Multi-modal Face Anti-spoofing Attack Detection Challenge at CVPR2019

Ajian Liu[1], Jun Wan[2]*, Sergio Escalera[3], Hugo Jair Escalante[4], Zichang Tan[2]
Qi Yuan[5], Kai Wang[6], Chi Lin[7], Guodong Guo[8], Isabelle Guyon[9], Stan Z. Li[1,2]
[1]M.U.S.T, Macau; [2]NLPR, CASIA, UCAS, China; [3]CVC, UB, Spain
[4]INAOE, CINVESTAV, Mexico & ChaLearn, USA; [5]Beihang University, China; [6]UESTC, China
[7]USC, CA, USA; [8]Baidu Research, China; [9]U. Paris-Saclay, France & ChaLearn, USA
{ajianliu92,hugo.jair}@gmail.com, {jun.wan,szli}@nlpr.ia.ac.cn

Abstract

Anti-spoofing attack detection is critical to guarantee the security of face-based authentication and facial analysis systems. Recently, a multi-modal face anti-spoofing dataset, CASIA-SURF, has been released with the goal of boosting research in this important topic. CASIA-SURF is the largest public data set for facial anti-spoofing attack detection in terms of both, diversity and modalities: it comprises $1,000$ *subjects and* $21,000$ *video samples. We organized a challenge around this novel resource to boost research in the subject. The Chalearn LAP multi-modal face anti-spoofing attack detection challenge attracted more than 300 teams for the development phase with a total of 13 teams qualifying for the final round. This paper presents an overview of the challenge, including its design, evaluation protocol and a summary of results. We analyze the top ranked solutions and draw conclusions derived from the competition. In addition we outline future work directions.*

1. Introduction

As an important branch of biometric recognition, face recognition (FR) is being increasingly used in our daily life for tasks such as phone unlocking, access authentication and control, and face-payment [5, 34]. Because of its wide applicability and usage, FR systems can be an attractive target for identity attacks. For instance, unauthorized people trying to get authenticated via face presentation attacks (PAs), such as a printed face photograph (print attack), displaying videos on digital devices (replay attack), or 3D masks attack. These PAs make face recognition systems vulnerable, even if they achieve near-perfect recognition performance [1]. Therefore, face presentation attack detection (PAD), commonly called face anti-spoofing, is a critical step to ensure that FR systems are safe against face attacks.

*Corresponding author

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

Table 1. Team and affiliations name are listed in the final ranking of this challenge.

State-of-the-art face PAD algorithms [17, 15] have achieved high recognition rates in the intra-testing (*i.e.*, training and testing with the same dataset). However, they generally show low performance when a cross-testing (*i.e.*, training and testing data come from different datasets) scenario is considered. Therefore, face PAD remains a challenging problem, mainly due to lack of generalization capabilities of existing methods. This is largerly due to the fact that current face anti-spoofing databases have not enough subjects (≤ 170), or lack from fruitful samples ($\leq 6,000$ video clips) [25] compared with image classification [8] or face recognition databases [34], which severely limits the type of methods that can be used to approach the PAD problem (*e.g.* deep learning models). Another missing feature in existing datasets (e.g., [9, 5]) is the availability of multi-modal information. This sort of extended information may be very helpful for developing more robust anti-spoofing methods. The above mentioned problems seriously hinder novel technology developments in the field.

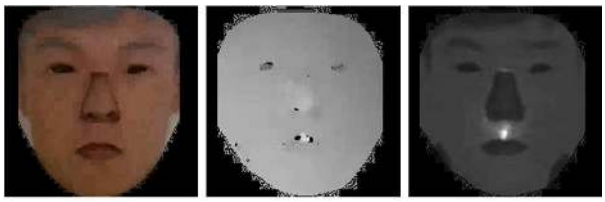

Figure 1. A processed Attack 5 sample, more shown in [25]

In order to deal with previous drawbacks, a large-scale multi-modal face anti-spoofing dataset, called CASIA-SURF [25], has been collected. The data set consists of $1,000$ different subjects and $21,000$ video clips with 3 modalities (RGB, Depth, IR). Based on this dataset, we organized the *Chalearn LAP multi-modal face anti-spoofing attack detection challenge* collocated with CVPR2019. The goal of this competition was to boost research progress on the PAD, in a scenario where plenty of data and different modalities are available. The challenge was run in the Codalab[1] platform. More than 300 academic research and industrial institutions worldwide participated in this challenge, and finally thirteen teams entered at the final stage. A summary with the names and affiliations of teams that entered the final stage are shown in Table 1. Interestingly, compared with the previous challenges [4, 6, 2], the majority of the final participants (ten out of thirteen) of this competition come from the industrial community, which indicates the increased importance of the topic for daily life applications.

To sum up, the contributions of this paper are summarized as follows: (1) We describe the design of the *Chalearn LAP multi-modal face anti-spoofing attack detection* challenge. (2) We organized this challenge around the CASIA-SURF datsaset, proving the suitability of such resource for boosting research in the topic. (3) We report and analyze the solutions developed by participants. (4) We point out critical points on the face anti-spoofing detection task by comparing essential differences between a real face and a fake one from multiple aspects, discussing future lines of research in the field.

2. Challenge Overview

In this section we review the organized challenge, including a brief introduction of the CASIA-SURF dataset, the evaluation metric, and the challenge protocol.

CASIA-SURF. The CASIA-SURF dataset is, to the best of our knowledge, the largest existing one in terms of subjects and videos [25]. Each sample of the dataset is associated to three modalities captured with an Intel RealSense SR300 camera. Each sample comprises 1 live video clip, and 6 fake video clips under different attacks (one attack way per fake video clip, shown in 1). A total of $1,000$ subjects and $21,000$ videos were captured for building this dataset.

We relied on this dataset for the organization of the *ChaLearn Face Anti-spoofing Attack Detection Challenge*. Accordingly, the CASIA-SURF data set was processed as follows. (1) The dataset was split in three partitions: training, validation and testing sets, with 300, 100 and 600 subjects, respectively. This partitioning corresponds to 6,300 (2,100 per modality), 2,100 (700 per modality), 12,600 (4,200 per modality) videos for the corresponding partitions. (2) For each video, we retained 1 out every 10 frames to reduce its size. This subsmapling strategy results in: 148K, 48K, 295K frames for training, validation and testing subsets, respectively. (3) The background except face areas from original videos was removed to increase the difficulty of the task.

Evaluation. In this challenge, we selected the recently standardized ISO/IEC 30107-3[2] metrics: Attack Presentation Classification Error Rate (APCER), Normal Presentation Classification Error Rate (NPCER) and Average Classification Error Rate (ACER) as the evaluation metrics, these are defined as follows:

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

$$NPCER = FN/\left(FN + TP\right) \tag{2}$$

$$ACER = \left(APCER + NPCER\right)/2 \tag{3}$$

where TP, FP, TN and FN corresponds to true positive, false positive, true negative and false negative, respectively. APCER and BPECER are used to measure the error rate of fake or live samples, respectively. Inspired by face recognition, the Receiver Operating Characteristic (ROC) curve is introduced for large-scale face Anti-spoofing detection in CASIA-SURF dataset, which can be used to select a suitable threshold to trade off the false positive rate (FPR) and true positive rate (TPR) according to the requirements of real applications. Finally, The value TPR@FPR=10^{-4} was the leading evaluation measure for this challenge. APCER, NPCER and ACER measures were used as additional evaluation criteria.

Challenge protocol. The challenge was run in the CodaLabplatform, and comprised two stages as follows:

Development Phase: (*Started: Dec. 22, 2018 - Ended: in March 6, 2019*). During this phase participants had access to labeled training data and unlabeled validation samples. Participants could use training data to develop their models, and they could submit predictions on the validation partition. Training data was made available with samples labeled with the genuine and 3 forms of attack (4,5,6).

[1] https://competitions.codalab.org/competitions/20853

[2] https://www.iso.org/obp/ui/iso

Whereas samples in the validation partition were associated to genuine and 3 different attacks (1,2,3). For the latter dataset, labels were not made available to participants. Instead, participants could submit predictions on the validation partition and receive immediate feedback via the leader board. The main reason for including different attack types in the training and validation dataset was to increase the difficulty of FAD challenge.

Final phase: (*Started: March 6, 2019 - Ended: March 10, 2019*). During this phase, labels for the validation subset were made available to participants, so that they can have more labeled data for training their models. The unlabeled testing set was also released, participants had to make predictions for the testing partition and upload their solutions to the challenge platform. The considered test set was formed by examples labeled with the genuine label and 3 attack types (1,2,3). Participants had the opportunity to make 3 submissions for the final phase, this was done with the goal of assessing stability of their methods. Note that the CodaLab platform defaults to the result of the last submission.

The final ranking of participants was obtained from the performance of submissions in the testing sets. To be eligible for prizes, winners had to publicly release their code under a license of their choice and provide a fact sheet describing their solution.

3. Description of solutions

The face anti-spoofing problem has been studied for decades. Some previous work [20, 29] attempted to detect evidence of liveness in samples (*i.e.*, eye-blinking). Other works were based on contextual information [21, 16] (*i.e.*, attack material and screen moir). As deep learning has proven to be very effective in many computer vision problems, CNN-based methods are also present now in the face PAD community [10, 17, 15]. They treat face PAD as a binary classification problem, achieving remarkable performance in intra-testing evaluation.

For the organized challenge, no team used traditional methods for FAD, such as detecting physiological signs of life, like eye blinking, facial expression changes and mouth movements. Instead, all submitted face PAD solutions relied on model-based feature extractors, such as ResNet [12], VGG16 [26], etc. In the rest of this section we describe the methods based on the ranking order (except Baseline and Vivi) on the testing data set developed by the participants that made it to the final stage; a summary is provided in Table 2.

Baseline. Before the challenge, we built a strong baseline for approaching the task, our goal was to have a method of competitive performance for this datast. A detailed description of the baseline in provided in [25]. In short, we considered the face anti-spoofing problem as a binary classification task (fake *v.s* real) and conducted the experiments based on the ResNet-18 [12] classification network. In order to make full use of the characteristics between different modalities, inspired by [13], we proposed the squeeze and excitation fusion method that uses the "Squeeze-and-Excitation" branch to enhance the representational ability of the different modalities' feature by explicitly modelling the interdependencies among different convolutional channels.

VisionLabs. This method used a modified network architecture as in [25]. As shown in Figure 2, the RGB, Depth and IR inputs were processed by separate streams followed by concatenation and fully-connected layers. Unlike [25], they used aggregation blocks (AGG res2, AGG res3, AGG res4) to aggregate outputs from multiple layers of the network. Then, they pre-train network weights on four different tasks for face and gender recognition, and fine-tune these networks separately on the training set of the CASIA-SURF. It is worth noting that they split the training set into three folds according to different attacks present in the training subset to increase robustness to unknown attacks. Finally, the outputs of three networks were combined by averaging to produce results on the final validation and test sets.

ReadSense. This team relied on local features. They used a shallow SEresnext [13] to classify the multi-modal face images based on image patches in variant scale. To further improve the performance, a multi-stream fusion network with three-modal images was utilized. The fusion network was trained from scratch with RGB, Depth and IR data at the same time. Moreover, data augmentation was applied and modalities were randomly dropped during training. For optimization, they followed a cyclic cosine annealing learning rate schedule[14] which yielded better performance.

Feather. The main idea of this team's solution was to process multi-modal images sequentially in a cascaded network, as shown in Figure 3. Participants considered that depth information plays a key role between live and spoof faces based on the fact that live faces have face-like depth, *e.g.*, the nose is closer to the camera than the cheek in frontal-view faces, while faces in print or replay attacks have flat or planar depth, *e.g.*, all pixels on the image of a paper have the same depth to the camera. Furthermore, IR data was adopted in at end of network which measures the amount of heat radiated from a face which can provides strong error correction for reducing FP (false positive) samples greatly. Therefore, the process is divided into the following two stages: **Stage 1:** Four ensemble networks with depth modal as input respectively and output the scores of classification by voting. **Stage 2:** A MobileLiteNet followed by stage 1 which takes the IR modal as input to judge the fake samples further. The basic networks of these two phases are Fishnet [27] and MobileNetv2 [24] respectively.

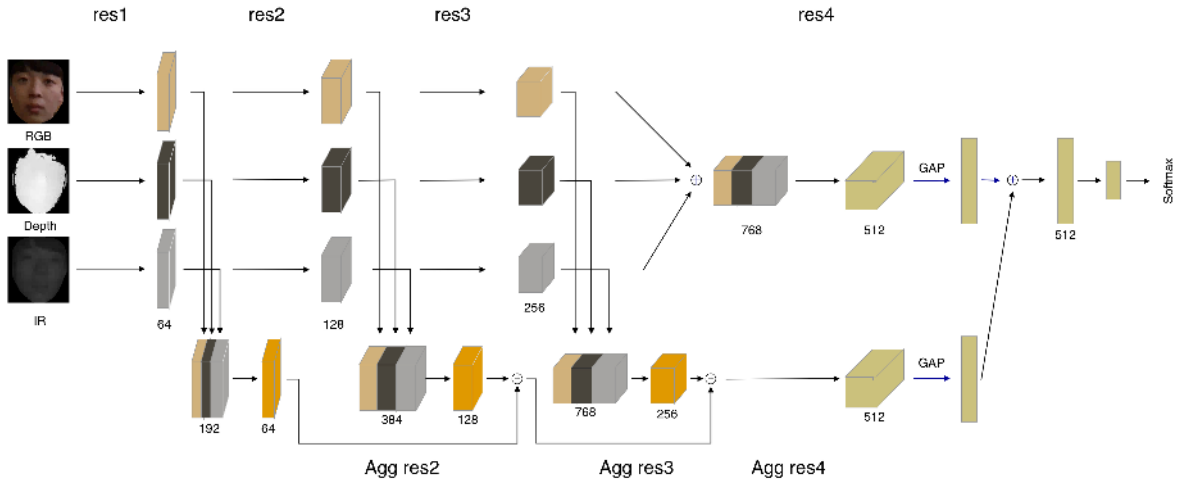

Figure 2. Provided by VisionLabs team. Deep layer aggregation architecture of VisionLabs.

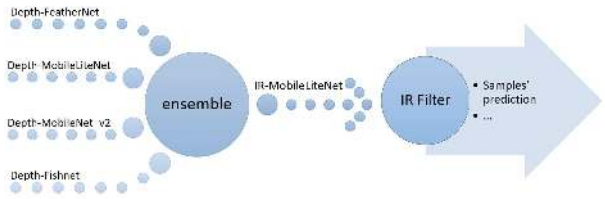

Figure 3. Provided by Feather team. The network structure of Feather team, in which the FeatherNet and MobileLiteNet are modified by Fishnet [27] and MobileNetv2 [24] respectively.

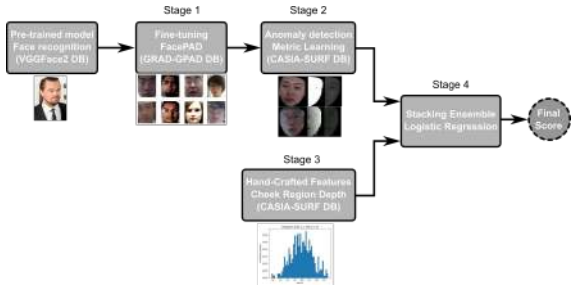

Figure 4. Provided by GradiantResearch team. General diagram of the GradiantResearch team.

Hahahaha. Their base model is a Resnext [33] which was pre-trained with the ImageNet dataset [8]. Then, they fine-tune the network on aligned images with face landmark and use data augmentation to strengthen the generalization ability.

MAC-adv-group. This solution used the Resnet-34 [12] as base network. To overcome the influence of illumination variation, they convert RGB image to HSV color space. Then, they sent the features extracted from the network into a fully-connected layer and a binary classification layer.

ZKBH. Analyzing the training, valid and test sets, participants assumed that the eye region is promising to get good performance in FAD task based on an observation that the eye region is the common attack area. After several trials, the input of the final version they submitted adopted quarter face containing the eye region. Different from prior works that regard the face anti-spoofing problem as merely a binary (fake *v.s* real) classification problem, this team constructed a regression model for differentiating the real face and the attacks.

VisionMiracle. This solution was based on the modified shufflenet-V2[18]. The feature-map was divided into two branches after the third stage, and connected in the fourth stage.

GradiantResearch. The fundamental idea behind this solution was the reformulation of the face presentation attack detection problem (face-PAD) following an anomaly detection strategy using deep metric learning. The approach can be split in four stages (Figure 4): **Stage 1:** using a pre-trained model for face recognition and apply a classification-like metric learning approach in GRAD-GPAD dataset [7] using only RGB images. **Stage 2:** they fine-tune the model obtained in Stage 1 with the CASIA-SURF dataset using metric learning for anomaly detection (semi-hard batch negative mining with triplet focal loss) adding Depth and IR images to the input volume. Once the model converged, they trained an SVM classifier using the features of the last fully connected layer (128D). **Stage 3:** they trained an SVM classifier using the normalized histogram of the depth image corresponding to the cheek region of the face (256D). **Stage 4:** they performed a simple stacking ensemble of both models (Stage 2 and Stage 3) by training a logistic regression model with the scores in the training split.

Team Name	Method	Model	Pre-trained data	Modality	Pre-process	Additional FAD dataset	Fusion and Loss function
VisionLabs	Fine-tuning Ensembling	Resnet-34 [12] Resnet-50 [12]	Casia-WebFace [34] AFAD-Lite [19] MSCeleb1M [11] Asian dataset [35]	RBG Depth IR	Resize	No	Squeeze and Excitation Fusion Score fusion SoftmaxWithLoss
ReadSense	Bag-of-local feature Ensembling	SEresnext [33]	No	RBG Depth IR	Crop image patches Image augmentation	No	Squeeze and Excitation Fusion Score fusion SoftmaxWithLoss
Feather	Ensembling	Fishnet [27] MobileNetv2 [24]	No	Depth IR	Resize Image adjust	Private FAD data	Score fusion SoftmaxWithLoss
Hahahaha	Only using depth images	Resnext [33]	Imagenet [8]	Depth	Data augmentation Aligned faces	No	SoftmaxWithLoss
MAC-adv-group	Features fusion	Resnet-34	No	RBG Depth IR	Transfer color space	No	Features fusion SoftmaxWithLoss
ZKBH	Using regression model	Resnet-18	No	RBG Depth IR	Crop image Image augmentation	No	Data fusion Regression loss
VisionMiracle	Modified shufflenet-V2	Shufflenet-V2 [18]	No	Depth	Image augmentation	No	SoftmaxWithLoss
Baseline [25]	Features fusion	Resnet-18	No	RBG Depth IR	Resize Image augmentation	No	SoftmaxWithLoss
GradiantResearch	Deep metric learning	Inception resnet v1 [28]	VGGFace2 [3] GRAD-GPAD [7]	RBG Depth IR	Crop image Image augmentation	No	Stacking ensemble Logistic regression
Vipl-bpoic	Attention mechanism [31]	ResNet-18	No	RBG Depth IR	Control positive and negative sample ratio	No	Data fusion Center loss [30] SoftmaxWithLoss
Massyhnu	Ensembling	9 Softmax classifiers	No	RBG Depth IR	Resize Transfer color space	No	Color information fusion SoftmaxWithLoss
AI4all	Only using depth images	Vgg16 [26]	No	Depth	Resize Image augmentation	No	SoftmaxWithLoss
Guillaume	Multi-Channel CNN	LightCNN [32]	Yes	Depth IR	Resize	No	Data fusion SoftmaxWithLoss
Vivi	A dense-cross-modality-attention model	Densenet [36]	Yes	RBG Depth IR	Image augmentation Transfer color space	Private FAD data	Features fusion Score fusion SoftmaxWithLoss

Table 2. Summary of the methods for all participating teams.

Vipl-bpoic. This team focused on improving face anti-spoofing generalization ability by proposing an end-to-end trainable face anti-spoofing model with attention mechanism. Due to the sample imbalance, they assign the weight of 1:3 according to the number of genuine and spoof faces in Training set. Subsequently, they fuse the three modal images including RGB, Depth and IR into 5 channels as the input of ResNet-18 [12] which integrated with the convolutional block atttention module. The center loss [30] and cross-entropy loss are adopted to constrain the learning process in order to get more discriminative cues of FAD finally.

Massyhnu. This team paid attention to color information

NN1	NN1a	NN2	NN3	NN4	Val TPR @FPR=10e-4	Test TPR @FPR=10e-4
✓					0.9943	-
	✓				0.9987	-
		✓			0.9870	-
			✓		0.9963	-
				✓	0.9933	-
✓		✓			0.9963	-
✓		✓	✓		0.9983	-
✓		✓	✓		0.9997	-
✓		✓	✓	✓	1.0000	-
	✓	✓	✓	✓	**1.0000**	**0.9988**

Table 4. Provided by VisionLabs team. The results on the valid and test sets of the VisionLabs team, different NN modules represent different pre-trained Resnet [12].

fusion and ensemble learning [23, 22].

AI4all. This team used VGG16 [26] as the backbone for face PAD.

Guillaume. Their method consists in a Multi-Channel convolutional Neural Network (MC-CNN) taking a face images of different modalities as input. Near-infrared and depth images only have been used in their approach. The architecture of the proposed MC-CNN is based on the second version of the LightCNN [32] containing 29 layers. Also, the pretrained LightCNN model is used as a starting point for their training procedure. The training consists in the fine-tuning of the low-level convolutional layers of the network in each modalities, and in learning the final fully connected layers.

Vivi. A dense-cross-modality-attention model was trained by using the Depth, RGB and IR dataset. In this network, a dense connected structure was used in every single modality and the cross-modality attention mechanism was designed to transfer information from different modalities. After the cross-modality backbone was designed, they used the paddle-auto-ml [3] tool to search for the hyperparameters of the network such as channel numbers and kernel sizes. In addition, they collected a large amount of data in three modalities same with CASIA-SURF.

4. Challenge Results

In this section, we present the results obtained by the thirteen teams that qualified to the final phase. Then, the effectiveness of proposed algorithms are analyzed. Finally, we point out some limitations of the algorithms proposed by participating teams.

4.1. Challenge Results Report

In order to evaluate the performance of solutions, we adopted the following metrics: APCER, NPCER, ACER and TPR in the case of FPR=10^{-2}, 10^{-3}, 10^{-4} respectively, and the scores retained 6 decimal places for all results. The scores and ROC curves of participating teams on the testing partitions are shown in Table 3 and Figure 5 respectively. Please note that although we report performance for a variety of evaluation measures, the leading metric was TPR@FPR=10^{-4}. It can be observed that the best result (VisionLabs) achieves TPR=99.9885%, 99.9541%, 99.8739% @FPR=10^{-2}, 10^{-3}, 10^{-4}, respectively, and the TP = 17430, FN = 28, FP = 1, TN = 40251 respectively on the test data set. In fact, different application scenarios have different requirements for each indicator, such as in higher security access control, the FP is required to be as small as possible. While, a small FN value is more important in the case of troubleshoot suspects. Overall, the results of the first eight teams are better than the baseline method [25] when FPR = 10^{-4} on test data set.

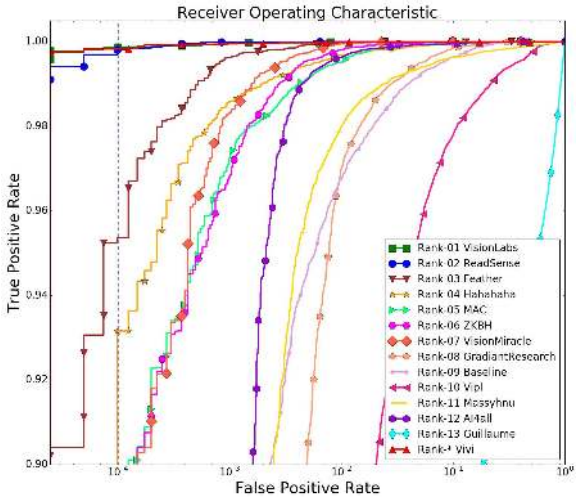

Figure 5. ROC curves of final stage teams on test set.

4.2. Challenge Results Analysis

As shown in Table 3, the results of the top three teams on test data set are clearly superior to other teams, revealing that ensemble learning has an exceptional advantage in deep learning compared to single model solutions under the same conditions, such as in Table 4 and Table 2. Simultaneously, analyzing the stability of the results of all participating teams' submission from the ROC curve 5, the three teams are significantly better than other teams on testing set (*e.g.*, TPR@FPR=10^{-4} values of these three teams are relatively close and superior to other teams). The team of ReadSense applies the image patch as input to emphasize the importance of local features in FAD task. The result of FN = 1 shows that the local feature can effectively prevent the model from misclassifying the real face into an attack one, shown in the blue box of Figure 6. Similarly, Vivi and Vipl-bpoic introduce the attention mechanism into FAD

[3] http://www.paddlepaddle.org/paddle/ModelAutoDL

Team Name	FP	FN	APCER(%)	NPCER(%)	ACER(%)	TPR(%)			data set
						@FPR=10e-2	@FPR=10e-3	@FPR=10e-4	
VisionLabs	**3**	27	**0.0074**	0.1546	0.0810	99.9885	**99.9541**	**99.8739**	test
ReadSense	77	**1**	0.1912	**0.0057**	0.0985	**100.0000**	99.9427	99.8052	
Feather	48	53	0.1192	0.1392	0.1292	99.9541	99.8396	98.1441	
Hahahaha	55	214	0.1366	1.2257	0.6812	99.6849	98.5909	93.1550	
MAC-adv-group	825	30	2.0495	0.1718	1.1107	99.5131	97.2505	89.5579	
ZKBH	396	35	0.9838	0.2004	0.5921	99.7995	96.8094	87.6618	
VisionMiracle	119	83	0.2956	0.4754	0.3855	99.9484	98.3274	87.2094	
GradiantResearch	787	250	1.9551	1.4320	1.6873	97.0045	77.4302	63.5493	
Baseline	1542	177	3.8308	1.0138	2.4223	96.7464	81.8321	56.8381	
Vipl-bpoic	1580	985	3.9252	5.6421	4.7836	82.9877	55.1495	39.5520	
Massyhnu	219	621	0.5440	3.5571	2.0505	98.0009	72.1961	29.2990	
AI4all	273	100	0.6782	0.5728	0.6255	99.6334	79.7571	25.0601	
Guillaume	5252	1869	13.0477	10.7056	11.8767	15.9530	1.5953	0.1595	
Vivi*	7	15	0.0173	0.0859	**0.0516**	99.9828	99.9484	99.8282	

Table 3. Results and rankings of the final stage teams, the best indicators are bold. Note that the results on the test set are tested by the model we trained according to the code submitted by the participating teams.(∗ indicates Vivi is affiliated with the sponsor and does not participate in the final ranking).

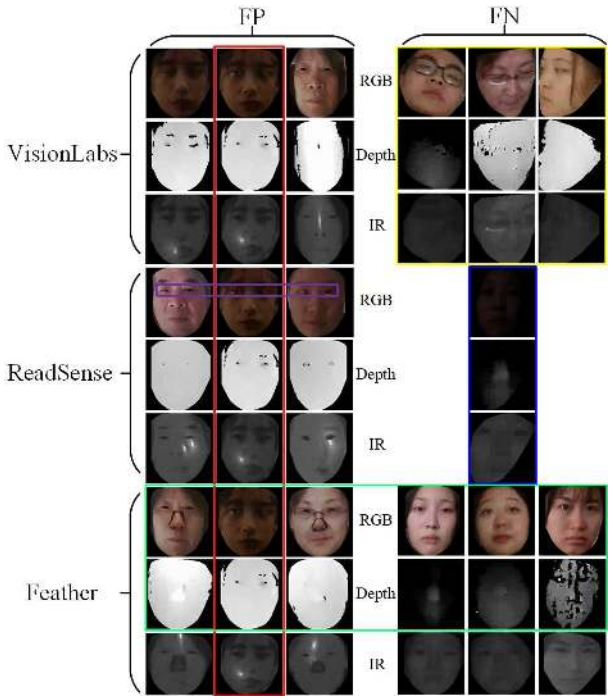

Figure 6. Mistaken samples of the top three teams on the Testing data set, including FP and FN. Note that the models were trained by us.

task. Since different modalities have different advantages: the RGB data have rich details, the Depth data is sensitive to the distance between the image plane and the corresponding face, and the IR data measures the amount of heat radiated from a face. Based on this characteristic, Feather uses a cascaded architecture with two subnetworks to study CASIA-SURF with two modalities, in which Depth and IR data are learnt subsequently by each network. Some teams consider face landmark (*e.g.*, Hahahaha) into FAD task, and other teams (*e.g.*, MAC-adv-group, Massyhnu) focus on the color space conversion. In stead of binary classification model, ZKBH constructs a regression model to supervise the model to learn effective cues. GradiantResearch reformulates the face-PAD as an anomaly detection using deep metric learning.

Although these methods have their own advantages, there are still some shortcomings in the code reproduction stage of the challenge. As described before, CASIA-SURF is characterized by multi-modal data (*i.e.*, RGB, Depth and IR) and the main research point is how to fuse the complementary information between these three modalities. However, many teams apply ensemble learning that is a way of Naive Halfway Fusion [25] in fact, which cannot make full use of the characteristics between different modalities. In addition, most of the ensemble methods use greedy manner for model fusion, including constantly increase the model if the performance does not decrease on the valid set in Table 4, which inevitably brings additional time consumption and instability. In order to demonstrate the shortcomings of the algorithm visually, we randomly selected 6 misclassified samples for each of the top three teams on the test set, of which the FP and FN are 3 respectively, as shown in Figure 6. Notably, the fake sample in the red box was simultaneously misclassified into real face by the three teams, where the clues were visually seen in the eye portion of the color modality. From the misclassification samples of the VisionLabs team, face pose is the main factor leading to FN samples (marked by a yellow box). As for the FP samples of ReadSense, the main clues are concentrated in the eye region (shown in the purple box). However, image patches applied by this team as the input of network, which is easy to cause misclassification if the image block does not contain an eye region. Only Depth and IR modal data sets were used by Feather team, resulting in misclassified samples that can be recognized by the human eyes easily. As

shown in green box, obvious clues which attached on the nose and eyes region in the color modal data sets are discarded by their algorithm. Overall analysis, the three teams have better recognition performance than Attack 1, 3, 5 for Attack 2, 4, 6 (performing a bending operation on the corresponding former) [25]. It shows that the bending operation used by simulating the depth information of the real face is easily detected by the algorithms. Last but notable, from the FP samples of the three teams, the misclassified samples are mainly caused by Attack 1, indicating that the sample with some regions are cut from the printed face can bring the depth information of the real face, but introducing more cues which can prove itself is fake one.

5. Open Issues and Opportunities

5.1. Critical Issues and Breakthrough Point

Face PAD remains a challenging problem due to lack of generalization and far from meeting the requirements of practical applications, mainly in the following aspects:

Intra-testing. Results vary greatly across different testing set scales, such as the performance gap between the same team on validation and test set (except the top three teams) listed in Table 3.

Inter-testing. Existing PAD algorithms rely heavily on the data used in training phase, and is easily affected by different attack types, acquisition devices and spoofing mediums presented in other datasets.

An important reason for the poor generalization ability is also used by most of the participating teams in this challenge (see Table 2). It is a CNN with softmax loss might discover arbitrary cues, such as spot or screen bezel of the spoof medium, that are not the faithful spoof patterns. When these cues disappear during testing, these models will fail to distinguish fake *v.s* real faces and result in poor generalization [17].

Therefore, the supervision should be designed from the essential differences between live and spoof faces, such as the rPPG signals (*i.e.*, heart pulse signal) which can reflect human physiological signs. From the perspective of image imaging, the depth information of face image has essential differences between real and fake face due to the real face is taken from one shot, while the fake one belongs to second imaging from a print or replay attack which has flat or planar depth. From the perspective of light reflection, the imaging light of real and fake face image comes from diffuse and specular reflection respectively, which may result a difference in the noise distribution between the real and fake images. Finally, from the perspective of multi-frame, information between video frames of a live sample is different from a fake video clip, especially in the face of static images or print attacks.

5.2. Future Work and Opportunities

In order to take full advantage of this multi-modal dataset, as future work, we plan to define a series of cross-modal testing protocols that are different from this challenge, *e.g.*, training on RGB images, and testing on Depth or IR modal data sets. The original intention of cross-modal protocol design is to guide the model to learn the relevant information between different modalities for the same category, *e.g.*, real or fake. Further, we will focus on improving the generalization ability of face PAD algorithms by designing supervision information that mentioned in section 5.1.

In addition, as the attack techniques are constantly upgraded, some new types of PA have emerged, *e.g.* 3D masks or custom-made silicone masks, which are more realistic in terms of texture and depth information than traditional 2D PAs, such as photos or video-replay attacks. In fact, a substantial portion of 2D PAD methods are rendered inoperative when 3D facial masks are introduced for attacks [1]. Therefore, we plan to collect a 3D masks dataset including head-mounted mask and face silicone models to push the research for countering 3D face mask attack.

6. Conclusion

We organized the *Chalearn LAP multi-modal face anti-spoofing attack detection challenge* based on the CASIA-SURF dataset and running on the CodaLab platform. Three hundred teams registered for the competition and thirteen teams made it to the final stage. Among the latter, teams were formed by ten companies and three academic institutes/universities. We described the associated dataset, and the challenge protocol including evaluation metrics. We reviewed in detail the proposed solutions and reported the results from both development and final phases. We analyzed the results of the challenge, pointing out the critical issues in FAD task and presenting the shortcomings of the existing algorithms. Future lines of research in the field have been also discussed.

7. Acknowledgement

This work has been partially supported by Science and Technology Development Fund of Macau (Grant No. 0025/2018/A1), by the Chinese National Natural Science Foundation Projects #61876179, #61872367, the Spanish project TIN2016-74946-P (MINECO/FEDER, UE) and CERCA Programme / Generalitat de Catalunya, and by ICREA under the ICREA Academia programme. We acknowledge Surfing Technology Beijing co., Ltd (www.surfing.ai) to provide us this high quality dataset. We also acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

References

[1] S. Bhattacharjee, A. Mohammadi, and S. Marcel. Spoofing deep face recognition with custom silicone masks. 2018. 1, 8

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

[3] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018. 5

[4] M. M. Chakka, A. Anjos, S. Marcel, R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, et al. Competition on counter measures to 2-d facial spoofing attacks. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–6. IEEE, 2011. 2

[5] I. Chingovska, N. Erdogmus, A. Anjos, and S. Marcel. Face recognition systems under spoofing attacks. In *Face Recognition Across the Imaging Spectrum*, pages 165–194. Springer, 2016. 1

[6] I. Chingovska, J. Yang, Z. Lei, D. Yi, S. Z. Li, O. Kahm, C. Glaser, N. Damer, A. Kuijper, A. Nouak, et al. The 2nd competition on counter measures to 2d face spoofing attacks. In *2013 International Conference on Biometrics (ICB)*, pages 1–6. IEEE, 2013. 2

[7] A. Costa-Pazo, D. Jimnez-Cabello, E. Vazquez-Fernandez, J. L. Alba-Castro, and R. J. Lpez-Sastre. Generalized presentation attack detection: a face anti-spoofing evaluation proposal. 2019. 4, 5

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 4, 5

[9] N. Erdogmus and S. Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *BTAS*, pages 1–6, 2014. 1

[10] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *JVCIR*, 38:451–460, 2016. 3

[11] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: Challenge of recognizing one million celebrities in the real world. *Electronic Imaging*, 2016(11):1–6, 2016. 5

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 4, 5, 6

[13] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[14] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017. 3

[15] A. Jourabloo, Y. Liu, and X. Liu. Face de-spoofing: Anti-spoofing via noise modeling. *arXiv preprint arXiv:1807.09968*, 2018. 1, 3

[16] J. Komulainen, A. Hadid, and M. Pietikainen. Context based face anti-spoofing. In *BTAS*, pages 1–8, 2013. 3

[17] Y. Liu, A. Jourabloo, and X. Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398, 2018. 1, 3, 8

[18] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 116–131, 2018. 4, 5

[19] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. 5

[20] G. Pan, L. Sun, Z. Wu, and S. Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *ICCV*, pages 1–8, 2007. 3

[21] G. Pan, L. Sun, Z. Wu, and Y. Wang. Monocular camera-based face liveness detection by combining eyeblink and scene context. *Telecommunication Systems*, pages 215–225, 2011. 3

[22] F. Peng, L. Qin, and M. Long. Ccolbp: Chromatic co-occurrence of local binary pattern for face presentation attack detection. In *2018 27th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–9. IEEE, 2018. 6

[23] F. Peng, L. Qin, and M. Long. Face presentation attack detection using guided scale texture. *Multimedia Tools and Applications*, pages 1–27, 2018. 6

[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018. 3, 4, 5

[25] A. L. C. Z. J. W. S. E. H. S. Z. W. S. Z. L. Shifeng Zhang, Xiaobo Wang. A dataset and benchmark for large-scale multi-modal face anti-spoofing. *arXiv:1812.00408v2*, 2018. 1, 2, 3, 5, 6, 7, 8

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3, 5, 6

[27] S. Sun, J. Pang, J. Shi, S. Yi, and W. Ouyang. Fishnet: A versatile backbone for image, region, and pixel level prediction. In *Advances in Neural Information Processing Systems*, pages 762–772, 2018. 3, 4, 5

[28] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 5

[29] L. Wang, X. Ding, and C. Fang. Face live detection method based on physiological motion analysis. *Tsinghua Science & Technology*, 14(6):685–690, 2009. 3

[30] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 5

[31] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the Euro-*

pean Conference on Computer Vision (ECCV), pages 3–19, 2018. 5

[32] X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 5, 6

[33] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017. 4, 5

[34] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 1, 5

[35] J. Zhao, Y. Cheng, Y. Xu, L. Xiong, J. Li, F. Zhao, K. Jayashree, S. Pranata, S. Shen, J. Xing, et al. Towards pose invariant face recognition in the wild. In *CVPR*, pages 2207–2216, 2018. 5

[36] Y. Zhu and S. Newsam. Densenet for dense flow. In *2017 IEEE international conference on image processing (ICIP)*, pages 790–794. IEEE, 2017. 5