Review of Face Presentation Attack Detection Competitions

Zitong Yu, Jukka Komulainen, Xiaobai Li and Guoying Zhao

Abstract Face presentation attack detection (PAD) has received increasing attention ever since the vulnerabilities to spoofing have been widely recognized. The state of the art in unimodal and multi-modal face anti-spoofing has been assessed in eight international competitions organized in conjunction with major biometrics and computer vision conferences in 2011, 2013, 2017, 2019, 2020 and 2021, each introducing new challenges to the research community. In this chapter, we present the design and results of the five latest competitions from 2019 until 2021. The first two challenges aimed at evaluating the effectiveness of face PAD in multi-modal setup introducing near-infrared (NIR) and depth modalities in addition to colour camera data, while the latest three competitions focused on evaluating domain and attack type generalization abilities of face PAD algorithms operating on conventional colour images and videos. We also discuss the lessons learnt from the competitions and future challenges in the field in general.

Zitong Yu
Center for Machine Vision and Signal Analysis, University of Oulu, Finland
e-mail: zitong.yu@oulu.fi

Jukka Komulainen
Visidon Ltd, Finland
e-mail: jukka.komulainen@kapsi.fi

Xiaobai Li
Center for Machine Vision and Signal Analysis, University of Oulu, Finland
e-mail: xiaobai.li@oulu.fi

Guoying Zhao
Center for Machine Vision and Signal Analysis, University of Oulu, Finland
e-mail: guoying.zhao@oulu.fi

1 Introduction

Presentation attacks (PAs) [32], commonly referred to also as spoofing, pose a serious security issue to biometric systems. Automatic face recognition (AFR) systems, in particular, are easy to be deceived, *e.g.*, using images of the targeted person published on the web or captured from distance. Many works, such as [53], have concluded that face recognition systems are vulnerable to sensor-level attacks launched with different presentation attack instruments (PAI), such as prints, displays, and wearable 3D masks. The vulnerability to PAs is one of the main reasons for the lack of public confidence in AFR systems, especially in high-security applications, such as mobile payment services, which has created a necessity for robust solutions to counter spoofing.

One possible solution is to include a dedicated face presentation attack detection (PAD) component into AFR systems. Face PAD, commonly referred to also as face anti-spoofing (FAS) or liveness detection, aims at automatically differentiating whether the presented face biometric sample originates from a bona fide subject or an artefact. Based on the used imaging modalities, face PAD schemes can be broadly categorized into unimodal and multimodal based methods. Unimodal face PAD systems usually exploit efficient visual features extracted from conventional colour (RGB) camera data for binary classification (*i.e.*, bona fide vs. attack), thus they can be easily deployed in most practical AFR scenarios but with limited accuracy. In contrast, multi-modal methods [71] introduce some additional imaging modalities (*e.g.*, depth, near-infrared (NIR), or thermal infrared sensor data) that can capture specific intrinsic differences between the bona fide and attack samples but with extra hardware costs. For example, the depth maps obtained from 2D printout and display face artefacts using 3D sensors usually have flat and close-to-zero distributions in facial regions.

Despite the recent progress in deep learning based face anti-spoofing methods [52, 84] with powerful representation capacity, it is difficult to tell what are the best or most promising feature learning based approaches for generalized face PAD. Along with the development in manufacturing technologies, it has become even cheaper for an attacker to exploit known vulnerabilities of AFR systems with different kinds of face artefacts, such as a realistic 3D mask made of plaster. Simulating AFR scenarios with various attacks, environmental conditions, acquisition devices and subjects is extremely time-consuming and expensive, but the domain shifts caused by such covariates have a significant impact on the face PAD performance. These issues have been already explored with several public unimodal datasets, such as [3, 7, 49, 74, 95]. However, these benchmarks have been yet rather small-scale in terms of number of subjects, samples and AFR scenarios and, consequently, the corresponding evaluation protocols have been too limited (*e.g.*, in terms of unknown PAs and acquisition conditions in the test set). Moreover, there have been no public benchmarks and protocols for evaluating multi-modal face PAD schemes until recently.

Competitions play a key role in advancing the research on face PAD and provide valuable insights for the entire face recognition community. It is important to organize collective evaluations regularly in order to assess, or ascertain, the current state of the art and gain insight on the robustness of different approaches using a common platform. Also, new, more challenging public datasets are often collected and introduced within such collective efforts to the research community for future development and benchmarking use. The quality of PAIs keeps improving as technology (*e.g.*, 2D/3D printers and displays) gets cheaper and better, which is another reason why benchmark datasets need to be updated regularly. Open contests are likely to inspire researchers and engineers beyond the field to participate, and their outside the box thinking may lead to new ideas on the problem of face PAD and novel countermeasures.

In the context of face PAD, altogether eight international competitions [45, 1, 5, 8, 43, 46, 58, 94] have been organized in conjunction with major biometrics and computer vision conferences in 2011, 2013, 2017, 2019, 2020 and 2021, each introducing new challenges to the research community. In this chapter, we focus on analysing the design and results of the five latest competitions [45, 43, 46, 58, 94] from 2019 until 2021, while an extensive review of the first three competitions [1, 5, 8] can be found in [38]. The key features of the five most recent face PAD competitions are summarized in Table 1.

Table 1 Summary of the recent five face PAD competitions organized from 2019 until 2021.

Competition	Modality	Highlight	Limitation
CVPR2019 challenge [45]	RGB, depth, NIR	First multi-modal face PAD challenge	With only print attacks
CVPR2020 challenge [43]	RGB, depth, NIR	Cross-ethnicity & cross-PAI testing	Testing with only print and mask PAs
ECCV2020 challenge [94]	RGB	Largest dataset with rich (43) attributes	Limited domain shift in testing set
IJCB2021 LivDet-Face [58]	RGB	Unseen testing set with rich (9) PAIs	No training set provided
ICCV2021 challenge [46]	RGB	Largest 3D mask dataset & open set protocol	Limited (only three) mask types

The multi-modal face anti-spoofing challenge organized in 2019 (referred to as CVPR2019 challenge) [45] provided an initial assessment of multi-modal countermeasures to various kinds of print attacks by introducing a precisely defined test protocol for evaluating the performance of the face PAD solutions with three modalities (*i.e.*, colour, depth and NIR). In 2020, the cross-ethnicity face anti-spoofing challenge (referred to as CVPR2020 challenge) [43] extended the previous CVPR2019 challenge with several new factors (*e.g.*, unseen ethnicities and 3D mask attacks), and having separate competition tracks for unimodal (colour) and multi-modal (colour, depth and NIR) data. While the datasets used in these first two contests contained a limited number of subjects and samples, the CelebA-Spoof Challenge in 2020 (referred to as ECCV2020 challenge) [94] provided an assessment on the performance of face PAD methods on the largest publicly available unimodal (colour) benchmark dataset. The LivDet Face 2021 liveness detec-

tion competition (referred to as IJCB2021 LivDet-Face) [58] focused on the challenging domain generalization issues and unseen attacks, with separate competition tracks for methods using colour image and video data. In order to bridge the gap between competitions (*i.e.*, laboratory conditions) and real-world application scenarios, the test data was practically concealed as only few samples of attacks were provided. Finally, to further evaluate the performance of face PAD approaches under challenging 3D mask attacks, the 3D high-fidelity mask attack detection challenge in 2021 (referred to as ICCV2021 challenge) [46] was conducted on the largest publicly available 3D mask dataset with a novel open-set evaluation protocol.

The remainder of the chapter is organized as follows. First, we will recapitulate the organization, solutions as well as results of the five most recent face PAD competitions in Section 2. Then, in Section 3, we will discuss the lessons learnt from the competitions and future challenges in the field of face PAD in general. Finally, Section 4 summarizes the chapter, and presents conclusions drawn from the competitions discussed here.

2 Review of Recent Face PAD Competitions

We begin our review by first introducing two multi-modal face PAD competitions, namely CVPR2019 and CVPR2020 challenges, in Sections 2.1 and 2.2, respectively, where the latter one included also a competition track for unimodal (colour) data. Then, three latest unimodal (colour) face PAD competitions, namely ECCV2020, IJCB2021 LivDet-Face and ICCV2021 challenges are reviewed in remaining Sections 2.3, 2.4 and 2.5, respectively.

2.1 Multi-Modal Face Anti-spoofing Attack Detection Challenge (CVPR2019)

The first three face PAD competitions [38] organized in conjunction with International Joint Conference on Biometrics (IJCB) 2011, International Conference on Biometrics (ICB) 2013 and IJCB2017 focused on photo (*i.e.*, both printed and digital) and video-replay attack detection relying on small-scale datasets (*i.e.*, PRINT-ATTACK [5], REPLAY-ATTACK [8], OULU-NPU [1]) for training, tuning and testing. To be more specific, these datasets have an insufficient number of subjects (< 60) and data samples ($< 6,000$ videos) compared with databases used in the field of image classification, *e.g.*, ImageNet [14] and face recognition, *e.g.*, CASIA-WebFace [80], which severely limits the development and testing of data-driven deep model based approaches for generalized face PAD. Also, due to the lack of variation in the face PAD datasets, the deep models have been suffering from overfit-

ting and learning database-specific information instead of generalized feature representations capturing the disparities in the inherent fidelity characteristics between bona fide samples and different kinds of face artefacts. Another missing feature in previous face PAD competitions has been the availability of multi-modal facial information in addition to conventional visible light colour (RGB) data. This kind of extended range imaging information might be very helpful for developing more robust face PAD methods for practical real-world AFR applications. In order to address the limitations of previous competitions, the Chalearn multi-modal face anti-spoofing attack detection challenge[1] [45] was held in conjunction with the Conference on Computer Vision and Pattern Recognition (CVPR) in 2019. The competition was based on a newly collected large-scale multi-modal face anti-spoofing dataset, namely CASIA-SURF [90, 91], which consists of 1,000 subjects and 21,000 video clips in three modalities (colour, depth and NIR). The goal of this competition was to push the research progress in AFR applications, where plenty of data and multiple modalities can be considered to be available.

Table 2 Teams and affiliations listed in the final ranking of the CVPR2019 challenge [45].

Ranking	Team Name	Affiliation
1	VisionLabs	VisionLabs
2	ReadSense	ReadSense
3	Feather	Intel
4	Hahahaha	Megvii
5	MAC-adv-group	Xiamen University
6	ZKBH	Biomhope
7	VisionMiracle	VisonMarcle
8	GradiantResearch	Gradiant
9	Vipl-bpoic	ICT, CAS
10	Massyhnu	Hunan University
11	AI4all	BUPT
12	Guillaume	Idiap Research Institute
invited team	Vivi	Baidu

The CVPR2019 challenge was run in the CodaLab[2] platform and consisted of two phases: development phase (December 22, 2018 – March 6, 2019) and final phase (March 6, 2019 – March 10, 2019). More than 300 academic and industrial institutions worldwide participated in this challenge, and finally 13 teams entered into the final stage. A summary with the names and affiliations of these teams is presented in Table 2. Compared with the previous competitions [1, 5, 8], the majority of the final participants (10 out of 13) of this competition came from the industry, which indicates the increased need for reliable liveness detection products in daily life applications. Furthermore,

[1] `https://sites.google.com/qq.com/face-anti-spoofing/welcome/challengecvpr2019`

[2] `https://competitions.codalab.org/competitions/20853`

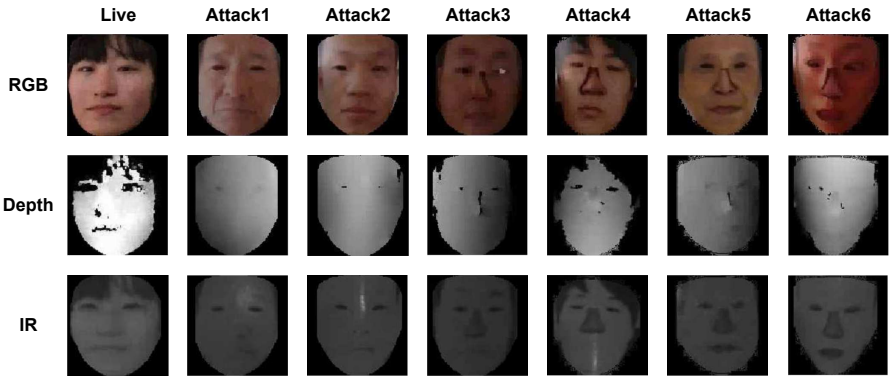

Fig. 1 Samples of a live face and six kinds of print attacks from the CASIA-SURF dataset [90, 91].

one highlight of the CVPR2019 challenge is that the three top-performing teams (VisionLabs[3], ReadSense[4], and Feather[5]) released their source code in GitHub and summarized their approaches in the related CVPR workshop papers [56, 64, 72, 89], enhancing the fairness, transparency, and reproducibility of the solutions so that they can be easily facilitated by the face recognition community.

2.1.1 Dataset

The CASIA-SURF dataset [90, 91] was the largest face PAD database in terms of number of subjects and videos at the time of the CVPR2019 challenge. Each sample of the dataset was associated with three modalities (colour, depth and NIR) captured using an Intel RealSense SR300 camera. Samples from each subject consist of one live video clip and one video clip of each six different attack presentations. A total number of 1,000 subjects and 21,000 videos were captured to build the dataset. Representative samples of bona fide and attack samples across the three modalities are illustrated in Fig. 1.

The CASIA-SURF dataset considers six different kinds of print attacks, where a person is holding:

- **Attack 1:** A flat face photo from which the eye regions are cut.
- **Attack 2:** A curved face photo from which the eye regions are cut.
- **Attack 3:** A flat face photo from which the eye and nose regions are cut.

[3] https://github.com/AlexanderParkin/ChaLearn_liveness_challenge

[4] https://github.com/SeuTao/CVPR19-Face-Anti-spoofing

[5] https://github.com/SoftwareGift/FeatherNets_Face-Anti-spoofing-Attack-Detection-Challenge-CVPR2019

• **Attack 4:** A curved face photo from which the eye and nose regions are cut.

• **Attack 5:** A flat face photo from which eyes, nose, and mouth regions are cut.

• **Attack 6:** A curved face photo from which the eye, nose, and mouth regions have been cut.

The samples in the CASIA-SURF dataset were pre-processed for the competition as follows: 1) the dataset is split in three subject-disjoint partitions: train, validation, and test sets, with 300, 100 and 600 subjects, respectively, when the corresponding number of videos is 6,300 (2,100 per modality), 2,100 (700 per modality) and 12,600 (4,200 per modality), 2) only every tenth frame from each video was selected to reduce the size of the competition dataset, which resulted in 148K, 48K and 295K frames for the three subsets, respectively, and 3) to mitigate the effect of pre-processing methods (*e.g.*, face detection and alignment) and limit the problem of face PAD to the actual facial information, the background information was masked out pixel-wise from original data, thus only pre-cropped aligned facial images were provided for each modality.

2.1.2 Evaluation Protocol and Metrics

In order to focus on the generalization to unknown attacks, the organizers provided only a part of the CASIA-SURF dataset for training, *i.e.*, for each subject only a subset of the PA types was available. Hence, the participants were given about 30K frames for training and 9.6K frames for validation. Note that the attacks in the test set differ from the attacks in the training set, therefore a successful model should avoid intra-attack overfitting, which was a common issue in the earlier face PAD competitions. The challenge comprised development and final stages. The detailed protocols are described as follows.

Protocol in development phase: *(December 22, 2018 – March 6, 2019).* During the development phase, the participants had access to the labelled training and unlabelled validation samples. Training data included the bona fide samples and three kinds of PAs (4, 5, 6), whereas the validation data consisted of bona fide samples and three other types of PAs (1, 2, 3). The participants were able to submit predictions on the validation partition and receive immediate response via the leaderboard using the CodaLab platform. As it can be observed from Fig. 1, the attacks (4, 5, 6) in the validation set differ in appearance (partial cuts in eyes, nose, and mouth regions) from attacks (1, 2, 3), which made the task of face PAD challenging.

Protocol in final phase: *(March 6, 2019 – March 10, 2019).* During the final phase, the labels for the validation subset were made available to the participants, so that they could leverage the additional labelled data for tuning to alleviate the domain gap between different attack types. The

participants had to make predictions on the unlabelled test partition and upload their solutions to the CodaLab platform. The considered test set was formed from bona fide samples and three kinds of PAs (1, 2, 3). The final ranking of the participants was determined based on the performance of the submitted systems on the test set. To be eligible for prizes, the top solutions had to release their source code under a licence of their choice and provide a fact sheet describing their solution. All codes would be re-run and verified by the organizing team after the final submission phase. For the sake of reproducibility and fairness, the final ranking of the teams was based on the verified results.

Evaluation metrics: The recently standardized ISO/IEC 30107-3[6] [32] metrics, including Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER) and Average Classification Error Rate (ACER) were adopted as part of the used evaluation criteria. They can be formulated as:

$$APCER = FP/(FP+TN), \tag{1}$$

$$BPCER = FN/(FN+TP), \tag{2}$$

$$ACER = (APCER+BPCER)/2, \tag{3}$$

where TP, FP, TN and FN correspond to true positive, false positive, true negative and false negative, respectively. APCER and BPCER are used to measure the error rates of attack or bona fide samples, respectively. Similarly to the common metrics in AFR systems, the Receiver Operating Characteristic (ROC) curve was also considered for examining a suitable operating point trade-off in the False Positive Rate (FPR) and True Positive Rate (TPR) regarding the requirements of real-world biometric applications. Finally, the operating point of TPR@FPR=10^{-4} was selected as the leading evaluation measure for the CVPR2019 challenge, while the ACER was used as an additional evaluation criterion.

2.1.3 Results and Discussion

In this subsection, we summarize all the face PAD solutions reaching the final stage in terms of method keywords, backbone models, pretraining data, modalities, fusion schemes, and loss functions. Finally, the overall results are analysed and discussed.

Summary of the participating solutions: Table 3 summarizes the face PAD solutions of the 13 participating teams and the baseline method. Different from the previous three competitions (*i.e.*, IJCB2011 [5], ICB2013 [8] and IJCB2017 [1]), none of the final teams used traditional face PAD methods, such as hand-crafted image quality/texture descriptors [2], and liveness

[6] `https://www.iso.org/standard/67381.html`

Table 3 Summary of the face PAD methods for all participating teams and baseline method [45]. 'SE' denotes Squeeze-and-Excitation [26] and 'BCE' denotes binary cross-entropy.

Team	Method	Backbone model	Pre-training data	Modalities	Fusion scheme and loss function
VisionLabs [56]	Fine-tuning Ensembling	ResNet34 [26] Resnet50 [26]	CASIA-WebFace [80] AFAD-Lite [54] MSCeleb1M [24] Asian dataset [96]	RGB Depth NIR	SE Fusion Score fusion BCE loss
ReadSense [64]	Bag-of-local features Ensembling	SE-ResNeXt [78]	No	RGB Depth NIR	SE Fusion Score fusion BCE loss
Feather [89]	Ensembling	Fishnet [66] MobileNetV2 [62]	No	Depth NIR	Score fusion BCE loss
Hahahaha	Depth only	ResNeXt [78]	Imagenet [14]	Depth	BCE loss
MAC-adv-group	Feature fusion	ResNet34	No	RGB Depth NIR	Feature fusion BCE loss
ZKBH	Regression model	ResNet18	No	RGB Depth NIR	Data fusion Regression loss
VisionMiracle	Modified Shufflenet-V2 Depth only	Shufflenet-V2 [51]	No	Depth	BCE loss
Baseline [91, 90]	Feature fusion	ResNet18	No	RGB Depth NIR	SE fusion BCE loss
GradiantResearch	Deep metric learning	Inception [67]	VGGFace2 [4] GRAD-GPAD [10]	RGB Depth NIR	Logistic regression ensemble
Vipl-bpoic [72]	Attention mechanism [76]	ResNet18	No	RGB Depth NIR	Data fusion Center loss [75] BCE loss
Massyhnu	Ensembling	9 softmax classifiers	No	RGB Depth NIR	Colour information fusion BCE loss
AI4all	Depth only	VGG16 [65]	No	Depth	BCE loss
Guillaume	Multi-Channel CNN No RGB	LightCNN [77]	Yes	Depth NIR	Data fusion BCE loss
Vivi	Dense cross-modal-attention model	DenseNet [97]	Yes	RGB Depth NIR	Feature fusion Score fusion BCE loss

cues like eye blinking, facial expression changes and mouth movements [55]. Instead, all the submitted face PAD solutions relied on data-driven model-based feature extractors, such as ResNet [26] and VGG16 [65]. Furthermore, most of the approaches were multi-modal, combining two or three modalities, while only three teams (Hahaha, VisionMiracle and AI4all) relied on unimodal (depth-based) PAD solution. It can be seen from the last column in Table 3 that several kinds of multi-modal fusion strategies (*e.g.*, input-level data fusion, feature-level Squeeze-and-Excitation (SE) [26] fusion and score-level fusion) were used. Regarding the use of pre-training data, two teams (VisionLabs and GradiantResearch) leveraged pre-trained models from related face analysis tasks (*e.g.*, face recognition models on CASIA-WebFace [80] and face PAD models on GRAD-GPAD [10]) to mitigate the issues with overfit-

ting. It is worth to note that all three top-performing solutions (VisionLabs, ReadSense and Feather) adopted ensemble strategy to aggregate the predictions from multiple variant models.

Table 4 Results and rankings of the stage teams [45] at the final stage. The best results are bolded. (∗ denotes Vivi that is affiliated with the sponsor and did not participate in the final ranking).

Team Name	FP	FN	APCER(%)	BPCER(%)	ACER(%)	TPR(%)@FPR=10e-4
VisionLabs	**3**	27	**0.0074**	0.1546	0.0810	**99.8739**
ReadSense	77	**1**	0.1912	**0.0057**	0.0985	99.8052
Feather	48	53	0.1192	0.1392	0.1292	98.1441
Hahahaha	55	214	0.1366	1.2257	0.6812	93.1550
MAC-adv-group	825	30	2.0495	0.1718	1.1107	89.5579
ZKBH	396	35	0.9838	0.2004	0.5921	87.6618
VisionMiracle	119	83	0.2956	0.4754	0.3855	87.2094
GradiantResearch	787	250	1.9551	1.4320	1.6873	63.5493
Baseline	1542	177	3.8308	1.0138	2.4223	56.8381
Vipl-bpoic	1580	985	3.9252	5.6421	4.7836	39.5520
Massyhnu	219	621	0.5440	3.5571	2.0505	29.2990
AI4all	273	100	0.6782	0.5728	0.6255	25.0601
Guillaume	5252	1869	13.0477	10.7056	11.8767	0.1595
Vivi*	7	15	0.0173	0.0859	**0.0516**	99.8282

Result analysis: The results and ROC curves of the participating teams on the test data are shown in Table 4 and Fig. 2, respectively. It can be observed that the winning team (VisionLabs) achieved TPR=99.8739%@FPR=10^{-4}, and the FN = 27 and FP = 3 on the test set. In fact, different application scenarios have different requirements for each indicator. For example, to meet the higher security needs of an access control system, the FP is required to be as small as possible. With respect to this criterion, VisionLabs performed very well as only three attack samples were misclassified as bona fide. In contrast, a small FN value is crucial from usability point of view, where the team ReadSense achieved the best result (FN=1) due to the effectiveness of local patch inputs. In overall, the first eight teams were performing better than the baseline method [90, 91] in terms of FP and TPR@FPR=10^{-4}, indicating the valuable outputs and insightful solutions of this challenge.

As shown in Table 4, the results of the three top-performing teams on the test set were clearly superior compared with the other teams. By combining Table 3 with Table 4, we can conclude that ensemble learning performed more robustly compared to single-model based solutions under the same conditions. The ROC curves of all the participating teams are illustrated in Fig. 2. It can be seen that three teams (*i.e.*, VisionLabs, ReadSense, and Vivi) were significantly better than other teams on the test set. For instance, the TPR@FPR=10^{-4} values of these three teams are relatively close to each other and superior compared to the other teams. The characteristics of the

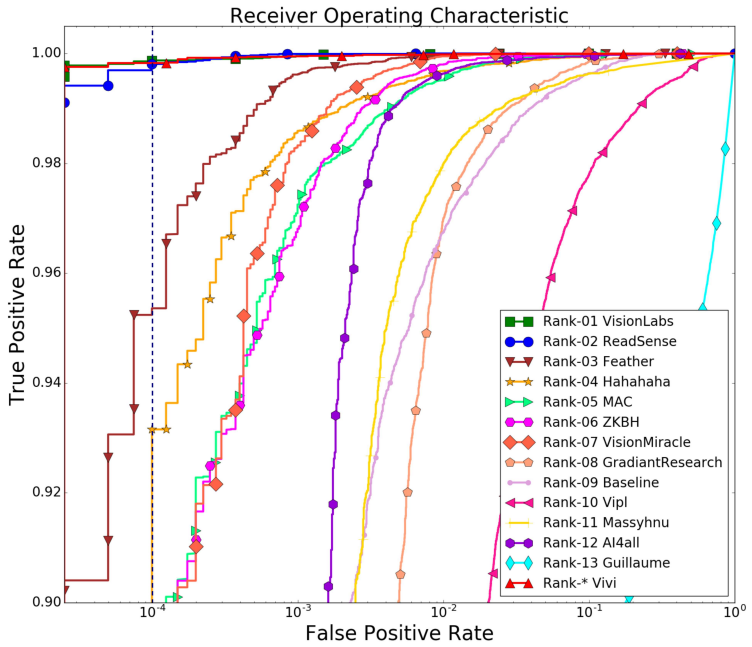

Fig. 2 ROC curves of the teams on test set at the final stage [45].

three modalities are different, as the colour data is rich in details, the depth data is sensitive to the distance, while the NIR data captures better the face-specific skin reflectance properties, for instance. Therefore, the teams Vivi and Vipl-bpoic introduced attention mechanisms into the face PAD task, enforcing the models to focus on different informative regions among three modalities. Similarly, the team Feather used a cascaded architecture with two subnetworks, where two modalities of the CASIA-SURF dataset (*i.e.*, depth and NIR data) were examined subsequently by each network. Some teams considered also facial landmarks (*i.e.*, Hahahaha) and colour space conversions (*i.e.*, MAC-adv-group and Massyhnu) for PAD. Instead of conventional binary classification based PAD model, the team ZKBH constructed a regression model to supervise the model to learn local cues in the eye regions. In order to generalize better to unseen attacks, the team GradiantResearch reformulated face PAD as an anomaly detection problem using deep metric learning.

Discussion: Although most of the proposed solutions achieved superior performance compared with the provided multi-modal SE fusion based baseline, there were still some limitations in the CVPR2019 challenge. Originally, the main research question was to explore new efficient multi-modal fusion schemes for combining the colour, depth and NIR modalities. However, no novel or otherwise insightful multi-modal fusion strategies were proposed in

the end. Most of the teams applied very simple data-level and score-level fusion in a greedy search manner, which is likely to fail when evaluating a method on an unknown multi-modal dataset. Furthermore, the two top-performing teams adopted the feature-level SE fusion strategy directly from the baseline method, which again was not very fruitful from the multi-modal challenge point of view. Many of the top-performing solutions exploited an ensemble of multiple models to boost the performance. However, while pushing the efficiency, model ensembling also increases the complexity of the whole solution, which is not practical in real-world conditions, especially considering the limitations of mobile and embedded platforms. Finally, some solutions were considerably inspired by the human-observed priors in the CASIA-SURF dataset (*e.g.*, the apparent discrepancy in the eye and nose regions between bona fide and attack samples), which were easily fooled by the cut paper attacks with similar shapes in these regions. Based on the aforementioned observations, the problem of designing more generalized multi-modal face PAD solutions capturing specific intrinsic fidelity characteristics between bona fide and attack samples remains an open issue.

2.2 Cross-Ethnicity Face Anti-spoofing Recognition Challenge (CVPR2020)

The racial bias in face PAD methods was not explicitly explored until it was demonstrated in [44] that the PAD performance of deep models can vary widely on test samples with unseen ethnicity. To alleviate the racial bias and ensure the reliability of face PAD methods among different populations, the CASIA-SURF Cross-ethnicity Face Anti-Spoofing (CeFA) dataset [44] along with the Chalearn Cross-ethnicity Face Anti-Spoofing Recognition Challenge [43] were established.

The cross-ethnicity face PAD challenge comprised unimodal (*i.e.*, colour) and multi-modal (*i.e.*, colour, depth and NIR) competition tracks, which were collocated with the Workshop on Media Forensics[7] at CVPR2020. Similarly to the previous multi-modal challenge, both the unimodal[8] and multi-modal[9] tracks were run simultaneously using the CodaLab platform. The competition attracted 340 teams in the development stage, with 11 and eight teams finally entering the actual evaluation stage for the unimodal and multi-modal face PAD tracks, respectively. A summary of the names and affiliations of teams that entered the final stage as well as their final rankings are shown in Tables 5 and 6 for the unimodal and multi-modal tracks, respectively. From the tables, it can be seen that most participants came from industrial insti-

[7] `https://sites.google.com/view/wmediaforensics2020/home`

[8] `https://competitions.codalab.org/competitions/22151`

[9] `https://competitions.codalab.org/competitions/22036`

tutions, indicating the increasing need for reliable and robust PAD systems in practical AFR applications. Interestingly, the team VisionLabs was not only the winner of the unimodal track of the CVPR2020 challenge, but also the winner of the earlier multi-modal CVPR2019 challenge [45]. In addition, the team BOBO from University of Oulu (the authors' team) proposed novel Central Difference Convolution (CDC) [86, 87] and contrastive depth loss (CDL) [73] methods for feature learning, achieving the first and second place in multi-modal and unimodal tracks, respectively.

2.2.1 Dataset

The CASIA-SURF CeFA [44] was the largest face anti-spoofing dataset at the time of the CVPR2020 competition, covering three ethnicities (*i.e.*, Africa, East Asia and Central Asia), three modalities (*i.e.*, colour, depth and NIR), 1,607 subjects, and four different of PA types (*i.e.*, prints, video-replays, 3D print and silica gel masks). The multi-modal videos were captured using an Intel RealSense SR300 camera with resolution of 1280 × 720 pixels for each video frame at 30 frames per second. The data was pre-processed in similar

Table 5 Names, affiliations and rankings of the participating systems in the unimodal track [43].

Ranking	Team Name	Affiliation
1	VisionLabs	VisionLabs
2	BOBO	Zitong Yu, University of Oulu
3	Harvest	Jiachen Xue, Horizon
4	ZhangTT	Zhang Tengteng, CMB
5	Newland_tianyan	Xinying Wang, Newland Inc.
6	Dopamine	Wenwei Zhang, huya
7	IecLab	Jin Yang, HUST
8	Chuanghwa Telecom Lab.	Li-Ren Hou, Chunghwa Telecom
9	Wgqtmac	Guoqing Wang, ICT
10	Hulking	Yang, Qing, Intel
11	Dqiu	Qiudi

Table 6 Names, affiliations and rankings of the participating systems in the multi-modal track [43].

Ranking	Team Name	Affiliation
1	BOBO	Zitong Yu, University of Oulu
2	Super	Zhihua Huang, USTC
3	Hulking	Qing Yang, Intel
4	Newland_tianyan	Zebin Huang, Newland Inc.
5	ZhangTT	Tengteng Zhang, CMB
6	Harvest	Yuxi Feng, Horizon
7	Qyxqyx	Yunxiao Qin, NWPU
8	Skjack	Sun Ke, XMU

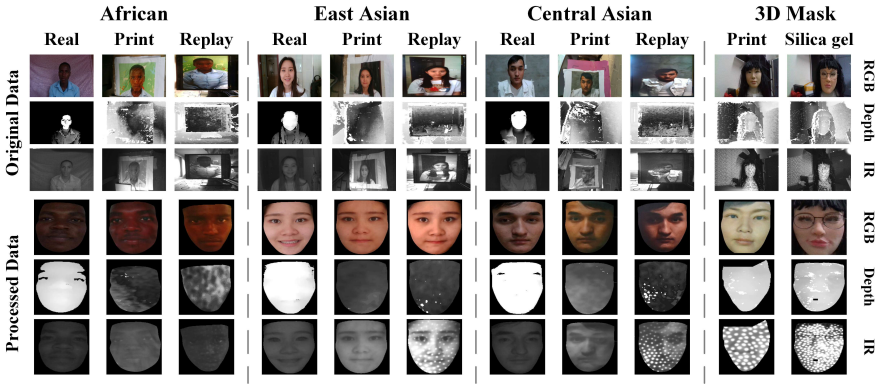

Fig. 3 Samples from the CASIA-SURF CeFA dataset [44], consisting of 1,607 subjects, three different ethnicities (*i.e.*, Africa, East Asia, and Central Asia), four PA types (*i.e.*, print, video-replay, 3D print and silica gel mask) and three modalities (*i.e.*, colour, depth, and NIR).

way as in the previous CVPR2019 multi-modal challenge [45] (see, Section 2.1.1). The CASIA-SURF CeFA was the first public dataset designed for exploring also the racial bias of face PAD methods. Some samples of the CASIA-SURF CeFA dataset are shown in Fig. 3.

The main motivation of CASIA-SURF CeFA dataset is to serve as a benchmark that allows evaluating the generalization of PAD methods across different ethnicities, PAIs and modalities under varying scenarios using four specific protocols:

- **Protocol 1:** Cross-ethnicity generalization of PAD methods is evaluated by using one ethnicity for training and validation, while the two remaining ones are used as unseen ethnicities for testing.
- **Protocol 2:** Cross-PAI generalization of PAD methods is evaluated by using print or video-replay attack for training and validation, while the remaining three attacks are used as unknown PA types for testing.
- **Protocol 3:** Cross-modality generalization of PAD methods is evaluated by using one modality for training and validation, while the two remaining ones are used as unknown modalities for testing.
- **Protocol 4:** Cross-ethnicity and cross-PAI generalization of PAD methods is evaluated simultaneously by combining the first two protocols, *i.e.*, using one ethnicity and PAI for training and validation, while the remaining two ethnicities as well as three PAIs are used for testing.

The most challenging Protocol 4 was adopted for ranking the methods in both unimodal and multi-modal tracks of the competition. As shown in Table 7, this protocol consists of three subsets: training, validation, and test sets, containing 200, 100, and 200 subjects for each ethnicity, respectively. Note that the remaining 107 subjects correspond to the 3D masks attacks.

Table 7 Protocols and statistics of the Protocol 4 of the CASIA-SURF CeFA [44] dataset, where 'A', 'C', and 'E' denote Africa, Central Asia, and East Asia, respectively.

Subset	Subjects (one ethnicity)	Ethnicity 4_1	Ethnicity 4_2	Ethnicity 4_3	PAIs	# images (RGB) 4_1	# images (RGB) 4_2	# images (RGB) 4_3
Train	1-200	A	C	E	Video-Replay	33,713	34,367	33,152
Valid	201-300	A	C	E	Video-Replay	17,008	17,693	17,109
Test	301-500	C&E	A&E	A&C	Print+Mask	105,457	102,207	103,420

Since there are three ethnicities in CASIA-SURF CeFA, in total three sub-protocols (*i.e.*, 4_1, 4_2 and 4_3 in Table 7) were adopted in the CVPR2020 challenge. In addition to the racial variation, the unknown PAIs introduced in the test sets made the competition even more challenging.

2.2.2 Evaluation Protocol and Metrics

The challenge comprised development and final stages. The detailed protocols are described as follows.

Protocol in development phase: *(December 13, 2019 – March 1, 2020).* During the development phase, the participants had access to the labelled training set and unlabelled validation set. Since the Protocol 4 of the CASIA-SURF CeFA dataset used in this competition comprised three sub-protocols (see Table 7), the participants first needed to train a model for each sub-protocol and then predict the scores for each corresponding validation set. Finally, the participants had to merge the predicted scores of the three sub-protocols and submit the resulting final scores to the CodaLab platform, where an immediate response was seen in the public leaderboard.

Protocol in final phase: *(March 1, 2020 – March 10, 2020).* During the final phase, the labelled validation set and the unlabelled testing set were released. The participants could first utilize the labels of the validation set for model selection to improve the generalization on the test data. All results of the three sub-protocols were made publicly available online in terms of APCER, BPCER, and ACER. Like with the OULU-NPU dataset [3] used in the IJCB2017 competition [1], the mean and variance of evaluated metrics across the three different sub-protocols were calculated and included in the final results.

Note that in order to fairly compare the performance of different submitted systems, the use of external training datasets or pre-trained models was explicitly prohibited in the CVPR2020 challenge. All participants were encouraged to release their source codes under feasible licences and to provide a fact sheet describing their solution. All codes would be re-run and verified by the organizing team after the final submission phase. For the sake of re-

producibility and fairness, the final ranking of the teams was based on the verified results.

Evaluation metrics: Similarly to the previous CVPR2019 competition [45], also in the CVPR2020 challenge, the standardized ISO/IEC 30107-3 [32] metrics (*i.e.*, APCER, BPCER and ACER) were considered as the main evaluation criteria (see, Section 2.1.2). Also, ROC curves were included for visualization purposes and additional result analysis. The final rankings were based on the ACER metric on the test set because it has been widely used for evaluating the performance of face PAD systems in the literature and majority of the previous face PAD competitions. The ACER threshold was determined by calculating the Equal Error Rate (EER) operating point on the validation set.

2.2.3 Results and Discussion

In this section, we first summarize the methods and report the results of the unimodal track, and then analyse the solutions as well as the results of the multi-modal track. Finally, we provide general discussion on the proposed algorithms and competition.

Solutions of the unimodal (colour) track: Table 8 summarizes the face PAD solutions of the teams participated in the unimodal track. The source codes of ten teams, including VisionLabs[10], BOBO[11], Harvest[12], ZhangTT[13], Newland-tianyan[14], Dopamine[15], IecLab[16], Chungwa-Telecom[17], Wgqtmac[18], and Hulking[19], were made publicly available. It was not surprising that every team adopted end-to-end learning based approaches due to the strong representation capacity of modern deep models. Regarding the model inputs, most of the teams used the provided facial colour images directly, while the winning team VisionLabs considered two kinds of pre-processing methods for dynamic inputs (*i.e.*, optical flow [28] and rank pooling [20] images). As for the backbone networks, only the team Dopamine adopted spatio-temporal 3D convolutional neural network (CNN) model, while the others relied on 2D CNNs (mostly ResNet). Most of the solutions treated face PAD as a binary classification problem via simple bi-

[10] https://github.com/AlexanderParkin/CASIA-SURF_CeFA

[11] https://github.com/ZitongYu/CDCN/tree/master/FAS_challenge_CVPRW2020

[12] https://github.com/yueyechen/cvpr20

[13] https://github.com/ZhangTT-race/CVPR2020-SingleModal

[14] https://github.com/XinyingWang55/RGB-Face-antispoofing-Recognition

[15] https://github.com/xinedison/huya_face

[16] https://github.com/1relia/CVPR2020-FaceAntiSpoofing

[17] https://drive.google.com/open?id=1ouL1X69KlQEUl72iKHl0-_UvztlW8f_l

[18] https://github.com/wgqtmac/cvprw2020.git

[19] https://github.com/muyiguangda/cvprw-face-project

nary cross-entropy (BCE) loss, but few teams (*i.e.*, BOBO, Harvest, and ZhangTT) considered pixel-wise depth loss, temporal continuous L1 regression loss, and multi-class softmax cross-entropy (CE) loss, respectively. It is interesting to see from the last two columns of Table 8 that most of the solutions leveraged dynamic cues but did not adopt complex model ensemble strategy.

Table 8 Summary of the top-performing solutions in the unimodal track of the CVPR2020 challenge, where 'S/D' indicates Static/Dynamic, 'OF' and 'RP' for optical flow [28] and rank pooling [20], 'BCE', 'CDC', and 'CDL' binary cross-entropy, central difference convolution, and contrastive depth loss, respectively.

Team Name	Method (keywords)	Input	Backbone	Loss function	S/D	Ensemble
VisionLabs	Creating artificial modalities	OF+RP	SimpleNet [57]	BCE loss	D	No
BOBO	CDC, CDL, Attention	RGB	CDCN [87]	Depth loss	S	Yes
Harvest	Motion-aware labels	RGB	ResNet101	L1 loss	D	Yes
ZhangTT	Quality tensor	Grayscale	ResNet [26]	4-class CE loss	D	No
Newland-tianyan	Subtracted neighborhood mean	RGB	5-layer network	BCE loss	D	No
Dopamine	Multi-task learning	RGB	ResNet100	BCE + face ID loss	S	No
IecLab	Authenticity+expression features	RGB	3DResNet [25]	BCE loss	D	No
Chunghwa-Telecom	Bag of local features	RGB	MIMAMO-Net [13]	BCE loss	S	Yes
Wgqtmac	Warmup strategy	RGB	ResNet18	BCE loss	S	No
Hulking	Frame vote module	RGB patch	PipeNet [20]	BCE loss	D	No
Dqiu	-	RGB	ResNet50	BCE loss	S	No
Baseline	Hybrid feature fusion	RGB+RP	SD-Net [44]	BCE loss	D	No

Results of the unimodal (colour) track: The final results of the 11 participated teams are shown in Table 9. The final ranking was based on the mean ACER computed over the three sub-protocols. The EER thresholds from the validation set are also reported in Table 9. The threshold values for the best-performing algorithms were either extremely large (*e.g.*, more than 0.9 for BOBO) or small (*e.g.*, 0.01 for Harvest), except for VisionLabs's algorithm that was more stable with threshold values around 0.5. Visionlabs achieved the highest accuracy in detecting the PA samples (APCER = 2.72%), while Wgqtmac's algorithm obtained the best results in terms of BPCER (0.66%). In overall, the first ten teams were performing better than the baseline method [44] in terms of ACER. The top three teams obtained excellent ACER values below 10%, and the team VisionLabs achieved the first place with a clear margin.

The ROC curves of the three sub-protocols are given in Fig. 4 to further analyse the trade-off between APCER and BPCER, *i.e.*, tuning the operating point according to the requirements of a given real-world application. The results of the winning team VisionLabs (blue curve) on all three sub-protocols are clearly superior compared to others, indicating the benefits of optical flow based motion clues and rank pooling images in improving the

Table 9 The results of the unimodal track of the CVPR2020 challenge [44]. Avg±Std denotes the mean and variance computed across the three sub-protocols. The best results are shown in bold.

Team Name	Threshold	FP	FN	APCER(%)	BPCER(%)	ACER(%)	Rank
VisionLabs	0.34±0.48	2±2	21±9	**0.11±0.11**	5.33±2.37	**2.72±1.21**	1
BOBO	0.97±0.02	129±67	10±2	7.18±3.74	2.50±0.50	4.84±1.79	2
Harvest	0.01±0.00	85±47	55±10	4.74±2.62	13.83±2.55	9.28±2.28	3
ZhangTT	0.9	97±37	75±31	5.40±2.10	18.91±7.88	12.16±2.89	4
Newland-tianyan	0.67±0.11	282±239	44±62	15.66±13.33	11.16±15.67	13.41±3.77	5
Dopamine	0.07±0.11	442±168	10±12	24.59±9.37	2.50±3.12	13.54±3.95	6
IecLab	0.40±0.07	597±103	24±2	33.16±5.76	6.08±0.72	19.62±2.59	7
Chunghwa-Telecom	0.86±0.06	444±93	76±34	24.66±5.16	19.00±8.69	21.83±1.82	8
Wgqtmac	0.80±0.22	928±310	2±3	51.57±17.24	**0.66±0.94**	26.12±8.15	9
Hulking	0.76±0.08	810±199	78±53	45.00±11.07	19.50±13.27	32.25±3.18	10
Dqiu	1.00±0.00	849±407	116±48	47.16±22.62	29.00±12.13	38.08±15.57	11
Baseline	1.00±0.00	1182±300	30±25	65.66±16.70	7.58±6.29	36.62±5.76	

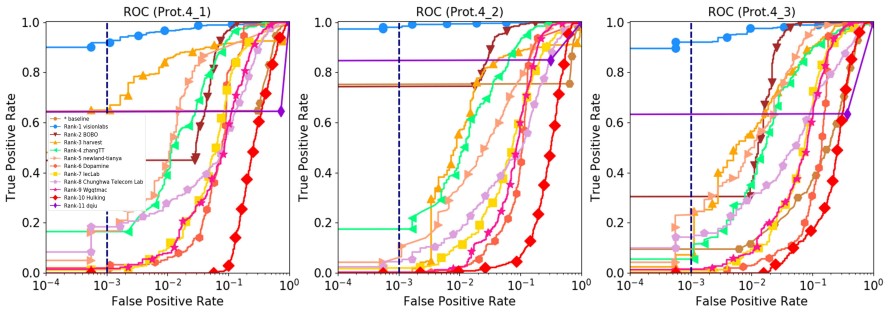

Fig. 4 The ROC curves for the 12 teams participating in the unimodal track of the CVPR2020 challenge [44]. From left to right are the ROC curves for protocol 4_1, 4_2 and 4_3, respectively.

generalization performance. However, the TPR value of the remaining teams decreases rapidly as the FPR reduces (*e.g.*, TPR@FPR=10^{-3} values for these teams are almost zero). In addition, although the ACER of the team Harvest was worse than that of the team BOBO, its TPR@FPR=10^{-3} was significantly better than that of BOBO. It was mainly because the values of FP and FN samples for the team Harvest were relatively close to each other (see Table 9).

Solutions of the multi-modal track: Table 10 summarizes the face PAD solutions of the teams that participated in the multi-modal track. The

source codes of seven teams, including BOBO, Super[20], Hulking[21], Newland-tianyan[22], ZhangTT[23], Qyxqyx[24], and Skjack[25], were made publicly available. Most of the teams exploited all the three modalities (colour, depth and NIR) for feature and score level fusion, except for the teams ZhangTT and Harvest who considered only depth and NIR modalities. There were no teams using data level fusion strategy. As for the architectures and loss functions, teams BOBO and Qyxqyx adopted MM-CDCN [83] and DepthNet [49] with pixel-wise supervision, while the other teams relied on ResNet with BCE loss. In contrast to the unimodal track, the use of static cues and ensemble models were popular in the multi-modal track.

Table 10 Summary of the top-ranked solutions of the multi-modal track in the CVPR2020 challenge.

Team Name	Modality	Fusion	Backbone	Loss function	S/D	Ensemble
BOBO	RGB, Depth, NIR	Feature & score level	MM-CDCN [83]	Depth loss	S	Yes
Super	RGB, Depth, NIR	SE fusion in feature level	ResNet34/50	BCE loss	S	Yes
Hulking	RGB, Depth, NIR	Feature level	PipeNet [79]	BCE loss	D	No
Newland-tianyan	Grayscale, Depth, NIR	Score level	Resnet9	BCE loss	S	No
ZhangTT	Depth, NIR	Feature level	ID-Net	BCE loss	S	Yes
Harvest	NIR	No fusion	-	Triplet loss	S	No
Qyxqyx	RGB, Depth, NIR	Score level	DepthNet [49]	BCE+BinaryMap loss	S	Yes
Skjack	RGB, Depth, NIR	Feature level	Resnet9	BCE loss	S	No
Baseline	RGB, Depth, NIR, RP	Feature level	PSMM-Net [44]	BCE loss	D	No

Results of the multi-modal track: The results of the eight teams participating in the final stage are shown in Table 11. The team BOBO team achieved the best performance in terms of BPCER = 1.00% and ACER = 1.02%, and the team Super ranked second with a minor margin ACER = 1.68%. It is worth noting that the team Newland-tianyan achieved the best results in terms of APCER (0.24%). Similarly to the unimodal track, most of the participating teams had relatively large EER thresholds calculated on the validation set, especially the teams Super and Newland-tianyan with threshold values of 1.0, indicating that the samples would be easily classified as anomalies. In addition, it can be seen that the ACER values of the four top teams were 1.02%, 1.68%, 2.21% and 2.28%, all outperforming the best performance (2.72%) reported in the unimodal track. This suggests that the

[20] `https://github.com/hzh8311/challenge2020_face_anti_spoofing`
[21] `https://github.com/ZhangTT-race/CVPR2020-SingleModal`
[22] `https://github.com/Huangzebin99/CVPR-2020`
[23] `https://github.com/ZhangTT-race/CVPR2020-MultiModal`
[24] `https://github.com/qyxqyx/FAS_Chalearn_challenge`
[25] `https://github.com/skJack/challange.git`

Table 11 The results of the multi-modal track in the CVPR2020 challenge [44]. Avg±Std indicates the mean and variance across the three folds and the best results are shown in bold.

Team Name	Threshold	FP	FN	APCER(%)	BPCER(%)	ACER(%)	Rank
BOBO	0.95±0.02	19±11	4±2	1.05±0.62	**1.00±0.66**	**1.02±0.59**	1
Super	1.0±0.00	11.33±7.76	11±6	0.62±0.43	2.75±1.50	1.68±0.54	2
Hulking	0.98±0.02	58±35	4±4	3.25±1.98	1.16±1.12	2.21±1.26	3
Newland-tianyan	1.00±0.00	4±4	17±12	**0.24±0.25**	4.33±3.12	2.28±1.66	4
ZhangTT	0.87±0.07	56±51	17±17	3.11±2.87	4.41±4.25	3.76±2.02	5
Harvest	0.92±0.04	104±84	13±12	5.77±4.69	3.33±3.21	4.55±3.82	6
Qyxqyx	0.95±0.05	92±142	26±23	5.12±7.93	6.66±5.86	5.89±4.04	7
Skjack	0.00±0.00	1012±447	47±45	56.24±24.85	11.75±11.37	33.99±7.08	8
Baseline	0.39±0.52	872±463	62±43	48.46±25.75	15.58±10.86	32.02±7.56	

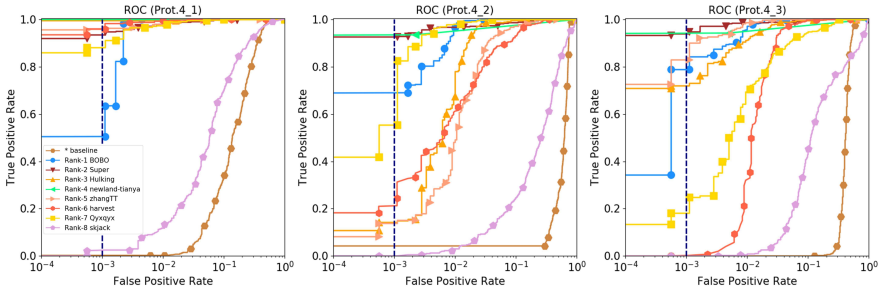

Fig. 5 The ROC curves of nine teams in the multi-modal track [44]. From left to right are the ROC curves on protocol 4_1, 4_2 and 4_3, respectively.

additional modalities are indeed useful in improving robustness of face PAD under the challenging cross-ethnicity and cross-PAI conditions.

The ROC curves of the solutions in multi-modal track are shown in Fig. 5. From the Table 11 and Fig. 5, we can find that even though the ACER values of the top two algorithms were relatively close, the stability of the team Super (brown curve) is better than that of the team BOBO (blue curve). For instance, the TPR@FPR=10^{-3} values for Super and Newland-tianyan were better than that of BOBO on all three sub-protocols. In other words, compared with the BCE loss based solution, the depth-wise supervision in team BOBO's solution might cause larger bias between metrics ACER and TPR@FPR=10^{-3}.

Discussion: From Tables 9 and 11 of the competition results, we can find that the EER thresholds computed on the validation set for both unimodal and the multi-modal track were generally high, indicating that the proposed algorithms might easily make over-confident or biased decisions. The reason behind this might be two-fold: 1) the biased distributions of the CASIA-

SURF CeFA dataset, *e.g.*, as the environments for the attack samples were more diverse, while bona fide samples were usually recorded indoor, and 2) the lack of generalization when the algorithm faces unknown PA types and ethnicities. Moreover, rethinking the evaluation metric for ranking the solutions is necessary. It can be seen from the both tracks that some solutions (*e.g.*, the team BOBO) could achieve excellent ACERs but, on the other hand, unsatisfying ROC curves, especially at operating points with low FPR.

2.3 CelebA-Spoof Challenge on Face Anti-Spoofing (ECCV2020)

Despite both the CVPR2019 [45] and CVPR2020 [43] competitions were successful in benchmarking the generalization of unimodal and multi-modal face PAD methods in challenging settings, the amount of data (number of images $< 150,000$ and subjects ≤ 1000) and domain diversity (only indoor conditions) of these two previous contests were still limited for evaluating the performance of FAS methods 'in the wild'. Recently, a large-scale face PAD dataset, namely CelebA-Spoof [93], containing 625,537 face images of 10,177 subjects, was released. It is still the largest publicly available face PAD dataset in terms of the number of images and subjects. Leveraging the CelebA-Spoof dataset, the CelebA-Spoof Challenge on Face Anti-Spoofing [94] was organized in conjunction with the European Conference on Computer Vision (ECCV) 2020 Workshop on Sensing, Understanding and Synthesizing Humans[26]. The goal of this competition was to boost the research on large-scale face anti-spoofing.

The ECCV2020 challenge was also hosted in the CodaLab platform[27]. After registering to the competition, each team was allowed to submit their models to the Amazon Web Services (AWS) and allocated with one 16 GB Tesla V100 GPU to perform online evaluation on the hidden test set. The encrypted prediction files, including the results for each data sample in the hidden test set, were sent to the teams via an automatically generated email after their requested online evaluation was finished. The teams were required to upload their encrypted prediction files to the CodaLab platform for ranking the algorithms.

The ECCV2020 challenge lasted for nine weeks from August 28, 2020 to October 31, 2020. During the contest, the participants had access to the public CelebA-Spoof dataset and were restricted to use only the public CelebA-Spoof training dataset for building their models. The results of the challenge were announced on February 10, 2021. A total number of 134 participants registered for the competition, while 19 teams made valid submissions in the end. The details and results of top five teams are shown in Table 12.

[26] `https://sense-human.github.io/`

[27] `https://competitions.codalab.org/competitions/26210`

It is surprising to see that all top three teams were from industry, indicating even increasing attention on PAD for real-world AFR applications. It is worth noting that the top three teams achieved TPR=100%@FPR=5*10^{-3}, indicating the effectiveness of the solutions for large-scale face PAD on the CelebA-Spoof dataset.

Table 12 Final results of the top-5 teams in the ECCV2020 challenge.

Ranking	Team	User	Affiliation	TPR @$FPR = 10^{-3}$	TPR @$FPR = 5*10^{-3}$	TPR @$FPR = 10^{-6}$
1	ZOLOZ	ZOLOZ	ZOLOZ	1.00000	1.00000	**1.00000**
2	MM	liujeff	Meituan	1.00000	1.00000	0.99991
3	AFO	winboyer	Meituan	1.00000	1.00000	0.99918
4	k_	k_	-	0.99973	0.99927	0.98026
5	SmartQ	SmartQ	-	0.99963	0.99872	0.96938

Fig. 6 Representative samples and attributes in the CelebA-Spoof dataset [93].

2.3.1 Dataset

The ECCV2020 challenge employed the CelebA-Spoof dataset [93] for training and evaluation purposes. The CelebA-Spoof is a large-scale face PAD dataset that has 625,537 images corresponding to 10,177 subjects, including 43 rich attributes on face, illumination, environment, and PA types. Bona fide facial images were selected from the CelebA dataset [50] but they were also manually examined to find and remove possible "attack" samples, including posters, advertisements and artistic drawings. The corresponding attack

samples with four different PAIs (*i.e.*, 2D print, cut paper, video-replay and 3D paper mask) were collected and annotated to form the CelebA-Spoof database. Among the 43 rich attributes, 40 attributes describe the bona fide images, including all facial components and accessories (*e.g.*, skin, nose, eye, eyebrow, lip, hair, hat, and eyeglasses), while the three remaining attributes describe the attack samples, including PAI, environments (*e.g.*, indoor and outdoor) and illumination conditions (*e.g.*, strong, weak, back, dark and normal). Some typical samples in the CelebA-Spoof database are shown in Fig. 6.

2.3.2 Evaluation protocol and metrics

The training, validation, and test sets of the original CelebA-Spoof database were split into subject-disjoint folds with a ratio of $8:1:1$. The compilation of the hidden test data devised for the ECCV2020 challenge was as the same as for the public test set. All the teams participating in the competition were restricted to train their algorithms using only the training subset of the publicly available CelebA-Spoof dataset, thus the use of both proprietary and public external datasets was explicitly forbidden.

Unlike in the two previous contests (*i.e.*, CVPR2019 [45] and CVPR2020 [43]), TPR@FPR based evaluation criteria were adopted for the ECCV2020 challenge. The TPR@FPR=5^{-3} determined the final ranking but also TPR@FPR=10^{-3} and TPR@FPR=10^{-4} values were reported. In the case if the TPR@FPR=5^{-3} for two submitted algorithms were the same, the one with higher TPR@FPR=10^{-4} would rank better.

Table 13 Summary of the top-ranked solutions [94] in the ECCV2020 challenge.

Team	Input	Model	Ensemble strategy
ZOLOZ	Face	FOCUS, AENet, ResNet, Attack types, Noise Print	Heuristic voting strategy
MM	Face + Patches	CDCN++, LGSC, SE-ResNet50, EfficientNet-B7, SE-ResNeXt50	Weight-after-sorting
AFO	Patches	CDCN, CDC-DAN, SE-ResNeXt26, Light-weighted Network	Weighted summation

2.3.3 Results and Discussion

In this section, we first analyse the top three solutions as well as their results, and then we discuss the algorithms and challenge in general.

Analysis on the top three solutions: Table 13 summarizes the FAS solutions of the top three teams. It is not surprising that all the best-performing solutions exploited ensembles of multiple deep models to achieve more robust performance. As the use of external training data in addition to the competition dataset was explicitly forbidden, the features of a single model can

easily overfit and learn specific attack cues, while stacking the feature representations of multiple (deep) models can be more generalizing to alleviate this issue. To be more specific, the two best teams ZOLOZ and MM utilized more advanced ensembling strategies compared with the team AFO considering only straightforward weighted summation. The team ZOLOZ proposed a heuristic voting scheme at the score level to form robust combinations of different models, whereas the team MM proposed a novel 'weighting-after-sorting' strategy based on particle swarm optimization (PSO) [36] algorithm for their model ensembles. All three teams utilized at least five different deep model architectures, of which some (*e.g.*, FOCUS [94], AENet [93], Noise Print [33], CDCN [87], CDCN++ [87], LGSC [19], CDC-DAN [94]) were aiming at capturing fine-grained (pixel-wise) fidelity characteristics, while some others (*e.g.*, ResNet [26], SE-ResNet [29], EfficientNet [70], ResNeXt [78]) focused on extracting semantic cues that could be complementary to improve face PAD performance. As for the model inputs, both whole facial images and local image patches were utilized by team MM, while the team ZOLOZ and AFO considered only the whole facial images and local image patches, respectively. It can be seen from the Table 12 that all top three teams achieved excellent PAD performance, reaching TPR>0.999@FPR=10^{-6}, indicating the effectiveness of deep multi-model ensembles on the competition data.

Discussion: Although the aforementioned best solutions achieved very promising results on the CelebA-Spoof dataset, there were still some shortcomings with the ECCV2020 challenge. The hidden test set is rather similar compared with the training data because the CelebA-Spoof dataset was simply divided into subject-disjoint training, development and test folds, thus not explicitly taking into account specific known issues related to domain generalization or unknown PAs. Furthermore, there were no restrictions on the size or number of deep models, which was also disappointing from real-world deployment point of view. Finally, compared with previous two face PAD challenges (*i.e.*, CVPR2019 [45] and CVPR2020 [44]), detailed ablation studies (*e.g.*, impact of each sub-model prior stacking) were missing, as well as the source codes of the solutions were not made public available in the ECCV2020 challenge, thus limiting the transparency and, consequently, usefulness of the whole competition to the FAS community.

2.4 LivDet-Face 2021 – Face Liveness Detection Competition (IJCB2021)

Recent literature surveys (*e.g.*, [52, 84]) have concluded that both handcrafted and deep features yield in satisfying classification performance in identifying known PAIs but often fail to detect unknown PAIs and more sophisticated face artefacts. Therefore, continuous efforts are necessary to up-

date face anti-spoofing algorithms to detect rapidly evolving PAs. Although earlier the CVPR2020 cross-ethnicity face PAD challenge considered also a cross-PAI setting (*i.e.*, training on the video-replay attacks and testing on the print and mask PAIs), the types and quality of the unknown PAIs were still limited from the generalized PAD point of view. To address this issue, the LivDet-Face 2021 – Face Liveness Detection Competition [58] was organized in conjunction with IJCB2021.

The registration for the IJCB2021 LivDet-Face competition began on February 15, 2021 and ended on April 25, 2021, while the final submission deadline was April 30, 2021. The objective of the competition was to evaluate the performance of the state-of-the-art face PAD algorithms against traditional and novel PAIs. The competition had two separate tracks for image and video data, and the competitors were allowed to participate in both tracks. Different from all previous competitions, IJCB2021 LivDet-Face contest did not provide any specific training dataset to the participants, thus the competitors were free to use any proprietary and/or publicly available data to train their algorithms, replicating more realistic and challenging practical AFR application scenarios.

Both academic and industrial organizations were welcome to participate in IJCB2021 LivDet-Face competition anonymously or non-anonymously. In total, thirty international teams registered to the competition, including ten submissions for the image track and six submissions for the video track. Finally, six submission could be successfully tested by the organizers for the image track and five submission for the video track. Unsuccessful tests were due to software issues, which were communicated with the participants.

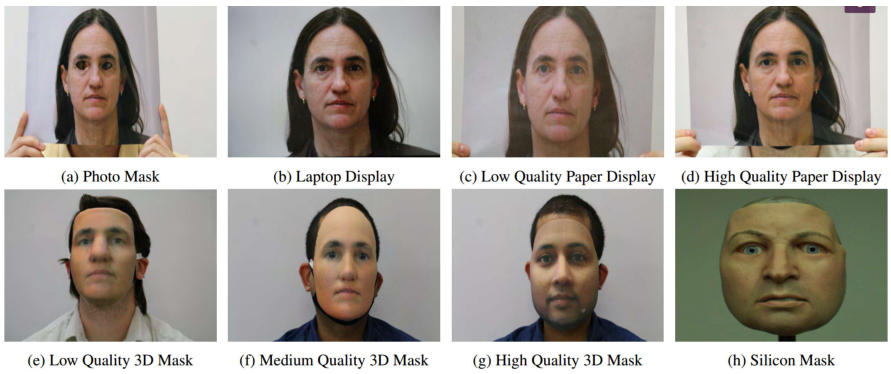

Fig. 7 Samples of each PA type present in the IJCB2021 LivDet-Face test set [58].

2.4.1 Dataset

No official training dataset was shared by the organizers of the IJCB2021 LivDet-Face competition. Instead, participants were encouraged to use any data available to them (*i.e.*, from both public and proprietary sources) to train and tune their algorithms. The organizers shared only few (no more than two) examples of the known PAIs to familiarize the competitors with the test dataset, while the remaining samples of the disclosed PAI types were considered as unknown to the competitors. The test dataset used in the IJCB2021 LivDet-Face competition was a combination of data from two of the organizing institutions: Clarkson University (CU) and Idiap Research Institute. The dataset consisted of 724 images (135 bona fide and 589 PAI samples) and 814 videos (125 bona fide and 689 PAI samples) for the image and video tracks, respectively. The data was collected from in total 48 live subjects using altogether five different sensors (digital single-lens reflex (DSLR) camera, iPhone X, Samsung Galaxy S9, Google Pixel and Basler aA1920-150uc). The length of the videos in the test dataset was up to six seconds. Eight PAIs for the image track and nine PAIs for the video track were included in the dataset (see Table 14 for a summary and Fig. 7 for typical examples for each PAI type).

Regarding the 2D PAIs, 100 low-quality (LQ) print paper, 100 high-quality (HQ) photo paper attacks, 100 static display (SD) and video-replay (VR) attacks on laptop screen, and 100 2D photo mask attacks were collected from 25 live subjects using four different sensors. Specifically, the 100 video-replay attacks were used only as an unknown PAI for the video category of the competition, thus were not introduced in the few validation samples that were shared with the competitors. In addition to the 2D attacks, three different qualities of 3D masks (low, medium, and high) as well as silicon mask attacks were included in the test dataset for the both competition tracks. A total number of 24 images and 24 videos of low-quality (LQ) 3D masks were created corresponding to six live subjects. Also, 12 medium-quality (MQ) 3D mask and 12 high-quality (HQ) images/videos corresponding to three live subjects were included in the test dataset. The HQ 3D masks were kept as an unknown PAIs for the competitors in the test dataset. In total, 141 image/video samples of wearable 3D silicon masks were collected using five different sensors.

2.4.2 Evaluation protocol and metrics

During the IJCB2021 LivDet-Face competition, at least two samples of the majority of the considered PAIs (except the high-quality 3D masks and video-replay attacks) were shared with the competitors of both image and video tracks as a small validation set to fine-tune their algorithms. The performance of an algorithm for each sample was determined by an output ("liveness")

Table 14 Summary of the test dataset used in the IJCB2021 LivDet-Face [58].

Class Types of PAIs	Images	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

score ranging between 0 and 100 with a threshold of 50, while the score of 1000 indicates undetected samples. The test samples with scores less than 50 were classified as PA, whereas the scores of 50 and above were classified as bona fide. Most of the competitors normalized their output scores at their end and provided a score of 0, 100 or 1000 (if undetected) based on their classification output. If the submitted algorithms provided a score of 1000 for the PAs, the result was considered as a correct decision as the algorithm was able to reject an attack, thus is not included in attack presentation classification errors. A score of 1000 for bona fide samples were considered as incorrect, thus was accumulated in bona fide classification errors.

Following the recommendations of the ISO/IEC 30107-3 [32] standard, APCER, BPCER, and ACER (see, Section 2.1.2) were used as the evaluation metrics like in the CVPR2020 challenge. Since all algorithms were required to deliver normalized liveness scores in the range of $0 - 100$, $t = 50$ was used as the decision threshold to calculate the APCER and BPCER. The final ranking of teams was based on the ACER calculated over all the test samples.

Table 15 Summary of the solutions of in IJCB2021 LivDet-Face [58].

Team	Training/validation	Model	Ensemble strategy
Fraunhofer IGD	CRMA for validation	12 models (DeepPixBis, ResNeXt, etc.)	FDR weights
SiMiT Lab	Replay Mobile, SiW, Oulu-NPU, 3DMAD for training	DeepPixBis, EfficientNet-B7	Mean score fusion
CLFM	-	CDCN	-
FaceMe	-	3 models (DepthNet, Digital signal processor, etc.)	-
little tiger	Glint360k for pre-training	5 models (ResNet50, ResNext26, CDCN, etc.)	-
NTNU Gjøvik	SWAN, CASIA-FASD, NTNU-Silicon Mask for training	6 models (Resnet18, Resnet50, InceptionV3, 9 VGG1, VGG16, Alexnet) with 2 linear SVM	Majority voting

2.4.3 Results and Discussion

In this section, we first analyse the solutions of both tracks. Then, the result analysis of the image and video tracks is presented. Finally, we discuss the algorithms and the challenge in general.

Analysis on the solutions: Table 15 summarizes the FAS solutions of the six teams. The teams SiMiT Lab and NTNU Gjøvik utilized several public datasets for training, including the 2D PAI (*e.g.*, Replay-Mobile [9], SiW [49], Oulu-NPU [3], SWAN [59], CASIA-FASD [95]) and 3D mask attack (*e.g.*, 3DMAD [17] and NTNU-Silicon Mask [60]) data. The team Fraunhofer IGD improved the generalization of their algorithm by using the 50 attacks and bona fide samples from the Real Mask Attack Database (CRMA) [18] as unknown development data to tune the decision threshold.

Like in the previous three challenges discussed already in this chapter, most of the solutions in IJCB2021 LivDet-Face competition used ensembles of multiple deep models to achieve more robust performance. The teams Fraunhofer IGD, SiMiT Lab, FaceMe, little tiger and NTNU Gjøvik stacked 12, 2, 3, 5 and 6 models in their submitted systems to make the final PAD decision. Three kinds of models were considered in these solutions: 1) hand-crafted models with digital signal processing, 2) pixel-wise supervised models (*e.g.*, DeepPixBis [21], DepthNet [49] and CDCN [86]), and 3) BCE supervised models (*e.g.*, ResNet [26], ResNeXt [78], VGG [65], Inception [69] and Alexnet [40]). The team Fraunhofer IGD adopted Fisher discriminant ratio (FDR) weights [11] for 12 models to get a combined face PAD decision, while the team NTNU Gjøvik considered a simple majority voting strategy to make decision from six models, and the team SiMiT Lab fused the scores from two models with shuffled patch-wise supervision [35]. Some competitors (*e.g.*, teams CLFM and FaceMe) were lacking these kinds of details in their method descriptions, thus making it impossible to draw conclusions on some factors in our result analysis, including data fusion approaches and training/tuning data.

Table 16 Results of the image track in the IJCB2021 LivDet-Face competition. The APCER is respectively calculated for each type of PAI, and then averaged for final ACER calculation.

Team	Paper		Replay	2D Mask	3D Mask				BPCER(%)	ACER(%)	Ranking
	LQ	HQ			LQ	MQ	HQ	Silicon			
Fraunhofer IGD	**0**	24	45	14.7	4.17	8.33	**14.29**	**16.31**	**15.33**	**16.47**	1
CLFM	6.06	**10**	8	**5.88**	**0**	16.67	21.43	34.75	24.08	18.71	2
FaceMe	22.22	11	**3**	11.76	66.67	66.66	50	57.45	16.06	20.72	3
little tiger	41.41	52	4	58.82	54.17	25	28.57	82.98	21.17	33.92	4
SiMiT Lab	7.07	18	43	15.68	16.66	**0**	42.85	80.85	51.09	42.05	5
Anonymous	78.78	86	77	89.21	87.5	83.33	100	98.58	16.79	49.35	6

Results of the image track: Table 16 summarizes the results of the image track. The team Fraunhofer IGD was the winner, obtaining the lowest ACER = 16.47%, followed by the team CLFM with a narrow margin in ACER = 18.71%. The winning team Fraunhofer IGD achieved the lowest BPCER = 15.33% among the six competitors. The six competitors achieved highly varying performances across the different PAI types. The algorithm submitted by the team Fraunhofer IGD detected all the low-quality paper display attacks, and the algorithm by CLFM successfully detected all the low-quality 3D mask samples. Team CLFM's algorithm also performed the best with APCER = 10% for high-quality photo paper display samples but achieved an unsatisfying BPCER = 24.08%. The team FaceMe, who achieved third place in the image track, achieved the best APCER = 3% for laptop display samples and second-best BPCER = 16.06%. The team SiMiT Lab successfully detected all the medium-quality 3D mask samples with APCER = 0%, which was best among all the competitors, but achieved also the worst BPCER = 51.09%.

By comparing the performance of the top two ranked solutions in the image track, it is obvious that the models performed better against low-quality PAIs than higher quality PAIs. The team Fraunhofer IGD obtained APCER = 0% for the low-quality paper display and APCER = 24% for high-quality paper display. Similarly, the team CLFM obtained APCER = 6.06% for low-quality paper display, and APCER = 10% for high-quality paper display. The same trend can be observed for the different quality of 3D face masks. The team Fraunhofer IGD obtained APCER = 4.17% for low-quality 3D masks compared with APCER = 8.33% for medium-quality 3D masks, APCER = 14.29% for high-quality 3D masks and APCER = 16.31% for high-quality silicon masks. Similarly, the team CLFM achieved APCER = 0% against low-quality 3D masks compared with APCER = 16.67% for medium-quality 3D masks, APCER = 21.43% for high-quality 3D masks and APCER = 34.75% for high-quality silicon masks.

Table 17 Results of the video track in the IJCB2021 LivDet-Face competition.

Team	Paper		Replay		2D Mask	3D Mask				BPCER(%)	ACER(%)	Ranking
	LQ	HQ	SD	VR		LQ	MQ	HQ	Silicon			
FaceMe	8	10.10	18	16	6.93	40	45.45	38.46	9.22	14.29	**13.81**	1
Fraunhofer IGD	**1**	25.25	29	9	1	4	9.09	**0**	12.77	16.67	14.49	2
CLFM	4	**4.04**	**8**	**1**	**0**	**0**	27.27	7.69	**1.42**	39.68	21.49	3
NTNU Gjøvik-V1	50	59.60	83	75	18.81	36	18.18	46.15	21.28	**4.76**	26.51	4
NTNU Gjøvik-V2	5	9.09	32	20	1	**0**	**0**	**0**	33.33	51.59	34.05	5

Results of the video track: Table 17 summarizes the results of the five solutions in the video track of the IJCB2021 LivDet-Face competition. The team FaceMe was the winner with the ACER = 13.81% followed by the

team Fraunhofer IGD with a narrow margin in ACER = 14.49%. The lowest BPCER = 4.76% was achieved by the team NTNU Gjøvik. The team CLFM performed well in detecting paper and video-replay attacks (with APCER < 3.30% for all scenarios) but ranked third due to the bad BPCER performance (39.68%). The team NTNU Gjøvik-V2 performed well against 3D face mask attacks, achieving APCER = 0% for the three different types of 3D masks but also the worst BPCER (51.59%).

It can be observed that the top two solutions performed better against low-quality PAIs compared to higher-quality PAIs. For example, the performance of the team Fraunhofer IGD against low-quality paper display was APCER = 1%, while APCER = 25.25% was obtained against the high-quality paper display attacks. The same trend can be observed for the different quality of 3D face masks as well. The performance of the team Fraunhofer IGD against low-quality 3D masks was APCER = 4%, which was obviously better than that of the medium-quality (APCER = 9.09%) and high-quality silicon masks (APCER = 34.75%). However, the performance of high-quality 3D masks was better than that of any other 3D mask category with APCER = 0%.

Discussion: Compared with the three earlier competitions [45, 43, 94] introduced in this chapter, a significant degradation in the overall performance can be observed. This can be due to several factors, such as: 1) increased complexity in the test dataset with nine different PAI types, 2) introduction of three novel attack types with limited availability, or not covered at all, in the public datasets, 3) lack of specific competition training dataset, *i.e.*, choice of training data up to the competitors, 4) domain shift between the training and test conditions in terms of environmental factors, sensors, quality of PAIs, and the introduction of unknown PAIs. The results of this competition indicate that generalized face PAD is still far away from a solved research problem.

Despite its important findings on the current state of face PAD 'in the wild', the IJCB2021 LivDet-Face competition still had some shortcomings. First, as there were no pre-defined training sets and, consequently, no restrictions on the diversity and scale of the private or public datasets, it is unfair to evaluate the performance of the different approaches. It is simply impossible to explore the hidden reasons behind the differences in performance and to tell if an algorithm actually better than another one, or is it in fact a matter of the amount and quality of training data. Second, no ablation studies or open-source code for most of the solutions were provided, which again limits the transparency and, consequently, usefulness of the results and findings to the FAS community. Finally, as the complexity of the proposed systems was not limited, mainly huge ensembles of deep models were adopted by the participants, while none of the teams proposed interesting novel efficient face PAD approaches.

2.5 3D High-Fidelity Mask Face Presentation Attack Detection Challenge (ICCV2021)

As seen in two recent competitions (*i.e.*, CVPR2020 [43] and IJCB2021 LivDet-Face [58]), face PAD performance of state-of-the-art methods drops significantly under unknown 3D mask attacks. However, the previous competition datasets contained in general only a limited number of samples and types of 3D facial masks, thus there is a large gap in between the existing benchmarks and 3D mask attack detection in real-world conditions. To alleviate the threats posed by 3D mask attacks and to improve the reliability of face PAD methods under emerging types of 3D mask attacks in various scenarios, the 3D High-Fidelity Mask Face Presentation Attack Detection Challenge [46] was organized in conjunction with the International Conference on Computer Vision (ICCV) in 2021 using the very recently constructed CASIA-SURF High-Fidelity Mask (HiFiMask) dataset [47].

The ICCV2021 challenge was conducted using the CodaLab platform[28], attracting 195 teams from all over the world. A summary with the names and affiliations of teams that entered the final stage of the contest is shown in Table 18. Again, the majority of the final participants came from industrial institutions, and all the six best-performing teams represented companies. This indicates clearly that mask attack detection is no longer limited to academic research but also a crucial problem in real-world AFR applications. The results of the top three teams were far better than the baseline results [47], thus greatly improving the performance of 3D high-fidelity mask attack detection.

2.5.1 Dataset

HiFiMask [47] is currently the largest 3D face mask PAD dataset, consisting of 54,600 videos corresponding to 75 subjects with three skin tones, including 25 subjects in yellow, white, and black, respectively. The database contains three high-fidelity masks for each identity, which are made of transparent, plaster and resin materials, respectively. During the acquisition process, six complex scenes were considered for recording the videos (*i.e.*, white light, green light, periodic three-colour light, outdoor sunshine, outdoor shadow, and motion blur). For each scene, there are six videos captured under different lighting conditions (*i.e.*, normal, dim, bright, back, side, and top) to explore the impact of directional lighting. Periodic lighting within [0.7, 4] Hz in the first three scenarios (see, the first three columns of the last row in Fig. 8 for examples) tries to mimic the natural human pulse variations to fool remote photoplethysmography (rPPG) based mask detection technology (*e.g.*, [42]). Finally, seven mainstream imaging devices (*i.e.*, iPhone11, iPhone X, MI10, P40, S20, Vivo and HJIM) were utilized for recording the videos in order to

[28] `https://competitions.codalab.org/competitions/30910`

Table 18 Names, affiliations and final ranking of the teams participating in the ICCV2021 challenge.

Ranking	Team Name	Leader Name	Affiliation
1	VisionLabs	Oleg Grinchuk	Visionlabs.ai
2	WeOnlyLookOnce	Ke-Yue Zhang	Tencent Youtu Lab
3	CLFM	Samuel Huang	FaceMe
4	oldiron666	Zezheng Wang	Kuaishou Technology
5	Reconova AI-LAB	Mingmu Chen	Reconova Technology
6	inspire	Jiang Hao	Bytedance Ltd.
7	Piercing Eyes	Hyokong	National University of Singapore
8	msxf_cvas	Liang Gao, MaShang	Consumer Finance Co.,Ltd
9	VIC_FACE	Cheng Zhen	Meituan
10	DXM-DI-AI-CV-TEAM	Weitai Hu	Du Xiaoman Financial
11	fscr	Artem Petrov	Peter the Great St. Petersburg Polytechnic University
12	VIPAI	Yao Xiao	Zhejiang University
13	reconova-ZJU	Zhishan Li	Zhejiang University
14	sama_cmb	Yifan Chen	Chinese Merchants Bank(CMB)
15	Super	Yu He	Technische Universität München
16	ReadFace	Zhijun Tong	ReadFace
17	LsyL6	Dongxiao Li	Zhejiang University
18	HighC	Minzhe Huang	Akuvox (Xiamen) Networks Co., Ltd.

ensure high resolution and imaging quality corresponding to modern mobile devices. The original videos were not provided due to huge amount of data. In order to decrease the size of the dataset, the organizers sampled every tenth frame of each video and applied a fast face detector [92] to remove most of the background information from the sampled video frames, thus the final competition data consisted of coarsely pre-cropped facial images. Some typical samples of pre-processed video frames in the HiFiMask dataset are presented in Fig. 8.

In order to increase the difficulty of the competition and meet the real-world deployment requirements, an 'open-set' test protocol was utilized to comprehensively evaluate the discriminative and generalization power of face PAD algorithms. In other words, the training and development sets contained only subsets of common mask types and operating scenarios, while there were more general mask types and scenarios in the test set. Thus, the distribution of test set was more complicated compared to the training and development sets in terms of mask types, scenes, lighting, and imaging devices. Such 'open-set' protocol considers explicitly both 'seen' and 'unseen' domains as well as mask types for evaluation, which is also more valuable from real-world face PAD deployment point of view.

As shown in Table 19, every skin tone, part of mask types, such as transparent and resin materials (1, 3), part of scenes, such as white light, outdoor sunshine and motion blur (1, 4, 6), part of lighting conditions, such as normal, bright, back and top (1, 3, 4, 6), and part of imaging devices, such as iPhone

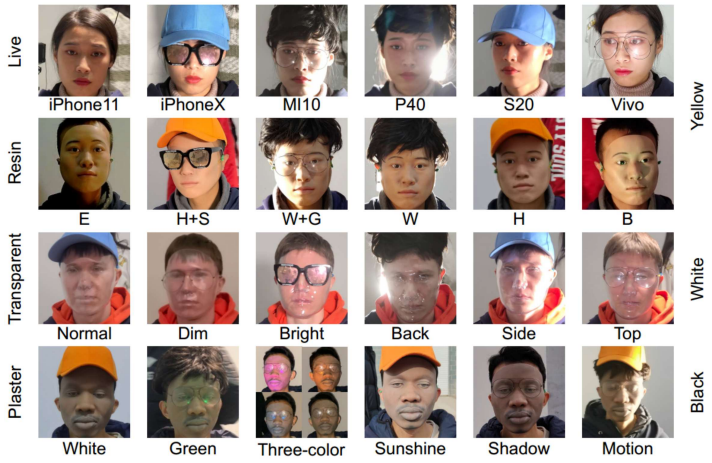

Fig. 8 Samples from the HiFiMask dataset [47]. The first row shows six kinds of imaging sensors. The second row shows six kinds of appendages, of which E, H, S, W, G, and B are the abbreviations for 'empty', 'hat', 'sunglasses', 'wig', 'glasses', and 'messy background', respectively. The third row shows six kinds of illumination conditions, and the fourth row represents six deployment scenarios.

Table 19 Statistical information for the protocols used in the ICCV2021 challenge. Note that 1, 2 and 3 in the third column mean 'transparent', 'plaster' and 'resin' masks, respectively. The numbers in the fourth, fifth, and sixth columns are explained in Section 2.5.1.

Subset	Subject	Mask type	Scene	Light	Sensor	# Live num.	# Mask num.	# All num.
Train	45	1&3	1&4&6	1&3&4&6	1&2&3&4	1,610	2,105	3,715
Dev	6	1&3	1&4&6	1&3&4&6	1&2&3&4	210	320	536
Test	24	1∼3	1∼6	1∼6	1∼7	4,335	13,027	17,362

11, iPhone X, MI10, P40 (1, 2, 3, 4) are included in the training and development subsets. All skin tones, mask types, scenes, lighting conditions and imaging devices are present in the test subset. For clarity, the organization of the dataset and quantity of videos for each sub-protocol of the challenge are shown in Table 19.

2.5.2 Evaluation protocol and Metrics

The challenge comprised two stages as follows:

Development phase: *(April 19, 2021 – June 10, 2021).* During the development phase, the participants had access to the labelled training data and unlabelled development data. The samples in the training set were labelled with the bona fide, two types of masks (1, 3), three types of scenes (1, 4, 6), four kinds of lighting conditions (1, 2, 4, 6) and four imaging sensors (1, 2,

3, 4). The labels of the validation data were not provided to the participants in development phase. Instead, participants could build their models on the labelled training data and then submit their predictions on the development data and receive immediate feedback via the competition leaderboard in the CodaLab platform.

Final phase: *(June 10, 2021 – June 20, 2021)*. During the final phase, the labels for the development set were also made available to the participants and the unlabelled test set was also released. The competitors had to make their predictions on the test samples and upload their solutions to the challenge platform. The organizers did then rerun the best-performing algorithms, and the final ranking of the participants was obtained based from the verified results on the test data. To further facilitate the outcome and findings of the competition to the face PAD community, the best-performing teams were encouraged to make their source code publicly available and provide fact sheets describing their solutions.

Evaluation metrics: Similarly to the previous two competitions (*i.e.*, CVPR2020 [43] and IJCB2021 LivDet-Face [58]), the ISO/IEC 30107-3 [32] standardized metrics APCER, BPCER, and ACER were selected as the evaluation criteria, and the ACER was the leading evaluation criterion for ranking the submitted systems in the ICCV2021 challenge. The ACER threshold on the test set was determined based on the EER operating point on the development data.

Table 20 Summary of the solutions in the ICCV2021 challenge [46]. 'DSFD' and 'DBO' denote dual shot face detector [41] and deep bilateral operator [81], respectively.

Team Name	Pre-processing	Backbone	Branch	Loss
VisionLabs	DSFD detector	EfficientNet-B0	6	BCE
WeOnlyLookOnce	DSFD detector	ResNet12	2	3-class CE with label smoothing
CLFM	Crop mouth region	CDCN++	1	BCE
Oldiron666	Whole face	Resnet6 (SimSiam)	1	BCE+MSE+Contrastive loss
Reconova-AI-Lab	RetinaFace detector	ResNet50+YoLoV3-FPN	3	BCE+Focal loss
inspire	RetinaFace detector	SE-ResNeXt101	1	BCE+MSE+Contrastive loss
Piercing Eye	Face detection	CDCN	2	Depth regression+BCE
msxf cvas	Face detection+alignment	ResNet34	1	4-class CE loss
VIC FACE	Face detection	CDCN with DBO	1	Depth regression
DXM-DI-AI-CV-TEAM	-	DepthNet	1	BCE with meta learning

2.5.3 Results and Discussion

In this section, we first summarize solutions in the 3D Mask face PAD challenge. Then, we provide our result analysis. Finally, the algorithms and the challenge are discussed in general.

Analysis of the solutions: Table 20 summarizes the solutions of the most teams participating in the ICCV2021 challenge. The source code of the winning team VisionLabs was released[29] while the detailed descriptions as well as the ablation studies of the top three ranked solutions can be found in [6, 22, 31]. Different from all the previous competitions, the ICCV2021 challenge accepted only the results obtained with single (deep) model based systems, thus ensemble strategy with multiple models was explicitly prohibited. As a result, several solutions (*e.g.*, the teams VisionLabs, WeOnlyLookOnce, Reconova-AI-Lab and Piercing Eye) developed multi-branch architectures, which aimed at capturing more diverse PAD-specific feature representations. Also, team VIC FACE proposed a novel architecture in its solution, which integrated the deep bilateral operator in the original CDCN [85] in order to learn more intrinsic features via aggregating multi-level bilateral macro and micro-texture information. Most of the teams adopted DSFD [41] or RetinaFace [15] detector for pre-processing to localize more fine-grained facial region and to filter out partial low-quality face mask attacks. The teams considered face PAD mainly as a binary classification task using BCE loss or as a depth regression problem, but two teams (*i.e.*, WeOnlyLookOnce and msxf cvas) also forced the models to learn more mask type-aware features via fine-grained multiple class CE loss.

Result analysis: The results and ranking of the top 18 teams are shown in Table 21. The ACER performance of the top three teams were relatively close (within $< 3.3\%$). One major reason behind this is that the original pre-processed images in the HiFiMask dataset correspond to very coarsely cropped facial regions, including also outliers (*i.e.*, non-facial samples), due to the limited accuracy of the used efficient face detector [92]. Thus, the additional pre-processing with DSFD and mouth region cropping was very beneficial in removing the outlier samples affecting negatively the training of deep models and focusing on the details in the actual facial information discriminating attacks from bona fide samples. The team VisionLabs achieved the best BPCER $= 2.33\%$ with only 101 FN samples, while WeOnlyLookOnce had the lowest APCER $= 1.858\%$ with 242 FP samples, indicating that the features from multiple facial region based branches and vanilla/CDC branches benefited the spoof cue representations. Moreover, the ACER performance of all the teams were evenly distributed ranging from 3% to 10%, which not only indicates the rationality and selectivity of the 3D mask face PAD challenge but also demonstrated the value of the HiFiMask dataset for future face PAD studies.

Discussion: Some general observations on 3D mask face attack detection can be concluded based on findings of the ICCV2021 competition: 1) the accurate facial localization and subregion information was very beneficial, focusing on the actual discriminative local facial details and avoiding learning and extraction of irrelevant features for face PAD, and 2) multi-branch-based

[29] `https://github.com/AlexanderParkin/chalearn_3d_hifi`

Table 21 The final results and team rankings of the ICCV2021 challenge [46]. The best results are shown in bold.

Team Name	FP	FN	APCER(%)	BPCER(%)	ACER(%)	Rank
VisionLabs	492	**101**	3.777	**2.330**	**3.053**	1
WeOnlyLookOnce	**242**	193	**1.858**	4.452	3.155	2
CLFM	483	118	3.708	2.722	3.215	3
oldiron666	644	115	4.944	2.653	3.798	4
Reconova-AI-LAB	277	276	2.126	6.367	4.247	5
inspire	760	176	5.834	4.060	4.947	6
Piercing Eyes	887	143	6.809	3.299	5.054	7
msxf_cvas	752	232	5.773	5.352	5.562	8
VIC_FACE	1152	104	8.843	2.399	5.621	9
DXM-DI-AI-CV-TEAM	1100	181	8.444	4.175	6.310	10
fscr	794	326	6.095	7.520	6.808	11
VIPAI	1038	268	7.968	6.182	7.075	12
reconova-ZJU	1330	183	10.210	4.221	7.216	13
sama_cmb	1549	188	11.891	4.337	8.114	14
Super	780	454	5.988	10.473	8.230	15
ReadFace	1556	202	11.944	4.660	8.302	16
LsyL6	2031	138	15.591	3.183	9.387	17
HighC	1656	340	12.712	7.843	18	

feature learning was a widely used framework by the participating teams, which benefits from the shared low-level feature learning, capturing diverse separated multi-branch features for generalized 3D mask description. However, there were still some shortcomings in the ICCV2021 challenge. The fidelity and the collection of the 3D mask types were still limited. For instance, the face mask in the fourth row and second column of Fig. 8 has a very artificial appearance, while also more 3D attack types, such as paper and silicone masks, and wax faces should be considered. Furthermore, only every tenth frame for each video was provided. Such a low frame rate makes it impossible to recover facial physiological signals and, consequently, to study rPPG based 3D mask detection [48, 82], for instance.

3 Discussion

All the recent five competitions were successful in consolidating and benchmarking the current state of the art in face PAD. In the following, we provide

general observations and further discussion on the lessons learnt, model architectures and potential future challenges.

3.1 General Observations

It is apparent that the used datasets and evaluation protocols, and also the recent advances in the state of the art, reflect the trends in face PAD schemes seen in the different contests. The algorithms proposed in the context of the first two multi-modal competitions (*i.e.*, CVPR2019 [45] and CVPR2020 [43]) exploited the evident visual cues that we humans can observe in the multi-modal imaging data of the CASIA-SURF and CASIA-SURF CeFA datasets, including plain/structural discrepancies in the depth modality, material surface reflection differences in the NIR modality, and natural human movement in the dynamic modality. Most of the solutions adopted exhausted model and hyperparameter selection strategies for each modality and ensembling the best combination of different modalities and models for robust PAD performance in the multi-modal setting, *i.e.*, TPR=99.8739%@FPR=10e-4 on the CASIA-SURF and ACER = 1.02% on the CASIA-SURF CeFA datasets. In these two competitions, the performance gains relied mostly on powerful deep models and ensembling strategies, while the essence of multi-modal fusion was not truly explored. This gave, disappointingly, limited insight to the multi-modal face PAD community, thus it is necessary to rethink these kinds of multi-modal PAD settings, especially from the efficient fusion point of view, in the upcoming benchmarks.

In the latest two competitions (*i.e.*, IJCB2021 LivDet-Face [58] and ICCV2021 [46]) on unimodal colour camera based face PAD, generalized and intrinsic bona fide/attack cues and motion analysis were hardly used. The proposed RGB data based features were overfitting in the training data, thus generalized poorly to the test sets. One reason to the unsatisfactory performance is that the test sets included more unseen high-fidelity 3D mask types (*e.g.*, plaster masks in the ICCV2021 [46] and high-quality 3D masks in the IJCB2021 LivDet-Face [58] challenge). Although it was nice to see a diverse set of advanced deep learning based systems and further improved versions of the provided baseline method, it was a bit disappointing that entirely novel generalized face PAD solutions, *e.g.*, for zero-shot unseen attack (especially against 3D masks) detection, were not proposed. The best performances in the IJCB2021 LivDet-Face (ACER = 16.47%/13.81% for image/video tracks) and the ICCV2021 (ACER = 3.053%) competitions were still limited, considering the needs of real-world use cases. However, as seen in the CVPR2020 competition [43], the major issues with domain shift and unseen PAs can be at least partially alleviated by introducing depth and NIR sensors as these additional modalities provide more accurate 3D shape information for 2D

PAI detection and material reflection information for discriminating realistic 3D masks among other face artefacts from genuine human skin.

Unlike with the first three face PAD contests (*i.e.*, IJCB2011 [5], ICB2013 [8] and IJCB2017 [1]), it is interesting to observe that most of the participants came from the industry in the competitions organized from 2019. For example, the majority of the final participants (10 out of 13) of the CVPR2019 challenge [45] and (8 out of 11) of the unimodal track of the CVPR2020 challenge [43] came from industrial institutions. Also, in the most recent ICCV2021 challenge [46], the majority (13 out of 18) of the final participants came from companies, and the six top-performing teams were all from industry, indicating the steadily growing interest and need for practical and reliable face PAD solutions in commercial AFR products.

3.2 Lessons Learnt

The competitions have given valuable lessons on designing databases and test protocols, and competitions in general. In the CVPR2019 [45] and ECCV2020 [94]) challenges, the best performances on the test data reached TPR=99.8739%@FPR=10e-4 and TPR=100%@FPR=10e-6, respectively. However, the problem of face PAD has not been solved as the error rates (ACER = 16.47%/13.81%) of the more recent IJCB2021 LivDet-Face competition reveal that the state of the art in face anti-spoofing still suffers from significant generalization issues in unknown operating conditions. Therefore, we should also rethink the design of evaluation protocols in competitions like the CVPR2019 [45] and ECCV2020 [94]) challenges. In the CVPR2019 challenge, the provided validation data had similar distribution with the test data, and the validation set was also allowed to be included in training the face PAD models. Furthermore, the training, validation, and testing sets were split in subject-disjoint manner in the ECCV2020 challenge when the domain covariates (*e.g.*, biometric sensors and lighting conditions) and PAIs are too similar. As a result, even using the off-the-shelf deep models with powerful representation learning capacity, the participants could easily train an ensemble of several overfitting models in these two competitions, performing well not only on the training and validation sets but also on the test data. It is necessary to mimic the requirements of real-world biometric applications and to design and capture even more challenging and larger scale datasets with unknown domains and unseen PA types in the test sets.

TPR@FPR was utilized in the CVPR2019 and ECCV2020 challenges as the ranking criteria, while ACER was adopted in the three remaining recent face PAD competitions. Despite being widely used in large-scale biometric evaluations (*e.g.*, face recognition), TPR@FPR was first utilized in face PAD competitions due to emerging larger scale face PAD datasets (*e.g.*, the CASIA-SURF [91] with 295,000 frame samples and CelebA-Spoof [93] with

over 62,000 samples), which made the calculation of TPR@FPR=1e-4 or even TPR@FPR=1e-6 possible. In most of the competitions (4 out of 5), the results were reported using the mainstream metrics APCER, BPCER, and ACER recommended in ISO/IEC 30107-3 standard [32]. However, the selection of score threshold value for computing the ACER is worth discussing from both contest and database design point of views. For the CVPR2019 [45], CVPR2020 [43] and ICCV2021 [46] challenges, the ACERs for the test sets were determined by the EER operating point on the validation sets, while in the IJCB2021 LivDet-Face [58]) evaluation, the fixed threshold of 50 (normalized output liveness scores from 0 to 100) was utilized directly, and it was up to the participant how the scores were normalized within the valid range of liveness scores. The latter evaluation method with fixed range of liveness scores and ACER threshold sounds intuitive from interpretability and real-world applications point of views, and adequate considering the lack of specific validation set in the IJCB2021 LivDet-Face competition. The former approach seems to be more reasonable but with some drawbacks. For instance, significant discrepancies in performance can be observed among the ranked solutions when comparing their results using the ROC related TPR@FPR metrics and ACER. As an example, the second-best solution in terms of ACER ranking in Table 9 performed poorly in terms of TPR@FPR=1e-3 in Fig. 4. In practical biometric applications, the suitable operating point depends highly on the application context. However, when looking at a single ACER value, the misclassification rates between bona fide and attack classes are often ignored, even though they are a crucial piece of information. A method performing well in terms of ACER at the selected threshold might suffer from severely imbalanced APCER/BPCER ratio. For instance, one of the metrics might clamp to zero, while the other one can be relatively high. In this case, the ACER fails to point out that the PAD system is able to detect all attacks but rejects many bona fide samples (or vice versa), thus making it impossible to judge its performance trade-off in the APCER and BPCER, *i.e.*, security and usability. Examples of such system behaviour can be seen in Table 9, where the teams VisionLabs and Wgqtmac show *0.11%/5.33%* and *51.57%/0.66%* APCER/BPCER ratios, respectively.

The CVPR2019 [45], CVPR2020 [43] and ICCV2021 [46] challenges can be considered to be more transparent and fair due to the richer amount of publicly available details about the participating teams and the best-performing solutions, whereas the ECCV2020 [94] and IJCB2021 LivDet-Face [58] contests reported only limited information about the teams and evaluated solutions. In general, competitions should encourage the participants to provide authentic public details on the registered teams and open-source implementations, as well as detailed ablation analysis of the best-performing solutions. The authentic team information is useful for avoiding malicious registration situations where teams consisting of similar members from the same institution could make more submission entries, thus less cost of trial and error. The open-source codes and detailed ablation analysis would benefit the

reproducibility of the solutions and mitigate the possibility of cheating by using manually annotated competition test data. The best solution to prevent "data peeking" would be to keep the evaluation set, including unseen test conditions, inaccessible during algorithm development phase and to conduct independent (third-party) evaluations, in which the organizers run the provided executables or source codes of the submitted systems on the competition data. Another option would be to hide some "anchor" samples from the development set (with randomized file names) in the evaluation data and releasing the augmented test set once the development set scores have been submitted (fixed), as done in the BTAS 2016 Speaker Anti-spoofing Competition [39]. The scores of the anchor videos could be used for checking whether the scores for the development and test sets have been generated by the same system. At minimum, the organizers should be able to retrain and rerun the best-performing models following the official competition protocols to check if the submitted solutions have been calibrated, or even trained, on the test set, and determine the final ranking of the teams based on the verified results.

The ICCV2021 challenge [46] provided also a fairer evaluation of the proposed algorithms per se by limiting the influence caused by differences in the amount of training data and number of ensembled (deep) models. The use of the same training data and a fixed number of models would be fairer for comparing different algorithms. Otherwise, the competitions just assess and ascertain how far the participants can push the face PAD performance with "black-box" methods on the specific benchmark, while not gaining actual insight in the effectiveness of the different proposed algorithms under the same conditions. For instance, we observed that the performance gains with the best solutions of the CVPR2019 [45] and ECCV2020 [94] challenges were largely due to the use of large-scale pretraining data and huge ensemble models, respectively. Another option would be to conduct separate ablation studies on the competition test data by evaluating the solutions trained on the same training data in order to find out at least the impact of data, especially in the case of proprietary datasets.

The CASIA-SURF [90, 91] and the CASIA-SURF CeFA datasets [44] used in the CVPR2019 [45] and CVPR2020 [43] challenges, respectively, provide pre-cropped and aligned facial images for each modality. Furthermore, the background information has been pixel-wise masked out to mitigate the effect of different face detection and alignment methods and limit the problem of face PAD to the actual facial information instead of exploiting the domain-specific contextual cues. The findings of the ICCV2021 challenge [46] suggest, however, that the use of proper pre-processing (*e.g.*, face or facial attribute localization) can, in fact, significantly improve the PAD performance. The best-performing teams used additional pre-processing steps to focus on the actual discriminative facial details and specific subregion information and to avoid learning and extraction of irrelevant features for face PAD. Data pre-processing needs definitely further attention in future because it is an understudied subject in face PAD research and an important component in

complete face PAD solutions used in real-world AFR applications. While artificially restricting the original facial images and videos into pre-cropped bounding boxes or pixel-wise masked regions mitigates the issues related to sharing and working with huge datasets, and exploiting dataset-specific contextual cues for better PAD performance, the heavily pre-processed data can also limit the novelty of proposed solutions and usability of the dataset for experimental analysis during the competitions and, more importantly, in future studies in the research field. Therefore, the research community needs to find means for providing large-scale face PAD datasets also with unprocessed facial images and videos in order to be able to evaluate complete face PAD solutions and conduct comprehensive ablation studies.

Table 22 ACER (%) performance comparison of the single model architectures used in the CVPR2019 [45], CVPR2020 [43], IJCB2021 LivDet-Face [58], and ICCV2021 competitions [46].

Model	CVPR2019 challenge multi-modal	CVPR2020 challenge multi-modal	CVPR2020 challenge unimodal	IJCB2021 LivDet-Face unimodal	ICCV2021 challenge unimodal
VGG16 [65]	0.6255	-	-	-	-
ResNet18 [26]	2.4223	-	26.12	-	-
ResNet34 [26]	1.1107	1.68	-	-	5.562
ResNet101 [26]	-	-	9.28	-	-
ResNeXt101 [78]	0.6812	-	-	-	-
SE-ResNeXt101 [29]	0.0985	-	-	-	4.947
DenseNet121 [30]	**0.0516**	-	-	-	-
Inception [68]	1.6873	-	-	-	-
ShuffleNet-V2 [51]	0.3855	-	-	-	-
EfficientNet-B0 [70]	-	-	-	-	**3.053**
DepthNet [49]	-	5.89	-	-	6.31
CDCN [87]	-	**1.02**	**4.84**	**18.71**	5.054
CDCN++ [87]	-	-	-	-	**3.215**

3.3 Summary on Model Architectures

Model architectures play a vital role in extracting PAD-specific feature representations, thus it is worth investigating how to select suitable deep models for unimodal and multi-modal face PAD in light of the competition results. Table 22 shows the performance of the methods based on a single model architecture (*i.e.*, without ensemble models) in the CVPR2019 [45], CVPR2020 [43], IJCB2021 LivDet-Face [58], and ICCV2021 competitions [46]. In the multi-modal challenges (see the second and third columns in Table 22), the meth-

ods based on the DenseNet121 and CDCN backbone models achieved the best results. Compared with the ResNet family (*e.g.*, ResNet18, ResNet34, and ResNeXt101), the DenseNet121 extracts denser contextual semantic features, which benefits from the hidden feature representations between the different modalities. The generic backbones supervised by BCE loss focus on high-level semantic feature representations, while the CDCN with central difference convolution and multi-level fusion module is based on pixel-wise supervision, aiming at capturing more fine-grained disparities in the intrinsic fidelity characteristics between bona fide samples and face artefacts. In the unimodal challenges, the EfficientNet-B0 and CDCN performed the best in the IJCB2021 LivDet-Face and ICCV2021 challenges, respectively. It is interesting to notice that both of these two architectures have been discovered using automatic neural architecture search (NAS) [16], indicating the promising potential of NAS in searching optimal task-aware architectures also for face PAD. Despite the differences in hyperparameter settings, the architectures highlighted in bold in Table 22 can be recommended and treated as valuable prior knowledge in selecting models for the upcoming unimodal and multi-modal face PAD challenges.

In the ECCV2020 and IJCB2021 LivDet-Face competitions, most of the top teams considered ensemble models combining different architectures. For instance, the best-performing solution in the ECCV2020 challenge (see Table 13) consists of five kinds of models (*i.e.*, FOCUS, AENet, ResNet, Attack type, Noise Print) while 12 different models (*e.g.*, DeepPixBis, ResNeXt, etc.) are combined in the winning solution of the IJCB2021 LivDet-Face competition (see Table 15). It can be concluded that ensemble strategy works the best in these kinds of competition settings, but the contribution of each model on the final performance remains unclear due to lack of proper ablation studies. However, one common feature among these ensemble solutions combining mixed architectures is that both generic, high-level backbones (*e.g.*, ResNet [26] and ResNeXt [78]) with BCE loss and PAD-specific architectures (*e.g.*, DeepPixBis [21] and AENet [93]) with pixel-wise supervision are used. Thus, the feature representations from these two different approaches seem to compatible and complementary towards more generalized face PAD solutions.

3.4 Future Challenges

The test cases in the current competitions measuring the generalization across different covariates are still rather limited. Especially, the domain diversity should be increased, as the samples in most (4 out of 5) of the competitions were recorded in indoor office locations with no more than three ethnicities. Regarding the PA species, recently introduced challenging partial face attacks (*e.g.*, half 3D mask, makeup, and tattoo) have not been yet considered

in face PAD contests. Another one issue is the imbalanced long-tailed data distribution across different PAs, where some more challenging attack types (*e.g.*, high-fidelity masks) have usually been represented only with a few samples due to their high manufacturing costs. A large-scale test set 'in the wild' with diverse domain conditions and PA types, as well as more balanced data distribution, will be eventually needed to achieve more realistic evaluation settings.

Most of the existing face PAD competitions have not considered the efficiency or costs of the proposed solutions, as no constraints on either model size and number of models have been given. As a result, the best-performing algorithms have been usually ensembles of several deep models, which gives insight how robust PAD performance can be reached with the current state-of-the-art methods on the competition data. However, blindly pushing to maximum detection performance without any restrictions encourages the participants to exploit and combine off-the-shelf components instead of trying to invent truly novel and effective solutions that could be also deployed in real-world applications on mobile and embedded platforms with restricted resources. Although the organizers of the latest ICCV2021 challenge [46] explicitly informed that results obtained with fusion of deep models are rejected and the computational cost of a single model should be less than 100G FLOPs, there are still no evaluation metrics for measuring the trade-off between accuracy and efficiency.

Despite two multi-modal face PAD competitions have been already conducted, it is a bit disappointing that only few solutions pursued to introduce new kinds of advanced multi-modal fusion algorithms. Many major manufacturers have already included multi-modal camera systems into their products, including mobile devices and laptops, thus there is an urgent need to explore novel multi-modal fusion algorithms instead of just ensembling models with different modalities. Furthermore, not all the modalities are always available, as the selection of multi-modal data sources depends on the deployment scenario in question. Thus, it would be useful to investigate performance in settings where a model is trained on multiple modalities but evaluated on partial or arbitrary combinations of the modalities. Among emerging imaging technologies, depth and NIR cameras have been already considered in the two multi-modal competitions, thus it would be interesting to include also more advanced sensors, such as short-wave infrared (SWIR) [27]) or even hyperspectral imaging [34], in the upcoming collective evaluations.

Apart from conventional PAs, two kinds of physical adversarial attacks (*i.e.*, recognition and PA aware) could be considered for generic face PAD. For example, special printed eyeglasses [63], hats [37] and stickers [23] synthesized by adversarial generators have been demonstrated to effectively circumvent deep learning based AFR systems when worn by an attacker. Besides recognition-aware adversarial attacks, adversarial print and replay attacks [88] with specific perturbation injected before physical broadcast have been developed to fool face PAD systems. Therefore, it can be expected to

be necessary to consider a diverse set of physical adversarial attacks in future competitions. In addition to PAs, there are many vicious digital manipulation attacks (*e.g.*, "deepfakes" [61]) that can be applied against AFR systems. Despite the differences in generation techniques and visual quality, some of these attacks still have coherent properties and artefacts. In [12], a unified digital and physical face attack detection framework is proposed to learn joint representations for coherent attacks. Therefore, another interesting challenge to tackle in upcoming contests assessing the robustness of face biometric systems would be to simultaneously detect both digital and physical attacks.

4 Conclusions

Competitions play a vital role in consolidating the recent trends and assessing the state of the art in face PAD. This chapter introduced the design and results of the five latest international competitions on unimodal and multi-modal face PAD organized from 2019 until 2021. These contests have been important milestones in advancing the research on face PAD to the next level, as each competition has offered new challenges to the research community and resulted in novel countermeasures and new insight. The industrial participants have dominated most of these competitions, which indicates the strong need for robust anti-spoofing solutions in real-world applications. The first two face PAD competitions (*i.e.*, CVPR2019 [45] and CVPR2020 [43]) provided initial assessments of the state of the art in multi-modal face PAD algorithms under mainstream PAs in diverse conditions, while the three most recent face PAD competitions (*i.e.*, ECCV2020 [94], IJCB2021 LivDet-Face [58] and ICCV2021 [46]) benchmarked conventional colour (RGB) camera based PAD algorithms on larger scale datasets, challenging unseen attacks, and high-quality 3D mask attacks, respectively. Although several solutions proposed in the first two multi-modal competitions achieved satisfying PAD performance, more comprehensive multi-modal datasets and evaluation protocols on generalized PAD are still needed, especially considering the situation in which one or more modalities are missing in the test phase. In contrast, none of the systems proposed in the context of the latest two unimodal competitions managed to achieve satisfying PAD performance under unseen attacks detection in unknown operating conditions. Thus, more diverse, larger scale unimodal face PAD datasets are still needed to develop and evaluate more robust learning based algorithms.

Acknowledgements This work was supported by the Academy of Finland for Academy Professor project EmotionAI, ICT 2023 project and Infotech Oulu.

Index

A

Attack Presentation Classification Error Rate (APCER) 8
Average Classification Error Rate (ACER) 8

B

Bona Fide Presentation Classification Error Rate (BPCER) 8

C

Competition 3

D

Domain shift 30

E

Ensemble model 12, 42

F

False Positive Rate (FPR) 8

M

Multi-modal face PAD 2

N

Neural Architecture Search (NAS) 42

O

Operating point 39

P

Pixel-wise supervision 42
Pre-processing 35, 40
Presentation Attack Detection (PAD) 2
Presentation Attack Instrument (PAI) 2

R

Receiver Operating Characteristic (ROC) 8

T

Third-party evaluation 40
True Positive Rate (TPR) 8

U

Unimodal face PAD 2

Glossary

ACER: Average Classification Error Rate, 8
AFR: Automatic face recognition, 2
APCER: Attack Presentation Classification Error Rate, 8

BCE: Binary Cross-Entropy, 17
BPCER: Bona Fide Presentation Classification Error Rate, 8

CDC: Central Difference Convolution, 13
CDL: Contrastive Depth Loss, 13
CE: Cross-Entropy, 17
CNN: Convolutional Neural Network, 16

DSLR: Digital Single-Lens Reflex, 26

EER: Equal Error Rate, 16

FAS: Face Anti-Spoofing, 2
FDR: Fisher Discriminant Ratio, 28
FN: False Negative, 8
FP: False Positive, 8
FPR: False Positive Rate, 8

NAS: Neural Architecture Search, 42
NIR: Near-Infrared, 2

PA: Presentation Attack, 2
PAD: Presentation Attack Detection, 2
PAI: Presentation Attack Instrument, 2

ROC: Receiver Operating Characteristic, 8
rPPG: Remote Photoplethysmography, 31

SE: Squeeze-and-Excitation, 9
SWIR: Short-Wave Infrared, 43

TN: True Negative, 8
TP: True Positive, 8
TPR: True Positive Rate, 8

References

1. Boulkenafet, Z., Komulainen, J., Akhtar, Z., Benlamoudi, A., Samai, D., Bekhouche, S.E., Ouafi, A., Dornaika, F., Taleb-Ahmed, A., Qin, L., et al.: A competition on generalized software-based face presentation attack detection in mobile scenarios. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 688–696 (2017) 3, 4, 5, 8, 15, 38
2. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. IEEE Transactions on Information Forensics and Security **11**(8), 1818–1830 (2016) 8
3. Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., Hadid, A.: OULU-NPU: A mobile face presentation attack database with real-world variations. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 612–618 (2017) 2, 15, 28
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: VGGFace2: A dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 67–74 (2018) 9
5. Chakka, M.M., Anjos, A., Marcel, S., Tronci, R., Muntoni, D., Fadda, G., Pili, M., Sirena, N., Murgia, G., Ristori, M., et al.: Competition on counter measures to 2-d facial spoofing attacks. In: International Joint Conference on Biometrics (IJCB) (2011) 3, 4, 5, 8, 38
6. Chen, S., Yao, T., Zhang, K., Chen, Y., Sun, K., Ding, S., Li, J., Huang, F., Ji, R.: A dual-stream framework for 3d mask face presentation attack detection. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 834–841 (2021) 35
7. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: International Conference of the Biometrics Special Interest Group (BIOSIG) (2012) 2
8. Chingovska, I., Yang, J., Lei, Z., Yi, D., Li, S.Z., Kahm, O., Glaser, C., Damer, N., Kuijper, A., Nouak, A., et al.: The 2nd competition on counter measures to 2d face spoofing attacks. In: International Conference on Biometrics (ICB) (2013) 3, 4, 5, 8, 38
9. Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., Marcel, S.: The replay-mobile face presentation-attack database. In: International Conference of the Biometrics Special Interest Group (BIOSIG) (2016) 28
10. Costa-Pazo, A., Jiménez-Cabello, D., Vazquez-Fernandez, E., Alba-Castro, J.L., López-Sastre, R.J.: Generalized presentation attack detection: a face anti-spoofing evaluation proposal. In: International Conference on Biometrics (ICB) (2019) 9
11. Damer, N., Opel, A., Nouak, A.: Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution. In: European Signal Processing Conference (EUSIPCO), pp. 1382–1386 (2014) 28
12. Deb, D., Liu, X., Jain, A.K.: Unified detection of digital and physical face attacks. arXiv preprint arXiv:2104.02156 (2021) 44
13. Deng, D., Chen, Z., Zhou, Y., Shi, B.: MIMAMO Net: Integrating micro-and macro-motion for video emotion recognition. In: AAAI Conference on Artificial Intelligence (AAAI), vol. 34, pp. 2621–2628 (2020) 17
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009) 4, 9
15. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: RetinaFace: Single-shot multi-level face localisation in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5203–5212 (2020) 35
16. Elsken, T., Metzen, J.H., Hutter, F.: Neural architecture search: A survey. Journal of Machine Learning Research **20**(1), 1997–2017 (2019) 42

17. Erdogmus, N., Marcel, S.: Spoofing face recognition with 3d masks. IEEE Transactions on Information Forensics and Security **9**(7), 1084–1097 (2014) 28
18. Fang, M., Damer, N., Kirchbuchner, F., Kuijper, A.: Real masks and fake faces: On the masked face presentation attack detection. arXiv preprint arXiv:2103.01546 (2021) 28
19. Feng, H., Hong, Z., Yue, H., Chen, Y., Wang, K., Han, J., Liu, J., Ding, E.: Learning generalized spoof cues for face anti-spoofing. arXiv preprint arXiv:2005.03922 (2020) 24
20. Fernando, B., Gavves, E., Oramas, J., Ghodrati, A., Tuytelaars, T.: Rank pooling for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 773–787 (2016) 16, 17
21. George, A., Marcel, S.: Deep pixel-wise binary supervision for face presentation attack detection. In: International Conference on Biometrics (ICB), pp. 1–8 (2019) 28, 42
22. Grinchuk, O., Parkin, A., Glazistova, E.: 3d mask presentation attack detection via high resolution face parts. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 846–853 (2021) 35
23. Guo, Y., Wei, X., Wang, G., Zhang, B.: Meaningful adversarial stickers for face recognition in physical world. arXiv preprint arXiv:2104.06728 (2021) 43
24. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: Challenge of recognizing one million celebrities in the real world. Electronic Imaging **2016**(11) (2016) 9
25. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d CNNs retrace the history of 2d CNNs and ImageNet? In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6546–6555 (2018) 17
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016) 9, 17, 24, 28, 41, 42
27. Heusch, G., George, A., Geissbühler, D., Mostaani, Z., Marcel, S.: Deep models and shortwave infrared information to detect face presentation attacks. IEEE Transactions on Biometrics, Behavior, and Identity Science **2**(4), 399–409 (2020) 43
28. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial intelligence **17**(1-3), 185–203 (1981) 16, 17
29. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018) 24, 41
30. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708 (2017) 41
31. Huang, S., Cheng, W.H., Cheng, R.: Single patch based 3d high-fidelity mask face anti-spoofing. In: IEEE/CVF International Conference on Computer Vision (ICCV), pp. 842–845 (2021) 35
32. ISO: ISO/IEC JTC 1/SC 37 Biometrics: Information technology biometric presentation attack detection part 1: Framework. In: https://www.iso.org/obp/ui/iso (2016) 2, 8, 16, 27, 34, 39
33. Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: Anti-spoofing via noise modeling. In: European Conference on Computer Vision (ECCV), pp. 290–306 (2018) 24
34. Kaichi, T., Ozasa, Y.: A hyperspectral approach for unsupervised spoof detection with intra-sample distribution. In: IEEE International Conference on Image Processing (ICIP), pp. 839–843 (2021) 43
35. Kantarcı, A., Dertli, H., Ekenel, H.K.: Shuffled patch-wise supervision for presentation attack detection. In: International Conference of the Biometrics Special Interest Group (BIOSIG) (2021) 28
36. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE International Conference on Neural Networks (ICNN), vol. 4, pp. 1942–1948 (1995) 24

37. Komkov, S., Petiushko, A.: AdvHat: Real-world adversarial attack on arcface face id system. In: IEEE International Conference on Pattern Recognition (ICPR), pp. 819–826 (2021) 43
38. Komulainen, J., Boulkenafet, Z., Akhtar, Z.: Review of face presentation attack detection competitions. In: S. Marcel, M.S. Nixon, J. Fierrez, N. Evans (eds.) Handbook of Biometric Anti-Spoofing: Presentation Attack Detection, pp. 291–317. Springer (2019) 3, 4
39. Korshunov, P., Marcel, S., Muckenhirn, H., Gonçalves, A.R., Mello, A.G.S., Violato, R.P.V., Simoes, F.O., Neto, M.U., de Assis Angeloni, M., Stuchi, J.A., Dinkel, H., Chen, N., Qian, Y., Paul, D., Saha, G., Sahidullah, M.: Overview of BTAS 2016 speaker anti-spoofing competition. In: IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS) (2016) 40
40. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Neural Information Processing Systems (NIPS) (2012) 28
41. Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., Wang, C., Li, J., Huang, F.: DSFD: dual shot face detector. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5060–5069 (2019) 34, 35
42. Li, X., Komulainen, J., Zhao, G., Yuen, P.C., Pietikäinen, M.: Generalized face anti-spoofing by detecting pulse from face videos. In: IEEE International Conference on Pattern Recognition (ICPR), pp. 4244–4249 (2016) 31
43. Liu, A., Li, X., Wan, J., Liang, Y., Escalera, S., Escalante, H.J., Madadi, M., Jin, Y., Wu, Z., Yu, X., Tan, Z., Yuan, Q., Yang, R., Zhou, B., Guo, G., Li, S.Z.: Cross-ethnicity face anti-spoofing recognition challenge: A review. IET Biometrics **10**(1), 24–43 (2021) 3, 12, 13, 21, 23, 30, 31, 34, 37, 38, 39, 40, 41, 44
44. Liu, A., Tan, Z., Wan, J., Escalera, S., Guo, G., Li, S.Z.: CASIA-SURF CeFA: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1179–1187 (2021) 12, 13, 14, 15, 17, 18, 19, 20, 24, 40
45. Liu, A., Wan, J., Escalera, S., Escalante, H.J., Tan, Z., Yuan, Q., Wang, K., Lin, C., Guo, G., Guyon, I., Li, S.Z.: Multi-modal face anti-spoofing attack detection challenge at CVPR2019. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1601–1610 (2019) 3, 5, 9, 10, 11, 13, 14, 16, 21, 23, 24, 30, 37, 38, 39, 40, 41, 44
46. Liu, A., Zhao, C., Yu, Z., Su, A., Liu, X., Kong, Z., Wan, J., Escalera, S., Escalante, H.J., Lei, Z., et al.: 3d high-fidelity mask face presentation attack detection challenge. In: IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 814–823 (2021) 3, 4, 31, 34, 36, 37, 38, 39, 40, 41, 43, 44
47. Liu, A., Zhao, C., Yu, Z., Wan, J., Su, A., Liu, X., Tan, Z., Escalera, S., Xing, J., Liang, Y., et al.: Contrastive context-aware learning for 3d high-fidelity mask face presentation attack detection. arXiv preprint arXiv:2104.06148 (2021) 31, 33
48. Liu, S.Q., Lan, X., Yuen, P.C.: Multi-channel remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. IEEE Transactions on Information Forensics and Security (2021) 36
49. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 389–398 (2018) 2, 19, 28, 41
50. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: IEEE International Conference on Computer Vision, pp. 3730–3738 (2015) 22
51. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: ShuffleNet v2: Practical guidelines for efficient CNN architecture design. In: European Conference on Computer Vision (ECCV), pp. 116–131 (2018) 9, 41
52. Ming, Z., Visani, M., Luqman, M.M., Burie, J.C.: A survey on anti-spoofing methods for facial recognition with RGB cameras of generic consumer devices. Journal of Imaging **6**(12), 139 (2020) 2, 24

53. Mohammadi, A., Bhattacharjee, S., Marcel, S.: Deeply vulnerable: a study of the robustness of face recognition to presentation attacks. IET Biometrics (2017) 2
54. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output CNN for age estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4920–4928 (2016) 9
55. Pan, G., Sun, L., Wu, Z., Lao, S.: Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In: International Conference on Computer Vision (ICCV) (2007) 9
56. Parkin, A., Grinchuk, O.: Recognizing multi-modal face spoofing with face recognition networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019) 6, 9
57. Parkin, A., Grinchuk, O.: Creating artificial modalities to solve RGB liveness. arXiv preprint arXiv:2006.16028 (2020) 17
58. Purnapatra, S., Smalt, N., Bahmani, K., Das, P., Yambay, D., Mohammadi, A., George, A., Bourlai, T., Marcel, S., Schuckers, S., et al.: Face liveness detection competition (LivDet-Face)-2021. In: IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10 (2021) 3, 4, 25, 27, 31, 34, 37, 39, 41, 44
59. Ramachandra, R., Stokkenes, M., Mohammadi, A., Venkatesh, S., Raja, K., Wasnik, P., Poiret, E., Marcel, S., Busch, C.: Smartphone multi-modal biometric authentication: Database and evaluation. arXiv preprint arXiv:1912.02487 (2019) 28
60. Ramachandra, R., Venkatesh, S., Raja, K.B., Bhattacharjee, S., Wasnik, P., Marcel, S., Busch, C.: Custom silicone face masks: Vulnerability of commercial face recognition systems & presentation attack detection. In: IEEE International Workshop on Biometrics and Forensics (IWBF) (2019) 28
61. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to detect manipulated facial images. In: IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 44
62. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520 (2018) 9
63. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: A general framework for adversarial examples with objectives. ACM Transactions on Privacy and Security (TOPS) **22**(3), 1–30 (2019) 43
64. Shen, T., Huang, Y., Tong, Z.: FaceBagNet: Bag-of-local-features model for multi-modal face anti-spoofing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019) 6, 9
65. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 9, 28, 41
66. Sun, S., Pang, J., Shi, J., Yi, S., Ouyang, W.: FishNet: A versatile backbone for image, region, and pixel level prediction. In: Advances in Neural Information Processing Systems, pp. 762–772 (2018) 9
67. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-ResNet and the impact of residual connections on learning. In: AAAI Conference on Artificial Intelligence (AAAI) (2017) 9
68. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 41
69. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016) 28
70. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning (ICML), pp. 6105–6114 (2019) 24, 41

71. Wan, J., Guo, G., Escalera, S., Escalante, H.J., Li, S.Z.: Multi-modal face presentation attack detection. Synthesis Lectures on Computer Vision **9**(1), 1–88 (2020) 2
72. Wang, G., Lan, C., Han, H., Shan, S., Chen, X.: Multi-modal face presentation attack detection via spatial and channel attentions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019) 6, 9
73. Wang, Z., Yu, Z., Zhao, C., Zhu, X., Qin, Y., Zhou, Q., Zhou, F., Lei, Z.: Deep spatial gradient and temporal depth learning for face anti-spoofing. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 13
74. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security **10**(4), 746–761 (2015) 2
75. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European Conference on Computer Vision (ECCV), pp. 499–515. Springer (2016) 9
76. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: CBAM: Convolutional block attention module. In: European Conference on Computer Vision (ECCV), pp. 3–19 (2018) 9
77. Wu, X., He, R., Sun, Z., Tan, T.: A light CNN for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security **13**(11), 2884–2896 (2018) 9
78. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern recognition (CVPR), pp. 1492–1500 (2017) 9, 24, 28, 41, 42
79. Yang, Q., Zhu, X., Fwu, J.K., Ye, Y., You, G., Zhu, Y.: PipeNet: Selective modal pipeline of fusion network for multi-modal face anti-spoofing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 644–645 (2020) 19
80. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv preprint arXiv:1411.7923 (2014) 4, 9
81. Yu, Z., Li, X., Niu, X., Shi, J., Zhao, G.: Face anti-spoofing with human material perception. In: European Conference on Computer Vision (ECCV), pp. 557–575 (2020) 34
82. Yu, Z., Li, X., Wang, P., Zhao, G.: TransRPPG: Remote photoplethysmography transformer for 3d mask face presentation attack detection. IEEE Signal Processing Letters (2021) 36
83. Yu, Z., Qin, Y., Li, X., Wang, Z., Zhao, C., Lei, Z., Zhao, G.: Multi-modal face anti-spoofing based on central difference networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 650–651 (2020) 19
84. Yu, Z., Qin, Y., Li, X., Zhao, C., Lei, Z., Zhao, G.: Deep learning for face anti-spoofing: A survey. arXiv preprint arXiv:2106.14948 (2021) 2, 24
85. Yu, Z., Qin, Y., Zhao, H., Li, X., Zhao, G.: Dual-cross central difference network for face anti-spoofing. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 1281–1287 (2021) 35
86. Yu, Z., Wan, J., Qin, Y., Li, X., Li, S.Z., Zhao, G.: NAS-FAS: Static-dynamic central difference network search for face anti-spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(9), 3005–3023 (2020) 13, 28
87. Yu, Z., Zhao, C., Wang, Z., Qin, Y., Su, Z., Li, X., Zhou, F., Zhao, G.: Searching central difference convolutional networks for face anti-spoofing. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 13, 17, 24, 41
88. Zhang, B., Tondi, B., Barni, M.: Attacking CNN-based anti-spoofing face authentication in the physical domain. arXiv preprint arXiv:1910.00327 (2019) 43
89. Zhang, P., Zou, F., Wu, Z., Dai, N., Mark, S., Fu, M., Zhao, J., Li, K.: FeatherNets: convolutional neural networks as light as feather for face anti-spoofing. In:

IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2019) 6, 9
90. Zhang, S., Liu, A., Wan, J., Liang, Y., Guo, G., Escalera, S., Escalante, H.J., Li, S.Z.: 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) 5, 6, 9, 10, 40
91. Zhang, S., Wang, X., Liu, A., Zhao, C., Wan, J., Escalera, S., Shi, H., Wang, Z., Li, S.Z.: A dataset and benchmark for large-scale multi-modal face anti-spoofing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 919–928 (2019) 5, 6, 9, 10, 38, 40
92. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S.Z.: Faceboxes: A CPU real-time face detector with high accuracy. In: International Joint Conference on Biometrics (IJCB) (2017) 32, 35
93. Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., Liu, Z.: Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In: European Conference on Computer Vision (ECCV), pp. 70–85. Springer (2020) 21, 22, 24, 38, 42
94. Zhang, Y., Yin, Z., Shao, J., Liu, Z., Yang, S., Xiong, Y., Xia, W., Xu, Y., Luo, M., Liu, J., et al.: Celeba-spoof challenge 2020 on face anti-spoofing: Methods and results. arXiv preprint arXiv:2102.12642 (2021) 3, 21, 23, 24, 30, 38, 39, 40, 44
95. Zhang, Z., Yan, J., Liu, S., Lei, Z., Yi, D., Li, S.Z.: A face antispoofing database with diverse attacks. In: International Conference on Biometrics (ICB), pp. 26–31 (2012) 2, 28
96. Zhao, J., Cheng, Y., Xu, Y., Xiong, L., Li, J., Zhao, F., Jayashree, K., Pranata, S., Shen, S., Xing, J., et al.: Towards pose invariant face recognition in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2207–2216 (2018) 9
97. Zhu, Y., Newsam, S.: Densenet for dense flow. In: IEEE International Conference on Image Processing (ICIP), pp. 790–794 (2017) 9