Real Masks and Fake 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
Email:meiling.fang@igd.fraunhofer.de

The ongoing COVID-19 pandemic has lead to massive public health issues. Face masks have become one of the most efficient ways to reduce coronavirus transmission. This makes face recognition (FR) a challenging task as several discriminative features are hidden. Moreover, face presentation attack detection (PAD) is crucial to ensure the security of FR systems. In contrast to growing numbers of masked FR studies, the impact of masked attacks on PAD has not been explored. Therefore, we present novel attacks with real 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 intra- and cross-database scenarios. We also evaluate the vulnerability of FR systems on masked attacks. The experiments show that real masked attacks pose a serious threat to the operation and security of FR systems.

***Index Terms*—Face presentation attack detection, COVID-19, Masked face, Face recognition, Biometric security**

I. Introduction

Since the SARS-CoV-2 coronavirus outbreak and its rapid spread worldwide, wearing a mask has become one of the most efficient ways to protect and prevent the widespread of virus infections. However, in crowded scenarios like airports or identity checks, taking off the mask for face recognition (FR) increases infection chance. As a result, researchers have shown an increased interest in the effect of face masks on the performance of FR systems [1], [2]. Their experimental results showed that pre-COVID-19 FR algorithms suffer degradation in accuracy due to masked faces. Therefore, several attempts have been made to improve the FR performance on masked faces to fit the current situation [3], [4]. Additionally, even before the COVID-19 pandemic, face occlusion [5], [6] has been studied for a long time. However, no previous face Presentation Attack Detection (PAD) studies have been conducted exclusively on the occlusion or facial masks. With the popularity of FR systems, attackers use presentation attacks (PAs) to the FR systems and attempt to impersonate someone or obfuscate their identity. Even though FR algorithms achieved a remarkable improvement, most FR systems are still vulnerable to PAs, such as printed images, replay videos, or 3D masks. So far, very little attention has been paid to the vulnerability analysis of FR systems.

Almost all the countries across the globe have supported the use of masks to minimize the spread of the virus. Hence, we believe that wearing masks in public will be an essential health measure and a new norm even after the COVID-19 pandemic. To avoid frequently pulling off the masks in crowded scenarios, a very recent approach [4] is proposed by Li *et al.*to utilize a cropping and attention-based network to improve the performance of masked face recognition. As shown in their class activation visualization (CAM) maps, three of seven methods paid too much attention to the mask, while the rest methods focused well on the area around the eyes yet still included small mask areas. Because more FR systems will further target the masked face problem, the PAD research field calls for the need to collect masked attack data and the vulnerability analysis of FR systems for masked attack faces. Since face PAD has emerged as a crucial technique to protect face recognition security, most of the recent face PAD databases [7], [8], [9], [10] have been focused on enlarging the diversity in subjects, sessions, or sensors to enhance the generalizability and reliability of PAD techniques from the root data limitation. However, these PAD databases have an obvious shortcoming as no masked PAs are provided. Much uncertainty still exists about the relationship between the performance of PAD techniques and masked attacks. To fill such gaps, researchers require well-studied masked attack data for further developments. The main contributions in this work are:

- The novel Collaborative Real Mask Attack Database (CRMA) is presented. The bona fide samples are collected in realistically variant collaborative face capture scenarios [2]. Based on bona fide subjects, we generate three types of mask attacks for each subject and each attack category (print and replay): subject without a mask on, with a mask on, and with a real mask placed on print images or replay videos (samples in Fig. 1). For the creation of such attacks, three electronic tablets with high-resolution and three capture scales are used. Additionally, we design three experimental protocols for exploring the effect of masked attacks on PAD performance.
- Extensive experiments are conducted to explore the effect of real masks (on attack faces) and masked faces attacks and bona fide samples on the face PAD behavior. To support the comprehensive evaluation, seven PAD algorithms comprising of texture-based, deep-learning based, and hybrid methods are selected to evaluate the performance and generalizability in intra- and cross-database scenarios under three mask-related protocols. The quantitative and qualitative analysis both reveal that masked bona fides and PAs dramatically decrease the performance of PAD

algorithms. Moreover, deep-learning based methods perform worse on real mask attacks than mask faces attacks in most cases.
- An in-depth vulnerability analysis of FR systems is presented. We evaluate three deep-learning based FR techniques to three types of mask attacks. 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 Sec. II. Then, our novel CRMA database are described in detail in Sec. III. The used face PAD algorithms, the used FR systems, and evaluation metrics are introduced in Sec. IV. Sec. V discusses the PAD results and analyzes the vulnerability of FR systems. Finally, the conclusion is presented in Sec. VI.

II. Related Work

This section reviews the most relevant prior works to ours from three perspectives: face PAD databases, face PAD algorithms, FR techniques and vulnerability analysis.

Face PAD Databases: Data resources have become especially important since heading into the deep learning era, 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 were previously released, e.g., NUAA [11], CASIA-FAS [12], Replay-Attack [13], MSU-MFSD [16], OULU-NPU [7], and SiW [8], all consisting of 2D print/replay attacks. In addition, SiW-M [9] and Celeb-Spoof [10] databases provided multiple types of attacks like makeup, 3D mask or paper cut. Moreover, some multi-modal databases are publicly available: 3DMAD [14], Mssproof [15], CASIA-SURF [17], CSMAD [18]. These databases contribute to the significant progress of PAD research without a doubt, e.g., CeleA-Spoof database collected images from various environments and illuminations with rich annotations to reflect real scenes. However, these databases also have weaknesses: 1) the multi-modal databases have high hardware requirements and cannot be widely used in daily life, 2) some databases like CASIA-MFS [12] and MSU-MFS [16] 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 Celeb-Spoof [10] are relatively up-to-date, but lack any consideration of the real face mask attack to fit the current COVID-19 pandemic. Hence, we collect the CRMA database to fill the gap between these databases and the ongoing COVID-19 pandemic, and at the same time, ensure the generalizability and compatibility with real scenarios. The CRMA database can be used to better analyze the effect of 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 above mentioned databases is listed in Tab. I.

Face PAD Methods: In recent years, there has been an increasing amount of studies in the face PAD field. These studies can be broadly grouped into three categories based on features: textured based methods, deep-learning based methods, and hybrid methods. Texture features, such as Local Binary Pattern (LBP) [19] and Binarized Statistical Image Feature (BSIF) [20], project the faces to a low-dimension embeddings. Määättä *et al.*[21] 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 bona fide or attack. The LBP features extracted from different color spaces [22] were further proposed to utilize the chrominance information. They achieved competitive results on Replay-Attack [13] (Equal Error Rate (EER) value of 0.4%) and CASIA-FAS [12] (EER value of 6.2%) databases. Furthermore, Boulkenafet *et al.*[23] organized a face PAD competition based on the OULU-NPU database and compared 13 algorithms provided by participating teams and one color-LBP based baseline. In this competition, the GRADIANT algorithm, fusing color, texture, and motion information, achieved competitive results in four evaluation protocols. In addition to the hand-crafted feature 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 fine-tuned Inception-v3 network and the baseline method. Consequently, we chose re-implement the baseline and CPqD method in this paper (details in Sec. IV-A), while the GRADIANT and MixedFASNet are discarded in our work as they do not provide enough details for re-implementation. Deep-learning based methods have been pushing the frontier of face PAD research and have shown remarkable improvement in PAD performance. Lucena *et al.*[24] presented an approach, named FASNet, that a pre-trained VGG16 is fine-tuned by replacing the last fully-connected layer. The FASNet network achieved great performance on 3DMAD [14] and Replay-Attack database [13]. Recently, George *et al.*[25] proposed an approach that utilized a pixel-wise supervision on output maps forced the CNN to learn shared representation using information from different patches. DeepPixBis [25] outperformed not only state-of-the-art algorithms in Protocol-1 of OULU-NPU database (e.g.,1.6% ACER by auxiliary and 0.42% by DeepPixBis) but also achieved much 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 choose the FASNet and DeepPixBis (details in Sec. IV-A) to study the effect of the real mask and masked face attacks on 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. Moreover, FR techniques [26], [27], [28] 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. 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

Database	Year	# Subjects	# Data (BF/attack)	Capture devices (BF/attack)	Display devices	Modality	Attack type
NUAA [11]	2010	15	5105/7509 (I)	Webcame	-	RGB	1 Print
CASIA-FAS [12]	2012	50	150/450 (V)	Two USB cameras, Sony NEX-5	iPad	RGB	1 Print, 1 Replay
Replay-Attack [13]	2012	50	200/100 (V)	MacBook 13 / iPhone 3GS, Cannon SX150	iPhone 3GS, iPad	RGB	1 Print, 2 Replay
3DMAD [14]	2013	17	170/85 (V)	Microsoft Kinect	-	RGB/Depth	1 3D Mask
Msspoof [15]	2015	21	1,680/3,024 (I)	uEye camera	-	RGB/IR	1 Print
MSU-MFSD [16]	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 168B	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 [10]	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)	Various 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**

TABLE I: 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 one database containing subjects wearing face masks and real face mask attacks. The details of our CRMA database is introduced in Sec. III.

accuracy, e.g., even the best of the 89 algorithms had error rates between 5% to 50%. It is worth noting that the masks they used in experiments were synthetically created, not real masks. Damer *et al.*[2] presented a real mask database to simulate a realistically variant collaborative face capture scenario. They also explored the effect of wearing a mask on FR performance and concluded that the face masks significantly reduce the accuracy of algorithms. Mohammadi *et al.*[29] contributed empirical evidence to support the claim that the CNN-based FR method is 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 the custom-mask based PAs. However, the vulnerability of FR systems on masked face attacks has not been investigated. Therefore, in this work, we selected three CNN based FR algorithms for further FR-vulnerability analysis on masked face attacks: the state-of-the-art ArcFace [26], SphereFace [27], and VGGFace [28]. These algorithms are discussed in more detail in Sec. IV-B.

III. The Collaborative Real Mask Attack Database (CRMA)

The CRMA database [1] can serve as a supplement to the previous databases in Tab. I, and due to the COVID-19 pandemic, it can better reflect the current real-world PAD performance. The data also presents novel attack scenarios that were not previously studied. Fig. 3 introduces the general statistical information of the CRMA database. This database contains 62% males and 38% females. The attack M0, M1, and M2 ratios are 30%, 60%, and 10%, respectively. 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 0.05% and 0.30%. This section begins with the collection of bona fide samples, which an early version of it is presented in [2], then we introduce the process of attack creation. Additionally, three evaluation protocols are introduced to explore the effect of real masks in FR and PAD systems.

A. Collection of bona fide samples

Damer *et al.*[2] recently presented a database with the subjects wearing face masks to explore the FR performance on masked faces, motivated by the current COVID-19 pandemic. This database simulated a collaborative yet varying scenario, e.g., unlocking devices or identity verification at automatic border control gates. The initial version of this database contains 24 participants, while the current number of participants in this paper extends to 47 by further data collection efforts. The data were captured indoors, each at their residence during home-office. Furthermore, each participant was asked to collect data for three days (not necessarily consecutive) and for three different scenarios each day: 1) face with no mask and no additional electric illumination, 2) face with a mask on and no additional illumination, 3) face with a mask on and electric light on. As a result, nine videos are recorded by each participant. In our study, we focus on the impact of masks on PAD performance, while the effect of illumination variation is neglected for now. The bona fide data is divide into two categories: face without a mask on is noted as M0 (3 videos per subject), and face with a mask is marked as M1 (6 videos per subject). It is worth noting that this database simulated a collaborative and varying scenario. The mask types, capture environments, illuminations, and capture devices of each participant are various, only eyeglass is removed.

[1]The CRMA database is not publicly available due to 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. 1: Example bona fide and attack samples in CRMA database. The print (PR) and replay (RE) M2 attacks, including a real part (the mask), are novel and were not addressed in previous works.

B. Creation of the presentation attacks

Several FR databases tried to collect data under various harsh conditions, such as poor lighting, strong occlusion, or low resolution. Such databases tried to reproduce what might happen in the real-world scenario when a legitimate user obtains authorization [30]. On the contrary, the attackers will use highly sophisticated artifacts, such as high-resolution images or videos, to maximize the success rate when trying to impersonate someone. For that reason, our presentation attacks were captured in a windowless room where all lights were on. In addition, 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. Besides, the capture devices and display images/tablets were stationary when collecting data. The videos were captured with 1920×1080 resolution. Also, each video has a minimum length of 5 seconds, and the frame rate is 30 fps. The presentation attack instruments (PAIs) in this database can be roughly grouped into two categories: print attack and replay attack. The attack data in each PAIs (See samples in Fig. 1) are divided into three types: 1) the displayed bona fide samples with no face mask (M0), 2) the displayed bona fide samples with a face mask on (M1), 3) the displayed bona fides with no face mask, but a real mask was placed on it to simulate a participant wearing a mask (M2). However, the face area sizes are slightly inconsistent because the videos were recorded by participants themselves. To reproduce the look of wearing a mask in the real-world, and include mask variations, We cropped five masks to fit most of the faces (See Fig. 3). The details of each PAI present as the following:

- **Print image attack**: In print attack, an attacker tries to fool the FR system using a printed photo. The 35th frame of each video from each participant was printed out for attack because the face in the video tends to stabilize after the first second. Therefore, we obtain nine photos per subject. Then, the above mentioned three tablets were used to capture these 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 contain all areas (100%) of the photos, the second scale consists of most areas (80%) of the original photos, and the third scale focuses on the face area (60%) as much as possible. In addition to solely collecting attack data from printed images, we also collected data from real face masks overlaid on photos (i.e., the previously defined M2). Theoretically, the 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 are generated for each subject, i.e., a total of 4,230 videos for 47 subjects in print PAI.
- **Replay video attack**: In replay attack, 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 inclusion of M2 type data, respectively. The difference is that these tablets were also be used as display devices (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 M0 and M1 groups, 18 videos of M2.), i.e., a total of 8,460 videos in this attack subset.

C. Evaluation protocols

To study the possible effect of face masks and attacks with real masks on the performance of PAD and FR systems, we designed three protocols for further experimental analysis. These three protocols are based solely on the face masks, and in this paper, we disregard other factors, such as types of device, illuminations, and capture scales. However, it is worth mentioning that the CRMA database is also suitable for the further targeted verification of face PAD algorithms under different scenarios. 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). The gender was balanced as much as possible between the these three sets. Tab. II provides more information about three protocols. The detailed description of three protocols are:

- **Protocol 1 (P1)**: The first protocol tries to simulate most of the existing databases, where subject samples wear no face mask. Then, the generalizability of face PAD algorithms is evaluated. The training and development set only contains the videos of M0 (of bona fides and attacks), i.e., videos of M1 (of bona fides and attacks) and M2 (of attacks) are not used for training. In the test set, videos of M0, M1, and M2 will be separately evaluated. In this case, M1 and M2 can be seen as unknown data/attack types.
- **Protocol 2 (P2)**: On the contrary, the second protocol is designed to validate the performance of face PAD algorithms when artifacts of M1 and M2 are learned. Therefore, the training and development set contains

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

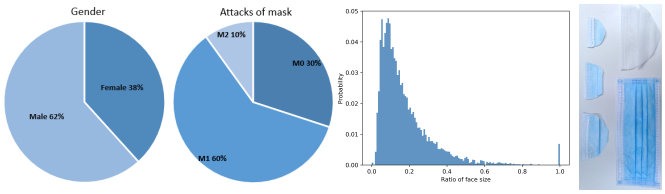


Fig. 3: The statistics of the subjects and the mask shapes of M2 samples in the CRMA database. From left to right: gender, mask types of attacks (M0, M1, M2), the histogram shows the probability distribution of the face size ratio, the applied mask shapes.

all scenarios of videos. The videos in the test set are evaluated separately similarly to P1.

- **Protocol 3 (P3)**: However, until now, the effect of M2 on PAD performance is still unclear. Neural Networks are often considered a black box because we do not know precisely what kind of abstract features it considers. Hence, it is interesting to investigate how a deep learning-based algorithm deals with an unknown attack video containing only partial artifacts. The training and development set in P3 include bona fide and attack videos of M0 and M1, while the test set contains known videos of M0 and M1 and unknown attack videos of M2.

Since the data in this database is video sequences and the numbers of videos between bona fide and attack classes are imbalanced, we sampled 60 frames from a bona fide video and 5 frames from an attack video to reduce the data bias. In addition to the different frames sampling, we also adapt the class weights inversely proportional to class frequencies to reduce the overfitting in the training phase (details can be seen in Sec. IV). In the test phase, the final classification decision is determined by averaging the prediction scores of all sampled frames.

IV. Experiments

This section first describes the adopted face PAD algorithms for the investigation of masks. Later on, several FR algorithms are introduced for further vulnerability analysis, followed by the evaluation metrics. In both PAD and FR experiments,

Protocol	Set	Subjects	Types of mask	# BF (V)	# Attack (V)
P1	Train	1-19	BF: M0, attacks: M0	57	1569
	Dev	20-29	BF: M0, attacks: M0	30	810
	Test	30-47	BF: M0, M1, attacks: M0, M1, M2	162	4860
P2	Train	1-19	BF: M0, M1, attacks: M0, M1, M2	171	5130
	Dev	20-29	BF: M0, M1, attacks: M0, M1, M2	90	270
	Test	30-47	BF: M0, M1, attacks: M0, M1, M2	162	4860
P3	Train	1-19	BF: M0, M1, attacks: M0, M1	171	4617
	Dev	20-29	BF: M0, M1, attacks: M0, M1	90	2430
	Test	30-47	BF: M0, M1, attacks: M0, M1, M2	162	4860

TABLE II: The detailed information of three protocols for exploration the possible effect of face masks. Bona Fide is denoted as BF and V refers to video.

the widely-used Multi-task Cascaded Convolutional Networks (MTCNN) [31] technique is adopted first to detect and crop the face.

A. Face PAD algorithms

A competition [23] was carried out in 2017 to evaluate and compare the generalization performances of face PAD techniques under some real-world variations. In this competition [23], there were 14 participating teams, including the organizers that contributed several state-of-the-art approaches. We chose two methods from them ((as previously discussed in Sec. II)), LBP-based baseline and CPqD, and we included additional solutions. Together, we re-implement a total of seven face PAD algorithms in this study, which can be categorized into three groups: hand-crafted features, deep-learning features, hybrid features. For further cross-database evaluation scenarios, we use three publicly available databases mainly involving 2D PAs (details in Sec. II): CASIA-FAS [12], MSU-MFS [16], and OULU-NPU [7] in the competition. A brief description of the adopted methods are provided below:

- **LBP (baseline):** The LBP is the baseline method [23], which provided by the competition organizers, that utilizes the color texture technique. The face in a frame is first detected, cropped, and normalized into a size of 64×64 pixels. Second, an RGB face is converted into HSV and YCbCr color spaces. Third, the LBP features are 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 is computed by averaging the output scores of all frames.

• **CPqD:** The CPqD is based on the Inception-v3 network [32] and the above LBP baseline. The last layer of the pre-trained Inception-v3 model is replaced by a fully connected layer and a sigmoid activation function. The faces in RGB frames are detected, cropped, and normalized into 299×299 pixels. These face images are utilized as inputs to fine-tune the Inception-v3 model. The model with the lowest EER on the development set among all ten training epochs is selected. A single score for a video is obtained by averaging the output scores of all frames. To further improve the performance, the final score for each video is computed by fusing the score achieved by the Inception-v3 model and the score obtained by the LBP baseline.

• $\mathbf{Inception_{FT}}$ and $\mathbf{Inception_{TFS}}$: Since the CPqD uses the Inception-v3 [32] network architecture as the cornerstone, we also report the results of fine-tuned Inception-v3 model, named $\mathrm{Inception_{FT}}$. In addition to the fine-tunned model, we train the Inception-v3 model from scratch for performance comparison, named $\mathrm{Inception_{TFS}}$. In the training phase, the binary cross-entropy loss function and Adam optimizer with a learning rate of 10^{-5} are used. The output scores of frames are averaged to obtain a final prediction decision for each video.

• $\mathbf{FASNet_{FT}}$ and $\mathbf{FASNet_{TFS}}$: FASNet [24] used transfer learning from pre-trained VGG16 model [33] for face PAD. They used on ImageNet [34] dataset pre-trained VGG16 model as a feature extractor and modified the last fully connected layer. The newly added fully connected layers with sigmoid function were then fine-tuned for the PAD task. This fine-tuned FASNet is referred to $\mathrm{FASNet_{FT}}$, similar to the Inception-v3 network methods, we also train the FASNet from scratch with name $\mathrm{FASNet_{TFS}}$. The input images are the detected, cropped, and normalized RGB face frames with the size of 224 pixels. The Adam optimizer with the learning rate of 10^{-4} is used for training as defined in [24]. To deal with the imbalanced data problem, data augmentation techniques and class weights are utilized. To further reduce overfitting, early stop technique with the patience of 5 and maximum epochs of 30 is used. The resulting scores are averaged to obtain a final score for each video.

• **DeepPixBis:** George *et al.*[25] proposed a densely connected network framework for face PAD with binary and deep pixel-wise supervision. This framework is based on the DenseNet [35] architecture. Two dense blocks and two transition blocks with a fully connected layer with sigmoid activation will produce the binary output. We use the same data augmentation technique (horizontal flip, random jetter 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 the [25] for the training. In addition to data augmentation, we apply class weight and early stopping technique to avoid overfitting. The final score for each video is computed by averaging the scores of frames.

B. Face Recognition algorithms

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

• **ArcFace:** ArcFace [26] introduced an additive angular margin loss function to obtain highly discriminative features for FR. We choose this algorithm because ArcFace consistently outperformed the state-of-the-arts. For example, ArcFace achieved the 99.83% on Labeled Face in the Wild (LFW) [30] and 98.02% on Youtube Faces (YTF) [36] dataset. The pre-trained ArcFace model [2] in our study is based on ResNet-100 [37] architecture and refined on MS-Celeb-1M [38] dataset (MS1M-v2). The output template is a 512-dimension feature vector extracted from '*fc1*' layer of ArcFace.

• **SphereFace:** Liu *et al.*[27] proposed a deep hypersphere embedding approach (SphereFace) for FR task. SphereFace [27] utilized the angular softmax loss for CNNs to learn angularly discriminative features. This method also achieved competitive performance on LFW [30] (accuracy of 99.42%) and YTF [36] dataset (95.00%). Since only 20-layer SphereFace [3] trained on CASIA-WebFace [39] dataset is officially provided, we use the 512-dimension representation extracted from this pre-trained model as a template.

• **VGGFace2:** The first version of VGGFace [40] is based on 16-layer VGG network, while the second version of VGGFace (VGGFace2) [28] adopt ResNet-50 [37] as the backbone architecture. Moreover, VGGFace2 dataset contains 3.31 million images, while initial VGGFace consists of 2.6 million images. Therefore, we use the second version in this study that a ResNet-50 network trained on VGGFace2 dataset [28] [4] for extracting the 512-dimension templates.

The vulnerability of each FR system on M1/M2 attacks is analyzed based on three scenarios. Regardless of which scenario, the references are scenarios-specific bona fide videos captured in the first day, while bona fide videos from the second and third days or attack videos are selected as probes. The three cases including the division of scenario-specific references and probes are described with results in details in Sec. V-C. Once the templates for the face images are obtained, we use the Cosine-similarity as recommended in [26], [27], [28] to compute the similarity scores between references and probes.

C. Evaluation metrics

The metrics following ISO/IEC 30107-3 [41] standardization are used to measure the performance of PAD algorithms: *Attach 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 the attack in a specific scenario. APCER and BPCER

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

reported in the test set are based on a pre-computed threshold in the development set. In our study, we use a BPCER at 10% (on development set) for obtaining the threshold (denoted as $\tau_{BPCER10}$). Additionally, *Half-Total Error Rater* (HTER) corresponding to the half of the summation of BPCER and APCER is used for the cross-database evaluation. Noticeably, we compute a threshold in the development set of the training database. Then, this threshold is used for determining HTER value in the test database. The Detection EER (D-EER) values where APCER and BPCER are equal are also reported in the cross-database scenarios. For further analysis on PAD performance, Receiver Operating Characteristic (ROC) curves are also demonstrated.

To measure the performance of FR techniques, the *Genuine Match Rate* (GMR), referring to the proportion of correctly matched genuine samples, is used at 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, *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 the 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 the similarity scores lower than the $\tau_{FMR@0.01}$ are not matched. Moreover, the recognition score-distribution histograms are shown for more details. Apart from these metrics, the EER value, where FMR equals to FNMR, is computed to compare FR algorithms.

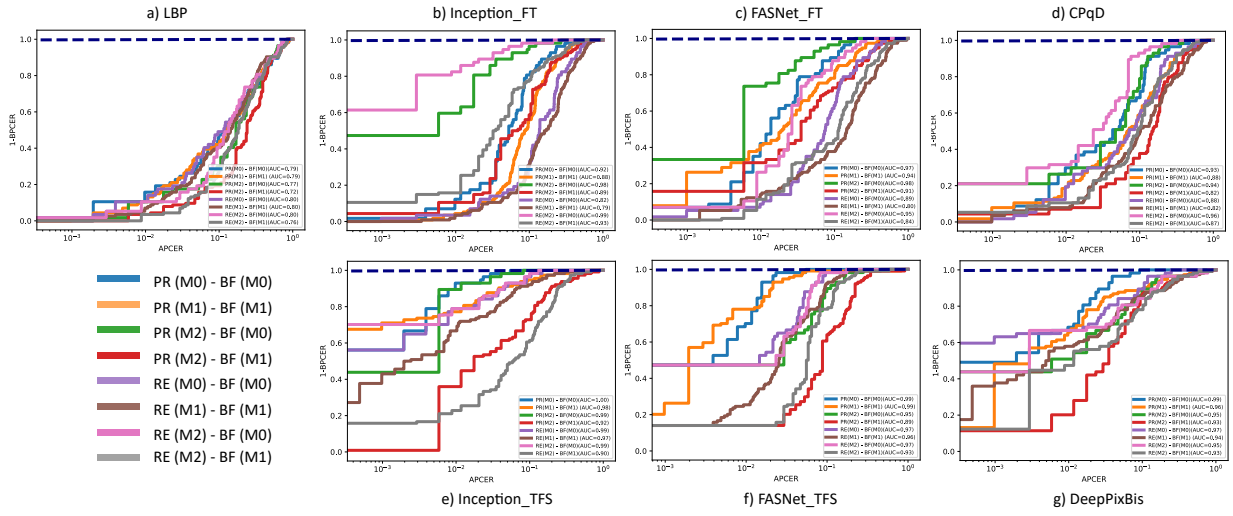
V. Results and Discussion

A. The Intra-database evaluation

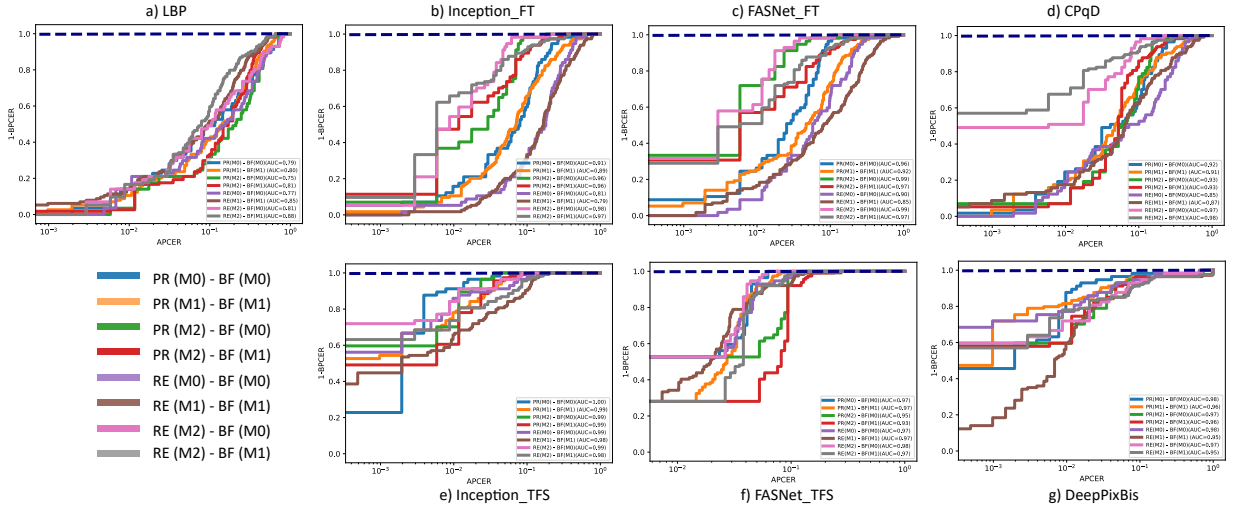
This subsection reports the $\tau_{BPCER10}$ determined APCER and BPCER results in the CRMA database by following the above defined three protocols. Tab. III shows the comparison of seven PAD methods, while Fig.4 presents the ROC curves for each method and each protocol. The observation of each protocol are described below:

- **Experiments in P1:** This protocol emulates the pre-COVID-19 PAD scenarios, in which subjects normally do not wear a mask. Therefore, P1 can be considered the most challenging task due to the unknown testing on M1 and M2 data. As shown in Tab. III, the BPCER values of masked bona fide samples are much higher than unmasked ones, but relatively, most PAD systems achieve higher APCER values on the masked attack samples (either M1 or M2). Moreover, it is interesting to note that networks trained from scratch and the DeepPixBis approach work worse on attack M2 than M1. These observations are consistent with the ROC (Fig. 4). Red curves generated by print-M2 and BF-M1 and grey curves obtained by replay-M2 and BF- M1 possess significantly smaller areas under curves in five of all seven methods. Furthermore, training a network from the first layer boosts the overall performance. Consequently, we rationalise 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 on attack samples.
- **Experiments in P2:** This protocol emulates the known attack scenario where masked bona fide, as well as masked M1 and M2 attacks, are learned in the training phase. We can observe the following points in Tab. III: first, despite that the masked bona fide samples are still more difficult to classify correctly than unmasked ones in most cases, the difference in BPCER between M0 and M1 becomes smaller compared to P1. This indicates that more data is able to improve the performance of the models. This is also consolidated with the observation in ROC curves. In particular, $\mathrm{Inception}_{\mathrm{TFS}}$, $\mathrm{FASNet}_{\mathrm{TFS}}$ and DeepPixBis achieve significant progress (larger areas under curves). Second, six of the seven methods perform worse on the masked printed face (M1 or M2), while five of the seven algorithms show more inferior results on unmasked replay attacks. Moreover, M2 in print PAI achieves higher APCER values than M1 by the training from scratch approaches. One possible reason for the different results between print and replay attacks is the specular reflection. Because attack data were collected in windowless labor with all electric lights on, tablets easily reflect the light than the printed paper and this reflection is difficult to avoid. The face masks also might leak light when placed on an electric tablet, but this does not appear when applied on a printed paper.
- **Experiments in P3:** The third protocol aims to evaluate the generalizability of algorithms on unknown M2 attack. For bona fide samples, we can draw a similar conclusion to P2, but the algorithms perform inconsistently with P1 and P2 on attack samples. In this protocol, the highest APCER values of most PAD algorithms appear on either M1 or M2 attacks in both print and replay PAI. Additionally, the traditional LBP method, $\mathrm{Inception}_{\mathrm{FT}}$, $\mathrm{FASNet}_{\mathrm{FT}}$, and hybrid CPqD method that achieve relatively poor results on M0 or M1 attack samples may have proved to be unable to learn or extract sufficient discriminative features. Moreover, even though the methods learning from the first layer ($\mathrm{Inception}_{\mathrm{TFS}}$, $\mathrm{FASNet}_{\mathrm{TFS}}$ and DeepPixBias) achieve impressive results on M0 and M1 attacks, they generalize not well on unknown M2 attacks. The BPCER values in M0 and M1 are also much higher due to the variations in mask types, illumination, background, and capture sensors.

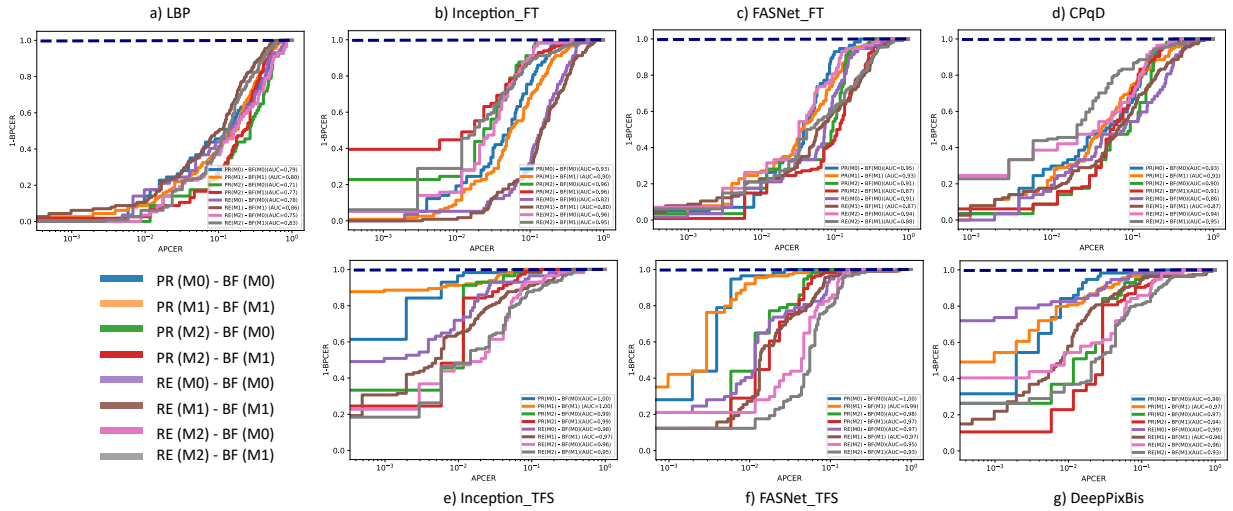
To qualitatively analyze and interpret the deep-learning based methods, Score-Weighted CAM [42] technique is adopted to localize the discriminative areas in face images. The rows from top to bottom correspond to the $\mathrm{Inception}_{\mathrm{FT}}$, $\mathrm{Inception}_{\mathrm{TFS}}$, $\mathrm{FASNet}_{\mathrm{FT}}$, $\mathrm{FASNet}_{\mathrm{TFS}}$, DeepPixBis. Fig. 5a shows the results of P1 (the example subject is in the test set). $\mathrm{Inception}_{\mathrm{FT}}$ mainly focuses on the nose, including nearby partial masks, while $\mathrm{Inception}_{\mathrm{TFS}}$ pays more attention to the upper region of the face. Similarly, $\mathrm{FASNet}_{\mathrm{TFS}}$ reduces the attention on masks and increases the concentration around the forehead. The DeepPixBis concentrates great around the eyes for both M0 and M1 bona fides. However, for attack samples, the attention seems focused on the left eye and partial masks. In general, masks are noticed by all networks. The results of P2



(a) ROC in P1



(b) ROC in P2



(c) ROC in P3

Fig. 4: ROC curves for all PAD methods in three protocols *test* sets. For each protocol and each method, eight curves focusing on masks with Area Under the Curve (AUC) values are plotted to represent the different settings that include PR(M0)-BF(M0), PR(M1)-BF(M1), PR(M2)-BF(M0), PR(M2)-BF(M1) in print attack and RE(M0)-BF(M0), RE(M1)-BF(M1), RE(M2)-BF(M0), RE(M2)-BF(M1) in replay PAI. Red curves (PR(M2)-BF(M1)) and gray curves (RE(M2)-BF(M1)) pose significantly smaller AUC values by most PAD methods on P1. Moreover, $\text{Inception}_{\text{TFS}}$, $\text{FASNet}_{\text{TFS}}$ and DeepPixBis achieve higher AUC values on P2 and P3 than on P1 due to more masked BF/attack data.

Protocol	Method	Threshold @ BPCER 10% in dev set							
		BPCER (%)		APCER (Print) (%)			APCER (Replay) (%)		
		M0	M1	M0	M1	M2	M0	M1	M2
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**

TABLE III: Intra-dataset evaluation. P1: Training on BF-M0, Attack-M0. P2: Training on BF-M0, BF-M1, Attack-M0, Attack-M1, Attack-M2, P3: Training on BF-M0, BF-M1, Attack-M0, Attack-M1.

and P3 for the same subjects can be seen in Fig. 5b and Fig. 5c. We noticed that 1) the attention areas of fine-tuned networks hardly change in three protocols due to the fixed weights of layers before the last classification layer. 2) $\text{Inception}_{\text{TFS}}$ in P2 seems focused on the upper face, including much more eye region than in P1. 3) $\text{FASNet}_{\text{TFS}}$ in P2 concentrates much more on applied real masks than in P3 where training without M2 data. 4) DeepPixBis still works well on bona fide, but for attack samples, its attention seems to be distracted to the edge of images. Despite the fact that DeepPixBis produces correct decisions, this observation raises a serious concern about its reliability and generalizability. This concern is confirmed in the following cross-database evaluation. DeepPixBis obtains generally worse cross-database results than other two training from scratch networks (details see Tab. IV and Tab. V in Sec. V-B). Finally, looking at attention maps in all protocols for this identity, we can find that except for the misclassified samples (with red boxes) that appear on print/replay M0, print M2 attacks are easier to be incorrectly detected as bona fide than M1 attacks.

To further understanding of the above quantitative and qualitative results, we provide additional t-SNE plots for visualize the learned features in the supplementary material. These plots consolidate our findings here that 1) masked bona fide samples are more probable to detected as bona fide by the pre-COVID-19 PAD algorithms. 2) attacks with real masks on presentations are more accessible detected by PAD systems as bona fide than attacks with masked faces.

B. The Cross-database evaluation

In this subsection, we perform cross-database experiments to explore the generalizability of these PAD algorithms on the masked data in the CRMA database. Because the PAIs in the CRMA database are print and replay attacks, we select three popular publicly available databases containing the same PAIs: CASIA-MFS [12], MSU-MFS [16], and OULU-NPU [7] to demonstrate the evaluation. Moreover, two experiments are conducted for cross-database testing. First, the models, trained on the training set of three publicly databases, are evaluated on the test set in the CRMA database. In addition, the results tested on their own test set are also reported (as shown in Tab. IV). Conversely, in the second experiment, the models trained on P1 of the CRMA database are evaluated separately on the test set of these three databases (results in Tab. V). In both cross-database scenarios, We use the $\tau_{BPCER10}$ decision-threshold that computed in the development set of training database as priori to determine the APCER, BPCER and HTER value of the test database.

As shown in Tab. IV, the performance in the cross-database setting is poor for all models in general. Even though deep-learning based methods achieved great results on their own test sets, they generalize significantly worse on masked bona

(a) P1 (b) P2 (c) P3

Fig. 5: Examples for attention maps generated by ScoreCAM of different PAD algorithms and different protocols. The rows from top to bottom on each protocol correspond to the $\text{Inception}_{\text{FT}}$, $\text{Inception}_{\text{TFS}}$, $\text{FASNet}_{\text{FT}}$, $\text{FASNet}_{\text{TFS}}$, DeepPixBis. The columns from left to right on each protocl refers to the BF-M0, BF-M1, PR-M0, PR-M1, PR-M2, RE-M0, RE-M1, RE-M2. The faces with red boxes are misclassified.

Trained on	Method	Threshold @ BPCER 10% in dev set of trained database											
		Tested on same dataset (%)			Tested on our dataset (%)								
		D-EER	BPCER	APCER	BPCER		APCER (Print)			APCER (Replay)			
					M0	M1	M0	M1	M2	M0	M1	M2
CAISA-FASD	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

TABLE IV: Cross-database evaluation 1: trained on three publicly available databases and tested on the CRMA database.

fide samples, e.g., most BPCER values for M1 are close to the 100%. On the contrary, most algorithms achieve lower APCER values of masked M1 and M2 than unmasked M0 attacks. Such results indicate that the model trained on the databases without masked data cannot handle the case of wearing a mask, i.e., cannot fit the ongoing COVID-19 pandemic. A subject with a mask on has a high probability of being detected as an attack by PAD systems, even if this subject is a bona fide. Besides, it is interesting to note that most trained on MSU-MFS models achieve better testing results on the CRMA database than models trained on OULU-NPU database. This may be due to the partially similar statistic information between the MSU-MFS and presented CRMA databases. First, the gender distribution between them is almost the same (38% female and 62% male in CRMA, 37% female and 63% male in MSU). Second, the MSU-MFS database is also collected by a mixture of sensors, including webcam, tablets, and mobile phones, while OULU-NPU used only multiple smartphones. The results of the second cross-database experiment are reported in the Tab. V. It can be seen that trained from scratch networks outperform traditional LBP or fine-tuned networks in most cases. The models trained on P1 of the CRMA database achieved better results than models trained on P2 and P3 in the OULU-NPU database (13.12% HTER on P1, 14.48%, and 17.92% on P2 and P3). We can conclude that even without masked data, the CRMA database still possesses great diversity in sensors

Train	Method	Threshold @ BPCER 10% in dev set of CRMA database					
		CASIA-FASD (%)		MSU-MFSD (%)		Oulu-NPU (%)	
		D-EER	HTER	D-EER	HTER	D-EER	HTER
P1	LBP	41.25	47.19	41.67	40.28	32.50	38.96
	$\text{Inception}_{\text{FT}}$	37.19	42.19	36.81	34.03	24.90	24.06
	CPqD	**36.25**	46.56	34.03	31.94	22.50	23.23
	$\text{FASNet}_{\text{FT}}$	47.50	56.56	40.97	44.44	18.33	18.33
	$\text{Inception}_{\text{TFS}}$	40.00	51.88	36.81	**29.17**	**9.17**	17.40
	$\text{FASNet}_{\text{TFS}}$	48.44	**41.88**	**20.83**	36.81	**9.17**	**13.12**
	DeepPixBis	47.50	49.06	41.67	39.58	21.56	24.23
P2	LBP	46.25	45.63	45.83	44.44	30.83	30.83
	$\text{Inception}_{\text{FT}}$	**34.69**	**34.69**	45.83	42.36	20.83	37.19
	CPqD	40.00	45.63	45.14	47.92	21.67	29.27
	$\text{FASNet}_{\text{FT}}$	50.00	55.94	40.97	39.58	19.27	26.77
	$\text{Inception}_{\text{TFS}}$	37.50	47.50	**25.00**	**30.56**	**13.96**	**14.48**
	$\text{FASNet}_{\text{TFS}}$	47.19	44.38	37.50	31.25	16.67	22.81
	DeepPixBis	46.25	46.88	45.83	38.89	15.94	28.96
P3	LBP	47.50	46.56	45.83	44.44	31.67	32.08
	$\text{Inception}_{\text{FT}}$	**37.50**	**33.44**	41.67	40.28	20.10	33.96
	CPqD	43.75	47.50	45.83	43.75	21.46	28.12
	$\text{FASNet}_{\text{FT}}$	53.75	55.31	41.67	40.97	15.94	**17.92**
	$\text{Inception}_{\text{TFS}}$	42.50	55.62	**25.00**	30.56	15.00	40.73
	$\text{FASNet}_{\text{TFS}}$	46.69	49.38	28.47	**27.78**	**13.96**	23.75
	DeepPixBis	43.75	42.81	42.36	36.81	16.46	28.65

TABLE V: Cross-database evaluation 2: trained on different protocols of the CRMA database and tested on three public databases.

and environments to boost the performance of vanilla models. For example, $\text{FASNet}_{\text{TFS}}$, which train VGG16 from scratch, achieves competitive results (13.12% HTER) on OULU-NPU database [23], [7]. Besides, as mentioned earlier, even though DeepPixBis achieves much better results than other methods in intra-database evaluations, it performs mostly worse than other deep-learning methods, especially cross-testing on MSU-MFSD database.

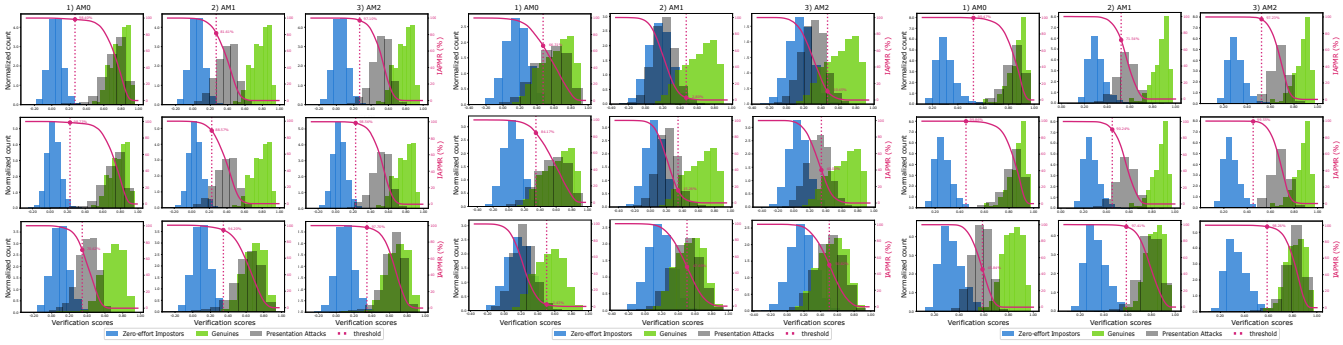
C. The Vulnerability of Face Recognition

The vulnerability of each FR system on M1/M2 attacks is analyzed based on three cases. In the first case (M0-M0), we use the bona fide unmasked samples captured on the first day to enroll subjects in the FR system. Then, the enrolled samples are compared against bona fide M0 samples captured on the second and third day of the same subjects (to compute genuine scores), as well as of other subjects (for zero-effort imposter (ZEI) scores). Once genuine and ZEI comparison scores are obtained, the operating threshold is computed by $\tau_{FMR@0.01}$ threshold. To focus on the effect of masks, we group the print and replay attacks into three categories: AM0 (subjects without mask), AM1 (subjects with the mask on), and AM2(real mask placed on attack presentations). Finally, the probe masked samples of these categories are compared against the enrolled data of the same subjects separately. In the second case (M0-M1), the difference is that bona fide M0 data captured on the second and third day are used to compare against enrolled bona fide M0 samples and then obtain the corresponding genuine and ZEI scores. In the third case (M1-M1), subjects are enrolled in the FR systems using bona fide masked faces captured on the first day. Such enrolled references are also compared against the masked bona fide samples captured on the second and third day to obtain their genuine and ZEI scores.

The performance and vulnerability of each FR system is summarized in Tab. VI. SphereFace [27] obtains relatively low IAPMR values, however, its GMR values are also much lower than ArcFace [26] and VGGFace [28]. In general, the IAPMR values of all three FR systems are roughly close to their GMR values. Specifically, FR systems are vulnerable to the AMO when bona fide M0 samples are used to enroll systems and vulnerable to the AM1/AM2 when M1 data are used as enrollment reference. Comparing the vulnerability analysis results on AM1 and AM2 in all three cases and all FR systems, we note that the IAMPR values of the AM2 are always significantly higher than the IAPMR values of AM1. This indicates that applying real masks on the 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 the histograms of the similarity score distribution in the three scenarios and three FR systems (see Fig 6). In the

References	Attack Probes	ArcFace[26]			SphereFace [27]			VGGFace [28]		
		EER	GMR	IAPMR	EER	GMR	IAPMR	EER	GMR	IAPMR
M0 - M0	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]
M0 - M1	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]
M1 - M1	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]

TABLE VI: The performance and vulnerability of FR systems. The GMR and IAPMR values have been computed based on the $\tau_{FMR@0.01}$ threshold. 95% confidence intervals for the IAPMR values are shown in brackets.



(a) ArcFace (b) SphereFace (c) VGGFace

Fig. 6: The similarity score distributions by off-the-shelf FR networks: ArcFace [26], SphereFace [27], VGGFace [28]. The rows from top to bottom represent three scenarios: M0-M0, M0-M1, M1-M1 as shown in Table VI. The columns from left to right in each subplot refers to AM0, AM1, AM2. The green, blue, and grey correspond to the similarity scores of bona fide, zero-effort imposter, attacks. The red curve describe the IAPMR values. In each plots, it can be seen that grey histograms have more overlaps with green in AM2 than in AM1. This indicates that attacks with real masks placed on presentations are more difficult to detect correctly by FR systems than attacks with masked faces.

histograms, green refers to genuine scores, blue presents ZEI scores, and grey presents attack verification scores. The ideal situation is that no overlap between the green and the other two histograms. Fig. 6a shows the score distributions of ArcFace [26], where the rows from top to bottom present M0-M0, M0-M1, M1-M1 cases and columns from left to right refer to AM0, AM1, and AM2. It can be seen that 1) the verification scores of attacks are higher than scores of ZEI in all cases. 2) the scores of AM0 attacks and genuine scores almost overlap in the M0-M0 and M0-M1 scenarios, while the scores of AM1/AM2 attacks have a lot of overlapping areas with genuine scores in the M1-M1 scenario. 3) for all cases, the scores of AM2 has more overlaps with genuine scores than AM1. Similar observations can be found in the Fig. 6b for SphereFace and Fig. 6c for VGGFace. These observations are consistent with the findings in the previous Tab. VI.

Overall, these results indicate that 1) a user wearing the face mask has a high probability of being detected as the attacker by the current face PAD systems. 2) attacks with applied masks are slightly difficult to detect correctly than attacks with masked faces when using the deep-learning based PAD methods. 3) FR systems pose a higher vulnerability for attacks with real masks than a masked face.

VI. Conclusion

In this work, we presented a new large-scale face PAD database, **CRMA**, including the novel real mask attacks and masked face attacks. It consists of 13,113 high-resolution videos and has a great diversity in capture sensors, displays, and capture scales. To study the effect of the masked attack on PAD algorithms, we designed three experimental protocols. The first protocol measures the generalizability of the current PAD algorithms on unknown masked bona fide and M1/M2 attack data, while in the second protocol, masked data are included in the training phase to compare PAD performances. Additionally, the third protocol aims to evaluate the generalizability of algorithms only on unknown PA that real mask attacks (M2 attacks). The extensive experiments were conducted on these protocols and under intra- and cross-database scenarios. The results showed that pre-COVID-19 PAD algorithms have a high possibility of detecting masked bona fide samples as attackers. Furthermore, the vulnerability of FR systems on masked attacks was analyzed. The experiments demonstrated that all state-of-the-art FR systems are more vulnerable to attacks with real masks placed on presentations (M2) than attacks with masked faces (M1).

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.

REFERENCES

[1] M. L. Ngan, P. J. Grother, and K. K. Hanaoka, "Ongoing face recognition vendor test (frvt) part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms," 2020.

[2] N. Damer, J. H. Grebe, C. Chen, F. Boutros, F. Kirchbuchner, and A. Kuijper, "The effect of wearing a mask on face recognition performance: an exploratory study," in *BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, online, 16.-18. September 2020*, ser. LNI, A. Brömme, C. Busch, A. Dantcheva, K. B. Raja, C. Rathgeb, and A. Uhl, Eds., vol. P-306. Gesellschaft für Informatik e.V., 2020, pp. 1–10.

[3] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic," *Measurement*, vol. 167, pp. 108 288 – 108 288, 2020.

[4] Y. Li, K. Guo, Y. Lu, and L. Li, "Cropping and attention based approach for masked face recognition," *Applied Intelligence*, 2021.

[5] W. Wan and J. Chen, "Occlusion robust face recognition based on mask learning," in *2017 IEEE International Conference on Image Processing, ICIP 2017, Beijing, China, September 17-20, 2017*. IEEE, 2017, pp. 3795–3799.

[6] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 773–782.

[7] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and 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 2017, Washington, DC, USA, May 30 - June 3, 2017*. IEEE Computer Society, 2017, pp. 612–618.

[8] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 389–398.

[9] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4680–4689.

[10] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu, "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12357. Springer, 2020, pp. 70–85.

[11] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds., vol. 6316. Springer, 2010, pp. 504–517.

[12] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *5th IAPR International Conference on Biometrics, ICB 2012, New Delhi, India, March 29 - April 1, 2012*, A. K. Jain, A. Ross, S. Prabhakar, and J. Kim, Eds. IEEE, 2012, pp. 26–31.

[13] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group, Darmstadt, Germany, September 6-7, 2012*, ser. LNI, A. Brömme and C. Busch, Eds., vol. P-196. GI, 2012, pp. 1–7.

[14] E. Nesli and M. Sébastien, "Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect," 2013.

[15] I. Chingovska, N. Erdogmus, A. Anjos, and S. Marcel, "Face recognition systems under spoofing attacks," in *Face Recognition Across the Imaging Spectrum*, T. Bourlai, Ed. Springer, 2015.

[16] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 4, pp. 746–761, 2015.

[17] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li, "CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing," *IEEE Trans. Biom. Behav. Identity Sci.*, vol. 2, no. 2, pp. 182–193, 2020.

[18] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing deep face recognition with custom silicone masks," in *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018*. IEEE, 2018, pp. 1–7.

[19] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[20] J. Kannala and E. Rahtu, "BSIF: binarized statistical image features," in *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, November 11-15, 2012*. IEEE Computer Society, 2012, pp. 1363–1366.

[21] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *2011 International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–7.

[22] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*. IEEE, 2015, pp. 2636–2640.

[23] 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, and A. Hadid, "A competition on generalized software-based face presentation attack detection in mobile scenarios," in *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*. IEEE, 2017, pp. 688–696.

[24] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. de Alencar Lotufo, "Transfer learning using convolutional neural networks for face anti-spoofing," in *Image Analysis and Recognition - 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5-7, 2017, Proceedings*, ser. Lecture Notes in Computer Science, F. Karray, A. Campilho, and F. Cheriet, Eds., vol. 10317. Springer, 2017, pp. 27–34.

[25] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*. IEEE, 2019, pp. 1–8.

[26] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4690–4699.

[27] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 6738–6746.

[28] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising face across pose and age," in *International Conference on Automatic Face and Gesture Recognition, 2018.*, 2018.

[29] A. Mohammadi, S. Bhattacharjee, and S. Marcel, "Deeply vulnerable: a study of the robustness of face recognition to presentation attacks," *IET Biom.*, vol. 7, no. 1, pp. 15–26, 2018.

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

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

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las*

Vegas, NV, USA, June 27-30, 2016. IEEE Computer Society, 2016, pp. 2818–2826.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[34] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 248–255.

[35] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2261–2269.

[36] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*. IEEE Computer Society, 2011, pp. 529–534.

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

[38] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9907. Springer, 2016, pp. 87–102.

[39] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *CoRR*, vol. abs/1411.7923, 2014.

[40] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, vol. 1. BMVA Press, 2015, pp. 41.1 – 41.12.

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

[42] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*. IEEE, 2020, pp. 111–119.

Supplementary material

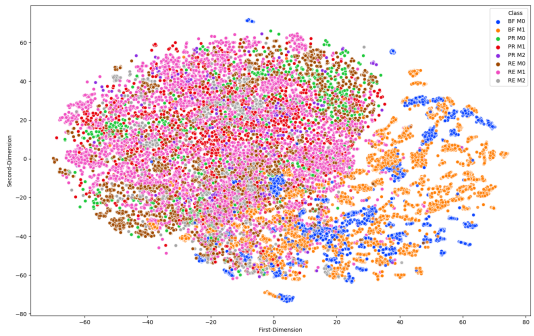
Paper title: Real Masks and Fake Faces: On the Masked Face Presentation Attack Detection

To further observe and deeper understand the discriminative features between bona fide and PAs in the CRMA database, we present here, as supplementary material, the visualized features of face samples from seven classes: BF-M0, BF-M1, PR-M0, PR-M1, PR-M2, RE-M0, RE-M1, RE-M2. Here, BF refers to bona fide, while PR and RE correspond to print and reply. Fig. 8 shows the differential results of fine-tuned $\mathrm{Inception}_{\mathrm{FT}}$ and $\mathrm{FASNet}_{\mathrm{FT}}$, and trained from scratch $\mathrm{Inception}_{\mathrm{TFS}}$ and $\mathrm{FASNet}_{\mathrm{TFS}}$ on the first protocol. Furthermore, the results of DeepPixBis on three experimental protocols are selected to present in Fig. 7 as the DeepPixBis method outperforms other PAD algorithms in the CRMA database.

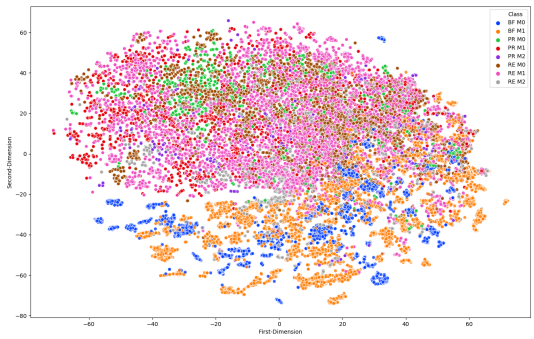
First, by observing the t-SNE plots from $\mathrm{Inception}_{\mathrm{FT}}$ and $\mathrm{FASNet}_{\mathrm{FT}}$ methods in Fig. 8, we can find that the unmasked bona fide, unmasked print attacks and unmasked replay attacks are tightly grouped, while the masked bona fide and masked M1 samples in both PAIs are clustered. Simultaneously, the print M2 and replay M2 attack samples tend to form a compact distribution. Moreover, it is interesting to note that M2 attacks are enclosed by the previous M0 BFs/PAs and M1 BF/PAS clusters. This indicated that 1) fine-tuned networks are unable to learn the discriminative features between bona fide and attacks. 2) M2 attacks, which have artifacts and live features together, are more difficult for networks to make a correct PAD decision. Second, it can be seen in the t-SNE subplots from $\mathrm{Inception}_{\mathrm{TFS}}$ and $\mathrm{FASNet}_{\mathrm{TFS}}$ that compared to fine-tuned networks, trained from scratch networks perform better. Because the features of bona fide M0 and M1 are in a group, print attacks in a group, and replay attacks in a group. However, 2D features of M2 attacks in print and replay PAIs are still closer than other attack types when employing $\mathrm{FASNet}_{\mathrm{TFS}}$. Overall, M2 attacks are slightly difficult to detected correctly than M1 attacks by trained from scratch networks, even though the bona fide and attacks are more separate than using fine-tuned networks. Third, by taking a look at results on P1 by DeepPixBis method in Fig. 7a, almost half of the bona fide M1 data is mixed with attack data. Besides, on the P1 plot, print M0 data is close to the replay M0, print M1 is close to the replay M1, and grouped M2 attacks in both PAIs. It should be noticed again that M2 attacks are surrounded by the bona fide and other attack data. In Fig. 7b and Fig. 7c, bona fide and attack data are more separate. However, some replay M2 attacks are still mixed inside bona fide groups on the P2. Additionally, more M2 attacks are close to the bona fide samples on the P3 than P2. A possible reason is that M2 attacks are not learned in the training phase of P3. All the above findings are consistent with observations in our main work.

Together these results provide important insights that 1) fine-tuned networks have poorer generalizability on the unknown masked bona fide and attack data than trained from scratch networks. 2) masked bona fide samples are more probable to detected as bona fide by the pre-COVID-19 PAD algorithms. 3) attacks with real masks on presentations (M2) are more accessible detected by PAD systems as bona fide than attacks with masked faces (M1).

(a) Feature visualization of DeepPixBis method in P1

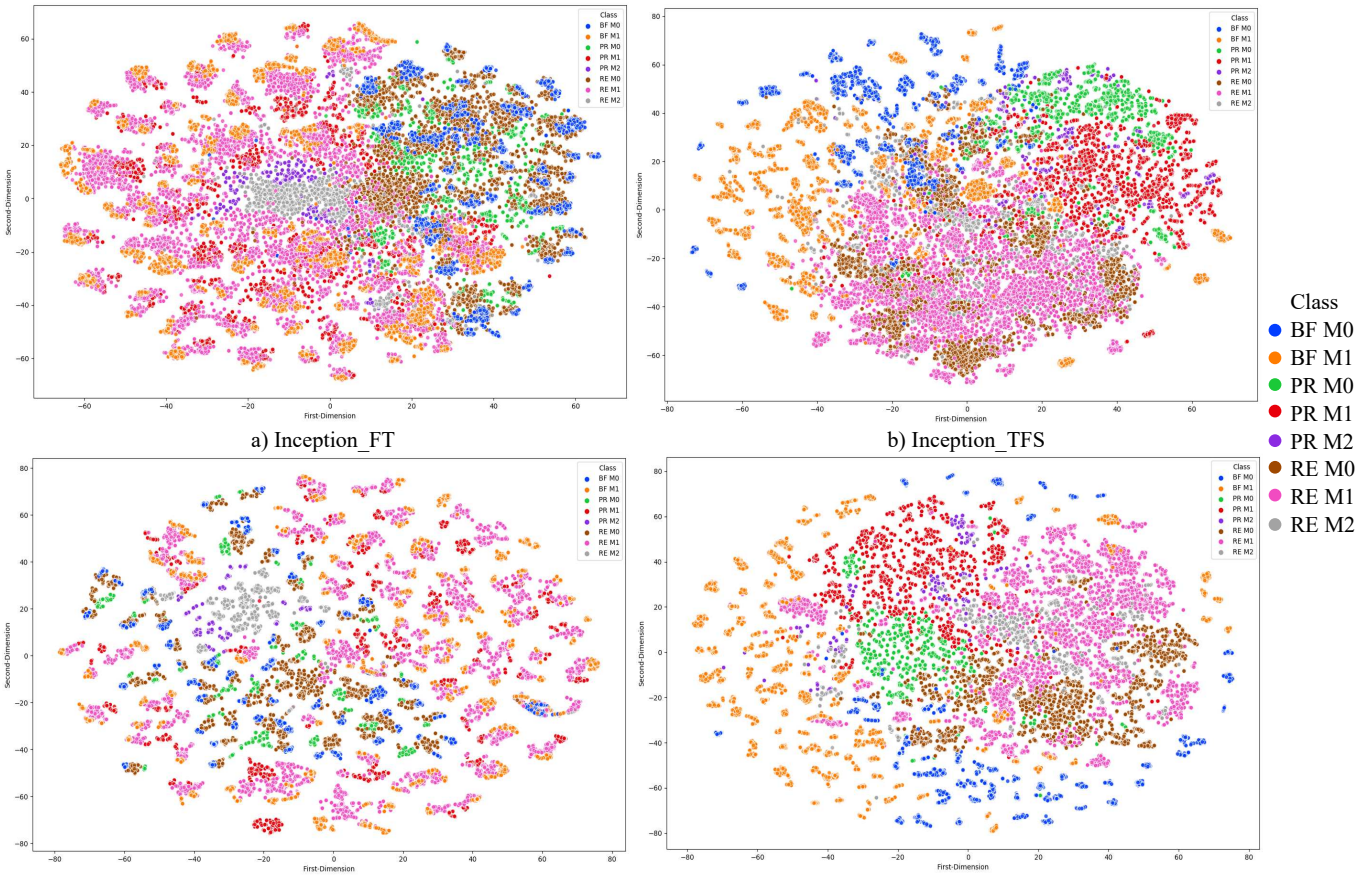


(b) Feature visualization of DeepPixBis method in P2



(c) Feature visualization of DeepPixBis method in P3

Fig. 7: t-SNE plots for DeepPixBis method on three protocols in our CRMA database. It can be seen that the unknown M2 attacks are enclosed by bona fide samples and other types of attacks.



a) Inception_FT b) Inception_TFS

c) FasNet_FT d) FasNet_TFS

Fig. 8: The t-SNE plots of fine-tuned $\mathrm{Inception}_{\mathrm{FT}}$ and $\mathrm{FASNet}_{\mathrm{FT}}$ (in left column), and trained from scratch $\mathrm{Inception}_{\mathrm{TFS}}$ and $\mathrm{FASNet}_{\mathrm{TFS}}$ (in right column) on the first protocol which targets the unknown masked bona fide and two types of masked attacks. These plots show that fine-tuned $\mathrm{Inception}_{\mathrm{FT}}$ and $\mathrm{FASNet}_{\mathrm{FT}}$ cannot discriminate the features between bona fide and attacks. The 2D features representing the face images seem to be grouped based on the existence of face masks. Moreover, M2 attacks are surrounded by bona fide samples and other types of attacks.