Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Pattern Recognition 123 (2022) 108398

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Real masks and spoof faces: On the masked face presentation attack detection

Meiling Fang[a,b,*], Naser Damer[a,b], Florian Kirchbuchner[a], Arjan Kuijper[a,b]

[a] Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany
[b] Mathematical and Applied Visual Computing, TU Darmstadt, Darmstadt, Germany

ARTICLE INFO

Article history:
Received 16 February 2021
Revised 22 October 2021
Accepted 23 October 2021
Available online 26 October 2021

Keywords:
Face presentation attack detection
COVID-19
Masked face
Face recognition
Biometric security

ABSTRACT

Face masks have become one of the main methods for reducing the transmission of COVID-19. This makes face recognition (FR) a challenging task because masks hide several discriminative features of faces. Moreover, face presentation attack detection (PAD) is crucial to ensure the security of FR systems. In contrast to the growing number of masked FR studies, the impact of face masked attacks on PAD has not been explored. Therefore, we present novel attacks with real face masks placed on presentations and attacks with subjects wearing masks to reflect the current real-world situation. Furthermore, this study investigates the effect of masked attacks on PAD performance by using seven state-of-the-art PAD algorithms under different experimental settings. We also evaluate the vulnerability of FR systems to masked attacks. The experiments show that real masked attacks pose a serious threat to the operation and security of FR systems.

1. Introduction

Since the SARS-CoV-2 coronavirus outbreak and its rapid global spread, wearing a mask has become one of the most efficient ways to protect and prevent getting infected with COVID-19. However, for identity checks in crowded scenarios such as at airports, removing the mask for face recognition (FR) increases the chance of infection. Wearing masks in public might be an essential health measure and a new norm even after the COVID-19 pandemic as most countries support the use of masks to minimize the spread of the virus. As a result, researchers have shown an increased interest in the effect of face masks on the performance of FR verification [1–3]. The results of their studies have shown that pre-COVID-19 FR algorithms suffer performance degradation owing to the masked faces. However, attacks compromising the security and vulnerability of FR systems for subjects wearing face masks have so far been overlooked. In this study, security refer to the presentation attacks (PAs). Attackers can use PAs to spoof FR systems by impersonating someone or obfuscating their identity. Common PAs include printed photos/images, replayed videos and 3D masks [4,5]. Driven by the ongoing COVID-19 pandemic, presentation attack detection (PAD) [6] has encountered several understudied challenges when facing masked faces. Current face PAD databases [7–9] only contain printed images or replayed videos in which subjects were not wearing face masks. Therefore, there is uncertainty about the relationship between the performance of PAD techniques and PAs with face masks. Moreover, the vulnerability of FR systems to masked attacks remains unclear. To overcome such gaps, researchers require well-studied masked PAs.

In this study, we design and collect three types of attacks based on masked and unmasked face images collected realistically and collaboratively [2,3]. The bona fide samples were divided into categories of BM0 (subjects wearing no masks) and BM1 (subjects wearing masks). AM0 data are unmasked print/replay attacks, which are commonly used data in most current PAD databases. AM1 data include print/replay attacks, where live subjects wore masks. In addition, we provide a novel partial attack type, called AM2, where a real medical mask is placed on printed photos or replayed videos to simulate the subject wearing a mask. This is motivated by our assumption that AM2 might be a challenging attack as it contains both bona fide and attack presentations that may confuse PAD and/or FR systems. The data samples are presented in Fig. 1. The main contributions in this study are:

- The novel Collaborative Real Mask Attack Database (CRMA) is presented. Three types of PAs, called AM0 (unmasked face PA), AM1 (masked face PA), and AM2 (unmasked face PA with a real masked placed on the PA) (as shown in Fig. 1), were created for both print and replay presentation attack instruments (PAIs). To create such attacks, three electronic tablets with high-

* Corresponding author.
E-mail address: meiling.fang@igd.fraunhofer.de (M. Fang).

https://doi.org/10.1016/j.patcog.2021.108398
0031-3203/

Fig. 1. Example bona fide and attack samples in the CRMA database. Based on the presence of face masks, bona fides are grouped into BM0 (without mask) and BM1 (with mask) categories. The novel attacks are grouped into AM0 (spoof face without mask), AM1 (spoof face with mask), AM2 (spoof faces covered by real masks).

resolution and three capture scales are used. Additionally, we designed three experimental protocols to explore the effect of masked attacks on PAD performance.

- Extensive experiments are conducted to explore the effect of bona fide samples, masked faces attacks, and real masks (on spoof faces) on the face PAD behavior. To support the comprehensive evaluation, seven PAD algorithms comprising of texture-based, deep-learning-based, and hybrid methods were selected to evaluate the performance and generalizability in intra- and cross-database scenarios under three mask-related protocols. Both quantitative and qualitative analyses revealed that masked bona fides and PAs dramatically decreased the performance of PAD algorithms. Moreover, deep-learning-based methods perform worse on real mask attacks than mask-face attacks in most cases.
- An in-depth vulnerability analysis of FR systems is presented. We evaluated three deep-learning-based FR techniques for three types of PAs. The experimental results indicate that these three FR networks exhibit significantly higher vulnerabilities to the real mask attacks than masked face attacks.

We provide a brief review of relevant works in Section 2. Then, our novel CRMA database is described in detail in Section 3. The face PAD algorithms and FR systems used in this study are introduced in Section 4. Section 5 introduces the three PAD protocols and PAD evaluation metrics, and then discusses the PAD results. Section 6 describes the three FR experimental settings, used FR metrics, and analyzed the vulnerability of FR systems. Finally, conclusions are presented in Section 7.

2. Related work

This section reviews the most relevant prior works from three perspectives: face PAD databases, face PAD methods, and FR and vulnerability analysis. At the end of each part, the difference between our work and prior work is pointed out.

Face PAD Databases: Data resources have become especially important ever since the advent of deep learning, because machine-learning-based algorithms have the risk of underfitting or overfitting on limited data. Given the significance of good-quality databases, several face PAD databases have been released, such as NUAA [10], CASIA-FAS [11], Replay-Attack [12], MSU-MFSD [13], OULU-NPU [7], and SiW [8], all consisting of 2D print/replay attacks. In addition, SiW-M [9] and CelebA-Spoof [14] databases provide multiple types of attacks such as makeup, 3D mask, or paper cut. Moreover, some multimodal databases are publicly available: 3DMAD [15], Mssproof [16], CASIA-SURF [17], and CSMAD [18].

These databases undoubtedly contribute to the significant progress of PAD research. For example, the CeleA-Spoof database comprises images from various environments and illuminations with rich annotations to reflect real scenes. However, these databases also have weaknesses: 1) the multimodal databases have high hardware requirements and cannot be widely used in daily life; 2) some databases such as CASIA-MFS [11] and MSU-MFS [13] cannot satisfy the current needs because of the lower quality of the outdated acquisition sensors; 3) Oulu-NPU [7], SiW [8], SiW-M [9], and CelebA-Spoof [14] are relatively up-to-date, but they do not consider PAs with real face masks to fit the current COVID-19 pandemic. Hence, we collect the CRMA database to fill the gaps in these databases in the context of the ongoing COVID-19 pandemic; furthermore, we ensure the the database is generalizable and compatible with real scenarios. The CRMA database can be used to better analyze the effect of a real mask on PAD performance and the vulnerability of FR systems for novel attacks, such as placing a real mask on an attack presentation. Detailed information related to the databases mentioned above is presented in Table 1).

Face PAD Methods: In recent years, there has been an increasing number of studies in the field of face PAD [19–21]. These studies can be broadly grouped into three categories: texture-based methods, deep-learning-based methods, and hybrid methods. Texture features, such as local binary pattern (LBP) [22], project the faces to a low-dimensional embeddings. Määttä *et al.*[23] proposed an approach using multi-scale LBP to encode the micro-texture patterns into an enhanced feature histogram for face PAD. The resulting histograms were then fed to a support vector machine (SVM) classifier to determine whether a sample is a bona fide or attack. The LBP features extracted from different color spaces [24] were further proposed to utilize chrominance information. They achieved competitive results on Replay-Attack [12] (equal error rate (EER) value of 0.4%) and CASIA-FAS [11] (EER value of 6.2%) databases. Furthermore, Boulkenafet *et al.*[25] organized a face PAD competition based on the OULU-NPU database and compared 13 algorithms provided by participating teams and one color-LBP-based method (referred to as baseline in [25]). In this competition, the GRADIANT algorithm fused multiple information, that is, color, texture, and motion. The GRADIANT achieved compet-

Table 1
The summary of face PAD databases, including our CRMA database information for brief comparison. It should be noted that our CRMA database is the only database containing subjects wearing face masks and real face mask attacks. The details of our CRMA database are presented in Section 3.

Database	Year	# Subjects	# Data (BF/attack)	Capture devices (BF/attack)	Display devices	Modality	Attack type
NUAA [10]	2010	15	5105/7509 (I)	Webcame	-	RGB	1 Print
CASIA-FAS [11]	2012	50	150/450 (V)	Two USB cameras, Sony NEX-5	iPad	RGB	1 Print, 1 Replay
Replay-Attack [12]	2012	50	200/100 (V)	MacBook 13 / iPhone 3GS, Cannon SX150	iPhone 3GS, iPad	RGB	1 Print, 2 Replay
3DMAD [15]	2013	17	170/85 (V)	Microsoft Kinect	-	RGB/Depth	1 3D Mask
Msspoof [16]	2015	21	1,680/3,024 (I)	uEye camera	-	RGB/IR	1 Print
MSU-MFSD [13]	2015	35	110/330 (V)	MacBook Air, Google Nexus 5 / Cannon 550D, iPhone 5s	iPad Air, iPhone 5s	RGB	1 Print, 2 Replay
Oulu-NPU [7]	2017	55	1,980/3,960 (V)	6 smartphones	Dell 1905FP, Macbook Retina	RGB	2 Print, 2 Replay
SiW [8]	2018	165	1,320/3,300 (V)	Cannon EOS T6, Logitech C920 webcam	iPad Pro, iPhone 7, Galaxy S8, Asus MB 168 B	RGB	2 Print, 4 Replay
CASIA-SURF [17]	2018	1000	18000/3000 (I)	RealSense camera	-	RGB/IR/Depth	5 Papercut
CSMAD [18]	2018	14	88/160 (V)	RealSense, Compact Pro, Nikon P520	-	RGB/IR/Depth/LWIR	1 silicone mask
SiW-M [9]	2019	493	660/1630 (V)	Logitech C920, Cannon EOS T6	-	RGB	1 Print, 1 Replay, 5 3D mask, 3 Makeup, 3 Partial
Celeb-Spoof [14]	2020	10,177	202,559/475,408 (I)	Various cameras/ 20 smartphones, 2 webcams, 2 tablets	PC, phones, tablets,	RGB	3 Print, 3 Replay, 1 3D mask, 3 Paper Cut
CRMA	2021	47	423/12,690 (V)	Webcams/iPad Pro, Galaxy Tab S6, Surface Pro-6	iPad Pro, Galaxy Tab S6, Surface Pro-6	RGB	1 Print, 3 Replay, **1 Real mask**

itive results in the four evaluation protocols. In addition to the texture-based GRADIANT approach, deep-learning-based method (MixFASNet) or hybrid method (CPqD) also achieved lower error rates in all experimental protocols. CPqD fused the results from the fine-tuned Inception-v3 network and the color-LBP-based method (referred to as the baseline in [25]). Consequently, we chose to re-implement the color-LBP and CPqD methods in this study (details in Section 4.1), while the GRADIANT and MixedFASNet are discarded in our work because they do not provide sufficient details for re-implementation. Deep-learning-based methods have been pushing the frontier of face PAD research and have shown remarkable improvements in PAD performance. Lucena *et al.*[26] presented an approach called FASNet in which a pre-trained VGG16 is fine-tuned by replacing the last fully connected layer. The FASNet network achieved excellent performance on 3DMAD [15] and Replay-Attack databases [12]. Recently, George *et al.*[27] proposed training a network with pixel-wise binary supervision on feature maps to exploit information from different patches. DeepPixBis [27] outperformed the state-of-the-art algorithms in Protocol-1 of the OULU-NPU database (0.42% ACER) but also achieved significantly better results than traditional texture-based approaches in the cross-database scenario. Considering the popularity of PAD techniques and the ease of implementation, we also chose FASNet and DeepPixBis (details in Section 4.1) to study the effect of the real mask and masked face attacks on the PAD performance.

Face Recognition and Vulnerability Analysis: As one of the most popular modalities, the face has received increasing attention in authentication/security processes, such as smartphone face unlocking and automatic border control (ABC). Moreover, FR techniques [28–30] have achieved significant performance improvements, and many personal electronic products have deployed FR technology. However, the ongoing COVID-19 pandemic brings a new challenge related to the behavior of collaborative recognition techniques when dealing with masked faces. Collaborative data collection refers to a subject actively attending to use the FR systems, such as unlocking personal devices or using an ABC gate, in contrast to uncollaborative capture scenario where the user does not intentionally use the FR service, such as in the case of surveillance. The National Institute of Standards and Technology (NIST) [1] provided a preliminary study that evaluated the performance of 89 commercial FR algorithms developed before the COVID-19 pandemic. Their results indicated that digitally applied face masks with photos decreased the recognition accuracy; for example, even the best of the 89 algorithms had error rates between 5% and 50%. It is worth noting that the masks used in the experiments were synthetically created. Damer *et al.*[2,3] presented a real mask database to simulate a realistically variant collaborative face capture scenario. Each participant was asked to simulate a login scenario by actively looking toward a capture device, such as a static webcam or a mobile phone. Our attack samples were created and collected based on the masked face data, which refers to the bona fide samples in the PAD case (as described in Section 3). They also explored the effect of wearing a mask on FR performance and concluded that face masks significantly reduce the accuracy of algorithms. Mohammadi *et al.*[31] provided empirical evidence to support the claim that the CNN-based FR methods are extremely vulnerable to 2D PAs. Subsequently, Bhattacharjee *et al.*[18] presented the first FR-vulnerability study on 3D PAs. The experiments also clearly showed that CNN-based FR methods are vulnerable to custom 3D mask PAs. However, the vulnerability of FR systems to PAs with face masks has not been investigated. Therefore, in this study, we selected three CNN-based FR algorithms for further FR-vulnerability analysis on masked face attacks: the state-of-the-art ArcFace [28], SphereFace [29], and VGGFace [30]. These algorithms are discussed in more detail in Section 4.2.

3. The collaborative real mask attack database (CRMA)

Our proposed CRMA database[1] and can serve as a supplement to the databases in Table 1, and because of the COVID-19 pandemic, it can better reflect the possible issues facing real-world PAD performance. The CRMA database includes 1) both unmasked (BM0) and masked (BM1) bona fide samples collected in a realistic scenario [2,3], 2) conventional replay and print PAs created from faces not wearing a mask (AM0), 3) replay and printed PAs created from masked face images (AM1), and 4) novel PAs where the PAs of unmasked faces are covered (partially) with real masks (AM2), as shown in Fig. 1. Damer et al. [2,3] collected data to investigate the effect of wearing a mask on face verification performance. For PAD, such data are considered bona fide. The data presented in this study build on an extended version of the data introduced in [2,3], by creating and capturing different types of PAs based on the bona fide data captured in [2,3]. As a result, the bona fide data in this work are an extended version of the one introduced in [2,3] and the attack data presented here are completely novel and have not been previously studied.

Fig. 3 introduces the general statistical information of the CRMA database. This database contains 62% males and 38% females. The attack AM0, AM1, and AM2 ratios are 30%, 60%, and 10%, respectively, as will be described later in this section. Additionally, we count the frequency of the proportion of the face size in the video. The histogram shows that the proportion of the face areas in the videos is mostly between 5% and 30%. This section first describes the bona fide samples provided by [2,3], and then introduces our process of attack sample creation and collection.

3.1. Collection of bona fide samples

To explore the FR performance on masked faces, Damer *et al.*[2,3] recently presented a database where the subjects wearing face masks.

This database simulates a collaborative environment in which participants collect videos by actively looking towards the capture device. During this process, the eyeglasses were removed when the frame was considered very thick following the International Civil Aviation Organization (ICAO) standard [32]. The videos were captured by the participants at their residences while working from home. Therefore, the types of face masks, capture devices, illumination, and background were varied. For PAD, these videos are classified as bona fides and will be used later to create attack samples.

The final version [3] of this database contains 47 participants. Each subject recorded a total of nine videos over three days with three different scenarios for each day. In contrast to the study by Damer *et al.*[2], which examined the effects of both face masks and illumination variations, we focused only on the impact of face masks on PAD performance. Hence, in our study, the bona fide videos are divided into two categories: a face without a mask on is denoted as BM0 (three videos per subject), and a face with a mask is marked as BM1 (six videos per subject) (as shown in the right column of Fig. 1).

3.2. Creation of the presentation attacks

Most FR databases tried to collect data under various harsh conditions, such as poor lighting, strong occlusion, or low resolution. Such databases attempted to reproduce what might happen in a real-world scenario when a legitimate user obtains authorization [33]. In contrast, attackers use highly sophisticated artifacts, such as high-resolution images or videos, to maximize the success rate when impersonating someone. For this reason, we first collect the PAs in a windowless room where all lights are on. Second, three high-resolution electronic tablets were used in the acquisition process: 1) iPad Pro-(10.5-inch) with the display resolution of 2224×1668 pixels, 2) Samsung Galaxy Tab S6 with the display resolution of 2560×1600 pixels, 3) Microsoft Surface Pro-6 with the display resolution of 2736×1824 pixels. In the process of collecting data, the capture devices and displayed images/tablets were stationary. The videos were captured with a resolution of 1920×1080. In addition, each video had a minimum length of 5 seconds, and the frame rate was 30 fps. This work focuses on the two common PAIs, print and replay attacks, due to their ease of creation and low cost. The attack data in each PAI (see the samples in Fig. 1) are divided into three types: 1) the spoof face with no face mask (AM0), 2) the spoof face with a mask on (AM1), and 3) the spoof face with no face mask, but a real mask was placed on it to simulate a participant wearing a mask (AM2). However, the size of the face area in each video is slightly inconsistent because the videos were recorded by the participants themselves. To reproduce the appearance of wearing a mask in the real world, we cropped five face masks to fit most of the faces (see Fig. 3). The five masks are three small blue surgical masks, one slighter bigger white face mask, and one uncropped mask. When placing the mask, we select a suitable mask according to the size of the face in the printed image or video, aiming to cover the nose to the chin area and the cheeks without exceeding. The details of each PAI are as followings:

Print image attack: In print PAI, an attacker tries to fool the FR system using a printed photo. Considering the instability of the face during the first second, such as the participant pressing the recording button or adjusting the sitting position, the 35^{th} frame of each bona fide video was printed out as an attack artifact. Therefore, we obtained nine photos per subject. The three tablets mentioned above were used to capture the photos. Furthermore, to increase the diversity and variety of the data, each tablet captured three videos for a photo with three scales (see examples in Fig. 2). The captured videos using the first scale contained all areas (100%) of the photos, the second scale consisted of most areas (80%) of the original photos, and the third scale focused on the face area (60%) as much as possible. In addition to collecting attack data solely from printed images, we also collected data from real face masks overlaid on photos (i.e., the previously defined AM2). Theoretically, real masks will reduce the region of artificial features and increase the complexity and mixture of the features in the collected attack data. Eventually, 90 print attack videos were generated for each subject, that is, a total of 4230 videos for 47 subjects in print PAI.

Replay video attack: In replay PAI, an attacker tries to obtain the authentication by replaying a video. The three common points of the collection process between print and replay PAI are the use of three tablets, the use of three scales, and the process of AM2 data creation, respectively. The difference is that these tablets were also used for capturing displays of videos (see examples in Fig. 2). While one tablet was replaying the video, the other two tablets were used to capture the data. As a result, each subject corresponded to 180 replay attack videos (162 videos of the AM0 and AM1 groups, 18 videos of AM2.), i.e., there were a total of 8460 videos in this attack subset.

4. Experimental algorithms

This section first describes the adopted face PAD algorithms for the investigation of masked face attacks. Subsequently, three FR al-

[1] The CRMA database is not publicly available because of privacy regulations. However, the database will be (1) available for assisted in-house research use by collaborators and partners in the research community; (2) bending the legal authorization by the data collection institute, the data will be submitted to be included on the Open Science BEAT platform (www.beat-eu.org).

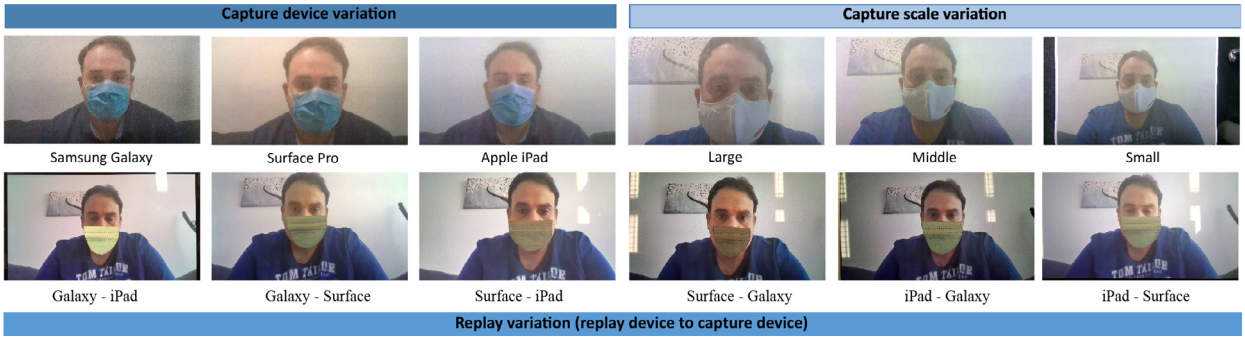


Fig. 2. Different capture variations in the CRMA database. The top left shows the videos captured by different devices. The top right shows the different capture scales. The bottom shows the six cross-device types of replay attack settings.

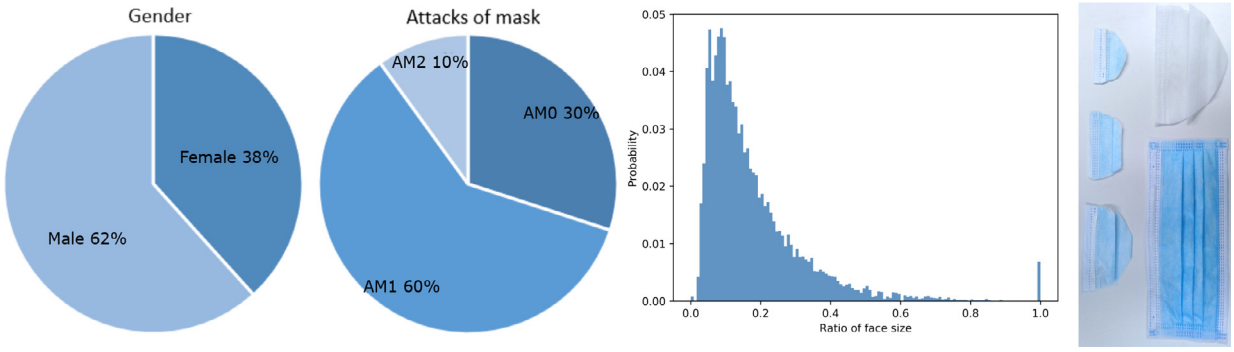


Fig. 3. The statistics of the subjects and the used mask shapes for creating AM2 samples in the CRMA database. From left to right: gender, mask types of attacks (AM0, AM1, AM2), the histogram shows the probability distribution of the face size ratio and the applied mask shapes.

gorithms were introduced for further vulnerability analysis. In both PAD and FR experiments, the widely used multi-task cascaded convolutional networks (MTCNN) [34] technique was adopted to detect and crop the face.

4.1. Face PAD algorithms

A competition [25] was carried out in 2017 to evaluate and compare the generalization performance of face PAD techniques under real-world variations. In this competition [25], there were 14 participating teams together with organizers that contributed to several state-of-the-art approaches. We chose two methods ((as previously discussed in Section 2)), the LBP-based method (referred to as the baseline in [25]), and hybrid CPqD, and included additional solutions. We re-implemented a total of seven face PAD algorithms in this study, which can be categorized into three groups: hand-crafted features, deep-learning features, and hybrid features. For further cross-database evaluation scenarios, we used three publicly available databases, mainly involving 2D PAs (details in Section 2): CASIA-FAS [11], MSU-MFS [13], and OULU-NPU [7] in the competition. A brief description of the adopted methods is provided below:

- **LBP:** The LBP method is referred to as baseline method in [25] provided by the competition organizers that utilized the color texture technique. The face in a frame is first detected, cropped, and normalized to a size of 64×64 pixels. Second, an RGB face was converted into HSV and YCbCr color spaces. Third, the LBP features were extracted from each channel. The obtained six LBP features are then concatenated into one feature vector to feed into a softmax classifier. The final prediction score for each video was computed by averaging the output scores of all the frames.
- **CPqD:** The CPqD is based on the Inception-v3 network [35] and the above LBP method. The last layer of the pre-trained Inception-v3 model was replaced by a fully connected layer and a sigmoid activation function. The faces in the RGB frames are detected, cropped, and normalized to 299×299 pixels. These face images were utilized as inputs to fine-tune the Inception-v3 model. The model with the lowest EER on the development set among all 10 training epochs was selected. A single score for each video was obtained by averaging the output scores of all frames. To further improve the performance, the final score for each video was computed by fusing the score achieved by the Inception-v3 model and the score obtained by the LBP method.
- **Inception$_{FT}$** and **Inception$_{TFS}$**: Since the CPqD uses the Inception-v3 [35] network as the basic architecture, we also report the results of fine-tuned Inception-v3 model, named Inception$_{FT}$. In addition to the fine-tuned model, we trained the Inception-v3 model from scratch for performance comparison, named Inception$_{TFS}$. In the training phase, the binary cross-entropy loss function and Adam optimizer with a learning rate of 10^{-5} were used. The output scores of the frames were averaged to obtain a final prediction decision for each video.
- **FASNet$_{FT}$** and **FASNet$_{TFS}$**: FASNet [26] used transfer learning from pre-trained VGG16 model [36] for face PAD. They used a pre-trained VGG16 model as a feature extractor and modified the last fully connected layer. The newly added fully connected layers with a sigmoid function were then fine-tuned for the PAD task. This fine-tuned FASNet is referred to as FASNet$_{FT}$, similar to the Inception-v3 network methods, and we also train FASNet from scratch with the name FASNet$_{TFS}$. The input images are the detected, cropped, and normalized RGB face frames with a size of 224×224 pixels. The Adam optimizer with a learning rate of 10^{-4} was used for training, as defined in [26]. Data augmentation techniques and class weights are utilized to deal with imbalanced data problems. To further reduce overfitting, an early stop technique with a patience of 5 and maxi-

mum epochs of 30 was used. The resulting scores were averaged to obtain the final score for each video.

- **DeepPixBis:** George *et al.*[27] proposed a densely connected network framework for face PAD with binary and deep pixel-wise supervision. This framework is based on DenseNet architecture [37]. Two dense blocks and two transition blocks with a fully connected layer with sigmoid activation produce a binary output. We used the same data augmentation technique (horizontal flip, random jitter in brightness, contrast, and saturation) and the same hyper-parameters (Adam optimizer with a learning rate of 10^{-4} and weight decay of 10^{-5}) as defined in [27] for the training. In addition to data augmentation, we applied the class weight and an early stopping technique to avoid overfitting. The final score for each video was computed by averaging the frame scores.

4.2. Face recognition algorithms

For FR systems, trained CNNs are typically used as feature extractors. The feature vector extracted from a specific layer of an off-the-shelf CNN was used as the template to represent the corresponding input face image. Then, the resulting templates were compared with each other using similarity measures. To provide a vulnerability analysis of the FR systems to our novel masked attacks, we adapted the following three FR algorithms:

- **ArcFace:** ArcFace [28] introduced an additive angular margin loss function to obtain highly discriminative features for FR. We chose this algorithm because ArcFace consistently outperformed state-of-the-art methods. ArcFace achieved 99.83% on Labeled Faces in the Wild (LFW) [33] and 98.02% on YouTube Faces (YTF) [38] dataset. The pre-trained ArcFace model[2] in our study was based on the ResNet-100 [39] architecture and trained on the MS-Celeb-1M [40] dataset (MS1M-v2). The output template is a 512-dimension feature vector extracted from the '*fc1*' layer of ArcFace.
- **SphereFace:** Liu *et al.*[29] proposed a deep hypersphere embedding approach (SphereFace) for FR task. SphereFace [29] utilized the angular softmax loss for CNNs to learn angularly discriminative features. This method also achieved competitive performance on LFW [33] (accuracy of 99.42%) and YTF [38] datasets (95.00%). We extract the face representation with 512-dimension from a pre-trained 20-layer SphereFace model.[3]
- **VGGFace2:** The first version of VGGFace is based on 16-layer VGG [36] network, while the second version of VGGFace (VGGFace2) [30] adopt ResNet-50 [39] as the backbone architecture. In this work, we use the second version that a ResNet-50 network trained on VGGFace2 dataset [30][4] for extracting the 512-dimension templates.

The vulnerability of each FR system to attacks was analyzed based on three scenarios. Regardless of the scenario, the references are scenarios-specific bona fide videos captured on the first day, while bona fide videos from the second and third days or attack videos were selected as probes. The three cases, including the division of scenario-specific references and probes, are described with the results in detail in Section 6. Once the references for the face images are obtained, we use the Cosine-similarity as recommended in [28–30] to compute the similarity scores between references and probes.

[2] The official ArcFace model: https://github.com/deepinsight/insightface.
[3] The official SphereFace model: https://github.com/wy1iu/sphereface.
[4] The VGGFace2 model: https://github.com/WeidiXie/Keras-VGGFace2-ResNet50.

5. Analysis of PAD performance

This section first describes three protocols designed to investigate the effect of masked attacks on PAD performance under different training settings. Second, the PAD evaluation metrics used were introduced for further analysis. Third, quantitative results were reported according to different PAD protocols. Finally, qualitative analysis and visualization are presented and discussed.

5.1. PAD evaluation protocols

5.1.1. PAD protocols for the CRMA database

In this study, three protocols are provided to study the impact of masks on the performance of PAD solutions under different training settings. Other factors, such as various devices, illumination, and capture scales, are outside the scope of this study. These three protocols try to answer three questions separately: 1) Does the PAD algorithm trained on unmasked data generalize well on the masked bona fides and attacks, that is, can the previously trained model be adapted to the present-day situation? 2) Does the PAD algorithm designed before the COVID-19 pandemic still work efficiently if it is trained on additional masked data? 3) Will a network that has learned masked face attacks be confused by real masks that obscure the spoof face? Hence, we split 47 subjects in the CRMA database into three subject-disjoint sets: the training set (19 subjects), the development set (10 subjects), and the testing set (18 subjects). Gender was balanced as much as possible between the three sets. Table 2 provides more information about three protocols. A detailed description of three protocols is as follows:

Protocol-1 (P1): This protocol demonstrates the generalization performance of the PAD solutions trained on unmasked data. The training and development sets contain only videos of subjects without masks (such as data in most current PAD databases). The trained model was then tested on the data using face masks. More specifically, only BM0 and AM0 data were used for training, while BM1, AM1, and AM2 were considered unknown mask data.

Protocol-2 (P2): In contrast to protocol-1, which focuses on generalizability on unseen mask data, the second protocol is designed to evaluate the performance of PAD algorithms when masked data has been learned in the training phase. In this protocol, the training, development, and testing sets include masked and unmasked bona fides (BM0, BM1), masked and unmasked attacks (AM0, AM1), and spoof faces with real masks (AM2).

Protocol-3 (P3): Until now, the effect of AM2 on PAD performance is still unclear. AM2 is a special attack type that a real face mask is placed on spoof faces, which means it contains only partial artifacts (i.e., unmasked face spoofing region) compared to AM1, which carries entire artifacts (i.e., spoofed face and mask). Therefore, this protocol attempts to answer the following question: If the network has learned the masked attacks AM1, can this trained model not be confused by a real mask and perform similarly on the attack covered by a real mask AM2? Consequently, the training and development sets include bona fides BM0 and BM1, and attacks AM0 and AM1, while AM2 is an unknown attack in the testing set.

Because data in the CRMA are video sequences and the number of videos between bona fide and attack classes are imbalanced, we sampled 60 frames from a bona fide video and five frames from an attack video to reduce data bias. In addition to different frame sampling, we also adapt the class weights inversely proportional to the class frequencies to reduce overfitting in the training phase. In the test phase, a final classification decision was determined by averaging the prediction scores of all sampled frames.

Table 2
The detailed information of three protocols for exploration of the possible effect of face masks. The bona fide is denoted as BF. The test data is the same in the three protocols, while the types of training and development data are different.

Protocol	Set	Subjects	Types of masks	# BF videos	# Attack videos
P1	Train	1–19	BM0, AM0	57	1569
	Dev	20–29	BM0, AM0	30	810
	Test	30–47	BM0, BM1, AM0, AM1, AM2	162	4860
P2	Train	1–19	BM0, BM1, AM0, AM1, AM2	171	5130
	Dev	20–29	BM0, BM1, AM0, AM1, AM2	90	270
	Test	30–47	BM0, BM1, AM0, AM1, AM2	162	4860
P3	Train	1–19	BM0, BM1, AM0, AM1	171	4617
	Dev	20–29	BM0, BM1, AM0, AM1	90	2430
	Test	30–47	BM0, BM1, AM0, AM1, AM2	162	4860

Table 3
The PAD performance of different PAD solutions in three protocols (as described in Section 5.1). The bold number in each protocol and each method refers to the highest BPCER on BM0 and BM1 data and the highest APCER value between AM0, AM1, and AM2 in the two PAIs, respectively. The higher BPCER values for BM1 (in comparison to BM0) indicate that subjects wearing masks tend to be classified falsely as attacks.

Protocol	Method	Threshold @ BPCER 10% in dev set							
		BPCER (%)		APCER (print) (%)			APCER (replay) (%)		
		BM0	BM1	AM0	AM1	AM2	AM0	AM1	AM2
P1	LBP	1.75	**4.39**	**80.12**	72.61	71.93	**74.95**	67.76	73.98
	$\text{Inception}_{\text{FT}}$	19.30	**84.21**	**10.33**	3.80	2.92	**27.19**	5.81	0.88
	CPqD	7.02	**47.37**	**18.52**	7.80	15.79	**31.77**	11.19	10.23
	$\text{FASNet}_{\text{FT}}$	12.28	**56.14**	**7.02**	1.36	2.92	**20.37**	12.21	9.65
	$\text{Inception}_{\text{TFS}}$	7.04	**48.25**	1.36	0.00	**1.75**	**7.50**	0.34	7.02
	$\text{FASNet}_{\text{TFS}}$	7.02	**29.82**	1.95	0.49	**15.20**	**8.09**	4.64	7.89
	DeepPixBis	19.30	**28.95**	1.56	1.56	**5.85**	3.61	4.05	**6.43**
P2	LBP	**26.32**	11.40	31.38	**44.44**	36.84	**36.74**	34.39	28.95
	$\text{Inception}_{\text{FT}}$	1.75	**7.02**	**35.28**	30.80	11.70	**54.09**	52.17	10.23
	CPqD	3.51	**7.89**	27.49	**30.41**	16.37	**46.20**	44.50	10.23
	$\text{FASNet}_{\text{FT}}$	1.75	**17.54**	10.72	**12.77**	5.85	**30.60**	28.09	3.80
	$\text{Inception}_{\text{TFS}}$	8.77	**18.42**	0.78	1.56	**2.34**	3.90	**5.23**	2.63
	$\text{FASNet}_{\text{TFS}}$	14.04	**29.82**	4.09	3.41	**9.36**	**4.69**	2.88	3.80
	DeepPixBis	**29.82**	24.56	0.78	0.19	**1.75**	0.10	**1.86**	0.88
P3	LBP	**22.81**	9.65	35.28	**48.15**	47.95	38.50	36.79	**42.40**
	$\text{Inception}_{\text{FT}}$	1.75	**8.77**	24.17	**24.37**	11.70	46.69	**47.14**	14.04
	CPqD	**7.02**	**7.02**	20.66	**28.95**	21.64	41.23	**41.52**	17.84
	$\text{FASNet}_{\text{FT}}$	5.26	**21.93**	14.04	9.94	**26.71**	**22.62**	19.88	20.47
	$\text{Inception}_{\text{TFS}}$	21.05	**21.93**	0.19	0.00	**1.17**	1.56	2.34	**4.97**
	$\text{FASNet}_{\text{TFS}}$	22.81	**34.21**	0.39	0.29	**2.34**	3.41	2.20	**6.43**
	DeepPixBis	17.54	**24.56**	0.78	0.68	**2.92**	0.88	1.91	**6.43**

5.1.2. PAD protocols for cross-database scenarios

In addition to the intra-database scenario on our CRMA database, we also perform cross-database experiments to explore the generalizability of these PAD algorithms on masked data. Because the PAIs in the CRMA database are print and replay attacks, we selected three popular publicly available databases containing the same PAIs: CASIA-MFS [11], MSU-MFS [13], and OULU-NPU [7] to demonstrate the evaluation. We conducted two cross-database experiments. In the first cross-database scenario, the PAD solutions trained on three publicly available databases were evaluated on the test set of the CRMA database. In addition, the results tested on their own test sets are also reported (as shown in the left block in Table 4). The first setting is similar to protocol-1 of the CRMA intra-database scenario, as no masked data are seen in the training phase. Therefore, the first cross-database setting is also used to answer the first question: does the PAD algorithm trained on unmasked data generalize well on masked bona fides and attacks? Conversely, in the second cross-database experiment, models trained on different protocols of the CRMA database were evaluated separately on publicly available databases. This experimental setting can help us understand the CRMA database values beyond face masks, such as the diversity of masks/sensors/scales. However, the second scenario does not support the main study of the work and is provided only for completeness; thus, the results are reported in the supplementary material. In both cross-database scenarios, we use the $\tau_{BPCER10}$ decision threshold computed on the development set of the training database as a priori to determine the APCER, BPCER, and HTER values of the test database.

5.2. PAD evaluation metrics

The metrics following the ISO/IEC 30107-3 [41] standard were used to measure the performance of the PAD algorithms: *Attack Presentation Classification Error Rate* (APCER) and *bona fide presentation classification error rate* (BPCER). APCER is the proportion of attack images incorrectly classified as bona fide samples in a specific scenario, while BPCER is the proportion of bona fide images incorrectly classified as attacks in a specific scenario. The APCER and BPCER reported in the test set were based on a pre-computed threshold in the development set. In our study, we use a BPCER at 10% (on the development set) to obtain the threshold (denoted as $\tau_{BPCER10}$). Additionally, *half-total error rater* (HTER) corresponding to half of the summation of BPCER and APCER is used for the cross-database evaluation. Noticeably, we computed a threshold in the development set of the training database. Then, this threshold was used to determine the HTER value in the test database. The detection EER (D-EER) value, where APCER and BPCER are equal is also reported in the cross-database scenarios. For further analysis of PAD performance, receiver operating characteristic (ROC) curves were also demonstrated.

Table 4
Cross-database evaluation 1: the model trained on three publicly available databases is used to test on the CRMA database. This cross-database scenario is similar to protocol-1, as no masked data is seen during the training phase. Italic numbers indicate the lowest error rate on their own test set, and bold numbers indicate the highest error rate in the bona fide and each PAI category. The results show that despite good performance on their own test set, these trained models do not generalize well to masked bona fides and attacks.

Trained on	Method	Threshold @ BPCER 10% in dev set of trained database											
		Tested on the same dataset (%)			Tested on our CRMA dataset (%)								
CAISA-FASD		D-EER	BPCER	APCER	BPCER		APCER (Print)			APCER (Replay)			
					BM0	BM1	AM0	AM1	AM2	AM0	AM1	AM2
	LBP	7.50	6.25	8.75	38.60	**56.14**	**42.11**	24.76	18.13	**60.72**	34.59	22.51
	$\text{Inception}_{\text{FT}}$	10.00	8.75	15.00	21.05	**38.60**	**35.48**	5.95	16.96	**69.49**	47.44	15.50
	CPqD	6.25	11.25	3.12	38.60	**65.79**	**31.97**	12.38	8.77	**53.22**	23.06	14.62
	$\text{FASNet}_{\text{FT}}$	8.75	12.50	4.38	15.79	**90.35**	**44.83**	2.14	23.98	**64.13**	5.76	22.81
	$\text{Inception}_{\text{TFS}}$	*0.00*	*1.25*	*0.00*	12.28	**20.08**	**61.60**	40.35	49.71	**90.35**	83.19	59.65
	$\text{FASNet}_{\text{TFS}}$	1.25	3.75	0.62	21.05	**75.44**	**60.23**	19.49	38.60	**70.86**	16.32	45.61
	DeepPixBis	1.25	6.25	*0.00*	35.09	**66.67**	**70.57**	36.65	56.73	**57.99**	29.26	42.98
MSU-MFSD	LBP	4.17	*4.17*	4.17	98.25	**100.00**	0.58	**0.68**	0.00	**3.22**	2.25	0.00
	$\text{Inception}_{\text{FT}}$	20.14	20.81	16.67	**50.88**	25.44	47.95	**56.04**	52.05	31.19	**48.85**	44.15
	CPqD	4.17	*4.17*	4.17	98.25	**100.00**	0.19	**0.39**	0.00	**1.46**	1.56	0.00
	$\text{FASNet}_{\text{FT}}$	13.19	26.39	4.17	43.86	**85.96**	**32.55**	2.63	0.58	**42.50**	13.39	2.34
	$\text{Inception}_{\text{TFS}}$	4.17	8.33	1.39	80.70	**94.74**	**0.19**	0.00	0.00	**8.58**	0.78	2.05
	$\text{FASNet}_{\text{TFS}}$	*0.00*	8.44	*0.00*	91.23	**100.00**	0.00	0.00	0.00	**7.70**	0.00	0.29
	DeepPixBis	*0.00*	*4.17*	*0.00*	**82.46**	80.70	0.00	**0.10**	0.00	10.33	**10.36**	5.26
Oulu-NPU	LBP	8.33	7.50	10.21	40.35	**67.54**	**35.28**	25.54	13.45	**26.12**	10.89	13.74
	$\text{Inception}_{\text{FT}}$	15.00	16.67	11.04	61.40	**87.72**	**11.50**	5.85	8.77	**12.38**	2.39	1.46
	CPqD	8.33	9.17	3.54	57.89	**89.47**	**9.55**	3.70	1.17	**10.14**	1.03	0.58
	$\text{FASNet}_{\text{FT}}$	3.23	*1.67*	4.38	49.12	**73.68**	**33.92**	27.10	8.77	**22.81**	8.99	3.80
	$\text{Inception}_{\text{TFS}}$	4.17	3.33	6.46	80.07	**100.00**	**22.81**	0.78	2.34	**3.22**	0.00	0.00
	$\text{FASNet}_{\text{TFS}}$	5.10	11.67	3.33	70.18	**99.12**	**46.98**	18.03	19.88	**8.09**	0.39	0.29
	DeepPixBis	*2.29*	2.92	*0.00*	66.67	**98.25**	**44.64**	11.21	4.68	**10.23**	0.10	0.58

5.3. Analysis of protocol-1

Protocol-1 represents the pre-COVID-19 PAD scenarios, in which subjects normally do not wear a mask, and demonstrates the generalization performance on masked data. Therefore, protocol-1 is considered the most challenging task because of the unseen BM1, AM1, and AM2 data.

Table 3 describes the results of the different protocols on the CRMA database. The bold numbers indicate the highest BPCER values between BM0 and BM1 and the highest APCER values between AM0, AM1, and AM2 in each PAI. By observing the first block, P1, in Table 3, the BPCER values of masked bona fide samples are much higher than those of unmasked ones; however, most PAD systems achieve lower APCER values on the masked attack samples (either AM1 or AM2). The higher classification error rates on masked bona fide and the lower error rates on masked attacks are intuitively conceivable. When the model has not seen faces wearing a mask before, it is more inclined to falsely classify such a masked bona fide sample (BM1) as an attack.

Moreover, it is interesting to note that networks trained from scratch and the DeepPixBis approach work worse on attack AM2 than AM1. These observations are consistent with the ROC (Fig. 4). The red curves generated by printed AM2, bona fide BM1, and gray curves obtained by replay AM2 and bona fide BM1 possess significantly smaller areas under the curves in five of the seven methods. Furthermore, training a network from scratch improves the overall performance. The possible reason for those observations is that learning from scratch is more efficient for obtaining discriminative features between bona fide and artifacts. On the contrary, such approaches might be confusing when applying realistic masks to attack samples.

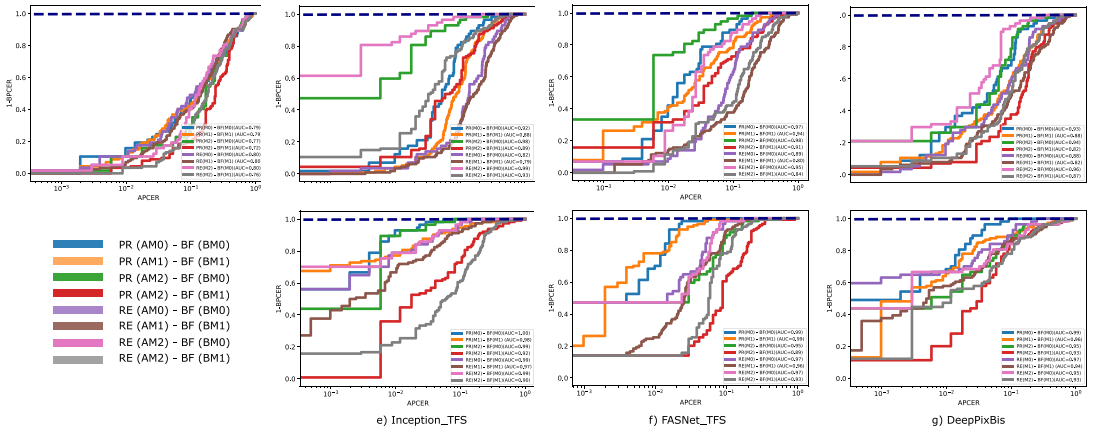
In addition to the intra-database scenario, the first cross-database experiment (introduced in Section 5.1.2) can be seen as similar to protocol-1, as both scenarios study PAD methods that PAD solutions trained on unmasked data and tested on the CRMA database. In Table 4, the bold BPCER number is the highest BPCER (between BM0 and BM1) for each PAD method. The bold APCER number is the highest APCER (between AM0, AM1, and BM2) for each PAD method in print and replay attacks, respectively. This bolding is performed to show which samples are more difficult to classify correctly. We observed that the performance in the cross-database setting was relatively poor for all models. Even though deep-learning-based methods achieved great results on their own test sets, they generalize significantly worse on masked bona fide samples; for example, most BPCER values for BM1 are close to 100%. In contrast, most algorithms achieve lower APCER values on masked AM1 and AM2 than unmasked AM0 attacks, which is consistent with the observation of protocol-1 from the intra-database scenarios.

In general, the experimental results of the intra-database protocol-1 and the first cross-database scenario results answer the first posed question (in Section 5.1.1) by showing that models trained only on unmasked data cannot properly classify images of masked faces. A subject with a mask on has a high probability of being falsely detected as an attack by PAD systems, even if this subject is bona fide.

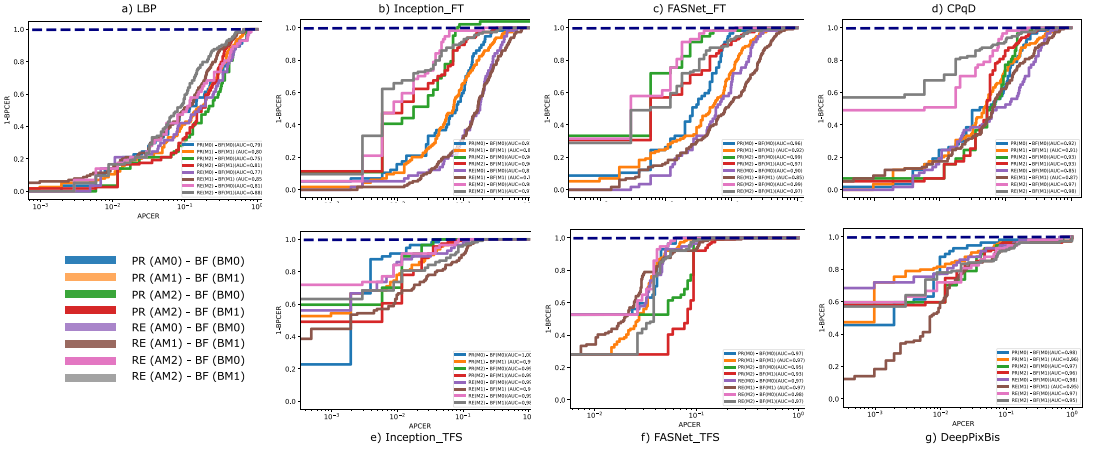
5.4. Analysis of protocol-2

Protocol-2 targets the performance of PAD algorithms on masked data when both unmasked and masked samples are used in the training phase. As shown in Table 3, we can observe the following points: First, despite the fact that the masked bona fide samples are still more difficult to classify correctly than unmasked ones in most cases, the BPCER value of BM1 behaves more similar to its behavior on the BM0 in protocol-2 than in protocol-1. Moreover, the BPCER values of BM0 and BM1 in protocol-2 decreased in most cases compared with the results of protocol-1. For example, the BPCER value of BM1 achieved by $\text{Inception}_{\text{FT}}$ was 84.21% in protocol-1 and 7.02% in protocol-2. This finding indicates that learning the masked data is helpful in improving the performance of PAD methods.

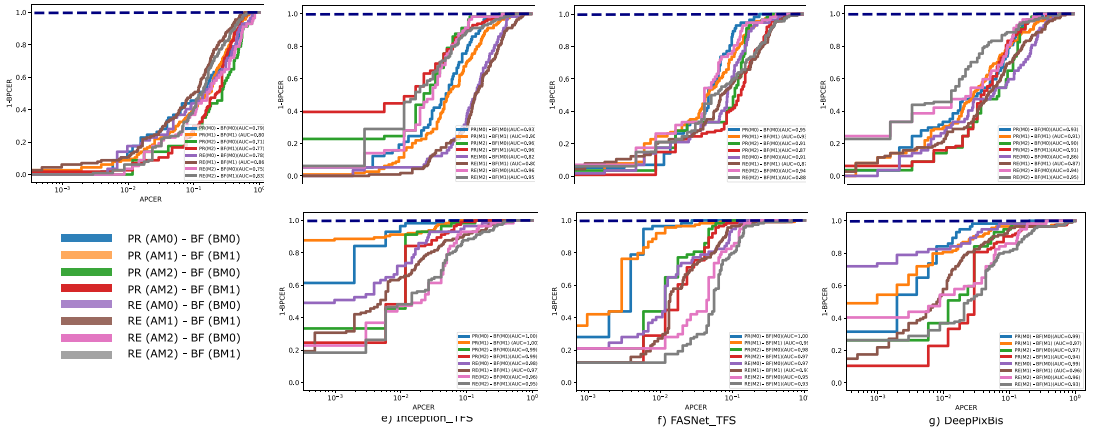
This is also consistent with the observation in the ROC curves (by comparing the ROCs in protocol-1 and protocol-2 in general).



(a) ROC curves in protocol-1



(b) ROC curves in protocol-2



(c) ROC curves in protocol-3

Fig. 4. ROC curves for all PAD methods in three protocols. Eight combinations between bona fide and attack (testing data) are represented for each method in each protocol: PR(AM0)-BF(BM0), PR(AM1)-BF(BM1), PR(AM2)-BF(BM0), PR(AM2)-BF(BM1) in print PAI and RE(AM0)-BF(BM0), RE(AM1)-BF(BM1), RE(AM2)-BF(BM0), RE(AM2)-BF(BM1) in replay PAI. The x-axis and y-axis are APCER and 1 - BPCER, respectively. The red curves (PR(AM2)-BF(BM1)) and gray curves (RE(AM2)-BF(BM1)) show significantly smaller AUC values by most PAD methods on protocol-1. Moreover, Inception$_{TFS}$, FASNet$_{TFS}$, and DeepPixBis achieve higher AUC values on protocol-2 and -3 than on protocol-1 might be due to the masked data in the training phase.. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In particular, Inception$_{TFS}$, FASNet$_{TFS}$, and DeepPixBis achieved significant progress (larger areas under the curves). Second, six of the seven methods performed worse on the masked printed face (AM1 or AM2), while five of the seven algorithms showed inferior results for unmasked replay attacks. Moreover, AM2 in print PAI achieves higher APCER values than AM1 by training from scratch approaches. One possible reason for the different results between print and replay attacks is specular reflection. Because attack data were collected in windowless labor with all electric lights on, tablets easily reflect the light compared to the printed paper, and

this reflection is difficult to avoid. The real face masks might also leak light when placed on an electric tablet, but this does not appear when applied on printed paper.

In general, the experimental results of the intra-database protocol-2 answer the second question (in Section 5.1.1), which addresses the performance changes of the current PAD algorithms after complementary learning on the masked data. Based on the above findings, we can conclude that the PAD algorithms still perform worse on masked bona fides (BM1) than on unmasked faces (BM0), even when the PAD solutions are trained on masked data.

5.5. Analysis of protocol-3

Protocol-3 investigates the generalizability of the model trained on data that includes masked face attacks (AM1) when tested on the masked face attacks where a real mask is placed on top of the attack (AM2). For bona fide samples, we draw a similar conclusion to protocol-1 and protocol-2, stating that masked bona fide samples have a higher probability of incorrectly being classified as attacks. However, the experimental results show differences in attack detection behavior (APCER) between protocol-3 on one side and protocols-1 and -2 on the other side. In this protocol, the highest APCER values of most PAD algorithms appear on either the AM1 or AM2 attacks in both print and replay PAIs. Second, the traditional LBP method, $\text{Inception}_{\text{FT}}$, $\text{FASNet}_{\text{FT}}$, and the hybrid CPqD method that achieve relatively worse results on AM0 or AM1 attacks than other methods may have proved to be unable to learn or extract sufficient discriminative features. Third, although the other methods, such as learning from scratch $\text{Inception}_{\text{TFS}}$ and $\text{FASNet}_{\text{TFS}}$ or custom designed DeepPixBis achieve impressive results on seen AM0 and AM1 attacks, they generalize not well on unseen AM2 attacks. These observations answer the third question stated in Section 5.1.1 by stating that a network trained on masked face attacks (AM1) tends to produce confusing decisions on AM2, where a real mask is placed on an attack face.

5.6. Qualitative analysis and visualization

To qualitatively analyze and interpret the deep-learning-based methods, the score-Weighted CAM [42] technique was adopted to localize the discriminative areas in face images. The rows from top to bottom correspond to $\text{Inception}_{\text{FT}}$, $\text{Inception}_{\text{TFS}}$, $\text{FASNet}_{\text{FT}}$, $\text{FASNet}_{\text{TFS}}$ and DeepPixBis. Fig. 5(a) shows the results of protocol-1 (the example subject is in the test set). $\text{Inception}_{\text{FT}}$ mainly focuses on the nose, including nearby parts of the masks, whereas $\text{Inception}_{\text{TFS}}$ pays more attention to the upper region of the face. Similarly, $\text{FASNet}_{\text{TFS}}$ reduces the attention paid to the masks and increases the concentration around the forehead. DeepPixBis concentrates around the eyes for both unmasked (BM0) and masked (BM1) bona fides. However, for attack samples, attention seems to be focused on the left eye and partial masks. In general, masks are noticed by all networks. The results of protocol-2 and protocol-3 for the same subjects are shown in Fig. 5(b), and Fig. 5(c). We noticed that 1) the attention areas of fine-tuned networks hardly change in the three protocols because of the fixed weights of layers before the last classification layer. 2) $\text{Inception}_{\text{TFS}}$ in protocol-2 appears to focus on the upper face, including many more eye regions than in protocol-1. 3) $\text{FASNet}_{\text{TFS}}$ in protocol-2 concentrates much more on applied real masks than in protocol-3 where training without AM2. 4) DeepPixBis still works well on bona fide, but for attack samples, its attention seems to be distracted to the edge of images. Although DeepPixBis produces correct decisions, this observation raises a serious concern about its reliability and generalizability. This concern was confirmed by the cross-database evaluation. DeepPixBis generally obtains worse cross-database results than the other two training from scratch networks (details

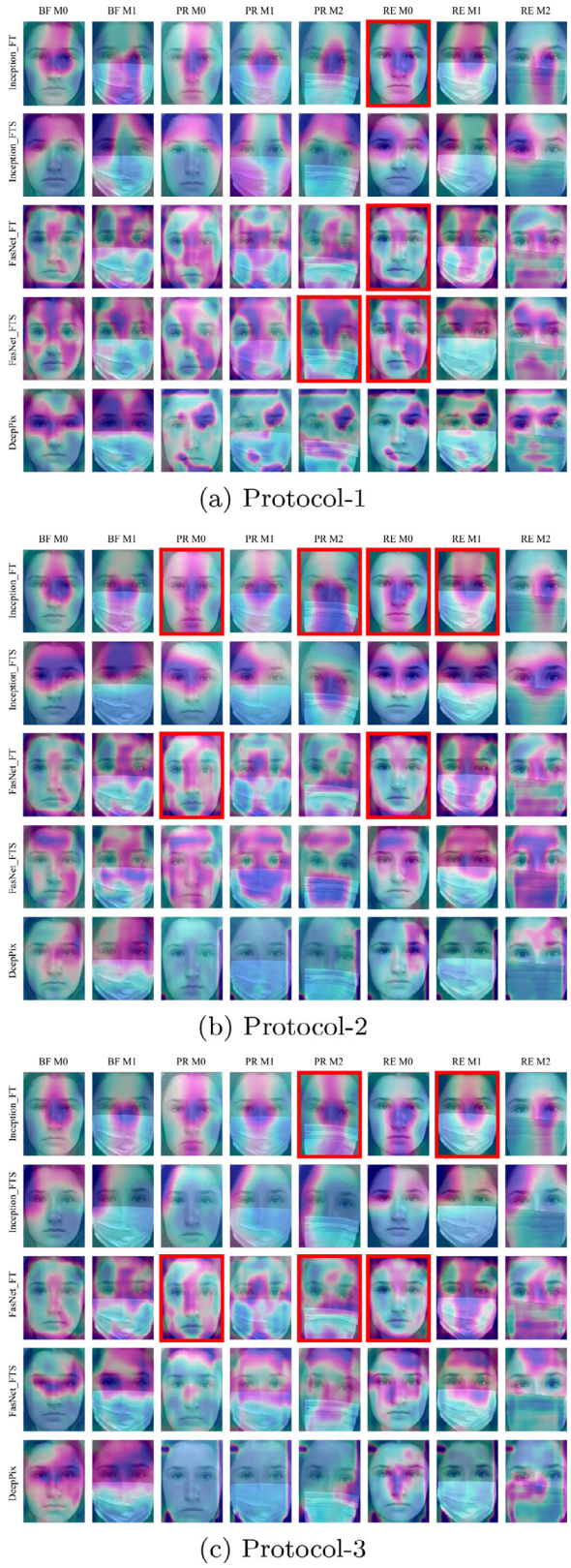


Fig. 5. Examples for attention maps generated by ScoreCAM of different PAD algorithms and different protocols. The rows from top to bottom in each protocol correspond to $\text{Inception}_{\text{FT}}$, $\text{Inception}_{\text{TFS}}$, $\text{FASNet}_{\text{FT}}$, $\text{FASNet}_{\text{TFS}}$, and DeepPixBis. The columns from left to right in each protocol refer to BM0, BM1, PR-AM0, PR-AM1, PR-AM2, RE-AM0, RE-AM1, RE-AM2. Faces with red boxes are misclassified. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

see Table 4). Finally, looking at attention maps in all protocols for this identity, we notice that except for the misclassified samples (with red boxes) that appear on print/replay AM0, print AM2 attacks are more easily to be incorrectly detected as bona fide than AM1 attacks. This finding is in line with the previous quantitative evaluation that AM2 attacks may confuse the PAD, even if the network has been trained by masked face attacks.

To further understand the above quantitative and qualitative results, we provide additional t-SNE plots to visualize the learned features in the supplementary material. These plots consolidate our findings here that 1) masked bona fide samples are more likely to be detected as bona fide by the pre-COVID-19 PAD algorithms. 2) attacks with real masks placed on the attacks (AM2) are more falsely detected by PAD systems as bona fides than attacks with masked faces (AM1).

6. Analysis of FR vulnerability

6.1. Experimental settings

The vulnerability of each FR system on each type of PA is analyzed based on three experimental settings. In the first setting BM0-BM0, we use the bona fide unmasked samples captured on the first day as references. Then, the references are compared against bona fide BM0 samples captured on the second and third days of the same subjects (to compute genuine scores), as well as of other subjects (zero-effort imposter (ZEI) scores). Once genuine and ZEI comparison scores are obtained, the operating threshold is computed using the $\tau_{FMR@0.01}$ threshold. Finally, the probe samples of each type of PA were compared against the reference of the same subjects separately. In the second setting BM0-BM1, the difference is that bona fide BM1 data captured on the second and third days are used for comparison against references BM0 and then obtain the corresponding genuine and ZEI scores. In the third setting, BM1-BM1, the bona fide masked faces captured on the first day are references for each subject. Such references are also compared against the masked bona fide samples captured on the second and third days to obtain their genuine and ZEI scores.

These three experimental settings are provided to enable addressing the following four questions: 1) When having an unmasked reference and we use a decision threshold that does not consider masked comparisons (BM0-BM0), how vulnerable are FR systems to the three types of attacks in CRMA (AM0, AM1, and AM2)? 2) When having an unmasked reference and we use a decision threshold based on unmasked-to-masked comparisons (BM0-BM1), how vulnerable are FR systems to the three types of attacks in CRMA (AM0, AM1, and AM2)? 3) When having a masked reference and we use a decision threshold based on masked-to-masked comparisons (BM1-BM1), how vulnerable are FR systems to the three types of attacks in CRMA (AM0, AM1, and AM2)? Additionally, we address the fourth question: 4) will the vulnerability of FR systems be different when facing the AM1 and AM2 attacks?

6.2. Evaluation metrics

To measure the performance of FR techniques, the *genuine match rate* (GMR), which refers to the proportion of correctly matched genuine samples, is used at the fixed false match rate (FMR). GMR is equal to 1 minus the false non-match rate (FNMR). Moreover, to analyze the vulnerability of FR algorithms for our masked attacks, the *imposter attack presentation match rate* (IAPMR) corresponding to the proportion of PAs accepted by the FR system as genuine presentations is adopted. IAPMR also follows the standard definition presented in ISO/IEC 30107-3 [41]. The threshold for GMR and IAMPR is defined by fixing the FMR at 1% (denoted as $\tau_{FMR@0.01}$). The probe images with similarity scores lower than the $\tau_{FMR@0.01}$ are not matched. Moreover, the recognition score-distribution histograms are shown in Fig. 6, 7, and 8. In addition

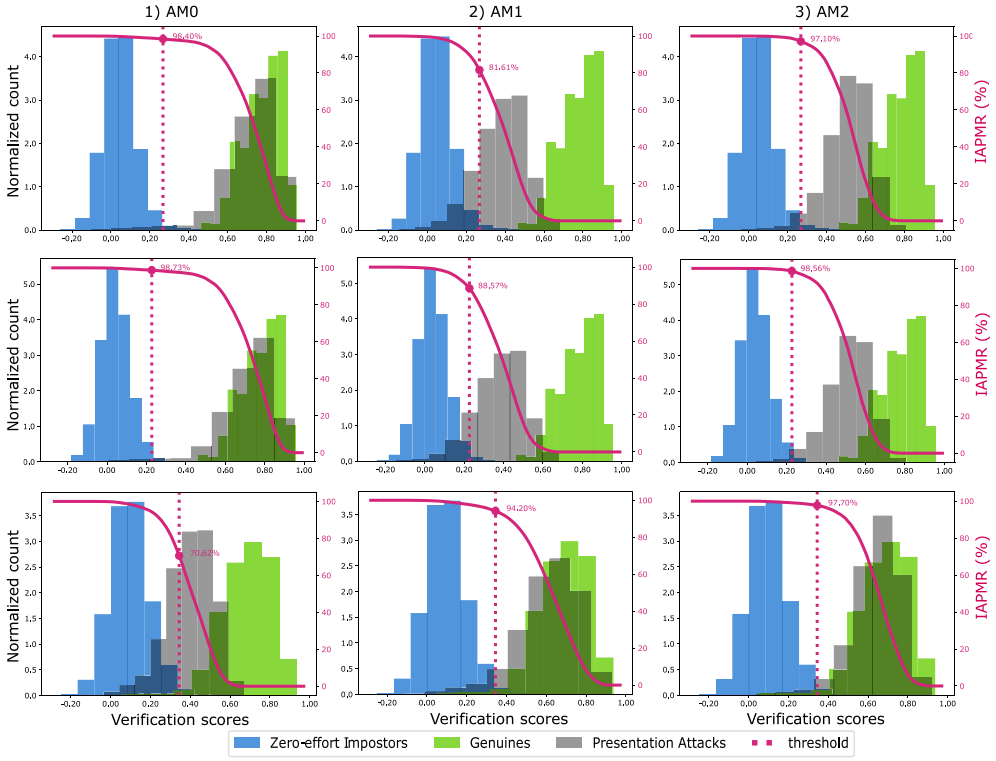


Fig. 6. The similarity score distributions by off-the-shelf ArcFace [28]. The rows from top to bottom represent three experimental settings: BM0-BM0, BM0-BM1, BM1-BM1, as shown in Table 5.

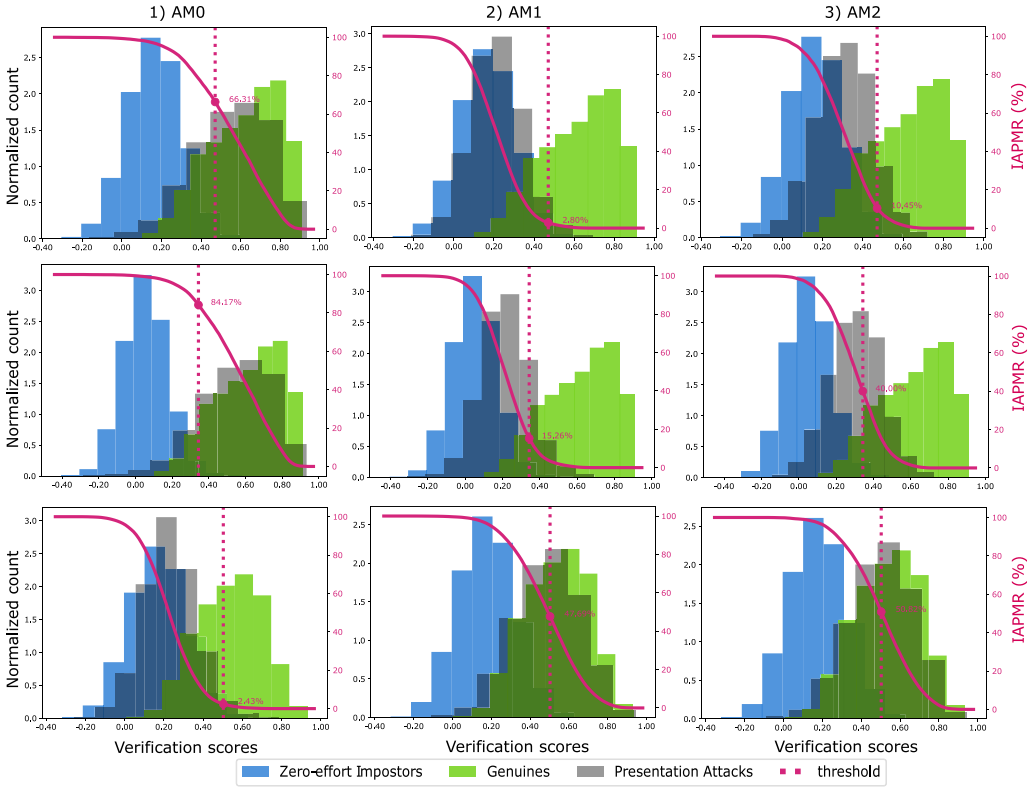


Fig. 7. The similarity score distributions by off-the-shelf SphereFace [29].

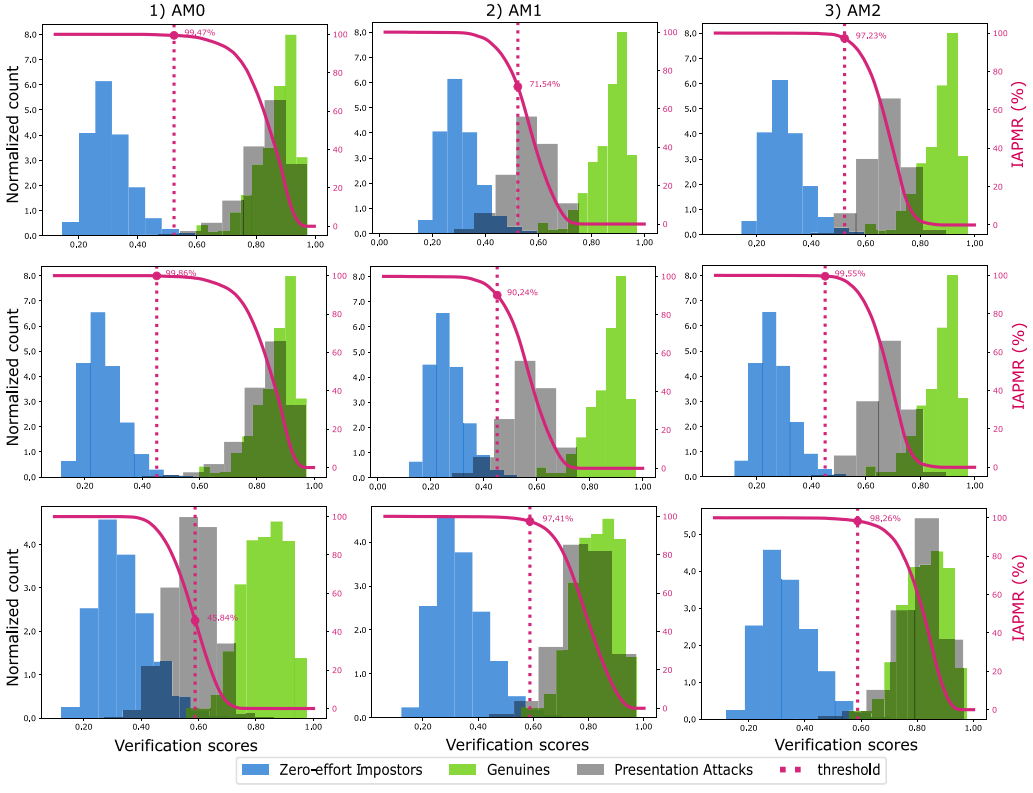


Fig. 8. The similarity score distributions by off-the-shelf VGGFace [29].

Table 5
The performance and vulnerability of FR systems. The GMR and IAPMR values were computed based on the $\tau_{FMR@0.01}$ threshold. The 95% confidence intervals for the IAPMR values are shown in parentheses.

Settings	Attack Probes	ArcFace[28]			SphereFace [29]			VGGFace [30]		
		EER	GMR	IAPMR	EER	GMR	IAPMR	EER	GMR	IAPMR
BM0 - BM0	AM0	0.00	100	98.40 [98.22, 98.56]	8.57	75.85	66.31 [65.69, 66.93]	0.12	100	99.47 [99.37, 99.56]
	AM1			81.61 [81.24, 81.97]			2.80 [2.65, 2.96]			71.54 [71.12, 71.96]
	AM2			97.10 [96.77, 97.41]			10.45 [9.89, 11.03]			97.23 [96.91, 97.53]
BM0 - BM1	AM0	2.25	96.56	98.73 [98.58, 98.88]	22.83	19.99	84.17 [83.68, 84.64]	2.29	94.2	99.86 [99.80, 99.90]
	AM1			88.57 [88.27, 88.86]			15.26 [14.92, 15.60]			90.24 [89.96, 90.51]
	AM2			98.56 [98.33, 98.78]			40.00 [39.09, 40.91]			99.55 [99.41, 99.67]
BM1 - BM1	AM0	1.00	99.00	70.62 [70.19, 71.04]	13.13	59.33	2.43 [2.29, 2.58]	0.85	99.46	45.84 [45.38, 46.31]
	AM1			94.20 [94.04, 94.35]			47.69 [47.36, 48.02]			97.41 [97.30, 97.51]
	AM2			97.70 [97.49, 97.89]			50.82 [50.16, 51.48]			98.26 [98.08, 98.43]

to these metrics, the EER value, where FMR equals FNMR, is computed to compare the FR algorithms.

6.3. Analysis of FR results

The performance and vulnerability of each FR system are summarized in Table 5. SphereFace [29] obtains relatively low IAPMR values; however, its GMR values are also much lower than those of ArcFace [28] and VGGFace [30]. In general, the IAPMR values of all three FR systems were close to their GMR values. Specifically, FR systems are vulnerable to unmasked attacks when unmasked bona fide samples are used as references (the settings BM0-BM0 and BM0-BM1), and vulnerable to the masked attacks when the reference is masked bona fide data. Comparing the vulnerability analysis results for AM1 and AM2 in all three cases and all FR systems, we note that the IAMPR values of AM2 are always significantly higher than those of AM1. This indicates that applying real masks on attack presentations can further reduce the performance of FR systems. This might be due to the fact that the AM2 attacks possess more realistic features than AM1. To further verify this assumption, we provide histograms of the similarity score distribution in the three scenarios and three FR systems (see Fig. 6, 7, and 8). In the histograms, green refers to genuine scores, blue represents ZEI scores, and gray represents attack verification scores. The ideal situation is that there is no overlap between the green and the other two histograms. Fig. 6 shows the score distributions of ArcFace [28], where the rows from top to bottom represent BM0-BM0, BM0-BM1, BM1-BM1 cases and columns from left to right refer to AM0, AM1, and AM2 attacks. It can be seen that 1) the verification scores of attacks are higher than the scores of ZEI in all cases. 2) The scores of AM0 attacks and genuine scores almost overlap in the BM0-BM0 and BM0-BM1 settings, while the scores of AM1/AM2 attacks have many overlapping areas with genuine scores in the BM1-BM1 setting. 3) for all cases, the scores of AM2 have more overlaps with genuine scores than AM1. Similar observations can be found in Fig. 7 for the SphereFace, and Fig. 8 for VGGFace. These observations are consistent with the findings presented in Table 5.

Overall, these results indicate that 1) FR systems are more vulnerable to unmasked attacks compare to masked attacks when the references are unmasked faces, 2) when the threshold is computed based on the unmasked-to-masked comparison, the vulnerability of FR systems becomes higher for both masked or unmasked attacks, 3) when the reference is masked, FR systems are more vulnerable to masked attacks in comparison to the FR systems having unmasked references. Another important finding is that 4) FR systems pose a higher vulnerability for spoof faces with real masks placed on them (AM2) than a masked face attack (AM1). Such observations raise concerns about the security of FR systems when facing masked attacks.

7. Conclusion

We studied the behavior of PAD methods on different types of masked face images. To enable our study, we presented a new large-scale face PAD database, CRMA, including the conventional unmasked attacks, novel attacks with faces wearing masks, and attacks with real masks placed on spoof faces. It consists of 13,113 high-resolution videos and has a large diversity in capture sensors, displays, and capture scales. To study the effect of wearing a mask on the PAD algorithms, we designed three experimental protocols. The first protocol measures the generalizability of the current PAD algorithms on unknown masked bona fide or attack samples. In the second protocol, masked data are used in the training phase to measure the performance of PAD solutions where the face masks are known. The third protocol investigates the generalizability of models trained on masked face attacks when tested on attacks covered by a real mask. Extensive experiments were conducted using these protocols. The results showed that PAD algorithms have a high possibility of detecting masked bona fide samples as attackers (median BPCER value for BM1 in protocol-1 is 48.25%). Moreover, even if the PAD solutions have seen the masked bona fide data, the PAD algorithms still perform worse on masked bona fide samples compared with unmasked bona fides. Furthermore, the PAD solutions trained on masked face attacks (AM1) do not generalize well on attacks covered by a real mask (AM2). For example, the APCER values achieved by DeepPixBis increased from 0.62% for AM1 to 2.92% for AM2 in print attack and from 1.92% for AM1 to 6.43% for AM2 in replay (protocol-3). In addition, we performed a thorough analysis of the vulnerability of FR systems to such novel attacks. The results indicate that FR systems are vulnerable to both masked and unmasked attacks. For example, when the reference images and system threshold are based on unmasked faces (BM0-BM0), the IAPMR values for unmasked attacks (AM0), masked attacks (AM1), and attacks covered by a real mask (AM2) are 98.40%, 81.60%, and 97.10%, respectively. This leads to the interesting observation that all the investigated FR systems are more vulnerable to attacks where real masks are placed on attacks (AM2) than attacks of masked faces (AM1).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

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 Center for Applied Cybersecurity ATHENE.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.patcog.2021.108398.

References

[1] M.L. Ngan, P.J. Grother, K.K. Hanaoka, Ongoing Face Recognition Vendor Test (FRVT) Part 6B: Face recognition accuracy with face masks using post–COVID-19 algorithms, Technical Report, 2020.

[2] N. Damer, J.H. Grebe, C. Chen, F. Boutros, F. Kirchbuchner, A. Kuijper, The effect of wearing a mask on face recognition performance: an exploratory study, in: BIOSIG - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, in: LNI, volume P-306, Gesellschaft für Informatik e.V., 2020, pp. 1–10.

[3] N. Damer, F. Boutros, M. Süßmilch, F. Kirchbuchner, A. Kuijper, Extended evaluation of the effect of real and simulated masks on face recognition performance, IET Biom. 10 (5) (2021) 548–561.

[4] S. Jia, G. Guo, Z. Xu, A survey on 3d mask presentation attack detection and countermeasures, Pattern Recognit. 98 (2020) 107032.

[5] S. Jia, C. Hu, X. Li, Z. Xu, Face spoofing detection under super-realistic 3d wax face attacks, Pattern Recognit. Lett. 145 (2021) 103–109.

[6] Z. Yu, J. Wan, Y. Qin, X. Li, S.Z. Li, G. Zhao, NAS-FAS: Static-dynamic central difference network search for face anti-spoofing, IEEE Trans. Pattern Anal. Mach. Intell. 43 (9) (2021) 3005–3023.

[7] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, A. Hadid, OULU-NPU: A mobile face presentation attack database with real-world variations, in: 12th IEEE International Conference on Automatic Face & Gesture Recognition, FG, IEEE Computer Society, 2017, pp. 612–618.

[8] Y. Liu, A. Jourabloo, X. Liu, Learning deep models for face anti-spoofing: Binary or auxiliary supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2018, pp. 389–398.

[9] Y. Liu, J. Stehouwer, A. Jourabloo, X. Liu, Deep tree learning for zero-shot face anti-spoofing, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2019, pp. 4680–4689.

[10] X. Tan, Y. Li, J. Liu, L. Jiang, Face liveness detection from a single image with sparse low rank bilinear discriminative model, in: European Conference on Computer Vision, ECCV Proceedings, Part VI, volume 6316, Springer, 2010, pp. 504–517.

[11] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, S.Z. Li, A face antispoofing database with diverse attacks, in: 5th IAPR International Conference on Biometrics, ICB, IEEE, 2012, pp. 26–31.

[12] I. Chingovska, A. Anjos, S. Marcel, On the effectiveness of local binary patterns in face anti-spoofing, in: BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group, in: LNI, volume P-196, GI, 2012, pp. 1–7.

[13] D. Wen, H. Han, A.K. Jain, Face spoof detection with image distortion analysis, IEEE Trans. Inf. Forensics Secur. 10 (4) (2015) 746–761.

[14] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, Z. Liu, Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations, in: European Conference on Computer Vision, ECCV, Proceedings, Part XII, volume 12357, Springer, 2020, pp. 70–85.

[15] N. Erdogmus, S. Marcel, Spoofing in 2d face recognition with 3D masks and anti-spoofing with kinect, in: IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS, IEEE, 2013, pp. 1–6.

[16] I. Chingovska, N. Erdogmus, A. Anjos, S. Marcel, Face recognition systems under spoofing attacks, in: Face Recognition Across the Imaging Spectrum, Springer, 2016, pp. 165–194.

[17] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H.J. Escalante, S.Z. Li, CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing, IEEE Trans. Biom. Behav. Identity Sci. 2 (2) (2020) 182–193.

[18] S. Bhattacharjee, A. Mohammadi, S. Marcel, Spoofing deep face recognition with custom silicone masks, in: 9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS, IEEE, 2018, pp. 1–7.

[19] X. Song, X. Zhao, L. Fang, T. Lin, Discriminative representation combinations for accurate face spoofing detection, Pattern Recognit. 85 (2019) 220–231.

[20] Y. Jia, J. Zhang, S. Shan, X. Chen, Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing, Pattern Recognit. 115 (2021) 107888.

[21] S. Fatemifar, S.R. Arashloo, M. Awais, J. Kittler, Client-specific anomaly detection for face presentation attack detection, Pattern Recognit. 112 (2021) 107696.

[22] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[23] J. Määttä, A. Hadid, M. Pietikäinen, Face spoofing detection from single images using micro-texture analysis, in: IEEE International Joint Conference on Biometrics, IJCB, IEEE Computer Society, 2011, pp. 1–7.

[24] Z. Boulkenafet, J. Komulainen, A. Hadid, Face anti-spoofing based on color texture analysis, in: IEEE International Conference on Image Processing, ICIP, IEEE, 2015, pp. 2636–2640.

[25] Z. Boulkenafet, J. Komulainen, Z. Akhtar, A. Benlamoudi, D. Samai, S.E. Bekhouche, A. Ouafi, F. Dornaika, A. Taleb-Ahmed, L. Qin, F. Peng, L.B. Zhang, M. Long, S. Bhilare, V. Kanhangad, A. Costa-Pazo, E. Vázquez-Fernández, D. Pérez-Cabo, J.J. Moreira-Perez, D. González-Jiménez, A. Mohammadi, S. Bhattacharjee, S. Marcel, S. Volkova, Y. Tang, N. Abe, L. Li, X. Feng, Z. Xia, X. Jiang, S. Liu, R. Shao, P.C. Yuen, W.R. de Almeida, F.A. Andaló, R. Padilha, G. Bertocco, W. Dias, J. Wainer, R. da Silva Torres, A. Rocha, M.A. Angeloni, G. Folego, A. Godoy, A. Hadid, A competition on generalized software-based face presentation attack detection in mobile scenarios, in: IEEE International Joint Conference on Biometrics, IJCB, IEEE, 2017, pp. 688–696.

[26] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, R. de Alencar Lotufo, Transfer learning using convolutional neural networks for face anti-spoofing, in: Image Analysis and Recognition - 14th International Conference, ICIAR, volume 10317, Springer, 2017, pp. 27–34.

[27] A. George, S. Marcel, Deep pixel-wise binary supervision for face presentation attack detection, in: International Conference on Biometrics, ICB, IEEE, 2019, pp. 1–8.

[28] J. Deng, J. Guo, N. Xue, S. Zafeiriou, Arcface: Additive angular margin loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2019, pp. 4690–4699.

[29] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, Sphereface: Deep hypersphere embedding for face recognition, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2017, pp. 6738–6746.

[30] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, A. Zisserman, Vggface2: A dataset for recognising faces across pose and age, in: 13th IEEE International Conference on Automatic Face & Gesture Recognition, FG, IEEE Computer Society, 2018, pp. 67–74.

[31] A. Mohammadi, S. Bhattacharjee, S. Marcel, Deeply vulnerable: a study of the robustness of face recognition to presentation attacks, IET Biom. 7 (1) (2018) 15–26.

[32] International Civil Aviation Organization, Technical Report: Portrait Quality (Reference Facial Images for MRTD), Technical Report, 2018.

[33] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments, Technical Report, University of Massachusetts, Amherst, 2007.

[34] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: IEEE Conference on Computer Vision and Pattern Recognition,CVPR, IEEE Computer Society, 2016, pp. 2818–2826.

[36] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations, ICLR, 2015.

[37] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2017, pp. 2261–2269.

[38] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, in: The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR, IEEE Computer Society, 2011, pp. 529–534.

[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, IEEE Computer Society, 2016, pp. 770–778.

[40] Y. Guo, L. Zhang, Y. Hu, X. He, J. Gao, Ms-celeb-1m: A dataset and benchmark for large-scale face recognition, in: European Conference on Computer Vision, ECCV, Proceedings, Part III, volume 9907, Springer, 2016, pp. 87–102.

[41] International Organization for Standardization, ISO/IEC DIS 30107-3:2016: Information Technology Biometric presentation attack detection P. 3: Testing and reporting, Technical Report, 2017.

[42] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, X. Hu, Score–cam: score-weighted visual explanations for convolutional neural networks, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops, IEEE, 2020, pp. 111–119.

Meiling Fang received her master degree from Karlsruhe Institute of Tech- nology (2019). Since August 2019, she is a researcher at Fraunhofer IGD. Her research interests are in the fields of machine learning, computer vision, and biometrics.

Naser Damer is a senior researcher at Fraunhofer IGD. He received his PhD from TU Darmstadt (2018). He is a principal investigator at the ATHENE Center, serves as an AE of TVCJ, lectures at TU Darmstadt, and represents the German Institute for Standardization (DIN) in ISO/IEC SC37 standardization committee.

Florian Kirchbuchner received his master degree from TU Darmstadt in 2014. Since 2018, he has been Head of the Department for Smart Living & Biometric Technologies at the IGD. Mr. Kirchbuchner has been the spokesperson for the Fraunhofer Alliance Ambient Assisted Living AAL since 2019.

Arjan Kuijper is a member of the management of Fraunhofer IGD, responsible for scientific dissemination. He holds the Chair in mathematical and applied vi- sual computing with TU Darmstadt. He is the author of over 300 peer-reviewed publications, the AE of CVIU, PR, and TVCJ, and the Secretary of the IAPR.