CASIA-SURF: A Large-Scale Multi-Modal Benchmark for Face Anti-Spoofing

Shifeng Zhang, Ajian Liu, Jun Wan, *Senior Member, IEEE*, Yanyan Liang, *Senior Member, IEEE*, Guodong Guo, *Senior Member, IEEE*, Sergio Escalera, Hugo Jair Escalante, and Stan Z. Li, *Fellow, IEEE*

***Abstract*—Face anti-spoofing is essential to prevent face recognition systems from a security breach. Much of the progresses have been made by the availability of face anti-spoofing benchmark datasets in recent years. However, existing face anti-spoofing benchmarks have limited number of subjects (≤170) and modalities (≤2), which hinder the further development of the academic community. To facilitate face anti-spoofing research, we introduce a large-scale multi-modal dataset, namely CASIA-SURF, which is the largest publicly available dataset for face anti-spoofing in terms of both subjects and modalities. Specifically, it consists of 1, 000 subjects with 21, 000 videos and each sample has 3 modalities (*i.e.*, RGB, Depth and IR). We also provide comprehensive evaluation metrics, diverse evaluation protocols, training/validation/testing subsets and a measurement tool, developing a new benchmark for face anti-spoofing. Moreover, we present a novel multi-modal multi-scale fusion method as a strong baseline, which performs feature re-weighting to select the more informative channel features while suppressing the less useful ones for each modality across different scales. Extensive experiments have been conducted on the proposed dataset to verify its significance and generalization capability. The dataset is available at https://sites.google.com/qq.com/face-anti-spoofing/welcome/challengecvpr2019?authuser=0.**

***Index Terms*—Face anti-spoofing, large-scale, multi-modal, dataset, benchmark.**

I. Introduction

FACE anti-spoofing aims to determine whether the captured face from a face recognition system is real or fake. With the development of deep Convolutional Neural Networks (CNNs), face recognition [1]–[8] has achieved near-perfect recognition performance and already has been applied in our daily life, such as phone unlock, access control and face payment. However, these face recognition systems are prone to be attacked in various ways including print attack, video replay attack and 2D/3D mask attack, causing the recognition result to become unreliable. Therefore, face Presentation Attack Detection (PAD) [9], [10] is a vital step to ensure that face recognition systems are in a safe reliable condition.

In recent years, face PAD algorithms [26], [28] have achieved great performances. One of the key points of this success is the availability of face anti-spoofing benchmark datasets [11], [12], [16], [17], [25], [26]. However, there are several shortcomings in the existing datasets as follows:

- **Number of subjects is limited.** Compared to the large existing image classification [29] and face recognition [30] datasets, face anti-spoofing datasets have less than 170 subjects and 60, 00 video clips as shown in Table I. The limited number of subjects is not representative of the requirements of real applications.
- **Number of modalities is limited.** As shown in Table I, most of the existing datasets only consider a single modality (*e.g.*, RGB). For these existing available multi-modal datasets [13], [19], they are very scarce including no more than 21 subjects.
- **Evaluation metrics are not comprehensive enough.** How to compute the performance of algorithms is an open issue in face anti-spoofing. Many works [17], [25], [26], [28] adopt the Attack Presentation Classification Error Rate (APCER), the Normal Presentation Classification Error Rate (NPCER) and the Average Classification Error Rate (ACER) as the evaluation metric, in which APCER and NPCER are used to measure the error rate of fake or live samples, and ACER is the average of APCER and NPCER scores. However, in real applications, one may be more concerned about the false positive rate, *i.e.*, attacker is

Manuscript received September 27, 2019; revised December 16, 2019 and February 3, 2020; accepted February 3, 2020. Date of publication February 12, 2020; date of current version March 30, 2020. This work was supported in part by the Chinese National Natural Science Foundation Projects under Grant 61961160704, Grant 61876179, and Grant 61872367, in part by the Science and Technology Development Fund of Macau under Grant 0008/2018/A1, Grant 0025/2019/A1, Grant 0019/2018/ASC, Grant 0010/2019/AFJ, and Grant 0025/2019/AKP, in part by the Spanish Project under Grant TIN2016-74946-P (MINECO/FEDER, UE), in part by the CERCA Programme/Generalitat de Catalunya, and in part by the ICREA under the ICREA Academia Programme. This article was recommended for publication by Associate Editor T. Hassner upon evaluation of the reviewers' comments. *(Shifeng Zhang and Ajian Liu contributed equally to this work.) (Corresponding author: Jun Wan.)*

Shifeng Zhang and Jun Wan are with the National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: shifeng.zhang@nlpr.ia.ac.cn; jun.wan@nlpr.ia.ac.cn).

Ajian Liu and Yanyan Liang are with the Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China (e-mail: ajianliu92@gmail.com; yyliang@must.edu.mo).

Guodong Guo is with the Institute of Deep Learning, Baidu Research and National Engineering Laboratory for Deep Learning Technology and Application, Beijing 100085, China (e-mail: guoguodong01@baidu.com).

Sergio Escalera is with the Computer Vision Center, Universitat de Barcelona, 08007 Barcelona, Spain (e-mail: sergio@maia.ub.es).

Hugo Jair Escalante is with the Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico, and also with the Computer Science Department, CINVESTAV-Zacatenco, Mexico City 07360, Mexico (e-mail: hugojair@inaoep.mx).

Stan Z. Li is with Westlake University, Hangzhou 310024, China, and also with the Institute of Automation Chinese Academy of Sciences, Beijing 100190, China.

Digital Object Identifier 10.1109/TBIOM.2020.2973001

TABLE I
COMPARISON OF THE PUBLIC FACE ANTI-SPOOFING DATASETS (∗ INDICATES THIS DATASET ONLY CONTAINS IMAGES, NOT VIDEO CLIPS, ⋆ IS SHORT FOR SEEK THERMAL COMPACT PRO SENSOR, – INDICATES THAT THIS TEAM IS NOT COUNTED)

Dataset	Year	# subjects	# videos	Camera	Modal types	Spoof attacks
Replay-Attack [11]	2012	50	1,200	VIS	RGB	Print, 2 Replay
CASIA-MFSD [12]	2012	50	600	VIS	RGB	Print, Replay
3DMAD [13]	2013	17	255	VIS/Kinect	RGB/Depth	3D Mask
I^{2}BVSD [14]	2013	75	681∗	VIS/Thermal	RGB/Heat	3D Mask
GUC-LiFFAD [15]	2015	80	4,826	LFC	LFI	2 Print, Replay
MSU-MFSD [16]	2015	35	440	Phone/Laptop	RGB	Print, 2 Replay
Replay-Mobile [17]	2016	40	1,030	VIS	RGB	Print, Replay
3D Mask [18]	2016	12	1,008	VIS	RGB	3D Mask
Msspoof [19]	2016	21	4,704∗	VIS/NIR	RGB/IR	Print
SWIR [20]	2016	5	141∗	VIS/M-SWIR	RGB/4 SWIR bands	Print, 3D Mask
BRSU [21]	2016	50+	–	VIS/AM-SWIR	RGB/4 SWIR bands	Print, 3D Mask
EMSPAD [22]	2017	50	14,000∗	SpectraCam^{TM}	7 bands	2 Print
SMAD [23]	2017	–	130	VIS	RGB	3D Mask
MLFP [24]	2017	10	1,350	VIS/NIR/Thermal	RGB/IR/Heat	2D/3D Mask
Oulu-NPU [25]	2017	55	5,940	VIS	RGB	2 Print, 2 Replay
SiW [26]	2018	165	4,620	VIS	RGB	2 Print, 4 Replay
WMCA [27]	2019	72	6,716	RealSense/STC-PRO⋆	RGB/Depth/IR/Thermal	2 Print, Replay, 2D/3D Mask
CASIA-SURF	2018	**1,000**	**21,000**	RealSense	RGB/Depth/IR	Print, Cut

treated as real/live one. These aforementioned metrics can not meet this need.

- **Evaluation protocols are not diverse enough.** All the existing face anti-spoofing datasets only provide within-modal evaluation protocols. To be more specific, algorithms trained in a certain modality can only be evaluated in the same modality, which limits the diversity of face anti-spoofing research.

To deal with these aforementioned drawbacks, we introduce a large-scale multi-modal face anti-spoofing dataset, namely CASIA-SURF, which consists of $1,000$ subjects and $21,000$ video clips with 3 modalities (RGB, Depth, IR). It has 6 types of photo attacks combined by multiple operations, *e.g.*, cropping, bending the print paper and stand-off distance. Some samples and other detailed information of our dataset are shown in Fig. 1 and Table I. Comparing to these existing face anti-spoofing datasets, the proposed dataset has four main advantages as follows:

- **The most subjects.** The proposed dataset is the largest one in term of number of subjects, which is more than 6× boosted compared with previous challenging face anti-spoofing dataset like Spoof in the Wild (SiW) [26].
- **The most modalities.** Our CASIA-SURF is the only dataset that provides three modalities (*i.e.*, RGB, Depth and IR), and the other datasets have up to two modalities.
- **The most comprehensive evaluation metrics.** Inspired by face recognition [31], [32], we introduce the Receiver Operating Characteristic (ROC) curve for our large-scale face anti-spoofing dataset in addition to the commonly used evaluation metrics. The ROC curve can be used to select a suitable trade off threshold between the False Positive Rate (FPR) and the True Positive Rate (TPR) according to the requirements of a given real application.
- **The most diverse evaluation protocols.** In addition to the within-modal evaluation protocols, we also provide the cross-modal evaluation protocols in our dataset, in which algorithms trained in one modality will be evaluated in other modalities. It allows the academic community to explore new issues.

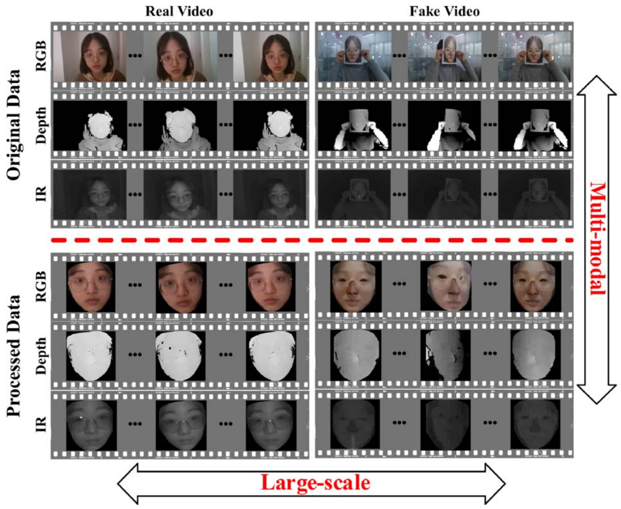


Fig. 1. The CASIA-SURF dataset. It is a large-scale and multi-modal dataset for face anti-spoofing, consisting of $492,522$ images with 3 modalities (*i.e.*, RGB, Depth and IR).

Besides, we present a novel multi-modal multi-scale fusion method as a strong baseline to conduct extensive experiments on the proposed dataset. Our new fusion method performs feature re-weighting to select the more informative channel features while suppressing the less useful ones for each modality across different scales. To sum up, the contributions of this paper are three-fold:

- Presenting a large-scale multi-modal face anti-spoofing dataset with $1,000$ subjects and 3 modalities.
- Introducing a new multi-modal multi-scale fusion method to effectively merge the involved three modalities across different scales.
- Conducting extensive experiments on the proposed CASIA-SURF dataset to verify its significance and generalization capability.

Preliminary results of this work have been published in [33]. The current work has been improved and extended from the conference version in several important aspects. (1) We provide the cross-modal evaluation protocols in our dataset for the academic community to explore new issues. (2) We improve the multi-modal fusion method in our previous work from one scale to multiple scales for better performance. (3) Some additional experiments are conducted and we noticeably improve the accuracy of the baseline in our previous work. (4) All sections are rewritten with more details, more references and more analysis to have a more elaborate presentation.

II. Related Work

Face anti-spoofing has made great progress with the proposal of new datasets in recent years. This section first summarizes the existing face anti-spoofing datasets and then reviews some representative methods

A. Dataset

Most of existing face anti-spoofing datasets only contain the RGB modality, including Replay-Attack [11], CASIA-FASD [12] and SiW [26]. With the popularity of face recognition in mobile phones, there are also some RGB datasets recorded by replaying face video with smartphone, such as MSU-MFSD [16], Replay-Mobile [17] and OULU-NPU [25].

As attack techniques are constantly upgraded, some new types of attacks have emerged, *e.g.*, 3D [13] and silicone masks [34]. These attacks are more realistic than traditional 2D attacks and the drawbacks of visible cameras are revealed. Fortunately, some new sensors have been introduced to provide more possibilities for face PAD methods, such as depth, muti-spectral and infrared light cameras. Kim *et al.* [35] introduce a new dataset to distinguish between facial skin and mask materials by exploiting their reflectance. Kose and Dugelay [36] propose a 2D+3D face mask attack dataset to study the effects of mask attacks. 3DMAD [13] is recorded using Microsoft Kinect sensor and consists of Depth and RGB modalities with 3D masks. Another multi-modal dataset is Msspoof [19], containing visible and near-infrared images of real accesses and printed spoofing attacks with ≤ 21 objects.

However, existing face PAD datasets have two main limitations: 1) They have limited number of subjects and samples, resulting in a potential over-fitting risk; 2) Most of existing datasets only include the RGB modality, causing substantial failures when facing new types of attacks (*e.g.*, 3D mask).

B. Method

Previous face PAD works [37]–[40] attempt to detect the evidence of liveness (*e.g.*, eye-blinking). Some works are based on contextual [41], [42] and moving [43]–[45] information. To improve the robustness to illumination variation, some algorithms adopt HSV and YCbCr color spaces [9], [10], as well as Fourier spectrum [46]. All of these methods use handcrafted features, such as LBP [11], [47], [48], HoG [47]–[49] and GLCM [49]. They achieve a relatively satisfactory performance on small public face anti-spoofing datasets.

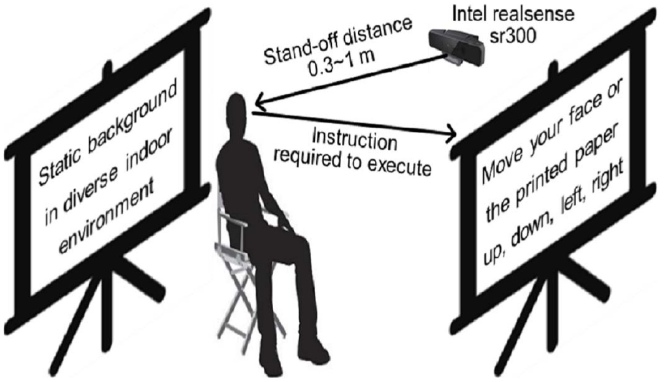

Intel realsense sr300
Stand-off distance 0.3~1 m
Static background in diverse indoor environment
Instruction required to execute
Move your face or the printed paper up, down, left, right

Fig. 2. Illustrative sketch of recording setups in the CASIA-SURF dataset.

Some fusion methods have been proposed to obtain a more general countermeasure effective against a variation of attack types. Tronci *et al.* [50] propose a linear fusion of frame and video analysis. Schwartz *et al.* [49] introduce feature level fusion based on a set of low-level feature descriptors. Other works [51], [52] obtain an effective fusion scheme by measuring the independence level of two anti-counterfeiting systems. However, they only focus on score or feature level, not modality level, due to the lack of multi-modal datasets.

CNN-based methods [26], [28], [53]–[56] have been presented recently. They treat face PAD as a binary classification and achieve remarkable improvements. Liu *et al.* [26] design a network to leverage Depth map and rPPG signal as supervision. Jourabloo *et al.* [28] solve the face anti-spoofing by inversely decomposing a spoof face into the live face and the spoof noise pattern. However, they exhibit a poor generalization ability due to the over-fitting to training data, even adopting transfer learning to train a CNN model [54], [55] from ImageNet [29]. These works show the need of a larger PAD dataset.

III. CASIA-SURF Dataset

Existing datasets involve a limited number of subjects and modalities, which severely impedes the development of face PAD with higher recognition to be applied in problems, such as face payment or unlock. In order to address these aforementioned limitations, we collect a new large-scale and multi-modal face PAD dataset namely CASIA-SURF. To the best our knowledge, the proposed dataset is currently the largest face anti-spoofing dataset, containing $1,000$ Chinese people in $21,000$ videos with three modalities (RGB, Depth, IR). Another motivation for creating this dataset, beyond pushing the further research of face anti-spoofing, is to explore the performance of recent face anti-spoofing methods when considering a large amount of data. In this section, we will give the detailed introduction of the proposed dataset, including acquisition detail, attack type, data preprocessing, statistics description, evaluation metric and protocol.

A. Acquisition Detail

Fig. 2 shows the diagram of data acquisition procedure, *i.e.*, how the multi-modal data is recorded via the multi-modal camera in diverse indoor environment. Specifically, we use the

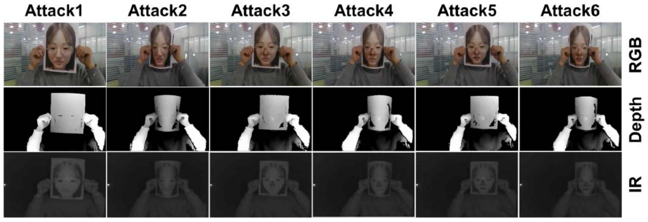


Fig. 3. Six attack styles in the CASIA-SURF dataset.

Intel RealSense SR300 camera to capture the RGB, Depth and InfraRed (IR) videos simultaneously. During the video recording, collectors are required to turn left or right, move up or down, walk in or away from the camera. Moreover, the performers stand within the range of 0.3 to 1.0 meter from the camera and their face angle is asked to be less 30^0. After that, four video streams including RGB, Depth, IR, plus RGB-Depth-IR aligned images are captured using the RealSense SDK at the same time. The resolution is 1280×720 for RGB images and 640×480 for Depth, IR and aligned images. Some examples of RGB, Depth, IR and aligned images are shown in the first column of Fig. 4.

B. Attack Type

We print collectors' color pictures with A4 paper to obtain the attack faces. The printed flat or curved face images will be cut eyes, nose, mouth areas or their combinations, generating 6 different attack ways. Thus, each sample includes 1 live video clip and 6 fake video clips. Fake samples are shown in Fig. 3. Detailed information of the 6 attacks is given below.

- Attack 1: One person hold his/her flat face photo where eye regions are cut.
- Attack 2: One person hold his/her curved face photo where eye regions are cut.
- Attack 3: One person hold his/her flat face photo where eye and nose regions are cut.
- Attack 4: One person hold his/her curved face photo where eye and nose regions are cut.
- Attack 5: One person hold his/her flat face photo where eye, nose and mouth regions are cut.
- Attack 6: One person hold his/her curved face photo where eye, nose and mouth regions are cut.

C. Data Preprocessing

Data preprocessing is widely used in the face recognition system, such as face detection and face alignment. Different pre-processing methods would affect the face anti-spoofing algorithms. To focus on the face anti-spoofing task and increase the difficulty, we process the original data via face detection and alignment. As shown in Fig. 4, we first use the Dlib [57] toolkit to detect face for every frame of RGB and RGB-Depth-IR aligned videos, respectively. Then we apply the PRNet [58] algorithm to perform 3D reconstruction and density alignment on the detected faces. After that, we define a binary mask based on non-active face reconstruction area from previous steps. Finally, we obtain face area of RGB image via point-wise product between the RGB image and the RGB binary mask. The Depth (or IR) area can be calculated via the point-wise product between the Depth (or IR) image and the RGB-Depth-IR binary mask. After the data pre-processing stage, we manually check all the processed RGB images to ensure that they contain a high-quality large face.

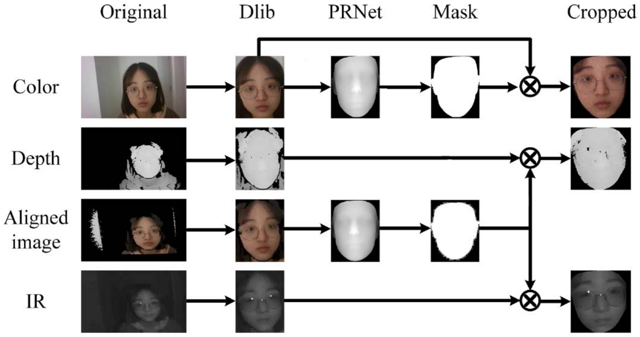


Fig. 4. Preprocessing details of three modalities of the CASIA-SURF dataset.

TABLE II
STATISTICAL INFORMATION OF THE PROPOSED CASIA-SURF DATASET

	Training	Validation	Testing	Total
# Subject	300	100	600	1,000
# Video	6,300	2,100	12,600	21,000
# Original image	1,563,919	501,886	3,109,985	5,175,790
# Sampled image	151,635	49,770	302,559	503,964
# Processed image	148,089	48,789	295,644	492,522

D. Statistics Description

Table II presents the main statistics of the proposed CASIA-SURF dataset. (1) There are $1,000$ subjects with variability in terms of gender, age, glasses/no glasses and indoor environments. Each one has 1 live video clip and 6 fake video clips. (2) Data is divided into three subsets. The training, validation and testing subsets have 300, 100 and 600 subjects with $6,300$ ($2,100$ per modality), $2,100$ (700 per modality), $12,600$ ($4,200$ per modality) videos, respectively. (3) From original videos, there are about 1.5 million, 0.5 million, 3.1 million frames in total for training, validation, and testing subsets, respectively. Owing to the huge amount of data, we select one frame out of every 10 frames and form the sampled set with about $151K$, $49K$, and $302K$ for training, validation and testing subsets, respectively. (4) After removing non-detected face poses with extreme lighting conditions during data prepossessing, we finally obtain about $148K$, $48K$, $295K$ images for training, validation and testing subsets in the CASIA-SURF dataset.

The information of gender statistics is shown in the left side of Fig. 5. It shows that the ratio of female is 56.8% while the ratio of male is 43.2%. In addition, we also show age distribution of the CASIA-SURF dataset in the right side of Fig 5. One can see a wide distribution of age ranges from 20 to more than 70 years old, while most of subjects are under 70 years old. On average, the range of [20, 30) ages is dominant, being about 50% of all the subjects.

E. Evaluation Protocol

We select the live faces and Attacks 4, 5, 6 as the training subset, while the live faces and Attacks 1, 2, 3 as the validation

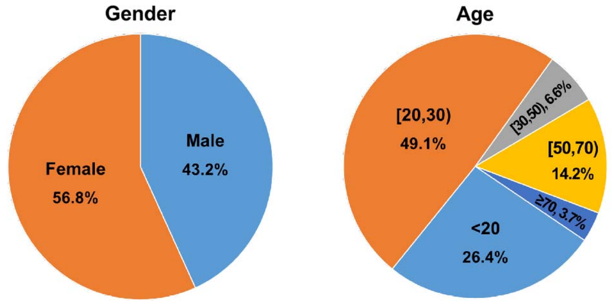


Fig. 5. Gender and age distribution of the CASIA-SURF dataset.

and testing subsets. This makes the ratio of flat/curved face and the extend of cut organ different between training and evaluation in order to increase the difficulty. The validation subset is used for model and hyper-parameter selection and the testing subset for final evaluation. Our dataset has two types of evaluation protocol: (1) **within-modal evaluation**, in which algorithms are trained and evaluated in the same modalities; (2) **cross-modal evaluation**, in which algorithms are trained in one modality while evaluated in other modalities.

F. Evaluation Metric

Following the face recognition task, we use the ROC curve as the main evaluation metric. ROC curve is a suitable indicator for the algorithms applied in the real world applications, because we can select a suitable trade-off threshold between FPR and TPR according to the requirements. Empirically, we compute TPR@FPR=10^{-2}, 10^{-3} and 10^{-4} as the quantitative indicators. Among them, we regard TPR@FPR=10^{-4} as the main comparison. Besides, the commonly used metric ACER, APCER and NPCER are also provided for reference.

IV. Proposed Method

Before showing some experimental analysis on the proposed dataset, we first built a strong baseline. We aim to find a straightforward architecture that provides good performance on our CASIA-SURF. Thus, we regard the face anti-spoofing problem as a binary classification task (fake *v.s* real) and conduct the experiments based on the ResNet-18/34 [59] classification network. ResNet-18/34 consist of five convolutional blocks (namely res1, res2, res3, res4, res5), a global average pooling layer and a softmax layer, which are relatively shallow networks with high classification performance.

A. Naive Halfway Fusion

CASIA-SURF 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. We use a multi-stream architecture with three subnetworks to study the dataset modalities, in which RGB, Depth and IR data are learnt separately by each stream, and then shared layers are appended at a point to learn joint representations and perform cooperated decisions. The halfway fusion is one of the commonly used fusion methods, which combines the subnetworks of different modalities at a later stage, *i.e.*, immediately after the third convolutional block (res3) via the feature map concatenation. In this way, features from different modalities can be fused to perform classification. However, direct concatenating these features cannot make full use of the characteristics between different modalities.

B. Squeeze and Excitation Fusion

The three modalities provide with complementary information for different kind of attacks: RGB data have rich appearance details, Depth data are sensitive to the distance between the image plane and the corresponding face, and IR data measure the amount of heat radiated from a face. Inspired by [60], we propose the Squeeze and Excitation Fusion (SEF) module to fuse features from different modalities. As shown in Fig. 6(b), this module first adds a branch[1] to obtain the channel-wise weights for each modality, then re-weights the input features and finally combines these re-weighted features together. Comparing to the naive halfway fusion that directly combines the features from different modalities, the SEF performs modality-dependent feature re-weighting to select the more informative channel features while suppressing less useful features from each modality.

C. From Single-Scale to Multi-Scale SEF

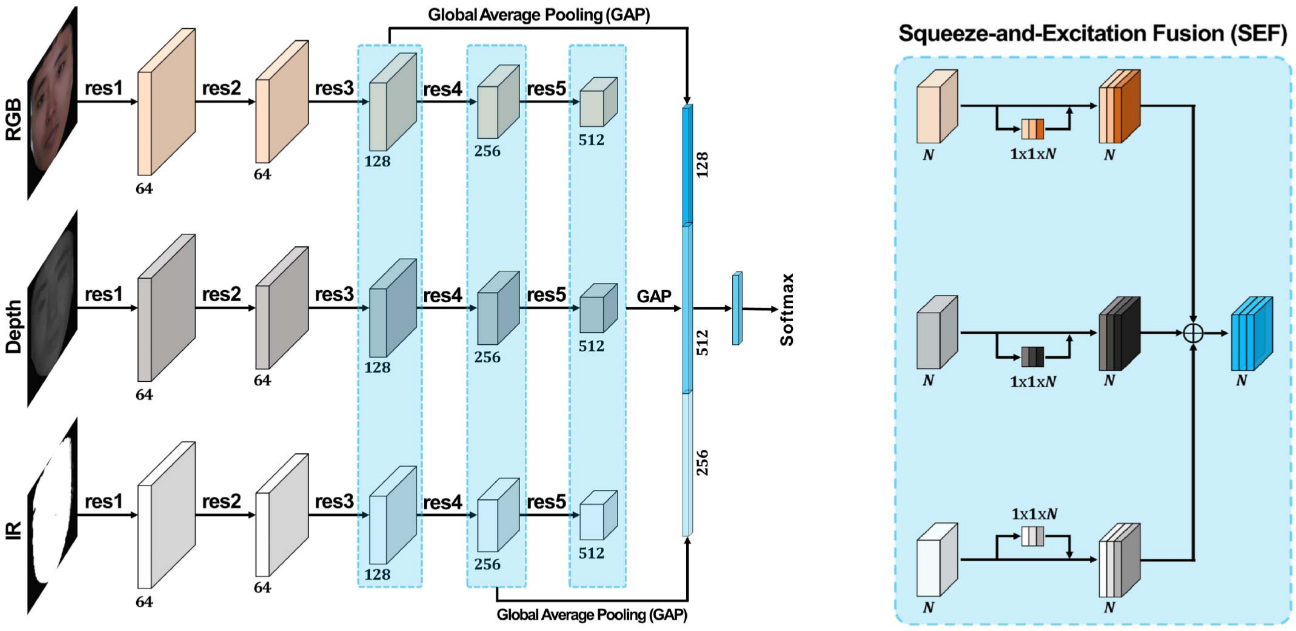
In our previous work [33], we only apply the SEF module on one of the scales in the ResNet-18 network, *i.e.*, the SEF module is appended after the res3 block to fuse features from different modalities and the subsequent blocks are shared. As is well-known, in convolutional neural networks, the high-level layer has a large receptive field with strong ability to represent semantic information, but has low resolution with weak ability to represent detailed information. While the low-level layer has a small receptive field with weak ability to represent semantic information, but has large resolution with strong ability to represent detailed information. For the anti-spoofing task, it is better to fuse deep features with strong semantic information and shallow features with detailed information to globally and locally determine whether a face is real or fake.

However, the single-scale SEF does not make full use of features from different levels. To this end, we extend the SEF from single scale to multiple scales. As shown in Fig. 6(a), our proposed method has a three-stream architecture and each subnetwork is feed with the image of different modalities. The res1, res2, res3, res4 and res5 blocks from each stream extract features from different modalities. After that, we first fuse features from different modalities via the SEF after res3, res4 and res5 respectively, then squeeze these fused features via the Global Average Pooling (GAP), next concatenate these squeezed features and finally use the concatenated features to predict real and fake.

V. Experiments

In this section, we firstly describe the implementation details, secondly verify the effectiveness of the proposed

[1] It is the same as the "Squeeze-and-Excitation" branch [60], composed of one global average pooling layer and two consecutive fully connected layers.


Global Average Pooling (GAP)
RGB
Depth
IR
res1
res2
res3
res4
res5
64
128
256
512
GAP
Softmax
Global Average Pooling (GAP)
Squeeze-and-Excitation Fusion (SEF)
N
1x1xN

(a) Structure of the proposed method

(b) Squeeze-and-Excitation Fusion (SEF)

Fig. 6. (a) Each stream uses ResNet-18/34 as backbone, which has five convolution blocks (*i.e.*, res1, res2, res3, res4, res5) to extract features of each modal data (*i.e.*, RGB, Depth, IR). We first fuse features from different modalities via SEF after res3, res4 and res5 respectively, then squeeze these fused features via GAP, next concatenate these squeezed features and finally use the concatenated features to predict real and fake. (b) Illustration of SEF.

TABLE III

EFFECTIVENESS OF THE PROPOSED METHOD. ALL RESNET-18 MODELS ARE TRAINED ON TRAINING SET AND EVALUATED ON TESTING SET. THE TESTING SET HAS 57, 710 TRIPLES (*i.e.*, RGB, DEPTH, IR IMAGES), COMPOSED BY 17, 458 REAL FACE TRIPLETS AND 40, 252 FAKE FACE TRIPLETS

Method	TPR (%) / # TP			APCER (%)	NPCER (%)	ACER (%)
	@FPR=10^{-2} (# FP≈403)	@FPR=10^{-3} (# FP≈40)	@FPR=10^{-4} (# FP≈4)			
Halfway fusion	89.1 / 15,555	33.6 / 5,866	17.8 / 3,108	5.6	3.8	4.7
SEF	96.7 / 16,882	81.8 / 14,281	56.8 / 9,916	3.8	1.0	2.4
+ Data augmentation	97.8 / 17,074	84.8 / 14,804	66.2 / 11,557	3.7	0.5	2.1
+ Addition operation	98.7 / 17,231	93.2 / 16,271	73.5 / 12,832	2.8	0.3	1.5
+ ImageNet pretrain	99.4 / 17,353	95.8 / 16,725	81.4 / 14,211	2.3	0.3	1.3
+ Multi-scale fusion	99.7 / 17,406	97.4 / 17,004	92.4 / 16,131	1.9	0.1	1.0
+ Stronger backbone	**99.8 / 17,423**	**98.4 / 17,179**	**95.2 / 16,620**	**1.6**	**0.08**	**0.8**

method, thirdly present a series of experiments to analyze the CASIA-SURF dataset in terms of number of modalities and subjects, fourthly conduct the cross-modal evaluation and finally present the generalization capability of the proposed dataset.

A. Implementation Detail

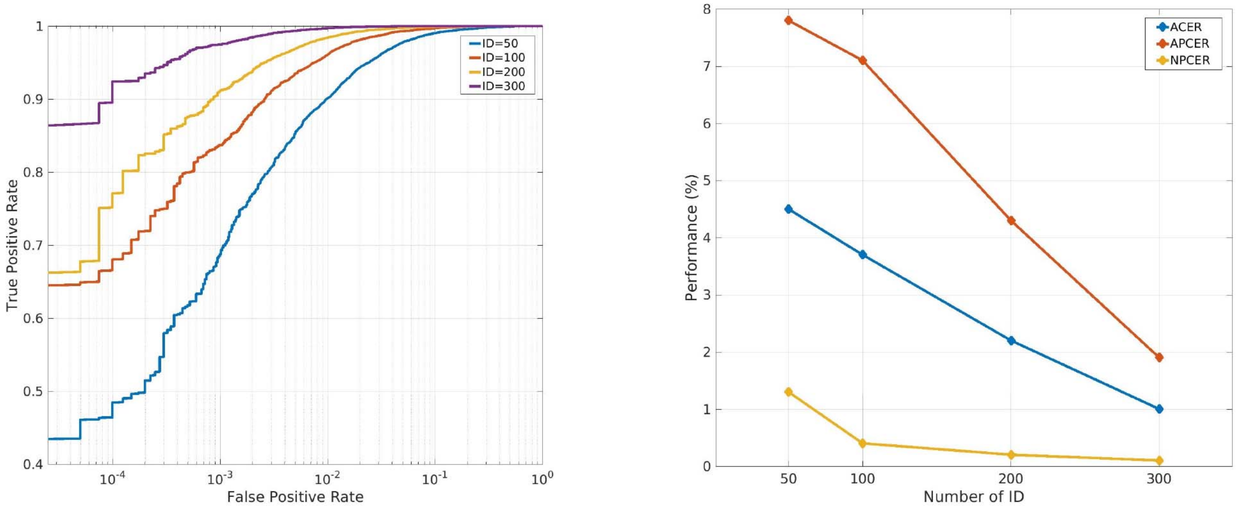
We resize the cropped face region to 112 × 112, and use random flipping, rotation, resizing, cropping and color distortion for data augmentation. For the CASIA-SURF dataset analyses, all models are trained for 40 epochs and the initial learning rate is 0.01, decreased by a factor of 10 after 20 and 30 epochs, respectively. All models are optimized via the Adaptive Moment Estimation (Adam) algorithm on 2 TITAN X (Maxwell) GPU with a mini-batch 256. Weight decay and momentum are set to 0.0005 and 0.9, respectively.

B. Model Analysis

As listed in Table III, we carry out some ablation experiments on the CASIA-SURF dataset to analyze our proposed method. For a fair comparison, we use the same settings except for the specific modification. In the conference version of this work [33], we have verified the effectiveness of the single-scale SEF module, which improves the TPR@FPR=[10^{-2}, 10^{-3}, 10^{-4}], APCER, NPCER, ACER from 89.1%, 33.6%, 17.8%, 5.6%, 3.8%, 4.7% to 96.7%, 81.8%, 56.8%, 3.8%, 1.0%, 2.4%, respectively. At this stage, the commonly used metrics APCER, NPCER and ACER are very promising, but TPR@FPR=[10^{-2}, 10^{-3}, 10^{-4}] have a big space to improve, especially for TPR@FPR=10^{-4}. To this end, we explore some strategies as shown in Table III to further improve the performance: (1) adjusting some hyper-parameters of data augmentation increases TPR by 1.1%, 3.0%, 9.4% for FPR=10^{-2}, 10^{-3}, 10^{-4}; (2) replacing the concatenation operation in the SEF module with the addition

TABLE IV
EFFECT OF NUMBER OF MODALITIES. ALL MODELS ARE BASED ON RESNET-18 AND TRAINED ON THE CASIA-SURF TRAINING SUBSET AND TESTED ON THE TESTING SUBSET WITH ONE, OR TWO, OR THREE MODALITIES

Modality	TPR (%)			APCER (%)	NPCER (%)	ACER (%)
	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
RGB	51.7	27.5	14.6	40.3	1.6	21.0
Depth	96.8	86.5	67.3	6.0	1.2	3.6
IR	62.5	29.4	15.9	38.6	0.4	19.4
RGB&Depth	97.1	87.5	71.1	5.8	0.8	3.3
RGB&IR	87.4	60.3	37.0	36.5	0.005	18.3
Depth&IR	99.4	95.2	81.2	2.0	0.3	1.1
RGB&Depth&IR	99.7	97.4	92.4	1.9	0.1	1.0



(a)

(b)

Fig. 7. (a) ROC curves of different training subset size in the CASIA-SURF dataset. (b) Performance *vs.* training subset size in the CASIA-SURF dataset.

operation boosts TPR@FPR=[10^{-2}, 10^{-3}, 10^{-4}] by 0.9%, 8.4%, 7.3%; (3) using ImageNet pretrained model brings 0.7%, 2.6%, 7.9% improvements for TPR@FPR=[10^{-2}, 10^{-3}, 10^{-4}]; (4) extending the SEF from single scale to multiple scales improves TPR@FPR=[10^{-2}, 10^{-3}, 10^{-4}] to 99.7%, 97.4%, 92.4%; (5) applying a stronger backbone from ResNet-18 to ResNet-34 has 0.1%, 1.0%, 2.8% improvements for TPR@FPR=[10^{-2}, 10^{-3}, 10^{-4}]. Besides, the APCER, NPCER and ACER are also improved from 3.8%, 1.0%, 2.4% to 1.6%, 0.08%, 0.8% after using these new strategies. Notably, the newly proposed multi-scale SEF achieves the most significant improvement 11.0% for TPR@FPR=10^{-4}, demonstrating its effectiveness.

C. Dataset Analysis

The proposed CASIA-SURF dataset has three modalities with $1,000$ subjects. In this subsection, we analyze the effect of the number of modalities and subjects.

Effect of number of modalities: As shown in Table IV, only using the prevailing RGB data, the results are 51.7%, 27.5%, 14.6% for TPR@FPR=[10^{-2}, 10^{-3}, 10^{-4}] and 40.3%, 1.6%, 21.0% for APCER, NPCER, ACER. In contrast, simply using the IR data, the results can be improved to 62.5% (TPR@FPR=10^{-2}), 29.4% (TPR@FPR=10^{-3}), 15.9% (TPR@FPR=10^{-4}), 38.6% (APCER), 0.4% (NPCER) and 19.4% (ACER), respectively. Among these three modalities, the Depth data achieves the best performance, *i.e.*, 96.8%, 86.5%, 67.3% for TPR@FPR=[10^{-2}, 10^{-3}, 10^{-4}], 6.0% for APCER and 3.6% for ACER. By fusing the data of arbitrary two modalities or all three ones, we observe an increase in performance. The best results are achieved by fusing all the three modalities, improving the best results of single modality from 96.8%, 86.5%, 67.3%, 6.0%, 0.4%, 3.6% to 99.7%, 97.4%, 92.4%, 1.9%, 0.1%, 1.0% for TPR@FPR=[10^{-2}, 10^{-3}, 10^{-4}], APCER, NPCER, ACER, respectively, demonstrating the necessity of multi-modal dataset.

Effect of number of subjects: As described in [61], there is a logarithmic relation between the amount of training data and the performance. To quantify the impact of having a large amount of training data in PAD, we show how the performance grows as training data increases in our benchmark. For this purpose, we train our baselines with different sized subsets of subjects randomly sampled from the training subset. This is, we randomly select 50, 100 and 200 from 300 subjects for training. Fig. 7(a) shows ROC curves for different number of subjects. We can see that the TPR is better when more subjects are used for training across different FPR. When FPR=10^{-4},

TABLE V
CROSS-MODAL EVALUATION. ALL MODELS ARE BASED ON RESNET-18 AND TRAINED ON CASIA-SURF TRAINING SET AND TESTED ON TESTING SET

Modality		TPR (%)			APCER (%)	NPCER (%)	ACER (%)
Training	Testing	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
RGB	Depth	16.8	1.6	0.1	82.9	0.8	41.8
RGB	IR	4.0	0.2	0.02	73.8	0.4	37.1
Depth	RGB	6.9	2.1	0.7	42.4	38.6	40.5
Depth	IR	6.0	1.4	0.3	3.7	86.5	45.1
IR	RGB	4.4	0.4	0.04	93.9	4.9	49.4
IR	Depth	0.09	0.01	0.001	60.2	95.9	78.1

the best TPR of 300 subjects is higher about 15% than the second best TPR result (ID=200), showing that using more data will achieve better performance. In Fig. 7(b), we also provide with the performance of ACER, APCER and NPCER under different number of subjects. Their performances are getting better when more subjects are considered.

Difficulty of different types of attacks: Since the Attacks 1, 2, 3 are selected as the validation and testing subsets, thus we evaluate the trained model (*i.e.*, TPR@FPR=10^{-4} is 92.4%) on Attack 1, Attack 2 and Attack 3, respectively. The corresponding TPR@FPR=10^{-4} performances are 94.4%, 92.9% and 86.4%. The difference between Attack 1 and Attack 2 is whether the fake face is curved. Attack 2 is more challenging than Attack 1, indicating that the curved fake face is more difficult than the flat fake face. The difference between Attack 1 and Attack 3 is whether the fake face is cutout. Attack 3 is more challenging than Attack 1, indicating that the cutout fake face is more difficult than the intact fake face.

D. Cross-Modal Evaluation

We introduce the cross-modality evaluation protocol for the academic community to explore new issues. Although there are no real world scenarios for this protocol until now, if algorithms trained on a certain modality data are able to perform well on other modalities data, this will greatly enhance their versatility for different scenes with different devices. We aim to provide this cross-modal evaluation protocol for those possible real-world scenarios in the future. In this protocol, one of RGB, Depth and IR modalities is used for training, and then evaluate on the remaining modalities. As shown in Table V, the model only trained on the RGB, Depth or IR modality is evaluated on the Depth and IR, RGB and IR, RGB and Depth modalities, respectively. All the results are far away from satisfactory, even worse than random guesses. The reason behind these poor results is the large differences between different modalities data. Therefore, it is a challenging task and deserves further study in academic community.

TABLE VI
EVALUATION RESULTS ON FOUR PROTOCOLS OF OULU-NPU

Prot.	Method	APCER (%)	NPCER (%)	ACER (%)
1	CPqD [64]	2.9	10.8	6.9
	GRADIANT [64]	1.3	12.5	6.9
	FAS-BAS [26]	1.6	**1.6**	1.6
	FAS-Ds [28]	1.2	1.7	**1.5**
	STASN [65]	1.2	2.5	1.9
	FAS-TD-SF [63]	**0.8**	10.8	5.8
	FAS-TD-SF (CASIA-SURF)	2.7	2.5	2.6
2	MixedFASNet [64]	9.7	2.5	6.1
	FAS-Ds [28]	4.2	4.4	4.3
	FAS-BAS [26]	**2.7**	2.7	2.7
	GRADIANT [64]	3.1	1.9	2.5
	STASN [65]	4.2	**0.3**	**2.2**
	FAS-TD-SF [63]	3.6	3.8	3.7
	FAS-TD-SF (CASIA-SURF)	**2.7**	1.6	**2.2**
3	MixedFASNet [64]	5.3±6.7	7.8±5.5	6.5±4.6
	GRADIANT [64]	2.6±3.9	5.0±5.3	3.8±2.4
	FAS-Ds [28]	4.0±1.8	3.8±1.2	3.6±1.6
	FAS-BAS [26]	2.7±1.3	3.1±1.7	2.9±1.5
	STASN [65]	4.7±3.9	**0.9±1.2**	2.8±1.6
	FAS-TD-SF [63]	3.1±1.8	6.6±9.4	5.3±4.4
	FAS-TD-SF (CASIA-SURF)	**2.4±1.5**	2.2±3.8	**2.3±2.6**
4	Massy_HNU [64]	35.8±35.3	8.3±4.1	22.1±17.6
	GRADIANT [64]	**5.0±4.5**	15.0±7.1	10.0±5.0
	FAS-BAS [26]	9.3±5.6	10.4±6.0	9.5±6.0
	FAS-Ds [28]	5.1±6.3	6.1±5.1	**5.6±5.7**
	STASN [65]	6.7±10.6	8.3±8.4	7.5±4.7
	FAS-TD-SF [63]	7.0±5.3	20.0±24.8	13.5±10.9
	FAS-TD-SF (CASIA-SURF)	8.7±5.6	**5.8±8.0**	7.2±5.8

E. Using CASIA-SURF for Pre-Training

The CASIA-SURF dataset contains not only RGB images, but also the corresponding Depth information, which is indeed beneficial for Depth supervised face anti-spoofing methods [26], [62]. Thus, we adopt FAS-TD-SF [62] as our baseline to evaluate the generalization capability of the proposed dataset. We first pre-train the model on CASIA-SURF and then fine-tune with the concerned dataset including Oulu-NPU [25], SiW [26] and CASIA-MFSD [12]. This model is termed as FAS-TD-SF (CASIA-SURF).

Oulu-NPU dataset: It is a high-resolution dataset, consisting of 4, 950 real access and spoofing videos with many real-world variations. This dataset contains 4 evaluation protocols to validate the generalization of methods: Protocol 1 evaluates on the illumination variation; Protocol 2 examines the influence of different attack medium, such as unseen printers or displays; Protocol 3 studies the effect of the input camera variation; Protocol 4 considers all the factors above, which is the most challenging. As shown in Table VI, using the proposed dataset to pre-train our baseline method FAS-TD-SF significantly improves its ACER performance, *i.e.*, from

TABLE VII
EVALUATION RESULTS ON THREE PROTOCOLS OF SIW

Prot.	Method	APCER(%)	NPCER(%)	ACER(%)
1	FAS-BAS [26]	3.58	3.58	3.58
	STASN [65]	-	-	1.00
	FAS-TD-SF [63]	**1.27**	0.83	1.05
	FAS-TD-SF (CASIA-SURF)	**1.27**	**0.33**	**0.80**
2	FAS-BAS [26]	0.57±0.69	0.57±0.69	0.57±0.69
	STASN [65]	-	-	0.28±0.05
	FAS-TD-SF [63]	0.33±0.27	0.29±0.39	0.31±0.28
	FAS-TD-SF (CASIA-SURF)	**0.08±0.17**	**0.25±0.22**	**0.17±0.16**
3	FAS-BAS [26]	8.31±3.81	8.31±3.80	8.31±3.81
	STASN [65]	-	-	12.10±1.50
	FAS-TD-SF [63]	7.70±3.88	7.76±4.09	7.73±3.99
	FAS-TD-SF (CASIA-SURF)	**6.27±4.36**	**6.43±4.42**	**6.35±4.39**

TABLE VIII
EVALUATION RESULTS ON DIFFERENT CROSS-TESTING PROTOCOLS

Method	Training	Testing	HTER (%)
Motion [52]	Replay-Attack	CASIA-MFSD	47.9
LBP [52]	Replay-Attack	CASIA-MFSD	57.6
Motion-Mag [40]	Replay-Attack	CASIA-MFSD	47.0
Spectral cubes [66]	Replay-Attack	CASIA-MFSD	50.0
CNN [57]	Replay-Attack	CASIA-MFSD	45.5
STASN [65]	Replay-Attack	CASIA-MFSD	**30.9**
FAS-TD-SF [63]	SiW	CASIA-MFSD	39.4
FAS-TD-SF [63]	CASIA-SURF	CASIA-MFSD	37.3

5.8% to 2.6% in Protocol 1, from 3.7% to 2.2% in Protocol 2, from 5.3% to 2.3% in Protocol 3, and from 13.5% to 7.2% in Protocol 4. Without bells and whistles, our method achieves the lowest ACER in 2 out of 3 protocols. We believe that other state-of-the-art methods can be further improved by using our CASIA-SURF as the pre-training dataset.

SiW dataset: It contains more live subjects and has three protocols used for evaluation, please refer to [26] for more details of the protocols. Table VII shows the comparison between three state-of-the-art methods on the SiW dataset. FAS-TD-SF generally achieves better performance than FAS-BAS, while our pre-trained FAS-TD-SF on CASIA-SURF can further improve the performance across all protocols. Concretely, the performance of ACER is reduced by 0.25%, 0.14% and 1.38% in Protocol 1, 2, and 3 respectively when using the proposed CASIA-SURF dataset as pre-training. The improvement indicates that pre-training on the proposed dataset supports the generalization on data containing variabilities in terms of (1) face pose and expression, (2) replay attack mediums, and (3) cross Presentation Attack Instruments (PAIs), such as from print attack to replay attack. Interestingly, it also demonstrates our dataset is also useful to be used for pre-trained models when replay attack mediums cross PAIs.

CASIA-MFSD dataset: It contain low-resolution videos with resolution 640 × 480 and 1280 × 720. To further evaluate the generalization capability of the proposed dataset, we perform cross-testing experiments on this dataset, *i.e.*, training on the proposed CASIA-SURF and then directly evaluating on the CASIA-MFSD dataset. State-of-the-art methods [40], [51], [56], [65] are listed for comparison, which use the Replay-Attack [11] dataset for training. Results in Table VIII show that the model trained on the CASIA-SURF dataset performs the best among all models.

VI. DISCUSSION

Why not collect video replay attacks? In the design stage of the proposed dataset, we found that replay videos are presented black in depth images, *i.e.*, pixels in depth images are zero because of the same depth value for replay videos. It means that replay video attacks are easy to be recognized by means of depth data. This is why the developed dataset contains only the print attack and not the video replay attack. Besides, there are many other ways of attacking and we plan to continuously include more presentation attack ways (*e.g.*, 3D masks).

Why use the ROC curve as the evaluation metric? As shown in Table III, accurate results are achieved on the CASIA-SURF dataset for traditional metrics, *e.g.*, APCER=1.6%, NPCER=0.08%, ACER=0.8%. However, APCER=1.6% means about 2 fake samples from 100 attackers will be treated as real ones. This is below the accuracy requirements of real applications, *e.g.*, face payment and phone unlock. To decrease the gap between technology development and practical applications, the ROC curve is more suitable as the evaluation metric for face anti-spoofing to reflects whether algorithms meet the requirements of a given real application.

VII. CONCLUSION

This paper builds a large-scale multi-modal face anti-spoofing dataset namely CASIA-SURF. It is the largest one in terms of number of subjects, data samples, and number of visual data modalities. Comprehensive evaluation metrics, diverse evaluation protocols, training/validation/testing subsets and a measurement tool are also provided to develop a new benchmark. We believe this dataset will push the state-of-the-art in face anti-spoofing. Furthermore, we proposed a multi-modal multi-scale fusion method, which performs modality-dependent feature re-weighting to select the more informative channel features while suppressing the less informative ones for each modality across different scales. Extensive experiments have been conducted on the CASIA-SURF dataset to verify the generalization capability of models trained on the proposed dataset and the benefit of using multiple visual modalities. In the further, we plan to continuously increase the diversity of the dataset by including more presentation attack modalities (*e.g.*, 3D masks) and more subjects (*e.g.*, different ethnicity). On the other hand, we also plan to study heterogeneous face anti-spoofing using the cross-modal evaluation protocol.

ACKNOWLEDGMENT

The authors gratefully acknowledge Surfing Technology Beijing co., Ltd (www.surfing.ai) to capture and provide them this high quality dataset for this research.

REFERENCES

[1] S. Zhang, C. Chi, Z. Lei, and S. Z. Li, “Refineface: Refinement neural network for high performance face detection,” 2019. [Online]. Available: https://arxiv.org/abs/1909.04376.

[2] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, "Mis-classified vector guided softmax loss for face recognition," in *Proc. AAAI*, 2020 pp. 1–8.

[3] S. Zhang, L. Wen, H. Shi, Z. Lei, S. Lyu, and S. Z. Li, "Single-shot scale-aware network for real-time face detection," *Int. J. Comput. Vis.*, vol. 127, nos. 6–7, pp. 537–559, 2019.

[4] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8231–8238.

[5] S. Zhang, X. Wang, Z. Lei, and S. Z. Li, "FaceBoxes: A CPU real-time and accurate unconstrained face detector," *Neurocomputing*, vol. 364, pp. 297–309, Oct. 2019.

[6] X. Wang, S. Zhang, Z. Lei, S. Liu, X. Guo, and S. Z. Li, "Ensemble soft-margin softmax loss for image classification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 992–998.

[7] S. Zhang *et al.*, "Improved selective refinement network for face detection," 2019. [Online]. Available: https://arxiv.org/abs/1901.06651.

[8] S. Zhang, X. Zhu, Z. Lei, X. Wang, H. Shi, and S. Z. Li, "Detecting face with densely connected face proposal network," *Neurocomputing*, vol. 284, pp. 119–127, Apr. 2018.

[9] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1818–1830, Aug. 2016.

[10] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face antispoofing using speeded-up robust features and fisher vector encoding," *Signal Process Lett.*, vol. 24, no. 2, pp. 141–145, 2017.

[11] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Proc. Biometrics Electron. Signatures*, 2012, pp. 1–7.

[12] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *Proc. Int. Conf. Biometrics*, 2012, pp. 26–31.

[13] N. Erdogmus and S. Marcel, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with kinect," in *Proc. Int. Conf. Biometrics Theory Appl. Syst.*, 2014, pp. 1–6.

[14] T. I. Dhamecha, R. Singh, M. Vatsa, and A. Kumar, "Recognizing disguised faces: Human and machine evaluation," *PLoS ONE*, vol. 9, no. 7, 2014, Art. no. e99212.

[15] R. Raghavendra, K. B. Raja, and C. Busch, "Presentation attack detection for face recognition using light field camera," *Trans. Image Process.*, vol. 24, no. 3, pp. 1060–1075, 2015.

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

[17] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *Proc. Biometrics Electron. Signatures*, 2016, pp. 209–216.

[18] S. Liu, B. Yang, P. C. Yuen, and G. Zhao, "A 3D mask face anti-spoofing database with real world variations," in *Proc. CVPRW Workshops*, 2016, pp. 100–106.

[19] I. Chingovska, N. Erdogmus, A. Anjos, and S. Marcel, "Face recognition systems under spoofing attacks," in *Face Recognition Across the Imaging Spectrum*. Cham, Switzerland: Springer, 2016.

[20] H. Steiner, A. Kolb, and N. Jung, "Reliable face anti-spoofing using multispectral SWIR imaging," in *Proc. Int. Conf. Biometrics*, 2016, pp. 1–8.

[21] S. Holger, S. Sebastian, K. Andreas, and J. Norbert, "Design of an active multispectral SWIR camera system for skin detection and face verification," *J. Sensors*, vol. 2016, pp. 1–16, Nov. 2016.

[22] R. Raghavendra, K. B. Raja, S. Venkatesh, F. A. Cheikh, and C. Busch, "On the vulnerability of extended multispectral face recognition systems towards presentation attacks," in *Proc. Int. Conf. Identity Security Behav. Anal.*, 2017, pp. 1–8.

[23] I. Manjani, S. Tariyal, M. Vatsa, R. Singh, and A. Majumdar, "Detecting silicone mask-based presentation attack via deep dictionary learning," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 7, pp. 1713–1723, Jul. 2017.

[24] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face presentation attack with latex masks in multispectral videos," in *Proc. CVPRW Workshops*, 2017, pp. 81–89.

[25] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "OULU-NPU: A mobile face presentation attack database with real-world variations," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 612–618.

[26] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proc. CVPR*, 2018, pp. 389–398.

[27] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Trans. Inf. Forensics Security*, early access, doi: 10.1109/TIFS.2019.2916652.

[28] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014. [Online]. Available: https://arxiv.org/abs/1411.7923.

[31] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "SphereFace: Deep hypersphere embedding for face recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2017, pp. 6738–6746.

[32] X. Wang, S. Wang, S. Zhang, T. Fu, H. Shi, and T. Mei, "Support vector guided softmax loss for face recognition," 2018. [Online]. Available: https://arxiv.org/abs/1812.11317.

[33] S. Zhang *et al.*, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 919–928.

[34] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing deep face recognition with custom silicone masks," in *Proc. Int. Conf. Biometrics Theory Appl. Syst.*, 2018, pp. 1–7.

[35] Y. Kim, J. Na, S. Yoon, and J. Yi, "Masked fake face detection using radiance measurements," *J. Opt. Soc. America A*, vol. 26, no. 4, pp. 760–766, 2009.

[36] N. Kose and J.-L. Dugelay, "Countermeasure for the protection of face recognition systems against mask attacks," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2013, pp. 1–6.

[37] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic Webcamera," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[38] L. Wang, X. Ding, and C. Fang, "Face live detection method based on physiological motion analysis," *Tsinghua Sci. Technol.*, vol. 14, no. 6, pp. 685–690, 2009.

[39] K. Kollreider, H. Fronthaler, and J. Bigün, "Verifying liveness by multiple experts in face biometrics," in *Proc. CVPR Workshops*, 2008, pp. 1–6.

[40] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh, "Computationally efficient face spoofing detection with motion magnification," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 105–110.

[41] G. Pan, L. Sun, Z. Wu, and Y. Wang, "Monocular camera-based face liveness detection by combining eyeblink and scene context," *Telecommun. Syst.*, vol. 47, nos. 3–4, pp. 215–225, 2011.

[42] J. Komulainen, A. Hadid, and M. Pietikainen, "Context based face anti-spoofing," in *Proc. Int. Conf. Biometrics Theory Appl. Syst.*, 2013, pp. 1–8.

[43] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection using 3D structure recovered from a single camera," in *Proc. Int. Conf. Biometrics*, 2013, pp. 1–6.

[44] M. De Marsico, M. Nappi, D. Riccio, and J.-L. Dugelay, "Moving face spoofing detection via 3D projective invariants," in *Proc. Int. Conf. Biometrics*, 2012, pp. 73–78.

[45] S. Kim, S. Yu, K. Kim, Y. Ban, and S. Lee, "Face liveness detection using variable focusing," in *Proc. Int. Conf. Biometrics*, 2013, pp. 1–6.

[46] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of Fourier spectra," *Proc. SPIE*, vol. 5404, pp. 296–303, Aug. 2004.

[47] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor," in *Proc. Int. Conf. Biometrics*, 2013, pp. 1–6.

[48] J. Maatta, A. Hadid, and M. Pietikainen, "Face spoofing detection from single images using texture and local shape analysis," *IET Biometrics*, vol. 1, no. 1, pp. 3–10, 2012.

[49] W. R. Schwartz, A. Rocha, and H. Pedrini, "Face spoofing detection through partial least squares and low-level descriptors," in *Proc. Int. Joint Conf. Biometrics*, 2011, pp. 1–8.

[50] R. Tronci *et al.*, "Fusion of multiple clues for photo-attack detection in face recognition systems," in *Proc. Int. Joint Conf. Biometrics*, 2011, pp. 1–6.

[51] T. de Freitas Pereira, A. Anjos, J. M. De Martino, and S. Marcel, "Can face anti-spoofing countermeasures work in a real world scenario?" in *Proc. Int. Conf. Biometrics*, 2013, pp. 1–8.

[52] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel, "Complementary countermeasures for detecting scenic face spoofing attacks," in *Proc. Int. Conf. Biometrics*, 2013, pp. 1–7.

[53] L. Feng *et al.*, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *J. Vis. Commun. Image Representation*, vol. 38, pp. 451–460, Jul. 2016.

[54] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *Proc. Int. Conf. Image Process. Theory Tools Appl.*, 2016, pp. 1–6.

[55] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 10, pp. 2268–2283, Oct. 2016.

[56] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," 2014. [Online]. Available: https://arxiv.org/abs/1408.5601.

[57] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1–4. Dec. 2009.

[58] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 557–574.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[60] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

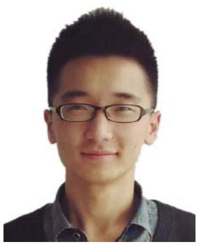
[61] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 843–852.

[62] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, and Z. Lei, "Exploiting temporal and depth information for multi-frame face anti-spoofing," 2018. [Online]. Available: https://arxiv.org/abs/1811.05118.

[63] Z. Boulkenafet *et al.*, "A competition on generalized software-based face presentation attack detection in mobile scenarios," in *Proc. IJCB*, 2017, pp. 688–696.

[64] X. Yang *et al.*, "Face anti-spoofing: Model matters, so does data," in *Proc. Comput. Vis. Pattern Recognit.*, 2019, pp. 3507–3516.

[65] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4726–4740, Dec. 2015.

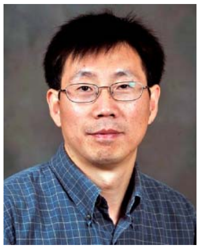


Shifeng Zhang received the B.S. degree from the University of Electronic Science and Technology of China in 2015. He is currently pursuing the Ph.D. degree with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science. His research interests include computer vision, pattern recognition, especially with a focus on object detection, face detection, pedestrian detection, and video detection.

Ajian Liu received the B.E. degree from the College of Physics and Information Engineering, Shanxi Normal University, Shanxi, China, in 2015, and the master's degree from the College of Information and Computer, Taiyuan University of Technology, Shanxi, in 2018. He is currently pursuing the Ph.D. degree with the Faculty of Information Technology, Macau University of Science and Technology. His main research interests include deep learning and face anti-spoofing.

Jun Wan (Senior Member, IEEE) received the B.S. degree from the China University of Geosciences, Beijing, China, in 2008, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, in 2015. Since January 2015, he has been a Faculty Member with the National Laboratory of Pattern Recognitions, Institute of Automation, Chinese Academy of Science, China, where he currently serves as an Associate Professor. His main research interests include computer vision, machine learning, especially, for gesture and action recognition, and facial attribution analysis.

Yanyan Liang (Senior Member, IEEE) received the B.S. degree from the Chongqing University of Communication and Posts, Chongqing, China, in 2004, and the M.S. and Ph.D. degrees from the Macau University of Science and Technology, Macau, China, in 2006 and 2009, respectively, where he is currently an Assistant Professor. He has published over 30 papers related to pattern recognition, image processing, and computer version. He is also researching on smart city applications with computer vision. His current research interests include computer vision, image processing, and machine learning.



Guodong Guo (Senior Member, IEEE) received the B.E. degree in automation from Tsinghua University, Beijing, China, the Ph.D. degree in pattern recognition and intelligent control from the Chinese Academy of Sciences, Beijing, and the Ph.D. degree in computer science from the University of Wisconsin–Madison, Madison, WI, USA. He visited and worked in several places, including INRIA, Sophia Antipolis, France; Ritsumeikan University, Kyoto, Japan; Microsoft Research, Beijing, China; and North Carolina Central University. He is an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University (WVU), Morgantown, WV, USA. He authored a book *Face, Expression, and Iris Recognition Using Learning-Based Approaches* (2008), co-edited 2 books, *Support Vector Machines Applications* (2014) and *Mobile Biometrics* (2017), and published over 100 technical papers. His research interests include computer vision, biometrics, machine learning, and multimedia. He received the North Carolina State Award for Excellence in Innovation in 2008, the Outstanding Researcher from 2017 to 2018 and from 2013 to 2014 at CEMR, WVU, and the New Researcher of the Year from 2010 to 2011 at CEMR, WVU. He was selected the "People's Hero of the Week" by BSJB under Minority Media and Telecommunications Council in 2013. Two of his papers were selected as "The Best of FG'13" and "The Best of FG'15," respectively.

Sergio Escalera received the Ph.D. degree from Computer Vision Center, UAB in 2018. He leads the Human Pose Recovery and Behavior Analysis Group. He is an Associate Professor with the Department of Mathematics and Informatics, Universitat de Barcelona. He is also a Member of the Computer Vision Center, UAB. His research interests include the visual analysis of humans, with special interest in affective and personality computing. He has been awarded with ICREA Academia. He is the Vice-President of ChaLearn Challenges in Machine Learning and the Chair of IAPR TC-12: Multimedia and Visual Information Systems.

Hugo Jair Escalante is a Researcher Scientist with the Instituto Nacional de Astrofisica, Óptica y Electrónica, INAOE, Mexico. He has been the Director of ChaLearn, a nonprofit organization dedicated to organizing challenges since 2011. He has been involved in the organization of several challenges in computer vision and automatic machine learning. He is Reviewer with the *Journal of Machine Learning Research* and PAMI. He has served as the Competition Chair and the Area Chair of venues like NeurIPS, PAKDD, and IJCNN.

Stan Z. Li (Fellow, IEEE) received the B.Eng. degree from Hunan University, the M.Eng. degree from the National University of Defense Technology, and the Ph.D. degree from Surrey University. He is currently a Professor and the Director of Center for Biometrics and Security Research, Institute of Automation, Chinese Academy of Sciences. He was an Associate Professor with Nanyang Technological University. He was with MSRA as a Researcher from 2000 to 2004. His research interests include image and vision processing, pedestrian recognition, and biometrics. He has published more than 300 papers in international journals and conferences, and authored and edited 8 books. He was an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is acting as the Editor-in-Chief of the *Encyclopedia of Biometrics*. He served as the Program Co-Chair for ICB 2007, 2009, 2013, 2014, 2015, 2016, and 2018, and has been involved in organizing other international conferences and workshops in the fields of his research interest. He was elevated to IEEE Fellow for his contributions to the fields of pedestrian recognition, pattern recognition, and computer vision. He is a member of the IEEE Computer Society.