	SMR	BLP	M1P	M2P	M12P	SMP
Illumination	No	No	Yes	both	No	No	No	Yes	Both	No
Number of captures	480	480	480	960	480	960	960	960	1920	960

TABLE 1 An overview of the database structure and number of images in each subset

Abbreviations: BLP, baseline probe; BLR, baseline reference; M1P/M2P/M12P, mask probe of the first, second, and combined capture scenarios, respectively; M1R/M2R/M12R, mask reference of the first, second, and combined capture scenarios, respectively; SMP, simulated mask probe; SMR, simulated mask reference.



FIGURE 1 Samples of the collected database from the three capture types (BL, baseline; M1, first scenario; M2, second scenario; SMP, simulated mask probe)

behind A-softmax loss is to learn discriminative features from the face image by formulating the softmax as an angular computation between the embedded feature vectors and their weights.

ArcFace achieved state-of-the-art performance scores on several face recognition evaluation benchmarks such as LFW 99.83% and YTF 98.02%. ArcFace introduced additive angular margin loss to improve the discriminative ability of the face recognition model. We deployed ArcFace based on ResNet-100 [22] architecture pretrained on a refined version of the MS-Celeb-1M data set [23] known as MS1MV2.

VGGFace is one of the earliest face recognition models to achieve competitive face verification performance on LFW (99.13%) and YTF (97.3%) face benchmarks, doing so by using a simple network architecture trained on a public database of 2.6 million images of 2600 identities). The network architecture is based on the VGG model [24] trained with softmax loss and fine-tuned using triplet loss [25].

We used the MegaMatcher 12.1 SDK [18] from the vendor Neurotechnology. We chose this COTS product because Neurotechnology achieved one of the best performances in the recent NIST report that addressed vendors' face verification products [26]. The face quality threshold was set to zero for probes and references to minimize neglected masked faces. The full processes of detecting, aligning, feature extraction, and matching are part of the COTS, and thus we are unable to provide their algorithmic details. Comparing two faces by the COTS method produces a similarity score.

For the three academic systems, the widely used multi-task cascaded convolutional networks (MTCNN) [27] solution is employed, as recommended in [17], to detect (crop) and align (affine transformation) the face.

ArcFace and SphereFace networks process the aligned and cropped image input and produce a feature vector of size 512 from the last network layer. The VGG model produces a feature vector of size 4096 from the third-to-last output layer as recommended by the authors [16]. Two faces are compared by calculating the distance between their respective feature vectors, which is calculated as Euclidean distance for ArcFace and VGGFace features, as recommended in [15, 16], and as cosine distance for SphereFace features, as recommended in

[17]. The Euclidean distance (dissimilarity) is complemented to illustrate a similarity score, and the cosine distance shows a similarity score by default.

5 | EXPERIMENTAL SETUP

This section presents the set of experiments conducted, evaluation metrics employed, and setup used to create the synthetic masks.

5.1 | Experiments and evaluation metrics

As a baseline, we start by evaluating face verification performance without any mask-wearing influence. This evaluation performs an N:N comparison between the data splits BLR and BLP (BLR-BLP). In the subsequent step, we perform an N:N evaluation between the data subsets BLR and M12P (BLR-M12P) to measure verification performance when the probe subject is wearing a mask. These experiments are conducted on the four face recognition solutions being evaluated. Additionally, to evaluate real masked-to-masked face verification, we consider the M12R subset as a reference and perform an N:N comparison between M12R and M12P (M12R-M12P). We also evaluate the performance of the real masked references against the simulated masked probes (BLR-SMP), and we evaluate the simulated masked references against the simulated masked probes (SMR-SMP).

The effect of having probe subjects wear a face mask is studied by illustrating the imposter and genuine distributions of the BLR-M12P (mask) along with the baseline BLR-BLP comparisons. This enables deeper analyses of the distribution shifts caused by wearing a mask. To study performance in verifying not-masked references to masked references versus masked-to-masked pairs, we plot the imposter and genuine comparison score distributions of the BLR-M12P (baseline) comparisons along with the imposter and genuine score distributions of the M12R-M12P. To compare the verification performance for the not-masked references with simulated and real masked faces, we plot the imposter and genuine

comparison score distributions of the BLR-M12P and BLR-SMP pairs. To compare the verification performance of the simulated masked probes with the simulated mask references and not-masked references, we plot the imposter and genuine comparison score distributions of the BLR-SMP and SMR-SMP pairs. Additionally, we present the mean of imposter comparison scores (I-mean) and the mean of the genuine comparison scores (G-mean) for each experiment to analyse the comparison score shifts quantitatively.

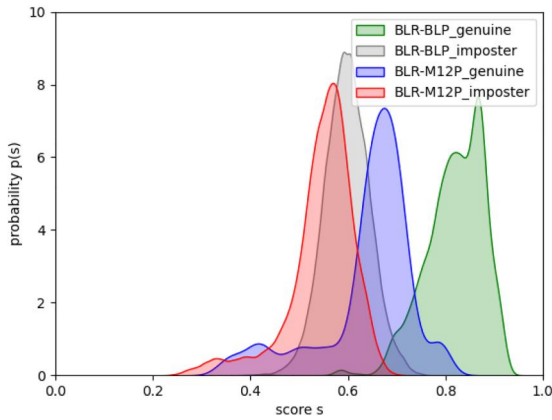
We also report verification performance metrics based on the ISO/IEC 19795-1 [28] standard. As an essential pre-processing step, face detection can be affected by the strong change in appearance induced by wearing a face mask. To capture that effect, we report the failure-to-extract rate (FTX) for our experiments. FTX measures the rate of comparison where feature extraction is not possible, and thus a template is not created. Beyond that, and for comparisons of successfully generated templates, we report a set of algorithmic verification performance metrics. From these, we report the equal error

rate (EER), which is defined as the false non-match rate (FNMR) or false match rate (FMR) at the operation point where they are equal. We additionally present the FNMR at different decision thresholds as an algorithmic verification performance metric by reporting the lowest FNMR for FMR $\leq 1.0\%$, $\leq 0.1\%$, and $\leq 0\%$, namely FMR100, FMR1000, and ZeroFMR, respectively.

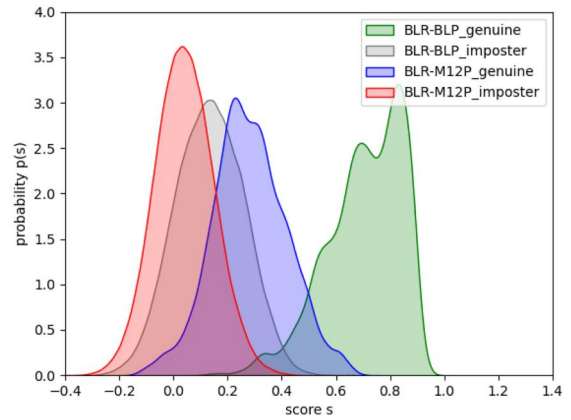
Further, we enrich our reported evaluation results by reporting the FDR to provide an in-depth analysis of the separability of genuine and imposter scores for different experimental settings. FDR is a class separability criterion used by [29, 30], and it is given by:

$$FDR = \frac{(\mu_G - \mu_I)^2}{(\sigma_G)^2 + (\sigma_I)^2},$$

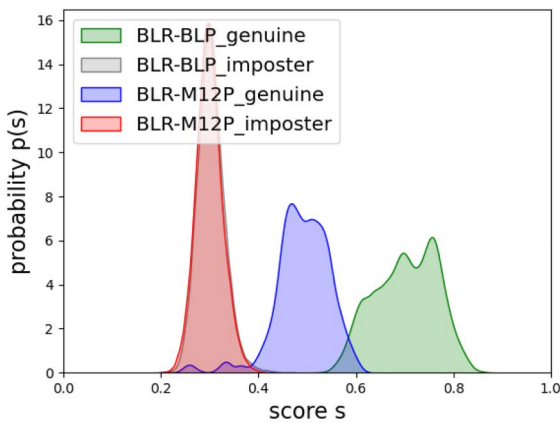
where μ_G and μ_I are the genuine and imposter scores' means, and σ_G and σ_I are their standard deviations. The larger the



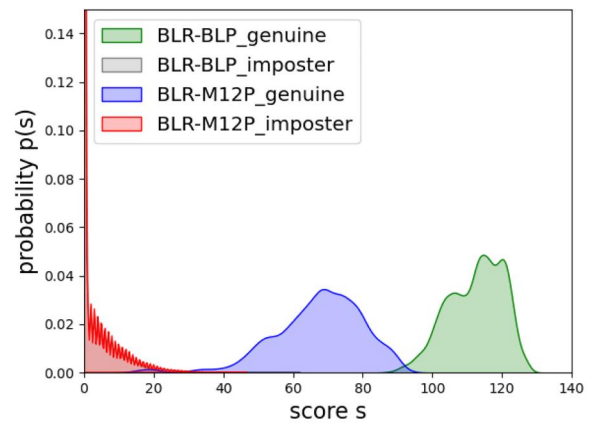
(a) VGG: BLR-BLP and BLR-M12P.



(b) Sphere: BLR-BLP and BLR-M12P.



(c) ArcFace: BLR-BLP and BLR-M12P.



(d) COTS: BLR-BLP and BLR-M12P.

FIGURE 2 The comparison score (similarity) distributions comparing the ‘baseline’ BLR-BLP genuine and imposter distributions with those of the distributions including ‘masked’ face probes (BLR-M12P). The shift in the genuine scores towards the imposter distribution is clear when faces are masked for all the investigated systems. (a) VGGFace, (b) SphereFace, (c) ArcFace, and (d) COTS

TABLE 2 Verification performance measures FDR, G-mean, and I-mean achieved by VGGFace on the different experimental setups

VGG	EER	FMR100	FMR1000	ZeroFMR	G-mean	I-mean	FDR	FTX
BLR-BLP	2.3795%	4.6561%	9.9513%	19.2483%	0.819001	0.596985	8.6816	0.000%
BLR-M12P	18.9209%	47.5859%	67.8121%	86.9508%	0.639760	0.544981	0.6378	4.736%
M12R-M12P	20.9248%	51.3110%	76.9665%	84.9871%	0.693083	0.569062	0.6135	4.736%
BLR-SMP	7.6525%	14.1156%	54.2517%	70.4082%	0.695138	0.576018	3.0399	5.114%
SMR-SMP	3.6799%	5.8282%	23.0061%	42.9448%	0.800212	0.610588	1.3912	5.371%

Note: The performance degradation induced by the masked face probes.

Abbreviations: BLP, baseline probe; BLR, baseline reference; EER, equal error rate; FDR, Fisher discriminant ratio; FMR, false match rate; FTX, failure-to-extract rate; M12P, mask probe of the combined first and second capture scenarios; M1R/M2R/M12R, mask reference of the first, second, and combined capture scenarios, respectively; SMP, simulated mask probe; SMR, simulated mask reference.

TABLE 3 Verification performance measures FDR, G-mean, and I-mean achieved by SphereFace on different experimental setups

SphereFace	EER	FMR100	FMR1000	ZeroFMR	G-mean	I-mean	FDR	FTX
BLR-BLP	2.5237%	3.5856%	9.1958%	43.1286%	0.703659	0.132542	9.2316	0.00000
BLR-M12P	15.6472%	57.7897%	79.1916%	96.8344%	0.280806	0.039197	1.8983	4.736%
M12R-M12P	25.5317%	78.5115%	91.3690%	99.5655%	0.533365	0.266994	0.9161	4.736%
BLR-SMP	15.4323%	31.6846%	46.3397%	76.9503%	0.415802	0.04813	2.6162	5.114%
SMR-SMP	12.7381%	23.9176%	39.1283%	68.7882%	0.55719	0.343760	2.332	5.371%

Note: the performance degradation induced by the masked face probes.

Abbreviations: BLP, baseline probe; BLR, baseline reference; EER, equal error rate; FDR, Fisher discriminant ratio; FMR, false match rate; FTX, failure-to-extract rate; M12P, mask probe of the combined first and second capture scenarios; M1R/M2R/M12R, mask reference of the first, second, and combined capture scenarios, respectively; SMP, simulated mask probe; SMR, simulated mask reference.

TABLE 4 Verification performance measures FDR, G-mean, and I-mean achieved by ArcFace on different experimental setups

ArcFace	EER	FMR100	FMR1000	ZeroFMR	G-mean	I-mean	FDR	FTX
BLR-BLP	0.0000%	0.0000%	0.0000%	0.0000%	0.702020	0.302603	33.4823	0.000%
BLR-M12P	2.8122%	3.3917%	4.2793%	11.0741%	0.490206	0.298841	9.3521	4.736%
M12R-M12P	4.7809%	4.7959%	4.9427%	99.7933%	0.624205	0.318468	7.6235	4.736%
BLR-SMP	1.1652%	1.3002%	5.4551%	13.0724%	0.507047	0.300292	12.3798	5.114%
SMR-SMP	0.2732%	0.0000%	0.6151%	6.9195%	0.675107	0.330339	6.4712	5.371%

Note: The performance degradation induced by the masked face probes.

Abbreviations: BLP, baseline probe; BLR, baseline reference; EER, equal error rate; FDR, Fisher discriminant ratio; FMR, false match rate; FTX, failure-to-extract rate; M12P, mask probe of the combined first and second capture scenarios; M1R/M2R/M12R, mask reference of the first, second, and combined capture scenarios, respectively; SMP, simulated mask probe; SMR, simulated mask reference.

TABLE 5 Verification performance measures FDR, G-mean, and I-mean achieved by COTS on different experimental setups

COTS	EER	FMR100	FMR1000	ZeroFMR	G-mean	I-mean	FDR	FTX
BLR-BLP	0.0000%	0.0000%	0.0000%	0.0000%	112.3174	1.6304	145.2178	0.000%
BLR-M12P	1.0185%	1.0747%	1.6454%	5.3296%	67.3309	1.6667	29.9121	0.000%
M12R-M12P	0.0417%	0.0000%	0.0248%	3.2992%	102.4084	6.73598	37.8482	0.000%
BLR-SMP	0.6002%	0.6337%	0.6337%	0.8684%	74.3175	1.7565	30.6757	0.000%
SMR-SMP	0.9322%	0.1081%	0.8917%	1.4317%	104.7186	5.8493	45.3178	0.000%

Note: The performance degradation induced by the masked face probes.

Abbreviations: BLP, baseline probe; BLR, baseline reference; EER, equal error rate; FDR, Fisher discriminant ratio; FMR, false match rate; FTX, failure-to-extract rate; M12P, mask probe of the combined first and second capture scenarios; M1R/M2R/M12R, mask reference of the first, second, and combined capture scenarios, respectively; SMP, simulated mask probe; SMR, simulated mask reference.

FDR value is, the higher the separation between the genuine and imposter scores, and thus verification performance and its generalizability are expected to improve. Furthermore, we report the receiver operating characteristic (ROC) curves for different experiments to illustrate the algorithmic verification performance at a wider range of thresholds (operation points) for the different face recognition systems considered.

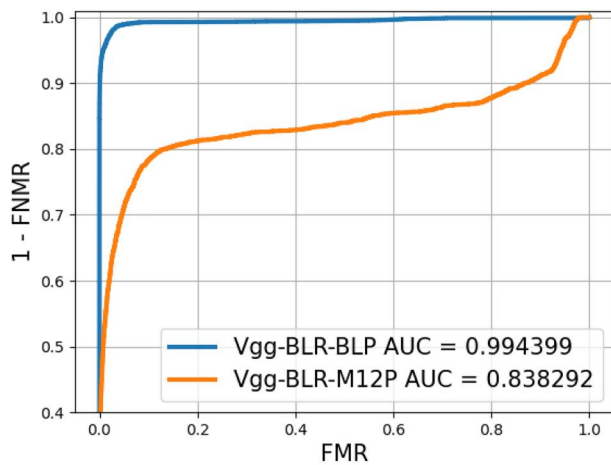
5.2 | Simulated mask

We use the synthetic mask generation method described by the NIST report [14]. The synthetic generation method depends on the Dlib toolkit [31] to detect and extract 68 facial landmarks from a face image. Based on the extracted landmark points, a face mask of different shapes, heights, and colours can be drawn on the face images. The detailed

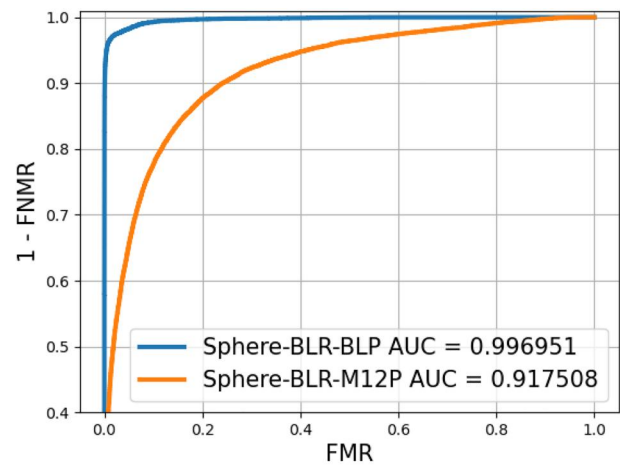
implementation of the synthetic generation method is described in [14], and the open-source implementation can be found under [32] as provided by Boutros et al. [33]. The synthetic mask generation method provided in [33] offers different face mask types with different heights and coverages. To generate a synthetic mask database, we first extract the facial landmarks of each face image. Then, for each face image, we generate a synthetic masked image of the mask type C and the colour blue, described in [14, 33].

6 | EVALUATION RESULTS

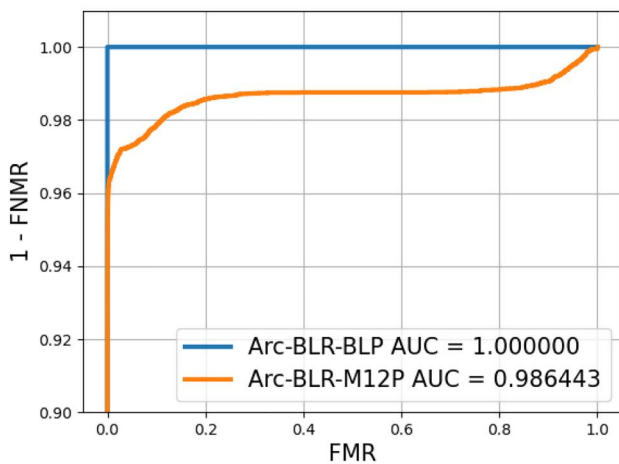
The evaluation results are based on the presented database and the experimental setup. These results are discussed in this section in the form of answering practical questions on the issue of masked face recognition.



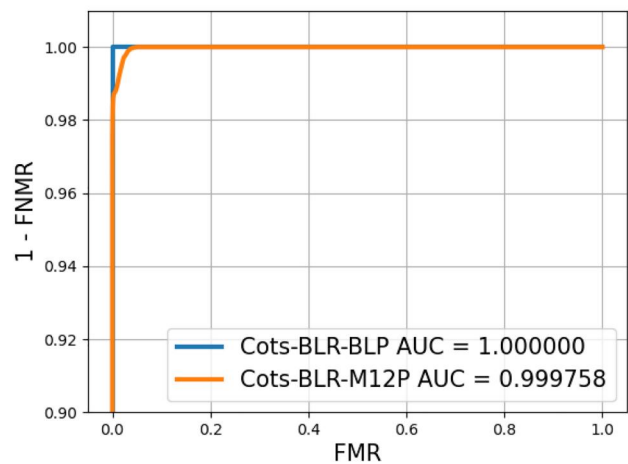
(a) VGG - mask probe vs no mask probe.



(b) SphereFace - mask probe vs no mask probe.



(c) ArcFace - mask probe vs no mask probe.



(d) COTS - mask probe vs no mask probe.

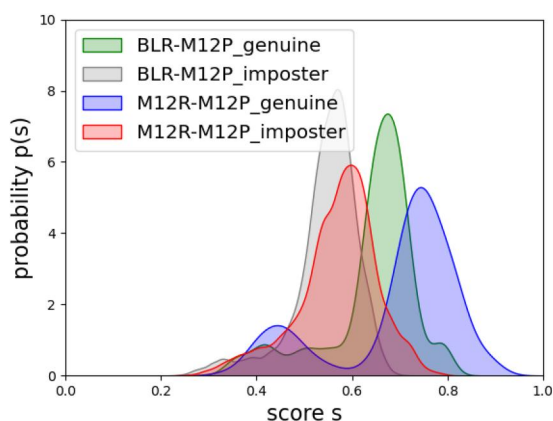
FIGURE 3 The verification performance of the four investigated systems—VGGFace (a), SphereFace (b), ArcFace (c), and COTS (d)—is presented in the form of receiver operating characteristic (ROC) curves. For each system, two curves are plotted to represent the settings that include ‘masked’ face probes (BLR-M12P) and the unmasked baseline (BLR-BLP). The area under curve (AUC) is also listed for each ROC curve. As in Tables 2–5, the negative effect of masked probes is apparent in the performance of the VGGFace ArcFace, SphereFace, and COTS models

6.1 | How does wearing a face mask affect face verification performance when the reference is not masked?

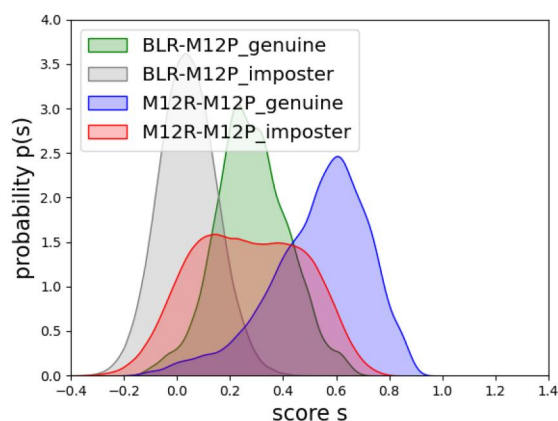
The comparisons between the baseline (BLR-BLP) imposter and genuine score distributions and the experiments with masked face probes (BLR-M12P) on the four face recognition solutions considered are presented in Figure 2. One can notice in all experimental setups that, when comparing unmasked references with masked face probes, the genuine score distributions strongly shift towards the imposter distributions in comparison with the baseline BLR-BLP setup. This points out an expected decrease in performance and general trust in the verification decision, as the separability between imposter and genuine samples decreases. On the other hand, the imposter score distributions do not seem to be affected by the masked probes (BLR-M12P) in comparison with the unmasked

baseline (BLR-BLP) in the better performing COTS and ArcFace. In the lower performing SphereFace and VGGFace, imposter score distributions do shift towards the genuine distributions, however, this shift is significantly smaller than the genuine distribution shift. This points out that, given a preset decision threshold, having masked probes will result in a larger increase in FNMR values in comparison with the increase in the FMR values.

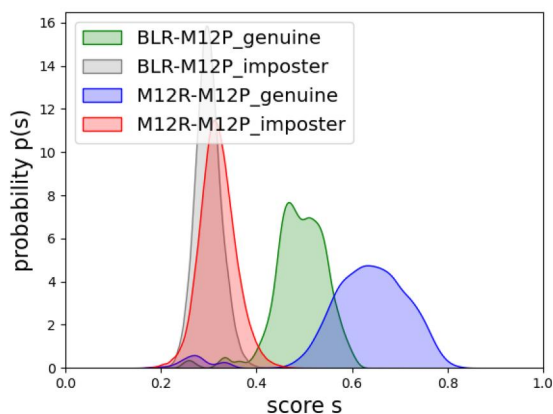
Tables 2–5 list the achieved verification performance, given by the different evaluation metrics, on all experimental setups by the VGGFace, SphereFace, ArcFace, and COTS solutions, respectively. For all academic systems, wearing a face mask affected the ability to detect the face properly, resulting in a higher than zero (as in the baseline) FTX. The FTX values for the three solutions are identical as they all use the MTCNN network for face detection and alignment. The COTS solution was able to produce results for all comparison pairs and thus



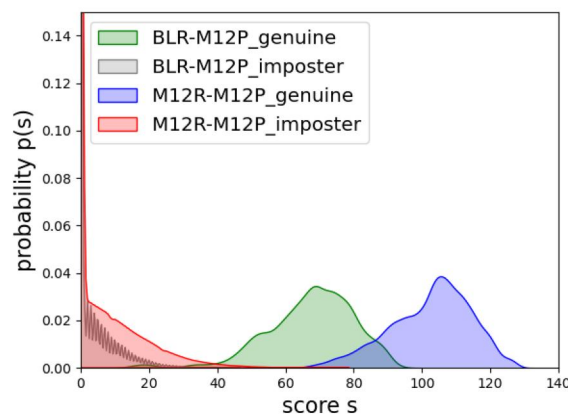
(a) VGG: BLR-M12P and M12R-M12P.



(b) Sphere: BLR-M12P and M12R-M12P.



(c) ArcFace: BLR-M12P and M12R-M12P.



(d) COTS: BLR-M12P and M12R-M12P.

FIGURE 4 The comparison score (similarity) distributions from comparing the no-mask-to-mask ‘baseline’ BLR-M12P genuine and imposter distributions to those of the distributions including ‘masked’ face references and probes (M2R-M12P). In the COTS model, separability between the genuine and imposter distributions appears to be higher when comparing masked references with masked probes than when comparing not-masked references with masked probes. (a) VGGFace, (b) SphereFace, (c) ArcFace, and (d) COTS

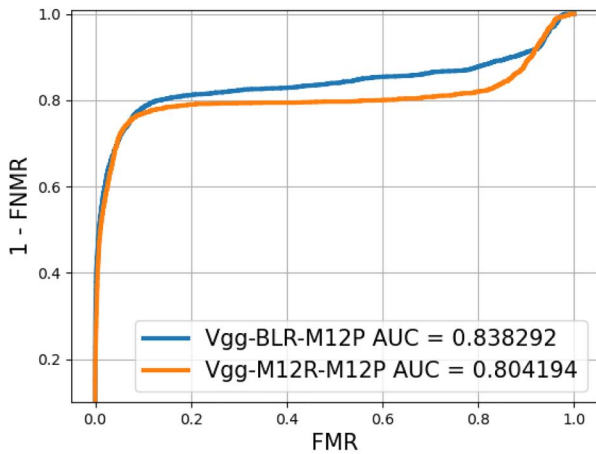
achieved an FTX of 0%. However, it must be noted that the quality thresholds in the COTS were set to zero for both, the reference and probe images.

The verification performance metrics (EER, FMR100, FMR1000, ZeroFMR) reported for the COTS, ArcFace, SphereFace, and VGGFace are negatively affected when the probe faces are masked (BLR-M12P), see Tables 2–5. The reduction in the performance is much more apparent in the VGGFace and SphereFace solution in comparison with the COTS and ArcFace. For all systems, the G-mean values decreased significantly when considering the masked probes. This, despite the relatively small size of the evaluation data, indicates a strong negative effect of the masks on the face recognition performance. On the other hand, the I-mean values when considering the masked faces, in comparison with the baseline (BLR-BLP), were only slightly changed under the COTS and ArcFace solutions. For the SphereFace and

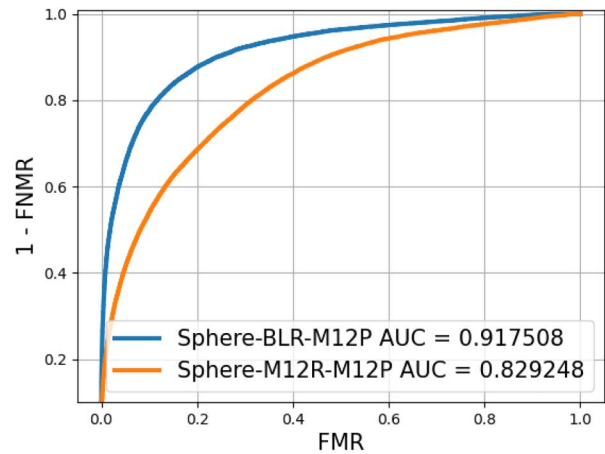
VGGFace solutions, the I-mean values are affected, although not as drastically as the G-mean. The separability measure FDR was also significantly affected across the four systems by having a masked probe.

To show verification performance over a wider range of operation points, Figure 3 presents the ROC curves for the different experimental settings for each of the four investigated systems. Similar conclusions to those established from Tables 2–5 can be made. The COTS, ArcFace, VGGFace, and SphereFace verification performances are affected by the masked face probes.

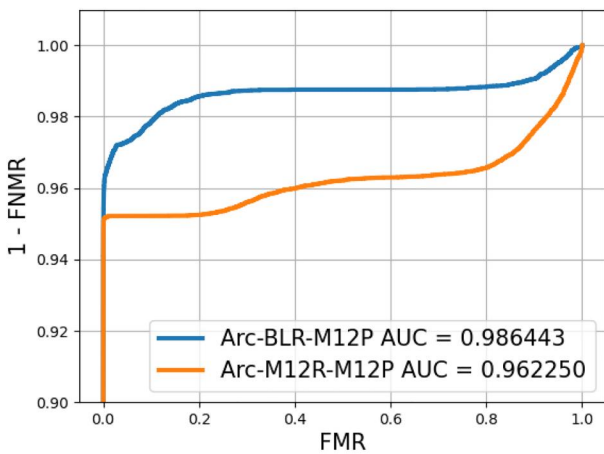
As an answer to this subsection title, having a masked face probe does significantly affect the verification performance of the investigated face recognition systems. This effect is seen more in the large shift in the genuine score values in comparison with that in the imposter score values.



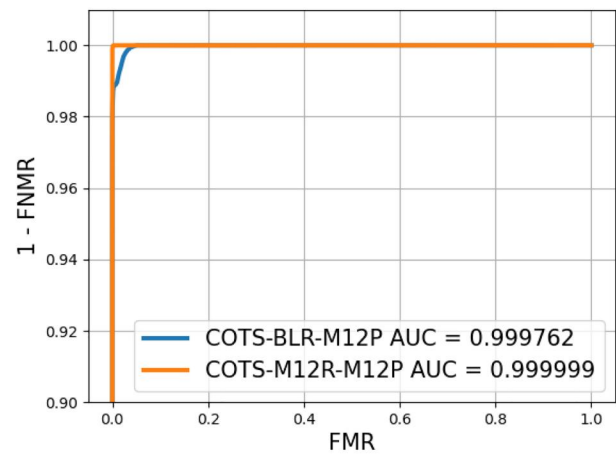
(a) VGG - mask vs no mask reference (masked probe).



(b) SphereFace - mask vs no mask reference (masked probe)



(c) ArcFace - mask vs no mask reference (masked probe).



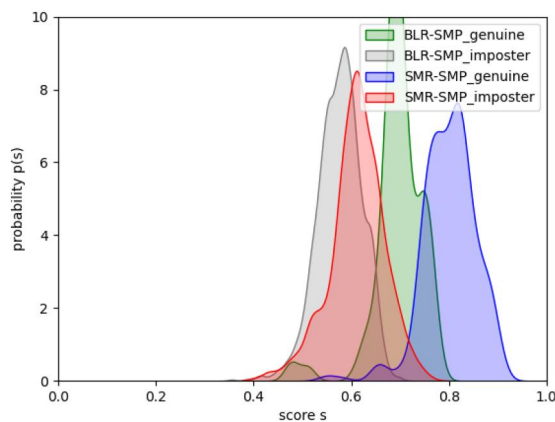
(d) COTS - mask vs no mask reference (masked probe).

FIGURE 5 The verification performance of the four investigated systems—VGGFace (a), SphereFace (b), ArcFace (c), and COTS (d)—is presented in the form of receiver operating characteristic (ROC) curves. For each system, two curves are plotted to represent the settings that include not-masked references and ‘masked’ face probes (BLR-M12P) as baselines and the masked references and masked probes (M12R-M12P). The area under curve (AUC) is also listed for each ROC curve. As in Tables 2–5, the negative effect of both masked references and probes in comparison with only masked probes is apparent in the performance of the academic solutions VGGFace, ArcFace, and SphereFace

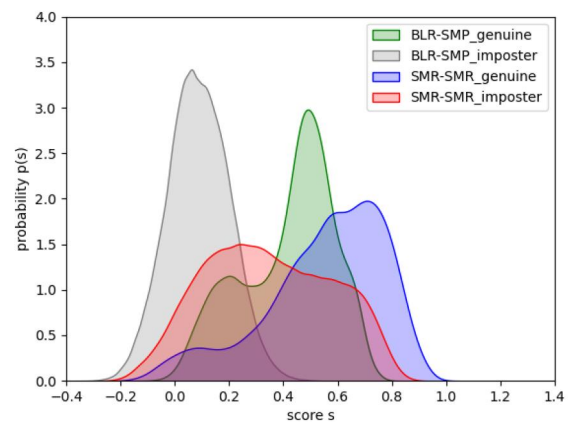
6.2 | Does having a masked reference enhance verification performance when the probe is masked?

Figure 4 presents a comparison between the baseline where the reference is not masked and the probe is masked (BLR-M12P) on one side, and also the case where the reference is masked (M2R-M12P). In both cases, the genuine imposter separability appears low. Both the genuine and imposter scores shift to the higher values when the reference is masked, in comparison with not-masked reference. This might be due to the fact that the masked area of the face appears more similar when covered by a mask in both the compared images. This is visible in all the four investigated face recognition solutions. This is also supported by the G-mean and I-mean values presented in Tables 2–5.

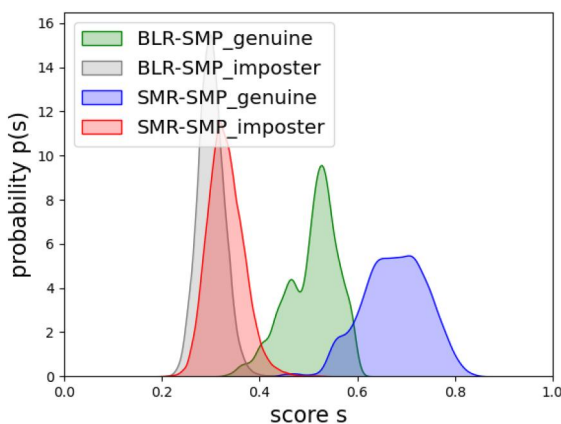
The verification performance of the three academic solutions presented in Tables 2–4 indicates lower performance when the probe is masked and the reference is masked, in comparison with not-masked reference. This case is more apparent in the SphereFace and ArcFace, and to a less degree for the VGGFace solution. The FDR separability measure indicates a similar trend. This performance drop is also visible on a wider range of operation points in the ROC curves in Figure 5 for the three investigated solutions. On the other hand, the COTS solution performs better when the probe is masked and the reference is masked, in comparison with not-masked reference, see Table 5. This is also supported by the FDR value in the table and on a wider operation range in Figure 5. Moreover, one should not neglect the large difference in the imposter and genuine score values between both cases, BLR-M12P and M12R-M12P.



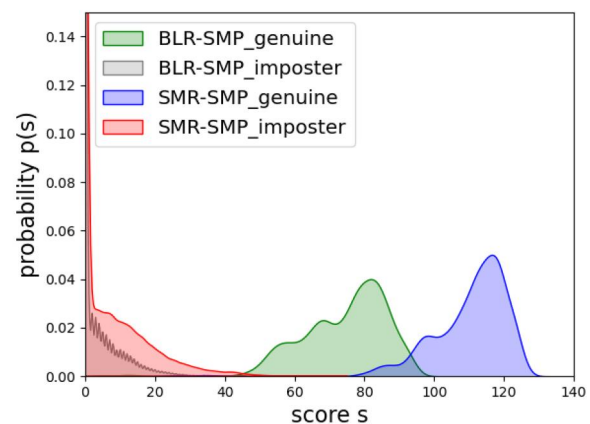
(a) VGG: BLR-SMP and SMR-SMP.



(b) Sphere: BLR-SMP and SMR-SMP.



(c) ArcFace: BLR-SMP and SMR-SMP.



(d) COTS: BLR-SMP and SMR-SMP.

FIGURE 6 The comparison score (similarity) distributions comparing the not-masked references with simulated masked probe ‘baseline’ BLR-SMP genuine and imposter distributions to those of the distributions including ‘simulated’ masked face references and probes SMR-SMP. The genuine scores of comparing simulated mask probes and references SMR-SMP are generally higher than those of the real masked references BLR-SMP in the investigated systems. (a) VGGFace, (b) SphereFace, (c) ArcFace, and (d) COTS

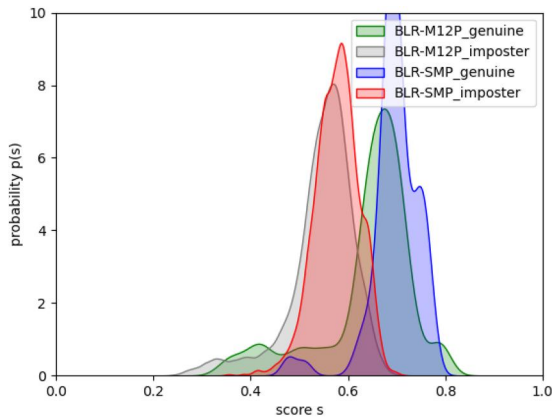
As an answer to this subsection title, having a masked face reference does not enhance verification performance with a masked probe, in comparison with having a not-masked reference, on most investigated systems. This eliminates the need of having multiple references (masked and not masked) in face verification galleries, at least when the academic investigated systems are considered, which are not optimized for masked faces. On the other hand, the COTS investigated system performs better when both the probe and imposter are masked.

6.3 | Does the answer to our last question (Section 6.2) also apply to simulated masks?

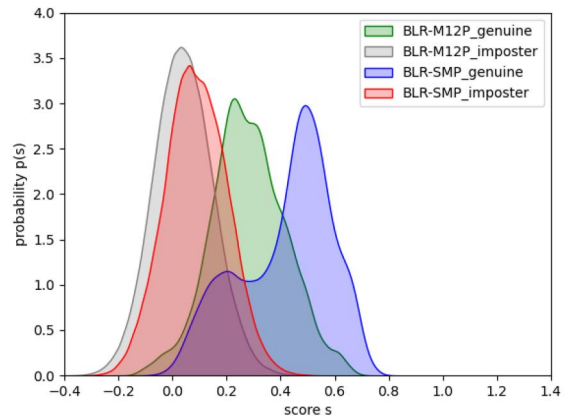
Figure 6 presents a comparison between the baselines where the references are not masked and the probes are synthetically masked (BLR-SMP) on one side, and also the case where the references are synthetically masked (SMR-SMP). Both the genuine and imposter scores seem to slightly shift to the higher

values when the references are synthetically masked, in comparison with not-masked reference in the investigated systems. This is also supported by the G-mean and I-mean values presented in Tables 2–5. The verification performance presented in these tables indicates slightly better performance of the three academic solutions when the references are also synthetically masked. The COTS, on the other hand, performs better when only the references are not masked. These observations are in contrast to the observations made in Section 6.2 where the investigated masks were real. This might be due to the high similarity (identical) in the simulated masks, which leaves only the visible part of the face as a comparison information source. This is supported by the large increase in G-mean (in comparison with increase in I-mean) in the SMR-SMP case in Tables 2–5.

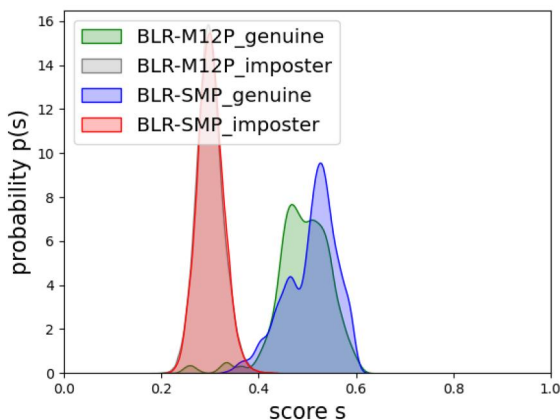
As an answer to this subsection title, unlike the real masks, verifying two synthetically masked faces do perform slightly better than comparing a not-masked reference with a synthetically masked face on the academic solutions. This different conclusion, in comparison with the situation with



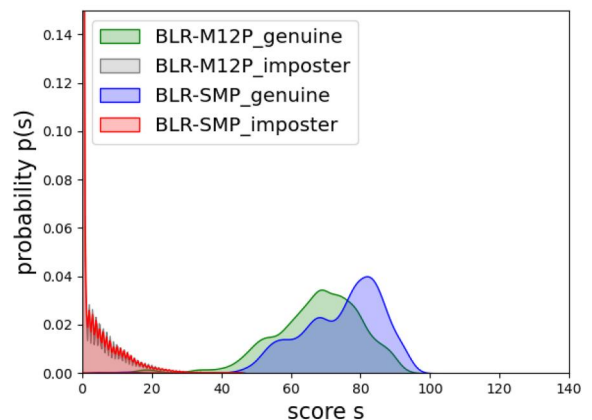
(a) VGG: BLR-M12P and BLR-SMP.



(b) Sphere: BLR-M12P and BLR-SMP.



(c) ArcFace: BLR-M12P and BLR-SMP.



(d) COTS: BLR-M12P and BLR-SMP.

FIGURE 7 The comparison score (similarity) distributions from comparing the not-masked references with masked probe ‘baseline’ BLR-M12P genuine and imposter distributions to those of the distributions including ‘simulated’ masked face probes BLR-SMP. The genuine scores of the simulated mask probes (SMPs) are generally higher than those of the real masked score M12P in the investigated systems, (a) VGGFace, (b) SphereFace, (c) ArcFace, and (d) COTS

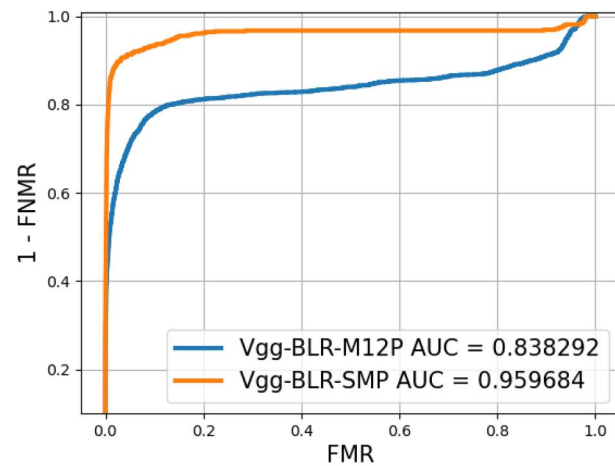
real masks in Section 6.2 do indicate the less-than optimal use of simulated masks as a replacement of real masks in such evaluations, which will be measured in more details in the next section.

6.4 | Does evaluating face recognition performance with simulated mask probes truly reflect the real mask scenario?

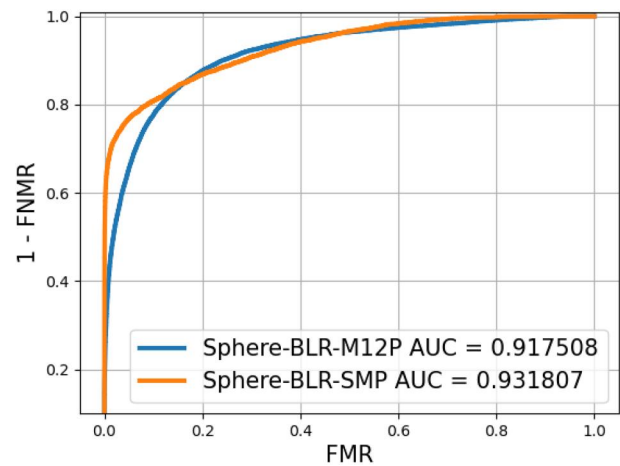
Figure 7 presents the comparison between the baselines where the reference is not masked and the probe has a real mask (BLR-M12P) genuine and imposter score distributions and the case where the probe is synthetically masked (BLR-SMP) on the four face recognition solutions considered. It is noticeable in all experimental setups that, when the probes are

synthetically masked, the genuine score distributions shift to higher values in comparison with the real mask probes setup. On the other hand, the imposter scores are not significantly shifted, this indicates an enhancement in the performance and general trust in the matcher decision, as the separability between genuine and imposter samples increase in the case of synthetically masked probes.

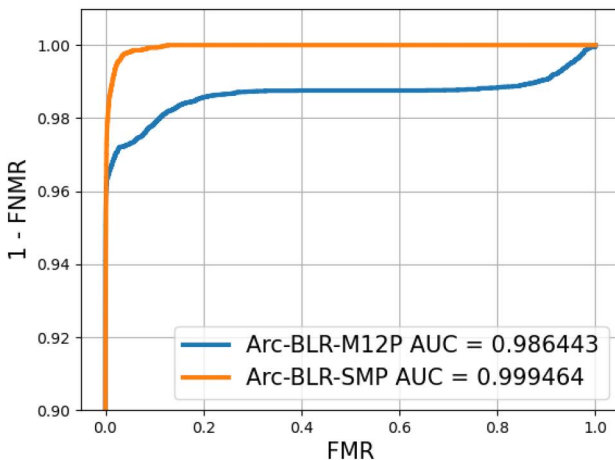
Tables 2–5 present the lower verification performance achieved by the real masks in comparison with the simulated masks. This is also indicated by the lower FDR values and lower G-mean values in the case of real masks. Similar verification performance indications can be noticed on a wider operation point in the ROC curves presented in Figure 8. A main exception to this is the COTS performance at high FMR values, where the SMPs score a higher FNMR than the real masked probes, as shown in Figure 8-d.



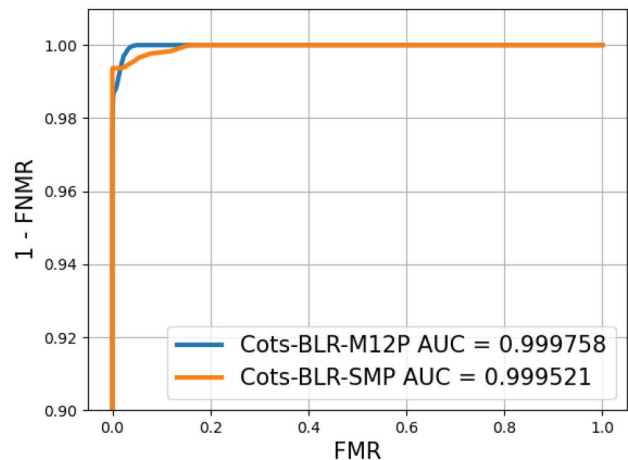
(a) VGG - real vs. simulated mask probes.



(b) SphereFace - real vs. simulated mask probes.



(c) ArcFace - real vs. simulated mask probes.



(d) Cots - real vs. simulated mask probes.

FIGURE 8 The verification performance of the four investigated systems—VGGFace (a), SphereFace (b), ArcFace (c), and COTS (d)—is presented in the form of receiver operating characteristic (ROC) curves. For each system, two curves are plotted to represent the settings that include not-masked references and real ‘masked’ face probes (BLR-M12P) as baselines and the same reference but with simulated masked probes (BLR-SMP). The area under curve (AUC) is also listed for each ROC curve. As in Tables 2–5, the higher performance with simulated probes in comparison with real masked probes is apparent in the performance of the VGGFace, ArcFace, and SphereFace, which is not the case for COTS on the full operation range

As an answer to this subsection title, synthetically generated masks do not reflect the effect of wearing a real face mask on face recognition. Variations in the simulated mask shape, colour, and texture might be seen as an option to this issue, but it would be difficult to capture the full variation scale of real masks and their interaction with other environmental factors. Such factors can range from personal wear preferences, to environment illumination, and up to customized mask designs.

In general, the effect of wearing face masks on the face recognition behaviour is apparent on all investigated systems. This renders the current face recognition solutions in strong need for evaluation under these new challenges. This also motivates building face recognition solutions that are adaptive to masked faces and pushes for realistic masked face evaluation platforms.

7 | CONCLUSION

This work presented an extensive study on the effects of mask-wearing on face recognition performance in collaborative scenarios. The main motivation behind this effort is the widespread use of face masks as a preventive measure in response to the COVID-19 pandemic. We presented a specifically collected database captured in three different sessions with and without wearing a mask. This database is augmented with additional subsets where the masks are synthetically added. We analysed the behaviour of three widely studied academic face recognition solutions in addition to one commercial solution. Our study indicated the significant effect of wearing a mask on comparison score separability between imposter and imposter comparisons in all the investigated systems, and thus the effect on verification performance. Moreover, we questioned, among other conclusions, the suitability of simulated masks in the realistic evaluation of face recognition performance when processing masked faces.

ACKNOWLEDGEMENT

This research work has been funded by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Centre for Applied Cybersecurity ATHENE.

Open Access funding enabled and organized by Projekt DEAL.

ORCID

Naser Damer  <https://orcid.org/0000-0001-7910-7895>

Fadi Boutros  <https://orcid.org/0000-0003-4516-9128>

Arjan Kuijper  <https://orcid.org/0000-0002-6413-0061>

REFERENCES

1. Opitz, M., et al.: Grid loss: detecting occluded faces. In: Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, October 11-14, 2016, Proceedings, Part III. vol. 9907 of Lecture Notes in Computer Science. pp. 386–402. The Netherlands Springer, (2016) https://doi.org/10.1007/978-3-319-46487-9_24
2. Song, L., et al.: Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, pp. 773–782. Seoul (2019). <https://doi.org/10.1109/ICCV.2019.00086>
3. Damer, N., et al.: The effect of wearing a mask on face recognition performance: An exploratory study. In: BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, vol. pp. 16–18 Gesellschaft für Informatik e.V., online (2020). <https://dl.gi.de/20.500.12116/34316>
4. Brömme, A., et al. (eds.) BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, vol. pp. 16–18. Gesellschaft für Informatik e.V., online (2020). <https://dl.gi.de/handle/20.500.12116/34315>
5. Damer, N., et al.: On the generalization of detecting face morphing attacks as anomalies: Novelty vs. outlier detection. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–5. IEEE, Tampa (2019). <https://doi.org/10.1109/BTAS46853.2019.9185995>
6. Damer, N., et al.: Realistic dreams: Cascaded enhancement of GAN-generated images with an example in face morphing attacks. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–10. IEEE, Tampa (2019). <https://doi.org/10.1109/BTAS46853.2019.9185994>
7. Damer, N., Dimitrov, K.: Practical view on face presentation attack detection. In: Wilson, R.C., Hancock, E.R., Smith, W.A.P. (eds.) Proceedings of the British Machine Vision Conference 2016, BMVC 2016, 19–22. BMVA Press, York (2016). <http://www.bmva.org/bmvc/2016/papers/paper112/index.html>
8. Damer, N., et al.: Crazyfaces: Unassisted circumvention of watchlist face identification. In: 9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, pp. 22-25. IEEE, Redondo Beach, CA, USA (2018). <https://doi.org/10.1109/BTAS.2018.8698557>
9. Damer, N., et al.: Deep learning-based face recognition and the robustness to perspective distortion. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 3445–3450. IEEE, Beijing (2018). <https://doi.org/10.1109/ICPR.2018.8545037>
10. Damer, N., Samartzidis, T., Nouak, A.: Personalized face reference from video: Key-face selection and feature-level fusion. In: Ji Q., B. Moeslund T., Hua G., Nasrollahi K. (eds.) Face and Facial Expression Recognition from Real World Videos. FFER 2014. Lecture Notes in Computer Science, vol. 8912. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-13737-7_8
11. Ge, S., et al.: Detecting masked faces in the wild with lle-cnns. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, pp. 426–434. IEEE Computer Society, Honolulu (2017). <https://doi.org/10.1109/CVPR.2017.53>
12. Wang, Z., et al.: Masked face recognition dataset and application. arXiv. <https://arxiv.org/abs/2003.09093> (2020)
13. Anwar, A., Raychowdhury, A.: Masked face recognition for secure authentication, arXiv. <https://arxiv.org/abs/2008.11104> (2020)
14. Ngan, M.L., Grother, P.J., Hanaoka, K.K.: Ongoing face recognition vendor test (frvt) part 6b: face recognition accuracy with face masks using post-covid-19 algorithms, National Institute of Standards and Technology, Gaithersburg (2020)
15. Deng, J., et al.: Additive angular margin loss for deep face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pp. 4690–4699. Long Beach (2019). June 16-20, 2019
16. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC). BMVA Press, Swansea (2015)
17. Liu, W., et al.: SphereFace: deep hypersphere embedding for face recognition. In: IEEE conference on computer vision and pattern recognition, pp. 6738–6746. CVPR 2017, Honolulu (2017). <https://doi.org/10.1109/CVPR.2017.713>
18. Neurotechnology: Neurotechnology megamatcher 12.1 sdk. HAL-Inria, https://www.neurotechnology.com/mm_sdk.html
19. Huang, G.B., Ramesh, M., Berg, T., Learned.Miller, E.: Labelled faces in the wild: A database for studying face recognition in unconstrained environments, pp. 07–49. University of Massachusetts, Amherst (2007)

20. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity CVPR 2011, pp. 529–534. IEEE (2011). <https://doi.org/10.1109/CVPR.2011.5995566>
21. Yi, D., et al.: Learning face representation from scratch. CoRR (2014). [abs/1411.7923](https://arxiv.org/abs/1411.7923). Available from: <http://arxiv.org/abs/1411.7923>
22. He, K., et al.: Deep residual learning for image recognition IEEE conference on computer vVision and pattern recognition, CVPR 2016, Las Vegas, pp. 770–778. IEEE Computer Society, NV, USA (2016). June 27-30, 2016 Available from. <https://doi.org/10.1109/CVPR.2016.90>
23. Guo, Y., et al.: Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In: Proceedings, Part III Computer Vision - ECCV 2016 - 14th European conference, pp. 87–102. Amsterdam (2016). Available from: https://doi.org/10.1007/978-3-319-46487-9_6
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International conference on learning Representations, ICLR 2015. Conference Track Proceedings, San Diego (May 7-9, 20152015). <http://arxiv.org/abs/1409.1556>
25. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: unified embedding for face recognition and clustering. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, pp. 815–823. IEEE Computer Society, Boston (2015). <https://doi.org/10.1109/CVPR.2015.7298682>
26. Patrick, G., Ngan Mei, H.K.: Ongoing face recognition vendor test (frvt). NIST Interagency Report. National Institute of Standards and Technology, Gaithersburg (2020)
27. Zhang, K., et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal. Process. Lett. 23(10), 1499–1503 (2016)
28. Mansfield, A.: Information technology–biometric performance testing and reporting–part 1: Principles and framework. pp. 19795–1. ISO/IEC (2006)
29. Poh, N., Bengio, S.: A study of the effects of score normalisation prior to fusion in biometric authentication tasks. IDIAP (2004)
30. Damer, N., Opel, A., Nouak, A.: Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution In: 2014 22nd European Signal Processing Conference (EUSIPCO), pp. 1382–1386. IEEE, Lisbon (2014).
31. King, D.E.: Dlib-ml: A machine learning toolkit. J. Mach. Learn. Res. 10, 1755–1758 (2009). <https://dl.acm.org/citation.cfm?id=1755843>
32. Boutros, F.: Simulated mask implementation in python. <https://github.com/fdbtr/MFR> (2021)
33. Boutros, F., et al.: Unmasking face embeddings by self-restrained triplet loss for accurate masked face recognition. arXiv. <https://arxiv.org/abs/2103.01716> (2021)

How to cite this article: Damer, N., et al.: Extended evaluation of the effect of real and simulated masks on face recognition performance. IET Biom. 10(5), 548–561 (2021). <https://doi.org/10.1049/bme2.12044>