


REVIEW

Cross-ethnicity face anti-spoofing recognition challenge: A review

Ajian Liu¹ | Xuan Li² | Jun Wan³  | Yanyan Liang¹ | Sergio Escalera⁴ |
Hugo Jair Escalante^{5,6} | Meysam Madadi⁴ | Yi Jin² | Zhuoyuan Wu⁷ |
Xiaogang Yu⁷ | Zichang Tan³ | Qi Yuan⁷ | Ruikun Yang¹ | Benjia Zhou¹ |
Guodong Guo⁸ | Stan Z. Li^{1,3,9}

¹Faculty of Information Technology, Macau University of Science and Technology, Taipa, Macau, China

²School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China

³National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

⁴Universitat de Barcelona and Computer Vision Center, Barcelona, Spain

⁵Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico

⁶Department of Computer Science, CINVESTAV-Zacatenco, Mexico City, Mexico

⁷School of Software, Beihang University, Beijing, China

⁸Baidu Research and National Engineering Laboratory for Deep Learning Technology and Application, Institute of Deep Learning, Beijing, China

⁹Westlake University, Hangzhou, China

Correspondence

Jun Wan, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China.
Email: jun.wan@ia.ac.cn

Funding information

Chinese National Natural Science Foundation Projects; Science and Technology Development Fund of Macau, Grant/Award Number: 0025/2018/A1, 0008/2019/A1, 0019/2018/ASC, 0010/2019/AFJ, 0025/2019/AKP; Key Project of the General Logistics Department, Grant/Award Number: ASW17C001; Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE)

Abstract

Face anti-spoofing is critical to prevent face recognition systems from a security breach. The biometrics community has achieved impressive progress recently due to the excellent performance of deep neural networks and the availability of large datasets. Although ethnic bias has been verified to severely affect the performance of face recognition systems, it still remains an open research problem in face anti-spoofing. Recently, a multi-ethnic face anti-spoofing dataset, CASIA-SURF cross-ethnicity face anti-spoofing (CeFA), has been released with the goal of measuring the ethnic bias. It is the largest up to date CeFA dataset covering three ethnicities, three modalities, 1607 subjects, 2D plus 3D attack types and the first dataset including explicit ethnic labels among the recently released datasets for face anti-spoofing. We organized the Chalearn Face Anti-spoofing Attack Detection Challenge which consists of single-modal (e.g. RGB) and multi-modal (e.g. RGB, Depth, infrared) tracks around this novel resource to boost research aiming to alleviate the ethnic bias. Both tracks have attracted 340 teams in the development stage, and finally, 11 and eight teams have submitted their codes in the single-modal and multi-modal face anti-spoofing recognition challenges, respectively. All of the results were verified and re-ran by the organizing team, and the results were used for the final ranking. This study presents an overview of the challenge, including its design, evaluation protocol and a summary of results. We analyse the top-ranked solutions and draw conclusions derived from the competition. Besides, we outline future work directions.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. *IET Biometrics* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

1 | INTRODUCTION

Face anti-spoofing aims to determine whether the captured face from a face recognition system is real or fake. It is essential to protect face recognition systems from malicious attacks, such as a printed face photograph (i.e., print attack), displaying videos on digital devices (i.e., replay attack) or even 3D attacks (i.e., face mask). Therefore, the presentation attack detection (PAD) task is a critical stage for visual face recognition systems which has been widely applied in financial payment, access control, phone unlocking and surveillance. Some early temporal-based face PAD works [1–4] attempt to detect the evidence of liveness (e.g., eye-blinking), which require a constrained human interaction. However, these methods become vulnerable if someone presents a replay attack or a print photo attack with cut eye/mouth regions. Other works are based on static texture analysis [5,6]. However, these algorithms are not accurate enough because of the use of handcrafted features, such as LBP [7–9], HoG [8–10] and GLCM [10], that do not necessarily capture the most discriminative information associated to the data. Recently, CNN-based face PAD methods [11–16] have shown impressive progress due to the excellent performance of deep neural networks [11,14,15,17] and the availability of large datasets [15,18–22]. Although these methods achieve near-perfect performance in intra-database experiments, they are still vulnerable when facing complex authentication scenarios. In particular, ethnic bias has been verified to severely affect the performance of face recognition systems [23–24], representing an open research problem in face anti-spoofing.

We have verified in our own previous work [22] that state-of-the-art (SOTA) PAD algorithms do suffer from severe ethnic bias. For example, the average classification error rate (ACER) metric values vary widely on the test samples with different ethnicities for the same algorithm. To alleviate the ethnic bias and ensure that face PAD methods are in a safe reliable condition for users of different ethnicities, Liu et al. [22] introduced the largest up-to-date cross-ethnicity face anti-spoofing (CeFA) dataset, covering three ethnicities, three modalities, 1607 subjects and 2D plus 3D attack types. Some samples of the CASIA-SURF CeFA dataset are shown in Figure 1. Four protocols were defined to measure the effect under varied evaluation conditions, such as cross-ethnicity, unknown spoofs or both of them. To the best of our knowledge, CeFA is the first dataset including explicit ethnic labels among the published datasets for face anti-spoofing. Additionally, they provided a baseline including two aspects to alleviate the above bias: (1) a static–dynamic fusion mechanism applied in each modality (i.e., RGB, Depth, infrared [IR] image) and (2) a partially shared fusion strategy is proposed to learn complementary information from multiple modalities.

Leveraging on the CeFA dataset, we organized the *Chalearn Face Anti-spoofing Attack Detection Challenge* comprising *single-modal* (e.g., RGB) and *multi-modal* (e.g., RGB, Depth, IR) tracks collocated with the Workshop on Media Forensics at

CVPR2020. The goal of this challenge was to boost research on facial PAD aiming to alleviate the ethnic bias. Both tracks, single-modal (<https://competitions.codalab.org/competitions/22151>) and multi-modal tracks (<https://competitions.codalab.org/competitions/22036>), were run simultaneously on the Codalab platform. The competition attracted 340 teams in the development stage, with 11 and 8 teams entering the final evaluation stage for the single-modal and multi-modal face anti-spoofing recognition tracks, respectively. Summaries with the names and affiliations of teams that entered the final stage are shown in Tables 1 and 2 for the single-modal and multi-modal tracks, respectively.

Compared to previous challenges on related topics [25–28], the algorithms of all participating teams were based on deep learning and did not require external resources (e.g., additional datasets and pre-trained models). This was a rule established in the challenge that not only provides a fairer evaluation scenario but also brings benefits for reproducibility and algorithm implementation in practical applications. To sum up, the contributions of this study are summarized as follows:

- We describe the design and organization of both tracks of the *Chalearn Face Anti-spoofing Attack Detection Challenge*, which is based on the CASIA-SURF CeFA dataset and was run on the CodaLab platform
- We provide a complete description of solutions developed in the context of the challenge
- We point out critical points on face anti-spoofing detection by comparing essential differences between a real face and a fake one from multiple aspects, also discussing future lines of research in the field

2 | CHALLENGE OVERVIEW

In this section, we describe the organized challenge, including a brief introduction to the CASIA-SURF CeFA dataset, evaluation metrics and the challenge protocol.

2.1 | CASIA-SURF CeFA

CASIA-SURF CeFA [22] is the largest up-to-date CeFA dataset, covering 3 ethnicities, 3 modalities, 1604 subjects, and 2D plus 3D attack types. More importantly, it is the first public dataset designed for exploring the impact of cross-ethnicity in the study of face anti-spoofing. Some samples of the CASIA-SURF CeFA dataset are shown in Figure 1.

The main motivation of CASIA CeFA dataset is to serve as a benchmark to allow for the evaluation of the generalization performance of new PAD methods. Concretely, four protocols were originally introduced to measure the robustness of methods under varied evaluation conditions: (1) cross-ethnicity (Protocol 1), (2) cross-PAI (Protocol 2), (3) cross-modality (Protocol 3), and (4) cross-ethnicity and cross-PAI (Protocol 4). To make the competition more challenging, we adopted

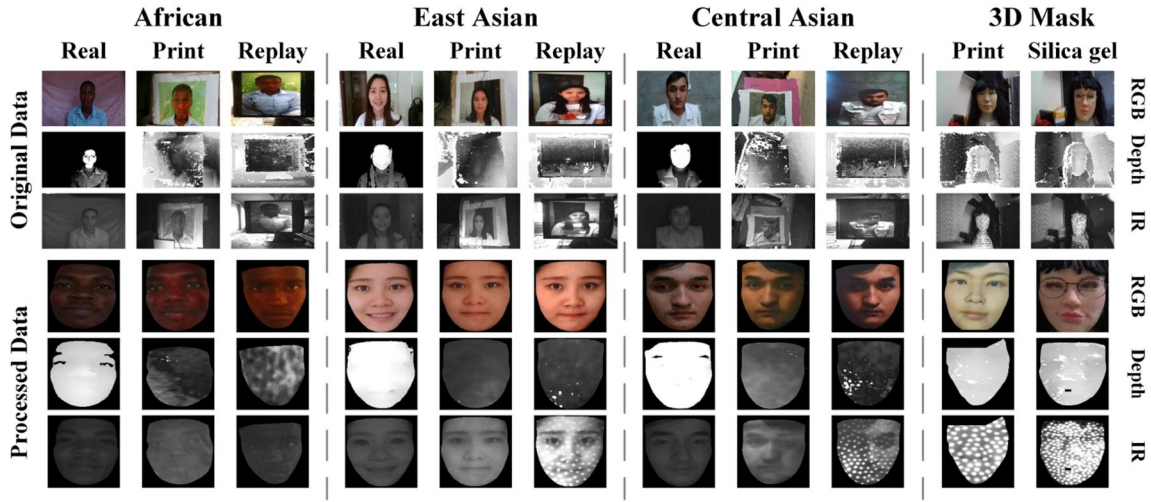


FIGURE 1 Samples of the CASIA-SURF cross-ethnicity face anti-spoofing dataset. It contains 1607 subjects and three different ethnicities (i.e., Africa, East Asia and Central Asia), with four attack types (i.e., print attack, replay attack, 3D print and silica gel attacks). IR, infrared

TABLE 1 Team and affiliations name are listed in the final ranking of this challenge (single-modal)

Ranking	Team name	Leader name, affiliation
1	VisionLabs	Alexander Parkin, visionlabs
2	BOBO	Zitong Yu, OULU unv.
3	Harvest	Jiachen Xue, Horizon
4	ZhangTT	Zhang Tengting, CMB
5	Newland_tianyan	Xinying Wang, Newland Inc.
6	Dopamine	Wenwei Zhang, huya
7	IecLab	Jin Yang, HUST
8	Chuanghua Telecom Lab.	Li-Ren Hou, Chunghwa Telecom
9	Wgqtmac	Guoqing Wang, ICT
10	Hulking	Yang, Qing, Intel
11	Dqiu	Qiudi

TABLE 2 Team and affiliations name are listed in the final ranking of this challenge (multi-modal)

Ranking	Team name	Leader name, affiliation
1	BOBO	Zitong Yu, OULU unv.
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

Protocol 4 in this challenge, which is designed by combining conditions of Protocols 1 and 2. As shown in Table 3, it has three data subsets: training, validation and testing sets, which contain 200, 100 and 200 subjects for each ethnicity, respectively. Note that the remaining 107 subjects are 3D masks. To fully measure the cross-ethnicity performance of the algorithm, one ethnicity is used for training and validation, and the remaining two other ethnicities are used for testing. Since there are three ethnicities in CASIA-SURF CeFA, a total of three sub-protocols (i.e., 4_1, 4_2 and 4_3 in Table 3) are adopted in this challenge. In addition to the ethnic variation, the factor of PAIs is also considered in this protocol by setting different attack types in training and testing phases.

2.2 | Evaluation metrics

In this challenge, we selected the recently standardized ISO/IEC 30107-3 (<https://www.iso.org/obp/ui/iso>) metrics for evaluation: attack presentation classification error rate (APCER), normal presentation classification error rate (NPCER) and ACER; these are defined as follows:

$$\text{APCER} = \text{FP} / (\text{FP} + \text{TN}) \quad (1)$$

$$\text{NPCER} = \text{FN} / (\text{FN} + \text{TP}) \quad (2)$$

$$\text{ACER} = (\text{APCER} + \text{NPCER}) / 2 \quad (3)$$

where TP, FP, TN and FN correspond to true positive, false positive, true negative and false negative, respectively. APCER and NPCER 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 CeFA

TABLE 3 Protocols and statistics

Track			Subjects (one ethnicity)	Ethnicity			PAIs	# Num.img (rgb)		
S	M	Subset		4_1	4_2	4_3		4_1	4_2	4_3
		Train	1–200	A	C	E	Replay	33,713	34,367	33,152
		Valid	201–300	A	C	E	Replay	17,008	17,693	17,109
		Test	301–500	C & E	A & E	A & C	Print	105,457	102,207	103,420

Note: A, C and E are short for Africa, Central Asia and East Asia, respectively. Track (S/M) means the single/multi-modal track. The PAIs means the presentation attack instruments.

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.

2.3 | Challenge protocol

The challenge was run in the CodaLab platform, and comprised two stages as follows:

2.3.1 | Development phase (started in 13 December 2019 and ended in 1 March 2020)

During this phase, participants had access to the labelled training subset and unlabelled validation subset. Since the protocol used in this competition (Protocol 4) comprises three sub-protocols (see Section 2.1), participants first need to train a model for each sub-protocol, then predict the score of the corresponding validation set, and finally, simply merge the predicted scores and submit them to the CodaLab platform and receive immediate feedback via a public leader board.

2.3.2 | Final phase (started in 1 March 2020 and ended in 10 March 2020)

During this phase, labels for the validation subset and the unlabelled testing subset were released. Participants can firstly take the labels of the validation subset to select a model with better performance, then they can use this model to predict the scores of the corresponding testing subset samples, and finally, submit the score files in the same way as the development phase. We made public all results of the three sub-protocols online; these include the obtained values of APCER, BPCER and ACER. Like Boulkenafet et al. [20], the mean and variance of evaluated metrics for these three sub-protocols are calculated for the final results.

Note that to fairly compare the performance of participants' algorithms, this competition does not allow the use of other training datasets and pre-trained models. To be eligible for prizes, winners had to publicly release their code under a licence of their choice and provide a fact sheet describing their solution. Besides, the code was re-run and all of the results were verified by the organizing team after the final

phase ended, the verified results were used for the final ranking.

3 | DESCRIPTION OF SOLUTIONS

In the final ranking stage, there were 19 teams submitting their code and fact sheets (for your reference, these are available in this link: http://www.cbsrlia.ac.cn/users/jwan/fact_sheet/spoofing_fact_sheet-cvprw2020.zip) for evaluation. According to the information provided, in the following, we describe the solutions developed by each of the teams, with detailed descriptions for top-ranked participants in both single-modal (RGB) and multi-modal (RGB, Depth, IR) face anti-spoofing recognition challenge tracks.

Tables 1 and 2 show the final ranking for both tracks. It can be seen from these tables that most participants came from the industrial community. Interestingly, the VisionLabs team was not only the winner of the single-modal track, but also the winner of the *Chalearn LAP multi-modal face anti-spoofing attack detection challenge at CVPR 2019* [28]. In addition, the BOBO team designed central difference convolution (CDC) [29] and contrastive depth loss (CDL) [30] for feature learning, and achieved second and first place in both single-modal and multi-modal tracks, respectively.

3.1 | Single-modal face anti-spoofing challenge track

3.1.1 | Baseline

We provided a baseline for approaching this task via designing a SD-Net [22] which takes Resnet18 [31] as the backbone. As shown in Figure 2, it contains three branches: static, dynamic and static–dynamic branches, which learn hybrid features from static and dynamic images. For static and dynamic branches, each of them consists of five blocks (i.e., conv, res1, res2, res3 and res4) and one global average pooling (GAP) layer, while in the static–dynamic branch, the conv and res1 blocks are removed because it takes fused features of res1 blocks from static and dynamic branches as input.

For dynamic image generation, a detailed description is provided in [22]. In short, we compute its dynamic image online with rank pooling using K consecutive frames. Our selection of dynamic images for rank pooling in SD-Net is

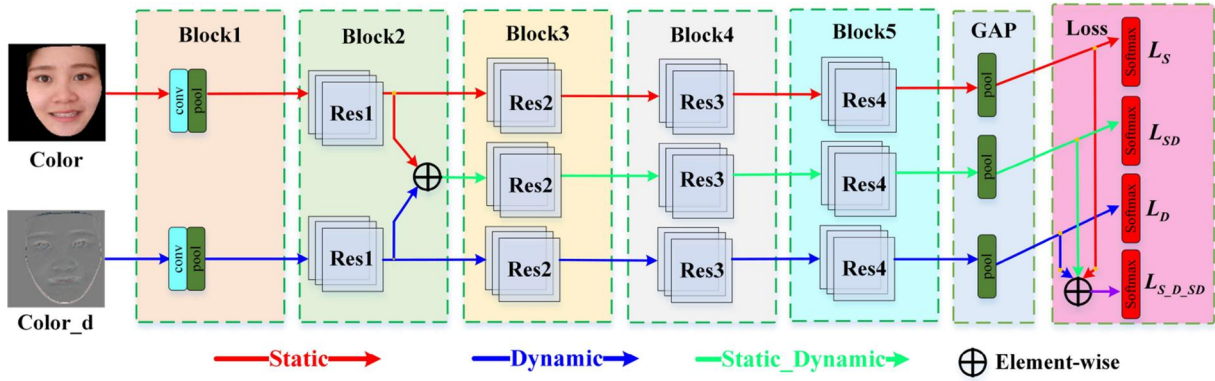


FIGURE 2 The framework of SD-Net. The figure is provided by the baseline team and ranked NO.11 in single-modal track

further motivated by the fact that dynamic images have proved its superiority to regular optical flow [32,33].

3.1.2 | VisionLabs

Due to high differences in the train and test subsets (i.e., different ethnics and attack types), the VisionLabs team used a data augmentation strategy to help train robust models. Similar to previous works which convert RGB data to HSV and YCbCr colour spaces [34], or Fourier spectrum [35], they decided to convert RGB to other ‘modalities’, which contain more authentic information instead of identity features. Specially, the Optical Flow and RankPooling are used as shown in Figure 3.

As shown in Figure 3, the proposed architecture consists of four branches where two branches are used for dynamic images via a dynamic pooling algorithm, and the left two branches are used for the optical flow images. For optical flow modality, they calculated two flows between the first and last images of RGB video as well as between the first and second images. For the rank pooling modality, they used the rank pooling algorithm [33] where different hyper-parameters used to generate two different dynamic images.

Formally, a RGB video with K frames is represented by $\{X_i^k\}$, where $i = 0, \dots, K-1$ and $t = \{0, 1\}$ is the label (0 – fake, 1 – real). Then for each RGB video, they sample $L = 16$ images uniformly, obtaining $\{X_j^k\}$, where $j = 0, \dots, 15$. Then, they remove black borders and pad image to be square of size (112, 112). Then they apply intensive equal colour jitter to all images, emulating different skin colours.

As shown in Figure 3, they apply four ‘modality’ transforms: rank pooling ($\{X_j^k\}$, $C = 1000$), rank pooling ($\{X_j^k\}$, $C = 1$), Flow (X_0^k, X_{15}^k), Flow (X_0^k, X_1^k), where C is the hyper-parameter for SVM in the rank pooling algorithm [33]. The code of rank pooling was released in <https://github.com/MRzzm/rank-pooling-python>. These transforms return four tensors with sizes $3 \times 112 \times 112$, $3 \times 112 \times 112$, $2 \times 112 \times 112$ and $2 \times 112 \times 112$ respectively. Further, the features of each modal sample are extracted by an independent network (namely SimpleNet and its structure depicted in Figure 3) with size of $d = 256$ and all features are concatenated

to get a tensor of shape $4 \times d$. Then they apply Max, Avg and Min pooling among the first dimension and concatenate results to get $3 \times d$ tensor. Finally, a binary cross-entropy is adopted in their network. The code of VisionLabs was released in https://github.com/AlexanderParkin/CASIA-SURF_CeFA.

3.1.3 | BOBO

Most CNN-based methods [11,12,14,36] only treat face anti-spoofing as a binary classification task, and train the neural network supervised by a softmax loss. However, these methods fail to explore the nature of spoof patterns [15], which consist of skin detail loss, colour distortion, moire pattern, motion pattern, shape deformation and spoofing artefacts. To relieve the above issues, similar to Wang et al. [30], the BOBO team adopts depth supervision instead of binary softmax loss for face anti-spoofing. Different from Wang et al. [30], they design a novel CDC [29] and a CDL for feature learning and representation.

The structure of the depth map regression network based on CDC is shown in Figure 4. It consists of three blocks, three attention layers connected after each block and three down-sampling layers followed by each attention layer. Inspired by the residual network, they use a short-cut connection, which is concatenating the responses of Low-level Cell (Block1), Mid-level Cell (Block2) and High-level Cell (Block3), and sending them to two cascaded convolutional layers for depth estimation. All convolutional layers use the CDC network which is followed by a batch normalization layer and a rectified linear unit activation function. The size of input image and regression depth map are $3 \times 256 \times 256$ and $1 \times 32 \times 32$, respectively. Euclidean distance loss (EDL) is used for pixel-wise supervision in this work which is formulated:

$$L_{EDL} = \|D_p - D_G\|_2^2, \quad (4)$$

where D_p and D_G are the predicted depth and ground-truth depth, respectively.

EDL applies supervision on the predicted depth based on pixel one by one, ignoring the depth difference among adjacent

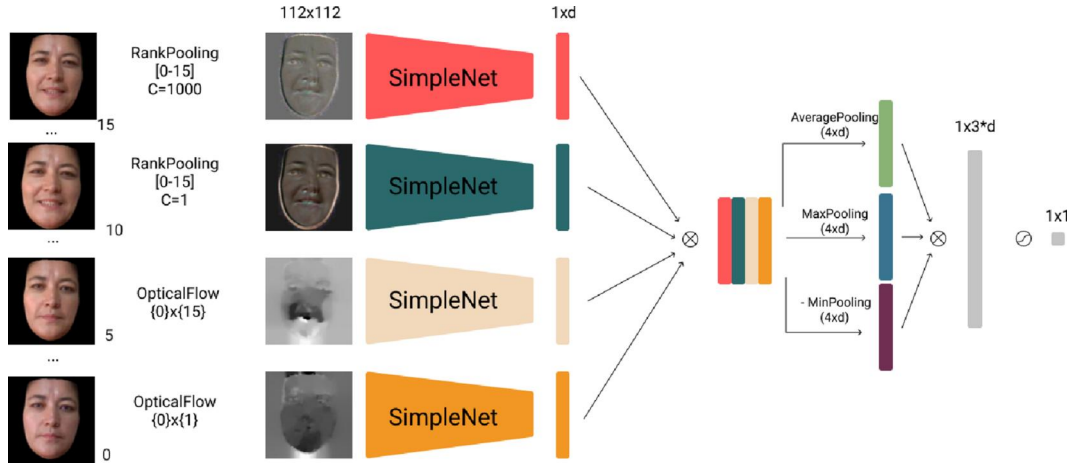


FIGURE 3 The framework is provided by the VisionLabs team. Note that the SimpleNet architecture: four blocks of Conv 3×3 – BatchNorm – Relu – MaxPool of sizes 16, 32, 64 and 128, followed by Conv 5×5 with 256 filters. The figure is provided by the VisionLabs team and ranked NO.1 in single-modal track

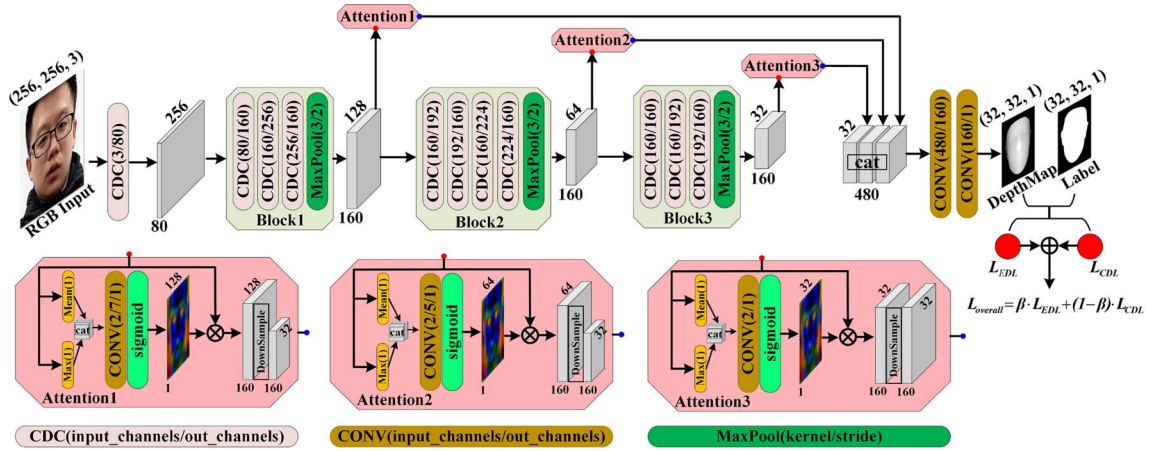


FIGURE 4 The framework of regression network. The figure is provided by the BOBO team and ranked NO.2 in single-modal track. CDC, central difference convolution

pixels. Intuitively, EDL merely assists the network to learn the absolute distance between the objects to the camera. However, the distance relationship of different objects is also important to be supervised for the depth learning. Therefore, one proposed the CDL to offer an extra supervision, which improves the generality of the depth-based face anti-spoofing model:

$$L_{CDL} = \sum_i \|K_i^{CDL} \odot D_P - K_i^{CDL} \odot D_G\|_2^2, \quad (5)$$

where K_i^{CDL} is the i^{th} contrastive convolution kernel, $i \in [0, 7]$. The details of the kernels can be found in Figure 5.

Therefore, the total loss L_{overall} employed by this team is defined as follows:

$$L_{\text{overall}} = \beta \cdot L_{EDL} + (1 - \beta) \cdot L_{CDL}, \quad (6)$$

where β is the hyper-parameter to trade-off EDL loss and CDL loss in the final overall loss L_{overall} . Finally, their code is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_0, \begin{bmatrix} 0 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_1, \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_2, \begin{bmatrix} 0 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}_3, \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}_4, \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}_5, \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix}_6, \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}_7$$

FIGURE 5 The kernel K_i^{contrast} in contrastive depth loss

publicly available in https://github.com/ZitongYu/CDCN/tree/master/FAS_challenge_CVPRW2020.

3.1.4 | Harvest

It can be observed from Table 3 that the attack types of the spoofs in the training and testing subsets are different. The Harvest team considered the motion information of real faces is also an important discriminative cue for face anti-spoofing attack detection. Therefore, how to effectively learn the motion

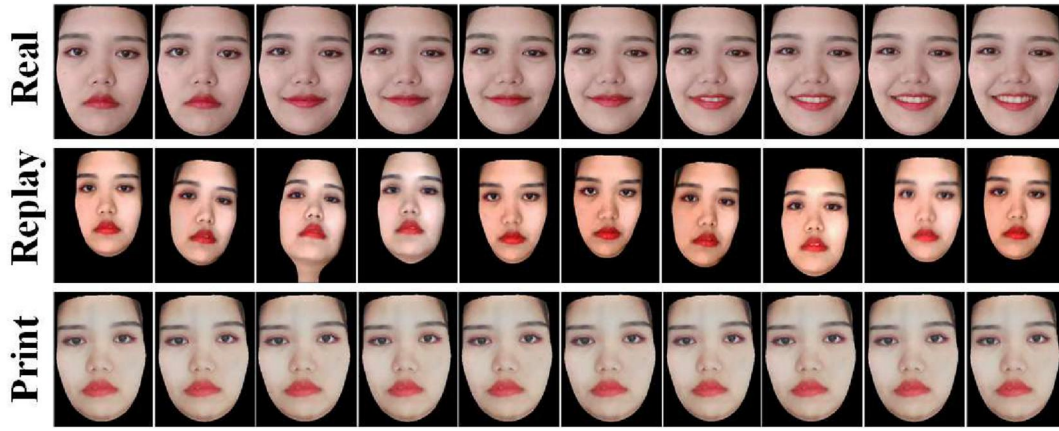


FIGURE 6 Visual comparison of real face, replay attack, print attack motion information for Harvest team

information of real faces from the interference motion information of the replay attack is a key step. As shown in Figure 6, the live frame displays obvious temporal variations, specially, in expressions, while there are very little facial changes in the print spoof samples for the same subject, which inspires the Harvest team to capture the subtle dynamic variations by re-labelling live sequence. Suppose the labels of spoof and live samples are 0 and 1 respectively. They define a new temporal-aware label via forcing the labels of the real face images in a sequence to change uniformly from 1 to 2, while the spoofing faces stay 0. Let $X = \{x_1, x_2, \dots, x_n\}$ denote a video containing n frames, where x_1 and x_n represent the first and final frames, respectively. They encode this implicit temporal information by reformulating the ground-truth label, such as

$$gt_i = 1 + \frac{i}{n}, \quad (7)$$

where the genuine label grows over time. Note that they do not encode the temporal variations in the spoof video due to their irregular variations in sequence. As shown in Figure 7, the overall framework consists of two parts as follows:

- (1) In the training stage, they encode inherent discriminative information by re-labelling live sequence
- (2) In inference stage, they aggregate the static-spatial features with dynamic-temporal information for sample classification. Finally, combined with the strong learning ability of backbone, their method achieved third in the single-modal track and the code is publicly available in <https://github.com/yueyechen/cvpr20>

3.1.5 | ZhangTT

Similar to the SD-Net in baseline [22], this team proposes a two-branch network to learn hybrid features from static and temporal images. They call it quality and time tensor, respectively. As shown in Figure 8, they take the ResNet [31] as the

backbone for each branch and use the single frame and multi-frame as the input of the two branches. Specially, the quality tensor and time tensor are first sent to a normal 7×7 receptive field convolution layer for preliminary feature extraction. After feature extraction by three independent blocks, a higher level expression quality feature map and time feature map were obtained. Then the quality feature and the time feature are concatenated together to form a new feature map for final classification with a binary cross-entropy loss function. The blocks in this work are the same as the ResNet block [31].

For data pre-processing, they first discarded the colour information by converting the RGB modality to grayscale space and then used histogram equalization to mitigate the skin-tone gap between ethnicities. Finally, they adopted the following four strategies to reduce the difference between replay and print attacks: (1) They regard face anti-spoofing work as a classification task for four classes instead of two. The four considered categories are live-invariable (label 0), fake-invariable (label 1), live-variable (label 2) and fake-variable (label 3), respectively; (2) dithering each channel of the attack sample solves the problem of consistency of each frame of the print attack; (3) to enhance the robustness, consider randomly superimposing Gaussian noise and superimposing gamma correction on each channel of the time tensor; and (4) to discriminate the texture difference, the first channel of the time tensor is separately identified and recorded as the quality tensor. It is sent to the network to extract features without noise superposition. Their code is publicly available in <https://github.com/ZhangTT-race/CVPR2020-SingleModal>.

3.1.6 | Newland_tianyan

This team mainly explores single-modal track from two aspects of data augment and network design. For data augmentation, on the one hand, they introduced print attacks in the training set by randomly pasting study textures on real face samples. On the other hand, they performed random rotation, movement, brightness transformation, noise and fold texture addition on

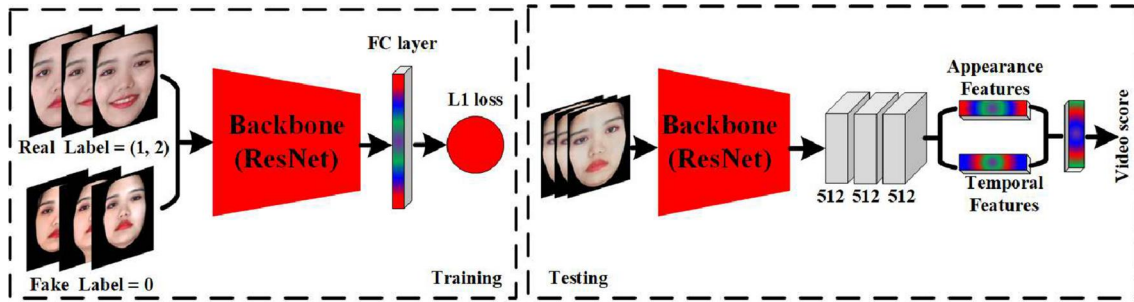


FIGURE 7 The framework of training and testing phases for Harvest. The figure is provided by the Harvest team and ranked NO.3 in single-modal track. IR, infrared

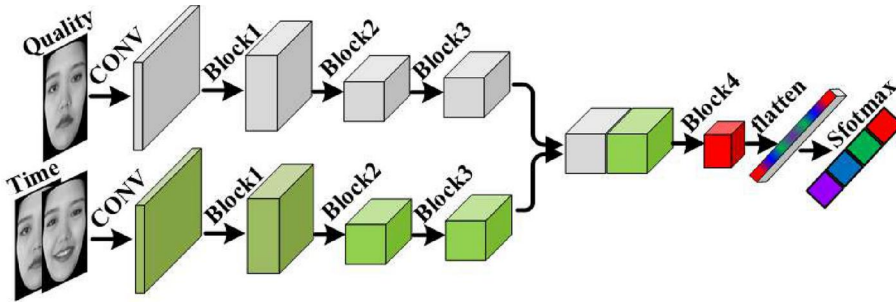


FIGURE 8 Architecture of the proposed model for the single-modal track. The figure is provided by the ZhangIT team and ranked NO.4 in the single-modal track

the same frame of real face to simulate the case that there is no micro expression change for the print attack. For network design, this team used a five-layer sequence network which takes 16 frames of samples as input to learn the temporal features. To improve the generalization faced with different ethnicities, the images were subtracted from the neighbourhood mean before sending to the network due to the samples of different ethnicities vary widely in skin colour. Their code is publicly available in <https://github.com/XinyingWang55/RGB-Face-antispoofing-Recognition>.

3.1.7 | Dopamine

This team uses face ID information for face anti-spoofing tasks. The architecture is shown in Figure 9, a multi-task network is designed to learn the features of identity and authenticity simultaneously. In the testing phase, these two scores are combined to determine whether a sample is a real face. They use the softmax score from the real/fake classifier and the feature computed by the backbone network (Resnet100) to compute the minimal similarity between the same person. In theory, the feature similarity score of the attack sample is close to 1, and the real face is close to 0. Their code is publicly available in https://github.com/xinedison/huya_face.

3.1.8 | IecLab

This team uses feathernet and 3DResNet [37] to learn the authenticity and expression features of the samples, and finally

merged the two features for anti-spoofing tasks. Their code is publicly available in <https://github.com/1relia/CVPR2020-FaceAntiSpoofing>.

3.1.9 | Chuanghwa telecom lab

This team combines subsequence features with Bag of local features [38] within the framework of MIMAMO-Net (<https://github.com/wtomin/MIMAMO-Net>). Finally, the ensemble learning strategy is used for feature fusion. Their code is publicly available in https://drive.google.com/open?id=1ouL1X69KIQEUI72iKH10_UvztdW8f_1.

3.1.10 | Wgqtmac

This team focused on improving face anti-spoofing generalization ability and proposed an end-to-end trainable face anti-spoofing approach based on deep neural network. They choose Resnet18 [31] as the backbone and use a warmup strategy to update the learning rate. The learnt model performs well on the developing subset. However, it is easily overfitted on the training set and gets worse results on the testing set. Their code is publicly available in <https://github.com/wgqtmac/cvprw2020.git>.

3.1.11 | HULKING

The main role of PipeNet proposed by this team is to selectively and adaptively fuse different modalities for face

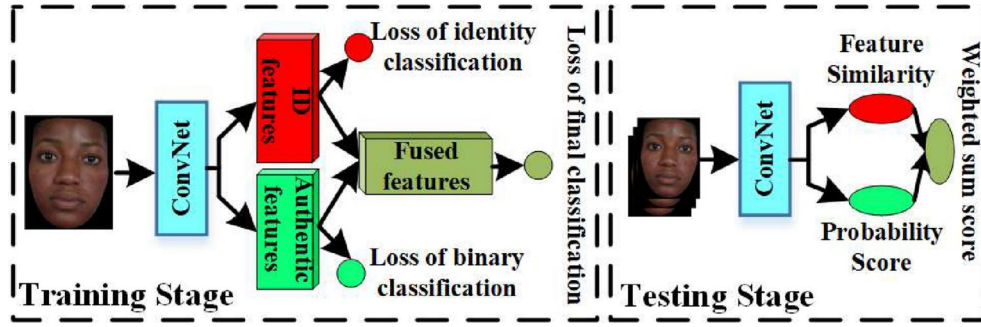


FIGURE 9 The architecture of the multi-task network for face anti-spoofing. The figure is provided by the Dopamine team and ranked NO.6 in the single-modal track

anti-spoofing tasks. Since the single-modal track only allows the use of RGB data, the team's method has limited performance in this challenge. We detail the team's algorithm in Section 3.2. Their code is publicly available in <https://github.com/muyiguangda/cvprw-face-project>.

3.1.12 | Dqiu

This team treats the face anti-spoofing as a binary classification task and uses Resnet50 [31] as the backbone to learn the features. Since no additional effective strategies were used, no good results were achieved on the testing set.

3.2 | Multi-modal face anti-spoofing challenge track

3.2.1 | Baseline

In order to take full advantage of multi-modal samples to alleviate the ethnic and attack bias, we propose a novel multi-modal fusion network, namely PSMM-Net [22]. As shown in Figure 10. It consists of two main parts: (a) the modality-specific network, which contains three SD-Nets to learn features from RGB, Depth and IR modalities, respectively; and (b) and a shared branch for all modalities, which aims to learn the complementary features among different modalities. To capture correlations and complementary semantics among different modalities, information exchange, and interaction among SD-Nets and the shared branch are designed.

There are two main kind of losses employed to guide the training of PSMM-Net. The first corresponds to the losses of the three SD-Nets, that is, colour, depth and IR modalities, denoted as $\mathcal{L}^{\text{color}}$, $\mathcal{L}^{\text{depth}}$ and \mathcal{L}^{IR} , respectively. The second corresponds to the loss that guides the entire network training, denoted as $\mathcal{L}^{\text{whole}}$, which bases on the summed features from all SD-Nets and the shared branch. The overall loss \mathcal{L} of PSMM-Net is denoted as follows:

$$\mathcal{L} = \mathcal{L}^{\text{whole}} + \mathcal{L}^{\text{color}} + \mathcal{L}^{\text{depth}} + \mathcal{L}^{\text{IR}} \quad (8)$$

3.2.2 | BOBO

For the multi-modal track, as shown in Figure 11, this team takes three independent networks (backbone) to learn the features of the three modalities (e.g., RGB, Depth, IR). Therefore, the entire structure consists of two main parts: (a) the modality-specific network, which contains three branches (the backbone network of each modality branch is not shared) to regress depth maps of RGB, Depth and IR modalities, respectively; and (b) a fused branch (via concatenation) for all modalities, which aims to learn the complementary features among different modalities and output final depth map with the same size ($1 \times 32 \times 32$) of the single-modal track. Similar to the single-modal track, the CDL and CDE loss functions are used in a multi-modal track in the form of weighted sums.

As the feature-level fusion strategy (see Figure 11) might not be optimal for all protocols, they also try two other fusion strategies: (1) input-level fusion via concatenating three-modal inputs to $256 \times 256 \times 9$ directly, and (2) score-level fusion via weighting the predicted score from each modality. For these two fusion strategies, the architecture of single-modal CDCN (see Figure 4) is used. Through comparative experiments, they concluded that the input-level fusion (i.e., simple fusion with concatenation) might be sub-optimal because it is weak in representing and selecting the importance of modalities. Therefore, this final result is combined with the best sub-protocols results (i.e., feature-level fusion for protocol 4_1, while score-level fusion for protocol 4_2 and 4_3). Specially for score-fusion, they weight the results of RGB and Depth modalities averagely as the final score (i.e., $\text{fusion_score} = 0.5 \times \text{RGB_score} + 0.5 \times \text{depth_score}$). This simple ensemble strategy helps to boost the performance significantly in their experiments.

3.2.3 | Super

CASIA-SURF CeFA is characterized by multi-modality (i.e., RGB, Depth, IR) and a key issue is how to fuse the complementary information between the three modalities. This team explored multi-modal track from three aspects: (1) Data pre-processing, (2) Network construction, and (3) Ensemble strategy design.

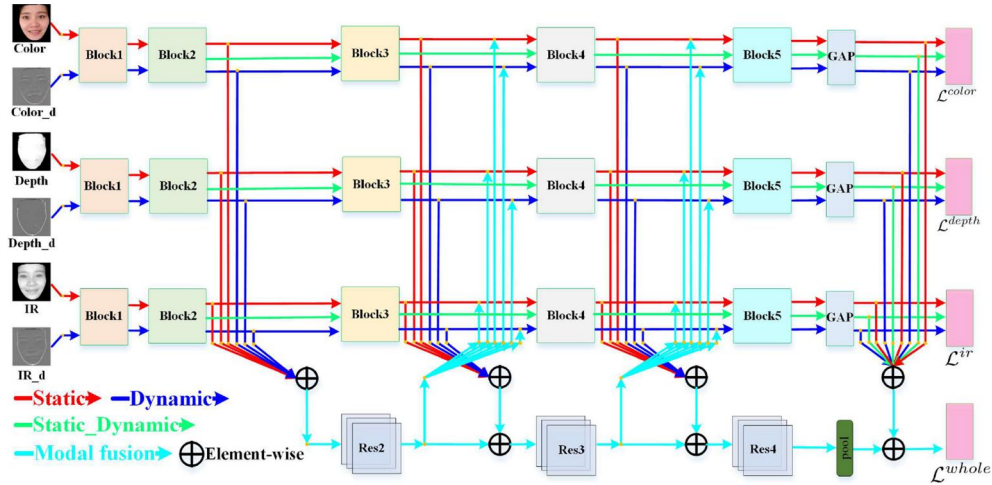


FIGURE 10 The framework of PSMM-Net. The figure is provided by the baseline team and ranked NO.8 in multi-modal track

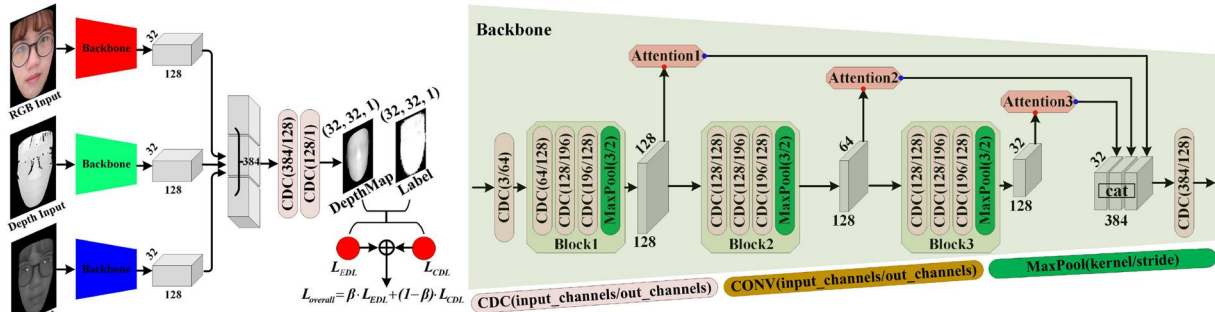


FIGURE 11 The framework of regression network for three modalities. The figure is provided by the BOBO team and ranked NO.2 in multi-modal track. CDC, central difference convolution

Since the dataset used in this competition retained the black background area outside the face, this team tried to remove the background area using the histogram threshold method to mitigate its interference effect on model learning. To increase the diversity of training samples, they use random rotation within the range of $[-30^\circ, 30^\circ]$, flipping, cropping and colour distortion for data augmentation. Note that the three modalities of the same sample are maintained in a consistent manner to obtain the features of the corresponding face region.

Inspired by Zhang et al. [21] which employs the ‘Squeeze-and-Excitation; Block (SE Block) [39] to re-weighting the hierarchy features of each modality, this team takes a multi-stream architecture with three subnetworks to study the dataset modalities, as shown in Figure 12. We can see that the RGB, Depth and IR data are learnt separately by each stream, and then shared layers are appended at a point (Res-4) to learn joint representations. However, the single-scale SE block [39] does not make full use of features from different levels. To this end, they extend the SE fusion from a single scale to multiple scales. As shown in Figure 12, the Res-1, Res-2 and Res-3 blocks from each stream extract features from different modalities. After that, they first fuse features from different modalities via the

SE block after Res-1, Res-2 and Res-3, respectively, then concatenate these fused features and sending them to aggregation block (Agg Block), next merging these features (including shared branch features after the GAP) via element summation operations similar to [40]. Finally, they use the merged features to predict real and fake. Differently from [40], they add a dimension reduction layer before the fully connected layer for avoiding the overfitting.

To increase the robustness to unknown attack types and ethnicities, they design several new networks based on the basic network shown in Table 4. Such as the Network A with a dimension reduction layer and without SE fusion after each res block. While the Network B and C are similar to [21,40] respectively. For the IR_ResNet50, it uses the improved residual block which aims at fitting the face recognition task. In the experiments, they found that different networks performed differently under the same sub-protocol. Therefore, they selectively trained these networks according to different sub-protocols and get the final score via averaging the results of selected networks. Their code is publicly available in https://github.com/hzh8311/challenge2020_face_anti_spoofing.

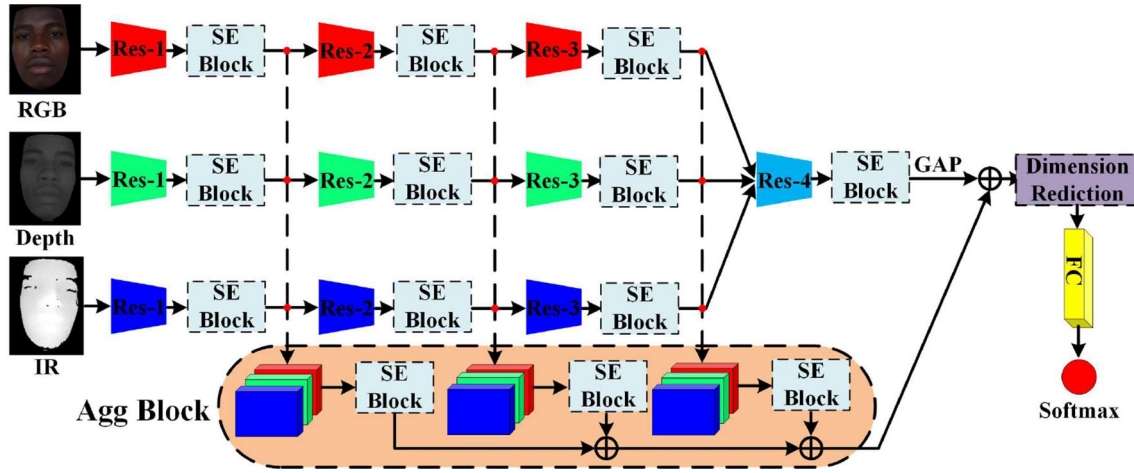


FIGURE 12 The framework of Super team. The ResNet34 or IR_ResNet50 as the backbone. The figure is provided by the Super team and ranked NO.2 in the multi-modal track. SE, Squeeze-and-Excitation

TABLE 4 The networks ensemble ways adopted by Super team. Each network carries functions marked by ✓

Network	Backbone	SE block	Dimension reduction	Agg block
A	ResNet34		✓	✓
B	ResNet34	✓		✓
C	ResNet34	✓		
D	IR_ResNet50			✓

3.2.4 | Hulking

As for this team, they propose a novel Pipeline-based CNN (namely PipeNet) fusion architecture which taking modified SENet-154 [39] as the backbone for multi-modal face anti-spoofing. Specifically, as shown in Figure 13, it contains two modules, namely selective modal pipeline (SMP) module and limited frame vote (LFV) module for the input of multiple modalities and sequence video frames, respectively. We can see that the framework contains three SMP modules, and each module takes a modal data (i.e., RGB, Depth, IR) as input. Taking the RGB modality as an example, they first use one frame as input and randomly crop it into patches, then send them to *ColorPipeline* which consists of data augmentation and feature extraction operations. They use a fusion strategy, which is concatenating the responses of *ColorPipeline*, *DepthPipeline* and *IRPipeline*, and sending them to *FusionMoudle* for further feature abstraction. After the linear connection, input all frame features of the video to the LFV module, and iteratively calculate the probability that each frame sample belongs to the real face. Finally, the output is a prediction for real face probability of the input face video.

3.2.5 | Newland_tianyan

For multi-modal track, this team uses two independent ResNet-9 [31] as backbones to learn the features of Depth

and IR modal data respectively. Similar to the single-modal track, the inputs of depth branch are subtracted from the neighbourhood mean before entering the network. In addition to data augment similar to the single-modal track, they transferred the RGB data of real samples to grey space and added light spots for data augment. Their code is publicly available in <https://github.com/Huangzebin99/CVPR-2020>.

3.2.6 | ZhangTT

A multi-stream CNN architecture called ID-Net is proposed for the multi-modal track. Since the different feature distributions of different modalities, the proposed model attempt to explore the interdependence between these modalities. As shown in Figure 14, there are two models trained by this team which one is trained using only IR as input and the other using both IR and Depth as inputs. Specially, a multi-stream architecture is designed with two sub-networks to perform multi-modal features fusion and the feature maps of two sub-networks are concatenated after a convolutional block. The final score is a weighted average of the results of two models. Their code is publicly available in <https://github.com/ZhangTT-race/CVPR2020-MultiModal>.

3.2.7 | Harvest

Different from other teams, they pay more attention to the network structure, this team mainly explores data pre-processing and data augmentation to improve the generalization performance. Through experimental comparison, they found that IR modal data are more suitable for face anti-spoofing task. Therefore, in this multi-modal track, only the IR modal data participate in model training. Similar to the team Super, they first use the face detector to remove the background area outside the face. Concretely, they use a face detector to detect face region of

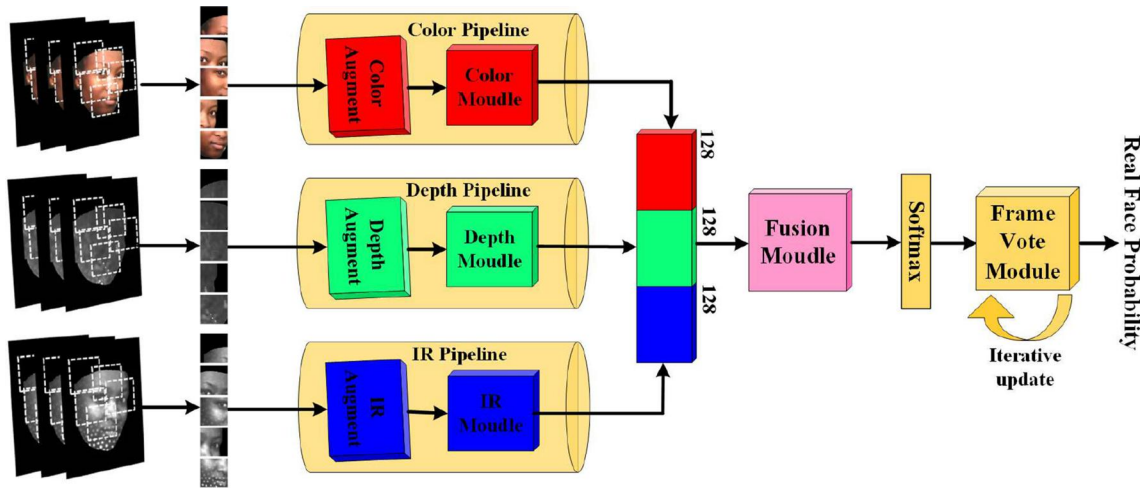


FIGURE 13 The overall architecture of PipeNet. The figure is provided by the Hukling team and ranked NO.3 in the multi-modal track

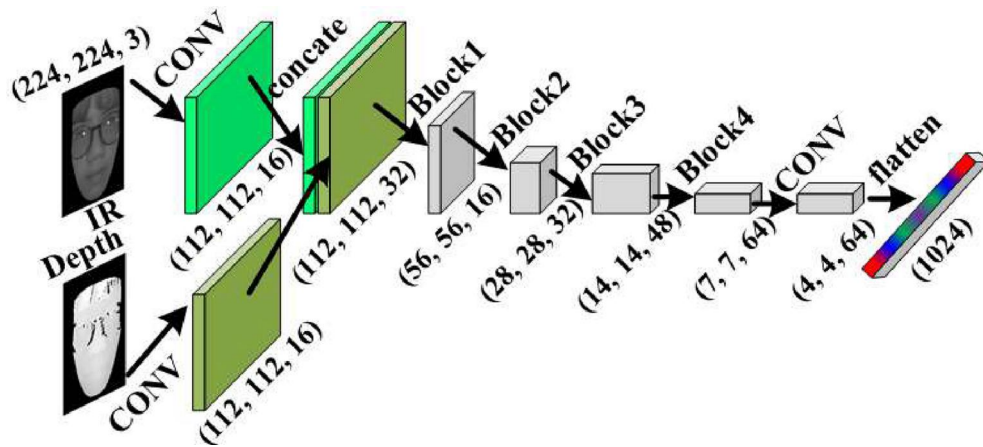


FIGURE 14 Architecture of the proposed for the multi-modal track. The figure is provided by the ZhangTT team and ranked NO.5 in the multi-modal track. IR, infrared

interest (ROI) with RGB data, and then mapping these ROIs to IR data to get the corresponding face position. Since only IR modal data are used, more sample augmentation strategies are used in network training to prevent overfitting. Such as the image is randomly divided into patches in an online manner before sending it to the network. Besides, they tried some tricks including triplet loss with semi-hard negative mining, sample interpolation augmentation and label smoothing.

3.2.8 | Qyxqyx

Based on the work in [15], this team adds an additional binary classification supervision to promote the performance for multi-modal track. Specifically, the network structure is from [15,41], and the additional binary supervision is inspired by [42]. As shown in Figure 15, taking the RGB modality as an

example, the input samples are supervised by two loss functions which are a binary classification loss and a regression loss after passing through the feature network. Finally, the weighted sum of the binary output and the pixel-wise regression output as the final score. Their code is publicly available in https://github.com/qyxqyx/FAS_Chalearn_challenge.

3.2.9 | Skjack

The network structure is similar to team Super. They use ResNet-9 [31] as the backbone and fuse the RGB, Depth and IR features after the res-3 block, then a 1×1 convolution operation is used to compress the channel. Since there are no additional novel innovations, the team's algorithm did not perform well in this competition. Their code is publicly available <https://github.com/skJack/challenge.git>.

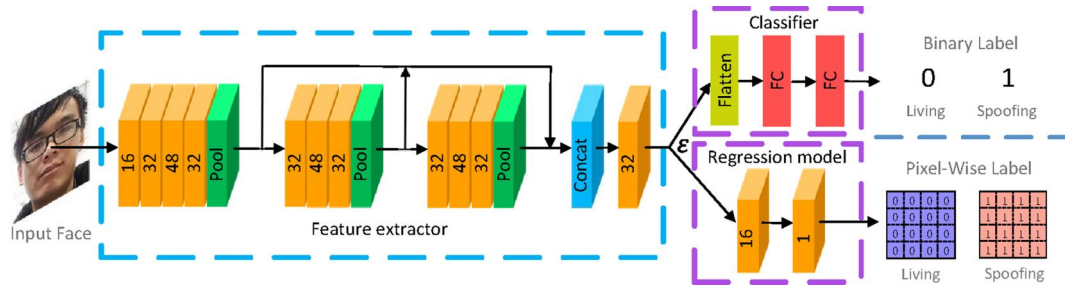


FIGURE 15 The supervision and the network of Qyxqyx team. The orange cube is convolution layer. The pixel-wise binary label in their experiment is resized into 32×32 resolution. The figure is provided by the Qyxqyx team and ranked NO.7 in the multi-modal track

4 | CHALLENGE RESULTS

In this section, we first report the results of the participating teams from the perspective of both single-modal and multi-modal tracks, and then analyse the performances of the participants' methods. Finally, the shortcomings and limitations of these algorithms are pointed out.

4.1 | Challenge results report

4.1.1 | Single-modal (RGB) track

Since the single-modal track only allows the use of RGB data, the purpose is to evaluate the performance of the algorithms on a face anti-spoofing system with a VIS camera as the acquisition device. The final results of the 11 participating teams are shown in Table 5, which includes the three considered indicators (e.g., APCER, BPCER and ACER) on three sub-protocols (e.g., 4-1, 4-2 and 4-3). The final ranking is based on the average value of the ACER on three sub-protocols (smaller means better performance). At the same time, we report the thresholds for all algorithms to make decisions on real faces and attack samples. The thresholds of the top three teams are either very large (i.e., more than 0.9 for BOBO) or very small (i.e., 0.01 for Harvest), or have very different thresholds for different sub-protocols (i.e., 0.02 vs. 0.9 for VisionLabs). In addition, VisionLabs achieves the best results on APCER with a value of 2.72%, meaning that the algorithm can better classify attack samples correctly. Whilst, Wgqtmac's algorithm obtains the best results on the indicator of BPCER (0.66%), indicating that it can better classify real face. Overall, the results of the first 10 teams are better than the baseline method [22] when ranking by ACER. The VisionLabs team achieved the first place with a clear advantage.

4.1.2 | Multi-modal

The multi-modal track allows the participating teams to use all the modal data. The purpose is to evaluate the performance of the algorithms on anti-spoofing systems equipped with

multi-optic cameras, such as the Intel RealSense or Microsoft Kinect sensor. The results of the eight participating teams in the final stage are shown in Table 6. BOBO team's algorithm gets first-place performance, such as APCER = 1.05%, BPCER = 1.00% and ACER = 1.02%. While the team of Super ranks second with a slight disadvantage, such as ACER = 1.68%. It is worth noting that Newland_tianyan's algorithm achieves the best results on the APCER indicator with a value of 0.24%. Similar to the conclusion of the single-modal track, most of the participating teams have relatively large thresholds which are calculated on the validation set, specially the Super and Newland_Tianyan teams with the value of 1.0 on three sub-protocols, indicating that these algorithms treat the face anti-spoofing task as an anomaly detection. In addition, we can find that the ACER values of the top four teams are 1.02%, 1.68%, 2.21%, and 2.28%, which are better than the ACER of the first place of the single-modal track, such as 2.72% for the team of VisionLabs. It shows the necessity of our multi-modal track in improving accuracy in face anti-spoofing task.

4.2 | Challenge results analysis

In this section, we analyse the advantages and disadvantages of the algorithm performance of each participating team in detail according to different tracks.

4.2.1 | Single-modal

As shown in Table 3, the testing subset introduces two unknown target variations simultaneously, such as the different ethnicities and attack types in training and testing subsets, which pose a huge challenge for participating teams. However, most teams achieved relatively good results in the final stage compared to baseline, specially the top three teams get ACER values below 10%. It is worth mentioning that different algorithms have their own unique advantages, even if the final ranking is relatively backward. Such as the value of BPCER of Wgqtmac'team is 0.66%, meaning about one real sample from 100 real faces will be treated as fake ones. While, APCER = 0.11% for the team of VisionLabs indicates about

TABLE 5 The results of single-modal track

Team Name	Method (keywords)	Prot.	Thre.	FP	FN	APCER(%)	BPCER(%)	ACER(%)	Rank
VisionLabs	OpticalFlow,	4_1	0.02	4	21	0.22	5.25	2.74	1
	RankPooling,	4_2	0.90	0	12	0.00	3.00	1.50	
	Data augment,	4_3	0.10	2	31	0.11	7.75	3.93	
	SimpleNet	Avg \pm Std	0.34 \pm 0.48	2 \pm 2	21 \pm 9	0.11 \pm 0.11	5.33 \pm 2.37	2.72 \pm 1.21	
BOBO	CDC, CDL, EDL,	4_1	0.95	201	10	11.17	2.5	6.83	2
	Multi-level cell,	4_2	0.99	120	8	6.67	2.0	4.33	
	Attention moudle,	4_3	0.99	67	12	3.72	3.0	3.36	
	Depth supervision	Avg \pm Std	0.97 \pm 0.02	129 \pm 67	10 \pm 2	7.18 \pm 3.74	2.50 \pm 0.50	4.84 \pm 1.79	
Harvest	Motion cues,	4_1	0.01	31	48	1.72	12.0	6.86	3
	Relabelling live,	4_2	0.01	116	51	6.44	12.75	9.6	
	Sequence,	4_3	0.01	109	67	6.06	16.75	11.4	
	ResNet	Avg \pm Std	0.01 \pm 0.00	85 \pm 47	55 \pm 10	4.74 \pm 2.62	13.83 \pm 2.55	9.28 \pm 2.28	
Zhang'IT	Quality tensor,	4_1	0.9	103	74	5.72	18.5	12.11	4
	Time tensor,	4_2	0.9	132	45	7.33	11.25	9.29	
	Data	4_3	0.9	57	108	3.17	27.0	15.08	
	Pre-processing	Avg \pm Std	0.9	97 \pm 37	75 \pm 31	5.40 \pm 2.10	18.91 \pm 7.88	12.16 \pm 2.89	
Newland_tianyan	Data augment,	4_1	0.77	34	117	1.89	29.25	15.57	5
	Temporal feature,	4_2	0.7	513	11	28.5	2.75	15.62	
	Neighbourhood	4_3	0.55	299	6	16.61	1.5	9.06	
	Mean	Avg \pm Std	0.67 \pm 0.11	282 \pm 239	44 \pm 62	15.66 \pm 13.33	11.16 \pm 15.67	13.41 \pm 3.77	
Dopamine	ID information,	4_1	0.02	325	6	18.06	1.5	9.78	6
	Multi-task,	4_2	0.22	367	24	20.39	6.0	13.19	
	Score fusion,	4_3	0.01	636	0	35.33	0.0	17.67	
	Resnet100	Avg \pm Std	0.07 \pm 0.11	442 \pm 168	10 \pm 12	24.59 \pm 9.37	2.50 \pm 3.12	13.54 \pm 3.95	
IecLab		4_1	0.33	696	21	38.67	5.25	21.96	7
	3D ResNet,	4_2	0.45	606	26	33.67	6.5	20.08	
	Fueature fusion,	4_3	0.45	489	26	27.17	6.5	16.83	
	Softmax	Avg \pm Std	0.40 \pm 0.07	597 \pm 103	24 \pm 2	33.16 \pm 5.76	6.08 \pm 0.72	19.62 \pm 2.59	
Chunghwa-Telecom	Subsequence	4_1	0.87	538	44	29.89	11.0	20.44	8
	Feature,	4_2	0.93	352	113	19.56	28.25	23.9	
	Local feature,	4_3	0.79	442	71	24.56	17.75	21.15	
	MIMAMO-Net	Avg \pm Std	0.86 \pm 0.06	444 \pm 93	76 \pm 34	24.66 \pm 5.16	19.00 \pm 8.69	21.83 \pm 1.82	
Wgqtmac		4_1	0.85	1098	1	61.0	0.25	30.62	9
	ResNet18,	4_2	1.0	570	7	31.67	1.75	16.71	
	Warmup strategy,	4_3	0.56	1117	0	62.06	0.0	31.03	
	Softmax	Avg \pm Std	0.80 \pm 0.22	928 \pm 310	2 \pm 3	51.57 \pm 17.24	0.66 \pm 0.94	26.12 \pm 8.15	

(Continues)

TABLE 5 (Continued)

Team Name	Method (keywords)	Prot.	Thre.	FP	FN	APCER(%)	BPCER(%)	ACER(%)	Rank
Hulking		4_1	0.81	635	138	35.28	34.5	34.89	10
	PipeNet,	4_2	0.82	1027	37	57.06	9.25	33.15	
	Softamx	4_3	0.67	768	59	42.67	14.75	28.71	
		Avg \pm Std	0.76 ± 0.08	810 ± 199	78 ± 53	45.00 ± 11.07	19.50 ± 13.27	32.25 ± 3.18	
Dqiu		4_1	1.0	1316	142	73.11	35.5	54.31	11
	ResNet50,	4_2	1.0	567	60	31.5	15.0	23.25	
	Softmax	4_3	1.0	664	146	36.89	36.5	36.69	
		Avg \pm Std	1.00 ± 0.00	849 ± 407	116 ± 48	47.16 ± 22.62	29.00 ± 12.13	38.08 ± 15.57	
Baseline	Static and	4_1	1.0	1331	7	73.94	1.75	37.85	*
	Dynamic features	4_2	1.0	1379	27	76.61	6.75	41.68	
	Features,	4_3	1.0	836	57	46.44	14.25	30.35	
	RankPooling	Avg \pm Std	1.00 ± 0.00	1182 ± 300	30 ± 25	65.66 ± 16.70	7.58 ± 6.29	36.62 ± 5.76	

Notes: Avg \pm Std indicates the mean and variance operation and best results are shown in bold. *, it means the proposed baseline method.

Abbreviations: ACER, average classification error rate; APCER, attack presentation classification error rate; CDC, central difference convolution; CDL, contrastive depth loss; EDL, Euclidean distance loss; FP, false positive; FN, false negative.

one fake sample from 1000 attackers will be treated as real ones.

To fully compare the stability of the participating team's algorithms, similar to [21], we introduce the receiver operating characteristic (ROC) curve in this challenge which can be used to select a suitable trade-off threshold between false-positive rate (FPR) and TPR according to the requirements of a given real application. As shown in Figure 16, the results of the top one team (VisionLabs) on both three sub-protocols are clearly superior to other teams, revealing that using optical flow method to convert RGB modal data to other sample spaces can effectively improve the generalization performance of the algorithm to deal with different unknown factors. However, the TPR value of the remaining teams decreased rapidly as the FPR reduced (e.g., $\text{TPR@FPR} = 10^{-3}$ values of these teams are almost zero). In addition, we can find that although the performance of ACER for Harvest team is worse than that of the BOBO team, the performance of the $\text{TPR@FPR} = 10^{-3}$ is significantly better than the BOBO team. It is mainly because the FP and false-negative (FN) samples of the Harvest team are relatively close (see from Table 5).

Finally, for the top three teams, we randomly selected some mismatched samples as shown in Figure 17. We can see that most of the FN samples of the VisionLabs team are real faces with large motion amplitude, while the most of FP samples are 3D print attacks, indicating that the team's algorithm has correctly classified almost all 2D attack samples. In addition, due to the challenging nature of our competition dataset, such as it is difficult to distinguish the real face from attack samples without the label, the BOBO team and the Harvest team did not make correct decisions on some difficult samples.

4.2.2 | Multi-modal

From the Table 6, we can find that the ACER values of the top seven teams are relatively close, and the top four teams are better than VisionLabs (ACER = 2.72%) in the single-modal track. It indicates that the complementary information between multi-modal datasets can improve the accuracy of the face anti-spoofing algorithm. Although Newland_Tianyan ranked fourth in ACER, they achieved the best results on the APCER indicator (e.g., APCER = 0.24%). It means the smallest number of FP samples among all teams. In addition, from the Table 6 and Figure 18, we can find that although the ACER values of the top two algorithms are relatively close, the stability of the Super team is better than the BOBO, such as the values of $\text{TPR@FPR} = 10^{-3}$ for Super and Newland_Tianyan are better than BOBO on both three sub-protocols. Finally, we can find from the Figure 19 that the FP samples of the top three teams contain many 3D print attacks, indicating that their algorithms are vulnerable to 3D face attacks.

5 | OPEN ISSUES AND OPPORTUNITIES

In this section, we will first summarize some common issues that were identified in this challenge, then analyse some of the corresponding causes, and describe some feasible solutions to alleviate these problems in combination with practical applications. Finally, we formulate the future work based on the CASIA-SURF CeFA dataset.

TABLE 6 The results of Multi-modal track. Avg \pm Std indicates the mean and variance operation and best results are shown in bold

Team Name	Method (keywords)	Prot.	Thre.	FP	FN	APCER(%)	BPCER(%)	ACER(%)	Rank
BOBO	CDC, CDL, EDL,	4.1	0.98	6	2	0.33	0.5	0.42	1
	Feature fusion	4.2	0.95	25	3	1.39	0.75	1.07	
	Score fusion,	4.3	0.94	26	7	1.44	1.75	1.6	
	Depth supervision	Avg \pm Std	0.95 \pm 0.02	19 \pm 11	4 \pm 2	1.05 \pm 0.62	1.00 \pm 0.66	1.02 \pm 0.59	
Super	Data pre-processing,	4.1	1.0	9	11	0.5	2.75	1.62	2
	Dimension reduction,	4.2	1.0	5	17	0.28	4.25	2.26	
	SE fusion	4.3	1.0	20	5	1.11	1.25	1.18	
	Score fusion	Avg \pm Std	1.0 \pm 0.00	11.33 \pm 7.76	11 \pm 6	0.62 \pm 0.43	2.75 \pm 1.50	1.68 \pm 0.54	
Hulking	PipeNet,	4.1	0.96	31	0	1.72	0.0	0.86	3
	SENet-154,	4.2	1.0	99	5	5.5	1.25	3.37	
	Selective modal Pipeline	4.3	1.0	46	9	2.56	2.25	2.4	
	Limited frame Vote	Avg \pm Std	0.98 \pm 0.02	58 \pm 35	4 \pm 4	3.25 \pm 1.98	1.16 \pm 1.12	2.21 \pm 1.26	
Newland_tianyan	Resnet9,	4.1	1.0	0	3	0.0	0.75	0.37	4
	Data pre-processing	4.2	1.0	4	26	0.22	6.5	3.36	
	Neighbourhood mean	4.3	1.0	9	23	0.5	5.75	3.12	
	Data augmen	Avg \pm Std	1.00 \pm 0.00	44	17 \pm 12	0.24 \pm 0.25	4.33 \pm 3.12	2.28 \pm 1.66	
Zhang'IT	ID Net,	4.1	0.94	0	19	0.0	4.75	2.37	5
	Feature fusion	4.2	0.9	66	34	3.67	8.5	6.08	
	Score fusion	4.3	0.79	102	0	5.67	0.0	2.83	
		Avg \pm Std	0.87 \pm 0.07	56 \pm 51	17 \pm 17	3.11 \pm 2.87	4.41 \pm 4.25	3.76 \pm 2.02	
Harvest	Data pre-processing,	4.1	0.87	13	4	0.72	1.0	0.86	6
	Data augment,	4.2	0.93	180	28	10.0	7.0	8.5	
	Only IR	4.3	0.96	119	8	6.61	2.0	4.31	
	Semi-hard negative mining	Avg \pm Std	0.92 \pm 0.04	104 \pm 84	13 \pm 12	5.77 \pm 4.69	3.33 \pm 3.21	4.55 \pm 3.82	
Qyxqyx	Binary supervision	4.1	0.98	1	53	0.06	13.25	6.65	7
	Pixel-wise regression,	4.2	0.98	19	8	1.06	2.0	1.53	
	Score fusion		0.89	257	19	14.28	4.75	9.51	
		Avg \pm Std	0.95 \pm 0.05	92 \pm 142	26 \pm 23	5.12 \pm 7.93	6.66 \pm 5.86	5.89 \pm 4.04	
Skjack	Resnet9	4.1	0.0	1371	2	76.17	0.5	38.33	8
		4.2	0.01	1155	46	64.17	11.5	37.83	
	Softmax	4.3	0.0	511	93	28.39	23.25	25.82	
		Avg \pm Std	0.00 \pm 0.00	1012 \pm 447	47 \pm 45	56.24 \pm 24.85	11.75 \pm 11.37	33.99 \pm 7.08	
Baseline	SD-Net,	4.1	1.0	413	109	22.94	27.25	25.1	*
	A shared branch,	4.2	0.17	1340	23	74.44	5.75	40.1	
	PSMM-Net	4.3	0.02	864	55	48.0	13.75	30.87	
	Fusion	Avg \pm Std	0.39 \pm 0.52	872 \pm 463	62 \pm 43	48.46 \pm 25.75	15.58 \pm 10.86	32.02 \pm 7.56	

Notes: *, it means the proposed baseline method. Bold value means the best performance under a specific evaluate metric (i.e., ACER).

Abbreviations: ACER, average classification error rate; APCER, attack presentation classification error rate; CDC, central difference convolution; CDL, contrastive depth loss; EDL, Euclidean distance loss; FP, false positive; FN, false negative.

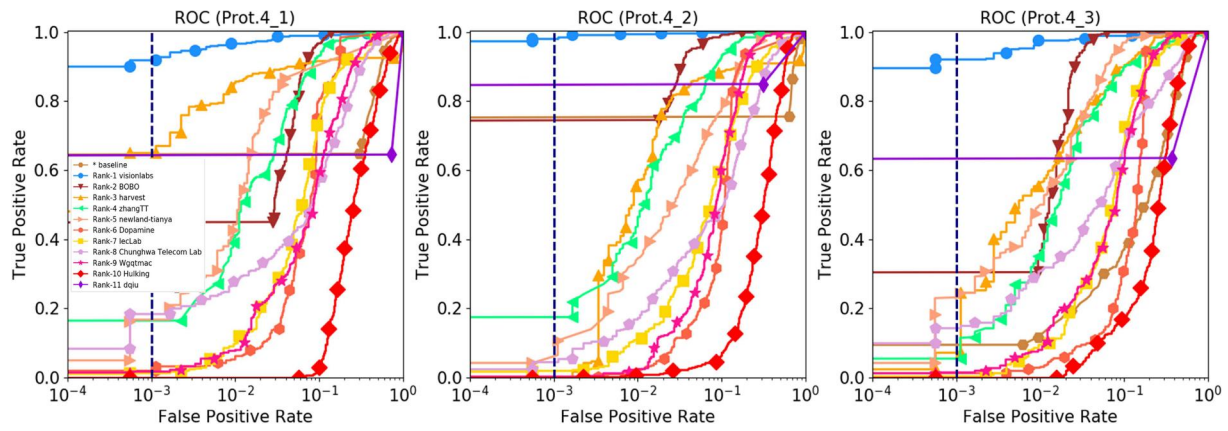


FIGURE 16 The ROC of 12 teams in a single-modal track. From left to right are the receiver operating characteristics on protocol 4_1, 4_2 and 4_3

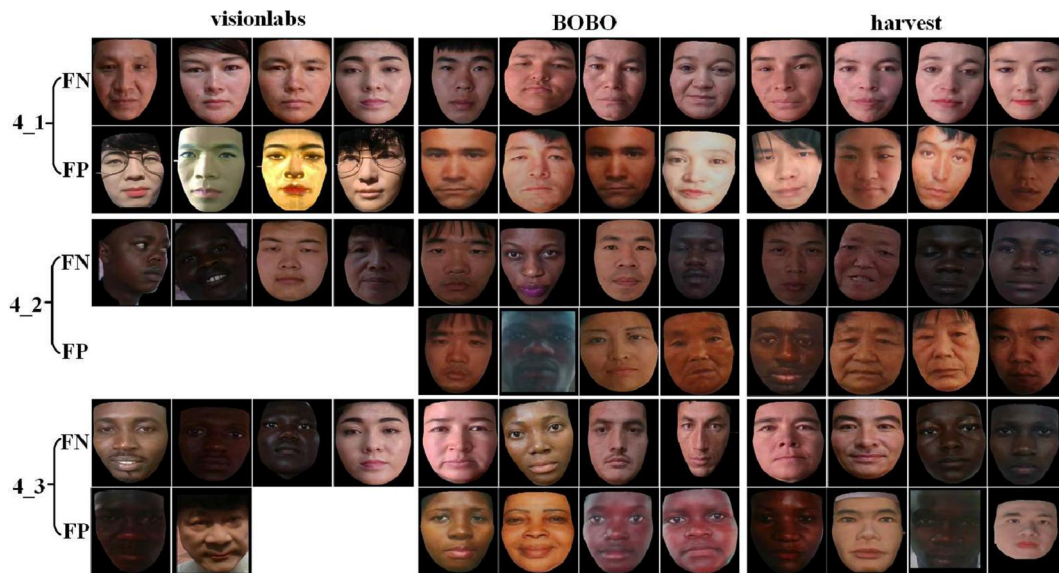


FIGURE 17 The mismatched samples of the top three teams in the single-modal track. FN and FP indicate false negative and false positive, respectively

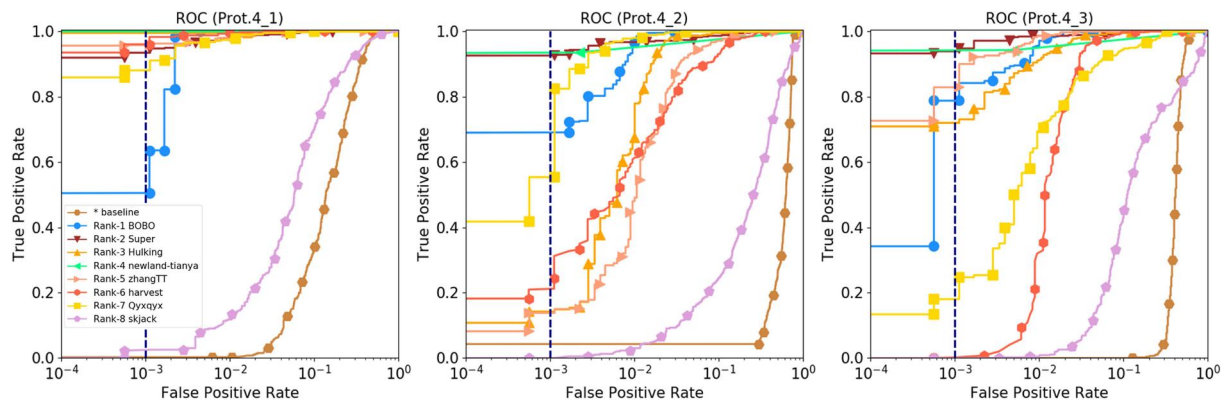


FIGURE 18 The ROC of nine teams in the multi-modal track. From left to right are the receiver operating characteristics on protocol 4_1, 4_2 and 4_3

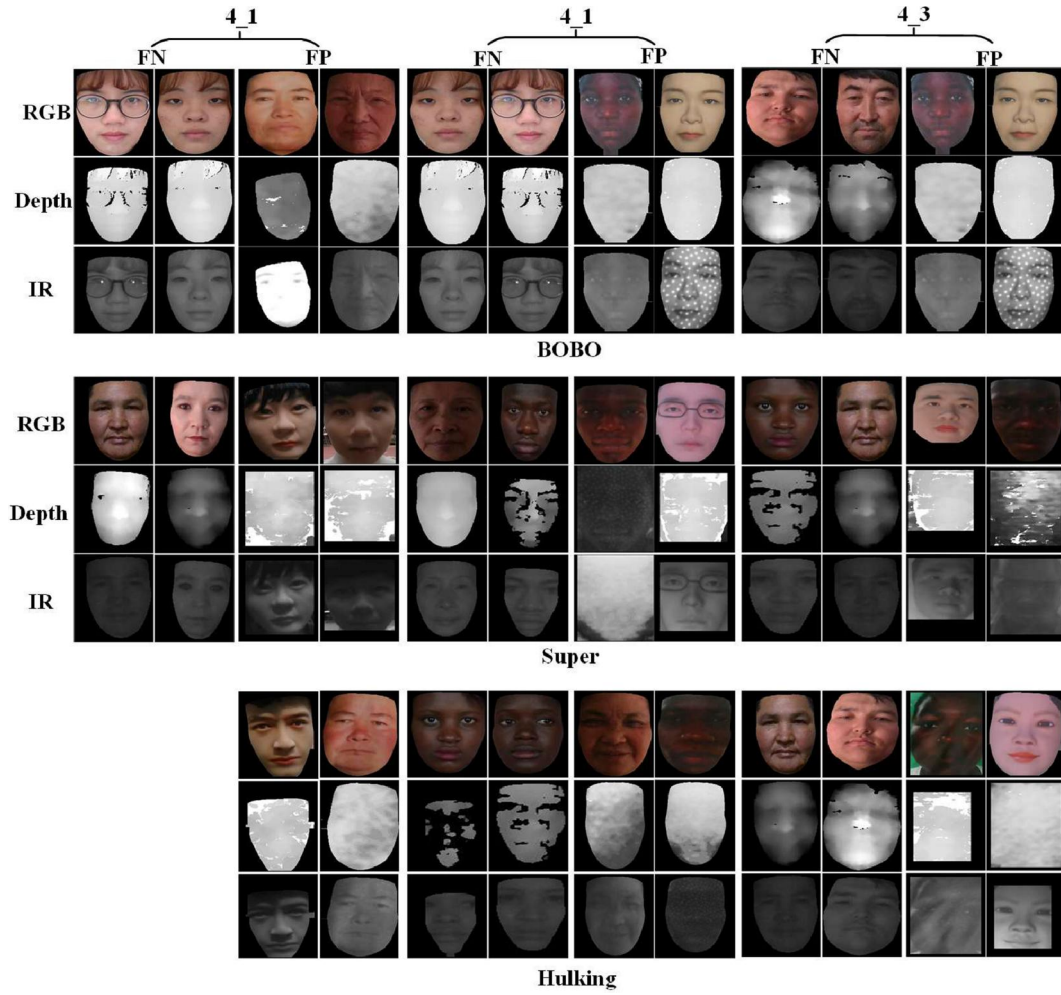


FIGURE 19 The mismatched samples of the top three teams in the multi-modal track. FN and FP indicate false negative and false positive, respectively

5.1 | Critical issues and breakthrough point

From Tables 5 and 6 of the competition results, we can find that the threshold for both single-modal and the multi-modal track is generally high. The meaning of the threshold in our challenge is the minimum probability that a sample will be classified as a real face. For instance, the thresholds on three sub-protocols reach to one for the team of dqu in single-modal track and the top-ranked teams (Super, Huking and Newland_Tianyan) in the multi-modal track. These over-confidence problems mean that some attack samples will be judged as real faces with high probability, which is unreasonable in practical applications. We analyse the following three reasons responsible for this problem: (1) caused by the task itself. The nature of the face anti-spoofing task is a binary classification task. If the sample scale is small and lacks diversity, it can easily lead to extreme thresholds. This phenomenon is also found in other binary classification tasks, such as face detection, (2) caused by different collection environments for positive (real face) and negative samples (spoof). For example, the attack samples of the same subject

are collected under multiple lighting conditions, while the real face is collected only in indoor environments, and (3) caused by the lack of generalization performance when the algorithm faces unknown attack types and ethnicities. According to the characteristics of the testing protocol that contains two unknown variables (i.e., cross-PAIs and cross-ethnicity) in training and testing phases, some teams design networks and loss functions pay more attention to the motion information of real face and replay attack in the training phase, and treat any unseen static-samples (including spoofs and real faces) in the testing phase as abnormal information (spoofs), resulting in poor generalization ability in cross-PAIs. Other teams have subtracted different neighbourhood mean values according to different ethnicities to alleviate the interference caused by skin colour differences. However, in the face of unknown ethnic samples, the inability to subtract the appropriate neighbourhood mean causes classification errors. In summary, poor generalization performance (i.e., unable to correctly classify unknown real samples and attack types) causes the classification threshold to be too large or too small. To alleviate this problem, we propose feasible solutions from the

three aspects of data collection, training strategy, and algorithm design. CASIA-SURF CeFA is the largest up to date face anti-spoofing dataset and contains various attack types and attack environments, such as the attack types include print attacks and replay attacks under multiple lighting conditions. However, the diversity of the device and environment for collecting real face samples is limited.

It inevitably brings the problem of sample imbalance. Therefore, the CASIA-SURF CeFA dataset should consider supplementing some real samples including acquisition equipment and shooting environment. Whilst, an effective training strategy is to balance the positive and negative proportions of samples in each batch during the training process. Finally, a binary cross-entropy loss might discover arbitrary cues, such as spot or screen bezel of the spoof medium, that are not the faithful spoof patterns. 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.

5.2 | Future work and opportunities

Face anti-spoofing based on multi-modal datasets attracts increasing research interests. However, the gap exploration between sensing patterns of different face modalities remains an open research problem in face anti-spoofing. Some previous works [22,43] have been verified the existence of performance deviations of the SOTA algorithms in different face modalities. At the same time, they designed a testing protocol to measure the degree of modal bias, such as the Protocol three in CASIA-SURF CeFA [22]. Similar to heterogeneous face recognition (e.g., NIR-VIS [44–46]), which refers to matching faces across different modalities (or sensing patterns), we cast the face anti-spoofing task as a heterogeneous face matching problem. In this way, the discrimination information of other modal samples can be used to assist the learning of RGB modal data. And after the model is trained, there is no need to load other modal samples during the testing phase.

Since the existing datasets for training and verification are collected in VIS spectrum, the use of samples of additional modalities (e.g., Depth or IR) to assist the learning of RGB modal data while without extra modalities in testing phase is interesting in the practical applications. On the other hand, CASIA-SURF [21] and CASIA-SURF CeFA [22] are multi-modal face anti-spoofing datasets and each sample contains three paired modalities, which may provide us with the possibility to study heterogeneous face anti-spoofing.

6 | CONCLUSION

We organized the *Chalearn Face Anti-spoofing Attack Detection Challenge at CVPR2020* based on the CASIA-SURF CeFA dataset with two tracks and running on the CodaLab platform. Both tracks attracted 340 teams in the development stage, and finally, 11 and eight teams have submitted their

codes in the single-modal and multi-modal face anti-spoofing recognition challenges, respectively. We described the associated dataset, and the challenge protocol including evaluation metrics. We reviewed in detail the proposed solutions and reported the challenge results. Compared with the baseline method, the best performances from participants under the ACER value are from 36.62 to 2.72, and 32.02 to 1.02 for the single-modal and multi-modal challenges, respectively. We analysed the results of the challenge, pointing out the critical issues in PAD task and presenting the shortcomings of the existing algorithms. Future lines of research in the field have been also discussed.

ACKNOWLEDGEMENTS

This work was supported by the Chinese National Natural Science Foundation Projects #61961160704, #61876179, Science and Technology Development Fund of Macau (No. 0025/2018/A1, 0008/2019/A1, 0019/2018/ASC, 0010/2019/AFJ, 0025/2019/AKP), the Key Project of the General Logistics Department Grant No. ASW17C001, the Spanish project PID2019-105093GB-I00 (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya, and by ICREA under the ICREA Academia programme, Science and Technology Development Fund of Macau (No. 0025/2018/A1, 0008/2019/A1, 0019/2018/ASC, 0010/2019/AFJ, 0025/2019/AKP). We acknowledge Surfing Technology Beijing co., Ltd (www.surfing.ai) to provide us this high-quality dataset. Finally, we thank all participating teams for their participation and contributions, and special thanks to VisionLabs, BOBO, Harvset, ZhangTT, Newland_tianyan, Dopamine, Hulating, Super, Qyxqyx for their guidance in drawing figure. Hugo Jair Escalante was supported by CONACyT under project grant CB-2017-2018 A1-S-26314.

ORCID

Jun Wan  <https://orcid.org/0000-0002-4735-2885>

REFERENCES

1. Pan, G., et al.: Eyeblick-based anti-spoofing in face recognition from a generic webcam. Paper presented at International Conference on Computer Vision, 2007
2. Wang, L., Ding, X., Fang, C.: Face live detection method based on physiological motion analysis. Tsinghua. Sci. Technol. (2009)
3. Kollreider, K., Fronthaler, H., Bigun, J.: Verifying liveness by multiple experts in face biometrics. Paper presented at Conference on Computer Vision and Pattern Recognition Workshops, 2008
4. Bharadwaj, S., et al.: Computationally efficient face spoofing detection with motion magnification. Paper presented at Conference on Computer Vision and Pattern Recognition, 2013
5. Pan, G., et al.: Monocular camera-based face liveness detection by combining eyeblink and scene context. Tsinghua. Sci. Technol. 47, 215–225 (2011)
6. Komulainen, J., et al.: Context based face anti-spoofing. Paper presented at Conference on Biometrics: Theory, Applications and Systems, 2013
7. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: Proceedings of the International Conference of Biometrics Special Interest Group (2012)
8. Yang, J., et al.: Face liveness detection with component dependent descriptor. Paper presented at International Conference on Biometrics, 2013

9. Maatta, J., Hadid, A., Pietikainen, M.: Face spoofing detection from single images using texture and local shape analysis. *IET Biom.* 1(1), 3–10 (2012)
10. Schwartz, W.R., Rocha, A., Pedrini, H.: Face spoofing detection through partial least squares and low-level descriptors. Paper presented at International Joint Conference on Biometrics, 2011
11. Feng, L., et al.: Integration of image quality and motion cues for face anti-spoofing: neural network approach. *J. Vis. Commun. Image Represent.* (2016)
12. Li, L., et al.: An original face anti-spoofing approach using partial convolutional neural network. Paper presented at Image Processing Theory Tools and Applications 6th International Conference on, IEEE, 2016
13. Patel, K., Han, H., Jain, A.K.: Secure face unlock: spoof detection on smartphones. *IEEE Trans. Inf. Forensics Secur.* (2016)
14. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. *arXiv* (2014)
15. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: binary or auxiliary supervision. Paper presented at Conference on Computer Vision and Pattern Recognition, 2018
16. Jourabloo, A., Liu, Y., Liu, X.: Face de-spoofing: NTI-spoofing via noise modeling. *arXiv* (2018)
17. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. Paper presented at Conference on Neural Information Processing Systems, 2012
18. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. Paper presented at International conference of the Biometrics Special Interest Group, 2012
19. Zhang, Z., et al.: A face antispoofing database with diverse attacks. Paper presented at International Conference on Biometrics, 2012
20. Boulkenafet, Z., et al.: OULU-NPU: mobile face presentation attack database with real-world variations. Paper presented at International Conference on Automatic Face and Gesture Recognition, 2017
21. Zhang, S., et al.: A dataset and benchmark for large-scale multi-modal face anti-spoofing. Paper presented at Conference on Computer Vision and Pattern Recognition, 2019
22. Liu, A., et al.: CASIA-SURF CeFA: Benchmark for Multi-modal Cross-ethnicity Face Anti-spoofing, Paper presented at Winter Conference on Applications of Computer Vision, (2021)
23. Are Face Recognition Systems Accurate? Depends on Your Race (2016) <https://www.technologyreview.com/s/601786>
24. Wang, M., et al.: Racial faces in the wild: reducing racial bias by information maximization adaptation network. Paper presented at International Conference on Computer Vision, October 2019
25. Chakka, M.M., et al.: Competition on counter measures to 2-d facial spoofing attacks. Paper presented at IEEE International Joint Conference on Biometrics, 2011
26. Chingovska, I., et al.: The 2nd competition on counter measures to 2d face spoofing attacks. Paper presented at 2013 International Conference on Biometrics (ICB), pp. 1–6. IEEE (2013)
27. Boulkenafet, Z., et al.: A competition on generalized software-based face presentation attack detection in mobile scenarios. Paper presented at IEEE International Joint Conference on Biometrics (IJCB), pp. 688–696. IEEE (2017)
28. Ajan Liu, J.W., et al.: Multi-modal face anti-spoofing attack detection challenge at CVPR2019. Paper presented at Conference on Computer Vision and Pattern Recognition Workshop, 2019
29. Yu, Z., et al.: Searching central difference convolutional networks for face anti-spoofing. Paper presented at Conference on Computer Vision and Pattern Recognition, 2020
30. Wang, Z., et al.: Deep spatial gradient and temporal depth learning for face anti-spoofing. Paper presented at Conference on Computer Vision and Pattern Recognition, 2020
31. He, K., et al.: Deep residual learning for image recognition. Paper presented at Conference on Computer Vision and Pattern Recognition, 2016
32. Wang, J., Cherian, A., Porikli, F.: Ordered pooling of optical flow sequences for action recognition. Paper presented at Winter Conference on Applications of Computer Vision, pp. 168–76. IEEE (2017)
33. Fernando, B., et al.: Rank pooling for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(4), 773–87 (2017)
34. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face spoofing detection using colour texture analysis. *Trans. Inf. Forensics. Secur.* 11(8), 1818–1830 (2016)
35. Li, J., Wang, Y., Tan, T., et al.: Live face detection based on the analysis of Fourier spectra. Paper presented at Conference on Biometric Technology for Human Identification, 2004
36. Patel, K., Han, H., Jain, A.K.: Cross-database face antispoofing with robust feature representation. Paper presented at Chinese Conference on Biometric Recognition, 2016
37. Hara, K., Kataoka, H., Satoh, Y.: Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and Imagenet?, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 6546–6555. (2018)
38. Shen, T., Huang, Y., Tong, Z.: FaceBagNet: Bag-of-local-features model for multi-modal face anti-spoofing, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019
39. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. Paper presented at Conference on Computer Vision and Pattern Recognition, 2018
40. Parkin, A., Grinchuk, O.: Recognizing multi-modal face spoofing with face recognition networks. Paper presented at Conference on Computer Vision and Pattern Recognition Workshops, 2019
41. Qin, Y., et al.: Learning Meta Model for Zero- and Few-Shot Face Anti-spoofing. Association for Advancement of Artificial Intelligence (AAAI) (2020)
42. George, A., Marcel, S.: Deep Pixel-Wise Binary Supervision for Face Presentation Attack Detection. *CoRR.* abs/1907.04047 (2019) <http://arxiv.org/abs/1907.04047>
43. Zhang, S., et al.: CASIA-SURF: large-scale multi-modal benchmark for face anti-spoofing. *arXiv.* 1908.10654 (2019)
44. Yi, D., et al.: Face Matching Between Near Infrared and Visible Light Images, 523–530. Paper presented at International Conference on Biometrics, (2007)
45. Lei, Z., Li, S.: Coupled spectral regression for matching heterogeneous faces. Paper presented at Conference on Computer Vision and Pattern Recognition, pp. 1123–1128, (2009)
46. Lezama, J., Qiu, Q., Sapiro, G.: Not Afraid of the Dark: NIR-VIS Face Recognition via Cross-Spectral Hallucination and Low-Rank Embedding (2016)

How to cite this article: Liu A, Li X, Wan J, et al. Cross-ethnicity face anti-spoofing recognition challenge: A review. *IET Biome.* 2021;10:24–43. <https://doi.org/10.1049/bme2.12002>