2021 IEEE International Joint Conference on Biometrics (IJCB) | 978-1-6654-3780-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCB52358.2021.9484359

Face Liveness Detection Competition (LivDet-Face) - 2021

Sandip Purnapatra[1] †, Nic Smalt[1], Keivan Bahmani[1], Priyanka Das[1], David Yambay[1], Amir Mohammadi[2], Anjith George[2], Thirimachos Bourlai[3], Sébastien Marcel[2], Stephanie Schuckers[1],
[1]Clarkson University, USA, [2]Idiap Research Institute, Switzerland, [3]University of Georgia, USA,
†purnaps@clarkson.edu

Meiling Fang*[4], Naser Damer*[4], Fadi Boutros*[4], Arjan Kuijper*[4], Alperen Kantarci *[5], Başar Demir *[5], Zafer Yildiz *[5], Zabi Ghafoory *[5], Hasan Dertli *[6], Hazım Kemal Ekenel *[5], Son Vu*[7], Vassilis Christophides*[7], Liang Dashuang*[8], Zhang Guanghao*[8], Hao Zhanlong*[8], Liu Junfu*[8], Jin Yufeng*[8], Samo Liu*[9], Samuel Huang*[9], Salieri Kuei*[10], Jag Mohan Singh*[11], Raghavendra Ramachandra*[11],
[4]Fraunhofer Institute for Computer Graphics Research IGD, Darmstadt, Germany,
[5]SiMiT Lab, Istanbul Technical University, Istanbul, Turkey, [6]Sodec Apps, Istanbul, Turkey
ETIS lab, Paris, France, [7]CY Cergy Paris Universite , ENSEA,
ETIS lab, Paris, France, [8]Vision Intelligence Center of Meituan, China, [9]FaceMe, Taiwan,
[10]CLFM, Taiwan, [11]Norwegian University of Science and Technology, Gjøvik, Norway
*Competitors

Abstract

*Liveness Detection (LivDet)-Face is an international competition series open to academia and industry. The competition's objective is to assess and report state-of-the-art in liveness / Presentation Attack Detection (PAD) for face recognition. Impersonation and presentation of false samples to the sensors can be classified as presentation attacks and the ability for the sensors to detect such attempts is known as PAD. LivDet-Face 2021 * will be the first edition of the face liveness competition. This competition serves as an important benchmark in face presentation attack detection, offering (a) an independent assessment of the current state of the art in face PAD, and (b) a common evaluation protocol, availability of Presentation Attack Instruments (PAI) and live face image dataset through the Biometric Evaluation and Testing (BEAT) platform. The competition can be easily followed by researchers after it is closed, in a platform in which participants can compare their solutions against the LivDet-Face winners.*

1. Introduction

Face recognition systems are widely used for human identity recognition across the government, and the industry for various reasons, including, but not limited to, ease of use, convenience and competitive accuracy across other biometric modalities. Despite the high accuracy of current face recognition algorithms, the overall reliability of face recognition systems depends also on their capability to detect presentation attacks, a process also known as Presentation Attack Detection (PAD) [8]. Popular presentation attacks include printed face photos, replay face videos and face masks, which demonstrate a security risk for unattended face recognition systems [14]. Often presentation attacks are carried out with malicious motives, such as concealing the real identity, impersonating the real identity and enrolling a virtual identity in a face recognition system [9]. These challenges can be mitigated with hardware and software-based presentation attack detection (PAD) systems. An ideal system should accept all the genuine or live samples being presented and reject all the impersonation attempts successfully. To achieve that goal, some software-based PAD algorithms perform binary-classification between attack & bona-fide presentations and employ popular machine learning techniques, which depend on hand-crafted features like local binary patterns (LBP), phase quantization and histograms of oriented gradients. More recently deep learning-based algorithms have been utilized which are based on convolutional neural networks (CNNs) trained on many live and presentation attack (PA) examples [13] [7] [6]. Our literature survey indicates that both handcrafted and deep neural network-based ap-

*https://face2021.livdet.org/

978-1-6654-3780-6/21/$31.00 ©2021 IEEE

proaches yield high classification performance for correctly identifying Presentation Attack Instruments (PAIs) when the performance of these algorithms is tuned with known PAIs. However, these algorithms have certain drawbacks and often fail to detect unknown PAIs and more challenging or sophisticated presentation attacks. Continuous efforts are necessary to update PAD algorithms to detect rapidly evolving presentation attacks.

LivDet-Face is an international competition and the first face liveness detection competition of the LivDet series to access the state-of-the-art in face PAD with independent evaluation of the submitted algorithms on unseen face presentation attacks.

The most significant contributions of this paper and the LivDet-Face 2021 competition are:

- A report on the present state-of-the-art in face PAD based on independent testing of **eleven algorithms submitted** to the competition organizers for both **image category and video category**
- Dataset prepared in accordance to Fast ID Online (FIDO) Biometric Requirements [17] and all algorithms were evaluated by standard PAD metrics as defined by International Organization for Standardization (ISO) [10]
- The largest spectrum of PAIs used till date, to the best of our knowledge, in all face PAD competitions- **nine different PAIs** constitute the test dataset for LivDet-Face with each category captured with four different sensors.
- Introduction of **three novel PAIs**: high-quality 3D-mask, flexible 3D silicon masks and video display sample of live subjects
- **Initiation of LivDet-Face Competition**, i.e., the competition benchmark will be available to all researchers through the BEAT platform [1] after the competition is concluded, to allow testing of all future algorithms with LivDet-Face 2021 protocol, **without revealing the test data** .

2. Performance Evaluation Metrics

LivDet-Face 2021 employs two basic PAD metrics for evaluation which follows the recommendations of ISO/IEC 30107-3 [10]:

- **Attack Presentation Classification Error Rate (APCER)**, the proportion of attack presentations of the same PAI species incorrectly classified as bonafide presentation, *i.e*. PAI classified as live, and
- **Bonafide Presentation Classification Error Rate (BPCER)**, the proportion of bonafide presentations classified as attack presentations, *i.e*. live classified as PAI.

Both the APCER and BPCER metrics are used to evaluate the algorithms. ISO also recommends using the maximum value of APCER when multiple PA species (or categories) are present in case of system-level evaluation, which is primarily designed for industry applications. For this competition, our goal is to consider the detection of all PAIs, and not to rank the algorithms submitted by the competitors from the worst- to the best-performing PA. Thus, in the LivDet-Face 2021 competition, we evaluated performance based on weighted average of APCER over all PAIs:

- **Weighted Average of APCER** ($APCER_{average}$), which is the average of APCER over all PAIs and weighted by the number of samples in each PAI category, as reported in Table 1.
  For the **purpose of competition ranking**, the Average Classification Error Rate (ACER) was computed to select the best performer
- **Average Classification Error Rate (ACER):** the average of $APCER_{average}$ and BPCER.

Note that ACER has been deprecated in ISO/IEC 30107-3:2017 [10] in the industry-related PAD evaluations.

3. Face PAD efforts in last five years

Our review of the literature of the Facial PAD competitions suggest a wide range of software and hardware-based solutions. Two of the literature surveys in this area highlight the most recent state-of-the-art in facial-PAD evaluation [14] [9]. In this section we have summarized the known facial-PAD competitions for the last five years.

3.1. CelebA-Spoof Challenge

CelebA-Spoof Challenge 2020 was an algorithm-based competition, organized to boost research on face anti-spoofing. The CelebA-Spoof dataset offers 625,537 images collected from 10,177 subjects with various sensors and different lighting conditions, however sophisticated high level PAIs were not part of the competition dataset. In the test set of the competition, there were less variety and level of PAIs. The competition had a total of 19 competitors, however, the publication mentions the results of five competitors [22]. The organizers evaluated the performance of the submitted algorithms with True Positive Rate (TPR) for three different levels of False Positive Rate (FPR) - 10^{-3}, $5*10^{-3}$ and 10^{-6}. The best TPR was 100% for all FPR and 98% for FPR=10^{-6}.

The main difference of this competition with LivDet-Face 2021 is that the LivDet competition have used standard PAD evaluation metrics defined by ISO [10], dataset constitute more variable and higher quality PAIs to evaluate the competition results; in addition LivDet-Face was organized for both image and video categories.

3.2. Generalized Software-based Face Presentation Attack Detection in Mobile Scenarios

The main objective of this competition, held in 2017, was to evaluate and compare the performance of mobile face PAD algorithms under real-world variations. The competition training dataset included a total 4950 real and fake access videos, collected from front-facing cameras of six different smartphones with two different levels of PAIs used for testing. The performance of the competitive algorithms were tested with four different PAD protocols. Protocol-I was designed to evaluate the performance of the algorithms with unseen environmental, illumination and backgrounds. Protocol-II evaluated the performance for the PAIs created with different printers or displays. Protocol-III evaluated the performance in a sensor interoperability scenario, where the algorithms were trained with the videos collected from five smartphones and tested with the video collected using the rest of the smartphones. Protocol-IV evaluated the performance of the algorithms simultaneously with the previous three protocols. The best performance of the algorithms were for protocol-II, with ACER performance equal to 2.5% [2].

In comparison, LivDet-Face 2021 did not share any training dataset with the competitors. The dataset of the LivDet-Face 2021 competition also included high-quality PAIs including sophisticated level-C PAIs i.e. different variety of 3D face masks. LivDet Face dataset was not limited to mobile PAD scenarios as the dataset was collected using high-quality smartphone and DSLR camera.

3.3. ChaLearn Face Anti-spoofing Attack Detection Challenge

Held in 2019, this competition leveraged on the publicly available face anti-spoofing dataset - CASIA-SURF [21] with 21,000 videos from 1,000 subjects and each sample having 3 modalities- RGB spectrum, Depth and InfraRed. All 3 modalities were used with the motivation about how to fuse the complementary information between the three modalities. The competition reported on thirteen algorithms with the winning team performing with an APCER of 0.0074% and Normal Presentation Classification Error Rate (NPCER) of 0.15%. Unlike LivDet-Face, this competition made the dataset for training, validation and unlabelled test data available to the participants.

LivDet Face 2021 differed from this competition in terms of data and dataset- LivDet Face 2021 considered only RGB images; the training/ testing data was not available to the participants- replicating a more real life challenge scenario.

The PAI types, levels and evaluation metrics used in the above mentioned competitions are different from the LivDet-Face 2021 competition. The LivDet competition added three novel PAI types, more higher quality PAIs and a variety of sophisticated PAI types in the test database. The test database was prepared conforming with the FIDO PAD test procedures [17] to make the testing scenario more challenging and standardized in comparison to the other competitions.

3.4. LivDet-Face 2021

The LivDet-Face 2021 competition is the first LivDet competition on face PAD and is co-organized by three institutes, namely: the Clarkson University (USA), the Idiap Research Institute (Switzerland) and the University of Georgia (USA). Previously, LivDet has organized liveness detection competition for fingerprint and iris, more details can be found in [12]. The objective of the competition was to evaluate the performance of the state-of-the-art facial PAD detection algorithms against traditional and novel PAIs. The competition had two categories: *Image*, and *Video*. Competitors were given the chance to participate in both image and video category of the competition. International academic and industrial institutions were encouraged to participate in the competition. For the LivDet-Face 2021 competition no official training dataset was offered – the competitors were free to use any proprietary and/or publicly available data to train their algorithm. The LivDet-Face 2021 competition focused on the evaluation capabilities of the state-of-the-art algorithms **to generalize to uncertain circumstances**.

For both Image and Video category of the competition there were nine PAI types i.e. laptop display, photo mask, low and high-quality paper display, video display of live subjects, low, medium and high-quality 3D masks and wearable and 3D silicon masks. While, at least two samples of majority of the competition PAIs for both the categories were shared with the competitors as a validation dataset to fine-tune their algorithms, the overall test samples and two of the PAI types i.e. high-quality 3D masks and video display of live subjects, were not revealed to the competitors. The performance of the algorithm for each sample were determined by a output score ranging between 0 to 100 with a threshold of 50. A score of 1000 indicates undetected samples. Test samples with scores less than 50 were classified as PAI and scores of 50 and above were classified as live. Most of the competitors normalized the score outputs at their end and provided scores as a 0, 100 or 1000 (if undetected) based on their detection. If the submitted algorithms provided a score of 1000 for the PAIs then it was considered as a correct decision as the algorithm was able to reject PAIs and is not considered as an attack presentation classification error. A score of 1000 for bonafide samples were considered incorrect and was included as part of BPCER calculation. All of the evaluations reported in this publication were completed by

the competition organizers and were *not* self-reported by the competitors.

4. Experimental Protocol

4.1. LivDet-Face 2021 competition participation

International academic and industrial organizations were welcome to participate anonymously or non-anonymously in LivDet-Face 2021 competition. Non-anonymous competitors were given the opportunity to participate in the publication as co-authors. A total of thirty teams registered for the competition from across the globe. The organizers received a total of ten submissions for the image category and six submissions for the video category. Among the image category submissions, six could be successfully tested and for the video category submissions, five could be successfully tested by the organizers. Unsuccessful tests were due to software issues and the reasons were communicated with the competitors.

4.2. Dataset for LivDet-Face 2021

Training dataset For LivDet-Face 2021 competition, no official training dataset was shared by the organizers. Instead, the organizers were encouraged to use any data available to them (both from public and proprietary sources) to train their algorithms successfully. Additionally, the competition organizers shared two or three examples of the known PAIs to familiarize the competitors with the test dataset. The rest of the samples of the disclosed PAI types were considered as unknown for the competitors.

Test dataset The testing dataset used in this competition was a combination of the data from two of the organizers: Clarkson University (CU) and Idiap Research Institute. The dataset consisted of 724 images (135 live and 589 PAI samples) for image category and 814 videos (125 live and 689 PAI samples) and were collected using overall five different sensors (DSLR, iPhone X, Samsung Galaxy S9, Google Pixel, Basler aA1920-150uc) from overall 48 live subjects, as summarized in 1. The video lengths of the test dataset were up to six seconds. Eight PAIs for image category and nine PAIs for video category were included in the dataset:

- **Paper Displays:** A total of 100 low-quality paper and 100 high-quality photo paper images were collected for image category and a total of 100 low-quality paper and 100 high-quality photo paper videos were collected from 25 live subjects using four different sensors.
- **Laptop Display:** 100 samples of laptop screen displays for both the competition categories were collected from 25 live subjects using four different sensors.
- **2D Photo Masks:** The portion of the eyes of the high-quality photo paper face images were cut out and put on a subject's face like a mask. A total of 100 samples for image category and 100 samples for video category were collected from 25 live subjects using four different sensors.
- **3D Masks:** Photographs of the front and sides of a live subject were used to make a software-based 3D model of the face and masks were printed using 3D printers. Based on the printing quality, three different qualities of 3D masks (low, medium and high), included in the test dataset for both competition categories. The low-quality masks were created using a 3D volumetric regression model [11] and only require one frontal image. However, these masks have less prominent facial features. The medium-quality masks were created using three live images (one frontal and two sides) through the *FaceGen Modeller software and have moderately better and more prominent facial features than the low-quality 3D masks [4]. Finally, the high-quality 3D masks were crated by researchers using professional 3D stitching software and using 60 images of a live subject from different angels. The high-quality masks have very prominent or life-like facial features. A total of 24 images and 24 videos of low-quality 3D masks were created from six live subjects and a total of 12 medium-quality 3D mask images and 12 videos, created from three live subjects were included in the test dataset. The high-quality 3D masks were kept as an unknown PAI type from the competitors and the test dataset had 12 high-quality 3d mask images and 12 videos, created from three live subjects in the test dataset. The face masks were created from six live subjects and images were collected with four different sensors.*
- ***Silicon Masks** A total of 141 image and video samples of the wearable and 3D silicon masks were collected using five different sensors (DSLR, iPhone X, Samsung Galaxy S9, Google Pixel and Basler aA1920-150uc).*
- ***Video Display** A total of 100 video display samples were collected from 25 live subject's videos using four different sensors, for the video category competition. Replays of live videos where subjects were blinking or moving their heads, were collected. These videos were used as unknown PAI for the video category of the competition and were not part of the validation dataset which were shared with the competitors.*

4.3. LivDet-Face 2021 Competition Algorithms

For the LivDet-Face 2021 competition, there were six teams competing for the image category and five for the video category of the competition. All competitors were provided with the opportunity to present their results anonymously to this competition. The competing teams were given the opportunity to submit the description of their submitted algorithm. The descriptions are provided below.

Fraunhofer IGD: Team Fraunhofer IGD submitted their

(a) Photo Mask (b) Laptop Display (c) Low Quality Paper Display (d) High Quality Paper Display

(e) Low Quality 3D Mask (f) Medium Quality 3D Mask (g) High Quality 3D Mask (h) Silicon Mask

Figure 1: Example images of all presentation attack types present in the LivDet-Face 2021 test dataset.

Table 1: Test Dataset Summary

Class	Types of PAIs	Total Images	Total Videos	Sensors
Live	-	135	125	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel
PAI	Laptop Display (DL)	100	100	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel
PAI	Photo Mask (PM)	100	100	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel
PAI	Low-Quality Paper Display	100	100	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel
PAI	High-Quality Paper Display	100	100	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel
PAI	Low-Quality 3D Mask	24	24	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel
PAI	Medium-Quality 3D Mask	12	12	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel
PAI	High-Quality 3D Mask	12	12	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel
PAI	Silicon Mask	141	141	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel, Basler acA1920-150uc
PAI	Video Display (VD)	-	100	DSLR, iPhone X, Samsung Galaxy S9, Google Pixel

algorithm for both image and video category of the competition. Their algorithm adopts three different detection strategies trained on multiple groups of databases. This detector ensemble approach is unified with Fisher-discriminative ratio (FDR) weights [3] to achieve the face presentation attack detection decision. The first model is a strategy based on the DeepPixBis [6] under pixel-wise supervision, and the second strategy is based on an off-the-shelf ResNetXt network [19] trained on ImageNet for simple binary classification. The face is detected by MTCNN [20] and then resized to 224x224 as inputs of these two models. The third strategy is based on a lightweight model that takes advantage of the difference between bona-fide and attack in the frequency domain. Fast Fourier Transform is used to calculate the discrete Fourier Transform of the input face image. The result along with the face image is then fed to a lightweight model and produces the binary decision. To make the solution more robust, the team used the FDR to weigh the 12 models. In the training process, the models are trained with a maximum of 25 epochs and the data are re-sampled to keep the bona-fide-attack ratio as close to 1:1. Several augmentation techniques, i.e. horizontal flip, rotation, shift, and cutout, were used to avoid over-fitting. The decision threshold of 50 is set by using the attacks and bona-fide samples in the Real Mask Attack Database (CRMA) [5] as a development unknown (not used in the training) data.

The algorithm submitted for the video category is the same as the image-based algorithm, where a final result of a video is a fusion of the results of multiple frames. Up to a hundred frames are picked from each video and each of these frames is analyzed as a single image as described above, resulting in a PAD score. The resulting PAD scores are fused in a simple mean-rule score-level fusion to produce a PAD score for the video.

Istanbul Technical University (SiMiT Lab): Team SiMiT Lab's algorithm, called Shuffled Patch Wise Supervision, is a special training method for liveness detection. The main contribution is creating different input data than the previous approaches. Patch-wise supervision forces the model to detect liveness by using small clues of the given patch. It allows the model to be robust to out-of-sample data. The stitched different face patches from different peo-

ple instead of using full-face images. For example, one single image consists of 49 different persons' face images with 32x32 patch size. Pixel-wise supervision is used for training the proposed model in DeepPixBis [6]. The team used the EfficientNet-B7 backbone to create a 2D feature map from 224x224 input images that are a combination of different patches. The initial feature map of the model has 14x14x224 dimensions (width x height x channel). Then by using a 1x1 convolution, a 14x14x1 map were generated to check if the patches are bona-fide or fake with a score between 0 and 1. Since, the input image consists of 49 patches, each patch corresponds to 4 cells in a 14x14 map. The calculated mean score of this prediction map created a liveness score between 0 to 1. Binary cross-entropy loss on the 14x14 map were used as objection function. In the test time, the algorithm did not create shuffled faces, instead fed the model with the given input face image and created a 14x14 map that shows the model prediction for each part of the face, instead used the mean score of this 14x14 map. The submitted algorithm was trained with Replay Mobile, SiW, Oulu-NPU, and 3DMAD dataset.

CLFM: Team CLFM submitted their solutions for both image and video category. CLFM created a model with 12 layers for this competition. In the model, central difference convolution is used to replace traditional convolution. Also, several attention modules are introduced to make the model perform better. The input images were resized to 56*56 and the model was trained from scratch using several public datasets.

FaceMe: Team FaceMe submitted their solutions for both image and video category. In the algorithm of team FaceMe, there are 3 sub-projects. One sub-project is a depth based neural network classification and images were resized to 128x128 as an input; One sub-project is a frame detector based on digital signal processor; One sub-project is a frame detector based on a neural network and uses images resized to 224x224 as an input. This project arbitrates a result with the score of each sub-project.

Vision Intelligence Center of Meituan (little tiger): Team little tiger's proposed method is a fusion method of five models based on whole image or cropped face image. There are two models based on the whole image, the first one is a binary classifier model with backbone of resnet50. In the training stage of this model, patches are randomly cropped from the input image with sizes of 224 * 224, while in the inference stage, the center region with the same size as input is cropped to get the confidence score. Another whole image-based model, which is also a binary classifier with backbone of resnet50, used beside the ordinary data augmentation strategy. The model also uses constrained mix-up operation which can only be used between the same categories to do data augmentation before training. During training, besides the common classification loss, the contrast loss is also added to supervise the learning of the network. In addition, there are three models based on face images. One model uses resnext26 as the backbone to train a binary classification model. In the training phase, it randomly cropped a 224 * 224 block from the face-based image as input, while in the test phase, it takes 9 different patch blocks from the whole cropped face image, and then uses the comprehensive results of 9 blocks predicted by the trained model to improve the prediction accuracy. Another model is an improved Central Difference ConvolutionalNetwork(CDCN), which has a custom defined dual attention structure in original CDCN, thus it is called as CDC-DAN (central difference revolutionary dual attention network), the input of this network is 128*128 block that is randomly cropped from the origl cropped face image, and the output is a feature map with the size of 16*16. The model is also a binary classifier that uses a pre-training weight produced by contrast learning on Glint360k dataset, and contrast loss is also added supervise the learning of network in the training stage.

NTNU Gjøvik: Team NTNU Gjøvik submitted two algorithms for the video category of the competition and the algorithms use the framework based on [18] Hierarchical Spherical Linear Interpolation (SLERP) of deep learning feature vectors followed by training a Linear SVM for PAD Classification. The deep learning feature vectors are extracted from existing networks trained on the Imagenet dataset followed by SLERP to generate a single feature vector, and we train two Linear SVMs. The first linear SVM is trained by extracting features from Resnet-18 (pool5 layer), Resnet50 (average pool layer), and Inception-v3 (average pool layer), giving a 2048-dimensional feature vector. The second linear SVM is trained by extracting features from VGG-19 (fc6 layer), VGG-16 (fc6 layer), and Alexnet (fc6 layer), giving a 4096-dimensional feature vector. The per-frame predictions obtained by the Linear SVMs are majority voted to generate a video-level decision. The current training set for the submission includes SWAN [15], CASIA-FASD [23], and NTNU-Silicon Mask [16] dataset.

5. Results

In this section the performance of the competing algorithms of both the categories: (1) image and (2) video, are discussed. The performance has been evaluated with APCER for each of the PAI categories and BPCER for the live category. Both the APCER and BPCER are evaluated at the threshold of 50, which was announced prior to the competition. A summary of the error rates for both image and the video category are provided in Table 2 and Table 3. The performance comparison of the algorithms of the image category based on Receiver Operating Characteristics (ROCs) are shown in Figure 2. The same could not be done for the video category as most of the output scores of the

algorithms were binary.

LivDet-Face 2021 Image category: Team Fraunhofer IGD is the winner based on the lowest ACER = 16.47%, closely followed by Team CLFM with ACER = 18.71%. The winning team's algorithm achieved the lowest BPCER = 15.33% among the six competitors. All the six competitors achieved variable performance for each type of PAI. The algorithm submitted by team Fraunhofer IGD detected all the low-quality paper displays and the team CLFM's algorithm successfully detected all the low-quality 3D mask samples. Team CLFM's algorithm also performed best with APCER = 10% for high-quality photo paper display samples but they achieved a BPCER = 24.08%. Similarly, team FaceMe, who achieved third position in the image category competition achieved APCER = 3%, the best for laptop display samples and they achieved BPCER = 16.06%. Team ITU's algorithm successfully detected all the medium-quality 3D mask samples with APCER = 0% which is best among all the competitors, although they achieved BPCER = 51.09%. The lowest APCER was achieved by the team UL for the high-quality 3D masks with APCER = 100% and closely for the silicon masks as well with APCER = 98.58%. The live face detection performance of team anonymous-1 with BPCER = 16.79% is third best among the six competitors.

Comparing the performance of the algorithms of the two best competitors from the image category it is evident that the algorithms performed better against low-quality PAIs than higher quality PAIs. Team Fraunhofer IGD performed with APCER = 0 % for the low-quality paper display against APCER = 24% for high-quality paper display. Similarly, team CLFM performed with APCER = 6.06% for low-quality paper display against APCER = 10% for high-quality paper display. The same trend can be observed for the different quality of 3D face masks. Team Fraunhofer IGD's performed with APCER = 4.17% for low-quality 3D masks compared to APCER = 8.33% for medium-quality 3D masks, APCER = 14.29% for high-quality 3D masks to APCER = 16.31% for high-quality silicon masks. Similarly, Team CLFM performanced with APCER = 0% against low-quality 3D masks, compared to APCER = 16.67% for medium-quality 3D masks, to APCER = 21.43% for high-quality 3D masks and APCER = 34.75% for high-quality silicon masks.
For the PAI category performance comparison, the second ranked team, CLFM, on average performed better in Level A and Level B type PAIs compared to the first ranked team Fraunhofer IGD. Team Fraunhofer's algorithm performed well for the sophisticated Level C type PAIs compared to Level A and Level B types.

LivDet-Face 2021 Video category: Team FaceMe is the winner of the video category of the competition with the ACER = 13.81% and was closely followed by the team Fraunhofer IGD, with ACER = 14.49%. Team FaceMe also had the lowest BPCER = 14.29% compared to team Fraunhofer IGD BPCER = 16.67%. The lowest BPCER = 4.76% was achieved by team NTNU Gjøvik. Team CLFM performed well to detect the PAIs with average APCER = 3.30% but they ranked third as their algorithm performed with a BPCER = 39.68%. Team Fraunhofer IGD achieved lowest APCER among the PAIs with APCER = 0% for medium-quality face masks and APCER = 1% in low-quality photo display and photo masks. Team CLFM's solution also performed well against live video display attacks and low-quality face masks with APCER = 0%. Team NTNU Gjøvik performed well against 3D face mask attacks and achieved APCER = 0% for the three different types of 3D masks.

Comparing the performance of the algorithms of the top two competitors from the video category– FaceMe (first) and Fraunhofer IGD (second), it can be observed again that the algorithms performed better against low-quality PAIs than higher quality PAIs. Team FaceMe's performance for the low-quality paper display was APCER = 8% against the high-quality paper display where APCER = 10.10%. Similarly team Fraunhofer IGD's performance against low-quality paper display was APCER = 1% against the high-quality paper display with APCER = 25.25%. The same trend can be observed for the different quality of 3D face masks as well. Team FaceMe's performance against low-quality 3D masks was APCER = 40% compared to medium-quality 3D masks with APCER = 45.45%. But the team's performance is marginally better against high-quality 3D masks with APCER = 38.46% and to high-quality silicon masks with APCER = 9.22%. Team Fraunhofer IGD's performance against low-quality 3D masks was APCER = 4%, compared to medium-quality 3D masks with APCER = 9.09%, and to high-quality silicon masks with APCER = 34.75%. But Fraunhofer IGD's performance of high-quality 3D masks was better than any other 3D mask categories, with APCER = 0%.

For performance comparison of the video category for each PAI type, the first ranked team FaceMe performed better for the level A and B type PAIs compared to level C type PAIs. In comparison, team Fraunhofer IGD's algorithm performed well against sophisticated Level C type PAIs compared to Level A and Level type. It should be mentioned that team FaceMe is the winner of the video category based on the weighted average APCER, even though their APCER performance differed significantly from the second placed team for the sophisticated level C type PAIs. The weighted average of the APCER score was close to the second ranked team because of the significantly low number of the level C type PAIs in the test dataset compared to the level A and B type PAIs.

Table 2: Facial-PAD Competition Summary: Image category PAD results for all competitors

Competitor Name	Presentation Attack Instruments Level Types								Overall Performance		
	Level A		Level B		Level C				$APCER_{average}$	BPCER	ACER
	Paper Display		Display Attacks	2D Masks	3D Face Masks						
	LQ	HQ	DL	PM	LQ	MQ	HQ	Silicon			
Fraunhofer IGD	0	24	45	14.70	4.17	8.33	14.29	16.31	17.61	15.33	16.47
CLFM	6.06	10	8	5.88	0	16.67	21.43	34.75	13.33	24.08	18.71
FaceMe	22.22	11	3	11.76	66.67	66.66	50	57.45	25.40	16.06	20.72
little tiger	41.41	52	4	58.82	54.17	25	28.57	82.98	46.67	21.17	33.92
SiMiT Lab	7.07	18	43	15.68	16.66	0	42.85	80.85	33.01	51.09	42.05
Anonymous-1	78.78	86	77	89.21	87.5	83.33	100	98.58	81.90	16.79	49.35

Table 3: Face PAD Competition Summary: Video category PAD results for all competitors

Competitor Name	Presentation Attack Instruments Level Types									Overall Performance		
	Level A		Level B			Level C				$APCER_{average}$	BPCER	ACER
	Paper Display		Display Attacks		2D Masks	3D Face Masks						
	LQ	HQ	DL	VD	PM	LQ	MQ	HQ	Silicon			
FaceMe	8	10.10	18	16	6.93	40	45.45	38.46	9.22	13.33	14.29	13.81
Fraunhofer IGD	1	25.25	29	9	1	4	9.09	0	12.77	12.32	16.67	14.49
CLFM	4	4.04	8	1	0	0	27.27	7.69	1.42	3.30	39.68	21.49
NTNU Gjøvik-V1	50	59.60	83	75	18.81	36	18.18	46.15	21.28	48.26	4.76	26.51
NTNU Gjøvik-V2	5	9.09	32	20	1	0	0	0	33.33	16.52	51.59	34.05

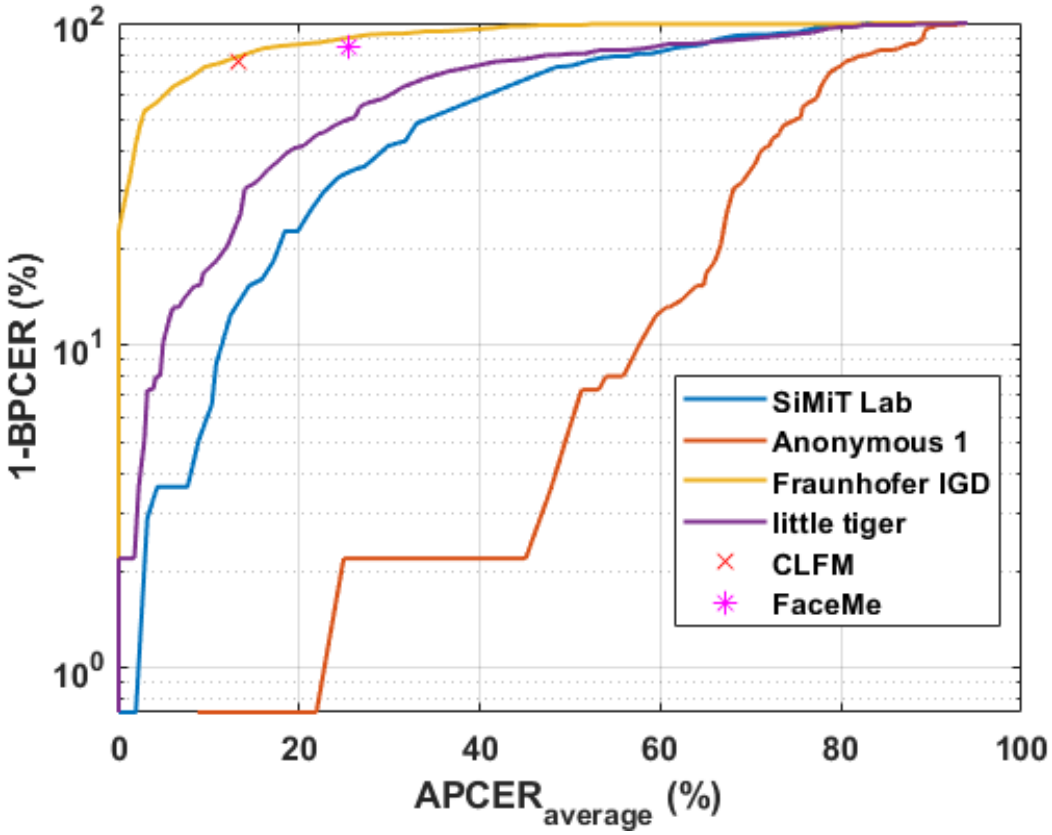

Figure 2: ROC curves for all six algorithms for the image category of the competition presenting the overall performance on samples representing all eight PAIs. The overall APCER is evaluated based on ($APCER_{average}$).

The results from both the categories of the LivDet-Face 2021 competition depicts the performance difference of the algorithms against high-quality PAIs. Most of the algorithms performed poorly against high-quality PAIs than low-quality PAIs. Also, from the results, it is evident that the state-of-the-art face detection algorithms are vulnerable against presentation attacks and specially against sophisticated level-C presentation attacks, with all six of the image category and two of the video category algorithms have high APCER scores against such attacks.

6. Conclusion

The LivDet-Face 2021 competition featured multiple new additions to the evaluation of face presentation attack detection: (a) employed three novel PAIs (high-quality 3D mask, flexible 3D silicon masks and video display sample of live subjects), (b) provided a comparative analysis of the six state-of-the-art algorithms in image categories and five in video categories. The winning algorithm of the image category achieved an ACER of 16.47% (APCER averaged over all PAIs = 17.61% and BPCER = 15.33%). The winning algorithm of the video category achieved an ACER of 13.81% (APCER averaged over all PAIs = 13.33% and BPCER = 14.29%). While the competitors were not provided any training dataset, a small example dataset was shared with the competitors. This allowed the competitors flexibility to use any publicly available or proprietary dataset and testing their developed algorithms in a real-world uncertain scenario of different PAIs of different attack types, different environment and sensors.

We note degradation of the overall performance of the competitors than the two of the recent competitions mentioned in the literature [2] [22]. This overall degradation can be contributed by multiple factors:

a) increased complexity in the test dataset of both image and video category: nine different PAI types were employed in the competition;
b) introduction of three novel attack types with limited or no access to large-enough public dataset for the PAIs;
c) no training dataset was offered, and that training choice was left to be decided by competitors;
d) the results may reflect variability between the training and the test dataset in terms of environmental factors, sensors, quality of PAIs, and the introduction of "unknown" PAIs.

The results from this competition indicate that face PAD has still a long way to go and is far from a fully solved research problem. Large differences in accuracy among the evaluated algorithms, stress the importance of access to large and diversified training dataset, encompassing many PAIs. We believe that this competition, and the benchmark dataset to be available to researchers via the BEAT platform, will contribute to our efforts as a biometric community to improve biometric system security and confidence.

7. Acknowledgement

This material is based upon the work supported in part by the National Science Foundation under Grant No. 1650503, the Center for Identification Technology and Research (CITeR), and the European Union's Horizon 2020 research and innovation program, AI4EU, under grant agreement 825619.

References

[1] Biometrics Evaluation and Testing (BEAT) . https://www.idiap.ch/software/beat/. Accessed: 2021-06-28. 2

[2] Z. Boulkenafet, J. Komulainen, Z. Akhtar, A. Benlamoudi, D. Samai, S. E. Bekhouche, A. Ouafi, F. Dornaika, A. Taleb-Ahmed, L. Qin, et al. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 688–696. IEEE, 2017. 3, 9

[3] N. Damer, A. Opel, and A. Nouak. Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution. In *2014 22nd European Signal Processing Conference (EUSIPCO)*. IEEE, 2014. 5

[4] FaceGen Modeller. available at: http://facegen.com/modeller.html, 2021. 4

[5] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper. Real masks and fake faces: On the masked face presentation attack detection. *arXiv preprint arXiv:2103.01546*, 2021. 5

[6] A. George and S. Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019. 1, 5, 6

[7] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15, 2019. 1

[8] L. J. González-Soler, M. Gomez-Barrero, and C. Busch. On the generalisation capabilities of fisher vector based face presentation attack detection. *arXiv preprint arXiv:2103.01721*, 2021. 1

[9] A. Husseis, J. Liu-Jimenez, I. Goicoechea-Telleria, and R. Sanchez-Reillo. A survey in presentation attack and presentation attack detection. In *2019 International Carnahan Conference on Security Technology (ICCST)*, pages 1–13. IEEE, 2019. 1, 2

[10] ISO/IEC 30107-3. Information technology – Biometric presentation attack detection – Part 3: Testing and reporting, 2016. 2

[11] A. S. Jackson, A. Bulat, V. Argyriou, and G. Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE international conference on computer vision*, pages 1031–1039, 2017. 4

[12] LivDet Organizing Team. Livdet website. available at: http://livdet.org/. 3

[13] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar. Detecting silicone mask-based presentation attack via deep dictionary learning. *IEEE Transactions on Information Forensics and Security*, 12(7), 2017. 1

[14] R. Ramachandra and C. Busch. Presentation attack detection methods for face recognition systems: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 50(1), 2017. 1, 2

[15] R. Ramachandra, M. Stokkenes, A. Mohammadi, S. Venkatesh, K. Raja, P. Wasnik, E. Poiret, S. Marcel, and C. Busch. Smartphone multi-modal biometric authentication: Database and evaluation. *arXiv preprint arXiv:1912.02487*, 2019. 6

[16] R. Ramachandra, S. Venkatesh, K. B. Raja, S. Bhattacharjee, P. Wasnik, S. Marcel, and C. Busch. Custom silicone face masks: Vulnerability of commercial face recognition systems & presentation attack detection. In *2019 7th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE, 2019. 6

[17] S. Schuckers, G. Cannon, E. Tabassi, M. Karlsson, and E. Newton. Fido biometrics requirements. *Population*, 5:2–1, 2019. 2, 3

[18] J. M. Singh, R. Ramachandra, and C. Busch. Hierarchical interpolation of imagenet features for cross-dataset presentation attack detection. In *Intelligent Technologies and Applications: Third International Conference, INTAP 2020, Grimstad, Norway, September 28–30, 2020, Revised Selected Papers 3*. Springer International Publishing, 2021. 6

[19] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 5

[20] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10), 2016. 5

[21] 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 Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020. 3

[22] Y. Zhang, Z. Yin, J. Shao, Z. Liu, S. Yang, Y. Xiong, W. Xia, Y. Xu, M. Luo, J. Liu, et al. Celeba-spoof challenge 2020 on face anti-spoofing: Methods and results. *arXiv preprint arXiv:2102.12642*, 2021. 2, 9

[23] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li. A face antispoofing database with diverse attacks. In *2012 5th IAPR international conference on Biometrics (ICB)*, pages 26–31. IEEE, 2012. 6