A Dataset and Benchmark for Large-scale Multi-modal Face Anti-spoofing

Shifeng Zhang[1*], Xiaobo Wang[2*], Ajian Liu[3], Chenxu Zhao[2],
Jun Wan[1†], Sergio Escalera[4], Hailin Shi[2], Zezheng Wang[5], Stan Z. Li[1,3]
[1]NLPR, CASIA, UCAS, China; [2]JD AI Research; [3]MUST, Macau, China
[4]Universitat de Barcelona, Computer Vision Center, Spain; [5]JD Finance
{shifeng.zhang,jun.wan,szli}@nlpr.ia.ac.cn, ajianliu92@gmail.com
{wangxiaobo8,zhaochenxu1,shihailin,wangzezheng1}@jd.com, sergio@maia.ub.es

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 visual 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 a measurement set, evaluation protocol and training/validation/testing subsets, developing a new benchmark for face anti-spoofing. Moreover, we present a new multi-modal fusion method as baseline, which performs feature re-weighting to select the more informative channel features while suppressing the less useful ones for each modal. 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/`.*

1. Introduction

Face anti-spoofing aims to determine whether the captured face of a face recognition system is real or fake. With the development of deep convolutional neural network (CNN), face recognition [2, 6, 34, 46, 52] has achieved near-perfect recognition performance and already has been applied in our daily life, such as phone unlock, access control, face payment, etc. However, these face recognition systems are prone to be attacked in various ways, including print attack, video replay attack and 2D/3D mask attack, which cause the recognition result to become unreliable. Therefore, face presentation attack detection (PAD) [3, 4] is a vital step to ensure that face recognition systems are in a safe reliable condition.

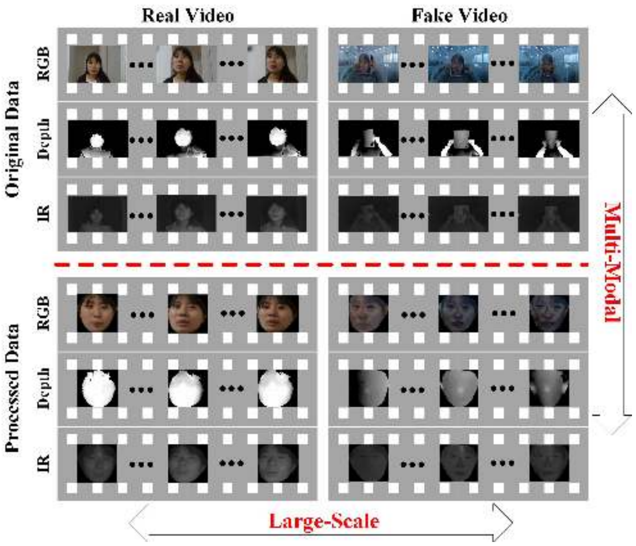

Figure 1. The CASIA-SURF dataset. It is a large-scale and multi-modal dataset for face anti-spoofing, consisting of $492,522$ images with 3 modalities (*i.e.*, RGB, Depth and IR).

Recently, face PAD algorithms [20, 32] have achieved great performances. One of the key points of this success is the availability of face anti-spoofing datasets [5, 7, 10, 32, 48, 53]. However, compared to the large existing image classification [14] and face recognition [51] datasets, face anti-spoofing datasets have less than 170 subjects and $60,00$ video clips, see Table 1. The limited number of subjects does not guarantee for the generalization capability required in real applications. Besides, from Table 1, another problem

[*] These authors contributed equally to this work
[†] Corresponding author

Dataset	Year	# of subjects	# of videos	Camera	Modal types	Spoof attacks
Replay-Attack [7]	2012	50	1200	VIS	RGB	Print, 2 Replay
CASIA-MFSD [53]	2012	50	600	VIS	RGB	Print, Replay
3DMAD [15]	2013	17	255	VIS/Kinect	RGB/Depth	3D Mask
MSU-MFSD [48]	2015	35	440	Phone/Laptop	RGB	Print, 2 Replay
Replay-Mobile [10]	2016	40	1030	VIS	RGB	Print, Replay
Msspoof [9]	2016	21	4704*	VIS/NIR	RGB/IR	Print
Oulu-NPU [5]	2017	55	5940	VIS	RGB	2 Print, 2 Replay
SiW [32]	2018	165	4620	VIS	RGB	2 Print, 4 Replay
CASIA-SURF (Ours)	2018	**1000**	**21000**	RealSense	**RGB/Depth/IR**	Print, Cut

Table 1. The comparison of the public face anti-spoofing datasets ($*$ indicates this dataset only contains images, not video clips).

is the limited number of data modalities. Most of the current datasets only have one modal (*e.g.*, RGB), and the existing available multi-modal datasets [15, 9] are scarce, including no more than 21 subjects.

In order to deal with previous 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 of the dataset are shown in Figure 1. As shown in Table 1, our dataset has two main advantages: (1) It is the largest one in term of number of subjects and videos; (2) The dataset is provided with three modalities (*i.e.*, RGB, Depth and IR).

Another open issue in face anti-spoofing is how performance should be computed. Many works [32, 20, 5, 10] adopt the attack presentation classification error rate (APCER), bona fide presentation classification error rate (BPCER) and average classification error rate (ACER) as the evaluation metric, in which APCER and BPCER are used to measure the error rate of fake or live samples, and ACER is the average of APCER and BPCER scores. However, in real applications, one may be more concerned about the false positive rate, *i.e.*, attacker is treated as real/live one. Inspired by face recognition [31, 45], the receiver operating characteristic (ROC) curve is introduced for large-scale face anti-spoofing in our dataset, which can be used to select a suitable trade off threshold between false positive rate (FPR) and true positive rate (TPR) according to the requirements of a given real application.

To sum up, the contributions of this paper are three-fold: (1) We present a large-scale multi-modal dataset for face anti-spoofing. It contains $1,000$ subjects, being at least 6 times larger than existing datasets, with three modalities. (2) We introduce a new multi-modal fusion method to effectively merge the involved three modalities, which performs modal-dependent feature re-weighting to select the more informative channel features while suppressing the less useful ones for each modality. (3) We conduct extensive experiments on the proposed CASIA-SURF dataset.

2. Related work

2.1. Datasets

Most of existing face anti-spoofing datasets only contain the RGB modalitiy. Replay-Attack [7] and CASIA-FASD [53] are two widely used PAD datasets. Even the recently released SiW [32] 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 [48], Replay-Mobile [10] and OULU-NPU [5].

As attack techniques are constantly upgraded, some new types of presentation attacks (PAs) have emerged, *e.g.*, 3D [15] and silicone masks [2]. These 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.* [23] aim to distinguish between the facial skin and mask materials by exploiting their reflectance. Kose *et al.* [28] propose a 2D+3D face mask attacks dataset to study the effects of mask attacks. However, associated data has not been made public. 3DMAD [15] 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 [9], containing visible (VIS) and near-infrared (NIR) images of real accesses and printed spoofing attacks with ≤ 21 objects.

However, existing datasets in the face PAD community have two common 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 [7, 53]. 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 PAs (3D and custom-made silicone masks).

2.2. Methods

Face anti-spoofing has been studied for decades. Some previous works [36, 43, 25, 1] attempt to detect the evidence of liveness (*i.e.*, eye-blinking). Another works are based on contextual [37, 26] and moving [44, 13, 22] information. To improve the robustness to illumination variation, some algorithms adopt HSV and YCbCr color spaces [3, 4], as well as Fourier spectrum [29]. All of these methods use handcrafted features, such as LBP [35, 8, 50, 33], HoG [50, 33, 40] and GLCM [40]. They are fast and 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*. [42] proposed a linear fusion of frame and video analysis. Schwartz *et al*. [40] introduced feature level fusion by using Partial Least Squares (PLS) regression based on a set of low-level feature descriptors. Other works [11, 27] obtained 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.

Recently, CNN-based methods [16, 30, 38, 49, 32, 20] have been presented 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*. [32] designed a network architecture to leverage two auxiliary information (Depth map and rPPG signal) as supervision. Amin *et al*. [20] introduced 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 exhibited a poor generalization ability on the cross-testing due to the over-fitting to training data. This problem remains open, although some works [30, 38] adopted transfer learning to train a CNN model from ImageNet [14]. These works show the need of a larger PAD dataset.

3. CASIA-SURF dataset

As commented, all existing datasets involve a reduced number of subjects and just one visual modality. Although the publicly available datasets have driven the development of face PAD and continue to be valuable tools for this community, their limited size 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 current limitations in PAD, we collected a new face PAD dataset, namely the CASIA-SURF dataset. To the best our knowledge, CASIA-SURF dataset is currently the largest face anti-spoofing dataset, containing $1,000$ Chinese people in $21,000$ videos. Another motivation in creating this dataset, beyond pushing the research on face anti-spoofing, is to explore recent face spoofing detection models performance when considering a large amount of data. In the proposed dataset, each sample 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 Figure 2. Detailed information of the 6 attacks is given below.

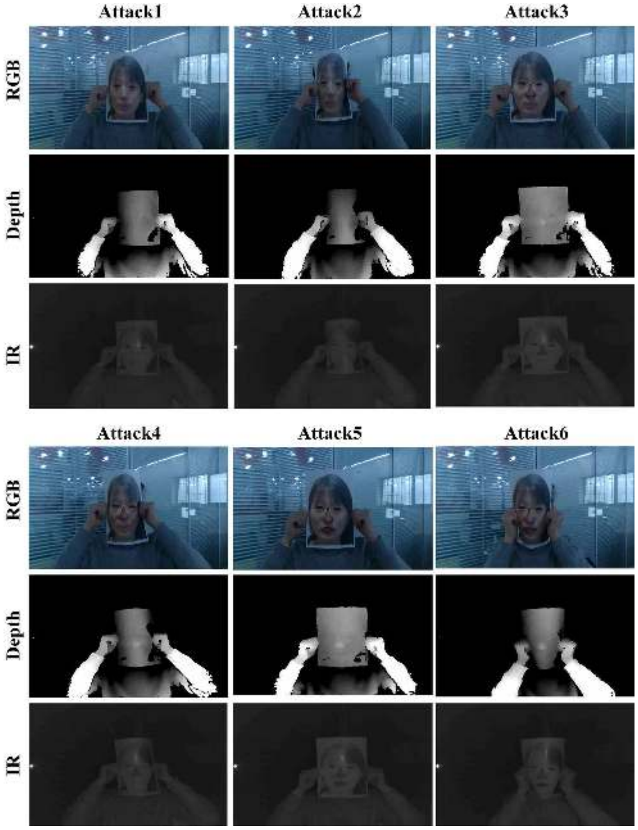

Figure 2. Six attack styles in the CASIA-SURF dataset.

- 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 eyes and nose regions are cut from the printed face.
- Attack 4: One person hold his/her curved face photo where eyes and nose regions are cut from the printed face.
- Attack 5: One person hold his/her flat face photo where eyes, nose and mouth regions are cut from the printed face.

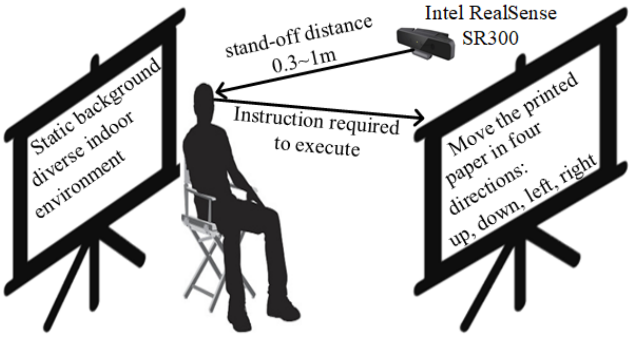
Intel RealSense SR300
stand-off distance 0.3~1m
Instruction required to execute
Static background diverse indoor environment
Move the printed paper in four directions: up, down, left, right

Figure 3. Illustrative sketch of recordings setups in the CASIA-SURF dataset.

- Attack 6: One person hold his/her curved face photo where eyes, nose and mouth regions are cut from the printed face.

3.1. Acquisition details

We used the Intel RealSense SR300 camera to capture the RGB, Depth and Infrared (IR) videos simultaneously. In order to obtain the attack faces, we printed the color pictures of the collectors with A4 paper. During the video recording, the collectors were required to do some actions, such as turn left or right, move up or down, walk in or away from the camera. Moreover, the face angle of performers were asked to be less 30^0. The performers stood within the range of 0.3 to 1.0 meter from the camera. The diagram of data acquisition procedure is shown in Figure 3, where it shows how the multi-modal data was recorded via Intel RealSence SR300 camera.

Four video streams including RGB, Depth and IR images were captured at the same time, plus the RGB-Depth-IR aligned images using RealSense SDK. The RGB, Depth, IR and aligned images are shown in the first column of Figure 4. The resolution is 1280×720 for RGB images, and 640×480 for Depth, IR and aligned images.

3.2. Data preprocessing

In order to create a challenging dataset, we removed the background except face areas from original videos. Concretely, as shown in Figure 4, the accurate face area is obtained through the following steps. Given that we have a RGB-Depth-IR aligned video clip for each sample, we first used Dlib [24] 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 Figure 4. After face detection, we applied the PRNet [17] algorithm to perform 3D reconstruction and density alignment on the detected faces. The accurate face area (namely, face reconstruction area) is shown in the third column of Figure 4. Then, we defined 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 Figure 4. Finally, we obtained face area of RGB image via pointwise product between RGB image and RGB binary mask. The Depth (or IR) area can be calculated via the pointwise product between Depth (or IR) image and RGB-Depth-IR binary mask. The face images of three modalities (RGB, Depth, IR) are shown in the last column of Figure 4.

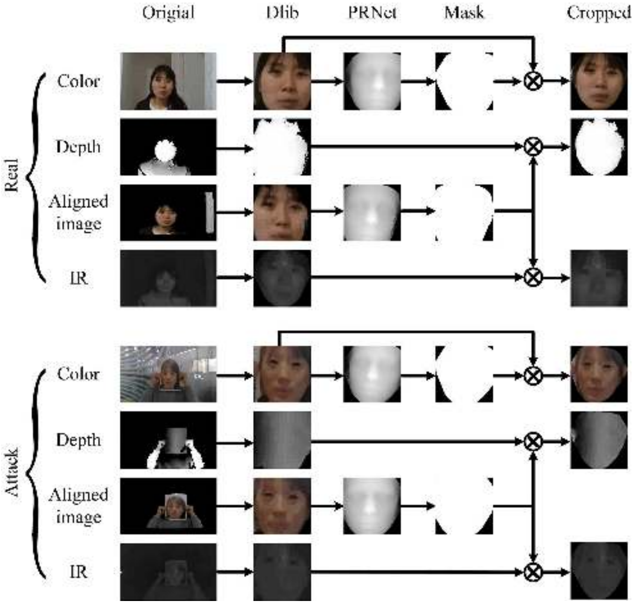

Figure 4. Preprocessing details of the three modalities of the CASIA-SURF dataset.

3.3. Statistics

Table 2 presents the main statistics of the proposed CASIA-SURF dataset:

(1) There are $1,000$ subjects and each one has a 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 split in three sets: training, validation and testing. The training, validation and testing sets 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 set.

	Training	Validation	Testing	Total
# Obj.	300	100	600	1000
# Videos	6,300	2,100	12,600	21000
# Ori. img.	1,563,919	501,886	3,109,985	5,175,790
# Samp. img.	151,635	49,770	302,559	503,964
# Crop. img.	148,089	48,789	295,644	492522

Table 2. Statistical information of the proposed CASIA-SURF dataset.

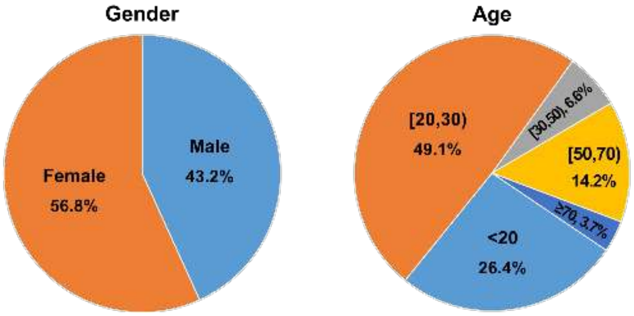

Figure 5. Gender and age distribution of the CASIA-SURF dataset.

(3) From original videos, there are about 1.5 million, 0.5 million, 3.1 million frames in total for training, validation, and testing sets, respectively. Owing to the huge amount of data, we select one frame out of every 10 frames and formed the sampled set with about $151K$, $49K$, and $302K$ for training, validation and testing sets, respectively.

(4) After data prepossessing in Sec. 3.2 and removing non-detected face poses with extreme lighting conditions, we finally obtained about $148K$, $48K$, $295K$ frames for training, validation and testing sets on the CASIA-SURF dataset, respectively.

All subjects are Chinese, and the information of gender statistics is shown in the left side of Figure5. 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.

3.4. Evaluation protocol

Intra-testing. For the intra-testing protocol, the live faces and Attacks 4, 5, 6 are used to train the models. Then, the live faces and Attacks 1, 2, 3 are used as the validation and testing sets. The validation set is used for model selection and the testing set for final evaluation. This protocol is used for the evaluation of face anti-spoofing methods under controlled conditions, where training and testing sets belong to the CASIA-SURF dataset. The main reason behind this selection of attack types in the training and testing sets is to increase the difficulty of the face anti-spoofing detection task. In this experiment, we show that there is still a big space to improve the performance under the ROC evaluation metric, especially, how to improve the true positive rate (TPR) at the small value of false positive rate (FPR), such as FPR=10^{-5}.

Cross-testing. The cross-testing protocol uses the training set of CASIA-SURF to train the deep models, which are then fine-tuned on the target training dataset (*e.g.*, the training set of SiW [32]). Finally, we test the fine-tuned model on the target testing set (*e.g.*, the testing set of SiW [32]). The cross-testing protocol aims at simulating performance in real application scenarios involving high variabilities in appearance and having a limited number of samples to train the model.

4. Method

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

4.1. 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.

4.2. 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 the IR data measure the amount of heat radiated from a face. Inspired by [19], we propose the squeeze and excitation fusion method that uses the “Squeeze-and-Excitation” branch to enhance the representational ability of the different modalities’ feature by explicitly modelling the interdependencies between their convolutional channels.

As shown in Figure 6, our squeeze and excitation fusion method has a three-stream architecture and each subnetwork is feed with the image of different modalities. The res1, res2 and res3 blocks from each stream extract features from different modalities. After that, these features

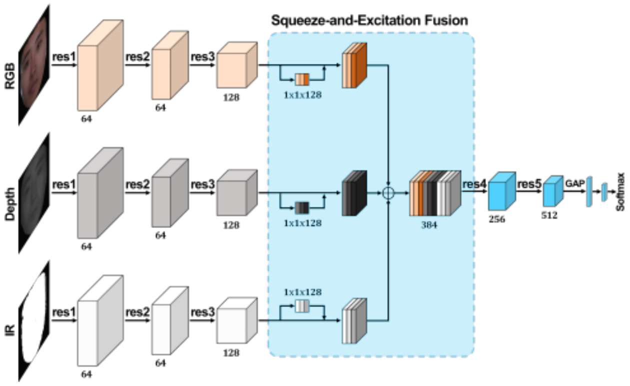

Figure 6. Diagram of the proposed fusion method. Each stream uses ResNet-18 as backbone, which has five convolution blocks (*i.e.*, res1, res2, res3, res4, res5). The res1, res2, and res3 blocks extract features of each modal data (*i.e.*, RGB, Depth, IR). Then, these features from different modalities are fused via the squeeze and excitation fusion module. Next, the res4 and res5 block are shared to learn more discriminative features from the fused one. GAP means the global average pooling.

are fused via the squeeze and excitation fusion module. This module newly adds a branch for each modal and the branch is composed of one global average pooling layer and two consecutive fully connected layers. The squeeze and excitation fusion module performs modal-dependent feature re-weighting to select the more informative channel features while suppressing less useful features from each modality. These re-weighted features are concatenated to define the fused multi-modal feature set.

5. Experiments

This section describes the implementation details, evaluates the effectiveness of the proposed fusion method, and presents a series of experiments to analyze the CASIA-SURF dataset in terms of modalities and number of subjects. Finally, the generalization capability of a baseline model trained with the CASIA-SURF dataset is evaluated/fine-tuned when tested on standard face anti-spoofing benchmarks.

5.1. Implementation details

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 $2,000$ iterations with 0.1 initial learning rate, and decreased by a factor of 10 after $1,000$ and $1,500$ iterations. All models are optimized via Stochastic Gradient Descent (SGD) 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.

5.2. Model analysis

We carry out an ablation experiment on the CASIA-SURF dataset to analyze our proposed fusion method. For evaluation, we use the same settings except for the fusion strategy to examine how the proposed method affects final performance. From the results listed in Table 3, it can be observed that the proposed fusion method achieves TPR=96.7%, 81.8%, 56.8% @FPR=10^{-2}, 10^{-3}, 10^{-4}, respectively, which are 7.6%, 48.2% and 39.0% higher than the halfway fusion method, especially at FPR=$10^{-3}, 10^{-4}$. Besides, the APCER, NPCER and ACER are also improved from 5.6%, 3.8% and 4.7% to 3.8%, 1.0% and 2.4%, respectively. Compared with halfway fusion method, we show the effectiveness of the proposed squeeze and excitation fusion method.

5.3. Dataset analysis

The proposed CASIA-SURF dataset has three modalities with $1,000$ subjects. In this subsection, we analyze modalities complementarity when training with a large number of subjects.

Effect on the number of modalities. As shown in Table 4, only using the prevailing RGB data, the results are TPR=49.3%, 16.6%, 6.8% @FPR=10^{-2}, 10^{-3}, 10^{-4}, 8.0% (APCER), 14.5% (NPCER) and 11.3% (ACER), respectively. In contrast, simply using the IR data, the results can be improved to TPR=65.3%, 26.5%, 10.9% @FPR=10^{-2}, 10^{-3}, 10^{-4}, 1.2% (NPCER) and 8.1% (ACER), respectively. Notably, from the numbers, one can observe that the APCER of the IR data increases by a large margin, from 8.0% to 15.0%. Among these three modalities, the Depth data achieves the best performance, *i.e.*, TPR=88.3%, 27.2%, 14.1% @FPR=10^{-2}, 10^{-3}, 10^{-4}, and 5.0% (ACER), respectively. 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 TPR=88.3%, 27.2%, 14.1% @FPR=10^{-2}, 10^{-3}, 10^{-4}, 5.1% (APCER), 1.2% (NPCER) and 5.0% (ACER) to TPR=96.7%, 81.8%, 56.8% @FPR=10^{-2}, 10^{-3}, 10^{-4}, 3.8% (APCER), 1.0% (NPCER) and 2.4% (ACER), respectively.

Effect on the number of subjects. As described in [41], 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 set. This is, we randomly select 50, 100 and 200 from 300 subjects for

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
Proposed fusion	**96.7**	**81.8**	**56.8**	**3.8**	**1.0**	**2.4**

Table 3. Effectiveness of the proposed fusion method. All models are trained in the CASIA-SURF training set and tested in the testing set.

Modal	TPR (%)			APCER (%)	NPCER (%)	ACER (%)
	@FPR=10^{-2}	@FPR=10^{-3}	@FPR=10^{-4}			
RGB	49.3	16.6	6.8	8.0	14.5	11.3
Depth	88.3	27.2	14.1	5.1	4.8	5.0
IR	65.3	26.5	10.9	15.0	1.2	8.1
RGB&Depth	86.1	49.5	10.6	4.3	5.6	5.0
RGB&IR	79.1	50.9	26.1	14.4	1.6	8.0
Depth&IR	89.7	71.4	24.3	**1.5**	8.4	4.9
RGB&Depth&IR	**96.7**	**81.8**	**56.8**	3.8	**1.0**	**2.4**

Table 4. Effect on the number of modalities. All models are trained in the CASIA-SURF training set and tested on the testing set.

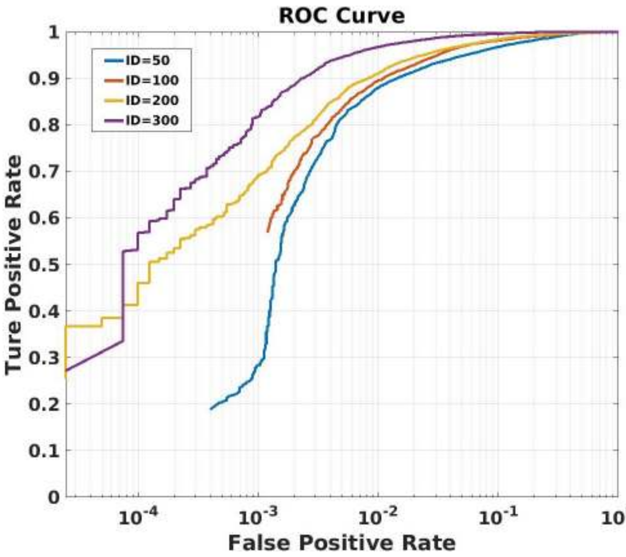

Figure 7. ROC curves of different training set size in the CASIA-SURF dataset.

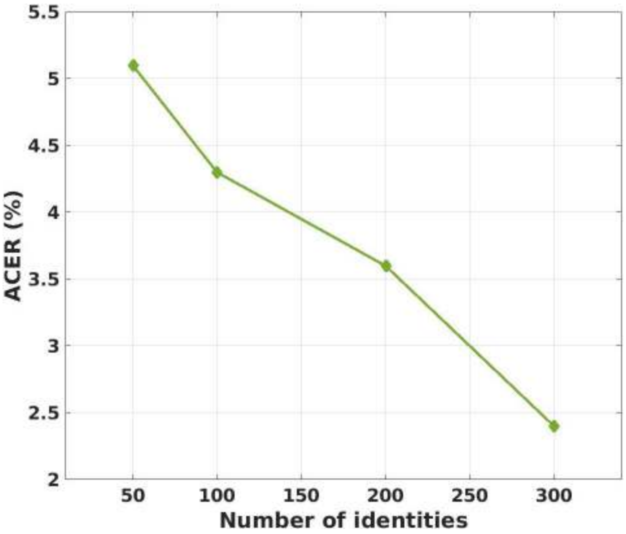

Figure 8. Performance *vs.* training set size in the CASIA-SURF dataset.

training. Figure7 shows the ROC curves for different number of subjects. We can see that when FPR is between 0 to 10^{-4}, the TPR is better when more subjects are used for training. Specially, when FPR=10^{-2}, the best TPR of 300 subjects is higher about 7% than the second best TPR result (ID=200), showing the more data is used, the better performance will be. In Figure 8, we also provide with the performance of APCER when a different number of subjects is used for training. The performance of ACER (average value of the fake and real error rates) is getting better when more subjects are considered.

5.4. Generalization capability

In this subsection, we evaluate the generalization capability of a model trained using the proposed dataset when tested/fine-tuned on the SiW [32] and CASIA-MFSD [53] 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 [32, 47]. Thus, we adopt FAS-TD-SF [47] as our baseline for the experiments.

SiW dataset. Two state-of-the-art methods (FAS-BAS [32] and FAS-TD-SF [47]) on the SiW dataset are selected for comparison. We use the RGB and Depth images from the proposed CASIA-SURF dataset to pre-train the FAS-TD-

SF CNN model, and then fine-tune it in the SiW dataset. Table 5 shows the comparison of these three methods. FAS-TD-SF generally achieves better performance than FAS-BAS, while our pre-trained FAS-TD-SF in CASIA-SURF (FAS-TD-SF-CASIA-SURF) can further improve the performance of PAD on both protocols[1] 1, 2 and 3. Concretely, the performance of ACER is superior about 0.25%, 0.14% and 1.38% when using the proposed CASIA-SURF dataset in Protocol 1, 2, and 3, respectively. The improvement indicates that pre-training in the CASIA-SURF 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.

Prot.	Method	APCER(%)	BPCER(%)	ACER(%)
1	FAS-BAS [32]	3.58	3.58	3.58
	FAS-TD-SF [47]	1.27	0.83	1.05
	FAS-TD-SF-CASIA-SURF	1.27	0.33	**0.80**
2	FAS-BAS [32]	0.57±0.69	0.57±0.69	0.57±0.69
	FAS-TD-SF [47]	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 [32]	8.31±3.81	8.31±3.80	8.31±3.81
	FAS-TD-SF [47]	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 5. Evaluation results in three protocols of SiW.

CASIA-MFSD dataset. Here, we perform cross-testing experiments on the CASIA-MFSD dataset to further evaluate the generalization capability of models trained with the proposed dataset. State-of-the-art models [12, 1, 39, 49] used for comparison used Replay-Attack [7] for training. We then train the FAS-TD-SF [47] in the SiW and CASIA-SURF datasets. Results in Table 6 show that the model trained in the CASIA-SURF dataset performs the best.

Method	Training	Testing	HTER (%)
Motion [12]	Repaly-Attack	CASIA-MFSD	47.9
LBP [12]	Repaly-Attack	CASIA-MFSD	57.6
Motion-Mag [1]	Repaly-Attack	CASIA-MFSD	47.0
Spectral cubes [39]	Repaly-Attack	CASIA-MFSD	50.0
CNN [49]	Repaly-Attack	CASIA-MFSD	45.5
FAS-TD-SF [47]	SiW	CASIA-MFSD	39.4
FAS-TD-SF [47]	**CASIA-SURF**	**CASIA-MFSD**	**37.3**

Table 6. Cross testing results on different cross-testing protocols.

6. Discussion

As shown in Table 3 and Table 4, accurate results were achieved in the CASIA-SURF dataset for traditional metrics, e.g. APCER=3.8%, NPCER=1.0%, ACER=2.4%. However, this shows an error rate of fake samples of 3.8% and an error rate of real samples of 1.0%. Thus, 3.8 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. Table 5 also demonstrates a similar performance in the SiW dataset. In order to push the state-of-the-art, in addition to large datasets, new evaluation metrics would be beneficial. The ROC curve is widely used in academic and industry for face recognition [31]. We consider the ROC curve to be also appropriated to be used as evaluation metric for face anti-spoofing.

As shown in Table 3 and Table 4, although the value of ACER is very promising, the TPR at different values of FPR is dramatically changing, being far from the standard required in real applications, e.g. when FPR=10^{-4} the TPR is 56.8%. Similar to the evaluation of face recognition algorithms, the TPR when FPR is about 10^{-4} or 10^{-5} would be meaningful for face anti-spoofing [21].

7. Conclusion

In this paper, we presented and released a large-scale multi-modal face anti-spoofing dataset. The CASIA-SURF dataset is the largest one in terms of number of subjects, data samples, and number of visual data modalities. We believe this dataset will push the state-of-the-art in face anti-spoofing. Owing to the large-scale learning, we found that traditional evaluation metrics in face anti-spoofing (*i.e.*, APCER, NPECR and ACER) did not clearly reflect the utility of models in real application scenarios. In this regard, we proposed the usage of the ROC curve as the evaluation metric for large-scale face anti-spoofing evaluation. Furthermore, we proposed a multi-modal fusion method, which performs modal-dependent feature re-weighting to select the more informative channel features while suppressing the less informative ones. Extensive experiments have been conducted on the CASIA-SURF dataset, showing high generalization capability of models trained on the proposed dataset and the benefit of using multiple visual modalities.

8. 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 JDGrapevine Plan in the JD AI Research, 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.

[1] For more details of the protocols, please refer to [32].

References

[1] Samarth Bharadwaj, Tejas I Dhamecha, Mayank Vatsa, and Richa Singh. Computationally efficient face spoofing detection with motion magnification. In *CVPR*, 2013.

[2] Sushil Bhattacharjee, Amir Mohammadi, and Sébastien Marcel. Spoofing deep face recognition with custom silicone masks. In *BTAS*, 2018.

[3] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *TIFS*, 2016.

[4] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face antispoofing using speeded-up robust features and fisher vector encoding. *SPL*, 2017.

[5] Zinelabinde Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. Oulu-npu: A mobile face presentation attack database with real-world variations. In *FG*, 2017.

[6] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Selective refinement network for high performance face detection. In *AAAI*, 2019.

[7] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, 2012.

[8] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, 2012.

[9] Ivana Chingovska, Nesli Erdogmus, André Anjos, and Sébastien Marcel. Face recognition systems under spoofing attacks. In *Face Recognition Across the Imaging Spectrum*. 2016.

[10] Artur Costa-Pazo, Sushil Bhattacharjee, Esteban Vazquez-Fernandez, and Sebastien Marcel. The replay-mobile face presentation-attack database. In *BIOSIG*, 2016.

[11] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *ICB*, 2013.

[12] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. Can face anti-spoofing countermeasures work in a real world scenario? In *ICB*, 2013.

[13] Maria De Marsico, Michele Nappi, Daniel Riccio, and Jean-Luc Dugelay. Moving face spoofing detection via 3d projective invariants. In *ICB*, 2012.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[15] Nesli Erdogmus and Sebastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *BTAS*, 2014.

[16] Litong Feng, Lai-Man Po, Yuming Li, Xuyuan Xu, Fang Yuan, Terence Chun-Ho Cheung, and Kwok-Wai Cheung. Integration of image quality and motion cues for face anti-spoofing: A neural network approach. *JVCIR*, 2016.

[17] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[20] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. *arXiv preprint*, 2018.

[21] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016.

[22] Sooyeon Kim, Sunjin Yu, Kwangtaek Kim, Yuseok Ban, and Sangyoun Lee. Face liveness detection using variable focusing. In *ICB*, 2013.

[23] Youngshin Kim, Jaekeun Na, Seongbeak Yoon, and Juneho Yi. Masked fake face detection using radiance measurements. *JOSA A*, 2009.

[24] Davis E. King. Dlib-ml: A machine learning toolkit. *JMLR*, 2009.

[25] Klaus Kollreider, Hartwig Fronthaler, and Josef Bigun. Verifying liveness by multiple experts in face biometrics. In *CVPR Workshops*, 2008.

[26] Jukka Komulainen, Abdenour Hadid, and Matti Pietikainen. Context based face anti-spoofing. In *BTAS*, 2013.

[27] Jukka Komulainen, Abdenour Hadid, Matti Pietikäinen, André Anjos, and Sébastien Marcel. Complementary countermeasures for detecting scenic face spoofing attacks. In *ICB*, 2013.

[28] Neslihan Kose and Jean-Luc Dugelay. Countermeasure for the protection of face recognition systems against mask attacks. In *FG*, 2013.

[29] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K Jain. Live face detection based on the analysis of fourier spectra. *Biometric Technology for Human Identification*, 2004.

[30] Lei Li, Xiaoyi Feng, Zinelabidine Boulkenafet, Zhaoqiang Xia, Mingming Li, and Abdenour Hadid. An original face anti-spoofing approach using partial convolutional neural network. In *IPTA*, 2016.

[31] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017.

[32] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, 2018.

[33] Jukka Maatta, Abdenour Hadid, and Matti Pietikainen. Face spoofing detection from single images using texture and local shape analysis. *IET biometrics*, 2012.

[34] Amir Mohammadi, Sushil Bhattacharjee, and Sébastien Marcel. Deeply vulnerable: a study of the robustness of face recognition to presentation attacks. *IET Biometrics*, 2017.

[35] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 2002.

[36] Gang Pan, Lin Sun, Zhaohui Wu, and Shihong Lao. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In *ICCV*, 2007.

[37] Gang Pan, Lin Sun, Zhaohui Wu, and Yueming Wang. Monocular camera-based face liveness detection by combining eyeblink and scene context. *Telecommunication Systems*, 2011.

[38] Keyurkumar Patel, Hu Han, and Anil K Jain. Secure face unlock: Spoof detection on smartphones. *TIFS*, 2016.

[39] Allan Pinto, Helio Pedrini, William Robson Schwartz, and Anderson Rocha. Face spoofing detection through visual codebooks of spectral temporal cubes. *TIP*, 2015.

[40] William Robson Schwartz, Anderson Rocha, and Helio Pedrini. Face spoofing detection through partial least squares and low-level descriptors. In *IJCB*, 2011.

[41] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *ICCV*, 2017.

[42] Roberto Tronci, Daniele Muntoni, Gianluca Fadda, Maurizio Pili, Nicola Sirena, Gabriele Murgia, Marco Ristori, Sardegna Ricerche, and Fabio Roli. Fusion of multiple clues for photo-attack detection in face recognition systems. In *IJCB*, 2011.

[43] Liting Wang, Xiaoqing Ding, and Chi Fang. Face live detection method based on physiological motion analysis. *Tsinghua Science & Technology*, 2009.

[44] Tao Wang, Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection using 3d structure recovered from a single camera. In *ICB*, 2013.

[45] Xiaobo Wang, Shuo Wang, Shifeng Zhang, Tianyu Fu, Hailin Shi, and Tao Mei. Support vector guided softmax loss for face recognition. *arXiv preprint*, 2018.

[46] Xiaobo Wang, Shifeng Zhang, Zhen Lei, Si Liu, Xiaojie Guo, and Stan Z Li. Ensemble soft-margin softmax loss for image classification. In *IJCAI*, 2018.

[47] Zezheng Wang, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, and Zhen Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. *arXiv preprint*, 2018.

[48] Di Wen, Hu Han, and Anil K Jain. Face spoof detection with image distortion analysis. *TIFS*, 2015.

[49] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint*, 2014.

[50] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In *ICB*, 2013.

[51] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint*, 2014.

[52] Shifeng Zhang, Longyin Wen, Hailin Shi, Zhen Lei, Siwei Lyu, and Stan Z Li. Single-shot scale-aware network for real-time face detection. *IJCV*, 2019.

[53] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z Li. A face antispoofing database with diverse attacks. In *ICB*, 2012.