CASIA-SURF: A Large-scale Multi-modal Benchmark for Face Anti-spoofing

Shifeng Zhang*, Ajian Liu*, Jun Wan†, *Member, IEEE,* Yanyan Liang, *Member, IEEE,*
Guogong Guo, *Member, IEEE,* Sergio Escalera, *Member, IEEE,* Hugo Jair Escalante, 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/chalearnfacespoofingattackdete/.**

***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]–[5] 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) [6], [7] is a vital step to ensure that face recognition systems are in a safe reliable condition.

Shifeng Zhang, Jun Wan and Stan Z. Li are with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation Chinese Academy of Sciences (CASIA) and University of Chinese Academy of Sciences (UCAS), Beijing, China (e-mail: {shifeng.zhang, jun.wan, szli}@nlpr.ia.ac.cn).

Ajian Liu, Yanyan Liang and Stan Z. Li are with the Macau University of Science and Technology (MUST), Macau, 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 (e-mail: guoguodong01@baidu.com).

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

Hugo Jair Escalante is with the Instituto Nacional de Astrofsica, ptica y Electrnica, Puebla 72840, Mxico (e-mail: hugojair@inaoep.mx).

*Equal contribution. † Corresponding author.

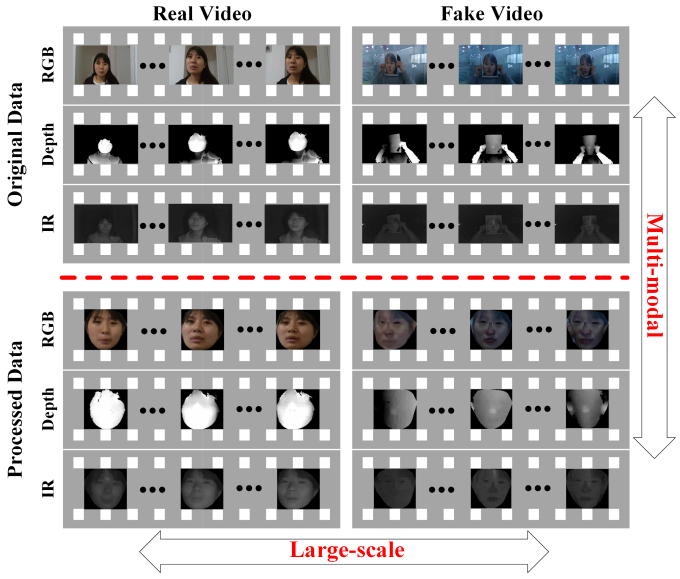


Fig. 1. The CASIA-SURF dataset. It is a large-scale and multi-modal dataset for face anti-spoofing, consisting of $492,522$ images with 3 modalities (*i.e.*, RGB, Depth and IR).

In recent years, face PAD algorithms [15], [16] have achieved great performances. One of the key points of this success is the availability of face anti-spoofing benchmark datasets [8], [9], [11], [12], [14], [15]. However, there are several shortcomings in the existing datasets as follows:

- **Number of subjects is limited.** Compared to the large existing image classification [17] and face recognition [18] 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 [10], [13], 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 [12], [14]–[16] 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

TABLE I
COMPARISON OF THE PUBLIC FACE ANTI-SPOOFING DATASETS (∗ INDICATES THIS DATASET ONLY CONTAINS IMAGES, NOT VIDEO CLIPS).

Dataset	Year	# of subjects	# of videos	Camera	Modal types	Spoof attacks
Replay-Attack [8]	2012	50	1200	VIS	RGB	Print, 2 Replay
CASIA-MFSD [9]	2012	50	600	VIS	RGB	Print, Replay
3DMAD [10]	2013	17	255	VIS/Kinect	RGB/Depth	3D Mask
MSU-MFSD [11]	2015	35	440	Phone/Laptop	RGB	Print, 2 Replay
Replay-Mobile [12]	2016	40	1030	VIS	RGB	Print, Replay
Msspoof [13]	2016	21	4704*	VIS/NIR	RGB/IR	Print
Oulu-NPU [14]	2017	55	5940	VIS	RGB	2 Print, 2 Replay
SiW [15]	2018	165	4620	VIS	RGB	2 Print, 4 Replay
CASIA-SURF (Ours)	2018	**1000**	**21000**	RealSense	**RGB/Depth/IR**	Print, Cut

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 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 $6\times$ boosted compared with previous challenging face anti-spoofing dataset like SiW.
- **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 [19], [20], 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.

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 [21]. 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 in recent years and lots of methods have been proposed with the help of face anti-spoofing datasets. In this section, we first summarize the existing face anti-spoofing datasets and then review some representative methods

A. Dataset

Most of existing face anti-spoofing datasets only contain the RGB modality, including the two widely used PAD datasets Replay-Attack [8] and CASIA-FASD [9]. Even the recently released SiW [15] dataset, collected with high resolution image quality, only contains RGB data. With the widespread application of face recognition in mobile phones, there are also some RGB datasets recorded by replaying face video with smartphone, such as MSU-MFSD [11], Replay-Mobile [12] and OULU-NPU [14].

As attack techniques are constantly upgraded, some new types of presentation attacks have emerged, *e.g.*, 3D [10]

and silicone masks [1]. These attacks are more realistic than traditional 2D attacks. Therefore, the drawbacks of visible cameras are revealed when facing these realistic face masks. Fortunately, some new sensors have been introduced to provide more possibilities for face PAD methods, such as depth cameras, muti-spectral cameras and infrared light cameras. Kim *et al.* [22] introduce a new dataset to distinguish between the facial skin and mask materials by exploiting their reflectance. Kose *et al.* [23] propose a 2D+3D face mask attack dataset to study the effects of mask attacks. However, associated data has not been made public. 3DMAD [10] is the first publicly available 3D masks dataset, which is recorded using Microsoft Kinect sensor and consists of Depth and RGB modalities. Another multi-modal face PAD dataset is Msspoof [13], containing visible and near-infrared images of real accesses and printed spoofing attacks with ≤ 21 objects.

However, existing datasets in the face PAD community have two main limitations. First, they all have the limited number of subjects and samples, resulting in a potential over-fitting risk when face PAD algorithms are tested on these datasets [8], [9]. Second, most of existing datasets are captured by visible camera that only includes the RGB modality, causing a substantial portion of 2D PAD methods to fail when facing new types of attacks (*e.g.*, 3D and custom-made silicone masks).

B. Method

Face anti-spoofing has been studied for decades. Some previous works [24]–[27] attempt to detect the evidence of liveness (*e.g.*, eye-blinking). Another works are based on contextual [28], [29] and moving [30]–[32] information. To improve the robustness to illumination variation, some algorithms adopt HSV and YCbCr color spaces [6], [7], as well as Fourier spectrum [33]. All of these methods use handcrafted features, such as LBP [34]–[37], HoG [36]–[38] and GLCM [38]. They achieve a relatively satisfactory performance on small public face anti-spoofing datasets.

Some fusion methods have been proposed to obtain a more general countermeasure effective against a variation of attack types. Tronci *et al.* [39] propose a linear fusion of frame and video analysis. Schwartz *et al.* [38] introduce feature level fusion by using Partial Least Squares (PLS) regression based on a set of low-level feature descriptors. Other works [40], [41] obtain an effective fusion scheme by measuring the level of independence of two anti-counterfeiting systems. However, these fusion methods focus on score or feature level, not modality level, due to the lack of multi-modal datasets.

CNN-based methods [15], [16], [42]–[45] have been presented recently in the face PAD community. They treat face PAD as a binary classification problem and achieve remarkable improvements in the intra-testing. Liu *et al.* [15] design a network architecture to leverage two auxiliary information (Depth map and rPPG signal) as supervision. Amin *et al.* [16] introduce a new perspective for solving 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 in the cross-testing due to the over-fitting to training data. This problem remains open, although some works [43], [44] adopt transfer learning to train a CNN model from ImageNet [17]. These works show the need of a larger PAD dataset.

III. CASIA-SURF DATASET

As mentioned before, all existing datasets involve a limited number of subjects and up to two modalities. Although these publicly available datasets have driven the development of face PAD and continue to be valuable tools for this community, their limitations severely impede 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 in PAD community, 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 recent face spoofing detection models performance 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

The diagram of data acquisition procedure is shown in Fig. 2, where it shows how the multi-modal data is recorded via the multi-modal camera in diverse indoor environment. Specifically, we use the Intel RealSense SR300 camera to capture the RGB, Depth and InfraRed (IR) videos simultaneously. During the video recording, the collectors are required to do some actions, such as turning left or right, moving up or down, walking 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.

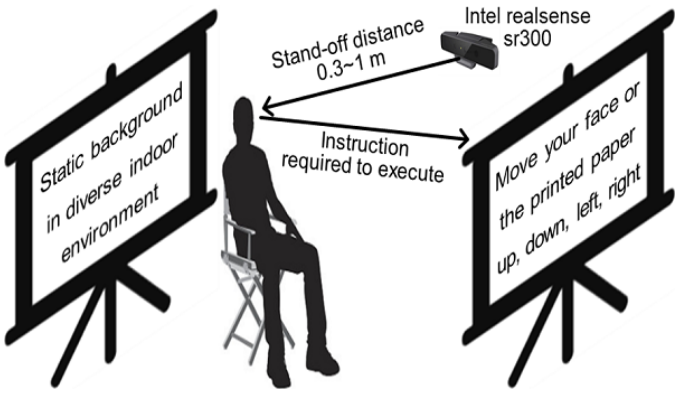


Fig. 2. Illustrative sketch of recording setups in the CASIA-SURF dataset.

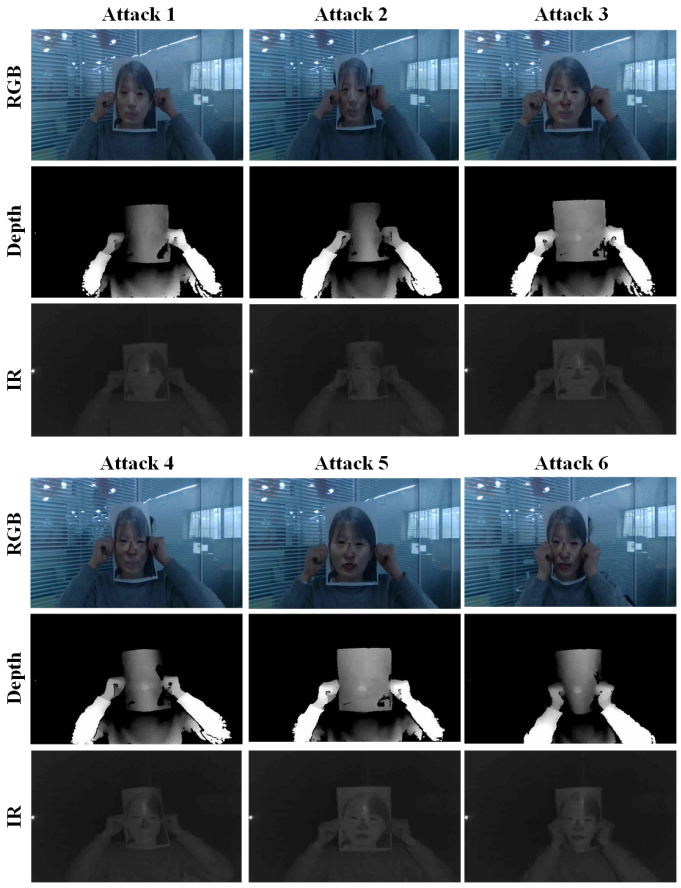


Fig. 3. Six attack styles in the CASIA-SURF dataset.

B. Attack type

We print the color pictures of the collectors with A4 paper to obtain the attack faces. In this way, each sample in the proposed dataset includes 1 live video clip and 6 fake video clips under different attack ways (one attack way per fake video clip). In the different attack styles, the printed flat or curved face images will be cut eyes, nose, mouth areas or their combinations. Finally, 6 attacks are generated in the CASIA-SURF dataset. 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 from the printed face.
- Attack 2: One person hold his/her curved face photo where eye regions are cut from the printed face.
- Attack 3: One person hold his/her flat face photo where eye and nose regions are cut from the printed face.
- Attack 4: One person hold his/her curved face photo where eye and nose regions are cut from the printed face.
- Attack 5: One person hold his/her flat face photo where eye, nose and mouth regions are cut from the printed face.
- Attack 6: One person hold his/her curved face photo where eye, nose and mouth regions are cut from the printed face.

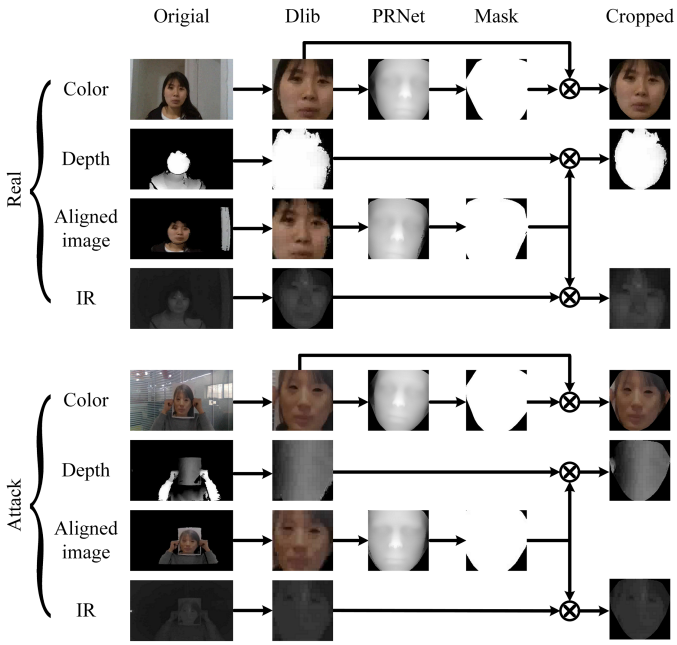


Fig. 4. Preprocessing details of three modalities of the CASIA-SURF dataset.

C. Data preprocessing

In order to create a challenging dataset, we remove the background except face areas from the original videos. Concretely, as shown in Fig. 4, the accurate face areas are obtained through the following steps. Given that we have a RGB-Depth-IR aligned video clip for each sample, we first use the Dlib [46] toolkit to detect face for every frame of RGB and RGB-Depth-IR aligned videos, respectively. The detected RGB and aligned faces are shown in the second column of Fig. 4. After face detection, we apply the PRNet [47] algorithm to perform 3D reconstruction and density alignment on the detected faces. The accurate face area (*i.e.*, face reconstruction area) is shown in the third column of Fig. 4. Then, we define a binary mask based on non-active face reconstruction area from previous steps. The binary masks of RGB and RGB-Depth-IR images are shown in the fourth column of Fig. 4. 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. The face images of three modalities (RGB, Depth, IR) are shown in the last column of Fig. 4.

D. Statistics description

Table II presents the main statistics of the proposed CASIA-SURF dataset. (1) There are $1,000$ subjects and each one has one live video clip and six fake video clips. Data contains variability in terms of gender, age, glasses/no glasses and indoor environments. (2) Data is divided into three subsets: training, validation and testing. The training, validation and testing subsets have 300, 100 and 600 subjects, respectively. Therefore, we have $6,300$ ($2,100$ per modality), $2,100$ (700 per modality), $12,600$ ($4,200$ per modality) videos for its corresponding subset. (3) From original videos, there are about

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

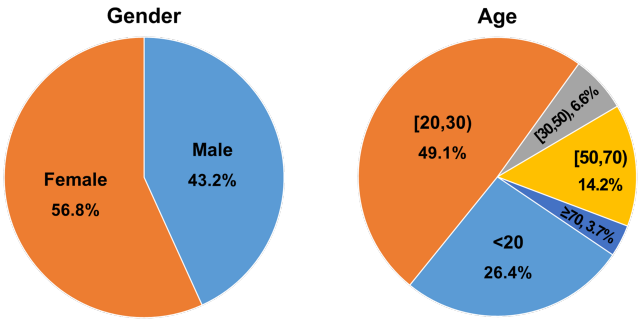


Fig. 5. Gender and age distribution of the CASIA-SURF dataset.

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.

All subjects are Chinese and 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 metric

Following the face recognition task, we use the ROC curve as the main evaluation metric for the proposed dataset. 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.

F. Evaluation protocol

To increase the difficulty, we select the live faces and Attacks 4, 5, 6 as the training subset, while select the live faces and Attacks 1, 2, 3 as the validation and testing subsets. The validation subset is used for model and hyper-parameter selection and the testing subset for final evaluation. There are two types of evaluation protocol in our dataset: (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.

IV. PROPOSED METHOD

Before showing some experimental analysis on the proposed dataset, we first built a strong baseline method. We aim at finding a straightforward architecture that provides good performance on the proposed CASIA-SURF dataset. 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 [48] 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.

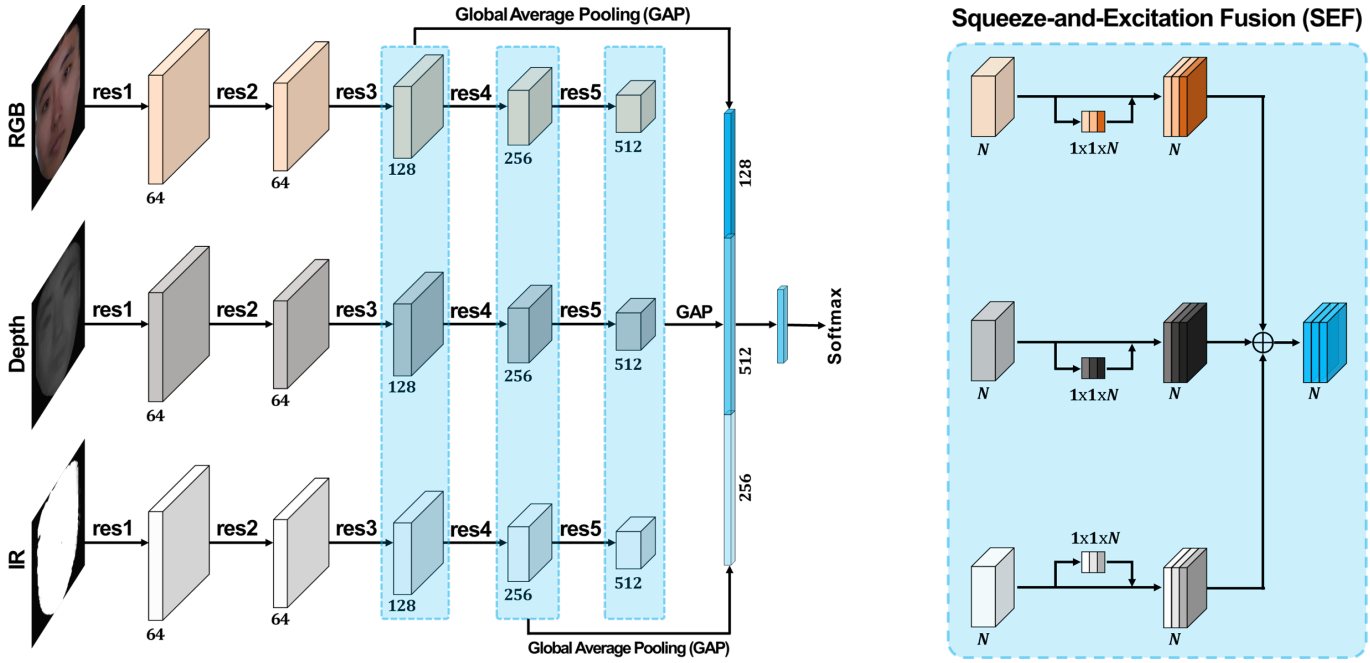
A. Naive halfway fusion

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 [49], 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.

[1]The branch is the same as the "Squeeze-and-Excitation" branch [49], composed of one global average pooling layer and two consecutive fully connected layers



(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.

C. From single-scale to multi-scale SEF

In our previous work [21], 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. The single-scale SEF is not able to make full use of features from different scales. 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 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 [21], 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 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)

TABLE III
EFFECTIVENESS OF THE PROPOSED METHOD. UNLESS OTHERWISE STATED, ALL MODELS ARE BASED ON RESNET-18 AND TRAINED ON THE CASIA-SURF TRAINING SUBSET AND TESTED ON THE TESTING SUBSET.

Method	TPR (%)			APCER (%)	NPCER (%)	ACER (%)
	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
Halfway fusion	89.1	33.6	17.8	5.6	3.8	4.7
SEF	96.7	81.8	56.8	3.8	1.0	2.4
+ Data augmentation	97.8	84.8	66.2	3.7	0.5	2.1
+ Addition operation	98.7	93.2	73.5	2.8	0.3	1.5
+ ImageNet pretrain	99.4	95.8	81.4	2.3	0.3	1.3
+ Multi-scale fusion	99.7	97.4	92.4	1.9	0.1	1.0
+ Stronger backbone	**99.8**	**98.4**	**95.2**	**1.6**	**0.08**	**0.8**

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.

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

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

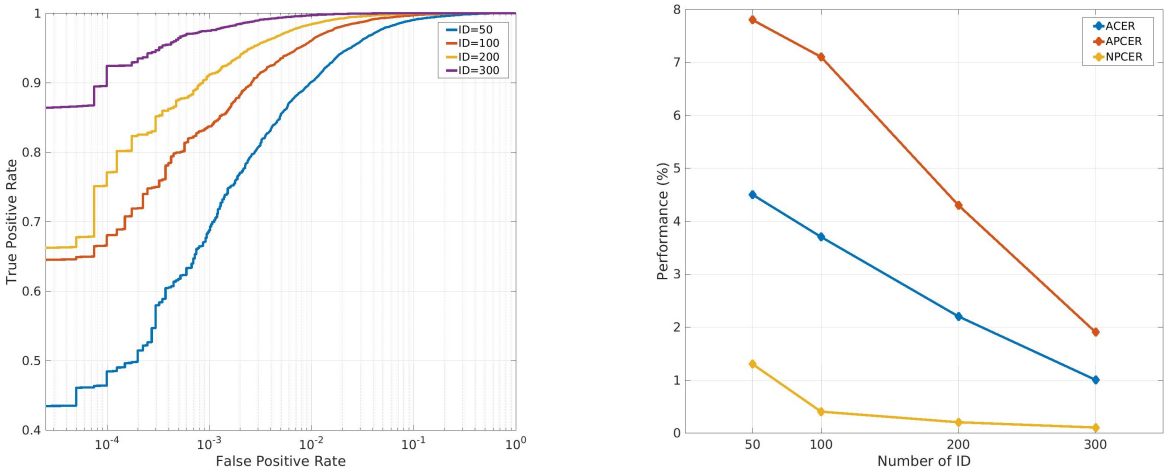
The proposed CASIA-SURF dataset has three modalities with $1,000$ subjects. In this subsection, we analyze the effect of the number of modalities and subjects.

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 the three ones, we observe an increase in performance. Specifically, 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 [50], there is a logarithmic relation between the amount of training data and the performance of deep neural network methods. 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 the 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. Specially, when FPR=10^{-4}, the best TPR of 300 subjects is higher about 15% than the second best TPR result (ID=200), showing the more data is used, the better performance will be. In Fig. 7(b), we also provide with the performance of ACER, APCER and NPCER when a different number of subjects is used for training. Their performances are getting better when more subjects are considered.

D. Cross-modal evaluation

Applications from the real world usually face an emergency problem that face images are captured from different modalities. This is the heterogeneous face recognition [51], [52] task, which involves matching two face images from alternate imaging modalities, such as an IR image to a RGB image



(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.

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

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

or a Depth image to a RGB image. In these applications, heterogeneous face anti-spoofing is needed, which means that face anti-spoofing algorithms are trained on one modality data and used on other modalities data. Thus, we introduce the cross-modal evaluation protocol for heterogeneous face anti-spoofing. In this protocol, models trained in one modality will be evaluated in other 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, heterogeneous face anti-spoofing is a challenging task and deserves further study in academic community.

E. Generalization capability

In this subsection, we evaluate the generalization capability of the proposed dataset on the Oulu-NPU [14], SiW [15] and CASIA-MFSD [9] datasets. 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 [15], [53]. Thus, we adopt FAS-TD-SF [53] as our baseline for the generalization experiments.

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. To verify the generalization capability of the proposed dataset, we first use the RGB and Depth images from our CASIA-SURF dataset to pre-train the FAS-TD-SF model, and then fine-tune it on the Oulu-NPU dataset. The results are 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 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 4 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 [15] for more details of the protocols. Two state-of-the-art methods (FAS-BAS [15] and FAS-TD-SF [53]) on the SiW dataset are selected for comparison. We verify the generalization capability of our dataset via pre-training FAS-TD-SF on CASIA-SURF

TABLE VI
EVALUATION RESULTS ON FOUR PROTOCOLS OF OULU-NPU.

Prot.	Method	APCER (%)	NPCER (%)	ACER (%)
1	CPqD [54]	2.9	10.8	6.9
	GRADIANT [54]	1.3	12.5	6.9
	FAS-BAS [15]	1.6	**1.6**	1.6
	FAS-Ds [16]	1.2	1.7	**1.5**
	FAS-TD-SF [53]	**0.8**	10.8	5.8
	FAS-TD-SF (CASIA-SURF)	2.7	2.5	2.6
2	MixedFASNet [54]	9.7	2.5	6.1
	FAS-Ds [16]	4.2	4.4	4.3
	FAS-BAS [15]	**2.7**	2.7	2.7
	GRADIANT [54]	3.1	1.9	2.5
	FAS-TD-SF [53]	3.6	3.8	3.7
	FAS-TD-SF (CASIA-SURF)	**2.7**	**1.6**	**2.2**
3	MixedFASNet [54]	5.3±6.7	7.8±5.5	6.5±4.6
	GRADIANT [54]	2.6±3.9	5.0±5.3	3.8±2.4
	FAS-Ds [16]	4.0±1.8	3.8±1.2	3.6±1.6
	FAS-BAS [15]	2.7±1.3	3.1±1.7	2.9±1.5
	FAS-TD-SF [53]	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 [54]	35.8±35.3	8.3±4.1	22.1±17.6
	GRADIANT [54]	**5.0±4.5**	15.0±7.1	10.0±5.0
	FAS-BAS [15]	9.3±5.6	10.4±6.0	9.5±6.0
	FAS-Ds [16]	5.1±6.3	6.1±5.1	**5.6±5.7**
	FAS-TD-SF [53]	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

TABLE VII
EVALUATION RESULTS ON THREE PROTOCOLS OF SIW.

Prot.	Method	APCER(%)	NPCER(%)	ACER(%)
1	FAS-BAS [15]	3.58	3.58	3.58
	FAS-TD-SF [53]	1.27	0.83	1.05
	FAS-TD-SF (CASIA-SURF)	1.27	0.33	**0.80**
2	FAS-BAS [15]	0.57±0.69	0.57±0.69	0.57±0.69
	FAS-TD-SF [53]	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 [15]	8.31±3.81	8.31±3.80	8.31±3.81
	FAS-TD-SF [53]	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 [55]	Repaly-Attack	CASIA-MFSD	47.9
LBP [55]	Repaly-Attack	CASIA-MFSD	57.6
Motion-Mag [27]	Repaly-Attack	CASIA-MFSD	47.0
Spectral cubes [56]	Repaly-Attack	CASIA-MFSD	50.0
CNN [45]	Repaly-Attack	CASIA-MFSD	45.5
FAS-TD-SF [53]	SiW	CASIA-MFSD	39.4
FAS-TD-SF [53]	**CASIA-SURF**	**CASIA-MFSD**	**37.3**

and then fine-tuning on SiW. Table VII shows the comparison of these three methods. 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 [27], [45], [55], [56] are listed for comparison, which use the Replay-Attack [8] 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.

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 increasing 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.

VIII. Acknowledgements

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

References

[1] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing deep face recognition with custom silicone masks," in *BTAS*, 2018.
[2] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Selective refinement network for high performance face detection," in *AAAI*, 2019.
[3] A. Mohammadi, S. Bhattacharjee, and S. Marcel, "Deeply vulnerable: a study of the robustness of face recognition to presentation attacks," *IET Biometrics*, 2017.
[4] X. Wang, S. Zhang, Z. Lei, S. Liu, X. Guo, and S. Z. Li, "Ensemble soft-margin softmax loss for image classification," in *IJCAI*, 2018.
[5] S. Zhang, L. Wen, H. Shi, Z. Lei, S. Lyu, and S. Z. Li, "Single-shot scale-aware network for real-time face detection," *IJCV*, 2019.
[6] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face spoofing detection using colour texture analysis," *TIFS*, 2016.
[7] ——, "Face antispoofing using speeded-up robust features and fisher vector encoding," *SPL*, 2017.
[8] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *BIOSIG*, 2012.
[9] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *ICB*, 2012.
[10] N. Erdogmus and S. Marcel, "Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect," in *BTAS*, 2014.
[11] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *TIFS*, 2015.
[12] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *BIOSIG*, 2016.
[13] I. Chingovska, N. Erdogmus, A. Anjos, and S. Marcel, "Face recognition systems under spoofing attacks," in *Face Recognition Across the Imaging Spectrum*, 2016.
[14] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *FG*, 2017.
[15] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *CVPR*, 2018.
[16] A. Jourabloo, Y. Liu, and X. Liu, "Face de-spoofing: Anti-spoofing via noise modeling," *arXiv*, 2018.
[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.
[18] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv*, 2014.
[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017.
[20] X. Wang, S. Wang, S. Zhang, T. Fu, H. Shi, and T. Mei, "Support vector guided softmax loss for face recognition," *arXiv*, 2018.
[21] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *CVPR*, 2019.
[22] Y. Kim, J. Na, S. Yoon, and J. Yi, "Masked fake face detection using radiance measurements," *JOSA A*, 2009.
[23] N. Kose and J.-L. Dugelay, "Countermeasure for the protection of face recognition systems against mask attacks," in *FG*, 2013.
[24] G. Pan, L. Sun, Z. Wu, and S. Lao, "Eyeblink-based anti-spoofing in face recognition from a generic webcamera," in *ICCV*, 2007.
[25] L. Wang, X. Ding, and C. Fang, "Face live detection method based on physiological motion analysis," *TST*, 2009.
[26] K. Kollreider, H. Fronthaler, and J. Bigun, "Verifying liveness by multiple experts in face biometrics," in *CVPR Workshops*, 2008.
[27] S. Bharadwaj, T. I. Dhamecha, M. Vatsa, and R. Singh, "Computationally efficient face spoofing detection with motion magnification," in *CVPR*, 2013.
[28] G. Pan, L. Sun, Z. Wu, and Y. Wang, "Monocular camera-based face liveness detection by combining eyeblink and scene context," *TCS*, 2011.
[29] J. Komulainen, A. Hadid, and M. Pietikainen, "Context based face anti-spoofing," in *BTAS*, 2013.
[30] T. Wang, J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection using 3d structure recovered from a single camera," in *ICB*, 2013.
[31] M. De Marsico, M. Nappi, D. Riccio, and J.-L. Dugelay, "Moving face spoofing detection via 3d projective invariants," in *ICB*, 2012.
[32] S. Kim, S. Yu, K. Kim, Y. Ban, and S. Lee, "Face liveness detection using variable focusing," in *ICB*, 2013.
[33] J. Li, Y. Wang, T. Tan, and A. K. Jain, "Live face detection based on the analysis of fourier spectra," *BTHI*, 2004.
[34] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, 2002.
[35] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *BIOSIG*, 2012.
[36] J. Yang, Z. Lei, S. Liao, and S. Z. Li, "Face liveness detection with component dependent descriptor." in *ICB*, 2013.
[37] J. Maatta, A. Hadid, and M. Pietikainen, "Face spoofing detection from single images using texture and local shape analysis," *IET biometrics*, 2012.
[38] W. R. Schwartz, A. Rocha, and H. Pedrini, "Face spoofing detection through partial least squares and low-level descriptors," in *IJCB*, 2011.
[39] R. Tronci, D. Muntoni, G. Fadda, M. Pili, N. Sirena, G. Murgia, M. Ristori, S. Ricerche, and F. Roli, "Fusion of multiple clues for photo-attack detection in face recognition systems," in *IJCB*, 2011.
[40] 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 *ICB*, 2013.
[41] J. Komulainen, A. Hadid, M. Pietikäinen, A. Anjos, and S. Marcel, "Complementary countermeasures for detecting scenic face spoofing attacks," in *ICB*, 2013.
[42] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung, and K.-W. Cheung, "Integration of image quality and motion cues for face anti-spoofing: A neural network approach," *JVCIR*, 2016.
[43] 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 *IPTA*, 2016.
[44] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *TIFS*, 2016.
[45] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *arXiv*, 2014.
[46] D. E. King, "Dlib-ml: A machine learning toolkit," *JMLR*, 2009.
[47] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *ECCV*, 2018.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.
[50] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *ICCV*, 2017.
[51] B. F. Klare and A. K. Jain, "Heterogeneous face recognition using kernel prototype similarities," *TPAMI*, 2012.
[52] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li, "Coupled discriminant analysis for heterogeneous face recognition," *TIFS*, 2012.
[53] Z. Wang, C. Zhao, Y. Qin, Q. Zhou, and Z. Lei, "Exploiting temporal and depth information for multi-frame face anti-spoofing," *arXiv*, 2018.
[54] Z. Boulkenafet, J. Komulainen, Z. Akhtar, A. Benlamoudi, and A. Hadid, "A competition on generalized software-based face presentation attack detection in mobile scenarios," in *IJCB*, 2017.
[55] 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 *ICB*, 2013.
[56] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *TIP*, 2015.